Editor-in-Chief

Simon Asher Levin, Moffett Professor of Biology, Princeton University, Princeton, New Jersey, USA
Associate Editors

Robert Colwell, University of Connecticut, Storrs, Connecticut, USA

Gretchen Daily, Stanford University, Stanford, California, USA

Jane Lubchenco, Oregon State University, Corvallis, Oregon, USA

Harold A. Mooney, Stanford University, Stanford, California, USA

Ernst-Detlef Schulze, Universität Bayreuth, Bayreuth, Germany

G. David Tilman, University of Minnesota St. Paul, Minnesota, USA
International Editorial Advisors

Dan Cohen, Hebrew University of Jerusalem, Jerusalem, Israel

Rita R. Colwell, National Science Foundation, Arlington, Virginia, USA

Francesco di Castri, National Research Center of France, Montpellier, France

Paul R. Ehrlich, Stanford University, Stanford, California, USA

Thomas Eisner, Cornell University, Ithaca, New York, USA

Niles Eldredge, American Museum of Natural History, New York, New York, USA

Paul Falkowski, Rutgers University, New Brunswick, New Jersey, USA

Tom Fenchel, University of Copenhagen, Helsingoer, Denmark

Diana H. Wall, Colorado State University, Fort Collins, Colorado, USA

Madhav Gadgil, Indian Institute of Science, Bangalore, India

Stephen Jay Gould, Harvard University, Cambridge, Massachusetts, USA

Francesca Grifo, American Museum of Natural History, New York, New York, USA

Masahiko Higashi, Kyoto University (deceased), Kyoto, Japan

Yoh Iwasa, Kyushu University, Fukuoka, Japan

John H. Lawton, Imperial College at Silwood Park, Ascot, Berks, United Kingdom

Sir Robert May, University of Oxford, Oxford, United Kingdom

Ortwin Meyer, Universität Bayreuth Bayreuth, Germany

Norman Myers, Consultant in Environment and Development, Headington, Oxford, United Kingdom
Michael J. Novacek, *American Museum of Natural History, New York, New York, USA*

Sir Ghillean Prance, *Royal Botanic Gardens, Richmond, Surrey, United Kingdom*

Michael Rosenzweig, *University of Arizona, Tucson, Arizona, USA*

Nigel Stork, *Research Center for Tropical Rainforest, Ecology and Management Cairns, Queensland, Australia*

Monica G. Turner, *University of Wisconsin, Madison, Wisconsin, USA*

Marvalee H. Wake, *University of California, Berkeley, Berkeley, California, USA*

Brian H. Walker, *Commonwealth Scientific and Industrial Research Organization, Lyneham, Australia*

Edward O. Wilson, *Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts, USA*
Biodiversity is the totality of the inherited variation of all forms of life across all levels of variation, from ecosystem to species to gene. Soon after the term was introduced at the first National Forum on Biodiversity in 1986, and after it began its rapid spread around the world, there occurred a reconfiguration in the way much of the science of biology is conceived. Where previously comparative biology had been almost entirely focused on the fundamentals of classification, evolution, behavior, and ecology, now it was augmented by a wide range of analyses from the social sciences. Where taxonomy and biogeography had been marginalized through the middle half of the twentieth century, now they moved back toward center stage. And where extinction had been little more than a phenomenon recognized and lamented, now it became a major concern of science. Much of ecology shifted toward the study of biodiversity's role in the assembly and maintenance of ecosystems. A growing number of economists, political scientists, and bioethicists took up the issue as part of their scholarly agenda. From this mix the discipline of conservation biology was born, and the Society of Conservation Biology became one of the fastest growing organizations in modern science. In 1992 the Rio Summit catapulted biodiversity to global prominence, from which most of the nations of the world endorsed the Convention on Biodiversity and have since used it as a guideline for conservation programs.

The new biodiversity initiative gave organismic and evolutionary biologists a global mission worthy of their science. It confirmed for those who labored in the vineyards that, as medicine is to molecular and cellular biology, the environment is to organismic and evolutionary biology. The first is responsible for personal health, and the second for planetary health. The additional evidence adduced moreover made clear that we are in the midst of an episode of massive extinction, unprecedented since that closing of the Mesozoic Era—and that scientists must lead the attempt to save the Creation.

The articles in the Encyclopedia of Biodiversity are unusually eclectic, yet organized by a set of easily articulated goals. They are the following: to carry the systematics and biogeography of the world fauna and flora toward completion; map the hot spots where conservation will save the most biodiversity; orient studies of natural history to understand and save threatened species; advance ecosystems studies and biogeography to create the needed principles of community assembly and maintenance; acquire the knowledge of resource use, economics, and politics to advance conservation programs based on sustainability; and enrich the ethic of global conservation in terms persuasive to all.

The road ahead, down which we must urgently travel, will be smoothed by the exponential growth of information and a growing public awareness and support. The Encyclopedia of Biodiversity will serve as an important knowledge base to guide this supremely important effort.

Edward O. Wilson
Museum of Comparative Zoology
Harvard University
Preface

The science of biodiversity has become the science of our future. Our awareness of the disappearance of biodiversity has brought with it a long-overdue appreciation of the magnitude of our loss, and a determination to develop the tools to protect our future. This encyclopedia brings together, for the first time in its completeness, study of the dimensions of diversity with examination of the services that biodiversity provides, and measures to protect it.

The entries in the encyclopedia have been arranged alphabetically, but the coverage is designed functionally. At the core is a comprehensive survey of biodiversity, across taxonomic groups and ecological regions. The emergence of biodiversity is then placed in an evolutionary perspective, as background for an understanding of current trends. Particular attention is given to the loss of services—for example, in fisheries, forestry and climate mediation—that are derived from natural systems. These are placed in an economic framework through a comprehensive set of papers that address problems of valuation, costs, and benefits, and develop a framework for prioritizing actions. Finally, a review is given of institutions and other mechanisms that exist and are needed for the preservation of biodiversity and, with it, the services that humans derive from nature.

The background for understanding biodiversity is to be found in the fossil record, and in the evolutionary patterns and trends that it reveals. The encyclopedia hence discusses these patterns, the origins of biodiversity, the effects of geological events, the mechanisms of evolution, and the uniqueness of the evolutionary process, with implications for conservation and restoration. The essential processes in macroevolution are those of speciation and extinction, which together govern the dynamics of diversity at higher levels of organization. These are given extensive coverage, both from a mechanistic and from a historical perspective, and provide an essential context for understanding the rest of the contributions.

The classification of organisms into species and higher taxa, and the elucidation of the mechanisms of natural selection, were the essential intellectual advances that allowed the development of the science of biodiversity. Carl Linnaeus introduced a systematic framework for understanding phylogenies, which continues to provide the foundation for evolutionary studies today; and Charles Darwin’s great legacy—the theory of evolution by natural selection—is the essential organizing principle for understanding the processes that gave rise to the patterns Linnaeus recognized. The encyclopedia provides unmatched taxonomic coverage of the organization of diversity into taxonomic groups and complements that with an extensive examination of ecosystems by biogeographic region and by functional type. These chapters elucidate latitudinal trends, life zones, species-area relationships and the distribution of diversity within and among ecological communities.

Throughout the core chapters, there is a healthy balance between empirical facts and conceptual theories. Such theories help to illuminate principles that cross systems and levels of organization, and transform the study of biodiversity into a science. Basic ecological constructs, such as the habitat and the niche, are given extensive treatment, as are key ecological mechanisms such as competition, predation, herbivory, parasitism and mutualism. These treatments are complemented by exploration of fundamental evolutionary mechanisms related to local differentiation, aspect diversity, sex, and recombination, and especially theories of extinction.

With these foundational chapters in hand, one can turn to the contemporary problems in biodiversity and compare today’s rapid rates of change to the historical patterns. Key chapters examine agriculture, fisheries, and forests, their importance to human needs, and their status and trends in response to changing land-use patterns, population growth, overexploitation, and climate
change. Threatened and endangered species are discussed in detail, with relation to the consequences of the spread of invading species.

The utilization of nature's bounty for food, fiber and fuel provides some of the most obvious benefits of biodiversity to humanity. Equally important, however, are the things that are less well appreciated: the potential for the discovery of new pharmaceuticals that can improve human health, the role of biodiversity in pollinating crops and wild species, and the importance of natural systems in regulating climate, mediating nutrient fluxes, and sequestering carbon as well as toxic materials. Each of these services provides humans with direct and indirect benefits, and somehow we must find ways to weigh these benefits, along with the ethical and aesthetic values we place on natural systems and biodiversity, to provide priorities for action. Only recently have economists recognized the importance of such issues as intellectual challenges essential to our survival on the planet. Much of biodiversity is exploited by humanity as part of a global commons, in which one does not pay in fair measure for extracting parts, or affecting the commons otherwise through land use or pollution. Economists have come to realize, along with ecologists and others, the magnitude of the externalities involved. When such externalities are involved, the market does not function as it must to maintain the resource, and new measures are needed if the sustainability and resilience of these resources are to be preserved. There is as yet, however, no ecological equivalent to the power of financial institutions, such as the Federal Reserve Board in the United States, to modify individual incentives sufficiently to maintain regional or global stability in the system of interest. In this encyclopedia, some of the most enlightened and thoughtful economists turn their attention to the economic challenges, and discuss the mechanisms and institutions that might be needed.

Together, the state-of-the-art entries in this encyclopedia tell an exciting story of how biodiversity arose, continues to arise, and is maintained. It is a story of a complex, self-organizing system—the biosphere—whose pieces can be examined individually, but cannot be understood outside the context of the whole. It is also a story of the coevolution of the biosphere and Homo sapiens, the first species whose own activities can feed back to influence the evolution of the biosphere on time scales that could lead to its own demise. The articles in the encyclopedia can be used as material for a wide spectrum of courses, tracing the history of the emergence of biodiversity from its origins to the challenges we face today.

This has been a massive effort, but one of the most rewarding I have ever undertaken. So many people have played a role that it is difficult to know where to begin. The project began through the initiative of Scott Bentley at Academic Press, and then was managed flawlessly by Chris Morris at AP. I cannot recall ever having dealt with an editor who operated more professionally than Chris, who combined a true vision and enthusiasm for the project with a sense of economic realities and the energy and insight to make the whole project work. In this he was ably assisted by outstanding Academic Press staff, especially Naomi Henning, Nick Panissidi, and Ann Marie Martin. At every step, it has been a pleasure to work with Academic Press, and I especially single out Chris for his fantastic and scholarly efforts.

At the next stage, the Editorial Boards were terrific in generating and commenting on ideas, suggesting authors, and critiquing contributions. More than 400 authors then adopted our view of the importance of the project, accepted the task of writing, and produced timely and comprehensive articles that make this Encyclopedia like no other source available today. To all of these, I extend my thanks and congratualtions.

And finally, special acknowledgment and gratitude are due my wife, Carole, and my assistant, Amy Bordvik. Carole put up with the late nights and obsessiveness that were essential to the process, and Amy worked tirelessly and without complaint through the whole long process. To them, any expression of thanks is insufficient.

Simon Levin
Princeton University

References
Dedication

Dedicated to the memory of three encyclopedia authors, Takuya Abe, Masahiko Higashi, and Gary Polis, and their colleagues Shigeru Nakano and Michael Rose, who perished March 27, 2000 in a tragic boating accident while on a research trip in Baja California. Masahiko Higashi was also a member of the Board of International Editorial Advisors.
ACID RAIN
AND DEPOSITION

George R. Hendrey
Brookhaven National Laboratory

I. Acid Deposition
II. Causes of Acid Rain
III. Precipitation Chemistry
IV. Effects
V. Regulation

GLOSSARY

acid deposition The combination of acid rain plus dry deposition; a term preferred over "acid rain."
acid rain Rain, fog, snow, sleet, or hail with pH less than 5.6.
aerosols Fine particulate matter suspended in the atmosphere, with diameters less than 5.5 μm.
alkalinity The acid-neutralizing capacity (ANC) of water: \[ \text{ANC} = [\text{HCO}_3^-] + [\text{CO}_3^{2-}] + [\text{OH}^-] - [\text{H}^+] \]
cation exchange capacity The total of exchangeable cations that a soil can absorb.
dry deposition Deposition of dry pollutants from the atmosphere including gases and aerosols.
macrophytes Vascular plants, mosses, liverworts, and macro-algae.
metric ton 1000 kg.
periphyton Community of organisms dominated by algae growing on submerged surfaces.
phytoplankton Microscopic plants that live suspended in the water column.

I. ACID DEPOSITION

A. Acid Rain

In the mid-1970s the existence of highly acidic rain became widely known because it appeared to be reducing biodiversity through acidification of surface waters. This ecological problem was linked to emissions of compounds of sulfur and nitrogen from fuel combustion that are oxidized in the atmosphere to form sulfuric acid (H₂SO₄) and nitric acid (HNO₃) and related compounds that make precipitation very acidic, commonly referred to as "acid rain." Large, national-scale research projects have since found that over large areas of eastern North America and northern Europe, the deposition of these acids and related substances has led to extensive acidification of lakes and streams and the extinction of populations of fish from many surface waters. High-elevation forests are injured by acid deposition and buildings and monuments are corroded. Phenomena related to acid deposition reduce atmospheric visibility and impact human health. This knowledge has led to the regulation of air pollutants that is effective in reducing some of these problems. The most comprehensive
source of information on this subject is the report series of the U.S. National Acid Deposition Assessment Program (NAPAP) published in 1990.

B. Dry Deposition

Dry deposition occurs when, in the absence of condensed water droplets, acid-forming substances in the atmosphere are deposited as gases and dry particles. Dry deposition may be in the form of a gas, such as SO2, or in the form of a fine, dry aerosol particle such as ammonium sulfate [(NH4)2SO4]. In landscapes receiving this deposition, runoff water from acid rain adds to the dry-deposited materials, making the combination more acidic than the falling rain alone.

C. Acid Deposition

Acid deposition, a term preferred over acid rain, is the combination of acid rain plus dry deposition. The most important chemical species of acid deposition are hydrogen ion (H+), oxides of sulfur (SOx) and nitrogen (NOx), including the strong acid anions sulfate (SO42−), nitrate (NO3−), and chloride (Cl−), and ammonium (NH4+). These substances are dissolved in liquid water (rain or fog) or adsorbed onto frozen water (snow, sleet, or hail) so that the hydrogen ions (H+) are dissociated from the acid anions.

Controls on the emissions of SO2 already in place in both North America and Europe are reducing acid deposition. NOx emissions and deposition, however, continue to increase. With these two opposing trends, there has been only a slight decrease in the acidity of "acid rain."

D. Acidity and the pH Scale

Pure water is a very weak acid (H2O ⇌ H+ + OH-) and the concentrations of H+ and OH− are equal. The amount of H+ present in pure water under standard conditions (20°C, 1 atm pressure) is 1 ten-millionth of a gram of H+ in a liter of water (0.0000001 M), or 10−7 moles per liter of water (mol/liter). Acidity is measured on the pH scale expressed as the negative logarithm of the H+ concentration. Thus, pure water has a pH of 7. An acid concentration 10 times greater than pure water can occur if acid-forming anions are present. This solution will have one-millionth of a gram of H+ in water, or 10−6 mol/liter, and the pH is 6. Thus, each whole pH unit lower represents a 10-fold increase in acidity. Over most of the eastern United States and other areas receiving acid deposition the pH of rain is in the range 4.1–4.8. Of the anions associated with precipitation acidity, SO42− accounts for about 60% and NO3− for about 40%.

II. CAUSES OF ACID RAIN

Acid deposition has been occurring for a long time. In 1856, Robert Angus Smith, who was chief alkali inspector for Britain, wrote, "It has often been observed that the stones and bricks of buildings, especially under projecting parts, crumble more readily in large towns where coal is burnt. . . . I was led to attribute this effect to the slow but constant action of acid rain." Smith was concerned about air pollution and soot in Manchester, England. In the mid-nineteenth century, sulfuric fumes from the burning of coal in homes and factories reacted with water in the air to produce a dilute solution of sulfuric acid that attacked limestone and lime-based mortar in brickwork. Smith's acid rain problems tended to be local in scale. Chimneys in those days were low and their smoke spread out at low elevation across cities and towns. The problem that Smith described led to a gradual increase in the heights of smoke-stacks to allow the dissipation of smoke and fumes over larger areas, reducing the concentration from any particular source at ground level. This strategy for dealing with air pollutants in general prevailed into the middle of the twentieth century.

Today, electric utility plants account for about 70% of annual SO2 emissions and 30% of NOx emissions in the United States. Mobile sources (transportation) also contribute significantly to NOx emissions. More than 22 Tg (terragrams = 1 million metric tons) of SO2 are emitted into the atmosphere each year in the United States, and 180 Tg are emitted globally.

A. SO2

SO2 is the principal form of anthropogenic sulfur emission and it is released primarily by combustion of fossil fuels. SO2 dissolves in water droplets where it can be oxidized to H2SO4. This has a low vapor pressure and tends to form aerosol particles. These aerosols can form salts with Ca2+, Mg2+, or NH4+ and can become nuclei for the condensation of water and formation of clouds. The residence time of sulfur in the atmosphere is controlled by the processes that deposit it to the ground. About half of the sulfur burden of the atmosphere is removed by dry deposition, although the ratio of dry to wet deposition varies widely. The total amount of sulfur emitted into Earth's atmo-
sphere in 1985 (the reference year) was 90 Tg (calculated as elemental sulfur, equivalent to 180 Tg of SO2) from all sources (Fig. 1). By 1990, global anthropogenic emission of sulfur was 85 Tg (170 Tg as SO2). Emissions of SO2 in the United States peaked in 1977 at 32 Tg. By 1985, U.S. emissions of SO2 had declined to 25 Tg (Table I). The largest source of SO2 is electric power plants, accounting for 69% of U.S. SO2 emissions. More than 90% of these power plant emissions are from combustion of coal.

Natural sources of sulfur emissions globally contribute as much as 7% of total sulfur emissions. Dimethyl sulfide released from the oceans is oxidized in the atmosphere to sulfate and may account for 60% of these natural emissions. Volcanism (20%), decomposition processes in soils and plants (13%), and coastal wetlands (3%) are other sources. In eastern North America and northern Europe and Britain, natural sources of sulfur emissions are of little importance as sources of SOx and NOx, accounting for less than 1% of regional sulfur emissions according to Environmental Protection Agency (EPA) studies.

**TABLE I**

<table>
<thead>
<tr>
<th>Source</th>
<th>SO2 (Tg/yr)</th>
<th>NOx (Tg/yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electric utilities</td>
<td>14.6</td>
<td>6.13</td>
</tr>
<tr>
<td>Nonutility combustion</td>
<td>2.4</td>
<td>2.98</td>
</tr>
<tr>
<td>Nonferrous smelters</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Residential/commercial</td>
<td>0.6</td>
<td>0.64</td>
</tr>
<tr>
<td>Other industrial processes</td>
<td>2.1</td>
<td>0.63</td>
</tr>
<tr>
<td>Transportation</td>
<td>0.8</td>
<td>7.61</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>21.1</strong></td>
<td><strong>10.21</strong></td>
</tr>
</tbody>
</table>

*From NAPAP (1990).*

**B. NOx**

Human activities have more than doubled the emissions of fixed nitrogen to the atmosphere, surpassing the total of all natural sources. The primary form emitted by fuel combustion is NOx. The largest single anthropogenic source is the transportation sector (40%), with fossil-
fired utilities (30%) a close second. Anthropogenic emissions of NO\textsubscript{x} have risen more or less steadily from about 3 Tg released in the Year 1900. By 1985, 36–41 Tg of NO\textsubscript{x} was emitted globally (Fig. 2), with more than 22 Tg emitted in the United States alone. About 30% was from electric utilities and 43% from the transportation sector in that year. Natural sources of NO\textsubscript{x} to the atmosphere, which may contribute to the formation of NO\textsubscript{3}\textsuperscript{-} and H\textsubscript{2}SO\textsubscript{4}, are less well-known. There are natural sources of NO\textsubscript{x} emissions in soil, lightning, and stratospheric injections that account for 6, 5, and 0.1%, respectively, of the total of anthropogenic plus natural sources of NO\textsubscript{x} emissions in the United States according to NAPAP reports (1990).

NO\textsubscript{x} chemistry is complex and leads to the formation of nitric acid (HNO\textsubscript{3}). Nitric acid gas can react with aerosols such as sea salt, adsorb onto particles of soil, or react with ammonia to form NH\textsubscript{4}NO\textsubscript{3}. Ammonia is emitted to the atmosphere in urban and agricultural areas largely due to human activities. The rate at which nitric acid is deposited from the atmosphere as a dry gas is much faster than the deposition velocity of NH\textsubscript{4}NO\textsubscript{3}; thus, the presence of ammonia facilitates the long-range transport of NO\textsubscript{x}.

### III. PRECIPITATION CHEMISTRY

Wet deposition is relatively easy to collect and to evaluate. Most of the wet-deposited pollutants arrive in just a few major precipitation events. Dry deposition is a slower and more continuous process, but it is quite difficult to measure and local factors that alter wind turbulence and seasonal factors are important to the accuracy of measurements. On a regional basis, wet and dry deposition are approximately equal; but in urban areas or near to major emission zones dry deposition may be considerably greater than wet deposition.

Precipitation samples are collected in areas of the world that are remote from sources of SO\textsubscript{x} such asPoint Barrow in Alaska, Mauna Loa in Hawaii, and at the South Pole, by the Global Trends Network (GTN). In such remote areas, the average pH of precipitation is closer to 5.0 than to 5.6, which is the pH value that might be expected from an equilibrium of atmospheric CO\textsubscript{2} in pure water. Apparently, natural sources of acidity (e.g., oceanic or wetland emissions of sulfur) reduce pH below this expected value. It is also clear, however, that anthropogenic pollutants, SO\textsubscript{4}\textsuperscript{2-} and NO\textsubscript{3}\textsuperscript{-}, contrib-

![FIGURE 2](image-url) Annual nitrogen oxide emissions as nitrogen on a 1° × 1° latitude/longitude grid (1000 kg/year) (Canadian Global Emissions Interpretation Centre, a joint initiative of Canadian ORTECH Environmental, Inc., and Environment Canada). See also color insert, this volume.
ute to this acidification and no area of the world is free of anthropogenic pollutants. Eastern North America and northern Europe are receptor regions downwind from large area sources. Compared to the remote regions, these receptor regions receive nine times more \( \text{SO}_4^{2-} \), 14 times more \( \text{NO}_3^- \), seven times as much \( \text{NH}_4^+ \), and six times as much \( \text{H}^+ \). The sources of these contaminants are the upwind emissions from industrial and population centers. NAPAP (1990) reports that all forms of precipitation over much of eastern North America, on average, are quite acidic (Fig. 3). Mean annual “wet” precipitation (weighted by the volume of each precipitation event) was in the range pH 5.0–4. Individual rain episodes with pH near 3.0 are observed in the northeastern United States. There is great variability in the amount deposited across continental areas. For example, the average annual deposition of sulfur species (Fig. 4) at Argonne, Illinois (in 1983–1987), was 23.6 kg/ha, whereas at Pawnee, Colorado, it was 1.7 kg/ha. In northern Europe, including Britain, all of Scandinavia in the north and down to mid-France and northern Italy, and east to the border of Russia, the annual average pH of precipitation was below 4.9 in 1985.

\( \text{HNO}_3 \) is deposited as a dry gas from the atmosphere onto vegetation or other moist surfaces. The rate of deposition velocity of the nitrate aerosol, ammonium nitrate (\( \text{NH}_4\text{NO}_3 \)). Thus, the presence of ammonia facilitates the long-range transport of \( \text{NO}_x \) (Fig. 5).

IV. EFFECTS

A. Forests

There are numerous examples of forest dieback related to local sources of pollution. For example, \( \text{SO}_2 \) emissions at near-ground level from a copper smelter in Sudbury, Ontario, killed forests, grasses, and soil organisms and created a local landscape that some called a moonscape. Similar situations exist around point emission sources elsewhere. Acid deposition, however, is a problem associated with the long-range transport of
pollutants, with receptor areas hundreds or even thousands of kilometers from the emission sources.

Throughout Europe, forests are plagued by thinning of the topmost branches, called “crown-thinning.” In Eastern Europe, where high-sulfur coal has been consumed in prodigious quantities, areas in which high concentrations of atmospheric pollutants (especially SO₂) occur have undergone significant forest injury and dieback, although the extent of damage is not well quantified. In the western part of Germany, many forest declines appeared to be occurring in the 1970s that were attributed in the popular press to acid rain. The term “Waldsterben” (forest death) was used to describe the situation. This led to increased public pressure for environmental protection in general and for research on topics relating to acid deposition in particular. Large-scale surveys of forest condition were carried out in Germany where there were many regional declines. The overall conclusion of the surveys, however, was that less than 20% of western German forest area was classified as damaged and no large-scale deforestation was occurring. In fact, it was found that the rate of forest stress seemed to be decreasing and surveys showed that Norway spruce (Picea abies) injury was reported to be 9% less in 1988 than in 1983.

Partly due to forest problems in Germany, it was suspected that forest productivity and the health of forest in North American ecosystems might be compromised by acidification, acting either directly on vegetation or through changes in forest soils. In Canada, 49% of the land is covered with forests, as are 33% of the United States and 21% of Mexico. NAPAP organized a Forest Response Program in 1985 to address issues of forest damage in general and the role that acid deposition might play in such damage. Similar research activities were carried out in other countries, including Canada and Norway, in which forests cover extensive areas in regions most heavily impacted by acid deposition.

It is known that forests are impacted by a variety of stresses and it is often difficult to isolate specific causes of local forest decline. Fire, insect pests, microbial infestations, poor management practices, and even natural aging can act alone or together (Table II) in causing forest decline. Severe forest declines have occurred in the past. For example, during the period 1871–1885 an estimated 50% of mature spruce trees in the Adirondack Mountains died from unknown causes. Another example is that of “fir waves,” in which patches of balsam fir (Abies balsamea) in the Appalachian region die out.
in a wave-like pattern across the landscape. It was thought that a potentially stressing condition, such as heavy loading of a region by acid deposition, might cause a general weakened condition that makes forests more susceptible to other problems (Barnard et al., 1990).

In the United States red spruce (Picea rubens Sarg.), high-elevation spruce trees that populate the ridges of

<table>
<thead>
<tr>
<th>Stress</th>
<th>Affected area (10^6 ha)</th>
<th>Nature of impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wildfire</td>
<td>1.7</td>
<td>$730 million damage to timber resource in 1985</td>
</tr>
<tr>
<td>Gypsy moth</td>
<td>0.7</td>
<td>Tree defoliation concentrated in southern New England and mid-Atlantic states</td>
</tr>
<tr>
<td>Mountain pine beetle</td>
<td>&gt;1.6</td>
<td>Mortality often severe in infested stands</td>
</tr>
<tr>
<td>Southern pine beetle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spruce bud worm</td>
<td>2.1</td>
<td>Tree defoliation and mortality concentrated in Maine</td>
</tr>
<tr>
<td>Western spruce bud worm</td>
<td>5.1</td>
<td>Tree defoliation and mortality, primarily in the Rockies, intermountain region, and the Pacific Northwest</td>
</tr>
<tr>
<td>Fusarium rust</td>
<td>6.1</td>
<td>$35 million in lost timber value through volume and quality reduction in southern pines</td>
</tr>
<tr>
<td>Dwarf mistletoe</td>
<td>9.0</td>
<td>Annual timber volume loss of 10.8 million m^3 in the West</td>
</tr>
<tr>
<td>Root diseases</td>
<td>&gt;6.0</td>
<td>Annual timber value loss of 6.6 million m^3 in the West; substantial but unquantified impact in other regions</td>
</tr>
</tbody>
</table>

the Appalachian Mountains from Maine to Georgia, has undergone a period of dieback of 25–50% from the 1930s through 1989. The dieback is associated with severe winter injury that kills the terminal and lateral shoots, and the repetition of this injury can lead to overall stress, susceptibility to injury from fungi and insects, reduced growth, and tree death. NAPAP concluded that acid deposition has contributed to this dieback, but the mechanisms of injury are uncertain.

Many experimental studies of acid deposition effects on trees have been conducted, including exposure to SO₂, ammonium-sulfate aerosols, artificial acid rain, or acid rain plus ozone. In general, these experimental studies did not show that a significant negative effect that might stress forests or reduce forest growth was caused by pollutant levels similar to those associated with acid deposition. In reviewing all the evidence concerning acid deposition effects on forests, NAPAP concluded that (i) most forests in eastern North America are exposed to acidic deposition and to elevated concentrations of ozone but do not show signs of unusual growth loss or tree decline; (ii) spatial and temporal patterns of tree health and productivity are not consistently related to estimated levels of pollutant exposure; and (iii) except for red spruce at high elevations, there is no general deterioration in the health or productivity of eastern forests and no consistent relationship between forest health and atmospheric deposition. It was also concluded, however, that there were indications of stress to forests that should be monitored carefully, including the following: (i) Ambient ozone levels are affecting plant physiology in some species, (ii) acid deposition and ozone can interact in tree injury, and (iii) alteration of forest element cycles may affect species composition.

B. Crops

Scientists have been studying the effects of air pollutants on plants for many years. Research in this area accelerated after the discovery of ozone as a constituent in the atmosphere. (Effects of exposure of plants to elevated ozone, although related to the issue of acid rain, are not included in the context of this article.) There are many examples of plant injury due to acute fumigation by air pollutants from local sources. Whereas there is convincing experimental evidence that acid deposition can damage crop plants, reports of crop loss due to acid deposition (excluding local point sources of fumigation) are as scarce as hens’ teeth. Since the mid-1960s, more than 5000 research reports have been published dealing with the topic, including acute and chronic exposures of plants to SO₂, various forms of NOₓ, and elevated H⁺ as well as other contaminants. This extensive research effort is probably more a reflection of the importance of crops to both human nutrition and the agricultural economics than to any observation of crop loss due to long-range transport of pollutants. The conclusions reached by NAPAP (1990) are as follows: (i) Conditions capable of causing acute injury to vegetation as a result of exposure to present-day levels of gaseous air pollutants (including ozone) are rare, occur only during unusual conditions of atmospheric stability, and are confined to a limited number of areas, and (ii) acute injury to vegetation due to acidic deposition is virtually unknown.

C. Soils

The characteristics of natural terrestrial ecosystems are largely determined by the properties of their soils. In northern forest soils there is a layer of humus containing exchangeable bases overlying the mineral soil. When H⁺ is deposited with acid deposition into the forest, or generated by growth of vegetation, the humus layer and soil minerals can retain this H⁺ and release an equivalent amount of base-forming cations (Ca²⁺ and Mg²⁺) that generate alkalinity. In this way, much of the acid input may be neutralized for as long as the exchange capacity lasts. In sufficiently moist areas precipitation can leach base-forming cations as fast as the rate of primary mineral weathering of these ions, causing soils to become acidic due to natural processes. As soil acidity increases, aluminum (Al) becomes increasingly soluble but also adsorbs onto clay minerals, and soil solution Al can hydrolyze to increase soil H⁺ concentrations.

Acid deposition will have an acidifying effect or will leach base cations (Ca²⁺, Mg²⁺, K⁺, and Na⁺) or both in soils with low cation exchange capacity (CEC). In areas where sulfate adsorption in soils is low, such as in the Adirondack Mountains, sulfur deposited from the atmosphere behaves more or less conservatively and passes through soils into lakes and streams as SO₂. However, the total amount of strong acid anion must be balanced by an equivalent amount of cations. If CEC is depleted, then Al³⁺ is mobilized by H⁺ exchange. Al³⁺ and some of the H⁺ associated with acid rain enter the runoff water, thus maintaining the charge balance, and the water is acidified.

Recent studies by the U.S. Geological Survey (1999b) and others have found that calcium is being depleted...
from forest soils in the eastern United States as a consequence of both acid deposition and uptake by roots. In agricultural soils, agronomic practices of tilling, fertilization, and liming are far more important factors in altering soil chemistry than acid deposition. Additions of HNO₃ by acid deposition even may be beneficial for forest growth since nitrogen is frequently a growth-limiting nutrient though benefits from H₂SO₄ deposition are viewed as minimal.

D. Ground-water

The total atmospheric load of acids exceeds the ability of soils to provide bases in many areas of Europe. This is especially true in areas underlain by slow-weathering, base-deficient rocks, such as granite, gneiss, quartzite, and sandstone. In these areas with continued acid deposition, base saturation (the fraction of CEC occupied by exchangeable base cations) can be expected to decrease steadily and eventually approach zero. This will result in acidification of ground-water.

Norway is highly impacted by acid deposition, with average precipitation pH in the range 4.3–4.5. Ground-water in many areas is unusually acidic, with pH in the range 5.2–5.7. Such water is quite corrosive for copper pipes. In Denmark, the pH of deep well water decreased from 6.5 to 5.6 between the 1950s and 1980s. The Hartz Mountains of Germany also receive very acidic precipitation. Sulfate concentrations in ground-water there have risen from 5 mg/liter in the 1960s to a current value near 20 mg/liter and the water from 33 springs has high concentrations of metals (cadmium, 0.1–2.0 μg/liter; zinc, 50–150 μg/liter; and nickel, 5–20 μg/liter). In some cases, spring-water pH is less than 4.

E. Surface Waters

Acidification of surface waters is defined as a decrease in alkalinity, or acid neutralizing capacity (ANC). As acids are added to water, the H⁺ increases and ANC is reduced. The most significant impact of acid deposition is that on surface waters, in which it causes acidification and ecological damage in many thousands of lakes and streams. In some sensitive waters, fish species such as brook trout have been completely eradicated. This is one of the few environmental impacts that have been clearly demonstrated for “acid rain”, and it is politically important because it has resulted in the loss of fisheries and recreational value, which people can readily understand.

Unpolluted surface waters sensitive to acidification generally are found to be in the range pH 6–7 with low ANC. Watersheds with significant amounts of carbonate minerals can readily buffer inputs of acid by creating alkalinity. Watersheds with soils low in minerals with Ca²⁺ or Mg²⁺ have little ability to generate ANC. When unpolluted waters have ANC < 100 microequivalents per liter (μeq/liter) they are classified as sensitive to acidification.

Most surface water acidification is due to the deposition of sulfate that provides a long-term and rather steady base-load of strong acid anions. Nitrate is important in the episodic acidification associated with particular precipitation events or snowmelt, which can dramatically increase acidity of lakes and streams. In “brown” waters, the concentration of acids, humic materials is high and also contributes to acidity. However, acidification of surface waters—that is, the change in chemistry over time in many “sensitive” areas of the world—is clearly the result of excessive SO₄²⁻ concentrations due to acid deposition and not a consequence of the presence of natural organic acids.

Several factors interact to make the waters or a region susceptible to acidification due to inputs of strong acids. The most important of these are (i) proximity to emissions sources, (ii) regional meteorological patterns, (iii) bedrock geology, and (iv) topography (NAPAP, 1990). The most heavily impacted regions are located downwind from large emissions sources in Great Britain and northern Europe and of the central industrial region in the United States and Canada. The importance of geographic location and wind direction is illustrated by the fact that the very sensitive waters of northern Minnesota and southwestern Ontario are not yet as severely impacted as are the waters of the Northeast and maritime provinces. There is evidence that acidification is occurring in this region. Acid deposition is emerging as a significant problem in Asia, but there is scant information available on actual or potential ecological consequences there.

1. Surface Water Surveys

Large-scale, statistically based surveys of lakes and streams have been conducted in several countries to evaluate actual and potential impacts of acid deposition on surface waters. In most cases, these surveys were designed to investigate waters in regions thought to be sensitive to acidification because of high rates of acid deposition or in regions having waters with low ANC. In these sensitive areas, concentration of SO₄²⁻ is strongly correlated to wet SO₂ deposition. The U.S. National Surface Water Survey found that
all of the sampled lakes and streams with pH < 5.5 or ANC < 0 occur in areas receiving precipitation with pH < 5.0 and wet SO4\(^{2-}\) deposition less than 10 kg/ha/year. Furthermore, acidic lakes in which SO4\(^{2-}\) is the dominant anion are not found in regions receiving wet SO4\(^{2-}\) deposition less than 10 kg/ha/year. The Norwegian national acid deposition effects project found that in southern Norway, where acid deposition is great and soils are both thin and base deficient, 75% of the lakes are acidic and SO4\(^{2-}\) from acid deposition is the dominant anion.

A survey of 85106 lakes was carried out in 10 regions of Canada and 50% of these lakes were found to be sensitive to acidification (ANC < 100 μeq/liter). In some areas, up to 84% of the lakes were found to be sensitive and as many as 60% of the lakes in some areas were very sensitive (ANC < 50 μeq/liter). Acidic lakes, those with ANC < 0 μeq/liter, comprised 5% of all lakes in the sample, and up to 24% in one region were acidic.

Nitrogen as well as sulfur deposition can contribute to chronic and episodic acidification of surface waters. Unlike SO4\(^{2-}\), however, NO3\(^{-}\) is usually conserved within watersheds because of plant uptake of N. Exceptions to this rule, however, are seen in those areas of the world in which NO3\(^{-}\) deposition is unusually great. In streams of southwestern Norway, NO3\(^{-}\) concentrations exceed 10 μeq/liter and nitrate can make up over 10% of strong acid anions.

Trends in precipitation and stream-water chemistry were examined at eight precipitation monitoring stations during the period 1984–1996 by the U.S. Geological Survey (1999a). In the northeastern United States, results indicate that decreases in atmospheric deposition of SO4\(^{2-}\) have resulted in decreased precipitation acidity.

2. Episodic Acidification

Episodic acidification of surface waters occurs as a consequence of acidic snowmelt and acidic rain events. Snow accumulation is one mechanism by which strong acid anions may be stored and concentrated within watersheds receiving acid deposition. Accumulated contaminants in the winter snowpack can be released at the onset of melting so that 30–80% of the SO4\(^{2-}\) received over a period of months may be released from the snow-pack with the first 30% of the meltwater. Thus, early snowmelt runoff waters in areas such as southern Norway and the Adirondack Mountains carry pollutant loads that are greatly elevated. Regions with heavy snowfall can be especially susceptible if the rate of acid deposition is high.

F. Marine Waters

In some areas, the amount of nitrogen in soils, from agricultural fertilizers and acid deposition, exceeds the needs of vegetation and NO3\(^{-}\) is discharged in surface waters. In Scandinavia, NO3\(^{-}\) accumulated in the snow-pack is discharged so quickly by melt-water that it is not taken up by vegetation. In such cases, rivers carry NO3\(^{-}\) to estuaries and bays. Acid deposition also falls directly on marine waters, increasing the loading of nitrate. Chesapeake Bay, for example, receives 30–50% of its nitrogen from acid deposition and this contributes to eutrophication. In Scandinavia, acid deposition has contributed to excessive nitrogen in marine waters that appears to cause phytoplankton blooms.

G. Aquatic Biota

Acid deposition, by acidifying surface waters, causes widespread ecological damage (Table III). There is a widespread misconception, however, that acidified lakes and streams are "dead." The fact is that even the most acidified surface waters have many organisms. Species of protozoa and insects are found at pH 2.0, rotifers and Cladocera occur at pH 3.0, and even some fish are found at pH 3.5. Acidified waters are not "dead;" they can be full of life—but this is life run amok in ecosystems severely out of balance.

1. Microbial Communities

Abnormal accumulations of coarse organic matter are observed on the bottoms of some acidified lakes and dense felt-like mats of fungal hyphae can cover much of the bottom areas. The accumulation of debris and fungal mass both seal off the mineral sediments from interaction with the overlying water and hold organically bound nutrients that would have become mineralized and available if normal decomposition had occurred. Reductions in nutrient availability may have a negative feedback effect on microorganisms, further inhibiting their activities. Acidification can also inhibit microbial nitrogen cycle activities. Reduction of microbial decomposer activities may also have a direct effect on invertebrates feeding on microbial biomass associated with decomposing litter, further inhibiting litter removal and nutrient regeneration. Bacteria respond to acidification gradually, with no clearly delineated thresholds above pH 3.5. Treatment of lakes with lime raises pH and causes rapid decomposition of the organic debris and fungal mat and increases in bacteria in the water, indicating that microbial communities were inhibited at low pH.
TABLE III


2. Aquatic Plants

Freshwater ecosystems are supported by photosynthesis within the water body and by inputs of organic debris from the surrounding land. Primary production, the synthesis of living material from inorganic elements by photosynthesis, is carried out in freshwaters by a wide variety of plants, including leafy macrophytes, mosses, and algae.

i. Phytoplankton

Phytoplankton are microscopic plants that live suspended in the water column. Phytoplankton communities are usually quite diverse, with typically several dozen species. Evidence concerning the impact of acid deposition on phytoplankton comes from the large synoptic lake surveys in North America and Europe (particularly Scandinavia), from experiments in which the chemistry of lakes was changed intentionally to evaluate acidification impacts, and from studies of artificially enclosed "mesocosms" in which variables such as pH and nutrient concentration can be manipulated and the species composition controlled. These studies demonstrate that decreasing pH lowers species richness and diversity. Simplification of phytoplankton communities is especially acute over the range of pH 6–3.

ii. Periphyton

Periphyton is the material growing on submerged surfaces in freshwaters. It is dominated by microalgae that often form long filaments or sheers that can cover the sediments, plants, or other objects in water. The Periphyton can become a complex community of algae,
bacteria, fungi, and a variety of invertebrates. Periphyton species richness decreases with increasing acidity.

A striking phenomenon is the proliferation of attached algae in both streams and lakes, with increasing acidity. Common water macrophytes, such as Lobithullia doritarna and Isoetes laevigata, are festooned with filamentous algae and the bottoms of acidic streams may be covered with attached algae. Such increases in algal mass occur despite reduced specific rates of photosynthesis, indicating factors other than a preference for low pH are allowing algae to accumulate. Several ecological factors appear to contribute to algal proliferation at sub-optimal pH: decreased microbial activity, reduced competition among algal species allowing only the most acid tolerant to proliferate, and reduced grazing by invertebrates.

iii. Macrophytes
Aquatic macrophytes, including the vascular plants, mosses, liverworts, and macro-algae, are important elements of aquatic ecosystems. Macrophytes help to stabilize sediments and shorelines, form breeding grounds for some fish and many invertebrate species, and are a food source for waterfowl and mammals such as beavers and moose. Swedish limnologist Ole Grahn and colleagues (1974, 1977) studied acidification in Swedish lakes. Acid deposition decreased pH from 5.6 to 4.8 and brought about a regression of communities including Lobithullia, whereas communities dominated by the aquatic moss Sphagnum expanded from average coverage of about 8% to cover half of the littoral zones in a period of just 6 years. Sphagnum has a significant ion exchange capacity that results in the sequestration of Ca\(^{2+}\) and Mg\(^{2+}\), thus withdrawing cations from the water. The extensive moss mats covered much of the lake bottom and reduced both mineralization and exchange between the sediments and the overlying water. Large mats of Sphagnum are infrequently observed in North American lakes. Changes in macrophyte communities in acidified lakes may also be associated with other chemical changes, such as the availability of Ca\(^{2+}\). Raising the pH of lakes and increasing the Ca\(^{2+}\) supply by liming dramatically reduced Sphagnum communities.

3. Invertebrates
Lakes and streams that are not impacted by acidification have a diverse set of invertebrates with many species of insects, worms, crustaceans, and mollusks. In clear, unpolluted streams with moderate alkalinity in the pH range 6–8, there may be 70–90 species, of which a few are plentiful. As pH decreases below 5.7–5.4, so do the numbers of species. Mayflies, caddis-flies, crustaceans, and mollusks become rare or even disappear from the community. Changes in other elements of the ecosystem can alter its food supply, and changes in the faunal community may increase or decrease predation on a particular invertebrate species. There are critical pH thresholds below which survival of a particular species is greatly reduced. Not only the acidity of the water but also the concentrations of beneficial elements such as calcium and potassium and the concentration of toxic metals, particularly dissolved aluminum, are critical features in the responses of invertebrates to acidification and can greatly influence the rate of mortality at low pH.

i. Zooplankton
Zooplankton are small (normally less than 2 mm long) aquatic invertebrates, including copepods, cladocerans (water fleas), and rotifers, living in the water columns of lakes or slow-moving streams. Some are herbivores grazing on phytoplankton and some are predatory carnivores, and they are an important food source to fish and waterfowl. Synoptic surveys of hundreds of surface waters in Scandinavia and North America found that the number of zooplankton species in a water sample is highly correlated to pH. Several species of Cladocera and Rotifera are seen to increase in abundance with decreasing pH. Thus, zooplankton density (animals per liter of water) is not as sensitive to pH as is species richness since the more tolerant species can increase in number to replace missing species. In some acidified lakes there is a shift toward large-bodied zooplankton predators that may be due to decreased predation by fish, with the fish having been excluded due to acidification. This increases predation on smaller zooplankton. There is ample evidence that population-level changes are linked to increasing concentrations of Al\(^{3+}\) and reproductive failure.

ii. Macroinvertebrates
The aquatic macroinvertebrates are normally highly diverse assemblages of organisms. They are ecologically important to healthy ecosystems, assisting in the breakdown of litter and detritus, as grazers of algae, as predators of other invertebrates and juvenile stages of fish, and as a food source to fish and waterfowl. Surveys of macroinvertebrates in hundreds of lakes and streams in areas receiving large inputs of acid deposition clearly show that species richness declines sharply with increasing acidity. Several species of mayflies, amphipods, crayfish, and virtually all snails and clams are quite sensitive to low pH and are lost from the fauna of acidified waters. Species richness, diversity, and biomass decrease with decreasing pH. This is evident even
in the pH range 7.0–6.0. A few species are very tolerant of both low pH and elevated aluminum concentration. In acidified lakes, fish predation is reduced or eliminated altogether by the disappearance of the fish and the acid-tolerant and predatory water boatmen and back-swimmers (Hemiptera) may become important predators of other invertebrates. The amphipod *Gammarus lacustris* is absent from waters with pH lower than 6.0. Acidification experiments show that the progression through larval stages of *Lepidurus arcticus* is retarded with increasing pH and toxicity is complete at pH 5.5.

Impacts of acid deposition on lake ecosystem have been studied experimentally by David Schindler and colleagues at Canada’s Experimental Lakes Area. They intentionally acidified whole lakes over a period of several years from near neutral to pH near 5.0. Changes in macroinvertebrate communities became apparent even as pH changed from 6.8 to 5.9. Species numbers were reduced and others became more abundant as pH continued to decrease. At pH 6.0–5.8, the freshwater shrimp *Mysis relicta* became extinct. At pH 5.1, the crayfish *Orconectis virilis* became extinct, apparently due to a combination of factors including the inability to calciﬁy their shells, reproductive failure, and direct toxicity to juveniles.

Stoney streams normally contain a rich assemblage of macroinvertebrates. When ANC is moderate and pH is approximately 6 or higher, there may be 70–90 taxa present. When stream acidity is lower than 5.5, many of these taxa are scarce or absent. Mayﬂies, some caddis ﬂies, mollusks, and crustaceans are the most sensitive. The fauna is impoverished by acidification and may contain only half the numbers of taxa found in unacidified soft-water streams. Experimental acidification of streams has demonstrated detrimental impacts on macroinvertebrates including reduced numbers of species. Some intolerant species drift downstream to avoid the acidified waters and in this way can be eliminated from the acidified stream reach. In headwater streams, in which acidification is most severe, re-colonization would be unlikely.

4. Fish

In 1926, fisheries biologists noted that there was a widespread reduction in the catch of salmon in the major rivers of southern Norway, and in 1939 acid deposition was identiﬁed to be the cause. In seven rivers (mean pH 5.1) of this impacted region, 130 metric tons of Atlantic salmon were taken in 1900. Atlantic salmon were reduced and others became more abundant as pH increased from 6.8 to 5.9. Species numbers disappeared from the region. Lakes from which fish populations were lost had lower pH, higher concentrations of aluminum, and lower concentrations of calcium. Many lakes in the region have been studied intensely and fish kills associated with episodic acidification during acidic rain events and snowmelt are observed in some of the lakes in which fish stocks are declining.

Surveys in many areas show a strong relationship between species richness and lake pH, including lakes in Norway, Sweden, The Netherlands, Scotland, the LaCloche Mountains of Ontario, the Adirondack Mountains of New York, northern Wisconsin, and the Upper Peninsula of Michigan. Acidification problems in the United States and Canada may be greater than is indicated by large-scale surveys because they tend to miss episodic acidification events. Lakes and streams throughout North America, including high-elevation lakes in the West, experience such events. Many have low ANC and are therefore sensitive to acidification. Episodic acidification causes ﬁsh kills and can severely damage entire year-classes of fish. In the Adirondack Mountains 70% of all sensitive lakes are at risk of episodic acidification. In the mid-Appalachian region, 50% of sensitive streams, or seven times the number of chronically acidic streams, can become acidified by such episodes.

Both low pH and elevated Al$^{3+}$ concentrations are known to cause these impacts through the loss of the ability to regulate body salts and leakage of salts through the gills. Recruitment failure, due to effects on all stages from egg to adult, is an important mechanism for the loss of populations of fish.

5. Other Animals

Much less is known about the impact of acid deposition on other animals, such as amphibians (frogs and newts), birds, and mammals. Many species of amphibians are
declining throughout the world, but the causes are not obvious and large-scale species declines have not been clearly linked to acid deposition. Fish-eating birds are impacted by losses of fish populations. Elevated aluminum concentrations are associated with decreased reproduction in passerine birds. The concentration of cadmium is elevated in the internal organs of large herbivores in areas of North America and Scandinavia where surface water acidification is a problem. Elevated concentrations of mercury in fish in these areas may lead to contamination of otters and mink.

H. Materials
The problem of corrosion due to air pollutants has been known for 150 years or more. Angus Smith (1852) noted,

“The presence of free sulfuric acid in the air sufficiently explains the fading of colours in prints and dried goods, the rusting of metals, and the rotting of blinds. It has been observed that the lower portions of projecting stones in buildings were more apt to crumble away than the upper; as the rain falls down and lodges there and by degrees evaporates, the acid will be left and the action on the stone be much increased.”

Acid deposition contributes to corrosion of many types of materials, including painted surfaces, metals and carbonate stone (limestone and marble), masonry, carbon steel, zinc, nickel, and some paints and plastics. Both wet and dry deposition participate in the corrosion process. This is particularly a problem for limestone and marble buildings and monuments throughout the world. Monuments and buildings, such as the Taj Mahal, have suffered extensive damage. The great Gothic churches, such as the Cologne Cathedral and Notre Dame in Paris, as well as more ancient structures such as the Colosseum in Rome are melting away. Many structures that have withstood normal weathering processes for 1000 years or more are, in recent times, suffering extensive damage, as are newer buildings such as the U.S. Capital Building in Washington, DC.

I. Health
Sulfur dioxide can have serious health impacts on people. Persons with asthma can experience difficulty in breathing when exposed to SO₂ while exercising for as little as 3 minutes. Studies of air pollution episodes in London and New York in the mid-twentieth century found that among the elderly, the very young, and those with pre-existing respiratory disease, increased mortality followed exposure to average ambient SO₂ concentrations of >0.19 ppm for 24 hours. In other epidemiological studies, the U.S. EPA found that persons living within 20 km of large point sources of SO₂ emissions were at risk from such episodes. Lowering sulfate aerosol levels will reduce the incidence and the severity of asthma and bronchitis. Reductions in NOₓ and O₃ emissions are also expected to have a beneficial impact on health effects. The Clean Air Act and subsequent amendments resulted in reductions of SO₂ emissions. Consequently, air quality has improved. Nevertheless, approximately 46 million people in the northeastern United States continue to be exposed to air quality that does not meet EPA’s health-based air standards for one or more of the six criteria pollutants.

J. Visibility
Emissions that cause acid rain also reduce transparency of the atmosphere and decrease atmospheric visibility. The aesthetic properties of outdoor scenery in parks and forests such as the Shenandoah and the Great Smoky Mountains are noticeably reduced by hazy air. Particles with diameters less than 2.5 μm, dominated by sulfate and ammonium in eastern North America, account for 75–95% of visibility reduction. In the western United States, the sulfate contribution is less: 20–50% in rural areas and 10–20% in urban areas. A measure of atmospheric visibility is the visual range, which is the distance over which one can see. In the U.S. Southwest, the median value is about 130 km. On the U.S. Pacific and Atlantic coasts the median visual range is 20–50 km. Summertime haziness has generally increased in the eastern United States since the late 1940s, and this is largely due to increased sulfate aerosols. The trend is not uniform, however; haziness increased most in the Southeast. This has been a gradual process so that most people think that a slightly whitish haze on a clear, sunny day is normal. This haziness is what Stephen Schwartz has called the “white house effect” and it is a consequence of sulfate aerosols in the size range of 0.1–1 μm diameter. The sulfate is nearly all from oxidation of SO₂ emitted by fuel combustion. These aerosols act as condensation nuclei for water and the formation of clouds. Under some conditions aerosols may be reduced to 1% of their usual concentration by convective upward movement of air, cloud formation, particle scavenging, and precipitation. Such conditions make the air unusually dry. On these rare days the sky seems unusually blue, and this
ACID RAIN AND DEPOSITION

is a hint of what our ancestors could see on most sunny days.

V. REGULATION

An early attempt to limit precursors of acid deposition globally was the 1979 Geneva Convention on Long-Range Transboundary Air Pollution. This established emissions limits for sulfur and nitrogen that have in general been met. In the late 1990s emissions of SO\(_2\), in Europe were approaching half the amount emitted in the 1970s. In the United States, the Clean Air Act and subsequent amendments have brought about large reductions in SO\(_2\) emissions. By 1996 the annual wet SO\(_4^{2-}\) deposition over much of the eastern United States declined by 10–25%. Despite progress in reducing emission of SO\(_2\) in North America and Europe, the global problem of acid deposition is not likely to disappear. In Asia, emissions of SO\(_2\) are expected to triple in the period 1990–2010.

See Also the Following Articles

AIR POLLUTION • ATMOSPHERIC GASES • FOREST ECOLOGY

Bibliography


I. Historical Introduction
II. Two Common Definitions
III. Evidence for Adaptation
IV. Critique of Adaptationism
V. Adaptation after Adaptationism

GLOSSARY
adaptation One or two of the following: a beneficial construct produced by an omnipotent being, the process of change established by natural selection, and a biological character that gives increased Darwinian fitness.
adaptationism The doctrine that all important evolutionary processes are dominated by natural selection, and that all significant biological characters increase an organism’s fitness.
biological altruism Behavior of an organism such that the fitness of another organism is increased while its own fitness is decreased.
clutch size The number of eggs a bird lays in its nest at one time.
epistasis Interactions between genes at different chromosomal locations in the determination of phenotypic character values.
fitness Net reproductive output, discounted for any lack of viability.
genetic drift Accidents of segregation and recombination causing evolutionary genetic change.
group selection Selection between different populations or sub-populations based on attributes of the entire group, where these attributes usually are either selected against or not favored at the level of individual selection.
heterozygote An individual having two different alleles at a genetic locus.
hominid A great ape from the lineages most closely related to humans, where this may be a lineage ancestral to humans.
inbreeding The mating of close biological relatives.
individual selection Selection driven by differences in the net reproduction of individual organisms.
industrial melanism Selection for darker pigmentation as a result of industrial pollution, particularly in moths and butterflies.
linkage disequilibrium Nonrandom association of alleles on chromosomes.
meiotic drive Preferential segregation of a parasitic gene during gamete production.
phenotype The manifest biological character(s) of a particular organism.
phylogenetic Pertaining to evolutionary ancestry.
recombination The shuffling of gene combinations in the production of gametes, possibly by the physical breaking and rejoining of pieces of chromosomes.
segregation Allocation of genetic variants ("alleles") to different gametes during sexual reproduction.
telegogy The imputation of goal-directed behavior or structures.

**ADAPTATION** consists of one or two of the following: a beneficial construct produced by an omnipotent being, the process of change established by natural selection, and a biological character that gives increased Darwinian fitness.

### I. HISTORICAL INTRODUCTION

#### A. Classical Times

The concept of adaptation is older than any scientific concept of evolution, and certainly older than Darwin's theory of natural selection. The founder of academic biology, Aristotle, gave adaptive explanations for many of the features of the living and nonliving world. Thus, the webbed feet of a frog can be said to be "for" efficient swimming, and thus they can be explained as an illustration of the universe being well made. This type of reasoning was commonplace in classical culture, which often assumed some type of benign natural order.

It is significant, however, that adaptive reasoning had its critics even in classical times. Lucretius, one of the most important of classical proto-scientists, was scathing about the wholesale imputation of function to body parts, even when such inferences were regarded as "common sense." Such criticisms of the concept of adaptation have waxed and waned ever since.

#### B. Pre-Darwinian Christendom

Biblical theology gave arguments about adaptation a new cast. The assumption that there was a single, benign, omnipotent Creator made the existence of well-constructed organisms a natural assumption. From the beneficence of the Creator, each organism must have been given the specific characteristics best suited to its role in the Creation as a whole. Indeed, this concept of benign, and efficient, creation was extended to physics, especially by Isaac Newton, an avid believer. The orbits of the planets were thereby interpreted as evidence of some type of adaptation, or suitedness, to a divine plan.

This universal adaptation gave rise to some interesting paradoxes for pre-Darwinian scientists. Did the Creator adapt organisms to the physical universe or was the physical universe created to fit the organisms? Was life on Earth based primarily on the chemistry of water and organic molecules because that was the biochemistry that could work on this particular planet? Or was the planet constructed by the Creator to fit the biochemistry that He already had in mind? How could such questions ever be resolved?

#### C. The Darwinian Theory of Adaptation

Darwin's theory of evolution by natural selection provided a natural solution to the two problems of what adaptations were and how they occurred. In Darwin's theory, selection operating on heritable variation increased the frequency of individuals bearing attributes, "adaptations," which gave them increased fitness. Fitness in turn was to be defined as the net reproductive rate of individuals bearing attributes.

Darwin's theory had many novel features for the biology of his time. First, it involved no omnipotent Creator beneficently organizing the arrangements of life. Second, there were no inner drives or teleologies shaping the process of organic change to an adaptive end, unlike the scheme of Lamarck and others, who were more influenced by Aristotle than was Darwin. Third, there was no overarching pattern to Darwin's process of evolution, and therefore adaptations might occur higgledy-piggledy, whenever selection made mortality or reproduction hinge on a particular attribute. All these features made Darwin's new biology of adaptation distasteful to many of the older generation of Victorian biologists, who were highly teleological in their thinking when not avidly creationist.

### II. TWO COMMON DEFINITIONS

The nature of Darwinian theory instilled a large degree of ambiguity in the term adaptation. For a creationist, there is no process of adaptation, only what the Creator made and its beneficent nature. For the Darwinian, like the Lamarckian, there is necessarily a process of adaptation—a process by which adaptation is brought about. Then there is the product of adaptation as a process, which is called an adaptation as well.

There has been controversy regarding which of these basic meanings of the term is the true one or the correct one. However, we can follow Ernst Mayr, or Karl Pop-
A. Adaptation as a Process

The concept of adaptation as a process derives from the theory of natural selection. Therefore, a deeper consideration of this incarnation of adaptation requires study of natural selection. One of the basic intuitive expectations that most evolutionary biologists have is that natural selection should lead to the evolution of increased Darwinian fitness. However, this is not universally true. Mutation, segregation, recombination, meiotic drive, and frequency-dependent selection can force natural selection to produce a decrease in fitness. Usually, this decrease in fitness is temporary. For example, if the heterozygote, at a locus with two alleles, is the most fit genotype, then if the population is initially composed entirely of heterozygotes segregation will cause an immediate decrease in fitness. However, this effect is confined to the first generation. Subsequent generations will have increasing or stable mean fitness as natural selection brings the population to the stable gene frequency equilibrium, at which mean fitness will also be at a local maximum. Analogous processes can occur with other genetic processes, such as recombination. Again, in some cases, mean fitness does not continue to decline. However, there are also cases in which fitness may continue to decrease, and natural selection never produces a recovery in mean fitness. Meiotic drive occurs when some genes pervert segregation rates in their favor so that, despite being deleterious, they spread through populations. Another situation in which mean fitness can decrease is brought about by natural selection. The fitness of a mating with males and females can be a nonlinear function of the genotypes of the two organisms mating. If such nonlinearities are sufficiently severe, natural selection on fertility can actually drive fitness to increasingly lower levels, at least in theory. No prominent empirical examples of this process are currently known, however.

The point that these examples serve to make is that although it is conventional in evolutionary biology to expect improved adaptation from the action of natural selection, there is no absolute warrant, either in theory or in fact, for this assumption. Theory and experiment both indicate that a process of adaptation is usually brought about by natural selection. However, it is not always brought about by natural selection. When natural selection does act, however, to establish a process of adaptation, what can we say about that process? This is one of the three major research projects of evolutionary biology (the other two being the inference of phylogeny and the study of the genetic material used by evolution). Thus, our understanding of adaptation as a process is undergoing continual upgrading as our understanding of natural selection improves.

At the most basic level, however, there are some essential features of adaptation by natural selection that can be considered as well established. There is no general or consistent pattern to natural selection. Specific populations may undergo very intense selection for a short period of time. One of the best studied examples is the recent work on the evolution of Darwin's finches on the Galapagos, particularly the effects of drought on bill size (Grant, 1986). On the other hand, for most populations, it is usually very difficult to detect the action of natural selection. It is either too weak or too variable in direction (Abrahamson and Weis, 1997). Some of the cases in which natural selection can be readily detected as working in each generation to produce adaptation involve human disturbances—ecological events that are unlikely to reflect the evolutionary situation of populations that have been spared artificial disruption. The classic example of this scenario is the evolution of wing camouflage in the moths of industrial Europe in which natural selection was generated because soot blackened the tree trunks on which these moths rested, making the light-colored moths stand out against a black background. The very artificiality of this case, however, underscores the point that we do not normally find such cases of unequivocal selection when we study natural populations.

This leads to the next major point about adaptation as a process: It is difficult to detect. Therefore, adaptation as a process tends to be assumed by evolutionary biologists more than it is actually demonstrated. Also, the teasing out of the mechanistic particulars of adaptation as a process is almost never accomplished. This central problem has led to a pervasive weakness in the scientific analysis of adaptation as a process, with unfortunate consequences.

B. Adaptation as a Product of Evolution

The view of adaptation as a product of evolution does not logically require that it be a product of natural selection. An adaptation can arise evolutionary from selection on some other character(s), or it might occur from some nonselective process, such as inbreeding or genetic drift. Thus, for example, a spider's web might
have been evolved because of selection for prey capture, but it may also constitute an adaptation that enables spiders to obtain water from dew condensing on the web. This raises the following question: If adaptation is divorced from the process by which it arose, then how is it to be distinguished from the other characteristics of an organism?

The conventional solution to this problem is to define adaptations as those products of evolution, however generated, that enhance the fitness of the organism. Nominally, this requires that fitness be measured with and without the character(s) that is presumed an adaptation. This is a difficult enterprise for two reasons.

First, it is often difficult to perform the surgery, or other manipulation, required to make organisms without the adaptation in question. Recently, however, it is in precisely this area in which significant progress has been made in studies of adaptation (e.g., Sinervo and Basolo as cited in Rose and Lauder, 1996). Evolutionary biologists are now successfully ablating tissues and grafting on additional body parts in order to test the fitness consequences of the possession or loss of particular structures that are being evaluated for their status as adaptations. Manipulation of clutch size by removal or addition of eggs has been a traditional method in studies of vertebrate life history adaptations. Research in this area now manipulates fertility and egg size using a variety of techniques, including microsurgery. The resources of modern molecular biology are likely to give evolutionary research even more power to manipulate phenotypes.

Second, the measurement of fitness is difficult in most organisms. In organisms that reproduce strictly by dividing in two, without sex, fitness can be measured fairly easily from estimates of viability between bouts of fission. In every other kind of organism, sex and variable numbers of offspring make the estimation of fitness extremely difficult. Perhaps the worst character of all in the estimation of fitness is male mating success. This difficulty arises because the attribution of maternity is usually fairly secure, whereas the attribution of paternity is often pure speculation. This is an area in which the recent findings of behavioral ecology suggest considerable grounds for pessimism. Pairs of birds, for example, may indeed remain together for life, sharing the tasks of caring for young, foraging for food, and nest construction. However, molecular genetic analysis of pedigrees frequently reveals that the “monogamous” female has had sex with another male of the species, while the male has himself dallied. Similar patterns are well-known from human paternity cases. There are also species that are either highly promiscuous, such as chimpanzees, or ejaculate gametes externally, such as most fish. In these species, there are no mated pairs to keep track of over the long term. For these reasons, estimating the Darwinian fitness of an individual with a particular phenotype is often extremely difficult, if not practically impossible.

The fallback position of many biologists, especially functional morphologists, comparative physiologists, and behavioral ecologists, has been to use a surrogate for fitness. Such surrogates include mechanical efficiency, conservation of metabolic energy, and the number of copulations. The assumption is usually made that such surrogate measures will always be positively correlated with fitness. When they improve, fitness should increase. Unfortunately, it is precisely these characters that will show diminishing returns rather than a stable, positive correlation with fitness. Mechanical efficiency is patently not the only impact of structure on fitness. Structures may be costly to develop, or they may impede movement. Evolution is unlikely to maximize each and every “design feature” of an organism, even if there were no genetic constraints preventing the realization of any particular phenotype. Therefore, the expedient of using surrogates for fitness is not likely to be reliable in many cases.

If fitness cannot be accurately measured, and surrogates for fitness cannot be relied on, it is difficult to see how the concept of adaptation as a product of evolution can be used in most cases. There are pleas to the effect that some characters are so intuitively beneficial that they cannot reasonably be denied the status of adaptations. Legs must be adaptations for terrestrial locomotion, large brains must be adaptations for life as a tool user, and so on. However, limbs may be used for many functions, not just locomotion. The hominid brain has also been explained as an adaptation for social behavior, not the use of tools. Supposedly obvious cases become far from obvious once all possible scientific interpretations are taken into account.

III. EVIDENCE FOR ADAPTATION

If both basic definitions of adaptation are allowed, then there are two different lines of evidence for the existence of adaptation. The first is simply the action of natural selection. If adaptation is the process of natural selection, then any evidence for such selection is in turn evidence for adaptation. The second line of evidence is supplied whenever there are data showing an increase in fitness when a particular character is acquired. Together, the accumulated evidence bearing on both of
these points helps establish the importance of adaptation as a feature and as an outcome of the evolutionary process.

A. Evidence for Natural Selection
If Darwin generally lacked evidence for natural selection in nature, modern evolutionary biology has supplied an abundance of such evidence (Endler, 1986), including classic studies of industrial melanism and recent studies of drought selection in Darwin's finches (Grant, 1986). However, there are many other examples of natural selection in the wild, dating back to W. R. F. Weldon's study of carapace width in estuarine crabs in the 1890s. Indeed, natural selection is such an obvious feature of the living world that it is now considered in discussions of such practical medical problems as the prescription of antibiotics and the treatment of the human immunodeficiency virus (Freeman and Herron, 1998). Thus, the general principle that there is a process of adaptation involving natural selection is not in any reasonable doubt.

The evidential problems instead concern the importance of the process in any particular instance. The idea of an adaptive process shaping the course of evolution is very attractive because it can be used to support the interpretation of evolutionary change in terms of natural selection. However, as discussed previously, the demonstration that such a process is occurring is usually very difficult. Also, the possibility that other evolutionary processes are involved—processes that do not involve adaptation by natural selection for the character of interest—cannot be dismissed out of hand. This renders most casual post hoc invocations of natural selection essentially dubious. Whatever the specific features of natural selection, casually invoking it as an explanation for all features of life is no longer reputable behavior in evolutionary biology.

This means that, although there are some specific studies that provide excellent evidence for adaptation by natural selection, in most cases scientists are not in a position to interpret an evolutionary process as being driven by natural selection. It may be allowed as a possibility, but further study is usually required before a particular evolutionary change can be considered as being brought about by natural selection, even when such an interpretation seems intuitively natural.

B. Evidence for Increased Fitness
Even if it is difficult to establish the nature of the evolutionary process, surely the products of evolution are easier to categorize as adaptive? For the reasons discussed previously, however, it is often difficult to make an accurate determination concerning whether or not the possession of a particular character increases fitness. In particular, it is not enough to show that a particular function (e.g., locomotion) has been improved, perhaps by a longer hind-limb, because such demonstrations do not define the effect on fitness as a whole. A particular function could be improved while fitness is reduced.

Currently, some of the best demonstrations of adaptation come from the field of behavioral ecology. Of particular value have been manipulative experiments which change the behavior or morphology of study animals and plants. These studies have supplied many instances in which artificially created deviants have demonstrably reduced fitness (Sinervo and Basolo as cited in Rose and Lauder, 1996). There is much potential for the study of molecular biology to extend the power of manipulation in the study of adaptation, particularly with genetic transformation and the insertion of genes with artificially inducible expression. Some of these studies have measured the effects on adult survival of gene insertions (Fleming and Rose, 1996). Fitness could also be measured in such experiments.

An alternative approach is to measure the relationship between the variation of a character and fitness in polymorphic populations. Much of modern evolutionary quantitative genetics collects data of this kind. One of the central concerns in these studies is the delimitation of optima for fitness as a function of quantitative characters.

Finally, artificial selection can be used to generate perturbed values for selectable characters. If artificial selection is then relaxed, and the original character state was adaptive, natural selection should drive the character to its original state. This has been observed in only a few cases (Service et al., 1988). Such patterns of reversion are expected to occur especially when there are trade-offs between functional character, such that high values of one character are associated with low values of other characters. This situation is particularly important for the use of surrogate measures for fitness.

IV. CRITIQUE OF ADAPTATIONISM
A. Adaptationism
The record of classical thinking, such as that of Aristotle, illustrates the extent to which the human mind is attracted to the idea of beneficial organization in
the natural world. Many classical scholars believed this even when they had no particular scientific theories to buttress the concept. Darwinian evolution is thus an almost irresistible temptation for those who wish to infer function in the living world. Darwinism guarantees a role for natural selection in evolution, and it guarantees the existence of adaptation among the characters of organisms. However, it does not guarantee that selection and adaptation must be everywhere prepotent, at all times, and in all respects.

Nonetheless, there is a variant form of Darwinism that flourished particularly in the 1950s and 1960s—a variant that assumed that all the attributes of an organism are shaped by natural selection to the end of increased fitness. In this version of Darwinism, now called "adaptationism," all characters are adaptations and all nontrivial evolutionary processes are driven by natural selection. In effect, this school of thought made the study of evolution tantamount to the study of adaptation.

Among the effects of adaptationism on scientific practice was the notion that there must always be an adaptive explanation for every organ, structure, or behavior. Therefore, if an adaptive explanation for a particular structure has not been found, greater efforts must be made to discover its adaptive value. Alternative evolutionary processes (genetic drift, inbreeding, meiotic drive, etc.) must not be considered until all possible adaptive explanations have been tried and found wanting.

During its heyday, adaptationism put adaptation at the center of evolutionary biology, and to some extent at the center of all biology. The many theoretical and experimental problems facing the study of adaptation were minimized or dismissed altogether.

B. The Rejection of Adaptationism

From the late 1960s until the early 1980s, adaptationism suffered a series of blows from which it has yet to recover. The first of these was the detection of a vast amount of molecular genetic variation, first by protein electrophoresis and later by DNA sequencing. The significance of this finding for adaptationism is that most species appear to have far more segregating genetic variation than is likely to be explicable in terms of natural selection. Therefore, natural selection probably is not prepotent at the molecular level. The current scientific consensus is that many of the alleles that arise and eventually become fixed during evolution are merely neutral variants of already extant alleles. A great deal of genetic evolution has occurred, but much of it has not been driven by natural selection.

A second event was the publication of Adaptation and Natural Selection by George C. Williams in 1966. One of the common evasions of the adaptationists was to invoke group selection when they could not explain a particular character in terms of individual selection. Thus, many of the social behaviors of colonially nesting birds were explained in terms of adaptations for group selection. Williams pointed out that, usually, these explanations were highly dubious. He argued that the inference or explanation of adaptations required greater restraint, particularly regarding social behavior. This undercut group selection, one of the ways in which adaptationists had been able to discover adaptations underlying seemingly maladaptive behavior, such as biological altruism. In so doing, Williams also helped expose the extent to which adaptationism was based more on dogma than on well-founded science.

The third, and culminating, event in the decline of adaptationism was the publication of the paper, "Saunder's of San Marco," by Stephen Gould and Richard Lewontin (1979). In this paper, Gould and Lewontin hold up for ridicule the adaptationist assumption that there is a history of selection for every significant attribute of an organism. They follow Voltaire in his satirizing of such intellectual figures as Spinoza, particularly their boundless belief that "this is the best of all possible worlds," except that Gould and Lewontin satirize the adaptationist assumption of an all-powerful beneficent natural selection.

These events essentially undermined adaptationism as a dominant movement within evolutionary biology. Adaptationists remain scattered throughout biology, including such fields as molecular biology, comparative physiology, and systematics. However, the powerful hold that they had on evolutionary biology in the 1950s was broken.

V. ADAPTATION AFTER ADAPTATIONISM

Although adaptationism was clearly in error with regard to the universality of adaptation in the living world, its deposition brought with it an overreaction. Many evolutionary biologists effectively rejected the concept of adaptation as a whole. They refused to work on the problem and they criticized those who did. Since 1980, the study of phylogeny has become the central concern of evolutionary biology. For some, the study of adaptation is now a marginal, somewhat disgraceful, practice within biology as a whole.
Replacing the study of adaptation was the study of "constraints." Constraints in evolutionary biology are factors that prevent the achievement of an optimal adaptive outcome. Constraints have been discovered promiscuously in the evolutionary machinery: lack of genetic variation, linkage disequilibrium, too much environmental variation, too little environmental variation, epistasis, temporally variable selection, spatially variable selection, and so on. Evolutionary biology went from a doctrine in which adaptation was everywhere to a doctrine in which adaptation had disappeared to be replaced by paralyzing constraints.

In the 1990s, there was some stabilization of views on the topic of adaptation. Less of a pariah among evolutionary topics, an edited volume titled Adaptation was published by Rose and Lauder in 1996. Evolutionary biologists were spending more time using experimental and other techniques that could test for adaptation rather than simply assuming its presence or absence.

The comparative study of adaptation was greatly improved by an infusion of phylogenetic techniques. For example, if it is hypothesized that the gill structure of a fish species is an adaptation to a new way of life in salt water, but two species that had evolved in fresh water exclusively also have this gill structure, then the phylogenetic information indicates that the basic adaptive hypothesis is not correct.

Laboratory selection is currently used more often to study adaptation, with greater replication and greater attention to designs that can be used to make inferences about selection. The great advantage of performing selection in laboratories is that selective processes can be studied with greater statistical power and control, compared to the "experiments" of nature, all of which are unique and uncontrolled. For example, instead of studying the water physiology of two desert insect species compared to the "experiments" of nature, all of which are unique and uncontrolled. For example, instead of studying the water physiology of two desert insect species, evolutionary biologists select insects under conditions of desiccation using replicated selection lines and controls maintained free of desiccation (Bradley et al., 1999). Instead of dealing with possible historical accidents that might have differentiated species in the wild, selected laboratory populations provide good material for critically testing theories of adaptation to particular environmental conditions.

Another major development in the study of adaptation has been the use of natural populations, especially manipulated natural populations, in studies that approximate laboratory experiments. Reznick and Travis (cited in Rose and Lauder, 1996) have studied guppy evolution in the streams of Trinidad. Multiple streams pass in parallel through highly uniform drainage systems, giving the streams isolated guppy populations—populations that evolve under effectively identical conditions. This experimental system has provided tremendous opportunities for the study of adaptation in the wild with both replication and controls.

The study of adaptation now proceeds with much more skepticism than in the past. Simultaneously, empirical methods have been greatly improved. The prospects have never been brighter for a genuine scientific analysis of adaptation, as opposed to the blithe speculations of the past.

See Also the Following Articles

**ADAPTATION**

DARWIN, CHARLES • EVOLUTION, THEORY OF • INBREEDING AND OUTBREEDING • PHENOTYPE, A HISTORICAL PERSPECTIVE • PHYLOGENY • RECOMBINATION

Bibliography


ADAPTIVE RADIATION

Rosemary G. Gillespie,* Francis G. Howarth,† and George K. Roderick*
*University of California, Berkeley and †Bishop Museum

I. History of the Concept
II. Nonadaptive Radiations
III. Factors Underlying Adaptive Radiation
IV. Are Certain Taxa More Likely to Undergo Adaptive Radiation Than Others?
V. How Does Adaptive Radiation Get Started?
VI. The Processes of Adaptive Radiation: Case Studies
VII. The Future

GLOSSARY

adaptive shift A change in the nature of a trait (morphology, ecology, or behavior) that enhances survival and/or reproduction in an ecological environment different from that originally occupied.
allopatric speciation The process of genetic divergence between geographically separated populations leading to distinct species.
character displacement Divergence in a morphological character between two species when their distributions coincide in the same ecological environment compared to overlap of the character in question in the two species when they are geographically separated.
convergence The evolution of similar characters in genetically unrelated or distantly related species, often as the result of selection in response to similar environmental pressures.
ecological release Expansion of habitat, or ecological environment, often resulting from release of species from competition.
founder effect Random genetic sampling in which only a few “founders” derived from a large population initiate a new population. Since these founders carry only a small fraction of the parental population’s genetic variability, radically different gene frequencies can become established in the new colony.
key innovation A trait that increases the efficiency with which a resource is used and can thus allow entry into a new ecological zone.
natural selection The differential survival and/or reproduction of classes of entities that differ in one or more hereditary characteristics.
sexual selection Selection that acts directly on mating success through direct competition between members of one sex for mates or through choices made between the two sexes or through a combination of both modes.
sympatric speciation The process of genetic divergence between populations occupying the same geographic range leading to distinct species.
taxon cycle The repetitive pattern by which widespread dispersive stage I populations or species give rise to more restricted and specialized stage II populations or species; subsequent divergence leads to stage III local endemics.
Numerous definitions of adaptive radiation have been proposed. Almost all incorporate the idea of diversification in ecological roles, although they differ in their emphasis on relative rates of proliferation. Here, we propose a definition that seeks to be general but at the same time removes any implication of process: Adaptive radiation is a pattern of species diversification in which different species within a lineage occupy a diversity of ecological roles, with associated adaptations.

I. HISTORY OF THE CONCEPT

Beginning with the work of Darwin (1859) on the Galapagos fauna, the concept of adaptive radiation, in terms of diversification of ecological roles by means of natural selection, has been recognized. The term was first used by Osborn (1902) in describing parallel adaptations and convergence of species groups on different landmasses. Subsequently, it was developed as a major tenet for arguments presented in the modern synthesis by Huxley (1942). Simpson (1953), working on paleontological data, discussed the importance of key innovations in triggering adaptive radiation. For a detailed history of the concept of adaptive radiation, see Givnish (1997). Much recent information has been added, particularly during the past decade with the rise of molecular methods (Givnish and Sytsma, 1997).

II. NONADAPTIVE RADIATIONS

The term “nonadaptive radiation” has been used to describe situations in which species proliferation has not been attended by diversification of ecological roles (Gittenberger, 1991). When proliferation is simply a consequence of isolation, with isolated sibling species maintaining similar ecological affinities, then the radiation cannot be considered “adaptive.” As will be described later, isolation has been invoked to explain the initial divergence of taxa in some radiations (e.g., Galapagos finches and cichlid fish), with the adaptive phase not occurring until recently diverged sibling species become sympatric. However, there are some cases of nonadaptive radiation, with many allopatric and ecologically similar species. Most of these radiations are caused by changes in topography that, instead of opening up new habitats, have served simply to isolate a previously more widespread species. For example, isolated mountaintops and other continental refugia have allowed species long periods of evolution in isolation, without any ecological change. This may lead to patterns of considerable genetic distance between morphologically similar species from different isolates (Schneider and Moritz, 1999). Similarly, diversification of snails on islands has frequently been attributed to topographical isolation (e.g., Crete (Gittenberger 1991) and Madeira (Cameron et al., 1996)). In general, it appears that (i) nonadaptive radiation occurs if there is isolation without any novel ecological opportunity and (ii) coexistence of species within a lineage will not occur in nonadaptive radiations but is a primary characteristic of adaptive radiations.

III. FACTORS UNDERLYING ADAPTIVE RADIATION

The common requirement for triggering adaptive radiation is the opening up of ecological space. This may be allowed by intrinsic factors, i.e., something that changes in the organism to allow radiation to occur; for example, evolution of tolerance toward noxious plant chemicals (Farrell and Mitter, 1994; Mitter et al., 1988). Alternatively, it may occur as a result of extrinsic factors; for example, it has been reported to occur in geological history after an influx of nutrients into the system (Vermeij, 1995), in recent evolutionary time when new islands are colonized (Wagner and Funk, 1995; Liebherr and Polhemus, 1997), and in ecological time when a new habitat opens (Rainey and Travniano, 1998).

For ancient radiations, it is often difficult to determine the relative importance of intrinsic and extrinsic factors in allowing adaptive radiation. Factors associated with such radiations include (i) coincidence (after a slight delay) with major extinction episodes (Sloan et al., 1986) and (ii) radiation of a group frequently starting from a small, unimpressive set of species from an earlier period. For example, fossil ammonites (shelled cephalopod mollusks) reveal episodes of tremendous proliferation and extinction through the Devonian, Triassic, Jurassic, and Cretaceous (Fig. 1; Lehmann, 1981). Echinoderms show a similar pattern, originating in the Ordovician and undergoing small radiations until all but one lineage went extinct by the end of the Permian (Smith, 1984). These then radiated extensively in the Triassic–early Jurassic, and the current diversity of forms remains similar to what arose at that time. The great placental radiation (>4300 species) has
been attributed partly to the extinction of many reptilian groups at the end of the Cretaceous (Simpson, 1953). The parallel adaptive radiation of marsupials in Gondwana has also been attributed to the Cretaceous extinctions and subsequent opening of ecological space (Springer et al., 1997). However, within each lineage (placentals and marsupials) key innovations may have been involved. The radiation of ungulates and ruminants is associated with the opening up of the savannas (Fig. 2) but would not have happened if the organisms did not develop the morphological and physiological features necessary to exploit the habitat. Similarly, the radiation of the diprotodontians appears to have commenced in the Eocene and may have been promoted by a key adaptation for herbivory (Springer et al., 1997).

The actual basis for radiations subsequent to extinction episodes is still a subject of debate, particularly because coincidence between extinction events and subsequent radiations is generally poor. Vermeij (1995) argued that there is a stronger coincidence of species diversification episodes with increases in nutrient input into the biosphere.

We consider factors underlying species proliferation under two headings: intrinsic factors, and the concept of "key innovations," and extrinsic factors, including environmental change and colonization of isolated landmasses.

A. Intrinsic Factors: Key Innovations

Simpson (1953) suggested that the evolution of a suite of traits, or key innovations, that increase the efficiency with which a resource is used might allow species to enter a "new" adaptive zone, and the ecological opportunity thus allowed might promote diversification. The concept of the key innovation is an essential element in hypotheses of the evolution of specialization and subsequent adaptive radiation in herbivorous insects. However, the nature of key innovations is not often clear. In an attempt to define more clearly the concept, Berenbaum et al. (1996) examined cytochrome P450 and its relation to the adaptive radiation of butterflies. They found high levels of diversification in substrate recognition sites between species that do not share the same set of host plants; the reverse was true for those species that do share host plants. This result was taken to indicate that specialization may necessitate conservation of this region of the genome and could therefore be considered a key innovation.

Many attributes of species have been proposed as key innovations, or characteristics that have allowed diversification and proliferation. They generally involve the development of features that modify biotic interactions. Particular examples include the development of features that allow species to "escape" predation...
ory pressure of insects) and subsequent development of tolerance to the toxin in insects which allows them to radiate onto the plants. Symbioses are another possible "evolutionary innovation," allowing the abrupt appearance of evolutionary novelty (Margulis and Fester, 1991). They provide a possible avenue through which taxonomic partners can enter into a new set of habitats unavailable to one or both of the symbiotic partners alone. One of the most famous examples is the radiation of ruminants in the African savannas, which has been attributed partially to the development of gut endosymbionts and the concomitant ability to digest cellulose. Among the Foraminifera, Norris (1996) showed that photosymbiosis appeared in the fossil record in synchrony with the taxonomic differentiation of three of the dominant surface water foraminifera groups in the Paleocene and early Eocene. This radiation was not paralleled in the asymbiotic sister group. Symbiosis was suggested to provide a jump-start for diversification by providing the ecological opportunity.

In their classic paper, Ehrlich and Raven (1964) examined how interacting species in themselves may create ecological opportunity, and hence periodically enhance evolutionary rates, through a broad "coevolutionary" response. They hypothesized that, when plant lineages are temporarily freed from herbivore pressure via the origin of novel defenses, they enter a new adaptive zone in which they can undergo evolutionary radiation. However, if a mutation arose in a group of insects that allowed it to feed on one of these previously protected lineages of plants, it would also be free to diversify in the absence of competition. Ehrlich and Raven envisioned this as a step-like process in which the major radiations of herbivorous insects and plants have arisen as a consequence of repeated opening of novel adaptive zones that each has presented to the other over evolutionary history. This idea, termed the "escalation/diversification" hypothesis (Berenbaum and Feeny, 1981), has been supported by the work of Farrell and colleagues, who have studied insect diversification in the context of host plants (Fig. 3). Repeated evolution of angiosperm feeding in phytophagous beetles is associated with an increased rate of diversification (Farrell, 1998). Similarly, there is consistently greater diversity among plants in which latex or resin canals have evolved as protection against insect attack (Farrell et al., 1991).

Since Ehrlich and Raven, there has been a tremendous amount of research on the role of coevolution in dictating patterns of diversification. One of the major avenues that this research has taken is the study of the extent to which the phylogenetic order of divergence among herbivores or parasites corresponds to that among their hosts as a result of "parallel diversification" (Farrel and Mitter, 1994). A strongly corresponding evolutionary history might suggest a coevolutionary re-
ADAPTIVE RADIATION

FIGURE 3: Phylogeny estimate of Tetraopes beetles based on morphology and allozymes, compared to literature-based relationships of host plants. Host Asclepias shows an apparent progression toward increased complexity and toxicity of cardenolides, perhaps representing escape and radiation. From Farrell and Mitter, 1994.

How do the coevolution arguments invoke adaptive radiation? The situation that Ehrlich and Raven (1964) envisioned was one in which the host radiated prior to exploitation and subsequent radiation by the herbivore ("escape and radiation"). This might be considered analogous to the opening up of an array of ecological opportunities every time the innovation arose for either "escape" or "exploitation." The established diversity of hosts could provide the necessary diversity of ecological roles and associated adaptations. Where coevolution involves parallel diversification, adaptive radiation may not be involved. In particular, parallel diversification might be considered analogous to geographic separation, with divergence of the host causing isolation of the herbivore. This might then be considered a nonadaptive radiation. On the other hand, parallel diversification might cause an escalation in responses, with enhanced toxicity and reciprocal tolerance evolving in step-like progression (Berenbaum, 1983; Farrell and Mitter, 1994). In this latter scenario, adaptive radiation can be implicated for both the herbivore and the host.

B. Extrinsic Factors

Speciation rates are generally considerably higher in novel environments, whether a lake in the middle of a continent or an island in the middle of the ocean (Schluter, 1998; Fig. 4).

1. Environmental Change

Environmental change has frequently been implicated in species radiations, with the opening up of new habitat. Diversification has frequently been suggested to occur under stressful conditions [e.g., for the origin of angiosperms (Shields, 1993) and the recent diversification of mole rats (Nevo et al., 1984)]. However, any novel environment in which the organism is subjected to a new selective regime could be considered "stressful." In other words, an organism that successfully in-

sponse between the host and the herbivore or parasite. However, there appear to be few cases, at least among insect–plant interactions, in which the phylogeny agreement is precise. In most cases there has been periodic transfer of species to more distantly related hosts.
Per capita speciation rates of clades in novel environments (○) and in closely related "control" lineages inhabiting other environments (●). Rate estimates (y, left axis) are plotted against a dummy variable, the median of y-values. The solid line indicates y = x; points above the line therefore exhibit high rates. Rates were calculated from phylogenies based on allozyme frequencies. Time is measured in units of genetic distance (D). The calculation of rate y assumes exponential growth of species number: y = ln(N(0))/t, where N is the number of extant species in a clade and t is its estimated time of origin. Corresponding times required for species number to double are indicated on the right; the number of species in a clade doubles after ln (2)/y time units. From *Endless Forms: Species and Speciation*, ed. by D. J. Howard and S. H. Barlow, © 1998 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.

The evolution and adaptive radiation of the African cichlids (Fig. 5) appear to have been initiated by environmental change. Geological activity 20 million years ago (mya) caused the rivers in the area to become progressively meandric and swampy while still connected to the Zaire hydrological system. Over time, a mosaic of small, shallow, and isolated lakes developed, and finally the drainage system became closed and the lakes deepened (approximately 5 mya). The diversification of cichlid fish appears to have been initiated when river species moved into the swamp (Sturmbauer, 1998), and then successive radiations were associated with the development of protolakes and subsequently deep lakes.

In geological history, environmental changes appear to form the basis of the Phanerozoic revolutions (Vermeij, 1993): Increasing temperature and nutrient supplies as a result of submarine vulcanism may have triggered later Mesozoic and perhaps early Paleozoic diversification episodes. Similar factors may underlie the iterative radiations of ammonoids throughout the geological record (Dommergues et al., 1996). Each radiation appears to have originated from a few taxa, which went on to produce a wealth of morphological diversity. Although well documented for ammonoids, the pattern of iterative radiation is extremely rare in the fossil record. Within the total morphospace defined for the first three Jurassic stages, the radiation corresponds to a
ADAPTIVE RADIATION

string of "events" separated by episodes of morphospace collapse. During each event, only a portion of morphospace (≤45%) was filled; some morphs were reiterated, but each event had its own particular derived morphs.

2. Colonization of Isolated Landmasses (Particularly Oceanic Islands)

Isolated landmasses that have never been in contact with a source provide abundant opportunity in terms of newly available "niche space" for those taxa that manage to colonize them. Isolated archipelagos are generally considered in a different category of adaptive radiations, although they are really just special cases of environmental change: the appearance of a new environment, which happens to be isolated in the ocean. This has much in common with the formation of, for example, a lake in the middle of a continent. In either case, newly created habitats that are isolated from a source of colonists provide an extraordinary opportunity for adaptive radiation. Both the novelty and the isolation are key features in allowing adaptive radiation in such areas. If a new habitat appears in close proximity to other such habitats, it will be colonized by taxa from those habitats. Species diversity patterns will then match closely the predictions of the MacArthur–Wilson model of island biogeography (MacArthur and Wilson, 1967); that is, species diversity patterns will be governed by ecological processes. As isolation from the source of colonists increases, fewer taxa will be able to colonize the new habitat, and the low rate of colonization may provide sufficient time for species diversification to occur. Adaptive radiations are most likely to occur at the extreme ends of the dispersal range of a given taxon (Whittaker, 1998).

IV. ARE CERTAIN TAXA MORE LIKELY TO UNDERGO ADAPTIVE RADIATION THAN OTHERS?

Are species predisposed to undergo adaptive radiation because of a broad environmental tolerance, generalized feeding patterns, or perhaps some proclivity to develop novel associations? This question has been developed by some authors. For example, Adler and Dudley (1994) compared patterns of adaptive radiation among birds and butterflies in the insular Pacific: Birds have undergone extensive adaptive radiation, whereas butterflies have not. They argued that speciation in butterflies may be constrained by the mechanics of insect–plant coevolution that prevents rapid diversification. However, this argument is not well supported because other insects with similar coevolutionary ties have undergone some of the most spectacular insular adaptive radiations known. It appears that almost any group of organisms is capable of undergoing adaptive radiation upon being provided ecological opportunity that it can exploit.

V. HOW DOES ADAPTIVE RADIATION GET STARTED?

A. Initiation of Adaptive Radiation: Genetic Changes

1. Founder Events

The establishment of species in new environments inevitably involves sampling from the parent population. The size of the sample that can build a new population can be very small (cf. founder effects), although it need not necessarily be so. In particular, if, subsequent to colonization, a very small number of individuals were to proliferate rapidly, there would be little subsequent loss in genetic variability (Nei et al., 1975). Consequently, the deleterious effects of inbreeding are largely mitigated. However, because the genes represented in the founding population are only a small sample of the original population, genetic drift may be pronounced.

The nature of genetic changes during shifts in population size, particularly those experienced during or after population bottlenecks, has been the subject of considerable controversy in recent years. Clearly, a crash in population size as a result of a genetic bottleneck or founder event will cause allele frequencies at some loci to differ from those of the parent population because of accidents of sampling (Templeton, 1980). The debate concerns the nature of genetic changes that occur subsequent to the bottleneck, during the period of population growth. Traditional arguments suggested that founder events may trigger rapid species formation (Carson and Templeton, 1984). However, recent arguments have largely refuted the contribution of founder events to reproductive isolation (Barton, 1990).

Other possible changes during founder events are due to genetic reorganization. Carson (1990) proposed that blocks of loci are destabilized when a newly founded colony undergoes a flush of exponential growth, during which time selection is relaxed and recombinants that ordinarily have low fitness survive.
Release of additive genetic variance through change in epistatic interaction may allow formation of novel recombinants (Goodnight, 1988). Recent work has questioned the importance of epistatic interactions and the nature of “destabilization” among blocks of loci during periods of reduced population size. Slatkin (1996) discussed changes that might occur subsequent to reduction in population size in terms of conventional population genetics. Genetic drift will be the primary force causing genetic changes during the bottleneck. However, when the population starts to grow, the influence of genetic drift is expected to become weaker so that selection is most effective during and immediately after the period of rapid demographic expansion. On the other hand, Otto and Whitlock (1997) argued that, although it is true that the probability of fixation of beneficial alleles present during the bottleneck is increased during the subsequent period of population growth, a large proportion of alleles are lost during the bottleneck and few new mutations can occur while the population is at small size. The resultant effect of these opposing forces is that the number of beneficial mutations fixed per generation remains virtually unchanged by the bottleneck. Nevertheless, selection subsequent to a genetic bottleneck has an important effect on alleles that are initially rare and that would tend to be lost to stochastic events in populations of constant size (Slatkin, 1996). In addition, there is evidence that stress on a genetic system (e.g., as a result of a population crash) may activate transposable elements (Carson, 1990), which can exercise a mutagenic effect by interrupting structural or regulatory regions of the genes into which they become integrated. Therefore, mutation rates may actually be higher in founder populations, providing raw material on which selection can act, and which in turn could lead to a rapid recovery of genetic variability. There are thus two processes which can be associated with colonization events: (i) possible genetic changes/ restructuring within the population and an expected loss of genetic variability and (ii) the subsequent recapturing of genetic variation during population expansion as a result of selection (Slatkin, 1996).

2. Rapid Proliferation and Hybridization

Differential mixing of characters during segregation of populations and species may occur as a result of hybridization of newly divergent taxa (Harrison, 1993). Behavioral changes during founder events may facilitate hybridization because it has been suggested that sexual interactions may lose specificity subsequent to a founder event (Kaneshiro, 1989). Closely related heterospecifics may therefore hybridize and/or introgress subsequent to colonizing a new landmass, but the extent of genetic exchange may differ between regions of the genome (DeSalle and Giddings, 1986). Differences may be particularly pronounced between character sets involving nuclear and extranuclear DNA, such as mitochondrial or chloroplast DNA. Extranuclear DNA differs from nuclear DNA because of its greater sensitivity to the effects of each founder event due to its smaller effective population size (one-fourth) relative to nuclear DNA, attributable to its transmission primarily through the female line, and its existence as a single copy (Avise, 1991). Analysis of extranuclear DNA information will result in a gene genealogy, but it is likely to provide an incomplete history of the organisms if much hybridization has taken place. The occurrence of hybridization may explain differences that have been found between nuclear and mitochondrial phylogenies, particularly for the Hawaiian Drosophila (DeSalle et al., 1997). Indeed, natural hybridization with the formation of fertile hybrids has been documented between closely related species of Drosophila on the youngest island of the Hawaiian Islands (Carson, 1989). Among silverwords, hybridization has been implicated as an important element in the adaptive radiation of the group (Baldwin, 1997).

Among the Galapagos finches, recent molecular data have failed to distinguish species limits, at least for the morphologically defined ground and tree finch species: Individuals representing different morphologically identified species are intermingled. This may be explained by interspecific hybridization and/or sorting of haplotypes (Freeland and Boag, 1999). In the case of ground and tree finches, both explanations may apply. The incomplete species differentiation within these taxa may be taken as an indication that adaptive radiation is currently ongoing. Failure of molecular data to distinguish species limits has been found in several other adaptive radiations. For the cichlid fish of Lake Malawi, it has been suggested that speciation is occurring faster than alleles can become fixed within a species (Moran and Korstjeld, 1993). Among Hawaiian crickets, Shaw (1996) comments on a discrepancy between phylogenies generated on the basis of song (Otto, 1994) with that generated from mtDNA variation. Although the discrepancy could be explained by problems with current taxonomic boundaries, and problems with mtDNA lineage sorting, she argues that hybridization and introgression appear to be the most likely explanation.

Although the number of species in which hybridization has been documented is currently small, it is likely to increase as researchers accumulate phylogeographic

---

**ADAPTIVE RADIATION**

---
knowledge based on multilocus molecular data for different radiations.

B. Initiation of Adaptive Radiation: Ecological Changes

Populations frequently initiate a cycle of change in abundance and distribution upon colonization of a novel habitat or an unoccupied set of niches, during which they undergo ecological release, expand their range, and adopt a more generalized habit (Cox and Ricklefs, 1977). After their initial rise to dominance as widespread generalists, these taxa may subsequently be competitively displaced from much of their original range by younger relatives (Darlington, 1937). Wilson’s (1961) “taxon cycle” was used to describe such changes in the distributional pattern of Melanesian ants. He suggested that widespread, dispersive populations give rise to many more restricted and specialized species. Regular changes in ecological and geographical distribution have been recognized in many island systems (Cox and Ricklefs, 1977). Ecological release may be the precursor to adaptive radiation. This argument has been supported by recent experimental evidence: Rainey and Travisano (1998) studied adaptive radiation experimentally using the bacterium Pseudomonas fluorescens. This interesting bacterium is known to evolve rapidly in novel environments, with evolutionary differences detectable in the morphology and ecological affinity of the phenotype. An isogenic population of one morph was propagated in (i) a spatially heterogeneous environment and (ii) a spatially homogeneous environment. In the spatially heterogeneous environment, extensive morphological diversification was found to occur within 3–10 days, resulting in three dominant morphs. Wilson’s (1961) “taxon cycle” was used to describe such changes in the distributional pattern of Melanesian ants. He suggested that widespread, dispersive populations give rise to many more restricted and specialized species. Regular changes in ecological and geographical distribution have been recognized in many island systems (Cox and Ricklefs, 1977). Ecological release may be the precursor to adaptive radiation. This argument has been supported by recent experimental evidence: Rainey and Travisano (1998) studied adaptive radiation experimentally using the bacterium Pseudomonas fluorescens. This interesting bacterium is known to evolve rapidly in novel environments, with evolutionary differences detectable in the morphology and ecological affinity of the phenotype. An isogenic population of one morph was propagated in (i) a spatially heterogeneous environment and (ii) a spatially homogeneous environment. In the spatially heterogeneous environment, extensive morphological diversification was found to occur within 3–10 days, resulting in three dominant morphs. This was taken as evidence for diversifying selection. In the spatially homogeneous environment, no morphological variation was found.

VI. THE PROCESSES OF ADAPTIVE RADIATION: CASE STUDIES

What are the factors underlying diversification in an adaptive radiation? This question has fascinated biologists for many years. However, until recently, the characters involved in the adaptive radiation necessarily had to be used as the basis for phylogenetic inference. Given that adaptive radiations are characterized by tremendous levels of convergence in almost every morphological and ecological character, it has been very difficult to interpret evolutionary processes. As Givnish (1997) notes, “any rigorous, noncircular study of adaptive radiation must be based on a phylogeny that has been derived independently of the traits involved in that radiation.” Recent advances in molecular techniques, although not without problems of their own, provide an opportunity to obtain such an independent assessment of evolutionary history (see review of arthropod radiations in Hawaii in Roderick and Gillespie, 1998). The following sections outline case studies from a range of taxa, most of which have used a combination of morphological and molecular techniques in an attempt to gain some understanding of processes underlying adaptive radiations.

A. Galapagos Finches

There are currently 14 recognized species of Darwin’s finches in six genera, which have evolved from a common ancestor (Fig. 6; Lack, 1947; Grant, 1986). Of these, 13 live in the Galapagos Islands. Based on morphological, behavioral, and ecological data, they have been divided into three lineages: First, the ground finches, Geospiza (6 species), which are found in more arid areas of the archipelago and feed on seeds on the ground. Three of the species which are considered the most “finch-like” differ primarily in body and beak size and are known as the large (G. magnirostris), medium (G. fortis), and small (G. fuliginosa) ground finches. The 3 other species of ground finches have longer beaks. Two feed on cactus flowers and pulp as well as seeds and are known as the large (G. contortirostris) and small (G. scandens) cactus ground finches. Finally, the sharp-beaked ground finch (G. difficilis) supplements its diet with the eggs and blood of other birds and reptile ticks. Second, the tree finches, which are found mostly in trees and shrubs, are divided into 3 genera: Cactospiza [the woodpecker finch (C. pallida) and the mangrove finch (C. heliodora)], Camarhynchus [the large tree finch (C. pittacus)], the medium tree finch (C. pauper), and the small tree finch (C. parvula), and Platyspiza [the vegetarian finch, P. crassirostris]. All except P. crassirostris are insect eaters. Finally, the warbler-like finches, which are small with slender beaks, are in 2 genera: The warbler finch (Cerithidea olivacea) catches insects like a warbler, and the Cocos finch (Pinaroloxia inornata) is the only Darwin finch that lives outside the Galapagos Archipelago. It appears to have colonized Cocos Island from the Galapagos. Darwin’s finches share common features of nest architecture, egg pattern, and courtship displays. They
differ in song, morphology, and plumage. Based on morphology, allozyme, and DNA sequence data, the warbler finch *C. olivacea* appears to be closest to the ancestral form. However, recent molecular data indicate that the Cocos finch *P. inornata* is closest to the tree finches of the Galapagos (Sato et al., 1999; Petren et al., 1999). Also, the vegetarian finch appears to be ancestral to both the tree and ground finches rather than being a member of only the tree finch group.

Grant (1986) proposed a three-step process to speciation to explain the adaptive radiation of the group: (i) The ancestral species arrives in the archipelago; (ii) the species spreads to other islands and as a result of this, there will probably be some selection and hence differentiation in size because the islands are different; and (iii) members of the original and derived populations encounter each other, and as a result of competition for food and selection against intermediates character displacement causes rapid divergence in feeding structures between the species when they come together. Therefore, the radiation appears to be based on (i) the isolation of the archipelago, which provided ecological opportunity; (ii) considerable distances between the different islands, which has led to infrequent interisland exchange; and (iii) different environments on the different islands, which have selected for different feeding niches both within and between islands.
B. Hawaiian Drosophila

The Hawaiian Drosophilidae represent some of the most striking examples of adaptive radiation known for any group anywhere (Kaneshiro, 1988), with 337 species in the genus Drosophila (pomace flies), 19 in Idomyia (picture-winged flies), 122 in Scaptomyza, and 11 in Titanocharta (Hardy, 1965; Hardy and Kaneshiro, 1981). Based on the premise that founder events are the most important mechanism of speciation in the Hawaiian Drosophila, Kaneshiro (1983) proposed models suggesting the importance of sexual selection in driving species proliferation in these insects. The Hawaiian Drosophilidae, particularly the males of the picture-winged species, often have ornately patterned wings as well as unusual modifications of the mouthparts and legs. The extraordinary manifestations of these features in the male picture-winged species are frequently accompanied by elaborate courtship behavior. In classic studies of picture-winged Drosophila, Kaneshiro (1983) found evidence for asymmetrical sexual isolation: Females from a geologically older island were found to be highly discriminating in terms of mate choice and would not mate with a closely related species from a younger island. In contrast, females from species on the younger island readily accepted males from the older island species. Based on these observations, Kaneshiro proposed that courtship requirements are relaxed during the early stages of colonization of a new island; sexual behavior may then become less constrained and simpler, and there may be more intraspecific variability. In such circumstances, there would be strong selection for less discriminating females because of the difficulties in finding mates and reproducing when the population size is small. Intrasexual selection may then operate to cause divergence of the sibling species during isolation because of a shift in the distribution of mating preferences during the founder/flush cycle (Carson, 1986). The result would be a shift in the mating system, which could then be fixed at a new “equilibrium.”

Evolution of the Hawaiian Drosophila is commonly treated as an example of sexual selection influencing speciation (Carson, 1986). The nature of this influence is not entirely clear. Drosophila heteroneura and its close relative, D. silvestris, are partially sympatric in forests on the island of Hawaii, where they occasionally hybridize. The species are distinguished by the much broader head of D. heteroneura. Male aggression and male courtship are the major determinants of male mating success. Accordingly, it appears that male head width is subject to sexual selection through mate choice, although this, and perhaps related sexually dimorphic traits, may not be involved in behavioral isolation (Boake et al., 1997).

A recent molecular phylogeny of picture-winged Hawaiian Drosophila showed that clades are characterized according to whether they breed on fungi, leaves, fruit, or bark (Kambysellis and Craddock, 1997). Suites of reproductive characters in particular, ovarian egg and ovipositor traits appear to have evolved together (Kambysellis, 1993), and adaptive shifts to new breeding sites appear to have been important at least in the early diversification of the group and in some recent speciation events (Kambysellis and Craddock, 1997). The interplay of such ecological shifts with sexual selection in allowing diversification of Hawaiian Drosophila remains unresolved.

C. Hawaiian Honeycreepers

quent morphological differentiation. The second factor considered to be involved in the diversification of these fish is sexual selection, which allows reproductive isolation and hence diversification. The basis for this assertion is their trophic apparatus, which is unique in the possession of two sets of jaws—one in the mouth for sucking or scraping and the other in the throat for macerating, crushing, or piercing. These structures can be modified according to the diet of the fish. The second factor considered to be involved in allowing reproductive isolation and hence diversification of the fish is sexual selection, which allows subsequent morphological differentiation.

D. African Cichlids
Lacustrine fish represent some of the most spectacular cases of adaptive radiation in vertebrates (McCune, 1997), with the best known being cichlids. Cichlids are spiny-rayed freshwater fishes that reach their most abundant diversity in Africa, particularly in the great east African lakes of Victoria (300–500 species), Malawi (300–500 species), and Tanganyika (approximately 400 species) (Fig. 5). In each of these lakes the fish exhibit spectacular diversity in trophic morphology, including specialist algal scrapers, planktivores, insec-tivores, piscivores, paedophages, snail crushers, and fin biters. In addition to their trophic diversity, they exhibit a striking array of color patterns. They also show complex mating behaviors, polyanandrous mating systems, and a tendency to breed in leks. Several factors are considered to be involved in the diversification of these fish. One suggests that morphological adaptation is the primary event underlying speciation (Liem, 1973). The basis for this assertion is their trophic apparatus, which is unique in the possession of two sets of jaws—one in the mouth for sucking or scraping and the other in the throat for macerating, crushing, or piercing. These structures can be modified according to the diet of the fish. The second factor considered to be involved in allowing reproductive isolation and hence diversification of the fish is sexual selection, which allows subsequent morphological differentiation.

E. Hawaiian Silverswords
The Hawaiian silversword alliance has been considered “the best example of adaptive radiation in plants” ( Raven et al., 1992). It consists of 30 species in three genera (Wilkesia, Dubautia, and Argyroxiphium) but with one common ancestor. Life form diversity includes trees, shrubs, mat plants, monocarpic and polycarpic rosette plants, cushion plants, and vines that occur across a broad environmental spectrum from rain forests to desert-like settings (Carr, 1985). Major ecological shifts have accompanied speciation in each of the major island-endemic lineages (Fig. 8; Baldwin, 1997). The estimated minimum rate of diversification in the silversword alliance (0.56 ± 0.17 species per million years) is comparable to, or higher than, rates of several more ancient continental groups that have been regarded as examples of adaptive radiation (Baldwin and Sanderson, 1998).

F. Lizards in the Caribbean
Currently, 139 species of Caribbean lizards are recognized, 80% of which occur in the islands of the Greater Antilles (Cuba, Hispaniola, Jamaica, and Puerto Rico)
Losos (1992) and Losos et al. (1998) found that on each of the islands lineages have diversified in such a way as to occupy a range of ecological roles, with as many as 11 species occurring sympatrically. Different species live, for example, on twigs, in the grass, or on tree trunks near the ground. Twig anoles tend to be slender with short legs and tails, whereas the trunk-ground anoles are stocky with long legs and poorly developed toepads. Losos and colleagues recognized six habitat specialists and called these “ecomorphs,” which are recognizable on the basis of morphological measurements. The same set of ecomorphs are found on each island (with a few exceptions): Four occur on all islands (trunk-ground, trunk-crown, crown-giant, and twig), one occurs on all islands except Jamaica (grass-bush), and one occurs only on Cuba and Hispaniola (trunk). Phylogenetic studies using mitochondrial DNA indicate that the same ecomorphs have evolved independently on each island, i.e., they appear to have arisen as a result of one-to-one convergence of the same set of ecomorph types on each island (Losos, 1992). However, the sequence by which they evolved differs on each island. Such tests of convergence are among the strongest available for evaluating the premise that the number and types of coexisting species are locally determined. The conclusion from this research was that “adaptive radiation in similar environments can overcome historical contingencies to produce strikingly similar evolutionary outcomes” (p. 2115, Losos et al., 1998).
<table>
<thead>
<tr>
<th>Section</th>
<th>Characteristics</th>
<th>Cuba</th>
<th>Hispaniola</th>
<th>Jamaica</th>
<th>Puerto Rico</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Crown</td>
<td>Large body, large toe pads</td>
<td><em>Anolis equestris</em></td>
<td><em>A. ricordii</em></td>
<td><em>A. garmianii</em></td>
<td><em>A. carieri</em></td>
</tr>
<tr>
<td>Upper Trunk/Canopy</td>
<td>Large toe pads, can change color</td>
<td><em>A. porcarius</em></td>
<td><em>A. chlorocyamus</em></td>
<td><em>A. grahamii</em></td>
<td><em>A. evermanni</em></td>
</tr>
<tr>
<td>Twig</td>
<td>Short body, slender legs &amp; tail</td>
<td><em>A. angusticeps</em></td>
<td><em>A. immotus</em></td>
<td><em>A. valencieni</em></td>
<td><em>A. occultus</em></td>
</tr>
<tr>
<td>Midtrunk</td>
<td>Long forelimbs, vertically flattened body</td>
<td><em>A. losiana</em></td>
<td><em>A. dirichius</em></td>
<td>none found</td>
<td>none found</td>
</tr>
<tr>
<td>Lower Trunk/Ground</td>
<td>Stocky body, long hind limbs</td>
<td><em>A. sagrei</em></td>
<td><em>A. cybotes</em></td>
<td><em>A. lineatipes</em></td>
<td><em>A. gundlachii</em></td>
</tr>
<tr>
<td>Grass/Bush</td>
<td>Slender body, very long tail</td>
<td><em>A. altamorensis</em></td>
<td><em>A. obsoni</em></td>
<td>none found</td>
<td><em>A. pulchelius</em></td>
</tr>
</tbody>
</table>
G. Sticklebacks in Deglaciated Lakes of Canada

Deglaciated lakes of coastal British Columbia, Canada, harbor a tremendous diversity in the form of undescribed sibling species of fish. There are numerous examples of species pairs in many fish families, and the repeated occurrence of such sympatric pairs has been attributed to novel ecological opportunity provided by deglaciation and recolonization from relatively depauperate faunas. In particular, the radiation of species of three-spined sticklebacks (a species complex currently classified under *Gasterosteus aculeatus*) includes some of the youngest species known—less than 13,000 years. No more than two species occur in any one lake, but pairs of species in different lakes appear to have evolved completely independently of other pairs. Results based on morphological, ecological, and molecular data suggest that species have diverged as a result of parallel bouts of selection for alternate trophic environments (Fig. 10; Taylor et al., 1997). Natural selection has been implicated as the major cause of evolution (Schluter, 1994; Schluter and McPhail, 1993). This study has been very important in highlightig the role of divergent natural selection as a mechanism underlying adaptive radiation—a view held widely by naturalists in the earlier part of the century although, until recently, without much support (Schluter, 1996). Ecological character displacement appears to underlie diversification in the sticklebacks, with the evolution of reproductive isolation a by-product of resource-based divergent natural selection.

H. Hawaiian Tetragnatha Spiders

The long-jawed, orb-weaving spider genus *Tetragnatha* comprises 295 described species worldwide. The adaptive radiation of the genus in Hawaii has been uncovered only recently, with a total of 28 species described and many more undescribed (Fig. 11; Gillespie et al., 1998). This radiation spans a huge spectrum of colors, shapes, sizes, ecological affinities, and behaviors. Phylogenetic analyses have been performed on the “spiny-legged” clade of Hawaiian *Tetragnatha*, a lineage that has adopted a vagile, cursorial predatory strategy. On each island, species can be characterized as “green leaf-dwelling,” “maroon,” and “gray/black bark-dwelling.” However, species on any one island are generally most closely related to each other, and each of the different ecomorphs appears to have evolved independently on the different islands (Gillespie et al., 1997). There appears to have been a one-to-one convergence of the same set of ecomorph types on each island.

I. Hawaiian Swordtail Crickets

Thirty-five species of *Laupala* (Gryllidae) are found in rain forests across the Hawaiian archipelago, and up to 4 species coexist on any given island (Otte, 1994).
Adaptive radiation of Hawaiian Tetragnatha spiders. Shown are representative species from a lineage that has abandoned web building ("spiny-leg" species) and a large lineage of web-building species that show tremendous morphological and ecological diversity. Spiny-leg species: From top, T. bryanti and T. hamalae from wet forests on E. Maui; T. pilosa from wet forests on Kauai. Web-building species: Five from wet forests on E. Maui (clockwise from bottom left): T. filiciphilia, T. stelarobusta, T. paludicola, T. euryphila, and T. trituberculata; (bottom right) undescribed species from wet forests on Hawaii.

common with other crickets, courting males "sing" to attract females by rubbing their forewings together. Each species has a unique song, and females respond preferentially and move toward the source of their species-specific song. It appears, therefore, that the song serves as an important mate recognition signal and variation in the acoustic system may serve as a basis for behavioral isolation between species. Phylogenetic analyses based on morphological (Otte, 1994) and molecular (Shaw, 1996) characters indicate that extensive intraisland species formation has occurred in this group. Differentiation between closely related species is associated with changes in song. These features of the Laupala system suggest that, unlike the Hawaiian Drosophila, founder events have not played a major role in the initial stages of species diversification (Shaw, 1996). Rather, differentiation appears to occur through the interaction of sexual selection on genetically well-struc-
pears to be important in dictating the array of species at a site.

K. Bacteria in Culture
As described previously, experimental studies on a bacterium, *Pseudomonas fluorescens*, have provided fascinating insights into possible processes underlying adaptive radiation. *Pseudomonas fluorescens* was found to diversify rapidly from an isogenic population under conditions of environmental heterogeneity. Moreover, selection appeared to be the primary force maintaining diversity in the heterogeneous environment. The evolution of variant forms was found to follow a predictable sequence (cf. Caribbean lizards, stickleback fish, and *Tetragnatha* spiders), and competition is inferred to maintain the variation (Rainey and Travisano, 1998).

L. Conclusions
Three primary mechanisms have been inferred to underlie species proliferation in adaptive radiations: (i) geographical differentiation and subsequent character displacement (on the basis of either ecologically or sexually important characters) in Galapagos finches and African cichlids; (ii) sexual selection, with or without genetic bottlenecks, in Hawaiian *Drosophila* and *Laupala* crickets; and (iii) divergent natural selection based on ecological shifts in Hawaiian *Tetragnatha* spiders, plants in the Hawaiian silversword alliance, lizards in the Caribbean, *Partula* land snails in the Pacific, stickleback fish in deglaciated lakes, and bacteria in culture.

VII. THE FUTURE
Adaptive radiation, as a phenomenon, has tremendous research potential. The existence of a suite of closely related species adapted to exploit different habitats or lifestyles allows one to make comparative studies on the processes of speciation and selection in natural populations. Molecular systematics is providing a much better understanding of the evolutionary history of groups of closely related species and provides the opportunity for testing mechanisms underlying adaptive radiation (Givnish, 1997). The phylogenetic hypotheses provide a framework for examining the evolution of specific morphological, ecological, behavioral, and physiological adaptations and the circumstances in which they have arisen and have allowed adaptive radiation.

Adaptive radiations have recently become the focus of studies in conservation biology because they are frequently, particularly on islands, associated with high frequencies of endemism. As a corollary to the high endemism, many of the species that make up an adaptive radiation are often very rare and are characterized by very high extinction rates. Far more extinctions have been documented from islands than from continents: Of known extinctions, those from islands comprise 38% for mammals (most of which are absent from remote islands), 80% for mollusks, and 83% for birds (Whittaker, 1998). Similarly high extinction rates are known for radiations of lacustrine fish. The high rate of extinction of these narrowly endemic species has been greatly accelerated in recent years by the introduction of alien species into island environments. The plight of these extraordinary sets of species has largely been ignored because attention has focused on the devastation of forests in South America and Asia. However, setting aside even small reserves would serve to protect many species on islands from extinction. The microcosmal nature of species swarms makes such efforts at least feasible. However, without immediate action, few of the world’s most spectacular radiations will survive far into the next millennium.

See Also the Following Articles
COEVOLUTION • DARWIN, CHARLES • DISPERSAL BIOGEOGRAPHY • DIVERSITY, MOLECULAR LEVEL • ENDEMISM • HABITAT AND NICHE, CONCEPT OF • POPULATION GENETICS

Bibliography
ADAPTIVE RADIATION


ADAPTIVE RADIATION


I. Introduction

II. Evolutionary Approaches to Environmental Aesthetics

III. Environmental Information, Problem Solving, and Survival

IV. Negative Emotional Responses

V. Positive Emotional Responses

VI. Aesthetics and Biodiversity

VII. Aesthetics and Cognition

GLOSSARY

aesthetics The field of investigation, from philosophical and psychological perspectives, that attempts to discover the rules and principles that govern the sense of beauty and ugliness.

beauty The qualities of a perceived or imagined object whereby it evokes feelings of admiring pleasure.

cognition Act or process of knowing and understanding.

habitat Type of environment in which an organism lives, grows, and reproduces.

phobia Obsessive fear or dread of an object or situation.

symmetry The correspondence in size, form, and arrangement of parts on opposite sides of a two- or three-dimensional object.

AESTHETIC FACTORS are those characteristics of a given object or situation that evoke a certain emotional response, either a sense of beauty, attractiveness, pleasure, symmetry, order, and so on, or, conversely, of ugliness, disorder, menace, disgust, or the like. Generally speaking, the aesthetic preferences that humans display in response to their environment, in such contexts as mate choice, food patterns, and habitat selection, have been shaped by evolutionary experience and reflect suitable solutions for survival and reproductive success.

I. INTRODUCTION

Humans have strong emotional responses to living organisms and to natural and human-modified environments. Depending on the circumstances and the needs of an individual, nature may evoke feelings of awe, respect, fear, loathing, longing, nostalgia, excitement, challenge, and belonging. These powerful emotions influence how we respond to nature, how we attempt to manipulate it, and why we care about it.

Strong emotional responses in all organisms, including humans, evolved because, on average, they increased fitness, that is, they improved the survival and reproductive success of the individuals that expressed them. For example, those of our ancestors who did not enjoy food and sex certainly were more poorly represented in subsequent generations than those who did enjoy and hence sought out food and sexual partners. Similarly, individuals who chose ecologically inferior environments in which to live should have...
been less well represented genetically in subsequent generations than individuals who made better habitat choices.

However, this truism does little more than establish general guidelines for investigating why human emotional responses to components of the environment are so varied. Nor does it take us very far in improving our understanding of how different emotional responses to varied environmental components resulted in fitness-improving behaviors. More specific theories and analyses are needed.

II. EVOLUTIONARY APPROACHES TO ENVIRONMENTAL AESTHETICS

Why human emotional responses to nature would have influenced fitness is obvious. Our ancestors lived in environments devoid of modern conveniences. Their health, survival, and reproductive success depended on their ability to obtain and use environmental information wisely. They had to know how to interpret signals from the physical and biological components of the environment, and to adjust their behavior in response to them. They needed to evaluate habitats and the resources provided over space and time, and to adjust their use of habitats accordingly. Our ancestors also needed to be able to detect the presence of other humans, their effects on resources, and the direct dangers they posed or the help they might provide.

Making these evaluations and decisions was and still is difficult because the environment provides a far richer array of information than can possibly be assimilated and synthesized. Adaptive behavior requires selective attention to components of the environment that strongly influence fitness. In fact, the complex behavior of organisms would be impossible in the absence of neural filters that emphasize or deemphasize aspects of information emanating from the environment. These preselection filters embody evolutionarily stored knowledge about the world that enables us to construct hypotheses capable of describing and understanding the external world. Extensive experiments on artificial intelligence have clearly shown that the rapid and efficient learning of a language performed by nearly every human child is quite impossible in the absence of preformed neural structures (Dennett, 1995).

These prefilters evidently express themselves in all organisms as a sense of aesthetics and some form of logical analysis (Marchetti, 1998). In other words, aesthetic responses, rather than being a recently acquired capacity of little evolutionary significance, are fundamental to the ways in which organisms learn about the world and adapt to it. The English word aesthetic is derived from the adjectival form of the Greek aisthētikos, which means "to perceive." Thus, "aesthetic pleasure" means literally "pleasure associated with or deriving from perception." The central problem in aesthetics is to explain why pleasure and disgust have evolved to be so strongly associated with perception and recognition of certain kinds of objects.

The evolution of preselecting filters is reflected in what is called biologically prepared learning (Seligman, 1970). Biologically prepared learning theory asserts that evolution has predisposed humans, as well as individuals of other species, to learn easily and quickly, and to retain associations or responses that foster survival when certain objects or situations are encountered. An important corollary is that even though modern societies have greatly reduced the real danger posed by the objects of fears and phobias, fear and avoidance responses may nonetheless persist because selection against those responses is weak (Ulrich, 1993, Wilson, 1984). Biologically prepared learning in humans should be evident only for stimuli that have had significant influences on survival and reproductive success during our evolutionary history. Preselected responses should arise quickly when the appropriate conditioning stimuli appear, and these responses should be unusually resistant to extinction or forgetting. Adaptive responses to natural stimuli do not necessarily appear spontaneously or in the absence of learning, but they may.

Although aesthetic inquiries have been carried out since the middle of the eighteenth century, attention has been directed almost exclusively to responses to human creations rather than to nature. The neglect of nature is a by-product of the view that has dominated Western thinking for centuries—namely, that aesthetic experiences are molded primarily by cultural symbols and art forms. As a result, the study of aesthetics has been viewed as the domain of artists and philosophers. Attempts to explore the biophysical bases of aesthetic responses to environments were regarded, and still are by some people, as both futile and ideologically dangerous.

Culture and learning clearly exert strong influences on the ways humans perceive and respond to environmental information and they have important impacts on the symbolisms we attach to natural objects (Appleton, 1990; Schama, 1995). But attempts to understand and interpret human aesthetic responses to environmental features without asking why they evolved have failed.
For example, people who studied the concept of beauty from a nonbiological perspective assumed that beauty was an intrinsic property of objects. They therefore looked for, and expected to find, correlations between the characteristics of objects and human aesthetic responses to them. This attempt largely failed because, as an evolutionary perspective immediately suggests, beauty is not an intrinsic property of the objects that we call beautiful. Rather, it is the product of interactions between traits of objects and the human nervous system that evolved so that we regard as beautiful those objects having properties that, if positively attended to, result in improved performance in some aspect of living (Appleton, 1975). Conversely, we regard as ugly those objects that should be avoided or destroyed. In essence, an evolutionary perspective suggests that concepts such as beauty and ugliness should be viewed from a functional rather than a structural perspective. In other words, emotional responses are best studied by asking “How did these responses help us solve problems?”

III. ENVIRONMENTAL INFORMATION, PROBLEM SOLVING, AND SURVIVAL

Individuals of all species use information to make decisions that enhance their survival chances. The array of information presented to an individual has both components that are highly relevant to survival and reproductive success, and components that can be ignored with few or no adverse consequences. In addition, which components are important varies with time, location, and the needs of the individual.

Progress in the study of environmental aesthetics and problem solving has been aided by the development of systems for classifying environmental information into categories that roughly correspond to the kinds of decisions that organisms must make (Heerwagen and Orians, 1993). The basic problems that people (and other animals) must solve are: (1) protecting themselves from being injured or killed by other people or dangerous animals; (2) avoiding being attacked by parasites and disease-causing organisms that may be acquired from people or from other sources; (3) protecting themselves from adverse physical conditions; (4) acquiring enough of the right kinds of food; and (5) choosing good-quality associates for reproduction, foraging, protection, and achieving higher social status.

One useful classification divides information into categories based on the type of objects being identified. Some of these objects, such as food, water, and refuges, are resources. Gathering information about these resources should be pleasurable and exciting. Other objects, such as potentially dangerous animals, competitors, human enemies, and hazardous physical objects, pose dangers. Some of them are life-threatening, but others, although potentially dangerous, can be studied and overcome. Gathering information about them should be accompanied by feelings of fear, anxiety, and apprehension rather than pleasure (Table I).

Another classification of environmental information is based on the time frame over which the information is relevant. Although time is a continuous variable, humans long ago discovered that it is heuristically useful to divide time into categories. From the perspective of the study of environmental aesthetics, time is most usefully divided into categories that correspond to the time frames over which decisions about them matter (Table II). Some environmental information signals events of temporary significance. Still other information signals seasonal changes that are associated with shifts in the types and locations of resources that will be available in the near future. Still other information signals seasonal changes that are associated with shifts in the types and locations of resources that will be available in the near future. Finally, other information arrives from objects that, measured in terms of human lifetimes, are permanent.

A third classification is based on the sensory mode in which the information arrives. What information can be perceived obviously depends on the sensory capacity.
TABLE II

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
<th>Decisions affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short term (minutes to hours)</td>
<td>Weather changes (thunderclouds, wind)</td>
<td>Seeking shelter, maintaining outdoor activities</td>
</tr>
<tr>
<td></td>
<td>Appearance of dangerous animals, enemies</td>
<td>Immediate defensive actions</td>
</tr>
<tr>
<td></td>
<td>Appearance of valuable prey</td>
<td>Immediate hunting activities</td>
</tr>
<tr>
<td></td>
<td>Illumination changes</td>
<td>Moves to appropriate locations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>for spending the night</td>
</tr>
<tr>
<td>Seasonal</td>
<td>Day-length changes, vegetative growth, flowering, precipitation changes</td>
<td>Shifts of hunting sites, planting and harvesting of crops</td>
</tr>
<tr>
<td>Multiyear changes</td>
<td>Vegetation succession, erosional changes (river meanders, lake sedimentation)</td>
<td>Shifts of hunting sites, movement of villages</td>
</tr>
<tr>
<td>Long term (decades to centuries)</td>
<td>Topography</td>
<td>Development of traditions</td>
</tr>
</tbody>
</table>


of an organism and its neural processing system. The first, and still the most important sensory capacity in the living world as a whole, is chemical sensitivity. Molecules arrive by diffusion, augmented by movements of the medium. All chemicals are sensed when they come into contact with an organism’s body, but, based on how fast they are carried by air currents, we distinguish verbally between the senses of taste and smell. The chemical knowledge possessed by a single bacterial cell exceeds what the most sophisticated organic chemist can measure. And because the sense of taste is based on chemical reactions, noses can be scaled down almost to the size of molecules.

Light sensitivity also evolved early in life’s history on Earth. Bacteria evolved both photosensitivity—light-sensitive spots—and photosynthesis. The former vastly increased the distance over which objects could be detected; the latter generated nearly all of the energy in today’s biosphere. Because light travels in straight lines, at the speed of light, vision potentially provides accurate and time-specific information about shapes of objects and their location in space. Sound has similar properties to light, but location of objects is not given precisely. Animals have evolved complex ways of inferring locations of sources of sounds by comparing events at the two ears.

Nervous systems evolved in part because, by building up parallel computing units, they enabled organisms to speed up and amplify responses to sensory inputs. For motile organisms, speed of response is often extremely important. Parallel computing units also are the units in which preselected filters are stored.

Responses to environmental information are complex because the significance of objects and events varies with the location and needs of observers. An approaching storm may be welcomed by a farmer whose crops are wilting from lack of rain, but despised by members of a family enjoying their first picnic of the summer. A vivid sunset enjoyed by people relaxing on the deck of their home may bring fear to a person walking on an African savanna too far from home to reach it before dangerous nocturnal predators become active. For these reasons, simple correlations between features of objects and aesthetic evaluations are unlikely ever to be discovered, although correlations are likely between types of objects, contents, and aesthetic responses.

Three functional concepts—prospect, refuge, and hazard—have guided recent approaches to the role of environmental aesthetics in problem solving (Appleton, 1975). Prospect refers to the ability of an individual to gather information about an environment with which to evaluate its characteristics and decide how to use it. Environments high in prospect offer rich opportunities for evaluation; environments low in prospect offer fewer opportunities. Refuge refers to the degree to which an environment provides security for an individual from negative agents while exploration and information gathering are taking place. Hazard refers to the dangers to which an individual would be exposed during information-gathering activities.

These concepts have been applied primarily to the initial evaluation and exploration of unfamiliar environments, but when combined with more recent develop-
ments in evaluating environmental information, they are readily applied and extended to a rich array of circumstances in which people must solve problems. Emotional responses that assist in solving these problems necessarily include both positive and negative components. Because they are simpler and easier to study, negative responses are discussed first.

IV. NEGATIVE EMOTIONAL RESPONSES

Many people have strong fearful reactions—phobias—to a wide variety of objects or situations (snakes, spiders, heights, closed spaces, open spaces, blood). The set of objects that evoke fearful responses appears to be the same in all industrialized societies for which data are available (Ulrich, 1993), but few data exist on the prevalence of phobias in developing nations and nontechnological societies. The objects and situations that evoke fearful responses are generally ones that have been associated with threatening situations during human evolutionary history. Precipitous cliffs are dangerous if approached too closely. Closed spaces offer few escape routes; people in wide open spaces are vulnerable to attacks by enemies and dangerous predators. Venomous and predatory animals have been significant sources of human injury and mortality for many millennia. Interestingly, even though they account for trivial numbers of injuries and deaths in modern societies, spiders and snakes are the objects of the most prevalent phobias in Western societies today.

Research on human twins has provided convincing evidence that generic factors play major roles in a wide range of human traits, including animal phobias and fear of open spaces. Imaginative research, especially by psychologists in Sweden and Norway, shows that biologically prepared learning plays a significant role in the acquisition and retention of phobias (see Ulrich, 1993, for an overview of experimental results). The most informative experiments employ a Pavlovian conditioning approach to compare the development of defensive and aversive conditioned responses (learned through repeated exposure or experience) to slides of fear-relevant and fear-irrelevant neutral stimuli. Defensive responses are assessed by recording autonomic nervous system indicators such as skin conductance and heart rate. Experimenters condition initial defensive responses by showing either fear-relevant (snakes or spiders) or neutral stimuli (geometric figures) paired with an aversive stimulus (the unconditioned stimulus), usually an electric shock intended to mimic a bite. This phase of the experiments makes it possible to compare the speed and strength of acquisition of defensive/aversive responses to fear-relevant and neutral stimuli.

Following an initial acquisition phase, the same stimuli are presented many times without reinforcement of an electric shock. This “extinction” phase allows comparison of the rate of extinction of the defensive/aversive response acquired earlier. The general result is that conditioned responses are usually, but not always, acquired more quickly, but that responses to snakes and spiders are reliably more resistant to extinction than responses to neutral stimuli. These responses cannot be the result of prior cultural reinforcement, because conditioned aversive responses to familiar modern dangerous stimuli, such as hand guns and frayed electrical wires, extinguish more rapidly than conditioned responses to snakes and spiders. In addition, aversive responses to fear-relevant natural stimuli can be elicited merely by telling a person that a shock will be administered. Aversive responses to fear-irrelevant natural stimuli cannot be elicited in this manner. People also acquire much more persistent defensive reactions when watching an experimenter’s reactions to fear-relevant than to fear-irrelevant stimuli. Similar results have been obtained by exposing rhesus monkeys to fear-relevant (toy snakes and crocodiles) and fear-irrelevant stimuli (toy rabbits).

Even more striking are the results of “backwardmasking” experiments, in which slides are displayed subliminally (for 15–30 milliseconds) before being “masked” by a slide of another stimulus or setting. Even though the subjects are not consciously aware of having seen the stimulus slide, presentations of natural settings that contain snakes or spiders elicit strong aversive/defensive reactions in nonphobic persons. If a previous conditioning has already occurred, a masked subliminal presentation is sufficient to elicit defensive responses to the feared stimulus (O’hanan and Soares, 1994).

Thus, a rich array of experimental results demonstrates that aversive responses develop more rapidly and persist longer to fear-relevant than to fear-irrelevant natural stimuli. In addition, defensive/aversive responses can develop to natural threat stimuli even though subjects are unaware that they have seen them. Such responses do not develop to neutral or fear-irrelevant stimuli. The patterns of these responses demonstrate the powerful role of biologically prepared learning, and they make adaptive sense.
V. POSITIVE EMOTIONAL RESPONSES

Research and theory are less well developed for positive emotional responses to environmental stimuli than for negative ones. No studies of the genetics of positive emotional responses have been carried out, and positive conditioning studies are relatively rare. This is in part due to the fact that positive Pavlovian conditioning experiments are typically more difficult to conduct than aversive conditioning experiments. In addition, positive emotional responses are very complex and difficult to quantify.

A. Symmetry and Beauty

Extensive information has been gathered on emotional responses to the symmetry of objects. A rich body of literature demonstrates that humans find symmetrical objects, including abstract patterns, woody plants, and human bodies—especially faces—aesthetically pleasing. Moreover, sexual preferences for symmetrical facial features appear to be similar across human cultures. Other vertebrates also are sensitive to very small asymmetries (see Møller and Swaddle, 1997, for a review of the human and animal literature). Conversely, asymmetries typically evoke negative emotional responses, the most extreme of which is selective infanticide of asymmetrical newborn babies in many human cultures.

There are several nonmutually exclusive reasons why positive responses to symmetry are adaptive. One is that symmetrical animals perform better physically than asymmetrical animals, just as symmetrical objects, such as bows, arrows, axes, boats, autos, and airplanes, function better than asymmetrical ones. Asymmetrical individuals may have been exposed to severely stressful environments that disrupted normal development, thereby rendering them less functionally adequate. For example, alcoholic mothers give birth to children with greater developmental asymmetries than do mothers who consumed less alcohol. Thus, it is not surprising that choices of objects to use, the design of objects, and choices of social partners have evolved to favor use of symmetrical objects, including abstract patterns, woody plants, and human bodies—especially faces—aesthetically pleasing. Moreover, sexual preferences for symmetrical facial features appear to be similar across human cultures. Other vertebrates also are sensitive to very small asymmetries (see Møller and Swaddle, 1997, for a review of the human and animal literature). Conversely, asymmetries typically evoke negative emotional responses, the most extreme of which is selective infanticide of asymmetrical newborn babies in many human cultures.

Issues of spatial scale are important during habitat selection. At small spatial scales, the primary decisions involve food and shelter. The key decisions are whether to accept or reject a specific food item or shelter. At medium spatial scales, the key decisions are in which activity to engage and in which patches in the environment those activities will be carried out. At still larger spatial scales, decisions center on whether to initiate major changes in where to carry out activities, that is, to migrate or to shift the base of operations.

Theoretical explorations of these decisions typically include the development of “ideal” models, that is, models that assume perfect knowledge on the part of the decision maker. These models specify the best that an organism can do in solving a problem (e.g., obtaining the most food per unit hunting time). They are unrealistic because knowledge is always incomplete, but they are valuable for assessing the marginal value of additional information and for determining how well an animal using rather simple “rules of thumb” would perform compared to an “ideal” animal. Many of these rules of thumb include aesthetic responses that guide decisions. It is useful to begin with a consideration of medium-scale responses and then turn to microscale responses.

Habitat selection, viewed at medium spatial scales, has served as a perspective for a number of studies on human aesthetic responses to landscape features (Heerwagen and Orians, 1993). Habitats occupied by humans during most of our evolutionary history rarely
provided resources that were reliable long enough to enable permanent occupation of sites. Frequent moves through the landscape were the rule even though traditional sites might be revisited on an annual basis. Because relatively few generations have passed since humans started to live in mechanized and urban environments, evolutionarily based response patterns of humans to landscapes are unlikely to have been substantially modified since the rise of industrialized, urban societies.

Human responses to environmental cues vary with a person's age, social status, and physiological state. Nevertheless, positive responses to indicators of the presence of food, water, shelter, and protection from predators are general. So are negative responses to potential hazards, such as inclement weather, fire, dangerous predators, and barriers to movement. Although no direct evidence yet exists for genetic influences on these responses, a number of evolutionary hypotheses have generated predictions, some of which have been tested experimentally.

One approach is based on the fact that Homo sapiens evolved in African savannas and only recently has invaded other continents. Therefore, landscape features and tree shapes characteristic of high-quality African savannas are expected to be especially attractive to humans today. This hypothesis has been tested by determining the responses of people to tree shapes and by examining the features of "aesthetic environments," that is, those environments, such as parks and gardens, that are designed to make them attractive (Ortans, 1986).

The shapes of trees that dominate savannas are good predictors of the resource-providing capacities of those environments. Therefore, people evolved to find more pleasing the shapes of trees that were prominent in environments that provided the highest-quality resources rather than shapes of trees that dominated poor-quality habitats. Trees that grow in the highest-quality African savannas have canopies that are broader than they are tall, trunks that bifurcate close to the ground, and layered canopies. People on three continents preferred Kenyan Acacia tortilis trees that had highly or moderately layered canopies, lower trunks, and higher canopy width/tree height ratios than trees with narrow canopies and trunks that bifurcated higher above the ground (Heerwagen and Ortans, 1993). Similarly, college students in Australia, Brazil, Canada, Israel, Japan, and the United States preferred trees with broad spreading crowns over conical and columnar trees (Sommer and Summit, 1996).

The changes that landscape architects recommend to their prospective customers are another source of data on human responses to landscapes. Humphrey Repton, an eighteenth-century British landscape architect, presented his clients with "before" and "after" drawings of their estates (Repton, 1807). Because he presumably wished to encourage rather than discourage potential clients, Repton changed landscapes by creating more savannalike scenes, by increasing visual access and penetrability of closed woods, by opening up distant views to the horizon (i.e., increased prospect), by adding refuges and cues signaling ease of movement, and by adding evidence of resource availability, particularly large mammals (Heerwagen and Ortans, 1993).

Although a love of flowers is a pervasive human trait, it is not obvious why an omnivorous primate should take flowers to hospitals, bring them to dinner parties and house warmings, and annually spend billions of dollars on them. Nor is it obvious, given that flowers did not evolve their forms and colors to please us, why we should find them so aesthetically attractive. An evolutionary perspective suggests that flowers evoke strong positive feelings because they have long been associated with food resources. Because flowers precede fruits, flowering plants provide excellent cues to timing and locations of future resources. In addition, flowers may attract animals, especially birds, that are potential human prey. In species-rich environments, paying attention to flowering plants may particularly enhance resource acquisition abilities in the future. Until the nineteenth century, honey was the only natural source of sugar; beekeeping is an ancient human enterprise.

No studies have investigated which traits of flowers evoke strong positive feelings, but the obvious changes produced in many species of flowers by artificial selection—increased size and duplication of floral parts—result in flowers similar to those that historically produced large nectar rewards. Anecdotal evidence suggests that strongly asymmetrical flowers, which usually produce large nectar rewards, are generally more attractive than symmetrical ones, suggesting a possible exception to the general human preference for symmetrical objects. Future studies of human aesthetic responses to flowers are likely to provide interesting results.

C. Foods and Food Selection

Eating is essential for survival. Foraging animals make two major types of decisions: what items are acceptable as food, and which acceptable items are actually eaten during a foraging bout. Aesthetic responses are especially prominent in the first of these decisions. The strong emotional responses associated with foods are...
not surprising because eating requires voluntary ingestion of foreign objects. To eat, the body’s generally aversive responses designed to protect health must be overcome; vomiting and diarrhea are the only postigestion defenses. These decisions are especially difficult for omnivores, who must ingest a variety of foods to achieve a balanced diet in varied environments where the array of available foods differs dramatically over space and time. Omnivores are expected to evolve only general food preselectors and should develop systems for classifying organisms with respect to their potential value as food.

People accept or reject foods for complex reasons (Table III). Direct sensory responses—tastes good or tastes bad—are important, but they are only one component of acceptability. People also make decisions on the bases of anticipated consequences, both physiological and social, of eating a type of food. In addition, particular foods may acquire idiosyncratic features that limit their acceptability to particular situations or cause them to be categorically rejected. Because of the variety of reasons that influence which foods are in the set of acceptables, human cuisines differ more than would be expected simply from knowing what edible resources are available in the environments in which social groups live (Rozin, 1996).

Despite the powerful idiosyncracies of human diets, a few generalizations have emerged. First, most items that evoke disgust are of animal origin. Plant parts rarely evoke disgust in any culture. On the other hand, “inappropriate” items are primarily vegetable in origin (Rozin and Fallon, 1981). Disgust may be an adaptation that deterred our ancestors from eating animal tissues, such as feces, rotting meat, and soft internal parts, all of which commonly harbor large numbers of potentially harmful microorganisms. Microorganisms have the ability to multiply rapidly, so there is no safe dose for ingesting them. Disgust may represent evolutionarily programmed intuitive microbiology. Perhaps the most interesting feature of human cuisines is the stability of the major spices and sauces that characterize them. Humans are remarkably conservative in their food habits and are typically reluctant to try new foods or to abandon familiar ones. Traditional flavorings are high-priority culinary items; immigrant groups go to great lengths and expense to procure them in foreign settings. The deliberate manipulation of food by adding ingredients that reliably alter its taste is a uniquely human behavior. No other animals are known to do so. Which flavorings are used probably evolved in relation to what was available in the environment, but once established they are remarkably persistent. Certain flavorings probably signal that the food is safe. For example, the French refused to eat potatoes until Parmentier, an eighteenth-century French agriculturist, showed how to prepare them in familiar ways with familiar seasonings: butter, cheese, and herbs.

Cross-cultural similarities in which spices are used have evolved because spices inhibit or kill food-spoiling microorganisms. The most widely used spices all have strong antimicrobial properties (Billing and Sherman, 1998); mixes of them, which are common in many traditional recipes, are even more powerful. Not surprisingly, given that disease-causing organisms are more abundant in tropical than in temperate regions, the proportion of traditional recipes containing antimicro-

---

**TABLE III**

<table>
<thead>
<tr>
<th>Primary reason</th>
<th>Acceptance</th>
<th>Rejection</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Sensory</td>
<td>Tastes good</td>
<td>Tastes bad</td>
</tr>
<tr>
<td>B. Anticipated Consequences</td>
<td>No adverse reaction</td>
<td>Allergic reactions, gastrointestinal upset</td>
</tr>
<tr>
<td>Short-term physiological effects</td>
<td>Societal concepts of “healthy foods”</td>
<td>Societal concepts of “unhealthy foods”*</td>
</tr>
<tr>
<td>Long-term physiological effects</td>
<td>Results in acceptance by group (smoking, social drinking)</td>
<td>Causes rejection by group (foods not eaten by that group, e.g., pig in some mideastern cultures)</td>
</tr>
<tr>
<td>Social effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Ideational Appropriateness</td>
<td>Appropriate ritual foods</td>
<td>Inappropriate nonedible contaminants (swil, words)</td>
</tr>
</tbody>
</table>

bial spices is inversely correlated with latitude. Alternative hypotheses—that spices provide micronutrients, disguise the taste and smell of spoiled foods, or increase perspiration and, thus, evaporative cooling—are not supported by the extensive data base (Billing and Sherman, 1998).

D. Restorative Responses

If aesthetic responses evolved because they enabled people to better solve life's problems, exposure to high-quality environments should be restorative, that is, it should reduce feelings of tension and stress. Stress reduction consistently emerges as one of the key benefits reported by recreationists in wilderness areas. Restoration from stress is also reported as a key benefit from time spent in urban parks with savannalike vegetation and water (Schroeder, 1989). Patients recovering from surgery in hospitals with either views of natural vegetation or simulated views that depict natural scenes with water recover more rapidly and have less postoperative anxiety than patients with no access to natural views or who are presented with simulations of abstract designs. Many studies have shown that even a brief exposure to nature, real or via photographs, leads to positive emotional feelings, reductions in stress, and better performance on demanding tasks (see Ulrich, 1995, for a review). Clearly, the affiliative responses people have to nature have important implications for the design of work and living spaces and healthcare facilities that are just beginning to be implemented.

VI. AESTHETICS AND BIODIVERSITY

Although people are strongly attracted to living organisms, it is less clear that the attractiveness of an environment is consistently positively correlated with the number of species in it. On the one hand, for example, the most highly evolved garden traditions—European formal gardens and Japanese gardens—are based on just a few species of woody plants. Landscape designers generally do not like the gardens of botanists because they are cluttered up with plants of too many species! Scenes of environments that contain a jumble of plants of many species receive low scores in psychological tests. Subjects report that they are too difficult to interpret; it is difficult to determine how to enter and use them.

On the other hand, people take great pleasure in finding as many species of birds as they can on a given day and from assembling "life lists" of species they have seen. Generally, seeing more species is better than seeing fewer species. Journeys of hundreds or even thousands of miles to see a rare species not on one's life list are not uncommon in today's mobile society. People are also powerfully attracted to the unusual—rare species, individuals outside the normal range of the species, or individuals present at unusual times of the year or in unusual habitats.

Because these familiar patterns of human behavior have been subjected to remarkably little formal study, we can only speculate about why they have evolved. That environments with intermediate levels of biological complexity should be preferred over both simpler and more complex environments makes sense, because the range of resources present in an environment and the ability to find and use those resources probably peak at intermediate levels of complexity. Simple environments have too few resources; complex ones have so many that choosing among them becomes difficult. Developing a suitable classification system to guide responses to the components of complex environments may be especially difficult. People may have evolved to respond to rare and unusual events because they provide new information about the state of the environment. Not all novel events are associated with something important, but it may be best to pay attention to them to find out if they are rather than to ignore the signals. Novel events may indicate that current patterns of use of the environment should be altered.

VII. AESTHETICS AND COGNITION

The human mind evolved into its current form long before the invention of agriculture and the dawn of the Industrial Revolution. It evolved its special characteristics in the service of our hunter-gatherer ancestors, who almost daily faced serious challenges from their physical, biological, and social environments. During the past century, many scientists found it difficult to explain why the demands of functioning in a preagricultural environment should have favored the evolution of the vast complexities of the human mind, which is capable of the many feats of which we are so proud. The apparent paradox arises because of a failure to appreciate that succeeding as a forager is probably more complicated than playing chess or doing calculus. To outrun nature, people need to use some form of intuitive scientific thinking, develop abstract conceptions, think about the future, compete in a rich social environment, and plan tactics. Aesthetic responses are basic components in all of these responses.
Mathematicians describe their search for important theorems as an aesthetic experience; they think of proven theorems as being beautiful. In fact, scientists in all fields regularly describe their models, experiments, and results in aesthetic terms. Thus, aesthetic sensibilities are imbedded deep in the human mind, having evolved in the service of choice of mates, habitats, and foods. An aesthetic sense functions as a holistic filter that helps the mind search efficiently for good solutions. It can do this because the filter has been molded by countless evolutionary experiences. Perhaps one of the most satisfying results of having a mind that was molded by natural selection is that we generally enjoy doing things that are good for us to do.

See Also the Following Articles

SOCIAL AND CULTURAL FACTORS • HISTORICAL AWARENESS OF BIODIVERSITY • HUMAN IMPACT ON BIODIVERSITY • OVERVIEW • LITERARY PERSPECTIVES ON BIODIVERSITY • TOURISM, ROLE OF

Bibliography


AFRICA, ECOSYSTEMS OF

J. M. Lock
Royal Botanic Gardens, Kew

I. Introduction
II. Major Environmental Factors of Continental Africa
III. Major Phytogeographic and Ecoclimatic Zones of Continental Africa
IV. Major Ecosystems of Tropical Africa
V. Madagascar
VI. The Future for African Ecosystems

GLOSSARY

fynbos Habitat type in southern Africa that is characterized by thickets and low shrubs, in which fire plays a dominant role in ecosystem maintenance. Plant endemism is particularly high in these areas.

miombo Woodland habitat type widespread in south-central Africa, characterized by numerous species of the tree genus Brachystegia and Isoberlinia, which form nearly closed canopies. Fire is an annual event in this habitat, which supports relatively low populations of large mammals.

phytochorion (pl. phytochoria) Region within which a substantial proportion of the flora is endemic.

THE HUGE CONTINENT OF AFRICA STRADDLING THE EQUATOR, extending to 37°N and 35°S. It has no marginal oceanic trenches and subduction zones and so lacks the extensive mountain ranges of the Americas, but much of the southern half of the continent is a high plateau close to 1000 meters above sea level, broken only by the southern extension of the Great Rift Valley and the somewhat lower basin of the Congo (Zaire) River. Its great latitudinal range gives it an enormous variety of climates, and this variability is reflected in an extreme diversity of ecosystems. There are three major climatic zones: two at the extreme north and south of the continent, where the main rainfall season is the winter, and one in central Africa, where the rains fall mainly in the hot summer season. The two winter rainfall areas are distinct from each other and from the tropical parts of Africa. These regions are briefly considered here; a fuller treatment appears in the entry for Mediterranean Ecosystems.

I. INTRODUCTION

White (1983) classified the vegetation of Africa into a number of phytochoria—regions within which a substantial proportion of the plants are endemic. These regions are also useful in helping to define zoogeographical regions, and comparisons can also be made with neighboring continents. Thus the Mediterranean phytochorion is most closely related floristically to southern Europe and the Middle East. The Somali–Maasai phytochorion, which occupies the Horn of Africa and regions

Encyclopedia of Biodiversity, Volume 1
Copyright © 2001 by Academic Press. All rights of reproduction in any form reserved.
south to central Tanzania, shows floristic relationships with parts of Arabia and the western part of the Indian subcontinent. On the other hand, about 80% of the plant species that occur in the Guineo–Congolian phytochorion, which occupies the Congo Basin and extends westwards to Liberia, are endemic. The Cape Region is also extremely rich in species, almost all of which occur nowhere else.

The large island of Madagascar, off the southeast coast of Africa, has been isolated from the rest of the continent for over 100 million years and its flora and fauna are very different indeed from those of Africa. The ecology of Madagascar is treated, in brief outline, in Section V.

II. MAJOR ENVIRONMENTAL FACTORS OF CONTINENTAL AFRICA

A. Geology

Much of southern tropical Africa is a plateau lying at or near 1000 m above sea level, most of which has not been submerged since the Tertiary period. This plateau is formed of Pre-cambrian ("basement complex") rocks, mainly igneous and much metamorphosed, which have been heavily and deeply weathered. The topography of much of the plateau is gently rolling, often with isolated rocky inselbergs, and sometimes with flat-topped hills capped with secondary ironstone, a product of prolonged weathering in high temperatures.

In eastern Africa the plateau has been cut by the two branches of the Great Rift Valley, which extends from Israel to the Zambezi River. Rifting seems to have begun in the Miocene, and was (and still is) accompanied by volcanic activity. In Ethiopia great sheets of basalt were erupted and now form the Ethiopian Highlands. Scattered volcanoes line the branches of the Rift Valley; some, such as Elgon and Kilimanjaro, are now extinct or dormant, and others, such as Ol Doinyo Lengai (Tanzania) and Nyiragongo (Zaire), are still active. The Rift Valley contains numerous lakes, some relatively shallow and often saline (e.g., Lake Chala (Ethiopia), Lakes Turkana and Bogoria (Kenya), and Lakes Natron and Eyasi (Tanzania)) and others deep and freshwater (Lakes Albert and Edward (Uganda/Zaire), Lake Tanganyika, and Lake Malawi). The blocks between the branches of the Rift Valley have tilted as part of the same tectonic disturbances. This has impeded or even reversed the flow of rivers, and led to the formation of extensive shallow lakes, such as Lake Victoria, and huge areas of swamp.

B. Climate

The climate of Africa (excluding the extreme north and south) is determined by the movement of the Inter-Tropical Convergence Zone (ITCZ). As the apparent position of the sun moves north and south with the seasons, the area where the sun is immediately overhead is heated more than the areas to the north and south. The heated air rises, and air is drawn in from north and south to replace it. The convergence of these two air masses, and the rising of the heated air, causes rain. The rain belt tends to lag behind the sun, so the rain belt reaches its farthest north in July–August, rather than at the summer solstice (June 21st), and its farthest south in January–February, not at the winter solstice (December 21st). Away from the equator, the rainfall produced by the ITCZ as it moves in one direction merges into that produced during its return, so that there is a single rainy season. Near the equator, however, there is a tendency for the two passages of the ITCZ to be separated by a dry season, giving two rainy seasons and two dry seasons in each year, in the wettest regions these dry gaps hardly occur so that rainfall occurs more or less throughout the year.

Superimposed on this basic seasonal pattern is variation in the total annual rainfall. As a rule, totals decline away from the equator, and also from west to east. In East Africa, air coming from the north-east has passed over the dry Arabian land-mass, not the sea as in most of the rest of the continent, and this explains the lower rainfall. Areas close to the sea tend to be wetter than inland areas, except where, as in south-western Africa, cold sea currents offshore produce foggy conditions but low rainfall. Extremes include Sierra Leone, where the coastal capital, Freetown, has over 250 cm of rain each year, most of it falling in four months. Debundscha Point, on the seaward margin of Mt. Cameroon, receives over 10 m of rain in some years; this comes from the combination of a warm sea and rapidly rising ground close to the shore. Altitude also affects rainfall; totals appear to increase up to about 3000 m and then decline; the upper regions of Mt. Kilimanjaro are a virtual desert with probably less than 25 cm of rain each year. At middle and higher altitudes, the rain is supplemented by clouds from which droplets of moisture condense onto leaves and drip to the ground, and here humidity is extremely high for much of each day.

C. Soils

The soils of much of Africa are developed on very ancient land surfaces and are therefore the product of
long weathering. Forest soils, developed under high rainfall, are generally very strongly leached, rich in clay, and yellowish brown in color. Drier forests have redder, somewhat less nutrient-poor soils. Savanna soils are often brownish and, again, rich in clay. On the older plateaux there are often regular sequences of soils ("ca-
tenac"), with leached sandy or gravelly soils with kaolin-
ite clay minerals on the hill and ridge tops, finer-grained soils on the hill slopes, and deep, dark-colored clays in the valleys. These valley soils, which shrink and crack as they dry and swell again when they are wetted, become extremely sticky and difficult to work when they are seasonally flooded or waterlogged. These are the so-
called "cotton-soils," rich in montmorillonite clay min-
erals, in which motor vehicles so easily become bogged.

As a result of prolonged weathering, most of the soils of the African plateau are nutrient-poor, with phos-
phorus in particular often being in short supply. Some of the more volatile elements, such as nitrogen and sulfur, can be lost in the smoke in the regular dry season fires.

In a few areas, such as the "copperbelt" of southern Zaire, northern Zambia, and the Great Dyke in Zim-
babwe, there are outcrops of soils rich in metals such as copper, cobalt, and chromium. These produce soils that are toxic to many plants. Often trees are scarce or absent, and the vegetation is mainly composed of grasses and a range of specialized herbs, often very local in their distribution.

**D. Fire**

The highly seasonal climates of most of tropical Africa provide ideal conditions for widespread vegetation fires. During the wet season, grass growth is rapid, but during the dry season the aerial parts of the grasses dry out and burn very easily. Most of those parts of Africa with a seasonal climate are burned by vegetation fires every year. The exceptions are areas with a rainfall too low to produce the necessary volume of fuel, or so high that the tree cover is continuous and dense, preventing the growth of enough grass to provide fuel. The carbon dioxide produced by this biomass burning is highly significant in the annual atmospheric carbon dioxide budget of the world. There can be little doubt that fire has been important in African vegetation patterns for a very long time. Natural fires, at long intervals, can be started by lightning, as well as by volcanic activity. Once humans began to use fire—and some have sug-
gested that this may have been as much as a million years ago—fire frequency would have increased. The acquisition of the ability to make (not just use) fire would have increased this frequency further, and an-
other increase surely came when safety matches became widely available.

The question of the influence of fire on the distribu-
tion of vegetation types in Africa has been extensively discussed. On the one hand there are the views of work-
ers such as A. Aubreville, who considered fire to be a major destructive force that had probably been respon-
sible for the destruction of vast areas of the drier forests of Africa. The converse view, that fire has been a feature of African vegetation for so long that it is now a funda-
mental part of many ecosystems, is probably more widely held. Though fire may sharpen boundaries be-
tween vegetation types, it does not greatly affect their distribution. Assessment of the importance of fire also depends on viewpoint. Foresters generally disapprove of fire and expend much effort in its prevention, whereas pastoralists use fire extensively to remove old grass and to stimulate the development of young growth, as well as to check development of woody plants. Foresters, who want trees, generally recommend that fires should be started early in the dry season, when the fuel is still not completely dry ("early burning"). Pastoralists, who want grass, prefer a fire late in the dry season when the fuel is completely dry and damage to woody growth is maximal ("late burning").

The presence of specific adaptations to fire in a num-
er of plants is a further argument for fire having been an important factor for a long time. Several trees, such as the shea-butter tree (*Vitellaria paradoxa*) and species of *Combretum* and *Pterocarpus*, have seeds that, when they germinate, produce what appears to be a radicle that pushes into the ground. However, close examina-
tion shows that at least part of this is hollow and is actually formed by fusion of the stalks of the cotyledons (seed-leaves), which remain within the seed. The termi-
nal bud, which lies between the bases of these stalks, is thus carried down below ground level so that it is at least partially protected from fire. Once well buried, the bud starts growth and the shoot breaks out through the side of the apparent root. It elongates and eventually emerges from the ground, but not before producing a number of reduced scale-like leaves, each with a bud in its axil. If the terminal bud is burned away, the lateral buds in the axils of the scale leaves start to develop, giving the seedling another chance of establishment.

Many of the grasses of the fire-swept regions have a long awn on their grains. This awn is hygroscopic, twisting and untwisting with daily changes in humidity. The tip of the grain is sharply pointed and bears stiff hairs. If the point enters a crack, the hairs prevent it from coming out again, and the regular movements of
the awn will tend to drive it deeper into the soil, until it is protected from the main heat of the fire.

E. Past Climatic and Environmental Fluctuations

Tropical Africa has not escaped the climatic fluctuations of the last 2 million years. However, interpretation of changes in Africa is complicated by the effects of rifting and volcanism, particularly in the east, as well as by the presence of many isolated mountains and mountain ranges. Evidence is also relatively scarce because the peat and other organic deposits that are widespread in cooler regions do not accumulate nearly so widely in tropical lowlands because high temperatures speed decomposition processes. The peat deposits that do exist are mainly on mountains, and so may not present a representative picture of the much more extensive lowlands.

However, there are other kinds of evidence available, such as the extent of active and inactive (fossil) sand dunes. In West Africa the fossil dunes extend 400–600 km south of the present limit of active dunes; the arid period during which these fossil dunes formed appears to have been 12,000 to 20,000 years ago. South of the equator, the Kalahari Sands extend far to the north and east of any presently active dunes, reaching the Zaire River in the north. It appears that the most recent period of active dune growth coincided with that north of the equator. Lake levels, as marked by raised beaches, can also provide evidence, at least of times when lake levels stood higher than they do now. The study of plant distributions can also sometimes suggest former connections between similar but now isolated vegetation types. For instance, several tree genera with most of their species in West Africa have one or more species in the Eastern Arc montane forests of Tanzania. This suggests a connection between these forests areas in the distant past.

In general, periods marked by cold in high latitudes are correlated with periods of drier and cooler climate in tropical Africa. During these periods, forested regions contracted and “savanna” regions spread. In the driest times, forests seem to have contracted into a number of refugia, in Sierra Leone and Liberia in West Africa, as well as Cameroon and Gabon, eastern Zaïre, and eastern Tanzania. It is now agreed that most of the evidence shows that there was an arid period between 12,000 and 20,000 years ago. Prior to this there were certainly similar fluctuations, perhaps of greater magnitude, but their dating and extent are still somewhat in dispute. Since 12,000 years ago there have also been changes; the wettest period seems to have been between 8000 and 4000 years ago. At this time, Neolithic pastoralists inhabited large areas of the Sahara and left rock paintings showing elephants, griffes, and antelopes in areas now far too dry to support them.

III. MAJOR PHYTOGEOGRAPHIC AND ECOCLIMATIC ZONES OF CONTINENTAL AFRICA

A. Mediterranean North Africa

This region borders the Mediterranean Sea, with winter rainfall and hot dry summers. It includes Morocco, Algeria, Tunisia, Libya, and Egypt. The wettest part is the west, and here also there are the high mountains of the Atlas range. Thousands of years of human settlement, agriculture, and grazing of domestic animals have greatly altered the ecosystems. In Roman times there were certainly lions (Panthera leo) (and therefore a substantial prey population) and probably elephants (Loxodonta africana) in this region, but all are now gone.

The wetter parts were probably originally covered with forest, but this is now represented only by tiny fragments. Celtis australis and Pistacia atlantica may have been important trees in the original forests. The drier forests were (and in places still are) dominated by evergreen oak (Quercus ilex), which casts a dense shade in which few other species can grow, or by cork oak (Quercus suber). There are also coniferous forests of species such as Aleppo pine (Pinus halepensis) and North African cedar (Cedrus atlantica). Other parts are covered by scrub, similar in physiognomy to the chaparral of California and the fynbos of the Cape Region of South Africa, made up of shrubs with small hard (sclerophyllous) leaves, such as the kermes oak (Quercus coccifera), wild olive (Olea europaea), and, in very degraded sites, the dwarf palm Chamaerops humilis. The shrubs are fire-resistant, sprouting from the base after fires. The gaps between the cushions and, lower down, between the sclerophyllous shrubs support a rich herb flora including many annuals (Fabaceae are abundant and diverse) and many plants springing from underground bulbs or corms (geophytes). These grow during the late winter and spring, flower, and then dry up in the baking heat and drought of the summer. At high altitudes, between 2800 and 3800 m, there is often a low scrub made up of spiny cushions, sometimes graphically referred to as hedgehog heath.
B. The Cape Region of South Africa

The Cape Region also has a Mediterranean climate, although, being in the Southern Hemisphere, it enjoys a hot dry summer when North Africa is having a cool wet winter. Like North Africa, thicket and low scrub (known locally as fynbos) are the main physiognomic vegetation types, and fire is a regular influence on the vegetation. Many species appear to be fire-adapted. Some species, such as the red-flowered, lily-like species of *Cyrtanthus*, flower only after fires, stimulated either by chemicals in the smoke or by the greater daily fluctuations in soil temperature that follow removal of the vegetation cover. Others, such as species of *Leucadendron* (a *Protea* relative), retain their seeds on the parent plant and only release them after fire. The range of fire-adapted species in fynbos, and the many ways in which they respond to fires, suggests that fire has been a feature of this vegetation type for a very long time.

The region is extremely diverse geologically and has numerous isolated mountain ranges. These factors, combined with long isolation, have given rise to an extraordinary diversity of plant species—estimates vary from about 7000 species in the 71,000 km² of the region (White, 1983) to 8600 species in an area of 91,000 km² (Cowling and Richardson, 1995). The genus *Erica* (heaths) has over 500 species in the Cape Region; other extremely diverse genera include *Aspalathus* (Fabaceae, 250 species) and *Muralius* (Polygalaceae, 100 species). The family Proteaceae, including 85 species of *Protea*, is prominent among the larger woody plants, and is associated with an endemic pollinator, the Cape Sugarbird (*Promerops cafer*). The lower-lying and more level areas have largely been converted to agricultural land, but the mountain ranges continue to provide refuges for the endemic flora, although invasion by woody species introduced from other regions with a similar climate is a major problem.

Originally there was a rich fauna of large mammals but these were heavily hunted by European settlers. Some still survive in reserves, but the quagga (a form of zebra, *Equus quagga*) is extinct, and the bontebok (*Damaliscus dorcas dorcas*) and white-tailed gnu (*Connochaetes gnou*) survive only on enclosed farms.

C. Tropical Africa

The region known as tropical Africa takes in most of the continent. The vegetation is determined by the climate, which is highly seasonal over much of the region. There are virtually no parts of Africa without some kind of a dry season; truly ever-wet climates like that of Singapore (where two weeks without rain is a drought) are virtually absent. The major exception lies along the equator in West Africa; here the dry season lasts a month or less. The equatorial regions of eastern Africa, however, lie within the rain-shadow of the Arabian landmass, and here, even on the equator, rainfall is low. Furthermore, there tend to be two rainy seasons rather than one. Both of these vary in intensity, and the gaps between them also vary in length, so that this region tends to suffer more than most from periodic droughts.

Tropical Africa can be divided into two parts. South and east of a line from Ethiopia to the mouth of the Zaire River, most of the land forms a dissected plateau lying at about 1500 m above the sea. This is split from north to south by the Great Rift Valley, which extends from Israel through the Red Sea, then across Ethiopia, Kenya, and into Tanzania. A western branch runs south through Uganda, along the western side of Tanzania, and ends in Malawi. Between the branches of the Rift Valley, the land surface has tilted in places, disrupting river flows and producing the huge but shallow Lake Victoria as well as the extensive swamps of Uganda. To the north and west of the high plateau, the general land surface is much lower. Here again rocks of the basement complex underlie most of the region, but younger rocks are found here and there. Most of tropical Africa is covered by woodland and various forms of savanna, with forest occupying the basin of the Zaire River, and drier bushlands, thickets, and grasslands in the equatorial regions of eastern Africa.

IV. MAJOR ECOSYSTEMS OF TROPICAL AFRICA

A. Forest

True tropical forest is confined in Africa to two main blocks: the basin of the Zaire (Congo) River, extending north and west through Cameroon and Gabon into southern Nigeria and east to the borders of Uganda, and a region farther to the west stretching from western Ghana through Ivory Coast and Liberia into Sierra Leone. Elsewhere, isolated forest patches are found in and around the Ethiopian Highlands, and along the East African coast. Most forests are found where annual rainfall exceeds 120 cm and where the dry season is no more than four months long. At the margins, small differences in water availability can greatly affect the vegetation; forest may extend far into grassland along the banks of rivers as “gallery forest,” and non-forest
African forests are poor in species compared to those of South-east Asia and the Amazon Basin, but they are still species-rich places. Surveys in Ghana found forests in which a 25-by-25-m patch contained more than 200 higher plant species. These surveys also showed that the wettest forests are the most species rich. They do not, however, contain the tallest trees, which are found in drier, partially deciduous forest, perhaps because the soils there are not so leached of nutrients as those in the wetter areas. These drier forests are, at least in Ghana, richest in the number of timber trees. The most prominent plant families in the forest canopy are the mahoganies (Meliaceae) and legumes (Fabaceae). There are also ungulates in the rain forests of Africa, ranging in size from the tiny royal and dwarf antelopes (Neotragus pygmaeus and N. batesi, 23–30 cm at the shoulder) and the taxonomically very distinct water chevrotain (Hyemoschus aquaticus, only 30–40 cm), through several species of duiker (Cephalophus, small antelopes), to the large bongo (Tragelaphus eurycerus). The Ituri forests of eastern Zaïre harbor, the okapi (Okapia johnstoni), whose closest relative is the giraffe. All of these are browsers on the leaves of forest shrubs and herbs.

The largest forest animal is the elephant. Forest elephants are usually recognized as belonging to a separate subspecies from those of open country; they are smaller and tend to have smaller and straighter tusks. Elephants are voracious feeders; their inefficient digestive systems mean that they must feed for a large proportion of each day to obtain enough nourishment. They are particularly fond of clearings in forest, because here more of the foliage is close to the ground and accessible, and by concentrating their feeding in such areas they may prevent tree development and perpetuate the clearings. Elephants also eat large quantities of fruit. Plants such as Pandania octosa, Deypatra, and Balantia wilsoniana have such large seeds that it is hard to know what animal other than an elephant could possibly disperse them. Attempts to germinate seeds of Panda that have not passed through an elephant have been unsuccessful. The effects on such plants of the widespread decline in elephants may be severe.

**B. Seasonal Tropical Vegetation**

Much of tropical Africa is occupied by vegetation that develops under a climate in which the year is divided into dry and wet seasons. The vegetation of these areas is of many kinds, ranging from woodland with an almost...
closed canopy at one extreme to dry open grasslands at the other. All of these vegetation types have been referred to as savanna, but recently there has been a tendency to attempt to discard this term. It has been argued that a word that can mean almost anything from closed forest to open grassland is too vague to have any utility in ecological discussions. White's classification avoids the term, and instead uses “woodland,” “wooded grassland,” “grassland,” “bushland and thicket,” and “shrubland” to define the various physiognomic vegetation types more precisely. These terms will be used here, but where the term ‘savanna’ has wide currency, as in much of West Africa, it will be used.

In tropical seasonal vegetation, water is abundant and growth can be rapid during the wet season, but in the dry season the grass rapidly dries out and becomes flammable; fires are frequent and often annual. It is somewhat paradoxical that the fiercest fires occur where the rainfall is highest—and where, therefore, biomass production during the growing season is highest. Over much of the savanna regions of Africa there is one wet season and one dry season each year, but near the equator, particularly in East Africa, there can be two wet and two dry seasons each year (see Section II,B), so that the seasonal cycle is six months rather than twelve. The herds of wild ungulates and their associated predators for which Africa is so famous are all found in areas of seasonal tropical vegetation. Densities vary enormously, now largely because of human pressures, but there can be no doubt that densities varied greatly before human impacts became significant.

1. Woodland

In West Africa, where the isohyets (lines of equal rainfall) are more or less parallel to the coast, there are belts of vegetation that follow the isohyets. These were classified many years ago by the great French botanist Auguste Chevalier into three savanna types: Guinea, Sudan, and Sahel. In West Africa, use of the term ‘savanna’ is long established, and Chevalier’s zones are still recognized and used today. The southernmost zone, with the highest rainfall and the shortest dry season, is the Guinea Zone.

Woodland has been called by many different names in the African ecological literature. Tree savanna is perhaps the most widely used, as well as the more descriptive “tall grass–low tree savanna.” In West Africa, woodland is referred to as “Guinea Savanna,” and in francophone countries it is often called “forêt claire.” This is a confusing term, because it implies that woodland is essentially the same as true forest (“forêt dense”). This is not the case; surveys in Nigeria of adjacent forest and woodland have shown that they have very few species in common. Physiognomically, woodland consists of a single tree layer, sometimes with an almost closed canopy. Lianas are absent, as are epiphytes. Grass covers the ground beneath the trees.

The trees are often misshapen because the frequent fires kill the growing points and cause branching. Because grass productivity is high, fires are more intense than in other zones. The bark of many of the trees is thick and furrowed, which insulates the delicate growing tissue (cambium) from the heat of fires. Regeneration is difficult as young seedlings are in the hottest part of the fire. Some species, such as the shea-butter tree (Vitellaria paradoxa), have specialized germination mechanisms that may help to overcome this. Some grass seeds, mainly those from the subfamily Andropogoneae, also have a mechanism for evading fires (see earlier discussion).

The greatest obstacle to the development of the woodlands of Africa (many would say their most valuable conservator) is the presence of the tsetse fly (Glossina morsitans, G. palpalis, and other species). This biting fly is the vector of trypanosomiasis, a lethal protozoan disease of cattle and of sleeping sickness in humans. Wild game animals are immune to trypanosomiasis, although they can carry the protozoan in their blood. Control of the fly has been attempted in many areas, using methods of varying destructiveness to the ecosystem. The fly requires shaded places in which to rest, so tree clearance has been used. Aerial spraying with DDT and other non-specific insecticides has now been replaced by highly specific treatment with persistent insecticides of the undersides of large tree branches, where the flies rest. Other attempts to break the cycle of transmission have included the shooting of wild ungulates. Researchers continue to look for effective and environmentally sound ways of tsetse control. An alternative that has been tried in some areas is to leave the flies alone and harvest the wild game instead of the cattle. This has an advantage in that each member of the diverse ungulate community occupies a different niche, so that the carrying capacity and yield are, at least in theory, higher than those from a monoculture of cattle.

The Guinea Savanna woodlands of West Africa have as their characteristic tree species Lophira alata (Ochnaceae) and Daniellia oliveri (Fabaceae). Slightly drier woodlands have species of Isoberlinia (Fabaceae). The main grasses are species of Andropogon and Hyparrhenia, both with awns on their grains. These large grasses stand 2–3 m high when mature. Many agronomists have been misled into thinking that this high
standing crop also implies a high animal production potential, but this is not so. When the grasses are at their tallest, they are made up largely of stem, which is hard, woody, and of very low nutritional value. After a fire, and at the beginning of the rains, nutritious foliage is produced, but this does not last long.

This extreme seasonality in the quality and availability of grass may go a long way towards explaining the relative scarcity of large wild animals in this zone. Variability of grass may go a long way towards explaining the foliage is produced, but this does not last long. A fire, and at the beginning of the rains, nutritious foliage is produced, but this does not last long.

The corresponding zone to the Guinea Savanna south of the equator is the miombo woodland. "Miombo" is the local name for one of the commonest trees, *Isoberlinia tomentosa* (Fabaceae), but by far the most widespread and abundant trees belong to the genus *Brachystegia* (Fabaceae). There are about 25 species of this genus in the woodland regions of southern-central Africa, all fairly similar to one another, mostly variable, and difficult to identify. Miombo covers vast areas of southern Tanzania, Zambia, northern Zimbabwe, southern Zaire and south-eastern Angola. Miombo is physiognomically rather different from Guinea Savanna; the trees are usually taller, often with an almost closed canopy, and the grass is somewhat shorter. However, fires are virtually an annual event. Although at first sight miombo woodland can appear rather monotonous, over much of its extent it shows considerable local variations related to topography. The ancient land surface on which it grows is weathered into a pattern of gentle hills and valleys. The tops of the hills and ridges are dry and well drained, but there are often rock outcrops (inselbergs) around which water accumulates so that the vegetation around them is tall and dense. The rocks themselves, if not sloping too steeply, often bear a shallow turf that dries out completely during the dry season, but during the wet season it becomes saturated and supports many small ephemeral plants such as species of *Urticaria* (bladderworts), *Xyris* (yellow-eyed grasses), and various sedges (Cyperaceae). There are often mats of the sedge *Aforostrips*, which can dry out completely and then rehydrate and resume growth when water is once again available. In southern Africa there can be clumps of the shrub *Myriophyllum fluviatile*, the resurrection plant, which behaves in the same way (i.e., it is potliko-hydric). These rocky hills are often home to a small and specialized antelope, the kipspringer (*Oreotragus oreotragus*), and to rock hyraxes (*Procavia johnstoni* and *Heterohyrax brucei*).

The typical miombo woodland occurs on the hill slopes, but even this is not uniform. Termite mounds, built by species of *Macrotermes*, can be up to 4 m high and 10 m in diameter. These termites are fungus gardeners, they collect dead plant material from a wide area and carry it into the center of the nest. Here it is formed into a honeycomb-like structure that is colonized by fungi that can break down the cellulose and lignin in the wood. The termites feed on the fungal hyphae and are thus indirectly nourished by the wood. A by-product of this activity is the concentration of mineral nutrients and fine soil fractions in the mound so that, somewhat paradoxically, it becomes a favorable habitat for plant growth. Termite mounds bear much denser woody vegetation, often casting enough shade to exclude grasses and therefore fires.

The valley bottoms in the miombo ecosystem are seasonally waterlogged, and this prevents the growth of most trees. Tall grassland occupies this habitat, the grasses often forming well-spaced tussocks between which small plants can grow early in the wet season. The soils of these valley grasslands are often black clays that are extremely sticky when wet and very hard when dry.

As in the Guinea Savanna, large mammal biomass is low in miombo. There are, however, many more species than in Guinea Savanna. Roan antelope occur, as well as the closely related sable antelope (*Hippotragus niger*) with its spectacular scimitar-shaped horns. Hartebeest (*Alcelaphus buselaphus*) occur as do elephant where there is permanent water within reasonable range. The spectacular greater kudu (*Tragelaphus strepsiceros*), with long, spirally twisted horns, is very much an animal of the miombo. Herds of Cape buffalo (*Syncerus caffer*) can be found, again where permanent water lies within a reasonable distance.

The annual cycle of the miombo woodland follows the rains. The trees often begin to produce their new leaves before the first rains, presumably drawing on reserves of water from deep in the soil. This flush of new foliage is often reddish, perhaps because of secondary compounds that make the leaves unpalatable to predators. Flowering takes place fairly early in the wet season, giving the often bulky pods and seeds time to develop before the dry season begins. The pods ripen and burst to scatter their seeds during the dry season. The extent to which the leaves are shed varies from place to place and year to year according to the degree of drought and the intensity of the fire.

Drier woodlands are found both to the north and
south of the equator. To the north, they fall into the zone called "Sudan Savanna" by Chevalier. This zone has a longer dry season and a lower annual rainfall than the Guinea Savanna. Typical tree species include Parkia filicoida and Piliostigma thomningii, but this zone has a much higher human population and is more cultivated than the Guinea Savanna. There are a number of possible explanations for this; maybe all play a part. Because it is farther from the coast, it was less heavily raided and depopulated during the time of the slave trade. The relatively smaller growth of grass makes land clearance and maintenance easier, and the lower rainfall means that the soils are less intensively leached and therefore rather more fertile. Whatever the reason, this zone now supports high human populations who cultivate groundnuts, sorghum, and various kinds of millet. Many of the remaining trees survive because they are of some use; Parkia seeds and the pulp that surrounds them are a useful food, as are the leaves and fruit-pulp of the baobab tree (Adansonia digitata).

In southern Africa the corresponding zone is probably mopane woodland, which is not very intensively settled or cultivated. It is dominated by a single tree species, Colophospermum mopane (Fabaceae: Caesalpinioideae). Mopane grows in areas that are hotter and drier than those occupied by miombo. It is unusual in its family in being normally wind pollinated. Mopane woodland can support a wide range of large mammals, including elephant, black rhino (Diceros bicornis), eland, and impala (Aepyceros melampus).

2. Bushland and Thicket

The drier lowlands of eastern Africa and the Horn of Africa are occupied by woody vegetation that is low-growing (less than 10 m tall), often spiny, and leafless for a substantial part of the year. The rainfall in this region is always somewhat unpredictable, tending to fall in two wet seasons in each year, either or both of which may fail. The most obvious components of the vegetation are species of Acacia (Fabaceae: Mimosoideae). Almost all of these bear paired spines, sometimes hooked and very sharp. Another genus that is abundant and diverse here is Commiphora (Burseraceae); all the trees in the genus have a resinous scent and one is the main source of myrrh. Another genus of the region, Boswellia (Burseraceae), yields the resin known as frankincense.

Many other tree and shrub species also occur so that this vegetation type is very rich in species. Few of these are common to the wetter woodland regions, and the Somalis–Maasai area is recognized by White as a separate phychochorton, that is, a region with a distinctive flora. The floristic relationships of this area lie with southern Arabia and the north-western part of the Indian peninsula rather than with the rest of Africa. Somalia and northern Kenya in particular are rich in endemic plant species, perhaps partly because of long isolation, but also because of the existence of specialized habitats such as regions where the underlying rock is either limestone or gypsum (calcium sulfate) and the coastal strip near Obbia (Somalia), where low-growing plants cover fossil sand dunes just inland from the coast. The animals of this area are also distinctive. Grevy's zebra (Equus grevyi) and the wild ass (Equus africanus) are virtually confined to this part of Africa, and the latter is very scarce. The very handsome and distinct reticulated giraffe (Giraffa camelopardalis reticulata) is confined to Somalia, southern Ethiopia, and northern Kenya. The beira (Dorcatragus megalotis), a specialized long-legged duiker, is confined to Somalia. The four species of dinkon (Madoqua), tiny antelopes that live alone or in pairs in bushland, in territories marked by dung piles, are virtually confined to this vegetation type. Several species of gazelle are also confined, or almost so, to this region. Finally, two remarkably long-necked gazelles, the dibatag (Ammodorcas clarkei) and the gere-nuk (Lithocranus walleri) occur only here; both are exclusively browsers, stretching their long necks up into the bushes and often standing on their hind legs to reach taller growth. Their narrow muzzles allow them to pick out small leaves from between dense twigs or thorns.

The islands of Socotra and Abd-el-Kuri lie to the east of the tip of the Horn of Africa. Both have undoubtedly been isolated for a very long time and have no indigenous large mammals (although introduced livestock now abound). This has allowed the development of a remarkable flora that, although clearly part of the Somalia–Maasai Region, includes numerous endemics such as Dendrosicyos socotranus (a tree-forming representative of the cucumber family), Dorsenia gigas (Moraceae), and Adenium socotranum (Apocynaceae).

3. Grasslands

Many of the grasslands of tropical Africa are the product of some special condition of the soil that prevents the growth of trees; seasonal waterlogging, shallow soils subject to extreme seasonal droughts, and high concentrations of metals such as copper and cobalt all lead to local grasslands. However, near the equator in eastern Africa there are extensive areas of grassland, sometimes with scattered flat-topped acacia trees (Acacia tortilis) or thicket clumps. Most of these areas lie between 1000 and 2000 m above sea level. The commonest and most
prominent grass is the red oat grass (*Themeda triandra*). This species thrives under a regime of annual burning and light grazing, but it is vulnerable to overgrazing. Dense tussock-forming grasses such as *Sporobolus pyramidalis* tend to replace it if the grazing pressure is excessive.

These grasslands have long been the territory of pastoral peoples such as the Maasai, and it is possible that these people, who are well aware of the effects of fire on vegetation, have used it over the millennia to alter the balance from woodland or thicket to grassland on which to pasture their cattle. These peoples have also long lived in close proximity to huge populations of wild ungulates and their predators. These ungulate populations often make seasonal migrations to make best use of their range, such as that which straddles the Tanzania–Kenya border in the Serengeti and Mara region. The main species in these migratory populations are wildebeest (*Connochaetes taurinus*) and the plains zebra (*Equus quagga boehmi*); other more sedentary species include Thomson’s gazelle (*Gazella thomsonii*) and Grant’s gazelle (*Gazella granti*). In wetter areas, with access to permanent water, other sedentary species such as Cape buffalo, elephant, and waterbuck (*Kobus ellipsiprymnus*) are found. The Serengeti migration is the best-known, but in southern Sudan there is also a huge migratory system involving tiang (*Damauls cus lunatus*), Mongalla gazelle (*Gazella rufifrons albonata*), and white-eared kob (*Kobus kob leucotis*). These species spend the wet season in the tall grasslands towards the Uganda border. When the dry season begins, they move northwards and feed on the swamp grasses, such as *Oryza* (wild rice) and *Echinochloa*, exposed by the retreat of the Nile River floods. This migration crosses the line of the partially dug Jonglei Canal—now abandoned due to civil war—which forms a significant obstruction.

The biomass of ungulates supported by some of these grasslands is very high, particularly where there are two rainy seasons each year and therefore at least some forage of good quality at all times. The Western Rift Valley in Uganda and eastern Zaire is typical; here large herds of elephant, hippopotamuses from the lakes, and Cape buffalo form the main part of the biomass, but other species such as Uganda kob (*Adenota kob thom- awiti*), waterbuck, and warthog (*Phacochoerus africanus*) are also common.

In recent years, following the widespread destruction of the larger animals, the smaller ones have increased in numbers. The killing of many of the elephants has also led to widespread vegetation changes. *Acacia* trees, whose seedlings were formerly so regularly browsed and burned that they remained in a suppressed state in the grass, are now regenerating in many places, and thicker clumps dominated by the thorny scrambler *Capparis tomentosa* (*Capparidaceae*) are now spreading and coalescing to produce larger areas of thicket that are becoming home to giant forest hogs (*Hylochoerus mein- errzigwageni*), a species more commonly found in montane forests.

In all of these regions with numerous herbivores, there are high populations of predators. Lions are the most conspicuous, living in prides made up of one or more females and their offspring with at least one mature male in attendance. The females collaborate in hunting, and any kill is shared by all the pride. When a new male displaces another, his first action is to kill any cubs in the pride still dependent on their mothers. This action rapidly brings the lionesses into estrus and allows the new male to start passing on his genes with the minimum of delay.

Spotted hyenas (*Crocuta crocuta*) can also be very common and have been shown to function not only as scavengers but also as highly efficient predators in their own right. They live and hunt in matriarchal groups that hold and defend communal territories. The cheetah (*Acinonyx jubatus*) is a highly specialized and solitary cat, entirely dependent on its speed for running down prey. Its jaws are relatively weak, and many of its kills are lost to lions or hyenas against which it has virtually no defense. Finally, there is the wild dog (*Lycaon pictus*). This is another species that lives in groups; hunting is by a prolonged chase in which the members of the pack take turns in the lead. The leaders snap and tear at the hind end of the prey, eventually bringing it down. Wild dogs are short-lived, however, and rely on frequent large litters to maintain their numbers. They usually have a very wide home range and in many areas they are now very scarce or absent, and may well, with the cheetah, be Africa’s most endangered carnivores.

The African grasslands are also home to large ground-living birds; some, like the ostrich (*Struthio camelus*), are flightless and others, like the kori bustard (*Ardeotis kori*), almost so. The large mammal populations provide food for scavengers; for instance, a carcass on the Serengeti Plains will attract four or five species of vulture. Vultures spend much of their day soaring on thermals; they watch both the ground and each other so that if one spots prey and starts to descend, others quickly follow.

4. Shrublands

Shrubland occupies the driest areas; under conditions of even lower rainfall the individual shrubs grow farther
and farther apart until the land is best referred to as desert. The shrubs are generally between 10 cm and 2 m in height, and are of many different species and families. Members of the Acanthaceae, particularly the genera *Baularia* and *Elephantis*, are often prominent. Many are spiny, and all are highly facultative and irregular in their production of leaves and in their flowering and fruiting, often not reproducing every year but only when rainfall is exceptionally high or prolonged. Under these conditions, many annual grasses also appear between the shrubs. Perennial drought-resistant grasses often occur among the shrubs, and some authorities believe that the natural state in these ecosystems is a drought-resistant grassland, which has been converted to a shrubland by overgrazing. However, at least some shrublands grow on soils developed from limestone or gypsum and so may be partly edaphically controlled.

C. Deserts

There are two main areas of real desert in Africa: the Sahara to the north of the equator and the much smaller Namib of southwestern Africa. Parts of the Horn of Africa are also extremely dry, but only small areas such as the Danakil Depression can be considered as true deserts, if true desert is defined as an area in which plants grow only where there is extra water either from springs or from runoff. At least some rain falls over much of the Sahara each year. The exception is the eastern end, in the Nile Valley, where no rain may fall for many years in succession. Over much of the rest, enough rain falls in most years to produce a thin ephemeral vegetation in more favorable areas. Grasses are the main component of this vegetation, which dries out quickly to form a standing hay that is a valuable food resource for both wild animals and the domestic animals of nomadic herders.

Formerly several ungulate species were not common in the desert: the addax antelope (*Addax nasomaculatus*) and the scimitar-horned oryx (*Oryx dammah*) being the largest, with several smaller species of gazelle. Most of these have extremely efficient water-conserving strategies and are capable of surviving without drinking. Human population increase, with concomitant competition for space, grazing, and water, and the increased availability of firearms, mean that all of these species are now scarce and endangered. Scattered over the desert are oases where springs from underground aquifers provide year-round water. Most of these oases support permanent human settlements, often dependent on sparse annual crops as well as dates produced by the date palms (*Phoenix dactylifera*) that thrive in these environments.

The Namib Desert in southwestern Africa is also very dry, at least near the coast, but the dryness is mitigated by frequent fogs arising from the cold Benguela Current immediately offshore. Many of the plants and animals in this region survive by collecting water from mist. Furthermore, a few rivers cross the coastal strip, forming linear oases in which water is almost always available. Animals such as oryx (*Oryx gazella*), and even a few herds of elephants, manage to survive in this harsh environment.

One plant family, Mesembryanthaceae, has speciated enormously in this region, and there are hundreds of species. Many of these are very simple in structure, producing clumps of short shoots each bearing just one pair of large thick leaves at any one time. The lower parts of these leaves are buried in the soil, thus avoiding the extreme heat of the surface, and the exposed surface is often transcurrent, allowing light to reach the buried part of the leaves. The exposed parts of the leaves are also often colored and textured just like the surrounding stones, which camouflage the plants, the members of one genus, *Lithops*, are known as ‘living stones.’ One truly remarkable plant grows in this area and nowhere else in the World—Welwitschia mirabilis. This is not a flowering plant, and neither is it a true conifer. Each plant consists of a thick woody stem, mostly buried in the ground, bearing two huge strap-shaped leaves that grow continuously from the base, lie on the ground, and wear away at the tips.

D. Montane and Afroalpine Ecosystems

Mountains in Africa are of two kinds: relatively recent volcanics such as Mts. Cameroon, Kilimanjaro, Elgon, and Kenya and the Ethiopian Highlands, and old upthrust portions of basement complex such as the Ruwenzori Mountains and numerous lower ranges in Tanzania such as the Usambara, Uluguru, and Uzungwa Mountains.

Mountain slopes were originally forested, but because of their fertile soils in many places the forests have been cleared for cultivation. However, the higher forests tend to be cool and misty and much less attractive for agriculture, and some of these survive. Trees such as *Podocarpus* and the mountain bamboo *Arundinaria alpina* are common. Epiphytic bryophytes, lerns, and orchids are abundant. Above the montane forest is a belt of “giant heath” formed from large species of *Erica* that attain 10 m in height. As in the montane
forest, epiphytic mosses and liverworts are very abundant, often forming dense matts covering the trunks and branches of the giant heaths. Above the heath zone lies the afroalpine zone proper. The climate here has been described as “summer every day and winter every night.” Hot sunny days (this region is often above the cloud zone) are followed by nights during which the temperature plunges to well below freezing, and frost and ice forms on the ground. Here giant groundsel (Dendroseres) and giant lobelia grow scattered in a low shrubland of everlastingings (Helichrysum) and Alchemilla. The giant groundsel and giant lobelias consist of rosettes of huge leaves borne at the ends of sparsely branched stems. (A very similar growth form is found in the genus Espeletia of the Andean paramos.) During the night, the leaves fold upwards to form a dense mass around the delicate terminal buds, thus protecting them from the cold. Water and mucilage accumulate in the rosette and also help to prevent the buds from freezing. The old leaves of some species do not fall, but accumulate as an insulating blanket around the stems; other species have thick corky bark that functions in a similar way.

Animals are generally sparse in montane forests and in the afroalpine zone. Various antelopes move up here from the surrounding savannas, as do elephants and the occasional buffalo. With these come predators, par- from the surrounding savannas, as do elephants and in the afroalpine zone. Various antelopes move up here

Wetlands

The wetlands of Africa may be seasonal or permanent. Warping and subsidence of the earth’s crust has produced several extensive wetland areas, such as the inland delta of the Niger River in Mali, the Sudd of the Nile River in the Sudan, the inland delta of the Okavango in Botswana, and the Bangweulo swamps of northern Zambia. Many permanent swamps are dominated by the giant sedge Cyperus papyrus. Papyrus requires at least some water movement and a reasonable supply of nutrients to thrive; when it does, it attains heights of 3 m or more. Papyrus swamps are very species poor. There are few other plants, most of them climbers like the purple morning-glory, Ipomoea, and the yellow-flowered pea, Vigna luteola. A swamp antelope, the sita- tunga (Tragelaphus spretus), shelters in papyrus but finds little food there. Some birds, such as the golden- crowned gonorlek (a shrike) and Carruthers’ cisticola (a small brown warbler) spend their whole lives in papyrus swamps, but most of the herons and other waterbirds are found along channels and at the swamp margins.

Temporary pools are a common feature of the sea- sonal region of Africa. A pool will fill early in the rainy season, sometimes from rain and sometimes from overspill from rivers. Fish usually reach all but the most isolated pools; the air-breathing African lungfish can survive in a mucous cocoon in the mud for many months, emerging when water returns. Catfish (Clarias) can move over land through wet grass, and species of killifish survive as resistant eggs in the mud, hatching when wetted.

Floating plants like the Nile cabbage (Pistia) and water lilies (Nymphaea), as well as submerged plants such as bladderworts (Utricularia) and hornwort (Ceratophyllum), appear from seed in the mud. In game reserves and parks hippopotamuses (Hippopotamus amphibius) often move into the pools and churn them into mud, as well as adding nutrients in their droppings. At the end of the wet season the pools start to dry up. Then flocks of herons, egrets, and storks arrive to feed on the fish trapped in the shrinking patches of water, and the hippo move back to permanent water or die.

Seasonally flooded grasslands are found around the edges of many of the swamp regions of Africa. Some of them, like those around the Sudd of southern Sudan, and the Kafue River flats of Zambia, are very large. Among the grasses, species of Echinochloa are often common; at least some of these are very tolerant of flooding and their stems can elongate and form a floating mat when water levels are high. Perennial wild rice (Oryza longistaminata) is also common. Many of these areas are occupied by pastoral groups such as the Dinka of southern Sudan. They are cattle herders and move seasonally to follow the rise and fall of the flood. There are also wild singelates, the Nile lechwe (Kobus megac- eros) is confined to the grasslands at the edges of the Sudd swamps, and the red lechwe (Kobus leche) is found in various seasonally flooded grassland areas of Zambia. These animals also follow the rise and fall of the flood.

F. Lakes

Africa has numerous freshwater lakes, some of them very large. They are either shallow or deep lakes. First,
there are the deep lakes of the Rift Valleys: Lakes Tanganyika (33,000 km², maximum depth of 1460 m), Malawi (704 m), Edward, and Albert. These lakes are permanently stratified. In the temperate zones, low winter temperatures cool the surface waters to 4°C, at which temperature water is densest and sinks, carrying dissolved oxygen to the bottom of the lake. In the tropics there are no significant seasonal variations in temperature, the surface water is always warmer than the depths, and there is little circulation, so that the deeper layers are anaerobic and lifeless. In spite of this, the oxygenated upper layers have a rich fauna and are highly productive. These lakes are ancient, and speciation within them has been rapid. The diversity of molluscs in Lake Tanganyika, for instance, parallels that of the sea, with thick-shelled species occurring on exposed coastal rocks and species with long spines on the shells living on soft mud bottoms. The fish faunas also parallel those of the sea, with some species living mainly near the shore, while others exploit the zooplankton and other food in the open water.

Then there are the shallower lakes, such as Lakes Victoria (69,000 km², maximum depth of 90 m), Kyoga, Chad, and Bangweulu. In these lakes the stirring action of the wind is enough to circulate the water and carry oxygen to the lake bottom. Most of the shallow lakes are relatively young in geological terms, but some of the groups of fishes in them have speciated explosively to produce "species flocks," each of whose members exploits a different, often extremely specialized, niche.

In Lake Victoria, which is less than a million years old, the cichlid mouth-brooding genus Haplochromis is represented by more than 150 species, including plant-eaters, snail-eaters, and fish predators. More bizarrely, there are also species that live by biting off scales and portions of the fins of others, by pulling mayfly larvae there from their burrows in dead wood, and by (probably) sucking the eggs and young from the mouths of brooding females. Sadly this astounding diversity is under extreme threat following the introduction of the Nile perch (Lates niloticus) to the lake. This is a voracious predator, formerly found only in the Nile system below the barrier of the Murchison Falls, but now introduced to Lake Victoria with the intention of increasing fishing yields. Water hyacinth (Eichhornia crassipes) has also been introduced to Lake Victoria and is altering the ecosystem enormously by forming extensive mats at the edge of the lake, which make access for fishing difficult. (These mats may also, however, be providing nursery areas for young fish.) Lake Malawi has a similarly remarkable diversity of cichlid fishes and, so far, neither Nile perch nor water hyacinth has been introduced.

Finally, within the Rift Valley there are a number of lakes that are highly alkaline or saline. Soda (sodium carbonate) is obtained commercially from Lake Magadi in Kenya, and salt (sodium chloride) from Lake Katwe in western Uganda. The soda lakes of the western Rift Valley in Ethiopia, Kenya, and Tanzania, as is often the case with extreme environments, support relatively simple ecosystems in which a few species are present in great abundance. A planktonic blue-green alga, Spirulina, often forms almost a soup in the alkaline water. Planktonic copepods (small crustaceans) and chironomid (midge) larvae feed on the algae. Lesser flamingos (Phoeniconaias minor) also feed on the algae by filtering water through their beaks. Greater flamingos (Phoenicopterus ruber) have more widely spaced filters in their beaks, and feed mainly on the larger copepods and midge larvae, often stirring them out of the mud with their feet. They are not as numerous as the lesser flamingos. Both species migrate up and down the Rift Valley in response to changes in water levels and algal concentration. They breed in huge colonies when water levels are right and food availability is maximal.

G. Coastal Ecosystems

Most of Africa's seashores are sandy or muddy; rock is rare and generally confined to isolated headlands. Coral reefs occur along the east coast from the Red Sea to Mozambique, but not in the west, where cold upwelling water and turbidity virtually exclude them.

The strand lines of sandy shores support open communities of creeping plants such as Ipomoea pes-caprae (Convolvulaceae), Camarmondia rosea (Fabaceae) and Retama maritima (Cyperaceae). The first two have large seeds that float in seawater and are probably thus dispersed; both species occur widely in the Old World tropics. On the landward side of this grow salt-resistant bushes of species such as Sophora inhambanensis and Pemphis maritima, often forming dense thickets. Such thickets may grade into forest, or, in parts of West Africa, into a wind-swept grassland with bush clumps sculpted by the wind. Some of the grassland plants exist here as prostrate ecotypes that maintain their prostrate growth form in cultivation.

More sheltered shores often support mangrove swamps, particularly in the neighborhood of the mouths of large rivers such as the Niger and the Rovuma (Tanzania). West African mangrove swamps are relatively poor in species; Avicennia and species of Rhizophora are frequent here. The east coast is richer in
mangrove ecosystem is extremely productive; the breathing roots of the mangroves support communities of seaweeds and oysters, while the mud surface teems with crabs and small fishes. The whole zone is a rich nursery for fish.

Some of the more sheltered sandy flats behind reefs, particularly in northern Kenya and southern Somalia, have extensive beds of seagrasses (Cymodocea, Halodule, and Thalassodendron). These are not true grasses, although their flat, parallel-sided leaves are similar, but are closer to the pondweeds (Potamogetonaceae). Some grow intertidally, but many extend well below low-tide level and may form extensive lawns on sandy substrates. Here they are grazed by dugongs (Dugong dugon). These marine vegetarian mammals are regularly hunted and are becoming scarce and endangered almost throughout their range.

V. MADAGASCAR

Madagascar has been isolated from Africa for at least 140 million years, and from India for around 88 million years. It has a rich flora in which about 80% of the species are endemic, and in at least some cases their closest relatives occur in Southeast Asia rather than Africa. The isolation of Madagascar predated the major adaptive radiation of mammals that has occurred in Africa, and it lacks large grazing animals—although a dwarf hippopotamus appears to have become extinct less than 1000 years ago—and large carnivores.

At present the largest wild mammal is the bush pig (Potamochoerus larvatus), which is believed to be a recent arrival. A group of primitive primates, the lemurs, has radiated into all the habitats on the island. Fossils show that they were formerly even more diverse than they are now; some extinct forms were much larger than any modern species. The largest carnivore is the fossa, a large mongoose-like animal that climbs trees well and is a specialist predator on lemurs. There are also fossil and subfossil remains of giant flightless birds (Aepyornis), the last of which seem to have become extinct only a few hundred years ago. Humans reached the island from the east perhaps 1300 years ago. At first their settlements were confined to the coast, but later they spread inland and colonized the central plateau. Humans have had a dramatic effect on the island’s natural vegetation and habitats.

Flowering plants were in their earliest stages of evolution when the island became isolated, and a high percentage of species (80% in the legumes, Fabaceae), many genera, and some families are endemic. The vegetation of the island is very diverse. On the eastern side, rainfall is high and there is little or no dry season. Here there are tropical rain forests, very diverse in composition, with no single species being dominant.

Differ from the forests of mainland Africa in their lower stature (23–30 m), lack of large emergent trees, the abundance of small palms in the understory, and the frequent occurrence of climbing bamboos. These forests have been considerably reduced by clearance for agriculture, and only scattered fragments remain. Secondary forests are widespread, often characterized by the distinctive traveler’s tree (Ravenala madagascariensis), with a single stem crowned by huge leaves arranged like a fan in two opposite rows. Higher up the forest takes on a more montane aspect; the trees are shorter and more branched, and epiphytic ferns and mosses are abundant. The highest mountains support a montane thicket of small-leaved ericoid shrubs such as Erica (Ericaceae), Stoebe (Asteraceae), and everlasting (Helichrysum, Asteraceae).

A drier form of forest or woodland seems also to have occupied much of the central plateau, but only tiny fragments remain, and these are under intense pressure from fire, agriculture, and wood cutting for charcoal. The commonest tree is tapia (Uapaca bojeri), which may owe its survival to its fire resistance. One legume genus with two species, Pelteura, has recently been described from forest fragments in this zone; only three specimens and no living plants are known and it seems likely that the genus was extinct before it was described.

The forests of the central plateau have largely been replaced by a species-poor grassland that provides little protection to the soil from erosion so that gullies are widespread and deep.

In the western half of the island, dry deciduous or semideciduous forest survives here and there, particularly in limestone areas, which have often weathered to produce an inhospitable landscape of sharp ridges and pinnacles ("tsingy") that is very difficult of access and unsuited to any kind of agriculture.

The southern end of the island, particularly in the west, is very dry, and here a peculiar thorn forest is found in which the endemic cactus-like family Didierriacae is common. Lemurs (sifakas) live in this thorny forest. This remarkable vegetation type is threatened by agriculture, particularly sisal cultivation, by grazing, and by cutting for charcoal production. Perhaps because of the absence of large grazing animals, members of several plant families have developed a growth form in which leaves are absent and photosynthesis is carried out in the flattened stems. Several members of the family Fabaceae show this feature.
Another plant growth form perhaps more widely developed in the dry parts of Madagascar than in any other region is the “bottle-tree,” in which a thick and swollen trunk supports a rather small crown. The genus *Adansonia*, with one species in Africa (the baobab) and one or two in Australia, has seven species in Madagascar. The flame-tree (*Dorstenia regia*), now an extremely widespread ornamental tree in the tropics, is one of ten *Dorstenia* species in Madagascar, with just one other in tropical Africa. Several of the Madagascar species are bottle-trees.

Some of the richest habitats in Madagascar are the rocky outcrops, perhaps because they are sheltered from fires and grazing animals. Numerous endemic species of *Aloe* and succulent spurges (*Euphorbia*), as well as strange single-stemmed spiny succulents (*Pachypodium*), are common on these rocky outcrops and make them striking refuges for the remarkable flora of this isolated island.

**VI. THE FUTURE FOR AFRICAN ECOSYSTEMS**

What of the future? The human species has been present in Africa for longer than in any other continents. For much of the time, however, humans would have lived as part of the normal ecosystem—essentially part of the wild fauna. Exactly when they began to rise to dominance and to influence the composition and distribution of ecosystems is hard to say. There is some evidence that humans have been using fire (although probably not making it) in Africa for as long as half a million years. Fire is one of the most potent forces for change in tropical vegetation and an increase in fire frequency caused by humans may well have shifted the balance between woody and herbaceous vegetation towards the latter. The southward spread of pastoral peoples, with their knowledge of the power of fire to produce new grass and control trees, seems to have begun by at least 2000 years B.C., perhaps in response to increasing Saharan aridity, and may have been an important force for vegetation change.

Domestication of crops may well have begun at the same time. African plants that have been domesticated include yams (* Dioscorea*), sorghum, upland rice (*Oryza*), and cowpea (*Vigna*), with oil palm (*Elaeis*) in the forest zone. Bananas (*Musa*) probably arrived from Asia in the first millennium A.D., and New World crops such as maize (*Zea*), cassava (*Manihot*), tomato (*Lycopersicum*), and peppers (*Capsicum*) did not arrive until the fifteenth or sixteenth centuries A.D.

Most of the people of Africa survive by subsistence agriculture, or by growing crops that are sold to the rapidly increasing town populations. Shifting cultivation, in which a piece of ground is cleared, cultivated for a few years, and then abandoned for a fallow period of varying length, is the traditional way of exploiting the nutrient-poor soils characteristic of much of Africa. Although this practice is often attacked as wasteful and destructive, it is a very satisfactory mode of land use, so long as population densities remain low and a long fallow period is possible. However, once populations increase, the length of the fallow period falls, as do yields. This increases pressure to find new agricultural land, such as may be opened up in forest by logging activities. The extraction roads that allow people into the forest and the clearings made during timber cutting provide sites for settlement. It is often said that most of the forest loss in Africa is not caused by timber extraction, but rather by the subsequent settlement.

Human populations are increasing throughout Africa, in some nations at alarming rates. It is easy to overlook that an annual increase of 3.5% implies a population doubling every 15 years, and such rates are found in many African countries. Probably a majority of Africans are aware of the problem, but the absence of state care for the old is a considerable incentive to produce numerous children; even if child mortality is high, at least one or two will survive to provide care for the parents in their old age. Improvements in health care tend to come before reductions in birth rates, leading to lower death rates and longer life expectancies. A rise in population increases the pressure on land, and therefore on natural ecosystems. Cultivators spread farther into areas moist enough for agriculture, and irrigation schemes push out the cultivable boundaries. Pastoralists increase their flocks and herds, which are viewed as cash on the hoof, whose numbers tend to increase until drought or disease cuts them back. Some of the more natural ecosystems, such as those of seasonally dry regions with high wild animal populations, are enclosed in national parks or game reserves, but unless these are very large they are invariably inadequate to accommodate migratory species or those, like the elephant, that normally range over a very wide area. Within the parks and reserves, tourist pressure can also be a problem. Political instability and the wider availability of firearms are other threats. All of the larger antelopes of the Sahara and its fringes are now endangered because of hunting with guns and motor vehicles, as well as competition with domestic stock for forage and water. The migratory animal populations of southern Sudan...
have been greatly reduced by hunting with the many firearms now in the region. In the 1970s and 1980s, elephants were hunted throughout Africa for their ivory.

Undisturbed forest is now a rare commodity in the whole of West and East Africa, only in the Congo River Basin are there tracts still in their original state, protected by their size and inaccessibility. Exploitation of forests for timber will no doubt continue and probably intensify, not only for timber for export and internal use but also for charcoal to fuel the cookers of the cities. Additionally, more areas will be cleared for plantation agriculture and for cash crops such as oil palm and cocoa. The woodlands will continue to be heavily exploited for firewood and charcoal, and more favorable sites will be cleared and converted to large-scale agriculture, if soil fertility can be maintained. Fire frequency in the grasslands will continue to increase as the human population increases, although the spread of agriculture tends to reduce fire by fragmenting the area that can be burned.

Yet here and there are signs that change may be taking place. In West Africa there is a tradition of preserving a patch of forest near every village to provide a burial ground, a home for the ancestral spirits, and a source of plants for medicine and poles for building. These sacred groves provide refugia for some forest species. In parts of the Guinea Savanna zone, schemes for sustainable exploitation of the woodlands for firewood and charcoal have been developed and may be taking hold. Overall, however, the future for many plant and animals in Africa, particularly those of the forests, is still bleak.

See Also the Following Articles

- DESERT ECOSYSTEMS
- FIRES, ECOLOGICAL EFFECTS OF
- MEDITERRANEAN-CLIMATE ECOSYSTEMS
- NEAR EAST ECOSYSTEMS
- TROPICAL ECOSYSTEMS

Bibliography

I. Introduction

II. Exotic Species Introductions in the United States

III. Exotic Species Introductions in the United Kingdom

IV. Exotic Species Introductions in Australia

V. Exotic Species Introductions in South Africa

VI. Exotic Species Introductions in India

VII. Exotic Species Introductions in Brazil

VIII. Conclusions

GLOSSARY

**competition** Use or defense of a resource, such as a desirable growing site for a plant and a food source or shelter for an animal, by an individual that reduces the availability of the resource for another individual.

**competitive exclusion** Extirpation or extinction of one species by another in a given area through competition for resources.

**exotic (nonindigenous) species** A species that occurs in an area outside of its historically known range and that has been introduced into a new habitat or ecosystem, either intentionally or accidentally.

**native (indigenous) species** A species that occurs naturally in a given area.

THE INTENTIONAL—AND ACCIDENTAL—INTRODUCTION OF VARIOUS BIOLOGICAL SPECIES INTO AGRICULTURE started with the origin of farming about 10,000 years ago. Since then, more than 400,000 species have been moved from one region of Earth to another. Introduced species, like corn, wheat, rice, and other food crops, as well as domestic cattle, poultry, and other livestock, now provide more than 98% of the world’s food supply, at a value of more than $5 trillion per year. Other introduced species have been used for landscape restoration, biological pest control, sport, pets, and food processing.

I. INTRODUCTION

Thousands of species have been intentionally introduced to new regions for explicit agricultural purposes, such as the growing of crops and pest control. Yet many species are also accidentally introduced, and these biological “invaders” end up doing harm to the areas they invade. They can cause major economic losses in agriculture, forestry, and other segments of the world economy if crops or other products are destroyed. One recent study reported that the more than 50,000 species that have been introduced into the United States since Columbus’ time have resulted in approximately $137 billion in damages (Pimentel et al., 2000a). In addition, introduced species often have substantial negative im-
pacts on global ecological integrity; for example, the presence of an alien invader can result in the extinction of many native species.

It is difficult and complicated to estimate the full extent of the ecological damage to world agriculture caused by exotic species and the number of species extinctions they have caused, because only approximately 20% of the estimated 10 to 15 million species in the world have been described and catalogued. Nonetheless, about 400 of the 958 species on the Threatened or Endangered Species List in the United States are considered at risk primarily because of competition and predation by nonindigenous species. Other species are continually threatened by hybridization with alien species and/or ecosystem changes caused by biological invaders. Estimation of the economic impacts of nonindigenous pests affecting world agriculture is also difficult; however, sufficient data are available to quantify some aspects of the economic impacts on agriculture. This article assesses the magnitude of environmental impacts and some of the economic costs associated with the diverse nonindigenous species that have invaded world agriculture.

II. EXOTIC SPECIES INTRODUCTIONS IN THE UNITED STATES

Most plant and vertebrate animal introductions in U.S. agriculture have been intentional, whereas most invertebrate animal and microbe introductions have been accidental. In the recent past, the rate of and risk associated with exotic species introductions have increased enormously because the human population is growing so rapidly; this ever-increasing number of humans has substantially altered, and continues to alter, the global environment (Pimentel et al., 2000b). In addition, large numbers of people are traveling faster and farther and more goods and materials are being traded among nations all over the world (Bryan, 1996), so the opportunity for even more species introductions is on the rise.

A. Nonindigenous Plants

1. Introduced Crop Plants

Most alien plants now established in the United States were introduced for food, fiber, and/or ornamental purposes. An estimated 5000 plant species have escaped these realms and now exist in U.S. natural ecosystems. This number is substantial when compared with a total of about 18,000 species of native U.S. plants.

Ornamental plants are the largest group of nonindigenous plants that have become established in nature. These plants include those found in cultivated gardens, such as purple loosestrife (Lythrum salicaria). Of the approximately 25,000 cultivated plant species, mostly ornamentals brought into Florida, more than 900 have become established in surrounding natural ecosystems (Simberloff et al., 1997). More than 3000 plant species have been introduced into California. New York state also has a large percentage of introduced plant species: 1082 exotics versus 1940 native species.

In crop systems, including forage crops, an estimated 500 introduced plant species have become weed pests. Most of these were accidentally introduced with crop seeds, from ship ballast, or from various imported dead and live plant materials. Some of these accidentally introduced weeds, like yellow rocket (Barbarea vulgaris) and Canada thistle (Cirsium arvense), have become major weed pests in U.S. gardens and commercial agriculture.

Most of the 5000 nonindigenous plants established in U.S. natural ecosystems have displaced some native plant species. Alien grasses introduced as forage species have diminished the health of natural ecosystem functions as a result of nutrient losses, changes in microclimates, altered vegetation succession, and increased incidence of fires. Nonindigenous weeds are spreading and invading approximately 700,000 ha of U.S. wildlife habitat each year (Babbitt, 1998).

One of these weed pests is the European purple loosestrife, which was introduced in the early nineteenth century as an ornamental garden plant. It has been spreading at a rate of 115,000 ha per year and is changing the basic structure of most of the wetlands that it has invaded. The dominant, monotypic stands of purple loosestrife have reduced the biomass of 44 native plants and endangered the wildlife that depends on these native plants.

2. Economic Costs of Introduced Plants

In U.S. agriculture, weeds cause a reduction of 12% in crop yields; in economic terms, that represents an approximate $33 billion loss in crop production annually, based on the potential crop value of all U.S. crops of more than $267 billion per year. Using the estimate that about 73% of the weeds are nonindigenous, it follows that around $24 billion of these crop losses are due to introduced weeds. In addition, approximately $4 billion in herbicides are applied to U.S. crops, of which about $3 billion can be attributed to nonindigenous weed control.

U.S. pastures provide about $10 billion in forage
crops annually, with total losses due to weeds estimated at nearly $2 billion. Since about 45% of these weeds species are nonindigenous, I estimate that forage losses due to these weeds are close to $1 billion each year.

Some introduced weeds are actually toxic to cattle and wild ungulates. For example, leafy spurge (Euphorbia esula) contains a toxic substance that, when eaten by cattle, acts as an irritant, emetic, and purgative and can cause dysentery and sometimes death. Some animals, after unpleasant experiences, learn to avoid eating the plant. If the vegetation on grazing land is composed of just 10 to 20% leafy spurge, the cattle will not graze there and the value of the land as forage is lost. Total losses resulting from the prevalence of this weed on western rangelands account for nearly $110 million per year due to toxic effects and reduced grazing by the cattle (this cost is included in the $1 billion in losses mentioned earlier) (OTA, 1993).

Several nonindigenous thistles also replace desirable native plant species in pastures, rangelands, and forests and similarly reduce cattle grazing. According to United States Secretary of the Interior Bruce Babbitt (1998), ranchers spend about $5 billion each year to control invasive nonindigenous weeds in pastures and rangelands; despite those expensive efforts, though, these weeds continue to spread.

Lawn, garden, and golf course management costs about $36 billion per year. A significant proportion of these management costs is related to pest control. In addition to the high cost of this commercial pest control, it is estimated that about $1.3 billion (included in the $36 billion) is spent on outdoor residential weed, insect, and disease pest control each year. Given the high percentages of nonindigenous species, I estimate that about $300 million is spent on residential nonindigenous weed control, and an additional $1 billion is invested in nonindigenous weed control on golf courses. Therefore, the total economic cost of introduced weeds on U.S. cropland, pastureland, and other uses is about $35 billion annually (Table I).

### B. Nonindigenous Vertebrates

1. **Mammalian Introductions**

   Introduced mammal species include dogs, cats, horses, burros, cattle, sheep, pigs, goats, monkeys, and deer (Layne, 1997). Several of these animals escaped or were released into the wild, and many have become pests by either preying on native animals, grazing on vegetation, or intensifying soil erosion.

   For example, goats (Capra hircus) introduced on San Clemente Island, California, are responsible for the extinction of eight endemic plant species and the endangerment of eight other native plant species on the island (Kurdila, 1995). Horses (Equus caballus) and burros (Equus asinus) released in the western United States have reached populations of up to 50,000 animals. These animals graze heavily on native vegetation, allowing nonindigenous annuals to displace native perennials. Furthermore, burros inhabiting the northwestern United States compete for the primary food sources of native bighorn sheep and seed-eating birds, thereby reducing the abundance of these native animals (Kurdila, 1995).

   Overall, large populations of feral horses and burros are costly in terms of lost forage for livestock and wild-
life. Based on a minimum of $10 per animal-unit-month (AUM) per hectare, and assuming 10 ha of pasture per animal per year, these large animals exact damage of at least $5 million per year. This cost does not include the extensive environmental damage to ecosystems, which can be up to 10 times this amount when increased soil erosion and the reduction of native plants and associated native animals are considered.

Pigs (Sus scrofa), which are native to Eurasia and North Africa, have been introduced into some U.S. parks (like the California coastal prairie) for hunting, and have substantially changed the vegetation in those places. Feral pigs are also a serious problem for U.S. agriculture. In Hawaii, more than 80% of the soil is bare in regions inhabited by pigs (Kurdila, 1995). This disturbance allows annual plants to invade the overturned soil and intensifies soil erosion. Pig control per park in Hawaii, with about 1500 pigs per park, costs about $150,000 each year (based on $100 spent per pig for control). Assuming that the three parks in Hawaii have similar pig control problems, the total is $450,000 per year just in Hawaii (R. Zuniga, personal communication).

Pigs also have become a serious problem in Florida, where the feral pig population has risen to more than 500,000. In a few cases these invaders are beneficial—for example, pigs serve as a food source for the endangered Florida panther (Layne, 1997). For the most part, though, in Florida and elsewhere, pigs damage hay, grain, peanut, soybean, cotton, and various vegetable crops. Pigs also damage fences, livestock watering devices, ponds, and young livestock. In addition, they transmit and are reservoirs for serious diseases that affect both humans and livestock, like brucellosis, pseudobrucellosis, and trichinosis.

Nationally, there are an estimated 4 million feral pigs (Pimentel et al., 2000a). Control costs for feral pig range from $36 per pig for poison control to $140 per pig for hunting with dogs. Estimated damages are at least $300 per pig per year. Assuming that 4 million feral pigs inhabit the United States, the yearly cost is about $800 million per year. This is a conservative estimate because pigs cause significant environmental damage that is not easily translated into dollar values.

In addition to large animals like horses and pigs, many small mammals have been introduced both accidentally and intentionally into the United States. Some examples of these species are the European (black or tree) rat (Rattus rattus), Asiatic (Norway or brown) rat (Rattus norvegicus), house mouse (Mus musculus), nutria (Myocaster coypus), European rabbit (Oryctola-
gus cuniculus), cat (Felis cattus), and dog (Canis familiaris) (Layne, 1997).

Some rodents have become serious pests on farms, in industries, and in homes (Layne, 1997). On farms, rats and mice are particularly abundant and destructive. On poultry farms there is approximately 1 rat for every 5 chickens (D. Pimentel, unpublished data). Assuming this ratio, the total rat population on U.S. poultry farms may easily number more than 1.4 billion. Even with a decline in the rat population since these observations were made, I estimate that the number of rats on poultry and other farms is still approximately 1 billion. Given an estimated 1 rat per person, there are an estimated 250 million rats in U.S. urban and suburban areas. All told, there is a total of around 1.25 billion rats in the United States.

Various studies report that an individual rat consumes and/or destroys from 210 to 740 kg of grain each year. If it is conservatively assumed that each rat consumes and/or destroys grain and other materials valued at $15/yr, then the total cost of destruction by rats in the United States is more than $19 billion per year. In addition, rats cause fires by gnawing electric wires, pollute foodstuffs, and act as major vectors in the transmission of several diseases, including salmonellosis and leptospirosis, and to a lesser degree plague and murine typhus.

The Indian mongoose (Herpestes auropunctatus) was first introduced into Jamaica in 1872 for the biological control of rats in sugarcane plantations. It was soon introduced to Puerto Rico, other West Indian islands, and Hawaii. The mongoose was effective in reducing the Asiatic rat population in some areas, but with reduced competition the European rat returned to being a major pest in sugarcane fields. Moreover, the mongoose preyed heavily on ground-nesting birds and subsequently reduced their numbers. It also preyed on amphibians and reptiles that were beneficial to biological pest control, resulting in a minimum of 7 to 12 extinctions in Puerto Rico and other islands of the West Indies. In addition, the mongoose emerged as the major vector and reservoir of rabies and leptospirosis in Puerto Rico and other islands. Based on public health effects, poultry losses, extinctions of amphibians and reptiles, and the elimination of many native birds, we estimate that the mongoose causes approximately $50 million in damages each year in Puerto Rico and the Hawaiian Islands (D. Pimentel, unpublished data; R. Zuniga, personal communication).

Finally, most dogs introduced into the United States have been domesticated, although some have escaped into the wild. Some of these feral dogs run in packs
and kill deer, rabbits, and domestic cattle, sheep, and goats. It is estimated that feral dog packs in Texas cause more than $5 million in livestock losses each year. Dog packs have also become a serious problem in Florida (Layne, 1997). Assuming $5 million in damages for the other 47 continental states, total losses in livestock kills per year would be approximately $10 million per year.

2. Exotic Birds

Of the 97 bird species introduced into the United States, only 5% are considered beneficial; a majority (56%) are pests (Temple, 1992). Several nonindigenous bird species, including chickens and pigeons, were introduced into the United States for agricultural purposes.

The common myna (Acridotheres cristatellus), introduced into Hawaii for biological control, helped control pest cutworms and armyworms in sugarcane. However, it became the major disperser of seeds of the introduced pest weed Lantana camara. To cope with the weed problem, Hawaii was forced to introduce insect biocontrol agents for the weed (Kurtila, 1995).

The English or house sparrow (Passer domesticus) was introduced into the United States intentionally in 1853 to control the canker worm. By 1900, the birds themselves were considered pests because they damage plants around homes and public buildings and consume wheat, green corn, and the buds of fruit trees. About 70% of the house sparrow's diet consists of grains (the introduced alfalfa weevil accounts for the other 30%). Furthermore, English sparrows harass robins, Baltimore orioles, yellow-billed cuckoos, and black-billed cuckoos. They also displace bluebirds, wrens, purple martins, and cliff swallows from their nest sites, and are associated with the spread of about 29 diseases that affect both humans and domestic livestock.

Similarly, European starlings (Sturnus vulgaris) are serious pests and are estimated to occur at densities of more than 1 per hectare in agricultural regions. They are capable of destroying approximately $2000 worth of cherries per hectare. In grain fields, starlings consume about $6 worth of grain per hectare. Therefore, assuming approximately $5/ha for all damages to agriculture crop production in the United States, the total economic loss due to starlings would total approximately $800 million per year. In addition to these economic damages, starlings also have some negative environmental impacts. For example, starlings are aggressive and have displaced numerous native bird species, which can disrupt precarious balanced ecosystems. Starlings also have been implicated in the transmission of 25 diseases, including parrot fever and other diseases in humans and livestock.

The domestic pigeon (Columbia livia) has been introduced for agricultural production and it has invaded most cities of the world, including U.S. cities. Pigeons are considered a nuisance because they foul buildings, statues, cars, and sometimes pedestrians with their droppings, and because they feed on agricultural grain. Each pigeon consumes an average of 15 kg of grain per year (Smith, 1992). The economic impacts caused by fouling are estimated to be at least $9 per pigeon per year (based on $9 in control costs per pigeon). Assuming there is 1 pigeon per hectare in urban areas or approximately 0.5 pigeon per person in urban areas, common pigeons cause at least $1.1 billion in damages each year. This estimate does not take into consideration pigeons' role as reservoirs for over 50 serious human and livestock diseases, including parrot fever, ornithosis, histoplasmosis, and encephalitis.

Thus, if we assume $800 million per year in economic losses from starlings, $1.1 billion per year from pigeons, and $200 million for house sparrows and other birds, the damages from nonindigenous pest birds are estimated to be $2.1 billion per year.

3. Nonindigenous Amphibians and Reptiles

The cane toad (Bufo marinus) and bull frog (Rana catesbeiana) were introduced into Florida, Puerto Rico, Hawaii, and other warm regions, in some cases for the biological control of pest insects. However, the cane toad, with its toxic skin glands, has proved lethal to birds, dogs, cats, and other mammals.

C. Nonindigenous Invertebrates and Microbes

Approximately 4600 arthropod species (2600 species in Hawaii and more than 2000 in the continental United States) have been introduced into the United States. Eleven earthworm species and nearly 100 aquatic invertebrate species also have been introduced (OTA, 1993). More than 95% of these introductions were accidental—many invertebrate species gained entrance via plant introductions, in soil, or in water ballast from ships.

1. Insects and Mites

Approximately 1000 nonindigenous insect and mite species are pests in crops, stored-food products, and structures. Hawaii has about 5200 identified native insect species, and an additional 2600 introduced insect species (Howarth, 1990). Introduced insects account for 99% of the pest insects in that state. In addition to
Florida's 11,500 indigenous insect species, 949 immigrant species have invaded the state (42 species were introduced for biological control). In California, the 600 introduced species are responsible for 67% of all crop losses. Some of the California pests include the cottony cushion scale (Icerya purchasi) and alfalfa weevil (Hypera postica).

Each year pest insects destroy about 13% of potential crop production at a cost of about $33 billion in U.S. crops. Considering that about 40% of the pests were introduced, we estimate that these alien pests cause about $13 billion in crop losses each year. In addition, about $1.2 billion in pesticides are applied for all insect control each year. The portion applied against introduced pest insects costs approximately $0.5 billion. Therefore, the total cost for introduced nonindigenous insect pests is approximately $13.5 billion annually.

Lawn, garden, and golf course management activities cost about $36 billion annually. More than $3 billion per year is related to the expenses of pest control. Assuming the presence of 40% nonindigenous insect pests, I estimate the control costs in lawns, gardens, and golf courses to be about $1.5 billion each year.

Other introduced insect species have become pests of livestock and wildlife. For example, the imported red fire ant (Solenopsis invicta) kills poultry chicks, lizards, snakes, and ground-nesting birds. In some areas fire ants are extremely abundant, with as many as 367 nests per ha. The estimated damage to livestock, wildlife, and public health caused by fire ants in Texas is estimated to be $300 million. An additional $200 million per year is invested in controlling these ants. Assuming equal damages in other such infested southern states, the fire ant damages total approximately $1 billion per year.

2. Nonindigenous Earthworms

There are approximately 70 native U.S. earthworm species. In disturbed habitats, 11 species of introduced earthworms have reduced the numbers of some of the native species. Although detailed data are not available concerning impacts of nonindigenous earthworms on U.S. earthworm species and soil quality, earthworms are generally beneficial to soil productivity and formation in agriculture.

3. Microbe Introductions in Agriculture

Although some microbes were intentionally introduced into the United States for wine and cheese making, most were accidentally introduced and some have become serious pests. The number of microbe species introduced into the United States cannot be estimated because several thousand species can exist in a single gram of soil. Only a small fraction of all of these species have been identified.

More than 100 species of microbes have been intentionally introduced for processing wine, beer, cheese, and other foods. In addition, about 50 microbes have been introduced for the biological control of pest insects. A strain of Bacillus thuringiensis (BT) was introduced and has been used extensively to control pest caterpillars. Other strains of BT have been developed to control beetles and mosquitoes.

An estimated 121 species of microbes, mostly introduced inadvertently in seeds and other parts of host plants, have become major crop pests in the United States. U.S. crop losses to all plant pathogens total approximately 12% of crops planted, a loss of $33 billion per year. Approximately 65% of these crop losses—an estimated $21 billion per year—are attributable to nonindigenous plant pathogens. In addition, $0.72 billion is spent annually on fungicides to control plant pathogens; approximately $0.5 billion of this goes toward the control of nonindigenous plant pathogens.

As mentioned earlier, lawn, garden, and golf course management activities have an annual cost of about $36 billion; most of this is spent for pest control. Assuming that about $3 billion of management costs are related to plant pathogen control and 65% of the pests are nonindigenous pathogens, we estimate the costs caused by introduced plant pathogens in lawns, gardens, and golf courses to be $2 billion each year. Therefore, the damages caused by nonindigenous pathogens and the attempts to control them total about $23.5 billion annually.

When all of the foregoing economic costs in crop and pastureland losses and for pest control for nonindigenous weeds, mammals, birds, insects, and plant pathogens are combined, the annual total is approximately $96 billion to the U.S. economy.

III. EXOTIC SPECIES INTRODUCTIONS IN THE UNITED KINGDOM

A. Nonindigenous Plants

There are a total of 26,000 introduced plant species in the United Kingdom. An estimated 25,000 species are established in U.K. botanical gardens alone, 14,000 species are cultivated as commercial horticultural crops, and there are 6000 species of noncultivated aliens. There are 1169 species of exotic plants known to be established in natural ecosystems. The native flora of the United
Kingdom numbers only 1515 species (Crawley et al., 1996).

Most of the alien plants occur in relatively few types of habitats. More than 80% of the alien species are present in waste ground areas, urban sites, roadway sides, and agricultural habitats. An estimated 63% of the alien plants occur in hedges and scrub areas; croplands and gardens harbor about 43% of the alien species (Crawley et al., 1996). Finally, about 40% of alien species occur in rock walls and woodlands. It is interesting that plant communities like grazed, mesic grasslands and native Pinus sylvestris woodlands contain no alien plant species.

In U.K. agriculture, weeds cause an average reduction of about 10% in crop yields; however, this loss can be as high as 32% in some crops. In economic terms, about $2.8 billion in potential crop production is annually lost due to weed infestations. Given the estimate that about 43% of the weeds are alien, it follows that $1.2 million of the crop losses are due to introduced weeds (see Table 1).

B. Nonindigenous Vertebrates

1. Introduced Mammals

The total number of mammalian species in the United Kingdom is 54; 17 of these are alien species. The mammal introductions include domestic animals like dogs, cats, cattle, horses, sheep, and pigs, as well as other nondomesticated mammals. Some of the species that were intentionally or accidentally introduced into the United Kingdom include the gray squirrel (Sciurus carolinensis), European rabbit, North American mink (Mustela vison), brown rat, and black rat (Lever, 1994). All of these animals, except for cattle and horses, have escaped into the wild and are now well established.

In the United Kingdom, rodents are serious pests on farms, in industries, and in homes. The estimated number of rats are estimated at 5 per person (based on the rat-to-person ratio in the United States), including rats on farms (Pimentel et al., 2000a). Thus the total number of rats is 295 million. Using the same value of $15 for food and other goods damaged per rat, the cost of rats in the United Kingdom is about $4.5 billion annually.

The European rabbit is also abundant in the United Kingdom, with densities in some areas of up to 30 rabbits per hectare. Assuming approximately 10 rabbits/ha on the 7 million ha of cropland in the United Kingdom with an estimated $11 damage per rabbit, the total annual economic damages from European rabbits are $800 million/yr. They are reported to reduce wheat production by about 5% to 8%, and to reduce forage production for livestock by about 20%. With 11 million ha of pasture for livestock, we estimate these losses to be $400 million/yr. Thus, the total damage from the European rabbit in the United Kingdom is $1.2 billion/yr.

2. Introduced Birds

Out of the 542 bird species in the United Kingdom, 47 are alien species. Some of the introduced species include the Canada goose (Branta canadensis) and the little owl (Athene noctua). One of the 47 introduced bird species is currently causing major ecological and/or economic problems—the common pigeon. This bird does invade agriculture, but is most common in cities and towns (Lever, 1994). Pigeons are a particularly serious problem for reasons discussed earlier, namely, the pollution of city surfaces with their droppings and their consumption of grain in agriculture.

Assuming there is 1 pigeon per hectare, or 0.5 pigeon per person in urban areas (as in the United States), then there are approximately 30 million pigeons in the United Kingdom. The estimated damage that a pigeon causes is a minimum of $9 per year to crops and structures in cities and towns (Pimentel et al., 2000a). Therefore, pigeon damages are estimated to be at least $270 million/yr. This does not include the role of pigeons as reservoirs for over 50 human and livestock diseases, including parrot fever, ornithosis, histoplasmosis, and encephalitis. Pigeons are also responsible for transmission of at least three diseases to U.K. poultry, including Newcastle disease.

C. Nonindigenous Invertebrates and Microbes

1. Introduced Arthropods

There are approximately 23,000 native species and 1700 introduced species of arthropods in the United Kingdom. An estimated 1500 species are of economic importance; about 169 alien species are considered pests. An estimated 30% of the crop losses in the United Kingdom are associated with the introduced arthropod pests, as compared with about 40% in the United States (Pimentel et al., 2000a).

Arthropods damage and/or destroy approximately $2.8 billion in crops in the United Kingdom each year, based on the average of 10% crop losses per year. With about 30% of these losses due to introduced arthropods, they cause an economic loss of about $840 million/yr.

2. Introduced Crop Plant Pathogens

An estimated 74% of the plant pathogens in the United
Kingdom were introduced when seeds and other crop parts were brought into the country for agriculture. Approximately $8 billion is lost to all pests in crop production; about 8% of total potential production is lost to plant pathogens at a cost of about $2.3 billion/yr. If 74% of these crop losses are due to introduced plant pathogens, then about $1.7 billion/yr is associated with introduced microbes in crops.

IV. EXOTIC SPECIES INTRODUCTIONS IN AUSTRALIA

A. Nonindigenous Plants

There are approximately 20,000 vascular plant species in Australia (D. Pimentel, unpublished data), including an estimated 1952 alien species. The rate of alien species introductions into Victoria, Australia, alone has been 5 to 6 species per year during the past century. Many of these species have become weeds and have invaded a wide range of environments. The invasive plants are a serious problem in both agricultural and wild ecosystems, where they disrupt key natural ecosystems, alter fire regimes, and reduce the resources for native animals. An estimated 60% of the weed species in crops in Australia are alien (based on a survey of major weeds in cereal crops).

The introduced blackberry (Rubus procerus) alone causes $77 million worth of damages to crop production each year. The indirect and direct losses due to all weeds in pastures are estimated to be $970 million/yr. Weeds are estimated to cause about $4 billion in total damages in crops and pastures. Therefore, invasive plants cause approximately $2.4 billion/yr in losses to agriculture (see Table 1).

B. Nonindigenous Vertebrates

1. Introduced Mammals

There are presently 20 introduced mammals in Australia, including the European rabbit, hares (Lepus europaeus), water buffaloes (Bubalis bubalis arnee), cats, dogs, foxes (Vulpes vulpes), sheep, goats, cattle, horses, camels (Camelus dromedarius), pigs, and donkeys. In comparison, the number of native mammals is 227. The populations of some of these introduced mammals in the Northern Territory are quite high: buffaloes, 340,000; horses, up to 300,000; donkeys, up to 140,000; and camels, up to 30,000. The populations of these herbivores are much too high, but the costs of implementing a control program are also very high. For example, the cost of controlling buffalo is nearly $30 per animal.

The herbivorous animals significantly reduce the vegetative cover by overgrazing and this intensifies soil erosion and encourages annual plants and inedible shrubs in pastures (Lever, 1994). Surprisingly, the most serious pest herbivorous mammal in Australia is not one of these large-hoofed herbivores, but rather the European rabbit. In Tasmania, the rabbit population in 1952 reached a density of up to 250 rabbits/ha. The total number of rabbits in Australia ranges from 200 to 300 million. Approximately 15 rabbits consume the equivalent pasture forage of one sheep (Emmerson and McCulloch, 1994). The impact of rabbits on sheep production per year is estimated to be $110 million, including reduced sheep production and rabbit control costs. If we assume a conservative estimate of only 0.5 rabbit/ha for cropland and pastureland and also assume that each rabbit causes a minimum of $5 damage, then on the 465 million ha of Australian cropland and pastureland, rabbits are causing at least $1.2 billion/yr in damages.

Feral pigs are also a serious problem in Australia, as they damage fences and spread animal diseases, including tuberculosis, brucellosis, rabies, and foot-and-mouth disease (Lever, 1994). The number of feral pigs ranges from 4 to 20 million. These pigs damage crops, kill lambs, and damage the natural environment; they are estimated to cause at least $80 million/yr in damages (Emmerson and McCulloch, 1994).

The house mouse damages crops, houses, farm machinery, and livestock production. Estimates for these annual losses range from $50 to $100 million/yr. Far more serious than the house mouse, though, are the invading brown and black rats. In the United States, there are 4.6 rats per person (Pimentel et al., 2000a). Assuming the same number of rats per person in Australia and that each rat causes $15 of damage, then for a human population of 18 million, rats are causing approximately $1.2 billion/yr in damages in Australia.

Introduced pet cats and feral cats are also a serious problem, especially for native bird, mammal, and amphibian populations. There are an estimated 18 million feral cats in Australia and each cat is estimated to kill 8 birds per year. Assuming that 144 million birds per year are killed, and that each bird has a value of $30, then damages from cats alone reach about $4.3 billion/year.

2. Introduced Birds

Australia has 850 bird species, of which about 70 are alien species. Most of the introductions have been inten-
tional, including the English starling, English house sparrow, and common pigeon. These birds are often restricted to cities and towns. Pigeons cause similar problems in Australia as they do elsewhere.

3. Introduced Reptiles and Amphibians
There are about 700 species of reptiles and amphibians in Australia, but only 2 of them are nonindigenous. One of these introduced amphibians, the cane toad (Bufo marinus), was introduced as a biological control agent for insect pests in sugarcane fields. Unfortunately, the toad has become a pest itself because it is poisonous to dogs, cats, and other mammals that attack it.

C. Nonindigenous Invertebrates

1. Introduced Arthropods
There are an estimated 108,000 arthropod species in Australia, with 54,000 species of native and nonnative insects and mites and 10,000 mollusk species. Crop losses due to insects and mites are estimated at 10.7% of Australia’s gross potential production of $22 billion/yr. An estimated 36% of the pest arthropods in Australia are alien species. Based on arthropod-caused crop losses of about $2.4 billion/yr, the exotic pests account for losses of $860 million/yr. In addition, three exotic insects and mites cause $228 million/yr in damages to the wool industry alone. Thus, we estimate that exotic insect and mite species in Australia cause losses of at least $1 billion/yr.

2. Introduced Crop Plant Pathogens
If the total potential crop production in Australia is $22 billion/yr and about 15.2% of crop losses are due to plant pathogens, the economic costs of these pathogens total about $3.3 billion/yr. Because a large number of plant pathogens are introduced with crop seeds and other plant parts, an estimated 82% of the plant pathogens in crops are believed to be alien species (based on plant pathogens in field crops). Therefore, Australia loses $2.7 billion/yr in crops from exotic plant pathogens.

3. Microbe Introductions Affecting Livestock
There are an estimated 44 exotic animal diseases that could infect livestock in Australia. One exotic livestock disease, sheep pox, costs the wool industry an estimated $21 million/yr (D. Pimentel, unpublished data).

V. EXOTIC SPECIES INTRODUCTIONS IN SOUTH AFRICA

A. Nonindigenous Plants
There are 24,000 plant species in South Africa, including an estimated 8750 alien species. Most of the alien species were introduced from South America and Australia and have in turn invaded a wide range of environments. A total of 273 species of introduced plants in South Africa are serious weeds in crops (Bromilow, 1995).

Reduced crop production due to all weeds is 16.6% of potential crop production and totals $3.7 billion/yr in losses (total potential agricultural production is $22 billion/yr). Assuming that 67% of the weeds in crops are alien (Bromilow, 1995), then the total loss in crop production due to alien weeds is $2.5 billion/yr (see Table 1).

Two of the most serious plants invading pasturage include the shrub Lantana camara and the cactus plant Opuntia ficus-indica. In addition to these two weeds, there are approximately 800 alien weed species out of a total of 1604.

B. Nonindigenous Vertebrates

1. Introduced Mammals
There are 16 species of introduced mammals in South Africa, including the European rabbit, hare, water buffaloes, cats, dogs, sheep, goats, cattle, horses, camels, pigs, and donkeys. The total number of mammal species in South Africa, including alien species, is 247. As in Australia, the herbivore populations in South Africa are much too high for the resources that are available, and they significantly reduce the vegetative cover by overgrazing. This intensifies soil erosion and encourages some annual plants, weeds, and indelible shrubs to take over the pasturage (Pimentel et al., 2000b).

Feral pigs are also a serious problem; they damage fences and spread disease in South Africa much as they do in Australia. The estimated control cost per feral pig in the United States is about $100 (Pimentel et al., 2000a), about the same as in South Africa.

Rats are a serious problem in South African agriculture as well as in urban areas. There are an estimated 4.6 rats per person in South Africa (based on data from the United States; Pimentel et al., 2000a), which has a human population of 39 million people. Thus, there are about 179 million rats in South Africa; each rat is assumed to cause $15 in damages per year. From these estimates, losses from rat damage total $2.7 billion/yr.
2. Introduced Birds
Of the 725 bird species in South Africa, only 8 are alien species. Some of the introduced species include the English starling, common pigeon, Indian myna, and English house sparrow. Only the starling and pigeon are currently causing economic problems, and their damage is restricted to cities and towns. The problems with starlings and sparrows are very similar to the negative effects that these birds have elsewhere; namely, they pollute structures with their droppings and cause agricultural losses by consuming grains and fruits. In addition, pigeons, starlings, and sparrows are known reservoirs and vectors of up to 50 different diseases of humans and livestock.

C. Nonindigenous Invertebrates and Microbes
1. Introduced Arthropods
An estimated 80,000 species of insects, 6000 species of spiders, and numerous other arthropod species exist in South Africa. About 20% of these are believed to be exotic. One of the most serious invaders is the Argentine ant (Linepithema humile), which causes major problems because it destroys native vegetation, including endangered plants. The same ant also negatively affects native ants and other species of arthropods by competing with them for limited resources.

Insect and mite pests in agriculture cause 16.7% or $3.7 billion in losses of potential crop production each year. Because approximately 45% of the insect and mite pests are exotic, the economic losses to exotic pests are estimated to be $1.7 billion/yr.

2. Introduced Crop Pathogens
Approximately 85% of the plant pathogens that attack crops in South Africa are nonnative species (based on an assessment of diseases of fruits and vegetables). Most of these pathogens were introduced with the introduction of crops into South Africa. In forests, around 69% of the pathogens are exotic. Plant pathogens in South Africa cause an estimated 15.6% or $3.4 billion/yr loss of the potential crop production. Since 85% of the pathogens are exotic, crop losses to exotic species total $2.9 billion/yr.

3. Microbe Introductions Affecting Livestock
Several serious livestock diseases, including tuberculosis, brucellosis, East Coast fever, anthrax, and rinderpest, infect livestock and other animals in South Africa (D. Pimentel, unpublished data). Estimates suggest that these livestock diseases are causing losses of around $1 billion/yr (Table 1).

VI. EXOTIC SPECIES INTRODUCTIONS IN INDIA

A. Nonindigenous Plants
There are 45,000 plant species in India, and an estimated 18,000 of these are alien species (Saxena, 1991). Many of the alien species have become weeds and have invaded a wide range of environments. Several weed species have been introduced along with the introductions of new crops. Weeds are estimated to cause a 30% loss in potential crop production, which totals about $54 billion/yr. Assuming that 42% of the weeds in crop production are alien, the total cost associated with invading weeds is $22.8 billion/yr (see Table 1).

Lantana camara, a shrub introduced from South America as an ornamental plant, is a major weed in India, and it has invaded most pasturclands (13.2 million ha). Lantana is toxic to cattle and the cost of controlling it is $70/ha. Only about 4% of India's land area is in pasture, yet total damage per year from Lantana is $924 million.

B. Nonindigenous Vertebrates
1. Introduced Mammals
There are approximately 30 species of nonnative mammals in India, including cats, dogs, sheep, goats, pigs, axis deer (Axis axis), house mice, and rats. The total number of mammal species in India, including alien species, totals 320 (Pimentel et al., 2000b). As in other countries, the introduced herbivorous animals significantly reduce the terrestrial cover by overgrazing, resulting in intensified soil erosion and increased invasion of pastureland by inedible weeds and shrubs.

Rats number at least 2.5 billion in India, or about 2.7 rats per person (Pimentel et al., 1999). They attack crops in the field and are estimated to reduce potential crop yields by about 2%. In addition, rats are especially serious pests of stored grain supplies. Various studies report that an individual rat consumes and/or destroys about 210 kg of grain per year in India, and up to 740 kg of grain per year in Pakistan. Rats are estimated to destroy about 12 million tons of grain per year in India (D. Pimentel, unpublished data), in addition to damaging other foods, goods, and structures. In India I esti-
mated that each rat causes at least $10 of damage per year, and thus they are responsible for at least $25 billion annually (Pimentel et al., 2000b). In addition, rats are major vectors of and carriers of 38 human and livestock diseases. An average of 250,000 people die each year from the plague in India.

2. Introduced Birds
The number of bird species in India is 1221, including migrants and vagrants. Introduced species number only 4, and include the English house sparrow, common pigeon, black francolin, and Alexandrine parakeet. In neighboring Pakistan, sparrows reduce potential wheat yields by 170,000 tons at a cost of $26 million/yr (D. Pimentel, unpublished data). Based on these data, losses in India are calculated to be $50 million/yr. Also, the common pigeon is a problem in agriculture, as well as in cities and towns, where it consumes grains and fouls buildings, statues, cars, and the occasional unlucky pedestrian. Pigeons are also involved in the spread of about 50 diseases that affect humans and livestock.

C. Nonindigenous Invertebrates and Microbes

1. Introduced Mollusks
An estimated 1500 species of land mollusks exist in India; several of these are exotic. The giant African snail (Achatina fulica) destroys from 0.29 to 4.3 g of grain per day per snail from eight major crops.

2. Introduced Arthropods
There are more than 30,000 species of arthropods in India. An estimated 560 mite species and about 600 species of insects are crop pests. About 30% of the pest species are introduced arthropods, and as a group they reduce potential crop production by 18.7%. Based on total potential crop production in India of $181 billion/yr, crop losses to alien arthropods total $10.2 billion/yr.

3. Introduced Crop Plant Pathogens
Plant pathogens reduce potential crop production in India by approximately 16%, at a total cost of $29 billion/yr. There are approximately 30,000 species of plant pathogens in India, and about 74% of the major pathogens are exotic species (based on the major plant pathogens in vegetable crops). Thus, the total cost of invading plant pathogens to crops in India is about $21.4 billion/yr.

4. Microbe Introductions
Affecting Livestock
Several major diseases of livestock cause significant losses in India, most significantly foot-and-mouth disease. During 8 months in 1996, nearly 50,000 cases were reported at an estimated cost of $17,000/yr.

The combined cost of crop losses that can be attributed to nonindigenous plants, mammals, arthropods, and plant pathogens is $80 billion. Thus the negative effect of exotic species on Indian agriculture rivals that found in the United States.

VII. EXOTIC SPECIES INTRODUCTIONS IN BRAZIL

A. Nonindigenous Plants
Of the 55,000 plant species in Brazil, an estimated 21.1%, or 11,605, are alien species. Many of the alien species have become weeds and invaded a range of environments. In crop production, alien species make up 73% of the weed species. Weeds are estimated to destroy about 13.4% of Brazil’s potential crop and pasture production, or about $12.3 billion/yr (see Table I).

B. Nonindigenous Vertebrates

1. Introduced Mammals
The number of introduced mammals in Brazil is estimated to be 30, including cats, dogs, sheep, goats, cattle, horses, pigs, and donkeys. The total number of mammal species in Brazil, including alien species, is 428.

On the outskirts of São Paulo in the community of Taboão de Sera, estimates suggest that there are 12,500 feral dogs and 4600 feral cats (Pimentel et al., 2000b). The feral dogs attack livestock and other animals. Feral cats are also a serious problem, as they destroy birds and other native animals in Brazil.

It is estimated that there are about 320 million rats in Brazil. Assuming that each rat causes $10 in damages, the total cost of rats is estimated to be $3.2 billion/yr.

2. Introduced Birds
The number of native bird species in Brazil is 1635. Only three, the English house sparrow, common waxbill (Estrilda astrild), and common pigeon, are introduced species. The sparrow and pigeon populations cause major ecological and/or economic problems. The common waxbill is not considered a serious problem because it feeds primarily on the introduced guinea
grass (Panicum maximum) and does little damage to this fast-growing grass (D. Pimentel, unpublished data).

C. Nonindigenous Invertebrates and Microbes

1. Introduced Arthropods

Invertebrate species number more than 100,000 in Brazil; about 70,000 are arthropod species. About 14.4% of potential crop production is destroyed by insects and mites, and approximately 35% of these pests are exotic. The calculated loss of crops to exotic insects and mites is estimated to be $4.6 billion/yr.

2. Introduced Crop Plant Pathogens

There are an estimated 100,000 species of microbes in Brazil; around 75% of the microbes that attack crops are exotic. Like most countries, most of these plant pathogens were introduced via alien crop species. Plant pathogens are estimated to cause 13.5% in crop losses each year. If 75% of the plant pathogen species are exotic, estimated losses from alien species total $18.3 billion/yr.

VIII. CONCLUSIONS

With approximately 400,000 nonindigenous species in various nations worldwide, if even a small fraction of these invaders are harmful, significant agricultural problems can result. While nearly all of our crop and livestock species are nonindigenous and have proven essential to the viability of the world's agriculture and economy, exotic species invasions do result in many negative financial effects. This article shows that the economic damages of these alien species to the agricultural economies of just six nations (including, in some cases, control costs) amount to approximately $247 billion each year. If this figure is extrapolated to all nations based on costs per person, the total cost of exotic species to world agriculture would be about $943 billion/yr.

Precise economic data for some of the most ecologically damaging biological invaders are not available. Mammals, including horses, water buffaloes, and goats, have been responsible for reducing the productivity of pastures and rangelands, as well as the extinction of many plant species in various ecosystems. In other areas, feral pigs and feral dogs have had serious impacts on agriculture, but as of yet few data are available concerning these animals.

Most of the exotic species in the world have been introduced within the past 70 years. Accelerated international trade and travel, plus rapidly increasing human numbers, guarantee that the threat from nonindigenous species invasions is still growing. The true challenge lies not in assigning precise dollar values to exotic species losses but in preventing further biological introductions and the resultant damage to managed and natural ecosystems. Although policies and practices to prevent the accidental and intentional introduction of exotic species are improving, the world's nations are woefully short of allocating sufficient resources to the problem in proportion to the risks. Both natural and managed ecosystems need to be protected from additional costly damages resulting from introduced species.

See Also the Following Articles

ENDANGERED ECOSYSTEMS • EXTINCTION, CAUSES OF • RANGE ECOLOGY • INTRODUCED SPECIES, EFFECT AND DISTRIBUTION • PESTICIDES, USE AND EFFECTS OF • PLANT INVASIONS

Bibliography


Pimentel, D., McNair, S., Janecke, J., Rightmier, J., Simmons, C.,


I. Introduction
II. The Intensification of Crop Production
III. The Implications of Intensification for Crop Genetic Diversity
IV. The Commercialization of Crop Production
V. The Implications of Commercialization for Crop Genetic Diversity
VI. Conclusions

GLOSSARY

alleles Variants of nucleotides within the genome that cause changes in a given protein.
crop genetic diversity The diversity of the sets of genes carried by different individuals within a crop species. It occurs in the form of nucleotide variation within the genome.
genetic diversity The diversity of the sets of genes carried by different organisms. It occurs in the form of nucleotide variation within the genome. When this variation causes a change in a given protein, the variants are called alleles. Allelic variation occurs at various genetic loci or gene positions within a chromosome.
genetic erosion The loss of genes from a gene pool attributed to the elimination of populations caused by factors such as the adoption of high-yielding varieties, farmers’ increased integration into the market, land clearing, urbanization, and cultural change.
genetic variation The allelic variation that occurs at various genetic loci or gene positions within a chromosome. Genetically variable loci are termed polymorphic or are said to show polymorphism.
landraces These have a complex nature; therefore, it is not possible to give an all-embracing definition. Landraces are the originally adapted but variable crop populations on which farmers based their selections. They are geographically or ecologically distinctive populations which are conspicuously diverse in their genetic composition both between populations and within them. They have certain genetic integrity and are recognizable morphologically; farmers have names for them and different landraces are understood to differ in adaptation to soil type, time of seeding, date of maturity, nutritive value, use, and other properties.
loci Gene positions within a chromosome. Allelic variation occurs at various genetic loci.
spatial and/or temporal patterns of genetic variation The systematic changes in the alleles occurring at specific loci along spatial and/or temporal dimensions.
CROP GENETIC DIVERSITY IS THE BASIS of food production. Its pattern has changed significantly with the intensification and commercialization of agriculture. There is disagreement about the implications of these changes for the future. This article explores the nature of the links between the modernization of crop production and patterns of crop genetic diversity in the rice, maize, and wheat fields of the developing world.

I. INTRODUCTION

During the past 200 years, and in a more accelerated way since 1960, the pattern of crop genetic diversity in the fields of the developing world has changed fundamentally. The germplasm that dominates the area planted with major cereals has shifted from the locally adapted populations that farmers historically selected from seed they saved—often called “landraces”—to the more widely adapted seed types produced by scientific plant breeding programs and purchased by farmers—often called “modern varieties.” These yield-enhancing seed types enabled the intensification of agriculture in areas of the world with high population densities. Initially, they diffused through the environments best suited for their production, spreading later—and unevenly—into less favored areas.

Economic growth, increasing incomes, and migration of labor from rural to urban areas lead to the commercialization of agriculture. Commercial crop production is characterized by the controlled application of water and other purchased inputs, such as fertilizer, fungicides, or pesticides to modern seed types. Sustained yield increases in commercial agriculture have depended not only on continued modifications in crop management practices but also on successive genetic improvements that are accomplished through recombining and exchanging diverse sources of germplasm. The spatial and temporal diversity among modern varieties in farmers’ fields is determined in large part by the economic factors affecting their profitability and the performance of agricultural research institutions and seed industries.

In contrast, the genetic content and the geographical distribution of landrace populations are influenced more by natural selection pressures and the seed and crop management practices of traditional farming communities. Typically, it is not profitable to grow them with large amounts of purchased inputs since they are often grown principally for home consumption of food or feed and the size of their commercial market is limited. Seed markets for these varieties are generally local and rules for the exchange for seed, grain, and other products are more personal.

Today, the modernization of agriculture in the developing world is incomplete. In some areas with rain-fed, upland, or heterogeneous growing environments, landraces still dominate crop area or both modern varieties and landraces coexist. As the process of modernization continues, more farmers will be integrated into specialized product markets. Whereas others will work outside the farm and eventually leave it. Even then, local or export demand for certain quality attributes may provide market-based incentives for continued cultivation of landraces. In some places, there will be no alternative but to grow landraces for food since there may be no urban target for migration.

This article outlines some of the major implications of the modernization of agriculture for crop genetic diversity, focusing on the world’s three principal cereals—rice, wheat, and maize. Wheat is a major tradable crop grown principally in temperate zones over a wide geographic area and often on relatively large, mechanized farms of industrialized countries. The world’s main wheat-producing regions are southern Russia and the Ukraine, the central plains of the United States and adjacent areas in Canada, northwest Europe, the Mediterranean basin, southwestern Australia, and, in the developing world, India, China, and Argentina. Rice is the most important food grain in the developing world. Almost all the world’s rice is both produced and consumed on numerous small farms within Asia, with less than 5% of production entering trade. India, China, and Indonesia are the major producers. Rice is the world’s most widely grown cereal in terms of growing environments and, after wheat, the most extensively traded cereal. Although virtually all the wheat and rice traded internationally are destined for human consumption, most of the maize is used to feed livestock. Maize is a very important part of the human diet in Africa, parts of Latin America, and Asia. Major maize-producing regions include the United States and parts of Europe, Mexico, Brazil, Argentina, South Africa, India, and China.

The next section summarizes the factors driving the intensification of agriculture. Section III explains the implications of intensification for the patterns of genetic variation in the fields of the developing world. Section IV discusses how commercialization affects these pat-
II. THE INTENSIFICATION OF CROP PRODUCTION

A. What is the Intensification of Crop Production?

The intensification of agriculture refers to the increase in output per unit of land used in production, or land productivity. Population densities, expressed by the ratio of labor to land, explain much about where and under which conditions this process has occurred. The transition from low-yield, land-extensive cultivation systems to land-intensive, double- and triple-crop systems is only profitable in societies in which the supply of uncultivated land has been exhausted. The process of agricultural intensification has been observed in traditional and modern agricultural societies. The movement from forest and bush fallow systems of cultivation to annual and multi-crop cultivation systems, whereby plots of land are cultivated one or more times per year, has generally been influenced by increasing population densities in traditional societies. It is no accident that the modern seed-fertilizer revolution has been most successful in densely populated areas of the world, where traditional mechanisms for enhancing yields per unit area have been exhausted.

Intensive cultivation will also be observed in areas with lower population densities provided that soil conditions are suitable and markets are accessible. Intensification occurs in the less densely populated areas for two reasons: (i) higher prices and elastic demand for output imply that the marginal utility of effort increases—hence, farmers in the region will begin cultivating larger areas, and (ii) higher returns to labor encourage migration into well-connected areas from neighboring regions with higher transport costs. Examples of regions with low population density but intensive, market-oriented production are the central plains of Thailand and parts of South America's southern cone.

If the conditions described are not present, labor and other costs associated with intensive agriculture are substantially higher than its incremental economic returns.

Intensification of land use and the adoption of yield-enhancing technologies have occurred in traditional and modern agricultural systems. In the case of Africa, Pingali et al. (1997) documented the movement from shifting cultivation to permanent agriculture with increases in population densities and improvements in market infrastructure. As land became scarce, traditional farming communities across sub-Saharan Africa began to extract increasingly higher levels of output from their land through investments in land improvements and soil fertility management. The intensification of traditional farming systems is a process that the more densely populated regions of Asia had been through several decades and in some cases centuries earlier. The application of modern scientific innovations in the form of high-yielding seeds and fertilizer allowed the extremely land-scarce regions of Asia to achieve levels of land productivity that were not possible through the exclusive reliance on traditional systems of farmer innovation.

B. Seed Technical Change: Traditional and Modern

The genetic evolution of cultivated crops is closely interwoven with the evolution of human civilization and crop husbandry. The recognition of wild species of cereal crops was first accomplished by primitive societies of hunters and gatherers, who harvested them for food. The domestication of wild species occurred when societies of shifting cultivators first began to cultivate food crops. Sedentary or permanent agricultural systems led to the emergence of ancestors of modern-day landraces of cereal crops. Although landraces have evolved over thousands of years under farmer management of natural selection, varieties have been bred by plant breeding programs for about 100 years. The first high-yielding maize hybrids were developed about 50 years ago. The high-yielding varieties of rice and wheat with semidwarf stature were developed less than 50 years ago and have been successfully adopted only in intensive agricultural production systems. Modern cereal cultivars have developed through three main phases of selection: (i) subconscious selection by the earlier food growers in the process of harvesting and planting, (ii) deliberate selection among variable material by farmers living in settlements and communities, and (iii) purposeful selection by professional breeders using scientific methods.

The main achievement of the first phase was to make the crop more suitable for planting and harvest by humans, threshing or shelling, and consumption. Higher germination rates, more uniform growing periods, resistance to shattering, and palatability were some of the achievements of this effort. In the second phase, many
In the next section, the implications of agricultural intensification to crop genetic diversity are outlined. First, crop genetic diversity is defined. Some historical perspective is then provided on the spatial and temporal distributions of genetic variation in rice, wheat, and maize. Current estimates of areas planted with modern and landrace varieties of these cereals are reported: Evidence on the genetic diversity of modern varieties and landraces follows.

III. THE IMPLICATIONS OF INTENSIFICATION FOR CROP GENETIC DIVERSITY

Crop genetic diversity broadly defined refers to the genetic variation embodied in seed and expressed when challenged by the natural and human selection pressures that shape the environment in which it grows. In applied genetics, diversity refers to the variance among alternative forms of a gene (alleles) at individual gene positions on a chromosome (loci), among several loci, among individual plants in a population, or among populations (Brown et al., 1990). Diversity can be measured by accessions of seed held in gene banks, lines or populations utilized in crop-breeding programs, or varieties cultivated by farmers (cultivars).

The relationship between precise quantitative measurements and what can be casually observed among the plants growing in a field, and between these measurements and what can be observed in other environments, is indirect. Crop genetic diversity cannot be literally or entirely observed at any point in time; it can only be indicated with reference to a specific crop population and analytical perspective. To understand the implications of agricultural intensification to crop genetic diversity, it is first necessary to gain a spatial and temporal perspective of the variation in crop plants.

A. Current Spatial Patterns of Genetic Variation

The adoption of modern cereal varieties has been most widespread in land-scarce environments and/or in areas well connected to domestic and international markets. Even in these areas, the profitability of modern variety adoption has been conditioned by the potential productivity of the land under cultivation. For instance, while modern rice and wheat varieties spread rapidly through the irrigated environments, their adoption has been less spectacular in the less favorable environments—the drought-prone and high-temperature environments for wheat and the drought- and flood-prone environments for rice. Maize, as discussed later, has an even spottier record in terms of farmer adoption of modern varieties and hybrids. For all three cereals, traditional landraces continue to be cultivated in the less favorable production environments throughout the developing world.

Recent estimates of the areas planted with modern varieties and landraces are shown in Table 1. Based on data from a global wheat survey conducted by the Centro Internacional de Mejoramiento de Maiz y Trigo (CIMMYT) and a national crop variety survey conducted in Mexico, the area planted with modern wheat varieties increased almost fourfold from 1961 to 1981, from 6.0 million hectares (ha) to 24.0 million ha. Similarly, maize landings increased more than fivefold, from 3.8 million ha in 1961 to 20.3 million ha in 1981. Area planted with modern varieties of rice increased similarly, from 0.8 million ha to 4.7 million ha. In 1981, modern varieties of rice were cultivated on 1 million ha in China and 0.9 million ha in India, the two countries with the largest populations of the world. The area planted with modern varieties of rice in China increased more than fivefold, from 0.2 million ha in 1961 to 1 million ha in 1981. Modern rice varieties are grown with increased levels of fertilizer and a controlled water supply, these varieties performed significantly better than the varieties they replaced. Initially, they spread rapidly throughout many of the irrigated zones of the developing world where rice and wheat cultivation was concentrated and where population densities were high. Later, more widely adapted descendants of these varieties spread gradually into less favorable environments, including rain-fed areas with relatively modest production potential. Their diffusion was faster in the plains and valleys, diminishing up the hillsides and in more heterogeneous environments. The term modern varieties is also used to refer more exclusively to semidwarf varieties of rice and wheat.

Conservationists who observed the popularity of the green revolution varieties expressed concern for the possible loss of valuable genetic resources and narrowing in the genetic stock that serves as the basis for crop improvement (Frankel, 1970; Harlan, 1992; Hawkes, 1983). In the next section, the implications of the intensification of agriculture for crop genetic diversity are outlined. First, crop genetic diversity is defined. Some historical perspective is then provided on the spatial and temporal distributions of genetic variation in rice, wheat, and maize. Current estimates of areas planted with modern and landrace varieties of these cereals are reported: Evidence on the genetic diversity of modern varieties and landraces follows.
TABLE I
Percentage Distribution of Rice, Wheat, and Maize Area by Type of Germplasm in the 1990s

<table>
<thead>
<tr>
<th>Region</th>
<th>Wheat</th>
<th>Maize</th>
<th>Rice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Semidwarf</td>
<td>Tall</td>
<td>Landraces/ Other</td>
</tr>
<tr>
<td></td>
<td>improved</td>
<td>improved</td>
<td>unproved</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>66</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>West Asia/North Africa</td>
<td>66</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>Asia</td>
<td>86</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Latin America</td>
<td>90</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>All developing countries</td>
<td>81</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Industrialized countries</td>
<td>93</td>
<td>45</td>
<td>Trace</td>
</tr>
</tbody>
</table>


*Data for rice in sub-Saharan Africa are West Africa only.

(Agromon) in 1997, approximately 80% of the wheat area in the developing world was sown to semidwarf varieties, with the remainder split almost equally between improved tall varieties and landraces or varieties with unknown ancestry. The relative importance of tall wheat varieties remains greater in the industrialized than in the developing countries, probably for reasons related to wheat-growing environment and management practices.

Most of the extensive area sown to wheat landraces is found in parts of Turkey, Iran, Afghanistan, and Ethiopia. Pockets of landrace diversity for special traits may also be found throughout the Mediterranean region (Morocco, Tunisia, Syria, Egypt, Cyprus, Portugal, Spain, and Italy) as well as in Asia, Southeast Asia, and Tibet, although field research and data analysis would be needed to confirm these hypotheses. Work in Mexico, Guatemala, Honduras, Peru, and Bolivia indicates the presence of unique landrace varieties that were probably introduced by Spanish immigrants in the sixteenth century. No significant areas are believed to be sown to wheat landraces in the former Soviet Union or in Eastern Europe, despite the historical importance of these areas for major progenitors such as the so-called “Turley” and “life” wheat. The loss of landrace populations in the former Soviet Union was no doubt a consequence of collective agriculture that was practiced in the region for several decades.

Approximately three-fourths of the rice area in Asia, which produces most of the world’s rice, is thought to be sown to semidwarf varieties. Semidwarf varieties dominate the irrigated rice ecosystems of Asia and cover substantial areas in the favorable rain-fed lowlands (M. Jackson and G. Khush, personal communication). In the more marginal rain-fed lowland environments, the deep-water environments, and the upland rice environments, farmers continue to grow landraces adapted to those particular environmental niches. Although more accurate data concerning rice landraces in Asia are now being compiled at the International Rice Research Institute (IRRI), the most comprehensive data from a national-level sample survey conducted by the Department of Agricultural Extension in Bangladesh (1996–1997) suggests that although farmers in that country still grow large numbers of named varieties including landraces (more than 300), more than 20 modern varieties cover nearly 60% of the total rice area.

In West Africa, semidwarf varieties cover essentially all the irrigated lowlands and one-third of the rain-fed lowlands, but landraces and other improved varieties appear to occupy most of the area in uplands, mangroves, and deep-water flooded areas. Particularly in the deep-water floating rice environment near Mopti in Mali, Birnin Kebbi in Nigeria, and in northern Guinea are found Oryza glaberrima (African rice) and tall traditional O. sativa (Asian rice) cultivars. The West African Rice Development Association (WARDA) has recently developed hybrid crosses of O. sativa and O. glaberrima species. In East Africa, the area in Madagascar seems to be approximately evenly split between traditional and modern varieties. In Latin America, tall and landrace varieties cover most of the upland area, whereas few are found in the irrigated and rain-fed lowlands. In the uplands of Asia, traditional varieties still domi-
nate, with "pockets" of important diversity in the Bastar Plateau (Madhya Pradesh, eastern India), parts of northern Bangladesh and the Chitrangol Hill Tract in that country, northern states of Myanmar, almost all of Laos, and parts of Cambodia (M. Jackson and G. Khush, personal communication). Oka (1988) reported that the genetic diversity in landraces of Asian rice was most prevalent in the area extending over Assam, Bangladesh, Burma (now Myanmar), Thailand, Laos, and Yunnan, China. Although Oka also reported that diversity was relatively high in Indonesia, which lies outside the center of domestication, much of Indonesian ricelands are now planted with modern varieties.

Data from a global maize surveys conducted by CIMMYT in 1992 and 1997 indicate that, relative to wheat and rice, far less of the maize area in the developing world is planted with maize types released by plant breeding programs. Hybrids appear to occupy an increasing proportion of the area planted with modern maize types, but in zones such as Latin America, most of the maize area is still planted with landraces.

At least some of the area listed under landraces is planted with populations that result from the genetic integration of modern varieties with landraces when farmers save seed or plant seed of different types in adjacent fields. Similarly, a substantial proportion of the maize area in sub-Saharan Africa is planted with advanced generations of improved varieties whose seed farmers could not afford to replace on a regular basis introduced with landraces brought to the continent with the slave trade several centuries ago. Although there appears to be a high proportion of maize landraces grown in West Asia and North Africa, the total area in maize is limited in this region and many of these are likely of unknown origins. In the industrialized world, a negligible percentage of maize area is planted with either improved open-pollinated varieties or landraces. These are specialty maize or "heritage" varieties grown for fresh consumption, popcorns, or ornamental corns such as those marketed on holidays in the United States.

The data in Table 1 confirm that although much of the genetic variation in the world's wheat and rice area is shaped today by the efforts of modern plant breeders, a substantial part of the genetic variation in the maize grown in the developing world remains in the hands of farmers, some of which are among the poorest in the world. In some zones of production, such as Turkey and Iran for wheat or parts of Latin America for maize, fairly large contiguous areas may be planted to landraces. Often, in regions of crop domestication and diversity, landraces persist "as patches and islands of farming systems" (Brush, 1993, p. 246). Harlan (1992, p. 147) invoked the term "microcenters of diversity" to describe relatively small regions, 100–500 km across, in which may be packed an astonishing variation within extensive areas of apparent uniformity in parts of Turkey, the Caucasus, Iran, and Afghanistan. Oka (1986) called the Jeypore Tract in India a microcenter of diversity in Asian rice. Jackson (personal communication) refers to some of the upland rice areas of Asia as "pockets" of diversity.

The change in the crop genetic landscape from predominantly traditional to largely modern patterns of genetic variation occurred during the past 200 years and at an accelerated rate since the 1960s. Whether the change to modern varieties has resulted in a narrowing of genetic diversity remains largely unresolved for many reasons, which are discussed in the following sections and further in Smale (1998) and Wood and Lennie (1997).

B. Comparing Genetic Diversity in Landraces and Modern Varieties

A major problem in assessing whether genetic narrowing has occurred with the replacement of landraces by modern varieties is the magnitude of the sampling and measurement effort that would be required to test the hypothesis in a meaningful way. In most cases it would be impossible to locate reliable samples of the landraces originally grown in an area now planted with modern varieties since this process occurred over time and unevenly across environments.

In some sense, the genetic diversity of landraces and modern varieties is incomparable by definition since its structure is distinct for each. Hawkes (1983) wrote that landraces, which are mixtures of genotypes, "could not even be called varieties." He called the range of genetically different varieties available to breeders "the other kind of diversity" (pp. 100–101). Harlan (1992) described a landrace as "an integrated unit" of "component genotypes" that have adjusted to one another over the generations as well as to the local environment, both natural and man-made (p. 148). A field planted with a rice or wheat landrace may be viewed by a plant breeder as a mixture of several lines but viewed by the farmer as one single variety because of its recognizably distinct agronomic, processing, or consumption characteristics.

Vaughan and Chang (1992) described traditional rice types that are mixtures or composites of morphologically distinguishable types grown together deliberately.

The breeding system of the crop plant also affects the structure of genetic variation. Landraces generally contain some heterozygous material, though the extent of segregation is considerably greater for open-pollinated than for self-pollinated species. Maize is a cross-
pollinating species, and maize in Mexico is often cited as an example of deliberate manipulation of the composition of landraces by farmers. Landrace varieties of maize evolve continuously through the purposeful mixing of seed lots of the same varieties or introgression by farmers, as well as inadvertently when fields of maize planted with different varieties flower simultaneously. A single “race” or race complex, as understood by maize geneticists, contains numerous genetically distinct farmers’ varieties.

The distinction between modern varieties and landraces can also be blurred for predominantly outcrossing crops such as maize, making it difficult to determine which population is under study. Many small-scale, subsistence-oriented maize farmers promote hybridization between improved varieties and landraces by growing them together or in neighboring fields and producing what farmers in Mexico call “crossed” varieties. Also termed “rustication” or simply “adaptation,” this process may enable improved varieties to fit better the need of local farmers.

These points imply that comparing counts of landraces and modern varieties as an index of genetic narrowing may not make sense. They also imply that even if reliable samples of the landraces originally cultivated in an area could be obtained, analyses comparing their genetic diversity might provide only part of the answer regarding genetic narrowing. Although the landrace in the farmers’ field is a heterogeneous population of plants, it is derived from generations of selection by local farmers and is therefore likely to be local in adaptation. The plants of a modern variety are uniform but the diverse germplasm in the genetic background may enable them to adapt more widely. The diversity in a modern variety may not be expressed until challenged by the environment. On the other hand, the landrace may carry an allele that occurs rarely among modern varieties and is a potentially valuable source of genetic material not only for the farmer that grows it today but also for future generations of producers and consumers.

C. When Did Genetic Narrowing Occur?

Another problem in assessing the relationship of modern varieties to genetic narrowing is the temporal point of reference. Porceddu et al. (1988) described two major stages of genetic narrowing in wheat during modern times. The first occurred in the nineteenth century when scientific plant breeding responded to the demand for new plant types. Farming systems emerged that were based on the intensive use of land and labor, livestock production, and the use of organic manure. Changes in cultivation methods favored genotypes that diverted large amounts of photosynthates into the ear and grain. Bell (1987) reports that the engineering innovations of the late nineteenth century led to the establishment of extensive wheat-growing areas in North America, Australia, and parts of South America. Mechanization of agriculture dictated uniformity in plant type.

According to Porceddu et al. (1988), a second stage of narrowing occurred in the twentieth century, when genes were introduced to produce major changes in plant type. Use of the dwarfing genes Rht1 and Rht2, for example, conferred a positive genotype-by-environment interaction in which yield increases proved greater given a certain combination of soil moisture, soil fertility, and weed control. Varieties carrying these dwarfing genes were developed by N. Borlaug with the national breeding program in Mexico and later by the International Maize and Wheat Improvement Center (CIMMYT). They became known as the green revolution wheats.

Evenson and Gollin’s (1997) summary of the history of rice breeding suggests a process of continual expansion and narrowing of the genetic pool. Organized breeding efforts probably date earlier than 1000 AD in China. Modern efforts can be traced to the late nineteenth century in several parts of Asia. In temperate east Asia, the first significant advances were made by Japanese farmers and scientists when they developed relatively short-statured and fertilizer-responsive cultivars. Known as the ronō varieties, these belonged to the japonica class of rice and were widely cultivated in Japan as early as the 1860s. During the Japanese occupation of Taiwan in the early part of the twentieth century, Japanese scientists sought to adapt these varieties to the more tropical conditions of Taiwan. At the same time, researchers in tropical Asia were seeking more productive varieties of rice from the indica and javanica classes of rice. After World War II, the United Nations Food and Agricultural Organization initiated a program to cross indica rice with japonica as a means of increasing rice yields, culminating in the formation of the IRRI and the green revolution varieties of rice.

To Vaughan and Chang (1992), genetic narrowing in modern rice began early in this century. Development projects, population increases, and forest clearing in Asia were the primary causes of the loss of wild and cultivated rice landraces. In the Mekong Delta, the replacement of traditional deep-water rice by irrigated rice occurred with drainage and irrigation schemes that were introduced during the French colonial period.
are popular. Because such innovations often produce varieties that the breadth of materials grown by farmers precisely a plant breeding program may lead to a narrowing of 

agriculture, today's broadening of the genetic pool in rice breeders. As these examples suggest, in modern classes of rice extended the gene pool accessible to 

to increase rice yields by crossing 

japonica and indica classes of rice extended the gene pool accessible to rice breeders. As these examples suggest, in modern agriculture, today's broadening of the genetic pool in a plant breeding program may lead to a narrowing of the breadth of materials grown by farmers precisely because such innovations often produce varieties that are popular.

D. Trends in Genetic Diversity of Modern Varieties of Rice, Wheat, and Maize

Part of the concern for genetic narrowing is based on the perception that, with time, conventional plant breeding practices inevitably restrict the genetic base of modern varieties. The evidence from studies on the parentage of modern varieties lends little support to the view (Witcombe, 1999). In an analysis of genealogies of 1709 modern rice varieties, Evenson and Gollin (1997) found that although a variety released in the 1960s had 3 landraces in its pedigree, recent releases have 25 or more. The complexity of rice pedigrees, in terms of parental combinations, geographical origin, and number of ancestors, has expanded over time. A similar pattern has been shown for about 800 wheat varieties released in the developing world since the 1960s (Smale, 1997). The average number of distinct landraces found in bread wheat pedigrees increased from approximately 20 in the mid-1960s to about 50 in 1990.

Slovmand and de Lacy (1999) analyzed the distance among coefficients of parentage for a historical set of CIMMYT wheat varieties during the past four decades. Their results show a rate of increase in genealogical diversity that is positive but decreases over time, with marked expansion in genealogies from 1950 to 1960 and gradual flattening through the 1990s. If progenitors were recycled and reused, the distance among them would decrease over time and the slope of the line would be negative.

D. Trends in Genetic Diversity of Modern Varieties of Rice, Wheat, and Maize

Part of the concern for genetic narrowing is based on the perception that, with time, conventional plant breeding practices inevitably restrict the genetic base of modern varieties. The evidence from studies on the parentage of modern varieties lends little support to the view (Witcombe, 1999). In an analysis of genealogies of 1709 modern rice varieties, Evenson and Gollin (1997) found that although a variety released in the 1960s had 3 landraces in its pedigree, recent releases have 25 or more. The complexity of rice pedigrees, in terms of parental combinations, geographical origin, and number of ancestors, has expanded over time. A similar pattern has been shown for about 800 wheat varieties released in the developing world since the 1960s (Smale, 1997). The average number of distinct landraces found in bread wheat pedigrees increased from approximately 20 in the mid-1960s to about 50 in 1990.

Slovmand and de Lacy (1999) analyzed the distance among coefficients of parentage for a historical set of CIMMYT wheat varieties during the past four decades. Their results show a rate of increase in genealogical diversity that is positive but decreases over time, with marked expansion in genealogies from 1950 to 1960 and gradual flattening through the 1990s. If progenitors were recycled and reused, the distance among them would decrease over time and the slope of the line would be negative.

Less evidence is available worldwide on trends in the pedigrees or ancestry of maize varieties than for rice and wheat, in part because this information is confidential in an increasingly privatized industry. Following the epidemic of corn blight in the U.S. crop in 1970, the National Research Council (1972) concluded that the genetic base of maize in the United States was sufficiently narrow to justify concern. Day and coworkers (1984) found that during the 10 years following the 1970 epidemic, breeders had broadened their germplasm pools. Molecular markers, like genealogies, can be used to construct indicators of the latent diversity in a set of crop populations. Using molecular markers, Donini et al. (2000) concluded that there is no objective evidence to support the assertion that modern plant breeding has reduced the genetic diversity of U.K. wheats since 1930. Recent molecular evidence for a set of CIMMYT wheats indicates that genetic distance has been maintained among major parents and popular varieties during the past 30 years (unpublished data). Since many of the varieties of spring bread wheat grown in the developing world have a combination of CIMMYT and locally bred materials in their ancestry (Heisey et al., 1999), these data represent a lower bound on actual genetic diversity. Furthermore, the genetic diversity that is accessible to conventional plant breeders today includes not only spring bread wheat, of course, but also wheat types with different growing habits, close relatives, and wild grasses. Techniques of biotechnology may traverse the species barriers faced by conventional breeders.

IV. THE COMMERCIALIZATION OF CROP PRODUCTION

A. What Is Commercialization of Crop Production?

Economic growth, urbanization, and the withdrawal of labor from the agricultural sector have led to the increasing commercialization of agricultural systems.
Subsistence-oriented monoculture food production systems give way to a diversified market-oriented production system. Agricultural commercialization means more than the marketing of agricultural output: It means that product choice and input use decisions are based on the principles of profit maximization. Commercial realignment of agricultural production occurs for the primary staple cereals and for the so-called high-value cash crops. Commercialization of agricultural systems leads to greater market orientation of farm production, progressive substitution of nontraded inputs in favor of purchased inputs, and the gradual decline of integrated farming systems and their replacement by specialized enterprises for crop, livestock, poultry, and aquaculture products (Pingali, 1997).

On the demand side the process of agricultural commercialization is triggered by rapid income growth and the consequent diversification in food demand patterns. A slowdown in income-induced demand for rice and for coarse grains is accompanied by a shift of diets to bread and higher valued foods such as meat, fruit, and vegetables. These dietary transitions are induced by the growth in per capita income and by the rapid migration of population to urban areas. The need to provision the rapidly growing cities of the world also acts as an impetus for the transformation of food production systems.

On the supply side, growing factor scarcities contribute to the demise of subsistence agricultural systems. Although growing land and water scarcity can be compensated for with increasing scientific knowledge and farmer management, farmer time required for sustenance is increased. Food production systems will become increasingly scarce. The collapse of subsistence systems will come about because of the competing demands for farmers’ time. Although the speed of the structural transformation differs substantially across countries they are all moving in the same direction.

B. Seed Industries

As the orientation of crop production shifts from subsistence toward commercial objectives, the locus of crop improvement and seed distribution moves from individual farmers toward an organized seed industry composed of specialized private and public organizations. In terms of an increased reliance on commercially produced seed, this has occurred substantially faster for maize than for rice and wheat. In a stylized depiction of the maize seed industries in developing countries, subsistence production is characterized by open-pollinated varieties improved through farmer selection and on-farm seed production with local seed markets governed by custom. In a fully commercial system, the predominant seed type is a hybrid that is purchased annually. Seed is traded globally and is a product of specialized research that is both privately and publicly funded. The exchange of seed and the genetic resources used to improve it are enabled and protected by strict forms of intellectual property rights.

For rice and wheat, which are self-pollinating crops, the incentives for privatization of research have not always been as strong as those for maize, although this depends on the institutional and economic context. In industrialized countries, profound changes in science and in intellectual property protection during the past 20–30 years have been associated with a higher rate of investment in agriculture by the private sector than the public sector and a shift in the composition of private investment from agricultural machinery and processing into chemical research and plant breeding. Although privatization is greatest in the maize seed industry, it is also occurring in wheat to a lesser extent and particularly in Europe. There is very little private sector rice breeding anywhere in the world. Almost all the research on rice has been conducted by the public sector, and most of this has taken place in Asia. In the developing world, there is increasing privatization of the maize seed industry but rice and wheat remain primarily public.

In commercial systems of rice, wheat, and maize, recent changes in the structure of the seed industry are likely to have implications for the utilization of modern patterns of genetic variation. The global seed industry has integrated both vertically (within production processes) and horizontally (among production processes) into “life science” firms that combine seed, chemical, and pharmaceutical businesses. As part of this structural change, firms are engaging in strategies to ensure more exclusive proprietary rights, including, for the first time in history, patents on genetically modified organisms.

The implications of these changes for the exchange and utilization of the genetic resources that are used in breeding modern varieties are unknown. Efforts are under way to harmonize intellectual property regimes globally through international trade agreements, but differences between developed and developing countries, as well as among developing and developed countries, pose challenges. Small public plant breeding programs in developing countries are not on the same footing with respect to investments and legal clout as the life science conglomerates. The maize, wheat, and rice industries are likely to be affected in different ways given the nature of economic incentives associated with seed reproduction. Patents are only one type of intellec-
tual property right; in addition to intellectual property
rights, national seed regulatory systems will have a
strong impact on farmers’ access to seed and the conser-
vation of plant genetic diversity.

V. THE IMPLICATIONS OF
COMMERCIALIZATION FOR CROP
GENETIC DIVERSITY

Agricultural commercialization influences the extent of
crop genetic diversity in two ways: (i) through changes
in land use patterns and (ii) through crop choice
changes in the irrigated as well as the rain-fed environ-
ments. The organization and management of food pro-
duction systems in both the irrigated and the rain-fed
environments are affected by economic growth. The
opportunity cost of family labor can be expected to
increase equally in the high- and low-potential areas
since the populations in both environments are re-
sponding to nonagricultural employment opportuni-
ties. The declining viability of subsistence production
systems can also be expected to be similar. The move-
ment from subsistence to market-oriented, rain-fed pro-
duction could follow a general pattern: (i) the abandon-
ment of highly drought-prone environments, especially
in areas where the opportunities for groundwater ex-
ploration are limited; (ii) the shift from small subsis-
tence farms to mechanized cultivation of large farms;
and (iii) where dry season water supplies are available,
increased areas under vegetables, feed grain and fodder
crops, and other high-valued crops. Cereal crop produc-
tion would continue to have a comparative advantage
in the rain-fed environment, primarily because of the
high cost of modifying the environment in order to
make it suitable for noncereal crops. Low-input, low-
yield cereal production systems, rice and maize, grown
on consolidated holdings may emerge as the most viable
option for the rain-fed environments. The irrigated en-
vironments, on the other hand, would shift from being
predominantly under cereal monoculture to a highly
diversified production system. The implications of these
changes in crop choice on genetic diversity in the irri-
gated and rain-fed environments are discussed in the
following sections.

A. Implications When Intensification
Has Occurred

In the irrigated rice and wheat production zones of
the developing world, commercialization has had little
impact on crop genetic diversity beyond that of agricul-
tural intensification. Since the modern varieties grown
there are varieties rather than hybrids, their seed is
saved and spreads now, as in the past, from farmer to
farmer. This will change if hybrids, or certain types of
transgenic varieties that require annual seed purchase,
are developed. It may also change if the demand for
labor-saving technologies such as herbicides leads to
the use of herbicide-tolerant varieties. These must be
purchased annually to prevent the carryover of weed
seed.

The effects of commercialization on maize crop ge-
genetic diversity, independent of those associated with
agricultural intensification, are much more pronounced
because of the historical importance of hybrids relative
to improved, open-pollinated varieties. The distinctive
biological properties of maize plants (in particular, their
propensity for open pollination and their tendency to
segregate) make it difficult for farmers to maintain the
genetic purity of maize seed saved from their own har-
vest. Commercial maize growers are therefore depen-
dent on reliable external sources of affordable seed in
a way that growers of self-pollinated rice and wheat are
not. The reliance on the seed industry will continue to
grow for maize if farmer use of genetically engineered
seed increases in importance in the future.

The survival of landrace diversity for cereal crops in
the high-potential, irrigated environments would de-
depend on farmer incentives for maintaining that diver-
sity. To a large extent, farmer incentives to do so would
depend on the market demand for the unique quality
characteristics that are present in some of the landraces.
The importance of Basmati rice in the irrigated produc-
tion zones of India and Pakistan is an example of how
market demand for quality can influence the survival of
traditional varieties and landraces. Even where modern
varieties are used exclusively, diversity within the plant
has increased over time, as discussed previously, by the
introduction of new gene pools through breeding.

B. Implications When Intensification Has
Not Occurred

The areas of the developing world where modern varie-
ties are not widely grown are typically marginal for
production of the crop or are inadequately served by
markets and infrastructure. In some areas, agricultural
research has been unable to produce varieties demonstr-
ating an obvious yield advantage or commercial seed
systems have not had the incentive to do so because of
the small size of the market or fluctuating effective
demand.
As suggested previously, this may remain the case for sizable portions of the developing world's maize area. The proportion of cultivated area that is irrigated is far less for maize than for rice and wheat, whereas the use of purchased inputs remains modest. Most of the farmers who grow maize in developing countries face difficult and variable maize production environments and cultivate it with the primary objective of meeting subsistence requirements. These farmers have little incentive to make investments in fertilizer, pesticides, and other modern means of coping with disease and weather since their traditional varieties do not respond as well to these as modern varieties. In many of these production zones, it is not easy to breed well-adapted materials and there are few profits to be earned for seed companies.

In the most difficult environments, commercialization is likely to lead to the complete abandonment of crop production, as has already occurred in parts of Asia. When there are limited opportunities for migration but environments are too marginal for specialized agricultural production to be profitable, farmers may remain on small landholdings and grow landraces for subsistence.

Even when a zone may be suitable for the production of modern varieties, the development of commercial seed systems is not sufficient to ensure that they will replace landraces in the near future because markets are imperfect. In some local communities, the specific varietal traits demanded by farmers (grain quality, fodder, and suitability for a certain soil type) cannot be obtained through the production of modern varieties or procured through impersonal market transactions so that farmers must rely on their own production or that of nearby farmers for their supply of a valued attribute. The specialized uses of certain landrace varieties for medicinal purposes, rituals and festivals, and culinary practices have been extensively documented. Small-scale farmers’ choice to grow more than one variety simultaneously is likely to reflect their need to address numerous concerns that no single variety can satisfy (Bellon, 1996). Farmers often choose to grow both landraces and modern varieties. Zimmerer (1999) found that the capacity of farmers to grow diverse food plants (including maize) in Peru and Bolivia depends on whether they can cultivate them in combination with commercially developed, high-yielding varieties. Vaughan and Chang (1992) noted that the rapid changes that occur with natural calamities are more likely to have a greater impact on the loss of rice genetic diversity than the farmer-driven, incremental changes that are regularly occurring, many of which enhance diversity (Dennis, 1987). Meng et al. (1998) concluded that multiple factors, including missing markets, yield risk, grain quality, and agroclimatic constraints, influence the probability that a Turkish household will grow a wheat landrace; a change in any single economic factor is unlikely to cause farmers to cease growing it.

Viewed in the conventional microeconomic literature as partial adoption, this observed pattern has been explained theoretically through attitudes toward risk and uncertainty, nonexisting markets, and differential soil quality or nutrient response combined with fixity or rationing (Meng, 1997; Smale et al., 1997). Although treated as a transitional period to full adoption (or replacement), the coexistence of modern varieties and landraces may represent an equilibrium if one or several of these aspects persist despite economic change. Then there are locally based economic incentives for farmers to continue to grow landraces.

Even when the pressures for market integration are strong, the coexistence of modern varieties and landraces may also persist with certain types of market-based incentives, as discussed previously. In the early phases of economic growth when rural populations move to urban areas, market integration exerts pressures for uniformity in the attributes of coarse grains. Localized preferences diminish in favor of cheaper, bulk-marketed grains. The elasticity of demand for staple grains declines as income increases, and it is sometimes negative. Generally, rice substitutes for maize, and wheat substitutes for rice.

The income elasticity of demand for attributes of the grains may be higher, however, than the income elasticity of demand for the cereal (Pingali et al., 1997). For example, a notable pattern of rice consumption is that, with growing incomes, people express preferences for higher quality rice once their calorie needs have been met. High-income consumers spend more on rice by paying higher prices for varieties with preferred eating quality which they substitute for the lower quality variety consumed when the their income levels were lower. In Asia, traditional varieties are generally of higher quality and fetch premium prices in the market. Thailand still grows low-yielding traditional rain-fed varieties extensively for the export market. When the income level was low, South Korea used to grow the modern “tongil” variety, but this was replaced by relatively low-yielding traditional japonica rice as consumers expressed preference for japonicas by offering higher prices. In response to these market signals farmers are eager to grow even low-yielding, high-quality rice because the higher prices more than compensate for their lower yields. Because rice scientists have had limited...
success in developing high-yielding cultivars with better eating quality, the price difference between the standard- and high-quality varieties has been increasing in Asian markets.

The post-industrial agricultural economy is characterized by growth in demand for an array of increasingly specialized goods and services. Some product quality attributes are associated with features of the production process (organic/inorganic). Some are extrinsic (origin or effects on animal welfare), whereas others cannot be discerned without laboratory tests (genetically modified organisms). The elasticity of demand for such attributes is likely to increase with very high levels of income. Under these conditions, global market integration may provide market-based incentives for continued cultivation of patches of diverse landraces. In addition to their demonstrated private economic value to the farmers who grow them, some of these hold potential for niche markets and exports.

### VI. CONCLUSIONS

In areas of the world that are more favorable for agricultural intensification, the conversion from landrace varieties to modern varieties has been almost complete for rice, wheat, and maize. In some of the remaining crop-growing environments, and especially in relatively small pockets of crop diversity called micro-centers, farmers who are linked to commercial agriculture through labor markets or specialized product markets continue to grow landraces—often in combination with modern varieties. Some cannot obtain or afford to purchase seed on a routine basis, so they both purchase and save the seed of modern types. Others purposefully adapt modern varieties to their own conditions by saving and selecting the seed, genetically integrating modern and landrace types.

The structure of genetic diversity is distinct for modern varieties and landraces. Both are essential to the future food supply. Conservationists propose that landrace diversity must be maintained not only in preserved stocks called ex situ collections but also in situ, or in the fields of farmers. The future of landrace cultivation appears to be uncertain. Although some view the replacement of landraces by modern varieties as an inevitable product of agricultural commercialization, idiosyncratic growing environments and consumer preferences may provide economic incentives for their continued cultivation by farmers—although on a limited scale. It appears unlikely that modern varieties of rice, wheat, and maize will entirely replace landraces in the near future, although it is difficult to postulate about equilibrium areas planted with each type of germplasm since the equilibrium itself shifts with technical and economic change.

Continued genetic improvement does not necessarily lead to loss of genetic diversity in areas where modern varieties dominate, especially when access to germplasm is relatively unrestricted and innovative plant breeding strategies may be employed. Access to diverse sources of germplasm is therefore of great importance to the success of public and private breeding programs for the supply of varieties in modern agriculture. The continued advances in yield potential that are a necessary (although not a sufficient) condition for alleviating hunger are thought to depend on increasingly complex combinations of genes and novel alleles. Landraces and wild relatives have served as repositories for resistance to biotic and abiotic stress when these are absent in advanced breeding materials. Even in the parts of the world where the "ancient patterns of diversity" may still be found, access to the products of modern plant breeding is often integrated economically or genetically to generate more resilient and sustainable systems.

### See Also the Following Articles

- **AGRICULTURE, SUSTAINABLE**
- **AGRICULTURE, TRADITIONAL**
- **CROP IMPROVEMENT AND BIODIVERSITY**
- **GENETIC DIVERSITY**
- **HERBICIDES**
- **PESTICIDES, USE AND EFFECTS OF**

### Bibliography


AGRICULTURE, SUSTAINABLE

G. Philip Robertson and Richard R. Harwood
Michigan State University

I. Elements of Sustainability
II. Development of the Sustainability Concept
III. Indicators of Sustainability
IV. Examples of Sustainable Cropping Systems
V. Conclusions

GLOSSARY

bush-fallow cropping system An agronomic system in which soil fertility is maintained by allowing native vegetation to regrow following several years of cropping.

conservation tillage Tillage that reduces soil disruption in order to conserve organic matter and water and reduce erosion.

ecosystem services Services provided by an ecosystem to the organisms inhabiting the system or living nearby.

legumes Plants belonging to the family Leguminosae, many of whose members can form symbiotic associations with nitrogen-fixing bacteria.

mixed farming system An agronomic system that incorporates different combinations of herbaceous crops, trees, and animals within a single farm unit.

nitrogen fixation The transformation of atmospheric nitrogen into a form usable by plants.

resource conservation The protection and enhancement of resources on which sustainability depends.

rotation The sequence of crops grown in a single field.

soil organic matter Material in soil containing organic carbon derived from the decomposition of mainly plant residue.

trophic complexity The number and types of organisms that feed at different trophic levels within a community; also known as food web complexity.

SUSTAINABLE AGRICULTURE describes a food and fiber production system that is economically viable, environmentally safe, and socially acceptable over long periods.

I. ELEMENTS OF SUSTAINABILITY

Strictly defined, sustainability denotes any system capable of persisting. Because persistence depends on scale, the term "sustainable agriculture" is entirely scale dependent. An agricultural practice that is sustainable at the scale of an individual field may lack sustainability at the larger farm scale if the inputs required to maintain stable production eventually exceed the capacity of the farm to provide them. Likewise, farm-scale sustainability must be evaluated in the context of local and global regions, and global sustainability requires consideration of global output vs the long-term costs of that output and the ability of global resources to accept those costs.

Of course costs are not simply economic but also social and environmental: If a cropping system requires
large inputs of fertilizer that leak from the system to pollute groundwater drinking supplies and distant coastal fisheries, a system that may be sustainable at the field scale becomes unsustainable at the farm and regional scale—even though the long-term supply of fertilizer is stable and the economic cost of fertilizer is easily borne by higher grain production.

Sustainable agriculture must therefore be defined not just in terms of its long-term economic productivity but also in terms of its environmental and social costs. Many of these costs are value-laden. For a society that values family farms, for example, sustainability must be evaluated in terms of the impact of farming practices on the social structure of rural communities: Do the capital expenses associated with large-scale production systems exclude small producers? For a society that values biodiversity, sustainability must be evaluated in terms of the impact of farming practices on wildlife health and habitat. These considerations take sustainability arguments into sometimes contentious territory because different evaluators may have different social or cultural values. What is socially acceptable in one nation or to one segment of society may be socially unacceptable to another, and these differences must be clearly defined when evaluating sustainability.

Therefore, it is possible to define sustainable agriculture as any suite of agronomic practices that are

- economically viable,
- environmentally safe, and
- socially acceptable.

These elements provide an operational definition of agricultural sustainability. There are many ways to blend them into an overarching definition, and many authors—and policymakers—have done so. One should always keep in mind the subjective nature of these amalgamations, however, and that sustainability is a relative term. Although it may be difficult to foresee what ultimately will be sustainable, one can usually judge whether one set of practices is likely to be more sustainable than another at a societal scale.

II. DEVELOPMENT OF THE SUSTAINABILITY CONCEPT

Sustainable agriculture as a descriptive term evolved toward common usage in the United States in the early 1980s as a mixture of concepts, ideas, values, and development direction that many believe is a vision for what agriculture should be. As with most visions, it has had strong impetus from a small group of critics of the conventional paradigm, notably Wes Jackson in his 1980 book *New Roots for Agriculture,* and by 1984 at a conference at Michigan State University the term was in widespread use (Edens et al., 1985). By 1991, the term had been fully "legitimized" as evidenced in the National Research Council's (NRC) documentation of the many U.S. Department of Agriculture, university, and other programs in sustainable agriculture throughout the United States and abroad (NRC, 1991).

Many of the philosophical roots for sustainable agriculture that coalesced in the 1980s were traced from as far back as the Greek and Roman philosophers by Harwood (1990). The public agenda debate over sustainable agriculture of the 1980s embodied many of the older concepts but certainly not any one in its entirety. The notion of land ownership, cultivation, and personal connectedness to the land as a grounding for personal responsibility, morality, and sense of purpose can be found in the writings of English philosopher John Locke in the 1600s, those of Thomas Jefferson in the late 1700s, and recently in the poetry and books of Wendell Berry (1977). Wes Jackson applied the concept to the 1990s in his book *Becoming Native to This Place* (1994), which focused on homecoming and on being native, i.e., developing an ecological literacy in a holistic sense, through being, living, and having personal experience with one's ecosystem. The values in having a sense of place, a connectedness, have changed from an earlier focus on moral and political values; these values have nevertheless become a part of the vision.

The concepts of conservation and natural resource preservation are woven throughout sustainability, derived from the work of Aldo Leopold and others earlier in this century. Louis Bromfield (1947) and the prolific writings of Faulkner (1943, 1952) had major impact. The writings of Faulkner had a significant influence on both J. I. Rodale and his son, Robert, who carried the notion further to one of regenerative agriculture (Rodale, 1983), defined as an agriculture which not only maintains the natural resource base but also restores and increases its productive potential.

Organic, biodynamic, and the many schools of thought and broad literature surrounding these terms made major contributions to the sustainability debate. The "humus farming" school, with its focus on management of soil organic matter and the interconnectedness of soil health, plant health, and that of animals and humans, was a foundation for this philosophy. Several
schools of organic practice originated in Europe and England, but the movement was popularized as organic agriculture in the United States by J. I. Rodale in his widely read book Pay Dirt (1945).

Each of these philosophical roots contained the notion of a holistic structure of agriculture—an interconnectedness between people, other living organisms, and the soil. Organic agriculture today focuses on the living soil, on optimizing the use of biological processes and on avoiding the use of synthetic chemicals and fertilizers. Most sustainable advocates agree with a biological focus and hope to reduce but not necessarily eliminate chemical use.

Major impetus for the new vision also evolved from the 1960s era of intense agricultural development. The magnitude of the chemical revolution with its newly available pesticides, herbicides, and fungicides plus the availability of concentrated, more easily handled fertilizers very much narrowed the agricultural development focus to the high productivity of major cereal crops. The very real specter of massive starvation in China and hunger and food deficits in India, Bangladesh, and many other countries drove the agricultural development of the 1960s. The low cost of oil and gas and prosperous northern economies all combined for infrastructure development, high agricultural inputs, and the well-known green revolution of this period. A backlash of environmental impact, embodied in Rachel Carson's Silent Spring (1962), the oil crisis of the early 1970s, and concern over a growing gap between rich and poor people and nations all set the stage for the new vision. The senescence and decline of rural communities in the United States during the 1970s and 1980s added urgency to the social dimensions of agriculture. The narrowness and focus of the development debate of the 1960s set the stage for a new, broader vision that included revisiting the old and much introspection amid assurances of decreasing real food prices and near-term global food security.

The 1980s and early 1990s produced a plethora of sustainable agriculture development writings. Future Horizons: Recent Literature in Sustainable Agriculture (Hegyes and Francis, 1997) is an excellent review of this broad literature. The terms "low-input sustainable" and "alternative agriculture" appeared frequently and are still used by many non-American authors, particularly with reference to organic production systems. The American Journal of Alternative Agriculture was founded in the mid-1980s, and in 1989 the NRC's report, Alternative Agriculture, brought the name "alternative agriculture" to prominence. Alternative agriculture was defined, in these sources, as any food or fiber production that has

- a more thorough incorporation of natural processes;
- reduced use of off-farm inputs, with less harm to the environment and consumers;
- a more productive use of the biological and genetic potential of plants and animals;
- a better match between cropping patterns and the physical capacity of lands; and
- an improved emphasis on conservation of soil, water, energy, and biological resources.

Alternative agriculture is not synonymous with organic agriculture (which completely avoids synthetic chemical inputs), but they share many of the same farm management practices and approaches.

The 1980s and 1990s have led us from a primary focus on engineering and chemistry in agriculture toward a greater emphasis on biology—from an age of "alchemy" to the age of "algeny" (Rifkin, 1983). The adoption of the term "agroecology" signaled a reemphasized trend in holistic thinking and analysis. A shift in emphasis within the field of ecological science toward managed ecosystems in the late 1980s added significant perspective. Entomologists, agronomists, and other scientists began to use an ecological, process-oriented approach. Stephen Gliessman's Agroecology: Ecological Processes in Sustainable Agriculture (1998) is one of the latest books on this topic. The infusion of ecological thinking has added clarity to our understanding of overlays of subsystems, of spatial hierarchies, and of how complexity makes up an agricultural ecosystem. It has added to the understanding of multiple functions of an agricultural system and to the notion of ecosystem services. Most importantly, agroecology has taken the analysis of sustainability to a process level of understanding that allows us to understand gradients of change in production systems across time and space.

Giving greater voice in development direction to farmers has been another significant dimension in sustainable agriculture. It was very apparent that much, if not most, of alternative agriculture had originated with farmers. Philosophical direction had been heavily influenced by farmer—writers. The resurgence of on-farm research in the developing world of the 1970s (Harwood, 1979) was followed by similar emphasis with farmer collaborators in the United States in the 1980s and 1990s. Farmers have been increasingly invited to
serve on steering committees and research grant award committees for sustainable agriculture projects.

Farm family well-being and that of their rural communities has been another major area of merger and inclusion in the sustainability vision. The many links and interdependencies between human and community development and sustainable agriculture are being considered in the structure and design of food systems. In 1986, Dahlberg presented a comprehensive and thought-provoking collection of writings on social, economic, and structural issues. The importance of farm size, community interaction, and the global structure of the food system is seen to be critical to both social and economic well-being. Heffernan, as cited in the Dahlberg book and in more recent publications (Heffernan, 1997), forcefully makes the point that with globalization of capital markets and the resulting centralization of control and ownership (of both input supply and product handling and processing) have come a reduction in market competition, a shift in balance of economic power away from the producer, and a replacement of farm-level production instability with greater macro-economic instability in the marketplace. Much of the current literature on the social dimensions of sustainable agriculture is found in Agriculture and Human Values, the Journal of the Agriculture, Food, and Human Values Society.

Economic dimensions of sustainable agriculture have been typically associated with whole-farm studies comparing organic with "conventional" farms in the late 1970s, exemplified by paired comparison studies in the Midwest (Locheretz et al., 1981). Other, single-farm studies have been reported, such as in the case studies of the NRC's report Alternative Agriculture (1989). Most of these studies have shown organic farms, in most years, to be as profitable as conventional farms. Most of the sustainable agricultural production research of the past two decades has focused on comparisons of crop rotations and use of cover crops and other systems component practices.

There has been a growing crescendo of voices critical of the failure to account for the side effects, the "external costs" of conventional agriculture. These external factors impact communities, the environment, and human health. There is increasing criticism of the emerging structure, which includes farm scale, the patterns of movement of food, and the corporate concentration of input supply and processing on a global scale. A significant component of the sustainable agriculture debate concerns the desirability of some portion of foods being of local origin, with the size of that portion related to location and community development status (Shuman, 1998; Harwood, 1998).

Long-standing schools of thought and practice in agriculture, influenced by changes in science and technology and in food system structure, have thus provided much of the content of the present-day sustainable agriculture agenda or vision. The amalgam of ideas and component factors has provided an extremely rich background from which development direction can be modified. The breadth of the agenda and the level of dissatisfaction with the current system point to a very major underlying problem of the global, monetarily driven process that is directing and fueling current change. Most of the sustainable agricultural debate concerns differences in goals and in ethical and value dimensions between farmers, agriculture as a sector, and national, international, and global interests as pointed out so clearly by Dahlberg (1989). Many of the resources used and managed by agriculture and many of the services and outputs from agriculture that are critical to ecological and human well-being lie outside the monetary process that is currently driving global agriculture. If the marketplace does not put value on those dimensions and the political process either cannot or will not value them, there is a high level of disarray. Much of today's sustainable agriculture debate revolves around these value differences. Many of these sustainability issues are deeply imbedded in the current public debate over genetically modified organisms.

III. INDICATORS OF SUSTAINABILITY

What agricultural practices are sustainable? This is an area of intensive research. Sustainable practices must meet the three criteria defined in Section I: They must be economically viable, environmentally safe, and socially acceptable. There is no single prescription for sustainability; sustainable practices will vary by cropping system, local environment, and socioeconomic system. Nonetheless, emerging research results suggest that locally sustainable systems tend to be more resource conservative than less sustainable systems and tend to rely less on external subsidies and more on internal ecosystem services.

A. Resource Conservation

Resource conservation means that those resources on which sustainability depends are conserved and even enhanced by agronomic management. Soil organic mat-
ter is a good example of an ecosystem resource that is easily reduced without effective management. Soil organic matter declines rapidly in almost all cropping systems following initial cultivation—typically to 40–60% of original values within a few decades. However, soil organic matter is a valuable resource, providing habitat and energy for soil organisms, a soil structure favorable for plant growth and water retention, and a chemical structure favorable for nutrient retention.

The loss of soil organic matter is often associated with a need for greater external inputs. Cropping practices that conserve or enhance soil organic matter buildup will invariably enhance the environmental and often the economic sustainability of cropping systems. Crops grown in high-organic matter soils have a better water and nutrient environment than similar crops grown in soils that are depleted in organic matter, and thus they may require fewer external inputs for the same productivity. Additionally, less soil erosion and lower runoff from high-organic matter soils better protects downstream environments from agronomic impact. Therefore, cropping practices that conserve soil organic matter can be considered more sustainable than those that do not.

Often, however, there are trade-offs that require any specific conservation effort to be evaluated in the overall context of sustainability. For example, conservation tillage typically slows or stops soil organic matter loss and thus can be considered a resource-conserving, sustainable cropping practice. However, tillage controls weeds in cropping systems, and in the absence of tillage weed control is typically left to herbicides, which have environmental and economic costs different from those of tillage. Is the maintenance of soil organic matter as sustainable in light of a more intensive reliance on herbicides? Ideally, such trade-offs can be minimized. For example, winter cover cropping can also reduce soil organic matter loss and additionally can reduce nitrate leaching and suppress weeds, without the need for additional herbicide. Nevertheless, each cropping practice must be evaluated in a whole-system context to adequately evaluate its contribution to a system’s sustainability.

B. Ecosystem Services

Ecosystem services are those services provided by an ecosystem to the organisms inhabiting the system or living nearby. Unmanaged systems provide such services as a matter of course. Farms can likewise provide such services to organisms within the farming system as well as to organisms in the surrounding landscape and to local communities. Services such as pollination, water retention and groundwater recharge, a particular light environment, or food sources—whether provided directly by other organisms or indirectly by their effects on local environmental conditions—are integral to the functioning of healthy ecosystems. In modern cropping systems many services provided by the original ecosystem prior to its conversion to agriculture have been suppressed or ignored in favor of services provided by external inputs. In a nitrogen-poor native ecosystem, biological nitrogen fixation by native legumes such as clover (Trifolium spp.) might be a principal source of fixed nitrogen; modern cropping systems rely almost exclusively on industrially fixed nitrogen provided as inorganic fertilizer. In a native or unmanaged system, insect herbivory is suppressed largely by trophic and structural complexity that enables insect and vertebrate predators to keep plant pests at bay. In most modern systems insect pests are controlled with insecticides, which also kill insect predators. Managing a cropping system with legumes or with greater plant diversity (either within fields or adjacent to field edges) would allow the ecosystem to provide more of the services now provided via external inputs. Legume cover crops can reduce the need for external nitrogen, and greater plant diversity can provide the structural complexity and refugia needed to support predator populations in otherwise monospecific landscapes.

Just as for practices intended to enhance resource conservation, practices established to reintroduce or enhance existing ecosystem services need to be evaluated on the basis of their total net contribution to sustainability. Although nitrogen fixation by legumes can lower the need for fertilizer inputs and benefit soil organic matter buildup as well as provide winter habitat for predaceous insects, there is no evidence that legume-fixed nitrogen is conserved more tightly than fertilizer-derived nitrogen. Thus, there may be no downstream environmental benefit associated with this ecosystem service. Likewise, animal manure produced on-farm and recycled back to the field may be less conserved than fertilizer nitrogen if the manure is added out of sync with plant nutrient demands. Ongoing research is helping to identify ways in which management can add ecosystem services that both enhance resource availability and reduce the environmental costs of agriculture. At the societal scale there is ongoing debate on how to value services provided by farms to their neighboring communities. Should farmers be compensated for managing their land in ways that provide
services to local, regional, and national communities? Such payments occur in some parts of Canada, the United States, and Europe today.

IV. EXAMPLES OF SUSTAINABLE CROPPING SYSTEMS

A. Bush Fallow Rotation Systems

Perhaps the best documented example of a locally sustainable cropping system is the bush-fallow rotation, also known as swidden and slash and burn agriculture, indigenous to many cultures prior to the advent of continuous cropping systems several hundred years ago, and still evident in the humid tropics today. In the absence of population change, the bush-fallow system allows a tract of forest or savanna to provide food with few subsidies other than human labor.

In these systems, a small section of native vegetation is cut and cropped. Crop nutrient needs are met by the decomposition of soil organic matter and perhaps by leguminous crops. Sufficient pest protection is provided by crop rotation, complex crop mixtures, and the proximity of fields to native vegetation. Weed suppression is performed by hand.

Once soil nutrients are depleted to levels that significantly compromise crop productivity, the plot is abandoned to “bush fallow” and another plot is cleared from native vegetation and cropped. Meanwhile, the newly fallowed plot is undergoing secondary succession with attendant recovery of soil nutrient stores. By the time several more plots have been sequentially cut, cropped, and fallowed, the original plot will have recovered much of its original fertility and be ready to be cleared and cropped again.

Such agronomic systems are sustainable indefinitely as long as each cropped area is allowed sufficient time to recover its original fertility. However, when land becomes scarce because of development or population growth or both, the system can quickly fail. Native vegetation brought out of fallow too quickly will provide soil fertility for only a portion of the former cropping periods, so the crop portion of the rotation will either be shorter or yield less, forcing more of the native vegetation to be brought out of fallow earlier than planned in order to feed a growing population. Eventually, little native fallow will remain and crops will be grown continuously on soils that now lack much of their former fertility. One of today’s greatest agronomic challenges is providing nitrogen and phosphorus to cropping systems that until recently have been in bush-fallow rotations, especially in sub-Saharan Africa where fertilizer is largely unavailable and most food is grown on small holdings of a few hectares. The maintenance of soil quality and adequate levels of soil organic matter to provide it are major concerns of tropical agronomists.

B. Mixed Farming Systems

The successor to simple bush-fallow systems is mixed farming systems that have several production enterprises of different herbaceous crops, trees, animals, or combinations of crops and animals. In less developed or unstable economies requiring a high level of local community, and farm family self-reliance, the production of a wide array of goods was primarily to meet family and local market needs for an ensured, year-round supply of food, fuel, and building materials. Farm and landscape-level diversity optimized stability within local environments and increased the resiliency of the system to a wide variety of disturbances. The diversity of land use provided a wide range of ecosystem services, including precipitation management, groundwater recharge, wildlife habitat, an environment usually conducive to adequate pest–predator balance, and some mitigation of harsh climatic conditions. The mixed plant community provided shade, wind protection, privacy, and many other, often seasonal, assorted products and services. This range of outputs has recently been termed the multifunctional character of agricultural land (Food and Agricultural Organization, 1999) shown in Fig. 1.

As infrastructure and markets develop, the need for a broad range of products and services decreases. When the costs of adverse environmental impacts such as groundwater contamination are not internalized, or when farmers are not rewarded for ecosystem services that their farms provide to the community or region, they do not include such values in their farm enterprise unless they are motivated and willing to make an altruistic contribution. Many farmers, in fact, do this now, but ultimately the more narrowly focused economic marketplace rules. Today’s farms in highly developed economies frequently have a level of product and land use specialization that is well below an acceptable standard for long-term environmental and resource sustainability (Fig. 1B). In other words, the production base, the environment, and its ecosystem have not been stabilized and are being degraded. With continued market evolution, farmers may be increasingly compensated for the full range of ecosystem services as well as actual product output that they provide (Shuman, 1998; Soule and Piper, 1992) (Fig. 1C).

A more immediate incentive is to add crops and/
An increasing amount of information on the efficiencies of specific technologies for integration is becoming available in the scientific literature. The reduction in input requirements is often a key part. There is less direct research information on the relationship of many of these practices on environmental loading. An exception is the wealth of data on reduced soil erosion as a result of reduced or zero tillage. Currently, the predictive models of loss of pesticides, nutrients, or crop or animal residues are rudimentary. Direct measurements of loss from alternative rotations and use of cover crops are very difficult, expensive, and location specific. These rotation and cover crop practices are widely acknowledged as being fundamental to sustainability. Their efficiencies are being quantified with respect to yield, input reduction, and soil quality and the prevention of soil loss. Michigan data show, for example, that wheat in rotation loses less than 20 kg N/ha per year via groundwater leaching. Well-fertilized continuous corn averages 50 kg N/ha per year. Most U.S. farmers use at least a two-crop rotation.

Animal integration in crop systems is declining in the United States. Poultry and turkeys are increasingly produced in specialized production facilities not located on the farms where their feed is produced. They are usually located in areas where agricultural land is available for manure application, often on a contract "disposal" basis. The level of crop or animal diversity that is appropriate on a farm to balance the market forces for specialization with the need for biological efficiency and ecosystem maintenance is very situation specific. As enterprise integration increases with an effective level of appropriate enterprises and their effective management (Fig. 2, technology T.), agricultural output can be maintained at a much higher level for a given amount of ecosystem disturbance. In other words, sustainable agriculture can maintain productivity at a much lower level of ecosystem disturbance. Very large-scale operations tend to have less diversity, in part because of the greater difficulty of managing diverse enterprises. Crop and animal management requires numerous and often frequent decisions to be made as conditions change that are often stimulated by visual, difficult to measure changes. The frequent presence and sensitivity of the manager, the experience in production management, and the ability to make decisions place limits on the scale of highly diversified operations. Every farm owner experiences this tension.

On a global scale, under conditions of high population density, small farms, and the need for producing a wide array of products in often marginal production environments, a very diverse type of farm enterprise
in only a few in which large tracts of land remain is there cattle only on farms. In the humid tropics the mixture of crops, trees, and animals (agrisilvo-pastoral systems) represent the great majority of farms. Where human population is relatively high (>300–500 persons per square kilometer) in rural areas, if there is poverty combined with modest levels of rainfall (less than 1000 mm per year) and/or cool temperatures for part of the year, fuel for cooking and heating becomes a problem. Resource degradation and loss of production potential often occur as the standing stocks of carbon (particularly in trees) and eventually the soil carbon stocks are reduced as crop and animal residues are burned. The system rapidly loses crop nutrient holding and recycling capacity, and its ability to intercept and retain rainfall decreases.

The immigrant Japanese farms of the Brazilian area are an interesting example of high diversity, maintenance of a high-standing stock of carbon, and extremely high market value on a small land area (Fig. 4). This is similar in many ways to Indonesian agroforestry systems in Central Sumatra. Most developing country mixed farms have a larger portion of cash grain crops.

V. CONCLUSIONS

Our global food and fiber production systems are undergoing an enormous transformation, driven by rapid advances in the sciences of biotechnology, engineering, and food processing and chemistry. The increasing centralization in the manufacture of agricultural inputs and in the collection and processing of food is causing huge economic and social change. Many social and political values are being challenged. The global marketplace forces for product uniformity and the geographical concentration of its production are driving farmers toward a level of specialization that results in their farms having much less diversity of crop and animal enterprises on the landscape than that desirable for the maintenance of many ecosystem services. Markets have not yet matured to adequately value these services, especially those that affect environmental quality, nor in most places have governments established disincentives in order to protect them. The greatest challenge to sustainability, with its many economic, environmental, and social dimensions, is the lack of public awareness, vision, and will to implement necessary changes. In some cases, research is needed to clarify the value of alternative strategies and to provide additional options for sustainable management.
For the present, and in the aggregate, the world has enough food. In many if not most cases it is not being produced sustainably, i.e., in a manner that is economically viable, environmentally benign, and socially acceptable to many who are affected by its production. On the other hand, new research is showing that sustainable cropping systems can be designed to operate effectively, using ecological knowledge to substitute for some of the management options now provided by external inputs, and in a way that has a less adverse environmental and social impact than conventional management.

See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • AGRICULTURE, TRADITIONAL • BIODIVERSITY AS A COMMODITY • ECOSYSTEM SERVICES, CONCEPT OF • HERBICIDES • NITROGEN AND NITROGEN CYCLE • PESTICIDES, USE AND EFFECTS OF • SOIL CONSERVATION

Bibliography


I. Introduction

II. Biodiversity Features of Traditional Agroecosystems

III. The Complex Nature of Traditional Farmers’ Knowledge

IV. The Ecological Services of Biodiversity in Traditional Agroecosystems

V. Preserving the Biodiversity of Traditional Agroecosystems

VI. Using Biodiversity-Based Strategies to Support Traditional Agriculture

VII. Conclusions

GLOSSARY

agroecosystem A simplified natural ecosystem subjected to exploitation for purposes of food and fiber production.

biodiversity Diversity of microbial, animal, and plant species in an ecosystem that performs distinct ecological functions and services.

ethnoecology Study of the various forms of traditional environmental knowledge that are characteristic of specific ethnic groups and that translate into natural resource management.

polyculture Intensive growing of two or more crops either simultaneously or in sequence on the same piece of land.

sustainable agriculture Form of agriculture that is environmentally sound, culturally sensitive, socially acceptable, and economically viable.

traditional agriculture Indigenous form of ecologically based agriculture resulting from the coevolution of local cultural and environmental systems.

TRADITIONAL AGRICULTURE IS AN INDIGENOUS FORM OF FARMING that results from the coevolution of local cultural and environmental systems. It exhibits a high level of ecological rationale expressed through the intensive use of local knowledge and natural resources, including the management of agrobiodiversity in the form of diversified agricultural systems.

I. INTRODUCTION

One of the salient features of traditional farming systems throughout the developing world is their high degree of biodiversity. These traditional systems have emerged over centuries of cultural and biological evolution and represent the accumulated experiences of indigenous farmers interacting with the environment without access to external inputs, capital, or modern scientific knowledge (Chang, 1977; Grigg, 1974). Using inventive self-reliance, experiential knowledge, and locally available resources, traditional farmers have often developed farming systems that generate sustained yields (Har-
wood, 1979). In Latin America alone, more than two and a half million hectares are under traditional agriculture in the form of raised fields, polycultures, and agroforestry systems, documenting the successful adaptation of these farming practices to difficult environments (Altieri, 1991).

Many of these traditional agroecosystems, still found throughout the Andes, Meso-America, and the lowland tropics, constitute major in situ repositories of both crop and wild plant germplasms. From an agroecological perspective, these agroecosystems can be seen as a continuum of integrated farming units and natural or semi-natural ecosystems where plant gathering and crop production are actively pursued. Plant resources are directly dependent on management by human groups; thus, both species and genetic diversity have evolved in part under the influence of farming practices shaped by particular cultures and the forms of sophisticated knowledge they represent (Nabhan, 1983).

Perhaps the key to understanding how traditional farmers maintain, preserve, and manage biodiversity is to recognize the complexity of their production systems. Today, it is widely accepted that indigenous knowledge is a powerful resource in its own right and complementary to knowledge available from Western scientific sources. Therefore, in studying such systems, it is not possible to separate the study of agricultural biodiversity from the study of the culture that nurtures it.

This article explains the features of the biodiversity inherent in traditional agroecosystems, and the ways in which farmers apply local knowledge to manage such biodiversity to satisfy subsistence needs and to obtain ecological services. Traditional agriculture is rapidly disappearing in the face of major social, political, and economic changes. The conservation and management of this agrobiodiversity will be possible only if they are linked to the preservation of the cultural diversity and economic viability of the local farming populations.

II. BIODIVERSITY FEATURES OF TRADITIONAL AGROECOSYSTEMS

Traditional farming systems commonly support a high degree of plant diversity in the form of polycultures and/or agroforestry patterns (Chang, 1977; Clawson, 1985). This strategy of minimizing risk by planting several species and varieties of crops stabilizes yields over the long term, promotes diet diversity, and maximizes returns even with low levels of technology and limited resources (Harwood, 1979). Such biodiverse farms are endowed with nutrient-enriching plants, insect predators, pollinators, nitrogen-fixing and nitrogen-decomposing bacteria, and a variety of other organisms that perform various beneficial ecological functions.

Traditional multiple-cropping systems provide as much as 15–20 percent of the world food supply (Francis, 1986). Polycultures constitute at least 80 percent of the cultivated area in West Africa and predominate in other parts of Africa as well (Norman, 1979). At the same time, much of the production of staple crops in the Latin American tropics occurs in polycultures. More than 40 percent of the cassava, 60 percent of the maize, and 80 percent of the beans in the region grow in mixtures with each other or other crops (Francis, 1986). Polycultures are also very common in parts of Asia where upland rice, sorghum, millet, maize, and irrigated wheat are the staple crops. Lowland (flooded) rice is generally grown as a monoculture, but in some areas of Southeast Asia farmers build raised beds to produce dryland crops amid strips of rice (Beets, 1982).

Tropical agroecosystems composed of agricultural and fallow fields, complex home gardens, and agroforestry plots commonly contain well over 100 plant species per field, and these are used as construction materials, firewood, tools, medicines, livestock feed, and human food. Examples include multiple-use agroforestry systems managed by the Huastecs and Lacandones in Mexico, the Bora and Kayapo Indians in the Amazon River basin, and many other ethnic groups who incorporate trees into their production systems (Wilken, 1977).

In the Latin American tropics, home gardens are a highly efficient form of land use, incorporating a variety of crops with different growth habits. The result is a structure similar to that of tropical forests, with diverse species and a layered physical configuration (Denevan et al., 1984). In Mexico, for example, Huastec Indians manage a number of fields, gardens, and forest plots that may harbor a total of about 300 species. Small areas around their houses commonly average 80–125 useful plant species, mostly native and medicinal plants. Huastec management of the noncrop vegetation in these complex farm systems has influenced the evolution of individual plants and the distribution and composition of the crop and noncrop communities.

In these “forestlike” agricultural systems, nutrient cycles are tight and closed. In traditional shaded coffee plantations (where Inga and Erythrina are common tree
species), the total nitrogen input from the decomposition of shade tree leaves, as well as from litter and symbiotic fixation, can be well over ten times higher than the net nitrogen output in the coffee harvest, which usually averages 20 kg/ha/yr. Clearly, the system amply compensates for the nitrogen loss by harvest with a subsidy from the shade trees. In Mexico, farmers encourage the growth of native leguminous trees in cultivated fields (Wilken, 1977). From Puebla and Tehuacán south through Oaxaca, farms with light to moderately dense stands of mesquite (Prosopis spp.), guaje (Leucaena esculenta), and guamuchil (Pithecellobium spp.) are a familiar sight. Stand density varies from fields with only a few trees to virtual forests with crops planted beneath them. A slightly different practice is found near Ostuncalco, Guatemala, where rigorously pruned sauco (Sambucus mexicana) stumps dot maize and potato fields. Saucos leaves and small branches are removed annually, scattered around individual crop plants, and then chopped and interred with broad hoes. Local farmers claim that crop quality and yields in the sandy volcanic soils of this region depend on the annual application of this method (Wilken, 1977).

Many traditional agroecosystems are located in centers of crop diversity, and thus contain populations of variable and adapted landraces as well as wild and weedy relatives of crops (Harlan, 1976). Clawson (1986) described several systems in which tropical farmers plant multiple varieties of each crop; this practice supports both intraspecific and interspecific diversity, and also enhances harvest security. For example, in the Andes, farmers cultivate as many as 50 potato varieties in their fields (Brush, 1982). Similarly, in Thailand and Indonesia, farmers maintain a diversity of rice varieties adapted to a wide range of environmental conditions, and they regularly exchange seeds with each other (Grigg, 1974). The resulting genetic diversity heightens resistance to diseases that attack particular strains of the crop and enables farmers to exploit different microclimates and to derive multiple nutritional and other uses from the genetic variation among the species.

Many plants within or around traditional cropping systems are wild or weedy relatives of crop plants. In fact, many farmers “sponsor” certain weeds in or around their fields that may have positive effects on soil and crops, or that serve as food, medicines, ceremonial items, teas, soil improvers, or pest repellents. In the Mexican Sierra, the Tarahumaras Indians depend on edible weed seedlings or “quelites” (e.g., Amaranthus, Chenopodium, Brassica) in the early season from April through July, a critical period before crops mature from August through October. Weeds also serve as alternative food supplies in seasons when maize or other crops are destroyed by frequent hail storms (Ibey, 81). In barley fields, it is common for Tlaxcalan farmers to maintain Solanum mezzianum at levels up to 4500 plants/ha; this yields about 1300 kg of fruit, a significant contribution to agricultural subsistence (Altieri and Trujillo, 1987).

Farmers also derive other benefits from weeds, such as increased gene flow between crops and their relatives. In Mexico, when the wind pollinates maize, natural crosses occur with wild teosinte growing in the field borders, resulting in hybrid plants. Certain weeds are used directly to enhance the biological control of insect pests, as many flowering weeds attract predators and parasites of pests to their pollen and nectar. Other farmers allow weeds such as goosegrass (Eleusine indica) in bean fields to repel Empoasca leafhoppers, or wild Lupinus as a trap plant for the pestiferous scarab beetle (Macrotyloma sp.), which otherwise would attack corn (Altieri, 1993).

However, diversity is maintained not only within a cultivated area. Many farmers maintain natural vegetation adjacent to their fields, and thus obtain a significant portion of their subsistence requirements through gathering, fishing, and hunting in habitats that surround their agricultural plots. For the Purhepecha Indians who live around Lake Pátzcuaro in Mexico, gathering is part of a complex subsistence pattern that is based on multiple uses of their natural resources. These people use at least 224 species of native and naturalized vascular plants for dietary, medicinal, household, and fuel needs (Caballero and Mapes, 1985).

Depending on the level of biodiversity of closely adjacent ecosystems, farmers accrue a variety of ecological services from surrounding natural vegetation. For example, in western Guatemala, the indigenous flora of the higher-elevation forests provide valuable native plants that serve as a source of organic matter to fertilize marginal soils, for each year farmers collect leaf litter from nearby forests and spread it over intensively cropped vegetable plots to improve tilth and water retention. Some farmers may apply as much as 40 metric tons of litter per hectare each year; rough calculations indicate that a hectare of cropped land requires the litter production of 10 ha of regularly harvested forest (Wilken, 1977).

Clearly, traditional agricultural production commonly encompasses the multiple uses of both natural and artificial ecosystems, where crop production plots and adjacent habitats are often integrated into a single agroecosystem.
III. THE COMPLEX NATURE OF TRADITIONAL FARMERS’ KNOWLEDGE

Ethnoecology is the study of the natural world knowledge systems of indigenous ethnic people. This knowledge has many dimensions, including linguistics, botany, zoology, craft skills, and agriculture, and is derived from the direct interaction between humans and their environment. In such a system, cognition and perception select the most adaptive and useful environmental information, and this “successful” knowledge is preserved from generation to generation through oral or experimental means. Indigenous peoples’ knowledge about soils, climates, vegetation, animals, and ecosystems usually results in multidimensional productive strategies (i.e., the use of multiple ecosystems with multiple species), and these strategies generate (within certain ecological and technical limits) the food self-sufficiency of farmers in a region (Netting, 1993).

Captivated by the ecological intricacies of these traditional agricultural systems, many scientists are now beginning to show interest in them. As scientists search for ways to remedy the deficiencies of modern agriculture, they recognize that indigenous farmers’ knowledge may hold vital information for the future of world agriculture. After centuries of cultural and biological evolution, these farmers have developed locally adapted, complex farming systems that have helped them to sustainably manage a variety of environments and to meet their subsistence needs, without depending on modern agricultural technologies.

For many agricultural scientists, four aspects of these traditional knowledge systems are relevant: knowledge of the environment, folk taxonomies, knowledge of farming practices, and the experimental nature of traditional knowledge (Altieri, 1987).

A. Knowledge of the Environment

Indigenous knowledge about the physical environment is often very detailed. Many farmers have developed traditional calendars to control the scheduling of agricultural activities, and many sow according to the phase of the moon, believing that there are lunar phases of rainfall. They also cope with climatic seasonality by utilizing weather indicators based on the phenologies of local vegetation.

Soil types, degrees of soil fertility, and land-use categories are also discriminated in detail. Soil types are commonly distinguished by color, texture, and sometimes taste. Shifting cultivators usually classify their soils based on vegetation cover. In general, peasants identify soil types based on the nature of the peasant’s relationship to the land (Williams and Ortiz-Solario, 1981). Aztec soil classification systems were very complex, recognizing more than two dozen soil types identified by origin, color, texture, smell, consistency, and organic content. These soils were also ranked according to their agricultural potential, which was used in both land-value evaluations and rural census. Today, Andean peasants in Coporaque, Peru, recognize four main soil classes, where each class has specific characteristics matching the most adequate cropping system (Brush, 1982).

B. Biological Folk Taxonomies

Many complex knowledge systems that are used by indigenous people to group together plants and animals have been well documented (Berlin et al., 1973). The traditional name of a plant or animal usually reveals that organism’s taxonomic status, and researchers have found that, in general, there is a good correlation between folk taxa and scientific taxa.

The classification of animals, especially insects and birds, is widespread among indigenous farmers. Insects and related arthropods have major roles as crop pests, as causes of disease, as food, and as medicinal products, in addition to their importance in local myth and folklore. In many regions of the world, agricultural pests are tolerated because they also constitute agricultural products; that is, traditional agriculturalists may consume plants and animals that would otherwise be considered pests (Brokensha et al., 1980).

Ethnobotanies are the most commonly documented folk taxonomies (Alcorn, 1984). The ethnobotanical knowledge of certain campesinos in Mexico is so elaborate that the Tzeltal, Purepecha, and Yucatan Mayans can recognize more than 1200, 900, and 500 plant species, respectively (Toledo et al., 1985). Similarly, !Ko bushwomen in Botswana were able to identify 206 out of 211 plants collected by researchers (Chambers, 1983), while Hanunoo swidden cultivators in the Philippines could distinguish over 1600 plant species (Grigg, 1974).

C. Knowledge of Farming Practices

As more scientific research is conducted, many of the traditional farming practices once regarded as primitive or misguided are being recognized as sophisticated and appropriate. For example, when confronted with specific problems of slope, flooding, drought, pests, dis-
In general, traditional agriculturalists have adjusted to environmental constraints by concentrating on a few characteristics and processes that incorporate the following structural and functional elements (Chesman, 1998; Altieri and Anderson, 1986):

a. They combine high species numbers and structural diversity in time and space (through both vertical and horizontal organization of crops).

b. They exploit the full range of microenvironments (which differ in soil, water, temperature, altitude, slope, fertility, etc.) within a field or region.

c. They maintain closed cycles of materials and wastes through effective recycling practices.

d. They rely on the complexity of biological interdependencies, resulting in some degree of biological pest suppression.

e. They rely on local resources plus human and animal energy, thereby using low levels of technology input.

f. They rely on local varieties of crops and incorporate the use of wild plants and animals. Production is usually for local consumption. The level of income is low; thus the influence of noneconomic factors on decision making is substantial.

The Experimental Nature of Traditional Knowledge

The strength of traditional people's knowledge is that it is based not only on acute observation but also on trial and error and experimental learning. The experimental approach is very apparent in the selection of seed varieties for specific environments, but it is also implicit in the testing of new cultivation methods to overcome particular biological or socioeconomic constraints. In fact, Chambers (1983) argued that farmers often achieve a richness of observation and a fineness of discrimination that would be accessible to Western scientists only through long and detailed measurement and computation.

Yet only recently has some of this traditional knowledge been described and documented by researchers. The evidence suggests that the finest discrimination develops in communities where the environments have great physical and biological diversity and/or in communities living near the margins of survival (Chambers, 1983). Also, older community members possess greater, more detailed knowledge than younger members (Klee, 1980).

IV. THE ECOLOGICAL SERVICES OF BIODIVERSITY IN TRADITIONAL AGROECOSYSTEMS

In traditional agroecosystems, complex and diversified cropping systems are vital because the interactions among crops, animals, insects, and trees result in beneficial synergisms that optimize soil fertility, pest control, and productivity (Altieri, 1995; Harwood, 1979; Richards, 1983). Among the ecological services are the following:

1. By interplanting, farmers take advantage of the capacity of cropping systems to reuse their own stored nutrients. The tendency of some crops to deplete the soil is counteracted by interplanting other crops that enrich the soil with organic matter. Soil nitrogen, for example, can be increased by incorporating legumes in the crop mixture, and phosphorus assimilation can be enhanced by growing crops with mycorrhizal associations.

2. The complex structure of traditional agroecosystems minimizes crop loss to insect pests through a variety of biological mechanisms. The intercropping of diverse plant species provides habitats for the natural enemies of insect pests as well as alternative host plants for pests. For example, a crop may be planted as a diversionary host to protect other more susceptible or more economically valuable crops from serious damage. The diversity of crops grown simultaneously in polycultures helps prevent the buildup of pests on the comparatively isolated plants of each species. Where shifting cultivation is practiced, the clearing of small plots from secondary forest vegetation also permits the easy migration of natural pest predators from the surrounding forest.

3. Increasing the species and/or genetic diversity of cropping systems is a key strategy to minimize losses from plant diseases and nematodes (types of roundworms that are among the most widespread and damaging of agricultural pests). The mixing of different crop species or varieties can delay the onset of diseases, reduce the spread of disease-carrying spores, and modify environmental conditions such as humidity, light, temperature, and air movement so that they are less favorable to the spread of certain diseases.
Many intercropping systems prevent competition from weeds, chiefly because the large leaf areas of their complex canopies prevent sufficient sunlight from reaching sensitive weed species. In general, the extent to which weeds present a problem depends on the type of crops and the proportion of the different species grown, their density, where they are planted, the fertility of the soil, and management practices. Weed suppression can be enhanced in intercrop systems by adding crop species that inhibit weed germination or growth. Crops such as rye, barley, wheat, tobacco, and oats release toxic substances into the environment, either through their roots or from decaying plant material. Such toxins inhibit the germination and growth of some weed species such as wild mustard (Brassica spp.) and poppy.

The integration of animals (e.g., cattle, swine, poultry) into farming systems, in addition to using them for milk, meat, and draft needs, adds another trophic level to the system, making it even more complex. Animals are fed crop residues and weeds with little negative impact on crop productivity, and this serves to turn otherwise unusable biomass into animal protein. Animals also recycle the nutrient content of plants by transforming them into manure. Furthermore, the need for animal feed broadens the crop base to include plant species that are useful for conserving soil and water (Beijbesjö et al., 1982). Legumes are often planted to provide quality forage, but they also improve the nitrogen content of soils. Integrated crop–livestock systems usually take the form of a crop–pasture rotation in which the pasture phase “charges” the system with nutrients and organic matter and the cropping phase “extracts” the accumulated nutrients. This balances biomass and nutrient inputs and outputs.

V. PRESERVING THE BIODIVERSITY OF TRADITIONAL AGROECOSYSTEMS

As many rural societies undergo the conversion from a subsistence economy to a cash agricultural economy, the loss of biodiversity in their ecosystems is mounting at an alarming rate. Because many peasants are directly linked to the market economy, external economic forces are increasingly influencing production by favoring genetically uniform crops and mechanized and/or agrochemical practices. Many landraces and wild plant relatives are being abandoned, which may cause them to become relic populations or even extinct. In some areas, land scarcity (mostly a result of uneven land distribution) has forced changes in land use and agricultural practices, which in turn have caused the disappearance of habitats that formerly maintained useful noncrop vegetation, including wild progenitors and weedy forms of crops (Altieri et al., 1987).

In many parts of the world, genetic erosion is occurring at a fast pace because farmers are having to quickly change their farming systems because of economic, technical, and social pressures. As farmers adopt high-yield modern varieties (HYVs), they often subdivide their farming systems into commercial (mostly devoted to HYVs) and subsistence sectors, growing native varieties in the latter. The greatest loss of traditional plant varieties is occurring in lowland valleys close to urban centers and markets (Brush, 1986).

Given these destructive trends, many scientists and development workers have emphasized the need for in situ conservation of native crop genetic resources and the environments in which they occur (Prescott-Allen and Prescott-Allen, 1981). However, most researchers believe that in situ preservation of landraces would require a return to or the preservation of microcosms of traditional agricultural systems, which some regard as an unacceptable and impracticable proposition (Franke and Soulé, 1981). Nevertheless, the maintenance of traditional agroecosystems may be the only sensible strategy to preserve in situ repositories of crop germplasm. Although most traditional agroecosystems are undergoing some process of modernization or drastic modification, the conservation of crop genetic resources can still be integrated into agricultural development, especially in regions where rural development projects can still be integrated into agricultural development, especially in regions where rural development projects preserve the vegetation diversity of traditional agroecosystems and are anchored in the peasant rationale to utilize local resources and their intimate knowledge of the environment (Akcorn, 1984; Nabhan, 1983).

Previous recommendations for in situ conservation of crop germplasm emphasized the development of a system of village-level landrace custodians (a farmer curator system) whose purpose would be to continue growing a limited sample of endangered landraces native to the region (Mooney, 1983). One suggestion for preserving crop-plant diversity was for governments to set aside carefully chosen 5- by-20-km strips of land at as few as 100 sites around the world where native agriculture is still practiced (Wilkes and Wilkes, 1972). But given the increasing impoverishment and lack of income-generating alternatives for many rural populations in less developed countries, a proposition of this kind is clearly unrealistic since it fails to address the subsistence needs of these populations. In many areas where the urgent short-term goal of the local people is...
survival, diverting the limited land available for conservation purposes per se might prove totally inappropriate. A more feasible approach would be to support sustainable farming systems that incorporate native crops and wild/weedy relatives within and around production fields, as well as appropriate technologies aimed at upgrading food production for self-sufficiency (Altieri and Merrick, 1987). Such efforts would ensure that germplasm preservation remains linked to the economic and agricultural viability of local populations.

If biodiversity conservation is to succeed among small farmers, conservation goals and rural development efforts must be integrated to give equal importance to local resource conservation, food self-sufficiency, and equitable market participation. Any attempt at in situ crop genetic conservation must struggle to preserve the agroecosystem in which these resources occur (Nabhan, 1983). In the same vein, preservation of traditional agroecosystems cannot be achieved unless the sociocultural stability of the local community is also assured (Altieri, 1995).

An examination of effective grassroots rural development programs in less developed countries suggests that the process of agricultural improvement must (a) utilize and promote autochthonous knowledge and resource-efficient technologies, (b) emphasize use of local and indigenous resources, including valuable crop germplasm as well as essentials like firewood resources and medicinal plants, and (c) remain a self-contained, village-based effort with the active participation of the local people (Altieri, 1987). The subsidizing of a peasant agricultural system with external resources (e.g., pesticides, fertilizers, and irrigation water) can bring high levels of productivity, but such a system would then be sustainable only at high external cost and would depend on the uninterrupted availability of commercial inputs. In contrast, an agricultural strategy based on a diversity of plants and cropping systems can bring moderate to high levels of productivity through the manipulation and exploitation of the resources internal to the farm and can be sustainable at a much lower cost and for a longer period of time (Gliessman, 1998).

VI. USING BIODIVERSITY-BASED STRATEGIES TO SUPPORT TRADITIONAL AGRICULTURE

By understanding the common features of traditional agriculture, such as the capacity to bear risk, the use of biological folk taxonomies, and the production efficiencies derived from multiple and symbiotic crop mixtures, agricultural scientists have been able to develop technologies that support the needs and circumstances of specific groups. While subsistence farming generally lacks the potential for producing a meaningful marketable surplus, it does ensure food security. Many scientists wrongly believe that traditional systems do not produce more because hand tools and draft animals put a ceiling on productivity. However, where productivity is low, the cause appears to be social, not technical. When the subsistence farmer succeeds in providing food, there is no pressure to innovate or to enhance yields. Yet research shows that increased productivity is possible when traditional crop and animal combinations are adjusted and when labor and local resources are used more efficiently (Pretty, 1995).

As the inability of the Green Revolution to improve production and farm incomes for the very poor became apparent, growing enthusiasm for established, traditional agricultural practices generated a renewed quest in the developing world for affordable, productive, and ecologically sound technologies that could enhance small farm productivity while conserving resources. In the Andean altiplano, development workers and farmers have reconstructed a 3000-year-old indigenous farming system at an altitude of almost 4000 m. These indigenous farmers were able to produce food in the face of floods, droughts, and severe frosts by growing crops such as potatoes, quinoa, oca, and amaranthus in raised fields or “waru-warus,” which consisted of platforms of soil surrounded by ditches filled with water (Browder, 1989).

Technicians have now assisted local farmers in reconstructing 10 ha of these ancient farms, with encouraging results, which later led to a substantial expansion of the area under warus. For instance, yields of potatoes from waru-warus can surpass yields from chemically fertilized fields. Recent measurements indicate that waru-warus produce 10 tons of potatoes per hectare compared to the regional average of 1–4 tons/ha.

This combination of raised beds and canals has proven to have remarkably sophisticated environmental effects. During droughts, moisture from the canals slowly ascends the crop roots by capillary action, and during floods, furrows drain away excess runoff. Waru-warus also reduce the impact of temperature extremes. Water in the canal absorbs the sun’s heat by day and radiates it back by night, thereby helping protect crops from frost. On the raised beds, nighttime temperatures may be several degrees higher than in the surrounding area. The system also maintains its own soil fertility. In the canals, silt, sediment, algae, and organic residues
Advantages of using terraces are that they minimize risks in times of frost or drought, reduce soil loss, amplify the cropping options because of microclimate and hydraulic differences, and thus improve crop yields. Yield data from new bench terraces showed a 43–65 percent yield increase in potatoes, maize, and barley compared to yields of those crops grown on sloping fields. One of the main constraints of this technology is its high labor intensity, requiring about 350–300 worker-days per hectare for the initial building of the terraces. Such demands, however, can be buffered when communities organize and share tasks (Browder, 1989).

Another example of how a biodiversity-based approach can support or even resurrect traditional agriculture is occurring on Chiloe Island in southern Chile. This is a secondary center of origin of potatoes, and development workers are currently tapping the ethnobotanical knowledge of elderly female Huilliche Indians in an effort to slow genetic erosion and to recover some of the original native potato germplasm. They intend to provide impoverished farmers with locally adapted varieties that can produce without the use of agrochemical fertilizers. After surveying several agroecosystems on Chiloe, technicians collected hundreds of samples of native potatoes still grown by local farmers, and with this material, and in collaboration with farmers, they established community seed banks where more than 120 traditional varieties are grown year after year and are subjected to selection and seed enhancement. In this way, an in situ conservation program has been initiated involving farmers from various rural communities, thus ensuring the active exchange of varieties among participating farmers. As more farmers become involved, this strategy will provide a continuous supply of seeds to resource-poor farmers and will also create a repository of vital genetic diversity for future regional crop improvement programs (Altieri, 1995).

VII. CONCLUSIONS

A key conclusion that emerges from the relevant anthropological and ecological literature is that, when not disrupted by economic or political forces, indigenous modes of food production generally preserve rather than destroy biodiversity and natural resources. In fact, in any particular region, capitalist development through the promotion of large-scale, energy-intensive, commercial agriculture is bound to deplete natural resources more than some of the existing traditional systems. A number of studies have proven that many traditional agricultural systems are highly sustainable and productive, offering an alternative to the capital-
intensive agriculture currently promoted by many development and governmental agencies. Besides employing crop diversity, traditional farmers use a set of practices that often cause minimal land degradation. These include the use of terraces and hedgerows in sloping areas, minimal tillage, mulching, small field sizes, and long fallow cycles (Grigg, 1974; Brush, 1982; Richards, 1985; Netting, 1993). It is clear that this more traditional strategy is both ecologically informed and environmentally sound, as the agricultural practices that are most likely to endure are those that deviate least from the native diversity of the natural plant communities within which they exist (Altieri, 1993; Gleissman, 1988).

This assessment of traditional subsistence agriculture does not romanticize its origins or practitioners, nor does it consider development per se to be detrimental. The intention is rather to stress the demonstrated value of traditional agriculture in the preservation of biodiversity, native crop diversity, and the adjacent vegetation communities (Toledo, 1980). Basing a rural development strategy on traditional farming and ethnobotanical knowledge not only assures the continual use and maintenance of valuable genetic resources, but also allows for the diversification of peasant or other indigenous subsistence strategies (Alcorn, 1984; Caballero and Mapes, 1985), which is a crucial issue in times of economic uncertainty.

The study of traditional agroecosystems and the ways in which indigenous peoples maintain and use biodiversity can facilitate the discovery of valuable agroecological principles, which in turn can contribute to the development of more sustainable agroecosystems and biodiversity conservation strategies in both developed and less developed countries.

See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • AGRICULTURE, SUSTAINABLE • ETHNOBIOLOGY AND ETHNOECOLOGY • GRAZING, EFFECTS OF • INDIGENOUS PEOPLES, BIODIVERSITY AND • LAND-USE PATTERNS, HISTORIC • TRADITIONAL CONSERVATION PRACTICES

Bibliography


Principles, which in turn can contribute to the development strategy on traditional farming and ethnobotanical knowledge not only assures the continual use and maintenance of valuable genetic resources, but also allows for the diversification of peasant or other indigenous subsistence strategies (Alcorn, 1984; Caballero and Mapes, 1985), which is a crucial issue in times of economic uncertainty.

The study of traditional agroecosystems and the ways in which indigenous peoples maintain and use biodiversity can facilitate the discovery of valuable agroecological principles, which in turn can contribute to the development of more sustainable agroecosystems and biodiversity conservation strategies in both developed and less developed countries.
of Crop Genetic Resources: A Report to the International Board for Plant Genetic Resources. IBPGR, Rome.


I. Sources and Environmental Pathways of Air Pollutants
II. Mechanisms of Impacts on Biodiversity
III. Evidence of Impacts on Biodiversity
IV. Air Quality Guidelines to Protect Biodiversity
V. Concluding Remarks

GLOSSARY

acute toxicity A damaging effect caused by a single short period of exposure to high concentrations of a pollutant.
available (or bioavailable) concentration The concentration of a contaminant in soils or surface waters which can be taken up by the target organism.
background concentration The concentration of a contaminant in an environment which has not been measurably influenced by anthropogenic sources.
chronic toxicity A damaging effect caused by a long period of exposure to low or moderate concentrations of a pollutant.
contamination The presence of elevated concentrations of a toxic substance, compared with normal ambient concentrations.
deposition The rate of influx of a substance from the atmosphere, usually expressed as mass per unit area of ground.
detoxification Reduction in the toxic effect of a pollutant by its chemical or biological transformation to another less toxic chemical.
emission The release of a contaminant to the environment.
photochemical Chemical reactions with rates which are increased by radiation of particular wavelengths.
phytotoxic Damaging to plants.
pollution The presence of a toxic substance at a concentration that can cause adverse effects.
wet deposition The influx of contaminants from the atmosphere in precipitation (i.e., rain, snow, mist, fog, and clouds).

AIR POLLUTION results from a range of sources, including industry, power generation, transport, and domestic sources. Its local impacts on biodiversity can be dramatic close to large point sources of emissions, but significant impacts on biodiversity over much wider areas can also result from the long-range transport of pollutants. This article provides a brief overview of the mechanisms of air pollution impacts on biodiversity and of the evidence of such impacts in the field. It also considers the issues connected with establishing credible air quality guidelines to protect biodiversity.
I. SOURCES AND ENVIRONMENTAL PATHWAYS OF AIR POLLUTANTS

A. Historical Overview

The impact of air pollutants derived from human activity on vegetation and animals has been recognized for many centuries. The English diarist John Evelyn wrote a pamphlet on air pollution in London, published in 1661, in which he noted, "It is the horrid smoake which kills our Bees and Flowers abroad, suffering nothing in our gardens to bud, display themselves or ripe." There is also evidence from historical records of changes in biodiversity as a result of pollution. For example, until the early 1800s the dominant vegetation of the upland bogs of the Peak District in the southern Pennines of England was Sphagnum moss species. However, with the onset of the industrial revolution in the surrounding valleys, pollutant levels in the region increased rapidly. The Sphagnum species began to disappear, and today the area is dominated by cotton grass (Eriophorum vaginatum). Overall, there has been a loss of species diversity as species associated with the ombrotrophic Sphagnum species, such as sundews (Drosera species) and marsh andromeda (Andromeda polifolia), have also been lost. There is little doubt in this case that the change in community composition is a result of the direct elimination of the sensitive Sphagnum species by increased sulfur deposition.

During the past century, the range and quantities of chemicals discharged into the atmosphere from industry, transport, agriculture, energy production, domestic heating, and many other human activities have increased dramatically. Once discharged into the atmosphere, these compounds are physically dispersed in the atmosphere and may undergo chemical transformations that alter their potential environmental impact. The environmental impact of air pollutants will depend on their concentration in the environment or on the dose received or accumulated by the target organism. At low doses, the environment may be contaminated but no adverse effects ensue; normally, the chemical is only referred to as a pollutant if it has some environmental impact.

B. Sources, Distribution, and Effects of Major Air Pollutants

Air pollutant problems vary greatly in their spatial scales. Some are very local in character, with the environmental impact of the pollution restricted to the immediate vicinity of, for example, a road or a factory. Other problems are regional in character as a result of the long-range transport of pollutants such as acid rain and tropospheric ozone. Similarly, pollutant impacts may vary on different temporal scales. Some impacts, for example, are the result of an accidental release of large pollutant concentrations, which may cause an immediate impact on biodiversity, and from which there may be a slow and gradual recovery, whereas others are the result of an accumulation of pollutant deposited over years or even decades.

This chapter provides an overview of the ways in which pollutants can affect ecological processes and

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Major sources</th>
<th>Major impacts</th>
<th>Scale of effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfur dioxide (SO₂)</td>
<td>Power generation, industry, domestic and commercial heating</td>
<td>Forest decline, elimination of lichens and bryophytes</td>
<td>Local</td>
</tr>
<tr>
<td>Nitrogen oxides (NOₓ) and ammonia (NH₃)</td>
<td>Power generation and transport (NOₓ), intensive agriculture (NH₃)</td>
<td>Altered plant growth and enhanced stress sensitivity, soil acidification and eutrophication</td>
<td>Local, regional</td>
</tr>
<tr>
<td>Acid deposition</td>
<td>Secondary pollutant formed from SO₂ and NOₓ</td>
<td>Soil and freshwater acidification; forest decline</td>
<td>Regional</td>
</tr>
<tr>
<td>Ozone (O₃)</td>
<td>Secondary pollutant formed from hydrocarbons and NOₓ</td>
<td>Reduced plant growth; forest decline</td>
<td>Regional</td>
</tr>
<tr>
<td>Toxic metals (e.g., lead and cadmium)</td>
<td>Smelting industry, transport (lead)</td>
<td>Reduced soil microbial activity, reduced soil invertebrate populations</td>
<td>Local, regional</td>
</tr>
<tr>
<td>Persistent organic pollutants (POPs)</td>
<td>Industry, fuel combustion, pesticide use</td>
<td>Bioaccumulation in food chain</td>
<td>Local, global</td>
</tr>
</tbody>
</table>
biodiversity. Inevitably, it is not possible to provide a comprehensive account of the effects of the vast range of contaminants emitted into the atmosphere by human activity; therefore, this chapter will concentrate on those air pollutants for which the greatest evidence exists of impacts on biodiversity. Table I summarizes the major sources of these pollutants, their major ecological impacts, and the spatial scale of their impacts.

### C. Deposition and Environmental Pathways

The pathways by which pollutants enter ecosystems may have direct relevance to the nature of their impacts. Where pollutants enter as gases, they may be taken up directly into foliage, through stomata, or directly inhaled by animals. However, in other cases pollutants may enter ecosystems in rainfall, in occult precipitation (cloud or mist, in which pollutant concentrations are often higher than in rain), or as particles.

The pathways of aerially deposited pollutants through ecosystems and the mechanisms by which they affect biodiversity are varied. In general, there are four groups of air pollutants which are of significant concern in terms of biodiversity and which differ generically in the nature of their pathways from the atmosphere to the site of biological impact:

1. Acidifying and eutrophying pollutants, such as sulfur dioxide, nitrogen oxides, and ammonia, which may also be deposited as sulfate, nitrate, and ammonium in wet deposition: These gases may have direct effects on organisms, but over longer periods they can also lead to acidification of soils and fresh waters and to changes in their nutrient status, with implications for biodiversity.

2. Photochemical oxidants, primarily ozone: These compounds are not emitted directly into the atmosphere but are secondary pollutants formed as a result of reactions involving nitrogen oxides and volatile organic compounds. These reactions require sunlight and high temperatures, and hence photochemical oxidants are characteristic of warmer climates. The main impacts of ozone on biodiversity are through direct uptake into, and effects on, leaf tissue.

3. Metals and other inorganic contaminants, which are deposited primarily in rainfall or in particulate matter. Atmospheric deposition is only one source of these chemicals; they can also result from soil contamination or from discharges into marine or freshwater systems. The impacts of metals result primarily from their accumulation in soils at concentrations which are toxic to soil organisms or plant roots or through leaching into fresh waters.

4. Persistent organic pollutants (POPs): These are of concern because of the potential for significant bioaccumulation in the food chain. As for metals, emissions to the atmosphere are not the only source of these chemicals. However, unlike for metals, there is evidence that the atmosphere can act as a pathway to redistribute these compounds through "global distillation," in which compounds volatilize at ambient temperatures in warmer parts of the planet and are then redeposited at cooler latitudes. Thus, it is the bioaccumulation of these compounds in the polar regions which is of greatest concern; for example, Antarctic fish have been reported to contain concentrations of certain POPs which are as high as those in North Sea fish.

### D. Current and Future Global Distribution of Air Pollution Emissions

In the past few decades, measures have been taken in Europe and North America which have reduced atmospheric emissions of many pollutants. Although recovery of biodiversity in ecosystems affected by these pollutants may take time, and may not result in the same community as that originally present in a particular location, major new impacts on biodiversity due to large increases in atmospheric emissions are unlikely in these regions. In contrast, the past two decades have seen large increases in atmospheric emissions in parts of Asia, Africa, and Latin America, and continued increases are projected. It is estimated that sulfur dioxide ($SO_2$) emissions in 2010 will be higher in Asia than the combined emissions of Europe and North America. These projections are based on projected growth in population and GDP and assume that no significant measures to reduce emissions will be taken in Asia, Africa, or Latin America. Similar changes in the global patterns of emissions of nitrogen oxides have been projected for the first decades of this century.

### II. MECHANISMS OF IMPACTS ON BIODIVERSITY

#### A. Long-Term and Short-Term Effects on Individual Organisms

Before considering the impacts of air pollution at a community level, it is essential to assess the ways in
which individual plants and animals respond to air pollutants. The impact of a pollutant on any individual organism is complex and may involve many factors. The most important of these factors is the dose of the pollutant received. This will partly depend on the concentration of the pollutant in the atmosphere, or other relevant medium, and the duration of exposure to it. Short-term effects of air pollution exposure at high concentrations result in acute toxicity which is usually characterized by direct damage to exposed tissue and visible foliar injury on vegetation. In contrast, long-term effects of air pollution exposure, which may result from much lower concentrations, result in chronic toxicity, which is usually characterized by alterations in physiology, growth, and reproduction.

However, pollutants may be present in the environment without actually being taken up by the organism. For example, when stomata are closed at night, or under drought conditions, less air pollution can penetrate to the internal leaf tissue, although pollutants may still be deposited on the leaf surface. Similarly, the impact of metals deposited from the atmosphere on soil microorganisms is believed to relate most closely to the free ion concentration in soil solution; for metals such as lead, which are tightly bound to the soil matrix, this may only be a small fraction (the bioavailable fraction) of the total soil concentration.

In general, the relationship between pollutant concentration, or dose, and the response of an organism is classified as follows:

1. When the pollutant is present in sufficiently high concentrations, the organism may be killed outright.
2. At lower concentrations, the organism is able to survive, but its performance is adversely affected. For example, rates of growth or reproduction may be reduced, or changes may occur in the patterns of development or in resource allocation.
3. At even lower concentrations, the physiology and growth of the organism may be unaffected under optimal environmental conditions, but the pollutant may cause subtle morphological, physiological, or behavioral changes which lead to altered tolerance of other environmental stresses.
4. Finally, in the case of certain pollutants, there may be positive effects at low concentrations. For example, certain toxic metals, such as copper and zinc, are essential micronutrients in animal diets, whereas sulfur dioxide can stimulate plant growth at low concentrations, especially in soils which are sulfur deficient.

This range of different biological responses, depending on the severity of the air pollution stress, means that different types of mechanisms exist for impacts on biodiversity. Thus, for 1, direct elimination of species may lead directly to a reduction in biodiversity. For 2, reductions in biomass and population growth rates may affect the longer term viability of populations and lead to local extinctions; however a more likely outcome is a shift in competitive balance between species, which may lead to a changed community composition. In the case of 3, the pollutant has no direct effect on biodiversity but may exacerbate, or mitigate, the impacts of other stress factors. Finally, for 4, although the effect of the air pollutant may superficially be positive, increases in growth of one species may alter the competitive balance and lead to reduced populations or local extinctions of other species. Air pollution may also have effects on some species indirectly through its impacts on other components of the ecosystem. For example, elimination of forest cover through direct effects of air pollution will cause large changes in the microclimate of the forest floor, with consequences for ground vegetation and soil biodiversity.

There is a large degree of variation between species in their sensitivity to particular pollutants; there is also often substantial genetic variation in response within species. This variation broadly relates to the ability of the organism to restrict pollutant uptake or, once it has been taken up, to detoxify, metabolize, or sequester the pollutant. Within the same genotype, other factors such as age and growth stage may also influence sensitivity to air pollutants. Finally, it is important to realize the dynamics of this response, with pollutant exposure frequently inducing adaptive biochemical, physiological, or morphological responses which lead to a reduction in its adverse effects.

B. Interactions between Air Pollutants and Other Environmental Factors

The effect of a given dose of an air pollutant may also depend on other environmental conditions which can modify responses to the pollutant in many ways. First, they may modify pollutant dose; for example, soil water stress may lead to stomatal closure and thus to a reduced uptake of air pollutants by plants and hence reduced pollutant damage. Second, environmental factors may reduce the capacity of an organism to detoxify and assimilate pollutants; for example, it is well established from both controlled experiments and field observa-
tions, that SO$_2$ is more phytotoxic when plants are growing at low temperatures or under low light conditions. Third, exposure to pollutants may lead to changes in the morphology and/or physiology of the organism which make it more sensitive to environmental stresses; for example, increased deposition of sulfate or ammonium ions to conifer seedlings has been shown to increase their sensitivity to cold stress.

Air pollutants rarely occur alone, and the responses to pollutant mixtures may be very different from those to the individual constituent pollutants. These interactions can operate in several different ways. First, two pollutants taken up together may have a greater effect than would be expected from knowledge of the effects of each pollutant—a so-called synergistic response. For example, inhalation of SO$_2$ and particles together often has a synergistic effect on humans and animals, whereas uptake of SO$_2$ and nitrogen dioxide (NO$_2$) together often has synergistic effects on vegetation. Although synergistic interactions between air pollutants have received considerable attention, there are many instances in which the effect of a pollutant mixture is not different from that of the individual pollutants or indeed there is a reduced effect—an antagonistic interaction. Second, over the longer-term, deposition of one pollutant can affect the uptake and impacts of another contaminant; for example, there is evidence that freshwater acidification resulting from deposition of sulfate and nitrate can cause increased bioaccumulation of metals such as lead, cadmium, and methylmercury in fish and birds. Finally, problems such as forest decline may result from a combination of air pollutants which make it more sensitive to environmental stresses; for example, increased deposition of sulfate and nitrate can cause increased bioaccumulation of metals such as lead, cadmium, and methylmercury in fish and birds. Improvements in air quality have led to a reversal of this evolutionary trend, with a decreased proportion of darker forms in the population.

Evolution of tolerance to air pollution has also been demonstrated in vegetation. Studies of emission sources in both the United States and the United Kingdom have shown that tolerance to SO$_2$ in annual and perennial species increases with increasing concentrations closer to the source. There is also evidence that decreasing urban SO$_2$ concentrations are associated with a loss of this tolerance in grass species, suggesting that the SO$_2$-tolerant genotypes are at a selective disadvantage in the absence of the pollutant. Where the pollutant is more widely dispersed, it is more difficult to demonstrate the phenomenon, although there is no reason to suppose that it does not occur. Thus, recent work in the United Kingdom has shown both temporal and spatial associations between ozone exposures and the ozone tolerance of local populations of the annual species Plantago major which is strongly suggestive of an evolution of tolerance.

There is little evidence of the impact of evolution of air pollution tolerance on the overall genetic variation, or fitness, of a population. Key factors are the strength of the selective pressure from air pollution and the frequency of resistant genotypes. Where pollution stress is high and the frequency of resistant individuals is low, and a large proportion of the population is eliminated, the effect on overall genetic variation in the population may be substantial. However, with less severe pollution levels, the effect on genetic variation may be much less.

D. Effects on Interactions between Organisms

When an air pollutant is present at a concentration which affects the physiology, growth, or reproduction of individual organisms, it is clear that the potential exists to influence the outcome of competition between species. In the case of competition between plant species, several simple experiments with air pollutants such as SO$_2$ and ozone (O$_3$) have shown, as expected, that when a pollution-sensitive and pollution-resistant species (often a clover and grass species, respectively) are grown together, the presence of the pollutant shifts the balance in favor of the latter. However, although these simple experiments clearly demonstrate that pollutants can modify the outcome of plant competition, it is doubtful whether they provide much guidance regarding how pollution modifies interactions between plant species in real communities. For example, the vertical stratification of the plant community, which cannot be readily
reproduced in simple competition experiments, may have a major effect on community responses. Field studies of forest stands which have been damaged by industrial pollution typically report a decline in density of the dominant overstory tree species. However, in some cases this can be accompanied by an increase in density in the lower canopy and in the shrub layer as a result of the release of competitive suppression.

Air pollution is also known to alter the relationships between plants and insect herbivores. For example, high concentrations of nitrogen oxides are probably a major factor in the high numbers of aphids, and other phytophagous insects, found alongside major roads. Figure 1 illustrates some of the potential interactions between air pollution, insect pests, and host plants. At high concentrations, there may be direct effects on the insects, but at lower concentrations the effect is mediated primarily through chemical changes in the host plant; for SO₂ and NO₂, the effects on amino acid composition may be particularly important. Effects on the natural enemies of insect herbivores are another possible mechanism of response. These primary interactions can, as Fig. 1 demonstrates, lead to a range of secondary interactions affecting both plant performance and populations of other species at higher trophic levels.

Air pollutants can similarly modify the interactions between plants and fungal pathogens. Many pathogens are very sensitive to pollutants such as SO₂, and the absence of certain diseases producing clear visible symptoms, such as tarspot on sycamore (Rhytisma aceri-num), has been proposed for use as a bioindicator of elevated concentrations of SO₂. Chemical and biological

---

changes on the leaf surface induced by pollutants may critically affect the performance of fungal pathogens and other leaf microflora. Similarly, acid deposition can lead to acidification of tree bark, affecting the range of epiphytic lichen species found.

Pollutant exposure can have an adverse effect on mycorrhizal associations, which are crucial to the stress tolerance (including heavy metal tolerance) and competitive ability of many plant species. This may occur because of direct effects of changes in soil chemistry on some mycorrhizal types due, for example, to mobilization of aluminum by acid deposition or accumulation of heavy metals. Alternatively, exposure to air pollutants above-ground may influence mycorrhizal activity indirectly through changes in carbon partitioning to the roots.

III. EVIDENCE OF IMPACTS ON BIODIVERSITY

The previous section summarized the key factors which may influence the impact of air pollution on biodiversity and provided some simple examples. In this section, some examples of actual impacts of air pollution on biodiversity in the field are discussed in more detail to illustrate the types of effects which have been found. It should be noted that the evidence proving the link between exposure to air pollution and loss of biodiversity is variable in quality. Since experimental manipulations of actual communities over the timescales of decades over which these effects may develop are impossible, it is generally necessary to rely on field evidence. Where the effects are severe and localized, spatial associations between pollutant deposition and biodiversity may be readily established and, especially when the species present before operation of a local source are known, a casual link may readily be inferred. However, where the concern is related to regional-scale deposition of air pollutants at more moderate levels and the effects on biodiversity are more subtle, possibly involving interactions between several factors, the causal link is much more difficult to establish. In such cases, simple cause-effect relationships do not exist, and we are reliant on establishing a balance of probabilities, or using a precautionary approach, in establishing the need for intervention to reduce pollutant emissions to protect biodiversity. The examples discussed in the following sections illustrate a range of cases, from local to regional.

A. Effects of SO2 and Metals

Near Large Smelters

It is clear that when air pollutants are present in high enough concentrations they can cause the complete elimination of all plant species. Near the large Sudbury smelter in Ontario, for example, in the early 1970s areas devoid of vegetation occurred up to 8 km from the source, and species numbers and productivity were reduced up to 20–30 km from the smelter. Figure 2 illustrates the relationships between cover, species numbers, diversity index of overstory vegetation, and the distance from the smelter; similar relationships were found for ground vegetation, although the trend with distance was not as strong. These effects were primarily due to the combined effects of SO2 and heavy metal emissions from the smelter. There was also evidence of acidification and heavy metal accumulation in lakes near the smelter, with adverse effects on the numbers of species of plankton, macrophytes, and fish. During the past 20 years, there has been a reduction in emissions of both SO2 and heavy metals from the smelter, and some reinvasion is evident, particularly of metal-tolerant ecotypes of certain grass species.

Similar patterns of decreases in species diversity and productivity have been reported near other large point sources of SO2 and metals. It is of interest to compare the results found near these sources, which, unlike Sudbury, have minimal SO2 emissions. One of the most intensively studied areas is near the brassworks at Gumbo, Sweden, where metals have accumulated in the surrounding soil over a period of three centuries. Figure 3 shows the relationship between soil copper concentrations and the numbers of taxa found for different groups of organisms; it should be noted that lead and zinc are associated with copper, and it should not be assumed that copper is the sole causal factor. Vascular plant species are relatively unaffected in terms of numbers of taxa, although cover and growth are reduced near the smelter; in contrast, mosses, earthworms, ground lichens, and macrofungi are all significantly reduced in terms of numbers of taxa at higher metal concentration.

Patterns of recovery as emissions decrease reflect the pathways of the pollutants. For example, a series of studies in woodlands near a major metal smelter in southwest England revealed that most of the deposited metal accumulated in the soil compartment, but that there are contrasting patterns of concentration change in the soil profile in response to decreasing emissions for four metals. In the case of cadmium and zinc, concentrations in the surface horizons have decreased since
the mid-1970s and there is evidence of a wave of higher concentrations passing down the soil profile. In contrast, the concentrations of lead and copper, which are highly bound to the soil matrix, have continued to increase in the surface layers, and there is no evidence of movement down the profile. The impact of smelters also depends on soil and climate. For example, the impact of two large nickel smelters in the Kola peninsula, which produce large emissions of metals and are estimated to be major European sources of SO2, may be exacerbated by their location in the Arctic, where nutrient turnover rates are low and the low temperatures are well documented to increase the sensitivity of vegetation to SO2.

### B. Effects of Air Pollution on Lichen Biodiversity

The impacts of air pollution, and SO2 in particular, on lichen species provide another illustration of large-scale changes in biodiversity. The disappearance of many lichen species from European and North American cities during the past century has been well documented; in the United Kingdom, between 30 and 90% of species were lost from areas in which air pollution was a dominant factor. Many of these species were highly sensitive to direct effects of relatively low concentrations of SO2. In contrast, other species are relatively tolerant, and several different scales have been developed to map SO2 concentrations based on the occurrence of particular groups of lichen species.

In many of these cities, SO2 concentrations have decreased dramatically in recent years, but the reinvasion of lichens in response to this decline has been patchy and variable. In London, for example, it has been found that some relatively pollution-sensitive lichens have reinvaded more quickly than species which are more tolerant. This variable pattern of recovery means that the biindicator scales based on lichen species distributions can no longer provide a reliable basis...
for mapping SO₂ concentrations; the lichen communities are no longer in equilibrium with the pollutant levels. There are several possible explanations for these different rates of recovery. One factor is certainly the speed of dispersal of vegetative propagules into areas of decreasing SO₂; certain species (so-called “zone-skippers”) are able to disperse much more rapidly than others. Another important factor is changes in bark chemistry: Recovery of epiphytic species seems to be more rapid on tree species such as ash and willow, which have a relatively high pH bark, than on species with acidic bark, such as oaks. Once bark has been acidified, there may be a delayed recovery of the lichen flora as SO₂ concentrations decrease because bark chemistry changes more slowly than atmospheric concentrations.

There is also evidence of other pollutants affecting the wider distribution of epiphytic lichens. Acid deposition in more remote areas of the United Kingdom and Sweden has been linked to the loss of sensitive cyanobacterial lichens, such as certain Lobaria species; as in the urban situation, these effects are more marked on tree species with acidic, poorly buffered bark. Increased levels of ammonium deposition have also been associated with an increase in the cover and distribution of nitrophilic species and a loss of more acidophilic species in The Netherlands. Thus, although there is little doubt from field evidence that there are currently considerable changes in lichen distributions in many areas of western Europe, a complex mixture of causal factors is involved that includes interactions between changing patterns of pollutant deposition, sub-
strate chemistry, dispersal mechanisms, and other factors.

C. Effects of Nitrogen Deposition on Heathland Communities

A recent example of changes in biodiversity which may be linked to atmospheric pollution, but are not due to direct adverse effects of these pollutants, comes from The Netherlands. Here, in recent decades, heathlands dominated by ericaceous shrubs, such as Calluna vulgaris, have been replaced in many areas by acid grassland communities dominated by Molinia caerulea and Deschampsia flexuosa. A major cause of this change in community structure has been identified as increased nitrogen deposition due primarily to increased emissions of ammonia from intensive agriculture.

Here, the primary mechanism postulated is a shift in the competitive balance between Calluna and the grass species; experimental studies clearly demonstrate that young plants of the grass species are better able to respond to increased levels of nitrogen and thus can out-compete Calluna. However, the field situation is more complex because of the effect of competition for light; once a Calluna canopy is established, invasion by grasses is unlikely, even when there are high levels of nitrogen availability. Thus, it has been suggested that the increased nitrogen deposition also acts by increasing the sensitivity of Calluna to biotic and abiotic stress factors, which will lead to canopy breakdown; in particular, outbreaks of heather beetle infestations, which are favored by higher leaf nitrogen concentrations, can cause widespread defoliation of the Calluna canopy. Finally, it is likely that management practices are also significant; there have been reduced rates of grazing and sod-cutting in recent years, and these traditional management practices have been important in maintaining the low nutrient status of the heathland communities. Thus, in this case, changes in species composition may have resulted from a complex mixture of factors, including pollutant deposition, management practices, competition for light and nutrients, and the effects of climatic stress and insect herbivores.

D. Effects of Ozone on Forest Ecosystems

The three previous cases involved strong spatial gradients in air pollution concentrations due to industrial, urban, and agricultural sources of pollution. In each case, the major pollutants responsible for the observed effects were primary pollutants. In contrast, the secondary pollutants involved in the final two cases are more regional in their distribution, and the spatial and temporal gradients in pollutant exposure are often confounded by other factors.

Where these pollutants are present in high concentrations, the evidence linking cause and effect is strong. For example, the area where the impacts of ozone stress on forest community composition have been most intensively studied is in the San Bernardino mountains, which surround the city of Los Angeles. Effects of ozone pollution, generated from pollutant emissions in the city, began to be observed on the native forest community in the 1960s. The most dominant species of these mixed-conifer forests prior to European settlement were ponderosa pine (Pinus ponderosa) and Jeffrey pine (Pinus jeffreyi) because of their tolerance of the frequent wildfires; however, these have also proved to be the most sensitive species to ozone. In many of these areas, both species have shown severe foliar injury and reduced needle longevity. These are associated with reduced radial growth or even years with missing growth rings.Trees affected by ozone are more susceptible to attack by bark beetles, which are often the direct cause of mortality. Outbreaks of bark beetles are associated with drought years and high ozone concentrations. Regeneration in these forests is greater for trees such as white fir or cedar species, which are more resistant to ozone, although at some higher elevation sites these and other conifer species do not naturally regenerate and the area may become dominated by shrubs. The patterns of change in community composition are confounded by the role of fire since current fire exclusion policies favor replacement of ponderosa and Jeffrey pine by more fire-sensitive species which also happen to be more ozone tolerant.

Experimental studies of the effects of ozone on other tree, shrub, and herbaceous species have clearly demonstrated that, when grown alone, many of these species show significant effects of ozone at concentrations which are found in the field. Systematic surveys of visible leaf injury in the United States have clearly demonstrated the presence of such injury on many native forest and herbaceous species, but the significance of such symptoms for long-term effects on biodiversity is unclear. For example, a recent ozone fumigation study of an early successional forest community showed that ozone decreased species richness, diversity, and evenness; however, the species which dominated the community in the highest ozone treatment was blackberry, which is considered to be highly ozone sensitive on the basis of visible injury responses. The previous examples serve to illustrate that, even where pollutant concentrations are high enough to cause significant levels of visible leaf injury, the short-
term and long-term effects of ozone on canopy composition may be complex and not readily predictable because of the interactions with management practices, biotic and abiotic stresses, and competition between species. Thus, assessment of the long-term effects in other forest ecosystems in which declines of sensitive species have been attributed to the effects of ozone, such as the declines in fir vitality which have been reported recently in the mountains near Mexico City, will need to take account of local factors which may significantly modify responses to the pollutant.

E. Effects of Sulfur and Nitrogen Deposition on Forest Ecosystems

The long-term effects of ozone on biodiversity are difficult to predict because we lack an understanding of the key mechanisms involved at an ecosystem level. In contrast, the effects of deposition of sulfate, nitrate, and ammonium can be understood more mechanistically in terms of their effects in causing soil acidification or eutrophication and in terms of nutrient cycling. Thus, any analysis of the causes of any change in forest vitality needs to consider atmospheric, edaphic, and biological factors influencing nutrient cycling as well as their interaction with other stress factors and with forest management. Figure 4 illustrates the key interactions between these factors. In the case of ozone, the analysis of impacts on forest ecosystems is focused on the direct effects of the air pollutant on the forest canopy, as modified by site factors, pests and diseases, and climate. In contrast, when considering the impacts of sulfur and nitrogen deposition, direct effects on the forest canopy are of minor concern except at very high rates of deposition. Rather, it is the long-term effects on soil, soil solution, and soil leachate chemistry, and their implications for plant growth and vitality, which are crucial.

In soils in which buffering is dominated by cation exchange, increased acid deposition may lead to exchange and leaching of base cations, such as calcium and magnesium, down the soil profile and beyond the rooting zone. In more acidic soils, buffering is dominated by aluminum exchange, and acid deposition can increase levels of available aluminum, which can be directly toxic to fine roots or restrict uptake of other mineral nutrients. Soil acidification may also increase the availability of other toxic metals and can have a direct impact on rates of litter decomposition. There is strong evidence of acidification and base cation depletion of forest soils in western Europe during the past three decades (Fig. 5), and there is little doubt that this


Acidification is due to atmospheric deposition rather than the effects of forest growth. In several of these areas, there is evidence of adverse effects on the canopy density of major forest species such as beech and Norway spruce; these are often associated with mineral nutrient deficiencies, as demonstrated by experiments in which symptoms have been temporarily reversed by soil liming and fertilization.

However, the picture is complicated by evidence that forest growth rates are actually increasing at many European sites. Although increasing atmospheric CO₂ concentrations may contribute to this effect, there is considerable evidence that increasing rates of nitrogen deposition are a key factor, especially because growth at many of these sites is nutrient limited. The longer-term sustainability of these increased growth rates is uncertain. Experimental manipulations of nitrogen inputs to forest ecosystems at a series of sites throughout Europe during a period of 6 years have shown consistent effects on nitrate leaching, but little effect on foliar chemistry or root/mycorrhizal vitality. In contrast, field studies in Switzerland during a 10-year period have associated increased nitrogen deposition with increasing foliar nitrogen concentrations in beech and spruce foliage. There is evidence that these changes may be associated with an increased susceptibility to attacks by fungal diseases and insect pests. Growth stimulation due to increased nitrogen deposition can increase the demand for mineral nutrients such as magnesium and where these nutrients are present in limited supply can lead to foliar nutrient imbalances. Hence, the impacts of increased regional sulfur and nitrogen deposition will depend both on local site factors and on the timescale over which these effects are considered.

Most of the discussion of the consequences of soil acidification and eutrophication associated with increased sulfur and nitrogen deposition has focused on the implications for forest growth and vitality. However, this deposition may also have important consequences for woodland ground flora composition. For example, there have been reports from several areas in Europe of changes in species composition in the ground flora of deciduous forests during the past two or three decades. As expected, some basophilic or neutrophilic species have decreased in frequency or been lost, whereas some acidophilic and nitrophilic species have increased in frequency.

IV. AIR QUALITY GUIDELINES TO PROTECT BIODIVERSITY

Air pollution does not respect boundaries, and the long-range transport of air pollutants means that measures to protect biodiversity in specific areas, such as national parks or nature reserves, may have little value. Thus, many of the national parks in the United States are affected by ozone pollution, and many sites of high conservation in upland areas of the United Kingdom have been affected by acidification. It is clear that the prevention of adverse effects on biodiversity requires measures to reduce emissions and to manage air quality in the regions of concern.

However, complete control of atmospheric emissions is impossible, and the costs of decreasing pollution emissions typically increase greatly as the degree of removal increases. Thus, the following question often arises: How far should we go in reducing pollutant emissions in order to protect biodiversity? Ideally, it would be possible to define air quality standards for biodiversity, as for human health, and devise cost-effective emission control programs to ensure that these standards are met. In Europe, for example, the concept of a critical load has been introduced and is defined as “a quantitative estimate of an exposure to one or more pollutants below which significant harmful effects on specified sensitive elements of the environment do not occur according to present knowledge.” The critical load refers to the pollutant input in all forms of deposition; however, analogously, critical levels, defined in
terms of atmospheric concentrations over a given averaging time, have also been defined for pollutant gases. It is important to appreciate that the critical load concept is based on a precautionary approach and does not relate to current levels of impact; the concern rather relates to the longer term cumulative effects of deposition. Thus, where current deposition rates exceed critical loads, action to reduce emission rates may still be justified to prevent future damage, even though field studies demonstrate no adverse effects on biodiversity of current deposition rates.

Although the critical load concept has proven to be of value in providing an effects-based approach to international negotiations on transboundary air pollution in Europe, it is important to appreciate the difficulties in defining thresholds, in terms of pollutant deposition or concentration, for adverse effects on biodiversity. These difficulties arise from three major sources:

1. There are conceptual problems in defining what is meant by a “significant harmful effect,” both scientifically and in terms of societal judgment. For example, are small shifts in species composition of urban lichen communities, changes in species composition in the microfauna of a woodland floor which do not affect nutrient cycling, or small changes in the genetic composition of populations significant harmful effects. This issue often needs to be addressed in areas where other human impacts, for example, through changes in land use, have already changed substantially the structure and composition of plant and animal communities.

2. Although the adverse effects of air pollution may be clear in situations in which they are present in high concentrations and have dramatic effects, as a “threshold” concentration is approached the effects become gradually more subtle and are difficult to detect in field observations. Furthermore, in many cases the threshold concentration may be close to the natural background concentration of the pollutant or there may be no obvious threshold. Experimental methods may be adequate for assessing direct effects on single organisms, but the long-term effects of air pollution on complex communities are beyond the scope of current experimental approaches.

3. Effects on biodiversity of chronic exposures to air pollution may be the result of the long-term accumulation of pollutants or long-term chemical changes in soils, vegetation, or waters. There is no obvious experimental method of directly testing the consequences of these cumulative effects for biodiversity. The alternative approach, which has been adopted to define critical loads to prevent long-term damage to ecosystems, is to define a critical chemical concentration for biological effects in the relevant medium (typically soil or fresh water) and then to develop mathematical models to estimate the level of atmospheric deposition which would eventually lead to this chemical criterion being exceeded.

V. CONCLUDING REMARKS

This chapter summarized some key aspects of air pollutant impacts on biodiversity. Although the text has, for the sake of clarity, focused on a limited range of pollutants, many of the same principles and ideas apply in the case of other air pollutants and other ecological situations. Although in the past many situations occurred in which high concentrations of pollutants were the dominant factor causing local changes in biodiversity, most of the situations of concern today tend to involve more widely dispersed pollutants at lower concentrations, the effects of which may only become apparent over many years or even decades. In such cases, the pollutant is no longer the single dominant factor but one of a range of biological, climatic, and edaphic factors which may influence biodiversity. If we are to fully understand the role of air pollution in such situations, it is essential that we gain further understanding of the ways in which pollution can interact with these other factors and base our analysis of pollutant impacts more clearly within an ecological framework. It is also vital that we develop more effective methods to identify those areas of the planet in which current or future emissions of air pollutants are a threat to biodiversity.

See Also the Following Articles

ADAPTATION • ATMOSPHERIC GASES • GREENHOUSE EFFECT • NITROGEN AND NITROGEN CYCLE • POLLUTION, OVERVIEW

Bibliography


ALPINE ECOSYSTEMS

Christian Körner
University of Basel

I. The Alpine Life Zone
II. Habitat Diversity, a Key to Alpine Plant Diversity
III. Plant Diversity in the Alpine Life Zone
IV. Diversity of Alpine Animals and Microorganisms
V. Potential Functions of Alpine Biodiversity
VI. Alpine Biodiversity and Global Change

GLOSSARY

alpine Refers to the life zone above the climatic high-elevation treeline, irrespective of latitude. Though originating in the Alps (a pre-Roman word for high mountains), the term is applied globally. The reader should be aware that this term is often used in a much wider sense in common language, and is also applied to regions with mountains in general, including settlements and resorts, which is not the meaning here.
apomixis A very common but “hidden” mode of clonal propagation by seeds, the embryos of which are 100% genetic copies of the source plant. Seeds are produced without fertilization, but often pollination is required to induce apomictic seed production. Apomictic plants also reproduce sexually, but these are very rare events.
clonal growth A vegetative mode of propagation and expansion of plants by runners, tillers, or plant fragment dispersal, which is very important in the alpine life zone. (See also apomixis.) Clonal plants also produce sexual offspring by seed, but their clonal propagules often show higher survival. All clonal offspring of one source plant have the same genome and hence belong to the same “genet.”
ecotype or ecotypic Refers to genetic (evolutionary) differentiation within a given species (a specific “race”) that reflects an obvious advantage in a given environment. Ecotypic traits are retained when individuals are transplanted into a different environment where these traits have no advantage.
life-form The size and stature of a plant under natural life conditions. Environmental constraints can cause the life-form to differ substantially from that of a plant that develops under more favorable conditions, where genotype morphology, the “growth form,” finds full expression. For instance, the growth form “tree” may be modified to the life-form “shrub” in the alpine zone.
microclimate The climate that plants, small animals, and soil microbes experience, and that differs substantially from the “macroclimate” reported by weather stations. This difference is related to surface warming by the sun or cooling at night, as well as wind shelter effects, and is largely driven by relief, exposure, ground cover, and plant stature.
treeline Also known as the forest line, this describes the high-elevation limit of (usually fragmented) for-
I. THE ALPINE LIFE ZONE

The lower boundary of the alpine life zone is, by definition, the natural climatic high-elevation treeline. Where a treeline is missing, as is the case in some dry continental areas or because of deforestation, the elevation of the nearest existing natural treeline is taken as a guideline. This is nothing more than a practical convention. The treeline and hence the lower end of the alpine zone do not form a sharp boundary, for patches of stunted trees and alpine plants often intermingle. The common climatological denominator of this boundary is a mean temperature during the growing season of 5–7°C; this temperature applies to alpine habitats worldwide (see the discussion in Körner, 1999). The length of the growing season varies from 12 months in the tropical alpine zone to merely 4–6 weeks in alpine snowbed vegetation at higher latitudes.

Between 5 and 6 million km² fall into the alpine life zone (ca. 5% of the vegetation-covered land area of the globe). Because not all of this alpine area is covered by vegetation—some consists of bare rock, rock fields, scree, and glaciers—the vegetated alpine land area is estimated to be 3 to 4 million km² (see Körner, 1995).

Two-thirds of the global alpine area is situated in the temperate and subtropical zones and only 10% occurs in the tropics (Table 1). Partly this is because mountains in the tropics need to be very high, at least 3600 m, to permit tropical alpine vegetation to occur, whereas at the polar circles mountain heights of only 600 m are required to support alpine habitat. These are the approximate treeline elevations of the respective climatic zones. At about 4000 m, treeline reaches its highest elevations in the subtropics (Fig. 1). At high latitudes (>65° and 70°N), alpine vegetation merges with arctic tundra. Despite a number of common taxa and the overwhelming influence of low mean temperature, the arctic tundra life zone is very different from the alpine zone in terms of climate, land surface structure, and vegetation, hence several authors (e.g., D. Love and W. D. Billings) have recommended that alpine vegetation not be referred to as “alpine tundra.” The upper limits of vascular plant occurrence are commonly 1000 to 1500 m above the lower limits of the alpine zone (the treeline), but some extreme high-elevation outposts of higher plants are found up to 4400 m in the temperate zone and up to 6200 m in the subtropics, where the uppermost individual of a higher plant was found in the Himalayas (Miehe, 1997). An important feature of alpine life is its isolation. Mountaintops with their high biological diversity represent islands or archipelagos surrounded by lowlands, where most alpine organisms cannot survive.

Various aspects of plant diversity in high mountain systems have been reviewed previously. A selection of such synthetic papers or volumes is included in the Bibliography. Full bibliographic references to original studies (mentioned in the text by authors’ names only) can be found in Körner (1999).

<table>
<thead>
<tr>
<th>Climatic zone</th>
<th>Latitude range</th>
<th>Percent of total area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subarctic/subantarctic alpine</td>
<td>&gt;60°N, &gt;50°S</td>
<td>23%</td>
</tr>
<tr>
<td>Temperate alpine</td>
<td>40–60°N, 35–50°S</td>
<td>33%</td>
</tr>
<tr>
<td>Subtropical—meridional alpine</td>
<td>25–40°N, 20–35°S</td>
<td>34%</td>
</tr>
<tr>
<td>Tropical alpine</td>
<td>0–25°N, 0–20°S</td>
<td>10%</td>
</tr>
</tbody>
</table>
II. HABITAT DIVERSITY, A KEY TO ALPINE PLANT DIVERSITY

A. Climatic Microhabitats

There is a common belief that the alpine life zone is very hostile to plants and animals and that low temperatures restrict life activities, including productivity. This generalization is wrong for two reasons: (1) as will be shown, temperatures are not always that low, and (2) this belief reflects our human perspective of what is “cold.” For organisms adapted to alpine life conditions the temperatures they experience are not necessarily “cold.” In fact, if temperatures were higher, most of these organisms would suffer or even die, many of them because they would be outcompeted by other species that do better at such higher temperatures.

The most important characteristic of the alpine life zone is its fragmentation into a multitude of microhabitats created by relief, exposure, and slope, which interact with solar radiation and wind to cause soil moisture, temperature, and substrate quality to vary enormously over very short distances. It is this diversity of microhabitats, the steepness of the terrain in particular, that makes the alpine life zone so different from arctic tundra, and that is responsible for its much greater organismic richness. Across a few meters one can easily find snowbed plants, just emerging from melting snow, and succulent plants on outcrops, which perform crassulacean acid metabolism (CAM) as do many hot desert plants. In fact, these succulents may regularly experience temperatures close to 50°C on steep south slopes under full midday sun, even at high latitudes.

In addition to these microclimatic determinants imposed by land surface structure, plants themselves influence their micro-environment (Fig. 2). Depending on life-form, plants may decouple themselves from ambient air conditions. Under direct insolation, compact cushion plants or prostrate dwarf shrubs have been shown to warm up to tropical temperatures. These life-

form-dependent microclimates largely disappear under thick clouds or at night, but storage of warmth in the topsoil is also strongly influenced by the type of plant cover, and so is radiative cooling during clear nights.

Hence, the climate that alpine plants experience can be very different from what might be expected based on elevation alone or data measured at weather stations. Such weather data monitor the temperature, humidity, and windspeed that occur outside the calm boundary layer into which alpine plants (leaves in particular) and their animal and microbe partners are commonly nested. A widespread assumption is that alpine plants are small and prostrate because their growth is temperature limited. The reality is that, because alpine plants are small (and mostly stay small when grown at higher temperatures), they may periodically escape the cold and, at least during sunny hours, experience radiative warming, and thus are not always colder than lowland plants. This is well reflected in their thermal optima for photosynthesis, which (in the temperate zone) were

![Figure 2: The diversity of plant morphologies creates diverse microclimates. Note the different temperatures in upright and prostrate leaves of two species in the Ecuadorian Andes. (After M. D. Diemer, in Körner, 1999.)](image-url)
shown to be similar to those of lowland plants. In contrast, night-time temperatures may be prohibitive for plant growth at alpine elevations, and thus limit structural investments of assimilates acquired during the day. Because of the strong link between plant life-form and microclimate, alpine biodiversity can be understood only if one accounts for plant structural diversity.

B. Diversity of Substrates

Alpine microhabitats may belong to a suite of different land surface structures and soils, the ten most important of which are common to all mountains:

- exposed rock terraces and rock crevices
- block fields
- scree and mixed scree/rock slopes or flats
- drained ridges and plateaus
- periodically wet depressions, gullies, or snowbeds
- gentle slopes with relatively stable soil
- steep slopes with creeping soil
- flats with cryogenic structure (e.g., hummocks, polygons)
- mires or other wet ground
- springs, water flows, or lakes

These ten habitat types are not exhaustive. Special habitats not found everywhere include sand drifts or dunes and salt flats (in some semiarid subtropical mountains), disturbed surfaces due to animal trampling, rest places, or burrowing, avalanche tracks, and man-made landscapes such as pastures.

Each of these types of habitats may be found at different exposures to sun and wind (so the number of substrate types multiplies with the number of microclimatic conditions, despite some redundancy), and each of these combinations in turn may be found on very different parent rock material such as calcareous or siliceous, mixed metamorphic, or volcanic material. Depending on wetness and elevation, plants may have converted the top layers of the substrate to an extent that it becomes chemically independent from the parent rock (e.g., humic turf of pH 3.7 overlaying calcareous bedrock). Varying degrees of erosion and soil formation will then enhance the spatial heterogeneity of the substrate. The availability of soil nutrients and the extent of soil development are strongly relief driven, but plant life-forms also determine local nutrient retention. By forming dense cushions or tough tussocks, plant litter and thus the precious nutrients contained in it are prevented from being blown or washed away. When trapped beneath or strongly attached to the plant, litter can decompose and nutrients can recycle locally. Plant life-form, root and rhizome structure in particular, also influences substrate stability and soil formation (see Section V).

To understand alpine biodiversity, it is important to be aware of this extensive microenvironmental patchiness of the alpine life zone, and the capacity of plants to influence their life conditions substantially. Taken together, the foregoing combinations of microenvironmental conditions yield hundreds of very specific niches, each preferred by a different combination of species.

III. PLANT DIVERSITY IN THE ALPINE LIFE ZONE

A. Diversity of Morpho-types

Though the alpine life zone is not as rich in life-forms as a humid tropical forest, the diversity of morpho-types found here is surprisingly high. There are ten principal groups of life-forms, eight of higher plants and two of cryptogams, irrespective of whether individuals perform clonal growth (see later). The first four groups are most important:

- low stature or prostrate woody shrubs;
- graminoids such as grasses and sedges, many forming tussocks;
- herbaceous perennials, often forming rosettes; and
- cushion plants of various types.

Less common or of more regional importance are:

- giant rosettes of tropical mountains;
- geophytes, mainly confined to mountains with a pronounced seasonality;
- succulents, with both stem and leaf succulence; and
- annuals (sometimes biannuals), which become quite rare at high elevations.

The remaining two life-forms are cryptogams, that is, desiccation-tolerant, non-flowering plants:

- bryophytes ("mosses"), in some areas also ferns and lycopsids; and
- lichens (including fruticose, foliace, and crustaceous).

These life-forms, in mixtures of varying abundance of
each, compose the “alpine vegetation.” In addition, algae and fungi play an important role. The diversity of plant structures is further enhanced by various modes of clonal propagation, which become increasingly important as elevation increases. The following is a short list of the diversity of clonal structures (Fig. 3):

- tussock graminoids
- stoloniferous graminoids
- mat- or cushion-forming forbs
- stoloniferous or rhizomatous forbs
- creeping dwarf shrubs
- prostrate dwarf shrubs
- viviparous plants
- accidental clonal plants (fragmentation by external forces)

These typologies do not account for plant height or degree of horizontal spreading, both of which vary considerably among species and microclimates. Overall, the diversity of plant stature in the alpine life zone is perhaps nearly as large as the taxonomic diversity (see

**FIGURE 3** Diversity of clonal growth in alpine plants. The examples shown here include (A) graminoid tussocks, (B) cushions with adventitious roots and potential later fragmentation, (C) mats of rhizomatous forbs or graminoids, (D) mats of proliferating forbs that retain a primary root, (E) stoloniferous plants, and (F) “viviparous” plants producing bulbils or floral plantlets. (After H. Hartmann, from Körner, 1999)
Halloy and Mark, 1996). It is obvious that these structural features are significant functional attributes, strongly associated with microhabitat preference and/or microhabitat tolerance.

B. Diversity of Physio-types

The term physio-type is used here to circumscribe the physiological attributes of alpine plants, which may be as diverse as the variety of life-forms. Certain morphological attributes are functionally linked to physio-types. The following is only a brief summary. Readers with an interest in this field will find an extended treatment of the subject in Körner (1999).

Rates of plant photosynthesis, how photo-assimilates are invested in the plant body, nutrient use, water relations, stress resistance, and secondary metabolites all vary substantially among alpine species. The predominance of physical limitations to growth, particularly at the uppermost limits where plants can grow and survive, may be expected to narrow the spectrum of possibilities, in the ultimate case by permitting only one way to survive. Surprisingly this is not so. Even in habitats that by all standards can be regarded as "extreme," one can find a suite of physio-types (often associated with specific morpho-types). This is a most important point for the study of alpine plant adaptation. The selection of a single species for study inevitably will produce data with no generalization potential. "The alpine physio-type" does not exist. One randomly selected species or small group of species may represent a very special case, and be all but "typical."

A good example is the way plants invest in biomass. Although this finds expression in morphology, the quantitative aspects of it are directly related to plant metabolism. Plants may favor roots, stems, storage organs (all three are net sinks for carbon), or leaves (the net source of carbon). How plants invest is key to the understanding of whole-plant carbon balance and to growth and reproduction. Co-occurring species, equally successful in terms of abundance and often found together in the same habitat, may represent the left and right tails of a frequency distribution of these traits at a common attitude (Fig. 4). There seems to be a multitude of ways to cope with the demands of life conditions even at extremely high elevations.

Another example is the way alpine plants construct their leaves in terms of leaf dry matter investment per unit leaf area. "Expensive" (thick) leaves contain a lot of dry matter per unit leaf area, whereas "cheap" (thin) leaves contain little. Again the diversity of this trait expressed as "leaf mass per area" (LMA, or its reciprocal specific leaf area, SLA) is very large. There is nothing like a typical alpine SLA (Fig. 5; for a full account of such leaf traits in a global comparison see Körner et al., 1989). The same applies to leaf nitrogen concentration and mobile carbohydrates, although there are slight overall trends with increasing elevation for SLA to decrease and nitrogen and mobile carbohydrate concentrations to increase (leaves becoming more "expensive"). However, such means across many species need to be treated with care, given the great diversity. It cannot be concluded that alpine plants have lower SLA when a substantial fraction of all studied alpine species does not fit this pattern. The species from different elevations that are included or excluded from such community subsamples will always affect the mean.

The same applies to any other physiological trait that has been studied in more than one species. For example,
freezing resistance varies enormously in alpine plant species. According to W. Larcher, some species can survive any low winter temperature that could occur on Earth (less than \(-70^\circ\text{C}\)), and others are killed by only \(-12^\circ\text{C}\). Freezing resistance is one of the few traits where this diversity of responses could be explained by habitat characteristics, in this case the predictability of snow cover in winter. But for most other traits such clear-cut causal links have not been found. It rather seems there are many different ways to cope with similar problems. These different traits may be “non-functional” in the sense of not being critical for survival and reproduction under “normal” situations, but some may become decisive under very specific and rare conditions through which they might have been selected for as advantageous (see the discussion of the ecological function of diversity in Section V).

C. Diversity of Reproduction

Recruitment via seed is rather risky in alpine environments because the seedling needs to establish in an often short season and on a potentially hostile substrate. Hence the most common strategies for alpine plants to survive and persist are long life and clonal propagation. The production of seed itself does not seem to be a problem and alpine plants produce lots of viable seed in most years, as has been known from the beginning of alpine plant research in the nineteenth century (see review by Körner, 1999). For a functional understanding of alpine biodiversity, seed production is very important, as rare as successful seedling establishment might be. It is through sexual reproduction that the genetic diversity is retained and by which the ecotypic differentiation of traits, essential in the fragmented and ever-changing alpine environment, develops. The most important pollinators are flies; at lower alpine elevations, bumblebees, solitary bees, and butterflies become more prominent (with decreasing importance by order); a detailed analysis for the southern Andes was done by M. K. Arroyo). Roughly one-fourth of alpine plant species are wind pollinated (such as Poaceae, Cyperaceae, and Polygonaceae).

Current evidence suggests that outbreeding is the dominant form of sexual reproduction, but most alpine plants can also self (which secures some seed production if pollen transfer fails because of bad weather), and some are even obligatory selfers. The most extreme forms of retaining the parental genome is apomixis, where embryos are genetic copies of the source plant. Surprisingly, both clones and apomicts exhibit high intraspecific genetic diversity, indicating that sexual reproduction also occurs in such plants. Through genetic fingerprinting, T. Steinger and coworkers recently showed that some of these obligate clonal alpine plants can produce very large genets and be thousands of years old, but clones of different genets were found to be intermingled. Yet clonal propagation must not be seen as a substitute to sexual propagation. Both the sexual and clonal modes occur simultaneously for most of the time. There is no indication of genetic depauperation within species with increasing elevation, but this is a field that needs more research. Current knowledge does not suggest that alpine plant diversity is limited by reproductive constraints.

It is well established through classic transplantation experiments in California, the Rocky Mountains, and the Alps that ecotypic differentiation does occur among populations from different elevations, so that genotypes of the same species from high elevation differ in stature, phenology, and physiology from lower-elevation genotypes when grown in a common garden (see review by Clements et al., 1950). Hence, in addition to species diversity, discussed next, there is also a high degree of within-species diversity.

D. Taxonomic Diversity

The total alpine flora of the globe consists of approximately 8000 to 10,000 species of flowering plants (Körner, 1995). These belong to about 100 (±10) families and 2000 genera. Hence, one-fourth to one-third of all plant families of higher plants have representatives in the alpine life zone. Assuming a total known global flora of 250,000 species, alpine species contribute about 4% of the global plant species diversity. Given that only 3–4% of the global land area that is suitable for at least some plant growth falls in the alpine zone, global alpine plant diversity per unit land area corresponds to the global mean of other biomes. Although such gross means need to be treated with great caution, one can at least conclude that plant species diversity in the alpine life zone is comparatively high, in view of the generally less luxurious life conditions found there.

Whether high mountains are biodiversity hot spots (Barthlott et al., 1996) is a question of scale. The preceding numbers strictly refer to the alpine life zone as defined in Section I. If one selects a census area that includes the full spectrum of elevations (e.g., a cross section through a whole mountain range), it may encompass almost all life zones on Earth from humid tropical forest at the bottom to glacier forefields at the top. High mountains, those in tropical latitudes in particular, indeed represent an incredible compression of biomes. Climatic and vegetation zones separated by several thousands of
kilometers at sea level may be found within 50 km or less horizontal distance on the flanks of the major subtropical or tropical mountain systems. However, the topic of life zone compression in mountains exceeds the framework of this article, which is restricted to the discrete alpine part of mountain vegetation.

A single mountain system such as the Rocky Mountains, the Alps, the Caucasus, the Venezuelan part of the Andes, or the mountains of New Zealand commonly has an alpine flora consisting of 600 to 1500 species. This needs to be considered in view of a total flora of the arctic tundra of about 1000 to 1500 species (depending on how sub-species are ranked). In this respect, biodiversity of the alpine life zone is outstanding. A distinct mountain region such as the Teton Range in Wyoming, the Snowy Mountains of Australia, or the central part of the Swiss Alps commonly contains roughly 200–400 alpine species, a number that is surprisingly constant across the globe (most ranges harbor around 250 species). On a single sample plot (e.g., 100 × 100 m) one may find one-third of the total regional flora. An analysis of the Swiss Alps by T. Wohlgemut revealed that alpine plant species diversity increases with the size of the observation area up to 20 km², but beyond that species numbers level off.

Plant species diversity generally decreases with increasing elevation (Fig. 6a), although there may be intermittent peaks where two altitudinal life zones merge. Again this is a question of scale. The decline of species diversity is particularly impressive in the uppermost part of the alpine life zone (Fig. 7).

The diversity of cryptogams, mosses and lichens in particular, is relatively high in the alpine life zone. Species numbers may be of the same order of magnitude as for vascular plants, depending on humidity. With greater humidity one generally finds higher numbers of cryptogams. In some parts of temperate zone and subarctic mountains the biomass of fruticose lichens in alpine grassland is double that of higher plants. Since mosses and lichens are desiccation tolerant, they can occupy bare rock and scree and are among the first organisms to initiate humus accumulation on raw substrates. Although cryptogams have almost no elevation limits as long as there is some bare substrate (rock lichens are reported at 7200 m elevation in the Himalayas), the number of cryptogam species also declines with greater elevation, as is shown in Fig. 6b for bryophytes.

Besides the major role of microhabitat diversity discussed earlier, there are also historical reasons for alpine plant diversity. While the lowland flora changed owing to global climatic changes, such as after the ice ages, a fraction of the then cold-adapted lowland flora migrated to high elevations and enriched the existing older stock of alpine species. This is how the edelweiss (Leontopodium alpinum) became a recent addition to the alpine flora of the Alps when the ice retreated. Agakhanjanz and Breckle (1995) suggested that in some mountains (e.g., the Pamirs) tectonic uplift matched time constants of speciation and contributed to the autochthonous species richness.

In summary, the alpine life zone is rich in plant species that belong to a multitude of morphological, physiological, and reproductive types and that inhabit a broad spectrum of microhabitats, reflecting both his-

**FIGURE 6**  Altitudinal trends in species diversity of higher plants in the Alps (a, after J.-P. Thuriillat and A. Schütz) and bryophytes (b, after P. Geissler and C. Vellutini). (Adapted from Körner, 1999.)
torical and even geological events. Because alpine plants are small, species diversity may exceed 50 species per 1 m² of land area, which is among the highest in the world. Alpine plants also provide a varied diet of food for alpine animals and microbes, the diversity of which will be briefly touched upon in the following section.

IV. DIVERSITY OF ALPINE ANIMALS AND MICROORGANISMS

A. Animal Diversity

The great variety of animals from mites to birds, and the greater fragmentation of expertise in animal sciences, may explain why there has been no attempt at a global synthesis of animal diversity in the alpine life zone. Franz (1979), in his German book on high mountain ecology, referred to a great number of otherwise scattered observations, and Meyer and Thaler (1995) attempted a summary for invertebrates, with a focus on the Alps. The following brief account leans heavily on the latter publication.

The only resident vertebrates in the uppermost part of the alpine zone (often called the nival zone) are voles, with the record finding by S. R. P. Halloy at ca. 6000 m elevation on Suenampa volcano in the Argentinian Andes. Snow voles are very abundant and active even above 3200 m in the Alps. At these elevations birds are visitors, but only a few hundred meters lower they are resident as the plant cover becomes more regular. In the Argentinean Andes, ducks breed at 4250 m elevation (personal observation at an alpine lake in the Cumbres Calchaquies). Snakes, lizards, and frogs are reported to occur up to 1000 m above treeline. Large mammal grazers (e.g., ibex in the Alps and guanacos in the southern Andes) occur at almost any alpine elevation, except for peaks surrounded by glaciers or inaccessible cliffs. These grazers profoundly influence the development of alpine plant diversity. In the lower and middle alpine belts, wild herbivores have been replaced by domestic herbivores in many areas (sheep, goats, yaks), which continue to exert selective pressure on vegetation. Some of these animals graze the highest fragments of vegetation, as was reported for yaks, which were found grazing above 5000 m in the dry part of the Himalayas (Miehe, 1997).

Quantitative data for some important groups of invertebrates have been compiled by Meyer and Thaler (1995). The major losses of taxonomic groups as elevation approaches the upper limits of closed vegetation are those of earthworms, gastropods, grasshoppers, Hy-

menoptera, and beetles (Fig. 8), with the latter showing the greatest decline already below treeline. By contrast, flies, spiders, and springtails remain quite abundant (at 300 m above treeline, one-fifth of the total regional number of 200 springtail species were found in the Tirolan Alps). According to L. W. Swan, a salticid spider species was collected in the Himalayas at an altitude above any plant growth, possibly living on the aerial import of small arthropods from lower elevations. Only 2% of the 400 species of spiders of the central Alps of Tirol (Austria) are regularly found in the uppermost alpine zone. The more open vegetation becomes, the more invertebrate life is linked to the occurrence of compact plant forms, such as cushion plants.

In their overview, Meyer and Thaler noted that among invertebrates the herbivore species numbers declined more rapidly than species numbers in higher trophic levels. The reduced species number at "extreme" habitats is often balanced by greater individual densities per species. Overall, animal species diversity in the alpine life zone follows similar elevational trends as in plant species diversity, but in terms of the total numbers of alpine species, animals may exceed plants by factors of 5 to 10 (this is a personal guess). Establishing such diversity ratios would be an important contribution to the understanding of alpine biodiversity in general.

B. Microbial Diversity

As the altitude above treeline increases, animals make a smaller contribution to plant litter decomposition and soil formation and microbes become more important, even as their species numbers decline steadily. Diversity of soil fungi decreases with altitude and mycorrhization, the important plant root–fungus sym-
biosis, also decreases (Fig. 9). All known types of mycorrhizae occur in alpine soils: ectomycorrhizae (e.g., on Salix, Dryas, Polygonum, and Kobresia spp.), ericoid mycorrhizae (Ericaceae), vesicular–arbuscular (VA) mycorrhizae (most forbs, grasses, and some sedges), and even orchid mycorrhizae. Non-mycorrhizal plant species can also be found (Gardes and Dahlberg 1996). So-called vesicular–arbuscular mycorrhizal and dark-septate hyphae of unknown taxonomic affiliation are found even in the highest rock and scree habitats, though at greatly reduced abundance. Completely isolated plants above 3000 m in the Alps were found to be largely free of mycorrhizae, but the dark-septate hyphal root fungi could still be seen (for references see Körner, 1999).

The most robust of all organisms, the bacteria plus some unicellular fungi, retain a high diversity and abundance in the alpine life zone. F. Schimper and coworkers isolated 130 different strains of microorganisms in alpine environments of the western and eastern Alps that can survive and multiply at 0°C. Of these, 77% were bacteria, 20% yeasts, and only 3% hyphal fungi. A very detailed analysis of bacterial diversity on Niwot Ridge, Colorado, by R. L. Mancinelli showed that Pseudomonas and Bacillus were the most abundant genera. Dinitrogen-fixing as well as nitrifying and denitrifying bacteria are abundant in alpine soils. Bacteria have no alitudinal limits as long as some organic dust and short spells with liquid water occur, and Swan reported a number of taxa isolated from substrate collected at 8400 m altitude on Mt. Everest, the environment on Earth that he thinks is most comparable with that of Mars.

Taken together, these observations indicate that mycorrhizae are a common element of alpine plant life, being more prominent in lower alpine elevations and in infertile soils and becoming rare only with isolated plants on high mountain peaks, where soils have little carbon. Rich bacterial life is found even at the highest elevations, indicating a capacity for metabolism under the most extreme conditions.

V. POTENTIAL FUNCTIONS OF ALPINE BIODIVERSITY

Among the theories that attempt to explain the functional significance of organic diversity, the insurance theory seems to be most relevant for the alpine life zone. In simple terms it says that a species-rich, functionally partly redundant organic “work force” ensures the functional integrity of ecosystems even if some of the organisms die out. A complete loss of species by extreme stress or through pathogens would be nearly irreversible in steep alpine terrain, because it is the presence of at least some species that prevents the soil from being washed away. The sustained functional integrity of alpine ecosystems is inevitably tied to the presence of soil, and this presence is inevitably dependent on the roots and rhizomes that hold it. Strong-rooted plants are the keystone elements for the preservation of the alpine ecosystem in steep terrain. A high diversity of species is commonly associated with a high diversity of rooting patterns, which in combination create the mechanical strength required to hold the soil.

B. Messerli and J. D. Ives estimated that 10% of the global human population depends directly, and 40% of the population indirectly, on mountain ecosystems, and thus the stability of their upper part, the alpine zone, is of critical significance to society. Supplies of drinking and irrigation water, as well as the safety of hydroelectric schemes and transport routes, depend on intact upslope soil conditions, and alpine biodiversity helps to ensure ecosystem health. Of course alpine biodiversity also provides other things. Alpine meadows and fellfields are very attractive landscapes for human recreation, and with their rich biodiversity they are of prime conservation value; in many regions they represent the last undisturbed natural areas. Because the alpine life zone is represented across the globe, it is also ideally suited for global monitoring of biological responses to atmospheric change.

VI. ALPINE BIODIVERSITY AND GLOBAL CHANGE

Global change has many facets, all related to the ever-increasing use of resources and land area because of human population growth and the increasing consump-
tion of goods. The most severe impacts on alpine ecosystems are (1) land use practices, (2) potential global warming with associated changes of snow cover and permafrost, and (3) increasing wet nitrogen deposition. Other aspects of global change, such as atmospheric CO₂ enrichment by itself, increasing ultraviolet radiation due to ozone layer depletion, and air pollutants other than nitrogen loading, seem to be of minor significance on a global scale (see review by Körner, 1999).

Land use, in particular the intensification of pasturing or the reverse, the abrupt abandonment of former, traditionally pastured alpine terrain, exerts the greatest influence. The steeper the terrain, the more critical these effects become. Overgrazing generally diminishes plant diversity and ruins the protective plant cover within a few seasons, with self-repair of the ecosystem commonly occurring at a much slower rate than soil erosion. In contrast, moderate and well-managed grazing can increase biodiversity and create short and dense swards of vegetation that are extremely robust against erosion and increase catchment water yield. Some of the biodiversity hot spots of the Alps, the Caucasus, and the Himalayas are traditional pastures in the lower alpine belt. There is no way of maintaining these systems by "let alone" grazing strategies, because herds of domestic animals tend to crowd certain areas and ignore others. Maintaining these intact alpine grasslands is very important for at least three reasons: (1) their often very impressive biological richness, (2) their continued potential as a "clean" food source, and (3) their positive influence on water yield. The abrupt abandonment of pastures leads to an unstable transition phase that may last for a century before new, well-adapted communities of species are able to return. A later reversal back to the previous biodiverse pasture-land is often impossible within a reasonable time and with affordable effort, because the biological structure and the soils of these pastures took many hundreds or thousands of years to develop, but are rapidly converted.

Climate warming will affect alpine biodiversity in subtle ways because of the great variety of intermingled microhabitats (see Section II). Although Grabherr et al. (1995) documented some upslope migration of species, the more important changes possibly happen among microhabitats, with new niches filled by species and other niches abandoned (Fig. 10; Gottfried et al., 1998). These mosaics of life conditions represent a certain margin of safety against the loss of alpine species in a slightly warmer overall climate. However, the abundance of species will change as the abundance of their microhabitats changes. Effects on snow cover and snow duration will be more critical than temperature per se.

FIGURE 10 Global warming and alpine plant diversity: upslope migration of species or niche filling? Both responses are likely and will alter local species abundance. Niche filling, that is, "horizontal reallocation," will precede longer-term changes of the position of whole vegetation belts.

Exceptions to these scenarios are mountains that are not high enough, in which case the current alpine biota will find no upslope escape if it gets warmer.

Since alpine vegetation is well adjusted to cope with low soil nutrient levels, the regional increase of soluble nitrogen deposition will influence biodiversity. More vigorous and nitrogen-demanding species are likely to gain space over slower-growing, smaller species. Since these more vigorous species are commonly also less resistant to stress, nitrogen deposition can increase the sensitivity of certain ecosystems, while others (e.g., pioneer vegetation) may profit. It was shown that even minute additions of nitrogen fertilizer—less than is contained in many places in lowland rain-water—can create drastic changes in the alpine flora (Körner, 1999).

In summary, the greatest risk of loss of biological diversity in the alpine life zone is human land use. However, land use can also contribute to the maintenance of highly diverse and stable ecosystems in the lower alpine belt if sustainable management practices are applied.

See Also the Following Articles

ARCTIC ECOSYSTEMS • EUROPE, ECOSYSTEMS OF • GRAZING, EFFECTS OF

Bibliography


AMAZON ECOSYSTEMS

Ghillean Prance
Royal Botanic Gardens, Kew

I. Nonflooded Terra Firme Ecosystems
II. Transition Ecosystems
III. Floodplain Ecosystems
IV. Amazonian Savannas
V. Montane Fringing Ecosystems
VI. Secondary Forest
VII. Conclusions

GLOSSARY

caatinga Within the Amazon region this is applied to open forest on white sand, which occurs in the basin of the Rio Negro. The term is also applied to the semidesert region of northeast Brazil; it is an indigenous word from the Tupi language meaning an open place.
capoeira Brazilian term for secondary forest on cleared ground.
cerrado The Brazilian term for the large area of savanna and savanna forests that are the dominant vegetation of the Planalto of Central Brazil.
igapó Vegetation periodically flooded by acidic black water or clear water rivers.
inselberg Granitic domes that rise above the forest in the older geological formations to the north and south of the alluvial plains of Amazonia and elsewhere.
llanos The term used in Colombia and Venezuela for the extensive grass-dominated, open savannas of the Orinoco river region.
oligotrophic Term applied to water and soils that are particularly poor in nutrients, such as the weathered white sand soils of Amazonia.
rivers in Amazonia The three main types of rivers in the Amazon region are known as white, black, and clear water. White water rivers are muddy with much suspended sediment and are neutral or only slightly acidic. Black water rivers are dark because of dissolved tannic matter and are acidic (pH about 4). Clear water has neither mud nor humic matter and is usually slightly acidic.
tepui Venezuelan term for the sandstone table mountains of the Guayana Highland or Lost World region.
terra firme Brazilian term for areas that are above the level of periodic inundation by the rivers.
va´rzea Brazilian term for vegetation periodically flooded by white water rivers.

MOST PEOPLE THINK OF THE AMAZON REGION as being covered by a uniform green carpet of rain forest. This is far from reality, because Amazonia is a complex mosaic of ecosystems varying from typical tall rain forest to open grassland savannas to scrubby vegetation on white sand that resembles the chaparral of western California. Even when one type of vegetation is studied, many local variations occur depending on local edaphic
conditions, the flooding regime, or rainfall. This great
diversity of ecosystems has been defined in different ways
by different authors depending on their approach to
the subject. For example, some authors have concen-
trated on the physiognomy and structure of the vegeta-
tion and others, such as Holdridge et al. (1971), on
the life zones that define vegetation based on climatic
factors such as rainfall, temperature, and altitude. In
this account, the primary separation of ecosystems is
based on floristics, that is, the individual species that
make up an ecosystem, and on their physiognomy
and structure.

In addition to the previously mentioned factors that
account for the variety of ecosystems, the history of the
region must also be taken into consideration. Over time
the Amazon ecosystems have by no means been stable.
The distribution of the different types of vegetation
has fluctuated with changes in worldwide climate. The
distribution and species composition of forest and non-
forest biomes changed continuously during the Pleisto-
cene and earlier epochs. What today is continuous forest
was once broken up into isolated blocks interspersed
with deciduous forest and savanna ecosystems that later
coaalesced when the climate again became favorable for
forest. These changes have added to the overall com-
plexity of the current Amazon ecosystems. The princi-
pal vegetation formations are described here.

I. NONFLOODED TERRA
FIRME ECOSYSTEMS

The primary division of Amazonian ecosystems is be-
tween those that are subject to periodic flooding by the
seasonal rise and fall of river water levels and those
that are above the flood level. In Brazil the nonflooded
land is known as terra firme, and since this term is
widely applied in the literature about Amazonian vege-
tation it is also used here. Within the Amazon region
there are forests and savannas on both terra firme
and floodplain.

A. Rain Forest on Terra Firme

This is the single most widespread vegetation type of
Amazonia and covers approximately 50 percent of the
region. Rain forest occurs where there is heavy rainfall,
usually over 2000 mm/yr, that is not markedly seasonal.
Some of the other types of vegetation described here
occur in areas with a marked dry season and/or with
less rainfall, for example, on the southern fringes of
Amazonia where transition forest and semideciduous
forest occur.

The rain forest on terra firme is characterized by
its dense closed canopy at about 25–35 m above the
forest floor. Above the canopy a number of tall
e emerge tree species rise up to about 50 m. The
dense canopy means that light (about 3 percent)
penetrates to the forest floor and the herbs and shrubs
growing there are adapted to living in low light. The
periodic sunflecks are of considerable importance to
those species and to the seedlings of trees. Tree
species range from light demanders that grow quickly
where light is abundant, but that survive poorly in
shade, to shade-tolerant species whose seedlings can
survive and grow in low light.

Rain forest on terra firme is most notable for its
amazing diversity of tree species. There have now been
numerous inventories of trees in different parts of Ama-
zonia. The highest number of trees of 10 cm diameter
or more was recorded in western Amazonia at Cuyabeno
in Ecuador, where Valencia et al. (1994) found 307 tree

<table>
<thead>
<tr>
<th>Site</th>
<th>No. of species</th>
<th>No. of trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuyabeno, Ecuador</td>
<td>283 (117 lianas)</td>
<td>630</td>
</tr>
<tr>
<td>Yanomono, Peru</td>
<td>162</td>
<td>967</td>
</tr>
<tr>
<td>San Carlos de Rio Negro, Brazil</td>
<td>83</td>
<td>844</td>
</tr>
<tr>
<td>Beni, Bolivia</td>
<td>94</td>
<td>649</td>
</tr>
</tbody>
</table>
species on a single hectare. Inventories have listed from 81 to 307 species depending on the soil, rainfall, and the seasonality of the climate (Table I). When the total number of woody species in a hectare is counted, there can be as many as 550 species.

The floristic composition is not uniform and some species have very local distributions. As a result, the lowland forest of Amazonia has been divided into a number of distinct phytogeographic regions based on the distribution of species. The first important subdivision of the region was offered by Ducke and Black (1953): this was later modified slightly by Prance (1977). The seven major phytogeographic regions can be explained by the history of the region and by the present-day climate. For example, the Atlantic coastal region has a wet almost aseasonal climate, but farther inland around Santarem, Brazil, there is a region with a much drier climate and a strong dry season. Farther to the west near the Andes in Peru, the climate is much wetter and less seasonal. Each of these regions falls into different phytogeographic zones. Figure 1 shows the phytogeographic regions with the distributions of some plant species superimposed.

While some species have extremely restricted local ranges and help to define the different regions, others are widespread and occur throughout the entire region, for example, Caryocar glabrum, Licinia heteromorpha, and Parinari excelsa. The distribution of individual species is also dependent on altitude, soil, drainage, and topography, as well as climatic factors. There is not much structural or physiognomic difference in terra firme forest between the different phytogeographic regions, but the canopy height can vary considerably from one place to another.

Between 1972 and 1983, the RADAMBRASIL project surveyed the entire Brazilian Amazon using side-scanning radar. It distinguished two main types of forest on terra firme based on topography: forest in the lowlands on relatively flat ground and hill forest on undulating land between 250 and 700 m. The dense lowland forest occurs mainly on the area of Tertiary and Quaternary sediments between the major rivers on land below 250 m. This forest usually has a canopy at about 30 m and frequent emergent trees up to 50 m, for example, the Brazil nut (Bertholletia excelsa) or the angelim pedra (Dintzia excelsa), the largest tree in the forest. The un-
derstory has relatively few shrubs and consists mainly of young individuals of canopy tree species. Above 250 m the lowland forest tends to be replaced by hill forest, which has a lower canopy and its emergents reach to about 35 m tall. The canopy is more open and consequently the understory is often much thicker and has a greater diversity of shrubs and herbs. The physiognomy of hill forest is far more varied than that in lowland forest, and there are often marked differences between slopes and valleys in both physiognomy and species composition, adding to the total species diversity of the region.

Most trees of the terra firme forest have large heavy fruit or seeds that drop to the ground. After this initial dispersal by gravity they are often carried farther by secondary dispersal agents such as rodents and other terrestrial animals. In contrast to savanna and riverine vegetation, most terra firme forest species are not adapted to long-distance dispersal. This must affect the rate of expansion of the forest into potentially colonizable habitats following climate change or human-caused forest destruction. The various species of primates in the canopy are important dispersers of seeds as well as seed predators.

B. White Sand Oligotrophic Formations

The edaphic factor that has the most striking effect on vegetation on terra firme is white sand soil. This habitat is scattered throughout the region, but especially between the Rio Branco and the Rio Negro in Brazil. These areas are characterized as podzols, or leached pure white sand, known as regosols in Brazilian soil terminology. The origin of these sands is either from the erosion of sandstone mountains in the northern part of Amazonia or from uplifted former river or sea beaches throughout the region.
earlier, Gesneriaceae and with epiphytes. Besides the epiphyte families mentioned particularly tortuous branches that are heavily loaded with epiphytes. Orchidaceae, Araceae, and pteridophytes on the tortuous branches of the trees.

In addition to the almost continuous area of campina forest in the upper Rio Negro, there are many small isolated patches of campina especially in the lower Rio Negro around Manaus. These form part of the transition zone between Amazonian terra firme forest and the branching is tortuous and does not form a closed canopy. As a result, the ground vegetation is abundant and rich and it contains many endemic species, especially in the families Bromeliaceae, Marantaceae, and Rapateaceae. There are also abundant epiphytic Orchidaceae, Araceae, and pteridophytes on the tortuous branches of the trees.

In some campinas there are open areas where bare sand occurs that is often covered by the blue-green alga Stigmosoma tomentosum, as well as small islands of shrubs and low trees with abundant lichens around them. There is good evidence from the presence of pottery shards that these open areas were cleared by indigenous peoples about 800 years ago. The recolonization by vegetation has been extremely slow because of the lack of nutrients in the soil, and so many open patches still remain today.

A study of a campina island near Manaus showed that three-fourths of the species of trees and shrubs are adapted to long-range dispersal by birds, bats, or wind. Therefore, campina species tend to have a much greater capacity for long-distance dispersal than species of the terra firme forest.

White sand areas occur in many other places throughout Amazonia. One large area is that of Serra do Cachimbo on the border between the Brazilian states of Para and Mato Grosso. This area has a similar physiology to the campinas of the Rio Negro, but the species composition differs considerably and there are a number of species that are endemic to the sands of Cachimbo. To the west of this region there are white sand areas in Amazonian Peru and Colombia and in the Brazilian state of Acre. In the Guianas, white sand areas are abundant and are covered by forest or by open savanna. Dense forests on white sand in Guyana are dominated by the legume species Eperua falcata and two other species of Eperua and are termed wallaba forest. To the south of Amazonia there are extensive areas of white sand vegetation in the Chapada de Parecis in Rondônia, where many of the same genera occur as in the Rio Negro region, for example, Abelboda, Clusia, Humiria, Paepalanthus, Retiniphylum, Ternstroemia, and Xyris.

The other white sand forest formation that occurs in Amazonia is called restinga. This low scrub habitat along the seacoast of Brazil is found on sand dunes beyond the influence of salt water and inland from beaches or mangrove forest. Within Amazonia, restinga is not extensive but it extends southward around the coast to southern Brazil. Restinga is a rather uniform scrub characterized by the presence of Chrysobalanus icaco, a species that also occurs in similar habitats in West Africa. Common herbs include the ubiquitous Ipomoea pes-caprae and the sedge Bulbostylis capitulata; shrubs include Hibiscustiliacus and Manihot triflora.

C. Dry Semideciduous Southern Fringe Forests and Cerradão

This formation occurs predominantly in southeastern Amazonia on the border with the cerrado (the savanna) of the Central Brazilian Plateau. This type of dry forest also extends into the cerrado region, where it is termed cerradão. In this region the climate is much more seasonal, drier, and less humid. As a result, the trees lose some of their leaves in the dry season to make the forest semideciduous in marked contrast to the evergreen terra firme forest. Dry forests occur in small patches and do not occupy an extensive area. A small amount of dry forest also occurs in northern Amazonia in Roraima state and in southern Guyana. The dry forests really form part of the transition zone between Amazonian high forests and regions of savanna and are interspersed with the various types of transition ecosystems described in the following section.

The semideciduous forest is not rich in endemic species. Some of the common species that are widespread and characteristic of this formation include Bow-
**dichia virgíolídes, Centrostigma macrophyllum, Combréutum leprosum, Erythrina ulei, Lajosia pucari, Magnolia glabrescens, Attalea speciosa, Physocalymma scaberrimum, and Vochysia haenholiana.**

A study of an area of this type of forest in southern Amazonia at Sarare in Mato Grosso found that it contained many species typical of Amazonian terra firme, such as Enterolobium schomburktii, Hernandia guianensis, Parkia spp., Schizolobium amazonicum, Simarilia amara, and Spondias lutea, mixed with others more characteristic of the forests of São Paulo, such as Aspidosperma, Myroxylon, and Poeppigia procera. There are also a number of species in this belt of transition forest that do not occur in central Amazonia within the dense forest region. The best known of these species is mahogany, Swietenia macrophylla, which occurs around the western and southern fringes of Amazonia. The genus Myroxylon is another example of this type of distribution. The other region that contributes to the semideciduous ecosystem is the cerrado region of central Brazil. Species such as Apúdia leocarpia, Erythrina ulei, Platyophyloia foliosa, and Physocalymma scaberrimum are typical cerrado species that extend their ranges into the semideciduous forest.

**II. TRANSITION ECOSYSTEMS**

The transition forests form a belt of vegetation between the terra firme forests of Amazonia and the savannas of Roraima and Rupununi to the north and the cerrado of central Brazil to the south.

Transition forests mostly occur in regions with a strongly seasonal climate, but the boundary between them and the tall rain forest is not a clear-cut line. Rain forests extend farther into areas with seasonal forests that are on soils where the periodic water stress is compensated for by good water retention. This can occur because of the physical properties of the soil or the proximity of rivers. Conversely, seasonal forests can extend into areas with little dry season in places that easily become dry, such as ridge crests and coarsely textured soils, and also on oligotrophic soils. Thus the superimposition of a climate map on that of vegetation will not necessarily correspond well and the forest boundaries are reticulate. The rain forest to seasonal forest ecotone still remains one of the least-studied ecosystems of Amazonia, yet it is in this boundary where most deforestation is occurring because of its accessibility to the developed south of Brazil and because the soil there tends to be more fertile than under evergreen forests. It is also easier to farm in a seasonal region when there is a time to burn the cut forest and to dry crops. In addition to the semideciduous forest, there are various, more specialized formations that tend to be dominated by one or a few species only. The three principal types of transition forest are described in the following.

**A. Babassu Palm Forest**

Babassu is a species of palm, Attalea speciosa, that dominates the transition forests of the southeastern fringe of Amazonia in the Brazilian states of Maranhão, Tocantins, and Para. In a babassu palm forest, the trees are widely spaced and do not form a closed canopy. The tree height is from 10 to 25 m and many pure stands of Attalea occur. As a result, species diversity is extremely low and there are few endemics. Attalea speciosa is a fire-resistant species and this is probably the reason for its abundance in a transition region to savanna. The fires that occur naturally in the cerrado probably help to reduce the competition from other species and allow babassu to dominate. The extent of babassu forest has probably been increased by deforestation both historically by indigenous peoples and more recently by clearance for farming. Babassu occurs as a component of semideciduous forest, and when the latter is cleared, it is often one of the few species that survives the burning, hence the recent increase in area covered by babassu forest. Babassu is also an extremely useful species and oil is extracted commercially from the kernel of the fruit. The woody endocarp is used to make charcoal or as a fuel for factories, and the outer mesocarp is used as an edible flour.

**B. Liana Forest**

As the name implies, in a liana forest woody vines or lianas predominate. It is open forest with well-spaced trees that are often completely entwined by lianas. In Brazil this formation is known as cipoal or mata de cipo (cipo being the Brazilian word for liana). The tree species in liana forest are some of the commoner species of terra firme forest, but they occur at a much lower diversity. Babassu palm also occurs frequently in liana forest, but there is a far greater quantity of lianas here, particularly in the plant families Bignoniaceae, Dilleniaceae, Leguminosae, Malpighiaceae, and Menispermaceae. To walk through this forest, one has to cut one's way through a thicket of tangled liana stems.

Liana forest is especially abundant in the area be-
tween the Rios Tapajós, Xingu, and Tocantins in southern Pará, Brazil, but it also occurs in small patches in Roraima state in the north. The exact reason for the occurrence of this type of forest has not yet been established. One study showed that the soil under liana forest did not differ from that of nearby terra firme forest. It is thought to occur in regions where the forest has been disturbed by either human intervention or natural climate cycles.

Some of the typical Amazonian tree species that occur in liana forest include Astronium gracile, A. lecointei, the Brazil nut (Bertholletia excelsa), Elizbetha paraensis, and Sapunis marmieri.

C. Bamboo Forests of Western Amazonia

Bamboo forest is an open type of transition forest that occurs mainly in southwestern Amazonia in Amazonian Peru, Bolivia, and the Brazilian state of Acre. In Acre, bamboo forest is abundant and covers a large area. In this forest various species of bamboo (Bambusa subgenus Guadua and the genus Merostachys) are extremely frequent in both the understory and the canopy, where they reach up to 30 m. Bamboo forest contains relatively few tree species and those that do occur are mostly common species of the terra firme forest. The bamboos are interspersed with trees, which lend support to their arching branches. By branching the bamboos spread far over the forest canopy and can easily be picked out in aerial photographs. Botanist Thomas Soderstrom measured a bamboo culm of 29.77 m in length with a side branch of 9.69 m and a secondary branch that was a further 4.96 m. The belt of transition forest that extends across the southern fringe of Amazonia is dominated by bamboos to the west, lianas in the central part, and babassus palms to the east.

III. FLOODPLAIN ECOSYSTEMS

Almost 10 percent of the Amazon region is subject to periodic flooding due to the rise and fall of river levels in most of the region and backup of the rivers from high tides in the delta region. River levels are very seasonal and the magnitude of change is considerable. In central Amazonia around Manaos, the Rio Negro and Rio Solimões undergo an average annual change in water level of 10 m, and the difference between extremely high flood years and the extreme low level is 13 m. The forests beside the rivers are flooded to a depth of several meters for five to seven months each year. The two main flooded ecosystems are those of the white water rivers called várzea and those of the black and clear water rivers called igapo.

A. Várzea Forest

Várzea forest is formed on soil that is flooded by the alluvial-rich white water rivers. The soils are therefore much more fertile than those of the terra firme because of the annual deposit of alluvial matter. For this reason, much of the várzea forest has been replaced by agricultural systems that are more sustainable than those on terra firme. Seasonal várzea is forest that is periodically flooded by the seasonal rise and fall of river level, and tidal várzea is flooded twice a day by fresh water as a result of backup caused by the tide in the Atlantic Ocean. There are few differences between seasonal and tidal várzea and most plant species can occur in both types.

Várzea forest is tall and is physiognomically quite similar to terra firme forest, but it is much less diverse in species. Buttressed trunks are common in várzea forest trees, and lianas are also common. The várzea often has a rich understory with a large number of species of Zingiberaceae, Marantaceae, and Heliconiaceae. In upper Amazonia the várzea forests are much richer in species composition than those of lower Amazonia. Várzea forests tend to have high riverbanks that form natural levees. Some of the typical species are Buchenavia macrophylla, Celea pentandra (the kapok tree), Caesalpa benthamii, Gustavia hexapetala, Hevea brasiliensis (the rubber tree), Hura crepitans, Piranha trifoliata, Rheedia brasiliensis, and Virola surinamensis, as well as numerous species of palms such as Astrocarum humboldtianum.

In the lower Amazon, where the river is very wide, especially between Itacoatiara in Amazonas and Monte Alegre in Pará, the várzea forest is often accompanied by large meadows of grassland known as canarana (Echinochloas spectabrle being the dominant species). This region has narrow banks of forest on higher ground beside the rivers and extensive grass meadows on the lower, more flooded ground farther away from the river margins; the lowest areas form lakes. When the waters recede and the lakes dry up, the open ground quickly becomes covered with grass, increasing the area of the meadows. The arboareal species that are most common in this type of várzea are Bombax munguba, Calycophylum spruceanum, Crateva benthamii, Cordia tetrandra, Pithecellobium multiflorum, Mantinga calabura, and Saltis martiana var. humboldtiana. The latter species of willow occurs mainly on muddy beaches that are in the process of formation.
B. Igapo

The black water and clear water rivers form a very different ecosystem than do the white water rivers. In this case there is little or no sediment and the water is acid and nutrient poor. In black water rivers the water contains much humic matter. The rivers flow from either sandy or rocky terrain. The largest black water river is the Rio Negro and the three major clear water rivers are the Russ Tapajos, Tocantins, and Xingu. In these rivers in the dry season the banks are usually gradually sloping sandy beaches with trees on them, but in the rainy season they are transformed into an inundated igapo forest where only the canopy is above water. Characteristic species include Alchornea castanifolia, Carapa guianensis, and Astrocaryum jauari are frequently used as bait. The seeds of most species of the igapo and varzea forest are dispersed either by water or by fish. Some of the shrubs that grow on riverbanks, such as Mystisia dubia or the cannu camu, Alchornea castanifolia, and Couepia uiti, produce fruit as the river level rises. The fruit become submerged and are then readily available to their fish dispersal agents, who pluck them from the trees.

The physiognomy of the igapo forest is usually rather different from that of the terra firme and varzea forests. This forest is of a lower stature and the trees are more scattered. Many species are confined to the igapo forest but local endemism is rare. One interesting local endemic, Polygenanthis amazonicus, occurs on the sandy beaches near Manaus in Pará, and it is still of uncertain plant classification.

Near Manaus, where the large black water Rio Negro and the white water Amazon (called the Rio Solimões above Manaus) merge, there are extensive areas of inundated igapo forest where only the canopy is above water. Characteristic species include Alchornea castanifolia, Carapa guianensis, and Astrocaryum jauari. Two common palms of the Rio Negro are the piassaba, Leopoldinia piassaba and the closely related Leopoldinia pulchra.

The seeds of both varzea and igapo species are often light and have many different flotation mechanisms, such as spongy tissue, oil bodies, hollow areas, or a light mesocarp. In some cases it is the seed that floats, and in others the entire fruit floats to ensure dispersal. The lakes of Amazonia have an extensive macrophyte flora. White water lakes beside the main Amazon River that are cut off from river flow in the dry season are the habitat of the world’s largest water lily, Victoria amazonica. This spectacular plant is pollinated by scarab beetles that the lily traps inside its flowers for 24 hours. There are many species of Pontederiaceae, including the water hyacinth Eichhornia crassipes, which has become a noxious weed in many other parts of the tropics. Another common floating aquatic in the same family is Rausia rotundifolia.

In most lakes the aquatic flora can be divided into the free-floating species, those that are rooted to the bottom of the lake, and those growing around the margin of the lake. In the free-floating group there are several aquatic ferns such as Azolla, Salvinia, Ceratopteris pterioides, and Marsilia polycarpa. Others include a floating spurge, Phyllanthus flavus, a legume, Neptania oleracea, and Ludwigia helminthorrhiza in the willow-herb family or Onagraceae. Some of the rooted aquatics are Limnocharis flava, Pachyrrhiza dulcis, Capernia castaneifolia, Aeschynomene radiata, Ludwigia decurrens, and Sphacelaea zeylanica. Growing around the lake margin are common species such as Cassia occidentalis, Cyperus mutisii, Parapodium fasciculatum, Hyppis parkeri, and Diodea hantzch. When the lake level rises, large mats of floating vegetation, mainly the grasses Panicum and Echinochloa, are washed out into the river and can be seen floating downstream.

1. Lakes

Not all of the flooded ground is covered by the tall, species diverse, varzea and igapo forests. There are many lakes scattered throughout Amazonia and some of them are extensive, such as Lago Manacupuru in Amazonas and Lago Grande in Pará. In the upper Amazon in Peru and in the upper reaches of some of the main Amazon tributaries, oxbow lakes are also common.

The lakes of Amazonia have an extensive macrophyte flora. White water lakes beside the main Amazon River that are cut off from river flow in the dry season are the habitat of the world’s largest water lily, Victoria amazonica. This spectacular plant is pollinated by scarab beetles that the lily traps inside its flowers for 24 hours. There are many species of Pontederiaceae, including the water hyacinth Eichhornia crassipes, which has become a noxious weed in many other parts of the tropics. Another common floating aquatic in the same family is Rausia rotundifolia.

In most lakes the aquatic flora can be divided into the free-floating species, those that are rooted to the bottom of the lake, and those growing around the margin of the lake. In the free-floating group there are several aquatic ferns such as Azolla, Salvinia, Ceratopteris pterioides, and Marsilia polycarpa. Others include a floating spurge, Phyllanthus flavus, a legume, Neptania oleracea, and Ludwigia helminthorrhiza in the willow-herb family or Onagraceae. Some of the rooted aquatics are Limnocharis flava, Pachyrrhiza dulcis, Capernia castaneifolia, Aeschynomene radiata, Ludwigia decurrens, and Sphacelaea zeylanica. Growing around the lake margin are common species such as Cassia occidentalis, Cyperus mutisii, Parapodium fasciculatum, Hyppis parkeri, and Diodea hantzch. When the lake level rises, large mats of floating vegetation, mainly the grasses Panicum and Echinochloa, are washed out into the river and can be seen floating downstream.
2. Permanent Swamps and Buritizal

In a few places in Amazonia the water never drains and permanent swamp occurs. The permanent swamp forest has a low species diversity and covers only a small area of the region. The soil under the forest is a eutrophic humic gley and in some areas of dystrophic humic gley palm swamp occurs. This is usually dominated by species of Mauritia or Euterpe. There are several extensive areas of *M. flexuosa* palm swamp that are locally called buritizal. These occur in eastern Amazonia, in the state of Maranhão, in central Amazonia in the basin of the Rio Jutai, and in the west near Iquitos, Peru. Since almost all parts of both *Mauritia* and *Euterpe* are useful, these oligarchic or one-species forests have great potential for sustainable use. *Euterpe oleracea* occurs in large quantities in the swamp forests of Marajó Island in the Amazon delta, where it is harvested extensively for both its fruit and as a source of palm hearts or palmito. *Mauritia flexuosa* also occurs in the gallery forests of the Roraima savanna and in várzea forest and is not confined to the permanent swamp forest.

3. Pirizal

Pirizal refers to a type of vegetation that occurs in small restricted areas of eastern Amazonia. A pirizal is a shallow lake or pond with stagnant water and a large number of rooted plants that emerge above the water. The commonest species are the large sedge, *Cyperus giganteus*, and *Ihala gynacalata*, an aquatic member of the Marantaceae. There are also a number of typical floating aquatic species such as species of *Cabomba*, *Eichhornia*, *Limnophila*, *Nymphaea*, *Salvinia*, and *Sagittaria*. The *Mauritia* palm often grows around the margins. The pirizals occur as enclaves in the midst of dense forest, mainly in the state of Amapá in an area along the coast, especially in the region of Maragá. There are also sedge-dominated communities in the Amapá called carazal. These are dominated by *Diplasis karataefolia*, the local name of which is cartá.

D. Mangrove Forest

The mangrove forest of Amazonia occurs as a narrow littoral belt in the coastal area that is subject to salt water inundation by tidal movements. Amazonian mangrove forests are species poor and have a rather uniform formation. The commonest species is the red mangrove, *Rhizophora mangle*, which occurs in the saltiest water. *Rhizophora racemosa* occurs in less salty brackish water, and the intermediate hybrid species *R. harrisonii* occurs in the area in between. All three species are found in the mangroves of the Amazon coast.

The other common species of this formation are *Avicennia nitida*, the white mangrove, *Laguncularia racemosa*, and *Conocarpus eremerus*. *Avicennia* grows inland for a considerable distance upriver into the zone of fresh water, whereas all three species of *Rhizophora* do not extend beyond the influence of salt water. In areas of clay salt beaches the small grass *Sparina brasiliensis* plays an important role in stabilizing the soil as it occurs in thick dense carpets. Some of the species that are common around the fringes of Amazonian mangrove are *Amomna palustris*, *Hibiscus tiliaceus*, *Pithecellobium cockleatum*, and *Pterocarpus officinalis*. The Atlantic coastal mangrove forests of South America are species poor in comparison to those along the Pacific coast.

IV. AMAZONIAN SAVANNAS

A. Nonflooded Savannas and Cerrado

Scattered throughout the rain forest region of Amazonia are a number of patches of open grassland or savanna. Some of these cover small areas and others are quite extensive. The largest ones are the Roraima–Rupununi savanna in Guyana and northern Brazil, which covers 54,000 km², the Sipaliwini savanna on the border between Brazil and Suriname, the Pampas del Heath in Peru, and the Humaitá savannas in the south of Amazonas state. Amazonia is bordered to the south by the largest savanna area of South America, the cerrado of the Central Brazilian Plateau. There are also banks of cerrado that extend into the rain forest region of southern Amazonia, but the isolated Amazonia savannas are a rather different formation from the cerrado. The Amazonian savannas have less local endemism and species diversity than the cerrado, few suffrutescent (woody plants with underground branches and short aerial shoots which die back after fires and resprout), and less tortuously branched trees, indicating that fire is not as strong a factor as it is in the drier cerrado.

The cerrado region of central Brazil has an extremely well defined dry season, the air humidity is often low, and the soils are very deep. As a consequence, roots are also deep and are specialized to reach water at considerable depth when the upper layers of soil are dry. When the topsoil is scraped off by a tractor, immediately there is a large amount of sprouting from vegetative organisms such as xylopodia. In contrast, in Amazonian savannas the roots are more superficial and there are not many vegetative parts in the soil to allow vegetative reproduction. Also, the climate in Amazonia is more humid and the relative humidity of the air is higher. Furthermore, in Amazonian savannas there is never
such a dense arboreal cover as in the more closed types of savanna, whereas in the cerrados, dense closed vegetation (cerrado) is common and is similar to the dry semideciduous forest.

The two ecosystems have many species in common, for example, *Antonia ovata, Byrsonima verbascifolia, Caesalpinia pilulifera, C. parviflora, Qualea grandiflora, Salvertia convalarioidora, and Tabebuia caruaba.* The wide distribution of these and other species is evidence of a historically more continuous distribution of savannas during previous periods of drier and cooler climate. However, it has been shown that many savanna species have a much greater capacity for long-distance dispersal than rain forest trees. Bird- and wind-dispersed diaspores are common in savanna ecosystems, enabling distribution into isolated islands of savanna.

Most of the larger areas of savanna, such as those of Roraima, the savannas of the Rio Purus region of Pará state, and the extensive savannas around Santarém, also in Pará, occur within the drier and more seasonal regions of Amazonia. A belt of drier seasonal climate extends from Roraima to Santarém with under 2000 mm of rainfall and a strong dry season. Other savannas, especially some of the smaller patches, occur for edaphic reasons, such as an underlying hardpan.

There are a large number of savannas in Amazonian Venezuela and these were subdivided into three types by Otto Huber (1982): (1) grassy llanos and llanos-type savannas confined to the north of the region, (2) grassy inundated savannas of the Manapare–Paracoto basin, and (3) the true Amazonian savannas of the central and western part of Amazonas Territory. Unlike many of the Brazilian savannas, the Amazonian savannas of Venezuela have a high rate of endemism. They are mainly of edaphic, pre-Quaternary origin, whereas the llanos and inundated savannas are relic areas from the Pleistocene arid periods. The savannas of Amazonia are therefore not just one single ecosystem.

Other savannas also vary greatly in their botanical diversity and degree of endemism. The Roraima, Sipaliwini, and Humátia savannas have few endemic species, whereas a few savannas have a rich and diverse flora. Most notable are those of the Rio Cururu region of Pará, which occur over sandstone. The sandstone savannas of the Guayana region of Venezuela are also richer in endemism than those of central Amazonia. Savannas that occur over special edaphic conditions are generally much richer than savannas occurring in terrain with similar conditions to those of the surrounding forest. These edaphic savannas are probably older and more stable over time.

Along the northern fringes of Amazonia, other extensive savannas are the llanos of Colombia and the contiguous Gran Sabana of Venezuela. There are also extensive upland savannas in the Departments of San Martín and Madre de Dios in Peru. Overall, savanna covers approximately 15 percent of the Amazon region and so it adds considerable species diversity. Many of the savanna ecosystems are under threat by overuse for farming and agriculture. For example, much of the cerrado region is being destroyed to plant soy and the Roraima–Bupununi savannas are being intensively burned by cattle ranchers. In the dry El Nino year of 1998, the fires spread into the rain forest, causing extensive destruction.

### B. Flooded Savannas of Eastern Amazonia

In the delta region of Amazonia in the state of Amapá and in the eastern half of Marajó Island, there are large areas of inundated savanna or campos de várzea as they are called locally. Other inundated savannas occur farther inland in Pará, especially between the Rio Xingu and Tapajós, and in the basin of the Rio Madeira in Amazonas, but these are much smaller than those of Marajó and Amapá. There are many species of grass in the flooded savannas, such as *Eragrostis hypnoides, E. glomerata, Panthera prostrata, Paspalum orbiculatum,* and *P. guianense.* But the main difference between flooded and nonflooded savannas is the greater frequency of sedges (Cyperaceae) such as *Cyperus giganteus* in the wettest areas and *C. leucae,* C. inferus, and *Scleria geniculata* in the less flooded areas. Shrubs are less abundant in flooded savanna, but in areas that are grazed by cattle *Artemesia artemisifolia* becomes so abundant that it can completely dominate an area because it is unpalatable to cattle. *Ipomoea fischeri,* a shrub that is toxic to cattle, also increases on grazed land for the same reason. Other species characteristic of flooded savannas include *Aeschynomene sensitiva,* *Alternanthera philoxeroides,* *Capronia fusiula,* *Cassia reticulata* (known locally as mata pasto or pasture killer because of its tendency to invade pasture), *Clitoria trichotoma,* *Justicia obtusifolia,* *Mimosa pigra,* *Montrichardia lintifolia,* *Phaeolus limarus,* *Rhabdadenia macroura,* *Sesbania exasperata,* and *Thalictro glandulosa.*

In the flood season, various floating aquatic species, such as *Ceratopteris pteridioides,* *Eichhornia azurea,* *Neptunia oleracea,* *Pistia stratiotes,* and *Salvinia radula,* multiply with great rapidity.

### V. Montane Fringing Ecosystems

Around the northern and southern edges of Amazonia, on the older pre-Cambrian formations of the Guiana
and Brazilian crystalic shields, there are a number of hills and mountains, and to the west Amazonia is bordered by the Andes Mountains. These upland areas, which enclose the lowland Amazon basin on three sides, add considerably to the biological diversity of the Amazon ecosystem. Small changes in altitude or topography can cause significant changes in the vegetation type, physiognomy, species composition, and climate.

A. The Sandstone Tepuis of the Guayana Highland

An area with a complex mixture of vegetation types is the Guayana Highland, which is dominated by dramatic sandstone table mountains or tepuis. The mountain slopes are covered by lower and upper montane forest and often there is an abrupt change to open savanna and swamp formations on the summits of the mountains because of the sheer sandstone cliff faces. The highest mountains have less forest on their summits and more open vegetation except for gallery forest along the streams. The lower mountains have considerable areas of forest on their summits. For example, on Cerro Guaiquinima in Venezuela, one of the largest sandstone tepuis, 40 percent of the summit is covered by forest. Both tall forest and relatively low dense forest, as well as intermediate types, occur on Guaiquinima. There is a strong relationship in species content between the forest on the summit of this and other tepuis and the forest of the neighboring lowlands. Species endemism is high on the sandstone tops of the Guayana Highlands, and many genera have a series of closely related species with one occurring on each of the larger mountains, for example, the pitcher plant genus *Heliamphora* or the shrub *Tepuianthus*.

B. Granite Inselbergs

Scattered throughout the Guianas and northern Amazonian Brazil are a series of granitic outcrops or domes that rise above the rain forest to a height of 300–800 m. In some cases these hills are covered by dense forest (*e.g.*, Palunlouimeéempeu and Mitraka), but most of them are characteristic inselbergs with low scrub forest and open areas of exposed rocks. These are obviously islands of a special type of vegetation surrounded by rain forest. The tops of the inselbergs are well drained and become very dry in the dry season, and so the vegetation is often dominated by sclerophyllous plants or species with other adaptations to drought, such as the orchid *Cyrtopodium andrenonii* with large pseudobulbs that store water or various cacti (*Eppiphyllum* and *Melocactus*). The dominant shrub is usually a species of *Clusia* and other terrestrial orchids include *Epipendrum nocturnum* and *Encyclia tosona*, and the bromelad *Pitcairnia gyxestis* is also common. Many of the plant species on the inselbergs of the Guianas are confined to that habitat. There has been considerable speciation in adaptation to the summit of inselbergs with their arid dry season conditions and very humid rainy season conditions. However, in contrast to tepuis, there has been much less specialization between the different inselbergs, and local endemics are rare.

C. Low Hills within Amazonia

In addition to the tepuis and inselbergs scattered throughout the Amazon region are several lower hills that can have distinct vegetation. These are more frequent to the north of the Amazon River and near the Brazilian Shield region and are mostly outliers of the ancient shields. Small changes in altitude can have extremely important effects on vegetation type, physiognomy, species composition, and climate. For example, outside Amazonia, small low hills in the arid northeast of Brazil tend to accumulate cloud and therefore are covered by tall moist forest called brejo. The brejo forests contain many Amazonian species. The different effects between larger and smaller mountains (known as the Massenerhebung effect), as well as latitude, on the vegetation type are quite striking in the Neotropics. Consequently, there are many small patches of montane or cloud forest on small outcrops throughout the lowland region. Also the altitudinal limits of the different forest types vary considerably depending on the effects of local climate, soil, and latitude. Some of the species and genera that occur on the higher mountains also occur on these areas, for example, *Retinophyllum* and *Pagamea* in the Rubiaceae.

D. The Andes Foothills

The Amazon region is bordered on the west by the Andes, which rise abruptly out of the lowland rain
forest. Unlike the tepuis and inselbergs, the Andes are a relatively recently formed mountain chain and have a distinct zonation of the vegetation according to altitude. The forests of the eastern slopes fall into the ecosystems of Amazonia and add considerable variety of both vegetation types and species.

The lowland forest extends up the mountains to between 700 and 1000 m. Above that altitude it is replaced first by lower montane, and then by montane vegetation types. The montane forests can generally be divided into three zones: lower montane forest, upper montane forest, and subalpine forest. Lower montane forest begins between 700 and 1200 m. It is quite similar to lowland rain forest, but a large number of species drop out and are replaced by more upland species. This forest tends to be lower in stature than that of the surrounding lowland area, and it has fewer woody vines and fewer buttressed trees. Vascular epiphytes are common and mesophyll leaf types are predominant. Many of the lowland taxa persist into this zone, such as species of Livistona (Chrysoisolanaeaceae) and Eschweilera (Lecythidaceae), but a number of distinctly highland elements also enter the lower montane forest, for example, in the Colombian Andes, Brunelia com- ocladia, Alchornea bogotensis, and Cinchona calycina. Palms such as Euterpe purpurea and Wettinia clados- podyx are quite abundant, and woody vines include Anomopernum occidentale and species of Passiflora and Paullinia. This zone typically extends upward to between 1800 to 2400 m depending on latitude and local conditions.

The upper montane forest begins at 1800–2400 m and may extend in places up to 3400 m. It is usually of lower stature than the lower montane forest, with a predominance of microphyllous trees. Vascular epiphytes are still common and woody lianas are rarer. An increasing number of species characteristic of higher altitude enter the flora, for example, Brunelia occiden- tals, Weimannia balliiiana, Symplocos pichindensis, and the vines Hydrangea peruviana and Liabum mega- cephalum.

The subalpine forest, including the dwarf elfin forest, is of frequent occurrence in the Andes and extends locally into Central America. It is microphyllous or, at its altitudinal extreme, nanophyllous or dwarf-leaved. Few vascular epiphytes and climbers occur, but there are abundant bryophytes and lichens. This formation may extend up to 3800 m in some places, above which it is replaced by the herbaceous alpine formations such as paramo and puna, which are not discussed here. The subalpine forest has few predominantly lowland genera but many that are characteristic of the highlands, for example, species of Weinmannia (Canoniaceae) and Gynoxys (Asteraceae), Brunellia (Brunellaceae), Clusia (Clusiaceae), Refaria (Ericaceae), Micromeria (Melastoma- taceae), and Rhamnus (Rhamnaceae).

A most interesting type of dry forest occurs in small isolated dry valleys that are scattered along the Andes due to local weather conditions. They are significant as a link between the larger arid areas, and they were probably more nearly connected during the dry periods of the Pleistocene and early Holocene. Important genera in these valleys include Acacia, Bursera, Cereus, Cardiopodia, Cercidium, Prosopis (P. chilensis and P. limensis), and many Cactaceae.

E. Campo Rupestre

The campos rupestres, or open formations on rocks, are generally confused with open savanna and the orchard savanna on terra firme, but they are quite different physiographically and floristically. They develop on rocks and in rocky terrain, because they suffer drought in the dry season despite the equatorial humid climate, and because there is no possibility for water retention and all rainfall runs off immediately. Typical species of the cerrado do not occur here, such as Curatella americana, Hancornia speciosa, Salvertia convallaria- dora, and Qualea grandiflora. Instead there are various species of Byrsonima, Chusia, Norantea, and Vellozia, certain Bromeliaceae (Bromelia, Dyckia, and others), and many Ericaceae (Parpalanthus and Syngonan- thus). Lichens are also frequent and cover many of the rocks.

Campos rupestres are also quite common in central Brazil, where they are often confused with cerrados. For example, Serra do Cipó in Minas Gerais, much cited in the literature because of its interesting landscape and large number of Velloziaceae, is mostly composed of campo rupestre. In Amazonia, Serra do Cachimbo, in the southwest of Para, has an area of campo rupestre that covers about 16,000 km². Another large patch is the campo of the Rio Curiuri, in the Tapajós River basin. At Atiramba, on the Rio Trombetas, both savanna and campo rupestre occur.

The canga or ironstone vegetation, which occurs on Serra dos Carajás over the iron deposit, is also a special form of campo rupestre. On this formation, various extra-Amazonian species are common, such as Pilocarpus microphyllus, which occurs also in the northeastern states of Piauí and Maranhão; the extra-Amazonian ge- nus Callisthene is represented by C. microphyllus of central Brazil, which occurs only within Amazonia on Serra Carajás. Also found on Carajás is the curious and
The branches of this thick-barked treelet sometimes lengthen and turn into vines. The Amazonian montane forests also show a strong affinity to campo rupestre and some could even be classified as floresta rupestre (i.e., forest on rock). In the mountain areas there has been a great accumulation of organic matter because, although the forests sometimes endure short dry periods, they are maintained by the high air humidity and mist.

VI. SECONDARY FOREST

Unfortunately, secondary forest is of increasing importance in Amazonia owing to the amount of felling of primary forest that is taking place. Obviously the species that now dominate secondary areas had a natural distribution prior to the occurrence of man-made secondary forest. A few secondary forest species are found throughout the tropics, and others grow in open spaces beside rivers, on old landslides, in gaps caused by tree falls or dead trees, or where forest has been felled by severe storms. Because the seeds and stumps have not been destroyed by fire, regenerated vegetation of naturally cleared areas is different from that in areas burned by man. Secondary forest species do not generally occur in the transition regions between forest and savanna. Modern vegetation maps show an increasingly large area of a new man-made vegetation and also of agricultural areas, but these are not discussed here. Some of the characteristic genera of secondary forests are Cecropia, Byrsonima, and Vismia. In Brazil, secondary forest is known by the term capoeira.

VII. CONCLUSIONS

This summary of the principal vegetational formations that occur in Amazonia shows the considerable diversity that accounts for the approximately 30,000 species of vascular plants within the region. Many of the ecosystems described here could be subdivided further into many different local communities. A knowledge of this ecosystem diversity and the distribution of species is an essential tool for the conservation and sustainable use of the region. To preserve the diversity of Amazonia, it is vital to conserve areas of each different type of vegetation because of the number of unique species of plants and animals that each ecosystem contains.
Amphibians are tetrapod vertebrates. They differ from the other tetrapods (the reptiles, birds, and mammals) in that their eggs are anamniotic; they are relatively simple and are enclosed in a jelly capsule.

GLOSSARY

anamniotic Eggs are not surrounded by the complex membranes that distinguish the amniotic eggs of reptiles, birds, and mammals.
cloaca The common chamber in which the reproductive, excretory, and digestive tracts of amphibians unite before exiting the body.
metapopulation A group of local populations among which individuals migrate relatively frequently; however, the rate of migration is slow enough that the populations fluctuate independently.
paedomorphosis Reproduction while retaining at least some larval characteristics.
tetrapods The terrestrial vertebrate classes amphibians, reptiles, birds, and mammals, so named because they primitively possess four legs.

I. THE EVOLUTIONARY HISTORY OF AMPHIBIANS

Amphibians first appeared during the Late Devonian, about 360 million years ago (Ma). There is a general consensus that all amphibians shared a common ancestor, a sarcopterygian (fleshy-finned) bony fish in the class Osteichthyes. All sarcopterygians have paired fins with limb-like bones, and many exhibit other anatomical features, such as lungs, that ally them with the tetrapods. It is not likely that either of the extant groups of sarcopterygians, the dipnoans (lungfishes) and crossopterygians (lobe-finned fishes), contained the ancestor of the amphibians. The earliest major radiation of terrestrial vertebrates occurred during the Carboniferous Period (ca. 335 Ma). By the end of the Triassic, about 200 Ma, nearly all of the large ancestral amphibians were extinct. The subclass Lissamphibia, the modern amphibians, appeared during the Triassic and is the only group that has survived to the present.
The ancestry of modern amphibians is poorly understood because there is a sparse fossil record linking primitive amphibians to the three modern orders. The earliest fossil Lissamphibian is *Triadobatrachus massinoti* from the early Triassic, about 230 Ma. *Triadobatrachus* is similar to the modern frogs but is not considered to belong to the order Anura. The Anura (frogs) and Gymnophiona (caecilians) appeared during the Early Jurassic (about 190 Ma), whereas the Caudata (salamanders) appeared in the Middle Jurassic, 170–150 Ma. Although the fossil evidence is sparse, phylogenetic analyses of shared derived morphological characters and of molecular characters strongly suggest that the Lissamphibia share a common early amphibian ancestor and that the Lissamphibia and the amniote tetrapods (reptiles, birds, and mammals) originated from different early amphibians. It is possible, however, that the Lissamphibia is a polyphyletic group, with one or more of the modern orders having an independent origin from the subclasses Temnospondyli or Microsauria.

II. HISTORICAL BIOGEOGRAPHY AND CURRENT DIVERSITY OF MODERN AMPHIBIANS

Modern amphibians are found in all continents except Antarctica. However, the three modern orders show disparate patterns of fossil and recent diversity that are associated with their different histories on the landmass of Pangaea and subsequently on landmasses associated with Gondwana and Laurasia. The evolution of salamanders is linked closely with landmasses derived from Laurasia. All fossil and living species occur in the Northern Hemisphere, with the exception of the lungless salamanders (family Plethodontidae) which invaded South America relatively recently. In contrast, the caecilians and most frogs are associated predominantly with the southern Gondwanan landmasses. The caecilians occur throughout the tropics, except in Madagascar and to the east of Wallace’s Line in Australasia. They are the least diverse group of living amphibians, with five families and 165 species. The greatest diversity of species occurs in northern South America and central America. Frogs are the most widespread of the three orders. Although the breakup of Pangaea probably isolated salamanders on the Laurasian continents and many groups of frogs on the Gondwanan continents, several clades of frogs have dispersed to all continents and frogs are now absent only from Antarctica.

There are now more described species of amphibians than there are of mammals (Glaw and Kohler, 1998;

### Table 1: Richness of Modern Amphibian Taxa

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Families</th>
<th>Genera</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anura</td>
<td>23</td>
<td>334</td>
<td>4204</td>
</tr>
<tr>
<td>Caudata</td>
<td>10</td>
<td>81</td>
<td>411</td>
</tr>
<tr>
<td>Gymnophiona</td>
<td>6</td>
<td>33</td>
<td>165</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>428</td>
<td>4780</td>
</tr>
</tbody>
</table>


Table 1: More than 400 salamander species are currently known in 10 families, with the highest diversity found in North America. One family, the Plethodontidae of North and South America, contains more than half the known salamander species. The frogs are placed in 25 families, and more than 4200 species are known, 80% of which are in tropical regions. It is likely that at least several hundred species remain to be described. The greatest diversity of frogs occurs in South America, but other tropical areas such as Southeast Asia, New Guinea, Madagascar, and central Africa also have highly diverse frog faunas.

Four of the five most diverse frog families, the Helidae, Ranidae, Bufonidae, and Microhylidae, are each found across several continents—distributions that reflect successful dispersal. The most diverse family is the Leptodactylidae, which is restricted to South and Central America and contains 22% of living frogs. It contains the genus *Eleutherodactylus*, which is the most diverse vertebrate genus with more than 530 species.

III. BASIC MORPHOLOGY AND FUNCTIONAL ANATOMY OF MODERN AMPHIBIANS

All modern amphibians have complex glandular skins and most lack scales. Their skins are kept moist by the secretions of mucous glands, whereas granular glands produce a variety of toxins that serve to deter predation and may also help to protect from microbial infections. They shed their skins periodically and usually consume them as they are shed. Their color patterns are produced by xanthophores, iridophores, and melanophores. Many species are able to alter the shape of pigment cells and the distribution of pigments within them and can rapidly change color.

The eyes of most species contain photoreceptors of several types, probably giving at least limited color vision, they can be covered by moveable eyelids, and they...
focus by moving the lens using ciliary muscles. Many frogs and salamanders have a wide binocular field of vision and have good depth perception. Many species use olfaction in the detection and capture of prey and have well-developed olfactory systems and Jacobson's organs. Larval amphibians and adults of aquatic species have lateral line systems. The amphibian auditory apparatus differs from those of other terrestrial vertebrates. Their middle ears have a columella bone (also found in reptiles and birds and modified in mammals) which functions in the reception of high-frequency sounds. A second bony element in the middle ear, the operculum, receives low-frequency vibrations from the air and substrate via a muscular connection to the pectoral girdle.

Most terrestrial amphibians respire (in part) with lungs and have largely separate pulmonary and systemic blood circulations. Their hearts typically have three chambers—two atria that receive blood from the body and the lungs and one ventricle that serves to pump blood to the lungs and the body. Although the ventricle is usually single, a combination of partial septa and the high viscosity of blood reduce mixing of the blood streams flowing to the lungs and the remainder of the body. Many species carry out a considerable fraction of gas exchange across the moist skin; for the lungless salamanders (family Plethodontidae) this is the primary mode of respiration.

The kidneys of amphibians function in maintaining water and ionic balance. In species with aquatic larval stages, most water and ion balance needs are reversed at metamorphosis; the freshwater larvae must cope with a hypo-osmotic environment, to which they lose ions and from which they gain water, whereas the terrestrial stages must cope with rates of evaporative water loss that are typically high because of their moist skins. Aquatic individuals tend to excrete dilute solutions of ammonia, whereas terrestrial animals excrete more concentrated solutions of urea, although no amphibian excretes fluid urine that is more concentrated than the blood plasma. Water uptake in almost all amphibians does not involve drinking but is accomplished by active or passive transport of water across the skin.

Amphibians are ectotherms, relying on the external environment as a source of body heat. All aquatic forms and many terrestrial forms are also poikilotherms or thermoregulators, with body temperatures that do not differ from their environments. Some species may control their body temperatures to some extent by selecting microhabitats that provide appropriate temperatures, and a few species of frogs periodically increase their body temperatures by basking in sunlight. The ability of most terrestrial amphibians to thermoregulate by basking is limited by their need to conserve water while maintaining a moist skin. Some species of frogs and salamanders can be active at temperatures approaching 0°C, whereas others can tolerate body temperatures well above 40°C. Some temperate species can survive exposure to temperatures below 0°C by supercooling. A few species can survive partial freezing by a combination of ice nucleating proteins that encourage freezing of extracellular fluids and high intracellular glucose concentrations that prevent intracellular water from freezing and keep cells from dehydrating.

A. Special Features of Caecilians

Caecilians are aquatic and burrowing animals that superficially resemble large earthworms. Adults range from approximately 10 to 150 cm in length. They have elongate bodies with distinct annuli, which are grooves delineating their body segments. They are limbless, and their tails are reduced or absent. Their eyes are reduced and are covered by skin. They are unique among the Chondrichthyes in possessing dermal scales, which occur in the annuli of some species. Their skulls are heavily ossified and completely roofed. Caecilians possess a unique chemosensory organ, the tentacle, which extends a short distance from the surface of the head, emerging from a skull opening between the eyes and the nostrils.

B. Special Features of Salamanders

Salamanders are typically four-limbed animals with relatively long tails that superficially resemble lizards but lack epidermal scales and claws. Salamanders range from about 30 mm to 2 m in total length. Their limbs are relatively small and are reduced or lost in some terrestrial and aquatic species. Their skulls typically show the loss of many bony elements. Salamanders lack external ears and, with the exception of weak distress calls in some species, do not vocalize. The most diverse group of salamanders, the Plethodontidae of the Americas, is limbless. Salamanders of many groups exhibit various degrees of paedomorphosis. Salamander larvae are carnivorous and usually have well-developed external gills.

C. Special Features of Anurans

Anurans are invariably four-limbed, and terrestrial juveniles and adults completely lack true tails. Adults range from about 1 to 30 cm in length. Their hindlimbs and feet are greatly elongated. The radius and ulna of the forelimb and the tibia and fibula of the hindlimb are fused. There are no more than nine trunk...
vertebrae, and most modern frogs lack free ribs. The caudal vertebrae are fused into a rod-shaped urostyle that is associated with the elongate pelvis. Most of these features are probably the result of adaptation for lightweight and strength for jumping and show interesting convergences with similar adaptations in birds. The lightweight skulls of frogs are large relative to their body size and lack many bones. All but one species lack teeth on the dentary bone of the lower jaw. The tongue is attached at the front of the mouth in most species and is flipped forward rapidly to capture prey. The ears of frogs follow the general amphibian model, with the addition of an external tympanic membrane in many species.

The skins of frogs depart from the usual amphibian pattern in several ways. Some species have additional types of glands: lipid glands that secrete lipids which reduce rates of evaporative water loss or breeding glands that produce sticky secretions which adhere the male to the female during amplexus. Skin lipids and other as yet poorly understood modifications of the skin allow some treefrogs to achieve rates of evaporative water loss as low as those of some lizards. Frogs also have distinct differences between their dorsal and ventral skins. The ventral skin usually has fewer mucous glands, and many terrestrial species have an area called the pelvic patch in which the skin is unusually thin and highly vascularized. Frogs sit with the pelvic patch in contact with moist substrates to increase the rate of water uptake. Burrowing desert frogs of several lineages form cocoons by repeatedly shedding their skins and retaining the shed layers to reduce evaporative water loss while they are buried.

Frogs have a variety of adaptations to structures other than the skin for maintaining water balance. Most species store copious quantities of dilute urine in the bladder when water is available and withdraw water from this pool to replace evaporative losses. Desert species that inhabit brackish water can allow high concentrations of urea to accumulate in the body fluids. A few species can excrete nitrogenous wastes as uric acid, which minimizes the water loss when they are excreted.

IV. REPRODUCTIVE BIOLOGY AND LIFE HISTORIES OF MODERN AMPHIBIANS

The “typical” life history of the Lissamphibia includes a complex life cycle in which eggs are deposited in freshwater habitats, where larvae grow and develop, eventually metamorphosing into terrestrial juveniles and leaving the water. Although this is a common pattern found in all three modern orders, it is far from universal. One of the major features that sets amphibians apart from the remainder of terrestrial vertebrates is their extremely wide range of life histories and modes of reproduction, many of which occur in all three extant orders. The presence of a wide range of life histories and reproductive modes suggests that the “typical” amphibian life cycle does not reflect a failure to adapt to the terrestrial environment but rather serves as an adaptation that allows female amphibians to produce very large numbers of eggs with a small investment in each and to exploit freshwater habitats for larval growth and development.

A. Caecilians

Limited observations have been made on the reproductive biology of caecilians and no information on courtship behavior is available. It appears that fertilization is internal via protrusion of the male cloacal wall. Although vocalizations have been reported for some species, it is not known if they are linked to reproduction. Caecilians are oviparous or viviparous. In viviparous species gestation may take up to a year with reproduction occurring only every 2 years, and nutrition can be supplied to the young within the oviduct. Oviparous species lay terrestrial eggs but the larvae can be aquatic or can complete development within the egg. Oviparous caecilians produce more offspring than viviparous species. In oviparous species, parental care of eggs is common.

B. Salamanders

Many salamanders conform to the “typical” amphibian complex life cycle. Most produce aquatic eggs and larvae, which metamorphose into terrestrial juveniles. As in the frogs, there is a great deal of variation beyond this basic mode. Larval salamanders are relatively similar to adults; this has allowed species belonging to several groups to evolve the ability to reproduce while retaining larval characters, either facultatively or obligately. An example of the complexity and variability of salamander life histories is the life history of red-spotted newts (Notophthalmus viridescens, family Salamandridae). Their life history begins with aquatic eggs which hatch into aquatic larvae. The larvae typically develop within one summer and metamorphose into a morphologically distinct terrestrial juvenile stage, the eft. The eft stage lasts from 1 to 8 years, depending on temperature and food availability. Efts undergo a second trans-
AMPHIBIANS, BIODIVERSITY OF

formation into adults, which then return annually to water to breed. Larvae that encounter very favorable conditions and grow rapidly can retain some larval morphological characteristics, remain permanently aquatic, and reproduce as paedomorphic adults. Larvae that encounter slightly less favorable conditions can bypass the eft stage and metamorphose directly into the adult morphology.

Courtship in salamanders can be quite elaborate, incorporating chemical, visual, and tactile cues. During the breeding season males of aquatic species can develop enlarged fins and become more brightly colored. There may be other morphological changes to the head glands, tooth structure, musculature, and skin during the breeding season. Many salamanders have specialized glands that secrete compounds used as olfactory signals during courtship.

Salamanders have internal or external fertilization. External fertilization involves deposition of the sperm on the egg mass. Salamanders with internal fertilization either transfer sperm directly from male to female or exhibit a unique form of sperm transfer in which a bundle of sperm (the spermatophore) is deposited by males. The spermatophore is picked up in the female cloaca and stored in a special structure in the cloaca, the spermatheca. The stored sperm can remain viable until ovulation, which may occur from a few days to 2.5 years later. In live-bearing salamanders the sperm enters the oviduct, in which the eggs are internally fertilized.

Both male and female salamanders of some species exhibit parental care (Fig. 1A), although it is practiced mostly by terrestrial breeders and may serve to protect the eggs from predation, fungus infection, and desiccation. Adults attend clutches for up to 9 months. In a few species communal clutches are attended by aggrega-
tions of females, and eggs are fertilized by many different males.

C. Anurans

The most common reproductive strategy in frogs involves a complex life cycle with externally fertilized aquatic eggs which produce highly specialized larvae—tadpoles. The tadpoles grow and develop for some period in the water and then undergo a radical metamorphosis and emerge as terrestrial juvenile frogs. However, frogs have evolved a remarkable variety of reproductive strategies, most involving a trend toward removal of eggs and larvae from the aquatic environment. Pough et al. (1998) described 29 combinations of egg deposition sites, including still water and fast-flowing water, terrestrial nests, and on or in the body of the male or the female, and they also described tadpole development, ranging from the typical free-swimming feeding larva to direct development inside the oviduct. Because amphibian eggs lack a water-resistant shell, the greatest diversity of reproductive modes occurs in the humid tropics where eggs can survive for long periods in the permanently moist terrestrial environment. Complex behaviors including egg guarding and embryo transport (on the dorsum, in the mouth, and in Rheobatrachus species in the stomach), and unusual morphological structures such as skin pockets to provide protection for developing embryos, are associated with many reproductive modes. Several species bear live young, and at least one of these, Nectophrynoides occidentalis, actually provides oviductal nutrition to its developing embryos.

In tropical regions where conditions for reproduction are favorable throughout the year, breeding can be aseasonal, and females may lay multiple clutches in one year. In more temperate or high-altitude regions breeding is typically strongly seasonal, occurring only during short periods of the year when temperature and rainfall reach critical levels. Under these conditions females generally lay only one clutch of eggs each year. Male frogs vocalize mainly to attract mates (Fig. 1B) and to advertise their presence and sometimes their status to other males. Calls are species-specific and each female's brain is tuned to respond only to males of her own species. In most frogs, males possess a single or double vocal sac which serves as a resonator and in at least some species as a sound radiator. Vocal sacs also conserve energy by allowing passive reinflation of the lungs as the vocal sac contracts after a call. Some male frogs, such as the Australian torrent frog (Litoria nanos), lack a vocal sac but can still produce a surprisingly loud call. Environmental conditions such as temperature affect vocalizations. At colder temperatures notes and pulses are produced at slower rates, but the length of the call increases. The dominant frequencies in the calls of most frogs are lower than 3000 Hz, although those of some small species are higher. Within species, variation in the dominant frequencies, pulse rates, and durations of calls often reflects male body size; therefore, the call may indicate male quality as well as male location. Females of many species use these characteristics to choose their mates from among competing males. Males vocalize from species-specific locations which can be in water, on or beneath the ground, in vegetation from near ground level to high in trees, and even under water (several species including African clawed frogs, Xenopus).

Female frogs do not have a vocal sac and very few vocalize. Some female frogs produce a scream when distressed, and reproductively active females of some species call softly in response to male advertisement calls.

The posture of frogs during the fertilization of eggs is called amplexus; in most species this involves the male grasping the female from above (Fig. 1C). The exact posture adopted depends on the morphology and relative size of the male and female. The two most common positions involve the male grasping the female in front of the back legs (inguinal amplexus) or the front legs (axillary amplexus). Amplexus is aided in many species by specialized patches of skin called nuptial pads on the forelimbs of males. Pairs remain in amplexus while the male sheds sperm onto the eggs as they are released by the female. Fertilization in nearly all frogs is external but several species accomplish internal fertilization by cloacal apposition. The frog Ascaphus truei, commonly called the tailed frog, breeds in fast-flowing streams of the Northwest of North America and carries out internal fertilization using the "tail," which is actually an intromittent organ formed from an extension of the male's cloaca.

V. ECOLOGY AND FUNCTIONAL MORPHOLOGY OF LARVAL AMPHIBIANS

A. Caecilians

Relatively little is known about the larvae of caecilians. They are more similar to adults than are those of frogs or salamanders. Externally they closely resemble adults


but have gill slits and fins. Free-living caecilian larvae have long external gills and a lateral line system. Their mouth and dentition resemble those of adults. They lack the tentacle organ that appears on the head of adults; this appears at metamorphosis.

**B. Salamanders**

Salamander larvae are much more similar to adults than are the tadpoles of frogs. Larval salamanders have external gills that are not completely covered by an opeculum. Some embryonic salamanders have paired lateral projections from the head called balancers; in some species these persist for a short period following hatch ing. Most species possess well-developed fore- and hindlimbs through most of the larval period. Their bodies are laterally compressed compared to those of adults, and their tails are also relatively thinner and deeper. Their skins contain lateral line organs (neuromasts) and are thinner and less glandular than those of adults. Their dentition is different from adults, and their tongues are rudimentary. Their eyes lack lids. Larval salamanders are almost all carnivorous, usually feeding on zooplankton and larval insects. The larvae of larger species can also feed on small vertebrates. The larvae of some species have alternative morphologies; the typical morphology is usually a planktivorous carnivore, but when conditions are favorable some individuals develop relatively larger heads and more powerful jaws, adopting a “cannibal” morphology that allows them to prey on small vertebrates, including their siblings.

As in larval frogs, there is considerable variation among and within species in rates of larval growth and development. Both the minimum and the maximum rates for salamanders are slower than those for frogs. Salamanders can take from 6 weeks to 5 years to complete larval development. Within species, rates respond to both temperature and food availability, and salamanders have the additional option, apparently not available to frogs, of changing the relative rates of development of somatic and reproductive structures so that they mature sexually without losing all larval characters.

**C. Anurans**

The tadpole larva of frogs are highly specialized for growth and development in the aquatic environment. They have an oval head–body region and a long tail, which is laterally compressed and includes a central area of musculature and dorsal and ventral fins of thin, lightly vascularized tissue. The tail is supported only by a notochord. Despite their very different body form, they swim and turn as rapidly and efficiently as fishes of similar body sizes. They feed using an elaborate pumping mechanism that is very different from the oral and branchial morphology of adults. This mechanism transports water through the mouth and pharyngeal cavity, where food particles are removed by branchial filters and entrapment in strands of mucus. Some tadpoles can remove particles as small as 0.126 μm from the water. Water is ejected through the nostrils of most species and through the spiracle, which is usually a single, tubular structure leading out of the opercular chamber and can be located midventrally or on the left side of the body. The unique mouthparts of tadpoles typically include an oral disc with transverse rows of keratinized “teeth” that are used to scrape particles into suspension. Keratinized sheaths on the jaws provide cutting and biting surfaces. The oral apparatus is variously modified and sometimes allows attachment to the substrate via suction (Fig. 1D). The relatively long, coiled intestine fills most of the body cavity. Tadpoles are typically thought of as microphagous herbivores that feed on algae and small parts of higher plants, but most species will feed on animal material when it is available. Tadpoles often scavenge on dead animals in the water and frequently prey on amphibian eggs.

Tadpoles often hatch with external gills, which are quickly covered over by a fold of epithelium, forming the opercular chamber. Before the opercular chamber forms, many species do not swim but attach to a substrate using adhesive organs located posterior to the mouth. Tadpoles lack limbs at hatching. The rear legs usually develop slowly, starting as limb buds at the posterior end of the body and developing over a long period. The forelimbs develop within the opercular chamber and are visible only after they erupt fully formed through the opercular wall at the onset of metamorphosis. A typical pattern of tadpole growth and development would include 1 to a few days as a nonswimming hatching with exposed gills followed by several weeks to months as a swimming and feeding larva. During this period, the tadpole grows dramatically; many species increase their body mass by hundreds of times and some by thousands. During this period, the hindlimbs begin to grow and slowly develop through the remainder of larval life. Metamorphosis begins with the eruption of the forelimbs and involves drastic changes to all elements of the structure and function of the body. Reorganization of the mouth and digestive tract allows a switch from
larval filter feeding to adult carnivory. The tail fins and muscularia are broken down and reabsorbed. Most elements of the chondrocranium are reshaped and realigned, and the branchial apparatus ceases to be a support for gills and takes on a role as support for the adult tongue. The external and middle ears form, and in all but the aquatic frogs of the family Pipidae the lateral line system disappears. The eye changes in size and structure, the photopigments of the rods change from porphyropsin to rhodopsin, and eyelids appear. The axial skeleton is reorganized and many elements are ossified. The lungs, which develop and begin to function during the larval period of many species, enlarge and take on their role as a major respiratory structure. The complexity of the skin increases, with the epidermis increasing from two layers of cells to five or six, many with specialized functions. The kidney, which in tadpoles is relatively simple and excretes excess water and ammonia, becomes more complex to serve its new function of conserving water and excreting urea. The gonads differentiate at about the time of metamorphosis.

Rates of growth and development of tadpoles are typically highly variable within species, responding to environmental temperature, food availability, and the density of tadpoles of their own and other species. Many species that inhabit unpredictable environments, such as temporary ponds, can have larval periods from 2 weeks or less up to months. Some species regularly spend 1 year or more (maximum 3 years) as larvae. The interaction between rates of growth and development in tadpoles has produced a rich literature that examines how and why this interaction is controlled. In general, it appears that rates of growth control rates of development during the earlier part of the free-living tadpole stage (Wilbur and Collins, 1973). Larvae that are growing slowly develop more slowly than fast growers, and changes in growth rate caused by changes in environmental conditions are mirrored by alterations in developmental rate. Fast-growing tadpoles tend to metamorphose both earlier and at larger sizes than slower growing individuals. Very slow growers appear to regulate their developmental rate so that they metamorphose near a species-specific minimum size after a larval period that may vary greatly in length. This flexibility in relative rates of growth and development is lost late in the larval period at a point that may vary among taxa. Theory suggests that the regulation of these rates ultimately responds to natural selection acting in a way that depends on the relative rates of growth and survival in aquatic and terrestrial environments. These ideas have been applied to life-history transitions in many nonamphibian taxa, including fishes, insects, and plants.

Because frogs typically deposit large numbers of eggs during relatively short breeding seasons, densities of tadpoles are often high. The success or failure of a cohort of tadpoles in a typical temporary freshwater habitat depends on a highly complex and unpredictable set of factors, including the density of tadpoles of their own species and other species, which control the degree of intraspecific and interspecific competition for food and space; the number and species of predators present; and the duration of the aquatic phase of the habitat. Competition and predation are both controlled by the choice of time and place of breeding by adult frogs and to some extent by microhabitat selection within habitats by tadpoles. The outcomes of species interactions involving tadpoles can be altered by changes in the timing of breeding, and microhabitat selection by tadpoles can depend on the species and sizes of other tadpoles present.

Tadpoles are preyed on by a wide variety of vertebrate and invertebrate predators, for whom they constitute a valuable resource. Major predators include fishes, salamanders and salamander larvae, and the aquatic larvae of insects such as dragonflies, damselflies, and beetles. Vulnerability of tadpoles to predators typically decreases as the tadpoles grow and develop, and many tadpoles exhibit short-term behavioral responses to predators, such as decreasing their levels of activity or switching from midwater feeding to substrate feeding, that appear to decrease their vulnerability. Most of these responses also decrease the tadpoles’ rates of growth and development, leading to trade-offs that have been explored by behavioral ecologists.

VI. BEHAVIOR AND ECOLOGY OF POSTLARVAL AMPHIBIANS

The ecological breadth of the three classes of modern amphibians is reflected in their geographical distributions (Savage, 1973). The caecilians are restricted by both thermal and water requirements to relatively low latitudes and elevations, and they do not occur where mean annual temperatures are less than about 12°C or total annual precipitation is less than about 1000–2000 mm. The salamanders have less restrictive ecological requirements and are distributed across a broader range of habitats, occurring from low to high latitudes and
elevations. Their major limitation is clearly moisture; they do not occur in areas that have prolonged dry seasons and only rarely in areas with total rainfalls less than 1000 mm per annum. The anurans as a group can tolerate wide ranges of both temperature and water availability, and they occur at all but the highest latitudes and elevations and in all but the driest deserts. Almost all adult amphibians are carnivores that ingest invertebrate and vertebrate prey small enough to be swallowed whole. Amphibians are mostly either sit-and-wait predators or active foragers, but none engage in cursorial pursuit of prey. Sit-and-wait predators locate their prey primarily using vision, whereas active foragers often use olfaction.

Salamanders use a variety of methods for prey capture. Many larvae, and adults of aquatic taxa, are suction feeders, using rapid depression of the floor of the throat to pull in water and prey. Terrestrial salamanders usually feed by extending their large, fleshy tongues, which adhere to prey and pull it into their mouths. The tongue of salamanders is attached at the base and is protruded and elongated by muscles and fluid tension. The degree of attachment of the tongue and the length to which it is protruded vary among taxa. Captured prey are ingested whole. Terrestrial anurans also capture prey by protruding their tongues. These are usually attached near the front of the lower jaw and are protruded by literally flipping them forward and downward. Prey that adhere to the tongue are drawn into the mouth and swallowed whole.

Amphibians generally appear to grow throughout their lives, but rates of growth decline drastically after reproductive maturity is attained. In captivity, many species can live for decades, and even in nature some species live for extended periods. In general, amphibian life spans appear to be limited more by environmental hazards than by aging and senescence.

Many species of frogs and salamanders occupy relatively small home ranges during the nonbreeding season and migrate to breeding habitats for reproduction. The nonbreeding home range is aggressively defended by some species. Some species return to their natal ponds to breed, whereas others may simply migrate to a suitable body of water. In temperate regions, some species may also regularly migrate to overwintering sites. Amphibians use celestial navigation, light polarization, and the earth’s magnetic field as means of orientation during migrations. There is considerable movement among local populations in many species, particularly by anurans. Some of this movement is dispersal by juveniles, but some is due to longer range movements by terrestrial adults. At least one species of frog, Bufo marinus, can move great distances over relatively short periods. The range of the introduced Australian populations of this species has expanded by approximately 30 km per year for an extended period, and the animals that arrive first in new habitat are adults. Detailed studies of the local movements of adult B. marinus suggest that they are nomadic. Many mark-recapture studies of frogs have found high rates of disappearance of marked frogs and of the appearance of unmarked animals, suggesting that adults of many species may at least occasionally disperse to new areas.

The relatively high rates of migration between local populations found in many amphibian species suggest that groups of local populations often form metapopulations. It is important to recognize this because the dynamics of metapopulations are governed by different factors than those of local populations, and the conservation of metapopulations requires a different approach than the conservation of local populations.

Because amphibians typically produce relatively large numbers of relatively small eggs, their populations can increase rapidly in size when reproduction is successful. It is likely that populations of most amphibians normally fluctuate fairly widely over time (Alford and Richards, 1999). These normal fluctuations may include fluctuations of local populations to extinction, followed by relatively rapid reconcentration by immigrating from adjacent local populations belonging to the same metapopulation. It is likely that for many species the persistence of most local populations has little bearing on whether the regional metapopulation will persist. However, there may be a few critical local populations that either serve as reliable sources of colonists or serve as stepping stones for migration between more widely separated local populations. Identifying these local populations and conserving them will be crucial for ensuring the long-term persistence of many species.

VII. AMPHIBIAN CONSERVATION

A. Human Uses of Amphibians

Amphibians have featured prominently in many human cultures through stories, art, and poetry. In urban areas they are frequently found coexisting successfully with humans in parks and garden ponds. Frogs are an important source of protein in some subsistence cultures. In affluent countries, frogs are imported for consumption in gourmet restaurants. Hundreds of mil-
lions of frogs have been exported from Southeast Asia and the Indian sub-continent, resulting in increasing insect pest populations. Frogs have also become model organisms in ecological, embryological, physiological, and genetic research.

Amphibian skin contains a wide variety of chemicals, including complex amines, alkaloids, and polypeptides, some of which have pharmacological properties. Some skin toxins are effective against amphibian bacterial and fungal infections, and the skin of the South American frog *Epipedobates tricolor* contains a constituent, epibatidine, that blocks pain 200 times more effectively than morphine. Poison dart frogs of the family Dendrobatidae harbor many exceptionally toxic skin compounds and one species, *Phyllobates terribilis*, contains sufficient toxin in a single frog to kill several adult humans. This toxin is smeared on darts used by the Choco Indians of western Colombia for hunting monkeys and other large game. There are connections between frog diets and the presence of skin toxins; some dendrobatid frogs with high toxin levels in the wild gradually lose their toxicity when held in captivity. Although indigenous cultures have recognized the toxic and medicinal properties of frog skin for centuries, their potential for development as medicines using scientific methods has only recently emerged as a significant field of research.

**B. Amphibians as Components of Ecosystems**

Amphibians form a vital link in many food chains. They represent the highest vertebrate biomass in some ecosystems and occupy an intermediate position in the food chain. Aquatic larval amphibians are herbivorous to omnivorous (Anura) or carnivorous (Caudata and Apoda) and are significant prey items for a wide variety of vertebrate and invertebrate predators. Tadpoles in lakes and ponds often reach extremely high densities and can have a significant impact on nutrient cycling within these aquatic habitats. Because anuran larvae feed on algae and other aquatic material, they play an important role in the transfer of plant energy to predators of tadpoles. Adult amphibians feed on a wide variety of live food. Most are generalists that consume any living creature smaller than their gape size. However, some have specialized to feed exclusively on narrow ranges of food items, such as worms, ants, and even snails. Predators of amphibian eggs, tadpoles, and adults include other amphibians, spiders, insects, mammals, birds, and reptiles (especially snakes). Because of their important role in ecosystems, population declines or extinctions of amphibians may have significant impacts beyond the amphibian species affected.

**C. Amphibian Diversity and Levels of Threat**

There are approximately 5000 amphibian species (Table 1), and new species are being discovered every year. Amphibians occupy all continents except Antarctica, and they are found in habitats ranging from arid deserts and saline mangrove swamps to tropical rain forests and mountain peaks more than 4000 m high. More than 80% of amphibians are found in the tropics, with an estimated 44% of the world’s amphibians in the tropical Americas. Many tropical regions, including New Guinea, South-east Asia, and parts of northern South America, have been inadequately surveyed, and there is no doubt that many previously unknown species will be discovered in these areas.

The 1996 International Union for the Conservation of Nature (IUCN) *Red Data Book* lists 5 amphibian species as extinct and 124 as threatened, which represents 25% of the species for which assessment of conservation status has been undertaken. However, this information is influenced to some degree by the research and assessment focus in affluent countries and may change substantially when comparable research is undertaken in other geographic regions. There is little doubt that extensive habitat loss in tropical regions is causing the extinction of poorly known and undiscovered species.

**D. The Problem of Amphibian Population Declines**

Well-documented declines and disappearances of amphibian species and entire suites of species occurred over wide areas in the 1980s and 1990s. All amphibian populations fluctuate, and assessing the significance of downward trends in amphibian populations has been difficult. However, the widespread loss of species and populations in a relatively short time frame, including dramatic extinctions of previously abundant species in protected areas such as national parks, is evidence that declines are real.

Many causes have been postulated for amphibian declines, including habitat loss and modification, predation, environmental toxicity, disease, immunosuppression, ultraviolet radiation, changes in climate or weather patterns, and combinations of these. No single cause
has been identified and declines need to be assessed case by case. In some areas populations of a suite of species have declined while other ecologically similar species have not been affected. Populations of some species have declined at high altitudes but remained unaffected at low altitudes. Elucidating the causes of these declines remains a difficult and complex problem. Determining the causes of declines can be difficult. In areas of extensive urbanization, such as Europe and North America, declines or extinction of amphibian populations or species through habitat loss have clearly occurred. However, understanding rapid population declines in relatively pristine montane rain forests is a more challenging problem. A major hindrance to our understanding of population declines is the lack of information about amphibian aetiology; how populations behave, and the extent to which they operate as metapopulations. Experimental ecology aimed at testing hypotheses about population declines will be vital for identifying causal factors. The role of diseases in amphibians is poorly understood, and many diseases have only recently been documented. Although monitoring of populations over time is essential to understand population behavior, it will not identify the causes of amphibian population declines. The amphibian decline problem is currently the focus of much research effort, as reviewed by Alford and Richards (1999).

VIII. CONCLUSION: THE AMPHIBIAN SUCCESS STORY AND ITS FUTURE

Contrary to popular opinion, modern amphibians are not a relictual remnant of the ancestors of other tetrapods but are a highly successful group in their own right. There are more species of amphibians than there are of mammals. Modern amphibians occur in nearly all of Earth’s terrestrial habitats, from within the Arctic Circle to tropical deserts. Groups of modern amphibians that need to conserve water have evolved impermeable skins, cocoons, and the ability to excrete uric acid. Groups that need to breed outside water have evolved a startling array of reproductive adaptations; amphibians have the widest range of reproductive modes of any tetrapods. These include aquatic eggs and larvae, many species with terrestrial eggs, and truly viviparous species in which the mother provides nutrition in addition to the yolk during development. This diversity indicates that the typical reproductive pattern, with aquatic eggs and larvae, must not represent a constraint that has limited their success. It probably represents a successful adaptation that allows the exploitation of aquatic habitats by terrestrial species and allows a much higher fecundity than is available to species that must provision their eggs with enough yolk for complete development. The ability to respond to environmental challenges has allowed the modern amphibians to persist and flourish during and through periods of dominance of terrestrial habitats by other tetrapod groups. They have outlived the early dominant amphibians, several waves of dominance by reptiles including the dinosaurs, and many radiations and extinctions of mammals. The current concern regarding declines and disappearances of many amphibians is justified because it may be an early manifestation of a general crisis in biodiversity. It seems likely, though, that as long as terrestrial habitats continue to accommodate vertebrate life, some amphibians will persist.

See Also the Following Articles

Birds, Biodiversity of • Fish, Biodiversity of • Mammals, Biodiversity of • Reptiles, Biodiversity of

Bibliography


ANTARCTIC ECOSYSTEMS

Peter Convey
British Antarctic Survey, Natural Environment Research Council

I. Antarctica
II. Antarctic Terrestrial Biota
III. Present-day Biodiversity
IV. Colonization Processes
V. Future Trends

GLOSSARY

ablation Direct transfer of water molecules from ice crystals to the vapor phase, without transition through the liquid phase.
anhydrobiosis Ability of certain organisms to lose all detectable water from their bodies under certain conditions, yet remain viable when subsequently rehydrated.
cryoturbation Process by which soil particles and stones are moved and mixed by the frequent formation and subsequent melting of ice crystals in the soil column.
cryptogam Plants without apparent reproductive organs; plants that reproduce by spores.
fellfield Dry, windswept habitat of cold regions comprising mineral soil, gravels, stones, and rock, dominated by cryptogams (especially mosses and lichens) and, outside the Antarctic, by compact cushion-forming phanerogams and short grasses.
microclimate The climate experienced over a scale of millimeters to centimeters, within which the major-
I. ANTARCTICA

A. Introduction to the Continent

Antarctica is distinct among the Earth's continents, not least because it is the only one without a long-term history of human occupation and influence. It is a large continent, with greater surface area than Australia. However, 99.5% of the continental area is covered permanently by snow or ice. Terrestrial ice-free ecosystems are limited to exposed nunataks, cliffs, and seasonally exposed snow- and ice-free areas, with habitats generally being more extensive and exposure of longer duration on a ring of sub-Antarctic islands surrounding the continent. Ice-free terrestrial habitats are patchily distributed over the continent, and are best represented in coastal regions, particularly along the Antarctic Peninsula, and in the major mountain ranges inland. Many ice-free areas are necessarily small in extent and isolated from others by distances of tens to hundreds of kilometers. However, in the McMurdo Dry Valleys region of southern Victoria Land, ice-free areas of several hundred square kilometers exist. Antarctica is isolated from other continental land-masses by the 1000-km-wide Drake Passage south of Tierra del Fuego and by 4000–5000 km of the Southern Ocean that lies between it and Australia and South Africa.

Antarctica plays an important controlling role in the world climate, through its reciprocal influences on atmospheric and oceanic circulation patterns. For much of the year, Antarctica maintains a negative energy balance, losing heat energy to the atmosphere. This acts to maintain the continental ice cap, which on average is 2 km, and in places up to 4 km, deep. This ice mantle gives Antarctica the highest average altitude of any continent. It is also the coldest continent, having recorded the lowest surface temperature on Earth (−89.6°C), and is the windiest. Despite the fact that the vast continental ice cap develops from snow precipitation, large areas of Antarctica are classified as cold deserts, with extremely low precipitation rates (much of which is directly lost by ablation, rather than entering terrestrial habitats). There is glaciological evidence supporting wide fluctuations in glacial extent throughout the Pleistocene and even in the last few centuries, parallel to widely fluctuating global climate sources of heat energy and leading to gradual cooling and formation of the continental ice cap. The extent of glaciation has varied widely both geographically and over time during the last 20 million years, with evidence of local warmer periods as recently as 1–3 million years ago being sufficient for the development of areas of cool temperate rain forest dominated by southern beech (Nothofagus spp.).

The Pleistocene (Ice Age) glaciation mirrored that experienced in the Northern Hemisphere. At its maximum, the continental ice cap was considerably thicker than at present, and the continent was surrounded by extensive floating ice shelves. The island groups of the South Shetlands, South Orkneys, and South Georgia were similarly the centers of large ice caps that extended tens of kilometers into the surrounding ocean. At this time, terrestrial habitats would probably have been limited to isolated inland nunataks; most, and probably all, current lowland (coastal) habitats would have been obliterated by ice. Bathymetric models suggest that, even allowing for lower sea levels, floating edges of ice shelves would not have been interspersed with exposed terrestrial habitats. There is glaciological evidence supporting wide fluctuations in glacial extent throughout the Pleistocene and even in the last few centuries, parallel to widely fluctuating global climate sources of heat energy and leading to gradual cooling and formation of the continental ice cap. The extent of glaciation has varied widely both geographically and over time during the last 20 million years, with evidence of local warmer periods as recently as 1–3 million years ago being sufficient for the development of areas of cool temperate rain forest dominated by southern beech (Nothofagus spp.).

The Pleistocene (Ice Age) glaciation mirrored that experienced in the Northern Hemisphere. At its maximum, the continental ice cap was considerably thicker than at present, and the continent was surrounded by extensive floating ice shelves. The island groups of the South Shetlands, South Orkneys, and South Georgia were similarly the centers of large ice caps that extended tens of kilometers into the surrounding ocean. At this time, terrestrial habitats would probably have been limited to isolated inland nunataks; most, and probably all, current lowland (coastal) habitats would have been obliterated by ice. Bathymetric models suggest that, even allowing for lower sea levels, floating edges of ice shelves would not have been interspersed with exposed terrestrial habitats. There is glaciological evidence supporting wide fluctuations in glacial extent throughout the Pleistocene and even in the last few centuries, parallel to widely fluctuating global climate sources of heat energy and leading to gradual cooling and formation of the continental ice cap. The extent of glaciation has varied widely both geographically and over time during the last 20 million years, with evidence of local warmer periods as recently as 1–3 million years ago being sufficient for the development of areas of cool temperate rain forest dominated by southern beech (Nothofagus spp.).

C. Recognized Biogeographic Zones

Antarctic terrestrial environments include those found on the continent itself and also on a ring of Southern Ocean islands south of approximately 50° latitude (Fig. I). In this area terrestrial ecosystems with widely varying characteristics are to be found and a number of systems of classification have been proposed (see R. I. L. Smith, in Laws, 1984; Longton, 1988), some applying only to the Antarctic and others attempting to relate similar zones in both Antarctic and Arctic polar latitudes. Currently, three biogeographical zones are recognized: the sub-Antarctic, the maritime Antarctic, and the continental (or frigid) Antarctic. Although elements
of their biota and environmental characteristics overlap, ecosystems of the three zones are distinctively different.

1. Sub-Antarctic

The Sub-Antarctic zone consists of a ring of isolated sub-Antarctic islands and island groups lying at high latitudes in the Southern Ocean. With the exceptions of South Georgia, Heard Island, and McDonald Island, these lie either close to or north of the "Polar Frontal Zone" (formerly known as the Antarctic Convergence), a circumpolar oceanic feature formed where cold Antarctic surface waters sink below warmer sub-Antarctic waters. Sub-Antarctic island climates are strongly oceanic and, with the occasional exception of South Georgia, they are not influenced by seasonal pack or fast ice. Temperature variation is reduced by the surrounding cold ocean, with mean air temperatures of most being low but positive year-round, and precipitation is high (Table I).

Some islands remain extensively glaciated (South Georgia and Heard Islands) and formed the center of much larger ice caps extending well away from current coastlines at the height of Pleistocene glaciation. However, the majority are now or have been only partially glaciated. The age of these islands and availability of terrestrial habitats vary widely. South Georgia is composed of continental crust of Upper Jurassic/Lower Cretaceous age, once attached to the southern Andes. Kerguelen may be up to 100 million years old, but has never had a continental association, while Macquarie Island is formed from a segment of oceanic crust raised above the surface 3–5 million years ago. The remaining groups of Crozet, Prince Edward, and Marion Islands are younger, the latter apparently of fairly recent origin (ca. 300,000 years) and the former of Miocene origin. Some authors also include more northerly island groups (Falkland Islands, Gough Island, Iles Amsterdam and St. Paul, and New Zealand’s outlying groups...
TABLE I
Typical Air and Microhabitat Temperature Ranges Experienced in Summer and Winter in Each of the Antarctic Biogeographical Zones

<table>
<thead>
<tr>
<th>Antarctic zone</th>
<th>Months with positive mean air temperatures</th>
<th>Air temperature range (°C)</th>
<th>Microhabitat temperature range (°C) (exposed habitat minima in parentheses)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Antarctic</td>
<td>6 to 12</td>
<td>–2 to +25</td>
<td>–5 to +30; &gt;–2 (–10)</td>
</tr>
<tr>
<td>Maritime</td>
<td>1 to 4</td>
<td>–10 to +15</td>
<td>–5 to +30; –20 to +3 (–43)</td>
</tr>
<tr>
<td>Antarctic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continental Antarctic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coastal</td>
<td>0 to 1</td>
<td>–20 to +10</td>
<td>&lt;–10 to +30; (–37)</td>
</tr>
<tr>
<td>Inland (deserts and nunataks)</td>
<td>0 to 1</td>
<td>–15 to +9</td>
<td>–20 to +12; –41 to –16</td>
</tr>
</tbody>
</table>

* Few data are available from microhabitats in the continental Antarctic, particularly during winter; considerably more negative minima would be expected in sites without snow cover.

of Snares, Campbell, Bounty, and Chatham) either within the sub-Antarctic or in a separate “mild Antarctic” classification. More correctly these should be classified as southern cool temperate or ocean temperate, containing faunal and floral groups not otherwise represented in the sub-Antarctic.

2. Maritime Antarctic
The maritime Antarctic also experiences significant oceanic influence on climate, but on a more seasonal basis than in the sub-Antarctic. The zone includes the western coastal regions of the Antarctic Peninsula to Alexander Island (ca. 72°S), the Scotia Arc island groups of the South Shetland, South Orkney, and South Sandwich Islands, and the isolated islands of Bouvetøya and Peter I Øy (see Fig. 1). Although in a biological sense the South Sandwich archipelago and Bouvetøya are within the maritime Antarctic zone, they are unusual in being geologically recent (1–3 million years old) volcanic islands. Most show evidence of recent or current volcanic activity, with unique biological communities associated with limited areas of geothermally warmed ground. Together with analogous areas on Deception Island (South Shetland Islands) and Mounts Erebus, Melbourne, and Rittmann (Victoria Land, continental Antarctic), such geothermally associated biota are exceptional within Antarctica.

Maritime Antarctic mean air temperatures are marginally positive for 1–4 months of the year; however, both summer maxima and winter minima are buffered to some extent by the surrounding ocean, the surface temperature of which varies annually between ca. −2 and +1°C (see Table I). Short periods of positive temperature and thaw are experienced in all winter months. Precipitation is generally high through a combination of oceanic influence and typical westerly airflow, but varies widely between sites and throughout the year.

3. Continental Antarctic
The continental Antarctic comprises most of the continental area of Antarctica, including all of East (or Greater) Antarctica, the Balleny Islands, and the eastern side of the Antarctic Peninsula. With the exception of the extensive ice-free cold desert region of the McMurdo Dry Valleys, terrestrial habitats of this zone are of limited extent and great isolation. In addition to the cold deserts, they include two basic habitat classes—exposed coastal regions superficially similar to those of the maritime Antarctic and inland nunataks. Temperature regimes are more extreme than those of the maritime Antarctic, with positive mean air temperatures achieved for <1 month in coastal locations and never inland. Air temperatures rarely if ever become positive even for short periods and, unlike in the maritime Antarctic, never in winter.

As in the other zones, microhabitat temperatures do not track air temperatures closely, the two most significant influences being warming due to absorption of solar radiation in summer and protection from winter extremes by snow cover. However, many continental Antarctic habitats also experience a period of prolonged darkness with no thermal input. Few data are available, but those habitats with no winter protection (e.g., exposed mountain ridges and cold deserts) are expected to track closely the winter air temperature patterns and minima.
II. ANTARCTIC TERRESTRIAL BIOTA

A. Habitats

In the maritime and continental zones, with the exceptions of steep cliffs and exposed mountain ridges, most terrestrial habitats are covered seasonally by snow and/or ice, which has the benefit of buffering them from extreme temperature lows and fluctuation, and also from wind abrasion. There is an important difference between the sub-Antarctic and other zones. While habitats in the latter zones may be free of seasonal snow cover for periods of only weeks up to perhaps 5 months (those in the most extreme continental sites may not be exposed at all in some seasons), many sub-Antarctic islands experience intermittent or no periods of snow cover, which is often restricted to higher altitudes. There is much variation, but even on the coldest islands such as South Georgia (with up to 6 months of annual snow cover in lowland sites), mean air temperatures do not drop far below zero, and sub-nivean microhabitat temperatures are often sufficient to allow year-round invertebrate activity.

Overwinter microhabitat temperatures in maritime and continental zones are normally low enough to arrest physiological processes and activity. For a period in the austral spring, absorption of solar energy by soils or plant communities may be sufficient to allow the formation of a sub-nivean air space. This acts almost as a “greenhouse,” within which temperatures are high enough to allow positive levels of net photosynthesis by microflora, lichens, and mosses and feeding activity by invertebrates (Convey, 1996; B. Schroeter et al., in Lyons et al., 1997). This feature effectively increases the active season length for these groups.

Antarctic soils are generally poorly developed, with low organic content. Again, there is a clear dichotomy between soils of the sub-Antarctic, and soils in the other two zones, with only the latter possessing a widespread and permanent permafrost layer. Brown soils, resembling those in temperate regions, are widespread in the sub-Antarctic zone but are associated only with the larger stands of higher plants (phanerogams) in the maritime Antarctic; they are not present in the continental zone. Likewise, deep peat deposits have developed under extensive valley bog communities in the sub-Antarctic, and have been radio-carbon-dated as forming soon after the end of the Pleistocene glaciation (10,000–20,000 years b.p.). Moss peat deposits are much more restricted in the maritime zone (5000–6000 years b.p.) and are not found in the continental Antarctic. Particularly in the maritime and continental zones, the formation, development, and stability of soils are heavily influenced by cryoturbation processes, leading to large areas of patterned ground (including soil/stone polygons, stripes, and solifluction lobes).

B. Terrestrial Fauna

Terrestrial vertebrates form a very small element of the natural fauna of Antarctica, and these few are limited to sub-Antarctic islands. They include a single endemic insectivorous passerine and freshwater ducks on South Georgia and/or Kerguelen. There are many records of vagrant birds, particularly from sub-Antarctic islands but also from the maritime Antarctic. The majority of these have obviously been carried far off course and succumb rapidly to the extreme conditions or predators. There are no naturally occurring mammals, reptiles, or amphibians, although human contact has led to the deliberate or accidental introduction and naturalization of a variety of mammals (see the following).

In contrast, marine vertebrates are plentiful in Antarctica, and penguins, seabirds, and seals spend significant periods ashore to breed, rest, and molt; often in colonies comprising tens of thousands or even millions of individuals. Most colonies are coastal, but birds such as the Antarctic and snow petrels and south polar skuas also breed on nunataks up to several hundred kilometers inland. These concentrations of marine vertebrates have direct influences on local terrestrial ecosystems through trampling and deposition of guano and carcasses. Aerosol dispersal of chemical nutrients from guano extends the influence of these colonies far beyond their physical boundaries.

In the absence of terrestrial vertebrates, the faunas of each Antarctic zone consist mainly of invertebrate groups (Table II). Those of sub-Antarctic islands include representatives of “higher” pterygote (winged) insects, although many of these show behavioral or morphological (brachyptery and aptery) traits that remove the ability to fly—a feature necessary to reduce the risk of being dispersed away from favorable habitats by frequent high winds. Many invertebrate groups well known in temperate areas are not found in the sub-Antarctic (e.g., Odonata and Trichoptera) or are represented by very few species (e.g., Isopoda, Araneae, Lepidoptera, Hymenoptera, and Hemiptera). The most diverse higher insect groups are Diptera and Coleoptera. Molluscs and annelid worms (Enchytraeidae) are also present in sub-Antarctic ecosystems, but there are very few records and no detailed studies available. Diverse communities of micro-arthropods (Acari and Collembola) and micro-invertebrates (Nematoda, Tardigrada,
TABLE II
Dominant Biotic Components of Typical Ecosystems of Each of the Antarctic Biogeographical Zones

<table>
<thead>
<tr>
<th>Zone</th>
<th>Flora</th>
<th>Fauna</th>
<th>Microflora</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Antarctic</td>
<td>Phanerograms and cryptograms; open-fellfield (cf. maritime zone) at higher altitude; alien species</td>
<td>Arthropods (esp. Insecta); microarthropods (Acari, Collembola); microinvertebrates (Nematoda, Tardigrada, Rotifera); alien vertebrates and invertebrates</td>
<td>Various groups</td>
</tr>
<tr>
<td>Maritime Antarctic</td>
<td>Closed and open cryptogamic communities; phanerograms very limited extent</td>
<td>Microarthropods (Acari, Collembola); microinvertebrates (Nematoda, Tardigrada, Rotifera); Diptera very limited</td>
<td>Algal and cyanobacterial mats; foliose algae; protozoa</td>
</tr>
<tr>
<td>Continental Antarctic</td>
<td>Bryophytes and lichens, of very limited extent</td>
<td>Microinvertebrates (esp. Nematoda, Tardigrada, Acarida, Collembola) more limited</td>
<td>Algal and cyanobacterial mats; protozoa; endolithic fungi; algae, and cyanobacteria</td>
</tr>
</tbody>
</table>

and Rotifera) are also present in this zone, but have generally received less critical attention than the larger and more obvious groups. Maritime Antarctic faunas include fewer taxonomic groups. The only higher insects represented are two chironomid midges (Diptera). The most-studied members of ecosystems in this zone are free-living soil microarthropods (Acari and Collembola), although micro-invertebrate groups (Nematoda, Tardigrada, and Rotifera) are also well represented numerically. All of these groups are known by a small number of species (see the following section), but population densities are often high, in the range 10^1–10^2 individuals m^-2, comparable with or greater than such densities in many temperate ecosystems. The arthropod fauna is subject to very low predation pressure, with a maximum of only two predatory mite genera (including three species) present in any ecosystem. Food webs therefore have a very simple structure. Microinvertebrate groups include species from genera with characteristic trophic preferences (e.g., algivory, bacterivory, fungivory, and nematophagy), however, the specific dietary preferences of Antarctic taxa are virtually unknown.

The terrestrial fauna of the continental Antarctic represents a further simplification of that found in the maritime zone. No higher insects are present, and microarthropods only reach the dominance and abundance seen in the maritime zone in the much more limited areas of vegetation and (vertebrate) nutrient enrichment. Instead, in the continental zone microfauna such as nematodes assume dominance. In the most extreme continental cold deserts, simple food webs consist of as few as 1–3 nematode species, only one of which may be predatory (Freckman and Virginia, 1997). These cold desert soils are faunistically the most sterile on Earth.

C. Terrestrial Flora
Changes in floral composition and patterns of diversity between the three Antarctic biogeographical zones mirror those of the faunal communities (see Table II). Smith (in Laws, 1984) provided the most authoritative description and comparative treatment of Antarctic floras, and it remains valid despite many subsequent advances in detailed taxonomic knowledge. The flora of the sub-Antarctic is richest and most diverse, with well-developed phanerogamic and cryptogamic floras. These show strong affinities with either South American (Fuegian) or Australasian (New Zealand) floras. Trees and shrubs are not present (with the exception of the diminutive dwarf shrub Coprosma on Macquarie Island) and attempts to introduce several species to sub-Antarctic islands have been unsuccessful.

Sub-Antarctic vegetation can be described as "tundra-like." Although the vegetation is superficially similar, and climate, land-form, and pedology are comparable to those of Arctic tundra, the two regions differ in several important respects. Sub-Antarctic soils do not contain a permafrost layer, and the only woody plants are dwarf-shrub-like rhizomatous herbs (Acaena). The vegetation is dominated by large rosette-forming herbs ("megaherbs") and tall grasses, and includes a significant proportion of ferns. Phanerogamic vegetation is best developed in low-altitude coastal regions. As conditions worsen rapidly with increasing altitude, these plants are progressively replaced by a cryptogamic fell-
field flora that is closely similar (and sharing many species) to that found in the more extreme maritime Antarctic. The most favorable habitats of the maritime Antarctic are dominated by closed cryptogamic communities of carpet- and turf-forming mosses. Development of extensive vegetation is limited to a narrow altitude range (up to ca. 150 m) in coastal regions. Beyond this range, in inland areas, at higher altitudes, and at more exposed coastal sites, more open fellfield communities occur, consisting of several sub-formations of cushion- and turf-forming mosses, liverworts, and crustose, foliose, and fruticose lichens. Phanerogams are represented by only two species (hairgrass, Deschampsia antarctica, and pearlwort, Colobanthus quitensis), which are also found in the sub-Antarctic and across a wide latitudinal range in the South American Andes. Both are particularly sensitive to current climate amelioration in the Antarctic Peninsula region (see the following) and are undergoing rapid population increases.

Cryptogam-based maritime Antarctic vegetation is fragile and, lacking roots, sensitive to physical disturbance, for example, in the vicinity of penguin and seal colonies. Such areas tend to be dominated by the foliose alga Prasiola crispa. One consequence of the rapid recovery of the Antarctic fur seal (Arctocephalus gazella) populations in this century, following their near extinction through hunting and similar drastic human-mediated reductions in Southern Ocean cetacean populations (with which the seals are thought to compete for food), has been the occupation of new summer resting and molting sites within terrestrial habitats of the maritime Antarctic. The cryptogamic communities of these areas have been unable to withstand the twin pressures of excessive seal trampling and nutrient enrichment and have largely been destroyed in many coastal lowlands. Continental Antarctic vegetation is categorized within the same system as that of the maritime zone, although lichen sub-formations predominate. No phanerogams are present, the extent of closed cryptogamic communities is much more limited, and peat formations are absent.

**D. Microbial Systems**

Microbial autotrophs form the basis of polar terrestrial ecosystem processes (Vincent, 1988; Friedmann, 1993; Wynn-Williams, 1996), playing pivotal roles in the processes of primary colonization and stabilization of mineral soils, which allow secondary colonization and succession by other microbiota, plants, and Metazoa. The most significant groups involved in primary colonization are autotrophic cyanobacteria and algae, which enable secondary invasion by bacteria, fungi, and protozoa (a polyphyletic group whose most significant members in Antarctica are heterotrophic flagellates, gymnamoebae, testate amoebae, and ciliates). At the level of microbial ecosystems, there is not a sharp divide between “terrestrial” and “aquatic” (stream/lake) habitats. In addition to the largely edaphic (on or within soils), epiphytic (on surfaces of living plants and lichens), and epilithic/hypolithic (on exposed surfaces or undersurfaces of rocks) habitats typically occupied by faunal and floral communities, microbial ecosystems may also be found in cryophilic (between ice crystals in melting snow) and endolithic (within the surface few millimeters of rock matrix) habitats. The latter habitat is further divided into chasmooendolithic (within fissures and cracks open to the rock surface) and cryoendoendolithic (within tiny cavities of the rock matrix). These cryptic habitats, as found in the cold deserts of continental Antarctica, are thought to represent one limit to biological existence on Earth, and as such they have been proposed as models to assist the development of exobiological assessment techniques (Wynn-Williams, 1996). Microbial ecosystems present a visibly greater dominance in more extreme terrestrial habitats, although the same groups are present and important in all three Antarctic zones. Groups such as algae and cyanobacteria may become visible via the formation of filaments and mats, both within water bodies, such as streams and lakes, and on/in the surface layers of damp soils. Such communities are well represented within the maritime zone, and are often a climax community of the continental zone (see Table II). In the most extreme cold desert habitats of the latter zone, no detectable life survives on the surface of soils or rock. In these habitats, microbiota (algae, cyanobacteria, and fungi) have retreated to the endolithic niche, with cells existing in the interstitial spaces between rock crystals.

**E. Freshwater Systems**

Freshwater ecosystems of the Antarctic possess a very simple structure when compared with those of the Arctic and lower latitudes. Perhaps most striking is the absence of fish. The faunas of sub-Antarctic fresh waters include a single predatory diving beetle (found only on South Georgia), several Crustacea (cladocerans, copepods, ostracods, and anostracans), adventitious microarthropods, nematodes, tardigrades, rotifers, and protozoans. Ecosystem structure in the maritime Antarctic
is even more simplified: the plankton include one copepod (*Boeckella poppei*) and one anostracan (*Branichinecta gaini*) herbivore, the predatory copepod *Parabroteas sarsi*, benthic cladocerans and ostracods, and the microscopic groups. Two Crustacea (*Daphniopsis studeri* and *Acanthocyclops mirnyi*) are recorded from the continental zone.

The trophic impact of metazoan groups in lakes of all three Antarctic zones is currently thought to be minimal, although data are lacking and there is disagreement over the importance of metazoan predators in the sub-Antarctic lakes of South Georgia. With this exception, top-down grazing control is reduced and many ecosystems are thought to be driven by bottom-up forces. In particular, the “microbial loop” assumes great importance—a microbial food web consisting of phytoplankton, bacteria, and protozoans (Laybourn-Parry, in J. Lyons et al., 1997).

Lakes in the maritime and continental zones are seasonally or permanently covered by ice. Some continental lakes are meromictic or hypersaline. Unlike terrestrial ecosystems, some lakes are ancient systems thought to be hundreds of thousands of years old. At the extreme, Lake Vostok, a 200-km-long, 500-m-deep lake beneath at least 3 km of the continental ice cap, provides the intriguing possibility of harboring “ancient” microorganisms and other groups that have been effectively isolated since the formation of a permanent ice barrier (Karl et al., 1999).

### III. PRESENT-DAY BIODIVERSITY

#### A. Within Antarctica

Attempts to catalog and compare the biodiversity of the three Antarctic biogeographical zones are hindered by two fundamental problems: a lack of adequate (or comparable) sampling coverage, and taxonomic uncertainty (particularly the likelihood of extensive synonymy), both of which apply in varying extent to all groups encountered. Currently, sufficient data do not exist to allow rigorous comparisons for any microbial groups, beyond general evidence for lower diversity in the continental versus maritime and sub-Antarctic zones. The taxonomy of prokaryotic microbiota is poorly documented anywhere in the world, and the recent steps to develop molecular phylogenies are hard to reconcile with earlier classification systems, although efforts are accelerating. Both in the Antarctic and worldwide, a small proportion of prokaryotes have been described, with virtually no knowledge of species distribution. Thus, although it is possible to conclude from molecular comparisons that “the majority of Antarctic prokaryotes diverged from their nearest known non-Antarctic relatives long before a stable ice-sheet developed in Antarctica,” it is not possible to estimate the time of colonization of the continent (Frantzen, in Wynn-Williams, 1996).

The biodiversity of groups for which there are reasonable data in each of the Antarctic zones is summarized in Tables III and IV. These data illustrate a general trend of reducing diversity, and loss of specific faunal and floral groups, along a transect of increasing environmental extremes from the sub-Antarctic to the continental Antarctic. Although there have been several proposals of similar trends occurring within zones (particularly along the Antarctic Peninsula within the maritime Antarctic), the true picture appears to be more complex, with biodiversity being related to the presence of suitable microhabitats, which themselves become reduced in extent at higher latitudes.

#### B. Bipolar Comparison

Environmental conditions experienced at high northern polar latitudes may be compared broadly with those of...
### Table IV

<table>
<thead>
<tr>
<th>Group</th>
<th>Sub-Antarctic</th>
<th>Maritime Antarctic</th>
<th>Polar Antarctic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Protozoa</strong></td>
<td>83</td>
<td>ND</td>
<td>13</td>
</tr>
<tr>
<td><strong>Rotifera</strong></td>
<td>ND</td>
<td>ND</td>
<td>13</td>
</tr>
<tr>
<td><strong>Tardigrada</strong></td>
<td>&gt;20</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td><strong>Nematoda (Oligochaeta)</strong></td>
<td>22</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td><strong>Mollusca</strong></td>
<td>ND</td>
<td>ND</td>
<td>13</td>
</tr>
<tr>
<td><strong>Nematoda (nonmarine)</strong></td>
<td>48</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td><strong>Crustacea (terrestrial)</strong></td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Crustacea (terrestrial but including meromictic lakes)</strong></td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Insecta (total)</strong></td>
<td>210</td>
<td>35</td>
<td>49</td>
</tr>
<tr>
<td><strong>Collembola</strong></td>
<td>39</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td><strong>Millepoda</strong></td>
<td>61</td>
<td>23</td>
<td>34</td>
</tr>
<tr>
<td><strong>Dipera</strong></td>
<td>44</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Colleopera</strong></td>
<td>40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Arachnida (total)</strong></td>
<td>167</td>
<td>36</td>
<td>29</td>
</tr>
<tr>
<td><strong>Araneida</strong></td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Pseudoscorpiones</strong></td>
<td>140</td>
<td>36</td>
<td>29</td>
</tr>
</tbody>
</table>


**Large changes are likely with future research due to current lack of sampling coverage, expertise, and/or synonymy.**

**ND, number of representatives of group unknown.**

the sub-Antarctic and maritime Antarctic zones, although with the exception of the Greenland ice cap, no Arctic climatic equivalence with the continental Antarctic is possible. Terrestrial habitats of the Arctic consist of the northern fringes of large continental landmasses surrounding a shallow ocean, whereas the Antarctic is itself a continental mass surrounded and isolated by a large extent of cold ocean. This simple geographical difference is sufficient to drive the large climate differences found between the two polar regions at any specific latitude.

Geography also leads to a fundamental difference in colonization patterns between the two regions. With continuous southwards land connection to North America, Europe, and Asia, colonization of Arctic regions following post-Pleistocene deglaciation was a simple process. In contrast, the extreme isolation of terrestrial and freshwater habitats in all Antarctic zones from plausible refugia or other continental sources of colonists, combined with the likely obliteration of all coastal habitats during glaciation, led to much lower colonization rates. As a result, contemporary levels of diversity are one or more orders of magnitude lower in the Antarctic than the Arctic. This statement holds true for many faunal, floral, and microbial groups (e.g., Longton, 1988; Danks, 1990; Wynn-Williams, 1996; Fogg, 1998).

### C. Origin and Antiquity of Biota

As mentioned earlier, low-lying coastal terrestrial habitats of the maritime and continental zones are likely to have been obliterated during glaciation. Even in the sub-Antarctic, where some islands experienced incomplete glaciation, many terrestrial ecosystems currently based around complex plant communities did not exist more than 10,000–20,000 years ago, as evidenced by glaciological models and maximum radiocarbon dates obtained from peat deposits.

Because a large proportion of extant biota in all three zones is found in low-lying coastal areas, it is unlikely that they could have survived glacial maxima in refugia in situ (or in higher-altitude nunatak sites where, even now, they are not represented). Rather, their presence must be a result of post-glacial colonization from refugia on southern oceanic islands and/or continental sources. Identification of such refugia is at an early stage, and is restricted by the general lack of detailed biogeographical data for most groups, including a lack of relevant information from potential source regions of Southern Hemisphere continents.

As recent colonists, there is little evidence of evolutionary adaptation by biota to the extreme environmental challenges posed by the Antarctic. Even though many studies demonstrate the ecophysiological abilities of Antarctic organisms (see Block, 1990), all features identified to date are plesiotypic for the taxonomic group concerned. Their possession appears to be a prerequisite for colonization and establishment rather than an evolutionary response to conditions experienced. However, at least among the Acari, Collembola, and Nematoda, species endemic to one or more of the Antarctic zones are present, indicating that evolution has occurred since colonization.

Not surprisingly then, the biota of Antarctic zones is most closely related to that of other Southern Hemisphere continents. However, a number of examples of plant and microbial groups with bipolar or alpine distributions are known (see Longton, 1988; P. A. Broady, in Wynn-Williams, 1996). These are groups with known ability to disperse via aerial propagules. Gross similarities between soil faunal communities of both polar regions...
(the dominance of herbivorous/detrivorous Acari, Collembola, and other micro-invertebrates) simply represent a consequence of the generalist function of these groups in soil ecosystems worldwide. Nunataks and inland ice-free areas in the continental Antarctic have, in contrast, provided evidence of the continued existence of a relict fauna predating Antarctic glaciation (Acari, via patterns of specific and generic endemism; Marshall and Pugh, 1996), and biological support for the sequence of events involved in the breakup of the Gondwanan supercontinent (Tardi-grada, McInnes and Pugh, 1998).

IV. NATURAL COLONIZATION PROCESSES

A. Current Colonization

Aerobiological sampling programs carried out in Antarctica demonstrate the presence of viable colonizing propagules in the aeraspora, from both ‘local’ (Antarctic) and distant sources. Propagule densities are several orders of magnitude lower than obtained in comparable temperate and tropical studies. Most such records refer to microbial or lower plant groups (bacteria, cyanobacteria, fungi, algae, and bryophytes) and lichens; there are no verified records of the capture of living invertebrates from the air column.

Direct observational evidence of continuing contemporary colonization on both short- and long-distance scales, is given by the discovery of developing populations on previously barren ground, and of previously-unrecorded species in known sites. Particularly compelling evidence is provided by the colonization of geothermally-heated ground associated with the few, very isolated, volcanically active sites spread around Antarctica. These sites are recent, short-lived, and of limited area. Their bryophyte (and presumably algal and microbial) communities contain species not found elsewhere in the Antarctic, with postulated ancestral populations at much lower latitudes in South America or New Zealand. Evidence for recent natural colonization events involving the Antarctic terrestrial fauna is mostly lacking or equivocal. The dominant groups (in terms of diversity) in all three zones are members of the soil fauna, including micro-arthropods (mites and springtails) and other micro-invertebrates (nematodes, tardigrades, rotifers, etc.). All of these are small, inconspicuous, and easily overlooked, with many Antarctic distributional records being the results of the specific collecting efforts of individual taxonomic experts. Thus, for any of these groups, the detail of coverage both within and between Antarctic zones is very patchy, and therefore confirmation of any record as a ‘new’ colonization event is virtually impossible.

Several sites in the maritime Antarctic and in Victoria Land (continental Antarctica) have hosted individual studies since the 1960s that are detailed enough to provide baseline micro-arthropod biodiversity data. However, with the exception of two detailed studies of the maritime Antarctic South Sandwich Islands separated by 33 years, no sites have received sufficient subsequent attention to allow comparisons to be made. This isolated volcanic archipelago possesses an entirely colonized micro-arthropod fauna, with no endemic species. Comparison of the two studies of this fauna is illuminating, as both identified several taxa of sub-Antarctic (South Georgian) origin, but very few of these were in common. In particular, species of oribatid mite (Edwar-deytes and Austropia spp.) recorded in the earlier study and springtail (Cryptopygus caccus) in the latter were sufficiently widespread not to have been overlooked by either. This ‘crusoe’ pattern of records is suggestive of frequent colonization events and short-term establishment, followed by a high probability of extinction.

B. Mechanisms of Colonization

Colonization mechanisms may be classified into (1) directed active movement, (2) assisted transport by other species or nonliving debris, and (3) passive transport by air or water currents.

No members of the invertebrate Antarctic terrestrial fauna show evidence of directed migratory behavior and the first option is discounted in explaining current or short-term future patterns of Antarctic biodiversity. However, several vertebrate and invertebrate species with well-known active dispersal characteristics are being increasingly recorded as vagrants at sub-Antarctic and even maritime Antarctic locations, where arriving individuals may survive for long periods. Indeed, several Diptera and Lepidoptera have become established on sub-Antarctic Marion Island in recent years, and it is reasonable to expect other dispersing species to follow.

Assisted transport of non-parasitic invertebrates, plant seeds, and other propagules by vertebrates (birds), other invertebrates, and debris (e.g., driftwood) has been proposed as an important mechanism of Antarctic colonization by several authors, but no direct evidence exists to support this route for groups other than microbes. Transport is most likely to be rapid and via bird
species that spend time associated with land-masses during the austral winter, such as skuas, gulls, and sheathbills (thus coming into contact with terrestrial biota), rather than marine mammals (seals) and birds that spend the winter period wholly at sea (penguins, albatrosses, and smaller petrels).

Passive transport in either air or water currents is also an attractive option, but again with little supporting evidence available for some groups. Some arthropods, such as oribatid mites (Acari) and springtails (Collembola), show rafting behavior or high tolerance of submersion in sea-water, and have large numbers of littoral or marine taxa. Others (e.g., prostigmatid mites) show no such ability. Transport in the air column presents the twin challenges of low temperature and desiccation. These conditions are most likely to be survived by organisms with specific dispersal (lichen soredia and moss spores) and/or resistant stages (tardigrade "tuns," anhydrobiotic nematodes, rotifers, and protozoan cysts), and so air transport is unlikely to be a viable mechanism of fortuitous transport for other groups.

C. Dispersing Propagules

Investment in features favoring dispersal is not a characteristic of the "adversity-selected" life histories typical of Antarctic terrestrial plants and animals. Rather, features of these life-history strategies include factors such as extended life histories, low reproductive output, investment in survival features, and development of behavioral and morphological traits that reduce the chance of "accidental" dispersal. Nevertheless, all Antarctic organisms require the ability to disperse over a range of scales, from the millimeters to centimeters required to move between or colonize areas of favorable microhabitat, through the meters to kilometers required to support local colonization of ice-free ground within the Antarctic, to the hundreds or thousands of kilometers necessary to allow colonization from lower latitudes. The wide distribution of most Antarctic species in suitable habitats, particularly of the maritime and continental zones, combined with the recent age of such habitats, argues for the success of dispersal of these species, despite the paucity of observational data.

Various Antarctic microinvertebrates (e.g., protozoans, rotifers, tardigrades, and nematodes) have the potential to disperse in a resistant desiccated (anhydrobiotic) state, and then simply rehydrate and resume activity upon arrival in a favorable microhabitat. Likewise, algal and cyanobacterial masses and mosses survive long periods of unfavorable conditions in a desiccated state, and are subject to fragmentation and wind dispersal. Lichens, bryophytes, and fungi possess sexually- and asexually-produced dispersing propagules. All of these are present in aerobiological samples collected in Antarctica, with evidence of both local and distant sources.

D. Atmospheric/Meteorological Circulation Patterns

The oceanic currents and typical pattern of strong westerly airflows encircling the Antarctic continent between approximately latitudes 40° and 60°S normally provide effective barriers to north–south movement of colonizing organisms or propagules by extending likely transit times to such an extent that survival may be impossible. There are several records of South American Nothofagus trunks stranded on beaches on the Antarctic Peninsula and maritime Antarctic archipelagos, but these may have circumnavigated the Southern Ocean before deposition. However, rapid transport of pollen of South American origin into the maritime Antarctic does occur, and is associated with large cyclonic air masses that infrequently create a north–south airflow. Transit time is then 1–2 days, with synoptic weather records suggesting that such events occur approximately once every 18 months. There is also clear evidence, from analysis of ice and lake sediment cores covering a long timescale, of deposition events involving volcanic ash or biological particles that originate from identifiable sources at lower latitudes in the Southern Hemisphere.

V. FUTURE TRENDS

A. Anthropogenic Influences

Both the continent and surrounding Southern Ocean were immune from human impact until recent times. Since the early nineteenth century, however, marine ecosystems have been devastated, at first by the uncontrolled exploitation of marine mammal (seal and whale) populations and, more recently, by the continuing impact of various fisheries either directly on their target species or on non-target bycatch. There is no reason to expect these ecosystems to return to their pre-disturbance states.

Human impact on the terrestrial environment commenced with visits of sealing and whaling vessels to sub-Antarctic and maritime Antarctic islands, followed by the first landings on the Antarctic Peninsula and
continent in the late nineteenth and early twentieth centuries. Since the Second World War, many scientific research stations have been erected around the Antarctic, which brought the development of their logistic support operations and the importation of personnel and materials into the biome. This has been followed since about 1975 by rapidly increasing ship- and air-supported tourist operations. Antarctic terrestrial ecosystems are fragile and sensitive to disturbance, often existing on the same areas of ice-free ground where research stations are established and where large concentrations of wildlife attract tourist operations. Therefore, human presence has inevitably disturbed and destroyed local areas of terrestrial habitat.

Even before the expansion of direct impacts, the consequences of human activity were detectable throughout Antarctica (including the nearshore marine environment) in the form of chemical pollutants, whose increase is recorded throughout the industrial era, and radioactive decay products arising from the use or atmospheric testing of nuclear weapons. Both are dispersed to the continent by the same atmospheric circulation patterns as proposed for biotic dispersal.

1. Introduction of Alien Organisms

The most obvious direct biological impact of human activities occurs via the introduction of alien organisms. Accidental introductions and deliberate transplant experiments have shown that a wide range of flora, fauna, and microbes are capable of surviving and establishing viable populations in all Antarctic zones (Friedmann, 1993; Pugh, 1994). So far, the greatest impact has been on sub-Antarctic ecosystems, which have the longest record of human influence and the least extreme climate. Here, various vertebrates have been introduced both accidentally (rats and mice) and deliberately (fish, chickens, rabbits, cats, pigs, sheep, moufflon, cattle, and reindeer) to all major islands. In some cases, the consequences of these introductions are irreversible, as endemic species have been eradicated, while in others (particularly the presence of rats on larger islands) effective control measures appear to be impracticable. The ecological impact of some introductions appears negligible (e.g., trout to Kerguelen, mice to South Georgia) but, in general, appropriate studies have not been made. Introduced species may have both direct (e.g., predation of bird eggs and terrestrial invertebrates, trampling and grazing of plants) and indirect impacts (e.g., alteration of habitat structure leading to changes in species dominance or behavior) on native species.

Introductions of invertebrates are less well documented. There have been no deliberate introductions of invertebrate species to Antarctic sites. Rather, all such aliens have been introduced accidentally, with stores, food, equipment, or domestic animals associated with human settlement. Pugh (1994) estimated that 13.5% of ca. 520 Acari species reported from Antarctica (mainly the sub-Antarctic) originated from other continents, the majority with human assistance. Although many alien species have persisted only as long as human settlement continued, a number have become established (particularly in the sub-Antarctic, but also in the maritime Antarctic), with some evidence of competitive displacement of native species and the introduction of new trophic interactions to terrestrial ecosystems.

Introductions of flowering plants, bryophytes, and microbes have also accompanied human occupation. Again, the sub-Antarctic zone has been most affected—for instance, South Georgia now has more persistent alien than native flowering plants within its flora (23 versus 17 species, although many of the aliens have very restricted distribution and minimal impact). Transplant experiments have demonstrated that several sub-Antarctic and temperate vascular plants are capable of long-term survival in maritime Antarctic conditions, with one example of an accidental introduction (Poa pratensis) surviving for at least 30 years on the Antarctic Peninsula. Very few data on microbial introductions exist, although the process is inevitable via human transport (see Broady, in Wynn-Williams, 1996) and complete control measures are impracticable. Given the apparent evolutionary isolation of Antarctic prokaryotes (see earlier discussion), the potential for damage to this unique biological resource should not be underestimated.

2. Regional and Global Environmental Change

The identification and prediction of global climate change via global circulation models (GCMs) currently receive a great deal of attention from scientists, politicians, and the general public. Although details differ, most GCMs predict that any climate (temperature) amelioration will be both greatest and most rapid at high latitudes.

a. Temperature

There is clear evidence of rapid regional warming trends from long-term temperature records maintained at several maritime and sub-Antarctic research stations, with increases on the order of 1–1.5°C between 1950 and 2000. It is not yet possible to state conclusively whether this represents a regional or global process. There is limited evidence available from a small number of
coastal sites in the continental Antarctic of a parallel process occurring, although it is not yet clear how any global processes will affect the main continental ice mass.

b. Water

Another important prediction of GCMs is changing patterns of precipitation, which alter the water input to terrestrial habitats. Detailed predictions are not available for Antarctica, although it is recognized that water is possibly the single most important factor limiting distribution of Antarctic terrestrial biota. Water availability from precipitation is modulated by seasonal temperature variation. At a local scale, water availability to terrestrial ecosystems can undergo drastic change as a result of factors as simple as the complete exhaustion of a snow bank or, conversely, the increased release of water from melting ice fields. These effects may be caused by changes in either the timing or amount of snow accumulation, or the duration of positive summer temperatures (i.e., in combination with temperature amelioration).

At a larger scale, there is abundant evidence of rapid glacial retreat in the sub-Antarctic and maritime Antarctic zones. If this process includes the increased rate or duration of melting during summer, it will clearly influence the rate of water input to terrestrial ecosystems. Conversely, a combination of more rapid melting of winter snowfall (a finite resource) and extended summer length may lead to decreased water availability as the summer progresses.

c. Ultra-violet Radiation

A recent anthropogenically generated influence on Antarctic and possibly global ecosystems—the ozone hole—has existed since the early 1980s. This depletion of the Earth’s protective stratospheric ozone layer is caused annually in the austral spring over Antarctica by a light-catalyzed reaction between ozone and pollutants concentrated at high altitudes and latitudes by atmospheric circulation patterns. Its formation leads to a drastic reduction in the ozone concentration over Antarctica and high southern latitudes, allowing a disproportionate increase in the exposure of terrestrial (and shallow marine) ecosystems to potentially damaging shorter-wavelength UV-B radiation. Radiation levels experienced at any site vary depending on the pattern of movement of the ozone hole and modulation by factors such as clouds and snow cover, but at worst exposure levels comparable to those of the tropics are experienced. This represents a very rapid change in a biologically important variable over a period of less than 20 years.

B. Biological Consequences of Climate Change

Despite a predictive literature (e.g., Voytek, 1990; Wynn-Williams, 1994) and increasing numbers of published studies of species in the Arctic region (e.g., Oechel et al., 1997; Henry, 1997), identification of the effects of climate change processes in the Antarctic terrestrial environment is only now starting to receive critical attention (see Kennedy, 1995; Convey, 1997), and it is currently impossible to predict the detailed trajectory of change for any specific ecosystem. Knowledge of the physiological and life-history characteristics of individual species indicates that extant Antarctic terrestrial biota exhibit sufficient physiological and ecological scope or flexibility to absorb and even benefit from both the direct and indirect (e.g., changes in nutrient availability) effects of predicted levels of change. However, how individual species responses may be integrated at the community or ecosystem level is unknown in Antarctic systems (but see Freckman and Virginia, 1997). Any climate amelioration is likely to influence Antarctic ecosystems and biodiversity further by easing the constraints that limit colonization and establishment of exotic species.

The effects of temperature amelioration are expected to be greater at high-latitude sites, which already have a very restricted thermal energy budget and where the relative importance of a small temperature increment will be of greater significance. Thus an expected consequence of amelioration (assuming continued or increased water availability) is to increase the effective length of season and/or thermal energy budget of terrestrial ecosystems, which should in turn lead to increased growth, reproductive rates, and population sizes of both photosynthetic autotrophs and heterotrophic microbial and invertebrate species. The limited evidence available from Antarctic studies supports this contention, with increasing growth rates and more rapid completion of development reported in experimental studies of higher plants and invertebrates, combined with field observations of rapid areal increases in plant communities (Convey, 1997). Parallel studies from the Arctic are more equivocal, where groups such as aphids showed large population increases (11-fold increase in numbers of overwintering eggs) in response to realistic temperature manipulations, while other detritivorous soil arthropod populations did not show a detectable response (Hodkinson et al., 1998).
Some climate change processes are likely to have deleterious effects on Antarctic biota. In particular, any decrease or total loss of water input to ecosystems could lead to local extinctions and drastic changes in ecosystem structure (via both physical and biological effects). Likewise, should the maximum exposure to increased UV-B radiation surpass the tolerances of exposed biota, particularly colonizing microbiota with crucial roles in soil stabilization that permit secondary colonization by other organisms, some terrestrial habitats may become barren.

Because of the extreme isolation of terrestrial habitats in all three biogeographical zones from potential source populations in the lower latitudes, initial responses to climate change processes will be restricted to the extant biota. However, climate amelioration will lead to both an increased area of available terrestrial habitat and a longer "window of opportunity" within which to complete the dual processes of colonization and establishment. Eventually, this would likely lead to increases in biodiversity and/or shifts in community composition, with or without human assistance, and in turn to greater trophic complexity and the inclusion of higher trophic levels. The resistance of extant Antarctic biota to these processes is unknown. Yet these simply structured ecosystems do present ideal conditions with which to test important ecological theories relating to community stability and species redundancy.

See Also the Following Articles

ARCTIC ECOSYSTEMS • AUSTRALIA, ECOSYSTEMS OF • DISPERSAL, BIOGEOGRAPHY • SOUTH AMERICA, ECOSYSTEMS OF

Acknowledgments

I thank the British Antarctic Survey (BAS) for resources, many BAS staff for discussions that have contributed to the development of this paper, and two anonymous referees. Dr H. J. Peat extracted biogeographical information contained in Table II from databases held by the BAS Biological Sciences Data and Resources Centre.

Bibliography

I. Aquaculture and Biodiversity

II. Conclusion

GLOSSARY

aquaculture: The farming of aquatic organisms, including fish, mollusks, crustaceans, and aquatic plants. Farming implies some sort of intervention in the rearing process to enhance production, such as regular stocking, feeding, or protection from predators. Farming also implies individual or corporate ownership of the stock being cultivated. For statistical purposes, aquatic organisms which are harvested by an individual or corporate body which has owned them throughout their rearing period contribute to aquaculture, whereas aquatic organisms which are exploitable by the public as a common property resource, with or without appropriate licenses, are the harvest of fisheries.

broodstock: Fish or shellfish from which a first or subsequent generation may be produced in captivity, whether for growing as aquaculture or for release to the wild for stock enhancement.

farming intensity: In a broad continuum, extensive systems are those which are closest to natural fisheries, requiring minimal inputs and offering relatively low yields, whereas intensive systems require a large amount of inputs to maintain an artificial culture environment, with high yields. Between these extremes are the varying degrees of semi-intensive aquaculture, where definitions are less distinct: (i) extensive aquaculture does not involve feeding of the organism, (ii) semi-intensive aquaculture involves supplementation of natural food by fertilization and/or the use of feeds, and (iii) intensive aquaculture is when the culture species is maintained entirely by feeding with nutritionally complete diets.

feed conversion: The efficiency of farmed animals to incorporate given feed into biomass. Feed conversion is usually expressed in terms of the feed conversion ratio of weight of diet used to fish/shellfish flesh biomass produced. The ratio is affected by the relative moisture content of both feed and aquaculture product as well as the metabolic characteristics of the farmed species, farming techniques, and husbandry.

seed: A term used to describe eggs, larvae, postlarvae, or juveniles (fry and fingerlings) stocked into aquaculture production systems.

spawner: Mature individual of a stock responsible for reproduction.

AQUACULTURE, the aquatic counterpart of agriculture, has grown rapidly in recent decades to become one of the most important means of obtaining food from the sea. Impacts of aquaculture on biodiversity arise from the consumption of resources, such as land (or space), water, seed, and feed, their transformation...
into products valued by society, and the subsequent release into the environment of wastes from uneaten food, fecal and urinary products, and chemotherapeutics as well as microorganisms, parasites, and feral animals. Negative effects may be direct, through release of eutrophating substances, toxic chemicals, the transfer of diseases and parasites to wild stock, and the introduction of exotic and genetic material into the environment, or indirect through loss of habitat and niche space and changes in food webs. Today, large quantities of fish are caught to produce fish meal—the main ingredient in feed—which may result in overfishing and affect marine food chains, including marine mammals and top carnivores. In some types of aquaculture, fish and shrimp larvae are caught in the wild to be used as seed. This may also result in bycatches of large amounts of other larvae, representing losses to capture fisheries and biodiversity. Large areas of critical habitats such as wetlands and mangroves have been lost due to aquaculture siting and pollution, resulting in lowered biodiversity and recruitment to capture fisheries. The magnitude of biodiversity loss generally increases with scale, intensity of resource use, and net production of wastes, but it is very much dependent on which species is cultured and the method of cultivation. In some cases aquaculture may increase local biodiversity, e.g., when ponds are constructed in dry areas and with integrated aquaculture.

I. AQUACULTURE AND BIODIVERSITY

A. The Aquaculture Process

Aquaculture is an economic human activity that uses and transforms natural aquatic resources into commodities valued by society, e.g., fish, shrimp, mussels, and seaweed. In doing so it may impact on biodiversity, essentially due to the consumption of resources, the transformation process (aquaculture), and the production of wastes (Fig. 1).

Contrary to common belief, technical and economic inputs, such as construction materials, energy, and labor, form only a small part of the inputs needed for aquaculture. The main and critical inputs are instead natural resources. Together with nature’s services they ultimately determine the limits for the local and global expansion of aquaculture. The magnitude and type of resource use and impacts of aquaculture are very much dependent on which species is cultured, the farming methods used, and the intensity of farming (Box 1 and Table 1).

The aquaculture process in itself may affect biodiversity as a result of disturbance through increased road and boat traffic. High densities of farmed fish and food often attract predators and scavengers such as wild fish, gulls, and seals. These can come into conflict with farmers and may be killed, either accidentally (entanglement

FIGURE 1 Diagram summarizing direct and indirect effects on biodiversity of the aquaculture process through the use of resources and the generation of wastes. Details are given in the text.
in nets) or deliberately (shooting and trapping), or if they become established they may displace sensitive local species.

B. Feed

Whereas extensive aquaculture of herbivorous species such as carp and filter-feeding bivalves uses natural production and semi-intensive farming is supplied agricultural wastes and some feed, intensive aquaculture of fish and shrimp uses formulated feeds based on fish meal and fish oils (Boxes 1 and 2 and Table 1). Many intensive and semi-intensive aquaculture systems use two to five times more protein, in the form of fish meal, to feed the farmed species than is ultimately harvested.

Feed requirements place a strain on wild fish stocks, and currently about one-third of the total harvest of capture fisheries is used to produce fish meal, one-third of which is used by the aquaculture industry. This may result in overfishing of small pelagic species, affecting marine food chains and ultimately marine mammals and top carnivores.

C. Land

Land is needed for building fish or shrimp ponds, whereas fish cages, pens, and mussel and seaweed farms occupy space in lakes and the sea. Globally, fish ponds are usually sited in agricultural land and this arguably contributes to landscape and floral and faunal diversity. In Europe, however, unproductive, boggy areas of agricultural land have often been used, and since such boundary ecosystems, or ecotones, may serve as reserves for species in areas otherwise surrounded by monocultures of crops, this may reduce biodiversity. Large areas of tropical coastal wetlands and mangroves have been converted to fish and shrimp ponds, resulting in lowered biodiversity and recruitment to fisheries. When the full range of ecological effects associated with mangrove habitat loss are accounted for, the net production in fish and shrimp aquaculture may be negative (Box 3).

It is increasingly being realized that the direct land and space resources occupied by agricultural or industrial production facilities are a side issue; what really
### Table 1

<table>
<thead>
<tr>
<th>Farming intensity</th>
<th>Natural resource use</th>
<th>Relative importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>extensive</td>
<td>construction materials (ropes, wood); insignificant</td>
<td>highly significant; global; 7.2 mmt</td>
</tr>
<tr>
<td>molluscs (mussels, oysters)</td>
<td>construction materials (ropes, wood, bags/nets); insignificant</td>
<td>highly significant; global; 8.6 mmt</td>
</tr>
<tr>
<td>shrimp, crabs and lobster</td>
<td>land, (sea)water, seed (wild)</td>
<td>insignificant; tropics/sub-tropics; intensification widespread (&lt;0.1 mmt)</td>
</tr>
<tr>
<td>fish (carps, tilapias, salmonids, etc.)</td>
<td>land, (rain)water, seed (usually wild capture)</td>
<td>insignificant; of local importance in tropics only; limited to poorer sectors in a few societies (&lt;0.1 mmt)</td>
</tr>
<tr>
<td>semi-intensive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrimp, crabs and lobster</td>
<td>land (significant), seawater (insignificant), seed (wild; significant), feed (increasingly significant)</td>
<td>significant; tropical; ~0.8 mmt</td>
</tr>
<tr>
<td>fish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>intensive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrimp, crabs and lobster</td>
<td>land (significant), seawater (insignificant), seed (wild; significant), feed (significant)</td>
<td>insignificant and decreasing; tropics and sub-tropics; ~0.3 mmt</td>
</tr>
<tr>
<td>fish</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Matters is the area required to supply the resources that sustain the activity (see Box 6).

### D. Water

Aquaculture requires large amounts of clean water to support the farmed animals, replenish oxygen, and remove wastes. In land-based systems, aquaculture does not only borrow water, returning it a more degraded form, but also consumes it or accelerates its loss from surface to groundwater or the atmosphere. Thus, by creating ponds, especially in areas of poor (sandy/loam) soils or high temperatures, evaporation and seepage are increased and as much as 1–3% of the fish pond volume may be lost in this way each day. Such losses may be particularly significant in arid or semi-arid areas of the world, such as Israel, where fish pond design and management practices have had to be changed in order to reduce surface water losses. Conversely, the incorporation of a fish pond into small rural farms has been shown to improve water conservation.

### E. Seed

Seed (eggs, larvae, and fry) is required to stock fish and shellfish farms. Although most inland and temperate coastal aquaculture relies on hatchery-produced fry and fingerlings, tropical marine fish and shrimp culture is still largely dependent on wild broodstock or juveniles, often creating demands that adversely affect wild stocks. Moreover, it is not only the target species which may be at risk since large bycatches of other larvae are killed, representing losses to capture fisheries and biodiversity (Box 4).

### F. Impacts of Wastes

The term “wastes” in the current context is used to mean not only food, fecal and urinary products, and chemicals but also microorganisms, parasites, and feral animals which may be introduced to new environments along with aquaculture and may escape from aquaculture facilities. These have implications on biodiversity (Fig. 1 and Box 5). Uneaten food and fecal and urinary wastes may lead to eutrophication and oxygen depletion, the magnitude of which is dependent on the type and size of operation and the nature of the site, especially size, topography, and water exchange. Chemotherapeutants, including antimicrobial compounds and pesticides, are used in intensive marine fish culture to control bacterial and parasitic diseases. Controls on use
are increasingly strict, especially in northern Europe and North America. In shrimp farming, many chemicals are used, largely without regulation. Farmers use a range of vitamins, immunostimulants, disinfectants, and chemotherapeutics and employ chemicals for pond soil and water treatment. The impacts of these chemicals are largely unknown, although there is evidence that microbial biodiversity can be affected by antibiotics, and antiparasitic chemicals may have implications for biodiversity.

G. Assessment of Resource Use and Carrying Capacity

To reduce the risk of resource constraints and impacts on biodiversity, a shift to aquaculture production systems that use less resources and emit wastes that do not exceed the assimilative capacity of the environment must occur. We also need to recognize and manage nature's life support on which economic development and human welfare depends. One way of identifying the demands for natural resource and ecosystem services of aquaculture is to estimate the ecosystem area—the ecological footprint—functionally required to support the activity (Box 6).

The footprint concept has proven to be very useful in illuminating the nonpriced and often unperceived work of nature that forms the basis for economic activities such as industrial aquaculture. When problems appear in fish ponds or fish cages, people tend to look at what is going on inside the pond or cage, not realizing that the farm is part of a much larger surrounding ecosystem, with which it interacts. Surrounding ecosystems provide the feed, seed, clean water, and other necessary natural resources and ecosystem services including waste assimilation. This work of nature sets the limits on how much can be cultured without compromising biodiversity or creating pollution or disease problems. Exceeding the carrying capacity of the environment is a major reason why many aquaculture developments, such as intensive shrimp farming, have had problems or even failed.

H. Improving Aquaculture Practices

The aquaculture industry has made considerable achievements in areas in which environmental impacts have negatively affected production quantity and quality. Some feeds have been formulated with lipids instead of proteins as an energy source in order to reduce nitrogen excretion by fish (Austín, 1997). Plant proteins with lower phosphorous levels are also used in feeds to reduce water pollution since most fish meal contains more phosphorous than fish can assimilate (Runsey, 1993). In addition, the use of extrusion processes to develop high-quality fish meals has improved protein digestibility and produced more buoyant feeds that reduce waste (Bottling, 1991), whereas the so-called “smart” feeding systems that monitor appetite and adjust feeding regime accordingly reduce food wastage considerably.

Many other systems have also been created in response to environmental damages and consumer concerns. In Orkney, Scotland, salmon produced with certified organic labels are now available for sale on the international market (Seafarm Business Review, 1998; Aarset et al., 1998).

In some cases, the development of polyculture and integrated farming systems that make more efficient use of inputs and generate less waste, thus adding to net fish supplies locally and regionally at relatively minor environmental and social expense, may be a viable option (Troell et al., 1997) (Box 7).

These examples indicate the positive directions in which the industry is heading. However, in cases in which awareness of environmental consequences is low, or in which the impacts do not affect profitability of the farmers and retailers, there is little incentive for the industry to regulate itself. Whereas economics and markets are promoting the adoption of technical innovation among different sectors of the industry, the policymakers must also take responsibility by incorporating proper incentive structures, developments of integrated coastal zone management, environmental impact assessment, and other organizational and institutional frameworks, including trade rules and regulations.

II. CONCLUSION

Since the principle of aquaculture is to reroute flows of energy and matter from the ecosystem into those species that we culture, aquaculture, like agriculture, will always affect the environment to some extent. This interaction and alteration of supporting environment is unavoidable, but it should not be done in a fashion that deteriorates the capacity of ecosystems to sustain social and economic development. Recent research has revealed that aquaculture systems developed in the past few decades are not an independent food-producing process. They are strongly coupled to nature’s subsidies and services to sustain production.

Species such as shrimp, salmon, and other intensively reared carnivorous species in coastal areas present problems for wild fish stocks since feed requirements alone lead to a net loss in fish protein for these species
farmed fish and shellfish production was approximately 2 million tons and largely confined to areas of Asia. During the past 50 years, global aquaculture production has grown nearly 20-fold to 36 million tons in 1997. In the 10-year period ending in 1997, production of farmed fish increased, resulting in a more than doubling in weight and value, as has its contribution to world fish supplies (FAO, 1999). The aquaculture sector today produces 16.7 million tons of finfish (49% of the total), 1.1 million tons of crustaceans (3%), 8.5 million tons of mollusks (25%), and 7.7 million tons of aquatic plants (23%). Fish produced from farming activities, or aquaculture, now accounts for more than one-fourth of all fish directly consumed by humans. More than 300 species of fish and shellfish are farmed; the range includes giant clams that obtain most of their nutrients from symbiotic algae, various species of carp that are largely herbivorous, and Atlantic salmon and marine fish species that are carnivorous (Williams, 1996). Aquaculture typically involves the enclosure of a species in a secure system under conditions in which it can thrive. Interventions in the life cycle range from exclusion of predators and control of competitors (extensive aquaculture) to enhancement of food supply (semi-intensive) or even the provision of all nutritional requirements (intensive). Intensification also implies increasing the number of individuals per unit area, which requires greater use and management of inputs and a greater reliance on technology and fossil energy.

Production practices and their impacts on aquatic ecosystems vary widely across species (Table 1). Mollusks such as scallops and mussels are generally farmed along subtidal or intertidal coastlines where wild-caught or hatchery-reared seed are grown in bags set on the sea bottom or on stakes and suspended ropes. The animals rely entirely on prevailing supplies of plankton and organic particles for food. A range of systems—ponds, tanks, or cages—are used to farm finfish. The majority of carp and other freshwater species farmed in the tropics and sub-tropics are herbivores/omnivores and are grown in fertilized by supplemental feeds (see Box 2). In contrast, most diadromous and marine finfish, including both tropical and temperate species, are farmed intensively in floating net cages and are reliant on nutritionally complete fish meal and fish oil-based diets. Crustacean farming is dominated by

---

**Box 1**

**Aquaculture Development and Practices**

Aquaculture, the farming of aquatic plants and animals, is many thousands of years old. Nevertheless, it must be regarded as a largely post-World War II phenomenon. In 1950, global
penacid shrimp. Shrimp postlarvae, either derived from captured wild parents or directly collected from the sea, are stocked in coastal ponds and grown for 3 or 4 months until they reach marketable size. The more intensive shrimp systems depend mainly on formulated pellet feeds, aeration to replenish dissolved oxygen, and pumped seawater to dilute pollutants and flush out harmful metabolites.

Box 2

Fisheries for Fish Meal as Feed

Extensive aquaculture, such as that of seaweed and mollusks, is solely reliant on ambient nutrient/food supplies. The traditional semi-intensive managed fish and shrimp production systems that predominate in the tropics and subtropics also use little or no fish meal. Herbivorous and omnivorous carps and tilapias are grown in ponds to which nutrient-rich materials are often added to stimulate growth of algae and other food items. On-farm feeds, based largely on cheap, locally available agricultural by-products such as rice bran, augmented by household scraps and perhaps small amounts of fish meal are used to supplement the food in the ponds. However, in the intensive production systems that predominate in temperate aquaculture the farmed animals are solely reliant on fish meal and fish oil-based diets.

Only 2% of the 10 most commonly farmed species, catfish and carp, require less fish as inputs than is ultimately harvested (Naylor et al., 2000). For the others, more fish is required as feed inputs than is ultimately produced: 2 kg of wild fish is used to produce 1 kg of farmed salmon, trout, and marine shrimp, and 4.5 kg of wild fish is used to produce 1 kg of marine finfish and eel (Tacon, 1996). The culturing of these species will thus lead to a net loss in fish protein.

Human consumption of seafood is 103 million metric tons (mmt), of which 65 mmt of fish, crustaceans, and mollusks and 1 mmt of seaweed come from capture fisheries, whereas 29 mmt of fish and 8 mmt of seaweed are from aquaculture (Fig. 2) (Tacon, 1998). Global fish capture is much higher—123 mmt—but 27 mmt of this is discarded as bycatch (Alverson et al., 1994), and 30 mmt of it is used for fish meal production. An additional 2 mmt of processing scraps from aquaculture and fisheries is also converted into fish meal (Pike, 1998). Two-thirds of total fish meal supplies (22 mmt) are currently used in feed for chickens, pigs, and other animals, and one-third (approximately 10 mmt) is converted into aquaculture feeds (Tacon, 1998; Pike, 1998). However, an increasing proportion will probably be used for aquaculture feeds as supplies are unlikely to expand and as aquaculture production continues to grow and production methods of pond fishes in major producer countries such as China intensify.

Diet for salmonids, seabass, and sea bream and other carnivores are largely composed of fish meal and fish oil. Although it may be possible to replace much of the fishmeal used in intensive fish diets with oilseed proteins (Stickney et al., 1996), requirements for essential amino acids, especially cystine and methionine, will continue to have to be met from other sources such as fish meal. It remains to be seen whether commercial oilseed protein-based diets can be developed in an industry in which the product is competing with many others for customer attention and in which profit margins are increasingly being squeezed. Depending on source and inclusion rate, oilseed meals can compromise palatability, growth (Stickney et al., 1996) and, possibly, profitability. Any decrease in palatability or diet digestibility may aggravate waste loadings to the environment (Beveridge et al., 1999). The issue of fish oils is perhaps even more pressing than that of fish meal. Aquatic carnivores are poor at using carbohydrate to supply energy requirements, relying instead on protein and lipid (Cowey and Sargent, 1977). The substitution of fish oils with vegetable oils in freshwater carnivorous or omnivorous fish diets is possible (Buzzi et al., 1997). However, there are limitations with regard to marine carnivores such as salmon because they require n-3 highly unsaturated fatty acids which currently can only be derived in commercial quantities from fish oils (Bell et al., 1986).

The use of fish meal and fish oil in livestock and aquaculture feeds reduces aquatic productivity through a series of food web interactions. Four of the top 5 and 8 of the top 20 capture species are used for reduction to fish meal (FAO, 1998). All are small, pelagic fish, including anchoveta, Chilean jack mackerel, Atlantic herring, chub mackerel, Japanese anchovy, round sardinella, At-
lantic mackerel, and European anchovy. Many fishes exploited for feed are overfished and are strained by climatic variability associated with El Nino—Southern Oscillation events (FAO, 1997; NRC, 1999). Although the impact of pelagic fisheries depletion has been little studied (Botsford et al., 1997; NRC, 1999), it is thought to reduce available food supplies for marine predators including valuable species consumed by humans (Folke and Kaustsky, 1989; Fischer et al., 1997). In Europe, recent crashes of North Sea capelin and herring stocks have been attributed to overfishing and may have caused the loss of other wild fish stocks (e.g. cod) and the starvation of seals and seabird chicks (Folke and Kaustsky, 1989; Vader et al., 1990). Declining capelin populations in the western Gulf of Alaska are implicated in the decrease of harbor seal and sea lions in the early 1980s (Hansen, 1997). A strong interaction between anchoveta and seabird and mammal populations has also been well documented for the Peruvian upwelling system (Pauly, 1987).

Mollusk culture does not use compound feeds; however, very dense farming of mussels and bivalves in semi-enclosed coastal areas may in exceptional cases reduce fisheries stocks by shortening the linkages comprising the food web. In Rio Arosa, Spain, for example, overgrazing of the phytoplankton population by filter-feeding mussels resulted in zooplankton starvation and the subsequent collapse of the sardine fishery (GESAMP, 1991). This type of ecological disruption, combined with other environmental impacts from culture activities such as conversion of land, habitat modification, and pollution, may lead to net reductions in aquatic productivity over time.

Box 3

Habitat Modification and Biodiversity

Coastal ecosystems, such as mangroves, seagrass beds, and coral reefs, provide habitats and nursery areas for many fish and invertebrate species important to marine fisheries. Almost one-third of the world’s marine fish species are found on coral reefs, and fish catch from reefs contributes to human fish consumption approximately 10% globally and much higher in developing countries (Weber, 1993). Harvests of fish, crustaceans, and mollusks from mangrove-lined lagoons and estuaries—excluding mangrove-associated species captured elsewhere—accounted for more than 1% of global fish catch in 1986 (Matthes and Kapetsky, 1988). In the Southeast Asian region, mangrove-dependent species, excluding trash fish, account for approximately one-third of wild fish landings annually (Sasekumar et al., 1994). Moreover, mangroves are closely linked to habitat conditions of coral reefs and seagrass beds through the biophysical interactions in the coastal seascape (Ogden and Gladfelder, 1983; Rönnbäck, 1999).

A positive relationship between fish/shrimp landings and mangrove area has been documented throughout the tropics (Pauly and Ingles, 1986; Rönnbäck, 1999). Mangroves serve as nurseries providing food and shelter for many commercial fish and invertebrate species caught in coastal and offshore fisheries (Robertson and Duke, 1987; Primavera, 1998). In Southeast Asian mangroves, these juveniles belong predominantly to the penaeid shrimps and the fish families Ariidae, Mugilidae, Centropomidae, Serranidae, Carangidae, Leiognathidae, Lutjanidae, and Sciadidae (Matthes and Kapetsky, 1988; Sasekumar et al., 1994). Other commercial species are associated with mangroves as opportunistic or sporadic foragers. Moreover, mangroves also harbor a wide array of nonmarketed fish, crustacean, and mollusk species, whose subsistence harvest constitutes an important protein source for coastal communities.

Since the 1400s, hundreds of thousands of hectares of mangroves have been transformed into milkfish ponds in Indonesia and the Philippines. In recent decades, shrimp farming has been responsible for a significant share of the conversion of coastal and supratidal areas, e.g., 102,000 ha of mangrove forests in Vietnam in 1983–1987 (Tuan, 1997) and 65,000 ha in Thailand in 1961–1993 (Menasveta, 1997). The loss in wild fisheries stocks due to habitat conversion associated with shrimp farming is potentially very large. If 900 kg of fish, 160 kg of penaeid shrimp, 74 kg of sergestid shrimps, and 25 kg of mangrove mud crabs are harvested annually nearshore for each hectare of mangrove (Rönnbäck, 1999), the 65,000 ha of mangroves converted to shrimp ponds in Thailand result in an annual loss to nearshore Thai fisheries of approximately 58,500 mt of fish, 10,400 mt of penaeid shrimp, 4800 mt of sergestid shrimps, and 1600 mt of
mud crabs. With an average productivity of 3000 kg/ha in Thai shrimp ponds (Rosenberry, 1998), 300 g of fish, 53 g of penaeid shrimp, 25 g of sergestid shrimps, and 8 g of mud crabs are lost from capture fisheries for every kilogram of shrimp cultured in these ponds. These estimates only partially include fish, prawns, crabs, and mollusks that are harvested from mangroves and that contribute to onsite fisheries production. Neither are the off-site fisheries supported by the biophysical interactions between mangroves and other coastal ecosystems such as seagrass beds and coral reefs accounted for in this analysis. Moreover, because penaeid shrimp sales generate most of the revenues from mechanized trawling in developing countries, shrimps (and indirectly mangroves) effectively subsidize commercial harvesting of fish by these vessels (Turner, 1977; Bennett and Reynolds, 1993), including species that do not use mangroves as habitat. The loss to capture fisheries may be higher than fish and shrimp production in mangrove ponds if the full range of ecological effects associated with habitat conversion are accounted for.

**Box 4**

**Wild Capture of Larvae and Spawners for Aquaculture**

The farming of shrimp and fish depends on larvae collected from the wild or reared in hatcheries from eggs of wild broodstock or spawners, thereby putting additional pressure on marine fisheries (Fig. 2). The quantities of bycatch associated with such wild catches are directly proportional to the natural abundance of the target species for culture. For example, milkfish Chanos chanos constitute only 15% of total fry (excluding great numbers of ambassids, clupeids, and engraulids) in daily inshore collections by seine net (Bagarinao and Taki, 1986). In the Philippines and Indonesia milkfish culture is a major industry which relies heavily on natural seed. The annual utilization of approximately 1.7 billion wild fry for stocking in Philippine milkfish ponds (Bagarinao, 1998) means a corresponding loss of nearly 10 billion fry of other fish species.

In India and Bangladesh, where ponds are predominantly stocked with wild *Penaeus monodon* seed, up to 160 fish and other shrimp fry are discarded for every penaeid shrimp collected from estuarine waters (Primavera, 1998). The same low abundance of larval *P. monodon* applies to adult stages—for trawl catches, 0.14% in the Arafura Sea, Indonesia (Evans and Wahju, 1996), 0.56% in peninsular Malaysia (Ahmad, 1995), and 0.48–0.85% in Guimaras, the Philippines (Primavera, 1995); 0.43–0.68% of commercial fisheries in the Andaman Sea and the Gulf of Thailand (Vibhakar, 1980); 2.2% of prawn landings in the east coast of India (Rao et al., 1993); and 8.78% of fish corral catches in Batan Bay, the Philippines (Motoh, 1981).

Although the development of hatcheries for cultured shrimp and fish species may have reduced dependence on wild seed (and their mangrove nurseries), it has also increased demand for wild-caught mature (spawners) and immature (broodstock) adults. Because the species is so rare, wild collection of *P. monodon* broodstock and spawners may lead to large amounts of bycatch. Overexploitation of the adults and larvae of both target and incidental shrimp species could be a cause of declining stocks. The staggering quantities of bycatch destroyed by a yearly seed collection of 1 billion *P. monodon* in southeast Bangladesh could have major consequences for biodiversity and capture fisheries production (Deb et al., 1994). In West Bengal, India, where shrimp seed collection constitutes a significant fishery, the contribution of adult shrimp to fisheries landings decreased from 14.4% in 1970–1971 to 9.1% in 1989–1990 (Banerjee and Singh, 1993).

**Box 5**

**Impacts of Wastes, Chemicals, Diseases, and Feral Animals**

Nutrient release into marine ecosystems from coastal and nearshore aquaculture operations, especially shrimp and salmon, has been well documented (Gowen and Bradbury, 1987; Briggs and Funge-Smith, 1994; Beveridge, 1996). Similar problems exist in intensive cage culture of carp and tilapia in freshwater ponds, but semi-intensive, extensive, traditional, and polyculture pond systems generally assimilate wastes internally (Beveridge, 1984; Kautsky et al., 1997) and often
enhance water quality in the case of integrated carp/agriculture/livestock systems (Kestemont, 1995).

Certain chemicals added to culture systems, such as antibiotics, chlorine, and anti-fouling paints, also harm farmed and wild fish populations. For example, wild fish and shellfish that seek food and refuge near scallop farms in Chile and salmon farms in Scotland contain large amounts of copper and zinc, respectively. Cultured milkfish in the central Philippines have been observed with tin levels of 0.4 mg/kg dry weight (DW) in muscle and 2.3–2.8 mg/kg DW in liver (Coloso and Borlongan, 1999). The acceptable daily intake of total tin for humans is only 0.0005 mg/kg body weight (FAO/WHO, 1991). The high fish levels of tin and triphenyltin (0.3–0.4 mg/kg DW in muscle and 0.6–0.7 mg/kg DW in liver) are striking because application of organotin molluscicides, popular among milkfish pond culturists, had been discontinued (in the wake of a government ban) for 6–12 months in these farms at the time samples were taken.

Further impacts of aquaculture operations on ocean and coastal fisheries can be linked to invasions of feral species (Gausen and Moen, 1991; McKinnell et al., 1997; Gross, 1998; Soto, 1998), spread of diseases and parasites (Rosenthal, 1980; Welcomme, 1998; Chew, 1990), release of nutrients and organic matter (Folke et al., 1994; Costa-Pierce, 1996), and discharge of chemicals and antibiotics (Beveridge, 1984; Alderman et al., 1994; Weston, 1996; GESAMP, 1997). These factors—combined with increasing reliance on external feeds, habitat conversion, and bycatch—negatively affect fisheries through qualitative and quantitative changes in spawning and nursery grounds, species composition, and food webs of adjacent ecosystems (Fig. 2).

In some cases, fish culture reduces wild fish stocks through the introduction and invasion of exotic species. Tilapias have had a long history of purposeful (for aquaculture and fisheries) and accidental introductions to 90 countries and territories. Limited information available on the environmental impacts of these introductions indicates that the presence of tilapias is only one of many factors contributing to the decline of native fish species in some waters in the Philippines and Bangladesh (Pullin et al., 1997). Atlantic salmon from culture facilities have escaped within the geographic range of wild Atlantic salmon, as well as in Pacific waters, and are now found as far north as the Bering Sea and as far south as Chile. Increasing evidence suggests that these escapes may have direct genetic impacts on wild populations through hybridization (Gross, 1998). Larger numbers of escapes also increase the likelihood of hybridization between farmed Atlantic salmon males and wild females in populations that are locally endangered or close to extinction (Slaney et al., 1996; Gross, 1998). In addition to consequences for the population gene pool and fitness, there are many potential ecological impacts associated with feral fish. Atlantic salmon may compete extensively with wild salmon species for food and space, disturb native spawning sites, and introduce new diseases and parasites into wild populations (Beveridge et al., 1994). Numerous wild salmon stocks in Norwegian rivers have been infected by an ectoparasitic trematode through the release of juvenile farmed salmon (Folke and Kautsky, 1989).

Diseases are prevalent in other aquaculture species as well, especially marine shrimp. Worldwide transfers and introductions of the few preferred culture species, including P. monodon, Litopenaeus vannamei, and Marsupenaeus japonicus, were numerous in the early decades of commercialized shrimp culture. At the peak of Taiwanese shrimp production in 1982–1986, yearly imports from Southeast Asia of more than 100,000 live P. monodon broodstock supported hatchery production (Chin, 1988). Such introductions and transfers may lead to competition with endemic fauna, genetic introgression with local fauna, and introduction of pathogens and parasites (Beardmore et al., 1997).

Recently, the introduction of shrimp postlarvae and broodstock from areas affected by the whitespot syndrome virus (WSSV) and Taura syndrome virus (TSV) was often followed by the rapid spread of these major pathogens throughout most of the shrimp-growing regions in Asia and Latin America, respectively. A native of Asia, where it has caused multimillion-dollar shrimp crop losses, the WSSV has been detected in wild and cultured shrimp (L. setiferus, L. vannamei, L. stylirostris, and Farfantepenaeus duorarum) in Texas and South Carolina (Lightner et al., 2000). The virus was probably introduced by release of untreated wastes from plants processing imported Asian shrimp into coastal waters and by use of imported shrimp as bait in sports fishing or as
fresh food for rearing other aquatic species (Lightner et al., 1997). Another major shrimp virus, the infectious hypodermal and hematopoietic necrosis virus (IHHNV), is believed to have been introduced to the Americas from Asia through the importation of live *P. monodon* in the early 1970s (Lightner et al., 1999). In the Philippines, IHHNV prevalence in various wild populations of the giant tiger prawn has been correlated with shrimp culture intensification and mangrove status (Belak et al., 1999). Lower viral incidence in wild shrimp has been found in sites with primary mangroves and no major aquaculture industry, whereas higher levels have been observed in areas with intensive shrimp farms and severely degraded mangroves. Wild populations had a significantly lower overall IHHNV incidence of 51% compared to total infection in captive *P. monodon* reared from second- and third-generation hatchery fry.

In contrast to shrimp and salmon, comparatively few diseases have been reported for carps, tilapia, and milkfish, particularly from extensive and other low-density culture systems. The current trend toward intensification in rearing ponds and cages, however, may create stressful conditions through deterioration of water quality, excessive stocking, and polluted water inflow that predispose the fish to disease. For example, pond discharges of nutrients from feces and uneaten food may alter nutrient and organic matter concentrations, stimulate blooms of toxic algae, and generate self-pollution that is conducive to disease outbreaks.

**Box 6**

**The Ecological Footprint: A Tool for Assessing Resource Use in Aquaculture**

To reduce the risk of resource constraints and biodiversity impacts, we need to shift to aquaculture production systems that consume less resources and release less waste within the assimilative capacity of the environment. We also need to recognize and manage nature's life support on which economic development and human welfare depends. One way to identify human demands for natural resource and ecosystem services is by estimating the functional ecosystem area—the ecological footprint—required to support human activities.

The concept has proven useful in illuminating the nonvalued (or monetized) and often unrecognized work of nature that forms the basis for economic activities such as industrial aquaculture. When problems beset aquaculture operations, solutions focus on the pond or cage unit, and it is not considered that the farm is part of a much larger ecosystem with which it interacts. Surrounding ecosystems provide the feed, seed, clean water, and other required resources and services including waste assimilation. This unvalued work of nature sets the limits to culture levels without compromising biodiversity or causing pollution or disease problems. Many aquaculture developments, such as intensive shrimp farming, have encountered problems or even failed because they exceeded the carrying capacity of the environment.

A detailed illustration of the footprint concept is provided by a study on semi-intensive shrimp farming in a coastal mangrove area in Colombia. This study estimated that the spatial ecosystem support or footprint required to produce food inputs, nursery areas, and clean water as well as to process wastes was 35–190 times the surface area of the farm (Fig. 3) (Larsson et al., 1994). The mangrove nursery area required to produce the shrimp seed for stocking was the largest support system covering up to 160 times the pond area. If located close to the farm, the same mangrove area could also supply natural

![Figure 3](image-url)
food inputs (4.2 m² per m² shrimp pond area) and absorb polluting nutrients (2-22 m² per m² pond area) in the farm effluents. Feed pellets form a major input to a shrimp farm, and a marine area of 14.5 m² was needed to catch the fish, and an additional agricultural area of 0.5 m² for the vegetable ingredients used in feed pellet manufacturing was also needed. Finally, 7.2 m² was needed for providing clean lagoon water to the ponds, and 0.8-2.5 m² of forest area per square meter shrimp pond area was needed to sequester the CO₂ of fossil fuel burning at the farm.

Footprint size will change with farming intensity, i.e., a higher stocking density will require more food inputs and produce more wastes (Folke and Kautsky, 1992). Pressure on local ecosystems can be reduced to some extent by importing some inputs (e.g., feeds) from other areas and by investing in shrimp hatcheries, the costs of which can be absorbed in the budget. Although producing seed in hatcheries reduces large amounts of shrimp and fish larvae bycatch that would otherwise be recruited to fisheries (Primavera, 1998; Ronnback, 1999), it also increases demand for wild-caught spawners and broodstock.

Other services, however, such as clean water supply and waste assimilation, must be located close to the farming area. This may be no problem up to a certain level of farming intensity, but the whole operation may collapse when the dynamic carrying capacity of the local environment is exceeded unless extensive and costly pipelines and water treatment facilities are built. The footprint concept provides an early warning device when the level of carrying capacity is being approached. Integrated farming technologies that recirculate resources and wastes within the farm may be one way of reducing the footprint (see Box 7).

Because the surrounding mangrove resources and the lagoon water supply were becoming limiting, the Colombia study recommended no further shrimp farm expansion in the area and preservation of remaining mangroves. According to available information, shrimp farming has not expanded much since the study was made in 1992 and the farms have not been struck by disease problems as severely as has occurred in many other areas.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Resource production</th>
<th>Waste assimilation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salmon cage farming, Sweden</td>
<td>40,000-50,000</td>
<td>—</td>
</tr>
<tr>
<td>Tilapia cage farming, Zimbabwe</td>
<td>10,000</td>
<td>115-275</td>
</tr>
<tr>
<td>Salmon tank system, Chile</td>
<td>—</td>
<td>16-180</td>
</tr>
<tr>
<td>Shrimp farming (semi-intensive)</td>
<td>34-187*</td>
<td>—</td>
</tr>
<tr>
<td>Colombia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shrimp farming (semi-intensive)</td>
<td>2-22</td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>20</td>
<td>—</td>
</tr>
<tr>
<td>Mussel rearing, Sweden</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Values are area of footprint per area of activity (m²/m²)
(Data from Folke et al. 1998).

Ecological footprints have been established for many aquaculture commodities (Table II). The data illustrate that the coastal and marine support areas needed for resource inputs and waste assimilation range from as large as 50,000 times the cultivation area for intensive salmon cage farming to negligible for extensive and semi-intensive tilapia pond farming. The latter uses discards from fisheries and agricultural waste as the sole inputs (Table II).

In its current stage, the concept is useful for communicating the importance of viable ecosystems to farmers and policymakers. Although it is a static measure in its current form, the footprint concept reveals the hidden requirements for ecosystem support and locates (or situates) fisheries and aquaculture within a larger ecosystem framework. Human activities that seem separate from nature would not function without ecosystem support.
Integrated Aquaculture

Commercial aquaculture is usually practiced using large monocultures. Like all types of monocultures, large-scale, and especially intensive aquaculture will result in negative environmental impacts (see Box 5) and make the surrounding waters less suitable or even unusable for other purposes, including the further culturing and harvesting of natural stocks. The problem with these types of cultivations is their throughput characteristics (Daly and Cob, 1989; Folke and Kautsky, 1992), in which large amounts of wastes are released beyond the capacity of the environment to recycle. The fact that treatment of effluents usually involves a high degree of technology and therefore high costs implies that release of untreated water is the rule rather than the exception, especially in many developing countries. Traditional pond cultures of herbivorous fish species (e.g., carp in China), however, have been viable for centuries and their mere existence is proof of sustainable integrated farming systems. Here, raising poultry and livestock is integrated with fish culture, and the principal linkages between raising poultry or livestock and aquaculture are animal manure and other agriculture waste products. Compost is used to fertilize the pond water for proliferation of natural organisms as natural feeds for fish from juvenile to adult. Such systems utilize fish species low in the food web, making efficient use of inputs, and generate little waste. The production also adds, at relatively minor environmental and social expense, to net fish supplies at local and regional levels.

However, increased world demand for seafood products cannot be met by such traditional extensive production systems, and as previously mentioned current modern, intensive aquaculture practices cannot be regarded as a suitable alternative. As an alternative to monocultures and high-tech pollution treatment solutions, the practice of ecological engineering (Mitsch, 1989) may offer solutions to increasing production without deteriorating the environment. The concept has recently gained new interest in an aquaculture context with negative environmental effects being remedied by species integration for nutrient trapping or recirculation. Wastewater from land-based fish and shrimp cultivation and open fish cage cultivation has proven to be a suitable nutrient source for culture of seaweeds and bivalves. The systems have different characteristics depending on how waste is emitted to the surrounding ecosystems (Troell et al., 1999), implying that one solution applied in one type of system may not be optimal in another.

Filter feeders such as mussels and oysters have successfully been integrated with fish aquaculture (Jones and Iwama, 1991; Shigel et al., 1993) and the technique is also being practiced in shrimp aquaculture (Hopkins et al., 1993). Filter feeders in such an integrated system benefit from release of particles, from waste feed and feces from the farm, and from the stimulation of bacterial production and phytoplankton cells. In addition to generating additional income for the farmer, such an integrated system will also secure income by resulting in a more diversified production (Troell et al., 1999). In addition to output of particulate wastes, aquaculture also releases dissolved nutrients, and generally less than one third of the nutrients added through feed will be removed through harvest in intensive fish and shrimp farming (Gowen et al., 1991; Primavera, 1994; Briggs and Funge-Smith, 1994). Many studies have verified that wastewater from intensive and semi-intensive tank or pond mariculture systems is suitable as a nutrient source in seaweed production (Neori et al., 1991; Bouschmann et al., 1994, 1996; Jimenez et al., 1994; Krom et al., 1995; Neori, 1996; Troell et al., 1999), and that integration with seaweeds can significantly reduce the loading of dissolved nutrients to the environment (Lin et al., 1993; Phang et al., 1996). The choice of commercially attractive seaweed species also increases profitability for the farmer in integrated cultivation. In more open culture systems such as cage farming, the continuous exchange of water makes waste disposal difficult to control, but studies have shown beneficial effects from the integration of seaweeds with such culture (Hirata et al., 1994; Petrill et al., 1993; Troell et al., 1997).

The development of integrated techniques in which seaweeds and filter feeders are used as biofilters is in its infancy and further research on ecologically sound production systems is needed. Even though the development of closed, intensive recirculating systems seems to be of high priority (e.g., shrimp culture), the use of biofilters could
See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • ENERGY FLOW AND ECOSYSTEMS • FISH STOCKS

Bibliography


ARACHNIDS

Jonathan A. Coddington* and Robert K. Colwell†
*Smithsonian Institution; †University of Connecticut

I. Overview of Arachnida
II. Araneae
III. Scorpiones
IV. Opiliones
V. Smaller Arachnid Orders
VI. Acari

GLOSSARY

booklung A one to four pairs of abdominal respiratory organs consisting of a thin, multifolded membrane (the book’s “pages”) over which blood circulates and that is open to an air-filled cavity on the outside, itself open to the exterior via a spiracle. Gases passively diffuse back and forth across the membrane.

chelicerae (chelate) The first pair of preoral appendages. They are at most three segmented, usually two, and usually the distal segment acts against the penultimate to grab or hold prey or objects. If the basal segment has a finger-shaped outgrowth against which the distal segment operates, the chelicerae are chelate (as in scorpions or harvestmen). If not, the chelicerae are subchelate, as in spiders and tailless whip scorpions. In parasitic mites the chelicerae are modified into piercing styli.

monophyly A true, historical, evolutionary lineage consisting of an ancestor and all of its descendants; defined by shared, derived characters.

ovoviviparous Young are born alive, but the mother simply retains eggs within her body until they hatch.

paraphyly A group consisting of an ancestor and only some of its descendants. Defined by primitive characters.

pedipalps The second pair of preoral appendages. They are multisegmented and primitively leg-like. They may be raptorial or sensory (like antennae) or used as walking legs.

phoresy A method of long-range dispersal in which the dispersing animal attaches itself to another animal (e.g., beetle, wasp, or bird) that carries the disperser along with it until the disperser drops off or disembarks.

diphylly A group in which the most recent common ancestor of the included taxa is excluded from the group. Defined by convergent, nonhomologous characters.

spermatophore A chitinous container produced by the male to hold sperm. It may be attached to the substrate for the female to find or passed to the female from the male during mating.

spinnerets Usually three, rarely four pairs of modified terminal abdominal appendages in spiders bearing one to hundreds of hollow spigots from which silk is drawn.

tracheae A system of hollow, branched or unbranched air-conducting tubes used for respiration, opening via abdominal spiracles. They may or may not extend into the cephalothorax or legs.

trichobothria Long, delicate, slender setae set in broad, shallow innervated sockets in the cuticle. Trichobothria are sensitive to vibration or near-field air movement and are a major sense organ of arachnids.
Arachnida is a class of the huge phylum Arthropoda. Familiar arachnids are spiders, scorpions, ticks, mites, and harvestmen, but arachnids include many lesser-known terrestrial arthropod groups as well.

I. OVERVIEW OF ARACHNIDA

The known diversity of arachnids is approximately 640 families, 9000 genera, and 93,000 species (Table 1), but there are many thousands of new mite and spider species still undescribed and hundreds to thousands of undescribed species in the remaining orders. Together with the marine horseshoe crabs (Xiphosura) and sea spiders (Pycnogonida), arachnids comprise the arthropod subphylum Chelicerata, named for the characteristic first pair of preoral appendages, the chelate, or pinching, mouthparts. In some arachnid groups the chelicerae are further modified into venomous fangs or piercing stylets to suck body or plant fluids. Arachnids are the only terrestrial chelicerates. Along with the insects, arachnids are by far the most species-rich, abundant, and widely distributed terrestrial arthropods. Acarologists (scientists who study mites and ticks) estimate that there may be as many mite species as beetles, implying that total extant arachnid diversity may exceed 1 million species. Arachnids are an important component of every terrestrial ecosystem, but, apart from several specialized mite lineages, none are aquatic or marine. Although most arachnids share many ancestral similarities in body plan and lifestyle, many extremely specialized groups exist, especially among the mites (Acari).

Although arachnids are commonly mistaken for some sort of peculiar insect, the groups are quite distinct and only distantly related. Arachnids have four pairs of walking legs rather than three (except the young stages of mites and the related Ricinulei), only two (not three) major body parts, and simpler chelate mouthparts rather than the more complex feeding apparatus of insects. The anterior body part is specialized for locomotion and the posterior for digestion and reproduction. Arachnids lack the wings, antennae, and compound eyes usual in insects. In many groups the first pair of walking legs (e.g., amblypygids, uropygids, schizomids, solifuges, palpigrades, and many mites) are elongate and function in much the same way as insect antennae. Figure 1 depicts a consensus view of the position of arachnids in the arthropod evolutionary tree.

Arachnids and hexapods (which insects dominate) differ in fundamental ways, possibly because their marine ancestors were already distinct lineages in the Silu-
rian and both colonized land independently. They therefore solved the fundamental challenges of terrestrial existence (support, breathing, water balance, reproduction in dry environments, and nitrogenous waste management) in different ways. The arachnid skeleton is hydraulic; arachnids, except scorpions and pseudoscorpions, lack extensor muscles at key joints. Instead, the animal pumps blood into the limb to extend it. The basic arachnid uses two (or four) pairs of book lungs (gas-permeable, gill-like membranes with blood on one side and open to the air on the other) to exchange carbon dioxide for oxygen rather than tracheae. Many arachnids possess rudimentary tracheal systems that supplement or replace book lungs, but ventilation is passive, not active. Unlike insect tracheae, arachnid tracheae generally do not ramify throughout the entire body or penetrate inside body cells. For all these reasons, insect tracheae more efficiently deliver oxygen directly to tissue, and insects in general can lead more energy-intensive lifestyles than arachnids. Arachnids have significantly lower metabolic rates than other terrestrial arthropods, especially insects. Whereas male insects transfer sperm directly to the female using an intromittent organ, in most arachnids males either ejaculate onto a special structure and carry the sperm mass with a specialized appendage until the female is encountered or they deposit the sperm mass in a specially built receptacle (spermatophore) fixed to the substrate, which the female picks up. Major exceptions include the harvestmen, astigmatic mites, and spider mites, which transfer sperm to the female by means of an intromittent organ, and the water mites (Prostigmata: Hydracarina), which transfer sperm directly by opposing the male gonopore to the female gonopore.

Arachnids are peculiar among animals in using guanine (three nitrogen atoms per molecule) as well as the much more common uric acid (two nitrogen atoms per molecule) to eliminate nitrogenous wastes. Insects have compound eyes, which provide relatively excellent vision. Arachnids lost compound eyes early in their evolutionary history but retain usually one to five (commonly four) pairs of simple eyes, much inferior in acuity to compound eyes. Schizomids, palpigrades, ricinuleids, mites, and other mainly litter-dwelling arachnid groups are nearly always blind. Vision is much less important to arachnids than vibration. Many structures (e.g., sense organs, trichobothria, and lyriform organs) are specialized to detect minute vibrations and slight air currents.

Arachnids are also peculiar in that species of most arachnid groups digest food externally. They have a strong pumping stomach that rhythmically vomits and sucks digestive juice through a preoral cavity formed by the basal articles of the pedipalps back and forth over their prey. The process continues until only the hard, indigestible parts of the prey remain. Only liquids or very small particles are actually ingested. Some major groups of arachnids have internal digestion, however. Opilionids and several groups of mites are particulate feeders, and parasitism of plants, vertebrates, and invertebrates has arisen repeatedly in the mites. External digestion is a major obstacle to life in fresh water or the sea. Arachnids are also notable for their ability to withstand starvation. Fasts of weeks or even months are routine for larger arachnids. Some scorpions and mygalomorph spiders live for more than 1 year without food, and adult soft ticks (Argasidae) can survive for years without feeding.

A. Reproduction and Growth

Like all arthropods, arachnids grow by molting their exoskeletons and expanding the larger skin beneath with blood pressure before it hardens into the usual tough covering. The number of molts to maturity varies widely between 3 and 10–12; five is perhaps the most common. Life spans also vary greatly. The majority of mites and spiders live less than 1 year, but several years is common among the larger forms, and mygalomorph spiders can live 20–30 years in captivity. Some arachnids cease molting at adulthood but others continue to molt periodically until death.

The ancestral reproductive pattern of sperm transfer is via spermatophore, modified in spiders, harvestmen, mites, and ricinuleids. Except among permanently social species and many mites, the sex ratio is equal, and parthenogenesis is rare. In the vast majority of species, males and females meet only to mate; cohabitation and parental care are uncommon. Nevertheless, various spiders, scorpions, schizomids, uropygids, and amblypygids may carry and feed their young, and in some harvestmen the male cares for the eggs in a specially built nest. Pseudoscorpion females nourish eggs with secretions from their bodies, and scorpions bear only live young. Females commonly guard their eggs until they hatch, but the young are usually abandoned soon thereafter. Mites are more diverse in reproductive strategies than noncarine arachnids.

B. Ecology

Despite huge numbers of species (Table 1), arachnid biology is coherent in many ways (even though exceptions to nearly every generalization exist). Most arachnid orders consist of fluid-feeding predators, and predation still dominates these groups today. However, opilionids and two of the three orders of mites are
particulate-feeders on detritus, fungi, and small invertebrates. Additionally, parasitism of vertebrates, invertebrates, and plants has arisen numerous times within the Acari and radiations within these lineages account for most of the 43,000 described species of mites.

Non-acarine arachnids tend to be at the top of the terrestrial invertebrate food chain wherever they occur. At one site in Israel, mites comprised 35% of the total soil arthropod population; in the Amazon ranges from 35 to 55% have been reported. Some harvestmen eat dead or decaying animal or plant material. Arachnids are generally nocturnal, despite numbers of diurnal harvestmen, spiders, and mites. Nocturnal forms hide in dark crevices and burrows during the day; several orders are morphologically specialized to inhabit small spaces. With the exception of the wind spiders (Solifugae), arachnids tend to be torpid and sedentary—none fly, for example, nor do any move constantly like ants or other active insects. The basic arachnid forages with a "sit and wait," solitary strategy. They move rarely and wait for prey to encounter them. Prey is then seized with a quick strike and immobilized. Highly organized social systems are known only in a few spiders, but loose aggregations (spiders, harvestmen, and pseudoscorpions) are not uncommon, usually in response to high prey density or limited refugia. Most arachnids are well adapted to last for long periods (weeks or months) without food; in the laboratory some have survived for more than 1 year. Only jumping spiders among arachnids have notably good vision; otherwise, they orient primarily via vibrations and touch.

Arachnids occupy all terrestrial habitats—deserts, forests, tundra, grasslands, mountaintops, soils, litter, caves, etc. Hydrachnid mites (approximately 5000 species) are important components of most freshwater ecosystems; other mites are parasites of marine organisms, whereas others inhabit marine sediments, including the deepest oceanic trenches. Otherwise, arachnids are exclusively terrestrial. A few groups, such as scorpions and wind spiders, conserve body water as well as any arthropod and thus tend to dominate in deserts. The majority needs moist conditions to survive. Schizomids, palpigrades, and ricinuleids are apparently restricted to the interstices of moist tropical and subtropical leaf litter or equally constant and moist habitats.

C. Phylogeny and Taxonomy

The most commonly encountered arachnids are spiders, scorpions, harvestmen, and mites, but the class contains seven smaller groups of terrestrial arthropods less familiar to the general public (Fig. 1). The largest and heaviest arachnid is the African scorpion Pandinus imperator, which may reach a length of 18–20 cm. The smallest are perhaps the gall mites at 80 μm. No group of arachnids is well-known by vertebrate, butterfly, or vascular plant standards. Popular manuals are available for only a smattering of the most common species of spiders in Europe, North America, and Japan; all others require technical literature and specialist knowledge to identify. Myriad species are undescribed and undiscovered; at best, the approximately 93,000 known arachnid species are but one-third of the probable total, and probably much less. Most undescribed arachnid species are mites.

The study of arachnids is called arachnology. The principal international scientific society for nonacarines has approximately 600 members but many more belong only to regional societies. The taxonomy of arachnids is still a monumental task and an obstacle to better ecological and biotechnological understanding of arachnids, but the number of arachnid taxonomists is small and decreasing, and new students are not being trained. There are no comprehensive arachnology texts appropriate for university teaching (three exist for Acari), although modern "biologies" are available for spiders, scorpions, solifuges, pseudoscorpions, and various aspects of mite biology.

D. Paleontology

Arachnida were among the earliest terrestrial animals. The marked similarity between fossil and recent forms in overall body plan and morphology suggests few changes over hundreds of millions of years. The earliest sites for terrestrial arachnids are Early Devonian (400 Ma) and Late Silurian (414 Ma); the extinct arachnid order Trigonotarbidida figures prominently, but mites are also present. The fossil record of all arachnids is comparatively poorly known. The 13 living and extinct orders are still known from less than 50 major time horizons since the Silurian; gaps remain more common than fossils. Arachnids seem to have invaded land in the Silurian and reached a pinnacle of ordinal diversity by the Carboniferous. The latest dates for the extinct orders Trigonotarbidida and Phalangiotarbi are Permian and Carboniferous, respectively. No order seems to have succumbed to the end Cretaceous event that eliminated the dinosaurs. The earliest arachnid fossils are aquatic scorpions from the Late Silurian. Spiders, pseudoscorpions, terrestrial scorpions, and mites are known from the Devonian. The scorpion-like Eurypterida, which are sister to true Arachnida, also may have become extinct at the Permio-Triassic boundary; the youngest fossils are also Carboniferous. The phylogeny in Fig. 2 implies that many arachnid clades must predate all scorpion fossils; an alternate opinion, based on less
evidence, is that scorpions instead are sister to eurypterids, and that the phylogeny of Fig. 2 will be shown to be incorrect. Small size may explain the rarity of schizomid and palpigrade fossils, but mites in the Devonian show that small arachnid fossils can persist. Given the sparse record, it may well be that all arachnid orders were distinct and essentially modern in appearance by end Silurian–Early Devonian times. The most detailed arachnid fossils are “inclusions” in amber, which is fossilized gum or tree sap. The organisms became caught in the wet sap and were fossilized along with the sap. Cenozoic amber arachnid fossils 40–80 million years old can mostly be placed in modern genera; in some cases, the fossils are difficult or impossible to distinguish from modern species. Ecological relations are also preserved: Parasitic mites occur on their insect hosts, for example, thus documenting the antiquity and diversity of mite parasitism.

II. ARANEAES

Araneae is the second largest arachnid order with 108 families, 3200 genera, and approximately 37,000 species described (second to mites). Spiders are distinguished from other arachnids by their silk-producing spinnerets at the end of the abdomen and the prosomal poison glands exiting through their chelicerae modified as fangs (Fig. 3). Spiders are among the very few animals that use silk throughout their entire lives. A narrow stalk joins the abdomen to the prosoma, allowing great flexibility and precise orientation of the abdominal spinnerets. Pedipalps are leg-like and short. Colors are predominantly dull tans, browns, and blacks, but spiders are occasionally very colorful, even iridescent. The abdomen of most spiders shows no trace of ancestral segmentation, unlike that of other arachnids. The Neotropical tarantula Theraphosa leblondi (Theraphosidae) is the largest spider at about 10 cm in body length. The smallest spiders are the tiny orb-weaving Symphytognathidae; adults are less than 1 mm long.

A. Ecology

Spiders are very abundant. One calculation estimated 5 million animals per hectare in an English meadow. Another found 29,000 per cubic meter in an English sewage treatment plant. These are extreme values, but “average” non-desert habitat probably supports at least 1 and as many as 800 spiders per square meter. Many spider species disperse by ballooning. Ballooning spiders spin a silk line until it is caught by the wind and lifted aloft—potentially for hundreds or thousands of kilometers. One study estimated that 216,000 spiders per hectare may balloon into tilled fields during the growing season. Point diversities per hectare may vary from 100 species in moderate temperate zones to 600 or more in the wet tropics. Boreal diversities are less, perhaps 20–50 species per hectare.

Spiders are the only animals to use silk throughout their lives. Silk is one of the strongest and toughest natural fibers known and compares favorably with the best man-made filaments. Eight different kinds of spider silk have been discovered, but the maximum made by
any single species is seven. Orb web spinners, for example, make two varieties of stiff, tough silk for weight-bearing structural fibers and safety lines, cement to fix silk to itself and substrate, sticky silk or glue to capture prey, rubbery silk to carry the sticky silk in the web, specialized silk for eggsacs, and thin, weak silk spun as multiple fibers to wrap prey, cradle eggs, and for other general purposes. All spiders are capable of spinning silk as soon as they leave the eggsac, and all make at least safety lines ("draglines") and the cement to attach them to substrate. Approximately half of spider species spin webs to capture prey. Web architectures are taxonomically specific and provide many clues for reconstructing spider evolution. The remaining spiders are ambush predators such as the crab spiders, which lurk inside flowers to attack pollinators, tube or retreat dwellers that forage in the very limited area at the burrow mouth, or vagabond predators such as wolf, ground, or jumping spiders.

B. Reproduction and Growth

Reproduction in spiders requires the male to ejaculate sperm onto a specially constructed sperm web. He then sucks up the sperm into specialized sperm transfer organs at the pedipalpal tips. Only adult males have such structures; their form is usually species-specific. With his palps filled, the male then searches for females. Females are usually more sedentary than males. Life is short for adult males, both because of predation and because they eat little as adults. The duration of courtship and copulation varies from seconds to days, but it ends with the insertion of the male pedipalp and transfer of sperm into the female gonopore. Females store the sperm and sometime later construct a silken eggsac into which 1–2500 eggs are placed. Eggs are fertilized only as they exit the female’s body. Parental care is rare and highly variable, ranging from simple guarding of eggsac to actively feeding the babies. In at least one crab spider the female dies as the juveniles emerge from the eggsac, which then eat their mother’s body. About 20 species of spiders are extremely social. Males are rare (1:40–100), generations overlap, food is shared, and prey capture and brood care are cooperative. Colony sizes range from a dozen to several thousand individuals.

C. Phylogeny and Taxonomy

Spiders are the seventh largest zoological order on Earth (after Coleoptera, Hymenoptera, Lepidoptera, Diptera, Hemiptera, and Acari), and of these they are the only one for which all taxonomic literature is fully cataloged. Catalogs greatly facilitate all kinds of research because scientists can easily determine the current taxonomic status and history of any described species and thus decide whether a given specimen belongs to a described species or not. At higher taxonomic levels (approximately, family) a basic, first-draft phylogeny is nearly complete. At the species level the easiest species-specific characters are found in the male and female genitalia. Species boundaries in spiders are generally clear-cut. However, the species taxonomy of spiders is based overwhelmingly on morphology. If more costly and sophisticated molecular methods were routinely applied, the number of distinguishable spider “taxa” would certainly increase.

The fundamental phylogenetic division in spiders is between the primitive mesothelae (spinnets towards the middle of the abdomen) and the opisthos heteles (spinnenets terminal). Within opisthosetheles there are again two basic groups, the mygalomorphs (tarantulas and their allies) and araneomorphs (so-called “true” spiders). Mesothes and mygalomorphs are not particularly diverse at the species level; araneomorphs currently include 94% of all known spider species and this disparity will certainly increase. Within araneomorphs the basal taxa are a few relictually distributed families in north and south temperate regions. Araneomorph haplogynes are diverse, but again comprise relatively few species. The araneomorph Entelegynae includes the bulk of modern spiders. Seven spider families currently contain more than 1000 species—all are entelegyne.

During a recent 39-year period, an average of 314 new species were described per year (12,200 total), but an annual average of 104 old names were synonymized, for a net gain of approximately 200 species per year (8800 total). Estimates of total spider diversity range from 76 to 170,000. Lower estimates mostly extrapolate from the proportions of new versus known species in taxonomic publications or are based on comparisons to well-known groups. The higher estimates take into account that many regions, particularly those richest in spider species, are disproportionately undercollected. In any case, the real diversity of spiders will never be known because a potentially great fraction will certainly go extinct before being discovered, much less described.

1. Major Lineages
a. Mesothelae

Liphiididae is the only extant family and is limited to areas of Southeast Asia and Japan. Only a few dozen
species are known, but some are common where they occur. Liphistiids retain many primitive morphological features, such as eight (rather than six or fewer) segmented spinnerets that insert anteriorly rather than terminally on the semi-segmented abdomen. Their biology may likewise represent the ecological “ground plan” for spiders. Liphistiids live in silk-lined tubes equipped with rudimentary trap doors in banks and cave entrances. Sometimes, silk “trip lines” lead away from the burrow entrance to extend the sensory radius of the animal. They are nocturnal, ambush predators. They live for 5–8 years, are remarkably sedentary, and consume a catholic diet of mainly walking prey.

b. Mygalomorphae

Mygalomorphs include the tarantulas or baboon spiders (Theraphosidae), trap-door spiders (Ctenizidae, Actinopodidae, Migidae, etc.), purse web spiders (Atypidae), funnel web spiders (Hexathelidae), and several other families with no common name. Mygalomorphs number 15 families and approximately 2200 species, but several of the families are para- or polyphyletic. A more realistic number is 20–30 “family-level” groups. Like mesothelae, mygalomorphs tend to live in burrows and forage at the burrow entrance or for a very limited distance around it. Some theraphosids are arboreal and spin elaborate silken retreats. Diplurids make extensive webs but are virtually unique among mygalomorphs. The venomous Australian funnel web spiders (Atrax and Hadronyche; Hexathelidae) were responsible for many deaths until an antivenin was developed in the 1980s. The large theraphosid baboon spiders are not seriously venomous to humans, despite their popular reputation.

c. Araneomorphae

Araneomorphs include approximately 94% of known spider species. Even the most primitive araneomorphs are very different from mygalomorphs and mesothelae. Basal araneomorphs tend to be much smaller, and most are obligate web spinners with elaborate spinnerets capable of making adhesive “cribellate” silk. Cribellate silk is adhesive due to the extremely fine threads drawn from the cribellum, the very modified and fused anterior pair of spinnerets. Adhesive silk makes feasible a greater variety of web architectures, and these basal araneomorphs spin elaborate catching webs. The sister group of all remaining araneomorphs is the family Hypochilidae; its dozen or so species are limited to the Appalachian Mountains, a few places in western North America, and equally restricted sites in China.

i. Haplogynae

Haplogynae comprises 17 families of spiders of diverse habits and worldwide distribution. Filistatidae are cribellate web-spinning spiders. Their web architecture is not much different from that of the mygalomorph diplurids—a sheet that narrows to a silk-lined retreat in a tube or crevice. Diplurids make dry silk webs. The majority of haplogynes are leaf-litter specialists and are vagabond, webless predators. The venomous brown recluse spiders (Loxosceles spp.) are haplogyne scorpions. The sister group of Loxosceles in southern Africa (Sicarius) has also been implicated in medically serious bites.

ii. Entelegynae

The remaining 70 families of spiders are the Entelegynae. Entelegynes share many evolutionary novelties. Females have a convoluted abdominal plate protecting their gonopore, which male genitalia must navigate successfully to achieve insemination. Sperm are stored in a unique “flow-through” system so that the female reproductive tract has two apertures to the outside. Female entelegynes also make special silk used exclusively in egg sacs, although its exact role is unknown. The lateral eyes possess a canoe-shaped tapetum that in some spiders enables orientation via polarized light. Although the higher phylogeny has been worked out for many entelegyne lineages, some very large ones remain unstudied, and the relationships between entelegyne lineages are also controversial.

Lycosidae includes 11 families of mainly hunting spiders, some with common names: lynx, wolf, fishing, or tropical wolf spiders. Lycosidae (wolf spiders) and Pisauridae (fishing spiders) are common, cosmopolitan lycosid spiders. Lycosoids occur in all terrestrial habitats, and some are semiaquatic in their ability to run across the surface of the water or dive beneath the surface. Web spinning is rare among lycosoids; some may have regained it after evolutionary loss. Most species are vagabond predators or, occasionally, tube dwellers. Active at night, they move sporadically or wait until prey approaches and then attack with powerful front legs and chelicerae. They are built strongly and run and jump with agility. The South American ctenid *Phoneutria* is venomous to humans.

Dionycha includes 17 families of spiders with two tarsal claws and a tuft of hair rather than the more common three claws and no claw tufts. The monophyly of Dionycha is by no means certain. Dionycha are also hunting spiders and have habits similar to those of lycosoids. Crab spiders (Thomisidae) wait for insect prey in flowers. Jumping spiders (Salticidae), the largest spider family, can be very brightly colored and often prefer to jump rather than walk. Their vision is superior.
to any other arachnid; salicids are the only sizable spider lineage that is strictly diurnal.

Orbiculatiae includes 14 families of spiders and approximately 12,000 species. Most orbicularians spin prey-catching webs, but a few groups have secondarily lost the web-spinning habit. The primitive web architecture seems to be the orb—the classic spider web of radially symmetric, stiff, dry spokes supporting a spiral of sticky silk—but more orbicularian species have lost or modified the orb architecture than retained it. Web spiders rely exclusively on webs for prey capture. Araneidae (common orb weavers), Linyphiidae (sheet weavers), and Theridiidae (cobweb weavers) are the largest families. The venomous widow spiders (Latrodectus) are theridiids and are distributed worldwide; several species are spread by humans and are now cosmopolitan.

III. SCORPIONES

About 1256 species of scorpions are currently known in 156 genera and 18 families (Fig. 4). Scorpions are one of the better collected arachnid groups so that huge increases in diversity are not as likely as in mites or spiders. Estimates of total diversity run as high as 7000 species. All scorpions have large, obvious pedipalps modified as pinces, both body regions are broadly joined, the distal abdomen is narrowed into a flexible tail bearing a venomous stinger at the end, and ventrally. The abdomen bears a pair of comb-like sensory appendages known as pectines. Colors vary from translucent to brown or black. Curiously, they fluoresce under ultraviolet light, a discovery that has galvanized recent field research on these animals. The longest is Hadogenes troglodytes at 21 cm, but Pandinus imperator is nearly as long and much heavier. The external morphology of recent scorpions is impressively similar to Silurian fossils. Formerly scorpions were thought to be the sister group of all other arachnids because they closely resemble the extinct marine eurypterids that are the sister group of all arachnids. Better analysis of morphological data (Fig. 2), weakly corroborated by molecular evidence, suggests that scorpions are more deeply imbedded in the arachnid clade and merely retain many primitive features. The issue is controversial.

Scorpions are the only arachnids with a narrow post-abdomen ("tail") terminating in a venomous stinger. The stinger is most often used for defense, although scorpions will sting large or strong prey. The sting of most scorpions is painful—like wasp or hornet stings—but not dangerous. Characteristically, scorpions with slender pedipalps are more prone to sting their prey, whereas those with robust pedipalps tend to crush prey. The Central American genus Centruroides, Brazilian Tityus, and Old World Androctonus, Leiurus, Mesobuthus, and Parabuthus are very venomous and medically important. In Mexico, Centruroides spp. sting 300,000 and kill 1000 people annually; Androctonus, Leiurus, and Mesobuthus kill thousands annually in Egypt and Pakistan alone. Excepting ticks that spread disease, scorpions are by far the most dangerous arachnids to humans. Scorpions are correspondingly prominent in mythology and folklore (e.g., the zodiacal constellation "Scorpio"). Scorpion venoms typically contain multiple low-molecular-weight proteinaceous neurotoxins. Scorpion blood inactivates scorpion venom, but if the venom is injected directly into a nerve, the animal rapidly dies. Parabuthus transvaalicus and P. villosus squirt venom to damage corneas, like spitting cobras.

A. Ecology

Scorpions are most diverse in deserts or similar dry areas, although they are reliably present in moist ecosystems if the temperature is not too cold. They now occur
on all major landmasses except Antarctica. Favored habitats are burrows, under bark, stones, or logs, or inside small crevices. Burrows may be as deep as 40–80 cm, serving to escape the hot daytime temperatures in deserts. Because they like hard substrates and dry conditions, scorpions adapt well to human structures. Most scorpions are ground dwelling, but many species are arboreal, especially in the neotropics. In canopy fogging at four Amazonian sites in Peru, all of approximately 100 specimens were Buthidae (J. Ochoa, personal communication). A few are limited to lightless caves.

Scorpions are almost invariably nocturnal, although the African *P. villosus* is predominantly diurnal. The eyes seem to detect luminosity at best. Prey movements are detected by tarsal sense organs at distances up to 15 cm, and prey are attacked in a single motion. At distances up to 30 cm prey are located through orientation responses. The large, pincher-like pedipalps immobilize prey; thereafter, pieces are torn off by the chelicerae and digested in the pre-oral cavity before being sucked into the gut. Scorpions can be important consumers in some communities. In Israel, *Scorpio maurus* annually ate an average of 11% of the isopod population, which was not the only item in their diet. At moderate densities of 1.5 kg/ha, *Urodacus yaschenkoi* ate an annual average of 7.9 kg/ha of prey. Cannibalism and predation by other scorpion species are thought to be the most important sources of mortality, but other top invertebrate predators (e.g., spiders) and vertebrates are also important scorpion predators. Generally, mortality is highest immediately after birth, lower for intermediate-aged animals, and high for adults (e.g., 65, 30, and 60%, respectively, per year for the Australian *Urodacus manicus*). Scorpion mortality is particularly high among males because of their high activity levels and mobility during the breeding season. Cannibalism by females is a significant cause of male death. Biased adult sex ratios of 1.2–1.4:1 are typical. Communal behavior, however, does occur. For example, communal groups up to 15 individuals of the Brazilian *Ophiocanthurus cayaporum* cooperate to construct and occupy communal chambers in the center of termite mounds. The African *Heterometrus* spp. also construct and share a communal burrow, inhabited by individuals of various ages. “Piles” of 20–30 individuals of *Centruroides exilicauda* are found in the winter months. Groups of 5 individuals of *Mesobuthus martensi*, all of the same age and all with their heads oriented toward a central spot, have been found under wet rocks in the intertidal zone.

The vast majority of scorpion species are subtropical or tropical. Point diversity (the number of species sympatric at one site) peaks in subtropical deserts and is particularly high (6 or 7, with a maximum of 12) in Baja California. Two to three species per site is more usual. *Vaejovis littoralis* reaches unusually high densities (8–12/m²) in the drift line along the Gulf of California. The North American *Paruroctonus boreus* occurs as far north as British Columbia and Alberta, and the European *Euscorpius germanus* reaches the southern Alps. Even tropical scorpions sometimes inhabit extreme conditions; *Orobothriurus crassimanus* was collected at 5560 m in the Peruvian Andes.

B. Reproduction

Reproduction in scorpions is via a spermatophore attached to the substrate. The male completes production of the spermatophore inside his body, deposits the sperm inside, and attaches the spermatophore to the substrate, all while holding on to the female during preliminary courtship. The spermatophore is “spring-loaded” and catapults the sperm mass into the female gonopore when a lever is touched. Scorpions are exclusively ovoviviparous or viviparous. The 1–105 young are born live and cling to the mother for the first few molts. A few species are parthenogenetic. Scorpions live 4 or 5 years (rarely 8); they do not molt as adults.

C. Phylogeny and Taxonomy

The higher classification of scorpions has changed dramatically as classical data have been reinterpreted phylogenetically. The old system proposed a few, huge, polyphyletic families about which nothing much in general could be said; now 16 or 18 families with increasingly coherent biologies are recognized. Species limits in scorpions are often difficult because scorpion genitalia are usually not species-specific. By tradition, scorpion taxonomists use the subspecies category more than most arachnologists. About 150 subspecies are recognized in addition to the 1260 species, but because these are easily distinguished they are probably distinct species. A classic example is the 25 non-overlapping, fully distinct subspecies of *Scorpio maurus*. Species-level taxonomic characters include the surface sculpturing of the exoskeleton, morphometric data, the number and position of pedipalpal trichobothria, and the hemispermatophores—internal male structures that produce the spermatophore.

1. Major Lineages

The basal division in Scorpionidae is between the bu- thoids and remaining scorpions. New and Old World
buthids are also distinct lineages. Scorpioidoids and the vaejovoid–chaetoid lineage are the remaining major scorpion lineages. Chaerilidae and Pseudocactidae (Chaerillus, 21 species; Pseudochactus ovchinnikovi, from Kazakhstan) are monogenic and enigmatic; they are like none of the other scorpion families and their relation to other major lineages is obscure. They may be basal buthoid groups.

Buthidae is the largest and most widely distributed scorpion family with approximately 74 genera and 531 species. Buthidae is most diverse in the African tropics and Palearctic regions. Buthids tend to have slender, elongate pincers, a robust tail, and usually a tubercle under the sting. All scorpions considered dangerous to humans are buthids. Buthids are also the most diverse ecologically and occupy humid, mesic, and dry habitats. The small family Microchaeridae is an Afro-tropical buthid segregate with two genera and six species.

The Scorpionidae (36 genera and 355 species) is a large, monophyletic lineage that includes Bothriuridae, Diplocentridae, Heterocroniidae, Hemiscorpiidae, Ischnuridae, Scorpionidae, and Urodactylidae. Scorpionidae lack a tubercle under the sting and the sides of the sternum are parallel. The family contains the genus Scorpio from the Mediterranean and Near East, much mentioned in classical Greek, Egyptian, and Christian myths, and Pandinus, the giant African black scorpion. The longest and heaviest scorpions are scorpionids, which are exclusively Old World.

The relationships and monophyly of the vaejovoid–chaetoid linage (43 genera and 493 species: Chactidae, Euscorpiidae, Iuridae, Scorpioptidae, Superstitionidae, Troglotayosidae, and Vaejovidae) are the most problematic areas of scorpion higher taxonomy and phylogeny. The large families Iuridae and Chactidae, in particular, are doubtfully monophyletic, although each includes many clearly valid groups. Together, the vaejovoid–chaetoid lineages comprise about one-fourth of known scorpion species, including the most common species in North America.

**IV. OPILIONES**

The described world fauna of Opiliones (harvestmen or daddy long legs) comprises approximately 44 families, 1554 genera, and about 5000 species (Fig. 5). The largest harvestman is Troglomus torosus at 2.2 cm long. The anterior and posterior body regions are broadly joined and the abdomen is rather short, giving the body a wider and rounder appearance than those of other arachnids. The second pair of walking legs is usually the longest. Opilions have just two eyes (cave or litter groups are sometimes blind). All harvestmen have a pair of glands that open via large pores on the anterolateral margins of the body; the function of their secretions is apparently diverse. Harvestmen are the only arachnids in which males have a penis. Females have a long, flexible, and extensible ovipositor (as do many mites). The most common group in north temperate regions, the Phalangioidea (daddy long legs), have soft and flexible bodies, weak mouthparts, and extremely long legs (commonly 15 times the body length), but just as many harvestmen are fantastically armored with bizarre, huge chelicerae, raptorial pedipalps, and short, stiff legs. Others are mite-like, and still others are dorsoventrally flattened. Colors are usually subtle patterns of brown, gray, or black, but tropical forms can rival anything seen in spiders or mites. In the long-legged forms, the distal leg tip is divided into numerous false segments that form a prehensile tip. It can be wrapped around objects to achieve a very firm and adaptable grip. The long second legs may double as feelers.

The respiratory system is exclusively tracheate. The very long-legged Phalangioidea have accessory spiracles on distal leg articles. Touch and vibration perception, as in most other arachnids, seems to be the dominant sense. The eyes at best distinguish light and dark and direction of light. Opilions consume a broader diet than any other arachnids other than mites. The basic pattern is predation, but some, for example, specialize on snails, which are otherwise rarely consumed by noncarmine arachnids. Opilions also are known to eat dead insects, fruit, and decaying vegetable matter. Unlike other arachnids, harvestmen can ingest solid particles, as shown by sclerotized bits in their excrement.
Many harvestmen defend themselves against attack by shedding legs. One study of two species in Louisiana found that about half the animals had lost one leg, but less than 10% lacked three. Seven-legged harvestmen seem to survive and function as well as intact animals. A shed leg continues to jerk and twitch attractively for minutes, permitting the harvestman to escape. All harvestmen have paired “repugnatorial” glands on the front margin of the body. When legs are pinched a droplet of fluid appears at the orifice which may be dabbed on an attacker with a leg or allowed to evaporate. The secretions can also help aggregating species to find each other, and some show broad antibiotic and anti-fungal activity, presumably useful to litter and soil-dwelling forms. Quinones are a major ingredient. Soil-dwelling harvestmen and the short-legged Laniatores are slow compared to long-legged harvestmen.

**A. Ecology**

In general, harvestmen prefer moist, or at least not xeric, environments. The mite-like Cyphophthalmi live in dark leaf litter, caves, or under stones. The largely tropical and usually short-legged Laniatores move slowly over vegetation or the forest floor. The usually long-legged Eupnoi can be anywhere, but their very long legs with prehensile tips are specialized for crossing the large gaps between the leaves of trees, shrubs, and herbs. The northern European opilionid fauna comprises approximately 24 species and that of North America approximately 235 species. In temperate regions, diversities of more than a dozen species per hectare are uncommon.

**B. Reproduction**

Uniquely among non-acarine arachnids, sperm transfer occurs via the male penis. Mating occurs quickly and seemingly without courtship. The male faces the female and pushes the penis underneath her body or between her chelicerae and into the gonopore. After insemination, the sexes separate and continue their solitary wanderings. Females use their long, flexible ovipositors to deposit eggs in suitably protected areas. Trogulids deposit their eggs only in empty snail shells, and other groups oviposit beneath stones, deep into soil, underneath bark, or in bore holes in plant stems left by insects, usually abandoning the eggs once laid. Newly hatched animals are active and resemble adults. Five to eight molts to maturity are common. An unusual reversal of sex role occurs in the Panamanian harvestman Zygocephalus albus marginis. Males fight to occupy existing mud nests or construct their own. Females wander between nests, courting the males, mating, and ovipositing in a series of nests. They have nothing more to do with the offspring. Males accumulate eggs of different ages and from different females and defend the eggs against conspecifics and ants.

**C. Phylogeny and Taxonomy**

The phylogeny of Opilionidae has recently been clarified at the superfamily level, but many additional changes are expected in familial arrangements. Some families seem to be based only on primitive features (e.g., Travuniidae, Phalangodidae, and Triarachnychidae). Eupnoi and Dysnoi classically formed the suborder Palpatores, but increasing evidence indicates that this taxon is paraphyletic. The number of recognized families has approximately doubled in the past 20 years. At the species level the morphology of male genitalia is especially diagnostic.

1. **Major Lineages**

   a. **Cyphophthalmi**

   This suborder contains five (or six) families and about 100 species. It is sister to all remaining Opilionidae (the “Phalangidae”). Cyphophthalmi are eyeless, live in deep moist leaf litter or caves, and range in size from 1 to 7 mm. The animals have a hardened plate covering the entire dorsal surface, and they resemble mites. Siro is North American and European, but the family also occurs in Southeast Asia, Turkey, Japan, Mexico, and South Africa. Life spans of up to 7 years have been reported.

   b. **Eupnoi**

   This group contains two superfamilies, including the classic daddy long legs (Phalangioidea: Phalangiidae, Sclerosomatidae, Megalopacidae, and Neoptilidae) of soft-bodied, long-legged harvestmen. Phalangium opilio is common around buildings and introduced throughout the world. The sclerosomatid Leiobunum spp. are common in North American and European forests, in which they move easily across the upper vegetation. They are predators and scavengers. Cadidida includes only one or two families of harvestmen with enlarged eyes and short legs, sometimes common on tree trunks.

   c. **Dysnoi**

   This group also includes two superfamilies, but has many fewer species than Eupnoi. Ischyropsalidoidea
contains three families and just seven genera. Ischyropsalis is European and feeds on snails. Troguloidae contains four families, two monotypic and several genera in the remaining families. Nemostoma is common in caves. Trogulidae are peculiar harvestmen that look like giant, flattened mites. Legs are very short. Trogulids live under stones and in leaf litter; snails are an important part of their diet.

d. Laniatores

This group is morphologically more diverse than the previous groups and tends to be more diverse in the tropics, including many colorful species. Three superfamilies are recognized: Travunioidea (4 families), Oncopodoidae (1 family), and Gonyleptoidae (18–20 families). Colloquially, laniatorids are known as “short-legged” harvestmen because the most common laniatorids do have short legs, but some agoristenids, cosmetids, gonyleptids, caelopigines, progonyleptoidelines, and mitohatines have legs comparable in length to those of Palpatores. The Gonyleptoidae contains 18–20 families with raptorial pedipalps for prey capture and enlarged fourth coxae with spreading appendages—perhaps a defense against being swallowed whole by predators or dragged down the burrow of a parasitoid. Gonyleptidae are typical heavily armored and spiny, often colorful, and usually larger than 4 mm. They are exclusively Neotropical and, with more than 100 genera, are one of the largest families. The closely related Cosmetidae superficially resemble gonyleptids—large, heavily armored, often colorful, and slow-moving New World harvestmen. Phalangodidae is a large cosmopolitan family of more than 150 genera defined mainly by the features they lack. For years it has been used to file taxa with no obvious relatives, and therefore its biology makes little sense. Although most diverse in the northern neotropics, their distribution includes southern North America. Their pedipalps are flattened with spiny margins.

V. SMALLER ARACHNID ORDERS

Amblypygi (whip spiders or tailless whip scorpions), Uropygi (whip scorpions or vinegaroons), Schizomida (no common name), and Palpigradi (micro-whip scorpions) are all small, unfamiliar arachnid groups that are closely related to each other and to spiders (Fig. 2). Like primitive spiders, all have two pairs of abdominal book lungs, although the second pair is missing in the tiny Schizomida and palpigrades lack both. The first pair of walking legs is elongate and feeler-like, with false articulations in the terminal articles to promote flexibility. Pseudoscorpiones (pseudoscorpions), Solifugae (wind spiders), and Ricinulei (no common name) are a miscellany of remaining arachnid orders related to various orders already discussed, as illustrated in Fig. 2.

A. Amblypygi

Approximately 125 species of amblypygids in 20 genera and 5 families are known (Fig. 6). The American Acanthephyrus coronatus is the biggest, at about 4.5 cm in body length. The usual adult body size is about 4–6 cm. The first walking legs of amblypygids are enormously long, and fully stretched a large animal can span 50 cm. Whip spiders have no tail, and their pedipalps are modified into fierce-looking, spiny raptorial appendages. Amblypygids are easily recognized by their extremely long front legs and raptorial pedipalps. Colors are dull brown or black. The body is flat and leg insertion twisted so that limbs fold in the same plane as the body (like a crab), permitting the animals to edge through thin openings such as cracks in hollow trees. They move sideways more easily than forward or back. Amblypygids hunt by drifting their long front legs gently over the surface around them to locate prey and using their raptorial pedipalps to pounce. Like spiders, all the abdominal ganglia have migrated into the prosoma in which, fused, they form a brain. Reproduction is via a spermatophore. Like uropygids, amblypygids females carry their egg clutches inside a membrane of dried mucus glued to their ventral abdomens. The 15–50 young hatch and remain inside this membrane until they have undergone their first molt. The young cling to the mother for a short time.

FIGURE 6 Whip spider or tailless whip scorpion (Amblypygi: Phrynidae, Phrynus sp.) (photograph by Jonathan A. Coddington). See also color insert, this volume.
Whip spiders live in subtropical and tropical areas, in forests and often in caves. They are exclusively nocturnal and fairly common. During the day they hide in hollow trees or logs, under loose bark, or under large logs. Only one species lives in leaf litter and is not known to burrow. Their diet seems to be a broad range of smaller arthropods in their environment.

B. Uropygi

About 100 species of whip scorpions are known in 16 genera and one family, Thelyphonidae (Fig. 7). At 7.5 cm, the largest is Mastigoproctus giganteus of North America, but 3–5 cm is usual. Whip scorpions are easy to recognize by the long posterior whip or flagellum (highly modified terminal abdominal segments). Colors are brown to black. Uropygids have defensive anal glands that accurately spray an acid-smelling fluid at attackers. The smell explains the common name “vina-garoon.” The fluid of M. giganteus is 85% acetic acid but it also contains substances to reduce the surface tension of the epicuticle so that the acetic acid can spread widely and penetrate. Vinegaroons are supposedly not sensitive to their own spray. Reproductive habits are known only for a few species, but presumably sperm transfer occurs via a spermatophore glued to the substrate in all species. Some species have a lengthy courtship—10 hr to several days. Females keep their 12–40 eggs attached to their ventral abdomen. Female Thelyphonus build a deep burrow and do not feed while guarding the eggs for 4 or 5 weeks. Uropygids may live 6–8 years or even longer.

Uropygids are tropical to subtropical animals. They hide in leaf litter, burrows, under logs and rocks, and inside dark crevices or holes during the day, emerging at night to hunt. Not much is known about their prey, but presumably it consists of other small ground-dwelling insects, arachnids, and crustaceans, which they crush between their pedipalpal segments prior to ingestion.

C. Schizomida

About 200 species in 30 genera and two families (Protoschizomidae and Hubbardiidae) are known (Fig. 8). Schizomids (there is no common name) are most like tiny uropygids, but the abdominal flagellum is short (three segments but rarely five). Colors are usually light yellow to tan to dull green. Assiduous searching in moist tropical litter usually turns up a schizomid. The largest schizomid known is Agastoschizomus lucifer at about 12.7 mm long, but 3 mm is typical. Most schizomids lack eyes entirely and live in moist tropical or subtropical leaf litter, under stones, in logs, moist crevices, caves, and so on. Reproduction is via a spermatophore attached to the substrate, after which females lay 6–30 eggs that are glued to the ventral female abdomen until the young emerge. At two Amazon sites schizomid abundances ranged from 5 to 110 animals per square meter per month.

Like the uropygids they have been reported to produce a defensive chemical smell. They can move backwards rapidly and have enlarged femora apparently used to hop backwards.

D. Palpigradi

There is only one family (Eukoeneniidae) with five or six genera and about 80 species. Palpigrades (micro-
whip scorpions) are tiny, light yellow to white, delicate soil and leaf litter specialists with a cosmopolitan distribution. The largest is *Eukoenenia draco* at 2.8 mm long, but 1 or 2 mm is typical. They resemble the young of Uropygi. Like whip scorpions, they have a wide predorsum and a multi-segmented whip-like postabdomen. Like many soil organisms, palps are most of the organ systems required by large animals that live in drier and less constant environments. They lack eyes, respiratory organs, and a circulatory system but have innervated setae that detect vibrations. Most species are known from the tropics; several palps are apparently live in intertidal or shallow marine habitats. *Eukoenenia janetscheki* was the only species found at two Amazon sites, but it was fairly abundant (30–120 animals per square meter per month).

E. Ricinulei

One family (Ricinoididae), three genera, and 53 species of ricinuleids are known (Fig. 9). Ricinuleids, like schizomids and palps, are soil-dwelling specialists. The largest is *Ricinoides afer* at 10 mm long, but most are 3–5 mm. Unlike the preceding forms, ricinuleids are heavily armored with a thick exoskeleton and a kind of visor or flap (the coccus) that folds over the mouthparts. Colors are reddish to brown or black. Males have third legs modified for sperm transfer. Some have “eye-spots” that are able to sense light. Females lay one or two eggs at a time, which they may carry about with them. Adults may live 10 years. Formerly thought to be extremely rare, ricinuleids are not uncommon in Neotropical leaf litter. They are also known from Africa, but not from the Asian tropics.

F. Pseudoscorpiones

About 3100 species of Pseudoscorpiones (or “false scorpions”) have been described, classified in approximately 400 genera, 24 families, and two suborders (Euproctidae and Icthyosauridae) (Fig. 10). The North American fauna comprises about 350 species. One local fauna in the Amazon comprised nine species and Mediterranean ecosystems around Perth, Australia, and other Old World tropical moist ecosystems are comparable (M. Harvey, personal communication). Pseudoscorpions look like tiny scorpions without the tail and pectines. Colors are tan to dark brown or black. Their chelate pedipalps can have venom glands and the chelicerae have silk glands. The largest is *Gyrocus titanus* at 12 mm long, but most are 2 or 3 mm. They are flattened for life in crevices, usually under bark or in leaf litter, but a few forms live on ocean shores where they are truly intertidal. Numerous species live among the hairs of mammals where they feed upon other arthropod parasites. At least one species has the body, legs, and palps modified for living on a Celebes rat. *Chelifer cancroides* is cosmopolitan and lives in houses where it hunts book lice and other small insects. Pseudoscorpions are sometimes found in aggregations of dozens of individuals. Reproduction, as in most other arachnids, is via a spermatophore. The female builds a silken nest.
and secretes a "brood sac" attached to her body and in which she nourishes the 3–40 embryos with maternal secretions. Juveniles leave the mother soon after hatching. Pseudoscorpions are famous for their tendency to disperse by phoresy. Mature females (rarely males or juveniles) use their pedipalps to hold onto larger insects, which carry them from place to place as they fly about. Phoretic specimens in fossil amber show that this behavior is at least 10 million years old.

G. Solifugae

About 1065 species of solifuges are known, grouped in about 153 genera and 12 families (Fig. 11). The greatest known diversity occurs in Namibia. Wind spiders (sun spiders or camel spiders) look somewhat like hairy, stout, fast-running spiders, but their chelicerae are fearsomely enlarged and both the leg-like pedipalps and the long and slender first legs are used as feelers. The fourth femurs bear ventral sensory structures shaped like inverted mushrooms. Colors are tan to sandy yellow. The largest is Galeodes caspius at 7 cm, but 2 or 3 cm is usual. Wind spiders can run extremely fast (53 cm/sec) for short bursts, but like most arachnids they cannot sustain rapid locomotion for long periods. The solifuge respiratory system is only tracheate and is comparable in complexity to that of insects. They are generally nocturnal and exclusively carnivores and, like any of the larger arachnids, will take small mice, lizards, amphibians, and so on if the opportunity arises. A few species feed only on termites. Unlike most other arachnids, solifuges are active, cursorial predators. Wind spiders are most common in tropical and subtropical dry or desert regions, but they are absent from Australia. Many inhabit burrows, where they may stay up to 9 months of the year, depending on rainfall. Reproduction is via spermatophore, which the male produces during courtship and transfers into the female gonopore with his chelicerae. The female digs a deep burrow to deposit egg masses that may contain 5–164 eggs. Females may lay 1–5 egg masses. Some females remain with the eggs until hatching. Newly hatched wind spiders are gregarious and remain in the burrow for the first two molts.

VI. ACARI

Mites and ticks are classified as Acari, from a Greek word meaning a thing too tiny to be divided (Fig. 12). At more than 45,000 species, Acari is already the most diverse arachnid order, but acarologists estimate that between half a million and more than 1 million species actually exist. Despite being far more abundant and diverse than any other order of arachnids, mites and ticks are nonetheless much less familiar to most people than spiders, harvestmen, or scorpions. Likewise, scientific knowledge of mites lags far behind our understanding of most other arachnid groups. The primary reason for both public and scientific ignorance of mites is their small size.

Since we sometimes find them on ourselves and our pets, because they are relatively large, and because of their role as serious vectors of human disease, the ticks are probably the most familiar group of mites. In fact, ticks (which have their own name only because of their

FIGURE 11 Wind spider or solifuge (Solifuga: Eremobatidae) (photograph by Jonathan A. Coddington). See also color insert, this volume.

FIGURE 12 Free-living yellow moss mite (Acari: Penthalaenidae, Stereocerus sp.) (photomicrograph by David Evans Walter; scale bar = 100 µm).
importance to humans) represent a small offshoot on the evolutionary tree of mites. Other kinds of mites are vastly more common and diverse than the ticks and include not only other parasites but also followers of many other ways of life. In this article, unless otherwise distinguished, the term mites is intended to include ticks and will be used interchangeably with Acanthida.

The Acanthida are distinguished from other arachnids principally by the lack of conspicuous body segmentation, the broad union of the leg-bearing part of the body (the podosoma) with the posterior part (the opisthosoma) to form a single unit (idiomisoma), the aggregation of the mouthparts into a distinct anterior body region (the gnathosoma), and the first active immature stage, the larva, having only three pairs of legs. The gnathosoma is not a true head, like that of an insect, in that it possesses neither eyes nor antennae and does not contain the brain. (Most mites have no eyes at all, but those that do bear them on the idiosoma.) The brain is situated anterior to the stomach in the idiosoma.) Although usually compact in form, like that of a spider mite or tick, the overall shape of the mite body varies greatly. Species that live in polypore fungi or hair follicles are elongated to fit their habitat, and mites found deep in soil tend to be worm-like. Ticks and some other mites can tolerate enormous distention of the body wall to accommodate food.

Like other chelicerate arthropods, mites have no antennae, but in many groups the first pair of legs carry chemosensory and tactile organs and are not used, or little used, for locomotion. The pedipals of integrated laterally into the gnathosoma, are also primarily sensory in function. The chelicerae, the primary food-getting organs, vary greatly in form, in accord with the varied feeding habits of mites. In many species, the chelicerae of males are so modified for sperm transfer that adult males scarcely feed.

A. Ecology

Mites are found in virtually every habitat on Earth. On land, legions of free-living mites populate soils in all habitats, often numbering in the hundreds of thousands per square meter in the soil (50,000–250,000 per square meter, down to 5 cm depth, in forests; 20,000–100,000 in grasslands; and 500–1000 in deserts). In many soils, including deserts, mites can be found at depths up to 10 m, where they follow plant root systems. Although the habits and behavior of many soil mites remain a mystery, it is known that soil mites fill many ecological roles. Many eat decaying organic matter (and the microorganisms it contains); others consume fungi, algae, and bacteria; and still others are predators, feeding on nematodes, other mites, or various stages of insects. Oribatid mites in the soil facilitate plant growth by dispersing the spores of mycorrhizal fungi, which form mutually beneficial and often essential associations with higher plants.

Mites are common and often abundant on plants, on which many species feed on leaf, stem, or root tissue by piercing cell walls. Others specialize on plant reproductive parts, including nectar, pollen, and fruit. Still other plant-dwelling mites are fungivores (eating fungi on leaves or stems) or predators, mostly of plant-feeding mites. Unlike life in the soil, life on plant surfaces constantly threatens mites with desiccation. Hiding from direct sun on the underside of leaves offers some protection because the “boundary” layer of still air is kept relatively moist by plant transpiration.

In addition, many plants produce domatia—hollow, mite-sized cavities or pockets in the leaf. The adaptive function of domatia is controversial, but experiments have shown that removal of domatia from some plants that normally have them reduces the density of beneficial predatory mites and increases the density of herbivorous mites.

Some fungus-feeding mites have specialized cavities or pouches in the body wall (sporotheceae) in which they carry spores or propagules of fungi that they depend on for nourishment.

Parasitic and phoretic mites exploit most terrestrial vertebrate species, most other arthropod groups, mollusks, and even marine invertebrates (Fig. 13). Most species of birds, for example, generally support many “feather” mite species found nowhere else. Four or five different, host-specific species inhabit different feather
types (flight feathers, tail feathers, or body feathers), whereas different species live on the surfaces of feathers, others in the interior lumen of the quills, others on or in the skin, and yet others in the nasal passages. One particular species of parrot hosts more than 20 species of feather mites. Bird nests support additional, phoretic species that feed on debris or on other nest associates. Both parasitic and phoretic species generally colonize young birds while in the nest and travel with the host until it builds its own nest. Similar sets of species inhabit the bodies of mammals and reptiles and, to a lesser degree, amphibians.

Mites are also common associates of insects and other invertebrates. Some are parasitic and deleterious to their hosts, such as the tracheal mite and Varroa mite parasites of honeybees or moth ear mites. Some are commensal (e.g., feeding on insect waste products). Others are believed to benefit the host, perhaps by consuming other harmful mites, parasites, nematodes, or decaying matter that would otherwise nourish fungi or bacteria that might threaten the host or its young. Some insects have acarinaria, special pockets in the outer body wall, that are routinely occupied by phoretic mites. The very existence of acarinaria, which have no other known function other than to house and transport mites is evidence for a net beneficial relationship between these mites and their insect hosts. Parasitic mites have been shown to be capable of mediating horizontal gene transfer between insect host species.

All ticks require blood from a vertebrate host for development and reproduction but live off the host when not feeding or mating. Chiggers and many other mite species must live a portion of their lives, and others all their lives, as parasites of animals. Other mites exploit animal hosts not for food but strictly for transport—a phenomenon called phoresy (see the following section). Parasitism is thought to have evolved, in some lineages, from phoretic associations.

Mites are common and diverse inhabitants of freshwater and marine habitats. In fresh water, adults and nymphs prey on other mites, crustaceans, and aquatic insects. The larvae of aquatic mites are commonly parasitic on insects and crustaceans, and nymphs and adults of unionicol mites parasitize bivalve mollusks. Other aquatic mites graze algae and fungi growing on aquatic plants. Mites are common on seashores, including the intertidal zone. The sea is home to predatory and algae-feeding mites. Some mites have been found at depths up to 7000 m at hydrothermal vents, and one group lives in the gut of sea cucumbers. Mites are the only arachnids found in the ocean depths.

Closer to home, even the cleanest office, classroom, or home normally supports dust mites (e.g., Dermatophagoides spp.), which live on organic matter in house dust. Some mites are stored-product pests, feeding on grains, cheese, and other dry foods. Even more interesting, as you read this encyclopedia there is a very good chance that the follicles and sebaceous glands of your forehead or eyelids house normally benign, microscopic mites (Demodex spp).

B. Reproduction and Dispersal

Reproductive patterns in mites are remarkably diverse. Many mites are ordinary diploids (having an equal number of chromosomes from each parent) with two sexes, requiring fertilization by the male for females to reproduce. In other groups, males are absent, and females bear female young from unfertilized eggs. In still other mites (e.g., spider mites), females require sperm from a male to produce (diploid) daughters, but sons are haploid (they have only the maternal chromosomes) since they are produced from unfertilized eggs. In other species (e.g., some phytoseid mites), both sexes are diploid and all eggs require fertilization to develop, but only the mother's chromosomes are active in males (called maternal genome loss).

Modes of sperm transfer are diverse. Some mites transfer sperm directly through copulation, using legs, palps, or other specialized body parts to accomplish the transfer. Other mites rely on spermatophores placed on the substrate, with or without active enticement of the female to take up the sperm.

Development of the egg takes place in most mites after oviposition. In some species (e.g., moth ear mites and hummingbird flower mites) the eggs are laid communally. In other mites, the young leave the mother's body as active larvae. In some mites (e.g., Siteroptes graminum) sisters and brothers both hatch and mate within the mother's body, and both mother and sons die when the fertilized daughters emerge—truly material for a Greek tragedy.

The development of most mites includes one or more active stages between egg and adult. (Some Heterostigmata have no active immature stages.) In nearly all groups, the first active immature stage, called the larva, has three pairs of legs, and later stages, including the adult, have four pairs. (The plant-parasitic Ergophyoidae are exceptional in having two pairs of legs in all active stages, and females of many species of insect-parasitic Podapolipidae have three, two, or even only one—the anterior—pair of legs.) The number of nymphaal stages between larva and adult varies from none to three, depending on the group. Several groups
of mites have metabolically inactive stages that can endure stressful periods. Life expectancy among mites also varies widely. Some mites live only about 1 week or so (e.g., hummingbird flower mites), whereas others require several years to complete the life cycle (e.g., certain ticks, orbibatid mites, and water mites).

Because mites are so small and (unlike most insects) are wingless, selection for effective dispersal from one food patch to another has produced many special adaptations. Parasitic mites, for example, must have a way to find a new host—or ensure that their offspring do—before the current host dies or becomes unsuitable. In the case of plant-feeding mites, some species disperse aerially, either by simple passive transport in the wind or, in the case of some spider mites (Family Tetranychidae), by “ballooning” on strands of silk. (The silk-producing glands in these mites are not homologous to those of spiders.) Other plant-feeding mites disperse phoretically on plant-visiting animals. For example, nectar- and/or pollen-feeding insects and birds on all continents carry nectar- and/or pollen-feeding mites as hitchhikers—a way of life that has many independent evolutionary origins.

Some groups of mites have special, non-feeding stages adapted for phoretic dispersal. In others, adults and sometimes later nymphal stages mount phoretic hosts to reach new hosts or habitats when local mite population density becomes high, when the host is ready to disperse (e.g., emerging adult insects), or when a food plant becomes unsuitable.

C. Phylogeny, Taxonomy, and Current Knowledge

Only 90 species of Acari were known to Linnaeus, all of which he placed in the genus Acarus. Currently, perhaps 45,000 species in 431 families and 3672 genera have been described compared to 30,000 species estimated 30 years ago. The rate of description of new species is still very high—several hundred species per year—and every systematic acarologist has dozens or hundreds of undescribed species sitting on shelves in vials of alcohol, awaiting description and classification. In fact, the taxonomic impediment to full knowledge of the Acari is monumental:—The number of practicing, modern mite systematists is no more than 50, and the rate of training of new systematists is alarmingly low. Moreover, exploration and species discovery in tropical forests (especially the forest canopy), lakes, and streams is still in its earliest stages. Most estimates of the total world fauna of Acari range from 500,000 to more than 1 million species, but some acarologists even suggest that mites rival insects in worldwide number of species. In any case, mites represent by far the most species-rich and ecologically diverse arachnid group. Table 1 gives summary data on mite diversity.

1. Major Lineages

Given the highly incomplete knowledge of the Acari, it is not surprising that their phylogeny, classification, and nomenclature are far from settled. It is generally recognized, however, that the Acari comprise three lineages that diverged from one another in very ancient times: the Opilioaraciformes and Parasitiformes (which together constitute the Anactinotrichida) and the Acariiformes (which constitute the Actinotrichida). Their relationships to one another and to non-acarine arachnid groups have been debated for decades, but the current opinion appears to be that the Acari, so defined, is a monophyletic group.

The taxonomic rank of acarine lineages is also controversial. The tradition among zoologists in general and arachnologists is to treat Acari as an order, but acarologists now treat it as a subclass or even a class, with subsidiary mite taxa then being orders or superfamilies. Clearly, this is not the place to attempt a resolution of this controversy. The taxonomy in the following sections reflects the prevailing acarological point of view that major groups should rank as orders, despite the incongruity with the rest of this article.

a. Order Opilioaraciformes

The small order Opilioaraciformes (with 17 species described of an estimated 85–170 total) with a single family (Opilioaracidae) are large mites (1.5–2.3 mm in length) that somewhat resemble small harvestmen. Some genera occur in dry climates, often under litter or stones, and others in tropical forest litter, where they feed as omnivores or predators.

b. Order Parasitiformes

The Order Parasitiformes includes three Suborders: Holothyrida, Mesostigmata, and Oxidida.

i. Suborder Holothyrida.

The Holothyrida is a small group (with 32 species described, of an estimated 160–320 total) of large (2–7 mm) saprophagous and predaceous mites of temperate and tropical forest litter, known only from Pacific-Indian Ocean islands and the Australasian and Neotropical regions.

ii. Suborder Mesostigmata.

The cosmopolitan suborder Mesostigmata, in contrast, is extremely rich in species (with 11,632 species described of an estimated
100,000–200,000 total) and diverse in habits, ranging in size from 0.2 to 4.0 mm. Mesostigmatic mites range from saprovores (eating dead or decaying organic matter), fungus feeders, predators living in soil, litter, beach wrack, dung, or rotting wood, to pollen and nectar feeders. Repeatedly, mesostigmatic lineages have evolved close relations with other arachnids (spiders, amblypygids, and scorpions), myriapods, insects, and vertebrates and their nests. Some are endoparasitic (living inside vertebrate respiratory tracts), and others are ectoparasitic. Many others are phototactic, with a variety of feeding habits. Some prey on mite and insect pests of orchards and stored food products. Some of these beneficial predators have been genetically engineered to be more acaricide resistant, enhancing integration of biological and chemical control of pest mite species.

iii. Suborder Ixodida. With 880 species described, of an estimated 1000–1200 total, the ticks (suborder Ixodida) are taxonomically the best-known major group of Acari primarily because of their medical and economic importance. Ticks are considerably larger than most other Acari; adults range in size from 1.7 to 12.7 mm in the unfed state. All ticks feed on the blood of terrestrial vertebrates as ectoparasites. (Sea snakes also have ticks.) Ticks are capable of carrying and transmitting to humans and their domesticated animals more kinds of disease-causing organisms than any other group of blood-feeding arthropods. These agents include viruses, spirochaetes, rickettsiae, anaplasmas, bacteria, piroplasmas, and filariae. Important tick-borne human diseases include Russian spring–summer encephalitis, Colorado tick fever, African relapsing fever, Lyne disease, Rocky Mountain spotted fever, Siberian tick typhus, Q-fever, monocytic and granulocytic ehrlichiosis, Kyasanur forest disease, and tularemia.

c. Order Acariformes

The order Acariformes comprises three suborders—Prostigmata, Astigmata, and Oribatida—with a fourth group, the Endostigmata (120 described species, perhaps 1200–2400 total) of uncertain rank and affinities but often treated as Prostigmata.

i. Suborder Prostigmata. Prostigmatic mites (with 17,000 described species of an estimated 320,000–640,000 total) range in size from 0.1 to 2 mm, with a few of the giant red velvet mites as large as 16 mm, and vary widely in body form, color, habitat, and feeding adaptations. They include predatory and omnivorous species living in organic strata of soils and on algae, lichens, mosses, trees, and shrubs. The Prostigmata also include obligately plant-feeding groups, including the spider mites (Tetranychidae), false spider mites (Tenuipalpidae), and eriophyoid mites, many of which are serious economic pests of field crops, orchards, and greenhouse plants, in some cases acting as vectors of damaging plant viruses. Some groups inhabit the soil; species of the family Nematalyidae have been found at depths of up to 3 m in sand dunes. Species of at least five families of prostigmatic mites are known from Antarctica, and others live in caves. Some prostigmatic groups live in the sea, where they are predaceous or algivorous. Many families of this order are specialized for life in springs, streams, rivers, waterfalls, lakes, bogs, or aquatic interstitial habitats of all descriptions. Some groups live in thermal waters; mites of the genus Thermacarus inhabit hot springs, thriving in water up to 50°C, in which the larvae parasitize amphipods. As in the Mesostigmata, parasitism has arisen repeatedly in independent lineages within the Prostigmata. Some are ectoparasites of slugs, scorpions, insects, or vertebrates; some are parasitoids of eggs or larvae of insects; and some are endoparasites of insect respiratory and reproductive organs, vertebrate respiratory tracts, the quills of birds, the skin of turtles, the mantle cavity of mollusks, the guts of echinoderms, or the gills of decapod crustaceans. A few species in the order are vectors of human disease (scrub typhus), and some cause mange and other skin diseases in domesticated animals. Demodex spp., which live (generally benignly) in hair follicles and sebaceous glands of the human face, are prostigmatic mites, as is the straw itch mite Pyemotes triquetri, which can cause severe skin lesions in humans.

ii. Suborder Oribatida. The Oribatida is remarkable in being the only major group of mites in which a great diversity of species (11,000 described species of an estimated 33,000–110,000 total) has evolved without the evolution of parasitism. Although a few species are phoretic as adults on insects, the rest of this vast radiation are free-living inhabitants of the soil, forest litter, tree holes, bark, twigs, leaves, mosses, lichens, algae, freshwater vegetation, bogs, and the intertidal zone. Most oribatid mites feed on fungi, but some consume dead woody material or algae, and a few are predators on nematodes, rotifers, and other small invertebrates. Oribatida (and many species classified as Astigmata, which are actually derived oribatids) feed on particulate matter. Through their feeding, they help to maintain soil structure. Unknown for other mites, many oribatid species sequester calcium and other minerals in their thickened cuticle. As with Prostigmata, oribatid species are also found on continental Antarctic.
Some oribatids are intermediate hosts of cestode tape-worms whose final hosts are herbivorous mammals. Oribatid mites range in size from 0.2 to 1.5 mm.

Suborder Astigmata. The Astigmata, closely allied to the Oribatida, are primarily associates of arthropods, vertebrates, and other animal groups, although a few species are free-living in all life stages (with 4500 described species of an estimated 90,000–180,000 total). This lineage is thought to be derived from within the Oribatida, but a revised classification, with names of major groupings to reflect this relationship, has yet to be promulgated. A key evolutionary innovation of the Astigmata is the nonfeeding, immature stage especially adapted for phoretic dispersal or tolerance of adverse conditions. Astigmatic mites include fungus feeders, a few plant feeders, a few predators, and mites with mouthparts specialized for filter feeding in water. Some feed on algae in the intertidal zone, in water-filled tree holes, or in sap exudates. Several groups inhabit the nests of insects, mammals, and birds as ectoparasites or commensals. The specialized deutonymphs in several families are parasitic in the hair follicles of mammals. The insect hosts of some species produce acariniaria. One group lives as commensals on the gills of hermit crabs. A large radiation of astigmatic mites lives on feathers and within quills of birds as well as in the avian respiratory tract and air sacs. Another group has radiated among mammals as parasites of the skin, hair, follicles, respiratory passages, ears, and even (rarely) the digestive tract. Some astigmatic mites feed in dung or carrion, and others are important pests of stored food products, including not only grains but also stored meat and dried fish. This order also includes house dust mites (Dermatophagoides spp.), the itch or scabies mite (Sarcoptes scabiei), and mange mites.

Acknowledgments

We gratefully thank Valerie Behan-Pelletier, Clifford Desch, Mark Harvey, Gustavo Hormiga, Adriano Kury, Evert Lindquist, Roy Norton, Barry O'Connor, Lorenzo Prendini, Heather Proctor, Jeffrey Shultz, William Shear, and David Walter for comments on the manuscript, whole or in part. Estimates of mite species numbers are based on Walter and Proctor (1999).

See Also the Following Articles

ARTHROPODS, AMAZONIAN • CRUSTACEANS • INSECTS, OVERVIEW • INVERTEBRATES, FRESHWATER, OVERVIEW • INVERTEBRATES, MARINE, OVERVIEW • INVERTEBRATES, TERRESTRIAL, OVERVIEW • MYRIAPODS

Bibliography


I. Introduction

II. Diversity, Ecology, and Physiology of the Archaea

III. Origin and Evolution of the Archaea

GLOSSARY

acidophiles From the Latin acidus (sour) and the Greek philos (loving). Includes organisms that grow optimally at low pH.

Archaea One of three domains of life. From the Greek archaioi (ancient, primitive). Prokaryotic cells; formerly called Archaeobacteria.

Bacteria One of three domains of life. From the Greek bacterion (staff, rod). Prokaryotic cells; formerly called Eubacteria.

Crenarchaeota One of two kingdoms of organisms of the domain Archaea. From the Greek cren (spring, fountain), for the resemblance of these organisms to the ancestor of the Archaea, and archaios (ancient). Include sulfur-metabolizing, extreme thermophiles.

Eukarya One of three domains of life. From the Greek eu (good, true) and karia (nut; refers to the nucleus). Eukaryotic cells; formerly called Eucaryotes.

Euryarchaeota One of two kingdoms within the domain Archaea. From the Greek eury (broad, wide), for the relatively broad patterns of metabolism of these organisms, and archaios (ancient). Includes halophiles, methanogens, and some anaerobic, sulfur-metabolizing, extreme thermophiles.

halophiles From the Greek halos (salt) and philos (loving). Includes organisms that grow optimally at high salt concentrations.

hyperthermophiles From the Greek hyper (over), therme (heat), and philos (loving). Includes organisms that grow optimally at temperatures higher than 80°C.

Korarchaeota Proposed third kingdom within the domain Archaea. From the Greek kora (young man), for the early divergence of this group during the evolution of the Archaea, and archaios (ancient). Includes a small group of ribosomal RNA sequences retrieved from geothermally heated sediments.

mesophiles From the Greek mesos (middle) and philos (loving). Includes organisms that grow optimally at temperatures between 20 and 50°C.

methanogens Strictly anaerobic Archaea that produce (Greek gen: to produce) methane.

phylogeny The study of the evolutionary relationships among organisms or genes.

ribosomal rRNA Universally distributed molecule among cellular life forms. Widely used to infer the evolutionary relationships among organisms.

thermophiles From the Greek therme (heat) and philos (loving). Includes organisms that grow optimally at temperatures between 50 and 80°C.

Archaea (formerly Archaeobacteria) are a group of microscopic organisms that constitute a phy-
logenetically separate domain of life distinct from the other two domains, the Bacteria (Eubacteria) and Eucarya (Eukaryotes). Archaea are generally, although not exclusively, found living in extreme terrestrial or aquatic environments such as hot springs, deep-sea thermal vents, or under conditions of extreme pH and/or concentrated salinity. Because of their adaptation to unusual environments, modern Archaea may provide insights into the early history of life on Earth.

I. INTRODUCTION

A. Historical Background

For more than 50 years, the notion that there were two basic kinds of living organisms, prokaryotes and eukaryotes, was generally accepted. The placement of a specific organism within one or the other group was based on overall morphological and phenotypic similarities. In 1965, Zuckerkandl and Pauling for the first time suggested that sequences of molecules could be used to reconstruct evolutionary history. Thus, they opened the way to molecular phylogeny, a discipline that studies the evolutionary relationships among organisms or genes. Using molecular phylogeny, in the late 1970s Carl Woese and co-workers found evidence that life consisted not of two but of three distinct groups of organisms: eukaryotes and two kinds of prokaryotes, the Eubacteria and the Archaeobacteria. Their evidence was based on the phylogenetic analysis of a single molecule, the small-subunit ribosomal RNA (SSU rRNA), which is now generally considered an excellent molecule for studying the evolutionary relationships among organisms. The tripartite view of life was formally proposed by Woese et al. in 1990 in the form of three domains: the Eucarya (formerly Eukaryotes), Bacteria (formerly Eubacteria), and Archaea (formerly Archaeobacteria).

Although Archaea resemble Bacteria morphologically, they differ in a variety of cellular and genetic features. From a cellular and biochemical standpoint, Archaea have some unique characteristics, such as the composition of their cell wall and membranes. From a genetic standpoint, Archaea have a unique combination of characteristics once thought to be exclusive to either the Bacteria or Eucarya. Recent analysis of genomic sequences revealed that there is a tendency for archaean genomes to be “grab-bags” or chimeras of both bacterial and eukaryotic sequences.

Since many Archaea thrive under unusual environmental conditions that are lethal to most organisms, such as high temperature, high salinity, or extreme pH, they provide experimental models to study adaptations to extreme environments. From an evolutionary standpoint, the study of modern Archaea may offer valuable insights into the nature of the evolution of biological processes and the origin of life.

B. Taxonomy and Phylogeny

Molecular phylogenetic studies involving the comparison of SSU rRNA sequences revealed that the domain Archaea consists of two kingdoms: the Crenarchaeota and the Euryarchaeota (Fig. 1). Physiologically characterized members of the Crenarchaeota include extremely thermophilic microorganisms belonging to the order Thermoproteales (e.g., Thermoproteus and Pyrobaculum) and to the proposed order Ignicoccales (e.g., Pyrodictium and Desulfurococccus), respectively, and thermoacidophilic microorganisms belonging to the order Sulfolobales (e.g., Sulfolobus and Acidithiobacillus). The Euryarchaeota comprise a metabolically versatile group, which includes all the methanogenic Archaea (orders Methanopyrales, Methanococcales, Methanobacterales, Methanomicrobiales, and Methanosarcinales), the extreme halophiles (order Halobacterales), the extreme thermophiles (orders Thermococcales and Archaeoglobales), and the extremely acidophilic microorganisms belonging to the order Thermoplasmales.

The recent application of molecular approaches to survey microbial diversity in natural environments has provided a new perspective on the distribution and evolutionary relationships of the Archaea. A small group of SSU rRNA sequences retrieved from geothermally heated sediments were found to be different enough from all other archaean rRNA sequences to warrant the proposal of a new kingdom within the Archaea, namely, the Korarchaeota (see Fig. 3). Furthermore, the Crenarchaeota, once thought to be restricted to high-temperature environments, have been found to be ubiquitously distributed, challenging earlier evidence that these microorganisms were confined to high-temperature ecological niches.

II. DIVERSITY, ECOLOGY, AND PHYSIOLOGY OF THE ARCHAEA

In general, the Archaea fall within three general physiological types: (i) thermophiles that live at high tempera-
FIGURE 1  Unrooted phylogenetic tree showing the relationships among representative species of Archaea. Branch lengths (indicated by horizontal distances from nodes) are proportional to the evolutionary distances determined by comparative analyses of SSU rRNA sequences. The scale represents the expected number of changes per sequence position.

... (some thermophiles are adapted to live in extremely low pH conditions and are known as thermoacidophiles); (ii) methanogens that inhabit strictly anaerobic environments and convert carbon dioxide and simple organic molecules to methane; and (iii) halophiles that inhabit highly saline environments and are characterized by a chemoorganotrophic, aerobic metabolism. These three major groups of Archaea are adapted to live in unusual environments and are commonly known as extremophiles. However, molecular surveys have recently revealed that Archaea are much more diverse than previously known, and that they are also widespread in more common biotopes. Non-extremophilic Archaea have been found in marine plankton, terrestrial soils, and freshwater and marine sediments. Because these newly discovered Archaea have resisted cultivation to date, their physiology and metabolic potential are unknown.

A. Thermophilic Archaea

Extremely thermophilic Archaea, or hyperthermophiles, comprise a group of microorganisms which are adapted to grow at temperatures higher than 80°C. Hyperthermophilic Archaea are generally restricted to environments in which geothermal energy is available, such as hot springs, solfataras, geothermally heated marine sediments, and submarine hydrothermal vents.
These environments are rich in sulfur and sulfides, and consequently many of the thermophiles have a sulfur-dependent metabolism. Extremely thermophilic Archaea can carry out a variety of respiratory processes and in most cases elemental sulfur is used as either an electron donor or an electron acceptor. Elemental sulfur is formed from geothermal $\text{H}_2\text{S}$ either by spontaneous oxidation of $\text{H}_2\text{S}$ with $\text{O}_2$ or through the reaction between $\text{H}_2\text{S}$ and $\text{SO}_3$. Submarine volcanic environments include shallow (2–50 m) and deep-sea hydrothermal vents (to depths exceeding 3700 m), where the pressure, even at shallow depths, can raise the boiling point of water sufficiently to select for organisms capable of growth at temperatures higher than 100°C.

Among the Crenarchaeota, *Pyrodictium* and *Pyrolobus* (order Igneococccales) are chemolithotrophic sulfur-dependent hyperthermophiles whose maximum growth temperatures of 110 and 113°C, respectively, represent the upper temperature limits for life known so far. *Pyrodictium* is a strict anaerobe and grows on $\text{H}_2$ and $\text{S}^8$. *Pyrolobus* is unusual in that it is capable of reducing both $\text{NO}_3$ and $\text{S}_2\text{O}_3^{2-}$ to $\text{NH}_4^+$ and $\text{H}_2\text{S}$, respectively, with $\text{H}_2$ as the electron donor. *Desulfurococcus* and *Staphylothermus* (order Igneococccales) are phylogenetically clearly separate from the *Pyrodictium* group (Fig. 1). These coccolid or disc-shaped organisms have an optimal growth temperature higher than 85°C and, in contrast with the *Pyrodictium* group, a maximum temperature not higher than 100°C. They can grow chemolithotrophically by sulfur reduction to $\text{H}_2\text{S}$ or heterotrophically by sulfur respiration of various organic substrates. *Thermoproteus* and *Thermofilum* (order Thermoproteales) are rod-shaped hyperthermophiles that grow in mildly acidic conditions at temperatures up to 95°C. They are both strict anaerobes that can grow chemolithotrophically on $\text{H}_2$ or chemooorganotrophically on complex carbon substrates with $\text{S}^8$ as an electron acceptor. *Pyrobaculum aerophilum* (order Thermoproteales) is a rod-shaped hyperthermophile capable of aerobic respiration in the presence of very low oxygen concentrations (~0.3%) and nitrate reduction under strictly anaerobic conditions.

Sulfataras are terrestrial volcanic environments which abound in sulfur and are gassed by emanating steam carrying $\text{CO}_2$, $\text{H}_2$, and $\text{H}_2\text{S}$, with temperatures up to 100°C. Thermoacidophilic members of the order *Sulfolobales* inhabit acidic solfatara and oxidize $\text{H}_2\text{S}$ or $\text{S}^8$ to $\text{H}_2\text{SO}_4$ using organic carbon or fixing $\text{CO}_2$ as a carbon source. *Sulfolobus* grows at, or close to, the surface of acidic solfatara at temperatures between 75 and 85°C and low pH. No thermoacidophilic, aerobic sulfur oxidizers of the *Sulfolobus* type have been isolated from deep-sea hydrothermal vents. This is related to the steep gradient that occurs between the rising hot, anoxic hydrothermal fluid and the cold, oxygenated ambient seawater, which does not provide an ideal niche for these type of organisms. *Acidianus* (order Sulfurospirales) is a close relative of *Sulfolobus* that grows at an optimum temperature of 85°C and pH 2. *Acidianus* is a more versatile organism than *Sulfolobus* in that it is capable of growth under both aerobic and anaerobic conditions. During aerobic growth, $\text{S}^8$ is oxidized to $\text{H}_2\text{SO}_4$, whereas under anaerobic conditions $\text{H}_2$ is oxidized and $\text{S}^8$ is concomitantly reduced to $\text{H}_2\text{S}$.

Among the Euryarchaeota, *Pyrococcus* and *Thermococcus* (order Thermococcales) are two closely related, coccolid, sulfur-reducing hyperthermophiles that differ primarily in their optimal growth temperatures (100 and 88°C, respectively) (Fig. 2). These organisms are obligate anaerobic chemooorganotrophs that utilize proteins and other complex organic mixtures and reduce $\text{S}^8$ to $\text{H}_2\text{S}$. *Archaeoglobus* (order Archaeoglobales) is an extremely thermophilic, sulfate-respiring organism which has been isolated from shallow and deep-sea hydrothermal environments. *Archaeoglobus* grows at temperatures up to 95°C using $\text{H}_2$, lactate, or complex organic mixtures as electron donors and reducing $\text{SO}_4^{2-}$ and $\text{S}_2\text{O}_3^{2-}$ to $\text{H}_2\text{S}$. *Ferroglobus* (order Archaeoglobales) is a coccolid, hyperthermophilic organism capable of oxidizing $\text{Fe}^{2+}$ at neutral pH under anaerobic conditions. The order Thermoplasmales includes two moderately thermophilic organisms (optimum temperature growth approximately 60°C), *Thermoplasma* and *Pyrococcus*, isolated from coal refuse piles and acidic solfatara. *Thermoplasma* is an acidophilic, cell wall-less Archaeon that resembles bacterial *Mycoplasma* species. It can grow aerobically and anaerobically. Under anaerobic conditions there is a requirement for $\text{S}^8$, which is reduced to $\text{H}_2\text{S}$. *Pyrococcus* is an extremely acidophilic, heterotrophic organism adapted to grow optimally at pH 0.7. Table 1 summarizes some characteristics of the thermophilic Archaea.

**B. Methanogenic Archaea**

Methanogenic Archaea are characterized by their metabolic capability of producing methane. Methanogenesis is a strictly anaerobic respiratory means of metabolism that produces cellular energy in the form of ATP through the reduction of carbon dioxide, formate, or CO (CO$_2$-reducing methanogens), methanol or methylamines (methylo trophic methanogens), or acetate
(acetoclastic methanogens), respectively, to methane. Typical methanogenic reactions are

\[ 4\text{H}_2 + \text{CO}_2 \rightarrow \text{CH}_4 + 2\text{H}_2\text{O} \quad \text{(CO}_2\text{-reducing methanogens)} \\
4\text{CH}_3\text{OH} \rightarrow 3\text{CH}_4 + \text{CO}_2 + 2\text{H}_2\text{O} \quad \text{(Methylo trophic methanogens)} \\
\text{CH}_3\text{COOH} \rightarrow \text{CH}_4 + \text{CO}_2 \quad \text{(Acetoclastic methanogens)}

Most methanogens are capable of autotrophic growth using molecular hydrogen to reduce CO\text{2}\ in a multistep reaction that requires coenzymes unique to this group of organisms, such as methanofuran, methanopterin, F\text{vOH}, F\text{vSH}, and coenzyme M. Methanofuran is a low-molecular-weight phenol derivative involved in the first step of methanogenesis. Methanopterin is a pterin-containing coenzyme which resembles the vitamin folic acid and serves as a C1 carrier during the reduction of CO\text{2}\ to CH\text{4}. Coenzyme F\text{vSH} is a flavin derivative involved in redox reactions. The oxidized form of F\text{vSH} has a characteristic blue-green fluorescence at 420 nm and is very useful in recognizing methanogens microscopically. Coenzymes M and F\text{vOH} play a crucial role in the final step of the methanogenesis as part of the methyl reductase system. F\text{vOH} is a yellow, soluble, nickel-containing tetrapyrrrole that, unlike F\text{vSH}, does not fluoresce.

All methanogenic Archaea are strictly anaerobic and occupy anoxic habitats, such as sediments, marshes, waterlogged soils, submarine hydrothermal vents, and the digestive tracts of animals. Methanogens are usually abundant in environments depleted in Fe\text{3+}, NO\text{3-}, and SO\text{4-2} because these substrates stimulate growth of anaerobic bacteria that can outcompete them. Therefore, methanogenesis is usually limited in sulfate-rich marine environments because the methanogens have to compete with sulfate-reducing bacteria for H\text{2}. Since oxidation of H\text{2} with sulfate as the electron acceptor is thermodynamically more favorable than when CO\text{2}\ is the electron acceptor (as in methanogenesis), sulfate reducers are usually favored. However, since methylo trophic methanogens can use noncompetitive substrates (such as methylamines, which are inaccessible to sulfate-reducers), they can be detected in some sulfate-rich environments. Most methanogens have been found to contain the nitrogenase protein complex; therefore, they are also capable of anaerobic nitrogen fixation under certain growth conditions.

The methanogenic Archaea is a diverse group with many species; it is the largest taxonomic group of Archaea. The order Methanobacteriales includes mainly rod-shaped methanogens that grow by CO\text{2}\ reduction, although a few species of the genus Methanosphaera are cocci that grow only by using H\text{2} to reduce methanol to methane. Methanobrevibacter species are very short rods which use H\text{2} or formate to reduce CO\text{2}\ to CH\text{4}. The genus Methanothermus includes extremely thermophilic methanogens isolated from terrestrial geothermal springs. They can grow at 88\textdegree C on CO\text{2}\ and H\text{2}.
### TABLE I

Habitat and General Characteristics of Thermophilic Archaea

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Order</th>
<th>Genus</th>
<th>Habitat</th>
<th>Morphology</th>
<th>Temperature (°C)</th>
<th>Optimum pH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crenarchaeota</td>
<td>Igneococcales</td>
<td><strong>Pyrodictium</strong></td>
<td>Coastal and deep-sea hydrothermal vents</td>
<td>Disc-shaped</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Pyrolobus</strong></td>
<td>Deep-sea hydrothermal vent, Mid-Atlantic Ridge</td>
<td>Lobed cocci</td>
<td>105</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Desulfurococcus</strong></td>
<td>Hot springs and solfataras</td>
<td>Sphere</td>
<td>85</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Staphylothermus</strong></td>
<td>Submarine hydrothermal vent, Italy</td>
<td>Spheres in clumps</td>
<td>92</td>
<td>98</td>
</tr>
<tr>
<td>Thermoproteae</td>
<td></td>
<td><strong>Thermoproteus</strong></td>
<td>Hot springs and solfataras</td>
<td>Rod</td>
<td>88</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Thermofilum</strong></td>
<td>Solfataras</td>
<td>Rod</td>
<td>88</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Pyrococcus</strong></td>
<td>Submarine hydrothermal vents and solfataras</td>
<td>Rod</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td>Sulfolobales</td>
<td></td>
<td><strong>Sulfolobus</strong></td>
<td>Acidic solfataras</td>
<td>Lobed sphere</td>
<td>75–83</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Acidianus</strong></td>
<td>Acidic solfataras</td>
<td>Sphere</td>
<td>85–90</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Desulfurolobus</strong></td>
<td>Acidic solfataras</td>
<td>Lobed sphere</td>
<td>80</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Stygiolobus</strong></td>
<td>Hot spring, Azores</td>
<td>Lobed sphere</td>
<td>80</td>
<td>89</td>
</tr>
<tr>
<td>Enhydroarchaeota</td>
<td>Thermococcales</td>
<td><strong>Pyrococcus</strong></td>
<td>Coastal and deep-sea hydrothermal vents</td>
<td>Sphere</td>
<td>100</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Thermococcus</strong></td>
<td>Coastal and deep-sea hydrothermal vents</td>
<td>Sphere</td>
<td>88</td>
<td>98</td>
</tr>
<tr>
<td>Archaeoglobales</td>
<td></td>
<td><strong>Archaeoglobus</strong></td>
<td>Coastal and deep-sea hydrothermal vents</td>
<td>Coci</td>
<td>83</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Ferroglobus</strong></td>
<td>Submarine hydrothermal vent, Italy</td>
<td>Coci</td>
<td>85</td>
<td>95</td>
</tr>
<tr>
<td>Thermoplasmata</td>
<td></td>
<td><strong>Thermoplasma</strong></td>
<td>Coal refuse pile, solfataras</td>
<td>Sphere, filaments</td>
<td>59</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Picroplasma</strong></td>
<td>Acidic solfataras, Japan</td>
<td>Coci</td>
<td>60</td>
<td>65</td>
</tr>
</tbody>
</table>

The order Methanomicrobiales encompass a genetically diverse group of CO₂-reducing methanogens that grow at temperatures ranging from 30 to 60°C.

The order Methanosarcinales includes morphologically diverse organisms such as Methanosarcina, which form irregular spheroid bodies occurring alone or in aggregates of cells, and Methanobrix, which forms sheathed rods. Both Methanobrix and Methanosarcina are acetoclastic methanogens, although the latter can also use methanol and methylamines as substrates for methanogenesis.

The order Methanococcales includes mesophilic and extremely thermophilic methanogens characterized by irregular cocoid morphology. Methanococcus jannaschii was isolated from a deep-sea hydrothermal vent system where it grows at 88°C on CO₂ and H₂ released in volcanic gases.

*Methanopyrus* (order Methanopyrales) is an extremely thermophilic methanogen isolated from hydrothermally influenced sediments, and it is capable of autotrophic growth at temperatures up to 110°C. Table II summarizes some characteristics of the methanogenic Archaea.

### C. Halophilic Archaea

The halophilic Archaea occur in environments characterized by high salinity. Salt-neutral lakes, saline soils, solar evaporation marine salterns (areas where sea salt is produced), and subsurface saline deposits are among the most common habitats for halophiles. Most of the halophilic Archaea are red or orange because of the presence of carotenoid pigments in the cell envelope. Frequently, due to their abundance, they impart a red color to the brine. Alkaliphilic halophiles are found in soda lakes, which are highly alkaline environments
### TABLE II

<table>
<thead>
<tr>
<th>Order Genus</th>
<th>Morphology</th>
<th>Substrates</th>
<th>Temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methanobacteriales</td>
<td>Long rods</td>
<td>H$_2$ + CO$_2$, formate</td>
<td>35–40</td>
</tr>
<tr>
<td>Methanobacterium</td>
<td>Short rods</td>
<td>H$_2$ + CO$_2$, formate</td>
<td>30–38</td>
</tr>
<tr>
<td>Methanobrevibacter</td>
<td>Cocci</td>
<td>Methanol + H$_2$ (both needed)</td>
<td>36–40</td>
</tr>
<tr>
<td>Methanosphaera</td>
<td>Rods</td>
<td>H$_2$ + CO$_2$</td>
<td>83–88</td>
</tr>
<tr>
<td>Methanoothermus</td>
<td>Irregular cocci</td>
<td>H$_2$ + CO$_2$, pyruvate + CO$_2$, formate</td>
<td>35–40</td>
</tr>
<tr>
<td>Methanococcales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methanococcus (mesophilic sp.)</td>
<td>Irregular cocci</td>
<td>H$_2$ + CO$_2$</td>
<td>88</td>
</tr>
<tr>
<td>Methanococcus (thermophile sp.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methanomicrobiales</td>
<td>Short rods</td>
<td>H$_2$ + CO$_2$, formate</td>
<td>40</td>
</tr>
<tr>
<td>Methanothermobacter</td>
<td>Irregular cocci</td>
<td>H$_2$ + CO$_2$, formate</td>
<td>30–37</td>
</tr>
<tr>
<td>Methanospirillum</td>
<td>Spirilla</td>
<td>H$_2$ + CO$_2$, formate</td>
<td>30–40</td>
</tr>
<tr>
<td>Methanoplanus</td>
<td>Plate-shaped cells</td>
<td>H$_2$ + CO$_2$, formate</td>
<td>32–40</td>
</tr>
<tr>
<td>Methanoculleus</td>
<td>Coccus</td>
<td>H$_2$ + CO$_2$, formate</td>
<td>37–60</td>
</tr>
<tr>
<td>Methanosarcinales</td>
<td>Large irregular cocci in packets</td>
<td>H$_2$ + CO$_2$, methanol, methylamines, acetate</td>
<td>35–50</td>
</tr>
<tr>
<td>Methanolobus</td>
<td>Irregular cocci in aggregates</td>
<td>Methanol, methylamines</td>
<td>30–40</td>
</tr>
<tr>
<td>Methanocylindricus</td>
<td>Irregular cocci</td>
<td>Methanol, methylamines</td>
<td>50</td>
</tr>
<tr>
<td>Methanococcales</td>
<td>Irregular cocci</td>
<td>Methanol, methylamines</td>
<td>23–35</td>
</tr>
<tr>
<td>Methanospirillum</td>
<td></td>
<td>Methanol, methylamines, methyl sulfides</td>
<td>26–36</td>
</tr>
<tr>
<td>Methanobrevibacter</td>
<td>Long rods to filaments</td>
<td>Acetate</td>
<td>35–60</td>
</tr>
<tr>
<td>Methanopyrus</td>
<td>Rods in chains</td>
<td>H$_2$ + CO$_2$</td>
<td></td>
</tr>
</tbody>
</table>

whose high pH (8 to >12) is due to high levels of carbonate. Halophiles require at least 1.5 M NaCl for growth, and optimal salt concentrations are usually in the range of 2–4 M NaCl.

Neutrophilic and alkaliophilic halophiles both belong to the order Halobacteriales. The members of the neutrophilic group, represented by the genera Halobacterium, Halococcus, Haloarcula, and Haloferax, grow optimally under conditions of high magnesium and sodium concentrations (0.5 and 4.0 M, respectively) at pH ranging from 5 to 8. Alkaliphile halophiles (genera Natronobacterium and Natronococcus) grow optimally at low magnesium concentrations and pH of approximately 10.

Halophilic Archaea have evolved several physiological adaptations that permit their growth in habitats with salt concentrations that cause cellular dehydration and protein denaturation in other organisms. Halophilic Archaea such as Halobacterium resist high salt concentrations by pumping large amounts of K⁺ from the environment into the cell such that the concentration of K⁺ inside the cell is greater than the concentration of Na⁺ outside the cell. This mechanism allows Halobacterium to remain in positive water balance and avoid dehydration.

The halophilic Archaea are aerobic and grow heterotrophically using carbohydrates, alcohols, organic acids, and amino acids. Halobacterium species are normally aerobic but can grow anaerobically in the presence of light. Under oxygen-limiting conditions and in the presence of light, Halobacterium inserts large amounts of a protein called bacteriorhodopsin into the cytoplasmic membrane. This purple pigment, which adsorbs light strongly at approximately 570 nm, acts as a light-driven proton pump and leads to the establishment of an electrochemical membrane potential. Its equilibration can be accompanied by the generation of ATP. Thus, bacteriorhodopsin facilitates a special kind of photophosphorylation. The energy obtained by this mechanism complements that obtained from aerobic substrate oxidation. Table III summarizes some characteristics of the halophilic Archaea.


### TABLE III

**Habitat and General Characteristics of Halophilic Archaea**

<table>
<thead>
<tr>
<th>Order</th>
<th>Genus</th>
<th>Species</th>
<th>Habitat</th>
<th>Morphology</th>
<th>Physiological characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Halobacteriales</td>
<td>Halobacterium</td>
<td><em>H. salinarum</em></td>
<td>Salted fish, hypersaline lakes</td>
<td>Rod-shaped</td>
<td>Amino acids, phototrophic</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>H. mediterranei</em></td>
<td>Salters</td>
<td>Rod/disk-shaped</td>
<td>Carbohydrates</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>H. halobacterium</em></td>
<td>Salters</td>
<td>Rod-shaped</td>
<td>Nitrate respiration</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>H. lacusprofundi</em></td>
<td>Deep lake, Antarctica</td>
<td>Rod-shaped</td>
<td>Carbohydrates</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>H. sodomense</em></td>
<td>Dead Sea</td>
<td>Ploemorphic</td>
<td>Photosphoric</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>H. marismortui</em></td>
<td>Death Valley</td>
<td>Carbohydrates</td>
<td>Nitrate respiration</td>
</tr>
<tr>
<td>Halococcus</td>
<td></td>
<td><em>H. marinae</em></td>
<td>Dead Sea</td>
<td>Cocci</td>
<td>Amino acids, Nitrate respiration</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>H. salinarum</em></td>
<td>Salters</td>
<td>Carbohydrates</td>
<td></td>
</tr>
<tr>
<td>Natrialba</td>
<td></td>
<td><em>N. gregoryi</em></td>
<td>Soda lakes</td>
<td>Rod-shaped</td>
<td>Organic acids, pH 9.5–10</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>N. occultus</em></td>
<td>Soda lakes</td>
<td>Cocci</td>
<td>Amino acids, pH 9.5–10</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>N. asiatica</em></td>
<td>Sea sand</td>
<td>Rod-shaped</td>
<td>Nonpigmented, extreme halophile</td>
</tr>
</tbody>
</table>

---

**D. Nonextremophilic Archaea**

The application of SSU rRNA-based molecular techniques to survey natural microbial populations has recently revealed that Archaea are virtually ubiquitous. Therefore, our view of the Archaea as highly specialized microorganisms adapted to survive in extreme environments is gradually changing to that of a very diverse group of organisms, including more moderate representatives.

Non-extremophilic Archaea have been identified in a variety of environments, such as temperate and cold marine planktonic habitats, freshwater sediments, soils, terrestrial subsurface environments, and permanently cold deep-sea sediments, and in association with several marine metazoans. To date, none of the non-extremophilic Archaea have been isolated in pure culture; therefore, their phenotypic and physiological characteristics remain unknown. However, their ubiquitous distribution and high phylogenetic diversity suggests a wide range of ecological and physiological adaptations.

---

The distribution and phylogenetic affinity of the non-extremophilic Archaea are summarized in Table IV.

---

**III. ORIGIN AND EVOLUTION OF THE ARCHAEA**

Archaea occupy a pivotal phylogenetic position between the two other domains of life, Bacteria and Eukarya (Fig. 3). With the recent sequencing of the entire genomes of several Archaea, it has become evident that archaerial genomes share bacterial and eukaryotic features. Furthermore, the availability of an increasing number of gene sequences from organisms belonging to the three domains of life allows the identification of a "universal" set of protein families present in all three domains. This universal set of proteins can then be used as an estimate of the genome content and the cellular processes that were present in the last universal
TABLE IV
Distribution and Phylogenetic Affiliation of Nonextremophilic Archaea

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Phylogenetic affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marine Habitats</td>
<td></td>
</tr>
<tr>
<td>Surface and deep waters (up to 3000 m)</td>
<td>Crenarchaeota, Euryarchaeota</td>
</tr>
<tr>
<td>Temperate coastal sediments (12 m)</td>
<td>Crenarchaeota, Euryarchaeota</td>
</tr>
<tr>
<td>Low-temperature deep-sea sediments (1500 to 4500 m)</td>
<td>Crenarchaeota, Euryarchaeota</td>
</tr>
<tr>
<td>Temperate microbial mats at deep-sea hydrothermal vents</td>
<td>Crenarchaeota, Euryarchaeota</td>
</tr>
<tr>
<td>Antarctic low-temperature surface waters</td>
<td></td>
</tr>
<tr>
<td>Salt marsh</td>
<td>Euryarchaeota</td>
</tr>
<tr>
<td>Associated with Marine Metazoans</td>
<td></td>
</tr>
<tr>
<td>Gut of abyssal holothurian Oenocochlora mutabilis (4870 m)</td>
<td>Crenarchaeota</td>
</tr>
<tr>
<td>Digestive tract of fish</td>
<td></td>
</tr>
<tr>
<td>Tissues of sponge Axinella mediterranea (10–20 m)</td>
<td>Crenarchaeon Crenarchaeum symbionum</td>
</tr>
<tr>
<td>Freshwater Habitats</td>
<td></td>
</tr>
<tr>
<td>Lake Sediments</td>
<td>Crenarchaeota, Euryarchaeota</td>
</tr>
<tr>
<td>Terrestrial Habitats</td>
<td></td>
</tr>
<tr>
<td>Soils</td>
<td>Crenarchaeota, Euryarchaeota</td>
</tr>
<tr>
<td>Subsurface paleosol (188 m)</td>
<td>Crenarchaeota</td>
</tr>
<tr>
<td>Contaminated aquifer</td>
<td>Crenarchaeota, Euryarchaeota</td>
</tr>
<tr>
<td>Rice roots</td>
<td>Crenarchaeota, Euryarchaeota</td>
</tr>
</tbody>
</table>

Cytocrome oxidase, which catalyzes the reduction of oxygen to water and acts as a redox-linked proton pump, is the key enzyme of aerobic metabolism. This enzyme is present in members of both the Crenarchaeota (Sulfobus, Acidanuvus, and Pyrobaculum) and the Euryarchaeota (Halobacterium and Natronobacterium) as well as in many thermophile Bacteria (e.g., Thermus and Aquifex). It is plausible that in the early atmosphere, prior to the advent of oxygenic photosynthesis, microaerophilic organisms similar to the present-day archaeon Pyrobaculum may have been able to thrive under extremely low oxygen concentrations, known to have existed as a result of the photolysis of water.

The ability to use sulfate as a terminal electron acceptor is a characteristic common to many mesophiles and thermophilic bacterial species and to the thermophilic archaeon Archaeoglobus. In addition, a sulfite reductase-type protein has been identified in Pyrobaculum. The enzyme dissipates the sulfite reductase catalyzes the six-electron reduction of (l)glutathione to sulfide, which is the central energy-conserving step of sulfate respiration. Phylogenetic evidence suggests an ancestral origin of sulfate respiration, a finding which is consistent with the notion that sulfates of magmatic origin were common in the Archaean time. In addition to sulfate reduction, Archaeoglobus produces very small quantities of methane. Factor F430 and methanopterin, two coenzymes involved in methanogenesis, are also present in Archaeoglobus, although this organism appears to lack other cofactors normally present in methanogens. This apparently intermediate type of metabolism between methanogenesis and sulfur reduction suggested that Archaeoglobus may represent a transitional form in the diversification of Archaea from sulfur-respiring to methane-producing and halophilic Archaea. The branching position of Archaeoglobus in the SSU rRNA tree is consistent with this interpretation (Fig. 1).

Sulfur respiration has also been proposed to be an ancestral mode of energetic metabolism on the base of two considerations: (i) Volcanic-derived S0 was probably one of the most abundant electron acceptors in the early atmosphere, and (ii) the capacity to reduce S0 to H2S is common among anaerobic, hyperthermophilic members of both Archaea and Bacteria.

A. Early Respiratory Processes
The presence of genes encoding for homologous proteins in organisms belonging to both domains Archaea and Bacteria allows for the reconstruction of cellular processes that were present early in the history of life on Earth. In particular, terminal oxidases belonging to oxygen, nitrate, sulfate, and sulfur respiratory pathways are present in both domains.

B. Adaptive Features to High Temperature
Archaea exhibit a strong adaptive capacity to extremely high temperatures. Since microorganisms are isothermal, they have to develop strategies to avoid thermal
stress. Thus, in order to survive at temperatures in excess of 100°C, hyperthermophiles have to adapt their cell inventory to function optimally at the temperatures in the niches they occupy. Many different mechanisms have been hypothesized by which hyperthermophiles can simultaneously retain stability and plasticity at high temperatures.

One important mechanism of thermoadaptation is the biosynthesis of thermostable proteins and enzymes that maintain sufficient structural integrity to allow optimal catalytic efficiency at high temperatures. Thermostable proteins are often characterized by a very efficient packing density in the hydrophobic core of the molecule as well as by shorter connecting elements (loops) between regions of secondary structure. An elevated number of ionic interactions has frequently been found in highly thermostable proteins. Ionic interactions among monomers are thought to stabilize multimeric proteins at high temperatures. Overall, these characteristics lead to decreased flexibility in the polypeptide chain, a required feature that compensates for increased thermal fluctuations at high temperatures.

Most living beings need the strand-opening potential of negative DNA supercoiling to allow transcription and other DNA-dependent processes. In bacterial nucleoproteins and in eukaryal chromosomes, DNA has a negative superhelicity as a result of topoisomerase activity and wrapping of DNA around histone cores, respectively. Histone-like proteins may play an important role in certain thermophilic Archaea, protecting DNA against thermal denaturation and degradation. The DNA-binding proteins isolated from both *Thermoplasma acidophilum* and *Methanothermus fervidus* stabilize double-stranded DNA molecules *in vitro* by increasing their melting temperatures. A novel class of topoisomerases has been identified in both hyperthermophilic Archaea (e.g., *Sulfolobus*) and Bacteria (e.g., *Thermotoga*). Since these enzymes catalyze positive supercoiling into the DNA, they have been referred to as reverse gyrase. Reverse-gyrase activity is widely distributed in hyperthermophiles and appears to be, at least at critical regions, a requirement to prevent DNA denaturation at high temperatures. Thus, the reverse gyrase may be considered as an ancient trait retained by present-day hyperthermophilic Archaea.

C. Evolution

Hyperthermophilic Archaea have been proposed as analogs for the early life on Earth. If life arose and evolved under high-temperature, reduced, sulfur-rich conditions, then the requirement of thermophilic Archaea for high temperature environments as well as their predominantly anaerobic, sulfur-metabolizing phenotype
suggest they are good analogs to test such a theory. Two independent lines of evidence suggest that the phenotype of the extant hyperthermophiles may represent remnants of the ancestral Archaea. First, the hyperthermophilic trait is the only one found in both archaean kingdoms, the Crenarchaeota and the Euryarchaeota. Second, most hyperthermophilic lineages are more slowly evolving than any other archaean lineage. This is particularly evident when the rate of evolution of the SSU rRNA genes (indicated as the branch length in the phylogenetic tree) is considered (Fig. 1). Mesophilic methanogens, halophiles, and the recently discovered non-extremophilic Archaea appear to evolve much faster than hyperthermophiles (Fig. 1). Furthermore, mesophilic methanogens, such as members of the Methanococcales, evolve more rapidly than their extremely thermophilic relatives. Therefore, the assumption that more slowly evolving lineages (thermophilic) tend to retain more of the ancestral traits than do rapidly evolving lineages (non-thermophilic) suggests a high-temperature ancestry of the Archaea, which later adapted to lower temperatures.

A different theory supporting a non-thermophilic ancestor to extant life forms has been recently formulated on the base of the strong correlation between the G + C content of rRNA genes and the optimal growth temperature of microorganisms. The analysis of rRNA genes from hyperthermophiles suggests that survival in high-temperature conditions requires a high G + C content in rRNA genes. However, the estimated G + C contents of ancestral organisms appeared to be incompatible with life at high temperature and suggested a moderate environmental growth temperature for the universal ancestor. This theory leads to the conclusion that present-day thermophily evolved from mesophilic organisms via adaptation to high temperature.

In principle, the nature of an ancestral organism may be inferred from the distribution of homologous traits among its descendants. However, the reconstruction of early evolutionary histories is complicated by phenomena such as horizontal gene transfer (gene swapping) among organisms. As new microbial genomes become available, it has become increasingly evident that extensive gene swapping, even among distantly related organisms, may have occurred. This tends to blur the history of the early stages of life on Earth; therefore, the nature of the last universal ancestor remains elusive. At the same time, comparative functional genomics is providing a new perspective of the characteristics of the universal ancestor. Because of their evolutionary history, Archaea are likely to play a critical role in our understanding of the early events in the evolution of cells and the origin of life.

See Also the Following Articles

BACTERIAL BIODIVERSITY • BACTERIAL GENETICS • EUKARYOTES, ORIGIN OF • MICROBIAL BIODIVERSITY • ORIGIN OF LIFE, THEORIES OF • PSYCHROPHILES • THERMOPHILES, ORIGIN OF

Bibliography


ARCTIC ECOSYSTEMS

Terry V. Callaghan,* Nadya Matveyeva,† Yuri Chernov,‡ and Rob Brooker¶

*University of Sheffield, †Abisko Scientific Research Station, ‡Komarov Botanical Institute, 
¶Severtzov Institute of Evolution, and Institute of Terrestrial Ecology

I. Introduction
II. Definitions and General Environmental Characteristics of the Arctic
III. Patterns of Biodiversity in the Arctic
IV. Controls of Biodiversity
V. Consequences of Biodiversity
VI. Threats to Biodiversity in the Arctic
VII. Conclusion

GLOSSARY

Arctic Geographically, the region lying to the north of latitude 66.7°N, but environmentally, and in the context of this article, the region to the north of the climatically controlled northern latitudinal treeline which corresponds approximately to the mean July isotherm of 10°C.

permafrost The phenomenon of water which is permanently at or below 0°C. Usually, but not always, the water is in the solid state. The Arctic is characterized by the presence of large, continuous areas of permafrost that have the form of lower soil layers that are permanently frozen with a shallow (usually <1 m) "active layer" which freezes and thaws each year and accommodates belowground biological activity.

refugia In the context of the Arctic, these are land areas which were not covered by ice sheets or glaciers during the last glaciation. Consequently, some biota could survive there and recolonize adjacent areas when the ice retreated. Therefore, these areas are associated with high biodiversity and endemism. Refugia usually occurred in coastal areas (now continental shelves) and ice-free mountain tops or "nunataks.”

subarctic The ecotone (ecological boundary zone) connecting the treeless tundras in the north with the taiga or coniferous boreal forests in the south. The area is characterized by the presence of scattered, deciduous or coniferous trees of low stature and is sometimes termed "forest tundra." (Other Arctic vegetation zones are described in the text.)

treeline Here, denotes the northern latitudinal and altitudinal distribution limit of vegetation in which trees are a conspicuous but not necessarily a dominant element. Individual, isolated trees or "oases" of trees in environmentally benign areas are considered to be beyond the treeline.

tundra This is a type of vegetation characteristically occupying the Arctic. However, the term is used in many ways, from characterizing individual plant communities of the Arctic which consist of scattered shrubs and sedges to characterizing all vegetation above the altitudinal treeline and between the latitudinal treelines and the poles in both hemispheres. In this article, the term is used in the Russian sense to characterize Arctic vegetation lying between the taiga and the region of the polar deserts.

THE ARCTIC REGION is one of the world's last great wilderness areas where human impacts on terrestrial
ecosystems are relatively light. Although species diversity is generally lower than at more southerly latitudes, the diversity of animals and plants, communities, and landforms are surprisingly rich. Patterns of biodiversity are strongly coupled with the wide variety of Arctic environments because of relatively low interference by man. The Arctic is therefore far from uniform. This article introduces the reader to current patterns of biodiversity in terrestrial Arctic ecosystems and discusses aspects of the physical, historical, and also biotic environments that have shaped these patterns. We then seek to highlight the importance of the Arctic’s biodiversity for ecosystem function and provision of resources for human welfare. Finally, we outline the numerous threats to the Arctic’s biota and the major challenges to our further understanding of its responses to change.

I. INTRODUCTION

Although Arctic environments have undergone dramatic changes over millions of years, the twentieth century has been associated with particularly rapid changes in many aspects of the physical environment and particularly with rapid cultural and sociological changes in the north leading to the increasing exploitation and fragmentation of wilderness areas. Because future environmental changes are predicted to be even greater during this century, it is becoming increasingly necessary to document, monitor, and understand the biological resources of the Arctic. Understanding the patterns, causes, and consequences of biodiversity as highlighted by Chapin and Körner (1995) throughout the Arctic’s lands is an important aspect of this challenge. However, the Arctic should not be seen as a region remote from the more populated regions of the world at lower latitudes: Several hundred million birds of many species overwintering in temperate regions migrate to summer nesting grounds in the Arctic, and the functions of many Arctic ecosystems in sequestering the greenhouse gas carbon dioxide in soils has contributed to cooling the earth’s surface since the end of the last ice age.

This article introduces the reader to current patterns of biodiversity in terrestrial Arctic ecosystems and presents further details of the physical, historical, and biotic environments that have shaped these patterns. We then show the importance of the Arctic’s biodiversity in terms of ecosystem function and its provision of resources for human life support and welfare. Finally, we show how the Arctic’s biota are threatened by numerous factors and discuss the major challenges to further understanding the biota.

II. DEFINITIONS AND GENERAL ENVIRONMENTAL CHARACTERISTICS OF THE ARCTIC

The Arctic is difficult to define, basically because of two confounding factors: the sensitivity of flora and fauna to low temperatures by the formation of mists. On land, topography also affects temperatures locally. As one moves closer to the North Pole, progressively smaller differences in topography become more important in determining temperature. Often, therefore, the definition of the Arctic depends on the subject and the scientific discipline. Because all biological processes depend on chemophysical reactions, temperature is an important determinant of biological activity. Also, because nearly all food webs ultimately depend on primary producers which require light for photosynthesis, the strong seasonal variations in photoperiod are also important for biological activity. For biologists, therefore, definitions that incorporate the sensitivity of flora and fauna to low temperatures and specific photoperiods are particularly appropriate. Thus, the 10°C summer isotherm (where the mean monthly temperature is at or below 10°C) and the northern limit of forests (the latitudinal treeline), which both represent the southern boundary of the “low Arctic” in Western terminology and the “tundra region” in...
Figure 1. Contrasting concepts of vegetation zones of the Arctic according to Bliss and Matveyeva (1992) and Alexsandrová (1980). The low and high Arctic are commonly used in the West, whereas the geobotanical zonation is used in Russia.

Russian terminology, are commonly used to define the Arctic (Fig. 1).

The extent of the Arctic is totally dependent on its definition. Using botanical definitions for the Arctic, Bliss and Matveyeva (1992) calculate that the Arctic comprises approximately 7.5 million km², which is approximately 5.5% of the earth’s land surface. The Arctic stretches over 33 degrees of latitude from 84°N in Greenland to 51°N in Hudson Bay, Canada (see Jonasson et al. in Nuttall and Callaghan, 2000). Almost 2 million km² is permanent snow and ice. Distances around the Arctic decrease dramatically towards the North Pole. For example, the circumference of the equator is 40,076 km, whereas that of the 70° parallel is only 13,752 km. This has great implications for connectivity between animal and plant populations and species.

The complexity of defining the Arctic is amplified when the Arctic is subdivided into different zones. Al-
though it is clear that the Arctic is not a homogeneous environment, definitions of subzones conflict between Eastern and Western scientific traditions (Iliess and Matveyeva, 1992). In the West, a generally accepted classification system is as follows:

1. Taiga in the south, consisting of predominantly taiga forest, and forest tundra (sparse trees, often with low stature) at its northern edge, near the 10–12°C isotherm for July
2. The low Arctic consisting of tundra vegetation (low thicket-forming shrubs—sedge, tussock—dwarf shrub, and mire communities)
3. The high Arctic consisting of polar semidesert vegetation in the south (cryptogam—herb, cushion plant—cryptogam, and mire communities which do not cover all of the ground) and polar desert (herb-cryptogam communities which cover only approximately 5% of the ground) in the far north where mean July temperatures vary from 6°C in the south to only approximately 2°C in the north. Precipitation in the north is less than 50 mm per year and falls mainly as snow.

In a generally accepted Russian classification, the taiga zone remains the same, but northwards a broad tundra zone classification replaces the low Arctic and polar semidesert subzones of the Western classification. The Russian classification is as follows:

1. The taiga as before
2. The tundra zone, divided into the southern tundra subzone with low shrub—sedge, tussock—dwarf shrub, and mire communities; a typical tundra subzone with sedge—dwarf shrub and ploygonal mire communities; and an Arctic tundra subzone in the north consisting of dwarf shrub—herb communities
3. The polar desert zone, characterized by cryptogam—herb communities

Despite these classifications, the “subarctic” region is often referred to and represents the ecotone between the tundra and taiga and the forest tundra. In fact, there is a continuous gradient of environmental severity within the Arctic from the boreal forest zone at its southern boundary to the polar deserts of the far north, even if this is interrupted in some places by mountain chains and water bodies. There are great temperature variations throughout the Arctic. From the southern boundary where the mean July temperature is 10°C and there are more than 1000 degree days >0°C, there is a decrease to the high Arctic where the mean July temperature can be less than 2°C with just over 110 degree days per year. Precipitation also decreases toward the north, from over 1000 mm per year to approximately 50 mm per year. Low temperatures and low precipitation exert critical limitations on the productivity of Arctic ecosystems (Jonasson et al. in Nutall and Callaghan, 2000). Net primary production varies from approximately 1 g m⁻² year⁻¹ in polar deserts to 150–800 g m⁻² year⁻¹ in the Subarctic. In the low Arctic, values can range between 100 and 1200. Although there is a good correlation between temperature and productivity, the mechanisms are complex. In many areas, it has been suggested that the most important aspect of temperature is indirect in that it constrains the rate of nutrient cycling in cold arctic soils, thereby resulting in nutrient limitation throughout most of the Arctic (Jonasson et al. in Nutall and Callaghan, 2000).

The great spatial differences in temperature regimes throughout the current Arctic are accompanied by even greater variations over time. Arctic landmasses and oceans are relatively young and have been formed by the migrations of landmasses from southerly latitudes. Also, much of the Arctic’s land and oceans have been covered even in summer by ice, but only during the past 1.8 million years. The last glacial period, which ended approximately 10,000 years ago in much of the Arctic, displaced terrestrial, freshwater, and marine species southwards from much of the Arctic. A similar process occurred over large lowland areas of the Russian Arctic when they were entirely inundated by the marine transgression (Aleksandrova, 1980; Yurtsev, 1997).

The Arctic’s current biota can be seen, therefore, as relatively young and, on land, a depauperate remnant of previous floras and faunas that existed in preglacial eras. For example, forests covered areas of Greenland, Svalbard, and the Canadian high Arctic Archipelago during the Tertiary Period, ending approximately 3 million years ago (this signified a warmer Arctic but also an Arctic landmass that was a few degrees latitude further south). The megafauna, such as the mammoth, of the Russian Arctic and extensive tundra steppe communities disappeared during the early Holocene.

Although much of the current Arctic floras immigrated, and the Arctic during the Holocene as the ice retreated, some Arctic areas, “refugia” and “nunataks,” remained ice free and supported biota for periods long enough to facilitate the establishment of endemic taxa. The extent to which species colonized the Arctic during the Holocene from such refugia, or from immigration from the south, is unclear. One example of a species thought to have survived in a refugium on Svalbard and then spread out is Pedicularis dasynatha (Odasz as quoted by McGraw in Chapin and Korner, 1993).

During the recolonization of the Arctic, the continu-
ity of landmasses allowed a movement of biota into and out of and around the Arctic, e.g., across the Bering land bridge. This connectivity together with the decreasing longitudinal distances toward the pole allowed, and continues to allow, the mobility of many organisms such as marine mammals, migrating caribou, and particularly migratory birds. Such mobility adds considerable complexity to understanding biodiversity in the Arctic because some of the causal factors are operative when migratory organisms are in other latitudes.

Despite this general connectivity, barriers to dispersal of species occur throughout the Arctic. In Alaska, the Brooks mountain range runs approximately east–west and separates the taiga forest zone to the south from the tundra zone in the north. In the Canadian Arctic, much of the Northwest Territories is an archipelago and in Greenland, particularly along the east coast, glaciers flowing to the sea interrupt the continuity of ice-free coastal lands. In Fennoscandia, the mainland is represented mainly by subarctic areas in the mountains, and high Arctic environments are found only on the Svalbard archipelago. In the Russian Arctic, the Ural mountains running north–south separate the vast landmass longitudinally. However, on the Taimyr Peninsula, there is probably the greatest latitudinal continuity of land reaching into the Arctic.

III. PATTERNS OF BIODIVERSITY IN THE ARCTIC

It is not possible to state the exact number of species in the Arctic because of a lack of taxonomic knowledge (particularly for the Protozoa, Nematoda, terrestrial Oligochaeta, many Acari taxa, and insect families in the orders Diptera, Hymenoptera, and Lepidoptera) and incompatible synonyms used in different countries. However, some estimates of general species number can be made (Matveyeva and Chernov in Nutall and Callaghan, 2000; Table I).

The number of species in the majority of plant and animal classes that are of importance in the higher Arctic environments amounts, on average, to approximately 2.5% of the worldwide species number for a given taxon (Table I). The majority of the higher terrestrial faunal taxa are represented in the Arctic, although on a species number basis the Arctic contains less than 1% of the world’s fauna. In comparison, the angiosperms, the most advanced plant taxon, have fewer species in the Arctic than the cryptogams (Table I). Thus, the Arctic’s biota is not just an extremely depauperate replication of the world’s biota but also has a very distinctive structure as exemplified by the proportions of the highest taxa of plants and animals.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Group</th>
<th>Number</th>
<th>% of world biota</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammals</td>
<td>55</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Birds</td>
<td>200</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>Insects</td>
<td>3000</td>
<td>0.3–0.4</td>
<td></td>
</tr>
<tr>
<td>Diptera</td>
<td>1600–1800</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Bees</td>
<td>350</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Springtails</td>
<td>400–300</td>
<td>7.0–8.0</td>
<td></td>
</tr>
<tr>
<td>Spiders</td>
<td>300</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Mites</td>
<td>700</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Other groups (^a)</td>
<td>600</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Total estimate</td>
<td>6800–7000</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Modified from Matveyeva and Chernov in Nutall and Callaghan (2000).

\(^b\) Amphibians (4), Centipedes (10), Mollusks (12–24), Worms (70), and Nematodes (~300).
Classes that are well represented in, and are therefore particularly well adapted to, Arctic environments constitute approximately 1–3% or more of global diversity. These include birds, springtails, mosses, and lichens. The largest classes of global biota (e.g., insects and flowering plants) have the lowest level of representation in the Arctic at 0.3–0.4 and 0.8%, respectively. In contrast, relatively small groups such as springtails (Collembola) and horsetails (Equisetales) include 7, 8, and 23%, respectively. At lower taxonomic levels—orders, families, and genera—some taxa are endemic to the Arctic. For instance, all species of the order of the diving birds Gaviiformes are found in the Arctic.

Although the proportion of various taxa that are found in the Arctic may be very small, there is a common misunderstanding that this equates to low species numbers in Arctic communities (Chapin and Körner, 1993). This is not necessarily the case. For example, the wide ecological distribution of many Arctic plants, together with their small stature, results in high numbers of species per unit area in certain communities. Thus, on Taymyr, Matveyeva (Matveyeva and Chernov as cited in Nutall and Callaghan, 2000) recorded from 110 to 182 species per sample plot (100 m²). There were 40–50 species within 1 m² and up to 25 species within 1 dm². Even in the polar desert, approximately 110 to 182 species per sample plot (100 m²) were recorded from approximately 250 in the south to approximately 2 in the north. Analogous decreases in plant diversity from the polar deserts to the taiga boundary decrease from 12°C at the taiga and the taiga and polar desert 2°C, whereas in the boreal forest belt a comparable 10°C. Changes in mean July temperature decreases from 12°C to 2°C, whereas in the boreal forest belt a comparable 10°C change in mean July temperature occurs over almost 2000 km, a range across which there are three natural life zones. The gradient of summer temperature of 12°C at the treeline of the Taymyr Peninsula to 2°C in the polar desert is associated with a decrease in the number of vascular plants from approximately 250 in the south to approximately 30 in the north. Analogous decreases in plant diversity in Canada also exist (Rannie, 1986). A similar pattern occurs in the animal world, for instance, in day butter-
Hypoarctic Optima in the

The group characterizes the southern tundra but usually in local habitats and wet depressions.

- Soil algae, the mosses Hylocomium splendens and
- the lichens Cetraria nigricans, the seaweed
- Phyllospadix eburnea, and
- the herb Eriophorum vaginatum

The common raven Corvus corax, the peregrine falcon Falco peregrinus, the white wagtail Motacilla alba, and the northern wheatear Oenanthe oenanthe

TABLE II

<table>
<thead>
<tr>
<th>Category</th>
<th>Distribution</th>
<th>Plants</th>
<th>Birds</th>
<th>Mammals and invertebrates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperarctic</td>
<td>Polar desert and in the northernmost part of the tundra zone</td>
<td>The dwarf shrubs Salix polaris and S. arctica (this group is relatively small, but it has an important value in the subdivision of the tundra zone into subzones)</td>
<td>The black-bellied plover Calidris hiemalis, the lesser golden plover Pluvialis dominica, and the dunlin Calidris alpina and C. minutilla</td>
<td>The lemming Dicrostonyx groenlandicus and the ermine Mustela erminea</td>
</tr>
<tr>
<td>Euarctic</td>
<td>Northern part of the tundra zone, rare in the southern part</td>
<td>The dwarf shrubs Salix polaris and S. arctica (this group is relatively small, but it has an important value in the subdivision of the tundra zone into subzones)</td>
<td>The black-bellied plover Calidris hiemalis, the lesser golden plover Pluvialis dominica, and the dunlin Calidris alpina and C. minutilla</td>
<td>The lemming Dicrostonyx groenlandicus and the ermine Mustela erminea</td>
</tr>
</tbody>
</table>
| Hemiarctic | Throughout the tundra zone, but most frequent in the middle | Most of the dominant plant species: the grasses Arctophila folia (Alpine fen grass), the sedges Carex bigelowii/arctisbirica and Carex staggersi, the lichens Cetraria nigricans, the seaweed Phyllospadix eburnea, and
- the herbs Eriophorum vaginatum
- the moss Polytrichum juniperinum
- the lichens Cetraria delina | The lemming Dicrostonyx groenlandicus and the ermine Mustela erminea | The lemming Dicrostonyx groenlandicus and the ermine Mustela erminea |
| Hypearctic | Optima in the southern tundra subzone | This group characterizes the southern tundra subzone, the shrubs Betula nana and
- the sedge Eriophorum vaginatum
- the lichens Cetraria nigricans, the seaweed Phyllospadix eburnea, and
- the herbs Eriophorum vaginatum

The pygmy Lapland longspur Calcarius lapponicus, the lesser golden plover Pluvialis dominica, and the dunlin Calidris alpina and C. minutilla | The vole Microtus oeconomus, the Flajellated collared lemming Pseudomyx Heniorynchus, the long-legged willow warbler Phylloscopus collybita, and
- the spiders and M. torquatus, and
- the amphibians

Polyarctic

Tundra and far to the north in the tundra but usually in local habitats and wet depressions.

- Soil algae, the mosses Hylocomium splendens and
- the lichens Cetraria nigricans, the seaweed
- Phyllospadix eburnea, and
- the herb Eriophorum vaginatum

The common raven Corvus corax, the peregrine falcon Falco peregrinus, the white wagtail Motacilla alba, and the northern wheatear Oenanthe oenanthe

The wolf Canis lupus, the ermine Mustela erminea, the weasel Mustela nivalis, the vole Microtus oeconomus, the Flajellated collared lemming Pseudomyx Heniorynchus, the long-legged willow warbler Phylloscopus collybita, and
- the spiders and M. torquatus, and
- the amphibians

Zonal boreal

Not abundant and constrained to the south of the Arctic in bogs habitats such as river valleys, south-facing slopes, and wet areas.

- True species of Larix, the orchid Cypripedium calceolus, the shrubs
- Salix myrtyloides, and
- the herbs Dactylina ramulosa, D. maritima

The forest birds Tardus and"muscovus and T. pilaris (ibises),
- chiff-chaffs Pycnonotus sinensis and P. cellicia, and river
- ducks Aythya nova, A. penelope, and A. crecca

Remi and the walrus, Olaus groenlandicus

Zonal Arctic

Polyzonal Taiga and far to the north in the tundra but usually in local habitats and wet depressions.

- Soil algae, the mosses Hylocomium splendens and
- the lichens Cetraria nigricans, the seaweed
- Phyllospadix eburnea, and
- the herb Eriophorum vaginatum

The common raven Corvus corax, the peregrine falcon Falco peregrinus, the white wagtail Motacilla alba, and the northern wheatear Oenanthe oenanthe

The wolf Canis lupus, the ermine Mustela erminea, the weasel Mustela nivalis, the vole Microtus oeconomus, the Flajellated collared lemming Pseudomyx Heniorynchus, the long-legged willow warbler Phylloscopus collybita, and
- the spiders and M. torquatus, and
- the amphibians

* Compiled from information in Matveyeva and Chernov in Nutall and Callaghan (2000).
flies (Fig. 2) and spiders. At a larger scale, there is a general decrease in biodiversity with an increase in latitude from temperate to arctic ecosystems (Fig. 2). In this case too, decreasing biodiversity is associated with a gradient of decreasing temperature with increasing latitude, as discussed previously.

In the plant kingdom, patterns of decrease in biological diversity toward the North Pole differ among taxa. Some diminish their significance in the biota or even disappear (Ericoids), others change proportionally to the level of general diversity (Fabaceae and Rosaceae), and a third type retain a high level of biological diversity in the Arctic where their proportion becomes higher (Saxifragaceae and Brassicaceae). Not only is there a decrease in plant species and family diversity with increasing latitude but also the paucity of life forms within the plant kingdom of the Arctic results in a simpler vertical vegetation structure than that of the forested regions further south. In the Arctic, two- or three-layered vegetation, with the height of the tallest shrub layer up to 2 m, changes into the nearly two-dimensional plant cover of polar deserts where most of the biota is concentrated into a thin film of less than 5 cm above the ground and no more than 5 cm below the ground.

Because any species belongs to a particular life (or growth) form, the changes in species composition inevitably leads to changes in the proportion of life forms...
in a wide sense. The proportion of rooted species (flowering plants) which control their water content to merely adnate species (cryptogams) decreases from approximately 1:2 at the treeline to about 1:5 in the polar desert. Tall shrubs are common in the south of the tundra but absent in the north. Dwarf shrubs are widely spread throughout the tundra zone but absent in polar desert ecosystems. The proportion of long rhi zomatous herbs decreases while that of tap-rooted plants increases in the northern direction. Loose caespitose species are replaced by dense tufted cushions or mat-forming species (Fig. 3). Within a species, there is also a tendency to form more compact growth forms in the north compared to the south.

Latitudinal patterns of biodiversity within the Arctic may reflect global trends. Passerines and other bird taxa that are important in the tropics (e.g., kingfishers, woodpeckers, parrots, hummingbirds, and pigeons) decrease successively as a proportion of the avifauna from the equator to the poles. Passerine birds comprise about 60% of the avifauna in temperate and tropical forests, about 50% up to the northern taiga landscapes, 40% in forest–tundra, and 20% in the northern part of the tundra zone. In contrast, the proportion of Charadriiformes (40%) and goose-like birds (20%) in the Arctic avifauna is high and decreases toward the south. Charadriiformes are reduced to 20% in the north taiga and 10% in the south, and they contribute only 3% in the tropics.

B. Circumpolar Diversity

The Arctic’s biota has a relatively large number and proportion of circumpolar species compared with circumboreal species. One of the obvious reasons for this is the decreasing circumference of latitudes toward the poles. Circumpolar vascular plants comprise about 45–50% of the extant flora in the south, 60–65% in the

FIGURE 3 Growth forms of tundra plants, with example species. Plant functional type classifications basically follow the growth forms depicted (reproduced with permission from Webber, 1978).
north of the tundra zone, and more than 70% in the polar desert.

Less than half of the Arctic birds may be considered circumpolar because many species formally considered as circumpolar have large gaps in their distribution. Thus, about 20% of the Arctic avifauna are true circumpolar species, and only 10% of the tundra terrestrial fauna are circumpolar species. Only 2 of 30 tundra Tipulidae species have true circumpolar distributions. Only 1 species is conditionally circumpolar among about 40 leaf-eater beetles (Chrysomelidae) known in the Arctic. In two very common plant genera, Saxifraga and Dryas, which each include about 40 species in the Arctic, only 25% (10 in each) have true circumpolar distributions. The rich genus Pedicularis has 6 circumpolar species out of 24, but 5 of these have large gaps in Europe, Greenland, or eastern North America. In the Ranunculus genus, 9 of 25 species are circumpolar. Only 1 of the 10 species of Dryas in the Arctic, Dryas punctata sensu stricto, has a truly circumpolar distribution.

Species that form monotypic genera have the most pronounced circumpolar distribution. Examples from the animal world are arctic fox (Alopex lagopus), snowy owl (Nyctea scandiaca), old squaw duck (Clangula hyemalis), and calliphorid (Boreellus atriceps) and from the plant world the grasses Arctophila fulva, Dupontia fisheri, and Pleurogonon sabineti, the herbs Koenigia islandica and Oxyria digyna. Oxyria digyna and K. islandica have enormous geographic ranges, being circumpolar and with extensions far south: K. islandica is bipolar.

Most species have distributions narrower than circumpolar. In general, there are five main types of longitudinal distribution. Species can have (i) true circumpolar distribution or, if represented on both continents, they can have (ii) amphib-Atlantic or (iii) amphib-Pacific (Beringian) distributions or they can occur only on (iv) Eurasian or on (v) North American continents. Within each continent, there are an additional five subregions, giving 10 zonations in total (Porsild, 1959): They can have a wide geographic range (Eurasian species or North American species) or occur in a more restricted area (e.g., Siberian or Alaskan species).

C. Effect of Topography on Biodiversity

The macrolevele that is represented by Arctic lowlands and mountain ranges is the main cause of interland-
A series of filters has selected, and is selecting, species and genotypes which occur in the Arctic. (Walker in Chapin and Körner, 1995). The first filter is the presence of a species in the region, the second set of filters is the biogeography of species within the Arctic, and the third set of filters is internal filters within communities and the environment. Körner (in Chapin and Körner, 1995) adds a time dimension to the geographical filters. For example, the current range of species and communities in the Arctic has been determined by survivors from the thinning of past populations and floras during glacial periods together with the species which immigrated into the Arctic during the Holocene. In his alpine examples, Körner describes a second filter or sieve, grazing and freezing temperatures, and a third filter which acts at the microscale and specifically selects plant species which can tolerate certain microclimates, soil disturbance levels, and moisture regimes.

The next set of filters, as suggested by Körner, act at the microscale and consist of numerous interactions among plants in communities, between plants and herbivores, and between all organisms and the microenvironment. These fine-scale filters select for organisms which are preadapted to Arctic environments or which have developed specific adaptations. Among the animals, preadaptations can be seen in traits common in arctic–alpine distributions (i.e., with populations in alpine areas to the south of the Arctic) and in those from nutrient-poor bogs of the Chukotka Peninsula, well to the north of the tree line; ferns, orchids, and umbellifers near warm springs on Disko Island; and dwarf shrubs in some inner fjord areas of Svalbard. Steep south-facing river banks combine warm and well-drained habitats with the possibility of migration along the river corridor and are another focal point of community diversity, often supporting trees within the tundra. Animals also create diversity hot spots, when they are dead, by bringing back nutrients to their nesting places or dens. Examples are bird cliff communities, fox and wolf dens, and lemming gardens. Such outliers represent potential “inocula” for recolonization (discussed later) and plant community change in a future warmer climate and support animal species with more southerly distributions. This process may have occurred about 12,000 or 13,000 years ago in Beringea when birch rapidly spread and created a “birch zone.”

VI. CONTROLS OF BIODIVERSITY

Localized topographical extremes can be associated with extreme species distribution outliers. Deep valleys with thick snow cover in winter and warm weather in summer at latitude 75°N on Taymyr have dense willow thickets of Salix lanata up to 2 m tall. Warm-water springs in Alaska, Chukotka, and middle-west Greenland create localized warm microclimates and many stands in the surroundings of these hot springs support extreme range disjunctions of southern, even boreal, species. Examples include stands of Populus balsamifera in the northern foothills of the Brooks Range and in the Chukotka Peninsula, well to the north of the tree line; ferns, orchids, and umbellifers near warm springs on Disko Island; and dwarf shrubs in some inner fjord areas of Svalbard. Steep south-facing river banks combine warm and well-drained habitats with the possibility of migration along the river corridor and are another focal point of community diversity, often supporting trees within the tundra. Animals also create diversity hot spots, when they are dead, by bringing back nutrients to their nesting places or dens. Examples are bird cliff communities, fox and wolf dens, and lemming gardens. Such outliers represent potential “inocula” for recolonization (discussed later) and plant community change in a future warmer climate and support animal species with more southerly distributions. This process may have occurred about 12,000 or 13,000 years ago in Beringea when birch rapidly spread and created a “birch zone.”

In the Arctic, during the Pleistocene, it could be argued that the first filter was environmental in that only freezing-tolerant plant species could survive in a region which became “arctic” in character. Later, during the Holocene, the next set of filters could be seen as relating to migration and dispersal abilities in that many species followed the retreating ice margins northwards, whereas some glacial relics spread outwards from their refugia. It is possible that this filter could at least partially explain the high abundance of cryptogams, which have easily dispersed spores, relative to vascular plants in the Arctic. In general, however, it can be assumed that the main set of filters on current Arctic biodiversity are environmental rather than constraints on migration and the “available species pool.” This assumption is supported because, for example, latitudinal treelines are often situated more to the south now than during the earlier Holocene (although trees never reached the high Arctic, even during the climatic optimum). Also, herbivore diversity and population sizes are limited by low primary production resulting primarily from low nutrient availability. Any local increase in nutrient availability (e.g., through animal activity) dramatically stimulates biodiversity and productivity.

The next set of filters, as suggested by Körner, act at the microscale and consist of numerous interactions among plants in communities, between plants and herbivores, and between all organisms and the microenvironment. These fine-scale filters select for organisms which are preadapted to Arctic environments or which have developed specific adaptations. Among the animals, preadaptations can be seen in traits common in boreal forest animals. In plants, preadaptations are particularly evident in those with arctic–alpine distributions (i.e., with populations in alpine areas to the south of the Arctic) and in those from nutrient-poor bogs of the temperate region. Specific adaptations of plants to the Arctic environment have been considered to be few. A constraint on adaptation rate has been the generally young nature of the Arctic flora, the longevity of many plants, and the sporadic successful completion of reproduction and seedling recruitment in areas of closed vegetation (Callaghan and Emanuelsson, 1985).

The fine-scale filters on biodiversity arising from biotic interactions were mentioned earlier. Plant competition is not thought to be a major force displacing species in the Arctic. Indeed, species removal experi-
ments often result in little compensatory growth of remaining species. Instead, facilitation and succession are important and might explain to some extent the similarity of plant aggregations throughout large areas of the Arctic. If there were any competitive “tensions” within these aggregations, they would be expected to have broken down over time or over geographical ranges.

Herbivory is only sometimes a filter on biodiversity. Perhaps in general, herbivores do not eat particular plant species to extinction and, indeed, can stimulate plant diversity. In other cases, such as grazing by snow geese in the Hudson Bay area, high population pressure can lead to ecological cascades ending with loss of plant populations and ultimately soil erosion. Similar dramatic impacts on vegetation can occur when population peaks of two different herbivores, both feeding on the same plant, occur within a short period of time. Analogically, the nature of tundra vegetation in Fennoscandia can be produced when subarctic birch forests are defoliated by insects and regenerating trees are browsed by reindeer. Also, browsers such as the megafauna of the Pleistocene and early Holocene had the ability to maintain open vegetation (Zhimov et al. in Chapin and Körner, 1995), whereas grazers can control dominance in plant communities, e.g., by suppressing moss growth. Although such impacts can be dramatic at the local scale, the generally wide distributions of many Arctic species ensure survival of the species. Survival of rare species and ecotypes, however, has a greater risk.

V. CONSEQUENCES OF BIODIVERSITY
A. Environmental Consequences of Biodiversity

One way in which we can readily consider the importance of biodiversity within processes is to understand the potential impact of biodiversity decline. Primary successional processes are indicative of the potential impact of biodiversity loss. One of the most thoroughly studied successional series within the Arctic is that present at Glacier Bay, Alaska (Chapin et al., 1994). The retreat of the glaciers during the past 200 years has led to the development of a series of communities that have been colonized by species from nearby riparian systems. The activity of burrowing animals such as ground squirrels or arctic foxes may provide exposed areas of disturbed soil that act as potential refugia for plant primary colonizers. This is an example of faunal diversity promoting floral diversity.

As succession proceeds, the potential for interaction between organisms increases because their zones of influence are more likely to overlap with increasing organism density (this is especially true of plants). Because of an increase in the number of species, the potential complexity of interactions within a community also increases. Therefore, there may be a parallel increase in the level of biodiversity and the degree and complexity of interactions between individuals. This type of pattern might be found along the latitudinal gradient from high Arctic polar semidesert toward more productive tundra environments. Types of interaction include direct resource-based interactions (e.g., competition) and indirect interac-
tions, possibly involving higher trophic levels. For example, there is evidence that herbivores may encourage the development of vascular plants by preventing the formation of dense moss mats in certain tundra environments (Zhimov et al. in Chapin and Korner, 1995; R. van der Wal, personal communication). Because of the mobile nature of larger herbivores (e.g., reindeer and musk ox), their importance in the functioning of a community at any given time may not be clear. As in the case of plant primary colonizers, the key role that they play, and their importance in terms of maintaining diversity at both the species and the community level, may only be apparent at particular points in time or space.

Facilitative interactions may be of particular importance within Arctic communities. In the case of plant interactions, facilitation may act through the amelioration of environmental severity and may occur within all successional stages, including mature communities. Facilitative plant species may be keystone species within an environment; for example, on areas of frost heave-disturbed scree the growth of spreading plant species with persistent nets of rhizomes or roots has been shown to stabilize the scree and to allow the development of other plant species within the stabilized zone. A critical factor in this case is the morphology of the colonizing plant. Other types of plant growth form may have different facilitative effects. Although all plants may provide shelter to their neighbors, there are certain species and growth forms, especially cushion plants, which can develop in the absence of this initial shelter and can raise temperature differentials between their meristems and the air by more than 25°C. They subsequently provide shelter for others. In one alpine example, five species of cushion plant, including Silene acaulis, together “hosted” 93 other species of plants. In another example, grasses growing in moss cushions in the high Arctic grew dramatically faster than those growing outside moss cushions (Fig. 4). Diversity of form (Fig. 3) is therefore perhaps one aspect of Arctic biodiversity that may be critical for ecosystem function.

Differences in plant growth form also strongly affect the movement of resources within an environment. Spreading rhizomatous species that maintain physiological connections between individuals may move nutrient resources away from localized resource-rich areas. Species with spreading root systems but a localized canopy may in turn concentrate resources within a given area. Similarly, differences in the behavior of animals may also lead to variability in resource distribution and resource movement throughout the environment. During the winter period when their grazing is confined to roots, rhizomes, and the stem bases, lemmings can accumulate within their nesting areas small reservoirs of nutrients in plant tissue and droppings. Other grazing species such as geese may redistribute nutrients through the environment.

Not only are nutrient distribution processes dependent on the variability in life form, and hence diversity, but also nutrient cycling and soil conditions (e.g., temperature, moisture, and pH) may be strongly related to the diversity of organisms within an ecosystem. Hobbie (in Chapin and Korner, 1995) describes how differences in plant traits may directly and indirectly influence biogeochemical processes within tundra plant communities. For example, species with high productivity rates (generally deciduous and graminoid species) may increase the rate of nutrient turnover, woody species may increase the sequestration of carbon, and graminoid species may attract herbivores and increase the degree of herbivore nutrient cycling. Interspecific variability in rooting depth and leaf area influence soil moisture conditions, which in turn may influence soil nutrient availability, soil biota diversity, ion uptake, and soil pH. Leaf litter tissue quality differences may also influence soil microbial processes and decomposition rates. Mixtures of litter from more than one arctic plant growth form, particularly herbs and dwarf shrubs, have been shown to decompose at a faster rate than single-species litters (H. Quested et al., personal communication). The diversity of a plant community will therefore have important consequences for soil environmental conditions and terrestrial nutrient cycling which
is a major constraint on productivity of Arctic ecosystems. Species diversity may also affect atmospheric environmental conditions by influencing soil–atmosphere exchange processes. Verville et al. (1998) found that sedges contribute to methane emission from Arctic wet meadow communities by acting as a conduit for methane release from anaerobic soils (rather than by producing methane directly). In contrast, mosses tended to limit the evolution of methane (Fig. 5).

B. Natural Resources and Environmental Exploitation

At the most simplistic level, biodiversity can be considered a natural resource. Arctic ecosystem biodiversity, in terms of species number (i.e., alpha diversity), is not high. One key future of Arctic biodiversity as a resource is its unique components, i.e., species (and processes) that are confined to Arctic ecosystems. However, existence per se is not commonly considered to be a resource, and we tend to regard resources as items for exploitation by ourselves, either currently or in the future. Certain types of human exploitation of Arctic natural resources are limited by biodiversity and by productivity, namely, those in which an organic crop is removed from the environment. Biodiversity determines the number of resources available and productivity limits the rate at which they can be exploited. For humans to live within the Arctic using extant natural resources, they must travel over a wide area to obtain and concentrate sufficient energy, nutrients, and raw materials. The Saami people of northern Fennoscandia follow their reindeer because productivity in general is very low and it is necessary for the reindeer herds to move over a vast area in order to obtain sufficient resources. This is true both for the semidomesticated reindeer herds of the Saami and for the wild caribou herds of North America. One impact of low biodiversity in this context may be the necessity of exploiting one species intensively, such as the Saami exploitation of reindeer, rather than being able to switch exploitation between species on a seasonal basis.

Nonindigenous human exploitation of wildlife has involved hunting to accumulate raw material for manufacturing processes (e.g., furs and blubber). The replacement of many animal-derived products with synthetic materials derived from oil has reduced the pressure on Arctic animal populations. However, hunting of Arctic wildlife still occurs, either by indigenous peoples, such as the Saami or Inuit, or for sport. This latter type of hunting is just one particular branch of the development of tourism within the Arctic. Ecotourism depends heavily on the wildlife of the Arctic. The biodiversity of the Arctic, including unique Arctic species, and the paucity of species and barren nature of high-latitude Arctic systems in particular (i.e., their inherently low biodiversity) may be features of Arctic environments that are attractive to visitors. In this case, biodiversity can be an exploitable resource, and as ecotourism develops perhaps the monetary value attached to the biodiversity of the Arctic will increase.

A more intangible natural resource dependent on Arctic biodiversity is the potential for future acquisition of knowledge. The low biodiversity and process simplicity within Arctic ecosystems makes them a valuable resource for testing ecological theories. Arctic ecosystems may provide a second type of scientific resource, i.e., they might contain undiscovered or unrecognized useful natural substances. Although the low biodiversity of Arctic ecosystems suggests that the number of possibly useful substances is less than that in diverse
ecosystems (e.g., rain forest), we may find unique sub-
tances that have evolved in response to the Arctic
vironment. For example, it may be that the Arctic
cludes us with a source of natural antifreeze chemi-
cals. Similarly, the mechanism whereby some Arctic
pecies can tolerate extreme anoxia might be important
understanding the problems associated with the sus-
ceptibility of some trees and crops to waterlogging.

Mineral resources are another natural product ex-
tracted from Arctic environments. However, in this case
the problem of exploitation is not one of low biodiver-
sity but rather that of exploiting the resources without
damage to biodiversity, as discussed in the following
section.

VI. THREATS TO BIODIVERSITY IN
THE ARCTIC

The Arctic is currently undergoing rapid changes: cul-
tural and sociological changes resulting in exploitation
of natural and nonrenewable resources, changes in cli-
mate, and changes in pollution—both local and trans-
boundary (Nutall and Callaghan, 2000). Many of these
changes are predicted to accelerate in the future and
these have implications for changes in biodiversity
within the Arctic.

Extractive industries (oil, nickel, etc.) have resulted
in large local impacts on species, including reductions
in the diversity of soil microorganisms (Evdokimova in
Callaghan et al., 1995). Examples of impacts of mineral
extraction are found in towns of the Kola Peninsula,
but there has been surprising resilience of the vegetation
and pollution-tolerant forms are being actively selected
for (M. Kozlov, personal communication). Oil spills in
Alaska and the Perchora region of Russia have had
dramatic effects on local populations of both plants and
animals. However, the large and widespread distribu-
tions of Arctic biota provide a system in which changes
in biodiversity from this source over wide areas are
unlikely. More serious is the slow recovery potential of
Arctic vegetation which exacerbates disturbances to
ecosystems.

Harvesting of natural resources on land in the Arctic
consists of hunting, collecting (e.g., berries), and rein-
deer husbandry. Effects of hunting have probably had
serious impacts on a small proportion of Arctic biota
over the Holocene. It is likely that the large northern
grazers and herbivores such as the mammoth were
hunted to extinction at the beginning of the Holocene,
whereas hunting within the past 300 years probably
hastened the demise of the great auk. Following the
advent of firearms in the north, wolf populations have
been reduced to just a few animals in the subarctic of
Fennoscandia. Currently, international agreements
protect some species such as polar bear and their popu-
lations have increased. However, the health of many
species of birds which migrate to the Arctic is deter-
mimed more by hunting practices in the overwintering
grounds than in the Arctic. This is more difficult to leg-
islate.

Changes in reindeer husbandry practices in Feno-
sascandia during the past 300–400 years from hunting
to herding and, recently, from herding to almost farm-
ing in some areas have had impacts on predators
(hunted by husbanders) over large areas and on grazing
pasture productivity and biodiversity. Restrictions on
the nomadic lifestyles of reindeer husbanders both in
Fennoscandia (due, for example, to constraints of na-
tional boundaries) and in areas of the Russian Arctic
(due to imposition of a sedentary lifestyle for women
and children) have resulted in large-scale impacts on
vegetation with an often dramatic reduction in lichen
cover. Changes in lichen biodiversity remain to be docu-
mented but are likely to be pronounced.

Pollution in the Arctic is still generally less than in
more southerly regions and the pathways of pollutants
into the Arctic and their concentrations within Arctic
biota have been well documented recently (Reiersen in
Nutti and Callaghan, 2000). However, little is known
about the sensitivity of Arctic biota to the wide range
of pollutants which are found there.

During the 1.8 million years of the Pleistocene, the
Arctic has undergone severe climatic fluctuations asso-
ciated with glacial and interglacial periods. Even within
the last interglacial period of the Holocene, climate
has changed greatly, resulting in a current cool period
which is associated with lower and more southerly tree-
lines in many places in the Arctic than found in the
early Holocene. Since the Little Ice Age, however, which
ended in the north Atlantic region about 150 years
ago, mean annual temperatures have been increasing.
Since at least 1960, temperatures have been changing
throughout much of the Arctic. In continental Alaska
and eastern Canada, and also in central Siberia, mean
annual temperatures have risen by up to 1.5°C per
decade. However, in eastern Canada and midwest
Greenland, temperatures have been decreasing by the
same amount. Overall, though, the Arctic's temperature
has increased and there has been a concurrent increase
in the "greenness index" of vegetation in the north, as
seen by satellite images, which suggests an increase in
the growing season of approximately 11 days over the
past 20 years. The recent increases in temperature in the Arctic have been associated with increased pest outbreaks in the taiga, increased frequency of forest fires, and a warming of the permafrost with an increase in disturbance events.

General circulation models which predict future climates for various concentrations of greenhouse gases all agree that future warming will be greatest in the Arctic, and particularly during wintertime. Because of the strong correlations between biodiversity and temperature discussed previously, it is expected that future warming will have a large impact on biodiversity in the Arctic. Experiments in a range of Arctic habitats which simulate warmer conditions show that one of the first effects of warming on plant communities is a change in the dominance of vascular plant species existing when the manipulations started and an increase in biomass (Chapin et al., 1995; Press et al., 1996). However, as the vascular plants increase in biomass, lichen biomass and cover decrease and some moss species (e.g., Hylocomium splendens) also decrease in abundance. In the case of lichens, the decreases seen in the experiments parallel decreases found along natural geographical gradients of temperature. The major implications of the experiments are therefore that, over the medium term, warming will change the dominance among vascular plant species, the abundance of lichens and mosses will decrease, and the immigration of more southerly species will not happen. Over the longer term, clearly biodiversity of vascular plants will increase but that of cryptograms will decrease. In addition to this decrease in lichens is the impact of changing reindeer husbandry discussed previously. Overall, the Arctic’s important role as a reserve of primitive plants is threatened, and because there is such a great cover of lichens in the Arctic, this will have an effect on their biodiversity at a global level.

In contrast to predicted decreases in biodiversity of the mosses and lichens, warming may increase diversity of all plant groups in some places. Glacier forefield expansion as a result of glacier retreat during warming and the open ground of polar deserts and semi-deserts offer open niches for the establishment of new plants. Because colonization of these habitats is usually by seeds and spores, colonization can expand the local species and subspecific genetic diversity. This is in marked contrast to closed areas of vegetation in the mid- and subarctic where plant establishment is mainly via clonal growth. However, even here, increases in disturbance due to thawing permafrost are likely to open up niches and allow new genets to establish.

VII. CONCLUSION

Although biodiversity in the Arctic is low compared with that in other regions, a surprisingly varied array of landscapes, ecosystems, species, ecotypes, and topographic micropatterns exist. Relationships between diversity and environment are superficially simple because land-use impacts are relatively few in the Arctic. However, beneath the surface, interactions between the low temperatures of the Arctic and its biota are complex: Temperature has many direct effects on the biota but also controls nutrient cycling in cold soils which severely constrains the abundance of vascular plant species and the organisms which depend on them. Importantly, interactions between plants and climate are twodirectional: Plant growth forms facilitate dramatic differentials to occur between plant and air temperature, whereas tundra soils have sequestered the greenhouse gas carbon dioxide throughout much of the Arctic. Currently, rapid sociological and environmental changes are occurring in the Arctic. Although low biodiversity is compensated for to some extent by large and widespread populations, the vast wilderness areas are nevertheless under numerous threats. Our major challenge is to further document (Conservation of Arctic Flora and Fauna, 1996) and monitor the biota in the remote Arctic wildernesses and to understand both how biodiversity patterns are formed and how they contribute to the functioning of ecosystems. Only then can we hope to establish conservation strategies that can cope with rapid environmental change.

Acknowledgments

We gratefully acknowledge Dr. C. Jonasson, D. Hollingworth, and C. Aune, who kindly helped with the preparation of some of the figures.

See Also the Following Articles

ANTARCTIC ECOSYSTEMS • BOREAL FOREST ECOSYSTEMS • LATITUDE, COMMON TRENDS WITHIN

Bibliography


ARTHROPODS (TERRESTRIAL), AMAZONIAN

Joachim Adis

Max-Planck-Institute for Limnology, in cooperation with National Institute for Amazonian Research

I. Taxa and Number of Species
II. Origin and Geographical Distribution of Species
III. Species Richness: Causes and Maintenance
IV. Biotope and Habitat Specificity of Species
V. Morphological Species
VI. Genetical Species
VII. Prospects

GLOSSARY

arboreal Tree inhabitants ("arboreal").
assemblage Collective occurrence of several individuals representing the same species or several species without interspecific relationships.
bionomics Mode of life of a species.
biotope Characteristic living space, of distinguishable nature compared to its environment, of a (bio)community [also "(bio)coenosis"].
bivoltine Species with two generations per year.
community Collective occurrence of individuals representing several species with, at least in part, interspecific relationships (also "coenosis").
flood resistance Submersion ability of weeks or months.
flood tolerance Submersion ability of few hours up to several days.
habitat Characteristic living space or site of a species.
plastron Thin layer of air, held by specific body structures, into which oxygen from the surrounding water is added by means of diffusion to the same extent as oxygen is withdrawn by breathing.
plurivoltine Species with several generations per year (also "multivoltine").
quiescence Dormancy in which inhibition of development depends directly on environmental factors.
terrestrial Soil inhabitants.
Univoltine (Bivoltine) Species with one (two) generations per year.

ESTIMATION OF THE biodiversity in terrestrial arthropod taxa of the "bioregion" Amazonia depends on solid data of species, collected from different biotopes and habitats, that include comparative long-term information on their bionomics and ecology. Some examples, mostly deriving from studies in Central Amazonia near Manaus, are available and clearly show the difficulties in achieving a more general statement.
I. TAXA AND NUMBER OF SPECIES

Approximately 2% of the world's Arachnida and 3% of the Myriapoda live in Amazonia. However, Ricketts represent 28%, Schizomida and Scolopendromorpha 9%, and Psocopoda more than 7% of species known worldwide in the respective taxa (Table I). To estimate the extant species in Amazonia is difficult, especially in those cases in which numbers differ by two or three different orders of magnitude from the species described worldwide. For mites which, like spiders, are considered a mega- or hyper-diverse taxon in the Arachnida, opinions of taxonomists differ substantially. This is also due to different methods used to derive such estimates (Platnick, 1999). In millipedes, approximately 250 species have been described from the Amazon drainage basin but 5000 are estimated to exist (Table I). This is based on the assumption that most genera originated in the Andes, where most of the known species now occur but there have been few sampling efforts.

Mega-diversity and high densities are attributed to Neotropical insects, in particular those inhabiting forest canopies. The dominant taxa in the canopy of non-flooded primary forests in Central Amazonia are Formicidae and Diptera. In studies by Guerrero in 1995 and 1996 near Manaus (Reserva Ducke), these two taxa represented 52 and 10%, respectively, of the 325 arthropods/m² obtained on average by fogging the canopy of 40 trees (nine species of the families Sapotaceae and Lecythidaceae) once each during the dry and rainy seasons. In another study in this reserve, two canopies of a widely distributed Amazonian tree, Goupia glabra Aubl. (Celastraceae; height 38 and 45 m) were fogged in intervals of 6 or 24 months (1991–1994). A total of 95 ant species were found on a single tree. In comparison, there are 105 ant species recorded from all of Germany. Of the 124 ant species obtained from both trees, more than one-third were represented by singletons, i.e., one specimen per species. Data also indicate a biotic interaction between predatory ants (probably Crematogaster spp.) and the gall-building Cecidomyiidae and between Cecidomyiidae and the parasitic Hymenoptera. This was not determined from previous studies in the tropical canopy (Adis et al., 1998a).

Comparative to the high density and biodiversity of Formicidae and Diptera in tree canopies is that of Acari and Colembola on the forest floor. These two taxa represent between 50 and 80% of the 26,000–74,000 terricolous arthropods extracted per square meter from 14-cm soil depth in non-flooded “terra firme” forests of Central Amazonia. However, due to taxonomical difficulties in both groups, data on species diversity are scarce and based on adult animals, although just as many immatures have been caught. The highest number of oribatid mites, 97 morphospecies, were reported by Wunderle from a primary forest near Pucallpa, Peru. In the Manaus area, Franklin found 71 and 74 morphospecies of oribatid mites in a primary forest on yellow latosol and on white sand, respectively, and 57 morphospecies in a 3-year-old secondary forest on yellow latosol (previous primary forest that was cut and burned). The ratio between the total number of specimens and morphospecies decreased accordingly, being 53.3, 36.3, and 22.8, respectively. In the same forests, Oliveira obtained 74, 65, and 65 morphospecies, respectively, of Collembola (53–64% represented Isotomidae and 7–16% Entomobryidae), but specimens: morphospecies ratio was lowest in the primary forest on white sand (49.5) compared to the primary and secondary forests on yellow latosol (67.6 and 61.8, respectively). Statistical analysis of data indicated that morphospecies similarity is dependent on both the nature of the soil (in springtails) and human disturbance (in oribatid mites). Only 32 oribatid morphospecies were obtained from pastures near Manaus.

II. ORIGIN AND GEOGRAPHICAL DISTRIBUTION OF SPECIES

Little is known about the centers of evolution and dispersal for arthropod taxa which occur in Amazonia. Several pathways are currently being discussed.

A. Origin in the Neotropics with a Generally Wide Range of Geographical Distribution (Chilopoda: Scolopendromorpha)

No scolopendromorph taxon of generic level is known as Amazonian endemic as it is at least present in other areas of the Neotropical realm. Studies of Schileyko showed that only 23% of the total 52 species currently known in Amazonia, representing 6 of the total 11 genera, seem to be endemic. The wide geographical distribution might be related to the relatively large size, high mobility, swimming ability, and predacious mode of life of an ancient group.
| Arthropods (Terrestrial), Amazonian | 251 |

| TABLE I |
| Families, Genera, and Species of Taxa Representing the Arachnida and Myriapoda in the World and in Amazonia (up to 1999) as Well as Estimates of Existing Species (Fossil Taxa Not Included) |

<table>
<thead>
<tr>
<th>Species</th>
<th>Families</th>
<th>Genera</th>
<th>Described</th>
<th>Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arachnida</td>
<td>World</td>
<td>&gt;566</td>
<td>&gt;8869</td>
<td>&gt;92,529</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>&gt;136</td>
<td>&gt;480</td>
<td>&gt;1592</td>
</tr>
<tr>
<td>Araneae</td>
<td>World</td>
<td>106</td>
<td>±3200</td>
<td>±37,000</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>&gt;65</td>
<td>&gt;300</td>
<td>&gt;1000</td>
</tr>
<tr>
<td>Puligradi</td>
<td>World</td>
<td>2</td>
<td>6</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Uropod (Thelyphonida)</td>
<td>World</td>
<td>2</td>
<td>16</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Amblypygi</td>
<td>World</td>
<td>5</td>
<td>20</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>1</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Schizomida</td>
<td>World</td>
<td>2</td>
<td>31</td>
<td>196 (219)</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>1</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Solifuge</td>
<td>World</td>
<td>12</td>
<td>153</td>
<td>1065 (1115)</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pseudoscorpiones</td>
<td>World</td>
<td>24</td>
<td>430</td>
<td>3100</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>12</td>
<td>31</td>
<td>75</td>
</tr>
<tr>
<td>Ricinulei</td>
<td>World</td>
<td>1</td>
<td>3</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Opiliones</td>
<td>World</td>
<td>44</td>
<td>&gt;1554</td>
<td>&gt;4559</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>13</td>
<td>73</td>
<td>160</td>
</tr>
<tr>
<td>Cyphopodites</td>
<td>World</td>
<td>6</td>
<td>28</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Palpatores</td>
<td>World</td>
<td>15</td>
<td>&gt;250</td>
<td>1000–2000</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>1</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Laniatores</td>
<td>World</td>
<td>23</td>
<td>1276</td>
<td>3460</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>9</td>
<td>66</td>
<td>136</td>
</tr>
<tr>
<td>Scorpiones</td>
<td>World</td>
<td>18–20</td>
<td>156</td>
<td>1250–1500</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>4</td>
<td>12–14</td>
<td>68–111</td>
</tr>
<tr>
<td>Acari</td>
<td>World</td>
<td>350–422</td>
<td>3300–4000</td>
<td>45,000</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>35</td>
<td>50</td>
<td>150–300</td>
</tr>
<tr>
<td>Myriapoda</td>
<td>World</td>
<td>158 (159)</td>
<td>2167</td>
<td>&gt;15,056</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>28 (29)</td>
<td>&gt;94</td>
<td>&gt;423</td>
</tr>
<tr>
<td>Chilopoda</td>
<td>World</td>
<td>20 (21)</td>
<td>325</td>
<td>&gt;3196</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>8 (9)</td>
<td>26</td>
<td>&gt;115</td>
</tr>
<tr>
<td>Scutigeromorpha</td>
<td>World</td>
<td>1</td>
<td>16</td>
<td>&gt;80</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Lithobiomorpha</td>
<td>World</td>
<td>2</td>
<td>95</td>
<td>1500</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Craterostigmomorpha</td>
<td>World</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Scolopendromorpha</td>
<td>World</td>
<td>3</td>
<td>33 (32)</td>
<td>515–616</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>2</td>
<td>11</td>
<td>52</td>
</tr>
<tr>
<td>Geophilomorpha</td>
<td>World</td>
<td>13 (14)</td>
<td>180</td>
<td>1100</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>4 (5)</td>
<td>13</td>
<td>60</td>
</tr>
<tr>
<td>Diplopoda</td>
<td>World</td>
<td>131</td>
<td>1800</td>
<td>11,000</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>16</td>
<td>55–60</td>
<td>250</td>
</tr>
<tr>
<td>Pauropoda</td>
<td>World</td>
<td>5</td>
<td>29</td>
<td>700</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>2</td>
<td>8</td>
<td>52</td>
</tr>
<tr>
<td>Symphyla</td>
<td>World</td>
<td>2</td>
<td>13</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Amazonia</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

*Note: Numbers in parentheses include forthcoming descriptions and revisions.*

*9 Modified with permission from Ads and Harvey (2000).*
B. Origin in the Andes and/or the Guyanian Shield and Subsequent Dispersal into the Amazon Basin
(Arachneognatha: Meinertellidae;
Coleoptera: Carabidae;
Diplopoda: Polydesmida)

Representatives of the Meinertellidae probably advanced from the Andes into the lower situated forests of the Amazon Basin. During this process, the formerly petrophilous animals, inhabiting mountain floors, adapted to an arboriculous mode of life in the trunks and canopies of trees. They even acquired a vertical jumping ability on the tree trunk, which was found for the first time in the arboriculous species Neomachillettus aditi in a black-water inundation forest of Central Amazonia. In these periodically flooded forests the Meinertellidae succeeded to colonize the forest floor as well. Flood resistance of eggs (observed in N. scandens) and the determination of egg development by the flood pulse (a quiescence duration correlated with the period of inundation) are prerequisites for a potential distribution on the waterway (on logs or floating meadows). Neomachillettus aditi was also found in an inundation forest above Letícia (Columbia) and might have reached Central Amazonia on the Rio Solimões-Amazonas. Similarly, the Neomachillettus species which currently inhabit the Caribbean and the coastal regions of the United States might originally have been transported by ocean currents from the mouth of the Rio Amazonas and the Rio Orinoco to these areas. An actually waterproof, resistant blastoderm cuticle and a presumably long duration of egg development during transportation (1 year or more, as in European species) certainly favor a geographical distribution on waterways. An interruption of the quiescence due to drying of the means of transport used, a rapid eclosion of juvenile animals, and a fast achievement of maturity (all observed in N. scandens) favor the new colonization of a biotope as well. Useful for this colonization is the swimming ability and a temporal survival on the water surface (flood tolerance), which was reported for several species of the Machiloidea (Adis, 1992; Snurk, 1984).

Erwin and Pogue presented biogeographical maps with centers of endemicism, based on taxa distribution and cladograms, for sister groups of the arboriculous carabid beetles genus Agra from Neotropical forests. Patterns of species evolution are directed, in part, from the North Atlantic coast (including the northern Guayanian shield) to upper Amazonia (palmata lineage) and from the North Atlantic coast to either upper and middle Amazonia or lower Amazonia (erythropus lineage). According to Erwin (1991), these hypothetical "centers and corridors of radiation" can be discovered and targeted for long-term protection through analyses of diverse arthropod groups (beetles in particular) and detection of congruent patterns among radiation lineages.

C. Origin in Non-flooded and/or Floodplain Forests of the Amazon Basin
(Scorpiones, Chilopoda: Geophilomorpha)

Many scorpion species apparently require strict environmental conditions and thus are found only in narrow zones of distributions. Very few species currently known from eastern Amazonia are present in western Amazonia. One of the four principal "corridors of distribution" (or radiations) in South America is located in Central Amazonia. Rupture zones and eco-barriers between these corridors are attributed to geo-historical events. For instance, the rupture zone of distribution found in the region of the Orinoco delta might be due to the proposed existence of a large inland lake in Central Amazonia during the late Pleistocene and early Holocene (Lourenço, 1994).

In the Geophilomorpha, half of the approximately 100 Neotropical species representing the genera Schen-dyllops, Pectinilungis, and Ribautia seem to be endemic to areas identified as Pleistocene forest refuges (see Section III) based on evidence from woody angiosperms (France, 1982). Valleys of the major rivers (Amazon, Orinoco, Paraná, and Paraguay) are suggested to be the main routes of subsequent distribution, even to high-altitude sites (Pereira et al., 1997).

D. Origin in Treeless Floodplains of the Amazon Basin
(Coleoptera: Carabidae)

According to Erwin, the main centers of evolution in carabid beetles are open floodplains along rivers or lakeshores in the equatorial tropics with somewhat stable conditions. Taxon pulses which effect these flying and highly vagile waterside generalists caused, by means of adaptations, their dispersal in three directions: (i) colonization of tree-tops (canopy specialists), (ii) colonization of mountains (altitude specialists), and (iii) colonization of temperate regions (climate specialists). For the colonization of temperate regions, seasonality is considered a basic requirement. Seasonal changes, primarily due to periodical precipitation (° primary
or ultimate ecolfactor), cause seasonal inundations in riverine habitats (e.g., Central Amazonia) or dryness of the soil due to a decreasing river level (e.g., East Zaire). Both situations cause an interruption of the reproduction in carabid beetles during these unfavorable conditions. They induce a gonad dormancy which is regulated by changing temperatures in the habitat (= secondary or proximate ecolfactor). For this reason, Central Amazonian inundation forests have been postulated to represent one of the locations at which seasonality has been generated in carabids and other arthropods (Adis, 1992; Erwin, 1998).

III. SPECIES RICHNESS: CAUSES AND MAINTENANCE

Biological diversification in the Neotropics is the subject of ongoing and frequently antagonistic discussions. Several theories have been proposed to explain the high species richness of Amazonian plants and animals, mostly in a deductive sense, based on selected groups. Studies on distribution patterns aim to detect centers of endemism. It is hoped that biogeographical maps, now also based on cladistic methods, will elucidate evolutionary patterns among these centers. However, it is accepted that data bases are still insufficient (with the possible exception of those for birds) for the type of detailed study needed to understand distribution patterns for most Neotropical groups. Full knowledge of distribution will not be obtained until we understand what sets the limits, for example, the edges of species distribution (Vanzolini and Heyer, 1988).

Past assessments of this type have been based mostly on the presence of Pleistocene refugia (see Section II), also called the “Hafer effect.” This is the theory that Neotropical forest fragmentation occurred during northern ice advances due to Milankovitch cycles and the temporary existence of forests and grasslands (savanna), called “island” refugia, that caused temporal centers of endemism (Hafer as cited in France, 1982; Hafer, 1996).

Studies on 162 species of Neotropical butterflies (Nymphalidae) by Brown indicated the existence of 50 principal forest centers of evolution and endemism. Smaller areas, recognizable as 38 forest refuges, presumably acted during the last major dry, cold spell in the Quaternary (20,000–13,000 years B.P.). Studies by Erwin and Pogue on arboricolous carabid beetles, adapted to life in Neotropical forest canopies (Genus Agra), showed that species group distribution is much broader.

In this case, areas considered to be centers of endemism are geographically larger than that in butterflies and vertebrates (France, 1982; Vanzolini and Heyer, 1988).

Hafer’s forest refuge model as stated deals only with non-flooded terra firme forests, their fragmentation, and/or replacement by savannas (sensu lato). There has been no indication that riverine forests disappeared along the lengths of rivers, but rather it appears that savannas were behind such forests as they are today. For this reason, Erwin and Adis consider Central Amazonian floodplain forests, especially along black-water rivers in the Negro delta (called igapô), to be short-term refuges and long-term evolutionary centers (France, 1982; Adis, 1984). Flood cycles greatly influence speciation rate and dispersion of arthropods. Evidence from Rio Tarumã Mirim, an affluent of the Rio Negro above Manaus, suggests alternation of two types of inundation: annual flooding and continental water table rise during the past million years. The area examined originally was covered by a non-flooded terra firme forest. Periods of high sea level caused a back-up of main riverine courses in Amazonia. During these periods, the affluent Tarumã Mirim was high enough that annual water fluctuations of the Rio Negro led to an inundation lasting several months in the forest under study. The first inundation was several million years ago. Since that time, various periods of high sea level (according to Fairbridge) have occurred (e.g., 250,000, 170,000, and 85,000 years ago; Fig. 1), with annual flooding of 10,000–30,000 years duration. Approximately 6000 years ago, the sea level reached a height of about 5 m below mean sea level. Since then, the examined forest has been inundated again. Flood cycles were responsible for extant vegetation formation in the igapô. During high sea level, the (lower) bank of the Rio Negro suffered longer periods of inundation and, according to Sioli, possible formations of large “rias.” Igapô forests were drowned (at least partially) and backed up minor tributaries which subsequently had annual flooding (e.g., Rio Tarumã Mirim; Fig. 1A). They replaced non-flooded terra firme forests that previously occupied the area. Intermittent dry spells during high sea level periods may have caused formation of savanna and non-flooded terra firme forest refugia (Fig. 1B), as postulated by Hafer. During low sea level periods, annual inundation occurred only along main rivers (e.g., the Rio Negro); this has been the case for at least 1 million years, probably including all glacial advances of the Pleistocene and perhaps even since the Andes took their present configuration. Non-flooded vegetation on terra firme (“uplands”) spread extensively, forcing igapô forests along affluents to retreat to main rivers
Intermittent dry periods, according to Haffer, again caused savanna formation with possible non-flooded forest refugia on terra firme and "igapó forest refuges" (Fig. 1D).

Flood cycles have a dramatic impact on inundation forest fauna. This is especially important when two kinds of flooding patterns manifest themselves through time. In the greater region of Amazonia, there could have been periods favoring lakeshore species, then inundation forest species, and then riparian species (see discussion of carabid beetles, see Erwin and Adis as cited in France, 1982). Large waterbodies and drying afluenls, with possibly additional regression of non-flooded forests on terra firme in both cases, cause interruption and isolation of igapó forests. Species vagility then seems to become a crucial factor in sorting out and isolating gene pools. Thus, small barriers may be highly disruptive to populations of individuals with low vagility, particularly non-flying soil arthropods. Supposing that intense competition and predation promote rapid speciation (especially in igapó forests, in which flooding mixes terrestrial and canopy biotas during half the year) and that the isolation of igapó forests is sufficiently long for rapidly evolving species, this model represents another key to Amazonian species richness.

In fact, recent data indicate that (i) igapó forests retain many endemic species; (ii) the fauna of igapó forests is different in composition and activity patterns when compared to that of adjacent non-flooded terra firme forests; and (iii) in igapó forests across the Rio Negro and its tributaries, taxonomically different subspecies and species are found.

A third theory considers the high plant and animal diversity of Amazonian non-flooded terra firme to be a response to extremely low nutrient concentrations in geographically impoverished ecosystems (Fittkau, 1973; Klinge, 1973). Species richness and simultaneously occurring high diversity of moist tropical ecosystems do not reflect high nutrient supply but rather a mode of adaptation to continuous restriction of nutrients or food substances under otherwise permanently favorable living conditions. Plants and animals are assumed to act as highly efficient "nutrient traps." During evolution of multiple life forms, shortages must have been of significant importance at all times.

A model group supporting this theory are the carabid communities on fig fruit falls. Paarmann and Adis sampled 8962 beetles on 65 fruit falls from 10 fig species between 1991 and 1996 in a non-flooded terra firme forest (Reserva Ducke) near Manaus. Eight of the 36 species collected represented the spermatophagous genus Notiobia, which accounted for 92% of all carabid beetles obtained. Only one Notiobia species was dominant. The abundance distribution of the eight Notiobia species was very similar on fruit falls of the two most common fig species as well as on fruit falls of the remaining fig species. The dominance structure varied considerably between individual fruit falls. Also, the abundance distribution changed during the course of a single fruit fall. Only 2 of the 8 Notiobia species were found to be specialized fig seed feeders, able to reproduce only on fig fruit falls (Vaniteck et al., 1994). The remaining 6 species of this genus use fig fruit falls as stepping stones between fruit falls of their host trees, which are separated by time and space.

Another theory for the high within-community diversity of tropical arthropods is that continuous stochastic local disturbances in nature are assumed to prevent the achievement of any long-term equilibrium (climax) state. These so-called stochastic non-equilibrium models assume that the presence of a species at a vacant site is important. This may represent an advantage against all species that arrive later. Niche overlaps are assumed to be very common within species-rich communities. As a result, neither successional nor a climax community can emerge. This is in contrast to the so-called deterministic equilibrium models, which are based on the ecological niche. Each organism maintains a defined position in its environment and, driven by competition, the system goes through defined successional stages and a structurally predictable climax equilibrium results (Linsenmair, 1990).

Examples which clearly support one or the other mechanism are not available for Amazonian arthropods. For example, concerning the diversity of the seed-feeding ground beetle community of the genus Notiobia, they are specialized on certain seeds as food for a successful larval development (deterministic process) and yet the non-fig seed specialists consume fig seeds to survive periods of food shortage. The necessity to do so limits the tendency to specialize. If stochastic processes do play a role in maintaining

**FIGURE 1** Postulated formations and shifts of black-water inundation forests along main rivers and afluenls of Central Amazonia within the past 100,000 years due to changes in sea level and climate (see text for further explanation; modified with permission from Adis, 1984 with kind permission from Kluwer Academic Publishers).
the diversity of the seed-feeding Notiobia species, this role is probably less important. Thus, they are "chance specialists" adapted to the unpredictability of fruit falls. The fig seed specialists belong to the "touring group," moving from one fig fruit fall to the next, depending on a sufficient density and an annually equal distribution. The aseasonal fruiting of figs is caused by their pollination biology.

IV. BIOTYPE AND HABITAT SPECIFICITY OF SPECIES

Conservation or sustainable use of Amazonian ecosystems demand information on biodiversity. Due to the lack of basic comparative data on arthropods, recent recommendations mostly relate to the high diversity and rarity of species found in certain regions, which are called areas of "high biological value" (Anonymous, 1991; Dinerstein et al., 1995; Kress et al., 1998), rather than to specific indicator species of biotopes and/or habitats revealing endemism and possible centers of radiation. Two long-term studies on pseudoscorpions and tiger beetles demonstrate the validity for such information.

Pseudoscorpions were studied intensively in Central Amazonian biotopes between 1975 and 1990. The 35,000 specimens collected represented 26 genera and 60 species (Adis and Mahnert, 1990). Of these, 29 species (48%) occurred exclusively in non-flooded terra firme forests and 25 species (42%) were restricted to floodplain forests. Species from terra firme forests were more tericolous, with 69% living in litter and soil. Species from floodplain forests were predominantly arboricolous, with 88% living in the trunk and/or canopy region. Differences in habitat selection are attributed to the impact of the flood pulse in Amazonian floodplains, i.e., long-term inundation of 5-7 months duration. Only 11 species were found in forests inundated by white-water (Rio Solimões-Amazonas) compared to 20 species found in black-water forests (Rio Negro). This is attributed to the differences observed between biotopes (flora and soil structure) as well as between the two river systems (current and sediment load; Adis as cited in Junk, 1997).

Data bases such as these help to answer general questions about specific biotopes. For example, the terrico-

---

**TABLE II**

Species of Tiger Beetles (Carabidae; Cicindelinae) Sampled in White-Water Floodplains (Varzea) of the Rio Solimões-Amazonas and on Non-Flooded Terra Firme Near Manaus, Central Amazonia

<table>
<thead>
<tr>
<th>Species</th>
<th>Varzea floodplains</th>
<th>Terra firme upland</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cylindera (Plectographa) sutoralis (Fabricius) (O, d)</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Megacephala (Tetracha) sobrina punctata Castetsianu (O, n)</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Megacephala (Tetracha) spinosa (Brullé (O, n))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Megacephala (Flauroamia) aequinoctialis Dejean (O, n)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Megacephala (Platanius) kugli Chaulok (O, n)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odontochila confusa (Dejean) (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pentacentra (Mesacanthina) cribra Brullé (O, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pentacentra (Pentacnem) egregia (Chauloi) (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pentacentra (Poecliochila) lacordairei Gory (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anitra septulica (Fabricius) (O, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bristia (Cicindela) argyriata (Fabricius) (O, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bristia (Cicindela) pretiosa (Dokhourioli) (O, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cerastus varius (Gory) (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. benthamia (Stenocentum) asperulum Bates (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. benthamia formicarium (Fabricius) (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cylindera (Cylindera) masto (Klug) (O, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Megacephala (Tetracha) bilunata Klug (O, n)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odontochila cayennensis (Fabricius) (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odontochila chrysas (Fabricius) (O, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odontochila luripes (Dejean) (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odontochila marginiguttata (Dejean) (O, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odontochila nigrovaria (Horn) (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pentacentra (Poecliochila) lacordairei Gory (F, d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pentacentra (Poecliochila) ventralis (Dejean) (O, d)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Abbreviations used: F, forest; O, open areas; d, diurnal; n, nocturnal.

*Modified from Adis et al. (1990b).
lous pseudoscorpions of terra firme forests on white-sand soil in Central Amazonia (called campinarana) represent the species spectrum of primary terra firme forests on yellow latosol and show no endemic species (Adis and Mahnert, 1993). This reinforces the geological results that indicate that campinarana does not represent forests along former river-beds which have dried up, as previously suggested, but rather the final stage of podzolization (i.e., the transformation of clayey latosols to white-sand podzols by long-term weathering and leaching processes).

Of the 24 species of tiger beetles (Carabidae: Cicindellinae) studied between 1991 and 1998 in the Manaus area (Table II), 15 were found in non-flooded terra firme areas. Of these, 7 species lived in forests. Five of them, being diurnal, inhabited the floor (Fig. 2: CRA, OCA, OLU, ONI, and PLA), and 2 species were canopy dwellers (CAS and CFO). Eight species, 1 nocturnal (MBI) and 7 diurnal (ASE, BAR, BPR, CMO, OCH, OMA, and PVE), lived in open areas. Three of these were found on mostly bare latosol (BAR, OCH, and PVE). The other 5 species occurred on mostly bare white sand (ASE, BPR, CAR, CMO, and MBI). The species assemblage in open areas indicated an impact by human intervention: Pentacoma ventralis (PVE) is known as a species that inhabits extensive cleared forest sites and Odontocheta marginiguttata (OMA) occurs in secondary forests. Another 9 tiger beetle species were recorded from white-water floodplains (várzea; Table II). Of these, 3 diurnal species inhabited inundation forests (Fig. 3: OC, PE, and PL). Six species, 2 diurnal (CS and PC) and 4 nocturnal (MA, MK, MF, and MS), lived in open areas. Four were found on bare beaches (CS, PC, MP, and MS). The other 2 species (MA and MK) inhabited grass-grown areas behind the beaches which were mainly composed of annual and perennial aquatic grasses. Only 1 species, Pentacoma lacordaii, occurred in the forests of both non-flooded terra firme and floodplains (Table II).

Data clearly show the biotope and habitat specificity of these tiger beetle assemblages. Moreover, the lifecycle of species on non-flooded terra firme differed greatly from those inhabiting floodplains, which indicates a high adaptation to the respective biotope and habitat. Data such as these might also help to elucidate questions of general interest. For example, the dichotomy of forest or open-habitat assemblages might be important for understanding tiger beetle distribution in Amazonia. Similar patterns were reported by Heyer for frogs east of the Andes (Vanzolli and Heyer, 1988). Data also support the potential use of tiger beetles as bioindicators for monitoring the degradation and regeneration of Amazonian forests. In Venezuela, forest-floor species assemblages changed significantly with the degree of forest disturbance, and each stage of disturbance was characterized by a particular subset of species (Rodriguez et al., 1998).

**V. MORPHOLOGICAL SPECIES**

In millipedes, taxonomy is mainly based on the structure of genitalia in males, the gonopods. However, in the process of adaptation to new environments, "pioneers" apparently change their external structure more quickly than their gonopods, which results in difficulties in defining species. According to Hoffman (1990), the diversification of the external body form with little modification of gonopod structures in the order Polydesmida seems to be associated with those taxa (usually genera) which seem to have recently occupied a new area or biotope independently of close relatives and are indulg-
ing in a burst of adaptive radiation unimpeded by sibling competition. It is presumed that after an initial period of great diversification some selection of successful lineages would occur. During this time, body form types, stabilized by adaptational factors, would tend to remain constant. The generic variability, however, might be expressed much more rapidly in structures such as genitalia, which are less directly influenced by environmental constraints. Thus, a period would ensue in which these species groups (or genera) would consist of species readily distinguishable by highly distinctive gonopods. Such a stasis would endure for a long time unless it were again possible for a fragment of a group (probably migrant populations of one or two species) to escape into a new area or new ecological niche and begin the cycle anew. This situation seems to be different from the punctual evolution of new clades in that two basically different character systems alternate in phase with the chronological status of the organisms in "new" and "old" territory and thus underscore the principle that phylogeny cannot be adequately interpreted outside the context of biogeography.

New insights derived from studies on the Neotropical genus Pycnotropis support these postulations. Among the 26 species known, only 3 are restricted to Amazonian white- and mixed-water inundation forests, whereas the remainder are found in non-flooded terra firme forests. Biogeographically, the history of Pycnotropis can be viewed as one implying the origins in the western Andes, the postulated center of origin (and radiation) of the entire subfamily Amphelininae, with subsequent waves of downstream dispersal throughout the Amazon Basin. The present-day distribution of Pycnotropis is the result of numerous, apparently ongoing vicariance events involving repeated gene flows downstream of the Amazon and its tributaries (radiation corridors) from source areas, with colonization and recolonization of inundation forests and/or non-flooded terra firme habitats. The genus is still in a stage of very active speciation because at least some of its species display pronounced variation in both external and gonopod characters. This is supported by preliminary tests concerning genetic variation in Pycnotropis tida, a species that has been described twice due to variation in mor-
phological characters. The pattern of a somewhat "unsettled" speciation process seems to contrast the one observed in the chelodesmid genus *Camptomorpha*, which likewise is species rich in the Andes and abundantly represented in the Amazon Basin. It is, however, a good example of a taxon in the "stable body-variable gonopod" stage, with each species being distinct morphologically (Golovatch et al., 1998).

VI. GENETICAL SPECIES

Comparative phenological studies on arthropod species which inhabit both floodplain and non-flooded terra firme forests in Central Amazonia suggested an univoltine life cycle of populations in annually inundated forests and a plurivoltine mode of life of populations in terra firme forests. Representatives of both populations were morphologically alike; hence, they represented the same species (e.g., in Pseudoscorpiiones, Diplopoda, Symphyla, and Archaeognatha).

In the case of *Neomachilis scabdens*, a member of the Meinertellidae (Archaeognatha, Insecta), the two populations in terra firme and floodplain forests were first postulated to be "biotope-specific races." Morphological characters that would justify separation into two species could not be found by the taxonomist, at least with traditional research methods. However, protein analyses by means of electrophoresis (testing 15 enzymes) revealed that there is no gene flow between populations from floodplain and terra firme forest types, even when they are only 50 m distant from each other. Thus, this "species" was proposed to represent two different species based on eco-phenological and genetic characteristics. Without this genetic background, problems arise if the two genetical species occur sympatrically because, due to the intermixture of data, the eco-phenological differences are hardly or not at all recognizable. This was the case in a mixed-water inundation forest near Manaus (Wolf and Adis, 1992). Until now, culturing and breeding of Amazonian Meinertellidae was unsuccessful; thus, interbreeding experiments could not be performed.

The polydesmidan millipede *Pyenotropis tida* (see Section V) represents a second example. In the Manaus area, populations which inhabit secondary terra firme forests breed throughout the year, whereas populations of white- and mixed-water inundation forests show a defined, univoltine reproduction during the non-aquatic phase (Vohland and Adis, 1999). The genetic data were obtained from allozyme analyses (testing 14 enzymes by electrophoresis) and from a specific satellite DNA. In this case, results suggested that individuals from the two biotopes represented populations of the same species, although genotypic structures among and within local populations indicated processes of ongoing genetic differentiation (Bachmann et al., 1998). A successful interbreeding thereafter reinforced the genetical results.

For taxonomists, the use of genetic techniques and the consideration of ecological as well as ethological data represent the final, although mandatory, stage of analysis in systematic biology. In the case of Amazonian Pseudoscorpiones and Symphyla, both genetic and breeding exercises have not been successful.

VII. PROSPECTS

Amazonian terrestrial arthropods are special: They have adapted in some way to their environment. Only detailed studies at the species level will reveal the "strategies" which have evolved. For example, there are (i) grasshoppers on aquatic macrophytes in the Rio Amazonas with spiny front legs, like mantids, that follow their prey into tree canopies during high water (*Phlugis teres*, Tettigoniidae); (ii) a small flood-resistant "terrestrial" millipede which lives up to 11 months submerged in inundation forests along the Rio Negro, breathing under water by means of a piastron and feeding on algae (*Gonographis adisi*, Fyrgodesmidae); and (iii) larvae of a carabid beetle (normally a predaceous taxon) which accomplish their development feeding on fig seeds (*Notiobia flavicinctus*) but apparently depend on larvae of their sister species (*N. pseudolimbipennis*) which are able to open the seed shells with their morphologically larger mandibles. Some insights into this fascinating world have been achieved since Bates voyages on the "River Amazonas." However, it took about 150 years to adopt his suggestion to study 1 acre of rain forest in more detail. Two approaches are currently being taken. One is to develop an All-Taxa-Biodiversity-Inventory and the other to develop an All-Biota-Taxon-Inventory (Platnick, 1999). Both enterprises require an intensified training of tropical taxonomists and eco-entomologists.

Acknowledgments

I gratefully acknowledge the valuable comments received on a draft of this article by Terry L. Erwin, Sergei I. Golovatch, Richard
See Also the Following Articles

AMAZON ECOSYSTEMS • ARACHNIDS • BEETLES • FOREST CANOPIES, ANIMAL DIVERSITY • HYMENOPTERA • MYRIAPODS

Bibliography


ASIA, ECOSYSTEMS OF

Elgene O. Box and Kazue Fujiwara
University of Georgia and Yokohama National University

Glossary

angiosperms Flowering vascular plants in which the seeds, which are produced from the ovules of the flowers, are enclosed in fruits developed from the matured ovaries; the largest class of vascular plants, as opposed to cryptogams (mosses, etc., which do not produce seeds) and conifers (in which the seeds are not enclosed in fruits).

biodiversity Biological diversity, i.e., the variety of living organisms in an area, including the variety in genes, species, functional types of organisms, and ecosystems.

biome A major terrestrial ecological community and landscape type, characterized by more or less uniform physiognomy of its potential natural vegetation and with characteristic fauna and flora, such as the tropical rainforest, the warm deserts, or the temperate grasslands.

boreal Pertaining to the northern high latitudes (but not polar), regions which contain large continental landmasses and thus continental climates with moderate summers but long, severely cold winters (from Greek boreas, the north wind).

ecosystem A community of organisms (plants, animals, and microbes) and their physical environment interacting as an ecological unit, such as a lake, a wetland, a forest, or an agricultural landscape.

endemic Occurring only within a comparatively restricted geographic range within a specific region.

global circulation The regular global circulation pattern of the earth's atmosphere, which generates the world's basic climate types; the system involves an Intertropical Convergence zone of low pressure near the equator, subtropical high-pressure belts near the Tropics of Cancer and Capricorn, trade winds flowing from these high-pressure belts toward the equatorial low, and westerly winds in the midlatitudes.

monsoon system A wind system covering the eastern half of Asia in which winter cooling of the large landmass produces strong, stable high pressure, with clear skies and outward flow of cold, dry air, and summer warming produces low pressure, drawing wet air masses, with clouds and rain, inward from the adjacent oceans.

Pleistocene glaciation The expansion of ice sheets over the large northern continents, especially northern...
Europe and North America (excluding the north-west) during the “Ice Ages.”

Potential natural vegetation The vegetation cover which would develop naturally in an area and become stable (not replaced by a subsequent stage) if all outside disturbances were eliminated.

Species richness The total number of species of an area or ecosystem.

Temperate Pertaining to the climates and landscapes of the midlatitude regions, which are seasonally warmer and cooler, with winter frost even in most coastal areas and at least partial dormancy or collapse of the vegetation.

Species richness The total number of species of an area or ecosystem.

Species richness The total number of species of an area or ecosystem.

Temperate Pertaining to the climates and landscapes of the midlatitude regions, which are seasonally warmer and cooler, with winter frost even in most coastal areas and at least partial dormancy or collapse of the vegetation.

Tropical Astronomically, the region lying between the Tropic of Cancer (23° ¹⁄₂ N) and Tropic of Capricorn (23° ¹⁄₂ S); more generally, involving the climates and landscapes characteristic of this region, which are essentially frost free in the lowlands and permit biological activity as long as water is available.

Vegetation The total plant cover of an area.

Zonation The tendency of climate, soil, and natural landscape types to occur in distinct latitudinal zones (e.g., tropical, subtropical, and temperate), as generated by the global atmospheric circulation system.

Asia is the world’s largest continent, extending from the Mediterranean Sea in the west to the Japanese Archipelago in the east, and from polar landscapes in northern Siberia to hot deserts in the southwest and tropical regions south of the Himalaya and China. Tropical Asia includes perhumid equatorial rainforest climates as well as seasonally wet and dry tropical savannas, rainforest woodlands, and monsoon forests, all extending to some degree from India to the Philippines and East Indies. Dry regions of Asia include an eastward extension of the hot, subtropical Sahara desert into southwestern Asia, large expanses of higher latitude interior deserts with cold winters, and also a small “temperate” region and an east–west strip of temperate grassland (steppe) which spans almost the entire continent. The main areas of temperate forests are in East Asia, with deciduous and mixed forests in the north, rich evergreen “laurel” forests in the south, and extensive areas of mostly secondary pine forests throughout. Mountain areas of Asia have many endemic species, especially relict conifers, and the mountains of East Asia represent the only truly large mountainous area in the warm-temperate and humid subtropical zones of either hemisphere. Northern Asia was less glaciated than Europe or North America and contains enormous expanses of relatively diverse Siberian conifer forests, polar and upland tundra, and Lake Baikal (the world’s deepest). Wetlands in Asia include mangroves, coastal strand forests, salt marshes and estuaries, and terrestrial swamp forests, marshes, and bogs, most of which are highly productive and represent critical habitat for terrestrial as well as aquatic animals. Despite long histories of human habitation and landscape alteration, Asia retains the highest biodiversity of any continent, due not only to its size but also to its climatic and topographic complexity and its complex geological and evolutionary history. This biodiversity is threatened in Asia as elsewhere by human overpopulation and overdevelopment.

I. ASIA THE REGION

Asia is traditionally separated from Europe, with which it shares the world’s single largest land-mass Eurasia (Fig. 1). Even by this cultural definition, however, Asia remains the world’s largest continent and extends almost halfway around the world, from a small western coastline on the Mediterranean Sea (about 25°E in Turkey) eastward to the Bering Strait (170°W) separating it from Alaska. Asia also spans almost the entire latitudinal range of the Northern Hemisphere, from a northern coastline in the Arctic zone to islands lying slightly south of the equator in Indonesia. Asian superlatives include the world’s

- highest mountain (Mt. Everest, 8848 m) and mountain range (Himalaya);
- largest highland plateau (Tibet);
- lowest terrestrial elevation (Dead Sea, −400 m);
- largest (Caspian Sea, 143,200 km²) and deepest (Lake Baikal, 1620 m) lakes; and
- coldest, most continental region except for Antarctica (northeastern Siberia), with mean monthly temperature ranging from 20 to −60°C and extremes to −80°C.

Also found in Asia are

- the world’s third (Borneo) and fifth (Sumatra) largest islands;
- four of the world’s seven rivers over 5000 km long (Yangtze, Ob’-Irtysh, Yenisey-Angara, and Huang-Hei);
- one of the world’s three regions with essentially no rainfall (Tarim Basin); and
- two of the world’s three stations with more than 10 m of average annual rainfall (windward slopes
ASIA, ECOSYSTEMS OF

FIGURE 1 Location and main physiographic features of Asia. Asia extends almost halfway around the world, from Turkey and Arabia in the west to the Bering Strait in the east. It also extends from the Arctic Ocean in the north to slightly south of the equator in Indonesia. Middle Asia extends across the Turanian Basin from the Caspian Sea to the Tien Shan and Altai mountains and is largely steppe and desert. Central Asia includes the high, cold Tibetan Plateau and the desert basins to its north, on into the steppes of Mongolia. Siberia (Ussurian Russia) is the mostly forested region north of the east-west mountains and Middle Asian deserts. Other regions and the geological development of Asia are described in the text.

In addition, Asia has been home to some of the oldest human cultures and to areas of very long, continuous human habitation, especially in the Middle East, India, and Southeast and East Asia. Asia represents the largest part of Laurasia, the northern half of the supercontinent Pangaea, which existed in the late Paleozoic and broke up during the Mesozoic (Table I; Raven and Axelrod, 1974). Eurasia began to separate from North America around 180 million years ago, but contact was reestablished in the Cretaceous when Siberia approached Alaska, forming Beringia. At about this time, southwestern Asia also began to be influenced by northward-moving Africa, a portion of Gondwana, the southern part of Pangaea. The present-day geologic and physiographic struc-
### TABLE I
Some Major Events in Asia’s Geological and Evolutionary History

<table>
<thead>
<tr>
<th>Era</th>
<th>Event</th>
<th>Geological and climatic events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paleozoic</td>
<td><strong>Silurian</strong> (425–400 m BP)</td>
<td>Climate similar to present?</td>
</tr>
<tr>
<td></td>
<td><strong>Carboniferous</strong> (345–280 m BP)</td>
<td>First vascular and land plants</td>
</tr>
<tr>
<td></td>
<td><strong>Permian</strong> (260 m BP)</td>
<td>Forests of ferns, horsetails, etc.</td>
</tr>
<tr>
<td>Mesozoic</td>
<td><strong>Triassic</strong> (225–185 m BP)</td>
<td>Pangaea, glacialiation at S pole</td>
</tr>
<tr>
<td></td>
<td><strong>Jurassic</strong> (185–135 m BP)</td>
<td>Ginkgos and cycads dominant</td>
</tr>
<tr>
<td></td>
<td><strong>Cretaceous</strong> (135–65 m BP)</td>
<td>“Greenhouse” climate, high sea level</td>
</tr>
<tr>
<td></td>
<td><strong>Major separations</strong></td>
<td>Gondwana (130–100 m BP)</td>
</tr>
<tr>
<td></td>
<td><strong>Major contacts</strong></td>
<td>SW Asia and Africa (100 m BP)</td>
</tr>
<tr>
<td></td>
<td><strong>Tertiary Period</strong></td>
<td>Eurasia and N. America (180 m, initially)</td>
</tr>
<tr>
<td></td>
<td><strong>Quaternary Period</strong></td>
<td>Asia and N. America (80 m, not north)</td>
</tr>
<tr>
<td></td>
<td><strong>Pleistocene</strong></td>
<td>First primates</td>
</tr>
<tr>
<td></td>
<td><strong>Pliocene</strong></td>
<td>Most modern plant taxa, herbaceous plants abundant</td>
</tr>
<tr>
<td></td>
<td><strong>Holocene</strong></td>
<td>Temperate and high-latitude soils and biomes redevelop</td>
</tr>
</tbody>
</table>

*Times are in millions (m) or thousands (k) of years before present (BP). Sources: Pearson (1995), Bridges (1990), and Raven and Axelrod (1974).*

...
Over most of this area anyway. The “summer monsoon" global atmospheric circulation, would occur in summer
air masses in off the adjacent Pacific and Indian Oceans, continental warming causes rising air, which draws wet
but extremes are not far below the means. In summer, heavy winter snow can delay the warming
of the landmass to such an extent that the summer monsoon “fails” in places such as India, with disastrous
effects on crops.

The Philippines and East Indies are less affected by the monsoon and more by the seasonal shift in global
circulation. Near the equator, warm humid conditions continue throughout most of the year, with only a
short dry season if any at all. The length and degree of the dry season increase away from the equator, reaching
several months in the northern Philippines but also in some eastern parts of Indonesia (Lesser Sunda Islands
as well as eastern Java and southernmost Borneo and Sulawesi).

Southwestern Asia is not affected by the monsoon but rather by the pervasive effect of the subtropical
high-pressure belt. This belt of descending dry air migrates north–south seasonally, creating “Mediterranean” climates with dry summers in Turkey and throughout the Mediterranean borderlands. To the
south, however, subtropical high pressure remains the
dominant influence throughout the year, causing the desert belt across North Africa, the Arabian Peninsula,
and as far east as the Thar desert in Pakistan and western
India.

Northern Asia (Siberia) is dominated by its high latitude and large land area, resulting in ultracontinental
climates with short summers and long severe winters
rivaling the northern polar region in degree of cold.

II. BIODIVERSITY

Angiosperms are thought to have originated in South-
east Asia, and indeed Asia is thought to contain many
other foci of more recent biotic radiation as well. During the Pleistocene, when many species were lost from
heavily glaciated Europe and North America, glaciation
in Asia was less extensive, especially in the east. The Scandinavian ice sheet extended east to the Urals, but
Siberia remained largely ice-free except in the northwest
and in the mountainous northeast. A large Pleistocene
refuge for temperate Asian plants was also readily avail-
able in southern China. These climatic (and related sea-
level) changes, plus the physiographic complexity of
### TABLE II

Global Climatic Zonation in Lowlands, with Regions in Asia

<table>
<thead>
<tr>
<th>Zone</th>
<th>Temperature extremes</th>
<th>Significance</th>
<th>Asian examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical</td>
<td>No frost or other &quot;cold&quot; temperatures ever</td>
<td>Sensitivity of many tropical plants to non-freezing cold</td>
<td>South Asia and East Indies</td>
</tr>
<tr>
<td>Subtropical</td>
<td>Occasional frost or near frost, not every year and not below about (-1^\circ C)</td>
<td>Frost sensitivity of tropical evergreen and most other tropical plants</td>
<td>Southern China, Taiwan, southern Asian foothills, northern Arabian Peninsula</td>
</tr>
<tr>
<td>Warm-temperate</td>
<td>Light to moderate frost, every year or nearly so; absolute minima not (&lt; -15^\circ C)</td>
<td>Annual leaf-changing broad-leaved evergreens tolerant, (sub)tropical evergreens may not</td>
<td>Eastern, inland SE and SSW China, southern Japan, Mediterranean region</td>
</tr>
<tr>
<td>Temperate</td>
<td>Significant frost every year, occasional temperatures below about (-15^\circ C)</td>
<td>Coldness tolerance limit for evergreen broad-leaved plants (e.g., internal ice formation)</td>
<td>Northern China and Japan, Korea, interior Asia from Mongolia to Turkey</td>
</tr>
<tr>
<td>Cool-temperate</td>
<td>Moderate to significant frost every year, plus cool summer; minima can be (&gt; -15^\circ C) if oceanic</td>
<td>Growing-season warmth marginal for many typical temperate plants, including deciduous</td>
<td>Hokkaido (minima (&lt; -15^\circ C))</td>
</tr>
<tr>
<td>Boreal</td>
<td>Cool, short summer and long severe winter; absolute minima (&lt; -15^\circ C), perhaps (&lt; -40^\circ C)</td>
<td>Growing season insufficient for most deciduous trees (exceptions: larch, birch, etc.)</td>
<td>Most of Siberia, northern Manchuria</td>
</tr>
<tr>
<td>Polar</td>
<td>Short summers below (10^\circ C) and long severe winters below (0^\circ C), extremes (&lt; -40^\circ C) unless oceanic</td>
<td>Growing season too cold for wood-producing enzymes, no trees or large shrubs</td>
<td>North-coastal Siberia, island of Arctic Ocean</td>
</tr>
</tbody>
</table>

Asia, all contributed to the development and maintenance of high biotic diversity in Asia, both in and within genera. Among vascular plants, for example, both Siberia and temperate east Asia generally show higher diversity than comparable latitudes in Europe and North America.

The most complete compilation of recent estimates of both species richness and the protection status of taxa and ecosystems is titled Global Biodiversity (Groombridge, 1992), from which summaries will appear throughout this article. Species richness and endemism of plants in the different countries and regions of Asia are summarized in Table III. China and Malesia are thought to have about 30,000 plant species each (most not in common). India may have 15,000 species, and tropical and subtropical Asia may have a total of at least 50,000 species. By comparison, Latin America may have as many as 85,000 species, but this includes the region from Mexico to temperate southern South America. Within Asia, China has the largest number of plant species at 30,000, followed by Indonesia (20,000), India (15,000), Malaysia (12,000) and Thailand (12,000). A special feature of East Asia is its large number of endemic plant families, many of which are monospecific, as shown in Table IV. Species richness is lowest in the dry areas of the Arabian Peninsula and mainland southwestern Asia, but rates of endemism are high in some of these drier areas. Inventories are incomplete in all areas, with only small Brunei claiming more than 20% completion. No inventory statistics at all were available for Southeast Asia, except for Thailand and Myanmar.

Estimates of species richness and endemism in four major groups of animals, but without indication of inventory status, are provided in Table V (from Groombridge, 1992). Species richness is generally lowest in the drier areas, but the number of reptiles is still high. Numbers of birds are also fairly high in mainland southwestern Asia. Other sources of biodiversity data include the United Nations Environment Programme Global Biodiversity Assessment (Heywood, 1995) and the various publications of the World Resources Institute.

### III. REGIONALIZATION AND BIOMES OF ASIA

Global regionalization of climatic types, as related to causal mechanisms (latitude and global circulation) and
### Table III: Plant Species Richness and Endemism in Asia

<table>
<thead>
<tr>
<th>Location</th>
<th>Angiosperms</th>
<th>Gymnosperms</th>
<th>Ferns</th>
<th>Endemics</th>
<th>Percentage endemism</th>
<th>Date</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabian Peninsula</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahrain</td>
<td>193</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1991</td>
<td>c</td>
</tr>
<tr>
<td>Kuwait</td>
<td>234</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1991</td>
<td>c</td>
</tr>
<tr>
<td>Oman</td>
<td>1,018</td>
<td>3</td>
<td>14</td>
<td>74</td>
<td>7.1</td>
<td>1991</td>
<td>c</td>
</tr>
<tr>
<td>Qatar</td>
<td>220</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1991</td>
<td>c</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>1,729</td>
<td>8</td>
<td>22</td>
<td>34</td>
<td>1.9</td>
<td>1991</td>
<td>2 c</td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>340</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>—</td>
<td>1991</td>
<td>c</td>
</tr>
<tr>
<td>Yemen (P.D.R.)</td>
<td>1,373</td>
<td>3</td>
<td>41</td>
<td>98</td>
<td>4.1</td>
<td>1991</td>
<td>2 c</td>
</tr>
<tr>
<td>Yemen (Arab Rep.)</td>
<td>959</td>
<td>1</td>
<td>14</td>
<td>77</td>
<td>7.9</td>
<td>1991</td>
<td>c</td>
</tr>
<tr>
<td>Southwest Asia (mainland)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyprus</td>
<td>1,630</td>
<td>12</td>
<td>20</td>
<td>86</td>
<td>5.2</td>
<td>1977–1989</td>
<td>1 c</td>
</tr>
<tr>
<td>Iraq</td>
<td>2,914</td>
<td>7</td>
<td>16</td>
<td>107</td>
<td>3.5</td>
<td>1986–1986</td>
<td>1 c</td>
</tr>
<tr>
<td>Israel</td>
<td>2,284</td>
<td>8</td>
<td>15</td>
<td>135</td>
<td>6.7</td>
<td>1982–1984</td>
<td>1 c</td>
</tr>
<tr>
<td>Jordan</td>
<td>2,200</td>
<td>6</td>
<td>6</td>
<td>—</td>
<td>—</td>
<td>1982–1985</td>
<td>2 c</td>
</tr>
<tr>
<td>Turkey</td>
<td>8,472</td>
<td>22</td>
<td>83</td>
<td>2,881</td>
<td>30.9</td>
<td>1988</td>
<td>3 e2</td>
</tr>
<tr>
<td>South Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td>3,000</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1972</td>
<td>2 c</td>
</tr>
<tr>
<td>Bhutan</td>
<td>5,446</td>
<td>22</td>
<td>—</td>
<td>50–100</td>
<td>1.4</td>
<td>1991</td>
<td>3 c1</td>
</tr>
<tr>
<td>British Indian Ocean Territory</td>
<td>100</td>
<td>1</td>
<td>—</td>
<td>0</td>
<td>0</td>
<td>1971</td>
<td>e1</td>
</tr>
<tr>
<td>India</td>
<td>13,000</td>
<td>—</td>
<td>1,000</td>
<td>5,000</td>
<td>31.3</td>
<td>1983–1984</td>
<td>2 e2</td>
</tr>
<tr>
<td>Maldives</td>
<td>260</td>
<td>2</td>
<td>13</td>
<td>5</td>
<td>1.8</td>
<td>1983</td>
<td>1 c</td>
</tr>
<tr>
<td>Nepal</td>
<td>6,300</td>
<td>23</td>
<td>450</td>
<td>315</td>
<td>4.3</td>
<td>1978–1982</td>
<td>2 c</td>
</tr>
<tr>
<td>Pakistan</td>
<td>4,917</td>
<td>21</td>
<td>—</td>
<td>372</td>
<td>7.9</td>
<td>1986</td>
<td>2 e</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>2,900</td>
<td>—</td>
<td>314</td>
<td>900</td>
<td>28.0</td>
<td>1982–1985</td>
<td>2 c</td>
</tr>
<tr>
<td>Southeast Asia (mainland)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cambodia</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laos</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td>7,000</td>
<td>—</td>
<td>—</td>
<td>1,071</td>
<td>15.3</td>
<td>1984</td>
<td>4 e2</td>
</tr>
<tr>
<td>Indonesia</td>
<td>20,000</td>
<td>—</td>
<td>2,500</td>
<td>15,000</td>
<td>66.7</td>
<td>1984</td>
<td>4 e3</td>
</tr>
<tr>
<td>Malaysia</td>
<td>12,000</td>
<td>—</td>
<td>900</td>
<td>—</td>
<td>900</td>
<td>1991</td>
<td>3 e3</td>
</tr>
<tr>
<td>Philippines</td>
<td>8,000</td>
<td>31</td>
<td>900</td>
<td>5,500</td>
<td>39.3</td>
<td>1982–1991</td>
<td>3 e2</td>
</tr>
<tr>
<td>Singapore</td>
<td>1,000</td>
<td>2</td>
<td>166</td>
<td>1</td>
<td>0.1</td>
<td>1989–1991</td>
<td>1 e1</td>
</tr>
<tr>
<td>East Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>30,000</td>
<td>200</td>
<td>2,000</td>
<td>18,000</td>
<td>95.9</td>
<td>1990</td>
<td>3 e2</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>1,800</td>
<td>4</td>
<td>180</td>
<td>25</td>
<td>1.3</td>
<td>1978–1991</td>
<td>2 e2</td>
</tr>
<tr>
<td>Japan</td>
<td>4,700</td>
<td>42</td>
<td>630</td>
<td>2,000</td>
<td>37.2</td>
<td>1987</td>
<td>1 c</td>
</tr>
<tr>
<td>Korea (North)</td>
<td>2,808*</td>
<td>—</td>
<td>—</td>
<td>107</td>
<td>14.0</td>
<td>1976–1985</td>
<td>2 c</td>
</tr>
<tr>
<td>Korea (South)</td>
<td>2,608*</td>
<td>—</td>
<td>—</td>
<td>224</td>
<td>14.0</td>
<td>1976–1985</td>
<td>2 c</td>
</tr>
<tr>
<td>Mongolia</td>
<td>2,722</td>
<td>—</td>
<td>—</td>
<td>229</td>
<td>10.1</td>
<td>1984</td>
<td>1 c</td>
</tr>
<tr>
<td>Taiwan</td>
<td>2,993</td>
<td>20</td>
<td>365</td>
<td>—</td>
<td>[23]</td>
<td>1982–1991</td>
<td>1 c</td>
</tr>
</tbody>
</table>

* The data are regrouped by region from those presented in the Global Biodiversity report (Groombridge, 1992, p. 80) compiled by the World Conservation Monitoring Centre (WCMS). The last two columns refer to the date and status of country inventories, using codes from the original source.

† Percentage endemism is calculated from the data, unless in square brackets.

4 Date is the date of the information provided to the WCMS.

A Data includes percentage completion (first number) and method (counted/estimated): 1, <5%; 2, 5–10%; 3, 10–15%; 4, 15–20%; 5, >20%; c, counted; e, estimated based on any available information and comparable floras.

Data for North and South Korea are combined; no data are available for combined Yemen.
natural ecosystems, has been portrayed perhaps best by the widely used global system of Heinrich Walter (1985). This and the major large-area landscape types, called biomes (e.g., tropical rain forest and temperate grassland), provide a geographic framework for regionalizing the ecosystems of Asia. An attempt is made in Table VI to juxtapose the main biome types, the climate type in which they occur, the corresponding Bailey ecoregion class, and the corresponding International Union for the Conservation of Nature biogeographic provinces in Asia. Note that the “Savanna” ecoregion also includes the region of tropical moist and dry deciduous forests (“monsoon forests”) that stretches across South and Southeast Asia and which corresponds to the larger tropical wet–dry climatic regions of Africa and South America, with their areas of tropical deciduous forest, woodland, and savanna. Tropical deciduous forests are not recognized in some newer classifications because they have been so completely converted into savanna. Areas of tropical deciduous forest still remain in Asia, however, especially in the Ghats Mountains of India and from the Deccan Plateau to interior Southeast Asia. Note also that one must be careful with Russian and Chinese terminologies which use terms such as “Hot/Warm Continental” and “subtropical” for temperate forest regions which have cold to severely cold winters.
### Table V

Species Richness and Endemism of Higher Animals in Asia

<table>
<thead>
<tr>
<th>Region</th>
<th>Mammals</th>
<th>Birds</th>
<th>Reptiles</th>
<th>Amphibians</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Species known</td>
<td>Endemic species</td>
<td>Species known</td>
<td>Endemic species</td>
</tr>
<tr>
<td><strong>Mediterranean–Middle Asia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afghanistan</td>
<td>123</td>
<td>0</td>
<td>436</td>
<td>0</td>
</tr>
<tr>
<td>Cyprus</td>
<td>21</td>
<td>0</td>
<td>80</td>
<td>2</td>
</tr>
<tr>
<td>Iran</td>
<td>140</td>
<td>4</td>
<td>—</td>
<td>1</td>
</tr>
<tr>
<td>Iraq</td>
<td>81</td>
<td>1</td>
<td>145</td>
<td>1</td>
</tr>
<tr>
<td>Israel</td>
<td>—</td>
<td>2</td>
<td>109</td>
<td>0</td>
</tr>
<tr>
<td>Jordan</td>
<td>—</td>
<td>0</td>
<td>132</td>
<td>0</td>
</tr>
<tr>
<td>Lebanon</td>
<td>92</td>
<td>0</td>
<td>124</td>
<td>0</td>
</tr>
<tr>
<td>Syria</td>
<td>—</td>
<td>0</td>
<td>163</td>
<td>0</td>
</tr>
<tr>
<td>Turkey</td>
<td>116</td>
<td>0</td>
<td>284</td>
<td>0</td>
</tr>
<tr>
<td><strong>Arabian Peninsula</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahrain</td>
<td>—</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>Kuwait</td>
<td>—</td>
<td>0</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Oman</td>
<td>46</td>
<td>3</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>Qatar</td>
<td>—</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>—</td>
<td>1</td>
<td>59</td>
<td>0</td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>Yemen</td>
<td>—</td>
<td>1</td>
<td>—</td>
<td>8</td>
</tr>
<tr>
<td><strong>Tropical Asia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td>109</td>
<td>0</td>
<td>334</td>
<td>0</td>
</tr>
<tr>
<td>Bhutan</td>
<td>109</td>
<td>0</td>
<td>448</td>
<td>0</td>
</tr>
<tr>
<td>British Indian Ocean Territory</td>
<td>—</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>Brunei</td>
<td>342</td>
<td>0</td>
<td>399</td>
<td>0</td>
</tr>
<tr>
<td>Cambodia</td>
<td>117</td>
<td>0</td>
<td>325</td>
<td>0</td>
</tr>
<tr>
<td>India</td>
<td>317</td>
<td>38</td>
<td>969</td>
<td>69</td>
</tr>
<tr>
<td>Indonesia</td>
<td>513</td>
<td>163</td>
<td>1,359</td>
<td>258</td>
</tr>
<tr>
<td>Laos</td>
<td>173</td>
<td>0</td>
<td>481</td>
<td>1</td>
</tr>
<tr>
<td>Malaysia</td>
<td>264</td>
<td>14</td>
<td>301</td>
<td>4</td>
</tr>
<tr>
<td>Maldives</td>
<td>—</td>
<td>0</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>Myanmar</td>
<td>320</td>
<td>8</td>
<td>807</td>
<td>4</td>
</tr>
<tr>
<td>Nepal</td>
<td>167</td>
<td>1</td>
<td>629</td>
<td>1</td>
</tr>
<tr>
<td>Pakistan</td>
<td>191</td>
<td>3</td>
<td>476</td>
<td>0</td>
</tr>
<tr>
<td>Philippines</td>
<td>166</td>
<td>90</td>
<td>395</td>
<td>172</td>
</tr>
<tr>
<td>Singapore</td>
<td>57</td>
<td>1</td>
<td>118</td>
<td>0</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>86</td>
<td>12</td>
<td>223</td>
<td>20</td>
</tr>
<tr>
<td>Thailand</td>
<td>251</td>
<td>3</td>
<td>616</td>
<td>2</td>
</tr>
<tr>
<td>Vietnam</td>
<td>273</td>
<td>3</td>
<td>638</td>
<td>12</td>
</tr>
<tr>
<td><strong>East Asia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>394</td>
<td>62</td>
<td>1,100</td>
<td>63</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>38</td>
<td>0</td>
<td>107</td>
<td>0</td>
</tr>
<tr>
<td>Japan</td>
<td>90</td>
<td>29</td>
<td>&gt;250</td>
<td>20</td>
</tr>
<tr>
<td>Korea (North)</td>
<td>—</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>Korea (South)</td>
<td>49</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>Mongolia</td>
<td>—</td>
<td>6</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>Taiwan</td>
<td>62</td>
<td>13</td>
<td>160</td>
<td>19</td>
</tr>
</tbody>
</table>

*As in Table III, the data are regrouped by region from those presented in the Global Biodiversity report (Groombridge, 1992, p. 139).*
### TABLE VI
Biomes, Ecoregions, and Biogeographic Provinces of Asia

<table>
<thead>
<tr>
<th>Biome regions</th>
<th>Climate</th>
<th>Ecoregions</th>
<th>Biogeographic province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical Rain forest</td>
<td>I</td>
<td>Tropical rain forest</td>
<td>Malabar rain forest, Bengal rain forest, Burman rain forest, Indochinese rain forest, South Chinese rain forest, Malay rain forest</td>
</tr>
<tr>
<td>Tropical deciduous forest (mont, dry types)</td>
<td>II</td>
<td>Savanna</td>
<td>Indian–Ganges monsoon forest, Burman monsoon forest, Thailand monsoon forest, Mahanadi monsoon forest, Coromandel monsoon forest, Ceylon monsoon forest, Deccan thorn forest</td>
</tr>
<tr>
<td>Tropical islands</td>
<td>—</td>
<td>—</td>
<td>Laccadives Islands, Maldives and Chagos Islands, Cocos–Keeling and Christmas Islands, Andaman and Nicobar Islands, Sumatra, Java, Lesser Sunda Islands, Celebes, Borneo, Philippines</td>
</tr>
<tr>
<td>Subtropical desert</td>
<td>III</td>
<td>Tropical/subtropical desert</td>
<td>Arabian desert, Thar desert</td>
</tr>
<tr>
<td>Mediterranean scrub</td>
<td>IV</td>
<td>Mediterranean</td>
<td>Mediterranean sclerophyll, Japanese evergreen forest, Taiwan Bryley Islands</td>
</tr>
<tr>
<td>Laurel forest (evergreen broad-leaved)</td>
<td>Vc</td>
<td>Humid-subtropical</td>
<td>Chinese subtropical forest, West Anamalian temperate forest</td>
</tr>
<tr>
<td>Subtropical islands</td>
<td>—</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Summergreen forest (temperate deciduous)</td>
<td>VI</td>
<td>Hot/warm continental</td>
<td>Manchu–Japanese mixed forest, Oriental deciduous forests, West Anamalian temperate forest, Kamchatkan</td>
</tr>
<tr>
<td>Temperate grassland</td>
<td>VII</td>
<td>Temperate steppe</td>
<td>Mongolian–Mancurian steppe, Pamir–Tien Shan steppes, East Siberian taiga, East European taiga</td>
</tr>
<tr>
<td>Temperate desert</td>
<td>VIIa</td>
<td>Temperate desert</td>
<td>Anatolian–Iranian desert, Iranian (steppe) desert, Turanian desert, Takla Makan–Gobi desert, Tibetan desert</td>
</tr>
<tr>
<td>Boreal forest</td>
<td>VIII</td>
<td>Subarctic</td>
<td>West European taiga, East Siberian taiga, Low Arctic tundra, High Arctic tundra, Arctic cold desert</td>
</tr>
<tr>
<td>Tundra</td>
<td>IX</td>
<td>Tundra</td>
<td>Caucasian–Iranian highlands, Akas highlands, Pamir–Tien Shan highlands, Hindu Kush highlands, Himalayan highlands, Schumann–Yunnan highlands</td>
</tr>
<tr>
<td>Multi-zonal mountains</td>
<td>X</td>
<td>Mountains</td>
<td>喀喇昆仑–塔里木山高原, Altai highlands, Pamir–Tien Shan highlands, Hindu Kush highlands, Himalayan highlands, Schumann–Yunnan highlands</td>
</tr>
<tr>
<td>Lakes systems</td>
<td>—</td>
<td>—</td>
<td>Aral Sea, Lake Baikal</td>
</tr>
</tbody>
</table>

* Biome regions are denoted by names commonly used in regional and global treatments of terrestrial vegetation and biomes (Archibold, 1995; Walter, 1985; and the "Ecosystems of the World" book series).
* Climate types are those of Walter (1985): I, equatorial; II, tropical–summer–rain; III, subtropical arid; IV, Mediterranean; Vc, warm–temperate (humid east sides); Vm, marine west-coast (perhumid); VI, typical temperate (humid); VII, temperate continental (subhumid); VIIa, temperate arid; VIII, boreal; IX, polar; X, highlands (unclassified).
* Ecoregions are from Bailey (1996). See also the world map from 1999.
* Biogeographic provinces are from the classification by Udvardy (1975).
Using the synthesis presented in Table VI, Asia can be divided into four main regions of bioclimates and associated ecosystems:

1. Tropical Asia, including the tropical rain forest and monsoon forest regions, the tropical islands, and the various elevational belts of tropical mountains
2. the dry region of subtropical southwestern and interior Middle and Central Asia, including the subtropical and interior deserts, the temperate grasslands, the Mediterranean-type ecosystems in the southwest, and the drier mountain ranges
3. East Asia, the monsoonal but extra-tropical region of subtropical and warm-temperate evergreen broad-leaved forests ("laurel forests") and of temperate deciduous forests, with the more humid uplands and mountain ranges in these areas
4. Boreal and polar Asia, including the boreal forests of Siberia, northeasternmost China, and northern Mongolia and the tundra landscapes across the northern Siberian coastal plain and in the mountain areas of eastern Siberia.

Treatment of the various ecosystems of Asia will follow this framework, with an emphasis on vegetation since it provides the basic physical and trophic structure of ecosystems. Wetlands, coastal ecosystems, and some artificial ecosystems will be treated separately.

Estimates of vegetation cover type, percentage cover, and carbon storage, by country and region, are summarized in Table VII (Groombridge, 1992). The most extensive forest cover is in eastern to southern Asia and in Siberia, with non-forest landscapes dominating southwestern Asia. These data probably overestimate the extent of remaining forest.

Allotted space does not permit a detailed description of regional vegetation composition or even a more complete bibliography. Some general treatments, from which one can find references to more regional descriptions, include the following:

Tropical: Champion and Seth (1968), Mani (1974), and Whitmore (1984)
Dry region: Walter and Box (1983) and Zohary (1973)
Mountains: Chen et al. (1986), Troll (1972), and Walter (1974)

For animals, one might start with the extraordinary syntheses of Schaller (1998).

IV. TROPICAL ASIA

Tropical Asia is the region of mainly low-lying topography and warm, at least seasonally humid climates which extends from India to the Philippines and East Indies (i.e., excluding all of dry southwestern Asia). Much of the region is coastal lowland, and the highest point is on an island, at Mount Kinabalu (4130 m) in northern Borneo. The region can be divided climatically into two main zones:

1. The perhumid equatorial zone, covering most of the East Indies and southern Philippines, the Malay Peninsula, and extending along the windward southwest-facing coastlines of mainland southern Asia to southern Kerala and Sri Lanka
2. The seasonally wet–dry zone to the north, mainly across interior southern Asia from India to Vietnam and the northern Philippines, but also including the eastern Sunda Islands (6°–11°S) and some other island areas, including rain shadows such as eastern and northern Sri Lanka

The natural biome of the equatorial zone is tropical rain forest and other humid tropical evergreen forests, whereas the wet–dry zone contains natural landscapes ranging from moist and dry deciduous forests to savanna and small areas of dry evergreen forest mainly in Thailand. Other ecosystems, such as swamp forests, kerang (sclerophyll scrub), and viny limestone forests, occur on particular substrates. These main characteristic vegetation types of tropical Asia and their topographic or substrate affinities are summarized in Table VIII. Characteristic vegetation belts in mountains are summarized in Table IX (Whitmore, 1984).

A. Humid Tropical Forest

Tropical rain forest is tall, multi-layered, evergreen broad-leaved forest characterized by many epiphytes, climbing plants, and a tremendous richness of tree species with dark green, thin-coriaceous leaves, many with attenuated "drip tips" for better drainage but otherwise all looking very similar. The air is constantly humid, but the canopy is exposed each day to high solar radiation, resulting in a drier canopy microclimate and a more perhumid microclimate in the forest understory. Many tree species respond to this difference with somewhat smaller, thicker canopy leaves and larger, thinner (and often darker) understory leaves, all on the same tree. In true tropical rain forest, the only seasonality is the
TABLE VII
Estimates of Vegetation Type and Percentage Cover in Asia*

<table>
<thead>
<tr>
<th>Region</th>
<th>Coastal aquatic (%)</th>
<th>Major wetlands (%)</th>
<th>Deserts/ semideserts (%)</th>
<th>Polar/ alpine (%)</th>
<th>Grass/ shrub (%)</th>
<th>Crop and settlements (%)</th>
<th>Interrupted woods (%)</th>
<th>Major forests (%)</th>
<th>Carbon (kg/m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>4</td>
<td>2</td>
<td>13</td>
<td>12</td>
<td>20</td>
<td>11</td>
<td>17</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>Asia</td>
<td>4</td>
<td>1</td>
<td>16</td>
<td>9</td>
<td>24</td>
<td>17</td>
<td>10</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Middle East</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyprus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iran</td>
<td>1</td>
<td>30</td>
<td>41</td>
<td>7</td>
<td>17</td>
<td>4</td>
<td>1</td>
<td>14</td>
<td>0.8</td>
</tr>
<tr>
<td>Iraq</td>
<td>30</td>
<td>34</td>
<td>33</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Israel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40</td>
<td>30</td>
<td>30</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Jordan</td>
<td>44</td>
<td>49</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Lebanon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td>7</td>
</tr>
<tr>
<td>Syria</td>
<td>18</td>
<td>41</td>
<td>36</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td>3</td>
<td>9</td>
<td>37</td>
<td>18</td>
<td>25</td>
<td>7</td>
<td>1</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>Arabian Peninsula</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kuwait</td>
<td>9</td>
<td>91</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Oman</td>
<td>13</td>
<td>46</td>
<td>34</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Qatar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td>0.9</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>1</td>
<td>62</td>
<td>33</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>3</td>
<td>76</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Yemen</td>
<td>7</td>
<td>11</td>
<td>74</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>USSR (former)</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>26</td>
<td>10</td>
<td>8</td>
<td>21</td>
<td>26</td>
<td>3.2</td>
</tr>
<tr>
<td>South Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afghanistan</td>
<td>11</td>
<td>16</td>
<td>62</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td>7</td>
<td>3</td>
<td>42</td>
<td>25</td>
<td>20</td>
<td>20</td>
<td>4</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td>Bhutan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td>25</td>
<td>19</td>
<td>31</td>
<td>3.7</td>
</tr>
<tr>
<td>India</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>44</td>
<td>23</td>
<td>14</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>Nepal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23</td>
<td>35</td>
<td>42</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td>1</td>
<td>2</td>
<td>21</td>
<td>7</td>
<td>46</td>
<td>13</td>
<td>9</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>32</td>
<td>42</td>
<td>3</td>
<td>13</td>
<td>10</td>
<td>10</td>
<td>11</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Southeast Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brunei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>67</td>
<td>2.3</td>
</tr>
<tr>
<td>Cambodia</td>
<td>7</td>
<td>6</td>
<td>19</td>
<td>5</td>
<td>65</td>
<td>6.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indonesia</td>
<td>24</td>
<td>9</td>
<td>4</td>
<td>9</td>
<td>14</td>
<td>40</td>
<td>5.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>7</td>
<td>86</td>
<td>9.1</td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>33</td>
<td>45</td>
<td>6.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myanmar</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>16</td>
<td>23</td>
<td>52</td>
<td>6.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philippines</td>
<td>32</td>
<td>2</td>
<td>22</td>
<td>13</td>
<td>31</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td>41</td>
<td>7</td>
<td>49</td>
<td>5.3</td>
<td></td>
</tr>
<tr>
<td>Vietnam</td>
<td>17</td>
<td>1</td>
<td>4</td>
<td>30</td>
<td>49</td>
<td>6.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>East Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>1</td>
<td>1</td>
<td>14</td>
<td>22</td>
<td>21</td>
<td>17</td>
<td>3</td>
<td>18</td>
<td>2.4</td>
</tr>
<tr>
<td>Japan</td>
<td>21</td>
<td>1</td>
<td>18</td>
<td>5</td>
<td>18</td>
<td>27</td>
<td>30</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>Korea (North)</td>
<td>12</td>
<td>18</td>
<td>14</td>
<td>16</td>
<td>39</td>
<td>3.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korea (South)</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>20</td>
<td>24</td>
<td>22</td>
<td>3.4</td>
</tr>
<tr>
<td>Mongolia</td>
<td>32</td>
<td>32</td>
<td>1</td>
<td>5</td>
<td>32</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>1.4</td>
</tr>
<tr>
<td>Taiwan</td>
<td>29</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>5.3</td>
<td></td>
</tr>
</tbody>
</table>

* Estimates are from the Global Biodiversity report (Groombridge, 1992, p. 231), based mainly on Olson et al. (1983).

A given tree may be seen to be flowering, flushing new leaves, dropping individual or small bunches of old leaves, and carrying ripening or ripe fruit simultaneously on different branches. As one proceeds away from the equator, subtle climatic cues begin to appear: seasonal periods of less difference between day and night. As a result, the main phenological functions of the trees tend to occur simultaneously.
TABLE VIII

<table>
<thead>
<tr>
<th>Climate</th>
<th>Soil water</th>
<th>Zone</th>
<th>Elevation</th>
<th>Soil type</th>
<th>Forest type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everwet</td>
<td>Dry land</td>
<td>Inland</td>
<td>Lowland (&lt;1200 m)</td>
<td>Zonal</td>
<td>Lowland rain forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1200–1500 m</td>
<td></td>
<td>Lower montane rain forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1500–3000 m</td>
<td></td>
<td>Upper montane rain forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3000 m to treeline</td>
<td></td>
<td>Subalpine forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lowland (moistly)</td>
<td>Podzolized sands</td>
<td>Heath forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Limestone</td>
<td>Limestone forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ultrabasic rock</td>
<td>Forest on ultrabasic rock</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Coastal</td>
<td>Samal</td>
<td>Brach vegetation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Saltwater</td>
<td></td>
<td>Mangrove forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Brackish water</td>
<td>Oligoegrype</td>
<td>Brackish-water forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fresh water</td>
<td>pea</td>
<td>Peat swamp forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wet perennially</td>
<td></td>
<td></td>
<td>Freshwater swamp forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wet periodically</td>
<td></td>
<td></td>
<td>Seasonal swamp forest</td>
</tr>
<tr>
<td></td>
<td>Wet–dry</td>
<td>Moderate dry season</td>
<td>Inland</td>
<td></td>
<td>Most deciduous forest</td>
</tr>
<tr>
<td></td>
<td>Increasing seasonal deficit</td>
<td></td>
<td></td>
<td></td>
<td>Dry deciduous forest</td>
</tr>
</tbody>
</table>

* From Whitmore (1975), based on Steenis (1950).

Precipitation and slight seasonal temperature variations. The result is a "tropical seasonal evergreen forest" which is still tall and very diverse but in which pheno-
logical functions become synchronized and more sensi-
tive plant types become somewhat less abundant, such as epiphytes with only aerial roots. Further from the equator, as the tropical dry season begins to become more pronounced, the forest becomes a "tropical semi-
evergreen forest" containing some deciduous trees and
still fewer epiphytes (Walter, 1985). In most tropical
trees, deciduousness appears to be a facultative charac-
teristic. This is demonstrated by two facts:

1. Many tropical tree species are evergreen in one part of their range but lose their leaves in a drier part.
2. Many individual tropical trees have been ob-
served to keep almost all their leaves in wetter years
and lose most or all of their leaves in drier years.

Temperate tree species, on the other hand, often die
within a few years if planted in the tropics, apparently
due to lack of the seasonal cues necessary for vernal-
ization.

Tropical rain forest has its largest extent in Asia
from peninsular Malaysia through the East Indies and

TABLE IX

<table>
<thead>
<tr>
<th>Belt</th>
<th>Elevation (m)</th>
<th>Floristic zone</th>
<th>Main taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper montane</td>
<td>1500–2100</td>
<td>Montane ericaceous</td>
<td>Ericaceae, Myrtaceae, conifers</td>
</tr>
<tr>
<td>Lower montane</td>
<td>1200–1800</td>
<td>Oak-laurel</td>
<td>Fagaceae, Lauraceae</td>
</tr>
<tr>
<td></td>
<td>760–1200</td>
<td>Upper dipterocarp</td>
<td>Shorea, Dipterocarp</td>
</tr>
<tr>
<td></td>
<td>300–790</td>
<td>Hill dipterocarp</td>
<td>Lowland + Shorea curtissi</td>
</tr>
<tr>
<td></td>
<td>0–360</td>
<td>Lowland dipterocarp</td>
<td>Many dipterocarp, especially Dipterocarpus,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Shorea, Dryobalanops aromatica</td>
</tr>
</tbody>
</table>

* From Whitmore (1975; 1984) for mountains of Malaya, modified from Symington (1943). The oak-laurel
zone corresponds closely to subtropical/warm-temperate laurel forests further north (Olson, 1995).
southern Philippines. It can also be found, however, along the mainland coast from peninsular Thailand and Myanmar to Bangladesh, in Kerala and southwestern Sri Lanka, and northward through the mountains of Southeast Asia, Myanmar, and Assam to southeastern China (Yunnan) and the foothills of the eastern Himalaya (Whitmore, 1984). In these more northerly areas, extending beyond the Tropic of Cancer, the rain forest takes on a more subtropical character but maintains high diversity and its multi-layered structure, with abundant epiphytes and lianas.

Tropical rain forest in general is composed of many tree species from families such as the Lauraceae, Rubiaceae, Moraceae, Meliaceae, and Leguminosae, the latter two usually with compound leaves and somewhat smaller leaflets than the typical rain forest leaf. Tropical rain forest reaches its greatest species richness in Southeast Asia, but here another family, the Dipterocarpaceae, provides more of the rain forest species than does any other family elsewhere. Some of the most important tree genera in Malaysia, Indonesia, and adjacent mainland areas include Dipterocarpus, Shorea, Hopea, Dryobalanops, and Vatica (Dipterocarpaceae); Mitragnya and other Rubiaceae; trees with large buttress roots such as Heritiera (Sterculiaceae); Aplasia and Dysopyrum (Meliaceae); Mangifera and other anacards; Ficus and Artocarpus (Moraceae); other trees with large edible fruits such as Durio (Bombacaceae); Kosmopsis and other legumes; and other families such as Dilleniaceae, Flacourtiaceae, and the Hymelieae. Most genera are mainly tropical. The higher understories contain trees of the same physiognomy but also various palms (e.g., Licuala), even climbing palms (rattans), and other climbers such as stranglers (especially Ficus spp.). The ground layer is largely composed of frutetuous but entirely herbaceous monocots such as Heliconia and Alpinia (both Zingiberi-
dae). Tall bamboos, of genera such as Calamus and Phyllostachys, may form colonial stands where the forest has been disturbed frequently and opened, especially in mountains. Secondary vegetation may also involve larger-leaved, deciduous, fast-growing arborescents from the characteristic pioneer genera Macaranga and Mallotus (analogous to Cecropia in the tropical Ameri-
cas). Some other characteristic secondary genera in-
clude Melastoma, Glochidion (Euphorbiaceae), and Toma (Ulmaceae).

Birds, amphibians such as tree frogs, reptiles such as tree-dwelling lizards, and countless insects are the main groups of animals in tropical rain forests. Mammals are less numerous but include both larger ground-dwellers such as rodents and flying mammals (i.e., bats). Most animals are probably insectivorous, although many are also mixed feeders, including various mammals which range between the canopy and the ground. Some larger ground animals are herbivores, eating mainly plant roots and fallen fruit. Ants are an especially interesting and important group for their role in dispersing plant diaspores and in decomposition and nutrient cycling.

B. Monsoon Forest

(Moist Deciduous Forest)

Monsoon forest is not always a well-defined term but is used most commonly for the tall tropical "moist de-
ciduous forest" with fairly large, soft (malacophyllous) deciduous leaves which occur in areas with essentially perfumid rain forest conditions during a wet season which is longer than the dry season. Since the "rain-
green" leaves wilt when soil water is depleted (about a month after the end of the rainy season), there are generally no bright colors in the tropical "autumn" (with the exception of Excocarpus agallocha in man-
groves). In the tropical "spring," however, many trees bloom brightly in the month before the rainy season, cued by increasing temperatures, and the new leaves of many species are also bright reddish before enlarging and receiving their full dose of chlorophyll. Moist decid-
uous forest is not as species-rich as the rain forest but may still reach canopy heights of 30 m and have more species than temperate forests.

Most deciduous forest occurs in the Western Ghats (Kerala) and in those parts of lowland and mountainous eastern India, Myanmar, and Southeast Asia which have relatively short dry seasons (Champion and Seth, 1968). In the Western Ghats, one of the most important trees is teak (Tectona grandis), which grows very straight but has unusually large, soft leaves very subject to insect herbivory. Other important trees include Shorea spp., Terminalia tomentosa, Lagerstroemia lanceolata, Grewia tiliaefolia, and Pterocarpus marsupium. To the east one encounters sal (Shorea robusta), a dipterocarp and an-
other of the most important forest trees of tropical Asia, economically and ecologically. Moist deciduous forests with Shorea continue eastward from the Eastern Ghats through Assam and Myanmar into parts of Southeast Asia. The composition and biogeography of the moist as well as dry deciduous forests of India are well de-
scribed by Mani (1974), who provides photographs of various forest areas which are now probably gone. Other moist deciduous forests occur in scattered mountain areas of Sri Lanka and the East Indies but have been largely destroyed over most of their area.
C. Tropical Dry Forest, Scrub, and Savanna

‘Dry deciduous forest’ occurs where the dry season is at least as long as the wet season and involves shorter, somewhat more open forests dominated by trees with usually smaller leaves, including many Leguminosae. Species richness and leaf area are lower than in monsoon forest or rain forest, and canopy height is usually not more than about 13 m. In still drier areas, dry forest may grade into dense but shorter thorn-scrub, dominated by sprawling, thorny arborescents with sparse foliage of quite small leaves (or leaflets), many of which may be Leguminosae. In other drier areas, especially on relatively flat terrain, trees may be only widely scattered in savannas dominated by tall to short seasonal grasses. In most areas, all three of these vegetation structures may be interspersed in mosaic landscapes, with forest now usually replaced by more extensive areas of degradation stages: thorn-scrub, sclerophyll scrub, and derived savanna. This kind of rain-green mosaic landscape occurs over much of the Deccan Plateau and Ganga Valley, in the ‘dry belt’ of interior Myanmar, and in much of interior Southeast Asia. It also occurs on the drier northern to eastern side of Sri Lanka and in drier parts of the northern Philippines and the East Indies, especially in the Lesser Sunda Islands of eastern Indonesia. An interesting variant of this rain-green response to the tropical dry season is represented by the somewhat open ‘dry evergreen forest’ of interior Thailand.

The gradual transition between moist and dry deciduous forest occurs in particular across the Deccan Plateau, from moist deciduous forest in the Western Ghats to thorn-scrub and savanna in the rain shadow immediately eastward and to dry deciduous forest, and again to moist deciduous forest further eastward. Teak and sal still occur in the dry forest, but trees such as Dacryodes melanoxylon (ebony), Anogeissus latifolia (Combretaceae), leguminous taxa such as Acacia (especially A. catechu) and Cassia, and low arborescents such as Zizyphus species become more important. This is also the home of the evergreen but semi-parasitic sandalwood tree (Santalum album). Further east, in interior Southeast Asia, the main type of dry deciduous forest is ‘dry dipterocarp forest,’ reaching over 20 m in height and with canopy species such as Dipterocarpus tuberculatus, D. intricatus, Shorea obtusa, and Shorea harryana. In addition, dry evergreen forest dominated by evergreen dipterocarps such as Hopea ferrea and Shorea heryana covers significant areas, especially in Thailand.

D. Humid Tropical and Subtropical Mountains

The main truly tropical or subtropical humid mountain areas in Asia are the lower slopes of the Himalaya (south side), the Western and Eastern Ghats in peninsular India, the mountains of Southeast Asia, and the smaller ranges and isolated volcanic peaks of the East Indies, the Philippines, and Sri Lanka. A general zonation scheme for the vegetation of mountains in tropical Asia was shown in Table IX. As one ascends any mountain range, two climatic changes occur which together quickly reduce the effect of any dry season which may exist in the lowlands.

1. Temperature decreases (approximately 6°C per kilometer), reducing potential water loss to evapotranspiration.
2. Humidity and precipitation increase, at least up to the quasi-permanent cloud belt, which is usually 1000–1500 m above the base of the mountains (lower in more humid areas).

Thus, a tropical dry season in the surrounding lowland may disappear completely with a climb of 1000 m or less. At some elevation, generally between 800 and 1200 m, the lowland forest is replaced by a belt of tropical montane forest reminiscent of the evergreen laurel forest of the warm-temperate zone and dominated by temperate taxa such as Ericaceae and Ilex, plus subtropical taxa such as Myrsinaceae, Symplocaceae, Rubiaceae, and understory Lauraceae and Theaceae. From the lowland forest to this montane forest, leaf size typically changes from normal mesophyll to a mix of smaller notophyll and microphyll (Ohsawa, 1995). Above this there may be a low, dense, distinctly microphyll cloud forest of laurophyll treelets covered by mosses and other epiphytes. Of the truly tropical Asian mountains (i.e., excluding the upper parts of the Himalaya), however, only Mt. Kinabalu is high enough for an alpine belt and it is only rocky (old volcanic plug) at its top. Mount Kinabalu (+4110 m, northern Borneo) is the highest mountain between the Himalaya and New Guinea (geologically part of Australia). Due to its location near the equator, Mt. Kinabalu contains taxa from both hemispheres, including some distinctly Australian elements such as Leptospermum and Platycladus. Due to the discontinuous nature of the Southeast Asian landmass and its fluctuating biogeographic history, Mt. Kinabalu also represents a true biogeographic ‘island in the sky’ and one of the most interesting goals of any knowledgeable geobotanist. The lower slopes are cov-
ered by lowland and montane (subtropical) rain forest, followed by laurel forest rich in Syzygium and other amphi-tropical taxa. Near the probably edaphic treeline (around 3300 m), on shallow soil, is a belt of low, open forest dominated by Leptospermum, with extremely small (leptophyll) leaves reminiscent of Australian forms and the Kan conversion of New Zealand. Other interesting features include 30-cm-wide, brown, foul-smelling, fly-pollinated flowers of Rafflesia and the diversity of Nepenthaceae species, with their insectivorous pitchers hanging on vines.

The mountains of mainland Southeast Asia are less described but generally show the same vertical zonation as described elsewhere (Whitmore, 1984). Evergreen Castanopsis, Lithocarpus, and Cyclobalanopsis (evergreen Quercus), dominants in the laurel forests of East Asia, extend far into the tropics in Southeast Asia to form extensive montane forests. Satellite data have suggested that some relatively large areas of montane rain forest may remain.

In the Western Ghats of Kerala, mostly deciduous forest forms a lower montane forest belt, followed upward by semi-evergreen transition and then evergreen laurel cloud forest. The same zonation probably occurred on other mountains of southern India and Sri Lanka but has been largely destroyed, converted especially for tea plantations. A similar zonation occurs on the lower southern slopes of the eastern Himalaya but involves more subtropical species.

V. DRY REGIONS OF WEST AND INTERIOR ASIA

The dry regions of Asia include the subtropical deserts of southwestern Asia, the adjacent region of Mediterraneanto-type ecosystems from Turkey to Israel, the Asian portion of the temperate grassland corridor from the Ukraine to Mongolia and northern China, and the semi-desert and desert regions of both Middle Asia (Turkestan) and Central Asia (Gobi and northwest China as well as the high, dry Tibetan Plateau). Major mountain ranges separate some of these and adjacent regions, especially the Himalaya, which blocks moist monsoonal air from the south, and the Tien Shan system, which blocks moist air from the Mediterranean and thus separates summer-dry Middle Asia from summer-rain Central Asia. Most of these dry climates are also continental, reaching coastal areas only in the Middle East. The Mediterranean region contains landscapes ranging from small areas of sclerophyll forest and montane conifer forest to sclerophyll shrublands and dwarf-scrub. The temperate grasslands range from tall-grass prairies and meadow-steppes in the north to drier, open steppe in the south, following the north–south moisture gradient which obtains across most of the continent. These steppe eventually grade southward into the temperate semi-deserts and deserts.

A. Subtropical Deserts

Deserts (sensu stricto) have essentially no vegetation, whereas semi-deserts have scattered plants. The subtropical and zone of Asia contains large areas of extreme desert: the Rub al-Khali (empty quarter) of the southwestern Arabian Peninsula, some areas of the Syrian desert (in Iraq and northern Arabia, not Syria), and the Dasht-e-Kavir and smaller areas of interior Iran. The rest of the subtropical desert area, from the Middle East to westernmost India, is mostly semi-desert, with widely scattered xeromorphic dwarf-shrubs, desert grasses, and ephemeral herbs on coarse, often gravelly or even rubbed substrates.

Desert conditions may have existed since the earlier Paleozoic. Whether the desert flora arose this early, or in the Miocene (expanding during the colder, drier Pliocene and Pleistocene), the desert flora of Asia has probably been closely related to that of northern Africa for a long time. Although total species richness is low and perhaps because there are essentially no true native stem-succulents, the vegetation of the Middle Eastern deserts shows a wide variety of adaptations of form and seasonality to avoid dehydration (Walter, 1985; Zohary, 1973). Some dwarf-shrubs are evergreen but greatly reduce their leaf area at the beginning of the dry season (e.g., Artemisia monosperma, Reaumuria palestina, Salsolea villosa, Stara pedra, and Zygophyllum dumosum). Some are wintergreen shrubs, losing their leaves in summer (e.g., Lycium arabicum and Anagyris foriata), whereas some (generally dwarf-shrubs) shed their leaves but retain evergreen stems, as in Retama and Calligonum species. Some have grayish, soft, deciduous leaves but keep smaller leaves during the summer (e.g., Artemisia herba-alba). Some are totally leafless, such as Ephedra. Perhaps most interesting are the leafless shrubs with modular stems and green, photosynthetic but disposable bark segments (e.g., Haloxylon persicum and Anabasis articulata). Many of these desert dwarf-shrubs extend well into the region of cold-winter continental deserts, in Middle and even Central Asia. Where more water is available, there may be scattered trees such as acacias, deciduous thorn-shrubs...
Hejaz contains dry juniper woodlands (southern and central Iran. The montane belt of the desert zone are the Hejaz (Yemen and southwestern season, migrating to northern Eurasia in summer. be dependent on desert ecosystems during the cooler night, as well as ectothermic lizards and other reptiles. living in underground burrows and foraging only at supplied to their brains. There are also many rodents, pant to moisten nasal sinuses, thus cooling the blood supplied to their brains. There are also many rodents, living in underground burrows and foraging only at nigh, as well as ectothermic lizards and other reptiles. Most birds spend less time in deserts but may still be dependent on desert ecosystems during the cooler season, migrating to northern Eurasia in summer.

The two main mountain ranges in the subtropical desert zone are the Hejaz (Yemen and southwestern Saudi Arabia) and the Zagros and other mountains of southern and central Iran. The montane belt of the Hejaz contains dry juniper woodlands (Juniperus).

B. Mediterranean Ecosystems

Mediterranean-type climates have cool winters with some precipitation and long, dry summers with no significant precipitation perhaps for several consecutive months. As a result, the primary growing season is restricted to springtime, when temperatures are increasing and soil water is readily available. Plant activity may continue into the summer as long as root systems can still reach deeper soil water. A shorter, secondary growing season often occurs in the autumn, when west-erlies again bring some precipitation but before temper-atures drop too low. The most characteristic plant adap-tation to this climate is woody vegetation with deep root systems and hard (sclerophyll) evergreen leaves, often with waxy cuticles and/or oily secondary com-pounds which retard water loss from exposed surfaces. This type of 'mediterranean sclerophyll' vegetation oc-curs in all five of the world's mediterranean-climate regions and represents a striking example of convergent evolution. The mediterranean region of Eurasia lies mainly between southern Europe and North Africa but extends eastward along the coastal rim of western and southern Turkey, to Cyprus, and eastward from Syria, Lebanon and Israel to more continental Iraq, the south-ern Caucasus region, and Iran.

Three general types of mediterranean landscapes can be recognized, differentiated primarily by total water availability. Where sufficient soil water can be reached throughout the year, forests composed of sclerophyll oaks such as Quercus ilex covered the Mediterranean borderlands (more in the more humid west) before they were cut by the ancient Romans to build their navies. Where less but still some water is available through the summer, perhaps at greater depth, the landscape is dominated by the characteristic dense sclerophyll woodlands and scrub denoted by their French name maquis. Where soil water is not available through the summer, more open stands of smaller, summer-decidu-ous dwarf shrubs dominate landscapes generally called garrigues (which may be patchy and may include some sclerophylls). Other plant types, however, also occur in mediterranean ecosystems on permanently more moist sites, including more water-demanding, evergreen laurel, myrtle, oleander, and even soft-leaved, winter-deciduous shrubs.

The Mediterranean region of Eurasia has been settled and modified by humans for several millennia, with consequent massive soil erosion and other degradation. As soil was lost, both soil water-holding capacity and local precipitation were reduced, resulting in replace-ment of forest and woodland by various types of gar-rigues in typical degradation stages.

In Mediterranean mountains, where winter tempera-tures may become too cold for sclerophyll, dry forests of pines are common, especially in Turkey. Cypresses (Cupressus spp.) and true cedars (Cedrus) may also occur, the latter occurring only in the mediterranean mountains of Lebanon (C. libani) and Cyprus (C. brevi-folia), in the Atlas Mountains of northwestern Africa (C. atlantica), and on the drier western part of the Hima-laya (C. deodara). Summergreen trees may also appear in mediterranean mountains and may form deciduous forests in sub-mediterranean borderlands (mainly in Europe), in continental areas with colder winters (such as the southern Caucasus and northern Iran), and on particular biotopes such as continuously moist flood-plains.

Many of the familiar mediterranean plants of the more humid western Mediterranean do not extend into the Asian region to the east. In particular, the stem-evergreen 'broom' form of the western Mediterranean and Atlantic Europe (genera Genista, Cytisus, etc.) ap-pears not to be important in southwestern Asia. Some species, however, extend throughout the western and eastern Mediterranean areas, such as Quercus coccifera, Ceratonia siliqua, Juniperus oxycedrus, Myrtus communis, and Pistacia lentiscus. Others are replaced to the east
by similar species, such as Quercus calliprinos, Qu. tiba-
burmensis, Pistacia palestina, and Pinus brutia. On the
other hand, some species widespread throughout the
Mediterranean region may have come originally from
the east, such as Aleppo pine (Pinus halepensis). The
vegetation of Mediterranean Asia is well described by
Zohary (1973, pp. 130-165).
C. Temperate Grasslands
In temperate Asia, especially in the west but also in the
east, precipitation decreases southward while tempera-
tures increase. As a result, climatic dryness increases
southward and the temperate grasslands of Asia occur
primarily in a rather narrow corridor that extends east–
west across the continent, from the Black Sea region to
northeastern China, between the boreal forest to the
north and drier semi-desert and desert areas to the
south. The Russian-derived term “steppe” normally
means a short, discontinuous grassland in English-lan-
guage ecology but is traditionally used in Asia for all
grasslands. The Asian grassland zone can be divided
as follows:

1. A discontinuous strip of “forest steppe” in the
north (mainly Populus and Betula), analogous to the
aspen “grove belt” of North America
2. A zone of tall-grass “meadow steppe” with many
colourful forb species, analogous to the tall-grass prairies
of North America
3. A zone of shorter, more open steppe, less devel-
oped north of the Black Sea but widening east of the
Caspian and extending southward toward the desert
areas of Turkestan and the Gobi

This zonation, its soils, and steppe ecosystems have
been well described in the Russian ecological literature,
much of which is essentially unknown in the West.
In China also, scientific study of the grasslands (and other
vegetation regions) goes back well before the Commu-
nist period but is unknown in the West.

The Eurasian steppe corridor is broken, largely by the
Altai Mountains, into the Russian–Siberian steppes
and the grasslands of Mongolia and northern China
(Lavrenko as cited in Walter, 1974, pp. 163–165). In
the western section, the steppe is composed especially
of Striga species, such as Striga capillata, S. lesingiana, S.
sterophylla, and other important grasses such as Festuca
sulcata, Koeleria gracilis, Dactylis glomerata, Poa botry-
oides and other Poa spp., various Agropyron spp., Bro-
mus inermis, and species of Carex. Especially character-
istic of the western steppe are the many small geophytes
(“epheroides”; Walter, 1974) from genera such as Tu-
lipa, Crocus, Ornithogalum, Bulbocodium, and Hyacin-
thella as well as Poa bulbosa. Tall forbs in the meadow-
steppe include Anemone patens, Adonis vernalis, Salvia
nuttans, S. superba, and Delphinium confluens. These are
accompanied by some shorter shrubs and semishrubs,
including the genera Sisyrinchium, Carex, and Artemisia.
The main tree species in the forest-steppe, a wooded
tall-grass prairie, are Quercus robur, Betula verrucosa,
and Populus tremula.

In the mountains of Middle Asia, montane steppes
occur below 2000 m and taller meadow-like grasslands
above, including a park-like forest-steppe transition.
These may involve grasses such as Alopecurus pratensis
and D. glomerata and tall forbs, such as Delphinium
confluens, Ligularia altaica, Aquilegia sibirica, and Scab-
iosa alpina.

The grasslands of Central Asia, east of the Altai
Mountains, involve mainly the Mongolian steppe and
the grasslands of Manchuria. In the Mongolian steppe
the most important grasses are Stipa species such as S.
kyrylovii, and Aneurolepidium chinense, plus K. cristata,
F. sulcata, and P. botryoides from the west. These are
joined by xeric shrubs such as Caragana species and
Artemisia frigida. The Manchurian grassland is largely a
meadow-steppe dominated by Aneurolepidium chinense,
along with A. pseudo-agropyron and Stipa baikalensis,
plus forbs such as Tanacetum sibiricum (Compositae)
and xeric shrubs such as Artemisia sibirica (Walter,
1974). On the lower slopes of the southern Hinggan
Mountains is a forest-steppe transition to montane for-
est of Larix dahurica, Betula platyphylla, and more tem-
perate forests of Quercus mongolica, Tilia mongolica,
and Pinus tabuliformis.

The development of the world's temperate grasslands
is closely linked with populations of large ungulates,
such as the well-known bison of North America. In
Europe some of the larger animals, most highly endan-
gered, include Przewalski's horse (Equus przewalskii),
the Saiga antelope (Saiga tatarica), the Zheiran gazelle
(Gazella subgutturosa), and the wild camel (Camelus
ferus). The ranges of these and many smaller grassland
animals also extended into the adjoining semi-desert
and desert zones. Smaller animals include many ro-
dents, such as jereboas, gerbils, and small hampsters.

D. Temperate Semi-Deserts and Deserts
The temperate deserts and semi-deserts of Asia, with
cold winters, occur south of the temperate grasslands
in three general areas (Walter and Box, 1983):
1. Middle Asian Deserts

A characteristic of the sandy deserts of Middle Asia is the dominance of both small, typical and much larger xeromorphic shrubs, some leafless and some at least partly evergreen. Reaching 2 m in height, leafless Artemisia pauciflora, A. terre-albae, Halostachys caspica, Anabasis salsa, Halocnenum strobilaceum, and Kochia prostrata, along with steppe and desert grasses such as Festuca salicata, Stipa capillata, Koeleria cristata, and Agropyron repens. In areas of flat terrain and finer soil, mini-geophytes such as Poa bulbosa form vernal carpets of miniature flowers (Walter, 1974, p. 234).

2. Tangaraxia, Gobi, and Tarim Basin

Dzungaria is the region east of the Dzungarian Gate, a gap in the Tien Shan–Altaï mountain system, and extends across northern Xinjiang. This corridor between Middle Asia and the Gobi, as well as the Gobi itself, represents a region of rockier substrates and deserts dominated by typical desert shrubs such as Artemisia, Anabasis, and Calligonum. The most important grasses are Stipa-bunch-grasses. Isolated areas of Populus diversifolia, Tamarix ramosissima, and occasionally H. ammodendron may occur along streams. The Takla Makan (Tarim Basin), on the other hand, is a large sand sea within the horseshoe (open to the east) formed by the Tien Shan on the north, the Pamirs in the center, and the Karakoram and Kunlun ranges on the south. Together they form one of the most effective rain shadows in the world, and precipitation has very rarely ever been recorded in the Tarim Basin. There is little vegetation on the shifting sand areas, but runoff from the mountains does generate intermittent and permanent streams which extend into the basin. Along these streams one can find riparian strips of the phreatophytic Tamarix ramosissima as well as the widespread Populus diversifolia and Ulmus pumila. In some areas there are even extensive marshes of Phragmites, Typha, and Scirpus species, which are especially important for the large numbers of birds which overwinter in India and migrate to Siberia for the summer.

3. Tibetan Plateau

The high, dry Tibetan Plateau (average elevation approximately 4000 m) is a region of dry steppes, alpine tundra, and high-altitude desertiferous vegetation. Most familiar genera from Central Asia (and Middle Asia) are represented, including Caragana, Reaumuria, Salsola, Atriplex, Artemisia, Phragmites, and Festuca. Some familiar species also extend into the Tibetan highlands, including Erotes erato and Kochia prostrata. East Asian elements are important only in the east but include Gentiana, Primula, Salsifraga, Saxifraga, and Rhododendron. Holarctic elements are few, but some are very abundant, especially in the alpine mats and marshes: Kobresia, Carex, Eriochloa, Eriophorum, and Juncus. Where mois-
VI. HUMID MONSOON ASIA
(EXTRATROPICAL)

Monsoon Asia is the region dominated by the seasonally alternating monsoon wind system, which brings wet oceanic air masses onshore in the summer and blows dry and cold outward from the continental interior in winter (mainly eastward and southward). The region lies east of the Altai–Tien Shan mountains and extends from southern Siberia, through China, to Korea and Japan in the east and to tropical Asia in the south. In most of this region the only major mountains run east–west, so mean winter temperatures are low but outbreaks of even colder Siberian air are largely excluded. Except toward the interior, summer brings adequate rainfall throughout for forest growth. In fact, in Japan and some parts of eastern China, there is rarely any climatic drought and forests can be very mesomorphic. Humid monsoon Asia (i.e., forested East Asia) can be divided into two regions: (i) temperate deciduous forest in the typical-temperate and cool-temperate bioclimatic zone of evergreen broad-leaved forests and (ii) evergreen broad-leaved “lauraceous forest in the typical-temperate and cool-temperate areas to the south. Pine forests are important secondary landscapes in both areas. Mountains in East Asia are especially interesting for the rich variety of endemic conifer taxa which they contain.

A. Laurel Forest (Evergreen Broad-Leaved Forest)

Laurel forests are evergreen broad-leaved forests of warm-temperate and humid-subtropical climates, dominated by trees with intermediate-sized, dark green (shade-tolerant), thin-coriaceous but mesomorphic leaves (laurophylls) such as are especially characteristic of the laurel family (Lauraceae). Forests dominated by laurophyll trees are rather dark and somber, with low light levels below the canopy, and are evergreen from top to bottom. In some respects laurel forests can be thought of as an extension of the tropical rain forests into the warm-temperate zone. East Asia contains the world’s largest area of such forests, due to its abundant rainfall and winters without severely low temperatures, at least in coastal areas or on islands. Laurel forests occur generally south of about 35° N latitude but extend to 38° N on both sides of Japan. Counterparts occur in southern Brazil and small areas of eastern Australia, as well as moist depressions (“bay forests”) in the southeastern United States. Cool-temperate analogs occur in New Zealand and Tasmania.

In East Asia, laurel forests occur in southeastern China, in drier southwestern China (Sichuan and Yunnan, with different but largely vicariant species), in southern Japan, and in a small strip across southernmost Korea (Wang, 1961). In Southeast Asia, laurel forests also ascend into the mountains to form montane forests just above the tropical seasonal evergreen forests of the lowlands (Whitmore, 1984; Ohsawa, 1995). The main canopy tree genera of laurel forest are Persea (= Machilus), Cinnamomum, Brachychiton, etc. (Lauraceae); Castanopsis, evergreen Quercus (= Cyclobalanopsis), and Lithocarpus (Fagaceae); Schima (Theaceae), Ilex (Aquifoliaceae), Michelia (Magnoliaceae), and others, all with very similar laurophyll physiognomy. Understory trees and arborescents are largely from the same families. Lauraceae tend to be especially important in more perhumid (e.g., coastal) areas, whereas evergreen Fagaceae tend to become more important inland, especially in southwestern China.

In humid East Asia, the largest species turnover between the polar region and the tropics occurs within the bioclimatic zone of evergreen broad-leaved forests—without major change in the forest physiognomy. This occurs in the Okinawa Islands of Japan and in southeastern China, as temperate species abruptly disappear, including many laurophyll tree species, and are replaced by an essentially tropical flora which includes many new tree genera with essentially the same evergreen broad-leaved structure.

B. Temperate Deciduous Forest

The temperate deciduous forests of East Asia cover the northern half of Japan (including Hokkaido), most of Korea, and the northeastern part of China, especially Manchuria. This region is smaller than the region of evergreen broad-leaved forest and smaller than the deciduous forest region in eastern North America. Field estimates of primary productivity in Japan suggested that the deciduous forests were less productive than in North America and less productive than some montane conifer forests of Japan. Nevertheless, the deciduous forests of East Asia appear to have more species and genera, including almost all genera found in eastern...
The deciduous forests of Japan are especially rich, due perhaps to the perhumid maritime climate and total lack of water stress during the growing season. Deciduous forests cover most of northern Honshu, dominated mainly by Fagus crenata but with many other taxa, such as Fraxinus lanuginosa, Tilia japonica, and Magnolia obovata. Understories are rich with maples (e.g., Acer mono, A. japonicum, and A. palmatum) and other mesic understory elements such as Styrax obassia, Sorbus alnifolia, Cercidiphyllum japonicum, Aesculus turbinata, Callicarpa japonica, Hamamelis obtusata, Viburnum, and Euonymus. One interesting transitional forest type occurs in parts of northeastern Honshu and involves co-dominance by F. crenata and Abies firma, a temperate-zone fir. Forests of Fagus japonica generally occur only in the areas of transition to warm-temperate forest. An especially characteristic feature of the deciduous Fagus forests of Japan is the prevalence of short but broad-leaved bamboos of the genus Sasa (and recently also Sasamorpha), which form dense understories 0.5–1 m high where there is continuous snow cover. Other common herb-layer taxa include Disporum, Viola, Chimonophila, Mitchella, Carex, and ferns. Floodplain and other especially moist forests are dominated by ashes (Fraxinus sieboldiana and F. spathulata), elms (e.g., Ulmus davidiana), Pterocarya rhoifolia, and other typical floodplain taxa (e.g., Alnus hirsuta). Also belonging to the temperate zone are the hemlock (Tsuga sieboldii) forests which occur most commonly on rocky substrates at the base of lower mountain slopes. Drier areas (still without water deficits) are generally covered by deciduous oak forests, including interior areas, lower mountains (e.g., Qu serrata), and most of Hokkaido (Qu mongolica var. grosseserrata).

The deciduous forests of Japan occur in a distinctly more continental climate, in which oaks (e.g., Qu mongolica), and also Carpinus, become more important. Much of the Korean landscape is still recovering from the Korean War, after which the southern part was seeded with North American pines. In deciduous forest areas of both Japan and Korea, as in the evergreen forest region, native pines such as Pinus densiflora play a very important role in successional landscapes.

2. Manchuria and Eastern China

The forests of northeastern China and the Amur valley have a shorter growing season than in Japan but are still very rich. Canopy trees may include such species as Qu mongolica, Fraxinus mandshurica, Tilia amurensis and T. mandshurica, Ulmus macrocarpa and U. davidiana, Phellodendron amurense, Maackia amurensis, Juglans mandshurica, Acer mono, Betula dahurica and B. costata, and Populus koreana and P. ussuriensis. A common temperate conifer in these forests is Pinus koraiensis. Understories may include trees of mainly the same species plus shrubs such as Corylus heterophylla, Deutzia amurensis, Euonymus pauciflora, Philadelphus schrenkii, Syringa amurensis, and Viburnum spp.

Further south, in northern, eastern, and into southern China, the deciduous and evergreen broad-leaved forests are almost totally destroyed by millennia of human habitation. As a result, it is difficult even for local scientists to identify the boundaries of the two forest regions. In hill and mountain areas of northern China (Liaoning, Shandong, Hebei, and Beijing) remaining mature forest types are dominated by Quercus mongolica, Qu. hupehensis, and other Quercus species along with Celtis, Fraxinus, Tilia, Betula, and Ulmus as subordinate canopy elements. Understories may have elements such as Lindera obtusiloba, Corylus heterophylla, Lespedeza bicolor, and Spiraea trilobata. Betula and Populus forests occur as secondary types. Quercus variabilis forests occur as transitional forests south of the main deciduous forest region.

Among the most interesting deciduous forests are those involving various species of beech (Fagus), generally covering small areas at elevations from 1300 to 2200 m in central eastern China (Zhejiang to Guizhou provinces). Deciduous trees in interior southern China may have only short leafless periods. This is also one of very few areas in the world in which deciduous and evergreen broad-leaved trees coexist in semi-evergreen broad-leaved forests.

C. Pine Forests

Pine forests represent extensive landscapes in East Asia due to long histories of continuing forest cutting. Until after World War II, many of the evergreen broad-leaved forest areas were managed as coppice forests. Other areas, however, were burned and cleared much earlier. In hilly and mountainous terrain, soil erosion followed clearing and broad-leaved forests could no longer recover. These areas usually became pine forest and were maintained by intermittent natural and induced burning of the understorey. Much of southern Japan, the Korean peninsula, and upland warm-temperate and subtropical China are now covered by such pine forests.

In Japan, Korea, and some parts of northeastern China, these secondary pine forests are mainly domi-
nated by *Pinus densiflora*. In eastern and southeastern China, pines with more warm-temperate to subtropical distributions become dominant, especially *P. massoniana* and *P. tabuliformis*. In southwestern China, forests of *P. armandii* occur in Sichuan and on the Yunnan plateau, whereas subtropical pine forests of *P. kesiya* (cf. *P. hians*) become important in the mountainous transition southward into lowland Xishuangbanna and the mountains of northern Thailand. In addition to pine, the warm-temperate to subtropical "Chinese fir" (*Cunninghamia lanceolata*) is also important in secondary landscapes and, like pine, is widely planted.

### D. Mountains of Humid East Asia

The mountains of East Asia represent the only large mountainous area in the warm-temperate and humid subtropical zones of either hemisphere (although smaller areas do occur in Mexico, southern Brazil, eastern Australia, and northern New Zealand). The various mountain systems of humid East Asia, including at least the more humid eastern Himalaya, thus represent areas of unusual diversity and endemism, especially among what appear to be relict conifers. Endemic conifer genera include *Amentotaxus*, *Cathaya*, *Cephalotaxus*, *Keteleeria*, *Cunninghania*, *Fokienia*, *Glyptostrobus*, *Sciadopitys*, *Taiwania*, and *Thuja*, at least some of which once occurred more widely. In addition, more widespread conifer genera such as *Abies*, *Chamaecyparis*, *Pinus*, *Taxus*, *Thuja*, and *Tsuga* have relatively high species diversity in these areas. In most of the area, the lowland biome is evergreen broad-leaved forest, which may extend well into the montane belt in most places, with well-developed Rhododendron belts and a more warm-temperate character. As winter becomes colder upward, summergreen deciduous forest may occur, but the forest may also grade directly into evergreen high-mountain conifer forest. None of these warmer mountain systems is high enough to have an alpine belt except the Himalaya.

In the western Himalaya, montane conifer forests are dominated by *Pinus roxburghiana* and the deodar cedar (*Cedrus deodara*), a preferred ornamental tree throughout the world. To the northwest, in the Tien Shan and Altai systems, the montane conifer forests are simpler, dominated often by *Picea schrenkiana* but including forests of shade-tolerant *Pinus sibirica* as well as boreal *Picea* and *Abies* larches, and *Pinus* sylvestris (Walter, 1974; Walter and Box, 1983, Chap. 7). In western Tibet, the main conifers are *Abies delavayi* and *A. webbia*, along with *Picea likiangensis*. This is also the area of the dawn redwood (*Metasequoia glyptostroboides*), first identified from fossils, only discovered in the wild in the early 1900s but now widely planted as a street tree throughout China.

Extensive evergreen oak forests, involving *Quercus semicarpifolia* and many other species, still occur on the southern slopes of the Himalaya and may represent a somewhat unique warm-temperate forest biome. At higher elevations, *Betula albosinensis*, *Acer tetramerum*, and many other deciduous species form belts of summergreen forest in some areas. Subalpine conifer forests may contain species such as *Picea likiangensis* (better known from Manchuria), *Abies squamata* and *A. georgei*, *Tsuga dumbosa*, and even the deciduous *Larix pendula*. The best developed Rhododendron belts are in the eastern Himalaya.

On the lower parts of the eastern Tibetan Plateau, *P. likiangensis*, *A. squamata*, and *A. georgei* form tall coniferous forests, followed upward by *Abies faxoniana*, *Pinus asperata*, and *P. novicii* (Archibold, 1995). *Larix* may be admixed, and *Sabinia* and *Capsaspora* form shorter woodlands in drier areas. Toward the Yunnan Plateau to the southeast, *Tsuga yunnanensis* becomes important, with *Picea brachytyla* and *Abies fabri* at higher elevation. As elevation increases, the eastern Tibetan Plateau (mainly in Sichuan) also contains two of China’s biological treasures: the range of the Giant Panda at the Wolong Preserve and E-meishan (3099 m), one of China’s four sacred mountains, with a well-developed zonation (and footpath) from laurel forest to subalpine conifer forest of *A. fabri*. To the south, at elevations not higher than about 1500 m on the Yunnan Plateau, smaller areas of laurel forest remain in the Ailao Shan and other ranges southward into Xishuangbanna.

In the transition region of eastern China (summergreen to evergreen laurel forest), two other accessible mountain systems also have well-developed zonations, which are botanically well described: Huang-Shan in southeastern Anhui province and Tianmu-Shan in northern Zhejiang province. The mountains of Taiwan are especially steep, and the montane laurel forests are consequently relatively well preserved.

In northeastern China, along the Korean border, the vertical zonation of the Changbai-Shan can be described as follows:

- 0–300 m: summergreen deciduous forest
- 300–1100 m: mixed deciduous and conifer forest (*P. horaiensis* plus much secondary forest of *Betula platyphylla*)
- 1100–1700 m: coniferous forest (*Picea jezoensis* var. *komarowii*, *P. horaiensis*, with *Abies nephrolepis*, *Acer*, *Betula*, *Sorbus*)
Further to the north, larch forests dominated by Larix gmelinii occur in the northern Da Hinggan Mountains near the Siberian border (Missouri Botanical Garden, 1983; Wang, 1961).

Montane forests and zonation in Korea can perhaps best be seen on Mt. Seolag (1308 m). The zonation is similar to that described for Changbai-Shan but with a greater importance of deciduous Quercus and Carpinus forests at low elevation and the lack of an alpine belt.

The mountain vegetation of Japan has been very extensively studied and described in the evergreen as well as the deciduous forest zones (Numata, 1974). Among the most impressive native coniferous forests of Japan are the tall forests of Cryptomeria japonica occurring on Yakushima island and in some areas with more than 3000 mm of precipitation on Shikoku and Honshu. These forests could be called temperate rain forests. Forests of Chamaecyparis obtusa/Juliana are also tall and impressive. Other montane conifers in Japan include Abies homolepis, A. mariesii, and A. veitchii mainly on Honshu; A. sachalinensis mainly on Hokkaido; Picea jezoensis (Hokkaido) and P. glehnii; Larix leptolepis (on old lava); Tsuga sieboldii (lower mountains) and subalpine Ts. diversifolia; Thuja standishii; and Thujaoschistosolidae (endemic genus).

In most of northeastern Asia, the subalpine belt is characterized by dense stands of the krummholz species Pinus pumila, which also occurs in Siberia. In more maritime areas approaching the boreal zone, such as Hokkaido, conifers may be replaced by forests of birch, especially Betula ermanii.

VII. BOREAL AND POLAR ASIA

Boreal and polar Asia extends from the Ural Mountains to the Pacific Ocean and from islands in the Arctic Ocean to the transition to temperate climates in the south, approximately along the southern boundary of Siberia. In Europe, this transition takes the form of mixed boreo-nemoral forests of deciduous broad-leaved trees and conifers, narrowing eastward from Poland through Moscow and on to the Urals around Sverdlovsk. In Asia, this transition is a more discontinuous zone of mixed forests and forest-steppe, running from the southern Urals past the southern end of Lake Baikal and on to northeastern China, where it widens again between the Siberian larch taiga to the north and the deciduous forests of Manchuria. In the middle section, the transition is mainly a narrow belt of aspen (Populus tremula) and birch parklands between subboreal or boreal forests to the north and the grasslands of Mongolia to the south.

During the Pleistocene, much of Siberia remained unglaciated but covered by tundra and cold steppe vegetation, which extended southward to the grasslands with no intervening coniferous forest zone. The boreal forest which now covers most of Siberia is thus a relatively recent phenomenon. This boreal forest is mainly an evergreen coniferous forest in the west but opens to lighter larch (Larix) forest and woodland in the east, where winters are most severe. The west-Siberian lowland is a vast complex of boreal forest and wetlands, whereas northeastern Siberia is quite mountainous. The larch region of interior eastern Siberia contains the most continental climates in the world, with mean July temperatures usually exceeding 20°C but with mean January temperatures generally near or below −30°C, reaching −51°C at Oymyakon and Verkhoyanski in the northeast. The transition to polar tundra involves a zone of open forest-tundra woodland, which reaches its farthest point north (in the world) in the Taimyr Peninsula of northwestern Siberia. Polar tundra forms a narrow belt along the northern coastal plain completely across Eurasia.

A. Boreal Forest

If the subboreal zone is included, four main types of boreal forest can be recognized (Walter, 1974):

1. Mixed forests of summergreen broad-leaved trees and conifers (mainly Picea and Pinus), along with localized pine forests
2. Dense forests of spruce (Picea), fir (Abies), and five-needle pine (Pinus sibirica)
3. More open forests and woodlands of Scots pine (Pinus sylvestris)
4. Generally open forests and woodlands of larch (Larix), a genus of deciduous conifers

Of these, the dense forests of spruce, fir, and Siberian pine are generally called “dark taiga,” whereas the more open forests and woodlands of pine and larch are called “light taiga.”
Boreo-nemoral mixed forests, with *Acer*, *Tilia*, *Ulmus*, *Fraxinus*, *Quercus*, *Picea*, *Pinus*, etc., are better developed in eastern Europe than in Asia but reappear in Manchuria and the Russian Far East and some smaller areas such as the subboreal region on the west side of Lake Baikal. In more continental areas, *Betula* and *Populus* may be the only important broad-leaved trees, and pines may be the main conifers, especially if the soil is coarse and less humic. At the southern end of Lake Baikal (32°–54°N, 450 m), boreo-nemoral mixed forests can be found in moist alluvial lowlands, whereas dark taiga, with larch and five-needle pine, occurs on uplands higher than about 700 m.

Dark taiga is composed of shade-tolerant conifers, mainly Siberian spruce (*Picea obovata*), five-needle Siberian pine (*P. sibirica*), and Siberian fir (*Abies sibirica*). All three of these extend from the Urals into southeastern Siberia (if not completely to the Pacific). Dark taiga is limited geographically by the extreme winter cold, yielding to deciduous larch taiga generally where mean January temperature falls below about –30°C (extremes to below –60°C). The range of dark taiga is thus mainly European Russia through western Siberia and in a narrowing zone into southeastern Siberia. Within this range, dark taiga occurs over most of the area, with more open forests of *P. sylvestris* confined mainly to sandy, upland rock, or other drier substrates. In some smaller areas, mainly in mountains, other species of *Picea* and *Abies* appear as well as other five-needle pines such as *P. cembra* and *P. pumila*, often occurring as subalpine krummholz. Larches, often large, also occur widely as companion species within the dark taiga.

Scots pine (*P. sylvestris*) has the widest distribution of any tree species in Eurasia and covers large areas of light taiga throughout the European and Asian boreal zone and occurs in mixed forests in Europe, the Caucasus, and northern Manchuria. Boreal pine forests are usually open and have understories of two main types: Cladonia and similar lichens (plus mosses) on drier sites such as sand, and ericads such as Vaccinium on more moist sites.

Most of the boreal zone of central and eastern Siberia is covered by extensive open forests and woodlands of larch, especially *Larix dahurica* but also other species with smaller ranges. *Larix sibirica* is more important as a companion in the dark taiga of western Siberia. As in many higher latitude mountainous regions, the larch woods may occur mainly on slopes while flat valley bottoms remain largely treeless, a pattern attributed to cold-air drainage. The ground vegetation in these areas and in the larch stands as well is composed largely of boreal and tundra dwarf-shrubs such as *Vaccinium uliginosum*, *Ledum*, *Dryas*, and *Arctous*, plus abundant mosses and lichens. In moist areas shrubs from the genera *Juniperus*, *Salix*, and *Betula* often become important understory elements.

The larch woods of interior northeastern Siberia in particular occur where annual precipitation is generally under 300 mm and sometimes under 200 mm. Although the growing season is thus a period of climatic water deficit, the trees survive from soil water which is effectively rationed to them over the course of the summer by the slow thawing of the soil (*Walter, 1974, p. 84*). Larches are outcompeted by evergreen conifers in some boreal areas but have no tree competitors in the coldest parts of eastern Siberia. Larch-dominated landscapes thus extend from the polar treeline across much of Siberia southward to Mongolia and into inland areas of the Russian Far East (Primorye and Khabarovsk provinces) and mountains of northermost Manchuria.

Mountains within the boreal zone may quickly reach treeline. On lower slopes, however, the boreal forest may be somewhat different from the surrounding flatter lowlands. In the Ural Mountains, forests on the wetter western slopes generally involve *P. sylvestris*, *Abies sibirica*, *Picea abies*, *Betula pubescens*, and krummholz of *P. pumila*, whereas forests on the drier eastern slopes are mainly *P. sylvestris*. In less extreme, maritime areas of eastern Siberia, the deciduous larch taiga may be replaced again by evergreen conifers, such as *Abies nephrolepis* around the Sea of Okhotsk and *A. sachalinensis* on Sakhalin island (also northern Japan). Stunted forests of *B. cemani* appear in the most maritime areas (e.g., Kamchatka, Sakhalin, and northern Japan).

### B. Polar and Mountain Tundra

Polar ecosystems, generally called tundra, occur across the northern coastal plain of Eurasia but also in both upland and high-mountain situations, such as the lower mountains of boreal northeastern Siberia and the high Tibetan Plateau. Upland tundra in the boreal zone represents the first upward vegetation belt and is thus commonly called montane tundra. In high mountains, tundra-like ecosystems occur above the alpine treeline and are generally called alpine tundra. In all three situations, these are treeloss landscapes dominated by mosaics of the following basic associations:

1. Largely evergreen, microphyllous dwarf-scrub on locally somewhat higher, drier areas
2. Wet graminoid vegetation in the lowest areas, generally flat, marshy depressions
3. Mixes of evergreen and summergreen plants, herbaceous and dwarf-shrub, including mosses and lichens, on the broad, slight slopes between the higher and lower areas.

4. Mainly mosses and lichens on the most extreme areas, including the coldest but also the most exposed upland areas where little snow remains.

In mountains, two additional types can be identified. One is the "snow valleys" (Scheentälchen), depressions with deep snow accumulations which have shorter snow-free seasons and specialized floras composed of plants which can complete their life cycles in periods as short as one month in summer. The other is the extensive alpine mats, often dominated by Kobresia, found in drier high-mountain areas such as the Tien Shan, Hindu Kush, and Tibetan highlands, generally above 3000 m.

The polar tundra extends from extreme northern Fennoscandia across northern Russia to coastal north-eastern Siberia in a narrow strip which is widest on the Taimyr peninsula. Due to continentality effects, however, the Taimyr is also the area where boreal forest extends farthest north, reaching about 72.3°N (as larch forest) north of Khantanga. On the flat coastal topography, herbaceous tundra is perhaps most extensive. The polar tundra can be represented zonally, however, in four subzones from south to north:

1. The transitional forest–tundra of largely dwarf-shrub tundra with scattered but still tall individual conifers (mainly larches in the east and spruces in the west).

2. Dwarf-shrub tundra, dominated largely by shrub forms from summergreen tree genera, mainly Betula, Salix, and Abies.

3. Largely herbaceous tundra with few dwarf-shrubs, but still a continuous vegetation cover including many mosses and lichens.

4. The discontinuous High Arctic cold-desert with vegetation restricted to more scattered individuals of a few vascular plants (especially Dryas spp.) and various lichens and mosses.

The forest–tundra and dwarf-shrub tundra zones of the lowlands are better represented in western Siberia, but montane analogs are fairly widespread in the lower mountain areas east of the Lena River in eastern Siberia. Cold-desert, on the other hand, is a bit more widespread in the mountain areas east of the Lena River in eastern Siberia. Montane analogs are fairly widespread in the lower lowlands are better represented in western Siberia, but polar tundra, perhaps due to the lack of permafrost and the greater number of microhabitats in the more heterogeneous terrain of high mountains. Many more localized species usually occur. Some important genera of Asian high mountains which are not important if present at all in polar tundra include Gram, Gentiana, and Leontopodium.

Despite its relatively low productivity, the polar tundra supports many animal species and in surprisingly large numbers, at least in summer. Caribou (reindeer) winter in the boreal zone and migrate to the tundra during the summer. Birds migrate from much greater distances away, some from as far as tropical Asia. Other birds, such as prarigman, overwinter near the polar zone. Mammals such as polar bears and some small rodents (e.g., lemmings) also live year-round in the polar zone.
Wetlands include mangroves, other coastal ecosystems such as strand forests, beach vegetation, salt marshes, deltas, and estuaries and terrestrial wetlands such as swamp forests, marshes, bogs, and (sensa lato) even lakes and streams. Most wetlands are highly productive ecosystems when not badly degraded, have important economic and social benefits for local human populations, and perform useful ecological services such as cleansing wastewater. Wetlands also support a wide range of animal life, often far beyond their own borders, as in the case of migratory birds, fish, and shellfish.

Mangroves are salt-tolerant forests growing in the intertidal zone, thus flushed by saltwater tides once or twice daily but also influenced by the inflow regime of freshwater runoff from the land. Behind many mangroves are so-called “back mangroves” which are less influenced by wave energy, generally less salty, and contain different and usually more species. Mangroves have developed in part by trapping fine sediment, so that the mangrove substrate is usually deep mud (but also shallower mud and sand). As a result, mangroves are home to abundant burrowing shrimp, crabs, and other crustaceans as well as spiders, insects, and some birds. The total number of plant “mangrove species” in the world has generally been estimated at fewer than 100, involving especially the genera Rhizophora and Avicennia as well as particularly Asian genera such as Bruguiera, Ceriops, and Laguncularia (Rhizophoraceae); Sonneratia, and Aegiceras. Back mangroves contain additional taxa such as Launaea arborescens, Xylocarpus, the mangrove palm (Nipa fruticans), Acutantis, the widespread fern Acrostichum aureum, and even the buttressed tree Heritiera littoralis (Sterculiaceae) in some cases. In areas protected from direct hits by typhoons, mangrove forests grow to about 35 m in height. In more exposed areas, however, and in areas with shallower or hard mud, mangroves are reduced to only a few meters in height. In areas with higher salinity (drier climates and less freshwater inflow), mangroves are also shorter and may have saline lagoons behind them instead of more productive back mangroves.

Mangroves in Asia are considered to be the most diverse in the world and covered about 7.7 million ha at the time of the Global Biodiversity census (Groombridge, 1992). As shown in Table X, most of this area was in Indonesia (about half in Sumatra), Malaysia (about half in Sabah), Bangladesh, the Philippines, Vietnam, India, Thailand, and Pakistan (see complete statistics in Groombridge, 1992, p. 325). Another 3 million ha belonged to Indonesia but are in Irian Jaya (New Guinea), which geologically is part of Australia, not Asia. These statistics are old, and mangroves have been declining steadily and often rapidly, being converted for development, urban sprawl, or mining (tim); cut for wood, or simply degraded by overuse.

On higher-energy tropical and subtropical coastlines, where the substrate is usually sand, the first beach pioneer inland from the water line is usually the creeping vine Ipomoea pes-caprae (Convolvulaceae). Behind this the first stable vegetation is often a dense thicket of large-sclerophyll shrubs, Scaevola (Goodeniaceae) and sometimes Mesemvryum (Boraginaceae), plus some pandans and various smaller species which elsewhere might pass as weeds. On many sandy tropical shorelines, coconut palms (Cocos nucifera) have established naturally or been planted and cultivated. Strand forests may develop behind beaches and on raised coral terraces, dominated by tree species such as Pongamia pinnata, Ficus microcarpa, and Bischotia javanica.

At higher latitudes, where winter cold precludes tropical taxa and the previously mentioned species, salt marshes are the most common shoreline ecosystem. These are also highly productive ecosystems, rivaling tropical rain forests in some cases, and can cover quite wide areas along more concave shorelines, where the intertidal range is higher and thus reaches farther inland. In east Asia, salt marshes extend as far north as the Russian Far East north of Vladivostok. Of course, estuaries also extend to higher latitudes and have been the basis for human settlement in coastal areas for millennia because of their abundance of fish and shellfish. All of these coastline ecosystems, though narrow, are important as buffers between the ocean and the more widespread, typical forests behind them.

Inland, the western Siberian lowland (between the Urals and the Yenisey River, about one-third of all of Siberia) is mostly a vast boreal swamp and may rank as the world’s largest single wetland. Boreal landscapes in general, however, are often slowly shifting mosaics of forested uplands and lower lying fens and bogs. In the cold climate most bogs accumulate peat, which raises the land surface such that former depressions become convex surfaces fed only by nutrient-poor precipitation rather than groundwater. These are the highly acidic “raised bogs” which contain mainly eutrophic dwarf-shrubs (e.g., Oxyurus and Andromeda), sedges (e.g., Eriophorum), and mosses, especially Sphagnum species. Around the edges of these (and other) bogs may be stunted birches, larches, and other trees tolerant of saturated soils and high acidity. Although individu-
ally small and not stable over long time scales, these bogs cover a very large area and store much carbon, which may be released to the atmosphere as global warming warms and dries these areas.

Bogs and (less acidic) fens also occur in the temperate zone to the south but especially in moist, cooler areas such as northern Japan. More important in drier climates are probably the freshwater marshes, which are of great importance to birds migrating between wintering areas in tropical Asia (e.g., India) and summer breeding grounds in Siberia. Such marshes may occur in depressions as well as along streams, especially in the deserts of western China, the grasslands of northern China and Mongolia, and in Middle Asia and the highlands of Tibet. These marshes are composed primarily of reeds (Phragmites), cattails (Typha), sedges (Cyperus, Scirpus, and Carex), and similar herbaceous plants.

Among lakes and streams, the important ecological distinction is between more aerated, nutrient-rich flowing waters and standing waters which can become quite oligotrophic, especially in cooler climates. Lake Baikal, in southern Siberia, is the world’s deepest lake (1620 m) and contains about 20% of all the world’s (unfrozen) fresh water. The lake contains about 600 plant species (35% endemic) and 1200 animal species (60% endemic). It is also large enough to produce a distinct limnoclimatic atmosphere which includes species from boreal and polar areas occur in small disjunct populations. Another well-known lake in Asia is Dal Lake in Kashmir, which has a flat bottom, is only a few meters deep, and has been filling rapidly over the past decades with sediment from the surrounding mountains, which have been denuded largely since the 1950s. This and many other lakes are also extremely important for migrating birds. In temperate and tropical Asia, the large rivers are especially important, even though their ecosystems are often badly degraded or at least greatly altered by introductions of exotic species.

### Table X

Mangrove Areas in Asia

<table>
<thead>
<tr>
<th>Country</th>
<th>Total mangrove area (ha)</th>
<th>Number of protected mangrove areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>67,000</td>
<td>24</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Japan (Ryukyu Islands)</td>
<td>400</td>
<td>4</td>
</tr>
<tr>
<td>Taiwan</td>
<td>174</td>
<td>4</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brunei</td>
<td>7,000</td>
<td>3</td>
</tr>
<tr>
<td>Cambodia</td>
<td>10,000</td>
<td>0</td>
</tr>
<tr>
<td>Indonesia (including Irian Jaya)</td>
<td>4,351,011</td>
<td>152</td>
</tr>
<tr>
<td>Malaysia</td>
<td>630,000</td>
<td>99</td>
</tr>
<tr>
<td>Myanmar</td>
<td>517,000</td>
<td>6</td>
</tr>
<tr>
<td>Philippines</td>
<td>400,000</td>
<td>99</td>
</tr>
<tr>
<td>Singapore</td>
<td>1,400</td>
<td>2</td>
</tr>
<tr>
<td>Thailand</td>
<td>208,003</td>
<td>17</td>
</tr>
<tr>
<td>Vietnam (South)</td>
<td>370,000</td>
<td>2</td>
</tr>
<tr>
<td>South and Southwest Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andaman Islands</td>
<td>(50,000)</td>
<td>29</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>410,000</td>
<td>3</td>
</tr>
<tr>
<td>India</td>
<td>306,000</td>
<td>9</td>
</tr>
<tr>
<td>Iran</td>
<td>23,117</td>
<td>3</td>
</tr>
<tr>
<td>Maldives</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>Pakistan</td>
<td>240,950</td>
<td>2</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>120,000</td>
<td>9</td>
</tr>
<tr>
<td>Arabian Peninsula</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahrain</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>Saudi Arabia, Oman, Qatar</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>—</td>
<td>6</td>
</tr>
<tr>
<td>Yemen</td>
<td>—</td>
<td>3</td>
</tr>
</tbody>
</table>

* Data are from Groombridge (1992, p. 323).
### TABLE XI

**Most Seriously Threatened Wetlands in Asia**

<table>
<thead>
<tr>
<th>Country</th>
<th>Wetlands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangladesh</td>
<td>Chalan Beel*, Haor Basin of Sylhet and Eastern Mymensingh, The Sundarbans, Wetlands in Pabna Wildlife Sanctuary, Chokoria Sundarbans*</td>
</tr>
<tr>
<td>Bhutan</td>
<td>Bumthang Valley</td>
</tr>
<tr>
<td>Burma (Myanmar)</td>
<td>Ironwaddy Delta</td>
</tr>
<tr>
<td>China</td>
<td>Yangtze Marshes, Shujin Hu, Shengjin Hu and the lower Yangtze Lakes, Shengjin Hu, Xi Jiang (Pearl River) Delta*, Yueshu Hu (Kurlyk Nor) and Kuerheko Hu</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Banyunus Must River Delta, Muara Cimandeh*, Sukoharjo, Cilegon and Sagara Anakan, Danau Banaran, other swamps of Barito Basin*, Banas Senturan, Wetlands in Manusa, proposed national park, Waaser and Rawa Buru</td>
</tr>
<tr>
<td>Japan</td>
<td>Shonan-Fujinabe Tidal Rais and Inner ls Bay, Lake Shimp and Lake Nakaizumi</td>
</tr>
<tr>
<td>Korea</td>
<td>South Ganghwa and North Yongyang muillasses, Mudflats of South Yongyang and adjacent islands, Namyang Bay, Anan Bay, Kim, Manleyung and Tangin estuaries</td>
</tr>
<tr>
<td>Malaysia</td>
<td>Selish Ked-swa swamp forest, Klung Islands, Pelau Ketan*, Kapur Forest Reserve, North Selanger swamp forest, Marintaman Manglaong*, Temposak Plant, Lawas mangroves, Temoan-Sumilang mangroves, Limbong mangroves, Malasum swamp forest, Sarawak Mangrove Forest Reserve</td>
</tr>
<tr>
<td>Nepal</td>
<td>Begaun Tal*</td>
</tr>
<tr>
<td>Pakistan</td>
<td>Khaba Lake*, Stranda Lake*, Hawkes Bay/Sandspit Beaches and adjacent creeks, Clifton Beach, Korangi and Ghoro Creeks, The Outer Indus Delta</td>
</tr>
<tr>
<td>Philippines</td>
<td>Pangawan Wetlands*, Manila Bay*, Laguna de Bay, Tayabas Bay including Pagbilao Bay, Inabanga Coast, Maran, Kaluwanan and Cagsaya Bays*, Apsan marsh, Lake Leonard*, Dowas Golf, Leguanan marsh</td>
</tr>
<tr>
<td>Singapore</td>
<td>Serington estuary*</td>
</tr>
<tr>
<td>Taiwan</td>
<td>Taru estuary, Tungmih (Ton-Shou) mangroves*</td>
</tr>
<tr>
<td>Thailand</td>
<td>Golf of Thailand, Pak Phunang estuary, Ph Phrum</td>
</tr>
<tr>
<td>Vietnam</td>
<td>Red River Delta, Red River estuary, Mekong Delta, Nam Can mangrove forest</td>
</tr>
</tbody>
</table>


* Sites are considered to be already too degraded to merit any special conservation effort.
A listing of the most seriously threatened wetlands in Asia is shown in Table XI. Some sites are denoted as already too degraded to merit any special conservation effort. Many other wetlands are not on this priority list but were or have since become threatened to the point that their biodiversity as well as their benefits to local human societies are being rapidly lost.

### IX. ARTIFICIAL ECOSYSTEMS

One other class of ecosystems, especially important in long-settled Asia, is artificial ecosystems such as those involving tree plantations, urban landscapes, rice paddies, cropland, managed ponds, etc. Plantations of native conifers are especially extensive in the mountains of Japan, but much more widespread, especially in tropical Asia, are plantations of exotic eucalyptus, Australian pine (Casuarina), and other trees not from any part of Asia. Although these plantations, which mainly provide fuelwood, may take the pressure off more natural, native forests, they also damage the soil such that it no longer supports the normal diversity of soil and understory plants, animals, and microbes, especially decomposer organisms which break down and cycle minerals.

Rice paddies and cropland do support some degree of diversity among species which are commonly associated with certain types of management regimes. These largely weedy species may not be species of stable native ecosystems, but they do form somewhat consistent communities. Ponds and other water bodies managed for fish raising, etc. are also common in warmer parts of Asia. The species diversity is low and involves mainly exotic, weedy species such as Myriophyllum or water hyacinth (Eichhornia crassipes).

Urban landscapes are extensive and ever-growing in Asia as elsewhere. Urban ecosystems may involve alleys and other plantings for landscape amelioration but may also involve seminatural wooded, grassy, and aquatic communities. Some of the methods for building seminatural forests in densely built-up areas are quite well developed, especially in Japan (Miyawaki et al., 1987).

### X. CONSERVATION STATUS

The conservation status of natural and even of modified but functioning, quasinatural ecosystems in Asia varies widely. Some vast tracts of boreal forest appear un-
### TABLE XIII
Total Numbers of Threatened Plant and Vertebrate Species by Country

<table>
<thead>
<tr>
<th>Country</th>
<th>Plants</th>
<th>Mammals</th>
<th>Birds</th>
<th>Reptiles</th>
<th>Amphibians</th>
<th>Fish</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mediterranean–Southwest Asia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afghanistan</td>
<td>4</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cyprus</td>
<td>45</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Iran</td>
<td>301</td>
<td>15</td>
<td>20</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Iraq</td>
<td>1</td>
<td>19</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Israel</td>
<td>5</td>
<td>8</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Jordan</td>
<td>752</td>
<td>11</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lebanon</td>
<td>5</td>
<td>4</td>
<td>15</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Syria</td>
<td>11</td>
<td>4</td>
<td>15</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Turkey</td>
<td>1044</td>
<td>9</td>
<td>18</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><strong>Arabian Peninsula</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahrain</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Kuwait</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Oman</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Qatar</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>2</td>
<td>9</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yemen</td>
<td>134</td>
<td>6</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Tropical Asia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td>15</td>
<td>15</td>
<td>27</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bhutan</td>
<td>13</td>
<td>15</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>British Indian Ocean Territory</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Brunei</td>
<td>40</td>
<td>9</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Cambodia</td>
<td>11</td>
<td>21</td>
<td>13</td>
<td>6</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>India</td>
<td>1336</td>
<td>79</td>
<td>17</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Indonesia</td>
<td>70</td>
<td>49</td>
<td>13</td>
<td>3</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Laos</td>
<td>5</td>
<td>25</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td>322</td>
<td>25</td>
<td>39</td>
<td>12</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Maldives</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Myanmar</td>
<td>—</td>
<td>23</td>
<td>42</td>
<td>10</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Nepal</td>
<td>35</td>
<td>22</td>
<td>20</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pakistan</td>
<td>14</td>
<td>15</td>
<td>25</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Philippines</td>
<td>159</td>
<td>12</td>
<td>39</td>
<td>6</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Singapore</td>
<td>19</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>220</td>
<td>7</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Thailand</td>
<td>68</td>
<td>26</td>
<td>34</td>
<td>9</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Vietnam</td>
<td>338</td>
<td>28</td>
<td>34</td>
<td>8</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><strong>East Asia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>350</td>
<td>40</td>
<td>83</td>
<td>7</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Japan</td>
<td>45</td>
<td>31</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Korea (North)</td>
<td>0</td>
<td>5</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Korea (South)</td>
<td>35</td>
<td>16</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mongolia</td>
<td>0</td>
<td>9</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Taiwan</td>
<td>95</td>
<td>4</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Data are from the Global Biodiversity report (Groombridge, 1992, p. 239).*

touched, but one cannot fly over these areas without seeing smoke (often industrial) rising from what appear to be very isolated places. Other forest areas have almost completely disappeared. It has been estimated that less than 1% of the natural evergreen broad-leaved forest area of Japan remains, and natural broad-leaved as well as conifer forests in China are confined entirely to mountain areas, where they now are also facing encroachment. Tropical forests have also largely disappeared, remaining over large areas mainly in Southeast
Asia (see Table VII). The especially well-developed tropical rain forest of Borneo, with canopy heights commonly near 50 m, have largely been cut within the past 20 years, mostly for export to Japan. Some estimates of original and remaining area of tropical rain forest are shown in Table XII, but the data for remaining area (all pre-1990) must be overestimates.

Country totals of threatened plant and vertebrate species, and also fish, in Asia are summarized in Table XIII. Although these statistics reflect different levels of coverage and different methodologies, they give an indication of the magnitude of the problem of threats to biodiversity by changing conditions—mostly changes in land use by humans. Large numbers of threatened plants are recognized only in Turkey in the west, in Yemen on the Arabian Peninsula, in China in east Asia, and in India, Malaysia, Vietnam, and Sri Lanka in tropical Asia—although many others must also be threatened in other areas. The pattern for animals is more even across countries and regions, but with Indonesia or China recognizing the largest numbers of threatened species in all five categories.

Species and ecosystems are endangered throughout Asia, as elsewhere. For further reading, see in particular the various recent summaries of the status of conservation efforts and strategies in all parts of the world (Hoyt, 1994; Heywood, 1995) as well as more standard treatments of the general problems of wildlife conservation.

See Also the Following Articles

AFRICA, ECOSYSTEMS OF • EUROPE, ECOSYSTEMS OF • NEAR EAST, ECOSYSTEMS OF • NORTH AMERICA, ECOSYSTEMS OF • SOUTH AMERICA, ECOSYSTEMS OF

Bibliography


ATMOSPHERIC GASES

Donald J. Wuebbles
University of Illinois at Urbana-Champaign

I. Introduction
II. The Changing Composition
III. The Importance of Stratospheric Ozone
IV. Air Pollutants in the Lower Atmosphere
V. The Future and Atmospheric Gases

GLOSSARY

atmospheric lifetime Timescale characterizing the rate of removal of a gas from the atmosphere, generally defined as the time to remove 63.2%, the e-folding time.
greenhouse gases Gases like water vapor and carbon dioxide that can affect climate through their absorption and reemission of terrestrially emitted infrared radiation.
mixing ratio Ratio of the concentration of a gas to the concentration of air. It can be defined in terms of volume (e.g., molecules per m³) or mass (e.g., g per m³).
oxidizing capacity Self-cleansing ability of the atmosphere through oxidative reactions.
ozone Molecule composed of three oxygen atoms that is extremely important to life on Earth because of its absorption of solar ultraviolet radiation; ozone is also a greenhouse gas.
photochemical source or sink Production or loss of an atmospheric gas involving photodissociation and/or chemical reactions.
photodissociation Destruction of a molecule through absorption of solar radiation and subsequent breaking of one or more chemical bonds.

THE EARTH'S ATMOSPHERE IS COMPOSED OF A MIXTURE OF GASES THAT WE CALL AIR. The most abundant of these gases, molecular nitrogen and molecular oxygen, are not being greatly affected by human activities and are unlikely to be so in the foreseeable future. However, many of the other gases, like carbon dioxide and ozone, are being affected by human activities. Although there are only trace amounts of these gases in the atmosphere, many of them are extremely important to life on Earth and to biodiversity. This article describes many of the important atmospheric gases, their role in affecting the planet, and the factors affecting their changing concentrations in the atmosphere.

I. INTRODUCTION
As seen in Table I, the bulk of the Earth's dry atmosphere is composed of molecular nitrogen (N₂, 78.08%), molecular oxygen (O₂, 20.95%), and argon (Ar, 0.93%). While the remaining gases comprise only about 0.03% of the Earth's atmosphere, many of them are extremely important to the climate and habitability of our planet. These gases are the primary focus of this article. The concentrations of many of these gases are also being affected by human activities.
Some of the gases listed in Table I, like ozone (O₃), affect the transmission of solar radiation in the atmosphere. Others, like carbon dioxide (CO₂), affect the absorption of terrestrially produced infrared radiation and are called greenhouse gases. These types of gases are therefore linked to the physical climate system of the Earth. Increasing concentrations of greenhouse gases have resulted in concerns about global warming. Some trace gases, such as sulfur dioxide (SO₂) and hydrogen sulfide (H₂S), can affect climate through their role as precursors to the production of atmospheric particles. The resulting sulfate particles can scatter solar radiation, thus preventing some of it from reaching the Earth’s surface. These particles can also serve as condensation nuclei in the formation of clouds.

The ozone layer is a term that refers to the distribution of ozone that is naturally formed in the stratosphere. This layer protects life on Earth from harmful levels of solar ultraviolet radiation. Chlorofluorocarbons (CFCs) and other chlorinated and brominated halocarbons are emitted by a variety of human activities. Atmospheric measurements have clearly corroborated theoretical studies showing that the chlorine and bromine released from the destruction of these halocarbons in the stratosphere have been reacting to destroy significant amounts of ozone over the last few decades.
Gases like hydroxyl (OH), ozone, and hydrogen peroxide (H₂O₂) control the oxidizing capacity of the atmosphere, and destroy many pollutants emitted into the atmosphere through chemical reactions. Hydroxyl has been referred to as the "atmospheric vacuum cleaner." It is generated primarily by the interactions of water vapor, ozone, and solar ultraviolet radiation, and is destroyed primarily through reaction with carbon monoxide (CO) and methane (CH₄).

Emissions of trace gases in urban regions can result in atmospheric pollution referred to as photochemical smog. For example, ozone, a potential cause of health problems in polluted regions, can be formed in the air by chemical reactions following emissions of nitrogen oxides (NOₓ) and volatile organic gases (VOCs), including hydrocarbons. NOₓ is the sum of the concentrations of nitric oxide (NO) and nitrogen dioxide (NO₂), while NO also includes other reactive nitrogen gases like nitric acid (HNO₃).

Missing from Table I is water vapor, a gas with highly variable concentration, whose amount in the atmosphere is dependent on evaporation and hydrological cycle processes associated with weather. The concentration of atmospheric water vapor can be affected at the local scale by human activities but is not thought to be greatly affected at the global scale by human activities. Biodiversity affects the concentrations of atmospheric trace gases and can also be affected by changing concentrations of atmospheric gases. The biosphere is an important component in the biogeochemical cycles of carbon, nitrogen, and sulfur. As a result, the sources and sinks of many atmospheric gases are strongly affected by changes in the biosphere. For example, deforestation is contributing to the increase in concentration of carbon dioxide: Local air pollution, changes in climate, and changes in stratospheric ozone can affect the biosphere and have an impact on biodiversity. For example, acid rain or acidification resulting from emissions of nitrogen- and sulfur-containing gases can affect landscapes in many ways.

II. THE CHANGING COMPOSITION

Without human intervention, concentrations of many atmospheric gases would be expected to change slowly. Ice core measurements of the gases trapped in ancient ice bubbles indicate that this was the case before the last century. However, since the beginning of the Industrial Age, emissions associated with human activities have risen rapidly. Agriculture, industry, waste disposal, deforestation, and especially fossil fuel use have been producing increasing amounts of carbon dioxide, methane, nitrous oxide (N₂O), chlorofluorocarbons, and other important gases. Because of these increasing emissions, atmospheric levels of these gases have been building at an unprecedented rate, resulting in effects on ozone, another radiatively important gas, and raising concerns regarding the impact of these gases on climate. Although all of these gases are having some influence on climate, carbon dioxide is of particular interest in the concerns about global warming. In addition, changes in the human-related emissions and atmospheric concentrations of small particles or aerosols are also influencing climate. The following discussion focuses on the changing concentrations and budgets of a number of these gases.

A. Carbon Dioxide

Of all the greenhouse gases, carbon dioxide has been undergoing the largest changes in concentration. It is also the gas of most concern to analyses of the potential human effects on climate. Accurate measurements of atmospheric CO₂ concentration began in 1958 at the Mauna Loa Observatory in Hawaii. Figure 1 shows that the annually averaged concentration of CO₂ in the atmosphere has risen from 316 ppm (parts per million) in 1959 to 364 ppm in 1997. The CO₂ measurements exhibit a seasonal cycle, which is mainly caused by the seasonal uptake and release of atmospheric CO₂ by terrestrial ecosystems. The average annual rate of
ties are primarily responsible for emissions of CO₂: fossil fuel use, which released about 6.0 GtC (gigatons of carbon; a gigaton is 10¹⁰ tons) into the atmosphere in 1990, and land use, including deforestation and biomass burning, which may have contributed about 1.6 ± 1.0 GtC in addition to that from fossil fuels.

Why has the atmospheric concentration of CO₂ increased so dramatically? Analyses with models of the atmosphere–ocean–biosphere system of the carbon cycle indicate that human activities are primarily responsible for the increase in CO₂ over the past century. Two types of human activities are important: emissions of carbon dioxide from fossil fuel use, which released about 6.0 GtC during the period from 1980 to 1990, and land use, including deforestation and biomass burning, which may have contributed about 1.6 ± 1.0 GtC in addition to that from fossil fuels.

Human-related emissions from fossil fuel use have been estimated as far back as 1751. Before 1863, emissions did not exceed 0.1 GtC/yr. However, by 1995 they had reached 6.0 GtC/yr, giving an average emission growth rate slightly greater than 3% per year over the last two and a half centuries. Recent growth rates have been significantly lower, at 1.8% per year between 1970 and 1993. Emissions were initially dominated by coal. Since 1983, petroleum products have been the main source of emissions despite their lower carbon content. The regional pattern of emissions has also changed. Global emissions were once dominated by Europe and North America, but now developing nations are providing an increasing share of emissions. In 1995, developing countries (including China and India as part of the Annex I nations being used in international policy discussions) accounted for 48% of global emissions.

2. Emissions from Land Use Changes

The biosphere holds approximately 560 GtC in the form of aboveground biomass, and an additional 1200 GtC in soils and detritus. These pools form the principal reservoirs from which terrestrial systems can exhaust or sequester carbon. Evaluations of carbon releases from vegetation and soils based on changes in land use indicate that, since 1800, land use practices have decreased carbon storage in vegetation and soil by about 170 Gt. Estimates of the net flux of carbon into the atmosphere in 1980 due to land use range from 0.4 to 2.5 GtC/yr. For the decade of the 1980s, tropical emissions from changes in land use averaged 1.6 ± 1.0 GtC/yr.
B. Methane

Although methane’s atmospheric abundance is less than 0.5% that of CO₂ on a molecule-by-molecule basis, a molecule of CH₄ is approximately 50 times more effective as a greenhouse gas in the current atmosphere than CO₂. When this is combined with the large increase in its atmospheric concentration, methane becomes the second most important greenhouse gas of concern to climate change. Based on analyses of ice cores, the concentration of methane has more than doubled since preindustrial times. The current globally averaged atmospheric concentration of methane is about 1.73 ppmv (Fig. 2). Continuous monitoring of methane trends in ambient air from 1979 to 1989 indicates that concentrations have been increasing at an average of about 16 ppbv (parts per billion by volume) or 1% per year.

FIGURE 2 (Top) Globally averaged atmospheric methane (CH₄) concentrations (ppbv, parts per billion by volume) derived from the NOAA Climate Monitoring and Diagnostics Laboratory air-sampling sites by E. Dlugokencky and colleagues. The solid line is a deseasonalized trend curve fitted to the data. The dashed line is a model-estimated (that accounts for methane emissions and loss in the atmosphere) calculated trend that fit to the globally average values. (Bottom) Atmospheric CH₄ instantaneous growth rate (ppbv/yr), which is the derivative with respect to the trend curve shown in the first panel.
In the late 1980s, rates of increase fell to about 10 ppbv/yr; they declined even further in the early 1990s before stopping entirely at some locations in 1992. In 1994, global methane growth rates recovered to about 8 ppbv/yr. There are a number of ideas to explain these rapid changes in the growth rate, ranging from suggested reductions in emissions from anthropogenic or natural sources to a slowing in the rate of CH₄ removal. However, the best indications are that the sharp increase and the subsequent dip in the early 1990s are connected to changes in atmospheric chemistry and temperature induced by the 1991 Pinatubo eruption. Yet the cause of the longer-term global decline in methane growth is still not well understood, although it may be that much of the earlier rapid increase in methane emissions from agricultural sources is now slowing down. A recent increase in the trend of methane during the last one and a half years (not shown in Fig. 2) to −10 ppbv/yr is also not fully understood but may be associated with the response of wetlands to increasing global temperatures over the last several years.

1. The Methane Budget

Methane emissions come from a number of different sources, both natural and anthropogenic. One type of human-related emissions arises from biogenic sources from agriculture and waste disposal, including enteric fermentation, animal and human wastes, rice paddies, biomass burning, and landfills. Emissions also result from fossil fuel-related methane sources such as natural gas loss, coal mining, and the petroleum industry. Methane is emitted naturally by wetlands, termite mounds, other wild ruminants, oceans, and hydrates. Based on recent estimates, current human-related biogenic and fossil fuel-related sources for methane are approximately 275 TgCH₄/yr and 100 TgCH₄/yr, or 67% of the total sink. The remainder of the CH₄ is removed through reactions with soil (30 TgCH₄/yr, or −9%) or transport to the stratosphere (40 TgCH₄/yr, or −7%).

C. Nitrous Oxide

Nitrous oxide is a greenhouse gas that on a molecule-to-molecule basis is 200 times more efficient than CO₂ in absorbing infrared radiation. Also, through reactions with excited oxygen atoms, N₂O is the primary source for stratospheric ozone loss. Among the most potent halocarbons in the current atmosphere are the chlorofluorocarbons (CFCs). One

D. Chlorofluorocarbons and Other Halocarbons

Halocarbons are greenhouse gases that can contribute to climate change, and they are also largely responsible for stratospheric ozone loss over recent decades. Because of their dual impact, they are of particular concern to environmental health. Among the most potent halo- carbons in the current atmosphere are the chlorofluoro- carbons CFC-11 (CFCl₃) and CFC-12 (CF₂Cl₂). One
molecule of CFC-11 or CFC-12 in the atmosphere is, respectively, 12,400 and 15,800 times more effective as a greenhouse gas than one molecule of CO₂. With the exception of the naturally occurring portions of CH₃Cl and CH₃Br emissions, all the halocarbons in the atmosphere are man-made. Their inertness and long lifetimes have made them attractive chemicals for use as propellants, refrigerants, fire retardants, and other industrial applications. Natural sources of CH₃Cl from ocean surface waters and wood-rotting fungi may account for as much as half of atmospheric CH₃Cl, with biomass burning (some of which is natural) being the other primary source. Roughly half of the CH₃Br comes from the oceans, while biomass burning and its use as a soil fumigant account for the majority of its remaining emissions into the atmosphere.

Halocarbons containing chlorine and/or bromine are of particular concern with regard to destruction of stratospheric ozone. Bromine and chlorine effectively catalyze ozone destruction cycles. These halocarbons, plus others such as the perfluorocarbons (PFCs) and the hydrofluorocarbons (HFCs) that contain fluorine instead of chlorine, also have the potential to affect climate change since these chemical species characteristically have strong infrared absorption features. Measurements of CFCs and other compounds by the NOAA Climate Monitoring and Diagnostics Laboratory at sites throughout the world are shown in Fig. 3. CFC-11 and CFC-12 have the largest atmospheric concentrations, at 0.26 and 0.53 ppbv, respectively. The tropospheric concentrations of both of these gases were increasing at about 4% per year in the early 1990s, but have now slowed appreciably. The concentrations of CFC-11 and several other controlled halocarbons have already started to decline. The use of these compounds has diminished greatly, with all but essential applications being banned by the beginning of 1996. The atmospheric concentrations of several other halocarbons have, until recently, been growing at an even faster rate than CFC-11 and CFC-12. For example, the concentration of CFC-113 (C₂F₃Cl₃) was increasing about 10% per year in the early 1990s but has also slowed greatly, with a current concentration of about 0.08 ppbv. Abundances of the shorter-lived controlled compound methyl chloroform (CH₃CCl₃), have declined appreciably. HCFC-22 (CHF₂Cl), a refrigerant often found in home air conditioners, has gained increased use as a soil fumigant account for the majority of its remaining emissions into the atmosphere.

Measurements of CFCs and other compounds by the NOAA Climate Monitoring and Diagnostics Laboratory at sites throughout the world are shown in Fig. 3. CFC-11 and CFC-12 have the largest atmospheric concentrations, at 0.26 and 0.53 ppbv, respectively. The tropospheric concentrations of both of these gases were increasing at about 4% per year in the early 1990s, but have now slowed appreciably. The concentrations of CFC-11 and several other controlled halocarbons have already started to decline. The use of these compounds has diminished greatly, with all but essential applications being banned by the beginning of 1996. The atmospheric concentrations of several other halocarbons have, until recently, been growing at an even faster rate than CFC-11 and CFC-12. For example, the concentration of CFC-113 (C₂F₃Cl₃) was increasing about 10% per year in the early 1990s but has also slowed greatly, with a current concentration of about 0.08 ppbv. Abundances of the shorter-lived controlled compound methyl chloroform (CH₃CCl₃), have declined appreciably. HCFC-22 (CHF₂Cl), a refrigerant often found in home air conditioners, has gained increased use as a replacement for CFCs and its concentration has been increasing at about 5–6% per year since 1995. The measured concentrations of several other replacements are also increasing.

All of the fully halogenated chlorofluorocarbons have long atmospheric lifetimes. The lifetime of CFC-11 is about 30 years, while the lifetime of CFC-12 is about 102 years. The atmospheric lifetimes of HFCs (hydrochlorofluorocarbons), HCFCs, and other halocarbons containing hydrogen tend to be much shorter than those of the CFCs. Because of these shorter lifetimes, less of the replacement compounds containing chlorine and bromine reach the stratosphere and they have less effect on ozone than the CFCs.

Bromine is more effective at destroying ozone than chlorine. Therefore, early rapid increases in atmospheric concentrations of bromine-containing halons, most notably Halon-1301 (CF₃Br) and Halon-1211 (CF₂BrCl), have caused concern. Despite their control under the 1997 Montreal Protocol on Substances That Deplete the Ozone Layer, the concentrations of several halons, including H-1211 and H-1301, continue to increase. Primary destruction of these compounds occurs through photolysis, resulting in long atmospheric lifetimes (65 years for H-1301 and 20 years for H-1211). However, these compounds currently have small atmospheric concentrations, about 4 pptv (parts per thousand by volume) or less, and hence their contribution to the absorption of infrared radiation is considered minimal.

E. Other Gases

1. Carbon Monoxide

In addition to its potential effects on health in polluted areas, carbon monoxide affects the oxidizing capacity of the atmosphere through its reactivity. The reaction of carbon monoxide with OH is the primary sink for atmospheric OH. This reaction also yields an additional source of the greenhouse gas CO₂. The global emissions of CO are still poorly understood. Sources of carbon monoxide include incomplete combustion processes (complete combustion yields CO₂ rather than CO). Emissions of CO from fossil fuel combustion peak between 30 and 60 degrees north latitude. More than 70% of the CO from the biomass burning source is emitted in tropical regions. Chemical decomposition of methane and other hydrocarbons is also an important source. It is estimated that about two-thirds of current atmospheric CO results from human activities. Total sources are about 1100 TgC per year with a large uncertainty range. As a result of its short lifetime (2–3 months) and the high spatial variability of its sources, the atmospheric concentration of CO varies greatly in time and space. Annually averaged concentrations of CO peak at about 0.2 ppbv at high northern latitudes; minimum concentrations of about
0.06 ppmv occur throughout most of the Southern Hemisphere. Long-term trends up to 1990 suggest that CO concentrations had been increasing in the Northern Hemisphere at about 1% per year, with little evidence of an increase in the Southern Hemisphere. After 1990, concentrations appear to have decreased.

2. Non-methane Hydrocarbons
Nonmethane hydrocarbons (NMHCs) in the presence of nitrogen oxides (NOx) contribute to the formation of tropospheric ozone and stratospheric H2. They can also react with nitrogen species to produce the atmospheric gas peroxyacetyl nitrate (PAN), which is a long-lived reservoir of reactive nitrogen. Oxidation of NMHCs is an additional source of CO and ultimately CO2.

There are many anthropogenic and natural sources emitting NMHCs into the atmosphere. Anthropogenic sources result from all aspects of human activity, including chemical manufacturing, vehicle exhaust, food
processing, refuse disposal, biomass burning, and energy production. Global emissions from anthropogenic activities are estimated at roughly 140 Tg/yr. Emissions from the natural sources are less well known.

3. Nitrogen Oxides

Emissions of nitrogen oxides have been a cause of concern because of their role as primary pollutants in photochemical smog and their contribution to acid wet and dry deposition. Nitrogen oxides are also important because of their indirect effect on climate through their role in affecting global ozone concentrations. Although NO species are relatively short-lived, they can react chemically with hydrocarbons to produce PAN. The constituent PAN provides a reservoir for nitrogen oxides that can be transported long distances to affect ozone chemistry well downstream from the sources. The lifetime of PAN depends strongly on temperature.

There is significant uncertainty in the sources of nitrogen oxides, with the total source being estimated to be roughly 53 Tg (N)/yr. The largest sources of reactive nitrogen in the troposphere are fossil fuel combustion, biomass burning, lightning discharges, microbial activity in soils, aircraft emissions, and transport from the stratosphere. Emissions of nitrogen oxides from combustion of fossil fuel have increased globally at 1–2% per decade during this century, resulting in increased tropospheric concentrations particularly over continents in the lower atmosphere and in the flight corridors used by commercial aircraft. Nitrogen dioxide is an important absorber of visible solar radiation, and it could affect climate directly if tropospheric and/or stratospheric concentrations continue to increase.

4. Sulfur Gases

Emissions of sulfur dioxide and other gases can result in the formation of aerosols that can affect climate. Aerosols affect climate directly by absorption and scattering of solar radiation and indirectly by acting as cloud condensation nuclei (CCN). A variety of analyses indicate that human-related emissions of sulfur, and the resulting increased sulfuric acid concentrations in the troposphere, may be cooling the Northern Hemisphere sufficiently to compensate for much of the warming expected from greenhouse gases. Volcanic emissions can influence climate for short periods (1 to 3 years) through emissions of sulfur dioxide into the lower stratosphere.

Over half of the sulfur dioxide (SO₂) emitted into the atmosphere comes from human-related sources, mainly from the combustion of coal and other fossil fuels. Most of these emissions occur in the Northern Hemisphere. Analyses indicate that anthropogenic emissions have grown dramatically during this century. Other SO₂ sources come from biomass burning, from volcanic eruptions, and from the oxidation of di-methyl sulfide (DMS) and hydrogen sulfide (H₂S) in the atmosphere. DMS and H₂S are primarily produced in the oceans. Atmospheric SO₂ has a lifetime of less than a week, leading to formation of sulfuric acid and eventually to sulfate aerosol particles. Gas-to-particle conversion can also occur in cloud droplets; when precipitation doesn’t occur soon, the evaporation of such droplets can then leave sulfate aerosols in the atmosphere.

During periods of low volcanic activity, carbonyl sulfide (COS) is thought to be responsible for the maintenance of the sulfuric aerosol layer found in the lower stratosphere. Natural emissions explain most of the COS in the present atmosphere, while the relatively long atmospheric lifetime (about 2 years) of COS explains why much of it reaches the stratosphere before its conversion to sulfuric acid aerosol. However, if sources of COS (or its precursor, CS₂) were to increase dramatically, the background aerosol layer concentration would increase, with significant implications for climate.

III. THE IMPORTANCE OF STRATOSPHERIC OZONE

Ozone, O₃, is composed of three oxygen atoms and is a gas at atmospheric pressures and temperatures. Most of the ozone (about 90%) exists in the stratosphere, the layer of the atmosphere about 10–50 km above the Earth’s surface. The remaining ozone is in the troposphere, the lower region of the atmosphere extending from the Earth’s surface up to roughly 10 km at midlatitudes and 16 km in the tropics. Figure 4 shows a typical integrated column of ozone, referred to as the total ozone column, as a function of latitude and season based on the satellite observations from the TOMS (Total Ozone Mapping Spectrophotometer) instrument. Despite the fact that the primary production of ozone occurs in the tropics and the mid-latitudes, the largest amounts of ozone are found at high latitudes as a result of the pole-ward transport of ozone by atmospheric dynamical processes. The large decrease in ozone over springtime Antarctica, termed the Antarctic ozone “hole,” can also be seen in Fig. 4.

Ozone in the troposphere and stratosphere is chemically identical, but it has very different effects on life
on the Earth depending on its location. Stratospheric ozone plays a beneficial role by absorbing solar ultraviolet radiation (UV-B), thus preventing biologically harmful levels of UV radiation from reaching the Earth’s surface. It is the absorption of solar radiation by ozone that explains the increase in temperature with altitude in the stratosphere. Concerns about increased UV-B from the decreasing levels of ozone have been the driver for policy actions to protect the ozone layer.

Ozone is also a greenhouse gas, with a large infrared absorption band in the atmospheric window, at 9.6 μm. It is the balance between the solar and infrared radiative processes that determines the net effect of ozone on climate. Increases in ozone in the stratosphere above about 30 km tend to decrease the surface temperature as a result of the increased absorption of solar radiation, effectively decreasing the solar energy that would otherwise warm the Earth’s surface. Below 30 km, increases in ozone tend to increase the surface temperature, and the infrared greenhouse effect dominates in this region.

Closer to the Earth’s surface, ozone displays its destructive side. Ozone is a strong oxidizer. Hence, direct exposure to high levels of ozone has toxic effects on human health and plants. Although ozone is a major component of photochemical smog in urban areas, this ozone is generally not thought to be a significant contributor to the global ozone budget. Balloon measurements suggest that tropospheric ozone at the global scale has been increasing.

A. Stratospheric Ozone Trends

Concentrations of ozone in the stratosphere result from chemical production and destruction processes in combination with transport processes. Production of ozone in the stratosphere results primarily from photodissociation of oxygen molecules. The destruction of ozone occurs mainly through catalytic reactions with other gases, such as chlorine and bromine. The total amount of ozone in the stratosphere will remain fairly constant (relative to the well-recognized seasonal variations) as long as there is no change in the destruction rate and the transport of ozone out of the stratosphere. However, increasing inputs of chlorine and bromine into the stratosphere over the past few decades have changed this balance.

Measurements of ozone by satellite and ground-based instruments over the last several decades indicate that stratospheric ozone levels have decreased significantly. Amounts of ozone in the global atmosphere have decreased globally by more than 5% since 1970. Figure 5 shows satellite and ground-based measurements of the change in ozone from 60°S to 60°N latitudes. Satellite measurements began in 1978, while a reasonably representative global network of ground-based sta-
FIGURE 5  Deviations in total ozone with time relative to January 1979 from various ground-based (bottom panel) and satellite measurements (top panel). The data are area-weighted over 60°S to 60°N. Based on the World Meteorological Organization 1998 assessment of ozone depletion.

Significant decreases in the total (integrated column of) ozone are found in both the Northern and Southern Hemispheres at the middle and high latitudes, with no significant change in the tropics. Much larger decreases in total ozone are found at latitudes greater than 60°, particularly in the Southern Hemisphere, as discussed later. Measurements indicate that the total ozone column in the Northern Hemisphere has decreased by 1.3% per decade in the summer months and 2.7% per decade in the winter months since 1969. Satellite and ozonesonde (balloon) data sets indicate that ozone is particularly decreasing in the lower stratosphere, accounting for a major fraction of the trend in total ozone, although there is also significant ozone destruction occurring in the upper stratosphere. At the end of the ozone record in Fig. 5, there is an upturn in ozone associated with the recovery from the effects of the Mount Pinatubo volcanic eruption in 1991. Also, in the top part of Fig. 5, the switch to the new satellite in 1996 resulted in a bias that was not yet corrected at the time of the international assessment in 1998. When this article was written, there was no indication of a recovery in ozone due to the control measures on CFCs and other halocarbons.

Beginning in the late 1970s, a special phenomenon began to occur in the springtime over Antarctica, referred to as the Antarctic ozone “hole.” A large decrease in the total ozone, now over a 60% decrease relative to pre-hole levels, has been observed in the springtime (September to November) above Antarctica. Joe Farman of the British Antarctic Survey and co-authors documented this rapid springtime decrease in Antarctic ozone over the ozone measurement station at Halley Bay, Antarctica, attracting the attention of the scientific community. Decreases in the total ozone column of more than 50% were found compared with historical values observed by both ground-based and satellite techniques. Measurements made in 1987 indicated that more than 95% of the ozone over Antarctica at altitudes from 13 to 22 km had disappeared during September and October. More recently, the Antarctic ozone holes since 1992 have been quite comparable, being the biggest (areal extent) and the deepest (minimum amounts of ozone overhead), with ozone being locally depleted by more than 99% between about 14–19 km in October.

B. Causes and Consequences of Stratospheric Ozone Depletion

The connection between potential environmental effects and man-made chlorofluorocarbons was first raised by Drs. Mario Molina and F. Sherwood Rowland in 1974, when they suggested that chlorine from these compounds could destroy stratospheric ozone. Research findings since then have continued to support the significant role that these compounds play in affecting the global distribution of ozone. In addition, it is recognized that other gases containing chlorine and bromine, which are even more reactive with O₃ than chlorine, are also affecting ozone. To a lesser extent, the increasing concentrations of other gases like CO₂, CH₄, and N₂O are also affecting stratospheric ozone.

The inverse relationship between changes in ozone and UV-B is well established by both theoretical analyses and observations. A number of studies have shown
that the corresponding increase in UV-B at the ground resulting from ozone depletion can lead to increased incidences of skin cancers, cataracts, and other effects on humans and animals.

The recognition of the deleterious effect of chlorine and bromine on ozone spawned international action to restrict the production and use of CFCs and halons to protect stratospheric ozone. These included the 1987 Montreal Protocol on Substances That Deplete the Ozone Layer and the subsequent 1990 London Amendment, 1992 Copenhagen Amendment, and the 1997 Montreal Amendment. The agreements mutually called for reduction of CFC consumption in developed countries. A November 1992 meeting of the United Nations Environment Programme held in Copenhagen resulted in substantial modifications to the protocol because of the large observed decrease in ozone, and called for the phase-out of CFCs, carbon tetrachloride (CCl₄), and methyl chloroform (CH₃CCl₃) by 1996 in developed countries. As part of this, the United States, through the Clean Air Act, has also eliminated the production and import of these chemicals. Production of these compounds is to be totally phased out in developing countries by 2010, while production of halons in developed countries was stopped in 1994. Human-related production and emissions of methyl bromide are not to increase after 1994 in developed countries, with total elimination by 2005.

C. Projected Trends in Ozone

Projected changes in globally averaged total ozone, using several assumptions about future emissions of CFCs, halons, and their replacements, have been evaluated using models of atmospheric processes. With the original Montreal Protocol provisions, there would still have been a significant reduction in total ozone, as much as 15% by 2050 relative to 1980 levels, according to some analyses. It is only under the London and Copenhagen Amendments provisions, which call for the complete phase-out of CFCs and halons and other halocarbons, that the ozone reduction trend is reversed. The largest ozone reductions are reached at about 1998 to 2000. After this, ozone begins to recover, although it is not until about 2050 that the 1980 level of global total ozone would again be expected. Thus, according to these models, it will roughly be the middle of the next century before the chlorine and bromine in the stratosphere are reduced to levels corresponding to those when the Antarctic ozone "hole" first began. Also, without any control measures and assuming unfettered growth, the global mean total column ozone would decrease about 30% by 2050, decreasing further with time.

IV. AIR POLLUTANTS IN THE LOWER ATMOSPHERE

The elevated concentration of oxidants in urban regions is a difficult problem in many parts of the world. In the United States, progressively tighter emissions controls have been implemented since the late 1970s to control production of ozone. Most of these controls have been aimed at emissions of NOₓ and VOCs, particularly the former. Nonetheless, many urban areas still fail to meet the federal standards. In addition, there is growing concern that enhanced levels of ozone concentrations in rural areas downwind of urban areas are resulting in harmful effects on agricultural and forest ecosystems. Part of the difficulty is understanding how emissions of hydrocarbons from the biosphere affect ozone levels in urban areas.

V. THE FUTURE AND ATMOSPHERIC GASES

Changes in climate associated with increasing concentrations of greenhouse gases could have significant effects on the biosphere and on biodiversity. Corresponding changes in the biosphere could affect biogeochemical cycles and the sources and sinks of atmospheric gases, leading to further changes in the climate. Such interactions with the biosphere and with biodiversity are inadequately considered in current climate projections. The effects of such interactions require much more study.

Stratospheric ozone levels should largely recover over the next half century. However, the timing of this recovery will largely depend on the emissions of halocarbons and on the emissions of other gases that influence stratospheric chemistry. Changes in the oxidative capacity could also change the amounts of gases reaching the stratosphere and thus the rate of and extent of recovery. Climatic change would also likely influence stratospheric temperature and winds, thus further affecting the rate of ozone recovery. The long-term impacts of the reduced levels of ozone on the biosphere and on biodiversity remain poorly understood.

Attempts to control pollution in urban regions continue to increase. However, there is still much to learn before such controls can be fully effective. The U.S. National Academy of Sciences has recently stated that too little is known about the transport and deposition of atmospheric gases, including toxins and nutrients, to the biosphere and their interaction with biota.
study of the rates of chemical exchange between the atmosphere and ecosystems will provide better quantitative estimates of atmospheric chemical impacts on the biosphere and biospheric emissions to the atmosphere.

See Also the Following Articles
AIR POLLUTION • CARBON CYCLE • CLIMATE CHANGE AND ECOLOGY, SYNERGISM OF • CLIMATE, EFFECTS OF • GREENHOUSE GASES • NITROGEN AND NITROGEN CYCLE • ULTRAVIOLET RADIATION

Bibliography
AUSTRALIA, ECOSYSTEMS OF

Raymond L. Specht* and Alison Specht†
*The University of Queensland and †School of Resource Science and Management, Southern Cross University

I. Australian Ecosystems
II. Evaporative Aerodynamics
III. Major Plant Communities in Australia
IV. Diagnostic Species in Australian Floristic Groups
V. Biogeographic Regions in Australia
VI. Diagnostic Floristic Groups in Australian Biogeographic Regions
VII. Species Richness of Australian Ecosystems
VIII. Community Diversity of Australian Ecosystems
IX. Conclusions

GLOSSARY

biogeographic region A region that contains a "homogeneous" suite of Floristic Groups sorted by some classificatory program.

community diversity Number of ecosystems and associated plant and animal species per 1° latitude by 1° longitude grid-cell.
diagnostic species Species that emerge from semi-quantitative analysis as being the key species associated with a floristic group.
evaporative aerodynamics Study of the influence of aerodynamics on the development of both leaf structure and foliage distribution throughout a plant community.
floristic group A plant community that contains a "homogeneous" suite of species sorted by some classificatory program.

DISTINCTIVE ECOSYSTEMS FOUND IN AUSTRALIA range from the arid lands to the perhumid evaporative zone in both the temperate south and the tropical north of the continent. After applying the classificatory program TWINSPAN to almost 5000 ecological surveys, the floristic components of Australia's plant communities were sorted into 338 Floristic Groups. Eight biogeographic regions, with 45 subdivisions, were defined. The species richness (the number of species per hectare) of overstory plants is related to the annual biomass production of that stratum. Because annual growth of the understory stratum depends on the transmission of solar radiation through the overstory, species richness of this stratum is higher in tropical than in temperate climates, and in the early stages of secondary succession when the overstory stratum is poorly developed. The species richness of non-arboreal vertebrates parallels that of the understory stratum; species richness of epi-
Australian ecosystems depend on the annual supply of leaf litter, largely from the overstory. The diversity of Floristic Groups (per 1° latitude by 1° longitude grid-cell) is related to the annual biomass production of overstory foliage in each evaporative climate. As annual shoot growth is markedly reduced by the low-phosphate status of the relictual Gondwanan soils of the northern, central, and western parts of the continent, the diversity of major plant communities in these areas is lower than that found on the richer soils of the east.

I. AUSTRALIAN ECOSYSTEMS

Australia is a continental mass lying south of the equator between latitudes 11° and 44°. The climate varies markedly with geographical position: summer rainfall predominates in the north, winter rainfall is characteristic of the south, and the eastern seaboard shows a more general distribution of rainfall throughout the year. Marked evaporative gradients from the humid coastal fringe to the arid inland are found in the south, east, and north of the continent. The structure of the vegetation varies along these evaporative gradients: from the dense rainforests in eastern Australia, to the tall eucalypt forests of the south-east and south-west, to the eucalypt forests/woodlands, with either grassy or heathy understory, distributed throughout the continent. Heathlands and grasslands are found in tropical to temperate climes, the mallee eucalypt vegetation is found on the calcareous soils of the south, and a variety of other vegetation structures are found in the vast arid zone and in the alpine, coastal, and wetland landscapes of the continent.

Most families and many genera now found in the Australian flora developed over 100 million years ago, before the Gondwanan supercontinent began to break up. Vast expanses of infertile lateritic soils developed in the humid/perhumid climate of that time. The onset of aridity in the early Tertiary, some 50 million years ago, caused fragmentation of the humid/perhumid Gondwanan vegetation, enabling its less humid elements to expand and to evolve, eventually to cover much of the continent.

The vegetation of today comprises those taxa that were able to survive as the evaporative climate became drier. A distinctive heathy vegetation developed on the nutrient-poor soils of lateritic origin, characteristic of a large part of the original Gondwanan super-continent. The attributes that enabled most of the Gondwanan flora to survive over the last 50 million years also facilitate their rapid regeneration from underground organs and epicormic buds whenever a disturbance such as a fire occurs.

II. EVAPORATIVE AERODYNAMICS

As most of the continent of Australia experiences a dry season, both the distribution of foliage in the overstory and understory and its eco-physiological attributes result from the growth of a multitude of shoot apices (in both overstory and understory strata) in equilibrium with the evaporative aerodynamics of the atmosphere flowing over and through the plant community (Specht, 1972, Specht and Specht, 1999). The amount of available water affects the vertical growth of each foliage shoot (and the number of essentially identical leaves produced annually), but not the horizontal cover of the foliage throughout the plant community (Specht and Specht, 1989a); this is determined by the evaporative power of the atmosphere, a relatively constant factor from year to year.

After any perturbation, the horizontal cover of the foliage and its eco-physiological characteristics are rapidly restored in both the overstory and understory of an evergreen plant community. The ratio of the actual to potential evapotranspiration (per hectare) from the plant community is linearly related to the amount of water available during each month; the slope of this linear regression is called the Evaporative Coefficient (Specht, 1972). Values of this community-physiological constant have been computed for meteorological stations throughout Australia and the resulting isolines are mapped in Fig. 1.

III. MAJOR PLANT COMMUNITIES IN AUSTRALIA

The great number of tree species (20 to 140) that co-exist in subtropical and tropical plant communities in Australia, from the arid to the perhumid evaporative zones, makes the definition of statistically ‘homogeneous’ suites within these vegetations a challenge, one that is not faced by temperate plant ecologists who deal with simple systems with relatively low numbers of overstory species per hectare (1 to 10 in number). This has necessitated the adoption of computer techniques. TWINSPAN Floristic Groups have been defined in all Australian plant formations (Specht et al., 1993). To accomplish this, species lists reported for plant communities throughout the continent over the last 70 years...
FIGURE 1  Map of Australia showing isolines of the Evaporative Coefficient (k). Climatic zones may be defined as follows.—Arid Zone, \( k < 0.35 \); Semi-arid Zone, \( k = 0.35-0.45 \); Subhumid Zone, \( k = 0.45-0.65 \); Humid Zone, \( k = 0.65-0.75 \); Perhumid Zone, \( k > 0.75 \). (All values of k are \( \times 10^{-2} \) per mm of soil water available per month.) (After Specht, 1972; Specht and Specht, 1999)

were collated in 16 large data banks and arranged according to structural formation (Table I). These formations were then analyzed by TWINSPLAN (M. O. Hill, Ecology 61, 237–249, 1973) on the basis of the presence or absence of a species in each list. TWINSPLAN first sorted tropical/subtropical plant communities in northern Australia from the temperate/montane plant communities in southern Australia. The program then sorted the tropical from the subtropical, and the temperate from the montane. Further subdivision separated the humid plant communities from the subhumid, and so on for plant communities that had developed in different micro-environments. The number of TWINSPLAN Floristic Groups and their subdivisions that were recorded in each 30° latitude by 30° longitude grid-cell are shown in the Conservation Atlas of Plant Communities in Australia (Specht et al., 1995). The diversity of Floristic Groups is greatest in the perhumid zone, and least in the arid evaporative zone.

TWINSPLAN analysis of the sixteen data banks resulted in the definition of 338 overstory Floristic Groups. A further 60 understory Floristic Groups have been recognized in open-structured plant communities (Specht et al., 1993).
### Table I

Structural Characteristics of Major Plant Formations in Australia

<table>
<thead>
<tr>
<th>Vegetation formation</th>
<th>Climatic zone</th>
<th>Stand height (m)</th>
<th>Foliage projective cover (%)</th>
<th>Leaf specific weight (mg cm⁻²)</th>
<th>Understory</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overstory with trees</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closed-forest</td>
<td>Subhumid</td>
<td>20–10</td>
<td>100–70</td>
<td>8–16</td>
<td>Regeneration saplings, shade plants</td>
</tr>
<tr>
<td></td>
<td>Subhumid</td>
<td>&gt;20–10</td>
<td>70–30</td>
<td>10–20</td>
<td>Grassy, heathy</td>
</tr>
<tr>
<td></td>
<td>Semi-arid</td>
<td>3–2</td>
<td>70–30</td>
<td>18–30</td>
<td>Grassy, heathy</td>
</tr>
<tr>
<td></td>
<td>Arid</td>
<td>3–2</td>
<td>30–10</td>
<td>20–50</td>
<td>Grassy, hummock grass, chenopods</td>
</tr>
<tr>
<td><strong>Treeless vegetation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heathland</td>
<td>Semi-arid</td>
<td>2–0.5</td>
<td>70–10</td>
<td>Sclerophyllous</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Montane</td>
<td>0.25</td>
<td>30–10</td>
<td>Sclerophyllous</td>
<td></td>
</tr>
<tr>
<td>Tussock grassland</td>
<td>Semi-arid</td>
<td>1–2</td>
<td>70–30</td>
<td>Herbaceous</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;0.25</td>
<td>30–10</td>
<td>10</td>
<td>Herbaceous</td>
<td></td>
</tr>
<tr>
<td>Hummock grassland</td>
<td>Semi-arid</td>
<td>1–0.5</td>
<td>70–10</td>
<td>Sclerophyllous</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>30–10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a* From Specht et al. (1995) and Specht and Specht (1999).  
*b* Leaf/leaflet area decreases, and leaf specific weight increases, in the 1°C temperature gradient from tropical to subtropical to temperate regions.  
*c* In the tropics and subtropics, these are emergents and strangler figs, or lamas and epiphytes.

### IV. Diagnostic Species in Australian Floristic Groups

The plant species which define the floristic groups (Specht et al., 1993) determined by TWINSPAN in the 29 Australian plant formations (Table I) are listed below. Each floristic group has been given a code: capital letters for the formation (e.g., RF = closed-forest (rainforest), T = open-forest and woodland), followed by numerals. The numerals refer to the dichotomous subdivisions in the TWINSPAN classification: within a formation, groups having consecutive odd–even pairs of numerals are most closely related (e.g., RF35 and RF36). In species-rich formations, such as closed-forests, subdivisions of the primary TWINSPAN groups are shown by the numbers in parentheses.


RF35(140) Drypetes deplanchei  
RF35(141) Endiandra glauca - Polyscias australiana  
RF35(142) Calophyllum australis - Castanospermum australe

RF36(143) Polyalthia nitidissima  
RF36(144) Agathis robusta - Diospyros pentamera  
RF36(145) Myristica insipida

RF36(146) Polyscias australiana - Podocarpus neriifolius

RF37(147) Arctostaphylos alexandrae - Ficus racemosa  
RF37(75) Acronychia acidula - Dendrocnide photinophylla

RF37(148) Arctostaphylos alexandrae - Castanospermum australe

RF37(149) Musgravea heterophylla - Normanbya nor-manbyi

RF38(76) Bellschmidia toram - Endiandra sankeyana

RF38(308) Cryptocarya macquorniana - Flindersia bourjotiana

RF38(309) Balanos australis - Lomatia fraxinifolia  
RF39 Laccospadix australis - Stegnthera macrorrana


RF40(160) Austromyrtus acmenoides - Litsea lefebura-Sloanoea woollsi

RF40(163) Helicia ferruginea - Litsea lefebura

RF40(165) Helicia ferruginea - Litsea lefebura
<table>
<thead>
<tr>
<th>RF40(322)</th>
<th>Rhodamnia argentea - Syzygium francisii</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF40(323)</td>
<td>Cryptocarya obovata - Slaneea woollii</td>
</tr>
<tr>
<td>RF40(324)</td>
<td>Acmena ingens - Zanthoxylum brachyantheum</td>
</tr>
<tr>
<td>RF40(325)</td>
<td>Schizomeria ovata - Argyrodendron acanthodes</td>
</tr>
<tr>
<td>RF41(165)</td>
<td>Cupaniopsis anacardioides</td>
</tr>
<tr>
<td>RF41(167)</td>
<td>Celtis paniculata - Cryptocarya tripilinervis</td>
</tr>
<tr>
<td>RF41(328)</td>
<td>Dendrocnide spp. - Ficus macrophylla - Toona ciliata</td>
</tr>
<tr>
<td>RF41(329)</td>
<td>Acmena smithii - Ficus coronata - Toona ciliata</td>
</tr>
<tr>
<td>RF41(332)</td>
<td>Cryptocarya glaucescens - Mischocarpus pyriformis</td>
</tr>
<tr>
<td>RF42(168)</td>
<td>Orites excelsa - Sloanea woollii</td>
</tr>
<tr>
<td>RF42(169)</td>
<td>Nothofagus moorei - Cryptocarya foveolata - Orites excelsa</td>
</tr>
<tr>
<td>RF42(170)</td>
<td>Nothofagus moorei - Tristaniopsis laurina</td>
</tr>
<tr>
<td>RF42(171)</td>
<td>Doryphora sassafras</td>
</tr>
<tr>
<td>RF43(86)</td>
<td>Acmena smithii - Ficus coronata - Eupomatia laurina</td>
</tr>
<tr>
<td>RF43(87)</td>
<td>Acacia melanoxylon - Acmena smithii</td>
</tr>
<tr>
<td>RF44(88)</td>
<td>Cryptocarya floydii - Pittosporum undulatum</td>
</tr>
<tr>
<td>RF44(178)</td>
<td>Podocarpus elatus - Syzygium oleosum</td>
</tr>
<tr>
<td>RF45(180)</td>
<td>Araucaria cunninghamii - Argyrodendron trifoliatum - Cryptocarya budwillii</td>
</tr>
<tr>
<td>RF46(181)</td>
<td>Murraya ovatifoliolata - Maytenus cunninghamii</td>
</tr>
<tr>
<td>RF46(182)</td>
<td>Aglaia sapidina - Hydriastele wendlandiana</td>
</tr>
<tr>
<td>RF46(183)</td>
<td>Calophyllum sil - Ficus hispida - Ilex arnhemensis</td>
</tr>
<tr>
<td>RF47</td>
<td>Brachychiton chillagoe - Terminalia aridicola</td>
</tr>
<tr>
<td>DS8(16)</td>
<td>Cupaniopsis parvifolia - Diospyros humilis - Owenia venosa</td>
</tr>
<tr>
<td>DS8(17)</td>
<td>Cupaniopsis parvifolia - Maclura cochinchinensis</td>
</tr>
<tr>
<td>DS9(18)</td>
<td>Archidendropsis thozetiana - Cupaniopsis wadsworthii</td>
</tr>
<tr>
<td>DS9(19)</td>
<td>Archidendropsis thozetiana - Brachychiton rupicola</td>
</tr>
<tr>
<td>DS10(20)</td>
<td>Pleiogynium timorense - Brachychiton rupicola</td>
</tr>
<tr>
<td>DS11(21)</td>
<td>Alchornea ilicifolia - Ficus tripilinervis</td>
</tr>
<tr>
<td>DS11(22)</td>
<td>Cupaniopsis parvifolia</td>
</tr>
<tr>
<td>DS11(23)</td>
<td>Cupaniopsis parvifolia - Araucaria cunninghamii</td>
</tr>
<tr>
<td>DS12(24)</td>
<td>Mischocarpus pyriformis - Alyxia ruscifolia</td>
</tr>
<tr>
<td>DS12(25)</td>
<td>Mischocarpus pyriformis - Araucaria cunninghamii</td>
</tr>
<tr>
<td>DS13(26)</td>
<td>Araucaria cunninghamii - Alectryon tomentosus - Cleistanthus cunninghamii</td>
</tr>
<tr>
<td>DS13(27)</td>
<td>Araucaria cunninghamii - Alectryon tomentosus - C. laevigata</td>
</tr>
<tr>
<td>DS14(28)</td>
<td>Cryptocarya macdonaldii - C. laevigata</td>
</tr>
<tr>
<td>DS14(29)</td>
<td>Cryptocarya macdonaldii</td>
</tr>
<tr>
<td>DS15(30)</td>
<td>Araucaria cunninghamii - Calamus muelleri - Wilkiea macrophylla</td>
</tr>
</tbody>
</table>

AUSTRALIA, ECOSYSTEMS OF


<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MRF1 Aglaia sapidina - Hydrastele wendlandiana</td>
</tr>
<tr>
<td>MRF2 Acacia auriculiformis - Carpentaria acuminata</td>
</tr>
<tr>
<td>MRF3 Amentosperma claviflorum - Macaranga involucrata</td>
</tr>
<tr>
<td>MRF4 Calophyllum sil - Ficus hispida - Ilex arnhemensis</td>
</tr>
<tr>
<td>MRF5 Amentosperma claviflorum - Litsea breviumbella</td>
</tr>
<tr>
<td>MRF6 Oudlandeana nove-guineensis - Xanthostemon eucalyptoides</td>
</tr>
<tr>
<td>MRF7 Alsosperma ternata - Calophyllum sil</td>
</tr>
<tr>
<td>MRF8 Alsosperma ternata - Alyxia ruscifolia</td>
</tr>
<tr>
<td>MRF9 Aglaia radula - Diaspyros martiana - Seacome elliptica</td>
</tr>
<tr>
<td>MRF10 Barringtonia acanthurula - Livistona benthamii</td>
</tr>
<tr>
<td>MRF11 Ficus racemosa - Melaleuca argentea</td>
</tr>
<tr>
<td>MRF12a Canarium australiannum - Celtis philippensis</td>
</tr>
<tr>
<td>MRF12b Alstonia spectabilis - Denhamia obscura</td>
</tr>
<tr>
<td>MRF13 Ficus coronata - F. vitres - Glocidium apodogynum</td>
</tr>
<tr>
<td>MRF14 Brachychiton collinus - Callicarpa candicans</td>
</tr>
<tr>
<td>MRF15 Brachychiton spectabilis - Ziziphus quadrilocularis</td>
</tr>
<tr>
<td>RF47</td>
</tr>
<tr>
<td>RF16(132)</td>
</tr>
<tr>
<td>RF16(32)</td>
</tr>
</tbody>
</table>

Semi-deciduous closed-forests—subtropical eastern Australia (P. I. Forster, P. D. Bostock, et al., Queensland Herbarium, Brisbane, 1991)

AUSTRALIA, ECOSYSTEMS OF

Araucaria cunninghamii - Maclura cochinchinensis - Smilax australis

Alectryon forsythii - A. subdentatus - Notelaea micracarpa


Alectryon forsythii - A. subdentatus - Notelaea micracarpa


T187a Eucalyptus leucophloia

T189 Eucalyptus microneura

T346 Eucalyptus campanulata - E. laevopinea - E. cypellocarpa - E. obliqua

T346b Eucalyptus saligna - E. microcorys

T351 Eucalyptus campanulata - E. saligna

T352 Eucalyptus amplifolia - Angophora subvelutina

T360 Eucalyptus chloroclada - Angophora costata subsp. leiocarpa

T361 Eucalyptus chloroclada - E. sideroxylon

T362 Eucalyptus microcarpa

T363a Eucalyptus microcarpa - E. populnea

T364 Eucalyptus populnea - E. orgadophila

T365 Eucalyptus melanophloia

T366 Eucalyptus cephalocarpa - E. globoidea

T367 Eucalyptus obliqua

T368 Eucalyptus cullenii - E. shirleyi - Corymbia erythrophloia

T369 Corymbia papuana - C. erythrophloia

T371 Corymbia papuana - C. polycarpa - Eucalyptus platyphylla

T372 Eucalyptus tectifica - Corymbia grandifolia - C. foelscheana

Eucalypt open-forests and woodlands—subtropical eastern Australia (R. L. Specht and D. W. Drake, unpubl. data)

T346 Eucalyptus campanulata - E. laevopinea - E. cypellocarpa - E. obliqua

T346b Eucalyptus saligna - E. microcorys

T351 Eucalyptus campanulata - E. saligna

T352 Eucalyptus amplifolia - Angophora subvelutina

T360 Eucalyptus chloroclada - Angophora costata subsp. leiocarpa

T350 Eucalyptus piperita - E. sieberi - Corymbia gummifera

T40 Eucalyptus cordata

T664 Eucalyptus nitida

T665 Eucalyptus cephalocarpa - E. globoidea

T666 Eucalyptus viminalis

Eucalypt open-forests and woodlands—montane south-eastern Australia (R. L. Specht and D. W. Drake, unpubl. data)

T8 Eucalyptus coccifera

T106 Eucalyptus dealbata - E. macrocarpa subsp. nov.

T706 Eucalyptus pilularis - Corymbia gummifera - Angophora subsp.

T707 Eucalyptus paniculata - E. proptinqua

Eucalypt open-forests and woodlands—montane south-eastern Australia (R. L. Specht and D. W. Drake, unpubl. data)

Eucalyptus coccifera

Eucalyptus dealbata - E. macrocarpa subsp. nov.

Eucalyptus pilularis - Corymbia gummifera - Angophora subsp.

Eucalyptus paniculata - E. proptinqua

Eucalyptus dealbata - E. macrocarpa subsp. nov.

Eucalyptus pilularis - Corymbia gummifera - Angophora subsp.

Eucalyptus paniculata - E. proptinqua

Eucalyptus dealbata - E. macrocarpa subsp. nov.

Eucalyptus pilularis - Corymbia gummifera - Angophora subsp.

Eucalyptus paniculata - E. proptinqua

Eucalyptus was formerly regarded as a subgenus of Eucalyptus.
T79 Eucalyptus brookeriana - E. delegatensis
T148 Eucalyptus obliqua
T149 Eucalyptus delegatensis - E. regnans - E. obliqua
T156 Eucalyptus delegatensis - E. pauciflora - E. rubida
T157 Eucalyptus delegatensis - E. pauciflora - E. rodwayi
T334 Eucalyptus pauciflora - E. rubida
T335 Eucalyptus pauciflora - E. stellulata

Eucalypt open-forests and woodlands—temperate south-western Australia (R. L. Specht and A. J. M. Hopkins, unpubl. data)

T801 Melaleuca leucadendra - M. viridiflora


H7 Corymbia ferruginea - Jacksonia odontoclada - Plocotrachne pungens
H7a Calytrix estipulata - Grevillea pteridifolia - Persoonia junciata
H12 Asteromyrtus lyxicephala - Jacksonia thesioidea
H13 Neofabricia myrtifolia - Neoeospera bambusa
Baecfrut Baeckea frutescens
Leptpurp Leptospermum purpureascens


H40 Allocasuarina littoralis - Alphitonia excelsa
Heathlands and related shrublands—temperate south-eastern Australia (R. L. Specht and W. E. Drake, unpubl. data)

- Melaleuca nodosa - Gahnia sieberana - Blechnum indicum
- Leptospermum polygalifolium - Phyllota phylicoides
- Leptospermum polygalifolium - Platysace linearifolia
- Xanthorrhoea spp.

Heathlands and related shrublands—temperate south-eastern Australia (R. L. Specht and W. E. Drake, unpubl. data)

- Melaleuca uncinata
- Banksia ornata - Xanthorrhoea caespitosa - Allocasuarina pusilla
- Xanthorrhoea spp.


- Carex gaudichaudiana - Myriophyllum pedunculatum
- Carex gaudichaudiana
- Baeckea gunniana - Epacris breviflora - Carex gaudichaudiana
- Baeckea gunniana - Epacris paludosa - E. lanuginosa
- Oxylobium alpestre - Olearia phlogopappa - Phebalium ovatifolium
- Grevillea australis - Epacris paludosa - Poa spp.
- Leucopogon montanus - Poa spp.
- Podocarpus lawrencei
- Microtrobos niphophilus - Epacris serpyllifolia
- Abrotanella forsterioides - Richea scoparia
- Carpha rodwayi - Oreobolus pumilio - Phyllobium colensoi
- Diselma archeri - Microcachrys tetragona
- Gaultheria depressa - Senecio leptocarpus
- Epacris serpyllifolia - Orites revoluta - Richea sprengelioides
- Epacris serpyllifolia - Richea scorparia - Senecio leptocarpus
- Eucalyptus vernicosa - Nothofagus cunninghamii


- Allocasuarina humilis - Callithamnus quadri fragulus - Dryandra nivea
- Allocasuarina humilis - Dryandra nivea - Hakea lissocarphe
- Banksia attenuata - B. menziesii
- Allocasuarina acutivalvis

Tusock grasslands (R. L. Specht and M. P. Bolton, unpubl. data)

- Astrebla elymoides - Isoelema membranaceum
- Allocasuarina cinerea
- Allocasuarina novifolia
- Allocasuarina acutivalvis

Acacia vegetation—subhumid, subtropical eastern Australia (R. L. Specht, M. P. Bolton, and A. Specht, unpubl. data)

- Acacia harpophylla - Casuarina cristata
- Acacia acutifolia - Acacia catenulata
- Acacia strickleyi
- Acacia cambagei
- Acacia pendula - Atriplex nummularia
- Acacia argyrodendron
- Acacia sparsiflora - A. burrowii
- Acacia sutherlandii - Astrebla spp.
- Acacia aneura - Eucalyptus populnea - Eremophila mitchellii
- Eucalyptus terminalis - Acacia ligulata - A. dictyophleba
- Eucalyptus terminalis - Acacia dictyophleba - Senna artemisioides
- Acacia georginae
- Acacia ligulata - Eucalyptus populnea - Eremophila mitchellii
- Acacia aneura - Senna spp. - Eremophila spp.
- Acacia acradenia - Senna artemisioides
- Acacia kempeana - A. aneura
- Acacia hampsleyana - A. aneura
- Acacia acradenia - A. tenissima - Senna artemisioides
- Acacia gunnii
- Eucalyptus gamophylla - Acacia spp.
- Eucalyptus gunnii
DA62a Eucalyptus gongylocarpa - Triodia basedowii
DA63a Eucalyptus gamophylla - Acacia spp. - Thryptome mene maisonnevii
DA100g Eucalyptus dichromophloia - Acacia lysiphloia - Acacia spp.
DA101c Acacia pachycarpa - Eucalyptus dichromophloia
Casdec Allocasuarina decaisneana - Triodia basedowii
DAstow Acacia stowardii - Senna spp. - Eremophila spp.
Erem Eremophila spp. - Dodonaea spp. - Senna spp. - Acacia spp.
Hummock grasslands—Australian Arid Zone (R. L. Specht and E. E. Hegarty, unpubl. data)
HG1 Triodia pungens
HG2 Triodia basedowii
HG3 Triodia wiseana
HG4 Triodia intermedia - Eriachne obtusa
HG5 Triodia inutilis - Triodia spp.
HG6 Plectrachne schinzii
HG7 Triodia clandoni - T. irritans
HG8 Triodia basedowii - Zygochloa paradoxa
HG9 Triodia irritans
HG10 Triodia hubbardii
Chenopod low shrublands—southern Australian Arid Zone (R. L. Specht and A. Specht, unpubl. data)
C4 Atriplex vesicaria - Maireana sedifolia
C7 Myoporum platycarpum - Atriplex vesicaria
C13 Alectryon oleifolius - H11006 Atriplex vesicaria - H11006 Maireana spp.
C20 Halosarcia spp. (inland salt marsh)
C21 Atriplex vesicaria - Maireana aphylla - M. planifolia
C22 Casuarina pauper - Maireana pyramidalis
C23a Atriplex vesicaria - A. nummularia - Acacia pendula
C24 Acacia papycarpa - Maireana sedifolia
Aquatic vegetation—tropical and subtropical northern Australia (R. L. Specht and M. P. Bolton, unpubl. data)
A12 Monochoria cyanea - Potamogeton tricarinatus
A15/26 Caldesia oligococca - Eleocharis dulcis - Nymphaea gigantea
A27 Caldesia oligococca - Nymphaea gigantea
A28 Typha domingensis - Caldesia oligococca
A29 Typha domingensis
Restio Leptocarpus spathaceus - Restio tetraphyllus
A16 Eleocharis acuta - Isoetes fluitans - Lepilaena spp.
A17 Lemna minor - Isoetes muelleri
A18/19 Eleocharis acuta - E. sphacelata
A40/45 Eleocharis acuta - E. sphacelata - Typha spp. - Marilea sp.
A41 Eleocharis acuta - E. sphacelata - Typha spp. - Myrio-phylhum sp.
A42 Eleocharis acuta - E. sphacelata - Typha spp.
A46 Eleocharis acuta - Typha spp. - Potamogeton crispus
A47 Eleocharis acuta - Typha spp.
Baumea Baumea juncea - B. rubiginosa
Gahnia Gahnia filum - G. trifida
CD8 Olearia axillaris - Capparis spinosa - Spinifex longifolius
CD12 Banksia integrifolia - Casuarina equisetifolia - Spinifex sericeus
CD13 Calophyllum inophyllum - Spinifex sericeus
CD14 Calophyllum inophyllum - Casuarina equisetifolia - Spinifex longifolius
CD15 Casuarina equisetifolia - Argasia argentea - Spinifex longifolius
CD18 Olearia axillaris - Acacia cyclops - Spinifex hirsutus
CD19 Allocasuarina verticillata - Olearia axillaris - Spinifex sericeus
CD20 Banksia integrifolia - Leptospermum laevigatum - Leucopogon sp.
CD21 Banksia integrifolia - Leptospermum laevigatum - Myopororn sp.
CD22 Banksia integrifolia - Leptospermum laevigatum - Cunapteris sp.
CD23 Banksia integrifolia - Casuarina equisetifolia - Iso-lepis nodosa
Pisonia Pisonia grandis
CW3 Avicennia marina - Aegiceras corniculatum - Halosarcia indica
CW3d/10d Casuarina glauca
CW6 Rhizophora stylosa - Avicennia marina - Aegiac-
Itis annulata
Nypa Nypa fruticans
Melaleuca Melaleuca acacioides
V. BIOGEOGRAPHIC REGIONS IN AUSTRALIA

The presence or absence of TWINSPAN Floristic Groups in each 1° latitude by 1° longitude grid-cell of the Australian continent was analyzed using the PATN micro-computer Pattern Analysis program. The symmetric form of the Kulczynski measure with $\beta = 0.1$ was used to determine the association between community groups in each grid-cell. This has been found to produce the most useful classification for such ecological applications (D. P. Faith, P. R. Minchin, and L. Belbin, *Vegetatio* 69, 57–68, 1987) and in this case produced the most ‘sensible’ results. The results are plotted as a dendrogram in Fig. 2. The resulting distribution of the biogeographic regions identified by the foregoing analysis is plotted in Fig. 3.

The first subdivision of the Kulczynski Symmetric dendrogram (see Fig. 2) separated the arid and semi-arid biogeographic regions of Western Australia, Northern Territory, South Australia, and north-western Victoria. Three biogeographic regions (AN2–4) were defined in the northern section of the arid zone, an area that receives mostly summer rainfall. The biogeographic regions of Central Australia (AC) and the Ashburton (AN1) of Western Australia receive both summer and winter rainfall. Although summer rains occur erratically, winter rainfall is predominant in the southern arid zone (A1a–b, A2a–b, A3, A4a–b) and in the semi-arid climatic zone (SW1–2 and S1–3). The TWINSPAN Floristic Groups on the mid-Tertiary limestone sediments from the Nullarbor Plain (A3) across South Australia (S1 and S2) into north-western Victoria (S3) are closely related.

The complex of TWINSPAN Floristic Groups (A1a–b) in the southern section of the arid zone of Western Australia are closely interrelated (and probably should not be separated); they are affiliated with the vegetation of the Ashburton region (AN1), where summer rainfall predominates. The northern section (A2a–b) of the arid zone in South Australia is distinctly separated from the southern arid vegetation (A4a–b) in which outliers of the mallee open-scrub (dominant in the semi-arid climate of S1–3) are to be found.

The second subdivision of the Kulczynski Symmetric dendrogram differentiated both the vegetation of the humid/perhumid section of south-western Western Australia (SW1–2, with a Mediterranean-type climate) and all the biogeographic regions in eastern Australia (where rainfall is evenly distributed throughout the year). Biogeographic regions were defined in the humid/perhumid climates of Tasmania to south-eastern Queensland (T, Tasmania; SV, Southern Victoria; and E1–3, Eastern Australia). Subhumid biogeographic regions were defined in central Queensland (CQ1–3) and in central New South Wales (CN), with related biogeographic regions (BT, the Barkly Tableland; CC, the Channel Country; and A5, Far Western New South Wales) in the semi-arid to arid zones of both states.

The third subdivision of the Kulczynski Symmetric dendrogram defined the biogeographic regions in northern Australia, where summer rainfall alternates with winter drought. The biogeographic regions of Cape York Peninsula (NE1) and the Gulf of Carpentaria (NE2) in Queensland are closely related to the wet tropical region (NE3) that extends southwards along the coast toward the tropic of Capricorn.

The TWINSPAN Floristic Groups of the humid/sub-humid section of the Kimberley (N3) and the Top End of the Northern Territory (N1–2) are interrelated and were differentiated from the inland, semi-arid/sub-humid biogeographic regions (NI1–3) across northern Australia and the Hamersley Ranges (HR) in Western Australia.

VI. DIAGNOSTIC FLORISTIC GROUPS IN AUSTRALIAN BIOGEOGRAPHIC REGIONS

Australian biogeographic regions (shown by alphanumeric codes in Figs. 2 and 3) were defined using the classificatory program PATN on the basis of the spatial distribution of the TWINSPAN Floristic groups. The PATN regions and their associated TWINSPAN groups are shown below.

N1 (humid north-western W.A.) — RF34(69), MRF12a, T371, T372
N2 (humid N.T.) — MRF12a, T187a, T371, T372, T373a, G36, H7
N3 (humid Cape York Peninsula and Gulf, Qld.) — RF35(140), RF35(141), T370, T372, T373a, G18, H12, H13
N4 (Broome-Derby, W.A.) — RF16(132), T372

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>Humid north-western Western Australia</td>
</tr>
<tr>
<td>N2</td>
<td>Humid Northern Territory</td>
</tr>
<tr>
<td>N3</td>
<td>Humid Cape York Peninsula and Gulf</td>
</tr>
<tr>
<td>N4</td>
<td>Broome-Derby, Western Australia</td>
</tr>
</tbody>
</table>
FIGURE 2  Dendrogram showing the relationships of the biogeographic regions of Australia derived from a Kulczynski Symmetric Pattern Analysis of the major TWINSPAN Floristic Groups recorded in 1° latitude by 1° longitude grid-cells throughout Australia. (From Specht et al., 1999; Specht and Specht, 1999.)

N5 (southern Kimberley region, W.A.)—RF16(132), T187a, T371, T372
N6 (subhumid N.T.)—T187a, T371, T372, H7
N7 (semi-arid N.T.)—T187a, T372, DA85b
N8 (Pilbara region, W.A.)—T372, DA100g
NE1 (subhumid Cape York Peninsula, Qld.)—T188, T189, T368, T369, T371, T714, T801, G18
NE2 (perhumid north-eastern Qld.)—RF35(142), RF36(144), RF36(145), T368, T369, T371, T710, T712, T713, T714, T715, T801
Figure 3. Biogeographic regions of Australia derived from a Kulczynski Symmetric Pattern Analysis of the major TWINSPAN Floristic Groups recorded in 1° latitude by 1° longitude grid-cells throughout Australia. (From Specht et al., 1995; Specht and Specht, 1999.)

NE3 (subhumid north-central highlands, Qld.)—T188, T189, T368, T369, T714, DA8b
NE4 (subhumid central Qld.)—RF46(184), T188, T364, T352b, T369, T714, T716a, T717, HW11, DA8a
NE5 (subhumid south-eastern Qld.)—T188, T369, T363a, T364, T365, T716a, T717, Calcolu, HW11
BT (Barkly Tableland, Qld.)—T188, G8, G19
CC (Channel Country, Qld.)—T188, DA18, DA19, C20
SE1 (humid south-eastern Qld.)—RF45(180), RF41(332), T354, T707, T710, H82, H166, H167
SE2 (perhumid Border Ranges, N.S.W.—Qld.)—RF42(169), RF40(322), RF40(325), T346, T715, T710, H82, H166, H167
SE3 (central to northern coasts and highlands, N.S.W.)—T334, T347, T354, T710, T716b, H82, H166, H167
SE4 (south-eastern and montane N.S.W.—Vic.)—RF25, T38, T79, T716b, T334, T347, T348, T349, T350, AV4, AV15, H170
SE5 (south-western to south-eastern Vic.)—RF24, T331, Euclyt, T341, T342, T666, T667, G69, H170, H171
SE6 (Tasmania)—RF7, RF13, T36, T148, T149, Euclyt, T666, T667, AV14, AV13, H170, H171
SE7 (northern tablelands, N.S.W.)—T338, T339, T362, T363a, T717, Calcolu
SE8 (southern and central tablelands, N.S.W.)—T190, T339, T362, T363a, T716b, Calcula
SE9 (humid Mt. Lofty Ranges and South East District, S.A.)—T190, T328, T329, T330, T362, T363a, T366a, T667, H170
S1 (Nullarbor Plain, S.A.)—M13, M45, C4, C20
S2 (northern Eyre and Yorke Peninsulas, S.A.)—M25, M45, C24, DA14a, H44b
S3 (subhumid S.A.)—T190, T328a, T329, T330, T366a, T366b, T667, M10, M25, M45, H44a, H44b
SW1 (humid to perhumid south-western W.A.)—SW4a, SW20, SW43a, SW47, SW84, SW85
SW2 (subhumid to semi-arid south-western W.A.)—SW22, SW46, SW96a, SW97a, SW100, SW196
A1a (south-eastern Pilbara District, W.A.)—C20, DA14a, DA22a, DA31, DA62, DA63a
A1b (southern part of Murchison District to Great Victoria Desert, W.A.)—C20, DA22a, DA62, SW197
A1c (northern part of Murchison District to Great Victoria Desert, W.A.)—C20, DA22a, DA62, DA63a
A1d (southern N.T.)—C20, DA41a, DA46a
A1e (Great Victoria Desert, S.A.)—C20, C7, C13, C20, C21, DA44a
A2a (Great Sandy Desert, W.A.)—C20, DA100g, DA101c, HG6
A2b (Tanami Desert, N.T.)—C20, DA47, HG6
A3a (northern S.A.)—C20, DA21, DA41a, DA46a
A3b (Simpson Desert, S.A.)—C20, HG2, HG6
A4a (Exmouth Gulf, W.A.)—C20, DA14a
A4a2 (north-eastern Wheat Belt, W.A.)—C20, DA14a, SW100, SW196
A4b (Goldfields, W.A.)—C20, DA14a, DA62, SW26, SW31, SW90a, SW101b
A5a (Nullarbor Plain to Flinders Ranges, S.A.)—C4, C7, C13, C20, C21, DA44a
A5b (North East District, S.A.)—C4, C7, C13, C20, C22, DA44a, T366a
A6 (western N.S.W.)—C4, C20, C22, C23a, DA41a, T366a

VII. SPECIES RICHNESS OF AUSTRALIAN ECOSYSTEMS

The potential for annual foliage growth increases in the 10°C temperature gradient from temperate to tropical Australia (Table II). The numbers of both stems and species recorded in the overstory (per hectare) increase exponentially (Fig. 4a; after Specht and Specht, 1993, 1994).

### TABLE II

<table>
<thead>
<tr>
<th>Climate</th>
<th>Stand height (m)</th>
<th>Stand density (stems ha⁻¹)</th>
<th>Foliage project cover (%)</th>
<th>Annual shoot growth (kg ha⁻¹)</th>
<th>Overstory (spp. ha⁻¹)</th>
<th>Understory (spp. ha⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical</td>
<td>30–(49)–24</td>
<td>&gt;3000–2100</td>
<td>100–65</td>
<td>12,900</td>
<td>140</td>
<td>150</td>
</tr>
<tr>
<td>Perhumid</td>
<td>24–17</td>
<td>2000–800</td>
<td>65–49</td>
<td>7620</td>
<td>35</td>
<td>50</td>
</tr>
<tr>
<td>Subhumid</td>
<td>17–14</td>
<td>800–450</td>
<td>45–35</td>
<td>5700</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Semiarid</td>
<td>14–10</td>
<td>450–200</td>
<td>35–25</td>
<td>4800</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>Arid</td>
<td>&lt;10</td>
<td>&lt;200</td>
<td>&lt;25</td>
<td>3600</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Subtropical</td>
<td>30–(50)–22</td>
<td>4500–1200</td>
<td>100–65</td>
<td>10,300</td>
<td>90</td>
<td>120</td>
</tr>
<tr>
<td>Perhumid</td>
<td>22–13</td>
<td>1100–300</td>
<td>65–45</td>
<td>6500</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td>Subhumid</td>
<td>15–12</td>
<td>500–300</td>
<td>45–35</td>
<td>4400</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>Semiarid</td>
<td>12–6</td>
<td>300–200</td>
<td>35–25</td>
<td>3500</td>
<td>6</td>
<td>33</td>
</tr>
<tr>
<td>Arid</td>
<td>≤8</td>
<td>≤200</td>
<td>≤25</td>
<td>2800</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>Temperate</td>
<td>30–(60)–20</td>
<td>350–400</td>
<td>100–65</td>
<td>4800</td>
<td>10</td>
<td>43</td>
</tr>
<tr>
<td>Perhumid</td>
<td>20–14</td>
<td>400–300</td>
<td>65–45</td>
<td>2700</td>
<td>5</td>
<td>43</td>
</tr>
<tr>
<td>Subhumid</td>
<td>14–11</td>
<td>300–230</td>
<td>45–35</td>
<td>1000</td>
<td>4</td>
<td>44</td>
</tr>
<tr>
<td>Semiarid</td>
<td>11–7</td>
<td>250–200</td>
<td>35–25</td>
<td>1100</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>Arid</td>
<td>≤7</td>
<td>≤200</td>
<td>≤25</td>
<td>600</td>
<td>2</td>
<td>40</td>
</tr>
</tbody>
</table>

* Stand heights of emergent eucalypts, with high nitrate-reductase activity, at the edge of rainforest stands.
* The species richness of understories in perhumid to humid climates contains a mixture of regenerating overstory species as well as shade-loving vascular plants and cryptogams.
Figure 4 (a) Species richness (number of species per hectare) of the overstory increases exponentially as annual shoot growth of this stratum increases. Species richness ($N$) increases exponentially (1) in the evaporative gradient from the arid zone (Evaporative Coefficient $k = 0.35 \times 10^{-3}$) to the perhumid zone (Evaporative Coefficient $k = 1.0 \times 10^{-3}$), and (2) in the 10°C temperature gradient from temperate to subtropical to tropical Australia. (From Specht and Specht, 1993.) Warm temperate

\[
\ln(OverstoryN) = 0.21 - 2.05k (n = 16, r^2 = 0.71); Subtropical OverstoryN = 0.01 + 4.95k (n = 19, r^2 = 0.93); Tropical OverstoryN = 0.92 + 4.00k (n = 29, r^2 = 0.98). \]

(b) The low amount of solar radiation that penetrates the overstory to reach the understory in the temperate latitudes of Australia has little effect on species richness (number of species per hectare) of this stratum. In the tropics and subtropics, the relationship of species richness of the understory to annual shoot growth (of the upper stratum) is similar to that shown by the overstory. (After Specht and Specht, 1993; 1994; 1999)

Photosynthesis in overstory leaves, and hence potential annual foliage growth, depends primarily on the amount of solar radiation incident on the plant community, a level that declines from tropical to temperate Australia (Specht and Specht, 1999). In the open-structured plant communities of the wet–dry tropics and sub tropics, a high percentage of direct-beam solar radiation reaches the understory through the gaps in the overstory canopy; species richness (the number of species per hectare) of the understory is higher than that of the exposed overstory (Fig. 4b), and follows the exponential curve shown for the overstory in Fig. 4a. In more temperate latitudes, solar radiation reaches the Earth at a lower angle, and consequently the amount of radiation that reaches the understory is low (Specht et al., 1992; Specht and Specht, 1999). Species richness of the understory in temperate latitudes increases only slightly from the semi-arid to the humid zone, with the proportion of cryptogamic species rising strongly in the most humid climates.

On nutrient-poor soils (residual lateritic soils, arenaceous, acid granitic, and serpentinitic soils, Quaternary sands, etc.), common in many parts of Australia, the establishment of a symbiotic association of mycorrhizae with seedling rootlets of Myrtaceae is reduced as available soil water decreases (J. P. Burrell, Aust. J. Bot. 29, 747–764, 1981), thus affecting the density of stems in the overstory. As soil depth decreases, less soil water is available for evapotranspiration, the transport of phosphate and other nutrient ions in the transpiration stream is reduced, and the growth of the overstory is limited (Banannoll et al., 1997)—thus the foliage projective cover of the overstory, the annual shoot growth (per hectare), and the stand height are all reduced (Fig. 5a). As more solar radiation penetrates the gaps between the trees due to the reduction in overstory growth, the foliage projective cover of the understory increases; the sum of the foliage covers of overstory and understory strata remains a constant in equilibrium with the evaporative aerodynamics (Fig. 1) of the area (Fig. 5b; Specht and Morgan, 1981; Specht, 1983; Specht et al., 1990). The species richness of the overstory on nutrient-poor soils is low, below that on higher-fertility soils in the same evaporative regime (Specht and Specht, 1998b), whereas the species richness of the understory increases.
as more solar radiation penetrates the overstory (Fig. 5c).

The shading effect of the overstory on the species richness in the understory has been demonstrated when long-lived shrubby species, such as Banksia (Proteaceae), regenerate from seed or resprouts in a secondary succession following the fire that regularly razes Australian heathlands (Fig. 6). Application of phosphate fertilizer to these nutrient-poor ecosystems results in more rapid growth of overstory shrubs to the detriment of the understory species (Specht and Specht, 1988c).

The species richness of vascular plants in both overstory and understory strata is related to the amount of solar energy fixed as annual shoot growth. The community-physiological processes that produce these predictable values of alpha diversity need urgent scientific investigation. The species richness of vascular plants found in the understory stratum (Fig. 7) and thus must also be related to the potential energy fixation (Branthwaite et al., 1985; P. C. Catling, in Specht, 1988; Cody, 1994a, 1994b; Specht and Specht, 1999). It would also appear from the limited data available that the species richness of epigaeic invertebrates collected in pit-fall traps decreases as the annual input of leaf litter from the overstory decreases from the humid to the semi-arid climatic zone in the Mediterranean-type climate of southern Australia (Table III; after Greenslade and Majer, 1985, and in Specht, 1988).
AUSTRALIA, ECOSYSTEMS OF...

TABLE III

<table>
<thead>
<tr>
<th>Locality</th>
<th>Vegetation formation</th>
<th>Evaporative Coefficient</th>
<th>Annual leaf litter (kg ha(^{-1}))</th>
<th>Epigaeic invertebrates (no. groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dwellingup, W.A. (32°43′S, 116°04′E)</td>
<td>Eucalypt open-forest</td>
<td>0.33 × 10(^{-2})</td>
<td>1300</td>
<td>19</td>
</tr>
<tr>
<td>Kuitpo, S.A. (33°13′S, 138°47′E)</td>
<td>Eucalypt woodland</td>
<td>0.45 × 10(^{-2})</td>
<td>1050</td>
<td>13</td>
</tr>
<tr>
<td>Wyperfield, Vic. (35°35′S, 142°00′E)</td>
<td>Eucalypt open-scrub</td>
<td>0.42 × 10(^{-2})</td>
<td>930</td>
<td>12</td>
</tr>
<tr>
<td>Wyperfield, Vic. (35°35′S, 142°00′E)</td>
<td>Heathland</td>
<td>0.40 × 10(^{-2})</td>
<td>850</td>
<td>10</td>
</tr>
</tbody>
</table>

*a Invertebrate groups—Acarina, Araneae, Isopoda, Amphipoda, Diplopoda, Chilopoda, Collembola, Insecta (Thysanura, Blattodea, Isoptera, Dermaptera, Orthoptera, Psocoptera, Hemiptera, Homoptera, Thysanoptera, Collembola, Deuterostomia, and Hymenoptera including Formicidae).

*b After Greenslade and Majer (1985) and Greenslade and Majer, in Specht (1988).

*c Monthly actual/potential evapotranspiration per millimeter of available water.

related with the optimal annual shoot growth for the evaporative climate of the area (Fig. 8a). The number of vertebrate species also appears to be correlated with the optimal shoot growth of plant communities in the grid-cell (Fig. 8b).

As foliage growth is reduced by the dearth of phosphate ions available in the nutrient-poor lateritic soils of northern, central, and western Australia [also in the Fleurieu Peninsula (35°34′S, 138°20′E) and on Kangaroo Island (35°30′S, 137°13′E) of South Australia and in the Brisbane Ranges (37°55′S, 144°20′E) of Victoria], stem density and species richness of the overstory trees

FIGURE 7

(a) The number of species of small mammals recorded in ecosystems in tropical and temperate Australia plotted against the number of species of vascular plants recorded in the understory of each plant community. (b) The number of species of lizards, snakes, and frogs recorded in ecosystems in temperate Australia plotted against the number of species of vascular plants recorded in the understory of each plant community. (After Specht and Specht, 1999; Braithwaite et al., 1985 and R. C. Catling in Specht, 1988)
IX. CONCLUSIONS

The vegetation of Australia in the Cretaceous–Early Tertiary, fifty to a hundred million years ago, had developed on infertile lateritic soils which extended across the Gondwanan super-continent of which Australia was part. The tropical and temperate rainforests, savanna and healthy floras of Australia therefore show strong family and generic (sometimes specific) affinities with the floras of India, Africa and, to a lesser extent, with South America (Specht, 1981).

Although Australia was at this time contiguous with Antarctica at latitude 60°–65° S, palaeo-oxygen analyses of sea-bed sediments indicate that the climate was subtropical with a mean annual temperature of 19.5°C (Specht et al., 1992). Increasing aridity during the Tertiary, as well as the movement of the Australian Tectonic Plate northward towards the equator at 66 mm per year, imposed great stresses on the Australian vegetation. Remnants of the Gondwanan flora persisted as fragments in more humid habitats. In addition, a great

![Image](image-url)
diversity of taxa evolved during the Early Tertiary to occupy the greater part of the continent which today experiences varying degrees of seasonal drought (Specht, 1981; Specht and Specht, 1990). The convergence of the Australian Plate with the Sunda Plate about 15 million years ago allowed some migration of the Australian flora northward onto Sundaland, but little, if any, migration southward from southeast Asia.

The evaporative and frictional forces of the atmosphere, flowing over and through the vegetation across the Australian continent, moulded foliage distribution and structure within all plant communities; distinctive plant formations resulted from the perhumid to the arid evaporative zone, from the tropical to the temperate region. Species richness (per hectare) of vascular plants in the overstory of Australian plant communities is correlated with the annual shoot growth (per hectare) of that stratum. Similarly, the species richness of vascular plants in the understory is related to annual shoot growth in that stratum—an attribute influenced by the amount of solar radiation transmitted through the overstory stratum.

The species richness (per hectare) of non-arboreal vertebrates is correlated with species richness of the understory stratum. Species richness of epigeic invertebrates appears to depend on the annual input of leaf litter from the overstory foliage. The diversity of TWINSPAN Floristic Groups and associated vertebrates is also correlated with the optimal annual shoot growth (per hectare) of the overstory stratum of Australian plant communities.

The community-physiological processes that determine the species richness of vascular plants, vertebrates, and invertebrates and the diversity of Floristic Groups need urgent scientific investigation before they will be better understood.

See Also the Following Articles

AFRICA, ECOSYSTEMS OF • ANTARCTIC ECOSYSTEMS • ASIA, ECOSYSTEMS OF • SOUTH AMERICA, ECOSYSTEMS OF

Bibliography


I. Recognizing Gaps in Knowledge
II. Unraveling Phylogenetic Diversity
III. Diversity Extends to the Strain Level
IV. Phylogenetic Diversity and Distribution of Phenotypic Traits
V. The Diversity of Symbiotic Prokaryotes
VI. The Diversity of the Uncultured Free-Living Organisms
VII. Conclusions

GLOSSARY

domains The highest taxonomic rank defined to classify organisms into Archaea, Bacteria, and Eucarya, which differ from each other in fundamental genomic and phenetic properties.

endosymbiont Specialized form of symbiosis in which one partner thrives within cells, lumen, or tissues of the host organism; most obligate endosymbionts belong to the group of uncultured organisms.

epigenetic level Study of molecules which are the product of gene expression.

homology Denoting common ancestry. Structures, processes, sequences, behaviors, etc. are said to be homologous if there is evidence that they are derivations from a common ancestral structure. In molecular biology the term indicates a significant degree of similarity between DNA or proteins.

phylogeny Natural relatedness among life forms; the science of ordering the genealogy of organisms into a family tree.

prokaryotes Life forms that are members of the domains Archaea and Bacteria as opposed to Eucarya, comprising organisms with a cell nucleus.

rDNA Genes coding for rRNA that play a fundamental role in the translation process; the most thoroughly studied molecule in prokaryotic cells; used in comparative phylogenetic studies.

uncultured prokaryotes Organisms, the presence of which has been detected by molecular methods in the environment but they have not been cultured under artificial laboratory conditions.

BACTERIAL DIVERSITY comprises the total variability of prokaryotic life on Earth, covering all genomic, phenetic, phylogenetic, and ecological variations from the level of an individual strain to the community that evolved over a time span of probably more than 3.5 billion years. Determination of the extent of diversity is mainly restricted to cultured prokaryotes because the vast majority of strains are not yet accessible for subsequent research. Strains involved in obligate symbiotic
and parasitic relationships with their eukaryotic hosts and free-living strains for which appropriate culture conditions have not been developed can at best be detected by molecular techniques. This article summarizes results of recent approaches that have broadened the bacteriologist’s view about the immense richness of prokaryotic diversity.

I. RECOGNIZING GAPS IN KNOWLEDGE

Due to progress in methodologies and concepts, facets of biodiversity have no been covered equally in different areas in bacteriology, but progress must be viewed as a logical consequence of available technologies. Introduction of groundbreaking methods is usually followed by a period of increased knowledge in those fields for which the methods were developed. Numerous examples exist in microbiology, such as the elucidation of (i) the ultrastructure of the prokaryotic cell, following the development of the electron microscope; (ii) metabolic and biochemical pathways, following the introduction of the isotope label technique and enzymology; (iii) anaerobic organisms, following the development of appropriate anaerobic cultivation technology; and (iv) cell constituents, following the introduction of the amino acid analyzer, gas chromatography, high-performance liquid chromatography, and thin layer chromatography. Recently, the introduction of high-resolution gel electrophoresis led to one- and two-dimensional fingerprinting methods for proteins, ribonucleic acids (RNAs), and deoxyribonucleic acids (DNA); application of restriction enzymes, cloning strategies, and polymerase chain reaction (PCR) technology led to improved sequence analysis of genes and genomes; and cloning of PCR fragments of environmental DNA, in situ hybridization, and the development of gradient gel electrophoresis revealed a larger spectrum of prokaryotic diversity than previously known. The latter insights have been used to develop strategies in which DNA, isolated from environmental samples, is expressed to yield a spectrum of novel enzymes not detected in strains available from biological resource centers. As a result, taxonomists learned that the vast majority of prokaryotic species are still undetected, and physiologists, biochemists, and geneticists can expect to find novel lines of descent containing organisms that express fundamental deviations from currently known biochemical pathways. New models will help to better understand the structure and function of living matter and the role of prokaryotes in maintaining the biosphere.

The number of validly described species of animals, plants, and lower eukaryotes is approximately 400 times larger than the number of prokaryotic species (1.600,000:4,000). This fraction of bacterial and archaean species is surprisingly low considering that prokaryotic species evolved eons ago, exploring and occupying any niche that has been investigated for the presence of prokaryotic organisms. Insects, on the other hand, which comprise more than 1 million species, evolved late in evolution—less than 600 million years ago during the Cambrian. Certainly, differences in the species definition that exist between the biological species of higher evolved eukaryotes and the pragmatically defined prokaryotic species contribute to the tremendous discrepancy, but it has been shown convincingly that the vast majority of prokaryotic strains which are part of the free-living microbial community species have not been cultured. Endosymbiotic prokaryotes which are not free living but firmly associated with eukaryotic cells are another source of uncultured organisms. The limitation of recognizing the richness of microbial organisms is most likely due to the use of a restricted spectrum of enrichment media which selects for a very narrow spectrum of organisms that compete best under artificial laboratory conditions.

The discussion of diversity issues and the increasing awareness of the importance of microorganisms in maintaining the biosphere is embedded in discussion of the implementation of the articles of the Convention on Biological Diversity (CBD). The driving force behind the CBD has been the recognition of the possibility of a significant reduction or loss of biological diversity at source by human activities, and the preamble admits a general lack of information and knowledge regarding biological diversity. Nowhere is the lack of knowledge more acute than for microbial diversity (prokaryotes, fungi, yeasts, and heterotrophic and autotrophic protists) and there is a widespread appreciation among microbiologists that cultured microorganisms represent a very small, not necessarily ecologically important, fraction of natural microbial diversity.

Considering the total number of nucleotides per genome, far more recognizable diversity can be seen at the molecular level than at the epigenetic level. Compared to the total number of described species, the number of organisms analyzed in genome sequencing projects is still small (about 2.5%), but this portion will increase rapidly and genomic screening and data mining will develop into dominating biological disciplines in the future. However, taxonomists will not lose interest
in the more traditional properties of a cell because of the demanding process of describing species and genera. The polyphasic approach to classification requires information about morphology, ultrastructure, metabolism, chemical structure of cell constituents, and genomic features. Thus, from more than 100 years of bacterial classification, an enormous wealth of phenotypic and epigenetic data have accumulated and are of benefit to bacterial taxonomy, which many scientists regard as the mother of biological sciences. However, taxonomists are concerned that this spectrum of diversity will have to be sacrificed for an approach that considers molecular/genomic data more appropriate for delineating taxonomic ranks. Indeed, this cut will happen sooner or later, but it does not mean that the avalanche of genomic data will bury all activities directed toward the elucidation of the phenotype.

II. UNRAVELING PHYLOGENETIC DIVERSITY

Whereas the past century unraveled the diversity of epigenetic properties, i.e., those characters that are the result of gene expression, the next century will prove a tremendous wealth of information on genomic properties. The terms genomics and reverse genetics have been coined for the strategy of obtaining value-added information from sequences of genomes and genes. Data mining and meganetworking programs will replace the simple search for similarities between a few homologous genes, and it can be expected that the importance of phenotypic traits will be reintroduced as a consequence of the desire to understand the horizontal and vertical flow of genes and their regulation and expression.

The combination of information of a (small) portion of the genome with phenotypic properties is nowhere exemplified better than in modern classification strategy of prokaryotes. The first component, the backbone of the system, is provided by the primary structure of homologous molecules which have accompanied the organisms since early evolution and they can be ranked according to their evolutionary history. The number and nature of sequence differences among proteins and genes coding for rRNA and proteins allow the recognition of pairs or groups of organisms which evolved from a common ancestor, and the order in which lineages have emerged in time facilitates decisions about the grouping of organisms.

The main problem that has emerged in phylogenetic studies during the past 10 years is the question of whether to place emphasis on the genealogical relationships derived from comparative analyses of a single homologous gene, a gene cluster, or even large parts of the genome. It has been demonstrated convincingly that the historic fate of ribosomal (r)DNA genes does not necessarily represent the fate of other genes because the phylogenetic branching patterns of different genes may show significant deviations. Analysis of the presence and position of genes in completely sequenced prokaryotic and eukaryotic genomes and horizontal gene transfer among genomes with a high plasticity has played an enormous role in designing and shaping early evolved organisms, and the chimerical genomic structure of descendants of early evolving organisms has been proven beyond doubt. The 16S rDNA is the most widely analyzed molecule in phylogenetic studies because of its alternating change of degrees of sequence conservatism that allows the recognition of most distant relationships and moderate and close relationships. The availability of a database comprising more than 10,000 sequences of prokaryotic strains makes it easy to either unambiguously affiliate a new isolate to one of the >4000 species or to postulate the finding of a new taxon in case no highly similar match is obtained with the isolate and a recognized species. Analysis of rDNA is the gold standard for analyzing phylogenetic relatedness; Schleifer and Ludwig (1989) provided excellent evidence that the rDNA data are closely matched by results of comparative analyses of other conservative molecules responsible for central cell functions. Second, the phylogenetic branching patterns provide a scenario in which clusters of related organisms also share a high portion of epigenetic properties. The resolution power of the 16S rDNA is restricted, however, because neither very early nor very recent evolutionary events are well resolved. Branching patterns are influenced not only by the treeing algorithms applied for inferring phylogenetic relationships but also by the size of the database. A phylogenetic tree is a dynamic construct which will to some extent change its topology with any new sequence added. Consequently, it is difficult for the scientist to judge whether the phylogenetic tree generated is a true reflection of the evolutionary history of the analyzed molecule. If the tree is generated for the purpose of making conclusions on the fate of genomic and epigenetic characteristics of the organisms concerned, conclusions should be underpinned by the analysis of additional evolutionary markers.

Increasingly, genes with a resolution higher than 16S rDNA are sequenced for determining intragenetic and intraspecies relatedness (e.g., genes hsp65 and hsp70
coding for heat shock proteins and gyrA or gyrB). Traditionally, and still required in the polyphasic approach to classification, DNA–DNA hybridization studies are performed at the level of highly related species and at the interspecies level, but the laborious experimental burden and lack of a cumulative database excludes wide application.

III. DIVERSITY EXTENDS TO THE STRAIN LEVEL

In the past, selected physiological reactions and computer-assisted techniques were applied to elucidate intraspecific relationships. This goal can also be achieved by generating patterns from DNA, RNA, and proteins that represent a strain-specific fingerprint or a barcode-type profile. Spontaneous mutations occur at an average rate of about $10^{-9}$ per gene and generation in prokaryotes but vary to a great extent in genetic loci. The extent to which strains of a species differ from each other also depends on the relative evolutionary time during which strains evolved from each other. Because prokaryotic species are man-made constructs established to facilitate taxonomy, strains included in a species may vary significantly from each other in phylogenetic depth and hence in the extent to which macromolecules diverged. The complexity of the patterns obtained depends on the size, length, and degree of conservation of the macromolecule of choice and the tools used to cleave, amplify, hybridize, and separate these markers. Separation of DNA fragments by pulsed field electrophoresis, probing of defined genes with labeled probes, visualization of bands via computerized, laser-analyzed densitometer scanning, and the use of fully automated, reproducible techniques have improved the resolution and the monitoring part of the analyses significantly. Pattern identification is rapid, discriminating, and applicable to any species for which DNA, RNA, and whole cell proteins can be isolated. Providing the potential for discrimination, isolates and reference strains that exhibit a high degree of pattern similarity can be considered related. These techniques complement traditional typing methods used mainly in the clinical environment, such as serotyping, biotyping, and phage typing. The decision of whether a strain with a unique pattern actually belongs to a described species or should be described as a new species requires more quantitative methods at the genomic level that allow one to measure the degree of relatedness.

IV. PHYLOGENETIC DIVERSITY AND DISTRIBUTION OF PHENOTYPIC TRAITS

The phylogenetic and phenotypic separation of the domain Archaea from the domains Bacteria and Eucarya (Fig. 1) is the most exciting result since the introduction of comparative sequence analysis by Woese and Fox in the mid-1970s. The presence of two prokaryotic domains in which members are defined by clearly different genomic and phenotypic properties has changed fundamentally the hypothesis on the dichotomy of life forms and revolutionized ideas about the evolution of the eukaryotic cell. Domains were introduced by Woese, Kandler, and Wheelis to denote that these primary lines of descent constitute higher entities than the traditional eukaryotic kingdoms. The most significant epigenetic differences among members of Archaea and Bacteria are the compositions of the cell wall and fatty acid and the modification pattern of RNA.

Figure 1 schematically depicts the tripartition of the domain Archaea that guided the description of three kingdoms, the Euryarchaeota, the Crenarchaeota, and the Korarchaeota, for some uncultured organisms. No kingdoms have been described for the rich phylogenetic structure of the domain Bacteria, outlined in Fig. 2, because of the significantly large number of organisms and lineages involved, which are not always well separated from each other. The order at which these lineages evolved is of low statistical significance but the phylogenetic composition of organisms within these lineages can be recovered by analyses of other genes, such as those coding for 23 rRNA, 5S rRNA, ribosomal proteins, ATPase, elongation factors, and heat shock protein HSP70. In some lineages certain characteristics are indeed of phylogenetic significance, such as morphology and/or ultrastructural features (Thermotogales, Planctomycales, Verrucomicrobiales, Spirochaetales, and Myxobacteria), chemotaxonomic properties such as cell wall composition (Thermotogales/Desmococcus, Clostridium, and Actinobacteria) or lack thereof (Planctomycales), and physiology, i.e., the composition of the photosynthetic apparatus (Chloroflexus, Chlorobiales, and cyanobacteria). Most lineages, however, have a wide variation of morphological, chemical, ultrastructural, and biochemical diversity, some traits of which may have been acquired in the course of their evolution by horizontal gene transfer, whereas others may have evolved as a response to occupying new environmental niches (e.g., autotrophic and chemolithoautotrophic forms).
and hence of taxonomic value—lost their significance in classification when their polyphyletic origin was demonstrated or when these properties were found to be of little genomic stability. In general, this is true for morphology, spore formation, the relationship to oxygen, the presence of a photosynthetic apparatus, gliding motility, and many other characters. In modern classification, these properties are no longer used as the sole basis for the description of higher taxa, such as families, orders, and classes; today, the rational for a higher taxon is primarily the distinct phylogenetic grouping of its members, whereas their phenotypic description may be rather broad.

A good example of the inability of phenotypic properties to serve as phylogenetic markers is provided by members of the class Proteobacteria, which includes the majority of Gram-negative bacteria. This class is highly diverse with respect to physiological and morphological properties. Morphological diversity ranges from simple spherical forms to the highly complex fruiting bodies of myxobacteria. Physiologies include chemolithooauto- trophy, photosynthesis, fermentation, anaerobic respiration, and nitrogen fixation. To reliably affiliate a new isolate to a described genus by these properties is improbable; the chance of doing so is increased by the presence of unique biochemical properties, such as nitrate or ammonium oxidation (Nitrobacter in the α-subclass and Nitrosomonas and relatives predominantly in the β subclass, respectively) and sulfate reduction (Desulfovibrio and relatives, δ subclass). Sulfur and sulfate oxidation as well as anaerobic photosynthesis are poor phylogenetic markers because members of Thiobacillus are found in the α, β, and γ subclasses and the sulfurless photosynthetic organisms occupy different sublines of descent within the α and β subclasses.

As a consequence of the recognized discrepancies
between the phylogeny of prokaryotic taxa and their previous taxonomic treatment, the recent phylogeny-oriented classification system of genera and higher taxa was developed to primarily match the 16S rDNA data.

Newer textbooks have adopted the modern approach, in which, by providing the phylogenetic framework, the origin and evolution of certain phenotypic traits may be better understood than expressed by traditional superficial lumping.

V. THE DIVERSITY OF SYMBIOTIC PROKARYOTES

Endosymbiotic associations, recognized more than a century ago, initially concentrated mainly on the elucidation of the origin of chloroplasts. Later, the importance of eukaryote–prokaryote relationships was recognized for associations between plants and nitro-
gen-fixing bacteria, e.g., legumes and *Rhzobia* species, monocots and *Asgospirillum* species, *Casuarina* sp. and *Frankia* sp., ferns and *Anabena* sp., and the prokaryotic origin of plant mitochondria. However, most endosymbiotic relationships between microorganisms and their eukaryotic hosts were mainly descriptive and the taxonomic affiliation of the vast majority of uncultured microorganisms remained virtually unknown. In contrast, microbial partners participating in nonobligate symbiotic relationships were identified long before the molecular era; there are numerous associations for the ectosymbiotic microbiota of the rumen, intestine, gut, and skin.

### A. Identification of Symbionts

Most endosymbiotic bacteria are defined as “as yet uncultured” organisms, many of which do not exist as a pure culture within the host’s tissue. Hence, genes coding for rRNA cannot be isolated selectively but must be identified within clone libraries consisting of PCR-amplified rDNA using prokaryote-specific PCR primers and total DNA extracted from the plant, animal, or even the prokaryotic cell. If the host contains a single symbiotic partner only, the clone library will consist exclusively of the one unique rDNA insert; if the association is more complex, the clone library will contain phylogenetically different inserts. Authentication of the symbiont and verification of the location of the putative endosymbiont within the host’s tissue is performed by fluorescence in situ hybridization techniques as elegantly developed by Stahl, Amann, and coworkers. In most cases the 16S rDNA sequence will provide sufficient unique nucleotide stretches to allow generation of symbiont-specific oligonucleotide probes.

Not until 1982 was *Prochloron didemni*, the endosymbiont of the ascidian *Lissoclinum patellum*, identified by molecular techniques. Ten years later, a wealth of information was available on the molecular phylogeny of symbionts and endosymbionts from a broad spectrum of eukaryotic hosts, ranging from protozoa to vertebrates. The availability of a large database, consisting of thousands of 16S rDNA sequences of free-living prokaryotic species, facilitates the search for the phylogenetic affiliation of the more than 500 16S rDNA sequences available for host-associated bacteria. Fundamental questions about the identity of the prokaryotic partner, the evolution of symbiotic relationships, and the mechanisms of symbiont transmission can now be addressed.

### B. The Evolutionary Origin of Symbionts

Most endosymbionts investigated to date originate from ancestors within the domain Bacteria, in which they are found in a few main lines of descent (open arrowheads in Fig. 2). The ability to thrive in certain anaerobic protozoa of the genera *Metopus*, *Plagiopyla*, and *Trimyema* appears to be widespread among methanogenic Archaea (kingdom Euryarchaeota). No endosymbionts have been described to share a common ancestry with those prokaryotes which define the most deeply branching lineages, such as the *Crenarchaeota*, domain Archaea, and *Aquificales*, *Thermotogales*, and other branches comprising thermophile and phototrophic organisms of the domain Bacteria. Because symbionts and nonsymbionts share more than 80% 16S rDNA sequence similarity it can be concluded that the invasion of the eukaryotic host by prokaryotic cells must have occurred less than 2 Gy ago. Figure 3 shows recent evolutionary events by plotting geological time against 16S rDNA similarity values determined for the origin of organisms defined by key physiological types (i.e., oxygen-generating photosynthesis by cyanobacteria, fermentative metabolism in facultative anaerobic bacteria, and origin of respiration chain in aerobic bacteria). Thus, as derived from 16S rDNA similarity values, the origin of endosymbionts correlates with the origin of the eukaryotic cell and endosymbiosis has occurred repeatedly in (perhaps all) eukaryotic lineage.

In order to study the history of symbiotic associations, the phylogenetic trees of hosts and their symbionts should be compared. Few data are available, with the most convincing study being that on the endosymbiotic
tropical unicid bivalves and deep sea bivalves. The topology of the symbiont Buchnera aphidicola tree is completely concordant with host phylogeny based on morphology. The fossil and biogeographic time points for the aphid phylogeny have been used by Wilson and Bauman to calibrate the 16S rDNA of the closely related endosymbionts (>95% similarity). The value of 1% fixed substitutions per 25–30 million years determined for the symbionts of aphids is similar to the value of 1% per 50 million years determined on the basis of a broader range of nonobligate symbiotic relationships (e.g., Rhizobium/legumes, Photobacterium/fish, and enterobacteria/mammals) and to the value of 1% per 60 million years for the past 500 million years (Fig. 3).

The host’s advantage of the association has been unraveled in a few cases, such as the removal of hydrogen produced from hydrogenosomes of ciliates by archaean methanogenic endosymbionts, provision of nutritional carbon to the host bivalves by sulfur-oxidizing gill symbionts, or essential amino acids to aphids by their endosymbionts. Application of the PCR techniques has allowed the elucidation of the transmission route of symbionts in ovaries, testis, and gill tissue of tropical lumicid bivalves and deep sea bivalves.

C. Phylogenetic Affiliation of Endosymbionts

The majority of endosymbionts cluster phylogenetically most closely with Gram-negative free-living bacteria. It can therefore be deduced that Gram-negative bacteria are the most successful candidates for forming symbiosis, including obligate endosymbiotic associations. A few nonproteobacterial symbioses have been described, such as those found between wood-eating cockroaches and termites and spirochetes and Gram-positive host-associated bacteria, such as the microspatulate Pasteuria penetrans (in root-knot nematodes), Epulopiscium (a fish symbiont), and Frankia (nitrogen fixing on Casuarina and relatives). The association between Spingobacterium comitans, a member of the Bacteroides/Flavobacterium phylum, and the myxobacterium Chondromyces croatus is an example of a prokaryote–prokaryote symbiosis. Endosymbionts of Archaea are members of the kingdom Euryarchaeota and they are, in contrast to those of Bacteria, organisms containing a Gram-positive cell wall. These anaerobic symbionts, originating from ancestors of the families Methanoccicicaceae, Methanocobacaceae, and Methanosarcicaceae, have been identified in several genera of termites and protozoa. The basis of the symbiotic interaction appears to be hydrogen transfer from host to endosymbiont, which can use the gas for methanogenesis.

1. Symbionts of the Proteobacteria

Proteobacteria embrace organisms known to have close associations with eukaryotic hosts as pathogens, pathogeness bacteria, incompatibility bacteria, symbionts, and organelles (such as the mitochondria of plants which evolved from α proteobacterial ancestors). Endosymbionts and nonobligatory associates of the same host may belong to different phylogenetic groups, indicating that the same host is susceptible to more than a single invasion process and that not all symbiotic relationships result in obligate endosymbiosis.

a. Symbionts of the α Proteobacteria

The α subclass of Proteobacteria contains a wide spectrum of organisms that are closely associated with eukaryotic cells. Prime examples are members of the nitrogen-fixing genera Rhizobium, Sinorhizobium, Bradyrhizobium, Mesorhizobium, and Azorhizobium. Pathogens include Afpia, Brucella, Bartonella, Rickettsia, Ehrlichia, Orientia, and Anaplasmata. A highly related cluster of host-associated organisms, including symbionts of insects, cytoplasmically inherited bacteria such as the parthenogenesis bacteria (PB), and the cytoplasmatic incompatibility bacteria (CIB) is related to Wolbachia pipientis. Other members of this group are the PB and CIB of Culex and Drosophilidae. Because of the degree of 16S rDNA similarity among cultured strains, these symbionts must be considered members of the same species. These relationships demonstrate that the common ancestor of Wolbachia and relatives invaded a broad spectrum of insect hosts which, as shown by the high values of up to 99% 16S rDNA similarity, must have occurred recently in evolution.

b. Symbionts of the β Proteobacteria

Members of the β subclass of Proteobacteria encompass a wide range of mainly pathogenic plant- and animal-associated bacteria, such as members of Burkholderia, Azorhizobium, Rhizobium, Bordetella, Kingella, Eikenella, and Neisseria. Also included in this subclass is the kinetoplast of Crithidia. The diversity of endosymbionts, however, is rare and restricted to the endosymbionts of the mealy bugs, which are moderately related members of the genera Ralstonia and Burkholderia.

c. Symbionts of the γ Proteobacteria

By far the higher number of endosymbionts of insects and vertebrates are members of the γ subclass of Proteos-
bacteria. In addition, this taxon contains a wide spectrum of animal and human pathogens and nonobligate symbionts such as members of Enterobacteriaceae, Legionellaceae, Pasteurellaceae, Vibrionaceae, Pseudomonas (sensu stricto), and Acinetobacter. Many endosymbionts cluster according to the phylogenetic rank of their hosts, which may be indicative of coevolution events: (i) The primary endosymbionts of giant ants, aphids, tsetse-tse, and the sweet potato white fly form a phylogenetically coherent cluster within the radiation of enterobacteria; (ii) the symbionts of fish with light organs, such as the deep-sea anglerfish and the flashlight fish, cluster with different members of Vibrio; (iii) two moderately related subclusters consist of the gill symbionts of bivalves. These organisms are remotely related to methylotrophic bacteria. The latter symbionts are sulfur-oxidizing organisms which provide their hosts with nutritional carbon. Nitrogen-fixing symbioses are not restricted to plants but the gland of an invertebrate shipworm contains large numbers of γ proteobacteria which possess the ability to digest cellulose and fix nitrogen.

VI. THE DIVERSITY OF THE UNCULTURED FREE-LIVING ORGANISMS

For more than a century, assessment of prokaryotic species has been evaluated by the culturing approach. The number of different growth media is unknown, but all aimed to recover the largest possible diversity of organisms. However, it is not the enrichment and isolation but rather the lack of cheap and reliable molecular identification methods that still slow microbiologists in their attempts to classify strains to the species level and to describe new prokaryotic species. The number of novel strains that has been eliminated during the isolation process, or which were only included in a biased search for specific properties before they disappeared in nonpublic resource collections, cannot be counted. It is unknown how many strains have been investigated in parallel. Although this problem may one day be overcome by more facilitated species definitions and the availability of a global network of biological information, microbiologists are currently confronted with the problem that there is a remarkable difference in the number and morphology of organisms in natural samples with enrichment cultures and isolated colonies. Staley and Konopka introduced the phrase “great plate anomaly” to indicate that only a small fraction of prokaryotic species observed under the microscope will grow under artificial laboratory conditions.

A. The Vast Majority of Prokaryotic Species Have Not Been Cultured

Pace and colleagues first suggested that rRNA sequences could be used to characterize natural communities without the need to culture. The delay in publishing the first studies by the research groups of Giovannoni and Ward was largely due to the need to develop robust and simple technologies whereby 16S rRNA sequences could be recovered from complex mixtures of environmental nucleic acids and then individually sequenced. Since then, there have been numerous 16S rRNA- and 16S rDNA-based studies in which sequences were analyzed to explore microbial diversity in different environments. Most of these studies differ in details of methodologies, such as isolation of nucleic acids, PCR amplification conditions, and source of cloning vectors and ligation enzymes, but in one aspect the outcome of all of these studies was similar: The vast majority of the more than 1200 environmental partial 16S rDNA sequences, deposited in public databases, were not identical with and often not even similar to the homologous sequences of described species accessible in the extensive databases of cultured bacteria (Fig. 2, solid arrowheads). Also, the sequences were rarely identical to sequences obtained from strains that were isolated in parallel to the molecular work from the same environment. As judged from the low degree of sequence similarity, one could even conclude that many sequences are indicative of the presence of higher taxa. This finding reinforced the previously mentioned idea that the vast majority of species have not yet been isolated.

Additional molecular techniques have been developed for understanding the composition of microbial communities that reach beyond the mere assessment of phylogenetic relationships of clones and strains. Among others, methods comprise (i) the application of gradient electrophoresis of PCR-amplified 16S rDNA sequences to facilitate recognition of changes of populations in time and space, (ii) the development of biological probes to detect the presence of genes encoding metabolic enzymes or to identify bacterial species directly in an environmental sample, (iii) the development of biosensors to determine microprofiles of inorganic compounds, (iv) flow cytometry and cell sorting to enumerate and separate groups of organisms according to size and taxon specificity, and (v) subtractive hybridiza-
tion to facilitate comparative analysis of environmental samples, and (vi) extension of the database to include genes other than 16S rDNA (e.g., nif genes).

The environments discussed in the following sections were selected because they provide the largest database of phylogenetic information on uncultured organisms. Many other environments have been investigated, but the overall picture of prokaryotic diversity is less focused; these include, fresh water, paddy fields, marsh oil, marine plants, bioremediation sites, bioleach reactors, or the multistructured associations between prokaryotes and eukaryotic cells.

1. The Marine Environment
   a. The Uncultured Archaea
      The use of archaean-specific 16S rDNA PCR primers and subsequent analysis of clone libraries from DNA of oxygenated coastal surface waters and oligotrophic open-ocean samples by the groups of DeLong and Fuhrman revealed the widespread occurrence of two types of archaean diversity. Analyses of different marine sites, including surface and subsurface waters of the Pacific, Atlantic, Antarctic coastal waters, offshore slope regions, and a deep-sea marine holothurian, have subsequently confirmed the affiliation of archaean sequences to these two groups. Clone sequences of the first group contained 16S rDNA genes of Crenarchaeota organisms found at depths 100 to 4800 m, whereas others were found at different marine sites at more shallow depths. Organisms from which sequences originated constitute a significant component of approximately 5–14% of the marine picoplankton assemblages and were novel and closely related to the archaean "Crenarchaeum symbiosum," a symbiont of the sponge Axinella mexicana. At a lower level of relatedness these sequences form a separate branch at the basis of the kingdom Crenarchaeota, where they show some distant relatedness to a group of sequences isolated from sediment organisms of the Obsidian Pool, Yellowstone.

A few sequences have been described to belong to the second sequence group originating from Euryarchaeota and are distantly related to the terrestrial species Thermoplasma acidophilum. Except for a holothurian archaean sequence, they have been retrieved from the same environment from which the crenarchaeotal sequences were obtained, but mostly at lower depth (0–100 m). The high sequence similarity of members of this group from geographically separate sites contrasts the finding by Munson and colleagues, who demonstrated the high degree of phylogenetic diversity between archaean sequences retrieved from 16S rDNA clone libraries generated from material sampled in marsh sediment samples and adjacent vegetative marshland in the United Kingdom. In this study, clone sequences formed about 15 different phylogenetic groups, each of which was highly to distantly related to cultivated Euryarchaeota species.

The origin of marine archaea is unclear because no representative has been cultured. One may argue that these prokaryotes originate from dormant stages of released commensals or symbionts of marine invertebrates. This view, however, is contradictory to the high cell number of living organisms of this group—up to 1×10^6 of the total community. Disturbed deep-sea sediments, which have not been investigated, may be another possible source for these organisms. In contrast, euryarchaeal clone sequences retrieved from surface water material may be coastal and even terrestrial origin, considering that there is a specific relationship of these sequences to some of the DNA retrieved from coastal salt marsh.

b. The Uncultured Bacteria
   Phylogenetic analysis of rDNA from phytoplankton of the Sargasso Sea by Giovannoni, Britschgi, and coworkers revealed that the majority of the mostly eubacterial sequences were novel to systematists. Some clones represented oxygenic photrophs and could be assigned to the Synchococcus group of the Cyanobacteria. Groups of related sequences of mainly the α and γ Proteobacteria were identified and defined as SAR (Sargasso) groups, showing distant relatedness to cultured bacteria of Shewanella, Vibrio, and Oceanospirillum. Analysis of the phytoplankton from the north Pacific near Hawaii by Schmidt and colleagues led to the unexpected result that the population was very similar to that of the Atlantic Ocean. Some sequences were similar to those of common cultured marine organisms (e.g., Vibrio, Pseudalteromonas, and Chromatium), whereas others represented novel lineages which were distantly related to the Fibrobacter group and Chlorobium.

Following studies extended the range of sampling sites in the Atlantic and Pacific Oceans, the Antarctic Sea, and the Mediterranean Sea. Basically, the findings of the first studies were confirmed in that many of the new sequences were highly related, although not identical, to those defined earlier, irrespective of the location. The ecological role of the hitherto undescribed organisms remains unresolved, although their general physiological capacity is probably similar to that of described species because α Proteobacteria
encompass mainly lithotrophic and oligocarbohydrate organisms, whereas γ-Proteobacteria and cytophages exhibit strong hydrolyzing and degrading capacities. The few sequences of Gram-positive bacteria are probably of terrestrial origin because they are mainly found in coastal regions and in sediments which must be regarded as a deposit mainly for endospore-forming organisms.

B. Hot Spring Environments

Research in hot spring environments has been stimulated by the discovery of Thermus aquaticus as a species of high biotechnological value. The community of mats especially provided an excellent comparison of microbial composition as assessed by selective culture techniques introduced by Brock and Castenholz. Results of the sequence investigations by Ward, Weller, and colleagues clearly demonstrated that even such a rather closed system inhabits a phylogenetically diverse community. Consistent with the findings in the marine environment, none of the recovered sequences closely resembled sequences from cultured taxa isolated from similar environments. Sequences of cultured organisms believed to constitute a major component of the mat community, such as Synechococcus lividus or Chloroflexus aurantiacus, were not recovered. Analysis of prokaryotic DNA from hot springs located in Yellowstone National Park by Pace and coworkers revealed an unexpectedly large number of distinct bacterial sequences, indicative of novel main lines of descent. Archaeal sequences were determined to be specifically related to sequences from the Crenarchaeota. A few sequences showed high similarities with 16S rRNA sequences of cultivated Archaea, (e.g., Desulfurococcus mobilis, Pyrobaculum islandicum, and Thermostoﬁlum pondens) but were not identical to any. The archaeal sequences from the hot spring environment were not closely related to the novel archaeal sequences retrieved from the marine environment, which appear to possess a similar position intermediate to the Crenarchaeota. Results of these studies indicated that the domain Archaea possesses a third line of descent, the kingdom Korarchaeota, suggesting that not only the phyllogenetic but also the physiological diversity of the Archaea are significantly larger than reflected by the few cultured representatives.

C. Soil Environment

In contrast to the early extensive work on different marine sites and hot springs, determination of microbial biodiversity in soil was delayed until appropriate methods were developed that circumvented methodological difficulties such as the isolation of PCR-able DNA from humic acid-containing soil and semi-quantitative cell recovery. Despite these shortcomings studies by Liesack and Triplett clearly demonstrated that soil samples from the Southern and Northern Hemispheres contain a rich and varied bacterial flora, including non-thermophilic archaean members from the kingdom Crenarchaeota. The first soil sample investigated by sequencing and probing 16S rDNA genes was located in a subtropical, moderately acidophilic, and forested environment in Queensland, Australia. Unexpectedly, only a few sequences were obtained from commonly isolated soil organisms, such as Streptomyces, although members of this genus were cultured in large numbers from the same soil sample. The reasons for the low representation of Gram-positive bacteria in libraries generated with universal 16S rDNA primers are not known, but this may be explained by the cell wall structure of resting and dormant cells which fail to disintegrate under mild enzymatic lysis. By using actinomycete-specific primers, Embley and colleagues were indeed able to demonstrate the presence of a rich diversity of these taxa in soil. Alternatively, Gram-positive bacteria may be a minor (numerically) component of the soil flora, with selective isolation exaggerating their numbers. Some Australian soil sequences were closely related to those of nitrogen-fixing species of the α-Proteobacteria, but the majority of clone sequences represented novel groups that were only remotely related to known taxa, e.g., Planctomycetes, Actinobacteria, Verrucomicrobia, Acidiphilium, and Thiothrix. Several actinobacterial clone sequences, retrieved from rDNA of different soil types and different geographical locations, formed two clusters grouping remotely with members of Rhodococcus, Acidithiobacillus, and Arthrobacter. In addition to their presence in Australian forested soil, their occurrence was verified in a hot spring (Australia), geothermally heated soil (New Zealand), paddy fields and soybean fields (Japan), cultivated soil (Mexico), peat bog and garden soil (Germany), grassland soil (The Netherlands), and forest soil (Finland) (Fig. 4). These sequences formed fractions of 1–23% of the respective clone libraries and one of these organisms constituted about 6% of the metabolically active part of a Dutch grassland soil community as shown by ribosomal RNA analysis. It can be deduced that these uncultured organisms...
FIGURE 4 Distribution of 16S rDNA clone sequences, representing uncultured bacteria, within the radiation of 16S rDNA sequences of some cultured bacteria. Origin of clone sequences: MC, Australian soil; PAD, Japanese paddy field; FIE, Japanese soy bean field; NH, north Pacific Ocean. The scale bar indicates 10 estimated changes per nucleotide position.

are distributed worldwide and play a physiologically important role in the soil ecology.

VII. CONCLUSIONS

Despite tremendous progress in the elucidation of prokaryotic diversity, many pitfalls have been identified which influence the composition of sequences in a clone library and hence these data can be used neither to quantify nor to qualify the composition of communities. Any estimation of the relation of cultured and as yet noncultured organisms is nothing more than a guess and not supported by scientific data. Most environmental analyses have revealed a heterogeneous mixture of deep and shallow branching lineages, very few of which have shown close relationships to cultured taxa. Many lineages are very closely related to each other, some of this diversity is due to microheterogeneity at the level of rrn operons within a single cell, whereas others may represent true strain diversity.

Although the contribution of rDNA and rRNA to microbial ecology must be considered significant, one molecule alone cannot nearly cover all facets of microbial ecology. Not only must the function of an ecosystem be deduced from analyses of genes expressed through rRNA, mRNA, and proteins, and not only should the network of broad physiological interactions be verified by in vitro reconstitution of isolates but also all these strategies must include data on thorough physical and chemical analysis of the natural sample. Because ecological interactions are performed by strains and not by species, the 16S rDNA in not the appropriate tool to unravel the diversity of this high phylogenetic level. Strain diversity may thus be one or more magnitudes higher than reflected by the analysis of such an evolutionary conserved gene. Ecological niches must be defined and the difficulty in doing so is increased with the complexity of the sample. Soil samples, for example, are homogenized in that during the isolation of DNA many individual microniches with their individual populations are destroyed. In order to understand ecological interactions, population sizes must be known, strain richness and strain abundance must be assessed, the biochemical diversity of strains must be recognized by applying functional probes for the detection of the expression of specific genes, and physical probes must be applied for the assessment of the chemical and physical conditions of the environment. Modern environmental studies must maintain the isolation component because biotechnological exploitation of cosmid libraries is as desirable as the increase in knowledge of the evolution, phylogeny, and ecology of pure cultures.

See Also the Following Articles

ARCHAEA, ORIGIN OF • BACTERIAL GENETICS • EUKARYOTES, ORIGIN OF • MICROBIAL BIODIVERSITY, MEASUREMENT OF • THERMOPHILES, ORIGIN OF

Bibliography


I. TAXONOMY

Eubacteria (henceforth denoted simply bacteria) are a genetically highly diverse group of organisms and have been found in virtually every environment examined. Bacteria have been isolated from hot springs, deep-sea hydrothermal vents, inside Antarctic rocks, and as endosymbionts in many organisms. There is greater genetic diversity among Eubacteria than in beetles, insects, or arthropods—taxa notable for genetic diversity. This diversity results from both a long evolutionary history (more than 3.5 billion years) and the effects of both selection and genetic drift. Unlike eukaryotes, much of the genetic diversity of bacteria is largely unknown with the vast majority of species not characterized. Until the advent of molecular tools, only those bacteria culturable in the laboratory could be easily studied, and such bacteria may be no more than 0.1% of the total diversity.
Gram staining has historically been used to classify bacteria into two broad groups. Bacteria which retain the crystal violet–iodine complex used in gram staining are gram positive and those which do not are gram negative. Differences in gram staining result from differences in cell wall structure, and as such gram staining generally does correlate well with actual phylogenetic differences. However, based on 16S rDNA trees, there are 19 currently recognized main bacterial lineages (Table I). Of the 19 lineages, three phyla (Proteobacteria, gram-positive bacteria, and Cyanobacteria) comprise more than 90% of the known genera and contain the majority of described species. Most characteristics which previously had been used in bacterial taxonomy (e.g., nitrogen fixation, prototrophy, sulfur reduction, motility, budding reproduction, and spore formation) are of little value in determining phylogenetic relationships. Similarly, phylogenetic distance is not a good predictor of the mode of life in bacteria and extraordinarily different bacterial phenotypes can be closely related.

Assessing bacterial diversity is difficult because the biological species concept cannot be easily applied to most bacteria. According to the biological species concept, a species is made up of those individuals which interbreed in nature. Excluding bacterial species undergoing natural transformation, recombination is too infrequent to be a defining species characteristic for bacteria. The commonly used alternative adopted by bacterial taxonomists is based on the tendency for double-stranded DNA to “melt” into single strands when heated. Homoduplex DNA, which is double-stranded DNA from the same bacterial strain, will tend to have a higher melting temperature than heteroduplex DNA, which is double-stranded DNA from different strains, due to DNA sequence differences between the two strains. The greater the sequence differences between strains, the lower the melting temperature of the heteroduplex DNA. Two bacterial strains are considered to be members of a single species if the difference between the melting temperatures of the double-stranded homoduplex DNA and heteroduplex DNA is not more than 3°C and if they share no less than 70% identical DNA base pairing (sequence similarity). This alternative species criteria reflects a sequence divergence of no less than 7 or 8% and would group together naturally transformable bacterial species that can be differentiated by the biological species concept. If applied, the criteria would also fail to differentiate many eukaryotic species, such as humans, chimpanzees, and gorillas. However, based on such criteria one would estimate that there are between $10^9$ and $10^{12}$ bacterial species (Dykhuizen, 1998).

II. GENETIC ARCHITECTURE

Genetic architecture refers to the physical and informational structure of the genome. How is the genome organized and how is the encoded information expressed? Across all bacterial lineages much of the genetic basis of heredity is the same. Bacteria are haploid and all genetic material in bacteria is encoded by double-stranded DNA molecules using four bases: adenine, guanine, cytosine, and thymine. The sequence of bases in DNA provides the information necessary for RNA and protein structure. Nevertheless, genetic differences among bacteria abound and it is these differences that allow bacteria to be reliably differentiated. In contrast, although phenotypic differences are also prevalent, they
are of substantially less value in differentiating bacteria, and over evolutionary timescales bacterial phenotypes are highly variable. The ease at which evolutionary transitions can be made and the great variety of habitats in which bacteria are found indicate that the bacterial genetic architecture, as a whole, is able to respond to selection relatively easily. The persistent and sometimes large differences in genetic architecture among bacterial species appears to engender minimal constraints on bacterial diversification.

Against this backdrop of diversity, *Escherichia coli* is the best characterized bacterium. Much of the initial work on bacterial metabolism and genetics was done with *E. coli*, and all subsequent work has used *E. coli* as a comparison. The article will follow a similar convention.

### A. Physical Structure

The physical structure of bacterial genomes is strikingly divergent from that of eukaryotes. Even so, the diversity among bacterial species in genome size, organization, and method of replication may be as large or larger.

#### 1. Size

Bacterial genomes are typically small: *E. coli* strain K12 has a genome size of 4.6 million bases (mb) in comparison with the 12 mb of the eukaryotic yeast *Saccharomyces cerevisiae* and 3300 mb for humans (Fig. 2). The small size of bacterial genomes reflects the simplicity of the bacterial cells but also appears to be due to the lack of "junk" DNA in bacterial genomes. Bacterial genes are generally packed closer together than in most eukaryotes with relatively few intergenic spacer regions.

In *E. coli*, ~85% of the chromosome is composed of RNA and protein coding regions. Among bacteria, smaller genome size is often found in parasites and obligate endosymbionts. For example, *Mycoplasma genitalium* has a genome of approximately 9 mb. Endosymbionts and parasites can lose some metabolic pathways and depend on their host for those gene products. Chlamydia, which are obligate animal parasites, depend on their host for energy generation because they are incapable of either synthesizing adenosine triphosphate or reoxidation of reduced nicotinamide adenine dinucleotide phosphate. However, many pathogenic bacterial species do not have especially small bacterial genome sizes. *Mycobacterium tuberculosis*, the bacterium causing tuberculosis, has a genome size of 4.4 mb, which is slightly larger than the 4.2-mb genome of free-living soil bacterium *Bacillus subtilis*.

#### 2. Chromosomes

Bacterial chromosomes are located in a nucleoid, a distinct cytoplasmic structure, in which double-stranded DNA is coated with histone-like proteins. Most bacteria appear to have a single large circular chromosome, but this is not universal. Many species have multiple chromosomes, such as *Rhodobacter sphaeroides*, which has one chromosome of 3.0 mb and a second of 0.9 mb, and *Burkholderia cepacia*, which has three chromosomes with respective sizes of 3.6, 3.2, and 1.1 mb. Other species have linear chromosomes, such as the spirochete *Borrelia burgdorferi*, which is the causative agent of Lyme disease, and the gram-positive *Streptomyces coelicolor*. In addition, many bacteria contain extrachromosomal elements such as plasmids (see Section II, A, 3). The genome of *B. burgdorferi* is composed of an ~0.9-mb linear chromosome and ~19 linear and circular plasmids with a total size greater than 0.56 mb.

The type of chromosome topology has strong effects on chromosome replication. Unlike in eukaryotes, bacterial chromosomes have a single site for initiation of replication. In *E. coli*, replication of the single circular chromosome is initiated at the oriC site (origin of replication) and proceeds in both directions in the same semiconservative fashion as in eukaryotes. Gene organization of the origin of replication region is evolutionarily conserved across many bacteria lineages (e.g., *E. coli* of the phylum Proteobacteria and gram-positive *B. subtilis*) suggesting early fixation of a general mechanism for chromosome replication. DNA replication con-
times around the circular chromosome until both replication forks meet in the terminus region, which forms a barrier to replication fork progression. Specific AT-rich sites (Ter sites) within the terminus region impede replication from one direction, but not the other, and are oriented so that genome replication can be completed (Fig. 3). In contrast, the linear chromosome of *B. burgdorferi* is bounded by telomeres akin to eukaryotic chromosomes. Telomeres function to seal the ends of chromosomes, stabilizing the chromosome. *Borrelia burgdorferi* telomeres consist of inverted repeat sequences and covalently closed single-stranded hairpin loops—a very different structure from that of the terminus region of circular chromosomes. As in circular genomes, chromosome replication occurs bidirectionally, but many other details including the location of the origin of replication are unknown.

Two other aspects of chromosome replication are particularly noteworthy. First, because initiation of chromosome replication can occur repeatedly prior to its completion, in circular chromosomes it has been shown that genes closer to the origin of replication are present in higher numbers than genes closer to the terminus in rapidly growing cells (Fig. 3). Hence, rapidly growing cells are partial or merodiploids, even though after growth ceases the resulting bacterial cells are haploid. Second, chromosome partitioning among the two daughter cells is a highly accurate process (loss occurs in less than 0.03% of cells) and until quite recently was thought to be a passive process coupled directly to growth of the bacterial cell prior to binary fission. Instead, chromosome partitioning appears to involve active movement, as occurs in eukaryotic cells, of the oriC regions toward opposite poles of the forming daughter cells.

3. Plasmids

Plasmids are double-stranded DNA molecules ranging in size from ~1 kilobase (kb) to hundreds of kilobases that are physically independent of the major bacterial chromosome(s) and replicate independently. Most known plasmids are circular, although linear plasmids are found in some bacterial species. In part, the distinction between plasmids and bacterial chromosomes is historical, dating from when all bacterial cells were thought to contain a single chromosome per cell. The modern distinction is that chromosomes contain essential genes as opposed to plasmids which may contain genes expressed only under particular conditions (inducible genes). Genes carried on plasmids may code for a variety of traits, including antibiotic resistance, virulence factors, nutrient catabolism, conjugative gene transfer (see Section III,B), and antimicrobial compounds (e.g., antibiotics). Some plasmids, known as cryptic plasmids, code for no known function other than their replication. Depending on the bacterial host, genes encoded, and environmental conditions, plasmids may be parasites, commensals, or beneficial symbionts.

Plasmids can be present in single copy or up to many hundreds of copies per cell. Different plasmids use different mechanisms for replication, some with multiple origins of replication. There are two general types of replication: # and rolling circle replication (RCR). oriC-type replication is an example of # replication, in which one or two RNA primers are made for subsequent DNA replication. Unidirectional replication occurs if there is one primer and bidirectional if there are two. In RCR, a break is made in the plasmid DNA at a specific site (dso, double-strand origin) on one of the DNA strands (the + strand), which is subsequently “rolled-off” while a new strand is synthesized in its place. RCR plasmids are small and none have been found larger than 10 kb. A tendency for genetic rearrangement of RCR plasmids may be the cause for the apparent size limitation.

Regardless of the type of replication, copy number is maintained by negative feedback acting on replication initiation by one of two mechanisms. In the first, the plasmid produces a small diffusible RNA that inhibits an essential step of replication. The RNA concentration, determined by the rates of RNA production and degradation, regulates the timing of replication. In the second, a set of directed repeats (known as interons) are located within the origin of replication and serve as binding sites for a replication initiation protein. Plas-
mids having the same initiation control are said to be incompatible because negative feedback on copy number prevents two or more plasmids of the same initiation type from being maintained in the same bacterial cell. The degree of initiation type similarity causing incompatibility varies depending on the precise nature of control. A single base change can alleviate incompatibility for some types of diffusible RNA control, whereas interon incompatibility is more tolerant of base pair substitutions.

B. Informational Structure

Diversity among bacteria in the informational structure of the genome exists at almost all levels (e.g., GC bias and codon usage). An exception is the operon, a type of gene organization common to all bacteria.

1. GC Bias

The GC:AT (guanine + cytosine to adenine + thymine) ratio of nucleotide base composition varies greatly among bacterial genera, from 23% GC in Mycoplasma to 79% GC in Micrococcus, strongly contrasting with the smaller ranges of invertebrates (32–45%), vertebrates (40–49%), and plants (33–48%). Also unlike many warm-blooded vertebrates, GC base composition is relatively uniform over bacterial genomes and is not partitioned into homogeneous stretches known as isochores. The diversity of GC bias across bacterial lineages is conserved at the genus level, suggesting that base composition is generally stable over long periods of evolutionary time. There are two hypotheses for diversity of base composition across bacterial species. One hypothesis for the diversity in GC bias is selective constraint. For example, thermally stable amino acids that are encoded by GC-rich codons may be preferred in thermophilic bacteria. The second hypothesis is mutational pressure. Direct measurements of mutational pressure must be interpreted with care.

2. Codon Usage

The genetic code is almost completely universal across all life, with a few minor differences in some bacteria such as Mycoplasma and in bacterial-derived lineages such as mammalian mitochondria. More surprising is the variation in codon usage. As mentioned previously, the genetic code is very degenerate, with 61 codons coding for only 20 primary amino acids. Given the degeneracy in the genetic code, codon usage would be expected to be random for synonymous codons. However, synonymous codon usage varies across bacterial lineages, genes, and gene segments. Different bacterial lineages have highly biased codon usage, even taking into account GC bias. A large fraction of codon bias is associated with gene expression because highly expressed genes generally contain codons with abundant cognate tRNAs, which differ among species.

Differences in codon usage also exist for the initiation codon. Approximately 90% of E. coli genes have AUG as the ribosome start site for mRNA translation. In E. coli GUG or UUG can also serve as initiation sites (in this order of frequency and translational efficiency) but have a lower translational efficiency. The use of a non-AUG start codon is thought to be an adaptation to limit the production of a particular gene product in E. coli. For example, the use of a non-AUG start codon for adenylate cyclase limits the production of this protein, which is toxic at high levels. In contrast, approximately 30% of B. subtilis genes have non-AUG initiation, the order of frequency and translational efficiency is AUG > UUG > GUG, and the initiation codon has less influence on gene expression.

3. Gene Organization and Expression

At the most fundamental level, bacterial gene organization and expression is no different among bacterial species or from that of eukaryotes. All genes can be categor-
rized into approximately three classes: protein-coding genes, RNA-specifying genes, and nontranscribed regulatory genes. Genes of the first two classes undergo transcription by a RNA polymerase, and a protein product is made from genes of the first class by rRNA translation of a mRNA transcript. However, there are numerous differences between eukaryotic and bacterial gene organization and expression. Unlike in eukaryotes, which contain multiple RNA polymerases, all bacterial RNA (mRNA, tRNA, and rRNA) are synthesized by a single RNA polymerase. Splicing of mRNA transcripts to remove introns, which are intervening noncoding regions of a gene, appears to be extremely rare in bacteria but is common in eukaryotes. The lack of a nuclear membrane in bacteria allows mRNA translation to begin prior to completion of transcription. Bacterial genes are organized into physically linked structures called operons.

a. The Operon

An operon is a transcriptional unit which minimally contains signals for DNA to RNA transcription start and stop sites but which also usually contains regulatory elements as well as RNA to protein translation-specific sites (Fig. 4). A monocistronic operon contains only a single gene, but most genes occur in polycistronic operons encoding multiple genes associated with a related function. A cistron is a DNA segment encoding a polypeptide and genes in a polycistronic operon are jointly regulated. At the most simple level, an operon has many typical elements. A repressor/activator binding site allows for the binding of a trans-acting factor which affects RNA polymerase binding that may itself be modulated by corepressors/inducers. The promoter region binds the RNA polymerase and a promoter-specific additional element, a σ factor, which is required for recognition and binding to the transcription initiation site. Since a single bacterium contains a variety of σ factors, global regulation of different sets of proteins can be achieved by expression of different σ factors. The most abundant σ factor found in all bacteria, σ^{70}, is essential for vegetative growth and its DNA binding is affected primarily by two DNA sequence sites, regions approxi-

![Operon structure and mechanisms of transcriptional regulation](image)
mately 35 and 10 bases upstream of the transcription start site. Of the more than 130 promoters in E. coli that have σ70 binding promoters, 69% are repressible, 49% are activatable, and 17% contain both repressor and activator sites.

Once the RNA polymerase has made a short piece of RNA of about 12 bases, the σ factor dissociates from the RNA polymerase and elongation of the mRNA transcript continues until terminated in one of two ways. Termination may occur by a combination of a GC-rich sequence followed closely by a run of adenines. The GC-rich region is thought to form a stem loop structure which causes the polymerase to pause and decreases the affinity of the RNA polymerase to the nascent RNA. The run of A residues results in the nascent RNA having a poly-U RNA, and A-U hybrid base pairing is relatively unstable. A second type of termination is dependent on the RNA binding protein Rho. Rho is thought to track behind RNA polymerase and terminate transcription if the polymerase pauses.

Operons encoding genes to be translated contain one or more start codons and ribosome binding sites (often known as Shine–Dalgarno sequences). Translation is initiated by the binding of the small ribosomal subunit (30S) with associated factors to the mRNA start codon and ribosomal binding site, a region approximately five bases upstream of the start codon which is complementary to the 16S rRNA component of the 30S subunit. Binding to the initiation codon by the appropriate tRNA carrying the bound cognate amino acid then leads to binding of the 50S large ribosomal subunit and formation of the bacterial 70S ribosome. Posttranscriptional regulation of expression can occur at both the ribosome binding site and the initiation codon, by the binding of translational repressors, making either or both the ribosome binding site and initiating codon inaccessible. Once begun, translation continues at an average rate of 15–20 amino acids per second until the ribosome reaches a stop codon, at which it pauses and binds a release factor which causes the completed polypeptide to be released, and the ribosome is free to dissociate from the mRNA.

b. Global Regulation

Unlike many eukaryotes, bacteria are extraordinarily proficient at altering their phenotype by gene regulation. By alteration in their patterns of gene expression, bacteria are able to take advantage of new nutrient sources, respond to harsh environmental conditions, and take alternative developmental pathways. The classification of gene regulation in bacteria is hierarchical. Operons allow for the coordinated expression of related genes, such as those for lactose utilization (e.g., the lac operon). Regulons allow for coordinated expression of multiple operons in response to specific substrates, such as those for the four glycerol utilization operons in E. coli. In the glycerol regulon, a single repressor negatively inhibits expression of the four independent operons altering glycerol uptake and metabolism under aerobic and anaerobic conditions. Regulatory units composed of groups of operons and regulons are sometimes known as modulons. Modulon regulation is generally in response to some aspect of environmental change and the component genes typically have related tasks or activities. The aforementioned σ factors are one mechanism of global regulation, of which E. coli has at least six (Table II). All but one of the E. coli σ factors (σ70) have four regions of sequence conservation. All known bacterial σ factors have sequence similarity with either σ70 or σ54.

c. Mutational Regulation

High-frequency mutations in the genome can effectively regulate gene expression. In E. coli, the mutation rate per gene per generation is typically between $2 \times 10^{-8}$ and $2 \times 10^{-9}$, but specific sequences (contingency loci) may have up to a $10^{5}$ higher mutation rate. Moreover, many contingency loci usually alternate between one of two states so that the variety of outcomes is pro-

### TABLE I

<table>
<thead>
<tr>
<th>Main Bacterial Lineages (Phyla)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidobacterium</td>
</tr>
<tr>
<td>Bacteroides/Cytophage</td>
</tr>
<tr>
<td>Chlamydiae</td>
</tr>
<tr>
<td>Chloroplast</td>
</tr>
<tr>
<td>Fibrobacter</td>
</tr>
<tr>
<td>Gram-positive bacteria</td>
</tr>
<tr>
<td>Leptospirillum</td>
</tr>
<tr>
<td>Pelagibacter</td>
</tr>
<tr>
<td>Synechocystis</td>
</tr>
<tr>
<td>Thermodesulfobacterium</td>
</tr>
<tr>
<td>Verrucomicrobiales</td>
</tr>
</tbody>
</table>
### Table II
Factors of *Escherichia coli*

<table>
<thead>
<tr>
<th>Name</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>σ^{21}</td>
<td>Iron citrate transport</td>
</tr>
<tr>
<td>σ^{24}</td>
<td>Flagellar gene expression</td>
</tr>
<tr>
<td>σ^{32}</td>
<td>Temperature acclimation</td>
</tr>
<tr>
<td>σ^{38}</td>
<td>Nongrowth stationary phase conditions</td>
</tr>
<tr>
<td>σ^{54}</td>
<td>Nitrogen use</td>
</tr>
<tr>
<td>σ^{70}</td>
<td>Vegetative growth</td>
</tr>
</tbody>
</table>

* The numerical designation refers to the molecular weight of the σ factor.

The enhanced mutation rate increases the likelihood that at least one bacterial cell of a related population will be expressing the appropriate phenotype if environmental conditions change. Two mechanisms are known: intracellular recombination and strand slippage of DNA repeats. Recombinational regulation of cell surface components occurs in many bacteria, typically in endosymbionts avoiding host immune responses. Phase variation in *Salmonella typhimurium* is the most well-known system and is due to an invertible promoter. In one promoter orientation one surface antigen (HagA) is produced, and in the other a second antigen is produced (HagB). The promoter is flanked by a pair of repeat sequences which facilitates site-specific inversion.

Strand slippage is the insertion or deletion of nucleotides during DNA replication. This can alter the spacing between regulatory elements of an operon, enhancing or eliminating gene expression. In *Bordetella pertussis*, the bacteria causing whooping cough, a run of C nucleotides in the promoter region of a surface antigen (*fim* genes) results in frequent strand slippage and displacement of the activator binding site. Strand slippage can also result in frameshift mutations by the insertion or deletion of nucleotides other than in sets of three. All codons downstream of frameshift mutations are out of the correct reading frame, preventing expression of the protein. Runs of short DNA repeats of two, four, five, or longer greatly enhance the likelihood of strand slippage. The antigenic variation of *Neisseria gonorrhoeae* proteins essential for infection (*Opa*) is maintained by a five base repeat, CTCTC.

### III. GENETIC EXCHANGE

Genetic exchange in bacteria was demonstrated some time ago, but it was not until quite recently that it was thought to have evolutionary importance for most species. The effects of recombination are widespread, affecting the genetic structure of bacterial populations, patterns of coevolution, and gene–gene interaction.

Even so, for many species recombination is sufficiently rare that gene linkage is maintained within species. There are three mechanisms for genetic exchange among bacteria: transduction, conjugation, and transformation. All three mechanisms involve the one-way transfer of genetic material from a donor to a recipient but differ substantially in the means and control of transfer. Transduction is controlled by viral genes, conjugation by plasmids and transposons, and transformation by the bacterial chromosome.

#### A. Transduction

Transduction is the transfer of bacterial genetic material from one bacteria to another by a virus (a bacteriophage). Normal phage reproduction involves subversion of the host cell to produce viral-specific DNA and protein and release of infective phage containing phage DNA from the host cell. Occasionally, host cell DNA is incorporated into a phage so that the phage acts as a "vector" transferring DNA from one bacterial cell to the next. The transduced DNA can then be incorporated into the host genome by recombination. The likelihood of transduction depends on the sequence similarity between the host chromosome and the transduced DNA. Incorporation of exogenous DNA, brought into the cell by transduction or other means, drops off exponentially even with only 1% sequence divergence.

Two types of transduction are known—specialized and general—and differences in the life cycle of phage determine the possible routes for transduction to occur. Specialized transduction results when a hybrid host–viral DNA molecule is packaged into the virus, most commonly by the action of an integrative temperate phage. Temperate phages have two alternative pathways for reproduction: (i) a lysogenic pathway in which the phage genome integrates in the host genome and is replicated passively as part of the host genome and (ii) a lytic pathway in which the phage genome replicates independently and leads to the production of infective phage and lysis of the host cell. Transition from lysogenic to lytic pathways involves excision of the phage from the host DNA, and imprecise excision can result in packaging of host DNA adjacent to the site of phage DNA integration. In contrast, generalized transduction occurs as a result of errors in DNA packaging. For some phage, packaging of phage DNA is initiated at phage-specific sequences known as pac sites, and pac-like sites
in the host DNA can result in its packaging into phage instead of the phage DNA. Since generalized transduction results from packaging errors, it can result from both temperate or virulent phage (phage having no lysogenic state) and can readily result in transduction of any portion of the host DNA, not just DNA adjacent to a phage DNA integration site.

B. Conjugation

Conjugation is the direct transfer of genetic material from a donor cell to a recipient cell via cell-to-cell contact. The genes transferred are usually those required for conjugation and are typically encoded on self-transmissible genetic elements such as conjugative plasmids. A variety of such elements have been identified, the most well-studied being the F plasmid of E. coli. Cells bearing an F plasmid (F+ cells) extend one to three hair-like structures (sex-pili) which contact F− cells (F plasmid-free cells) and then pull together the F+ and F− cells. Upon contact of the F+ and F− cells, DNA replication of the F plasmid is initiated and plasmid transfer from the F+ donor to the F− recipient begins while plasmid rolling circle replication in the donor is ongoing. If mating is not interrupted, for example, by vigorous agitation, then the entire plasmid is transferred and the recipient becomes F+.

Genes other than those specifically for self-transmission can also be transferred by conjugative plasmids in several ways. First, R plasmids encode not only conjugal machinery but also antibiotic resistance genes. Multiple antibiotic resistances can be encoded by a single plasmid so that antibiotic resistance can quickly spread, as has been commonly observed in hospitals. Second, plasmid mobilization, the transfer of non-conjugative plasmids, can be achieved by co-occurrence of conjugative and non-conjugative plasmids in the same bacterial cell. Most plasmids appear to encode the necessary genes for mobilization even if they do not encode the complete conjugative DNA transfer apparatus. The conjugative plasmid provides those functions essential for transfer of the non-self-transmissible plasmid. Third, transfer of host genes can occur as many conjugative plasmids integrate into their bacterial host genome. Bacterial strains carrying an integrated F plasmid transfer bacterial host genes to F− cells at a high frequency and are known as Hfr (high frequency of recombination) strains.

Although conjugative plasmids are frequently the cause of conjugation in gram-negative bacteria, conjugation in gram-positive bacteria is often due to conjugal transposons. A transposon is a transposable genetic element that is able to move directly from one DNA sequence site to other sites within the genome of a bacterial cell. At a minimum, a transposon contains only those elements necessary for it to move about the genome, but transposons may contain other genes. Even without conjugative ability, transposons may carry antibiotic genes and thus provide a rapid mechanism for movement of such genes to and from conjugative plasmids. However, some transposons are able to move from cell to cell via conjugation, although little is known about their mechanism of intercellular transfer. Gene transfer by this method is likely to be an important factor in the spread of antibiotic resistance because all known conjugative transposons characterized to date carry a tetracycline resistance gene.

C. Transformation

Genetic transformation results from uptake of free DNA from the environment and its integration into the host genome. Techniques of artificial transformation of bacteria are a cornerstone of modern molecular biology, but some bacterial genera (e.g., Streptococcus, Neisseria, Bacillus, and Haemophilus) have evolved mechanisms for frequent natural transformation. In laboratory cultures of B. subtilis, natural transformation and recombination begins as bacteria exhaust available resources. Cells become competent, able to pick up, bind, and internalize exogenous free DNA. There are approximately 50 DNA binding sites on the surface of a component cell and DNA binding and uptake is independent of DNA sequence and source. Once bound, DNA rapidly undergoes double-strand cleavage and DNA uptake commences —1 or 2 min after binding. However, unlike binding and internalization, integration of exogenous DNA is severely reduced by sequence divergence. Internalized DNA that is homologous (having similar or identical base sequence with the host chromosome) can be integrated into the host chromosome. Heterologous DNA (differing in base sequence) is degraded and has no heritable effect on the recipient. The sensitivity of integration to heterologous DNA is high and is sufficient to distinguish different Bacillus species. Mechanisms of transformation differ among other bacterial species. Transformation in Streptococcus pneumoniae appears similar to that in B. subtilis, except that competence is induced in response to a diffusible protein (competence factor) secreted by S. pneumoniae cells. In Haemophilus influenzae, specialized vesicle-like structures, transformosomes, extend from the cell membrane for DNA recognition and uptake. Only ho-
mologous DNA is bound to competent cells, with discrimination relying on the presence of an 11-base pair DNA sequence that occurs approximately every 4 kb in the *H. influenzae* genome.

### D. Evolutionary Effects of Genetic Exchange

In the absence of genetic exchange, bacteria evolve as a series of asexual lineages. A non-recombining bacterial cell, and all the genetic material encompassed within that cell, is derived from a single immediate parental cell and all its non-recombining descendants are similarly ultimately derived exclusively from the same parental cell. A population of non-recombining bacteria initiated from a single cell is thus composed of lineages related in a dichotomously branching tree-like fashion. The strict lineal descent of non-recombining bacteria greatly affects the dynamics and effects of adaptation. If an advantageous mutation occurs in a non-recombining population, the selective benefit conferred by the mutation affects the cell as a whole. The cells carrying the beneficial mutation can increase in frequency, displacing cells not carrying the advantageous mutation. Like the advantageous mutation, the other genes of the original mutant cell increase in frequency as well, "hitchhiking" on the selective benefit of the mutation. If the mutation fixes in the population, the complete linkage of the advantageous mutation with the rest of the genome carrying the mutation will have simultaneously have caused the genetic diversity of the population to plummet. Only those cells descended from the first cell carrying the advantageous mutation persist in the population. Hence, a series of selective sweeps by advantageous mutations, known as periodic selection, can maintain low genetic diversity within nonrecombining bacterial populations. The purging of genetic diversity will occur even if multiple adaptive mutations appear since the mutations in different cells cannot be brought together into a single cell by recombination. In addition, the rate of adaptation within a population of nonrecombining bacteria is limited because multiple advantageous mutations cannot be simultaneously fixed (Fig. 5) (de Visser et al., 1999).

In contrast, multiple beneficial mutations can be fixed in sexually recombining populations and the effects of multiple selective sweeps on extant genetic variation are much reduced. Recombination eliminates the tight linkage of a beneficial mutation to the rest of the genome, with the amount of hitchhiking determined by the rate and extent of recombination. These substantial effects of genetic recombination on evolution have greatly enlivened the study of recombination in bacteria since recombination in bacteria generally appears to be very rare but of large effect.

Evidence of genetic recombination has been found in all bacterial lineages examined at the sequence level, even if recombination has not been observed. The most extensive surveys of recombination have been done in

![Figure 5](image_url)  
**Figure 5** The rate of fixation of three beneficial genes in (a) an asexual and (b) a sexual population. The different patterns denote three different beneficial genes and the width of an area represents gene frequency. (a) In asexual populations, fixation of beneficial genes occurs sequentially due to the lack of recombination. (b) In sexual populations, the fixation of multiple beneficial alleles can occur simultaneously.
E. coli, a bacterial species not known to undergo natural transformation. Divergence among E. coli lineages is 50 times more likely due to recombination than mutation (Guttmann and Dykhuisen, 1994). As much as 13% of the E. coli strain K12 genome may have been acquired from bacteria differing in GC bias from E. coli, and such foreign DNA has accumulated at a rate approximately equivalent to that introduced by point mutations (Lawrence and Ochman, 1997). Genetic exchange is likely to have been a major source of variation in chromosome size among E. coli isolates, which differ by almost 20% (Berghorsson and Ochman, 1998). Although it has often been argued that bacteria exist simply as a series of asexual lineages, with the spread of antibiotic resistance being an exception due to intense selection, genetic exchange has had widespread effects on the genetic structure of bacteria examined so far.

The effects of genetic exchange in many bacteria differ from those of obligate sexual organisms. The combination of selection and generally far less frequent genetic recombination in bacteria maintains linkage disequilibrium, which is the occurrence of gene combinations more frequently than expected by independent assortment (Hauhold et al., 1998). One outcome is the evolutionary effects on operon structure. The tight linkage of the genes comprising an operon allows for the transfer of a complete metabolic function to bacteria in a single event of genetic exchange. Within a single bacterial cell, selection is relatively weak for persistence of the tight linkage of an operon. A second outcome is the potential for rapid spread of a selectively important trait, such as antibiotic resistance, even though environmental conditions are far from ideal. Third, bacteria are phenotypically plastic, allowing them to take advantage of temporally varying environments. Bacteria can drastically alter their requirements for growth, including changes from heterotrophy to prototrophy and from aerobic to anaerobic growth. Many bacteria are able to grow despite rapid changes in osmosis, pH, and temperature, and their ability to withstand rapid changes by ceasing growth. If conditions are sufficiently poor, many can enter a dormant metabolic state, such as a spore, that can remain viable for decades or longer without water or nutrients.

Finally, as mentioned at the beginning of the article, bacterial phenotypes are evolutionarily plastic and adaptation can be finely tuned to environmental conditions. For example, E. coli grown in a spatially and temporally constant environment for 800 generations will diverge from a single lineage into three distinct lineages when provided with glucose as a limiting resource. The dominant lineage is a glucose specialist, the second is an acetate specialist that excretes acetate and glycerol as waste products, and the third is a glyceral specialist. By optimizing different steps in the catabolism of a single limiting resource, otherwise identical bacterial genotypes can coexist in the same environment (Rosenzweig et al., 1994) but can also be seen in nature (e.g., antibiotic resistance).

Thus, it is clear that the evolution of bacterial diversity is due to at least four reasons. First, bacteria are small. This allows them to take advantage of microenvironments arising from spatial heterogeneity—environments that are abundant but incapable of supporting metazoan populations. Such environments include different animal and plant hosts. Adaptation to microenvironments can be readily observed in laboratory cultures of bacteria in as little as 3 days (Rainey and Travisano, 1998). Their small size also allows them to grow to large populations (10^8 cells or more).

Second, environment-specific adaptation can occur readily since genetic recombination is too infrequent to impede adaptation. Frequent recombination could prevent specialization to different niches as the genes involved in specialization to different niches are mixed with one another by recombination. However, even though genetic exchange of selectable important genes can readily be demonstrated, ecological specialization can rapidly evolve. This has been observed in laboratory populations of bacteria several times (Rosenzweig et al., 1998) but can also be seen in nature (e.g., antibiotic resistance).

Third, bacteria are phenotypically plastic, allowing them to take advantage of temporally varying environments. Bacteria can drastically alter their requirements for growth, including changes from heterotrophy to prototrophy and from aerobic to anaerobic growth. Many bacteria are able to grow despite rapid changes in osmosis, pH, and temperature, and their ability to withstand rapid changes by ceasing growth. If conditions are sufficiently poor, many can enter a dormant metabolic state, such as a spore, that can remain viable for decades or longer without water or nutrients.

Finally, as mentioned at the beginning of the article, bacterial phenotypes are evolutionarily plastic and adaptation can be finely tuned to environmental conditions. For example, E. coli grown in a spatially and temporally constant environment for 800 generations will diverge from a single lineage into three distinct lineages when provided with glucose as a limiting resource. The dominant lineage is a glucose specialist, the second is an acetate specialist that excretes acetate and glycerol as waste products, and the third is a glyceral specialist. By optimizing different steps in the catabolism of a single limiting resource, otherwise identical bacterial genotypes can coexist in the same environment (Rosenzweig et al., 1994) but can also be seen in nature (e.g., antibiotic resistance).

Thus, it is clear that the evolution of bacterial diversity is due to at least four reasons. First, bacteria are small. This allows them to take advantage of microenvironments arising from spatial heterogeneity—environments that are abundant but incapable of supporting metazoan populations. Such environments include different animal and plant hosts. Adaptation to microenvironments can be readily observed in laboratory cultures of bacteria in as little as 3 days (Rainey and Travisano, 1998). Their small size also allows them to grow to large populations (10^8 cells or more).

Second, environment-specific adaptation can occur readily since genetic recombination is too infrequent to impede adaptation. Frequent recombination could prevent specialization to different niches as the genes involved in specialization to different niches are mixed with one another by recombination. However, even though genetic exchange of selectable important genes can readily be demonstrated, ecological specialization can rapidly evolve. This has been observed in laboratory populations of bacteria several times (Rosenzweig et al., 1998) but can also be seen in nature (e.g., antibiotic resistance).

Third, bacteria are phenotypically plastic, allowing them to take advantage of temporally varying environments. Bacteria can drastically alter their requirements for growth, including changes from heterotrophy to prototrophy and from aerobic to anaerobic growth. Many bacteria are able to grow despite rapid changes in osmosis, pH, and temperature, and their ability to withstand rapid changes by ceasing growth. If conditions are sufficiently poor, many can enter a dormant metabolic state, such as a spore, that can remain viable for decades or longer without water or nutrients.

Finally, as mentioned at the beginning of the article, bacterial phenotypes are evolutionarily plastic and adaptation can be finely tuned to environmental conditions. For example, E. coli grown in a spatially and temporally constant environment for 800 generations will diverge from a single lineage into three distinct lineages when provided with glucose as a limiting resource. The dominant lineage is a glucose specialist, the second is an acetate specialist that excretes acetate and glycerol as waste products, and the third is a glyceral specialist. By optimizing different steps in the catabolism of a single limiting resource, otherwise identical bacterial genotypes can coexist in the same environment (Rosenzweig et al., 1994) but can also be seen in nature (e.g., antibiotic resistance).

Thus, it is clear that the evolution of bacterial diversity is due to at least four reasons. First, bacteria are small. This allows them to take advantage of microenvironments arising from spatial heterogeneity—environments that are abundant but incapable of supporting metazoan populations. Such environments include different animal and plant hosts. Adaptation to microenvironments can be readily observed in laboratory cultures of bacteria in as little as 3 days (Rainey and Travisano, 1998). Their small size also allows them to grow to large populations (10^8 cells or more).

Second, environment-specific adaptation can occur readily since genetic recombination is too infrequent to impede adaptation. Frequent recombination could prevent specialization to different niches as the genes involved in specialization to different niches are mixed with one another by recombination. However, even though genetic exchange of selectable important genes can readily be demonstrated, ecological specialization can rapidly evolve. This has been observed in laboratory populations of bacteria several times (Rosenzweig et al., 1998) but can also be seen in nature (e.g., antibiotic resistance).

Third, bacteria are phenotypically plastic, allowing them to take advantage of temporally varying environments. Bacteria can drastically alter their requirements for growth, including changes from heterotrophy to prototrophy and from aerobic to anaerobic growth. Many bacteria are able to grow despite rapid changes in osmosis, pH, and temperature, and their ability to withstand rapid changes by ceasing growth. If conditions are sufficiently poor, many can enter a dormant metabolic state, such as a spore, that can remain viable for decades or longer without water or nutrients.

Finally, as mentioned at the beginning of the article, bacterial phenotypes are evolutionarily plastic and adaptation can be finely tuned to environmental conditions. For example, E. coli grown in a spatially and temporally constant environment for 800 generations will diverge from a single lineage into three distinct lineages when provided with glucose as a limiting resource. The dominant lineage is a glucose specialist, the second is an acetate specialist that excretes acetate and glycerol as waste products, and the third is a glyceral specialist. By optimizing different steps in the catabolism of a single limiting resource, otherwise identical bacterial genotypes can coexist in the same environment (Rosenzweig et al., 1994) but can also be seen in nature (e.g., antibiotic resistance).

Thus, it is clear that the evolution of bacterial diversity is due to at least four reasons. First, bacteria are small. This allows them to take advantage of microenvironments arising from spatial heterogeneity—environments that are abundant but incapable of supporting metazoan populations. Such environments include different animal and plant hosts. Adaptation to microenvironments can be readily observed in laboratory cultures of bacteria in as little as 3 days (Rainey and Travisano, 1998). Their small size also allows them to grow to large populations (10^8 cells or more).

Second, environment-specific adaptation can occur readily since genetic recombination is too infrequent to impede adaptation. Frequent recombination could prevent specialization to different niches as the genes involved in specialization to different niches are mixed with one another by recombination. However, even though genetic exchange of selectable important genes can readily be demonstrated, ecological specialization can rapidly evolve. This has been observed in laboratory populations of bacteria several times (Rosenzweig et al., 1998) but can also be seen in nature (e.g., antibiotic resistance).

Third, bacteria are phenotypically plastic, allowing them to take advantage of temporally varying environments. Bacteria can drastically alter their requirements for growth, including changes from heterotrophy to prototrophy and from aerobic to anaerobic growth. Many bacteria are able to grow despite rapid changes in osmosis, pH, and temperature, and their ability to withstand rapid changes by ceasing growth. If conditions are sufficiently poor, many can enter a dormant metabolic state, such as a spore, that can remain viable for decades or longer without water or nutrients.

Finally, as mentioned at the beginning of the article, bacterial phenotypes are evolutionarily plastic and adaptation can be finely tuned to environmental conditions. For example, E. coli grown in a spatially and temporally constant environment for 800 generations will diverge from a single lineage into three distinct lineages when provided with glucose as a limiting resource. The dominant lineage is a glucose specialist, the second is an acetate specialist that excretes acetate and glycerol as waste products, and the third is a glyceral specialist. By optimizing different steps in the catabolism of a single limiting resource, otherwise identical bacterial genotypes can coexist in the same environment (Rosenzweig et al., 1994) but can also be seen in nature (e.g., antibiotic resistance).
See Also the Following Articles
ARCHAEA, ORIGIN OF • BACTERIAL BIODIVERSITY • EUKARYOTES, ORIGIN OF • NUCLEIC ACID BIODIVERSITY

Bibliography


Detecting linkage disequilibrium in bacterial populations. Genetics 138, 1541.


I. INTRODUCTION

In terms of numbers of species, arthropods are the largest phylum of living organisms, insects are the largest group within the arthropods, and beetles are the largest order among the insects. Beetles presently comprise about a quarter of all described, extant organisms. This numerical dominance led the British biologist J. B. Haldane to answer the question of what one might learn about the mind of the Creator from a study of his creation with the famous reply: “An inordinate fondness for beetles.”

In 1982, T. L. Erwin used samples of beetles to estimate that the number of species of living organisms may be 30 million or more. The ensuing controversy (especially Stork, 1988) rekindled interest in the more general question of the magnitude of global biodiversity (May, 1988) and how to measure it. These discussions did not question the usual assumption that beetles represent the largest part of global biodiversity, although mites and nematodes may in fact be more diverse than beetles. More recently and narrowly, a proposal to “explain” the diversity of beetles has itself engendered controversy. Both questions are reviewed here.

II. WHAT BEETLES ARE

The order Coleoptera is usually divided into four suborders and 160–170 families depending on the author. The suborders Archostemata and Myxophaga each con-
The Adephaga includes about a dozen families, with most of the species in the large family of ground beetles Carabidae. All other beetles—150 families and 90% of the 340,000 species—are classified as Polyphaga. Somewhat more than two-thirds of the species belong to the eight largest families discussed here, each with more than 10,000 species, but there are 30 other families that have between 1000 and 10,000 species (Table 1). The most thorough review of the biology of members of the order is found in Crowson (1981).

Because they are well sclerotized, there is a good fossil record for beetles. Beetle-like insects first appeared in the Lower Permian, modern families are known from the Cretaceous, and extant genera date from the Oligocene and Miocene. Ecologically, beetles are the most diverse order of insects, which is a primary factor in their great species richness. There are large families of beetles that live in fresh-water habitats, that are predaceous and parasitic, and that feed on fungi and detritus, but the largest proportion of described species feed in some way on plants. Among these phytotrophicus beetles, the largest number of species are wood-borers; a much smaller number feed directly on leaves, and some feed on seeds.

### III. DIVERSITY—HOW MANY BEETLES ARE THERE?

Gaston (1991) has reviewed current estimates of the number of insect species described, favoring those of J. F. Lawrence for beetles. Lawrence (1982) put the numbers of beetles at 340,500 species divided into about 25,000 genera and placed in 169 families. His estimate of the number of species falls in the middle of estimates (290,000–370,000) cited by Gaston, and roughly equal to the sum of the number of species in the three next largest insect orders (Diptera, 98,000–120,000 species; Lepidoptera, 112,000–165,000 species; Hymenoptera, 100,000–130,000 species): Estimates of the total number of living species, described and undescribed, are difficult to relate to the number of described species for two major reasons. First, most taxonomic research has been carried out in the north temperate regions of the world on organisms that live there, whereas most insect species are tropical. Second, different groups of insects have received differential amounts of study from scientists, both within the beetles and between beetles and the other insect orders. This article briefly reviews what is known about the major beetle families, considers explanations for the success of the order, reviews a few ecological and biogeographic patterns in beetle diversity, and then discusses the problems in estimating what we don’t yet know about beetle diversity.

As stated earlier, eight families of beetles currently account for more than half of all beetle species. However, unequal efforts in the collection and study of different groups very likely mean that the relative sizes of these families will change when the less well studied taxa become better known. Some of the better-studied groups have already begun to be used to test hypotheses of mechanisms that might promote or maintain beetle diversity; for example, whether speciation in plant-feeding beetles parallels specialization in their hosts. The eight major families are considered in order of their traditional phylogenetic placement.

1. Carabidae

The ground beetles (30,000 described species, including the tiger beetles, sometimes placed in the separate family Cicindelidae) are primarily predatory. Although some carabid groups are very rich in tropical regions, others do not show as great an increase in diversity there as do other beetles. Despite the vernacular family name, tropical forms are commonly arboreal. Tiger beetles are well studied taxonomically, have been used to test ecological and evolutionary hypotheses, and have been suggested as indicator species for biodiversity studies (Pearson and Cassola, 1992).

2. Staphylinidae

The rove beetles (30,000 species) are both numerous in species and diverse ecologically, including predatory, fungus, and detritus feeders. The staphylinids are the least well known family of beetles, with perhaps only 10% of extant species described, and may eventually prove to be the largest family. They are abundant in canopy fogging samples in tropical regions.

3. Scarabaeidae

The scarabs (25,000 species) are primarily phytotrophicus, but with one group well known as dung beetles. Because they are popular with collectors, they are among the best known of the beetle groups. The actual number of extant species is unlikely to be more than twice as many as those already described.

4. Buprestidae

The jewel beetles (15,000 species) include a majority of wood-boring species and several groups of leaf-miners. Although larger species are popular with collectors, the
### TABLE I

Diet and Numbers of Species and Genera of the Larger Families of Beetles

(More Than 1,000 Species)

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Genera</th>
<th>Species</th>
<th>A</th>
<th>P</th>
<th>D</th>
<th>F</th>
<th>Ph</th>
<th>W</th>
<th>S</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adephaga Carabidae</td>
<td>1500</td>
<td>30,000</td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Carabidae</td>
<td>120</td>
<td>3000</td>
<td></td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polyphaga Hydrophilidae</td>
<td>125</td>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydrophilidae</td>
<td>200</td>
<td>3000</td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydrophilidae</td>
<td>230</td>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydrophilidae</td>
<td>75</td>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydrophilidae</td>
<td>1500</td>
<td>30,000</td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydrophilidae</td>
<td>430</td>
<td>15,000</td>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydrophilidae</td>
<td>600</td>
<td>9000</td>
<td></td>
<td></td>
<td></td>
<td>Ph</td>
<td>W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydrophilidae</td>
<td>190</td>
<td>1200</td>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeroidea</td>
<td>125</td>
<td>1200</td>
<td></td>
<td></td>
<td></td>
<td>Ph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>450</td>
<td>15,000</td>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>130</td>
<td>3000</td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>30</td>
<td>2500</td>
<td></td>
<td></td>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>120</td>
<td>3500</td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>490</td>
<td>4500</td>
<td></td>
<td></td>
<td></td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>100</td>
<td>3000</td>
<td></td>
<td></td>
<td></td>
<td>Ph</td>
<td>W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>100</td>
<td>1400</td>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>100</td>
<td>3000</td>
<td></td>
<td></td>
<td></td>
<td>Ph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>120</td>
<td>3000</td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>1700</td>
<td>18,000</td>
<td></td>
<td></td>
<td></td>
<td>Ph</td>
<td>W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>4000</td>
<td>35,000</td>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>4000</td>
<td>35,000</td>
<td></td>
<td></td>
<td></td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>325</td>
<td>2600</td>
<td></td>
<td></td>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>100</td>
<td>2200</td>
<td></td>
<td></td>
<td></td>
<td>Ph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>250</td>
<td>1500</td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>325</td>
<td>2300</td>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histeridae</td>
<td>4500</td>
<td>30,000</td>
<td></td>
<td></td>
<td></td>
<td>Ph</td>
<td>W</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Data from Lawrence (1982). Families listed in boldface account for more than two-thirds of all beetle species.

*A = aquatic, D = detritus, F = fungi, L = leaves, P = predatory, Ph = phytophagous, S = seeds, W = wood-boring.
smaller species are primarily subtropical and tropical and very poorly known. The genus Agrilus may be the largest genus of living organisms with 2000 described species and many more undescribed—there are nearly 350 undescribed species among more than 600 known from México alone. At La Selva Biological Station in Costa Rica, 24 of 29 Agrilus (83%) are undescribed, of 206 Buprestidae, 147 (71%) are leaf-miners and 62% are undescribed. The family should more than double in size when completely known.

5. Tenebrionidae
The darkling beetles (18,000 species) are phytophagous and have been relatively popular and well studied in temperate and subtropical regions. I have seen no discussion of its current taxonomic status or estimates of undescribed species.

6. Cerambycidae
The long-horned beetles (35,000 species) are wood-borers, very popular among collectors, and well studied. Of 274 species collected at La Selva, only 23 (8%) are undescribed, although another 28 (10%) are still undetermined. It is possible that the number of species will double when the rich South American fauna is more fully known. This and the following two families constitute the Phytophaga and are the largest beetle families in described species.

7. Chrysomelidae
The leaf beetles (35,000 species) are by far the largest group of beetles that primarily feed on the leaves of plants. Some groups within the family are relatively well studied, even within tropical areas (Cassidinae), whereas others are not (Alticinae). Members of the family are common in canopy samples and have been shown to speciate in parallel with their plant hosts. The success of beetles overall has been attributed to the success of this and the other two families in the Phytophaga. When the total chrysomelid fauna is known, it will probably double the number of currently known species.

8. Curculionidae
The weevils (50,000 species) are the largest family of living organisms. Although they are primarily phytophagous and associated with angiosperm plants in a variety of ways, many (and perhaps most) are wood-borers (many Cryptorhynchini, Molytini, and Zygoptini) rather than leaf-eaters, in some cases cultivating fungi in their tunnels (Platyphrynini and some Scolytinae). The great majority of weevils are endophytic; the few that feed directly on leaves often do so as leaf-miners. Sampling at tropical localities has yielded large numbers of weevils. At Barro Colorado Island in Panama, three years of light trapping (Wolda et al., 1998) yielded 1240 species of Curculionoidea (excluding Platypodinae and Scolytinae, but including about 200 species from four other families). Sampling at La Selva has collected well over a thousand species of weevils, including 504 in the largest tribe, Zygopini, of which 425 (84%) are undescribed. The number of species is probably two to three times those presently known.

IV. EXPLANATIONS FOR THE SPECIES RICHNESS OF BEETLES
Farrell (1998) recently claimed to have verified an earlier proposal that the great species richness of Coleoptera is due to the association of the largest lineages within the Phytophaga (Cerambycidae, Chrysomelidae, Curculionidae) with angiosperm plants. Indeed, beetles feed on angiosperms in a number of ways. The narrowest form of phytophagy is to feed on leaves (foliivory), but very few beetles do that, primarily Chrysomelidae and leaf-miners in several families. As foliivores, the Lepidoptera are much more successful (most of 140,000 species). A few phytophagous beetles feed on seeds (Bruchidae, some Curculionidae), but most are wood-borers (Buprestidae, Cerambycidae, many Curculionidae). Perhaps more than qualities of the host itself, it is the ability of beetles to feed endophagously in wood and other plant tissues that makes them so successful. Anderson (1995) has attributed the success of weevils, the most species-rich lineage, to the evolution of the rostrum as an aid to oviposition in plant tissues. It is more difficult for Diptera and Lepidoptera to feed endophagously because adults lack chewing mouthparts with which to exit plant tissues.

Another problem with the Farrell hypothesis is that even if the Phytophaga were removed from the beetles, the remaining taxa are still twice as numerous as any other insect order. If the Staphylinidae are as species rich as has been suggested (300,000 species), that family alone would be more than twice as large as the currently known Phytophaga and about the same size as estimates of total Phytophaga. The success of beetles is almost certainly due to their use of so many major resources, not just their intensive use of one of them. Of the 8 largest beetle families, the Carabidae are primarily predaceous and the Staphylinidae are partly or largely so; 10 of the other 30 families with more than 1000 species are also primarily predaceous. Five of the larger...
families and some Staphylinidae feed on fungi, which is another major adaptive zone for beetles. Aquatic habitats are used by two large families of beetles and only by Diptera among the other three major orders. Ecological samples of canopy insects, although dominated by phytophages, include significant portions of species of other feeding types. If only half of all beetle species feed on angiosperm plants, it could be argued that beetles are in fact less successful than expected, for the reason that angiosperms constitute a much larger resource in terms of productivity and biomass than the other trophic levels in which beetles feed with equal success. In sum, there is no single simple explanation for the great diversity of beetles. Although association with angiosperms has obviously been a major factor in the diversification of beetles, it has been so in a very complex way that may be due less to the plants as a resource and more to the features of the beetles ("preadaptations") that allowed them to use the plants endophagously.

V. TAXONOMIC AND ECOLOGICAL PATTERNS

Most large taxa have greater numbers of species in tropical regions than in temperate ones, and that pattern is likely to hold for families of beetles. Increase in species numbers at lower latitudes is not usually a matter of uniform increases in all genera within the taxon, but rather of changes in dominance among taxa. For example, in the Buprestidae, wood-boring genera dominate in temperate and subtropical regions both north and south of the equator, but leaf-mining genera dominate near the equator. Among the leaf-miners, the genera Brachys and Taphrocerus dominate in temperate and subtropical regions, but the genera Hylaeogena, Leiopleura, and Pachyschelus dominate near the equator. In the Chrysomelidae, the subfamily Clytrinae has a peak in species richness in the subtropics, as does the zygoiptine weevil genus Cylindrocopturus. More complex relationships between species diversity and ecological diversity are suggested by some data. The wood-boring Cerambycidae increase in absolute species richness as one moves towards the tropics, but F. T. Hovore (unpublished data) has found that species richness relative to potential woody plant host species diversity actually decreases at lower latitudes because of a greater increase in the number of woody plant taxa (Fig. 1). This is not unexpected in that many tropical plant species are rare, and Southwood and others have shown that rarer plants have fewer host-specific herbivores than do more common plants. In the tropical Cerambycidae, many of the species are associated with a few larger, more abundant host plant genera such as Inga and Ficus, and host generalists may be common. Mimicry is more common in tropical beetle species than in temperate members of the same taxon. The proportion of species participating in mimicry complexes has been shown to increase with decreasing latitude in Agrilus (Buprestidae) mimicking the ant genus Zacryptocerus, and in zygoiptine weevils mimicking a variety of models. On the other hand, mimicry of and by clytrine chrysomelids is most frequent in the subtropical regions of México, where the clytrines are most species rich.

VI. HOW MUCH DON'T WE KNOW?—LOCAL FAUNAS

Estimates of how many beetle species exist in various families and the contribution of beetles to global biodiversity should be placed in the context of our extensive, basic ignorance of the ecology and distribution of beetles at local and regional scales. Except perhaps for England, there is no place on earth of any size and ecological diversity where one can make a complete list of all beetle species. In areas where checklists have been attempted, the larval biology and population ecology of many or most species are completely unknown. Lists of described species are incomplete, and a significant number of species are undescribed in all but the most
popularly collected taxa. A few small taxa have been surveyed both intensively at a few sites and more or less extensively at scattered sites (Pearson and Cassola, 1992). In general, recent extrapolations to global diversity estimates have been made from samples of local faunas.

Beetles occupy such a wide variety of ecological roles in communities that no single method of sampling is adequate, even when the taxonomic scope of the beetles studied is relatively narrow. In recent years, Malaise trapping and canopy fogging have become popular, but flight intercept traps, sweep netting, blacklighting, Berlese sampling, pitfall traps, bating, beating, systematic rearing, and other methods have been used as well. Malaise traps sample active, flying insects (e.g., Mordellidae), and beetles in Malaise samples have been used as indicators of specific habitats. Light trapping (Wolda et al., 1998) is successful in sampling cryptorhynchine and molytine weevils, but is poor for diurnally active zygoptines and baridines. Canopy fogging has been widely used in recent years to sample beetle faunas. Several major projects have attempted to inventory the arthropod faunas of single, diverse localities. The Arthropods of La Selva (ALAS) project in Costa Rica, for example, uses a combination of four standardized sampling techniques (Malaise trapping, canopy fogging, litter sampling, and light trapping), as well as group-specific collecting methods, to survey a variety of arthropods including several beetle groups. A similar variety of sampling techniques were employed in Sulawesi (Hammond, 1993). These and other such projects in tropical regions have been slow in producing more than partial faunal lists because of the massive taxonomic problems associated with rich tropical faunas that have been largely unstudied (see the following sections). The current ecological concept of metapopulations raises questions about the meaning of local faunas by suggesting that local diversity is a dynamic function of regional diversity. The idea that some “sink” populations are maintained by regular immigration from other localities is popularly collected taxa. A few small taxa have been surveyed both intensively at a few sites and more or less extensively at scattered sites (Pearson and Cassola, 1992). In general, recent extrapolations to global diversity estimates have been made from samples of local faunas. 

The problem of rare species has both ecological and taxonomic–evolutionary implications. Rare species are difficult to sample and difficult to interpret when they are sampled. Apparent rarity can be due to several factors, including methodological artifacts, disturbance, or "tourists," seasonal phenology, cyclic populations, or true biological rarity. Furthermore, small populations may be characteristic of tropical species. Three years of light trapping by Wolda at Barro Colorado Island, Panama, yielded over 95,000 weevil specimens, but 28% of the 1239 species were represented by a single specimen. In smaller samples at six other localities in the same study, unique specimens represented 39–51% of the species. Canopy fogging samples show comparably high proportions of rare species. Reliable taxonomic decisions about what is or is not a valid biological species often require series of specimens to interpret variation; adequate series for studies are even more necessary when specimens from several localities are involved. Many described tropical species are known only from unique type specimens. The question of what is a species has been justifiably raised in discussions of biodiversity at regional and global geographic scales. Erwin (1982) startled the scientific world by calculating that 1200 beetle species (955 counted, 206 weevils estimated, then "rounded up" to 1200) sampled from 19 individual trees of Luehea semmanni implied that there were 30 million species of insects worldwide. His extrapolation depended on three important quantities: his estimate of the proportion of host-specific beetles, the fraction found exclusively in the canopy, and the assumption that all tree species have the same number of herbivores. The last of these is certainly not true—rare tree species have fewer host-specific insects (see earlier discussion).

Empirical studies that sample both the canopy and lower levels in the forest have not shown a large and/or distinct canopy fauna. Gaston (1991) interviewed taxonomic specialists of many insect taxa, including beetles, and concluded that the proportions of undescribed species in tropical samples were high but not high enough to warrant an estimate of 30 million species (if 1 million species are known and 30 million exist, undescribed species would comprise 97% of samples). Other means of estimating global species richness have been used. Rates of description of new species of beetles (“trend lines”) have been extrapolated to an asymptote, but this method incorrectly assumes constant levels of scientific study and publication. May (1988) compared observed with theoretical distributions of body size and abundance to estimate the numbers of undescribed species, primarily those in the smaller size classes. It is certainly true that larger organisms are more likely to be collected and described than
are smaller organisms. In each of the other major orders of insects, it is the smaller species that are least well known; for Lepidoptera, it is the microlepidoptera; for Hymenoptera, the smaller parasitic wasps in the Chalcidoidea; for Diptera, the small Nematocera and Phoridae. In beetles, Gaston showed that larger beetles were described earlier and that the average size of the species being described decreases as one nears the present. Of the zygoptine weevil species identified at La Selva, 57% of weevils larger than 4 mm in length are undescribed (88 measured), but 81% of those less than 4 mm are undescribed (146 measured). Methods of estimating unsampled species from theoretical abundance frequency distributions depend on the mode of the distributions being defined, which is not usually possible for tropical samples. It seems to be true for all estimates of global beetle or insect species numbers, regardless of the sampling method used, that what is not known is so large compared to what is known that a precise estimate is impossible.

VIII. WHAT DO WE KNOW AND HOW LIKELY ARE WE TO EVER KNOW?

Beetles are a very large group of organisms and are diverse both taxonomically and ecologically. The 169 families and 340,000 described species currently comprise somewhat less than half of all known insects and about a quarter of all living organisms. About 40 families of beetles have more than 1000 described species, and 8 families have more than 10,000 species. About half of all beetle species feed on plants, usually as endophages on woody or other support tissue, but there are a large number of predatory species, as well as lesser but significant numbers of fungivores, detritivores, and aquatic forms. This ecological diversity precludes simple explanations for the diversity of beetles. However, many beetle families show general patterns of distribution, for example, greater diversity in tropical latitudes and the greater importance there of antipredator defenses such as mimicry. These crude generalizations aside, we actually know rather little about most beetles. Because of the popularity of certain beetle groups, the various families have been unequally studied. Parts or all of the families Carabidae, Scarabaeidae, Buprestidae, Tenebrionidae, and Cerambycidae have been relatively better studied, whereas the three largest families (Staphylinaeidae, Chrysomelidae, and Curculionidae) are the most poorly studied. Ironically, as we have become aware of the existence of large numbers of undescribed species by using new sampling methods in previously little-collected tropical areas, the number of professional taxonomists able to name and describe species has been declining. Many positions for taxonomists have been lost at most major museums, funding for taxonomic research has declined, and the trend in research has been toward genetic analyses for the purpose of elucidating phylogenetic relationships rather than describing species. Money has become available for sampling and conducting species inventories, especially in threatened tropical areas, but not for processing and studying the material collected beyond simple sorting of a few target or local taxa and counting of their putative "morphospecies." May (1988) showed that, although beetles are the largest group of organisms, they have the lowest rate of publications per described species of all organisms except nematodes. He expressed amazement that there is, in fact, no complete worldwide catalog of insects that would allow one to know how many are actually described (hence the varying "estimates" of even the numbers of described species).

At the level of local or regional faunas, there are few sites outside of Western Europe where beetle faunas are more than partially known. The richest areas—the tropics, and especially the Neotropics—are virtually unknown faunistically outside of samples taken at a few widely scattered sites. Many large regions of the tropics have never been significantly collected for any beetle groups. Costa Rica's National Biodiversity Institute (INBio), including the ALAS project at La Selva, has made the most extensive collections toward a regional fauna of any tropical country, but their collections favor larger species, are only partially mounted and sorted, and are only beginning to be studied. Overall, then, there is no good estimate of global species richness for beetles and few data on which to base such an estimate. As the magnitude of how much we still don’t know becomes clear, there are serious questions about whether we will ever understand the extent of beetle diversity in light of the declining number of taxonomically trained scientists and the possibility that many species will go extinct because of human activities before they are ever collected or studied. The decline in the number of scientists who are able or willing to describe species is unlikely to be halted. Furthermore, widespread beetle extinction over the next century is especially likely in the face of the accelerating environmental destruction of the richest sites in the tropics to accommodate growing human populations and, even more significantly, to maintain Western European and North American standards of living. Unfortunately, it
is a safe guess that we will never completely know how many beetles there were.

See Also the Following Articles
ARTHROPODS, AMAZONIAN • INSECTS, OVERVIEW • INVERTEBRATES, TERRESTRIAL, OVERVIEW

Acknowledgments
I thank F. T. Hovore for reviewing this manuscript and allowing me to publish some of his data. I have been supported by the Research Committee of the UCLA Academic Senate and by grants in support of the ALAS Project (National Science Foundation Grants BSR 9025024, DEB 9401069, and DEB 9706979).

Bibliography
BIODIVERSITY AS A COMMODITY

Geoffrey Heal
Columbia University and Stanford University

I. Biodiversity and Productivity
II. Biodiversity and Insurance
III. Biodiversity and Genetic Knowledge
IV. Biodiversity and Ecosystem Services
V. Biodiversity and Markets
VI. Conclusions

GLOSSARY

first theorem of welfare economics: If all goods are private and all private and social costs are equal, then an economy with a complete set of competitive markets operates in a way that is Pareto efficient.

insurance: A contract under which a person or institution buys the right to be compensated in the event of a specified harmful occurrence. The payment, referred to as the insurance premium, is paid before it is known whether the harmful event will occur.

intellectual property rights: The rights of developers of ideas and techniques to require payment for their use by others and to prevent their use by others unless such payment has been made.

invisible hand: The term used by Adam Smith to describe the capacity of a decentralized market system to attain efficiency.

pareto efficient: A pattern of operation of an economy is said to be Pareto efficient if it is impossible to change its operation so that everyone gains or at least someone gains and no one loses.

public good: A good whose consumption is nonrival and nonexcludable. Nonrival means that one person's consuming it does not preclude another from doing likewise. Nonexcludable means that the provider of the good cannot ensure that only those who have paid can benefit from its provision. Knowledge is a public good: My knowing something does not conflict with your knowing the same fact, and those who develop knowledge cannot ensure that only people and institutions that have contributed to the costs can benefit. An apple, in contrast, is a private good: If I eat it, you cannot. Also, apple producers can ensure that only those who pay for them can eat them.

social costs: The total costs to society of an action. These may exceed (or in some cases be less than) the private costs, which are the costs of that action to the individual or institution executing it.

WHAT IS THE NATURE OF BIODIVERSITY as an economic commodity and why does it matter? How would its conservation contribute economically to our well being? I consider three issues: Why is biodiversity important from an economic perspective? What kind of commodity is it? and Does our usual economic mechanism, the market system, have the capacity to appreciate the economic value of biodiversity? I first characterize biodiversity from an economic perspective, and then
consider the capacity of our main economic institutions to realize the value of biodiversity and ensure that it is treated in a way commensurate with its importance.

Many distinguished scientists have argued strenuously for the conservation of biodiversity as a significant human priority (Ehrlich, 1988; Raven and Williams, 1999; Wilson, 1988). Because economic activity is the main driver of biodiversity loss, such conservation would undoubtedly have important economic implications. It would require economic changes and would certainly be associated with economic costs, although these could be more than offset by the gains. The proposal implicit in these arguments is in effect that we—society—should "buy" biodiversity by changing our economic modus operandi and incurring conservation costs.

What are we buying if we make such a deal? What is the nature of biodiversity as an economic commodity and why does it matter? How would its conservation contribute economically to our well being? These are the themes I address in this article. I consider three issues:

What are the economic functions of biodiversity?
What kind of commodity is biodiversity?
Does our usual economic mechanism, the market system, have the capacity to appreciate the economic value of biodiversity?

In other words, I first characterize biodiversity from an economic perspective, and then consider the capacity of our main economic institutions to realize the value of biodiversity and ensure that it is treated in a way commensurate with its importance. With regard to the first of these tasks, assessing why biodiversity is important economically, I draw extensively on recent literature in ecology, which I do with some trepidation because this is not my field of expertise. However, this is unavoidable: A serious analysis of the economic contributions of biodiversity has to draw on scientific understanding of how diversity contributes to the functioning of the natural environments that provide crucial infrastructure to human societies.

I begin with an attempt to explain what biodiversity contributes, in economic terms, to human societies. Why is biodiversity important? The reasons can be classified under the following headings. Biodiversity provides or enhances:

- Ecosystem productivity
- Insurance
- Knowledge
- Ecosystem services

There is some overlap between these concepts, but nevertheless they are helpful as a guide to thinking through the issues. All of them are economically important categories.

I. BIODIVERSITY AND PRODUCTIVITY

How does biodiversity contribute to productivity? There is experimental evidence that plant systems with more biodiversity are on average more productive than those with less biodiversity. A good illustration of this is work done by David Tilman at the University of Minnesota (Tilman and Downing, 1994; Tilman et al., 1996, 1997). He planted many similar plots of land with a variety of grassland plants, some with many species and some with a much smaller number. Each plot was planted with the same mix year after year, and several indicators of plot performance were recorded, including the amount of biomass grown and the proportion of the nutrients available that were taken up by the plants. Biomass refers to the total dry body weight of the plants. It is a measure of the amount of carbon from the atmosphere that is photosynthesized into carbohydrate. Tilman and others performing similar studies found that, on average, during a period of about 20 years, plots with a more diverse collection of species performed better than those with a less diverse collection.

What does more or less diverse mean here? There are two dimensions to diversity: diversity of functional groups, or of plant types, and diversity of plant species within a functional group. Plants are classified into functional groups on the basis of their intrinsic physiological and morphological characteristics, such as whether they fix nitrogen, have three carbon or four carbon photosynthetic pathways, or are woody (Tilman et al., 1997). These characteristics influence the plant's resource requirements, seasonality, and life history. A key aspect of diversity is measured by the number of different functional groups represented by the plants on the plot. This is called, not surprisingly, functional diversity. Species diversity refers to the number of different species within each functional group, or to the total number of species present. This latter measure is sometimes called diversity per se. Clearly, there is a correlation between diversity per se, the total number
of different species present, and functional diversity. One cannot add more species without eventually adding more functional groups as well. Another related determinant of productivity is the composition of the functional groups present: Productivity may depend not only on the number of such groups represented but also on their identities because some groups may be more important than others in contributing to productivity or resilience.

The average amount of biomass grown per year on a plot of a given size increased with the diversity of functional groups represented, as shown in Fig. 1. The increase leveled off after a certain point. Tilman and co-workers also found more nutrient uptake and better soil quality on plots with a more diverse collection of plant species. Furthermore, the plots that were more diverse in this sense were also more robust in the face of climatic fluctuations. It appears from this work that both functional diversity and species diversity are important in maintaining productivity and resilience. Having functionally similar plants that respond differently to environmental fluctuations contributes to resilience. It ensures that whatever the environmental conditions, there will be plants of a given functional type that thrive under those conditions. The functional composition of the community will therefore not be changed by environmental fluctuations. In contrast, if the members of different functional groups respond differently to environmental fluctuations, then these fluctuations will alter the functional composition of the community and therefore its ecological characteristics, and it will be less resilient in the face of such fluctuations (Chapin et al., 1997). Chapin et al. (p. 303) argue that

no two species are ecologically redundant, even if they appear similar in their ecosystem effects under one particular set of environmental conditions.

What is the mechanism behind these results? There is still dispute about this. One possibility is that each plant type has a range of climatic conditions to which it is best suited. Climate varies from year to year in terms of temperature, rainfall, and in many other ways, and if a plot contains only one plant type then in many years it will not have any plants well adapted to the climates of those years. If, however, a plot carries many types, then in most years there will be some that are well adapted to the climate of that year, and on average its productivity will be greater. This is another illustration of the old proverb, "Don't put all your eggs in one basket." Analytically, it illustrates the same point as the economic argument for holding a diversified portfolio of stocks. The more diversified the portfolio, the less vulnerable one is to conditions that are bad for particular stocks or stocks in a particular sector of the economy. A robust portfolio should have stocks that do well in times of growth, stocks that do well in times of high interest rates, and stocks that do relatively well in times of recession. It should have different types of stocks. Diverse plots will also be more resilient to climatic variation, as a part of the same phenomenon. This is related to the insurance role of biological diversity, here manifested as higher average productivity. An alternative explanation for the greater productivity of biological diversity, here manifested as higher average productivity. An alternative explanation for the greater productivity of biological diversity, here manifested as higher average productivity. An alternative explanation for the greater productivity of biological diversity, here manifested as higher average productivity.
below is another layer of even more shade-tolerant
trees or shrubs. With such an arrangement of diverse
species, bright light falls on those that most need it,
and the light that passes through their canopy and is
of lesser intensity then falls on plants well suited to
it. Total photosynthesis—that is, conversion of carbon
dioxide in the air to carbohydrates with the aid of light
energy—will be greater under such a regime than in a

These arguments show that diversity is important in
ensuring the productivity and robustness of natural
plant communities, and therefore of the ecosystems
that are based on them. Diversity also helps natural
ecosystems to make the best adjustments to conditions
that vary over time or over space. Without the appro-
riate level of diversity, natural ecosystems cannot ad-
just to natural variations in the environment.

Through its role as the raw material in plant breed-
ing, biodiversity also contributes substantially to the
productivity of agricultural systems. New and higher
yielding plant and animal varieties are generated from
the natural variation in plants and animals. The large
increases in grain yields of the “green revolution” of
the 1960s and 1970s, which were responsible for keep-
ing food output increasing in parallel with populations
in developing countries, were largely achieved by use
of genetic diversity in the plant populations. Estimates
suggest that as much as $1 billion has been added to
the value of the U.S. agricultural output each year for
the past half century as a result of plant breeders’ use
of genetic diversity. Specifically, in the past half century
there has been a doubling in yields of rice, barley, soy-
beans, wheat, cotton, and sugarcane; a threefold in-
crease in tomato yields; and a quadrupling in yields of
maize, sorghum, and potato (U.S. Congress Office of
Technology Assessment, 1987). All this has been based
on and derived from genetic variability in the underly-
ing plant populations. In economic terms, this variabil-
ity is an asset—and one that has yielded a high return
at little cost. Diversity also contributes to productivity
in agriculture through the practice of crop rotation.
Changing from one crop to another from one year to
the next can enhance soil productivity by increasing
the nutrients in the soil. For example, rotating a crop
such as wheat with a legume that fixes nitrogen can
prevent long-term nitrogen loss and reduce the need
for nitrogen fertilizers. This type of crop rotation was
characteristic of traditional agricultural practices in
medieval Europe.

It is important to note that although biodiversity
contributes to the productivity of both natural ecosys-
tems and agricultural systems, it does so through differ-
ent mechanisms. Natural systems benefit directly from
a diverse mix of species: Agricultural systems benefit
from the existence of a pool of genetic variability on
which breeders can draw. Agricultural systems are usu-
ally monocultures, consisting of a single species grown
intensively over a large area. Its growth is supported
by applications of water, fertilizers, pesticides, and weed
killers. Farmers manage cropland so as to ensure that
crop growth is not limited by lack of water or nutrients,
and that the main food crop does not have to compete
with other species or with pests during its growth.
Farmers create and maintain an artificial environment
and then plant a crop that is optimally adjusted to this
environment—an approach that is radically different
from the natural growth process.

II. BIODIVERSITY AND INSURANCE

A dramatic illustration of the insurance role of biodiver-
sity comes from the recent history of rice production.
The prosperity and comfort of literally billions of people
depends on the rice harvest. In the 1970s, a new virus,
the grassy stunt virus, carried by the brown plant hop-
per, threatened the Asian rice crop. This appeared to
be capable of destroying a large fraction of the crop
and in some years destroyed as much as one-fourth of
the crop. Developing a form of rice resistant to this
virus became of critical importance. Rice breeders suc-
ceded in this task with the help of the International
Rice Research Institute (IRRI) in the Philippines. The
IRRI conducts research on rice production and holds
a large seed bank of seeds of different varieties of rice
and the near relatives of rice. In this case, the IRRI
located a variety of wild rice that was not used commer-
cially but which was resistant to the grassy stunt virus.
The gene conveying resistance was transferred to com-
mercial rice varieties, yielding commercial rice resistant
to the threatening virus. This would not have been
possible without genes from a variety of rice that was
apparently of no commercial value. Without this vari-
ety, the world’s rice crop, one of its most important
food crops, would have been seriously damaged. Interest-
ingly, the variety of wild rice that was resistant to
the virus was found in only one location—a valley that
was flooded by a hydroelectric dam shortly after the
IRRI found and took into its collection the critical rice
variety. The same situation was repeated later in the
1970s, and similar situations have occurred with other
food crops, particularly corn in the United States (My-
ers, 1997). There is every reason to expect that events
such as these will recur regularly: Planting large areas
with genetically identical plants greatly increases the chances that once a disease starts it will spread with dangerous speed through the entire area and crop. A report by the Committee for Agriculture, Science and Technology (1999, p. 13) emphasized this point:

Because of the increasingly high densities and large areas over which they are now grown, both livestock and crop plants are continually acquiring new diseases and pests, and existing diseases and pests are continually evolving new strains that overcome the defenses of particular breeds or strains. This is exacerbated by the accidental transport of diseases around the world. These diseases and pests destabilize agricultural systems. For instance, areas of western Minnesota and eastern North and South Dakota no longer can produce viable wheat and barley crops because of new strains of scab and vomit toxin for which no crop varieties have sufficient genetic resistance. Indeed, catastrophic attacks of disease, invasions of insects, and climatic extremes have caused wholesale crop destruction and ensuing famines whenever crops had insufficient diversity to provide at least some plants with the ability to withstand the assaults. Outbreaks of avian flu in the Chesapeake Bay area regularly result in rigorous quarantines of poultry houses due to the extremely high density of poultry farms in this region.

Disease problems, as old as agriculture, are recorded in myth and in written history, and still exist. Red rust on wheat in Roman times, mass poisoning from ergot-tainted rye during the middle ages, the Irish potato famine of the 19th century, and the Southern corn leaf blight in 1970 all were due to insufficient biodiversity in the affected crops. The severity of the 1998 Hong Kong chicken epidemic was likely exacerbated by the lack of diversity in disease resistance as well as by the high chicken densities in the production facilities.

The continual accrual of new diseases can be countered only if breeders can find sufficient genetic diversity within a crop or its relatives. Even the full complement of natural genetic variation, though, may not be sufficient to stop some diseases. Consider, for instance, the impacts of chestnut blight, an introduced disease that devastated what was once the dominant tree of the eastern United States, but which now occurs only as rare stump sprouts. Despite the vast geographic expanse and genetic diversity of the native North American chestnut, there is no known genetic resistance to its pathogen. Similarly, in vast areas of west and central Africa, livestock genetic resistance to the debilitating effects of trypanosomiasis is found only in a few unproductive local breeds. Despite massive efforts, the genetic mechanisms governing this resistance are not yet well understood.

A lethal disease of corn, or wheat or rice, were it to appear, would devastate agriculture and human society. The only insurance that society has against such a catastrophe is biodiversity. Genetic diversity within a crop plant or animal species and its relatives might allow resistant strains to be discovered and used. Similarly, a diversity of potential food plants might allow another species to become an effective substitute for a major crop species that was lost to disease.

These cases illustrate clearly the insurance role of biodiversity. It is an important defense against disaster in the form of new diseases. The pathogens that cause disease are evolving continually, in an attempt to outwit our defenses against them. A clear example of this phenomenon is the evolution of antibiotic resistance among bacteria. The bacteria that cause several once common diseases in humans are now showing resistance to their principal controls, to the great concern of public health authorities. The same is happening with the pathogens that cause disease in crops and in commercial animals. Without reserves of genetic variability we may not be able to develop varieties of our agricultural crops and animals that can resist these new disease varieties. Indeed, it is precisely genetic variability in the pathogens that allows them to develop resistance. Genetic variability means that some of the disease-causing pathogens are naturally relatively unaffected by our defenses against them, which may be in the form of weed killers, insecticides, or vaccinations for livestock. These more resistant specimens are the ones that survive and from which new subsequent generations are bred. Therefore, pathogens use against us the mechanisms that we will use against them if we preserve and use genetic diversity. Without this diversity, we have disarmed unilaterally in the war against our most threatening enemies.

There is another important role for the insurance provided by biodiversity—to provide variability that could be critical in responding to the environmental changes wrought by humans. Human activity is changing the climate and the seas level, and it is making many more local changes in the environment. A hotter climate
may require different crop varieties. An increase in sea level may lead to increased salinity in groundwater, and therefore to a need for crop varieties that are salt tolerant. A good example of the value of diversity in the context of a changed environment is the evolution of plants that have grown on mine wastes in the United Kingdom. These wastes are rich in heavy metals and are poisonous to most plant species (Antonovics et al., 1971). The existence of a pool of genetic diversity allows us to find plants that could tolerate these poisons and even help to remove them from the soil. We are making changes to the global environment on an unprecedented scale, and biodiversity might be critical in allowing us to respond to the consequences of these changes. Population growth and environmental change mean that we now need the insurance provided by biodiversity more than ever before.

III. BIODIVERSITY AND GENETIC KNOWLEDGE

The third reason I previously gave for the importance of biodiversity is that it is a source of knowledge. We can learn from natural organisms how to make chemicals that have important and valuable properties. A good example is provided by the polymerase chain reaction (PCR). This reaction is central to culturing DNA specimens for analysis—as in forensic tests used in trials such as the O. J. Simpson trial, and in many processes central to the biotechnology industry. Culturing requires an enzyme that is resistant to high temperatures. Enzymes with the right degree of temperature resistance were found in hot springs in Yellowstone National Park, and the heat resistance of these was then used to create an enzyme that could be used to culture DNA specimens. This enzyme is now central to the rapidly growing biotechnology industry. There are many more less complex examples. In fact, 37% by value of the pharmaceuticals sold in the United States are or were originally derived from plants or other living organisms (Carte, 1996). Aspirin is derived from the bark of willow trees. The bark of yew trees has been used to derive a drug that is effective against ovarian cancer (Stierle et al., 1993). A derivative of the rosy periwinkle flower is being used to cure childhood leukemia. The key point is that certain plants and animals are known to produce substances that are highly active pharmacologically. Plants that live in insect-infested areas produce substances that are poisonous to insects, and these have been used as the basis for insecticides. Some snakes produce venom that paralyzes parts of the nervous system, and others produce venom that reduces blood pressure. Other insects produce anticoagulants. All of these have been adapted for medical use. There is little that is new in these observations: They form the basis for many traditional medicines, which rely heavily on plants. Shakespeare refers to this in Romeo and Juliet (II: iii):

Of mickle is the powerful grace that lies
In herbs, plants, stones and their true qualities:
For nought so vile that on earth doth live
But to the earth some special good doth give,
Within the infant rind of this weak flower
Poison hath its residence and medicine power.

IV. BIODIVERSITY AND ECOSYSTEM SERVICES

I previously mentioned the role of genetic diversity in providing raw material for selective breeding, the traditional way of developing new crop or animal varieties that are more productive, more disease resistant, hardier, or more desirable in some other way. I have also mentioned its role in ensuring the productivity of ecosystems, and in ensuring their robustness against diseases and pathogens. There are other more complex ways in which biodiversity is essential to the proper functioning of ecosystems and to the delivery of the ecosystem services on which human beings are so dependent.

There are cases in which the full diversity of organisms in an ecosystem is required for that system to function and to provide services to human societies, and the removal or addition of even a single type of organism can have extraordinarily far-reaching consequences. "Keystone species" provide a convincing illustration. Ecologists use the term keystone species to describe a species whose removal will cause an entire ecosystem to change substantially. A widely cited example is that of sea otters on the California coast. The removal of sea otters as a result of hunting them for their pelts led to far-reaching and undesirable changes in the California coastal ecosystems. Sea otters eat sea urchins, which in turn graze on kelp plants. Without control of the urchin population by otters, the urchins will destroy the kelp beds, completely changing the marine coastal environment. Removal of otters led to a greatly impoverished coastal environment, which was restored in part to its original state with a ban on otter
Biodiversity as a commodity

hunting. Another example of the role and impact of a keystone species is provided by the removal of kangaroo rats from an area of the Chihuahuan desert, which led to a threefold increase in the yields of grasses and to far-reaching changes in the desert ecosystem. In this case, the rats had played a key role by eating seeds and disturbing the soil, and their removal consequently changed the plant balance (Power et al., 1996).

Not only can the removal of a species lead to major changes in an ecosystem but also the introduction of a new species (a so-called exotic species) can lead to a profound transformation of the system. A dramatic example is the introduction of the rinderpest virus into East Africa in 1890. This initially attacked domestic and wild cattle and then spread. By 1892, 95% of the wildebeest in the Serengeti region had died, in addition to most of the domestic cattle. Wildebeest are one of the main grazers and also the main food sources for carnivorous predators (lions, leopards, and hyenas) in the Serengeti; therefore, their virtual elimination led to profound changes in the system. In the 1930s, the introduction of a vaccination against rinderpest reestablished the original system (Aber and Melillo, 1991). The point of these examples is that we cannot easily tell a priori which species are essential and which are not. There is often a risk that an apparently small change in a set of species will have effects far beyond those initially anticipated. The degree of interdependence between different species is great; therefore, human beings may depend on many more species than we would expect from a first analysis of the situation. Abelard (an eleventh-century French theologian) suggested that any organism has a role to play and a reason for existing: "Whatever is generated is generated by some necessary cause, for nothing comes into being except there be some due cause for it" (as quoted in National Research Council, 1999). Also, John Donne (as quoted in Hemingway, 1940), an English metaphysical poet of the seventeenth century, wrote that

No man is an island, entire of itself; every man is a piece of the continent, a part of the maine; if a clod be washed away by the sea, Europe is the less, as well as if a promontory were, as well as if a manor of thy friends or thine own were; any man's death diminishes me, because I am involved in mankind, and therefore never send to know for whom the bell tolls: it tolls for thee.

There is an ecological equivalent to this: No species is an island, entire of itself, not even Homo sapiens. Any species' extinction may diminish us because we depend on many species. To repeat: The loss of even apparently unimportant species can have immensely costly consequences because of the complex patterns of interdependence between species. In the end, the loss of an apparently small and unimportant group of species could threaten the provision of ecosystem services that are essential to humanity. The distinguished biologist E. O. Wilson once said of microbes that "We need them but they don't need us." This is why many scientists see a serious risk in the current rate of species extinction. They cannot be precise about the dangers involved but nonetheless believe that there is a real risk of costly consequences. To give this point some substance, I mention a possible relationship between the extinction of passenger pigeons and the introduction of Lyme disease into American society. When Europeans first arrived in the United States, the passenger pigeon was probably the most abundant bird in the country. Its population was estimated in the billions. It traveled around in flocks of hundreds of thousands, flocks so large that their passing darkened the sky for many minutes at a time. By 1914 they were extinct, annihilated by a combination of hunting and destruction of the habitat that they needed for survival. It seemed unbelievable that an animal so abundant could be reduced to extinction so fast. A possible connection between this extinction and the emergence of Lyme disease events has recently been proposed. A letter to Science in 1998 made the following suggestion (Blockstein, 1998, p. 1831):2

There is another possible twist to the complicated ecological chain of events presented by Clive G. Jones et al. (Reports, 13 Feb., p. 1023) whereby the incidence of Lyme disease might increase following population increases of mice allowed by a big mast year of acorns. A major competitor of deer and mice for these bumper crops has been absent from the eastern deciduous forests for a century. The extinct passenger pigeon (Ectopistes migratorius) was a nomadic wanderer that specialized on a diet of the superabundant, but unpredictable, crops of mast. With a population estimated at 2 to 5 billion, concentrated in enormous flocks, passenger pigeons congregated wherever there were huge crops of mast. The birds were so efficient at detuning the woods of nuts that many observers noted that native wildlife and feral hogs could not find sufficient food after a pigeon flock had passed through. Is it possible that, in the

2 I am grateful to Paul Ehrlich for this reference.
presence of passenger pigeons, the population explosions of mice in mast years, reported by Jones et al., would have been less likely. Could the outbreaks of Lyme disease in the late 20th century have been a delayed consequence of the extinction of the passenger pigeon?

The point here is that passenger pigeons ate acorns and beechnuts, both of which were abundant in the forests of the northeastern United States. The demise of the pigeons led to an increase in the food available for other animals that ate these, including mice. Mice are the main breeding ground and hosts of the parasites that cause Lyme disease, and it is reasonable that the explosion of food for mice led to an increase in their population and thus in the population of Lyme disease vectors. (Abundant acorn crops always lead to increases in the population of mice.) The disease vectors transfer from mice to deer, which browse in the same forests and on the same foods and then move across territory likely to be used by humans, grazing on grass on lawns and fields. Therefore, the extinction of passenger pigeons could have been instrumental in causing the spread of Lyme disease to humans. This illustrates well the extraordinary complexity of the web of life and of the connections between different species and between species and human welfare. No one could reasonably have anticipated this connection between passenger pigeons and Lyme disease. No analysis of the consequences of the loss of this bird could have anticipated such an outcome. Indeed, the bird was so abundant that it must have been difficult to anticipate that human activity could drive it extinct.

The message of these examples is that it is difficult to foresee the consequences of a change in the biodiversity of an ecosystem. Even an apparently small change can lead to dramatic alterations in the system's ability to function and to provide the services on which human beings are dependent. There is another aspect of this phenomenon. A particular role in an ecosystem may be played at different times or in different circumstances by quite different plants or animals. The type of tree that stabilizes soil on a north-facing slope at a certain latitude may not grow on a south-facing slope at that latitude so that a different species is needed there to maintain the physical stability of the system. As a consequence, the set of species required for a certain type of ecosystem to function may vary greatly from region to region. In fact, we know of no single subset of species that on their own would serve to operate all ecosystems and provide all ecosystem services in all regions of the planet. Therefore, diversity in a given location may increase productivity and ecosystem functions in that location, whereas diversity at the regional or global level is actually necessary for the operation of important ecosystems in all geographic regions. Although individual species may possibly be redundant in some locations, it is possible that on the global scale few if any are really redundant. A clear statement on this topic is given by Chapin et al. (1997, p. 505):

The abundance of species with similar ecological effects should give stability (resistance and resilience) to ecosystems in the face of increasingly rapid human-induced environmental change. Loss of a keystone species or of all species in a major functional group will, by definition, have large ecosystem effects. Efforts to identify and protect such species and groups often yield demonstrable near-term benefits. Of increasing concern is the loss of species that have similar ecosystem effects but differ in their environmental responses. Loss of such species may reduce ecosystem resilience and the capacity to adjust to ever-increasing rates of environmental change. This latter role of diversity is not adequately represented in current international conventions, but it may be one of the most important mechanisms by which we sustain the long-term functioning of ecosystems and the services they provide to society.

V. BIODIVERSITY AND MARKETS

Returning to the economic questions, to what extent can we hope to commercialize these contributions of biodiversity? Obviously, they are economically important, but can the market capture them? Can the economic contributions of biodiversity be used to generate incomes to the owners of biodiversity that will provide them with incentives to conserve it? Can they make conserving tropical forests more attractive than clearing them? This is a critical question: Forest owners do not conserve forest because they are important to humanity. Rather, if they conserve them, if they do so because they can profit from so doing. Therefore, we need to know whether the important services provided by biodiversity can be the basis for profits from forest conservation or for conservation of sources of biodiversity more generally. In this context, I again examine the categories under which I classified the contributions
of biodiversity:

- Increasing productivity
- Providing insurance
- Providing knowledge
- Maintaining ecosystem services

One of the most fundamental insights into the operation of a competitive market economy is that under certain conditions it will align individual and social interests and provide incentives that lead to an efficient outcome. In the famous words of Adam Smith in his *An Enquiry into the Nature and Causes of the Wealth of Nations* (1777),

> Every individual ... neither intends to promote the public interest, nor knows how much he is promoting it. He intends only his own security, his own gain. And he is in this led by an invisible hand to promote an end which was no part of his intention. By pursuing his own interest he frequently promotes that of society more effectively than when he really intends to promote it.

This is a beautiful metaphor: Market forces are an invisible hand, steering us to act in the interests of society as a whole when in fact we only seek to promote our own interests. Recently, this insight was formalized and made more precise via some important propositions from economic theory, including the First Theorem of Welfare Economics. These propositions state that if all goods are private goods (i.e., there are no public goods) and the private and social costs of all activities are equal, then a competitive market economy is Pareto efficient (i.e., operates so as to leave unexploited no possibilities for mutual gain). This is a remarkable result and provides the basis for economists’ belief in the efficacy of market systems and the desirability of market-based approaches to economic organization. However, in the context of biodiversity the restriction that all goods be private is critical.

What is the public–private good distinction? A public good is one that has two properties: My consumption does not interfere with yours, and the provider cannot prevent nonpayers from benefiting from the good. Such goods are said to be nonrival and nonexcludable, in contrast with private goods, which are both rival and excludable. A seat at the opera is a private good: If I sit in it you cannot, and the management can certainly exclude nonpayers from seeing the performance. Law and order, in contrast, is a public good: It benefits everyone in the region in which it is enforced and the benefits cannot be restricted to those who have contributed to its costs.

Markets are not good at providing public goods. Their nonexcludability makes it difficult for the provider to earn a good return on the costs of providing them. Knowledge is an interesting and relevant example: It is naturally a public good because it can be passed costlessly from one person to another and enjoyed by all of them, even though none of them, or perhaps only the first, paid for it. Hence the existence of intellectual property rights, instituted as a means of ensuring some return on the generation of knowledge. Because of the difficulty in appropriating the returns on their provision, markets tend to underprovide public goods relative to an economist’s concept of what is efficient for society. Consequently, they have traditionally been provided by the public sector. Some of the services provided by biodiversity are public goods, although biodiversity does not fit the traditional mold for public goods completely. The extent of biodiversity is not something that can be determined by the public sector because it is the result of literally billions of land-use choices throughout the world. It is also strongly influenced by issues such as climate change, which are again driven by billions of heating and transportation choices. In fact, biodiversity has been called a privately produced public good. For the remainder of this argument, the key point is that the First Theorem fails in the context of public goods, and indeed it has long been recognized that markets will underprovide public goods relative to the level that would be required for efficiency.

One further basic economic point is needed—a familiarity with the diamond–water paradox and the limitations of market prices as indicators of “importance” to society. Discussion of this paradox helps us to be clear that the price of a good does not reflect its importance in any overall social or philosophical sense. Very unimportant goods can be valued more highly by the market than (have higher prices than) very important goods. The classic illustration of this is the diamonds and water paradox, which perplexed economists through the eighteenth and nineteenth centuries until its resolution by Alfred Marshall. The point here is that water is clearly more important to human society than diamonds, but diamonds trade in the market at prices far in excess of those fetched by water. Why? Marshall’s answer was simple and is now part of common knowledge: Price is set by supply and demand. The market price is the price at which the amount supplied is also the amount demanded. In the case of water, the supply (at least in Marshall’s time) was so large as to exceed the amount that could possibly be demanded at any
price. Consequently, the price was zero: Water was free. Now, of course, the demand for water has increased greatly as a result of population growth and increasing prosperity, whereas the supply has remained approximately constant so that water is no longer free. For diamonds, being naturally scarce, the desire for ownership always exceeded that which could be accommodated naturally. The market price was high as a result of competition between rich people for the few diamonds available.

What are the implications for biodiversity? Simply that even if it is of great importance to society, and is not a public good, it will not necessarily be possible to convert this importance into value in the market place. The balance of supply and demand will be critical here, as shall be shown in the context of the commercialization of the genetic knowledge inherent in biodiversity.

A. Productivity

To the extent that diversity increases productivity in agricultural systems, we would expect that farmers would be willing to pay for it. Arable farmers achieve some of the benefits of diversity by crop rotation, i.e., rotating between a series of different crops in successive years. The different crop types make different demands on the soil and contribute different nutrients to it. However, the range of crops used for this purpose is quite limited and does not contribute in any substantial way to the conservation of biodiversity. The other mechanism through which diversity contributes to productivity in agriculture is via its contribution to the breeding of new plant species that are better adapted to emerging conditions or more resistant to new diseases. At this point, the productivity and knowledge roles of biodiversity merge, and I comment on the possibility of commercializing the knowledge role later. These comments on biodiversity's contribution to knowledge will also apply to the contribution that biodiversity makes to human societies via the breeding of new varieties.

There appears to be some appreciation of the benefits of diversity in tropical agriculture, in which there is a tendency to grow several crops together or to grow crops in a way that conserves the original forest. Traditionally, coffee was grown as an understory plant beneath high tropical forest trees: This benefits the coffee, which is a shade-tolerant plant, takes full advantage of the light available in the region, and avoids the need to destroy the main forest trees to make land available for coffee growing. This practice also allowed other commercial crops to be grown with the coffee, such as citrus fruits and avocados, thus allowing farmers to diversify their risks. In cases such as this there is a contribution to the preservation of diversity because of the conservation of the forest. Studies have shown that forests converted for production of shade-grown coffee retain a very large proportion of their original biodiversity and that growing in this way is less expensive per pound produced than plantation growing. This cost difference reflects in part the greater productivity of diverse ecosystems and the more effective cycling of nutrients in these plant communities (the total yield of coffee per hectare, however, is less; Perfecto et al., 1996). In this case, it seems that there can be some conditions under which the productivity enhancements of diverse systems can be realized commercially, with attendant benefits for biodiversity conservation.

B. Insurance

Insurance is clearly something for which there is a demand. Most of us insure our homes and our cars and have health insurance. Therefore, perhaps the insurance role of biodiversity is one for which people will pay. The difficulty here is that until recently this insurance has been provided as a public good—indeed, as a global public good. Consider in the light of the public-private good distinction the conservation of rice varieties by the IRRI mentioned previously and in particular the use of one of these to provide a defense against the grassy stunt virus. In cases such as this, the insurance was provided as a public good. It was available to all for the cost of buying the new variety of rice incorporating resistance to the grassy stunt virus. You did not have to pay an insurance premium to benefit from the insurance. The developers of the new variety could not exclude from using it those who did not contribute to its development by paying insurance premiums.

At this point, I again digress briefly into economics. Risk, such as the risk of the destruction of a part of a crop, is managed in a market system by contingent contracts. A contingent contract is a contract that pays a specified amount if and only if a specified contingency occurs. Buying the contract reduces the risk to you that is associated with the specified event. If it occurs and has negative consequences, then you are to some degree compensated by the payment made to you under the contract. An insurance contract is a classic example of a contingent contract: You buy it by paying a premium and it entitles you to a payment if and only if a specified event—the insured peril—occurs. It is crucial for the efficient management of risks that the payment through which one purchases a contingent contract should not be contingent. You pay the insurance premium whether you use the insurance
or not: The only uncertainty or contingency is whether you can make a claim under the policy.

It follows that the problem of the public good nature of the insurance provided by the existence of new varieties could not be overcome by charging premium prices for them. This price would only be paid after the product had been developed and needed. If the new variety were never needed, then this price premium would never be paid. No contingent contract is being sold in this framework.

There is a possibility that the public good nature of the insurance will change. Public goods can become private goods through either institutional or technical change. A good example is television broadcasting. Until about 10 years ago, this was a public good par excellence: A broadcaster could not exclude from viewing a program anyone in the reception area so that the non-excludability property of public goods held, and of course there is no rivalry in consumption. My viewing a TV broadcast in no way interferes with your viewing. The development of scramblers changed this. A TV broadcast can now be scrambled so that it can only be viewed by those who have purchased a descrambler. Broadcasters can now exclude those who have not paid from viewing their programs, which are therefore no longer public goods. A public good has been privatized.

Developments in the area of intellectual property rights for agricultural biotechnology could change the situation in a similar way for the insurance value of biodiversity. These may lead to the privatization of a hitherto public good. Crop developers are increasingly patenting genes developed to enhance the properties of crops, including such properties as their taste, productivity, and insect resistance. Most agricultural biotechnology companies have a large and growing portfolio of patents on genes and on genetically modified plant varieties. They are also aggressively defending these patents, to the extent of developing and introducing "terminator genes" that will ensure that the properties conveyed by patented genes will not be transferred to offspring of the plant. Consequently, users cannot breed from seed that they have bought but must purchase more from the supplier. It is possible that plant breeders with such a tight hold on their intellectual property will be able to extract from users a great enough return to justify substantial investment in biodiversity conservation. The aggressive enforcement of intellectual property rights might act here like scramblers did with TV broadcasts, effectively privatizing a previously public good. Were it to happen, this would increase substantially the incentives for biodiversity conservation, but possibly at the cost of restricting access to the latest agricultural technology to those with the ability to pay a premium.

There is already a good illustration of the possible side effects of privatization of genetic knowledge. The bacteria Bacillus thuringiensis (Bt) produces a toxin that kills many crop pests and is itself biodegradable. For this reason, organic farmers use it as a pesticide: Because it is biodegradable it leaves no dangerous residues on the crops on which it is used. Monsanto and Novartis recently incorporated genes from Bt into transgenic crops. The presence of these genes in the plants means that the plants benefit from the defense provided by the Bt bacteria, which is the production of proteins that are toxic to the main pests of cotton and corn. Because the Bt genetic defense against pesticides is now widely used, it is possible that pests will develop resistance to it. Resistance develops faster the more widely a defense is used. Therefore, within a few years there may be generations of crop pests that are immune to this way of defending crops, i.e., immune to the Bt toxin. Monsanto and Novartis will seek to develop variants on the Bt genes, and with their scientific and financial resources they may succeed. If so, they will sell a new generation of transgenic plants with defenses against the new generation of pests. However, these will be their proprietary products, covered by their patents. Because of the development of resistance to the Bt toxin, organic farmers will no longer be able to use Bt as a harmless pesticide. They will be forced to buy proprietary defenses against pests. This is a clear illustration of the two factors mentioned previously: development of intellectual property rights in genetic knowledge leading to both stronger incentives to conserve and to develop further and also leading to a restriction of access. In the case of transgenic crops including the Bt defenses, Monsanto recently announced that all growers of Bt corn will be required to grow plots of nonengineered corn that are at least 20% of the size of the engineered crops. The aim here is to provide sufficient nonengineered corn that the development of resistant pests will be delayed or possibly even prevented. ("Monsanto Concession," 1999).

C. Knowledge

Here, I discuss the role of biodiversity as a source of knowledge. There are two different areas of application here: the development of medical products by the pharmaceutical industry and the development of new or better crops by the agricultural biotechnology industry.

In applications of biodiversity to the development of pharmaceuticals there has already been some progress
toward commercialization. Recognition of the likelihood that tropical plants contain chemicals that could be forerunners of pharmaceuticals has led most major drug companies to pursue bioprospecting as a way of finding new pharmacologically active substances to serve as a basis for drug development. Typically, they have sought these compounds in the tropics, in areas where there is extensive inter species competition, or in other extreme areas. They have been willing to pay quite substantial sums for access to these regions, and they have made deals with host countries that involve giving them a royalty on the products that might eventually be based on their prospecting. Such royalties could be large relative to the incomes of the countries concerned. Merck, Inc., one of the largest pharmaceutical companies in the United States, has an agreement with a Costa Rican agency called InBio (Institute Nacional de La Biodiversidad) for bioprospecting rights in Costa Rica. The terms of the agreement are that Merck paid InBio a fixed sum, $1.35 million, to be used for forest conservation in exchange for the right to receive samples collected by InBio and to use these as the basis for new product development. Should any of them prove commercially successful, Merck will pay InBio a royalty on the revenues generated. Similar agreements are in place between other U.S. pharmaceutical companies and regions of Central and South America.

The discovery of the PCR enzyme and the agreement between Merck and InBio and several other drug discoveries based on plants from developing countries led to a wave of optimism, perhaps excessive, about the potential commercial value of in situ biodiversity in developing countries. What in fact is the commercial potential here? There is no question that pharmaceutical- and agricultural products of great human and commercial value have been and will in the future be developed from the biodiversity in tropical countries. The key question is how much of this will be returned as a reward for the conservation of the originating biodiversity. In answering this question, one must take note of several points. The first point is related to the discussion of the diamond–water paradox: Not everything that is important will have value in the marketplace because there is nothing associated with being important that rules out the possibility that the supply may exceed the demand. Some commentators have seen this as a critical issue in the market for biodiversity-based genetic information. There are literally millions of organisms in the world that might provide genetic information, and if we do not know which ones will provide valuable information and which will not, then the supply of potential genetic leads is huge, possibly greater than pharmaceutical companies can process. In such a situation, the market price for such leads would be near zero. However, recent calculations suggest that in a small number of the world's biodiversity hot spots, bioprospecting rights may be worth as much as $9000 per hectare. This is small in relation to the amount that might ultimately be derived from drug sales but still large relative to other uses of the land. In fact, it is approximately a century's worth of ranching income (Rausser and Small, 1998). The key insight in these calculations is that prior knowledge of the nature of the ecosystems in a location can improve estimates of the probability of finding commercially interesting compounds there and can suggest where search would be profitable and where success is unlikely. In this case, not all organisms or all leads are equivalent: Some have much greater chances of success than others. Prior scientific information can change the odds of success from 1 in 10,000 to approximately an order of magnitude better. In practical terms, this means that developing countries can clarify the commercial attractions of their biodiversity by performing research on the ecosystems of which it is a part. This is similar to a country with potential oil reserves engaging in basic geological prospecting before seeking to negotiate leases for oil development. The results may be positive or negative, but either way they will give the country a better view of its prospects. In those cases in which the research is positive, the impact on the value of prospecting rights could be large.

Another important point is that an immense amount of human skill and expertise are needed to develop a plant specimen into a commercial drug. Typically, there will be a minimum of 10 or more years of work by hundreds of skilled people working with millions of dollars of sophisticated equipment. Unfortunately, very few plant extracts actually produce drug leads (i.e., contain pharmacologically active compounds with no obvious ill effects)—probably less than 1 in 10,000. Of these few, very few become commercial drugs—less than 1 in 100. On average, perhaps 1 in 250,000 samples collected leads to a commercial drug. ("When Rhetoric Fits Reality," 1998).

Therefore, the chance of any individual bioprospecting operation leading to a commercially valuable drug is very small indeed. Also, even if it does produce a drug, tens or even hundreds of millions of dollars will have to be invested, with significant chances of failure. Developments in biotechnology are currently altering this picture. They are reducing the time needed for testing and development and giving greater insights into the kinds of chemicals likely to be successful. By
Biodiversity as a Commodity

reducing the costs of drug development based on bioprospecting, they are making bioprospecting more attractive. For example, the cost of screening 10,000 samples for pharmaceutical potential 10 years ago would have been $6 million. Today it is $150,000 (Reid et al., 1993). Simultaneously, advances in knowledge are also making more effective alternative methods of drug development, based on understanding of the cellular and genetic mechanisms of disease. In total, the picture that emerges is one of heavily guarded optimism. Bioprospecting does have economic value, and technological developments may be increasing that value. However, in the short term we cannot expect great sums of money to flow to the conservation of biological diversity because of bioprospecting possibilities.

There is another problem to be overcome in establishing an income from bioprospecting—a problem with intellectual property rights. The same plant may occur in several different regions, and the same or similar chemicals may occur in different plants. Therefore, the same or similar drugs may be derived by different routes from different plants or different geographic regions. Research toward a commercial product has to be well under way before it is patentable; therefore, there is always a risk of being blocked by a prior patent.

There is also a risk, regarding the conservation of biodiversity, that biodiversity is valuable but leads to no direct commercial application. In a recent article on the value of marine bioprospecting, Carle (1996, p. 284) stated,

Although many of these products are not likely to become therapeutics, the information gained from studying them is likely to lead to the development and understanding of novel molecular targets, which may in turn lead to the development of new therapeutic agents.

This is a classic statement of the importance of basic knowledge. Basic knowledge is a public good. Its importance is great but it is not patentable and not something that can be appropriated by a typical bioprospecting contract with an element of royalty payment or revenue sharing. The development of basic knowledge is typically publicly funded precisely because its economic benefits, although potentially immense, are difficult to appropriate. More constructive thinking about how to realize the undoubted importance of bioprospecting in terms of income for conservation is needed. A further illustration of this point is provided by the example of taxol mentioned earlier. Taxol is a promising antitumor agent in breast and ovarian cancers that can be extracted from a fungus that lives in the phloem (inner bark) of the Pacific yew tree. Taxol was first isolated from the tree itself, but the tree is relatively rare and slow growing, and it produces little taxol; therefore, a search for other sources was initiated (Stierle et al., 1993). Ultimately, little in the way of economic returns may flow to the regions in which taxol was discovered.

Biodiversity can also be applied to the development of new or better crops: To date, this has probably been the most important commercial application of biodiversity. As noted previously, the existence of a pool of genetic variation provides plant and animal breeders with the raw materials for developing new varieties and more productive or resilient varieties of existing varieties. The existing varieties of a commercially important crop are usually the property of a commercial firm or of a research facility and are protected by patents. For example, the University of California owns the patents of many varieties of strawberries. These patents cover varieties that are best suited to different soil types and different weather conditions, that are least prone to spoiling during transport and storage, and so on. For most other commercial crops, the varieties are owned by seed companies, whose main asset is often the intellectual property represented by their ownership of patents to widely used varieties. In this context, the market can certainly recognize the value of biodiversity, provided that it is of the type that seems likely to contribute to the development and refinement of commercially important crops. Unusual variants of commercial crops, such as early varieties of wheat, corn, or soybeans, would qualify, and possibly so too would their near genetic relatives. However, biodiversity more broadly would probably not derive a value through this process, even though genes from unrelated plants might enhance the commercial potential of existing crops.

D. Ecosystem Services

Perhaps the most promising approach is to consider selling the services of natural ecosystems and using the revenue to provide incentives for conserving the biodiversity that supports them. Selling services provided by natural ecosystems can potentially provide incentives for the conservation of these systems and thus indirectly of biodiversity. For example, watersheds provide economically valuable services for which there is a market, and indeed recognition of this has already led to the conservation of significant forest areas (Chichlinsky and Heal, 1998). Some commentators (Reid, 1998) have suggested that as much as 15% of the earth's land area serves as watersheds for large cities and so
could legitimately be conserved on the basis of the watershed services that it supplies. Many of these watersheds are areas of considerable biological uniqueness and their conservation would be a major advance for the conservation of biodiversity. Their functioning as watersheds probably depends substantially on the continuation of their current levels of biodiversity.

Ecotourism is based on the preservation of intact ecosystems and the more appealing elements of biodiversity to be found in some of these systems, and revenues from this are providing powerful incentives for the conservation of several important ecosystems (Chichilnisky and Heal, 1998; Heal, 2000a). Insofar as ecotourism conserves tropical ecosystems, it contributes to the conservation of biodiversity. In certain regions, particularly southern Africa, ecotourism is making a major contribution to the conservation of biodiversity. In Angola, Botswana, Kenya, Malawi, Mozambique, Namibia, South Africa, Tanzania, Zambia, and Zimbabwe, approximately 18% of the total land area is now devoted to the support of wildlife. This is both a significant amount in total and a massive increase relative to 15 or 20 years ago (Bond, 1993; Cumming, 1990a, and b; Cumming and Bond, 1991). This has led to an increase in the populations of several previously endangered species and to more robust breeding populations of many important birds and mammals. All of this has been driven by the fact that in many parts of this region land devoted to ecotourism and sport hunting can earn a higher return than land devoted to more conventional agriculture. Figure 2 illustrates these phenomena: It shows for Zimbabwe both the growth increase in land whose primary use is the support of wildlife and the favorable economic returns to wildlife conservation relative to agriculture in parts of the region. (Regions III–V are either arid regions or wetlands.) Although the growth of ecotourism as an economically important activity has increased most in southern Africa, this region is not unique in being able to generate conservation incentives by tourism. Central and South America, and parts of Asia, are now benefiting from ecotourism, which is providing significant returns to the conservation of certain endangered species in these regions (Heal, 2000b; Freese, 1999; Honey, 1999).

Finally, there is a real prospect of commercializing some of the carbon sequestration services of forests under the terms of the Kyoto Protocol, which provides for compensation for some carbon sequestration activities via its provisions for joint implementation and through its clean development mechanism. I have estimated elsewhere (Heal, 2000b) that this could generate

![Figure 2](image-url)
an income as high as $50–100 per hectare per year, which is high enough to radically change the incentives for forest conservation and hence biodiversity conservation.

There are also other economic mechanisms through which the goods and services provided by tropical forests and their associated diversity might be marketed, although it is not clear whether these mechanisms can be applied generally. One is the sale of “nontimber forest products,” which are commercially valuable forest products that are not produced by tree cutting. They are therefore not timber, and their harvest and sale are compatible with conservation of the forest. They include various tropical fruits, vines that can be used as ropes, rattan (which grows among the trees), resins such as latex, and plants used as medicine by local populations. This last use is important. Approximately 4 billion people have no access to Western-style medicine and in times of sickness depend on plant extracts; this is known as ayurvedic medicine. This market is important in human terms, and although the amounts of money involved are small by comparison with those in Western medicine, they could still be enough to provide a significant incentive. Recent estimates suggest that in some tropical forests the collection and sale of nontimber forest products could yield as much as $60–140 per hectare per year, although many other studies have suggested significantly lower numbers (Grimes et al., 1994; Houghton and Mendelsohn, 1996; Peters et al., 1989). These are significant amounts. Given that they are sustainable, i.e., can be earned year after year, they are probably sufficient to justify conservation of forests even in the face of pressure for logging. However, currently there are a limited number of studies on this form of forest use so that the generalizability of these numbers is not clear, and whether the implied levels of harvesting are truly sustainable is also unclear.

The different mechanisms for generating income from natural ecosystems and the biodiversity that powers them are not mutually exclusive: The same land area could earn income by all mechanisms. A forest could obtain returns from carbon sequestration, bio-prospecting, nontimber products, managing a watershed, and ecotourism. In fact, the region of the Mata Atlantica (Brazilian coastal rain forest) inland from Rio de Janeiro is in a position to do exactly this. It manages the watershed for Rio in much the way that the Catskills region does for New York. It also manages the stream flow of the river Rio Paraiba do Sul, which provides most of the electric power for Rio via hydropower. These two services make it truly a major utility for Rio, with great economic value. Additionally, it supports a wide range of endemic species, sequesters carbon, and acts as a magnet for tourists. Currently, the region obtains a financial return only on one of these activities—ecotourism. In a case such as this, it is clear that the economic incentives for conservation could be immense, if we were only to do effectively what we already know how to do.¹

VI. CONCLUSIONS

What can we conclude about biodiversity, its economic importance, and the prospects for generating income based on this? Clearly, its importance is great, in the sense that it has been and remains a key contributor to human well-being. We understand enough of this contribution for its magnitude to be clear, although there are probably many aspects that are still unknown to us. Economically, the following is a key question: Can this importance be reflected in a commensurate income yielded by the conservation and use of biodiversity and of goods and services based on it?

In essence, much of what biodiversity contributes is, or has traditionally been, a public good. Its contribution to the functioning of natural ecosystems is non-excludable and nonrival, as is its insurance value. Also, knowledge is the quintessential public good. Despite the overwhelming publicness of much that biodiversity offers, there is some prospect of commercializing a limited part. Some of the services provided by natural ecosystems can be privatized and sold, generating a return to the conservation of the biodiversity that supports them. In the case of knowledge, the prospect of commercialization rests on the ability to establish intellectual property rights that will effectively privatize some of the public good aspects of its insurance and knowledge functions. Probably the most fundamental of these functions, like basic scientific research and development, will never be privatized and will thrive only with financial support via other mechanisms.

In summary, biodiversity is important economically. There are some market-based approaches to obtaining returns from biodiversity that can provide incentives for its conservation. If fully implemented, they would

¹ All of these measures are consistent with the United Nations Convention on Biological Diversity, to which many nations (although not the United States) have subscribed. This states in article 11 that “Each Contracting Party shall, as far as possible and appropriate, adopt economically sound measures that assist in the conservation and sustainable use of biological diversity.”
have a profound and positive impact. However, they would not solve the problem of biodiversity conservation. They would need to be supplemented by nonmarket measures such as the Endangered Species Act (ESA), the Convention on International Trade in Endangered Species (CITES), the Marine Mammals Protection Act, the Fishery Conservation and Management Act, and others. I am not suggesting that these are good examples of nonmarket policies: Indeed, in several cases they are not. The point is that they illustrate the type of option available as an alternative to the market. Of course, except for CITES, the previously listed measures are all domestic U.S. measures. To be really effective, we would need global equivalents—a Global Endangered Species Act, for example. Given the political complexity of the domestic ESA, this prospect should illustrate clearly the desirability of market-based policies whenever possible.

It is important to note that even regulatory approaches such as the ESA can give rise to market-based incentives if coupled with provisions for measures such as mitigation banking. In the case of the ESA, several states have modified this in a way that has proven highly effective and has introduced economic incentives where none existed in the initial formulation of the act. A good illustration is an agreement reached between the International Paper Corporation and the U.S. Fish and Wildlife Service concerning the red cockaded woodpecker. This bird is endangered and nests in forests owned by International Paper. The U.S. Fish and Wildlife Service and International Paper reached the following arrangement: A target number of breeding woodpecker pairs was agreed on, and provided that this number is attained or exceeded, then International Paper will be regarded as complying with the ESA, whatever modifications it might make in the habitat under its control. Furthermore, the agreement also provided that any surplus of breeding pairs over this number could be "banked." This means that it could be used by the company to offset ESA requirements with respect to red cockaded woodpeckers elsewhere, or title to the surplus could be sold to other landowners and used by them to gain some measure of exemption (Jorling, 2000; Noss et al., 1997). The important point is that the costs of compliance with the ESA have been reduced by this agreement, without reducing its effectiveness. Indeed, there are additional benefits: Because the production of nesting pairs over a target level is saleable, International Paper now actually has an economic incentive to encourage the endangered species, something it never had with a strict interpretation of the ESA. Recent reports suggest that International Paper may be able to sell banked breeding pairs for as much as $100,000 per pair (W. Coleman, personal communication). Similar mitigation banking systems have been put in place for wetlands, which have to be conserved under the Clean Water Act and which also provide important habitat for endangered or threatened species.

In the cases of the red cockaded woodpecker and wetlands, regulation has produced a market. The market occurs as a result of efforts to meet the regulations at minimum costs and has the positive effect of providing stronger incentives than the original regulation for the conservation of the endangered species or habitat. In principle, we do not need to go through the process of regulation to set up a market. The most obvious move from an economic perspective is instead to create a market in situations in which this seems needed by using the state as a buyer. In other words, in the case of the red cockaded woodpecker, the government could just pay landowners on whose land the woodpeckers breed. This immediately establishes an incentive for conservation without the political and other costs of regulation. Systems of this type are currently being used on a trial basis in parts of the United Kingdom. In regions in which hill farming is economically marginal, and has been subsidized by the European Community's Common Agricultural Policy, the farming subsidies have been replaced by subsidies for preserving the ecological integrity of the countryside. Specifically, farmers are paid to grow and maintain hedgerows, ponds, wetlands, wildflowers, and coppices that provide habitat for birds and mammals. The Financial Times, the European equivalent of the Wall Street Journal, recently commented in an editorial that in many cases it would make more economic sense to pay farmers to provide "ecological sustenance" rather than to grow food that is clearly surplus to the region's needs and cannot compete on world markets ("Too Much Food," 1998). This solution is less attractive than establishing a market in which consumers buy directly, as in the cases of ecotourism or watersheds, because the money spent by the government has to be raised as tax revenue from individuals or corporations. This can produce a loss of economic efficiency elsewhere in the economy. However, if the services provided by biodiversity are public goods, then there may be little alternative to having the public sector act as a buyer on behalf of society as a whole.

* This system, known as Tyr Cwmen, is in operation in parts of South Wales.
BIO DIVERSITY AS A COMMODITY

See Also the Following Articles

DEFORESTATION AND LAND CLEARING • ECONOMIC GROWTH AND THE ENVIRONMENT • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • ECOSYSTEM SERVICES, CONCEPT OF • KEYSPEC SPECIES • LAND-USE ISSUES • MARKET ECONOMY AND BIODIVERSITY • PLANT SOURCES OF DRUGS AND CHEMICALS

Bibliography


Stierle et al. (1993). Taxol and taxane production by Taxomyces
I. What Is Biodiversity?
II. Definition of Biodiversity
III. Genetic Diversity
IV. Species Diversity
V. Ecosystem Diversity
VI. Biodiversity: Meaning and Measurement
VII. Biodiversity: Changes in Time and Space
VIII. Loss of Biodiversity and Causes
IX. Maintaining Biodiversity
X. Contextual Variations of the Definition
XI. Implications of Variations in the Definition

GLOSSARY

biodiversity/biological diversity Species, genetic, and ecosystem diversity in an area, sometimes including associated abiotic components such as landscape features, drainage systems, and climate.
diversity indices Measures that describe the different components of biodiversity, such as species richness (alpha diversity), beta and gamma diversity, endemicity, and higher taxon richness.
ecosystem diversity Diversity of habitats, ecosystems, and the accompanying ecological processes that maintain them.
endemicity State of a species or other taxon being restricted to a given area, such as a specific habitat, region, or continent.
flagship species Charismatic or well-known species that is associated with a given habitat or ecosystem and that may increase awareness of the need for conservation action.
genetic diversity Genetic variety found within or among species; this diversity allows the population or species to adapt and evolve in response to changing environments and natural selection pressures.
keystone species Species that has a disproportionately greater effect on the ecological processes of an ecosystem, and whose loss would result in significantly greater consequences for other species and biotic interactions.
organismal (species) diversity Number and relative abundance of all species living in a given area.
species richness Absolute number of species living in a given area (also called alpha diversity), giving equal weight to all resident species.
use values Values that are obtained by using a natural resource, such as timber, fuelwood, water, and landscapes. These include direct, indirect, option, and nonuse values.

THE WORD BIODIVERSITY IS A MODERN CONTRACTION OF THE TERM BIOLOGICAL DIVERSITY. Diversity refers to the range of variation or variety or differences among some set of attributes; biological diversity thus refers to variety within the living world or among and between living organisms.
I. WHAT IS BIODIVERSITY?

The term "biodiversity" was first used in its long version (biological diversity) by Lovelock (1980) and is most commonly used to describe the number of species. Recognizing that conventional methods of determining, and separating species, were inadequate, others elaborated the definition by including the variety and variability of living organisms.

These reduced and simple definitions, which embrace many different parameters, have been much elaborated and debated in the last three decades (see Section II); upon this definition hangs the outcome of important scientific considerations, international agreements, conventions, conservation initiatives, political debates, and socio-economic issues. Indeed, while the word "biodiversity" has become synonymous with life on earth, the term is commonly used in the fields of politics and environmental technology in addition to various scientific disciplines (Ghillarav, 1996). The U.S. Strategy Conference on Biological Diversity (1981) and the National Forum on Biodiversity (1986) in Washington, D.C., were the critical debates in crafting a definition, and it was the proceedings from the latter, edited by E. O. Wilson, that "launched the word 'biodiversity' into general use" (Harper and Hawksworth, 1994).

In measuring biodiversity, it is necessary to deconstruct some of the separate elements of which biodiversity is composed. It has become widespread practice to define biodiversity in terms of genes, species, and ecosystems, for example, "the abundance, variety, and genetic constitution of native animals and plants" (Doddson et al., 1998). Biodiversity also encompasses all five living kingdoms, including fungi. However, biodiversity does not have a universally agreed on definition and it is often re-defined on each occasion according to the context and purpose of the author.

II. DEFINITION OF BIODIVERSITY

"Biodiversity" is a relatively new compound word, but biological diversity (when referring to the number of species) is not. Over the last decade its definition has taken a more reductionist turn. Possibly the simplest definition for biodiversity, lacking in specificity or context, is merely the number of species. Yet many have argued that biodiversity does not equate to the number of species in an area. The term for this measure is species richness (Fiedler and Jain, 1992), which is only one component of biodiversity. Biodiversity is also more than species diversity (simply called diversity by some authors), which has been defined as the number of species in an area and their relative abundance (Pielou, 1977).

DeLong (1996) offered a more comprehensive definition:

Biodiversity is an attribute of an area and specifically refers to the variety within and among living organisms, assemblages of living organisms, biotic communities, and biotic processes, whether naturally occurring or modified by humans. Biodiversity can be measured in terms of genetic diversity and the identity and number of different types of species, assemblages of species, biotic communities, and biotic processes, and the amount (e.g., abundance, biomass, cover, rate) and structure of each. It can be observed and measured at any spatial scale ranging from microsites and habitat patches to the entire biosphere.

This definition allows for modification according to the context in which it is used. Various authors have proposed specific and detailed elaborations of this definition. Gaston and Spicer (1998) proposed a three-fold definition of "biodiversity"—ecological diversity, genetic diversity, and organismal diversity—while others conjoined the genetic and organismal components, leaving genetic diversity and ecological diversity as the principal components. These latter two elements can be linked to the two major "practical" value systems of direct use (genetics and indirect use/ecological described by Gaston and Spicer (1998). Other workers have emphasized a hierarchical approach or hierarchies of life systems.

In contrast, some argue that biodiversity, according to the definition of biological, does not include the diversity of abiotic components and processes, and that it is inaccurate to identify ecological processes, ecosystems, ecological complexes, and landscapes as components of biodiversity. The term ecological, as used in the sense of ecological system (ecosystem), encompasses both biotic and abiotic components and processes. Therefore, ecological diversity is a more appropriate term for definitions that include the diversity of ecological processes and ecosystems. However, ecological processes, it has been argued, should be included in the definition of biodiversity, the reasoning being that "although ecological processes are as much abiotic as biotic, they are crucial to maintaining biodiversity." Similarly, a U.S. Bureau of Land Management advisory group included ecological processes in their definition of biodiversity in response to criticism that the Office
of Technology Assessment's (1987) definition did not consider ecosystem form and function. Other writers point out that even though ecological processes are often cited as being crucial to maintaining biodiversity (Reid and Miller, 1989; Noss and Cooperrider, 1994; Samson and Snopik, 1994), this does not warrant the inclusion of ecological processes into the meaning of biodiversity. For example, Reid and Miller (1989) and Agarwal (1992) distinguished between biodiversity and the processes and ecological diversity that maintain it. Nevertheless, the jargon word "biodiversity" is, by its very origin, fundamentally indefinable, being a populist word invented for convenience. Its invention has had beneficial effects by fuelling research projects, mainly in ecology and systematics, and scientists have been drawn into contributing to the debate by the need to show that biodiversity is useful to humans and necessary for the proper functioning of ecosystems. Conservation (i.e., management) of biodiversity is axiomatic to these two concerns and lies behind the scientific need to define the term within whatever context is appropriate, since no general definition will be suitable when applied across a range of situations.

Biodiversity conservation requires the management of natural resources, and this in turn requires the measurement of these resources. Biodiversity measurement implies the need for some quantitative value that can be compared. Among the first scientists to measure diversity were Fisher, Corbet, and Williams (1943), who approximated the frequency distribution of the species rep-resented by 1, 2, 3, . . . (and so on) individuals by the logarithmic series \( \alpha \, \alpha^x/2 \, \alpha^x/3 \, \alpha^x/4 \, \ldots \) where the constant \( \alpha \) has been found to be a measure of species diversity. Species diversity is low when the number of species is growing slowly with respect to the increase in number of individuals, and it is high when the number of species is growing quickly.

If the need to quantify biodiversity drives the fundamental meaning of biodiversity, the definition may be limited to that which can be readily measured given current understanding and technologies. Such a definition of biodiversity could change over time as ideas, technology, and resources for measuring diversity change. DeLong (1996) suggested that an operational "clause" should be added to the definition of biodiversity, namely, that "biodiversity is . . . as measured in terms of . . . ." This approach provides a link to management while distinguishing between what biodiversity is (a state or attribute) and how it is measured. It also allows the operational clause to be adjusted over time without changing the fundamental meaning of the term. A definition of biodiversity should portray the full scope of what the term means, not just what can be measured and managed. In contrast, monitoring or management objectives must be attainable to be effective. Recognizing the distinction between a definition and management objectives should reduce the confusion between the meaning of biodiversity and the objectives for achieving biodiversity goals.

Biodiversity is a broad totality and often embraces elements beyond species diversity or numbers. For example, a major debate in biological sciences over many decades has been that of pattern versus process, especially in systematics and evolutionary studies. Molecular biology and systematics have enabled ecologists to see that inferred history is important in framing appropriate questions, and this understanding has precipitated a real integration of these twin hierarchies—pattern (e.g., diversity) and process (e.g., evolution). Fundamental divisions remain, such as "straight" parsimony (i.e., pattern) versus maximum likelihood (i.e., process) in the phylogenetic interpretation of sequence data.

It is apparent that the term biodiversity still lacks consistent meaning within the field of natural resource management. Michael Soulé found it shocking that "we are still trying to define biological diversity after all of the efforts of the Office of Technology Assessment and E. O. Wilson's book, Biodiversity" (Hudson, 1991). It is still defined in different ways by different people; some characterize biodiversity as being a widely used term "having no unified definition" and others emphasize or limit the meaning of biodiversity to that of native biodiversity. Some writers have included human alterations of biological communities in the scope of biodiversity (Bryant and Barber, 1994). Angermeier (1994) argued that "the absence of a 'native' criterion within the definition [of biodiversity] severely compromises biodiversity's utility as a meaningful biological concept," reasoning that native biodiversity is more valuable than artificial diversity and should be the primary focus of conservation efforts. The conservation of native biodiversity appears to be the theme of biodiversity conservation texts (Wilson and Peter, 1988; Hunter, 1996). Conversely, others argued that an important component of biodiversity is maintained by traditional farming techniques. In the context of conserving biodiversity, Reid and Miller (1989) and Bryant and Barber (1994) discussed the importance of genetic diversity within species of cultivated plants. Biodiversity within agricultural plants is important for pest management in agroecosystems and sustainable agriculture.

An accepted fundamental definition of biodiversity is needed for conservation planning, as are effective
communication and co-operation within and among different countries, governments, agencies, disciplines, organizations, and private landowners. Co-operation among these entities has been identified as being necessary for the conservation of biodiversity (Babbitt, 1994). Knopf (1992) asserted that the definitions of biodiversity are “as diverse as the biological resource.” Definitions of biodiversity range in scope from “the number of different species occurring in some location” to “all of the diversity and variability in nature” and “the variety of life and its processes.” A more comprehensive definition is “the variety of living organisms, the genetic differences among them, the communities and ecosystems in which they occur, and the ecological and evolutionary processes that keep them functioning, yet ever changing and adapting” (Noss and Cooperrider, 1994).

This plethora of terms and definitions is one of the major stumbling blocks to reaching agreement in problem solving and decision making. If entities in a planning process view biodiversity in fundamentally different ways, agreement on management objectives and strategies for biodiversity conservation will be impaired. (Swingland, 1990).

The differences between these conceptual perspectives on the meaning of biodiversity, and the associated semantic problems, are not trivial. Management intended to maintain one facet of biodiversity will not necessarily maintain another. For example, a timber extraction program that is designed to conserve biodiversity in the sense of site species richness may well reduce biodiversity measured as genetic variation within the tree species harvested. Clearly, the maintenance of different facets of biodiversity will require different management strategies and resources, and will meet different human needs.

Even if complete knowledge of particular areas could be assumed, and standard definitions of diversity are derived, the ranking of such areas in terms of their importance with respect to biological diversity remains problematic. Much depends on the scale that is being used. Thus, the question of what contribution a given area makes to global biodiversity is very different from the question of what contribution it makes to local, national, or regional biological diversity. This is because, even using a relatively simplified measure, any given area contributes to biological diversity in at least three different ways—through its richness in numbers of species, through the endemism (or geographical uniqueness) of these species (e.g., Mittermeier et al., 1992), and on the basis of degree of threat. The relative importance of these three factors will inevitably change at different geographical scales, and sites of high regional importance may have little significance at a global level. None of these factors includes any explicit assessment of genetic diversity.

Although the word biodiversity has already gained wide currency in the absence of a clear and unique meaning, greater precision will be required of its users if policy and programs are to be more effectively defined in the future.

III. GENETIC DIVERSITY

Genetic diversity is reliant on the heritable variation within and between populations of organisms. New genetic variation arises in individuals by gene and chromosome mutations, and in organisms with sexual reproduction it can be spread through the population by recombination. It has been estimated that in humans and fruit flies alike, the number of possible combinations of different forms of each gene sequence exceeds the number of atoms in the universe. Other kinds of genetic diversity can be identified at all levels of organization, including the amount of DNA per cell and chromosome structure and number. Selection acts on this pool of genetic variation present within an interbreeding population. Differential survival results in changes of the frequency of genes within this pool, and this is equivalent to population evolution. Genetic variation enables both natural evolutionary change and artificial selective breeding to occur (Thomas, 1992).

Only a small fraction (<1%) of the genetic material of higher organisms is outwardly expressed in the form and function of the organism; the purpose of the remaining DNA and the significance of any variation within it are unclear (Thomas, 1992). Each of the estimated $10^{10}$ different genes distributed across the world’s biota does not make an identical contribution to overall genetic diversity. In particular, those genes that control fundamental biochemical processes are strongly conserved across different taxa and generally show little variation, although such variation that does exist may exert a strong effect on the viability of the organism; the converse is true of other genes. A large amount of molecular variation in the mammalian immune system, for example, is possible on the basis of a small number of inherited genes (Thomas, 1992).

IV. SPECIES DIVERSITY

Historically, species are the fundamental descriptive units of the living world and this is why biodiversity is very commonly, and incorrectly, used as a synonym of species diversity, in particular of “species richness,”
Biodiversity, Definition of

which is the number of species in a site or habitat. Discussion of global biodiversity is typically presented in terms of global numbers of species in different taxonomic groups. An estimated 1.7 million species have been described to date; estimates for the total number of species existing on earth at present vary from 5 million to nearly 100 million. A conservative working estimate suggests there might be around 12.5 million.

When considering species numbers alone, life on earth appears to consist mostly of insects and microorganisms. The species level is generally regarded as the most natural one at which to consider whole-organism diversity. While species are also the primary focus of evolutionary mechanisms, and the origin and extinction of species are the principal agents in governing biological diversity, species cannot be recognized and enumerated by systematists with total precision. The concept of what a species is differs considerably among groups of organisms. It is for this reason, among others, that species diversity alone is not a satisfactory basis on which to define biodiversity.

Another reason why a straightforward count of the number of species provides only a partial indication of biological diversity concerns the concept of degree or extent of variation that is implicit within the term biodiversity. By definition, organisms that differ widely from each other in some respect contribute more to overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar. The greater the interspecific differences (e.g., by an isolated overall diversity than those that are very similar.

VI. Biodiversity: Meaning and Measurement

A. Species Diversity

A. S. Corbet, upon analyzing a large collection of butterflies from Malaya, remarked on the decrease in number of new species with an increasing number of individuals. He thought that the resulting distribution could be described by a hyperbola, but R. A. Fisher, to whom Corbet sent his results, suggested that a negative binomial distribution would be much more appropriate.
A comprehensive understanding of the underlying mechanisms that result in a given pattern of species abundance become more significant (Whittaker, 1972). Species diversity measurement was thus clearly formulated more than 50 years ago and a particular index was proposed. Fisher et al. attempted to find some general "rule" or "law" according to which the numerical abundances of different species were related to each other. In many communities, the number of species with given abundance could be approximated by the log-normal distribution. If species are classified in accordance with their abundance in logarithmically increasing classes—so-called "octaves" (i.e., the first octave contains 1–2 individuals, the second contains 2–4 individuals, the third has 4–8, the fourth has 8–16, and so on)—then the number of species per "octave" shows a truncated normal distribution. If a sample contains a high number of species and individuals, we can usually obtain a log-normal distribution, and it is obviously more tractable than the logarithmic series.

MacArthur (1957) went further by proposing an interesting model that assumed that boundaries between niches in resource–niche hypervolume are set at random, whereas the relative abundances of species are proportional to these sections of hypervolume. This model became widely known as the "broken-stick" or MacArthur's model. The distribution of abundance prescribed by MacArthur's model is much "flatter" (i.e., the contrast between given species and the next in the sequence is less) than in the case of a logarithmic series (Chiarov, 1996).

It has become clear that there is no universal type of distribution of relative abundance that corresponds to all real communities, though such distributions change in the course of succession according to a particular pattern. The dominance of a few of the most abundant species is more pronounced at the early stages of succession, while later the species of intermediate abundance become more significant (Whittaker, 1972).

A comprehensive understanding of the underlying mechanisms that result in a given pattern of species abundance still eludes scientists.

Another line of species diversity studies was connected with the use of special indices proposed to measure diversity without reference to some hypothetical distribution of relative abundance. A great variety of indices were proposed that assess the number of species and the proportions in abundance of different species. Among others, there was the very popular index that is based on Shannon's formula derived from information theory.

\[ H = \sum p_i \log p_i \]

where \( p_i \) is the proportion of the total number of individuals that belong to the \( i \)th species.

In a seminal work on the measurement of diversity, Whittaker (1972) introduced the concepts of alpha, beta, and gamma diversity. The measurements just described, giving diversity values for single sites, are examples of alpha diversity. The beta and gamma diversity concepts relate to changes in diversity between sites at local (beta) and geographical (gamma) scales. An essential part of these relational concepts is the idea of species turnover—the degree to which species replace other species at different sites. For use in assessing the relative value of multiple sites for the conservation of biodiversity, the idea of species turnover is translated into the principle of complementarity (see Section VIII.A), which can be implemented in combination with a taxonomic diversity index.

### B. Taxonomic Diversity

Biodiversity measurements that measure genetic difference directly, or indirectly through use of the taxonomic (cladistic) hierarchy (Williams et al., 1991), are currently being used. The indirect taxonomic approach is more practical because we already have a "rule of thumb" taxonomic hierarchy (which is being steadily improved through the application of cladistic analysis, notably to molecular data), whereas reliable estimates of overall genetic differences between taxa are virtually non-existent (abridged from Vane-Wright, 1990).

Based on the shared and unshared nodes between taxa (equivalent to position in the taxonomic hierarchy), a number of taxonomic diversity indices have now been developed. Of these, the most distinct are root weight, higher taxon richness, and taxonomic dispersion. The first places highest individual value on taxa that separate closest to the root of the cladogram and comprise only one or relatively few species; in effect this gives high weighting to relict groups (Vane-Wright, 1996). Higher taxon richness favors taxa according to their rank and number of included species. Dispersion, the most complex of the measures proposed so far (Williams et al., 1991), endeavors to select an even spread...
of taxa across the hierarchy, sampling a mixture of high, low, and intermediate ranking groups.

For a given group these measures, together with simple species richness if desired, can be used to compare the biotic diversity of any number of sites. The measures can also be expressed as percentages. Thus a site with viable populations of all species in a group would have a diversity score of 100%, whereas a site without any species of the group in question would score zero. In reality, of course, most sites have only a selection of species, and so receive various intermediate scores. Such assessments allow us to compare all sites with each other, and rank them individually from highest to lowest diversity (Vane-Wright, 1996). However, if we then take some conservation action (such as conserving a particular site), the same measures are unlikely to be directly comparable for making a second decision (such as choosing a second conservation site). This is because, in most real situations at least, there will be considerable overlap in the presence of species at particular sites.

C. Community Diversity

Early ecologists did not confine themselves to measuring species diversity. They also tried to understand the relationship of diversity with other features of the community (e.g., Williams, 1964; Whittaker, 1972). The dependence of species diversity on the structural complexity of the environment was demonstrated (MacArthur and MacArthur, 1961), as was the role of predation (Addicott, 1974) and periodical disturbance (Sousa, 1979) in determining a given level of diversity. The relationship between the species diversity and standing crop of a community was also shown (Ghilarov and Timonin, 1972).

Margalef (1957) was the first to use the Shannon index (though expressed in a different form). He proposed to evaluate the level of community organization in terms of information theory. Margalef stimulated many ecologists to quantitatively measure the species diversity of different communities and/or of the same community in different stages of its development. At that time, there was a widespread belief that with a single numerical value, an assessment could be made of some very significant feature of community structure. Many ecologists believed that in measuring species diversity at the community level they were using an approach that was fundamental to an understanding of diversity (Ghilarov, 1996).

Ecologists have measured diversity either by estimating species richness (number of species) in an area, or by using one or more indices combining species richness and relative abundance within an area. Some attempts have also been made to measure change in species richness (species turnover) between areas. These solutions to the problem of measuring biodiversity are limited because species richness takes no account of the differences between species in relation to their place in the natural hierarchy. Moreover, relative abundance is not a fixed property of a species, for it varies widely from time to time and place to place. In many environments most taxa are virtually or even completely unknown.

Conservation biologists, or applied ecologists, have called for a measurement of diversity that is more clearly related to overall genetic difference. An example concerns the problem of differential extinction. In World Conservation Strategy (IUCN/UNEP/WWF, 1980), it is noted that “the size of the potential genetic loss is related to the taxonomic hierarchy because... different positions in this hierarchy reflect greater or lesser degrees of genetic difference.” The current taxonomic hierarchy provides the only convenient rule of thumb for determining the relative size of a potential loss of genetic material.

D. Synthesis

A model incorporating island biogeographic theory, species abundance, and speciation, and that produces a fundamental biodiversity number (\( \theta \)) that is closely associated with species richness and abundance in an equilibrium meta-population, has been proposed in Hubbell’s unified theory (1997). This model assumes zerosum community dynamics or a saturated, totally stochastic local community, which limits its application, but it advances the study of species richness and relative abundance if others can extend its usefulness to the nonequilibrium systems that characterize the real world.

VII. BIODIVERSITY: CHANGES IN TIME AND SPACE

A. Changes Over Time

The fossil record is very incomplete, which emphasizes the marked variation between higher taxa and between species in different ecosystems in the extent to which individuals are susceptible to preservation and subsequent discovery. Chance discovery has played a large part in compiling the known fossil record, and interpre-
tation by paleontologists of the available material is beset by differences of opinion. Thus, the record is relatively good for shallow-water, hard-bodied marine invertebrates, but poor for most other groups, such as plants in moist tropical uplands.

Two relevant points appear to be well substantiated. First, taxonomic diversity, as measured by the number of recognized phyla of organisms, was greater in Cambrian times than in any later period. Second, it appears that species diversity and the number of families have undergone a net increase between the Cambrian and Pleistocene epochs, although interrupted by isolated phases of mass extinction (few of which are reflected in the fossil record of plants).

B. Changes in Space

Species diversity in natural habitats is high in warm areas and decreases with increasing latitude and altitude; additionally, terrestrial diversity is usually higher in areas of high rainfall and lower in drier areas. The richest areas are tropical moist forest and, if current estimates of the number of microfaunal species (mainly insects) of tropical moist forests are correct, then these areas, which cover perhaps 7% of the world’s surface area, may well contain over 90% of all species. If the diversity of larger organisms only is considered, then coral reefs such as Bunaken (see earlier) and, for plants at least, areas with a Mediterranean climate in South Africa and Western Australia may be as diverse. Gross genetic diversity and ecosystem diversity will tend to be positively correlated with species diversity.

What are not fully understood are the reasons for the large-scale geographic variation in species diversity, and in particular for the very high species diversity of tropical moist forests. The origin of diversity through the evolution of species and the maintenance of this diversity both need more study before they are better understood. This will require consideration of the present and historic (in a geological or evolutionary sense) conditions prevailing in particular areas, principally climatic but also edaphic and topographic. Climatically benign conditions (warmth, moisture, and relative aseasonality) over long periods of time appear to be particularly important.

Climax ecosystems will be more diverse than areas at earlier successional stages, but an area with a mosaic of systems at different successional stages will probably be more diverse than the same area at climax provided that each system occupies a sufficiently large area of its own. In many instances, human activities artificially maintain ecosystems at lower successional stages. In areas that have been under human influence for extended periods, notably in temperate regions, maintenance of existing levels of diversity may involve the maintenance of at least partially man-made landscapes and ecosystems, mixed with adequately sized areas of natural climax ecosystems.

VIII. LOSS OF BIODIVERSITY AND CAUSES

Species extinction is a natural process that occurs without the intervention of humans since, over geological time, all species have a finite span of existence. Extinctions caused directly or indirectly by humans are occurring at a rate that far exceeds any reasonable estimates of background extinction rates, and to the extent that these extinctions are correlated with habitat perturbation, they must be increasing.

Quantifying rates of species extinction is difficult and predicting future rates with precision is impossible. The documentation of definite species extinctions is only realistic under a relatively limited set of circumstances, for example, where a described species is readily visible and has a well-defined range that can be surveyed repeatedly. Unsurprisingly, most documented extinctions are of species that are easy to record and that inhabit sites that can be relatively easily inventoried. The large number of extinct species on oceanic islands is not solely an artifact of recording, because island species are generally more prone to extinction as a result of human actions.

Most global extinction rates are derived from extrapolation of measured and predicted rates of habitat loss, and estimates of species richness in different habitats. These two estimates are interpreted in the light of a principle derived from island biogeography, which states that the size of an area and of its species complement tend to have a predictable relationship. Fewer species are able to persist in a number of small habitat fragments than in the original unfragmented habitat, and this can result in the extinction of species (MacArthur and Wilson, 1967). These estimates involve large degrees of uncertainty, and predictions of current and future extinction rates should be interpreted with considerable caution. The pursuit of increased accuracy in the estimation of global extinction rates is not crucial. It is more important to recognize in general terms the extent to which populations and species that are not monitored are likely to be subject to fragmentation and extinction (Temple, 1986).
Loss of biodiversity in the form of domesticated animal breeds and plant varieties is of little significance in terms of overall global diversity, but genetic erosion in these populations is of particular human concern in so far as it has implications for food supply and the sustainability of locally adapted agricultural practices. For domesticated populations, the loss of wild relatives of crop or timber plants is of special concern for the same reason. These genetic resources may not only underlie the productivity of local agricultural systems but may also, when incorporated into breeding programs, provide the foundation of traits (disease resistance, nutritional value, hardiness, etc.) that are of global importance in intensive systems and that will assume even greater importance in the context of future climate change. Erosion of diversity in crop gene pools is difficult to demonstrate quantitatively, but can be indirectly assessed in terms of the increasing proportion of world cropland planted to high-yielding, but genetically uniform, varieties. Genetic modification of organisms, varieties, or cultivars for food production, pharmaceuticals, and other products, which has caused concern in some countries but not others, may also contribute to the loss of biodiversity.

Humans exterminate species either directly by hunting, collection, and persecution or indirectly through habitat destruction and modification. Overhunting is perhaps the most obvious direct cause of extinction in animals, but it is undoubtedly far less important than the indirect causes of habitat modification in terms of overall loss of biodiversity. Hunting selectively affects the targeted species, as well as plant and animal species whose populations are subsequently affected either negatively or positively, and so it has important implications for the management of natural resources. Genetic diversity in a hunted population is liable to decrease as a result of mass production, for the desired economics of scale demand high levels of uniformity.

Sustained human activity will affect the relative abundance of species and in extreme cases may lead to extinction. This may result from the habitat being made unsuitable for the species (e.g., clear-felling of forests or severe pollution of rivers) or through the habitat becoming fragmented (discussed earlier). Fragmentation divides previously contiguous populations of species into small sub-populations. If these are sufficiently small, then chance processes lead to higher probabilities of extinction within a relatively short time. Major changes in natural environments are likely to occur within the next century as a result of changes in global climate and weather patterns. These will cause greatly elevated extinction rates.

**IX. MAINTAINING BIODIVERSITY**

A. *In Situ Conservation*

The maintenance of biological diversity is the sustainable management of viable populations of species or populations in situ or ex situ. The maintenance of a significant proportion of the world’s biological diversity only appears feasible by maintaining organisms in their wild state and within their existing range. This allows for continuing adaptation of wild populations by natural evolutionary processes and, in principle, for current utilization practices to continue. For such maintenance to succeed, it almost invariably requires enhanced management through the integrated, community-based conservation of protected areas.

Over the last thirty years, conservation biologists have struggled with the concept of the maintenance of biodiversity in highly diverse environments like rain forests. Analytical techniques (neural-net models) that allow us to reconstruct past distributions of forest types present an opportunity to predict past contractions and expansions of forest forms, and the likelihood of refugia surviving climate change. Such extrapolations must be treated with caution, as pollen samples from Brazil (for example) disproved modeling predictions that savanna grasslands should have been extant, when in fact tropical and temperate forests were present. Various authors also opposed the Pleistocene refugia hypothesis (Haffer, 1989) for the Amazon region because some evidence demonstrated the lack of rain forest fragmentation during that era. In the biogeographical zones of the Australian wet tropics, there is a strong correlation between diversity patterns and reputed rain forest refugia in both species and genetic diversity. However, this appears to have been caused by differential extinction rates in differently sized refugia rather than by allopatric speciation in the Pleistocene. Others have emphasized that a greater concentration on the Pliocene or before would be useful, since most tropical species radiations occurred before the Pleistocene.

The local-determination hypothesis of species diversity (Rosenzweig, 1995), which predicts similar species diversity in similar habitats, has also been challenged. In sister taxa of plants, the net diversification was significantly higher in Asia than in North America.
for genera shared between the two continents. Greater insights into the effects of current ecology on the local diversity of an area may be assisted by considering the relative ages of clades, which could establish species proliferation rates between regions, thus advancing the local versus regional diversity debate (Ricklefs and Schluter, 1993). They also tested the taxon cycle theory (Wilson, 1961) using phylogenies of bird species and showed that older species’ lineages had more restricted ranges, smaller habitat breadth, and more fragmented distributions, and were closer to extinction than younger species.

In efforts to conserve biodiversity, preserving genetic dissimilarity is often a higher priority than maintaining genes of considerable similarity. Recent work shows that genetic divergence in mammals increases from the headwaters to the mouth as a river gets broader and thus becomes a greater barrier to populations on opposite banks; this effect promotes species diversity through allopatric speciation. Headwater species are basal in the phylogeny, and shared haplotypes occur only at the headwaters; this research is a contribution to Wallace’s riverine diversification hypothesis in the Amazon basin.

A central question in the design of effective conservation programs is what geographical regions to protect in order to maintain the most biological diversity. The term biodiversity hotspot was coined by Myers (Myers, 1990) and most commonly refers to regions of high species richness. GAP analysis is used to identify gaps in existing protected area networks (Scott et al., 1993); it uses algorithms to select the minimum set of grid cells that encompass the unprotected species. Rarity and endemism have been used to define hotspots in bird conservation (Balmford and Long, 1994), and species richness and endemism have been used to rank countries (McNeely et al., 1990). Hotspots are also defined as those areas with the greatest number of threatened species.

In setting conservation priorities, assumptions are made that indicator groups (e.g., macro-organisms such as birds, mammals, and plants) are good predictors of biological diversity in general. Another question that arises is how best to analyze biodiversity information to generate accurate and useful analyses that will inform conservation decisions. On a large scale, some concordance is found between bird diversity across continents with insect diversity (Pearson and Cassola, 1992), and in endemism patterns across taxa (Lawton, 1994); but at a finer spatial scale this correlation begins to break down. Richness in genera and families are good predictors of species richness at a finer level (Balmford et al., 1996a, 1996b). However, species richness is not a good measure with which to identify hotspots for conservation because it overlooks rare species, although as the sample area for hotspots is increased, more rare species are included as a simple function of arithmetic progression. Rarity and endemism are efficient indices for selecting the most parsimonious number of sites, but compared to complementarity measures they are less useful in defining conservation priorities.

A good conservation measure is complementarity, where the species complement of a reserve or area is identified and then further sites are found that add the greatest number of new species; this is akin to the portfolio approach (Swingland, 1997). Another method using integer linear programming to choose the optimal set of sites (maximal-covering-location; Church et al., 1996) is limited to small datasets and does not achieve the greatest conservation gain for the fewest additional sites. Clearly, combining an ecosystem portfolio approach with a richness or endemism assessment would be effective, but differing approaches are needed according to the conservation goal and data availability.

B. Ex Situ Conservation

Viable populations of many organisms can be maintained in cultivation or in captivity. Plants may also be maintained in seed banks and germplasm collections; similar techniques are under development for animals (storage of embryos, eggs, and sperm, i.e., “frozen zoos”) but are more problematic. Ex situ conservation is extremely costly in the case of most animals, and while it would in principle be possible to conserve a very large proportion of higher plants ex situ, this would be feasible for only a small percentage of the world’s organisms. Furthermore, it often involves a loss of genetic diversity through founder effects and the high probability of inbreeding (Milner-Gulland and Mace, 1998).

X. CONTEXTUAL VARIATIONS OF THE DEFINITION

A. Derivation of “Biodiversity”

The definition of biodiversity put forth by the Office of Technology Assessment (1987) appears to be the most widely cited basis for other published definitions
(Scott et al., 1995). However, the OTA did not explain why they defined the term as they did, nor did they cite any supportive documentation. One problem with relying solely on authoritative sources for definitions of biodiversity is that different authorities have defined the term in fundamentally different ways.

Bio is derived from the Greek word bios, meaning life. Biological and biotic are terms that refer to life, living organisms, assemblages of living organisms, and the activities and interactions of living organisms. The scope of the term biological can be further understood in the context of components and processes that are considered biological. Defining biodiversity (i.e., diversity) is more difficult because it continues to be defined in several fundamentally different ways. In definitions of biodiversity, diversity has been characterized as (1) the number of different types of items, (2) the number of different types of items and their relative abundance, and (3) variety. Characterization of diversity in discussions of biodiversity has also included the structural complexity of landscapes (Huston, 1994).

### B. Classifying Biodiversity

The classification of biodiversity can be divided into those authors who consider biodiversity to be a state and those who believe that it is a measure of the state. Most authors have defined biodiversity as a state or attribute, for example, ‘biodiversity is the variety of . . .’ or ‘variety and variability of . . .’ (Noss and Cooperrider, 1994). Standard dictionaries have classified diversity as a state, condition, or quality (Soukhanov et al., 1988). Other definitions of biodiversity limited the scope of the attribute to explicit, quantifiable dimensions or measures, for example, ‘biodiversity is the number of . . .’ or ‘the number and relative abundance of . . .’ (Office of Technology Assessment, 1987). This emphasis on quantitative, operational definitions of biodiversity and criticisms of non-quantitative definitions (Angermeier, 1994; Hunter, 1996) may signal a potential shift in the classification of the term from an attribute to a measure of an attribute. In the ecological and natural resource management literature, Pielou (1977) and others have treated diversity as a one- or two-dimensional attribute of a community (e.g., diversity is ‘the number of’ or ‘the number and relative abundance of’). More recently, it has been defined as a measure or index of those attributes; for example, diversity is a ‘measure of . . .’ (Noss and Cooperrider, 1994). Operational definitions of biodiversity (Angermeier, 1994) provide impetus to define biodiversity in quantitative terms as Hunter (1996) recommended.

### C. Attributes of Biodiversity

Another way of delineating the meaning of a term is to list its characteristics, properties, qualities, and parts. Noss (1990) recognized three main attributes of biodiversity: composition, structure, and function. Composition addresses the identity and richness of biotic components, and the relative amount (e.g., abundance, cover, biomass) of each (Noss, 1990). Biotic components of ecosystems include genes, organisms, family units, populations, age classes, species and other taxonomic categories, trophic levels of animals (e.g., herbivores, predators), animal guilds and assemblages, plant communities, and interacting assemblages of plants, animals, and microorganisms (i.e., biotic communities).

Structural attributes of biodiversity refer to the various vertical and horizontal components of a community or landscape (Noss, 1990) and the organizational levels of plant and animal populations and assemblages (Gaston and Spicer, 1998; Hunter, 1996). Considering only biotic, vegetative components of a landscape, horizontal structure consists of the size, shape, and spatial arrangement and juxtaposition of different plant communities; vertical structure consists of the foliage density and height of different vegetation layers (Noss, 1990). Structure can also refer to population, age and trophic structure, and other levels of community organization (Hunter, 1996).

The inclusion of structure in the meaning of biodiversity provides linkages with other concepts, such as habitat diversity and the plant community concept, for both of which vegetation structure is an important differentiating attribute. Structure may have been left out of most definitions of biodiversity because the concept of biodiversity evolved from the concept of ecological diversity, which primarily focused on species diversity (Fisher et al., 1943). Interestingly, 20 years ago it was asserted that measurements of diversity should not exclude structural diversity even though the term is most often used in reference to species diversity. Diversity can also be used in reference to niche width and the structural complexity of habitats.

Biotic functions represent the third component of biodiversity, and these include processes such as herbivory, predation, parasitism, mortality, production, vegetative succession, nutrient cycling and energy flow through biotic communities, colonization and extinction, genetic drift, and mutation (Noss, 1990). Biotic
processes can be addressed in terms of the identity and number of different types of processes, as well as the rate (e.g., predation rate) at which each process operates. Diversity of biotic components and processes can be observed at many biogeographic scales, from microsites and larger-scale landscape elements (e.g., vegetation types, habitat types, range sites) to regional landscapes, biomes, continents, hemispheres, and the entire biosphere (Noss, 1990; Huston, 1994; Hunter, 1996). Although these are scales at which biodiversity can be observed, they are not necessarily scales of biodiversity because most include abiotic (e.g., geological) features. Biodiversity can also be observed at several organism-based scales, including individual organisms, populations, species, and assemblages (e.g., guilds and plant communities), which themselves can be observed at various biogeographical scales.

D. Biological Resource Asset and Management Objectives

The contextual variations in the definition of biodiversity depend on what use is being made of the biological resource asset (or bioasset), and thus the asset management objective. Biological resource values consist of direct use, indirect use, and option and non-use values. For the purposes of assessing potential use, they can be further classified as follows:

- **Direct use values of major extractive products.** Princially, this would include forestry for timber and commercial fisheries in the case of terrestrial and marine systems. Extraction of these products often involves substantial investment in capital equipment by large non-local firms, and the products are transported and sold in well-developed markets far from their original source.

- **Direct use values of "minor" extractive products.** These are naturally or semi-naturally occurring products that require labor-intensive gathering or harvesting activities, often carried out by local people. Examples include rattan, fuelwood, seaweed, wild foods, artisanal fisheries, aquarium fish, and medicinal herbs. These may be collected for sale, barter, or home consumption.

- **Direct use values that require the extraction of only a small amount of biological material for ex situ research or storage.** This includes extraction of material for biological inventories, germplasm banks, and industrial research. Extraction is often accomplished during short or long expeditions that traverse large areas to collect representative samples of biological material.

- **Non-use values.** By their very nature these values occur off site. The primary feature of these values is that they support or protect the basic functioning of the protected area. Examples include nutrient cycling, stabilization of soils in erosion-prone areas, coastal zone stabilization, and biological support to local ecosystems. As a result of their nature, the value of these on-site functions is likely to be a component of all the other direct and non-use values generated by the area.

- **Indirect use values that accrue off site.** The value of these functions—such as watershed protection, natural ecosystems protected as national parks in generating income from wildlife tourism, protection of fisheries’ nurseries and subsistence fisheries, and climate regulation—may be very large or very small depending on their relative importance to the support or protection of off-site economic activity.

- **Option values.** Because option values may be associated with each and every use value, they are considered only where they may be of potential significance in conjunction with the particular type of product or service.

- **Nonuse values.** By their very nature these values occur at a distance from the resource and require no extraction or physical interaction with the resource, for example, stewardship, ethics, cultural belief, and aesthetics.

These foregoing values are only indirectly related to biological diversity. That is, a certain level of species richness is required for these functions but there is not necessarily a direct correlation between the value of the ecosystem and its diversity. Thus, mangrove ecosystems are generally of lower species richness than adjacent lowland terrestrial forests, but in resource terms they are likely to be of comparable value. The savannas of eastern and southern Africa, which are of great importance in generating revenues from tourism, are less diverse than the moister forests in these countries, which have far less potential for tourism.
E. Cave Canem or the Precautionary Principle

At present, humans actively exploit a relatively small proportion of the world’s biological diversity. Many other potential, yet undiscovered, optional and non-use values of biodiversity exist. These factors support a precautionary approach to maintaining biological diversity. In this case, the precautionary principle argues that actions should be taken to prevent further loss of biodiversity and potentially irreversible consequences before all biological uncertainties are resolved. Yet in conserving biodiversity, there must come a point at which the projected costs required to protect and maintain it will outweigh any probable benefits.

If species are to be viewed as a resource, and their maintenance is to be cost-effective, conservation should concentrate on systems and areas rich in species, and on those species known to be useful, or regarded as having a high probability of being useful. Thus biodiversity and its conservation would be defined purely along operational or cost–benefit lines. This bioasset perspective on biodiversity would therefore rest upon economic arguments more than biological ones.

Biodiversity has been identified as important for ecosystem health, medicinal values, agricultural purposes, and aesthetic and recreational values (Noss and Coo-perrider, 1994). Noss (1990) characterized an operational definition as one that is responsive to real-life management and regulatory questions, adding that such a definition is unlikely to be found for biodiversity. Angermeyer (1994) referred to an operational definition in a similar way, and Hunter (1996) suggested that a quantitative definition is needed for monitoring biodiversity and developing management plans. On the other hand, some writers assert that the confounding of definition and application is partly to blame for the confusion over how biodiversity concepts can be practically implemented.

XI. IMPLICATIONS OF VARIATIONS IN THE DEFINITION

The need for an unequivocal and precise meaning of biodiversity that is scientifically sensible and universally applicable is imperative to help guide the design of policy and programs for the future, as well as to make critical decisions in the present. Currently, such a definition does not exist. As a concept, biodiversity is both ubiquitous and useful, particular and confusing, and for this reason it is constantly redefined on nearly every occasion.

One of the many reasons for this state of affairs is that the definition of biodiversity affects objectives in national and regional research and conservation management, and in international funding priorities. One could easily promote a timber extraction or non-timber forest product program that conserves species richness (i.e., numbers of species) at the expense of genetic diversity. Indeed, a current research program to stimulate or increase the range of tropical tree species not currently in trade, as a way to take the pressure off over-exploited species, may be misguided. It may lead to increased genetic as well as species impoverishment when foresters expand the number of species they take and select only the best and most mature specimens, thus removing the most productive and healthiest genetic stock.

Apart from the principal definitions of biodiversity discussed earlier, such as the highest number of species (i.e., species richness) and the highest level of species endemicity (Myers, 1990) or taxal endemicity (called critical faunas analysis), interpretations of pure or applied definitions are becoming more common within the vocabulary of conservation and biodiversity utilization when determining biodiversity management priorities. Some examples are national biodiversity programs that maintain “biodiversity portfolios”; biodiversity defined as flagship or keystone species diversity; viability modeling (population viability analysis); interpretations of pure or applied definitions are becoming more common within the vocabulary of conservation and biodiversity utilization when determining biodiversity management priorities; projects that focus on the feasibility of integrating the targeted species, assemblages, or ecosystems with the needs of local human populations and sustainable use; and (lastly) political exigency (Swingland, 1997). Although the conservation policy of a country may be driven by more pressing needs—family planning, education, politics, internal conflict, financial planning and investment, individual vested interests—current policy and decisions are also being made on the foregoing biodiversity bases rather than along strict academic lines. Endemicity and species richness are useful starting points in defining priorities on the global level, but without information on the possibility of extinction using viability modeling or population analysis, the urgency of a given conservation action cannot be assessed. Moreover, with the increasing emphasis on the integration of local people into conservation programs to minimize long-term costs and to provide a more stable basis for the people and their natural environment, the poten-
See Also the Following Articles

BIO DIVERSITY, ORIGIN OF • ECOLOGY, CONCEPT AND TERMINOLOGY OF • ECO SYSTEM, CONCEPT OF • GENETIC DIVERSITY • HAB IT AT AND NICHE, CONCEPT OF • LOSS OF BIODIVERSITY, OVERVIEW • MEASUREMENT AND ANALYSIS OF BIODIVERSITY • TAXONOMY, METHODS OF

Bibliography


1. Introduction
II. The Fossil Record
III. Current Patterns of Biodiversity
IV. Community Stability
V. The Future Record
VI. Conclusion

GLOSSARY

biodiversity A measure of the relative abundance of species or higher taxonomic units found in a certain area at a particular time. In ancient communities this is often simply referred to as the number of species or taxa found at a site, usually referred to as "species richness" in the ecological literature.

community A group of interacting organisms, representing multiple species in a given location.

complex adaptive systems A group of individuals or types that exhibit variability in which large-scale patterns emerge from small-scale interactions. The system is adaptive in the sense that it is subject to extrinsic selection that leads to a change in structure and/or dynamics over time.

evolution The morphological or genetic change in species over time. Small changes that do not lead to reproductive isolation among members of a group are referred to as "microevolution." Speciation, or the generation of new species, is generally referred to as "macroevolution."

fossil Anything found in strata of rocks or sediments that is recognized as the remains of an organism from a former geological time.

species A group composed of individuals that interbreed or potentially interbreed and are separated reproductively from other such groups.

taxon (taxa) A group of individuals representing a classification of organisms, such as genus, family, or order. Higher taxa are those above the species level.

EVOlution has led to an incredible diversification of forms that have originated from a single ancestor. Our current accounting of species on Earth is woefully lacking, as many of the entries in this encyclopedia have demonstrated. In addition, the fossil record, unfortunately, is grossly underrepresented by taxa that have existed on Earth since life began 3.5–4.0 billion years ago. The purpose of this article is to summarize our current understanding of the rise of Earth's biodiversity and to discuss how evolution has driven this process of the accumulation of taxa over time and how evolution influences communities, both past and present, in which organisms reside. The influence of evolution on integrated communities of diverse taxa and the likely changes that will take place in the near future certainly rank among the great frontiers facing students of biodiversity as we watch the profound test of the resilience of Earth's ecosystems to the enterprise of Homo sapiens.
I. INTRODUCTION

The earliest record of life on Earth is from approximately 3.6 billion years ago. Aside from a few difficult to test hypotheses about how life arose, a topic beyond the scope of this article, the fossil record clearly reveals that early life forms were single-celled organisms of the group Archaea with structures similar to modern bacteria. Earth at that time was far different than today and would have been uninhabitable to the vast majority of contemporary organisms. Our understanding of these life forms that have arisen over this time to the present relies mainly on the fossilized remains from organisms representing an array of groups that far exceeds those living today. Although the fossil record contains discontinuities and absences of taxa, recent evidence is sufficient to allow discussions of the evolution of communities, or groups of species, over time. This level of detail is astounding, but it is only suggestive of the myriad of information likely to be gained in the coming decades. In addition, recent development and use of computer models is enabling us to both understand these paleobiological data and make predictions of how evolution influences the dynamics of communities over time. I suggest that the coupling of our rapidly increasing knowledge of the fossil record and our use of computer modeling will help us to both understand and predict the influence of humans on patterns of biodiversity.

II. THE FOSSIL RECORD

A. What Does the Fossil Record Indicate?

In general, the fossil record is an outstanding record of changes in biodiversity, revealing both an increase and turnover in taxa over time. Data for marine organisms suggests that the diversity of families has generally increased during the Phanerozoic, which includes the past 540 million years, the time for which fossilized remains of multicellular organisms exist (Fig. 1; Sepkoski, 1991). A broader survey shows that diversity has increased exponentially for all known families, genera, and species during this period (Fig. 2; Benton, 1995). Increases in taxa, however, have been discontinuous, with major catastrophic losses of taxa in the form of mass extinctions, such as that which occurred at the end of the Cretaceous period that led to the loss of the dinosaurs. Thus, the accumulation of biodiversity in the form of numbers of taxa has been discontinuous and has exhibited a skewed distribution in the magnitude of the size of extinction events, ranging from relatively low-level background extinction rates to large-scale mass extinction events.

In general, we know that species are generally short-lived, averaging a duration on the order of 5 million years, with few taxa being long-lived. Most species are,
and are inferred to have been, composed of relatively few individuals locally distributed with few species being relatively abundant and widely distributed. We also know that most genera have few species, whereas most families have few genera. These patterns have led and continue to lead to high probabilities that any particular taxon is likely to be relatively short-lived as opposed to long-lived, where the mean duration of taxa exceeds the median duration.

1. The Number of Taxa Has Increased over Time

A clear and dramatic trend exists in the fossil record of an increase in diversity. This pattern exists for taxa ranging from the level of species to phyla, and it has been well documented for plants, invertebrates, and vertebrates. This pattern is seen, for example, in the number of marine families over time during the Phanerozoic eon, since about 540 million years ago (Fig. 1). The major additions to diversity in the fossil record include skeletonized marine invertebrates during the Cambrian, especially trilobites, brachiopods, and archaeocyathans. Corals, cephalopods, ostracods, crinoids, and star fish arose through the remainder of the Paleozoic, and bivalves, gastropods, echinoids, teleost fish, and marine reptiles arose during the Mesozoic. Diversity increased on land and included the evolution of vascular plants (Silurian and Devonian), gymnosperms (Carboniferous), and angiosperms (Jurassic). Insect diversity exploded in the late Paleozoic, followed by social insects in the Jurassic. Major clades of terrestrial vertebrates arose by filling vacant niches, including amphibians (Devonian), reptiles (Carboniferous), pterosaurs and dinosaurs (Triassic and Jurassic), birds (Jurassic), and mammals in the early Tertiary.

The pattern, however, is one of fluctuations over time in the shorter term. The number of species of relatively large herbivores and carnivores in North America during the past 44 million years, for example, appears to exhibit little long-term trend toward increased or decreased diversity (Fig. 3). Interestingly, however, these data suggest that the number of predators depends on the number of potential prey species over time. There is a significant correlation between the number of predators and prey which is consistent with the hypothesis that interspecific associations are interdependent within communities over time.

2. Extinction Events Are Followed by Evolutionary Radiations

Extinction and the origin of species defines the dynamics of the fossil record. Species, for instance, typically last on the order of 1–10 million years and appear to go extinct in a non-random fashion. Therefore, species arise and ultimately disappear throughout time, producing a general pattern of loss referred to as “background” extinction (Fig. 4).

Biodiversity at various levels of organization has increased during the Phanerozoic. There exist, however,
at least five catastrophic extinction events that have led to a decrease in the number of taxa, and therefore biodiversity, that greatly exceed the background extinction rate (Fig. 4). In particular, the extinction event at the end of the Permian (Fig. 1, “P”) resulted in a loss of a maximum estimated 96% of species on Earth (although this is likely an overestimate; Raup, 1994).

The overall pattern of extinction appears to follow a well-defined “kill curve,” proposed by David Raup, that indicates a skewed distribution of species loss ranging over the gradient where most time periods exhibit low rates of extinction (background) and few periods of very high rates of extinction (mass extinctions). These data, therefore, suggest a predictable pattern, based on probability, that describes species extinctions.

Following extinctions there is general radiation, or rapid evolutionary diversification, of taxa. This is clearly seen in Sepkoski’s database of marine families (Fig. 1) and even more clearly in Benton’s larger database of families from both terrestrial and marine systems following the five major extinction events (Fig. 2).

The rates of species origination have varied over time and are highest following periods of mass extinction and in times when new ecosystem became available, such as during the initial rise of diversity during the late Vendian and Cambrian (Figs. 1 and 2). Speciation was highest for terrestrial families during the Priabonian (late Eocene), most likely due to the increase in insect diversity.

B. How Good Is the Fossil Record?

The fossil record is far from a complete representation of all of Earth’s biotic history. There are only bits and pieces of evidence from organisms and we are able to date strata from which organisms derive only with varying levels of accuracy. The fossil record is better, in general, for recent strata and decreases with age. The fossils we do have, however, enable us to construct a logical progression in both complexity of form and diversity over time. The idea that evolution is progressive, however, has long been abandoned in light of the many examples of lineages that do not follow straight progressions of simple to more complicated forms over time. Moreover, extant species that appear ancient, such as the coelacanth, horseshoe crab, and cockroach, are neither unchanged from their ancestors nor is their longevity unexpected. By chance we expect variance in the longevity among lineages, with many short-lived species and few long-lived species.

Understandingly, only certain morphological characterics have been susceptible to fossilization. Of those that are fossilized we often find only disarticulated parts of organisms, such as a tooth, part of a shell, or perhaps a bone fragment, separated over varying distances. Soft-bodied organisms and soft tissues are rarely observed. To complicate matters, these fragments are susceptible to movement both spatially (e.g., down-stream after death) and vertically in the stratigraphy, which makes fossils appear to move temporally. This occurs when sedimentation rates are very low, allowing noncontemporaneous organisms to become fossilized together, or when layers are mixed, a process referred to as “time averaging” of the fossil record. However, taphonomy, or the study of postmortem behavior of organic remains, indicates that a vast majority of organisms are not transported out of their original habitats and therefore can be used to construct probable past communities. Evidence for this derives primarily from the observation of contemporary local death assemblages in which individuals rarely move across community boundaries.

Microfossils such as pollen are abundant in lake sediment and can be used for quantitative reconstruction of local plant communities. Pollen from sediments on the order of a few hundred thousand years has been used to reasonably describe assemblages. Little is known about the role that evolution played in plant communities from sediments more than 1 million years old. The pollen record during the Holocene (e.g., since the last glacial period ended about 16,000 years ago) has been widely and more confidently used to reconstruct past plant communities. Therefore, we remain perplexed with regard to the various assemblages of taxa that likely resided concurrently in more ancient communities, but for more recent times we have increasing confidence in our effort to reconstruct communities from fossil remains.

C. What Does the Fossil Record Indicate About Communities?

The interpretation of community structure in the fossil record is more problematic than documenting the change in taxa over time. The primary difficulty lies in determining whether fossilized taxa occurred contemporaneously. Postmortem movement of organisms and their parts appears to take place over relatively short distances and within time-averaged ranges of species, which is a pattern found for both marine and terrestrial taxa. Therefore, optimistically, it appears that co-occurring taxa through the fossil record likely represent evidence of past communities.
Evidence from pollen records, however, can produce a paradox in our interpretation of past plant communities. Over long periods of time, pollen deposition in lake sediments generally accurately reflects nearby plant communities. The distance-dependent deposition of pollen is generally correlated with the sum basal area of trees (cross-sectional area of tree diameters) within a few square kilometers of the surrounding area, although many factors influence the pollen collected from lakes and the surrounding vegetation, such as relative decomposition rates and the relative efficiency of a lake to capture pollen. Pollen records found in lake sediments, however, suggest that plant community change, in response to rapid climate change following the Wisconsin glacial period in North America, is the product of independently acting species. In addition, there is a consistent pattern that, among late Quaternary pollen records for basins of different sizes and from different regions (temperate and tropical), communities appear to be the product of independently acting species.

The rain of leaf litter and other debris, likewise, is well correlated in contemporary communities with the abundance of trees within local communities. Unfortunately, leaf litter in general rapidly decomposes and rarely provides evidence of past communities. The fossil record remains difficult to interpret with regards to community composition and dynamics over time. Much research is needed and is under way on the dynamics of communities over long periods of time. In particular, we need to resolve conditions that lead to the following paradox: Taphonomy of current assemblages suggests fossil taxa may represent contemporaneous communities, whereas pollen studies suggest that at least plant species have acted relatively independently over time. Additional data are needed to help accurately reconstruct the composition of past communities.

III. CURRENT PATTERNS OF BIODIVERSITY

Biodiversity is heterogeneously distributed on Earth, with the general pattern that species densities are greatest near the equator and decrease with increasing latitude. Hypotheses and evidence for these patterns abound. In the following sections, I briefly introduce how we define biodiversity, discuss very general trends in communities, and apply these to understanding biodiversity of communities over time.

A. How Do We Define Biodiversity?

Before discussing patterns of biodiversity change over time, I need to define diversity. In general, paleontologists consider biodiversity to be equivalent to the number of unique taxa. In the ecological literature, diversity generally incorporates both the number of species and equitability or the relative abundance of individuals among species. A simple hypothetical example is the relative diversities of two communities. Assume that both have 100 individuals within two species. One community has 99 individuals of one species and the second community has 50 individuals of both species. The first community is essentially a monoculture of species 1. The second community is considered a more diverse community, despite the fact that both communities have the same number of species present.

Accounting for equitability in the fossil record, however, is quite problematic because of the spotty nature of the preserved specimens. It must be recognized that fossils in an assemblage will not likely constitute a random sample of individuals from a community. Rather, the remains of fossil communities are likely composed of a skewed distribution of individuals representing only a subset of those species that contained structures susceptible to fossilization. Other species likely escaped fossilization completely or were missed in the sample from which the fossils were found. In addition, the subset that is found is more than likely to be small and therefore highly susceptible to the variance incurred from the chance occurrence of individuals of rare species. This is an example of the problem referred to as the “central limit theorem” in statistics in which sampling of a skewed distribution (the reality of the fossil communities) yields a normally distributed population (not an accurate representation of the fossil communities). In other words, we are most likely to find fossils of the most abundant species, yielding an underestimate of the true number of species that occurred at that time in that community. Some fossil assemblages, however, will contain samples of relatively rare taxa, making these taxa appear more abundant that they really were. In addition, it is difficult to assess whether forms represent truly different taxa. We are more likely than not to lump taxa together because of our inability to know whether similar fossils represent members of an interbreeding population.

B. Patterns of Species Change Over Time

Simple change over time can be expressed as a rate, similar to the change in a population over time. We
can represent this change in the number of taxa over time as $dN/dt = RN$, where $N$ represents the number of species in a genus or genera in a family, and $R = S - E$, where $S$ is the speciation rate and $E$ is the extinction rate. We may therefore assume $dN/dt = 0$ to represent a null hypothesis against which we test for significant changes in the rates of species additions or subtractions. There are several problems with this approach, however, since our null hypothesis is that all species are expected to eventually become extinct. Therefore, we must define a time constraint of interest, concluding that if $R < 0$ then the taxa of interest appear to be on the decline. Comparisons of the fossil record have been found to differ from this random null hypothesis. The database assembled by Sepkoski (Fig. 1) suggests that diversity increases logistically, whereas the data set assembled by Benton (Fig. 2) suggests diversity has increased exponentially during the past 600 million years.

IV. COMMUNITY STABILITY

Researchers in the field of ecology, and particularly conservation biology, have long been interested in the dynamics of communities and the rate of species change over time, also referred to as "species turnover." George Evelyn Hutchinson and his student Robert MacArthur were early advocates of the notion that stability of communities is positively related to the number of species residing in the community. This relationship, quantified by MacArthur and incorporated into the influential theory of island biogeography developed in collaboration with E. O. Wilson, portrayed ecological systems as having stable equilibria, resulting from the trade-off between colonization and extinction rates on islands. Recently, this theory has been extended to systems that, in general, function like islands.

Although intuitively attractive, the relation between diversity and stability may not exist, at least in some systems. In 1973, Sir Robert May polarized the field of community ecology by introducing a simple matrix model approach that suggested an inverse relationship between species diversity and community stability. Recently, empirical studies have suggested that diversity and stability are related. Research by David Tilman on grasslands in Minnesota, for instance, suggests that community resistance to drought is correlated with community diversity.

A. Communities as Complex Adaptive Systems

Recent developments in our understanding of community assembly, organization, and development over ecological and evolutionary time suggest that communities are more complex than previously believed. Classic approaches to understanding the dynamics of communities portrayed them as being composed of individual species that interact ecologically with little regard to spatial location or temporal dynamics. Recently, we have come to appreciate the intricacies of small-scale spatial and temporal partitioning of resources and interactions among co-occurring organisms. In addition, in the past 30 years we have come to appreciate the immense genetic diversity within species that enables species to adapt to biotic and abiotic factors that operate spatially within communities (Thompson, 1994).

Complex adaptive systems can be defined as groups of individuals that exhibit variability, interact with a subset of individuals within the global population (usually within some local neighborhood), and are suscepti-

C. Are Communities Saturated with Species?

In general, it appears that many communities are at or near a saturation level for the number of species they contain. The evidence for this derives indirectly from several sources, including the relationship of families depicted previously in the fossil record and from the ecological literature in which there is found a general inability of introduced species to successfully invade communities. Species introductions usually fail, and when they succeed species usually have little consequence on natural systems, despite the numerous celebrated examples of introduced species altering native communities. Occasionally, even with the loss of native dominant species such as the American chestnut tree to a fungal blight, the effect on natural processes may be equivocal.

The result of this contemporary observation suggests that communities may reach taxa saturation. If this pattern holds for prior paleocommunities, then this suggests that ecological and evolutionary processes operate to fill available niche space with either colonizing species or through the process of speciation. Inspection of either Sepkoski's or Benton's databases for the number of taxa over time suggests that the Earth is not at a carrying capacity. The following is the obvious question stemming from this: Where should we expect to find new species occurring, especially in light of the estimation that humans are currently consuming on the order of 40% of the Earth's productivity?
B. The Effect of Disturbance on Community Biodiversity: Evolutionary versus Ecological Time

The causes of various mass extinction events are not known, but hypotheses do exist. The most recent mass extinction at the end of the Cretaceous was most likely caused by an asteroid impacting the Earth, forming the Chicxulub crater on the north side of the Yucatan Peninsula. The causes of the other mass extinctions are unresolved but may be the result of similar massive disturbances. The probability distribution of taxa loss appears to follow what we expect to occur by chance and matches the loss of taxa due to random factors. Occasionally, we should expect to lose many species, whereas over most time periods only a few species should be expected to be lost. Therefore, the greater the disturbance, the greater the effect (reduction) on biodiversity.

The probability of species loss depends on many factors, including the size of the populations of species within the community. Smaller populations are more likely to go extinct than larger populations, a process referred to as demographic stochasticity. The Gambler's Ruin is a metaphor that is commonly associated with this process. The idea is that a gambler in a casino who has a limited number of betting units to lose (e.g., quarters for a slot machine or equivalents to the minimum bet at a black jack table) is more likely to go bankrupt simply by chance compared to a gambler with many betting units. Likewise, stochastic, or random, changes in population size lead to small populations being more likely to reach the absorbing state of zero individuals than a large population. Therefore, and not surprisingly, small populations are at greatest risk of extinction simply by chance.

Communities composed of species with small populations are therefore likely to suffer higher levels of species loss than communities composed of species with large populations. Raup (1994) argues that there appears to be little predictability in which species will be eliminated and that extinction does not appear to be selective. Species with narrow geographic ranges may be more susceptible to extinction, whereas species with large geographic ranges are not likely to experience disturbances so great as to influence their entire population.

Small populations typically exhibit lower genetic variability than large populations and are more likely to become locally extinct. Small populations, however, may experience more rapid evolution due to reduced genetic inertia. This may be seen in the founder effect, in which a few individuals begin a new population leading to allopatric speciation.

Without knowing the probability distribution of extrinsic factors causing large-scale mass extinctions, it is difficult to extrapolate from the fossil record if species that went extinct in fact had small populations. In addition, it does not appear that evolution during the past several hundred years has reduced the likelihood of extinction events. There is no clear reduction in the apparent risk of such mass extinctions with time in the fossil record. Using the "all things being equal" assumption, the fossil record (Figs. 1 and 2) suggests that, with a general trend toward increased diversity over time, the condition of having more taxa should lead to increases in the number of taxa lost due to disturbances. That is, if the probability of extinction of any particular species is constant, then we should see increased extinction rates as diversity has increased. However, as shown in Fig. 4, the background extinction rate appears to have decreased with time. There remains the need to evaluate the risk of extinction by determining factors such as the absolute abundance of individuals within species and how this influences extinction risk. For example, if increased diversity over time has led to smaller population sizes due to ecological partitioning of resources, we may anticipate a trend toward increased extinction rates.

Experimental work on laboratory and field communities suggests that decreasing biodiversity also leads to a reduction in ecological time of ecosystem function, such as total productivity and carbon dioxide sequestration (Tilman and Downing, 1994; Naeem et al., 1994). This suggests that a positive feedback mechanism may operate so that disturbed communities are more likely
to lose species in addition to those initially lost due to decreased population size.

C. Models of the Influence of Evolution on Ecological Systems

The effect of evolution on the likelihood of species extinction and community stability represents a fundamental problem in the field of evolutionary ecology and is of great importance to our ability to conserve species over long time periods. Theoretical work suggests that the process of evolution influences population and community dynamics (Hartvigsen and Levin, 1997; Doebeli, 1996). Hartvigsen and Levin, using a spatially explicit, individual-based model that accounts for genetic variability at the individual level, leading to evolution over time, show that the process of evolution in a model system can slow population dynamics, suggesting that the risk of species extinction declines due to reduced fluctuations resulting from demographic stochasticity. If this occurs in real systems, then we might expect a positive feedback mechanism to operate such that a population reduced in size is likely to experience a loss in genetic diversity. In addition to other known factors that increase extinction risk of small populations, including inbreeding depression and social dysfunction (Allee effect), a loss in genetic diversity may increase population fluctuations, thus increasing the likelihood that such a species will go extinct.

V. THE FUTURE RECORD

A. Documenting Earth's Current Mass Extinction Event

The extent of the current extinction event will only be known in retrospect. However, there are patterns that are recognizable today. For example, there exists a good record of species that have gone extinct during the past 400 years (Smith et al., 1993). A minimum of 485 animal species and 584 plant species became extinct during this period, with approximately half of these occurring during the twentieth century. This includes the loss of nearly 1% of all known bird species during the past 100 years, mostly due to human activity. This loss suggests an average species duration of about 10,000 years, which is two or three orders of magnitude shorter than has occurred on average in the fossil record (in other words, this represents an extinction rate two or three orders of magnitude greater than the background extinction rate in the fossil record). It is this rate of known extinction, compared to average extinctions in the fossil record, that suggests humans are the impetus for a mass extinction event.

B. The Big Question: How Will the Process of Evolution Influence Biodiversity Dynamics?

Evolution generates new species in the aftermath of mass extinction. This is well supported by the fossil record and appears, in geological time, to be rapid. The current extinction event taking place on Earth, however, appears to be more rapid than has occurred previously. Speciation will undoubtedly lead to the rise of new taxa but over time scales that are likely to be too long to have a noticeable effect in our lifetime.

The other role that evolution plays in biodiversity is its ability to influence the stability of communities. Little is known about this role of evolution. Empirical data for plant communities suggest that stability is correlated with biodiversity, such that more diverse communities are more resilient to disturbance. Natural communities are often composed of genetically variable individuals, which provide the foundation that enables species to evolve in response to biotic and abiotic factors over time. Also, communities with more species are in general more diverse genetically. If the time frame of environmental change can be matched by selection operating on genetic diversity, communities may resist disturbances by adjusting to changes over time. The analogy of species wandering over an "adaptive landscape" (sensu Sewall Wright) produces an image of communities functioning as complex adaptive systems. Ecosystems that function as complex adaptive systems may resist environmental fluctuations. If, however, environmental disturbances are abrupt, communities may either not be able to react quickly enough or suffer overloads that essentially shake species out of the community. There is evidence that communities harbor redundancy at the species level, but we have not yet determined patterns that identify which species are necessarily important or unimportant to community function. Therefore, the loss of species from communities is likely to decrease, if only slowly, the stability of communities.

Theoretical work suggests that the process of evolution may increase the resilience of communities to disturbance. This area of inquiry is in its infancy and will likely become an important and rapidly developed discipline in the following decades.
VI. CONCLUSION

Biodiversity, measured as the number of taxa, has increased on Earth over time. This pattern appears to follow either an exponential or a logistic growth pattern of taxa during the past 540 million years. The increase in taxa, however, has been discontinuous due to extinction events. There appears to be a probability distribution of extinctions with small background extinctions occurring frequently and large, catastrophic mass extinctions occurring rarely. Biodiversity has repeatedly and rapidly rebounded following such extinction events.

Biodiversity currently appears to be near a saturation level, based on the observation that most species that are introduced to areas fail to successfully become established. However, the pattern of increase in the number of taxa (e.g., families) over time does not appear to be at or near an asymptote (Fig. 2). The extreme alteration of the earth's biosphere by humans is likely to lead to a mass extinction event, currently on the order of 10–100 times the background extinction rate observed in the fossil record. In areas in which habitats are disturbed rather than destroyed, an explosive radiation of new species may occur, manifested, however, only over geological time.

The process of evolution may help stabilize communities. However, current human-induced worldwide disturbances are likely to lead to reduced population sizes and a resulting loss in genetic diversity, ultimately increasing the likelihood that more species will go extinct.

See Also the Following Articles
ARCHAEA, ORIGIN OF • DIVERSITY, COMMUNITY/REGIONAL LEVEL • EXTINCTION, RATES OF • FOSSIL RECORD • LIMITS TO BIODIVERSITY (SPECIES PACKING) • MASS EXTINCTIONS, NOTABLE EXAMPLES OF • NORTH AMERICA, PATTERNS OF BIODIVERSITY IN

Bibliography
I. The Fossil Record and Contemporary Biodiversity

II. Reconstructed versus Real Evolutionary Trees

III. Key Innovations, Sister Taxon Analysis, and the Origins of Adaptive Radiations

IV. Evolution within Lineages: The Comparative Method

GLOSSARY

key innovation A property of a species that promotes the process of speciation.
molecular phylogenies Phylogenies constructed from comparisons of homologous genetic material.
radiation All the species descended from a single common ancestor; tends to be a speciose group in comparison with others of a similar age.
reconstructed evolutionary trees Phylogenies linking contemporary species, therefore containing no information about lineages that have gone extinct.
sister taxa The expression has two uses, according to context. First, from a species perspective, the sister taxon of a particular species is the one with which it had the most recent common ancestor. Second, from the perspective of a node in a bifurcating phylogenetic tree, each daughter lineage gives rise to a set of contemporary species. The two sets are sister taxa with reference to the node.
The generation of biodiversity involves, in large part, the origins of species. The fossil record, though notoriously incomplete, provides the most direct evidence for tracing temporal changes in biodiversity, including both the origin and extinction of lineages. Biases in preservation, often confounded with differences in taxonomic interest among paleontologists, mean that much more is known about the history of some taxa than others. For example, May (1996) points out that while around 95% of fossil databases consist of shallow-water marine invertebrates, most of today's known species are terrestrial, with 56% being insects. At the same time, Benton and Storrs' (1996) analyses suggest that more is known about the fossil history of vertebrates, most of which have been terrestrial, than of echinoderms, which are marine.

Benton's (1993) extensive compilations of stratigraphic incidence records for different lineages reveal the extent of well-known mass extinctions, followed by the subsequent regeneration of biodiversity. The rises in biodiversity that follow mass extinctions are presumably a consequence of the exploitation of (and speciation into) vacant niches. Also it is clear that the demise of some radiations is accompanied or followed by the rise of others; dinosaurs and mammals are cases in point. The causes of such complementarities have been hotly debated for many years. Similarly, there are times when one radiation is accompanied by, or shortly follows, another. The angiosperm plants are a massive radiation, as are the beetles. Cause and effect have been firmly suggested: the radiation of angiosperm plants provided a new set of niches into which beetles could speciate (Farrell, 1998). Arguments such as this are frequently anecdotal in that they involve post hoc explanations that require further testing, which often proves very difficult. For example, while the angiosperm plant radiation can explain about one-half of the species number in the beetle radiation (there are about 300,000 known beetle species), the other half remains unexplained. As Barraclough et al. (1998) point out, the angiosperm radiation does not explain the large radiation of 30,000 mainly carnivorous and fungivorous beetle species in the Cucujoida or the equally large radiation of predatory beetle species in the Staphylinoidea.

II. RECONSTRUCTED VERSUS REAL EVOLUTIONARY TREES

The reasons for the origins of contemporary biodiversity are becoming better understood, largely as a consequence of advances in molecular genetics. Gene sequence analysis has become routine in many laboratories, and sequence data are deposited in openly accessible data banks. When equivalent sequences are available from different species, phylogenetic trees can be reconstructed. If a molecular clock has been operating, these trees show when in relative time any pair of species last shared a common ancestor. When the molecular clock can be dated, the trees can be calibrated in real time. Such phylogenies show the relationships among contemporary species, and contain no explicit information about extinct lineages. However, the structure of such trees can be used to demonstrate which evolutionary models are more likely than others to explain the origins of biodiversity in particular taxa.

When a model-based approach is used to analyze the origins of biodiversity, it soon becomes apparent how easily intuition can lead us astray. For example, the tree reproduced in Fig. 1 shows the relationships among contemporary salamander species from the genus Plethodon. The branching rate in the tree seems to have increased recently. This, however, does not necessarily mean that the net rate of speciation has increased. The simplest model to describe the generation of biodiversity is a constant-rates model in which the rate of speciation (lineage splitting) and the rate of extinction have been the same in all lineages at all times since the root of the tree. Obviously the rate of speciation will have been greater than the rate of extinction, or the tree would not have grown. Under such a model, if the rate of extinction is zero, a lineages-through-time plot based on the relationships among contemporary species would be a straight line with a slope, A, the speciation rate. As the rate of extinction (μ) increases, so does curvature of the line upward toward the present (Fig. 2). This means that the pletho-


Biodiversity Generation, Overview

FIGURE 2 The expected forms of an actual phylogeny (top line) and a reconstructed phylogeny (bottom line) produced by a birth–death process, in which the constant per-lineage speciation rate (λ) is greater than the constant per-lineage extinction rate (μ). Note that the two curves would be superimposed if the extinction rate was zero. The initial steep slope of the line for the actual phylogeny is a consequence of only a sample of phylogenies surviving to the present and those are likely to be the ones that got off to a flying start. Note also that as μ increases from zero toward λ, so does the change in slope of the reconstructed line, so that a steep upturn describes a phylogeny where μ is high in comparison with λ. (After Harvey, P. H., May, R. M., and Nee, S. (1994). Phylogenies without fossils. Evolution 48, 523–529.)

Biodiversity cannot increase indefinitely. Eventually, any adaptive radiation will run out of niches to occupy and net speciation rates will decrease. Such an effect is apparent when Sibley and Ahlquist's (1990) molecular phylogeny of birds is analyzed as a lineages-through-time plot (Fig. 4). Instead of the line steepening toward the present, there is a general leveling off. In the absence of fossil material, which is relatively sparse for birds, it is not possible to say whether, in fact, speciation rates have decreased through time, extinction rates have increased, or both. Sibley and Ahlquist's phylogeny is one of the earlier molecular phylogenies and is based on DNA–DNA hybridization data: DNA from different species is annealed, and then heated to dissociate the strands. The more similar that strands are in base composition, the higher the temperature at which they dissociate. More recent studies based on direct comparison of sequences largely support Sibley and Ahlquist's original phylogeny, which, for example, indicated for the first time that the Australian passerines are an independent radiation from other passerines. In addition to many sequence-based studies that lend support for the general phylogeny, one reports a slowing down in the net rate of speciation in eight out of nine genera studied.

If we are to better understand the reasons for the generation of biodiversity, it is critical that molecular phylogenies are properly related to the known fossil record and to our understanding of biogeographic events. For example, the molecular phylogeny of birds...
FIGURE 3  Contour plot of the likelihood surface for the *Plethodon* data (Figure 1). The maximum likelihood estimate, the peak of the surface, is marked by an X. The lines surrounding it approximate to the 90, 95, and 99% confidence limits, respectively. (After Nee, S., Holmes, E. C., May, R. M., and Harvey, P. H. (1995). Estimating extinction from molecular phylogenies. In *Estimating Extinction Rates* (J. H. Lawton and R. M. May, eds.) Oxford University Press, Oxford, United Kingdom, pp. 164–182.)

FIGURE 4  A lineages-through-time plot for the molecular phylogeny of birds. All families are represented, that is, all lineages that were present about 45 million years ago and have left descendants in the present day. Accordingly, we should not expect the leveling off of the curve under a constant-rate process: either speciation rate has been decreasing over time or extinction rate has been increasing. There is no evidence for sharp kinks in the curve, which would provide evidence for mass extinction events. For example, the mass extinction at the end of the Cretaceous about 65 million years ago is thought to have resulted in the loss of about 75% of species in many taxa. (After Nee, S., Mooers, A. Ø., and Harvey, P. H. (1992). The tempo and mode of evolution revealed from molecular phylogenies. *Proc. Natl. Acad. Sci. USA* 89, 8322–8326.)

reveals two aberrant clades that were speciating at a higher rate than that for others was slowing down. The two radiations resulted in the Passeriformes (seed-eating birds) and the Ciconiiformes (mainly shorebirds). The fossil record of both groups is poor, but it is clear that the radiations occurred at a time when landmasses were warmer and shorelines longer than previously. Though cause and effect may be inferred, the unique events described leave the way open for other possible explanations. One way to strengthen interpretations is to show similar increases in biodiversity associated with the similar environmental changes on repeated occasions. If it could be shown, for example, that bursts in speciation among shorebirds were repeatedly associated with increases in the length of shoreline, it would be less likely that a third unknown environmental factor was responsible for changes in shorebird speciation rates.

III. KEY INNOVATIONS, SISTER TAXON ANALYSIS, AND THE ORIGINS OF ADAPTIVE RADIATIONS

It is not necessary that increases in biodiversity occur in response to environmental changes: so-called key innovations may evolve that allow the invasion of currently unoccupied niches. Liem (1973, 1980) suggested that one such innovation is responsible for the much higher rates of speciation among cichlid fishes than among their sister taxa. Liem pointed out that a small shift in position of a single muscle attachment ultimately allowed the pharyngeal bones of cichlids to manipulate their prey items while still holding them. As a consequence, the premaxillary and mandibular jaws were freed to evolve along new routes that did not involve manipulating prey. Thus, the argument runs, allowed cichlids to evolve a whole new diversity of feeding mechanisms and there seems little doubt that the adaptive radiation of cichlids in African lakes, in the face of competition from other fish families, resulted from their evolved diversity of feeding mechanisms. However, the problem of nonreplication occurs: Lauder (1981) correctly pointed out that any other characteristic common to cichlid fishes but that differs from a sister group could, in principle, be the key innovation that resulted in high speciation rates. Conclusions would be strengthened if the same key innovation had been associated with high species diversity in several taxa.
The first example of a statistically supported successful sister group comparison showing a repeated correlate of high species diversity comes from the Mitter et al. (1988) analysis of phytophagy, the habit of feeding on vascular plants, among insects. There are considerable barriers to the evolution of phytophagy. Behavioral and morphological adaptations are required to reduce the risk of desiccation, to remain attached to host, and to deal with low-nutrient food. It had been argued that, once those barriers were overcome, the diversity of phytophages would be promoted by both the great diversity of plant species and of plant parts that could not be exploited by a single species, and by the absence of competitors. Mitter et al. performed 13 sister taxon comparisons between the numbers of species in a phytophagous clade versus its nonphytophagous sister clade. Since sister clades originated at the same time, they have had identical times for diversification. In 11 of the 13 comparisons, there were more species in the phytophagous clade than in the sister group, and in each of those cases the difference was greater than twofold. Sister taxon comparisons using newly established molecular phylogenies are beginning to allow the first real statistical tests of many long-standing hypotheses. For example, Darwin (1871) suggested that sexual selection by female choice might increase the rate of reproductive divergence between populations, and thereby increase the speciation rate of a clade. More than a century later, Barraclough et al. (1995) tested the idea using Sibley and Ahlquist’s phylogeny of passerine birds. Since it is generally accepted that mate choice is responsible for the evolution of sexual dichromatism in birds, Barraclough et al. were able to use sexual dichromatism as a measure of the importance of mate choice for a species. In significantly more than 50% of sister taxon comparisons, the clade with the higher proportion of sexually dichromatic species was more speciose than its sister clade, thereby supporting Darwin’s suggestion.

IV. EVOLUTION WITHIN LINEAGES: THE COMPARATIVE METHOD

It is clear that some characteristics of species can promote speciation itself. But species differ from each other in many ways: morphologically, anatomically, physiologically, behaviorally, and by life history. Making sense of the reasons for that diversity is the job of comparative biologists.

A. Fewer Causes Than Characters

Traits have not evolved independently from each other, and particular characteristics of the environment select for whole suites of traits. For example, among many mammalian orders such as the primates, females tend to be spatially distributed in relation to food resources. If food is in large clumps, females live in large groups, but if food is evenly distributed, females are often territorial. Males distribute themselves in response to female grouping patterns, having been selected to monopolize mating access to as many females as possible. When females live alone in territories, males follow sun and share territorial defense (monogamous, as in gibbons). If females live in small groups, then a single male may be able to defend a group of females from mating access by other males (single-male societies). But if females live in groups that are sufficiently large that they cannot be defended by a single male, then more than one male may join the group (multimale societies). Among monogamous species, the males and females have similar weaponry, such as canines, having been selected equally to defend their territory. In single-male and multimale species, males have enlarged canines relative to females of the species, so that they can defend the group of females, or at least any that are in estrus, against mating access from other males. However, in multimale species, there is always the potential for females to be mated by more than one male; males in such species have markedly enlarged testes as a consequence of selection to produce more sperm per ejaculate, thereby increasing the chances of success in sperm competition. Males of monogamous and single-male species are not subject to selection for sperm competition. Here, then, the distribution of resources influences differences in social behavior, weaponry, and testes size. Indeed, if we know a male primate’s testes and relative canine sizes, it is possible to say whether he belongs to a monogamous species (both small), a single-male species (large canines, small testes), or a multimale species (both large).

B. Identifying Independent Evolutionary Events

As mentioned in the previous section, there are many fewer explanatory factors than there are variable traits across species. How do we identify the factors that are responsible for the patterns of biodiversity that we see today? Species may share traits in common for two evolutionary reasons: they inherited them from a common ancestor or their ancestors independently evolved the traits. Closely related species share more traits...
through common ancestry, whereas more distantly related species share more through convergent evolution. This does not necessarily mean that traits cannot evolve rapidly in response to selection. For example, when a new niche appears, the species to successfully invade that niche is most likely to be one that came from a similar niche elsewhere. Very minor changes in phenotype may be sufficient to adapt the new species to its new home. However, the fact that closely related species are subject to similar selective pressures does mean that cross-species correlations can be misleading. For example, in a sample of vertebrates, we might find that those with feathers laid eggs while those with fur gave birth to live young. If the sample contains just birds and mammals, we should be wary of saying that the two sets of characters had responded to the same selective pressure. The way forward is to identify evolutionary independent origins of traits and niche occupancy. For example, we know that on many separate occasions bright coloration among insects has evolved together with distastefulness to predators. Because the two traits seem to have evolved in concert, and there is good reason to expect distasteful prey to advertise the fact to predators, we are more confident that they are part of the same adaptive complex.

For continuously varying characters, such as body size, a particularly elegant technique was developed by Felsenstein (1985) that allows comparative biologists to partition out independent evolutionary change and seek correlated evolution among traits. The key realizations are (1) that differences between each pair of sister taxa in a phylogeny evolved independently, and (2) likely ancestral character states could be estimated if characters had evolved according to a Brownian motion model of evolution. To determine whether characters had evolved independently of each other, it is simply necessary to plot sister taxa differences against each other. If one character evolved independently of another, differences between sister taxa would not be correlated (Fig. 5).

An example of the use of independent contrast analyses is Kelly and Woodward’s (1995) analysis of the correlates of variation in carbon isotope composition (δ13C) among plant species. Three cross-species correlates of δ13C had previously been identified: altitude, latitude, and growth form. When the data were analyzed using phylogenetically independent contrasts, the correlation with latitude dropped out, and when altitude was controlled for there was no correlation with growth form. Alternative interpretations of the reasons for species differences in δ13C were thereby reduced in number. Indeed, it was concluded that differences in atmospheric composition (CO2 and O2 partial pressures) may be sufficient to explain observed differences in carbon isotope composition.

See Also the Following Articles

ADAPTIVE RADIATION • BIODIVERSITY, ORIGIN OF • CLADOGENESIS • EVOLUTION, THEORY OF • EXTINCTION, RATES OF OF • FOSSIL RECORD • PHYLOGENY • SPECIATION, PROCESS OF

Bibliography


A GOOD FOSSIL RECORD OF LIFE ON EARTH is available since the beginning of the Cambrian Period about 545 million years ago. Over the interval between then and now, there has been a substantial increase in biodiversity interrupted by many extinction events which caused sharp but relatively brief dips in biodiversity. Modern-day biodiversity is close to the highest it has ever been.

Life began on Earth about 3.5 billion years ago in the Archean Eon, probably with the appearance first of self-reproducing RNAs and later of prokaryotic single-celled organisms. No actual fossils date from this far back, and so our knowledge of Archean life is sketchy at best. Approximately 2.5 billion years ago the last stocks of elemental iron in the earth's crust were converted to oxides as a result of photosynthesis by cyanobacteria, and the planet's atmosphere changed from reducing to oxidizing, making oxygen-breathing life possible. The Proterozoic Eon which followed saw the first appearance of eukaryotes (organisms with nucleated cells and mitochondria), sexual reproduction, and multicellular organisms, although the precise timing and even the order of these innovations are disputed. The earliest firm evidence of metazoan multicellularity dates back to about 575 million years ago.

The end of the Proterozoic at the Vendian--Cambrian boundary about 545 million years ago marks the start of the Cambrian explosion, a period of approximately 25 million years during which, for unknown reasons, multicellular life underwent a period of extraordinary diversification, producing a multitude of new evolu-
tionary lineages in a comparatively short time, includ-
ing almost all of the major metazoan body plans seen
today as well as many which have become extinct. The
fossil record from the Cambrian onward—an interval
known as the Phanerozoic Eon—is good by comparison
with that of the Precambrian largely because multicellu-
lar organisms, whose fossils are easier to find and study
than those of single-celled organisms, became nu-
merous.

The Phanerozoic fossil record consists of about a
quarter of a million known species. The vast majority
of these are marine species, mostly invertebrates such
as mollusks, brachiopods, corals, or foraminifera as well
as sea-dwelling plants. Terrestrial (i.e., land-dwelling)
organisms are far less well preserved because the deposi-
tional regimes under which fossils are formed are less
reliable and uniform on land. Even for marine species,
there are many biases in the fossil record, among which
the following are some of the most important:

1. Deposition can vary greatly from one time to an-
other. Even in the oceans, in which deposition is rela-
tively reliable, there are periods during which preserva-
tion is excellent and others in which it is poor. A sudden
period of poor preservation can even give the appear-
ance of a large extinction event because the number of
preserved organisms decreases substantially.

2. The “Pull of the Recent” is the name given to the
greater accessibility of recently formed rock and the
better preservation of fossils in these rocks. The Pull
of the Recent implies that we have a more complete
record of recent times than we do of times long past,
and this could produce the misleading impression of
an increase in diversity toward the present where none
exists. Despite this bias, it is still believed (as discussed
later) that there has been a real increase in biodiversity
over the course of the Phanerozoic.

3. The “monograph effect” refers to the sudden ap-
parent burst of new species resulting from the attention
paid by a particularly zealous researcher or group of
researchers to a particular section of the fossil record:
The thorough investigation of a short interval can turn
up many previously unknown species, making that in-
terval appear to be one of especial evolutionary activity
when in fact a similarly thorough investigation of an-
other interval would have turned up just as many spe-
cies. A similar effect is produced by the discovery of a
particularly rich and well-preserved fossil bed dating
from one particular time.

4. Although it is now possible to date fossils within
a few million years with high confidence, more accurate
dating in the fossil record is still problematic. Absolute
dates are derived from slowly decaying radioactive iso-
topes, whereas relative dates of different fossils are de-
rived from stratigraphic evidence, both geological and
from fiducial lineages (so-called “index fossils”). Some
geographic areas and intervals of time can be dated with
high accuracy—a million years or better—but others
are not nearly as good.

Despite these biases, the overall pattern of life during
the Phanerozoic survives the vagaries of the record and
is now reasonably well understood.

Figure 1 is a graph of (marine) biodiversity as a
function of time during the Phanerozoic. The horizontal
axis measures time before the present, whereas the ver-
tical one measures the number of known families of
marine animals at a succession of times. Each point
corresponds to one stratigraphic stage; stages are irregu-
lar intervals of time of average duration of about 7
million years which are based on widely accepted strati-
graphic features. The data are taken from the compila-

One of the most notable features of Fig. 1 is a sub-
stantial increase in the number of known families over
the course of the Phanerozoic, particularly in the 250
million years since the Late Permian mass extinction.
Overall, the number of known families increases by
more than a factor of 10 from the earliest Cambrian to
the present. In considering this increase, one must first
ask whether it is genuine or an artifact of biases in the
fossil record. As mentioned previously, there is certainly
a bias occurring because of the better preservation and
greater availability of fossils in more recent rocks. There is no doubt that some of the increase apparent in Fig. 1 is a result of this bias. On the other hand, the use of numbers of families as a measure of diversity, rather than species, tends at least partly to remove this bias since we need find only one species belonging to a certain family in order to establish the existence of that family at any given time. Therefore, even during periods of relatively poor species sampling, a family with a sufficiently large number of species can still contribute to our estimate of diversity. Furthermore, there are plausible biological reasons to believe in a real diversity increase. The increase seems to some extent to be the result of the evolution of organisms into new ways of life. The spread of life onto the land in the Silurian and Devonian is a particularly prominent example of this, but among marine organisms the trend is also clear. It seems reasonable to suppose that as life spreads into new environments, the total number of species (or families) which the planet can support will also increase.

Many models have been put forward for diversity increases in the fossil record. Benton (1995) suggested that the increase is an exponential one, indicating perhaps a stochastically constant rate of creation of new taxa from existing ones, with no detectable limit to ultimate levels of diversity. A more detailed analysis has been given by Sepkoski (1981), who suggested that the diversification of life occurred in three distinct phases, each typified by vigorous turnover within a particular subset of species or “fauna.” The first phase, during the Cambrian, was dominated by groups in which the rate of origination of new species was relatively high, such as trilobites, inarticulate brachiopods, and monoplacophorans. The second phase, lasting from the Ordovician to the end of the Permian, was dominated by groups with intermediate origination rates, such as crinoids, articulate brachiopods, and cephalopods. Finally, there was a third phase, lasting from the beginning of the Triassic up until the Recent, which was dominated by bivalves, gastropods, echinoids, and marine vertebrates—groups which have relatively low rates of origination. Sepkoski (1984) suggested that within each of the three faunas diversification proceeds initially exponentially but eventually saturates, implying the existence of equilibrium “carrying capacities” for the global ecosystem. The superimposition of this “logistic” growth within each fauna gives rise to the observed pattern of near-monotonic diversity increase in Fig. 1. A logistic diversification model was also explored by Courtillot and Gaudemer (1996) but without allusion to distinct faunas.

An important feature of the pattern of biodiversity shown in Fig. 1 which can be explained in terms of the logistic growth model is the apparent plateau in diversity throughout the greater part of the Paleozoic, from the Late Ordovician to the end of the Permian. [The same pattern is also present in genus-level data (Sepkoski, 1997), although there is in general more fluctuation in the numbers of genera so the plateau is not as clear.] One explanation for this plateau is that it represents the saturation phase of the logistic diversification of Sepkoski’s second fauna, which continues mostly uninterrupted for approximately 200 million year until the Late Permian mass extinction. This is not the only possibility, however; it has also been suggested that the plateau is more the result of high extinction rates in the later Paleozoic than it is the result of low origination rates (Stanley, 1999). In both cases, however, the ultimate limit on diversity is presumed to be some effective carrying capacity of the ecosystem.

Another notable feature of Fig. 1 is a dip in diversity at several points, particularly at the ends of the Devonian (D), Permian (P), and Cretaceous (K) periods. These dips correspond to major mass extinctions that can be seen more clearly in Fig. 2, which shows the percentage of extant families of marine animals becoming extinct in each stratigraphic stage of the Phanerozoic. The data are again taken from the compilation by Sepkoski (1992). We normally distinguish five major mass extinctions during the Phanerozoic. These “big five” are marked with arrows in Fig. 2. A sixth period of heightened extinction is visible in the Cambrian but is thought to be primarily an artifact of sampling biases.

<table>
<thead>
<tr>
<th>Time before present (in millions of years)</th>
<th>Percentage of families becoming extinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>200</td>
<td>30</td>
</tr>
<tr>
<td>300</td>
<td>40</td>
</tr>
<tr>
<td>400</td>
<td>50</td>
</tr>
<tr>
<td>500</td>
<td>60</td>
</tr>
</tbody>
</table>

**FIGURE 2.** Estimated extinction of marine animals in families per stratigraphic stage since the Cambrian as a percentage of the total number of families in existence. The arrows indicate the positions of the “big five” mass extinction events discussed in the text.
or carbon dioxide and sulfur release following large or a shift in ocean circulation driven by climate change, levels of oxygen in the oceans were low, and it is clear, however, that the sea level rose during this period. The Late Permian event are still a subject of debate. It much amount of research of the subject, the causes of extinctions are unclear. The leading theories suggest that changes in sea level and ocean anoxia, possibly triggered by global cooling or oceanic volcanism, were most likely responsible, although the impact of an extra- terrestrial body such as a comet has also been considered.

The Late Devonian extinction approximately 360 million years ago is complex and poorly understood. It is probably in fact composed of many separate events (as many as seven) spread over about 25 million years, including particularly notable extinctions at the ends of the Givetian, Frasnian, and Famennian stages. Overall, about 80% of living species died out in the Late Devonian. Particularly hard hit were corals, brachiopods, bryozoans, ammonoids, and fish. The causes of these extinctions are unclear. The leading theories suggest that changes in sea level and ocean anoxia, possibly triggered by global cooling or oceanic volcanism, were most likely responsible, although the impact of an extra-terrestrial body such as a comet has also been considered.

The Late Permian extinction approximately 250 million years ago was the largest extinction event of all time, killing approximately 99% of marine species and about 70% of land-dwelling ones. Like the end-Ordovician event, it seems to have been composed of two bursts, separated in this case by an interval of about 10 million years, with the second being the larger of the two. Notable extinction happened again among brachiopods, ammonoids, and corals as well as gastropods, echinoderms, and, unusually, insects. Despite an enormous amount of research of the subject, the causes of the Late Permian event are still a subject of debate. It is clear, however, that the sea level rose during this period, levels of oxygen in the oceans were low, and carbon dioxide levels were high. There is some suggestion that a cometary impact may have been involved, or a shift in ocean circulation driven by climate change, or carbon dioxide and sulfur release following large-scale volcanic activity. The Late Permian event had a profound effect on the terrestrial ecosystem which is still being felt today, a quarter of a billion years later. A particularly notable example among marine faunas is that of the bivalves, a relatively minor group during the Paleozoic that took advantage of the ecological vac- uum left by the extinction to establish a solid grip on shallow-water environments, leading to their dominance over the previously very successful brachiopods.

The end-Triassic extinction approximately 210 mil- lion years ago is probably the most poorly understood of the big five extinction events. It appears to have killed about 80% of species then living, either in one burst or possibly in two separated by about 20 million years. Major extinction is observed particularly among ammonoids, bivalves, gastropods, and brachiopods. Leading theories of the causes of the end-Triassic event are ocean anoxia, massive volcanism, or possibly a bo- lide impact.

The end-Cretaceous event, usually called the Creta- ceous–Tertiary or K–T event, has attracted the most popular interest of any extinction because it saw the end of the dinosaurs, but it was in fact the smallest by quite a wide margin of all the big five. The K–T event appears to have been a single pulse of extinction approx- imately 65 million years ago which wiped out about 70% of all species then living. In addition to the dino- saurs, it extinguished many other land-dwelling verte- brates, especially large-hooded ones, along with large numbers of (marine) bivalves, gastropods, and foramin- ifera. The proximal cause of the K–T event was almost certainly the impact of a large comet or meteor near the present site of the town of Puerto Chicxulub on the Yucatan peninsula in eastern Mexico, with an associated drop in sea level and possibly short-term cooling or heating or acid rain. It has also been noted that average extinction rates were higher than the historical average for some time in advance of the K–T boundary, indicat- ing that long-term environmental change may also have played a part.

Table I summarizes the fraction of families and spe- cies killed in each of the big five mass extinctions. In fact, the fraction of species killed cannot be measured directly from the fossil record since the preservation of individual species is too unreliable; the majority of species appear in only one stratigraphic stage so that no statistically significant estimate of their origination or extinction time can be made. The fraction of families killed can be calculated with reasonable certainty: Many families have reasonable coverage in the fossil record so that good estimates can be made of origination and extinction times. Estimates of species kill are then ex-
TABLE I
Extinction Intensities at the Family and Species Level for the Big Five Mass Extinctions of the Phanerozoic

<table>
<thead>
<tr>
<th>Extinction</th>
<th>Family loss (observed) (%)</th>
<th>Species loss (estimated) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>End Ordovician</td>
<td>26</td>
<td>84</td>
</tr>
<tr>
<td>Late Devonian</td>
<td>22</td>
<td>79</td>
</tr>
<tr>
<td>Late Permian</td>
<td>91</td>
<td>95</td>
</tr>
<tr>
<td>End Triassic</td>
<td>22</td>
<td>79</td>
</tr>
<tr>
<td>End Cretaceous</td>
<td>16</td>
<td>70</td>
</tr>
</tbody>
</table>

* Estimates of family extinction are obtained from directed analysis of the fossil record, whereas species loss is inferred using a statistical technique called “reverse rarefaction.” Data are taken from Jablonski (1995).

Another overall feature of the extinction record shown in Fig. 2 is a decline in the level of background extinction (i.e., not mass extinction) during the Phanerozoic. Like the increase in standing diversity discussed earlier, it is important to establish whether this decline is a real effect or the result of some bias in the record. It is difficult to imagine that a decrease in the number of families becoming extinct could arise from better preservation toward the Recent. Rather, one would expect the reverse since, if extinction were actually approximatively constant but more fossil families were preserved, one would expect an increase in apparent extinction. What is the explanation for the decrease? Many hypotheses have been put forward.

Flessa and Jablonski (1985) suggested that the decrease could be a result of increases in the average size of families. Following Raup (1979), data from living species are usually used for this purpose. Calculating the average probability of family kill gives us a “rarefaction curve” (Fig. 3) for the fraction of families or genera killed as a function of the fraction of species killed. “Reverse rarefaction” is the name given to the use of this curve “backwards” to estimate what fraction of species must have been killed in order to produce an observed level of family kill. This is the method which is used to calculate the data in Table 1.

Another overall feature of the extinction record shown in Fig. 2 is a decline in the level of background extinction (i.e., not mass extinction) during the Phanerozoic. Like the increase in standing diversity discussed earlier, it is important to establish whether this decline is a real effect or the result of some bias in the record. It is difficult to imagine that a decrease in the number of families becoming extinct could arise from better preservation toward the Recent. Rather, one would expect the reverse since, if extinction were actually approximatively constant but more fossil families were preserved, one would expect an increase in apparent extinction. What is the explanation for the decrease? Many hypotheses have been put forward.

Flessa and Jablonski (1985) suggested that the decrease could be a result of increases in the average size of families. Following Raup (1979), data from living species are usually used for this purpose. Calculating the average probability of family kill gives us a “rarefaction curve” (Fig. 3) for the fraction of families or genera killed as a function of the fraction of species killed. “Reverse rarefaction” is the name given to the use of this curve “backwards” to estimate what fraction of species must have been killed in order to produce an observed level of family kill. This is the method which is used to calculate the data in Table 1.

Another overall feature of the extinction record shown in Fig. 2 is a decline in the level of background extinction (i.e., not mass extinction) during the Phanerozoic. Like the increase in standing diversity discussed earlier, it is important to establish whether this decline is a real effect or the result of some bias in the record. It is difficult to imagine that a decrease in the number of families becoming extinct could arise from better preservation toward the Recent. Rather, one would expect the reverse since, if extinction were actually approximatively constant but more fossil families were preserved, one would expect an increase in apparent extinction. What is the explanation for the decrease? Many hypotheses have been put forward.

Flessa and Jablonski (1985) suggested that the decrease could be a result of increases in the average size of families. Following Raup (1979), data from living species are usually used for this purpose. Calculating the average probability of family kill gives us a “rarefaction curve” (Fig. 3) for the fraction of families or genera killed as a function of the fraction of species killed. “Reverse rarefaction” is the name given to the use of this curve “backwards” to estimate what fraction of species must have been killed in order to produce an observed level of family kill. This is the method which is used to calculate the data in Table 1.

Another overall feature of the extinction record shown in Fig. 2 is a decline in the level of background extinction (i.e., not mass extinction) during the Phanerozoic. Like the increase in standing diversity discussed earlier, it is important to establish whether this decline is a real effect or the result of some bias in the record. It is difficult to imagine that a decrease in the number of families becoming extinct could arise from better preservation toward the Recent. Rather, one would expect the reverse since, if extinction were actually approximatively constant but more fossil families were preserved, one would expect an increase in apparent extinction. What is the explanation for the decrease? Many hypotheses have been put forward.

Flessa and Jablonski (1985) suggested that the decrease could be a result of increases in the average size of families. Following Raup (1979), data from living species are usually used for this purpose. Calculating the average probability of family kill gives us a “rarefaction curve” (Fig. 3) for the fraction of families or genera killed as a function of the fraction of species killed. “Reverse rarefaction” is the name given to the use of this curve “backwards” to estimate what fraction of species must have been killed in order to produce an observed level of family kill. This is the method which is used to calculate the data in Table 1.

Another overall feature of the extinction record shown in Fig. 2 is a decline in the level of background extinction (i.e., not mass extinction) during the Phanerozoic. Like the increase in standing diversity discussed earlier, it is important to establish whether this decline is a real effect or the result of some bias in the record. It is difficult to imagine that a decrease in the number of families becoming extinct could arise from better preservation toward the Recent. Rather, one would expect the reverse since, if extinction were actually approximatively constant but more fossil families were preserved, one would expect an increase in apparent extinction. What is the explanation for the decrease? Many hypotheses have been put forward.

Flessa and Jablonski (1985) suggested that the decrease could be a result of increases in the average size of families. Following Raup (1979), data from living species are usually used for this purpose. Calculating the average probability of family kill gives us a “rarefaction curve” (Fig. 3) for the fraction of families or genera killed as a function of the fraction of species killed. “Reverse rarefaction” is the name given to the use of this curve “backwards” to estimate what fraction of species must have been killed in order to produce an observed level of family kill. This is the method which is used to calculate the data in Table 1.

Another overall feature of the extinction record shown in Fig. 2 is a decline in the level of background extinction (i.e., not mass extinction) during the Phanerozoic. Like the increase in standing diversity discussed earlier, it is important to establish whether this decline is a real effect or the result of some bias in the record. It is difficult to imagine that a decrease in the number of families becoming extinct could arise from better preservation toward the Recent. Rather, one would expect the reverse since, if extinction were actually approximatively constant but more fossil families were preserved, one would expect an increase in apparent extinction. What is the explanation for the decrease? Many hypotheses have been put forward.

Flessa and Jablonski (1985) suggested that the decrease could be a result of increases in the average size of families. Following Raup (1979), data from living species are usually used for this purpose. Calculating the average probability of family kill gives us a “rarefaction curve” (Fig. 3) for the fraction of families or genera killed as a function of the fraction of species killed. “Reverse rarefaction” is the name given to the use of this curve “backwards” to estimate what fraction of species must have been killed in order to produce an observed level of family kill. This is the method which is used to calculate the data in Table 1.

Another overall feature of the extinction record shown in Fig. 2 is a decline in the level of background extinction (i.e., not mass extinction) during the Phanerozoic. Like the increase in standing diversity discussed earlier, it is important to establish whether this decline is a real effect or the result of some bias in the record. It is difficult to imagine that a decrease in the number of families becoming extinct could arise from better preservation toward the Recent. Rather, one would expect the reverse since, if extinction were actually approximatively constant but more fossil families were preserved, one would expect an increase in apparent extinction. What is the explanation for the decrease? Many hypotheses have been put forward.
the highly volatile families, the time to extinction of the entire family, as a result of extinction of all its members, will be shorter than that for less volatile ones by precisely the ratio of the respective volatilities, meaning that the highly volatile families become extinct more quickly. Over the course of a long period of time, therefore, the families that remain will tend to be the less volatile, longer lived ones. Taxon sorting in effect produces a selection pressure at the family or other higher taxonomic level. Another possible explanation for the increase in the lifetimes of taxa and decrease in extinction rates is that species have really become more highly evolved toward the Recent and are better adapted to survive (Raup and Sepkoski, 1982). This explanation, however, is not a widely accepted one.

Although extinction has received the lion’s share of paleontological attention in recent years, origination is equally important in the study of Phanerozoic biodiversity. We have already discussed origination in the context of the “three faunas” view of the diversification of life. Figure 4 shows a detailed plot of the measured pattern of origination of families during the Phanerozoic. Many prominent features are visible. The major peaks in the Cambrian period are a reflection of the burst in metazoan diversity of the Cambrian explosion. Since the plot is one of origination rates as a fraction of standing diversity, however, the size of these peaks is partly just a result of low Cambrian diversity (Fig. 1). Another burst of origination is visible in the Ordovician. This burst was responsible for the corresponding major increase in diversity shown in Fig. 1. After the Ordovician, origination is more subdued; although substantial surges did take place in the Triassic, Jurassic, Cretaceous, and during the Cenozoic.

On the whole, origination tends to vary less in magnitude and to be less episodic than extinction (Foote, 1994). Nonetheless, it appears to be bursts of origination, rather than declines in extinction rates, which account for most of the major evolutionary radiations. The basic features of the Ordovician, Jurassic–Cretaceous, and Cenozoic radiations are as follows.

Origination during the Ordovician occurred particularly among groups such as brachiopods, bryozoans, and mollusks, increasing the dominance of these groups while groups such as trilobites shrank in importance. Although phyla had basically ceased to be produced by the Ordovician, the period led to substantial production of body plans at the level of classes and orders. The Jurassic–Cretaceous radiation produced substantial increases in the diversity of gastropods, bivalves, echinoids, and fishes—groups that were relatively minor during the Paleozoic. Ecological expansion in the form of the appearance of fundamentally new modes of life (Bambach, 1985) and the increasing occurrence of predation (Vermeij, 1977) have been advanced as important contributors to this pattern.

The Cenozoic radiation is certainly in part a product of the Pull of the Recent, but there are also genuine bursts of origination in the Paleocene and Eocene. In the marine realm, the diversity of veneroid bivalves, neogastropods, irregular echinoids, gymnothelmate bryozoans, and fishes increased substantially, whereas in the terrestrial realm the mammal and angiosperm radiations are particularly prominent.

Although mass extinctions are usually explained as the results of various kinds of exogenous stress on the ecosystem, such as climate change or bolide impact, originations seem to be more often the result of biological phenomena such as evolutionary innovations or changes in the structure of ecological assemblages. However, as a close inspection of Figs. 1, 2, and 4 reveals, origination is often intensified in the aftermath of mass extinction events, suggesting that physical disturbances do have an indirect effect on origination, probably through removal of incumbent species, through selectivity in extinction (Jablonski, 1989), and through particular patterns of rediversification of groups and repopulation of ecospaces during the recovery period following extinction. Recoveries typically take a few million years to run their course and reestablish previous levels of diversity and ecological heteroge-
of diversity. Understanding the extent of the Cambrian explosion hinges on a better understanding of such patterns. A paradigmatic example of recovery is the Triassic recovery which followed the end Permian (Fig. 4). A conspicuous spike in origination is observed beginning about 5 million years after the end of the extinction and continuing for about another 5 million years.

A long-term decline in origination over the Phanerzoic is as clear a feature of the fossil record as is the decline in extinction. Taxon sorting of the kind described earlier in the context of extinction is a plausible explanation (Gilinsky, 1994). Alternatively, the decline in origination may reflect a genuine decrease in the rate of evolutionary innovation as a result of either ecological or developmental restriction (Gilinsky and Bambach, 1987; Eble, 1999).

It is also instructive to examine origination rates at the level of higher taxa such as classes and orders. Origination at these levels reflects real innovations (to the extent that they are visible in the fossil record) rather than just sheer numbers of species. Indeed, there is no reason why the two need necessarily be correlated. During the Cambrian, for example, changes in total diversity (e.g., measured by the number of families) are modest, but it is in the Cambrian that one finds the most significant generation of new body plans—the "Cambrian explosion." This aspect of biodiversity having to do with numbers of evolving entities but with their distinction is often referred to as "morphological diversity." "Biodiversity," or simply "disparity" (Foote, 1997). Significant discrepancies between biodiversity and biodisparity can exist, such as faster morphological diversification early on in the evolution of clades, despite low levels of taxonomic diversity, and later deceleration despite continued increases in diversity. Further understanding of biodiversity in the fossil record may hinge on a better understanding of such patterns.

See Also the Following Articles

DIVERSITY, ORIGIN OF • FOSSIL RECORD • MASS EXTINCTIONS, CONCEPT OF • MASS EXTINCTIONS, NOTABLE EXAMPLES OF • ORIGIN OF LIFE, THEORIES OF • PALEOECOLOGY • SPECIES DIVERSITY, OVERVIEW

Bibliography


BIODIVERSITY-RICH COUNTRIES

José Sarukhán and Rodolfo Dirzo
Instituto de Ecología, Universidad Nacional Autónoma de México

I. Introduction
II. Facets of Biodiversity in Megadiverse Countries
III. Underlying Causes of Megadiversity
IV. Selected Megadiversity Countries
V. Threats to Biodiversity in Megadiverse Countries
VI. Sociodemographic Characteristics and Their Relation to Biodiversity
VII. Human and Institutional Capacity for Biodiversity Knowledge and Conservation
VIII. Case Studies of Managing Biological Information in Megadiverse Countries
IX. Corollary

GLOSSARY

**beta-diversity** Turnover of species within a small spatial scale resulting from highly differing ecological conditions.

**biogeographic regions** Major areas of the world where large assemblages of species originated and that differ considerably from other regions.

**CONABIO** The Mexican National Commission on Biodiversity.

**Convention on Biodiversity (CBD)** An international convention signed by 175 countries, which originated at the Earth Summit of Rio de Janeiro in 1992.

**cultural diversity** Variation of races, languages, and habits that define a given culture.

**database** Assemblage of information organized in tables with a logical structure. It is a fundamental tool used to gather, organize, and analyze biodiversity information.

**endemism** Confinement of a species or other taxonomic group to a given region, for example, an island or a country.

**ERIN** The Environmental Resources Information Network of Australia.

**INBio** The Biodiversity Institute of Costa Rica.

**megadiverse** An area or a country possessing a much larger proportion of species than would be expected by its extent, latitudinal position, and other factors.

**parataxonomist** Laymen and peasants trained to collect and to carry out preliminary identification of botanical and zoological specimens. They are widely used by INBio in Costa Rica.

**phylum** A major category, between kingdom and class, of taxonomic classification.

**primary production** Amount of energy produced by photosynthetic organisms in a community.


BIOLÓGICAL DIVERSITY IS NOT HOMOGENEOUSLY DISTRIBUTED AROUND THE GLOBE. The latitudinal variation in biological diversity is an obvious
example. However, superimposed on this and other natural trends of spatial variation is the heterogeneous distribution of diversity according to the geographical position of political entities, such as countries. A handful of countries in which biodiversity is particularly overrepresented constitute the so-called megadiverse countries. Obviously there is no reason for biodiversity to follow the artificial boundaries of countries. However, the reality is that such geopolitical entities exist, and they have particular ecological, historical, social, and economic structures that must be taken into account in any attempt to view, value, and conserve global biodiversity.

I. INTRODUCTION

The notion of megadiversity countries was first suggested by the well-known conservation biologist Russell Mittermeier, who developed it with an initial emphasis on tropical primates. Later it was extended to all types of ecosystems and several groups of organisms (Mittermeier et al., 1997). The concept of megadiversity countries is close to that of centers of diversity, which refers to the existence of areas with high biodiversity, particularly large numbers of species and a high concentration of endemic organisms. However, there are important distinctions. Centers of diversity are natural spatial units, and they may be recognized at several scales, such as local, regional, and global. Megadiversity countries, on the other hand, are spatial entities restricted within geopolitical limits and the recognition of variation at different scales can only be made within the countries' artificial boundaries or at the global scale.

It is not surprising that there is a lack of precise numerical coincidence between centers of diversity and megadiverse countries. For instance, at the global scale, the IUCN (World Conservation Union) Plant Conservation Office has recently recognized 234 centers of plant diversity, while the number of megadiverse countries, from the botanical point of view, is only about 10. However, some spatial overlap can be expected: of the total number of centers of plant diversity, close to 50% are located within these 10 megadiverse countries. In addition, the notion of megadiversity country implies several biopolitical connotations. One of the most evident of these is that such countries have a special responsibility to protect such high concentrations of biological resources. That such concentrations of the total biodiversity of the planet are located in a handful of countries constitutes a matter of global as well as national concern.

This article describes the salient facets of biodiversity that determine the existence of megadiverse countries, and then addresses the current threats that this diversity is experiencing. A discussion of relevant sociodemographic traits of megadiverse countries is presented, together with the repercussions for the conservation and management of natural resources. Special attention is given to the large indigenous human populations that inhabit most of these countries. Finally, we explore the factors that limit the capacity of many megadiverse countries to fully know their biological diversity and to be able to utilize that knowledge for the purpose of conserving and managing their natural ecosystems. We examine examples of how some countries have developed institutional efforts to compile, systematize, and utilize biodiversity information in order to convert it into conservation policy.

II. FACETS OF BIODIVERSITY IN MEGADIVERSE COUNTRIES

The first issue when describing the underlying causes for the concentration of biodiversity in specific countries is the establishment of criteria upon which the definition of megadiversity is based. Although several authors have used different approaches to rank areas according to their biodiversity, consideration of species richness predominates in this article, and following Mittermeier et al. (1997), megadiversity is defined on the basis of the following facets of biological diversity.

A. Species Richness

We base our assessment in terms of the species of plants and four groups of animals: mammals, birds, amphibians, and reptiles (which are, in general, the best known taxa). Species richness in these groups of organisms, and plant species in particular, can be taken as correlates of species richness of other, lesser-known organisms. For example, species richness of a critical group, insects (at least some groups), is highly dependent on plant species richness. When available, species richness of other groups of organisms was also used.
B. Concentration of Endemisms

The concentration of endemisms refers to the percentage of endemic species present among a country’s total resident species. This criterion is also based on the best-known groups of organisms: plants, mammals, birds, amphibians, and reptiles.

C. Diversity of Habitats

The diversity of habitats refers to the diversity of distinct ecosystems, usually considered to be the diversity of vegetation types. Closely related to this component is the concept of beta diversity, which is the spatial turnover of species between nearby or adjacent habitats. High beta diversity frequently results from the heterogeneity of ecological conditions. Commonly, significant species turnover also occurs at regional scales within a given vegetation type.

D. Presence of Tropical Forest Ecosystems

Given that tropical forests, particularly tropical rain forests, are widely recognized as the most species-rich terrestrial ecosystems, their presence has been used as another important facet of megadiversity. This criterion was used only in qualitative terms, with no attempt at quantifying the expanse of tropical forests per country.

E. Presence of Marine Ecosystems

The presence of marine ecosystems was also included because such systems hold the greatest diversity of taxa of higher rank. In particular, marine ecosystems are rich in phyla of animals: of the 33 known animal phyla, 28 are present in the marine realm and 13 are exclusive to it. In contrast, terrestrial ecosystems include 10 phyla, with only one endemic to it. In addition, marine ecosystems contain a great diversity of fish, which was used as a complementary criterion.

F. Cultural Diversity

The cultural diversity criterion addresses the diversity of ethnic groups (or autochthonous languages) that are present in a given country. The relevance of this component stems from the fact that there is a strong association between biological diversity and cultural diversity. Different cultural dimensions in different countries have played a significant role in the modes in which biodiversity is perceived, maintained, used, and appreciated.

III. UNDERLYING CAUSES OF MEGADIVERSITY

The next step is the analysis of the causes, proximal or ultimate, that determine the conjunction of the foregoing attributes in a given country. The predominant underlying causes that determine the existence of the megadiversity countries are latitudinal position, physical factors that increase primary production, historical factors, ecological diversity, and the presence of tropical rain forests.

A. Latitudinal Position

Latitudinal position is correlated with a number of ecological factors; by itself it is not a determining factor of high biodiversity. Rather, it is a proximal correlate of biodiversity: for almost all groups of organisms, species diversity increases toward equatorial latitudes. For example, on land, species richness and latitude are negatively correlated among mammals, birds, and trees (Fig. 1). Likewise, the number of species of ants in local regions increases from about 10 at 60°N latitude to about 2000 in equatorial regions.

It follows, therefore, that countries located at lower latitudes tend to be rich in species of many groups. Table 1 shows the contrast between tropical and temperate countries in terms of the number of plant and mammal species. Regardless of size, the sample of countries located close to tropical latitudes have plant and mammal diversities that are, on average, two times larger than those of the temperate countries. Evidently the tropical countries of this sample constitute a list of potential megadiversity candidates. Nevertheless, Table 1 shows that some of the temperate countries (e.g., Australia) reach values of species richness that are comparable to those of tropical countries. This implies that other factors besides latitude may be important correlates or causal factors of species richness. For example, in the case of Australia, its large territorial size and the fact that its northern region has tropical ecosystems account for its richness.

It may also be noted that historical factors could have played an important role in establishing the latitudinal
patterns of distribution of biological richness. High latitudes were covered by ice until relatively recently, whereas most equatorial latitudes have been relatively stable from the climatic point of view. This allows for a longer period of speciation and accumulation of taxa at lower than at higher latitudes.

The patterns of decrease in species richness with latitude in terrestrial ecosystems is paralleled in marine systems, although available information relates to other groups of organisms. For example, Arctic waters contain about 100 species of tunicates (sea squirts), temperate seas contain some 400 species, and waters at tropical latitudes include over 600 species. Likewise, the richness of species, genera, and families of bivalve mollusks peaks in tropical regions and declines with increasing latitude (Fig. 2). Although data are not readily available for individual countries, the similarity of latitudinal patterns in terrestrial and marine systems suggests that a similar trend exists on a per country basis, which again indicates that low-latitude countries are likely megadiversity candidates on the basis of marine species richness.

B. Physical Factors of the Environment Leading to Increased Primary Production

In many parts of the world, primary production increases with precipitation, available solar energy, and nutrient concentration (e.g., soil fertility). The highest terrestrial primary production is found in areas with high rainfall and year-round warm temperatures, whereas production decreases with lowered temperatures and the occurrence of frost. In marine settings, primary production increases with nutrient concentra-

<table>
<thead>
<tr>
<th>Tropical country</th>
<th>Plants</th>
<th>Mammals</th>
<th>Temperate country</th>
<th>Plants</th>
<th>Mammals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>50,000–56,000</td>
<td>524</td>
<td>Argentina</td>
<td>9,000</td>
<td>255</td>
</tr>
<tr>
<td>Colombia</td>
<td>45,000–51,000</td>
<td>456</td>
<td>Australia</td>
<td>15,638</td>
<td>282</td>
</tr>
<tr>
<td>Democratic Republic of Congo</td>
<td>11,000</td>
<td>413</td>
<td>Canada</td>
<td>2,920</td>
<td>163</td>
</tr>
<tr>
<td>Ecuador</td>
<td>17,600–21,100</td>
<td>271</td>
<td>Egypt</td>
<td>2,066</td>
<td>105</td>
</tr>
<tr>
<td>India</td>
<td>&gt;17,000</td>
<td>350</td>
<td>France</td>
<td>4,500</td>
<td>113</td>
</tr>
<tr>
<td>Indonesia</td>
<td>~37,000</td>
<td>513</td>
<td>Japan</td>
<td>4,700</td>
<td>186</td>
</tr>
<tr>
<td>Malaysia</td>
<td>15,000</td>
<td>286</td>
<td>Morocco</td>
<td>3,600</td>
<td>108</td>
</tr>
<tr>
<td>Mexico</td>
<td>18,000–30,000</td>
<td>450</td>
<td>South Africa</td>
<td>23,420</td>
<td>247</td>
</tr>
<tr>
<td>Peru</td>
<td>18,000–20,000</td>
<td>344</td>
<td>UK</td>
<td>1,550</td>
<td>77</td>
</tr>
<tr>
<td>Venezuela</td>
<td>15,000–21,070</td>
<td>288</td>
<td>USA</td>
<td>18,956</td>
<td>428</td>
</tr>
</tbody>
</table>
on each of the newly produced "provinces" (continents) produced many new taxa that were continental endemics and, indeed, much of the increase in species numbers of the Cenozoic era appears to be due to such continental isolation. As a consequence, several countries of South America (Brazil, Colombia), Africa (Democratic Republic of Congo), and Southeast Asia (Indonesia) are extremely rich in species, but they share few species in common because of their separation of many millions of years.

In general, areas that are geologically older have higher species richness than younger areas. Such geologically older areas have had more time to receive species dispersing from other regions of the world and more time for existing taxa to undergo adaptive speciation in response to the local conditions. The effect of geological age on species richness is well exemplified among marine ecosystems. The species richness of reeforming corals is several times greater in the Indian Ocean and western Pacific than in the Caribbean Sea and adjacent Atlantic Ocean, which are considerably younger than the former. A similar situation is the case of mangrove ecosystems, which typically contain a much higher number of arboreal mangrove species in the Old World than in the Neotropics. Marine species richness of countries associated with these older seas is therefore remarkably high, thus contributing to the existence of megadiverse countries in some of these areas. Australia is a clear example: its Great Barrier Reef contains about 8% of the world's fish species in just 0.1% of the global ocean surface area. The total diversity of fish species in Australia, about 4300, is equivalent to 20% of the world's total.

Regions rich in endemism tend to have experienced major events that caused fragmentation of the ranges of many species at approximately the same place. Such fragmentation events leading to geographic isolation include continental drift, mountain building, and sea level changes. Subsequent to their isolation, taxa may undergo evolutionary radiation in a given locality. For example, isolation and subsequent evolutionary radiation due to continental drift have been fundamental to the generation of a high degree of endemism in the biota of Australia, Papua New Guinea, and Madagascar. Mountain building that led to isolation, explosive radiation, and a high accumulation of endemism is well represented in countries such as Colombia, Peru, and Mexico. Changes in sea levels have been responsible for the remarkable species richness and concentration of endemism in Indonesia and also in Mexico.

Likewise, the remarkable diversity of fishes and other groups of organisms in large tropical lakes and rivers

C. Historical Factors

Historical factors are also critical in determining species richness and the concentration of endemisms. Over a long-term geological timescale, the continued rise in species richness is explained by the "provincialization" that accompanied the breakup of Pangaea and, later, Gondwanaland. Following their separation, evolution
is due to rapid evolutionary radiation in such isolated and usually productive ecosystems. Notable in this respect are countries such as Brazil, Colombia, and Peru.

As expected, islands, owing to their isolation, often have high proportions of endemic taxa. Australia, a continent-island of remarkable size, is one of the world's leaders in the endemism of mammals, birds, reptiles, and amphibians. Nevertheless, because islands frequently have a relatively depauperate biota, such high levels of endemism do not necessarily coincide with high richness of species, except on those islands where other driving factors are of exceptional importance, such as the large archipelago of Indonesia.

Another remarkable aspect related to biogeographic history is the occurrence of distinctive floristic and faunistic assemblages, which define the so-called biogeographic regions or provinces. Six of them have been identified on the basis of the geographic affinity as well as the presence and absence of particular taxa. Such biogeographic provinces reflect evolutionary histories and ecological affinities of plants and animals. Moreover, the hypothetical boundary lines between adjacent biogeographic regions constitute zones where biotas that evolved in different regions overlap with endemic taxa. The relevance of this qualitative aspect of biodiversity lies in the fact that a few countries of the world are located at the meeting ground of adjacent biogeographic provinces with such remarkable assemblages of species. In particular, three countries, Mexico, Indonesia, and China, each include in their territories a pair of the six major biogeographic regions of the world. The Neotropical and Neotropical regions meet in Mexico; the Oriental (or Indomalaysian) and Australoasiatic regions meet in Indonesia, and the Oriental and Palearctic meet in China. The case of the confluence of contrasting faunas of different origin in Indonesia is so striking that it stimulated the famous naturalist Alfred Wallace to develop the discipline of biogeography. Such confluence can be reflected in the fact that on a single tree can be seen monkeys, marsupials, cockatoos, and other birds. Several underlying factors of high biodiversity (see the following) coincide with this qualitative aspect in these three countries, but particularly in Mexico and Indonesia, which further increase their biological richness.

Another aspect of historical and biogeographic nature is the relative magnitude of the coastal/marine contour in relation to land area. Overall, when this ratio is high, some components of biodiversity, including ecological diversity and species richness, tend to be high. Though coastlines and the presence of seas are common to many countries, the ratio is particularly high in islands and some countries that are rich in biodiversity (Table II). Notable examples include Indonesia (an archipelago of several hundred islands), Australia, Madagascar, and Mexico. Table II shows that most of the megadiversity countries have coastline-to-area ratios that are relatively similar, with the exception of the Philippines and Indonesia, two countries constituted by archipelagoes, in which the ratios are particularly high. Mexico, in addition, includes extensive coastlines on two major oceans, the Pacific and Atlantic. Colombia is in a similar situation, although the ratio is lower. Mexico also has a sea of its own, the Sea of Cortez, which has resulted in significant levels of spe-

### Table II

<table>
<thead>
<tr>
<th>Country</th>
<th>Presence of coral reefs</th>
<th>Area (km²)</th>
<th>Coastline (km)</th>
<th>Ratio of coastline/area</th>
</tr>
</thead>
<tbody>
<tr>
<td>India</td>
<td>Yes</td>
<td>3,287,590</td>
<td>7,000</td>
<td>0.002</td>
</tr>
<tr>
<td>Philippines</td>
<td>No</td>
<td>300,000</td>
<td>36,289</td>
<td>0.120</td>
</tr>
<tr>
<td>Madagascar</td>
<td>Yes</td>
<td>587,041</td>
<td>4,628</td>
<td>0.008</td>
</tr>
<tr>
<td>Australia</td>
<td>Yes</td>
<td>7,741,220</td>
<td>25,760</td>
<td>0.003</td>
</tr>
<tr>
<td>Colombia</td>
<td>Yes</td>
<td>1,136,914</td>
<td>3,208</td>
<td>0.002</td>
</tr>
<tr>
<td>Brazil</td>
<td>Yes</td>
<td>8,511,965</td>
<td>7,491</td>
<td>0.0008</td>
</tr>
<tr>
<td>Mexico</td>
<td>Yes</td>
<td>1,958,201</td>
<td>11,592.77</td>
<td>0.005</td>
</tr>
<tr>
<td>Peru</td>
<td>No</td>
<td>1,285,213.6</td>
<td>2,414</td>
<td>0.001</td>
</tr>
<tr>
<td>Venezuela</td>
<td>Yes</td>
<td>916,445</td>
<td>2,800</td>
<td>0.003</td>
</tr>
<tr>
<td>Ecuador</td>
<td>Very small</td>
<td>275,800</td>
<td>2,237</td>
<td>0.008</td>
</tr>
<tr>
<td>China</td>
<td>No</td>
<td>9,596,561</td>
<td>14,500</td>
<td>0.001</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Yes</td>
<td>1,919,442</td>
<td>54,716</td>
<td>0.028</td>
</tr>
</tbody>
</table>
cies richness and endemism, of both marine and terrestrial species. This is why 51% of the marine mammal species of the world, and close to 80% of all sea turtle species, are present in Mexican coastal waters. It is worth noticing that no land-locked country, regardless of its size, is a significant megadiversity country.

D. Ecological Diversity

Ecological diversity is a significant driver of biodiversity, resulting from a variety of factors such as local variation in topography, soil conditions, altitudinal and latitudinal expanse, and overall climatic conditions, as well as historical and biogeographical backgrounds. Several of these factors have been touched upon already, thus only a synthesis and some additional points are necessary here.

Species richness and the generation of endemism tend to be greater where there is significant ecological heterogeneity that allows genetic isolation, local adaptation, and speciation to take place. Such situations can occur because of the existence of a series of isolated mountain peaks, large valleys or drainage systems that become isolated and separated into smaller systems, or areas that, being geologically heterogeneous and complex, produce a variety of soil conditions with well-defined boundaries between them. In addition to the promotion of speciation, these situations may provide the necessary ecological heterogeneity upon which a variety of communities can develop. This is reflected in two important aspects of the definition of megadiversity countries. The first is the occurrence of a rich diversity of vegetation types or ecosystems. The second is the occurrence of a high turnover of species between adjacent localities, leading to a spatially structured diversity of species, known as beta diversity.

Regarding the diversity of ecosystems and vegetation types, Colombia, Mexico, Indonesia, Peru, and India are remarkable in that their territories range from savannas (Colombia) and arid and semiarid ecosystems (the other countries) to evergreen tropical ecosystems, and from coastal vegetation types (e.g., mangroves) to alpine and sub-alpine ecosystems. Brazil, in addition to its great predominance of tropical rain forests, includes other tropical vegetation types such as semiarid cerrado and caatinga.

Regarding species turnover, available data on individual countries are extremely limited, largely because the rate at which the species composition of communities changes spatially depends on the range size of the species and the degree to which species are habitat specialists. However, it can be inferred from the diversity of vegetation types and the magnitude of environmental heterogeneity that the countries that are prolific in these attributes are likely to have high levels of beta diversity. One illustrative case is Mexico, where tropical dry forests expand along a geographic gradient from about 15° to 26°N latitude, in a variety of ecological conditions. At a given locality within this range, the number of arboreal species (with 1.0 cm or greater diameter at breast height) per 0.1 ha is high, with 75 species on average at 20 sites. However, the average proportion of species shared among sites is only 10%. This implies that the turnover of tree species is very high and that the same vegetation type is composed of a series of distinct assemblages of species.

E. Presence of Tropical Rain Forests

Several of the determining factors of species richness and endemism presented here are coincident for tropical rain forests and coral reefs (see Table II). In particular, the presence of tropical rain forests is a significant indicator of the potential for a country to be in the megadiversity group. The remarkable biodiversity of these ecosystems has prompted numerous researchers to investigate the underlying causes of their biodiversity, which include some additional and complementary factors to those described earlier. The most prominent are the following.

1. Climatic Stability

It has been argued that over geological time, tropical regions have had a more stable climate than temperate zones. Thus it is thought that the tropics might have high species richness because of the long time over which species could have accumulated, combined with a lower extinction rate in those stable environments. Such views are being challenged by studies (e.g., of pollen records) that suggest that tropical regions have not been so stable and that, moreover, marked climatic changes have affected the tropics and might have been responsible for the high species richness of some organisms.

2. Pest Pressures

Promoters of the idea that greater pressure from the high diversity of pests in the tropics contributes to species diversity argue that the lack of an unfavorable winter season may promote the persistence of natural enemies (e.g., herbivores and pathogens of plants). The increased pest pressure exerted by these ever-present natural enemies prevents a single, potentially competitively superior species, or group of species, from domi-
nating tropical communities, providing opportunities for numerous species to coexist at relatively low densities. Recent studies confirm that herbivores, for example, are crucial for the maintenance of species diversity in contemporary tropical forests. However, the evidence is limited regarding the role of pest pressure as a factor in generating species richness in these ecosystems.

3. Outcrossing Rates in Plants

The interbreeding with other (often numerous) individuals of the same species, as opposed to self-pollination, is thought to be greater in tropical ecosystems due to the abundance and diversity of tropical pollinators. In addition, the abundance of dioecious plant species (i.e., those having separate male and female individuals, and in which outcrossing is obligatory) is higher in tropical forests. The resulting higher rates of outcrossing may lead to higher levels of genetic variation and thus local adaptation and speciation.

IV. SELECTED MEGADIVERSITY COUNTRIES

The previous description gives some indication of the most likely nations to be considered megadiversity countries. A recent analysis by Mittermeier and Goettisch-Mittermeier (1997) examined 17 countries that collectively possess 66–75% of the planetary biodiversity (terrestrial, marine, and freshwater; Table III). However, not all of these countries comply with the requisites and facets of biodiversity as discussed here. For example, this list includes countries that, although prominent in one or more of the biodiversity facets considered, lack other facets. Some of the selected countries lack tropical rain forest ecosystems, or have a low diversity of ecosystems or cultures, or have relatively low species richness in some groups.

A more synthetic analysis by the same authors considered a shorter list of countries based on a hierarchical scoring of the five most prominent values (scored from 1 to 5) for species richness and degree of endemism of plants, mammals, birds, reptiles, and amphibians. Additionally, even though information is much more limited, consideration was also given to freshwater fishes, butterflies, and tiger beetles (family Cicindelidae). Thus, a country that had the highest value of any of these components was assigned a score of 5, the second highest was given the score of 4, and so on. The total sum of these scores gave a total value that could then be used to compare the different countries.

For example, Brazil had the highest values for four components: species richness of plants, mammals, and freshwater fishes, and plant endemism (i.e., $5 + 5 + 5 + 5 = 20$; see Table III). It also had three scores of 4 (species richness of amphibians and butterflies and endemism of amphibians, i.e., 12), four scores of 3 (species richness of birds and tiger beetles, and bird and tiger beetle endemism, i.e., 12), one score of 2 (bird endemism), and two scores of 1 (species richness and endemism of reptiles, i.e., 2). The total sum for Brazil is therefore 48. The corresponding values for the most diverse countries are shown in Table IV.

This synthetic analysis confirms the suggestions of the previous sections and can be summarized as follows. Brazil is the most prominent megadiversity country, followed by Indonesia and Colombia. At a considerable distance, Australia and Mexico follow in the fourth and fifth positions, respectively. The relative positions of these remarkable countries change depending on the facet of biodiversity that is considered. On the basis of species richness, the order is Brazil > Colombia > Indonesia > Mexico > Australia. If the emphasis is on endemism, the order is Indonesia > Brazil > Australia, and Madagascar supersedes Colombia > Mexico. The insular character and other attributes of Madagascar highlight its outstanding endemism and place this country in the overall sixth place, close to Mexico. The other countries in Table IV are notable for their species richness (Peru, seventh position; China, eighth position; Ecuador, eleventh position; and Venezuela, twelfth position) or their degree of endemism (Philippines, ninth position). India is remarkable in both species richness and endemism, although it ranks in the tenth position.

All seven of the most prominent megadiversity countries share the following common biological attributes. (1) They have tropical rain forest ecosystems within their territories. (2) They also have marine ecosystems and, to a varying degree, a high coast-to-land ratio (see Table II). (3) Their diversity of ecosystems is considerable and therefore their beta diversity is expected to be high. (4) From a sociological perspective, they share the characteristics of having a very rich diversity of cultures and, with the exception of Australia, of being developing countries and having considerable levels of threat to their biological diversity. Moreover, most of the megadiversity countries listed in Table IV share most of these characteristics. In particular, recent information on species of plants, birds, mammals, amphibians, and reptiles under several categories of extinction risk suggest that no less than 50% of the seriously threatened biological diversity at the global level is concentrated in the megadiversity countries.
**TABLE III**

Species Richness and Endemic Species (in Parentheses) of Plants and Five Groups of Vertebrates in Seventeen Megadiversity Countries*

<table>
<thead>
<tr>
<th>Country</th>
<th>Plants</th>
<th>Mammals</th>
<th>Birds</th>
<th>Reptiles</th>
<th>Amphibians</th>
<th>Freshwater fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>~50,000–56,000 (~16,500–18,500)</td>
<td>524</td>
<td>1622</td>
<td>468</td>
<td>517</td>
<td>&gt;3000</td>
</tr>
<tr>
<td>Indonesia</td>
<td>~37,000</td>
<td>515</td>
<td>1531</td>
<td>511</td>
<td>270</td>
<td>1400</td>
</tr>
<tr>
<td>(4.800–17,000)</td>
<td>(201)</td>
<td>(397)</td>
<td>(150)</td>
<td>(399)</td>
<td>(100)</td>
<td>(367)</td>
</tr>
<tr>
<td>Colombia</td>
<td>45,000–51,000</td>
<td>456</td>
<td>1815</td>
<td>520</td>
<td>583</td>
<td>&gt;1500</td>
</tr>
<tr>
<td>(10,000–15,000)</td>
<td>(28)</td>
<td>(125)</td>
<td>(133)</td>
<td>(368)</td>
<td>(169)</td>
<td>(468)</td>
</tr>
<tr>
<td>Mexico</td>
<td>18,000–30,000</td>
<td>430</td>
<td>1050</td>
<td>717</td>
<td>284</td>
<td>183</td>
</tr>
<tr>
<td>(10,000–15,000)</td>
<td>(140)</td>
<td>(125)</td>
<td>(133)</td>
<td>(169)</td>
<td>(169)</td>
<td>(169)</td>
</tr>
<tr>
<td>Australia</td>
<td>15,638</td>
<td>282</td>
<td>751</td>
<td>755</td>
<td>196</td>
<td>183</td>
</tr>
<tr>
<td>(11,458)</td>
<td>(210)</td>
<td>(355)</td>
<td>(616)</td>
<td>(169)</td>
<td>(169)</td>
<td>(169)</td>
</tr>
<tr>
<td>Madagascar</td>
<td>11,000–12,000</td>
<td>105</td>
<td>253</td>
<td>300</td>
<td>178</td>
<td>75</td>
</tr>
<tr>
<td>(8,800–9,600)</td>
<td>(77)</td>
<td>(103)</td>
<td>(274)</td>
<td>(170)</td>
<td>(170)</td>
<td>(170)</td>
</tr>
<tr>
<td>China</td>
<td>27,100–30,000</td>
<td>499</td>
<td>1244</td>
<td>367</td>
<td>274</td>
<td>1010</td>
</tr>
<tr>
<td>(~10,000)</td>
<td>(77)</td>
<td>(99)</td>
<td>(133)</td>
<td>(175)</td>
<td>(175)</td>
<td>(175)</td>
</tr>
<tr>
<td>Philippines</td>
<td>5000–12,000</td>
<td>201</td>
<td>556</td>
<td>193</td>
<td>63</td>
<td>330</td>
</tr>
<tr>
<td>(3,800–6,600)</td>
<td>(116)</td>
<td>(183)</td>
<td>(131)</td>
<td>(44)</td>
<td>(44)</td>
<td>(44)</td>
</tr>
<tr>
<td>India</td>
<td>&gt;17,000</td>
<td>350</td>
<td>1298</td>
<td>408</td>
<td>206</td>
<td>750</td>
</tr>
<tr>
<td>(&gt;7025–17,875)</td>
<td>(44)</td>
<td>(52)</td>
<td>(187)</td>
<td>(110)</td>
<td>(110)</td>
<td>(110)</td>
</tr>
<tr>
<td>Peru</td>
<td>18,000–20,000</td>
<td>344</td>
<td>1703</td>
<td>298</td>
<td>241</td>
<td>855</td>
</tr>
<tr>
<td>(3,300)</td>
<td>(46)</td>
<td>(109)</td>
<td>(98)</td>
<td>(~80)</td>
<td>(~80)</td>
<td>(~80)</td>
</tr>
<tr>
<td>Papua New Guinea</td>
<td>15,000–21,000</td>
<td>242</td>
<td>772</td>
<td>305</td>
<td>200</td>
<td>282</td>
</tr>
<tr>
<td>(10,000–16,000)</td>
<td>(57)</td>
<td>(85)</td>
<td>(79)</td>
<td>(134)</td>
<td>(134)</td>
<td>(134)</td>
</tr>
<tr>
<td>Ecuador</td>
<td>17,600–21,100</td>
<td>271</td>
<td>1599</td>
<td>374</td>
<td>402</td>
<td>&gt;44</td>
</tr>
<tr>
<td>(4,000–5,000)</td>
<td>(21)</td>
<td>(37)</td>
<td>(114)</td>
<td>(138)</td>
<td>(138)</td>
<td>(138)</td>
</tr>
<tr>
<td>USA</td>
<td>18,956</td>
<td>428</td>
<td>768</td>
<td>261</td>
<td>194</td>
<td>790</td>
</tr>
<tr>
<td>(4,036)</td>
<td>(101)</td>
<td>(71)</td>
<td>(90)</td>
<td>(126)</td>
<td>(126)</td>
<td>(126)</td>
</tr>
<tr>
<td>Venezuela</td>
<td>15,000–21,070</td>
<td>288</td>
<td>1360</td>
<td>293</td>
<td>204</td>
<td>1250</td>
</tr>
<tr>
<td>(3,000–8,000)</td>
<td>(11)</td>
<td>(45)</td>
<td>(57)</td>
<td>(75)</td>
<td>(75)</td>
<td>(75)</td>
</tr>
<tr>
<td>Malaysia</td>
<td>15,000</td>
<td>286</td>
<td>738</td>
<td>268</td>
<td>198</td>
<td>600</td>
</tr>
<tr>
<td>(6,500–8,000)</td>
<td>(27)</td>
<td>(11)</td>
<td>(68)</td>
<td>(57)</td>
<td>(57)</td>
<td>(57)</td>
</tr>
<tr>
<td>South Africa</td>
<td>23,420</td>
<td>247</td>
<td>774</td>
<td>299</td>
<td>95</td>
<td>153</td>
</tr>
<tr>
<td>(16,500)</td>
<td>(27)</td>
<td>(7)</td>
<td>(76)</td>
<td>(36)</td>
<td>(36)</td>
<td>(36)</td>
</tr>
<tr>
<td>Democratic Republic of Congo</td>
<td>11,000</td>
<td>415</td>
<td>1094</td>
<td>268</td>
<td>80</td>
<td>962</td>
</tr>
<tr>
<td>(3,200)</td>
<td>(38)</td>
<td>(23)</td>
<td>(33)</td>
<td>(53)</td>
<td>(53)</td>
<td>(53)</td>
</tr>
</tbody>
</table>

* Data from Mittermeier et al. (1997).

---

V. THREATS TO BIODIVERSITY IN MEGADIVERSE COUNTRIES

Megadiversity countries are experiencing significant alterations to their natural ecosystems, and this degradation is a serious threat to their biological resources. The main threats to biodiversity in these countries are the current patterns of land use that lead to deforestation and habitat fragmentation, and species overexploitation. These threats are summarized in the following sections.

A. Deforestation and Habitat Fragmentation

A commonly used measure of habitat destruction is the rate of deforestation. Table V presents compiled statistics on deforestation rates for the period 1980–
1989. Rates of deforestation for the 12 megadiversity countries range from 0.4% (Peru) to 2.3% (India and Ecuador) per year. Only four countries have deforestation rates below 1% per year. Even though these figures may seem low, for most countries they imply a significant loss of absolute area per year.

In Mexico, absolute rates of deforestation are on the order of 700,000 ha per year, with by far the greatest amount being tropical wet and dry forests, where most of the biological diversity of the country is concentrated. Moreover, this type of aggregated statistic conceals the fact that particular areas have considerably higher rates of deforestation. Deforestation rates in the tropical forests of southern Veracruz reach values of 4.3% per year. Should such rates continue, tropical forests of the region would be reduced to less than 9% of their original coverage in the first decades of the next century. Such drastic reductions of forest area may drive a significant proportion (about 50%) of the native species to local extinction, according to models derived for the relationship between species and land area on islands. In addition, rates of deforestation do not reflect the fact that the remaining areas of habitat frequently are left in the form of an archipelago of forest islands immersed in a matrix of transformed habitat. Forest fragmentation has been recognized as a significant factor leading to the local extinction of species, reduction of genetic variability, and disruption of several ecosystem processes.

The present rate of worldwide deforestation is close to 50,000 km² per year. The loss of species due to habitat destruction may be enormous if this rate is maintained over the next few decades. Botanist Peter Raven estimates that an average of 50,000 species may be lost per year during the next several decades, of which only some 7000 will have been recognized and named.

### Table IV

The Ranking of the Twelve Countries with Highest Megadiversity Based on Scores of Species Richness and Endemism (See Text for Details)

<table>
<thead>
<tr>
<th>Country</th>
<th>Species richness</th>
<th>Endemism</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>30</td>
<td>18</td>
<td>48</td>
</tr>
<tr>
<td>Indonesia</td>
<td>18</td>
<td>22</td>
<td>40</td>
</tr>
<tr>
<td>Colombia</td>
<td>26</td>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>Australia</td>
<td>5</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>Mexico</td>
<td>8</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>Madagascar</td>
<td>2</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>Peru</td>
<td>9</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>China</td>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Philippines</td>
<td>0</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>India</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Ecuador</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Venezuela</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

*From Mittermeier et al. (1997).*

### Table V

Demographic, Social, Physical, and Resource Utilization Parameters in Megadiverse Countries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Brasil</td>
<td>1714.8</td>
<td>1.6</td>
<td>23</td>
<td>851.2</td>
<td>57.3</td>
<td>4.9</td>
<td>6.7</td>
<td>194.27</td>
<td>0.7</td>
</tr>
<tr>
<td>Indonesia</td>
<td>2127.7</td>
<td>1.5</td>
<td>67</td>
<td>190.5</td>
<td>58.7</td>
<td>9.9</td>
<td>24.3</td>
<td>149.06</td>
<td>0.8</td>
</tr>
<tr>
<td>Colombia</td>
<td>37.8</td>
<td>1.5</td>
<td>28</td>
<td>113.9</td>
<td>43.9</td>
<td>3.4</td>
<td>13.5</td>
<td>17.22</td>
<td>1.7</td>
</tr>
<tr>
<td>Australia</td>
<td>19.2</td>
<td>1.3</td>
<td>—</td>
<td>771.3</td>
<td>18.8</td>
<td>6</td>
<td>4.6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Mexico</td>
<td>102.4</td>
<td>1.9</td>
<td>26</td>
<td>195.8</td>
<td>24.9</td>
<td>11.8</td>
<td>26.3</td>
<td>3.2</td>
<td>1.3</td>
</tr>
<tr>
<td>Ecuador</td>
<td>12.6</td>
<td>2.4</td>
<td>43</td>
<td>28.3</td>
<td>35</td>
<td>5.7</td>
<td>34.1</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Madagascar</td>
<td>17.3</td>
<td>3.2</td>
<td>74</td>
<td>58.7</td>
<td>39.5</td>
<td>4.4</td>
<td>42.1</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Peru</td>
<td>26.1</td>
<td>1.9</td>
<td>29</td>
<td>128.5</td>
<td>66</td>
<td>2.6</td>
<td>37.6</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Venezuela</td>
<td>24.2</td>
<td>2.1</td>
<td>8</td>
<td>91.2</td>
<td>32.9</td>
<td>3.3</td>
<td>5.9</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>China</td>
<td>1389.60</td>
<td>1.7</td>
<td>71</td>
<td>956.1</td>
<td>13.5</td>
<td>9.6</td>
<td>53.6</td>
<td>200.6</td>
<td>—</td>
</tr>
<tr>
<td>India</td>
<td>1022</td>
<td>1.8</td>
<td>74</td>
<td>528.75</td>
<td>20.8</td>
<td>50.5</td>
<td>28.9</td>
<td>262.8</td>
<td>2.3</td>
</tr>
<tr>
<td>Philippines</td>
<td>74.6</td>
<td>2.4</td>
<td>48</td>
<td>30</td>
<td>45.3</td>
<td>18.4</td>
<td>28.6</td>
<td>35.98</td>
<td>1.5</td>
</tr>
</tbody>
</table>

B. Overexploitation of Species

The overexploitation of species is much more difficult to document and no reliable, compiled statistics are available. Overexploitation of the majority of targeted marine fishes is very well documented, but this generally occurs in the "commons" of international waters. However, some isolated reports indicate that overexploitation is a significant threat to biodiversity in megadiverse countries. For example, in the Brazilian Amazon, studies by zoologist Kent Redford indicate that hunting of vertebrates (mammals and birds) can kill on the order of more than 14 million animals per year. In summary, the great biological richness stored in the megadiverse countries is seriously threatened and urgent measures are needed to deal with this problem, which is occurring on a global scale.

VI. SOCIO-DEMOGRAPHIC CHARACTERISTICS AND THEIR RELATION TO BIODIVERSITY

Of all the issues related to biological diversity, there are very few whose study makes any sense outside of the context of its relation with human populations. Indeed, human populations—their very existence, cultural evolution, size, and distributions—have been and still are intricately dependent and interactive with the biological diversity available to them in the regions where they live. Nowhere are these relations between population, cultures, and biologic diversity more evident than in the majority of the so-called "Third World" or "less-developed" nations. Most megadiverse countries share a number of human, cultural, social, and economic traits. These shared characteristics (total human population and growth rates, rural population and its distribution, cultural diversity, land ownership, and levels of scientific development) are discussed here.

Coincidentally, it is in these less-developed regions of the world where the richest biological diversity is present, and where most of the ancient and diverse human cultures and civilizations emerged. These areas presently support the greatest human cultural diversity, represented by the presence of large groups of indigenous peoples. In many megadiverse countries, indigenous groups comprise a high percentage of the total population, commonly ranging from 10% up to nearly 80% in the case of Papua New Guinea.

A measure of the cultural diversity of a country is the number of languages spoken by its people. A total of 6700 languages are spoken in the world, and over half are found in the 15 most biologically diverse countries (e.g., Papua New Guinea alone has over 800 languages). This relationship between biological and cultural diversity is explored by Victor Toledo elsewhere in this Encyclopedia. Usually such a relationship has resulted in cultural traits and land use practices that have increased biological diversity, mostly through the "invention" of new, cultivated species that have allowed the further development of those cultures and others around the world. As illustrated in Fig. 3, countries of both high biological and cultural diversity have usually been centers of origin of crops as well as of agricultural technologies.

A. Rural Populations and Territorial Distribution

Because of the presence of their many different cultural groups, and in stark contrast with most economically developed regions of the world, developing, biodiversity-rich countries have a large proportion of their total population living in rural areas. These groups are highly dependent on natural resources and are usually deeply knowledgeable of their uses and ecological traits. Additionally, in many cases a large proportion of these people live in small villages or hamlets, most of which are distributed within the high biodiversity regions in each country. Because of their close relationship with the surrounding environment and resources, most of these communities are intimately dependent on the conditions of the ecosystems in which they live for their quality of life and the maintenance of their well-being.

B. Land Ownership by Indigenous Groups

The two traits of large populations and predominantly rural living result in patterns of land ownership that may seem "complex" and inefficient to people from developed nations. These patterns are normally the result of well-established traditions and the forms of social organization and consensus building practiced in the communities. Communal land ownership is common, and it can take a variety of forms in different megadiverse countries. Another form of ownership is that of very small landholdings; this system is found in countries like India and China, where its effects are either positive or negative, for the rational use and conservation of biodiversity depending on management practices.

In most cases, indigenous groups own or control
extensive regions of natural forests and other ecosystems in the megadiverse countries. For example, the indigenous groups in the Amazon basin possess over 100 million ha of forested land. In Papua New Guinea, the many cultural groups that live there own close to 97% of the national territory. In Mexico, nearly 75% of all the areas listed as priorities for conservation are under communal ownership, mostly by indigenous or rural groups.

There is an abundant literature describing, with different levels of academic rigor, the benefits and disadvantages of communal land ownership in relation to the rational management and conservation of natural resources, particularly in biologically rich countries. After reviewing much of this literature and visiting the sites and talking to people, it is apparent to the authors that when a strong and healthy system of internal governance exists that allows participation of all interested parties, the community uses and conserves its resources in an admirable fashion. When those internal structures of governance break down for whatever reason, the communal system of land ownership is as destructive of its resources as any other system.

Because of the often large indigenous population in an area, their widespread territorial distribution, and the patterns of land ownership, any program of biodiversity use and conservation in most megadiverse countries must take these stakeholders into account. Such inclusion should not be undertaken as an undesirable and unavoidable necessity, but rather as a wise and intelligent approach that involves local people who are essential to any conservation program. In addition to being the owners of the areas whose resources the conservationists and governments wish to conserve, these indigenous groups normally possess an enormous wealth of valuable and ecologically informed knowledge. Such information is often the result of a long tradition of oral culture that has been passed from one generation to the next, and that makes them truly valuable partners in conservation, management, or restoration programs.

Although a great deal of traditional ecological knowl-
edge exists among the indigenous groups of megadiverse countries, the long-term sustainable use of many of those natural resources remains very uncertain and difficult. This problem is caused by a number of factors, including rising pressure from growing populations to acquire land, increased demands for food production, and the actions of economic interests, often of foreign origin, that are likely to degrade or destroy natural ecosystems and their resident biological diversity.

C. Levels of Scientific Development

Another characteristic that is common to the majority of biodiversity-rich countries (with the sole exception of Australia) is that they have a relatively incipient scientific infrastructure, which in many instances is reflected in the very small number of scientists (taxonomists, botanists, ecologists, etc.) that study their natural resources.

It is arguable to what extent this is the result of the history of colonization—another important characteristic shared by all megadiverse countries. All of these countries have been colonized by a European country for diverse lengths of time, from a few decades to several centuries. In these cases no significant effort was made by the European countries to promote the technical and scientific development of the colonies. Additionally, the subsequent development of these countries was characterized by a neglect of the traditional knowledge on natural resource management of their ethnic communities.

Very often the most important collections of flora and fauna of these countries are found in institutions, herbaria, and museums located in developed countries. This can play a significant role in limiting scientific training and consequently the national capacity to assess and manage natural resources. As described earlier, a great proportion of the world's terrestrial biodiversity is located in tropical forests in countries that lack the necessary scientific and managerial infrastructure. We believe that this is a major issue that must be addressed when considering the future of the conservation of the world's biological diversity. A detailed discussion of this point is presented in the remainder of this article.

D. A Case Study of Demographics, Cultural Diversity, and Rural Population: Mexico

The following are a few significant statistics for Mexico, but similar patterns are shared by many of the megadiverse countries discussed earlier. These carry high significance in influencing modes of sustainable resource utilization. At present around 6 million people (5 years of age and over) in Mexico have a first language that is not Spanish. This figure has increased between 1930 and 1995 from 2.3 million to 5.6 million persons, although the percentage they represent of the total population has decreased for the same period, from 16% to 6%. They represent at least 54 different ethnic groups, and speak nearly 230 languages. The rural population in Mexico (regardless of language spoken) constitutes about 24% of the total, a considerably lower figure than that of 71% at the beginning of the century; the greatest reduction occurred in the 1950s.

About 90% of all human settlements in Mexico (close to 200,000 registered in the 1995 census) have 1000 or fewer inhabitants (Fig. 4). This proportion is similar to or even higher than that of India, where 73% of all villages have less than 1000 inhabitants. The picture that emerges is clear: the rural population of Mexico is largely atomized, widely dispersed and organized in small villages, with different levels of community organization and various cultural identities. The smallest of these communities are often isolated from most social services and tend to maintain their traditional cultural and social organization traits. Many may succeed in maintaining these traits despite the efforts of external authorities or other sources to exert influence on their internal social organization schemes, or in their modes of production.

Regarding levels of scientific development, Mexico shares the general condition often found in less developed countries of having a limited scientific infrastructure and little natural resource management experience. However, for different historical reasons, efforts to advance the knowledge of the flora and fauna of Mexico go back to the middle of last century, when plant and animal collections were made, museums were organized, and a limited amount of institutional support was established. With time and the development of public institutions of higher education, many herbaria and museums cropped up in different regions of the country. This brought about the current situation in which taxonomic research, the botanical and zoological survey of the country, and other activities related to an understanding of biodiversity are fairly evenly distributed in many institutions across Mexico. This contrasts with the relatively heavy centralization in the metropolitan area of Mexico City of serious academic activities in other disciplines. The consequences of different levels of scientific development in megadiverse countries are discussed in the following sections when we de-
scribe the means by which these countries can develop important databases of information about their own biological diversity.

VII. HUMAN AND INSTITUTIONAL CAPACITY FOR BIODIVERSITY KNOWLEDGE AND CONSERVATION

With a few exceptions, a shared characteristic of megadiverse countries is that a very large proportion of the scientific collections that represent their rich biological diversity is deposited in institutions located in diversity-poor but economically rich nations. Most of the vast collections of the Earth’s biota (some 3 billion specimens, excluding microorganisms) are located in the largest natural history museums and herbaria of the world. These collections consist largely of specimens collected within biologically rich countries during many decades, or even centuries, by expeditions supported by the institutions of developed countries. Incidentally, perhaps no more than 1% of that vast pool of information can be accessed in automated fashion.

Many of these collections represent a heritage of very significant historical value, since they contain biological information of regions that have been seriously altered and that presently have lost much of their original floras and faunas. These natural history museums constitute a vast investment of effort and resources and are an invaluable repository of information.

Much of the reason why these collections are not located in their “resident” countries is because only about 6% of the world’s scientists live in the countries that house 80% of the planet’s biodiversity. Different estimations coincide in calculating the number of taxonomists in the world to be about 7000. If a similar ratio holds for taxonomists, this means that the megadiverse countries have fewer than 500 taxonomists. This would clearly be an underestimation, since the authors estimate that there must be at least twice as many taxonomists in the megadiverse countries. However, even with this higher figure, this represents a painfully small number of specialists who are unable to adequately study an enormously rich biota.

Comprehensive and scientifically sound biological collections, together with the necessary taxonomic expertise that must go hand-in-hand with the collections,
are a fundamental need to advance the ability of a country to know, understand, manage, and preserve its biological heritage. This was recognized at a high official level by the 175 signatories of the Convention on Biodiversity (CBD) during their fourth Conference of the Parties (COP) in Bratislava, Czech Republic, in May of 1998, who concluded that there exists what they called the "taxonomic impediment." This "impediment" results from the severe limitations of institutional and human infrastructure that affect the majority of megadiverse countries, which also happen to be the less economically and scientifically developed. These limitations are illustrated by the very small or nonexistent number of taxonomists, systematists, ecologists, and other specialists trained at a high level, and the absence or very limited number of zoological museums and herbaria that have reliable and long-term institutional support.

Table VI is a compilation of the herbaria (with the number of specimens they contain), botanical gardens (with the number of taxa reported under cultivation), and zoological collections in the 12 most biologically diverse countries. It is clear that the majority have a serious shortage of scientific collections of their floras and faunas, with the notable exception of Australia and to a lesser extent India, China, and Indonesia in Asia and Brazil, Mexico, and Colombia in the Americas.

It is worth noting that for developed countries of the temperate regions the situation is dramatically different. The United Kingdom, France, and the United States, for example, have between 10 and 100 times more herbaria and specimens housed in them than any of the 12 megadiverse countries. The comparison is almost the same for zoological collections and botanical gardens. Obviously, only a fraction of those specimens belongs to their national floras and faunas, which makes them, together with a few other European countries like the Netherlands and Germany, the most important repositories of the biological information of the majority of megadiverse countries.

At the Bratislava COP meeting, an agreement was reached on the imperative need to remove this "impediment" with the aim of fulfilling the objectives set by the CBD within a shorter period of time. However, the training of personnel at a high level, as is required for specialist taxonomists and systematists, and the extensive development of scientific collections (which in turn requires long-term institutional support) are complex, lengthy, and expensive processes. Most of the largest of such institutions in the world are the result of a tradition of many decades, and even centuries, of sustained institutional effort and economic support. This does not mean that countries that do not have, or that need to develop more, such human and physical infrastructures should abandon any attempts to embark themselves in this process. Yet it seems to be far more effective to simultaneously dedicate efforts to develop a strong capacity to acquire, systematize, interpret, and utilize the biological information already existing in the many local and foreign museums and herbaria. Examples of such efforts, although not abundant, demonstrate different strategies in how to bridge the gap of

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of herbaria</th>
<th>Accessions in herbaria ($10^6$)</th>
<th>Number of botanical gardens</th>
<th>Number of taxa in botanical gardens ($10^6$)</th>
<th>Number of zoological collections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>88</td>
<td>3.2</td>
<td>23</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>Indonesia</td>
<td>6</td>
<td>1.67</td>
<td>5</td>
<td>70</td>
<td>3</td>
</tr>
<tr>
<td>Colombia</td>
<td>23</td>
<td>0.65</td>
<td>13</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Australia</td>
<td>38</td>
<td>5.32</td>
<td>63</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>Mexico</td>
<td>52</td>
<td>2.3</td>
<td>35</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>Madagascar</td>
<td>2</td>
<td>0.08</td>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Peru</td>
<td>11</td>
<td>0.45</td>
<td>6</td>
<td>—</td>
<td>3</td>
</tr>
<tr>
<td>China</td>
<td>306</td>
<td>15.5</td>
<td>68</td>
<td>27</td>
<td>7</td>
</tr>
<tr>
<td>Philippines</td>
<td>9</td>
<td>0.29</td>
<td>9</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>India</td>
<td>52</td>
<td>3.7</td>
<td>68</td>
<td>85</td>
<td>27</td>
</tr>
<tr>
<td>Ecuador</td>
<td>10</td>
<td>0.26</td>
<td>2</td>
<td>0.5</td>
<td>3</td>
</tr>
<tr>
<td>Venezuela</td>
<td>15</td>
<td>0.63</td>
<td>7</td>
<td>1</td>
<td>15</td>
</tr>
</tbody>
</table>

knowledge about the biological diversity of a country. Some of these examples are presented in Section VIII.

VIII. CASE STUDIES OF MANAGING BIOLOGICAL INFORMATION IN MEGADIVERSE COUNTRIES

Most of the examples of development at the national level of the capacity to acquire, systematize, and utilize information about the biota of a country occur, paradoxically, in biodiversity-rich countries. As we have mentioned, by far the majority of the scientific information about the flora, fauna, and microorganisms that live in a region or country is housed in the systematic collections of plants and animals. Such information is complemented to a certain degree by works such as written floras and faunas, updated checklists of organisms, and detailed ecological accounts of groups of plants or animals. However, only the scientific collections contain the wealth of information that covers the historical dimension of which organisms used to live in what now are profoundly disturbed areas. A second dimension, the geographical, is also characteristic of scientific collections, found in the representation of specimens that records the spatial and ecological distribution of a species and its morphological and genetic variance within its range.

Despite the very considerable wealth of information contained in most scientific collections, the access to such information has been essentially restricted to the scientists (taxonomists and systematists) who work on and take care of those collections. However, there has been an intense effort in putting together different catalogs and directories for biodiversity information in several developed countries. These mostly take the form of accounts of collections for diverse organisms, catalogs of described species, lists of endangered species, bibliographic sources, or environmental studies.

A. The Comisión Nacional Para el Estudio y Uso de la Biodiversidad (CONABIO), Mexico

CONABIO was created in March 1992 as an interministerial organization with the mission of coordinating the actions and studies related to the knowledge and preservation of biological diversity, as well as promoting and stimulating scientific research activities for the exploration, study, protection, and utilization of biological resources. Its aims are to preserve Mexico’s ecosystems and to develop criteria for their sustainable management. Although CONABIO is placed within the Ministry of the Environment, Natural Resources and Fisheries (SEMARNAF) and receives funding from federal sources, it operates as a private trust fund (Fideicomiso Fondo para la Biodiversidad). This trust fund allows CONABIO to obtain, in addition to federal funds, economic support from contracts, donations, and services, which helps it to run a very efficient operation. The main functions of CONABIO are to:

1. Establish a national program for biological inventories.
2. Synthesize the information relating to biological resources in a permanently updated database.
3. Design and implement a National Biodiversity Information System.
4. Promote projects directed to the actual and potential use of biological resources that are both conventional and nonconventional.
5. Provide advice on technical and scientific aspects related to conservation and management of natural resources to government agencies, nongovernmental organizations, and industry.
6. Produce education materials for the general public, especially for younger people, and provide information about biological diversity and natural resources to all components of Mexican society.

In less than eight years, CONABIO has become the central Mexican institution that compiles data and generates information on biodiversity and natural resources, and in so doing it has been able to bridge the gap between scientific research efforts and the production of useful information that is relevant to policy making in the conservation of natural resources. One of its fundamental activities is the construction of a biodiversity database with the participation of most of the taxonomic community of Mexico. Currently, these data are housed in various institutions, as well as in the foreign herbaria and museums that hold large collections of Mexican plants and animals.

CONABIO does not conduct inventories or fieldwork on its own. Rather, it uses the biodiversity information already stored in national and foreign scientific institutions and funds research carried out by Mexican scientists in the areas that have been identified as priorities in its mandate. An extensive program for the repatriation of data on Mexican species housed in foreign institutions has begun and will continue over the coming years. The biodiversity database is constructed on the
basis of the information contained in the label of each specimen included, which is taxonomically validated by experts in the different groups of organisms and then precisely geo-referenced. At present the database includes information on over 2.5 million specimens, and is constructed into a geographic information system (GIS) containing a wide array of data on climate, geology, topography, soils, vegetation types, land use patterns, and other factors. The CONABIO home page is www.conabio.gob.mx.

B. The Environmental Resources Information Network (ERIN), Australia

ERIN is the oldest of all national entities devoted to compiling the biodiversity information of a country and building a database with that information. It has been a model and inspiration for other such systems around the world. The Australian government established it in 1989 as part of the Federal Department of the Environment. As is the case with CONABIO, ERIN's aim is to provide biodiversity and environmental information to assist planning and decision making in the fields of conservation and natural resource management. It also draws information from many sources and institutions and its database is constructed on the basis of specimen information and compiled into a very rich GIS. ERIN is the main source for an array of fairly sophisticated data analyses about the distribution, conservation, and management of the Australian biota. It has become an increasingly powerful analysis tool for many studies of the biogeography and ecology of the country's plant and animal species and the assemblages and ecosystems that they constitute. The ERIN home page is www.erin.gov.au.

C. The Instituto Nacional de Biodiversidad (INBio), Costa Rica

Started in October 1989, INBio has pioneered a substantial effort to survey the biodiversity of Costa Rica. It is a private, nonprofit organization, entitled to receive both governmental and private funds for its operation. Its Board of Directors and the General Assembly include representatives from different sectors of Costa Rican society who are stakeholders in the efforts to conserve the natural resources in the country.

One of INBio's central aims is to carry out a basic national biodiversity inventory. To this end, it has developed an active program of biological surveying of the major groups of organisms and is building scientific collections of the different groups. This effort has involved the deployment of a substantial number of field assistants (mostly laymen and peasants) who are qualified as "parataxonomists" and who do most of the collecting of specimens and the initial taxonomic identification. INBio currently maintains a biodiversity database that provides geo-referenced information (by latitude–longitude), particularly of the specimens that have been collected since it started operating. Another of its stated aims is to support biodiversity conservation and promote new opportunities for sustainable development within the social and economic contexts of Costa Rica.

A third area of INBio activity has been the prospecting for natural products in plants and animals of Costa Rica, which has been an especially active component within its various programs. The organization has a close collaboration with different governmental agencies in charge of the conservation and management of natural resources, and it has served as a constant source of advice and information to them in their programs. INBio's home page is www.inbio.ac.cr.

D. The South African Botanical Diversity Network (SABONET)

The South African Botanical Diversity Network was started in the early 1990s as a program closely linked to the National Botanical Institute (NBI) of South Africa. Its mission is to strengthen the level of botanical expertise, expand and improve herbaria and their collections, and stimulate closer collaborative links among botanists of the 10 countries that form the South African subcontinent (Angola, Botswana, Lesotho, Malawi, Mozambique, Namibia, South Africa, Swaziland, Zambia, and Zimbabwe).

A main objective of SABONET has been the training and support of professional botanists, taxonomists, horticulturists, and plant diversity specialists in these countries. A survey of all scientists working in the fields of botany in the region shows that there are nearly 250 botanists, of which a little over 100 work in institutions outside the area, mainly in Europe. Such personnel should be competent in inventorying, monitoring, and evaluating the botanical diversity of the region in the context of conservation needs and the threats to natural systems imposed by different human activities. To facilitate this, SABONET offers training courses and workshops, organizes botanical expeditions to poorly known areas of the region, and publishes a newsletter. SABONET also publishes a catalog that contains information about all research projects carried out in the region,
the botanists in charge of them, a taxonomic index of who is working in the different plant taxa, and a geographical index listing the areas where botanists are working. General inquiries can be addressed to ckw@nbipre.nbiac.za.

IX. CONCLUSIONS

The world's biological riches are unevenly distributed across the continents, and this distribution coincides with the same areas where many of our most ancient and diverse cultures developed. These areas of high biological diversity now face threats from the activities and growing demand for natural resources of today's much larger native populations, but also from demands that originate in distant developed countries. As never before, these resources and ecosystems are under direct and imminent threat. Yet we now have a more complete understanding of these ecologically diverse systems and the impacts of human activities on them on local, regional, and global scales.

Megadiverse countries find themselves pressed between the responsibility, and even the fundamental desire, to preserve their biological diversity and the inescapable need to attend to the living requirements and well-being of their growing populations. The difficulty lies in meeting both the justified demands for development and the responsibility for biodiversity conservation. We believe that in order for these countries to increase the possibilities of having a successful outcome in this "tug-of-war," serious consideration should be given to the following two points.

(1) It is impossible to make informed decisions about the conservation of a country's biological diversity if a scientifically sound platform of information does not exist. Such a platform is built with the training of specialists (taxonomists and systematists), the funding of extensive fieldwork, and the buildup of scientific collections. These are long-term (many decades long) enterprises that should be undertaken in a carefully planned manner. While these efforts are under way, a much faster and efficient way to collect and analyze information is to generate biodiversity databases using new field data and the vast amounts of existing information in herbaria and museums—all incorporated into an accessible geographic information system format. We believe that the examples discussed here offer an array of models that can be followed by other megadiverse countries to enable them to better understand the extent and characteristics of their biological wealth.

(2) In megadiverse countries with large indigenous and/or rural populations, the conservation of biological diversity will require an honest, well-intentioned, and equitable collaboration with the owners of the resources (indigenous people or rural communities) that are the object of conservation or management plans. If this is not done beginning with the initial stages of planning, there will be limited success, if any, in such plans. On the other hand, by carrying out successful collaborative programs using this approach, it will be possible to conserve a nation's biodiversity and the cultures, languages, and traditions of the indigenous peoples who own and inhabit the land. Sustained efforts toward achieving this type of partnership for conservation in megadiverse countries will pay off in the long run. Very simply, this "experiment" has seldom been done correctly and, once proven practicable and successful, it could then be adapted and applied to a variety of social, economic, and biological situations.

See Also the Following Articles

Biodiversity as a Commodity • Deforestation • Economic Growth and the Environment • Indigenous Peoples, Biodiversity and Social and Cultural Factors • Tropical Ecosystems

Bibliography


BIODEGEOCHEMICAL CYCLES

Paul G. Falkowski
Rutgers University

I. Introduction
II. The Distribution of Elements on Earth
III. Key Biogeochemical Reactions and Cycles
IV. Evolution of Biogeochemical Cycles

GLOSSARY

anion Negatively charged atom or molecule.
cation Positively charged atom or molecule.
chemoautotroph Capable of growing on inorganic substrates without light energy.
eolian Wind blown.
felsic rock Continental crustal rocks relatively rich in silicon and aluminum.
fixation To make nonvolatile (forming organic molecules from inorganic gases).
heterotroph Growth dependent on the oxidation of organic matter.
mafic rock Oceanic crustal rocks relatively rich in magnesium and iron.
orogenesis Formation of mountains.
photoautotroph Growth dependent on the conversion of light to chemical energy (photosynthetic).
spallation Thermally induced neutron and proton ejection following high-energy proton collision.

VIRTUALLY ALL NATURALLY OCCURRING ELEMENTS are found in living organisms. The biological assimilation, biochemical transformation, physical transport, and geological sequestration and mobilization of many elements are self-perpetuating and self-regenerating, leading to a “cycling” of the elements. Although the biogeochemical cycle for each element is unique, general reaction pathways can be discerned depending on the chemistry of the element, its role in metabolism, and its distribution on Earth.

I. INTRODUCTION

Like life itself, biogeochemical cycles are far from thermodynamic equilibrium, have evolved over hundreds of millions of years, and are interdependent, forming biogeochemical systems replete with feedback controls (Schlesinger, 1997). Biogeochemical cycles depend on, and co-evolved with, specific metabolic pathways. Hence, biogeochemical cycles depend on and are a selective force in metabolic (i.e., biological) diversity. Over geological time, biogeochemical cycles are responsible for altering the chemistry of the ocean, atmosphere, and terrestrial ecosystems such that rate-limiting reactions within key cycles modify the tempo and mode of evolution. Here, we examine some of the key biogeochemical cycles in the context of their evolution and biological diversity.
II. THE DISTRIBUTION OF ELEMENTS ON EARTH

A. The Origin of the Elements

The relative abundance of the crustal elements reveals a distribution of much more abundant light elements to lesser abundant heavier elements (Fig. 1). Although the two lightest elements, H and He, were formed approximately 16 billion years ago in the "big bang," all the heavier elements result from fusion of He nuclei or fusion/spallation (proton or neutron loss) reactions in stars (Degens, 1989). The fusion reactions involving 4He with itself, H, O, C, or N tend to form even-numbered atomic nuclei, whereas spallation and proton capture leads to odd-numbered nuclei. Additionally, because the nuclei of elements with paired protons are slightly more stable than those with an odd number, there is generally a larger relative abundance of even-numbered elements.

In the origin of our solar system approximately 4.6 billion years ago, elemental composition and planetary accretion were strongly influenced by the gravitational forces of the sun. Planet bodies closer to the sun contain relatively higher proportions of heavy elements than bodies further away. The four innermost planets are approximately three times denser than the outer planets and have solid-rock surfaces that contain a relatively high proportion of metals, especially iron and aluminum. In the accretion process, a further gravitational distillation occurred within the planets. On Earth, the abundant heavier elements, such as nickel and iron, tended to migrate toward the center of the internal gravitational field, whereas lighter elements tended to float above the metal core and, upon cooling, accumulated as a solid surface and crust. The lightest elements formed a gaseous phase. Almost all of the two lightest gases, H2 and He, escaped the gravitational field and diffused into interplanetary space during this initial period of Earth's history.

The composition of the gases in Earth’s atmosphere following accretion is not completely resolved but almost certainly contained high concentrations of CO2, N2, and H2O, HCl, and H2SO4. This gas composition is similar to that currently on Venus. Precipitation of minerals and formation of felsic rocks led to the condensation and upward migration of liquid, precipitable water that overlies vast regions of denser, mafic rocks.
Additional water may have been provided by meteoritic bombardment. Based on thermodynamic equilibrium calculations with crustal elements, the acidic gases or hydrated equivalents (e.g., HCl and H2SO4) would have solubilized mineral cations in the primordial ocean leading to a seawater dominated by Na+, K+, Mg2+, and Ca2+. The anion balance would have been supplied by vulcanism and outgassing from deep crustal sources. The dominant anions were Cl− and, to a lesser extent, SO42− and HCO3−. PO43− was probably present but to a lesser extent, and fixed inorganic N (as NH3 or NO3) was almost certainly very scarce. The pH of the early oceans was probably close to neutral or slightly acidic (pH 8–7.0). Additionally, radiogenically produced heat, the high concentrations of greenhouse gases, and geothermal activity would have provided a source of heat to maintain elevated temperatures in Earth’s early ocean history. From these conditions, life arose and evolved. Over the course of Earth’s history, there has been a continuous trajectory from the mildly reducing conditions that prevailed at the time of origin of the planet to the highly oxidizing conditions that prevail in the contemporary geological epoch. This oxidation trajectory is driven by biological processes but has also led to the increased diversity of life forms.

B. Phase State Transitions and Elemental Partitioning

The distribution of elements between the atmosphere, lithosphere, and hydrosphere depends, to first order, on the phase state of the element, its chemical reactivity, and its partitioning coefficients between water, the atmosphere, and the lithosphere (Table 1). Gases such as N2 and O2 partition between the atmosphere and liquid phase according to the solubility coefficient. CO2, H2S, and NO2 undergo hydration to form carbonic, sulfuric, and nitric acids. These acids deprotonate, according to the pH of the aqueous environment, forming anions. Consequently, the total solubility of these species in the aqueous phase exceeds that predicted from the vapor pressure alone, assuming the gases are chemically inert and obey idealized gas laws. Similarly, the distribution of elements between the aqueous phase and the lithosphere is dictated, to first order, by the solubility coefficient of the element, its reactivity or ionic form, and the saturation level. For example, all the alkali and alkaline earth metals (i.e., Li, Na, K, Cs, Ca, and Mg) are highly reactive with water and are highly soluble. Precipitation reactions of alkali earth metal cations with carbonate, sulfate, or phosphate anions result in mineral deposits in the solid phase. The movement of elements between the lithosphere, atmosphere, and hydrosphere is mediated by both purely physical/chemical processes and biological processes.

C. Physical/Chemical Transport Processes: Weathering

The flux of elements from continental rock sources to the atmosphere and oceans is promoted by weathering, which is the physical and chemical degradation of rocks to smaller physical pieces and soluble chemical constituents. The physical processes that promote weathering include erosion by water and wind, cracking through freeze–thaw cycles, and active abrasion by the movement of ice (e.g., glaciers), other rock formations (e.g., earthquakes and tectonic uplifting), or rocks (e.g., tidal abrasion). Ultimately, these processes lead to the formation of sediments that are transported by rivers to the ocean. In addition to physical weathering, chemical weathering is promoted by weak acids (e.g., carbonic and sulfuric) in precipitation as well as microbial and plant growth. Over geological time, the weak acids in precipitation remove alkaline cations, such as Mg and Ca, from carbonates and silicates, leading to the mobilization of these elements. Moreover, the production of weak organic acids by lichens, terrestrial plant roots, and the microbial degradation of decaying organic matter greatly promote chemical weathering. Thus, the invasion of land by terrestrial plants led to increased mobility of numerous elements and enhanced the fluxes of materials to the oceans. This results in a feedback between biogeochemical processes on land and in the ocean.

D. The Hydrological Cycle

Weathering is strongly dependent on the hydrological cycle, which is one of the earliest and most critical cycles on Earth. Water is both a solvent and a vehicle for the transport of elements. It is also critically necessary for life. Hence, the distribution of biota is critically related to the availability of water. By far, the oceans are the largest reservoir of water on Earth (Fig. 2). Flux of water into and from the oceans is driven by evaporation and precipitation, processes that are in turn driven by the heat budget of the planet. Over the oceans, evaporation exceeds precipitation; the net difference leads to precipitation over land. Because precipitation exceeds evaporation over land, there is a flux of water from land to the oceans in rivers. The atmospheric
<table>
<thead>
<tr>
<th>Element</th>
<th>Crystal ionic charge and radius</th>
<th>Continental crust (wt%)</th>
<th>Continental crust (Vol.%)</th>
<th>Oceanic crust (wt%)</th>
<th>Oceanic crust (Vol.%)</th>
<th>Ocean Water (wt%)</th>
<th>Ocean Water (Vol.%)</th>
<th>Average sediments (wt%)</th>
<th>Average sediments (Vol.%)</th>
<th>Atmosphere (wt%)</th>
<th>Atmosphere (mol%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>−2</td>
<td>1.32</td>
<td>46.40</td>
<td>93.04</td>
<td>43.80</td>
<td>92.57</td>
<td>47.61</td>
<td>91.32</td>
<td>86.0</td>
<td>99.0 as H₂O</td>
<td>23.15</td>
</tr>
<tr>
<td>Si</td>
<td>+4</td>
<td>0.42</td>
<td>28.13</td>
<td>1.04</td>
<td>24.00</td>
<td>0.93</td>
<td>24.40</td>
<td>0.86</td>
<td>1.28</td>
<td>1.28</td>
<td>1.28</td>
</tr>
<tr>
<td>Al</td>
<td>+3</td>
<td>0.51</td>
<td>8.13</td>
<td>0.36</td>
<td>8.76</td>
<td>0.63</td>
<td>6.03</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>Fe</td>
<td>+3</td>
<td>0.64</td>
<td>3.63</td>
<td>0.46</td>
<td>8.56</td>
<td>0.74</td>
<td>3.79</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Ca</td>
<td>+2</td>
<td>0.99</td>
<td>4.15</td>
<td>1.40</td>
<td>6.72</td>
<td>2.39</td>
<td>7.86</td>
<td>2.54</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Na</td>
<td>+1</td>
<td>0.97</td>
<td>2.36</td>
<td>1.31</td>
<td>1.94</td>
<td>1.13</td>
<td>1.36</td>
<td>0.72</td>
<td>1.09</td>
<td>1.09</td>
<td>1.09</td>
</tr>
<tr>
<td>Mg</td>
<td>+2</td>
<td>0.56</td>
<td>2.33</td>
<td>0.38</td>
<td>4.5</td>
<td>0.78</td>
<td>2.44</td>
<td>0.39</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>K</td>
<td>+1</td>
<td>1.33</td>
<td>2.09</td>
<td>1.75</td>
<td>0.83</td>
<td>0.73</td>
<td>2.00</td>
<td>1.61</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Ti</td>
<td>+4</td>
<td>0.68</td>
<td>0.54</td>
<td>0.05</td>
<td>0.90</td>
<td>0.09</td>
<td>0.68</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>Mn</td>
<td></td>
<td>0.095</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td>0.14</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>+5</td>
<td>0.53</td>
<td>0.03</td>
<td>0.14</td>
<td>0.02</td>
<td>0.013</td>
<td>0.16</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>S</td>
<td>+6</td>
<td>0.30</td>
<td>0.026</td>
<td>0.023</td>
<td>0.62</td>
<td>0.007</td>
<td>0.62</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>C</td>
<td>+4</td>
<td>0.16</td>
<td>2.91</td>
<td>0.013</td>
<td>0.28</td>
<td>0.002</td>
<td>0.28</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>Cl</td>
<td>−1</td>
<td>1.81</td>
<td>0.85</td>
<td>1.85</td>
<td>1.94</td>
<td>0.833</td>
<td>1.94</td>
<td>0.833</td>
<td>1.94</td>
<td>0.833</td>
<td>1.94</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>75.5</td>
<td>78.09 (N₂)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ar</td>
<td></td>
<td>1.28</td>
<td>0.93 (Ar)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

lifetime of water is extremely short—on the order of ca. 10 days. This lifetime compares with the average for the oceans of ca. 3000 years, whereas deep groundwater can have lifetimes exceeding 10,000 years. These differences in lifetimes are critical determinants of fluxes of elements and climate feedbacks.

E. Vulcanism and Orogenesis

The riverine and eolian (wind-blown) transport of sediments from land to the oceans would, over geological time, erode the continents to the peneplane—the lowest level achievable. Vulcanism and other orogenic processes counterbalance erosion and supply new crustal rock and atmospheric sources of minerals. In the subduction regions of converging tectonic plates, sedimentary rocks are injected back into the deeper crust, where they are reprocessed by heat and pressure, emerging again in uplifting regions and through volcanic eruptions. In addition to the formation of igneous rock, vulcanism provides significant sources of iron, sulfate, and CO2 to the atmosphere. These elements and materials are transported through the atmosphere and deposited over both the oceans and on land.

F. Biological Transport

Biological processes influence the distribution of elements primarily through phase state transitions. The chemical reduction of inorganic carbon to organic matter is a conversion of a gas to solid phase; the latter is often dissolved in an aqueous phase of the cell cytoplasm. For oxygenic photoautotrophs, the evolution of molecular oxygen leads to the formation of a gas from the liquid phase of water. Biological mediation of phase state transitions is critical for the movement of elements between the atmosphere, lithosphere, and hydrosphere. For example, CaCO3 is by far the largest reservoir of carbon on Earth (Table II). The formation of CaCO3 is the consequence of the biological precipitation by marine organisms and, over geological time, represents an important sink for atmospheric CO2. The oxidation of organic matter leads to the formation of CO2, which equilibrates with the atmosphere. Similarly, the biologically catalyzed oxidation of H2S to S or SO4 is a phase state transition that influences the partitioning of sulfur between aquatic ecosystems and the atmosphere. It should be noted that the biologically catalyzed phase state transitions need not be direct but can result from

FIGURE 2. Estimates of the global water cycle and its reserves. The accuracy of several of its components is poor, resulting in a closure error for the whole cycle of about a factor of 2. The obvious interactive nature of the cycle makes it impossible to reduce current closure errors without studying the whole cycle. (Reprinted by permission from Nature, Chahine, Moustafa T., The hydrological cycle and its influence on climate. Vol. 359, p. 373, copyright 1992, Macmillan Magazines Ltd.)
the indirect modification of the redox state of the environment. For example, the oxidation of H\textsubscript{2}S occurs spontaneously in the presence of O\textsubscript{2}; however, the O\textsubscript{2} is formed from oxygenic photoautotrophs.

G. Biological Assimilation

To accomplish phase state transitions, biological processes catalyze chemical and/or physical reactions. Reaction sequences require an assimilation of the elements. The elemental composition of organisms is dictated, to first order, by the biochemical machinery required for maintenance, growth, mobility, reproduction, and defenses against predation. The four fundamental biochemical constituents that comprise all organisms, namely, proteins, nucleic acids, lipids, and carbohydrates, are primarily synthesized from six light elements: H, C, N, O, P, and S. Of these, N, C, and O are generally reduced, P is oxidized, and S can be both reduced and oxidized. These six elements comprise approximately 98% of all living biomass. The remainder consists of approximately 20 ion-forming elements, including Na, Mg, Si, Cl, Ca, V, C, Mn, Fe, Co, Ni, Cu, Zn, Se, Mo, Sn, and I. The transition metals are frequently used in the mediation of electron transfer (redox) reactions.

III. KEY BIOGEOCHEMICAL REACTIONS AND CYCLES

A. Hydrolysis and Oxidation-Reduction Reactions

Two biologically mediated chemical reactions are key to understanding biogeochemical cycles, namely, hydrolysis and oxidation-reduction processes. Hydrolytic reactions chemically remove or add water to specific elements or molecules; the reactions are almost always accompanied by the transfer of protons and thus are pH-dependent processes. Hydrolytic reactions can be described in a general form by

\[
\text{AH}^+ + \text{OH}^- \rightarrow \text{A} + \text{H}_2\text{O}
\]

\[
\text{AOH}^- + \text{H}^+ \rightarrow \text{A} + \text{H}_2\text{O}
\]

Hydration/dehydration reactions can thermodynamically store or discharge energy and determine the reactivity and solubility of elements in the first series (H, Li, Na, etc.) and of phosphorous, (inorganic) carbon, and silica in the environment.

Redox reactions are characterized by the addition or loss of electrons, hydrogen atoms (but not simply a proton), or molecular oxygen. Redox reactions are the primary engines of life, and the consequences of redox chemistry strongly impact the distributions of oxygen, carbon, nitrogen, sulfur, and transition metals in the environment. Redox reactions are always paired; they must be an oxidant and a reductant. The reaction process can be generalized by half-cells:

\[
\text{A}_{\text{oxid}} + n\text{e}^- \rightarrow \text{A}_{\text{red}}
\]

\[
\text{B}_{\text{red}} \rightarrow \text{B}_{\text{oxid}} + n\text{e}^-
\]

where the overall reaction is

\[
\text{A}_{\text{oxid}} + \text{B}_{\text{red}} \rightarrow \text{A}_{\text{red}} + \text{B}_{\text{oxid}}
\]

For each half-cell, the tendency for a substrate to be oxidized or reduced is described by the Nernst equation:

\[
E = (E_0 + 2.3 \text{RT})/nF \log([\text{A}_{\text{red}}]/[\text{A}_{\text{oxid}}])
\]

where E is the redox potential (in volts), \(E_0\) is an arbitrarily accepted standard redox potential, \(F\) is the Faraday’s constant (~ 96,487 C/mol electron), \(n\) is the number of moles of electrons (Faraday’s) transferred in the half-cell reaction, \(R\) is the Boltzmann gas constant,
\begin{itemize}
\item \textbf{T} is temperature in Kelvin, and \(A_{\text{red}}\) and \(A_{\text{ox}}\) are the activities (or more commonly the concentrations) of the oxidized and reduced forms of the molecules, respectively. The Nernst equation describes the equilibrium condition for both electronic and ionic processes.
\item To maintain perpetual redox reactions, reductants consumed in one reaction must be regenerated in another (Table III). Hence, by the very fact that redox reactions are paired, a cycle is potentiated. However, other (Table III). Hence, by the very fact that redox reactions are paired, a cycle is potentiated. However, other
\end{itemize}

\begin{table}[h]
\centering
\begin{tabular}{lll}
\hline
\textbf{Reaction} & \textbf{\(E_V\)} & \textbf{\(\Delta G\)} \\
\hline
Reduction of \(O_2\) & \(2O_2 + 4H^+ \rightarrow 4H_2O\) & 0.812 & 29.9 \\
Reduction of \(NO\) & \(2NO + 4H^+ \rightarrow 4H_2O\) & 0.747 & 28.4 \\
Reduction of \(Mn^{2+}\) to \(Mn^{0}\) & \(MnO_2 + 8H^+ + 4e^- \rightarrow Mn^{0} + 4H_2O\) & 0.526 & 23.3 \\
Reduction of \(Fe^{3+}\) to \(Fe^{2+}\) & \(Fe(OH)_3 + 8H^+ + 6e^- \rightarrow 3H_2O + Fe^{2+}\) & 0.947 & 10.1 \\
Reduction of \(SO_2\) to \(H_2S\) & \(SO_2 + 10H^+ + 6e^- \rightarrow H_2S + 4H_2O\) & -0.221 & -5.9 \\
Reduction of \(CO_2\) to \(CH_4\) & \(CO_2 + 6H^+ + 6e^- \rightarrow CH_4 + 2H_2O\) & -0.244 & -3.6 \\
\hline
\end{tabular}
\caption{Thermodynamic Sequence for Reduction of Inorganic Substances by Hydrogen at pH 7.0 and 298°C.}
\end{table}

\begin{itemize}
\item Under mildly reducing conditions, photosynthetic carbon fixation is catalyzed by bacteria that couple the oxidation of \(H_2S\) or organic compounds to the reduction of \(CO_2\) to form organic compounds. The thermodynamic energy gradient in such systems is relatively low, requiring inputs of \(-0.4 \text{ eV} \) of photon energy. Such energy demands are satisfied in the infrared spectrum.
\item The precipitation of \(CaCO_3\) leads to the formation of \(CO_2\). In the absence of any biological activity, the geochemical processes would lead to a cycle of inorganic carbon between the three major reservoirs, namely, the atmosphere, oceans, and continents; however, biological processes interact in two critical ways to alevate carbon chemistry. First, biological pathways lead to the reduction and oxidation of carbon. Photobiological and chemosynthetic organisms oxidize alternative substrates to reduce carbon to the equivalent of carbohydrate. These reactions can be broadly classified in the reaction sequence
\end{itemize}

\begin{align*}
\text{H}_2O + \text{CO}_2 & \rightarrow \text{CH}_4O + A_j \\
\text{Ca}^{2+} + 2\text{HCO}_3^- & \rightarrow \text{CaCO}_3(s) + \text{CO}_2 + H_2O
\end{align*}

where \(A\) can be sulfur, oxygen, or organic compounds. This is an example of redox reactions applied to the formation of organic material.

\begin{itemize}
\item The weathering process is stimulated by the hydration of \(CO_2\) in the atmosphere in rain, forming carbonic acid. This is a self-limiting cycle; as \(CO_2\) rises, weathering accelerates leading to a depletion of the atmospheric inventory, which reduces weathering. In the oceans there is an excess of \(Ca^{2+}\), and an equilibrium in the inorganic carbon system is reached. This reaction series can be specified as
\end{itemize}

\begin{align*}
\text{H}_2O + \text{CO}_2 & \rightarrow \text{CH}_4O + A_j \\
\text{Ca}^{2+} + 2\text{HCO}_3^- & \rightarrow \text{CaCO}_3(s) + \text{CO}_2 + H_2O
\end{align*}
The long-term increase in $\delta^{18}O$ values reflects cooling of the deep ocean and growth of ice sheets at high latitudes. $\delta^{18}O = \left(\frac{^{18}O}{^{16}O}\right)_{\text{sample}} - 1$ where standard is PDB. (Reprinted by permission from Nature, Raymo, M. E., and Ruddiman, W. F. Techtonic forcing of late Cenozoic climate. Vol. 359: 117–122, copyright 1992, Macmillan Magazines Ltd.)

Consequence of biological mineralization in the ocean. Over geological time, the production of carbonates produced vast beds of chalk, limestones, and marbles. The deposition of carbonates accounts for the largest reservoir of carbon on Earth. The precipitation of carbonates is strongly influenced by the temperature of the ocean when and where the carbonates are formed. The isotopic fractionation of the stable isotopes of carbon ($^{13}C$ and $^{12}C$) is related to the atmospheric/oceanic inventory of CO$_2$. Hence, isotopic analyses of carbonates provide clues about both oceanic temperatures and carbon dioxide concentrations over geological time (Fig. 3).

Some groups of marine planktonic organisms, including diatoms and radiolarians, as well as several genera of higher plant grasses, precipitate silica to form hydratedopal. Silica is supplied to the oceans from the weathering of continental rocks and the subsequent flux of soluble silicic acid via rivers. Like carbonates, the precipitation of silicates over geological time leads to sedimentary deposits. In terrestrial ecosystems, several groups of higher plants incorporate silica from soils into stems and shoots.

Under basic conditions, calcium and phosphate ions will spontaneously precipitate to form apatites, which are often mined as a source of phosphate for fertilizers and chemical feedstocks. Vertebrates biologically control the precipitation of calcium phosphate to form bones and teeth. Approximately 85% of the phosphate and 95% of the calcium assimilated by vertebrates is incorporated in bones. The fossil record of these biomineral components provides a significant signature of vertebrate evolution but is not a significant component of crustal minerals.

### D. The Oxygen Cycle

One of the most important characteristics of Earth’s atmosphere is the relatively high concentration of free molecular oxygen (O$_2$). The mere presence of high concentrations of free molecular oxygen suggests that the chemistry of the planet is far from thermodynamic equilibrium. For all practical purposes, all the oxygen in the atmosphere originated from the photosynthetically mediated oxidation of water by oxygenic photautotrophs. The oxidation of water to form free oxygen requires an oxidant with a potential of at least 0.83 V at pH 7. The bacterial photosynthetic apparatus could not achieve such a high oxidation potential without evolutionary modification. These modifications included the incorporation of Mn as the transient electron acceptor in the photosynthetic reaction center, the alteration of proteins that permitted the oxidation of water, and the raising of the potential of the primary electron donor by $-0.4$ V. The latter was achieved by altering the pigment involved in the photosynthetic reaction from the relatively low energy levels that characterize bacteriochlorophylls to the higher energy levels found in chlorophyll a, a pigment that distinguishes cyanobacteria (i.e., blue-green algae) from all other photosynthetic bacteria.

The cyanobacteria are the oldest extant oxygenic photautotrophs. The exact timing of the origin of cyanobacteria is uncertain, but micropaleological evidence suggests that organisms with features remarkably similar to those of modern cyanobacteria existed in the Archean ocean at least 3.5 billion years before present. Based on geochemical evidence (primarily the oxidation of iron by molecular oxygen), it appears that oxidation of Earth’s atmosphere occurred over a relatively short
interval of ~100 million years, beginning ~2.2 billion years before present (Fig. 4).

Prior to the photosynthetic production of oxygen, all heterotrophic organisms oxidized organic carbon under anaerobic conditions. The biochemical oxidation of carbon under such conditions is thermodynamically inefficient, often leading to fermentation. For example, in the absence of oxygen, molecules such as acetate are potentially oxidized to methane and CO₂:

\[ \text{H}_4\text{COOH} \rightarrow \text{CH}_4 + \text{CO}_2 \]

where the methane is liberated to the environment.

The generation of oxygen led to thermodynamically favorable conditions for the oxidation of organic matter via aerobic respiration. The basic redox sequence for aerobic respiration is

\[ (\text{CH}_2\text{O})_n + \text{O}_2 \rightarrow \text{CO}_2 + \text{H}_2\text{O}. \]

This thermodynamic coupling between the photosynthetic reduction of organic carbon and the oxidation of water is reversed in mitochondria, in which oxygen is used as a sink for electrons and the oxidation of organic carbon is coupled to biochemical energy. Initially, the oxygen could be utilized by single-celled organisms via diffusion; subsequently, however, proteins and structures evolved that facilitated the diffusion and transport of O₂ and led to the evolution of metazoa.

E. Sequestration and Burial

From a biogeochemical perspective, the net evolution (i.e., accumulation) of molecular oxygen in the atmosphere means that oxidation of water and the (bio)geochemical reduction of free oxygen are not balanced, i.e., photosynthetic oxygen evolution must have exceeded oxygen consumption. Given the vast quantities of water on the surface of the planet, oxygenic photosynthesis could produce enormous quantities of oxygen without running out of substrate. When can such conditions occur and what are the oxygen-consuming reactions?

Prior to the net accumulation of oxygen in the atmosphere, oxygen produced by oxygenic photautotrophs in the oceans would have been exposed to several potential reductants, two of which are especially biogeochemically important, namely, sulfur and iron. After Cl⁻,
SO\textsuperscript{2−} is the most abundant anion in the ocean. Under the mildly reducing conditions of the Archean ocean, dissolved aqueous sulfur in the ocean would have been in equilibrium between a reduced form, probably sulfide (H\textsubscript{2}S), and the hydrated, oxidized form H\textsubscript{2}SO\textsubscript{4}. The latter (sulfuric acid) would equilibrate with the alkaline or earth alkaline elements of the first and second series, resulting in the formation of ionized sulfate salts. The precipitation of sulfate minerals (e.g., gypsum) could occur in shallow seas when evaporation leads to supersaturation and during periods when the influx of Ca from continental sources is high. The latter occurred during periods of rapid invasion of land by terrestrial plants (e.g., during the Triassic Period), where plant growth promoted weathering. Thus, photosynthetic evolution of oxygen led to the formation of vast quantities of sulfate via the oxidation of reduced sulfur anions; sulfur oxidation acted as a sink for molecular oxygen.

Under reducing conditions, iron, as Fe(II), is relatively soluble in seawater, whereas its oxidized counterpart, Fe(III), is virtually insoluble. The photosynthetic oxidation of the Archean ocean not only resulted in the oxidation of sulfur but also led to the oxidation and subsequent precipitation of iron oxides (rust) in ocean sediments. Thus, the biologically catalyzed oxidation of iron served as a second electron sink for molecular oxygen—a sink that ultimately came to represent the largest reservoir for the transition metal on Earth. The deposition of iron oxides would proceed until virtually no soluble iron remained in the ocean. The deposition of iron in Archean sediments is recorded in several deposits, the dating of which provides a record of the oxidation of the atmosphere (Fig. 5).

When the electron sinks in reduced iron and sulfur were consumed, oxygen could be consumed in respiratory reactions involving (photosynthetically) reduced carbon. If, however, there was a net evasion of photosynthetically derived oxygen to the environment (i.e., the oceans and atmosphere), then a stoichiometric amount of organic carbon must have avoided oxidation, i.e., organic carbon must have accumulated. The primary sink for organic carbon is marine sediment in which, in the absence of oxygen, carbon adsorbs to inorganic sedimentary particles and becomes buried in the sedimentation process. The sequestration and burial process occurred over hundreds of millions, if not billions, of years, and it resulted in the accumulation of massive amounts of organic carbon associated with sedimentary and lightly metamorphosed rocks (e.g., shales.)
TABLE IV

Annual and Seasonal NPP of the Major Units of the Biosphere, from CASA-VGPM\textsuperscript{a}

<table>
<thead>
<tr>
<th></th>
<th>Ocean NPP</th>
<th>Land NPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seasonal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>April to June</td>
<td>10.9 15.7</td>
<td>13.7 15.7</td>
</tr>
<tr>
<td>July to September</td>
<td>13.0 18.0</td>
<td>18.0 18.0</td>
</tr>
<tr>
<td>October to December</td>
<td>12.3 11.5</td>
<td>11.3 11.5</td>
</tr>
<tr>
<td>January to March</td>
<td>11.3 11.2</td>
<td>11.2 11.2</td>
</tr>
<tr>
<td>Biogeographic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oligotrophic</td>
<td>11.0 17.6</td>
<td>17.8 17.8</td>
</tr>
<tr>
<td>Mesotrophic</td>
<td>27.4 3.3</td>
<td>3.3 3.3</td>
</tr>
<tr>
<td>Eutrophic</td>
<td>9.1 3.1</td>
<td>3.1 3.1</td>
</tr>
<tr>
<td>Macrophytes</td>
<td>1.0 3.1</td>
<td>3.1 3.1</td>
</tr>
<tr>
<td>Savannas</td>
<td>16.6 4.4</td>
<td>4.4 4.4</td>
</tr>
<tr>
<td>Perennial grasslands</td>
<td>2.4 0.8</td>
<td>0.8 0.8</td>
</tr>
<tr>
<td>Broadleaf shrubs with bare soil</td>
<td>1.0 0.8</td>
<td>0.8 0.8</td>
</tr>
<tr>
<td>Tundra</td>
<td>0.8 0.8</td>
<td>0.8 0.8</td>
</tr>
<tr>
<td>Desert</td>
<td>0.5 0.2</td>
<td>0.2 0.2</td>
</tr>
<tr>
<td>Cultivation</td>
<td>8.0 8.0</td>
<td>8.0 8.0</td>
</tr>
<tr>
<td>Total</td>
<td>48.3 56.4</td>
<td>56.4 56.4</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Ocean data are averages from 1978 to 1983. The land vegetation index is from 1982 to 1990. All values are in petagrams of carbon (1 pg = 10\textsuperscript{15} g). Oligotrophic < 0.1 \(\mu g\text{ Chl L}^{-1}\); mesotrophic 0.1–1.0 \(\mu g\text{ Chl L}^{-1}\); eutrophic > 1.0 \(\mu g\text{ Chl L}^{-1}\). The macrophyte contribution to ocean production is not included in the seasonal totals. (Reprinted with permission from Field et al. (1998). “Primary production of the biosphere: integrating terrestrial and oceanic components.” Science 281, 237–240.)

F. Primary Production in the Contemporary World

The carbon remaining in photosynthetic organisms follows their own respiratory costs is potentially available for consumption by other organisms. This remaining carbon, or net primary production (NPP), provides the energy requirements of all ecosystems. The distribution of NPP between the terrestrial and marine ecosystems is given in Table IV. These data suggest that terrestrial ecosystems, containing approximately 99% of the plant biomass, account for approximately 55% of the global NPP, whereas marine ecosystems, containing approximately 1% of the plant biomass, account for 45% of the productivity. The huge discrepancy between NPP per unit biomass between these two ecosystems suggests that, on average, the lifetime of marine primary producers is less than 1 week, whereas the lifetime of terrestrial primary producers is on the order of a decade or more. On land, NPP is primarily limited by the availability of water, whereas in the oceans NPP is primarily limited by the availability of essential nutrients, especially fixed forms of nitrogen.

During the past several hundred thousand years, global NPP has remained relatively constant, whereas the distribution of NPP between terrestrial and marine ecosystems has waxed and waned with climatic shifts. Briefly, every 100,000 years or so, changes in Earth’s orbital relation to the sun lead to a small reduction in surface temperature, which is amplified (through unknown processes) to produce glacial periods. The advance of ice sheets in the Northern Hemisphere, combined with a reduction in liquid precipitation at low latitudes, leads to the loss of terrestrial NPP by approximately 30%. Simultaneously, however, enhanced upwelling fluxes of nutrients in the oceans combined with a strengthened eolian flux of iron stimulate oceanic

and cherts). A very small fraction of the buried organic carbon underwent alterations due to heat and pressure (i.e., diagenesis) to become fossilized carbon that literally fuels the industrialized world in the current geological epoch (Table II).
NPP, such that the change between the terrestrial loss and oceanic gain is approximately balanced. During interglacial periods, the situation reverses, with a retreat of the ice sheets and the relaxation of nutrient enrichment of the upper ocean to lead to the situation shown in Table IV for the contemporary world.

G. Oceanic Solubility and Biological “Pumps”

The oceans contain approximately 50-fold more organic carbon than the atmosphere, and on timescales of centuries the former reservoir controls the concentration of CO$_2$ in the latter. The vertical gradient of inorganic carbon in the oceans is “inverted”—that is, there is a higher concentration of inorganic carbon in the ocean interior than at the surface (Fig. 6). Because the ocean surface tends to equilibrate with the atmosphere, the vertical gradient is maintained by processes other than simple equilibrium diffusion. There are two major processes responsible for the observed gradient. First, cooling of waters at high latitudes increases the solubility of CO$_2$ while simultaneously increasing the density of the waters. The cold, CO$_2$-rich waters sink into the ocean interior, where they are transported by density discontinuities and the motion of the earth throughout the world oceans. The round-trip takes approximately 1000 years before the waters are again in contact with the atmosphere in the high latitude from whence they originated the journey. This “solubility” pump is responsible for approximately 50% of the inorganic carbon gradient. The other 50% is a consequence of biological activity. In the upper, lighted portion of the ocean, photosynthetic fixation of inorganic carbon by phytoplankton leads to the formation of organic matter, largely in the form of particulate materials. Gravity, acting on the particles, leads to a persistent “rain” of organic matter into the ocean interior. The organic matter is oxidized by heterotrophic bacteria and other organisms, leading to the production of inorganic carbon. Thus, this vertical flux of organic carbon and its subsequent oxidation effectively pumps up the ocean interior with inorganic carbon. The biological pump is maintained by the upward flux of inorganic nutrients from the ocean interior, especially fixed nitrogen and phosphate. Such nutrients are essential to supporting the photosynthetic activity of phytoplankton in the upper ocean.

H. The Nitrogen Cycle

All organisms contain nitrogen that is reduced to the equivalent of NH$_3$ (ammonia). The most abundant form of nitrogen in the environment is N$_2$, which at the temperature and pressure on Earth’s surface is a gas and forms 78% by volume of the planet’s atmosphere. Conversion of N$_2$ to NH$_3$ requires the addition of three protons and three electrons per atom of N; the balanced equation is

$$N_2 + 3H^+ + 3e^- \rightarrow 2NH_3$$

This reaction is kinetically limited at low pressure and without some catalytic enhancement there would be no ammonia in the environment. A small subset of bacteria evolved with the capability to “fix” N$_2$. Nitrogen fixation, or diazotrophy (“eating nitrogen”), is catalyzed by a protein complex (nitrogenase) which is irreversibly inactivated by oxygen. The enzyme complex normally contains an Fe subunit and a Mo, Fe subunit; however, V and Fe can substitute for Mo as mediators of electron transfer in some organisms. Gene sequence analysis suggests that nitrogenase evolved from a single common ancestor, and the enzyme is primarily found in eubacteria, although some Archea, such as methanogens, can also fix nitrogen. In vivo nitrogen fixation consumes approx-
In eubacteria, the supply of energy is either light (for N2-fixing cyanobacteria), or reduced carbon substrates (for heterotrophic bacteria). For methanogens it is H2. Whatever the energy source, N2 fixation is a metabolically expensive process, and hence, in the presence of high concentrations of fixed nitrogen, it usually is repressed. Many nitrogen-fixing bacteria live symbiotically with a host organism. One classical symbiosis is found in the Legumacea, which are higher plants such as peas and clover in which nitrogen-fixing bacteria are found in nodules attached to the plant root. In a second symbiosis, some insects, such as termites, contain nitrogen-fixing bacteria within their guts to supply essential amino acids to the host.

In most oxidized environments, fixed inorganic nitrogen is found in the form of nitrate (NO3−) or, to a lesser extent, nitrite (NO2−). These two oxidized forms of nitrogen are formed from ammonium via the action of nitrifying bacteria. Nitrification is a metabolic process in which the oxidation of ammonium can be coupled to the reduction of inorganic carbon to form organic molecules without light energy. This type of metabolic process is called chemosynthesis. For example,

\[ 2\text{NH}_3 + 2\text{CO}_2 + \text{O}_2 \rightarrow 2(\text{CH}_2\text{O}) + 2\text{NO}_2^- + 2\text{H}^+ \]

The pathway for ammonium oxidation is broken down between organisms that can oxidize the substrate to nitrite and those that further oxidize nitrite to nitrate. Nitrification is a metabolically inefficient process; one mole of NH3 is oxidized to produce each mole of organic carbon. Hence, the growth rate of nitrifying bacteria is generally low, whereas in the presence of O2 the rate of ammonium oxidation is generally rapid. Nitrification is confined to a small number of eubacteria; the gene sequence analysis of these organisms suggests the pathways evolved from single ancestral genes and are relatively conserved. Nitrate (NO3−) is a thermodynamically stable molecule under the oxidizing conditions of the contemporary world, and as such it is the most prevalent form of fixed nitrogen in the world oceans.

The cycle of nitrogen is completed when oxidized species of nitrogen, especially NO3−, are subsequently reduced to form N2O and N2. This anaerobic process, called denitrification, is mediated by a wide range of diverse bacteria and Archea. Denitrification is a respiratory pathway in which the terminal electron acceptor is NO3− or NO2−. Organisms generally reduce these two substrates “opportunistically,” that is, when they are deprived of oxygen. The reduction of nitrate to nitrous oxide and N2 represents a large loss of fixed nitrogen in the environment and largely occurs in anaerobic sediments of marine and lacustrine environments and in areas of soil rich in organic carbon that are periodically flooded (e.g., rice paddies and river deltas). Although approximately half of all the nitrogen fixed on Earth is due to natural biological activity in the environment, half is due to deliberate fixation by human activities (Fig. 7). Developed in the latter part of the nineteenth century by a German chemist, the Haber reaction permits the formation of ammonium from N2 by a high-temperature, high-pressure reduction in the presence of H2. This reaction is the primary source of fertilizer nitrogen for commercial crops, without which agricultural production would be much smaller than it is currently. However, approximately 23% of the applied nitrogen to agricultural crops is lost through denitrification, leading to an increase in N2O in the atmosphere. A significant fraction of the applied nitrogen in terrestrial ecosystems is further solubilized in waters, flowing either into groundwater or through surface waters to the sea. Most of this fixed nitrogen seldom makes its way to the open oceans; rather, due to the high rates of denitrification on continental margins (especially in sediments on the margins), the fixed nitrogen tends to be lost to the atmosphere through denitrification.

I. Profiles of Elements in the Sea

In the oceans there are three basic profiles to describe the distribution of elements. If an element is conservative, and is either extraordinarily abundant or does not interact with particulate matter (i.e., biological material), it will be homogeneously distributed. The soluble salts in seawater represent a vast quantity of alkaline metals, halides, and sulfates. These elements are so abundant relative to biological demands and biomass that they behave as conservative tracers; that is, their distributions are effectively uniform through the oceans. If an element is biologically reactive and can be absorbed in surface waters and subsequently transported by sinking to the ocean interior, the interior ocean can be enriched in the element relative to the surface. Such elements are said to have “nutrient-like” behavior. Examples of such elements include C, P, fixed N, Si, Fe, Cu, Ni, Cd, Hg, and the rare earth elements (REEs). It is important to note that simply because an element has a nutrient-like behavior it is not necessarily a true nutrient and vice versa. A REE, such as La3+, is highly reactive with sulfhydryl (−SH) groups found in most proteins and hence can be “scavenged” from the surface ocean by particles. In contrast, sulfate is so

...
abundant in the ocean that even though it is used as a nutrient in formation of proteins, the level of utilization is so small in comparison with its concentration that it behaves more like a conservative tracer than a nutrient. The third category of elemental distributions leads to an elevation in surface water and is primarily associated with an upper ocean source. One example is the vertical profile of O$_2$, which is generated by photosynthetic activity in the upper ocean and consumed in the sea interior by respiration.

**J. Sulfur**

Sulfur is an integral element in two amino acids, methionine and cysteine, and hence is an essential element for...
sulfate loading is excessive (e.g., when fossil fuel combustion is high), a small fraction of hydrated sulfate in rain is not buffered by cations, and the sulfate reacts with water to form a dilute solution of sulfuric acid (H₂SO₄), producing acidic rain. Note that the transport of sulfur from continental and aquatic sources to and from the atmosphere is a consequence of phase state transitions, redox reactions, and hydrolysis.

K. Phosphorus Cycle
Phosphorus is an essential component of all cells; the element forms the backbone of nucleic acids, without which cells cannot reproduce. Phosphorus is also incorporated in nucleotides, sugars, proteins, and lipids. In vertebrates, almost all the phosphorus is combined with calcium to form hydroxyapatite, which is the central molecule to build bones.

In the environment, phosphorous is found almost invariably in the oxidized form as hydrated phosphate. Mineral phosphate salts erode and weather from continental sources and are carried to aquatic ecosystems as soluble ions. There is no eolian phosphorous source and no significant transfer of phosphate from aquatic or terrestrial environments to the atmosphere. Unlike carbon, nitrogen, phosphate, or oxygen, phosphorus does not undergo significant biologically mediated redox reactions; the chemistry of phosphorus is solely based on hydrolysis.

L. Trace Metals
Transition metals play a key role in biological electron transfer (redox) reactions and biological processes in turn affect the availability of metals in the environment. The metals are incorporated into protein “scaffolds” in specific orientations and associated with specific ligands such that the electron transfer reactions are highly specified and optimized. The transfer of electrons to given substrates is often key to specific metals. For example, electron transfers involving O₂ frequently are mediated by Mn, whereas electron transfers involving N frequently are mediated by Mo. The availability of transition metals, in turn, is dictated by solubility and redox conditions. Under anaerobic conditions, Mn and Mo are often sequestered as sulfide precipitates, but they become more available under oxidizing conditions.

In contrast, Fe, which is a required electron transition metal in all organisms, is abundant under reducing conditions but scarce under oxidizing conditions. As previously noted, electron transfers through iron are essential to both anaerobic and oxygenic photosynthesis, heterotrophic respiration, and nitrogen fixation.
In some cases, anaerobic bacteria can oxidize Fe(II) to Fe(III), thereby precipitating iron as a ferric hydroxide in the sediment. Although iron is the most abundant transition metal in the earth’s crust, its concentration in the open ocean is extremely low, averaging no more than ~1 μM. To satisfy iron fluxes, eolian sources of iron are critical. Wind-blown dust from major desert areas supplies much of the iron to the open oceans. Consequently, open ocean areas, far removed from continental dust regions, have extremely low iron concentrations that result in a direct limitation of photosynthetic electron transport. Three such areas are the Southern Ocean, the eastern Equatorial Pacific, and the sub-Arctic Pacific. Ice core records suggest that during glacial periods eolian iron fluxes were much greater. The variations in dust supply appear to correspond with changes in the hydrological cycle; warm, wet climate regimes tend to reduce the eolian supply and hence may limit the biological sequestration of carbon in the oceans, whereas cold, dry periods may stimulate oceanic primary production. Mn is essential for oxygen evolution and, like iron, its concentration is extremely low in the oceans. The Mn requirements for biological activity are lower than those for Fe; consequently, there is no evidence of Mn limitation in either the oceans or the terrestrial ecosystems. Interestingly, Mn and Fe can exchange electrons spontaneously and, in the presence of a redox gradient established in sediments, can cycle between oxidized and reduced states. Mn has two main oxidation states, namely, Mn(II) and Mn(IV). Organic carbon and NH₄ can be oxidized by Mn(IV), forming CO₂ and N₂, respectively, whereas reduced Mn can be oxidized by O₂ and possibly by NO₃.

Many other transition metals have been co-opted for various biologically mediated electron transfers. Mo is often utilized biochemically in electron transfer reactions involving nitrogen. Most nitrogenases contain Mo, as do all nitrate reductases. Copper is frequently found in hydrophilic proteins that either catalyze single electron transfer reactions (e.g., plastocyanin) or coordinate oxygen binding (in the case of hemocyanin). The availability of these metals is inverse to those of Mn and Fe, namely, under oxidizing conditions, Mo and Cu are more abundant than under reducing conditions (Williams and Frausto da Silva, 1996).

IV. EVOLUTION OF BIOGEOCHEMICAL CYCLES

By definition, biogeochemical cycles are mediated in part by living organisms. The origins of biogeochemical cycles are presumably entwined with the origins of life. Although there is no consensus on the definition of life, to sustain life all living organisms must grow and reproduce. Growth is dependent on the availability of substrates and energy. C, N, P, H, O, and S are required substrates; often these substrates are combined (e.g., all organisms require phosphorous in the form of the phosphate anion, H₄PO₄⁻). All life forms either oxidize or reduce carbon; hence, carbon is either an electron sink or a source, depending on the thermodynamic basis of the organism’s metabolism. Given an external energy source, the redox reactions can run “uphill,” where an environmental redox gradient is used to thermodynamically provide the free energy for the subsequent carbon reduction. This metabolic strategy characterizes chemoautotrophs. In photoautotrophs, the energy is externally provided by light. All carbon-oxidizing organisms require an alternative electron sink, such as O₂, NO₃, and SO₄. All the basic biogeochemical cycles evolved more than 2 billion years ago in unicellular organisms. The evolution of the cycles is interdependent; that is, the requirement by all organisms to obtain reduced carbon for energy and growth created a selection pressure. Under moderately reducing conditions of the early earth, processes that led to N₂ fixation and methane production were more favorable than under oxidizing conditions; hence, these biogeochemical processes are among the earliest. Phylogenetic trees, constructed from 16S RNA sequence homology, suggest that the earliest organisms were nonphotosynthetic, thermophilic chemoautotrophs that are placed at the root branch between the Archa and Eubacterial kingdoms. These early organisms could have used inorganic substrates, such as H₂, H₂S, and Fe⁺², to reduce CO₂ to carbohydrate. Indeed, such organisms persist and thrive in deep-sea vents, in volcanic hot springs deep in Earth’s crust, and in other “extreme” environments in which liquid water and suitable oxidizable inorganic substrates are available. Chemoautotrophs are almost certainly the precursors of photosynthetic cells. The evolution of a photosynthetic process in a chemooautotrophic force considered of both the selective forces responsible and the mechanism of evolution.

Reductants for chemoautotrophs are generally deep in the earth’s crust. Vent fluids are produced in magma chambers connected to the atmosphere. Therefore, the supply of vent fluids is virtually unlimited. In the contemporary ocean, the chemical disequilibria between vent fluids and bulk seawater (which is highly oxidized) provides a sufficient thermodynamic gradient to continuously support chemoautotrophic metabolism. However, the redox gradient in early Earth’s
oceans did not exist prior to oxygenic photosynthesis. Moreover, magma chambers, vulcanism, and vent fluid fluxes are tied to tectonic subduction regions, which are transient features of Earth’s crust and hence only temporary habitats for chemoautotrophs. To colonize new vent regions, the chemoautotrophs would need to have been dispersed throughout the oceans by physical mixing. This same dispersion process would have helped ancestral chemoautotrophs exploit solar energy near the ocean surface. The evolution of photosynthetic organisms led to a highly redundant process; that is, many organisms are capable of reducing inorganic carbon to form organic materials utilizing solar energy. The ecological (i.e., “functional”) redundancy ensures a continuity of photosynthetic carbon fixation in all environments in which light and liquid water are available. Currently, there are approximately 20,000 species of aquatic oxygenic photosynthetic organisms belonging to 10 divisions; there are approximately 250,000 species of terrestrial (vascular) plants belonging to 3 divisions (Falkowski and Raven, 1997).

The evolution of oxygenic photosynthesis provided a niche for the co-evolution of ammonium and sulfide-oxidizing bacteria. The action of these organisms led to the chemoautotrophic production of NO\textsubscript{3} and SO\textsubscript{4}\textsuperscript{2-}, which are the most common forms of N and S found in the oceans and most terrestrial ecosystems today. The cycles of these elements are completed under anaerobic conditions, in which the oxidized forms of the elements are used as electron sinks in respiratory pathways. Under many conditions, fixed nitrogen limits primary production in both terrestrial and aquatic ecosystems. In contrast to photosynthesis, N\textsubscript{2} fixation is constrained to a relatively small number of bacteria, and hence the process is “functionally singular.” The lack of diversity among N\textsubscript{2} fixers and the subsequent low supply of fixed N to many ecosystems represents a bottleneck in biogeochemical cycles. In the oceans, N\textsubscript{2} fixers appear to be limited by the availability of Fe and P.

See Also the Following Articles

ATMOSPHERIC GASES • CARBON CYCLE • DIVERSITY, ORGANISM LEVEL • GEOLOGIC TIME, HISTORY OF BIODIVERSITY IN • GREENHOUSE EFFECT • NITROGEN AND NITROGEN CYCLE • THERMOPHILES, ORIGIN OF • VENTS

Bibliography

BIOGEOGRAPHY, OVERVIEW

Mark V. Lomolino
Oklahoma Biological Survey, Oklahoma Natural Heritage Inventory, and
University of Oklahoma

I. Introduction
II. Biogeography in the Twentieth Century
III. Biogeography and the Conservation of Biodiversity

GLOSSARY

biogeography Study of the geographic variation of nature, including variation in any biological characteristics (e.g., body size, population density, or species richness) on a geographic scale.
continental drift Model first proposed by Alfred Wegener that states that the continents were once united and then were displaced over the surface of the globe.
plate tectonics Study of the origin, movement, and destruction of the plates and how these processes have been involved in the evolution of Earth’s crust.
Pleistocene Geologic period from 2 million to 10,000 years before the present, which was characterized by alternating periods of glaciation events and global warming.
species composition Types of species that constitute a given community or sample.
species richness Number of species in a given community or sample.

BIOGEOGRAPHY HAS A LONG AND DISTINGUISHED HISTORY, and one inextricably woven into the historical development of evolutionary biology and ecology. Modern biogeography now includes an impressive diversity of patterns, each of which dealing with some aspect of the spatial variation of nature. Given this, few disciplines can be any more relevant to understanding and conserving biological diversity than biogeography.

I. INTRODUCTION

Traditionally, biogeography has been defined as the study of patterns in distributions of geographic ranges (Brown and Gibson, 1983). During the past three decades, however, this field has experienced a great surge in development and sophistication, and with this development the scope of the field has broadened to include an impressive diversity of patterns. Simply put, modern biogeographers now study nearly all aspects of the “geography of nature.” Biogeography now includes studies of variation in any biological feature (genetic, morphological, behavioral, physiological, demographic, or ecological) across geographic dimensions such as distance among sites or along gradients of area, elevation, or depth (see Brown and Lomolino, 1998).
A. Fundamentals of Biogeography

Despite the sometimes overwhelming complexity of the natural world, all biogeographic patterns ultimately derive from two very general features of nature. First, as we move along any dimension of the geographic template, environmental conditions tend to vary in a predictable manner. For example, more distant sites tend to be more dissimilar than adjacent sites, environments at higher elevations tend to be cooler and wetter than those at lower elevations, and larger areas tend to capture more solar energy and a greater diversity of environmental conditions than smaller areas. Second, all forms of life differ in their abilities to adapt to geographic variation in their environment. These differences, while including a great diversity of responses (e.g., physiological, behavioral, developmental, and evolutionary), ultimately influence the three fundamental processes of biogeography: immigration, extinction, and evolution. All the biogeographic patterns we study derive from nonrandom variation in these processes across geographic gradients and across individuals, populations, and species.

B. Early History of the Field

Biogeography has a long and distinguished history, and one inextricably woven into the historical development of evolutionary biology and ecology. The historical development of biogeography had its origins coincident with the Age of World Explorations by Europeans during the eighteenth and nineteenth centuries. Yet the study of geography of nature must be an ancient one. The European explorers were not the first to ask “Where did life come from, and how did it diversify and spread across the earth?” Aristotle asked these same questions, as did many others before and after him, when faced with accounts of strange forms of life from foreign lands.

The development of biogeography into a mature and respected field of science, however, required a much better understanding of variation in what we now call the geographic template and the associated variation in the natural world. It is by no minor coincidence that both evolutionary biology and biogeography developed in earnest during the Age of Exploration. Prior to this time, biologists had “discovered” and described less than 1% of plant and animal forms that we know today. Each new voyage or expedition added to the accumulated information on the earth’s environments and life-forms, and would eventually provide the raw material for the disciplines of evolution and biogeography. These disciplines are interconnected by the knowledge that selective pressures vary across space, and that all life-forms and their distributions are the product of natural selection.

The early explorers and naturalists did far more than just label and catalog their specimens. They soon, perhaps irresistibly took to the task of comparing their collections across regions, elevations, and other gradients of the geographic template. At the same time, they began to develop explanations for the similarities and differences among the biotas they studied. In fact, most of the persistent themes of the field of biogeography (Table I) were well established during the eighteenth and nineteenth centuries. To be sure, biogeography has made great strides during the twentieth century to become a mature and sophisticated science. It is important to acknowledge, however, that we owe a great deal to the many visionary explorers and naturalists who shared our fascination and asked the same questions about the geographic variation of nature.

1. Historic Explorations of the Eighteenth Century

While motivated to a large degree by the quest for money and power, the Age of European Exploration was also fueled by the call to serve God. It was widely believed that the Creator had placed on this earth a still unknown diversity of organisms—a divine zoo or garden of life. Accordingly, early European explorers believed that there was perhaps no greater way to worship God than to unlock the mysteries of creation. Yet with each new account of some distant biota came information that challenged the prevailing views of creation. Eventually, the growing body of knowledge would overturn the long accepted view that the earth, its climate, and its species were immutable, unchanging in both space and time. More immediately, however, biologists of the eighteenth century were struck by the

<table>
<thead>
<tr>
<th>Table I</th>
<th>Persistent Themes of Biogeography</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Comparing and classifying geographic regions based on their biotas.</td>
</tr>
<tr>
<td>2.</td>
<td>Reconstructing the historical development of biotas, including their origin, spread, and diversification.</td>
</tr>
<tr>
<td>3.</td>
<td>Explaining the differences in numbers as well as types of species among geographic areas.</td>
</tr>
<tr>
<td>4.</td>
<td>Explaining geographic variation in the characteristics of individuals and populations of closely related species, including trends in morphology, behavior, and demography.</td>
</tr>
</tbody>
</table>

astounding diversity of species. Such diversity presented them with two serious problems, one practical and the other conceptual. First, biologists urgently needed a systematic and generally accepted scheme for classifying the burgeoning wealth of specimens, one that would reflect the similarities and differences among the species. Second, it quickly became clear that there were just too many species to be carried by Noah's Ark. How could all the forms of life, now adapted to many distant and distinct regions across the globe, have originated and then spread from that one landing point?

Carolus Linnaeus (1707–1778), certainly one of the most prominent biologists of all time, took on both of these challenges. In fact, his system of binomial nomenclature is the system we continue to use today to classify organisms. Linnaeus also attempted to rectify the Biblical doctrine with what he and his contemporaries knew about the diversity and geography of nature. This was especially challenging because, like most of his colleagues, Linnaeus was sure that species were immutable. Given this, how could species adapted to a single site and climate (Noah's landing) have spread and become perfectly adapted to a suite of different environments (e.g., alpine tundra, coniferous forests, lowland forests, and grasslands)? Linnaeus' answer: Noah's landing had occurred along the slopes of Mount Ararat, a high mountain near the border of Turkey and Armenia. This mountain is so tall (reaching 16,853 ft above sea level) that along its slopes could be found a succession of environments and communities ranging from sub-tropical grasslands at the lower elevations to forests and alpine tundra at its summit. According to Linnaeus' hypothesis, each elevational zone harbored a distinct assemblage of animals, each immutable but perfectly adapted to their local environment. When the Flood finally receded, these animals then dispersed to eventually colonize their respective environments across the globe.

One of the foremost challenges to Linnaeus' views came from his contemporary Comte de Buffon (1707–1788), who believed that not only were climates mutable, but species were as well. How else could animals disperse across what are now inhospitable habitats to occupy their present ranges in such isolated regions of the globe? Buffon's theory of the origin and spread of life stemmed largely from his studies of living and fossil mammals, especially those of the Old and New World tropics. He was the first to realize that different regions of the globe, even those with the same environmental conditions, had distinct biotas. This observation was so fundamental that it eventually became biogeography's first law: Buffon's Law.

Like Linnaeus, Buffon also concluded that there was one "landing point," one site where all animals originated. However, this site, or region, was located far to the north of Mount Ararat, somewhere in the Arctic Circle where the early animals and their descendants could gain ready access to both the Old and New Worlds. This is where these life-forms survived the Flood during some earlier period when the earth's climate was much warmer, warm enough such that tropical environments could extend far poleward. Once the floods receded, animals spread southward into the continents and began to diverge in form as they became increasingly isolated on different landmasses.

Other biologists of the eighteenth century, including Joseph Banks and Johan Reinhold Forster, both of whom served as naturalist on voyages of Captain James Cook, were quick to confirm the generality of Buffon's Law: environmentally similar but isolated regions have distinct assemblages of plants and animals. Forster also discussed the relationship between regional floras and environmental conditions and, in turn, between plant and animal associations: two cornerstones of the field now known as ecology. Forster was also one of the first scientists to report that plant diversity increases as we move toward the equator, that islands have fewer plants than the mainland, and that the diversity of insular plants increases with island size and available resources. Later in the eighteenth century, Karl Wildenow (1765–1812) and one of his students, Alexander von Humboldt (1769–1859), confirmed and further generalized both Buffon's Law and Forster's. Toward the end of the century, Augustin P. de Candolle added the important insight that, not only is the distribution of organisms influenced by geographic variation in environments, but they also compete for limiting resources such as food, light, and water.

Therefore, by the beginning of the nineteenth century, biogeographers already had their first "law," they described and tested the generality of a number of related patterns about the geography of nature, and they offered some testable theories regarding those patterns. They were actively working on at least three of the four persistent themes of biogeography (Themes 1–3 in Table I). A number of biogeographers had abandoned the notion that species and climates were immutable. But for the field to advance and become a mature science, two additional, fundamental insights were needed. First, it required a mechanism for the mutability and adaptation of species. As many of us realize, Candolle's observations about competition and the struggle for existence were central to the development of the theory of evolution by natural selection.
These advances were to come in the latter part of the nineteenth century. Second, scientists had to recognize that the geographic template (i.e., the foundation for all of these patterns) also was mutable. That is, the size and relative positions of the continents and ocean basins have changed throughout the history of our planet.

2. Advances of the Nineteenth Century

Many of the most fundamental advances in biogeography, and evolutionary biology as well, have required advances in geology. Until the nineteenth century, the age of the earth was typically assumed to be just a few thousand years, way too brief to allow what we now know to be the requisite time for evolution of its plates and the species that have rafted on those plates. The collective work of nineteenth-century paleobiologists would push the age of the earth back hundreds of thousands and eventually millions of years before the present. Legendary geologists and paleobiologists such as George Lyell (1797–1875) and Adolphe Brongniart (1801–1876), through their studies of fossils, provided incontrovertible evidence for extinction and for changes in regional climate and the elevation of land. How else could they explain the existence of fossils that have no contemporary forms, of fossils from tropical species found in regions that are now temperate, and of shells and other marine fossils on present-day mountains? A theory of floating and drifting continents (now known as plate tectonics) would await discoveries of twentieth-century marine geologists, but their nineteenth-century forerunners understood that the earth was very old.
indeed, and it was mutable. Furthermore, if species (and many thousands of them) went extinct, then there had to be multiple periods of creation (or evolution) to compensate for those losses.

Again, these views of a mutable earth, mutable climate, and mutable species were essential for those attempting to classify biogeographic regions based on their respective assemblage of species (Theme 1), to reconstruct the origin, spread, and diversification of life (Theme 2), and to explain differences in numbers and types of species among geographic regions (Theme 3). However, well into the middle of the nineteenth century many scientists held stubbornly to the idea that not only were species immutable, but so were their distributions. Perhaps the most distinguished champion of this static view of biogeography was Louis Agassiz (1807–1873), who argued that species remain essentially unchanged at or near their sites of creation.

The static view was eventually overturned by the passionate and persuasive arguments of no one less than Charles Darwin. Not only did he propose a general theory for the diversification and adaptation of biotas (i.e., natural selection), but he was one of the world's first and foremost dispersalists and champions of dynamic biogeography. Through his observations during his circumnavigation of the globe on the *HMS Beagle* (1831–1836), his later experiments on dispersal of seeds by animals, and his general synthesis on the origin and distribution of life, Darwin convinced many of his colleagues that long-distance dispersal could account for many of the otherwise perplexing patterns of biogeography. Once he was joined by the likes of Asa Gray and Alfred Russell Wallace, Darwin and his colleagues were able to pull off a major paradigm shift in the field—from the static view of the earth and its species to the dynamic view of biogeography.
Yet among all of these visionary scientists it is Wal-
lace who is generally recognized as the father of zooge-
ography, and biogeography in general. While Darwin
argued passionately regarding long-distance dispersal
(even to the point of soundly criticizing his mentor,
Charles Lyell), most of his energies were devoted to-
ward developing and substantiating his theory of natu-
ral selection. On the other hand, biogeography was
17 tenets of the field that were developed by Wallace
and included in his seminal monographs *The Malay
Archipelago* (published in 1869 and dedicated to Dar-
win), *The Geographic Distribution of Animals* (1876),
and *Island Life* (1880). Five of these tenets of biogeogra-
phy are listed here:

1. Climate has a strong effect on the taxonomic simi-
larity between two regions, but the relationship is
not always linear.
2. The present biota of an area is strongly influenced
by the last series of geological and climatic events.
3. Competition, predation, and other biotic factors
play determining roles in the distribution, dis-
persal, and extinction of animals and plants.
4. When two large landmasses are united after a long
period of separation, extinctions may occur be-
cause many organisms will encounter new compet-
titors.
5. To analyze the biota of any particular region, one
must determine the distributions of its organisms
beyond that region as well as the distributions of
their closest relatives.

Using the latter approach and information provided
by over a century of naturalists, Wallace developed a
scheme of biogeographic regions (Fig. 1) that accurately
reflected the similarities and differences among biotas.
This same scheme, largely unchanged, is still used
today.

For obvious reasons, exploration and biogeographic
study of the marine realm have always lagged far behind
that of terrestrial systems. Yet by the middle of the
nineteenth century, biogeographers had made some sig-
ificant strides in studying this new frontier. Charles
Lyell discussed patterns of distribution of marine algae
in his seminal work *Principles of Geology,* first published
in 1830. Edward Forbes wrote the first comprehensive
monograph on marine biogeography in 1836, in which
he divided the marine realm into zoogeographic regions
based on latitude, depth, and animal assemblages. In
1897 the great British ornithologist and biogeographer
Philip Sclater, who produced a predecessor to Wallace's
biogeographic scheme, also developed a scheme for the
marine realm based on distributions of marine mam-
mals. Following the lead of earlier biogeographers and
also based on his own extensive field studies in south-
western North America, C. Hart Merriam (1894) devel-
oped a system of what he termed "life zones" that con-
firmed earlier observations that elevational changes in
vegetation were equivalent to those along latitudinal
gradients.

Finally, the countless specimens collected during the
late eighteenth and early nineteenth centuries enabled
others to begin to analyze geographic variation in char-
acteristics of individuals and populations (Theme 4).
C. L. Gloger reported in 1833 that, within a species,
individuals from more humid habitats tend to be darker
than those from drier habitats (Gloger's Rule). C. Berg-
mann (1847) found that in birds and mammals, popula-
tions from cooler environments tended to have larger
bodies than those from warmer environments (Berg-
mann's Rule). Also, J. A. Allen reported in 1878 that
birds and mammals inhabiting cooler environments
also tend to have shorter appendages (Allens' Rule).

II. BIOGEOGRAPHY IN THE
TWENTIETH CENTURY

A. Dynamics of the Geographic Template

Even the earliest human explorers appreciated the fact
that abiotic conditions vary as one moves from one
point on the globe to another. On land, precipitation,
temperature, seasonality, prevailing winds, soil condi-
tions, and a host of other important factors vary as we
move along transects of latitude, longitude, or altitude.
Similarly, in the aquatic realm, temperature, currents,
pressure, solar radiation, and concentrations of oxygen
and dissolved nutrients vary markedly within and
among ecosystems. Together, the variation in all of
these environmental characteristics combine to form
the geographic template, which influences all bioge-
ographic patterns.

Although a complete understanding of all aspects of
the geographic template may be a daunting and truly
impossible challenge, at a regional to global scale, geo-
graphic variation in environmental conditions is quite
regular and interpretable. On land, climatic conditions
vary in an orderly manner with latitude, elevation, and
proximity to mountain ranges or oceans (Fig. 2). Major
soil types (Fig. 3) also vary in a similar fashion, partially
because soil development is strongly influenced by local
climatic conditions, especially precipitation and tem-
FIGURE 2 Major climatic regions of the world. Note that these regions occur in distinct patterns with respect to latitude and the positions of continents, oceans, and mountain ranges. (After Strahler, 1973.)

As we shall see in later sections, such regular variation in environmental characteristics translates into nonrandom variation in biogeographic patterns of organisms, with each one adapted to slightly different environmental conditions. Such adaptations are, of course, the product of a long and complex evolutionary history: a series of innumerable interactions between organisms and their environments. With each successive generation, the abilities of descendants to respond and adapt to local environmental conditions change. Evolutionary change, however, is part of a never-ending battle because environmental conditions include other species, which are also evolving. Just as important, the geographic template has evolved throughout earth’s 4.5-billion-year history. Because species distributions and other aspects of their geographic variation are influenced by their interaction with the geographic template, a thorough understanding of its dynamics in space and time is essential if we are to understand any major biogeographic patterns.

By the opening of the twentieth century, each of the four persistent themes of biogeography was well established. Explanations for the major biogeographic patterns could now draw on insights from the rapidly growing field of evolutionary biology, as well as our knowledge of the other two fundamental biogeographic processes—immigration and extinction. In addition, biogeographers of the early twentieth century could tap a great wealth of information on geographic variation of biotas and of the environments that they inhabited. Obviously, a thorough knowledge of this underlying geographic template was essential for understanding patterns in distribution and variation among regions and isolated ecosystems. Yet to develop a more accurate and more comprehensive understanding of the major patterns and processes of biogeography, another major scientific revolution was required.

Biogeographers and most other natural scientists knew a great deal about contemporary environments,
and most of them appreciated the fact that climatic conditions had changed, sometimes dramatically, during earlier periods of earth’s history. Yet until the 1960s, most biogeographers clung to the belief that earth’s landforms and ocean basins remained fixed. During the twentieth century, acceptance of the theory of continental drift and plate tectonics revolutionized the field of biogeography as much as acceptance of the theory of natural selection and evolution had in the previous century.

1. Continental Drift and Plate Tectonics

Although imperceptible to most of us, the earth’s continents have moved, colliding at times and drifting apart at others: mountain ranges have formed and eroded away, seas have expanded and contracted, and islands have appeared and disappeared. These changes must have had profound effects on local and regional climates and, in turn, on the geographic distributions and variations of all forms of life on earth. As we will see in a subsequent section, the theory of continental drift and plate tectonics is a relatively recent advance. Yet, with the possible exception of the acceptance of the theory of natural selection, no other contribution has had more of an impact on the field of biogeography.

Plate tectonics is defined as the study of the origin, movement, and destruction of the earth’s plates and how these processes have been involved in the evolution of the earth’s crust. The theory of plate tectonics has achieved general acceptance among nearly all scientists and reigns as a unifying paradigm of both geology and biogeography. Yet until just three decades ago, relatively late in the development of these fields, champions of this theory were viewed as oddballs and heretics.

As with any other revolutionary theory in science, it is extremely difficult to pinpoint the origins of the theory of plate tectonics. The great geologist Charles Lyell entertained the idea during the 1830s and 1840s, but then abandoned it in favor of the accepted doctrine of the fixity of the continents and ocean basins. In their attempts to explain the affinities of biotas of now isolated continents, Lyell, Joseph Dalton Hooker,
and other “extensionists” of the nineteenth century hypothesized the periodic emergence of great land bridges that then allowed biotic exchange. Darwin, Wallace, and other members of the dispersal camp soundly criticized such views: nothing vexed Darwin more than those extensionists who created land bridges “as easy as a cook does pancakes.” In an uncharacteristically critical passage in one of his letters, Darwin complained to Charles Lyell of “the geological strides which many of your disciples are taking.... I fear you will not stop this, if there be a lower region of punishment for geologists, I believe, my great master, you will go there.”

As it turns out, neither the extensionists nor the dispersalists were correct. In most cases, the similarities among now isolated biotas were instead the result of “dispersal” of the continents themselves. Perhaps the first important evidence for what was at first referred to as the theory of continental drift was the configuration of the continents. That is, once geographers had developed relatively accurate maps, it became clear to some that opposite coastlines seemed to fit. In 1858 one of Lyell’s contemporaries, Antonio Snider-Pelligrini, may have been the first to demonstrate the geometric fit of the coastlines of continents on opposite sides of the Atlantic Ocean. Yet it wasn’t until 1908 and 1910 that an American geologist, F. B. Taylor, and a German meteorologist, Alfred L. Wegener, independently developed models describing the movements of the earth’s crust, along with the formation of mountain chains, island arcs, and related geologic features. Wegener continued to develop his model into a more comprehensive theory of continental drift, publishing his treatise in the 1920s. Wegener’s theory, however, included too many assumptions about geologic processes and patterns that would not be well established for another three or four decades. His theory also included factual errors, such as overestimating the rate of movement of the earth’s plates by perhaps two orders of magnitude. Finally, although he speculated on a potential mechanism, Wegener’s theory really lacked a plausible one that could somehow drive the massive plates...
about the earth like bits of ice on a pond in spring. It is perhaps one of History's most tragic ironies that, in his quest to discover this mechanism by exploring a volcanically active region of Greenland, Wegener perished in a snow storm.

Wegener's insights would not be widely appreciated for another three decades. Acceptance of the theory of continental drift and its maturation to become the more comprehensive theory of plate tectonics would require many additional insights from geographers, paleontologists, and especially marine geologists during the 1940s and 1950s. These scientists found that when they "rejoined" the continents based on their geometric fit, not only did their biotas seem to match up, but so did topographic features such as mountain chains, rock strata, and fossil and glacial deposits. Perhaps most critical to the acceptance of the theory of continental drift were the efforts by marine geologists following World War II to map the surface of the ocean basins. It soon became clear that beneath each ocean lay a system of ridges that were situated far offshore. As one moved away from these ridges, the seafloor became deeper and more ancient as well. Provided with these related clues, Herman Hess and his colleagues developed the theory of seafloor spreading: continental drift finally had an underlying mechanism (Fig. 5).

Eventually, paleomagnetic evidence would allow marine geologists to estimate the previous positions of the continents and develop reconstructions of the sequences of their movements and creation and the dissolution of previous continents. Biogeographers were now armed with not just the evidence, but also the mechanisms responsible for the dynamics of biotas and the geographic template itself (i.e., immigration, extinction, evolution, and plate tectonics).

2. Glacial Cycles of the Pleistocene

The great shifting, collision, and separation of earth's plates profoundly affected the distribution of its biota, both directly and indirectly. Not only did plate tectonics alter major dispersal routes among biotas, but it substantially changed both global and regional climates. As plates shifted across different latitudes, their local biota was exposed to major shifts in climatic conditions. Areas of what is now tropical Africa, South America, and Australia once were situated over the south pole and exposed to severe antarctic climates.

On a global scale, drifting continents also triggered great shifts from periods of global warming to those dominated by glacial conditions. Land absorbs substantially more solar energy than does water. Thus, global climates tended to be warmer when landmasses were situated near the equator, but cooled as they shifted poleward.

Yet global climates can change substantially even during periods too short for substantial shifting of earth's plates. For example, during the Pleistocene (roughly the past 2 million years), earth experienced many climatic upheavals. Rather than being caused by any shifts in plates (which must have been minor given the relatively short period), these climatic shifts were caused by periodic changes in characteristics of the earth's orbit (referred to as Milankovitch cycles; Fig. 6). These changes significantly altered the total amount of solar energy intercepted by the earth, ultimately causing the climatic reversals of the Pleistocene. During full glacial periods, global temperatures dropped by as much as 6°C and most landmasses beyond 45° latitude were covered with glaciers often 2 to 3 km thick. Because so much water was tied up in the glaciers, sea levels dropped by 100 to 200 m, thus uniting long-isolated biotas via temporary land bridges. For example, during the last glacial maximum, the region of Southeast Asia and Malaysia was united with Sumatra, Java, and Borneo to form Greater Sunda, while Australia and New Guinea formed the island continent of Sahul (Fig. 7).

Winds, ocean currents, and precipitation patterns also changed substantially between interglacial and glacial periods. With each climatic upheaval, environmental regimes shifted across both latitudes and elevations. Regions such as the American Southwest, which is now dominated by desert and xeric grasslands, were once covered with coniferous forests. Warming and drying conditions that led to the current interglacial period dramatically reduced these forests and caused them to shift toward the mountain peaks, where cool and relatively humid conditions prevail.

These and other events must have profoundly influenced the distributions of most if not all biotas. As Brown and Lomolino (1998) summarize, however, all the complex biogeographic dynamics of the Pleistocene were triggered by three fundamental changes in the geographic template:

1. Changes in the location, extent, and configuration of principal habitats.
2. Changes in the nature of environmental regimes (combinations of temperature, seasonality, precipitation, and soil conditions).
3. The creation and dissolution of barriers associated with changes in sea level or elevational shifts in habitats.
FIGURE 5  (A) During seafloor spreading, reversals in the earth's magnetic field are recorded as the magnetically sensitive, iron-rich crust cools. Differences in the widths of the magnetic stripes reveal differences in the duration of these polarity episodes and in the rate of seafloor spreading over time and among regions. (From Stanley, 1987.)  (B) The current model of plate tectonics includes the possibility that at least three forces may be responsible for crustal movements: (1) ridge push, or the force generated by molten rock rising from the earth's core through the mantle at the midoceanic ridges; (2) mantle drag, the tendency of the crust to ride the mantle much like boxes on a conveyor belt; and (3) slab pull, the force generated as subducting crust tends to pull trailing crust after it along the surface. (After Stanley, 1987.)

The responses of both terrestrial and aquatic biotas, while no doubt complex, also were of three types:

1. Some species shifted geographically with their optimal habitats.
2. Some species remained and adapted to the altered local environment.
3. Other species, unable to modify their ranges or ecological associations, suffered range contraction and eventual extinction.

The biogeographic dynamics of the Pleistocene remains one of the field's most active and interesting study areas. Recent advances in analyzing and dating fossil material continue to add to our ability to reconstruct the historical development of biotas (Theme 2).
FIGURE 6: Milankovitch cycles are periodic changes in the eccentricity, obliquity, and precession of the earth's orbit. Each of these changes influences the earth's interception of solar radiation; therefore, these cycles may have been largely responsible for the glacial cycles of the Pleistocene. (After Gates, 1993.)
FIGURE 7 The lowering of sea levels during glacial maxima of the Pleistocene caused the exposure of continental shelves and the formation of dispersal routes across four regions of the eastern Pacific: Sunda, Wallacea, Sahul, and Oceania. (White areas = land exposed during glacial maxima; dark shading = deep water (>200 m); possible dispersal routes are indicated by arrows). (After Fagan, 1990; Guilaine, 1991.)

and, in turn, understand major episodes of biotic inter-change and recent extinctions, as well as the current distributions of species.

B. Current Trends in Biogeography

1. Gradients in Species Diversity and Composition

In addition to biogeographic reconstructions, modern biogeographers continue to study an impressive diversity of patterns encompassing each of the four persistent themes of the field. During the middle of the twentieth century, many biogeographers focused on more general questions. Rather than dissecting and reconstructing the range of selected species, they examined trends in the total number of species, or what is often termed species richness. Though many patterns in richness have been studied, two have received the lion’s share of attention: the species–area and species–latitude relationships. Early explorers and naturalist of the seventeenth and eighteenth
centuries noted the tendencies for species richness to increase with area of a region or island, and be higher for tropical versus temperate, subarctic, and arctic biotas. Armed with data from many hundreds of additional surveys and with a battery of sophisticated statistical tools, twentieth-century biogeographers confirmed the great generality of these patterns and developed some relatively simple models to explain those patterns. Often these models were equilibrial, assuming that species richness resulted from the combined but opposing effects of processes such as immigration into an area (which added species) and extinction (which decreased species richness). MacArthur and Wilson’s equilibrium theory of island biogeography is perhaps the prototypic example of such a theory, and one that has dominated the field since its first articulation in the 1960s. Their theory was developed to explain both the species–area relationship and the species–isolation relationship (i.e., the tendency for species richness to decrease as one moves from near to more isolated islands). Simply stated, because immigration rates (the number of species new to an island) should decrease while extinction rate (loss of species already present) should increase as the island accumulates species, the island should eventually reach a level of richness at which immigrations balance extinctions. This equilibrrial level of richness should vary among islands: decreasing with isolation because immigration rates are lower for more isolated islands, and increasing with island area because populations on larger islands should be less prone to extinction.

2. Biogeography in the Twenty-first Century

The equilibrium theory stimulated many studies in biogeography and related fields of ecology, and has served as the paradigm of island biogeography for some four decades. Yet an increasing number of biogeographers are beginning to question its utility as a modern paradigm. Species richness is often influenced by speciation and disturbances (e.g., major storms and tectonic events), processes not included in MacArthur and Wilson’s original theory. Either the theory has to be expanded to include these processes, or it will be replaced by an alternative model—one that may eventually become the new paradigm of the field.

Whatever form such a model takes, it must be sophisticated enough to address the growing complexity of questions and patterns that we now study. MacArthur and Wilson’s model was primarily developed to explain patterns in richness along gradients of area and isolation. Yet modern biogeographers are now searching for a theory to explain patterns across other geographic gradients and, perhaps more important, to explain geographic trends in the types rather than just the numbers of species. Why do the proportions of large versus small, endothermic versus ectothermic, herbivorous versus carnivorous animals, woody versus herbaceous, or annual versus perennial plant species vary across geographic gradients? How are biogeographic reconstructions related to phylogenies? In what manner do gene frequencies vary with isolation or across other geographic gradients? How do the size and shape of geographic ranges vary with latitude and among taxonomic groups, and how do population density and other demographic parameters vary across the range of a species?

We may have reached a point at which our questions and appreciation for the complexity of nature have become too sophisticated for the relatively simple models that have dominated the field since the 1960s. If this is true, biogeography may be on the verge of a major scientific revolution, one that may well rival those triggered by the seminal insights of scientists such as Charles Darwin, Alfred Russel Wallace, Alfred Wegener, Robert H. MacArthur, and E. O. Wilson.

III. BIOGEOGRAPHY AND THE CONSERVATION OF BIODIVERSITY

Biogeographers study both the patterns and processes influencing the geographic variation of nature. We study not just how many species occur in a particular area, but why more are there than somewhere else and which ones are likely to be shared among areas. We study and attempt to develop explanations for what are now termed “hotspots,” regions of relatively high numbers and high endemicity of species. Biogeographers also study variation in the geographic template, including that associated with anthropogenic disturbances such as the spread of exotic species or the spatial patterns of deforestation. Many biogeographers study extinction and have demonstrated that it has a geographic signature: loss of species tends to be highest for the smallest and most isolated sites, namely, oceanic islands and fragments of once expansive habitats on the mainland. Given this, it becomes obvious that few disciplines could be any more relevant to understanding and preserving biological diversity than biogeography.

Our task, however, is far from a simple exercise of just applying what we already know. Indeed, only a small fraction (perhaps just 2 or 3%) of all extant species have been described, and we know precious little about...
the geographic distributions of most of those. What we do know often comes down to just general patterns for common species, but conservation biologists require detailed information on the rare species—those that may be the exceptions to most rules.

We have, however, made great progress in recent years in mapping and measuring the intensity (number of endemic species) of hotspots of biological diversity. In theory, these hotspots of biodiversity should receive the highest priority from conservation biologists, especially when they coincide with high levels of human activity. Yet even these approaches, generated by existing survey information and sophisticated geographic analyses, are based on a relatively limited number of surveys.

To develop more effective strategies for conserving global diversity, we still require a much more thorough understanding of the geographic variation of nature. A number of distinguished ecologists and biogeographers, including E. O. Wilson, have called for greatly accelerated efforts to map the diversity of life. With adequate support, within just a few decades we could greatly expand our knowledge of the distributions of most life-forms and, eventually, contribute to their conservation as well.

See Also the Following Articles

Bibliography

I. Bioprospecting: A Tool for Survival and a Source of Inspiration and Innovation

II. Economic Value and Benefit Sharing

III. Concepts and Practices of Bioprospecting: The Case of Costa Rica

GLOSSARY

bionics Science of systems, which function in the same way as or similar to living systems.

biopiracy Illegal appropriation or exploitation of genetic and biochemical resources.

chemical prospecting Search for natural compounds in wild living plants, animals, and microorganisms with a potential for the development of chemical products like pharmaceuticals, pesticides, cosmetics, or food additives.

combinatorial chemistry Laboratory methods to produce all possible combinations from various sets of chemical building blocks in a short period of time and to generate molecular diversity for the screening of new bioactive compounds.

gene prospecting Search for genes in wild living plants, animals, and microorganisms for the breeding or genetic engineering of plants, animals, and microorganisms in agriculture, fermentation, and cell culture for agricultural and industrial production.

INBio The Instituto Nacional de Biodiversidad (National Biodiversity Institute), a Costa Rican association created in 1989 to generate and disseminate knowledge and promote sustainable uses of biodiversity.

BIOPROSPECTING IS THE SYSTEMATIC SEARCH for genes, natural compounds, designs, and whole organisms in wild life with a potential for product development by biological observation and biophysical, biochemical, and genetic methods, without disruption to nature.

I. BIOPROSPECTING: A TOOL FOR SURVIVAL AND A SOURCE OF INSPIRATION AND INNOVATION

A. Traditional Bioprospecting

In 1991 a 5300-year-old mummy was discovered in the Tyrolean Alps. The “Ice Man” and his effects were extremely well preserved and his clothing and equipment were perfectly functional for the survival under the harsh climatic conditions of the Alps. They consisted of plant and animal fibers, oils, and waxes, put together by a highly skilled and sophisticated Neolithic culture. The “Ice Man” suffered from an intestinal endoparasite, the whipworm Trichuris trichiura, but he was...
already equipped with the corresponding medicine, the fruiting body of Piptoporus betulinus, which contains oils that are toxic to parasites and compounds that act as strong laxatives that would cause expulsion of the dead and dying worms and their eggs.

The search for and utilization of biological resources is as old as humankind and has been key to the survival, adaptation, and evolution of the human species. The perceptions and values of communities and individuals depending on and living in close proximity to wild and domesticated biodiversity are clearly different from those of the modern scientist engaged in bioprospecting. To many communities, biodiversity entails magic, religious, and ceremonial connotations in addition to the more mundane food, shelter, and medicine benefits derived from its domestication and utilization.

The forest has provided nourishment and shelter to many cultures through a myriad of products and services since early times. The Neotropical forests harbor the wild ancestors of such major food crops as manioc, cacao, and yams. Evidence of squash (Cucurbita moschata) farming in southwestern Ecuador over 10,000 years ago suggests that agriculture in the Americas may have begun in the rain forests (Piperno and Pearsall, 1998). Low population pressure and markets limited to exchange and barter enabled the original indigenous groups to lead a low-impact lifestyle. Hunting and gathering, plus domestication of common beans, maize, cacao, chilies, squash, and other crops, permitted not only the survival of autochthonous groups but also their social and cultural development, and, quite significantly, with a minimal impact on the natural environment. The arrival of Europeans in the Americas brought along agricultural practices that were to a great extent destructive to Neotropical ecosystems. These new farmers felt compelled to "conquer" nature and cleared increasingly large areas of the forest to plant subsistence crops. Both systems have survived to the present, with the latter practices and large-scale, chemical-intensive agriculture being dominant because of increased population pressure and the need for higher agricultural productivity.

Costa Ricans, like other inhabitants of Middle America, have used and continue to use medicinal plants (e.g., roots of Cephaelis ipecacuanha and Smilax spp.), fruits and nuts, palms, birds, animals, limes (for hunting pleasure and to protect domestic animals), ornamental plants, feathers, and building materials from the forest. The impacts of these practices, particularly logging over the last 50 years, have been significant in altering the landscape and call into question the tropical region's capacity for sustainable development.

In recent times, agronomists and professionals from related sciences have done considerable agricultural prospecting looking for landraces having agronomic advantages, particularly higher yields, pest and disease resistance, and adaptation to specific microhabitats. Gene prospecting of the resources managed, nurtured, and selected over centuries by indigenous communities has yielded considerable benefits to the world in terms of improved crop varieties and hybrids. Traditional plant breeding has slowed down in tropical countries due to lower research investment and unsuitable intellectual property protection, and this could diminish the opportunity of using the tools of the new biotechnologies.

Agricultural prospecting has resulted in the assembling of government, international, and private gene collections; however, the viability of these resources, mostly in tropical countries, has decreased over time. Some estimates suggest that over 70% of all germplasm collected since 1940 is no longer viable or the associated knowledge has been lost. In situ conservation, therefore, has become an attractive proposition to protect and utilize the remaining genetic resources.

Even plant breeding no longer depends only on genes from wild types and ancestors of modern crops; new improved varieties can be artificially generated through genetic engineering, for example, with designer genes that confer resistance to pests. Motifs in the amino acid sequences of the bactericidal and fungicidal magainin peptides from frog skin inspired genetic engineers to design resistance genes, which have scarcely anything in common with their natural ancestors (Nader and Hill, 1999).

B. Modern Bioprospecting

Modern science and technology have provided substitutes for some biological products: nylon, polyacryl, and polyester instead of cotton and wool; aniline dyes instead of natural colorants; and organophosphates or carbamates instead of plant-derived natural pesticides like rotenone, nicotine, and ryanodine. Yet human ingenuity still depends in great part on inspirations from wild life-forms to create these substitutes. In contrast to our forest-dwelling bioprospecting ancestors, we no longer need to use the biological original to create a final product; rather, through the study of these originals we are able to discover and unveil the principle of a biological function and then develop novel products on this basis.

Millions of years of evolution have created a wealth of structures and mechanisms at the molecular, cellular,
and macrostructural level, all of which function economically and interact to perfection. Nature provides solutions to most of life’s technical problems. Natural selection has imposed on living organisms the Min-Max Principle: a minimum of material and energy accomplishes a maximum of efficiency and stability. This makes biological prototypes particularly important for our future given the world’s decreasing resources and increasing environmental problems (Hill, 1997).

Box 1
Biodiversity offers three fundamental sources of inspiration to the modern scientist: chemicals, genes, and designs. Fields of applications include drug development, agrochemistry, and cosmetics (chemicals), development of recombinant pharmaceutical proteins, enzymes, and agricultural biotechnology (genes), and architecture, mechanical engineering, and sensor technology (designs) (Nader and Hill, 1999).

An example of how nature inspires inventors is the development of the first glider airplane. By analyzing the flight of the stork (the biological prototype), the principle of airlift was discovered and the functional principle of flight was reduced to the vault of the wings. Vaulted wings derived from nature became the basic structural feature of all airplanes (Hill, 1997). Given the three inspirational resources from living nature—chemicals, genes, and designs—bioprospecting can be differentiated into chemical, gene, and bionic prospecting (Fig. 1). To discover new biological prototypes, bioprospectors can evaluate traditional uses of plants, animals, and microorganisms, use random sampling, or, in the biorational approach, systematically analyze biological phenomena (Tamayo et al., 1997).

1. Chemical Prospecting
The major defense mechanisms of plants against herbivores rely on chemicals: “The world is not colored green to the herbivore’s eyes, but rather is painted morphine, L-DOPA, calcium oxalate, cannabinol, caffeine, mustard oil, strychnine, rotenone, etc.” (Janzen, 1975). The co-evolution of herbivores and their feeding plants has created diversity on both the species and the molecular levels (Ehrlich and Raven, 1964). Communication, competition, sexual attraction or rejection, and pollination are also based to a great degree on chemistry and have contributed to the development of diversity. To help conserve this chemical wealth for future generations, Thomas Eisner developed the concept of chemical prospecting, a collaborative effort among conservationists, scientists, the pharmaceutical industry, and biodi-
Although human ingenuity has created a huge variety of chemical compounds with exceptional pharmaceutical activity without inspiration from nature—like the anxiolytic benzodiazepines valium and librium—natural compounds and their derivatives are still of central importance in drug discovery. Table I lists the 30 best-selling pharmaceuticals (drugs and biologics) in 1997. Thirteen were either derived from or developed as the result of leads generated by natural products from wild plants, animals, and microorganisms. The antihypertensive proline derivatives enalapril (Vasotec) and lisinopril (Zestril) were derived from a peptide in the venom of the fer-de-lance (Bothrops jararaca or B. atrox). Diclofenac (Voltaren) belongs, like aspirin, to the group of nonsteroidal anti-inflammatory drugs, which were derived from the lead structure salicin from willow bark (Salix spp.). The antineoplastic paclitaxel (Taxol) was discovered in the screening program of the National Cancer Institute from the bark of the Pacific yew tree (Taxus brevifolia). The antiviral drug acyclovir (Zovirax) was developed using prior knowledge of cyto-

### Table 1

<table>
<thead>
<tr>
<th>Rank</th>
<th>Product</th>
<th>Product category</th>
<th>Company</th>
<th>U.S. $ (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Losec/Prilosec</td>
<td>Proton pump inhibitor</td>
<td>Astra/Astra Merck</td>
<td>3749</td>
</tr>
<tr>
<td>2</td>
<td>Zocor</td>
<td>Hypolipidemic</td>
<td>Merck &amp; Co.</td>
<td>3575</td>
</tr>
<tr>
<td>3</td>
<td>Prozac</td>
<td>Antidepressant</td>
<td>Eli Lilly</td>
<td>2539</td>
</tr>
<tr>
<td>4</td>
<td>Vasotec</td>
<td>ACE inhibitor</td>
<td>Merck &amp; Co.</td>
<td>2510</td>
</tr>
<tr>
<td>5</td>
<td>Zantac</td>
<td>H2-antagonist</td>
<td>Glaxo Wellcome</td>
<td>2254</td>
</tr>
<tr>
<td>6</td>
<td>Norvasc</td>
<td>Calcium antagonist</td>
<td>Pfizer</td>
<td>2217</td>
</tr>
<tr>
<td>7</td>
<td>Claritin</td>
<td>Antihistamine</td>
<td>Schering-Plough</td>
<td>1726</td>
</tr>
<tr>
<td>8</td>
<td>Augmentin</td>
<td>Penicillin antibiotic</td>
<td>SmithKline Beecham</td>
<td>1516</td>
</tr>
<tr>
<td>9</td>
<td>Zolofl</td>
<td>Antidepressant</td>
<td>Pfizer</td>
<td>1507</td>
</tr>
<tr>
<td>10</td>
<td>Seroxat/Paxil</td>
<td>Antidepressant</td>
<td>SmithKline Beecham</td>
<td>1474</td>
</tr>
<tr>
<td>11</td>
<td>Pravachol</td>
<td>Hypolipidemic</td>
<td>Bristol-Myers Squibb</td>
<td>1437</td>
</tr>
<tr>
<td>12</td>
<td>Ciprofloxacin</td>
<td>Quinolone antibiotic</td>
<td>Bayer</td>
<td>1378</td>
</tr>
<tr>
<td>13</td>
<td>Premarin</td>
<td>Hormonal product</td>
<td>American Home Products</td>
<td>1282</td>
</tr>
<tr>
<td>14</td>
<td>Mevacor</td>
<td>Hypolipidemic</td>
<td>Sankyo</td>
<td>1211</td>
</tr>
<tr>
<td>15</td>
<td>Procardia</td>
<td>Macrolide antibiotic</td>
<td>Abbott</td>
<td>1200</td>
</tr>
<tr>
<td>16</td>
<td>Novolin</td>
<td>Insulin</td>
<td>Novo Nordisk</td>
<td>1200</td>
</tr>
<tr>
<td>17</td>
<td>Pepcid</td>
<td>H2-antagonist</td>
<td>Merck &amp; Co.</td>
<td>1180</td>
</tr>
<tr>
<td>18</td>
<td>Sandimmune/Neoral</td>
<td>Immunosuppressant</td>
<td>Novartis</td>
<td>1166</td>
</tr>
<tr>
<td>19</td>
<td>Epogen</td>
<td>Antianemic</td>
<td>Angen</td>
<td>1163</td>
</tr>
<tr>
<td>20</td>
<td>Mercuror</td>
<td>Hypolipidemic</td>
<td>Merck &amp; Co.</td>
<td>1100</td>
</tr>
<tr>
<td>21</td>
<td>Imigran</td>
<td>Antimycosome</td>
<td>Glaxo Wellcome</td>
<td>1093</td>
</tr>
<tr>
<td>22</td>
<td>Procain/Eprex</td>
<td>Antianemic</td>
<td>Johnson &amp; Johnson</td>
<td>1070</td>
</tr>
<tr>
<td>23</td>
<td>Neupogen</td>
<td>Colony stimulating factor</td>
<td>Angen</td>
<td>1068</td>
</tr>
<tr>
<td>24</td>
<td>Adali</td>
<td>Calcium antagonist</td>
<td>Bayer</td>
<td>1063</td>
</tr>
<tr>
<td>25</td>
<td>Receptin</td>
<td>Cephalosporin</td>
<td>Roche</td>
<td>1062</td>
</tr>
<tr>
<td>26</td>
<td>Pralocal</td>
<td>Gastroprokinetic</td>
<td>Johnson &amp; Johnson</td>
<td>1050</td>
</tr>
<tr>
<td>27</td>
<td>Zestril</td>
<td>ACE inhibitor</td>
<td>Zeneca</td>
<td>1036</td>
</tr>
<tr>
<td>28</td>
<td>Voltaren</td>
<td>NSAID</td>
<td>Novartis</td>
<td>1019</td>
</tr>
<tr>
<td>29</td>
<td>Zovirax</td>
<td>Antiviral</td>
<td>Glaxo Wellcome</td>
<td>991</td>
</tr>
<tr>
<td>30</td>
<td>Taxol</td>
<td>Anticaner</td>
<td>Bristol-Myers Squibb</td>
<td>941</td>
</tr>
</tbody>
</table>

*Source: Wood Mackenzie's PharmaQuant, Edinburgh, U.K., Jan. 1999. Products derived from natural compounds from wildlife animals, plants, and microorganisms are printed in bold, and products derived from natural compounds are in italic.*
sine arabinoside, which was isolated from a Florida sponge (Alan Harvey, pers. comm.). Lovastatin (Mevacor) and its synthetic analog simvastatin (Zocor) are derived from the fungus Aspergillus terrestris, and anti-hypercholesterolemic drugs like pravastatin (Pravachol or Mevalotin) are natural compounds from Penicillium citrinum. Ceftriaxone (Rocephin) is a semisynthetic third-generation cephalosporin antibiotic (Cephalosporium sp.) and cyclosporin (Sandimmun/Neoral) is an immunosuppressant cyclic peptide from Tolypocladium inflatum. Co-amoxiclav (Augmentin) is a combination of the beta-lactamase inhibitor clavulanic acid from the bacterium Streptomyces lavuligerus and the semisynthetic antibiotic amoxicillin (Penicillium or Aspergillus spp.). Finally, clarythromycin (Biaxin/Klaricid) is a chemical derivative of the classic antibiotic erythromycin from the streptomycete Saccharopolyspora erythrea.

Random screening for new drugs from natural compounds comprises the collection of materials from plants and animals and the isolation of bacteria and fungi in large numbers. Extracts are prepared from these materials and then exposed to screening batteries with automated bioassay systems to discover novel biological activities. The bioactive principles are then purified and characterized by bioassay-guided fractionation. Since a large number of bioactive compounds will be already known and thus lack novelty, these need to be identified as early as possible in the discovery process by dereplication. Even if a compound is in the public domain, it might still be patentable in connection with a novel application. Another drug discovery route is provided by chemical screening, which fits the demand of modern high-throughput screening for large quantities of purified compounds. Without knowing their biological activity, extracts are first fractionated into pure compounds and these are then exposed to the bioassays.

Among the most notable developments in modern medicine were acetylcholinesterase (ACE)-inhibiting drugs like enalapril (Fig. 2a), captopril (Fig. 2b), lisinopril, and perindopril. These compounds were derived bioregionally from a peptide in the venom of a deadly tropical snake (Wyvarrant, 1988). The fer-de-lance (Bothrops atrox or B. jararaca) is a Neotropical pit viper that kills its prey by causing a drop in blood pressure (in the context of pharmaceutical development, this represents the observation of a biological prototype). Hypertension is a major health problem in most industrialized countries (the formation of an analogy between a medical problem and a biological phenomenon). A peptide in the venom inhibits an enzyme in the mammalian bloodstream, which converts the peptide angiotensin I to the hypertensive hormone angiotensin II. Analysis of the functional principle of this snake peptide at the molecular level and the discovery of the bioactive lead structure (a proline in combination with the carbonyl group of a glutamic acid) finally led to the synthesis of small molecules like captopril and enalapril (the technical realization of a biological principle). Thanks to drugs like the ACE inhibitors and, especially, these deadly snakes from the Neotropics, human deaths from stroke and heart attack have decreased by over 50% during the last two decades.

Many of today’s pharmaceutical industry multinational corporations became industrial giants by producing and selling products derived from ethnobotanical research. Prior to Bayer’s aspirin, American and European peoples treated fevers, inflammation, and pain with salicin-containing plants like willows and poplars. Novartis’ diclofenac, number 28 in the list of top-selling pharmaceuticals of 1997 (see Table I), was derived from the same salicin lead. Other examples are stabilizing muscle relaxants like tubocurarine, alcuronium, and pancuronium, which are routinely used during anesthesia. They were derived from curare, an arrow poison prepared from toxic plants by indigenous people in the Amazon and Orinoco valleys of South America. The anticancer drugs etoposide and temposide are derivatives of podophyllotoxin, a glycoside isolated from the roots of various species of the genus Podophyllum. These
plants possess a long history of medicinal use by early American and Asian cultures, including the treatment of skin cancers and warts. Cholesterol-lowering drugs like lovastatin, simvastatin, and pravastatin now have medical competition from a yeast extract from fermented Chinese red rice. Asians use these extracts traditionally as a food additive and the yeast produces lovastatin naturally.

Secondary metabolites from plants have been used traditionally in agriculture as natural pesticides, but they were pushed out of the market to a great extent following the advent of DDT. Other natural pesticides are nicotine from tobacco leaves, rotenone from roots of the genus Derris, quassin from the wood of Quassia amara, ryanodine from leaves, stems, and roots of the tree Rauvolfia serpentina, azadirachtins from the Asian neem tree, and pyrethrins from the perennial and Chrysanthemum cinerariifolium, originally from Dalmatia and Montenegro (Wink, 1993). However, the toxicity of chemically synthesized pesticides like organophosphates and chlorinated bicyclic compounds soon became evident and derivatives of natural insecticides, fungicides, and herbicides have gained increasing importance since then. Examples are the pyrethroids, the fungicides kresoxim-methyl (Fig. 3B) and azoxystrobin (Fig. 3C), and the herbicide phosphinothricin. Pyrethroids are derivatives of the pyrethrins, which were used traditionally as insecticides. Kresoxim-methyl and azoxystrobin are derived from strobilurin A (Fig. 3a), a natural fungicide that was discovered via biorational research on the mushroom Stobilurus tenacellus. Random screening led to the discovery of phosphinothricin in the culture supernatants of a streptomycete (Crouse, 1998).

2. Gene Prospecting

The selection of various phenotypes (and thus genes) has been used traditionally for plant and animal breeding, and modern gene technology offers many new opportunities. Numerous products derived from genes from wild life are already on or are close to entering the market (Tamayo et al., 1997).

A powerful protein to treat acute heart attack was discovered biorationally in the saliva of the common vampire bat, Desmodus rotundus. While feeding on the blood of their victims, these bats release analgesics and coagulation inhibitors with their saliva, including the Desmodus Plasminogen Activator (DPA), which dissolves thrombolytic blood clots and allows clot-free drinking. Recombinant human tissue Plasminogen Activator (tPA) has been approved as a therapeutic agent against heart attack in the United States and Europe, although researchers at Schering AG found that DPA is more efficient and safer for therapeutic application. Consequently other blood-sucking organisms like ticks, mosquitoes, and hookworms are currently under investigation in attempts to discover novel anticlotting, antiplatelet, and vessel-constricting principles.

One example of the value of traditional knowledge is the development of the recombinant pharmaceutical protein hirudin, a potent anticoagulant derived from the saliva of leeches (Hirudo medicinalis). Leech therapy has been used in traditional medicine in Europe and Asia for centuries as the only efficient treatment of thrombosis, thrombophlebitis, and hypertension.

Traditionally, humans have used enzymes for the production of cheese, softening of leather, and fermentation of tea and tobacco. In modern technology, enzymes have gained increasing importance in chemical engineering, in the processing of animal and human
food and textiles, as a component of laundry detergents, and for the diagnosis of diseases. Gene prospecting searches for enzymes with new substrate specificity and physicochemical characteristics, and gene technology allows their production in nearly unlimited amounts at low cost. Today random screening can be applied in the discovery process at high efficiency rates and without the need to isolate and cultivate microorganisms. DNA is isolated directly from water, soil, dung, carcass, or compost, then randomly cloned into expression libraries; the host bacteria are subsequently screened for enzyme activity in high-throughput screening robots on microtiter plates. From the hosts, the cloned DNA can be directly amplified and sequenced, and the genetic information used to construct highly productive host strains for the bioreactors.

Hydantoinases generate \( \alpha \)-amino acids, high-value fine chemicals for the production of half-synthetic antibiotics and pesticides. Particularly useful are thermostable hydantoinases, which were discovered by random screening of cultivated thermophilic bacteria from hot springs in Yellowstone National Park in the United States. Yellowstone was also the source of Taq polymerase from *Thermus aquaticus*, a crucial component of the polymerase chain reaction (PCR). Its inventor, Kary B. Mullis, plucked it in the late 1960s. Recently some users of PCR have replaced Taq polymerase by Pfu polymerase, which was isolated from *Pyrococcus furiosis* ("flaming fireball") and works best at 100°C. Proteases and lipases from psychrophilic bacteria, which decay whale cadavers in the deep sea, are of interest for producers of cold-wash laundry detergents (Madigan and Marrs, 1997). Food for poultry and pigs is increasingly processed with phytases from soil fungi like *Aspergillus niger*. This enables these monogastric animals to assimilate the phosphorous from plants and diminishes the amount of phosphorous that enters the environment in their manure. In the United States alone, the widespread application of phytase could prevent 82,000 tons of phosphorous from entering the environment each year and contributing to the eutrophication of rivers and lakes. Genes from wild plants, vertebrates, insects, and soil bacteria are used to create transgenic crops that carry resistance to pests and herbicides. They can be discovered biotecnologically, by random screening, or on the basis of knowledge from traditional uses.

Although frogs live in ponds infested with bacteria and fungi, they rarely get infected by these pathogens. This common knowledge led to the discovery of antimi-rsidal peptides, the magainins, which can be genetically expressed in transgenic plants and which confer resistance to fungi and bacteria. Random screening of actinomycetes from the soil of Cameroon led to the discovery of phosphinothricin acetylase, which confers resistance to the herbicide Basta in transgenic crops like soy bean, rape, and cotton. Resistance to insects can be achieved by the expression of insecticidal proteins from soil bacteria, like delta-endotoxin from *Bacillus thuringiensis* and cholesterol oxidase from *Streptomyces* sp. (Tamayo et al., 1997).

As a consequence of their absorptive mode of nutrition, fungi are able to exploit an almost infinite diversity of nutritional microniches. Recent studies suggest that considerable fungal diversity exists even in what would be considered a homogenes substrate. For example, Bills and Polsihook (1994a) adapted the particle filtration technique to maximize the diversity of fungi isolated from a substrate. Whereas traditional isolation techniques select for fast-growing species that utilize primarily simple carbohydrates, the particle filtration technique favors isolation of those fungi actively growing in the substrate. In their study, fungi were isolated from 1-ml particle suspensions of tropical rain forest leaf litter. From each sample, 80–145 different fungal species were isolated and the four replicates of the same substrate yielded between 300 and 400 different species of fungi. In a second study (Bills and Polsihook, 1994b), fungi were isolated from only one vascular plant host, *Heliconia mariae* J. D. Hooker. From 0.8-ml suspensions of filtered particles of leaves from four plants, they isolated about 50–100 fungal species per leaf sample. In both of these studies, approximately 40–60% of the fungi isolated were rare or unidentifiable. Clearly the fungi are far more diverse than is currently recognized, even by mycologists.

Metabolic pathway engineering also offers new agricultural opportunities. A gene for a 12-carbon thioseterase that transforms the common rape plant into a producer of high-laurate canola oil has been discovered in the undomesticated California bay (*Umbellulicaria californica*). This oil is a perfect substitute for palm kernel and coconut oil, which are high-value commodities to tropical countries. Already 70,000 acres of transgenic rape are under cultivation in the United States and Canada.

Developments like Laurical (the trade name for this high-laurate canola oil) displace the production of tropical commodities to the farm fields of developed countries, or even possibly to industrial bioreactors, posing...
a threat to the economic survival of millions of farm families in less developed countries. In Africa alone, $10 billion in annual exports are vulnerable to industry-induced changes in raw material prices and requirements. Most developments in plant biotechnology have been achieved with crops mainly cultivated in industrialized countries. This fact reduces the likelihood that farmers in developing countries will benefit from the new agricultural opportunities of gene technology (Tamayo et al., 1997).

3. Bionic Prospecting

Architects, designers, and engineers use prototypes from nature for construction and technical solutions. Josef Baxton, after seeing the radial ribbing in the veination of the Amazonian water lily (*Victoria amazonica*), designed on this basis the dome of the crystal palace in London. The inventor of Velcro fasteners was inspired by the stickiness of the seeds of a common farm weed, the burdock (*Arctium* spp.). The development of the echo-sounder was inspired by the ultrasonic orientation system of bats, and infrared sensors were based on the design of the thermosensitive pit organ of the rattlesnake. The latter developments are of particular importance for military use and thus it is not surprising that the concept of “bionics” was created in 1958 by the Air Force major Jack E. Steele as the “science of systems” that function in the same way or similar to living systems (Hill, 1997).

The discovery of new technical principles in living systems is far from being completed, and at least one major resource is still untapped: the insect world of the tropics. Studies by Hill et al. (in press) in Costa Rica on butterflies and grasshoppers have revealed a diversity of new coupling mechanisms and led to the definition of “tropical bionics” as a new branch in engineering. Insects became the most species-rich group of eukaryotes by their capability to adapt to ecological niches. They are not only excellent chemists and genetic engineers, but also masters in mechanical engineering, lightweight construction, and sensor technology. Figure 4 shows some of the morphological and physiological characteristics of a butterfly. Because of their diversity, insects are an enormous reservoir of potential technological developments.

Until recently, the development of technical solutions in engineering on the basis of biological prototypes was mainly driven by biorational observations. For example, sharks have a rough skin, and common sense suggests that rough surfaces should increase the resistance in water and thus slow the fish down. However, sharks swim at speeds of over 50 km/hr. This paradox led to the discovery that the structure of the shark squamae impedes microturbulence, which commonly emerges on smooth surfaces at high speed. The technical development of this principle led to the creation of new surfaces for airplanes that significantly reduce kerosene consumption. Another example of the biorational approach comes from the lotus, considered to be the flower of purity and cleanliness in Asian cultures. Tiny wax papillae on its surface act as a self-cleaning mechanism. This principle is currently being

**Figure 4** The butterfly: a bionic system and technical prototype (with variations derived from Hill et al., in press).
used to develop self-cleaning surfaces for buildings and cars. A new tool for the discovery and development process in bionics was recently developed by Hill (1997). Structures and mechanisms discovered in nature are systematically listed in catalogs and organized after the principal functions in engineering: connection, separation, formation, carriage, support, transmission, storage and blocking of materials, energy, and information. These catalogs of inspirations are available to the engineer to screen for technical solutions, the blueprints of which nature has already provided.

II. ECONOMIC VALUE AND BENEFIT SHARING

A. Biodiversity-Related Markets

Markets related to products derived from biodiversity are significant, as shown in Table II. This table does not include bionic applications like sensor technology and robotics.

The estimated growth of the world drug market is about 6% per year, from 1997 sales of U.S. $295 billion to $378 billion in the year 2001 (Thayer, 1998). To become a billion-dollar product, a newly developed pharmaceutical has to be better than similar products already on the market or reveal a new mode of action.

<table>
<thead>
<tr>
<th>Market</th>
<th>U.S. $ (billions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug market (1997)</td>
<td>295</td>
</tr>
<tr>
<td>Natural product-based drugs (1993)</td>
<td>50</td>
</tr>
<tr>
<td>Phytopharmaceuticals (1993)</td>
<td>12.4</td>
</tr>
<tr>
<td>Pesticides</td>
<td>30</td>
</tr>
<tr>
<td>Seeds (estimate for 2000)</td>
<td>50</td>
</tr>
<tr>
<td>Horticulture (U.K., 1994)</td>
<td>3.6</td>
</tr>
<tr>
<td>Enzymes (1996)</td>
<td>2.3</td>
</tr>
<tr>
<td>Cosmetics (U.S.A., 1994)</td>
<td>20</td>
</tr>
<tr>
<td>Natural cosmetics</td>
<td>0.5</td>
</tr>
<tr>
<td>Tourism (1995)</td>
<td>3400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>a From Thayer (1998).</td>
</tr>
<tr>
<td>b From ten Kate (1995).</td>
</tr>
<tr>
<td>c From Gruenwald (1995).</td>
</tr>
<tr>
<td>d From K. ten Kate, pers. comm.</td>
</tr>
<tr>
<td>e From Madigan and Marni (1997).</td>
</tr>
</tbody>
</table>

In the highly competitive drug market, innovation and production speed are paramount, and consequently in the United States R&D costs increased within ten years from U.S. $6 billion to $20.6 billion in 1998. The supply of screening batteries for the synthesis and extraction of new compounds consumes 11% of these R&D budgets and requires a workforce of 4400 professionals or 12.9% of all employees in R&D (Thayer, 1998). To compete with synthetic and combinatorial chemistry efforts, bioprospecting needs to provide large numbers of natural compounds at high speed, for example, by chemical screening, or untap new and unusual resources.

Global sales of U.S. $30 billion in 1995 makes agrochemistry the second largest biodiversity-related market. Because of the toxicity and environmental problems associated with synthetic chemicals, pesticides derived from natural compounds currently dominate the market (see earlier discussion), however, development of resistance in pests continually requires new and innovative products and creates demand for bioprospecting. Crop resistance to pests can be achieved by genetic engineering and this might, at least partially, substitute for some pesticide use in the future. The market for the seeds of these genetically engineered varieties is estimated to grow to U.S. $50 billion in the year 2000. Currently most genetic transformation is done with genes coding for insecticidal proteins from Bacillus thuringiensis. However, it is only a matter of time before crop resistance is overcome by pests, and therefore the discovery of new resistance genes from wild sources has become a central focus of bioprospecting efforts.

Other markets related to biodiversity are gaining increasing importance, like those for enzymes (see earlier), phytopharmaceuticals, seeds, and natural cosmetics. The worldwide market for phytopharmaceuticals derived from traditional medicinal sources is estimated at U.S. $12.4 billion, headed by products derived from ginseng, ginkgo, garlic, horse chestnut, and echinacea (Gruenwald, 1995). Nutraceuticals and functional foods offer a health advantage in addition to nutrition. Examples are Chinese red rice and cholesterol-lowering oils, the latter enriched in plant sterols and stanol, both by-products of the paper pulping industry and corn wet-milling operations. With an estimated market of $7.9 billion in 1997, these products are expected to compete effectively with cholesterol-lowering drugs like lovastatin. In general, the future market looks bright for products derived from nature, with an increasing demand for natural compounds and wild genes for use in genetic engineering. The same is true for bionics.

Recently, attempts have been made to put price tags
on natural resources, including biodiversity (Constanza et al., 1997). The value of sources of unique biological materials and products was estimated as U.S. $79 per hectare of forest per year. The value of yet undiscovered pharmaceuticals in tropical forests was estimated at U.S. $3–4 billion for a private pharmaceutical company, and as much as U.S. $347 billion dollars to global society as a whole. Although significant, these numbers do not reflect the real value of inspiration that can be derived from wild plants, animals, and microbes. Biodiversity provides the goods and services that are essential in supporting every type of human endeavor and thus enables societies to adapt to different needs and situations. Biological prototypes may be crucial for the survival of mankind in the future, and the “Great Hunger” in Ireland in the mid-1800s taught a bitter lesson. This potato late blight (Phytophthora infestans) caused the death of 1 million and emigration of another 1.5 million people.

B. Bioprospecting and Conservation

The initial meaning of prospecting was the search for precious metals and oil in the soil and subsoil. Then Thomas Eisner, from Cornell University, coined the term “chemical prospecting,” and more recently bioprospecting has been used to indicate the search for new sources of chemical compounds, genes, proteins, microorganisms, and other products that have economic potential and that can be found in the world’s diversity of plants, animals, and microorganisms.

For bioprospecting to be an ally of conservation, it must be based on sound principles and criteria, and must be linked to the concept of value. It was highlighted earlier that the present and potential value of bioprospectible designs, genes, and chemicals is enormous and justifies the use of intelligent conservation strategies. To be effective as an element of conservation, bioprospecting must contribute to the generation of wealth through R&D and must also alleviate poverty in biodiversity-rich regions. In this context, bioprospecting programs at the national level can help to develop national capacity and provide economic returns to conservation projects if investments are made in science, technology, and market research. However, this may not be sufficient to achieve sustainable development: communities inside or bordering protected areas must participate in conservation and must benefit from bioprospecting activities in their own regions. The utilization of biodiversity linked to added value at the local level appears to be a promising avenue to generate economic growth at the community level and to create incentives for conservation.

C. Biopiracy

Our world would be a totally different place today if crops and domestic animals had not been moved from their centers of origins and domestication to other parts of the world: Coffee to the Americas, potatoes to Europe, corn to Africa, and wheat to China are just a few examples of how germplasm has become the “patrimony of humanity.” The ratification of the Convention on Biological Diversity (CBD) and the development and implementation of property rights regimes have changed completely the prior existing scenarios on the spread and use of germplasm.

Today, in many cases genetic and biochemical resources are jealously guarded and are the cause of conflict (when appropriate and legally sound negotiations are absent) in many parts of the world. The term biopiracy has been coined to reflect the illegal appropriation or exploitation of genetic and biochemical resources. It must be added, however, that the lack of proper codes of conduct, lack of legislation, and the lack of national capacity to handle biodiversity utilization issues in the majority of biodiversity-rich countries are factors encouraging biopiracy. Even national researchers, when dealing with poorly defined or cumbersome legal procedures, tend to take shortcuts in their quest for useful genetic and biochemical resources.

D. Benefit Sharing Models and Bioprospecting Experiences

Perhaps the most controversial issue related to bioprospecting in tropical developing countries involves benefit sharing. This is true for the case of the Instituto Nacional de Biodiversidad (INBio) in Costa Rica, as well as for other models and experiences developed elsewhere.
The central questions are: What is equitable? Who benefits and how? Who makes the *rules*?

The issue of equitability is controversial and first arose with the initial perception that millions of dollars would flow back to the countries, organizations, and individuals involved in bioprospecting. The experiences of the last few years indicate that monetary benefits (unless royalties would materialize) to host countries, although significant, are limited in comparison to other less tangible benefits such as technology transfer, increased local scientific expertise, improvements in legal frameworks, and enhanced negotiating capacities. These less tangible benefits may be poorly understood or underappreciated by some segments of society, who quite rightly are interested in achieving a direct flow of resources and economic benefits to the local communities living near conservation areas.

Conditions are different in every country and therefore local solutions should be tailored to local circumstances. Inside Costa Rica’s government-protected wildlands, for example, there are no human settlements, in contrast to the case in many other biodiversity-rich countries, and therefore straightforward agreements developed by INBio with the government, the academic sector, and private companies do not need to address the more sensitive issues of indigenous knowledge and direct monetary compensation to communities. Current bioprospecting experience, such as in Costa Rica, indicates that benefits to society at large include an enhanced knowledge of the country’s biodiversity, and improved information dissemination to various sectors of society through guidelines and other publications, information available on the Internet, national and international workshops, and education programs. This scenario is a significant improvement to the information available on the Internet, national and international workshops, and education programs. Perhaps one of the more comprehensive and systematic bioprospecting programs has been the one carried out by the National Cancer Institute (NCI) of the United States. As of 1999 the NCI had major contract-supported programs to collect, identify, and screen plant products and fermentation broth. A modest marine acquisition program was established in 1972. Central objectives of the NCI have been cancer and more recently HIV research. Natural product extracts and other samples are screened in vitro against panels of human tumor cell lines and against HIV-infected cells. Acquisition of samples concentrates on marine and unusual microorganisms. Active natural components are isolated by bioassay-guided fractionation and only a handful of them will be pursued to complete structure elucidation. Known natural products are quickly identified via established dereplication.

Benefit sharing agreements go back to the year 1986, when the National Cancer Institute started its second plant acquisition and screening program. Through contracts with the Missouri Botanical Garden, New York Botanical Garden, and the University of Illinois at Chicago, collections in over 25 countries were made under the Letter of Intent and later the Letter of Collection, in which the NCI unequivocally stated its intent to deal with source countries in a fair and equitable manner. By inviting source country scientists to the NCI laboratories, these agreements include a technology transfer component as well as a commitment of the Institute to require successful licensees of their anticancer drugs to enter into further benefit sharing agreements with source country government agencies or organizations. Furthermore, license applicants are required to seek as the first source of supply the natural products from the source country. The letter also acknowledges the achievements and intellectual property rights of local collectors, taxonomists, and healers. In a further development of its benefit sharing policy, for several years the NCI has negotiated Memoranda of Understanding (MOU) with 13 source countries, including Costa Rica. These MOUs are cooperative research agreements and recognize that source country scientists and organizations are committed to performing an increasing share of the operations in-country, as opposed to the export of raw materials.

A successful example of this benefit sharing policy is the company Sarawak Medichem Pharmaceuticals, Inc., which was founded in 1996 as a joint venture between Medichem Research, Inc., a small pharmaceutical company based near Chicago, and the Sarawak state government. Sarawak Medichem is developing (+) calanolide A as an anti-HIV drug. Calanolides are potent inhibitors of retroviral reverse transcriptase and were originally found in leaves and twigs of the tree Calophyllum longatum, collected in Sarawak, Malaysia, in 1987, through the NCI contract with the University of Illinois at Chicago.

Australia is actively involved in several bioprospecting initiatives and has made significant attempts to add value and advance product development in the country. It has also developed strategic partnerships with Southeast Asian and South Asian countries to increase the availability of germplasm resources. Selected initiatives include those of AMRAD to do bioprospecting on the flora of the Northern Territory and the Australian...
random were active in the yeast assays, whereas 3.8% of the plants collected at the Virginia Tech ethnobotanical and random botanical collections using yeast-based assays. The Virginia Tech scientists conduct ethnobotanical and random botanical research on the flora of Queensland by local laboratories linked to ASTRa, and a new laboratory based in Lismore, New South Wales, that will address, among other areas, research and development on medicinal plants.

The International Cooperative Biodiversity Group (ICBG) Program is an experimental program that addresses the interdependent issues of drug discovery, biodiversity conservation, and sustainable economic growth. Funding from the Institutes of the National Institutes of Health (NIH), including the National Cancer Institute, supports awards of approximately $430,000 per year to the five initial initiatives. As of 1997, nearly 4000 species of plants and animals had been examined for biological activity in 13 different therapeutic areas. ICBG-supported projects include acquisition and analysis of natural products derived from biological diversity as potential therapeutic agents for diseases of concern, to both developed and developing countries, like AIDS, malaria, tuberculosis and other infectious diseases, cancers, central nervous system disorders, and heart disease.

The active ICBG projects have made variable progress in knowledge generation, local capacity building, training, and the identification of promising lead compounds. The one in Suriname, briefly reported here, involves the participation of the Virginia Technical Institute, the Missouri Botanical Gardens, Bristol Myers Squibb Pharmaceutical Research Institute, and the National Herbarium of Suriname. After conducting some 14,000 assays from 1993 to 1998 of more than 3300 extracts, Virginia Tech has identified 30 unique extracts that have activity and has isolated 20 chemical compounds that have bioactivity. The most interesting compounds discovered are a group of alkaloids from Eclipta alba that have good antifungal activity, better in some cases than the clinically used drug amphotericin B. However, these alkaloids also have weak cytotoxicity, and the decision was thus made not to develop them as antifungal agents.

Meanwhile, Bristol Myers Squibb has put more that3000 extracts through 32 screens each in six therapeutic areas, with the result that one promising compound (for anticancer activity) is still being tested. While looking for potential pharmaceutical products, project scientists conduct ethnobotanical and random botanical collections using yeast-based assays. The Virginia Tech researchers found there was a slight benefit to the ethnobotanical approaches: 2.8% of the plants collected at random were active in the yeast assays, whereas 3.8% of the shaman-identified plants were bioactive. This small difference is to be expected, given that the assays are not specific to how the shamans use the plants. The ICBG Suriname project has also been able to describe new plant species. This is relevant given that tropical forests have been the source of 60% of the anticancer drugs discovered in the last 10 years and offers a potentially powerful economic reason for preserving the forest, based on the $200 billion market for plant-derived drugs.

The experience of INBio as a participant in one of the initial ICBG projects has been positive in terms of knowledge generation and capacity building, however, it also indicates how the role of the industrial partner was not clearly defined nor negotiated at the outset. This issue has been explicitly raised by Chapela (1996), and in his opinion the five ICBG Consortia reflect the specific interests of the academic partners with little consideration of the objectives of the industrial partners. The interest of industry in the ICBG may be more closely related to discovering microorganisms (which were explicitly excluded) present in some of the samples to be collected (Chapela, 1998).

The isolation of active compounds from tropical plants with a history of medicinal use was the main strategy of Shaman Pharmaceuticals in their efforts to discover and develop novel products for human diseases. The company founded the Healing Forest Conservancy, a nongovernmental organization, explicitly to develop and implement a process by which to return benefits to countries and cultures that chose to contribute to its drug discovery program. Shaman contributed up to 15% of its drug discovery costs to fund projects or programs that were based on the expressed needs of the local communities. Shaman also provided medium-term reciprocity programs that included training for national scientists at Shaman's California laboratories, equipment and resources to conduct research in host countries, and the funding of studies to evaluate sustainable harvesting of indigenous medicinal plants. According to Shaman staff, the practical dilemmas of distributing fair compensation based on intellectual property rights are numerous, and therefore Shaman's approach involved an upfront return and at each further stage of the product development process. The recent and significant reorganization of Shaman in early 1998 will likely affect the ongoing benefit sharing experiences with the communities.

Leading botanical gardens have developed benefit sharing strategies to compensate the countries where botanical specimens are initially collected. Through their collaboration with the National Cancer Institute,
the New York Botanical Garden and the Missouri Botanical Garden voluntarily adopted policies on the research of natural products in the mid-1980s. Most notably the Missouri Botanical Garden committed itself to enter into commercial research agreements only with the provision that royalties will be paid to the source country and that the Garden itself will not receive any percentage of such royalties. In 1981 the New York Botanical Garden established the Institute of Economic Botany to focus a portion of the Garden’s research enterprise on applied botanical questions of great human concern.

One of its forest projects is the Belize ethnobotany project, begun in 1988, to conduct an inventory of the diversity of medicinal plants and to preserve this richness. In Europe, the Royal Botanical Gardens at Kew took the lead and its benefit sharing policy also includes those materials that had been collected prior to the ratification of the CBD.

A very innovative and comprehensive strategy is that of South Africa, where traditional healers, the government, and private and academic sectors are joining efforts to undertake bioprospecting. The initial investment (40% government and 60% private sector) of U.S. $10 million will fund the investigation of 23,000 plant species plus marine organisms and microorganisms. The plan is to develop international partnerships, however, most of the R&D will be done in-country.

III. CONCEPTS AND PRACTICES OF BIOPROSPECTING: THE CASE OF COSTA RICA

A. Nature-Based Initiatives and the Role of the National Biodiversity Institute

Costa Rica has developed a system of conservation areas in which over 25% of the national territory is allocated to nature reserves and national parks. Both the government and civil society have a role in conservation and have contributed to creating awareness of the importance of biodiversity at the national and international levels. In terms of economic benefits, Costa Rica derives its main source of income from nature-based tourism; this sector yielded over U.S. $830 million in 1998 and appears to be growing steadily. Surveys indicated that 70–75% of all tourists come to experience nature. This is an impressive development when compared to the traditional economic sectors of banana and coffee production $500 million and $350 million, respectively.

Other relevant initiatives to value biodiversity include payment for environmental services such as CO₂ fixation. A 5% fuel tax yields approximately $40 million yearly, which is invested in carbon sequestration programs to support the conservation of forests and reforestation. Serious consideration is also being given to charge users for the actual costs of water and electricity, and initial experiments are under way. Currently, water bills reflect essentially distribution costs and not the value of this essential commodity to society.

Costa Rica is considered to be a megadiversity country. Most of its territory has a diversity density of over 5000 species of vascular plants per 10,000 km². Only 10 other countries on earth (Nicaragua, Panama, Colombia, Ecuador, Peru, Bolivia, Brazil, China, Malaysia, and Papua New Guinea) have areas with similar diversity (Barthlott et al., 1996). However, of the total number of species estimated for the country, only a minor portion has been described. Vertebrates and vascular plants have been well studied, with over 90% of the species documented and characterized. For all other groups of organisms, including insects, algae, fungi, and bacteria, there are enormous gaps of knowledge.

The management and development of biological resources fall under the responsibility of the Ministry of Environment and Energy (MINAE), specifically the National System of Conservation Areas (SINAC), which is the institution in charge of conserving and promoting the sustainable use of the country’s natural resources. The eleven Conservation Areas in which the country has been divided constitute the various regions established by MINAE to carry out a decentralized management of biodiversity and natural resources, with the participation of the communities surrounding the protected areas.

Executive Presidential Decree No. 19153 of June 5, 1989, established a Biodiversity Planning Commission, with representatives from various governmental organizations, higher education institutions, and non-governmental conservationist organizations. This Commission recommended the creation of the National Biodiversity Institute (INBio) as a nonprofit, nongovernmental organization to serve the public good. INBio was legally registered on October 26, 1989, as a civil association governed by an Assembly of Founders and a Board of Directors.

INBio works under the premise that a tropical country will be able to conserve a major portion of its wild biodiversity if this biodiversity generates enough intellectual and economic benefits to make up for its maintenance. INBio and the Ministry of Environment and Energy established a collaborative agreement that...
allows INBio, within the existing legal framework, to
conduct biological inventories, biodiversity prospect-
ing, and management and distribution of Costa Rican biodi-
versity information.

One of INBio’s goals is to generate greater awareness of
the value of biodiversity and, thereby, promote its conser-
vation and sustainable utilization based on the demands
of national and international users. This goal is fulfilled
through the integration of the following processes:

• biodiversity inventory focused particularly on the
country’s protected wildlands;
• conservation for development;
• identification and promotion of sustainable uses of
biological resources;
• organization and management of information; and
• generation and transfer of knowledge.

The Biodiversity Inventory Program is carried out
in Conservation Areas by “parataxonomists.” They are
in charge of collecting, preparing, and mounting the
biological material with pertinent field information, and
local curators are responsible for the taxonomic classi-
fication of such material, with support from national
and international experts.

The Conservation for Development Program helps
strengthen the capacity of INBio and assists in the tasks
assigned by the National Conservation Strategy, which
is based on protecting, understanding, and using bio-
diversity in a sustainable manner.

The Biodiversity Information Management Program
processes and organizes the information generated by
the inventory and other national collections under vari-
ous formats, according to the needs of users. Thus, it
seeks to process and integrate texts, maps, photographs,
and quantitative information about Costa Rican biodi-
versity.

Natural resources constitute an essential source for
new medicines, agricultural products, coloring com-
pounds, resins, and other uses. Through the Biopros-
ppecting Program and in partnership with national and
international organizations, INBio carries out a system-
atic search of chemical compounds, genes, and micro-
and macroorganisms that may help find new products
for the benefit of humanity.

The need to disseminate biodiversity information
motivated the creation of the Social Outreach Program
as a means to distribute information to different sectors
of society. It includes production and publication of
printed and audiovisual material on biodiversity issues
for various users, special programs for elementary and
high school students, and information for visitors. This
program also cooperates with various national commis-
sions of private and public organizations.

B. Partnerships and Management Issues

Current bioprospecting approaches require a significant
degree of partnership at various levels. In Costa Rica,
INBio has developed a close alliance with the govern-
ment (Ministry of Natural Resources and Energy) that
makes bioprospecting development possible in the na-
tion’s Conservation Areas. This first level of partner-
ship is essential and is based on trust, understanding,
and clear-cut responsibilities and benefits (see Section III.D).

Another essential level of partnership involves the
academic sector. Universities, both local and interna-
tional, complement the required expertise in biology,
chemistry, microbiology, and biotechnology that is cen-
tral to the concept and need of adding value at the
national level. Last but not least, a strong alliance with
the private sector (conservation, pharmaceutical, ag-
ricultural, biotechnology, and fragrance companies)
allows market forces and the needs of society to deter-
mine the products and services to be pursued through
bioprospecting. The country has a limited capacity to
invest in science and technology, and therefore partners-
ships with the academic and private sectors are the only
way to accomplish many of the stated goals.

C. Mechanisms and the Legal Framework

Managing the complexity of the institutional interac-
tions and addressing the needs and expectations of the
various partners require careful planning and organiza-
tion. Costa Rica and INBio cannot provide a prescrip-
tion for bioprospecting success, but do offer 10 years
of experience full of drawbacks, challenges, and
achievements.

The requirements of a bioprospecting program are
relatively simple to state: vision, leadership, negotia-
tion, planning, understanding partners’ strengths and
expectations, teamwork, realistic research plans, and
sufficient local expertise in science and informatics. But
these elements are not necessarily simple to operation-
alize. It is widely perceived that being a small country
and having a well-educated workforce and the right
incentives make the process easier in a country like
Costa Rica. All this is probably true, however, there are
few comparisons available to ascertain the validity of
this hypothesis.
A country's legal framework, based on both theory and experience, sets the fundamental organization and incentives not only for conservation but also for bioprospecting. Table III highlights significant pieces of legislation that have made conservation and utilization of biodiversity possible in Costa Rica. The 1998 Biodiversity Law, a law of consensus among the various stakeholders, reflects in a significant way the country's prior experiences in conservation and bioprospecting.

### D. Criteria

Out of the 45 articles contained in the Convention on Biological Diversity (CBD), INBio and the National System of Conservation Area (SINAC) of the Ministry of Environment and Energy have emphasized the following ones in the design, approval, and implementation of bioprospecting agreements.

#### 1. Access

Signatory countries of the CBD have agreed to facilitate access to other members. INBio and SINAC comply with this principle provided that other agreed criteria (presented in the following sections) are an integral part of the access agreement.

#### 2. Compensation

In the INBio/SINAC strategy, compensation is negotiated at two different levels:

(a) All research costs related to a bioprospecting agreement (e.g., collecting, GIS, natural history extraction, determination of activity, etc.) are paid by the interested company. A contribution equaling 10% of all operational costs is transferred to SINAC for conservation purposes.

(b) In case a product based on a sample provided by Costa Rica reaches the market, royalties are payable in lieu of patent rights. Current patent laws in Costa Rica offer limited protection and are therefore unsuitable for products destined for the world market. Royalties vary significantly, depending mostly on added value generated within the country.

Literature surveys indicate that 1–15% royalty ranges are the norm. INBio's experience coincides in most cases with these figures. The transfer of resources from bioprospecting agreements to research and conservation programs is shown in Table IV, and the list of active bioprospecting agreements (both academic and with industry) is given in Table V.

#### 3. Transfer of Technology

Access to partner companies' research processes and bioassays of interest to INBio and its local partners is negotiated as part of the transfer of technology. INBio's current research capacity in bioprospecting is largely based on technology acquired through such agreements with the private and academic sectors.

#### 4. Training

Contributions that a host country can make in bioprospecting research and product development are essential.

### TABLE III

The Legal Framework for Biodiversity Conservation and Bioprospecting in Costa Rica

<table>
<thead>
<tr>
<th>Law/Convention</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forestry Law</td>
<td>1995</td>
</tr>
<tr>
<td>National Parks Law</td>
<td>1977</td>
</tr>
<tr>
<td>Wildlife Conservation Law</td>
<td>1992</td>
</tr>
<tr>
<td>Ministry of Environment Law</td>
<td>1990</td>
</tr>
<tr>
<td>Convention on Biological Diversity (CBD)</td>
<td>1993</td>
</tr>
<tr>
<td>General Environmental Law</td>
<td>1995</td>
</tr>
<tr>
<td>Biodiversity Law</td>
<td>1998</td>
</tr>
</tbody>
</table>

### TABLE IV

Contributions and Payments from Bioprospecting Agreements (U.S. $)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ministry of Environment and Energy (MINAE)</strong></td>
<td>110,040</td>
<td>43,400</td>
<td>66,670</td>
<td>51,092</td>
<td>95,196</td>
<td>24,160</td>
<td>390,558</td>
</tr>
<tr>
<td><strong>Conservation Areas</strong></td>
<td>86,102</td>
<td>203,135</td>
<td>153,555</td>
<td>192,035</td>
<td>126,243</td>
<td>29,579</td>
<td>790,649</td>
</tr>
<tr>
<td><strong>Public Universities</strong></td>
<td>460,409</td>
<td>126,006</td>
<td>46,962</td>
<td>31,265</td>
<td>34,694</td>
<td>326,486</td>
<td>1,025,822</td>
</tr>
<tr>
<td><strong>Other groups at INBio</strong></td>
<td>228,161</td>
<td>92,830</td>
<td>118,292</td>
<td>172,591</td>
<td>129,008</td>
<td>0</td>
<td>740,882</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>884,712</td>
<td>465,371</td>
<td>385,479</td>
<td>446,983</td>
<td>385,141</td>
<td>380,225</td>
<td>2,947,911</td>
</tr>
</tbody>
</table>

*Estimated amounts since 1991.

### TABLE V

<table>
<thead>
<tr>
<th>Partner</th>
<th>Year started</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Costa Rica</td>
<td>1991</td>
<td>General agreement</td>
</tr>
<tr>
<td></td>
<td>1998</td>
<td>Operation of an NMR unit</td>
</tr>
<tr>
<td>University of Strathclyde (U.K.)</td>
<td>1991</td>
<td>New pharmaceutical products from plant sources</td>
</tr>
<tr>
<td>Cornell University</td>
<td>1993</td>
<td>Drug discovery</td>
</tr>
<tr>
<td>Universidad Nacional, Costa Rica</td>
<td>1993</td>
<td>General agreement</td>
</tr>
<tr>
<td>Bristol Myers Squibb/Cornell University</td>
<td>1994</td>
<td>Inserts as source of new compounds for pharmaceutical industry</td>
</tr>
<tr>
<td>Instituto Tecnologico de Costa Rica</td>
<td>1994</td>
<td>General agreement</td>
</tr>
<tr>
<td>Granular Biot (New Jersey, USA)</td>
<td>1995</td>
<td>Fragrances and aromas from Costa Rican biodiversity</td>
</tr>
<tr>
<td>University of Massachusetts</td>
<td>1995</td>
<td>New insecticides</td>
</tr>
<tr>
<td>Diversa (San Diego, USA)</td>
<td>1995/1998</td>
<td>New enzymes from extremophile and other microorganisms</td>
</tr>
<tr>
<td>Government of Canada</td>
<td>1996</td>
<td>Debt for nature swap</td>
</tr>
<tr>
<td>INDENA (Italy)</td>
<td>1996</td>
<td>New antimicrobials for dermatological use from plant sources</td>
</tr>
<tr>
<td>EARTH/NASA/Other Latin American institutions</td>
<td>1997</td>
<td>Chaga space project: compounds against regulatory enzymes</td>
</tr>
<tr>
<td>University of Strathclyde (U.K.)</td>
<td>1997</td>
<td>Pharmaceutical products</td>
</tr>
<tr>
<td>Office for Scientific, Technical and Cultural Affairs</td>
<td>1998</td>
<td>MOSAICC Project</td>
</tr>
<tr>
<td>Phytera (Massachusetts, USA)</td>
<td>1998</td>
<td>Pharmaceutical products from plant sources</td>
</tr>
</tbody>
</table>

5. Sustainable Uses

INBio and its partners will not engage in extractive strategies should a forest resource become of interest to the market. Instead, they will pursue chemical synthesis approaches if possible, a domestication strategy if feasible and desirable. In the latter case, a species may be successfully domesticated and become a new agricultural crop or a new productive enterprise with the potential to generate economic development.

E. Lessons Learned, Challenges, and Opportunities

Costa Rica and INBio have gained considerably from bioprospecting agreements, through knowledge about the national biodiversity, transfer of technology training, and development of negotiating capabilities. However, we have also learned that bioprospecting is not the gold mine that was initially envisioned. There are no guarantees that royalties will materialize in the future and therefore they should not be included in decision-making and budgetary processing. Obviously, if they become a reality they will be a welcome incentive and support for conservation and research.

We have learned that partnerships and alliances are indispensable, that innovation must be at the forefront of our activities, and that controversy and misunderstandings are part of the daily routine. We have also learned, in order for history not to repeat itself, that Costa Rica must develop internal R&D activity to advanced product development processes and in certain cases to finished products. Private companies have shown a cyclical interest in natural products, and therefore bioprospecting has had and will continue to experience peaks and valleys of demand.

At INBio, bioprospecting is only one component of a much wider national biodiversity strategy. The
major factors responsible for INBio’s initial accomplishments have been solid political support, leadership and vision, a commitment to innovation, a diversified portfolio of goals, and strategic alliances (Mateo, 1996). Early bioprospecting efforts emphasized demand-driven agreements with pharmaceutical, biotechnology, agricultural, and fragrance companies. This line of activity will certainly be maintained and enhanced, and it should evolve toward higher levels of research and services that are performed within Costa Rica (the added-value contribution). In parallel, and given the particular conditions in Costa Rica, INBio will also be involved in the development of new biodiversity products (nutraceuticals, phytomedicines, ornamentals, etc.) with national industries and local communities. This strategy will involve a pilot product development phase followed by marketing of biodiversity-derived products. Cost recovery mechanisms will be designed to support new product development initiatives.

The recently born concept of modern bioprospecting is already being challenged by new developments in science and technology. In an endeavor to reduce risks and to increase the probability of success, companies are trying to substitute natural (incalculable) products with artificial (calculable) products wherever possible. The success rate in drug discovery is estimated to be about one product out of 10,000 screened chemicals or plant species for one field of diseases. To maintain 10% annual growth, or roughly $1 billion in new revenues, drug companies must bring two or three new products to the market each year, which is four to six times the current average achieved by the top producers. To speed up the discovery process, companies are investing in Ultrahigh Throughput Screening (UHTS) technologies, in which up to 100,000 compounds can be exposed each day to a growing array of bioassays. The number of extracts available from biological resources can scarcely satisfy this machinery and are now being challenged by the chemical diversity provided by combinatorial chemistry, which synthesizes all possible combinations of compounds from various sets of chemical building blocks.

It is too early to judge whether combinatorial chemistry will be as successful as natural sources in the generation of new drug candidates. Many scientists doubt that novel lead structures will result from these combinatorial efforts, which can compete with those evolved in nature during an evolutionary process of millions of years. However, the technology can be a powerful tool to complement natural sources by generating millions of derivatives from natural leads. It should be noted that many highly successful drugs, which have been introduced to the market only recently, are derivatives of natural leads that were discovered many decades ago. Examples are diclofenac (from salicin), clarithromycin (from erythromycin), and ceftriaxone (from cephalosporin).

In conclusion, the hunters, gatherers, and agriculturalists of the Old and New Worlds are, in some ways, the equivalent of the modern biochemical and gene bioprospectors. Perhaps the methods and strategies differ, yet the goals are essentially the same: survival, well-being, and spiritual and material advancement.

See Also the Following Articles

AGRICULTURE, TRADITIONAL • BREEDING OF PLANTS • ECOCOLOGICAL GENETICS • ECONOMIC VALUE OF BIODIVERSITY • GENE BANKS • PLANT SOURCES OF DRUGS AND CHEMICALS

Bibliography

International Service for National Agricultural Research (ISNAR), The Hague, Netherlands.


I. Flight: Constraints and Opportunities
II. Habitat and Bird Species Diversity
III. Geographical Variation in Bird Species Diversity
IV. Species Replacements throughout the World
V. Diversity on Islands
VI. Human Causes of the Loss of Avian Biodiversity

GLOSSARY

biomass The total dry mass of a population or community of animals or plants.
congenor A species that belongs to the same genus as the one under discussion.
elfin woodland Small, stunted tree growth characteristic of forests at higher elevations in warm, moist regions.
endothermic An animal whose body temperature is largely determined by its own metabolism rather than by the temperature of the environment.
epiphyte A plant that uses another plant, usually a tree, for support but not for nourishment.
niche The unique position occupied by a particular species in terms of the function that it performs within the community.

polyandrous A female that has more than one (male) mate.
polygynous A male that has more than one (female) mate.
trophic Relating to feeding.

THE LIFE OF BIRDS is based on flight, which has both constrained their diversification and opened ecological opportunities. It affects their size and their abundance. Their ecology is based on the exploitation of patchy and variable resources, migration being the most obvious manifestation of this. A few birds are flightless. The taxonomic diversity of birds is limited. Across all continents, bird species diversity is greater in wooded than in forested habitats, generally being related to habitat complexity and productivity. Coexisting species eat different foods and often use different parts of the habitat. Bird species diversity is greater where there is greater habitat diversity; it is less at higher altitudes; and it is generally less as one moves from the tropics toward the poles, with strong local modifications. It may be correlated with that of other animals, but not always. Bird data have been used to test hypotheses about the causes of the latitudinal gradient in biodiversity. There is evidence that it is partly dependent on tropical regions providing more ecological opportunities, more habitable area, and more utilizable energy. No species is found all over the world. Some closely related species replace each other geographically. Some parts of the...
world, mainly tropical, have unusually large numbers of endemic, restricted-range species. Geological and evolutionary history has resulted in the avifaunas of different regions of the world differing in richness and taxonomic origin. Unrelated species living in different places but similar habitats have often undergone evolutionary convergence. Birds are good colonists of islands, where they subsequently evolve in the presence of few (if any) other land vertebrates. More species are found on larger islands, on those that once formed parts of larger landmasses, and on those closer to sources of colonists. Islands that become cut off from the mainland or which are fragmented lose species because extinctions are not then sufficiently matched by colonizations. Humans have caused the extinction of many birds, in both historical and prehistorical times, especially on islands. The chief current threats to them are habitat loss, hunting, agricultural intensification, and the introduction of alien species to islands. Reducing these threats will be difficult. Avian biodiversity continues to be lost both by large-scale declines of many species in some parts of the world and by the introduction of species without their native ranges, which diminishes the differences in the bird communities of different lands.

I. FLIGHT: CONSTRAINTS AND OPPORTUNITIES

A. Birds as Flying Machines

The feather is one of the most remarkable structures ever to have evolved. It has allowed birds to become highly effective flying machines and to exploit resources in ways impossible for other vertebrates, although it has also placed constraints on them, considerably restricting their morphological diversification and thus limiting the ecological opportunities open to them.

Feathers provide highly effective insulation. Simply stated, the downy bases of the contour feathers trap a layer of air close to the body, and the integrity of this layer is preserved by the vanes of the feathers, which form a largely windproof and waterproof covering, like the tiles on a roof. Weight for weight, birds' plumage is considerably warmer and more waterproof than the hair and wool of mammals. Like mammals, therefore, birds are enabled to be endothermic, usually maintaining body temperatures well above their surroundings and thus leading active and independent lives.

The structure of feathers allows them to form strong but lightweight aerofoils, thus facilitating the evolution of wings. Since the wings are made up of many feathers, they are very adjustable, allowing effective and maneuverable flight. Because feathers have no blood supply, birds do not lose inordinate quantities of heat through their wings, as bats are in danger of doing. The bat wing is difficult to repair; in contrast, a damaged feather is renewed at the next molt. The strength of feathers allows birds to have long tails, which are used for both aerodynamics and social signaling, even though the bony and fleshy tail on which those feathers are mounted is reduced to a short stub.

The shape of wings varies considerably. Long, pointed wings provide high speed but little maneuverability and lift; they are characteristic of birds that need to fly fast or migrate long distances and which live in open places, such as plows Pluvialis. Short, rounded wings give maneuverability and lift but little speed; they are characteristic of woodland birds such as tits Parus. Large, broad wings allow birds to soar on rising air, covering long distances while patrolling the landscape for food; they are characteristic of large hawks, eagles, and vultures. Long, straight wings allow birds to glide on ocean winds; they are characteristic of albatrosses and other such seabirds.

Most vertebrates tend to have long, flexible bodies. Not so with birds: Flight requires a short, rigid trunk. This is achieved not only through the trunk being short and broad but also through the locking together of the skeletal elements, providing a firm base for the operation of the flight muscles while keeping skeletal weight down. Long bones are hollow, reducing weight. The compactness of the body is increased by having the major muscles of both wings and legs close to the center of mass, rendering locomotion mechanically efficient. Because the trunk is rigid, the neck must be long so that the head can be moved around. (The length of the neck is generally not apparent because of the covering of feathers.) Were birds to have a tidal airflow ventilating the lungs, this would be a major disadvantage, because in such a system there is a "dead space" lying between mouth and lungs into which air is drawn but never used. Birds have a through-flow so that all inspired air passes through the lungs. This is why they do not need to breathe as often as mammals of comparable size, even though they operate at the high metabolic rate that flight demands. Birds' hearts are larger than those of similarly sized mammals to underpin their rapid metabolism.

The long neck means that it is particularly important for the head to be light. Modern birds are toothless and do not have the heavy jaws typical of many reptiles and mammals. Although their beaks and tongues are capable of remarkable feats of manipulation (such as
finches husking seeds in seconds), this means that food is largely swallowed without chewing. To compensate, most species have part of the stomach modified as a muscular gizzard within which the food is ground. Indigestible elements of the food, such as bones, hair, arthropod exoskeletons, and husks, tend to be regurgitated in pellets.

As bats demonstrate, flight does not absolutely preclude viviparity; however, birds are the only vertebrate class in which no species bears live young. Weight restrictions result in most bats producing only one young per year, which prevents them from adopting "r-selected" lifestyles characterized by rapid reproductive rates. Some birds, in contrast, may produce well over 10 young each year. In those many species in which the young remain in the nest until fully grown, the parents’ ability to fly allows them nonetheless to forage over comparatively large areas.

Being so mobile in three dimensions, birds use vision and hearing as their dominant senses; they communicate mainly by sight and sound.

B. Size and Abundance

The constraints imposed by flight are nowhere as apparent as in the upper limit of bird size. The lifting power of wings depends on their surface area and on the cross-sectional area of the muscles powering them; however, the weight to be lifted depends on the volume of the body, so the ratio of weight to power increases with the size of the animal, imposing an upper limit on the size of flying animals. This appears to be approximately 15 kg for modern birds, exemplified by some swans, pelicans, and bustards, although some extinct species seem to have been appreciably larger. Flightless birds may be larger—more than 100 kg in the case of the ostrich Struthio camelus and approximately 450 kg in the extinct elephant bird Aepyornis maximus; however, even these are dwarfed by African elephants Loxodonta africana at 7 tons and blue whales Balaenoptera musculus at 100 tonnes. Flightless birds have no particular advantage over mammals and there are few of them.

Metabolic rate seems to impose similar lower limits on the size of birds and mammals: Kitti’s hog-nosed bat is 1.5–2.0 g, shrews such as the pygmy white-toothed shrew Suncus etruscus 2–3 g, and the bee hummingbird Calypte helene 1.6–3.0 g. The problem is that metabolic rates of small endotherms are disproportionately great, partly because their large surface to volume ratio results in disproportionate heat loss. Not only is it difficult for them to ingest enough food to maintain such high rates but also there is an upper limit to the rates imposed by the rate at which their hearts can beat. The 1200–1400 contractions per minute observed in the smallest species may be the maximum achievable.

Most birds are small—half of the species weigh less than 38 g (Fig. 1). They are larger toward the poles (Fig. 2), perhaps because it is easier for large birds to keep warm or because they are more able to migrate long distances. Within the same region, smaller birds are more common in wooded than in open habitats; the former give them better protection from weather and predators as well as provide many foraging opportunities among twigs and foliage, which heavy birds can
not exploit. Waterbirds tend to be large to aid heat conservation.

In many birds, males are slightly larger than females; they are particularly large in many polygynous species, which suggests that the reason for the difference is the advantage that larger males have in competing for females. Females tend to be larger in polyandrous species and also in species in which males perform aero-batic displays as part of courtship.

On a logarithmic scale, the abundances of bird species in a region are approximately normally distributed, with some skew because of a tail of very scarce breeders (Fig. 3). The same is probably true on a world scale. The Seychelles warbler Acrocephalus sechellensis was once confined to Cousin Island (29 ha), its population reaching a low of 50 pairs. Numerous other species confined to small, remote islands may have maintained populations of no more than a few hundred throughout their history. Various species have total world populations much less than this but usually as a result of human interference. Wilson’s petrel Oceanites oceanicus is probably the most common seabird, at 100 million or more. On land, the red-billed quelea Quelea quelea, of which individual flocks may number over 1 million, is almost certainly even more numerous; house sparrow Passer domesticus and startled Starlings are so widespread, partly through human introduction, that they may have total populations even greater than those of Q. quelea, which is confined to the more arid parts of sub-Saharan Africa.

At both local and regional scales, smaller birds are more abundant than larger ones (Fig. 4). Species that are more abundant in places where they occur occupy more sites within regions (Fig. 5) and have more extensive geographical ranges (Fig. 6), though on both scales colonial species (which occupy few sites but at high densities) weaken the relationship. The correlation tends to be weaker outside the breeding season, perhaps because many noncolonial species then form flocks. Various reasons for this relationship have been suggested, but the evidence favors none in particular.

In Britain at least, summer visitors tend to be smaller, on average, than residents, perhaps reflecting difficulties that small birds have in keeping warm in winter. They are substantially less abundant than residents of the same size (Fig. 4). (In North America, summer visitors are more abundant than residents; this is also true in Europe if one does not consider size.)

C. The Ecological Diversity of Birds

A 1-kg mammal can travel at sustained speeds of 6 km/hr, expending 118 kJ/km in doing so. A bird of the same size can travel at 47 km/hr, expending only 49 kJ/km. The power of flight thus allows birds to cover larger areas in the courses of their lives and this is the
key to their ecology: They tend to use scattered and variable resources much more than do earth-bound animals. As a result, they are able to occupy significantly narrower ecological niches, subdividing resources more finely. This may be the reason why, within the same geographical area, the abundance of individual species of nonflying mammals is very much greater than that of birds of the same size (Fig. 4). This, plus the fact that the average mammal is several times larger than the average bird, means that birds make up much less of the total biomass of terrestrial animals than do mammals (e.g., 13,000 tons for British resident birds versus 158,000 tons for mammals).

In contrast to such local restrictions on numbers, birds are able to occupy large geographical ranges because their powers of flight allow them to discover areas that provide them with the resources they require. Thus, local areas tend to be occupied by a higher percentage of the whole continental species pool of birds than is the case for other well-studied groups, such as mammals, amphibians, lizards, and butterflies.

Many waterbirds make seasonal use of water bodies that are impermanent or which freeze over in winter. On a shorter timescale, some species of birds (but few nonflying animals) move onto seashores in large numbers as tides recede, exploiting resources that are often rich but which are only briefly available. Pelagic seabirds use the efficiency of flight in another way, covering large areas in search of food that may be constantly available but is sparsely distributed. Outside the breed-
ing season, they may move over thousands of kilometers, often along definite tracks determined by winds, currents, and seasonal availability of food. Seabirds quickly discover shoals of fish that are close to the surface, such as at upwelling fronts of cold water, which shift unpredictably. Even those land birds that are resident in a small area are constantly moving about that area in ways that allow them to subsist on food that is either sparsely distributed (such as seeds in temperate grasslands in winter) or which occurs in patches, especially temporary patches (such as carcasses, fruiting trees, nectar-bearing flowers, ants and termite swarms, insects disturbed by ungulate herbivores, and small animals disturbed by fire). Outside the breeding season they may temporarily use habitat patches that are inadequate to support them while breeding.

Birds are able to exploit aquatic habitats not just because of their powers of flight but also because of the effectiveness of their plumage for insulation. This is particularly important for smaller animals (because of their high surface to volume ratio) and it is noteworthy that there are few aquatic mammals weighing less than several kilograms. Those that do must frequently leave the water to dry out their fur. Insulation is less important for larger animals, and terrestrial mammals have given rise to cetaceans, sirenians, and seals, apparently excluding birds from occupying these ecological zones.

The Australian central desert, lying between two rainfall belts, may experience occasional rain at any time, anywhere; approximately 30% of Australian birds are essentially nomadic, settling to breed quickly in places where rain has fallen. Although some desert birds can survive without drinking much free water, others cannot; rather, they are able to live in deserts by using their powers of flight to reach isolated drinking places. Sandgrouse (Pteroclididae) may make round-trips of over 100 km daily between their nest sites and water holes, with males transporting water back to the chicks that is soaked into the modified breast feathers.

Although flight allows birds to exploit patchy resources, it also restricts what they feed on. To keep weight down, birds have short guts, through which food passes quickly; therefore, they tend to feed on particularly nutrient-rich items, generally eschewing food that needs to be taken in bulk. Insects (and other terrestrial invertebrates) are taken by many species but, despite leaves making up such a large proportion of the biomass on land, comparatively few birds are specialist leaf eaters (Table I). Some grouse, geese and some ducks, some pigeons, the New Zealand ground parrot Strigops habroptilus, and the hoatzin Opisthocomus hoazin are exceptions: Most tend to feed selectively on the most nutritious leaves and to move their food through the gut rapidly.

Carrion is very sparsely distributed; therefore, only a few birds specialize on it, although there has been convergent evolution on this specialization in New and Old World vultures. Similarly, nectar is only found in sufficient quantities for birds to use in some parts of the world and efficient nectar-feeding requires anatomical and behavioral specializations so there are few families for which nectar is the primary food (and, as with the vultures, there has been a convergence of New and Old World forms: hummingbirds and sunbirds, respectively).

The various diets of birds are associated with a variety of bill shapes (Fig. 7). There are subtler variations: Frugivores with wider gapes take larger fruits, and finches with deeper bills take larger and harder seeds. Even within species, differences between individuals in bill size may be reflected in diet, the most extreme example being the African finch Pyrenestes ostrinus in which there are two genetically distinct forms specializing on seed diets that differ in hardness. The extinct Huia Heteralocha acutirostris of New Zealand fed on insects in dead wood, with the female probing into

<table>
<thead>
<tr>
<th>Food category</th>
<th>Primarily</th>
<th>Regularly</th>
<th>Sparingly if at all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green plants, buds</td>
<td>2</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Seeds</td>
<td>4</td>
<td>31</td>
<td>24</td>
</tr>
<tr>
<td>Fruit</td>
<td>19</td>
<td>34</td>
<td>21</td>
</tr>
<tr>
<td>Nectar</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Insects</td>
<td>50</td>
<td>58</td>
<td>32</td>
</tr>
<tr>
<td>Other terrestrial invertebrates</td>
<td>2</td>
<td>10</td>
<td>23</td>
</tr>
<tr>
<td>Litoral and benthic invertebrates</td>
<td>6</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Small vertebrates (&lt;5 kg)</td>
<td>4</td>
<td>14</td>
<td>29</td>
</tr>
<tr>
<td>Large vertebrates (&gt;5 kg)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fish, crustaceans, squid, etc.</td>
<td>20</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Carrion</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

a From Morse (1975).
BIRDS, BIODIVERSITY OF

FIGURE 7  Examples of beaks used for various ways of feeding: 1, flamingo, modified for filter feeding; 2, toucan, elongate for reaching suspended fruit and secondarily decorated for social signaling; 3, merganser, serrated for gripping fish; 4, woodpecker, chisel-shaped for probing wood; 5, crossbill, specialized for extracting seeds from cones; 6, hummingbird, elongate for reaching nectar; 7, warbler, fine for picking up insects; 8, puffin, deepened for carrying several fish at once and secondarily decorated for social signaling; 9, shrike, hooked for tearing prey; 10, swift, wide for catching insects in flight (illustration by Su Gough).

D. Migration

The most dramatic manifestation of birds' use of varying resources is migration. Huge areas of North America, Europe, and Asia become inhospitable to life in the winter, primarily because temperatures are low, the ground is snow covered, and waters are frozen. Endothermic animals that could perhaps survive the physical conditions are often unable to survive in these regions because their food supplies are inadequate. Therefore, there are relatively few permanently resident birds. In the summer, however, there is an abundance of food and this presents an opportunity that is taken by hundreds of millions of birds from south temperate, subtropical, and tropical regions that move north to breed. Approximately 87% of the species breeding at 80°N in eastern North America are summer migrants, the proportion declining to 12% at 25°N. The same north–south pattern is seen in Europe but the proportion is consistently about 17% lower at equivalent latitudes because the warmer winters of western Europe allow more species to be year-round residents than in eastern America.

Not all movements are north–south. The comparatively warm (and damp) winter conditions on the western edge of Europe allow many waterfowl and shorebirds to winter there (such that British wetland sites tend to be more species rich in winter than in summer); these birds not only move north in summer but also move east, into continental regions where winter conditions are intolerable. Summer visitors are generally more common in eastern than western Europe. Other species show altitudinal migrations, moving up into mountains to breed during short seasons of relative plenty. Within the tropics, savanna birds often have migrations that track the seasonally moving rainfall belts.

The seasonal variations in bird communities in particular areas may be complex. In the more benign parts of temperate regions, for example, such as the southern United States and Mediterranean Europe, winter visitors from the north and summer visitors from the south effectively replace each other. Thus, 52% of the wintering species at 25°N in eastern North America are visitors that have bred in the north; the proportion declines further north, to zero at 70°N. A similar pattern is seen in Europe, but there are about 10% fewer northern migrants at equivalent latitudes.

Summer migrants to North America mostly come from Mexico, Central America, the Caribbean, and northern South America rather than beyond the equator. These places are, if not forested, well wooded. Paleartic migrants come from similar latitudes, with 35% wintering wholly north of the equator and only 3% wholly south of it. In contrast to America, African migrants come almost entirely from savannas; these are not only much more extensive than forests in tropical Africa but also much more seasonal.
In much of Europe, summer visitors seem to occur particularly in grasslands and scrublands. In British and Irish woodlands, migrant passerines are most common in early successional stages and also in the upland oak-woods of Wales, where residents are less abundant than in lowland woods, migrant species are generally less abundant and less widespread than residents and use a less diverse array of nest sites. These patterns fit with residents, where they can survive the winter, preempting habitats, with migrants utilizing the summer surplus that the residents cannot fully take up.

There is much less land in the Southern Hemisphere than in the north. As a result, in both Africa and South America, although there is some migration between tropical and south temperate regions, it is less dramatic than the northern migration. As in the north, relatively few of the migrants cross the equator.

Although generalizations about migration can be made, there is much variation between species in details, as is apparent even from broad considerations of patterns of migration (Table II). Most migration is within-tropical, within-temperate, or tropical-temperate but there are a few species, such as American golden plover *Pluvialis dominica*, that move between temperate zones in the two hemispheres. At the extreme, some arctic terns *Sterna paradisaea* breed in the Arctic but winter in the Antarctic.

There is variation even within species: Populations (and even individuals within populations) may differ in whether they are resident or migratory and, if migratory, how far they migrate and by what routes. ‘Leapfrog’ migration is not uncommon. In the fox sparrow *Passerella iliaca*, for example, populations of Vancouver Island and the nearby mainland are resident; birds wintering in Oregon breed in the far southeast of Alaska and northern British Columbia; and those wintering in California breed further north and west, along the Alaskan peninsula.

### E. Flightless Birds

A few bird species are flightless. Given that it is the demands of flight that are responsible for the restricted morphological diversity of birds, it is not surprising that some of these flightless species show marked departures from the typical avian body plan.

Many species that have colonized remote oceanic islands have become flightless. Such islands hold few, if any, other terrestrial vertebrates so there is a variety of vacant ecological niches available for birds to exploit, including those based on resources that, being neither patchy nor unpredictable, do not demand the power of flight for their exploitation. Nor is flight needed as a means of escape in the absence of predators. Among the flightless species on remote islands, rails (Rallidae) feature prominently. Rails generally lead a sedentary existence, using their powers of flight mainly for long-distance migration and to find isolated habitat patches. Island rails have typically achieved flightlessness, as have some other species (ratites in particular), by neo-teny, i.e., the retention of juvenile features into adulthood. The key retained features are that the keel on the sternum (on which the massive flight muscles are inserted in flying birds) is absent (or much reduced), the legs are comparatively large, and the plumage is often looser than in typical birds, because the barbs that make up the vane of the feather lack the barbules that, in typical feathers, link adjoining barbs together to make the vane a continuous plate.

Among the flightless forms, one species, *Raphus cucullatus*, was the dodo of Mauritius and the two species of solitaire (*Pezophaps*) of Rodriguez and Reunion. Turkey-sized birds descended from doves, these species are typically pictured as exceptionally ponderous, although it has been argued that they played a marked annual cycle of fattening and that for much of the year they were far slimmer than usually portrayed. They had massive legs and bills and appar-

### TABLE II

Percentages of Birds (Excluding Seabirds) Breeding in the Western Palearctic According to the Way in which Their Range Changes in Winter

<table>
<thead>
<tr>
<th>Range Changes in Winter</th>
<th>Land</th>
<th>Freshwater</th>
<th>Coastal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range is unchanged</td>
<td>44</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>North edge of range moves south (range contracts)</td>
<td>4</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>South edge of range moves south (so range expands)</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Whole range moves south but summer and winter ranges overlap</td>
<td>18</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>Whole range moves south, summer and winter ranges do not overlap</td>
<td>32</td>
<td>16</td>
<td>46</td>
</tr>
<tr>
<td>Total number of species</td>
<td>300</td>
<td>63</td>
<td>26</td>
</tr>
</tbody>
</table>

* From Newton and Dale (1997).
ently occupied the niche of generalist herbivores, taking large seeds and foliage.

Apart from bats, New Zealand remained mammal free from its time of isolation from Australia 73–80 million years ago until colonized by man about 1000 years ago. In its forested landscape, moa (Dinornithidae) evolved. These were moderate to large birds, somewhat ostrich-like but much more sturdily built, and they usually held their necks horizontal rather than upright; they roamed scrubland and forests in search of seeds, fruit, and shoots. There were about 11 species, all extinguished by man. Kiwis (Apteryx), however, survive in small numbers. They are behaviorally and anatomically highly modified for feeding on soil invertebrates, with the nostrils on the tip of the elongate bill (unique among known birds). They are nocturnal. Apart from their locomotion being bipedal, they look remarkably mammal-like when rooting along the forest floor or across pastures in search of food. Another remarkable New Zealand species is the kakapo (Strigops habroptilus), an herbivorous flightless parrot; although a few other parrots are largely terrestrial, this is the only flightless species. The moa-like elephant birds (Aepyornithidae) of Madagascar also evolved in isolation, with their final extinction apparently coinciding with the arrival of man. What of the surviving large flightless birds—the ostrich (Struthio camellus) of Africa, the two rheas (Rheidae) of South America, the emu (Dromaius novaehollandiae) of Australia, and the three cassowaries (Casuarius) of Australia, New Guinea, and nearby islands? It is possible that all except the ostrich evolved in the absence of predators, although they have encountered them since; ostriches and rheas, living on open plains, can run at great speed. All have continued the bird habit of feeding on scattered food (fruit, seeds, flowers, shoots, and the more nutritious leaves, plus small animals in the case of emus and rheas), covering large foraging areas on their long legs and using the reach of their long necks to pick out food items.

The ecological gap left by the extinction of carnivorous bipedal dinosaurs led to the evolution of various giant flightless predatory birds but these seem to have been unable to survive competition from modern carnivorous mammals. Waterbirds use two different techniques to swim under water, both of which result in a predisposition to the evolution of flightlessness. Some use their feet to drive themselves through the water. Thus, in addition to retaining the massive wing muscles required for flight, these birds must also have particularly powerful hindlimbs. More important, efficient movement through the water requires an elongate body, conflicting with the benefit for flight of a short, compact body. The powers of flight are therefore reduced in many foot-swimmers; species in isolated situations have become flightless, such as the Galapagos cormorant (Nannopterum harrisi) and the Lake Titicaca grebe (Centropelma micropterus).

The second underwater swimming technique is to use the wings, as do the auks (Alcidae) of northern seas. Since water is denser than air, the wing feathers need to be short if they are not to bend too much. Indeed, not only can the whole wing be much smaller but also it must be; otherwise, the wing muscles would not be strong enough to move it. It is not surprising that the Great Auk (Alca impennis) and other large species now extinct were flightless. In the Southern Hemisphere, even greater specialization for swimming occurred in the penguins (Spheniscidae). Their flippers are mostly made of muscle and bone, with the feathers on them being as short as those on the rest of the body. To prevent the bird from being too buoyant, the downy layer is not as well developed as one might expect for animals living in cold water; instead, penguins may carry much subcutaneous fat. Their bones tend to be solid. The elongation of their bodies is so great that they have to stand upright when on land. It can clearly be seen how giving up flight has liberated them from the constraints of body form that apply to most birds.

F. The Taxonomic Diversity of Birds

Approximately 9700 species of birds are currently recognized, which is more than twice as many as there are species of mammals. The great number of avian species is remarkable given the restrictions on morphological diversity imposed by flight, but it is undoubtedly a consequence of flight. As already considered, the mobility that flight provides allows birds to occupy finely divided niches. It also allows them to colonize isolated islands and habitat patches, where the original colonists may evolve into new species. [Note that bats, similarly equipped with the powers of flight, resemble birds in terms of their abundance/size relationships (Fig. 4) and they are the second most species-rich order of mammals, with approximately 950 species].

Table III shows the orders of birds recognized in traditional classifications of birds and those determined on the basis of studies of DNA hybridization. Many of the differences are trivial: The four orders united into the Struthioniformes on the basis of the DNA evidence have always been recognized as closely related; the splits of several orders generally correspond with long-recog-
Traditionally Recognized Orders of Birds and Those Based on DNA Hybridization, with Numbers of Species in the Latter

<table>
<thead>
<tr>
<th>Traditional</th>
<th>Sibley and Ahlquist</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rheas</td>
<td>Rheiformes</td>
<td>10</td>
</tr>
<tr>
<td>Ostriches</td>
<td>Struthioniformes</td>
<td></td>
</tr>
<tr>
<td>Cassowaries</td>
<td>Casuariiformes</td>
<td></td>
</tr>
<tr>
<td>Kiwis and moa</td>
<td>Dinornithiformes</td>
<td></td>
</tr>
<tr>
<td>Tinamous</td>
<td>Tinamiformes</td>
<td>47</td>
</tr>
<tr>
<td>Geese, pheasants, guinea fowls, and quails</td>
<td>Galliformes</td>
<td>Galliformes</td>
</tr>
<tr>
<td>Guans and megapodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swan, grebes, ducks, and screamers</td>
<td>Anseriformes</td>
<td>Anseriformes</td>
</tr>
<tr>
<td>Woodpeckers and barbets</td>
<td>Piciformes</td>
<td>355</td>
</tr>
<tr>
<td>Javanans</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hornbills</td>
<td>Bucerotiformes</td>
<td>36</td>
</tr>
<tr>
<td>Rollers, motmots, kingfishers, and bee-eaters</td>
<td>Coraciiformes</td>
<td>Coraciiformes</td>
</tr>
<tr>
<td>Hoopos</td>
<td>Uropeliformes</td>
<td>10</td>
</tr>
<tr>
<td>Trogonidae</td>
<td>Trogoniformes</td>
<td>39</td>
</tr>
<tr>
<td>Mousebirds</td>
<td>Columbiformes</td>
<td>8</td>
</tr>
<tr>
<td>Cockatoos and hoatzin</td>
<td>Cacatuidae</td>
<td>143</td>
</tr>
<tr>
<td>Parrots</td>
<td>Psittaciformes</td>
<td>358</td>
</tr>
<tr>
<td>Swifts</td>
<td>Apodiformes</td>
<td>103</td>
</tr>
<tr>
<td>Hummingbirds</td>
<td>Trochiliformes</td>
<td>319</td>
</tr>
<tr>
<td>Turacos and plantain-eaters</td>
<td>Musophagiformes</td>
<td>23</td>
</tr>
<tr>
<td>Owls</td>
<td>Strigiformes</td>
<td>291</td>
</tr>
<tr>
<td>Nightjars and oilbirds</td>
<td>Caprimulgiformes</td>
<td></td>
</tr>
<tr>
<td>Doves and doves</td>
<td>Columbiformes</td>
<td>313</td>
</tr>
<tr>
<td>Sandgrouse</td>
<td>Pteroclidiformes</td>
<td>17</td>
</tr>
<tr>
<td>Shorebirds, plovs, terns, and alasks</td>
<td>Charadriiformes</td>
<td>Charadriiformes</td>
</tr>
<tr>
<td>Grebes</td>
<td>Pelecaniformes</td>
<td></td>
</tr>
<tr>
<td>Cormorants, pelicans, gannets, etc.</td>
<td>Falconiformes</td>
<td>Falconiformes</td>
</tr>
<tr>
<td>Harleks, eagles, falcons, secretary birds, etc.</td>
<td>Ciconiiformes</td>
<td>Ciconiiformes</td>
</tr>
<tr>
<td>Petrels, shearwaters, and albatrosses</td>
<td>Procellariformes</td>
<td>Procellariformes</td>
</tr>
<tr>
<td>Divers</td>
<td>Gaviiformes</td>
<td></td>
</tr>
<tr>
<td>Penguins</td>
<td>Sphenisciformes</td>
<td></td>
</tr>
<tr>
<td>Flamingos</td>
<td>Phoenicopteriformes</td>
<td></td>
</tr>
<tr>
<td>Perching birds</td>
<td>Passeriformes</td>
<td>9712</td>
</tr>
</tbody>
</table>


The classification based on DNA hybridization groups the orders of modern birds into two “infra-classes,” Struthioniformes plus Tinamiformes (Eoaves) and all the others (Neoaves). This is a long-recognized division based on the peculiar structure of the palate and other anatomical characteristics of the first group. Eoaves are often referred to as ratites, although strictly speaking this term should not include the tinamous since it refers to the condition of the sternum in the

- **Traditionally Recognized Orders of Birds and Those Based on DNA Hybridization, with Numbers of Species in the Latter**

A long-recognized division based on the peculiar structure of the palate and other anatomical characteristics of the first group. Eoaves are often referred to as ratites, although strictly speaking this term should not include the tinamous since it refers to the condition of the sternum in the.
struthioniform birds, which is flat, without the keel on which the flight muscles of most birds (including tinnamous) originate.

Both the traditional and the DNA-based classifications recognize more orders of birds than the 21 usually recognized for mammals. Given that birds are substantially less diverse than mammals, this is paradoxical. It reflects the fact that taxonomic categories at the same formal level in the hierarchy are not comparable in different groups and that ornithologists have adopted classifications that specialists in other disciplines would regard as inflated. (Entomologists would probably reduce most ornithological families to the rank of genus.)

The outstanding feature of avian taxonomic diversity is that 90% of the species belong to one order, the Passeriformes. In contrast, the most species-rich order of mammals (the rodents) comprises only 42% of the species. The passerines are “typical birds.” They are characterized by a particular form of the foot, well suited to perching, but have no obvious single characteristic that would appear to open up ecological opportunities in the way that the special gnawing incisors seem to have done for the rodents. Perhaps their numerical dominance is simply related to their small size, given that most birds are small (see Section I, B). An additional peculiarity is that the two suborders of passerines are disparate in size, with 80% of the species falling into the suborder Passeri (or oscines), the true songbirds, and only 20% in the suborder Tyranni (or suboscines). Speciation in songbirds often seems to involve the evolution of song differences between formally conspecific populations. Perhaps the complex vocal apparatus of oscines has promoted high rates of speciation, leading to this suborder (and thus the order to which it belongs) being remarkably species rich.

Another peculiarity of birds, particularly if one considers that avian orders are comparable to taxa of lower rank in other animal groups, is that only a few of the orders contain very few species. The Coliiformes is the only order with less than 10 species, whereas six mammalian orders have fewer than 10. The most taxonomically isolated bird species, generally given families of their own, include the secretary bird Sagittarius serpentarius, a large ground-living raptor of African savanna; the hoatzin Opisthocomus hoazin, a leaf-eating oddity related to cuckoos which inhabits permanently flooded forests in South America, with an enormous specialized crop for processing its unusual diet; the kagu Rhynochetos jubatus, a strange, short-legged flightless gruiform endemic to the forests of New Caledonia; and the oilbird Steatornis caripensis, a Neotropical caprimulgid that feeds on fruit by night using sonar to navigate around the dark caves into which it retreats by day.

Despite their peculiarities, none of these is as divergent from the rest of their class as is the aardvark Orycteropus afer, placed in an order all its own among the mammals.

II. HABITAT AND BIRD SPECIES DIVERSITY

A. Differences between Habitats in Bird Species Diversity

It is a matter of common observation that habitats in the same region differ in the species diversity of the bird communities living in them. Habitats with higher diversity also tend to hold larger numbers of individual birds. However, there have been few systematic surveys of bird communities across all the habitats in a region and even fewer with the extent of replication used for the data in Fig. 8, in which each habitat was represented by at least 12 sites (with a mean of 72 sites/habitat). This confirms the following generalization: Woodlands, scrub, and carr contain large numbers of species, whereas open downland contains very few. Machair, flat plains based on deposits of calcareous sand fringing parts of the western seaboard of northern Scotland, is a conspicuous exception, having large numbers of birds despite being completely unwooded. This may be the result of the extreme small-scale heterogeneity of machair lands. There may be a transition from mobile sand dunes to stabilized grassland, from very dry to very wet conditions, and from undisturbed to cultivated ground within distances of tens of meters.

The distributions of the points on the two axes of Fig. 8 are correlated: Woodland tends to have many of the more ubiquitous species (occurring in more than 75% of the sites) and the less ubiquitous (occurring in only 25–75% of sites), whereas open downland holds few of either. The scatter around the correlation, however, reveals further differences between habitats. For example, scrub is an ephemeral habitat, typically occupied by species that are good colonists and which therefore tend to be found in the majority of sites; as a result, two-thirds of the species found in more than 20% of scrub sites are actually found in more than 75% of sites.

B. Bird Species Diversity in Similar Ecosystems on Different Continents

If habitat is a major determinant of bird species diversity, one would expect to find similar levels of diversity in similar vegetation and landscape types on different
continents. Often one does, but sometimes one does not (Table IV). Even where total species numbers are similar, the numbers of species within individual feeding categories may differ. Thus, although numbers of bird species in most feeding categories are similar in Mediterranean scrub in Provence (France), Chile, and California, Provence has no nectarivores and only 4 granivores, in contrast with 3 and 14, respectively, in California and 2 and 6 in Chile. Again, the species richness of semi-arid habitats is similar in North America and Australia, but the bird communities are clearly different in detail, partly perhaps as a result of different evolutionary histories but partly because rainfall is more erratic in Australia than in North America. Thus, the habitats are similar but the overall environmental conditions differ significantly. Despite differences in species number, the number of species in the various feeding
categories tends to depend more on habitat itself than on what continent the habitat is in (Table V).

C. Bird Species Diversity and Habitat Complexity

Although it is not a universal rule, most studies have tended to confirm the discovery by R. H. and J. W.

**TABLE IV**
Numbers of Bird Species in Similar Habitats on Different Continents

<table>
<thead>
<tr>
<th>Habitat</th>
<th>No. of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peatlands</td>
<td>33 (Finland) 19 (Minnesota)</td>
</tr>
<tr>
<td>Desert</td>
<td>57 (Arizona) 61 (Argentina)</td>
</tr>
<tr>
<td>Shrub desert</td>
<td>5.5 (Australia) 6.3 (N. America)</td>
</tr>
<tr>
<td>Mediterranean scrub</td>
<td>30 (California) 20 (S. Africa)</td>
</tr>
</tbody>
</table>

* From a compilation by Schluter and Ricklefs (1993).

**TABLE V**
Numbers of Bird Species Recorded from Mediterranean Scrub and Wet Tropical Forests in Africa and South America, Grouped into Foraging Categories

<table>
<thead>
<tr>
<th>Foraging category</th>
<th>Forest Africa</th>
<th>South America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthropods from foliage</td>
<td>77</td>
<td>59</td>
</tr>
<tr>
<td>Seeds and fruit</td>
<td>20</td>
<td>49</td>
</tr>
<tr>
<td>Arthropods from sallies</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>Trunk, bark, etc.</td>
<td>13</td>
<td>31</td>
</tr>
<tr>
<td>Ground</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>Raptors and scavengers</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Nectar</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Aerial and crepuscular</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>163</td>
<td>209</td>
</tr>
</tbody>
</table>

MacArthur that bird species diversity is positively correlated with foliage height diversity, at least in scrub and woodland habitats (Fig. 9). That is, bird species diversity is greatest where there is an equal volume of foliage at all heights in the woodland. An early finding was that data points from Australia fell on the same line as those from the original study in North America. However, MacArthur discovered that data from Panamanian and Puerto Rican sites only fell on the same line if the measure of foliage height diversity was adjusted by changing the height subdivisions on which it was based. His interpretation was that Panamanian birds were effectively treating the forests as being divided into more layers than were North American birds, whereas Puerto Rican birds were treating them as being more coarsely divided. Similar variations have been found elsewhere and may be related to the productivity of the different areas. Where there are more resources available, species may divide them into narrower ecological niches; perhaps more simply, more resources allow more individual birds to live in an area and thus allow even relatively scarce species to maintain viable populations.

Along an altitudinal gradient from lowland rain forest at 400 m to elfin woodland at 3600 m in the Peruvian Andes, bird species diversity and foliage height diversity decline approximately in parallel. However, the magnitude of the decline differed according to the diets of the birds in question. It was more than fivefold for insectivores but only two- or threefold for frugivores, whereas the species diversity of nectarivores was similar at all altitudes. These differences may be explained in terms of the resources available to different groups: Whole groups of insects are absent at higher altitudes; less fruit is produced at higher altitudes but a higher proportion is available to the birds because there are fewer mammalian competitors than there are at lower altitudes; and because of the distribution of flowers on the vegetation, the availability of nectar only increases slightly with foliage height diversity and, although the annual production of nectar is greater at lower altitudes, it is highly seasonal and therefore can support no more birds than the less abundant but more constantly available nectar of the high-altitude forests. Thus, foliage height diversity is a crude measure of the diversity of ecological niches available to birds but it may become less satisfactory when one considers individual trophic guilds.

Since foliage height diversity is only a rough measure of niche diversity, it is easy to find other features of the habitat that influence bird species diversity. Thus, a survey of various sites in the Americas showed that although there was a good correlation between bird species diversity and foliage height diversity, the percentage vegetation cover of the ground also influenced the number of bird species. In other words, the total volume of vegetation was important, not just how it was structured. In hardwood forests of New Hampshire, for example, tree species differ in their value to birds: Some hold more food than others and some are easier to search (e.g., yellow birch Betula allegheniensis, with small leaves and short petioles); white ash Fraxinus americana has a distinct foliage structure that allows fewer foraging methods to be used than do other species. Furthermore, floristic
diversity is important: Even a few conifers in a hardwood forest allow in birds that specialize on such trees. In south Swedish forests the amount of standing dead timber (which is largely determined by how the forest is managed) chiefly determines bird species richness (Fig. 10), with the density of low trees and shrubs having only a slight effect.

The most extreme case of the influence of factors other than foliage height diversity on bird species diversity occurs in forests in Patagonia, in which the relationship between birds and foliage diversity appears to be reversed (Fig. 11). The reason for this is that the forests are of two quite different types and the type with the greater foliage height diversity (Nothofagus forest) has relatively few birds because it comprises patches that are isolated from other such forests by 2000 km of steppes and forests; furthermore, these forests are at high elevations, with harsh climates. The abundance of bamboo, which appears to support few insects, in the understory of some of the Nothofagus forests may also be important. Note that even these species-poor forests hold more species of birds than grasslands in the same region (Fig. 11), fitting the usual pattern.

Horizontal complexity is also important, with more varied places holding more bird species. Edges between habitats may be bird rich for this reason, although “edge effects” are variable both in nature and in cause. In Britain, woodland birds are generally more abundant and more diverse at edges, probably because foliage height diversity is greater there than deep in the woodlands, where the lower levels of the vegetation tend to be shaded out. On Finnish islands, bird species diversity is lower on the edges of woodlands than deeper within perhaps because of the importance of physical protection during the early nesting period, when most of the trees have not yet begun to leaf.

D. Foraging Niches and Habitat Structure
Classic studies of the foraging behavior of tits in broad-leaved woodlands in southern England were conducted by David Lack and colleagues. In summer, all species tend to feed largely on the glut of caterpillars in the
canopy; in winters during which the mast of beech *Fagus sylvatica* is abundant, they descend to the ground to feed on it. However, during the average winter their feeding niches are distinct. The great tit *Parus major* feeds mainly on the ground on beech mast, other large seeds, and larger insects. The blue tit *Parus caeruleus*, only about half the weight of the great tit and thus much more agile, frequently hanging upside down to reach its food, feeds high up in the trees, searching for small insects among the twigs, buds, and leaves. The less common marsh tit *Parus palustris*, although no heavier than the blue tit, has a larger beak and feeds on middle-sized insects, which it gathers in the shrub layer or from the twigs and branches of trees, lower down than where the blue tit feeds. Similar studies of other birds in many parts of the world have given similar results: Related species coexisting in the same forests and using broadly similar foods tend to differ in respect of the parts of the trees in which they forage. However, foraging height is not the only factor. Thus, the coal tit *Parus ater* in English broad-leaved woodland tends to forage at approximately the same height as the marsh tit, but it is a smaller species with a finer bill and it feeds on tiny insects that it finds in crevices on the bark. In addition, it is scarce in broad-leaved woods because it is a conifer specialist; its numbers are significantly enhanced when conifers are present—another example of the importance of floristic diversity for bird species diversity.

It is such differences in foraging site and food taken that allow different species to coexist in the same habitat. Introduced birds on Hawaii provide an unplanned experimental demonstration of how differences allow coexistence: Some species persisted for more than 10 years but then became extinct; extinction was particularly likely for species that had morphologically similar relatives (presumably taking similar diets) living in the same forest. Foraging niches sometimes, but not always, shift in apparent response to the presence or absence of competitors: Willow tits in northern Finland avoid the foraging positions of great tits *P. major* and crested tits *Parus cristatus* when in mixed flocks, but if either of these is absent from a flock the willow tits use the vacated niche.

Different strata in forests differ not just in which species occupy them but also in their overall species richness. In the example shown in Fig. 12, the forest floor, relatively uniform and simple in structure, supports few species—as does the high canopy, which is also relatively simple in structure. The middle layers support many more species because they are more complex, presenting opportunities for foraging on leaves, branches, bark, dead wood, epiphytes, termite nests, and accumulations of dead leaves in tree forks and for sallying from branches to capture insects flying in the open spaces below the canopy. Underlying the overall pattern of species richness are marked differences between feeding classes (Table VI): Mast is found either high in the canopy or fallen to the forest floor, so mast feeders occur only in these layers; gleaners are also

![Table VI](https://example.com/tables.png)

<table>
<thead>
<tr>
<th>Height class (m)</th>
<th>0–3</th>
<th>3–10</th>
<th>10–15</th>
<th>15–20</th>
<th>&gt;20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mast</strong></td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><strong>Fruit</strong></td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td><strong>Insects</strong></td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><strong>Glance</strong></td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td><strong>Sally</strong></td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td><strong>Bark</strong></td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td><strong>Fruit/predator</strong></td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td><strong>Fruit/insects</strong></td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

* From Terborgh (1980).
mostly found low down or high up, working over the continuous substrates of the forest floor or the canopy; and salliers tend to occupy middle levels, where the open structure allows them to see and pursue flying insects. The canopy is a much patchier habitat than the understory because the crowns of individual trees occupy large areas, whereas the understory is composed of much larger numbers of plants; thus, the canopy presents particular opportunities to adaptable generalists that can switch, for example, from feeding on fruit to feeding on insects depending on what is immediately available.

E. Bird Species Diversity and Succession

Where sheep grazing is light enough to allow it, scrub invades open hill land in Wales, eventually developing into mature sessile oak (*Quercus petraea*) woodland. This vegetational succession is accompanied by an almost monotonic increase in the numbers of both individual birds and bird species (Fig. 13). Most of the change involves increased vertical structure, although the latest stages result from gaps developing in the woodland canopy, with small patches of open ground and scrub. As chalk downland in southern England is succeeded by ever-denser scrub, species numbers increase monotonically from 1 or 2 per study site to approximately 15 (Fuller, 1982). More generally in Britain, the scrub that develops on downland and heaths and the carr that develops on rich fens hold more species overall and more of these species are found on a high proportion of individual sites (Fig. 8).

The increase in bird species diversity may slow in later stages of succession. In a subalpine northern American sere comprising meadow, aspen (*Populus tremuloides*), fir (*Abies lasiocarpa*), and spruce (*Picea engelmannii*), the number of bird species averaged over 3 years was 3, 17, 22, and 21, respectively. Indeed, although as yew *Taxus baccata* woodlands in southern England mature they gain species steadily (10 species at 60–80 years old and 20 species at 100–200 years), the succession from open scrub with small stands of trees to the young closed yew woodland is marked by a decline from 26 to 10 bird species. Polish deciduous forest stands developing after clear-cutting reach their maximum bird species diversity when 100–150 years old but do not do so monotonically. Following a peak at 15–20 years, diversity declines for the next 15–20 years. Local factors may be important in determining the relationship between bird species diversity and habitat gradients, such as the gradient from very low matorral (0.5 m high) to mature forest of holm oak (*Quercus ilex*). Whereas the number of bird species increases along the vegetation gradient in mainland France, it decreases in Corsica (Table VII) because Corsican forests are very restricted and the island simply lacks many forest species found on the mainland, where they occur in large blocks of continuous forest. An additional complication to the story of Mediterranean holm oak forests is that, although the number of bird species in an area is reduced by fire, the reduction is less than expected from the change in vegetation structure. Some characteristic forest birds persist in the open areas and matorral created by the fire either because individual birds simply occupy the same territories or because they spill over from populations in nearby unburned forest. Human management can produce habitats in which the normal pattern is disrupted. For example, when English woodlands are coppiced (the trees are cut to the ground but allowed to regenerate from the stumps), the dense, bushy regrowth is colonized by numerous species, but

<table>
<thead>
<tr>
<th>Stage</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corsica</td>
<td>29</td>
<td>24</td>
<td>32</td>
<td>24</td>
<td>23</td>
<td>18</td>
</tr>
<tr>
<td>France</td>
<td>11</td>
<td>13</td>
<td>11</td>
<td>19</td>
<td>26</td>
<td>23</td>
</tr>
</tbody>
</table>

BIRDS, BIODIVERSITY OF

FIGURE 14 Species richness of (a) birds, (b) butterflies, and (c) leaf-litter ants along a gradient of increasing habitat modification in Cameroon. 1, Near-primary forest; 2, old secondary forest; 3, partial manual clearance, plus plantation; 4, partial mechanical clearance, plus young plantation; 5, complete clearance, with young plantation; 6, manually cleared farm fallow (redrawn from Lawton et al., 1998).

many of these are lost as the coppice ages and the tall regrowth forms a dense canopy that shades out the majority of the lower layers, producing a uniform, structurally simple habitat.

Even where the relationship between bird species diversity and habitat follows the typical pattern, this may not be simply because structurally less complex habitats provide fewer niches. Thus, bird species richness in cleared areas in Cameroon is much less than that in the forest (Fig. 14), but this may be largely because the sources of open-ground species are savannas and grasslands that lie well beyond the surrounding forest so that the cleared areas may not have been fully colonized. (Note that this case also illustrates that just as different trophic groups of birds differ in their response to vegetation structure so do different taxa of animals generally: Butterfly species richness is affected by forest clearance in much the same way as that of birds but leaf-litter ants are more species rich in partially cleared and second-growth forests.)

In contrast to the absolute level of species diversities, the turnover of species during succession has been little studied. In Polish oak-hornbeam forests, there is a rapid shift in the early stages from species characteristic of open ground to those characteristic of scrub, but the rate at which species drop out of or enter the community slows as the succession progresses. In Finnish coniferous forests, in contrast, turnover peaks not at the start of succession but at 5–25 years.

F. Bird Species Diversity and Productivity

Many of the habitat differences that seem to influence bird species diversity are probably associated with differences in total biological productivity. Even with habitats of uniform physical complexity, such as the lakes of a region, differences in productivity are associated with large differences in the numbers of both individual birds and species present.

III. GEOGRAPHICAL VARIATION IN BIRD SPECIES DIVERSITY

A. The Relationship between Bird Species Diversity and Habitat Diversity

Different habitats have different bird communities. The greatest differences within geographical regions are between marine and terrestrial habitats since life on land and at sea requires different specializations. On land, the greatest differences are between bird communities of wooded areas and those of open country, just as these broad habitat categories show marked differences in overall species richness. However, the differences go beyond species richness: Grasslands differ from woodlands not just in that most of the woodland species are absent but also in that they have their own grassland specialists. The suites of species occurring in different habitats may be similar in size (number of species) but quite different in composition. For example, the numbers of nonmarine species occurring in more than 25% of sites in three British coastal habitats are 18 in sand dunes, 13 on rocky coasts, and 12 on salt marshes, but only 4 species are common to all three habitats and 6 common to sand dunes and rocky shores, whereas there are 7, 10, and 2 unique to sand dunes, rocky shores, and salt marshes, respectively.

Related species often differ in the habitats they occupy. For example, in Britain great tits *P. major* live in both deciduous and open conifer woods, blue tits *P.*
birds, biodiversity of

caeruleus and marsh tits P. palustris only in deciduous woods, and coal tits P. ater almost only in conifer woods.

In central Sweden, marsh tits again occupy deciduous woods and willow tits P. montanus live in conifers but on the nearby Åland Islands, from which marsh tits are absent, willow tits occupy both.

Given such differences, one would expect bird species diversity to be correlated with habitat diversity in an area. Although this is undoubtedly the case, it has rarely been documented. However, a study of the boreal birds occurring on isolated mountaintops and mountain chains in western North America showed that 91% of the variation in species number could be accounted for by differences in habitat diversity.

B. Bird Species Diversity and Altitude

The decline in bird species diversity with increasing altitude in the Peruvian Andes (see Section II,C) is typical not just of forested regions but also of open habitats and not just of the tropics but also of temperate and boreal regions. Everywhere, there are fewer species of birds at higher altitudes—unless the higher levels have escaped habitat destruction wrought by man in the lowlands or unless the contrast is between different habitats, such as lowland savanna and montane forest. One cause of such gradients is that the structural complexity and the floristic diversity of the habitat are reduced at higher altitudes. This is true of both forests and open habitats; indeed, forest gives way to open habitats at sufficiently high elevation.

The cooler, windier, and wetter conditions at high altitudes are no doubt responsible for the less luxuriant vegetation and probably also affect the birds directly so that only species capable of withstanding them can survive. Some mountains may also be so isolated from others that they resemble islands, lacking some species because they have never reached these mountains or because they have not recolonized following a chance extinction (see Section V).

C. Bird Species Diversity and Latitude

As with so many other groups, there are more bird species in the tropics than toward the poles (Fig. 15), although some groups do not conform to this generalization. Only 3 of the 17 species of penguins (Spheniscidae) breed within the tropics (where marine productivity is low) and sandpipers (Calidridinae) breed mainly in the boreal tundra and taiga (utilizing the summer flush of productivity, especially in wetter areas). The trend occurs in all habitats but may be steeper in forests than in savannas and prairies.

The broad trend is modified by local contingencies. In Argentina, the trend is steepest where forest gives way to savanna. In Africa, the reduction in species number as one moves away from the equator is steeper to the north than to the south, in accord with gradients of aridity and vegetation diversity. Northern finnish peat bogs hold more breeding bird species (and individuals) than southern ones, probably because they have willows (Salix) and therefore are more heterogeneous than the willow-free southern bogs. In Sweden, breeding passerine species are scarcer in northern than in southern bogs but waders are more diverse, with the northern bogs having more food-rich pools. In heavily urbanized locations in Finland, there is no latitudinal gradient in species diversity in winter because the birds are wholly dependent on man for food in such places. Mountain regions tend to be species rich because of their topographical and habitat diversity. Coastal areas differ from inland areas in diversity of nonmarine birds for reasons that are not clear: In North America they are comparatively poor in terms of breeding species, but in parts of Europe they are richer than inland squares.

Even at the continental scale, there are usually major departures from simple north–south trends. The pattern of aridity (and associated vegetation) largely determines bird species diversity in both Australia and Africa. In North America (Fig. 16), the east is richer in breeding bird species than the west, although the far west is...
richer than the Midwest, especially in the mountains and the southwestern desert. In the east, there is a midlatitude peak in richness. The latter is also seen in Europe. It is not simply a gross pattern that might be produced by differences in the proportion of richer and poorer habitats because it is observed within individual habitats.

Migration has major impacts on bird species diversity in many regions and habitats. It is substantially lower in the Arctic and much of the north temperate zone in winter than in summer and must vary considerably in African savannas because of migration both between Africa and the Palearctic and within Africa. This can modify geographical patterns of species richness. Thus, the difference between African forest and savanna is diminished, although by no means eliminated, in the northern winter because almost all the migrants move into the savanna; however, within the savanna the geographical pattern is maintained because areas rich in resident species are also areas to which most migrants move. In both Europe and eastern North America, the midlatitude peak of species richness in summer is eliminated in winter, when there is a simple north–south gradient.

D. Comparisons with Other Animals

There have been few direct comparisons between the geographical patterns of species richness of birds and those of other animals. In North America, a midlatitude peak in richness is also seen in mammals and amphibians, although the peak is further south, especially for amphibians. Reptiles show a simple north–south gradient. The decline in species richness in the north is steeper in mammals than in breeding birds, even steeper in amphibians, and steepest in reptiles. (Note, however, that many birds move out of the north in winter.)

Over North America as a whole, the species richness of birds and mammals are positively correlated but birds are not correlated with lizards and are negatively correlated with turtles and snakes. In Australia, there is a similar divergence: Birds are positively correlated with marsupials but negatively with lizards (which are species rich in hot, dry conditions, in which only specialist birds live).

Thus, although birds resemble other animals in showing a general tropical–polar gradient in species richness, local factors may affect them differently.

E. Causes of the Latitudinal Gradients in Species Diversity

Avian data have been widely used to illuminate possible causes of the tropical–polar gradient in species diversity. It is possible that it depends in part on there being more niches available in the tropics. Tropical vegetation is floristically more diverse and, in forests at least, structurally more complex. In North America, the midlatitude peak in bird species diversity is approximately matched by tree species diversity, as are the broad east–west trends. However, the relationship between the species diversity of birds and that of trees in North America is asymptotic; above a certain level, the richness of tree species appears not to influence that of birds. In Amazonian forests, 15% of bird species depend on habitats arising through dramatic fluctuations in river levels which do not occur in temperate regions. Furthermore, 34% of Amazonian birds belong to foraging guilds that are not found in temperate regions, such as obligate dead-leaf gleaners, obligate ant followers, and year-round frugivores; the rain forest’s richness and comparative constancy provide opportunities lacking in temperate forests. However, this is not the whole story because even where guilds are found in both Amazonian and temperate forests they contain fewer species in the latter.

The North American midlatitude peak in breeding bird species diversity has been interpreted in terms of migration. At low latitudes, there are many year-round residents but at midlatitudes there are fewer; therefore, the summer provides many opportunities for migrants to pour in and to exploit the seasonal abundance of productivity. There is indeed evidence that insect food,
although seasonally much more restricted in midlatitudes than in the south, is much more abundant when it does occur. Furthermore, it is migrants that feed on the insects that are active in summer that contribute most to the midlatitude peak; resident species such as woodpeckers, which have less seasonally variable food supplies, are not more numerous in midlatitudes than in the south. The hypothesis explains the decline in species richness in the far north (where there is also a summer flush of resources) in terms of constraints imposed by the shortness of the summer and the long journeys that migrants need to make to get there.

Another possibility is that the tropics are species rich because they include greater land areas, allowing more opportunities for speciation and reducing the likelihood of extinction. Although there is no correlation between the species richness of New World birds and land area at a given latitude outside the tropics, there is such a correlation if one excludes those species that are primarily tropical. It appears, therefore, that there is an underlying relationship between species richness and land area in the New World but that the very high richness of the tropics superimposes a wider gradient because some tropical species extend into subtropical and temperate regions.

An alternative hypothesis is that species richness is correlated with available energy in an area, either directly (warmer places allow animals to function more effectively) or indirectly, via productivity (higher productivity supports more individuals and therefore more species). Avian species richness in North America is correlated with potential evapotranspiration, which is a measure of available energy, but less well with actual evapotranspiration, which is more a measure of primary productivity. This supports the direct energy hypothesis. In Britain, temperature gradients are north–south in summer but, because of the influence of the North Atlantic Drift, east–west in winter. Data on the distributions of resident and wintering birds have been used to test the species energy hypothesis, predicting different patterns of species richness in summer and in winter. Early analyses provided support for the direct effect of energy, including indications that correlations with temperature were greater for smaller birds (which are more susceptible to low temperatures). Recent work has not confirmed this result. Nor has it found seasonal changes in distribution that match the predictions of the direct energy hypothesis. Summer temperature is the best predictor (of various climatic and topographical variables) of bird species diversity in Britain, in both summer and winter. This suggests that birds in winter depend largely on productivity stored from the summer (in the form of seeds, nuts, and invertebrates).

IV. SPECIES REPLACEMENTS THROUGHOUT THE WORLD

A. The Ranges of Birds

The barn owl Tyto alba is found in every continent of the world except Antarctica, and it is widespread in most of them. Most birds are much more restricted in terms of habitats, climate zones, and continents. Some occur on single small islands or isolated mountaintops. Differences in which species occur in which places are just as much a part of biodiversity as is the species diversity of individual places.

B. Species Replacements

Two species may have different geographical ranges because they live in different habitats that are found in different parts of the world. Alternatively, they may occupy similar habitats but simply have evolved in different places. Thirdly, they may originate in isolation from each other, subsequently come together, but be ecologically so similar that competition prevents them from coexisting, except perhaps in limited areas (Fig. 17). If they differ sufficiently in feeding habits or habitat, coexistence is possible but the extent of this varies between different sorts of birds, between breeding and winter quarters, and (within a taxon) between continents (Table VIII). The occurrence of closely related species pairs that overlap only partially, if at all, such as illustrated in Fig. 17 is common in Europe and is probably the result of the range of an original species being split in two by the great climatic fluctuations of the Pliocene and Pleistocene. The Great Plains of North America are another zone in which such pairs of species meet.

C. Centers of Endemism

The ranges of 27% of the species of land birds in the world extend less than 50,000 km². Such restricted-range species tend to occur in particular areas: Some 221 “endemic bird areas” (EBAs), each with two or more restricted-range species, have been identified. The concentration is such that 2% of the earth’s land surface contains 20% of the bird species. The tropics contain 76% of the EBAs but this may be no more than a reflection of the generally high species diversity in the tropics. Reflecting the importance of isolation in speciation, almost half of the EBAs are on islands, half of which are small and oceanic. Continental EBAs are often in mountainous regions, where both mountains and valleys can form barriers to the dispersal of species that
BIRDS, BIODIVERSITY OF

FIGURE 17 Ranges of nightingale *Luscinia megarhynchos* (dark shade) and sprosser *L. luscinia* (pale shade) in the breeding season (northern) and in winter (southern). Black areas show where the ranges overlap (redrawn from Cramp, 1988).

TABLE VIII

<table>
<thead>
<tr>
<th>Paras species (%)</th>
<th>Pass-Subtant passerine-migrants (%)</th>
<th>European birds (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N. America</td>
<td>Europe</td>
</tr>
<tr>
<td>Feeding</td>
<td>4</td>
<td>43</td>
</tr>
<tr>
<td>Habitat</td>
<td>11</td>
<td>32</td>
</tr>
<tr>
<td>Range</td>
<td>29</td>
<td>8</td>
</tr>
<tr>
<td>No contact</td>
<td>95</td>
<td>27</td>
</tr>
<tr>
<td>Unknown</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

*Data from Lack (1971).*
Many species were probably lost from these regions as a result of deterioration of the climate during the Pleistocene. Subsequently, the Caribbean Sea, the Mediterranean Sea, and the Sahara desert, although crossed by many migrants, have been partial barriers to recolonization from Africa and South America.

The Oriental region (the Indian subcontinent and Southeast Asia) has only 1 endemic bird family, although it remains somewhat isolated from the Palearctic by mountains. It has approximately 1000 species. Much of its avifauna originated largely in the Palearctic, although it has some affinities with the Ethiopian region (sub-Saharan Africa). The latter has 13 endemic families among its approximately 1500 species. It receives many Palearctic migrants during the northern winter and some of its resident species may have ancestrally originated in the Palearctic. The richness of forest birds in Africa is less than that of Central and South America, perhaps reflecting the disruption to African forests caused by dry periods in the Pliocene and Pleistocene.

The Australian region has a distinct avifauna—16 endemic families and approximately 1000 endemic species among the approximately 1600 bird species living there. Many of these belong to the Corvida, a major division of the Passeriformes revealed by DNA analysis, which appears to have originated in Australia. New Guinea and other nearby islands belong to this region, the somewhat unclear dividing line between the Australian and the Oriental regions lying between Lombok and Bali and between Celebes and Borneo, along a deep strait that has probably always separated these regions even when sea levels have been much lower than they are today. The ancestors of some Australian species have arrived from the north across this barrier, but others are derived from the south, at a time when Australia lay with the other Southern Hemisphere continents in the great landmass of Gondwanaland.

The most distinctive and most species-rich region is the Nearctic (South America), inhabited by 31 endemic families and more than 3000 species living in a variety of habitats ranging from the tropical to the far south temperate and from sea level to extensive mountains. It has a comparatively small proportion of oscines among its avifauna and seems to have had a long evolutionary independence from the rest of the world, although its passerines may be largely derived from North America.

E. Evolutionary Convergence

When unrelated forms occupy similar habitats in different parts of the world, they may evolve similar characteristics in adaptation to those habitats. Such evolutionary convergence between unrelated forms is widespread among birds: for example, the great auk, aimpennis and penguins (large, flightless seabirds using wings to swim under water; the little auk, Plautus alle and diving petrels, Pelecanoides (small seabirds that plunge into the depths to catch planktonic prey, using wings to both swim and fly); the New World vultures (related to storks) and the Old World vultures (related to hawks); the hornbills and toucans of the Old and New Worlds, respectively (both with long bills for reaching fruit in the canopy, with secondary use of the decorated bill for social signaling); and the sunbirds and the hummingbirds, also of the Old and New Worlds, respectively (small, brightly colored nectar feeders, with specialized tongues). More generally, the Australian Corvida include many groups of birds that have been considered in the past to be more closely related to various groups of passerines elsewhere in the world than to each other. Such conclusions depended on similarities that they shared with these other groups as a result of occupying similar ecological niches and thus being molded by similar selection pressures. These similarities are now recognized as convergent.

The extent of morphological convergence among whole communities has rarely been examined. A comparison of birds from Mediterranean communities in France, California, and Chile showed that they were no more similar to each other than each was to bird communities from a temperate part of France. Perhaps the environmental differences between temperate and Mediterranean France are too slight to promote significant morphological divergence.

V. DIVERSITY ON ISLANDS

A. The Colonization of Islands by Birds

Seabirds abound on islands partly because they are safer places to nest than mainlands (generally having fewer predators) but also because the oceans present no barriers to seabirds. In contrast, land and freshwater birds are less diverse on islands than on mainlands, although they have been much more successful in reaching islands than have other nonmarine vertebrates. Continental islands, which were once joined to larger landmasses before being separated by continental drift or sea-level rise, may retain some of their original species (perhaps much modified) but the evidence indicates that the more ancient of them have been colonized by many other species across the sea. Recent genetic evidence that kiwis are related to emus suggests, for
example, that their ancestors may have arrived from the north, perhaps by flying over the Tasman Sea. (Moa, in contrast, probably originate from an original Gondwanaland stock.) Oceanic islands, never joined to a mainland, must have been colonized over the sea. Va-grant individuals of a variety of species are commonly observed even on the most remote islands. The efficacy of such colonization is shown by the fact that those islands of Krakatau that were devastated by volcanic eruption in 1883 held as full a complement of species as undisturbed islands within less than 100 years. As another example, if one allows for differences in area (small islands tend to hold fewer species), oceanic is-lands are as rich in bird species as continental islands in the Gulf of California.

Even for birds, however, isolation appears to result in substantial loss of diversity (Fig. 18). Islands 250 km from New Guinea have only about half the number of bird species expected from their size, and those more than 800 km away have fewer than 10%. However, is this because birds do not reach remote islands? Perhaps they get there but do not find rich enough habitats because a sufficient wealth of plants and other animals has not arrived. The answer undoubtedly depends on the species and the island in question. Studies that focus on individual species, rather than just on counts of species number, show clearly that species differ mark-edly in colonizing ability. J. M. Diamond conducted extensive analyses of the bird faunas of the many satel-lite islands of New Guinea and found, in addition to simple distance effects, that islands that were connected by land bridges during periods of low sea level in the Pleistocene had more bird species than those that had not had such connections. Some species, apparently poor at crossing water, are found only on land-bridge islands, even though the never-connected islands have the same habitats available. Diamond divided the birds according to apparent colonizing ability. At one extreme are “supertramps,” often seen as vagrants, found on the most isolated islands that few other species have colonized, living at high densities, breeding rapidly, and having unspecialized habitat requirements. At the other extreme are species that are found only on islands that many other species have also reached; they are habitat specialists and live at low densities. When they reach an island these specialists appear to reduce the resources to such a degree that the generalist super-tramps cannot survive; the latter are therefore absent from species-rich islands. Thus, the supertramp dove *Ptilinopus solomonensis*, found on many small islands of the Bismarck Archipelago, is replaced on New Britain and New Ireland by three congeners—one in moun-tains, one in lowland forests, and the third in open lowland habitats.
B. Evolution of Island Birds

Island populations are genetically somewhat isolated (sometimes wholly isolated) from the mainland populations from which they sprang. The environment on the island is likely to differ from that on the mainland both physically and biotically (having fewer competitors and predators). As a result, there may be rapid evolutionary adjustment. Flightlessness has already been mentioned. Niches may be broadened. Thus, in most of western Europe, treecreepers Certhia familiaris are largely confined to conifers (and generally to high altitudes) by the presence of the closely related C. brachydactyla, which is a broad-leaved specialist, but in Britain and Ireland (and also in far eastern Europe and Asia), from which C. brachydactyla is absent, they inhabit woods of all sorts. Where studies of the relationship between bird species diversity and foliage height diversity have been conducted on islands, it has been found that the birds appear to divide the foliage into fewer layers than on the mainland; individual species have broader niches when there are fewer competitors. Another evolutionary change in some groups, such as ducks, is that island forms may lack distinctive male breeding plumages, either because there is less chance of hybridization when related species are absent or because demographic changes lead to behavioral modifications.

Island forms are often recognized as subspecies as a result of such changes; they may even become distinct species.

The chaffinch Fringilla coelebs is widespread in Europe and North Africa, inhabiting both coniferous and deciduous woods. It has the same habits on the Canary Islands of Fuera and La Palma, but on Tenerife and Gran Canaria it is restricted to deciduous woods; coniferous woods on those islands are occupied by the related F. teydea. These islands appear to have been colonized by F. coelebs, which then, adapting to the widespread coniferous woodlands there, evolved so much (into F. teydea) that when another F. coelebs invasion occurred the two species not only failed to interbreed but also each had a habitat to which it was better suited than the other and in which it could therefore persist. Such species pairs are known elsewhere. They represent a rare case of island bird communities being more diverse than those of mainland.

Studying the birds of the Lesser Antilles, R. E. Ricklefs and G. W. Cox modified the idea of a “taxon cycle,” earlier propounded by E. O. Wilson in respect of Australian ants to describe the distributional and evolutionary changes when a species colonizes an archipelago. Molothrus bonariensis, progressively colonizing the southern islands (from South America), represents an early stage of such a cycle; Tyrrannus dominicensis, found throughout the islands, represents the next; and then Loxigilla normis, with distinct subspecies on some islands. Finally, some of the islands have actually lost the colonizing species, either through chance extinction or through competition from later invaders: Dendroica adelaide, restricted to St. Lucia and Barbuda, is one such species.

Isolated archipelagos, with very depauperate faunas but diverse habitats and with some isolation between islands, are hotbeds of evolution for those species that do arrive. The adaptive radiation of the Hawaiian honeycreepers (Drepanididae) is a well-known example. Comprising 23 extant species (more than half the endemic bird species on Hawaii) and at least 8 extinct ones, they are derived from a single fringillid ancestor but with remarkably divergent feeding adaptations: typical finch beaks for seed-eating, parrot-like beaks, thin insect-eating beaks, and beaks elongated up to one-third the body length for sucking nectar. The akiapolaau Hemignathus wiliomi has a stout lower mandible with which it chips into decaying wood, holding out of the way its long, curved upper mandible, which it then uses to probe for insects in the exposed galleries. Darwin’s finches (Geospizinae) are even better known. The diversity of the Galapagos species of this subfamily is in sharp contrast to the existence of just 3 species on Cocos Island, about 700 km northeast. The latter, although it has a diversity of habitats, has not presented the opportunities for speciation that the many separate islands of the Galapagos have done, as a result, the Cocos finch is behaviorally flexible, utilizing a wide range of foraging opportunities.

C. The Species–Area Relationship in Birds

It is well-known that larger islands tend to hold more species (see Fig. 18). Birds have been used in many studies of the species–area relationship because they have reached most islands and their distributions are well-known. An especially steep relationship has been observed for islands in a Minnesota lake, apparently because the smaller islands studied were just too small (<0.5 ha) to support even single territories of many species. However, this is unlikely to account for the relationship observed in many studies of much larger islands. This has been widely interpreted in terms of the MacArthur–Wilson theory of island biogeography: As more species become established on an island, the rate at which new species colonize is bound to decrease...
(fewer of the arrivals are of new species); as the number of established species increases, the number going extinct also increases (there are more to go extinct); and at some point the colonizations and extinctions therefore reach an equilibrium. Since extinctions are less likely on larger islands (because populations are larger), the equilibrium number of species will be greater on larger islands. On well-studied islands, such as those around the British coast, there are certainly frequent colonizations and extinctions. These British islands are close to the source of colonists and are small, so this result is scarcely surprising, but it illustrates the dynamic nature of island avifaunas. The data also show that species differ in their probabilities of extinction; therefore, species-specific approaches are needed for a complete understanding.

The slope of the species–area relationship tends to be flatter for more isolated island groups because these are colonized only by the species with better powers of dispersal, which can quickly recoup islands from which they have gone extinct. On Bahamian islands, the slope is shallower for migrant species than for residents, the former being better colonists (and the slope is steeper for lizards, which are poor colonizers).

For species late in the taxon cycle, where extinctions have begun to take effect, the slope is steeper than for species early in the cycle, as the equilibrium theory would predict. Conversely, the slope was found to flatten in a group of Finnish islands following reduction in disturbance by man; this had previously led to many extinctions, especially on small islands. Although these observations are in agreement with the equilibrium theory, they do not rule out the importance of habitat. David Lack argued that larger islands held more bird species because they were more diverse, both physically and in terms of floristic richness. Statistical analyses designed to untangle the effects of habitat diversity and those of area per se found the former to be the chief cause of the species–area relationship in birds on the Isles of Scilly and on montane “islands” in the American Great Basin. Species richness of birds on the Lesser Antilles correlates independently with both habitat and area, in divergent contrast, bats correlate only with area and butterflies, reptiles, and amphibians only with habitat. It has been suggested that the difference is the result of birds (and bats a fortiori) having lower population densities than the other animals and therefore being more vulnerable to chance extinction. Similarly, the low species richness of the most isolated Bahamian islands seems to depend on both their isolation (many birds do not get there) and their habitat poverty (many arrivals find nowhere to live). Studies of forest fragments, which show the same sort of species–area relationship, indicate that part of the reason for the small number of species in small fragments is that they lack those species that are only found deep in forests.

Competition and resources mold island avifaunas. There is evidence that the species–area relationship is less steep for individual trophic groups than for whole avifaunas, which is in agreement with the idea that competition within the group reduces the ability of new species to colonize when an island already has many species or increases the likelihood of a new colonization leading to the extinction of an established competitor. Competition may also explain the paradoxical finding that migrant birds are more species rich on more remote Bahamian islands; the migrants, it is hypothesized, can reach all the islands but are less likely to establish on islands with residents already established (i.e., the less isolated islands).

D. Fragmentation as a Cause of Extinction

Nature has conducted experiments that reveal the importance of the area of islands in determining extinction rates and thus the numbers of species present at equilibrium. Thus, islands that were connected to New Guinea as a result of lower sea levels approximately 10,000 years ago now have fewer species than expected if they were part of mainland New Guinea (although they still have more species than do islands of comparable size that were never connected, suggesting that equilibrium has not yet been reached). Three New Guinean islands that were part of a single island approximately 10,000 years ago have fewer species than that single island would be expected to have, given its estimated size (although again they have more than expected for islands of their current size). The number of species that each has apparently lost is inversely proportional to its area, as the equilibrium theory predicts. The same is true for a suite of islands off the north coast of South America and the coast of Central America.

Man formed Barro Colorado Island during the construction of the Panama Canal early in the twentieth century. It has subsequently lost 50–60 species. Some of the losses may be the result of increased predation because man no longer hunts small and medium predators. It is striking, however, that the species that have gone extinct are those whose numbers appear to be particularly affected by the occasional periods of very dry weather that occur in the region. Average low population densities, in contrast, seem not to be a good
predictor of extinction from Barro Colorado. As long as a population does not fluctuate too much, it can persist even if its average numbers are small. Forests in most parts of the world have been fragmented by man and numerous studies have shown that fragmentation has resulted in the loss of species. This is not just the effect of random extinction, however, because it particularly involves the loss of forest-interior specialists. Indeed, direct observation in the highly fragmented woodlands of eastern England confirms that many more woodland species pass through small woodlots than stay to breed, showing that they would have no difficulty in colonizing such places if they were suitable. Blue tits P. caeruleus and great tits P. major breeding in such small woods produce fewer young, partly as a result of breeding later in the season, than do the same species breeding in larger woods. This may be because the small woods provide less shelter from the weather or because they lack sufficiently rich vegetation to produce adequate insect food for the birds. Whatever the cause, these fragmented woods support fewer birds than would a solid block of forest of the same total area.

In North America, Neotropical migrants also tend to avoid fragmented forests, much more than do resident species or temperate migrants. The fragmentation of forests may be the reason why so many Neotropical migrants have declined in numbers in recent decades.

E. Introduced Species as Colonists: A Challenge to the Equilibrium Theory?

Many birds have been introduced outside their native ranges and have flourished in their new homes. For example, 34 introduced land and freshwater species are established on New Zealand. Many of these have colonized offshore islands—up to 10 on a single island. This does not mean that those islands were not in equilibrium before these species arrived. What has happened is that the pool of species available to colonize the islands has been increased, raising the colonization rate and thus setting a new equilibrium. If New Zealand were capable of absorbing these species, does this mean that it was not at equilibrium? With only 60 resident, nonmarine indigenous species it was certainly depauperate, so one could argue that ecological niches were vacant. However, in undertaking the introductions, man was enhancing the colonization rate, thus shifting the equilibrium upward. Furthermore, many native species have been driven close to extinction in New Zealand, with some surviving only because of special protection measures. Although habitat loss and the devastations of introduced predators have been major causes of such near losses, competition from the introduced species may also have played a part.

VI. HUMAN CAUSES OF THE LOSS OF AVIAN BIODIVERSITY

A. Natural Causes of the Loss of Avian Biodiversity

The paleontological record shows that many birds have become extinct as part of the natural flux of life on Earth. The recent extinctions of an unnamed megapode on the Kermadec Islands (1876) and of the San Benedito rock wren Salpinctes obsOLEtus exsul (1952) were also natural, resulting from volcanic eruptions. That element of biodiversity that is made up of the differences between regions in the species that inhabit them must also have been eroded by natural colonizations. These natural processes, however, are slow—slow enough for extinctions to be approximately balanced by the origins of new species. Man has accelerated them manyfold, causing serious losses of avian biodiversity.

B. The History of Bird Extinctions

It is difficult to be sure whether a species has gone extinct, especially if it is scarce and lives in remote places. The Fiji petrel Pterodroma macgillivrayi, for example, was known from but a single specimen taken in 1839 until a second was found in 1984. With that caveat, Table IX shows the avian extinctions recorded in the past 400 years. These represent more than 1% of known birds, a proportion similar to that for mammals but otherwise greater than that for any other group of animals or plants.

Two things stand out from Table IX. First, extinctions peaked between 1850 and 1950 but have still continued at a high rate. Second, approximately 90% of them have been extinctions of island species, even though only 20% of bird species live on islands. One-third of the birds present on Hawaii 200 years ago are now extinct and two-thirds of the remainder are endangered. Two-thirds of the historically indigenous birds of the Mascarenes have been lost and more than 40% of the rest are endangered. Island species are partly vulnerable because their ranges are so small and some of the extinctions on continents have been of species
with particularly small ranges. Because birds that occur in few places tend to be uncommon even where they do occur (Figs. 3 and 6), they are especially vulnerable.

Hunting has caused many extinctions. As recently as 1945, Japanese troops eliminated the Wake Island rail Rallus wakensis. The case of the passenger pigeon Ectopistes migratorius is classic. Once numbered in hundreds of millions, it occurred in huge flocks that “darkened the skies.” Although habitat loss may have contributed to its decline, there is little doubt that it was extinguished by massive commercial hunting. Such uncontrolled exploitation contrasts with situations in which a community has been able to control the hunting, such as was the case on St. Kilda, where seabirds were a major resource whose exploitation was regulated. The great auk A. impennis and Dodo Raphus cucullatus, extinguished by visitors to their uninhabited island homes, provide a clear contrast and also illustrate other aspects of the vulnerability of many island birds: They were large, flightless, and had limited antipredator responses, thus making them easy and profitable to hunt. They probably had low reproductive rates and therefore were incapable of sustaining high levels of hunting.

The introduction of alien species has been a major cause of island extinctions. Worldwide, there is a clear correlation between the number of indigenous birds that have become extinct in a locality and the number of avian introductions established there. Some introduced species, especially mammals such as rats, cats, dogs, mongooses, and pigs, are important as predators. In the early 1960s, black rats Rattus rattus eliminated the last known population of the bush wren Xenicus longipes within 3 years of reaching Big South Cape Island, New Zealand, along with the island’s populations of four other birds. In 1894, the lighthouse keeper's cat caught 15 Stephen Island wrens Xenicus lyalli; these were the first specimens discovered—and also the last. Introduced monkeys are important predators at the nests of endemic birds on Mauritius. Only one introduced snake is known to have caused serious problems, but its effects have been dramatic. The brown tree snake Boiga irregularis seems to have arrived on Guam in the 1940s as an accidental hitchhiker. By 1986 it had extinguished six of the indigenous forest birds and reduced the other four to populations of less than 100; of these, the endemic Guam rail Rallus owstoni had been reduced to 100 in 1983 and it was extinct in the wild by 1987, although it survives in captivity. In contrast, none of the seven introduced birds on Guam has been obviously affected by the snake: They have effective antipredator behavior and live in habitats in which the snake is scarce. (Only one of the eight native species in such habitats has gone extinct.) It is noteworthy that few of the species extinguished by introduced predators lived on islands with indigenous rats or land crabs, which would have caused them to maintain their antipredator responses.

Introduced birds may out-compete natives, especially in the modified habitats that man has created on most islands since they may be adapted to such habitats in their own native ranges. Thus, the familiar birds in suburban and agricultural habitats of New Zealand are the familiar birds in suburban and agricultural habitats of western Europe. Competition may be genetic as well as ecological: Part of the cause of the extinction of the Atitlan grebe Podilymbus gigas may have been hybridization with the introduced pied-billed grebe P. podiceps.

The loss of natural habitats has been another major cause of extinctions. Much of it has been deliberate (the clearance of land for agriculture or human settlement), but much has been an accidental by-product of the introduction of alien herbivores, especially to islands where the native plants have evolved in the absence of such animals. Thus, the rabbit Oryctolagus cuniculus destroyed much of the vegetation on Laysan in the 20 years following its introduction in 1903; three of the five species of land birds disappeared with the plants. On some islands, soil erosion, following deforestation, has removed nest sites for burrowing seabirds.

### TABLE IX

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Continental</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Island</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>14</td>
<td>31</td>
<td>30</td>
<td>15</td>
<td>115</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>16</td>
<td>33</td>
<td>35</td>
<td>17</td>
<td>126</td>
</tr>
</tbody>
</table>

Finally, the impact of introduced diseases is largely unknown but native Hawaiian species, previously free of avian malaria, have been infected via introduced birds.

C. The Recent Prehistory of Bird Extinctions

It is well established that there were mass extinctions of large mammals following the colonization of continents by man the hunter. Vulnerable birds may have been extinguished at the same time, as may have birds dependent on the mammals. Fossils show the Californian condor Gymnogyps californianus once to have been widespread across North America, its restriction to the west presumably consequent on the reduction in numbers of large mammals brought about by man, dating from approximately 11,000 years ago. It now survives only through a captive breeding program, its final historical collapse also being a by-product of human activity—in this case, poisoning by the lead in carcasses abandoned by hunters and by poisons used to kill coyotes Canis latrans.

Colonization of islands mostly occurred later. In the Pacific, people had gone as far as the Solomon Islands by 30,000 years ago but did not reach western Melanesia and Micronesia until 4000 years ago, moving then through Fiji and Samoa, the Marquesas, Hawaii, and Easter Island, with New Zealand only colonized about 1000 A.D. European colonists followed in the sixteenth–eighteenth centuries. Some of these human colonizations eventually failed, perhaps because of their dramatic impact on the islands’ ecosystems. The early colonists hunted the animals they found, generally introduced alien species, and usually cleared forests. Their activities certainly led to loss of bird populations: On Hawaii, 60 endemic species are known only from bones; at just one site in New Zealand, 29 endemics are known only from bones; on Easter Island, bones of about 30 species of seabirds have been found from the early years of human occupation but only 1 is now found there (with about 10 others surviving on offshore islets) and all the indigenous land birds seem to have died out following the total deforestation of the island. D. W. Steadman reached the following conclusions about Polynesian land birds:

(i) The ranges of most living species are much smaller today than they were at the time of first human contact. (ii) Few volant species are naturally endemic to only one or two islands. (iii) Most species have become extinct in the past 3000 years. (iv) Most or all islands supported two or three species within a genus, unlike the situation today. (v) At least four formerly widespread genera now are gone from East Polynesia. (vi) Although modern distributions of Polynesian land birds continue to be analyzed as if they were natural, they do not furnish unbiased data for proposing or testing ecological models. (vii) Although some of the range losses of extant species could be restored with conservation efforts, we are centuries too late to preserve any true likeness to the original Polynesian avifauna.

On the basis of the fossil evidence, Steadman suggests that the 800 Polynesian islands inhabited by man have lost an average of 10 species or populations of birds, including 2 or 3 species of flightless rails endemic to each. Other authors have suggested that, because of the incompleteness of the fossil record, even these data may be underestimates. Whatever the details, the avifauna that survived to be further devastated by Europeans, Americans, and Japanese was a small part of that which the earlier colonists found. The marked deficiency of seabirds in the tropics, although probably partly a result of the low productivity of tropical oceans, may be a legacy of this destruction.

D. Current Threats to Birds

On the basis that they have undergone (or appear likely soon to undergo) rapid declines, that they have small and declining populations, that they leave small and declining ranges, or that they have very small populations, 11% of the species of birds in the world are classed as “threatened.” The percentage for mammals is similar; no other group of animals reaches this level of threat. Forest birds make up 65% of this total, and those of scrub and of wetlands each comprise 9%.

Most of the endangered species are tropical. At least for the New World, this is not just because there are more species in the tropics because the proportion of the avifauna that is threatened is also highest in the tropics (Fig. 19). This appears to be in part because the tropics hold many species with restricted ranges. (Throughout the world, about one-fourth of the threatened species are endangered simply by having very
small populations or ranges. In the New World, it is also because of the intensity of tropical deforestation. Of course, areas that hold relatively few threatened species may do so because those species that were especially vulnerable to the impact of man have already gone extinct; this may explain why there are fewer threatened species north of the equator in the Americas than in the south (Fig. 19), given the longer history of human interference in the north.

Habitat loss, especially the loss of tropical broad-leaved forest, is a threat to about half of the world's endangered birds. It is caused not only by the expansion of agriculture but also by mining and the construction of dams. In many parts of the world, afforestation with exotic species is a major problem: Being exotics, the introduced species offer homes to only some of the native animals. Furthermore, they are managed in ways that reduce the floristic and structural diversity of the plantations. Drainage of wetlands is a major problem in many regions, although one that can often be relatively easily reversed if there is the will to do so. About 8% of the world's endangered species are threatened by hunting and capture for the pet trade. In Europe, more than 30% of the species that have declined in recent years are thought to be affected by hunting or by related persecution (such as the killing of raptors that are viewed to be competitors with man). Much hunting is now recreational rather than for subsistence, but it is often more destructive than formerly because there are more people, because more of them have the leisure time to engage in hunting, because there is better technology (mouv nets, guns, and motor transport), and because more people can afford the technology. Coastal habitats in sunny places are steadily being destroyed by the development of tourist facilities. Even where beaches are left in a natural state, birds breeding on them suffer from increased disturbance.

In Europe, the greatest threat, affecting more than 40% of declining species, is the intensification of agriculture. Birds of the steppes are threatened by irrigation, whereas those of wet meadows are threatened by drainage. Fewer resources are available for wildlife when hay meadows are converted to silage and when stocking rates on pastures are increased. Insecticides and herbicides remove much potential bird food from crops. Plant breeding, the use of pre-emergence herbicides, and more effective machinery may allow autumn cultivation in areas where fields used to be left in stubble over winter, such that granivores could feed on grain spilled at harvest and on the seeds of weeds. Harvesting is more efficient. Fundamentally, more of the primary productivity is being harvested, so less is available for wildlife. Most of these changes depend on technological developments but their adoption has been promoted by systems of agricultural support, such as the Common Agricultural Policy of the European Union (EU), that concentrate on production rather than also taking account of social and conservation needs. Expansion of the EU to include countries of Eastern Europe, where agriculture is still relatively less intensive, is a major threat to birds and other wildlife.

Further agricultural changes may flow not only from technology and policy but also from global climate change. This will also affect natural habitats. Although vegetation and wildlife have adapted to the natural climate changes of the past 10,000 years, often apparently within decades, there are concerns that the changes now occurring will be substantially more disruptive because of their speed.

E. Reducing the Threats

Species have been pulled back from the brink. Use of DDT reduced the Mauritius kestrel to 4 birds in 1974 but 20 years later, through a program of captive breeding and reintroduction, there were more than 200. There have been several similar cases, but they have required very intensive care. More extensive measures are needed if all threatened birds are to be conserved. There are two broad approaches to such conserva-
tion. One is simply to control all of the threats listed previously, which would have a widespread impact. Some actions would be neither politically very difficult nor economically very costly, such as banning deliberate introductions, controlling accidental introductions, and eliminating introduced species that have become established. Most, however, would entail a direct conflict between the benefits to wildlife and the material wealth of burgeoning human populations.

The other remedy is to secure individual sites as biodiversity reserves. This would also be expensive because, although 20% of bird species occupy just 2% of the land surface of Earth, it would take a much larger area to include the other 80%. (Furthermore, because hot spots of diversity for different taxa do not closely coincide, even greater areas would be needed if more than birds were to be considered.) In any case, diversity hot spots are peculiar places; conserving them might maintain many species but it would not conserve the natural diversity of ecosystems. There is also the practical difficulty of ensuring that the correct locations are chosen. For example, after decades of trying to build up numbers of the red kite Milvus milvus in the Welsh fastnesses to which the British population had become restricted by nineteenth century persecution, conservationists concluded that it was probably in Wales not because the most suitable habitat was there but just because it had not been shot out there; populations subsequently introduced to eastern England and Scotland have flourished, indicating that the Welsh birds were probably living in relatively poor conditions.

F. Other Reductions in Local Diversity

Extinction is a particularly obvious way in which biodiversity is diminished but is merely the tip of an iceberg. For every species extinguished by man, there are many others whose numbers and ranges are substantially reduced. (There are also a few whose numbers have been much increased by human activity, such as the house sparrow Passer domesticus and starling S. vulgaris.) Changes in numbers are less easy to assess than extinctions and little is known about them in many parts of the world. There are good data, however, for much of North America, where it is clear that recent decades have seen major declines of many Neotropical migrants. It is not clear to what extent this is the result of habitat loss in the wintering areas or of fragmentation of woodlands on the breeding grounds. In Europe, changes in agriculture have been closely followed by declines in the populations of farmland bird specialists (Fig. 20). Studies of individual species are beginning to reveal the causes of these losses, which are different in detail for different species but all fundamentally related to agricultural intensification.

It is specious to argue that such changes are but a reflection of the natural flux of populations. Although systematic counts are not available for most species, except for recent years, it is clear from the extensive historical literature that British farmland birds have probably never before suffered losses on the scale of those recorded in the past few decades, even during the periods of enclosure and other intensive habitat
modification near the end of the eighteenth century and the development of high agriculture in the nineteenth century. The argument that the species characteristic of farmland are much more common than they were before much of western Europe was cleared of its forests is correct but incomplete. Forest clearance led to open-country birds replacing woodland birds, but the farmland birds now being lost are not being replaced. As a result, the total biomass of nonmarine birds breeding in Great Britain declined between 1968 and 1988 by 10% (39% if the pheasant Phasianus colchicus, which is widely and increasingly stocked for shooting, is omitted).

G. Losses of Geographical Diversity

Part of the diversity of life on Earth is the difference between places in their animal and plant communities. Introductions not only endanger biodiversity at the local scale (see Sections VI,B and VI,C) but also, however harmless they are to the indigenous fauna and flora, lessen biodiversity by reducing the difference between places. More than 130 species of birds have been naturalized outside their native ranges, i.e., they have been introduced and have established breeding populations. As a result, the range of some species has increased dramatically. The pheasant Phasianus colchicus, native in only a band stretching from the Caucasus to Japan, now occurs in most of Europe, temperate North America, and New Zealand, with scattered populations in Australia and various islands.

Sport is a major reason for introductions (even in regions with plenty of native game birds) and one-fifth of established introductions are of pheasants (Phasianidae). Some species have been introduced for food (such as the feral pigeon Columba livia), many to remind settlers of “home” (such as European songbirds across the globe), and some because they are beautiful (e.g., peafowl Pavo cristatus). Some have been introduced as agents of biological control, such as the house sparrows Corvus splendens, as house sparrows and house crows Corvus corone, released if markets collapse: This is the origin of the Australian population of the ostrich Struthio camelus.

Species that are closely associated with man, such as house sparrows and house crows Corvus splendens, have reached some places by riding on ships. Introductions of birds have been less damaging than those of mammals and fishes. Nonetheless, each one of them represents a direct reduction in the diversity of life throughout the world.

See Also the Following Articles

AMPHIBIANS, BIODIVERSITY OF • ENDANGERED BIRDS • FISH, BIODIVERSITY OF • MAMMALS, BIODIVERSITY OF • MIGRATION • REPTILES, BIODIVERSITY OF

Bibliography

1. Geography
2. Climate
3. Soils
4. Fire
5. Vegetation
6. Animals
7. Succession
8. Conservation

GLOSSARY

boreal A term derived from the Greek word for north reflecting the fact that this vegetation type, the boreal forest, occurs only in the Northern Hemisphere.

permafrost That part of the soil profile that is permanently frozen and which forms a barrier to water drainage often resulting in a wet surface condition.

podzol A soil type highly characteristic of boreal forests and developed as a consequence of podzolization.

podzolization The process of acid leaching whereby clay, organic particles, and mineral ions (primarily iron and aluminum) are carried downwards and deposited in the B soil horizon, leaving an impoverished and leached A horizon. This occurs as a consequence of low temperatures and precipitation in excess of the needs of evapotranspiration.

taiga Mostly used as a synonym for boreal forest, but more precisely it is a Russian word applied to Eurasian conifer forests described as damp and almost impenetrable. It is also defined as a coniferous forest with no admixture of nonconiferous tree species except Betula and Populus.

THE BOREAL FOREST, or Taiga, is a broad northern circumpolar belt of predominantly coniferous forests. It has many lakes and vast areas are covered by forested wetlands and bogs; bogs and meadows of varying sizes are found throughout the boreal zone. It is the world's largest vegetation type, occupying approximately 14 million km², or 8% of the world's continental ecosystems. It accounts for approximately 12% of the world's biomass. This article will describe the distribution of the boreal forest and the major factors influencing this distribution, such as climate, soils, and fire. The apparent uniformity of the forests throughout their range masks a real diversity that is evident at regional and local scales. This diversity, particularly of plants but also of animals, will be described.

1. GEOGRAPHY

The boreal forests are confined to the Northern Hemisphere. In North America the forest is a continuous vegetation belt stretching across the continent and spanning more than 10° latitude. Its northern limit is defined by transition to the treeless tundra; this limit...
extends from about 68°N in Alaska to 58°N on the west coast of Hudson Bay, and it reaches the Labrador coast at about 58°N (Larsen, 1980). In Fennoscandia it ranges from 56°N to 69°N (Esseen et al., 1997), with the northern limit fringing the northern coast of Norway. Across Russia, except for a small border, the northern boundary more or less follows the Arctic coast. In parts of Siberia the boundary is up to 500 km inland (Larsen, 1980) (Fig. 1). The southern limits of the boreal forest are more difficult to define because the boundaries are rarely sharp. In more oceanic areas the forest is bounded to the south by broad-leaved deciduous forest, and in continental areas transition is to parkland, dry grasslands, and semideserts. For example, in western Canada the southern transition is to subalpine forest, in central Canada it is to prairie grasslands, and in easterly regions it is to mixed deciduous forest.

II. CLIMATE

There are strong relationships between the climate and the soils of the boreal regions. The characteristic nature of the climate to a large extent dictates the nature of the soils, including the permafrost, and these ultimately determine the plants and animals that live there.

FIGURE 1 The circumpolar range of the boreal forest. About two-thirds of the area is in Eurasia. The sector in eastern Canada lies farthest from the North Pole (from Hare and Ritchie, 1972. Reprinted by kind permission of the author and the American Geographical Society).
A. General Limits

The climate of the boreal forest has been extensively documented by, for example, Hare and Ritchie (1972), Larsen (1980), Elliott-Fisk (1988), and Bonan and Stuart (1989). The boreal forest grows where winters are too long and summers too short to support temperate forests. This typically occurs when the growing season is less than 6 months and the frost-free period less than 4 months (or fewer than 4 months with temperatures higher than 10°C). The boundary between boreal forest and tundra corresponds approximately with (i) the line south of which the temperature is higher than 6°C for 4.5 months (Hare and Ritchie, 1972), (ii) the position of the July 13°C isotherm with marked departures in regions with montane or oceanic influences (Larsen, 1980), or (iii) where there are less than 30 days with a daily mean temperature higher than 10°C and where the cold season lasts 8 months (Walter, 1973).

B. Temperature

In more general terms, Elliott-Fisk (1988) describes the climate as cool, humid microthermal, with very cold winters of 7–9 months allowing persistence of snow cover during all but the brief (3 or 4 months), relatively cool summer season. For more than 6 months of the year, the mean temperature is below 0°C and net radiation is negative. Maximum summer temperatures generally reach the low 20s and winter minimums in the –50s. However, yearly variation can be extreme such as in Verkhoyansk (30°C in summer and –70°C in winter), but these extremes are moderated near the coast, e.g., in Umeå, Sweden (Fig. 2).

C. Precipitation

Generally, the boreal zone is characterized by having a high proportion of the annual precipitation falling as snow. Throughout most of the boreal forest annual precipitation is low, varying between 250 and 1000 mm (Fig. 2). There are extremes, however, with Fort Yukon in Alaska and Verkhoyansk in Siberia each recording 1800 mm annually. Some areas of western Norway with up to 2000 mm of precipitation annually have been classified as boreal rain forest. Because of low temperatures and short growing seasons, evaporation rates are low and drought is uncommon. When drought does occur, forest fires can ravage vast areas of boreal forest.

III. SOILS

Boreal forest soils are typically low in fertility and acidic, with a thin A horizon. The most characteristic soils are podzols. However, podzols occur in a wide range of climates, not only in boreal regions, and not all boreal regions are underlain by podzols. Climate, vegetation type, chemical composition of the substrate material, and topography are the major environmental influences that produce the typical boreal podzol. The combination of low temperatures and low pH impede decomposition processes and slow the rate of soil development. The soil surface may be covered by a mat of spruce needles up to 3–7 kg/m². This mat of acid, partly decomposed plant material is the mor litter layer. The slow but gradual decomposition of this layer continually releases a supply of organic acids that contribute to the leaching of organic particles and mineral ions (primarily iron and aluminum) from the surface soils.
and to the weathering of the parent material. In addition, soluble materials such as sodium, potassium, and calcium are washed out of the soil by water movement. As a consequence, the surface soils are relatively infertile and high in silica. In many cases in which the process of podzolization is prominent, the upper soil layers are gray or whitish in color. Beneath the leached layer is a zone in which materials leached downward by water accumulate, chiefly iron–humus complexes. These deposits may be cemented into a hardpan, sometimes thick and strong enough to prevent root penetration to the lower soils. In the more extreme boreal forest climates the subsoil is permanently frozen (the permafrost). The combination of nutrients being largely tied up in the litter layer, an infertile A horizon, hardpans and permafrost results in most boreal forest trees having a shallow root network.

IV. FIRE

Individual trees or local stands may be killed by windstorms, landslides, snow avalanches, erosion, ice storms, flooding, or insect attack, but the most important natural disturbance in the boreal forest is fire. Also, although the boreal forest is relatively impoverished floristically, much of the floristic diversity and large-scale vegetation mosaics are directly attributable to repeated burning. Longer periods of unseasonably drier and warmer weather lead to a greater probability of fire, and it is the few major fires during extreme fire years that account for the majority of forests burned; this inevitably has enormous consequences for the wildlife that inhabit the forests, particularly those unable to take refuge in water bodies. In northwestern Canada and Alaska, between 60 and 80% of all fires are less than 5 ha in area, but in extreme fire years individual fires can burn up to 200,000 ha. Because fires often burn all or most of the forest floor, they influence organic matter accumulation, soil temperature, and soil moisture, and through these they impact major ecosystem processes such as nutrient cycling, energy flow, and productivity. The natural cycle of fire frequency in boreal forests of North America ranges from an average of 50–200 years to up to 500 years in wetter parts of eastern Canada. Fire frequencies in northern Sweden average from 110 to 155 years but may be as high as 270 years. Fire frequency varies considerably, both within and between regions, resulting in a mosaic landscape with different forest patches in various stages of succession. Factors contributing to the probability of a fire include tree species composition, stand structure, soil conditions, the amount and moisture content of the fuel, exposure, topography, and climate; there is much disagreement regarding whether the probability of a fire increases with stand age. Pine forests burn, on average, twice as frequently as spruce forests. Local fire chronologies indicate that fire frequency is highest in the interior, more continental areas due to a higher frequency of thunderstorms and lightning strikes; most natural fires are caused by lightning strikes. In the boreal forest of Fennoscandia there is apparently a north to south gradient with a generally lower fire frequency toward the north.

V. VEGETATION

The boreal forest is the most continuous and extensive forest in the world. The North American and Eurasian forests are remarkably uniform in their appearance throughout their range, both in their physiognomic structure and in species composition. Typically, there is a simple canopy layer (15–20 m high) in which numerical dominance is maintained nearly everywhere by coniferous tree species belonging to four genera: spruce (Picea), pine (Pinus), fir (Abies), and larch (Larix) (Fig. 3). Species of juniper (Juniperus), cedar (Thuja), and hemlock (Tsuga) also occur. Next is a shrub layer (typically 1 or 2 m tall) supporting mainly broad-leaved deciduous species that are also frequently present as successional components of the forest; deciduous species rarely achieve dominance, except in some postfire successions and in the mountain birch forest in Fennoscandia. Generally, the deciduous species belong to the genera willow (Salix), birch (Betula), poplar (Populus), and alder (Alnus). The herb layer typically is poorly developed but it is enmeshed in a well-developed ground layer of mosses, liverworts, and lichens. Despite the uniformity of appearance, the tree species are unique to a particular continent, i.e., there are no circumboreal trees (Table 1). The species making up the shrub layer and ground layer may be very wide ranging, and many of the mosses and lichens are circumboreal. Bryophytes and lichens are typically a more common component of the forest floor vegetation than vascular plants. Bryophytes usually dominate on mesic and moist sites and their diversity in boreal forest is higher than in most temperate or tropical forests. Lichens also contribute significantly to plant diversity in boreal forests, especially on drier and northern sites, and particularly in the ground layer of pine heaths and on rocky ground.

Because there are few species of trees, the boreal forest gives an impression of monotony, but this is
misleading even though they are floristically impoverished compared to most other vegetation formations of the earth. Nevertheless, these forests are a complex mosaic of different plant communities, and they caused Cumming et al. (1996) to title their paper "Boreal Mixedwood Forests May Have No 'Representative' Areas." Some of the tree species are widely distributed (e.g., *Picea glauca* and *P. mariana*), but many other boreal species have more limited ranges, resulting in local and regional changes to forest composition. Throughout the North American boreal forest, a vegetation group dominated by *P. glauca* (sometimes codominated by *Abies balsamea*, *A. bijolia*, *Betula papyrifera*, or *Populus tremuloides*) is typically the climax community in relatively dry areas. A second group dominated by *P. mariana* occurs where soils are wetter, such as in bogs and muskegs. *Picea mariana* is not strictly characteristic of wet areas because it also flourishes after fire. In the European forests the Scots pine (*Pinus sylvestris*) is frequently dominant on drier soils and the Norway spruce (*Picea abies*) occurs on moister sites.

Although there are no abrupt transitions in the vegetation, changes in the dominant tree species, in the subdominant shrubs, and in the herbaceous layers can be detected. Various attempts have been made to survey and describe latitudinal and longitudinal gradients, based primarily on climatic distinctions, and to describe the associated vegetation assemblages. For example,
Larsen (1980) described 7 regions in the boreal forest of North America, and Rowe (1972) described 35 regions. However, rather than describe each of these, I focus on some of the recent classifications.

A. A Latitudinal Classification

Hare and Ritchie (1972) recognized three latitudinal zonal divisions of the boreal forest in North America: a southern closed forest, a central lichen woodland, and a northern forest—tundra ecotone.

1. Closed Forests

The closed forest dominates vast areas of the southern boreal forest, and it occurs on a wide range of soil types and topographies. These forests have a closed crown with a moist, deeply shaded floor. The spruce—feathermoss community is characteristic of this zone, with either white or black spruce dominant. The P. mariana—feathermoss forests have a fairly uniform, moderately dense tree stratum, with an almost continuous ground cover of bryophytes. In contrast, the P. glauca forests are more irregular and open, having an understory of broad-leaved shrubs and herbs and a patchy bryophyte distribution. The most common moss in the black spruce forests is Pleurozium schreberi, whereas Hylocomium splendens is more characteristic of the white spruce and mixed woodlands.

2. Lichen Woodland

These forests are more open with a discontinuous shrub layer and abundant lichens. The transition from the southern closed forests to these lichen woodlands is usually smooth or mosaic-like but can sometimes be quite abrupt as in northern Manitoba and Saskatchewan. In some cases, a Western Stereocaulon pauchale and an Eastern Cladonia stellaris-dominated woodland have been identified. Picea mariana and P. glauca are the dominant trees in these forests, with P. glauca declining to the north. Because these forests are more open, light penetration to ground level is much greater than in closed forests. Although these trees may be as tall (but less dense) as those in closed forests, the...
TABLE II

<table>
<thead>
<tr>
<th>Major Dominant, Codominant, or Abundant Species Identified in the Three Major Longitudinal Sections in Peinado et al.’s (1998) and La Roï’s (1967) Classifications of North American Boreal Forests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
</tr>
<tr>
<td>Peinado et al. (1998)</td>
</tr>
<tr>
<td>The western section (northeastern British Columbia, southern Yukon, and Alaska)</td>
</tr>
<tr>
<td>Abies lasiocarpa</td>
</tr>
<tr>
<td>Pinus contorta var. latifolia</td>
</tr>
<tr>
<td>Picea engelmannii</td>
</tr>
<tr>
<td>The central section (eastern British Columbia, Alberta, Saskatchewan, and Manitoba)</td>
</tr>
<tr>
<td>Abies lasiocarpa</td>
</tr>
<tr>
<td>Lonicera/Rubus pubescens/Lathyrus oenotheraeus</td>
</tr>
<tr>
<td>Abies bifoila</td>
</tr>
<tr>
<td>Lonicera/Vaccinium vitis-idaea-Geocaulon</td>
</tr>
<tr>
<td>Abies balsamea</td>
</tr>
<tr>
<td>The eastern section (all forests east of Lake Winnipeg, Manitoba)</td>
</tr>
<tr>
<td>Abies balsamea</td>
</tr>
<tr>
<td>Kalmia-Picea mariana</td>
</tr>
<tr>
<td>Pinus balsamiana</td>
</tr>
<tr>
<td>Abies Balsamea</td>
</tr>
</tbody>
</table>

additional light promotes branching to ground level. The additional light often promotes a discontinuous layer of heath plants such as crowberry (Empetrum nigrum) and bilberry (Vaccinium myrtillus).

3. Forest–Tundra Ecotone

Here, there are scattered and isolated trees, often deformed or prostrate, in a tundra landscape. Although this type of forest is located north of the limit of continuous forest, it is still considered boreal forest. The ecotone is more than 300 km wide in Quebec, up to 225 km wide in Central Canada, and it narrows at both its eastern and western ends.

B. A Longitudinal Classification

There is a remarkable similarity between the vegetation of the boreal forests in eastern North America and that in eastern Asia, with many identical or closely allied genera and sometimes species. These forests contain many different species, whereas those of the Euro-Siberian region contain few. Of all the tree species in the North American boreal forest, only P. glauca extends from the Bering Straits across Alaska and Canada to Newfoundland. Picea mariana, usually found only on poor or wet soils, is found at the timberline toward the Arctic, and Larix laricina is found in the continental regions. Only two species, P. sylvestris and P. abies, are of any real importance in the boreal zone of Europe. Only in eastern regions of Europe is P. abies replaced by the closely related Picea obovata, whereas additional species are being added to the forest (Abies sibirica, Larix sibirica, and Pinus sibirica; Table I). Moving east, spruce gradually declines from the forest until it is entirely absent in eastern Siberia.

Peinado et al. (1998) analyzed the vegetation of the North American boreal forests. They identified three major sections (Table II) and classified them as eight major groups (Table III). The three sections coincide well with La Roï’s (1967) classification for the same regions.

Qian et al. (1998) examined longitudinal patterns of plant diversity in the North American boreal forests, focusing specifically on the southern closed forests. The central section has a higher species and genera diversity than the western and eastern sections (Table IV). White spruce forests are always more diverse than black spruce forests; this is a reflection of the diversity of herbaceous plants and bryophytes and not the diversity of woody plants (Fig. 4). The diversity of white spruce forests is rather similar between western and eastern sections, but the diversity of black spruce forests is much higher in the west than in the east (Table IV). The diversity of bryophytes is remarkably consistent across the continent, but again there are more bryophytes in white spruce than in black spruce forests (Fig. 4).

VI. ANIMALS

The boreal forest is home to many animals. It is the winter home of the migratory caribou and reindeer and the permanent home of many others. The wolf and lynx are the major predators of the boreal forest. Some of
the best examples of population cycles in animals are described from the boreal forest regions; for example, lynx (Lynx canadensis), snowshoe hares (Lepus americanus), arctic ground squirrels (Spermophylus parryi), red squirrel (Tamiasciurus hudsonius), and boreal red-backed vole (Clethrionomys rutilus) in northern Canada and microtine rodents, owls, capercaillie (Tetrao urogallus), black grouse (Tetrao tetrix), mountain hare (Lepus timidus), and red fox (Vulpes vulpes) in Eurasia. The causal relationships of these cycles have not been fully explained but many of the North American examples are synchronized with the snowshoe hare cycle, and Eurasian examples are synchronized with microtine rodent cycles.

In North America, other inhabitants of the boreal forest include moose, black bear, grizzly bear, deer, wolverine, coyote, marten, beaver, porcupine, sable, voles of the genus Microtus, chipmunks, shrews, and

---

### TABLE III

Major Dominant, Codominant, or Abundant Species Identified in the Eight Groups in Peimado et al.'s (1998) Classification of North American Boreal Forests

<table>
<thead>
<tr>
<th>Group</th>
<th>Western</th>
<th>Central</th>
<th>Eastern</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td><em>Pinus glauca</em> and/or <em>Pinus contorta</em> var. <em>latifolia</em></td>
<td><em>Abies balsamea</em></td>
<td><em>Betula papyrifera</em></td>
</tr>
<tr>
<td></td>
<td><em>Populus tremuloides</em></td>
<td><em>Abies lasiocarpa</em></td>
<td><em>Betula alleghaniensis</em></td>
</tr>
<tr>
<td>II</td>
<td><em>Pinus mariana</em></td>
<td><em>Abies balsamea</em></td>
<td><em>Picea rubens</em></td>
</tr>
<tr>
<td>III</td>
<td><em>Alnus rugosa</em></td>
<td><em>Abies balsamea</em></td>
<td><em>Picea mariana</em></td>
</tr>
<tr>
<td>IV</td>
<td><em>Pinus banksiana</em></td>
<td><em>Abies balsamea</em></td>
<td><em>Picea glauca</em></td>
</tr>
</tbody>
</table>

---

### TABLE IV

Mean Number of Species and Genera in White Spruce and Black Spruce-Dominated Ecosystems in Three Geographic Sectors of the Boreal Forest of North America*

<table>
<thead>
<tr>
<th>White spruce ecosystems</th>
<th>Western North America</th>
<th>Central North America</th>
<th>Eastern North America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Genera</td>
<td>Species</td>
<td>Genera</td>
</tr>
<tr>
<td>Woody plants</td>
<td>13</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Herbaceous plants</td>
<td>28</td>
<td>25</td>
<td>33</td>
</tr>
<tr>
<td>Bryophytes</td>
<td>23</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>Vascular plants</td>
<td>41</td>
<td>35</td>
<td>21</td>
</tr>
<tr>
<td>All plants</td>
<td>65</td>
<td>54</td>
<td>77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Black spruce ecosystems</th>
<th>Species</th>
<th>Genera</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of plots</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Woody plants</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Herbaceous plants</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>Bryophytes</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>Vascular plants</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>All plants</td>
<td>53</td>
<td>42</td>
</tr>
</tbody>
</table>

*The boreal forest can be latitudinally divided into three subzones: a wide zone of closed forest in the south, a narrow belt of forest-tundra ecotone in the north, and a lichen woodland in the middle. These data cover almost the entire longitudinal range but focus on the southern closed forest (modified and adapted from Qian et al., 1998, by kind permission of the author and Kluwer Academic Publishers).
BOREAL FOREST ECOSYSTEMS

![Diagram]

**Figure 4** Triangular ordination showing sample plots in the white spruce and black spruce ecosystems of the North American boreal forest. Each point represents a single sample plot of 9 ha (300 × 300 m). Thirty-four of the plots were sampled in Picea glauca forest and 26 in Picea mariana. CNA, central North America; ENA, eastern North America; WNA, western North America (from Qam et al., 1998. Reprinted by kind permission of the author and Khuever Academic Publishers).

bats. Typical animals of the boreal forest vary slightly more from location to location than do the plants. The moose which browse on willow, birch, alder, and water plants, and the beaver which feeds on aspen, are widespread. Many birds also inhabit the boreal forest; for example, great horned owl, goshawk, spruce grouse, ruffed grouse, nuthatchers, junco, and warblers. Brown bears inhabit the boreal forest in Eurasia.

In the boreal zone of Eurasia, the diversity of mammalian herbivores is highest in the interior of the continent and declines to the east. Across Eurasia, species richness of mammalian herbivores is positively correlated to warm climate, the number of hardwood species, and the area of the boreal forest. Across North America, species richness of mammalian herbivores increases as the length of the growing season and the number of coniferous tree species increase (Fig. 5). Given this information, it appears that indirect measures of primary productivity as well as the number of tree species can accurately predict species richness of mammalian herbivores. Bird diversity decreases from west to east across both the North American boreal forests and the Eurasian boreal forests. In Fennoscandia the diversity of forest birds decreases northward; in Finland, this occurs only in pine forests and not in spruce.

It seems that the boreal forests of Canada, and possibly Russia, differ from those in northern Fennoscandia in that small herbivore biomasses reach much higher levels and are dominated by species of hare rather than voles. In addition, the densities of many fewer species in the boreal forests of Canada are correlated with the dominant herbivore relative to the situation in Fennoscandia. In Fennoscandia, where voles fluctuate, their changes in density are strongly correlated with that of hares, and those of two predominant alternative prey, red squirrels and voles, were not.

Tree death and decaying wood provide a variety of habitats for an enormous number of invertebrates. For example, in Sweden approximately 1000 species of beetle are dependent on dead trees. The most diverse fauna on snags is found during the first 2 years after the tree has died. Spruce logs have a more diverse invertebrate fauna than pine, but many invertebrates can inhabit both. Four typical stages in the succession of invertebrates on spruce logs in boreal forests have been described. Initial colonization is by bark beetles and other primary cambial eaters along with their associated parasitoids, predators, and detritivores. Subsequent stages have been described in detail by Esseen et al. (1997).

VII. SUCCESSION

Many studies of the boreal forest have shown a remarkably close fit between the plant community and the environment, especially with reference to moisture conditions and fertility level. This relationship is not static and communities are always in a process of change. The sequence of primary successional change is typically described as beginning in fens and then progressing as follows:

Fen → swamp → bog → muskeg → invading trees → forest

Fens develop under alkaline conditions on a peat substrate usually with standing water. Bogs are specialized communities of shrubs and herbs growing on a wet, acidic, peat substrate. The sequence becomes drier and more mineral poor as muskeg develops, and this is followed by tree invasion.
The usual sequence of secondary succession is described as

Disturbance → annual plants → perennial herbs and shrubs → larger shrubs and trees → forests

However, these descriptive sequences mask a dynamic system under the influence of frequent disturbances, and the boreal forest is better considered as a disturbance forest. This means that the concept of the vegetation climax is probably not applicable because most of the tree species in these forests are incapable of self-perpetuation in climax state. They are instead, adapted to fire, and the boreal forest is usually maintained in a pre-climax state by frequent fires. Fire is the primary disturbance factor interrupting the successional process, and the frequency and intensity of fires will determine the nature and stage of succession. Local topographic, soil, moisture, and microclimate conditions will determine the local fire regime, and if the local regional environment is patchy so also will be the frequency of fires. The consequence will be a vegetation mosaic, with each patch being at a different stage of succession and representing a different degree of recovery since the last fire. In this sense, the boreal forest climax is not a stand dominated by white or black spruce, but is a mosaic in which all stages of a successional or regeneration cycle are represented, i.e., the climax is a mosaic of successional patches. It has been argued that if fires could be excluded from a lichen woodland for at least 200 years, then it is likely, that a closed-forest spruce–feathermoss would develop; however, fires usually occur more frequently and so the lichen woodland is perpetuated.

Nevertheless, immediately following a fire some species are characteristic pioneer species. Among herbaceous plants, *Epilobium angustifolium* and *E. latifolia* (fireweeds) are the most obvious. Density and cover of *Salix, Betula, and Alnus* are typically high in early pioneer stages and greatly decrease in later stages when they are intolerant of the shady conditions imposed by the establishing *Pinus, Abies, Picea, and Populus*. Many questions remain concerning the nature of succession in the boreal forest. There are many regional differences, and local site conditions exert considerable influence of the successional process—both on its path and on its “final” form.

**VIII. CONSERVATION**

In both Canada and Russia, vast areas of boreal forest are being cut annually, and the situation is worse in Fennoscandia. Although wood cutting is the biggest danger to biodiversity, there are additional threats from mining, pollution, road building and dam construction. Clear-cutting vast areas of forest has profoundly altered the landscape structure in northern Sweden and in central Canada, resulting in habitat loss, habitat alteration, and fragmentation. Clearly, this will have mostly negative influences on both animal and plant biodiversity, even though some species will benefit in the short-term, i.e., weedy species such as *Epilobium angustifolium, Deschampsia flexuosa,* and *Calamagrostis purpurea* that are adapted to fire and other disturbances. Few data are available on the long-term effects of clear-cutting in the boreal forest, but studies from other forests suggest that herb communities do not recover in logging
cycles of 40–150 years. Rotational cutting and clear-cutting will inevitably influence, or eliminate, natural fire regimes and will lead to changes in tree species composition and forest structure, a general reduction in stage age, and reduced input from coarse woody debris (Esseen et al., 1997), all of which will have profound impacts on the natural biodiversity of the boreal forest.

Biodiversity must be preserved at all scales from the genetic variation within a population to heterogeneity occurring at the landscape level. There are two major approaches to preserving biodiversity in these areas: (i) sustainable management; and (ii) reserves of natural areas. The first approach is more applicable to areas that are already subject to intense management such as Fennoscandia, whereas the second approach can be applied where vast, relatively untouched areas still exist such as in parts of Canada and Russia.

A. Sustainable Management

To maintain biodiversity, we must also preserve or simulate the processes, mostly natural disturbances, that produced the heterogeneity in the first place. Microscale heterogeneity may be enhanced by gap disturbance and by coarse woody debris. Larger scale heterogeneity may be enhanced by fire or insect outbreak. Angelstam (1998) proposed a conceptual model as a guide to the maintenance and restoration of ecologically sustainable boreal forests. The model is based on the hypothesis that self-sustained forest ecosystems can be created, or recreated, and their biodiversity developed if forest management can simulate the composition and structure of boreal forest landscapes. This is done by introducing and maintaining disturbances that lead to naturally dynamic spatial and temporal patterns of forest regeneration. The model has already been implemented in planning systems of many large Swedish forest enterprises.

B. Reserves of Natural Areas

Reserves alone are not sufficient to conserve biodiversity in forests, but any such initiative for maintaining biodiversity in the boreal forest should include at least three components:

1. Large areas representative of southern closed-canopy forest, lichen woodland, the forest-tundra ecotone and each of the longitudinal elements within these must be protected.

2. In selecting these areas, attention must be given to the animal inhabitants, their abundance or rarity, and their migratory behavior, if any.

3. The decision-making process should consider the "floating reserve" strategy (Cumming et al., 1996) in which portions of a protected area could be periodically replaced in response to aging of components, unexpected large-scale disturbance, or refinements in conservation objectives.

Although much of the world's boreal forest still remains intact with little impact from man, the impact from man has been immense in many areas. In the new millennium, there will continue to be an increasing demand on the world's resources, including the boreal forests. Indeed, man's impact has already been substantial in Fennoscandia (less than 5% of the boreal forest remains in a natural or seminatural state) and is increasing in the North American forests. It is tempting to argue that these forests are the world's largest vegetation type, occupying vast areas (approximately 14 million km²) and account for approximately 12% of the world's biomass, so the incursion of man is likely to have little impact. As described in this article, although the boreal forest is less diverse than most of the other world's vegetation types, it is nevertheless a complex mosaic of patches with changing species composition both longitudinally and latitudinally, with no representative areas, and home to many species of mammals, birds, and other animals.

See Also the Following Articles

ARCTIC ECOSYSTEMS • FIRES, ECOLOGICAL EFFECTS OF • FOREST ECOLOGY • LATITUDE, COMMON TRENDS WITHIN • NORTH AMERICA, PATTERNS OF BIODIVERSITY IN • TIMBER INDUSTRY

Bibliography


BREEDING OF ANIMALS

David R. Notter
Virginia Polytechnic Institute and State University

I. Overview of Livestock Genetic Diversity
II. Conservation Planning: Coordinating the Use and Preservation of Livestock Genetic Resources
III. Management Strategies for Conservation of Livestock Genetic Resources
IV. Potential Impacts of New Technologies
V. Conclusions

GLOSSARY

breed “A homogenous, subspecific group of domestic livestock with definable external characteristics that enable it to be separated by visual appraisal from other similarly defined groups within the same species, or for which geographical separation from phenotypically similar groups has led to general acceptance of its separate identity” (Food and Agriculture Organization, 1995).

breeding line Any distinctive livestock population. The term may be synonymous with “breed” but is also often applied to somewhat distinct subpopulations within a breed or to breeding populations, either purebred or crossbred, that are of fairly recent origin. Breeding lines may exist within a breed, and newly created lines may become recognized breeds.

crossbreeding/crossbred Crossbreeding is the mating of animals of different breeds to produce commercial market animals or breeding animals. The resulting crossbred animals often exhibit improved fitness and performance compared to their purebred parents.

cryopreservation Long-term storage of gametes or embryos in liquid nitrogen.

ex situ preservation Maintenance of a breed outside the environment and agricultural production system(s) in which it was developed and used. Ex situ preservation often involves cryopreserved gametes and/or embryos but may also involve live animals.

germplasm Sources of hereditary material. In animal breeding the term is commonly applied to breeding animals and to fresh or frozen sperm cells, ova, and embryos.

in situ preservation Maintenance of a breed in the environment and agricultural production system(s) in which it was developed and used.

purebred An animal that is a member of a recognized livestock breed. Breed membership may be determined by pedigree records, geographical location, or knowledge of the breeding structure of the herd or flock. Animals that possess the full range of characteristics commonly associated with a specific breed are sometimes also designated as purebreds, but the designation must be recognized as subjective in these cases.

DOMESTIC ANIMALS make a critical contribution to human well-being as a source of food, fiber, power, and...
defines a livestock breed as a homogenous, subspecific group of domestic livestock with definable external characteristics that enable it to be separated by visual appraisal from other similarly defined groups within the same species, or ... for which geographical separation from phenotypically similar groups has led to general acceptance of its separate identity.

The concept of a livestock breed arose in nineteenth century England, where influential livestock breeders, including the noted Robert Bakewell, began to improve the native livestock of the region by controlled mating and selection of presumed superior individuals. Significantly, their activities included the establishment of "breed books" for most mammalian breeds which enumerated the pedigrees of animals within each breed so that breed membership was confirmed by records of ancestral ties to presumed elite foundation animals. These so-called "purebred" livestock thus had pedigree documentation of breed identity. Similar activities in poultry resulted in establishment of many avian breeds, although individual pedigree recording was not usually practiced for these smaller, more prolific species. This breed concept became widespread in Western Europe and in areas that were settled or influenced by Western Europeans. By the early twentieth century, a central part of livestock breeding in these areas was the identification and use of distinct, ostensibly superior, genetic types and the establishment of breeds to propagate these types.

The Western European definition of a breed had little meaning in most of Asia and Africa, where extensive pedigree recording did not appear to have arisen. However, selective breeding certainly did occur in these areas, and regionally distinct livestock types arose and were propagated. Isolation of these types was generally by distance or by selective matings based on obvious phenotypic traits such as color or horn shape rather than on pedigree information. However, pedigree information was occasionally accorded great value, as in the establishment of the Arabian horse. Today, the term breed has become globally pervasive and is used to describe any identifiable subpopulation possessing relatively distinct characteristics and maintained with a reasonable level of reproductive isolation.

I. OVERVIEW OF LIVESTOCK GENETIC DIVERSITY

A. Breeds of Livestock

Populations of domestic livestock are traditionally subdivided and isolated from other subpopulations by both distance and management. These subpopulations are generally known as "breeds." A single, widely accepted definition of a livestock breed does not exist, but livestock breeds are approximately analogous to the subspecies or races of wild species. Thus, the Food and Agriculture Organization of the United Nations (FAO) defines a livestock breed as:

- a homogenous, subspecific group of domestic livestock with definable external characteristics that enable it to be separated by visual appraisal from other similarly defined groups within the same species, or ... for which geographical separation from phenotypically similar groups has led to general acceptance of its separate identity.
in improving prolificacy can thus rely on this breed as a source of useful genes.

The traditional (i.e., mid-twentieth century) genetic structure of most livestock species was thus characterized by many regionally isolated subpopulations. In addition, several distinct subpopulations could often be identified within each region, maintained in more or less pure form by breeders who favored animals of a specific type. These dual mechanisms for maintaining reproductive isolation fueled development of levels of observed genetic diversity that are generally far in excess of those observed in wild species.

It would be incorrect, however, to view the livestock breeds as fixed, immutable genetic entities. Migration, or the exchange of genetic material, among breeds is not uncommon, and, even among breeds that maintain pedigree records, decisions to ‘open’ the herd book occasionally occur to allow incorporation of desirable animals from outside the breed. Breeds wax and wane in popularity; as some breeds merge, others are absorbed by more popular breeds, and a few simply lose favor and gradually disappear. Breed evolution is thus an accepted part of livestock breeding. This traditional genetic structure of livestock breeds was remarkably similar to that proposed by Sewell Wright as optimal for evolution in natural populations, with relatively high levels of reproductive isolation, modest effective population sizes, diverse selection pressures, and levels of migration that provided for periodic infusions of genes without seriously compromising the genetic integrity of the subpopulations.

Livestock breeds in most cases possess very significant amounts of within-breed genetic diversity. Although breeds can rightly be viewed as mildly inbred lines, they do not approach the levels of genetic uniformity commonly found in plant breeding lines in which self-fertilization and clonal propagation can strongly limit, or in some cases completely eliminate, genetic diversity. Thus, essentially all livestock breeds retain significant evolutionary potential and can undergo significant genetic changes in response to selection, either natural or artificial. Livestock breeds retain the capacity to change, sometimes radically, in response to changes in breeders’ preferences or market demands without losing the fundamental characteristics of the breed.

Finally, within each of the domestic species, a very large number of individuals exist that do not exhibit the full array of phenotypic characteristics necessary to confirm breed identity. For example, more than 90% of U.S. dairy cows would be classified visually as belonging to the Holstein breed, but at most one-third of these would be recorded as purebreds. The proportion of recorded animals is much lower in cattle used for meat production.

Many other animals are obvious crossbreds, showing mixtures of specific breed characteristics that suggest the possibility of predictable breed ancestry. Also, large numbers of animals are often truly nondescript, exhibiting a mosaic of various breed characteristics that point to a more catholic breed composition. The value of the crossbred or nondescript animals as a genetic resource is hotly debated. Populations are often large and well adapted to prevailing environmental conditions. Although adapted local breeds are often viewed as ‘contaminated’ by crossing with less well-adapted, imported breeds, these populations also provide opportunity for creation of new genetic combinations and for selective elimination of undesirable breeds. Thus, these animals represent a potentially useful genetic resource but lack the predictability of the pure breeds for use in commercial livestock production.

C. Current Trends in Livestock Diversity

Perhaps the most significant event in global genetic resource utilization in the late twentieth century has been the emergence of highly productive, widely distributed ‘global’ livestock breeds. These breeds, for the most part, arose in the developed nations of the temperate zones and have had a remarkable impact on livestock production worldwide. These breeds include the Holstein dairy cow, capable of producing an average of more than 30 kg of milk per day for 300 days; the modern broiler chicken, which can reach a market weight of 2 kg in less than 40 days and on less than 4 kg of feed; meat-type pigs with less than 5 mm of back fat at a body weight of 110 kg; and Australian Merino sheep with wool fiber diameters of 17–20 μm.

The development and continued improvement of these global breeds is viewed in many quarters as the capstone of modern animal breeding. Large population sizes, detailed recording of performance, and intensive selection fueled their development. Well-organized breeders’ groups and, recently, multinational corporations are involved in their propagation. The emergence of corporate animal breeding, in particular, allowed merging of genetic improvement, production, processing, and marketing activities and provided re-
sources for aggressive international marketing of germplasm. Within the developed nations, the global breeds have made an important contribution to maintaining low food costs, but many of the traditional livestock breeds in these areas have declined because they cannot compete with the emerging global breeds.

Because of their tremendous production potential, these breeds also caught the attention of government planners and officials in developing nations. Their global dissemination was seen as a magic bullet to improve animal productivity, analogous to the improvements in crop yields associated with the Green Revolution. Widespread importations of these breeds into developing nations had a negative impact on indigenous breeds, sometimes through outright breed replacement but more often through extensive and generally unregulated crossbreeding between imported and indigenous breeds. In rural areas and subsistence production systems, the contributions of these highly productive global breeds were often disappointing. When denied the high levels of feeding, housing, and veterinary care under which they were developed and forced to grapple with unfamiliar levels of disease and climatic stress, the global breeds and their crosses were often unable to forage, survive, and reproduce at acceptable levels and were ultimately less productive than the adapted indigenous types they were intended to replace.

Improved technology for transfer and use of genetic material contributed to the demise of many indigenous breeds. Use of frozen semen became commonplace in cattle in the latter half of the twentieth century and is now widespread for many livestock species. Use of frozen embryos is also widely practiced for cattle, sheep, and goats. These developments allowed efficient global exchange of genetic material and effectively ended the genetic isolation of many breeds, especially in Asia and Africa, where the concept of breed identity was less strong than in Europe and North America. Greater ease of transport of live animals also facilitated establishment of populations of foreign breeds and replacement or contamination of native breeds, especially for the smaller or more fecund species such as poultry and pigs. Thus, modern reproductive technologies today permit foreign breeds to have a much greater and more rapid impact, and in some cases to completely replace native breeds.

Improved technology for transfer and use of genetic material contributed to the demise of many indigenous breeds. Use of frozen semen became commonplace in cattle in the latter half of the twentieth century and is now widespread for many livestock species. Use of frozen embryos is also widely practiced for cattle, sheep, and goats. These developments allowed efficient global exchange of genetic material and effectively ended the genetic isolation of many breeds, especially in Asia and Africa, where the concept of breed identity was less strong than in Europe and North America. Greater ease of transport of live animals also facilitated establishment of populations of foreign breeds and replacement or contamination of native breeds, especially for the smaller or more fecund species such as poultry and pigs. Thus, modern reproductive technologies today permit foreign breeds to have a much greater and more rapid impact, and in some cases to completely replace native breeds.

Urbanization has concentrated animal production in periurban areas to allow convenient access to population centers and led to opportunities for more intensive production. These changes have occurred despite concerns over economic disenfranchisement of farmers in rural areas, air and water pollution and waste disposal in periurban production, and the long-term sustainability of these production systems. Intensive animal production systems in periurban areas have also expanded use of high-production foreign breeds at the expense of adapted native types. The competitive advantage of these intensive periurban systems may decline once the full costs of waste disposal and environmental monitoring are incorporated into production costs, and a return to the more traditional mixed-farming systems and breed may be needed.

Expanded use of grain for animal feed would have been inconceivable to many international observers a decade ago but is now a reality. China has the highest aggregate consumption of meat in the world, and global grain production is projected to keep pace with demands for use in animal feed for at least the next 25 years. The result of these increases in use of grain for animal feed has been a need for rapid improvement in efficiency of grain use which has been difficult to achieve in a timely manner using traditional breeds. However, if trends in use of grain for animal feed slow or reverse in the latter years of the twenty-first century, these traditional breeds may again be needed.

Animal welfare concerns may also influence the use of livestock breeds. In Europe, concerns over issues of animal welfare in intensive production systems have prompted a return to less-intensive conditions and a corresponding need for animals that produce well under those conditions.

II. CONSERVATION PLANNING: COORDINATING THE USE AND PRESERVATION OF LIVESTOCK GENETIC RESOURCES

Animal breeders are often confronted by an apparent dilemma in the design of livestock breeding programs. On the one hand, maximizing livestock productivity relies on the identification and propagation of superior genetic types. Rates of genetic improvement are generally proportional to the intensity of selection. Thus, propagation of small numbers of elite parents yields greater rates of progress than retention of larger numbers of less desirable individuals. Indeed, most of the accomplishments of modern animal breeding have re-
synthesized from improving the accuracy of genetic evaluation of prospective parents and increasing the impact of these selected parents by use of artificial insemination, embryo cloning and transfer, and other advanced reproductive technologies. Successes of modern animal breeding programs have, in some cases, been spectacular, as exemplified by the emerging global breeds of poultry, swine, and dairy cattle. The global proliferation of industrial production systems has relied primarily on these species because of their greater efficiency of use of harvested feeds. In other species, such as the grazing ruminants, successes have been more modest, largely because these animals are generally produced under extensive conditions and required to interact closely with their environment to harvest forages, but still have been substantial. Despite these successes, demands for further increases in rates of animal production to supply an expanding human population continue to increase, and pressure to utilize the highly productive global breeds to meet these demands is tremendous.

Improvement-oriented animal breeders view global genetic diversity as a resource to be used in pursuit of improved animal productivity. The existing array of livestock breeds is recognized as potentially valuable but primarily as a consumable resource to be integrated into elite populations as appropriate and molded into ever-improving commercial populations. This viewpoint, described as the “utilizationist view” by the U.S. National Academy of Science, emphasizes the potential value of the unique genes and gene combinations found in the various breeds as raw material for breed evolution and emphasizes the value in maximizing that evolution. In the utilizationist view, there may well be a need for a few dozen different pig breeds to meet the demands of different global production environments and markets, but a commitment to maintain all of the 353 currently recorded pig breeds would be seen as excessive and counterproductive.

In contrast to this position, a more “preservationist view” which tends to accord individual breeds a position similar to that of an endangered species and proposes similar management strategies to sustain them. The preservationist view emphasizes the unique history and presumed genetic distinctness of individual breeds and takes breed preservation per se as its goal. Blending of breeds to generate improved, adapted commercial populations is acknowledged to be necessary and desirable, but only against a backdrop of secure and relatively stable populations of the contributing breeds. Retention and continued use of existing breeds in their traditional environments and production systems is promoted, even if public subsidies are required to ensure that use. Synthesis and rectification of the utilizationist and preservationist views are badly needed but have yet to fully occur in the animal breeding community. The utilizationists have, to some extent, become victims of their successes, especially in poultry, dairy cattle, and pig production. Aggressive and widespread sampling and comparative evaluation of breeds in the mid-twentieth century led to the establishment of today’s global breeds and to the replacement of many of the middle-level breeds that were widely represented in commercial production a few decades ago. Although large numbers of locally adapted, relatively lowly productive breeds of poultry and swine still exist globally, the elite global breeds have become so differentiated from these stocks that there is now thought to be minimal opportunity for these breeds to contribute genetic material to the elite global breeds through traditional breeding methods. Thus, industrial stocks of chickens and turkeys have experienced no outside contributions from other breeds in more than 25 years. A similar situation appears to be evolving in modern pig populations, and, at least in temperate regions, the Holstein dairy cow is rapidly becoming the preeminent, and in many cases exclusive, dairy breed. Increasingly, the elite global breeds are viewed as dependent on existing reservoirs of intrabreed genetic diversity for future adaptation and improvement.

In contrast, the preservationist position has been made untenable by the events of global economic and social homogenization. In an increasingly cosmopolitan, interdependent, and rapidly changing world, continued use of 5000 livestock breeds will not occur and probably is not necessary to maintain adequate reservoirs of genetic diversity. However, a responsible fallback position which can ensure the identification, retention, and ready access to the core genetic diversity of each domestic species is badly needed. A reasonable synthesis and integration of conservation, preservation, and improvement activities have been achieved in the plant breeding community, as described elsewhere in this encyclopedia. Facilities for long-term seed storage exist and have both a broad mandate and the budgetary support to acquire and preserve samples representing the full range of diversity within individual species of crop plants and their wild relatives (Box 1). The importance of these programs is acknowledged by both public agencies and private breeding companies. Although there are fundamental differences between plants and animals in the population structure of breeding materials and in the technologies that can be used for genetic resource preservation and use, a more proactive approach to animal germplasm conservation is needed.
Use of Wild Relatives in Livestock Breeding

In contrast to plants, use of wild relatives as a source of genetic material for industry breeding programs in livestock is limited. In most cases, domestic animals represent distinctive species and wild progenitors of these species are no longer found. However, several recent developments suggest that wild relatives may be a useful source of genetic material and that the pool of useful wild relatives may be broader than originally thought.

In bovines, hybrids of domestic cattle with gaur (Bos gaurus), banteng (B. javanicus), and yak (B. mutus) are common in Asia and make important contributions to livestock production. In North America, hybrids of cattle and bison (Bison bison) were evaluated in Canada in the 1960s but with limited success because of near-complete infertility of hybrid males. Recent work in the United States, however, yielded enough fertile hybrid males to allow establishment of herds of animals, known as beefalo, possessing a mixture of genes from cattle and bison. These animals were claimed to possess some desirable adaptational and meat quality traits but are today primarily a novelty.

In western China, genes from the wild ibex (Capra ibex) have been introduced into domestic cashmere goats. The yield of cashmere from ibex is very low, but the cashmere fibers are very fine, averaging about 12 µm in diameter. Typical cashmere goat fibers are 16 or 17 µm in diameter, and the value of the fiber is related to its fineness. Introggression of genetic material from ibex, followed by selection for yield and fineness of cashmere, led to development of animals with 12.5–25% ibex genes that possess increased fineness of the cashmere with near-normal yield. Similar crossing programs have been proposed using the wild vicuña to improve fiber quality in domestic alpaca.

In the Republic of South Africa, there is a need for hardy and well-adapted livestock for use in communal and other low-input production systems. In pigs, crosses between an endangered domestic pig breed, the Kolbroek, and the wild bush pig (Potamochoerus porcus) led to production of fertile offspring with improved hardiness, foraging ability, and capacity to digest fiber. This interspecific cross also indicates the potential to use more distant crosses as a source of genetic material.

Wild relatives of domestic livestock have also contributed to development of genetic maps. Crosses of domestic pig with wild boar (Sus scrofa) were used to develop gene maps for pig, and crosses with red jungle fowl (Gallus gallus) were used in mapping the chicken genome.

In plant breeding, wild relatives of crop plants have long been recognized as a source of genes for disease and pest resistance. Recent work with wild relatives of tomato and rice suggest similar opportunities to improve production traits. Use of gene maps to identify useful genetic material and of improved methods to introduce this material into domestic stocks using either biotechnology or conventional breeding methods now exist for plants and will be developed for animals.

III. MANAGEMENT STRATEGIES FOR CONSERVATION OF LIVESTOCK GENETIC RESOURCES

Leadership in conservation of livestock genetic resources at the global level has come from the Food and Agriculture Organization of the United Nations (FAO), beginning with many Animal Production and Health Papers and continuing with the current Domestic Animal Diversity Information System (accessible via the worldwide web at http://www.fao.org/dad-is). In particular, the FAO Global Strategy for the Management of Farm Animal Genetic Resources, and the associated Guidelines for Development of National Farm Animal Genetic Resource Management Plans, provide for international communication and cooperation in managing domestic animal diversity (Box 2).

A. Inventory, Description, Characterization, and Assessment of Degree of Endangerment

An understanding of the status of domestic animal genetic diversity is a prerequisite to its successful management, and the inventory, description, and comparative characterization of livestock breeds is a key management activity. The FAO World Watch List for Domestic Animal Diversity, currently in its second edition, summarizes the contents of the FAO Global Databank, with
The production of crossbred animals by mating parents of different breeds is common in livestock breeding. Crossbred animals are often superior to their purebred parents and therefore more desirable for commercial farmers. The desirability of crossbred livestock arises from two sources: heterosis and complementarity.

Heterosis, also known as hybrid vigor, is normally manifested as an increase in fitness of crossbreds relative to their purebred parents. Thus, crossbred animals commonly are more fertile, more disease resistant, and better able to cope with environmental stresses than would be predicted from the average fitness of their purebred parents. As a result, productivity traits such as growth and milk or egg production are also usually improved in crossbred animals. In simple terms, purebred individuals commonly show modest levels of inbreeding as a result of the restricted matings required to genetically fix the defining characteristics of the breeds (e.g., colors and horn shape), and inbreeding commonly results in associated modest reductions in fitness. Crossing of purebred animals relieves accumulated inbreeding and provides an economically significant “kick” in performance. Significantly, however, effective use of hybrid vigor requires the crossing of unrelated parent breeds and is maximally expressed only in the first-generation cross. Thus, hybrid vigor can be maximally exploited and reliably captured only when parents of the original breeds are maintained as purebreds and crossed in very specific ways.

Complementarity is a characteristic of the production system which arises when different breeds play different and appropriate roles in crossbreeding systems. In low-input or extensive production systems, adaptation of breeding females to the production environment is critical. If the environment is harsh, the productive capacity of indigenous breeds is often low, thereby allowing demands for nutrients and other inputs to be synchronized with their limited supply. In meat production, mating of females of indigenous breeds to males of more productive and heavily muscled breeds can increase the value of the offspring while maintaining high levels of adaptation in the breeding females. Additionally, the benefits of having a well-adapted mother and of hybrid vigor often permit the crossbred offspring to perform at acceptable levels and increase overall productivity. However, successful use of complementarity again requires maintenance of the adapted, indigenous breeds. Replacement of the original indigenous breeding females with crossbreds results in losses in both adaptation and hybrid vigor, often with serious negative effects on productivity.

Proper use of crossbreeding thus relies on breed diversity. Access to arrays of both adapted and highly productive breeds allows synchronization of the genetic characteristics of both breeding females and crossbred offspring to diverse production environments and markets. However, proper use of crossbreeding also requires that breeds be maintained in their proper roles in the system. In developing nations, in which control over matings is often limited or nonexistent, use of highly structured crossbreeding systems has proven difficult, and initial improvements in productivity in first-generation crosses, as well as the genetic integrity of the indigenous breeds, have been lost through indiscriminate mongrelization of adapted and unadapted breeds. Thus, crossbreeding has significant benefits when an appropriate management infrastructure exists but can have devastating effects on indigenous breeds when not adequately managed.

The FAO Global Databank lists 3019 breeds representing 14 domestic species of mammals and 863 breeds representing 11 avian species (Tables I and II). Of these, data on population size and status are available for 2191 mammalian and 733 avian breeds, whereas the remaining breeds possess only cursory information on breed characteristics and distribution. Reporting of breed information is not consistent throughout the world. European breeds dominate the databank with 1172 mammalian and 516 avian entries, primarily because of a long history of concern over genetic diversity in Europe and because Europe was the cradle for the concept of breed formation. African breeds, in contrast, are almost certainly underreported, with only 291 mammalian and 105 avian breeds listed. The total number of these breeds at risk because of limited population size, uncontrolled crossbreeding, or declining use.
TABLE I
Global Summary: Mammalian Breeds at Risk, by Species

<table>
<thead>
<tr>
<th>Species</th>
<th>No. of breeds</th>
<th>On file</th>
<th>With population data</th>
<th>Endangered</th>
<th>Critical</th>
<th>At risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buffalo</td>
<td>72</td>
<td>95</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Cattle</td>
<td>787</td>
<td>982</td>
<td>77</td>
<td>98</td>
<td>135</td>
<td></td>
</tr>
<tr>
<td>Yak</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Goat</td>
<td>351</td>
<td>267</td>
<td>31</td>
<td>13</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Sheep</td>
<td>920</td>
<td>656</td>
<td>85</td>
<td>34</td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>Pig</td>
<td>553</td>
<td>285</td>
<td>39</td>
<td>30</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>Ass</td>
<td>77</td>
<td>24</td>
<td>2</td>
<td>7</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td>384</td>
<td>277</td>
<td>87</td>
<td>33</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>Dromedary</td>
<td>50</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Bactrian camel</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Alpaca</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Llama</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Guanaco</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Vicuña</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3019</td>
<td>2191</td>
<td>324</td>
<td>177</td>
<td>501</td>
<td></td>
</tr>
</tbody>
</table>

*See Table III for definitions.

b Sum of endangered and critical breeds taken from FAO (1995).

The breeding of animals is thus projected to be about 4000 and the total number of avian breeds likely exceeds 1000. To place these numbers in context, the number of breeds for these 14 domestic species of mammals approximately equals the number of mammalian species on Earth.

To be useful, inventory information must be augmented with some objective assessment of degree of

TABLE II
Global Summary: Avian Breeds at Risk, by Species

<table>
<thead>
<tr>
<th>Species</th>
<th>No. of breeds</th>
<th>On file</th>
<th>With population data</th>
<th>Endangered</th>
<th>Critical</th>
<th>At risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicken</td>
<td>606</td>
<td>612</td>
<td>227</td>
<td>47</td>
<td>274</td>
<td></td>
</tr>
<tr>
<td>Domestic duck</td>
<td>62</td>
<td>94</td>
<td>18</td>
<td>1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td>31</td>
<td>20</td>
<td>6</td>
<td>3</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Muscovy duck</td>
<td>14</td>
<td>13</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Domestic goose</td>
<td>39</td>
<td>55</td>
<td>18</td>
<td>10</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Guinea fowl</td>
<td>22</td>
<td>17</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Partridge</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Pheasant</td>
<td>8</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Quail</td>
<td>24</td>
<td>23</td>
<td>0</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Pigeon</td>
<td>19</td>
<td>16</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Otter</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>863</td>
<td>733</td>
<td>276</td>
<td>96</td>
<td>372</td>
<td></td>
</tr>
</tbody>
</table>

*See Table III for definitions.

* Sum of endangered and critical breeds taken from FAO (1995).
TABLE III
Criteria for Determining Degree of Endangerment for Breeds of Domestic Livestock

<table>
<thead>
<tr>
<th>Category</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not at risk</td>
<td>Total number of breeding females is greater than 1000 and total number of breeding males is greater than 20, or Population size approaches 1000, the percentage of females being bred to males of the same breed is near 100%, and the overall population size is increasing</td>
</tr>
<tr>
<td>Endangered</td>
<td>Total number of breeding females is between 100 and 1000 or the total number of breeding males is between 5 and 20, or Overall population size is slightly below 100 and is increasing and the percentage of females being bred to males of the same breed is more than 80%, or Overall Population size is slightly more than 100 but is decreasing and the percentage of females being bred to males of the same breed is less than 80%</td>
</tr>
<tr>
<td>Critical</td>
<td>Total number of breeding females is less than 100 or the total number of breeding males is less than 5, or Overall population size is slightly more than 100 but is decreasing and the percentage of females being bred to males of the same breed is less than 80%</td>
</tr>
<tr>
<td>Extinct</td>
<td>It is no longer possible to recreate the breed population from live animals and/or cryopreserved gametes or embryos</td>
</tr>
<tr>
<td>Unknown</td>
<td>Information on population size is not available</td>
</tr>
</tbody>
</table>

endangerment of the populations involved. The FAO guidelines for determining breeds at risk are shown in Table III. Based on these guidelines, about 23% of mammalian and 51% of avian breeds with population data are considered at risk. Furthermore, only 27% of mammalian breeds and 48% of avian breeds that are at risk have active programs in place to promote their conservation. The combined total of 30% of breeds at risk, if extrapolated to the projected 5000 global breeds, results in a figure of about 1500 breeds that are vulnerable to loss or to serious erosion of genetic diversity. Because many of these endangered breeds are found in harsh production environments with minimal care and management, they can reasonably be anticipated to possess unique genetic adaptations and disease-resistance characteristics. However, in common with many endangered species, they are at risk of disappearing before they can be adequately evaluated and characterized.

B. Preservation Strategies

Conservation of livestock genetic resources will almost certainly involve both in situ preservation of breeding herds in agricultural production systems and ex situ preservation of animals, gametes, and embryos in reserves and cryogenic repositories. Issues of current utility, degree of endangerment, and cost of preservation will determine the method(s) of choice.

1. In Situ Strategies

Strategies that are based on effective use of livestock genetic diversity in commercial and/or subsistence production systems are most likely to provide the security necessary to allow breeds to remain viable and to maintain the population sizes necessary to ensure continued improvement. Thus, a key activity in livestock genetic resource conservation is the comprehensive appraisal of the productive merit of the various breeds under current productive conditions. Most of the world’s domestic animals are found in the developing nations, and nearly 2 billion people obtain at least part of their daily livelihood from livestock. Rapid changes in production conditions have occurred in some of these nations, but the number of animals that are used in subsistence production systems remains very large. Levels of veterinary care, supplemental feeding, and shelter from climatic stresses often remain limited, and adaptational characteristics of indigenous livestock remain important to many farmers. Studies throughout the world have confirmed that when environmental conditions are harsh, imported temperate breeds of high production potential often experience increased levels of mortality and morbidity and reduced reproductive success, and as a result they may be less productive per unit time or per unit of inputs than indigenous breeds. Often, advisers recommend changes in the production environment, even though farmers often do not have the financial resources or market incentives to do so. A more responsible strategy, and certainly one that is more consistent with conservation of livestock diversity, is to focus on the retention and improvement of local breeds. The implementation of modern livestock improvement procedures in indigenous breeds may permit them to be improved in parallel with gradual improvements in production conditions while retaining important adaptational characteristics. An example of such an approach is found in India, where Operation Flood was begun more than 30 years ago to enhance milk production and marketing opportunities for smallholders in western India. The goal
was to improve the nutrition of the people of India by increasing supplies of buffalo and cattle milk. Certainly, Operation Flood and its successor projects have been a quantitative success. India is now the world’s number 1 producer of milk. More important, this success was achieved without recourse to establishment of high-volume Western dairy production systems or a rapid infusion of poorly adapted foreign germplasm. Instead, emphasis was placed on establishing marketing conditions that would reward efficient producers and provide incentives for improved production practices. Only recently has this identification of superior germplasm within the indigenous breeds of cattle and buffalo and the strategic use of imported breeds become a part of this program.

In other cases, however, the pace of social change has been so rapid that indigenous breeds in the developing nations cannot meet the challenges. Under these conditions, pressures on indigenous breeds can become extreme very quickly, and careful analysis is required to identify appropriate generic resource development activities. Developments in pig production in China provide an example. China has the world’s most diverse array of pig breeds, with at least 50 and perhaps as many as 100 distinct breeds. Almost all evolved in situations in which grain was primarily used for human food. The pigs therefore became adapted to the use of fibrous by-product feeds from grain and vegetable production. They also were relatively slow growing (to synchronize their nutrient requirements to their limited food supply), quite fat (because animal fat was an important energy source for peasant farmers), and, in some cases, highly prolific. In the 1980s, government policies in China were changed to provide greater access to grain for animal feed. Under these conditions, the indigenous pig breeds were markedly inferior to the faster growing, leaner, and more feed-efficient Western breeds. Government-sponsored importations and entry of multinational pig breeding companies into the market resulted in widespread crossbreeding of indigenous pigs with Western breeds. An extensive network of government artificial insemination (AI) centers for pigs facilitated the use of imported breeds, and the development of more sophisticated urban markets heightened preferences for leaner pork. By the early 1990s, many of the AI stations no longer provided boars of the local breeds.

The events in Chinese pig breeding were fully rational responses to changed government policy and to changes in the economy and the nature of society. Both farmers and consumers benefited from the changes. However, the result has been to endanger a whole array of local breeds, many of which possess globally unique characteristics.

In developed nations, focus on a declining number of elite breeds in relatively intensive production systems has likewise endangered many traditional breeds. The situation is particularly acute in Europe, and greater cognizance of the problem exists there. The European Association of Animal Production Working Group on Animal Genetic Resources in 1993 listed 877 European breeds of cattle, sheep, goats, pigs, and horses. Of these, less than half (412 breeds) were considered secure. The remainder were classified as “potentially endangered” (162 breeds), “minimally endangered” (89 breeds), “endangered” (43 breeds), or “critically endangered” (138 breeds). No information was available for 13 breeds.

In response to this imminent contraction in livestock genetic diversity, many nations of Western Europe have initiated programs to maintain endangered breeds. Although often using a combination of in situ and ex situ techniques, high priority is placed on in situ preservation. The importance of livestock breeds as a cultural and historical resource reflecting the heritage of the nation or as a component of unique and perhaps themselves endangered agroecosystems is particularly recognized in Western Europe. The contribution of livestock breeds to public welfare is sometimes referred to as “landscape value,” and public resources are increasingly being directed toward breed conservation activities. These programs may involve maintenance of endangered breeds on public facilities or direct payments to farmers who maintain endangered breeds.

Throughout the developed nations, “grassroots” organizations play a significant role in in vivo preservation of endangered livestock breeds. Non governmental organizations (NGOs) have for many years provided leadership for conservation of livestock diversity. Examples include the Rare Breeds Survival Trust in the United Kingdom, the American Livestock Breeds Conservatory in the United States, and the Canadian Foundation for the Conservation of Farm Animal Genetic Resources. The success of these organizations attests to the potential to involve committed private individuals in breed conservation. They have been most successful in wealthy countries in which significant private resources can be exploited, but the identification of private patrons to aid in breed conservation can also occur in developing nations. The success of these organizations relies on members who are committed to, and educated about, the technical aspects of genetic conservation. Otherwise, well-meaning but misdirected breed-
ing policies may result in losses of genetic diversity. Public–private partnerships to provide members of these NGOs with technical expertise can be particularly beneficial.

2. Ex Situ Strategies

Ex situ preservation usually involves cryopreservation of gametes and embryos but may also involve live animals kept in farm parks, on research farms, or in other noncommercial settings. In the developed nations, farm parks have increased in popularity and often contribute to meaningful maintenance of biodiversity through their association with responsible grassroots organizations. Live-animal, ex situ preservation can also occur at publicly funded facilities, generally under conditions that at least approximate those found in commercial agriculture but which often do not replicate the particular conditions under which the breeds evolved and were traditionally used. In these situations, selection for unique adaptational characteristics is released but with proper breeding management and adequate population size, the key genetic characteristics of the breed can be retained for many generations. Costs of ex situ live-animal preservation are high, however, involving feeding and daily care of breeding animals. The risks associated with live-animal ex situ programs are therefore also high, especially in developing nations in which sustained funding for long-term conservation programs may not be available and in which risks of social and political upheaval are greatest.

The use of cryopreserved gametes, embryos, and tissues is a more common form of ex situ preservation. In farm animals, sperm cells can be successfully frozen and stored for future use in all species, although success rates from use of frozen sperm vary considerably among species. Techniques for collection, cryopreservation, and subsequent use of sperm cells are relatively well developed, a single collection provides a relatively large number of gametes, and multiple samples are relatively easy to obtain. For these reasons, cryopreserved sperm cells are the most common material used for ex situ preservation of endangered breeds.

Cryopreserved sperm are ideally suited to support in situ preservation activities. Storage of sperm from a wide sample of males of a breed provides future access to the genetic material of these representative foundation animals. Losses of genetic diversity in living populations can thus, if necessary, be restored by use of sperm from males of past generations.

However, cryopreserved sperm cells are not particularly efficient for regeneration of a breed that has become extinct. Sperm cells contain only a sample half of the animal’s DNA; therefore, restoration of an extinct breed from cryopreserved sperm requires a “grading-up” process in which sperm is used on females of a different breed over several generations to eventually create animals that have a majority of their nuclear genes from the cryopreserved breed. The efficiency of use of sperm cells to restore a breed depends on the generation time and the fecundity of the species. For example, restoration of 93.75% of the nuclear genes (i.e., four generations of upgrading) can be accomplished in about 3.5 years in pigs, and use of 100 sows initially could conservatively result in production of 1000 breeding females by generation 4. In contrast, in cattle a minimum of 10 years would be required using conventional breeding techniques to produce even a small number of animals possessing 93.75% of the nuclear genes of the preserved breed.

Also, although nuclear genes can be adequately preserved using frozen sperm, cytoplasmic DNA found in animal mitochondria are contributed only via the ovum and can thus be preserved only by storage of cryopreserved ova or embryos. Cryopreservation and subsequent in vitro fertilization of ova are not yet practical for domestic species. Embryo cryopreservation, however, is practical in cattle, sheep, and goat and permits preservation of the full genome, both nuclear and cytoplasmic. However, collection of embryos for cryopreservation is more difficult and expensive than collection of sperm cells and yields of embryos are much lower, often on the order of only two to eight embryos per mating. Also, embryos of pig and poultry cannot be reliably cryopreserved.

Current recommendations for ex situ preservation programs thus generally focus on extensive use of frozen sperm cells. However, small numbers of cryopreserved embryos (in species for which this is possible) and/or small populations of breeding females preserved under either in situ or ex situ conditions generally must also be maintained to serve as a source of cytoplasmic genes and to allow efficient regeneration of the breed. A store of cryopreserved embryos produced by mating 25 pairs of unrelated parents and bolstered by cryopreserved sperm from 25 unrelated males can effectively capture the genetic diversity of a breed for long-term preservation and future use.

Cryopreservation of somatic cells is also becoming common in ex situ preservation programs under the assumption that cloning of new individuals from cryopreserved adult cells will eventually become a reality. Use of cloning to support preservation of rare breeds will be discussed in Section IV.C.
IV. POTENTIAL IMPACTS OF NEW TECHNOLOGIES

A. Advances in Cryopreservation

Several technological advances in cryopreservation promise to enhance our ability to preserve biodiversity. Key research areas in cryopreservation include

1. Cryopreservation of pig embryos to allow capture of the full genome including mitochondrial DNA. This is especially important in Asia, where the number of endangered pig breeds is large.

2. Cryopreservation of poultry embryos and semen. In chickens, success rates from use of cryopreserved semen remain low. Better success rates would prompt greater use of frozen semen in the poultry industry, which would then facilitate storage and access to frozen gametes for conservation. Cryopreservation of fertilized eggs is likely to remain difficult. However, it appears possible to extract and cryopreserve the germinal disk of cells from the fertilized egg separate from the egg itself.

3. An improved understanding of why semen from some males and embryos from some breeds do not survive freezing well. Large individual differences exist in success rates for cryopreservation of semen and embryos. In commercial cattle breeding, there is considerable selection for "freezability" of semen, and because the populations are large this selection has little impact on the genetic diversity within the breed. However, endangered breeds are commonly represented by only a few males, and if semen from several of these cannot be satisfactorily cryopreserved, a significant segment of the biodiversity of the breed may be lost. Success rates for frozen embryos are also often lower in poorly characterized endangered breeds, and these breeds sometimes respond poorly to the hormonal regimen required for embryo collection. More reliable and robust techniques for semen and embryo recovery and cryopreservation are thus needed.

B. Gene Mapping

The development of comprehensive gene maps is occurring rapidly for several domestic species. These maps will facilitate screening of breeding animals and of cryopreserved material for unique genetic variants. Understanding of the molecular genetic control of quantitative traits is still in its infancy but is expanding rapidly. Identification of quantitative trait loci (i.e., regions of the DNA which influence performance) is beginning and promises to aid in the screening of rare breeds for useful genes. Improved gene maps will also aid in determining genetic relationships among breeds, thereby assisting in prioritization of candidates for preservation. Long-term conservation of the 5000 or more global breeds of mammalian and avian breeds is unlikely to occur, but knowledge of relationships among breeds would allow identification of a smaller "core collection" of breeds representing the bulk of the genetic diversity within each species.

C. Cloning

Cloning of farm animals from adult cells, if perfected and commercialized, can have far-reaching effects on both the use of biodiversity and the efficiency of its preservation. Certainly, the widespread use of genetically identical cloned individuals in commercial production systems would reduce the biodiversity present within those systems and increase the genetic vulnerability of the animals to specific diseases or other environmental stressors. Preliminary risk assessment studies are already underway to attempt to optimize the number of clonal lines that should be used within a herd or flock to minimize these dangers.

The ability to efficiently produce clonal offspring from adult animals (i.e., from individuals whose performance has already been characterized), however, has potential to facilitate synchronization of genetic resources to specific production conditions. In such a scenario, breeding animals that have demonstrated exceptional adaptation and productivity in a particular environment could be clonally reproduced for use in that specific environment.

The ability to produce clonal offspring from adult cells would have tremendous implications for preservation of endangered genetic resources. Large numbers of cells could be harvested from individual animals using minimally invasive biopsy techniques and cryopreserved for future use. Production of clonal offspring from these cryopreserved cells would recreate the full genetic complement (both nuclear and cytoplasmic) of the preserved animals. Furthermore, because sampling of somatic cells for cloning is relatively easy, large numbers of founder animals could be sampled, increasing the biodiversity present within the sample. Complications involved in harvesting and preservation of gametes and embryos would be circumvented and efficiency of storage of animal genetic material would approach that enjoyed in plants by storage of seeds.

Although technology for clonal propagation of adult animals cannot yet be relied on for preservation of animal genetic resources, preliminary results hold great
promise. Dolly, the sheep produced by cloning of adult mammary cells at the Roslin Institute in Scotland, represented a watershed event in this technology. Subsequent cloning of cattle and pigs from adult cells has occurred, and recent reports in the popular press (eagerly awaited in the scientific literature) indicate that easily accessible bovine cells obtained from the ear may be satisfactory for cloning. Already, one of the last individuals of an endangered New Zealand breed, the Enderby Island cow, has been cloned to attempt to save the breed.

V. CONCLUSIONS

An assessment of likely trends in global livestock genetic diversity must recognize that demands of population growth, economic development, and urbanization will require significant increases in animal productivity worldwide. The normal processes of breed evolution have clearly accelerated during the past 25 years. Breeds are disappearing at more rapid rates than in the past and the rate of new breed formation has not correspondingly increased. Many indigenous breeds are ill-suited to contribute to meeting those challenges in a timely manner and will likely become endangered or restricted to subsistence production systems. Global, temperate breeds, with their high production potentials, will contribute to world food production, placing further pressure on less-productive indigenous breeds. However, the adaptational characteristics of local breeds have great value in helping to synchronize animal genetic resources with local stressors and should contribute to development of new, improved breeds. Thus, a new synthesis of livestock genetic resources is needed in which genetic resources from a variety of sources are managed to maximize livestock productivity in sustainable production systems, and this synthesis must be accompanied by programs to protect breeds which may be endangered or replaced.

See Also the Following Articles

BREEDING OF PLANTS • CAPTIVE BREEDING AND REINTRODUCTION • CATTLE, SHEEP, AND GOATS, ECOLOGICAL ROLE OF • GENETIC DIVERSITY • INBREEDING AND OUTBREEDING

Bibliography


I. The First Farmers
II. Plant Breeding, Crop Varieties, and Biodiversity
III. Conclusions

GLOSSARY

biodiversity (biological diversity) Hereditary variation in life-forms at all levels of organization. Examples are diversity among wild species, among individuals within those species, among crop species, and among and within varieties of a given crop. crop varieties Named populations of crop plants that possess recognizable features and known utility for food, feed, or fiber.
durable resistance Inherited resistance to a disease or insect pest, with relatively long effective lifetime. It is usually imparted by several genes of individually small effect.
genotype Genetic constitution of an organism or group of organisms such as a plant or plant variety.
heterosis Hybrid vigor, exhibited when offspring are larger, more vigorous, and more productive than the parents.
monoculture Usually refers to growing a single uniform plant variety over a large area.
multilines Planned mixtures of different selections of the same variety that differ only in genes for disease or insect resistance.

polyculture Usually refers to growing several crops in mixture or in rotation, often with variability among the plants of a given crop population.
specific resistance Narrowly targeted resistance to a specific genotype of disease or insect pest. It is usually imparted by a single gene of large effect, and typically with a relatively short effective lifetime.
traits Recognizable characteristics of a crop variety, such as plant height, grain color, specific disease resistance, yield potential, or tolerance to heat and drought.
transgenes Genes imparted to an organism by means of biotechnology rather than sexual hybridization.

THE GOAL OF CROP AGRICULTURE IS TO REDUCE BIODIVERSITY in favor of increased and easily accessible food supplies. This was true for the first farmers and it is still true today. The reduced biodiversity brings advantages but also disadvantages to farmers—to food producers—because biodiversity is essential for the maintenance of ecosystem stability. The following discussion shows how farmers and plant breeders have dealt with this paradox through the millennia up to the present time, and how they can cope with it in the years to come.
E. O. Wilson defines biodiversity (biological diversity) as "the totality of hereditary variation in life forms, across all levels of biological organization, from genes and chromosomes within individual species to the array of species themselves and finally, at the highest level, the living communities of ecosystems such as forests and lakes" (Wilson, 1994).

**Box 1**

The biodiversity found among and within crop varieties is a subsection of the totality described by Wilson. Today's professional plant breeders usually describe that subsection as "genetic diversity."

I. THE FIRST FARMERS

Ten thousand years ago in western Asia, and later in eastern Asia and the Americas, hunter-gatherers began to specialize in gathering and nurturing a few favorite species of plants (Smith, 1995). Eventually the favored species were purposely planted in isolated plots, seed from the most desirable plants was replanted, and gradually domesticated crops were invented.

The domesticates represented only a few of the hundreds of species that had furnished food for the hunter-gatherers. The fields in which they were grown were cultivated and weeded, and the farmers did their best to eliminate all plants except the ones they favored. They depended not only on a greatly reduced number of species for the majority of their food supplies, but they also used only a portion of the total genetic variability within each species. They selected only those rare plants in a species that had the traits needed for domestication, such as nonshattering seeds or uniform germination. As a result, plant biodiversity in food production was greatly reduced.

But within each of the newly domesticated crops, farmers selected and reselected for different kinds of desirable plant and seed types, gradually increasing the diversity in the narrow base that had been provided by the original selections. Within a species, groups of related individuals were selected for special grain types, adaptation to particular heat and drought conditions, or for resistance to disease or insect pests. In this way the farmers developed what we now call crop varieties. The varieties were recognized individually for their unique constellations of useful traits, and often they were given names that described their most important features. Biological diversity, in this case, genetic diversity, had been increased within each crop species. This diversity came in part from recombination of traits in the original selections, in part from genetic mutation, and in part from accidental backcrossing to the still-existing wild species in those cases where they still grew in proximity to the farmers’ fields.

Although the reduction of biodiversity in favor of carefully tended plantings of specially selected varieties increased amounts and availability of food from favored species, it brought its dangers as well. Uniform and broad expanses of one crop enabled easy multiplication of diseases and insects especially adapted to the particular genetic constitution of the farmers’ varieties. Catastrophic attacks of disease, or invasions of insects, or unprecedented heat and drought, or unusually cold and wet growing seasons caused wholesale crop destruction and ensuing famines whenever the farmers’ crops had insufficient diversity to provide at least some plants or some species with ability to withstand the assaults. These problems, a consequence of lack of the right kinds of genetic diversity among and within crop varieties, are as old as agriculture. They are recorded in myth and in written history, and they still occur wherever crop agriculture is practiced. Plant breeders today use the term “genetic vulnerability” to describe the condition that results “when a crop is uniformly susceptible to a pest, pathogen, or environmental hazard as a result of its genetic constitution, thereby creating a condition for disaster” (Board on Agriculture and National Research Council, 1993).

The phrase “genetic vulnerability” is new but the condition is not. Red rust on wheat (Triticum spp.) in Roman times, mass poisoning from ergot-tainted rye (Secale cereale L.) in medieval Europe, potato late blight (Phytophthora infestans) in the nineteenth century, and the widespread Southern Corn Leaf Blight Race 1 epidemic of corn (Zea mays L.) in the United States in 1970 all were due to insufficient biodiversity for important genes in the affected crops (Large, 1962; Matossian, 1989; National Research Council, 1972). Each crop was uniformly susceptible to a critical disease—it was genetically vulnerable. As a consequence, yield and crop quality were drastically reduced. (All of these diseases were fungi. Wheat rust is caused by several species of Puccinia, ergot by Claviceps purpurea (Fr.: Fr.) Tul., potato late blight by Phytoph-
thorae infestans (Mont.) de Bary, and Southern Corn Leaf Blight Race T by Cochliobolus heterostrophus (Drechs.) Drechs., Race T.

Box 3
Each year the early Romans sacrificed a red dog to the corn god Robigus in hopes that he would accept it with pleasure and therefore protect their wheat crop from the destructive red rust that usually appeared at about the time of the ceremony, on April 23. Their wheat crop was genetically vulnerable (Large, 1962).

II. PLANT BREEDING, CROP VARIETIES, AND BIODIVERSITY

A. Applications of Biodiversity in Crop Agriculture

1. Genetic Diversity in Place—Spatial Diversity

Early farmers found that one way to protect themselves from the problems of catastrophic disease, insect, and environmental problems was to maintain some degree of biodiversity among and within their crop species. Planting several kinds of crops gave assurance that at least some of them could survive any specific attack of disease, insects, or bad weather. Planting several varieties of a given crop increased assurance that some varieties of that crop could be harvested. Furthermore, genetic variability (seen by the farmers as phenotypic variability, i.e., visual variability) within a crop variety gave some assurance that at least some plants in the variety would be resistant to a specifically targeted, virulent strain of disease or insect. Each of these kinds of biodiversity (genetic diversity) decreased the odds of starvation for subsistence farmers.

Box 4
Genetic diversity in place (in the field) decreases the odds of catastrophic crop failure and starvation for subsistence farmers.

2. Genetic Diversity in Time—Temporal Diversity

Commercial farmers in industrialized countries such as the United States depend more heavily on genetic diversity in time than on genetic diversity in place.

Farmers have always, sometimes without even trying, relied on a second kind of biodiversity, “genetic diversity in time” (temporal diversity). Seed planted in successive seasons was never quite the same as that from the previous year, for the combined onslaughts of disease, insects, and weather eliminated many susceptible plants. Only the survivors could produce seed, and their genotypes reflected tolerance to the biological and weather-related constraints of the previous growing season. Varieties changed continually, even in the absence of purposeful selection by the farmers. The sequential genetic changes represented continuing adaptation to prevailing constraints to yield.

Farmers made conscious selection for changes in their varieties as well, sometimes for desired grain or plant traits, sometimes for yield per se. They also looked at their neighbors’ varieties in the next village, the next valley, or even farther afield if transportation and communications allowed. Varieties were traded, experimental plantings were made, and new varieties often replaced older traditional varieties if farmers judged that the new ones did a better job of meeting their needs. This practice, of purposely producing “genetic diversity in time” via selection or variety replacement, is still common among farmers in areas that are little affected by professional plant breeding.

Today’s professional plant breeders and modernized farmers work as a team to employ genetic diversity both in place and in time to provide protection against pest and weather problems, but they use tactics that are different from those employed by yesterday’s farmer/breeders. They take advantage of modern technology and newer kinds of societal organization (Duvick, 1984). Planted areas are larger, crops are more uniform, replacements are more frequent, and sources of new genetic diversity come not from the neighboring village but from anywhere in the world. Plant breeders ceaselessly produce new varieties with increased yield, needed new kinds of pest resistance, adaptation to new cultural regimes, or better tolerance to bad weather (Duvick, 1996). Farmers continually experiment with the new varieties, choose the best of them, and plant
them extensively, often so extensively that the stage
is set for another explosive round of increased pest
problems or too-uniform susceptibility to unexpected
weather problems. Following such upsets, farmers im-
mediately switch to other varieties, choosing from the
ever-present catalogs of new introductions. Modern
communication and transportation facilitate such rapid
changes. They would not have been possible in ear-
tier times.

3. Genetic Diversity in Reserve

Genetic diversity in reserve has several layers,
each layer farther away in time from use on the
farm, but also broader in the amount of biodiver-
sity that it can provide.

There is a caveat—a warning—about reliance on
genetic diversity in time. It will work only when
there is sufficient "genetic diversity in reserve" to
supply a continuing stream of new varieties. The
farmer-breeders of earlier times used their heteroge-
neous varieties as a reserve of genetic diversity as
well as for active food production. They had no other
ready reserve, no seed company offering a diverse
supply of new varieties each year. But these subsistence
farmers could also look farther afield to varieties
grown by neighbors, and sometimes they might select
from progeny of an occasional outcross to wild or
weedy crop relatives. So, ready reserves and backup
reserves of biodiversity were available and were used
by subsistence farmers, but the scope usually was
limited to the immediate locale and the current sea-
son's crops. Many of today's subsistence farmers still
use these methods of conserving and using genetic
reserves; they have no other options.

Today's commercial farmers, served by full-time pro-
fessional plant breeders, have much broader reserves.
The breeders have a supply of hundreds or even thou-
sands of experimental varieties for each crop and each
adaptation zone. They are grown in performance trials
every year, enabling breeders to supply well-tested new
varieties when needed. These are "front-line reserves." The
plant breeders have other reserves, farther to the rear.
These are "breeding pools" of materials in the process of hybridization and selection for production
of new experimental varieties.

Behind the breeding pools reserve is a very broad,
highly diverse set of varieties and wild and weedy rela-
tives from all around the world, stored in long-term seed
banks such as the regional USDA plant introduction
stations and the National Seed Storage Laboratory at
Ft. Collins, Colorado. Table I lists some of the world's
largest seed banks and shows the size of their collec-
tions.

Finally, available but not easily accessed are the
highly diverse materials now growing on farms and in
the wild in all parts of the world. Collecting expeditions
can gather materials from these sources, paying especial
attention to "centers of diversity" for a given species,
and make them available to plant breeders for long-term
breeding programs with the ultimate goal of adding to
the biodiversity of materials offered to farmers as new
crop varieties. Table II lists major and minor centers
of diversity worldwide and notes some of the important
crops in those centers.

Genetic diversity in reserve thus has several layers,
each layer farther away in time from use on the farm,
but also broader in the amount of biodiversity that it
can provide. As new varieties are introduced, become
popular, and then eventually (and always) show unsus-

### Table I

<table>
<thead>
<tr>
<th>Country/center</th>
<th>Crop</th>
<th>Number of accessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>All</td>
<td>557,000</td>
</tr>
<tr>
<td>China</td>
<td>All</td>
<td>400,000</td>
</tr>
<tr>
<td>Russia</td>
<td>All</td>
<td>325,000</td>
</tr>
<tr>
<td>India</td>
<td>All</td>
<td>76,800</td>
</tr>
<tr>
<td>Japan</td>
<td>All</td>
<td>60,000</td>
</tr>
<tr>
<td>BRII</td>
<td>Rice</td>
<td>86,000</td>
</tr>
<tr>
<td>ICRISAT</td>
<td>Sorghum, millet, chickpea, pea, pigeon pea</td>
<td>86,000</td>
</tr>
<tr>
<td>ICARDA</td>
<td>Cereals, legumes, forages</td>
<td>77,000</td>
</tr>
<tr>
<td>CIMMYT</td>
<td>Wheat, maize</td>
<td>73,000</td>
</tr>
<tr>
<td>CIAT</td>
<td>Common bean, canava, forages</td>
<td>66,000</td>
</tr>
<tr>
<td>IITA</td>
<td>Cowpea, rice, root crops</td>
<td>60,000</td>
</tr>
<tr>
<td>AVRDC</td>
<td>Vegetable crops</td>
<td>38,300</td>
</tr>
<tr>
<td>CIP</td>
<td>Potato, sweet potato</td>
<td>12,000</td>
</tr>
</tbody>
</table>

ions" typically are individual crop varieties. BRII, International Rice Research Institute; ICRISAT, International Crops Research Institute for the Semi-Arid Tropics; ICARDA, International Center for Agricul-
tural Research in the Dry Areas; CIMMYT, Centro Internacional de Mejoramiento de Maíz y Trigo; CIAT, Centro Internacional de Agri-
cultura Tropical; IITA, International Institute for Tropical Agricultu-
ture; AVRDC, Asian Vegetable Research and Development Center, C.I.P., Centro Internacional de la Papa.*
**TABLE II**

<table>
<thead>
<tr>
<th>Major Centers</th>
<th>Important food plants native to the center</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ethiopia</td>
<td>Barley, castor bean, coffee, flax, onion, sorghum, wheat</td>
</tr>
<tr>
<td>2. Mediterranean</td>
<td>Aquagamous, beets, cabbage, lettuce, grapes, oats, olive</td>
</tr>
<tr>
<td>3. Asia Minor</td>
<td>Alfalfa, barley, carrots, cherry, date palm, pear, wheat</td>
</tr>
<tr>
<td>4. Central Asiatic</td>
<td>Apple, apricot, cantaloupe, cotton, grapes, onion, peas</td>
</tr>
<tr>
<td>5. Indo-Burma</td>
<td>Cotton, cucumber, eggplant, lemon, orange, pepper (black), rice</td>
</tr>
<tr>
<td>6. Thailand, Malaysia, Indonesia</td>
<td>Banana, breadfruit, coconut, ginger, grapefruit, sugarcane, yam</td>
</tr>
<tr>
<td>7. China</td>
<td>Buckwheat, peach, radish, rhubarb, rice, soybean, tea</td>
</tr>
<tr>
<td>8. Mexico–Guatemala</td>
<td>Bean (common), corn, cacao, cotton, pepper (red), squash, sweet potato</td>
</tr>
<tr>
<td>9. Peru–Ecuador–Bolivia</td>
<td>Bean (lima), cacao, cotton, guava, papaya, potato, tomato</td>
</tr>
</tbody>
</table>

**Minor Centers**

| 10. Southern Chile | Potato, strawberry (Chilean) |
| 11. Brazil–Paraguay | Brazil nut, cashew, cassava, peanut, pineapple |
| 12. United States | Sunflower, blueberry, cranberry, Jerusalem artichoke |

*Adapted from G. Wilkes (1983). Current status of crop plant germplasm. CRC Critical Reviews in Plant Science 1, 133–181. These centers are the place of origin and/or early domestication of most of the world’s food crops. They still contain a very large share of the global genetic diversity for those crops.

Crop agriculture, from its beginning and continuing to the present time, entails a continual struggle to balance the imperative of reducing biodiversity to maximize yield with the imperative of increasing biodiversity to maximize stability. Farmers must constantly balance their commitment to grow only the most productive varieties of a few favored crops with their need to maintain sufficient biodiversity to protect against unexpected failures of the “best” varieties and the most favored crops.

The terms “monoculture” and “polyculture” sometimes are used to describe the two extremes of crop biodiversity. Monoculture refers to growing a single, usually uniform, variety over large expanses. Polyculture refers to the practice of growing several crops in rotation or in mixture, often in smaller areas and often with some variability within or among varieties. Although both terms are used with remarkable lack of precision and consistency, they do express the fact that crop culture and crop genotypes can vary widely across the global farming community.

Usually overlooked, but nevertheless unavoidably true, is the fact that even the most diverse polycultures provide only a fraction of the biodiversity present in “natural” ecosystems—ecosystems that are relatively untouched by human intervention. A polyculture cropping system, with two or three crops in one field and a certain amount of variability of genotype in each crop, is much closer to a monoculture than to any relatively pristine ecosystem, excepting perhaps a few deserts, tundras, or salt flats. Even relatively simple natural ecosystems, such as the tallgrass prairies of the American Midwest, contain hundreds of highly variable plant species compared to the half dozen or more that might be found in one of the more diverse agricultural polycultures.

But comparisons aside, the small amount of biodiversity provided to farming systems by biodiversity in place (spatial) and biodiversity in time (temporal) is essential and beneficial (Tilman et al., 1999). When properly managed it can make the difference between famine and plenty, or between economic ruin and a satisfactory income. The particular kinds and amounts of needed biodiversity vary by crop, by farming system, and by the norms and needs of societal and economic systems. This article reviews some of these issues, concentrating on the use of biodiversity for the development of field crops grown by commercial farmers in the United States. Its content draws heavily on an informal survey of a score of experienced professional plant breeders.

**B. Precarious Balance: Monoculture and Polyculture**

**Box 7**

The most diverse polycultures provide only a fraction of the biodiversity present in “natural” ecosystems, but they nevertheless may can make the difference between starvation and plenty in some farming systems.
BREEDING OF PLANTS

C. Contributions of Biodiversity to Today's Crop Varieties

1. Increased Yield

Box 8

As the relationship of base-broadening germplasm to that which is being improved decreases, so do the odds of producing successful new varieties, yet the odds increase that the infrequent successes will have a significant advantage in yield or other important traits.

Biodiversity in the form of genetically diverse breeding materials is essential for increasing the yielding ability of farm crops. Yield is the key trait for breeders of commodity crops such as wheat, soybeans (Glycine max (L.) Merrill), and corn. New varieties must yield more than those they are to replace, and they must do so repeatedly, season after season. Development of new varieties from the progeny of crosses of high-yielding varieties gives dependable but small increases in yield. The high-yield varieties often are related to each other. But through trial and error, breeders have found that the infusion of new germplasm from unrelated families can give large, stepwise increases in yielding ability. Sometimes the improvement in yield clearly is due to the purposive addition of traits that were lacking in older varieties, such as tolerance to heat and drought or resistance to a prevalent disease. But often the improvement is unpredicted and its basis is unknown. The only constant from the breeders' point of view is that one continually must broaden the germplasm base of any breeding population in order to keep the curve of yielding ability on an upward slope (Duvick, 1984).

Both intuition and scientific knowledge are used in choosing materials for the broadening of the genetic base. It is fairly easy to identify base-broadening materials that will have very little utility for variety improvement, but it is not easy to identify materials with high probability of giving large increases in yielding ability. Some breeders outcross only to elite but unrelated germplasm that is already adapted to the local growing conditions. Others also work with elite germplasm of widely different adaptation, and a few breeders infuse germplasm from seemingly undesirable sources, such as outdated farmer varieties or wild relatives of the crop species.

All of these methods work, but with widely varying rates of success. A rule of thumb is that as the relationship of the base-broadening germplasm decreases so do the odds of producing successful new varieties, but at the same time the odds increase that the infrequent successes will give large advantages in yield.

2. Hybrid Vigor

Box 9

Germplasm from diverse sources can give new kinds of heterosis and thus hybrids with novel improvements, including greater yield and stability of yield.

Biodiversity is essential for maintaining and increasing hybrid vigor. Field crops that are bred and sold as hybrids—such as corn, sorghum (Sorghum bicolor (L.) Moench), and sunflower (Helianthus annuus L.)—depend on the phenomenon of heterosis, or hybrid vigor, for part of their yielding ability. When heterosis is high, crosses between two inbred lines yield substantially more than either parent, even though the inbreds themselves may have been selected for high yield per se. (Inbred lines are uniform products of several generations of self-pollination.) Breeders sort inbred lines into “heterotic groups” based on performance of the inbreds in hybrid combination. Hybrids made by crossing inbreds from contrasting heterotic groups have more hybrid vigor, and in general higher yield, than hybrids made by crossing inbreds within a heterotic group. Breeders search continually for new heterotic groups, hoping to find new cross-group combinations with increased yield. Germplasm from diverse sources also can give new heterotic groups with useful traits that were not found in the earlier heterotic groups.

3. Disease and Insect Resistance

Box 10

Biodiversity both helps and hinders plant breeders as they develop new varieties that have improved resistance to pests and pathogens.

To plant breeders, nothing is more self-evident than their dependence on biodiversity to stay ahead of (or catch up with) continually changing and challenging problems with disease and insect pests. All too often, new biotypes of an insect pest or new races of a destructive fungus will overcome the genetic resistance in cur-
rently-grown varieties and all of their close relatives. Breeders then need to look for new resistance genes in far-away sources, particularly those with high genetic diversity. The broad range of genotypes available in traditional farmer varieties is an especially rich source of resistance genes.

a. Specific Resistance
For many crops such as soybeans and sorghum, numerous individual genes from exotic germplasm in all parts of the world will give high levels of resistance to diseases such as Phytophthora root and stem rot (from Phytophthora sojae; M. J. Kaufman & J. W. Gerdemann) or to insects such as greenbug (Schizaphis graminum Rondani). Such kinds of highly targeted resistance are strikingly effective, but they usually do not last for many seasons, for the pest species typically has a reservoir of genetic variants that can overcome the single genes for resistance. The variants multiply and soon the "resistant" crop varieties are susceptible. Such specific resistance, usually given by one or a few genes of large effect but often with short effective life-time, is sometimes called "vertical resistance" (van der Plank, 1963).

b. Durable Resistance
Heritable pest resistance also can be imparted by complex assemblages of genes, and insects and disease organisms are rarely able to overcome such kinds of resistance. The resistance is longer-lasting, and often is termed "durable" by plant breeders and others concerned with the ravages of disease and insect pests on crop plants. It also has been called "horizontal" resistance in contrast to single-gene "vertical" resistance (Simmonds, 1991; van der Plank, 1963). A disadvantage of durable resistance provided by complex assemblages of many genes is that breeders cannot easily move the assemblages from one plant or one variety to another. The gene assemblages tend to disassociate during the successive segregating generations that are needed to develop a variety, or they tend to bring in associated ("linked") traits from the donor parent that detract from the utility (often the yield) of the elite material that breeders wish to improve. But advantages of durable resistance are so great that many breeders have devoted much time and energy to incorporating the best gene combinations into their breeding stocks (Simmonds, 1993). Very often the best sources of multi-gene durable resistance are unadapted exotics such as farmer varieties and wild or weedy relatives of crop species. In these cases problems with linkage to undesirable traits are accentuated. Thus biodiversity both helps and hinders the efforts of breeders to impart durable pest and pathogen resistance to new varieties of crop plants.

4. Broad Adaptation

Theorists disagree about the utility of broad adaptation and its relationship to biodiversity. If a single variety fares well over a broad geographical range such as the full expanse of the U.S. Cornbelt, it offers more opportunity for the large-scale multiplication of disease or insect pests that can overcome its defenses. Large-scale plantings of a single variety increase the chances of genetic vulnerability.

On the other hand, it is self-evident that varieties could not have broad adaptation unless they contained multiple kinds of effective genetic defenses, able to counter the multiplicity of environmental and biological constraints of a wide range of localities. Broadly adapted varieties would appear to have greater internal biodiversity, that is, greater diversity of useful genetic systems in each plant.

Breeders say that varieties with maximum amounts of tolerance to the full range of environmental and biological stresses that could be expected in any specific locality are, therefore, broadly adapted as well. Selection for stability of performance over a variety of seasons in a single locality tends to produce varieties with "pre-adaptation" to a variety of localities, and such varieties often become popular in a much larger region than had been envisioned. It then is up to farmers and seed companies to decide how widely such varieties will be planted, and at what ratios in comparison to genetically different varieties with (perhaps) lower yields or lower stability of performance. Once again, the urge to maximize yields conflicts with the need to optimize biodiversity.

Breeders also agree that diverse germplasm sources are most likely (after a long and often frustrating period of hybridization and selection) to give varieties with maximum amounts of internal biodiversity, that is, with multiple defense systems that are the basis for broad adaptation. Biodiversity at the varietal level is thereby internalized to biodiversity at the genomic level.
5. New Traits

Box 12

Unlikely sources of diverse germplasm can provide valuable new traits.

In horticultural crops, and increasingly now in field crops, new traits may have great economic value. Almost invariably, breeders must go to genetically diverse sources for such new traits. Wild species related to tomato (*Lycopersicon esculentum* Mill.) contributed genes that enhanced the soluble solids and sugar content of tomatoes, traits that are of large commercial importance for certain uses of tomato. (Figure 1 illustrates the impressive amount of phenotypic diversity that exists among wild and domesticated tomatoes.) At the present time, an alliance of public and private corn-breeding organizations is introgressing tropical maize (*corn*) into elite temperate germplasm with the intention of finding new levels of oil content, or other new grain quality traits with commercial utility. Similarly, glandular trichomes (specialized leaf hairs) from a related wild species promise to give a new kind of insect resistance to potatoes.

6. Pleasant Surprises

Box 13

Linked genes often cause problems but sometimes they provide pleasant surprises.

One of the pleasures of plant breeding is to breed for one trait and discover, fortuitously, that the act of breeding has produced a new entirely unexpected but highly desirable trait. This was the case with cucumber (*Cucumis sativus* L.) when a breeder outcrossed to a variety from China in order to incorporate its virus resistance into adapted germplasm, and discovered that the variety also contributed an entirely new kind of color pattern to the fruit. The new pattern, a uniform fruit color, proved to be highly desirable for commercial use when enhanced with dark green background color. It now is incorporated into virtually all of the commercial cucumber varieties.

In another example, a corn breeder in Tennessee used germplasm from Cuban open-pollinated varieties to improve ear worm (*Heliothis zea* Boddie) resistance in hybrids for use in the Midsouth. An inbred line from that breeding program fortuitously had excellent...
resistance to a new strain of virus that swept across the southeastern United States. The resistance probably was derived from the Cuban germplasm. The inbred became the basis for a series of successful virus-resistant hybrids that were adapted to the Southeast. These examples demonstrate that gene linkage, which is often a problem, sometimes can provide pleasant surprises for plant breeders and the farmers they serve.

7. Transgenes and Key Biological Functions

In a sense, biotechnology has closed the circle of biodiversity as described by E. O. Wilson. The search for new kinds of useful biodiversity now can encompass “the totality of hereditary variation in life forms, across all levels of biological organization.”

The advent of biotechnology in the aid of plant breeding led some to say that plant breeders would no longer depend on diversity within the crop species for new advances. It was believed that genetic transformation (transfer of genes without benefit of sexual hybridization) could bring in powerful new genes as needed from any one of the biological kingdoms. Naturally occurring genetic diversity as found within a crop species would no longer be needed, or at least would become secondary in importance and use.

However, after 20 years of experience it now appears that biotechnology may have increased rather than decreased the value of biodiversity within a crop species and its close relatives. Although some genes of great promise have been moved from far afield into crop plant varieties (e.g., inserting insect resistance genes from bacteria to maize) and more will soon follow, the total number is not large.

On the other hand, genetic transformation has enabled biologists to learn more than was ever thought possible about how genes are regulated. Molecular biologists are learning how to increase or decrease the rate of formation of gene products (such as enzymes), and to control the time and place at which a gene is turned on. They increasingly are able to identify most of the genes in a crop species and are beginning to identify their functions. The goal is to find genes and gene systems of key importance in controlling important traits such as tolerance to heat and drought, disease and insect resistance, or time of flowering, and then to enhance their operation.

A great aid to acquiring this new knowledge is the ability to search through the genomes of a genetically diverse assemblage of local crop varieties, then to look farther afield to varieties adapted to other parts of the globe, then to wild or weedy relatives of the crop species, and eventually to unrelated species and even organisms from other biological kingdoms. Comparisons of like genes or like genetic systems at each level add incrementally to understandings of how the assemblages of genes function in a particular crop and give hints about how to make them function better. Biotechnology gives plant breeders the ability to bring in foreign genes from any part of the world of nature, but more importantly it helps them to improve the action of native genes in their original settings, based on insights gained from the study of similar genes in other organisms. In a sense then, biotechnology has closed the circle of biodiversity as described by E. O. Wilson by allowing plant breeders to manipulate “the totality of hereditary variation in life forms, across all levels of biological organization.” Biodiversity can contribute to agriculture more strongly than ever, thanks to biotechnology.

8. New Cultural Systems

At the other end of the spectrum from molecular biology is the biology of assemblages of whole plants and animals, as well as microbes, and how they interact with their environment. Although some people believe that the term “ecology” first described this holistic concept, the fact is that the term “farming” in whatever language has a precedence of about 10,000 years as a descriptor of at least one class of ecological systems. Farming is all about the biology of assemblages of plants, animals (including humans), and microbes and how they interact with their environment.

The first farmers also tested and developed ways to grow different crop varieties in combination with each other. Although their primary aim was to simplify farming and grow only the crops desired, many kinds of cropping systems were designed, each intended to give satisfactory food production for a particular locale and
life-style. We have inherited the results of those experiments, the failures and successes of these first farmers. Some of the methods still work for present needs, some have been reduced in importance or even largely abandoned, and others have been adapted in ways that disguise their origins.

a. Rotations

Crop rotations were essential before the advent of chemical-intensive agriculture. Farmers used rotations to “rest the soil,” to help in weed control, and, perhaps without clear understanding, to help control disease and insect problems. Biodiversity among crop species had obvious utility to the early farmers. Crop varieties were selected to fit rotations, and rotations were designed to fit the crop varieties.

Crop agriculture in the United States still uses rotations, but often they are greatly simplified, as in the common corn/soybean two-year rotation in the American Midwest. The purpose of rotations is still the same, namely, to enable successive crops to benefit from those that preceded them and, in turn, to benefit those that follow. Soybeans help corn to escape damage from the corn rootworm beetle (Diabrotica spp.) by providing a year in which the field is not infested with rootworm eggs and subsequent root-chewing larvae. In this way, farmers can avoid the use of chemical pesticides to control corn rootworm.

But biodiversity among populations of rootworm beetles has enabled some of them to wait an extra season before hatching. They emerge just in time for the corn crop following the soybean rotation. Other populations have developed the ability to feed and lay eggs in soybean fields, and they, too, produce larvae just in time for the next year’s corn crop. Farmers, plant breeders, and entomologists now are working to devise new rotational schemes or new kinds of plant varieties that might put the rootworm beetles at a new disadvantage. Biodiversity once again both helps and hinders agriculture, depending in this case on whether it is in the crop or in the pest. To use biodiversity to profit humans (and not rootworms) requires keen biological insights and knowledge of ecological interactions at many levels. Use of rotations to provide beneficial biodiversity is no simple matter.

There can be no mistake, however, that the utilization of specifically designed rotations adapted to today’s technology and economy can be cost saving and chemical saving, while also maintaining or increasing the yields and quality of the product. The principle works worldwide. In India’s Punjab State, for example, rice (Oryza sativa L.) yields following potato and maize were 36% higher than when rice followed wheat, a typical rotation in the Punjab.

b. Polycultures

Subsistence farmers, especially in the tropics, use polycultures routinely; that is, they grow mixtures of crops planted together in various patterns. The mixed crop plants complement each other by varying the time at which they draw on water or nutrients from the soil, or by adding beneficial nutrients for use by other crops (e.g., nitrogen-fixing beans (Phaseolus vulgaris L.) planted with corn), or by providing needed shade to other crops. Farmers in medieval Europe often grew mixtures of wheat and rye (“meitéll”) as insurance against the failure of wheat, which was the more delicate but more preferred of the two crops (Neveux et al., 1979).

Some researchers are experimenting with polycultures for use in modern commercial farming. They intend to produce special types of grain crops that can be grown in mixtures. Such polycultures could reduce the need for chemical fertilizer, increase the stability of production, provide non-chemical protection against disease and insect pests, and reduce soil erosion if the polycultures are composed of perennial plants. One group of researchers uses the tallgrass prairie ecosystem as a model. They intend to copy the salient features that give it stability, productivity, and, not least, the capacity for reducing soil erosion (Soule and Piper, 1992). The polyculture breeders intend to develop crop mixtures with useful degrees of intra- and inter-species diversity that also produce crop products that will meet standards of yield and quality for the needs of today’s farmers and the consuming public. Researchers agree that development of polyculture crops presents large challenges as compared to making improvements to one crop at a time. To simultaneously select several crops for individual and mutual enhancement when grown in one polyculture requires analyses that will test skills in statistics as well as in agronomy and plant physiology. But simply to attempt such a project will give benefits in knowledge that could well exceed the value of the final product.

For example, plant-rows (each plant-row is a monoculture of seed from a single plant) of wild species have shown great variability in resistance to certain fungal diseases. Some plant-rows show no symptoms whereas others are highly susceptible. This behavior contrasts with the fact that individual plants of the species show no or little damage in normal prairie settings. It would seem that protection afforded by biodiversity at the species level (both within and among species) was lost
when individual genotypes were planted as single-row monocultures. An alternative explanation is that disease-tolerant plants are hard to spot in natural prairie settings, and plant-rows simply make it easier to categorize the variation that exists in the wild.

The plant-row problems with disease testify to the great advances that had to be made by the first farmers when they put out similar isolated plantings of a single species in the first stages of domestication. The crop was deprived of intra- and inter-species protection and the farmers had to select (even if without special intent) for genetic systems that were powerful enough to give resistance without this protection.

c. Multilines

Modern science has shown that carefully planned mixtures or blends of different selections of the same variety, differing only in major genes for disease resistance, can provide reasonably durable protection against diseases. In oats (Avena sativa L.), for example, a series of resistance genes is backcrossed into a single variety, producing numerous strains, each with a different resistance gene. Each gene carries resistance to a specific race of a fungal disease such as crown rust (Puccinia coronata Cord.), and each gene carries a different kind of resistance. The several strains are then mixed and sown together as one crop, apparently homogeneous but in actuality genetically heterogeneous for resistance to the disease. Such a mixture is called a “multiline.” As noted, its diversity of resistance genes will cover all known races of the disease, and essentially provides “genetic diversity in place.” But as new races and new genes conferring resistance to them are found, new backcross strains of the oat variety can be produced and added to the original, thus breeders also can take advantage of “genetic diversity in time.”

III. CONCLUSIONS

Plant breeding—the development of plant varieties—simultaneously exploits and enhances biological diversity.

Plant breeding exploits biodiversity. Modern plant breeding could not exist without recourse to a continuing supply of biologically diverse populations at the variety, species, and family levels, and now (thanks to biotechnology) at any level in the world of nature.

Plant breeding enhances biodiversity. Plant breeders of all kinds—full-time professionals, farmer-breeders, and dedicated amateurs—add to the stock of genetically diverse organisms by the continual production of new and genetically diverse plant varieties with new, ecologically diverse adaptations. Such increased biodiversity is minuscule when compared to that existing in relatively pristine ecosystems, but it is essential for productive and stable agricultural production.

Yet plant breeding has also caused problems when breeders and farmers ignore or misunderstand the ways in which biodiversity contributes to ecological balance and crop productivity. Excessive dependence on simple solutions such as single-gene disease resistance has led to problems.

Plant breeding has given the greatest benefit when its products and their users took advantage of the beneficial interactions that accrue among diverse organisms at each level of complexity from individual genes to the landscape. Much remains to be learned about ways to make the greatest use of biodiversity at each level (e.g., spatial, temporal, or reserve), about how one kind can substitute for another, and when it is best not to substitute. Plant breeding epitomizes the duality of humanity’s interaction with biodiversity. We wish to alter and enhance it for the benefit of human needs and wants, but we also must avoid altering or depleting it in ways that destroy its benefits to us and to the world of nature—our home.

See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • AGRICULTURE, SUSTAINABLE • AGRICULTURE, TRADITIONAL • BREEDING OF ANIMALS • GENETIC DIVERSITY

Bibliography


I. OVERVIEW OF BUTTERFLY TAXONOMIC DIVERSITY

Although butterflies may be the best known group of insects, our understanding of their taxonomic diversity has two fundamental weaknesses. The first regards the recent decline of professional butterfly taxonomists and species level revisions in the past 50 years. This requires most species diversity estimates (from family to genus) to be derived from a literature that is out-of-date. The second is that the number of families and subfamilies of butterflies varies among classifications because, despite current interest in phylogenetic systematic methods and analyses, the relationships within the major groups are...
unresolved, particularly within Nymphalidae and Lycaenidae. Such is the state of butterfly taxonomy. Nevertheless, the "true" butterflies (superfamily Papilionoidea) may be placed conservatively into four families (Papilionidae, Pieridae, Nymphalidae, and Lycaenidae) that together include between 12,900 and 15,819 species. One useful framework for organizing butterfly taxonomic diversity is P. R. Ackery's (1984) synthesis of butterfly classification, which forms the basis of the following synopsis:

Family Papilionidae (Swallowtails, Fig. 1): a group of 500–600 species in three subfamilies, distributed worldwide but with most species being tropical; adults medium to large sized, both sexes have six walking legs bearing nonbifid tarsal claws; most are brightly colored, may be Batesian or Mullerian mimics, all feed on flower nectar, and males drink at wet soil (Fig. 1); caterpillars are herbivores, body smooth without hardened spines, and possess extrusible glands (osmeteria) that are unique among butterflies; pupa typically with a silk girdle at third thoracic segment.

Subfamily Baroniinae: one species (Baronia brevicornis) endemic to Mexico; caterpillars unique among the family by feeding on Fabaceae.

Subfamily Parnassinae: 40–50 species, mainly in north temperate mountains; host plant families include Crassulaceae, Fumariaceae, and Zygophyllaceae.

Subfamily Papilioninae (Fig. 1): more than 500 species placed into three or more tribes; distributed worldwide; most species are tropical. Many species with long hindwing tails; many with sexes strongly dimorphic; many involved in mimicry; species range from palatable mimics (Papilio and Eurytides) to unpal-
butterflies; Family Pieridae (whites, sulphurs, jezebels, cabbage butterflies: (e.g., Colias, Phoebis, Catopsilia, Aphrissa), and have vibratory papillae as sound producing organs when stimulated. Some are Batesian mimics of aposematic nymphalids and papilionids; some species undergo spectacular mass migrations (e.g., Delias, Mylothris, Pereute). Interspecies mimicry may be the original example from which Batesian mimicry theory was derived; host plants are in the Fabaceae.

Subfamily Pierinae (whites, jezebel, and cabbage butterflies): approximately 700 mostly tropical species, especially in Africa and Asia; some species have spectacularly bright, contrasting colors of white, red, and black (e.g., Delias, Mylothris, Preceata, and allies) suggesting unpalatability and warning coloration; some groups with dimorphic sexes in which females resemble unpalatable nymphalids and papilionids; some species undergo spectacular mass migrations (e.g., Astria, Appias, and Belenois), and a few species are crop pests (Pieris); host plant families include Capparidaceae, Brassicaceae, Santalaceae, and Loranthaceae.

Subfamily Coliadinae (sulphurs): 400–600 species well represented in temperate and subtropical regions; generally yellow or white with short, thickly scaled antennae; common in open areas, some migrate in large numbers and have been recorded out at sea far from land (Cobus, Phoebis, Catopsilia, Aphrissa, and Eurema); many species with patterns visible only in the ultraviolet.

Family Lycaenidae (hairstreaks, blues, coppers, and metal marks: Fig. 1): a group of 6000–6500 mostly tropical species in 10 subfamilies (the total number may change with future study), and accounts for nearly 50% of all butterflies; most species are small to very small, both sexes have six walking legs (except Riodinia), and frequently with alternating black and white bands on the antennae, the group displays a tremendous diversity of form, color, and life histories, adults feed on flower nectar, fruits, carrion, and honeydew, and a few species do not feed as adults; some African and Southeast Asian groups are involved in mimicry complexes; caterpillars most often slug-like, without hard spines or projections, but the Riodinidae shows an extensive variety of form, as a group, lycaenids have the widest diet breadth of all butterflies, and depending on the group caterpillars may feed on plants, other insects, or insect secretions; many caterpillars form intimate and complex symbiotic associations with ants and produce acoustical calls similar to ant calls, secretions that are harvested by ants, and chemicals that alter ant behaviors; pupae are typically round, seed-shaped, unadorned with projections, and may produce clicking or whirring sounds when stimulated.

Subfamily Riodinidae (metalmarks): 1200–1400 species that are almost exclusively Neotropical; arguably the most diverse of all lycaenoid groups with respect to adult and larval forms, and they are often treated as a distinct family (Riodinidae) divided into five subfamilies; adults feed primarily on flower nectar, some drink at wet soil, others at carrion; overall their life histories are poorly known, but as a group riodinid caterpillars appear to be mainly herbivores (Easelia, Meosemia, Anyclurus, Metacharis, Mesene, Symmachia, Emeis, Anteros, and Helicops), with a few carnivorous species on Homoptera (Alesa and Seribat); about one-third of the riodinids form symbioses with ants and have vibratory papillae as sound producing organs (Thrisbe, Audre, Lemonias, Synargis, Nymphidium, and Seribat (Fig. 21), riodinid myrmecophily is entirely Neotropical; host plant families include Araceae, Asteraceae, Bromeliaceae, Bombacaceae, Cactaceae, Clusiaceae, Dilleniaceae, Euphorbiaceae, Fabaceae, Lecythidaceae, Loranthaceae, Malpighiaceae, Marantaceae, Melastomataceae, Myrtaceae, Orchidaceae, Rubiaceae, Sapindaceae, and Zingiberaceae plus mosses, liverworts, and lichens.

Subfamily Stygininae: one enigmatic species (Styx in-fernalis) that is endemic to the Peruvian high Andes; the early stage biology is unknown.

Subfamily Lippeninae: a small African group of 30–40 species; some are Batesian mimics of aposematic nymphalids (Mimacraea and Ornitholuides); caterpillars feed on lichens and microscopic fungi (Durbania and Delenea).

Subfamily Eueomginae: a small group restricted to the Oriental regions; caterpillars may be gregarious, feed...
FIGURE 2. Butterfly caterpillars. Clockwise from upper left: (a) final instar caterpillar of *Morpho achilles* (Nymphalidae: Morphinae); (b) final instar caterpillar of *Tigridia acesta* (Nymphalidae: Nymphalinae); (c) Four frames showing details of the call production mechanism of *Thisbe irenea* (Lycaenidae: Riodininae). Upper left—top of caterpillar head, a vibratory papilla (arrow) and anterior margin of first thoracic segment. Upper right—granulations on top of caterpillar head. Lower left—detail of vibratory papilla. Lower right—head granulations at high magnification. (d) Final instar caterpillar of *Thisbe irenea* (Lycaenidae: Riodininae) being tended by ants. [All photos copyright P. J. DeVries.]

Subfamily Liphyrinae: a small group found in Africa, Australia, and Asia; adults have the proboscis partly or entirely atrophied; depending on the species, caterpillars feed entirely on insects (Homoptera) or on ant brood within ant nests (Liphyra and Aslauga).

Subfamily Miletinae: a moderate-sized group with most species in tropical Africa and the Orient and one in North America; caterpillars mainly feed on nymphs of Homoptera (*Feniseca*, *Spalgis*, and *Allotinus*), with a few species feeding on secretions of Homoptera (*Allotinus*) or ant regurgitations (*Thestor*); adults feed mainly on honeydew secretions of Homoptera.

Subfamily Curetinae: approximately 40 species restricted mainly to tropical Asia, all in the genus *Curetis*; caterpillars with tentacle organs as conspicuous, rigid cylindrical tubes; caterpillars feed on Fabaceae and have loose associations with ants.

Subfamily Theclinae (hairstreaks; Fig. 1): this very large and diverse group is found worldwide, but most species are tropical; caterpillars are mainly herbivores, some carnivores on Homoptera, and many associate with ants, but the life histories of most species remain unknown; host plant families include Anacardiaceae, Annonaceae, Asteraceae, Bromeliaceae, Chusiaeae, Cy- cadaceae, Euphorbiaceae, Fabaceae, Fagaceae, Gerania ceae, Lecythidaceae, Malpighiaceae, Malvaceae, Melastomataceae, Meliaceae, Myrtaceae, Oleaceae, Orchidaceae, Sapindaceae, Sapotaceae, Solanaceae, Sterculiaceae, Ulmaceae, and Verbenaceae.

Subfamily Lycaeninae (coppers): a group of 20–40 species found mainly in temperate regions, a few are tropical; caterpillars are herbivores on Polygonaceae, and some form loose associations with ants.

Subfamily Polyommatinae (blues): a large group with a worldwide distribution, most are pale, reflective blue above; caterpillars appear to always associate with ants, and their diets range from herbivores to carnivores, and some (*Maculinea*) feed on ant larvae and are clearly complex parasites and predators within ant nests; host plant families include Crassulaceae, Euphor biaceae, Fabaceae, Lamiaceae, Myrsinaceae, Oxalida ceae, Primulaceae, Rhamnaceae, Rutaceae, Santalaceae, Sapindaceae, Saxifragaceae, Selaginaceae, Sterculiaceae, Verbenaceae, Zingiberales, and Zygophyllaceae.

Family Nymphalidae (brush-footed butterflies; Figs. 1 and 2): a group of 4800–6000 species in 14 subfami-
lies embracing a prodigious variety of forms and sizes, with both sexes having four walking legs; the forelegs are greatly reduced (hence "brush-footed"); adults may be dull brown (Satyrinae and Nymphalinae), brightly colored (Nymphalidae and Brassoilinae), brilliantly iridescent due to physical properties of the wing scales (Morphininae and Apaturinae), or transparent (Ithomiinae and Satyrinae); some groups are entirely palatable, others highly distasteful, and some (Danainae, Ithomiinae, Heliconiinae, and Acraeinae) are extremely important mimetic models; some species (Charaxinae and Nymphalinae) mainly inhabit tropical forest canopies; adults may feed on flower nectar, pollen, rotting fruits, carrion, or do not feed at all; the caterpillars may bear many spines on the body, some also have head spines, whereas others are devoid of spines (Fig. 2); caterpillars are entirely herbivorous, and the particular groups exhibit strong associations with particular plant families; the pupae are typically suspended.

Subfamily Charaxinae: a group of 350–450 mainly tropical species; adults fly very fast with robust bodies; underside of wings typically camouflaged and leaf-like, some with brilliantly colored upperside (Agrias, Prepona, and Charaxes); all are palatable, with few Batesian mimics (Consal and Euxanthe); adults feed primarily on juices of rotting fruit, dung, and/or carrion (rarely flower nectar), and most inhabit the forest canopy; caterpillars have smooth bodies, often bearing a corona of head spines or projections; host plant families include Annonaceae, Celastraceae, Convolvulaceae, Euphorbiaceae, Fabaceae, Flacourtiaceae, Lauraceae, Myrtaceae, and Sapindaceae.

Subfamily Apaturinae: a small group of medium to large species that are mainly tropical (often placed within the Nymphalinae); males often with brilliant iridescence on wing upperside, the Neotropical species (Doxocopa) with proboscis and forelegs lime-green; adults are strong fliers and feed entirely on carrion and putrefying fruits; caterpillars are herbivores on Ulmaceae and have a smooth body, bifid tail, and head with a pair of strong horns.

Subfamily Satyrinae (saturys, wood nymphs, and browns): a cosmopolitan group of 1500–2000 mainly tropical species that are generally dull brown (some notable exceptions) with well-developed eyespots on the wings, most feed on juices of rotting fruits, some on fungi, and some on flower nectar in temperate regions; all are palatable, with only few clear examples of Batesian mimicry from Asia (Penthesia, Zephyra, and Elymnias); caterpillars are smooth with bifid tails, and some bear paired head projections; host plant families include Areaceae, Araceae, Cyperaceae, Heliconiaceae, Poaceae, and Selaginellaceae.

Subfamily Brassolininae (owls): 70–80 entirely Neotropical species ranging from medium to some of the largest butterflies known (Calligo) that fly at dawn and dusk, most with characteristic, large eyespots on the hindwing undersides, and they are common in butterfly conservatories; most species feed entirely on rotting fruit juices, but a few with a strongly reduced proboscis (Brassolis and Dynastor) may not feed; caterpillars have smooth bodies, often with dorsal pseudospines, bifid tails, and multiple horn-like projections on the head; host plant families include Areaceae, Bromeliaceae, Heliconiaceae, Musaceae, and Poaceae, and they can be pests in banana and palm plantations.

Subfamily Amathusiinae (fauns and duffers): an Indo-Australian group of approximately 80 species (sometimes placed in Brassolinae or Morphinae); medium to large butterflies that fly at dawn and dusk and feed on rotted fruit juices; the group is palatable (except for perhaps Varanavis) and shows little or no mimicry; caterpillars have smooth bodies, some with long, downy setae, bifid tails, often paired head horns; host plant families include Areaceae, Musaceae, Poaceae, and Smilacaceae.

Subfamily Morphinae (morps; Fig. 2): a group of 40–50 Neotropical species of medium to very large butterflies; the large reflective blue species (Morphe) are immediately noticeably in nature and in butterfly conservatories, but others that fly in the forest understory (Antirrhea and Caerottis) are seldom observed; all feed on rotting fruits, none are considered to be unpalatable, and none are mimics; a favorite of collectors and butterfly conservatories, surprisingly little is known of their natural history; caterpillars are covered with red, yellow, and green patterns, bear tufts of dorsal and lateral setae, possess bifid tails, and bear short, paired projections on the head (Fig. 2); host plant families include Areaceae, Brighamiae, Fabaceae, Menispermaceae, Poaceae, and Sapindaceae.

Subfamily Caligninae: a group represented by two to five species in the genus Caligna restricted to the Himalayan regions; it is apparent that Caligna forms mimicry complexes with Paranta (Danainae), but it is not clear if it is Batesian or Mullerian mimicry; caterpillars are smooth with short bifid tails and stout head horns and feed on Moraceae.

Subfamily Nymphalinae: a diverse, cosmopolitan group of more than 3000 species (sometimes split into Limenitinae and Nymphalinae) containing a tremendous range of size and color patterns; some are migratory (Vanessa, Eunica, and Sallya), some are unpalatable
models, some are palatable mimics, some show tremendous seasonal polymorphism (Janoua); some feed on flower nectar, others feed on rotting fruits and carrion; some groups inhabit tropical forest canopies (Euphaedra, Cymothoe, Panacea, Euphile, and Baeotus); some pass north temperate zone winters as adults (Nymphalis); caterpillars usually covered in spines, many with well-developed pairs of head spines (Fig. 2); pupa often with buld projections on head; host plant families include Acanthaceae, Caprifoliaceae, Convolvulaceae, Euphorbiaceae, Fabaceae, Flacourtiaceae, Lamiaceae, Loranthaceae, Moraceae, Plantaginaceae, Poaceae, Rubiaceae, Rutaceae, Salicaceae, Sapindaceae, Scrophulariaceae, Urticaceae, and Verbenaceae.

Subfamily Acræinae: a group of approximately 250 tropical African species, with a few in Asia and the neotropics; ranging from small to large, all are slow-flying unpalatable species that contain cyanogenic compounds, forming aposematic models for a variety of other groups, and are involved in Mullerian mimicry (Acræa, Rematixis, and Actinote); one species (Acræa encledon) is almost entirely female and reproduces via parthenogenesis; caterpillars (many which feed gregariously in communal nests) are densely covered in spines, but lack head spines; host plant families include Asteraceae, Passifloraceae, Sterculiaceae, Tiliaceae, and Urticaceae.

Subfamily Heliconiinae (passionflower butterflies, tigers, and crowns): a cosmopolitan group of approximately 150 species, with most in tropical Africa and Asia; ranging from small to large size, most are slow flying, conspicuously colored, and models in mimicry complexes; they feed on flower nectar, and some on pollen (Heliconius and Laparus) from which they can derive some of their unpalatability; caterpillars are densely spiny, bear paired head spines, and feed entirely upon plants in the Passifloraceae and allies—hence the name passionflower butterflies.

Subfamily Danainae (milkweed butterflies, tigers, and crowns): a cosmopolitan group of approximately 150 species, with most in tropical Africa and Asia; ranging from medium to large size, most are slow flying, conspicuously colored, and models in mimicry complexes (Danaus, Amuris, Idea, Tirumala, and Parantica); their distasteful nature derives from feeding as caterpillars on milkweeds, but they also acquire other chemical defenses by feeding as adults on flowers and wounds in certain plants containing particular alkaloids; males have well-developed, brush-like scent hairs they use during courtship; caterpillars are smooth, often patterned with alternating bands of black, white, and yellow, and many have dorsal pairs of fleshy tubercles; the pupa are frequently reflective gold or silver; host plant families include Apocynaceae, Asclepiadaceae, and Moraceae.

Subfamily Ithomiinae (ithomines and glass wings; Fig. 1): approximately 300 entirely Neotropical species; adults typically with a very small head, elongate wings, and species vary in color from transparent (Creia, Ithomia, and Pteronymia) to bright tiger-striped patterns (Mechanitis, Tithorea, and Melinaea) (see Fig. 1); they are involved in both Mullerian and Batesian mimicry, representing unpalatable models for many other groups; their unpalatable properties derive from larval host plants and chemicals acquired by adult feeding, and males typically possess a tuft of scent hairs between the wings that disseminate pheromones; caterpillars are smooth, often with fleshy tubercles, and may be brightly colored or cryptic; The pupa is often reflective gold or silver and squat; host plant families include Apocynaceae, Gesneriaceae, and Solanaceae.

Subfamily Tellervinae: a group of 6–10 Australasian species all in the genus Tellervo; they serve as a models for nymphaline and satyrine Batesian mimics; this group has been included within the Ithomiinae; the caterpillars resemble some Danainae and feed on Apocynaceae.

Subfamily Linythritaenae (beaks and snout butterflies): a cosmopolitan group of approximately 10 species recognized by the large erect palpi that form a “snout”; most species are well-known to periodically undergo spectacular migrations; the caterpillars somewhat resemble those of the Pieridae, and all feed on Ulmaceae.

II. EARLY STAGES AND HOST RELATIONSHIPS

Like all members of Lepidoptera, butterflies have four discrete stages in the life cycle (egg, caterpillar, pupa, and the adult), each with particular characteristics, behaviors, and requirements. Furthermore, to complete their life cycle, butterflies require a plant or insect host to feed on.

A. Egg

Butterfly eggs are laid either singly or in small to large clusters, either on or off the host, and the location where the egg is laid is typically important (Fig. 2). The eggshell frequently has an elaborate sculpturing that plays a role in respiration, and each major group of butterfly has its own form of egg.
B. Caterpillars

Depending on the group, the appearance of butterfly caterpillars may range from cryptic to aposematic, they may be covered with spines and/or hairs or appear to be naked, and their diets may include plant tissue or the flesh of other insects or they may feed entirely on secretions produced by other insects (Fig. 2). All caterpillars consist of three major sections: the head, thorax, and abdomen. The hardened head houses mandibles that function to shear off bits from their food. The head bears a gland that lays down silk that is grasped as the caterpillar moves forward, helping the caterpillar grip the subt Rate and also to secure rolled leaves in which they may shelter. Each thoracic segment bears a pair of true legs, whereas the 10 abdominal segments form the bulk of the body, housing the long gut. Caterpillars walk by using the prolegs (segments 3–6 and segment 10), and these function by hydraulic pressure and muscles. Segments 1–8 bear the external ring-like orifices (spiracles) that allow gas exchange with the atmosphere.

After reaching a particular size each caterpillar instar stops eating and undergoes a molt to the next instar; this is how caterpillars grow. Generally, but not always, there are five larval instars followed by a molt to the pupa, or chrysalis. Caterpillar growth is not so steady an increase in weight from first to final instar, rather, there is a dramatic fluctuation of weight between each molt. Newly molted caterpillars may weigh about the same as or even less than previous instars, but their weight will quickly increase and exceed that of previous instars. Caterpillar feeding behavior often differs among instars. In many groups, late-instar caterpillars may stop feeding during the day and then feed entirely at night. Alternatively, some lycaenid caterpillars may start life feeding on plants and, after molting to a later instar, they fall off the plant, are picked up by ants, and are carried into their nest where the caterpillars feed as carnivores on ant brood.

C. Pupa

When the final-instar caterpillar is fully grown, it stops eating and enters its molt to the pupa, or chrysalis. Within the pupa the caterpillar tissues are reconstructed into the adult by the process known as metamorphosis. The pupa attaches to a substrate by a series of hooks on the last segment (cremaster), and major groups typically have characteristic manners of pupation. For example, pupating with the head downward attached only by the cremaster is typical of Nymphalidae, but pupae of Papilionidae and some Pieridae attach by the cremaster with head upward and are restrained by a silk girdle. Pupae of Lycaenidae produce a whirring, clicking, or buzzing sound with a rasp and file system on the abdomen that may be a defense against predators. Sound production is also known in some nymphalid pupae. When mature, the pupa splits along its dorsal surface, and the adult ecloses. After eclosion, the adult normally hangs from the pupal shell or nearby with the wings suspended downward so that it can expand and dry its wings. If dried in a crumpled manner the wings are useless for flying, and the butterfly is effectively dead. As a rule, female butterflies are mated soon after or even before eclosion, exemplifying one of the most potent laws of evolution—nature abhors a virgin.

D. Adult

The adult butterfly (Fig. 1) is incapable of additional growth but is capable of flight, mating, and reproduction. Like all insects, the butterfly body is composed of the head, thorax, and abdomen. The most obvious features of the butterfly head are the large compound eyes composed of numerous facets (ommatidia) that cannot focus but are sensitive to movement, light, and certain colors. A pair of distally thickened antennae arise from between the eyes that vary in shape according to the group. The antennae function as sensory organs for finding food, mating, and balance during flight and are sensitive to volatile chemicals. Between the eyes there is a pair of appendages called labial palps, and between them lie the proboscis, a hollow tube composed of two interlocking halves that is coiled like a watch spring when not in use and can be extended for feeding. By virtue of having a proboscis, butterflies are restricted to a liquid diet that may include flower nectar, the juices of rotting fruit, carrion, dung, or semidigested pollen. Proboscis length may vary according to the group; in some species it is nearly vestigial, thus precluding feeding as adults (Brassolis and Liphyra), whereas in others the proboscis measures more than 1.5 times the length of the body (Eurybia), allowing them to take nectar from a wide range of flowers. The thorax is composed of three fused segments bearing the wings and legs, and it contains the muscles for locomotion and various internal organs. As in all insects, the adult butterfly has six legs, one pair per segment. Butterflies have four wings (two forewings and two hindwings) typically covered in scales that give butterflies their characteristic colors and patterns. The color patterns of butterflies result mainly from
the covering of scales that are arranged like overlapping roof tiles. There are three notable types of scales. The pigmented scales are colored by the deposition of melanin, pterins, or other chemicals. Structural scales generate blue, violet, copper, or green colors by reflecting particular wavelengths of incident light. Androconial scales store and disseminate chemical odors (pheromones) that are used in mating, and some of these scales may be physically transferred to the female during mating.

The abdomen houses the digestive and reproductive tracts, terminating in the reproductive organs (genitalia). Except for segments housing the genitalia, the abdomen can stretch when the gut becomes filled with food, and abdominal distention may be considerable in groups feeding on rotting fruits (Charaxinae). The penultimate abdominal segment of males bears two appendages (claspers) that open to expose the aedeagus (penis) and serve to grip the female's abdomen during mating. The female abdomen terminates in three openings: the anus, egg pore, and copulatory pore. The configuration of genitalia is used extensively in butterfly taxonomy.

E. Host Relationships

An important aspect in the butterfly life cycle is the ability of ovipositing females to find, and caterpillars to feed on, particular plants. The liaison with plants is so strong that many groups of butterflies only associate with particular taxonomic groups of plants; other plants are unacceptable to both caterpillars and ovipositing females. In the Lycaenidae this association may extend to particular species of ants or Homoptera. For example, caterpillars of milkweed butterflies (Danaineae) only feed on plant families containing milky latex, and those of the Heliconiinae use plants in the Passifloraceae. Such patterns of host association in butterflies gave rise to Ehrlich and Raven's classic paper that developed the concept of coevolution. On the whole, host association records for the Papilionidae, Pieridae, and Nymphalidae are much more complete than for the Lycaenidae.

III. BUTTERFLY–ANT SYMBIOSES

The ability to form symbiotic associations with ants (myrmecophily) occurs only within the Lycaenidae. Here, caterpillars provide food secretions to ants in exchange for protection against insect predators such as social and parasitic wasps. To form these symbioses, caterpillars have a suite of unique adaptations that may include organs (collectively known as ant-organs) to produce food secretions, volatile chemicals, and sound, all of which work in concert to modify ant behavior and enhance the protective attitude of ants toward caterpillars. Recent studies on ant-organs indicate that myrmecophily evolved at least twice in the butterflies: once in the Riodininae and independently in other lycaenid subfamilies.

The widespread trait of myrmecophily within the Lycaenidae and the fact that lycaenids account for approximately 50% of all butterfly species led Pierce (1984) to suggest that myrmecophily has amplified speciation rates in this group. Indeed, the diversity of life histories in myrmecophilous butterflies can be exceedingly complex, encompassing herbivores, carnivores, and those that feed as caterpillars only on secretions, and the associations with ants range from mutualistic to completely parasitic or predatory.

A. Food Secretions

Ants pay close attention to particular abdominal segments bearing ant-organs that produce food secretions, which in some species are known to have high concentrations of amino acids and sugars. In some Riodininae, these consist of a pair of organs (tentacle nectary organs) on the eighth abdominal segment that can be extruded individually or simultaneously. In all other lycaenid subfamilies, this organ is a single dorsal pore on segment 7 (dorsal nectary organ). Ants are so intent on obtaining the secretions that they constantly antennate the caterpillar to solicit more, and in many cases, this is a good example of a general rule among participants in symbiotic associations—’you scratch my back and I’ll scratch yours.’

B. Semiochemical Production

Myrmecophilous caterpillars may have extrusable glands that seem to produce volatile chemicals (semiochemicals or pheromones) that alter the behaviors of attending ants. Some Riodininae have a pair of extrusable glands on the third thoracic (anterior tentacle organs), whereas other lycaenids have a pair of glands on the eighth abdominal segment (tentacle organs). In both cases, when extruded from the body these organs do not produce a liquid secretion but rather the tip is modified with spines that gives the appearance of a tiny feather duster. These spines likely provide a larger surface area to disseminate volatile chemicals that may be similar to ant alarm pheromones. Instead
of anterior tentacle organs, a small group of Riodininae caterpillars (Theope) possess a corona of inflated setae around the head that appear to disseminate semiochemicals.

C. Call Production
The idea that caterpillars produce acoustic calls might seem unlikely. We now know, however, that myrmecophilous caterpillars produce substrate-borne calls that function in the formation and enhancement of their symbioses with ants, and these calls bear similarities to those produced by ants for communicating among themselves. In most Riodininae caterpillars calls are produced by a pair of mobile, grooved, rod-like appendages (vibratory papillae) arising from the distal edge of the first thoracic segment. An acoustical signal is produced when the grooves on the vibratory papillae grate against head granulations (Fig. 2). In other lycaenid caterpillars the call is produced by thickened bumps located ventrally between abdominal segments. Most concepts of insect communication suggest that acoustical calls evolved in a sexual context. However, caterpillar calls provide an example showing that, by forming symbiotic associations, the call of one species may evolve to attract another, unrelated species.

D. Ants and Caterpillar Associations
In general, myrmecophily in caterpillars occurs with a particular type of ant. A basic element among myrmecophilous caterpillars is that they typically form symbioses only with ant species that depend heavily on secretions as food—ants that also form symbioses with Homoptera and plants. Therefore, secretion-harvesting ants likely played a key role in the evolution of myrmecophily, whereas those ants that are predators or herbivores did not. An ecological consequence of the evolution with secretion-harvesting ants is that in any suitable contemporary habitat, a suite of caterpillar, plant, and Homoptera species all share the same species of ant symbionts. Among myrmecophilous caterpillars there are two main categories of ant association. The most widespread category comprises an association in which a particular caterpillar species may be tended by a suite of secretion-harvesting ant species. The other category comprises an association in which a caterpillar has an obligate association with a single species of ant. In this case, female butterflies may require the presence of a particular ant species to lay their eggs since caterpillars are unable to form symbioses with any other ant species. Furthermore, it is in these types of associations in which some caterpillars are adopted by the ants, taken into the nest, and become parasites or predators of their hosts.

IV. BUTTERFLY MIMICRY AND DIVERSITY
When several species of butterflies share conspicuously bright color patterns and fly together in the same habitat, this is called mimicry—a widespread and important antipredator defense. Fundamentally, mimicry occurs when one species closely resembles another species, and based on outward appearance one or both are avoided by predators. Virtually all major taxonomic groups of butterflies exhibit mimicry, but the warningly colored (aposematic) models are found predominately in particular groups of Papilionidae, Pieridae, and Nymphalidae, whose caterpillars feed on poisonous plants. Mimicry is a complex and subtle topic, but there are two basic types of mimicry in butterflies, both involving species that advertise conspicuous color pattern to predators: Batesian and Mullerian mimicry.

A. Batesian Mimicry
This is the phenomenon whereby one or more palatable species resemble one or more unpalatable model species, and the palatable species gain protection by duping predators. Here, predators avoid mimics because they resemble unpalatable models. The best known basic example of Batesian mimicry involves the palatable vice-roy (Limenitis archippus) and the unpalatable monarch butterfly (Danaus plexippus) in North America. However, Batesian mimicry occurs in all biogeographic regions, and one of the most intriguing examples involves the sex-limited mimicry of the wide-ranging African swallowtail, Papilio dardanus. Here, only the females are mimics. In this case, females are polymorphic with respect to their appearance, and their color patterns change geographically when flying with different model species. In other words, P. dardanus females from Kenya may look entirely different from those from Zaire or elsewhere, and in each instance they precisely mimic a different unpalatable model species. Thus, the type of female color pattern evolves in response to and depends on the local community diversity of the model species.

B. Mullerian Mimicry
This is the phenomenon whereby several unpalatable species (co-mimics) fly in the same area and share the
same color pattern. In this case, the association of similarly patterned unpalatable species is thought to reinforce the nasty experience predators have when eating butterflies of such a color pattern, thus educating predators to avoid it.

C. Mimicry and Diversity

Aposematic species are an important element of butterfly diversity, particularly in the tropics. For example, approximately 20–25% of all Papilionidae, Pieridae, and Nymphalidae in Costa Rica and Kenya are clearly aposematic models. Although accounting for a large percentage of all butterfly species, there are few documented examples of aposematic Lycaenidae.

Müllerian mimicry is primarily a tropical phenomenon and is an important part of tropical faunal diversity. This can be appreciated by considering the diversity of color patterns shown by many co-mimetic species and races of thionine and heliconine butterflies (plus many co-mimic moths and other butterfly groups) that converge across large areas of the neotropics. Even when the tightly overlapping distribution of only two co-mimetic species of Heliconius and their many precisely convergent color patterns are mapped over the Neotropical region, one cannot doubt the power mimicry plays in the evolution, organization, and diversification of tropical butterfly faunas. The numerous co-mimetic species of Acraeinae butterflies throughout the African region also serve as another potent example of the influence of mimicry on the diversity of butterfly communities. It is clear that butterfly mimicry evolved in the context of multispecies interactions which involved butterflies, their predators, and the plant and insect hosts they feed on as adults and caterpillars. Thus, mimicry serves to remind us that habitat disturbance and/or destruction may have important implications for the continued survival and future evolution of interacting species.

V. GEOGRAPHICAL PATTERNS OF BUTTERFLY DIVERSITY

One of the most frequently asked questions about butterflies is the following: How many are there? The answer is thought to be between 12,900 and 15,819 species. This is hardly a satisfying answer since the question has little biological context. In contrast, questions framed by comparing different geographical areas impart a sense of dynamics to butterfly diversity. For example, one might ask: Does butterfly species diversity change with respect to latitude? or How do different areas of Africa compare with respect to numbers of butterfly species? Indeed, comparing species numbers among areas shows that latitude, biogeographical region, area size, and taxonomic affinity all contribute to global butterfly species diversity. A caveat is in order about comparative species numbers: Sampling effort is almost always unequal among different areas, and the taxonomic accuracy of species counts or lists often varies among sites. Thus, it is prudent to use care when interpreting the generality of species diversity patterns among areas.

A. Latitudinal Gradients of Species Diversity

As is the case for most groups of terrestrial organisms, butterfly species diversity increases toward the equator. Comparing the North American and Neotropical faunas suggests that the neotropics has an estimated 10–12 times more butterfly species compared to North America (Table I). Comparing site diversity at different latitudes, however, provides a richer perspective of this phenomenon. A latitudinal transect using five well-known sites from the Americas makes it obvious that species numbers increase when moving from the north latitudes toward the equator (Table II), even though the smallest areas being compared are at the equator. The dramatic equatorial increase in species is exemplified by the fact that a mere 500 ha of lowland Ecuadorian forest has more butterfly species than all of North America.

The relative contribution of each butterfly family to site diversity also varies with latitude. From Table II, it is evident that the relative contribution of Papilionidae and Pieridae to total species diversity is about the same regardless of latitude, but the contributions of Nymphalidae and Lycaenidae increase dramatically between 20 and 0° latitude (Fig. 3). The increase in species diversity is...
TABLE II
Latitudinal Transect Showing Changes in American Butterfly Species Diversity

<table>
<thead>
<tr>
<th>Site</th>
<th>Latitude</th>
<th>Area</th>
<th>Papilionidae</th>
<th>Pieridae</th>
<th>Nymphalidae</th>
<th>Lycaenidae</th>
<th>Total species</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. J. Andrews Forest, Oregon</td>
<td>44°14'</td>
<td>6400 ha</td>
<td>5</td>
<td>7</td>
<td>27</td>
<td>23</td>
<td>62</td>
</tr>
<tr>
<td>Los Angeles Basin, California</td>
<td>34°4'</td>
<td>&gt;1 million ha</td>
<td>6</td>
<td>13</td>
<td>26</td>
<td>32</td>
<td>79</td>
</tr>
<tr>
<td>Estacion Los Tuxlas, Mexico</td>
<td>18°35'</td>
<td>700 ha</td>
<td>14</td>
<td>19</td>
<td>113</td>
<td>66</td>
<td>212</td>
</tr>
<tr>
<td>La Selva, Costa Rica</td>
<td>10°26'</td>
<td>1000 ha</td>
<td>16</td>
<td>26</td>
<td>219</td>
<td>181</td>
<td>442</td>
</tr>
<tr>
<td>Garza Cocha, Ecuador</td>
<td>0°29'</td>
<td>500 ha</td>
<td>24</td>
<td>23</td>
<td>314</td>
<td>312</td>
<td>676</td>
</tr>
</tbody>
</table>

The distribution of Nymphalidae is accounted for by subfamilies that reach their greatest diversity in the Amazon (Morphinae, Brassoimae, Charaxinae, Satyriinae, and Ithominae), whereas the increased contribution of Lycaenidae is due mainly to the subfamily Rodiniinae, which is also an Amazonian group.

B. Variation among Neotropical Sites
Species diversity is not equal among Neotropical forest sites. Nine well-known sites that have many species in common vary strongly in numbers of species and in the contribution of each family to species diversity (Table III). Here, species numbers may range from 212 in Mexico to 1199 at one Brazilian site. Among six Amazonian sites, numbers range from 676 to 1199 species. When averaged over the entire Table III, the contribution of each family to species diversity is as follows: Papilionidae, 3.8%; Pieridae, 4.8%; Nymphalidae, 45.2%; and Lycaenidae, 46.1%. It may be significant that these proportions approximately reflect those observed when comparing latitude (Fig. 3).

Comparisons among sites from the Americas point to the general relationship of the higher species richness that occurs at lower latitudes and to the marked variation among sites that share many species in common. These examples also show that the relative contribution of each family to species richness is not equal; the proportion of nymphalid and lycaenid species increases toward the equator, but the proportion of papilionid and pierid species appears relatively constant. At this point, we might ask if there are similar patterns outside the neotropics.

C. Regional Patterns of Species Diversity
Although most butterfly species are tropical, the number of species is not distributed equally among tropical regions (Table IV). As one might expect, those areas closest to the equator have a greater number of species, but there is also an effect of size on species diversity (Table IV). In this comparison, the continent of Africa has by far the most species, but it also encompasses the greatest geographical area and habitat types (deserts, mountains, forests, and savannas) of any area. Within Africa, there are more species in Zaire than in Kenya or Southern Africa, highlighting the ecological diversity in African climate and habitat types. Thus, the entire continent of Africa should contain more species than a geographic subset.

As in the neotropics, the relative contributions of families to total species numbers differ among areas of Africa (Table IV). Here, the Papilionidae and Nymphalidae contribute equally to total richness of both Kenya...
and Zaire, but Kenya shows an increased proportion of Pieridae and a decreased proportion of Lycaenidae. On the other hand, the proportional contributions of Papilionidae and Pieridae are reversed in the Zaire fauna. The Papilionidae and Nymphalidae in the Southern African fauna contribute less to the total than other mainland African areas, but nearly 55% of all Southern African butterflies are Lycaenidae. Overall, these proportional differences (Table IV) may reflect the ecological responses of taxonomic groups to differences in habitat types and/or the differences among regional taxonomists. Finally, in comparison to mainland Africa the fauna of Madagascar is interesting. Here, only 17% of all butterflies are lycaenids, but nymphalids constitute 67% of the total fauna, most of which are Satyrinae (i.e., 42% of all the butterflies). This example demonstrates how historical colonization and subsequent radiation of an island by one group can produce a fauna distinctly different in composition from that of the mainland.

Only a broad-brush comparison among areas of different sizes and regions is necessary to appreciate that, overall, butterfly diversity is highest in the neotropics (Table I–Table V). Despite the great disparity in geo-

### Table III

Comparative Diversity among Neotropical Sites

<table>
<thead>
<tr>
<th>Site</th>
<th>Papilionidae</th>
<th>Pieridae</th>
<th>Nymphalidae</th>
<th>Lycaenidae</th>
<th>Total species</th>
<th>Total area (ha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estacion Los Tuxlas, Mexico</td>
<td>14 (6.6)</td>
<td>19 (8.9)</td>
<td>133 (53.3)</td>
<td>68 (31.1)</td>
<td>212</td>
<td>700</td>
</tr>
<tr>
<td>Finca La Selva, Costa Rica</td>
<td>16 (5.6)</td>
<td>26 (8.8)</td>
<td>219 (49.9)</td>
<td>181 (40.9)</td>
<td>442</td>
<td>1,000</td>
</tr>
<tr>
<td>Garza Cocha, Ecuador</td>
<td>24 (3.3)</td>
<td>23 (3.7)</td>
<td>313 (46.8)</td>
<td>312 (46.3)</td>
<td>676</td>
<td>300</td>
</tr>
<tr>
<td>Jatun Sacha, Ecuador</td>
<td>29 (3.8)</td>
<td>23 (3.2)</td>
<td>326 (45.7)</td>
<td>345 (49.3)</td>
<td>690</td>
<td>600</td>
</tr>
<tr>
<td>Paquita, Maua, Peru</td>
<td>25 (3.8)</td>
<td>31 (3.6)</td>
<td>369 (43.3)</td>
<td>427 (50.1)</td>
<td>852</td>
<td>3,000</td>
</tr>
<tr>
<td>Tambopata, Peru</td>
<td>25 (3.4)</td>
<td>26 (3.3)</td>
<td>337 (42.3)</td>
<td>409 (51.3)</td>
<td>797</td>
<td>2,000</td>
</tr>
<tr>
<td>Caquandu, Brazil</td>
<td>31 (3.3)</td>
<td>31 (2.6)</td>
<td>423 (35.3)</td>
<td>715 (59.6)</td>
<td>1,100</td>
<td>2,000</td>
</tr>
<tr>
<td>Alto Jurea, Brazil</td>
<td>38 (3.7)</td>
<td>37 (3.6)</td>
<td>467 (45.0)</td>
<td>486 (47.8)</td>
<td>953</td>
<td>500</td>
</tr>
<tr>
<td>Cacaulandia, Brazil</td>
<td>30 (2.5)</td>
<td>31 (2.6)</td>
<td>423 (35.3)</td>
<td>715 (59.6)</td>
<td>1,138</td>
<td>2,000</td>
</tr>
<tr>
<td>Alto Jurua, Brazil</td>
<td>38 (3.7)</td>
<td>37 (3.6)</td>
<td>467 (45.0)</td>
<td>486 (47.8)</td>
<td>953</td>
<td>500</td>
</tr>
</tbody>
</table>

*a The percentage of the total species at a particular site is in parenthesis.

### Table IV

Variation in the Contribution of Each Family to Species Richness among Different Faunas

<table>
<thead>
<tr>
<th>Fauna</th>
<th>Papilionidae</th>
<th>Pieridae</th>
<th>Nymphalidae</th>
<th>Lycaenidae</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>All of Africa</td>
<td>80 (2.9)</td>
<td>145 (5.3)</td>
<td>1107 (40.6)</td>
<td>1397 (51.2)</td>
<td>2729</td>
</tr>
<tr>
<td>Zaire</td>
<td>46 (3.7)</td>
<td>120 (7.6)</td>
<td>607 (46.3)</td>
<td>393 (42.2)</td>
<td>1306</td>
</tr>
<tr>
<td>Kenya</td>
<td>27 (3.7)</td>
<td>87 (12.1)</td>
<td>353 (46.3)</td>
<td>271 (37.6)</td>
<td>720</td>
</tr>
<tr>
<td>Southern Africa</td>
<td>17 (2.3)</td>
<td>34 (7.2)</td>
<td>263 (53.6)</td>
<td>409 (54.9)</td>
<td>745</td>
</tr>
<tr>
<td>Madagascar</td>
<td>15 (4.9)</td>
<td>28 (10.7)</td>
<td>173 (56.8)</td>
<td>46 (15.7)</td>
<td>202</td>
</tr>
<tr>
<td>Australia</td>
<td>18 (6.3)</td>
<td>33 (12.6)</td>
<td>89 (30.6)</td>
<td>140 (30.3)</td>
<td>255</td>
</tr>
<tr>
<td>New Guinea</td>
<td>41 (3.2)</td>
<td>146 (11.6)</td>
<td>222 (28.3)</td>
<td>376 (54.0)</td>
<td>785</td>
</tr>
<tr>
<td>Malaysia</td>
<td>44 (8.8)</td>
<td>44 (8.8)</td>
<td>273 (53.9)</td>
<td>403 (52.3)</td>
<td>781</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>42 (10.0)</td>
<td>71 (18.9)</td>
<td>438 (41.9)</td>
<td>493 (47.2)</td>
<td>1044</td>
</tr>
</tbody>
</table>

*a The percentage of total faunal richness by each family is given in parentheses.
D. Species Diversity in Space and Time

When estimating species diversity in any habitat, the variables of time, space, and sample size can be profoundly important. For example, consider two samples taken in the same Wisconsin prairie—one for 7 days in mid-July and the other for 7 days in mid-January. Despite equal sampling effort, the number of species tallied in January would certainly be zero because no butterfly is known to fly in a prairie during midwinter. Here, it is easy to see that time of sampling is important for estimating species diversity.

The problem of sampling in space and time becomes more important in forest areas, especially tropical ones. Consider a recent study of fruit-feeding nymphalid butterflies conducted on 600 ha of lowland Ecuadorian rain forest that investigated how species were distributed in space and time. To ensure equal sampling, butterflies were simultaneously trapped in the canopy and understory for the first week of every month for a period of 1 year. The results showed that both time of sampling (month) and position of trap (canopy or understory) were extremely important in estimating species diversity at this site (Fig. 4). In other words, had the study sampled for only a few months, or only in the understory, species diversity would have been greatly underestimated. A similar study from a different site in Ecuador also demonstrated that a significant proportion of species diversity was accounted for by time of sampling and the forest canopy fauna. This study further showed that species diversity varied significantly over short distances within the same, seemingly uniform rain forest, and it highlights the importance of sample size in comparisons of species diversity (Fig. 5). Such studies illustrate the critical nature of sampling methods, space, time, and sample sizes in comparisons of butterfly diversity.

<table>
<thead>
<tr>
<th>Region</th>
<th>North America</th>
<th>Neotropics</th>
<th>Costa Rica</th>
<th>Africa</th>
<th>Australia</th>
<th>New Guinea</th>
<th>Malaysia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total species</td>
<td>471</td>
<td>1,044</td>
<td>2,729</td>
<td>2,78</td>
<td>785</td>
<td>761</td>
<td></td>
</tr>
<tr>
<td>% World fauna</td>
<td>3.4</td>
<td>38.8</td>
<td>7.6</td>
<td>19.8</td>
<td>2.0</td>
<td>5.7</td>
<td></td>
</tr>
</tbody>
</table>

* Total number of butterfly species for world fauna (13,753.5) and the neotropical region (5341) represent averages.
VI. BUTTERFLY DIVERSITY AND HABITAT DESTRUCTION

It is obvious that the evolution of butterfly diversity is based on historical and contemporary interactions with many species. These biological interactions include plant and/or insect hosts, co-mimics in Batesian and Mullerian mimicry complexes, predators, and parasites. Butterflies have also evolved within and adapted to a great many biomes, habitats, and microhabitats, ranging from the multilevels within lush tropical rain forests to starkly dry deserts and subarctic tundra.

Habitat destruction always has profound effects on the biological communities that inhabit them, and butterflies are no exception. Like all organisms, butterflies live, evolve, and diversify within dynamic biological systems, and as such they cannot be studied as art objects or protected as inventoried stock. To date, butterflies have served as tools for understanding the diversification of life on Earth and the fundamental interactions among species. However, our future understanding of butterfly diversity will depend on a renewed interest in studying them in the natural world and valuing the habitats in which they occur.

Acknowledgments

The author is indebted to P. Ackery, G. Austin, G. Beccaloni, S. Collins, H. Greeney, A. Heath, R. Hill, R. Lande, T. Larsen, D. Lees, J. Mallet, N. Pierce, K. Rozema, T. Walla, and especially C. M. Penz for providing information, encouragement, assistance, and boundless enthusiasm about butterfly diversity.

See Also the Following Articles

HYMENOPTERA • INSECTS, OVERVIEW • MOTHS • SPECIES COEXISTENCE

Bibliography


C₄ PLANTS

Rowan F. Sage
University of Toronto

1. Introduction
2. Physiological Significance
3. C₄ Biogeography
4. Taxonomic Diversity
5. The Rise of C₄ Photosynthesis
6. Climate Change and the Future of C₄ Photosynthesis
7. C₄ Vegetation and Global Biodiversity
8. Summary

GLOSSARY

bundle sheath Cell layer at the periphery of the vascular bundles in leaves. In C₄ plants, the reactions of the photosynthetic carbon reduction cycle are localized in this layer, in contrast to C₃ plants, where they occur in the mesophyll tissue. In comparison to C₃ species, the bundle sheath cells of C₄ species are small, with few chloroplasts.

Kranz anatomy Specialized anatomy of the C₄ leaf in which the bundle sheath tissue is enlarged and enriched with chloroplasts, whereas the mesophyll is reduced in size and often forms a lighter green halo around the bundle sheath. This produces a wreath-like (“Kranz” in the original German) appearance.

mesophyll tissue Photosynthetic cells that are located between the arrays of vascular bundles and bundle sheath cells of a leaf. In C₄ plants, PEP carboxylase and pyruvate-phosphate dikinase are localized in the mesophyll tissue, while Rubisco is absent.

NUE Nitrogen-use efficiency, a measure of photosynthesis or growth relative to the nitrogen content of the leaves or plant.

phosphoenolpyruvate carboxylase (PEPCase) PEP carboxylase, the initial carboxylation enzyme in C₄ photosynthesis.

photorespiration Biochemical process in which O₂ is assimilated by the oxygenation of RuBP, and CO₂ is given off in the metabolism of the products of RuBP oxygenation.

Rubisco Ribulose-1,5-bisphosphate carboxylase/oxygenase, the enzyme in all plants that catalyzes the formation of two phosphoglycerate molecules from RuBP and CO₂, and catalyzes the competing reaction of RuBP oxygenation.

WUE Water-use efficiency, a measure of the amount of photosynthesis or growth per unit of water lost in transpiration.

IN PLANTS, PHOTOSYNTHESIS OCCURS as one of three biochemical syndromes. Most of the estimated 270,000 terrestrial plant species employ C₃ photosynthesis, in which CO₂ is directly assimilated (fixed) in a reaction catalyzed by ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) (Fig. 1). Between 20,000 and 30,000 species utilize Crassulacean acid
metabolism (CAM) photosynthesis, a syndrome in which night-time fixation of CO₂ by phosphoenolpyruvate carboxylase (PEPCase) leads to the storage of carbon for fixation by Rubisco the following day. The third syndrome is C₃ photosynthesis, a CO₂ concentration mechanism also based on the initial carboxylation of phosphoenolpyruvate (PEP) by PEPCase. In C₃ plants, the C₃ biochemical pathway does not replace the C₂ pathway; instead it is a supplemental pathway that concentrates CO₂ in the tissue region where Rubisco and the C₃ photosynthetic biochemical system are localized. In so doing, inefficiencies associated with the C₂ pathway are largely avoided.

I. INTRODUCTION

Some 7600 advanced plant species have C₃ photosynthesis, making it the least used photosynthetic pathway in terms of species numbers. Despite representing less than 4% of all plant species, however, C₃ plants are ecologically significant in biomes that cover over 40% of the earth's landmass. They are well represented, if not dominant, in most tropical to warm-temperate grasslands, savannas, and salt marshes of the planet. C₃ plants are major contributors to agricultural productivity, particularly in tropical regions (Table 1). Economically significant C₃ species are currently increasing in importance because of their high productivity in tropical regions where human population growth is the most rapid. In a negative light, C₃ grasses that respond well to human-caused disturbance are serious threats to biodiversity because they aggressively spread where disturbance of tropical forests is extensive. Many non-weedy C₃ species are themselves threatened because the landscapes where they occur are often suited for agricultural development. Future threats to C₃ biodiversity also arise from rising atmospheric CO₂ levels and terrestrial eutro-

| TABLE 1 |
| Economically Significant C₃ Plants of the World

| Grasses | 
|-----------------|-----------------|
| Global crops | Zea mays (maize), Sorghum bicolor (sorghum), Pennisetum glaucum (pearl millet), Saccharum officinarum (sugarcane) |
| Regional cereals | Eleusine coracana (finger millet, Africa), Setaria italica (foxtail millets, N. Africa, Arabia), Digitaria exilis, D. decumbens, and D. iburua (foxtails, Africa), Brachiaria deflexa (annual fonio, Africa), Coix lacryma-jobi (adlay, India), Echinochloa colona (Sawa, India), Panicum sumatrense (sawa, India), Setaria pumila (kernels, India), Panicum somororum (sawit, Mexico) |
| Forages | Paspalum (5 spp.), Pennisetum (2 spp.), Panicum (2 spp.), Cenchrus ciliaris, Chloris gayana, Cyanoecylon daucylon, Digitaria decumbens, Melinis minutiflora, Sorghum albulum, Setaria anceps |
| Weeds | Cynodon dactylon (bermuda grass), Digitaria sanguinalis (crabgrass), Echinochloa crus-galli (barnyard grass), Echinochloa colona, Eleusine indica, Imperata cylindrica, Sorghum halepense (Johnson grass) |
| Turf grasses | Cynodon dactylon (bermuda grass), Zoysia japonica (zoysia), Pennisetum clandestinum, Paspalum spp. |
| Sedges (no major crops) | 
| Weeds | Cyperus excentius, Cyperus rotundus |
| Other | Cyperus papyrus (papyrus, central Africa) |
| Dicot | 
| Minor crops | Amaranthus caudatus (grain amaranth, Americas), Amaranthus tricolor (vegetable amaranth, Americas, India), Portulaca oleracea (purslane, S.E. Asia) |
| Weeds | Amaranthus (3 spp.), Heteropappus indicus, Portulaca oleracea, Chamaesyce (= Euphorbia) hirta, Salsola kali (tumbleweed of cowboy folklore), Tribulus terrestris (catthorn) |
| Other | Haloxylon spp. (saxaul, 4 species provided fuel and fodder along the silk roads of central Asia) |

phication, phenomena that can enhance the performance of C₃ plants over their C₄ competitors.

Because C₄ plants are significant to humans and ecosystem processes, they are important in discussions of global biodiversity. In such discussions, however, C₄ species cannot be treated as simply another set of plants. Unique features of the C₄ biochemistry produce physiological and ecological characteristics that are distinct from those of the C₃ vegetation that might otherwise occupy a landscape. These distinct characteristics in turn affect processes at the ecosystem or landscape level in a manner that might not occur if C₃ species were absent from the landscape. To enhance understanding of the role of C₄ plants for global biodiversity, this article provides a background perspective on the function, geography, taxonomic diversity, and evolutionary origin of the world’s C₄ biota.

II. PHYSIOLOGICAL SIGNIFICANCE

C₄ photosynthesis is often treated as a single biochemical pathway. This is an oversimplification. It is better regarded as a heterogeneous syndrome resulting from over 30 independent evolutionary modifications of C₃ leaf anatomy and biochemistry. Three distinct biochemical pathways and eight anatomical configurations have been described (Fig. 2). Together, these yield 14 recognizable patterns of C₄ photosynthesis. In addition, over a dozen species have biochemical characteristics intermediate between C₄ and C₃ photosynthesis. Given this variation, what general characteristics distinguish C₄ from C₃ plants?

A. Common Features

Four features are common to all C₄ species. First, PEP carboxylase activity is enhanced 5- to 50-fold over that of C₃ species, and all of it is restricted to photosynthetic mesophyll cells of the leaf and stem. PEPCase catalyzes the formation of oxaloacetic acid (OAA, a four-carbon or C₄ acid) from a three-carbon acceptor molecule (PEP) and bicarbonate (HCO₃⁻, formed when CO₂ dissolves in water) (Fig. 3). Second, Rubisco is localized in the bundle sheath cell layer that is generally located at the periphery of the vascular bundles of the leaf (see Fig. 2). Third, a decarboxylation reaction is catalyzed by one of three enzymes localized in the bundle sheath cells. The three decarboxylases are NADP-malic enzyme (NADP-ME), NAD-malic enzyme (NAD-ME), and PEP carboxykinase (PCK). Finally, the bundle sheath cells of C₄ plants are enlarged and have more chloroplasts when compared with C₃ species, while the mesophyll cells are reduced in number. In C₄ plants, one or two mesophyll cells are present for every bundle sheath cell; in C₃ species, more than four mesophyll cells are typically present for every bundle sheath cell.

CO₂ is concentrated in the bundle sheath when C₄ acids (malic or aspartic) derived from OAA diffuse from the mesophyll to the bundle sheath cells (see Fig. 3). There, the C₄ acids are decarboxylated to produce CO₂ and a three-carbon product (pyruvate or PEP). The CO₂ accumulates to very high levels, typically 5- to 10-fold over the atmospheric level, while the three-carbon residue diffuses back to the mesophyll cells where it will, if necessary, be converted to PEP with the expenditure of two ATP molecules per CO₂ fixed. The final CO₂ fixation step is the carboxylation of RuBP by Rubisco, using the CO₂ transported into the bundle sheath cells by the C₄ cycle. Metabolism of the resulting phosphoglycerate (PGA, the three-carbon or C₃ product of RuBP carboxylation) into sugars occurs in the same reaction sequence as in C₃ plants. None of the enzymes in the C₄ cycle are novel, as similar biochemical pathways control stomatal aperture, cellular pH, and respiratory metabolism in C₃ species. The novel changes in C₄ evolution are altered patterns of gene expression that in turn affect tissue development, enzyme regulation, and the localization of enzymes within the leaf.

B. Photosynthetic Enhancements

1. Functional Role

The C₄ pathway has two immediate benefits: first, it raises the CO₂ content in the bundle sheath to levels enabling Rubisco to operate close to its maximum activity. Second, the build-up of CO₂ around Rubisco greatly reduces the inhibitory process of photorespiration. In photorespiration, Rubisco binds O₂ to RuBP to produce the two-carbon compound phosphoglycolate (PG) and PGA (see Fig. 1). While the PGA can be immediately metabolized back to RuBP by the carbon reduction cycle, PG has no apparent value to the plant and must be metabolized to avoid its accumulation to toxic levels. Photorespiration is the metabolic process in which PG is converted to PGA, but in this process 25% of the carbon in the pool of PG molecules is lost as CO₂. In addition, for every two RuBP oxygenation events, one ATP and the equivalent of one NADPH is required to recover the remaining 75% of the carbon in the PG pool. Because of the loss of previously fixed carbon and the additional energy required to scavenge carbon from PG, photorespiration is potentially a significant inhibition on photosynthesis.
There are three requirements for photorespiration to occur at a high rate (Sage, 1999). First, atmospheric \( \text{O}_2 \) must be abundant to provide the substrate for RuBP oxygenation. The current \( \text{O}_2 \) level of 21% is sufficient to support high levels of oxygenation if the other preconditions are met. Second, \( \text{CO}_2 \) levels in the atmosphere must be less than about 500 ppm (Fig. 4A). Elevated levels of atmospheric \( \text{CO}_2 \) (\( \geq 1000 \) ppm) effectively inhibit the rate of RuBP oxygenation, while the current level of about 370 ppm allows for modest photorespiration at temperatures near 25°C. At present, the level of atmospheric \( \text{CO}_2 \) is anthropogenically enhanced by one-third over levels that persisted during the Holocene epoch of the past 10,000 years, and it is double the \( \text{CO}_2 \) level that predominated in the late Pleistocene (10,000 to 500,000 years ago). As a result of these lower \( \text{CO}_2 \) levels in recent geological time, photorespiration was higher in \( \text{C}_3 \) plants than it is today, and this likely...
reduced the photosynthetic potential of C₃ relative to C₄ vegetation.

The third requirement is warm temperatures. At low temperatures (<12°C), the rate of photorespiration is suppressed because Rubisco has low specificity for O₂ relative to CO₂. As temperature increases, the specificity of Rubisco for O₂ increases relative to that for CO₂ and the solubility of CO₂ relative to O₂ declines. Together,

**FIGURE 3** Diagrammatic representation of the C₃ photosynthetic cycle. The product of the PEP carboxylation reaction is oxaloacetic acid (OAA), which is converted to a transporable C₃ acid, either malic acid (NADP-ME type) or aspartic acid (NAD-ME and PCK). The decarboxylation reactions produce either pyruvic acid (PVA, produced by NADP-ME and NAD-ME) or PEP (produced by PEP carboxylase). The PVA returns to the mesophyll chloroplasts, where it is converted to PEP using ATP. CHL, chloroplast; tricose-P, tricose phosphate; PCR, photosynthetic carbon reduction cycle or C₄ cycle.

**FIGURE 4** (A) The relationship between photorespiratory inhibition of photosynthesis and chloroplast CO₂ content in C₃ plants at 21% atmospheric O₂. The brackets indicate the range of chloroplast CO₂ levels (C₃) that correspond to atmospheric CO₂ levels (C₄) of the current era (370 ppm), the preindustrial Holocene era (270 ppm), and the late Pleistocene epoch (200 ppm). (B) The relationship between chloroplast CO₂ and carboxylation potential (Vₐ – 0.5Vₐ) expressed relative to the maximum rate of Rubisco at CO₂ saturation. Responses modeled as described by Sage (1995).
these responses enhance RuBPy oxygenation such that at 35°C, photorespiration inhibits photosynthesis by over 20% in current atmospheric conditions and by over 30% at preindustrial CO₂ levels of 270 ppm (see Fig. 4A). By contrast, because of the high level of CO₂ in the bundle sheath, photorespiration in C₄ plants is marginal (<5%) at all temperatures (Table II).

Temperature also affects the kinetics of Rubisco in a manner that has significance for C₄ photosynthesis (Fig. 4B). At low temperature (<12°C), Rubisco operates close to CO₂ saturation in the current atmosphere. As temperature increases, the amount of CO₂ required to saturate Rubisco rises, such that in C₄ plants in air, the enzyme is operating well below its maximum rate above 30°C. By raising CO₂ levels in the bundle sheath, C₄ plants are able to operate Rubisco near its maximum, substrate-saturated activity (Vₘₐₓ) at all temperatures (see Fig. 4B).

Together, the ability to reduce photorespiration and operate Rubisco near CO₂ saturation confers substantial photosynthetic advantages on C₄ species in warm climates and low-CO₂ atmospheres. These differences are best demonstrated by comparing the CO₂ response of photosynthesis of C₃ and C₄ species of similar ecological habitat, for example, prairie grasses or agricultural weeds (Fig. 5A). In such comparisons, the C₄ species always have CO₂ compensation points of photosynthesis near 0 ppm in contrast to their C₃ counterparts, for which the CO₂ compensation point is 40 to 60 ppm. C₄ plants have a steeper initial slope of CO₂ assimilation versus rising CO₂ than do C₃ species, which combined with their lower CO₂ compensation point enables them to have a higher photosynthesis rate at low CO₂ (Fig. 5B). C₄ photosynthesis becomes CO₂ saturated at CO₂ levels near the current atmospheric value, whereas C₃

![TABLE II](image)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>C₃</th>
<th>C₄</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum photosynthesis rate</td>
<td>20–50</td>
<td>35–75</td>
<td>µmol m⁻² s⁻¹</td>
</tr>
<tr>
<td>Maximum daily production rate</td>
<td>10–40</td>
<td>40–80</td>
<td>g m⁻² d⁻¹</td>
</tr>
<tr>
<td>Maximum biomass yield</td>
<td>1–5</td>
<td>3–8</td>
<td>kg DM m⁻² yr⁻¹</td>
</tr>
<tr>
<td>Photosynthetic thermal optimum</td>
<td>15–30</td>
<td>30–40</td>
<td>°C</td>
</tr>
<tr>
<td>Photorespiration/photosynthesis @ 10°C</td>
<td>8%</td>
<td>1–5%</td>
<td></td>
</tr>
<tr>
<td>@ 25°C</td>
<td>25%</td>
<td>1–5%</td>
<td></td>
</tr>
<tr>
<td>@ 40°C</td>
<td>40%</td>
<td>1–5%</td>
<td></td>
</tr>
<tr>
<td>δ¹³C</td>
<td>-20 to -35</td>
<td>-10 to -15</td>
<td>per mil</td>
</tr>
<tr>
<td>Intercellular CO₂</td>
<td>200–300</td>
<td>100–200</td>
<td>ppm</td>
</tr>
<tr>
<td>Maximum leaf nitrogen (N)</td>
<td>150–250</td>
<td>80–160</td>
<td>mmol m⁻²</td>
</tr>
<tr>
<td>Relative Rubisco content</td>
<td>15–30</td>
<td>5–10</td>
<td>% of leaf nitrogen</td>
</tr>
<tr>
<td>Relative PEPCase content</td>
<td>&lt;1</td>
<td>3–6</td>
<td>% of leaf nitrogen</td>
</tr>
<tr>
<td>Maximum efficiencies:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water use (WUE)</td>
<td>1.5–2.5</td>
<td>3–5</td>
<td>g DM kg⁻¹ H₂O</td>
</tr>
<tr>
<td>Nitrogen use (NUE)</td>
<td>50–280</td>
<td>280–520</td>
<td>µmol CO₂ mol⁻¹ N</td>
</tr>
<tr>
<td>Radiation use (RUE)</td>
<td>1.7–1.9</td>
<td>2.5</td>
<td>g DM M⁻¹</td>
</tr>
</tbody>
</table>

*DM indicates dry matter. Developed from Larcher (1995), Brown (1999), and Sage and Pearcy (1999).*
TABLE III

| Energy Requirements for Carbon Metabolism to Triose Phosphate, and Associated Quantum Yields of C₃ and C₄ Plants at 30°C* |
|---|---|---|
| ATP | NADPH | 350 ppm CO₂ |
| No photorespiration | 3 | 2 | 0.078 |
| With 32% photorespiration | 5 | 3.2 | 0.053 |

C₃ photosynthesis

<table>
<thead>
<tr>
<th>NADP-ME and NAD-ME types</th>
<th>ATP</th>
<th>NADPH</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₃ cycle</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>C₄ cycle</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>25% leak rate</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>3% photorespiration</td>
<td>0.15</td>
<td>0.1</td>
</tr>
<tr>
<td>Total</td>
<td>5.05</td>
<td>2.1</td>
</tr>
<tr>
<td>PCK types (assuming 25% NAD-ME activity per unit of PCK activity)**</td>
<td>ATP</td>
<td>NADPH</td>
</tr>
<tr>
<td>C₃ cycle</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>C₄ cycle</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>25% leak rate</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>3% photorespiration</td>
<td>0.15</td>
<td>0.1</td>
</tr>
<tr>
<td>Total</td>
<td>4.71</td>
<td>2.1</td>
</tr>
</tbody>
</table>

350 ppm CO₂ at 30°C

* Adapted from Hatch (1987) and Kanai and Edwards (1999). The leak rate refers to fraction of pumped CO₂ that leaks out of the bundle sheath. The percentage photorespiration refers to the rate of photorespiration relative to photosynthesis.

** PCK-type plants require one ATP to power each PEP carboxykinase (PCK) reaction in the bundle sheath. This ATP is supplied by oxidation of malate in the bundle sheath mitochondria, with the malate coming from NAD-ME activity at a ratio of one NAD-ME event per three PCK events (Kanai and Edwards, 1999).

plants show substantial photosynthetic enhancement with CO₂ enrichment above the current atmospheric value. Thus, at elevated CO₂, C₃ species typically have higher rates than their C₄ competitors in short-term gas exchange studies. In warm environments, these differences translate into significant photosynthetic and growth advantages for C₄ species in the atmospheric conditions of the late twentieth and early twenty-first centuries. In crops and weeds, for example, C₄ species on average have greater photosynthetic capacity, higher daily production rates, and yield more biomass during a growing season than do C₃ species (see Table II).

2. The Cost of C₄ Photosynthesis

In C₃ plants without photorespiration, each CO₂ costs 2 NADPH and 3 ATP to fix it into carbohydrate end products (Table III). In C₄ plants, 2 NADPH and 4.7 to 5.7 ATP are needed, assuming that 25% of the pumped CO₂ leaks out of the bundle sheath (Kanai and Edwards, 1999). The higher ATP cost of C₄ photosynthesis reflects the energy required to pump CO₂ into the bundle sheath. As the rate of photorespiration increases in C₃ plants, the energy costs of photosynthesis rise, such that they are equivalent to that of C₄ plants at 25°C to 30°C, and greater above about 30°C (see Table III). Differences in the energy requirement of photosynthesis are demonstrated by comparing differences in the light-use efficiency (quantum yield) of C₃ and C₄ plants as a function of temperature (Table III and Fig. 6). C₃ species have superior quantum yields at cool temperatures, but lower quantum yield than C₄ species at warm temperature. Using quantum yields, it is possible to model the relative performance of C₃ versus C₄ photosynthesis as a function of CO₂ and temperature; as temperature increases, the CO₂ level at which C₄ species perform as well as C₃ plants increases (Fig. 7). For unexplained reasons, C₄ grasses have higher light-use efficiencies than do C₃ dicots, and thus begin to outperform C₃ species at cooler conditions and higher CO₂ levels.

3. Enhancement of Water- and Nitrogen-Use Efficiency

Second-order consequences arise as a result of the direct effects of the C₄ photosynthetic pathway on the catalytic efficiency of Rubisco. Because the C₄ cycle has a high capacity to concentrate CO₂ into the bundle sheath, even at low atmospheric CO₂ levels, C₄ plants will have less open stomata than C₃ species with the same photosynthetic rate. This reduces the rate of transpiration in C₄ relative to C₃ plants and leads to significant enhance-
ment in water-use efficiency. Typically, C₄ species have two to three times the water-use efficiency of ecologically similar C₃ species (see Table II). Even at very low stomatal apertures that would yield negligible rates of CO₂ assimilation in C₃ plants, C₄ plants are able to have significant rates of photosynthesis. This is most important in hot, arid regions, where evaporative demands are so high that anything but very low stomatal apertures could lead to unsustainable rates of water loss. High water-use efficiency also assists C₄ species in saline soils, apparently by reducing the volume of transpirational water from which salt must be filtered.

In addition to improved water use, C₄ plants have higher nitrogen-use efficiencies of photosynthesis, typically 1.5 to 3 times that of C₃ species (see Table II). Because C₄ plants operate Rubisco near CO₂ saturation and greatly reduce oxygenase activity, they require about 25% as much Rubisco as C₃ species of equal photosynthetic capacity, and have very low levels of photorespiratory enzymes. As a consequence, C₄ plants have about 30% less nitrogen per unit leaf area as do ecologically similar C₃ species. These differences promote superior performance of C₄ grasses in nitrogen-deficient soils, such as those of the prairie grasslands of temperate North America.

III. C₄ BIOGEOGRAPHY

Clear patterns are evident in the global distribution of C₄ plants that are related to their distinct photosynthetic responses to temperature and light. As a result of their superior photosynthetic performance at high temperature, and their lower quantum yields below 25°C, C₄ species have superior growth in warmer but not cooler conditions, yet perform poorly in deep shade relative to similar C₃ species. The primary requirements for the success of C₄ plants are therefore warmer temperatures and access to at least moderate light intensities. Aridity, soil nitrogen supply, and salinity also affect the success of C₄ plants, but these are secondary in that they become important in warm, illuminated habitats.

A. Primary Controls: Temperature and Light

The critical temperature affecting C₄ plant distribution is growth season temperature. In temperate zones, C₄ plants tolerate severe cold outside of the growth season as well as co-occurring C₃ species, and often tolerate low night temperature as well as their C₃ associates. During the growing season, however, leaf temperature must routinely rise above 25°C for C₄ plants to be present in a community. Where daily temperatures commonly rise above 30°C during the growing season, they generally dominate grass and sedge floras of open landscapes. When plotted on the basis of growth season temperature (which reflects night and morning temperatures as well as the afternoon high), C₄ plants dominate grassland floras and biomass above 20–22°C. Above growth season averages of 23–24°C, C₃ grass productivity in open grasslands is very low (typically less than 10% of total biomass). At the other end of the spectrum, C₄ biomass is rare to absent where the temperature at the peak of the growth season averages less than 12–14°C. Some C₃ species occur at low frequency in cold climates, but these are usually restricted to warm microsites where daytime temperatures rise above the regional average.

The control of temperature over C₄ distribution is demonstrated by altitude and latitude transects that show the contribution of C₄ grasses to regional grass floras and grassland productivity. Where C₄ species dominate lowland floras, they decline in frequency as elevation increases, becoming uncommon above 2000 to 3000 m. Along latitude transects, the transition between C₃- and C₄-dominated floras corresponds to temperate latitudes, as is demonstrated by a survey of grass floras of the world's oceanic islands (Fig. 8). Above >50° latitude, C₃ species are rare or absent, and below temperate latitudes (<30°), C₄ grasses dominate the grass flora of any given island. On larger tropical islands, where extensive forests form, a significant number of C₃ forest grasses are present, so that the C₄ contribution
may only be 70% of the grass flora. On smaller islands and atolls, where soils tend to be shallow, sandy, and drought-prone, grasses are often all C₄.

The second requirement for C₄ success is availability of moderate to high light intensity. Typically, this means that C₄ species require 30 to 50% of full sunlight intensities if they are to dominate a local habitat. In forest canopies where light intensities average less than 20% of full sun on a clear day, C₄ species rarely occur, regardless of latitude. Why C₄ species fail to dominate in low-light environments appears to result from a combination of factors. Shaded environments tend to be cool (<30°C at midday), so that high temperatures favoring C₃ species are not frequently encountered. In shade, much of the carbon that a plant acquires is during brief episodes of high light (sunflecks), and C₄ species appear to have a slightly reduced ability to exploit short sunflecks than do C₃ species. Also, because shaded environments tend to be cooler and more humid, C₃ plants often have more open stomates than they otherwise would. This allows for higher intercellular CO₂ levels in shade plants, thereby promoting greater photosynthetic efficiency.

The failure of C₄ species to do well in shaded settings has important consequences for the distribution of biomes around the globe. Except for a few desert shrubs that become arborescent with advanced age, and an unusual understory tree from Hawaii, there are no known C₄ trees or vines (Table IV), and relatively few C₄ herbaceous species survive in the interior of established forests. Thus, in warm to hot climates, the distribution of forest and woodland is the single greatest indicator of C₃ abundance and diversity. Where forests are present, the light environment will typically be too low to support a C₄ flora. Where forests are absent, either because of edaphic factors (e.g., shallow, nutrient-poor soils), severe abiotic stress (e.g., killing drought), mechanical stress (e.g., wind, waves, or flooding), logging, fire, or large animal activity, C₃ plants will be important components of the regional diversity. C₃-rich habitats include beaches along warm temperate to tropical coasts where wave action prevents woody dominance, tropical swamps and floodplains where flooding prevents dominance by a woody overstory, abandoned cropland at low latitudes, and warm temperate to tropical grasslands (Table V and Fig. 9). For example, in sub-Saharan Africa, the margins of lakes and slow-moving rivers are dominated by papyrus (Cyperus papyrus), a C₃ sedge. In the African savannas, elephants, giraffes, and fire aid in the reduction of C₃ Acacia trees that otherwise would dominate the landscape. Where elephants and fire become common, wooded landscapes are converted to C₄ grassland. Fire control and excessive poaching of elephants permit woodland formation, leading to a loss of the C₄-dominated grasslands.

B. Secondary Controls: Moisture, Nutrients, and Salinity

Aridity is often noted as directly supporting C₄ over C₃ biomass, because the higher WUE of C₄ species theoretically promotes their competitiveness in arid situations. In addition, interactions between aridity and other ecological factors, notably fire, are critically important. In the case of fire, aridity slows growth of woody species, thereby increasing their establishment time and the period during which they are vulnerable to fire injury. Aridity also enhances the probability of initial fire establishment, its subsequent rate of spread, and its intensity. In contrast to most woody species, C₃ grasses are typically fire tolerant, as their meristems are protected below ground or in fire-resistant tufts. Thus, the success of C₃ grasses in arid regions is promoted because fires typically kill or inhibit competing woody vegetation.

The WUE advantage of C₄ photosynthesis is perhaps most critical in very harsh locations, such as hot deserts in Africa, Asia, and northern Australia. Here, lack of rainfall combined with high daily temperature creates conditions where the transpiration potential may be greater than what a C₃ plant could support, particularly in the seedling stage. C₃ species not only survive, but they can thrive under thermal regimes that are harmful to most C₄ species. Consequently, C₃ plants are common if not dominant in the hot deserts of the earth. By contrast, the floras of deserts at high latitude or high elevation are exclusively C₃.
### TABLE IV
Principal Life-forms of C₃ Plants

<table>
<thead>
<tr>
<th>Life-form</th>
<th>C₃ families</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees</td>
<td>None</td>
<td><em>Haloxylon</em> spp. (xerophytic halophytes and psammophytes)*¹</td>
</tr>
<tr>
<td></td>
<td>Sub-trees (5 to 10)</td>
<td><em>Calligonum</em> spp. (xerophytic psammophytes)</td>
</tr>
<tr>
<td></td>
<td>(5-10)</td>
<td><em>Euphorbia</em> fouesi (small Hawaiian understory tree)</td>
</tr>
<tr>
<td>Vines</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Shrubs (5007)</td>
<td>Amaranthaceae</td>
<td><em>Alternanthera, Arvea, Froelichia, Gomphrena,</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Guilleminia, Tidestromia</em> (all xerophytic)</td>
</tr>
<tr>
<td></td>
<td>Chenopodiaceae</td>
<td><em>Anabasis, Atrophyles, Haloxylon, Salsola, Suaeda</em> (mostly xerophytic halophytes and psammophytes)</td>
</tr>
<tr>
<td></td>
<td>Euphorbiaceae</td>
<td><em>Chamaesyce</em> spp. (=<em>Euphorbia</em>, xerophytic shrubs)</td>
</tr>
<tr>
<td></td>
<td>Polygonaceae</td>
<td><em>Calligonum</em> spp. (xerophytic psammophytes)</td>
</tr>
<tr>
<td>Forbs (dicot herbs) (1000)</td>
<td>All dicot families</td>
<td>Summer desert annuals, xerophytic annual and perennial herbs of low latitudes, tropical and summer-active weeds, psammophytic herbs of low latitudes, tropical beach herbs</td>
</tr>
<tr>
<td>Graminoids (6000)</td>
<td>Poaceae</td>
<td>Grasses, mainly of low-latitude and warm temperate summer habitats</td>
</tr>
<tr>
<td></td>
<td>Cyperaceae</td>
<td>Sedges, mainly low-latitude and warm temperate wetlands</td>
</tr>
</tbody>
</table>

* Sub-trees refer to species that develop arborescence with age, but never rise above 12 m in height.
* Numbers in parenthesis refer to the estimated number of species in each life-form (after Sage et al., 1999b).
* Psammophytes are plants of sandy soils.

### TABLE V
A List of the Major Biomes of the World That Have High C₃ Plant Representation (Loosely Defined as >25% of the Vegetative Cover and a Potential for C₃ Dominance) or Have No C₃ Dominance and C₃ Cover is Negligible (<1%).

<table>
<thead>
<tr>
<th>Biome</th>
<th>Major Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>High C₃ representation</td>
<td></td>
</tr>
<tr>
<td>Tropical, subtropical grassland and savanna</td>
<td>South and Central America, Africa, India, S.E. Asia, Australia</td>
</tr>
<tr>
<td>Warm temperate grassland and savanna</td>
<td>Central and S.E. North America, N. Argentina, Australia</td>
</tr>
<tr>
<td>Arid steppe (low to middle latitudes)</td>
<td>S.W. North America, central Asia, Australia, Africa</td>
</tr>
<tr>
<td>Beach dunes and bluffs, warm temperate to tropics</td>
<td>Global</td>
</tr>
<tr>
<td>Tropical, subtropical wetlands (nonarborescent)</td>
<td>Global, especially South America, central Africa, S.E. Asia</td>
</tr>
<tr>
<td>Salt marsh (warm temperate to tropical)</td>
<td>Global, but more temperate owing to mangrove dominance in tropics</td>
</tr>
<tr>
<td>Salt desert (&lt;15° latitude)</td>
<td>W. North America, central Asia, central Australia</td>
</tr>
<tr>
<td>Desert and semideserts</td>
<td>Southwest North America, Africa, Australia, S.E. Asia</td>
</tr>
<tr>
<td>Disturbed ground (low latitudes, low altitude)</td>
<td>Global, more in arid regions</td>
</tr>
<tr>
<td>Always C₃ dominated</td>
<td></td>
</tr>
<tr>
<td>Forests (all, including arborescent wetlands)</td>
<td>Global</td>
</tr>
<tr>
<td>Tundra (all)</td>
<td>Polar latitudes, high mountains (all latitudes)</td>
</tr>
<tr>
<td>Heathlands</td>
<td>N. Europe, Canada, Russia</td>
</tr>
<tr>
<td>Cool temperate grasslands and savanna</td>
<td>Canada, N.W.A.S.A., S.E. Europe, S. Russia, Australia, S. Argentina, Tasmania</td>
</tr>
<tr>
<td>Montane grasslands</td>
<td>Global, elevations between 2000 and 4000 m</td>
</tr>
<tr>
<td>Mediterranean grasslands</td>
<td>California, Chile, S.Europe, Asia Minor, N. Africa, S. Africa, S.W. Australia</td>
</tr>
<tr>
<td>Mediterranean-type shrublands (shrubland)</td>
<td>California, Chile, S.Europe, Asia Minor, N. Africa, S. Africa, S.W. Australia</td>
</tr>
<tr>
<td>Temperate to boreal wetlands (nonarborescent)</td>
<td>Canada, N. Europe, Russia, New Zealand, Patagonia</td>
</tr>
<tr>
<td>Cold deserts and semideserts</td>
<td>W. North America, N. central Asia, Tibet</td>
</tr>
<tr>
<td>Salt marshes (&gt;50° latitude)</td>
<td>Global</td>
</tr>
<tr>
<td>Mangroves (mangrove swamps)</td>
<td>Tropical, subtropical latitudes</td>
</tr>
<tr>
<td>Disturbed ground (high latitude and altitude)</td>
<td>Global</td>
</tr>
</tbody>
</table>

reflecting the generally poor ability of $C_4$ species to compete in cold climates, regardless of any aridity that may be present.

The seasonality of precipitation is of great importance. Grassland ecosystems can be almost completely $C_3$ or $C_4$ depending on the timing of precipitation. Grasslands of Mediterranean climate zones that have cool-season precipitation (southern Europe, North Africa, California, Chile, southwestern Australia, and southwestern Africa) are heavily $C_3$ dominated, whereas at similar latitudes, temperate monsoon grasslands with significant summer precipitation are $C_4$ dominated. For example, the southwest tip of Africa near Capetown has a Mediterranean-type climate and $<10\%$ $C_4$ species in the grass flora. At Durban, on the eastern coast of South Africa at similar latitudes, the bulk of the precipitation falls during the warm season and the grass flora is $75\%$ $C_4$ (Fig. 10). Oceanic islands show a similar pattern. In the eastern Atlantic and Pacific at temperate latitudes, summer rain is rare and $C_4$ presence is low (see Fig. 8). On the western edges of these oceans, summer precipitation is abundant as are $C_4$ grasses.

Seasonality of precipitation is particularly important in hyperarid regions of the world. In southwestern North America, for example, the Mojave Desert in California receives 60 to 90% of its precipitation during the winter, and as a consequence it has a high proportion ($40\%$) of $C_4$ grasses in the grass flora. The Chihuahuan Desert in southwestern Texas receives the majority of its precipitation in summer, and $C_3$ grasses are infrequent ($20\%$ of the grass flora). Due in part to the low summer precipitation, the Mojave Desert lacks native grasslands, in contrast to the Chihuahuan Desert, where there are extensive grasslands dominated by $C_4$ species. In addition, two floristic seasons are often present in warm temperate deserts. In the deserts of the American Southwest, winter rains and mild temperatures support a rich flora of $C_3$ annuals that produce spectacular blooms during wet years. By contrast, summer rains produce a rich flora of $C_4$ annuals (Table VI). Because $C_4$ annuals typically produce inconspicuous flowers (Guo and Brown, 1997), the blooms of summer annuals fail to attract the attention that the cool-season blooms receive.

Moderate salinity tends to favor $C_4$ species, apparently because the inherently greater WUE reduces the salt load that the plants must deal with in the transpiration stream. In temperate to tropical latitudes, marshes and soils of moderate salinity are commonly dominated by $C_4$ species. For example, $C_4$ grasses of Spartina and Distichlis dominate salt marshes of coastal North America and Europe, and saltmarsh rangelands of central Asia and western North America are often dominated by $C_4$ species in the Chenopodiaceae (see Table V). Moderate salinity also appears to allow $C_4$ species to dominate landscapes in climate zones that may otherwise be too cold. Spartina marshes predominate in eastern Canada and northern Europe at latitudes where inland there is no $C_4$ dominance. In turn, in Mediterranean climate zones, $C_4$ salt grasses (Distichlis spp.) can dominate coastal bluffs and cliff ledges because higher salinity allows them to compete with the local $C_3$ flora during the cooler months of the year.
IV. TAXONOMIC DIVERSITY

A. Description

There are 7000 to 8000 C₄ plant species. In proportional terms, 3 to 4% of the world's terrestrial flora is C₄. All C₄ species are angiosperms, with no known C₄ ferns, conifers, lower vascular plants, or algae. Of the C₄ angiosperms, all are in advanced rather than primitive orders. Recent angiosperm phylogenies place the presence of C₄ photosynthesis in seven taxonomic orders, with the Zygophyllaceae remaining unclassified to order because of uncertain affinity (Fig. 11).

The distribution of C₄ photosynthesis within the angiosperm orders is dispersed, with no apparent pattern. Two monocot and five dicot orders include C₄ families. Within the dicots, the Eurosids I and II lines and Euasterids I and II lines each include an order containing C₄ families. With the exception of the Caryophyllalea (with eight families), each order has only one or two families containing C₄ species. In turn, many of these families include only a few C₄ genera. The most prolific C₄ families are the Poaceae (grasses) with almost 400 genera and 4600 species and the Cyperaceae (sedges) with 21 genera and about 1400 species (Table VII). The major C₄ dicot families are the Chenopodiaceae, Amaranthaceae, and Asteraceae, which together contain two-thirds of the C₄ dicot genera and species. With the exception of the Chenopodiaceae and Amaranthaceae, C₄ representation is generally low in the dicot families in which it occurs. Nearly 40% of the chenopods and 25% of the amaranth species are C₄, in contrast to 10 dicot families in which the C₄ syndrome is present in less than 5% of the species (Sage et al., 1999a).

B. Polyphyletic Origins

The widespread occurrence of C₄ photosynthesis in diverse angiosperm lineages was an initial observation indicating that C₄ photosynthesis is polyphyletic and of recent origin. Further support for this interpretation came from comparative biochemical and anatomical studies that describe the variation in decarboxylation steps and leaf anatomy between different C₄ lineages. From this work, over 30 independent origins of C₄ photosynthesis have been identified (Kellogg, 1999). Of the three decarboxylation pathways, the NADP-ME and NAD-ME pathways are evolutionarily unrelated, while the PCK pathway is derived from the NAD-ME pathway (Kellogg, 1999). Taxa with NAD-ME or PCK biochemistry are thus the result of separate evolutionary origins than NADP-ME taxa. On the basis of this understanding, it can be concluded that C₄ photosynthesis arose on multiple occasions within both families and individual genera. Of the 18 families with C₄
TABLE VI

Photosynthetic Pathway and Mean Density over a 10-Year Period for the Twelve Most Abundant Species of Summer and Winter Annuals on the Cave Creek Bajada, near Portal, Arizona (Chihuahuan Desert)

<table>
<thead>
<tr>
<th>Species</th>
<th>Photosynthetic pathway</th>
<th>Density (plants per 0.25 m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter annuals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haplopappus gracilis</td>
<td>C₄</td>
<td>36.11</td>
</tr>
<tr>
<td>Ernogonum diffusum</td>
<td>C₃</td>
<td>30.65</td>
</tr>
<tr>
<td>Ernogonum abertianum</td>
<td>C₄</td>
<td>30.25</td>
</tr>
<tr>
<td>Descaria pinnata</td>
<td>C₄</td>
<td>4.07</td>
</tr>
<tr>
<td>Erodium cicutarium</td>
<td>C₁</td>
<td>3.77</td>
</tr>
<tr>
<td>Cryptantha micrantha</td>
<td>C₃</td>
<td>2.55</td>
</tr>
<tr>
<td>Gilia sinuata</td>
<td>C₄</td>
<td>2.01</td>
</tr>
<tr>
<td>Baileya multiflora</td>
<td>C₄</td>
<td>1.56</td>
</tr>
<tr>
<td>Malacothrix fendleri</td>
<td>C₃</td>
<td>1.53</td>
</tr>
<tr>
<td>Spermocephalum echinata</td>
<td>C₄</td>
<td>1.41</td>
</tr>
<tr>
<td>Plantago parishii</td>
<td>C₁</td>
<td>1.38</td>
</tr>
<tr>
<td>Trigernon diversum</td>
<td>C₁</td>
<td>1.34</td>
</tr>
<tr>
<td>Summer annuals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bouteloua aristidoides</td>
<td>C₄</td>
<td>6.92</td>
</tr>
<tr>
<td>Molugo verticillata</td>
<td>C₃</td>
<td>3.82</td>
</tr>
<tr>
<td>Ernogonum abertianum</td>
<td>C₄</td>
<td>3.60</td>
</tr>
<tr>
<td>Molugo cernua</td>
<td>C₃</td>
<td>2.46</td>
</tr>
<tr>
<td>Pectis papina</td>
<td>C₃</td>
<td>2.27</td>
</tr>
<tr>
<td>Aristida adscensionis</td>
<td>C₁</td>
<td>2.20</td>
</tr>
<tr>
<td>Bebraria spicata</td>
<td>C₁</td>
<td>1.34</td>
</tr>
<tr>
<td>Chenopodium fremontii</td>
<td>C₃</td>
<td>0.87</td>
</tr>
<tr>
<td>Euphorbia fendleri</td>
<td>C₃</td>
<td>0.84</td>
</tr>
<tr>
<td>Baileya multiflora</td>
<td>C₃</td>
<td>0.72</td>
</tr>
<tr>
<td>Tidestromia laungiflorens</td>
<td>C₁</td>
<td>0.65</td>
</tr>
<tr>
<td>Panicum arizonicum</td>
<td>C₁</td>
<td>0.44</td>
</tr>
</tbody>
</table>

*Developed from Guo and Brown (1996).

photosynthesis, 5 (Amaranthaceae, Chenopodiaceae, Cyperaceae, Poaceae, and Portulacaceae) have genera that include NADP-MAE and NAD-ME species. In grasses, Panicum has all three biochemical subtypes; among dicots, two genera (Salsola and Portulaca) have the NAD-ME and NADP-ME subtypes, as well as containing C₄ members.

The extent to which C₄ photosynthesis is polyphyletic has become apparent in recent phylogenetic surveys. In the Cyperaceae, for example, four distinct C₄ lineages are present, each of which has a characteristic anatomy (Fig. 12). Fourteen C₄ sedge genera fall into the subfamily Cyperaeae, including the large genus Cyperus, which is about 60% C₄. Five genera are in the subfamily Abildgaardieae and two genera are in the Rhynchosporaeeae. The C₄ plants in each of these subfamilies are NADP-ME, but show distinct anatomical characteristics indicating independent C₄ origin (Soros and Brubel, 2000). The fourth C₄ line in the Cyperaceae is represented by Eleocharis in the otherwise C₃ sub-family Scirpeae. Eleocharis is NAD-ME, and it also has a unique C₄ anatomy.

Current models indicate at least three independent origins for NADP-MAE C₄ grasses, in the subfamilies Aristideae, Eriachne, and Panicoidae (Kellogg, 1999). Recent reports that Orctilia in the Chloridoideae is NADP-ME indicate a fourth independent origin in the Poaeae (Keeley, 1998). The NAD-ME subtype originated independently in the Panicoidae and Chloridoideae, and may have multiple independent origins within these subfamilies, because in both there is considerable anatomical and cellular variation.

For reasons that are not known, the vast majority of C₄ families lack C₄ photosynthesis, even though many of them occur in environments where the C₄ pathway confers substantial benefits. Notably, most of the largest families of flowering plants lack C₄ photosynthesis. The orchids (Orchidaceae), rushes (Juncaceae), legumes (Fabaceae), and Rosaceae have not produced any C₄ taxa, despite sharing habitat and having close taxonomic affinity with some successful C₄ lines. For example, rushes and sedges are sister families, and herbaceous legumes are one of the few C₄ groups that routinely occur with C₃ grasses in low-latitude grasslands and deserts. Also for reasons unknown, few families contain both CAM and C₄ lines. Only five are known to have both CAM and C₃ genera, and only one of these, the Portulacaceae, has both CAM and C₄ species within the same genera (Table VIII). Only one species, Portulaca oleracea, is known to be both CAM and C₄, which is a sharp contrast to many CAM species that operate in the C₃ mode when water is abundant.

An important question is why C₄ photosynthesis is concentrated in a relatively small number of taxa. Are there characteristics that predisposed these groups to develop C₄ photosynthesis? Ehleringer et al. (1997) suggest that grasses and other species with close vein spacing are more likely to develop C₄ photosynthesis than species (dicots) with wider vein spacing. They believe that plants with close vein spacing may have a greater probability of establishing the C₄ pattern of metabolite transport between mesophyll and bundle sheath cells. The habitat of the C₃ progenitors to C₄ taxa is also an important consideration. C₃ grasses are common in arid landscapes of high light, and C₃ chenopods are common in saline habitats over much of the earth. Adaptations contributing to success in open, high-light, and high-stress habitats may have predisposed species of these families to evolve traits
that facilitated subsequent development of a CO₂ concentration system. In addition, the reduction in biotic competition often found in severe environments may have enabled many intermediate forms to evolve, with some eventually developing the complete C₄ pathway.

V. THE RISE OF C₄ PHOTOSYNTHESIS

A. When Did C₄ Plants Evolve?

Three lines of evidence are used to evaluate when C₄ photosynthesis first appeared. These are: (1) fossilized grasses with Kranz anatomy and other characteristics found in modern C₄ taxa; (2) carbon isotopic ratios from soils, plant remains, or herbivore fossils that demonstrate C₄ presence on the landscape; and (3) phylogenetic estimates of divergence times for current C₄ taxa. The first two lines of evidence demonstrate presence of C₄ plants on a landscape but do not demonstrate actual time of origins. Phylogenetic analyses potentially indicate exact times of origin, assuming the biomolecular clocks used in divergence estimates are accurate.

1. Fossil Evidence

The oldest known C₄ fossils come from the Ricardo formation of southern California, dated from the middle
TABLE VII
The Occurrence of C4 Photosynthesis in Higher Plant Orders and Families

<table>
<thead>
<tr>
<th>Group</th>
<th>Genera statistics</th>
<th>Species statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># C4</td>
<td>Total</td>
</tr>
<tr>
<td>Dicotyledoneae (subclass)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asteraceae</td>
<td>8</td>
<td>1500</td>
</tr>
<tr>
<td>Brassicales</td>
<td>1</td>
<td>34</td>
</tr>
<tr>
<td>Caryophyllales</td>
<td>5</td>
<td>126</td>
</tr>
<tr>
<td>Caryophyllaceae</td>
<td>12</td>
<td>88</td>
</tr>
<tr>
<td>Chenopodiaceae</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>Molluginaceae</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>Portulacaceae</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Euphorbiaceae</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>Linnaeae</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>Polygonaceae</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>Saxifragales</td>
<td>1</td>
<td>250</td>
</tr>
<tr>
<td>Scrophulariales</td>
<td>1</td>
<td>250</td>
</tr>
<tr>
<td>Total dicot</td>
<td>15</td>
<td>90</td>
</tr>
<tr>
<td>Monocotyledoneae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alismatales</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Juncales</td>
<td>28</td>
<td>131</td>
</tr>
<tr>
<td>Poales</td>
<td>372</td>
<td>~800</td>
</tr>
<tr>
<td>Total monocot</td>
<td>3</td>
<td>401</td>
</tr>
<tr>
<td>Total C4</td>
<td>18</td>
<td>491</td>
</tr>
</tbody>
</table>

*Genera statistics are the number of genera having C4 species, total number of genera in the family, and percentage of genera in the family with C4 members. Species statistics refer to total number of C4 species estimated for the family, total species in the family, and percentage of those species that are C4. The percentage of all C4 refers to the number of C4 species in a taxonomic group divided by 7600, the estimated total number of C4 species. Estimates are from Søeg et al. (1999a). A '+' after the species number indicates an underestimate.

Miocene at approximately 12.5 million years ago (mya) (Cerling, 1999). These fossils show distinct C4 Kranz-type leaf anatomy that is characterized by enlarged bundle sheath cells and reduced numbers of mesophyll cells. In addition, carbon extracted from the samples has isotopic ratios indicative of C4 species. The C4 grasses at this location appear panicle, and are apparently not common, as only a small number of samples have been found despite repeated searching (Tidewell and Nabudiri, 1989). The only other definitive C4 fossils known from pre-Pliocene deposits are from the Ogalla formation of Northwestern Kansas (U.S.A.), dated at 5 to 7 mya (Thomasson et al., 1986). These fossils show distinct C4 leaf anatomy, and in other features closely resemble extant species in the Chloridoideae, a nearly pure C4 grass subfamily.

Probable C4 plant remains have been found in the Fort Ternan formation of Kenya, dated at 14.5 mya (Cerling, 1999). Because internal tissues have been lost in these remains, only morphological and cuticular features are useful. These indicate that the plants were Chloridoideae and Panicoideae grasses. The Chloridoideae specimens match modern C4 genera, and one has been identified as Distichlis, a widespread C4 genus. Of the Panicoideae fossils, one matches the C4 genus Stenochloa, which is now common in East Africa (Dugas and Retallack, 1993).

2. Carbon Isotope Discrimination
In contrast to the sparse record for early C4 plant fossils, there is abundant isotopic evidence documenting a widespread radiation of C4 plants at the end of the Miocene period 5 to 8 million years ago. C4 photosynthesis discriminates less against heavy isotopes of carbon than does C3 photosynthesis, such that C4 tissues have 8 to 15 more 13C atoms per hundred thousand 12C atoms than do C3 plants (see Table II). Using mass spectrometry, these differences can be readily detected in plant remains, or any material that originated from photosynthetic carbon.
Thus, fossilized organic matter, soil carbonates, and animal remains can be assessed to determine whether $C_3$ or $C_4$ species were predominant on the landscape at the time of fossilization.

Both soil and herbivore remains from around the world show a clear shift in isotopic ratios of 5 to 10 per mil between 5 and 10 mya (Fig. 13). In East Africa and Pakistan, isotopic shifts are dramatic, indicating near-complete replacement of $C_3$ with $C_4$ vegetation in as little as 2 million years at many sites. At higher
TABLE VIII

<table>
<thead>
<tr>
<th>Family</th>
<th>Total genera</th>
<th>$C_3$ genera</th>
<th>CAM genera</th>
<th>$C_4$ and CAM genera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aizoaceae</td>
<td>126</td>
<td>5</td>
<td>47</td>
<td>0</td>
</tr>
<tr>
<td>Asteraceae</td>
<td>1500</td>
<td>8</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Euphorbiaceae</td>
<td>300</td>
<td>1*</td>
<td>4</td>
<td>0*</td>
</tr>
<tr>
<td>Hydrocharitaceae</td>
<td>16</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Portulacaceae</td>
<td>20</td>
<td>2</td>
<td>6</td>
<td>2*</td>
</tr>
</tbody>
</table>

* Compiled from Sage et al. (1999a) and Smith and Winter (1996).

* $C_3$ members of the Euphorbiaceae are grouped here into the genus Chaenomele. Some treatments place these species into Euphorbia, a genus with CAM and $C_4$ species.

* The $C_3$ and CAM genera are Anacamptis and Portulaca. Only Portulaca is known to have species that switch between $C_3$ and CAM modes.

latitude, substantial variation is present in isotopic data about after 7 million years, indicating partial replacement of $C_3$ with $C_4$ species.

3. Phylogenetic Approaches

Fossil and isotopic analyses can detect when $C_4$ plants are well represented in regions where preservation can occur, but they may not detect actual times of origin if the species remain rare or occur in places such as arid zones where preservation is unlikely. To overcome this problem, molecular phylogenies from extant genera can be used to identify divergent times between $C_3$ and $C_4$ taxa. This approach relies on predictable rates of divergence in the DNA, RNA, and protein sequences of related $C_3$ and $C_4$ material. Using this approach, $C_4$ photosynthesis is estimated to have emerged over 25 mya (Kellogg, 1999). Maize (Zea mays) appears to have diverged from Pennisetum by 25 mya, and from sorghum (Sorghum halopense) by 18 mya. Because all three species are NADP-ME grasses sharing a common $C_4$ ancestor, their divergence is believed to have occurred after the origin of the $C_4$ pathway.

In contrast to grasses, there is little evidence for the origins of $C_4$ dicots. $C_4$ dicot fossils are unknown, and phylogenetic divergence estimates have yet to be presented. Estimating divergence times for $C_4$ origin is particularly difficult for dicots, given the high number of distinct and apparently recent origins. Based on the low number of $C_4$ dicot taxa and multiple appearances, Ehleringer et al. (1997) speculate that $C_4$ dicots evolved more recently than $C_4$ monocots, with most origins possibly occurring within the past 3 million years.

All current families containing $C_4$ photosynthesis arose no earlier than the late Cretaceous (65 to 75 mya). Grasses and sedges appear at the end of the Cretaceous period, and become common during the Eocene epoch (58 to 36 mya). Most dicot families that now contain $C_4$ species appear in the fossil record between 50 to 20 mya, a time when most orders and families of angiosperms arose. The Chenopodiaceae/Amaranthaceae complex appears to be the oldest of the large $C_4$ dicot families, with the earliest pollen dating to the late Cretaceous about 68 mya. Early Chenopodiaceae fossils are associated with marine/terrestrial boundary zones, indicating saline habitats. Where fossil material is present, the first members of these families appear to be $C_3$ species, with $C_4$ species appearing in more advanced lineages (Kellogg, 1999).

The origin of the plant families that now contain $C_4$ species is associated with a global cooling phase that occurred during the middle Tertiary period between 50 and 20 mya. In low to middle latitudes, this climatic shift produced semi-arid landscapes that by the middle Miocene (15 mya) were supporting grasslands and savannas that possibly contained isolated swards of $C_4$ grass. Similarly, sedges, chenopods, and members of other families that eventually developed $C_4$ species became more common after 35 mya (Collinson et al., 1993).

B. Driving Forces for the Evolution of $C_4$ Species

On the basis of the taxonomic and phylogenetic evidence, a likely scenario is that $C_4$ photosynthesis originated between the middle Oligocene and early Miocene (between 35 and 18 mya) when global aridification became pronounced, while the subsequent expansion to dominance occurred during the late Miocene about 8 mya. The identification of these times is important, because it allows us to address hypotheses concerning the environmental factors that promoted the rise of $C_4$ photosynthesis.

The modern association of $C_4$ plants with hot, dry conditions has led to suggestions that $C_4$ photosynthesis evolved in response to climate warming and aridification. Though the postulated origin of $C_4$ photosynthesis in the middle to late Tertiary is consistent with palaeoclimatic shifts to drier conditions, the global climate then was in general cooler than in previous periods. During the Cretaceous period (145 to 65 mya), for example, the earth was considerably warmer than in the middle to late Tertiary, but as far as we know there were no $C_4$ plants. Aridification alone is an inadequate explanation for the evolution of $C_4$ plants, because significant arid regions also existed long before $C_4$ species arose.
Moreover, in hot climates, modern distributions indicate that aridity is not a prerequisite for C₄ photosynthesis, as long as disturbances are present to check the woody vegetation. During the Cretaceous period, dinosaurs would have provided substantial disturbance, and geological cores demonstrate that there was an abundance of fire.

One important parameter that changed during the period when C₄ plants evolved is the atmospheric level of CO₂. During the Cretaceous, atmospheric CO₂ levels were 5 to 10 times higher than at present, and as a result the potential was low for photorespiration in C₃ plants (Fig. 14). Since then, CO₂ levels have declined, and with this the potential for photorespiration has increased. The origin of C₄ species corresponds to the period when atmospheric CO₂ levels are predicted to have fallen to the point where photorespiratory potentials at 30°C rise above 20% of photosynthesis (see Fig. 14). This reduction in CO₂ is proposed to have been the driving force for the origin of C₄ species.

A CO₂ control over the timing of C₄ origins is attractive because it unifies the various lines of evidence concerning C₄ evolution. Physiological studies demonstrate no major advantage of C₄ photosynthesis at elevated CO₂; instead, the extra energy cost of C₄ photosynthesis should be detrimental except at light saturation. Both growth and photosynthesis measurements predict a shift from C₃ to C₄ dominance at CO₂ levels modeled to have existed during the late Miocene when C₄ species expanded across tropical landscapes. Aridity
likely played a contributing role because stomatal closure to conserve water also reduces the supply of CO₂ for photosynthesis.

Where did the first C₄ species evolve? Based on physiological models and current biogeographic patterns, it is likely that C₄ photosynthesis first appeared in hot, arid, and saline situations. Such environments now have the greatest diversity of C₄ taxa, with a wide range of dicot and monocot species, including most woody C₃ species. As CO₂ levels declined during the Tertiary period, these settings would have been the first to favor CO₂ concentration systems, presumably because the combination of heat, drought, and salinity would have provided the greatest inhibition of nearby C₃ competitors.

VI. CLIMATE CHANGE AND THE FUTURE OF C₄ PHOTOSYNTHESIS

If C₄ species arose because of past global changes that favored photosynthetic CO₂-concentrating mechanisms, then it follows that anthropogenic increases in CO₂ would reduce the advantage that C₄ species had in the low-CO₂ atmospheres of recent geological time. Cooler, more temperate environments would be affected first, with C₄ species in warm, arid habitats being the least affected. It is unlikely, however, that widespread extinction of C₄ species will result from atmospheric CO₂ rise.

First, global warming accompanying CO₂ rise will offset advantages that the CO₂ increase might confer upon C₃ plants. C₄ species are generally better adapted to warmer climates than potential C₃ competitors, and thus will be in a better position to exploit additional warmth. Second, feedbacks are present in plants and ecosystems that constrain responsiveness to high CO₂. Central among these is nutrient availability. If nutrients are deficient, the ability of C₃ species to exploit CO₂ enrichment is substantially reduced, if not lost altogether. Most natural ecosystems are nutrient deficient, particularly those dominated by C₃ species.

Third, a common limitation in ecosystems where C₄ species occur is water. In water-limited situations, productivity of C₃ and C₄ species is enhanced by rising CO₂ because stomata of both close as atmospheric CO₂ rises, leading to water savings and longer growing seasons. The severity of drought is often greater during warm summer months when evaporative demand is greatest and C₃ species are more active. Thus, positive responses to high CO₂ are likely to be greater in summer-active C₄ species rather than the spring-active C₃ species from the same community. Warmer temperatures and more frequent aridity will also promote fires that strongly favor C₄ species because of negative effects on the woody C₃ vegetation.

Fourth, changes in seasonality will be critical to future C₃/C₄ dynamics. Where winters become milder, yet remain cool, C₃ species could be favored. Such a pattern has been observed in central Canada, where a recent rise in early spring temperatures has been associated with increased C₃ grass cover at the expense of C₄ species. Increased summer precipitation in arid regions would favor C₄ species by providing necessary moisture for growth during the high-temperature period favorable to C₄ species. Thus, Mediterranean zones that develop a summer monsoon could see a dramatic rise in the cover of C₄ species.

Because past climate change has been associated with CO₂ rise, paleoecological studies indicate how future C₃/C₄ dynamics might develop. Between 20,000 and 10,000 years ago, regional climates warmed 5° to 10°C and CO₂ levels rose 30%, thereby providing a similar situation to that predicted to occur in coming centuries. Theoretically, the CO₂ rise should have had a dramatic effect on C₃ relative to C₄ distributions, because the low-CO₂ level of 20,000 years ago (180 ppm) would
have been far more limiting for the performance of C\textsubscript{3} relative to C\textsubscript{4} plants. Paleocological studies indicate that in tropical climates, there was a wider distribution of C\textsubscript{4} species 20,000 years ago relative to 10,000 years ago and today. In central Africa and South America, isotopic and pollen data indicate that grasslands were more widespread 20,000 years ago than at present, which is consistent with a low-CO\textsubscript{2} effect favoring C\textsubscript{4} plants. However, these regions were drier than at present, which could also explain the pattern.

In temperate zones, little CO\textsubscript{2} effect on C\textsubscript{3} versus C\textsubscript{4} plant distribution is apparent during this time. In southwestern North America, for example, the distribution of C\textsubscript{3} and C\textsubscript{4}-dominated grass and shrub vegetation appears to remain stable over the past 20,000 years, possibly because the seasonality of precipitation has also changed little over this time (Connin et al., 1999). In contrast to the American Southwest, substantial expansion of C\textsubscript{4} grasses occurred at the end of the last ice age on the plains grasslands of central North America. Here, C\textsubscript{4}-dominated grasslands expanded northward between 12,000 and 8000 years ago, a time of little CO\textsubscript{2} change, but considerable climate warming.

VII. C\textsubscript{4} VEGETATION AND GLOBAL BIODIVERSITY

A. C\textsubscript{4} Grasses and Tropical Deforestation

Through the use of fire, humans have been major agents of vegetation change for over 2 million years. The most significant impact of burning has come in the past few thousand years, when humans increasingly used fire to maintain and expand agricultural lands. As a result, many of the grasslands and savannas of the world are derived from human-caused fires. In recent decades, expansion of C\textsubscript{4}-dominated grasslands has become pervasive in tropical regions, and in the process has become one of the more important threats to tropical biodiversity. This expansion has largely been driven by human population and economic growth, but the introduction of alien C\textsubscript{4} grasses into heavily impacted systems has promoted synergistic effects that often convert high-diversity forests to low-diversity grasslands.

Historically, agricultural practices in tropical regions consisted of shifting cultivation, in which forest plots would be slashed and burned, and the resulting open ground farmed for a few years and then abandoned. Plots were generally small enough and surrounding forest large enough to restrict the access of weedy grasses while providing large numbers of propagules from the nearby woodland. This ensured that the plots would readily undergo succession back to forest. In recent decades, however, higher human populations, technological improvements such as the use of chain saws, and high demand for timber and beef exports have promoted the widespread exploitation of tropical landscapes. As a result, the scale of land clearing has become far more extensive, with open areas favorable to C\textsubscript{4} grasses being produced at an increasing rate, while forests are becoming increasingly fragmented and influenced by the drier microclimates of the grassland areas. Coupled with these trends in land exploitation has been the introduction of cultivated or weedy C\textsubscript{4} grasses that are well adapted to human activities. In particular, aggressive growth of the weedy invaders following human disturbance has caused shifts in fire cycles, soil biogeochemistry, and food webs so that succession back to forest is often arrested and low-diversity grasslands become permanent fixtures on the landscape.

Numerous activities cause the deforestation of tropical forests, but three stand out as the major causes for forest decline. In each case, C\textsubscript{4} grasses play a key role.

1. Direct Clearing of Forest for Pasture

Forest clearing for pasture is particularly pervasive in Amazonia, where demand for low-cost beef is driving the expansion of the cattle industry into forested regions. Typically, primary forest with plant species numbering in the hundreds to thousands per hectare are converted to managed pastures with less than 10 plant species per hectare, most of which are economically useful C\textsubscript{4} grasses and their weedy associates.

2. Runaway Slash and Burn Agriculture

As human populations in forested areas have grown, slashed and burned plots have become larger and rotation times shorter. Invasive C\textsubscript{4} weeds become severe problems as these farming plots age, contributing to their abandonment after a few years. These weedy grasses are tough competitors, not only because they have high production rates, but also because their leaves decompose slowly due in part to having higher carbon-to-nitrogen ratios than C\textsubscript{3} leaves. Slow decomposition reduces soil nitrogen availability, thereby favoring the C\textsubscript{4} species that have higher nitrogen-use efficiency.

Many of the weedy C\textsubscript{4} invaders are pyrophytic; not only do they survive fire well, but they rapidly produce a dense canopy that readily dries to produce a highly flammable fuel. During dry episodes, the grass swards rapidly burn, producing fires that are lethal not only for forest seedlings establishing in the swards, but also...
for adjoining patches of forest. Once established, exotic grass swards will typically burn every 2 to 5 years. With each burn, more forest is destroyed, and fewer remnant forest species are present in the grassland. After a few cycles of burning, the forest seed bank is destroyed, and forest patches are too isolated to contribute to the seed rain over most of the grassland. At this point, the grassland may become a permanent fixture on the landscape.

Compared to the original diversity of hundreds to thousands of plant species per hectare, weedy grasslands are dominated by a few species of C₄ grasses, although they may include up to a few dozen species of minor herbaceous plants. In the extreme, weedy grasses can form dense stands that exclude all other plant species. In Southeast Asia and Indonesia, the C₄ grass *Imperata cylindrica* has become a widespread pest because of its ability to rapidly colonize wet soils cleared for farming or grazing, after which it forms highly flammable swards that kill all other species through crowding or burning. In Central and South America, the C₄ grasses *Melinis minutiflora*, *Panicum maximum*, and various *Pennisetum* species are causing loss of regional diversity for similar reasons.

3. Deforestation following Selective Logging of Tropical Forests

The demand for wood products and fuel has led to extensive logging operations throughout the wet tropics. Much of this logging is selective in nature, causing the forest to become disturbed but not destroyed. Although a selectively logged forest can succeed back to closed-canopy stands within a century, any logging increases the probability that they will burn owing to the increased level of downed woody debris (Fig. 15). Once burned, the level of woody debris increases as vegetation killed by the fire falls to the ground. This promotes a second fire within a few years, particularly if grasslands are nearby. Each fire tends to be more intense than the previous because of the increased level of woody debris, and each fire increases the probability of a subsequent burn. Critical to this cycle is the establishment of pyrophytic C₄ grasses in forest margins and gaps caused by repeated burning. These further accelerate fire cycles because the microclimate over the grass swards is warmer and less humid, thereby drying adjoining forests, and the grasses produce an abundance of fuel that carries fire into forest fragments.

The C₄ grass species contributing to biodiversity impoverishment across the globe are relatively small in number, and are disproportionately African in origin (Table IX). Part of the reason why African species are so successful is because they have had a long time to adapt to human activities. In contrast, grasses in the Americas and on oceanic islands have been exposed to humans for less than 1,400 years, and to intensive grazing regimes for only a few hundred years.

In addition to adaptation to humanity, unique features arising from C₄ photosynthesis help explain the expansion of these grasses at the expense of native species. C₄ photosynthesis promotes the high levels of primary production that make C₄ crops and forages so valuable to humans in warm tropical climates. This high productive potential enables C₄ grasses to rapidly establish in tropical locations relative to their C₃ competitors, such that they become the preferred pasture species, as well as the most aggressive weeds. The high water-use efficiency of C₄ grasses enhances tolerance of high evaporative demand and periodic drought that develops rapidly in tropical settings, typically after only a few days of no rain. The high nitrogen-use efficiency of C₄ grasses provides for an ability to produce greater amounts of biomass on infertile soils. Soil infertility often follows forest clearing because of leaching during heavy rainfall, nitrogen volatilization during fire, and nutrient exhaustion during cultivation. Importantly, the combination of these features allows C₄ grasses to rapidly produce large quantities of biomass that will soon provide fuel to promote intense, destructive fires.

B. Impoverishment of C₄-Dominated Landscapes

While certain C₄ species that are well adapted to human activity have greatly expanded their range in recent decades, most natural C₄-dominated ecosystems have been severely impacted by human activity. Many of the temperate C₄ grasslands of the world have already been destroyed, largely by human exploitation for agricultural production, fire suppression, or conversion to managed timber plantations. In North America, about 90% of the once widespread tallgrass prairie biome has been converted to row crop agriculture, forest, or pasture. Where stands of bluestem and switchgrass (*Sichotychium*, *Andropogon*, and *Panicum* spp.) once formed an ocean of C₄ grass from Illinois to Kansas in the central United States, maize and soybean fields now stretch to the horizon. In the Southeast, a wiregrass/long-leaved pine savanna once extended across as much as 60% of the coastal plain from Virginia to Texas. Now, about 93% of this ecosystem has been lost to timber plantations, crop fields, and pasture (Drew et al., 1998). In the western steppes and deserts, overgrazing coupled with climate change has contributed to conversion of
FIGURE 15 A conceptual model of forest conversion to grassland following logging. Dashed lines indicate recovery times if fire is excluded from the system. Arrested succession occurs by repeated burning, loss of soil fertility, and/or loss of the forest seed bank. (From Cochrane and Schulze, 1999, with modification to emphasize the role of C₄ grasses.) Copyright 1999 by the Association for Tropical Biology, P.O. Box 1897, Lawrence, KS 66044, USA. Reprinted by permission.

TABLE IX
Important Invasive C₄ Grasses of Natural Landscapes of the World*  

<table>
<thead>
<tr>
<th>Invaded region</th>
<th>Invaders</th>
<th>Region of origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>North America</td>
<td>Microstegium vincentum</td>
<td>Asia</td>
</tr>
<tr>
<td></td>
<td>Eragrostis curvula, Eragrostis lehmanniana, Cynodon dactylon, Digitaria spp., Echinochloa spp., Pennisetum spp., Sorghum halepense</td>
<td>Africa</td>
</tr>
<tr>
<td>Tropical America</td>
<td>Brachiaria spp., Hyparrhenia rufa, Melinis minutiflora, Panicum maximum</td>
<td>Africa</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>Imperata cylindrica</td>
<td>Africa</td>
</tr>
<tr>
<td>Australia</td>
<td>Cenchrus ciliaris, Eragrostis curvula, Melinis minutiflora, Pennisetum polystachyon</td>
<td>Africa</td>
</tr>
<tr>
<td>Oceania (Hawaii)</td>
<td>Andropogon virginicus, Schizachyrium condensatum</td>
<td>North America</td>
</tr>
<tr>
<td></td>
<td>Cenchrus ciliaris, Digitaria decumbens, Hyparrhenia rufa, Imperata cylindrica, Melinis minutiflora, Pennisetum clandestinum, Pennisetum purpureum</td>
<td>Africa</td>
</tr>
<tr>
<td>Europe, New Zealand</td>
<td>Spartina anglica (+ S. townsendii) in coastal salt marshes of Europe (this is a novel hybrid between S. alterniflora from the Americas and S. maritima from Europe)</td>
<td>Europe</td>
</tr>
</tbody>
</table>

* Reprinted from Sage et al. (1999b), with permission.
much of the arid \( \text{C}_4 \) grasslands to \( \text{C}_3 \) shrublands. The desert grasslands of Texas and New Mexico have been particularly affected, with most of the original grassland now converted to mesquite (Prosopis) and Acacia scrub. Although half of the shortgrass prairie in the high plains of the central states has been converted to dryland or irrigated farms, large expanses remain intact, but often in a degraded condition due to overgrazing.

Worldwide, similar patterns hold. Most productive areas once dominated by \( \text{C}_4 \) grasslands have already been developed into cropland, pasture, or timber plantations, while less productive areas are routinely subjected to unsustainable levels of grazing. Recent estimates suggest that 72–74% of the rangelands of Africa, Asia, and the Americas are in a moderate to very severe degradation category; in Australia, 55% of the rangelands are degraded (McNeely et al., 1995). Most of these rangelands once supported extensive \( \text{C}_4 \) grasslands. Land renewal efforts are under way in many of these areas, but the main focus is to establish a few productive species for pasture or wood production, rather than restoration of native \( \text{C}_4 \)-dominated ecosystems. Although this process differs from that driving deforestation, the ultimate cause—unsustainable exploitation by humanity—and the end result—species impoverishment—remain the same.

**VIII. SUMMARY**

\( \text{C}_4 \) photosynthesis is a polyphyletic solution to challenges imposed by reductions in atmospheric \( \text{CO}_2 \) and increasing aridification that occurred over the past 50 million years. In \( \text{C}_4 \) plants, the low atmospheric \( \text{CO}_2 \) levels of recent geological time promoted substantial inhibition of photosynthesis by photorespiration in hot conditions, particularly if water was also limiting. \( \text{C}_4 \) plants overcame these limitations by localizing Rubisco in the bundle sheath compartment into which \( \text{CO}_2 \) is concentrated. As a consequence, \( \text{C}_4 \) plants are highly productive, aggressive competitors under conditions promoting photorespiration in \( \text{C}_3 \) plants. This improved performance has enabled native \( \text{C}_4 \) species to dominate hot, open landscapes of low latitudes, and warm temperate regions receiving summer rain. It has also enabled invasive \( \text{C}_4 \) grasses adapted to human disturbance to become severe weeds, with important consequences for tropical biodiversity. Where humans disrupt forest cover and allow grass establishment, they establish conditions where invasive \( \text{C}_4 \) grasses can accelerate fire cycles and alter soil chemistry. In extreme cases, diverse tropical forests can be replaced by grasslands dominated by a few aggressive \( \text{C}_4 \) species. Because natural \( \text{C}_4 \)-dominated grasslands and savannas often occur on more productive landscapes, they too have been heavily exploited for agricultural purposes, resulting in habitat loss and impoverishment of the native \( \text{C}_4 \) flora of the planet.

**See Also the Following Articles**

- CARBON CYCLE • DEFORESTATION • DESERT ECOSYSTEMS • MEDITERRANEAN-CLIMATE ECOSYSTEMS • NITROGEN AND NITROGEN CYCLE • PHOTOSYNTHESIS, MECHANISMS OF • SLASH AND BURN FARMING

**Bibliography**


for CO₂ uptake among C₃ and C₄ plants. Plant Physiol. 73, 555–559.


I. Need for Captive Breeding and Reintroduction Programs
II. General Aspects of Captive Breeding and Reintroduction Programs
III. Difficulties with Captive Breeding Programs
IV. Genetic and Demographic Management of Captive Populations
V. Reintroducing Captive-Bred Animals to the Wild

GLOSSARY

allele One of two or more alternative forms of a gene.
effective population size The size of an ideal population that would have the same rate of increase in inbreeding or decrease in genetic diversity as the actual population under consideration. The effective population size is usually much less than the actual population size.
founder Wild-caught individual that contributes genetically to the captive population.
founder genome equivalent The number of equally contributing founders that would have produced the same genetic diversity found in an existing captive population if there had been no random loss of founder alleles.
genetic drift The variation of allele frequency from one generation to the next that occurs due to chance. Genetic drift leads to the loss of genetic variation in small populations due to the random loss of founder alleles during reproduction.
heterozygosity Average proportion of loci that are heterozygous (have two different alleles in an individual) in a population.
ideal population A hypothetical population widely used in population genetics theory. In this ideal population, the breeding sex ratio is equal, mating is random, generations do not overlap, selection and mutation do not occur, and the lifetime number of offspring produced by individual parents has a Poisson distribution.
mean kinship This value, calculated for every living member of a captive population, is the average kinship between that individual and all members of the population (including itself). Typically, living founders are excluded in the calculation of mean kinships. A population’s average mean kinship is the average of the mean kinships of all the individuals in the population.
reintroduction Releasing individuals of a species into an area where that species no longer occurs in an effort to reestablish a wild population. Reintroduced individuals may be captured from a healthy wild population in another area or may be derived from a captive population if there are no healthy wild populations remaining.
studbook A list of all the living and dead individuals in a captive population that contains information on the mother, father, date of birth, location, and other topics for each individual.
CAPTIVE BREEDING PROGRAMS enable zoos to exhibit many species of animals without capturing new individuals from the wild. Furthermore, the establishment of a captive breeding and reintroduction program may be the only hope for preserving a species that has been reduced to a very small number of individuals. Captive breeding programs are also used to support research and conservation education. A substantial research effort is often necessary to develop successful methods for raising a particular species in captivity and reintroducing it to the wild because these methods tend to be specific to a single species or a group of related species. However, methods for the demographic and genetic management of captive populations are similar across species. A captive breeding and reintroduction program should always be part of a larger conservation program for a species that also addresses the problems facing the species in the wild. Captive breeding and reintroduction programs require sustained long-term, adequately funded efforts and success is far from guaranteed.

I. NEED FOR CAPTIVE BREEDING AND REINTRODUCTION PROGRAMS

Captive breeding is the only choice for species that are extinct or nearly extinct in the wild. Nearly one-fourth of mammal species, 11% of bird species, and 25% of reptiles are threatened with extinction, according to the 1996 IUCN (World Conservation Union) List of Threatened Animals. Because most at-risk species are threatened by habitat loss and degradation, the most common conservation need is habitat protection and improvement. Captive breeding and reintroduction programs play a minor conservation role in comparison to protecting and improving habitat.

The need for captive breeding will undoubtedly increase because we are only just past the midpoint of the human population explosion, according to United Nations (UN) projections. The human population increased by a record-breaking 2 billion people in the past 25 years to a total of more than 6 billion, and the UN projects an increase of another 2 billion during the next 25 years with a total population of nearly 10.4 billion by 2100.

Due to limited space, staff, and funds, zoos will not be able to preserve populations of all animal species likely to become extinct in the wild. In the United States alone, the American Zoo Association (AZA) already maintains studbooks for 349 species. Of these, 118 species are managed under 89 interzoo captive breeding programs known as species survival plans (SSPs). The 2005 participating in these plans move individuals of the species concerned to other zoos as necessary to meet genetic and demographic goals specified in the plan. Of the remaining 231 species, 103 are covered by less formal population management plans. AZA anticipates that at least an additional 136 studbooks and 18 SSPs will be organized within the next 3 years. At least one-third of the SSPs are involved with reintroduction efforts. Some species may be maintained in captivity for long periods without the possibility of reintroduction. For example, Père David’s deer and the Mongolian wild horse survived in captivity many decades after their extinctions in the wild before the possibility of reintroducing them was even considered. Ultimately, a stable or decreasing human population with more sustainable patterns of resource use may allow the reintroduction of some species that have been preserved only in captivity.

II. GENERAL ASPECTS OF CAPTIVE BREEDING AND REINTRODUCTION PROGRAMS

The major goal of most captive breeding programs is to develop self-sustaining captive and/or wild populations. Even if a species is never reintroduced, a successful captive breeding program will supply zoos with animals to exhibit, thereby minimizing the need to collect them from the wild. Captive breeding programs also have considerable educational value because they are used to inform zoo visitors of the value of conserving biodiversity and to increase public interest in conservation issues. Animals maintained in captive breeding programs also support a variety of research programs. For example, some aspects of human medicine, human evolution, and the distinctiveness of the human genome require comparison of human DNA with chimpanzee and gorilla DNA, samples of which are normally obtained from captive populations.

Several organizations provide help with captive breeding and reintroduction efforts. The AZA coordinates captive breeding efforts in the United States and similar organizations exist in other countries. The AZA maintains two committees that give general advice on population management and reintroduction: the Small Population Management Advisory Group and the Rein-
Captive Breeding and Reintroduction

III. DIFFICULTIES WITH CAPTIVE BREEDING PROGRAMS

A. Failure to Breed in Captivity

Not all species breed well in captivity. The species that are easiest to breed in captivity are those that have management requirements similar to those for domestic animals or for other species with which zoos have had considerable experience. For example, zoos suddenly faced with the challenge of breeding California condors experienced little difficulty because they had been breeding Andean condors for many years. However, species with which zoos have had little prior experience may breed very poorly at first until zoos develop appropriate husbandry techniques. Thus, because husbandry techniques tend to be species specific, new captive breeding programs often require substantial research programs on behavior, reproductive biology, nutrition, genetics, or disease. Research on closely related species is also often helpful.

Poor reproduction in captivity is often due to behavioral problems caused by inadequate husbandry techniques. Because different zoos often have different degrees of success in breeding a particular species, important insights can often be gained by comparing the behavior and reproductive success of individuals kept under different conditions at different zoos. For example, a collaborative cross-zoo study indicated that more submissive male black rhinoceroses tended to sire more young and that reproductive success increased when the species was kept in larger cages. Space limitation or crowding may increase aggressiveness in male black rhinoceroses and therefore may reduce reproductive success. This suggests that dominant, aggressive males should be housed in zoos with large enclosures.

More invasive research work can also make important contributions. For example, studies on black-footed ferret reproductive biology have improved captive breeding techniques for this species. Until recently, only about 58% of mated black-footed ferret females produced young. Many of the other females became pseudopregnant, which means they underwent the hormonal changes typical of pregnancy without actually being pregnant. Now researchers know that yearling males do not produce sperm until 4–6 weeks later than older males. Many of the pseudopregnancies were the result of breeding these yearling males. Ensuring that males were actually producing sperm before using them for breeding has reduced the number of pseudopregnant females by about 20%.

B. Genetic and Behavioral Adaptation to Captivity

A captive population’s risk of extinction is increased by inbreeding and loss of genetic variation, and inbred animals with little genetic variation are less likely to survive when reintroduced to the wild than more outbred individuals. However, both inbreeding and loss of genetic variation are unavoidable in small, closed populations because all individuals in the population eventually become related to each other. A randomly mating small population loses genetic variation (heterozygosity) at a rate equal to

\[
\frac{1}{2N_e}
\]

each generation, where \(N_e\) is the effective population size.

Captive populations are routinely managed to minimize inbreeding and loss of genetic variation. Another threat to captive populations is that selective pressures in captive habitats are different than those in wild habitats; thus, captive populations tend to adapt to captivity, which can make it more difficult to reestablish a population in the wild.

Learned behavioral traits can degenerate in captivity more rapidly than genetic diversity. Traits that may degenerate in the captive environment include foraging skills, detection and avoidance of predators, and fear of humans. When captive-bred and wild-born individuals have been experimentally released in the same location, the captive-bred individuals have tended to survive for shorter periods due to lack of appropriate behaviors.
IV. GENETIC AND DEMOGRAPHIC MANAGEMENT OF CAPTIVE POPULATIONS

In contrast to husbandry methods, genetic and demographic management methods are similar for all captive populations. Genetic and demographic management of captive populations focuses on maintaining genetic diversity in order to minimize undesirable genetic changes due to selection in the captive environment, avoid deleterious effects of inbreeding depression, and maintain future options for genetic management.

A. Starting the Program

Ideally, the first step in the development of a captive breeding and reintroduction program is agreement among all concerned parties, such as agency personnel, nongovernmental conservation groups, and outside scientific advisers, that such a program would benefit a particular species. This step may be difficult because some people oppose captive breeding in general and because there are no precise guidelines for when to begin a captive breeding program. However, the IUCN recommends starting a captive population long before the wild population reaches the critical state in order to increase the probability that the captive breeding program will be a success.

Once in place, captive breeding programs have three phases. In the founding phase, the captive population is started. In the growth phase, the population rapidly increases to the final “target” population size specified by its managers. In the carrying capacity phase, the population is maintained at its target size and excess individuals may be reintroduced into the wild (Fig. 1). Management concerns change as the captive population progresses through these phases.

B. Management Concerns during the Founding Phase

The main management concerns during the founding phase are removing individuals for the captive population with minimal impact on the wild population, acquiring enough founders from the wild to achieve genetic goals, getting the species to breed reliably in captivity, and setting general goals and plans for the captive population. Ways to capture animals for the captive population include removing eggs from nests, using orphaned or injured animals, and capturing dispersing juveniles. Many birds (e.g., condors) will usually lay another egg to replace an egg that has been removed, and dispersing juvenile mammals often have high mortality rates in the wild. If the species is one that zoos do not know how to breed reliably in captivity, it is best to solve husbandry problems with only a few wild individuals or even animals of a closely related species.

Genetic goals for a captive population are usually specified in terms of the proportion of genetic variation (measured as heterozygosity) to be maintained for a specified time. A common goal is to maintain 90% of the genetic diversity of the source population for 100 years. However, some programs use other time frames. For example, the Guam rail and black-footed ferret programs are using the goal of “90% for 50 years” because of the short generation times for these species (Table I) and plans for the rapid establishment of several wild populations.

Once a genetic goal has been set, population genetics theory enables calculation of the number of founders needed for the captive population (the number of wild animals that must be captured and successfully bred) and the target population size (the number of individuals that must be maintained in captivity during the planning period). Planning to retain a higher proportion of genetic variation usually increases the target population size. For example, maintaining 90% of the ferret genetic variation for 50 years requires a population of only 300 individuals, whereas maintaining 92% would require a population of 2700. Increasing the length of
The goals and number of founders of captive breeding programs with reintroduction components

Table I

<table>
<thead>
<tr>
<th>Species</th>
<th>California condor</th>
<th>Black-footed ferret</th>
<th>Guam rail</th>
<th>Golden lion tamarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterozygosity goal (%)</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Length of program (years)</td>
<td>200</td>
<td>50</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>Number of generations</td>
<td>10</td>
<td>20</td>
<td>22</td>
<td>33</td>
</tr>
<tr>
<td>Target population size</td>
<td>150</td>
<td>500</td>
<td>150</td>
<td>550</td>
</tr>
<tr>
<td>Number wild-caught</td>
<td>14</td>
<td>18</td>
<td>21</td>
<td>69*</td>
</tr>
<tr>
<td>Number of contributing founders</td>
<td>13</td>
<td>10</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>Founder genome equivalences</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>12</td>
</tr>
</tbody>
</table>

*Species are listed in order of increasing number of generations in the program length (from Ralls and Ballou, 1992).

| Heterozygosity goal, program length, and target population size have not been officially adopted by program managers, other data from Kieler (1991).
| Founders with currently living descendants.
| The number of theoretically ideal founders taking into consideration loss of genetic diversity in the current captive population (Lacy, 1989).
| Includes the number of wild-caught tamarins acquired after the captive program was initiated in 1981 in addition to the number of founders and wild-caught individuals alive at the initiation of the program.

The planning period has a similar effect. For example, maintaining 90% of the ferret genetic variation for 100 instead of 50 years would increase the target population size from 500 to 1300 individuals.

Increasing the number of founders reduces the size of the target population needed to reach a particular goal. Twenty to 30 unrelated individuals are generally sufficient and reduce the cost of a captive breeding program. Unfortunately, many existing captive breeding programs have begun after the species' rate of reproduction and generation length was already too late to achieve this many founders. For example, the ferret population had only 10 founders. If it had been possible to obtain 25 founders, the target population size could have been reduced from 500 to 200 individuals. Although a small number of founders reduces the probability that a captive breeding program will be successful, it does not doom it to failure. Thus, the lack of an ideal number of founders does not justify abandoning or failing to initiate a captive breeding program.

The target population size also depends on the species' rate of reproduction and generation length. A smaller target population will be required to reach the genetic goal if the species can grow more rapidly each generation or if it has a longer generation time (because genetic variation is lost due to genetic drift each time individuals reproduce). The target population size may also be limited by practical considerations, such as the number of spaces available in zoos. Fewer zoos may be willing to participate in the program if the species is not attractive as an exhibit. Thus, the target population size may be a compromise between genetic and demographic factors and the limited resources available.

Once husbandry problems have been solved and the species is breeding well in captivity, the rest of the founders should be obtained as soon as possible. Unfortunately, the number of animals that must be captured from the wild is usually greater than the number of founders needed. Wild-caught animals may be related or fail to breed, or their descendants may fail to reproduce. For example, although 25 wild black-footed ferrets were captured, some died of distemper, some were known to be parents and offspring, and some failed to reproduce. Although ideally each founder will contribute an equal number of offspring to the captive population, those ferrets that did reproduce did so unequally, skewing their genetic contributions to the captive population. Ultimately, the ferret population was founded by the theoretical genetic equivalent of only 5 ferrets; that is, 5 founder genome equivalents.

C. Management Concerns during the Growth Phase

Management efforts during the growth phase center on getting the population to increase as rapidly as possible. Rapid growth has two benefits: It increases the captive population's chances of survival and it retains as much of the founders' genetic diversity as possible. Small captive populations are at higher risk of extinction due to many factors, including random demographic events.
CAPTIVE BREEDING AND REINTRODUCTION

(such as a succession of male births), inbreeding depression, and unpredictable events that can kill numerous individuals such as diseases, fires, hurricanes, and other catastrophes.

The standard SSP breeding strategy used in the United States is designed to maximize the retention of genetic diversity. This is accomplished by minimizing mean kinship among the members of the captive population (Ballou and Lacy, 1995). Breeding pairs are formed based on mean kinship, beginning with the individuals with lowest mean kinships, until the desired number of pairs is attained. Efforts are also made to avoid mating closely related individuals when forming new pairs. During the growth phase, this strategy is modified slightly to choose new pairs to minimize mean kinship as much as possible but breed all individuals in the population.

Although managers attempt to minimize mean kinship and inbreeding during the growth phase, rapid population growth takes priority over genetic concerns, particularly when the population is very small and the risk of extinction outweighs the risk of a few less-than-ideal matings. For instance, if a female rejects the genetically ideal mate, she may be allowed to mate with another male she prefers. This problem has occurred several times in the California condor population.

At some point during the growth phase, the captive population usually is divided into subpopulations housed in different breeding facilities. This reduces the risk that a catastrophe such as disease or fire will decimate the entire captive population. To ensure that each subpopulation is as genetically diverse as possible, each should have individuals descended from each of the founders.

D. Management Concerns during the Capacity Phase

Once the population has reached the target size, relatively few offspring may be needed each year to maintain it at that level. Thus, genetic concerns become more important and managers select which individuals to breed.

The number of offspring needed to maintain the captive population can be calculated by standard demographic techniques. Any “extra” offspring can be used for reintroduction. If there are more offspring than are needed for reintroduction, managers can prevent some adults from breeding either by using contraceptives or by housing them individually or in same-sex groups. For example, the captive golden lion tamarin population is now at its target size of about 350 individuals. Because tamarins often have twins or even triplets, these 350 individuals could potentially produce more than 400 offspring per year. However, only about 80 per year are needed to maintain the captive population and augment wild ones (the number that can be reintroduced to Brazil is limited by funding and habitat availability). Consequently, each year approximately 100 female tamarins are prevented from breeding by being housed with another female or a nonreproductive male or by using hormonal contraceptive implants.

There are two general strategies for producing the individuals to be reintroduced. If the date of a reintroduction effort can be scheduled well in advance and the species has a predictable breeding pattern (such as breeding once a year during the spring), males and females can be paired up for the specific purpose of producing excess young for that particular reintroduction. This approach has been used for Guam rails. However, if the date of a reintroduction effort is difficult to predict in advance (this may occur due to difficulties with funding or permits), animals for reintroduction can be selected from the existing population and breeding pairs can be set up to replace the reintroduced individuals.

In the early stages of a reintroduction program, reintroduction techniques are still being refined and mortality may be high. Thus, initially the most genetically expendable individuals are usually released. Later, emphasis will gradually shift to choosing individuals that are not closely related to the individuals already present in the wild population. This maximizes the genetic diversity of the wild population. The final genetic goal is to make the wild population as genetically diverse as the captive population.

V. REINTRODUCING CAPTIVE-BRED ANIMALS TO THE WILD

Ideally, the goals of all captive breeding plans would include reintroduction back to the wild. However, some species may be impossible to reintroduce due to lack of habitat or other problems. Furthermore, some species will be easier to reintroduce than others. Mark Stanley Price, former chairman of the IUCN Reintroduction Specialist Group, suggests that the species that may be easiest to reintroduce include large species with few predators, those living in herds or other social groups, those that tend to explore new habitat, and those that can tolerate a wide range of habitat conditions.
The Elements of a Successful Captive Breeding and Reintroduction Program

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>CAPTIVE BREEDING AND REINTRODUCTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ongoing research in behavior, genetics, physiology, nutrition, reproduction, and pathology</td>
<td>Genetic and demographic management of the population</td>
</tr>
<tr>
<td>Field studies</td>
<td>Self-sustaining viable captive population</td>
</tr>
<tr>
<td>Behavioral ecology studies (home range size, movements, habitat preferences, social organizations, mating system, feeding, and antipredator adaptations)</td>
<td>Genetic and demographic management of the population</td>
</tr>
<tr>
<td>Locating existing suitable habitat containing critical resources for reintroduction</td>
<td>Ongoing research in behavior, genetics, physiology, nutrition, reproduction, and pathology</td>
</tr>
<tr>
<td>Habitat preservation and management</td>
<td>Related to the reintroduction's success.</td>
</tr>
<tr>
<td>Protection of habitat from degradation and exploitation</td>
<td>Preventing habitat loss and degradation.</td>
</tr>
<tr>
<td>Restoration and management of degraded habitats</td>
<td>Ensuring habitat restoration and management.</td>
</tr>
<tr>
<td>Increase in or maintenance of the number of preserved areas</td>
<td>Conservation of natural habitats.</td>
</tr>
<tr>
<td>Conservation education for long-term support</td>
<td>Education for long-term support.</td>
</tr>
<tr>
<td>Professional training through academic studies, workshops, internships, courses, and fellowships</td>
<td>Professional development for conservation professionals.</td>
</tr>
<tr>
<td>Determining the most appropriate public relations and educational strategies through surveys</td>
<td>Surveying public relations and educational strategies.</td>
</tr>
<tr>
<td>Public relations educational efforts using appropriate mass media (e.g., television, radio, magazines, and newspapers)</td>
<td>Utilizing appropriate mass media for public relations.</td>
</tr>
<tr>
<td>Local community education, both formal and informal</td>
<td>Enhancing community education.</td>
</tr>
<tr>
<td>Preparation and reintroductions of animals</td>
<td>Preparing and reintroducing animals.</td>
</tr>
<tr>
<td>Choice of candidates and assessment of their characteristics for retrospective correlation with postrelease survival</td>
<td>Assessing candidate characteristics.</td>
</tr>
<tr>
<td>Training in survival techniques, including foraging and feeding, antipredator tactics, locomotion, and orientation</td>
<td>Training in survival techniques.</td>
</tr>
<tr>
<td>Adaptation to local conditions at release site (food, climate and temperature, and disease)</td>
<td>Adapting to local conditions.</td>
</tr>
<tr>
<td>Release and long-term monitoring to evaluate causes of death and basis for survival</td>
<td>Monitoring for survival.</td>
</tr>
</tbody>
</table>

* From Kleinman (1989).

The conditions under which captive-bred individuals are raised can be critical. The development of appropriate survival skills may require a skilled parent or a particular stimulus at some critical period during development. For example, adult ferrets prefer eating whatever they were fed when they were 2 or 3 months old, which is when ferrets develop their permanent teeth. Therefore, at the age of 2 or 3 months, captive ferrets should be fed Prairie dogs, their exclusive prey in the wild. Methods of reintroduction may also require research. For example, should the animals be released as social groups or as individuals? Should they be led after they are released and, if so, for how long? The answers to such questions depend on the particular species being reintroduced.
Reintroduced. Reintroduction programs using captive-bred individuals are usually expensive, lengthy, complex, and difficult. Thus, the decision to begin such a program should not be made lightly. A short checklist of the major factors that should be considered when deciding whether or not to reintroduce a species is illustrated in Table III with respect to three species of lion tamarins. Answers to the questions in the checklist indicate that reintroduction is appropriate for golden lion tamarins but not the other two species because the causes of their decline have not been eliminated and funds to support a reintroduction program are not available.

There are no generally accepted guidelines for declaring the success of a reintroduction effort. Beck and colleagues (1994) suggested two possible criteria: if 500 wild individuals survived without human support or if a formal population viability analysis predicted that the population would be self-sustaining. By these criteria, they judged that only 11% of 145 animal reintroductions were successful.

However, many of the reintroduction programs they considered are still in progress and it is too early to evaluate their success. For example, techniques for breeding and reintroducing black-footed ferrets have greatly improved and are still improving. Since 1986, the black-footed ferret program has released 873 captive-bred ferrets at five sites. The combined wild population is probably larger than the captive population, which is maintained at 240 breeding individuals. In 1998, more than 30 litters containing more than 100 pups were born in the wild.

Successful captive breeding and reintroduction programs require sustained long-term, adequately funded efforts. Research can solve many problems involved in successfully breeding a species in captivity and reintroducing it to the wild. However, the ultimate success of many programs, such as that for the ferret, will depend on whether or not we are able to preserve enough suitable habitat to sustain viable wild populations of the species.

See Also the Following Articles

- Breeding of Animals
- Breeding of Plants
- Conservation Efforts, Contemporary

---

### Table III

<table>
<thead>
<tr>
<th>Condition of species</th>
<th>Golden</th>
<th>Golden-headed</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Need to augment wild popn.</td>
<td>Yes</td>
<td>No</td>
<td>Yes (?)</td>
</tr>
<tr>
<td>Available stock</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>No jeopardy to wild popn.</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

### Environmental conditions

| Causes of decline removed | ? | No | No |
| Sufficient protected habitat | Yes (?) | No | Yes |

### Biopolitical conditions

| No negative impact for locals | Yes | ? | ? |
| Community support exists | Yes | Yes | Yes |
| NGOs/Government support/involved | 5 | 2 | 4 |
| Conformity with all laws/regulations | Yes | ? | ? |

### Biological and other resources

| Reintroduction technology known/in development | 4 | 5 | 3 |
| Knowledge of species’ biology | 3 | 1.5 | 3 |
| Sufficient resources exist for program | Yes | No | No |

| Recommended reintroduction/translocation? | Yes | No | No |

*Adapted from Kleiman (1990).*
Bibliography


I. General Features
II. Response of the Oceanic Sink to Climate Change
III. Response of the Terrestrial Sink to Climate Change
IV. History of the Carbon Cycle
V. Managing the Carbon Cycle
VI. Will the Sinks Increase, Remain the Same, or Decrease?

GLOSSARY

autotrophic respiration The respiration of photosynthetic organisms.
biomass The mass of biological material after removal of water by oven-drying at 70–100°C, often expressed as mass per area of ground surface.
BP Before the present, usually measured in billions of years (Ga) or millions of years (Ma).
El Nin˜o–Southern Oscillation Global climatological phenomenon, occurring every few years, involving reversal of trade winds, warming of parts of the ocean, and consequent widespread disturbances to the climate of several regions.
global circulation models (GCMs) Computer models of atmospheric circulation patterns and surface energy balance used for weather forecasting and to predict climates at regional and global scales. GCMs are so large that they must be run on supercomputers.
heterotrophic respiration The respiration of all those life forms that feed on photosynthetic organisms, including bacteria, fungi, and animals.
isotope Many elements exist in several forms, called isotopes. Carbon has eight isotopes, of which 12C forms 98.9% of naturally occurring CO2 and 13C forms 1.1%. Other carbon isotopes are unstable, although 14C has a half-life as long as 5730 years.
isotopic discrimination The tendency of chemical and physical reactions to “prefer” one isotope against another. The enzyme ribulose bisphosphate carboxylase oxygenase (Rubisco), which is responsible for capturing CO2 in photosynthesis, shows strong discrimination for 12CO2 against 13CO2.
lignin Heterogeneous carbon polymer associated with cellulose to form wood; resistant to decay by bacteria, and broken down by some fungi.
Q10 Mathematical index expressing the effect of temperature on respiration or decomposition: Q10 = 2 means that the rate doubles for a 10°C increase in temperature.
sequestration Term used to describe the uptake of carbon from a dilute source (the atmosphere) to a concentrated form (biomass).

CARBON IN THE FORM OF CO2 is exchanged by natural processes between the atmosphere and the land, and between the atmosphere and the ocean. The natural processes involved are photosynthesis, respiration, and
dissolution. The gains and losses may be viewed as a cyclic process, generally called the carbon cycle. The cycle is not in equilibrium. As a result of burning of fossil fuels and replacement of forest by farmland there is currently an anthropogenic input to the atmosphere of about 8 billion tonnes of carbon each year as CO$_2$. Of this, only about 3 billion tonnes of carbon appears in the atmosphere and contributes to global warming via the greenhouse effect. The remainder dissolves in the ocean, or is taken up by photosynthesis and stored as biomass or organic matter in the soil.

Photosynthesis is increased at elevated CO$_2$, so a negative feedback process may be operating whereby elevated CO$_2$ increases the capacity of the terrestrial ecosystems to act as a carbon sink. Nitrogen compounds emitted from terrestrial sources, especially motor engines and farmland, are deposited worldwide and may be enhancing the sink effect. But the long-term future of the sink is not assured. Global warming is likely to increase microbial decomposition of soil organic matter, causing an increased transfer of C from soil to atmosphere, hence reducing the sink.

There is now widespread realization, by world governments as well as scientists, that anthropogenic emissions of CO$_2$ and other trace gases are increasing the greenhouse effect and causing global warming. Consequently, there is renewed interest in the carbon cycle, and in managing both the terrestrial and ocean sinks to reduce the rate at which CO$_2$ increases in the atmosphere.

I. GENERAL FEATURES

The carbon cycle is a natural biogeochemical cycle in which carbon as CO$_2$ is transferred from the atmosphere to the land and ocean, where it resides in another form before returning to the atmosphere as CO$_2$. The principal processes involved in transfer from the atmosphere are the dissolution of CO$_2$ in the oceans and the uptake of CO$_2$ by the photosynthesis of green plants. The processes involved in return to the atmosphere are the release of CO$_2$ from the ocean in regions in which the surface of the ocean has become saturated with CO$_2$ and the oxidation of organic matter by respiration or fire, which essentially reverses the photosynthetic process:

\[
\text{photosynthesis} \rightarrow 6\text{CO}_2 + 12\text{H}_2\text{O} \rightarrow \text{C}_6\text{H}_{12}\text{O}_6 + 6\text{O}_2 + 6\text{H}_2\text{O}
\]

\[
\text{respiration and } \rightarrow \text{fire}.
\]

In photosynthesis, water is split and the 12 oxygen atoms produce six molecules of oxygen gas. Carbon dioxide is reduced to glucose (C$_6$H$_{12}$O$_6$), from which other biochemical constituents of biomass are synthesized.

The carbon cycle can be envisioned as a set of fluxes between major pools (Fig. 1). The pools differ in magnitude, and in the average time a carbon atom resides in them, and therefore the dynamic behavior of the system as a whole is likely to be complex. For example, an “average” carbon atom can be expected to reside in the atmosphere for about 5 years, whereas in the ocean the corresponding residence time will be hundreds of years.

An understanding of the carbon cycle is fundamental to our understanding of life because all biomass is carbon based. The overall carbon content of dry biomass is in the range 45–55%. The carbon cycle has become a subject of intense research in recent years, since the realization that anthropogenic emissions of CO$_2$ are large enough to force the cycle into a state of disequilibrium whereby the concentration of the gas in the atmosphere is increasing. Because carbon dioxide is the most important of the several “greenhouse gases” that absorb infrared radiation emitted from the planetary surface, its increase in concentration is capable of causing additional global warming. Indeed, there is overwhelming evidence that it is already doing so. The Framework Convention on Climate Change (United Nations, 1992)
states an ultimate aim: "to achieve stabilization of greenhouse gas concentrations . . . at a level that would prevent dangerous anthropogenic interference with the climate system."

In the early 1990s, it was noted that knowledge of the carbon cycle is still incomplete. In particular, it was shown that the annual anthropogenic emissions of carbon as CO2 exceeded the sum of the annual increase observed in the atmosphere plus that dissolving in the oceans. In other words, it was not possible to account for all the emitted CO2. It was postulated that there must be a substantial "missing sink" for CO2, amounting to about 2 billion tons of carbon (2 GtC). The sink was presumed to be the net photosynthetic uptake of the terrestrial vegetation. Therefore, in the 1990s there was a major research effort to resolve uncertainties in the location and magnitude of sinks. The challenge now is to predict the behavior of the carbon cycle over the next few decades, as CO2 and temperature continue to rise. The goal is to manage the carbon cycle through using scientific knowledge to develop policy.

The cycle shown in Fig. 1 is modified from Schimel (1995), updated to take into account new data on the rate of fossil fuel burning and reducing slightly the deforestation flux according to recent estimates. Units are in billions of tons of carbon, GtC (see Table I for conversion factors). In general, there is considerable uncertainty regarding many of the data on carbon stocks, largely the result of the heterogeneous nature of the earth's aquatic and terrestrial surface and the difficulty in comprehensively sampling the spatial variability.

Gases other than carbon dioxide are involved in the global carbon cycle, but they are quantitatively less important. Methane (CH4) and isoprene (C5H8) both have global emission rates estimated as about 0.3–0.5 GtC a⁻¹, much less than the respiratory fluxes of carbon dioxide. Carbon monoxide is formed during combustion, and also by oxidation of methane, but it reacts with the OH radical in the atmosphere to form carbon dioxide. The total production of carbon monoxide from all sources is estimated as 2.4 GtC a⁻¹ but its lifetime is only a few weeks before oxidation to CO2.

### A. Atmospheric Analysis

Some of the best information about the carbon cycle comes from the analysis of the CO2 concentration in the atmosphere, pioneered in the 1950s by Keeling, who established a CO2 observatory at Mauna Loa in Hawaii and first demonstrated the upward trend in the CO2 concentration (Fig. 2). Superimposed on the trend, there is an annual cycle whereby the concentration decreases during summer in the Northern Hemisphere.
and increases in the winter, with a minimum in October and a maximum in May. This summer decline is attributable to strong summertime uptake by photosynthesis of terrestrial vegetation (in the Southern Hemisphere, there is much less land and so a corresponding photosynthetic signal is not evident during the southern summer). Now, there is a network of remote stations whereby air samples are regularly taken in glass flasks and sent to a common laboratory for analysis. An important component of the analysis is the isotopic signal of O$_2$. This enables us to distinguish between ocean uptake and the photosynthetic uptake by C$_3$ plants because the latter discriminates against $^{13}$C, whereas the former does not. Recently, it has become technically possible to detect small changes between oxygen and nitrogen concentrations. This also provides a signal of photosynthesis because photosynthesis releases one molecule of O$_2$ for every CO$_2$ taken up, whereas dissolution in the ocean has no influence on O$_2$. In fact, just as the CO$_2$ concentration is increasing by a few parts per million (ppm) each year so also the O$_2$ concentration is decreasing. Fortunately, this is not cause for alarm because the O$_2$ concentration is very high (about 210,000 ppm). Ultimately, this technique of measuring changes in O$_2$ may prove to be the most sensitive method of detecting trends in photosynthesis at a global scale. Currently, most of the inferences have been made from CO$_2$ concentrations and isotopic fractions of $^{13}$C and $^{14}$C. Using data from the flask network coupled with knowledge of the anthropogenic emissions and atmospheric circulation, it is possible to calculate the latitudinal distribution of the terrestrial and oceanic carbon sink. Currently, there is not complete agreement between different groups of workers on the point of detail because each group uses its own approach to the calculation, but three conclusions emerge:

1. There is a large northern net sink of carbon, associated with uptake by the terrestrial vegetation and the ocean, usually estimated as 1 or 2 Gt of carbon per year.
2. In the equatorial latitudes there is a small net sink, but because deforestation accounts for an efflux of about 1.5 Gt of carbon per year there must be a biotic sink of opposite sign and about the same magnitude.
3. The pattern of sink distribution is not the same each year, being influenced by climatic phenomena and possibly by major volcanic eruptions (droughts associated with the El Niño–Southern Oscillation have been implicated as the main influence).

Overall, we can conclude from these studies that the net terrestrial sink is 2 or 3 GtC of carbon per year, and the ocean sink is likely to be about 2 GtC a$^{-1}$. In addition to the flask measurements mentioned previously, there is independent evidence for terrestrial carbon sinks. For example, temperate forests in Europe and North America are growing faster than ever before, and in the equatorial region it has been found that undisturbed, mature forests are accumulating carbon.

**B. Ocean Analysis**

The surface of the planet is largely covered by ocean. There are $3.6 	imes 10^8$ km$^2$ of ocean compared with $1.9 	imes 10^8$ km$^2$ of land. Carbon dioxide is readily soluble in water. At 1$^\circ$C the solubility is 1.16 m$^3$ of gas/m$^3$ of water, but this decreases to 0.85 at 20$^\circ$C. In solution, there is an equilibrium mixture of carbonic acid and bicarbonate and carbonate ions, which make up the fraction "dissolved inorganic carbon" in Fig. 1:

$$\text{CO}_2 + \text{H}_2\text{O} \rightleftharpoons \text{H}_2\text{CO}_3 \rightleftharpoons \text{H}^+ + \text{HCO}_3^- \rightleftharpoons 2\text{H}^+ + \text{CO}_3^{2-} \tag{2}$$

The proportions of each species depend on pH. At high pH the reactions shift to the right. At a pH of approximately 8, in the ocean most of the carbon is in the form of bicarbonate. In certain conditions of very high pH, carbonate predominates and calcium or magnesium carbonate are precipitated from the solution. Most of the world's limestone since Cambrian times consists of shells and exoskeletons of marine life. In present-day seas the main groups of organisms responsible are mollusks, corals, echinoderms, foraminifera, and calcareous algae.

Marine ecosystems play a role in the carbon cycle through the so-called "biological pump" acting as follows. Organisms occupy the well-mixed surface layers of the ocean and photosynthesize and grow at a rate which varies according to the nutritional state of the ocean. Dead biota and feces fall through the water column, some of them reaching the seabed, thus removing carbon from the surface layers and hence reducing the partial pressure of CO$_2$ there. This enables uptake of new CO$_2$ from the atmosphere. Thus, the ocean's sink strength is increased by biological activity. As the particles sink, decomposition occurs, with the release of inorganic carbon and nutrients. Attempts to directly measure the carbon transferred to the seabed in this particle "rain" by means of traps suggest that the downward flux of particulates is very low (Fig. 1). At great depth, the high pressure enables more CO$_2$ to be dissolved in the water than would otherwise be the case. The global ocean has currents that operate on various scales of time and space. Cold water subides in polar
regions, setting up a circulation. This tendency is enhanced by the fact that polar water is more salty and therefore more dense in the winter, as the fresh water is "frozen out" and added to the polar ice caps. A thermohaline circulation pattern is thus set up, with water subsiding in the polar regions and upwelling in tropical waters (Fig. 3). Patterns and rates of circulation have been obtained from studying the transport of material which entered the ocean as fallout from nuclear weapons tests in the early 1960s. The circulation is very slow compared with atmospheric circulation. In the equatorial zone the upwelling brings nutrient- and carbon-rich water to the surface and outgassing of CO₂ occurs.

Estimates of the sink strength of the ocean are made using knowledge of the CO₂ partial pressure in the surface waters. The rate of uptake or loss of CO₂ is proportional to the difference in partial pressure between the water and the air and to the exchange coefficient which depends on wind speed. Tens of thousands of such measurements are made annually. In high latitudes the surface waters are less than saturated with CO₂, so CO₂ dissolves from the atmosphere. In equatorial regions many waters are saturated with CO₂ and so there is an efflux to the atmosphere. On average, the ocean’s uptake is calculated as about 2 GtC per year.

A recent CO₂ flux map indicates strong sinks in the north Atlantic and Pacific and a source area in the mid-Pacific corresponding to an efflux of 0.5–1 GtC per year. There is likely to be considerable interannual variability in the sink strength of the ocean as a result of variations in currents, which influence sea surface temperatures and may influence the extent to which CO₂-rich water is brought to the surface. For example, the efflux from the equatorial waters is increased during El Niño years, when the surface temperature of the Pacific increases.

C. Uptake by Vegetation

Traditionally, ecologists have focused on measuring the biomass stocks and the net primary productivity (NPP, Table II). NPP is defined as the rate of net biological production of the phytomass, being the photosynthetic gains minus the losses of plant respiration. However, NPP is notoriously difficult to measure and is only one of the fluxes which are helpful in the analysis of the carbon cycle. The fundamental fluxes that we would like to know in addition to the NPP \( P_n \) are the net ecosystem productivity \( P_e \) and the net biome productivity \( P_b \). They may be defined and related as follows:

\[
P_e = P_n - R_h \tag{3}
\]

where \( P_n \) is the gross primary productivity, corresponding to the photosynthetic rate, and \( R_h \) is the respiration of the plant component, called the autotrophic respiration:

\[
P_a = P_n - R_h \tag{4}
\]

The net ecosystem productivity \( P_e \) expresses the net carbon flux to the ecosystem because it allows for respiratory losses of heterotrophs \( R_h \), including the animal and microbial consumption of the organic matter produced by plants. Micrometeorological methods, particularly eddy covariance, directly measure \( P_e \) and can be used to estimate \( P_n \) where \( R_h \) is available from measurements made in chambers attached to the soil surface.

At larger scales (geographical regions), another term is required to express the net flux after correction for disturbances and rate of change in land use which contribute to a leakage flux \( L \):

\[
P_b = P_e - L \tag{5}
\]

where \( P_b \) is the net biome productivity and the leakage term \( L \) includes processes such as changes in land use, erosion, and fire.

In 1996 the International Geosphere–Biosphere Programme launched FLUXNET to promote cooperation
TABLE II
Current Estimates of Global Carbon Stocks in Vegetation and Soils and the Range of Net Primary Productivity (NPP)

<table>
<thead>
<tr>
<th>Biome</th>
<th>Area (10^6 km^2)</th>
<th>Vegetation</th>
<th>Soils</th>
<th>Total</th>
<th>NPP (1 C ha^-1 year^-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical forests</td>
<td>17.6</td>
<td>212</td>
<td>216</td>
<td>428</td>
<td>11.0 (3.0–17.5)</td>
</tr>
<tr>
<td>Temperate forests</td>
<td>10.4</td>
<td>99</td>
<td>100</td>
<td>199</td>
<td>6.3 (2.0–12.5)</td>
</tr>
<tr>
<td>Boreal forests</td>
<td>13.7</td>
<td>88</td>
<td>471</td>
<td>559</td>
<td>4.0 (1.0–7.5)</td>
</tr>
<tr>
<td>Tropical savannas</td>
<td>22.5</td>
<td>66</td>
<td>264</td>
<td>330</td>
<td>4.5 (1.0–10.0)</td>
</tr>
<tr>
<td>Temperate grasslands</td>
<td>12.5</td>
<td>9</td>
<td>299</td>
<td>304</td>
<td>3.0 (1.0–7.5)</td>
</tr>
<tr>
<td>Deserts and semideserts</td>
<td>30.0</td>
<td>8</td>
<td>191</td>
<td>199</td>
<td>0.05 (0.0–0.1)</td>
</tr>
<tr>
<td>Tundra</td>
<td>9.3</td>
<td>6</td>
<td>121</td>
<td>127</td>
<td>0.1 (0.0–0.4)</td>
</tr>
<tr>
<td>Wetlands</td>
<td>3.3</td>
<td>15</td>
<td>225</td>
<td>240</td>
<td>0.9 (0.1–3.0)</td>
</tr>
<tr>
<td>Croplands</td>
<td>16.0</td>
<td>3</td>
<td>128</td>
<td>131</td>
<td>1.6 (0.2–3.0)</td>
</tr>
<tr>
<td>Total</td>
<td>135.6</td>
<td>466</td>
<td>2011</td>
<td>2477</td>
<td></td>
</tr>
</tbody>
</table>

*From WBGU (1998).*

for studies of CO₂ uptake by the terrestrial vegetation. Currently, about 70 science teams throughout the world are measuring the net uptake of carbon P, using sensors mounted above vegetation on towers. Most of them work in undisturbed forests. Conclusions so far are as follows:

1. Most of the boreal and temperate forests in the study are sinks for carbon, in the range 0.3–8 tC ha⁻¹ a⁻¹; however, most of the forests being studied are forests that have been managed for timber production and may not constitute ‘average’ or ‘representative’ forests.

2. The sink strength shows notable interannual variability, depending on the weather.

3. Old and undisturbed tropical forests are sinks (0.3–6 t C ha⁻¹ a⁻¹).

4. In all of these forests, the net carbon gain is a small difference between two large fluxes—the incoming photosynthesis and the outgoing respiration. Of the respiratory fluxes, the flux from the soil far exceeds the flux from the aboveground plant material.

Of these conclusions, the fourth is perhaps the most vital. Examples from three forest types are shown in Fig. 4. There is a remarkable similarity in the basic pattern of fluxes in all three forest types. In all of them, the total photosynthetic uptake is almost matched by the efflux of CO₂ from the soil.

D. Deforestation and Land Use Change

The most significant change in land use is the conversion of forest to grasslands and farmlands. In former times this conversion took place in Europe and North America. In recent years it has occurred in the tropics. The area of forests in different geographical regions according to Dixon et al. (1994) is 4.165 x 10⁹ ha, amounting to 359 GtC or three-fourths of the carbon stored in terrestrial vegetation (Table III). The rate of deforestation is 15–17 x 10⁶ ha per year, most of this being in the tropics. When tropical forest is cleared to prepare the land for pasture, there is an immediate loss of carbon as carbon dioxide during combustion and smoldering. Dixon’s data imply a worldwide loss of forest that corresponds to a carbon flux of 0.9 GtC per year, but most authorities now think the estimate should be higher (Fig. 1), perhaps as high as 2.0 GtC per year. In evaluating the deforestation flux it should be borne in mind that some forest disturbances including logging of large trees do not appear in satellite imagery, and this is one reason why Dixon’s estimate may be too low. This type of disturbance has been called “cryptic deforestation.”

Many nations have actually increased their forest area in recent years, but so far this represents only a trivial flux of carbon: 0.08 GtC per year for the United States, 0.05 for Canada, and 0.002 for Britain. In reality, the effort to manage carbon on a global scale, in order to reduce the rate of CO₂ increase, has scarcely begun. When forests are removed to make way for pastures or arable crops, there is an effect on the climate system resulting from the change in the biophysical properties of the land surface. Forests are rougher, they reflect less radiation than farmland, and they use more water. The climatic impact of deforestation has been estimated by running global circulation models with appropriate...
CARBON CYCLE

FIGURE 4  Carbon fluxes in three forest ecosystems: (a) humid tropical, (b) temperate, and (c) boreal. $G_p$, gross primary productivity; $R_t$, total respiration; $R_a$, autotrophic respiration; $R_h$, heterotrophic respiration; $R_l$, leaf respiration; $R_w$, aboveground wood respiration; $R_r$, root respiration; $D_{AG}$, aboveground detritus; $D_{BG}$, belowground detritus; $T$, belowground carbon translocation. Stocks of carbon and their annual increments are shown in bold: $AG$, carbon in aboveground biomass and $D_{AG}$ is the annual increment; $BG$, carbon in belowground biomass and $D_{BG}$ is the annual increment; $SOM$, carbon in soil organic matter and $D_{SOM}$ is its annual increment. Units are gC m$^{-2}$ a$^{-1}$ for fluxes and gC m$^{-2}$ for stocks (after Malhi et al., 1999).

parameterization of roughness, reflectance, and surface resistance to water loss. In tests with six different models of the Amazon basin, deforestation increased regional temperatures from 0.1 to 2.3°C, decreased precipitation from 15 to 640 mm year$^{-1}$, and decreased evaporation from 25 to 500 mm year$^{-1}$. It is too early to determine the reliability of these simulations.

E. Soils

The carbon stored in soil is the residue of thousands of years of NPP. There is uncertainty about the quantity of carbon stored as organic matter in the world’s soils. Soil surveys suggest that the figure exceeds that in the vegetation by two or three times (Tables II and IV). It
### Summary of Areas of Forests and the Carbon Stored as Biomass in the Above- and Belowground Plant Parts (not Soil Carbon)

<table>
<thead>
<tr>
<th>Location</th>
<th>Forest area (10^6 ha)</th>
<th>Change in forest area (10^6 ha)</th>
<th>Carbon stock, excluding soil (Gt C)</th>
<th>Change in pool (Gt C year⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High latitude</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russia</td>
<td>884</td>
<td>0.2</td>
<td>74</td>
<td>0.40</td>
</tr>
<tr>
<td>Canada and Alaska</td>
<td>488</td>
<td>0.3</td>
<td>14</td>
<td>0.08</td>
</tr>
<tr>
<td>Midlatitude</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>241</td>
<td>0.1</td>
<td>15</td>
<td>0.17</td>
</tr>
<tr>
<td>Europe</td>
<td>283</td>
<td>0.3</td>
<td>9</td>
<td>0.11</td>
</tr>
<tr>
<td>China</td>
<td>118</td>
<td>0.6</td>
<td>37</td>
<td>0.02</td>
</tr>
<tr>
<td>Australia</td>
<td>396</td>
<td>0.1</td>
<td>18</td>
<td>Trace</td>
</tr>
<tr>
<td>Low latitude</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>310</td>
<td>3.0</td>
<td>41</td>
<td>0.70</td>
</tr>
<tr>
<td>Africa</td>
<td>527</td>
<td>4.1</td>
<td>32</td>
<td>0.35</td>
</tr>
<tr>
<td>Americas</td>
<td>918</td>
<td>7.4</td>
<td>119</td>
<td>0.60</td>
</tr>
<tr>
<td>Total</td>
<td>4165</td>
<td>15.4</td>
<td>359</td>
<td>0.90</td>
</tr>
</tbody>
</table>

* The period was 1987–1990 (Source: Dixon et al., 1994).

### Estimates of the Pools of Soil Carbon

<table>
<thead>
<tr>
<th>Life zone group</th>
<th>Area (10^12 m²)</th>
<th>Carbon density (kg m⁻²)</th>
<th>Soil carbon (10^15 g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical forest, wet</td>
<td>4.1</td>
<td>19.1</td>
<td>78.3</td>
</tr>
<tr>
<td>Tropical forest, moist</td>
<td>3.3</td>
<td>11.4</td>
<td>60.4</td>
</tr>
<tr>
<td>Tropical forest, dry</td>
<td>2.4</td>
<td>9.9</td>
<td>23.8</td>
</tr>
<tr>
<td>Tropical forest, very dry</td>
<td>3.6</td>
<td>6.1</td>
<td>22.0</td>
</tr>
<tr>
<td>Temperate forest, warm</td>
<td>8.6</td>
<td>7.1</td>
<td>61.1</td>
</tr>
<tr>
<td>Temperate forest, cool</td>
<td>5.4</td>
<td>12.7</td>
<td>61.2</td>
</tr>
<tr>
<td>Boreal forest, wet</td>
<td>6.9</td>
<td>19.3</td>
<td>133.2</td>
</tr>
<tr>
<td>Boreal forest, moist</td>
<td>4.2</td>
<td>11.6</td>
<td>48.7</td>
</tr>
<tr>
<td>Tropical woodland and savanna</td>
<td>24.0</td>
<td>9.4</td>
<td>120.6</td>
</tr>
<tr>
<td>Temperate thorn steppe</td>
<td>3.9</td>
<td>7.6</td>
<td>20.6</td>
</tr>
<tr>
<td>Cool temperate steppe</td>
<td>9.0</td>
<td>13.3</td>
<td>119.7</td>
</tr>
<tr>
<td>Tropical desert bush</td>
<td>1.2</td>
<td>2.0</td>
<td>2.4</td>
</tr>
<tr>
<td>Warm desert</td>
<td>14.0</td>
<td>1.4</td>
<td>19.6</td>
</tr>
<tr>
<td>Cool desert</td>
<td>4.2</td>
<td>9.9</td>
<td>41.6</td>
</tr>
<tr>
<td>Boreal desert</td>
<td>2.0</td>
<td>10.2</td>
<td>20.4</td>
</tr>
<tr>
<td>Tundra</td>
<td>8.8</td>
<td>21.8</td>
<td>191.8</td>
</tr>
<tr>
<td>Subtotal for the study</td>
<td>1025.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cultivated land</td>
<td>21.2</td>
<td>7.9</td>
<td>167.5</td>
</tr>
<tr>
<td>Wetlands</td>
<td>2.8</td>
<td>72.3</td>
<td>202.4</td>
</tr>
<tr>
<td>Global soil carbon pool</td>
<td></td>
<td></td>
<td>1395.3</td>
</tr>
</tbody>
</table>

* Based on soil survey (Post et al., 1982).
is noteworthy that all major biomes contain substantial carbon stocks in the soil, even humid tropical forests in which rapid decomposition rates are observed in short-term experiments.

Soil organic matter can persist for a long time. Litter fall, exudates, leachates, and dead roots enter the soil and provide the substrate for the soil microflora. Decomposition is a multistage process, involving a succession of insects, bacteria, and fungi, operating on the various biochemical constituents of the detritus. The products of decomposition are $\text{CO}_2$, $\text{H}_2\text{O}$, and mineral ions. The litter are usually captured by the roots and mycorrhizas. The $\text{CO}_2$ accumulates in the pores of the soil and diffuses to the atmosphere as the heterotrophic respiration $R_h$, forming perhaps half of the total $\text{CO}_2$ flux from the soil. The remainder originates from root respiration.

As the dead remains of plants and animals (known as detritus or simply litter) are broken down they assume a different chemical and physical form, and when the cellular structure is no longer discernible the material is termed humus. Humus is chemically heterogeneous and contains some fractions which are extremely resistant to decay, particularly those which originate from lignin. The acid-soluble fraction of humus is termed fulvic acid, and the insoluble fraction is termed humic acid. These fractions are chemically heterogeneous. They are transported vertically in the soil by rain and they contribute to the long-lived soil organic matter termed the "recalcitrant fraction." In storms they may be transported laterally to rivers, to which they impart a brown coloration, and they may finally enter the sea.

Schlesinger presented an example of the stocks and pools in a grassland, showing typical pool sizes and resident times applicable to the various fractions (Fig. 5).

Current interest centers on the effect of temperature on the decomposition rate because the carbon balance of the terrestrial surface is very sensitive to increases in respiration, and it is expected that climate warming will increase the respiration rate. There are reports of ecosystems "flipping" from being sinks to being sources, and there is one report of an unexpectedly large respiration flux derived from "old" carbon in forest soils.

In laboratory studies the relationship between respiration and temperature can generally be approximated.
by an exponential function, with a $Q_{10}$ of approximately 2. However, much higher $Q_{10}$ values, ranging from 2 to 8, have been reported from chamber measurements of CO$_2$ efflux from the soil. This matter is currently one of great research interest because approximate calculations suggest that the soil efflux of carbon is likely to exceed the photosynthetic influx in a warmer world, thus causing complete loss of the terrestrial sink.

One small fraction of the soil carbon is elemental carbon derived from fire, occurring as fine deposits and as charcoal. This fraction is not decomposed by fungi and bacteria. In a regime of burning, elemental carbon can be expected to accumulate over centuries.

F. Rivers

The surface area of river and inland waters is not very great compared with that of land and ocean, and therefore the direct exchanges of carbon between rivers and atmosphere are not especially important. However, the rivers of the world are conduits between the land and ocean, carrying large amounts of carbon and nutrients to the sea. Estimates based on analyses of the organic material entering the sea from the world's rivers have been made based on survey data from a large proportion of the main rivers (Table V).

Particular inorganic carbon (PIC) originates mainly from the mechanical erosion of sedimentary rocks. Dissolved organic carbon (DOC) comes from soil leaching. Exudates from roots, or leaching from leaf canopies, releases simple organic compounds such as organic acids, sugars, and alcohols, which decompose rapidly (hours or days). However, humic materials have a long lifetime (hundreds and thousands of years), and are often present as colloids. In analysis, dissolved and particulate fractions are somewhat arbitrarily separated by a 0.5-mm filter. The captured part is called particulate organic carbon (POC). POC includes material derived from soil and peat by erosion, but there are also coarse fractions as the debris of plants living in the riparian zone.

From the viewpoint of the carbon cycle, the flux of carbon derived from the atmosphere (rather than geologically derived) is the relevant quantity. It is given as the sum of the atmospheric component of dissolved inorganic carbon (DIC) + DOC + atmospheric POC, and amounts to 0.48 GtC year$^{-1}$ (Table V). It is less than 1% of the global NPP.

G. Geological Processes

There is an immense quantity of carbon stored as carbonate rocks—perhaps as much as $65 \times 10^6$ Gt. The chemical forms predominating are CaCO$_3$, occurring in limestone in the mineral forms of calcite and aragonite; MgCO$_3$ (magnesite); CaMg(CO$_3$)$_2$ (dolomite); MnCO$_3$ (rhodochrosite); and ironstone or siderite (FeCO$_3$). Over geological timescales, carbonates have been precipitated both biotically and abiotically. Corals, gastropods, and clams have shells made of aragonite; foraminifera, and coccolithophores; some gastropods and clams

<table>
<thead>
<tr>
<th>Source</th>
<th>Age of carbon (years)</th>
<th>Flux (Gt C a$^{-1}$)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIC</td>
<td>Geological 10$^6$–10$^9$</td>
<td>0.17</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td>Geological 10$^6$–10$^9$</td>
<td>0.14</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Atmospheric 0–10$^7$</td>
<td>0.24</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOC</td>
<td>Soils 10$^7$–10$^8$</td>
<td>0.20</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pollutant 10$^7$–10$^8$</td>
<td>0.03</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CO$_2$ Atmospheric 0</td>
<td>0.02–0.08</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POC</td>
<td>Soil 10$^7$–10$^8$</td>
<td>0.10</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Algal 10$^7$–10$^8$</td>
<td>0.03</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pollutant 10$^7$–10$^8$</td>
<td>0.013</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geological 10$^7$–10$^8$</td>
<td>0.08</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*PIC, particulate inorganic carbon; DIC, dissolved inorganic carbon; DOC, dissolved organic carbon; POC, particulate organic carbon. Sensitivity to change is indicated by letters A–F as follows: A, land erosion; B, chemical weathering; G, global warming and ultraviolet radiation; D, eutrophication; E, organic pollution; and F, basin management (From Maybeck, 1993).
have shells made of calcite; and starfish have exoskeletons made of high-magnesium calcite.

Over long periods of time, carbonates and silicates in the marine sediment are subducted in the earth’s crust, where they are metamorphosed. In this process, primary minerals are created and CO₂ is released and returned to the atmosphere from volcanoes and fumeroles, currently at a very low rate (0.02–0.04 Gt a⁻¹) compared to that of fossil fuel emissions. This is sometimes known as the geologic carbon cycle. It takes 100–200 million years for the ocean crust to circulate.

The composition of volcanic gases is highly variable. For example, Kilauea in Hawaii was 2.3% by volume CO₂, Stromboli in Italy was 30%, and the eruptions from Lake Nyos in Cameroon may have been almost pure CO₂. Likewise, the total quantities of carbon from active volcanoes are very variable: Mt. St. Helens was 3 ktC day⁻¹, Etna was 30 ktC day⁻¹, and Nyiragongo was 45 ktC day⁻¹.

The weathering of rocks consumes CO₂. In the past, when CO₂ concentrations were higher, exposed rocks were weathered. Currently, the weathering process is more intense in the soil. There, CO₂ concentrations are much higher than the 0.036% found in the atmosphere, typically reaching several percent. This is the result of respiration of roots and microbes. Thus, carbonic acid H₂CO₃ and its dissociation products bicarbonate, carbonate, and hydrogen ions are formed (Eq. 2).

Carbonic acid weathers minerals in several ways, as illustrated in the following examples. In the first case, Na-feldspar is weathered and the carbon emerges as a product in the form of bicarbonate. Another product is the clay mineral kaolinite Al₂Si₂O₅(OH)₄:

\[
2\text{NaAlSi₃O₈} + 2\text{H}_₂\text{CO}_₃ + 4\text{H}_₂\text{O} \rightarrow 2\text{Na}^{++} + 2\text{HCO}_₃^- + 4\text{H}_₂\text{SiO}_₃ + \text{Al}_₄\text{Si}_₄\text{O}_₁₃\text{(OH)}₄
\]  

In the second example, calcium carbonate is weathered, and once more bicarbonate is produced:

\[
\text{CaCO}_₃ + \text{H}_₂\text{CO}_₃ \rightarrow \text{Ca}^{++} + 2\text{HCO}_₃^- 
\]

Bicarbonate produced in this way may enter streams and rivers, finally contributing to the DIC fraction of the carbon in the ocean.

II. RESPONSE OF THE OCEANIC SINK TO CLIMATE CHANGE

The ocean contains a large volume of water (1.3 × 10³ km³), with a correspondingly large capacity to dissolve CO₂. Given the high solubility of CO₂ in water, we would not expect the ocean to become saturated with CO₂ until it had accumulated in excess of 10⁵ Gt carbon, much more than it has today. However, exchange of gas between the atmosphere and the ocean is a slow process because only the surface layer, a few meters deep, is thoroughly mixed by wind action, and there is very poor mixing with the bulk of the ocean. This poor mixing is revealed from studies of the radium-
TABLE VI
Fossil CO\textsubscript{2} Emissions Breakdown\textsuperscript{a}

<table>
<thead>
<tr>
<th>Year</th>
<th>Annex 1 (GtC per year)\textsuperscript{b}</th>
<th>Non-Annex 1 (GtC per year)\textsuperscript{c}</th>
<th>Bunker fuels (GtC per year)\textsuperscript{c}</th>
<th>Total (GtC per year)\textsuperscript{c}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>3.91</td>
<td>2.09</td>
<td>0.11</td>
<td>6.11</td>
</tr>
<tr>
<td>1991</td>
<td>3.79</td>
<td>2.27</td>
<td>0.12</td>
<td>6.18</td>
</tr>
<tr>
<td>1992</td>
<td>3.71</td>
<td>2.24</td>
<td>0.14</td>
<td>6.09</td>
</tr>
<tr>
<td>1993</td>
<td>3.63</td>
<td>2.29</td>
<td>0.13</td>
<td>6.09</td>
</tr>
<tr>
<td>1994</td>
<td>3.68</td>
<td>2.48</td>
<td>0.13</td>
<td>6.29</td>
</tr>
<tr>
<td>1995</td>
<td>3.80</td>
<td>2.57</td>
<td>0.14</td>
<td>6.51</td>
</tr>
<tr>
<td>1996</td>
<td>3.87</td>
<td>2.66</td>
<td>0.14</td>
<td>6.67</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Fossil CO\textsubscript{2} emissions given are from fossil fuel combustion (including gas flaring), cement production, and other nonfuel carbon sources.

\textsuperscript{b} Annex 1 countries are Australia, Austria, Belarus, Bulgaria, Canada, Czechoslovakia, Denmark, European Economic Community, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, The Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russian Federation, Spain, Sweden, Switzerland, Turkey, Ukraine, United Kingdom, and United States.

\textsuperscript{c} Bunker fuels are those used in air and marine international transport.

clides released from bomb tests in the 1960s. Calculations based on the circulation rate suggest that it would take about 1000 years for equilibrium between the atmosphere and ocean to occur.

The response of the ocean sink to climatic change is not well established. It is presumed that in preindustrial times there was more or less a steady state, i.e., sink strength was zero. As the CO\textsubscript{2} concentration steadily increased from about 280 ppm in 1800 to present-day values in excess of 360 ppm, CO\textsubscript{2} began to be taken up by the ocean; however, despite the huge capacity available, the rate of uptake has been limited by the diffusional process at the sea surface so that there is essentially a lag in the response. Thus, if fossil fuel emissions were to abruptly cease, the sink would continue to operate for a long time (a century perhaps) because the concentration gradient driving the process would persist. This dynamic feature of the oceanic sink tends to dominate the more obvious short-term responses to temperature and CO\textsubscript{2} concentration.

The solubility of gases in water is a strong function of temperature, and it should be kept in mind in discussions of global change that all gases are considerably less soluble in water as the temperature increases (Table VII). Thus, a 1°C increase in temperature over the range 10–15°C is expected to reduce the sink strength by about 3%. A doubling of CO\textsubscript{2} concentration is expected to increase the diffusion gradient from atmosphere to sea, and hence the sink strength, by much more. Fluctuations in sea surface temperatures that result from climatic phenomena such as El Niño can be more than 1°C, but El Niño reduces the rate of upwelling of the Pacific Ocean, causing less efflux than in a normal year, and this overwhelms the expected effect of temperature. There are several other potentially critical processes. For example, increasing temperature will increase the melting rate of ice in the polar regions and thus increase the thermohaline cycle and the rate of upwelling in tropical waters. This effect is not well understood, is difficult to model, and is much more important than the direct effect of average temperature on the dissolution of CO\textsubscript{2}. It is becoming increasingly important to understand the marine biological processes in the context of

TABLE VII
Solubilities of Gases in Water at Atmospheric Pressure (101 kN m\textsuperscript{-2})\textsuperscript{a}

<table>
<thead>
<tr>
<th>Gas</th>
<th>Temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Carbon dioxide</td>
<td>1.16</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>0.018</td>
</tr>
<tr>
<td>Oxygen</td>
<td>0.037</td>
</tr>
<tr>
<td>Carbon monoxide</td>
<td>0.028</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Units: m\textsuperscript{3} of gas per m\textsuperscript{3} of water at the stated temperature.
the “sinking” of carbon and the operation of the biological pump. The pump is unlikely to be directly stimulated by CO₂, because the phytoplankton is believed to be rate limited by nutrients rather than by carbon. A factor which might stimulate the biological pump is the additional deposition of materials to the ocean. Approximate calculations of this effect have been made based on the Redfield ratio.

The Redfield ratio is derived from the chemical analysis of phytoplankton and expresses the atomic ratios of biomass. There tends to be a constant ratio of C : N : P such that marine biomass can be written as C₆₂₁N₃₁₆P谅. Because upwelling waters contain far more carbon relative to N and P than is required in the Redfield ratio, it is supposed that C concentration is not limiting. Thus, we may predict how much extra biomass might possibly be synthesized if the supply of rate-limiting N or P were enhanced, either by additional upwelling or by anthropogenic sources. It is well established that the deposition of nitrogen as nitrate and ammonium from anthropogenic sources has increased, particularly during the past few decades, thus augmenting the nitrate that was always present in the atmosphere as wind-blown dust derived from the land. The anthropogenic rates of dry and wet deposition are especially enhanced over the land of Europe and eastern North America, but atmospheric dispersal will have spread the emissions even to the ocean. It has been estimated that the “excess” nitrogen from anthropogenic sources could have augmented the nitrogen supply to the ocean by \(17 \times 10^{12} \text{ g N a}^{-1}\), which is the Redfield equivalent to an additional \(0.1 \times 10^{12} \text{ g C a}^{-1}\). Thus, the stimulation of the biological pump by anthropogenic nitrogen is currently only slight (see Fig. 1).

Of course, nitrogen is abundant in the atmosphere as \(\text{N}_2\), a gas which is sparingly soluble in water (0.0018 m³ nitrogen in 1 m³ water at 10°C). In fresh water, nitrogen fixation is achieved by aquatic nitrogen-fixing organisms, particularly the cyanobacteria. In seawater, nitrogen fixation proceeds at a slow rate as a result of deficiencies of micronutrients (notably Mo and Fe). A most interesting recent discovery has been that phytoplankton in the Pacific Ocean are stimulated to assimilate CO₂ because the phytoplankton is believed to have a higher rate of CO₂ fixation than the Redfield ratio. This suggests that the biological pump—an understating that marine biologists must provide.

III. RESPONSE OF THE TERRESTRIAL SINK TO CLIMATIC CHANGE

Insight into the carbon-accumulating capacity of ecosystems comes from studies of succession. For example, on glacial moraines in Alaska vegetation and soil develop over a period of several decades. In this period, discrete stages of vegetation occur as different life forms colonize the land. Initially, only lichens and mosses are able to withstand the resource-poor environment, but later grasses, herbs, and trees colonize. Some of the organisms have special attributes; for example, the lichens are capable of fixing atmospheric nitrogen, and the trees give rise to a spatial heterogeneity in microclimate that encourages biodiversity. Over 100 years the ecosystem aggrades, and the soil alone accumulates 0.4 tC ha⁻¹ a⁻¹ and about 0.02 tons of nitrogen ha⁻¹ a⁻¹. Thereafter, the rate of change is slow, and it is common to assume that a steady “mature” state is reached at which photosynthetic gains and respiratory losses are equal. Elsewhere, when agricultural land is abandoned, we observe a regrowth of vegetation—first as scrub then as forest. Examples from the tropics suggest an accumulation rate in the biomass of several tons of carbon ha⁻¹ a⁻¹, with a tendency to reach a ceiling after about 100 years. Unfortunately, no one has properly observed this ceiling, and it is possible that a steady state does not occur. In peatlands carbon can continue to accumulate for thousands of years until a profound disturbance occurs. This calls into question the normal assumption that vegetation reaches a steady state when it is mature.
In managed ecosystems, the natural tendency of eco-
systems to accumulate carbon in the juvenile stage is
exploited by farmers and foresters, who intervene by
picking, cutting, chopping, or burning. This interven-
tion creates much disturbance to the soil, and in some
cases losses of carbon to the atmosphere may be signif-
cant as a result. However, over entire crop cycles there
may well be a net accumulation of carbon in the soil
as humus and, in the case of forests, an accumulation
not only in the soil but also on the land as buildings,
furnishings, and other artifacts.

A. CO₂ Effects on Photosynthesis
and Growth

The response of photosynthesis to CO₂ and temperature
has been much studied in recent decades. Commercial
growers have long known that elevated CO₂ increases
productivity under greenhouse conditions, and good
measurements of photosynthesis of leaves, made in
short-term experiments as early as 1960 by Gaastra,
showed a strong positive effect of elevated CO₂. Those
who work on global modeling of the impact of elevated
CO₂ on plants like to use a single parameter \( \beta \) to de-
scribe the extent of the stimulation of NPP brought
about by doubling the CO₂ concentration, but plant
ecophysiologists see difficulties in this simplistic ap-
proach because it ignores many important factors. Hun-
dreds of species have been grown experimentally at
twice-normal CO₂ concentrations, but firm conclusions
about the value of \( \beta \) have been elusive. There are several
reasons for this. First, species differ in their response
to elevated CO₂, and for any single species there are
changes in sensitivity over the life cycle as acclimation
occurs. There are also likely to be variations according
to the nutrient and water supply, and there may be
complicating effects when several species are growing
together and thus competing with each other.

Experiments on the impact of elevated CO₂ on plants
are conducted using several experimental techniques.
Controlled environment rooms and outdoor chambers
cover an experimental area of a few square meters and
usually permit plants to be grown only for quite limited
periods (weeks to a few years) at elevated CO₂. Plants
may be rooted in natural soil, but more often they
have been planted in containers which have a restrictive
effect on the supply of nutrients and water. There are
exceptions, in which closed chambers for ecosystem
studies have operated over several years, especially in
tundra and grassland ecosystems. Sometimes, very
large, greenhouse-sized chambers have been used, as
in the case of the Biosphere-2 experiment in Arizona.

However, all chambers produce unwanted microclima-
tological effects, and to overcome these artificial influ-
ences plants have been exposed to high CO₂ in com-
pletely open conditions in the field, in their normal
cultivation conditions (the so-called FACE approach—
free air carbon dioxide experiment). Recently, FACE
experiments have been established at forest sites, and
it is expected that these will yield the most valuable
information about the response of major ecosystems to
eclimate change, although they consume large amounts
of gas and are thus costly to run. Finally, it has some-
times been possible to locate suitable natural sources
of CO₂, as fumaroles or CO₂ springs, where mature
vegetation is exposed permanently to elevated CO₂.
There are also drawbacks to this approach because the
concentration is uncontrolled, and there may be traces
of toxic gases.

A few conclusions may be drawn from the very large
and often conflicting literature:

1. In trees, the stimulatory effect of twice-normal
CO₂ on growth was reported in one review to be 38%
in coniferous trees and 63% in broad-leaved species.
However, the stimulation declines as the tree ages. This
effect of age has been shown in sour-orange, birch,
and Mediterranean oaks growing near high-CO₂ springs
in Italy.

2. In herbaceous crops, the stimulatory effect was
reported to be 16% on average, with some species being
stimulated by as much as 80%, whereas others showed
inhibition of growth.

3. There is often a downregulation of photosynthesis
when plants are grown at elevated CO₂. This takes
several forms. The most obvious is a decline in the
activity of the photosynthetic enzyme Rubisco (ribulose
bisphosphate carboxylase oxygenase). There may be a
reduction in the number of stomata per area of leaf.
Thus, a doubling of CO₂ concentration does not usually
produce a doubling of the photosynthetic rate: More
often the stimulation is in the range 20–60%.

4. Most plants show an increase in water use effi-
ciency (WUE) at elevated CO₂. WUE is defined as the
carbon taken up divided by the water used.

It is not clear how these conclusions are to be scaled
up to whole ecosystems. Models play an important role
in the use of experimental data to make predictions at
large scales, although much of the earlier data are not
useful in parameterizing the models because plants were
grown in wholly artificial conditions.
IV. HISTORY OF THE CARBON CYCLE

Carbon and other elements are produced in stars by nuclear fusion. Stars explode, and under the attractive force of gravity the debris from an exploded star (supernova) condenses to make a sun and planets. In our solar system this is believed to have happened 4.6 billion years ago, from the radiodated age of the oldest rocks found on the moon and as meteorites. Over a relatively short period, perhaps 100 million years, additional mass would have been accumulated from planetesimals and meteors, including meteors known as carbonaceous chondrites, which contain an abundance of the lighter elements including carbon. The carbon content of these is 0.5–3.6% and the nitrogen content is 0.01–0.28%. The planet would be hot inside, as a result of pressure and also as a result of radioactive decay of the larger atoms from the debris of the supernova. There would be considerably sorting and reprocessing of the elemental material to form minerals. An early planet would have no atmosphere, but outgassing from the hot interior would result in a mixture of gases containing water, nitrogen, carbon dioxide, hydrochloric acid, and sulfur dioxide—a mixture probably not unlike modern volcanic gas mixtures. Cooling would result in the water vapor condensing to form the ocean, although this could have been lost several times by meteorite bombardment which is believed to have been very intense until about 3.8 billion years ago. Gases such as SO₂ and HCl are extremely water soluble and form acids in aqueous solution. The acidity would have been neutralized by reaction with minerals at the surface. Nitrogen gas, N₂, is not very soluble in water and would have thus formed the bulk of the atmosphere. CO₂ is quite soluble in water (Table I) and would have eventually formed an equilibrium concentration.

The oldest fossils known to man are stromatolites, which are found as laminated calcareous deposits and are up to 3.5 billion years old. They are made up largely of filamentous organisms that, by comparison with modern analogs, are believed to be mainly the photo-synthetic prokaryotes cyanobacteria. The early communities may have resembled those which are now living in high temperatures in hydrothermal vents or "algal mats" in coastal regions. Photosynthesis may be even somewhat older than 3.5 Ga because the carbon isotopic signature of organic matter as old as 3.8 Ga shows δ¹³C to be depleted relative to its abundance in dissolved bicarbonate, a result of the greater affinity of the carbon-fixing enzyme Rubisco for ¹²CO₂. Two analogous sorts of photosynthesis may have occurred in ancient seas:

\[
CO₂ + 2H₂S + \text{sunlight} \rightarrow CH₂O + 2S + H₂O. \tag{8}
\]

Second, photosynthesis produced today by green plants, essentially the photochemical splitting of water to release oxygen:

\[
2H₂O + CO₂ + \text{sunlight} \rightarrow CH₂O + O₂ + H₂O. \tag{9}
\]

Oxygen accumulates only where the organic matter becomes buried and thus protected from decomposition or fire. It is presumed that the total oxygen that has ever been released is matched atom for atom by the storage of reduced carbon (coal, oil, and hydrocarbon gases), and it is approximately true that if all the known organic matter were to be burned the oxygen concentration would decrease to zero. In reality, biomass does not burn when the oxygen concentration decreases below about 15%. Lovelock noted that the current concentration of oxygen on Earth (21%) may be subject to homeostasis. Any more than 21% O₂ and biomass would burn uncontrollably, thus oxidizing more carbon and bringing the oxygen concentration back to the initial level. At much less than 21% O₂, fires do not burn, and thus oxygen can accumulate.

The oxygen took a long time to accumulate in the atmosphere as it reacted immediately with exposed minerals, especially iron, to produce Fe₃O₄. Banded deposits of Fe₃O₄ occur in rocks that are 2.5–3.0 Ga old. It was not until about 2 Ga ago that an oxygen-rich atmosphere developed, enabling aerobic heterotrophic life to evolve and forming a ultraviolet (UV)-shielding layer of ozone in the stratosphere. Low UV is an important condition for the colonization of the land.

So far, I have outlined the modern theory for the establishment of the carbon cycle and the generation of the oxygen-rich atmosphere, which distinguishes Earth from its sister planets Mars and Venus. These planets are lifeless, they have lost their oceans, and they contain an atmosphere which is almost pure CO₂.

Subsequent evolution of the carbon cycle, over a period of 2 billion years, is not well-known. A speculative reconstruction of partial pressures of CO₂ and O₂ has been made based on the oxidation state of iron and manganese in sedimentary and metamorphic rock and observations on the body size of certain groups of animals (Fig. 6). Other authors have examined the sedimentary record of organic carbon to calculate the rate of accumulation of carbon stocks during the past 0.6 Ga. It increased from a low value at 0.5 Ga to a peak...
in the Carboniferous. This is a time when the land flora was well developed over much of the earth, based on spore-bearing plants. Fish and mollusks were well established, and amphibians lived in tropical swamps. Reptiles were beginning to evolve.

As we consider recent times, it is clear that the carbon cycle has not been in a steady state over periods of tens of thousands of years. Fluxes between atmosphere and vegetation, and atmosphere and ocean, have occurred in response to changing temperature and changing vegetation cover. Evidence for the changing nature of the cycle in relation to temperature and vegetational cover comes from cores extracted from ice or firm in polar regions, especially in Greenland and Antarctica. The cores contain bubbles of trapped air, which are released when the core is thawed. Concentrations and isotopic ratios of CO₂ and CH₄ can be measured, and temperature can be estimated from the isotopic signature of oxygen in water. This type of record has shown clearly the correlative link between CO₂, CH₄, and temperature.

The most recent record, from Taylor Dome in Antarctica, is especially free from known artifacts and covers the postglacial period (the past 11,000 years). Following the retreat of the ice, the CO₂ concentration decreased slightly from 265 to 260 ppm. Thereafter, it increased to about 280 ppm before the period of industrialization and extensive deforestation. This pattern is mirrored by the isotopic signature of the carbon in CO₂, which at first increases and then decreases (Fig. 7). These changes in isotopic signature can only be achieved through massive changes in gas exchange with the land associated with terrestrial photosynthesis because exchanges with the ocean are not accompanied by isotopic discrimination. The changes from 11 to 7 ka BP are best interpreted as a consequence of the rapid growth and spread of vegetation when the temperature increased and the ice retreated. The climatic optimum for vegetation in the Northern Hemisphere, documented in the pollen record, occurred at about 7 ka BP. Thereafter, conditions became colder and drier; in Africa, for example, there was a change from grassland and savanna to desert. This seems to have been accompanied by a transfer of carbon from land to atmosphere to increase the CO₂. There may have also been other changes. For example, the increase in sea surface temperature would have caused a flux of carbon from the ocean to the atmosphere, but such changes are not accompanied by any fractionation because only terrestrial C₃ photosynthesis is capable of substantial fraction-
It is possible to conclude that changes in fluxes to and from the ocean were of minor importance during the period.

Rates of change of CO₂ concentration in the atmosphere, revealed by the ice core record, were always much slower than the current growth rate of 3.1 GtC a⁻¹.

V. MANAGING THE CARBON CYCLE

Responding to concerns that human activities are increasing atmospheric concentrations of greenhouse gases such as carbon dioxide and methane to levels which threaten to cause damaging global warming, most nations of the world signed the United Nations Framework Convention on Climate Change in 1992. They pledged to work toward stabilization of emissions. Following subsequent negotiations, the Kyoto Protocol (in late 1997) committed nations to legally binding reductions in emissions of six greenhouse gases: carbon dioxide, methane, nitrous oxide, hydrofluorocarbons, perfluorocarbons, and sulfur hexafluoride. These reductions will be counted from 2008 to the “commitment period” between 2008 and 2012. To date, 174 countries have ratified the treaty, and industrial countries are legally bound to decrease emissions by 5.2%. The European Union has agreed to reduce emissions 8% and the United States has agreed to 7%. Less industrialized countries have had difficulties in agreeing to any reductions. China, for example, has stated that it will not reduce emissions until it achieves the level of a medium-developed country. One notable industrialized country (Japan) is likely to achieve its target of 6% by building more nuclear reactors.

Under the terms of the treaty, reductions by the industrialized countries (called Annex 1 countries) may be met either by making a real reduction in emissions or by increasing the sequestration of some terrestrial ecosystems.

There are good prospects of reducing emissions in the short term by improvements in energy conservation and in the longer term by devising an energy strategy which is less dependent on fossil fuel. Since the energy crisis of the early 1970s, the developed world has been slow to embrace new energy technologies because there are still large supplies of fossil fuel. Even in Europe transportation is still dominated by the CO₂-emitting automobile rather than by public transport which emits less carbon per capita per kilometer traveled (Table VIII). The United States has a human population of 270 million, which is almost matched by a vehicle population of 210 million. The annual average mileage is 11,800; total gasoline consumption is 1.46 X 10⁹ gallons, and the carbon emissions exceed 1.8 tC a⁻¹ per head of population.

Transport, however, is often used as a scapegoat. It is the most visible form of CO₂ emission but not the largest. In the United Kingdom, the per capita emissions expected in the Year 2010 are as follows (tC a⁻¹): industry and business, 1.45; energy supply, 0.98; domestic, 0.78; transport, 0.65; agriculture and forestry, 0.43; and public sector, 0.16.

<table>
<thead>
<tr>
<th>Table VIII</th>
<th>Approximate Carbon Costs of Traveling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>Carbon cost (kg C/person/km)</td>
</tr>
<tr>
<td>Automobile</td>
<td>0.099</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Bus</td>
<td>0.024</td>
</tr>
<tr>
<td>Subway/metro</td>
<td>0.030</td>
</tr>
<tr>
<td>Train</td>
<td>0.009</td>
</tr>
<tr>
<td>Plane</td>
<td>0.087</td>
</tr>
<tr>
<td>Plane</td>
<td>0.044</td>
</tr>
</tbody>
</table>
A. Forest Sequestration

The rate at which CO$_2$ is increasing in the atmosphere could be moderated by (i) slowing down deforestation and protecting existing forests, (ii) expanding the forested area, and (iii) substituting fossil fuel-based products with those made of biomass.

The prospect of growing trees to sequester carbon is appealing for many reasons. First, the technology is established and "safe." Second, the "Kyoto" forests thus created may have significant other benefits, such as environmental enhancement, increase in biodiversity, and ultimate economic reward when timber is harvested. Forests are potentially useful sinks for carbon but vary greatly in their capacity to absorb carbon dioxide. Well-managed plantations show an uptake of 1–5 tC ha$^{-1}$ per year over their life span. A simple calculation, however, will show that the forests would need to be on a large scale to be globally useful. If we wanted to absorb the total annual emissions in excess of 6 Gt, and if we assume fast-growing plantations (5 tC ha$^{-1}$ per year), we would need more than $10^8$ ha of land, which is 40% more than the area of the entire Amazon basin. This argument shows that we must regard forests as only one of several measures that must be taken to reduce CO$_2$ concentrations in the atmosphere. Third, there is the question of what should be done with the trees at the time of harvest (50–150 years in the future). Some forest products (newspaper and chopsticks) are short-lived, but many others are long-lived (buildings, furniture, and carvings). A shift from the former to the latter is required. Timber could replace cement and concrete to save the CO$_2$ emissions that occur from the cement industry.

Currently, the Kyoto Protocol allows only certain types of forest. In order to qualify, a forest must be within an Annex 1 country and must be a new forest, constituting "afforestation, reforestation, and deforestation" since 1990. The accounting will be done in the commitment period defined as 2008–2012. Thus, if a country were to establish a large forest in 2000, and if the forest had grown and accumulated carbon by 2012, it would receive "forest credits" which could be set against its emissions. Under the terms of the treaty, these credits could be traded, in much the same way as stocks and shares are traded.

How would governments control emissions from industry? Currently, this is not very clear. There could be taxes on carbon emissions or (more likely) the relevant government agency would issue pollution permits. These permits would define the upper limit of greenhouse gas emissions, beyond which the industry would be in breach of the law. An industry with less than its limit would accumulate a carbon credit, which could be traded with another industry that exceeded its limit. Industries could acquire land for afforestation and thus receive forest credits.

The mode of operation of such schemes is still a matter for discussion. The Intergovernmental Panel on Climate Change is expected to report on the matter in 2000. The following general difficulties and specific loopholes in the operation of the protocol have been noted:

1. Many environmentalists believe that forest sequestration avoids the fundamental issues of how to reduce emissions and how to develop sustainable world economies. There is an important question of whether substantial reductions in emissions are possible without sacrificing economic competitiveness and growth. In the United States, for example, which has a reduction commitment of 7%, members of Congress who opposed the treaty were quick to draw attention to the likely economic impacts that might occur if the United States were to attempt to meet its emission reduction commitments.

2. The Kyoto forests need to be closely monitored, perhaps by satellite surveillance but also by site visits, to verify that carbon has indeed been sequestered as claimed. This requires measurements of belowground as well as aboveground carbon stocks. Currently, scientific knowledge of carbon accumulation over the whole lifetime of a forest stand, including planting, tending, and harvesting, is probably not adequate. What government agencies would be responsible for monitoring, and what would be the economic cost of monitoring?

3. There are loopholes in the protocol that permit counterproductive actions to occur. The discontinuity of commitment periods could be exploited by selling the older forests between periods, thus enabling the accounting period to start again. Moreover, forests that have been disqualified because they were planted in earlier times might now be prematurely felled so that planting can occur within the accounting period.

4. Another way to manage the carbon cycle is to reduce tropical deforestation, but this is currently not allowed for in the protocol. Deforestation is likely to continue at current rates and even accelerate. Would it not be better to explore the management of the carbon cycle by prevention of deforestation in the tropics? In fact, there are several such initiatives in the tropics that predate Kyoto. For example, power utilities, mining companies, automobile companies, and other commercial enterprises as well as the World Bank are paying...
farmers or government agencies in Bolivia, Mexico, Brazil, and Uganda for carbon sequestration at a rate of $10–20 U.S. dollars per ton of carbon sequestered. Some of these have established new forests. Others have simply prevented the deforestation that would have occurred without the project, counting this as carbon gained.

B. Ocean Sequestration

Sequestration in the ocean has been discussed, but currently our scientific knowledge of marine biogeochemistry is not adequate to predict the consequences of any large-scale actions that may be taken. Discussions have centered on the use of iron to fertilize the ocean, following the demonstration in the IronEx project that the phytoplankton in the mid-Pacific are limited by iron and that they "bloom" spectacularly when a solution of iron is experimentally added from a ship. Most of the carbon fixed in photosynthesis would be rapidly released by respiration, but dead biota and feces would sediment and contribute to the biological pump, thus reducing the outgassing of CO2. There is a secondary effect of stimulating the growth of phytoplankton. Some of these organisms produce dimethyl sulfide, which is chemically transformed in the atmosphere to aerosol sulfuric acid particles that form condensation nuclei for clouds. Thus, stimulating the phytoplankton might increase cloudiness over the ocean, which in turn would increase the reflectivity of the planet and cause a cooling effect. Although some people have been enthusiastic about the prospects of managing the carbon cycle in this way, it would be folly to attempt large-scale ocean fertilization in the absence of better knowledge of the functioning of marine ecosystems and the working of the biological pump.

C. Geological Sequestration

Sequestration of carbon in geological reservoirs is being explored. Technologies rely on the availability of a stream of CO2 which would otherwise be dispersed in the atmosphere because the cost of concentrating CO2 ab initio would be high. Target reservoirs include depleted oil and gas fields, aquifers, and the floor of the ocean. For example, in a process to recover methane from oil wells and deep coal seams (so-called coalbed methane), a stream of CO2 is used to flush out the methane. Instead of releasing it to the atmosphere, it can be sequestered in methane-bearing coal seams in which it is apparently stable. For example, it is estimated that the San Juan basin in the southwestern United States has a total capacity of about 2 GtC, and costs are said to be comparable to the cost of sequestration by forests. There are clear attractions to geological sequestration for oil and gas companies where appropriate geological strata exist near existing pipelines and a stream of CO2 is available.

VI. WILL THE SINKS INCREASE, REMAIN THE SAME, OR DECREASE?

In this final section, I raise questions that have not been addressed but which now seem to be urgent research questions.

1. What is the future of the terrestrial sink? Our understanding of the limits of the carbon sink is poor. In particular, it is difficult to predict the response of the sink over long periods of time, especially in future atmospheres of elevated CO2. It is clear that terrestrial photosynthesis responds to elevated CO2 in the range 360–700 ppm, and that in most cases growth responds to twice-normal CO2. However, at what CO2 concentration can this response be expected to saturate? It seems likely that nutrient supplies ultimately will become limiting in terrestrial ecosystems in the way that they are now limiting in the ocean. Models of plant growth suggest that climate warming will, in many cases, increase ecosystem respiration to a point where it exceeds photosynthesis, thus turning the sink into a source, and field observations suggest that this may have happened already in tundra ecosystems.

2. Species changes and acclimation to CO2: To what extent will ecosystems acclimate to elevated CO2 and temperature by species changes and reorganization, which may enable them to retain their sink strength? Answers to this question are likely to be suggested by studies of effects of past climates on vegetation, but current climatic changes are occurring at a much faster rate than before, and therefore the past may not (in this case) be the key to the future.

3. How secure is the soil carbon? The carbon content of the soil may be susceptible to climate warming and also possibly to oxidation on being disturbed by cultivation. The behavior of those pools which turn over slowly is not well understood, nor is the transport of organic material from the soil to the rivers and ocean.

4. How stable is the ocean? Questions of stability are important in the ocean carbon cycle because disturbance of the circulation patterns could result in large releases of stored carbon to the atmosphere or large
changes in uptake at the surface. It is thought, for example, that an increase in the rate of downwelling at the start of the last glacial epoch caused additional flux from the atmosphere. It is also thought by some scientists that the gulf stream, which brings warm water from the tropics to northern Europe, may have shifted substantially in the past. We do not know the stability of these circulations.

Very few authors have attempted to make predictions of global CO$_2$ concentrations and temperatures over periods of more than 50 years because the uncertainties are too great and technologies advance rapidly. An historical perspective illustrates this point. One hundred years ago in London, the environmental issue of the day was the inexorable accumulation of horse feces in the streets, hardly a matter of concern 50 years later in the post-Ford era. Attention then switched to atmospheric pollution, known as smog, caused by burning high-sulfur coal. Smog caused death and disease, but as a result of legislation the hazard was removed.

The track record of the scientific community in forecasting the future is not good. Scientists who have predicted “doom and gloom” in recent decades have generally been proved wrong. For example, the Club of Rome used a state-of-the-art computer model in 1972 to forecast the resource limitations of human population growth, and they forecast an abrupt downturn in the quality of life by 2000. However, economic and population growth are still occurring, despite their predictions to the contrary. They failed completely to predict the major environmental concerns about global warming that we are now discussing at the start of the new millennium.

See Also the Following Articles

- ATMOSPHERIC GASES
- BIOGEOCHEMICAL CYCLES
- DEFORESTATION
- ENERGY FLOW AND ECOSYSTEMS
- MARINE ECOSYSTEMS
- NITROGEN AND NITROGEN CYCLE
- PHOTOSYNTHESIS, MECHANISMS OF

Bibliography

CARNIVORES

Hans Kruuk
Institute of Terrestrial Ecology

I. Species Diversity
II. Size and Ecology
III. Foraging and Phylogeny
IV. Carnivore Guilds in Ecosystems
V. Carnivore Social Systems
VI. Changes in Diversity

GLOSSARY

basal rate of metabolism The minimum amount of energy spent by an adult animal that has not eaten recently, at normal body temperature during rest (usually sleep).

eutherian Mammal in which the embryo is attached to the mother by a placenta

fissipedia suborder of the Carnivora with divided toes.
guild Species in a community that use similar resources.
marsupial Mammal without placenta and with a pouch to carry the young.
monophyletic Derived from a single ancestor.
pinnipedia Suborder of the Carnivora with fin-like limbs (seals, sea lions, and walrus).
plantigrade Walking on the soles of the feet.

CARNIVORES are a highly varied group of closely related species. This article discusses the diversity of terrestrial carnivores, excluding the seals but including some marsupial species. Most of the terrestrial carnivores belong to one single order: Carnivora. Some of the taxonomic and evolutionary relationships are discussed as well as the social organizations, effects in ecosystems, and conservation status.

I. SPECIES DIVERSITY

Among the mammals that are broadly referred to as Carnivores there is a weasel of approximately 45 g and a polar bear of up to 700 kg, approximately 15,000 times heavier. There are species living almost permanently in water (sea otter), in trees (palm civet), or in deserts (fennec fox); some eat buffaloes, some eat beetles, and some eat bamboo. It is not surprising, therefore, that with such large variation and divergent adaptive features there is disagreement among researchers about the evolution and classification of species. Classification is based on morphological evidence (dental characteristics, anatomy of the skull base, and other morphological features) and on molecular genetic information.

This article will deal with the approximately 230 Fissiped carnivores, i.e., the terrestrial species of the order Carnivora and excluding the seals, sea lions, and walrus (Pinnipeds). It will also discuss a group of ecologically rather similar species of marsupials in Australia belonging to the families Dasyuridae and Thylacinidae.

The Carnivora are a monophyletic order, descended
from the family Miacidae approximately 60 million years ago in the Palaeocene. The order has two main branches, the dog-like Canoidea and the cat-like Fel- 
ioidea. The 33 seals and sea lions and walruses belonging to the Pinnipeds are sometimes included in the order Carnivora and sometimes given separate status. They evolved from the Canoidea, but there is disagreement about which terrestrial carnivores are their closest rela-
tives, the main candidates being the bears and the mustelids.

The Carnivora are generally divided into seven fami-
lies, although some taxonomists recognize more. The Canoidea include the Canidae (dogs) with 39 species, the Mustelidae (martens) as the largest carnivore family with 67 species, the Ursidae (bears and pandas) with 9 species, and the Procyonidae (raccoons) with 15 spe-
cies. The Felidae comprise the Felidae (cats) with 35 species, the Hyaenidae with 4 species, and the Viver-
ridae (mongooses and genets) with 66 species. The main taxonomic disagreements are over the Ursidae and the Viverridae. The two species of panda are often thought not to belong to the bears but to deserve a separate family or they may be included with the Procy-
onidae, and the Viverridae are often divided into two families—the Herpestidae (mongooses, 31 species) and the Viverridae (genets and civets, 35 species).

The dogs and foxes (Canidae) constitute a highly mono-

omorphous family. All are very dog-like with non-
retractile claws, and all are coursing predators, such as 
wolves, coyote, jackals, foxes, wild dog, bush dog, and 
maned wolf. All species have much in common in their 
ecology and social behavior, although they may vary 
from solitary to gregarious. Sizes vary between that of 
the large gray wolf (up to 80 kg) to that of the tiny 
fennec fox of little more than 1 kg. Canids occur on 
all continents, and with the dingo they even fielded an 
early introduction in Australia. The gray wolf is the 
evoluted more recently than the other carnivores and 
successfully colonized areas as far apart as the drift ice 
in the Arctic and the dense forests of the Old World 
and neotropics. Bears of colder regions hibernate. More 
than half of the species (five or 56%) are endangered 
or vulnerable, and all are in trouble in at least part of 
their geographic range.

Members of the raccoon family Procyonidae are all 
rather long-bodied animals, relatively small (up to 8 
kg), stockily built, plantigrade with nonretractile claws, 
and very short tails. The exception is the very aberrant red 
panda (a small arboreal species with a long tail), a 
species often classified as a procyonid. All but the polar 

tockily built, plantigrade with nonretractile claws. They 
are usually slimly 

built, some with retractile claws, and with a very distinc-

tive bouncing gait, but the family is quite variable and 
some species are stocky, such as the badgers. Also, 
in feeding and social behavior they vary greatly, from 
solitary predators on mammals to group-living animals 
feeding on earthworms or fish.

Stoats and weasels have been introduced on several 
islands outside their normal range, and the North Amer-
ican mink is now an abundant exotic in many places 
throughout Eurasia and South America. Some of the 
mustelids are almost extinct (e.g., the black-footed fer-
ret and the European mink), some were exterminated 
recently (sea mink), and several are in trouble (otters) 
or have just returned from the brink of extinction (such 
as the largest of them all, the sea otter). However, only 
seven species (18%) are listed as endangered or vul-

erable.

The bears and pandas (Ursidae) are comparatively 
large and some are huge (weighing up to 700 kg), 
stockily built, plantigrade with nonretractile claws, and 
very short tails. The exception is the very aberrant red 
panda (a small arboreal species with a long tail), a 
species often classified as a procyonid. All but the polar 

bear are mostly vegetarian; they are solitary animals 
with a fairly simple social organization. The bears 
evolved more recently than the other carnivores and 
are more specialized than the Procyonidae in feeding and social behavior; they vary greatly, from solitary to gregarious. Sizes vary between that of 
the large gray wolf (up to 80 kg) to that of the tiny 
fennec fox of little more than 1 kg. Canids occur on 
all continents, and with the dingo they even fielded an 
early introduction in Australia. The gray wolf is the 
evoluted more recently than the other carnivores and 
successfully colonized areas as far apart as the drift ice 
in the Arctic and the dense forests of the Old World 
and neotropics. Bears of colder regions hibernate. More 
than half of the species (five or 36%) are endangered 
or vulnerable, and all are in trouble in at least part of 
their geographic range.

Members of the raccoon family Procyonidae are all 
rather long-bodied animals, relatively small (up to 8 
kg), stockily 

building with nonretractile claws. They occur naturally only in the New World. In addition to the various species of 
raccoon throughout the Americas, there are also other 
Neotropical species, such as the coatis, the kinkajou, 
and the cacomistle. They are carnivorous and insectivo-

tous as well as herbivorous. Their social behavior is 
quite variable; most are solitary but some, for example, 

female coatis, live in large packs. Also, common rac-
coons may occupy winter dens in groups. None are 
classified as endangered, although some raccoon species 
are known from only one or two islands, on which 
they may be rare. In several areas of Eurasia common 
raccoons are now frequent as an introduced species.

The Viverridae are a large Old World family of small 
animals, occurring almost entirely in Africa and Asia. 
They are the most "primitive" carnivores, i.e., the closest
relatives of the carnivores’ original forebears, the Miacidae. The family includes many species of genet and civet, linsang, binturong, a large range of mongooses, meerkat, fossa, fanaloke, and others. Some taxonomists recognize the mongooses as a separate family—the Herpestidae (otherwise classified as a subfamily Herpestinae). The mongooses are long-bodied and look somewhat mustelid-like, with nonretractile claws; in contrast, the rest of the viverrids look and walk much more cat-like, with retractile claws. The fossa from Madagascar is unusually large (up to 20 kg), but all other Viverrids are much smaller, especially the mongooses. Most are solitary, but some mongooses and meerkats live in large packs. Most of their food is invertebrate, but they also take small mammals, etc. and occasional vegetable food.

Few viverrid species appear to be endangered (four or 6%); of these four, three are on Madagascar, an island beset by deforestation problems.

The Hyaenidae have only four species left; despite a fossil record showing large numbers in the geological past, and they occur only in the Old World. Three are relatively large species of 50 kg or more and are typical coursing animals with a dog-like build and nonretractile claws. Hyaenids are carnivorous; they often scavenge, although some also take vegetable food, and the smallest, the aardwolf, is a termite specialist. The largest species, the spotted hyena, is a gregarious hunter of large mammals, but the others are solitary. One of the four, the brown hyena, is classified as vulnerable as a species, but in many countries the others are also endangered.

Finally, there are the Felidae, the proper cats, a highly monomorphic family with many species. All are built as stalking predators, with retractile claws, and they range in size from fairly small to large, from serval and flat-headed cat and many other small species to lynx and tiger (up to more than 300 kg), various leopards, cheetah, and lion (up to 250 kg). All are carnivorous, but small species also eat many invertebrates. The social organization of all species is very similar—solitary and territorial, with the lion as the only group-living exception. Felids occur naturally on all continents except Antarctica and Australia. The wild cat Felis lybica has been domesticated and it has been introduced virtually everywhere in the world; it is now the most widespread carnivore.

There is concern about the conservation status of almost all cat species, especially the ones with desirable fur. Twelve (34%) are endangered or vulnerable, including all the large ones such as tiger, cheetah, jaguar, and various leopards, but not the lion.

The mammalian carnivore families outside the Carnivora proper are the Australian marsupials Dasyuridae and Thylacinidae. There are four dasyurid species that are very similar to Carnivora: three quolls or, as the Australians call them, “native cats” and the Tasmanian devil. They are relatively small, with the quolls similar in size and appearance to martens or mongooses and the Tasmanian devil more like a small badger in size and shape, and they are typically carnivorous, solitary, nonterritorial animals. The thylacine was the only species in its family and it was also the largest of these marsupial carnivores (comparable in size to a large jackal or coyote). It was exterminated very recently. The other, dasyurid species now have very much reduced geographical ranges and they are vulnerable.

II. SIZE AND ECOLOGY

There is a large diversity of sizes among carnivores, and the ecological implications of this variation are considerable. The basal rate of metabolism of carnivores tends to be greater than that of same-sized herbivores, and a large predator expends more energy than a small one; therefore, it has to capture more prey. However, the increase in metabolism with size is not linear, and when expressed as energy requirements per kilogram of body weight a large predator is more efficient than a small one (Fig. 1).

![Figure 1](https://example.com/f1.png)
The body mass of a carnivore, when combined with the group size in which it lives, is correlated with the size of its home range, although there is much variation (Fig. 2). Much of this residual variation is due to effects of habitat and to differences in diet (predators of vertebrates having larger ranges than insectivorous ones). There is no significant correlation between body mass and gregariousness, i.e., large carnivores are no more likely to live in groups than are small ones.

Brain size of carnivores increases with body size. Irrespective of this, a meat-eating, hunting carnivore or an omnivorous species generally has a relatively larger brain than an insectivorous one. Bears have relatively large brains, whereas Viverrids and Hyaenids have relatively small ones and the others have intermediate-sized brains.

In general, large carnivores take larger prey than do small ones, but there are important variations. First, some of the largest (bears) are mostly vegetarian. Second, the trend varies between carnivore families. Within the mustelids there is a nonsignificant negative correlation between predator body mass and prey size, so the predator–prey size principle does not apply. Both felids and canids show a strong positive correlation, but the slope of the relationship is different: The increase in prey size with predator body mass is much steeper in felids (Fig. 3).

There are many specific deviations from the previously discussed general patterns, but the trends are significant. Greater energetic demands on larger carnivores are met by a dependence on larger prey, which is likely to make the large predators more vulnerable. Many of the larger carnivores are threatened globally or in parts of their range, e.g., many of the large spotted cats and the tiger, wolves, and hyenas.

III. FORAGING AND PHYLOGENY

The diet of carnivores in general, and that of their marsupial equivalents in Australia, consists of vertebrate prey but also invertebrates and vegetable matter. The order Carnivora shows dental specializations enabling easy digestion of vertebrates, such as large canines and the carnassial shear, but in several species the molars have been further adapted to a more grinding function when eating plants.

Foraging or hunting behavior of carnivores consists of variations on a general theme. A search leads to detection and selection of a potential victim or other source of food, and it is followed by an approach which may contain elements of stalking and/or chasing. The actual capture of prey may include seizure, immobilizing, and killing, and this is followed by eating, taking food to cubs, or sometimes quietly caching it for later consumption. Parts of this sequence may be absent, and in fact most carnivores will just search and then eat small food items without much further ado. However, even if there is a full-blooded hunt, some species never stalk (e.g., dogs or hyenas), others never chase (e.g., cats), and some may not show any specific killing behavior but just eat (e.g., hyenas).

The chain of events aimed at the capture of prey may be broken off at any stage depending on circumstances; it can also be started at any stage. It is highly adaptive, depending on prey and environment, with the predator’s own motivation (its degree of hunger) apparently affecting especially the early searching stages of the hunting sequence.

Conspicuous in the predatory behavior of many carnivores is the phenomenon of surplus killing, i.e., kill-
The relationship between size of predator and prey. For each species of carnivore, mean body weight and main food category are given. 1, invertebrates; 2, small rodents; 3, rabbits and hares; 4, larger mammals less than 50 kg; 5, mammals more than 50 kg (reproduced with permission from Kruuk, 1986).

ing more than is required for immediate consumption. Classic examples are the fox in the hen house or gull colony or a lion among a cattle herd; it has been described for hyenas, polar bears, wolves, leopard, and others. Large numbers of animals were killed without being eaten. All these situations have in common a lack of defense by the prey: The prey may be immobilized by particular weather conditions, it may be penned in, or it may have lost its antipredator defense through domestication. In these cases, the predator is sated and no longer hungry, but hunger normally affects only the early stages of the carnivore hunting sequence, especially the search. If, for some reason, no search or stalk or chase is needed because the hunter suddenly finds itself close to the quarry, then the rest of the hunting sequence is put in train irrespective of hunger, and there is no inhibition to capturing and killing. Functionally, such events are wasteful from the carnivore’s point of view because they reduce prey availability without the predator getting the benefit.

Food caching (i.e., the storage of food in a hidden place) is a behavior pattern that limits this waste and utilizes the consequences of surplus killing. Many species do it, and it is in many different ways. In all canids it is highly stereotyped: A small hole is dug, and one single prey is dropped into it and covered with earth or vegetation by sweeps of the snout. In other families there are many variations on the caching theme: Leopards take a carcass high into a tree; spotted hyenas cache chunks of food in shallow water; brown and striped hyenas push it into a dense bush; stoats, mink, and other mustelids may make large stores by dragging numbers of prey into a single hole; wildcats may put remains of their quarry under a log; and some of the larger cats may cover a carcass with vegetation. The methods are consistent within the species, for canids within the family, and for martens within the subfamily, suggesting that caching has evolved in carnivores on several different occasions. There is no evidence of caching for the procyonids or viverrids, nor is it done by the marsupial carnivores.

Although food caching may use some of the surplus kills, it still does not utilize all the apparent waste. Not only are some of the caches never revisited by the perpetrator but also in the larger surplus kills only a small proportion of the victims are stored. There may be scores of dead gulls left by a fox and of gazelle left by spotted hyenas. In some carnivores caching is particularly highly developed. For instance, foxes are able to remember where they stored what, and they return preferentially to the more desirable cached items. Foxes also use some kind of bookkeeping system for
their caches, leaving a drop of smelly, long-lasting urine near those caches which they have emptied.

Foraging and hunting behaviors are features that differ with the phylogeny of the predator. Canids invariably are coursing and running predators, with an occasional semistalking approach of prey, and they often forage by "sniff and search." At the other extreme, felids search almost entirely dependent on vision and they approach their quarry in a highly concealed stalk or ambush. Most members of the other families, including the marsupial carnivores, show the canid sniff and search behavior pattern. Only the genets and civets have a stalking behavior that is similar to that of the felids.

When a relatively small vertebrate prey is caught, the felids, many of the mustelids, and civets will kill it by severing the spinal cord with their canines. Canids also have a specific killing method—violently shaking the prey. Felids kill large prey with a throat bite, or they suffocate it with a bite over the nose and mouth. In almost all other predator–prey interactions there is no specific killing or immobilizing behavior—the predator just eats the prey. Most species of carnivore have highly specific ways of dealing with a prey and its carcass, and it is often possible to distinguish afterwards which predator was responsible for a kill.

Some subfamilies have evolved extreme foraging specializations. The Lutrinae (otters) dive for fish and crabs using their tactile senses, but essentially their hunting is also based on the canid pattern. Several Melinea and Mellivorinae (badger) species pursue their prey by digging after it. Species such as the otters and African wild dogs use energetically demanding behavior to catch their quarry (a high investment and high reward strategy), which makes them especially vulnerable to fluctuations in prey density and to food loss to scavengers.

As a result of the variation in foraging methods between families and subfamilies, there are also phylogenetic patterns in the diet. Compared with many other species of mammals and birds, diet analysis for carnivores is relatively easy, although it has some serious problems. Food usually consists of clearly discrete items, which can be recognized and quantified. Furthermore, scats are often easy to find, no matter how elusive the animals may be, and scat analysis has become a major tool despite the difficulty of relating scat content to diet in a quantitative manner. Also, for many species it has proved possible to obtain direct, quantitative observations of predation and foraging behavior. The result is an extensive body of knowledge of carnivore diet, the important link in the relation between the predators and their environment.

One summary showed that of 111 species of carnivore (from all families), only 36% could be classified as predominantly meat eaters, i.e., taking more than 60% of their diet in the form of other mammals or birds. Indeed, in that analysis among representatives of families such as the bears, raccoons, and viverrids, there were no proper meat eaters at all, and they were found to feed on insects, vegetation, or a mixture of various food categories. Many species were called omnivorous if their diet did not include 60% of any one category of food. However, such a limit is quite arbitrary, and many important species were left out of the analysis because of lack of information. Therefore, it is an oversimplification, and one could also summarize the diet differently (Fig. 4): The majority of carnivore species will eat meat and will prey on other mammals at some time or other. However, the point had to be made that many other kinds of food are involved.

There is a clear importance of phylogeny in the diet: The food of a felid tends to be more like that of another felid than that of a canid and vice versa, and this applies to several families (Fig. 4). All families except the felids are intensive exploiters of vegetable and invertebrate food sources: The Feltidae are the most exclusively carnivorous—the ultimate predators. Only the felids, hyaenids, and canids feed significantly on large mammals, with an occasional exception in the other families. The viverrids, which are probable ancestors of felids and hyaenids, do not kill larger mammalian prey but they are either insectivorous or have a very mixed diet. The diets of the four hyenas are more different from each other than between the members of the other families, and their specializations range from wildebeest to termites or melons or carrion. The bears and pandas are vegetarians. Species in the raccoon family all have a mixed diet, including much vegetable matter. Canids have mostly varied diets, composed of insects, fruits, and mammals. Many of them are proper meat eaters, but even these often include some vegetable matter, quite unlike the cats.

Some of these family-specific trends are further refined in the subfamilies. The mostly meat-hunting mustelids also include two subfamilies with 9 badgers, which almost all feed on invertebrates and vegetation, and a subfamily of 13 otters that subsist on fish or crabs. Not only are there differences between families and subfamilies in the kinds of prey or vegetation which they select, but also the degree of specialization varies (Fig. 3): Specialization can affect an animal’s vulnerability to environmental change. Among the canids, each species uses an average of 6.5 (out of 10) major prey categories, each constituting at least 1% of its food. On the other end of the scale, each bear only uses 3.7 prey
CARNIVORES

FIGURE 4 The use of different foods by species from different families. Percentage of species in each family in which a particular food category constitutes more than 1% of diet. Veg, vegetable foods; inv, invertebrates; amp, amphibia and reptiles; rod, small rodents; lag, rabbits and similar size prey; s.mam, large mammals smaller than 50 kg; l.mam, large mammals larger than 50 kg; car, carrion (reproduced with permission from Kruuk, 1986).

categories and each felid 4.0, so they are much more specialized. The other carnivore families are intermediate. This specialization (dependence on a few resources) may make bears and felids more vulnerable to environmental change than canids.

The degree of specialization can be described only in the broadest of terms because it is difficult to measure and quantify. Terms such as omnivore, opportunist, generalist, and specialist have no absolute values, and they refer to animals which may be selecting from a variable set of availabilities. There is also a problem with definitions. For instance, the Eurasian badger is highly focused in its food selection in any one area, concentrating entirely on earthworms in northwestern
Europe, on rabbits in southern Spain, and on olives in northern Italy. There is no doubt that in each of these areas badgers are highly specialized, but their specializations are different in different places and overall they could be described as opportunists. Currently, there is no good quantitative descriptive term for such a pattern.

Nevertheless, despite inadequate terminology it is recognized that in any one place some species use many more different prey categories than others. For instance, a cheetah on an African savanna takes almost only small or medium-sized ungulates on open grassland, whereas a leopard in the same area is much more catholic in its tastes. It eats the same mammals but also much smaller mammals as well as birds and snakes, stalking them in the open and in dense bush or between the rocks. Along European streams a mink will eat small mammals, frogs, fish, birds, and insects, whereas along the same banks the (larger) otters feed almost exclusively on fish and frogs, clearly much more focussed. Such a broad comparative indication of specialization suggests the dependence of a predator on few or many prey categories, even though there may be difficulties with labeling as specialist or opportunist.

In general, because closely related animals often have similar food habits, food selection may be termed a conservative characteristic in the evolution of carnivores. There are many exceptions; for instance, among the meat-hunting felids there is a fishing cat, and the hyena family includes the aardwolf, which feeds on nothing but termites, the spotted hyena, which is an exclusive large ungulate hunter or scavenger, and the striped and brown hyenas, which are as catholic in their tastes as possible. However, these exceptions do not invalidate the overall importance of ancestry in the animals' environmental relationships.

### IV. CARNIVORE GUILDS IN ECOSYSTEMS

Closely related members of the same carnivore family tend to exclude each other, although there are important and puzzling exceptions. Of course, competition between natural populations of different carnivores is rarely observed, probably because those cases in which it occurred have long ago come to their natural conclusion, i.e., the demise of one of the contestants. However, when perturbations occur competition may be obvious.
For instance, the introduction of the American mink in Europe totally removed the European mink from most of its range. The famous wolf population of Isle Royal arrived there in the early 1950s, and it completely replaced a population of coyotes. There are several observations of wolves killing coyotes, apparently without the coyotes being eaten. In many other areas in North America, just one of the two species is found, although usually these places would appear to be suitable for both. Coyotes tend to replace red foxes.

Red foxes are aggressive to arctic foxes and exclude them when their areas overlap, but arctic foxes can feed and survive at much lower temperatures. Recent red fox increases (e.g., in northern Scandinavia) caused the demise of arctics over large areas. In the 1970s, when the black-backed jackals of the Serengeti were decimated by disease, the very similar and previously rare side-striped jackal population increased dramatically. Tigers are reported to often exclude leopards, and different species of otters exclude each other.

Also, species that are not closely related but have overlapping ecological niches may affect each other. For instance, the African wild dog disappeared from the Serengeti between the 1960s and 1990s probably partly because of an increase in populations of hyenas and lions.

Exceptions to this pattern of exclusions are the three species of African jackal. They are very similar and closely related; nevertheless, their geographical ranges overlap considerably and they can be seen eating from the same carcass.

There are other examples of perturbations or species introductions causing the disappearance of competing carnivores, but such events are relatively rare. Thus, it is often assumed that populations of predators each have their own ecological niche and rarely affect each other. However, the ecosystems that we see are end results, and often what we see may be the status quo long after earlier populations have been wiped out or have been prevented from moving in.

One way in which predator species adapt to an ecological niche in a particular habitat is through morphological or behavioral variation. When potentially competing species share a range, differences between them (e.g., in size) may be greater than those from animals that live in ranges not shared with the other species. This is known as character displacement. One of its consequences is that a guild of predators in any one area a character, such as body size or size of the canines, shows a fairly evenly spaced stepwise distribution among species (Fig. 6). The obvious effect of this is to minimize competition. Such divergence of morphological characters does not always occur, however, and its striking demonstration in British mustelids is contrasted with an almost absence in Serengeti jackals. The marsupial carnivores of Tasmania also show a clear character divergence.

The existence of such structured mechanisms, to avoid competition within predator guilds in ecosystems, suggests that resources for the carnivores may be at a premium. This in turn presents the possibility that the predators depress prey populations.

It is often argued that predators have little or no effect on prey numbers, as demonstrated by ecosystems such as the Serengeti with more than 25 carnivore species that coexist with the many prey in apparently stable populations. However, such an apparent stability is the result of interactions over a long period of time, and it is possible that many prey species were extinguished in the past and many predator–prey relationships have

---

**Figure 6** Even spacing of mean lengths of skulls of Mustelid species in Britain and Ireland, log scale. The mink has recently been introduced (reproduced with permission from Dayan and Simberloff, 1994).
The deliberate or accidental introduction of carnivores into ecosystems in which they were previously absent provides an opportunity to demonstrate effects of predation. Introduced predators (cat, dog, red fox, mongoose, American mink, stoat, and ferret) have caused extinctions and declines of endemic prey species throughout the world, including many birds, mammals, and reptiles. There is little doubt, therefore, that carnivores cause extinctions or prevent populations from becoming established, and they must have done so in the evolutionary past. On all continents the fauna would look very different from what it is now if there had been no carnivore predators.

Also in the predator–prey systems which currently exist, the effects of carnivore predation can be far-reaching. Predator removal often dramatically increases numbers in prey populations compared with those in suitable control sites, as demonstrated by foxes preying on partridges. Pest species such as rabbits can be limited in numbers by carnivores operating in conjunction with other mechanisms such as disease. Wolves may exercise major effects on populations of caribou, moose, and white-tailed deer, and spotted hyenas can have a major effect on wildebeest. Often, other limiting factors are involved simultaneously, e.g., predators may be the means whereby a herbivore population is maintained at the carrying capacity of the vegetation.

V. CARNIVORE SOCIAL SYSTEMS

The majority of carnivores are solitary and territorial. Their spatial organization is maintained by females (with or without offspring) defending a range against other females of the same species, and by males defending a (larger) range against other males. Male ranges may overlap with several females ranges. Territorial behavior involves scent marking and visual displays as well as direct aggression.

There are some important variations to this generalization. The marsupial carnivores are not territorial; therefore, instead of a regular spacing between individuals, they are more or less randomly distributed. This begs the question, not satisfactorily answered to date, why eutherian carnivores spend so much time and effort on risky territorial behavior if other species with similar ecology can do without.

The organization of canids is pair based rather than solitary; a male and a female share the same range. This is also the only family in which males usually assist in providing the offspring (by regurgitating food). The fact that such pair behavior is tied with the species' phylogeny suggests that in individual species it is not necessarily adaptive.

From the simple, territorial arrangement a group organization has evolved independently in several species in all carnivore families except the bears, and such group living has complicated land tenure considerably. Among the canids the most striking examples are the wolf, the wild dog, and the dhole in Asia, but in several other canid species offspring may also remain for 1 or 2 years, overlapping in time with subsequent litters of the basic pair. This has been described for several jackals, foxes and others.

African wild dogs live in extremely tight packs, often composed of 20 or more individuals, almost always close together, with on average twice as many adult males as females. The wolf has a very different organization. It also lives in sometimes large packs, but it does so in a much more "fusion–fission" type of society, with individuals coming and going, sometimes hunting in groups and sometimes alone, within the pack territory. In both species usually only one female per pack breeds. Group organizations in other carnivore families show as many different patterns as there are species, with, for instance, lions in permanent prides of up to 20 related females joined by small groups of males which are replaced every few years. Spotted hyenas live in female-dominated clans of up to 80 in group territories, and individual members may hunt or join with others or they may be solitary within the group territory. Cubs are cared for almost entirely by their own mothers. Both banded and dwarf mongooses occur in dense packs, with the sexes mixed, and all members of the pack care for all offspring; Eurasian badgers forage on their own and look after their own cubs only, but they live in group territories defended by both sexes. Female Eurasian otters, on the other hand, may occupy individual core areas within a group territory of five or six females, whereas males remain in their own ranges.

For the phenomenon of group living in carnivores, there are no striking, simple phylogenetic trends and no substantial effects of body size, climate, prey size, predation, or various other factors to explain the occurrence of packs, bands, or prides. It does not appear that social species have done any better or worse than solitary ones in terms of numbers or densities, nor is their future survival more or less endangered. There is, however, one set of environmental factors which appear to affect gregariousness, i.e., the distribution of food or of resources in general.

Currently, there is only one general hypothesis to
VI. CHANGES IN DIVERSITY

When the Carnivora first appeared, approximately 60–40 million years ago, there were other, similar animals already well established. For instance, the extinct order Creodonta included families such as the Hyaenodontidae, and there were various large marsupial predators such as the Thylacoleo or pouch lion. Proper carnivorous feeding and predation have evolved several times independently—at least twice among the marsupials (in the Borhyaenidae in South America and some Dasyuridae in Australia) and twice among placental mammals (in Creodonta and Carnivora).

The large carnivorous expansion from the Palaeocene onwards coincided with the evolution of angiosperms, flowering plants and grasses (evolving away from the ferns and gymnosperms such as conifers). This resulted in a large floral diversification, including savanna-type vegetations, and enabled the extensive diversity of ungulates and rodents to evolve. This in turn enabled the evolution of specialist predators.

Most of the other, noncarnivore large predators have now gone extinct, with the last creodonts occurring approximately 8 million years ago and the last really large marsupial carnivores occurring 2 million years ago, when species of Homo were already well established. Of the smaller marsupial carnivores, just a few small dasyurids still occur in Australia. However, while the other predators slowly disappeared, the order Carnivora diversified into a multitude of different families, genera, and species.

Why the creodonts disappeared, while at least initially carnivores thrived and probably replaced the creodonts, is a mystery. Creodonts and carnivores were closely related and the skeletal remains, such as vertebrae and the locomotory system, were similar. The difference between creodonts and carnivores was not much greater than the variation within these groups. However, obviously there is more to an animal than its skeleton, and the reasons for extinction may well have lain in other aspects of morphology, physiology, or behavior.

In the early stages of evolution the advantages appear to have lain with the creodonts, and only after 20 million years did the balance tip in favor of the Carnivora. The Carnivora have been four times as successful as the creodonts. The latter are known from 43 genera spanning approximately 45 million years, whereas Carnivora, excluding living and aquatic genera, are known from 218 genera over a span of more than 55 million years.

Carnivore species are also fewer now than they were in the geological past. In fact, most of the Carnivora have become extinct, and although there is still a rich complement of species, there were many more in the past. For instance, we know of 333 genera in the seven extant families of carnivores, of which 237 (71%) are extant. Many complete carnivore families have also disappeared, just like the creodonts and large marsupial predators.

An interesting phenomenon was the occurrence of saber-tooth species. Saber teeth evolved several times independently in different species, families, and even orders and included the Megantereon, Homotherium, and Machairodus, which were at least as big as a lion, and formidable felds such as the North American saber-tooth cat Smilodon. Some of these were present at the same time as early species of Homo, but all of them are extinct. They pose a difficult problem to palaeontologists. Large canines are used by today's carnivores for killing prey and for social purposes such as fighting opponents over territorial claims. However, were the extra-large saber teeth—the huge, flat daggers which were seemingly far too large for any jaw—used to kill extra-large prey, for opening carcasses, or what? In fossil assemblages it was always the very largest ones, the top predators, which sported saber teeth.

There is no likely explanation for saber teeth in the
acquisition of food. The fragile, sharp weapons, often with serrated inner edges, must have been quite useless against thick skin or on large bodies, with the gape of the owner being insufficient to use these canines effectively. However, saber teeth were obviously effective weapons; otherwise, they would not have evolved several times independently. However, they also disappeared again in all these cases. It is likely that saber teeth made use of some Achilles heel in their prey (of which we have no evidence today), but that in response the prey species evolved means of protection. It was an arms race which was eventually lost by the sabers, but we do not know who conquered and why.

Homo is a highly successful hunter, and mankind was and is in competition with Carnivora at several different levels. Homo also preys on and is preyed on by Carnivora. Therefore, are people the cause of carnivore extinctions? The question is often asked, and answers are far from straightforward. Many carnivore species extinctions occurred approximately 4 million years ago, at the time that hominids arrived on the scene, and many followed during the next 2 million years.

However, many extinctions occurred before man arrived, and at the time Homo’s arrival was not the only event which changed the environment. There were drastic changes in climate, for instance, approximately 3.2, 2.4, and 0.8 million years ago. The Pleistocene Period started at 0.8 million years ago with a massive climatic shift that coincided with the appearance in Eurasia of African species such as the lion, leopard, spotted hyena, and perhaps the major (although not the first) movement out of Africa of Homo. If such major dispersal events occurred in conjunction with climatic changes, it is equally likely that extinctions would have occurred.

For many of the carnivore extinctions, we cannot blame our own species with any conviction. Homo may have been closely involved or not at all, and perhaps environmental change rendered species more vulnerable to competition and predation by mankind. Of course, recent extinctions have been fairly well documented, and here mankind’s guilt is in no doubt. The Tasmanian wolf or Thylacine was exterminated by sheep farmers in the 1920s and 1930s. The North American sea mink was obliterated in the late nineteenth century for its fur, and the sea otter almost followed it into oblivion; it was barely saved and it has recovered fairly well. The “wolf” of the Falkland Islands was still there when Charles Darwin visited, but sheep farmers exterminated it and since the 1880s it has existed only in museums. Several carnivores were totally eradicated from Britain, including the brown bear and the wolf, whereas wildcat, polecat, and pine marten have only just managed to survive. There are long lists of carnivore extinctions from throughout the world, and there is no doubt that most of these were man induced.

Whatever caused the demise of carnivore species in the past, it is important to prevent it from occurring in the future. Several species are on the brink of extinction in the wild, including tiger, panda, European mink, and black-footed ferret, and many more face local extinction. In geological terms, the diversity of carnivores is decreasing extremely rapidly.

See Also the Following Articles

Bibliography


I. Introduction
II. Origin of the Concept of Carrying Capacity
III. Definitions of Carrying Capacity
IV. Do Populations Have Carrying Capacities?
V. Determining the Carrying Capacity of a Population
VI. Current Research on Carrying Capacity
VII. The Importance of the Concept of Carrying Capacity to Biodiversity
VIII. The Human Carrying Capacity
IX. Conclusions

GLOSSARY

density dependence The condition that environmental factors influence population growth rate in relation to population size. Density dependence usually is seen as an linear, inverse relationship between population growth rate and population density (i.e., population growth decreases as density increases) and may occur if individuals compete or predators are more effective as a prey population increases.
density independence The absence of environmental factors that influence population growth as a function of density. This may occur if mortality removes a fixed percentage of a population, independent of population size.
logistic growth Regulated population growth that follows the logistic equation \( \frac{dN}{dt} = rN(1 - N/K) \). Populations growing according to this equation increase rapidly at low densities and the growth rate decreases as they approach carrying capacity (K). population A group of individuals of a particular species that live in a region. A population is usually a subset of the entire species.

population regulation The constraint of positive population growth. The study of population regulation deals with the factors that cause this constraint, such as competition for food or predation.

population stability The tendency for populations to return to a previous size after a disturbance, such as reductions due to hunting or disease or increases due to immigration. Stable populations may be locally stable (return after small disturbances) or globally stable (return after severe or catastrophic disturbances).

CARRYING CAPACITY is the maximum number, density, or biomass of a population that a specific area can support. This number is likely to change over time and depends on changes in environmental factors (e.g., rainfall and temperature), resources (e.g., food, hiding places, and nesting sites), and the presence of predators, disease agents, and competitors over time. The concept of carrying capacity has been explicitly recognized for approximately 150 years and its use has waxed and waned during this time. Currently, the use of carrying capacity to describe any particular population is made only with great caution, although the concept remains
intuitive and fosters questions that address our fundamental understanding of what factors regulated populations over time and space.

I. INTRODUCTION

Populations, or groups of individuals within a species, change over time. There is general agreement among ecologists that population growth is bounded by biotic and abiotic environmental factors that result in approximate, maximum numbers of organisms that can be supported in different habitats. A population’s carrying capacity is difficult to measure and likely varies over time and through space. The concept of carrying capacity has played an important role in the fields of basic ecological research, wildlife management, and conservation biology.

The concept of carrying capacity also involves determining how many people Earth can support. Dialogues about human carrying capacity are often quite contentious and illustrate the difficulty surrounding the concept of carrying capacity. The concept of carrying capacity is alive and well, although some have argued that it should be abandoned altogether. On the surface, the concept is easy to understand and intuitive and, therefore, is likely to stay with us for some time. However, recent developments in our understanding of the dynamics of population change over time have greatly modified what is considered an area’s carrying capacity. Therefore, we need to recognize both the strengths and the weaknesses of the concept of carrying capacity.

II. ORIGIN OF THE CONCEPT OF CARRYING CAPACITY

Humans have long been aware of the limitations of their own population growth. As early as the Old Testament one can argue that the concept of population limitation was recognized. In the book of Genesis (28:3) are the following words: “And God Almighty bless thee, and make thee fruitful, and multiply thee, that thou mayest be a multitude of people.” Although rather vague, as quoted from the King James version, a “multitude” in English is defined as “a great number of people” (Merriam-Webster Dictionary) that eventually would be spread over the earth and would not be an ever-increasing population. The number at which a population reaches and remains sustainable is referred to as the “carrying capacity.”

Recognition of carrying capacity probably occurred long before written history began. It is likely that the earliest agriculturists, perhaps 10,000 years ago, were keenly aware of the number of mouths that an area could sustainably feed and that increasing numbers of people required increases in food production (i.e., area of land in cultivation). Long before agriculture, hunter-gatherer groups likely were aware of the sustainable number of members that regions could support, although it may be argued that high mortality rates inhibited these early populations from pushing the limits of sustainability. During difficult times populations reached or exceeded what we might think of as a carrying capacity, possibly imposing nomadic lifestyles in which groups had to intermittently move after local resources were depleted. Speculating on past population dynamics limits at the trouble with the concept of carrying capacity: It must represent a dynamic value that changes over time and is highly dependent on many interacting factors, such as environmental variability and, for early humans, what hungry animals awaited their forays. This makes the concept difficult to use.


“A struggle for existence inevitably follows from the high rate at which all organic beings tend to increase. Every being, which during its natural lifetime produces several eggs or seeds, must suffer destruction during some period of its life, and during some season or occasional year, otherwise, on the principle of geometrical increase, its numbers would quickly become so inordinately great that no country could support the product. Hence, as more individuals are produced than can possibly survive, there must in every case be a struggle for existence, either one individual with another of the same species, or with the individuals of distinct species, or with the physical conditions of life. It is the doctrine of Malthus applied with manifold force to the whole animal and vegetable kingdoms; for in this case there can be no artificial increase of food, and no prudential restraint from marriage. Although some species may be now increasing, more or less rapidly, in numbers, all cannot do so, for the world would not hold them.”
Malthus’ clear recognition of the importance of limitations of growth in populations helped Darwin to lay the foundation for his theory of natural selection, which is built on the premise that populations are regulated primarily by competition which leads to differential reproduction.

The earliest concise description of carrying capacity derives from Pierre Francois Verhulst, a Belgian who lived in the mid-nineteenth century. Verhulst, perplexed by accounts that the human population appeared to be increasing exponentially, derived a mathematical formula which he called the “logistic” equation that would account for a slowing in the population growth rate as a function of population size. The same relationship was rediscovered in 1920 by Raymond Pearl and Lowell Reed, who used the logistic equation to predict the population of the United States based on census data collected from 1790 to 1910 (apparently neglecting the fact that the area of the United States increased more than three-fold during this time). The resulting application of the logistic equation to U.S. census data led Pearl and Reed to greatly underestimate the U.S. population, predicting it would level off at about 197 million in the Year 2050 (Fig. 1).

The failure of Pearl and Reed to accurately predict the population of the United States, currently at about 275 million, reveals at least one important aspect of mathematical models. It indicates that, for the U.S. human population, some aspect of the logistic equation must be wrong. Two candidate problems include violations of the assumptions that the area of the United States remained constant and that the U.S. population lacked immigration. These differences between data and model predictions can help us to better understand the problem at hand. In this example, the violations of model assumptions leads us to re-evaluate the factors that influence population growth and ultimately regulate it. The prediction from the model also clearly did not hold for a variety of socioeconomic and human health reasons and may have been wrong for statistical reasons. Using the logistic equation (based on the filled data points in Fig. 1) to predict population size well into the future, without providing confidence limits, is clearly tenuous at best. These problems have led us to be more careful in our predictions of how populations change over time and whether a carrying capacity can be predicted from such data or even exists.

III. DEFINITIONS OF CARRYING CAPACITY

Carrying capacity is the maximum population that a given area can sustain. The measures commonly used include the number of individuals or the total biomass of a population, which are each highly dependent on differences in physiology and age structure among species and across large taxonomic groups. The use of the term carrying capacity has changed over time, but most models suggest that population growth is rapid when density is low and decreases as populations increase toward some maximum. In addition, any definition of this concept improves as we narrow the time and area for the population that we are studying. Population descriptions, therefore, are often depicted as densities, accounting for the number of individuals per unit area. Population density usually varies over time and from place to place. In practice, we generally use population size or density to describe carrying capacity, which is determined either by resource availability or by the influence of enemies (predators and/or pathogens).

Various definitions of carrying capacity arose in the twentieth century, ranging from the suggestion that carrying capacity is that level below which predators have no effect on a population to the population size which can be maximally supported in a given region (previously referred to as the “saturation level”). There also has been a distinction made between “ecological
CARRYING CAPACITY, CONCEPT OF

Carrying capacity, which refers to the limitation of a population due to resources, and a management-oriented, maximum sustainable yield for a population, referred to as an "economic carrying capacity," which is usually lower than ecological carrying capacity. These definitions clearly lead to difficulty for wildlife managers who have been preoccupied with attempting to determine whether populations are either too high or too low. These debates continue, as exemplified by range management decisions in Yellowstone National Park and issues regarding the increasing frequency of reintroduction programs of top predators.

Carrying capacity may best be expressed mathematically. One of the simplest forms of population change over time can be represented as the differential equation

\[
dN/dt = rN,
\]

where \(dN/dt\) represents the instantaneous change in a population over a short time period, \(r\) is the intrinsic growth rate of the population, and \(N\) is the size of the population. This yields what is often referred to as a "J" curve, or exponential growth (Fig. 2). In discrete time this relationship is referred to as geometric growth.

In 1838, Verhulst modified the exponential growth equation and derived the logistic equation that depicted population growth rate as being inversely related to population size. To slow population growth he added an additional term yielding

\[
dN/dt = rN(1 - N/K),
\]

where \(K\) is the population carrying capacity. The term \(1 - N/K\) slows growth rate linearly toward zero as the population \((N)\) approaches the carrying capacity \((K)\). This results in a sigmoidal S-shaped curve for an increasing population over time (Fig. 2). If the population exceeds \(K\) \((N > K)\), then \(1 - N/K\) is negative, causing growth rate \(dN/dt\) to be negative and the population to decline monotonically toward \(K\).

An important attribute to bear in mind is that the logistic equation is deterministic, meaning that if we use the equation to predict population size at the end of a fixed amount of time we will derive the same population each time we start the population over. This assumption is usually violated in field conditions in which random effects, such as accidental deaths, failure to find mates, or fluctuations in environmental conditions, are common. Therefore, it has been argued that we should not expect real populations to behave according to the logistic equation. This simple equation has been challenged repeatedly by critics without apparent damage. This resilience of a theory is rather rare in science, which is a discipline that prides itself on being able to quickly dispel hypotheses (or equations) given even a small amount of contradictory data. However, the intuitive nature of the idea that populations are regulated by factors such as food supply helps the logistic equation to remain a staple in ecological texts and classrooms. The reason this equation and carrying capacity \((K)\) endure is that the equation’s shortcomings help us better understand the dynamics of real populations, ensuring its utility for many years to come.

This simple equation has been challenged repeatedly by critics without apparent damage. This resilience of a theory is rather rare in science, which is a discipline that prides itself on being able to quickly dispel hypotheses (or equations) given even a small amount of contradictory data. However, the intuitive nature of the idea that populations are regulated by factors such as food supply helps the logistic equation to remain a staple in ecological texts and classrooms. The reason this equation and carrying capacity \((K)\) endure is that the equation’s shortcomings help us better understand the dynamics of real populations, ensuring its utility for many years to come.

This simple equation has been challenged repeatedly by critics without apparent damage. This resilience of a theory is rather rare in science, which is a discipline that prides itself on being able to quickly dispel hypotheses (or equations) given even a small amount of contradictory data. However, the intuitive nature of the idea that populations are regulated by factors such as food supply helps the logistic equation to remain a staple in ecological texts and classrooms. The reason this equation and carrying capacity \((K)\) endure is that the equation’s shortcomings help us better understand the dynamics of real populations, ensuring its utility for many years to come.

The discrete, or difference, form of the logistic equation yields a different prediction of population behavior compared to the previously described continuous version. In particular, the discrete form was the equation used by Sir Robert May to first describe how a simple, deterministic equation could produce chaotic population dynamics, a pattern that emerges when intrinsic growth is relatively high. This chaotic behavior appears to mimic realistic changes in populations over time. Several long-term data records conform to chaotic dynamics, including the change in the number of lynx captured over time in Canada (Fig. 3).

IV. DO POPULATIONS HAVE CARRYING CAPACITIES?

This question has been addressed using a variety of techniques, including observational types of studies that rely on long-term time series data sets such as the number of lynx captured over time (Fig. 3), highly controlled laboratory experiments (Fig. 4), and mathematical models to determine potential mechanisms through time series reconstruction. The short answer...
is that there likely are carrying capacities for most species but that determining these at any one point in time and space is quite difficult. Many controlled laboratory experiments have been published that show populations behaving in a fashion consistent with the logistic model (i.e., populations reach a carrying capacity). One of the earliest studies was completed by Gause in 1934 (Fig. 4). Most laboratory populations tend to increase and then reach some level at which they fluctuate around what might represent a carrying capacity. It is interesting to note that populations rarely exhibit a smooth transition between a growth phase and gradual reductions in growth rate ending in stable populations, despite the controlled environmental conditions. An additional caveat to consider is that prior studies that concluded population growth patterns differed significantly from the expected logistic growth likely suffered disproportionately during the review process and failed to make it into print.

Determining how natural populations change over time is surprisingly difficult. The first requisite information necessary to determine whether a population is at or near a constant size, considered here to be a proxy for the habitat’s carrying capacity, is simply the population’s size over time. This often has to be determined over long periods of time in ways that are accurate, reliable, and repeatable. In field studies, it is rare to have the luxury to repeatedly estimate population size, a technique allowing us to assess the accuracy of our estimates. Determining that a population fluctuates may represent real changes in populations or represent either natural variability (statistical “errors”) or actual errors in our estimates.

Assuming we overlook these shortcomings in our data, what do populations do? In general, populations usually fluctuate over time. We may be able to correlate these changes with biotic or abiotic factors or some function of the two with time. Sometimes the fluctuations cannot be distinguished from random noise. Some populations, including the classic examples of lynx, hares, and lemmings, cycle periodically (Fig. 3). The persistence of population cycles over long periods of time has led to great speculation regarding the factors that might lead to periodicity. Recent work suggests that simple causal mechanisms of cycling are unlikely and that a combination of random environmental factors and nonlinear, density-dependent factors influence populations.

V. DETERMINING THE CARRYING CAPACITY OF A POPULATION

Determining the carrying capacity of any particular population at a particular time is not trivial. Many different techniques have been suggested and tested, including three primary techniques that can be used to attempt to detect a change in population growth rate as a function of population size: mathematical modeling of specific mechanisms, tactical experimental tests in the laboratory and field, and statistical analysis of time series data. Ultimately, a combination of these techniques will enable us to understand the importance
of regulation in populations and the degree to which populations appear to be governed by a carrying capacity. I have already discussed the importance of the logistic equation and will briefly introduce the empirical approaches.

There has been much interest in and work completed to determine what factors regulate the change in growth of populations over time. The factors that slow population growth rate change over time and from location to location and differ for different species. Regulating factors also are likely to interact with each other, thus complicating the determination of a population's carrying capacity.

In a classic study, Davidson and Andrewartha in 1948 used a partial regression technique to analyze an experiment designed to test the relative influence of biotic and abiotic factors on regulating a small herbivorous insect population. They concluded that 78% of the population variance was due to abiotic or weather-related factors. In particular, the number of individual thrips in the spring was related mostly to the preceding autumn climate. This study was influential because it provided strong evidence that this population of thrips was regulated not by biotic factors such as competition or predation but rather by abiotic factors.

A second method used to detect the presence of density dependence on population regulation is the analysis of time series data. The best data are those that have been collected over consecutive years and that exceed the periodicity of both observed environmental and population fluctuations (generally >10 years). These data can be subjected to tests that investigate the relationship of change from year to year as a function of the population during the previous year or years in order to detect whether the population appears to be regulated. Such analyses, however, are unable to provide any information on the underlying mechanisms that might lead to population regulation. Therefore, time series analysis is an excellent exploratory tool that can be used to investigate the possibility that a population is regulated. This information can then suggest experiments designed to partition variance among potential candidate mechanisms.

Determination of a population's carrying capacity is best done through a combination of modeling, experimentation, and time series analysis. Research efforts, however, need to be directed toward investigating the underlying mechanisms that govern population regulation. Without an understanding of the relative importance of these regulating factors, it will be difficult to determine whether populations are regulated and whether we can detect a population's carrying capacity.

VI. CURRENT RESEARCH ON CARRYING CAPACITY

Two main areas of research continue to drive our quest to understand population regulation and the strength and importance of carrying capacity. The persistence of these questions indicates the need to clarify the mechanisms that influence population change over time.

A. Determining the Relative Strengths of Factors That Regulate Populations

Although some researchers have argued that populations are unregulated, most agree that negative feedback mechanisms operate on populations, resulting in decreased growth at high densities. This may occur through changes in the abundance of food, through increased predation or disease, or through a combination of these biotic factors and abiotic factors such as local climate. These factors may reduce birth rates or increase death rates, or both. Although there are circumstances in which these rates change at low population densities (e.g., the Allee effect, which states that very small populations are likely to decrease due to such factors as difficulty in finding mates or pollen limitation), their regulation at high densities is likely to be common. This change in birth and death rates as a function of density is referred to as "density dependence."

A population that is regulated has intrinsic, extrinsic, or a combination of these factors that slows population growth. Under such conditions a population's per capita growth rate decreases with increasing population size through reduced birth rates and/or increased death rates. This relationship, in logistic growth, is assumed to be linear. The existence of a carrying capacity, however, is not dependent on the shape of this function, so the violation of this linearity assumption does not weaken the concept of carrying capacity. A better understanding of this relationship, generally determined through carefully designed experiments, will help us understand the importance of regulation on population dynamics.
Krebs et al. (1995) suggested that hare and lynx cyclic population dynamics are likely influenced by different sets of factors, including food availability and predation driving the dynamics of hare populations and the lynx population is driven primarily by changes in the number of hares. In a more highly controlled experiment using three trophic levels, Hartvigsen et al. (1995) determined that plant performance was controlled by the interaction of top-down and bottom-up factors, including the level of plant resource availability and the presence or absence of herbivores and herbivore predators. These studies suggest that complex, interacting biotic and abiotic factors likely influence population dynamics.

B. Determining Population Carrying Capacity

The logistic growth equation attempts to model regulated population change over time and relies on several important assumptions, including the absence of time lags (population dynamics is independent of prior events), migration or immigration, genetic variability or selection, population age structure, and the fact that density dependence is linear (each individual added to the population has a similar effect on the population's per capita growth rate). Violations of these assumptions have been found in various populations and have led to more refined, realistic, and complicated forms of the logistic equation. In addition, the model assumes that carrying capacity ($K$) is constant over time and space. This assumption occasionally may be valid in situations in which a population is regulated by habitat availability. This might occur, for example, where the number or area of nesting sites is fixed. It is easy to conjure up situations, however, when this assumption would be violated over very small spatial or temporal scales. It is not likely that $K$ would be constant since populations are usually limited by resources, competitors, enemies, and often combinations of these factors that vary with the environment over time. Under these conditions changes in resource availability can influence population size directly or indirectly through its often non-linear effect on the population of competitors and/or predators.

In addition, there is great concern about the stability and persistence of threatened and endangered species (see Section VII). Work in this area has begun to recognize the importance of species interactions, immigration and emigration among subpopulations, the introduction of exotic species, and other factors that violate the assumptions of simple logistic growth. The movement of individuals among subpopulations enables the possibility of increased long-term persistence of populations by reducing large-scale fluctuations and the spreading of risk that a species will become extinct in the event that a single local population disappears (becomes extirpated). This area of research, referred to as “metapopulation biology,” involves determining long-term viability of these subdivided species, and there is currently much research activity in this area.

VII. THE IMPORTANCE OF THE CONCEPT OF CARRYING CAPACITY TO BIODIVERSITY

The concept of carrying capacity suggests that species are likely to have some upper limit to their population. If the upper limit is “hard,” then we expect populations to achieve this state and remain relatively constant. Populations, however, as demonstrated in Figs. 3 and 4, do not behave in such a simple fashion and have rather “soft” limits. As such, populations usually exhibit random, cyclic, or chaotic dynamics. These dynamics generally lead to increased chances that populations will reach the absorbing state of zero (become either locally extirpated or globally extinct).

We must be concerned about the dynamics of relatively small populations over time. The probability that a population will go extinct is generally related to the degree to which it fluctuates (population amplitude and frequency). Therefore, processes that cause populations to increase fluctuations are likely to lead to species loss and associated reductions in biodiversity. Thus, conservation efforts may be needed that will buffer populations and associated habitats from extreme fluctuations. Conservation efforts are often directed toward increasing a population's carrying capacity. It should be kept in mind, however, that constant environments also may lead to species losses. The intermediate disturbance hypothesis has gained much empirical support and suggests that the maximum number of species that an area can support occurs when disturbances are intermediate in either frequency or impact. We should be concerned that our management efforts do not reduce the carrying capacity of target species.
VIII. THE HUMAN CARRYING CAPACITY

The best estimates of human population indicate that it has continued to grow exponentially over recorded history, although the actual growth rate has changed over time. Attempts to fit data on the human population to the logistic equation have failed (Fig. 1), and current indications are that no human carrying capacity can be predicted from simple population statistics. However, we might ask whether our population growth rate is likely to slow down in the foreseeable future and, ultimately, reach a stable carrying capacity or whether it will overshoot its carrying capacity and eventually collapse. Joel Cohen (1995) found that estimates of the human carrying capacity have ranged between 1 billion and 1 trillion people, with the majority of estimates falling between 8 and 16 billion (the current population is about 6 billion). These estimates suggest that we are approaching an apparent limit for our species. Regardless of which estimate seems most appropriate as an upper limit for humans on Earth, the growth rate of our population will eventually slow to 0. This can occur as a result of increasing death rates and/or decreasing birth rates. I predict that as our population grows in the coming decades there will be an increase in mortality due to diseases. The effect of disease agents on controlling population growth will likely increase due to increases in human contact rates and rapid transmission rates, increasing evolution of drug resistance, and increasing virulence rates. These factors also may reduce birth rate, which of course presents a more pleasant alternative to slow population growth.

Can we avoid a population crash? I venture the guess that we cannot. Any long-term stabilization of the human population will require a decrease in the current global birth rate. We certainly cannot hope to achieve a relatively stable population without invoking a substantially higher death rate than the current rate, which is not a comforting thought. It is difficult to imagine, however, that the influence of disease will operate in a simple density-dependent fashion. Instead, it seems more plausible that diseases will "break out" more often with increasing population size and with larger scale consequences, bringing about a strong reduction in our population—a response seen in many other populations that have increased beyond their carrying capacities. One last area of hope is that individuals will lower their consumption rates, thereby adjusting the human carrying capacity. It is unlikely that Earth can support tens of billions of people with lifestyles matching those of people in the developed nations such as the United States. Therefore, there remains a chance that changes in human behavior will allow our population to gently transition toward a sustainable, zero growth rate population.

IX. CONCLUSIONS

The concept of carrying capacity has a history that spans at least thousands of years. The formal definition is about 150 years old and is generally coupled to the asymptotic population in the logistic growth equation ($K$). Critics argue that because of ongoing confusion and the multitude of definitions attached to the concept we would be better off to simply abandon the term. We also must be concerned that the term not be used to advance any particular political agenda associated with determining how large populations of any particular species "should" be in particular areas. This entry cautiously suggests that the concept remains useful. Since most populations are likely to be at least occasionally limited by factors that depend on the population's density, we need to continue advancing our knowledge of how populations behave and use this information to guide the design of laboratory and field experiments aimed at determining the mechanisms that regulate populations. Only by using the combination of field and laboratory techniques, grounded in a theoretical framework that has roots going back to the simple logistic equation, will we hope to understand and conserve populations, including our own, over long periods of time.

See Also the Following Articles

- POPULATION DENSITY
- POPULATION DYNAMICS
- POPULATION STABILIZATION (HUMAN)
- POPULATION VIABILITY ANALYSIS (PVA)
- SUSTAINABILITY, CONCEPT AND PRACTICE OF

Bibliography

CARRYING CAPACITY, CONCEPT OF


CATTLE, SHEEP, AND GOATS, ECOLOGICAL ROLE OF
Andreas Troumbis
University of the Aegean

I. Domestication of Ruminants: The Process That Changed the Face of the Earth
II. Ecosystem-Level and Global Effects of Large Domestic Animals
III. Biotic Relationships of Large Domestic Animals

GLOSSARY

grazing intensity  Frequency and closeness of grazing.
grazing pressure  Stocking rate, the units of grazing animals per land area.
overcompensatory growth  In grazed plants, the reaction to tissue removal by enhanced primary production compared to in undefoliated controls.

THE INTERACTION OF HUMANS WITH DOMESTICATED RUMINANTS has strongly influenced the cultural evolution of societies. Livestock breeding has both negative and positive effects on landscape and ecosystem processes and biodiversity. Ecological disturbances generated by domestic herbivores are defined at various spatial and temporal scales, from the micro-scale of the defoliation of an individual plant, to nitrogen redistribution at the pasture level, to global-scale phenomena such as desertification and toxification of the planet.

This article focuses mainly on the interactions of cattle, sheep, and goats with other biotic components of grazed ecosystems.

I. DOMESTICATION OF RUMINANTS: THE PROCESS THAT CHANGED THE FACE OF THE EARTH

The natural capacity of ruminants to digest hard plant tissue and to transform it into animal biomass is a service that has been provided by biodiversity to human societies for millennia. The cultural evolution of humans is strongly related to the exploitation of this service, which has influenced—and in turn has been influenced by—the pace, the pattern, and the distribution of societies. The first traces of domestication have been found in the Iranian mountains, where there are paleolithic settlements of sheep and goat herders dating from ca. 10,000 B.P., suggesting that domestic ruminant keeping precedes proper agriculture. Evidence of organized livestock breeding culminates between 8000 and 6000 B.P., at several sites in the Middle East and Anatolia. Ali Kosh, located at the head of the Persian Gulf in the ancient kingdom of Susiana (7000 to 6500 B.P.), is such a site where villages of farmers-breeders are found. Animals—and plants—were originally domesticated in a specific region, and then diffused during a 10,000 year process to many other parts of the world.

In the past, nomadic tribes and their herds of domesticated animals roamed freely to avoid drought, in search of good grazing land. In the rangelands of North America, Native Americans followed the wandering herds of buffalo, and it was not until domesticated cattle and sheep were introduced by European colonists that
the vegetation and soils of the plains and mountains were put under stress. The rise of national boundaries has created a barrier to many such movements. Seasonal transhumance is still a practice in Mediterranean areas, where especially flocks of sheep are moved between winter quarters in the mild rainy lowlands and summer quarters in rich alpine pastures.

On a global scale, diffusion was drastically accelerated by European expansion from the fifteenth century onward. At present, rangelands cover a large part of the world’s land surface (40%). Eighty percent of these rangelands are in arid and semiarid areas, where most of the 30 million humans who are directly dependent on livestock breeding for their subsistence live. In Africa, livestock breeding—mostly cattle and goats—is in the main source of income south of the Sahara, especially in countries like Sudan, Chad, and Somalia. This is also the case in much of Asia, in particular in Mongolia, the central Asian republics of the former USSR, and Tibet. Cattle and sheep are also economically important domestic animals in New Zealand, Australia, Europe, and parts of the United States.

The domestication of animal species, especially of ruminants, and the associated land management and breeding practices are historically important ecosystem disturbance agents through the effects of grazing, trampling, and digging. They have transformed, shaped, and selected landscape and vegetation types, and have had significant, often dramatic, effect on many ecosystem properties, including plant and animal diversity. Uncontrolled grazing on shrublands and the creation of pastures have gone hand in hand with the creation of agricultural land. Native plant and animal communities have been greatly modified by the introduction of domestic animals, either directly or through habitat alteration and landscape change. There are many well-documented examples from all around the world, but those that occurred on islands are the most informative for our purposes.

The direct effects of the introduction of domestic grazers on native faunas since prehistoric times are well described for the Mediterranean islands, where original faunas have been affected by species extinction and introductions promoted by humans. The modern mammalian faunas on the big islands of the Mediterranean (e.g., Corsica, Sicily, Crete, and Cyprus) replaced the entirely endemic Holocene faunas that existed before human arrival. After the spontaneous extinctions of the late Pleistocene, the surviving endemic herbivores (e.g., Episoriculus spp, Prolagus sardus, Rhagamys orthodon, Meridiopitymys henseli) were rapidly eliminated, even before A.D. 7000, either intentionally by hunting or indirectly through the effects of commensal species. The faunas of today are the result of a severe human-induced selection in favor of species belonging to the geographic and cultural universe of the human groups that immigrated to the islands, within which domestic livestock predominates.

In southern Greenland, the arrival of the Norsemen in ca. A.D. 1000, and the introduction of cattle have disturbed the fragile dwarf-shrub heaths, which had established slowly after the deglaciation period. The activities of Norsemen, but especially of their cattle, destroyed the equilibrium between climate and vegetation, causing severe erosion of the unstable soil. When the Norsemen disappeared in the fifteenth century, a new phase of soil stabilization began, but this is now being severely disrupted with the immigration of more humans, now as sheep-breeders, in the beginning of the twentieth century.

Madagascar’s highland region was once covered with evergreen forests dominated by about 20 endemic tree species. Beginning around A.D. 650, Indonesian settlers started to remove forests to create swidden fields. At about A.D. 1000, zebu cattle were introduced from Africa, further increasing the need to expand grassland at the expense of forests. These forests were being permanently replaced by a floristically impoverished steppe vegetation on ferrolic soils. By A.D. 1600, the highland forest had mostly disappeared: tree and humus removal had led to massive erosion, floods, water shortages, and faunal extinctions or endangerment.

Although much poorer in species richness, the faunas of the French subantarctic islands (Amsterdam, Saint Paul, Kerguelen, and Crozet Islands) show an analogous history following their discovery in the sixteenth and seventeenth centuries. Sheep, mouflon, and cattle are among the nine introduced species that thrive owing to lack of competitors and predators, and despite the small number of founding individuals. Herbivores have induced particularly significant changes in the nature and the structure of plant communities, leading to the extinction of endemic plants (e.g., Phylica nitida) or the degradation of the fragile peat-bogs that constitute the nesting sites of the rare Amsterdam albatross (Diomedea amsterdamensis).

In North America’s Great Basin Desert, changes in plant communities that occurred after the introduction of domestic livestock in the late 1800s resulted in unusual, unforeseen cascade effects on the interactions between native mammalian species. The establishment and dissemination of cheatgrass (Bromus tectorum) in the Great Basin has played a central role in this process. Evidence suggests that cheatgrass was introduced acci-
dentially as a grain contaminant at the end of the nineteen
teenth century, at the same time that large-scale domes
tic animal grazing began. Imported from Mediterranean
Europe and central and southwestern Asia, seeds of
cheatgrass exploited an ecological niche, as no native
annual was dominant in the Great Basin. Cattle, sheep,
and feral horses facilitated its establishment, for they
spread the seeds in the same areas that they disturbed.
Once established, cheatgrass promoted the likelihood
of fire to the detriment of the native species. In addition,
other factors, such as the effects of the lack of vesicular-
arbuscular mycorrhizae and selective lagomorph graz-
ing, have worked in concert to further establish cheat-
grass dominance. The ecological consequences of this
establishment have been an increase in fire frequency
and intensity, a decrease in species diversity, and a
landscape that is susceptible to severe erosion.

Ecosystem change resulted in the subsequent irrup-
tion of mule deer (*Odocoileus hemionus*) and the expan-
sion of mountain lions (*Felis concolor*). Domestic sheep
depredation is currently increasing in western North
America and is related to the expansion of suitable
mountain lion habitat, and consequently of the lion's
distribution and abundance. Furthermore, the expand-
sion of this predator's population has caused the severe
reduction of the porcupine (*Erethizon dorsatum*), an-
other North American native species.

Although environmental determinism is not the
dominant explanation of agricultural geography, the
limits of the distributions of domestic animals, espe-
cially of their races, are often attributed from a physio-
logical point of view to the physical environment, tem-
perature, and water availability on a regional scale.
However, it is also quite clear that the driving forces
behind some of these ecological changes stem largely
from cultural and social factors, including the size and
growth of human populations, which determine the
scale and the practice of stock-keeping activities.

II. ECOSYSTEM-LEVEL AND
GLOBAL EFFECTS OF
LARGE DOMESTIC ANIMALS

The ecological role of cattle, sheep, and goats, that
is, the effects they have on ecosystem processes and
biodiversity, is far more complex than the fact that they
remove biomass by eating grass and often browse, or
feed on, twigs, shoots, and leaves of other plant species.
The floristic composition and productivity of the grazed
ecosystem and the persistence of vegetation and of plant
and animal species, as well as physical and biogeochem-
ic processes, may suffer substantial changes because
of grazing animal-related factors, such as the severity
and frequency of grazing, species of animal, method of
prehension, treading, excreta deposition, and even the
saliva deposited on plants. In some environments, the
effects of grazing are quite predictable, whereas in oth-
ers such effects are much more difficult to predict.
Under certain conditions, grazing can act as a com-
pletely density- and species-independent disturbance,
and under other conditions the effects on plant can be
very selective and effects on soil properties and chemis-
try very heterogeneous in space.

A. Impacts on Vegetation

The impact of domestic ruminants on vegetation is
typically studied through comparisons between grazed
and ungrazed ecosystems. The dependent variables in
these studies are descriptors of vegetation structure,
commonly diversity and productivity. The independent
variables are ecosystem or environmental parameters
as well as grazing variables. Grazing variables are regu-
lated by humans. Depending on different grazing re-
gimes, the range of these variables extends from the
unregulated (by humans) grazing of native species, to
uncontrolled- or free-grazing of domestic ruminants,
to over-grazing situations, and to optimized stocking
rates for profitable animal mass gain per unit area of
grazed land.

A meta-analysis of data on the effects of grazing on
vegetation and soils, from more than 230 studied sites
worldwide, has shown that the typical symptoms of the
disturbances caused by domestic ruminants are changes
in species composition, changes in dominant species,
life-forms, and growth forms, changes in aboveground
net primary production, and, finally, effects on the rela-
tionship between species, aboveground net primary
production, root mass, and soil nutrients. In general,
herbivory prohibits the most productive species from
dominating and suppressing through competition the
less productive ones. The conceptual model of Milchua-
ras, Sala, and Lauenroth (1988) predicts the variation
in plant diversity in relation to grazing intensity along
gadients of moisture and evolutionary history of graz-
ing (Table I). According to this model, grazing should
have a greater effect on species composition in more
humid areas because adaptations of tall growth forms
capable of competing for light in a dense canopy are
opposite to those that provide resistance to or avoidance
of grazing. In contrast, plant adaptations to frequent
loss of organs from drought or herbivory are similar,
### Table I

The Plant Diversity–Grazing Intensity Relationship under Various Moisture and Grazing History Conditions, According to the Predictions of the Conceptual Model of Milchunas, Sala, and Lauenroth (1988)

<table>
<thead>
<tr>
<th>Moisture gradient</th>
<th>Moisture gradient</th>
<th>Moisture gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi arid</td>
<td>Subhumid</td>
<td></td>
</tr>
<tr>
<td>Evolutionary history of grazing</td>
<td>long monotonic (linearly decreasing)</td>
<td>unimodal, symmetric (parabolic)</td>
</tr>
</tbody>
</table>
| Field research offers evidence that herbivory by domestic ruminants may cause either an increase or a decrease in plant diversity. For example, in salt marsh habitats, heavy grazing eliminates sensitive species and produces a dense cover of graminoids in upper marsh coastal habitats. In other marshes, grazing produces bare patches that allow annuals and other low marsh species to invade upper marsh zonal communities. A retrogression in plant succession may occur in salt marshes and salt deserts because of heavy grazing. Intermediate levels of grazing by sheep, cattle, and horses could produce communities with the highest species richness and heterogeneity. At the other extreme of the aridity gradient, in a seasonally dry tropical savanna, species diversity was higher in a grazed area than in the neighboring area where grazers were permanently excluded. Furthermore, similarity in species composition between the grazed and ungrazed areas was very low. The introduction of domesticated animals to a pasture or rangeland invariably causes changes in species composition, leading to an increase in the abundance of those species that are less palatable to animals. Low-growing, prostrate growth forms are selected by grazing, and annuals and shrubs appear to be favored. Tall perennials decrease in abundance because this growth form, which offers competitive advantages in dense canopies free from grazing conditions, exposes the plants to selective herbivory pressure in grazed ecosystems. Between sites, aboveground primary productivity is negatively affected by grazing in the most productive ecosystems. Within-sites, lowlands are more affected than uplands. More productive lands are more likely to be preferred by grazing animals and breeders. Grazing has negative impacts on root systems. Intensive grazing removes the product of vegetative growth, which in turn reduces root growth. Under drought conditions, shallow roots cannot fully exploit subsoil moisture reserves, and the plant becomes stressed. B. Erosion, Desertification, and Land Degradation

Cattle are important agents of geomorphological change. The animals' impact on the landscape is to create bare soil by weakening the vegetation cover and then by breaking this cover down by trampling. On uplands, heavy grazing compacts the soil, reduces infiltration, increases runoff, and increases erosion and sediment yield. However, light and moderate grazing has effects that are much less significant. In riparian zones, grazing decreases erosional resistance by reducing vegetation and exposing more vulnerable substrate. Trampling directly erodes riverbanks, thus increasing turbulence and consequent erosion. Trampling also maintains and expands the area of bare soil upon which frost, rain, and wind act. Livestock keeping when practiced in a nonsustainable way is one of the major causes of desertification. Desertification is defined as land degradation in arid, semiarid, and dry subhumid areas resulting from temporary climatic crises, especially droughts that occur periodically, and harmful human activities in vulnerable ecosystems. Land degradation leads to a reduced capacity of dryland areas to produce useful outputs—
crops, fodder for grazing livestock, bush and tree cover—or to sustain wildlife. Degradation of dryland areas involves a range of processes: the erosion of soils through water and wind, falling levels of soil fertility and damage to soil structure, loss of vegetation cover and change in species composition, reduced availability and decline in the quality of water supplies, loss of wildlife, and a decline in the biological diversity of plant and animal life. Such processes reduce the productivity of crops and livestock systems in dryland areas, and increase the vulnerability to food crisis of populations depending on these resources.

The processes of desertification were first recognized in the Sahelian region of West Africa. Overstocking and overgrazing, firewood collection, and cultivation of unsuitable soils are responsible for 80% or more of the deserted lands in Africa. However, dryland degradation is now accepted as a worldwide problem. Over the last decade, surveys have been carried out on a global level to assess the extent of desertification and soil degradation by region, and they have produced significantly differing results. The GLASOD survey (Global Assessment of Soil Degradation), commissioned by the United Nations Food and Agriculture Organization (FAO), showed that 19.9% of drylands worldwide were suffering from desertification. In contrast, the United Nations Environment Programme (UNEP) usually quotes an estimate of 70% of the world’s dryland areas as suffering from some degree of desertification, with an estimated 900 million people worldwide at risk from problems of degradation. This survey includes not only areas affected by soil erosion, but also where a change in vegetation had occurred (e.g., where perennial grasses had been replaced by annuals). All regions of Africa have been affected by drought conditions and human pressures on land, as have parts of Mediterranean Europe, North America, Asia, and Latin America. In Spain alone, data from 1993 suggest that almost 1 million hectares (ha) of land are already considered as desert lands and another 7 million ha have been identified as being at high risk of desertification. In the United States, 90 million ha are considered to be affected by desertification.

The underlying factors that cause such adverse effects are many and various, and operate at different levels. At the local level, for example, inappropriate range, water, and livestock management practices may accelerate rates of erosion, especially on sloping land. Information continues to accumulate on the effects of digging deep boreholes in several regions of Africa. These boreholes discharge several liters of water per second, without the enforcement of any range management or land use policy, and have too often resulted in large concentrations of livestock (20,000 to 40,000 head) during the dry season. This destroys the range in the vicinity of the well over a radius of 20 km (125,000 ha) in one or two seasons (the stock rate being 10–15 times the carrying capacity). At the national level, government policies on land tenure and use determine whether production processes will be sustainable or not. During the second half of the twentieth century, the Autonomous Russian Republic of Kalmykia has undergone severe desertification. Under Soviet rule, rangelands were increasingly devoted to animal production, and pastures were converted to cropland in a campaign to increase crop productivity. Pastures were grazed at rates that were two or three times their sustainable production, saiga antelope (Saiga tatarica) populations and habitat greatly decreased, more than 17 million ha were subjected to wind erosion, 380,000 ha were transformed into moving sands, and 106,000 ha were ruined by secondary salinization and waterlogging. By the 1990s, almost 80% of the Republic had undergone desertification, and 13% had been transformed into a true desert.

At the international level, rising prices provide encouragement to breeders and farmers to produce more and, at the same time, raise revenues that can be invested to increase capacity. The African arid zone harbors about 55% of that continent’s 550 million head of livestock, and these livestock numbers have increased by 75% from 1950 to 1983, in spite of severe droughts that occurred in most arid zones in the early 1970s and 1980s. This is exponential growth of about 0.7% per year, compared to the 1.0–1.9% demographic growth rate of the African pastoralist population.

### C. Nutrient Cycling and Over-Enrichment in Grazed Ecosystems

Nutrient cycles in a grazed ecosystem do not differ from those in an ungrazed one in terms of individual element pools, flows, potential input and output pathways, and interrelationships among various nutrient pools (Fig. 1). On the contrary, the sizes of the pools, the rates of biogeochemical processes, the flows, and the residence time in the various pools of elements and nutrients are greatly affected by herbivores and the breeding practice and stock size. For example, in natural pastures, animals remove aboveground plant biomass (shoots, stems, and leaves) and deposit excrement; in managed pastures, where grazing intensity exceeds the natural carrying capacity, additional herbage and other fodder is provided to the animals and, thus, additional inputs of
CATTLE, SHEEP, AND GOATS, ECOLOGICAL ROLE OF 

FIGURE 1 Pools and fluxes of nitrogen (N) in a typical pasture. Nitrogen pools are shown in their relative average sizes: ORG-N, organic soil N (about two orders of magnitude larger than other pools); M-N, microbial biomass N; SF-N, soil faunal biomass N; INORG-N, inorganic (NH\textsubscript{4}, NO\textsubscript{2}, and NO\textsubscript{3}) soil N; PLANT-N, plant biomass N (ABG: shoots; BLG: roots); and HERB-N, large herbivore biomass N. The large rectangle represents field boundaries. For clarity, not all directions of flux are shown.

nutrients are entering the system. The cycle of nitrogen, most likely the most perplexing of all nutrients, is certainly disturbed under grazing conditions. In natural pastures, less litter returns to the soil, the topsoil organic matter declines, and the nutrient cycles are quantitatively disturbed. Grazers shorten the N cycle, increase the rate of N cycling, and cause significant redistribution of N, among pools and in space. On a daily basis, cattle excrete a somewhat constant amount of N in feces (8 lb/1000 lb of dry herbage consumed), the remainder being excreted in urine. Because they excrete more at night than during the day, bedding or watering places receive higher quantities of excreta-carried nutrients. Cattle and sheep retain a small fraction of the N consumed, and thus they have an impact on the control of N cycling that is out of proportion to their low biomass on an areal basis.

Areas of pasture that receive dung and urine may undergo marked changes in botanical composition. Urine spots and fecal pats contain the equivalent of 500 to 1000 lb N/acre, which is of different availability to plants: the N in dung is mainly in organic form, with a slow overall mineralization, whereas in urine N is present as urea, which is hydrolyzed rapidly to forms available to plants. The plant recovery of nutrients from excreta spots is rarely greater than 30%. Generally, the nitrogen in urine stimulates grass growth and the phosphorus in dung stimulates legume growth, especially on P-deficient soils. However, plants immediately beneath dung pats may be killed and urine occasionally scorches them.

The cycles of nitrogen, phosphorus, and other nutrients are not generally closed at the field or farm level. For nitrogen, six main pathways of loss have been described in grazed ecosystems: NH\textsubscript{3} volatilization, denitrification, wind and water erosion, NO\textsubscript{3} leaching, fire, and incorporation into animal biomass (exported or retained in situ). The relative importance of these pathways depends on environmental conditions and breeding practices. In humid and subhumid ecosystems, N losses are related to leaching, whereas volitilization is most important in semiarid regions, and the export of animal products and erosion in deserts.

Runoff from agricultural land and livestock feedlots is among the major "nonpoint" sources of nutrients entering marine and freshwater ecosystems, causing pollution and eutrophication. This has been shown on very large scales, as in the United States, where these widely dispersed activities are the major source of water pollution, as well as at the level of local ecosystems, such as lakes, rivers, and estuaries. In aquatic ecosystems, overenrichment with P and N causes a wide range of
problems, including toxic algal blooms, loss of oxygen, fish kills, loss of seagrass beds and other submerged aquatic vegetation, degradation of coral reefs, and loss of biodiversity—including species important to commercial and sport fisheries and shellfish industries.

Lake studies allow contemporary sediment and nutrient dynamics to be placed in a historical context so that trends and rates of change in catchment inputs may be calculated. Surface runoff from heavily grazed grassland has a high suspended sediment, ammonium nitrogen, and particulate phosphorus load. The combined effect of the long-term increase in the organic loading from livestock and the inorganic N and P load from fertilizers may be the source of nutrient enrichment in lakes. The literature offers several examples of eutrophication studies in lakes and streams; for example, in the already extinct karstic Jastericie Lake in Slovakia, in the coastal freshwater systems of Stapton Ley and Loe Pool in southwestern England, and in the streams draining the Mount Lofty ranges in South Australia, livestock breeding is the main source of organic N and P that affects the trophic status of surface waters.

Intensive animal production generally involves feeding large numbers of animals in small areas. For example, 4% of the cattle feedlots in the United States produce 84% of the cattle. Nutrients in manure can be recycled by applying the manure to cropland. However, the amount of manure generated by concentrated livestock operations often far exceeds the capacity of nearby croplands to use and retain the nutrients. Thus, excess fertilization and manure production cause a surplus of N and P, which accumulate in the soil. Some of this surplus is transported in soil runoff to aquatic ecosystems. Especially for N, the surplus is mobile in many soils, and much leaches into surface waters or percolates into groundwater. Indeed, the density of animals on land is directly related to nutrient outflow to aquatic ecosystems. Surplus N can also volatilize in nitrous oxide \((\text{NO}_2)\) and ammonia \((\text{NH}_3)\) forms to the atmosphere. The emission of ammonia from stored and land-applied manure can result in a significant loss of nitrogen for crop production. A high atmospheric concentration of ammonia can result in the acidification of land and water surfaces, cause plant damage, and reduce plant biodiversity in natural systems.

Nitrous oxide, together with carbon dioxide and methane, are considered to be major greenhouse gases. Besides emissions of these gases from the plant–soil compartment of the pasture soil system, \(\text{CO}_2\) and \(\text{CH}_4\), emissions from cattle and their excreta are significant components of the global fluxes. Methane emitted from dung is a very important contributor to the global methane budget; the corresponding figure is as high as 20%. Grazing animals on managed pastures and rangelands have also been identified recently as significant contributors to the global \(\text{N}_2\text{O}\) budget. This occurs because of the concentration of herbage N in urine and dung patches, and by the compaction of the soil due to treading and trampling. The limited amount of experimental data indicates that 0.1 to 0.7% of the N in dung and 0.1 to 3.8% of the N in urine is emitted to the atmosphere as \(\text{N}_2\text{O}\). Integral effects of grazing animals have been obtained by comparing grazed pastures with mown-only grassland. Grazing-derived emissions, expressed as a percentage of the amount of N excreted by grazing animals in dung and urine, range from 0.2 to 9.9% with an overall mean of 2%. Using the emission factor and data statistics from FAO for numbers of animals, the global contribution of grazing animals is estimated at 1.55 Tg \(\text{N}_2\text{O}-\text{N}\) per year.

III. BIOTIC RELATIONSHIPS OF LARGE DOMESTIC ANIMALS

A. Reactions of Plants to Grazing

The direct act of grazing represents a loss of organs or parts of organs to individual plants and an alteration of canopy structure to the community. Resistance of plants to grazing involves avoidance and/or tolerance mechanisms.

Plant tolerance reactions to the grazing event will depend on the capacity of individual plants to compensate for lost organs and the relative impact of the removal on competitive relationships in the canopy. Various species of animals graze differently because of their prehension organ anatomy and method, behavior, and diet preferences. For instance, cattle graze individual plant parts (leaves or stems) less selectively than do sheep and goats. The literature offers examples of strong interactions among intensity, severity, and other properties of grazing of different animal species and pasture (species) in terms of various components of plant fitness. Especially for grass species, the height, tiller number, survival, and reproduction are the fitness components that are most usually monitored in grazed plots/individuals and ungrazed controls.

Three alternative hypotheses on the effects of grazing intensity on plant growth and fitness have been proposed (Fig. 2). The first hypothesis predicts that net primary productivity shows a consistent decline as the intensity of grazing increases. The second hypothesis
predicts that plants compensate for tissue removal up to some level, beyond which productivity begins to decline. The third hypothesis, known as “overcompensatory growth,” states that within some levels of herbivorous feeding, plant productivity may be enhanced before declining beyond a threshold of grazing intensity. This hypothesis has generated an ongoing debate on the controversy of plant response to grazing, such as herbivore optimization and overcompensation.

Observations of mixed cattle and American elk grazing in high-elevation rangeland conditions in southwestern North America show that vegetation has deteriorated. Experiments in the North American tallgrass prairie suggest that overcompensation is a nonequilibrium plant response to grazing. Photosynthate that would be stored as reserves and used for root growth and flower and seed production is instead used to replace lost leaf area, thereby resulting in higher foliage productivity. However, under chronic grazing or mowing, vegetation is prevented from maintaining high nutrient and water uptake capacity (large root biomass) and accumulating reserves that allow overcompensation responses.

Herbivory by large animals is known to function as a selection pressure to increase herbivory resistance within plant populations by decreasing the frequency of genotypes possessing large erect canopies. Data on the trade-off between herbivory resistance and competitive ability in Schizachyrium scoparium confirm that herbivory by domestic cattle may function as a selection pressure to induce architectural variation in grass populations within an ecological time frame (ca. 25 years).

Comparative studies between species suggest that there is no consistent pattern of grazing effects on survival, reproduction, recruitment, and regeneration. In one study, five abundant native Kansas tallgrass prairie perennial forbs (Baptisia laciniata, Oenothera speciosa, Vernonia baldwinii, Solidago missouriensis, and Salvia azurea) were chosen to examine the effects of native (bison) and domestic (cattle) ungulates on plant growth and reproduction. The results show that their responses are complex and vary significantly among plant species, ungulate species, and plant life-history stages. Baptisia laciniata, O. speciosa, and V. baldwinii increased in growth and reproduction in grazed sites, indicating competitive release in response to selective grazing of the dominant warm-season matrix grasses. Species with reduced performance in grazed sites are likely to be negatively affected by disturbances generated by ungulate nongrazing activities, because none of the forbs studied was directly consumed by bison or cattle. Furthermore, the native and domestic ungulates differ significantly in their effects on forb growth and reproduction.

Among 18 species of shrubs and trees in southeastern Australia, 10 showed significant negative effects on recruitment and/or regeneration from present or past sheep grazing. In this case it was shown that the negative effect of sheep grazing on recruited seedlings must exceed that of natural thinning before overall regeneration is affected. Evidence that grazing by sheep and goat increases the total inclusive fitness in the Mediterranean shrub Anthyllis cytisoides has been reported; in this species, moderate grazing promotes growth, stability of vegetative structures, and adult survival, and a drop in seed production. Direct consumption of reproductive organs has significant effects on the overall fitness of the grazed species. In the annual wild wheat Triticum
dicoccoides, removal of maturing inflorescences by cattle reduces its fitness by 50%, estimated as the number of seeds produced per seedling. In this species, protection from grazing leads to a significant increase in the number of individuals that produce mature inflorescences as well as in the number of spikelets per inflorescence. In the clonal species Yucca elata, cattle browsing of inflorescences may reduce reproductive effort, but the most significant changes in population structure are due to the browsing of small caudices, including both genets and ramets. Thus, the expansion and dominance of “resistant” species on heavily grazed land may be a result not only of reduced competition, as is commonly assumed, but also of enhanced performance of some of these species.

Grazing avoidance mechanisms reduce the likelihood of defoliation by animals. For example, leaf surface chemistry of plants is related to the extent of defoliation by livestock. Data on cattle, sheep, and goat herbivory on tarbush (Flourensia cernua), an abundant but generally unpalatable Chihuahuan Desert shrub, support the hypothesis that secondary chemicals—mostly terpenes and phenolic compounds—in its resinous leaves may influence the diet selected by the ruminants. Although individual leaf surface compounds do not appear to greatly affect the degree of use of the plant by livestock, collectively these compounds may partially explain the differential herbivory on tarbush plants by livestock. Mimicry, both chemical and morphological, has also been suggested as a grazing avoidance mechanism for plants. Experimental tests with sheep showed that odor alone is not persistently effective in preventing herbivory, but that both taste and odor must be similar for one plant to successfully mimic another.

Spines and thorns are considered to be defense structures against herbivory by both wild and domestic animals. Experimental evidence from Acacia drepanolobium suggests that spine length is an inducible defense, with longer spines being produced by branches experiencing a greater level of herbivory. Examination of Acacia trees protected from herbivory for several years suggests that reduction in spine length eventually exceeds 70%. The effectiveness of spines and thorns as an antiherbivore defense in several plants in arid Australia did not vary with the evolutionary history of the herbivores (i.e., wild versus domestic). Furthermore, additional evidence shows that the interaction of ants of the genus Crematogaster and thorns of A. drepanolobium is a means of defense against browsing goats. This interaction causes the animal to stop feeding almost immediately, therefore keeping the amount of foliage lost to a minimum. It is hypothesized that the acacia–ant relationship evolved partly because of the pressure from browsing herbivores.

B. Seed Dispersal and Germination

Throughout the world, the presence of wild or domestic ruminant herbivores is correlated with the maintenance of high levels of plant diversity in natural and semi-natural grazed systems, where the grazing pressure does not exceed the carrying capacity of the vegetation. Zoochorous dispersal by herbivorous mammals has been verified repeatedly and its possible influence on the structure, function, and diversity of plant communities, mainly herbaceous, has been suggested. Cattle, sheep, and goats are important vectors for endo- and epi-zoochorous seed dispersal. On large scales, seed dispersal systems associated with domestic ruminants have been proved to be particularly favorable for the introduction of alien plant species, and herbivores have facilitated the naturalization and spread of many alien herbaceous species from their initial points of introduction. The case of central Iberian herbaceous species introduced into the Mediterranean-type zone of Chile is very relevant: almost 13% of the central Iberian herbaceous species are naturalized in Chile, with the endo- and epi-zoochorous species representing 21 and 23% of them, respectively. The expansion of the distribution of individual species is often correlated with the effects of livestock grazing on the viability and germination rates of seeds. For example, the calden (Prosopis caldenia Burkart) is a dominant tree of the xerophytic open forest in the semi-arid pampa of Argentina. Calden has gradually increased its distribution throughout the region during the past century as a result of livestock grazing in the pampa forest. Excreted calden seeds display a range of delayed germination responses. This variation would increase the probability of seed germination for a variety of environmental and site conditions.

The potential dispersal distances for adhesive seeds attached to the fur of cattle range from tens of meters to kilometers. The morphology of the seed's structure and its position on the animal's body influences the length of time that it remains attached to the fur.

Endozoochorous dispersal selects for traits that enhance ingestion and passage of viable seeds through the animal. Passage of buffalo grass (Buchloe dactyloides) seeds, one of the two dominant grasses of the North American shortgrass prairie, through cattle has a positive effect on germination and seedling growth from intact diaspores; the damage due to mastication is mini-
eral, and the retention time is from 1 to 5 days. This combination of retention time and movements of the animals influences the spatial expansion, distribution, and abundance of this species. Clear experimental evidence that the germinability of seeds increases significantly following their passage through the cattle gut has been provided for the legume *Bisorrula pellecinus*, which is greatly favored by its dispersal through cattle dung.

The amounts and diversity of viable seeds contained in the feces of domestic ruminants grazing in nature may be quite high. Based on germination trials, the number of seed species germinated in the dung of feral cattle feeding in savannas and floodplain wetlands in Rajasthan, India, was about 450 seeds per m², belonging to 35 species. Similar results have been obtained in Mediterranean open woodlands, where cow dung may yield as much as 70 seeds per gram of dry dung from a large number of species (about 75 species). Although ruminant species differ in their traits and feeding habits, dispersal seems to be mainly determined by seed production of the plant community.

Large isolated trees are a common feature of the agricultural landscape in humid tropical regions that were originally covered by rain forest. These isolated trees are used as a shade for cattle, but they also function as nursery plants for rain forest species if their seeds are distributed under their canopies by cattle, frugivorous birds, and bats. The same pattern is observed in lotswanan savannas: the analysis of seed pools under the canopy trees *Dichrostachys cinerea* showed that epizoochorous species such as *Tragus berteronius* were dominant at the cattle resting sites under trees.

Patch dynamics are also affected by seed zoocorous dispersal. In Mediterranean pastures, cattle-dispersed endozoocorous seeds germinate in manure and colonize the dung patches. The micro-succession involved is independent of the type of pasture. A small-scale spatial pattern results in which gaps of old dung are colonized by endozoochorous species. Thus, dung patches enhance the similarity between different communities when they are grazed, but also increase the variation within communities. In Australian sub-alpine woodland vegetation, the density of shrub seedlings in gaps varied considerably in space, but it is related to the dispersal of seeds and the trampling and browsing effects of domestic cattle.

Large seeds such as *Quercus* acorns are rapidly predated by large herbivores. However, in this case, experimental data seem to suggest that seedling emergence rate is inversely related to the intensity of predation on the acorn bank.

### C. Domestic Animals as a Food Source for Wild Predators

Domestic ruminants have largely replaced wild grazing animals over large areas on all five continents. Natural predators and scavengers of large mammals have easily adapted to this alternative food source. Population densities of cattle, sheep, and goats are equivalent and often higher than the densities of the wild grazers that have been replaced, as the animal keepers' tendency is to maximize herd or flock size and food may be supplemented at times of natural fodder shortage. Furthermore, domestication has usually produced heavier, slower-moving, less alert, and less experienced animals, which often have their ability to move impaired by various means and structures such as pens, fences, and tethering. Thus, domesticated animals are generally speaking easy prey for most natural predators. The way of life of pastoral societies around the world has been profoundly influenced by the constant strive for protection of their flocks and herds from natural predators.

Wolves (*Canis lupus*) are legendary in Western culture as predators of cattle, goats, and especially sheep. Today they have been exterminated or reduced to small populations over much of their former range in the Northern Hemisphere, but where they do exist they are still important predators of livestock. In regions like parts of southern Europe, where stock-raising is prevalent and wild prey are rare, their presence is still connected with the raising of stock animals. The wolves either prey directly on live individuals, most often stray ones, or feed on carcasses and offal. Similarly, dingos (*Canis familiaris*) can be serious predators of sheep in Australia, as are coyotes (*Canis latrans*) in North America. In Africa, Asia, and North and South America, large cats such as leopards (*Panthera pardus*), tigers (*Panthera tigris*), and pumas (*Puma concolor*) commonly prey on cattle, sheep, and goats. Big cats and the other canids differ from wolves in that their existence does not seem to rely on any significant extent on domestic animals. The big cats are especially dependent on extensive tracts of undisturbed habitat where natural prey is available, and consequently domestic animals are consumed opportunistically and usually only by certain individuals.

In many parts of Eurasia, both golden eagles (*Aquila chrysaetos*) and lammergeiers (*Gypaetus barbatus*, German for “lamb vulture”) have traditionally been regarded as significant predators of lambs, and this has
been the reason for their local extermination. In reality they kill very few lambs but commonly feed on dead lambs or the carcasses of older sheep. Both Old and New World vultures also include dead cattle, sheep, and goats in their diet. As in the case of the wolf, in parts of the world where wild ungulates are now rare, such as around the Mediterranean, these sources of food are essential for the survival of vulture populations. Although griffon vultures (Gyps fulvus) are not migratory, they perform movements between summer and winter quarters, closely following flocks of sheep as they move from high-altitude summer pastures to lowland grazing land in winter. For the larger eagles and vultures, open range grazing has a dual functionality in providing a food source (essential or supplemental) and by indirectly maintaining open habitats where such species are able to forage.

Cattle, sheep, and goats appear in the diet of the majority of other medium or large natural predators such as bears (Ursus spp.), foxes (Vulpes spp.), jackals (Canis spp.), other eagles (Aquila spp.), and buzzards (hawks, Buteo spp.) to a varying extent, depending on the region, the species of predator, and the method of stock-keeping. In most such cases, carcasses, offal, and stillborns are consumed rather than healthy, fully-grown adults.

D. Domestic Animals and Invertebrate Fauna

Both positive and negative effects of cattle, sheep, and goats on invertebrate communities of grazed ecosystems have been reported. In grazing systems where the carrying capacity of vegetation is not exceeded by grazing, as in the traditionally managed European mountain meadows or mixed woodlands, the maintenance of a landscape mosaics favors the persistence of high plant diversity and abundance of flowers and has positive consequences on insect population dynamics. Bumblebees, butterflies, syrphids, and other insect groups foraging on flowers depend on the landscape management in which domestic grazers play an important role. In central Spain, traditional landscape management for non-intensive grazing by goats and cattle favors the conservation status of the butterfly Euphydryas aurinia, which depends on patches of open oak woodlands mixed with open areas, where important nectar sources and larval foodplants are present. However, indirect effects on pollinators may appear from the removal of flowers by grazing; in Yucca elata, cattle preferentially consume inflorescences, which are found to be highly nutritious. This implies the possible local extinction of the yucca moth Tegeticula yuccasella, the exclusive pollinator of the plant.

In the shortgrass prairie of Colorado, in the United States, heavy grazing negatively affects the colony density and distribution of harvester ants (Pogonomyrmex occidentalis) on a broad scale, with important consequences to behavioral, community, and ecosystem processes. Browsing by goats has been proven to increase the damages from both tunneling (Diptera) and "blotch-making" (Lepidoptera) leaf miners on trees growing on the Aldabra Atoll.

Insects are particularly sensitive indicator taxa of land use (especially the Cicindelidae, Staphylinidae, and Carabidae). Domestic cattle influence the arthropod diversity through trampling, its intensity being more important than the type of trampling. In semi-arid tropical soils in Queensland, Australia, heavy grazing significantly affects Acari populations as well as the diversity and activity of termite species. The deterioration of soil hydraulic properties associated with cattle trampling at high stock rates is responsible for these negative effects on detritivorous termite activity in the topsoil.

A wide range of invertebrates, as well as bacteria and fungi, are involved in the breakdown and eventual decomposition of dung. Some are facultative generalist decomposers of organic matter, but there are also a large number of specialist dung consumers, such as dungflies (Scathophagidae) and dung beetles (Geotrupidae, Scarabaeidae). The array of dung-feeding invertebrates present in a given area is mainly dependent on the characteristics of the dung, that is, on the taxonomy of the dung producer and its diet rather than on a distinction between wild and domestic stock.

E. Domestic Animals and Birds and Small Mammals

Excluding predators and scavengers, the main influence of cattle, sheep, and goats on other animals occurs indirectly through the creation and maintenance of open pasture land, in particular in areas where the natural vegetation cover would be forest or scrub. Tall-grass meadows, short turf on stabilized sand dunes, alpine meadows, dry stony pastures, and all other forms of grazing land that exist on different soils and under different climatic regimes are all attractive to species that would otherwise be restricted to steppe, desert, or arctic habitats on a global scale or to the margins of wetlands, beaches, steep hill slopes, and woodland clearings on a more local scale. The creation of pastures has gone hand in hand with the creation of agricultural
land, which has also provided vast areas of open habitat. The expansion of both pasturelands and croplands began several thousand years ago, and the process of colonization and spread of open-country species into such newly created habitats is lost in prehistory and is largely a matter of conjecture.

Animal species closely connected with grazing land include mammals such as moles (Taipus spp.) and voles (Microtinae), birds such as larks (Alaudidae) and glow-worms (Charadriidae), and insects such as grasshoppers (Acrididae) and ants (Formicidae). Grazing is often employed as a management tool, for example, to create the very short turf required by wintering white-fronted geese (Anser albifrons) in Britain.

Other species may use cattle, sheep, or goats more directly in feeding. Two well-known cases are cattle egrets (Bubulcus ibis) and yellow wagtails (Motacilla flava). Cattle egrets commonly use cattle, as well as other domestic and wild ungulates, as a vantage point and vehicle, capturing large insects disturbed by the cattle as they move. Yellow wagtails may also do the same or they may feed on flies and other insects attracted to the animals themselves or their dung.

Negative impacts on other animal species may also occur. For example, when natural grazing land is used, domestic ruminants may compete with wild herbivores. Although wild herbivores may suffer, human interest is usually focused on the reverse impact of wild herbivores on domestic ones [e.g., rabbits (Oryctolagus cuniculus) and sheep in Australia].

Another negative impact on wildlife is the trampling of eggs and young of ground-nesting birds. This problem is more frequent in wetland habitats, where most bird species nest on the ground, and may be particularly damaging for colonially breeding species such as terns (Sternae) and avocets (Recurvirostra spp.). Although these birds generally select islets or other sites that are inaccessible to ground predators (e.g., foxes) and thus also inaccessible to livestock, problems may be caused when water levels change or when a herd or flock is intentionally transferred to such a site.

F. Domestic Ruminants as Hosts for Parasites and Pathogens

Domestic ruminants harbor significant numbers of internal and external parasites. Helminths, protozoa, and arthropods (especially ticks and mites) are the most common components of the parasitic communities, although many other species of occasional or temporary parasites, such as blood-sucking Diptera, also infect domesticated animals. Endoparasites, mainly gastrointestinal parasites, have received most attention in veterinary research because of their prevalence as disease-causing organisms that are economically important in rearing commercial livestock. Within a species, the taxonomic diversity of parasites varies between the various organs of the hosting organism. Thus, the digestive tract, the liver, the circulatory system, the respiratory tract, the skin and the subcutaneous tissue, the muscles and the tendons, the eyes, the central nervous system, and the serous cavities are affected by quite distinct species of parasites such as trematodes, cestodes, nematodes, protozoa, and arthropods. For example, in sheep and goats, the list of gastrointestinal parasites comprises 15 trematodes, 9 cestodes, 88 nematodes, and 23 protozoa; the parasitic fauna in the skin and the subcutaneous tissues comprises only 6 nematodes but 78 arthropods. Parasitic communities also differ between host animal species, often closely related races. Likely causes are differences in host suitability and feeding differences that affect the probability of transmission.

Besides the specificity of the host–parasite interactions, the diversity of the domestic ruminants' parasitic communities is strongly influenced by species–area and species–climate relationships. As an example, for nematodes it has been demonstrated that

(a) large areas of permanent pasture include a variety of microenvironments that are favorable to the development of the free-living stages of various species of nematodes and
(b) free-living stages are very susceptible to dryness and survive better in areas with heavy rainfall.

The widespread use of veterinary chemicals against endoparasites, pests, and other pathogens of domestic animals characterizes the high-input livestock breeding and production strategy. Avermectins, a relatively new class of broad-spectrum pesticides, are used widely to control livestock parasites. Following veterinary treatment, avermectins are eliminated in the livestock feces. The dung mesofauna potentially exposed to avermectin (or other active compound) residues includes insects, earthworms, springtails, mites, and nematodes. The effects range from acute toxicity in larvae and adults, through disruption of metamorphosis, to interference with reproduction. For example, at high drug concentration in the dung, larvae of the dipteran Cyclorrhapha are killed or paralyzed, while at lower levels their metamorphosis is inhibited. At very low concentrations of avermectins, well below levels occurring in feces after routine treatment, adult emergence is reduced and a
significant number of imagines show morphological abnormalities. Nematoceran Diptera are less sensitive than Cyclorrhapha, but larval and pupal development are affected at higher dose levels.

Dung mesofauna occupy a variety of different niches within the ecosystem and the faunal composition changes as the pats age. Some members of this fauna act in concert with soil microbial flora and assist in the breakdown of dung and consequent nutrient recycling on pasture-lands. A retardation in the rate of loss of biomass of dung pats from avermectin-treated cattle has been observed following the various forms of drug administration. Rare insects could be put at risk by the use of avermectins, especially those that breed exclusively in the dung of the herbivores on which avermectins are used.

The use of avermectins may also indirectly affect some species of vertebrates by depleting the quality and quantity of important food resources. The effects of any reduction in invertebrate food in livestock dung would be expected to be especially severe if it occurred at critical times for the vertebrates, such as during the breeding season or when newly independent young animals were foraging and fending for themselves. Insects that develop in livestock dung therefore have important, additional roles in the ecology of pasturelands other than aiding dung degradation processes. Livestock dung is an important feeding habitat for a number of vertebrate species. The potential for direct poisoning of vertebrates through the accumulation of avermectins following consumption of invertebrates containing residues would, on present knowledge, appear to be limited, but it should not be disregarded.

This issue has generated much controversy regarding the extent of the effects on pasture and rangeland biodiversity and ecosystem functioning. It has been suggested that ecotoxicological studies commonly disregard the veterinary use patterns of drugs and consequently they overestimate their negative effects at a populational level. Avermectins administration patterns in temperate regions indicate that peak periods of insect activity do coincide, the heterogeneous patterns of administration to livestock and the focus of treatment on young animals result in the deposition of feces that are predominantly free of avermectin residues. Results of large-scale, long-term studies indicate that, even under conditions of relatively high levels of avermectin use in cattle, the impact on non-target insect populations and their function is limited.

See Also the Following Articles

AGRICULTURE, TRADITIONAL • DESERTIFICATION • GRAZING, EFFECTS OF • GREENHOUSE EFFECT • RANGE ECOLOGY

Bibliography


I. Central America as a Biogeographic Region

II. Biodiversity of Central America: An Overview

III. Life Zones and Ecosystems

IV. Tropical Rain Forests

V. Seasonally Dry Tropical Forests

VI. Tropical Cloud Forests

VII. Temperate Forests

VIII. Other High-Elevation Ecosystems

IX. Central American Ecosystems in Light of Global Environmental Change

GLOSSARY

biogeography: The study of the distribution of plant and animal life in the earth's environment and of the biological and historical factors that produced this distribution.

geomorphology: The study of the surface configuration of the earth, especially the nature and evolution of current land forms, their relationships to underlying structures, and the history of geological activity as represented by such surface features.

land use change: The changes in the way land is used at a given location, for example when a forested land is converted to agricultural fields or when a forestry plantation is changed to other types of use.

paramo: This term refers to a cold, inhospitable and humid landscape located above the tree line.

plate tectonics: A modern geological theory of tectonic activity according to which the earth's crust is divided into a small number of large, rigid plates whose independent movements relative to one another cause deformation, volcanism, and seismic activity along their margins.

CENTRAL AMERICA is a relatively small area extending from the narrowest part of southern Mexico to the Panama–Colombia border. Despite its small area and restricted latitudinal extension, Central America contains a significant proportion of the biological diversity of the Earth. This is due to its complex geological history, current diversity of climates, and topographic heterogeneity. This is reflected in the diversity of ecosystems present in this region, which in turn contain a remarkable diversity of species of plants and animals of tropical and temperate affinity, as well as many endemic organisms. The ecosystems of Central America include tropical rain forests, seasonally dry tropical forests, tropical cloud forests, temperate forests, and other high-elevation ecosystems, particularly paramo and high-elevation grasslands. Such concentration of biodiversity in Central America is threatened by global environmental changes, particularly the current patterns of land-use leading to massive deforestation and habitat fragmentation. In addition, climatic changes are likely to affect the
distribution, functioning, and biodiversity of the eco-
systems of the region, particularly the temperate and
high-elevation ecosystems.

I. CENTRAL AMERICA AS A
BIOGEOGRAPHIC REGION

The definition of Central America varies considerably
among sources. The most popular notion of what con-
stitutes Central America is based on a geopolitical crite-
rion. According to this, Central America is that region
of the Western Hemisphere that constitutes the isthmus
lying between Mexico and South America, encom-
passing the countries of Guatemala, Belize, Honduras,
El Salvador, Costa Rica, and Panama. Some geographers
also include at least one (Jamaica) or most (Cuba, Haiti,
the Dominican Republic, and Barbados) of the Antillean
Islands as part of Central America, but the most ac-
cepted views only consider the continental part of the
region as Central America. However, even this widely
accepted geopolitical definition is not entirely satisfac-
tory for the purposes of this encyclopedia. A description
of the natural ecosystems of Central America requires
a more natural definition. Thus, this article will consider
a biogeographical and geomorphological approach, ac-
cording to which Central America is more naturally
defined as an elongated, tapering isthmus that begins
in the narrowest part of southern Mexico, at the isthmus
of Tehuantepec (Fig. 1). This narrow region divides
the area of volcanic rocks to the northwest from the
folded and considerably faulted structures of the more
conventional Central America to the east. The southern-
most limit of Central America is the valley of Rio Sucio
(the Astrato River) in Colombia, located just to the east
of the Panama–Colombia border, where the massive
Darien jungles begin.

An important biogeographic aspect derived from the
location of the area is that this narrow strip of land
currently connects the neotropics of South America
and the Nearctic zone of North America, two major
biogeographic realms of Earth. An important additional
implication is that Central America currently operates
as a corridor for terrestrial organisms from both realms
and as a barrier to marine organisms from the Caribbean
and the Pacific seas. However, such situation has not
always been the case. A variety of hypotheses have been
suggested regarding the possible configuration of the
area in the past. However, a consensus exists that from
the latter Mesozoic, until approximately 5 or 6 million
years ago, in the Pleocene, no continuous terrestrial
connection existed between South and North America
via Central America. Moreover, throughout the largest
part of the Cenozoic era (65 to approximately 5 million
years ago), the region comprised from Nicaragua to
the northernmost part of Colombia was probably an
archipelago similar to the Lesser Antilles of today, pro-
viding only, perhaps, an occasional pass for terrestrial
life forms. Most likely, the continuous connection that
Central America currently provides between the north
and the south has been in existence for only approxi-
mately the past 3 million years, and the mountainous
backbone of the region did not reach its present eleva-
tion until the latter part of the Cenozoic. This means
that the variety of habitats that currently characterize
the region is a rather recent phenomenon and that
the ecosystems and biodiversity of Central America in
general are a recent blend that has resulted from a
complex paleobiogeographic history. A review of the
paleobiogeographic history of the region is beyond the
scope of this article; however, some salient aspects that,
in addition to what I described previously, impinge on
the current biodiversity of Central America include the
following (Rich and Rich, 1983). Until approximately
3 million years ago the mammalian faunas of Central
American were composed predominantly of North
American elements and only after this time did a north–
south interchange begin with a significant incursion of
vertebrates from South America. It is possible that an
endemic small-mammal fauna might have developed at
about this time in Central America and southern North
America and dispersed toward the south. The greatest
affinity of Central American angiosperm (flowering
plants) floras with South America indicates that dis-
persal from the south was more pronounced for plants
than for animals, although Pleistocene cooling periods
promoted dispersal of typically montane plants toward
the south. In addition, at the height of the North Ameri-
can glaciation period in the Pleistocene, climatic condi-
tions favored the dispersal of grassland- and savanna-
adapted forms. Nevertheless, these ecosystems must
have become as fragmented as they are today by the
increasing precipitation of the latter part of this epoch.
Finally, it was not until these times, toward the end of
the Pleistocene, that the tropical rain forest ecosystem
became widespread in Central America. This has the
implications that this ecosystem, geologically speaking,
is very young in the region and that its expansion de-
creased the effectiveness of the terrestrial connection
between North and Central America. This has played
a determining role in the modern composition of the
biodiversity and ecosystems currently seen in Central
America.
II. BIODIVERSITY OF CENTRAL AMERICA: AN OVERVIEW

An important facet of the biodiversity of Central America is its diversity of natural ecosystems. Because of its complex geological history and intricate topography, the region of Central America includes a mosaic of ecological conditions, in which several of the major natural ecosystems of the earth are represented. Despite the fact the region has a restricted latitudinal range, of approximately 10° (from about 8° to 18° north), and the overall climate is of a marked tropical affinity, significant climatic variations occur due to the topographic complexity rather than to latitude. Accordingly, the ecosystems of Central America are related to three ma-
The country of Costa Rica harbors approximately 51,000 km², placing it near the equator. For example, within a latitudinal range of 8° N to 11° S, the region experiences a dramatic variety of ecosystems within a given area.

The plains and western mountain slopes of the Pacific coast receive approximately twice as much precipitation as those on the eastern mountain slopes of the Caribbean coast. As these winds flow up and over the high mountains of the region, they cause a marked increase in precipitation, which is then deposited on this side of the region. In contrast, the easterly winds of the Atlantic coast absorb much of the moisture, which is then deposited on this side of Central America as these winds flow up and over the high mountains of the region. As a result, the plains and eastern mountain slopes of the Caribbean coast receive approximately twice as much precipitation as the plains and western mountain slopes of the Pacific coast.

Such a mosaic of environmental conditions determines the existence, within a restricted geographical range, of a dramatic variety of ecosystems within a given latitudinal position. For example, within a latitudinal range of about 5° N to 11° S and an area of approximately 51,000 km², the country of Costa Rica harbors 12 distinct life zones (vegetation types), one more than the size of West Virginia. This sample of ecosystems ranges from the seasonally dry tropical forests in the Guanacaste province to the high-elevation grasslands (“páramos”) of the central part of the country and the wet tropical forests of the Atlantic coast. At a more local scale, a southwest–northeast transect of approximately 200 km in a straight line (i.e., not considering the ruggedness of the terrain) across the mountainous region of the state of Chiapas, in southern Mexico, harbors a collection of 12 distinct vegetation types: mangroves, palm communities, savannas, seasonally dry tropical forests, tropical moist and wet forests, cloud forests, deciduous sweet gums (i.e., *Liquidambar*), oak forests, oak forests, pine forests, grasslands, and high-elevation grasslands.

The conglomeration of natural ecosystems present in Central America is responsible for the fact that this region contains a significant proportion of the biodiversity of the planet in terms of species richness. The mere fact that the tropical rain forest, the most diverse biome of the planet, is widely distributed in the region implies that the biodiversity of Central America is of great significance. The presence of tropical rain forest is in fact one of the distinctive criteria defining the countries of megadiversity. Another important criterion for the definition of megadiversity is the presence of coastline and the ratio of coastline to land surface. The Central American coastline is approximately 5700 km (2900 km on the Pacific coastline and 2800 km on the Caribbean side); thus, there is a coastline surface area ratio of 0.011, which is comparable to that of some of the most important megadiversity countries. Undoubtedly, marine ecosystems and the presence of coral reefs (arguably the most species-rich marine ecosystem of Earth), including some of the most important coral reefs of the Western Hemisphere (located on the Caribbean coast of Central America), add to the diversity of the region.

The well-established gradient of decrease in species richness with latitude implies that, overall, the tropical ecosystems of Central America are not expected to be as species rich as their more equatorial counterparts. The available evidence indicates that this is the case. However, the diversity of ecological conditions, and therefore of distinct ecosystems, dictates that the identity of species and species composition change considerably among localities within this restricted area. Thus, species turnover (or beta diversity) is likely to be very high in this region. The scant information available for some areas of Central America (e.g., southern Mexico) suggests that indeed this is the case. However, biological inventories are still limited and the overall species richness and its geographic variation for many plant and animal groups are poorly known for the region. Thus, it is difficult to quantitatively assess the relative contribution of beta diversity to the overall biodiversity of Central America.

The available evidence indicates that this is the case. However, the diversity of ecological conditions, and therefore of distinct ecosystems, dictates that the identity of species and species composition change considerably among localities within this restricted area. Thus, species turnover (or beta diversity) is likely to be very high in this region. The scant information available for some areas of Central America (e.g., southern Mexico) suggests that indeed this is the case. However, biological inventories are still limited and the overall species richness and its geographic variation for many plant and animal groups are poorly known for the region. Thus, it is difficult to quantitatively assess the relative contribution of beta diversity to the overall biodiversity of Central America.
and Cedrela (cedro). Some Andean elements that belong to the paleooceanic realm, which comprises territories of the Andes, South America, South Africa, Australia, and New Zealand, are also represented in the Central American flora. These Andean elements include genera such as Podocarpus and the spectacular herbaceous plants of Guzmania.

The elements of the Neotropical realm are best represented in the tropical forests of the lowlands, whereas the Arcto-Tertiary and Andean elements are predominant in the vegetation of the temperate and cold highlands. Frequently, however, there is no clear demarcation between these types of floristic elements and in several localities of Central America there is a mixture of them producing, for example, remarkable tropical rain forests at elevations of 600–900 masl, which combine the typical tall tropical trees, lianas, palms, and tree ferns with deciduous forests, with a predominance of Nearctic elements, may include trees of clear tropical affinity such as Cecropia or Nectandra. Another remarkable feature of the Neotropical elements of Central American forests is that several of them comprise plants characteristic of the eastern United States, with a major interruption in their distribution in Texas but that penetrate into eastern Mexico and further south into Chiapas and Central America. Some of these are taxa as representative as Liquidambar styraciflua, Nyssa sylvatica, and Ostrya virginiana.

On the other hand, elements typical of the drier areas located toward the north of the Tehuantepec isthmus are absent in the state of Chiapas (and in most of tropical Central America), but they reappear further south in the dry lands of Guatemala (Departments of Zacapa, El Progreso, and Chiquimula). This is congruent with the suggestion, presented previously, that the climatic changes leading to the recent expansion of tropical rain forests must have fragmented the distribution of elements of arid and semi-arid ecosystems such as the arborescent cactus Myrtillocactus and trees such as Juliana asbriingensi and Aporopaneus paniculatus.

To this amalgamation of elements of tropical, temperate, and Andean affinity, an additional contingent of endemic elements must be added. Although some elements are distributed in most of Central America, others have a more restricted distribution within the zone. For example, although several species and genera typical of the floras of Guatemala and Belize extend their distribution into Chiapas, other genera of the rain forests of these two countries are not present in Chiapas or any other tropical forests of Mexico. This also indicates that for a portion of the flora of Central America the northernmost limit of distribution is Belize and northern Guatemala. On the other hand, the most important genera of trees of the temperate and cold ecosystems of Chiapas, including Pinus, Quercus, Liquidambar, Carpinus, Abies, Cupressus, Juniperus, and Tecomah, all of which are of northern affinity, find their southernmost distribution up to Guatemala or Nicaragua. The exception is the oaks, which extend well into the south, in the Colombian Andes. Likewise, several species of oak that in all probability evolved in Central American forests (e.g., Quercus hondurana, Q. poliacaulis, and Q. crispipolia) extend their distribution into Chiapas, which is their northernmost distribution.

A final aspect related to endemism has to do with the relatively recent discovery of new taxa as a result of biological surveys and taxonomic research in the region. In the past 10 years, two new endemic plant families have been discovered. One, the family Ticoendraceae, is a tree from Costa Rica, and the other is a herbaceous plant in the family Lacandoniaceae from the Lacandon forests in Chiapas. The latter is a truly revolutionary plant family in that the parasitic plants lack chlorophyll and the hermaphroditic flowers present the reproductive organs in a morphological disposition which is unknown among the flowering plants (male organs in the center of the flower and female organs in the periphery). Both taxa are of restricted distribution but Lacandoniaceae is known from a single population in a single locality. Further biodiversity exploration may yield additional, important taxonomic novelties.

In summary, although the degree of knowledge of biodiversity for the region is limited, the available information provides good evidence that the biological diversity of Central America is of special planetary significance. This is due not only to its quantitative aspects (e.g., diversity of species and ecosystems) but also because of the remarkable combination of life forms of different biogeographic affinity, together with endemic elements.

III. LIFE ZONES AND ECOSYSTEMS

The most synthetic and revealing way of describing the ecosystems of Central America is through the description of the vegetation. A premise for this is that the vegetation constitutes the most obvious descriptor of the ecosystem, constitutes the base of the food chain, and provides the structural matrix on which most com-
munities and populations of animals live or indirectly depend on.

A very popular classification system is widely used to describe the ecosystems of Central America. The system, based on the life zone concept elaborated by Holdridge (1947), combines three climatic characteristics of a given region—temperature, precipitation, and altitude—and an index of the availability of moisture, potential evapotranspiration, as a predictive model to forecast a corresponding vegetation type, which is termed a life zone. Although the Holdridge system is widely used, particularly in Costa Rica, the classification system developed by Miranda and Hernández-S. (1965) is the one most used for the Mexican vegetation. This system is based on the physiognomy (i.e., the appearance: life forms and height) and phenological character (e.g., whether the predominant plants are deciduous or evergreen) of the vegetation. The conceptual basis for this system is that vegetation physiognomy reflects the adaptive response of plants to the environmental characteristics where vegetation develops. Given the facts that these two systems do not coincide in vegetation nomenclature and that the number of ecosystem types resulting from both is very large and detailed, I present a system that attempts to capture the essential aspects of the two systems and provides a more simplified classification. In addition, this simplified system attempts to present a nomenclature that is comparable to that used more generally in the scientific community. The reader is advised to consult the original systems for a more detailed description of the ecosystems of Central America.

IV. TROPICAL RAIN FORESTS

The most distinctive aspect of this type of forest in Central America is, as in other regions, the great biological diversity of species. There are many examples. The Lacandon forests of Chiapas include approximately 4300 plant species, and the La Selva field station in Costa Rica harbors 1300 plant species and approximately 411 species of birds, 116 species of mammals, and 479 species of butterflies.

In terms of their physiognomy and structure, these forests are characterized by the presence of very tall trees (30–50 m) and by their evergreen vegetation—most of the plants retain their leaves year-round. Most of the trunks of the large trees, besides being straight, have diameters at breast height of between 30 and 150 cm. However, a few, such as the Ceibas trees, can have even thicker trunks and their height can reach beyond the forest canopy. These so-called emergent trees provide the typical appearance of a tropical rain forest as seen from afar. Another distinctive feature of these forests is the presence of irregular or undulating contours of the tree trunks of many species, particularly at the base where they become extended protrusions, approximately triangular in profile, which may play the role of support for the tree. The structures, called buttresses, can adopt several shapes or designs that can often serve as distinctive characteristics of particular species, such as in Dillnum guanense. The bark on the trunks, in general, is either smooth or scaly, with color ranging from light tones of Termitea amazonia and some figs to dark ones of the mahogany tree and shiny reddish hues of the guamo-limbo (Bursera).

Even when not clearly defined, from the canopy down there is a succession of layers of vegetation (stratification), from the subcanopy to the understory, with a spectacular profusion of life forms. The understory includes a layer of plants, most notably short palms, from less than 1 m up to 8 m. Several species of Chamaedorea exemplify this situation. In Central American rain forests the understory palms are typically accompanied by several species in the family Rubiaceae. At the ground level, the undergrowth is composed of ferns, several herbaceous plants such as ginger, some trailing plants, and the saplings or seedlings of the species found at higher levels. A direct consequence of this stratification and profusion of plants is that light diminishes as one moves from the top down to ground level, where only 1–3% of the available light reaches through. Such limited availability of light leads to a series of spectacular morphological and functional adaptations in the plants to compensate for the limitation of this critical resource. Such adaptations are in turn responsible for the existence of other physiognomic and structural features of the forest, including the occurrence of a great variety of climbing plants, both woody (lianas) and herbaceous (vines, mainly in the family Araceae), together with a variety of epiphytic (i.e., plants that live on top of other plants) and semi-epiphytic herbs, shrubs, and even trees. By far, lianas are the most conspicuous climbers since they can reach over 50 cm in girth and, depending on the locality, they can constitute 1–3 of each 10 trunk. In Central America, even some species of palms have evolved the habit of liana, as is the case for Domoncus and Chamaedorea elatior. The Araceae provide the most notable example of non-woody climbers, not only because they cover the trunks of the trees on which they climb and because of their abundance on the canopy but also because of their abundance as trailing plants at...
ground level, where they spread in search of trunks to climb. The vascular epiphytes, the most evident of which are the orchids, constitute a significant component of the biodiversity, given that the orchid family is commonly the most species-rich family in Central American forests. In addition to the orchids, the bromeliads constitute another important element of the epiphytic life form of these forests. Although bromeliads are not as species-rich, they are very abundant so that these two families constitute one of the most distinctive features of the physiognomy of the Central American rain forest.

A much less conventional type of climbing plant is that of the hemi-epiphytic trees, represented chiefly by the strangler figs, which also constitute a distinctive feature of the forests. Such plants germinate and take root on the branches of other large trees and then grow downwards at an impressive rate, producing many strong trunks that hold on the ground while covering the tree on which they established, forming a strangling network that eventually kills the initial support tree.

In addition to the impressive morphological adaptations described previously, most woody plants of the tropical rain forest have a series of ecological and functional attributes that allow them to deal with the problem of light limitation. Such attributes comprise a set of adaptations that fit the dynamic nature of the tropical rain forest. In essence, the strong winds that occur in these regions particularly during the rainiest season and winter months (November–February), overturn one or more of the giant trees of the canopy, creating so-called light gaps of up to 800 m². Depending on the locality, the incidence of such gap formation can be very intense, ranging from turnover rates (i.e., the number of years elapsed between two successive gaps) of 60 (e.g., in some Mexican forests) to 100 (e.g., in some Costa Rican forests) years. The result is the production of natural clearings with high light availability and higher temperature than in the shady understory. Many plant species produce dormant seeds equipped with photosensitive mechanisms that allow them to detect the amount and type of light of the gap and germinate in these conditions. This type of plant constitutes the group of the pioneer species, the most representative of which are Cecropia spp. and Ochroma, the balsa tree. The pioneer species display a very rapid growth and recolonize the gap by growing together with some of the young plants of the typical shade-tolerant species of the mature forest that might have been growing in the understory. The gap gradually begins to close, which demands the colonization of other species physiologically adapted to increasing levels of shading. These conditions make it impossible for the survival of the pioneer species that did not manage to grow and reproduce on time. The shade-tolerant species, provided with morphological, physiological, and growth mechanisms suited to scarce illumination, make their advance to the mature forest. This process of opening and closure of the forest creates a dynamic mosaic of vegetation with varying degrees of regeneration. In the tropical rain forests of Central America the mature phase is predominant and the most representative of the described physiognomy. In addition, in this phase, the leaves of the plants are predominantly dark green and sometimes shiny on the upper surfaces, tough in texture, and with sizes that range from medium (5–10 cm in length) to large (up to 1 m or more). As in other tropical rain forests, leaves have thin, tapering extensions on their tips, which are thought to serve to drain off the excess of water. Colorful, fleshy fruits are very abundant in the plants of the mature forest, but flowers tend to be mostly small and there is a predominance of light shades of white and green; bright and shiny colors are less abundant.

The complex and diverse vegetation serves as a matrix in which a highly diverse fauna is found, particularly for insects. Given the great diversity of plants and animals, a distinctive aspect of these forests is the complex network of interactions among different species, particularly between plants and animals: pollination, dispersal (the transportation of fruits and seeds by animals), and herbivory (the consumption of plant tissue by animals). As a result, the shapes, colors, and scents of flowers and fruits and seeds, and a wealth of toxic or attractive substances, determine the behavior, feeding, and sensory patterns of animals. These animals defoliate, seek out seeds, consume nectar and other fluids, guard and defend plants against herbivores, and even attempt to copulate with orchids that resemble and smell like female bees.

V. SEASONALLY DRY TROPICAL FORESTS

This kind of forest is found in warm climates, characterized by a long dry period (from 4 to 6 months), and at altitudes from sea level up to 1500 m. As a result, in physiognomic terms they are typified by the deciduous nature of their vegetation: green and luxuriant in the rainy season and grayish and leafless when it is dry. The highest trees rarely exceed 25 m, and the dominant...
flora comprises smaller trees, shrubs, and long-branched trees, although vines are also found. Commonly, two vegetation strata can be defined: the canopy, with trees 15–25 m tall, and the understory, with treeless 5–15 m tall and smaller shrubs. The ground layer is very sparse in general. Although lichens can be present, the scarcity of vines and epiphytes is notable, although they can be found in microhabitats with favorable conditions and along riverbeds. Some epiphytes, such as bromelids, can be found in abundance on trees growing on some suitable slopes. Another typical feature of these forests is the presence of numerous lichens attached to the branches of trees and shrubs, which acquire the colors of these lichens. In general, most woody plants have relatively small diameters, except in areas with greater soil moisture, such as in the vicinity of permanent or semipermanent bodies of water. As a consequence, Central American dry forests present two characteristic physiognomies: the seasonally deciduous forest associated with the hills, which is the most widespread type, and the riparian forest, smaller in extent, with larger trees that do not shed their leaves during the dry season. Frequently, riparian sectors of the dry forests include species of trees that are typically present in evergreen forests, including Brosimum alicastrum and several large figs. In addition to these two major variants of the dry forest, in rocky outcrops or as they extend into even drier climates or toward their northernmost distribution in Mexico, their average height decreases and there is an increased presence of thorny plants, mainly of the legume and cactus families. Another major distinctive feature of the trees in these forests is their bark, which in many species is smooth and shiny and frequently exfoliating (i.e., it peels off), as is the case of Jatropha or Bursera. In addition, several species present prominent thorns, as is the case of several members of the Bombacaceae family (Ceiba and Bombacopsis). Another distinctive aspect of the Central American dry forests is the abundance of species with small leaves, frequently of less than 5 cm in length, and many of the species present divided leaves composed of leaflets, although plants with entire or undivided leaves are also common. Several members of the legume family are representative of species with divided leaves. In addition, in marked contrast to the rain forest, dry forest plants typically present showy flowers, and many of the species produce their flowers in remarkable synchrony, particularly during the dry season when leaves are absent. These reproductive and vegetative rhythms are the norm among the plants of these forests, but notable exceptions occur, the most spectacular of which are those related to the patterns of leafing out. Although some species keep their leaves during the dry season, others such as Jacquinia pungens do not merely maintain their foliage during the harsh season but also lose it during the rainy season. The reasons for this leafing pattern are not fully understood by ecologists. Likewise, several species do not flower in mass or otherwise during the season of drought but rather at other times, a remarkable example of which is the Ipomoea tree.

The presence of endemic taxa in tropical dry forests is particularly high in these forests, in contrast to the rain forests, which share many of their species with South American forests. Endemic taxa in Central American forests include species and genera of a marked affinity with the dry lands of northern Mexico. In addition, at least one endemic life form is known to the tropical dry forests of Mexico—the Ipomoea tree referred to previously. This remarkable species is a member of the morning glory family, but one that has evolved as an arboreal, not a vine, life form.

Another significant contrast with the rain forest is the fact that dry forests have been studied to a much poorer extent throughout Central America. Nevertheless, the most recent review of this type of forest (Bullock et al., 1995) suggests that at least some aspects of the ecological complexity of these ecosystems are comparable to those of their rainy counterparts. Many complex plant–herbivore interactions, for example, are the same as those known for rain forests. Some of them, such as the famous mutualistic interaction between the bull's thorn Acacia shrubs/trees and their defending Pseudomyrmex ants, can be even more prominent in dry forests. In this remarkable interaction, the plants provide a home (the bull's thorns) and food (sugar produced in extrafloral nectaries and specialized, lipid-rich structures) for the ants. These, in turn, are very aggressive and attack potential herbivores of the plant and even remove the vegetation in the immediate vicinity of the plant, thus avoiding potential competition.

VI. TROPICAL CLOUD FORESTS

Although with considerable variation depending on the locality, elevation, and aspect, these forests owe their physiognomy, in general, to their enormous trees. Although not very rich in species, this group of trees is very abundant. Some of these may exceed 30 m, with trunks of up to 2 m in diameter at breast height, such as is the case of some oaks and the spectacular Talamanca...
Quercus candicans in the cloud forests of Chiapas. In the mature forest areas the cloud forest is clearly dominated by woody plants. Among these, the lianas, although not completely absent, are considerably scarce. A few of the trees lose their leaves during the driest time of the year, but because a high proportion of the trees do not lose their foliage or, if they lose it, they replace it very quickly, the forest as a whole can be regarded as an evergreen ecosystem. Although the cloud forests located in the southern parts of Central America are dominated by species of tropical affinity, toward their northern distribution, particularly in Chiapas, elements of boreal affinity become prevalent and intermingle with the tropical elements. This is the case for the oaks (e.g., *Quercus candelicans* or *Q. skinneri*) and especially for the sweet gum tree *Liquidambar styraciflua*. This noticeable tree, of approximately 30 m, extends in the Central American cloud forest from Honduras to Chiapas. Although the dominant trees of the more southern forests, such as in Costa Rica, are mostly 30–40 m tall, and a stratification is more clearly defined than in their more northern counterparts, the rest of the phytosociological characteristics are very similar. Although buttresses are present, they are not as developed as in the tropical rain forest trees. The understory and lower strata of the cloud forest are notable for the great profusion of palms and ferns, the most conspicuous of which are the spectacular tree ferns. In some areas these can reach up to 20 m. The ferns, including smaller species, tree ferns, and their immature forms, achieve a degree of abundance in the cloud forests that is hardly comparable to that of any other Central American ecosystem. Understory trees can be 8–15 m tall and frequently have leaning, crooked trunks and relatively long crowns. The shrub layer is fairly dense and can reach 2 or 3 m. The ground layer is considerably covered by ferns, *Selaginella*, and several broad-leaved herbs. Blooming coloration in several plants of these lower strata is quite common. Another distinctive feature of these forests is the great profusion of epiphytes and semi-epiphytic plants. Most of the trees are covered by dense associations of ferns, moss, orchids, bromeliads, peperomias, herbaceous vines, and other vascular plants such as the gesneriads. A spectacular fauna of characteristic birds, including quetzals, tanagers, and horned guans, accompanies such variety of plants.

Perhaps the most extensive cloud forest of Central America is that of Costa Rica. Nevertheless, this type of forest is of a very restricted extension and patchy distribution. As a result, and because of the relatively poor knowledge of its ecology and also because of its crucial role in retaining water, this Central American ecosystem requires attention and formal protection before it is converted into plantations and grasslands.

### VII. TEMPERATE FORESTS

Temperate-like forests of Central America include two major types (oak forest and pine forest), but in some parts, in which the dominant species overlap, a mixed oak–pine forest can be distinguished. Oak forests range from the tropical montane and premontane forest dominated by *Quercus* spp. and *Lauraceae* in Costa Rica to the clearly oak-dominated forests (“encinar”) of Chiapas, Mexico. The Costa Rican oak forests are evergreen communities of low to intermediate height, with trees 20–25 m tall and unbuttressed trunks and rough bark. The understory is open, with trees 4–15 m tall and with slender trunks and highly ramified crowns, with some tree ferns present. The understory is dominated by bamboo, particularly in the disturbed areas. The branches and trunks of trees are heavily covered with herbaceous epiphytes and moss. The oak forests are best represented in Chiapas, where they form distinct associations ranging from woodlands intermingled with the tropical rain forest (with *Quercus octopara*) or the cloud forest (*Q. candelicans* and *Q. skinneri*) to dense mono-dominated oak forests at elevations of approximately 3000 m, with trees 35 m in height (*Q. acatenangoensis*). The most widespread oak forests of Chiapas are those of intermediate elevations (700–2500 m). In the drier areas near the central depression of the state, these are low forests with grasses in the understory and they are sometimes adjacent to seasonally dry forests and savannas. In contrast, in areas of higher elevation and where winds are still loaded with water vapor, they are similar to cloud forests: dense, and loaded with epiphytes, particularly bromeliads and orchids. The most impressive forests of this type are those near the spectacular Montebello Lagoons. Other associations include the low oak forest of 2–4 m in height, which resembles a chaparral, near the city of Comitán.

The pine forests have their greatest distribution in the territory of Chiapas and extend, more sparsely and with fewer associations and diversity, toward the south to Honduras. With the exception of some fragmentary information for the pine forests of Chiapas, these Central American ecosystems are very poorly known scientifically. In Chiapas, these forests occupy most of the surface of temperate and cold areas, to which pines are very well adapted. This is due to the reduced surface of their leaves (needles), which are also well protected.
on their surface, allowing them to resist long periods of drought and low temperatures. In addition, the resin present in these plants’ tissue provides resistance to several pests and the thick bark of the trunk and branches help to reduce the effects of fire. In general, these forests present a clear dominance of a single tree species. They extend from intermediate elevations up to the timberline at approximately 4000 m and where rainfall does not exceed 1200 mm per year. However, even in areas of more precipitation, pine forests can be occupied by species with thin and flexible needles (Pinus strobus and P. ayacahuite). The most widespread pine forests, at elevations of 750–3000 m, are dominated by P. oocarpa, whereas P. hartwegi and P. rudis dominate the more restricted pine forests of the coldest regions (at elevations of 2900–4000 m).

In most cases, the understory is dominated by dense grassland. A significant, although seldom considered, aspect of the biodiversity of both oak and pine forests is the presence of a rich flora of shrubs and herbs, many of northern affinity but including several species endemic to these ecosystems.

This variety of Central American oak and pine forests, of peculiar and restricted distribution, is very poorly studied ecologically and from a biodiversity point of view, but these forests are disappearing at alarming rates, and most of what remains is heavily fragmented. It is likely that some significant aspects of their ecology and biodiversity have been very strongly disrupted or lost.

VIII. OTHER HIGH-ELEVATION ECOSYSTEMS

Two distinct ecosystems are present at the highest elevations in Central America: the paramo and the high-elevation grassland or zacatonal. The latter is an association of tall and dense grasses that extend in the clearings of pine and oak forests of the highest elevations. This indicates that their current extension is most likely secondary, that is, derived from human perturbation due to the considerable resistance of the grasses to the fires that sweep these regions. The height of these systems is 0.5–2 m and the dominant grasses include Epicampes macrura, Festuca amplissima, and Stipa ichu. The Central American paramo is the northernmost occurrence of the Andean paramo. In Costa Rica it is dominated by shrubs where drainage is adequate and by bogs where drainage is poor. Another portion of paramo occurs in Chiapas, at elevations of 4000 m or higher (above the timberline of the Tacana Volcano), where it is physiognomically similar to the zacatonal but the dominant grasses (Calamagrostis and Festuca) are shorter, sometimes forming a prairie. Frequently, the distance among grasses is considerable and there is much open space for the establishment of other plants, many of which barely rise above the level of the ground. This group of plants includes several members of the rose family (Alchemilla and Potentilla) and the daisy (Senecio and Gnaphalium) and cabbage (Draba) families. The small size and prostrate habit of the plants of this ecosystem are advantageous. They are able to take advantage of the heat of the layer of air close to the ground during daylight since the higher layers, even during daylight, are extremely cold and incompatible with their functioning and survival.

IX. CENTRAL AMERICAN ECOSYSTEMS IN LIGHT OF GLOBAL ENVIRONMENTAL CHANGE

Several aspects of global environmental change are predicted to have a profound effect on the biodiversity of the planet. Among these, climatic change has received the greatest attention, but recent information indicates that, in addition, the current patterns of land use (leading to deforestation and habitat fragmentation) may significantly affect biodiversity. Although information is limited, particularly for this region, it is of interest to review what the available models and information suggest regarding the situation of Central American ecosystems in the light of climatic change and land use patterns.

A. Climatic Change

General circulation models have been developed to forecast the changes in temperature and rainfall and their geographic variation under a given set of assumptions, mostly related to CO2 concentrations. These models suggest that the greatest changes of climate are to be expected at higher latitudes. Consequently, given the relatively low latitude of Central America, it could be expected that the natural ecosystems of this region would be affected to a minor degree. Indeed, the maps resulting from these models depict Central America as a region of little or no climatic change. Although this is an optimistic result, it is important to bear in mind that such maps and forecasts are, by necessity, of very gross resolution. Unfortunately, no detailed analyses
are available to assess the situation at a finer geographical scale. Nevertheless, a recent study on the expected changes in vegetation coverage resulting from climatic change in the forests of Mexico (Villers and Trejo, 1998) provides some useful insights, given that the study included the northern part of Central America (Fig. 1). In addition, the study considered the major Central American ecosystems described previously (i.e., tropical rain forest, tropical dry forest, cloud forest, pine, and pine–oak temperate forest). Ecosystem-response scenarios were developed with two models based on a doubling of CO2 concentration (the Geophysical Fluid Dynamics Laboratory and Canadian Climate Center models) and under the assumption of a homogeneous temperature increase of 2°C and 10% reduction in rainfall (Fig. 2). In synthesis, the results show that, in comparison to their current potential distribution, tropical forests of the warm lowlands would be the least affected: Seasonally-dry forests are expected to remain about the same, whereas rain forests are even predicted to increase. In contrast, temperate forests appear to be affected the most: Pine forests are predicted to disappear, whereas cloud forests would experience reductions of 45–75%. Moreover, the results were very consistent among the different models. In summary, whereas global models forecast little climate change impact on the ecosystems of the region, more detailed models suggest that this is likely the case for the tropical ecosystems of the warm/lowland climates but the temperate ecosystems are expected to be significantly impacted.

B. Land Use Patterns

The best available descriptive index of the patterns of land use is the rate of deforestation. Detailed studies on the rates of deforestation have been published only for some regions of Central America, particularly Costa Rica and the Mexican portion of the region. Studies on deforestation in the tropical rain forest of Lacandonia, Chiapas (Mendoza and Dirzo, 1999) show a considerable degree of variation in deforestation rates among localities, with a range of 0–8% and an average of 1.6% per year. Even if these values seem low, the estimated absolute deforestation rate for this area is 6286 ha per year.

The situation in Costa Rica is particularly revealing, given that it is a country that has made a significant effort to conserve its natural ecosystems. Despite this effort, the available statistics suggest that deforestation has been extremely high in this country. For example, studies from the early 1990s indicate that most of the country's ecosystems are heavily deforested and the percentages of area of habitat lost range from 32 (montane rain forest) to more than 90% (dry deciduous forest, lowland moist forest, premontane moist forest, and lower montane moist forest). The aggregated value for 11 of the 12 ecosystems (life zones) of the country is reported to be approximately 80%.

These isolated studies give the impression that deforestation in Central America is considerable and an aggregated analysis (World Resources Institute, 1991) confirms such impression and permits comparison of the overall situation of the region with that of other parts of world (Fig. 3). The estimated amount of percentage of forest converted to nonforest use in Central America is about 3.5 times greater than the corresponding estimate for the whole world and more than twice the value for the forested ecosystems of South America and Asia.

In synthesis, the information presented in this section indicates that even if climatic change will not
have a major impact on the tropical forests of Central America, forests of all types seem to be seriously threatened by the current patterns of land use. Moreover, the available information suggests that such threats might compensate for the potentially positive effects, or exacerbate the negative effects, of climatic change. Finally, the current patterns of land use and the expected effects of climatic change on the temperate forests might have a significant effect on the functioning of the ecosystems and overall structure and composition of the biodiversity of Central America. The relevant information to assess the consequences of the loss or degradation of the Central American ecosystems is very limited. In addition, such an analysis is beyond the scope of this article. Nevertheless, our knowledge of their biodiversity and the threats they are currently experiencing indicates that a significant effort toward their conservation and wise management is a major agenda for the scientific community and society at large.

See Also the Following Articles

Biodiversity-Rich Countries • North America, Patterns of Biodiversity in • Rainforest Ecosystems, Animal Diversity • Rainforest Ecosystems, Plant Diversity • South America, Ecosystems of

Bibliography


CLADISTICS

Ian J. Kitching, Peter L. Forey, and David M. Williams
The Natural History Museum, London

I. Relationships
II. Characters and Coding
III. Cladistic Analysis
IV. Cladogram Evaluation
V. Simultaneous and Partitioned Analysis

GLOSSARY

apomorphy A derived character or character state; if two or more taxa share apomorphies, these are referred to as synapomorphies. clade Group of taxa diagnosed as monophyletic by the discovery of homologies (or synapomorphies). cladogram Branching diagram specifying hierarchical relationships among taxa. cladogram support Tests that permit some evaluation of how well data fit a cladogram. consensus cladogram (tree) Branching diagram that summarizes the common branching patterns from two or more cladograms. homology Two characters passing the similarity, conjunction, and congruence tests are termed homologous; in cladistics, homology is synonymous with synapomorphy. homoplasy A character or character state acquired by parallel or convergent evolution that bears resemblance to a character in a different group. monophyly Relationship between taxa united by a synapomorphy. optimization Procedure for reconstructing the most parsimonious sequence of character change on a cladogram. parsimony General scientific principle that given alternative explanations or hypotheses for a set of observations or data, the most corroborated is that requiring the fewest ad hoc (ancillary or additional) hypotheses. plesiomorphy An apomorphic character or character state that specifies a more inclusive group than that under consideration.

CLADISTICS is a class of methods of biological classification that groups taxa hierarchically into discrete sets and subsets. This article presents the principles and concepts of cladistics and describes the principal analytical methods. The operations by which observations of organisms are coded for analysis are explained, followed by the methods for reconstructing the hierarchical relationships among taxa (usually expressed as branching diagrams termed cladograms). Statistics and principles for determining the degree of fit between data and cladograms are discussed, which permit choices to be made among competing cladograms.

I. RELATIONSHIPS

The basic concept of cladistics is that genealogical connections among organisms are expressed in relative
terms. Consider three taxa, A, B, C, whose genealogical relationships are as given in Fig. 1a. Taxa B and C are more closely related to each other than either is to taxon A because they share a common ancestor, x (which lived at time $t_1$), that is not shared with taxon A. Similarly, taxon A is more closely related to the group (B + C) because A, B, and C together share a unique common ancestor, y, that lived at an earlier time ($t_0$). In a real example (Fig. 2), the human and turkey are considered to share a unique common ancestor (w) that lived at time $t_3$. Similarly, the frog, turkey, and human are more closely related to each other than to either the perch or dogfish because these three taxa uniquely share a common ancestor $x$ that lived at time $t_2$. The human and turkey are called sister-groups. Likewise, in this example, the frog is the sister-group of (human + turkey). The aim of cladistic analysis is to infer the sister-group hierarchy of life-forms by analysis of characters and to express the results as branching diagrams. These diagrams are called "cladograms" because they identify a hierarchical arrangement of taxa based on homologies termed "clades."

A. Types of Characters

Two types of characters are recognized based on where they occur in the inferred phylogenetic history of a group (Fig. 1b). The character that occurs in the ancestor is termed "plesiomorphic" (near to the ancestral morphology) and the derived character is "apomorphic." Apomorphic and plesiomorphic are relative terms, that is, relative to a particular systematic problem. In Fig. 1b, character $a'$ is apomorphic with respect to character $a$, but plesiomorphic with respect to character $a''$. Cladistic analysis proceeds by identifying shared apomorphic characters or "synapomorphies." In Fig. 2, a four-chambered heart and endothermy are synapomorphies that suggest the human and the turkey share a unique common ancestor w. The cladogram implies that these two characters arose in ancestor w and were then inherited by both the human and the turkey. Synapomorphies may therefore be considered as evolutionary homologies. In contrast, the shared possession of internal nostrils and pentadactyl limbs by the human and turkey does not imply that they share a unique common ancestor because these attributes are also found in the frog. These shared primitive characters (or "sympleiomorphies") are inherited from an ancestor more remote than the most recent common ancestor of the human and the turkey. They are thus irrelevant to the hypothesis of a relationship between the human and the turkey. However, with respect to the more inclusive three-taxon problem comprising the frog, turkey, and human, internal nostrils and pentadactyl limbs are relevant. At this level, they are synapomorphies suggesting that these three taxa form a group with a common ancestry at x. Apomorphies occurring in only a single terminal taxon are termed "autapomorphies." In Fig. 2, these are prismatic cartilage (dogfish), spiny

![FIGURE 1](image-url)
A phylogenetic tree for five taxa of vertebrates. Three monophyletic groups are established using characters 1–6, while autapomorphic characters 7–12 diagnose the terminal taxa. The group Pisces is paraphyletic because one of its included members (the perch) is cladistically more closely related to Tetrapoda.

FIGURE 2. A phylogenetic tree for five taxa of vertebrates. Three monophyletic groups are established using characters 1–6, while autapomorphic characters 7–12 diagnose the terminal taxa. The group Pisces is paraphyletic because one of its included members (the perch) is cladistically more closely related to Tetrapoda.

Fin rays (perch), pedicillate teeth (frog), feathers (turkey), and hair and mammary glands (human). However, if a terminal taxon is itself a group, then its autapomorphies are also synapomorphies of its component taxa.

Thus, as with the cladistic meaning of relationship, characters are also relative, depending on the systematic problem under consideration. Furthermore, it should be stressed that characters are observations of the features occurring in organisms and the explication of their hierarchical distribution need not imply a particular theory of evolution.

B. Parsimony

Relationships among three taxa (as in Fig. 1a) can be resolved in three ways—A (B C), B (A C), and C (A B)—whereas for four taxa (as in Fig. 1b) there are fifteen possible fully resolved cladograms. In cladistic analysis, parsimony is the universal criterion for selecting among alternative hypotheses of character distribution. Characters are fitted onto alternative topologies and the cladogram that accounts for the greatest number of characters in the simplest way is chosen as the best hypothesis of relationships.

Suppose six characters are distributed among four taxa as shown in the taxon/character matrix in Fig. 3. Taxon A has none of the characters but the other three taxa each have a different complement. Characters 2 and 4 are autapomorphies because they are each present in only one of the taxa. They are uninformative for grouping taxa (they serve only to diagnose the terminal taxa). Characters 1, 3, 5, and 6 are potentially useful because they occur in more than one taxon. Given the
three taxa that have potentially informative information, there are three ways in which these taxa can be arranged dichotomously (Fig. 3b–d). If the characters are now placed onto each possible cladogram, according to the groups they specify, then three different results are obtained (Fig. 3e–g). In Fig. 3e, all characters except one appear only once. In this solution, character 6 must be assumed to appear twice, once in taxon B and once in taxon C, which are not sister-groups. In this example, the behavior of character 6 is homoplastic, that is, it occurs more than once on the cladogram and is said to be a homoplasy. In contrast, in the other two topologies (Fig. 3f and 3g), we must assume that two or more characters appear more than once. Hence, the cladogram in Fig. 3e accounts for the distribution of the characters in the most economical way and is thus the preferred solution.

The distribution of characters can also be regarded as the number of steps on a cladogram, which, in Fig. 3, is the number of instances where a character is gained. In Fig. 3e, this is seven, while the other cladograms (Fig. 3f and 3g) are more costly, requiring nine and eight steps, respectively. The concept of steps is actually a little more subtle than the sum of character gains because a character may appear at one point on a cladogram and then disappear at another point. For example, another explanation for the distribution of character 6 in Fig. 3e is to assume that it is gained by the group (B + C + D) and then lost in taxon D. Each change, whether gain or loss, is considered a step. In this example, both accounts of character change demand two steps and therefore both hypotheses of character change are equally parsimonious. The sum of the number of steps on a cladogram is termed the length of the cladogram, irrespective of whether the changes are gains or losses. The most parsimonious solution is also known as the optimal cladogram and the other cladograms (i.e., those requiring more than the mini-

mum number of steps to explain the character distributions) as suboptimal.

It is possible for a given set of characters to yield two or more equally most parsimonious cladograms. Then, we may prefer to accept one of the solutions based on other criteria, such as a closer agreement with the stratigraphic record or by differential weighting of one type of character change relative to another. Alternatively, we may simply accept that the conflict in the data is such that we cannot derive a unique most parsimonious solution. For certain purposes, we may choose to combine those components common to the different solutions to form a consensus cladogram.

C. Groups

Cladistics recognizes only monophyletic groups of organisms, which are those based on synapomorphies. Monophyletic groups are the only groups that can be circumscribed by objective boundaries. In evolutionary terms, monophyletic groups comprise the most recent common ancestor and all of its descendants. In Fig. 2, Amniota, Tetrapoda, Osteichthyes, and Gnathostomata are all monophyletic. Two other types of "groups" are sometimes referred to but these are not groups in the same sense as monophyletic groups. Paraphyletic "groups" are based on symplesiomorphy; in evolutionary terms, their members are linked by common ancestry but one or more of the descendants of the most recent common ancestor are excluded. In Fig. 2, Pisces (fishes) is a paraphyletic assemblage. Many taxa traditionally regarded as ancestral, such as fishes, reptiles, and green algae, are paraphyletic. Polyphyletic "groups" are based on homoplasy, that is, characters that are considered convergently derived and that cannot be inferred to have been present in the most recent common ancestor of the included taxa. In Fig. 2, an assemblage comprising the dogfish and the turkey (perhaps based on the observation that both lay eggs surrounded by a shell, although no one would claim such a homology) would be a polyphyletic group.

D. Cladograms and Phylogenetic Trees

A cladogram is a diagram that summarizes a pattern of character distribution. Usually, a cladogram is drawn as a branching diagram (e.g., Fig. 1). The nodes denote a hierarchy of synapomorphies but there is no necessary implication of ancestry and descent. Cladograms may also be written in parenthetical notation or illustrated as a Venn diagram (Fig. 4a), which conveys the same grouping information as a branching diagram. In contrast, phylogenetic trees include a time axis and embody concepts of ancestry and descent with modification. In phylogenetic trees, the nodes denote ancestors (known or hypothetical) and the branches imply character change. Several phylogenetic trees may be compatible with the pattern of character distribution implied by a cladogram (Fig. 4b). Some of these trees allow the possibility that one or more taxa are ancestral to others. Only the phylogenetic tree that assumes all nodes represent hypothetical ancestors has the same topology as
the cladogram. Thus, cladograms are more general than phylogenetic trees, which are precise statements about ancestry and descent.

II. CHARACTERS AND CODING

Opinions differ over the nature and discovery of taxonomic characters. One view holds that characters are properties of organisms that provide quantifiable variation. Alternatively, characters may be viewed as theories concerning two (or more) attributes, which may look different but are nevertheless considered the same. This latter view is embraced within a general understanding of homology, such that characters may be understood in the same manner as homology. There is a lack of agreement over what indicates the discovery of a character. However, all definitions of homology suggest that it concerns features that are similar in different taxa sharing a recent common ancestor. Such definitions satisfy as explanations but do not aid discovery.

A. Homology

Within cladistics, various tests assist in establishing homology. One view of characters is that they are identical in meaning and discovery to homology, and homology may be conceived as a series of three tests that apply to methods of comparison.

The similarity "test" suggests that without evidence to allow direct comparison of one feature with another, there would be no proposition of homology and, consequently, no concept of a character. This "test" is not exact and cannot be taken to imply "identity." For example, comparisons may consider the detailed similarity of any two stamens or the inferred similarity of mammalian stapes with gnathostome hyoid arches.

The conjunction test suggests that two features that co-occur in the same organism cannot be considered homologous. A familiar example, albeit contrived, is an angel with both wings and forearms. The two kinds of limbs in the same individual cannot be considered homologous. Many comparisons fail this test and are often associated with "homomony" or serial homology (e.g., the individual vertebrae of a single vertebral column or the abdominal appendages of arthropods).

The congruence test is considered the most exacting and refers to the support afforded to one homology by others. In other words, homologies are considered to have passed the test if there are other homologies that specify the same taxon. Congruence is actually an analytical procedure and is usually considered in terms of parsimony. However, it also points to another property of homology, that is, homology can never be proven. As data are accumulated, previously supported homologies may be overturned and new theories of homology established in their place.

B. Character Recognition

It is generally agreed that characters, however conceived, are based on observations. Stated simply, a feature (e.g., stamens, shoulder girdles, wings) is observed in a particular specimen and directly translated into the character. This approach may initially seem useful and would eventually lead to enumeration of all features of the specimen. However, the final list would not consist of "characters" but would be an inventory of "features," each being a descriptive element of the specimen and implying that such descriptions apply to all specimens of the same taxon. Each descriptive element contains a notion of theory.

Suppose the specimen examined is a rat. Initially, it would be straightforward to describe: head, body, limbs, tail, and so on. More detailed examination would reveal, for example, a vertebral column. We identify the vertebral column by drawing on knowledge of previous studies of rat anatomy and are able to confirm its detailed similarity to other vertebral columns. In so doing, we assimilate what is already known of vertebrates: they are animals with a vertebral column. If this process were performed for all features, then it would seem that all attributes of this single rat could uncover its place in the hierarchy of life, at every inclusive level. In this sense, taxonomic characters are very much like homologies: features that (potentially) specify a particular taxon. A vertebral column does not tell us it is a rat; it tells us that it is a vertebrate. In this sense, the vertebral column is a feature of any particular rat but only a characterize of vertebrates. This distinction identifies the general task of systematics: to identify the level at which various attributes are homologies (characters) diagnosing taxa.

C. Kinds of Characters

Characters are often thought of as comprising different types. Some refer to different numbers of a feature, and others refer to differences in structure. For example, variation in stamen structure in angiosperms encompasses both different forms of anthers and filaments and differences in their numbers. This exemplifies the distinction between quantitative and qualitative characters, the former usually being counts ("meristic characters") or measurements ("biometric characters"), the latter relating to structural differences. Quantitative
characters may be problematic for cladistics insofar as it can be difficult to render measurements and counts as meaningful homology statements. This is not to say that such characters are not useful, for they can serve to identify particular specimens. However, their use may be limited because they are not always amenable to cladistic analysis.

D. Characters as Phylogenetic Evidence

Although structural evidence is sought for cladistic purposes, the observed features themselves are not necessarily the characters. For example, some organisms have fins, others have arms, and yet others have wings. Studies of fins, arms, and wings show that they have certain parts in common as well as certain parts that are unique. These common properties might suggest an initial proposition that fins, arms, and wings are all examples of a single character, in this case "paired appendages." However, further details are needed to confirm this hypothesis.

Suppose the "wings" examined were from an insect such as a housefly. It might still seem reasonable to call them "paired appendages," but in this example there is nothing in the detailed anatomy to suggest that housefly wings and mammalian arms (or even bird wings) are in any way "the same." One conclusion is that the term "wings" is ambiguous when describing attributes of organisms but not when describing the function of these attributes (here, flight). Wings may indeed be considered as a part of an organism but flight is its assigned function (usually). A more reasonable conclusion is that "wings" is not a character at all but a functional attribute. Thus, the wing of a bird is better considered as a modified "paired appendage," modified for flight. The problem, however, is not simply semantic: if the comparison was made between a bird and a bat, then detailed anatomy does suggest that both are indeed "paired appendages" and that both are modified for flight. Our current understanding of vertebrates suggests that birds and bats do not form a monophyletic group. Hence, bird wings and bats wings would be considered two characters rather than one when interpreted on the cladogram (Fig. 5). Their "sameness" is captured as "forelimbs," their differences as wings of birds and wings of bats.

Consideration of the wings of birds, the forelimbs of tetrapods, and the fins of fishes together identifies a well-defined character ("paired appendages") with various manifestations. These manifestations might suggest particular taxonomic groups. There might be a taxon with "fins," a taxon with "arms," and a taxon with "wings." This, of course, was precisely the situation for many years: fishes (Pisces) have fins, tetrapods have arms, and birds (Aves) have wings (Table I). Derived from these observations is the implication that fins, wings, and arms are in some way connected, other than by being "paired appendages":

| fins → arms → wings |

Furthermore, that connection might be viewed in an evolutionary context, such that it represents the transformation of one manifestation (e.g., fins) into another (e.g., arms):

- Fins
- Standard fins (fins)
- Modified fins (arms)
- Modified arms (wings)

<table>
<thead>
<tr>
<th>TABLE I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon</td>
</tr>
<tr>
<td>Pisces</td>
</tr>
<tr>
<td>Tetrapods</td>
</tr>
<tr>
<td>Birds</td>
</tr>
</tbody>
</table>
The features are no longer structured as a series of alternatives, as in the first example (fins–arms–wings), but now specify a nested set of relationships. The proposition is that wings are really kinds of arms, and that arms are really kinds of fins, and fins represent the entire set of animals with "paired appendages." It is possible to interpret all of these taxa (fishes, tetrapods, and birds) as having fins. Hence "fins" is not a character of fishes but rather of gnathostomes (in this example, fishes + mammals + birds). Consequently, both the character "fins" and the taxon Pisces disappear. This view confirms the notion that characters are hypotheses drawn from observations rather than simply the observations themselves. Such hypotheses are identical to those made for general statements of homology. Superficially, the relationship between fins, arms, and wings may be considered identical to "fins → arms → wings," as was implied by Hennig (1966) in his concept of "transformation series." However, it is possible to view characters as more general, specifying particular relationships in terms of a definitive statement connecting to a taxon.

E. Character Coding

One significant outcome of theories relating to characters is how they might be represented numerically for cladistic analysis. For example, one might code each "character" (fins, wings, arms) separately (Table II, characters 2–4). This scheme reflects the "uniqueness" of each attribute but contains no information relevant to recognizing that the three observed forms are connected as "paired appendages." This approach is referred to as "absence/presence binary coding," because a positive value (usually 1) is assigned to the presence of a feature and a negative value (usually 0) is assigned to the absence of the feature. Alternatively, one might represent the same series of observations in a single column to signify their connection (as "paired appendages"), then assign each unique feature a separate value (Table II, character 1). This is "multistate coding" and considers the character to be composed of discrete states that bear some (usually unspecified) relationship to one another. Hence different values appear in the same column and are treated as dependent on the other values. This might not be seen as completely sufficient, as additional information would be needed to specify the exact nature of the connection. For instance, one might wish to specify that the "characters" are connected but that the nature of that connection is unknown. Choices of this nature relate to character optimization (see Section III,C).

To summarize, characters have their origin, but not their identity, in observations. Characters are what lead us to suspect that taxa exist (vertebral column → vertebrates) and hence are identical to conjectures of homology derived from empirical investigation of specimens. Homology is the relation that specifies taxa and that implies an intimate relationship between characters and taxa. Both are the results of analyses and are discovered by our investigation of features.

III. CLADISTIC ANALYSIS

A. Cladogram Construction

The original method of cladogram construction was proposed by Hennig (1950, 1966) and is thus known as Hennigian argumentation. In this approach, characters are first polarized into plesiomorphic and apomorphic states. The groups thus diagnosed by synapomorphies are then organized manually into a cladogram. However, this procedure can only find the most parsimonious cladograms when the data are free or nearly free of homoplasy (i.e., the fit of data to most parsimonious cladogram is perfect or nearly so). For larger and more complex data sets, computerized algorithmic methods become a necessity.

There are two main computerized approaches to cladogram construction. Exact methods guarantee to find the most parsimonious cladograms. The simplest exact method is "exhaustive search." First, three taxa are chosen and connected to form the only possible unrooted, fully resolved cladogram for these taxa. Then, a fourth taxon is selected and added to each branch to yield the three possible fully resolved, partial, unrooted cladograms for four taxa. A fifth taxon is then selected and added to each of the five branches on these three partial cladograms to yield the fifteen possible fully resolved unrooted topologies for five taxa. This process is continued, following every possible path of taxon addition, until all taxa have been added and all possible fully
resolved cladograms have been found. The lengths of these cladograms are then calculated and the shortest is chosen as the optimal solution(s). However, as the number of taxa increases, the number of cladograms to be examined rises exponentially and the time required for exhaustive search soon becomes unreasonable.

One exact method that does not require every possible cladogram to be evaluated is "branch-and-bound" analysis. In this approach, a preliminary cladogram is constructed and its length is set as the upper bound for subsequent searches. A procedure similar to an exhaustive search is then undertaken but at each step the length of the partial cladogram is recorded. Whenever this length exceeds the current upper bound, that partial cladogram is rejected (and so, consequently, are those complete topologies that would be derived from it by adding the remaining taxa). By this means, the number of topologies to be examined is reduced. Once all taxa have been added, the length of the complete cladogram is examined and if it is equal to the upper bound, then that topology is retained as a most parsimonious cladogram. However, should this cladogram be shorter than the current upper bound, then its length is substituted as a new upper bound. This important procedure allows subsequent partial cladograms to be rejected quickly and thus speed up analysis. This process continues until all possible paths have been examined, whence the set of optimal cladograms will have been found.

For large data sets (more than 30 taxa), even branch-and-bound analysis can be too time-consuming. In this case, approximate or "heuristic" methods are used. These approaches examine only a subset of all possible topologies and thus are not guaranteed to find the most parsimonious cladogram(s). However, they are faster than exact methods for large numbers of taxa and thus certainty of finding the optimal cladogram(s) is sacrificed for decreased computational time.

Heuristic analysis comprises two stages. In the initial building phase, a cladogram is constructed using a process of "stepwise addition." The order in which taxa are added is termed the "addition sequence" and there are various ways in which taxa may be added. Once a complete cladogram has been constructed, attempts can be made to improve upon it by performing a series of rearrangements called "branch-swapping." The cladogram is cut into two or more partial cladograms, which are then recombined in order to try to find new, shorter topologies. The efficiency of current branch-swapping algorithms in finding most parsimonious cladograms is very high, but it is always possible that they can become trapped in a local optimum. Thus, one should always be aware with heuristic analyses that shorter topologies than those reported may exist.

B. Character Polarization and Cladogram Rooting

Manually implemented cladistic methods, such as Hennigian argumentation, require synapomorphies to be identified in advance of cladogram construction. The process through which plesiomorphic and apomorphic characters are distinguished is termed "character polarization." Numerous criteria for polarizing characters have been proposed, but only two are now considered valid.

The first criterion, called "outgroup comparison," was classified by Nelson (1973) as an "indirect" method, because it draws upon evidence from a source (the "outgroup") that is external to the taxa under investigation (the "ingroup"). In its most basic form, polarization using outgroup comparison can be defined as follows: "For a given character with two or more states within a group, the state occurring in related groups is assumed to be the plesiomorphic state" (Warren and Wheeler, 1981: 5). This definition is adequate when all outgroup taxa share the same state, but it is insufficient if the outgroup taxa are heterogeneous. Maddison et al. (1984) further noted that it is inappropriate to estimate the state in the most recent common ancestor of the ingroup (the "ingroup node"). Rather, it is the state at the next most distal node, linking the ingroup to the first outgroup (the "outgroup node"). That should be estimated if the solution is to be globally optimal, and they described an algorithmic approach to such reconstructions.

In contrast, Nelson (1973) classified the "ontogenetic criterion" as a "direct" method, because its implementation relies on evidence from the ingroup taxa alone. It is defined as follows: "Given an ontogenetic character transformation from a character observed to be more general to a character observed to be less general, the more general character is primitive [plesiomorphic] and the less general character derived [apomorphic]." (Nelson, 1978: 327). For example, the embryos of both sharks and frogs have cartilaginous skeletons. However, this condition persists into the adult shark, but in frogs the cartilage is replaced by bone during ontogeny. In this example, a bony skeleton is observed to be less general (occurring only in the frog) and is thus interpreted as apomorphic. In this context, "more general" is defined as that occurring earlier in ontogeny.
As such, the more general character is not simply the more common (although this may often be the case) and the ontogenetic criterion does not equate with commonality. What is important is that the less general character is nested within the observed distribution of the more general character. This requirement is violated by ontogenetic sequences that are secondarily simplified through paedomorphosis or neoteny. Such an ontogeny cannot be interpreted as proceeding from the more general to the less general, and Nelson's criterion will not allow us to distinguish a secondarily reduced ontogeny from the plesiomorphic sequence. Then, we rely on congruence with other characters to make the distinction.

Most recent cladistic studies do not actually include a priori polarization of characters, but undertake “simultaneous, unconstrained analysis”: “simultaneous” because both ingroup and outgroup taxa are analyzed together, and “unconstrained” because outgroup taxon relationships are unspecified before analysis. Cladograms are then rooted between the outgroup node and the remaining outgroup taxa, at which point character polarities are established.

C. Optimization

Optimization is the process of determining the sequence of character state changes on a cladogram in order to test hypotheses of transformation. If the data include characters coded as multistate, then these may be interpreted according to many different optimality criteria, of which the best known are Wagner and Fitch optimization.

Wagner optimization (Fig. 6a–b) is used for “ordered” or “additive” multistate characters, in which
transformation between successive states are considered as incremental. Thus, the changes $0 \rightarrow 1$ and $1 \rightarrow 2$ each ‘cost’ the same number of steps (usually one), whereas the change $0 \rightarrow 2$ is considered to ‘pass through’ state 1 and thus costs two steps. Costs are asymmetrical, so that the changes $0 \rightarrow 1$ and $1 \rightarrow 0$ both constitute a single step (termed ‘free reversibility’)

Wagner optimization is implemented in two stages. First, the minimum number of steps for a character is determined. On a pass down the cladogram, from the most distal taxa toward the root, each of the internal nodes is assigned a ‘state set’ (Fig. 6a), which is defined as the intersection of the two derivative state sets. If the intersection is empty, then the smallest closed set that includes an element of each derivative state set is assigned. For example, the intersection of the state sets of taxon F (2) and taxon G (4) is empty. Thus, the smallest closed set, $(2–4)$, is assigned to the node linking these two taxa. In contrast, the intersection between this state set and that of taxon E is not empty. Both contain state 2, and this value is assigned to the node joining taxa E, F, and G. Unambiguous states are then assigned to internal nodes by a second pass, up the cladogram (Fig. 6b), to produce a “most parsimonious reconstruction.” If the state set of a node contains more than one value, then the node is assigned the value that is closest to that of the node of which it is a derivative. Thus, the node joining taxa F and G is assigned state 2, because this is the value of the next most distal node. The most parsimonious reconstruction has six steps because the character is ordered. As a result, the change from state 2 to state 4 along the branch leading to taxon G counts as two steps.

Fitch optimization (Fig. 6c–d) is used for ‘unordered’ or ‘nonadditive’ multistate characters in which transformations between any two states are considered equal. Thus, the changes $0 \rightarrow 1$, $1 \rightarrow 2$, and $0 \rightarrow 2$ all cost a single step. As with Wagner optimization, costs under Fitch optimization are freely reversible.

Fitch optimization follows similar procedures to Wagner optimization but with two differences. First, the state set assigned to an internal node is the union of the two derivative state sets (Fig. 6c). Second, when determining the most parsimonious reconstruction, a node with an ambiguous state set is assigned the value of the next most distal node if that value is an element of the ambiguous state set. Otherwise, an element is selected arbitrarily. This most parsimonious reconstruction (Fig. 6d) has four steps because the character is unordered. As a result, the change from state 1 to 3 along the branch leading to taxon D and the change from 2 to 4 along the branch leading to taxon G each count as a single step.

These procedures do not necessarily yield a unique most parsimonious reconstruction. For example, it is equally parsimonious to optimize the ordered character using the nodal state reconstructions shown in Fig. 6d for the unordered character. State 3 is assigned to the node joining taxa C–G, and is followed by a two-step change, $1 \rightarrow 3$, on the branch leading to taxon D and a one-step change, $1 \rightarrow 2$, on the branch joining taxa E–G. The steps on the branches joining taxa B–G and leading to taxon C would be lost, thus maintaining the most parsimonious length of six steps. This type of optimization, in which changes are placed onto the cladogram as far from the root as possible, is called ‘delayed’ or ‘slow’ transformation. In contrast, ‘accelerated’ or ‘fast’ transformation places changes onto the cladogram as close to the root as possible, as in Fig. 6b. When alternative most parsimonious reconstructions are possible, character optimization may lead to ‘spurious resolution,’’ in which groups appear to be resolved but have no unambiguous support in the data. Such groups are not strong hypotheses of relationship, and Nixon and Carpenter (1996) suggested that they should be eliminated wherever possible (without violating the minimum length requirement). Those cladograms that remain, which are both of minimum length and have all groups supported unambiguously by data, are termed ‘strictly supported cladograms’ and are the preferred topologies.

IV. CLADOGRAM EVALUATION

A. Character Fit

The preferred solution to a cladistic analysis is the most parsimonious cladogram because this represents the simplest explanation of the data with the number of ad hoc hypotheses of homoplasy kept to a minimum (see Fig. 3e). Consequently, the most basic measure for assessing the fit of data to a cladogram is cladogram length. The most parsimonious cladogram has best fit because it is the shortest; longer cladograms have poorer fit. However, the length of the most parsimonious cladogram is partly dependent on the absolute size of the data set from which it is derived. Larger data sets will necessarily yield longer cladograms than smaller data sets. A binary character will display perfect fit when it is placed on a cladogram with a single step. Homoplasy is manifest as an increase in the number of steps. The
amount of homoplasy implied by a character on a cladogram is measured by the ‘consistency index’ (CI), which is the ratio of the minimum number of steps required by the character (m = 1 for a binary character) to the observed number (s). In Fig. 3e, character 5 occurs only once and hence its $c_i = 1 (m/s = 1/1)$, whereas character 6 shows two steps and thus its $c_i = 0.5 (m/s = 1/2)$. The amount of homoplasy implied by the whole data set can be measured using the ‘ensemble consistency index’ (CI), which is the ratio of the minimum number of steps implied by all characters (M) to the length of a cladogram (S). For the cladogram in Fig. 3e, the $C_l = 0.86$ ($M/S = 6/7$).

There are three perceived problems with the consistency index as a measure of homoplasy. First, although uninformative characters do not add any grouping information to a cladogram, they will inflate its CI. However, this is of significance only when different data sets are being compared. Second, CI can never attain a zero value. A data set in which all possible informative characters occur in equal numbers (an ‘undecisive’ matrix) provides no evidence for preferring one cladogram to any other. Nevertheless, these cladograms will all have positive, nonzero CI values. Third, it has been observed empirically that CI decreases as the number of taxa increases, irrespective of change in the information content of the data. However, this is a recognized and expected property of the CI.

Although the consistency index is useful as a measure of the amount of homoplasy in a character or data set, it is indifferent to the pattern of fit on a cladogram. A binary character that occurs on two separate terminal branches will have the same CI value (0.5) as one that supports two separate groups of taxa. However, in the former case, the character contains no grouping information, while in the latter it is a synapomorphy (albeit homoplastic) for two groups of taxa. The amount of similarity in a character that is interpreted as synapomorphy is measured by the ‘retention index’ (ri). This is defined as $(g - s)/(g - m)$, where $s$ and $m$ are the same variables as for CI, and $g$ is the maximum number of steps that a character can show on any cladogram. For character 5 in Fig. 3e, the minimum and observed number of steps is one, and the maximum number is two. Hence its $r_i = 1 [(g - s)/(g - m) = (2 - 1)/(2 - 1)]$ and all similarity is interpreted as synapomorphy. In contrast, for character 6, the minimum number of steps is one, and the observed and maximum number is two. Hence its $r_i = 0 [(g - s)/(g - m) = (2 - 2)/(2 - 1)]$ and none of the similarity is interpreted as synapomorphy. The method can be extended to the whole data set as the ‘ensemble retention index’ (RI), which uses the summed values of $g$, $s$, and $m$ ($G$, $S$, and $M$, respectively). For the cladogram in Fig. 3e, the $R_l = 0.67 [(G - S)/(G - M) = (9 - 7)/(9 - 6)]$.

B. Character Weighting

The application of differential weights to characters has a long history in systematics. Methods of weighting can be divided into a priori and a posteriori procedures, depending on whether they are applied before or after cladogram construction.

A priori approaches to character weighting generally invoke beliefs that some characters are more important than others or use a particular model of evolution or character change, under which certain types of transformation are considered more or less likely than others. For example, it is common when analyzing nucleotide sequence data to downweight transition substitutions relative to transversions. Alternatively, changes in third codon positions may be disregarded (i.e., accorded zero weight) because they are considered much more likely than changes in first or second positions as a result of the redundancy of the genetic code. Numerous other models have been proposed and they are particularly frequent in the field of molecular systematics. However, such weighting schemes are justifiable only insofar as their underlying model is justifiable.

A posteriori weighting schemes are based on ‘cladistic consistency’ (i.e., the fit of characters to a cladogram) and characters with greater fit are accorded greater weight. One indication of a character's fit is the amount of homoplasy it shows on a cladogram. However, homoplasy does not imply that all similarity is uninformative and the proportion of similarity interpreted as synapomorphy also needs to be taken into account. Hence, both the consistency index (homoplasy) and the retention index (synapomorphy) can be used to estimate character weights. Farris (1989) suggested using the product of these two measures, the ‘rescaled consistency index’ (rc). By combining the $c_i$ and $r_i$ in this way, characters in which none of the similarity is synapomorphic ($r_i = 0$) receive zero weight, irrespective of their level of homoplasy. All other characters, which contain some amount of grouping information ($r_i > 0$), are differentially weighted according to their level of homoplasy. Using this approach, in Fig. 3e character 5 would receive a weight of $1 (c_i = r_i = 1)$, but character 6 would receive a weight of $0 (c_i = 0.5, r_i = 0)$. These weights are applied in a new analysis and the most parsimonious cladogram(s) obtained are used to estimate a new set of weights. This procedure is repeated until a stable set.
of both weights and most parsimonious cladograms is achieved; hence the name "successive approximations character weighting."

The level of homoplasy of a character may also be viewed as the number of extra steps required to fit it to a cladogram. If all extra steps in all characters are considered equal, then a linear fitting function is being applied to their relative cladistic consistency. For example, in Fig. 3e, the single step of character 3 is considered equal to each and either of the two steps of character 6. However, intuitively, we might consider that characters showing fewer extra steps are "better" than those showing more. The former can be assigned higher weights than the latter using a concave fitting function of relative cladistic consistency. This approach was implemented by Goloboff (1993) as "implied weighting," in which the weight \( W \) accorded to a character is calculated as \( W = K/(K + ES) \). \( ES \) is the number of extra steps shown by the character and \( K \) is the "constant of concavity." The value of \( K \) can be varied to weight more or less strongly against those characters with the most extra steps. As \( K \) decreases, these characters will receive progressively lower weights. For example, in Fig. 3e, character 6 will receive a weight of 0.85 when \( K = 6 \), but a weight of only 0.3 when \( K = 1 \) (the "perfect" character 3, which has no extra steps, receives the maximum weight of 1). The optimal cladogram is that for which the summed weights for all characters has the largest value.

C. Consensus Cladograms (Trees)

A cladistic analysis will often produce more than one most parsimonious cladogram as a result of contradictory signal in the data (homoplasy). The agreement between such "fundamental cladograms" (so-called because they are generated directly from the analysis of data) can be conveniently summarized by means of a consensus cladogram (usually referred to as a consensus tree). Several consensus methods have been proposed, of which the most widely used are "strict," "combinable components" (or "semistrict"), "Adams," and "majority-rule."

The strict consensus tree is the most conservative, because it includes only those groups (often referred to as "components") that are common to all the fundamental cladograms. For example, in the two cladograms shown in Fig. 7a and 7b, only groups ABC and DEF occur in both, and thus these are the only groups that appear in the strict consensus tree (Fig. 7d). Groups EF, AB, and BC are excluded because the first is lacking from Fig. 7b (where it is unresolved) and the other two are contradictory.

However, it is possible for a group to be lacking from one or more fundamental cladograms and yet be uncontradicted. For example, group EF in Fig. 7a does not conflict with the cladogram in Fig. 7b because it is one of the three resolutions possible for the group DEF. Combinable components consensus allows such non-replicated, but non-conflicting, groups to be included in the consensus tree, in addition to those groups in common (Fig. 7e). When all fundamental cladograms are fully resolved, with no spurious resolution due to ambiguous optimization, then the strict and combinable components consensus trees will be the same.

A problem with both strict and combinable components consensus is that a single taxon appearing in highly disparate positions on two cladograms is sufficient to collapse all intervening resolution. For example, taxon D in Figs. 7a and 7c appears as the sister-group to two different terminal taxon-pairs. Consequently, the strict consensus tree (Fig. 7d) is relatively unresolved. However, examination of Figs. 7a and 7c shows that taxon D is acting as a "rogue" taxon; that is, apart from its differing positions, the resolution of the remaining taxa is identical in the two cladograms. Such rogue taxa can be identified using Adams consensus analysis. On an Adams consensus tree, taxa in conflicting positions on the fundamental cladograms are placed at the most inclusive node they have in common; in other words, the consensus contains all intersecting sets of taxa common to the fundamental cladograms. However, as a result, it is possible for an Adams consensus tree to contain groups that are not found on any of the fundamental cladograms and thus they need to be interpreted with care. Sometimes taxa such as D would simply be deleted to give the "largest common pruned tree" as a consensus.

When the number of fundamental cladograms is large, strict consensus trees are often very poorly resolved and can be viewed as too restrictive. One method of increasing resolution of the consensus is to retain those groups that occur in a prespecified number of the fundamental cladograms. Typically, such majority-rule consensus trees will comprise those groups that occur in more than 50% of cladograms. The majority-rule consensus tree of the cladograms in Fig. 7a-c, shown in Fig. 7h, is fully resolved, despite the marked topological differences in its fundamental cladograms.

Regardless of their number, the most parsimonious cladograms found by cladistic analysis remain our best estimate of the relationships among the taxa under study. Because resolution is lost, most consensus trees
are less parsimonious than their fundamental cladograms. However, if the topological differences among the fundamental cladograms are due solely to ambiguous optimization, the strict consensus tree will also be of minimum length. Then, the strict consensus tree is also the strictly supported cladogram (Nixon and Carpenter, 1996) and is the preferred most parsimonious topology because it is the only cladogram that is both of minimum length and has all groups supported unambiguously by data.

D. Cladogram and Group Support

A number of statistics attempt to assign levels of support or confidence to the results of cladistic analyses. They can be divided into two categories: methods that address support for an entire cladogram and aim to determine whether there is any "significant" structure in the data, and methods that examine the support afforded to individual groups on a cladogram and attempt to distinguish those groups that are well supported from those that are not.

Methods aimed at assessing support for an entire cladogram all use the same general principle. The length of the most parsimonious cladogram obtained from the observed data set is compared with those derived from a large number of "phylogenetically uninformative" data sets, with the expectation that the former will be substantially shorter than any of the latter. Several definitions of "phylogenetically uninformative" data have
been proposed, including "statistically random" (scores in a data matrix are allocated at random), "undecisive" (all possible informative characters occur in equal numbers), and "randomly co-varying." The last of these forms the basis of the "permutation tail probability" (PTP) test. Covariation here is the degree to which characters are explicable by the same cladogram (i.e., congruence). A most parsimonious cladogram that contains a large amount of homoplasy may be derived from data exhibiting such poor covariation that randomly co-varying characters could produce a cladogram of equal length or shorter. The PTP test uses pseudoreplicate data sets in which character codes are randomly reassigned to taxa, with the restriction that the proportion of each code is maintained, and with each character treated independently. For example, for three taxa (A, B, and C) originally coded 0 and two (D and E) coded 1, permutation may reassign 0 to A, C, and D and 1 to B and E. This procedure is repeated to create a large number of such pseudoreplicates, for which the most parsimonious cladograms are then found. The PTP is defined as the proportion of all data sets (original plus permuted) that yield most parsimonious cladograms at least as short as the original data set and may be interpreted as the probability that a cladogram of this length could have arisen by chance.

The simplest measure of support for a particular group on a cladogram is branch length. However, homoplasy makes the assessment of branch support difficult and groups may appear to be better supported than they actually are. "Bremer support" is a more precise measure of clade support and is defined as the number of extra steps required to lose a clade from the strict consensus tree of near-minimum-length cladograms. When there is no homoplasy in the data, the Bremer support of a group is the same as its branch length. Otherwise, support is reduced to the extent that there are alternative equally parsimonious groupings. To calculate Bremer support, first those cladograms that are one step longer than minimum are found and the strict consensus tree formed from them and the most parsimonious cladogram. The process is repeated, adding a step at a time, until the group in question is lost from the consensus. The number of extra steps required to achieve this is the Bremer support for the group. If more than one most parsimonious cladogram is found initially, then the procedure begins with the consensus tree of these cladograms. Any group that may be a potential resolution of the consensus will have a Bremer support equal to 0.

The bootstrap seeks to estimate group support using pseudoreplicate data sets, which are formed by randomly sampling characters with replacement. The effect is to weight some characters and delete others, with the constraint that the total weight equals the original number of characters. A large number of such pseudoreplicates are generated and their most parsimonious cladograms are found. Conflict among these cladograms is assessed using a majority-rule consensus tree and the support for a group is estimated as the proportion of most parsimonious cladograms on which it is recovered. However, both the PTP test and bootstrap are questionable in a cladistic context. The PTP test's null hypothesis of randomly co-varying characters is contrary to the basis of cladistics. Characters as nested homology statements are intrinsically hierarchical and thus it is inappropriate to measure their performance against randomized characters from which this hierarchy has been removed. The bootstrap assumes that a data set represents a random sample of all possible characters. However, taxonomic characters are generally carefully selected with the aim of resolving the relationships of the taxa under study. Furthermore, it only requires a single synapomorphy to diagnose a clade. However, the random nature of the bootstrap process means that such a character may be represented in only a few pseudoreplicates and thus the group will not appear in the majority-rule consensus tree despite being unconstrued. Thus, bootstrap values are a one-sided test; recovered groups have some measure of support in the data, but groups that are not recovered cannot be rejected.

V. SIMULTANEOUS AND PARTITIONED ANALYSIS

It is generally recognized that data from many sources may be used in a cladistic analysis (e.g., morphological, physiological, behavioral, ecological, or molecular sequences) and that analyses of these data can yield different hypotheses of relationships. Simultaneous analysis (sometimes called a total evidence approach) combines all data, from whatever source, into a single taxon × character matrix for analysis. The resulting hypothesis of relationships is thus determined by character congruence (Fig. 8a). Alternatively, different classes of data may be analyzed separately and the resulting cladograms added together using a consensus method to extract the common phylogenetic signal. This is called the partitioned evidence approach (Fig. 8b) and the result is determined by taxic congruence. The reasoning
Simultaneous and partitioned analysis. In simultaneous analysis, all data are combined into a single matrix before analysis. In partitioned analysis, each data set is analyzed separately and the resulting cladograms are then "added" together using a consensus method.

underlying this approach is that different classes of data may reflect different evolutionary processes and so should be analyzed separately. A third alternative, known as conditional data combination, attempts to distinguish those conditions under which it would be best to keep data sets separate and conduct partitioned evidence analysis from those conditions under which it is more appropriate to conduct a simultaneous analysis. This approach estimates the degree of heterogeneity of phylogenetic signal among data sets, and if the heterogeneity is greater than might be explained by sampling error, then the data sets are analyzed separately.

See Also the Following Articles
CLADOGENESIS • DIVERSITY, TAXONOMIC VERSUS FUNCTIONAL • EVOLUTION, THEORY OF • SYSTEMATICS, OVERVIEW • TAXONOMY, METHODS OF

Bibliography
CLADOGENESIS

Christopher J. Humphries
The Natural History Museum, London

I. Diversification—Patterns and process
II. Anagenesis, cladogenesis, and stasis
III. Clades, grades, fossils, and ancestors
IV. Patterns of life
V. Rates of evolution
VI. Conclusions

GLOSSARY

adaptation Any genetically controlled characteristic that increases an organism’s fitness, usually by ensuring the organism to survive and reproduce in the environment it inhabits.

adaptive radiation The evolution of one or a few forms into many different species that occupy different habitats within new geographical areas or habitats.

anagenesis A pattern of evolutionary change involving the transformation of an entire population, sometimes to a state different enough from the ancestral population to justify renaming it as a separate species; also called phyletic transformation of an unbranched lineage of organisms which changes to such an extent as to be justifiably called new species or taxa. This is considered “biological improvement” (sensu Huxley, 1938), covering all types of change from detailed adaptation to generalized organizational advance.

clad The branch between the nodes on a cladogram or phylogenetic tree, the unit of cladogenesis.

cladogenesis branching evolution; the origin of diverging new forms from an ancestral lineage; the form of divergence and phyletic splitting, including speciation; adaptive radiation of species to major divergence of families and phyla. Evidence for cladogenesis is from those fossil groups that increase in the number of recognizable taxa over time.

genotype The genetic constitution of individuals, species and taxa.

grade The product and “unit” of anagenesis; paraphyletic group.

Linnean system/hierarchy The hierarchical arrangement of inclusive categorical ranks in which, apart from the lowest rank, subordinates members of all lower ranks; it contrasts with exclusive classifications (e.g., scala naturae).

monophyly (monophyletic group) A natural group that includes a most recent common ancestor and all of its descendants.

natural classification Systems of classification portraying as accurately as possible the entire pattern of life and its relationships.

ontogeny The developmental history of an organism from egg to adult. Includes embryogenesis that describes the generative phase and the allometric phase of growth and maturity.

paraphylly A group that includes a most recent common ancestor and only some, not all, of its descendants.
phenotype The characteristics of an individual and all its parts as an interaction between the genotype and the environment.

phylesis Line(s) of descent between ancestors and descendants.

phylogeny The inferred lines of descent showing genealogical relationships of organisms; often used to describe nested series of monophyletic taxa both in cladograms and phylogenetic trees.

polyphyly (polyphyletic group) A group that does not include the most recent common ancestor of all of its members.

punctuated equilibrium The morphological stasis of fossil species over long periods of time punctuated by seemingly instantaneous speciation and morphological change.

Scala naturae (Great Chain of Being) One of the most pervasive ideas in western thought derived from two concepts Plato’s principle of plenitude and the linear series of Aristotle and Plotinus (Panchen, 1992); in plain English, the belief in a linear progression from the simplest forms of life to the most perfect.

stasis (stasigenesis) The persistence of organisms through long geological periods of time that greatly resemble their fossil forebears. The “process” is used to recognize delimitable anagenetic units, or “grades.”

synapomorphy Shared derived characters; homologies of monophyletic groups.

tree of porphyry A tree classification constructed as a dichotomous key that at any rank divides on the basis of differential characters to give two taxa at the rank below. The net result is a comblike branching diagram intended to classify individuals within a general scheme.

THE TERMS ANAGENESIS (divergent evolution), cladogenesis (branching evolution or diversification), stasis (stasigenesis; constancy), grades, and clades are all associated with mechanisms and patterns of phyletic evolution. Anagenesis is evolutionary change in lineages through time, cladogenesis is diversification of clades by branching of lineages through time (Rensch 1939), and stasis (stasis) refers to persistence of lineages through time (Huxley, 1988). The distinctions came about through the belief that in evolution there were three main processes at work that led to “biological improvement,” to “diversification,” and to “ persistence.” Extinction, although seen by some to be of importance in evolution, is perceived as a failure to persist and is now treated as an important part of diversity assessments of life on earth. The concepts of phyletic evolution came to prominence in the neo-Darwinian synthesis of genetics and evolution, but their meanings changed when viewed through the lenses of different systematic perspectives. Grades, for example, once seen as extremely important with respect to the pattern of evolution in the scala naturae, are now seen as artifacts, described as paraphyletic and polyphyletic groups in the literature of cladistics where the emphasis lies in the analysis of clades and the recognition of sister groups. Furthermore, the processes of anagenesis and cladogenesis refer either to microevolutionary processes (speciation) or to patterns and rates of evolution through geological time calculated as rates of morphological, taxic, and molecular change.

I. DIVERSIFICATION—PATTERNS AND PROCESS

The centerpiece of evolutionary theory is that the environment varies with time and from place to place. Heritable variations that suit a particular environment are selected in situ, and so with time populations diverge and differentiate as each becomes adapted to its own conditions (Patterson, 1999). Darwin (1859) considered that such adaptive change led to diversity of form, with features thought to result from the process of diversification. Cetaceans and sea grasses, for example, are thought to be marine-adapted mammals and flowering plants, respectively, with land-borne ancestors. With the impact of tremendous changes in systematic thinking and population biology, Darwin’s theory of organic evolution is now considered by most people to be the only worthy explanation of the diversity and form of life on earth. Darwin’s theory has two quite different and distinct aspects. Evolution by common descent is used to describe the patterns of changes in diversity through time, and evolution by natural selection describes the processes leading to diversity as a measure of change over time.

The language of micro- and macroevolution emerged at a time when fossils were considered as of cardinal importance for understanding the patterns and processes of evolution, when evolutionary theory, which was making great strides over the past 30 years have been made in systematics, where fossils have been analyzed together with recent taxa.
The relationships have changed, and systematics is the key. Progress in evolutionary theory has largely been in mechanisms. The concepts of cladogenesis and anagenesis have consequently changed. For example, as interpreted in the light of phylogenetic systematics (or cladistics) grades pertain to polyphyly and paraphyly, and clades to monophyly. The origin of species or higher taxa by anagenesis, cladogenesis, or phylogenesis, and persistence through stasis, is associated more with process biology in macroevolution (see speciation processes) and particularly the assessment of patterns and rate of macroevolution.

A. The Synthetic Theory

The synthetic theory of natural selection (often described as neo-Darwinism) incorporates the original theory of natural selection of Darwin and Wallace (Darwinism) but also incorporates models of population genetics and the underlying bases of morphological and genetic heterogeneity within organisms. Classic genetic studies on model organisms such as Drosophila, Cepaea, and mimetic butterflies (see Panchen, 1992) have largely been concerned with genetic variation on different alleles in heterogeneous populations and the consideration of adaptive significance, which led to the rather stilling view that all change is adaptive. This has led to the idea that natural selection is essential for adaptive, anagenetic change and is a necessary precondition for cladogenesis (speciation). The processes of anagenesis, cladogenesis, and stasis are used to describe changes (or the lack thereof) in phenotypes and genotypes through changes in the environment both in space and time and by mutations and heritable variations within phylogenetic lineages. Patterns of diversity are described in phylogenetic schemes or genealogical trees through systematic analysis.

B. Speciation

As Huxley (1938) noted, Darwin (1839) recognized three processes at work “leading respectively to biological improvement, to diversification, and to persistence.” Improvement covers adaptation to specialized lifestyles, through phyletic transformation of lineages, resulting in the derivation of descendents species directly from ancestors. These are recognized as “transformational” or “evolutionary” species. New species arise as recognizable transformations within a lineage, by-products of the genetic properties of the lineage, and are thus recognized as arbitrarily divided segments of that lineage. Diversification covers all patterns and processes from micromorphological changes and subspeciation (microevolution), adaptive radiation (cladogenesis with adaptation), and macroevolution, the division into major phyla and higher taxa. Thus, the processes of divergence include lineage splitting, speciation, and adaptive radiation where genetic isolating mechanisms result in two or more descendants derived from a common ancestor. Persistence refers mostly to the idea that organisms of great geological age survive unchanged as living fossils into contemporary biotas. Little or no detectable change takes place through time such that the modern organisms resemble fossils from early geological periods. Examples include animals, such as Latimeria (Coelacanth), Nautilus, and Sphenodon, and plants, such as Ginkgo (Maidenhair tree), Wollomia (Wollemi Pine or “dinosaur tree”), and the Metasequoia (dawn redwood).

II. ANAGENESIS, CLADOGENESIS, AND STASIS

New species can either become modified through gradual change in an entire lineage in response to a changing environment or can emerge through diversification into two or more species through formation of internal and external isolating mechanisms. Anagenesis is gradual change in an entire lineage (Fig. 1). Division into two or more species is termed cladogenesis (Fig. 2). There are many modes of population differentiation and considerable debate as to their roles in speciation. Likewise the mechanisms for anagenesis and cladogenesis (speciation) are many and varied. Nevertheless, they all include some form of population differentiation, either by gradual change of genetic differences and natural selection (phyletic gradualism), or by abrupt punctational changes, involving chromosome inversions or translocations and rapid isolation between populations.
Cladogenesis

Figure 2
Cladogenesis. The transformation of a species or taxon into two (or more) species (or taxa) by branching along a lineage.

or by historical accidents, such as allopatric speciation by vicariance and isolation.

There is considerable controversy among the theories of diversification and how evolution proceeds. The Darwinian hypothesis is phyletic gradualism, whereby the same microevolutionary processes that lead to population differentiation cause ever-increasing divergence between populations (Fig. 3). Eventually, sufficient divergence has occurred for differentiation at species level to be recognized. Differentiation continues at a steady rate and new species originate by slow, gradual changes of ancestral species. The neo-Darwinian perspective is that evolutionary transformation takes place within species, or lineages, and the branching process of cladogenesis accounts for diversification but relatively small amounts of evolutionary change.

An alternative theory of evolutionary rates and speciation, punctuated equilibrium, was proposed by Eldredge and Gould (1972; Fig. 4). Evidence from paleontology on well-preserved fossils indicates long periods of stasis, where "species" remain relatively little changed over long periods of time. At other times there appears to have been rapid evolution and great morphological differentiation. Eldredge and Gould (1972) thus concluded that rates of evolution were not constant over time. They hypothesized that little evolutionary change occurs within species and that genetic changes within populations do not account for different species. Instead, the events of speciation account for evolutionary change, and short periods of "punctuation" were interspersed with little or no evolutionary divergence.

Phyletic gradualism and punctuated equilibrium represent the opposite extremes of a continuum. Recent research, especially on rates of molecular change, would suggest that evolutionary rates are clocklike, or at least "clocky," caused by periods of slow rates of phyleisis and rapid periods of cladogenesis. This is reflected in research of the 1980s and 1990s, which has concentrated on uncovering the patterns of divergence through systematic analysis (particularly cladistics). Sustained application of cladistic methods has determined nature's hierarchy and techniques such as maximum likelihood have been used to assess rates of change amongst phylogenetic trees.

III. CLADES, GRADES, FOSSILS, AND ANCESTORS

Systematic classifications are represented in branching diagrams, variously known as trees, phenograms, dendrograms, and cladograms depending on which systematic philosophy is being used. For most practicing biologists of the 20th century, modern systematics dates back to Linnaeus. He recognized that nature's hierarchy could be arranged into a series of named categories.
His system of categories, although still in use today, is somewhat arbitrary, based on the predefined levels of Regnum (Kingdom), Classis, Ordo, Genus, and Species, to which Phylum and family were later added (Panchen, 1992). All the members of the same level were given the same rank. As distinct from exclusive systems, such as military ranks, biological ranks are all inclusive and genuinely hierarchical, whether viewed agglomeratively from the bottom or divisively from the top, such that eventually all taxa belong to one group, life itself. The hierarchy is divergent, so that a taxon of, say, a specific rank, belongs to a taxon of higher rank and the taxon of this rank belongs to a taxon of even higher rank, and so on. The hierarchy is not necessarily symmetrical, because taxa of high rank can have everything from one isolated member to a great number of members. Elements of Linnean system are still in use today, especially nomenclature, but its avowedly artificial category hierarchy, expressed as a preordained ranking system, has become replaced by a whole succession of “natural” classification systems portraying as accurately as possible the entire pattern of life and its relationships.

A. Natural Classification

Almost a century of debate has questioned the meaning of “natural” classification systems but by the time that the Origin was published (Darwin, 1859), the form of representation had become a branching diagram of hierarchical clusters. To his credit, Darwin made the tree metaphor explicit in what was the only diagram published in the Origin (Fig. 5). This was not a diagram produced by clear-cut rules but a tree rooted in the past—a genealogical tree to denote phylogeny in a vertical direction and classification of relative distances of present-day organisms in the branching hierarchy. One of the most important issues that came from Darwin’s

**FIGURE 5** Darwin’s (1839) branching diagram from the Origin. Note that A–L represent the species of a genus. The small letters show degrees of divergence through time. The horizontal lines represent 1000 or more generations. The diverging lineages are shown with extinct populations at most generations in every lineage.
evolution by common descent was that phylogeny is the theory that underpins natural classification today. Fairly soon after the Origin appeared, "forests" of genealogical trees began to appear in the European literature. Perhaps the most memorable rendition of all schemes ever published is that of Haeckel (1874) who synthesized his interpretation of phylogeny and classification and the origin of man into a natural classification using Darwin’s tree metaphor (Fig. 6). Apart from its shape, the named parts of the tree embody two quite separate concepts. The shape is simple enough in that it depicts a multifurcating inclusive genealogy with the names of the branches indicating monophyletic groups (groups derived from a common ancestor). However, the names on the vertical axis, horizontal lines or “grades” of evolutionary achievement, embody a watered down version of the scala naturae, a belief in the progression from the lowest simplest forms of life (Monera) to the most “perfect” (human) through the series protista—metazoa—vertebrata—vertebrata—mammalia. The distinction was based on Haeckel’s twin concepts of the evolution of the foetus, ontogeny, and recapitulation in the evolution of the stem (phylogenesis). Arising out of former was the phylogenetic interpretation of the Tree of Porphyry as a step series (Stufenreihe) and distinguished from the ancestor series (Ahnenreihe) in palaeobiology (see Panchen, 1992).

By the turn of the 20th century, branching diagrams were largely depicted as phylogenetic trees to represent “Ahnenreihe” as “grades” in ancestor-descendant series and “Stufenreihe” to represent evolution by cladogenesis into a succession of adaptive zones (Panchen, 1992; Fig. 7). There have been endless variations of this theme in comparative biology and different emphases placed on grades and clades. For example, Bessey’s cactus (1913), deliberately emphasized grades and even interpreted extant supraspecific taxa, Ranales (Fig. 8a), as giving rise to other supraspecific taxa, Rosales (Fig. 8b), which in turn gave rise to Myrtales (Fig. 8c). At the other extreme, Janchen’s (1932) interpretation of Wettstein’s classification clearly emphasizes clades with grades added almost as an afterthought to indicate the distribution of floral evolution through gymnospermiae/angiospermiae, apetala/sympetala, and monocotyledony/dicotyledony (Fig. 9).

**B. Grades and Clades in Evolutionary Systematics**

These days we know that Bessey’s approach is untenable—one supra-specific taxon cannot give rise to another. The Ranales cannot be the ancestor of the Rosales, and even if we did consider them to have an ancestor-descendant relationship then the Ranales and Rosales would have to be classified in the same group rather than as depicted in Figure 8. Nevertheless, grades and clades are still established practice and appear in textbooks even today. The 20th century has seen an enormous output of different phylogenetic schemes and the persistence of grades has prevailed the methods of “evolutionary systematics.” The idea that grades are important in systematics so as to express patterns of anageneisis is the hardest notion to expunge. Grades are characteristic of classifications involving “satellite” groups considered sufficiently divergent to separate them at a high rank from their nearest relatives. The more “primitive” taxa are similarly rendered as a grade, a paraphyletic group in the language of phylogenetic systematics or cladistics.

To illustrate, perhaps the two most famous examples include human (Homo sapiens) considered superior to its primate relatives, and birds (Aves) considered greatly diverged from relatives in the class Reptilia (Fig. 10; see Mayr, 1974; Panchen 1992). The net effect of such schemes is that the phylogenetic reconstruction, based on character analysis, differs greatly from the written classification. Figure 10 illustrates this result in both examples. Thus for the diapsid reptiles the phylogenetic diagram shows ancestor A ancestral to two groups B (Diapsid reptiles) and C + D (crocodiles + birds). The classification differs in that group D (Aves) shows such considerable divergence along the horizontal character axis (the anagenetic component) that to express the information in a classification puts B + C together (diapsid reptiles + crocodiles) and D is kept separate (Aves). A similar story can be portrayed for orangutan + gorillas/chimps and Homo sapiens. The phylogenetic diagram shows C + D (gorillas/chimps and humans) as sister group to B (orangutan) but the classification puts B and C together and humans in a group D of their own. Mayr (1974) claimed that the anagenetic component should carry more weight because of many genes shared in common among birds and not found in the nearest relatives and because of the unique behavioral characters in man. Thus, groups represented by D are placed in higher grades and groups represented by B + C in lower grades.

**C. Paraphyly and Monophyly in Cladistics**

Cladistics demonstrates that grades are impossible to characterize as they do not express relationships between organisms. There is no way of knowing when
FIGURE 6: Haeckel’s (1874) "literal tree" (European oak) depicting human ancestry as interpreted by Darwin’s theory of evolution by common descent. The bifurcating branches indicate groups determined on embryology and the vertical axis (conflated with time) represents grades.
one grade starts and another stops. How many characters along a branch are required before it is possible to draw a line and say that one side represents a lower grade and the other side a higher grade? Related to this question is how many taxa belong to a grade group?

**FIGURE 7** Abel’s concept of Stufenreihe (clades A–D) and Ahnenreihe (grades a–d) (After Panchen, 1992; Fig. 3.6, p. 55). See text for explanation.

Hennig (1996) described a method to implement evolution by common descent by reconstructing phylogenies based on assessments of speciation (cladogenesis) and transformation of characters (now known as cladistics; see Methods of Systematics, and Cladistics). His most important contribution was to offer a precise definition of relationship and a technique for those relations to be discovered. A minimum of three taxa is necessary to express a relationship. For example, in Figure 11a taxa B and C are more closely related to each other, than either is to A, because they share a common ancestor not shared by A. Cladistic analysis finds monophyletic groups on the basis of uniquely derived, shared characters (synapomorphies).

Hennig showed that monophyletic groups are “natural groups,” that branching nodes expressed relations, and that synapomorphies are the only measurable quantities for determining pattern. Hence neither horizontal nor vertical branches (anagenesis) are meaningful for expressing relationships on cladograms or for the determination of groups. Vertical branches say nothing about time and the relative nesting of nodes on the cladogram provide only relative rather than absolute estimates of ordinal time. Relationships could just as easily be represented by nested sets or Venn diagrams (Fig. 11b).

Cladograms are synapomorphy schemes, induced from the most parsimonious distribution of characters to show sister-group relationships. In Figure 11a, taxa B and C represent one sister group nested in larger sister group, A and B + C. Cladograms are different from phylogenetic trees because they rely entirely on empirical data, taxa, and characters. They express only the general branching pattern of life because that is all that is available from analysis of taxa and form. The strongest support for this idea is that many phylogenetic trees can be hypothesized for the same cladogram irrespective of branch length (Fig. 12 b–c). Consequently, anagenesis, ancestors, and ancestor-descendant relationships are not directly available from character analysis but require models of one kind or another to arrive at answers to questions of rates of divergence. Cladograms are different from phylogenies or phylogenetic trees because they are isomorphic with the classification. Cladograms are consistent with name hierarchy and they can be recovered from written classifications. On the basis of this property, Hennig (1966) justifiably claimed that phylogenetic systematics provided the only truly general reference system consistent with the theory of evolution by common descent.

The task of systematics for Hennig was to understand natural relationships (monophyly, monophyletic groups) and rid the general reference system of polyphyly and paraphyly. Although there has been consis-
erable proliferation in the methods and sources of information in systematics, the main effect has been to put intense effort into cladogenesis by the discovery of clades or monophyletic groups. For the past 25 years or so systematics has concentrated on determining the pattern of life from its earliest beginnings to the highest nodes on the tree, especially as a result of massive strides in molecular biology. Programs in pattern analysis range from the minutiae of phylogeography within species and populations to the discovery of monophyletic clades throughout the entire history of life. On the process side have been intense efforts to discern the rates of macroevolution by calibrating what is known about fossil history and morphological evolution with what is known about base substitution rates in ubiquitous molecules.
IV. PATTERNS OF LIFE

The emphasis over the past 30 years has switched from worrying so much about whether one species or taxon gave rise to another species or taxon to emphasizing phylogenetic trees and cladistic patterns. The emphasis has switched entirely from ancestor-descendant stories to expressions of sister-group relationships. The emphasis on the overall morphology and anatomy of the phenotype has comparably given way to the use of molecular characters. Furthermore, the discovery that sequences of amino acids of protein chains and nucleic acids contain detailed sources of information has given impetus to the discovery of family trees across the entire hierarchy of life. By comparing homologous amino acid or gene sequences it has been possible to compile family trees using computerized cladistic analyses with thousands of taxa in one calculation. Through a massive empirical enterprise, involving literally thousands of published investigations, the stage has been reached when comparisons between bacteria, flowering plants, and mammals in the same analysis are a reality. The potential for observing the patterns on both a grand scale and at species level is dominating comparative biology in all disciplines. The output of data and cladograms is so vast and the information coming from a variety of museums and laboratories around the world that dedicated web sites are being built to makes sense of it all (see, for example, The Tree of Life web page at http://phylogeny.arizona.edu/tree/phylogeny.html).

The overriding message from all of this empiricism is that large-scale differences between humans, mice, elephants, and oak trees have the same sort of causes as closely related species (Patterson, 1999). It appears that at all scales of investigation the differences between organisms related by common ancestry owe their differences to divergence, the accumulation of mutations over great periods of time. The evidence for this homology comes from the fact that organisms share homologous features at many levels, molecular, structural, and physiological. At face value it would seem that gradual phylesis would explain changes in both micro- and macro-evolutionary events. Several evolutionists have different opinions about this. They can accept point mutations, gene duplications, and chromosome mutations to explain the smaller divergences in birds and butterflies but require other mechanisms to suggest the massive differences between fishes and land organisms or the original appearance of crustacea and green plants. The main reason for requiring macromutations for macro-evolution is that many features of major groups cannot be imagined to have come into being by small adaptive changes. The point often made is the existence of complex organs. Familiar questions are of what use is half an eye or half a lung? The answers are not readily apparent because the evidence is unclear. The intermediate steps cannot be imagined. By improving the accuracy of reconstruction of cladograms for the whole of life the potential exists for assessing whether the story is one of gradual phylesis or not.

V. RATES OF EVOLUTION

Rates of evolution have always been of interest to systematists and evolutionists alike, and one of the popular
CLADOGENESIS

myths is that the fossil record is the only direct evidence to track the course of evolution. During the history of paleontology many arguments have been constructed to link the fossil record to evolution because we know that fossils are the remains of once living organisms and they predate the rock sediments in which they lie (Panchen, 1992). Furthermore, the idea that the stratigraphic record encapsulates the historical record, and that most fossils are extinct but related to modern groups that extend a little way back in time, is the underlying justification for schemes to correlate the pattern of fossil record with evolution through time. It has long been considered possible to calculate different rates for a number of reasons. “Living fossils” that have survived from the geological past to the present-day show that most organisms in the fossil record were once living but are now extinct. Similarly, “Lazarus” taxa, found in an early geological period, which then disappear for millions of years to reappear in more recent periods, were considered to be a good line of evidence to support phylogenetic reconstructions with absolute, rather than relative time, intervals. The fact that many modern groups are modified older groups and can be traced back in time has reinforced evidence for the transmutation of fossil organisms into modern forms and hence the belief that rates of evolution can be calculated.

A. Morphological Evolution

Perhaps the most cited examples of evolution between fossils and modern taxa are the so-called missing links. Fossils, such as Archaeopteryx, were considered to show links between different major groups. The apparent intermediate nature of Archaeopteryx by having teeth, as well as derived features of feathers and wishbones, as found in all birds, was used to suggest that birds are descendants of dinosaur-like creatures. No exact sequence of fossils or ancestors could ever be found and it has been known for a long time that no particular group of fossils could give rise to a particular group of modern taxa. The search for ancestors has gradually switched from reconstructing history sequences of grades in a scala naturae to the modern approach of comparing dated fossils in stratigraphic sequences with well-corroborated cladograms (Smith, 1994).

There have been many reconstructions of rates of evolution through geological time in fossil and modern taxa of plants, vertebrates, and invertebrates (see Smith, 1991, 1993). In the case of horses, the technique was to see how well morphoclines of characters from the most primitive characters in the oldest taxa gradually changed into the derived characters of modern taxa. These were then examined to see how well these matched the chronocline (the age of fossils as determined by their appearance in the stratigraphic sequence). Simpson’s (1951) pioneering work traced lines of descent from the fossil Eohippus to Equus by creating a linked time series, or scala naturae, to measure the rates of morphological changes in skull, forelimb, and upper molar characters in a time period of 60 million years. The approach used was to literally measure the amount of change over a generalized phylogenetic tree after dates were assigned to the horse genera in the sequence (Fig. 13). Despite the great effort expended by Simpson, his results give only general approximations on the rates of divergence and the absolute dates.

FIGURE 13 Evolution of fossil horses. (A) represents a phylogram of increasing size, (B) changes in cheek teeth characters, (C) represents number of toes and foot mechanisms. After Simpson (1953, Fig. 34, p. 265).

\[
\text{FIGURE 13 Evolution of fossil horses. (A) represents a phylogram of increasing size, (B) changes in cheek teeth characters, (C) represents number of toes and foot mechanisms. After Simpson (1953, Fig. 34, p. 265).}
\]
are hard to accept because of a lack of a well-corroborated cladogram.

B. Rates of Molecular Evolution—Hemoglobin

The flood of molecular data accrued since the mid-1960s offers the prospect of more accurate assessments in the calculation of rates of evolution. Absolute rates of molecular evolution have been estimated from branching diagrams by calculating trees from various forms of sequence data and alignment of the cladograms to fossil or biogeographic events as fixed starting points. By calculating pairwise distances, the minimum ages of cladogenetic events across pairs of taxa can be estimated. These calculations assume that a ‘clock’ is operating within the molecules by gradual and regular substitution of nucleotides or amino acids constantly over time and that the clock can be calibrated to external events (e.g., the age of a fossil), which then reflect accurately the initial cladogenetic events.

Early estimates in the 1960s were made on chains of amino acids and earliest occurrence of fossils to calculate rates of divergence in proteins. Typical results for alpha haemoglobins (Table I) comparing recent organisms with their fossil relatives suggest that sharks, coelacanths, and lungfishes have hardly changed over 350 million years or more but mammals and birds have transformed totally from their fishy ancestors (Patterson, 1999). Patterson showed these calculations indicate that hemoglobin in humans and elephants differ by about 18% as does that of the starling and ostrich.

Also, humans and elephants diverged from the ancestral mammals at roughly the same time that ostrich and starling diverged from ancestral birds. In terms of amino acid differences, crocodiles differ from birds by about 33% and mammals by about 35%.

From the rough phylogenetic tree (Fig. 14), Patterson calculated from the distance data with a corrected time scale relative rates of amino acid substitution (Table I). Because there were more than 60% differences in the hemoglobin of sharks, as compared with humans, Patterson corrected for the error of multiple hits, because logically the real number of differences between the different taxa must be greater than those observed. The adjusted divergence times gave 340 myrs for the Platypus/mammal and crocodile/bird splits and about 70 myrs for the elephant/human and ostrich/starling divergence times.

It was from these kinds of results that questions of how natural selection could keep constant rates of divergence through time were asked. One such theory was the Red Queen hypothesis. It takes its name from Alice in Wonderland because every species has to keep on running just to keep in the same place. In other words, to compete and to survive all of nature’s vicissitudes during the evolutionary changes from shark to human, and between any other pair of species, it is necessary for hemoglobin, and presumably all other features, to adapt at a constant rate (Patterson, 1999).

C. Rates of Molecular Evolution—DNA

With the advent of nuclear and plastid sequences in the late 1970s and a routine availability of sequences

<table>
<thead>
<tr>
<th>Human</th>
<th>Elephant</th>
<th>Platypus</th>
<th>Ostrich</th>
<th>Starling</th>
<th>Crocodile</th>
<th>Lungfish</th>
<th>Coelacanth</th>
<th>Goldfish</th>
<th>Shark</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>18</td>
<td>28</td>
<td>30</td>
<td>29</td>
<td>33</td>
<td>59</td>
<td>49</td>
<td>48</td>
<td>50</td>
</tr>
<tr>
<td>18</td>
<td>—</td>
<td>43</td>
<td>52</td>
<td>48</td>
<td>48</td>
<td>64</td>
<td>52</td>
<td>48</td>
<td>47</td>
</tr>
<tr>
<td>28</td>
<td>43</td>
<td>—</td>
<td>52</td>
<td>45</td>
<td>45</td>
<td>34</td>
<td>51</td>
<td>36</td>
<td>37</td>
</tr>
<tr>
<td>30</td>
<td>52</td>
<td>52</td>
<td>—</td>
<td>56</td>
<td>53</td>
<td>26</td>
<td>91</td>
<td>56</td>
<td>33</td>
</tr>
<tr>
<td>30</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>—</td>
<td>53</td>
<td>59</td>
<td>91</td>
<td>57</td>
<td>52</td>
</tr>
<tr>
<td>30</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>—</td>
<td>26</td>
<td>56</td>
<td>56</td>
<td>26</td>
</tr>
<tr>
<td>30</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>—</td>
<td>26</td>
<td>56</td>
<td>26</td>
</tr>
<tr>
<td>30</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>—</td>
<td>26</td>
<td>52</td>
</tr>
<tr>
<td>30</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>—</td>
<td>32</td>
</tr>
<tr>
<td>30</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>—</td>
</tr>
</tbody>
</table>

The alpha chain contains 141 amino acids in most vertebrates, but it contains 142 in the goldfish and the lungfish has 143. Numbers to the upper right half are differences and those in the lower left are percentages. (From Patterson, 1999, Table 9.1.)

TABLE I Differences (Distances) in the Alpha Hemoglobin Chains of a Selection of Vertebrates
becoming available by the 1990s, it has become increasingly popular to infer divergence time estimates based on nucleotide sequences directly. Through a series of trials and errors it had become obvious by the mid-1990s that single point estimates of divergence times without any assessments of standard error were hardly worth calculating, and some greater degree of sophistication was needed (Waddell et al., 1999). Recent calculations have attempted to improve results by incorporating differential substitution rates for molecular sequence change, a wider range of reliable calibration points based on fossils and other sources of information, and incorporation of standard error values to factors that can be quantified. Waddell et al. (1999) applied these techniques to estimate rates using maximum likelihood models and the origins of major lineages of mammals and birds using all available published mitochondrial DNA protein sequences (including unpublished ones for elephants and birds (Figs. 15 and 16).

For mammals (Fig. 15) the results show that the well-studied horse/rhinoceros split diverged about 55 million years ago. The splitting time among carnivores is confidently shown to be more than 50 million years, and splits within the placentals are relatively old, in the case of armadillo and other mammals 169 mybp but within 2 standard errors of other estimates. Waddell et al. reported also that the whale/cow divergence at 65 mybp may be much older than previously assumed. One of the interesting findings is that the sampled splits between the main groups of ferocarnivores (carnivores, cetartiodactyls, perissodactyls, and pholidotes) took place prior to the critical Cretaceous/Tertiary boundary. The date of the vicariant separation of Africa and South America is coincident with the age of the cladogenetic event of the armadillo/elephant close relationship at around 122 mybp and could be a causal factor.

The data for birds (Fig. 16) was considered more controversial by Waddell et al. than for mammals. This was because the deeper calibration points normally used by other researchers between either crocodiles or mammals and birds were considered controversial in their assumptions that rates of evolution in birds had not changed much. They used the fossil anseriform, Probionyx, and placed it with the ducks (geese and swans) using a constraint of anseriform monophyly to have
at least one calibration point within birds. Using this information together with more detailed studies of anseriform relationships gave an anseriform stem lineage date of 58–78 (midpoint 68) mybp. The divergence times are indicated in Figure 16. The interesting findings are that the passerines seem to have split off from the rest of the birds at a much earlier date than normally believed and that the divergence time of ratites (rhea and ostrich) is much younger than generally believed.

IV. CONCLUSIONS

In an online *Encyclopaedia of Genetics*, Ernst Mayr recently considered cladogenesis and anagenesis as the two great phylogenetic processes of biology. He said that cladogenesis is the study of the origin and of the nature of the branching pattern of the phylogenetic tree. It deals with the various different methods by which the phylogenetic tree is reconstructed. Also, it includes the process of speciation, because every act of speciation adds a branch, no matter how short, to the phylogenetic tree. It is difficult to argue with such points of view because the definitions encapsulate so much of what we know and do in the modern synthesis. They are catchall statements with little or no precision in the contemporary scheme of things. The position taken here is that cladogenesis, and all of its related terminology, sits in some kind of neo-Darwinian time warp, which has become overtaken by huge developments in both systematics and evolutionary analysis. As Mayr points out, cladogenesis in a very particular sense refers to microevolution and speciation, both topics considered elsewhere in this encyclopedia. In cladistics all effort is concentrated on the patterns of clade distribution and it is generally accepted that grades and branch lengths are of no value in phylogenetic classification but are of interest for determining rates of change. Cladogenesis, anagenesis, and stasis are general descriptors of studies of rates in evolution and extinction, as expressed here in relatively recent calculations of ages of taxa and deeper evolutionary splits. In conclusion, the importance of cladogenesis and related terms...
have become uncritical in the light of relatively recent studies in systematics; rather they are words in an arcane language largely consigned to the history books.

See Also the Following Articles

ADAPTATION • ADAPTIVE RADIATION • CLADISTICS • EVOLUTION, THEORY OF • FOSSIL RECORD • PHYLOGENY • SYSTEMATICS, OVERVIEW • TAXONOMY, METHODS OF

BIBLIOGRAPHY


CLIMATE CHANGE
AND ECOLOGY,
SYNERGISM OF¹

Stephen H. Schneider* and Terry L. Root¹
*Stanford University; †University of Michigan

I. Synergisms of Climate Change and Ecology
II. Climate History: What Has Happened
III. Forces of Climate Change
IV. Climate Change Projections: What Will Happen?
V. Relevance of Climate Modeling to Regional Climate Change and Ecosystem Studies
VI. Examples of Ecological Responses to Climate Changes
VII. Three-way Linkages and Community Ecology
VIII. Climate Forecasts, Ecosystem Responses, and Synergistic Effects

GLOSSARY

climatic models Mathematical descriptions of the flows of energy, momentum, and materials around the atmosphere and oceans, typically run on large computers to simulate the earth’s climate naturally and when disturbed by greenhouse gases or other so-called forcings or disturbances.
global change disturbances Alterations or disturbances to the natural conditions of the global atmosphere (e.g., greenhouse gases produced by human activities) or to enough regions (e.g., largescale deforestation or the wide-spread introduction of exotic, invasive species in many places throughout the world) to have global impacts.
global warming The additional heating of the earth’s climate from the incremental injection of greenhouse gases to the atmosphere from human activities such as deforestation or fossil fuel consumption.
greenhouse gases Gasses in the atmosphere which selectively let more of the solar energy into and out of the atmosphere than they permit the transmission of long-wave infrared radiation. They contribute to the greenhouse effect, which traps radiant energy in the lower atmosphere and makes the earth warmer.
synergisms Interactions among many factors, which collectively may have a much larger or smaller effect than the sum of the effects of each factor acting independently.

¹Adapted from Schneider and Root (1998).

CLIMATE HAS VARIED SUBSTANTIALLY on geological timescales, and ecological systems have responded with large shifts in vegetation and animal patterns, including extinctions. Changes of about 5°C over many thousands of years have occurred and triggered the ecological responses. Human activities which clear land or burn fossil fuels inject greenhouse gases such as carbon dioxide (CO₂) or methane (CH₄) into the atmosphere. There has been a 30% increase in CO₂ and a 150% increase in CH₄ since the industrial revolution. Climatic theory—supported by climatic models run on supercomputers, suggests that several degrees Celsius of global warming is possible—even likely—during the
next century. Ecologists are concerned that this could substantially rearrange the ranges and interactions of many species. However, because of human land uses such as agriculture, urban settlements, or roads, species no longer have a free range in responding (e.g., by migrating) to natural climatic shifts. The combination of effects—synergism—among climate changes, land-use disturbances, the introduction of exotic species and artificial chemicals may collectively impact on wildlife and terrestrial systems much more significantly than if each of these disturbances were simply considered separately. Therefore, many scientists and others have serious concerns about the need to conserve wildlife and vegetation in the face of such plausible global change disturbances.

I. SYNERGISMS OF CLIMATE CHANGE AND ECOLOGY

The earth’s climate is vastly different now from what it was 100 million years ago when dinosaurs roamed the planet and tropical plants thrived closer to the poles. It is different from what it was 20,000 years ago when ice sheets covered much of the Northern Hemisphere. Although the earth’s climate will surely continue to change, climatic changes in the distant past were driven by natural causes, such as variations in the earth’s orbit or the carbon dioxide (CO₂) content of the atmosphere. Future climatic changes, however, will probably have another source as well—human activities. Humans cannot directly rival the power of natural forces driving the climate, for example, the immense energy input to the earth from the sun that powers the climate. We can, however, indirectly alter the natural flows of energy enough to create significant climatic changes. The best-known way people could inadvertently modify climate is by enhancing the natural capacity of the atmosphere to trap radiant heat near the earth’s surface—the so-called greenhouse effect. This natural phenomenon allows solar energy to reach the earth’s surface and warm the climate. Gases such as water vapor and CO₂, however, trap a much larger fraction of long wavelength radiant energy called terrestrial infrared radiation near the earth’s surface. This causes the natural greenhouse effect to be responsible for approximately 33°C (60°F) of surface warming. Thus, seemingly small human-induced changes to the natural greenhouse effect are typically projected to result in a global warming of 1–5°C in the next century (Intergovernmental Panel on Climate Change, 1990, 1996a). This could result in an ecologically significant change, which is why climatic considerations are fundamental in the discussion of the status and trends of ecological conditions.

We may already be feeling the climatic effects of having polluted the atmosphere with gases such as CO₂. Many activities associated with human economic development have changed our physical and chemical environment in ways that modify natural resources. When these changes, such as burning fossil fuels that release CO₂ or using land for agriculture or urbanization that causes deforestation, become large enough, significant global (worldwide) changes are expected. Such modifications can disturb the natural flows of energy in Earth systems and thus can force climatic changes. These disturbances are also known as global change forcings. Quantitative evaluations of the potential effect of human activities in creating global change are needed. Such evaluations are also central to potential policy responses to mitigate global changes (Schneider, 1997; Intergovernmental Panel on Climate Change, 1996c).

A. Synergisms

One of the most potentially serious global change problems is the synergistic or combined effects of habitat fragmentation and climate change. People fragment natural habitats for farmland, settlements, mines, or other developmental activities. If climate changes, individual species of plants and animals will be forced to adjust if they can, as they have in the past. It seems unlikely that all of the migrating species that survived the Ice Age would be able to safely reach refuges after migrating across freeways, agricultural zones, industrial parks, military bases, and cities of the twenty-first century. An additional complication arises with the imposition of the direct effects of changes in CO₂, which can change terrestrial and marine primary productivity as well as alter the competitive relations among photosynthesizing organisms.

One representative instance of synergism is that of the Kirtland’s warbler in northern Michigan; this species is restricted to a narrow area of jack pines that grow in sandy soil (Botkin et al., 1991). Forest gap models of growth and decline of jack pines indicate that this species will move north with warming, but the Kirtland’s warbler will not likely survive the transition. This bird nests on the ground under relatively young pines, and the soil to the north is not generally sandy enough to allow sufficient drainage for successful fledging of young (Cohn, 1989). Consequently, global warning could well doom the warbler to extinction in 30–60 years. This potential for extinction indicates how the already high rate of extinctions throughout the world would be exacerbated by climatic changes occurring more rapidly than species can adapt (Pimm, 1991; Peters and Lovejoy, 1992; Wilson, 1992).
The synergism question raises a controversial management problem of anticipating global change risk and responding by setting up interconnected nature reserves to ensure against some species becoming extinct in the event of climate changes. Alternatively, we could simply let the remnants of relatively immobile wildlife and natural plant communities remain in isolated reserves and parks as now exist. If we do opt for more environmental safeguards by interconnecting our parks, the question then becomes how do we interconnect the nature reserves. Priorities must be set and money made available for constructing natural corridors through which species can travel. For example, elevated sections of highways may be needed to allow for migration routes, similar to what was done for the caribou in the Arctic when the Alaskan pipeline was built. In order to examine such questions as these in scientific detail, it is first necessary to make a multi-disciplinary examination of the sub-components of the various aspects of climatology and ecology. We begin with a background discussion of climatic history, processes, modeling, and validation, as a prelude to focusing on ecological processes, which need to be examined in order to project possible synergisms among ecology and climate change.

II. CLIMATE HISTORY: WHAT HAS HAPPENED

Scientists can reconstruct the cyclical expansion and contraction of polar caps and other ice masses from ice core samples taken from Greenland and Antarctica. When snow falls on high, cold glaciers the air trapped between snow grains is eventually transformed into air bubbles as the snow is compressed into ice from the weight of subsequent accumulations. The ratio of two oxygen molecules with different molecular weights (O\textsuperscript{16} and O\textsuperscript{18} isotopes) is a proxy record for the temperature conditions that existed when the snow was deposited. From this, scientists have been able to determine that the ice buildup from 90,000 to 20,000 years ago was quite variable and was followed by a (geologically speaking) fairly rapid 10,000-year transition to the (current) climatically very stable Holocene Period. The Holocene is the 10,000-year interglacial period in which human civilization developed and modern plant and animal distributions evolved to their current states (Eddy and Oeschger, 1993). These ice cores also provide information on the presence of CO\textsubscript{2}, an important greenhouse effect gas. Carbon dioxide was in much lower concentrations during cold periods than in interglacials (which is similar for the greenhouse gas meth-
with the principles inculcated in this volume, these forms will not have been liable to much modification.

If this were true, the principal ecological concern over the prospect of future climate change would be that human land-use patterns might block what had previously been the free-ranging movement of natural communities in response to climate change. The Cooperative Holocene Mapping Project, however, incorporated multiple pollen types into its analyses, including not only boreal species but also herbs and more arid (xeric) species as well as oaks and other mesic species. They discovered that during the transition from the last Ice Age to the present interglacial, nearly all species moved north, as expected. During a significant portion of the transition period, however, the distribution of pollen types provided no analog associations to today's vegetation communities (Ovtepeck et al., 1992). That is, whereas all species moved, they did so individual by individual, not as groups. Consequently, the groupings of species during the transition period were often dissimilar to those present today (Fig. 1). The relevance of this conclusion is that in the future ecotypes will not necessarily move as a unit as climate changes (assuming there is enough time and space for such a migration). Past vegetation response to climatic change at an average rate of 1°C per millennium indicates that credible predictions of vegetation changes from comparable or even more rapid climatic changes projected from human activities cannot neglect transient (i.e., time-evolving) dynamics of the ecological system. Furthermore, caution should be exercised against relying too much on past conditions in forecasting future patterns resulting from global change forcings because the forecasted global average rate of temperature increase exceeds those rates typical of the last 120,000 years.

Future climates may not only be quite different from more recent previous climates but may also be different from those inferred from paleoclimatic data and from those to which some existing species are evolutionarily adapted. Therefore, possible future changes inferred from past changes can be taken only as a guide or a means to verify aspects of the forecasts of models of climate or ecosystem dynamics (Crowley, 1993; Schneider, 1993a). Such verification exercises may provide more credible forecasts of the effects of climatic change on animals.

III. FORCES OF CLIMATE CHANGE

The two basic categories of causes of climatic change are external and internal. However, these terms are defined relative to the focus of study: stating which components are external or internal to the climatic system depends on the time period and spatial scale being examined, as well as the phenomena being considered. External causes of climate change do not have to be physically external to the earth (such as the sun) but do occur outside of the climatic system. If our focus is on atmospheric change on a 1-week time-scale (i.e., the weather), the oceans, land surfaces, biota, and human activities that produce CO₂ are all external (i.e. they are not influenced much by the atmosphere in such a short time). If our focus is on 100,000-year ice age interglacial cycles, however, the oceans, ice sheets and biota are all part of the internal climatic system and vary as an integral part of the earth's environmental systems. On this longer scale we must also include as part of our internal system the "solid" earth, which really is not solid but viscous and elastic.

Fluctuations in heat radiated by the sun—perhaps related to varying sunspots—are external to the climatic system. Influences of the gravitational tugs of other planets on the earth's orbit are also external. Human-caused changes in the earth's climate could not perceptibly alter either one of these cycles.

Carbon dioxide and methane levels rise and fall with ice age cycles, which are clearly internal on a 10,000-year time-scale. However, on a 20-year scale these greenhouse gases become largely an external cause of climatic change, because small changes in climate have little feedback effect on, for example, humans burning fossil fuel.

Changes in the character of the land surface, such as those caused by human activities, are largely external. If vegetation cover changes because of climatic change, however, land surface change then becomes internal because changes in plant cover can influence the climate by changing albedo (reflectivity to sunlight), evapotranspiration, surface roughness, and relative humidity (Henderson-Sellers et al., 1993).

Snow and ice are important factors in climatic change because they have higher albedo (reflectivity) than warmer surfaces and, in the instance of sea ice, can inhibit transfer of heat and moisture between air and wet surfaces. Salinity, which affects changes in sea ice and in the density of seawater (which helps control where ocean waters sink), may also be an internal cause of climatic variation. The sinking and upwelling of ocean waters are biologically significant because the upwelling waters are often nutrient-rich.

Unusual patterns of ocean surface temperature, such as the El Nino, demonstrate the importance of internally caused climatic fluctuations because the atmospheric circulation can change simultaneously with ocean sur-
face temperatures. When the atmosphere rubs on the ocean, the ocean responds by modifying its motions and temperature pattern, which forces the atmosphere to adjust, which changes the winds, which changes the way the atmosphere rubs on the ocean, and so forth (Trenberth, 1992). As a result, air and water interact internally in this coupled system like blobs of gelatin of different size and stiffness connected by elastic bands or springs, all interacting with one another while also being pushed from the outside (by solar, volcanic, or human-caused change).

IV. CLIMATE CHANGE PROJECTIONS: WHAT WILL HAPPEN?

To predict the ecologically significant ways the climate might change, one must specify what people do that
modifies how energy is exchanged among the atmosphere, land surface, and space because such energy flows are the driving forces behind climate. Air pollution is an example of such a so-called societal forcing of the climate system. Estimating societal forcing involves forecasting a plausible set of human (or societal) activities affecting pollution during the next century. The next step is to estimate the response of the various components of the earth system to such societal forcings. The earth system consists of the following interacting subcomponents: atmosphere, oceans, cryosphere (snow, seasonal ice, and glaciers), and land-surface (biota and soils) systems.

Research in the field and in laboratories provides an understanding about various processes affecting the subcomponents of the earth system. This understanding can be put into mathematical expressions that, when combined, form a model of the behavior of particular components of the earth system. In practice, models of the atmosphere are connected to models of the oceans, ice, biota, and land surfaces to simulate the consequences of a scenario of societal forcing on climate and ecosystems. Controversy arises because both the societal forcing that will actually occur and the scientific knowledge of each subsystem is incomplete. Because models cannot be perfect replicas of the actual natural system, scientists must expend considerable efforts to test their models against the expanding base of field and laboratory data. This not only allows them to assess the credibility of current simulations but it also reveals improvements for the next generation of models.

A. Elements of Global Warming Forecasts

The societal driving forces behind global-warming scenarios are projections of population, consumption, land use, and technology. Typical twenty-first-century projections for human population size and affluence for less highly developed countries and more highly developed countries show drastic increases in population and wealth. When these factors are multiplied by the amount of energy used to produce a unit of economic product (the so-called energy intensity) and the amount of CO₂ emitted per unit of energy (the technology factor called carbon intensity), carbon emissions are predicted to increase severalfold during the next 100 years. It is very difficult to make such projections credibly; therefore, analysts disagree by as much as a factor of 10 about how much CO₂ will be emitted by 2100 (Johansson et al., 1993; Intergovernmental Panel on Climate Change, 1996a). Specific scenarios are debatable because the amount of carbon emitted through human activities will significantly depend on what kinds of energy systems will be developed and deployed globally and on what the standards of living will be during the next several decades, not to mention population growth.

To turn estimates of CO₂ emissions into estimates of CO₂ concentrations in the atmosphere—the variable needed to calculate potential climate changes—one must estimate what fraction of CO₂ emitted will remain in the atmosphere. This airborne fraction is most simply estimated at 50% because the amount of CO₂ buildup in the atmosphere each year (about 3 billion tons of carbon as CO₂) is about half the fossil fuel-injected CO₂. The atmospheric concentration of CO₂, however, should be computed by using carbon cycles models, which account for the time-evolving amounts of carbon in vegetation, soils, and oceanic and atmospheric subcomponents (Intergovernmental Panel on Climate Change, 1996a). The estimated CO₂ concentration can then be fed into computerized climatic models to estimate its effects on climate.

Climate prediction, like most other forecasts involving complex systems, generally involves educated guesses. Those attempting to determine the future behavior of the climate system from knowledge of its past behavior and present state basically can take two approaches. One approach, the empirical-statistical, uses statistical methods such as regression equations that connect past and present observations statistically to obtain the most probable extrapolation. The second approach, usually called climate modeling, focuses on first principles, which are equations representing laws believed to describe the physical, chemical, and biological processes governing climate. Because the statistical approach depends on historical data, it is obviously limited to predicting climates that have been observed or are caused by processes appropriately represented in the past conditions. The statistical method cannot reliably answer questions such as what would happen if atmospheric CO₂ increased at rates much faster than in the known past. Thus, the more promising approach to climate prediction for conditions or forcings different from the historic or ancient past is climate modeling. A significant component of empirical-statistical information, though, is often embedded into these models. (Washington and Parkinson, 1986; Root and Schneider, 1995). This often makes modelers uncomfortable about the validity of predictions of such models on unusual or unprecedented situations unless a great deal of effort is expended to test the models against current and paleoclimatic baseline data.

Climate models vary in their spatial resolution—that is, the number of dimensions they simulate and the
spatial detail they include. The simplest model calculates only the average temperature of the earth, independent of the average greenhouse properties of the atmosphere. Such a model is zero-dimensional: It reduces the real temperature distribution on the earth to a single point, a global average. In contrast, three-dimensional climate models produce the variation of temperature with latitude, longitude, and altitude. The most complex atmospheric models, the general circulation models, predict the time evolution of temperature plus humidity, wind, soil moisture, sea ice, and other variables through three dimensions in space (Washington and Parkinson, 1986).

B. Verifying Climate Forecasts

The most perplexing question about climate models is whether they can be trusted as a reliable basis for altering social policies, such as those governing CO₂ emissions or the shape and location of wildlife reserves. Even though these models are fraught with uncertainties, several methods are available for verification tests. Although no method is sufficient by itself, several methods together can provide significant, albeit circumstantial, evidence of a forecast’s credibility.

The first validation testing method involves checking the model’s ability to simulate the current climate. The seasonal cycle is a good test because temperature changes in a seasonal cycle are larger on a hemispheric average than the change from an ice age to an interglacial period (i.e., 15°C seasonal range in the Northern Hemisphere versus 5–7°C glacial–interglacial cycle). General circulation models map the seasonal cycle well. This supports the scientific consensus about the plausibility of global warming of several degrees in the twenty-first century. The seasonal test, however, does not indicate how well a model simulates slow processes such as changes in deep ocean circulation, ice cover, forests, soil carbon storage, which may have important effects on the decade- to century-long time scales over which atmospheric CO₂ is expected to double.

A second verification technique involves isolating individual physical components of the model and testing them against actual data. A reasonable model should reproduce the flow of thermal energy among the atmosphere, the surface, and space with no more than about a 10% error. Together, these energy flows make up the well-established natural greenhouse effect on Earth and constitute a formidable and necessary test for all models. A model’s performance in simulating these energy flows is an example of physical validation of model components.

A third validation method involves the model’s ability to reproduce the diverse climates of the past. This method is aided by recording instrumental observations made during the past few centuries and paleo-records that serve as a proxy for climatic conditions of the ancient earth, or even include testing the models’ ability to simulate climates of other planets (Kasting et al., 1988). Paleoclimatic simulations of the Mesozoic (Age of the Dinosaurs), glacial–interglacial cycles, or other extreme past climates help scientists understand the coevolution of the earth’s climate and living things (Schneider and Londer, 1984). As verification tests of climate models, they are also crucial to predicting future climates and changes in biological systems.

Using these techniques, much has been learned from examining the global climatic trends of the past century. The years 1997 and 1998 were the warmest on record for the lower atmosphere in the past century; at the same time, the stratosphere was at its coldest (Intergovernmental Panel on Climate Change, 1996a). These data are consistent with an enhanced greenhouse effect signal that might be anticipated from the greenhouse gas injections during the past 130 years, which saw a 25% increase in CO₂, a 150% increase in CH₄, and the introduction of human-generated heat-trapping chemicals such as chlorofluorocarbons and halons. Industrial activities since the 1950s have contributed to the increase of sulfur dioxide and other aerosol particles into the atmosphere, which reduce surface temperature by reflecting sunlight back to space. Although such cooling effects may have counteracted global warming by only several tenths of a degree, the hazes occur regionally and could be producing ecologically significant, unexpected regional changes in climate patterns (Schneider, 1994).

Although the 0.7 ± 0.2°C surface warming in the twentieth century is consistent with the human-induced greenhouse gas buildup, it is possible that the 0.7°C warming trend was wholly natural and that there was little or no contribution from the buildup of greenhouse gases (for estimates of the probability of human-induced global warming amounts, see Morgan and Keith, 1995). However, it is also not possible to rule out the counterfactual that, independent of the enhanced greenhouse effect due to human activity, there was a natural cooling fluctuation taking place during the twentieth century. If so, the world would then have warmed up much more than observed had we not had a fortuitous natural cooling trend. One could even speculate that the dramatic temperatures since the 1970s with global high temperature records reflect the termination of a natural cooling trend combined with the
rapid establishment of the expected enhanced greenhouse effect.

Studies (Santer et al., 1996) suggest that when aerosols and greenhouse gas forcings are combined, climate models more closely match 30 years of observations. Nevertheless, there is wide variability in predictions for a doubling of CO₂: Temperature changes as low as a 0.5°C warming to as high as a 5.0°C warming (Wigley and Raper, 1991) are all consistent with current observations. Several reasons exist for such a wide range of uncertainty: difficulty in knowing how to model delays in global warming because of the large heat capacity of the oceans; not knowing what other global change forcings may have opposed warming, for example, sulfate aerosols from burning high-sulfur coal and oil or undetectable changes in the sun’s light output before 1980; and large, unknown, internal natural climatic fluctuations. As mentioned previously, though, the ecologically important forecasts of time-evolving regional climatic changes are much less credible and require that ecologists use many alternative scenarios of possible climatic changes.

In summary, no clear physical objection or direct empirical evidence have contradicted the consensus of scientists (Intergovernmental Panel on Climate Change, 1990, 1996a) that the world is warming; nor has evidence emerged to contradict the substantial probability that temperatures will rise because of increases in greenhouse gases (Morgan and Keith, 1995). The evidence for current global warming forecasts is circumstantial, but it is sufficient enough that many researchers believe that recently observed climatic variations and human activities are probably related (Karl et al., 1995). The Intergovernmental Panel on Climate Change (1996a) carefully weighed the uncertainties and concluded that “the balance of evidence suggests that there is a discernible human influence on global climate.”

V. RELEVANCE OF CLIMATE MODELING TO REGIONAL CLIMATE CHANGE AND ECOSYSTEM STUDIES

Scientists who estimate the future climatic changes that are relevant to ecosystems have focused on the general circulation models that attempt to represent mathematically the complex physical and chemical interactions among the atmosphere, oceans, ice, biota, and land. As these models have evolved, increasingly more information has become available, and more comprehensive simulations have been performed. Nevertheless, the complexities of the real climate system still vastly exceed the general circulation models and the capabilities of even the most advanced computers (Intergovernmental Panel on Climate Change, 1990, 1996a). Simulating 1 year of weather in 30-min time steps with the crude resolution of 40 latitudinal lines × 48 longitudinal lines and 10 vertical layers—nearly 20,000 grid cells throughout the world—takes several hours on a supercomputer. This level of resolution, however, cannot resolve the Sierra Nevada of California and the Rocky Mountains as separate mountain chains. Refining the resolution to 50-km grid squares would so dramatically increase the number of computations that it would take approximately months of computer time to simulate weather statistics for 1 year.

Even the highest-resolution, three-dimensional general circulation model will not have a grid with nodes much less than 100 km apart within the foreseeable future; individual clouds and most ecological research (to say nothing of cloud droplets) occur on scales far smaller. Therefore, general circulation models will not be able to resolve the local or regional details of weather affecting most local biological communities or the importance of regional effects of hills, coastlines, lakes, vegetation boundaries, and heterogeneous soil (Root and Schneider, 1993). Nonetheless, it is important to have climatic forecasts and ecological response analyses on the same physical scales.

What is most needed to evaluate potential biological effects of temperature change is a regional projection of climatic changes that can be applied to ecosystems at a regional or local scale. Analyses of large, prehistoric climatic changes (Barron and Hecht, 1985; Budyko et al., 1987; Schneider, 1987; Cooperative Holocene Mapping Project, 1988) and historical weather analogs (Pittock and Salinger, 1982; Jager and Kellogg, 1983; Lough et al., 1983; Shabalova and Konnen, 1995) provide some insights into such changes. Historical weather analogs, however, since they are empirically and statistically based, rely on climatic cause-and-effect processes that probably differ from those that will be driven by future greenhouse gas radiative effects (Schneider, 1984; Mearns et al., 1990; Crowley, 1993). Consequently, ecologists turn to climatic models to produce forecasts of regional climatic changes for the decades ahead. How credible are such forecasts?

A. Regional Changes

Although the consensus among researchers about the plausibility of significant human-induced global climatic change is growing, no assessment (Intergovern-
mental Panel on Climate Change, 1996a) has suggested the existence of a strong consensus about how that global climatic change might be distributed regionally. For example, the world is not actually undergoing a dramatic and instantaneous doubling of CO₂, which is the hypothesis used in most standard computer model experiments applied to ecological assessments. Instead, the world is undergoing a steady increase in greenhouse gas forcing. Because that increase is heating the earth in a reasonably uniform way, one might expect a uniform global response, although this is far from likely. For example, the centers of continents have relatively low heat-retaining capacity, and the temperatures there would move relatively rapidly toward whatever their new equilibrium climate would be compared with the centers of oceans, which have high heat-retaining capacity. Tropical oceans, though, have a thin (about 50 m) mixed layer that interacts primarily with the atmosphere. It takes about 10 years for that mixed layer to substantially change its temperature, which is still much slower than the response time of the middle of the continents but is much faster than that of the oceans closer to the poles. At high latitudes, in places such as the Weddell or Norwegian seas, waters can mix down to the bottom of the ocean, thereby continuously bringing up cold water and creating a deep-water column for which a century or more is required to substantially change its temperature.

During the transient phase of climate change over the next century, therefore, one would expect the middle of continents, the middle of oceans, and the polar and subpolar oceans all to change toward their new equilibrium temperatures at different rates. Thus, the temperature differences from land to sea and equator to pole will evolve over time, which in turn implies that the transient character of regional climatic changes could be very different from the expected long-term equilibrium (Schneider and Thompson, 1981; Stouffer et al., 1989; Washington and Mehl, 1989). This does not imply that transient regional changes are inherently unpredictable, only that currently they are very difficult to predict credibly.

Even more uncertain than regional averages, but perhaps more important to long-term ecosystem responses, are estimates of climatic variability during the transition to a new equilibrium, particularly at the regional scale. These include estimates of events such as the frequency and magnitude of severe storms, enhanced heat waves, temperature extremes, sea-level rises (Titus and Narayanan, 1995), and reduced frost probabilities (Mearns, et al., 1984, 1990; Parry and Carter, 1985; Wigley, 1985; Rind et al., 1989). For example, there is a physical principle that evaporation increases dramatically as surface-water temperature increases. Because hurricanes are powered by evaporation and condensation of water, if all other factors are unchanged, the intensity of hurricanes and the length of the hurricane season could increase with warming of the oceans (Emanuel, 1987; Knutson, 1998). Such changes would significantly affect susceptible terrestrial and marine ecosystems (Doyle, 1981; O'Brien et al., 1992).

B. Downscaling Climate Predictions to Regional Effects

1. Empirical Mapping Techniques
Techniques exist that can translate the output of climate models so that it is closer to that of most ecological scales. One method that uses actual climatic data at both large and small scales can help provide maps that may allow small-scale analysis of large-scale climate change scenarios. For example, the Sierra Nevada of California or the Cascades in the northwestern United States are north–south mountain chains whose east–west dimensions are smaller than the grid size of a typical general circulation model. In the actual climate system, onshore winds on the Pacific coast would produce cool upslope and rainy conditions on the western slope and a high probability of warmer and drier conditions associated with that flow pattern on the downslope or eastern slope.

One regional map has been generated for Oregon (Gates, 1985) in which a high-resolution network of meteorological stations was used to plot temperature and precipitation isopleths based on observed climatic fluctuations at large (e.g., state-sized) scales. These maps show that the dominant mode of variation for this area is warm and dry on one side of the mountains and cold and wet on the other side. Although this empirical mapping technique seems appropriate for translating low-resolution, grid-scale climate model forecasts to local applications, a strong caveat must be offered. That is, the processes in the climate system that give rise to internal variability or natural fluctuations are not necessarily the same processes that would give rise to local deviations from large-scale patterns if the climate change were driven by external forces rather than an internal variation of the system. For example, the Oregon maps would indicate that if the grid box average temperature were warmer on the eastern slope, then it should be cooler and wetter on the western slope. This condition is the most probable regional situation for today's naturally fluctuating climate. However, if 50
years from now the warming on the eastern slope were, for example, a result of doubled atmosphere CO₂ causing an enhanced downward infrared radiative heating, then both eastern and western slopes would probably experience warming. Although the degree of warming and associated precipitation changes would not necessarily be uniform, an entirely different climatic change pattern would probably occur as opposed to what obtained from the empirical mapping technique if one used the naturally varying weather conditions existing today rather than the anthropogenically forced conditions of the twenty-first century (Schneider, 1993b).

Therefore, techniques to shrink climate forecasts that use current distributions of environmental variables at local scales and correlate them with current large-scale regional patterns will not necessarily provide a good guideline about how large-scale patterns would be distributed regionally. The reason is that the causes of the future change may be physically or biologically different from the causes of the historical fluctuations that led to the empirical maps in the first place. This caveat is so important that it requires scientists to use extreme caution before adopting such empirical techniques for global change applications.

2. Regional-Scale Models with General Circulation Model Inputs

Other techniques can still translate large-scale patterns to smaller scales, but these techniques are based on known processes rather than empirical maps for today's conditions. One such technique is to drive a high-resolution, process-based model for a limited region with the large-scale patterns produced by a general circulation model. In essence, this approach uses a mesoscale model (i.e., 10-50-km² grid cells) based on physical laws to solve the problem of translating general circulation model grid-scale averages into a finer scale mesh much closer to the dimensions of most ecological applications. Of course, even this mesoscale grid will still be too coarse to assess many impacts, necessitating further downscaling techniques. Neither are the problems of general circulation models entirely eliminated by mesoscale grids because they are bigger than individual clouds or trees. However, such methods do bring climate model scales and ecological-response scales much closer.

VI. EXAMPLES OF ECOLOGICAL RESPONSES TO CLIMATE CHANGES

Bringing climatic forecasts to ecological applications at local and regional scales is one way to bridge the scale gap across ecological and climatological studies. Ecologists, however, have also analyzed data and constructed models that apply over large scales, including the size of climatic model grids. A long tradition in ecology has associated the occurrence of vegetation types or the range limits of different species with physical factors such as temperature, soil moisture, land-sea boundaries, or elevation. Biogeography is the field that deals with such associations, and its results have been applied to estimate the large-scale ecological response to climate change.

A. Predicting Vegetation Responses to Climate Change

The Holdridge (1967) life zone classification assigns biomes (e.g., tundra, grassland, desert, or tropical moist forest) according to two measurable variables—temperature and precipitation. Other more complicated large-scale formulas have been developed to predict vegetation patterns from a combination of large-scale predictors (e.g., temperature, soil moisture, or solar radiation); vegetation modeled includes individual species (Davis and Zahniski, 1992), limited groups of vegetation types (Box, 1981), or biomes (Prentice, 1992; Melillo et al., 1993; Nelson, 1993). These kinds of models predict vegetation patterns that represent the gross features of actual vegetation patterns, which is an incentive to use them to predict vegetation change with changing climate, but they have some serious drawbacks as well. That is, they are typically static, not time-evolving dynamic simulations, and thus cannot capture the transient sequence of changes that would take place in reality. In addition, such static biome models occasionally make “commission errors”; that is, they predict vegetation types to occur in certain zones where climate would indeed permit such vegetation, but other factors such as soils, topography, or disturbances such as fire actually preclude it. Furthermore, local patterns may influence vegetation dynamics at scales not captured in some simulations, and seed germination and dispersal mechanisms are also either not explicitly simulated or simulated only crudely with such models. It is remarkable that they are still able to produce generalized maps of vegetation types which do indeed resemble current or even paleoclimatic patterns in a broad sense, but their details do not provide confident projections for future vegetation states. Fortunately, progress is being made to include some of the deficiencies mentioned previously, and so-called dynamical global vegetation models are being developed to treat the transient nature of vegetation change that would likely accompany climatic change.
B. Predicting Animal Responses to Climate Change

Scientists of the U.S. Geological Survey, in cooperation with Canadian scientists, conduct the annual North American Breeding Bird Survey, which provides distribution and abundance information for birds throughout the United States and Canada. From these data, collected by volunteers under strict guidance from the U.S. Geological Survey, shifts in bird ranges and abundances can be examined. Because these censuses were begun in the 1960s, these data can provide a wealth of baseline information. Price (1995) used these data to examine the birds that breed in the Great Plains. By using the present-day ranges and abundances for each of the species (Fig. 2a), Price derived large-scale, empirical-statistical models based on various climate variables (e.g., maximum temperature in the hottest month and total precipitation in the wettest month) that provided estimates of the current bird ranges and abundances (Fig. 2b). Then, by using a general circulation model to forecast how doubling of CO₂ would affect the climate variables in the models, he applied the statistical models to predict the possible shape and location of the birds' ranges and abundances (Fig. 2c).

Significant changes were found for nearly all birds examined. The ranges of most species moved north, up mountain slopes, or both. The empirical models assume that these species are capable of moving into these more northerly areas, provided habitat is available and no major barriers exist. Such shifting of ranges and abundances could cause local extinctions in the more southern portions of the birds' ranges, and, if movement to the north is impossible, extinctions of entire species could occur. We must bear in mind, however, that this empirical-statistical technique, which associates large-scale patterns of bird ranges with large-scale patterns of climate, does not explicitly represent the physical and biological mechanisms that could lead to changes in birds' ranges. Therefore, the detailed maps should be viewed only as illustrative of the potential for very significant shifts with different possible doubled CO₂ climate change scenarios. More refined techniques that also attempt to include actual mechanisms for ecological changes are discussed later.

Reptiles and amphibians, which together are called herptiles, are different from birds in many ways that are important to our discussion. First, because herptiles are ectotherms—meaning that their body temperatures adjust to the ambient temperature and radiation of the environment—they must avoid environments in which temperatures are too cold or too hot. Second, amphibians must live near water not only because the reproductive part of their life cycle is dependent on water but also because they must keep their skin moist because they breathe through their skin as well as their lungs. Third, herptiles are not able to disperse as easily as birds because they must crawl rather than fly, and the habitat through which they crawl must not be too dry or otherwise impassable (e.g., high mountains or super-highways).

As the climate changes, the character of extreme weather events, such as cold snaps and droughts, will also change (Karl et al., 1995), necessitating relatively rapid habitat changes for most animals. Rapid movements by birds are possible since they can fly, but for herptiles such movements are much more difficult. For example, R. L. Burke (personal communication) noted that during the 1988 drought in Michigan, many more turtles than usual were found dead on the roads. He assumed they were trying to move form their usual water holes to others that had not yet dried up or that were cooler. For such species, moving across roads usually means high mortality. In the long term, most birds can readily colonize new habitat as climatic regimes shift, but herptile dispersal (colonization) rates are slow. Indeed, some reptile and amphibian species may still be expanding their ranges north even now, thousands of years after the last glacial retreat.

R. L. Burke and T. Root (personal communication) have begun analyzing North American herptile ranges in an attempt to determine which, if any, are associated with climatic factors such as temperature, vegetation-greening duration, and solar radiation. Their preliminary evidence indicates that northern boundaries of some species ranges are associated with these factors, implying that climatic change could have a dramatic impact on the occurrence of herptile species. It could also alter the population genetics within species since there can be genetic differences among populations with respect to climate tolerance. Furthermore, several reptile and most North American turtle species could exhibit vulnerability to climatic change because the temperature experienced as they develop inside the egg determines their sex. Such temperature-dependent sex determination makes these animals uniquely sensitive to temperature change, meaning that climatic change could potentially cause dramatic range contractions. Many more extinctions are possible in herptiles than in birds because the forecasted human-induced climatic changes could occur rapidly when compared with the rate of natural climatic changes, and because the dispersal ability of most herptiles is very slow, even without considering the additional difficulties associated with human land-use changes disturbing their migration paths.
In general, animals most likely to be affected earliest by climatic change are those for which populations are fairly small and limited to isolated habitat islands. As a result of human-generated landscape changes, many reptiles now fall into this category, as do many other animals. Indeed, temperature-dependent sex-determined species are especially likely to suffer from extreme sex ratio biases, and therefore their sensitivity to rapid climate change appears potentially more severe than that of most other animals.

There are estimates that many small mammals living near isolated mountaintops (which are essentially habitat islands) in the Great Basin would become extinct given typical global change scenarios (MacDonald and Brown, 1992). Recent studies of small mammals in Yellowstone National Park show that statistically significant changes in both abundances and physical sizes of some species occurred with historical climate variations (which were much smaller than most projected climate changes for the next century), but there appear to have been no simultaneous genetic changes (Hadley et al., 1997). Therefore, it is likely that climate change in the twenty-first century could cause substantial alteration to biotic communities, even in protected habitats such as Yellowstone National Park. In addition, the biomass of macrozooplankton in waters off southern California has decreased dramatically as surface waters warmed (Roemmich and McGowan, 1995). Similarly, a study suggests that statistically the range of the Edith's checkerspot butterfly in western North America has shifted northward and upward in association with long-term regional warming trends (Parmesan, 1996).

C. Top-Down Approaches

The biogeographic approach previously summarized is an example of a top-down technique (e.g., Holdridge life zone classification), in which data on abundances or range limits of vegetation types or biomes are overlain on data of large-scale environmental factors such as temperature or precipitation. When associations among large-scale biological and climatic patterns are revealed, biogeographic rules expressing these correlations graphically or mathematically can be used to forecast changes in vegetation driven by given climate changes.

D. Bottom-Up Approaches

Another traditional analysis and forecasting technique is often referred to as bottom-up. Small-scale ecological studies have been undertaken at the scale of a plant or even a single leaf (Idso and Kimball, 1993) to understand how, for example, increased atmospheric CO₂ concentrations might directly enhance photosynthesis, net primary production, or water-use efficiency. Most such studies indicate increases in all these factors—increases that some researchers have extrapolated to ecosystems (Idso and Brazel, 1984; Ellsasser, 1990).

However, at the scale of a forest, the relative humidity within the canopy, which significantly influences the evapotranspiration rate, is regulated by the forest. In other words, if an increase in water-use efficiency decreased the transpiration from each tree, the aggregate forest effect would be to lower relative humidity. This, in turn, would increase transpiration, thereby offsetting some of the direct CO₂/water-use efficiency improvements observed experimentally at the scale of a single leaf or plant. Regardless of the extent to which this forest-scale feedback effect will offset inferences made from bottom-up studies of isolated plants, the following general conclusion emerges: The bottom-up methods may be appropriate for some processes at some scales in environmental science, but they cannot be considered credible without some sort of validation testing at the scale of the system under study.

E. Combined Top-Down and Bottom-Up Approaches

To help resolve the deficiencies of the top-down biome forest models mentioned previously, more process-based, bottom-up approaches such as forest gap models have been developed (Botkin et al., 1972; Pastor and Post, 1988; Smith et al., 1992). These models include individuals species and can calculate vegetation dynamics driven by time-changing climatic change scenarios.

---

**FIGURE 2** (a) Map of current range and abundance of the bobolink as determined from actual observations during the U.S. Geological Survey Breeding Bird Survey and (b) map of current range and abundance of the bobolink as estimated from the empirical–statistical model. The high correspondence in patterns between the maps in a and b suggests that this model reliably captures many of the features of the actual observed range and abundance of this species as depicted in the map in a. (c) Map of the forecasted range and abundance of the bobolink for climate change response of a model with doubled CO₂. This map illustrates the potential for very significant shifts that doubled CO₂ would cause.
However, the actual growth rate calculated in the model for each species has usually been determined by multiplying the ideal growth rate curve by a series of growth-modifying functions that attempt to account for the limiting effects of nutrient availability, temperature stress, and so forth. These growth-modifying functions for temperature are usually determined empirically at a large scale by fitting an upside-down U-shaped curve, whose maximum is at the temperature midway between the average temperature of the species' northern range limit and the average temperature of its southern range limit. Growing degree-days (related to temperature but not temperature per se) are used in this scenario.

In essence, this technique combines large-scale, top-down empirical pattern correlations into an otherwise mechanistic bottom-up modeling approach. Although this combined technique refines both approaches, it has been criticized because such large-scale, top-down inclusions are not based on the physiology of individual species and lead to confusion about the fundamental and realized niches (Facala and Hurd, 1993). (The fundamental niche is the ecological space in which a given species could theoretically survive—for example, if its competitors were absent—and the realized niche is where it actually exists.) The question is then, what limits the realized niche, particularly at the southern boundary? Furthermore, more refined models should include factors such as seed dispersal so that plant recruitment is related to the preexisting population and is not simply the result of a random number generator in the computer code.

F. Studies of More Refined Approaches

As noted previously, problems with the singular use of either top-down or bottom-up methods have led to well-known criticisms. For bottom-up models, the primary problem is that some of the most conspicuous processes observable at the smaller scales may not be the dominant processes that generate large-scale patterns.

Top-down approaches suffer because of the possibility that the discovered associations at large scales are statistical artifacts that do not, even implicitly, reflect the causal mechanisms needed for reliable forecasting. As Jarvis (1993, p. 121) stated, "A major disadvantage of a top-down model is that predictions cannot be made safely outside the range of the variables encountered in the derivation of the lumped parameter function."

A search of the literature (Wright et al., 1993; Root, 1994; Harte et al., 1995) provides examples of a refined approach to analyzing across large and small scales, which Root and Schneider (1995) labeled strategic cyclical scaling. This method builds on the combined techniques in which top-down and bottom-up approaches are applied cyclically in a strategic design that addresses a practical problem: in our context, the ecological consequences of global climatic change. Large-scale associations are used to focus small-scale investigations; this helps ensure that tested causal mechanisms are generating the large-scale relations. Such mechanism becomes the laws that allow more credible forecasts of the consequences of global change disturbances. Levin (1993) observed.

Although it is well understood that correlations are no substitute for mechanistic understanding of relationships, correlations can play an invaluable role in suggesting candidate mechanisms for (small-scale) investigation. (p. 14)

Strategic cyclical scaling, however, is intended not only as a two-step process but also as a continuous cycling process between large- and small-scale studies, with each successive investigation building on previous insights from all scales. This approach is designed to enhance the credibility of the overall assessment process (Vitousek, 1993; Harte and Shaw, 1995), which is why strategic is the first word in strategic cyclical scaling.

If the rate at which humans are injecting greenhouse gases into the atmosphere is not greatly decreased, there is a significant chance that the earth's climate will warm by several degrees Celsius by 2050 (Titus and Narayan, 1995). With this in mind, Root (1988) examined the biogeographic patterns of all wintering North American birds. She chose this group of species because birds are important parts of ecosystems and because of the availability of the necessary data. The National Audubon Society and the U.S. Geological Survey have volunteer forces amassed to aid in the collection of Christmas Bird Count data and Breeding Bird Survey data, respectively. By using Christmas Bird Count data, Root determined that for a large proportion of species average distribution and abundance patterns are associated with various environmental factors (e.g., the northern range limits of some species apparently may be limited by average minimum January temperature; Root, 1988, 1989; Repasley, 1991).

The following is the scaling question: What mechanisms (such as competition or thermal stress) at small scales may have given rise to the large-scale associations? Root first tested the hypothesis that local physiological constraints may be causing most of the particular large-scale, temperature-range boundary associations. She used published small-scale studies on the wintering
physiology of key species to determine that about half of the songbird species wintering in North America extend their ranges no farther north than the regions where, to avoid hypothermia during winter nights, they need not increase their metabolic rates more than approximately 2.5 times their basal metabolic rate (Root, 1988). Root embarked on a larger, regional study to determine whether the longer nights—hence, fewer hours of daylight available for foraging—or the colder temperatures in the more northerly locations are relatively more important. Preliminary results indicate that changing temperatures are more likely than day length to explain this effect (Root, unpublished data): Thus, global temperature changes would probably cause a rapid range and abundance shift, at least by selected bird species. Indeed, Root found significant year-to-year shifts in ranges and abundances; these shifts are apparently associated with year-to-year changes in winter temperatures. No claim is made at this point in the research for the generally of the preliminary results indicating strong and quantitative links between bird disturbances and climate. This example does permit, however, a clear demonstration of refined methods for cycling across scales to estimate ecological responses to climatic change.

VII. THREE-WAY LINKAGES AND COMMUNITY ECOLOGY

The anticipated changes in plant ranges will probably have dramatic effects on animals, both on the large biogeographic scale and on the local regional scale. The ranges of many animals are strongly linked to vegetation. For example, red-cockaded woodpeckers are endemic to mature longleaf pine and pine-oak forests (Mengel and Jackson, 1977), and the winter range of Sprague's pipit is coincident with bluestem, a grass (Root, 1988). Consequently, the ranges of various animals that rely on specific vegetation will change as the ranges of these plants shift, assuming that some other factor is not limiting them. If the climate changes more rapidly than the dispersal rates of the plants, it will result in extensive plant die-offs in the south or down slope before individuals can disperse and become established in the north or up slope. Thus, the ranges of animals relying on these plants could become compressed, and in some instances both the plants and the animals could become extinct. For instance, the red-cockaded woodpecker needs mature, living trees for nesting sites (Jackson, 1974), and if increasing temperature causes most large trees to die before the newly established dispersing trees grow large enough, then this woodpecker, federally listed as endangered, could easily become extinct.

Many animal species have ranges that are not directly limited by vegetation but are instead restricted by temperature. This is true for most ectotherms (insects and related arthropods, amphibians, and reptiles) as well as some endotherms (mammals and birds). For example, the eastern phoebe, a North American songbird, winters in the United States in areas with average minimum temperatures warmer than 4°C (Root, 1988). As the earth warms, those species directly limited by temperature will be able to expand northward as rapidly as their dispersal mechanisms will allow, again assuming other factors are not limiting them. The animals limited by vegetation will be able to expand their ranges only as rapidly as the vegetation changes. Consequently, the potential for significant disruption among communities is high. For instance, some animals may no longer be able to coexist because an invading species disrupts the balance between competing species or between predator and prey. Therefore, to understand the ecological consequences of global climatic change on animals, the three-way linkages among animals, plants, and climate must be understood. It is critical to realize that this is not simply a one-way process whereby climate influences biota but rather a three-way process because animals and plants affect each other and are affected by climate. At the same time, altered surface vegetation can affect climate because mid-continent summer precipitation is significantly influenced by water vapor from evapotranspiration (Ye, 1989; Salati and Nobre, 1991).

VIII. CLIMATE FORECASTS, ECOSYSTEM RESPONSES, AND SYNERGISTIC EFFECTS

A. Improve Regional Analysis, Study Transients, and Include Many Variables

The most reliable projections from climatic models are for global-scale temperature changes. Ecological impact assessments, however, need time-evolving (transient) scenarios of regional to local-scale climate changes: included are changes in precipitation; severe storm intensity, frequency, and duration; drought frequency, intensity, and duration; soil moisture; frost-free days; intense heat waves; ocean currents; upwelling zones; near-ground ozone; forest canopy humidity; and ultraviolet
radiation and total solar radiation reaching the surface, where photosynthesis is important. Data gathered at many scales and by coordinated volunteer and professional sources are needed for archives of these regional and local variables, which in turn can be used to develop and test models or other techniques for climatic forecasting.

B. Abrupt Climatic Changes

We have argued that sustained globally averaged rates of Earth and ocean surface temperature changes from the past Ice Age to the present were about 1°C per 1000 years. Alarmingly, this is a factor of approximately 10 slower than the expected changes of several degrees Celsius per 100 years typically projected for the twenty-first century due to human effects. We emphasize the words sustained globally averaged because comparably rapid regional variations have occurred. For example, about 13,000 years ago, after warm-weather fauna had returned to northern Europe and the North Atlantic, there was a dramatic return to ice age-like conditions in less than 100 years. This Younger Dryas climatic event lasted about 1000 years before the stable recent period was established (Berger and Labeyrie, 1987). The Younger Dryas was also accompanied by dramatic disturbances to plants and animals in the North Atlantic and Europe (Coope, 1977; Ruddiman and McIntyre, 1981). During the same period, dramatic shifts can be found outside of the North Atlantic region (e.g., Sevringhaus, 1998), but no significant climate change is evident in Antarctic ice cores. However, studies of the North Atlantic show that the warm gulf stream current deviated many degrees of latitude to the south and that the overall structure of deep-ocean circulation may have returned to near ice age form in only decades—a weakening of the vertical circulation known as the conveyor-belt current (Broecker et al., 1985).

Plausible speculations about the cause of the Younger Dryas center on the injection of fresh meltwater into the North Atlantic, presumably associated with the breakdown of the North American ice sheet (Boyle and Weaver, 1994; Paillard and Labeyrie, 1994). Could such a rapid change to the conveyor-belt current be induced today by pushing the present climatic system with human disturbances such as greenhouse gases or sulfur oxides? The potential for this is speculative, of course, but its possibility has concerned many scientists (Broecker, 1994, 1998; Rahmstorf, 1999). The prospect of climatic surprises in general is chilling enough to lend considerable urgency to the need to speed up the rate of our understanding, slow down the rates at which we are forcing nature to change, or both.

C. Adaptability

Our current inability to credibly predict time-evolving regional climatic changes has many implications, one of which concerns the adaptability of agricultural ecosystems. That is, any experience farmers might have with anomalous weather in, for example, the 2020s may not help them adapt to the evolving climate change in the 2030s because a transient climate change could differ dramatically over time. This would inhibit learning by doing, creating a potential lack of adaptability associated with the difficulty of reliably predicting regional climatic consequences (Schneider et al., 2000). Such rapid climate changes would be especially difficult for natural ecosystems to adapt to because habitats do not have the luxury of "choosing" to plant new seeds or change irrigation systems, soil tillage practices, or other agricultural practices.

D. Ecological Applications-Driven Climatic Research

Regional projections of climatic change arising from a variety of greenhouse gas and sulfur oxide emissions scenarios are essential for ecological applications. Such studies must stress the climatic variables most likely to have significant effects on biological resources. For example, extreme variability measures such as high temperature and low relative humidity are important for evaluating the risk of forest fires (Torn and Fried, 1992). Identifying such variables of ecological importance and communicating this information to climate scientists require close interdisciplinary, multi-institutional, and cross-scale research efforts to ensure that combinations of variables relevant to ecological applications receive research priority by climatologists. A focus of climate research toward changing climatic variability (Mearns et al., 1984, 1990; Rind et al., 1989) might be more useful for ecological impact assessments than the current focus among climatic modelers on climatic means.

E. Interactive, Multiscale, Ecological Studies Needed

Most ecological studies project the response of one species at small scales or shifts in biomes at large scales to an equilibrium, CO₂-doubled climate model (e.g., the
Vegetation/Ecosystem Modeling and Analysis Project, 1995). What is needed for more realistic and useful ecological impact assessments is a multi-scale, multi-species, multi-taxon analysis driven by regionally specific, transient climatic change forecasts. The construction of ecological forecasts models first requires large-scale data sets gathered locally by professional (e.g., U.S. Geological Survey landcover data sets) and volunteer (e.g., National Audubon Society Christmas Bird Count) workers. Without such data sets, virtually no credible progress is possible in determining large-scale patterns of associations among ecological and climatic variables. Small-scale studies informed by large-scale patterns are then needed to refine causal mechanisms underlying such large-scale associations, thereby testing the formulas used to make projections of various species or biome responses to hypothesized global changes. For example, Pacala and Hurnt (1993) suggested small- to medium-scale experiments to improve forest gap models. Their criticisms suggest that largely first-principles, bottom-up models may still be unrealistic if some top-down parameters (i.e., growth-modifying functions in the instance of gap models) are not appropriately derived from data at the scale at which the models are being applied (Root and Schneider, 1995).

One obvious truism emerges: Credible modeling required for forecasting across many scales and for complex interacting systems is a formidable task requiring repeated testing of many approaches. Nevertheless, tractable improvements in refining combined top-down and bottom-up techniques can be made. It will, however, take more than one cycle of interactions to reliably address the cross-scale and multi-component problems of ecological assessment—what we (Root and Schneider, 1995) have labeled strategic cyclical scaling.

See Also the Following Articles

CARBON CYCLE • CLIMATE EFFECTS OF • DIVERSITY, COMMUNITY/REGIONAL LEVEL • ENERGY USE, HUMAN • GEOLOGIC TIME, HISTORY OF BIODIVERSITY IN • GREENHOUSE EFFECT

Bibliography


CLIMATE, EFFECTS OF

F. I. Woodward
University of Sheffield

I. OBSERVATIONS

It is well established that the diversity of species increases toward the equator and that this observation is generally true for both animals and plants. The general term diversity indicates a measurement of the number of species which exist in a particular area. The measurement of species richness quantifies the number of species recorded for a particular area, with each species given an equal weighting. When species richness is adjusted to account for abundance, then species diversity is the appropriately termed measurement. Unfortunately, both measurements are often used interchangeably and with little precision. In this article, species richness is preferred, primarily because the aim is to investigate diversity at the global scale, when data are limited and, at best, may only include species lists, with some notional area of observation.

A. Species to Area Relationships

The area of ground within which species are counted plays a critical role in quantifying species richness. Rosenzweig (1995) provides an extensive account and examples of the relationships between species richness and area. One example is presented to illustrate the typical nature of the relationship for the flora of the British Isles (Fig. 1). Figure 1a shows the relationship for untransformed scales, with a particularly notable and rapid increase in species richness following small initial increments of area. Figure 1b presents the same data for log-transformed scales, in which the relationship is linear. Figure 1b indicates that the rapid increase in richness to about 1000 species occurs as the scale of observation increases from 0.1 ha to 1 million ha (10000 km² or 100 × 100 km).

All spatial scales contain data of ecological interest; however, in the context of the climatic control of diversity, the small areas are of less direct relevance. For example, the smallest scale in Fig. 1 of 0.1 ha (approximately 32 × 32 m) is too small to include many individual and mature trees. This is therefore a sampling prob-
lem. The problem lies in deciding when this species to area relationship might indicate a primary control by climate, over and above any problems of sampling. It can also be envisaged that if an area under study has low landscape diversity (e.g., it is a flat plain of similar soil type), then this will also influence the species to area relationship. Such environmental features must be incorporated into general species lists and cannot easily be dissected out from the major thrust of this article—the influence of climate. A comparison of species rich-
ness with scale for the British Isles and the fynbos of South Africa (Fig. 2) indicates very large differences in diversity, but the untransformed scale (Fig. 2a) shows that the differences between the two floras decrease much more slowly above areas of about $2 \times 10^7$ ha ($2 \times 10^5$ km$^2$ or $500 \times 450$ km). This indicates a useful order of scalar magnitude for comparing floras in the context of examining climatic control.

![Graph A](image)

**FIGURE 2** Species richness of the British Isles compared with the fynbos of South Africa: (A) untransformed axes and (B) log-transformed axes (regression lines from Rosenzweig, 1995).
B. Time Dependency

The species composition of an area endures changes with time, through mortality, disturbance, and migration. Disturbance of a moderate frequency can maximize species richness, producing landscapes in various stages of recovery from a disturbance event. An extreme disturbance event was the 1883 volcanic eruption which completely denuded the flora of the Krakatau Islands. A recent enumeration of higher plant species (Fig. 3; Whittaker et al., 1990) showed a monotonic increase in the species complement during the past 100 years. It is clearly not possible to assess the final species richness of these islands until the species complement reaches some asymptote. This example of primary succession is perhaps extreme in the contemporary world, although smaller scale recoveries from natural disturbance and larger scale recoveries from human disturbance are current widespread features. Human disturbance is generally considered to deplete species richness and push natural vegetation to early stages of secondary succession; however, over the large spatial scales considered for this article it has been assumed that species richness has not yet been markedly depleted by human activities.

The last ice age was also the last natural and very large-scale disturbance to the terrestrial biosphere. At that time, glaciers spread over large areas of North America, Europe, and Russia and clearly removed all species in their paths. The cooler glacial climate, in addition to the spread of glaciers, led to the equatorial migration of plant species. However, in Europe, for example, the Alps prevented significant migration further south, with significant extinctions at this barrier. The climatic amelioration to the current interglacial climate reversed the process which occurred during the ice age. However, reduced richness follows as a consequence of the irreplaceable species extinctions. The degree of this loss of richness can be addressed in part from paleoecological information. However, such information fails to detect all species, and therefore it is most likely that glaciations cause some species losses and reductions in regional richness, but it is not possible to assess the extent of these effects. In any event, these long-term effects, and the shorter term effects of disturbance, are incorporated in regional assessments of species richness and, because they are at least partially climatically related, they emphasize the impacts of contemporary climate.

C. Observations by Area

The impacts of area on species enumeration and on the relationship between climate and species richness are well illustrated (Fig. 4) using data provided by Williams (1964). Although the data are now outdated, in the sense that some larger areas are now better documented with greater richness, there is a clear distinction between areas which are wet and warm and those which are cold. Both the rate of increase in richness and the absolute richness with area are lower in the limiting, cold environment. Such observations conform with ob-
FIGURE 4  Species richness changes with area of observation and for areas with either wet and warm climates ( ), or cold climates ( ), with significant regression coefficients and constants (data from Williams, 1964).

CLIMATE, EFFECTS OF

D. Global-Scale Observations

The data shown in Fig. 4 are restrictive in terms of global spread and accuracy, primarily because of the incomplete estimates of species richness at the continental scale. A more current global-scale distribution of species richness is available (Barthlott et al., 1996) on the Internet at the following address: http://www.botanik.uni-bonn.de/system/biomaps.htm. This map attempts to present plant species diversity data at the global scale, after allowing for variations in recording areas, using a globally uniform species–area relationship. This is a useful approach to circumventing the problems of unequal sampling areas, but it does hide the fact that species–area relationships will change with location and climate (Fig. 4). Nevertheless, this distribution map provides an important and unified source of data for assessing relationships between species richness and climate.

1. Species Richness and Latitude

The first cut at assessing species richness is usually obtained by latitude. This spatial variable includes many environmental features, particularly solar radiation receipt, day length, and temperature, but it is a useful variable for assessing broad-scale controls on vegetation. Figure 5 indicates a spread of richness with latitude. In the northern latitudes, the maximum observed richness decreases with increasing latitude. A similar relationship can be defined for the Southern Hemisphere, except that there are significant exceptions. These cases include the southeast coastal region of Brazil, extending from Porto Allegre (Fig. 5, Brazil 3) in the south through Rio de Janeiro (Fig. 5, Brazil 1) to Vitoria (Fig. 5, Brazil 2) in the North. Additional sites of high diversity have been sampled for South Africa (The Cape) and Australia (Perth). The sampling scheme for global diversity does not cover all areas of high diversity, but these five sites in the Southern Hemisphere and, to a lesser extent, the Mediterranean region in the Northern Hemisphere clearly indicate that areas of unusually high diversity do exist. These areas are defined as unusual because they seem outside the typical latitudinally dependent maximum extents of diversity.

2. Species Richness and Climate

a. Temperature

The observations of Fig. 4 suggest that cooler climates are less diverse. This same result is observed at the global scale (Fig. 6), when mean annual climate (for the 1970s and sampled at a scale of 10000 km²) is compared with the Barthlott map of standardized species diversity (also adjusted to a 10000 km² scale). At
mean temperatures greater than about 5°C, species richness both increases markedly and becomes significantly more variable between individual sample points. The high-temperature, low-diversity sites are typically deserts or very arid, where it is clear that water supply is the critically limiting factor for plant survival. The areas of unusually high diversity in South Africa and Brazil are clear extremes in the temperature relationship, but the other sites shown in Fig. 5 are not distinct.

b. Precipitation

The relationship between richness and temperature is not simple and is also influenced by other climatic features, particularly precipitation. In contrast, 42% of the variance in richness (Fig. 7) is accounted for by a linear correlation between richness and annual precipitation. This association defines the obvious observation that diversity increases toward the tropics and is highest when warmth and high precipitation coincide. Four of the five recognized areas of high diversity remain as outliers from the remainder of the data set.

c. Net Primary Productivity

Precipitation is often taken as a surrogate of vegetation productivity in order to investigate any relationships between productivity, usually net primary productivity.
Although precipitation exerts a large effect on NPP, this is not the only influence and other climatic features such as solar radiation receipt, day length, growing season temperature, and annual minimum temperature can exert significant effects on NPP (Woodward 1987). Simulation models of NPP typically incorporate the effects of all these climatic variables to provide an integrated effect, seen as simulated NPP (Woodward et al., 1995).

In diversity studies, productivity is used and assessed in two different ways. For small field plot experiments, productivity is measured directly as the accumulation of biomass and is generally viewed as a consequence of diversity. On larger scales, intensive measurements of NPP are impossible and in this case productivity is generally simulated by some type of process model and it is more appropriate to consider diversity as a consequence of simulated NPP because this simulated value accounts for all variations in climate but is generally independent of any aspects of diversity.

In this case, a typical process model (Woodward et al., 1995) has been used. Annual NPP is the net fixation of carbon into plant biomass, after plant respiration is decremented from annual photosynthetic fixation. In ecological terms, NPP is a measure of the rate of production of biomass (usually over an annual interval) and which is available for consumers and for creating new plant structure. Simulation models can often provide more accurate estimates of annual production, particularly in multiseasonal environments, which would require many harvests for accurate assessments of NPP in the field.

Simulated NPP provides a tighter fit than annual precipitation to the observed variations in species richness at the global scale (Fig. 8). Three of the high-diversity sites, however, are notable outliers from the line of best fit. The site in South Africa is the most extreme, with a very high richness associated with a low NPP. Much of this area of vegetation is known as fynbos and is a heathland on very impoverished soils.

A further analysis of species richness in this area, and for the whole of southern Africa northwards to a latitude of 15°S, has been achieved by O’Brien et al. (1998). In this case, only the richness of woody species (native phanerophytes and plant heights of greater than 1.5 m) was considered. Woody species richness is very strongly correlated with NPP (Fig. 9) and the high-diversity site of Fig. 8 (arrow in Fig. 9) is not now distinguishable from the general trend. The high richness at this site is clearly due to herbaceous species and sub-shrubs and explanations for their high diversity may be related to factors not directly climatic, such as disturbance frequencies and types, the low nutrient status of the soils, and factors related to the history of the site, particularly rates of speciation and probabilities of species’ migrations.

II. THEORIES

The strong latitudinal gradients in species richness have long been recognized and have led to a wide range of theories to explain the observations. At the same time, relevant new data sets are being continually produced.

FIGURE 7 Changes in species richness with annual precipitation.
and analyzed in new ways. These data have been used to derive empirical or observational hypotheses about the controls of richness. In principle, if the empirical hypothesis is correct then it should also be in agreement with an appropriate theory. However, nature does not yield its secrets so easily and ambiguity seems to be the rule in that it is often difficult to differentiate between different theories and hypotheses when the differences in their consequences are often by degree rather than absolute. The following sections present current theories and hypotheses which address the relationships between climate and species richness.

A. Narrowing the Questions

Using Observations

The defined area for observing species richness exerts a major impact in the quantity of richness. A less obvious, but also important, consideration is the nature of the areas considered at increasing scale. Figure 10
indicates that islands show a greater increase in richness with area than do areas of the same size but nested within a region with no dispersal and migratory barriers. The species richness of any area is the net result of species immigration, extinction, and evolution. For an island, the rate of species immigration will be less than that for a land bridge, and as a consequence the island richness will be less because local species extinctions are not readily matched by migration. As the islands become larger, the areas for the arrival of migrants and the survival of extant species increase and so extinctions will be less and migrations greater and, as a consequence, richness will increase so that it is similar to that of the nested land sites. This example shows that the nature of a land form and its degree of isolation from continental sites will influence richness in a manner which can be quite independent of climate, and therefore it is critical to account for such responses.

An important contemporary corollary to the change in diversity with islands and mainland areas (Fig. 10) is that any process, such as human land clearance, which effectively makes small islands will inevitably lead to a loss of diversity.

The response shown in Fig. 10 also scales to larger areas (Fig. 11). In this case, the species to area relationship for the British Isles is much flatter than that for separate provinces and continents, from New Zealand, at the smallest, through Australia and Africa to the world. The flattest curve for the British Isles will include some component due to the net balance of extinctions and immigrations, plus a ready capacity for species to migrate throughout the British Isles, thus evening out, to some degree, the scale dependence of species richness. Different provinces and continents are much more isolated for immigration and also possess species complements determined by historical effects, not only past ice ages but also longer term speciation events. Climate will play and will have played a major part, therefore, in the contemporary species richness of these large areas. The climatic range of the British Isles, for example of temperature and precipitation, is small compared with that for a continent, but the range of soil conditions is wide from fertile alluvial soils to very low-nutrient bogs and chalk downs. Therefore, the species area curve (Fig. 11) for the British Isles also displays the influence of habitat diversity, particularly soil nutrient status on species richness, and the effect is small. It does not seem likely, therefore, that latitudinal gradients in soil nutrient conditions will play a major role in determining the gradients of richness seen, for example, with latitude (Fig. 5).

Observations by Rapoport (1982) may suggest a final constraint in understanding the climatic control of species richness. Rapoport noted the well-established observation that species richness declines with increasing latitude; however, this observation was supported by the observation that, on average, the geographical range of a species increased with latitude. Therefore, at high latitudes there are fewer species, but those which do occur are, on average, more widely dispersed. Subse-
sequent analyses (Gaston, 1999) have indicated that this relationship is neither universal nor simple. However, there is some consistency in the pattern from the tropics to high latitudes, particularly in the Northern Hemisphere.

B. Theories and Hypotheses

Theories and hypotheses for explaining the documented decline in diversity with latitude are based on one or more of the following fundamental controls:

- Competition for necessary resources
- Gap creation in vegetation by disturbance
- Energy availability
- Climatic extremes
- History

1. Competition

The competitive exclusion of species which require the same resource has been a theoretical underpinning to much of ecology, not just the climate control of diversity. Originating from simple experiments, it has proved easy to demonstrate the importance of competition in species interactions in response to changes in climate and in very small-scale experiments. It has not been so easy to demonstrate the importance of competition in defining the distributional limits of species on a larger field scale. Often, it seems that environmental heterogeneity in time and space can allow coexistence to occur, apparently circumventing direct competition for resources. However, competition is still considered an important mechanism in controlling survival and therefore diversity, but it does not appear to be easy to demonstrate competition in action.

It does seem that if competition is to play a major role in the control of species distributions, then it must do so as a secondary action to the primary location of climate. For example, consider the distribution of a boreal forest species in the cold of the high latitudes and a tropical species from the warmth of low latitudes. If competition plays a role in these distributions then it must favor boreal species at high latitudes and tropical species at low latitudes (i.e., competition is latitude dependent). Competition can only be latitude dependent through its impact on species’ interactions and this would either be through a climatic dependency of species’ contemporary or through historical responses. Therefore, it seems that competition is an integral and probably secondary part of a climatic control of diversity.

2. Disturbance

Creating gaps in vegetation by disturbance appears to provide a mechanism for avoiding direct competition and for providing bare ground and the opportunity for new species immigration. If some species are long-lived and can survive disturbance events, then they can co-
occur with shorter-lived species which depend on disturbance and therefore diversity can be maintained. If the rate of disturbance changed systematically with latitude, then this could provide an explanation for the decline in diversity with latitude. However, Runkle (1989) has shown that this is not the case, with similar rates of disturbance with latitude.

Runkle (1989) did propose a climatic-dependent combination of disturbance with competition as a theory for controlling diversity. The length of the growing season decreases with latitude, which is a direct consequence of the curvature of the earth, a decreased annual incidence of solar radiation with latitude, and a consequent decrease in annual temperature. Plants in general can only grow when the temperature exceeds the freezing point of water, and therefore this constrains the period of growth in climates in which freezing temperatures occur. Because disturbance rates are broadly similar with latitude, it is expected that gaps in the vegetation will be created during the winter period of no growth. In the more uniform, or aseasonal tropical climates gaps will also be created through the year, but these will be filled rapidly because propagules are generally available throughout the year for establishment. In the seasonal climate, in contrast, gaps will accumulate until the temperature threshold for growth is exceeded. In the short growing season there is likely to be intense competition to establish in the early season in these gaps. These plants must also reach an appropriate stage of maturity and winter tolerance by the end of the growing season for survival, so there will be no survival advantage for delayed establishment. In this situation, competition can act to amplify the limiting effects of a short growing season.

3. Energy Availability
Hutchinson (1939) suggested that the availability of energy may constrain species diversity. Studies have since shown strong positive correlations between diversity and evapotranspiration, which is often taken as a surrogate for both energy and NPP. Figure 8 shows a clear relationship between NPP and species richness, but in this case climates with high energy inputs of solar radiation but low precipitation will lead to low NPP. O’Brien et al. (1998) describe the same response where water availability tempers a direct response of diversity to energy. Therefore, the relationship between diversity and energy is not simple, whereas the relationship between NPP and richness is in general quite linear. However, the mechanistic nature of the relationship is quite unclear in either case.

In some cases, particularly in small-scale experiments with variable resource supply, NPP and diversity are often negatively correlated, with just a small number of rank growing species dominating energy capture and production. Such a response is opposite to that observed at the larger, global scale (Fig. 8) and, based on Fig. 1a, indicates that climatic control of diversity is difficult to determine when the area of observation is on the steeply rising curve of the species–area relationship. Nevertheless, small-scale experiments have often influenced ideas at the global scale. It is generally considered that sites of high NPP are highly competitive with only a small number of dominant species. However, highly productive tropical forests are also highly diverse in terms of both species and vegetation structure, even with limiting soil nutrient resources. The vegetation canopy may have many strata, with different species in each strata. High NPP requires a high supply rate of resources, particularly soil nutrients, water, and solar radiation, with no significant temperature limitations. On low-nutrient soils, the vegetation retains the majority of the soil nutrients, whereas water is in abundant supply. In these forests, turnover of plant parts (leaves, twigs, and small branches) is quite rapid, as is the decomposer system and thus also the rate of nutrient recycling. Therefore, the high NPP goes hand in hand with a high requirement and supply of nutrients to all component species. This appears to be the major mechanism by which high NPP can sustain high diversity. In addition, large areas of forest also transpire large quantities of water vapor, which returns as rain and also increases local cloudiness. The high NPP vegetation therefore exerts some control over its own climate, making it more equable and reducing energy supply.

4. Climatic Extremes
Climate appears to influence diversity by a range of mechanisms, such as by influencing turnover times of soil nutrients and through constraining the length of the growing season. However, it is also well-established that species have a range of minimum temperature tolerances which is much greater than the range of maximum temperature tolerances (Woodward, 1987). At the global scale, absolute minimum temperatures in vegetated areas of the world vary from about −80 to 20°C and species are appropriately adapted to the local absolute minimum temperature, even if this occurs only every 10–30 years. Therefore, minimum temperature tolerance increases with latitude. There also appears to be a cost involved in possessing a particular degree of temperature tolerance, with the cost probably reducing
growth capacity and competitive ability compared with those of species possessing a lower minimum temperature tolerance (Woodward, 1987). As the absolute minimum temperature decreases, so also does the number of species that can survive (Woodward, 1987).

The absolute minimum temperature decreases with latitude; therefore, this mechanism alone could account for the latitudinal trend in species richness (Fig. 5) if fewer species have evolved the capacity to endure increasingly lower temperatures. However, this does not account for the absence of low-temperature-tolerant species from warm climates; in addition, low absolute temperatures are also associated with shorter and cooler growing seasons. Therefore, the impact of climatic extremes on richness almost certainly occurs in parallel with other climatic limits, such as described by Runkle (1989; see Section II.B.2).

It is possible to use simple models to investigate the impacts of changes in absolute minimum temperature and competition on distribution and diversity. Figure 12 shows a typical output of a model in which species are designed with the capacity to survive to pre-determined minimum temperatures but have the capacity to survive at all higher temperatures. In a second case, the competitive ability of the species declines as temperatures increase above the low temperature cutoff for survival. In both cases of competition and no competition, richness increases with temperature. However, in the simulation with competition diversity increases less at temperatures above −20°C. This occurs because competition excludes those species with the lowest temperature tolerances from the highest temperatures. This does not occur in the simulation with no competition. Figure 13 illustrates the case for a species with a minimum temperature tolerance to −35°C. Competition exerts no effect on abundance to −30°C but at higher temperatures the abundance or occurrence of this species decreases consistently to zero at −5°C. Without competition there is still a reduction in the occurrence of this species with increasing temperature, and this occurs simply because the chance of occurrence decreases as the pool of species increases with temperature. Figure 12 demonstrates that richness decreases with decreasing temperature; however, the occurrences of the tolerant species increase, in line with the suggestion by Stevens (1989).

The model is simple but it readily predicts changes in diversity with temperature and also indicates that a reduction in richness will be associated with an increase in the abundance of species. Both examples (Figs. 12 and 13) also indicate that the effect of competition is primarily of degree, and therefore it will prove very difficult to differentiate between direct competitive effects and a simple lottery model as both temperature and the species pool increase.

5. History

It appears that competition, disturbance, NPP, and climatic extremes can all influence the way in which climate impacts on diversity. It also seems likely that we
cannot readily differentiate between these controls at the global scale. However, none of the models are capable of explaining the sites which are outliers in the general relationships between richness and latitude (Fig. 5), temperature (Fig. 6), precipitation (Fig. 7) and NPP (Fig. 8). Further analysis indicates that these areas all have high frequencies of endemic species and at the global scale species richness is positively correlated with the degree of endemism. The high degree of endemism in the fynbos of South Africa and kwongan of South West Australia appears to be strongly correlated with the degree of soil infertility. In these areas, richness decreases with increasing soil fertility. Both landscapes, however, are very infertile and ancient in geological terms, and it seems likely that the infertile soil limits the capacity both for alien species to invade and for greater genetic mixing and consequent lower levels of endemism than currently occur. The coastal Atlantic forests of Brazil are isolated from other areas by an inland mountain chain, indicating another method of limiting migration and cross-fertilization and enhancing species richness.

Processes associated with history and regional landscape structure seem to be the most parsimonious explanations for the high species richness in these areas. In the Cape floral kingdom of South Africa, there are more than 500 species in the genus Erica (Bond, 1989). The species are structurally very similar and it seems unlikely that their ecological characteristics are very different. Such situations suggest that the species concept is not an ideal measure of richness, particularly ecological richness. The indicated areas of high diversity (Fig. 5) certainly indicate high taxonomic richness, but it appears that this is hardly relevant to the climatic controls of richness.

III. EXPERIMENTS

Numerous experiments have investigated the relationship between species richness and productivity. However, through practical necessity these experiments have been carried out on a small scale, well within the range of rapidly increasing richness with area as shown in Fig. 1. This scale is not appropriate for the larger scale considerations discussed in this article. Indeed, Tilman (1999), a major contributor to the productivity–diversity debate with Grime (1997), has indicated that this understanding, developed from scales where sites are populated from the same regional species pool, is inappropriate for addressing larger scale latitudinal patterns.

One experimental approach which can, in part, be used to address the relationship between diversity and climate is that of a herbaceous productivity–diversity multi-site experiment distributed across Europe (Hector et al., 1999). Eight sites were distributed between 39° and 64°N of latitude and 8°W and 27°E of longitude.
In all cases, these grassland experiments were seeded from bare with a known number of species, from one to at least eight. At all sites, aboveground biomass, a measure of productivity, increased with species richness, in keeping with the correlation shown in Fig. 8. Although the expected reduction in biomass with short growing season (either through drought or low temperatures) was clearly demonstrated, it was not possible, in a 2-year experiment, to determine the effects of climate on the equilibrium species richness of the plots.

A. Genome Size

In addition to resolving the mechanisms by which climate controls species richness, it is also important to be predictive. For example, given a particular climate, what will be the diversity and what will be the characteristic behavior of the local species? An alternative question is, given a particular metric of a species, where might the species be distributed naturally and in how diverse a mixture of species? These questions cannot be resolved, however, there has been a continued interest in relating the size of the nuclear genome of a species in this way (Grime, 1998). Relationships between genome size and critical climatic tolerances of frost resistance, precipitation supply, and growing season have been found, as have relationships between genome size and plant growth dynamics. This is still clearly an area for further development, but in many ways the attraction to the approach is not the current certainty of its predictive skills but rather the way in which it allows questions to be asked and refuted or otherwise at the local to global scales.

IV. CONCLUSIONS

The impacts of climate on plant biodiversity are best studied on the large scale, in the order of hundreds of kilometers. This approach escapes the insoluble problems of disentangling the effects of scale, landscape structure, and negative small-scale relationships between productivity and diversity. Freed from these constraints, the large-scale nature of the study indicates that observations and correlations are the primary mechanisms for understanding the processes by which climate controls diversity. The processes are still not fully clear, but primary productivity, disturbance, the length of the growing season, and competition all appear to play important roles. There are some locations where species richness is much larger than the average. These areas are rich in endemic species which have probably originated and persisted because of migratory and reproductive barriers. In these areas, it is likely that species differ very little or at all in ecological characteristics, and in the context of this article such diversity, over and above the average, has little meaning.

See Also the Following Articles

Climate Change and Ecology, Synergism of • Competition, Interspecific • Disturbance, Mechanisms of • Energy Flow and Ecosystems • Latitude, Common Trends Within • Species-Area Relationships

Bibliography

EXPOSED OCEAN BEACHES ARE HARSH AND DYNAMIC ENVIRONMENTS controlled by waves, tides, and sediment type. Global patterns of macrofauna biodiversity on sandy beaches are tightly coupled to these physical features of the beach environment and can be reliably predicted on the basis of beach type.

I. INTRODUCTION

The world has entered a biodiversity crisis. The rapid loss of species as a consequence of human alteration of the biosphere has attracted widespread attention among ecologists and spawned a voluminous literature concerning the cataloging, description, prediction, rescue, and management of biodiversity. Marine ecologists have long been interested in explaining patterns in community diversity, especially in the benthos. Much of this effort has been directed at deciphering the role of biological interactions, such as competition, in maintaining diverse communities, and this debate has perhaps been most pronounced amongst workers interested in deep sea-floor and rocky shore environments. However, it has been suggested that competition is relatively unimportant in intertidal sediments (Peterson, 1979), where predation, recruitment, and other processes may exert greater influence on community structure (Reise, 1985).

Despite implicit assumptions of the underlying importance of physical environmental features and processes in influencing diversity in marine benthic com-
munities, it has been much more fashionable to look for biological explanations, especially in competitive interactions. This has perhaps been an oversight and abiotic factors have not been given adequate attention. This article addresses this shortcoming by examining a physically controlled ecosystem and determining to what extent biodiversity can be predicted by physical variables alone.

Sandy beaches dominate the ocean shorelines of all temperate and tropical continental coasts (Davies, 1972). Sandy beaches are devoid of any biological structures and their morphology and dynamics can be defined in terms of three interacting factors: waves, tides, and sand particle size. This simplicity should make understanding of such systems relatively elementary. However, relative to other shore types, beaches have been rather neglected by ecologists.

Here I describe the global range of wave-exposed, sandy beach types that can occur in response to changes in the values of their three defining variables, namely, waves, tides, and sand. I explore the extent to which the diversity and abundance of their intertidal macrobenthic communities can be explained by these factors, and then consider biological and biogeographic issues.

II. SANDY BEACH TYPES

Ocean beaches are defined by the interactions of the wave energy they experience, their tidal regimes, and the nature of the sand available for sorting and transport by the tides and waves. The simplest overall index of beach state is the beach slope, which is a product of the interaction among all three of these variables (Bascom, 1980; Short, 1996): beach face slopes flatten as wave energy increases, tide range increases, or particle size decreases, if other factors are kept constant. Thus the flattest beaches occur in microtidal regions of high wave energy and fine sand, and the steepest beaches occur in macrotidal regions with low wave energy and coarse sand. A range of beach morphodynamic types can be distinguished between these extremes.

In a microtidal regime, where beaches are wave dominated, three beach states can be recognized: reflective, intermediate, and dissipative. The reflective beach, characterized by a steep face and absence of a surf zone, occurs under a combination of coarse sand and gentle waves. The shoreward transport of sand, which occurs under these conditions, causes all sediment to be stored on the subaerial beach face; the reflective beach thus represents the accretionary extreme in beach states. Waves surge up the beach face, where they may break before being reflected back to sea (Fig. 1).

Dissipative beaches, in contrast, are a product of large waves moving over fine sand. This results in a flat beach face and wide surf zone. Waves break far out and dissipate their energy while traversing the surf zone as bores before expiring as swash on the beach face. Dissipative beaches, with their sand spread out over extensive surf zones, thus represent the erosional extreme in beach states. Between these two extremes, intermediate beaches are distinguished by the presence of surf zones that are smaller than in the dissipative situation and generally 20–100 m wide. The intermediate surf zone characteristically has well-developed bars (sandbanks) and channels with rip currents (see Fig. 1).

Beaches are not locked into a single morphodynamic state and respond to changes in wave energy by moving towards dissipative conditions during storms (and spring tides, which are of maximum range and occur during the new and full moons) and towards reflective conditions during calm weather (and neap tides, which are of minimum range and occur during the first and third quarters of the moon); that is, sand erodes or accretes on the beach face as wave height (and tide range) rises or drops.

There is a useful index that conveniently describes the state of a microtidal beach, that is, the extent to which wave energy is dissipated or reflected. Dean's parameter, also known as the dimensionless fall velocity, is given by:

\[
\text{omega} = \frac{\text{wave energy}}{\text{sand fall velocity}}
\]

where wave energy is given by modal breaker height (cm) divided by modal wave period (seconds) and sand fall velocity is the sinking rate (cm per second) of the mean sand particle size on the beach. Values for omega that are <2 generally indicate reflective beaches and values >5 indicate dissipative beaches.

The foregoing description of beach types adequately covers most microtidal situations, but increasing tide range complicates the picture and requires further explanation. Tides play a role essentially similar to that of waves in that increasing tide range tends to make beaches even more dissipative (Short, 1996). This occurs because increasing tide range allows the surf zone to work back and forth over a wider area. Indeed, fully reflective beaches will not occur when tide range exceeds 1–1.5 m. On beaches with larger tides reflective conditions can only occur at the top of the shore between the neap and spring high-water swash lines, an area reached by swash but not surf—this area is con-
COASTAL BEACH ECOSYSTEMS

FIGURE 1 Three morphodynamic states of microtidal beaches (HT = high tide, LT = low tide).

trolled by swash processes and is accretionary and steep, in contrast to the rest of the shore, where shoaling and breaking wave and bore processes operate.

Under large tidal regimes (mean spring range $>4$ m) beaches are generally tide dominated, whereas in intermediate situations (tide range 2–4 m) they are mixed and either waves or tides can dominate. A useful index of the relative importance of waves and tides is the relative tide range (RTR), which is given by the mean spring tide range divided by the modal breaker height. By combining the dimensionless fall velocity and RTR, a two-dimensional model of beach states can be produced (Fig. 2).

Besides beach slope, omega, and RTR, another useful index of beach morphodynamic conditions is the beach state index (BSI; McLachlan et al., 1993). BSI combines measures of wave energy, tide range, and sand fall velocity into one index:

$$BSI = \log(\omega) \times \left( \frac{\text{maximum tide range}}{0.8 \text{ m}} \right)$$

This index is rendered dimensionless by dividing the tide range by an equilibrium tide range of 0.8 m. The BSI has values between 0 and 2; microtidal, reflective beaches score $<0.5$, intermediate to dissipative beaches score $0.7–1.5$, and macrotidal ultradissipative beaches and sand flats generally score $>1.5$.

III. SAMPLING BEACH MACROFAUNA

Community studies of beach macrofauna make use of standard beach transect surveys. These typically involve quantitative sampling across the intertidal zone by excavating quadrats and passing the sand through a screen of 1-mm mesh, the optimum mesh size. Other mesh sizes, for example, 0.5 mm or 2 mm, have also been used and can significantly influence the results. The finer mesh size is more effective in sampling larval stages of macrofauna and can collect some larger meiofauna, but it is not practical in coarser sands.

In the context of this article, each transect survey represents one datum point in the sense that beaches are compared as if they were units. Data should only be used from transect surveys where the total sampling effort (assessed as total area excavated, i.e., levels $\times$
replicates × quadrat size) per beach exceeded 2 m². Smaller sample sizes introduce problems of undersampling, as illustrated by species/area relationships in Fig. 3 (Jaramillo et al., 1995). The absence of an asymptote in Fig. 3 means that total species richness is never sampled, but the curves indicate that, for microtidal beaches a sample area of 3 m² is effective, whereas macrotidal beaches require more. I will use richness rather than diversity indices because it is more conservative. Species richness (or number of species) in beach surveys is summed for all samples in a transect and abundance is usually calculated per running meter of transect, so results are expressed as the number of species per transect and the number of individuals per linear meter of transect.

IV. COMPOSITION AND ZONATION

The intertidal macrofauna of ocean sandy beaches is usually dominated by crustaceans, molluscs, and polychaetes, with other groups, such as insects, nemertean worms, echinoderms, anemones, and fishes, being of minor importance or restricted to the extreme upper or lower fringes. Crustaceans tend to be most successful in reflective conditions, where their great mobility enables them to cope with turbulence. Among the crustaceans, ocypodid crabs, hippid crabs, cirolanid isopods, a variety of amphipods, and psammophilic mysids are...
the most typical. Molluscs, both gastropods and bivalves, are successful over a wide range of beaches. Clams, in particular, may form large populations on high-energy dissipative beaches, where they can support commercial fisheries. Polychaetes tend to be the group that is most sensitive to beach state and are absent or scarce on reflective or coarse sand beaches. On lower-energy shores of fine sand, polychaetes can be particularly abundant and include predator/scavengers and deposit and suspension feeders.

Like all intertidal regions, sandy beaches display zonation of their macrofauna. Various models have been discussed, but the most applicable is that of Erik Dahl (see McLachlan and Jaramillo, 1995). This model identifies three zones on sandy beaches: a supralittoral zone characterized by ocypodid crabs in warm regions and talitrid amphipods in temperate areas; a littoral or midshore zone characterized by cirolanid isopods and spionid and opheliid polychaetes; and a lower shore or sublittoral fringe with many groups, including hippid crabs, mysids, haustorid and phoxocephalid amphipods, donacid clams, and nephryid worms. The lower zone tends to reduce or even disappear in very reflective situations, but it can expand to a broad, species-rich terrace under dissipative conditions. The supralittoral zone, above the high-water mark or drift line, is present on all shores. Thus, on most shores there is a clear gradient of increasing species richness as one moves downshore, typically from one or two species in the supralittoral to many species on the lower shore. This article focuses on between-beach diversity rather than variability within beaches.

V. GENERAL PATTERNS

McLachlan et al. (1981) were the first to demonstrate a relationship between beach type and biodiversity. Based on a survey of beaches around South Africa, they showed that the species richness of benthic macrofauna increased as beaches became flatter and particle size became finer (Fig. 4). They also demonstrated that faunal abundance increased exponentially as particle size decreased and beaches became flatter. At that time they were not aware of the beach state models that were first brought to the attention of ecologists two years later (Short and Wright in McLachlan and Erasmus, 1983). Since the South African coast is uniformly microtidal and subject to vigorous wave action throughout, these findings largely reflect the role of particle size in influencing microtidal beach state. The importance of beach slope, sand grain size, and wave exposure has also been stressed by other authors.

After taking into account the significance of morphodynamic models, a subsequent study that examined a wider range of beaches clearly showed a linear increase in species richness (and an exponential increase in abundance) per transect over a range of beach types from reflective to dissipative (Fig. 5; McLachlan, 1990). Taking this further, McLachlan et al. (1993) showed this trend to be global, based on data from four conti-

![Figure 4](species richness related to beach slope and grain size (in microns); early data from South Africa. (After McLachlan et al., 1981.)}
The first analysis of global patterns of species richness on sandy beaches. Grain size is in microns. (After McLachlan, 1990.)

McLachlan et al. (1996) then examined an even wider data set, including macrotidal beaches, and showed excellent correlation using the BSI index. Further confirmation has come from data sets from Australia (Fig. 7; Hacking, 1997), New Zealand, Brazil, Madagascar, Spain, the United States and Oman.

The foregoing studies have confirmed that, for a diverse range of exposed sandy beach environments on five continents, intertidal macrofauna species richness per transect increases linearly with increasing values of BSI. Thus the number of species accommodated on an intertidal beach increases in response to increasing tide range, increasing wave energy, and decreasing sand particle size. In simple terms, this means that species richness increases as beaches become flatter and wider.

The width of sandy beaches ranges from as little as 10 m in reflective, microtidal situations to hundreds of meters in dissipative tidal flats, thus representing an order of magnitude range in the length of an intertidal transect. The total area sampled by various authors, however, has fallen within a much smaller range. Total
sample area per transect in the studies discussed here falls between 2 m² and 4.5 m², with most cases being around 3 m². The study by Jaramillo et al. (1995) concluded that a total sample area of 3 m² was sufficient to recover 90% of the species on microtidal beaches but would result in undersampling on wider beaches, such as macrotidal flats. This implies that actual richness on the latter beaches is higher than recorded in Figs. 5 and 6 and supports the suggestion of Hacking (1997) that the response of species richness to BSI may not be linear but may rather tend to be exponential over beaches with high BSI values. More work is needed to clarify this.

VI. LATITUDINAL AND OTHER EFFECTS

McLachlan (1990) suggested that, because wave energy is highest in temperate areas, these regions would enjoy a preponderance of dissipative beaches and hence greater species richness. Subsequently, Dexter (1992) proposed that tropical beaches harbor less diverse faunas than temperate beaches. However, this was based on empirical data and did not consider the role of morphodynamics. More recent work comparing beaches of the same morphodynamic state has indicated that tropical beaches may in fact support greater species richness than temperate beaches of the same type; exceptional species richness has been reported for beaches in Queensland and Oman (Fig. 8). Thus, in addition to the three physical factors listed earlier, latitude also appears to play a role in determining species richness. Certainly, tropical beaches do not seem to be poorer in species than temperate beaches of similar types.

A final factor of importance to intertidal beach macrofauna communities may be beach length. In Australia, it has been shown that angling activity and success, and thus intertidal bait organism communities, may become richer as beach length increases; a non-linear relationship between beach length and species richness has been found for Chilean beaches, and studies showing clam abundance increasing with beach length also support this. Experience suggests that short ‘pocket’ beaches are often impoverished and this factor deserves further study.

VII. CAUSATIVE FACTORS

For ocean beaches, the number of species recorded in a single transect survey ranges from 1 to 40, if insects are excluded. Reflective beaches with coarse sand, dynamic swash action, and rapid drainage may be devoid of intertidal species and harbor only supralittoral forms,
whereas more dissipative beaches and flats may harbor 20–40 species at least. Loss of species under reflective beach conditions occurs mainly in the lower intertidal, not the supralittoral zone. It is in the lower intertidal zone where most species are added as the shore flattens towards the dissipative or tidal flat extreme. So the question must be raised: Why are so many species excluded towards the reflective beach extreme?

Considering only between-beach patterns, that is, if whole beaches are compared, the physical environment impinging directly on beach fauna differs primarily in two sets of variables. These two variables are (1) the movement of water over and through the beach face, which has been termed the swash climate, and (2) the sediment particle size range. This ignores the normal intertidal gradient of exposure, which is relevant to any within-beach study.

On most beaches, waves do not break directly in the intertidal but rather in the surf zone. It is only after transformation to bore and crossing the surf zone that wave energy reaches the beach face as swash. Thus, intertidal fauna experience the effects of waves, tides, and the surf zone transformation of these forces as the swash “climate” on the beach face. McLachlan (1990) suggested that it was not the beach state or morphodynamic type itself that was important for the fauna, but rather the swash climate associated with it.

There is a consistent relationship between beach type and swash climate features. Dissipative beaches are characterized by swash with extended periods and lengths, variable speeds, most swash activity below the effluent line (water table outcrop), and fairly laminar swash flow. Reflective beaches display the opposite swash features: the swash climate is extremely harsh; there are high swash speeds throughout the tidal cycle; and waves break directly in the intertidal, resulting in considerable turbulence and increased probability of animals being stranded above the effluent line where unsaturated sand might make burrowing difficult. Swash drainage through the sand is also rapid on reflective beaches and there is little inundation time for feeding. Physical stress in the swash zone on the beach face. Swash patterns and percolation of water through the beach face above the effluent line are closely coupled.

The coupling between beach state and swash climate is clear, but there is no published study undertaking a simultaneous examination of swash climate and fauna. However, recent work in New Zealand shows a strong correlation between community parameters and swash climate. Further, it has been demonstrated that the hallmarks of bivalves that are able to survive on reflective beaches are small size and high density, both adaptations for stability under turbulent conditions. The harsh swash climate on reflective beaches thus appears to select for relatively small size and high body density in bivalves, but may select for larger size and mobility in crustaceans.

It is now clear that swash is not the only physical factor of importance to beach macrofauna; sand particle size also plays a role. Ongoing work shows that small body sizes tend to be excluded in coarse sediments and various authors have demonstrated how sand particle size influences burrowing rate in a variety of sandy beach species—in most cases coarse sand makes burrowing difficult or impossible. Sandy beach isopods, for example, have been shown to select finer substrates. Coarse sand can cause vicious abrasion, especially of bivalve shells, and it reduces sand saturation by raising permeability and hastening drainage.

In the only study to examine faunal community response to swash and sediment factors over a range of each type, Gary Stephenson and I have concluded that both sediment particle size and swash climate directly influence the number of species of macrofauna on wave-exposed sandy beaches in northern New Zealand. Since both factors are closely related to beach type, this explains the good correlations between beach type and macrofaunal species richness. In general, conditions of fine sediment and long swash periods, with limited swash activity above the effluent line, appear most conducive to developing rich faunas. The key characteristic of this type of swash may, in fact, be its low degree of turbulence and laminar flow over fine, saturated sand, enabling even delicate forms to survive. Thus Stephenson and I propose that the swash exclusion hypothesis should be modified to a hypothesis of physical exclusion.
VIII. BIOLOGICAL FACTORS

There is little evidence of biological factors controlling intertidal beach macrofauna. Predation by birds, fishes, and invertebrates is well documented, indeed quantified (Brown and McLachlan, 1990), but it has not been demonstrated to influence species richness. The absence of an underlying base of competition probably precludes this. In such dynamic, three-dimensional environments, with highly mobile faunas, competition is limited and indicated in only a few exceptional cases.

One biological factor that can add to species richness is the input of seaweed wrack to a beach, if the input is not so great as to cause disturbance, smothering, and deoxygenation. In cases of moderate input, many wrack-associated species may be added to the usual beach fauna. A well-developed drift line can also add many insects. Progressing along the gradient from reflective, wave-dominated beaches to macrotidal flats where tidal factors are more important than waves, increasing stability of the substrate leads to the formation of permanent burrows and more complex communities with greater scope for biological interactions. At some point, as yet not defined, in the continuum between beaches and tidal flats, biological factors are likely to become more important than outlined here for wave-exposed beaches.

IX. DISCUSSION

Three paradigms provide perspective for our current understanding of the factors controlling the large-scale patterns of species richness of sandy beach intertidal benthic macrofauna.

The Autecological Hypothesis. The autecological hypothesis of I. Noy-Meir states that inhabitants of physically controlled environments respond independently to the physical environment. This implies that the presence or absence of a species on a beach will be little influenced by biological factors.

Beach Morphodynamic Models. Beach morphodynamic states can be defined by three physical factors: wave energy, tide range, and sand particles size. Increases in wave and tide energy and decreases in sand particle size all result in beaches becoming wider, flatter, and more dissipative.

Richness and Abundance Coupled to Beach State. Intertidal sandy beach macrofauna increase linearly (probably) in species richness and exponentially in abundance in response to the increasing dissipativeness of beaches, that is, from macrotidal reflective systems through high-energy dissipative systems to ultradissipative tidal flats.

Beyond these paradigms, two further hypotheses can be erected to explain the known patterns of sandy beach macrofauna biodiversity:

Latitude and Beach Length. Tropical beaches support more species than do temperate beaches of the same type, and long beaches appear to support greater abundance and more species than do short (or pocket) beaches. The former trend may be due to the greater species pool available to colonize beaches in the tropics, whereas the latter trend reflects the more favorable surf circulation patterns on long beaches, which promote greater retention of particulate primary production, better recruitment of planktonic larvae, and larger and therefore more resilient adult populations. Long beaches are also more persistent, that is, unlikely to erode away completely during storms, as can happen on pocket beaches.

Swash and Sand Control. Two physical factors define the immediate environment experienced by the beach macrofauna and may therefore control species richness: the swash climate and sand particle size. Reflective beaches have harsh swash climates in the sense of high turbulence, short swash periods, and rapid swash drainage, resulting in low effluent lines; beaches at the other extreme have more benign swash climates in the form of long-period swash/tidal bores with laminar flow over the beach face and low turbulence, keeping much of the beach saturated because of the slow drainage. All beach fauna can burrow into saturated sand and, in terms of grain size, fine to medium sands seem optimal for most beach fauna. Coarse sand appears to exclude small or delicate forms by crushing and abrasion and most species experience decreasing burrowing efficiency in coarse sands. Thus harsh swash climates and coarse sand associated with reflective beaches appear to exclude many species. The degree of exclusion of species by these factors will increase from the dissipative/macrotidal to the reflective/microtidal end of the continuum of beach morphodynamic types.

It has been stated that community structure in harsh
environments is in nonequilibrium, characterized by the decoupling of loose biotic interactions and the independent response of species to environmental variations. Thus populations and communities in stressful environments may respond primarily to the abiotic environment, that is, be highly stochastic and thus loosely structured. This is a restatement of the idea of a physically controlled community and the autecological hypothesis and appears to be highly relevant to the case of sandy beaches. The dynamic, three-dimensional nature of wave-exposed ocean beaches, their shifting populations, and the absence of biological structures (they are even devoid of permanent burrows in most cases) leave limited scope for biological controls of community structure. Furthermore, the evidence presented here has shown that an index based only on the three physical variables that define a beach (tides, waves, and sand) has high predictive power to estimate the species richness likely to be encountered on ocean beaches. Latitude, beach length, and biological interactions may contribute additionally, to a small extent, to explain some of the variability in the data, but their roles have yet to be conclusively demonstrated in high-energy beach environments.

Several lines of research could improve our understanding of patterns of sandy beach biodiversity and their control.

1. The BSI is a simple index developed partly by ecologists and fitted to a few data sets. Collaborative work between ecologists and geomorphologists, using all the data available at present and investigating new regions, could lead to refinement of this to an even more appropriate index (or indices) with greater predictive power.

2. Comparative work contrasting similar beaches in the tropics and temperate latitudes is needed to confirm conclusively whether tropical beaches do indeed support higher diversity and what controls this.

3. Once these processes are better understood, useful predictions could be made of global beach diversity patterns based purely on the global distribution of tide and wave regimes and sedimentology. This capability would be useful in predicting and identifying areas of high diversity and in planning coastal conservation strategies.

X. CONCLUSIONS

In the case of wave-exposed, ocean sandy beaches, the species richness of the intertidal (and immediately sublittoral) macrobenthic fauna is strongly controlled by, and predictable on the basis of, the physical nature of individual beach environments rather than global biogeographic patterns. Beaches that tend toward the ultra dissipative/tidal flat extreme—the product of large tide ranges, vigorous wave action, and fine sand—harbor most sandy beach species likely to be encountered in a region. Progressing from this end of the beach continuum towards reflective conditions, that is, decreasing tide range or wave energy or increasing sand particle size, results in beaches becoming steeper and narrower and the increasing exclusion of species. On highly reflective beaches, intertidal macrofauna may be absent and only supralittoral forms remain. These relationships are illustrated in simplified form in Fig. 9.
The mechanisms by which species are increasingly excluded as one moves towards reflective conditions appear to relate primarily to two environmental variables, the swash climate and sand particle size. Decreasing swash period, increasing swash turbulence, lower effluent lines, and coarse sand exclude delicate, less robust, and slower-moving species. At present our knowledge of how biogeographic factors influence species richness on sandy beaches is limited, but it appears that tropical beaches may support more species than do temperate beaches of similar morphodynamic type. There is considerable scope for further research to clarify these patterns and to decipher cause and effect relationships.

See Also the Following Articles
ESTUARINE ECOSYSTEMS • INTERTIDAL ECOSYSTEMS • LAKE AND POND ECOSYSTEMS • MARINE ECOSYSTEMS • RIVER ECOSYSTEMS

Bibliography

COEVOLUTION

Douglas J. Futuyma and André Levy
State University of New York at Stony Brook

I. Meaning and Varieties of Coevolution
II. Evidence of Coevolution
III. Competition between Species
IV. Responses to Common Predators
V. Enemy–Victim Interactions
VI. Special Features in Parasite–Host Coevolution
VII. Special Features in Herbivore–Plant Coevolution
VIII. Mutualism
IX. Major Consequences of Coevolution

GLOSSARY
character displacement Pattern of geographic variation in which a character differs more between sympatric than between allopatric populations of two species.
compeition Interaction between individuals of the same or different species, whereby resources (e.g., food, space, mating partners) used by one are made unavailable to others.
mutualism Symbiotic relation in which each of two species benefits from the interaction.
resistance An individual's capacity to reduce the damage inflicted upon it by an enemy.
specialization Evolutionary adaptation in a particular mode of life or habitat.
tolerance An individual's capacity to sustain damage by an enemy with limited decrease in fitness.
trade-off Negative correlation between traits, such that a benefit due to changes in the value of one trait is associated with a cost produced by changes in the value in another trait.
virulence Degree to which a parasite reduces the probability of survival or the reproductive capacity of an individual host. A relatively avirulent (benign) parasite has little impact on its host's fitness.

ALTHOUGH THE TERM COEVOLUTION is sometimes used to describe the joint evolution of different genes or characteristics within a species, it usually refers to evolutionary changes in two or more different species owing to ecological interactions between them. These interactions include competition, mutualism, and 'enemy/victim' interactions between predators and prey, parasites and hosts, and herbivores and plants. Even in this restricted sense—the subject of this article—coevolution has several meanings.

I. MEANING AND VARIETIES OF COEVOLUTION

(1) Patterns of correlated evolution, especially phylogenetic congruence, among species or other lineages are usually detected by phylogenetic analysis. In extreme cases, the phylogeny of one clade of species, such as
host-specific parasites, is congruent with that of a clade with which it interacts such as host species (Fig. 1a). Such congruence would require coincident speciation of hosts and associated parasites. Deviations from perfect congruence might be caused by speciation of a lineage but not its associate, by extinction of one but not the other, or by a parasite switching from one host to another (colonization) (Fig. 1b).

In some cases, the coincident speciation that underlies concordant phylogenies may be caused by the interaction between associated species. However, concordance might stem solely from a shared history of geographic isolation, without the interaction having played any causal role. Although phylogenetic concordance does not in itself provide evidence of reciprocal adaptation of the interacting species, it does imply extended historical opportunity for adaptation to have occurred. The greatest phylogenetic concordance has been described for some maternally transmitted endosymbiotic bacteria and their eukaryote hosts, and for certain ectoparasites (e.g., lice of gophers) that are primarily transmitted by contact among conspecific hosts. In many groups of parasites and herbivorous insects, related species are associated with related hosts, that is, members of a clade, but the phylogenies of the two clades are not congruent. This pattern often implies a long history of association, but one that has included shifts of parasite lineages among host lineages.

(2) Coevolution consisting of reciprocal adaptive responses has been rather arbitrarily classified as specific (or pairwise, or tight) coevolution and diffuse (or guild) coevolution. Specific coevolution is reciprocal adaptation of two species, independent of their interactions with other species, whereas diffuse coevolution occurs when evolutionary change in one species affects its interaction with two or more other species (e.g., a change in a host that affects two parasite species). Specific coevolution is easier to visualize, model, and detect. Diffuse coevolution is most likely to occur when a species’ response to selection by one interacting species is accompanied by genetically correlated effects on its interactions with other species. The term “diffuse coevolution” has also been used to describe nonadditive selection on one species by two or more other species, owing to an interaction between them or between their effects on the recipient species.

(3) Cospeciation may be induced by the interaction between species, but this probably requires special conditions. A likely case is in figs (Ficus) and the highly host-specific seed-eating wasps that pollinate them. Divergence between population of either a fig species or a wasp species in features governing the host preference of the wasp may engender divergence in the fig, resulting in reproductively isolated populations of both. Such cospeciation may result in concordant phylogenies.

(4) Escape-and-radiate coevolution describes a scenario in which an evolutionary lineage diversifies after it evolves a defense that breaks its association with enemies. Such a scenario was first proposed by P. R.
Ehrlich and P. H. Raven for plants and herbivorous insects. In this hypothesis, a plant lineage that evolves a novel defense sheds most of its herbivores, and therefore (by an unspecified causal inference) diversifies into many species. Only later do some herbivore species adapt to one or more of the plant species, and diversify in turn. Collectively, these hypothetical scenarios suggest that coevolution may increase both the numbers of species and the phenotypic diversity (disparity) of species by selecting for the many characteristics that affect their interactions. Coevolution may therefore be an important engine of biological diversity, but we shall shortly note that coevolution may also cause extinction. Often, the adaptation of one species to others is obvious, but reciprocal adaptation of the others is not. Because we seldom know a priori whether or not adaptation has been reciprocal, adaptations to ecological interactions are often loosely referred to as "coevolutionary," even if evidence for reciprocity is slight or lacking. Certainly the study of coevolution includes adaptations to other species, which may or may not prove to have adapted in turn.

II. EVIDENCE OF COEVOLUTION

A. Sources of Evidence for Coevolution

The study of coevolution comprises the same approaches as studying evolution in general. As in the broader field, the first forms of evidence for coevolution consisted of detailed natural history observations, descriptions of the diversity of adaptive structures that mediate ecological interactions, and comparison among populations and species. Charles Darwin published the first comprehensive illustration of how intrinsically species are adapted to one another and how structural complexity can be explained by coevolution in his description The Various Contrivances By Which Orchids Are Fertilized By Insects (1877). By comparing the shape of different orchid flowers and their associated pollinators, Darwin demonstrated that certain moth features are needed to successfully obtain nectar from the flower, features that are indeed exhibited by their specialized pollinators. By restricting nectar collection to a few pollinators, orchids increase the likelihood of cross-fertilization. Such is the correspondence between flower and pollinator shape that observing the 29-cm-long nectar-bearing spur of the Madagascan orchid Angraecum sesquipedale, Darwin predicted the existence of a pollinating moth with a proboscis of that length. Such a moth, Xanthopan morgani praedicta, was indeed discovered 40 years later. The description of patterns in plant use by lepidopteran larvae preceded the concept of escape-and-radiate coevolution. Higher taxa of butterflies often feed upon a single group of flowering plants. While some feed on more than one plant family, these tend to be closely related or have similar biochemistries. For instance, larvae of the butterfly subfamily Pierinae, or whites, feed predominantly on the families Capparaceae and Brassicaceae, which are closely related. Some whites also feed on members of the family Tropaealaceae that share with the other families the production of mustard oil glycosides and a rare fatty acid. These regularities imply an important role for plant secondary metabolites in determining butterfly host use. Given that these compounds affect herbivore behavior, acting often as deterrents, secondary chemistry may have constituted the key feature that allowed plant escape. Comparisons among conspecific populations have also been suggestive of coevolution. The coloration pattern of the butterfly Heliconius erato, thought to be a signal to predators indicating distastefulness, varies among populations in Central and South America. Strikingly, the wing coloration of H. melomene, an equally distasteful congener with distinct life history and host preference, varies geographically in parallel with H. erato. This pattern is thought to be an example of coevolution of mimicry between prey species that share a predator. Fritz Müller, a contemporary of Darwin, first suggested this particular model of coevolution to explain similarities in wing pattern among unpalatable butterfly species belonging to two distinct genera (Iulina and Thyridia). Müller also introduced the use of mathematical models to study the coevolutionary process. Modern mathematical and computer simulation models may incorporate population genetics, quantitative genetics, evolutionary game theory, and optimality theory. Mathematical modeling has proven useful in describing the dynamic of the interactions between species and in determining which conditions favor coevolution.

Although ecological interactions usually do not "fossilize," the analysis of paleontological records has provided some evidence of coevolution. For instance, the appearance in the Ordovician of predaceous cephalopods is associated with the simultaneous appearance of several defensive strategies on the part of their prey (e.g., strong sculpture and coiling in gastropods and shell-bearing cephalopods, spines in echinoderms), suggestive of diffuse coevolution between predators and their prey. The antiquity of certain
interactions may also be determined by inspecting fossils of extant species. Several plant families possess structures (domatia) that harbor mites, which attack plant enemies. Domatia similar to the modern form have been discovered in fossilized leaves from the Eocene, 35 million years ago!

The age of an association, or of interaction-related adaptations, can often be estimated from phylogenies with time calibration (e.g., using approximate molecular clocks or stem-group fossils). Molecular evidence from deep-sea vesicomyid clams and the sulfur-oxidizing endosymbiotic bacteria on which they depend for nourishment indicates that the interacting clades are both approximately 100 million years old. These two lineages appear to have been in close association since their origin and to have cosepeciated, as indicated by the remarkable level of congruence between their estimated phylogenies. Phylogenetic information also becomes relevant in testing whether a character is an adaptation for an ecological interaction or an ancestral feature that exists in the absence of the interaction.

In some instances it has been possible to document the particular genes that affect a species' interaction. H.H. Flor found several genes in flax (Linum usitatissimum) that provide resistance to the rust Melampsora lini. Rust virulence is determined by a set of complementary genes, in a one-to-one relationship. This study inspired the gene-for-gene model (see Section VI,A), which has become a paradigm of phytopathology. Most traits, however, have a complex genetic basis, involving many genes. Such complexity requires a quantitative genetic approach, which partitions the trait variation into genetic and environmental components. This approach has demonstrated that many of the traits relevant to interactions have genetic variability, that is, there is potential for coevolution. For example, the wild parsnip (Pastinaca sativa) and its most important associated herbivore, the parsnip webworm (Depressaria pastinacella), are thought to be engaged in coevolution mediated by the evolution of furanocoumarins and the insect's detoxifying mechanisms. May R. Berenbaum has documented genetic variation both in the production of furanocoumarins and in the webworm's ability to metabolize this group of plant toxins.

Quantitative genetics is also used to measure correlations between traits. The detection of negative genetic correlations is indicative of trade-offs between traits, such that selection for the increase in value of one trait leads to a decrease in value of the correlated trait. Trade-offs are of particular relevance in explaining evolutionary constraints, and particularly why species are specialized. Pea aphid (Acyrthosiphon pisum) clones collected from two crop plants (alfalfa and red clover) exhibited higher fitness when reared on the plant from which they had been collected, suggesting local adaptation. The negative genetic correlation in fitness across crops may constrain the evolution of generalist clones, as these would be outcompeted on either plant by crop-specialized clones.

Measurement of genetic variation and correlation between traits offers information on the genetic context in which selection can act. Correlations between traits and fitness suggest the form and direction of selection. In a greenhouse study, the wild parsnip exhibited a negative genetic correlation between concentration of several furanocoumarins and seed set, suggesting that the production of the chemicals may impose a cost to reproduction in the absence of the parsnip webworm. These negative correlations were not detected in the field, indicating that presence of furanocoumarins increases fitness in the presence of the herbivore. Ideally such studies are performed in a natural setting, as our ultimate interest is understanding how natural selection works in the wild, but one can use model systems in the laboratory, such as evolving populations of bacteria and bacteriophage.

Finally, studies of interacting species are commonly based on the analysis of single communities. However, most species are composed of many local populations, and increasing importance has been attributed to the geographical structure of species and their interactions. Across the distribution of an interaction one is likely to observe a mosaic of selection pressures as a result of variation in abiotic and biotic factors, and the particular demographic and genetic histories of local populations. Some localities may be coevolutionary hot spots, that is, sites of reciprocal coevolution, whereas in others selection may be unidirectional or act on neither species. The geographic variation in outcomes is further modified by gene flow among populations. Consequently, different degrees of coadaptation are to be expected among populations. Clearly our understanding of the dynamics of a species interaction requires the study of many communities and interpopulation processes.

A few cases that satisfy the requirements of long-term multipopulation studies have emerged recently and have reinforced our need for a geographic mosaic theory of coevolution. For instance, resistance and virulence structures of Linum and Melampsora (referred to
earlier), studied in New South Wales, sometimes vary dramatically across populations and time. The frequency of susceptible genotypes of flax will affect the local frequency of a particular strain of flax rust, but additional factors were found to play a role, namely, drift, extinction, and migration from neighboring populations. The geographic structure of flax and rust proved to be an essential factor in explaining the persistence of the interaction.

### B. Rates of Adaptation

Phylogenetic reconstructions and the fossil record provide some examples of ancient interactions between groups of organisms. Character mapping of feeding habits onto beetle phylogeny indicates that conifers and cycads are the ancestral host plants. The oldest fossils of herbivorous beetle taxa, from the Jurassic, are indeed ancestral forms of modern conifer and cycad feeders, suggesting that this association may have evolved 200 million years ago, making it the oldest insect/plant association known. Lichens, which are an intimate association between fungi and green algae, may be 100–200 million years old. Yet how often do new interactions establish themselves, how rapidly do species adapt to one another, and does that process occur on a timescale that is relevant to ecological processes?

The evolution of interspecific interactions can occur very rapidly, that is, during the course of 100 years. Examples include cases where recorded changes in the environment led to measurable changes in phenotype. In 1977, the Galapagos Islands suffered a severe drought during an El Niño event, causing an increase in the median size of available seeds. During the same event, the medium ground finch (Geospiza fortis) suffered strong selection for an increase in bill size, as only larger-billed birds could effectively feed on the available large seeds. This morphological shift occurred in a single generation. Many cases involve species of animals and plants introduced into new habitats. Populations of the apple maggot (Rhagoletis pomonella) that fed exclusively on native hawthorn have in the course of the past 100 years become adapted to a novel host, the introduced apple. The loss of interaction-related adaptations has also been observed to occur on a short timescale. Within the course of a 34-year study, guppies (Poecilia reticulata) artificially introduced into drainages in Trinidad, where natural predators are absent, lost defensive antipredator behaviors exhibited in the source populations. Although these examples unequivocally demonstrate that interspecific interactions can evolve rapidly, they do not provide examples of coevolution. Almost all examples describe the evolution of one species adapting to another, with little reference to reciprocal changes. One of the few exceptions is the effort to biologically control the rabbit (Oryctolagus cuniculus) in Australia. The introduction of the myxoma virus into Australia initially reduced the rabbit population drastically. Such strong selection led to the evolution in just a few generations of increased resistance in rabbits. This evolutionary change was accompanied by a decline in the virulence of the virus to an intermediate level. Although in this case evolutionary change was observed in both interacting species, this does not necessarily constitute a case of coevolution. The reduction in virulence did not occur as a result of changes in rabbit resistance, but rather because the initial strains of the virus were so virulent that they quickly killed the host, lowering the likelihood of transmission by the disease’s vector. More recent data, however, seem to indicate that the initially more virulent strains are presently increasing in frequency, perhaps as a result of increased rabbit resistance.

Rapid evolution is not, however, unconstrained. Each species is influenced by its phylogenetic history, and even populations of the same species will possess different evolutionary potential. To grasp how often new interactions are formed, one needs to understand these constraints. Rapidly evolving interspecific interactions can potentially affect the dynamics of community structure. Models of community ecology should not rest on the assumption that evolutionary dynamics occur on a larger timescale that is irrelevant to ecological processes. But it has yet to be determined under which conditions coevolution drives community dynamics and whether it has a stabilizing role.

### III. Competition Between Species

Hypotheses about the coevolution of competing species have long played a major role in both evolutionary biology and ecology. Most of the hypotheses are variations on the theme that species should evolve differences in resource utilization, thus reducing competition between them. For Darwin, the principle of “divergence of character” in response to competition, described in *The Origin of Species*, explained why an ancestral species should give rise to multiple, diverse descendant species. W. L. Brown and E. O. Wilson cited reduction of competition as a major cause of “character displacement,” the term they coined for a greater difference between sympatric than allopatric populations of two species. D.

---

**COEVOLUTION** 757
Lack and R. H. MacArthur provided famous examples of differences in diet and foraging behavior or trophic morphology among closely related species of birds, interpreting the patterns as stemming from avoidance of competition. G. E. Hutchinson cited instances of apparently constant ratios of the dimensions of trophic structures among sympatric congeners. MacArthur and R. Levins elaborated on this theme, developing a theory of limiting similarity that concerned the expected degree of niche differentiation among sets of potentially competing species, as part of a theory of the evolution of community structure that has had lasting influence. The importance of competition and of coevolution of competitors was the subject of vigorous debate in the 1970s and 1980s.

The most detailed recent models of coevolution of competitors (Taper and Case, 1992) assume that the size of a phenotypic feature, such as body size or beak depth, is correlated with the mean food type (e.g., prey size) that an individual uses. The evolution of the character, and thus of resource use, is modeled using quantitative genetic models for polygenic traits. (The character is approximately normally distributed, with specified additive genetic variance and heritability.) It is assumed that “a jack of all trades is master of none,” that is, that there exist trade-offs in adaptation so that each phenotype performs effectively over only a relatively small range of resources (its within-phenotype niche width, WPNW), and that a specialized phenotype (with smaller WPNW) is more efficient than a generalist. (For example, studies of finches have shown that the rate of energy intake from smaller versus larger seeds is greater for small-billed versus large-billed birds, respectively.) In the simpler models, the competitive effects of phenotypes on each other are symmetrical and are proportional to the degree of resource overlap, whether the phenotypes be the same or different species. Thus competition is more intense when two phenotypes are more similar.

The size–frequency distribution of resources is assumed to be unimodal. Thus the rarity of resources in the tails of the distribution selects against extreme phenotypes. Consequently, the mean phenotype of a solitary species evolves to match the peak of the resource curve. Frequency-dependent selection maintains phenotypic variance among sets of phenotypes remains approximately normal. It does not necessarily fit the frequency distribution of resources, some of which may be underutilized.

When applied to two or more species, these models generally support earlier suppositions. In each species, phenotypes that do not bear the combined burden of interspecific and intraspecific competition have higher fitness. Hence, the species evolve so that their phenotypic means are spaced apart (character displacement). Moreover, the phenotypic variance within each species at equilibrium is somewhat smaller than in a solitary species. At equilibrium, some overlap remains in the species’ resource utilization, for otherwise abundant food items of intermediate size would be left unused, creating selection for genotypes that can use them. Under some conditions, especially if two species initially overlap greatly in resource use, they may converge. Competitive exclusion, resulting in extinction, then becomes more likely. Moreover, if competitive effects are asymmetrical (e.g., if larger individuals reduce the fitness of smaller ones more than the converse), a species may converge toward another and “chase” it to extinction.

Facile invocation of evolutionary responses to competition to explain ecological patterns was severely criticized in the 1980s, resulting in more critical evaluation of evidence. Even under the new scrutiny, many data strongly support the theory of coevolution of competitors. For example, character displacement has been documented in many instances, such as sticklebacks in British Columbian postglacial lakes. Some lakes harbor two species, recently derived from a common ancestor, that differ in microhabitat and in morphological features associated with feeding. Other lakes have only one species, with intermediate behavior and morphology. Closely related sympatric species often differ more consistently than would be expected by chance. For instance, body size in bird-eating hawks (Accipiter) is correlated with average prey size, and throughout the world sympatric species of these hawks differ more in size than do random pairs of Accipiter species.

Such nonrandomness of sympatric assemblages might arise from either coevolution or ecological assembly (colonization followed by extinction of excessively similar species). Phylogenetic analyses of anole lizards in the West Indies have provided evidence of both processes. Many islands in the Lesser Antilles have both a large and a small species. These form distinct clades, so there is no evidence that the size of any one of these species has evolved in response to a sympatric congener. But each of the Greater Antilles has a monophyletic group of species that differ ecologically and morphologically from each other in a parallel pattern that is almost the same on each island. For example, each island has at least one species of “trunk anole,”
one “twig anole,” and one “crown giant.” Coevolution has taken much the same historical course in each case. Release from competition is thought to have important evolutionary consequences. “Ecological release” at a microevolutionary level is illustrated by cases in which only a single member of a genus occurs in a region, and there occupies a broader ecological niche than do species that coexist with congeners elsewhere. The single species of finch on isolated Cocos Island, northeast of the Galapagos archipelago, feeds on a much wider variety of items than does any of its many Galapagos relatives, and the sexual dimorphism in beak size and foraging mode is greater in the Hopiisanian woodpecker than in its continental relatives that are sympatric with other woodpecker species. At a macroevolutionary level, the spectacular diversification of African lake cichlids in the virtual absence of other fishes, the explosive diversification of mammals after the K-T mass extinction of large reptiles, the flowering of modern turtles after the extinction of primitive amphichelydian turtles, and many other examples are thought to show how diversification can be released when incumbent competition is alleviated.

IV. RESPONSES TO COMMON PREDATORS

A. Escape Space and Divergent Defenses

Competition is not the only possible explanation of resource partitioning and community assembly. For instance, species may interact indirectly via a shared predator. Predator density is highest where prey species coexist, and consequently a particular prey species will be maintained at lower density if it coexists with other prey species than if it is the sole victim. This form of predator-mediated apparent competition creates a pattern identical to that of direct competition. Coexistence is possible if prey species partition the resource they compete for, in this case an ecological space of reduced predation, known as an escape or enemy-free space.

Predators tend to increase their search intensity with increasing densities of prey. Search behavior can be improved by specialization in a foraging strategy, for instance, by attacking prey of a certain shape (i.e., similar search image) or prey that occupy particular sites, regardless of whether they belong to the same or different species. If faced with a diversity of prey patterns, they are most likely to form a search image for the most common pattern. This has been demonstrated for many bird and mammalian predators that use primarily visual cues. For example, if wild thrushes are presented with different combinations of two morphs of the snail Cepaea nemoralis, they will invariably preferentially attack the most common morph. The rarer individuals that differ from the ecological characteristics that the predator is attuned to will consequently experience lower predation rates. Such frequency-dependent selection within species, or apostatic selection, could lead to the divergence among prey species and an eventual increase in the diversity of forms and coloration patterns, that is, aspect diversity.

Evidence for this process is mostly circumstantial. R. E. Ricklefs and K. O’Rourke compared morphology, color, and behavioral characteristics of the moth species of a tropical and two temperate communities. Despite a higher number of moth species in the tropical site, the average similarity among moths within a community was the same in all three sites. They suggested that predation influences competition among moths for “escape space,” limiting the similarity of appearance in different moth communities. Cepaea nemoralis and C. hortensis are both polymorphic for shell coloration patterns and can profit from this aversion. Monarch butterfly larvae, Danaus plexippus, feed almost exclusively on milkweed, from which they sequester cardiac glycosides. These toxic compounds are retained in the adult butterfly, Limenitis archippus. Cleaner fish provide a variation on Batesian mimicry. In coral reefs in the Pacific, many fish allow cleaner fish, such as the sea blenny (Aspidontus taeniatus), to feed on parasites on their bodies and even in the interior of their mouths. The sabre-toothed blenny (Aspidontus taeniatus) mimics the white-and-black-striped coloration and swimming pattern of Labroides. By taking advantage of the passive behavior of fish toward the model, it is able to approach fish and bite off pieces of tissue. Labroides and Aspidon-
Mullerian mimicry refers to the convergence toward a similar pattern among unpalatable species. Faced with several undesirable species that look alike, a predator must learn a lower number of patterns to avoid. Evolution in all prey species leads toward a common pattern, and so warrants the designation of coevolution. One of the most striking cases of Mullerian mimicry, mentioned earlier, is the convergence between the neotropical butterflies Heliconius erato and H. melpomene. Despite differences in life history, these species share a common wing color pattern that varies geographically in parallel. One of the species, H. erato, is usually the most abundant where both species co-occur, raising the possibility that parallel evolution occurred by mere convergence of the rarer H. melpomene toward a common model. However, comparison between sympatric and allopatric populations of H. erato in Central America revealed that the width of the H. erato yellow hind-wing bar converges upon that of H. melpomene when in sympatry, suggesting that both species converge on each other.

V. ENEMY–VICTIM INTERACTIONS

The general heading of enemy–victim interaction includes a variety of antagonistic interactions, ranging from those in which the enemy (or predator) usually kills several individuals (prey) to sustain itself, to scenarios in which the enemy (or parasite) reduces its negative effects to one victim (host) during most of its lifetime. The interaction between plants and herbivores is multifurcated, depending strongly on the herbivore taxon, and does not conform nearly to the previous division. Herbivores can be more akin to predators or parasites depending on whether they consume several individuals, as do seed harvesters, or complete their life cycle on a single plant, as do many leaf-feeding insects. They may also be quite unlike either category, as are many grazers and browsers that feed on parts of many individual plants without necessarily killing any of them. Differences in the degree of specialization of the enemy and victim may enhance the asymmetry of the interaction. Most plants, for instance, are attacked by numerous herbivores and parasites, yet many of their enemies are relatively specialized on particular plant taxa. These distinctions and characterizations are relevant as the nature of the interaction determines the selection pressures imposed on both species, and the resulting coevolutionary process.

Enemy–victim coevolution has been envisioned as an arms race in which the exploiter evolves offensive traits that increase the strength (i.e., frequency, intensity) of the interaction, and the victim evolves counterdefenses to decrease it. In predator–prey interactions, for instance, one would focus on traits in that in some way affect the predator’s functional response, that is, its rate of predation. This includes everything from the predator’s detection ability and pursuit speed to the prey’s aposematic or cryptic coloration and escape speed. An evolutionary arms race would correspond, for instance, to a continuous improvement of a prey’s escape speed and the predator’s pursuit speed.

Clearly there are limits to such continuous escalation. There are physical limitations to improving an organism’s features, for example, to how fast an animal can run. Furthermore, investment in an interaction-related trait may imply costs. Development of the trait may require resources that could otherwise be invested in reproduction and other functions. For instance, elaborated morphological defenses (e.g., a thick shell or thorns) may imply an allocation cost expressed as a slower growth rate. Chemical defenses can act as feeding deterrents, but at high concentrations they may be autotoxic, implying a physiological cost. Additionally, some chemicals that act as deterrents toward generalists attract specialist enemies. Ecological costs produced by such genetic correlations limit a victim’s potential to adapt to its complete array of enemies. Traits tend to affect more than one function of the organism, and a compromise must be reached between its role in the different functions. For instance, Geerat Vermeij suggests that such an adaptational dilemma could have been responsible for the extinction of most ammonite cephalopods. The external shell of cephalopods, other than providing passive protection, also affects the speed of locomotion and the ability to compensate for changes in pressure during vertical movements. Toward the end of the Cretaceous, as predation and competition became more intense, no further improvement in these functions could be reconciled by evolution of the shell.

Theoretical work has shown that the existence of costs and density dependence (the effect of a species’ density on fitness of an individual of that species or of an interacting species) can hamper the occurrence of continuous reciprocal responses implicit in an arms race. Neither the prey nor the predator population is expected to always increase its investment in predation-
related traits as a response to an increased investment by its partner. On the one hand, depending on the shape of the predator’s cost–benefit function, the benefits reaped by an individual predator by investing more in predation can be insufficient to compensate for any additional costs. The predator’s capture rate may remain unaltered if the evolution of greater prey elusiveness is accompanied by an increase in prey density. On the other hand, models have shown that the evolution of greater prey-capture abilities can lead to a decrease in the predator’s equilibrium population size as a result of overexploitation of prey (Abrams, 1990). A smaller predator population can imply reduced selection for antipredator traits, and if the costs of defense become a more important selective factor than predation, the prey population may evolve a lower level of defense. So rather than a protracted arms race, one might expect enemy and victim to coevolve toward an intermediate stable state.

The models referred to previously focused on character values and population-dynamic parameters. A different class of models focuses on the genetics that underlie interaction-related characters, while usually ignoring changes in population densities. For example, gene–gene models of host–parasite interactions examine changes in genotypic frequencies at a virulence locus in the parasite and at a resistance locus in the host. Coevolutionary cycles are expected under a broad range of conditions. As the frequency of a virulent genotype increases, selection for resistant host genotypes becomes more intense. Frequency-dependent selection will generate fluctuations in genotypic frequencies, and both populations will remain polymorphic. Some polygenic models also produce stability, as either stable limit cycles or equilibrium points, particularly if the additive genetic variances of the characters are high in both interacting populations. Stability is most likely if the victim has a larger genetic variance and is under stronger stabilizing selection than the enemy is. A higher additive genetic variance allows the prey to respond to selection more rapidly than the predator.

The “Red Queen” model extends the notion of evolutionary arms race to the level of a community. According to these models, evolution of co-occurring species can lead to the continuous deterioration of a species’ environment, forcing it to constantly evolve just to avoid extinction. The name owes its origin to the character in Lewis Carroll’s Through the Looking Glass who explained to Alice that one must run as fast as possible just to stay in place. Analogously, a community at evolutionary equilibrium determined by the Red Queen principle contains a set of interacting species continually coevolving at rates that exactly balance each other. It has been hypothesized that Red Queen coevolution has selected for recombination and sexual reproduction, as only those populations with available genetic variation could have sustained continuous coevolution for extended periods.

The most plausible models of enemy–victim coevolution predict either evolution to an equilibrium or an oscillatory coevolutionary “chase” (rather than indefinite escalation of offensive and defensive properties). The clearest evidence for these models’ predictions should come from direct observation of long-term coevolution, which to date has been possible only with laboratory cultures of organisms with very short generations, such as bacteria and virulent bacteriophage. The most extensive coevolution observed in such experiments consists of fixation of a resistance mutation in the bacterial population, then of a countervailing “virulence” mutation in the phage, and finally of a second resistance mutation in the bacteria, resulting in stable coexistence and an apparent genetic equilibrium. The resistance mutations of the bacteria carried costs that reduced fitness in the absence of phage, but evolution at modifier loci later reduced the cost. The failure of the phage population to coevolve greater virulence was attributed to architectural constraints on these very simple organisms.

Although data from natural populations reveal little about the long-term dynamics of coevolution, they do show that it occurs. Perhaps the most famous instance is the evolution of greater resistance of rabbits to myxoma virus, released as a rabbit-control measure in Australia. This was coupled with evolution of somewhat lower virulence in the virus (see earlier discussion). Comparisons among geographic populations and species also provide evidence of coevolution. Parasites are commonly better adapted to their local host population than to other populations, as in the case of a microsporidian that infects the cladoceran Daphnia magna. Brook para- sitic cuckoos are polymorphic for egg color: females generally specialize on one or another species of host, and lay eggs that mimic those of the favored host. Some species of birds, especially those most frequently parasitized, reject foreign eggs if they can detect them. In some such species, rejection behavior is stronger in populations that experience parasitism by cuckoos than in those that do not. Thus, both the cuckoo and some of its hosts have evolved responses to the interaction.

Geographic populations of interacting species sometimes vary in parallel, although the causes of variation can be difficult to identify. For example, the parsnip webworm (Depressaria pastinacella), the sole special-
ized herbivore that has accompanied wild parsnip (Pastinaca sativa) from Europe to North America, exerts selection on the plant’s profile of several toxic furanocoumarins. Variation in furanocoumarin profile among populations of the plant is paralleled by variation in the differential capacity to metabolize these compounds in the associated populations of webworm. Although geographic variation in the insect is surely an adaptation to local plant characters, it is harder to show that variation among plant populations has evolved in response to selection by the insect.

Examples of adaptations of predators to prey or of the reverse abound. Yet in few instances has it been demonstrated that these adaptations have evolved in response to any one species. Many characteristics have a similar effect on several or many species of antagonists, and probably represent diffuse coevolution. For instance, the “Mesozoic marine revolution,” in which the evolution of diverse shell-crushing crustaceans and fishes was mirrored by the evolution of spines, thicker shells, and other defensive features in many lineages of molluscs, doubtless represents diffuse coevolution. It is difficult in such cases to show that an evolutionary change in any one species stimulated coevolutionary change in another.

VI. SPECIAL FEATURES IN PARASITE–HOST COEVOLUTION

Several topics loom large in the study of parasites and hosts that seem to play a lesser role in coevolution of animal predators and their prey.

A. Gene-for-Gene Systems

Whereas features that mediate predator–prey interactions, such as size, trophic structures, and cryptic coloration, are usually polygenic (quantitative) characters, specific loci for resistance have been identified in several species of plants. For each such resistance gene, a corresponding locus in a fungal pathogen confers “virulence,” meaning here the ability to attack and develop on a host with a specific resistance allele. (Most well-studied gene-for-gene systems involve crop plants, which has led some authors to suggest that resistance genes with large effects may be a result of the methods used to breed for resistance.) In the absence of pleiotropic costs, resistance alleles at all loci would be fixed in the host population and virulence alleles would be fixed in the parasite. If these alleles do have costs, then frequency-dependent selection can result in long-term fluctuations in allele frequencies, so that each species remains polymorphic and is continually adapting to the changing genetic constitution of the other. (This would be a clear instance of the Red Queen principle.) Populations of Australian wild flax differ in the frequencies of phenotypes resistant to various strains of flax rust, and associated rust populations likewise vary in virulence. The variation among populations may be an effect of Red Queen dynamics.

B. Selection for Sex

The supposition that parasite populations continually evolve has suggested to many authors that a major advantage of recombination and sexual reproduction may reside in the ability to generate novel, parasite-resistant genotypes. In a species of freshwater snail, for example, sexually reproducing females are more frequent than parthenogenetic females in habitats where the risk of infection by trematodes is greatest. Furthermore, among the several theories of sexual selection by female choice is the proposal that it is advantageous for females to choose males whose elaborate ornaments and behavior indicate that they are not debilitated by parasites, and thus are likely to father parasite-resistant offspring. Evidence from a few studies has been interpreted as support for this hypothesis, illustrating the broad ramifications that coevolution may have.

C. Evolution of Virulence and Avirulence

Often, though by no means always, parasites have less virulent effects on their normal host species than on novel, recently invaded species of hosts. This pattern might result from (1) evolution of resistance in a host species, (2) extinction of highly virulent species of parasites (“species selection”), because they drive their hosts to extinction, and/or (3) evolution within parasite species toward lower virulence. The evolutionarily naive often do not distinguish between the latter two causes, and imagine that parasites evolve avirulence because this will safeguard the host population and thus the perpetuation of the parasite species. However, natural selection is not prescient, and selection within populations can easily lead to fixation of “selfish” genotypes that enhance the risk of extinction. Evolution of the level of virulence depends on many factors.

For a parasite such as a virus, bacterium, or protozoan that reproduces on or in its host, each individual host carries a temporary group of parasites. Consider first a case in which each host carries only a single,
clonally propagating parasite genotype. The fitness of this genotype is measured by the number of uninfected hosts it infects, relative to other genotypes. This number is often proportional to the number of offspring produced within the host and capable of transmission. But greater numbers of parasites take a greater toll on a host, resulting in greater virulence. The fitness of a parasite genotype is thus set by a balance between the rate of reproduction within the host and the reduction in transmission due to death or debilitation of the individual host. If a parasite population evolves less than maximal virulence, it is only because this enhances transmission of progeny to new individual hosts, not because it preserves the host population for future use.

The optimal level of virulence depends partly on the mode of transmission, especially whether it is mostly vertical (from parent host to offspring) or horizontal (as in sexually transmitted parasites). The transmission rate of a vertically transmitted parasite is directly proportional to the number of offspring of its carrier, so such parasites should be relatively benign. However, a horizontally transmitted parasite, given opportunity for transmission, does not profit from its host’s further survival or reproduction. Many such parasites kill their hosts (e.g., baculoviruses of insects) or castrate them (e.g., some plant-pathogenic fungi).

The evolution of virulence is still more complex if individual hosts are typically coinfectected by multiple genotypes of the parasite. Within each host, the genotype with highest growth rate \( r \) (and highest virulence) has an advantage, but if individual hosts bearing this genotype die before transmitting it, groups of parasites dominated by less virulent genotypes will transmit the most offspring. Thus group selection favors lower virulence, and individual selection favors higher virulence. At equilibrium, the level of virulence is likely to be the higher than in the case of singly infecting parasites. Comparison of the virulence of species of nematodes that infect fig wasps either singly or multiply has supported this theory, and phage in laboratory cultures evolved lower virulence if they were vertically than if they were horizontally transmitted (Bull, 1994).

VII. SPECIAL FEATURES IN HERBIVORE–PLANT COEVOLUTION

A. Resistance, Tolerance, and Overcompensation

Herbivore impact on plant fitness can be intense, imposing selection for traits that reduce the effects of herbivory. One class of traits simply reduces the incidence of herbivory by escaping herbivores in space and/or time, that is, by decreasing the plant’s “apparent” \( r \). The idea is that abundant, long-lived plants in communities of low diversity will be more easily tracked down and fed upon by herbivores than ephemeral plants that establish short-lived populations in more diverse communities. Size of the plant may also be a factor, as small plants with simple architecture generally have lower insect diversity, perhaps by offering little protection to insects. Consistent with this hypothesis is the observation that among Capparales in Morocco, less damage by pierid butterfly larvae is incurred by species that fluctuate more in density. This suggests that greater constancy in population size and density, that is, greater predictability and apparentness, increases the likelihood of herbivory. Fluctuations in density could, however, be a result of adaptation of germination to specific environmental cues or other factors rather than pressures from herbivory.

Most plants must deal with the presence of herbivores, and possess a variety of mechanisms that either reduce herbivory (resistance) or ameliorate its effects (tolerance). The ability to tolerate damage, that is, to suffer a reduced impact of damage on fitness, depends both on extrinsic factors, such as resource availability, the timing of herbivory, and the parts affected, and on characteristics of the plant. Thus it is thought that plants with intrinsically higher growth rates, storage capacity, or higher number of active meristems will have greater opportunity for regrowth in face of biomass loss to herbivory. It is as yet unclear whether these traits are adaptive defenses or whether they evolved because of other selective forces, such as competition and resource assimilation.

Some experiments have suggested that under ideal conditions, damage at certain phenological stages and the presence of abundant resources could render plants more fit than had they not incurred damage, that is, they would overcompensate. For example, artificially damaged scarlet gilia (Ipomopsis aggregata) produced approximately twice as many flowers and fruits as undamaged plants. The root biomass of clipped plants was greater than that of controls, suggesting that by increasing root size plants were able to take up the extra nutrients necessary for the overcompensation response. This implies that herbivores would benefit plants by feeding on them. There is controversy over the design and results of these experiments, and overcompensation has not been observed in enough species to be considered a general phenomenon.

Greater attention has been given to plant resistance,
namely, to mechanisms that actively repel herbivores or that make the plant unpalatable or undesirable. Resistance mechanisms differ with respect to when they are expressed: many defenses are constitutive, that is, expressed regardless of herbivore presence, whereas others are absent until induced in the presence of herbivores. Induced responses curtail the costs associated with defenses, as these are incurred only when strictly necessary.

Some plants have morphological defenses, such as hard seed coats, thorns, or stinging hairs. However, a central role in defense has been attributed to secondary metabolites. These chemicals play no part in primary metabolism and groups of compounds are usually restricted to a few plant families. They have been found to affect herbivore behavior by decreasing a herbivore's feeding rate, deterring oviposition, slowing development, and reducing fecundity. Yet many of these compounds have also been found to affect bacteria, fungi, viruses, other plants, and other herbivores. Secondary compounds might just as well be an adaptive defense to these organisms, or even serve other ecological or physiological functions (e.g., UV protection, drought tolerance). Convincing evidence that insects are a driving force in the evolution of plant chemical defenses comes from studies that compare the pattern of natural selection on putative resistance factors in the presence and absence of herbivores. For example, in the presence of herbivores a population of Arabidopsis thaliana was shown to be under selection for increased levels of mustard oils (or glucosinolates) and trichome densities.

In contrast to resistance, plant tolerance does not affect the performance or fitness of the herbivore. Herbivores will not be under selection pressure to overcome this plant defense and so coevolution should not be expected, unless resistance and tolerance are correlated. Theoretical models do indeed predict a trade-off between these two forms of defense, but the limited evidence so far is inconclusive.

B. Specialist and Generalist Herbivores

Most herbivorous vertebrates and many insects are generalists, feeding on plants in many families. A more specialized diet, of plants in only a single family, genus, or even species, characterizes the majority of herbivorous insects and a few vertebrates (such as the koala, a specialist on Eucalyptus foliage). The advantages of a broad diet seem clear: an individual can feed despite spatial or temporal variation in the availability of any one food type. Theoretically, the possible advantages of specialization are clear, but evidence is often more ambiguous.

Commonly, closely related species of insects specialize on different plants. Assuming their common ancestor fed on one or more of these plants, the problem is to explain why some of its descendants, instead of expanding their diet, abandoned their ancestral host to which they were presumably well adapted in favor of new hosts. Several ecological advantages of such a switch have been suggested. The new host might provide escape from predators or parasites. Certainly specialization is maintained, if not originally caused, by the protection that many insects, such as milkweed-feeding monarch butterflies, gain by sequestering defensive plant compounds. Competition for resources, as we have noted earlier, may also select for specialization. In many species, mating occurs only on the host plant, and use of a single host may be the most reliable way to find mates. Host specialization will then be a consequence of positive frequency-dependent selection. There is indirect evidence for each of these hypotheses.

The most general explanation for specialization is likely to be trade-offs in the ability to find, handle, digest, or detoxify different kinds of plants. The secondary compounds that characterize higher taxa of plants are thought to play an especially important role, since many are toxic and/or deter feeding by nonadapted insects. Conversely, host-specific insects typically use some such compounds as feeding or oviposition stimuli. The hypothesis that related specialist insects usually feed on related plants because of the plants' chemical similarity has been supported by a study of Blepharida flea beetles: the beetles' phylogeny is more closely mirrored by chemical similarity than by phylogenetic affinity among the species of their Bursera host plants. Tests for trade-offs have had mixed results. There is little evidence that closely related generalist and specialist species differ in their efficiency of digestion of plant species that they both naturally feed on (Futuyma and Wasserman, 1981). There is some evidence that specialist species can find and recognize host plants faster than generalist species, owing to properties of the nervous system, and abundant evidence that closely related species are physiologically better adapted to their own than to each other's host plant species. However, since such differences may have evolved after these species became specialists, most research on trade-offs has compared genotypes within species. Only a few such tests have found clear evidence of trade-offs; different clones of pea aphids, for example, have high fitness on either pea
or alfalfa. In no case is the biochemical or physiological basis for such trade-offs fully understood.

VIII. MUTUALISM

In a mutualistic interaction, individuals of each of two species obtain a fitness benefit from the interaction between them. Long-term physical associations are referred to as symbioses. These include endosymbioses, in which microorganisms such as bacteria reside within their hosts. In some cases, endosymbionts have become, in effect, part of the host. Mitochondria and chloroplasts, for example, evolved from endosymbiotic bacteria early in eukaryote evolution.

Some mutualisms have doubtless evolved from casual, fortuitously beneficial interactions; such may be the origin, for instance, of the nectar glands of many plants that attract generalist ants, which attack herbivores and other intruders. Some such mutualisms have become obligate, as in the case of some Pseudomyrmex ants that inhabit and defend only certain species of Acacia trees that provide food bodies and special domiciles. In many cases, especially symbioses, mutualisms have arisen from parasitic interactions; for example, a complete spectrum from parasitism to mutualism is found in endophytic fungi (Clavicipitaceae). The conditions that favor evolution of parasitism into mutualism are much the same as those favoring the evolution of avirulence (see earlier). Specifically, the fitness of vertically transmitted symbionts depends on the fitness of their hosts, so a benefit provided to the host will enhance the symbiont's fitness. Even without vertical transmission, long-term (e.g., lifelong) association between individual partners satisfies the conditions for "reciprocal altruism." The survival and reproduction of a colony of Acacia-dwelling ants depends on the health of its host tree, so it pays to defend it, just as the fitness of the tree depends on nurturing its defending ant colony.

Many mutualisms are rather nonspecific, in that at least one of the species of partners holds a mutualistic relationship with several or many species. Most mycorrhizal fungi, for example, are not highly host-specific, and most species of pollinating or seed-dispersing animals visit many species of plants. There are, however, some exceptions, such as the fig wasps, each species of which develops as a seed predator of one of the several hundred species of fig and is the sole pollinator of that species. The vertically transmitted bacterial endosymbionts (Buchnera) of aphids are so intimately tied to their hosts that the phylogeny of the bacteria mirrors fully that of the aphids. Unique associations of lineages of aphids and bacteria have persisted for more than 100 million years.

Mutualism can generally be viewed as reciprocal exploitation. Mutualists do not provide benefits to their partners out of altruism; the benefit is either wrested from them (e.g., pollen consumed by pollinating bees) or it is a "payment" that ensures a reciprocal benefit. Thus potential conflict pervades most mutualisms, for selection often favors "cheater" genotypes that do not repay the benefit they receive. For instance, bees often rob nectar by chewing through the bases of flowers, so that they do not pollinate. Many orchids deceive pollinating insects, providing neither nectar nor consumable pollen, but instead produce scents that mimic the insect's sex pheromone and induce "pseudoocopulation" by male insects that effect pollination. Some orchids produce no reward whatever. Selection for cheating can destabilize a mutualistic relationship. For example, yucca moths (Tegeticula) oviposit in the ovaries of Yucca, where the larvae feed on some of the developing seeds. Most species of Tegeticula stereotypically (and "carefully") place pollen on the stigmas of flowers in which they lay, and so both the host plant and the moth's offspring benefit. However, several lineages of "cheater" Tegeticula have evolved from pollinating ancestors; they do not pollinate, but lay eggs in flowers that have been pollinated by those species that do.

IX. MAJOR CONSEQUENCES OF COEVOLUTION

If it is difficult to document and characterize coevolution between pairs or among small numbers of species, it is harder still to demonstrate the effects that coevolution has had on the properties of communities, ecosystems, or the history of biological diversity. For the most part, questions on so grand a scale have, so far, only tentative or even speculative answers.

A. Phenotypic Diversity

Although no quantitative data are available, coevolution, or at least adaptation to interspecific interactions, has unquestionably enhanced the phenotypic diversity ("disparity") among species. The diverse trophic structures displayed by related species of fishes, birds, and other animals partly reflect adaptation to the characteristics of the different arrays of prey species on which
they feed, and partly arise from selection due to inter-
specific competition, resulting in divergence and spe-
cialization in resource use. The diverse morphologies
in such adaptive radiations as the African lake cichlids
illustrate that ecological and morphological diversifica-
tion of a clade may be promoted by competition among
its member species, but may be constrained by competi-
tion from members of other clades, until such competi-
tion is relieved.

Predation and parasitism have doubtless enhanced
phenotypic diversity among prey. There is much cir-
sumstantial support for Ehrlich and Raven’s hypothesis
that chemical diversity among plant taxa has been
driven by herbivores (and, we would add, pathogens).
Intraspecific genetic diversity at major histocompatibil-
ity loci has almost certainly been maintained by selec-
tion imposed by pathogens, and the extraordinarily
complex immune system of vertebrates is testimony
to the potent selection that pathogens generate. The
Mesozoic evolution of shell-crushing predators called
forth an immense proliferation of defensive features in
the shells of molluscs.

The cooperation and conflict inherent in mutualism
have likewise had immense effects on phenotypic diver-
sity. The mycetocytes of aphids, the light organs in
some fishes and cephalopods, and the root nodules
of legumes are among the many special structures that
house symbiotic bacteria. The phenotypic diversity of
the Orchidaceae, the largest family of plants, lies largely
in the extraordinary variety of flower forms that attract
and manipulate diverse pollinating animals.

Although such examples provide undeniable evi-
dence that biotic interactions have had great evolution-
ary consequences, phenotypic disparity has seldom
been measured. Moreover, although the role of interac-
tions in these examples is obvious, the importance of
reciprocity (the “co” in coevolution) is not. Rigorous
tests of the effects of coevolution remain to be done.

B. Species Diversity

Variation in species diversity among clades, communi-
ties, ecosystems, and biomes and through geologic time
is affected by many factors. The relative importance of
coevolution on rates of speciation and extinction, which
together determine diversification rate, is unclear. Only
in exceptional cases, such as the divergence of figs and
their species-specific agaonid pollinators, is it likely
that coevolution directly induces speciation, that is,
reproductive isolation. It has often been suggested that
animal-mediated pollination and seed dispersal may
enhance rates of plant speciation by establishing repro-
ductively viable isolated populations. However, species
diversity of angiosperms is only slightly greater in
animal- than wind-pollinated families, although it is
greater in families that include both biotic and abiotic
seed dispersal than in families that use one mode exclu-
sively.

In principle, extinction rates can be increased or
decreased by coevolution, although either is very diffi-
cult to document in practice. If coevolution has effected
a net increase in diversification rate, it has probably
done so chiefly in three ways: (1) by shifts in resource
use, thus reducing competition; (2) by evolutionary
escape from predation, followed by radiation; and (3)
by evolution of specialized predators, parasites, and
mutualists following the evolution of new prey or hosts.
That is, diversification of one group of organisms begets
diversification of others, as Ehrlich and Raven postu-
lated for plants and herbivorous insects.

These hypotheses are supported by some modest
evidence. Farrell et al. (1991) found that clades of plants
with canals that deliver defensive latex or resin to sites
of herbivore attack are consistently more diverse than
their sister groups that lack these features. Similar sis-
ter-group comparisons among insects showed that the
evolution of herbivory has consistently been followed
by greater diversification. However, it is not known if
the diversity of herbivorous insects has been enhanced
in clades that have adapted to chemically novel, diverse
plant clades, as Ehrlich and Raven proposed. In western
Europe, the species richness of aphids and certain other
insects that feed on different plant families is correlated
with the number of species of plants in the family.
Whether or not such a correlation holds on a global
scale is not known.

C. Community Structure

The convergent evolution of ecologically “equivalent”
species, such as many Australian marsupials and their
placental counterparts in other regions, suggests that
independently evolving communities might converge
toward similar structure. Certainly, properties such as
vegetative morphology and the architecture of vegeta-
tion converge; for example, “Mediterranean” vegetation
(chaparral, matorral, maquis) is dominated by sclero-
phyllous, small-leaved shrubs in several parts of the
world. Such similarities, however, arise simply from
each species’ independent adaptation to physical envi-
ronmental factors. Whether or not community proper-
ties such as stability, food-web structure, or species
diversity converge as a result of coevolution is a differ-
ent, much more difficult question. Simple models of
interspecific competition and of food webs suggest that the coevolutionarily stable equilibria might be rather few, so that coevolution might yield some predictable structure (MacArthur and Levins, 1964). Whatever the ideal applicability of these models might be, however, the opportunity for such pervasive effects of coevolution may generally be rather slight. Gene flow among conspecific populations that interact with different ensembles of species may prevent finely tuned coevolution. Moreover, paleoecological studies have shown that throughout the Pleistocene, species have had highly individual histories of change in geographic distribution, so that many of today's assemblages of species are very recent. Except for specialized associations, as of host-specific parasites that have moved about with their hosts, there has been little time for coevolutionary adjustments in many of today's species assemblages. Nevertheless, paleontologists have documented rather steady levels of diversity at both global and local levels over vast periods of time (10^5–10^8 years), despite turnover of taxa (Jablonski and Sepkoski, 1996). These observations suggest that whatever convergence or constancy of community structure exists may be attributable more to purely ecological rules of community assembly rather than to coevolution.

An unusually clear example of convergent multispecies assemblages is provided by the anoline lizards mentioned earlier, which form monophyletic groups of morphologically and ecologically equivalent species on each of the Greater Antilles islands. However, similar habitats in different parts of the world generally are not very similar in species richness. For example, lizard diversity in Australia exceeds that in corresponding habitats in southern Africa. Nevertheless, in some instances variation in species richness among habitats shows similar patterns, lizards are more diverse in deserts than in wetlands in both Australia and Africa. Thus, habitats seem to have consistent effects on species coexistence, and perhaps on the evolution of resource partitioning (Ricklefs and Schluter, 1993).

Coevolution may also affect community properties such as stability, that is, the property of returning to equilibrium after a disturbance. The study of artifically constructed food webs has identified some patterns of interaction among component species that are conducive to stability. One such aspect is connectivity, that is, the ratio of actual to potential interspecific links. High levels of connectivity are associated with reduced stability, as a disturbance to one species will affect many other species in the community. Thus processes that reduce connectivity, or generate compartments within a community, such as a subset of species that interact predominantly with one another, contribute to stability. To the extent that coevolution between parasites and hosts and between plants and herbivores leads to increased levels of specialization (and thus compartmentalization), it can thus indirectly enhance community stability.

See Also the Following Articles

COMPETITION (INTERSPECIFIC) • DARWIN, CHARLES • EVOLUTION, THEORY OF • PARASITISM • PLANT-INSECT INTERACTIONS • RECOMBINATION • SPECIES COEXISTENCE

Bibliography


COMMONS, CONCEPT AND THEORY OF

Colin W. Clark
University of British Columbia

I. Introduction
II. Theory
III. Regulated Bionomic Equilibrium
IV. Managing Incentives Versus Managing Resources
V. Allocated Quotas
VI. International Common Property Resources
VII. Discounting the Future
VIII. Community Control
IX. Risk and Uncertainty
X. Biological Reserves
XI. Related Concepts

GLOSSARY

bionomic equilibrium In a biological common property resource, an equilibrium between resource production rate and exploitation rate, characterized by the dissipation of economic rents.
commons (or common property resource) Any resource asset that is not privately owned and controlled.
discount rate Rate (usually annual) at which future revenues are discounted to calculate a present value. Personal discount rates often exceed market rates by a wide margin.
economic rent The flow of net economic benefits (revenue minus costs) derived from the exploitation of a resource asset.
externality Cost or benefit imposed on others as the result of some economic activity.
Pareto efficiency An economic equilibrium is Pareto efficient if any change will decrease the economic well-being of at least one participant.
tragedy of the commons Process whereby a commons is overexploited because individual users cannot expect to realize the potential benefits of resource conservation.

IN MEDIEVAL ENGLAND THE TERM "COMMONS" referred to an unfenced area of land that, although privately owned (usually by the local manor lord), was by tradition available to specific persons (the commoners) for specific activities, such as pasturage or removal of building materials. The subsequent enclosures of the commons, for the sole benefit of their owners, reduced many former users to penury. In recent times, and especially following the publication of the famous article "The Tragedy of the Commons" (Hardin, 1968), the term has been used to refer to any resource asset that is not privately owned and controlled. This is the sense in which the term commons, synonymous with "common property resource," will be used here.
I. INTRODUCTION

Important contemporary examples of common property resources include the global atmosphere, the oceans, large lakes, rivers, forests, and fish and wildlife populations, including birds. Though not inevitable, the overexploitation of common property resources is always a potential threat, and often a frightening reality. Many current environmental problems can be traced to the working of a so-called tragedy of the commons.

The classification of resources as either common or private property is an oversimplification. Many gradations exist between a total lack of access restriction (i.e., open access) and complete individual control. For example, coastal marine resources within 200-mile fishing zones are now recognized as being under the sole jurisdiction of the coastal state, with foreign fishermen either excluded or subject to payment of fees. These resources are thus limited access, but (unless privatized) still common property by our definition. Also, private landowners seldom have full rights to do as they wish with their property—the state may retain mineral rights, for example, and zoning regulations may restrict land use and development.

The logic of the tragedy of the commons is straightforward, as will be explained later. Unfortunately, however, much of what has been written about common property resources suffers from vagueness and imprecision, leading to unnecessary confusion and controversy. The following hypothetical example can be useful as a basis for a general understanding of the economics of a commons.

Imagine a small lake, completely surrounded by the properties of two landowners. The lake contains a population of fish, and both landowners are avid fishermen. What will be the long-term outcome? The answer is not obvious and depends on various factors.

The first possibility is that the lake is so large, and the fish population so productive, that no amount of fishing by the two landowners has any appreciable effect on the population. In this ideal situation, by assumption nothing happens—the tragedy of the commons does not occur.

The opposite case is more instructive. Suppose, for example, that upon first taking up their properties, the two neighbors are able to catch all the fish they want. After a few years, however, they observe that fish are becoming more scarce. Also, the big ones that used to put up such a fight seem to have disappeared. Subsequent developments now depend on how the two landowners react. If they meet on amicable terms and agree to restrict their catches, or to return some of them to the lake, the dynamics of the exploited fish population may stabilize, permitting sustained if limited catches by both owners. It may take several years of trial and error before the maximum sustainable catch is determined—or they may decide to forego the maximum catch in favor of maintaining a large population of large fish.

An alternative solution might be to subdivide the lake into two fishing zones, one for each landowner. But this assumes cordial relations between the landowners, with each cooperating in not fishing in the other’s zone. If fish are able to swim freely between zones, such an agreement may fail to prevent overfishing. Privatization of property or resources is not always feasible, and when it is feasible it presumes mutual trust and cooperation, or external enforcement.

Instead of such amicable agreements, however, the owners may begin blaming each other for overfishing. Relations may then deteriorate, with both determined to catch all the fish they can find. One may buy a faster boat, hoping to outdo his neighbor, who then responds in kind, and so on. Such behavior between two neighboring individuals may seem petulant. But what if instead of two property owners, there were 100 cottages on the lake, all owned by avid fishermen? Besides increasing the pressure on the fish population, this would make the problem of agreeing over and enforcing catch restrictions much more difficult. Cheating by individuals would be hard to detect, and even harder to control.

This simple example illustrates many of the characteristic features of common property resources, including potential overexploitation, the necessity of mutual cooperation, and the increased difficulties of control as the number of joint owners increases. We return to these questions later.

II. THEORY

The theory of a common property renewable resource will be discussed here in terms of a specific example, a commercial fishery. The results and predictions apply, however, to most other biological resources, with minor modification.

The theory of bionomic equilibrium in an unregulated, open-access, common property fishery was initiated by Gordon (1954) (Fig. 1). Fishermen exert fishing effort $E$ on a certain fish population; fishing effort is a measure of fishing intensity, for example, the number of vessel-days per year, or (in a rod-and-reel fishery) the number of rod-hours per year. Figure 1 shows the...
Figure 1. Gordon's (1954) model of equilibrium revenue $R$ and cost $C$ in a fishery; $E^*$ bionomic equilibrium; $E^*$ effort level for maximum sustainable economic rent.

Total sustained annual revenue $R$ and total cost $C$ as functions of effort $E$. Cost increases in proportion to $E$, but revenue rises to a peak ("maximum sustainable yield") and then declines at higher levels of effort. An input–output function having this inverted-U shape is unconventional in economic theory; it reflects the underlying biology of the resource, which becomes increasingly unproductive as it is depleted by high levels of fishing effort.

Gordon's prediction is that the fishery will reach an equilibrium, called the bionomic equilibrium, at the effort level $E^*$ where revenues and costs are equal. The argument is as follows. First, for $E < E^*$, revenue exceeds costs and fishermen are making money in excess of their opportunity costs. This positive net revenue attracts additional entrants to the fishery, resulting in an increased effort level. On the other hand, if $E > E^*$, fishermen are losing money and some will leave the fishery. Equilibrium can only occur where $E = E^*$. At this bionomic equilibrium the fishermen are barely breaking even financially. (See Clark, 1990, for mathematical representation of these ideas.)

For levels of $E$ below $E^*$, revenues $R$ exceed costs $C$; the excess $R - C$ is called economic rent. Gordon's prediction can then be stated: in an open-access unregulated fishery all potential economic rents will be dissipated in long-run bionomic equilibrium. Additional predictions are that if the price of fish increases, or if the cost of fishing decreases, the effort level $E$ will increase, resulting in a more depleted fish population. What would normally be recognized as benefits to the fishing industry (higher prices, lower costs) are dissipated because of the tragedy of the commons.

What is not clear in the Gordon model is whether the fish population will survive in the long run, or be driven ultimately to extinction. In practice, very few fish species have been fished to extinction, although some marine mammals have been. On land, America's most abundant bird species, the passenger pigeon, was hunted to extinction in the nineteenth century. Today species such as rhinocerocoses, tigers, and certain birds are endangered by market hunters. If we agree that the Gordon theory applies to any such commons, then any species is likely to become endangered if the value of the last specimen caught exceeds the cost of capturing it. In a world of extremely affluent consumers and destitute hunters, this condition applies to an ever-widening class of wild species, both marine and terrestrial.

Gordon also identified the effort level $E^*$ yielding maximum economic rent (see Fig. 1). He pointed out the apparent anomaly that if the fishermen would only work less hard—reduce their effort—they would be better off! This argument may seem compelling, but the reality is that many, if not most, marine fisheries today remain closer to bionomic equilibrium than to rent maximization. It often seems as if the fishing industry is determined to suffer the tragedy of the commons even though impoverishment is the inevitable outcome. What explains this phenomenon?

To address this question we must consider the dynamics of the fish population (recall that Fig. 1 pertains only to equilibrium situations). For example, suppose that a certain fishery is currently at bionomic equilibrium $E$. A reduction of effort is needed in order to improve the situation. Unfortunately, such a reduction of effort will not lead to an immediate increase in catches—quite the reverse. At first catches will decrease, more or less in proportion to the decreased effort. The decreased catches will (it is hoped) allow the fish stock to increase over time. Eventually a new equilibrium will be reached, combining reduced effort with increased catches and increased rents. A simulated illustration of this process is shown in Fig. 2.

The principal message from the argument is that resource conservation necessarily involves a sacrifice of potential current benefits in order to achieve greater future benefits. From the economic point of view, conservation is thus a form of investment in the future.
FIGURE 2  Simulated annual catches following a reduction in effort at time \( t_0 = 10\) yr. At first catches are reduced, but eventually catches increase to a higher level than before, as the fish population recovers from overharvesting. (Discrete logistic model; \( r = 0.3 \text{ yr}^{-1}, K = 100,000 \) metric tons, \( E_{t_0} = 0.24 \text{ yr}^{-1}, E_{t_0} = 0.15 \text{ yr}^{-1} \).)

source, even though such activities would be individually profitable. The more successful the management program, in terms of protecting the resource, the greater will be the incentive for excessive exploitation by individuals. This dilemma is the real tragedy of the commons.

Although formulated here in terms of fish populations, the concepts of bionomic equilibrium and rent dissipation in common property resources are widely applicable. Overgrazing of grasslands, desertification, deforestation, soil erosion, and pollution are other important examples of overexploitation of the commons. Everywhere that resource stocks are used as common property, the incentive for the commons tragedy exists.

The participants in a tragedy of the commons are sometimes said to be motivated by short-term financial considerations—with “everyone following their own short-term interest” to quote from a recent publication. This portrayal of such resource users as greedy, shortsighted souls is often both unfair and inaccurate (although there probably is a tendency for such individuals to gravitate toward unregulated resource industries). In a commons situation, it is in each participant’s best interest to get as much as possible out of the resource before it is depleted by others. Any personal sacrifice for the sake of resource conservation is simply dissipated by the other users.

III. REGULATED BIONOMIC EQUILIBRIUM

How can the problems of common property resource exploitation be overcome? Let us first consider an approach that often does not, and perhaps cannot, succeed. Again, imagine a fishery initially at bionomic equilibrium (this is not essential for the discussion). A government agency is set up to manage the fishery. Realizing that the problem is excessive effort in the fishery, the managers introduce a limited-entry program, in which only a fraction of the previous participants are licensed to catch fish. (It goes without saying that, unless compensated in some way, the excluded fishermen will tend to resist this program.) At first the new system may appear to be successful—effort decreases, the fish population begins to recover, catch rates go up; the remaining fishermen are making money.

But now the economic dynamics of the commons begin to operate once again. By increasing their own levels of effort, individual fishermen can increase their current income. Some of the fishermen buy bigger boats, forcing others to do the same in order to compete. The managers must now bring in regulations to restrict the size of fishing boats. The owners then respond by “capital stuffing”—installing larger engines, more powerful winches, better electronic devices, and so on, all of which increase the fishing power of their vessels. These items in turn must also then be regulated. Eventually either the regulations become so cumbersome that economic rents are dissipated simply by the complexity of the management system, or else the regulations fail to prevent overexpansion of fishing capacity. In either case, the fishery is now in regulated bionomic equilibrium.

Such regulated bionomic equilibrium can nowadays be seen in many managed marine fisheries. Where the regulations are effective and strongly enforced, depletion of fish stocks may be prevented. In practice, however, the economics of the commons has often proved more powerful than the regulations. In a number of cases, intensively managed fisheries have collapsed from persistent overfishing, cod and groundfish on the Atlantic Coast of the United States and Canada being notorious examples. In most of these cases, the fishing industry itself persistently and adamantly opposed the imposition of the regulations that were required to preserve the fishery.

While all of this may seem quite irrational (unless you understand the economics of the commons), in actuality the situation is often more complicated—and
more perverse—than described here. To be specific, resource industries are often highly subsidized. Individuals can only benefit from the subsidy if they participate in exploiting the resource. The effect of the subsidy is to reduce the opportunity cost $C$, in extreme cases to near zero. The Gordon theory explains what happens: bionomic equilibrium corresponds to virtual resource extinction in such a situation. Exploiters become dependent on the subsidy, not the resource. Bionomic equilibrium is more devastating with subsidies than without them.

In the case of Canada's Atlantic cod fishery, the system of regulated, subsidized bionomic equilibrium proceeded to the ultimate conclusion with the almost complete disappearance of cod stocks in 1991. This collapse destroyed the basis of the economy of the entire province of Newfoundland; since 1991 the Canadian government has been paying over $1 billion per year as welfare to displaced Newfoundland fishermen and plant workers. Whether the Newfoundland cod will ever return is uncertain at the present time (1999).

IV. MANAGING INCENTIVES VERSUS MANAGING RESOURCES

Two general approaches to the management of common property resources can be identified: those directed toward management of the resource stock and those that alter the incentives of resource users. Methods aimed at managing the resource include total annual catch quotas, controlled access to the resource, restrictions on the technology of harvesting the resource, and so on. Methods that rely on altering the economic incentives of common property resource exploiters include user fees or royalties and individually allocated quotas (see the following). These methods may also be used in combination. As indicated by the foregoing discussion, management strategies that do not alter users' basic incentives tend to result in regulated bionomic equilibrium, with excessive inputs and persistent rent dissipation. Methods that alter incentives have the potential of overcoming these difficulties, although the persistent incentive to cheat must be recognized, and countered by monitoring and strict enforcement of the rules.

A third approach, privatization of the resource—usually involving geographical subdivision into individually manageable units—may be considered the ideal solution, since it allows users to become independent of each other and of excessive government interference. This is historically the usual method for dealing with nonfugitive resources, such as land and minerals. Such subdivision and privatization is not feasible, however, for many of today's remaining common property resources. In such cases, some form of centralized management seems necessary. We first consider the case in which the resource is located within the jurisdiction of a single state; the case of international common property resources is discussed in Section VI.

V. ALLOCATED QUOTAS

One method of altering the incentives of resource users is a system of allocated quotas, often sold under auction to the highest bidders. This method is commonly used in logging in national forests and in offshore oil drilling. Having obtained such a quota, the users' incentives are to harvest the quota at minimum cost. Persistent incentives toward cheating are thwarted by rigorous enforcement of the quotas. Resource rents are preserved, and divided between resource users and the government. The resource remains common (state) property, but specific user rights are privatized through the system of quotas. Multiple uses of the resource, as in national forests used for both timber production and recreational activities, can be achieved, although different user groups may attempt to influence management decisions in their favor.

The establishment of 200-mile national fishery zones in the 1980s rendered coastal marine resources subject to quota-based management systems, and individual catch quotas are indeed now in use in many such fisheries. Marine populations, however, provide unusual management difficulties, not the least of which is the fact that these resources are largely hidden from view until they are brought to the surface by net or hook. Also, fish populations undergo major natural fluctuations, which are poorly understood and hard to predict. In addition, many fish populations migrate, with migration routes that change in response to oceanic conditions. These features introduce large, irreducible levels of uncertainty into fisheries management; we discuss this topic further in Section IX.

VI. INTERNATIONAL COMMON PROPERTY RESOURCES

As an important example of an international common property resource, let us briefly consider the atmosphere. Although partitioning of the atmosphere is
clearly not possible, policies at the national level have had significant effects on atmospheric pollutant emissions. Norway, Sweden, Denmark, France, and Japan have employed taxes on the emission of sulfur dioxide since the 1970s. Some countries now use tradable quotas to control atmospheric emissions. Such policies affect incentives, and are thought to be far more efficient than direct quantitative regulations.

At the international level, over 100 nations have now ratified the 1987 Montreal Protocol requiring the phasing out of ozone-depleting chlorofluorocarbons and related substances. This example shows that, when evidence of severe harm is strong, international cooperation can be achieved. However, the example also illustrates the difficulties, especially at the international level, associated with common property resource management, as noncompliance, failure to ratify, and the proliferation of illegal trade in ozone-depleting substances have persisted. Although no international agency exists to enforce the Protocol, trade sanctions (or the threat thereof) have been used to encourage adherence to the rules.

Similar agreements, and methods of encouraging nations to follow them, apply to other international resource problems, such as greenhouse gas emissions, marine resource exploitation, ocean pollution, and the conservation of migratory birds. Although international management may lag behind the need, it is encouraging that the general principles of managing the commons do appear to be having an effect on the way in which the rules are being written.

VII. DISCOUNTING THE FUTURE

The view is sometimes expressed that privatizing a commons will automatically result in its conservation. Some economists have gone so far as to assert that the private owner of a renewable resource asset would never over-exploit it, except perhaps as a result of miscalculation. This assertion is totally false and is based on a failure to understand the economic dynamics of natural resource stocks.

To summarize the argument in a nutshell (see Clark, 1990, for details), the hypothetical owner of a renewable resource will calculate the total discounted present value of various options, including sustainable harvesting and depleting the resource for an immediate gain. Managing a certain resource sustainably might, for example, produce an annual income $I$, whereas eliminating the resource stock could yield an immediate gain $P$. The present value of the perpetual income stream $I$ is equal to $I/i$, where $i$ is the annual discount rate. If it happens that $P$ is larger than $I/i$, it will be more profitable to deplete the resource. By investing the proceeds at interest rate $i$, the owner will obtain a greater income than by harvesting it sustainably.

Many biological resources, such as whales, trees, and soil, have low annual rates of productivity, often less than 5% per annum. Removing more than 5% of the stock each year will deplete the resource, eventually to extinction. If the resource owner discounts future revenues at more than 5% per annum, he or she may deliberately choose to eliminate the resource for short-term gain.

Resource users may have reasons for discounting future income at rates higher than market interest rates. For example, they may need, or desire, an immediate cash flow. In addition, the future of the resource may be uncertain, enhancing the relative value of current harvests. Whatever the reason, a resource owner may decide to practice less conservation than may be consistent with sustainable use of the resource.

The same calculation applies to multiple users of a managed common property resource. It is a notorious fact, for example, that the fishing industry almost always demands larger quotas than recommended by government managers. The managers’ calculations may be based tacitly on zero discounting, but the industry almost always prefers to receive its income now.

VIII. COMMUNITY CONTROL

A subsidiary theme in the common property resource literature proposes that common property can be a good thing, not necessarily the stage of a tragedy. The idea is that if all so-called “stakeholders” can be assured a voice in the management of the resource, then agreement to achieve the optimum exploitation should follow. Doubtlessly this could be true in a tightly knit society in which defectors face ostracization or other severe penalties. Recent attempts to reach compromise agreements between users of forests or fisheries, however, have proven frustrating. The hypothetical example of the lake with two adjacent property owners suggests that resource conservation can depend strongly on individual behavior. Although decentralization can sometimes bring benefits, the likelihood that community control of biological resources will improve the prospects for conservation needs to be assessed critically in each instance.

As described earlier, individual community members may discount the future to such an extent that the
results differ little from the tragedy of the commons. Also, if there is any likelihood that some users will try to take more than their share of the resource, community control can quickly degenerate into a destructive commons scramble. Always a master of the rapier phrase, Hardin (1994) labeled those who preach the doctrine of community control as "commonists." In some cases, community control might resolve a local tragedy of the commons, but the number of individuals involved has to be fairly small, and all outsiders must be excluded. Furthermore, the community probably needs to be homogeneous in tradition and religion.

IX. RISK AND UNCERTAINTY
Risk and uncertainty are important in the management of many resources, including common property resources. Here, the following terminology is used: Risk occurs when the outcome of some process is not known in advance, but the probabilities of the various possible outcomes are known from past experience. Uncertainty occurs when these probabilities are not known, or not known accurately. Uncertainty can be reduced through experience and research, but in most situations it cannot be eliminated entirely, that is, it cannot be transformed into pure risk. Examples of irreducible uncertainty are common in resource management. To mention one example, we do not know for certain whether current global warming is the result of increased levels of greenhouse gases, nor do we know what impact greater concentrations of these gases will have.

In the case of fishery resources, uncertainty persists regarding current stock levels, relationships between stock and recruitment, and the overall effects of exploitation on future stock levels. Furthermore, marine ecosystems are not well understood, meaning that the implications of harvesting certain species cannot be predicted with certainty. In addition, uncertainty pertaining to future environmental conditions, such as global warming, only adds to the difficulty of foreseeing future trends in marine populations.

Reducing uncertainty (e.g., through scientific research), or reducing the effects of uncertainty (e.g., by adopting conservative harvest strategies), can be considered as forms of investment in maintaining the future viability of a renewable resource. The tragedy of the commons operates here as well—under common property conditions, resource exploiters ignore future uncertainties as they attempt to maximize current revenues from the resource. Management strategies that limit resource exploitation on the basis of scientific uncertainty are widely unpopular with fishermen and other common property resource users. Indeed, the argument is often advanced that uncertainty implies that the need to restrict resource use has not been proved.

X. BIOLOGICAL RESERVES
One response to the ever-growing pressure on biological commons is the establishment of protected biological reserves. While the benefits of reserves are apparent in terms of protecting biodiversity, the potential direct economic benefits of reserves have until recently been less appreciated. For example, large-scale marine reserves could help in preventing severe overexploitation, or collapse of important marine resources. Such marine reserves should be thought of as complementary to, rather than as substitutes for, other types of resource management. Unlike other management strategies, however, reserves can provide a positive, potentially "fail-safe" backup to traditional resource management.

XI. RELATED CONCEPTS
Closely related to the concept of a commons is the economist's concept of externalities (sometimes called spillover effects). An externality is a cost or benefit imposed on others (without compensation) as the result of some economic activity. Externalities can be positive (e.g., a homeowner paints her house to protect it from the weather, and neighbors enjoy the color scheme) or negative (smoke from the homeowner's fireplace chokes her neighbor). Users of common property resources impose negative externalities by reducing the stock of the resource available to other users. Environmental pollution is another common example of a negative externality. In this case, the quality of a resource, rather than the quantity, is reduced by pollution.

Another related economic concept is that of social cost. When a negative externality occurs, the agent that causes the externality incurs a personal cost (his or her private cost) that is less than the total social cost of the activity.

Yet another notion is that of free riders. When the users of a common resource agree to limit their individual impacts, others—the free riders—may continue to exploit the resource. Examples include poachers who illegally slaughter wildlife in protected areas, or nations that fail to ratify or honor international agreements. The dilemma here is that the greater the success in managing the commons, the greater is its attraction to
free riders. The problem of free riders explains the dual difficulties inherent in community control of the commons: exclusion of outsiders, and control of cheating by insiders. The failed experiments of national communism serve as strong warnings against undue optimism in these situations.

A basic theoretical prediction of economics, that competitive equilibria are Pareto efficient, is valid only in the absence of externalities. How to remedy the effects of negative externalities is the subject of welfare economics. The immense literature in this field cannot be succinctly summarized, but it can be stated that no fully satisfactory solution to the tragedy of the commons (negative externalities, if you prefer) has been discovered, and it may be that none exists. If not, the commons dilemma promises to become ever more serious as the world’s population continues to grow and exert increasing pressures on the biological systems that are essential to our very survival.

See Also the Following Articles

COMMONS, INSTITUTIONAL DIVERSITY OF • ENVIRONMENTAL ETHICS • FISH STOCKS • NATURAL RESERVES AND PRESERVES • POLLUTION, OVERVIEW • PROPERTY RIGHTS AND BIODIVERSITY • RESOURCE EXPLOITATION, FISHERIES • RESOURCE PARTITIONING

Bibliography

COMMONS,
INSTITUTIONAL
DIVERSITY OF

Elinor Ostrom

Center for the Study of Institutions, Population, and Environmental Change and Workshop in Political Theory and Policy Analysis, Indiana University

I. Introduction
II. Centralized Regimes and Biodiversity
III. Using Rules as Tools to Change the Structure of Incentives in Common-Pool Resources
IV. The Daunting Search for Better Rules
V. Experimenting with Rule Changes
VI. Self-Organized Resource Governance Systems as Complex Adaptive Systems
VII. The Advantages and Limits of Parallel Sets of Local Users in Policy Experiments
VIII. The Capabilities of Polycentric Systems in Coping with Tragedies of the Commons
IX. Conclusions

GLOSSARY

Common-pool resources Resources that include all ecosystems that are large enough such that excluding potential beneficiaries from their use is a nontrivial task and each individual's consumptive use (e.g., harvesting of a boatload of fish or a truckload of forest products) reduces what is available to others.
Governance systems Sets of rule configurations used to govern human–ecosystem relationships at operational, collective-choice, and constitutional levels of analysis.
Polycentric governance arrangements Complex, multi-tiered governance systems in which there is no single center of authority.

Rules Commonly understood, normative statements that specify who must, must not, or may take some action or affect some outcome at a particular node in a decision tree.

GOVERNANCE SYSTEMS USED TO REGULATE BIOLOGICAL COMPLEXITY need to be organized at multiple scales and linked effectively together. The design of rules must be viewed as policy experiments with a probability of failure. Highly decentralized governance systems possess both capabilities and limits, as do highly centralized systems. Polycentric systems tend to draw strength from both the lower levels of relatively autonomous units as well as the overarching large-scale units. Institutional diversity is viewed as essential to the preservation of biological diversity.

I. INTRODUCTION

Ecosystems characterized by high levels of biodiversity are complex systems with interactions occurring at multiple spatial and temporal scales. The task of enhancing and sustaining biodiversity through the design of institutions needs to take into account the complexity of the natural systems in which biodiversity occurs. Without effective institutions to limit who can use and regulate diverse harvesting and management practices, com-
mon-pool resources can be overharvested and even destroyed. When common-pool resources containing endemic or threatened species are destroyed, the amount of biodiversity in the world is reduced.

When complexity characterizes the nature of the systems to be governed, it is essential to think seriously about the complexity of the proposed governance systems. Without sufficient concern for sustaining complex governance systems, the very processes of regulating behavior to preserve biodiversity will produce the tragic and unintended consequence of destroying the complexity we are trying to protect. In his book *Design for a Brain: The Origin of Adaptive Behavior* (1960), W. Ross Ashby, an eminent biologist of an earlier era, developed the "Law of Requisite Variety." This law can be roughly stated as: Any regulative system needs as much variety in the actions that it can take as exists in the system it is regulating. Translated into the subject of this Encyclopedia, the law of requisite variety could be stated as: Any governance system that is designed to regulate complex biological systems must have as much variety in the actions that it can take as there exists in the systems being regulated. This is a tall order, but it is one to which we need to pay serious attention.

Unfortunately, much of contemporary policy analysis does not base recommendations on the law of requisite variety. Policy proposals related to the preservation of biodiversity tend to focus on two entirely different solutions: (1) the creation of effective market institutions and (2) the creation of national agencies to manage all common-pool resources within a large domain. Both markets and the state are important institutional mechanisms that should be part of the solutions of protecting biodiversity, but not the exclusive mechanisms. A third option, that of combining large numbers of relatively autonomous local governance systems for specific common-pool resources with larger-scale governance regimes—in other words, a polycentric system—has not been given sufficient attention.

This article focuses on monocentric and polycentric forms of regulation. Section II discusses the underlying presumption of substantial policy proposals, which is that strong, centralized regimes are essential to managing common-pool resources. Substantial empirical work challenges the validity of this presumption. Section III explores the complexity of using rules as tools to change the structure of commons dilemmas, presents the seven clusters of rules that affect the components of any action situation, and then describes the specific rules that are used in field settings by resource users and government agencies. An examination of the types of rules used in the field yields several important findings. First, the number of rules actually used in field settings is far greater than generally recognized. Second, the type of rules is also different. Boundary rules tend to include as co-appropriators of a resource those who are more likely to be trustworthy because they live permanently nearby and have a long-term stake in keeping a resource sustainable. Authority rules define rights and duties that are easy to understand, directly related to sustaining the biophysical structure of the resource, and easy to monitor and enforce. Some rules recommended in the policy literature are not found among the rules used by self-organized systems. This leads to the difficult search for optimal rules, which is discussed in Section IV.

Given the complexity of the process of designing rules to regulate the use of common-pool resources, I argue in Section V that all policy proposals must be considered as experiments. No one can possibly know whether a proposed change in rules is among the more optimal rule changes, or even whether a rule change will lead to an improvement. All policy experiments have a positive probability of failing. In Section VI, I draw on recent research by John Holland and colleagues at the Santa Fe Institute to discuss the attributes and mechanisms of a different form of general organization—a complex adaptive system—that is not the result of central direction. Complex adaptive systems cannot be understood if one tries to fit these systems into the image of an organization with a central director.

Section VII then shows how the parallel efforts by a large number of local resource users to search out and find local rule configurations may find better rule combinations over the long term, whereas top-down design processes are more limited in their capacities to search and find appropriate rules. All forms of decision making have limits. In Section VIII, I discuss the limits of a series of completely independent resource governance systems and the importance of building polycentric governance systems with considerable overlap to combine the strengths of parallel search and design processes with the strengths of larger systems in conflict resolution, acquisition of scientific knowledge, monitoring the performance of local systems, and the regulation of common-pool resources that are more global in their scope. The resulting polycentric governance systems are also not directed by a single center. They, too, are complex adaptive systems requiring policy analysts to change their fundamental views of organization in order to cope more effectively with tragedies of the
commons and many of the other problems facing modern societies.

II. CENTRALIZED REGIMES
AND BIODIVERSITY

A substantial number of contemporary policy proposals stress the desirability of central regimes for the regulation of natural resource systems. The "scientific management of natural resources," for example, which is regularly taught in many universities to future regulators of natural resources, presents fisheries and forests as relatively homogeneous systems that are closely interrelated across vast domains. This approach, as it has been applied to fisheries management, is described by Acheson et al. (1998: 391–392):

For those trained in scientific management, it is also an anathema to manage a species over only part of its range... [I]t is not rational to protect a species in one zone only to have it migrate into another area where it can be taken by other people due to a difference in regulations. Lobsters, for example, extend from Newfoundland to the Carolinas... From the point of view of the National Marine Fisheries Service, it makes sense to have a set of uniform regulations for the entire US coast rather than one for each state.

Contemporary policy analysts also share a belief in the feasibility of designing close-to-optimal rules to govern and manage species over a large domain by utilizing top-down direction. Since natural resources are viewed as relatively homogeneous, and interlinked and simple models exist of how they work, officials acting in the public interest are considered capable of devising uniform and effective rules for an entire region. Prescriptions calling for central governments to impose uniform regulations over most natural resources are thus consistent with much of contemporary policy analysis.

These prescriptions are not, however, supported by extensive research. An extraordinary number of field studies have found that local groups of resource users, sometimes by themselves and sometimes with the assistance of external institutional arrangements, have created a wide diversity of institutional arrangements for coping with common-pool resources when they have not been prevented from doing so by central authorities. These empirical studies document successful self-organized resource governance systems in diverse sectors in all parts of the world. Examples also exist of commons dilemmas that have continued unabated. One conclusion that can firmly be made in light of extensive empirical evidence is that overuse and destruction of common-pool resources is not a determinant and inescapable outcome when multiple users face a commons dilemma. Scholars have begun to identify the conditions of a resource, and of the users of a resource, that are most conducive to local users self-organizing to find solutions to commons dilemmas. Furthermore, the broad design principles that characterize robust self-organized resource governance systems that have resolved commons dilemmas for long periods of time have been identified (E. Ostrom, 1990) and found basically sound by other scholars.

Another important set of findings is that national governmental agencies have been notably unsuccessful in their efforts to design an effective and uniform set of rules to regulate important common-pool resources across a broad domain. Many developing countries nationalized all land and water resources during the 1950s and 1960s. The institutional arrangements that local users had devised to limit entry and harvesting lost their legal standing, but the national governments lacked funds and personnel to monitor the use of these resources effectively. Thus, common-pool resources were converted to de jure government-property regimes, but reverted to de facto open-access regimes. The perverse incentives of such open-access regimes are often accentuated, since local users have specifically been told that they will not receive any benefits if they adopt a long-term perspective and regulate their use of the resource. When resources that were previously controlled by local participants have been nationalized, state control has usually proved to be less effective and efficient, if not disastrous, than control by those directly affected. The harmful effects of nationalizing forests that had earlier been governed by local user groups have been well documented in Asia and in Africa. Similar results have occurred with regard to inshore fisheries taken over by state or national agencies from local control by the inshore fishermen themselves.

These findings challenge the theoretical foundations of much of contemporary policy analysis. A foundational belief underlying this analysis is that designing rules to change the incentives of participants is a relatively simple analytical task best done by objective analysts not intimately related to any specific resource. Analysts view most resources in a particular sector as relatively similar and sufficiently interrelated, and thus conclude that they need to be governed by the same set of rules. This is bolstered by the view that an organi-
zation itself requires central direction. Consequently, the multitude of self-organized resource governance systems are viewed as mere collections of individual agents each out to maximize their own short-term returns. The groups who have actually organized themselves are frequently invisible to those who cannot imagine organization without rules and regulations issued by a central authority.

In order to understand why designing rules to regulate the use of common-pool resources so as to sustain their long-term productivity and biodiversity is not a simple task, one has to change perspective somewhat. One has to begin to think through the process of using rules as tools to change the structure of incentives facing users of a common-pool resource.

III. USING RULES AS TOOLS TO CHANGE THE STRUCTURE OF INCENTIVES IN COMMON-POOL RESOURCES

With this change in perspective, we can think of users trying to understand both the biophysical structure of a common-pool resource and how to affect each other's incentives in order to increase the probability of sustainable use over the long term. Instead of being endowed with complete knowledge of how a common-pool resource reacts to use patterns over time, they have to explore and discover the biophysical structure of a particular resource that will differ on key parameters from similar resources in the same region. Further, they have to cope with considerable uncertainty related to the weather, complicated growth patterns of biological systems that may at times be chaotic in nature, and external price fluctuations that affect the costs of inputs and the value of outcomes. In addition to the physical changes that they can make in the resource, seven clusters of rules directly affect the components of their own action situations.² Specifically, the rules they can change include:

1. Boundary rules affect the characteristics of the participants.
2. Position rules differentially affect the capabilities and responsibilities of those in positions.

III. Authority rules affect the actions that participants in positions may, must, or must not do.
IV. Scope rules affect the outcomes that are allowed, mandated, or forbidden.
V. Aggregation rules affect how individual actions are transformed into final outcomes.
VI. Information rules affect the kind of information present or absent in a situation.
VII. Payoff rules affect assigned costs and benefits to actions and outcomes.

Given the nonlinearity and complexity of common-pool resource situations, it is rarely easy to predict what effect a change in a particular rule will produce. For example, a change in a boundary rule to restrict the entry of users simultaneously reduces the number of individuals who are tempted to break authority rules, but it also reduces the number of individuals who monitor others or contribute funds toward hiring a guard. Thus, the opportunities for rule breaking may increase. Further, the cost of a rule infraction will be spread over a smaller group of users and, thus, the harm to any individual may be greater. Assessing the overall effects of a change in boundary rules is a nontrivial analytical task. Instead of conducting such a complete analysis, users are more apt to use their intuitive understanding of the resource and each other to experiment with different rule changes until they find a combination that seems to work in their setting.

To understand the types of tools that are available to users somewhat better, let us examine in some detail the kind of boundary, authority, payoff, and position rules used in field settings. These four clusters of rules are the major tools used to affect the management of common-pool resources, while information, scope, and aggregation rules are utilized to complement changes induced by these four rules.

For the past 14 years, colleagues at or associated with the Workshop in Political Theory and Policy Analysis at Indiana University have studied a very large number of fisheries, forests, irrigation systems, and groundwater basins, as well as other common-pool resources. We have collected an immense archive of original case studies written by many scholars on all sectors in all parts of the world (Hess, 1999). We developed structured coding forms to help us identify the specific kinds of common-pool situations faced in the field, as well as the types of rules that users have evolved over time to try to govern and manage these resources effectively. In order to develop standardized coding forms, we read hundreds of cases describing how local common-pool resources were or were not regulated by a government

² This list of rules is the result of many years of theoretical and empirical work on a framework for conducting institutional analyses (see E. Ostrom et al., 1994: Chap. 2).
agency, by the users themselves, or by a nongovernmental organization (NGO).

A. Affecting the Characteristics of Users through Boundary Rules

The most frequent recommendation concerning boundary rules in the policy literature is to limit the number of persons allowed to appropriate from a common-pool resource so that the level of appropriation is reduced or to require users to obtain a license before harvesting. Boundary rules affect the types of participants with whom other participants will be interacting. If contingent cooperation is perceived to be a possibility, then one of the most important ways to enhance the likelihood of using reciprocity norms is to increase the proportion of participants who are well known in a community, have a long-term stake in that community, and would find it costly to have their reputation for trustworthiness harmed in that community. Reducing the number of users while opening the resource to strangers who are willing to pay a license fee, but who lack a long-term interest in the sustainability of a particular resource, may reduce the level of trust and willingness to use reciprocity and thus increase enforcement costs substantially.

We have identified 27 boundary rules as having been used in at least one common-pool resource somewhere in the world (E. Ostrom et al., 1989). While some systems use only a single boundary rule, many use two or three of these rules in combination. Boundary rules can be broadly classified in three general groups defining how individuals gain authority to enter and appropriate resource units from a common-pool resource. The first type of boundary rule relates to an individual's citizenship, residency, or membership in a particular organization. Many forestry and fishing user groups require members to have been born in a particular location. A second broad group of rules relates to individual ascribed or acquired personal characteristics. Other user groups may require that appropriation depends on ethnicity, clan, or caste. A third group of boundary rules relates to the relationship of an individual with the resource itself. Using a particular technology or acquiring appropriation rights through an auction or a lottery are examples of this type of rule. About half of the rules relate to the characteristics of the users themselves. The other half involve diverse relationships with the resource.

In a systematic coding of those case studies for which sufficient information existed about rules related to inshore fisheries in many parts of the world, E. Schlager coded 33 user groups out of the 44 groups identified as having at least some rules regarding the use of the resource. All 33 groups depended on some combination of 14 different boundary rules (Schlager, 1994: 258). None of these groups relied on a single boundary rule. Thirty out of 33 groups (91%) limited fishing to those individuals who lived in a nearby community, while 13 groups also required membership in a local organization. Consequently, most inshore fisheries organized by the users themselves restrict fishing to those individuals who are well known to each other, have a relatively long-term time horizon, and are connected to one another in multiple ways.

After residency, the next most frequent type of rule, used in two-thirds of the organized subgroups, involves the type of technology that a potential fisher must use. These rules are often criticized by policy analysts, since gear restrictions tend to reduce the "efficiency" of fishing. Gear restrictions have many consequences, however. Used in combination with authority rules that assign fishers using one type of gear to one area of the fishing grounds and fishers using another type of gear to a second area, they solve conflicts among incompatible technologies. Many gear restrictions also place a reduced load on the fishery itself and thus help to sustain longer-term use of the resource.

Other rules were also used. A scattering of groups used ascribed characteristics (age, two groups; ethnicity, three groups; race, five groups). Three types of temporary use rights included government licenses (three groups), lottery (five groups), and registration (four groups). Seven groups required participants to have purchased an asset such as a fishing berth, while three groups required ownership of nearby land. Schlager did not find that any particular boundary rule was correlated with higher performance levels across the groups. She did find, however, that the 33 groups who had at least one boundary rule tended to be able to solve common-pool problems more effectively than the 11 groups who had not crafted boundary rules.

Thus, many of the rich diversity of boundary rules used by users in the field attempt to ensure that the users will be relating to others who live nearby and have a long-term interest in sustaining the productivity of the resource. One way of coping with the commons is thus changing the composition of who uses a common-pool resource to increase the proportion of participants who have a long-term interest, who are more likely to use reciprocity, and who can be trusted. Central governments tend to use a smaller set of rules and some of these may open up a resource
B. Affecting the Set of Allowable Actions through Authority Rules

Authority rules are also a major type of rule used to regulate common-pool resources. We identified a diversity of authority rules used in field settings. Some rules involve a simple formula. Many forest resources, for example, are closed to all forms of harvesting during one portion of the year and open for extraction by all who meet the boundary rules during an open season. Most authority rules, however, have two components: an assignment and a basis. A fisher, for example, might be assigned to a fixed location (a fishing spot) or to a fixed rotational schedule, or a member of the founding clan may be assigned (allowed) to cut timber anywhere in a forest. In addition to the assignment used in an authority rule, most rules required a basis for the assignment. For example, a fisher might be assigned to a fixed location based on a number drawn in a lottery, on the purchase of that spot in an auction, or on the basis of his or her historical use.

If all of the bases were likely to be combined with all of the assignment possibilities, there would be 112 different authority rules (8 assignment formulas times 14 bases). A further complication is that the rules for one product may differ from those of another product in the same resource. In regard to forest resources, for example, children may be authorized to pick fruit from any tree located in a forest so long as it is for their own consumption, women may be authorized to collect so many headloads of dead wood for domestic firewood and certain plants for making crafts, while shamans are the only ones authorized to collect medicinal plants from a particular location in a forest. Appropriation rights to fish are frequently related to a specific species. Thus, the exact number of rules that are actually used in the field is difficult to compute since not all bases are used with all formulas, but many rules focus on specific products. A still further complication is that the rules may regularly change over the course of a year, depending on resource conditions.

Schlager (1994: 259–260) found that all 33 organized subgroups used one of five basic assignments in their authority rules. Every user group included in her study assigned fishers to fixed locations using a diversity of bases, including technology, lottery, or historical use. Thus, spatial demarcations are a critical variable for inshore fisheries. Nine user groups required fishers to limit their harvest to fish that met a specific size requirement, while seven groups allocated fishers to fishing spots using a rotation system and seven other groups only allowed fishing locations to be used during a specific season. Four groups allocated fishing spots for a particular time period (a fishing day or a fishing season).

An important finding—given the puzzles addressed in this article—is that the authority rule most frequently recommended by policy analysts is not used in any of the coastal fisheries included in Schlager's study. No attempt was made "by the fishers involved to directly regulate the quantity of fish harvested based on an estimate of the yield. This is particularly surprising given that the most frequently recommended policy prescription made by fishery economists is the use of individual transferable quotas based on estimates on the economically optimal quantity of fish to be harvested over the long run" (Schlager, 1994: 265). In an independent study of 30 traditional fishery societies, James Acheson and colleagues also noted the surprising absence of quota rules:

All of the rules and practices we found in these 30 societies regulate "how" fishing is done. That is, they limit the times fish may be caught, the locations where fishing is allowed, the technology permitted, and the stage of the life cycle during which fish may be taken. None of these societies limits the "amount" of various species that can be caught. Quotas—the single most important concept and tool of scientific management—is conspicuous by its absence (Acheson et al., 1998: 397).

Local inshore fishers, when allowed to manage a riparian area, thus use rules that differ substantially from those recommended by advocates of scientific management. Fishers have to know a great deal about the ecology of their inshore region, including spawning areas, nursery areas, the migration routes of different species, and seasonal patterns, just to succeed as fishers. Over time, they learn how "to maintain these critical life-cycle processes with rules controlling technology, fishing locations, and fishing times. Such rules in their view are based on biological reality" (Acheson et al., 1998: 405).

The diversity of rules devised by users greatly exceeds the limited authority rules recommended in textbook treatments of this problem. Users thus cope with the commons by a wide variety of rules affecting the actions available to participants and thus their basic set of strategies. Given this wide diversity of rules, it is particularly noteworthy that rules assigning users a right to a specific quantity of a resource are used so infrequently in inshore fisheries. (They are
used more frequently when allocating forest products, where the quantity available, as well as the quantity harvested, is much easier to measure.) To assign a user a specific quantity of a resource unit requires that those making the assignment know the total available units.

C. Affecting Outcomes through Payoff and Position Rules

One way to reduce or redirect the appropriations made from a common-pool resource is to change payoff rules so as to add a penalty to actions that are prohibited. Many user groups also adopt norms that those who are rule breakers should be socially ostracized or shunned and individual users tend to monitor each other's behavior rather intensively. Three broad types of payoff rules are used extensively in the field: (1) the imposition of a fine, (2) the loss of appropriation rights, and (3) incarceration. The severity of each of these types of sanctions can range from very low to very high and tends to start out on the low end of the scale. Inshore fisheries studied by Schlager relied heavily on shunning and other social norms and less on formal sanctions.

Passing rules that impose costs is relatively simple. The real difficult task is monitoring behavior to ascertain if rules are being broken. Self-organized fisheries tend to rely on self-monitoring more than the creation of a formal position of guard. Most inshore fishers now use shortwave radios as a routine part of their day-to-day operations, allowing a form of instant monitoring to occur. An official of a West Coast Indian tribe reports, for example, that “it is not uncommon to hear messages such as ‘Did you see so-and-so flying all that net?’ over the shortwave frequency—a clear reference to a violation of specified gear limits” (cited in Singleton, 1998: 134). Given that most fishers will be listening to their shortwave radio, “such publicity is tantamount to creating a flashing neon sign over the boat of the offender. Such treatment might be proceeded or followed by a direct approach to the rule violator, advising him to resolve the problem. In some tribes, a group of fishermen might delegate themselves to speak to the person” (cited in Singleton, 1998: 134).

Among self-organizing forest governance systems, creating and supporting a position as guard is frequently essential, since resource units are highly valuable and a few hours of stealth can generate substantial illicit income. Monitoring rule conformance among forest users by officially designated and paid guards may make the difference between a resource in good condition and one that has become degraded. In a study of 279 forest panchayats in the Kumaon region of India, Agrawal and Yadama (1997) found that the number of months a guard was hired was the most important variable affecting forest conditions. The other variables that affected forest conditions included the number of meetings held by the forest council (a time when infractions are discussed) and the number of residents in the village.

It is evident from the analysis that the capacity of a forest council to monitor and impose sanctions on rule-breakers is paramount to maintaining the forest in good condition. Nor should the presence of a guard be taken simply as a formal mechanism that ensures greater protection. It is also an indication of the informal commitment of the panchayat and the village community to protect their forests. Hiring a guard costs money... If there was scant interest in protecting the forest, villagers would have little interest in setting aside the money necessary to hire a guard (Agrawal and Yadama, 1997: 455).

Boundary and authority rules also affect how easy or difficult it is to monitor activities and impose sanctions on rule infractions. Closing a forest or an inshore fishery for a substantial amount of time, for example, has multiple impacts. It protects particular plants or fish during critical growing periods and allows the entire system time to regenerate without disturbance. Further, during the closed season, rule infractions are highly obvious to anyone as any user in the resource is almost certainly breaking the rules. Similarly, requiring users to use a particular technology may reduce the pressure on the resource, help to solve conflicts among users of incompatible technologies, and also make it very easy to ascertain if rules are being followed. Changing payoff rules is the most direct way of coping with commons dilemmas.

D. Affecting Outcomes through Changes in Information, Scope, and Aggregation Rules

Information, scope, and aggregation rules tend to be used in ways that complement changes in boundary, authority, payoff, and position rules. Individual systems vary radically in regard to the mandatory information that they require. Many smaller and informal systems rely entirely on a voluntary exchange of information and on mutual monitoring. Where resource units are valuable and the size of the group is large, more and more requirements are added regarding the information that must be kept by users.
V. EXPERIMENTING WITH RULE CHANGES

Instead of assuming that designing rules that approach optimality, or that even improve performance, is a relatively simple analytical task that can be undertaken by distant, objective analysts, we need to understand the policy design process as involving an effort to tinker with a large number of component parts (see Jacob, 1977). Those who tinker with any tools—including rules—are trying to find combinations that work together more effectively than other combinations. Policy changes are experiments based on more or less informed expectations about potential outcomes and the distribution of these outcomes for participants across time and space (Campbell, 1969, 1975). Whenever individuals agree to add a rule, change a rule, or adopt someone else’s proposed rule set, they are conducting a policy experiment. Further, the complexity of the ever-changing biophysical world combined with the complexity of rule systems means that any proposed rule change faces a nontrivial probability of error.

When there is only a single governing authority, policymakers have to experiment simultaneously with all of the common-pool resources within their jurisdiction with each policy change. And, once a change has been made and implemented, further changes will not be made rapidly. The process of experimentation will usually be slow, and information about results may be contradictory and difficult to interpret. Thus, an experiment that is based on erroneous data about one key structural variable or one false assumption about how actors will react can lead to a very large disaster (see Wilson et al., 1999). In any design process where there is substantial probability of error, having redundant teams of designers has repeatedly been shown to have considerable advantage (see Landau, 1969, 1973; Bendor, 1985).

VI. SELF-ORGANIZED RESOURCE GOVERNANCE SYSTEMS AS COMPLEX ADAPTIVE SYSTEMS

As discussed in Section II, the very concept of organization is closely tied for many scholars to the presence of a central director who has designed a system to operate in a particular way. Consequently, the mechanisms used by noncentrally directed systems are not always well understood. Many self-organized resource governance
systems are invisible to the officials of their own country or those from donor agencies. A classic example of this occurred in the Chitwan valley of Nepal several years ago when an Asian Development Bank team of irrigation engineers recommended a very large loan to build a dam across the Rapti River to enable the farmers there to irrigate their crops. What the engineering design team did not see was 85 farmer-managed irrigation systems that already existed in the valley and that had achieved relatively high performance. Most farmers in the Chitwan valley already obtained three irrigated crops a year as a result of their participation in the activities of these irrigation systems (see Benjamin et al., 1994).

In contrast to forms of organization that are the result of central direction, most self-organized groups—including the types of locally organized users of fisheries and forests discussed in this article—are better viewed as complex adaptive systems. These systems are composed of a large number of active elements whose rich patterns of interactions produce emergent properties that are not easy to predict by analyzing the separate parts of a system. Holland (1995: 10) views complex adaptive systems as "systems composed of interacting agents described in terms of rules. These agents adapt by changing their rules as experience accumulates." Such systems "exhibit coherence under change, via conditional action and anticipation, and they do so without central direction" (Holland, 1995: 38–39). Holland points out that complex adaptive systems differ from physical systems that are not adaptive and that have been the foci of most scientific effort. It is the physical sciences that have been the model for many aspects of contemporary social science. Thus, the concepts needed to understand the adaptivity of systems are not yet well developed by social scientists.

A. Properties and Mechanisms of Complex Adaptive Systems

No general theory of complex adaptive systems yet exists to provide a coherent explanation for processes shared by all complex adaptive systems. Biologists have studied many different adaptive systems but within separate fields of biology. Thus, even biologists have not recognized some of the similarities of structures and processes that characterize the central nervous system, the immune system, and the evolution of species. Recent work at the Santa Fe Institute has begun to identify central attributes, mechanisms, and processes used by all complex adaptive systems including both biological systems, as well as markets and other social systems that are not centrally directed.

It appears that all complex adaptive systems share four basic properties: nonlinearity, flows, diversity, and aggregation. The first three properties are self-evident and clearly characterize the types of self-organized resource governance systems discussed in this article. Aggregation refers to the "emergence of complex larger-scale behavior from the aggregate interactions of less complex agents" (Holland, 1995: 11). For example, many irrigation systems are divided into several tiers and multiple units at each of these tiers. All of the farmers on a field irrigation channel are responsible for distributing the water to this small channel as well as keeping it in good repair. All farmers whose field channels are served by a branch canal may send a representative to a branch canal organization that focuses its attention on the distribution of water among all branches and on the maintenance of the distribution channels. The branch canal organization may send a representative to a central committee who is responsible for the headworks that divert the water from a river into the system in the first place. The rules used on one branch canal or one field channel may be quite different than on others. There is no single center of authority for these systems that makes all relevant decisions on how to get water from the river to a farmer's field, but in many farmer-organized systems the water is distributed in an organized fashion and all of the waterworks are maintained as a result of the aggregation of decisions and actions at multiple levels.

In addition to these four attributes, complex adaptive systems also use three mechanisms that are key to the adaptive process itself. These include the use of tags, internal models, and building blocks.

1. The Use of Tags

Tagging is a universal mechanism for boundary formation and aggregation of units in complex adaptive systems. "Tags are a pervasive feature of complex adaptive systems because they facilitate selective interactions. They allow agents to select among agents or objects that would otherwise be indistinguishable" (Holland, 1995: 14). All of the types of boundary rules discussed here involve the specification of the tags that will be used to determine who is authorized to be a co-user from a common-pool resource. Residency, prior membership, and personal characteristics are attributes that already exist and are easy to use as boundary tags. Local governance systems rely heavily on tags that identify individuals who are already known to each other, who have a long-term stake in the sustainability of a re-
source, who have an incentive to build a reputation for being trustworthy, and who are thus likely to extend reciprocity rather than recalcitrance in dealing with joint problems.

Tags are also used extensively to mark locations in a resource, to warn rule infractors, and even to mark individual organisms that need to be treated in a special way. An example of the latter occurs along the Maine coast, where it is forbidden to harvest berried lobsters (those with eggs). Such lobsters are V-notched and returned back to the sea. Any other fisher who captures a V-notched lobster is also supposed to return it to the sea.

2. Internal Models

The users from a common-pool resource build internal models of the resource, of the relationships among the components of the resource, and of where their own actions positively or negatively affect one another and the resource. Among the shared lore for most fishing villages is a clear understanding of where fish breed, where young fish tend to cluster, the length of time it takes for fish to be mature and reproduce, the migration patterns of fish, the food chain in a location, and other information. Many inshore fishers develop their own maps of all of the fishing spots in their grounds. In an effort to reduce the interference of one boat with another boat's fishing, these are frequently defined so that if all "spots" are filled, all boats are still able to have a good chance to catch fish. These maps are then used in a variety of allocation rules that specify the basis for how any particular boat is assigned to a particular fishing spot. Users of forests also map their forests and may create refugia—sometimes as sacred forests—for sections of a forest that are particularly rich in biodiversity. By not harvesting from these refugia, they serve as a source of regeneration to other nearby locations that are disturbed through harvesting.

3. Building Blocks

Building blocks are ways of breaking down complex processes into small chunks that can be used in multiple ways and combined and re-combined repeatedly and at diverse levels. Once an authority rule that allocates resource units on some basis is determined, for example, using the same basis again to allocate responsibility for maintenance work is considered to be a fair allocation of benefits and costs in many cultures and is relatively easy to remember. On the large Chhattis Mauja farmer-organized system in Nepal, for example, water was originally allocated by the land area served. In the 1950s, the formula used for maintenance work was that each branch canal was responsible to send one person to work on the main canal for each 17 hectares of area it irrigated. "The term used for a person-day of labor for canal maintenance was kulara. Since the share of water a branch canal is entitled to receive is the same as the resource mobilization requirement, water allocation is now also referred to as 'so many kularas of water'" (Yoder, 1991: 7). As the system has grown, the total number of kularas has now been set at 177 shared among 44 branch canals. Voting rights are now also set in terms of kularas. "Therefore, a branch canal with five kularas was entitled to 5/177 of the water in the main canal, responsible to supply 5/177 of the resources mobilized for the irrigation, and had five of the total 177 votes in all important decisions" (Yoder, 1991: 7).

B. Changing Rules as an Adaptive Process

Given the logic of combinatorics, it is impossible—as we showed earlier—to conduct a complete analysis of the expected performance of all of the potential rule changes that could be made by the individuals who are served by a self-organized resource governance system and are trying to improve its performance. A similar impossibility also exists for many biological systems. Let us explore these similarities.

Self-organizing resource governance systems have two structures that are somewhat parallel in function to the concepts of a genotype and a phenotype in biology. Phenotypic structures characterize an expressed organism—how bones, organs, and muscles develop, relate, and function in an organism in a particular environment. The components of an action situation characterize an expressed situation—how the number of participants, the information available, and their opportunities and costs create incentives, and how incentives lead to types of outcomes in a particular environment. The genotypic structure characterizes the set of instructions encoded in DNA to produce an organism with a particular phenotypic structure. A rule configuration is a set of instructions on how to produce the structure of relationships among individuals in an action situation, which is also affected by the biophysical world and the kind of community or culture in which an action situation is located.

The evolution of social systems does not follow the same mechanisms as the evolution of species. As an evolutionary process, of course, there must be the generation of new alternatives, selection among new and old combinations of structural attributes, and retention of those combinations of attributes that are successful in a particular environment. In evolving biological sys-
tems, genotypic structures are changed through mechanisms such as crossover and mutation, and the distribution of particular types of instructions depends on the survival rate of the phenotypes that they produce in given environments. Instead of blind variation, however, human agents do try to use reason and persuasion in their efforts to try to devise better rules, but the process of choice always involves experimentation.

Rule configurations can be represented as a string of symbols that describe which rules are in effect in a particular location. In Table 1, the left-hand column lists a small subset of the rules typically found in the field. A “1” has been entered in a cell of a column to indicate when a specific rule is in use, and a “0” has been entered when the rule is not in use. Thus, the eight columns represent the simplified rule configurations for eight hypothetical locations.

Location A would be characterized as one at which the participants are required both to live in a local community and to use geographic assignments or a rotation system, but also at which there are no formal sanctions used or guards hired to monitor conformance to the entry rules and rotation system. Further, no records are kept, there are no refugia, and users can work individually or in teams. In contrast, Location C creates a refugia in which no appropriation may be undertaken, and does not allocate the remaining space to individual users. If there is a very good breeding and nursery area in this fishery, and it can be well demarcated so that it is obvious if someone has ventured into the refugia by error (or by intent), a rule system that does not limit fishing in any other way may be a very effective system for protecting the regenerative capabilities of a fishery. It is certainly less expensive than naming specific fishing spots, making an allocation of each fishing team to particular spots, and enforcing a more complicated rule system. At Location C, users also need to live in the local area, and they do face fines if a fellow user notices that they have appropriated from the refugia. Though a guard is not created, records are kept of any infractions noted through the mutual monitoring process.

Location H utilizes a much fuller array of rules. Users must belong to a local cooperative, even though they do not have to live in the local community. Both a refugia and a system for allocating time in which to conduct appropriation have been created, although, with the exception of the forbidden area of the refugia, the latter is not spatially defined. Payoff rules have been modified so that there are both fines for infractions as well as the possibility of losing rights to appropriate, if infractions are substantial. Records of infractions and of the time when someone appropriates are required, and individuals are constrained to working in teams rather than being allowed to appropriate as individuals.

Whether any or all of these rule configurations structured the incentives of users so that participants appropriate from their resource in a sustainable and efficient manner cannot be ascertained from learning about the rules alone. Whether a set of rules enhances performance depends on the structure of the biophysical system in that location and the willingness of participants to follow the rules. Whether the strategy of creating a refugia, as in C and H, works better than the geographic assignment, as in A, depends on how large the refugia is, how rapidly and extensively regeneration from the refugia spreads through the rest of the resource, and how well users understand each of these systems and try to make them work. All three rule systems are likely to work well in some locations and poorly in others.

With this form of representation we can begin to see how experimentation with rules may be similar to structural changes affecting other types of complex adaptive systems. Most systems are likely to start with one or two very simple rules. An obvious first candidate is to use tags to close the boundary to outsiders so that

<p>| TABLE I  |
| Rule Configurations and Appropriation Situations |</p>
<table>
<thead>
<tr>
<th>Rule configurations</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundary rules</td>
<td>A  B  C  D  E  F  G  H</td>
</tr>
<tr>
<td>Local residence</td>
<td>1  1  1  1  0  0  0  0</td>
</tr>
<tr>
<td>Authority rules</td>
<td>0  1  0  1  1  1  1  1</td>
</tr>
<tr>
<td>Payoff rules</td>
<td>0  0  0  0  0  0  0  0</td>
</tr>
<tr>
<td>Payoff rules</td>
<td>0  0  0  0  0  0  0  0</td>
</tr>
<tr>
<td>Position rules</td>
<td>0  0  0  0  1  1  1  1</td>
</tr>
<tr>
<td>Information rules</td>
<td>0  0  0  0  1  1  1  1</td>
</tr>
<tr>
<td>Scope rules</td>
<td>0  0  0  0  0  0  0  0</td>
</tr>
<tr>
<td>Aggregation rules</td>
<td>0  0  0  0  0  0  0  0</td>
</tr>
</tbody>
</table>

A “1” indicates when a specific rule is in use, and a “0” indicates when the rule is not in use. See text for details.
the likelihood of contingent cooperation and conformation to agreements will be enhanced. By only changing a few rules at the beginning, everyone can come to understand those rules while they are evaluating how they work. A second obvious candidate is to use the shared model of the environment built up through years of interaction in an environment to refine where harvesting should be undertaken and when. Space and time are obvious candidates for allocating access to resources in a manner that is relatively low in cost to sustain. If the community is small enough and shares common norms at a high enough level, creating formal sanctions, guards, records, and other rules may not be necessary. Thus, one can imagine a process in which a rule configuration of mostly zeros slowly converts over time to a scattering of ones and potentially to a rule configuration with many ones.

Changes in specific rules may come about through accident (such as forgetting a rule) or through specific collective-choice processes in which considerable time and effort are devoted to considering why performance needs to be enhanced and which rules might be changed. Since many users will have experience with more than one product (e.g., mushrooms from a forest), rules tested with regard to one product may also be applied with regard to others, if they are successful. Migration of individuals into a community brings individuals with repertoires of different rules used in other locations. Commerce with other groups lets users see and learn about others who may be doing better (or worse) than they in regulating a sustainable biodiverse and efficient resource system. Thus, a self-organized resource governance system with a higher level of immigration or greater communication with other localities is more likely to adapt and change rules over time than one in which few new ideas concerning how to use rules as tools are brought into the system. Trial-and-error processes may give relatively rapid feedback about rules that obviously do not work in a particular environment, but this is not always the case when the effect of human action on the environment has a long time delay. If all self-organized resource governance systems are totally independent and there is no communication among them, then each has to learn through its own trial-and-error process. Many will find that rules that they have tried do not work. Some will fail entirely.

The rate of change will differ among self-organized resource governance systems. As with all learning, the rate of change is an important variable that affects performance over time. If change occurs too rapidly, little is learned from each experiment before another experiment is launched. Respect for tradition and even religious mystification has been used to increase the retention of rules that are considered by at least some participants to work better. If the heavy hand of tradition is, however, too heavy and squelches innovation, a system that may have been well adapted to a past environment may find itself faltering as external changes occur without internal changes also occurring.

VII. THE ADVANTAGES AND LIMITS OF PARALLEL SETS OF LOCAL USERS IN POLICY EXPERIMENTS

Let us now discuss why a series of relatively autonomous, self-organized, resource governance systems may do a better job in policy experimentation than a single central authority. Among the advantages of authorizing the users of smaller-scale common-pool resources to adopt policies that regulate the use of these resources are:

- Local knowledge. Users who have lived and appropriated from a resource system over a long period of time have developed relatively accurate mental models of how the biophysical system itself operates, since the very success of their appropriation efforts depends on such knowledge. They also know others living in the area well and what norms of behavior are considered appropriate.
- Inclusion of trustworthy participants. Users can devise rules that increase the probability that others are trustworthy and will use reciprocity. This lowers the cost of relying entirely on formal sanctions and paying for extensive guarding.
- Reliance on disaggregated knowledge. Feedback about how the resource system responds to changes in actions of users is provided in a disaggregated way. Fishers are quite aware, for example, if the size and species distribution of their catch is changing over time.
- Better-adapted rules. Given the preceding, users are more likely to craft rules that are better adapted to each of the local common-pool resources than any general system of rules.
- Lower enforcement costs. Since local users have to bear the cost of monitoring, they are apt to craft rules that make infractions obvious so that monitoring costs are less. Further, by creating rules that are seen as legitimate, rule conformance will tend to be higher.
- Redundancy. The probability of failure throughout
an entire region is greatly reduced by the establishment of parallel systems of rule making, interpretation, and enforcement.

There are, of course, limits to all ways of organizing the governance of common-pool resources. Among the limits of a highly decentralized system are:

- Some users will not organize. While the evidence from the field is that many local users do invest considerable time and energy into their own regulatory efforts, other groups of users do not do so. There appear to be many reasons why some groups do not organize, including the presence of low-cost alternative sources of income and thus a reduced dependency on the resource, considerable conflict among users along multiple dimensions, lack of leadership, and fear of having their efforts overturned by outside authorities.

- Some self-organized efforts will fail. Given the complexity of the task involved in designing rules, some groups will select combinations of rules that generate failure instead of success. They may be unable to adapt rapidly enough to avoid the collapse of a resource system.

- Local tyrannies. Not all self-organized resource governance systems will be organized democratically or rely on the input of most users. Some will be dominated by a local leader or a power elite who only change rules that they think will give them further advantage. This problem is accentuated in locations where the cost of exit is particularly high and is reduced where users can leave when local decision makers are not responsible to a wide set of interests.

- Stagnation. Where local ecological systems are characterized by considerable variance, experimentation can produce severe and unexpected results, leading users to cling to systems that have worked relatively well in the past and to stop innovating long before they have developed rules likely to lead to better outcomes.

- Inappropriate discrimination. The use of identity tags is frequently an essential method for increasing the level of trust and rule conformance. However, tags based on ascribed characteristics can be the basis for excluding some individuals from access to sources of productive endeavor that has nothing to do with their trustworthiness.

- Limited access to scientific information. Whereas time and place information may be extensively developed and used, local groups may not have access to scientific knowledge concerning the type of resource system involved.

- Conflict among users. Without access to an external set of conflict-resolution mechanisms, conflict within and across common-pool resource systems can escalate and provoke physical violence. Two or more groups may claim the same territory and may continue to make raids on one another over a very long period of time.

- Inability to cope with larger-scale common-pool resources. Without access to some larger-scale jurisdiction, local users may have substantial difficulties regulating only a part of a larger-scale common-pool resource. They may not be able to exclude others who refuse to abide by the rules that a local group would prefer to use. In this situation, local users have no incentives to restrict their own use and watch others take away all of the valued resource units that they have not appropriated.

VIII. THE CAPABILITIES OF POLYCENTRIC SYSTEMS IN COPING WITH TRAGEDIES OF THE COMMONS

Many of the capabilities of a parallel adaptive system can be retained in a polycentric governance system. A polycentric system is one in which citizens are able to organize not just one but multiple governing authorities at differing scales (see McGinnis, 1999; V. Ostrom, 1997). Each unit may exercise considerable independence to make and enforce rules within a circumscribed scope of authority for a specified geographical area. In a polycentric system, some units are general-purpose governments and others may be highly specialized. Self-organized resource governance systems, in such a system, may be special districts, private associations, or parts of a local government. These are nested in several levels of general-purpose governments that also provide civil, equity, and criminal courts.

In a polycentric system, the users of each common-pool resource would have some authority to make at least some of the rules that are related to how that particular resource will be utilized, and thus would achieve most of the advantages of utilizing local knowledge, and the redundancy and rapidity of a trial-and-error learning process. On the other hand, problems associated with local tyrannies and inappropriate discrimination can be addressed in larger, general-purpose governmental units that are responsible for protecting the rights of all citizens and for the oversight of appro-
ropriate exercises of authority within smaller units of government. It is also possible to make a more effective blend of scientific information with local knowledge where major universities and research stations are located in larger units but have a responsibility to relate recent scientific findings to multiple smaller units within their region. Because polycentric systems have overlapping units, information about what has worked well in one setting can be transmitted to others who may try it out in their settings. Associations of local resource governance units can be encouraged to speed up the exchange of information about relevant local conditions and about policy experiments that have proved particularly successful. And, when small systems fail, there are larger systems to call upon—and vice versa.

Polycentric systems are themselves complex adaptive systems without one central authority dominating all of the others. Thus, there is no guarantee that such systems will find the optimal combination of rules at diverse levels that are optimal for any particular environment. In fact, one should expect that all governance systems will be operating at less-than-optimal levels, given the immense difficulty of fine-tuning any very complex, multifaceted system.

In the United States, there are many examples of dynamic, polycentric resource governance systems that display strong evidence of high performance. One example is the Maine lobster fishery, which is noteworthy because of the long-term, complementary roles adopted by both local and state governance systems. Maine is organized into riparian territories along most of the coast. Boundary rules and many of the day-to-day fishing regulations are organized by harbor gangs.

In order to go fishing at all, one must become a member of a “harbor gang,” the group of fishermen who go lobstering from a single harbor. Once one has gained admittance into such a group, one can only set traps in the traditional territory of that particular harbor gang. Members of harbor gangs are expected to obey the rules of their gang concerning fishing practices, which vary somewhat from one part of the coast to another. There is strong statistical evidence that the territorial system, which operates to limit the number of fishers exploiting lobsters in each territory, helps to conserve the lobster resource (Acheson et al., 1998: 400).

At the same time, the State of Maine has long established formal laws that protect the breeding stock and increase the likelihood that regeneration rates will be high. "At present, the most important conservation laws are minimum and maximum size measures, a prohibition against catching lobsters with eggs, and a law to prohibit the taking of lobsters which once had eggs and were marked—i.e. the 'V-notch' law" (Acheson et al., 1998: 400). Neither the state nor any of the harbor gangs has tried to limit the quantity of lobster captured. The state does not make any effort to limit the number of fishers since this is already done at a local level. However, the state has been willing to interfere when issues exceed the scope of control of local gangs. In the late 1920s, for example, when lobster stocks were at very low levels and many local areas appear to have had substantial compliance problems, the state took a number of steps—including threats to close the fishery—that supported informal local enforcement efforts. By the late 1930s, compliance problems were largely resolved and stocks had rebounded (although it cannot be shown that these two results are related, just correlated).

Recently, in response to changes that were breaking down the harbor gang system, the state formalized the system by dividing the state into zones with democratically elected councils. Each council was given authority over rules that have principally local impacts—trap limits, days and times fished, and so on. Interestingly, the formalization of local zones was followed, almost immediately, by the creation of an informal council of councils to address problems at a greater than local scale. It is expected that this council of councils will be formalized soon.

Today the state uses only about six patrol officers on the water to police the activities of 6800 lobstermen, all the other fisheries, and coastal environmental laws. During the 1990s, the lobster fishery has been growing substantially with increased yields. At the same time, there is strong evidence that the number of reproductive age females in Maine waters is very large and that the recruitment will continue at a high level.

IX. CONCLUSIONS

Widespread concern for conserving biodiversity is a reflection of the inherent and ecological value of such diversity. Those who advocate the necessity of sustaining biodiversity have frequently called for policy reforms that reduce the diversity of institutional arrangements for governing and managing complex common-pool resources. Yet, as Ashby long ago established, regulators need as much variety in their response capabilities as the systems they are regulating. To achieve biodiverse resource systems, we need to protect and enhance the institutional diversity of evolved gover-
nance regimes within the context of a broader multitiered, polycentric governance regime.

See Also the Following Articles

COMMONS, THEORY AND CONCEPT OF • GOVERNMENT LEGISLATION AND REGULATION • MARKET ECONOMY AND BIODIVERSITY

Acknowledgments

Support for the research on which this article is based has been provided by the MacArthur Foundation and the Latin American and Caribbean Regional Office of the Food and Agriculture Organization (FAO) of the United Nations. My thanks to Patty Dalecki and Tom Haris-Young for editing and preparing this manuscript.

Bibliography


I. Introduction

Interactions between species can be classified and defined in various ways. A very useful method of classification is to consider the "effect" that individuals of one species have on the population growth of another species and vice versa. We ask the question: In the presence of species A (1) increase its numbers (+), (2) not change its numbers (0), or decrease its numbers (−) relative to when species A is absent (−) or 0? The same question is asked of species A in the presence of species B. The answers are conveniently summarized in Table I.

Because of the symmetry there are only six types of interaction. These are frequently called (1) 00 neutralism, (2) +0 commensalism, (3) +− predator/prey, parasite/host, or herbivore/plant interactions, (4) 0− amensalism, (5) −− competition, and (6) ++ symbiosis. This article focuses on interactions of types (4) and (5) between individuals of different species. Amensal-
### TABLE I

A Classification of Two Species Interactions

<table>
<thead>
<tr>
<th>Effect of A on B</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of B on A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

is just an extreme example of asymmetric competition. This is competition, usually between a pair of species, in which the adverse effect of one species on the other is much greater than the reciprocal effect.

The negative effects on numbers are produced by individuals directly competing for essential resources (frequently food and space) that are in short supply. If the negative effect is caused by not getting enough of the limited resource, we call it exploitation competition. If the resource is not limiting, but individuals nonetheless harm each other in the process of obtaining it, we call it interference competition. An ecological situation can arise where two species appear to show the reciprocal negative effects associated with interspecific competition, but this is in fact the result of predation by a third species. This is called apparent competition.

Consider the situation shown in Fig. 1. A single species of predator attacks two species of prey. The predator/prey interactions are of \( + + \) type and therefore both species are adversely affected by the predator, and the predator is positively affected by both species of prey. This means that the positive effect that prey 1 has on the predator will, in turn, increase the negative effect on prey 2, and vice versa. The overall consequence of this is that there will appear to be a \( ++ \) interaction between prey 1 and prey 2, even if they are not competitors for any essential and limiting resource.

How have ecologists studied competition in order to see the patterns that this ecological process produces? In general, there are four approaches (in common with all studies in ecology and, in fact, all science). These are (1) mathematical models (equations, graphs, or computer simulations), (2) laboratory experiments (laboratory models), (3) field experiments (field models), and (4) field observations. Each approach has its advantages and disadvantages. As you progress from (1) through (2) and (3) to (4), the unit of study becomes more complex but more realistic. Approach (1) helps us to understand a process (or system) and suggests what kind of (2) and (3) studies may be useful. In turn, these may direct us towards particular kinds of field observation that will help confirm or reject the patterns suggested by (1), (2), and (3). In fact the study of competition (its historical development) progressed in part from (1) to (4) and it is convenient to do the same in this account of interspecific competition. The remaining sections illustrate these four approaches with three examples.

### II. COMPETITION MODELS

#### A. The Lotka–Volterra Model

Vito Volterra was an Italian mathematician who, inspired by his son-in-law Umberto D'Ancona (an eminent hydrobiologist), published work on competitive interactions in the 1920s and 1930s. A. J. Lotka was an American mathematician who published on similar topics, also in the 1920s and 1930s. Their names are associated with a basic model of interspecific competition that has had a major impact on how ecologists think about this species interaction.

The Lotka–Volterra model of interspecific competition is based on two other models of population growth, the exponential growth model and the logistic sigmoid growth model. Let us look first at exponential growth, that is, when there is no competition of any kind. Under such conditions the rate at which numbers change with time (\( \frac{dN}{dt} \)) can be represented by a per capita (per head) rate of increase multiplied by the number of individuals (\( N \)) in the population. The rate of increase used is the intrinsic rate of natural increase (\( r \)) and is equal to the difference between the per capita birth rate and the per capita death rate. In mathematical symbols

\[
\frac{dN}{dt} = rN
\]
In this simple model, the rate of increase does not change with density, that is, there is no density dependence and growth is unlimited. However, in most real populations, growth is a function of density ($fN$) and therefore the growth equation should be written as

\[ \frac{dN}{dt} = rN \]

which is the well-known logistic equation that produces sigmoidal limited population growth. It is important to remember that the term in brackets represents intraspecific density-dependent regulation. The inhibitory effect of one individual on population growth is $1/K$ and the inhibitory effect of $N$ individuals on population growth is $N/K$.

Now let us consider two species, each of which when living separately has its growth described by a logistic equation. However, when living together this logistic equation must be modified. In the following equations the two species are denoted by the subscripts 1 and 2. Let us consider first of all the population growth of species 1 when it competes with species 2. In terms of its own population growth, as described by the logistic equation, the inhibitory effect of one individual of species 1 on its own population growth is $1/K_1$, where $K_1$ is the carrying capacity of species 1. The inhibitory effect of every individual ($N_1$) of species 1 on its own population growth is therefore $N_1/K_1$. If species 2 was identical to species 1 in its effect on the population growth of species 1, then we could simply write the effect of every individual of species 2 on the population growth of species 1 as $N_2/K_1$. However, this is unlikely. It is more realistic to imagine that each individual of species 2 would have a greater or lesser effect on the population growth of species 1, and therefore we have to multiply $N_2/K_1$ by a constant that expresses this difference. This constant (usually called the competition coefficient and denoted by $\alpha_{12}$) describes the relative effect of an individual of species 2 compared to an individual of species 1. If an individual of species 2 was identical (from an ecological point of view), then $\alpha_{12} = 1$. Similarly, the effect of every individual of species 2 on its own population growth will be $N_2/K_2$ and the effect of every individual of species 1 on the population growth of species 2 will be $N_1/K_2$ multiplied by $\alpha_{21}$ (note the reversal of the subscripts).

To understand the competition coefficients ($\alpha_{12}$ and $\alpha_{21}$) more clearly, let us take an imaginary example in which competition is solely for a single food resource. Let us imagine that two species of desert mouse eat seeds. Mouse 1 consumes 100 seeds per day, and mouse 2 consumes 200 seeds per day. In terms of resource units (seeds), an individual of species 1 is half an individual of species 2, or two individuals of species 1 is one individual of species 2. Therefore, $\alpha_{12} = 2$ (the effect of adding one more individual of species 2 is like adding two more individuals of species 1) and $\alpha_{21} = 1/2$. 

![FIGURE 2](image-url)
We can now rewrite the two logistic growth equations for the two species (which incorporate intraspecific competition) to take into account interspecific competition:

\[
\frac{dN_1}{dt} = r_1 N_1 \left(1 - \frac{N_1}{K_1} - \frac{a_{12} N_2}{K_1} \right) \\
= r_1 N_1 \left( \frac{K_1 - N_1 - a_{12} N_2}{K_1} \right)
\]

\[
\frac{dN_2}{dt} = r_2 N_2 \left(1 - \frac{N_2}{K_2} - \frac{a_{21} N_1}{K_2} \right) \\
= r_2 N_2 \left( \frac{K_2 - N_2 - a_{21} N_1}{K_2} \right)
\]

This system of two species (the Lotka–Volterra model of two-species competition) will be at equilibrium when \( \frac{dN_1}{dt} = \frac{dN_2}{dt} = 0 \), that is, when both species are at their carrying capacities and population growth has stopped. Just considering for the moment species 1, this means that

\[
r_1 N_1 \left( \frac{K_1 - N_1 - a_{12} N_2}{K_1} \right) = 0
\]

Since \( r_1, N_1, \) and \( K_1 \) cannot be 0 (mathematically possible but ecologically uninteresting if species 1 cannot reproduce, is not present, or cannot exist in that habitat), this means that for species 1, \( K_1 - N_1 - a_{12} N_2 = 0 \) at equilibrium. Rearranging this equation we get \( N_1 = K_1 - a_{12} N_2 \), which is the equation of a straight line.
line and is called the species 1 zero isocline. On a graph of the numbers of species 1 \( (N_1) \) plotted against the numbers of species 2 \( (N_2) \), it represents all those combinations of \( N_1 \) and \( N_2 \) for which the population growth of species 1 \( (\frac{dN_1}{dt}) = 0 \).

Following a similar procedure for species 2 will produce the species 2 zero isocline, which is \( N_2 = \frac{K_2}{1 + \frac{N_1}{L_2}} \). On a graph of the numbers of species 2 \( (N_2) \) plotted against the numbers of species 1 \( (N_1) \), it represents all those combinations of \( N_1 \) and \( N_2 \) for which the population growth of species 2 \( (\frac{dN_2}{dt}) = 0 \).

There are four ways that these two straight lines can be placed on a graph of \( N_2 \) versus \( N_1 \), and these are illustrated in Fig. 3. These are known as phase plane graphs/diagrams. Each point on the graph represents a joint density of the two species. To each point in this graph we can attach an arrow, or vector, whose length and direction will indicate the dynamics of the system. Below its zero isocline, a species can increase in numbers. Above its zero isocline, a species will decline in numbers. Exactly on its zero isocline, a species will neither increase nor decrease. This means that the four graphs predict different outcomes for interspecific competition. In (a) either species can win depending on which area between the zero isoclines the population trajectory reaches. This depends on starting numbers and/or relative growth rates \( (r_1 \) and \( r_2) \). The point were the two lines cross represents an unstable equilibrium point. In (b) the two species coexist at the values of \( N_1 \) and \( N_2 \) were the two lines cross. This represents a stable equilibrium point. In (c) species 1 always wins and in (d) species 2 always wins.

Laboratory examples of one species winning (c and d) and coexistence (b) are provided by G. F. Gause’s experiments (see Section III,A). Examples of one species winning (c and d) and either species winning (a) are provided by T. Park’s experiments (see Section III,B). It is important to make two comments about the Lotka–Volterra model. First, it contains hidden variables. The competition coefficients \( (\alpha) \) are simply conversion factors to allow one species to be represented in the other species numerical equivalents. Species may compete for food and space and poison each other, but all of these mechanisms of competition are summarized and hidden within \( \alpha \). Second, because the real world (unlike most laboratory environments) is heterogeneous and patchy, even (a), (c), and (d) may result in coexistence. If within a habitat patch (a), (c), and (d) apply, there may only be one species per patch. But looking over many patches, both species will be seen to apparently “coexist.” This is the difference between what ecologists call ‘local’ and “regional” coexistence.

What do these graphs in Fig. 3 imply for competition and coexistence? Let us take graph (a) as an example (species 1 or species 2 wins depending on initial concentrations and/or the rates of increase). Using the end points of the lines, this graph can be defined as

\[
\alpha_{12} > \frac{K_2}{K_1} \quad \alpha_{21} > \frac{K_1}{K_2}
\]

or by dividing the two sides of the left-hand inequation by \( K_2 \) and the two sides of the right-hand inequation by \( K_1 \),

\[
\frac{\alpha_{12}}{K_1} > \frac{1}{K_1} \quad \frac{\alpha_{21}}{K_2} > \frac{1}{K_2}
\]

In competition terms this means that the effect of one individual of \( \text{species 2} \) on \( \text{species 1} \) is greater than the effect of one individual of \( \text{species 1} \) on \( \text{species 2} \).

In other words, each species reduces the \( \frac{dN}{dt} \) of the other species more than its own \( \frac{dN}{dt} \), which we can summarize as

\[
2 \text{ on 1} > 2 \text{ on 2} \quad 1 \text{ on 2} > 1 \text{ on 1}
\]

Interspecific effects for both species are stronger than intraspecific effects. For graph (b) (coexistence) these inequalities are

\[
2 \text{ on 1} < 2 \text{ on 2} \quad 2 \text{ on 2} < 1 \text{ on 1}
\]

which is the opposite of graph (a). Here each species reduces its own \( \frac{dN}{dt} \) more than that of the other species, and therefore interspecific effects for both species are weaker than intraspecific effects. This case has always intrigued ecologists, and the traditional explanation is that interspecific effects for both species should be weaker than intraspecific effects is “resource partitioning” (see Section II,B). However, there are other mechanisms that will produce this effect (see Section II,C).

For graph (c) (where species 1 always wins) the inequalities are

\[
2 \text{ on 1} < 2 \text{ on 2} \quad 1 \text{ on 2} > 1 \text{ on 1}
\]
which implies that the two species now do different things. Species 2 reduces its own \( \frac{dN}{dt} \) more than that of species 1, and species 1 reduces the \( \frac{dN}{dt} \) of species 2 more than its own \( \frac{dN}{dt} \).

For graph (d) (where species 2 always wins) the inequalities are

\[
2 \text{ on } 1 > 2 \text{ on } 2 \quad 1 \text{ on } 2 < 1 \text{ on } 1
\]

and we have the opposite of graph (c).

B. Resource Utilization Curves

The Lotka–Volterra competition model predicts that two potential competitors will coexist if they each affect the growth of the other more than they affect their own growth. That is, intraspecific effects are greater than interspecific effects for both species. This is summarized by the inequalities

\[
\frac{a_2}{K_1} < \frac{1}{K_1} \quad \frac{a_1}{K_2} < \frac{1}{K_2}
\]

Traditionally this led to the concept of resource separation (different niches) as a mechanism of coexistence. Clearly, if two species use different resources they will interact (compete) only with their own species, and therefore the preceding inequalities will be true. But partial resource partitioning has also been used widely as an explanation for coexistence. Field observations in particular (see Section V) have frequently found partial resource separation between species and accumulated these differences over several “niche dimensions” to convince the reader that enough separation exists to “explain” the coexistence observed. Unfortunately, most of these field studies do not calculate \( \frac{a_2}{K_1} \), \( \frac{a_1}{K_2} \), \( K_1 \), or \( K_2 \), and therefore we do not know if sufficient separation exists between the species to make \( \frac{a_2}{K_1} < \frac{1}{K_1} \) and \( \frac{a_1}{K_2} < \frac{1}{K_2} \). However, one attempt to quantify the “amount” of separation required is provided by a simple model, initially developed by Robert MacArthur (1972).

Imagine two species competing for a resource that can be visualized as varying continuously in one dimension, for example, insects or seeds of different sizes or foraging sites up a tree. Imagine also that the two species’ utilization of this resource follows a unimodal distribution, with a preferred area of use in the center and less preferred resource items on either side. Figure 4 shows this idea for two species using a resource \( j \), in which their resource utilization curves (\( u_1 \) and \( u_2 \)) are described by a normal distribution. Each species uses a different part of the resource, with some overlap. The distance between the two means is \( d \), and \( \sigma \) is the standard deviation. The two species are assumed to have the same carrying capacity and the intensity of competition is related to the area of overlap between the two distributions. With two normal distributions this is given by

\[
\alpha = e^{-\frac{d^2}{2\sigma^2}}
\]

and the relationship between \( \alpha \) and \( d/\sigma \) is shown in Fig. 5. For stable coexistence between two species,

\[
\text{FIGURE 4 Resource utilization curves used in the MacArthur model.}
\]

\[
\text{FIGURE 5 Relationship between } \alpha \text{ and the ratio } d/\sigma \text{ predicted from the MacArthur model.}
\]
$d$ should be approximately greater than unity, that is, $d$ (the distance between means) should be greater than $\sigma$ (the standard deviation of resource utilization). A field example that agrees with this amount of limiting similarity between species is provided by the foraging height relationships within antbirds (*Myrmotherula*), shown in Fig. 6. Unfortunately, it has been pointed out that $d/\sigma > 1$, and promoting coexistence may be a property of this type of model only with a one-dimensional resource. With competition in several dimensions and alternative resource utilization curves, lower values of $d/\sigma$ may be compatible with stable coexistence.

C. The Aggregation Model

Many insects exploit resources that are patchy, consisting of small, separate units, and that are ephemeral in the sense that they persist for only one or two generations. Such resources can include fruit, fungi, sap flows, decaying leaves, flowers, dung, carrion, seeds, dead wood, and small bodies of water held in parts of terrestrial plants (phytotelmata). This general view of insect ecology inspired the aggregation model of competition (Shorrocks et al., 1984; Atkinson and Shorrocks, 1981), which allows a competitively inferior species to survive in probability refuges. These are patches of resource (a single fungus, fruit, etc.) with no or a few superior competitors, that arise because the competing stages (usually larvae) have an aggregated distribution across the patches. These probability refuges are a permanent feature of such systems because patches, such as fungi, are ephemeral and aggregation increases mean crowding. Regional population density is limited by strong intraspecific competition in patches with high local density while low-density patches still exist (e.g., population size within a wood is limited by high density in some fungi, while other fungi still contain no or a few individuals). As with resource partitioning, coexistence is promoted because aggregation of the superior species increases its intraspecific competition and reduces interspecific competition.

In the aggregation model, the eggs of both insect species are independently distributed over the patches according to a negative binomial distribution, which has an exponent, $k$, inversely related to the degree of intraspecific aggregation. The use of the negative binomial and the assumption of independence have been justified for drosophilid flies. In the first version of the model, the parameter $k$ (level of aggregation) was constant and independent of density. This is not valid for real populations, but relaxing this assumption does not prevent coexistence. Within each patch, competition is modeled by a difference form of the Lotka–Volterra equations. The predictions of the aggregation model are that with $k$ of the negative binomial $<1$ (strong aggregation), it is virtually impossible for the competitively “superior” species to eliminate the competitively “inferior” species. Figure 7 shows the model’s results as a graph of “critical $k$” against $k$ of the negative binomial. Critical $k$ is the competition coefficient that the “superior species” must have in order to exclude the “inferior species.” Also shown on the graph are distributions of $k$ and $\alpha$ for drosophilid flies. For these flies it is clear that $k$ of the negative binomial is usually $<1$ and that competition coefficients are not sufficiently large to prevent competitive exclusion. For many animals exploiting ephemeral and patchy resources, this model therefore provides a viable alternative to traditional resource partitioning as an explanation for the coexistence of species. The two-species model has been extended to a many-species model (Shorrocks and Rosewell, 1987) and predicts average group sizes of about seven species coexisting on identical resources.
III. LABORATORY EXPERIMENTS

A. Yeast and Paramecium

In the early part of this century, G. F. Gause carried out a series of simple experiments with yeasts and protozoa that are instructive, are of historic interest, and set the experimental protocol for many of the so-called "bottle" experiments conducted in the 1940s, 1950s, and 1960s. His early experiments are detailed in his classic book *The Struggle for Existence* (Gause, 1934). He was mainly interested in testing the Lotka-Volterra model that had been put forward as a description of the competitive process, between species, a few years earlier.

The experimental design was simple. He grew two species, each alone and then together in mixed culture, to determine the course of competition under laboratory conditions. In his first experiments he used two species of yeast, which he called *Saccharomyces* and *Schizosaccharomyces*. However, we are not now sure of the exact taxonomic identity of the two species. He grew these species on an extract of brewer’s yeast, water, and sugar. The growth medium was not renewed, but growth ceased before the sugar resource was exhausted, apparently because of the accumulation of alcohol. These experiments with yeast are therefore an example of interference competition rather than exploitation competition. The main interest of these experiments lies in the calculation of the competition coefficients ($\alpha$), which Gause did by estimating all the variables in the Lotka-Volterra competition equations except $\alpha$.

The intrinsic rate of natural increase ($r_1$ and $r_2$) and the carrying capacity ($K_1$ and $K_2$) were estimated by fitting a logistic equation to the single-species growth curves. From the mixed cultures, he estimated both the numbers of each species ($N_1$ and $N_2$) and the growth rates ($dN_1/dt$ and $dN_2/dt$). By rearranging the Lotka-Volterra equations he obtained

\[
\alpha_1 = \frac{K_2 - dN_1/dt \cdot K_1}{r_1 \cdot N_1}
\]

\[
\alpha_2 = \frac{-dN_2/dt \cdot K_1}{r_2 \cdot N_2}
\]

Since all quantities on the right-hand side are known, he could estimate the competition coefficients ($\alpha_1$ and $\alpha_2$). Since $N_1$, $N_2$, $dN_1/dt$, and $dN_2/dt$ all vary with time, Gause selected three occasions during his experiments to estimate $\alpha$. He then took the mean of these three calculations. His experiments were also repeated under two environmental conditions, which he called anaerobic (yeasts grown in test tubes) and aerobic (yeasts grown in flasks). Since $\alpha$ is a measure of the effect that...
one individual of species 1 has on the population growth of species 2 (and vice versa), and with these species competition is thought to be mainly the result of alcohol poisoning, it should be possible to calculate $\alpha$ directly from the relative production of alcohol by the two species. Table II shows the values of $\alpha$ calculated from population experiments and relative alcohol production, under both anaerobic and aerobic conditions.

Two important points come out of these yeast experiments. First, the competition coefficients are not constant; they vary (quite markedly) with different environmental conditions. Second, relative alcohol production and its effect on growth are sufficient to explain the interference competition between these two species (particularly under aerobic conditions). Therefore, in this experiment, the competition coefficients are relatively simple, pure quantities and $\alpha_2 = 1/\alpha_1$.

In his later and perhaps more famous experiments, Gause used two species of single-celled protozoa called Paramecium. Initially he used what he called P. caudatum (1) and P. aurelia (2), although once again we are not sure about the precise taxonomy because these "species" are now known to be closely related groups of species. The Paramecium were grown in 5 cm$^3$ of Osterhau's medium (a mixture of salts) in test tubes at 26°C. Each day fresh food was added to the test tubes. The food was a bacterium grown on agar plates and added to the cultures using a platinum loop (to standardize the amount given). The bacterium used did not grow in Osterhau's medium. Each experiment was started with 20 Paramecium individuals on Day 0, and Gause estimated the number of individuals each day by sampling 10% of the medium. Since an individual of P. aurelia occupies only 39% of the volume of a single P. caudatum, he estimated the volume of a typical cell for each species and converted the estimated numbers to biomass (in volume). Following the same protocol as the yeast experiments, the two species were grown alone (four cultures of P. caudatum and three cultures of P. aurelia) and in mixed culture (three cultures); Fig. 8 shows the mean volumes, over time, for these experiments. In single culture, both species appear to grow to an equilibrium at around $K = 200$. In mixed culture, P. caudatum appears to suffer more from interspecific competition than does P. aurelia, although the latter also grows less well. Using the same procedure as for the yeast experiments, Gause was able to estimate the competition coefficients ($\alpha_1$ and $\alpha_2$), and these are given in Table III.

Notice that in the first few days the competition coefficient for the effect of P. aurelia on P. caudatum is negative. This implies that the effect of P. aurelia is actually positive at this stage. Given that $K_1 = K_2 = 200$, $\alpha_1 = 1.64$, and $\alpha_2 = 0.61$, then we can calculate $K_2/\alpha_2 = 122$ and $K_1/\alpha_1 = 328$ and therefore construct the Lotka–Volterra phase-plane for this experiment. This is shown in Fig. 9, along with the two-species population trajectory for the results in Fig. 8. These experiments with Paramecium appear to broadly confirm the predictions of the Lotka–Volterra model of interspecific competition, although the mortality imposed by the sampling procedure (10%) and the premature termination of the experiments result in the predicted equilibrium not being reached in 25 days.

In another set of experiments using P. aurelia and P. bursaria, Gause obtained population trajectories that suggested coexistence, although again the experiments were not run for long enough to reach an actual equilibrium. However, irrespective of the starting numbers, the two-species population trajectories converge on the same joint densities. Paramecium bursaria is interesting since it contains within its cell body unicellular green plants. The suggestion has been made that P. bursaria was able to exploit the bacterial food that had settled to the bottom of the experimental tube. This area of

### Table II

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Alcohol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaerobic</td>
<td>$\alpha_1$</td>
</tr>
<tr>
<td></td>
<td>0.64</td>
</tr>
<tr>
<td>Aerobic</td>
<td>$\alpha_1$</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
</tr>
</tbody>
</table>

### Table III

<table>
<thead>
<tr>
<th>Experiments</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>First days</td>
<td>$-1.64$</td>
<td>$+0.61$</td>
</tr>
<tr>
<td>Greater than fifth day</td>
<td>$+1.64$</td>
<td>$+0.61$</td>
</tr>
</tbody>
</table>
FIGURE 8 Growth in ‘volume’ of Paramecium caudatum and P. aurelia in both single and mixed culture (from Gause, 1934). Volumes are converted from number of individuals per 0.5 cc in the culture. Since the volume of P. caudatum was set at 1.00 and that of P. aurelia at 0.39, the volume for P. caudatum is equal to the number of individuals per 0.5 cc.

The microhabitat is depleted of oxygen (by the bacteria) and cannot be used by P. aurelia. However, P. bursaria carries its own oxygen supply (produced by its unicellular plants) and can use this part of the bacterial resource. Heterogeneity within this simple habitat, produced by the bacterial food, may have been sufficient to allow resource partitioning.

B. Flour Beetles

In a series of studies, Thomas Park carried out competition experiments between two species of flour beetle, Tribolium confusum (1) and T. castaneum (2). Both are frequently found as stored-product pests and they can be conveniently raised under laboratory conditions. Populations were kept in glass tubes with 8 g of medium (95% flour and 5% yeast). All adults, pupae, and larvae were counted every 30 days and both single- and mixed-species populations were kept. An interesting aspect of the experimental design, which proved instructive, was that populations were maintained at a number of temperatures (24°, 29°, and 34°C) and relative humidities (30% and 70%), and each type of population (single or mixed), under all six environmental conditions, was replicated 30 times.

1. Single-Species Results

Apart from T. castaneum at 24°C and 30% humidity, all single-species populations persisted, and Table IV shows the average population densities during the course of the experiment. The value of 3, for T. castaneum at 24°C and 30% humidity, is the average population density until extinction.

Using these average densities as an indication of success, Park compared the relative performance of these two species (alone) for each temperature/humidity combination. The results of this comparison are shown in Table V, where an asterisk indicates a statistically significant difference between the two species. Except for 34°C and 70% humidity, each of the physical environments significantly favors one species over another. These results are due to intraspecific competition and the effect of the physical environment on it.
2. Two-Species Results

Park found that while single-species populations persisted (with the exception of T. castaneum at 24°C and 30% humidity), mixed-species populations did not. In each population only one species persisted and the other was eliminated. However, unlike most previous “bottle” experiments (e.g., Gause’s experiments), it was not always the same species that was eliminated, even under the same environmental conditions. This “indeterminate” outcome of competition is one of the predicted possible outcomes of the Lotka–Volterra model. Table VI shows the percentage of replicates that T. castaneum won.

These experiments illustrate two interesting and possibly universal features of competition. First, the outcome of two-species competition can sometimes be predicted from the performance of each species on its own (i.e., 29°C/70% and 24°C/30%), but also sometimes it cannot (i.e., 34°C/70% and 24°C/70%). Second, the outcome of competition between two species can vary along an environmental gradient. Park’s experiments go from hot-humid conditions (34°C/70%) to cool-dry conditions (24°C/30%). At the hot-humid end, T. castaneum eliminates T. confusum each time. At the cool-dry end, T. confusum eliminates T. castaneum each time. In the middle region either species can win, but T. castaneum wins more often toward the hot-humid end, and T. confusum wins more often toward the cool-dry end.

If we imagine that the isoclines in the Lotka–Volterra phase-plane graphs can move their position depending on environmental conditions, then these Tribolium results are still understandable in terms of the Lotka–Volterra model. Figure 10 presents a series of phase-plane graphs, going from hot-humid (left) to cool-dry (right). The arrow(s) in each diagram indicate the equilibrium point. Notice that the species predicted to win

<table>
<thead>
<tr>
<th>TABLE IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Densities of Beetles Observed in Park’s Single Species Experiments</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>T. confusum (1)</td>
</tr>
<tr>
<td>70%</td>
</tr>
<tr>
<td>30%</td>
</tr>
<tr>
<td>T. confusum (2)</td>
</tr>
<tr>
<td>70%</td>
</tr>
<tr>
<td>30%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison of Park’s Single-Species Beetle Experiments</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>70%</td>
</tr>
<tr>
<td>30%</td>
</tr>
</tbody>
</table>

* Asterisks indicate that T. confusum and T. castaneum maintained significantly different population sizes.

<table>
<thead>
<tr>
<th>TABLE VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Percentage of Replicates That T. castaneum Won in Park’s Beetle Experiments</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>70%</td>
</tr>
<tr>
<td>30%</td>
</tr>
</tbody>
</table>
moves from always *T. castaneum* (left), to either species, to always *T. confusum*.

**FIGURE 10** A series of phase-plane graphs along an environmental gradient from hot-humid (left) to cool-dry (right). The arrow(s) in each diagram indicate the equilibrium point.

In the late 1960s and early 1970s, Francisco Ayala conducted a series of laboratory experiments, each using a pair of *Drosophila* species. Although five species were used in different combinations, the basic design and outcome of all of these experiments were very similar. One experiment, between *D. serrata* and *D. pseudoobscura*, will be sufficient to illustrate the apparent dilemma that emerged.

*Drosophila serrata* comes from Australia and *D. pseudoobscura* comes from North America. Notice therefore that since neither of these flies comes from the same continent, let alone the same community, Ayala was simply using the flies like analog computers to find out what might happen between flies in real communities.

At 25°C, *D. serrata* eliminates *D. pseudoobscura* in a few generations; at 19°C, *D. pseudoobscura* eliminates *D. serrata*. Ayala performed his experiments at 23.5°C, at which temperature the two species coexisted. Populations were started with 300 adult flies of each species and were maintained in a series of milk bottles with standard *Drosophila* medium (a kind of yeasty porridge). Adult flies feed off the yeasty surface of the medium, female flies lay their eggs onto the surface, and larvae feed within it. Third instar larvae pupate on the surface of the medium or on the side of the bottle. When an apparent equilibrium was reached in the mixed populations, single-species populations were established and maintained under the same conditions. From the population data, and assuming that the *Lotka–Volterra* model of competition describes what is happening between the two species, Ayala was able to calculate the competition coefficients, $\alpha_{12}$ and $\alpha_{21}$. He did this in the same manner as Gause, by rearranging the *Lotka–Volterra* equations:

\[
\frac{dN_1}{dt} = K_1 - \frac{dN_1/N_1}{r_1N_1 - N_1} - \frac{dN_2/N_1}{r_1N_2 - N_1}
\]

\[
\frac{dN_2}{dt} = K_2 - \frac{dN_2/N_2}{r_2N_2 - N_2} - \frac{dN_1/N_2}{r_2N_1 - N_2}
\]

However, by assuming that the two-species populations had reached an equilibrium ($dN_1/dt = dN_2/dt = 0$), these equations can be reduced to

\[
\alpha_{12} = \frac{K_1 - N_2}{N_2}
\]

\[
\alpha_{21} = \frac{K_2 - N_1}{N_1}
\]

where $K_1$ and $K_2$ equal the equilibrium population densities of each species on their own and $N_1$ and $N_2$ equal the equilibrium population densities for each species in the mixed populations.

All the pairs of species that Ayala studied coexisted for many generations, however, in every case he found that

\[
\alpha_{12} > \frac{K_1}{K_2} \text{ and } \alpha_{21} > \frac{K_2}{K_1}
\]

This is not the “case for coexistence” in the *Lotka–Volterra* model of competition. Intriguingly, Ayala did not conclude that the model was incorrect, but rather that the species could coexist even if they utilized the same resources.
COMPETITION, INTERSPECIFIC

FIGURE 11 The nature of density dependence in Ayala's model.

Of course, the Lotka–Volterra model incorporates the logistic equation, which assumes (as a first approximation) that the relationship between rate of increase and density is a straight line (see Fig. 2). However, Gilpin and Justice examined the relationship between birth/death and density in Ayala’s experiments and found that for these *Drosophila* it was frequently a curve. It is not surprising therefore that an apparent conflict had occurred between theory and practice. This led Gilpin and Ayala (1973) to develop a new Lotka–Volterra model that incorporated such ‘non-linear’ relationships. Their new model of competition has four parameters, one more than the Lotka–Volterra model, and explains 99% of the variance in Ayala’s experiments. For two species, the model can be written as

\[
\frac{dN_1}{dt} = r_1N_1 \left( 1 - \frac{N_1}{K_1} \right) - \left( \frac{N_1}{K_1} \right) \frac{N_2}{K_2}
\]

\[
\frac{dN_2}{dt} = r_2N_2 \left( 1 - \frac{N_2}{K_2} \right) - \left( \frac{N_2}{K_2} \right) \frac{N_1}{K_1}
\]

The new parameter (\(\theta\)) describes the ‘non-linear’ nature of density dependence. When \(\theta = 1\), the new model is identical with the Lotka–Volterra model. Figure 11 shows the nature of density dependence when \(\theta \neq 1\), \(< 1\), and \(> 1\). These modified Lotka–Volterra equations give zero isoclines, on an \(N_1\)-by-\(N_2\) graph, that are curves rather than straight lines. This leads to the intriguing possibility that more than one equilibrium may exist for a pair of competing species (Fig. 12).

IV. FIELD EXPERIMENTS

A. Barnacles

Barnacles compete for space in the inter-tidal zone of rocky shores. Because they are attached to the rock, they are ideal animals for experimental manipulation in the field. In the 1950s, Joseph Connell conducted a classic set of removal experiments to test for competitive interactions between two species of barnacle, *Chthamalus stellatus* and *Balanus balanoides*. The experiments were carried out on the Isle of Cumbrae in the Firth of Clyde, Scotland.

These two species occupy two separate horizontal zones (with a small area of overlap), with *Chthamalus* (which is more resistant to desiccation) higher up the shore. *Chthamalus* larvae settle on the shore in September/October, whereas *Balanus* settle in April/May. Connell selected eight areas for study, on different parts of the shore, and used old glass lantern slides (10.7 cm \(\times\) 8.2 cm) as quadrats on which he could mark the position of all individuals of both species of barnacle.

After the settlement of *Balanus* had stopped in early June, a census of the *Chthamalus* was made. Each quadrat was divided into two halves and from one randomly chosen half, all the *Balanus* individuals were removed. Survival of the *Chthamalus* individuals in both halves of the quadrats was then monitored for 12 months. Figure 13 shows three representative results from three of the quadrats. What is quite clear is that survival of *Chthamalus* is much better in those halves of the quadrats without *Balanus*. Direct observation showed that this was due to interference competition from *Balanus* individuals who smothered, undercut, or crushed...
Chthamalus individuals. Those Chthamalus individuals in the Balanus half of the quadrat that survived the competition were smaller than those that had not been subjected to this inter-specific competition, illustrating that competition affects fecundity as well as survival.

B. Ants and Mice

Seeds play a major role in the ecology of desert regions and the seeds of annual plants are the primary food of several distantly related taxa of specialized granivores, such as rodents, birds, ants, and beetles. In particular, rodents and ants are very similar in their utilization of seed resources.

In the desert near Tucson, Arizona, experiments demonstrated that ants and rodents take most of the seeds, harvest the same size and species, and collect them from the same microhabitat. Brown and Davidson performed a set of exclusion experiments to demonstrate competition between these two taxa. They established replicated, circular plots (36 m in diameter) in relatively level, homogeneous desert scrub, approximately 60 km northwest of Tucson. Two plots were subjected to each of the following treatments. (1) PLOTS were fenced to exclude seed-eating rodents. Those rodents present were removed by trapping. (2) Granivorous ants were removed by repeated application of insecticide to individual colonies. (3) Both of the preceding treatments were done. (4) None of the treatments were done (control experiment).

Over a period of several months, the ants were censused five times (colonies counted) and the rodents were censused twelve times (trapping). Table VII shows the "numbers" of ants and rodents in the different treatments at the end of the experimental period. In the "control" experimental plots both ants and rodents showed a decrease in "numbers" compared to their single-taxon plots. Both groups showed a negative effect, which was interpreted as competition for seeds. This suggestion is strongly supported by the fact that on plots from which both ants and rodents were removed, there was a 5.5 times increase in density of seeds. However, there was no difference in seed density between plots with ants + rodents, ants only, and rodents only.

C. Anolis Lizards

Anolis lizards are an important component of island terrestrial communities in the eastern Caribbean. In these communities, anoles substitute for the ground-feeding insectivorous bird guild and are major components of the animal biomass. Joan Roughgarden and her colleagues have demonstrated the presence of com-

<table>
<thead>
<tr>
<th>Rodents removed</th>
<th>Ants removed</th>
<th>Control</th>
<th>% increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ant colonies</td>
<td>543</td>
<td>318</td>
<td>70.8</td>
</tr>
<tr>
<td>Rodent numbers</td>
<td>—</td>
<td>144</td>
<td>18.0</td>
</tr>
</tbody>
</table>

petition between two species, Anolis gingivinus and A. wattsi. Both species are insectivorous, territorial as adults, and somewhat arboreal in habit.

On the island of St. Maarten, distributional evidence suggests that there is present-day competition. Anolis wattsi occurs only in the central hills of St. Maarten, while A. gingivinus occurs throughout the island including the central hills. This observation is interesting because all the relatives of A. wattsi on nearby islands occur throughout all elevations and habitats, including sea-level habitats. Two other observations are suggestive of a competitive interaction between these two species. Anolis gingivinus shows a lower abundance and shifts its perch position where it co-occurs with A. wattsi, relative to when it occurs alone. However, these observational data are not conclusive evidence of competition.

Roughgarden and colleagues therefore carried out some manipulative field experiments on a small cay close to St. Maarten. This very small island was essentially a limestone platform (100 m x 400 m), 15 m above sea level, with several vegetation groves consisting of sea grape, Manchaneel, and perennial grasses. This cay had a resident population of A. gingivinus, but lacked any A. wattsi. Table VIII shows the results of four experiments carried out over a three-year period in three of the vegetation groves. In the experiments in grove SK (1979) and LU, marked individuals of A. wattsi were released and the resident A. gingivinus were left undisturbed. Survival of A. wattsi was low; most of the introduced individuals disappeared and those that were left did not establish territories in the center of the grove, only at the periphery. In the experiments in grove SK (1981) and MK, between 40% and 50% of the resident A. gingivinus were removed prior to introducing A. wattsi. In these experiments, survival was twice that in the other experiments and some surviving A. wattsi did establish territories in the center of the grove. It seems likely therefore that the restricted distribution of A. wattsi (to the central hills) observed on St. Maarten is due to competitive exclusion of A. wattsi by A. gingivinus.

V. FIELD OBSERVATIONS

A. New England Warblers

Many field studies of competition actually look for resource partitioning (niche separation) that would reduce interspecific competition, relative to intraspecific competition, and therefore promote coexistence. Furthermore, these field studies frequently do not examine directly the resources used (types of seed or types of insect) but instead use surrogate measurements (bill size or foraging area). An example of this type of study is Robert MacArthur's work on New England warblers.

Five species of warbler (myrtle, black-throated green, Blackburnian, bay-breasted, and Cape May) inhabit the spruce forests of Maine and Vermont in the United States. All of these species nest in the forests of New England and winter in the Caribbean. The spruce forests in which they live and breed are uniform, without obvious variety. Their beaks are all the same size and shape and their stomach contents are approximately the same. In his study, MacArthur recorded how long each species spent foraging in each part of a tree. A tree was divided into 10-ft zones vertically and branches were divided into three horizontal zones: (1) the bare or lichen-covered base, near the trunk, (2) the middle zone of needles, and (3) the terminal zone of new needles or buds. He produced diagrams of trees showing the zones of foraging activity for each species of warbler (Fig. 14). The resulting foraging distributions suggested a rather subtle kind of resource partitioning (assuming that foraging in different parts of the tree gives access to different sections of the insect resource being used by these birds).

In addition to these zone differences, MacArthur also suggested differences in foraging behavior between the five species: “subjectively, the black-throated green appeared nervous, the bay-breasted slow and deliberate.” He quantified this by recording the length of time be-
between “use of wings” and was able to show that the black-throated green had a shorter time interval than either the blackburnian or myrtle, which in turn had a shorter time interval than the Cape May and bay-breasted. He also recorded the direction in which each species predominantly moved while foraging. This could be vertical (up and down the tree), tangential (around the tree), or radial (from the center of the tree outward and vice versa). Again he found differences between the species.

MacArthur’s work suggests that even when two species overlap in foraging space, they examine that space differently, and therefore use a different part of the available resource. The five warbler species are clearly searching for their insect resources in different spaces and in different ways, but whether this is sufficient to promote coexistence between potential competitors is difficult to tell. In fact, we have no direct evidence that these warblers are even competing for limited resources.

B. Pacific Island Birds

A number of field studies have used patterns in spatial distribution as evidence of the effects of competition on species co-occurrence. This type of field observation is illustrated by Jared Diamond’s study (1975) of land-breeding birds on the Bismarck Islands off the coast of Papua New Guinea. Several pairs of ecologically similar species (good candidates for potentially strong competitors) showed what Diamond called checkerboard distributions, for example, the two flycatchers *Pachycephala pectoralis* and *P. melanura dahli* (Fig. 15). These two species have mutually exclusive but interdigitating distributions, with only one of the two flycatchers present on any one island (some islands have neither species).
These checkerboard distributions are consistent with the idea that two ecologically similar species are excluding each other from islands because of strong competitive interactions.

Although Diamond did not find many examples of checkerboard distributions among his island birds, he did find other evidence of a more diffuse type of competition by plotting what he called incidence functions. These are graphs in which the probability of occurrence on an island \( J = (\# \text{ of islands with the species present})/ (\text{total } \# \text{ of islands}) \) is plotted against the number of species \( S \) on that island. Because of the species–area effect, \( S \) is also a measure of island size. Figure 16 shows the incidence function for the cuckoo-dove Macropygia mackinlayi. This is an example of a species that Diamond calls a "supertramp." These species are excellent colonizers but are poor at persisting in the diverse communities of large islands. This absence of supertramps from large islands implies strongly that competition is important in structuring these island bird communities.

C. African Wild Dogs and Hyenas

Many field observations of "competition" have relied heavily on changing species numbers over time. For example, one species arrives in an area (or a resident species increases its numbers), while another species declines in abundance. An early case was that of two species of crayfish in Russian streams, first commented upon by Gause in his book that reported his early Paramecium experiments (Gause, 1934). Another well-cited example is the invasion and spread of the American gray squirrel (Sciurus carolinensis) in Great Britain, followed by the decline and contraction of the native red squirrel (S. vulgaris). Yet another example is provided...
TABLE IX
Population Estimates for the Spotted Hyena in the Serengeti

<table>
<thead>
<tr>
<th>Year</th>
<th>Population estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1967</td>
<td>2207 ± 120</td>
</tr>
<tr>
<td>1977</td>
<td>3108 ± 432</td>
</tr>
<tr>
<td>1986</td>
<td>9214 ± 828</td>
</tr>
<tr>
<td>1991a</td>
<td>9500</td>
</tr>
</tbody>
</table>

* Estimate includes hyenas of the Mara.

by the increase in abundance of the spotted hyena (*Crocuta crocuta*) in the Serengeti-Mara ecosystem of Kenya and Tanzania and the subsequent decline of the African wild dog (*Lycaon pictus*).

Spotted hyenas are the most abundant large predator in the Serengeti-Mara savanna ecosystem. Between 1969 and 1976, the hyena population increased by 50%, probably as a response to the herbivore increase following the elimination of the rinderpest virus. Some estimates for the subsequent population of hyenas in the Serengeti are shown in Table IX. Over this period the wild dog showed a decline in numbers, with all remaining dog packs finally disappearing from the Serengeti in the early 1990s (Fig. 17). During the early 1970s, part of this decline in wild dog numbers was certainly due to hunting, as wild dogs were regarded as vicious vermin to be exterminated. However, the reason for the later decline is less certain, but competition with hyenas may have been important.

Both of these predators hunt in packs and run down their prey (mainly Thomson’s gazelle, wildebeest, and zebra). Hyenas frequently follow dog packs when they go hunting from a den (during the breeding season) and steal their kill. Dogs can defend a kill, but about four dogs are required to keep off one hyena. Although the ‘hood’ is probably not a limiting resource for large carnivores in the Serengeti, food in the form of ‘kills’ may be. Particularly during the breeding season, dog packs are hunting not only for themselves but also for the dominant bitch and her pups back at the den. Losing kills to competitors such as hyenas may have a serious effect on pup survival. Because the ability of wild dogs to defend kills depends on the ratio of dogs to hyenas, this effect may have become gradually more serious over the last 30 years. Additional evidence supporting this story comes from the correlation (\( r = -0.92, P = 0.01 \)) between hyena density and dog density over a series of ecosystems in eastern and southern Africa (Fig. 18).

VI. CONCLUSIONS

The Lotka–Volterra model can theoretically predict the outcome of interspecific competition between two species. Depending on initial population size (\( N_1 \) and \( N_2 \)), carrying capacity (\( K_1 \) and \( K_2 \)), and competition coefficient (\( a_{12} \) and \( a_{21} \)), either species can be the sole survivor, one species is always the sole survivor (competitive exclusion), or the two species will coexist. What is more, coexistence is assured if \( a_{12}/K_1 > 1/K_2 \) and \( a_{21}/K_2 > 1/K_1 \), which implies that intra-specific competition must be greater than inter-specific competition for both species. These predicted outcomes are ‘local’ outcomes (within a habitat patch such as a fungal fruiting body, a freshwater pool, a coniferous wood, or an island de-
pending on the scale of the environmental grain appropriate to the ecology of the organism). Both the aggregation model of competition and resource partitioning predict “local” competitive exclusion, with global coexistence using different local mechanisms that ultimately have the same effect on competitive interactions.

There is considerable evidence, from laboratory experiments, that these models could describe what happens to some real organisms. Graphs of \(N_1\) against \(N_2\) (using \(K\) and \(a\)) predict the correct competitive outcome for *Paramecium aurelia* and *P. caudatum*, while a small amount of environmental heterogeneity allows sufficient resource partitioning for *P. bursaria* to coexist with *P. aurelia*. The Tribolium experiments show that initial numbers can influence which species will win the competitive battle, and these same experiments, plus those with *Drosophila*, show that more complex relationships can still be incorporated into the same theoretical framework.

Lastly, there is some evidence, from field experiments and field observations, that these mechanisms might operate outside the computer and laboratory. We see patterns of replacement reminiscent of competitive exclusion (Bismarck birds and Serengeti carnivores) and niche differences consistent with resource partitioning (New England warblers). We also see numerical changes, after removal or introduction, that are consistent with competitive effects of one species on another (desert ants and mice, Scottish barnacles, and Caribbean lizards). However, the field data tend to fit the patterns predicted by theory rather than providing detailed parameter values for the models. Nonetheless, the ideas encapsulated within theoretical models, such as the Lotka–Volterra model, provide a believable framework within which to interpret complex species interactions in the real world.

**See Also the Following Articles**

*CARRYING CAPACITY, CONCEPT OF* • *COEVOLUTION* • *PREDATORS, ECOLOGICAL ROLE OF* • *RESOURCE PARTITIONING* • *SPECIES COEXISTENCE*

**Bibliography**


COMPLEMENTARITY

Paul Williams
The Natural History Museum, London

I. Complementarity and Area Selection
II. Representing Biodiversity for Conservation
III. Complementarity-Based Techniques
IV. Assessing Complementarity
V. Conclusions

GLOSSARY

efficiency The degree of success in reaching a conservation goal relative to the cost (or to some surrogate for cost, such as the number of areas). For goals such as maximizing the number of species represented for a particular cost, a distinction should be made between merely selecting areas on the basis of any records of species and choosing areas that maximize persistence of species in the long term.

flexibility The degree to which alternatives exist for one or more selected areas in the context of reaching a particular conservation goal. When seeking a set of areas to represent maximum diversity, flexibility for a selected area may be absent, incomplete (replacing the selected area while still reaching the conservation goal would require substitution of two or more areas or one or more areas of greater cost), or complete (other areas could be substituted, one-for-one by number or by cost, with the current choice).

goals An explicit and precise statement of conservation aims. Goals should express the values of those people who provide the mandate for conservation. Making the goal explicit allows efficiency to be measured as an aid to accountability. A statement should include which attributes are valued (such as genetic diversity), which surrogates for this value are actually surveyed (such as higher taxa, species, and threatened species), which areas are to be considered (such as land management units or grid cells), and how constraints of viability, threat, and cost are to be measured and accommodated. Goals are not universal but depend on people's values and their situations. Consequently, different goals may conflict, and areas necessary to meet one goal may be insufficient to meet broader goals.

irreplaceability A property of areas that include species (or other valued attributes) restricted to so few areas that all such areas would be needed in order to meet a conservation goal.

priority Areas that need the most urgent management action to avert threat in order to meet a conservation goal (such as to increase the probability of persistence of valued species or other attributes). Ranking of areas by priority may differ from their ranking by value.

representation The occurrence of species (or other attributes) within a set of selected areas. A distinction must be made between records of a species and areas with high probability of persistence for the species in the long term.

values Values are interpreted here in the broadest sense to include monetary and nonmonetary values. Biodiversity attracts many different values, which are not universal and which depend on social and economic situations. Consequently, different values may conflict.
COMPLEMENTARITY is a property of sets of objects that exists when at least some of the objects in one set differ from the objects in another set. This definition may be extended to include parts of a pattern, such as may be used to predict the distribution of the individual objects. It can be applied to genes, species, species assemblages, or any other attributes of the biota of areas, but it is not limited to these cases. Complementarity is closely related to the idea of spatial turnover (beta and gamma diversity) in the composition of biotas, and it arises because of differences in ecology and history among populations of organisms. For example, two biotas of the same size from areas with grassland and forest may be expected to have a higher proportion of different and complementary species (and therefore will represent more species in total) than would two biotas of the same size from different areas within the same kind of grassland. This article describes complementarity primarily in the context of its application to area-selection problems for biodiversity conservation, describing how complementarity can be used to ensure that as much valued biodiversity as possible is represented for conservation within constraints such as viability and cost. It reviews some of the techniques available and how other needs could be integrated.

I. COMPLEMENTARITY AND AREA SELECTION

A simple example of complementarity is illustrated in Fig. 1, which represents the mammal faunas of two areas. These faunas share one species, the lion, but each fauna also has species that are not shared. It is the species in one fauna that are not shared with the other fauna that are said to be "complementary" with respect to the other fauna.

A. Biological Basis of Complementarity

The biological causes of complementarity are the processes that lead to differences among biotas. At the species level, these processes are familiar from differences in ecology and history. In ecology, differences among species in their needs and responses to the environment, including responses to other species, lead to differences in species distributions near equilibrium. Unique events in history, particularly the gain or loss of barriers preventing any approach to large-scale ecological equilibrium, also contribute to these distributions, as studied in vicariance biogeography (see Section 1,E).

The description of patterns of complementarity at the species level is closely related to the description of spatial turnover, also known as beta or gamma diversity depending on the spatial scale being studied. There is a substantial literature on spatial turnover, documenting and exploring patterns of species differences, particularly along transects of environmental gradients.

Patterns of complementarity in biodiversity may be expected to be highly dependent on spatial scale because the size of area patches used to study patterns in the distribution of species (for example) has a strong effect on measures of their co-occupancy of these patches. For example, it has been shown that patterns in species richness may differ strongly across the same region when measured using different patch sizes (such as with different grid-cell sizes). Because complementarity also depends on the pattern of species co-occupancy of area patches, the pattern of complementarity may be strongly dependent on the spatial scale at which it is measured.

B. Need for Complementarity in Area Selection

Complementarity is an apparently simple idea that has considerable power to aid area-selection methods for
conserving as much biodiversity as possible. Although decisions on area selection will always benefit from experienced judgment, quantitative methods provide a framework of rules designed to make such complex decisions more rigorous and efficient and more open to public accountability. Other properties of these methods include their ability to identify flexibility for the planning process; speed to facilitate the exploration of alternative values, goals, data, and flexible solutions; their ability to deal with incomplete data; and simplicity to aid communication.

Efficiency is important because the area of land (or volume of sea and air) available for conservation is limited since there is often competition between conservation and other, incompatible land uses. However, efficiency should be considered in relation to the goal of ensuring effective long-term persistence of biodiversity. Efficiency comes from using complementarity to choose areas that encompass the most biodiversity within these constraints.

Accountability is important because conservationists should be acting in the broader interests of society, and therefore people should be able to see that their values are being acted on and that limited resources are not being wasted. This could become increasingly important as competition for limited land and resources becomes more intense and as conservationists are obliged to defend their choice of areas. Accountability comes from using quantitative methods in which rules are applied in a consistent manner, forcing people to be explicit about how they might resolve the often conflicting needs of large numbers of species. This makes it possible to trace precisely why one area is chosen in preference to another and, if necessary, to challenge the data or rules.

Complementarity may be particularly helpful in considering options when different land uses conflict because it can identify the unique and valued contribution of the biota in each area. However, the potential of complementarity-based methods will only be fully realized, as with all other area-selection methods, when important constraints such as viability and threat can be assessed realistically.

C. History of Complementarity in Area Selection

Complementarity-based methods were first described by Kirkpatrick (1983). The term "complementarity" was coined specifically in the context of area selection by Vane-Wright et al. (1991). The importance of the term is that it identifies a property essential to a group of optimizing techniques (and to more approximate heuristics), which has been useful in comparative studies with other approaches to area selection.

The impetus for the development of complementarity methods in conservation has been strengthened in the past 10 years by Agenda 21 and by the Biodiversity Convention. These policy initiatives gave a new emphasis to conserving the value of broader biodiversity in addition to the long-recognized need to conserve the relatively few, particularly threatened or vulnerable species. Conserving broader biodiversity presented several new challenges, particularly in keeping track of the very large numbers of components of which biodiversity is comprised.

Complementarity-based methods have been used more in the Southern Hemisphere (particularly in Australia and South Africa), where knowledge of the biota often derives from new surveys, some designed specifically for identifying potential conservation areas. In contrast, complementarity methods have been less popular in countries with a long natural history tradition and with a strong local knowledge base, where many locally important areas may already be well-known (although rigorous prioritizing among them or adding to them efficiently may still be difficult without using complementarity). Nonetheless, there is a growing interest in complementarity methods by the U.S. Gap Analysis Program, The Nature Conservancy, Conservation International, and biological recording schemes in Europe. One of the advantages of heuristic techniques for seeking complementary areas is that they are very simple so that they can be implemented easily, which is making them increasingly popular in biodiversity conservation research.

D. Use of Complementarity in Area Selection

Representing as much biodiversity as possible for conservation, given limited resources, essentially depends on the principle of complementarity. To simplify discussion, this article concentrates on the most popular surrogates for biodiversity value—the different species (rather than on genes or ecosystems).

For example, suppose we could choose only two areas from Fig. 2, and we had to do it to represent as much diversity as possible. It might be tempting to begin by choosing area 1 because it has the largest number of species (five). The next richest area is area 2. Areas 1 and 2 represent a total of seven species (a–g).
principle has been questioned for many years, and empirical studies show substantial variation.

Patterns of complementarity might be expected to differ among groups of organisms because the ecological and historical processes that have shaped their distributions differ. In cases in which these processes have been more similar, then a closer correspondence in patterns of complementarity may be expected. For example, some groups share closer ecological similarities in diet, habitat, and climatic tolerances. If some of them have diversified at the same time and in the same areas, they may also have shared similar phylogenetic and biogeographic patterns, as envisaged by vicariance biogeography. When seeking groups of species to use as indicators for use in area selection to represent other groups or broader biodiversity, shared patterns of complementarity within ecological or phylogenetic patterns are needed (Williams et al., 2000).

II. REPRESENTING BIODIVERSITY FOR CONSERVATION

A. Identifying Goals

A common goal of biodiversity conservation projects is to ensure the persistence of as much biodiversity as possible for the future. The basic form of the problem is one of optimizing the number of valued components of biodiversity (usually species) represented within a set of areas (whether these are to be protected or not), subject to constraints including viability, threat, and cost. These constraints have often been considered at a subsequent stage, in which the component areas of a proposed protected area network are refined, usually by moving to finer spatial scales. Ideally, however, all constraints could be included in the overall design of a network, if the candidate areas were realistic land management units and if the data were available. In principle, this could include appropriate treatments of ecological processes and climate change.

A simple representation goal used to illustrate the basic principles in many academic studies has been to achieve at least one representation of every included species within the reserve network. Alternatively, the goal could be to achieve any required number of representations (this number could also differ among species) or could be expressed in terms of population size, probability of occurrence, etc. For many studies to date, information on constraints has been less widely quantified (although perhaps because less effort was invested).

E. Complementarity and Indicators of Biodiversity

Data are rarely available for all the valued parts of biodiversity. One attempt to resolve this problem is to use data for one group as a surrogate or indicator for another or for all biodiversity. This relies on congruence in distributions between the indicator and indicated groups, not just in patterns of richness but also in patterns of complementarity. Whether congruence in complementarity between groups is realistic as a general
Viability and threat are often either ignored or treated only by using crude rules (see Section II,E). The cost constraint is usually represented by the number of area units required or sometimes by the area extent (see Section II,D).

One way of dealing with threat has been to include only the most threatened species for area selection (see Section III,D). This may amount to “fire fighting” (the reactive approach) as and when species become threatened, which may risk concentrating resources in the “wrong” place in the long term if threat changes unpredictably. The alternative proactive or strategic approach is to include all species for representation, which requires investing resources in at least surveying all the species, including those that are not currently threatened. The proactive approach does not assume that simple protection of one area is an adequate or appropriate action for all species. Representation goals may still differ among species and should reflect their different needs, although this would require a rigorous justification to make the process defensible.

An important part of specifying an area-selection goal may be to include the contribution of an existing conservation network. If more species can be considered to be adequately protected by the existing network, then only the complementary species need be targeted for additional representation. This is the idea behind “gap analysis,” which can be implemented using complementarity-based procedures.

### B. Minimum-Cost Sets

One kind of question is of the following form: “Which is the minimum set of areas within Madagascar required to represent the diversity of butterflies?”—to achieve at least one representation (or any other required number) for species within the area set. This “minimum set” form of the optimization problem has been stated formally as an integer-programming problem by Camm et al. (1996):

$$\begin{align*}
\text{min} & \quad \sum_{j \in J} x_j \\
\text{subject to} & \quad \sum_{i \in N_j} x_j \geq 1, \text{ for all } i \in I \\
& \quad x_j = (0, 1), \text{ for all } j \in J
\end{align*}$$

where $J = \{j \mid 1 \ldots n\}$ is the set of candidate areas for selection, and $I = \{i \mid 1 \ldots m\}$ is the set of species to be represented. $N_j$ is the subset of areas $j$ that contain species $i$. The selection variable $x_j = 1$ if the area $j$ is selected or 0 if it is not selected. The first constraint is that every species must be represented in at least one of the areas (or more areas for multiple representation goals), with one such constraint for each of the $m$ species. The second constraint states that selection is all or nothing for each of the $n$ areas.

Minimum sets aim to achieve complete representation, usually in the first instance by seeking representation of every species in at least one area (Fig. 3) or in more areas for multiple representation goals. There are many pitfalls with this approach, not only concerning the problems of viability and threat that are discussed later but also concerning the meaning of “complete” in this context. It is apparent that a single representation of a species will not necessarily represent all of the intraspecific variation. The problem becomes worse when more remote surrogates for biodiversity value are used. For example, when using ecoregion or community data, such as vegetation classifications, a single representation of every vegetation class would be very unlikely to represent every species, let alone all intraspecific variation. Of course, the only reliable solution to representing every difference is to include every area, although the premise of using area-selection methods is that competition with incompatible land uses limits the area available for conservation. Nonetheless, minimum sets for species at least have the advantage that they are expected to represent more species than any undirected selection for the same number of areas or cost.

### C. Maximum-Coverage Sets

Another kind of question takes the following form: “How can we choose 1% of the total area of Madagascar to represent the greatest diversity of butterflies?” This “maximum-coverage” form of the optimization problem has been stated formally as an integer-programming problem by Camm et al. (1996):

$$\begin{align*}
\text{max} & \quad \sum_{i \in I} y_i \\
\text{subject to} & \quad \sum_{i \in N_j} y_i \geq 1, \text{ for all } i \in I \\
& \quad y_i = (0, 1), \text{ for all } i \in I
\end{align*}$$

where $I$, $J$, $N_i$, and $x_j$ are defined as for the minimum set problem. In addition, constraint (2) is the number
of areas to be selected, \( k \). Because not all species will be represented, the constraints are modified, with \( y_i = 1 \) if species \( i \) is represented in at least one selected area or 0 if it is not represented.

Maximum-coverage problems (Fig. 4) may often prove to be the more realistic and widely applicable form of the problem because conservation goals are usually constrained more by budget and competing land uses than by ideas of complete representation. It is therefore unavoidable to have to face the "agony of choice" that not all valued biodiversity can be repre-
sented in areas selected for conservation (Vane-Wright et al., 1991).

D. The Cost Constraint

In the simplest formulation of conservation goals, the cost constraint on area selection can be represented by the number of areas, treating all areas as being of equal cost. A few studies have used the extent of areas of differing size as a measure of cost. Even when financial cost data have been used, they have rarely included a thorough consideration of all the appropriate cost components. These may include (but are not limited to) cost of area acquisition, costs of management to reduce threats, and (crucially) the opportunity costs that arise from income from any alternative uses of an area (e.g., forestry and agriculture) that is foregone when it is managed for biodiversity conservation (Faith and Walker, 1996).

When information on how cost varies among areas is available, it can be included in area selection by dividing the complementary part of the biodiversity value of areas by the cost of those areas to calculate the cost of acquiring a particular biodiversity increment.

The consequence of including cost data that differentiate among areas when seeking a minimum representative set is usually to decrease the cost, although the number of areas or the area required usually increase. For goals of finding a maximum-coverage set for a given cost, the number of areas and the amount of biodiversity represented usually both increase. If, however, sets are sought for increasing numbers of areas so as to maximize the numbers of species at each step, then the minimum cost may fluctuate strongly as the number of areas increases (because it may be cheaper to acquire more areas for representing the same number of species).

E. Other Constraints

Ultimately, effective biodiversity conservation is about ensuring the persistence of biodiversity values. This requires that appropriate treatment of viability and threat be integrated into area-selection methods. With branch-and-bound techniques (see Section III,B), multiple constraints can be included in the specification of the problem. With heuristics (see Section III,C), a variety of options for integration have been used. Those that are structured around a sequence of decisions (as opposed to combining or weighting scores for the different constraints) can have the advantage of better accountability because accounting for why an area is included or excluded at each stage is straightforward (Williams, 1998). For example, constraints may be imposed by applying a sequence of filters to the data (Bedward et al., 1992). To deal with viability, the simplest treatment is to include only those populations likely to be viable, although the necessary data to make this determination are rarely available. To deal with threat, areas in which species are known to be under severe threat might be excluded from consideration because they are deemed to be irretrievable, or the cost of averting threats might be included in the area cost.

III. COMPLEMENTARITY-BASED TECHNIQUES

In addition to the biological and social problems of accounting for viability and threat, area-selection methods have to deal with a technical problem of how to represent the most biodiversity (as species or any other valued aspects of biodiversity) with limited resources. Achieving the representation goals within the constraints described in the previous section is not entirely straightforward. For example, the size of the problem when seeking a minimum-cost set of areas cannot be calculated simply from the amount of data (this problem is described as not polynomial complete or "n-p complete"). Fortunately, the problem is familiar to mathematicians from many kinds of similar applications in operations research, in which a range of techniques for dealing with it have gained acceptance (exhaustive search, branch-and-bound, and heuristics).

The size of the area-selection problem may often be reduced before applying any of the techniques described later by simply removing redundant data. For example, if two areas have exactly the same species, then one of the areas can be removed from the data, providing a record is made that the data for the retained area represent either of the original areas. Similarly, if two species have the same distribution, then one species may be removed. This may help to speed up the procedure, but it is only applicable if the viability and threat treatment for the retained data and the removed data are identical. Similarly, widespread generalist species may not need protection to persist and could be excluded from an analysis.

A. Exhaustive Search

The procedure for exhaustive search is to test every possible combination of areas to find the optimum.
The advantage is that this provides an exact solution, guaranteeing that the one or more truly optimal results will be found. The disadvantage is that it is only applicable to problems with so few areas that it may be of little practical use. This is because the number of possible combinations to be tried becomes too large to be searched exhaustively, even with very powerful computers. For example, for a simple maximum-coverage problem, when seeking five areas to represent the most species from among 1000 possible candidate areas, there are $1000 \times 999 \times 998 \times 997 \times 996$, or nearly $10^{15}$ possible combinations of areas to be tested.

## B. Branch-and-Bound Algorithms

The procedure for branch-and-bound algorithms is to restrict the search to only a subset of all possible combinations of areas, but a subset within which the optimal solution is expected to lie. There are two advantages to branch-and-bound algorithms. First, they can give an exact solution with the truly optimal result. Second, because only a subset of area combinations is searched, this is faster than an exhaustive search and therefore may be applied to larger problems. There are two disadvantages. First, even searching a subset may require very large numbers of combinations to be searched, thus taking a long time. Advances in computer technology have reduced this to days or even hours for large problems, although this may still be too long when the decision-making process demands interactive analysis. Second, when seeking more complex representation goals, for example, involving quantitative area targets, the problem may not be solvable by this approach (Pressey et al., 1996). Branch-and-bound algorithms are also more complicated to implement, although commercial software packages are available.

The branch-and-bound technique begins by setting "bounds" between which the optimal solution must lie. This may be done by solving a simpler form of the problem (a linear-programming relaxation of the integer-programming, or presence-absence, problem), for which each species may be represented a fractional number of times. The technique is then to search "branches," or groups of possible area sets. Any area sets that exceed the bounds can be rejected, along with all of the possible but untested area sets remaining in the branch, thereby greatly reducing the size of the problem. The bounds are usually refined as the search proceeds.

## C. Heuristic Algorithms

The procedure for heuristic algorithms is to use relatively simple sets of rules that have been found, from experience with previous data sets, to give good approximations to optimal solutions. There are three advantages to heuristic algorithms. First, although some heuristics are known to be poor, others give results that are usually close to the optimum (Cusati et al., 1997) (hence, heuristics give "near-optimal" solutions) and distinctly better than those obtained from simple scoring techniques that do not use complementarity (Pressey and Nicholls, 1989). In practice, the suboptimality may be trivial compared to concerns about data quality and practical implementation for conservation (Pressey et al., 1996; see Section IV,A). Second, heuristics are by far the fastest complementarity-based techniques available, which makes them suitable for an interactive decision-making process. Third, because they consist of simple rules, they are easy for nonspecialists to implement, with or without a need for computers. There are two disadvantages. First, it is not guaranteed that they will find the optimal solution, only approximations to it. Second, because their relative efficiency depends not only on the choice of rules but also on the particular pattern of species distributions within each data set, it may be worthwhile to apply more than one set of heuristic rules in order to compare the results and select the best (this may still be much faster than a truly optimizing algorithm).

The simplest heuristic procedure is the greedy richness algorithm, which selects the area with highest richness in complementary species (i.e., richest in those not represented at previous steps) at each step or "iteration" of this selection rule until the selection goal is reached. This tends to work well for near-maximum coverage when choosing relatively few areas compared to the minimum set, but it is less efficient as the number of areas approaches the size of the minimum set (Cusati et al., 1997). Area choices for a minimum set are usually most constrained by the more restricted species; therefore, heuristic near-minimum set algorithms that focus on "rarity" (in the sense of the species frequency of occurrence among areas that are candidates for selection) tend to be more efficient. Two kinds of rarity-based algorithms have been used (see Section IV,C). First, weighted rarity algorithms (Kirkpatrick, 1983) are similar to the greedy richness algorithm but weight each species according to its frequency in order to favor choosing areas with many of the more restricted species. Second, progressive rarity algorithms (Margules et al., 1988) identify the rarest complementary species at each step and score areas for richness in only these species, moving progressively to the next less restricted species only after the most restricted species are all represented within selected areas (Fig. 5). Additional rules may be added to deal with ties, and which rule is used to make
each area choice can be recorded to aid accountability (Table I). This tends to work better for near-maximum coverage when choosing a number of areas that approaches the minimum set, but it is less efficient for very small numbers of areas (Fig. 6a). Which of these heuristic algorithms is most efficient depends on patterns in the data (see Section IV,A).

Heuristic algorithms may be supplemented with additional procedures designed to improve their efficiency. The most basic of these are redundancy tests, which are applied to area sets in order to identify and reject any areas (chosen at early steps of the algorithm) in which all of the species are found subsequently to be represented in other selected areas (chosen at later steps of the algorithm). In extended redundancy tests, the unique contributions of each selected area to the representation goal are identified (see Section IV,B). In some cases, it is then possible to substitute a single area for pairs of areas while still meeting the representation goal. Re-ordering a near-minimum area set of selected areas by complementary richness may be used as a heuristic technique to provide a series of approximations to maximum-coverage solutions (Table II). Again, the question of which heuristic provides the most efficient result may depend on patterns in the data and on how many areas are needed in comparison with the size of the minimum set. Therefore, if after reordering of the areas the results of cumulative species representation from the different algorithms are compared, then the more efficient maximum-coverage solution for a given number of areas can be chosen. These different area sets must not be mixed subsequently because maxi-
TABLE II
Near-Minimum Area Set: Reordered by Complementary Richness (No Within-Set Redundant Areas Detected): Summary of Species Representation by Area Sequence

<table>
<thead>
<tr>
<th>Choices</th>
<th>Species richness</th>
<th>Absolute</th>
<th>Increment</th>
<th>Cumulative</th>
<th>% Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step</td>
<td>Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Ecuador</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>23.26</td>
</tr>
<tr>
<td>2</td>
<td>Kashmir</td>
<td>9</td>
<td>9</td>
<td>19</td>
<td>44.19</td>
</tr>
<tr>
<td>3</td>
<td>Turkey</td>
<td>7</td>
<td>6</td>
<td>25</td>
<td>58.14</td>
</tr>
<tr>
<td>4</td>
<td>Michoacan</td>
<td>4</td>
<td>4</td>
<td>29</td>
<td>67.44</td>
</tr>
<tr>
<td>5</td>
<td>Central Bolivia</td>
<td>8</td>
<td>3</td>
<td>32</td>
<td>74.42</td>
</tr>
<tr>
<td>6</td>
<td>North California</td>
<td>4</td>
<td>3</td>
<td>35</td>
<td>81.40</td>
</tr>
<tr>
<td>7</td>
<td>Irkutsk</td>
<td>5</td>
<td>3</td>
<td>38</td>
<td>88.37</td>
</tr>
<tr>
<td>8</td>
<td>Afghanistan</td>
<td>5</td>
<td>1</td>
<td>39</td>
<td>90.70</td>
</tr>
<tr>
<td>9</td>
<td>Qinghai</td>
<td>8</td>
<td>1</td>
<td>40</td>
<td>93.02</td>
</tr>
<tr>
<td>10</td>
<td>Northeast India</td>
<td>3</td>
<td>1</td>
<td>41</td>
<td>95.35</td>
</tr>
<tr>
<td>11</td>
<td>Uzbelistani</td>
<td>4</td>
<td>1</td>
<td>42</td>
<td>97.67</td>
</tr>
<tr>
<td>12</td>
<td>Big Horn</td>
<td>4</td>
<td>1</td>
<td>43</td>
<td>100.00</td>
</tr>
</tbody>
</table>

*This is an example of the results of reordering the near-minimum set shown in Table 1 that represents all 43 species of bumble bees of the sibrica group among large equal-area (611,000 km²) grid cells. The areas are listed by complementary species richness; the order in which they contribute the largest number of previously unrepresented species. The complementary richness at each step or species increments are shown in the fourth column.

D. Prioritizing Areas

After selection of areas to meet species representation goals, ordering of areas (or reordering when using heuristic) may be used to rank the selected areas by priority. Priority is the urgency with which areas require management action for conservation to avert threats to the persistence of their biodiversity value. Therefore, priority ranking of areas may differ from their ranking by biodiversity value. The distinction between value and priority is important but often ignored.

The choice of criterion for assigning priority may depend on the particular situation. First, priority has been based on measures of threat, but only when threat is predictable and when suitable measures are available.

FIGURE 6 Species representation and relative efficiency. (a) Cumulative percentage of a group of 426 species of Oregon vertebrates represented within increasing numbers of hexagonal grid cells (each 635 km², as developed for the Oregon Gap Analysis Program) selected using maximum coverage (branch-and-bound algorithm), near-maximum coverage (progressive rarity algorithm), hot spots of richness, and hot spots of rarity (richness in the rare quartile of species), and areas chosen at random. Scores above the solid gray line (the percentage threshold score to the top 5% within 1000 randomly drawn scores) are significantly better than expected when choosing areas at random (data reanalyzed from Cauti et al., 1997). (b) Cumulative percentage of a group of 321 species of Madagascan butterflies represented within increasing numbers of quarter-degree grid cells (each approximately 170 km²) selected using near-maximum coverage (progressive rarity algorithm), hot spots of richness, and hot spots of rarity (richness in the rare quartile of species) and among areas chosen at random (data courtesy of David Lees). (c) Results from Fig. 6a recalced for relative efficiency between the maximum that can be achieved (using the branch-and-bound algorithm) and the median representation expected from selecting areas at random (relative representation efficiency: \( E_{\text{relative, max}} = 100 \times \frac{S_{\text{max}} - S_{\text{random, max}}}{S_{\text{max}} - S_{\text{random, min}}} \)). Scores above the solid gray line are significantly better than expected when choosing areas at random. (d) Results from Fig. 6b recalced for relative efficiency between an approximate estimate of the maximum coverage that can be achieved (using a progressive rarity algorithm) and the median representation expected from selecting areas at random. Because the heuristic algorithm is likely to be further from the optimum when very small numbers of areas are selected, the corresponding estimates of relative efficiency may be less reliable.
in order to ensure that the most threatened biodiversity is protected as soon as possible. Second, priority has been based on flexibility to ensure that the parts of biodiversity with fewest options for representation are represented as soon as possible (see Section IV.C). A potential (although unlikely) problem with this option is that unpredictable threats might conceivably still destroy the bulk of biodiversity in areas low on the list, despite their greater flexibility, before it can be protected. Third, when no other information is available, or when threats are intense and unpredictable, a precautionary approach has been to prioritize areas by reordering them in an approximation to a series of maximum-coverage solutions to ensure that the most biodiversity is protected as soon as possible.

IV. ASSESSING COMPLEMENTARITY

A. Efficiency

Much of the emphasis on using complementarity in area selection is a result of the fact that it can be used to represent more biodiversity with limited resources. In general, the more constraints placed upon the optimization problem, the greater the cost. The relationship between representation and cost has been identified with concepts of efficiency. However, efficiency should be considered in relation to not just representation of species records but also in the context of all the other constraints (viability, threat, etc.) necessary to reach the goal of ensuring the effective long-term persistence of biodiversity.

For minimum-set goals, the efficiency of species representation was measured by Pressey and Nicholls (1989) as the proportion of all areas that is not required to meet the goal:

\[ E_{\text{min}} = 1 - \left( \frac{X_{\text{min}}}{T} \right) \]

where \( E_{\text{min}} \) is the efficiency, \( X_{\text{min}} \) is the number or extent of areas needed for the minimum set, and \( T \) is the total number or extent of candidate areas for selection. This is an absolute measure in the sense that efficiency is judged relative to selecting all areas or none.

Because the absolute efficiency that can be achieved in any particular case varies with patterns in the data, it may be more useful to measure efficiency relative to what might actually be expected with those particular data. The upper bound to this is the maximum representation within the chosen constraints, the maximum-coverage set \( (s_{\text{max}}, \) for the given constraints \( c \). One useful lower bound for comparison is the representation that might be expected if areas were selected at random, without regard to their biota. In the past, efficiency of species representation has often been assessed informally by comparing species-accumulation curves for ordered sequences of selected areas. Some studies have also used simulated random draws of areas (without replacement) to estimate whether the representation achieved by an area set is significantly greater than would be expected by chance when selecting the same number of areas (Figs. 6a and 6c). This kind of assessment might be refined in many ways, particularly for dealing with existing networks of protected areas. One lower bound of interest might be the median score expected by random draws \( \left( s_{\text{rand}, \) for the same constraints, although other lower bounds could be substituted when appropriate. Therefore, for the representation within an observed set of areas \( (s_{\text{obs}, \) for the same constraints, the relative efficiency of any set of areas \( (E_{\text{obs}, \) for constraints \( c \) is given by

\[ E_{\text{obs}} = \left( s_{\text{obs}, \text{rand}} / s_{\text{max}} - s_{\text{rand}} \right) \]

In the simplest form of this measure, the number of areas can be used as a measure of the cost constraint, in which case efficiency is measured relative to what might be expected from the data when selecting a particular number of areas (Figs. 6b and 6d).

Higher efficiency in species representation was the original reason why Kirkpatrick (1983) used a complementarity-based technique for area selection rather than simple scoring of areas. This higher efficiency was demonstrated using the absolute measure of efficiency by Pressey and Nicholls (1989). A comparison of complementary areas with hot spots of richness and rarity (Table III) shows that using complementarity results in increased species representation and does so particularly for the most restricted species.

The maximum absolute efficiency in species representation that can be achieved using complementarity is determined by patterns in the data of co-occupancy of areas by species. For example, if all species were found in one area, then only one area would be needed as a minimum to represent them; however, if none of the species were found together, then as many areas as species would be needed (Vane-Wright et al., 1991). In a recent study, absolute efficiency was found to be increased by at least three factors (Pressey et al., 1999). First, efficiency increased with greater "nestedness" among species distributions (perfect nestedness is the pattern in which species present in any particular biota are present in all larger biotas, and where species absent
from any particular biota are absent from all smaller biotas). Second, efficiency increased when species were more broadly distributed. Third, efficiency increased with greater variation in the size (or cost) of areas.

B. Flexibility

Quantitative area-selection techniques frequently identify several alternative combinations of areas that meet their representation goals. This allows for flexibility in the final area choices (Pressey et al., 1993). In general, the more constraints placed on the optimization problem, the less the flexibility. In contrast, areas for which there are no flexible alternatives are "irreplaceable" if the representation goal is to be met.

When using exhaustive or branch-and-bound area-selection techniques, flexibility is identified directly as alternative optimal sets. When using heuristic techniques, however, it is done indirectly by identifying which species each selected area contributes uniquely to the representative set (Figs. 3 and 7). Any other area that shares these same "goal-essential" species (Rebelo, 1994) could be substituted in its place (Figs. 7 and 8a) and is therefore a fully flexible alternative (Fig. 8b). Flexibility can also be used to order selected areas by the number of options for representing their goal-essential complement, as shown by the color scale from black to light gray used for selected areas in Fig. 8b.

For truly optimal results, all constraints that can reasonably be quantified should be included when formulating the goal of the original area-selection problem. However, flexibility may be useful for exploring constraints that were not easily quantified or included as part of earlier steps in the area-selection procedure. Decision makers may find interactive exploration of flexibility particularly useful when asking "what if...?" questions. Interactive decision making is popular in other fields because it can use the graphical strengths of geographic information system technology to communicate large amounts of information quickly and easily.

C. Irreplaceability

Areas within a representative set for which there are no flexible alternatives (in order to meet a particular goal) have been described as "irreplaceable," an all or nothing property. However, "shades of irreplaceability" has also been used to refer to the relative number of alternatives that may exist for an area choice, depending on how widespread the species occurring there may be (Pressey et al., 1994).

Shades of irreplaceability concerns the frequency with which an area is required among all possible fully

<table>
<thead>
<tr>
<th>Species</th>
<th>Goal-essential species (minimum representation within area)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a b c d e f g h</td>
<td>2 4 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Area</th>
<th>Species</th>
<th>Goal-essential species (unique complement within minimum representative set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X X X X X X X</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X X X X X . .</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>. . . . X X X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X X X . . . .</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>. . . . X X X</td>
<td>Flexible area</td>
</tr>
</tbody>
</table>

FIGURE 7 Flexibility. Five areas have a biota of eight species. The minimum fully representative set for all species is a combination of areas 2 and 3. Area 2 uniquely contributes four species (a–d) to the goal of a single representation of each species, and area 3 uniquely contributes three (f–h) of these goal-essential species. Note that all three of the goal-essential species of area 3 (f–h) are shared by area 5, which is therefore a fully flexible alternative.
representative area sets, whatever the size or cost of the set. This depends on the rarest species in each area because the rarer a species, the fewer the alternative areas available to represent it. For this purpose, “rarity” of a species is measured from its frequency of occurrence among the areas that are candidates for selection. An approximate estimate of shades of irrereplaceability is “effective maximum rarity” (EMR), which depends on the rarest species in each area, but considering only the species that are complementary to areas with even rarer species (Pressey et al., 1994):

\[ \text{EMR} = 100/c_q \]

where \( c_q \) is the frequency of the rarest unrepresented species (i) that is present in an area (j). If no species are left unrepresented, then

\[ \text{EMR} = 100/C \]

where \( C \) is the total number of candidate areas in the data. This measure of irrereplaceability captures much of the relational property of complementarity among areas in a single score.

**FIGURE 8** Flexibility within a near-minimum set of quarter-degree grid cells selected to represent a group of 321 species of Madagascan butterflies at least once (ignoring other constraints). (a) Gray spots are the selected areas from Fig. 3. For selected area 13 in the northeast, the gray squares show richness in other cells of the four goal-essential species (Fig. 7) that justify its inclusion in the area set. Therefore, the other black square that shares all four of these goal-essential species is a fully flexible alternative (and could replace area 13), whereas the other gray squares are only partially flexible (more than one of these would be required to replace area 13). (b) The same map as in a, but showing the fully flexible areas for all selected areas. The black areas are either irreplaceable (Fig. 3) or have no fully flexible alternatives. The other gray areas are now groups of fully flexible alternatives, so only one area is required from each group. For example, the two black squares sharing all four goal-essential species from a are now shown in dark gray, as a group of fully flexible areas for selected area 13. From the number of fully flexible areas in each case, the total number of alternative near-minimum sets is estimated as \( 10 \times 2 \times 2 \times 2 \times 10 \times 15 \times 32 = 38,400 \) (see Williams et al., 2000) (data courtesy of David Lees).
Using the definitions at the beginning of this article, shades of irreplaceability is a measure of flexibility (Fig. 9) rather than a measure of the conservation value of an area (which would be measured in terms of an appropriate part of its biodiversity value). When the EMR score is applied in a stepwise manner to select areas, it is then almost identical to the progressive rarity heuristics described previously (Fig. 5). Indeed, the EMR score may suffer from similarly reduced efficiency if it is not also combined with appropriate redundancy checks, although more efficient procedures for estimating shades of irreplaceability are being sought. Once this is achieved, shades of irreplaceability will be very similar to flexibility but expressed in relation to all possible fully representative sets of areas rather than just minimum-cost sets of areas. Whichever term is used, this approach can be used as another way of prioritizing area choices against threat by their number of flexible options (see Section III.D).

D. Comparing Area Sets

Two kinds of assessments of area sets require special care. The first is the comparison of sets of complementary areas with other sets of areas. The second is counting the representation of species or area attributes other than those used to select the areas. The complication arises from flexibility among alternative complementary area sets, which may result in different degrees of overlap among area sets or in different degrees of representation of the other organisms. In both cases, the flexible areas cannot be ignored because they are being compared on the basis of attributes other than those for which they were selected, and which therefore may be distributed among areas quite differently. For example, in Fig. 10, the flexible sets of 50 areas selected for maximum coverage of all mammals and birds must by definition all have the same number of mammals and birds. However, when sets of 50 areas are selected for maximum coverage of just the flagship species, or just the large mammals, then any flexible sets may vary in the numbers of all mammals and birds represented.

When assessing representation of these other area attributes for a set of complementary areas that shows flexibility, an area set has to be treated as a population of area sets. This gives rise to a corresponding distribution of differing degrees of area coincidence or of species representation. If two complementary area sets are being compared and both show flexibility, then the full intersection of the two populations of sets has to be considered.

V. CONCLUSIONS

Complementarity allows the maximum amount of biodiversity value to be represented for conservation, sub-
FIGURE 10 Flexibility and representation of mammals and breeding bird species. Histograms of representation of all mammals and breeding birds within flexible sets of 50 one-degree grid cells (each approximately 11,025 km$^2$) in sub-Saharan Africa, each selected for maximum coverage of three different groups of species: upper, six flagship mammals; middle, five orders of larger mammals; and lower, all mammals and breeding birds. With the exception of the histogram at the bottom, which shows all flexible sets of 50 cells, scores are from a randomly drawn sample of 1,000 alternative sets from an estimated $2.12 \times 10^6$ flexible sets of 50 cells for flagship species and an estimated $6.54 \times 10^6$ flexible sets of 50 cells for larger mammals. Representation scores to the left of the dashed line are within the range expected when choosing 50 grid cells at random (adapted from Williams et al., 2000).

See Also the Following Articles
CONSERVATION EFFORTS, CONTEMPORARY • SPECIES-AREA RELATIONSHIP

Bibliography


COMPLEXITY VERSUS DIVERSITY

Shahid Naeem
University of Washington

I. The Complexity of Ecologically Diverse Systems
II. Measures and Consequences of Ecological Complexity
III. The Significance of Ecological Complexity
IV. Conclusions: The Disentangled Bank

GLOSSARY

biotic interaction When the growth or behavior of one species affects those of another species; such interactions can be antagonistic (e.g., competition for limiting resources, predation, herbivory, or parasitism) or facilitory (e.g., pollination or other forms of mutualism).

community Set of coexisting species in an ecosystem.

food web Set of trophic relationships among species in a community; schematically, the pattern often resembles a web of species each connected by trophic interactions with other species.

functional dependency Dependency of one species on another to complete a particular ecosystem process or function; an example would be the dependency of plants on decomposers and decomposers on plants to recycle nutrients.

phylogeny Pattern of evolutionary descent among species, generally illustrated as a “tree” in which nodes represent common ancestors and lengths of branches represent time since divergence; time since divergence is generally assumed to correlate with genetic relatedness.

process intricacy Complexity of temporal or spatial patterns of ecological processes such as population dynamics or production.

structural intricacy Complexity of patterns of links or connections among species in a community that are created by biotic interactions, shared pathways of nutrient and energy flow, or phylogenetic relationships.

trophic interaction Feeding relationship between two species; these include predation, herbivory, parasitism, bacterivory, frugivory, or any other interaction that involves individuals of one species consuming individuals or parts of individuals from another species; trophic interactions represent a subset of all biotic interactions in a community.

ECOLOGICAL COMPLEXITY AND BIOLOGICAL DIVERSITY are often presumed to be strongly correlated or even synonymous with one another, but there is no necessary relationship between these terms. Ecological complexity is a scale by which we compare ecosystems. At one end of the scale, systems that exhibit unpredictable behavior or many connections among its species are considered “complex.” At the other end of the scale, systems that exhibit highly predictable behavior or largely independent species are considered “simple.”
contrast, biological diversity is a measure of the extent of genetic and phenotypic variation found within a community. These two characteristics of ecosystems, complexity and diversity, though fundamentally related, are clearly distinct. Although diversity and complexity are not necessarily correlated, the potential for a community to exhibit complexity is likely to be associated with its diversity. Widespread changes in diversity experienced by most ecosystems have generated considerable interest in quantifying complexity and assessing whether current changes are affecting levels of ecological complexity in natural ecosystems. The resolution of issues surrounding the relationship between complexity and diversity is an active area in contemporary ecological research.

I. THE COMPLEXITY OF ECOLOGICALLY DIVERSE SYSTEMS

Hence, if certain insectivorous birds (whose numbers are probably regulated by hawks or beasts of prey) were to increase in Paraguay, the flies would decrease—then cattle and horses would become feral, and this would certainly greatly alter (as indeed I have observed in parts of South America) the vegetation: this again would largely affect the insects; and this, as we just have seen in Staffordshire, the insectivorous birds, and so onwards in ever increasing circles of complexity.

—CHARLES DARWIN, 1859

It is reasonable to believe that in following the numbers of orders in the fossil record we are indeed following the approximate overall course of ecological complication and diversification even though in a necessarily loose way.

—G. G. SIMPSON, 1969

Roughly speaking, we here take complexity to be measured by the number and nature of the individual links in the food web.

—R. M. MAY, 1974

The complexity of a system is in the eye of the beholder. It is measured by how well we understand causes, expect behaviours and, in praxis, achieve purposes. Hence, large numbers of variables, non-linear relationships among them, and the open nature of a system are important only to the degree they present barriers to understanding.

—C. S. HOLLING, 1985

A. The Complexity of Nature and the Entangled Bank

Darwin provided us with a wonderful metaphor for what is intuitively meant by "ecological complexity" in his famous closing passage to The Origin of Species. In this passage, Darwin envisioned a riverbank teeming with a diverse and entangled array of organisms all involved in a complex web of biotic interactions. At the heart of this seemingly complex system, however, lies evolutionary and ecological processes that produce and maintain such remarkable diversity. For Darwin, biodiversity and the complexity that accompanies it are what distinguish nature from the inanimate features and processes of our world.

The entangled bank metaphor highlights two main ingredients of ecological complexity. First, there is biodiversity itself, with a clear implication that greater diversity provides a greater potential for complexity. Second, and more importantly, there is the web of interactions among species, with an equally clear implication that more complex webs (e.g., greater numbers and kinds of biotic interactions among species) are indicative of more diverse communities.

This metaphor also highlights two distinct meanings of complexity. First, complexity refers to the intricacy of the pattern of connections among species. This is the meaning used in many studies that have related ecological complexity to ecological stability. The second meaning applied to the term complexity is when nature defies easy characterization and cannot be reduced to a finite set of processes and principles that govern its order. This second meaning is used by the "macroecological" perspective (Brown, 1995), the hierarchical perspective of nature (O'Neill et al., 1986), when considering nature as complex adaptive systems (Levin et al., 1997), or when nature is viewed as a self-organizing system (Holling, 1992).

Figure 1 illustrates these different kinds of complexities. The top row of Fig. 1 shows the dynamics of an ecosystem process as an example of an ecosystem property that may exhibit simple or complex behavior. This process could be production, decomposition, or any other process representing the aggregate activities of species in the community. The remainder of Fig. 1 shows three fundamental kinds of linkages among species in communities: community structure or the biotic interactions among species, phylogenetic structure or the pattern of phylogenetic relationships among species, and functional structure or the pattern of functional dependencies among species. The pattern of tem-
### COMPLEXITY VERSUS DIVERSITY

<table>
<thead>
<tr>
<th></th>
<th>Simple</th>
<th>Complex</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dynamics</strong></td>
<td>Disorder</td>
<td>Order</td>
</tr>
<tr>
<td>Process</td>
<td><img src="image" alt="Simple Process" /></td>
<td><img src="image" alt="Complex Process" /></td>
</tr>
<tr>
<td>Time</td>
<td><img src="image" alt="Simple Time" /></td>
<td><img src="image" alt="Complex Time" /></td>
</tr>
<tr>
<td><strong>Community</strong></td>
<td><img src="image" alt="Simple Community" /></td>
<td><img src="image" alt="Complex Community" /></td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td><img src="image" alt="Simple Structure" /></td>
<td><img src="image" alt="Complex Structure" /></td>
</tr>
<tr>
<td><strong>Phylogenetic</strong></td>
<td><img src="image" alt="Simple Phylogenetic" /></td>
<td><img src="image" alt="Complex Phylogenetic" /></td>
</tr>
<tr>
<td><strong>Functional</strong></td>
<td><img src="image" alt="Simple Functional" /></td>
<td><img src="image" alt="Complex Functional" /></td>
</tr>
</tbody>
</table>

**Figure 1** Simplicity and complexity in ecological systems. Dynamics are illustrated as temporal patterns in an ecological or evolutionary process, such as production, standing biomass, energy or nutrient flow, or rates of diversification. Community structure is illustrated as the number, type (e.g., competition or facilitation, shown as − or +, respectively), strengths (thickness of arrows), and arrangement of interactions (arrows) among species (circles). Phylogenetic structure is illustrated as phylogenetic trees in which every two species share a common ancestor. The vertical lengths of branches indicate distances, assumed to be a measure of genetic similarity or time since divergence. The complex phylogenies of more distantly related species contain more genetic information than would be found in the simple phylogeny of many closely related species. Functional complexity is illustrated as species whose contributions to ecosystem processes are either largely dependent (serially linked) on the activities of other species or exhibit a diverse array of interdependencies among one another. A simple system is one in which all species contribute to the functioning of an ecosystem in the same way (parallel redundancy) such that the loss of a species from a system does not affect its functioning.
poral or spatial variation of processes is known as process intricacy. The pattern of linkages, whether community, phylogenetic, or functional, is known as structural intricacy.

Whether considering process or structural intricacy, exhibiting either highly ordered states or highly irregular or disordered states is considered complex. As shown in Fig. 1, a temporal pattern for a specific ecological process that exhibits highly ordered, predictable, repeated series of oscillations would be considered complex. Yet a highly irregular, unpredictable, chaotic pattern of oscillation would also be considered complex when compared to a simple, mildly oscillating process. Similarly, Fig. 1 shows that a pattern of biotic interactions in which every species is linked to every species in the community by exactly the same kind of interaction with exactly the same strengths of interactions creates a symmetrical, highly ordered structure that would be considered complex. At the same time, a community that has a jumbled network of biotic interactions that vary in type and strength would also be considered complex when compared to a simple set of independent, noninteracting species.

Although ecological complexity is difficult to define precisely given the many different ways the term is used, when we consider the basic elements of ecological complexity it becomes apparent that diversity is not a necessary part of the definition of complexity. Note in Fig. 1, for example, that all nine structures shown are made up of six species. In spite of having the same number of species, these structures differ quite dramatically in complexity.

Despite the fact that diversity is not a necessary part of the definition of complexity, diversity and complexity are related in an elementary yet fundamental way. Few would debate that extremely depauperate ecosystems such as monoculture plantations of bananas are less complex than the extremely species-rich tropical rain forests they replaced, but for the bulk of ecological communities that lie somewhere between these extremes of biodiversity, determining just how complex a given system is can be quite difficult. For example, if a community contained 25 species of grasses while another contained only 10 species, but was made up of a mixture of 3 grasses, 4 forbs, and 3 shrubs, which community would be more complex? If a community consisted of 25 noninteracting species and another contained 5 species involved in a complex web of interactions, which would be more complex? If two communities each contained 10 species, but one exhibited chaotic dynamics while the other hovered close to a constant number of individuals per year in spite of repeated perturbations, which would be more complex?

Another way to look at the problem is to plot, simply for illustrative purposes, complexity against biodiversity as done in Fig. 2. Aside from the trivial points where there is no diversity, and therefore no ecological complexity, and where there is some diversity and there is some complexity, a variety of relationships could be drawn between the two points depending on one's perspective or experience. That is, complexity and diversity could either increase hand-in-hand or show any possible relationship that is bounded by a positive, asymptotic relationship between diversity and complexity on one end and an exponential relationship between diversity and complexity on the other end (see Fig. 2). Of interest would be whether some general relationship could be described or whether every ecosystem would show a unique relationship. If the latter is true, then knowing a system's diversity would provide no means for predicting its complexity and whatever ecological behaviors might be a function of complexity.

Because ecological complexity is believed to be an important part of biologically diverse ecosystems and because diversity is currently rapidly declining, the role of ecological complexity is of considerable concern to ecologists. Does biodiversity loss mean loss of ecological complexity? If so, what are the ecological consequences of losing ecological complexity? This article reviews ecological complexity, its definition, its measurement, its relationship to biodiversity, and what is known about its role in ecosystem functioning.

![Figure 2](image-url)
B. Concepts and Definitions

Concern over the potential alterations in ecological complexity caused by anthropogenic global change led the United Nations Environmental Program’s Global Change and Terrestrial Ecosystem (GCTE) Core Project to establish Focus Four in 1996 to establish a research agenda for examining the ecosystem consequences of changes in ecological complexity. This program uses the following definition of ecological complexity:

Ecological complexity represents biological diversity but in a broad sense, including not only species diversity but also diversity of ecosystems and landscapes, as well as genetic diversity within species. In addition, ecological complexity involves the diversity of trophic pathways and interactions, that is, the connectivity of ecosystems. We can envision ecosystems with similar diversity but contrasting complexity as a result of different organizational structures.

—Sala, 1996

The first part of this definition of biodiversity was put forth in Article 2 of the 1992 Convention on Biological Diversity. In this Convention, biodiversity was defined as “the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic systems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems” (Harper and Hawksworth, 1995). This virtual synonymy with the first part of the GCTE definition of ecological complexity clearly indicates how strong a sense there is that biodiversity and ecological complexity are tightly correlated.

The second part of the GCTE definition concerns biotic interactions, which equates linkages among species with complexity. “Complexity,” however, is not clearly defined, nor is it apparent from such a definition how one might go about measuring ecological complexity. Complexity is actually quite difficult to define and as a concept it has received far less attention than other subjects in the philosophy of science, but if we are concerned about its significance in ecological systems it would be valuable to begin by clearly delimiting just what exactly is meant by “ecological complexity.” In most instances, it is largely a matter of judgment that one thing is considered more complex than another, a judgment often based on a fair number of factors rather than just one or two. Thus it is not surprising, as discussed earlier, that species richness alone will fail to capture what is meant by ecological complexity.

Chapman (1985), in his review of a symposium on the concept of complexity, identified four distinct meanings of “complex” used in the conference and in the literature. First, there is what Chapman called “complication,” or the process by which a system becomes naturally complex, but such complexity is ultimately “reducible” to fundamental principles or mechanisms. Second, there is “complexification,” which, like complication, concerns reducible systems, but involves purposeful design (e.g., a watch or a banana plantation has a different kind of complexity than the solar system or a tropical rain forest). Third, there is “complexification,” which concerns irreducible systems (systems that are not composites of simpler parts). Finally, the fourth usage is “complexity” itself, which is simply a state of uncertainty concerning whether a system is reducible or not (irrespective of it being natural or of purposeful design). These decidedly ugly terms are not likely to catch on, but they do point to the variety of ways that ecologists, like other scientists, use complexity.

C. Community, Phylogenetic, and Functional Complexity

Typically, the web of interactions within a community, characterized by the number, nature, strengths, and structural pattern of biotic interactions among species, is often what is treated as ecological complexity. This can be the web of all interactions (community web, including competition, facilitation, parasitism, predation, and other interactions) or a subset of the community web, such as just trophic or feeding interactions (food web, including predation, herbivory, disease, or other consumer–resource species interactions).

In contrast to community complexity, however, phylogenetic complexity may be of interest when a community contains a diverse array of taxa. From the perspective of systematists or paleontologists, simply having many species that are phylogenetically diverse is evidence for complexity (Raven, 1996). A system of 15 species of grasses from the same genus might be considered less phylogenetically complex than a system of 10 species from 10 genera of grasses, which in turn might be less complex than a system with 5 species each from a different plant family.

Finally, complexity of an ecological system may be tied to its functioning as a biogeochemical system rather than to its population or biomass dynamics. That is, the flow of nutrients and energy through the system may be of interest rather than the dynamics of the
many populations found within it. Note that ecosystems function in the sense that they exhibit activity or processes such as nutrient and energy flow. "Function" need not imply purpose or design (an unfortunate connotation of the term "functioning").

From the foregoing, it is apparent that ecological complexity is a composite feature of ecosystems consisting of an extraordinary number of different factors, including population, phylogenetic, and functional characteristics of the species in the community. The full measure of the complexity of a single ecosystem would therefore include a large number of factors, most of which are difficult to quantify. In general, increases in any of the properties listed in Table I, whether associated with structural or process intricacy, tend to increase ecological complexity, whereas decreases tend to simplify them. Table I demonstrates that measures of biotic diversity per se, such as species richness or evenness, are likely to be inadequate measures of complexity, though they may serve as crude proxy measures. The next section considers some basic measures of complexity, but not surprisingly, it will be evident that such measures of complexity generally deal with only single dimensions of complexity.

II. MEASURES AND CONSEQUENCES OF ECOLOGICAL COMPLEXITY

A. Measures of Complexity

Because complexity is a scale by which we compare communities and ecosystems, quantifying or measuring complexity is important in providing accurate means of making comparisons. Measuring complexity, however, is difficult because complexity is difficult to define. Boulding (1985) nicely described the problem at a United Nations conference held in Tokyo on the theory of complexity. He stated:

Complexity is a very complex concept, which is not altogether surprising. We do have a vague concept of it as a somewhat linear property, both of structures and processes, in which we make judgments of "more" or "less." I am pretty sure I am more complex than an amoeba... Exact measurement of the concept eludes us and probably always will, for although we do make these estimations of it in terms of more or less, it represents a highly multidimensional reality that cannot be reduced to a linear measure.

In spite of the obvious difficulties Boulding outlined, ecologists and evolutionary biologists have nevertheless provided a number of measures of complexity, which are reviewed in the following sections. These measures can be used in conjunction with measures of diversity to assess the relationship between complexity and diversity.

B. Measures of Community Complexity

Measures of ecological complexity are closely allied with measures of diversity, which reinforces the common practice among ecologists of treating biodiversity and complexity as one and the same. At its simplest, one might consider diversity to simply be the number of species ($S$), or species richness. Such a measure is

---

**TABLE I**

<table>
<thead>
<tr>
<th>Term</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural intricacy</td>
<td>Numbers of populations, Numbers of species, Numbers of biotic interactions, Kinds of interactions, Strengths of interaction, Higher-order effects, indirect effects, Connectedness, Number of patches, Arrangement of patches, Immigration probabilities, Emigration probabilities, Number of functional groups, Number of functionally singular species (inverse of redundancy), Number of links among functional groups, Number of compound functional groups (e.g., omnivores), Phylogenetic distances among species, Phylogenetic moment</td>
</tr>
<tr>
<td>Process intricacy</td>
<td>Stability, Predictability, Reliability, Degree of nonlinearity, Succession, Population growth, Ecosystem processes, Extinction, Speciation</td>
</tr>
</tbody>
</table>
often felt to be inadequate because it does not reflect differences in commonness and rarity nor the degree of connectedness among species. Thus, many other measures of diversity are used that attempt to take linkages or potential linkages into account. For example, the Shannon-Weaver index \( H \) considers both species richness and the relative abundances of species (evenness) as the important components of diversity. \( H \) quantifies diversity by assuming that richness and evenness directly represent the potential number of community configurations that a particular assemblage of species can produce. The original index quantified information, summing logged probabilities of a system having particular configurations.

The configuration of a community is seldom known, let alone the variety of possible configurations it can occupy or the probabilities of taking on these configurations. One therefore uses proportional abundances \( p_i \) of species as a means of approximating probabilities of configurations, a strategy that some consider fairly unsound.

The formula using this approach is

\[
H = -\sum_{i=1}^{S} p_i \ln p_i
\]

where \( i \) refers to the \( i \)th species in the community.

A related measure of potential complexity that compares the full potential with the realized potential for configurations is called connectivity \( C \), or the proportion of actual interactions (links or connections) to total possible interactions, which is distinct from food web connectance (see later). The formula for connectivity is

\[
C = \frac{\sum_{i,j} a_{ij} p_i p_j}{\sum_{i,j} p_i p_j}
\]

in which \( p \) is the probability of a species being present, and \( a_{ij} \) is the probability of interaction when two individuals, each from a different species \( i \) and \( j \), encounter one another. In reality, identifying interactions in biological communities is very difficult and cannot be deduced from abundance data. Indeed, Margalef and Gutiérrez (1983) explored connectance by examining components and "interactions" in functional plans for construction toys because information of equivalent clarity from real communities was unavailable.

The Shannon formalism serves as the basis for the preceding and other measures of complexity and diversity, but this practice is questionable. The Shannon formalism quantifies complexity by estimating an upper bound for the number of configurations possible for a given set of entities as if all configurations would be functional. In reality, however, rules of assembly ultimately determine the actual limits of information that a set of entities can store and retrieve efficiently. For example, if one had a box of 26 unique letters (i.e., A–Z), each equally abundant, and a box of 4 unique nucleotides, each equally abundant, the Shannon formalism would consider the box of letters to have a greater capacity to store and retrieve information. Assembly rules for letters, however, must be followed to produce sentences containing useful information just as assembly rules for nucleotide sequences must be followed to produce useful proteins. It is quite possible that 4 nucleotides could code for a greater diversity of proteins in nature than 26 letters can code for words in English, the difference being attributable to the rules used in each system. Without knowing these rules, it is difficult to assess the actual information content for a collection of entities such as letters, nucleotides, or species. Although assembly rules for nucleotides and letters in human languages are well understood, assembly rules for species in communities, although well studied, are still poorly understood.

Appraisals of complexity based on static patterns of the distribution and abundance of species [Eqs. (1) and (2)] ignore flows of energy or nutrients and the dynamics of interacting populations, all of which are considered key elements of community complexity (see Fig. 1). Rutledge et al. (1976) adapted approaches based on information theory, as discussed earlier, to consider flows among species within a community. Their approach usefully distinguishes between and quantifies both flavors of complexity, that is, structured versus disordered patterns of flows among species.

Appraisals of community complexity may also be made using the pattern of biotic interactions that links species in a community. The notion that biotic interactions are the primary elements of structural intricacy (complexity) that determine community dynamics has been the dominant theme in community ecology for several decades, inspired by the original contributions of A. J. Lotka and V. Volterra. Mathematical treatments of complexity have explored the consequences of varying interactions by constructing model communities whose species richness, interaction coefficients, and signs and strengths of interactions vary randomly and by examining their stability features. The heart of this approach is to use the "community matrix" as a measure of complexity, a concept formalized by Richard Levins in 1968 and still very much in use today. The general
approach is to assume that all populations can be modeled by differential equations of the form

$$\frac{dx_i}{dt} = r_i x_i (K_i - x_i - \sum \alpha_j x_j) / K_i$$

where $r_i$ is the intrinsic rate of increase of the population of the $i$th species, $K_i$ is the carrying capacity of the $i$th species, $x_i$ is the population size of the $i$th species, and $\alpha_j$ is the effect of the $j$th species on the $i$th species. For a community at equilibrium, where

$$K_i = x_i + \sum \alpha_j x_j$$

the system of equations can be written as

$$AX = K$$

Here, $X$ is the column vector of species population sizes, $K$ is the column vector of species-specific carrying capacities, and $A$ is the community matrix

$$A = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \cdots \\ \alpha_{21} & 1 & \alpha_{23} & \cdots \\ \alpha_{31} & \alpha_{32} & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

The properties of this matrix can then be used in analyses of complexity. Although wonderfully appealing as a logical measure of community complexity, identifying and quantifying interactions in a community is a formidable task and requires that a community be at or near equilibrium, rarely the case in nature.

One concern over measures of structural complexity based on biotic interactions is that they must account for "keystone" species. Defined originally by R. T. Paine in 1966 as a predator species that fosters the coexistence of competing prey species and thereby promotes structural complexity in ecosystems, it has become colloquially used to describe any species that has a disproportionate impact on the community in comparison to other resident species (Liebold, 1996) or species that serve as ecological engineers (Jones et al., 1994).

Power et al. (1996) proposed that the distribution of "community importance values" (CIVs) may serve as a tool for comparing communities for this element of structural complexity. Formally, the $i$th species in a community has a community importance value ($V_i$) defined as

$$V_i = [(m_p - m_0)/m_0](1/p_i)$$

where $m_p$ = the quantitative property (e.g., production or rate of decomposition) of the intact community, $m_0$ = the quantitative property of the community with the $i$th species removed, and $p_i$ represents the proportional abundance of the $i$th species before it was removed. In practice, one would sequentially delete one species at a time, each time measuring the community response to such a loss. Eventually a distribution of CIVs would be obtained and statistics (e.g., the mean or modal CIV and its variance) could be used to compare one community with another. Communities with mean CIVs furthest away from 0 would be more complex than those whose distributions were not significantly different from a normal distribution with a mean of 0. Of course, such an experiment would be very difficult to conduct and Power et al. (1996) could identify only two such studies.

One approach to the intractability of contending with all species interactions is to focus on a subset of interactions. An example of such a focus is the study of feeding or trophic interactions. Typically, food web statistics serve to quantify food web structural complexity. These statistics include food chain length, connectance (the ratio of existing trophic links to possible links), compartmentation (the number of compartments or isolated subnetworks of interacting species in a food web), the ratio of predator species to prey species, and interval versus noninterval state (presence or absence of complex overlaps in resource sharing). The difficulty with all of these statistics is that they rest entirely on the quality of the set of food webs, many of which were not constructed for such analyses, and thus produce artifacts when subjected to such analyses.

C. Measures of Functional Complexity

From a functional perspective, complexity may be measured as simply the number of functional groups. Functional groups may be considered collections of species that are related by their ecosystem activities, though many other definitions have been used. For example, if a functional group were defined as the group of species sequestering carbon from the atmosphere by photosynthesis, then all macrophytes, algae, and cyanobacteria found in a lake would be considered part of this
group, in spite of their very different ecologies and evolutionary histories. This degree of aggregation is much larger than what is typically found in contemporary food webs (trophic groups or species), community matrices (interaction coefficients from species pairs), or high-resolution phylogenies (orders, families, genera, species, etc.).

In the same way that tallying up numbers of species fails to capture complexity, tallying up functional groups is unlikely to serve as an adequate measure of functional complexity. More important would be the degree of parallel versus serial dependency among species in an ecosystem. As in community complexity, network architecture is the focus for measuring complexity, but rather than linkages of biotic interactions, only linkages of dependencies among species and functional groups become important.

Ecosystem reliability is an example of an index that combines measures of serial and parallel dependency (Naeem, 1998). Serial dependency occurs when every species or every functional group is dependent on the other for the ecosystem to function (e.g., decomposers depend on producers for energy while producers depend on decomposers for nutrient cycling). Parallel dependency occurs when one or more species within a functional group substitute for each other upon the extinction of one of the members of the group. In its simplest form, ecosystem reliability may be formally represented as

\[ H(t) = \prod_{i=1}^{\ell} \left[ 1 - \prod_{j=1}^{S_i} (1 - e^{-\lambda}) \right] \]  

(5)

where \( \ell \) is the number of functional groups, \( S_i \) is the number of species in the \( i \)th functional group, and \( \lambda \) is the probability of extinction over a small increment in time, \( t \). Note that the reliability of an ecosystem is defined as 1 \( [H(0) = 1] \) and eventually decays in the absence of replacement of species by recolonization \( [H(\infty) = 0] \).

As in other measures of complexity, the information needed for this index is difficult to obtain. Species presence or absence, local extinction rates, and functional roles are often difficult to determine without considerable empirical work. Focusing on redundancy demonstrates, however, that if the properties being measured are the result of the collected activities of groups of similar (i.e., redundant) species, such as production, having many species could actually mean more steady levels of production (less complex behavior). Thus, the association between biodiversity and complexity of ecosystem functioning is dependent on the degree of species redundancy in the system.

**D. Measures of Phylogenetic Complexity**

Whereas community and functional complexity focus on structural aspects of ecosystems, phylogenetic complexity examines the structure of the phylogenies that relate species to one another within a community by their evolutionary history. As with community and functional complexity, the number of species is an important starting point, but the architecture of the evolutionary "tree" or "bush," as opposed to network architecture, serves as the primary basis for assessing phylogenetic complexity.

A straightforward measure of phylogenetic complexity might be simply the sum of the lengths in the tree that unifies all species in the set being examined or a simple modification of the Shannon-Weaver index in which species are weighted by their taxonomic distinctness. More informative, however, is the "phylogenetic moment" [defined in Eq. (6)] which provides a statistic by which communities might be compared. A difficult necessity for using this measure is that one has to have the complete phylogenetic tree at hand (referred to as tree \( T \)), including those species that might have gone extinct and might be missing from the local community being examined. The current community will consist of some subset of \( T \) (referred to as set \( R \)), and the local community that one is measuring will invariably be missing some species and contain an even smaller set of species (referred to as set \( r \)). If we assume that the tree (its nodes and branches) can be represented by a series of points in time, with a total of \( PT \) points for the whole tree, then the phylogenetic moment for this subtree \( (M_r) \)

\[ M_r = \sum_{p \in P} d_{pr} \]  

(6)

where \( d_{pr} \) is the distance between a point on the tree and the nearest species in the set \( r \). The smaller the phylogenetic moment is, the more closely the community represents the full phylogenetic structure (found in \( T \)). For example, in Fig. 3, although the total distances among the species in trees \( R_1 \) and \( R_2 \) are the same, using the phylogenetic tree \( T \), which includes extinct species \( F \) and \( G \) and missing but extant species \( B \), shows that \( R_1 \) would have a smaller phylogenetic moment than \( R_2 \) because \( C \) would have more near neighbors than \( D \). That is, \( A, C, \) and \( E \) will presumably contain genetic
interactions, their functional roles, and their phylogenetic relationships. As the next section shows, this means that we currently have only a partial picture of how complexity and biodiversity are related.

III. THE SIGNIFICANCE OF ECOLOGICAL COMPLEXITY

A. Complexity in Nature

Although diversity and complexity are practically universal features of ecosystems, we have much better information on diversity than on complexity. The estimation of existing diversity is, to be sure, challenging, but determining the distribution and abundance of species would only be the starting point for investigations of ecological complexity.

Unfortunately, the components of complexity are extraordinarily difficult to measure and for this reason empirical treatments have lagged behind the substantially greater progress made in theoretical explorations of the significance of ecological complexity. An ecologist never has a complete inventory of all species within a region, let alone their relative abundance, biotic interactions (direct, indirect, and higher order), phylogenetic relationships (at the regional level, including extinct species), functional roles, and population and process dynamics, or other details listed in Table 1.

In most cases, however, the question at hand is more focused and thus more tractable. If the question concerns population stability, then quantifying only community complexity may suffice. If the question concerns what sets of species are most likely to contain the bulk of genetic information stored within a regional biota, then quantifying phylogenetic complexity may be enough. If the question concerns the magnitude or rate of ecosystem functioning, then quantifying functional complexity may do.

What is known about the relationship between diversity and complexity in nature? Several important patterns that emerge from studies that have examined community, functional, and phylogenetic complexity are considered here.

1. Community Complexity

An early and influential work that examined community complexity was R. H. MacArthur’s 1955 study, inspired by E. Odum’s theory that the “amount of choice which the energy has in following the paths up through the food web is a measure of the stability of the community.” Using the Shannon-Weaver formula to model the num-
ber of alternative pathways and relate this to stability. MacArthur concluded that "stability increases as the number of links increases," though he was primarily considering resistance stability (remaining intact in spite of perturbation).

The theoretical relationship between complexity and stability is treated elsewhere in this volume, but it is worth noting here that this theory stimulated further investigations into the relationship between complexity and ecological processes, such as population dynamics. Later mathematical studies of local resilience stability (the ability to recover after perturbation) showed an inverse relationship between community complexity and resilience stability. Studies of connectivity and connectance (the nature and degree of connections among species) showed that resilience stability could either increase or decrease, depending on which trophic level was examined and the degree of connectance. Pimm (1979) also showed that complexity and stability could be positively related if one considered the deletion of species rather than resilience stability. Thus, those biologists who observed a positive association and those who observed a negative association between diversity and complexity more than likely differed in how they defined stability or how they defined complexity. Suffice it to say that there was agreement that complexity and stability were indeed related, but the nature of the relationship was dependent on one's definitions.

Although food web studies initially showed much promise as a means for examining the relationship between complexity (community structure) and diversity by focusing solely on trophic interactions, early findings did not fare well in the test of time. Polis and Winemiller (1996) summarize four patterns derived from early analyses of topographic or static food webs. First, there is constancy in the number of links per species in a food web irrespective of number of species. Second, the numbers of trophic levels seem to be invariably small (three or four). Third, the ratio of predator to prey species is relatively constant. Fourth, omnivory is rare. All of these patterns, however, have been challenged as artifacts.

Though potentially artifacts, food web patterns at least touched upon an obvious truth about communities—there are likely to be constraints in the structures of food webs. For example, at the simplest level, one cannot have consumers without resource species (i.e., no predators without prey) and there are likely to be stability or energy constraints that limit food chain lengths (Krauzinger and Morin, 1998). Currently, food web biology has gone far beyond the sorts of patterns that initially stimulated research in this field, but new patterns, if they exist, await the completion of a new set of higher-resolution food webs.

The main observation to be drawn from studies of community complexity is that complexity is never at its maximum in diverse ecosystems. For example, although the potential number of pairwise interactions is large for any community, few of these interactions may ever be realized. As Margalef (1985) states, "the considerations of artifacts [from mathematical studies of diversity and complexity] provide an elementary but effective approach to the assertion that no system is completely interlocked or connected." We may infer from this fact that complexity asymptotes with diversity, an important consideration in our quest to understand the relationship between diversity and complexity (see Fig. 2).

2. Ecosystem Functioning

Biodiversity and ecosystem functioning are treated elsewhere in this Encyclopedia, but here I examine their relationship to complexity and functioning. May (1974) noted in his "afterthoughts" to the second edition of his influential book that aggregate properties of communities, such as production (i.e., ecosystem functions), could show the opposite relationship with complexity that dynamic stability showed. Theoretical support comes from different approaches, but experimental tests of the idea are few. Support for this idea can be found in Tilman and Downing's (1994) study of experimental grassland plots, though some controversy surrounds this study. In their study, production recovery from a drought (resilience stability) increased with plant diversity. Two laboratory microcosm experiments using freshwater microbial communities by J. McGrady-Steed et al. and S. Naeem and Li (both conducted in 1997) showed that predictability and reliability of ecosystem functions, such as production and community respiration, can indeed be associated with biological diversity and complexity.

3. Phylogenetic Complexity

The relationship between phylogenetic complexity and biotic diversity is not well understood. Globally, estimates of origination and extinction of species suggest more or less steady increases in biotic diversity punctuated by aperiodic mass extinctions. This steady rise may be taken to mean that origination rates, over millions of years, have been outpacing extinction, except during periods of mass extinction, and that recovery from mass extinctions has always occurred. Such a history of the biota suggests, if true, that the biota can tolerate even very deep "prunings" of the tree of life. At a global level,
this suggests that random sets of species drawn from regional pools are likely to contain equivalent amounts of genetic information and that it takes severe levels of extinction to reach the point where insufficient genetic information is left in the remaining taxa to regenerate lost biodiversity. Because large-scale patterns in diversity across environmental gradients are similar across many taxa (Currie, 1991), we may expect that random samples from communities would result in phylogenies of similar complexity. Of course, since only a fraction of all species have been described, it is difficult to know exactly what the distribution and abundance of species are. As in community complexity and functional diversity, the true relationship between diversity and complexity is unknown, but it is likely to be asymptotic. That is, highly diverse systems may experience little change in phylogenetic complexity in the face of moderate levels of extinction.

V. CONCLUSIONS: THE DISENTANGLED BANK

Common usage of the term “complexity” in ecology implies that ecological complexity is a function of the phylogenetic, functional, and community structure of ecosystems. On the surface, ecological complexity and diversity seem so closely related that it is tempting to consider “biodiversity” as synonymous with “complexity.” If we consider complexity to be the structural and process intricacies of a system, however, it is clear that diversity, in terms of numbers and types of organisms in a community, shows no necessary relationship with any measure of complexity aside from the trivial fact that without any diversity there is no complexity. Thus, two communities identical in numbers of species could be quite different in community, phylogenetic, or functional complexity. Likewise, a highly diverse system could be more or less complex than a depauperate system.

Studies of complexity in nature suggest asymptotic relationships between complexity and species richness, where species richness represents one way to define biotic diversity independent of complexity. It is hard to imagine, for example, that tropical rain forests, where species often number in the thousands, will show much reduction in complexity in the face of the loss of a few species. Current rates of extinction, however, are estimated to be staggeringly high and many diverse systems, such as tropical ecosystems, are not suffering extinction of species within their boundaries so much as they are being transformed to depauperate, managed ecosystems.

Thresholds in the response of ecosystems to the loss of diversity are particularly of interest (Levin, 1997) and these may be hypothesized from current knowledge about the relationship between diversity and complexity. From what is known about nature, both theoretically and empirically, it is reasonable to postulate that both community and phylogenetic complexity show asymptotic relationships with diversity. Because functional complexity aggregates many species into small numbers of functional groups (e.g., all plants may be considered producers), we can postulate that the asymptote for functional complexity will be lower than that for community or phylogenetic complexity. When these are plotted together, as shown in Fig. 4, we can identify two thresholds of interest. At some point, admittedly difficult to define precisely without accurate measures of community complexity or phylogenetic complexity, one will continue to experience no appreciable loss in functional complexity, but ecosystem properties, such as dynamic stability, may change and genetic resources will decline. As diversity declines even further, functionality is eventually lost after the second threshold (the leftmost dotted line in Fig. 4) is crossed.

Managed ecosystems, such as aquaculture systems, farms, or monoculture tree plantations, most likely reside at the lower threshold, where the managers maintain the system at its minimal diversity. In general, these ecosystems may be functionally complete, containing all necessary functional groups to ensure that the desired ecosystem functions are maintained or, if not functionally complete, humans provide missing functions, such as fertilizer addition where nitrogen fixers have

![Graph showing hypothetical relationships between complexity and species richness.](image)

**FIGURE 4** Hypothetical relationships between complexity and species richness. $C \approx$ complexity. Vertical dotted lines indicate where the threshold in ecological consequences of declining ecological complexity might be identified. $C_c \approx$ community complexity; $F_c \approx$ functional complexity; $P_c \approx$ phylogenetic complexity.
been removed. Complex systems are always more difficult to manage, but there are likely costs and consequences for resorting to simple systems, many of which have to do with system stability, system utility, and system efficiency, as the material in this article reveals.

The appropriate metaphor for the process of human transformations of ecological systems may be the disentangling of Darwin's entangled bank. As the earth's ecosystems are steadily transformed to agroecosystems and aquaculture systems, there is no question that average levels of biodiversity in ecosystems will continue to experience sharp declines in local ontic diversity, which eventually leads to losses in complexity. In many cases, understanding the ecological consequences of this loss may require focusing on ecological complexity, for which the quantification of species richness is only the first step.

See Also the Following Articles

DIVERSITY, COMMUNITY/REGIONAL LEVEL • FOOD WEBS • GENETIC DIVERSITY • LANDSCAPE DIVERSITY • MEASUREMENT AND ANALYSIS OF BIODIVERSITY • SPECIES DIVERSITY, OVERVIEW

Bibliography


I. What Is a Model?
II. The Purposes of Modeling
III. Types of Models
IV. Limitations of Models
V. Some Tools of the Trade
VI. Applications
VII. Future Trends

GLOSSARY

aggregation Combining several potentially separate components of a system to simplify analysis.
dynamic model Mathematical description of a system that has components that vary in time.
multimodel Single integrated model that links together models taking different approaches.
parameter Constant in a model that must be estimated from data, or assumed to be of a particular value.

MODELS IN A VARIETY OF FORMS play a critical role in advancing our understanding of natural systems. Models abstract basic principles and derive the implications of such abstractions. This provides a method to analyze alternative hypotheses about natural system responses and the mechanisms that underlie these responses. This article presents an introduction to various modeling approaches and discusses how these have been applied to analyze species abundance and distribution.

I. WHAT IS A MODEL?

Just as storytellers can take their audience on trips to faraway places and provide a glimpse at life in different cultures, scientists tell their stories about the way the world works by making models. Such models never provide a complete view of how the world works, but do give us glimpses that help us to piece together interactions between different parts of the world and the processes that connect them. These models take many forms, some being mostly verbal, others mostly qualitative and graphical, some phrased in various mathematical forms, and still others set up as collections of rules within a computer program.

Models provide maps of varying levels of complexity to help us understand the topography of science. There are coarse road maps that provide merely the outline of major arteries for traffic, telling us nothing about buildings or other features of the landscape, but providing an overview of the linkages between key components of a system. More elaborate models show us the buildings and the infrastructure that links these buildings—the sewer and power lines. Even more complex models would indicate the humans in each building, their occupations, and the flow of money or capital...
goods between them. Similarly, models at many levels of complexity can be useful for addressing different questions in the study of biodiversity. A very coarse model might analyze the effect of land use change worldwide on total species richness. A more complex model could consider particular regions and analyze differentially the changes in land use within them and the associated changes in species numbers. A still more complex model could consider the local dynamics of individual species and from this elaborate the dynamics of species numbers.

II. THE PURPOSES OF MODELING

A. General Objectives of Models

Which type of model one constructs depends on the questions being asked and the availability of data to construct a reasonable model. Though there are many specific purposes for constructing models (Haezinger, 1996), these may be grouped into a few general objectives: Description, Mechanism, Prediction, and Control. These objectives are not mutually exclusive, so that descriptive and mechanistic approaches may be used to aid prediction and control.

1. Description

Sometimes all that might be desired is a simple description of a collection of data. For example, an average provides a single value to summarize a list of numerical data. This may be sufficient for some purposes that do not require a description of how much variation is in the data. To assess variation, a dispersion measure such as the variance would be needed. These summary statistics are coarse, ignoring many of the details in the data. Yet they allow us to easily comprehend major differences between different data sets. Extensive species lists within a region from two locations may be compared by considering just the total numbers of species in the two locations and the number of species in common between them. Such a summary may be sufficient for a comparison of the two locations, while ignoring details such as the diversity within the taxa included. Descriptive approaches may be much more complex than simply providing averages and variances. Indeed the field of exploratory statistics deals with methods to analyze and summarize multidimensional data (Jambu, 1991). A typical example would be methods of time-series analysis in which the histories of species numbers might be compared between two locations or correlated with the histories of anthropogenic actions in the location.

2. Mechanism

If the objective is to provide an understanding of how a particular system operates, then it is necessary to take account of the processes that govern the system. While all such mechanistic models are descriptive at some level, the point here is to deal with the basic physical, chemical, and biological processes operating in the system. This requires including those processes that operate at a spatial and temporal extent appropriate for the problem one is addressing, and ignoring others. Thus, analysis of how alternative global warming predictions would affect worldwide biodiversity might include the geographic variation in the temperature predictions at a spatial extent of hundreds of square kilometers, but would not doubt ignore the microclimate variation of every square meter. Even if it were possible to characterize the meter-by-meter temperature differences predicted by the alternative warming trends, the lack of available detail on the species present at this detailed spatial resolution limits the utility of including such detail. A discussion of such scaling issues is included elsewhere in these volumes.

3. Prediction

Predictive models are of two general types: those that attempt to project the behavior of the system based on certain explicit assumptions, and those that attempt to forecast the future behavior. The difference is between what might be true in the future if certain assumptions hold (projection) and what will be true in the future (forecasting) (Caswell, 1989). In many biological situations, the forecasting problem is not even attempted, as it would involve taking account of a wide variety of unpredictable abiotic phenomena (e.g., hurricanes and droughts). It is often feasible to construct a model to project the future dynamics of a system based on current observations and particular reasonable assumptions about the interactions in the system. The majority of population models (discussed elsewhere in these volumes) are of this form, in which abiotic influences are not included. These models can project the future behavior of the population based on the biotic forces of demography, genetics, and social structure within the population. Uncertainties associated with unpredictable phenomena can be taken into account by attempting to project just the mean and variance of the variables of interest (e.g., population size).

4. Control

When a system has one or several components that are under human control, either completely or in part, then
a model can be used to help determine how to apply such a control to meet certain objectives. Examples of controls are harvest quotas, limits on fertilization or pesticide application, flows from a dam, limiting importation of potentially harmful invasive species, and land use zoning regulations. Examples of objectives are maximizing biodiversity, minimizing population extinction probability, reducing the spread of nonnative species, and maintaining population size above some determined threshold (such as a minimum viable population size). Control models mostly focus on the dynamics of the system, with the simplest form of control being bang/bang, meaning on/off, such as allowing harvest in certain years and not allowing harvest in other years. A related objective is for control models to produce a relative comparison of alternatives in order to rank these alternatives according to some criteria (DeAngelis et al., 1998). Still other control models are used to analyze the physiological responses of individual organisms to varying environments and the homeostasis that can arise through these responses.

III. TYPES OF MODELS

Models can be physical, such as animal models used in drug testing and airplane models used in wind tunnels. In biodiversity contexts, microcosms and mesocosms, which are limited biological systems built in a laboratory setting, play this role. These are meant to mimic the key biotic forces interacting within a natural system, but are constructed at a spatial extent that allows for easy observation and controlled experimentation. They cannot include all of the components of the real system, but do allow for projection of how the real system might respond under particular perturbations. Physical models clearly are limited, particularly to organisms that are mostly sessile or have very short distance movements.

Mathematical models come in a wide variety of forms. Some are simply graphical relations that show the qualitative relationship between certain components of a system, mainly to demonstrate the shape of response and whether one component increases or decreases with another. An example would be the increase and then decrease in species diversity along a gradient from low to high frequency of disturbance (Fig. 1). Here, there is no attempt to predict at exactly what disturbance frequency the exact peak in species diversity occurs. Rather, the objective is to illustrate the qualitative behavior of diversity, representing the "intermediate disturbance hypothesis," which posits that highest diversity occurs at intermediate disturbance frequencies.

The majority of mathematical models in ecology deal with the dynamics of populations and communities. Such models consider the basic processes of birth and death, immigration and emigration, and competition and predation to elucidate general theories of population dynamics. Described using differential or difference equations, these models allow for projection of the long-term behavior of populations, as well as provide methods to project the within-population structure (age, size, genetic, etc.).

Computer models are quite varied in structure. First, all the standard mathematical models of populations and communities, constructed using differential or difference equations, may be implemented on computers. Indeed, since it may be quite difficult to develop analytic solutions for such models, analysis of their behavior often requires the use of numerical solution methods implemented on a computer. There are many computer models that, although they may have a description that is essentially mathematical in form, are really described by the code itself rather than an explicit set of mathematical equations. An example would be cellular automata models, one type of which consists of a two-dimensional lattice, with each point on the lattice having one of a number of states (Langton, 1988). The simplest situation would be each lattice point being occupied (e.g., in the 1-state) or unoccupied (in the 0-state). The model is then described by a set of rules that determines how the state of a lattice point changes from one time step to another, based on the states of surrounding lattice points. Such a cellular automation
may be used to mimic the spatial dynamics of populations, in which each lattice point represents a possible location of an individual. Alternatively, each lattice point can be interpreted as a local population, and the entire lattice then can follow the collection of such populations, called the metapopulation.

System simulation models are elaborate computer models that attempt to include most of the biotic and abiotic factors that affect the system. Many agricultural system models are of this type, and include the crop, its pests, soil nutrients, and weather conditions, among other factors. Some other types of computer models are described in later sections. In all cases, though the model is in essence specified by the code itself, it is very useful to have some graphical description of the major components of the model. One example is shown in Fig. 2. There are a number of general modeling software packages designed explicitly to aid construction of computer models through the use of graphical elements.

IV. LIMITATIONS OF MODELS

A. Trade-offs: Generality, Precision, Realism

No one model can do everything. In the process of deciding what components of a system to include, what processes to consider, and what spatial and temporal extent is appropriate, the model excludes part of reality. Modeling is a process of selective ignorance. We decide what to include and what to exclude. Part of the art of modeling is coming to grips with the issue of which details are important and which ones are not. In most cases the process is iterative, with a sequence of different models being tried until a model is arrived at that includes just the essential details necessary to address the problem of concern.

One view of the trade-offs in constructing a model is that no one model can be simultaneously general, precise, and realistic (Leyvins, 1968). As Fig. 3 illus-
trates, these properties may be viewed as points of a triangle. Generality implies that the model may be useful in many different natural systems. A realistic model is one with components, parameters, and variables that are all possible to estimate from observations. A precise model is one that produces quantitative, accurate descriptions of the natural system. Models for theory development, including most of the classical population and community models, are quite general, somewhat realistic, but lacking in precision. Descriptive models designed to mimic the response of particular systems tend to be quite precise, slightly realistic, and not at all general. Most regression models are of this type. They may provide an accurate portrayal of a particular system, for example, winter wheat growth in Nebraska, but are not transferable to other situations such as winter wheat growth in eastern Russia. System simulation models tend to be quite realistic, somewhat precise, but not very general. Control models take up various positions in the figure, depending on the level of precision desired.

B. Aggregation and Loss of Detail

A major factor that affects where a particular model fits into the scheme shown in Fig. 3 is the amount of aggregation included. Natural systems consist of many components that can be lumped together or disaggregated. Population models that use a single variable to represent the whole population must inherently ignore the within-population structure (e.g., age and size). Such a model would not be able to discriminate between a population with mostly small individuals and one with mostly large individuals, unless this within-population structure was assumed to affect the population’s growth characteristics in some manner. Aggregation implies a loss of detail that allows for quite general models. Often this gives model parameters that are not at all easily estimated from observations. Examples would be the growth rates in population models and the competition coefficients in community models. The less aggregation in the model, the more parameters there tend to be and more data are required in order to estimate them.

C. Uncertainties: Mechanisms, Data for Parameterization, Biotic and Abiotic Forcing Functions

The modeling process is limited by the information available. There may not be basic agreement on the mechanisms that are critical in the system of concern, so that any particular model includes only some of the mechanisms, or just partial information on these mechanisms. Many population models applied to vertebrates ignore social structure, despite evidence that it is often present in such populations. This is typically due to lack of understanding of the effect of such structure at population levels, and gives rise to another set of models designed specifically to investigate these effects.

Under situations in which the mechanisms are well understood, it may not be possible to accurately estimate model parameters because adequate data are lacking. There are a variety of statistical methods designed specifically to determine the optimal choice of parameters in such situations (Hilborn and Mangel, 1997). These methods may take account of parameters estimated either for similar species or for similar locations other than the one being considered. For example, it may be necessary to use observed clutch size distributions for one bird species in application to a similar species about which such information is lacking. Another uncertainty associated with natural system models arises due to the unpredictability of forcing functions such as weather and disturbance. If historical information is available on these functions, this may be used to estimate the stochastic effects of such forcing. These various uncertainties limit the detail at which models may be constructed, and thus limit the types of questions that may be addressed using models.

V. SOME TOOLS OF THE TRADE

A. Statistical Approaches

Statistical models usually have a descriptive objective rather than a mechanistic one. The parameters within
these models are directly estimated by choosing them in a manner that best fits a certain data set. Thus any particular statistical model is typically not very general in application to different systems. The structure of such models may be useful in a wide variety of different contexts. Regression models, which assume a particular mathematical relationship between variables and assume that errors in the data take a particular form, are widely applied. Numerous regressions have been estimated for species richness as a function of latitude, altitude, and rainfall (Huston, 1994). Discussion of statistical methods applied to estimation of population sizes and densities may be found elsewhere in these volumes.

B. Dynamic Models

Although many of the traditional models for populations and communities are in the form of dynamical systems (e.g., collections of linked differential or difference equations), often the types of analyses performed for these models are based around equilibrium assumptions. The objective is to find long-term asymptotic behavior. This may be a static equilibrium (e.g., population sizes approach a constant value through time) or a dynamic equilibrium (e.g., population sizes follow repeatable patterns through time) (Murray, 1989). While these situations may arise, many models produce behavior that does not have a long-term equilibrium structure. Another key objective is to determine the stability characteristics of any equilibria that arise, in the mathematical sense of determining whether a model that is perturbed from an equilibrium condition will return to it. The dynamics arising in all these models takes account of the basic demographics of the population, as well as interactions with other populations. Adding abiotic conditions such as temperature and rainfall, or adding spatial components, often requires that the analysis be done using numerical simulations.

C. Geographical Information Systems

The advent of remote-sensing methods using airplane cameras and satellite imaging has opened new possibilities for following and modeling the responses of the earth’s biota. A key tool that allows the use of such materials are Geographical Information Systems (GIS), which enable computers to graphically display the remote-sensing data as two-dimensional maps. Each image may represent one aspect of an underlying landscape, such as landcover or vegetation type. The image value at any particular location (or pixel) in the map is estimated using models that classify the output of the cameras or the multispectral scanners on satellites into types appropriate for the objective. These models require ground-truthing to ensure that the estimated value for a particular location matches what is actually present. GIS methods allow various spatially explicit components of a landscape to be combined by looking at different map layers (different images measuring different aspects of the landscape). A mathematical function is then applied that averages or applies thresholds to these various components. Estimates of regional and worldwide carbon uptake are obtained using such methods applied to vegetation maps, in which different carbon assimilation values are assigned to different vegetation types, linked with weather maps supplying temperature and rainfall patterns.

VI. APPLICATIONS

A. Habitat Suitability Indices

Habitat Suitability Evaluation Procedures (HEP) are a formalized methodology for impact assessment on wildlife habitat. These are based on Habitat Suitability Index (HSI) models, which attempt to summarize the site characteristics that affect the utilization of particular habitats by a variety of wildlife species. Numerous HSI models have been constructed, typically consisting of very simple regression-type models. The key habitat variables are often some measure of canopy cover in a variety of classes, diameter classes of trees and shrubs, tree stem densities, area of open water, and distance to forest cover, among others. The objective is to combine these variables, based on extensive field observations done in a correlative manner, to provide overall indices of suitability. HSIs are always indexes with values between zero and one, and they are assumed to be proportional to carrying capacity.

HSIs are based only on local habitat variables; they completely ignore any effects due to species interactions, except those due to indirect effects on related habitat variables. The models ignore the spatial interactions of habitat types across a landscape. This leads to difficulty in situations for which the size, shapes, edge effects, and neighborhood relationships have a greater effect on habitat preference than local forest composition and structure variables. The models also do not take account of the issue of presence/absence of a species, and thus ignore any historical influence on potential local abundance. HSIs are inherently static entities, so any dynamics they produce are driven completely
by changes in habitat variables and not by the inherent dynamics and demography in the species being considered. Despite these criticisms, HSI s are perhaps the most commonly used set of ecological models, in part because users realize their limitations and view them as a simplified tool to summarize a very complicated situation by a single number. Such simplification must result in a loss of information, but a key issue regarding HSIs is whether they indeed can be used as a useful predictor for abundance. A summary of the HSI modeling approach is given in Verner et al. (1986).

B. GAP Analysis
The GAP Analysis Program (GAP) is a nationwide comprehensive effort in the United States to inventory plant and animal species, computerize the results, and have the capability to analyze spatially the relationships among different taxa. GAP relies heavily on GIS methods, and has as a major objective the capacity to identify "gaps" in biodiversity. Such gaps are presumed to arise in locations that are expected, from biological knowledge of species requirements, to have certain mixtures of species present, but for some reason do not. The focus of these efforts is not on rare or threatened and endangered species, but rather on the status of ordinary species and their habitats in order to inform policymakers to improve their decisions. Much of the remote-sensing information cannot determine the presence or absence of particular species, because the technology can provide information only on basic vegetation components of a landscape, and landforms such as rivers, roads, and urban areas. Thus, an extensive ground effort is under way to survey the distribution of various species, map these, and relate these mapped distributions to habitat variables, land use, and ownership, and other related species distributions. The focus is on vertebrate species, basic floristic types whose presence can be estimated from satellite image analysis, and landform information.

GAP analysis does not attempt to explicitly model biodiversity, but rather uses the preceding surrogate measures as a method to assess spatial patterns in biodiversity. There are acknowledged taxonomic biases in this approach. The objective has been to provide basic methods to assess the impacts of the rapid land use changes that have occurred in the United States over the past several decades and how these impact biodiversity. Other limitations of this approach are the lack of detailed demographic information included for the species of concern, the static nature of the project, and the reliance on mapping methodologies that require extensive ground-truthing to ensure accuracy. Full documentation on the GAP may be found on the web at http://www.gap.uidaho.edu/gap. A basic reference is Scott et al. (1993).

C. Metapopulation Models
Both HEP and GAP are attempts to account for the spatial nature of biodiversity by including explicit maps of basic habitat variables. An intermediate approach between models that include all the spatial detail available from maps and those that ignore all spatial aspects of a system are metapopulation models. These models consider a landscape to be split up into a number of localized populations, called subpopulations, with the entire collection of these called a metapopulation (Gilpin and Hanski, 1997). Most of the biotic interactions driving population dynamics occur within the localized subpopulations, but there are exchanges of individuals between these subpopulations. This allows for differing environmental, demographic, genetic, and disease situations to be present in the subpopulations. Depending on the assumptions about transport of individuals between the subpopulations, these can be relatively isolated or closely coupled.

A clear advantage of metapopulation models is the ability to derive analytic results, such as equilibrium and stability behavior, as a function of the within-subpopulation characteristics and the between-subpopulation factors such as movement. The models are particularly appropriate for cases in which a landscape can be reasonably viewed as containing discrete patches of habitat suitable for the population, with the intervening regions not suitable. The level of detail in these models can be quite variable, with the simplest versions just treating subpopulations as either present of absent. More complicated models take account of demographics within each subpopulation or explicit details on the relative spatial locations of each subpopulation that affect movement between them. These can be used for a population viability analysis, in which the probability that the overall population will survive for varying time periods is estimated.

D. Individual-Based Approaches
All of the foregoing approaches include various levels of aggregation in their components in order to simplify the model. An alternative reductionist approach is to take account of differences between individuals within a population, allow the individuals to feed, grow, and interact, and from the aggregated behavior of these indi-
Individuals build an understanding of population-level responses (DeAngelis and Gross, 1992). These individual-based approaches are increasingly common thanks to advances in computer science and the increasing availability of data on behavior and physiology of species of particular interest. The advantages of these approaches include the ability to consider the effect of abiotic factors on populations through their direct impact on individual behavior and growth, to take account explicitly of spatial variation in habitat factors, and to deal with small populations in which individual differences within the population can have great impacts on population-level responses. Disadvantages of individual-based approaches include the requirement for a great amount of detailed data to realistically simulate individual behaviors, and the typical necessity of making numerous simulations to evaluate any particular response that may arise because of the stochastic nature of the models.

VII. FUTURE TRENDS

A. Multimodeling and Regional Assessment

Natural systems have many interacting components operating at a variety of temporal and spatial extents and requiring differing levels of detail to describe the interactions between them. One historical approach in ecology to model such systems is to break it into a number of compartments (often for different trophic levels, and sometimes with grouping within each trophic level) and consider the dynamics of each compartment with movements of energy, biomass, or nutrients among them. It is quite difficult to make these systems analysis approaches spatially explicit or to link them to GIS. Yet it is now becoming possible to link together a variety of different modeling approaches in order to best utilize the available data, with different resolutions at different trophic levels, and carry this out in a spatially explicit manner. Such multimodels may use very simple models similar to HSIs for some trophic components, more complex dynamical systems for certain populations with mostly very localized interactions, and individual-based models for organisms that move great distances and average over the spatial heterogeneity. One example of such an approach is the ATLLS (Across Trophic Level System Simulation) Project, which is an ongoing attempt to build a multimodel for estimating the biotic impacts of alternative water management plans on the Everglades of south Florida (DeAngelis et al., 1998).

Building multimodels requires extensive landscape data obtainable from remote-sensing and ground efforts. The approach is inherently dynamic, and thus requires methods to estimate the spatial dynamics of key environmental drivers, or else have available a history of this spatially that can be analyzed statistically. In the Everglades, the major driver is water, and both historical data and detailed hydrologic models are available to provide estimates for scenario evaluations. Without such data, assumptions must be made about the dynamics of the landscape. With the reduced cost of data storage and the development of standards for spatially explicit data, it is expected that future models will have available extensive time series of remotely sensed images to both calibrate multimodels and provide the opportunity to iteratively improve their predictive abilities. For problems at regional levels, such multimodels are a rational method to aid planning while taking account of the best scientific data at the variety of resolutions available.

B. Behavioral Dynamics of Species of Special Concern

Great strides are being made in improving our understanding of the conservation biology of rare, threatened, and endangered species throughout the world. In an effort to better estimate the responses of populations of these species, numerous remote-sensing methods have been employed to track the movements of individual organisms. These include radio-tracking devices implanted within sampled individuals, which allow explicit location and physiological data (e.g., body temperature) to be obtained regularly throughout the individual's life from either satellites, airplanes, or sampling locations on the ground. The technology has advanced so that implantation of such devices can be done rapidly with no determinable harm to the individual. When done for many individuals within a population, it is becoming possible to follow the behavioral dynamics of mixtures of individuals. This includes the ability to ascertain details of mating, territoriality, and aggressive interactions.

It is likely in the future that managers concerned with a particular species will be able to observe in real time the movements of many individuals within a population. Then they might apply spatially explicit modeling methods to project the response of the population to particular management alternatives. This scenario suggests that the simultaneous monitoring of other species that interact with the species of concern
would allow for the regular application of adaptive management methods.

Acknowledgments

The author appreciates the support of the U.S. Geological Survey, Biological Resources Division, through Cooperative Agreement No. 1445-CA09-95-0094 with The University of Tennessee. The statements, findings, conclusions, recommendations, and other data in this report are solely those of the author and do not necessarily reflect the views of the U.S. Geological Survey.

See Also the Following Articles

ECOSYSTEM FUNCTION MEASUREMENT • MEASUREMENT AND ANALYSIS OF BIODIVERSITY • REMOTE SENSING AND IMAGE PROCESSING.

Bibliography


CONSERVATION BIOLOGY, DISCIPLINE OF

Andrew P. Dobson* and Jon Paul Rodríguez†
*Princeton University and †Centro de Ecología, Instituto Venezolano de Investigaciones Científicas

1. Definition
2. Ecosystems and Communities
3. Species and Populations
4. Individuals and Genes
5. Future Challenges

GLOSSARY

Allee effect Biological phenomenon whereby the survival of the individuals of a population increases by spatial aggregation.

community All forms of life that coexists and interact with each other in a particular habitat.

ecosystem Complex dynamic system that arises by the integration of a biological community with the physical environment.

ex situ conservation Maintenance of live components of biodiversity physically removed from their original habitat.

extinction vortex Positive feedback loops that increase the risk of extinction of a population as it declines in size.

gene Functional unit of heredity; the portion of a chromosome of an organism that encodes for a particular protein.

gene bank Facility where genetic material is stored in the form of seeds, pollen, embryos, or semen, or, in some cases, as live plants or animals living in a field, a greenhouse or other installation.

habitat corridors or connections Strips of habitat that connect isolated habitat patches in a landscape transformed by human land use. Connections can be achieved through the conservation of existing habitat or by ecological restoration.

habitat restoration Active modification of the current state of a degraded habitat in order to return it to a former, preferred state.

in situ conservation Maintenance of biodiversity within the ecological and evolutionary dynamics of the ecosystem where it exists.

population A group of individuals with common ancestry that are much more likely to mate with one another than with individuals from another such group.

CONSERVATION BIOLOGY is one of the fastest-growing fields of modern scientific research. It is an applied discipline that integrates principles of natural and social sciences with the objective of achieving the long-term persistence of biodiversity on Earth. This article reviews current trends in conservation biology, using a hierarchical organization to present the principal domains of research. These domains range from the largest level of ecosystems and communities, to the intermediate level of species and populations, and down to the smallest level, that of individuals and genes.
I. DEFINITION

The current global crisis in the loss of biodiversity is the result of the immense success of one species, Homo sapiens, at the expense of the majority of most other species. Concern about the loss of biodiversity arises from spiritual, moral, and aesthetic motives, through economic rationales, to purely selfish reasons. Moreover, the growth of these concerns has fueled the rapid expansion and awareness of conservation biology into academia, government, industry, and other sectors of society worldwide. Although the first conservation biology textbooks and scientific journals date back to the early 1960s, conservation biology’s growth as a scientific discipline during the last two decades has been rapid and global.

Conservation biologists have an enormous agenda. Initially, they sought to understand the impact on biodiversity of the complex interactions between human societies and the natural world. They hope this understanding will provide guidelines for minimizing the negative effects of human actions on the persistence of biodiversity. Increasingly, the discipline has grown to recognize the great extent to which humans depend on biodiversity, and has begun to explore the services that natural ecosystems provide to humans. Ultimately, conservation biology aims to guarantee the persistence of natural landscapes, healthy ecosystems, species, populations, and genes, as well as the complex interactions among them and the dynamic processes that characterize them. This pursuit is necessarily trans-disciplinary, integrating the principles and practices of a variety of disciplines, principally ecology, population genetics, economics, and biogeography, but also increasingly anthropology, psychology, engineering, law, and public policy.

II. ECOSYSTEMS AND COMMUNITIES

The mandate of conservation biology covers a spectrum that runs from concern about the conservation of large, intact, functioning ecosystems, through the maintenance of viable indigenous human and ecological communities, to the preservation of the last surviving individuals of species in zoos and botanical gardens. Arguably, this spectrum may be extended to include the deoxyribonucleic acid (DNA) of currently extant or recently extinct species in museums. This spectrum includes both processes that are complex and only partially understood and ones that are more completely characterized.

The most economical and effective way to conserve biodiversity is through in situ protection that guarantees the persistence of all the processes that characterize biodiversity in a fully functioning ecosystem. Biogeochemical and hydrologic cycles, the flow of energy through food webs, and evolution by natural selection all require a complex web of interacting biological and physical elements. It is in general unrealistic to assume that these organisms and the processes they mediate can be removed from the natural setting and recreated in a zoo, botanical garden, theme park, or gene bank—for long-term persistence, they must be conserved where they stand.

The most popular instruments for in situ conservation are nature reserves and protected areas. According to the World Conservation Union (best known by its acronym, IUCN), a protected area is a section of land and/or sea especially dedicated to the protection and maintenance of biodiversity and of natural and associated cultural resources, which is managed through legal or other effective means. Protected areas come in many kinds and differ mainly in size and the degree of human intervention that is allowed to occur within them. At one extreme are large, strict nature reserves and wilderness areas, where little or no obvious modification of the landscape is permitted. Examples of these include the Bob Marshall wilderness in Montana, United States, and the Parima-Tapirapeco national park in southern Venezuela. At the other extreme are “managed resource” protected areas, which are devoted to the sustainable use of natural ecosystems, such as the networks of game reserves in Botswana or Spain. In all cases, the maintenance of ecological and evolutionary processes is recognized as a key motivation for the creation of such reserves. Ecosystem studies play an increasingly major role in understanding complex ecological systems, and they provide important guidelines for the management of protected areas.

Although there are very few examples of nature reserve systems that were based from their inception on ecological principles, conservation biologists have undertaken many ad hoc analyses of reserve systems that were originally designated for political, aesthetic, and occasionally scientific reasons. For example, Yellowstone National Park in the United States, the world’s first national park, was founded “For the benefit of the People” so that they could appreciate its spectacular and unique geothermal features. By chance, it also happens to contain a nearly complete set of the fauna and flora found in the North American Northern Rockies, and thus is also valuable from the perspective of the conservation biologist.
The first systematic approaches to examining the design of nature reserve systems originated in the application of R. H. MacArthur and E. O. Wilson's "theory of island biogeography" to fragmented landscapes. This theory postulates that the number of species on islands is the result of a dynamic equilibrium between the processes of extinction and immigration, such that as island size increases, extinction rate decreases, and as distance from the mainland increases, immigration rate decreases. As a result, a small island located far away from the mainland will achieve a smaller equilibrium number of species than a large island located near the mainland (Fig. 1). This led Jared Diamond to suggest a series of geometric arrangements that compared preferred to less optimal distributions of patches of protected habitat (Fig. 2).

The creation of a nature reserve may reduce or completely halt habitat conversion within its borders, but it is not likely to prevent land use changes in surrounding areas. As time passes, nature reserves become habitat "islands," and may effectively decrease in size while increasing in degree of isolation from the matrix of habitat that once included them. The theory of island biogeography predicts that the equilibrium number of species in habitat islands should decrease until either the effective size of the nature reserve stops decreasing or the effective distance from sources of immigrants stops increasing. Thus the effectiveness of many parks in the long-term conservation of biodiversity is compromised by processes that erode their boundaries and by those that increase their isolation.

A. The SLOSS Debate

With time, conservation biologists have become aware that the theory of island biogeography might offer additional insights to questions related to the design of nature reserve networks. For example, what sort of actions would the theory prescribe to minimize the rates of species loss? Alternatively, how can a reserve network maximize the number of species preserved? Is there an optimal shape for nature reserves, or can they be arranged in optimal geographic designs?

In attempting to answer some of these questions, conservation biologists initiated a controversy that came to be known as the single large or several small—or SLOSS—debate. The main point of contention was whether the conservation of biodiversity in a particular area was better served by one large reserve or by several small ones. Proponents of the "single large" approach argued that large reserves were able to house large population sizes, thus minimizing the risk of extinction of the species present. In small reserves, high extinction
rates would result in changes of the species assemblages in the reserve and lead to habitats that were very different from those initially targeted for conservation. Proponents of the "several small" approach demonstrated that several small reserves whose combined size equaled that of one single large reserve could include a larger proportion of the species present in a given region. Furthermore, several small reserves generate a set of independent duplicate subpopulations, which decreases the likelihood of extinction of the population as a whole from diseases or catastrophes.

Spatial considerations also play a major role in the size of areas set aside for endangered species. Indeed, the problem of incomplete consideration of spatial scale was at the heart of much of the SLOSS debate. For example, consider a small town that decides to set aside an area of habitat as a nature reserve. At this spatial scale, there is little difference in the species composition of similar-sized patches; however, a relatively small increase in the size of the patch is likely to capture a significantly larger proportion of the local biodiversity. In contrast, as we increase the scale of parks to the state, national, or even semicontinental scale, similar small increases in the area preserved will not significantly affect the species composition in the patch—but similarity between patches of habitat of similar sizes will decline. At this scale, it will be better to purchase, set aside, or otherwise preserve a variety of different patches of habitat. Indeed, at the national level of reserve planning, it is highly unlikely that the full wealth of a nation's biodiversity could be captured within a single reserve, unless that reserve approaches the size of the whole country or continent. Thus, the answer to the SLOSS debate may depend on the spatial scale considered.

Spatial and temporal considerations of scale are also central to our understanding of how ecosystems respond to natural and anthropogenic disturbances. The relationship between the spatial scale of a disturbance, the frequency of a disturbance, and the length of time it takes to recover from a disturbance will determine how readily degraded land can be recovered and used to increase the effective size of a preserved area. In habitats that require disturbances such as fire to maintain all their natural components of biodiversity, such as the New Jersey pine barrens in the eastern United States, or the banksia woodlands in western Australia, successful management will have to mimic the scale and frequency of natural disturbances. Some preliminary synthesis suggests that the recovery time from disturbance scales roughly with the square root of the area disturbed.

B. Gap Analysis

Modern reserve design methods seek to maximize the representation of the components of biodiversity in protected area networks while minimizing the resources needed to do so. For this approach to be effective, comprehensive information is needed on the geographical distribution of species from a wide range of different taxa.

The first step in a reserve design process is to select a series of surrogate measures of biodiversity and assess the existing network of nature reserves for how well it samples this biodiversity. This allows managers and researchers to identify components of biodiversity that are unprotected or not well represented in the current reserve network. The spatial configuration of these missing components defines priorities for future protected area designation. This technique, known as gap analysis, was pioneered by Australian scientists in the mid-1970s and popularized by U.S. governmental agencies in the decades that followed.

Gap analyses typically focus on well-known species or groups of species, and assume that their presence in a nature reserve network can be used as an indicator of the presence of other components of biodiversity as well. The efficacy of gap analyses relies on considering as much information as possible, and updating these results as new information becomes available. Sources of information need not be limited to biological data. Gap analyses may be carried out on the basis of climatic zones, landscape properties, or any subset of features in a geographical region and for any proportion of population size or geographical extent. In fact, given that for most regions of the world knowledge of the detailed distribution and abundance of species is quite limited, the best available data may be a combination of abiotic environmental parameters that are known to correlate with the biological species of interest. These data can then be integrated into a program of biological surveys that constantly adds new information to the database.

Figure 3 depicts the steps in a gap analysis. This process frequently employs "remotely sensed" data obtained from satellites and requires the use of geographical information systems (GIS). Maps of the distribution and abundance of plants and animals, vegetation types, topography, climate, and soil properties (among other possible data sources) are combined to generate overlay maps that quantify landscape heterogeneity. A map that combines existing protected areas is contrasted with the target features overlay...
map and the presence of these attributes in the nature reserve system is assessed. Thus, a final map with the distribution of those attributes that are not included in nature reserves allows for the identification of gaps in protection.

The definition of what constitutes adequate representation is the next challenge. Assessors may be interested in assuring that all species, communities, or habitats considered important are present at least once in a network of protected areas. After achieving this most basic objective, other criteria may be added that minimize costs or future threats, or that maximize the total area of the reserve system by considering alternative configurations of land. It may also be possible to select alternative configurations to reduce conflicts with alternative land uses for the area under consideration.

An important problem that arises is that even if a species is present in a protected area at the time of designation, this does not assure that this species will persist in the long term. Furthermore, the presence of a set of target species does not assure that their parasites, predators, prey, or mutualists are also included, or that other unrelated species will be present as well. The protection of a target species may not provide an adequate measure of the status of ecosystem services, such as the prevention of erosion or the pollination of crops. However, these are not necessarily weaknesses of the gap analysis process itself, but of the way in which it is implemented.

C. Reserve Selection Algorithms

The costs of administering a reserve network increase proportionally to the area protected. Two fundamental concerns of protected area designers are to efficiently allocate available resources, and to minimize the area required to protect as many natural features as possible. The efficiency of a reserve network can be described by the efficiency (E) index

\[ E = 1 - \left( \frac{x}{t} \right), \]

where \( x \) is the number or area of the sites required to achieve a particular representation target and \( t \) is the total number or area of sites to choose from. Thus, a network of protected areas is more efficient as the amount of land required to achieve the target decreases.

Reserve selection algorithms provide systematic means for designing efficient reserve networks. They follow a step-wise process for choosing a subset of sites from a list of proposed sites, such that the efficiency of the chosen subset is maximized. In other words, these algorithms allow for the selection of the minimum amount of land that is required to include all target species (or other attributes) at least once. There are three types of reserve selection algorithms. "Richness-based" algorithms begin the process by selecting the site containing the greatest number of species or attributes and sequentially adding those sites that include the greatest number of new species, until all species have been accounted for. "Rarity-based" algorithms select sites on the basis of rarity scores, favoring sites that are richest in species with restricted ranges. The third type of algorithm depends on linear-programming techniques, which sequentially search the database of available sites for replacement sites that increase the efficiency of the current set.

Although richness-based algorithms are the least efficient, they are the easiest to implement and have the advantage of including the principle of complementarity—that is, at each step of the process, the selected site is most complementary in terms of the features that it adds to the set. Rarity-based algorithms tend to be more efficient than richness-based algorithms, by focusing on species with restricted ranges, more ranging species tend to be included by default, and fewer sites are required to achieve fairly complete levels of coverage. Though linear-programming methods are ultimately the most efficient, they rely heavily on intensive computing and are prohibitive for very large data sets.

Reserve selection algorithms allow for the selection
of efficient reserve networks, but focus only on the adequate representation of a set of species or habitats within the landscape. They do not provide answers to questions related to the size, shape, or number of protected areas that are required to assure viable populations of different species, nor do they guarantee ecosystem functions and services. These more ambitious targets may only be achieved by expanding reserve networks as an interconnected system that allows the interaction of biological and physical components in a continuous landscape.

D. Corridors and Connectivity

Although overexploitation is responsible for the extinction of some species, the majority of species have declined because of the loss and fragmentation of their natural habitats, combined with the introduction of alien competitor, predator, pest, and pathogen species. It is these indirect effects of human development that have had the biggest impact on biodiversity. The crucial, and perhaps the only, way to minimize future impact is to reduce current and future rates of habitat loss and to explore ways of expanding current systems of reserves by the development of large-scale connections across the landscape.

Isolation of nature reserves and protected areas can often disrupt or break migration corridors that are used by seasonal residents of the park. The regular movements will range from the diurnal movement of snakes, birds, and many other organisms between feeding and nesting or brooding sites, to the larger annual migrations of large ungulates. In some instances, it is possible to maintain needed connectivity between habitat fragments by protecting natural corridors or stepping stones of appropriate habitat. In other cases, natural habitat and stopover points may be lost and enhancement of these converted habitats through ecological restoration may be the only option. The interactions between protected areas and developing or developed areas of human use outside them are often subtle and complex. Increased isolation is likely to lead to changes in the flow of nutrients and pollutants into and out of the protected area. Habitat conversion in the surrounding matrix facilitates the invasion of alien species that interfere with the species that the reserve was designed to protect.

Corridors and connections in the landscape have two major functions at the species level. First, they permit regular daily or seasonal movements, helping to ensure that different sub-sections of larger populations may have access to all the resources they require, while also maintaining the potential for all individuals in the population to successfully interbreed. Second, connectivity facilitates the dispersal of animals from their place of birth to their adult home range where they breed. At the regional or landscape level, this latter function is usually the most important, justifying, for example, the achievement of connectivity between mountain ranges in the western United States for the long-term conservation of fully functioning ecosystems.

The restoration of connectivity must occur at many scales. Most attention to date has been paid to local and regional connectivity as a way to redress the various forces threatening small populations. For example, amphibians must be able to move safely across a country lane during their annual migration to a breeding pond. However, many regional and inter-regional corridors have additional goals, such as accommodating the need for grizzly bears (Ursus arctos) to disperse safely between the Canadian Rockies and the Northern Rockies of the United States. In the face of global climate change and other major environmental changes, a substantial system of landscape connections is a major prerequisite for ensuring species persistence.

III. SPECIES AND POPULATIONS

Although conservation at the ecosystem level is the most effective way of conserving biodiversity, a significant proportion of research within the discipline of conservation biology has focused on conservation at the population level. This emphasis on species and populations partly reflects the constraints contained within legal mechanisms to preserve biodiversity; for example, both the Endangered Species Act in the United States and the global IUCN red lists focus on conserving species and/or populations. However, the species also offers a relatively well-defined (although not unambiguous) biological unit of study. Furthermore, the global biodiversity crisis is often most effectively communicated in terms of increasing extinction rates or numbers of threatened species. Therefore, awareness building and fund-raising tend to be more successful when focused on a species, rather than an ecosystem or a particular environmental problem.

It is often easier to raise funds or public support for charismatic species, such as elephants (Loxodonta sp.), chimpanzees (Pan troglodytes), or giant pandas (Ailuropoda melanoleuca), than for a generic habitat, such as a tropical or bamboo forest. More pragmatically, it is often possible to use the requirements of a single species to maintain the viable complexities of entire biological
communities or ecosystems. By focusing on either a target species that occupies a keystone position in an ecosystem or a species that requires large tracts of land to persist (often referred to as an "umbrella" species), many other additional species and their natural habitats are also protected. This is effectively the strength of the "jeopardy amendment" in the U.S. Endangered Species Act, which provides legislation that is designed to conserve habitat in order to ensure the viability of one specific species—as occurred with the northern spotted owl (*Strix occidentalis caurina*) and "old-growth forest" in the north-western United States. Similar species-centered approaches led to the creation of the Coskember Basin jaguar (*Panthera onca*) preserve in Belize, the Arabian oryx (*Oryx leucoryx*) sanctuary in Oman, the Laguna Brava vicuña (*Vicugna vicugna*) reserve in Argentina, the Wolong and Wawlang nature reserves for panda habitat protection in China, and the significant expansion of the El Guácharo national park in Venezuela to assure feeding habitat for a large oilbird (*Steatornis caripensis*) colony.

Population viability analyses are a valuable tool for conservation biologists concerned with studying the long-term persistence of particular species or populations. In the early 1980s, Mark Shaffer introduced the concept of the minimum viable population (MVP), as the minimum size required for an isolated population to have a 99% chance of remaining extant for the following 1000 years in a particular habitat. Shaffer’s work provided a quantitative framework for thinking about species conservation objectives and the timescales within which human actions should be considered. While focusing on the MVP concept, researchers began to realize that the key to conservation at the population level was understanding the impact of different stochastic and deterministic processes that determine extinction risk. By examining the sensitivity of the population to changes in these processes, it was argued that management actions could be taken to increase population persistence. Although there was an unfortunate tendency for many early analyses of viability to emphasize captive breeding as a means of enhancing persistence, this is now perceived as an option of last resort. Moreover, the notion of an MVP was progressively assimilated into a series of techniques that are known as "population viability analysis."

**A. Population Viability Analysis**

The central theme of population ecology is represented in a large canon of work that examines how populations respond to intrinsic and extrinsic factors and how these affect their long-term dynamics and persistence. Although deterministic, stochastic, environmental, demographic, and genetic processes can all have complex and interacting effects on extinction risk, these processes tend to operate in a hierarchical fashion determined by the size of the remnant population. Furthermore, habitat destruction and overexploitation can exacerbate intrinsic risk factors and accelerate population declines. A population viability analysis (PVA) explores the interactions between different known factors and assesses the likelihood that a population will become extinct within a specified time frame and under particular circumstances. PVAs generally require the use of computer simulations and simple mathematical models. The work of M. A. Burgman, S. Ferson, and H. R. Akçakaya (1993) provides the definitive introduction to this area.

PVAs focus on individual species, but they can also examine the potential consequences of different ecological factors, including interactions with other species, rates of habitat conversion, or management interventions that may be taken to improve population persistence. When integrated sensibly with wildlife or endangered species monitoring programs, PVA can be a powerful tool to assist in decision making within an adaptive management framework. The degree of complexity of a PVA will depend on the number of elements that a manager decides to consider, yet more complex models do not necessarily mean more useful results. Understanding the model structure and its sensitivity to varying parameter values may be more important that the modeling process itself. Numerous PVA software packages are available. Unfortunately, the different assumptions hidden within different packages mean that they can produce very different results from the same set of data. Thus, it is crucial to bear in mind that PVAs should not be used to estimate extinction risks per se. Instead, they are most useful for ranking different management options to determine those that increase the likelihood of persistence and that can be implemented at lower costs. They are also invaluable for indicating areas of ignorance about a species and thus focusing future research. The precautionary principle indicates that ignorance should always imply increased protection until a more complete set of demographic data on the species is available.

**B. The Demography of Small and Declining Populations**

The recognition that current losses of biodiversity constitute a global crisis has led to a growing interest in
the problems faced by small populations. By definition, all populations that go extinct first become small. Research in this area has emphasized the positive feedback between a number of mechanisms, each capable of increasing extinction risk as populations further decrease in size. The four principal mechanisms are (1) chance demographic events, (2) environmental stochasticity, (3) inbreeding depression, and (4) Allee effects due principally to the breakdown of the social benefits accrued by living in larger units. Unfortunately, in the 1980s this "cult of the extinction vortex" led to an emphasis on crisis management, while relatively less attention was focused on understanding the processes that initially caused the decline in a previously viable population. In an insightful review of population-level approaches to conservation biology, Graeme Caughey (1994) proposed that the discipline has developed along two relatively independent but complementary lines: the "small population paradigm," which has received more emphasis and addresses problems related to the persistence of small populations, and the "declining population paradigm," which considers the causes of population declines and how to reverse them.

In broad terms, "small population" conservation biologists have focused on finding ways of increasing the size of small populations and achieving the maintenance of a regular flow of individuals between fragmented populations. The application of these principles to the conservation of endangered species in the wild has been limited, but they have played a major role in influencing captive breeding practices, while providing some insights into the theory of how nature reserve systems might be designed.

In contrast, an increasing number of conservation biologists have focused their research on identifying the causes that lead to declines in species abundance or the contraction of the range of a species. The primary forces operating here are habitat loss and fragmentation, although threats posed by alien species, overexploitation, and pollution are also important. By examining populations at an earlier stage in their decline, it is hoped that strategies may be found for halting and reversing the decline. This paradigm is concerned with prescribing early, preventative actions that require minimal intervention that may be taken before a population becomes seriously endangered. At present, this approach has not succeeded in generating elegant theoretical principles to guide management decisions. This is mainly because population declines can be caused by numerous factors operating simultaneously on a community of populations. Not surprisingly, these effects are difficult to quantify. The key to future success in this area is carefully designed experimental and monitoring programs that systematically seek causes for population declines.

The most likely future for these two paradigms is eventual fusion. Clearly, both approaches are useful and contribute complementary insights to the advancement of the discipline.

IV. INDIVIDUALS AND GENES

All forms of life carry genetic information. In the long term, the capacity of any species to adapt to anthropogenic and nonanthropogenic changes in the environment will be related to the existence of genetic variation. In situ conservation programs are almost without exception the best option for conserving the genetic diversity that is a component of biodiversity. The protection of an ecosystem and its ecological processes provides the necessary conditions for evolution and natural selection to proceed. Additionally, protecting species in the wild is much more cost-effective than maintaining them in captivity. In the cases of some species, however, such as varieties of domesticated plants and animals and highly endangered species, ex situ conservation technologies are required. In these cases, the maintenance of genetic variation becomes a central topic of concern, especially if an endangered species is taken into captivity for future reintroduction into the wild.

A. Measuring Genetic Variability

Genetic information is encoded in genes, which are mainly composed of long chains of deoxyribonucleic acid, or DNA. Genes aggregate in the cells of every organism to form chromosomes. Variation in genetic information can be measured at any step of the biological process that transforms the sequence encoded in genes into its protein products. This information can then be used to quantify differences between organisms, populations, species, or any other taxa.

Methods for measuring genetic diversity focus on the variation of "markers," much in the way that gap analysis uses species or topographic variables to characterize the degree of spatial heterogeneity in a landscape. These markers are assumed to correlate in a broad sense with the genome as a whole; the more variable a population is at marker loci, the more variable it is overall. In the past, scientists were limited to examining indirect or "phenotypic" measures of genetic diversity through markers such as variable external morphology or coloration, intracellular concentration of secondary com-
pounds such as phenolics, cell pigments, or hydrocarbons, and variable protein products called "allozymes," which may be visualized using electrophoresis. In recent years, however, biologists have developed techniques for examining the DNA itself. This permits the use of markers such as the presence or frequency of specific base-pair sequences in DNA molecules. DNA fingerprinting, for example, allows the characterization of differences between individuals in a population on the basis of patterns in the variation of minisatelites, or relatively short (10–50 base-pairs long) repeat sequences that are distributed in tandem arrays throughout the nuclear genome. Variation in organelle DNA (such as mitochondria or chloroplasts) may also be examined. Different versions of these techniques are used to assay genetic differences within and between populations, species, or other taxonomic levels.

B. The Role of Zoos, Aquaria, Botanical Gardens, and Gene Banks

In extreme cases, such as when the last few remaining individuals of a species are under an imminent threat in the wild, or when a wild variety of a particular crop is being rapidly replaced by modern monocultures, the maintenance of populations in captivity may be the only short-term action to prevent their extinction. Zoos have played an important role in preventing the extinction of the Arabian oryx, the peregrine falcon (Falco peregrinus), and the Guam rail (Rallus owstoni), but in all cases the captive facilities were used to buy time while the agents of decline were being dealt with. Gene banks may be the last resource for hundreds of threatened varieties of domestic animals and plants. The Royal Botanical Gardens in Kew (United Kingdom) hold about 38,000 species in their collection; Brazil and Colombia are the only countries in the world that have more species within their territories.

The primary functions of these captive facilities are not to save endangered species, however, but to promote education and research. Conservation objectives are clearly most effectively and efficiently achieved by in situ rather than by ex situ approaches. Nonetheless, zoos, aquaria, and botanical gardens provide visitors with an opportunity for close interaction with a wide variety of plants and animals that are otherwise out of reach for most people. Zoos attract millions of people worldwide and are ideal facilities for educating them about the challenges and opportunities created by the present biodiversity crisis. They are also potentially important facilities for raising funds that can then be invested in conserving habitat in the regions where the animals and plants on display originally came from.

V. FUTURE CHALLENGES

The future of conservation biology lies in teamwork. The complexity of issues surrounding the conservation of biodiversity requires a trans-disciplinary approach that integrates knowledge and expertise from the different disciplines of natural and social sciences. Conservation biology will also become increasingly cross-institutional as governments, academia, and nongovernmental organizations work together on the same issues but with complementary approaches and points of view.

An excellent model is provided by the Intergovernmental Panel on Climate Change (IPCC), a global organization dedicated to assessing the available scientific, technical, and socio-economic information in the field of climate change. The IPCC combines experts from around the world into three working groups that focus on (1) the climate system, (2) impact and response options, and (3) economic and social dimensions of climate change. Although the process is slow and long, the IPCC has already produced two assessment reports and is expected to produce a third report by the year 2000. The topics covered by the IPCC reports range from the state of the art of climate modeling to benefit/cost analyses of choosing between different mitigation options. The process has not been free of flaws and criticisms, but has allowed the identification of issues on which there is consensus. Likewise, it has allowed the identification of topics where agreement has not been reached, due to either a lack of scientific data or uncertainty about the nature of the processes underlying global change.

One requirement of conservation biology in the future will be the design of biodiversity management programs and decision-making mechanisms that incorporate, rather than reject, uncertainty. Although our understanding of ecosystem function and human dependence on nature's services has greatly increased, it is still far from perfect. As in the case of global climate change, we might never understand every detail of the numerous processes that influence biodiversity. This cannot remain an excuse for inaction, however. The adoption of an adaptive management framework that embraces the precautionary principle will allow us to use the best science available to decide on issues of biodiversity stewardship. One of the most difficult issues may be the revision of current resource consump-
tion patterns, particularly in the developed world, and the design of new paradigms and technologies to attend to the needs of a growing global population. After particular actions are taken, their success may be monitored and adjusted in response to changing conditions, which arise because of either abrupt, unanticipated events or slowly evolving processes. With time, our understanding of biodiversity and our management decisions will both improve, generating creative solutions to reduce the impact of humans on Earth's systems while assuring the quality of life of all those who inhabit the planet.

See Also the Following Articles

CONSERVATION MOVEMENT, HISTORICAL • Ex Situ, in Situ CONSERVATION • ISLAND BIOGEOGRAPHY • KEYSTONE SPECIES • NATURAL RESERVES AND PRESERVES • POPULATION VIABILITY ANALYSIS (PVA) • RESTORATION OF BIODIVERSITY

Bibliography


CONSERVATION EFFORTS, CONTEMPORARY

Kristiina A. Vogt, Oswald J. Schmitz, Karen H. Beard, Jennifer L. O'Hara, and Michael G. Booth
Yale University

I. Introduction: New Concepts in Conservation
II. Analysis of Evolving Thinking in Conservation Biology
III. Tools for Managing and Conserving Species
IV. Future Directions and Challenges in Conservation

GLOSSARY

ecological integrity The balanced state of an ecosystem under normal environmental conditions, including the capacity of an ecosystem to absorb disruption and to recover from the disruption.
ecosystem engineers Animals (e.g., beavers) that modify an ecosystem by physically changing it (e.g., building dams) or by being present in such a great abundance (e.g., large grazing animals) that they have a disproportionately large impact on ecosystem functions.
ecosystem management Approach to management that attempts to maintain high ecosystem integrity while providing the services, uses, values, and products from that system for the long term. It explicitly integrates social, economic, and natural system sustainability.
invasive species Native and exotic species that are capable of displacing indigenous species or spreading into habitats where they were not common previously.
keystone species Species whose presence and abundance controls the integrity of a community or ecosystem and allows that system to persist within its natural range of environmental conditions.
legacy ecosystems Ecosystems that carry an intact “memory” of a change in some physical, chemical, or structural attribute or a change in species composition that modifies the resistance and resilience characteristics of the ecosystem.
resilience Rate at which an ecosystem is able to recover to conditions similar to those that existed prior to the imposition of a disturbance.
resistance Degree to which an ecosystem moves away from or is able to maintain current structure and function when faced with a disturbance.
restoration Re-establishment of similar structures and functions of an ecosystem or parts of an ecosystem that are no longer present because of past land uses or disturbances.
sustainability State that defines the biogeophysical, economic, social, cultural, and political thresholds in between which it is acceptable to continue to use and obtain services or products from a given piece of land. This definition requires all five factors to be considered when determining whether future generations will be able to acquire the same natural resources as the current generation.
sustainable development Meeting human needs with-
out damaging the ecosystems that produces these resources and at the same time providing equitable distribution and access to these resources around the world.

**system functions** Processes that occur in an ecosystem that are typically measured as changes in the carbon or nutrient cycles regulated by decomposers, consumers, or primary producers.

**umbrella species** Species that have either large habitat needs or have other requirements whose conservation results in many other species being conserved at the ecosystem or landscape level.

---

**THIS ARTICLE PRESENTS MANY OF THE ISSUES AND CHALLENGES** currently being faced in converting theories from the natural sciences into practical applications for conservation. As part of this transition in conservation, two philosophical changes have created the most problems in making conservation practical. The first is a movement from a species to an ecosystem focus. The second change has been the recognition that humans influence the success of conservation projects and have to be factored into the conservation formula. The human influences have produced a dual approaches to conservation that varies between developed and developing countries. In developing countries, sustainable development is being linked to conservation. In developed countries, the conservation focus has been on restoration of species and habitats and in developing tools to eradicate invasive species. Four tools being used for managing and conserving species will be discussed: models, landscape patterns, reserve design, and economic evaluations.

I. INTRODUCTION: NEW CONCEPTS IN CONSERVATION

The discipline of conservation biology arose from the need to apply science to the protection of species that are threatened or on the brink of extinction. In its early years, biologists and ecologists had to worry about refining the science that would be brought to bear on the cause of protecting endangered species on particular tracts of land. This resulted in the practice of conservation being very species sensitive. Nowadays, conservation biology has to resolve many conflicting demands that are simultaneously being placed on ecosystems services that are provided by particular tracts of land, as discussed in the work of J. F. Franklin, J. C. Gordon, N. Myers, G. Daily, and others. Saving endangered species is typically only one component of a larger task to protect the entire biotic communities (biodiversity) that make up whole ecosystems. This has placed an emphasis on putting into practice many of the holistically based philosophical ideas that were previously articulated by John Muir, Aldo Leopold, and other visionary thinkers.

The mounting worldwide conflicts over the conservation of biodiversity and other uses of the land base make it imperative to develop and test new models of conservation. For example, many conservation efforts in the tropics have reinforced the importance of factoring in humans as integral elements of conservation for these efforts to be successful. As documented by the United Nations and presented by N. Myers, fully one-half of the world’s biodiversity is locked up in tropical ecosystems where conflicts over resource uses are particularly intense. At the same time, human population size is at such a level in the tropics that the attendant population growth rates in these areas are double the world average. Such prodigious growth typically leads to a high level of poverty that makes conflicts over resource use and their protection particularly polarizing for resource managers and policymakers who are unable to satisfy all the stakeholders simultaneously. In these circumstances, even the best scientific knowledge about the ecology of groups of species will not provide the needed insight to protect biodiversity, since human demands on the natural resources will constrain the application of science to conservation. Therefore, the traditional model of conservation, whereby scientists continually refine and apply knowledge to the cause of species preservation, will be insufficient to meet the challenges of the future.

Additionally, the approaches to conservation have been evolving as new knowledge develops in population, community, and ecosystem and landscape ecology, as well as in social ecology. This knowledge has made us realize that the use of species demographic data is generally inadequate by itself for gauging conservation success and therefore for maintaining a species in the landscape. Species demographic data are effective in conservation when the species are keystone or umbrella species, or when they perform ecosystem engineering functions. Yet even then, knowledge of changes in population dynamics of a species does not necessarily provide a mechanistic understanding of the role that species play in ecosystems. Conservation has been moving towards holistic analyses of systems (i.e., ecosystem management) in which factors other than species biol-
ogy dictate the approaches used to maintain species in their habitats. L. Harris, M. L. Hunter, M. Mangel, D. Simberloff, W. W. Weeks, and many others have documented this shift toward a holistic approach to conservation. Holistic analysis also recognizes that humans are an integral part of the system and cannot be considered as external factors in conservation projects.

It is clear that in order for conservation projects to be successful, sustainable economies must be created at the local level, as demonstrated by G. E. Machlis, J. McNeely, M. Wells, K. Brandon, and others. This recognition has resulted in the explicit linking of conservation with international sustainable development projects, especially in the humid tropics. Projects that meld these dual objectives are commonly called Integrated Conservation and Development Projects (ICDPS). These projects vary in scale and scope but typically include one or more of the following topics as a central theme: biosphere reserves, eco-tourism ventures, non-timber forest product harvesting, and regional land use plans. Such projects are attempting to address the impacts of higher human population growth rates that are resulting in the development of economies that demand increasingly greater use of natural resources (e.g., tropical deforestation) to produce viable communities. The ultimate result of all of these changes is that humans are leaving a much larger "ecological footprint" on the land base and making it more difficult to practice conservation.

Conservation efforts also have to accept that species survival is heavily influenced by legacy ecosystems and landscapes. The end result of a legacy system is that it responds to future natural and human perturbations differently than do systems without these legacies, as shown by J. F. Franklin, J. J. Magnuson, T. A. Spies, P. M. Vitousek, K. A. Vogt, D. J. Vogt, and others. This is a problem for conservation since management and policy decisions may have to be based on science collected in non-legacy systems because that is the information that is available. Some factors contributing to the formation of legacy systems are:

- Past land uses and management practices that have fragmented the landscape and reduced habitat area, as documented by T. W. Clark, J. F. Franklin, R. T. T. Forman, and others.
- Introduction of invasive plant species that fix nitrogen, thus changing soil resource availability, as shown by P. M. Vitousek and Walker in Hawaii.
- Existence of disturbances (i.e., hurricanes) that affect succession occurring at longer temporal frequencies than a manager's typical time frame (recently documented by many researchers working in Puerto Rico).

The insufficiency of habitat area for managing threatened grizzly bears has been well documented by T. W. Clark, S. C. Minta, and others for the Greater Yellowstone Ecosystem. In this case, the logical approach would be to restore a sufficient area of suitable habitat for the grizzly bear. However, it is insufficient to deal with just the habitat to ensure the survival of grizzly bears. Indirectly, grizzly bears are also threatened by an exotic fungus that is causing the decline of Pitus albicaulis, whose seeds are an important food source for grizzlies. This underscores the need to recognize that conservation efforts are complicated by other factors regulating the functioning of a system that are not considered because they do not appear to have a direct bearing on a species.

Recently ecologists have come to understand the importance of disturbances that occur at long time frequencies (e.g., hurricanes) and how they affect the structure and drive the functioning of a system. Knowledge of these temporal dynamics is relevant for conservation, since the design of conservation efforts based only on consideration of intervals between disturbances can endanger an already threatened species. For example, when Hurricane Hugo swept over the tropical forests on Puerto Rico, the already small population of Puerto Rican parrots was dramatically reduced. These hurricanes are a reoccurring phenomenon there and have been recorded at a frequency of 21 years. The effect of these hurricanes on determining the structure and functioning of these forest systems during the intervals between hurricanes has been shown through the work of N. Brokaw, A. Covich, A. Lugo, F. N. Scatena, K. A. Vogt, D. Vogt, and others. Consequently, if ecosystem processes are studied only during the inter-disturbance intervals, we may overlook the important causal drivers for that system.

Five major philosophical changes are commonly identified when discussing why there is a need to adapt the old models and at times to find new models for conservation:

- First, there is the recognition that a focus only on species is inadequate to conserve species within habitats in perpetuity.
- Second, there is the realization that the human dimensional aspects of ecosystems cannot be ignored if conservation efforts are to succeed.
- Third, ecosystems are dynamic, so the maintenance of a particular type of habitat in the landscape will
require that several examples of that habitat exist elsewhere within that landscape but at different successional/developmental stages. This approach will require a landscape that is sufficiently large to include these stages. If the appropriate conservation area has been selected, when a particular habitat is lost at one location, its replacement will be developing at the same time in another part of the landscape.

- Fourth, conservation approaches will have to incorporate the tools of restoration ecology to reestablish the structural and functional aspects of native habitats required to maintain species. The restoration of damaged habitats may become a more central part of conservation practice, since centuries of human modification of the natural environment have resulted in inadequate habitat area or have sufficiently modified ecosystems. Because grizzly bear management in the Greater Yellowstone Ecosystem cannot be expanded beyond its present boundaries, there may be a need to consider restoration of habitats within the GYE if these habitats degrade due to the proliferation of invasive species. As the presence of invasive species increases within our landscapes, conservation biologists will have to consider using restoration tools as part of their management approach, since opportunistic invasive species frequently threaten the survival of native species.

- Fifth, disturbances (e.g., hurricanes and, droughts or floods associated with El Niño) at long time-scales can eliminate or modify the habitat set aside in nature reserves so that the species being conserved are unable to survive in that diminished or degraded habitat.

These five changes in philosophy are causing a reassessment of how conservation biologists approach the practice of conservation. These changes are partially being driven by new scientific understandings that are beginning to filter into the mind-set of conservation biologists. In this article, we present a case for the need to increase the acceptance of these concepts by conservation biologists so that conservation practices can become more effective. These five major philosophical shifts have already resulted in several changes in the practices being used to implement conservation efforts (the most common changes are summarized in Table I). Two of these shifts will be discussed in greater detail in the next section. These two points have been selected for greater analysis because they have not been as extensively discussed in the literature, as have the other points (the reader is encouraged to consult the Bibliography for further readings on these topics). Section III will present a brief discussion of four important tools of conservation in the context of these philosophical changes.

II. ANALYSIS OF EVOLVING THINKING IN CONSERVATION BIOLOGY

A. Species to Ecosystem Shift

Because of the historical emphasis on saving threatened and endangered species, conservation biology has had a largely singular focus on protecting those species and their habitats. This focus was founded on a simple ethic that humankind has a moral obligation to protect all living things on earth. This species-centric approach seemed to be relevant in the early phase of conservation because anthropogenic changes to the land base were geographically localized and, by today's standards, comparatively benign. However, current approaches in conservation are strongly based on the ecosystem, landscape, and/or human system. This shift resulted because of the failure of many conservation efforts that were based on the protection of species. Yet it is important to recognize that the single-species approach should not be eliminated from the suite of tools available in conservation in the push to assess systems more holistically. It would benefit conservation if a framework was developed that helped managers identify which systems would be most likely to have single or multiple species playing critical ecosystem roles. A brief discussion of the types of single-species focus that are useful in conservation biology follows.

In some ecosystems, the single-species approach will continue to be a powerful tool for conservation biologists. For example, the single-species approach may still be appropriate as a primary conservation tool when a single species functions as a keystone species, as an ecosystem engineer, or as an umbrella species or coarse filters (discussed by W. J. Bond, C. G. Jones, R. T. Paine, M. Soulé, D. Wilcove, and others). In some ecosystems, keystone species can play an essential role in regulating ecosystem structure, ecosystem function, and/or the maintenance of community diversity. In the event that keystone species are lost from the system, the result may be a disproportionately larger effect on some property of the system so that certain species can no longer be maintained in the landscape. However, the concept of keystones may be difficult to apply in conservation because these species are not always obvious or we have not developed a clear mechanism for identifying them in the system, and some are only keystones part of the
<table>
<thead>
<tr>
<th>Conservation focus</th>
<th>Previous practice</th>
<th>New practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservation focus</td>
<td>Single species based approach. Ex-situ conservation for single species (e.g., zoos, expensive reintroduction programs, captive breeding programs).</td>
<td>Umbrella species approach. Ecosystem based. In-situ conservation model preferred. Restoration of species and ecosystems, elimination of invasive species explicitly linked to conservation.</td>
</tr>
<tr>
<td>Integration of humans and human impacts on ecosystems in conservation</td>
<td>Protected areas management or reserves could be designed to exclude humans. Conservation practices strongly based on managing spatial distribution of ecosystems or habitat scales in landscapes since the major human impact has been fragmentation and loss of ecosystems.</td>
<td>Conservation is human ecosystem based—explicit integration of people as primary drivers of the success of protected areas or reserve management and a need for people to access conservation areas. Recognition of the temporal scales of human land-use activities and how they produce legacy ecosystems and landscapes that control the success of protected areas or reserves needed in conservation (states that few non-human-impacted ecosystems exist).</td>
</tr>
<tr>
<td>Human values</td>
<td>Species values dominant and used to drive conservation practices.</td>
<td>Recognition that the species is one of the multiple values existing in a system and a species focus may miss the &quot;real&quot; driving variables determining whether an established protected area or reserve will be successful. Incorporation of and determining trade-offs of all values desired out of a land as part of conservation.</td>
</tr>
<tr>
<td>Disturbances</td>
<td>Incorporate short-frequency disturbances. Conservation plan attempting to exclude or limit human extraction of resources from nature reserves. Humans classified as disturbance agents who are negative to conservation efforts.</td>
<td>Inclusion of decade-scale disturbances (e.g., hurricanes) for species conservation. Conservation plan includes human extraction of resources from nature reserves. Explicitly integrates invasive species as a major disturbance factor affecting conservation success.</td>
</tr>
<tr>
<td>Spatial scale of analysis</td>
<td>Reserve design and size varied with territory needs of each species. Landscape matrix analysis recognizing influences outside of habitat fragment will impact conservation. Identification of edge and interior regions of reserves and their influence on species composition. Assumed ability to transfer information from plot to produce regional maps of endangered species. Use of corridors and buffer zones to link habitat fragments and reserve networks. Reserve design focus—single large or several small reserves, reserve size, where to place reserves.</td>
<td>Reserve design and size based on the ecosystem and landscape and not on species territory needs. Recognition that linking habitat fragments with corridors is inadequate. Knowledge that human land-use legacies and disturbance impacts modify quality of edge and interior habitats. Inability to transfer plot data to produce regional maps of endangered species unless characterizes ecosystems into their diverse types within a vegetative community.</td>
</tr>
<tr>
<td>Models</td>
<td>Small-scale, data-intensive species and community models.</td>
<td>Less data-intensive, large-scale biodiversity models and explicit linking to GIS. Ecosystem models geared to predicting human land-use activities and the sustainability of the ecosystems needed for conservation.</td>
</tr>
<tr>
<td>Economic evaluation</td>
<td>Development of nonmarket values for species.</td>
<td>Need to address all the values in each environment. Recognition that conservation efforts should not be driven just by economic analyses</td>
</tr>
</tbody>
</table>
time and in certain habitat conditions. Once keystone species have been identified, the ability to use the single-species approach simplifies conservation efforts. W. G. Conway suggested that the existence of keystones in an ecosystem would facilitate integrating conservation with restoration efforts. When a single species has not been identified that provides a key structural or functional role, species become poor indicators of whether a system is going to degrade or whether the habitat can be maintained for the species needing to be conserved.

The use of a single-species focus in conservation is greatly facilitated when species function as ecosystem engineers. In any ecosystem, the existence of ecosystem engineers is easy to recognize because they modify some structural part of the ecosystem that can be readily detected. They do not have to be present in large numbers for the impact to be registered. In many examples, a few individuals will have a disproportionately large impact on directly changing plant successional processes and indirectly modifying carbon and nutrient cycles (as has been documented for beavers and moose by R. J. Naiman, J. Pastor, and others). In addition, ecosystem engineers may play keystone roles at particular stages of the development of a vegetative community (i.e., their impact would not be measurable at all times).

In conservation biology, an important concept that has been discussed for more than a decade is the identification of single species that can function as an umbrella species or as a coarse filter (see P. F. Brussard, T. W. Clark, C. R. Tracy, and others). Umbrella species are those species that have large habitat needs (e.g., spotted owl, red-cockaded woodpecker, and grizzly bear) or have other requirements such that when the system is managed to conserve that species, many other species will consequently also be conserved. The single-species approach has been most often justified through the concept of umbrella species.

Unfortunately, many ecosystems do not have a single species that can be monitored to provide an indication of how that system is functioning and whether a conservation project will be successful. Based on our current scientific understanding, many species are merely passive inhabitants of ecosystem in which they live. Instead, entire communities of species may play integral roles in ecosystem function (several studies by J. J. Ewel, J. H. Lawton, S. McNaughton, S. Naem, D. Tilman, and others have supported this idea). Biodiversity may then be a critical determinant of ecosystem productivity and stability—including the tendency to resist anthropogenic impacts and the ability to recover from impacts (i.e., resilience). However, the relationships between biodiversity and ecosystem productivity and/or stability cannot be generalized to all ecosystems. Strong relationships between these variables have been found in certain ecosystems (e.g., grasslands), but biodiversity cannot be automatically used as an indicator of ecosystem integrity (see K. H. Johnson, O. J. Schmitz, K. A. Vogt, and D. J. Vogt). For example, the biodiversity at the species level is a poor indicator of the resistance and resilience characteristics of woody perennial systems to disturbances since the response of the system occurs at levels other than species diversity (e.g., genetic diversity within a species).

The growing realization that biodiversity may be inextricably linked to ecosystem function has given rise to a more holistic, system-centric perspective for conservation. This new approach was codified by M. Mangel and others in 1996, when they defined a set of principles to which conservationists must adhere if we wish to successfully conserve wild living beings. These principles coincide fully with those articulated by J. K. Berry, K. A. Vogt, and others for successful ecosystem management. The principles of ecosystem management are:

1. focus on the sustainability of ecosystems, not on the output of products;
2. adopt a holistic understanding of the way all the parts are linked together in an ecosystem and the feedbacks among those linkages;
3. incorporate a long-term perspective and examine issues at a scale relevant to the functioning of the ecosystem; and
4. recognize that human values shape ecosystem structure and function in myriad ways that can constrain, promote, or reduce sustainability.

Because the primary goal of conservation is to sustain the habitat or ecosystem of species (as articulated by R. B. Primack), conservation biologists have already espoused the first principle of ecosystem management as articulated here. Most of the tools developed in conservation biology address this first principle but not the other three principles. Similar to what is still happening in ecosystem management (as discussed by J. C. Gordon, J. F. Franklin, K. A. Vogt, and others), conservation biologists are currently assessing the practicality of the different tools to implement the other three principles in conservation.

Out of the four principles listed here, the incorporation of people and their values into a holistic framework of analysis that is not solely human-oriented is one of the greatest challenges for both ecosystem management and conservation. Machlis has stated that conservation currently has no trained capacity to deal with the human
dimensions of biodiversity conservation. Most of the dialogue related to integrating the social sciences into natural resource management is presently at the theoretical level (see J. K. Berry and K. A. Vogt).

Conservation is now facing challenges similar to those previously experienced by natural resource managers attempting to implement ecosystem management under conditions of high scientific uncertainty. For example, natural resource management shifted from a resource-based emphasis (e.g., forest stands, fish stocks) to a more holistic, ecosystem-based approach that better enfranchises all stakeholders of a particular land or resource base. In the face of such a shift, managers were forced to embrace new management models (i.e., adaptive management) to deal with large-scale management problems under a scenario of high scientific uncertainty about the response of ecosystems to human activities. Because our scientific knowledge base is still fairly limited, few guidelines exist to help resolve conflicts arising from the need to determine the trade-offs between different natural resource uses. These conflicts are especially difficult to resolve since conservation biologists and ecosystem managers have to deal with resource scarcities.

The current view of J. C. Gordon is that ecosystem and adaptive management are both approaches that are attempting to respond to resource scarcities. Ecosystem management is attempting to respond to a scarcity of land. This land scarcity means that the exclusive uses of any given land area for one purpose is no longer possible because the production capacities of ecosystems are finite. This contrasts with adaptive management, in which managers are responding to scarcity of scientific information and knowledge. This scarcity makes it difficult to develop consensus on which tools are most appropriate to use for resolving the conflicts over resource uses. In a similar vein, conservation biologists are being forced to respond to scarcity and the resulting conflicts that can arise over the use of finite resources. However, conservation has yet to adopt a model that can adequately handle conflicts resulting from the scarcity of resources and simultaneously achieve natural and social system sustainability.

B. Direct and Indirect Influences of Humans on Conservation Efforts

Humans are strong influences on whether conservation projects are successful. These human impacts can be understood by analyzing the different approaches taken by conservation organizations to address conservation issues in developed and developing countries. In developing countries, humans are dependent on natural resource extraction from the environment in which they inhabit. A major deterrent to conservation has been the inability of these governments to satisfy all the resource uses, services, and values that are expected from the same land base by local, regional, national, and international stakeholders. Under these circumstances, the scarcity of land area and resources means that conflicts arise over the uses of natural resources that typically polarize the people inhabiting the environment against conservation organizations articulating a need to establish reserves to save species. These conflicts often occur in environments where the land base constraints limit the options that are available to satisfy all the needs of all stakeholders. Some of the direct impacts of humans on the environment appear to be easier to manage, since alternative solutions are possible through substitution of products, technological breakthroughs that make resource extraction more efficient, shifting to the use of different resources, alternative employment opportunities, and so on. The issues that are relevant for developing countries will be further discussed in Section II.B.1.

In developed countries, conservation organizations generally do not have to deal with indigenous communities who are living and surviving off the natural resources in areas designated for conservation. However, this does not mean that the problems identified in developing countries are not also found in the developed countries, rather they have not driven the approaches being used in conservation. In developed countries, greater emphasis has been placed on developing the tools to protect and restore habitats and species, and in eradicating invasive species that are threatening indigenous species composition and altering succession. The spread of invasive species and the concomitant loss of indigenous species is a problem driven by international trade and the globalization of the world. The changing transportation networks and the development of a global economy are contributing to an accelerated rate of the spread of invasive species that has not been documented prior to this time. In addition, the spread of invasive species frequently occurs in ecosystems that have been significantly modified by human land use activities and few good control mechanisms have been produced. According to P. M. Vitousek, humans have been modifying their environments at unprecedented rates during the last several decades. Because of the role that humans are playing in spreading species and in altering the health of ecosystems, one of the greatest challenges to the successful implementation of conser-
Conservation projects may be our ability to control the spread of invasive species. The impact of invasive species on conservation efforts in developed countries is discussed in Section II.B.2.

Some of the direct and indirect impacts of humans are much more difficult to manage in conservation since we are just beginning to understand these effects. Many of these effects occur at longer temporal and spatial scales, making it more difficult to explicitly link the cause-and-effect relationships so that tools could be identified to manage ecosystems and species in ecosystems. Few good solutions have been produced for even assessing and managing these constraints, and so useful tools are not available to transfer to conservation at this time. The impacts of humans on species loss and changing ecosystem structure and function will be further discussed in the following sections.

1. Humans and Sustainable Development as an Integral Part of the Conservation Formula

Within the last two decades, there has been an increasing understanding that conservation projects cannot be successful when implemented in isolation from neighboring human populations. Maintaining nature reserves has undergone a major shift from trying to keep people out of reserves to allowing some economic development by local communities as an integral part of conservation strategies (discussed in the works of K. Brandon, P. H. Buttel, R. J. A. Goodland, J. McNeely, M. L. Shaffer, M. Wells, and P. C. West). The response of the international community in dealing with this issue has been to design projects that explicitly link international aid to local communities with the establishment of nature reserves to conserve species. The goal of these projects has been to improve the livelihood of people living close to the natural resources so that they would not be dependent on extracting species or resources from the protected area. The strategy behind this approach was to find a balance between development and conservation by providing alternative approaches for people to survive and at the same time provide zones of lower human impact for conservation purposes.

This new model of conservation attempts to reconcile human pressures on resources in areas that are rich in biodiversity and therefore have been identified as high priority areas for conservation. This represents a major shift from the traditional conservation model, whose main focus was to lock up resources by eliminating human presence in nature reserves without considering its impact on the survival of the local communities excluded from these areas. Conservationists now recognize that successful application of their models will require explicit recognition and alleviation of human dependence on natural resources at the local level. It has become quite apparent that it was insufficient to set up borders and assume that people with few options would stay out of these areas.

In fact, humans have had an integral hand in shaping the very land that we now wish to conserve in many tropical countries. It is important to recognize that present forest conditions may be a relict of past management and land use activities of the indigenous and rural communities, so that eliminating them from the landscape may actually change the forest composition to one that may not be acceptable or desirable for conservation purposes. In wet tropical forests, humans have strongly influenced the resident biodiversity; indeed, the "virgin forests" and exceptionally rich biodiversity that are the focus of conservation efforts in the Amazon River basin are in part relics of old agricultural practices and management (see W. Balee, M. Pinedo-Vasquez, D. A. Posey, and H. Raffles).

The importance of the human dimension was especially reinforced for conservationists working in the tropics, where desperate people continued to harvest resources from forests set aside as nature reserves for conservation purposes (see works of K. Brandon, M. Dove, J. McNeely, and M. Wells). Many of the conservation organizations funding the establishment of protected areas quickly recognized that nature reserves could not be maintained if they were designed to exclude local communities without providing alternative sources of income. Most of the current projects designed to conserve biodiversity are encouraging local communities to use areas outside of reserves or at least along the buffer zones of reserves. This strategy resembles the UNESCO Man and the Biosphere (MAB) design for reserves, which is based on having concentric zones representing different intensities of human use—with the center or core area designated for conservation without human presence. However, shifting or confining people's use of a forest to particular zones without their cooperation is impossible in most developing countries. Often, land tenure or ownership does not exist for these local communities. Furthermore, these communities are accustomed to adjusting their resource acquisition efforts within the landscape both seasonally and by migrating to different locations.

As part of the new conservation approach, ecotourism and the harvesting of nontimber forest products have often been presented as two viable alternatives for local communities to generate income (see J. L. O'Hara). Yet so far these options have not proven to be effective.
in meeting the dual goals of biodiversity conservation and rural economic development (see M. Kremen, A. M. Merenlender, and D. D. Murphy).

Thus social factors can play a major role in determining the effectiveness of reserves, that is, the pressures of human activity and population growth will probably be the major determinants of whether a reserve will successfully conserve a given species (see A. Dobson). Sometimes the scientific knowledge of how much area is needed by a species cannot be implemented in the face of these social pressures. In these cases, there is nothing wrong with using the science to determine the necessary size of the conservation reserve, but managers must keep in mind that social and economic variables are exerting more control on whether the reserve will be able to protect its resident biodiversity.

2. Humans and Invasive Species

Within the last decade, many conservation organizations have implemented programs to control or eradicate invasive species, especially those that are exotic. This new objective is driven by the many field examples of invasive species disrupting or altering native ecosystem functions and structures, as well as their biodiversity. The significant economic costs of eradicating, or attempting to eradicate, invasive species have highlighted the difficulty of managing such species. One of the biggest challenges of future conservation efforts will be the ability to control or minimize the presence of invasive species, because they will hamper or limit the options that are available for the conservation of indigenous species and their intact ecosystems. The spread of invasive species will probably become more critical in the future as humans create the conditions or change disturbance cycles sufficiently to give invasive species the opportunity to dominate invaded habitats.

Humans have played a dominant role in the spread of invasive species around the world. Humans function as vectors of invasive species and/or alter the landscape in a way that enables invasive species to become better competitors than indigenous species. Many weeds, insects, and other animal pests are transported accidentally with produce and goods; food crops and horticultural varieties are often introduced to new environments intentionally for agriculture and landscaping. Likewise, there is a long history of humans introducing game animals to new territories. Once invasive species are introduced into a landscape, many of our land use activities have been implicated in facilitating their spread. There are several possible mechanisms by which the rate of invasive species spread is enhanced: (1) the combined effect of land use change and an invasive species may alter disturbance cycles to which indigenous species are adapted and (2) invasive species disrupt established abiotic resource availabilities within an ecosystem that results in the ecosystem conditions becoming more favorable for exotic species (see Office of Technology Assessment, 1993; as well as the work of J. Martinelli, J. Randal, P. M. Vitousek, M. H. Williamson, and others).

Although the invasion of ecosystems by alien species certainly has occurred throughout human history, the problems of alien invasion remain particularly challenging both for scientists studying patterns and processes of invasion and for ecosystem managers attempting to eradicate or control invaders. In many parts of the world, particularly on islands, even the greatest conservation efforts are ineffective at countering the proliferation of exotic species. Where invaders are well established, such as rats in New Zealand and melastome plants in Hawaii, invasive species are nearly impossible to control and conservation strategies need to explicitly manage invasive species that currently cannot be eliminated. In less severe cases, conservation practices may be able to protect biodiversity and limit the establishment and spread of exotic species.

What is needed to adequately handle invasive species and prevent their establishment is a framework for understanding what qualities predispose an ecosystem to invasion and what species characteristics make invaders so ecologically opportunistic. Such a framework must be developed from knowledge of rates and types of species invasions, establishment conditions, and mechanisms of species dispersal in new habitats (see work by M. J. Crawley, and others). Clearly, the inherent dispersal ability of a species is a key determinant of its invasion potential. With respect to dispersal, classic ecological theory identifies two general types of species: those that occupy stable habitats at relatively constant densities and those that are locally ephemeral and have extreme population dynamics. Naturally, the selection pressure for dispersal ability among this latter group is high, and species of this type are well represented among introduced species. Of the plants in this group, many are early-successional species such as grasses and legumes. Probably more important than species' natural dispersal mechanisms, however, are their associations with humans. For example, stocks of seed and produce shipped overseas often carry with them seeds of many weed species that germinate in their new environment. These shipments commonly include alien microbes and animal pests as well. Intentional introductions of species imported for agriculture and other human uses are common, as are introductions of organ-
isms employed as biological control agents to stanch the tide of previously introduced and invasive species.

Whether or not an introduced species establishes itself in a new ecosystem depends on its ability to meet a specific invasion criterion, or population growth rate, upon introduction. With an increasing body of formal ecological theory, calculations of invasion criteria may be more or less spatially and temporally explicit, including information about generational overlap and demographic and environmental stochasticity. Godfray and Crawley have provided a thorough review of invasion criteria, including temporal and spatial complexity. However, because of the complexity of invasion criterion models, it is often difficult to predict whether or not a specific introduced species will establish itself without actually conducting an invasion. The risk of such an experiment is understandably high, so in general it may be wise to assume that any introduction is irreversible.

Although some introduced plants and animals commonly fare poorly in new habitats, there are many examples of introduced species that better fit their invasion criteria in their new range than in their original environment. Frequently this increase in fitness in an invasive species follows the release from pressures of herbivory, predation, or a pathogen associated with the invader’s native environments. Alternately, an introduced species may fail to invade if it is a relatively poor competitor in its new environment or if it is preferentially grazed or predated compared to native species. An oft-cited example of success following herbivory release is that of the invasion of New Zealand by Monterey pine (Pinus radiata). This pine is a native of the Monterey peninsula of California, where it tends to grow as a small, scrubby tree. When imported to New Zealand as an experimental timber species earlier this century, it was discovered that the pine grew many times taller and much straighter, and was longer-lived than in its home range (see Ledger and Richardson). With the help of widespread planting, seedlings of P. radiata and other American pines and firs quickly escaped from where they had been originally planted. Today the spread of these wildings is argued to be the dominant ecological threat to many alpine and lowland forest communities in New Zealand.

In addition to a species’ potential for invasion, conservation scientists and practitioners must also consider the predisposition of a community to be invaded. Research by Crawley and others found that recently disturbed and less contiguous communities are more readily invaded than closed and late-successional communities. It is very significant that these characteristics are associated with areas of high human activity. Species already associated with areas of human development are consequently more likely to be introduced to new areas by human vectors, and the areas to which they are introduced tend to be more readily overwhelmed. Observations suggest that alien plant species richness correlates positively with proximity to human development and transport centers.

Once an invasive species has successfully colonized a new habitat, the ecological consequences of the invasion are of greatest concern to conservationists. Again, although prediction of the specific consequences of a particular invasion may be difficult, a large body of current ecological theory may be harnessed to make some generalizations. From a community ecology standpoint, the introduction of a new species could have a large number of direct and indirect effects on other populations depending on the strength of its competitive interactions and its trophic position. For example, an introduced plant that successfully competes for a limiting resource with another plant might cause population declines of a herbivore that grazes exclusively on the native plant. Alternately, if an invader does not necessarily exclude a competitor but is palatable to a herbivore that also grazes on the competitor, the herbivore population would be expected to increase and the population of the competitor would be expected to decrease, mostly as an indirect consequence of the invader. Of course, when an invader takes a new, high trophic position in a community (e.g., cats, mustelids, and cats on South Pacific islands), it is certain to have a dramatic effect on prey that have not evolved under the selective pressure of predation.

It is also important for conservationists to consider the consequences of invasive species on patterns of succession. As noted earlier, open and disturbed communities are commonly prone to invasion. If an invasive species or suite of species is characterized by an early-successional stage in its native range, its invasion following a disturbance may effectively arrest succession on the site, both excluding native species in the present sere and precluding species typical of later stages of succession. Several examples of this have been recorded in Hawaii with an invasive nitrogen-fixing plant and for warm-season grasses that are arresting succession (see works of P. Kareiva, P. M. Vitousek, P. Walker, and others).

Invasive species are also known to have dramatic effects on the abiotic environment of the newly invaded ecosystem. Forest invasions by perennial grasses have been shown to increase the frequency and intensity of
fire disturbances and thus shift the annual timing of these events. Invasive trees on forested watersheds have been known to transpire much more water than native trees, significantly altering forest hydrology (see R. M. Cowling).

Of ultimate importance is the degree to which the successful invasion of an organism affects directly and indirectly the community as a whole and the ecosystem functions associated with it. In many ecosystems, ecologists have observed that certain species, known as keystone species, play a critical role in shaping overall community structure. The exclusion of keystone species often results in a cascade of population shifts of directly and indirectly related species. R. Paine and others have illustrated this in a marine intertidal community by removing a top predator and noting a general decrease in total species diversity due to the overwhelming growth of a specific prey population. Considering such observations, conservationists must be particularly sensitive to the possibility that an invasive species may outcompete, predate, or otherwise exclude a keystone species, resulting in a much larger loss of biodiversity.

Because of the uncertainty in our understanding of invasive species, conservation ecologists will have to consider whether invasive species are present in areas being managed for species and their habitats. If invasive species are present, it will be important to consult the biological literature and determine whether and what types of effects may be attributed to that invasive. If these invasive species are shown to negatively affect indigenous species and ecosystems, it will be critical to eradicate small pockets of invasive species when they are found to prevent their establishment and spread to the point where they cannot be controlled. This suggests that conservation practices will have to aggressively manage invasive species even when their impacts have not been conclusively shown to detrimentally affect the target system. It will be critical for conservation to establish priorities and develop an effective strategy for preventing and handling invasive species. Several ecosystems in which conservation projects had to consider eradicating or controlling invasive species are given in Table II.

III. TOOLS FOR MANAGING AND CONSERVING SPECIES

When Michael Soulé, the “father of conservation biology,” was interviewed in 1994 and asked about the developments in conservation biology in the last decade, he suggested that there has not been a revolution in how conservation biology is practiced. Rather, he stated that conservation has been experiencing a gradual refinement in the principles and tools of the field. Soulé suggested that the major tools have been the application of theory from the fields of island biogeography, population biology, and community ecology, namely, consideration of the matrix landscape within which a habitat is embedded, the development of biological corridors, the development of new tools in molecular biology, and the development of integrated population viability analysis frameworks that include demographic, environmental, and genetic stochasticity. All of these are commonly used in conservation today.

Since the early 1990s, however, there has been a revolution in the philosophical basis of conservation biology. As discussed earlier, five philosophical changes have occurred in conservation, however, two have arisen to the forefront of discussions within the last decade. These two philosophical changes have also been the most difficult to convert into practical applications for natural resource management. One was the acceptance of ecosystem management by United States federal agencies to help resolve conflicts arising from public perceptions of poor management of natural resources. The second philosophical change was the acknowledgment of the importance of human dimensions in natural resource management. The tools for implementing these holistic approaches to ecosystem analysis are rapidly developing so that they can be practically applied in the field. Because these holistic approaches to species conservation are still evolving, many ecologists are not comfortable with them (see D. Wilcove). Despite these reservations, many conservation organizations are actively developing the ecosystem approach as their primary tool for managing species. By accepting these new approaches, conservation biology will be able to contribute to the development of application tools for studying systems holistically and for integrating the social and natural sciences. Unfortunately, conservation biologists will not be able to borrow well-tested tools from these other disciplines since consensus is still evolving on which suite of tools best allows theories to be converted into practical applications.

Approaches to conservation continue to evolve as new knowledge is produced in population, community, ecosystem, and landscape ecology, and in social ecology. Some of the change in conservation applications occurred from a greater understanding of several of the reasons why the Endangered Species Act has had limited success in halting species extinction or contributing to their recovery. The Endangered Species
<table>
<thead>
<tr>
<th>Conservation target</th>
<th>Invasive threat to conservation efforts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wet lowland and alpine ohia (Metrosideros polymorpha) forests, Hawaii, USA</td>
<td>Among numerous invasive species, the Latin American native plant Miconia calvescens grows rapidly to dominance over native stands where established. The “green cancer,” as it is known in Tahiti, reproduces prolifically with high success contributing to its rapid dispersal. Miconia threatens biodiversity at all levels in these forests because of the high degree of endemism typical of the island ecosystems.</td>
</tr>
<tr>
<td>Fraser fir (Abies fraseri) forests, Great Smoky Mountains National Park, Tennessee, N. Carolina, and Virginia, USA</td>
<td>The European balsam woolly adelgid (Adelges piceae) is a sucking insect that kills firs over several years by draining their sap. Already the invader has eliminated more than three-quarters of all spruce-fir forests in the southeast US, shifting dramatically microclimate, vegetation, and food-web dynamics. A number of plant and animal species have been listed on the federal endangered species list as a direct consequence of habitat destruction in the national park.</td>
</tr>
<tr>
<td>Wetland plant communities, Everglades National Park, Florida, USA</td>
<td>The Australian Melaleuca quinquenervia forms dense monotypic stands in marshlands, reducing native plant diversity by 60 to 80%. Presently it covers more than 450,000 acres, and its rate of spread has grown to 50 acres per day. It is especially problematic to conservation practitioners because stress and damage from fire and herbicide cause it to release abundant seeds.</td>
</tr>
<tr>
<td>Freshwater bivalve communities, Mississippi River basin, USA</td>
<td>The zebra mussel (Dreissena polymorpha), a recent invader from the Caspian and Black Seas, adheres to the shells of freshwater mussels—as many as 10,000 zebra to one native—and interferes with the feeding, growth, movement, respiration, and reproduction of these species. The Mississippi basin hosts more native mussels than any other river system in the world, a community largely responsible for nutrient cycling and sediment mixing in this system. Zebra mussels have no American parasites or predators.</td>
</tr>
<tr>
<td>Shortgrass prairies and rangelands, Montana, Wyoming, the Dakotas, and Nebraska, USA</td>
<td>Leaky spurge (Euphorbia esula), a Eurasian herb, is a grassland invader known to completely and quickly displace grassy prairies and forbs by overshadowing neighbors and outcompeting them for available moisture and nutrients. It is also known to reduce productivity of western grazing lands by 50 to 75%. The spurge is difficult to eradicate owing to its deep root systems, high seed production, and dispersal abilities.</td>
</tr>
<tr>
<td>Tropical forest birds, Guam (also Hawaiian and other tropical and subtropical communities)</td>
<td>The brown tree snake (Boiga irregularis), a native of several South Pacific islands and northern Australia, is a widely adaptable predator that exacts its heaviest toll on tree-nesting bird species. As dense as 12,000 individuals per square mile in some forest areas, the brown tree snake has already eliminated 9 of Guam’s 11 native forest birds and most of its exotic birds. Although not yet naturalized in Hawaiian forests or other likely systems, the nocturnal snake is noted for stealth and has been found on several occasions in cargo transported by air to Hawaii and North America.</td>
</tr>
<tr>
<td>Riparian woodlands, Sonoran Desert and elsewhere in southwest USA</td>
<td>The Eurasian tamarisk tree (Tamarix spp.) is the superior competitor for water in these fragile desert ecosystems and devastates soil structure and hydrology, thereby altering the habitat of numerous plants and rare birds, mammals, amphibians, and invertebrates. It has no native predators or diseases, and the selective pressures of widespread grazing on native cottonwoods and willows exacerbate its proliferation.</td>
</tr>
</tbody>
</table>
Act has tended to protect high-profile species rather than overall biodiversity and did not sufficiently protect habitat reserves for recovered species. The focus on species conservation as the primary tool has been shown to inadequately reflect whether a habitat was sustainable. Conservationists moved to an ecosystem approach when it became obvious that species by themselves were generally poor indicators of ecosystem function and resilience. A brief discussion of how the philosophical changes in conservation have been translated to field applications will highlight the shift occurring in conservation.

A. Models

Past modeling efforts focused on using data-intensive species population and community models, such as Minimum Viable Population (MVP) and Population Viability Assessment (PVA) models (see S. Bettsinger, M. S. Westphal, and others), to understand how to conserve species. These models focused on the single species without taking into account the ecological factors of the species' habitat. As discussed earlier, these types of approaches have not been adequate to predict the ability for conserving single species in ecosystems and landscapes. Recently, PVA models have been adapted to consider the spatial context, to better incorporate uncertainty and stochasticity, and to have more statistical rigor. There is also a growing attempt to develop wildlife habitat association models and link these with GIS databases to predict species diversity in actual geographic locations.

Models are also evolving as tools to help balance trade-offs between the design of conservation areas and conflicting natural resource uses. New models are using multiobjective programming to examine the multiple trade-offs that can exist in the design of a reserve (see K. D. Rothley, O. J. Schmitz, and others). To design bioreserve networks, these models are incorporating decisions that are made by animals for food and shelter, by people interested in recreation, and also by landowners wanting current income.

It is unlikely that any reserve or reserve network can satisfy the demands of all of the stakeholders involved with the land base within and/or surrounding reserves. So planners must find ways of reaching favorable compromises among competing demands. In many cases, however, reaching compromise solutions may require considering a multitude of options. For example, if a goal of reserve network design is to select 5 reserve areas from a set of 20 potential sites on a given land base, then one must decide among 15,504 possible networks of 5 reserve areas. The decision-making becomes even more complex when one must consider additional criteria such as size and shape of reserves, interconnectedness of reserves and the need for corridors, numbers of species preserved, locations of the reserves, and the cost of purchasing land (as discussed by G. Caughley, A. Gunn, and K. D. Rothley). Reserve planning has become increasingly sophisticated and it now enlists numerous algorithmic and optimization tools that help reconcile trade-offs among competing alternatives.

B. Species and Their Spatial Relationships in the Landscape

Much of the early theory development by conservation biologists was based on determining the relationships between species and their space (as first articulated by R. H. MacArthur and E. O. Wilson). The suggestion that habitat area was positively correlated to species richness provided arguments for designing larger natural reserve areas for maximizing the conservation of species. The size of a reserve area needed for effective conservation has been argued in the literature for more than 20 years without the development of a consensus. Part of the reason for a lack of consensus is that we know that area by itself is insufficient to characterize how many species can be maintained at a site (see work by J. F. Franklin and D. A. Saunders for discussion of spatial scale issues).

Even though space has been an important theoretical issue in conservation, the practical application of theory in conservation appears to have no dominant spatial scale of analysis. Therefore, theories and applications have not been linked at the spatial scale. A lack of defined spatial scale in applied conservation has been strongly influenced by the focus on specific animals as important species to conserve. The spatial scale at which conservation has been practiced has varied according to the territorial needs of the species of interest. This trend is apparent from a survey of key journals published in 1998 that tabulated the scales of analysis used by conservation biologists, community and population ecologists, and social ecologists. An amazingly high proportion of the articles published in the journal Conservation Biology did not even present the scale being used in the study; this suggests that spatial scale was not considered relevant by the authors for understanding their system. When a scale of analysis was presented, the smallest scale (<−0.01 ha) was as equally represented as the largest scale (>−10,000 ha). Again, this reflects the focus on animals.
This contrasts with much of the information collected from community and population ecology journals, where small-scale research is predominantly used to understand species relationships to ecosystem parameters. For example, plant ecologists who are developing the theory behind species dynamics focused on scales that were less than 0.01 ha in size. Some studies are being conducted at larger scales, but these larger scales of analyses are applied in the tropics and not other regions of the world. In the humid tropics, the presence of rare plant species has necessitated the establishment of permanent plots that are 50 ha in size (see P. Ashton, S. Hubbel, and others).

If conservation biologists are going to accept the ecosystem management paradigm, they will not be integrating the scales of analyses that are particular to ecosystem ecology. The watershed has been accepted as the relevant scale in ecosystem management (see FEMAT, 1993), but this is only one of the many scales that could be used in conservation. Selecting one scale of analysis for all systems is a real problem because scale is site dependent and must be chosen to reflect the driving variables controlling the functioning and resilience of that system.

C. Reserve Designs

Reserve design used to be based on defining the carrying capacity of an appropriate area needed by a species to maintain it in the landscape. This type of approach to defining the area was sufficient when there was an immediate need to set aside areas to protect species. However, this type of approach does not deal with the fact that the quality of the space makes a difference in how effectively it is used by the species—it is not just a question of total area and the amount of interior and edge environments. We realize that the distinction between interior and edge environments appeared to be easy to distinguish in the field, but in fact was not so clear when species other than birds were the species of conservation interest. How the quality of an environment changes with respect to the matrix landscape that is not in reserves is receiving considerable research attention today.

It is now universally acknowledged that the conservation of species requires the preservation of natural habitats within ecosystems. One fundamental concern is that if the rates of habitat loss due to resource exploitation continue at current levels, many habitats will disappear forever by the middle of the twenty-first century (see D. S. Hik, O. J. Schmitz, A. R. E. Sinclair, and others). The growing concern over the loss of habitat, and the concomitant loss of biodiversity, has generated intense interest in the use of reserves and reserve networks as conservation tools. For example, the Wildlife Conservation Society's flagship project, "Paseo Pantera," is an effort to consolidate and link protected areas throughout Central America. The intention is that these reserves will afford species an extensive refuge from exploitation.

The most fundamental strategy in reserve design is to circumscribe and set aside parcels of land that are key sources of biotic diversity, and to exclude all human activity within their boundaries. In this sense, reserves act as repositories for much of the world's biota in perpetuity. There are two main reasons why such a strategy alone is insufficient. First, natural habitats decline both within and outside of reserves through natural processes (e.g., succession) and as part of the differential impact of disturbances within the matrix landscape. Thus, habitat within protected areas will continue to change, albeit at slower rates than outside of protected areas. Consequently, the structure and function of those putatively protected ecosystems will not be sustained over the long term simply by delineating them as reserves (as discussed by A. R. E. Sinclair and others). Second, setting aside land to protect specific ecosystem types may no longer be possible because the land base has been degraded to a point where the complement of habitats needed for a sustainable ecosystem no longer exist. The solution is to adopt a new conservation strategy that promotes the reestablishment of habitat via habitat renewal and restoration at rates that offset natural habitat decline within a landscape framework. Accordingly, management must choose not only the size of land parcels to preserve, but also the size of area that balances habitat loss with habitat renewal. This may require the restoration of some ecosystem types and the eradication of invasive species.

Thus reserve designers must adopt a holistic perspective that considers habitat decline and renewal as consequences of succession and disturbances. They must also maintain a long-term, large-scale perspective in deciding how to balance renewal against loss. For example, consider a 100-ha parcel of old-growth forest that forms part of a reserve that is rapidly decreasing in area due to natural stand break-up (see B. C. Larson and C. Oliver). In order for old-growth forest to be sustained within the reserve, another 100 ha of younger successional forest must be available to replace the declining stand as that older forest stand is shrinking, not 100 or 200 years in the future. In this case, one must set aside much more land than 100 ha and protect a wider
variety of successional-age stands in order to have a sustainable old-growth forest ecosystem. In addition to these biologically motivated criteria for reserve design, there is growing recognition that human social factors outside of the reserves can play a major role in determining the conservation effectiveness of a reserve.

D. Economic Evaluations

In the past, economics was the dominant tool being used to build a market-based rationale for conserving species and ecosystems. Economics was used to show that conservation, and the resulting protection of ecosystem services, is more valuable than extraction or harvesting of resources. However, this overreliance on economics as the primary tool to build a rationale for conservation has produced mixed results (see C. H. Freese and others).

In the late 1980's, economists provided conservation biologists with what appeared to be an excellent tool to conserve tropical forests that were high conservation priorities. The analyses also provided different options for generating income by producing values for previously non-market-valued products. Historically, tropical forests had been undervalued when only timber and agricultural uses were used to estimate their net worth, so that there was no clear economic incentive to conserve these forests (see P. S. Ashton, T. Panayotou, C. M. Peters, and others). These authors have suggested that the lack of proper economic valuation has contributed to accelerating the rate of harvesting and conversion of forests for other uses. Therefore, a tactic to decrease deforestation rates was to determine the “real” value of the forest by including all other products that could be extracted from it (e.g., nontimber forest products and services like watershed protection). Using this analysis, the forest was no longer undervalued, and alternative products could be extracted that would provide income to local communities. This strategy is now widely promoted by the international development community (see K. Brandon, R. J. A. Goodland, and M. Wells). Its goal is to conserve forests and biodiversity in reserves while also generating income for local communities from the same reserve or surrounding areas.

This strategy of combining conservation with sustainable development was supported by studies reporting that extraction of nontimber forest products would return greater value in the marketplace than timber harvesting. In addition, this strategy was promoted as a new alternative to reduce the dependence of the local community on natural resources that were being protected in a newly formed conservation area. International funding organizations were interested in this strategy because it suggested that the goals of conservation were compatible with developing an economic base for local communities, who could forgo harvesting resources (trees, wildlife, etc.) within designated reserve areas if sufficient income could be generated from extracting other resources or from providing recreational opportunities from the forest. Unfortunately, it is not clear whether the goals of conservation and sustainable development are compatible. M. Dove has conducted research suggesting that this strategy should be approached with caution, since developing a total value for a forest does not protect it from being overharvested.

Some economic analyses have been criticized for producing unrealistic expectations of the value of what can be sustainably extracted from a forest (see Bawa, Godoy, and others). Biological inventory data do not indicate the amount of a product that can be harvested in an ecologically sound manner, and also significantly overestimate the expected income generated annually from extracting these products (Belsky, Pinedo-Vasquez, and Siebert). Indeed, placing an economic value on a resource can increase the demand for that product so that it is quickly overharvested; when this occurs, the species may be lost from the landscape where it previously was quite common (see J. L. O'Hara and M. Pinedo-Vasquez).

IV. FUTURE DIRECTIONS AND CHALLENGES IN CONSERVATION

The solutions to the problems facing conservation biologists will not be found solely by adopting new paradigms such as ecosystem management or sustainable development. Blending the theories developed in conservation biology with these new paradigms will potentially allow for better management of species within their landscapes. However, conservation biologists will not automatically find the tools that they need from the different disciplines. These new paradigms are currently struggling to develop tools that will be able to take theories and make them applicable to, for example, forest certification. Furthermore, conservation biologists have not developed structures similar to ecosystem management that would allow them to develop an adaptive management model (i.e., a way to allow scientific data to continuously modify policy links as better information became available). Without such a process or
framework to effectively link policy to science and management, it is difficult to integrate scientific findings into management itself. As stated by J. C. Gordon in 1999, "Often, policy is made with 'old' science, while interest groups confront the manager of public resources with 'new' science." One could speculate that "old" science is used because the tools have not been developed to implement or make the "new" science work at the ground level or in policy.

Conservation biologists will also have to confront the fact that humans have altered the environment on a global scale (see NRC, 1992). There is probably no place in the world that has not experienced human impact. In many cases, the current system that we study and want to conserve is something that may be an artifact of past human activities. If this is not recognized, it may be extremely difficult to conserve species because we really do not understand the driving variables that may control the resistance and resilience characteristics of the system. As protected area managers are increasingly called upon to generate funds through development projects such as ecotourism, timber harvest, and nontimber forest product extraction, it is important that the original goal of setting up these reserves as "genetic libraries" is not compromised.

Restoration is frequently suggested as a way to reestablish desired ecosystems and at the same time maintain the biodiversity of indigenous species and exclude invasive species. This is especially relevant considering the problems that conservation managers are facing with invasive species and altered landscapes. However, the science of restoration ecology is still evolving, and standard tools are not readily available for conservation purposes. A current practice has been to allow the development and elimination of a specific habitat with the assumption that an equivalent habitat will be recreated at another location. This approach has been commonly used in wetlands. Unfortunately, the successful restoration of a habitat that can then be used by an endangered species rarely results. For example, several authors (J. B. Zedler and R. Langis) found that a restored marsh that was supposed to support an endangered species was not suitable habitat for that species. They also found that after five years the marsh was still not supporting other species found in their reference marsh. Restoration ecology is still a young science, and it is not surprising that restoration tools have not had sufficient time to be tested in a variety of systems. Many restoration efforts are currently dealing with eradicating invasive species that have begun to dominate much of our landscape. Yet we have a minimal understanding of the impact of invasive species within landscapes and of the long-term ecological effects that they may exert in these newly modified environments.

A central challenge for conservation in the twenty-first century is to recognize that conservation efforts will be ineffective if the science in conservation biology fails to incorporate the human factor. Most models linking the human dimension to conservation are theoretical and few have managed to realistically link people and the environment. At this time, the tools are not readily available to link applications of the social and natural sciences, and it will require some innovative approaches to address this problem. Several other problems must also be explicitly addressed in conservation:

- New paradigms that can help shift conservation from a species focus to a holistic approach have good theoretical grounding, but there are few frameworks for implementation.
- Past land uses and their legacies have already modified the environment (i.e., habitat fragmentation, chemical pollution, and introduction of species that have become invasive) sufficiently that ecological theory or conservation models that ignore the prevalent effects of humans on the environment will not succeed in field application.

Conservation biology has done well in developing the science of understanding individual species in their habitat, performing spatial scale analyses of individuals, and modeling their activity within the landscape. It has not done well in the following areas:

- Incorporating the roles and needs of humans in conservation situations, especially at the local level. The values and principles of national and international organizations have largely driven conservation efforts.
- Developing frameworks for moving beyond the species focus and incorporating the dynamic nature of ecosystems into these frameworks.
- Focusing on the use of "easy" tools (i.e., GIS models) has restricted the diversity of conservation approaches and created a dependence on a small set of tools, thereby excluding the use of a more varied and innovative suite of tools.

See Also the Following Articles

BIODIVERSITY-RICH COUNTRIES • CONSERVATION MOVEMENT, HISTORICAL • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • ECOSYSTEM, CONCEPT OF • HOTSPOTS • INTRODUCED SPECIES • KEYSTONE
Bibliography


CONSERVATION MOVEMENT, HISTORICAL

Curt Meine
International Crane Foundation

I. The Background of Conservation
II. Modern Origins of Conservation
III. The Evolution of Twentieth-Century Conservation
IV. Emerging Themes in Conservation

GLOSSARY

anthropocentric Human-centered in perspective, especially with regard to the value of the natural world.
biocentric Valuing the existence and diversity of all biological species.
conservation biology An integrative approach to the protection and management of biological diversity that uses appropriate principles and experiences from the natural sciences, the social sciences, and various resource management fields.
preservation The school of conservation philosophy that emphasizes the protection of natural features and landscapes from human exploitation.
Progressive Era The period in American political history, especially coinciding with the presidency of Theodore Roosevelt (1901–1909), in which the conservation movement gained definition.
sustainable In general, capable of meeting economic goals in a manner that does not degrade the quality of the underlying ecosystem.
sustained yield The management and harvesting of renewable resources in a manner that does not exceed their rate of replacement or reproduction.
utilitarian The school of conservation philosophy that emphasizes the value of the natural world in contributing to human well-being, especially in terms of economic standard of living.

MODERN EFFORTS TO CONSERVE BIODIVERSITY have their foundations in older traditions of resource management and nature protection. This chapter traces the history of the conservation movement, focusing on those events and patterns that led to the emergence of biodiversity conservation from earlier utilitarian and preservation-oriented approaches. Because the conservation movement continues to redefine itself, this article concludes with a consideration of key themes from recent history.

I. THE BACKGROUND OF CONSERVATION

As a conscious goal of citizen action, public policy, and professional endeavor, conservation first gained coherence in the late 19th and early twentieth centuries. Over the decades, the conservation movement has evolved in response to varied forces: emerging findings in the natural sciences, new environmental threats, shifts in philosophical assumptions and aesthetic standards, novel technologies, expanding legal mandates, and changing social, economic, and political conditions.
The complex interplay of these forces has produced a movement that remains in flux and whose goals continue to evolve, even as the world’s ecosystems face increasing challenges to their integrity.

A. Prehistoric Precedents

The term conservation acquired its modern meaning in the early twentieth century, especially with the rise of the Progressive Era conservation crusade in the United States. However, as an expression of cultural commitment to an enduring relationship with the natural world, conservation has much deeper historic and prehistoric roots.

Prehistoric peoples did not live in a simple or constant state of peaceful coexistence with and within their natural surroundings. Converging lines of evidence from paleontology, paleoecology, archaeology, and anthropology suggest in fact a sobering picture of the human past: that the dispersal of the human population over the past 120,000 years has been accompanied by spasms of continental and insular extinction and other forms of environmental degradation. The best known of these prehistoric extinction events involves the disappearance of mastodons, ground sloths, camels, giant beaver, and other large North American mammal species coincident with a major wave of human colonization between 12,500 and 10,000 years ago. The same general pattern of extinction has been traced back over the centuries as humans moved out of Africa and Eurasia into Australia, Polynesia, and island groups around the world. The implications of this pattern of “dreadful synocation” are controversial, and research continues on the exact sequence of cause and effect, the interaction of probable causal factors, and the precise mechanisms of extinction (MacPhee, 1999).

In contrast to this record of prehistoric anthropogenic extinction, however, there is countervailing evidence from the many cultures that have achieved relatively sustainable ways of life. Native peoples in landscapes throughout the world developed sophisticated belief systems and land-use traditions that recognized their connections to and dependence on the nonhuman world. In the past, these traditions allowed people to adapt to even extreme environments throughout the world. The Maasai of East Africa, for example, practiced livestock herding practices in a manner that allowed them to coexist with the native megafauna of their semiarid homeland. Arctic dwellers coexisting with caribou herds and marine mammals, shifting cultivators and agroforesters in the humid tropics, small-scale farmers in temperate zones, fishing villages in coastal, riverine, and wetland environments: these and other traditional cultures developed social systems that balanced human needs against the capacity of the natural environment to meet those needs. Some of these traditions have survived into the modern era, though they must contend with increasing pressures from diminished resources, population growth, altered land tenure systems, and rapid economic and technological change.

The resource-use practices of native peoples had significant impacts on the biota, though these impacts varied in intensity over time and according to ecosystem type and taxonomic groups. People employed fire to modify vegetation and to concentrate game animals. They fished, hunted, and gathered in ways that affected species populations, animal behavior, and the dynamics of interspecific relationships. They domesticated plant and animal species, and altered natural hydrological systems through irrigation. In some cases—for example, the deforestation of Mediterranean watersheds and the development of hydraulic cultures in arid and semiarid lands—large-scale resource exploitation caused fundamental changes in biological diversity and ecosystem function. In other cases, the limited rate and scale of exploitation resulted in relatively minor long-term ecological impact.

Humanized landscapes, in sum, experienced varying degrees of anthropogenic environmental change. However, in many if not most native cultures, social mechanisms evolved to reinforce an attitude of respect and deference toward nature, to evoke nature’s bounty, and to sanction appropriate human use of that bounty. These mechanisms included hunting, planting, and herding rituals; stories and myths; fertility and birth control practices; recognition of sacred spaces; and the invocation of taboos and totems. The conservation movement might be characterized as modern society’s conscious effort to develop and exercise analogous social practices and restraints to guide its relations with the nonhuman world.

B. Historic Precedents

Contemporary environmental dilemmas have prompted modern scholars to reexamine conservation traditions in texts and stories from animist, Native American, Buddhist, Hindu, Islamic, Judeo-Christian, and other faith traditions. In the Judeo-Christian tradition, the biblical injunction to “fill the earth and subdue it, and have dominion over every living thing that moves upon
the earth" has often been cited as providing the justification and rationale for centuries of environmental exploitation in the Western experience. Only recently have biblical scholars returned to these textual sources to identify alternative traditions of stewardship and respect for creation.

Since ancient times, observers have recorded instances of environmental deterioration due to human action. Plato, for example, compared the deforested mountains of Attica to the "bones of a wasted body ... the richer and softer parts of the soil having fallen away, and the mere skeleton being left" (Dubos, 1980). Evidence of early conservation practices can be found in the history of both Western and non-Western cultures. These include efforts to protect particular species and special lands, to maintain populations of wild plants and animals, and to sustain the productivity of agroecosystems. Leopold (1933), for example, cited Marco Polo's narrative account of hunting restrictions and provisions for game bird populations in 13th century China as "the first clear record of a well-rounded system" of wildlife conservation.

European traditions of forestry and game-keeping date back to the Middle Ages and beyond. The establishment of royal game preserves and forests on the land estates of feudal Europe led to the development of customs and formalized laws regulating hunting and use of the forests. In the eleventh century, William the Conqueror set up in the newly conquered England "many deer preserves and also enacted laws/That whoever killed a hart or hind/Should be blinded./He also placed a ban on harts, also on boars" (Harrison, 1992). The high-handedness of the Norman conquerors in establishing privilege use of the forests gave rise to a venerable tradition of local resentment over centralized resource management. As codified in forest law, however, these measures allowed England's forests to survive and regenerate. On the continent, Germany and France in particular developed silvicultural systems and techniques that prevented wholesale destruction of the forest estate.

Europeans carried these "proto-conservation" traditions to their expanding colonial empires. The change in jurisdiction over natural resources, from native peoples to colonial governments, had profound implications for social systems and ecosystems alike. Colonial domination, coupled with rising populations and industrialization, disrupted traditional patterns of land tenure and resource use. In many regions, native peoples became increasingly alienated from their landscapes, while Western conservation ideas and practices were slow to adapt to the new environments.

In the New World, European colonists encountered a landscape of abundance that admitted profligate resource use. Unchecked resource exploitation was the norm as European settlement of North America proceeded. Despite this record, seeds of the later conservation movement were apparent in early efforts to protect wild game populations, forests, soils, and special natural features. Early measures to conserve game include decrees to protect the cahow (Pterodroma caahow) and green turtle (Chelonia mydas) in Bermuda (1616, 1622); the first closed hunting season on white-tailed deer (Odocoileus virginianus) in Massachusetts (1694); closed seasons on several species of game birds in New York (1708); and the institution of game warden systems in Massachusetts (1739) and New York (1741) (Matthiessen, 1959). The first federal game law, enacted in 1776, mandated closed seasons on deer in all the American colonies except Georgia. Through the 1800s the new American states intermittently passed laws establishing closed seasons, prohibiting hunting of nongame birds, and placing bounties on predators.

Exploitation of North America's extensive native forests—for conversion of land to agriculture, for construction materials, for charcoal and domestic fuel supplies—drove economic development through much of the colonial and early American era. In the late seventeenth and early eighteenth century, the British crown instituted the Broad Arrow policy, which reserved for the crown the right to cut the strategically important white pine (Pinus strobus) mast trees of the New England colonies. Although serving to restrict exploitation, these laws were intended more to secure the property against unauthorized timber cutting than to protect or perpetuate the forests per se. Closer in intent to later forest conservation measures was William Penn's 1681 dictum that, in the process of land development, one acre of forest should be maintained for every five acres cleared.

Little heed was paid to such recommendations. By 1745 Benjamin Franklin was noting the scarcity and high cost of fuelwood in Philadelphia. President James Madison in 1818 drew attention to the "injurious and excessive destruction of timber and firewood" as the most important and regrettable force shaping the rural economy of the United States. Such destruction continued through the 19th century. With the depletion of the eastern forests and the continuing settlement of interior North America, the focus of forest exploitation shifted to the Great Lakes region. In 1831 Alexis de Tocqueville, while visiting Michigan, noted that Americans were generally "insensible to the wonders of inani-
mate nature, and may be said not to perceive the mighty forests that surround them till they fall beneath the hatchet." The swift destruction of the Great Lakes forests in fact marked a turning point in forest conservation, and in the conservation movement generally (Williams, 1989).

A parallel pattern of resource degradation and early conservation response marked the process of agricultural development in North America. By 1776 the problem of soil erosion was evident to the new nation's leading statesmen. Thomas Jefferson experimented with contour plowing to retard soil erosion. Patrick Henry, speaking before the Virginia Assembly, stated that "Since the achievement of our independence, he is the greatest patriot who stops the most gullies." George Washington noted in a 1797 letter, "We ruin the lands that are already cleared and either cut down more wood if we have it, or emigrate into the western country" (Jackson, 1985). Over the next century, destructive agricultural techniques continued to affect widely varied ecosystems across North America (and other parts of the world as well).

The preservation impulse in conservation found special expression in the North American setting, as European settlers encountered scenic landscapes and natural features unlike any in their prior experience. Jefferson celebrated Virginia's Natural Bridge in his Notes on the State of Virginia. Niagara Falls, the Hudson River valley, and the prairies, mountains, and canyons of the American West became emblematic of the beauty to be found in New World landscapes, a quality seized upon by artists, promoters, and early conservationists alike. The artist George Catlin suggested after his first excursions in the West that portions of the continent's undeveloped lands were "worthy of our preservation and protection." As early as 1833, Catlin envisioned government reservation of "a magnificent park . . . a Nation's park, containing man and beast, in all the wild[ness] and freshness of their nature's beauty" (Nash, 1982). Such protected areas would not come into being until later in the nineteenth century, although the United States did move as early as 1832 to protect special natural features when it reserved Hot Springs in Arkansas as a federal holding.

What these disparate efforts had in common was not the conservation of biological diversity in the modern sense, but the protection or regulated exploitation of economically and aesthetically important components of the landscape or ecosystem. As such, they provided the foundation on which a more coherent conservation movement began to take shape in the latter half of the nineteenth century.

II. MODERN ORIGINS OF CONSERVATION

A. Prelude to a Movement

The conservation movement emerged in response to accelerated changes in demographic and environmental conditions, and in human perceptions and understanding of the natural world, through the eighteenth and nineteenth centuries. Exploration of the world's diverse ecosystems—from the earliest voyages of discovery through the New World scientific expeditions of Alexander von Humboldt, John and William Bartram, Meriwether Lewis and George Rogers Clark, John James Audubon, and others—contributed to a golden age of natural history studies. Establishment and adoption of the Linnean system of binomial nomenclature in the 1700s allowed for an unprecedented flourishing of taxonomic research (Mayr, 1982). This, in turn, provided critical foundations for the development of evolutionary theory in the work of Alfred Russell Wallace and Charles Darwin. Although the biological sciences had not yet developed field methods for comprehending the full diversity of life, they had begun to reveal the fundamental processes through which life in fact diversified and maintained itself.

These gains in systematics, biogeography, and evolutionary theory occurred even as the Industrial Revolution accelerated the scale, pace, and character of human environmental impacts. Through the 1800s, the advent of more efficient technologies resulted in increasingly intensive exploitation of forests, game populations, fisheries, agricultural lands, and river systems. Traditional resource management practices and established land tenure systems were abandoned or changed to fit the emerging economies of scale. Industrial pollution and the spread of invasive species became widespread problems for the first time.

Coincident with these scientific, cultural, and environmental changes, the Enlightenment and Romantic movements were altering Western conceptions of order, value, and beauty in the natural world. The natural philosophers of the Enlightenment stressed the smooth workings and stability of a mechanistic natural order. The Romantic philosophers and poets emphasized the unity and wholeness to be found in a spontaneously creative organic nature. Although offering different conceptions of nature, both encouraged human comprehension of natural objects and processes, and so laid the foundation for greater appreciation of human impacts upon the natural world. The writings of
Thomas Malthus and other early economic philosophers provided the basic framework for considering the interwoven fate of human population, human economies, and natural resources.

In Europe the Romantic movement drew heavily upon the experiences of New World explorers and settlers, the encounter with native peoples there and elsewhere, and the exposure to wild landscapes. In turn, adaptation of the Romantic impulse in the North American setting provided important literary and philosophical foundations for conservation. The strong American identification with wild nature founds its first full expression in the essays of the transcendentalists Ralph Waldo Emerson and Henry David Thoreau, the poems of William Cullen Bryant, and the novels of James Fenimore Cooper (Nash, 1982).

George Perkins Marsh's *Man and Nature: or, Physical Geography as Modified by Human Action* (1864) is widely regarded as the first landmark in modern conservation literature. A native Vermonter, Marsh saw in the destruction of New England's forests the latest expression of an ancient human tendency to "[derange] the original proportions between different orders of organic life." Drawing also on his extensive personal observations of long-term landscape change in the Mediterranean, Marsh argued that human actions had caused widespread disruption of the "harmonies" of the natural world. Marsh's reasoning followed lines that would sound familiar to later generations of ecologists and biodiversity conservationists. "All nature," he wrote, "is linked together by invisible bonds, and every organic creature, however low, however feeble, however dependent, is necessary to the well-being of some other among the myriad forms of life with which the Creator has peopled the earth" (Marsh, 1864).

**B. Gaining Definition**

Publication of Marsh's book provided direction to the conservation movement as it gained greater definition through the remainder of the 1800s and the first decade of the 1900s. In North America, dispossession of the American Indians, enactment of liberal land distribution policies, and the flow of settlers and capital into "virgin" landscapes resulted in an unprecedented wave of exploitation of natural resources. The effects were visible in ecosystems across the continent. These decades saw the virtual depletion of the extensive pine forests of the upper Great Lakes; overhunting of many game bird and mammal species by market hunters; rapid conversion of the Midwest's extensive prairies to agriculture; degradation of aquatic systems through overfishing, pollution, and hydrological changes; and widespread overgrazing of the semiarid western rangelands. Several well-recognized cases of species depletion and extinction—of the white pine, of waterfowl, of the bison (*Bison bison*), the Carolina parakeet (*Conuropsis carolinensis*), and the passenger pigeon (*Ectopistes migratorius*)—came to symbolize this era of unmitigated resource extraction and decline.

This period, however, also saw the first concerted efforts to address the causes and consequences of these massive changes. At first these actions tended to focus on particular issues, problems, or landscapes. Sportsmen led the campaign to rein in market hunting and to institute stronger game laws at the state and local levels. Many of the nation's most prominent sportsmen banded together in 1887 to form the Boone and Crockett Club, which became one of the first and most effective nongovernmental organizations to become involved in shaping national conservation policy. Exploitation of plume-bearing egrets and other birds for the millinery trade mobilized sportsmen, scientists, and nature lovers to work together for reform, and prompted formation of state Audubon Societies beginning in 1896. These moves culminated in passage of the Lacey Act of 1900. The first important piece of modern wildlife conservation legislation, it barred interstate shipment of wild animal species taken in violation of state laws, and soon succeeded in ending the plume trade (Dunlap, 1988).

Efforts to reform American land policy proved more difficult. Through the 1800s, the nation's land allotment policies had encouraged rapid immigration, settlement, and land conversion, often accompanied by corruption and land speculation. In the midcontinent's mesic landscapes, the system of land surveying and distribution facilitated the conversion of millions of acres of native savanna and grassland to crop and livestock production. These policies, however, were less tenable in the progressively drier lands of the American West, where farming was possible only near permanent watercourses, and livestock ranching required large land holdings. The geologist and explorer John Wesley Powell attempted in the 1870s and 1880s to devise a more appropriate tenure system that recognized the inherent environmental limitations and social requirements of the landscape, but the speculative momentum proved intractable. Only decades later did the U.S. government institute more careful land management policies on the nation's remaining public domain. However, Powell's innovations, especially his commitment to a strong role for science in land management, laid important founda-
tions for future private and public conservation planning (Stegner, 1954).

The rapid destruction of forests in the upper Great Lakes between 1865 and 1900 stimulated the most forceful and effective response among early conservation reformers. As this period began, forestry as a profession had not yet gained a foothold in the United States. Through the 1870s and 1880s, however, voices opposed to the destructive "cut and run" logging practices began to emerge. These reformers included botanists Charles S. Sargent, Asa Gray, and William H. Brewer; naturalists John Muir and Increase Lapham; the European-trained forester Bernard Fernow; Secretary of Interior Carl Schurz; and Franklin B. Hough, appointed as the nation's first government forestry official in 1876.

The forestry movement gained momentum quickly in the last decade of the 19th century. Forest activists working through professional scientific and forestry organizations campaigned for national legislation to reform public land laws and to prevent further forest destruction. These efforts culminated in 1891 with the adoption of legislation that included a provision allowing the U.S. president to set aside public land forests as forest reserves, the germ of the nation's current system of national forests. Within weeks of its passage, President Benjamin Harrison used the law to designate forest reserves in lands adjacent to Yellowstone National Park. In the next 10 years, 47 million acres of forest land were placed in reserves.

During this period, the loss of forests contributed also to the parallel movement to protect and preserve unusually scenic landscapes, special natural features, and wild spaces. Like the forestry movement, this channel of activism gained momentum in the decades following the American Civil War. In 1872 the U.S. Congress established Yellowstone National Park, the world's first. In 1885 New York created a state forest preserve in the Adirondack Mountains in order to protect its wild character as well as its watershed values. That same year Canada designated Banff National Park, its first. The impetus toward preservation drew heavily upon the enthusiastic nature writing of John Muir, who led the political effort that in 1890 resulted in designation of an enlarged Yosemite National Park in California and who in 1892 founded the Sierra Club (Fox, 1981).

At the end of the 19th century, the varied strands of conservation concern—over depleted wildlife, unprecedented deforestation, watershed degradation, inappropriate land development, loss of scenic values, and aesthetic quality—were connected only loosely. In the American experience, political corruption and inordinately concentrated wealth had typically accompa}ned these environmental changes. As the new century began, these overlapping social, economic, political, and environmental concerns brought forth a more consolidated movement that gave new meaning and power to the term conservation. However, the growing prominence of conservation also revealed inherent tensions in the nascent movement.

C. Progressive Era Conservation and the Utilitarian/Preservationist Split

When Theodore Roosevelt assumed the American presidency in 1901, the stage was set for a revolution in conservation policy. Roosevelt brought to the presidency a formidable command of the natural sciences, experience as a rancher and outdoorsman in the American West, a long-standing devotion to public policy reform, and boundless personal energy. In his first presidential address to the U.S. Congress, Roosevelt spoke at length of the importance of the nation's forests, stating that "We have come to see clearly that whatever destroys the forest ... threatens our well-being" (Pinchot, 1947). Conservation became a cornerstone of the Progressive political movement and, for the first time in history, a high national priority.

Roosevelt's partner in political innovation was forester Gifford Pinchot. Pinchot, the first American to receive formal training in forestry, was a friend of Roosevelt's who had been active in the complex politics of American forestry through the 1890s. At the time Roosevelt became president, Pinchot was in charge of the Division of Forestry in the U.S. Department of Agriculture. The two quickly forged a close alliance to strengthen the government's role in forestry. Their partnership during Roosevelt's presidency resulted in an increase in the total acreage of the forest reserves (renamed national forests in 1905) from 60 million to 151 million acres; transfer of these lands to the Department of Agriculture under Pinchot's jurisdiction; and creation, in 1905, of the U.S. Forest Service to administer the new national forests.

Pinchot's Forest Service embodied the Progressive spirit and approach to bureaucratic responsibilities. In contrast to Muir and others in the preservationist wing of the conservation movement, Pinchot conceived of the forests in utilitarian terms. The forests were not regarded as "reserves" to be "locked up," but as lands to be worked "for the greatest good of the greatest number for the longest time" (as the utilitarian credo put it). In practice, this meant that the forests were to be managed by a trained, professional workforce; that
scientific principles were to guide the efficient and sustained exploitation and processing of forest resources; and that the wealth derived from the forests was to be equitably distributed for the common good (Hays, 1959). Applied not only to forests, but to natural resources in general, the "resource conservation ethic" provided the dominant paradigm of the early movement (Callicott, 1991). In Pinchot's words, "The first great fact of conservation is that it stands for development" (Pinchot, 1910).

At the core of utilitarian conservation was the concept of sustained yield. As Aldo Leopold later (1933) observed, under the Progressive conservation banner "wild life, forests, ranges, and waterpower were conceived ... to be renewable organic resources, which might last forever if they were harvested scientifically, and not faster than they reproduced. 'Conservation' had until then been a lowly word, sleeping obscurely in the dictionary. The public never heard of it. It carried no connotation of woods or waters. Overnight it became the label of a national issue." As the leading edge of this movement, the Forest Service became a model not only for other resource management agencies but for Progressive-era government agencies in general.

Theodore Roosevelt's conservation activism extended beyond the public forests. During his presidency he established the nation's first wildlife refuge, at Florida's Pelican Island (1903); designated 50 additional wildlife refuges; used new presidential powers under the Antiquities Act of 1906 to set aside 18 national monuments (including the Grand Canyon); convened in 1908 a national Governor's Conference on the Conservation of Natural Resources; and created, also in 1908, a National Conservation Commission to provide continuing advice to the president and to work with the states on conservation policy. Early in 1909, Roosevelt and Pinchot sought to broaden the scope of conservation through a North American Conservation Conference and laid still more ambitious plans for a World Conservation Conference. Though the latter was unrealized, Pinchot and Roosevelt had succeeded in securing the place of conservation within both the political and geographic landscape.

Absent from many of the Progressive-era political initiatives were the voices of the increasingly influential preservationists and nature protectors. In contrast to the utilitarian views that Pinchot exemplified, adherents of the "romantic-transcendental preservation ethic" emphasized wild nature's aesthetic and spiritual values and the need to safeguard those values for future generations through strict prohibitions on development and manipulation (Callicott, 1991). The tensions between the utilitarian and preservationist approaches intensified as conservation assumed center stage politically. These tensions surfaced in the changing relationships among the principal players. Pinchot and John Muir had been friends, but by the late 1890s their differing approaches to management of the forest reserves caused a rift between them that would never heal. Few from the preservation camp participated in the landmark 1908 governors' conference on conservation.

Muir maintained a respectful relationship with Roosevelt. They camped together in Yosemite in 1903 and worked together in gaining protection for the Grand Canyon as a national monument in 1906. But the schism between the conservation factions continued to widen. It came to a head in the bitter political struggle over plans to dam the Tuolumne River in Yosemite National Park's Hetch Hetchy Valley. Waged over a period of 6 years, the battle culminated in 1913 with the adoption of national legislation providing federal support for the dam. Although Muir and his colleagues lost the battle, they had aroused a national constituency in favor of protection. The growing popular and political acceptance of preservation led directly to creation of the U.S. National Park Service in 1916 (Fox, 1981).

For all of the profound developments in conservation during the Progressive Era, scant attention had as yet been given to the state of the biotic diversity and ecological processes characteristic of either wild or more humanized landscapes. That attention would come only slowly, as conservation science, philosophy, policy, and practice co-evolved through the twentieth century.

III. THE EVOLUTION OF TWENTIETH-CENTURY CONSERVATION

The Progressive era advances fixed conservation permanently in the public mind and in public policy. The innovations of that period, however, served only to suggest the basic questions that would challenge conservationists throughout the twentieth century: What are we conserving? To what end? For whom? How? With what scientific information and understanding? With what social adjustments, economic mechanisms, and legal tools? Moreover, the conservation movement came out of the Progressive Era with the tension between the preservationist and utilitarian approaches still unresolved. The continuing efforts to answer those questions and to resolve that tension framed the century's technical and political advances in conservation.
A. Institutionalizing Conservation

In the United States, the first four decades of the twentieth century saw the consolidation of the Progressive era's conservation gains in the government, the private sector, the professions, and the academy. The spirit of reform that drove the conservation movement was now redirected toward the development of sound conservation policy and administration. New issues, however, arose to keep the movement invigorated and to expand its philosophical dimensions.

As conservation became institutionalized, it tended to follow the tenets of Piøch's resource conservation ethic. By the late 1930s, the basic principles of utilitarian resource conservation had been applied not only to forests, but to other “useful” components of the biota and the landscape: rangelands, game animals, sport and commercial fisheries, scenic areas, agricultural soils, river systems. New policies, laws, bureaucracies, academic disciplines, research and training programs, and professional societies arose to promote sustained yields of and from these various “resources.” Foresters, range managers, wildlife managers, recreation planners, and soil conservationists achieved their own professional identities in these years. Amateur conservationists, meanwhile, became increasingly involved in decision making through the formation of new citizen organizations.

Forestry. As an established field, forestry provided the model and foundation for the other emerging resource management professions in the United States. With those fields still lacking formal training and employment opportunities, forestry tended to attract professionals with broad conservation interests. However, institutionalization tended over time to narrow the scope of the field to silviculture and timber harvesting. This trend was magnified during these decades as the supply of timber taken from private lands diminished and the pressures on public forests increased.

Range management. In the American west, rapid settlement and exploitation had degraded vast stretches of arid and semiarid grassland in the late 1800s and early 1900s. Research and academic training programs followed in the wake of the deteriorating range conditions, bringing scientific management to rangelands much as forestry had brought such management to forests. Changes in land policy resulted by the late 1930s in stricter regulation of the public rangelands, cessation of privatization of the public domain, establishment of a new federal bureau (the Bureau of Land Management) to oversee these lands, and the emergence of range management as a discrete field.

Wildlife management. Although concern about the fate of wild animal populations had often galvanized the young conservation movement, wildlife management did not carve out its own professional niche until the 1930s. Continuing declines in game populations provided the initial impetus. To arrest these declines, game protectors at first focused on enactment of stricter hunting regulations, sporadic establishment of refuges, development of artificial propagation programs, control of predators, and the introduction of exotic game species. Only in the early 1930s did a revolutionary new approach, focused on the protection and restoration of habitats so that populations might sustain themselves, begin to take hold. (Fisheries management followed a roughly parallel pattern with regard to game fish species and habitats.) As principles from ecology filtered into the field, the term wildlife came into common usage, and the idea of management expanded to encompass the perpetuation of “nongame,” rare, and threatened species. Although the field tended to focus primarily on game species, it provided important foundations for efforts later in the century to conserve biological diversity generally.

Recreation. As the American population became increasingly urban and increasingly mobile, outdoor recreation became a more important component of conservation. Camping, fishing, hunting, hiking, and other outdoor activities continued to grow in popularity, even after the prosperity of the 1920s gave way to the Great Depression of the 1930s. Park agencies, at all levels of government, responded by establishing and developing areas of significant aesthetic and recreational value. Other land management agencies, such as the U.S. Forest Service, began to devote greater attention to the recreational values of the lands under their jurisdiction. Although there were exceptions, recreational developers tended to override or ignore problems associated with the protection or management of the biota, and they rarely incorporated scientific information in planning for increased tourism.

Soil conservation. The expansion and mechanization of agriculture in the early decades of the twentieth century brought widespread disruption to the agricultural landscape and economy. The disaster of the Dust Bowl in the southern Great Plains during the mid-1930s was the most highly visible consequence of forces that had degraded soils across North
America. Intensified exploitation of farm resources through land clearing, "clean farming," poor grazing management, and wetland drainage led to accelerated rates of soil erosion, degradation of pasture and forage vegetation, loss of wildlife habitat, and aggravated cycles of siltation and flooding. In response, the U.S. Congress in 1935 created the Soil Conservation Service (SCS) to work with farmers and other private landowners to adopt conservation practices. Under the leadership of Hugh Hammond Bennett, the SCS importantly emphasized watershed-scale approaches that integrated soil protection measures with other resource conservation goals.

Although increasingly defined by this enhanced structure of agencies, disciplines, and professions, the conservation movement remained a battleground for competing approaches and intense political activity. Even as conservation became professionalized, growing numbers of citizens became active as members of non-governmental conservation organizations. New national groups, including the Izaak Walton League, the National Wildlife Federation, and Ducks Unlimited, joined such older organizations as the Sierra Club and the Audubon Association as significant players in shaping conservation policy (Fox, 1981).

Meanwhile, the preservationist approach found renewed vigor as the campaign to protect wildlife, natural areas, and wildlands reemerged in the 1920s and 1930s. In these years, field zoologists such as Joseph Grimnell and Olaus Murie became leaders in the effort to reverse government antipredator policies. Early ecologists, under the leadership of Victor Shelford, created in 1917 a Committee on the Preservation of Natural Conditions within the Ecological Society of America. This committee provided the nucleus for what became, decades later (1950), The Nature Conservancy. In 1924, the U.S. Forest Service designated the nation's first "wilderness area" on a large roadless expanse of the Gila National Forest in New Mexico. In 1935, proponents of wilderness protection in the United States banded together as The Wilderness Society, the first conservation organization dedicated solely to the cause of wilderness protection. Although the wilderness protection movement still emphasized primarily the aesthetic, spiritual, and recreational value of wildlands, these moves signaled a shift toward greater recognition of the scientific and ecological value of undeveloped lands.

In the international arena, the years prior to World War II saw increased recognition of the global scope of conservation challenges and halting moves to institutionalize a response to those challenges. The western powers—particularly the United States, Britain, France, and Germany—continued to export forestry and national park programs to their colonies and protectorates. Early international conventions and treaties addressed the problems of protecting fur seals, sea otters, whales and other marine species, and migratory birds (Matthiessen, 1959). In 1913 an International Conference on the Protection of Nature in Switzerland attracted representatives from 16 nations. World War I prevented this effort from gathering momentum, but in the 1920s and 1930s intermittent steps were taken toward greater international cooperation: a second international conservation conference in Paris in 1923; establishment in 1928 of an International Office for the Protection of Nature in Brussels; an unprecedented international bird conservation conference in Paris in 1933; a Conference for the Protection of African Fauna and Flora in London, also in 1933; the first North American Wildlife Conference in 1936; establishment of the Pan-American Union, dedicated largely to conservation issues in the Western hemisphere, in 1940 (Nash, 1982). These set the stage for the expanded international programs that emerged in the aftermath of World War II.

B. Conservation Amid Crisis:
   The 1930s and 1940s

The social and environmental convulsions of the 1930s and 1940s fundamentally altered perspectives and priorities within the conservation movement. The Depression forced conservationists and nonconservationists alike to consider the connections between human economic systems and the sources of wealth in nature. In practical terms, the New Deal programs in the United States vastly increased governmental support for conservation programs, though often with little consideration of ultimate conservation goals. The experience of the Dust Bowl focused the attentions of an entire generation of conservationists and many natural scientists on the consequences of indiscriminate exploitation of sensitive lands, highlighting especially the systemic nature of the soil erosion problem. With the outbreak of World War II, conservation issues fell into the background of concerns, proving the difficulty of maintaining mindfulness of human-nature relationships when social crises erupt. As in no time since the Progressive Era, these years of crisis demonstrated the interwoven nature of social, economic, and environmental problems.

As a movement, conservation remained prominent
in North America and elsewhere, with growing popular and political acceptance. But for all of the recent advances, the conservation movement still suffered from its own internal crisis. Beneath the very active surface of conservation, there remained the underlying philosophical rift between the utilitarian and preservationist approaches. Neither approach adequately addressed such extensive problems as soil erosion and exhaustion, disruption of hydrological cycles, endangerment and loss of wildlife species, degradation of biotic communities (both terrestrial and aquatic), and the increased incidence of invasion by exotic pests and diseases. For a small but growing group of conservationists, the Dust Bowl (among other factors) stimulated a new effort to address the roots of conservation problems, one that would draw on elements of both the utilitarian and preservationist schools. Embracing the preservationist critique of human hubris in manipulating the natural world, the aesthetic appreciation of wild nature, the traditional commitment to wise use and stewardship of resources, and the value of science in the service of conservation, this cohort of conservationists sought to define a more robust synthesis to guide conservation action.

The incipient synthesis reflected a basic shift in conservation's scientific foundations. Theory and application—basic science and conservation practice—were meeting in ways they had not before. Developments in ecology and evolutionary biology over the previous decades had begun to inform conservation issues, even as contemporary problems forced conservationists to reexamine their scientific assumptions. Ecology was revolutionizing scientific understanding of the functioning of biological communities, landscapes, and systems. Evolutionary biology provided new perspectives on, for example, the adaptations, roles, and interactions of all forest species (in contrast to the basic descriptive botany, dendrology, timber physics, and forest mensuration upon which silviculture and forestry had previously rested).

For at least some conservationists, these scientific advances suggested a new need: to marry ecology and the various fields of resource management in the effort to sustain not just yields of particular commodities but the healthy functioning of the systems generally. The preeminent voice of this emerging approach was Aldo Leopold, who in the course of his career had applied ecological principles first to the conservation of forests, then to soils, game, and wildlife species, then ultimately to the land "as a whole." Writing in 1939, Leopold noted that ecology provided "a new fusion point for all the natural sciences" and that its emergence had placed the economic biologist in a peculiar dilemma: with one hand he points out the accumulated findings of his search for utility, or lack of utility, in this or that species; with the other he lifts the veil from a biota so complex, so conditioned by interwoven cooperations and competitions, that no man can say where utility begins or ends.... The only sure conclusion is that the biota as a whole is useful, and [the] biota includes not only plants and animals, but soils and waters as well (Leopold, 1939, p. 727.)

Leopold's expanded conservation philosophy, as finally expressed in his landmark essay "The Land Ethic" in *A Sand County Almanac* (1949), placed greater emphasis on the ecological integrity, diversity, and beauty of what he called "the biotic community" and rejected the view of nature as merely a collection of disaggregated natural resources. It shifted the role of human beings "from conqueror of the land community to plain member and citizen of it" (Leopold, 1949). This proposed shift, away from an anthropocentric toward a more biocentric approach, implied the need for fundamental changes within the various fields of resource management. Resource managers could no longer regard timber trees, game animals, water, soils, scenic vistas, or any other economically or aesthetically significant resource as separate entities. Rather, resources had to be seen as components within diverse systems, connected to and interacting in complex ways with other parts of the system (including human beings). It followed that the conservation professions had to develop more integrated approaches to resource management.

As an attempt to address conservation's underlying schism, Leopold's "evolutionary-ecological land ethic" would confront continuing challenges in the dramatically altered postwar environment (Callicott, 1991). However, it would also provide those in the conservation movement with new conceptual tools with which to meet those challenges.

### C. From Conservation to Environmentalism

The conservation movement emerged from World War II into a changed world. The war altered conservation's social and ecological context. In the world at large, the bonds of empire dissolved in the postwar decades as colonial nations gained their independence. The former colonial powers remained involved in the fate of their
erstwhile colonies as donors and providers of development assistance. This in turn altered the shape of the global conservation movement, as aid providers, international agencies, and nongovernmental agencies initiated the long struggle to somehow reconcile development and conservation priorities.

These changes in international relations occurred even as human population growth, land degradation, air and water pollution, and overexploitation of marine resources first emerged as global-scale conservation issues. The war had demonstrated the interrelated and global nature of modern conservation problems. Fairfield Osborn’s *Our Plundered Planet* (1948) and William Vogt’s *The Road to Survival* (1948) were among the first books to address in an integrated manner issues of global conservation and development. Reflecting this broader scale of concern, a wide spectrum of national and international governments and organizations met in 1948 and formed the International Union for the Protection of Nature (now known as the World Conservation Union) with the ambitious goal of preserving “the entire world biotic environment.” Shortly thereafter, the United Nations became involved in international conservation programs. These were early indicators of what became an increasingly active international response to conservation problems.

In the United States, the war years constituted a demographic divide, with important implications for the conservation movement. As veterans and their families resumed their lives at home, the postwar population boom fueled an economic boom that placed increasing demands on natural resources. To meet those demands, war-spawned technologies and industrial processes were quickly adapted for use in the peacetime economy. Over the next several decades, rings of suburbs expanded away from cities into rural lands and, along with dams, highways, and other infrastructural developments, continually reconfigured the landscape.

The postwar years brought a new generation of professionals into agriculture, forestry, wildlife management, and the other conservation fields. Conservation was no longer immune to the larger social and economic forces in the postwar environment. As the pressure to meet rising demands increased, the resource management professions became more specialized, more focused on commodity outputs, and more inclined to adopt technical solutions to conservation problems. Importantly, the war also changed the ways in which scientific research was supported and conducted, reinforcing the tendency toward specialization and utilitarian applications. The traditional foundations of conservation in systematics, organismal biology, population biology, ethology, and other field-oriented sciences suffered from relative neglect. Collectively these changes overwhelmed the integrated approach that had begun to emerge prior to the war, requiring conservationists to confront complex environmental problems within an increasingly rigid framework of disciplines and institutions.

However, the postwar years also gave rise to forces that countered these trends. New tools in the earth sciences provided greater scientific understanding of the interrelations within and among terrestrial, aquatic, and atmospheric systems. In the 1950s and 1960s, revolutions in fields ranging from genetics and evolutionary biology to atmospheric chemistry and geology began to reshape our understanding of the global biosphere and the human place within it. New communications technologies allowed such information to be more efficiently shared. Meanwhile, the indiscriminate adoption of other technologies—especially in the production of agricultural pesticides, industrial chemicals, and nuclear power—created new concerns over their impacts on human health and ecological systems. With the publication in 1962 of Rachel Carson’s landmark book *Silent Spring*, examining the biological impacts of DDT and other pesticides, the modern environmental movement began to assume an identity of its own, distinct from but still connected to the older conservation tradition.

Efforts to protect wildlands in the United States also intensified in the years following World War II. As pressures from urban and suburban expansion, resource extraction, water development, and motorized tourism increased, wilderness protection became a unifying goal among diverse conservation organizations. In particular, the defeat in 1955 of the proposal to build the Echo Park Dam in Dinosaur National Monument in Colorado indicated that the conservation movement had gained not only in cohesion, but in popular and political power. That unity led to the passage in 1964 of the Wilderness Act, providing for a strengthened national system of wilderness areas on the public lands of the United States.

The Wilderness Act was only one of many laws that emerged from this period of environmental awakening (much as earlier laws had institutionalized the conservation movement). Over the next decade, the U.S. Congress would enact a series of important environmental statutes, including the National Environmental Policy Act (1970), the Clean Air Act (1970), the Clean Water Act (1972), and the Endangered Species Act (1973). These far-reaching changes in environmental policy re-
lected a rising wave of popular support and organized political activism, symbolized by the observance of the first Earth Day on April 22, 1970.

D. The Reintegration of Conservation

With the emergence of the environmental movement, the context of conservation once again experienced dramatic changes. The years following Earth Day saw increasing acceptance, within the conservation professions and within society at large, of environmental values. Those values, however, were not easily translated into effective conservation action. The resource management professions, in seeking to address broader environmental concerns, found the tendency toward overspecialization difficult to overcome. Conservation programs tended still to focus on single species, or particular economic resources, or separate jurisdictions within a given landscape. As the long-term ecological impacts of such fragmented approaches to resource management became clear, the traditional conservation fields came under increasing public scrutiny. While many environmentalists pressed for change through legal means, the evolution of management philosophy, policy, and practice would require years of incremental change and adjustment.

At the international level, differences in perspective between the wealthier, developed countries of the North and the poorer, developing countries of the South likewise proved difficult to overcome. International conservation continued to make important gains through a series of conventions and treaties, including the Convention on Wetlands of International Importance (the “Ramsar Convention”) (1971), the Convention on International Trade in Endangered Species (CITES) (1975), the International Convention for the Conservation of Migratory Species of Wild Animals (the “Bonn Convention”) (1978), the United Nations Convention on the Law of the Sea (1982), and the United Nations Convention on Biological Diversity (1992). Again, however, international development policies would come to incorporate stronger conservation and environmental provisions only gradually, as the need to connect economic development and long-term environmental security became increasingly evident.

Even as these broad patterns of change in the human dimensions of conservation unfolded in the 1970s and the 1980s, the scientific foundations of conservation were shifting. Taxonomy and systematics provided more robust estimates of the extent of species diversity and of its actual and potential loss. Advances in island biogeography revealed the principles governing the spatial distribution, persistence, and extinction of species, with important implications for the establishment and management of protected areas. Genetics became an increasingly important component of conservation science as attention focused on the reproductive success of rare and endangered species and the viability of their populations, both in captivity and in the wild. Ecology moved away from its “classical paradigm,” which emphasized single, stable, deterministic equilibria, and toward a view of ecosystems that emphasized flux, uncertainty, and contingency. Disturbances such as fire, flooding, and drought were incorporated into our understanding of long-term ecosystem dynamics and were shown to be important factors affecting species diversity. Increasingly, conservation strategies required the integration of knowledge from the many branches of biological science, involving various levels of biological organization.

The need to rethink conservation across disciplinary lines was driven not only by changes in the foundational sciences, but by changes in the environment itself. By the late 1970s, scientists and conservationists were alarmed by the escalating loss of genetic, species, and ecosystem diversity at the global scale. Of special concern was the accelerated destruction of the species-rich forests of the humid tropics. The advances in island biogeography revealed that ecosystems of all types were being fragmented by human activity, while even the most effective protected areas were at risk due to their inadequate size and their isolation. Wildland managers increasingly understood that preservation alone was an inadequate management strategy, and that the loss of diversity and the disruption of ecological functions were intimately associated. Agricultural scientists, foresters, and other resource managers, too, were increasingly concerned about environmental degradation, the breakdown of ecosystem processes, the loss of diversity in cultivated landscapes, and the human costs associated with these changes.

These concerns prompted the emergence in the mid-1980s of a new synthetic discipline, conservation biology, specifically devoted to the integration of scientific disciplines and other fields in the effort to understand, maintain, and restore biological diversity. To this end, conservation biology has sought to address conservation problems within an evolutionary and ecology context, and to stimulate the traditional conservation professions to reassess their management methods and goals accordingly. The term biodiversity itself was coined in 1987 and has since been widely adopted by conservationists. The concepts of sustainability and sustainable development came into general usage during the same period, reflecting the complex challenge of
integrating long-term social, economic, and environmental factors in assessing human demands and impacts on ecosystems. The integrating impulse was also evident in the recent emergence of restoration ecology, landscape ecology, sustainable agriculture, environmental history, environmental ethics, ecological economics, and other interdisciplinary fields.

These recent changes in conservation science, philosophy, and policy have prompted many in the traditional resource management fields to adopt more integrated approaches. In practical terms, this means that as the conservation professions approached the new century, they sought to move away from narrow economic criteria for success and toward broader ecological standards; away from the mathematical *desideratum* of sustained yield of commodities and toward the more complex challenge of sustaining healthy ecosystems; away from the management of discrete components of the system and toward integrated ecosystem management; and away from an exclusive focus on the human goods and services provided by ecosystems and toward a focus on the biological diversity and ecological processes that sustain those goods and services. These changes have also begun to redefine the value of wilderness and to reintegrate wildlands within the broader landscape of conservation. Whereas wilderness as traditionally defined existed apart from the humanized landscape (even while being preserved to meet human recreational and aesthetic needs), wildlands at all spatial scales have become increasingly valued as repositories of biological diversity, as core protected areas and corridors within greater ecosystems, and as “controls” against which to compare the human impact on more intensively utilized lands.

These examples of increasing integration in conservation hearken back in many ways to the ecology-based synthesis that Aldo Leopold and others sought to articulate prior to the onset of the Cold War. Since then, the science of ecology has evolved, analytical tools and technologies have grown vastly more sophisticated, environmental threats have intensified, and the social and economic context of conservation has changed dramatically. Nevertheless, there is continuity between contemporary efforts to expand conservation’s cultural and natural connections and Leopold’s efforts to “enlarge the boundaries of the community to include soils, waters, plants, and animals” (Leopold, 1949).

IV. EMERGING THEMES IN CONSERVATION

As in the past, conservation remains responsive to new demands, new information, and new realities. The sweeping changes that have reshaped the conservation movement since the mid-1980s will continue to alter both its human dimensions and its practical field methods. Although we are still too much embedded in the present to recognize all these recent developments, we can identify important contrasts between emerging themes and conservation’s past practices.

In terms of its scientific foundations, conservation now seeks greater collaboration among the sciences, even while reaching beyond the sciences to build connections with philosophy, economics, history, literature, and the arts. Ecology continues to be the “fusion point” of the conservation-related sciences. In its conservation applications, ecology is bringing attention to a broader spectrum of species and to the processes that maintain diversity. To improve its chances of success, conservation will need to persevere in its efforts to marry ecology to the resource management professions, and to do so at various spatial and temporal scales. The integrating fields of conservation biology, landscape ecology, restoration ecology, sustainable agriculture, and ecological economics will need to resist the tendency toward specialization and reductionism to which even interdisciplinary fields are prone.

Emerging threats are making the conservationist’s task even more precarious. These include the rapid spread of invasive exotic species (especially as the human economy continues to globalize), the uncertain impacts of global climate change, the increased presence of genetically modified organisms in the landscape, and the seemingly relentless fragmentation of habitats and ecosystems. At the same time, the sciences have recently been called upon to illuminate the connections among biodiversity conservation, human health, and the well-being of human communities. Accordingly, the active weaving of conservation science into community-based ecological restoration programs promises to redefine the role and methodology of science in conservation.

The conservation movement continues to be shaped by overarching social and demographic trends. The growth in human population and in resource consumption rates remains a profound factor determining conservation’s future. To better inform that future, conservationists are seeking novel ways to integrate sustainable economies with effective protection, management, and restoration programs. In so doing, conservation planning efforts are expanding their reach across jurisdictional boundaries and across the landscape, recognizing the connections among wildlands, semiwild lands, “working” landscapes, and suburban and urban environments. Such cross-boundary forays have few precedents in conservation history and represent an important departure from the past.
In the past, conservation has also neglected human and cultural relationships to its own detriment. Increasingly conservationists recognize the need to integrate natural science, cultural traditions, and social relationships in the effort to protect and restore particular places. In seeking a better fit between the built environment and surrounding landscapes, conservationists have also begun to work more closely with architects, landscape architects, planners, and engineers to incorporate the emerging principles of conservation biology.

Meanwhile, the conservation movement itself continues to increase in diversity as individuals from varied backgrounds come into conservation as both professionals and amateurs. Local nongovernmental organizations are now regarded as key conservation players, and their role in conservation continues to grow. Although action at the international level remains necessary in addressing global threats to biodiversity and in supporting local initiatives, the assumption of greater local responsibility for ecosystem health is increasingly urgent. Community-based conservation projects, watershed restoration programs, land trusts, and other mechanisms are emerging to meet that need. As these experimental programs continue, they seek to create not just an enhanced conservation movement, but an enduring culture of conservation. In so doing, they extend the long history of increasing citizen involvement in conservation.

See Also the Following Articles
ECOSYSTEM, CONCEPT OF • ENVIRONMENTAL ETHICS • HISTORICAL AWARENESS OF BIODIVERSITY • INDIGENOUS PEOPLES, BIODIVERSITY AND • LITERARY PERSPECTIVES ON BIODIVERSITY • RELIGIOUS TRADITIONS AND BIODIVERSITY • STEWARDSHIP

Bibliography
I. What Is Domestication?
II. How Does Domestication Happen?
III. The Selection, Fixation, and Status of Amazonian Plant Domestication
IV. Final Comments

GLOSSARY

barreais Soils naturally enriched with organic matter and clay carried by the rain and/or a river.
caboclos Cultural and/or racial crossing between European and indigenous populations.
macaxeira Group of varieties of manioc (Manihot esculenta Crantz) with low levels of cyanidric acid. These varieties may be consumed without prior treatment to lower their toxicity, as must be done with other varieties of wild manioc. Synonyms: sweet manioc and aipim.
maromba Elevated flooring made of wood, used in flooded areas for placing cattle or plants during river flood periods.
reproductive isolation When a derived population is not able, for some reason, to breed and obtain fertile progeny with the original population.
wild manioc Manioces with a high level of cyanidric acid.

IN THIS ARTICLE, concepts are discussed relating to natural biodiversity used to form new crops and for the improvement of already domesticated species. The majority of examples come from underexploited plants of the Amazonian indigenous peoples, little known in the international scientific literature, but which are splendid examples of the capacity for breeding and selection of traditional communities. The article also gives curious reports of Amazonian Indian practices of management and interchange of germplasm from useful plants.

I. WHAT IS DOMESTICATION?

The evolution of a plant occurs when there is selection; this can be the result of human activity, which generally values attributes of interest to people in detriment to those necessary for the plant’s own survival in the wild. When this process reaches high levels of genetic modification, to the point that human aid is necessary for survival, the plant is considered to have achieved the stage of being domesticated (Harlan, 1975). Useful plants occur at various stages of domestication that require different degrees of management and cultivation. Generally, the more domesticated the plant, the more demanding it is as to degree of management and cultivation, but the fact of a plant being managed or cultivated does not imply that it is domesticated. However, any plant that has been managed or cultivated for a long time is inevitably undergoing modification toward domestication because it is placed in environmental conditions improved by humans, who reduce
or eliminate the usual processes of natural selection such as competition and predation. Domestication is related to the plant’s genetic response, management, and cultivation and human activity in dealing with the plant (Harlan, 1975).

This article will deal principally with practices still used by Indians and caboclos of Amazonia who may directly (conscious selection) and indirectly (unconscious selection) be inducing modifications in natural plants in the direction of their domestication.

II. HOW DOES DOMESTICATION HAPPEN?

The forces of selection, resulting from management practices or from cultivation per se and as induced by farmers, may be consciously or unconsciously provoked (Harlan, 1975). In conscious selection, the farmer sees some interesting attribute in a plant and seeks to propagate it in some way, be it by sexual means (seeds) or asexually (generally by trying to induce a portion of stem to take root), thus forming a new population. If the characteristic selected has good inheritability—that is, it can be transferred from one generation to another—the result is an improved population. In unconscious selection, as the name indicates, this selection does not occur but the populations are modified due to some human intervention, generally in the environment or harvest system, which ends up benefiting some segment of a population to the detriment of the others, generating a selected population. The classic example of unconscious selection is that of rice, which upon being harvested has seeds in its panicle that do not easily fall from the plant. When these seeds are planted, they carry alleles that favor this characteristic, which is selected increasingly more favorably with each harvest. Seeds of domesticated rice today do not fall as easily from the panicle as do those of more primitive varieties (Harlan et al., 1973).

The well-known “genetic drift” of the classic works on evolution is also very much associated with the process of domestication. Genetic drift occurs when a very small portion of a certain population is taken to another area that is isolated and there regenerates, but does so only with the descendants of the few individuals introduced therein (Fig. 1). Many times this process does not bring gains in productivity but does induce an increase in more or less stable varieties that end up forming the “local varieties.” This process is very common among farmers who have the habit of collecting only a few seeds from a given plant for planting on their farms or villages, generally far from the original population and isolated by some environmental barrier. This may be a large river or, in the case of plants typical of disturbed environments such as the plants domesticated in farm plots, a climax forest in which such species do not occur.

Currently domesticated plants were sometime ago wild plants which were in stages used, then managed, and finally cultivated (Clement, 1992a,b). The process of domestication, for the majority of species, took thousands of years. For corn, whose domestication process is the best known, it is estimated that the process took more than 3000 years (Iltis, 1986). If, for an annual plant, so much time was spent, one can imagine how long it took to domesticate the Amazonian fruit trees that are perennials. The appearance of a domesticated plant may also be sudden due to some mutation that provokes a very radical modification in the plant, such as alterations in the number of chromosomes. For example, cupuaçu (Theobroma grandiflorum Schum) is an Amazonian fruit highly valued for its pulp, but whose seeds constitute 32% of its total weight. A triploid variety that produces only pulp and no seeds has been found and is being propagated by grafting (Venturieri and Mendonça, 1985). Peach palm (Bactris gasipaes H. B. K.), considered to be the only palm tree domesticated in the Americas, is probably a hybrid between Bactris dahlgreniana Glassman and Guaduina insignis (Clement, 1988). Finally, there was the suppression of some gene or group of genes that made a species poisonous, as is the case with wild manioc, with a high level of cyanidric acid, producing macaíba with very low levels of cyanidric acid. Nowadays, with the help of advanced selection and biotechnology techniques, the process of domestication can be considerably shortened, but still 99% or more of the species currently domesticated were achieved by farmers before the “green revolution” (Harlan, 1975).

The greater the diversity of these plants, in terms of species as well as of varieties within each of these species, the greater would be the options of choices by humans.

For selection to be efficient there must be a reproductive isolation between the population selected and the original group from which it was taken. If there is no isolation, the new population again crosses with the original population and becomes similar to it, without achieving any gain. The more efficient this isolation, the more rapidly a given population may be domesti-
cated (Fig. 2). In indigenous peoples’ systems, this isolation is achieved by environmental barriers or management practices. In summary, to obtain a domesticated plant there must be (i) the chance for choice (genetic diversity), (ii) discovery of useful plants and obtaining of interesting varieties and (iii) reproductive isolation of the varieties selected so that they do not return to the wild state. How were these predicates achieved for Amazonian plants?

A. The Chance for Choice
(Genetic Diversity)

The Amazonian forest offers a grandiose biodiversity with an immense number of useful plants for domestication. Considering only the fruiting species (Benza, 1980; Cavalcante, 1991; Clement, 1999; Correa, 1926; Hoehne, 1946; Kerr and Clement, 1980; Kerr and Posey, 1984; Le Coint, 1947; Smith et al., 1992), one may prepare a list with more than 250 typically Amazonian species, some of which are internationally recognized, others with use restricted to Indians who extract them from the native forest, and other “little-known” species in general use among the civilized population that has undergone Indian influence. Certainly, if one considers the vastness of Amazonia with its 4 million km² and its incredible variability of plant species, estimated to be between 50,000 and 60,000 for the higher plants alone (Clement et al., 1982), and the current state of economic botany in the region, it may be concluded that this list may still be increased considerably. For example, in two expeditions specifically planned for collecting potential fruiting species in the Amazon region, 10 species were attained that had not yet been cataloged (Instituto Nacional de Pesquisas da
FIGURE 2 Diagram exemplifying the effect of reproductive isolation. A selected sample with large fruits is removed from an original population. (a) When the isolation is efficient—that is, when the new population is planted in a plot at a distance that avoids gene exchange with the original population—the new population will produce descendants with large fruits. (b) When isolation is not efficient, the new population may cross with individuals from the gene group that gave rise to it and reestablish the characteristics of the original population among its descendants. The more effective the reproductive isolation, the more rapidly genetic gains are achieved.

B. The Discovery of Useful Plants and Obtainment of Interesting Varieties

In the eyes of "civilized" society, it is difficult to understand the process of discovering uses for plants. The Indians experiment with almost everything new that they find in the forest or that is brought to them by white people. As presents, they like receiving new plant species and varieties, such as the exotic plants that have recently been introduced in their cultures. The Kayapó Indians are very fond of mango, an Asian tree, and plant the seeds everywhere (Kerr and Posey, 1984), as do the Nambiquara Indians. Some Indians describe methodologies for discovering the usefulness of plants by employing concepts that are difficult to understand; for example, an Indian explained to me that the discovery that the papo-de-peru plant (Aristolochia sp.) served as an antidote to snake venom was due to the fact that "its flower looked like the skin of a snake." It is common for them to risk their lives for discovery. I witnessed among the Hahaintesu group of the Nambiquara the ingestion of the aril of seeds of a Philodendron sp. (Fig. 3), the latex and fruit of which they recognized as very poisonous and capable of causing death. They would split the fruit and with tweezers carefully remove the seeds so that they were not contaminated by the latex and then place them in their mouths to suck the arils, doing so carefully so as not to bite the seeds which were later discarded. These Indians eat tarantulas (Setz, 1991) after singeing them to remove hairs that can cause irritations to the skin (in this case, in the mouth). They also cook and eat the tuberous stalk of a plant of the genus Zamia, which if eaten raw causes a strong headache (Fig. 4). Apparently, the process of discovering a plant's usefulness is by trial and error since the analogies between aspect and use are not logical. From the amount of information existing on edible and medicinal plants, one may imagine how many lives were lost and pains felt by these intrepid researchers.

Domesticated plants are generally described as having "appeared" through some myth. Among the Hahaintesu group of the Nambiquara there is a myth that

**FIGURE 3** Philodendrum sp. poisonous species whose seeds are removed from the fruit with the aid of tweezers and the aril is sucked carefully so as not to bruise the seeds, which are also toxic.

**FIGURE 4** Indian boy of the Hahaintesu group of the Nambiquara showing a plant of the genus Zamia, whose tuberous stalk can only be eaten cooked. When ingested raw, it causes a strong headache.
describes the origin of man, whose head was the gourd (Crescentia cujete L.), whose teeth were the corn (Zea mays L.), whose ear was the lima bean (Phaseolus lunatus L.), whose bones were the manioc (Manihot esculenta Crantz), and whose scrotum was the cara yam (Dioscorea trifida L.) (Fig. 5). In the Kayapó myth, corn was acquired from the stomach of a rat that lived in an enormous tree in the middle of an area that was being prepared for planting; the strength of the entire tribe was needed to fell the tree (Lukesch, 1976). Another Kayapó myth mentions that sweet potato, macaxeira, and banana were brought by the “daughter of the rain” who came down from the sky, became an ordinary Indian, married, had children, and during a time of

FIGURE 5 Species from the Namíquara myth about the appearance of cultivated plants cultivadas (a) the head is represented by the calabash (Crescentia cujete L.) (photo by Nigel J. H. Smith), (b) the teeth by corn (Zea mays L.), (c) the ear by the lima bean (Phaseolus lunatus L.), (d) the bones by manioc (Manihot esculenta Crantz), and (e) the scrotum by the cara yam (Dioscorea trifida L.).
hunger returned to the sky and brought back these species that were then cultured so that the tribe never again went hungry (Lukesch, 1976). Others describe useful plants as being inherited from a conquered people. The Tikuna recognize that many of the obviously domesticated fruits were already present when they arrived, possibly derived from the Omagua Indians, whom they expelled (Ailton Krenak, União das Nações Indígenas, personal communication), or they negotiated with other tribes (among the Hahaintesu group of the Nambiquara there is a variety of manioc called the “manioc of the Wassuso,” a tribe from the same linguistic group that is practically extinct: in 1985, there was only one elderly woman of this tribe).

The Indians’ diet is based on many plants and their varieties. One commonly finds up to 40 varieties of manioc per indigenous tribe (Kerr and Clement, 1980; Boster, 1984) and witnesses the dynamism these peoples have for renewing these varieties. Manioc is usually propagated by stakes (cloned) which avoids the combination between genotypes. However, if this were always the case, how would recruitment of new varieties for selection exist? How would one justify the great number of varieties observed among the Indians? Contrary to what is found in the literature, manioc is also propagated by seeds (Kerr and Posey, 1984). Plantules derived from seeds bring possibly new combinations resulting from crossing parents propagated by stakes. The Indians of the upper Rio Negro recognize plantules of manioc as derived from seed banks of old plantings in areas that have regenerated and been slashed and burned anew. This recognition is easy because the plantules derived from seeds display their cotyledons (Paulo Sodero Martins, Escola Superior de Agricultura “Luiz de Queirós,” personal communication).

The Hahaintesu group of the Nambiquara in 1985 possessed only nine varieties of macaxeira, almost all recently introduced—after their areas were demarcated and their extermination by whites was interrupted. The increase in varieties is an indication that peace may be giving the Indians a chance to recompose their genetic collection of plants (Boster, 1984). Generally, as in the case of birth control, the administration of genetic resources is directed by women (Kerr and Posey, 1984). In the case of the Hahaintesu group of the Nambiquara, an elderly Indian woman (as mentioned earlier, the last of the Wassuso, married to the Indian Papai) possessed a great knowledge and love for agricultural plants, and others frequently brought her new plants and varieties. This woman would plant them in different environments and evaluate them. Those that were approved she handed over to the rest of the tribe for planting.

This Indian woman and her garden fulfilled what would be the role of “curatorship and experimental station for genetic resources” in our society.

The Amazonian Indians highly value aesthetics, as expressed through their clothing, body decorations, and basketry, which are generally exuberant in both form and color. This sentiment is also reflected in the plants and varieties domesticated by them. Examples include peach palm (Bactris gasipaes Kunth) (Fig. 6A) and peppers of the genus Capsicum (Fig. 6C) and cubiu (Solanum sessiliflorum Dunal) (Fig. 6E), which have a large variety of colors, sizes, and shapes. Among Indians, selection is also made by productivity and taste. Peach palm, one of the best domesticated plants of the humid tropics, can produce up to 50 tons per hectare, with varieties selected for flour or oil and others that easily release the skin and have a refined flavor for cooking; there are also varieties with and without spiny trunks (Clement, 1988; Clement and Mora Urpi, 1987). Cubiu (S. sessiliflorum Dunal), a fruiting plant similar to tomato, produces up to 100 tons per hectare and has varieties with eight different shapes ranging from fusiform to semi-rounded, with colors ranging from yellow to purple (Silva Filho et al., 1993, 1999).

New varieties are also achieved through barter, raiding, and dowry. Groups of friendly Indians visiting from other tribes take varieties as presents. Rival Nambiquara trade varieties in a ritual that is difficult to understand. They pep at each other from opposite sides of a river, without exposing themselves so as not to be killed; they sing and call out challenges. Suddenly, they gain confidence and cross the river to leave presents that sometimes are repaid. After the exchanges they go back to being mortal enemies. Among the Alanteus group of the Nambiquara, after marriage is consented to for a woman, some compensation is required from the groom, which may take the form of labor, a bicycle, a radio, or some other article or group of articles. When the debt is paid, the husband takes his wife and with her some plants, which introgresses genotypes from other tribes. Kidnapping is very common among the indigenous tribes of Amazonia (Black et al., 1991). They kidnap children (the chief of the Ureú-au-au Indians, shown in a 1970s documentary about colonization in Amazonia along the BR-364 highway in the state of Rondônia, was a white man without any cultural traces of white civilization). Women are kidnapped, which seems to be a common practice among the Indians that does not annoy the kidnap victim for very long. Among the Hahaintesu group of the Nambiquara, we saw a kidnapped woman who was imprisoned in a straw hut from which escape would have been easy and was
guarded by her kidnapper, who stayed nearby. During the first days, she was aggressive toward those who approached her; two days later, we saw her laughing among the other Indian women. I later learned that she returned to the agricultural plots of her original tribe and brought back new varieties of manioc, and thus increased her standing with her new tribe.

Domesticated plants return to their wild state if the selective human pressure is relaxed (Harlan, 1975); therefore, there is a close relation between humans and domesticated plants. This relation is so intimate that gene introgression observed in Indian populations in Amazonia (Black et al., 1991) may be associated with the introgression also found among populations of domesticated plants, such as is observed in peach palm (B. gasipaes Kunth) (Picanço et al., 1999) and in caiauê (Elaeis oleifera Cortés) (E. Barcelos, Embrapa Amazonia Ocidental/Manaus, personal communication, and data partially published in Barcelos et al., 1999).

Among the Tikuna in the region near Leticia, Colombia, there is a tribal custom that reinforces plant selection: “All fruit or other edible product that is especially large or flavorful is divided among the tribe and seed planted.” This process practiced over many generations led, by massal selection, to an increase in desirable alleles, giving rise to the domestication of fruiting species such as peach palm (B. gasipaes Kunth; Fig. 6a), biribá (Rollinia mucosa Bail.; Fig. 6b), and abiu (Pouteria
III. THE SELECTION, FIXATION, AND STATUS OF AMAZONIAN PLANT DOMESTICATION

A. Reproductive Isolation

For one population to be different from another, it must fragment off from the original and in some manner maintain itself distinct in genetic terms: in other words, it should have very low or nonexistent chances of crossing with individuals of the population from which it originated. Any reproductive barrier is capable of permitting the fragmentation of a population and its differentiation into sub-populations. Many isolation mechanisms are known from cultivated plants: geographic and ecological separation, differences in time of blooming, self-fertilization, translocation races, polyploid races, gametophytic factors, cryptic chromosomal differences, and meiotic irregularities (Harlan, 1975). These mechanisms may occur simultaneously and be perfect, and if the sub-populations meet someday some degree of hybridization may occur. Sub-populations with potential for intercrossing are called varieties. When the isolation is perfect and there are no longer possibilities for a natural cross to occur, then with the generation of fertile progenies a new species has been formed.

B. The Environment

Practices capable of inducing modifications in the structure of a population are intimately related to the physiognomy of the site and to the ecological restrictions that each environment offers. In Amazonia, although there is a range of ecosystems (for climate, see Salati, 1985; for soils, see Jordan, 1985; and for vegetation, see Fries and France, 1985) and a varied number of singular strategies used by the Indians who practice agriculture on lands with very low trophic support—such as the use of barreiras near the rivers (Hill and Moran, 1983) and in the savannas, with the formation of “managed islands” or âpês of the Kayapó (Anderson and Posey, 1989)—agriculture in indigenous communities is basically practiced on (i) upland areas (terra firme) covered by forests and (ii) the river floodplains (várzeas).

On terra firme covered by forest, agriculture is practiced in cycles of felling–burning–planting–fallowing, also called slash-and-burn or swidden fallows. This is the most applied system of agriculture in Amazonia. A site with exuberant forest is located and the understory trees are cut first, followed by the large ones. Three months later, fire is applied. The ashes and organic matter formed by decomposing tree trunks nourish the soil for at most 4 years. When these soils are exhausted, cultivation of annual plants is abandoned and a new cycle must begin. Among Indians and many caboclo communities, this is not a complete abandonment because there is a subsequent recuperation stage of planting with useful perennial plants (Denevan et al., 1984). This system of apparently migratory agriculture has been very efficient in improving plants because each plot receives some seeds of selected plants. Because the plots are relatively isolated from each other as if islands within the forests and the majority of plants cultivated in them do not occur inside the forest, the latter serves as a barrier to gene flow (Clement, 1987) (Fig. 1). On terra firme, the soils are generally poor and need fertilizing, generally with the ashes resulting from the felling and burning of the original forest that covered them. On terra firme there are also areas that are cultivated continuously, although they are very rare. These are the terra-preta-do- índio (Indian black earth) anthropomorphic soils that are very fertile due to having been enriched by organic detritus deposited over countless generations by the Indians who inhabited them (Smith, 1980) (Fig. 7). These areas vary from 0.5 to 5 ha in area and due to their fertility are preferred even today by white people for planting. Along the Andirá and Paraná rivers, in the municipality of Barreirinha, AM, Brazil, sites with Indian black earth area are abundant. Some of them can be found in the municipal seat and in the communities of Araiú, Freguesia, Maloca, Terra-Preta-do-Limão, Terra-Preta-do-Castanhal, and Pedras Atalaia (Francisco das Chagas Rodrigues Junior and João Carlos dos Anjos Pimentel, residents of Barreirinha, personal communication and author’s observations).

Agriculture on periodically flooded lands is practiced on high and low várzeas and on riverbanks in cycles dictated by seasons of high and low waters. Despite this inconvenience, these lands are highly fertile and demand little labor for clearing and planting. Annual plants and a plantain-type of banana (Musa paradisiaca L., possibly the variety known as pacovi) with small bunches and an extremely short productive cycle are
cultivated (Cavalcante, 1991). On the high várzeas, where flooding rarely occurs, perennial plants that can survive some time with roots submerged and annual plants are cultivated. With some risk, perennial plants typical of terra firme are also cultivated on high várzeas (Smith, 1996).

These two groups of ecosystems have caused a great division between groups of species and varieties of domesticated plants. For the plants most domesticated among Indians, such as manioc (M. esculenta Crantz), corn (Z. mays L.), and lima bean (P. lunatus L.), short-cycle varieties were developed capable of being cultivated on the várzeas before the floods. Although plantain banana (M. paradisica L.) is an exotic and perennial species, the Indians have also selected varieties and cultivation systems of it that are adapted to the várzea. For plants propagated by stakes, such as manioc (M. esculenta Crantz), when floods arrive the young plants are removed from the soil and placed in a canoe or higher site for planting when the waters recede. In the case of banana plants, marombas (wooden structures elevated above the water level) are made in which the young plants are placed with a little soil and fertilizer so that they are sufficiently grown to be planted when the waters recede. There are also the tesoros of Marajó Island, which are hillocks made by the Indians in periodically flooded areas, where possibly corn was cultivated (Roosevelt, 1991). There is insufficient information about how these tesoros were cultivated and what species were utilized, but presumably there must have been a great food production capacity capable of having maintained a quite dense population. It is estimated that at the time of discovery, on terra firme there were 0.2 inhabitants per square kilometer and on the várzea 14.6 (Denevan, 1976)—numbers that are higher than those currently observed in Amazonia.

C. The Management System

The management system causes environmental changes that can influence the demography and genetic structure of a population. For there to be selection and consequent domestication, there must be differentiated reproduction and survival for the genotypes selected; therefore, management and domestication are two intimately linked factors. Forms of management generally influence domestication and it is through them that I intend to explain this process among Amazonian plants. Generally, the more elaborate the management demands of a plant, the greater its degree of domestication. In the case of Amazonian plants, depending on the degree of management to which they may be submitted, they may be classified as (i) collected, (ii) protected or planted with little human interference, and (iii) planted
with high human interference. There is also another distinct group—those managed unconsciously due to interference in the environment by humans with other purposes or due to dispersion of seeds through feces and/or discarding of non-edible parts.

Plants that are simply collected are left in their natural habitat without selection for any characteristic, and even if the fruits collected from more productive and better plants should be more disseminated there is no reproductive isolation and presumably the population maintains itself stable. The word “presumably” is used because no matter how subtle, human occupation is capable of introducing some modification in the natural populations of plants. The reduction, through hunting, of frugivorous animals such as agoutis (Dasyprocta leporina L., D. pyromelana Wagler, and D. fuliginosa Wagler), pacas (Agouti paca L.), and bats, which prey on and disperse seeds, may favor the increase in density of a certain species where the more productive plants might have greater chances of increasing the frequency of their alleles in a given location due to the increase in their density. The group of useful plants collected in the forest for medicinal purposes, food, or some other use may possibly comprise more than 1000 species.

The plants protected or planted with low human interference are those that occur naturally in the habitat and that are preserved during settling and later cared for or planted along trails or in clearings but that receive little care (Table I). Examples of plants protected from burning include the rucamá (Astrocaryum aculeatum G. F. W. Meyer and Astrocarum vulgare Mart.) and inajá [Maximiliana maripa (C. Serra) Drude] palms. Plants under low human interference are those generally cast along trails and in clearings, such as is done by the Kayapo Indians, who collect seeds and plantlets of useful trees and plant them along trails on which they usually travel or in clearings opened by tree falls (Kerr and Posey, 1984). They generally belong to the same habitat and although they undergo some selection, reproductive isolation, as in the previous case, is not efficient because these species also occur in forests and can interbreed. Occasional progress in the domestication process may be achieved because the selected plants, when planted together, have greater chances of intercrossing, which facilitates the maintenance of a higher frequency of the alleles selected. Since isolation is not very efficient, progress is very slow. This group is prone to receive greater human influence than is the case with the previous group because their density and fecundity are in most cases favored by the reduction of competition with the other plants. The reproductive isolation of these species, although more efficient than in the previous group, is also not very efficient and there is no directional selection of the plants; they are simply spared.

The dispersion of those planted with high human interference (Table II) is consciously done by humans and during their entire productive cycle they receive care and make up three distinct groups, according to the position they assume in the ecological succession of forest regeneration after burning. This succession is directed by humans toward a final floristic composition with a greater density of useful species than would be found in a forest without any history of human influence (Denevan et al., 1984; Anderson and Posey, 1989; Balsee, 1989). The first groups is that of the most domesticated species: These are annual plants that are colonizers and fast growers, with a high productivity of edible biomass. They are: manioc (M. esculenta Crantz), corn (Z. mays L.), and more recently, beans of the genus vigna. The second group is formed by semi-perennials species, sown after the plants in group 1: sweet potato (Ipomoea batatas L.) Lam., arrowroot (Maranta arundinacea L.), peanut (Arachis hypogaea L.), Lima bean (P. lunatus L.), cara yam (Dioscorea trifida L.), and pineapple (Ananas comosus (L.) Merrill). Those of the third group are arboreal species, generally fruit producers with rapid growth and the capacity for growing well in soils already leached by rains and exhausted by annual crops. These plants have a high capacity for production and deposition of their leaf biomass. They live for 80 years at the most and cannot bear competition with other trees of the forest. They are peach palm (Bactris gasipaes H. B. K.), cucuru (Pourowma cecropiojofia Mart.), umari do Amazonas (Foraquisa sertae Tat.), and biribá Rollinia mucosa (Jacq.) Baill. Also in this group are other trees that do not have a great capacity for regenerating the soil but are still able to grow well in poor soils: guava, cocoa, and abiu. After the planting of these categories, there is the reappearance of the group of plants protected/planted under low human interference. These make up the group in the final phase of regeneration in the direction of a climax forest. Sometimes, species of this group are planted, whereas others are only protected and require more specific ecological conditions, with some shading in the juvenile phase, humidity, and protection from heat. These plants are capable of competing with other trees of the forest and are part of the final climax stand achieved by the system: They are Brazilnut (Bertholletia excelsa H & B) and the piquiaceiros (Carinca villosum Pers.), amapa doce (Brosimum potabile Ducke), macaranuba (Manilkara...
TABLE I
Protected or Planted Plants at Low Levels of Human Care

<table>
<thead>
<tr>
<th>Names: English/Portuguese/Spanish</th>
<th>Latin name</th>
<th>Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papaya/na/ambah</td>
<td>Carica papaya L.</td>
<td>Leaves as vegetable</td>
</tr>
<tr>
<td>Guava/mambe</td>
<td>Psidium guajava L.</td>
<td>Fruits, fibers from leaves</td>
</tr>
<tr>
<td>Brazil nut/para</td>
<td>Bertholletia excelsa Humb. Et Bonpl.</td>
<td>Nuts; infusion of fruits as a medicinal fruit</td>
</tr>
<tr>
<td>Punta</td>
<td>Byrsonima crassifolia L. Rich.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Mora</td>
<td>Myrciaria dubia</td>
<td>Leaf as food</td>
</tr>
<tr>
<td>Guava</td>
<td>Psidium guajava L.</td>
<td>Fruit, leaves as vegetable</td>
</tr>
<tr>
<td>Brazil nut/para</td>
<td>Bertholletia excelsa Humb. Et Bonpl.</td>
<td>Nuts; infusion of fruits as a medicinal fruit</td>
</tr>
<tr>
<td>Maize</td>
<td>Zea mays L.</td>
<td>Seed as a nut</td>
</tr>
<tr>
<td>Coconut</td>
<td>Cocos nucifera</td>
<td>Oil from the pulp and endosperm, seed kernel for crafts</td>
</tr>
<tr>
<td>Potato</td>
<td>Solanum tuberosum L.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Cassava</td>
<td>Manihot esculenta Crantz</td>
<td>Root</td>
</tr>
<tr>
<td>番荔枝 (Euphorbia coryana)</td>
<td>Euphorbia coryana</td>
<td>Fruit, trunk for building</td>
</tr>
<tr>
<td>Cacao</td>
<td>Theobroma cacao L.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Black pepper</td>
<td>Piper nigrum</td>
<td>Fruit</td>
</tr>
<tr>
<td>Coffee</td>
<td>Coffea arabica</td>
<td>Fruit</td>
</tr>
<tr>
<td>Tobacco</td>
<td>Nicotiana tabacum</td>
<td>Fruit</td>
</tr>
<tr>
<td>Banana</td>
<td>Musa paradisiaca</td>
<td>Fruit</td>
</tr>
<tr>
<td>Mango</td>
<td>Mangifera indica</td>
<td>Fruit</td>
</tr>
<tr>
<td>Guava</td>
<td>Psidium guajava L.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Pineapple</td>
<td>Ananas comosus</td>
<td>Fruit</td>
</tr>
<tr>
<td>Guava</td>
<td>Psidium guajava L.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Cassava</td>
<td>Manihot esculenta Crantz</td>
<td>Root</td>
</tr>
<tr>
<td>Potato</td>
<td>Solanum tuberosum L.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Cassava</td>
<td>Manihot esculenta Crantz</td>
<td>Root</td>
</tr>
<tr>
<td>Tobacco</td>
<td>Nicotiana tabacum</td>
<td>Fruit</td>
</tr>
<tr>
<td>Coffee</td>
<td>Coffea arabica</td>
<td>Fruit</td>
</tr>
<tr>
<td>Black pepper</td>
<td>Piper nigrum</td>
<td>Fruit</td>
</tr>
<tr>
<td>Mango</td>
<td>Mangifera indica</td>
<td>Fruit</td>
</tr>
<tr>
<td>Pineapple</td>
<td>Ananas comosus</td>
<td>Fruit</td>
</tr>
</tbody>
</table>

continues
<table>
<thead>
<tr>
<th>Names: English/Portuguese/Spanish</th>
<th>Latin name</th>
<th>Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>—/bacaba açu, bacaba verdadeira/tungurahua, camou, manoco</td>
<td>Oenocarpus bacaba Mart.</td>
<td>Fruit, truck for building, leaves to cover houses</td>
</tr>
<tr>
<td>—/bacaba-de-leque/ —</td>
<td>Oenocarpus distintus Mart.</td>
<td>Fruit, truck for building, leaves to cover houses</td>
</tr>
<tr>
<td>—/bacah/bapora</td>
<td>Oenocarpus mapora Karsten</td>
<td>Fruit</td>
</tr>
<tr>
<td>—/bacai/bincas/ —</td>
<td>Oenocarpus minor Mart.</td>
<td>Fruit</td>
</tr>
<tr>
<td>—/feijão macaco/ —</td>
<td>Pachyrhizus tuberosum Lam.</td>
<td>Tubers as vegetable</td>
</tr>
<tr>
<td>—/pajúri-da-mata/ —</td>
<td>Farinaria moniana Auhl.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Wild passion fruit/maracujá suspiro, maracujá-de-rato, maracujá-do-mato/ —</td>
<td>Passiflora niitida H. B. K.</td>
<td>Fruit</td>
</tr>
<tr>
<td>—/camapu/ —</td>
<td>Physalis angulata L.</td>
<td>Fruit as a vegetable, roots as medicine</td>
</tr>
<tr>
<td>Backuri/bacuri/bacuri, matabazona</td>
<td>Platonia insignis Mart.</td>
<td>Fruit</td>
</tr>
<tr>
<td>—/umari, mara gordo/ —</td>
<td>Poraqueria parensis Ducke</td>
<td>Fruit</td>
</tr>
<tr>
<td>—/umari, umari do amazonas/ —</td>
<td>Poraqueria sericea Tul.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Canistel/rutie/calmo, canistel, yema de hueco</td>
<td>Pouteria macrophylla (Lam.) Eyma</td>
<td>Fruit</td>
</tr>
<tr>
<td>—/ucuquei/ —</td>
<td>Pouteria tucupi Fries et Schultes</td>
<td>Fruit</td>
</tr>
<tr>
<td>—/araça-pera/ —</td>
<td>Psidium acutangulum DC</td>
<td>Fruit</td>
</tr>
<tr>
<td>Guava/goiaba/guayaba</td>
<td>Psidium guajava L.</td>
<td>Fruit, medicinal tea from young leaves for diarrhea</td>
</tr>
<tr>
<td>—/aranha, araça-do-campo/ —</td>
<td>Psidium guajense Swartz</td>
<td>Fruit</td>
</tr>
<tr>
<td>—/bacuripi/iosi/ —</td>
<td>Rheedia brasiliensis (Mart.) Pl. et Tr.</td>
<td>Fruit</td>
</tr>
<tr>
<td>—/bacuripir/pacuripuzu</td>
<td>Rheedia macrophylla Pl. et Tr.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Hog plum, plum bush/tapereba, cajamirim, caja/monimba</td>
<td>Spondias mombin L.</td>
<td>Fruit, infusion from the bark for leprosy</td>
</tr>
<tr>
<td>—/pitombas/caráys-rola</td>
<td>Talsa escallenta (St. HIl.) Radlk.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Cocoa. chocolate tree/cacao/cacao</td>
<td>Theobroma cacao L.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Cupuaçu/cupuçu, cupa, cupuaçu</td>
<td>Theobroma grandiflorum (Willd. ex Spreng.) Schum</td>
<td>Fruit</td>
</tr>
<tr>
<td>verdadeiro/copoezur</td>
<td>Theobroma speciosum Willd.</td>
<td>Fruit</td>
</tr>
</tbody>
</table>

4 They receive some influence on the genetic structure of their population. The domestication process is accepted as being in progress but at a very early stage. Some of these species were possibly undergoing a more advanced process of domestication, but a possible retrogression on their status is accepted due to loss of varieties or by the people who used them (table structured from Denovan et al., 1984; Clement, 1999; Cavalcante, 1976; Smith, 1999; Smith et al., 1992).

jaguar,” “the tapir,” etc., and the individuals of each “people” have a specific cemetery where they should be buried. When Indians die, their companions take them to be buried in the cemetery of their respective people and thus they are able to visit the old sites, accompany the development of the “islands,” and maintain their localization in tribal memory.

The groves of babacu palms (Orbignya martiana Barb. Rodr.), which occupy approximately 200,000 km² in Brazil and support approximately 450,000 subsistence-level households in the north, northeast, and central west regions of Brazil through their multiple products (fibers, thatch, basketry, construction, palm heart and fruits for oil, charcoal, and animal feed) are an example of unconscious management. This palm tree regenerates profusely when the primary forest is burned. The plant survives fire well because it has a unique germination process. Upon germinating, the seed sends its hypocotyl underground to a depth of about 10 cm and then growth of the gemellae begins. It remains inside the soil, putting its leaves out aboveground. When the forest is burned, the plants that are still in this phase lose only their leaves but are capable of regenerating since the main bud is well protected from the fire. Due to the reduction in competition with the other plants that were destroyed by fire, the babacu plants end up dominating the environment (May et al., 1985a, b). Another example of unconscious
<table>
<thead>
<tr>
<th>Names: English/Portuguese/Spanish</th>
<th>Latin name</th>
<th>Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cashew/ceju/marabín</td>
<td><em>Anacardium occidentale</em> L.</td>
<td>Fruit, infusion from the bark for leucorhoea</td>
</tr>
<tr>
<td>Pineapple/abacaxi, ananás/piña</td>
<td><em>Ananas comosus</em> (L.) Merril.</td>
<td>Fruit, fiber from leaves</td>
</tr>
<tr>
<td>Soursop/graviola/guanábana, anona</td>
<td><em>Annona muricata</em> L.</td>
<td>Fruit, infusion from the leaves for snakebite</td>
</tr>
<tr>
<td>Peanut/amendoim/mani</td>
<td><em>Arachis hypogaea</em> L.</td>
<td>Seeds</td>
</tr>
<tr>
<td>Peach palm/pupunha/pépiba, pijnayo, chonaduro</td>
<td><em>Bactris gasipaes</em> H. B. K.</td>
<td>Fruit, trunks for building</td>
</tr>
<tr>
<td>Annatto/urucum/achioté</td>
<td><em>Bixa orellana</em> L.</td>
<td>Arielle from seeds as colorant and condiment, largely used for painting the body</td>
</tr>
<tr>
<td>Guinean arrow root/aria/dale dale</td>
<td><em>Calathea alouia</em> (Askl.) Lundl.</td>
<td>Tubercle</td>
</tr>
<tr>
<td></td>
<td><em>Canna edulis</em> Ker.</td>
<td>Tubercle</td>
</tr>
<tr>
<td>Sweet pepper/pimenta-doce/aji</td>
<td><em>Capsicum annum</em> L. var. minimum</td>
<td>Fruit as a vegetable, condiment, medicinal</td>
</tr>
<tr>
<td></td>
<td><em>Capsicum chinense</em> Jacq.</td>
<td>Condiment</td>
</tr>
<tr>
<td></td>
<td><em>C. brasiliannum</em></td>
<td>Condiment</td>
</tr>
<tr>
<td></td>
<td><em>C. baccatum</em> L.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Papaya, paw paw/mamao/lochosa, melon zapote</td>
<td><em>Carica papaya</em> L.</td>
<td>Liana as vegetable</td>
</tr>
<tr>
<td></td>
<td><em>Cissus gengyloides</em> Burch</td>
<td>Fruit domestic appliances</td>
</tr>
<tr>
<td>Calabash gourd/cueirica/pati</td>
<td><em>Crescentia ejugae</em> L.</td>
<td>Tubercle</td>
</tr>
<tr>
<td>Yam/cará/sicira/papa</td>
<td><em>Dioscorea trifida</em> L.</td>
<td>Fruit</td>
</tr>
<tr>
<td></td>
<td><em>Duguetia stenantha</em> R. E. Fries</td>
<td>Fruit</td>
</tr>
<tr>
<td></td>
<td><em>Eugenia stiptata</em> McVaugh</td>
<td>Fruit</td>
</tr>
<tr>
<td>Genipa/genipapo/lunão</td>
<td><em>Genipa americana</em> L.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Cotton/algodão brabo/algodao</td>
<td><em>Geernetium barbadense</em> L.</td>
<td>Fiber</td>
</tr>
<tr>
<td>Sweet potato/huata doce/camote</td>
<td><em>Ipomoea batatas</em> (L.) Lam.</td>
<td>Tubercle</td>
</tr>
<tr>
<td>Manioc/mandioca/yuca</td>
<td><em>Manihot esculenta</em> Crantz</td>
<td>Tubercle, leaves as vegetable</td>
</tr>
<tr>
<td>Arroz real/arroz real</td>
<td><em>Mamey arundinacea</em> L.</td>
<td>Tubercle</td>
</tr>
<tr>
<td>Tobacco/tabaco/tabaco</td>
<td><em>Nicotiana tabacum</em> L.</td>
<td>Leaves that are smoked or smashed as stimulant</td>
</tr>
<tr>
<td>Passion fruit/maracujá, maracaju, verdadeiro/granadilla</td>
<td><em>Passiflora edulis</em> Sims.</td>
<td>Fruit for a refrigerant and as a carminative</td>
</tr>
<tr>
<td>Giant granadilla/maracujá-aco, maracujá mamão/granadilla, granadilla real</td>
<td><em>Passiflora quadrangularis</em> L.</td>
<td>Fruit</td>
</tr>
<tr>
<td>Guarana/guarana/cupana</td>
<td><em>Paulinia cupana</em> H. B. K. var sorbilis (Mart.)</td>
<td>Seeds for infusion used as refrigerant and stimulant</td>
</tr>
<tr>
<td></td>
<td><em>Ducke and var. typical</em></td>
<td>Beans</td>
</tr>
<tr>
<td>Lima bean/feijão lima/—</td>
<td><em>Phaseolus lunatus</em> L.</td>
<td>Fruit</td>
</tr>
<tr>
<td></td>
<td><em>Pouteria caimito</em> (Ruiz et Pavon) Radlk.</td>
<td>Fruit</td>
</tr>
<tr>
<td></td>
<td><em>Quararibea cordata</em> (H. B. K.) Vischer</td>
<td>Fruit</td>
</tr>
<tr>
<td>Supote/sapota-do-peru/zapote, name colorado</td>
<td><em>Rollinia mucosa</em> (Jacq.) Baill.</td>
<td>Fruit as a vegetable</td>
</tr>
<tr>
<td></td>
<td><em>Solanum sessiliflorum</em> Dunal.</td>
<td>Leaves as vegetable</td>
</tr>
<tr>
<td></td>
<td><em>Spilanthes acmella</em> (L.) Murr.</td>
<td>Fruit, toasted seeds as a nut</td>
</tr>
<tr>
<td></td>
<td><em>Theobroma bicolor</em> Humb. &amp; Bonpl.</td>
<td>Tubers and leaves as a vegetable</td>
</tr>
<tr>
<td></td>
<td><em>Xanthosoma sagittifolium</em> (L.) Schott.</td>
<td>Seed as a grain, exes as medicinal</td>
</tr>
<tr>
<td></td>
<td><em>Z. mays</em> L.</td>
<td></td>
</tr>
</tbody>
</table>

*The domestication process is accepted as being in an advanced stage (table structured from Denovan et al., 1984; Clement, 1999; Cavalcante, 1976; Smith, 1999; Smith et al., 1992.)*
management is the sowing of plants collected near the huts in piles of trash (Fig. 8). This process is possibly what happened at the beginning of agriculture and the related sedentarism (Harlan, 1975). The plants thus unconsciously cultivated end up being cared for and recruited within an agricultural system.

All these systems of management and consequent selection have been applied since before the discovery to the present by the indigenous peoples, later by the caboclos, and recently by Japanese immigrants to Amazonia (Saragossi et al., 1988). This has maintained the preservation of already domesticated plants and their varieties and even the appearance or improvement of others. The Japanese farmers of Tomé-Açu and Castanhal and their descendants, using only massal selection, have been able to achieve better cultivars of cupuaçu (Theobroma grandiflorum Schum.) and açai (Euterpe oleracea Mart.). Vitola softwood from the estuary islands of the Amazon River, which in the past was predated as a source of extractivism, today is wisely managed as a consequence of market forces, and thus is also undergoing domestication by caboclos.

IV. FINAL COMMENTS

A domesticated plant, it should be remembered, is in this state because humans brought about its domestica-

tion. If human-induced selection pressures cease, these plants will not survive, or their descendants will need to recompose their defenses to face competition in a natural environment and in doing so generally lose the attributes selected by humans (Harlan, 1975). Many of the indigenous cultivars were selected not only for flavor or productivity but also for mystical reasons and are thus susceptible to disappearance as a result of changes in habits. In Amazonia, the Indians are being besieged by pressures toward deculturation from the "civilized" peoples who are occupying the region. In this process much of the existing knowledge on plant use is lost. For medicinal plants, this loss is even greater since most of the time their usefulness is not visible, as in the case of fruiting species. With the gradual loss by the Indians of the properties and uses of the plants, everything may simply be considered "jungle." Although Amazonia is considered one of the largest repositories of genetic resources in the world, there are still few efforts in recovery, collection, and evaluation of useful plants among the Indians. If humanity has these marvelous plants, delights in their flavor and their appearance, wears clothes and uses domestic utensils made from them, and uses them to cure pains and diseases, this is because the traditional peoples of Amazonia, very simple people, discovered them, selected them, and maintained them until now. The society that calls itself civilized still has not sufficiently recognized
this feat and its value, but it must some day find mechanisms for doing so or will regret its lack of action in the same way as we today regret the loss of the Alexandria library, burned down in the name of a culture that judged itself the owner of superior values.

Acknowledgments

I am enormously grateful to Indian agent Sydeney de Almeida and Nurse Maria Aurora (Lia) who kindly accompanied my group on our travels among the Namibiquara Indians and translated for us their customs and information about plants and myths. I also thank all the Indians and Amazon caboclos for having shared with us a little of their wonderful knowledge about their useful plants.

See Also the Following Articles

AGRICULTURE, TRADITIONAL • BREEDING OF PLANTS • DOMESTICATION OF CROP PLANTS • EDIBLE PLANTS • GENETIC DIVERSITY • INDIGENOUS PEOPLES, BIODIVERSITY AND

Bibliography

Environments: Amazonia (G. T. France and T. E. Lovejoy, Eds.),

na composição de quintais de três localidades de terra firme do
estado do Amazonas. In Ethnobiology: Implications and Applications
(D. Posey and W. L. Overal, Eds.). Proceedings of the first interna-

Setz, E. Z. (1991). Animals in the Xambiquara diet: Methods of
collection and processing. J. Ethnobiol. 11, 1-22.

variability of economic characters in 30 accessions of Cubiu (Solan-
umum sessiliflorum Dunal) evaluated in central Amazonia. Rev. Bras-
sileira Genet. 16(2) 409-417.

Silva-Filho, D. F., Noda, H., Clement, C. R., and Machado, F. M.
(1999). Cubiu (Solanum sessiliflorum): Conservação e melhoro-
mento de etnovariedades na Amazônia. Genet. Mol. Biol. 22(3),
708.

Smith, N. J. H. (1980). Anthropology and human carrying capacity in


Tropical Forests and Their Crops. Comstock/Cornell Univ. Press,
Ithaca, NY.


1987). Sâbitasu e Kwalhakanasu, duas fruteiras dos índios
Xambiquaras com potencial para a domesticação. Acta Amazônica
16/17, 1925.
CRUSTACEANS

Marjorie L. Reaka-Kudla
The University of Maryland

1. Summary and Definition of Biodiversity in Crustacea
2. The Setting
3. Biodiversity of Fossil and Modern Crustacea

GLOSSARY

antennules, antennae (pl.) Antennules are the first pair of sensory appendages on the head, and antennae are the second pair of sensory appendages on the head of crustaceans.
benthic, epibenthic Living at the bottom of the sea; epi = upon, thus epibenthic organisms live on top of sediment or hard substratum.
brackish Less salty than sea water.
chelate, subchelate, chela (s.), chelae (pl.) Chelate means having a pincher-like claw (usually due to an extension of the second from terminal segment beside the terminal segment, forming the claw). In subchelate forms, the terminal segment merely folds back on the second from terminal segment. The chela is the clawlike appendage.
commensal, symbiotic Living in association with (e.g., on or in) another organism.
crepuscular Active in dim light conditions, such as dusk or dawn.
dorsal, ventral Dorsal part of the organism is the "back," on the opposite side of the body from the mouth; ventral is the same side of the body as the mouth and legs for a crustacean.
extant Living.
furca (s.), furcae (pl.) Forked structure, usually associated with the telson in crustaceans.
gonochoristic Individual is either a male or a female throughout life.
hypersaline More salty than seawater.
interstitial Living among sand or silt grains; usually refers to an aquatic benthic environment, the intertidal, or in groundwater.
invertebrates Animals (usually used in reference to phyla or major taxa) that lack backbones.
macrobiota Organisms large enough to see, usually greater than 1 mm.
median, medial Along or toward the midline of the body.
phylopod A leaflike appendage.
pleon Abdominal appendage in crustaceans.
polyphylectic A group of species or lineages that have independent origins instead of being descended from a common ancestral group.
protogynous, protandrous Different types of hermaphroditism (in contrast to gonochoristic). Individuals in protogynous species develop into a reproductive female early in life, then change sex and become a reproductive male later in life. Individuals in protandrous species are functional males first, then females.
rami (pl.) A cuticular extension (usually relatively long and thin, can be leaflike) from the body (e.g., caudal...
rami = extention from the posterior segment of the body, usually paired in crustaceans).

setae (pl.) Bristle- or hairlike extensions of the cuticle that are characteristic of arthropods and related phyla.

taxon (s.), taxa (pl.) A group of organisms with shared similarities at any level of a taxonomic scheme (could be a species, genus, family, order, etc.)
telson Most posterior segment of the body in crustaceans, usually flattened and often armored.
thoracopod Appendage on the thorax (paired in crustaceans).

uniramous, biramous, triramous, polyramous An extension of the body (usually an appendage) having one, two, three, or many branches (e.g., triramous antennae have three main branches or flagellae).

vestigial An organ or structure that is present only in rudimentary form (probably lost from a previous condition).

---

TO UNDERSTAND THE BIODIVERSITY—and prospects for survival—of contemporary Crustacea, we need to know the number of species and higher taxa, and the range of architectural variation, habitats used, ecological roles, life history traits, and biogeographical patterns found in the group as it has developed over the last half-billion years.

1. SUMMARY AND DEFINITION OF BIODIVERSITY IN CRUSTACEA

In the last two decades of the twentieth century, it became evident to both the scientific and much of the lay public that the health of the earth’s living biota—the foundation of human existence—is seriously threatened by environmental changes, mostly generated by the exponentially increasing activities of humans. This realization was made more poignant by the simultaneous recognition that our knowledge of the diversity of the world’s animals, plants, fungi, and microorganisms is woefully inadequate even to assess the amount of diversity present and the degree of risk faced by different components of the global biota (Wilson, 1985, 1992).

The term “Biological diversity” and its shortened form “biodiversity” were introduced by Norse and McManus, (1980), Lovejoy (1980), and the National Forum on BioDiversity, held at the Smithsonian Institution (1986). Whereas “biological diversity” was first defined to include the concepts of genetic diversity (variability within species) and ecological diversity (number of species in a community of organisms; Norse and McManus 1980), the term “biodiversity” was formalized and popularized in the Proceedings of the National Forum (Wilson and Peters, 1988) to include an appreciation for the range of all of the different levels and types of biotic diversity found on earth, including the processes that generate and maintain diversity, and is thus followed here. “Diversity” sometimes implies concepts about the degree of evenness in abundance among species as well as number of species, but, for simplicity and convenience, the terms “diversity” and “biodiversity” will be used interchangeably here to indicate the number of species or higher taxa in a group and to encompass the range of structural variation, habitat characteristics, ecological roles, life history traits, and biogeographical characteristics of particular groups of organisms. The diversity of these features, including the uniqueness of certain adaptations or lineages, often can be comprehended best in the context of the ecological and evolutionary processes that generate these patterns of variation.

To properly evaluate biodiversity in Crustacea, we need to know the range of major ways of life of different groups of crustaceans, as indicated in their body plans. These architectural differences are reflected in the number of higher taxonomic groups within taxa and provide insights into the diversity of ecological processes in which the group participates.

Ecological diversity includes the range of habitats that the group occupies, whether the group lives in unique environments or species assemblages, and the diversity and importance of the ecological roles that taxa within the group play in their natural communities. For example, are the species low on the food chain (thus supporting the diversity of the rest of the community), do they affect community structure as top predators, do they determine diversity of the community by engineering the physical environment? Other important components of biodiversity include the phylogenetic history and uniqueness of lineages. Some groups of Crustacea represent the only surviving remnants of lineages that have persisted for half a billion years. Because of their early derivations, such groups are likely to be unique in their genetic, morphological, physiological, or ecological adaptations. Such a historical legacy merits special conservation attention. Similarly, the diversity of life history patterns is intimately related to genetic variability within and among populations. Some crustaceans show rapid maturation, short life spans, and massive reproductive effort, while others live for extended periods (often many years) and are morpHO-
logically and behaviorally adapted for strong interactions with other species. Some produce a few large offspring that develop directly into juvenile stages similar to the adults and do not disperse far from the parental population. Other species release huge numbers of tiny offspring, entirely different in form from the adults, that feed for extensive periods in the plankton and disperse great distances. Many of these patterns track within lineages and are correlated with body size. High variability among individuals within populations, frequently associated with wide dispersal of pelagic offspring and wide geographic ranges of species, may enable a population to persist in spite of environmental changes. However, high variability among populations (often associated with lower gene flow and isolated endemic populations) may increase the probability of losing entire genetic complements if local populations are eliminated by environmental change. Distributional patterns, particularly at the species level, are one of the most important and neglected aspects of biodiversity in most groups of organisms, including the Crustacea. The extent of the geographic range of a group, and of species within the group, are some of the most important predictors of extinction in the fossil record (Jablonski, 1991). Organisms that inhabit highly specialized habitats (especially when these habitats are geographically restricted), occupy high trophic levels, or are of large body size are likely to be vulnerable to extinction because of relatively low total population sizes and their dependence on a reliable food supply or special microhabitat (both of which fluctuate in many environments). Understanding the paleontological history of crustaceans is essential for understanding the present configuration and future prospects for crustacean biodiversity. The fossil record indicates that recoveries from major mass extinctions require about 10 million years, rendering the process of diversification irrelevant for the survival of humans and other biota over the next 500 years. However, understanding vulnerability to extinction in the past and in the present rapidly changing environment is crucial for assessing present crustacean biodiversity and sustaining it for future generations.

II. THE SETTING

A. Biodiversity in Major Global Habitats, with Special Reference to Marine Environments

Marine environments, the habitat in which crustaceans originated and which continues to sustain the greatest proportion of crustacean species, contain more higher-level biodiversity (different types of organisms represented, for example, as phyla, classes, orders, families, and genera) than other habitats because life arose in the sea, and many of these ancient lineages are still with us. Of plants, animals, and fungi large enough to see, 43 phyla live in the ocean and 28 live on land. Within the 33 animal phyla, 64% inhabit only the sea while only 5% are exclusively land dwellers, and 90% of all known classes are marine (May, 1994; Ray, 1991). It is important for humans to know and understand the diversity of these marine lineages, with all of their bizarre body plans and adaptations, in order to understand how life evolved and how we fit into it. Of more immediate urgency, marine organisms provide a critical source of protein for humans. This fact will become more important in the 21st century and beyond (especially in the world's tropical countries, which are experiencing rapid population growth and associated environmental degradation). Possibly even more significant in the long run, these marine groups provide a repository of genetic and chemical material that has been evolving, each in its own direction, for more than half a billion years. This genetic and chemical material is much more likely to be novel than the more recent, better-known lineages that invaded land.

Despite its huge expanse and the greater number of higher level groups it supports, the ocean contains fewer known species than the land. Only about 274,000 or 14% of the world's 1.9 million described species are known from the ocean (Reaka-Kudla, 1997). Of these, approximately 200,000 are macrobiota (vertebrates, invertebrates, and plants large enough to see) and 180,000 are macroinvertebrates (groups of marine invertebrates sufficiently big to see). Global freshwater environments contain approximately 44,000 total species (2.4% of all those described), 35,000 species of macrobiota and 26,000 species of macroinvertebrates. There are about 950,000 insects, 80,000 noninsect and noncrustacean (mostly chelicerate) arthropods, and 57,000 other terrestrial macroinvertebrates. Globally, about 100,000 species live symbiotically on or in other organisms (5.3% of all described species); roughly 73,000 of these probably are macrobiota (mostly invertebrates). If we exclude the noncrustacean terrestrial arthropods (chelicerates, insects, and their relatives) because of their unique specializations for life on land, there are approximately 336,000 described macroinvertebrate species in the marine, freshwater, terrestrial, and symbiotic environments. Of these, marine macroinvertebrates constitute 53.6%, freshwater macroinvertebrates 7.7%, terrestrial macroinvertebrates 17.0%, and symbiotic macroinvertebrates
21.7% of the total species. We can use these figures to estimate the numbers of crustacean species that occupy different types of habitats.

There are about 83,000 species of described crustaceans—2.3% of the 1.9 million described species and 12.8% of all macroinvertebrates. If crustaceans inhabit major types of environments in about the same proportions as other nonchelicerate and noninsect macroinvertebrate groups, probably about 23,000 described species of crustaceans live in marine and brackish coastal waters (figures rounded to nearest thousand); this represents about 12% of all species of marine macrobiota. Similarly, if crustaceans have invaded fresh water to about the same degree as other nonchelicerate and noninsect invertebrates, at least 3400 crustacean species likely inhabit fresh waters (rounded to nearest hundred); this represents about 10% of all described freshwater macrobiota. If crustaceans are about equally adapted to land as are other nonchelicerate and noninsect invertebrates, about 7300 crustaceans are terrestrial. This is less than 1% of the global terrestrial macrobiota. Since about 22% of all macroinvertebrate species are symbiotic, probably about 9,500 crustacean species are estimated to be symbionts (about 13% of global symbiotic macroinvertebrates). In fact, parasitism is rampant in some groups of crustaceans. Certainly these figures are underestimates—most likely great underestimates—of the true numbers of species in aquatic and particularly marine environments because these environments are so poorly known. If only 10 to 20% of all crustacean species have been described (discussed later), then there likely are about 200,000 to 450,000 total (known and unknown) species of crustaceans on earth.

The following observations suggest that the numbers of crustacean species, as well as those of other marine phyla, usually are underrepresented. Based on samples of 1597 species of soft-bottom marine macrofauna off the east coast of the North America (255 to 3494 m), Grassle and Maciolek (1992) calculated that the global deep-sea fauna, primarily because of the huge area it occupies, may include as many as 10 million species (mostly polychaete annelids, crustaceans, and mollusks). Most of these species are small and rare (90% of the species sampled comprised less than 1% of the individuals; 28% of the species in the entire fauna were collected only once). Reaka-Kudla (1997) calculated that global coral reefs contain between 1 and 5 million total species (though only about 100,000 described species).

Experts have estimated that about 21% of global crustacean species have been described and that about 26% of global molluscan species are known (Systematics Agenda 2000, 1994). These phyla represent the two most commercially important and thoroughly documented invertebrate groups. Only 10% of the relatively well-known isopod crustaceans collected from shallow sediments in southern Australia had been described (Poore and Wilson, 1993). Because of the great variation in the extent to which different regions of the oceans have been sampled, these scientists estimated that only 5% of marine invertebrates are known from the oceans overall. Calculations suggest that about 10% of the species on global coral reefs have been described. Scientific descriptions exist for about 17% of the species of algae. The microbiota is even more poorly known. Only 1%, 1 to 10%, 4 to 7%, and 2 to 3% of the total species of the viruses, bacteria, fungi, and nematodes have been described.

Although the ancient major groups (phyla, classes) of marine organisms are more different from each other than those on land, species within marine invertebrate groups often look alike, leading to underestimates of their true numbers. Recent molecular and genetic studies have shown that described species frequently include several different, previously unrecognized species (Knowlton, 1993).

Major new groups of organisms are still being discovered in the sea, whereas this is rare on land. At least three new invertebrate phyla (the largest category of related organisms) and many new major groups (classes, orders, families, genera, which represent increasingly smaller groupings of related organisms) of deep-sea organisms, picoplankton, marine viruses, and archaeabacteria (a new kingdom of one-celled organisms that often live in extreme environments) have been described over the past 20 years. In the crustaceans, a new class, the primitive Remipedia, were discovered in caves in the Bahamas and described only in 1981. A new subclasse of the Maxillopoda, the Tantulocarida (tiny ectoparasites of deep sea copepods, ostracods, and isopods) was erected in 1983. In the relatively well known malacostracan crustaceans, the Order Spelaeogrithacea was not discovered until 1957, the Order Amphionidacea recognized only in 1973, and the Order Mictacea established as late as 1985 (Schram, 1986). Crustacean biology, and marine biology in general, remains a realm of discovery.

Aquatic environments are difficult to see, sample, and study. The number of scientists who study invertebrates and the amount of scientific effort devoted to invertebrates other than insects is two orders of magnitude less than that devoted to vertebrates and one order of magnitude less than that devoted to plants (Gaston
and May, 1992). In addition, tropical environments, far from the location of most biologists, have been much less studied than those at higher latitudes; 80% of ecological researchers are based in North America and Europe in comparison to 7% in Latin America and tropical Africa. Identification of species in aquatic habitats is further reduced because of the inability to observe them easily and the frequent requirement for Scuba diving and extensive logistic support.

Moreover, most crustacean and other aquatic taxa are small in body size. Many aquatic species live within crevices, holes, or burrows in coarse (rocks, coral, rubble) and level (mud, sand) substrata. Many are crepuscular or nocturnal to minimize fish predation (Domínguez and Reaka, 1988), making them difficult to see, collect, and study. Small organisms of any environment always are poorly known and studied, but this is more extreme in aquatic environments, particularly in tropical marine environments, because of the remote localities and the difficult conditions for observing and collecting.

It seems likely that major advances in our knowledge of the extent of crustacean biodiversity will continue into the next centuries. These advances are imperiled, however, by a crisis in the number of systematists and taxonomists being trained (Feldmann and Manning, 1992; Reaka-Kudla et al., 1997; Wilson, 1985). Although we know a fair amount about the morphology, ecology, behavior, life history, and development of many of the crustacean groups, we know astonishing little about the facets of their biology that are most critical for assessing their biodiversity status, particularly the distribution and abundance patterns of individual species and the life history features that render the species vulnerable or resistant to extinction. More distressing, we are incapable of determining these aspects of the biology of given taxa without a reliable foundation of systematics and taxonomy. The understanding and conservation of crustacean biodiversity will depend on the implementation of major new funding and training programs that are directed toward systematics and taxonomy, and toward natural history museums which hold the collections necessary for such comparative studies.

B. Ecology of the Arthropod Radiation

Between 3 and 4 billion years ago, biopolymers (proteins, polysaccharides, nucleic acids), which had been formed as a result of physical processes, became enclosed in membranes, and the resulting cells developed the ability to replicate themselves. These duplicated cells became organized into multicellular organisms that could move about on their own about a billion years ago. An accelerating explosion of biotic diversity occurred sometime in the Precambrian (perhaps 700–800 million years ago), but fossils remain of this early proliferation are few, having been either poorly preserved or destroyed in subduction zones beneath migrating continents. By the beginning of the Paleozoic era (about 570 million years ago [mya]), almost all of the major phyla of animals already were present (by then being larger and more amenable to fossilization or preserved in continental rocks that are still available). The Cambrian marine environment (approximately 570–500 mya) was studded with sessile epibenthic forms such as stromatolites (aggregations of unicellular organisms, their secretions, and sand or other sedimentary particles); sponges and other now extinct sponge-like forms; upright branched and laminar bryozoans ("moss animals"); filamentous now extinct organisms called graptolites, brachiopods ("lamp shells"), and other sessile organisms with ciliated feeding arms; various types of cnidarians (hydroids, anemones, primitive types of corals, jellyfish); stalked crinoids (sea lillies) and other primitive echinoderms; and the sessile life stages of primitive forms of chordates (our own ancient forebears). Most of these sessile organisms fed on suspended particles near the seafloor.

Various other organisms crawled on the seafloor, many of them having developed segmentation (and other architectural features that allowed them to move more rapidly and feed more effectively) as well as external armor that protected them from the increasing tendency for their ilk to eat one another. In addition to many lineages (often with bizarre features unlike any organisms known today) that perished in the race for motility and are known only as fossils, the motile benthic fauna included the early mollusks (whose followers would diversify during the Paleozoic into major grazers, suspension feeders, burrowing sediment feeders, and large extraordinarily active predators). Also present were diverse groups of early echinoderms (that would become grazers, burrowing sediment feeders, and slow-moving predators as the Paleozoic progressed) and the early filter-feeding chordates (many of which would develop into large, increasingly active and important predators as jawless and then jawed fish during the early to mid-Paleozoic). At least as important for biodiversity in the long run, the Precambrian benthic marine environment abounded with a diverse array of unsegmented and segmented worms that provided the prototype for the development of the joint-legged animals, the arthropods.
Whether the arthropod phyla (uniramians [centipedes, millipedes, insects], onycophorans, tardigrades, chelicerates [horseshoe crabs, scorpions, spiders, ticks], pycnogonids, trilobites, pentastomids, and crustaceans) arose from only one group or independently from several groups of the early segmented annelid worms is hotly debated. The annelids and archerous groups originated in the Precambrian, and all of the major phyla were distinct by the time they were fossilized. The crustaceans, trilobites, and chelicerates appeared first in the fossil record, followed at a later time by the uniramians. Some experts have concluded, on the basis of functional morphology and developmental patterns, that the major arthropod groups probably all arose independently, and that the Crustacea are not closely related to the uniramians, trilobites, or chelicerates. Other scientists have argued, on the basis of morphology, that the trilobites are ancestral to both the chelicerates and crustaceans. New evidence from RNA and DNA sequences has brought forth various views, including the idea that the crustaceans are more closely related to the uniramians than to other arthropod groups. The question of how the arthropod groups evolved, then, remains far from solved at present.

Many workers concur that the onycophorans (velvet worms) and the uniramians have a common origin. Some workers have suggested a possible connection between two Cambrian marine fossils and the onycophorans, who are represented today by 80 species in warm moist terrestrial environments. The first myriapods (centipedes, millipedes) were fossilized in the Silurian (440–405 mya) and probably they or their ancestors were marine. Terrestrial myriapods and wingless insects are known from the Devonian (405–360 mya), and winged insects had arisen by the Carboniferous (360–290 mya). These underwent a continuing spectacular radiation that resulted in almost a million species of described insects that inhabit almost all terrestrial and many aquatic environments today. In addition to 32 living orders of insects, at least 10 other orders arose and went extinct in the late Paleozoic and early Mesozoic.

About 400 species of tardigrades (water bears) are known from marine (intertidal to deep-sea) environments, freshwater, hot springs, and semiterrestrial environments in plants and humus; the most primitive group is marine. A single fossil is known from Cretaceous (135–65 mya) amber. These minute organisms (most 0.1–0.5 mm, a few to 1.7 mm) occupy benthic, interstitial, water film, commensal, and parasitic habitats. They also have been been collected from some of the driest habitats on earth, the Antarctic dry valleys; they are capable of being freeze-dried, after which they resume metabolism when rehydrated.

Represented by the xiphosurans (horseshoe crabs) and the eurypterids (sea scorpions), the early chelicerates were marine and had appeared by the Cambrian. Eurypterids and xiphosurans were crawling scavengers and predators, crushing food between the bases of their legs. The anterior appendages (chelicerae) of some eurypterids were well developed and clawed. Many eurypterids were small but some grew to very large sizes (up to 3 m). These marine forms flourished during the Ordovician, Silurian, and Devonian (500–360 mya) and were abundant until the Permian (290–245 mya), after which the eurypterids became extinct and the xiphosurans were greatly diminished (only five species of horseshoe crabs remain today). Some eurypterids apparently invaded fresh water and may have become semiterrestrial. Related lineages, including the scorpions (the most primitive living arachnids) were present in the Silurian and invaded land very early, at least by the Carboniferous. Diversification of the Chelicerata resulted in a total of about 80,000 described, mostly terrestrial species living today.

The pycnogonids (sea spiders), considered a sister group to the Chelicerata, are known from the Devonian onward. They are represented by about 1000 species and 86 genera today, all marine. They are suctorial predators and occupy all oceans down to about 7000 m. Most are small (<10 mm), though the bodies of some deep sea forms are 20 to 30 mm with leg spans of up to 700 mm. The “protonymph” larva of some pycnogonids and the “trilobite” larva of some xiphosurans share some developmental features with the “protopsid” larva of trilobites.

The trilobites, along with the chelicerates, were major components of the motile benthic fauna of the early Paleozoic. The trilobites were exceedingly abundant during the Cambrian and Ordovician (600–440 mya) and persisted in somewhat lower numbers until the end of the Permian (245 mya), after which they became extinct. All of the 4000 species were marine. They were major scavengers and predators, crushing food between the bases of their legs, which operated as gnathobase jaws. Some were possibly sediment or suspension feeders. Most species were small (10–30 mm), some were moderately sized (30–60 mm), and a few giants (600–700 mm) are known. Most species were benthic and epifaunal or shallow infaunal, but some small (<10 mm) forms were planktonic.

The pentastomids, wormlike parasites that inhabit the respiratory system of terrestrial vertebrates, include about 95 species and are thought to be related to crusta-
ceans. They may have been a relatively late offshoot of early arthropod lineages, developing with or after the radiations of terrestrial vertebrates in the Carboniferous (amphibians), Permian (reptiles), and Mesozoic (reptiles, birds, mammals).

In the midst of this evolutionary theater rose the crustaceans. We shall return to the historical development of crustacean diversity after we are introduced to the major characters.

III. BIODIVERSITY OF FOSSIL AND MODERN CRUSTACEA

A. The Major Groups of Crustacea

1. Centers of Taxonomic Diversity

At least five major radiations in ancestral crustacean lineages produced the remipeds, cephalocarids, branchiopods, maxillopods, and malacostracans (Table 1). Scientists believe, either because of their primitive body forms or their fossil record, that all five major groups have ancient origins extending back at least as far as the beginning of the Paleozoic. The remipeds, cephalocarids, and branchiopods probably are impoverished relicts of ancient great lineages that did not fossilize well. Today, however, these three groups (along with a number of primitive malacostracan lineages) are marginalized in hypersaline, brackish, and fresh water (sometimes hot springs or ephemeral pools), caves, groundwater, or other isolated, seemingly suboptimal habitats. Collectively, these three classes contribute only 14% of the diversity in orders, 5% of the diversity in families, 2% of the generic diversity, and 2% of the species diversity within modern crustaceans. The remaining 2 classes—the maxillopods and especially the malacostracans—are the dominant groups in most major marine environments today and are responsible for much of the diversity in major body plans (e.g., as represented in subclasses; Table 1). The malacostracans alone constitute over half (53%, 59%, and 55%) of all present and past (fossilized) crustacean families, genera and species, respectively. The major groups of malacostracans have been subject to considerable turnover in the fossil record, however, since the lineage is ancient and many of their major evolutionary experiments (body plans at the level of order) have been lost to extinction. Whereas 80 to 100% of all the orders of branchiopods and maxillopods ever known are still extant, only 59% of malacostracan orders are still living. If their order did survive background extinction and the great mass extinctions (e.g., at the ends of the Paleozoic and the Mesozoic), however, malacostracan families have persisted (or are currently diversifying) at a higher rate (88% of all known families are still extant) than those of the maxillopods (76% living) and branchiopods (74% extant). Though only 67% of branchiopod genera still exist, very high percentages of genera in both the malacostracans (97%) and maxillopods (96%) are alive today.

Within the maxillopods, centers of taxonomic diversity are found in the copepods, ostracods (seed shrimp), and, to a lesser extent, cirripeds (barnacles). Copepods include 38% of the orders, 50% of the families, 54% of the genera, and 49% of the species in the maxillopods. Ostracods constitute 29% and 38% of the orders and families, and 36% and 44% of the genera and species in the maxillopods. Barnacles comprise 19% and 11% of the orders and families, and 10% and 6% of the genera and species in the group. Both ostracods and barnacles are ancient lineages whose fossil records extend back at least to the dawn of the Paleozoic.

Within the malacostracans, peracarids and eucarids dominate the group in terms of taxonomic diversity. Peracarids include 37% of the orders and 62% of the families, and at least 56% of the genera and 53% of the species, in the Malacostraca. The eucarids make up only 11% of the orders and 28% of the families, but at least 37% of the genera and 44% of the species within the group.

2. Architectural Diversity

To understand the diversity of Crustacea, it is necessary to come to grips with their major ways of life as reflected in their body plans. The range of morphological diversity among crustaceans is remarkable, considerably greater than that of the insects. Primitively, each segment supported a pair of appendages. Crustacean appendages have multiple joints along their length and are considered to have evolved from a biramous condition (divided into two branches), although both head and trunk appendages have lost one side of the branch in some groups. The primary mode of evolutionary change in the early diversification of crustaceans was by molding segments into functional units along the length of their body and reducing various morphological features, including both segments and appendages. The resulting fundamental body regions included a five segmented head (cephalon) and a long, segmented trunk that is usually divided into a thorax and abdomen.

The head was formed by telescoping five segments and their associated appendages into a fused unit that bears two pairs of antennae (a distinguishing character-
<table>
<thead>
<tr>
<th>Class</th>
<th>Subclasses</th>
<th>Orders</th>
<th>Families</th>
<th>Genera</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remipedia</td>
<td>—</td>
<td>2 (1 living), 4%</td>
<td>1 (1 living), 0.1%</td>
<td>6 (5 living), 0.1%</td>
<td>At least: 11 (10 living), &lt;0.1%</td>
</tr>
<tr>
<td>Cephalocarida</td>
<td>—</td>
<td>2 (1 living), 4%</td>
<td>2-3 (1-2 living), 0.2%</td>
<td>At least: 5 (4 living), 0.1%</td>
<td>At least: 836 (821 living), 2%</td>
</tr>
<tr>
<td>Branchiopoda</td>
<td>—</td>
<td>4-5 (4 living), 7%</td>
<td>At least: 35 (26 living), 4%</td>
<td>At least: 108 (94 living), 2%</td>
<td>At least: 18,386 (18,293 living), 43%</td>
</tr>
<tr>
<td>Maxillopoda</td>
<td>6 living</td>
<td>21 (18 living), 38%</td>
<td>355 (270 living), 43%</td>
<td>At least: 2162 (2077 living), 39%</td>
<td>At least: 8081 (about 8000 living), 44%</td>
</tr>
<tr>
<td>Ostracoda</td>
<td>6 (3 living), 29%</td>
<td>136 (35 living), 38%</td>
<td>At least: 776 (695 living), 36%</td>
<td>At least: 11 ± 42 (1138 living), 6%</td>
<td>At least: 23,890 (23,800 living), 59%</td>
</tr>
<tr>
<td>Mystacocarida</td>
<td>1 living, 5%</td>
<td>1 living, 5%</td>
<td>2 living, 0.1%</td>
<td>10 living, 0.1%</td>
<td>About 9000 living, 49%</td>
</tr>
<tr>
<td>Copepoda</td>
<td>8 living, 38%</td>
<td>177 living, 50%</td>
<td>1167 living, 54%</td>
<td>150 living, 1%</td>
<td>5 living, &lt;0.1%</td>
</tr>
<tr>
<td>Branchiura</td>
<td>1 living, 3%</td>
<td>1 living, 0.3%</td>
<td>4 living, 0.2%</td>
<td>150 living, 1%</td>
<td>5 living, &lt;0.1%</td>
</tr>
<tr>
<td>Tantulocarida</td>
<td>1 living, 5%</td>
<td>1 living, 0.3%</td>
<td>2 living, 0.1%</td>
<td>150 living, 1%</td>
<td>5 living, &lt;0.1%</td>
</tr>
<tr>
<td>Cirripedia</td>
<td>4 living, 19%</td>
<td>39 (33 living), 11%</td>
<td>At least: 211 (207 living), 10%</td>
<td>At least: 11 ± 42 (1138 living), 6%</td>
<td>At least: 23,890 (23,800 living), 59%</td>
</tr>
<tr>
<td>Malacostraca</td>
<td>2-7 (usually 5)</td>
<td>27 (16 living) (3 extinct living)</td>
<td>431 (383 living), 53%</td>
<td>At least: 3175 (3093 living), 59%</td>
<td>At least: 23,890 (23,800 living), 59%</td>
</tr>
</tbody>
</table>

In cases where number of taxa is controversial, the lower number was used in the total. In many cases, the number of certain taxa (especially genera and species) were not summarized in sources, so minimum numbers were estimated from the number of higher taxa (e.g., in a group with five families for which data on number of genera were not summarized, it was assumed that at least five genera [one per family] and at least five species [one per genus] were present). Therefore, these data probably considerably underestimate the number of genera and particularly the number of species in these classes (although total species exceed those previously estimated by other authors, e.g., Brusca and Brusca, 1990, who did not include paleontological data). Percentages horizontally aligned with classes (bolded) represent the total number in that taxon compared to the number for all Crustacea. Percentages horizontally aligned with subclasses (not bolded) represent the total number in that taxon compared to the total number in that class. Percentages are based on total (fossil and living) taxa and are rounded to the nearest whole number (except when <1%).
istic of crustaceans), the mandibles (jaws) that flank the mouth, and two pairs of accessory food handling appendages, the maxillules and maxillae. The antennules (first pair of antennae) are biramous in the primitive remipede. However, the antennules are uniramous in the cephalocarids, all of the branchiopods, all of the maxillopods (where present), and some of the eumalacostracans, while those of the primitive phyllocarid malacostracans and some of the eumalacostracans are biramous and those of stomatopods triramous, raising the question of whether or not the antennules of crustacean ancestral groups were all biramous. The antennules are reduced or lost in the parasitic branchiurans and tanaopoda and in very sessile adult cirripedes. The second antennular appendages (antennae) are biramous in most crustaceans but are uniramous in anostracans, some copepods, lepostracans, some but not other fossil phyllocarids, and some eumalacostracans. The antennae are reduced in the parasitic branchiura (modified for attachment) and lost in the anostracans, tanaopoda, and sessile adult cirripedes.

One or more of the anterior thoracic segments may be fused to the head (remipedes, isopods, amphipods), or the anterior thoracic appendages (maxillipeds) may be functionally associated with the mouthparts of the head (fossil kazacharthans, branchiopods; fossil liposuchids, ostracods, mystacocaridans, and copepod maxillipeds; stomatopod and possibly palaeostomatopod hoplocaridans; anaspidean synarcids; mysidans, loligostracids, tanaopoda, spatangiformaceans, cumaceans, mictaceans, isopods, and amphipod peracaridans; stomatopods and many eumalacostracans).

Reflecting very early divergence among groups of crustaceans, the division of the trunk into units that functioned in feeding (the anterior appendages), reproduction, and locomotion (often with accessories that aided respiration) was much more variable among major crustacean groups than the head region. The wormlike remipedes usually are considered to represent the most primitive crustacean morphology. Their isopodlike appendages are undifferentiated throughout the long trunk, followed by an anal segment with simple caudal rami. The trunk of the cephalocaridans, however, is divided into 8 thoracic segments, 11 abdominal segments, and an anal segment with caudal rami. The trunk region of both the branchiopods and maxillopods is variable among groups. Several groups with modified morphology possess simply a reduced trunk (cladocerans, conchostracans, ostracods, mystacocaridans, and adult tanaopoda). A thorax is present in the anostracans and anostracans (usually 11 segments), extinct kazacharthans, copepods (6 segments), branchiurans (4 segments), tantalocaridans (6 segments in juveniles), cirripeds (6 segments, reduced), and malacostracans (8 segments). The number of segments included in the abdomen is variable within several groups of branchiopods (anostracans, natostracans, fossil kazacharthans, conchostracans) and in juvenile tantulocaridans and ascothoracidan barnacles. The abdomen is reduced or lost, or segmentation obscured, in groups that are modified for specialized ways of life (cladocerans, the parasitic branchiurans and tantulocaridans, and most cirripeds, especially the parasitic forms). The primitive phyllocarid malacostracans all possess an abdomen with seven segments, in addition to a terminal anal segment with caudal rami. Adult hoplocaridans and eumalacostracans have lost one of the abdominal segments, however, and bear six abdominal segments plus an additional flattened anal segment termed a telson. Morphological and developmental evidence of a vestigial seventh abdominal segment (either at the front or the posterior of the abdomen) in the stomatopod hoplocaridans and in the mysidans, tanaopoda, and possibly the isopod eumalacostracans reflects another evolutionary tendency toward reduction of segments.

Trunk appendages in the remipedes are oarlike but different from the leaflike or phyllopodous appendages of the primitive branchiopods. Leaflike or phyllopodous appendages predominate in the cephalocarids, all of the branchiopods, and in the primitive malacostracans, including a number of extinct groups and the modern leptostracans. However, greatly modified or reduced numbers of trunk appendages (especially on posterior trunk segments) characterize the maxillopods (e.g., the ostracods, mystacocaridans, and copepods). Parasitic forms (e.g., the adult rhizocephalan barnacles) often lose all appendages. The copepods, branchiurans, and juvenile tantulocaridans all have biramous thoracic legs. In the malacostracans, however, some of the thoracic appendages are uniramous and some biramous in stomatopods, and all eight pairs of the thoracopods usually are uniramous (with some of these enlarged and clawed) in the eumalacostracans. Even some of these groups, however, show reduction in development of appendages toward the posterior of the thorax. Branched or layered structures arising from the base of the thoracic legs (epipods) serve as gills in the hoplocaridans and eumalacostracans.

Posterior trunk or abdominal appendages are reduced in most cephalocarids, branchiopods, and maxillopods (natostracans, conchostracans) or absent (cephalocarids, anostracans, cladocerans, ostracods, mystacocaridans, copepods, branchiurans, tantulocaridans,
and cirripeds). Like the remipedes, however, most of the malacostracans bear pleopods on the abdomen that often function in swimming (and sometimes facilitate mating). These may be present or absent in different fossil phyllocarids and are present, but reduced toward the posterior, in the leptostracans. Hoplocaridans have five pairs of well-developed pleopods that support tufy gills (palaeostomatopods, stomatopods). Some peracarid eumalacostracans (isopods) also bear breathing structures (gills or tracheae) on their pleopods. The eumalacostracans have pleopods, but their number and size often is reduced (particularly toward the posterior) or may be absent (anaspideans, some mysids, some thermosbaenaceans, some cumaceans, and some ano- murans). The sixth abdominal segment is fused to the telson in modern (but not ancient) tanaidaceans, as well as some cumaceans and isopod eumalacostracans. The hoplocaridans and most groups of eumalacostracans have a well-developed "tailfan" that is composed of uropods (bladelike appendages on the sixth abdominal segment) and a telson (a flattened and often armored posterior segment that may have one or more terminal projections but never bears causal rami).

The position of gonadal duct openings (gonopores) on either the ventral part of the body or the bases of the legs varies consistently among groups of crustaceans, reflecting the ancient divergence of different taxa. The female gonopore opens on the sixth thoracic segment, and the male gonopore on the eighth thoracic segment, in all malacostracans.

Both simple (ocelli) and compound (faceted) eyes are present in most taxa. One medial ocellus frequently occurs in larvae, and one or more ocelli are found in adults of some groups of branchiopods and maxillicopods such as the anostracans, notostracans, cladocerans, ostracods, mystacocarids, copepods, and branchiurans. Only ocelli (or ocellus-like structures) are present in the adults of the extinct kazacharhans, some cladocerans, mystacocarids, and copepods. Compound eyes often are positioned on stalks, as in the anostracan branchiopods, most malacostracans, including many phyllocarids (exclusive canadispans, leptostracans), the hoplocarids (exclusive aechronectids, stomatopods), extinct belotelsonideans, and most eumalacostracans (exclusive palhaco- caridaceans, some anaspideans, mysids, lophogast- tridans, thermosbaenaceans [nonfunctional], some tanaidaceans [on lobes], speleogriphaceans [on lobes], cumaceans [on a medial lobe], micuaceaans [nonfunc- tional on lobes], some isopods [on lobes], euphausia- ceans, amphionidaceans, dendrobranchiaceans, procariid shrimps [reduced], caridean shrimps, steno- podidean shrimps,thalassinidean shrimps, astacideans, palinurans, anomurans, and brachyurans). Eyes are reduced or absent in members of primitive lineages, sessile lineages, or species that inhabit caves or the deep sea (the remipedes, cephalocarids, many ostracods, tan- tulocarids,adult cirripeds, tanaidaceans, some speleogriphaceans, many cumaceans [especially females], some micuaceaans, some isopods, some amphipods, amphionidaceans, procariid shrimps, some caridean shrimps, some stenopodid shrimps, and some astaci- deans). The branchiuran maxillicopods have a particularly intriguing pair of compound eyes that project internally into a blood sinus suspended on a stalk composed of the optic nerves. Muscles move the eyes within the blood sinus.

Either a cephalic shield or a carapace may cover the head and part or all of the thorax. Remipedes, cephalo- carids, copepods, and tanulocarids bear a cephalic shield, while a "folded" or "bivalved" carapace characterizes the natostracans, extinct kazacharhans, cladocerans, conchostracans, ostracods, branchiurans, cirripeds, phyllocarids, hoplocarids, and most eumal- acostracans. However, these structures are reduced or absent in other groups (anostracans, mystacocarids, syncarid and most peracarid eumalacostracans). The carapace of cirripeds is modified into a fleshy, saddle-like mantle that may envelope most of the body and, in thoracicans, is covered with calcareous plates; eggs are brooded in the mantle cavity for various lengths of time. Conchostracans, most cladocerans, and some ostracods use their carapace as a brood chamber; one cladoceran broods eggs in its molted carapace. Thermosbaenaceans females possess a very enlarged, thin, blood-filled carapace that functions both as a brood chamber and for gaseous exchange. The enlarged, very thin carapace in female amphionidaceans may serve a similar function, though this has not yet been examined. The carapace extends laterally in decapods to enclose the base of the legs and gills, increasing the efficiency of water flow over the gills.

B. The Rise of Crustacean Diversity in the Fossil Record

All of the early continents, including Gondwana (modern South America, Africa, Madagascar, India, Australia, Antarcticia), Baltica (Scandinavia and northwestern Euro- pe), Laurussia (North America), Siberia, Kazakh- stania, and China, girdled the earth in tropical and subtropical locations. No land extended above 60° N or below 60° S during the Cambrian (570–500 mya), allowing mild warm conditions in most coastal areas. Developing from segmented wormlike forms, the earli-
east crustaceans must have developed sets of pronglike or leaflike legs to shuffle along the surface of the soft marine sediments. In addition, they secreted a tough, often calcified cuticle to protect them from the other benthic predators (nautiloid molluscs and trilobites, which were the most abundant benthic predators and scavengers in most marine habitats during the Cambrian). The earliest fossilized crustaceans included three orders of phyllocarid malacostracans (phyllodanous detritus feeders, scavengers, and generalized predators known from the Lower Cambrian); the thoracican goose barnacles (suspension-feeding maxillopods known from the Middle Cambrian); and the ostracods (scavenging, detritus-feeding maxillopods known from the Cambrian onward). Following the recent discovery of the predaceous Remipedia, which exhibit a very primitive arrangement of segmentation and appendages, many workers concluded that the remipedes, though not readily fossilized, represent one of the probable forms of the earliest crustaceans and likely were present in the Cambrian. In addition, the primitive morphology of the ascothoracian barnacles suggests that they also may have been present (though unfossilized) in the very early Paleozoic. Of these early groups, the canadapsid phyllocarid malacostracans did not persist into the Ordovician.

During the Ordovician (500–440 mya), part of the huge island continent of Gondwana slipped beneath what was then the south pole, but the climate remained generally warm and mild, with the northern reaches of Gondwana and most of the island continents situated in tropical or subtropical waters. Xiphosuran (horseshoe crabs) and eurypterid (sea scorpions) chelicerates joined the exceedingly abundant, ravenous trilobites on the seafloor. Nautiloids continued to diversify morphologically; jawless fish (encased in protective outer layers of bone and still not very mobile) proliferated. The ostracod crustaceans underwent their greatest radiation, and phyllocarid malacostracans (hymenocarid crustaceans, archaeostracans) were abundant; the hymenocarid crustaceans, however, disappeared after the Ordovician. A mass extinction closed the Ordovician.

The Silurian (440–405 mya) also was characterized by mild climates (except in the parts of Gondwana that rested over the pole), and a circumtropical current wound its way among the island continents and reefs composed of stromatoporoid sponges, rugose corals, and other sessile fauna. Eurypterids flourished, trilobites persisted in lesser numbers, myriapods and primitive arachnids (probably marine) arose, and jawed fish made their appearance. Thoracican goose barnacles (attached to a eurypterid) are recorded, ostracods flourished, and archaeostracan phyllocarid malacostracans were abundant.

During the Devonian (405–360 mya), predation from the flourishing eurypterids, increasingly motile jawed fishes, and nautiloid cephalopod mollusks dominated benthic ecology. Trilobites still occupied many benthic marine environments. Pycnogonids were present. In an era of mild climates and inland seas, fish and invertebrate groups radiated into freshwater environments. A circumtropical current coursed between Gondwana and the island continents. It was in these warm, sheltered, shallow coastal waters laden with archipelages that crustaceans began their great radiations. The fossil record shows that the archaeostracans (phyllodanous malacostracans) reached their greatest diversity, and the predaceous hoplocarids (palaeostomatopods, archaeostomatopods) made their first appearance. The ostracods began their second great radiation. The boring acrothoracian barnacles and the first predaceous, lobster-like decapod malacostracans appeared (Palaepalaeamon, allied either to the astacid crayfish and lobsters or palinurid lobsters). Detritus-feeding and scavenging conchostracan branchiopods originated, invading fresh water, and the detritus-feeding liposarcans (allied to the cephalocarids or branchiopods) flourished in hot springs. Fossils and primitive morphology suggest that other primitive detritus-feeding branchiopods (notostracans, anostracans) likely were present too. Another mass extinction impacted faunas during the Late Devonian.

The taxonomic, morphological, and ecological diversification of the Crustacea accelerated during the Carboniferous (Mississippian, Pennsylvanian; 360–290 mya) of the Late Paleozoic as the marine chelicerates and trilobites began their decline. The continents were beginning to congeal, setting the stage for the global climatic changes that likely contributed to the greatest mass extinction of all time at the end of the Paleozoic. Gondwana had largely completed its migration beneath the south pole and emerged to begin to fuse with Laurussia and Baltica (modern North America, Scandinavia, and northwestern Europe) over the equator. Though the southern reaches of Gondwana (modern South America) were still plagued with bouts of glaciation, the warm moist equatorial continents were awash with inland seas and swamps. Land plants had multiplied in diversity and biomass, their seed ferns, scale trees, and conifer trees producing great forests that decayed in coal swamps or washed out onto the continental shelves, enriching coastal sediments. Insects took flight (including some huge species), amphibians underwent their greatest radiation, and the first reptiles appeared. Ma-
rine organisms proliferated in the enriched coastal sediments. Sharks and jawed bony fishes were abundant and active. Nautioids declined, replaced by the emerging ammonoid cephalopod mollusks. Remipede crustaceans are recorded from the Lower Pennsylvanian. Conchostracan branchiopods radiated anew. Notostracan branchiopods probably were present. Acrothoracian barnacles are known from their borings in clam shells, and lepadomorph thoracican barnacles are recorded from this period. Early malacostracan lineages (the scavenging archaeostracan and hoplostracan phyllocarids; the scavenging aequihornetids and predaceous hoplocarids; the scavenging belotelsonidans, waterstonellideans, eocaridaceans, and paleocaridaceans) were exceedingly abundant and dominated the shallow marine fauna. Pygocephalomorphans and lophogastridans mysidaceans came on the scene, as well as most of the peracarid malacostracans (tanaisaceans, spelaecostracans, cumaceans, isopods, and probably other unfoossilized groups). Definitive palinurian decapods appeared, and the earliest lineages leading to the brachyuran crabs evolved. Several of the early malacostracan groups (hoplostracan phyllocarids, aequihornetid and palaeo-stomatopod hoplocarids, belotelsonidans) disappeared after the Carboniferous as the peracarid and eucarid malacostracans took center stage.

The Permian period (290–245 mya) brought one of the most significant geomorphological events in the history of the earth. All of the previous continental landmasses fused into a single gigantic continent, Pangaea, between 260 and 250 mya. The collision of Gondwana, Laurussia, Baltica, Kazakhstan, and Siberia is still visible in the mountain belts stretching across northern South America, northwestern Africa, eastern North America (the Appalachians), Britain and Scandinavia, and Eurasia (the Uralis and other ranges). By 250 mya, the ancient island continent of China also had joined the supercontinent. Pangaea now straddled the equator, anchored by Siberia near the north pole and Antarctica over the south pole. The circumtropical seaway was closed. The elevated landmasses produced a more extreme, cooler, continental climate than had been present throughout most of the Paleozoic (whether or not this was solely responsible for the mass extinctions is debated). In the cool, dry conditions, reptiles radiated, replacing the previously dominant amphibians, and the first mammal-like reptiles appeared. Many organisms, including worms, heart urchins, sea cucumbers, and crustaceans, mined the organic riches in the coastal sediments, throwing up sediment that buried and clogged the feeding structures of the epibenthic sessile filter feeders. Conchostracan branchiopods, ostracod maxillopods, primitive malacostracans (archaeostracan phyllocarids, the first leptostracan phyllocarids, waterstonellideans, eocaridaceans, palaeocaridacean syncarids, and pygocephalomorph mysidaceans), at least several peracarid malacostracans (tanaisids, cumaceans, isopods), and the first definitive astacid (crayfish- or lobster-like) eucarid malacostracans thrived in benthic marine environments. Brachiurans likely were present. The Paleozoic era ended, however, with the loss of 96% of all species. Many crustacean groups did persist through the Permo-Triassic biotic catastrophe, though, and underwent reradiation during the Mesozoic.

The early Triassic was a relatively barren environment both on the land and in the sea. The landmasses were high, cool, and dry, with widespread deserts. In coastal waters, no corals or reef communities are known. All of the trilobites, eurypterids, most of the nautiloid cephalopods, and many of the primitive fishes had been lost. All of the primitive phyllocarid malacostracans, except for the newly evolved leptostracans, were gone. All of the early hoplocarids had been eliminated except for a thin line, probably present but not fossilized, that would lead to the modern stomatopods. All of the late Paleozoic primitive malacostracans (the belotelsonids, waterstonellideans, eocaridaceans, the early palaeocaridacean syncarids, the pygocephalomorph mysidaceans) were now absent. The sessile epibenthic suspension feeding community (many of the bryozoans and brachiopods, many primitive echinoderms such as blastoids and most of the crinoids, primitive tabulate, and rugose corals) that had dominated Paleozoic benthic habitats had been largely obliterated. Most of the plankton was gone. As the Triassic progressed, however, mammal-like reptiles and the early groups of dinosaurs appeared, other reptiles diversified, and coniferous trees came to dominate the land. In the sea, the surviving remnants of the cephalopods began to diversify into modern groups, including predaceous squid and octopus-like forms. Scleractinian corals began to proliferate, even though they were subordinate to sponges, algae, and bryozoans as reef builders. Conchostracan branchiopods, ostracods and lepadomorph cirripeds, anaspidacean syncarids, lophogastrid mysidaceans, tanaisid and isopod peracarids, the first dendo-brachiate decapod shrimps, and astacid and palinurid decapods are recorded. Probably many more lineages were reestablishing themselves, though unfossilized, during this period (remipedes, cephalocarids, all of the branchiopod groups, many of the primitive maxillopods, other primitive eumalacostracans such as syncarids, and the mysidaceans, peracarids, and eucarids).
Another mass extinction, however, spread biological devastation in the Late Triassic.

It is thought that huge continental landmasses begin to accumulate heat beneath them, creating convection currents in the mantle that cause landmasses to drift apart approximately every 500 million years. In the Jurassic (210–140 mya), Pangea began to break apart. About 200 mya, Laurasia (North America and Eurasia) moved away from Gondwanaland, forming a nascent Atlantic Ocean in the west and a gulf between Asia and East Africa in the east. India drifted away from Antarctica toward Asia, and the south Indian Ocean began to form between Australia (still attached to Antarctica), Antarctica, and the joined southern tips of Africa and South America. A widening rift between South America and Africa was flooded with seawater near the end of the Jurassic. The northern drift of Laurasia created a pathway for a circumtropical current between the northern and southern continents. This shallow warm seaway, adorned with archipelagos, would be called the Tethys Sea, and it was the cradle from which most of the modern families of organisms that we know today would emerge.

The early Jurassic climate was cool, becoming warmer. High sea levels encroached on the continents, spreading shallow seas. Although mass extinctions challenged the faunas during the Middle and Late Jurassic, the Jurassic period is known for its spectacular radiation of reptiles, including the dinosaurs and pterosaurs. Modern groups of amphibians (frogs, salamanders) and birds arose. Primitive flowering plants and many of the modern groups of insects (flies, moths, bees, wasps, ants) diversified in the increasingly warm, mild climate. In the sea, marine reptiles must have been fearsome predators, crocodiles and turtles abounded, many of the modern families of bony fish were formed, corals acquired symbiotic algae that allowed them to grow up and away from sedimentation (such as that created by burrowing animals in the late Paleozoic) 10 times faster than they could in the absence of the photosynthetic symbiont. Reefs became more extensive and boring organisms more common. The Jurassic fossil record of Crustacea includes conchostracan and ostracod radiation (extinct Kazachthra) branchiopods, and shows a renewed ostracod radiation. The lepadomorph barnacles were present, and the balanomorph (acorn) barnacles made their first appearance. Many of the other groups (remipeds, cephalocarids, other branchiopods, other malacostracans) likely were present but not fossilized. In the Malacostraca, the fossil scudulid hoplocarids (descendants of the Devonian archaeostomatopods and progenitors of the modern stomatopods) arose. The last remaining vestige of the phyllocarids (the leptostracans) and some synbranchoids may have been present but not fossilized. The first mysidacean eumalacostracans appeared, tanaid and isopod peracarids were present, and many of the other peracarids likely were present but not preserved. The eucarid malacostracans burst on the scene, as shown by the first appearances of fossil caridean, stenopodidean, and thalasinidean shrimps, and anomuran and definitive brachyuran crabs. Dendrobanchiate shrimps, as well as astacidean and palinuran crayfishes and lobsters, also were present in the fossil record, completing the roster of all the major groups of eucarids known today. Despite the mass extinctions, no major lineages were lost during the transition from the Jurassic to the Cretaceous.

The tropical Tethyan seaway continued to girdle the earth and maintain mild conditions during the Cretaceous (140–65 mya), although the pathway was narrowing in several sites (e.g., northeastern and northwestern Africa). North America was still united with Eurasia despite a widening gulf invading northward from the opening Atlantic. India had not yet crashed into Eurasia, Madagascar was sliding northeastward from southern Africa, and Australia had begun to separate from Antarctica, beginning its trek toward the tropics. Africa and South America had left Antarctica behind as they migrated northward and severed their connections to each other. The Andes and Rocky mountains rose as North and South America ploughed westward and the Pacific plates dived beneath them. The continents were awash with spreading inland seas and swamps. The reptilian radiation was at its apex, spawning the largest land and marine reptiles ever known. Grasses appeared, and other flowering plants and insects diversified even further; coniferous gymnosperms declined.

In the sea, reefs began to accrete and proliferate on an unprecedented worldwide scale and reached their highest diversity ever. Scleractinian corals flourished, and in sheltered areas huge rudist bivalve mollusks (which, like the corals, may have obtained photosynthetic symbionts) committed massive deposits of calcium carbonate to reef structure, threatening the dominance of corals in the reef community. The teleostean (boney) fishes underwent major diversification, developing new body shapes that increased their maneuverability, calcareous mouth plates that crushed shelled and hard-bodied prey, and pincer-like jaws that could grab and pluck prey from the substratum. Beuthic octopuses with flexible arms and bodies could engulf and subdue prey. Crustaceans developed hardened specialized claws that could crush or cut prey. Motile and
sessile organisms seeking refuge from the burrowing organisms that caused extensive movement of benthic sediment (bioturbators) likely colonized the high, hard substratum provided by the corals and rudists. Boring organisms proliferated worldwide, for boring into the calcareous substratum provided protection from the increasingly diverse and dangerous predators. Individual bioeroders eventually died, however, leaving a vacant crypt. This three-dimensional microenvironment provided a safe refuge for a multitude of small sessile and motile invertebrates. The organisms that invaded the holes left by the borers, termed the crypto fauna, were constrained to small body sizes in order to occupy the protective holes, were plagued by predation whenever they emerged from the crevice to feed or mate, and had to develop special adaptations for acquiring and keeping the packed refuge space. All of these processes accelerated morphological and other adaptations in the benthic fauna. Small body size usually is associated with production of fewer, larger eggs, abbreviated development, and higher rates of speciation and extinction in motile invertebrates. These life history features likely facilitated diversification of other behavioral and morphological features, such as the deadly fighting behavior and armored bodies and appendages that developed in many of the stomatopods and decapods. Special adaptations (including coloniality, unique growth strategies, and chemical warfare) also developed in the sessile crypto fauna, promoting diversification and enabling them to better acquire and defend living space.

Among the crustaceans, the conchostracan branchiopods and ostracod maxillopods again are represented in the Cretaceous fossil record; the remipedes and all other modern groups of branchiopods and maxillopods likely were present but not preserved. Copepods, ascothoracid barnacles (as trace fossils), verrucomorph (wart) barnacles, and the squat modern eathamaloid acorn barnacles made their first appearance as fossils. Among the lower malacostracans, the leptostracan phyllocarids likely were present but unfossilized, and the predaceous stomatopod hoplocarids came into their own. The extinct sculids (bathyquiloids) and three of the major modern groups of stomatopods (the reef-dwelling gonodactyloids, mud-burrowing squillids, and sand-burrowing lysisquiloids) all appeared for the first time. The synarcids and mysidaceans probably were present but not recorded. Although their distribution suggests a more ancient derivation, the thermobasaneans are first known in the Cretaceous. Other peracarids and other eucarid shrimps (stenopodideans, thyasimideans) probably were present, but the tanidaeans and isopods, as well as the dendrogrammatid and caridean shrimps, definitively inhabited Cretaceous seas. The advanced eucarids (astacids, palinurans, ano- murans, brachyurans) were all present and continued their radiations begun in the Jurassic.

Following a mild mid-Cretaceous mass extinction, a significant mass extinction at the end of the Cretaceous eliminated the dinosaurs and many of the great reptiles (including the huge flying and marine reptiles), the ammonoids and many of the older cephalopod lineages, the reef-building rudist bivalves, and many groups of plankton. It is likely that the explosive impact of a giant asteroid contributed to this biotic disaster. The role of additional factors in the faunal turnover between the Mesozoic and the Cenozoic is debated. Mass extinctions often determine the biotic structure of the next era by filtering what major groups survive to radiate again in the postextinction periods. This was true for the mammals and birds, relatively minor groups in the Mesozoic that persisted through the mass extinction and then, in the absence of the great reptiles, diversified and dominated the ecology of terrestrial environments in the Cenozoic. For the crustaceans, however, the biotic upheaval at the end of the Mesozoic had far less impact than had the great extinction at the end of the Paleozoic. The kazacharid branchiopods, related to the notostracans, had been lost during the Jurassic, but the lineage continued and survives today. The hoplocarids also persevered. Despite the possible loss of the sculid stomatopods at the end of the Cretaceous, the bathysquilloid lineage persisted (though restricted today to deep continental shelf environments). The new stomatopod superfamilies generated in the Cretaceous endured and are common in warm shallow waters today. No other major crustacean lineages were lost. Their small body sizes (compared to those of some of the giant reptiles and cephalopods, for example) may have allowed crustaceans to persist in a time of great ecosystem disturbance and food limitation. Also, although the extent of the geographic range of an individual species confers resistance to background extinction, a broad distribution for the entire group is one of the most important factors enabling lineages to survive mass extinctions (Jablonski, 1991). Reaping the success of radiations in the warm, reef-studded seas of the Jurassic and Cretaceous, many crustacean groups likely had worldwide distributions that enabled them to survive the end of Mesozoic mass extinction.

The Cenozoic can be divided into the Tertiary (65–1.6 mya, including the Paleocene, Eocene, Oligocene, Miocene, Pliocene) and the Quaternary (1.6 mya-present, including the Pleistocene and Holocene). The Atlantic was now formed and widening. India crashed
CRUSTACEANS

into Eurasia, creating the Himalayan mountains. Extensive folding, resulting in the Alps, Caucasus, and other mountains, occurred as parts of Africa fused with Eurasia, forming Asia Minor and constricting the ancient Tethys seaway into the isolated, sometimes dry, Mediterranean and Black Sea areas. Mountain building continued in the Rockies, the Coast Range, and the Andes, eventually shifting the flow of the Amazon from westward to eastward. The early Tertiary climate was warm, wet, and mild. The great radiations of placental mammals began. Horses, camels, rhinoceroses, dogs, cats, whales, bats, and primates diversified in Eocene terrestrial environments. A gradual cooling trend began in the Oligocene. Extensive plains, grasslands, and new mammalian lineages (deer, elephants, antelopes, giraffes, and the ancestors of cattle and pigs) developed through the Miocene, and large mammalian carnivores radiated in the Miocene and Pliocene.

In the oceans, there were no extensive coral reefs in the Paleocene, although scleractinian corals had survived the mass extinction at the end of the Mesozoic. The Eocene brought a radiation of reef-building corals and many families and genera of fish, mollusks, and crustaceans that survive today. Extensive volcanism and tectonic movements in the first half of the Tertiary set the stage for island fringing reefs and the development of atolls in the Pacific.

The Tertiary produced a new radiation of the cirriped barnacles, including the coronuloids (which live as commensals on whales and turtles and appeared in the Paleocene) and the first modern intertidal balanoids (which originated in the Eocene and diversified again in the Miocene). The ostracods radiated once again during the Eocene; they also are represented in the Paleocene and Miocene fossil records. Among the branchiopods, the first definitive anostracans and cladocerans appeared during the Eocene. In the stomatopod hoplocarids, a possible sculcid bathysquillid, the squilloids, and the first pseudosquilloids are known from the Eocene; in addition, fossil gonodactylid and squilloids are recorded from the Miocene, and squilloids from the Pliocene. Following a late Paleozoic genesis, the isopod peracarid eumalacostracans radiated during the Eocene, and the first amphipod peracarids appeared; both of these groups are known also from the Oligocene and Miocene. Among the eucarid malacostracans, the caridean shrimps left fossils in the Oligocene and Miocene, and the thalasinidean shrimps are recorded from the Miocene; both groups hailed from the Jurassic. The astacid crayfish and lobsters, among the first of the Paleozoic decapods, are known from the Eocene. The brachyuran crabs, also from a likely Paleozoic heritage, underwent major radiations in both the Eocene and Pliocene, with many of the modern freshwater and terrestrial lineages developing in the latter epoch. Brachyurans are represented in the Paleocene and Miocene fossil record as well. Anomurans are known from the Oligocene and Miocene.

One of the most significant events in the late Tertiary was the gradual establishment of a complete barrier to circumtropical ocean circulation about 3 mya as mountain building along the west coasts of North and South America created a Central American isthmus that linked the two continents. This removed the ameliorating effect of the circumtropical current on the climate and created a much smaller, basin-like West Atlantic region as well as a long thin East Pacific region. Considerable extinction and divergence of new genera and species ensued in the New World tropical seas.

In spite of a minor mass extinction at the end of the Eocene and deteriorating ocean circulation in the Miocene and Pliocene, however, no major groups of crustaceans were lost during the Tertiary. The ostracod and cirriped maxillopods, the stomatopods, the isopod and amphipod peracarids, and the brachyuran decapods all underwent major radiations, particularly during the Eocene.

The Quaternary (Pleistocene 1.6–0.01 mya, Holocene 10,000 years ago–present) ushered in four major periods of glaciation, particularly on the northern continents (North America, Europe, North Asia) that had not been extensively glaciated during the Paleozoic and Mesozoic. Early humans appeared in the Pleistocene. Mountain building in the Cascades, Andes, Caucasus, and the Himalayan regions continued, and rifts in eastern Africa, Asia Minor, Baikal, and southeast Asia reflected continuing tectonic shifts. Worldwide sea levels receded with the advance of continental glaciers and advanced during warmer interglacial intervals. The current climate represents emergence from a colder glaciated interval about 10,000 years ago. The extent to which the frigid polar glaciation impacted climate and seawater temperatures at lower latitudes is debated. Most modern coral reefs (including Australia's Great Barrier Reef) are only about 10,000 to 15,000 years old and represent a thin surface veneer on a complex topography inherited from past reef communities as they adjusted to changing sea levels. The amphipod branchiopods, isopod and amphipod peracarids, and brachyuran crabs are among the relatively few crustacean groups that were preserved as fossils during the Pleistocene. Although morphological and distributional evidence suggests that some of the smaller and thinly skeletonized lineages were present much earlier, many
groups are recorded for the first time during the Holocene. These include the nectiopodan remipedes, cephalocaris, mystacocaris, branchuri, tanaocaris, ascopheran and rhizocephalan cirriped, bathynelaeacean syncaris, thermodbaenaceans, mictacaris, euphausiacaris, and amphipodaceans. Many other groups have radiated or renewed their diversification during the Holocene. Showing remarkable evolutionary resilience since the beginning of the Paleozoic, the ostracods are undergoing a fifth major radiation. The cirripeds show considerable current diversity. The stomatopods, peracaris (especially isopods and amphipods), and decapods (dendrobranchiate shrimps, and especially the caridean shrimps, thalassinidean shrimps, anomurans, and brachyurans) all are undergoing major modern diversification. Although no major lineages have been lost during the Quaternary, the remipedes, cephalocaris, branchuri, remaining lower malacostracan lineages (phyllocaris, syncarids) and some of the ancient eumalacostracan lineages such as the palinurans and astacids often inhabit peripheral refugia and appear to have their greatest eras behind them.

C. Ecological Diversity in Crustaceans

1. The Habitats That Crustaceans Occupy

The major types of environments inhabited by crustaceans can be separated into 16 major, often overlapping, categories. Of all of these crustaceans are most commonly found in benthic marine habitats, including both level bottom and coarse (rock and reef) habitats (35 groups). Brackish water (estuarial delses and lagoons, brackish marine caves and tidepools, salty inland pools, and brackish groundwater) is the next most common habitat for which 23 groups (39%) of crustaceans are well suited. Seventeen taxa (44%) are pelagic in the upper part of the ocean (<200 m), and 18 (46%) groups are found in the deep sea (200 to >6000 m), either as pelagic or as benthic species on level bottoms or around hydrothermal vents. Fifteen taxa (38%) inhabit permanent bodies of fresh water (ponds, lakes, streams, rivers), including both benthic and pelagic components. Eight groups (21%) contain species that live in intertidal environments (some of these among sand grains), and 8 (21%) inhabit other semiterrestrial or terrestrial conditions (usually moist conditions or decaying vegetation). Some isopods, some amphipods, and a few brachyurans are truly terrestrial, having achieved independence from water for reproduction. Eight taxa (including 5 highly specialized groups) have adopted parasitism (21%), and 7 groups (18%) include commensals that live on or in other organisms. Seven groups or sections of taxa (18%) flourish at high latitudes, 5 (13%) inhabit either marine or freshwater caves (with various degrees of specialization), and 5 (13%) carry a fragile existence in inland ephemeral freshwater and brine pools. Three taxa (8%) can tolerate hypersaline (very salty) conditions, 3 (8%) are known from hot springs, and 3 (8%) live a specialized existence in groundwater. At least 1 group lives interstitially among the meiofauna (organisms smaller than 0.5 mm) in subtidal benthic marine sediments. The number of major taxa that occupy each of the various types of habitat may but does not necessarily reflect the number of species in that environment. This analysis, instead, demonstrates the diversity of major adaptations (in morphology, physiology, and behavior) that crustaceans have developed to accommodate particular environmental challenges. These adaptations provide the underlying framework, overlain by other specializations, on which evolution can proceed in the long run.

The same environmental categories can be used to examine ecological diversity among the different crustacean groups. It is likely that those groups that occupy the greatest diversity of habitats are the ones that will be able to survive massive environmental change, since not all habitats are likely to be equally affected by most forms of environmental disturbance. As shown in the discussion of taxonomic diversity, the maxillipods and the eumalacostracans (particularly the peracaris and euracaris) are the champions in terms of ability to occupy a diversity of different environments. Following a history of evolutionary success that stretches from the earliest Paleozoic to the present, the ostracods inhabit seven different types of environments. The comparatively recent (late Mesozoic) but ubiquitous copepods also occupy seven major types of habitats, including exceedingly specialized parasitic ways of life, and the ancient thoracican barnacles live in five major habitat types. Among the lower malacostracans, the anastomacaris are surprising in the diversity of seemingly marginal types of environments that they inhabit (five: permanent fresh water, ephemeral freshwater pools, caves, semiterrestrial, groundwater), and the mysidaceans have radiated into seven major habitats and lifestyles. Even more remarkable, the thermodbaenacean peracaris, with only 11 species, inhabit five different types of environment as categorized earlier (benthic marine, inland salt ponds, fresh and salty groundwater, caves, hot springs). The tanaid peracaris and the anomuran encarids occupy five and six major types of habitats. The caridean shrimps and brachyuran crabs (decapod
eucarids) and the isopod and amphipod peracarids, however, inhabit more different types of environment than any other groups of crustaceans (seven, eight, nine, and nine types of habitat, respectively). Although many carideans and some brachyurans live commensally with other organisms, a fully parasitic existence is the only habitat that these eucarids have not exploited as fully as have some isopods and amphipods. Although caridean shrimps have been at least as successful in fresh water and are more diverse in pelagic environments than the brachyurans, they have failed to invade terrestrial habitats as completely as have the brachyurans, amphipods, and isopods. The isopods are the most successful terrestrial crustaceans, and they, along with most other peracarids, flourish in cold, high-latitude environments as well as the deep sea. Amphipods are less specialized for parasitism and terrestrial existence than the isopods. Like the amphipods and isopods, the carideans and brachyurans have radiated into all major marine environments, including the deep sea and its hydrothermal vents. The carideans likely invaded fresh water very early (Mesozoic), whereas the brachyurans undertook major invasions of freshwater and terrestrial habitats in the Late Tertiary and Quaternary.

A number of crustaceans also modify their microenvironment by commandeering empty shells or other structures, engineering tubes or burrows, and camouflaging themselves from pervasive predation in various ways. Anomurans, tanaids, and amphipods all inhabit vacated gastropod shells; anomuran hermit crabs also inhabit tusk shells, hollow twigs and bamboo, calcareous worm and vermetid mollusk tubes, hollow pieces of coral, and even bottle caps. Some anomurans plant anemones that function in camouflage and defense on their gastropod home. Some dromid brachyuran crabs carry a clam shell on their backs, and lathotid and majid brachyurans culture domes of sponge or decorate their carapace with algae. Tanaid and amphipod peracarids construct pebble tubes, and amphipods modify algal structures for domiciles or camouflage. One group of amphipods bores into wood, one group of isopods bores burrows in silt and clay, and acrothoracican barnacles bore into coral to make protective lodgings. Coral-dwelling stomatopods chip, scrape, and modify preexisting holes in coral or reef substratum to suit their needs, and most groups of crustaceans are well represented in preformed crevices and bioeroded holes in hard, especially calcareous, substrata (coral, limestone and reef basement, oyster and tube worm reefs). Many of these organisms, such as the thalassinideans, line and modify the preexisting crevice with silt and mucous for a better fit. Some groups of carideans snapping shrimps construct smoothly lined networks of burrows within sponges. Large species, such as stone crabs, coral crabs, spider crabs, spiny lobsters, and some stomatopods inhabit refuge space beneath bioeroded ledges and caverns on reefs. Several groups of stomatopods, astacid crayfishes and lobsters, thalassinidean shrimps, and brachyuran land crabs, fiddler crabs, and ghost crabs all excavate large (in some species very complex) burrows in mud and sand environments. All of these habitat modifications reflect the pervasive intensity of predation by epibenthic and benthic predators such as cephalopods (particularly octopuses), other crustaceans, and especially fishes. By increasing the complexity of the microenvironment, many of these structures increase the potential for commensal associations and heighten the capacity of the environment to support high species diversity.

2. The Role of Crustaceans in Biological Communities

Examination of the trophic position of crustacean groups indicates that most crustaceans have relatively generalized feeding habits and most are subject to strong predation. Their abundance, moderate size, and generalized carnivory likely exert top-down control in some communities. Also, as evidenced by almost universal morphological and behavioral adaptations against predation as well as their importance in the diets of many different predators, the availability of crustaceans as prey for other trophic levels probably is essential for the structure, function, and diversity of many aquatic communities. Cladocerans and copepods support higher trophic levels, including fishes, in freshwater habitats. When their ephemeral habitats are located, anostracans and notostracans are heavily consumed by birds. Anaspidacean syncarids are heavily preyed upon by freshwater fishes if the shrimp are dislodged from their refuges. Stomatopods, mysidans, tanaids, cumaceans, isopods, amphipods, dendobranchiate shrimps, caridean shrimps, anomurans, and brachyurans all are common in both epibenthic and benthic fish stomachs; cumaceans, isopods, and amphipods are particularly important in the deep sea. Dendobranchiate shrimps and mud-dwelling stomatopods are important in the diets of many large, commercially important fish, and euphausiaceans and pelagic anomurans are a major source of food for whales, squid, tunas, and seabirds.

Most crustaceans are either generalized predators and scavengers (30 taxa or portions of taxa), or detritus and suspension feeders (35 taxa or parts of groups). Many biologists consider the predaceous, scavenging
remipedes to exemplify the primitive feeding mode, but detrital suspension feeding, as seen in the phyllopodous cephalocarids, branchiopods, and some of the primitive phyllocarid malacostracans, obviously can be traced far back in crustacean history as well. Herbivory is not particularly well developed in crustaceans, although many of the phyllopodous groups (cephalocarids, branchiopods, leptostracans) may ingest phytoplankton and plant material along with other detritus. The suspension feeding ascothoracid, acrothoracid, and thoracic barnacles consume phytoplankton; and some ostracods, isopods, amphipods, and brachyurans feed on micro- or macroscopic plants (some of these species also eat detritus or other organisms as well). Euphausiaceans, however, are important grazers of phytoplankton at high latitudes, and copepods may consume more primary productivity than any other organism on earth because of their ubiquity and consumption of phytoplankton in the largest habitat on earth, the open ocean. The mouthparts of crustaceans are relatively easily modified for piercing and tearing flesh, so strict carnivory and ectoparasitic habits, with the potential for considerable impact on host populations, have evolved in a number of groups, including a cladoceran, branchiurans, tuntulocaridans, some ascothoracidian barnacles, isopods, and amphipods. Parasitism is most extremely developed in the endoparasitic copepods and rhizocephalan barnacles; many ascothoracidian barnacles and some isopods also live inside and consume the energy or tissues of their hosts. Large body size and armored claws that smash, crush, or shear hard-bodied prey make stomatopods, Palinura and astacid lobsters and crayfish, and anomuran and brachyuran crabs potent predators that, near the top of their benthic food webs, exert important effect on their local communities.

3. Life History Patterns of Crustaceans

Although most crustacean groups reproduce sexually, a number of groups, particularly in fresh water and high latitude environments, produce offspring without exchanging eggs and sperm. "Parthenogenesis," a form of asexual reproduction where eggs undergo development without fertilization, is found almost exclusively in the Branchiopoda and Maxillopoda. Other forms of asexual reproduction, such as budding of new individuals from a previous individual, do not appear to be present in the Crustacea, although regeneration of missing body parts is common. In the branchiopods, most anostracans reproduce sexually (and produce fertilized resistant eggs that are dormant during unfavorable conditions), but some populations are exclusively parthenogenetic. Parthenogenesis is common in both notostracans and cladocerans. Some notostracan populations are exclusively parthenogenetic, especially at high latitudes. Notostracans also reproduce sexually, producing fertilized resistant eggs that can withstand harsh conditions. Some species of cladocerans produce males and undergo sexual reproduction only during periods of environmental deterioration, such as autumn; males may be dwarfed and short lived. Some species of conchostracans are exclusively parthenogenetic, and other species may include both parthenogenetic and sexual individuals. In the maxillopods, some freshwater ostracods are parthenogenetic, and other species are both parthenogenetic and sexual. Some harpactocoid copepods are parthenogenetic. Some copepods produce resting eggs that are dormant for extended periods. Of the rare deep-sea tuntulocaridans, only females are known, though parthenogenesis is not proven. Evidence suggests parthenogenesis in some thoracican barnacles. Some terrestrial isopods are parthenogenetic.

Reduction of the prominence of males, as evidenced by short-lived, usually dwarfed males that frequently live on or in the female, is seen in the freshwater cladocerans, parasitic ascothoracidian barnacles (dwarf male free living or lives on female), boring acrothoracidian barnacles (lives on female), sessile thoracican barnacles (lives on hermaphroditic individual or female), parasitic rhizocephalan barnacles (short-lived dwarf male invades extruded female gonad), and parasitic isopods (lives on female). In the deep sea gnathiid isopods, the second maxillipeds are reduced to flaps that cover the mouth of the nonfeeding male. Few males are known (although sexual dimorphism appears in development) in the pelagic deep sea amphipodaceans, and populations of the primitive procaridid shrimps are composed mostly of females; these skewed sex ratios may be the result of differential mortality in males, reduction in the prominence of the male life cycle, or flexible sex determination (as is common in many crustacean groups) rather than a propensity for parthenogenesis.

The reproductive organs of crustaceans are relatively simple, and both testes and ovaries apparently develop readily, with dominance shifted from one to the other depending on hormonal and environmental conditions. The primitive remipede and cephalocarids are hermaphroditic. Although most sexually reproducing branchiopods are gonochoristic, some species of notostracans are hermaphroditic and likely undergo sex reversal. In the maxillopods, the sex of the sessile cirripeds is notably flexible. Although most of the parasitic ascothoracidian barnacles are gonochoristic, species in one group are simultaneous hermaphrodites and those in another group are protandrous hermaphrodites. In
others, the first arrival to a host becomes female, and
the next arrival becomes a male. Most of the sessile
thoracican barnacles are simultaneously hermaphro-
dritic or mixed hermaphroditic and separate sexes
dwarf males may reside either on females or hermaph-
roditic). Most acrothoracid and rhizocephalan barna-
cules, however, are gonochoristic. The other center of
sex change is found in the malacostracans. In the pera-
cardis, tanaidaceans usually are protogynous hermaph-
roditic, or the same population may have both go-
nochoristic and hermaphroditic individuals. Sex change
appears to be influenced by environmental factors, pos-
sibly even the ratio of males to females as affected by
predation. Most isopods are gonochoristic, but some
(epecially freshwater) lineages change sex. Both prot-
androus and protogynous species are known, and both
gonochoristic and hermaphroditic individuals may oc-
cur in the same population. Particularly in parasitic
isopods, sex is determined by the presence or absence
of another individual; first arrivals become females and
the second arrival becomes a dwarf male that resides
on the resident female. Most amphipods also are go-
nochoristic, but protandrous species are known, and some
species are composed of both protandrous and go-
nochoristic individuals. Although most caridean shrimps
have permanent separate sexes, some species are simul-
taneous hermaphrodites and some are protandrous.
Both gonochoristic and hermaphroditic individuals
may occur within some populations. Sexual differentia-
tion may be affected by both social and environmental
factors. For example, some species of snapping shrimp
live in colonies of up to 350 individuals in a sponge,
with only one large reproductive female, the "queen." A
host of smaller male or nonreproductive "workers"
defend the colony and maintain communal burrows.

Sexual reproduction often is accompanied by mor-
phological structures that have differentiated between
males and females, probably as the result of sexual
selection. Males often are adapted for greater motility
than females. In the ascothoracid barnacles, the fe-
male's body is enveloped entirely within the carapace,
while most of the swimming appendages are exposed
in males. Male tanaiids have well-developed pleopods,
but abdominal appendages are absent in females; males
also have larger eyes and more antennal sensory struc-
tures than females. Male cumaceans bear abdominal
pleopods while females do not, the antennae are re-
duced in females but not males, and the eyes are more
reduced in females than males. The carapaces of female
thermosbaenaceans and amphionidaceans are very en-
larged, while those of the males are reduced, exposing
swimming appendages. The last pair of thoracic legs
are absent in female amphionidaceans but not in males,
and the females, with reduced mouthparts and a vesti-
gial gut, do not feed.

Various appendages have been modified in many
groups to facilitate mating. The first pair of maxillae
(maxillules) were modified as a clasper in males of the
Devonian Liposarcans (allied to the cephalocarcids).
The antennae are modified for detecting or holding
females in male anostracans; some cladocerans, ostra-
cods, and cumaceans; and in many amphipods. The
size and shape of the antennules of male leptostracans
and euphausiaceans and of some species of copepods
are modified, probably for clasping females. Other sec-
ondary sexual characteristics include copulatory struc-
tures, which are common in crustaceans. Anostracans
have a pair of eversible penes on the first two abdominal
segments. The leaflike appendages of the first two trunk
segments are modified for copulation in conchostrac-
a; the first two trunk limbs of some genera of ostra-
cods are modified to assist mating, and the third trunk
appendages often are modified into copulatory organs.
The fourth trunk appendages, sixth thoracopods, and
second through fourth thoracopods of male mystacocar-
cids, copepods, and branchiurans, respectively, are mod-
ified for copulation. Abdominal appendages have been
fused into a single long median penis, the only re-
maining vestige of the abdomen, in thoracican barna-
cles. Paired penes also are fused into a single median
penis in the tanaulocarids and some isopods. A pair of
copulatory tubes are associated with the eighth thoraco-
pods in male stomatopods, and the eighth thoracopods
are modified, probably for copulation, in the bithynella-
cians. The bases of the eighth thoracopods in male
thermosbaenaceans bear a pair of well-developed penes.
Male miictaceans sport a copulatory process on the
eighth abdominal segment, and, although pleopods are
otherwise absent in both males and females, male miicta-
cians retain the second pair of pleopods, probably to
assist transfer of sperm. The anterior abdominal pleo-
pods are structurally modified to assist copulation in
many male malacostracans, including stomatopods,
mysidans, isopods, amphipods, and dendrobranchiate
shrimps. The anterior pleopods of both male and female
euphausiaceans are elaborately modified for copulation.

Several groups of crustaceans exhibit strong sexual
dimorphism in body size and other structures, includ-
ing the late Paleozoic pygoccephalomorph mysidaceans.
Sexual dimorphism can be observed among developing
juvenile amphionidaceans after stage 9. Male tanaiids
show pronounced sexual dimorphism in a variety of
sensory structures as well as relative size of the chelipeds.
The overall morphology of male and female ta-
nads becomes so highly modified as they molt into copulatory versus noncopulatory phases that they frequently have been misclassified as different taxa. Amphipods and caridean shrimps show similar variability of body size and shape between copulatory and noncopulatory stages, probably reflecting the impact of strong selection for attracting and defending mates against other suitors. Some groups of isopods (especially terrestrial lineages) show extreme dimorphism in size, with males greatly exceeding the size of females. The claws of many decapods (including some caridean shrimps, some anomurans, and particularly some thalassinidean shrimps, astacid crayfish and lobsters, and many groups of brachyuran crabs) are greatly enlarged in males. Either both or only one of the claws may be enlarged for fighting or behavioral displays. Mating or territorial displays often are associated with enlargement of only one claw, as in the fiddler crabs that attempt to attract females to their burrows with colorful choreographic waving displays. Though not sexually dimorphic, stomatopods also seek to attract mates by exposing species-specific colored eyespots and iridescent patterns on their spread maxillipods. Several groups, including land crabs and fiddler crabs, use sound vibrations (usually stridulating or thumping the walls of their burrow) to attract mates. Various other crustaceans stridulate (stomatopods, some isopods, palinurans, other brachyurans), but these sounds may function to deter attacking predators rather than to attract mates. Bioluminescent photophores on euphausiaceans may facilitate sexual identification, although this is not proven. One of the most sensational courtship displays in crustaceans is created by male ostracods that emerge from coral reef substratum about an hour after dark, emitting sequential pulses of bioluminescent material from the maxillary glands in a species-specific pattern as they ascend toward the surface. Different species deposit the discrete pulses of light in uniformly spaced vertical strings, suspend them in a line at a 45° angle from the substratum, or place sequential clusters at different heights. Females do not emit bioluminescent puffs, but emerge from the reef and follow the trail of lights to mate with the displaying males. A horde of nonluminescent males also follow the light trails, apparently to "sneak" matings with the pursuing females.

Most crustaceans produce round nonmotile sperm, although the sperm of some maxilliped groups are flagellated. The sperm, released either in seminal fluid or in packets of spermatophores, are deposited directly into the female genital openings or into a fused medial receptacle. Females may store sperm for long periods (up to several years in some astacid lobsters), although subsequent matings can dilute the sperm of a particular male.

Most crustacean matings are relatively brief (minutes to hours), but some crustaceans undertake long arduous courtships and maintain long-term pair bonds. Since little paternal care or mate provisioning is documented, most pair bonds probably function to ensure that another male does not also fertilize the female. Males of isopods, amphipods, and brachyuran crabs often sequester females, holding and carrying the smaller female for extended periods prior to copulation, which frequently follows female molting. In the violently aggressive and solitary stomatopods, mating is accompanied by an extended period of courtship (involving both visual displays and tactile stroking) that gradually mediates the tendency to smash the potential mate with the clublike or spurring raptorial appendages. The male and female reside together (either in the female's or male's burrow, depending on the taxonomic group) for a number of days, copulating repeatedly. After as much as a week, the female's aggressive levels rise and she evicts the male before laying the eggs. Commensal pinnotherid crabs ("pea crabs," which inhabit oysters and other organisms' burrows), parasitic isopods, stenopodid shrimps, and many groups of caridean shrimps usually are found in pairs, and probably maintain extended (seasonal or permanent) pair bonds. Some deep-sea stenopodid shrimps apparently colonize glass sponges as postlarvae, and the male and female (perhaps the opposite sex is induced in the later arrival) become permanently imprisoned in the crystalline gridwork of the sponge as they grow. These life-long pair bonds are the source of a romantic oriental legend.

Most crustaceans brood their young, but there is a surprising diversity in methods of treating the young, and some release the embryos free into the environment. The remipedes probably shed their embryos directly into the water (no brooding structures have been observed). In the branchiopods, the notostracans brood their embryos very briefly, then deposit them on the substratum, while most cladocerans release their embryos into the water. In the maxillipods, some ostracods release embryos directly into the water and some deposit them on the substratum. The interstitial mystacocarids apparently lay their embryos free among sand grains. Some copepods release embryos directly into water. Branchiurans leave their host to attach their embryos to a plant or rock. In the malacostracans, the extinct Late Paleozoic belotalsondeans, waterstonelli- deans, and palaeocaridaceans probably shed their embryos directly into the water, since no structural adaptations for brooding have been seen. Contemporary
anaspidean syncarids, and probably bathynellaceans, release their embryos directly into the environment. Contemporary stomatopods deposit their eggs in a free wheel-shaped disk that the female holds in her maxillipeds, grooming and defending them for several weeks. Most euphausiaceans and most dendrobranchiate shrimps shed their embryos directly into the water.

However, most crustaceans exhibit a variety of structural, physiological, and behavioral modifications to better protect their progeny. Cephalocarids brood one or two embryos at a time on anterior abdominal processes, and the allied Devonian lipostracans had a flap-like egg pouch. Among the branchiopods, the anostracans carry their eggs briefly in an egg sac before releasing them. Some cladocerans shelter their eggs in their carapace; the embryos even obtain nutrients from the sides of the brood chamber in some species. One species of cladoceran broods eggs in the molted carapace. Conchostracans carry their embryos on the thoracopods inside the bivalved carapace. Most ostracods brood their young inside the bivalved carapace. Copepods usually drag an egg sac behind the body, but some brood embryos on the sixth thoracopods. In the ascothoracid cirripeds, the female’s carapace is enlarged (enclosing her body) for brooding the young. Embryos are released at the nauplius stage or retained in the carapace through all larval stages up to the cyprid. The embryos of acrothoracic cirripeds hatch at an early stage but are brooded in the carapace through the cyprid stage. The body of thoracican barnacles is engulfed by a fleshy mantle cavity in which the embryos are brooded through the nauplius stage. In the malacostracans, leptostracan females brood embryos under their carapace until after two larval molts. As noted earlier, female stomatopods undertake extensive care and protection for their free egg mass within their cavities in coral or excavated burrows in sand or mud. Female mysids bear a brood pouch on the last two or three thoracopods. The enlarged, thin, blood-filled carapace of female thermostaenaeaceans serves as a brood chamber as well as for respiration. Similarly, the carapace of female amphipodaceans is thin and greatly enlarged, and it may be used for brooding. In this group, the first pair of pleopods, greatly enlarged and extending anteriorly beneath the carapace, are thought to assist in brooding the young. The brood pouch is particularly well developed in the tanaids, cumaceans, mictaceans, isopods, and amphipods. This “marsupium” is formed from expanded flaps, called oostegites, at the bases of thoracopods 3 to 6 in tanaids and cumaceans. Sometimes the oostegites are restricted to thoropod six in tanaids. Cumaceans, like isopods, bear an additional oostegite projecting toward the posterior from the maxillipeds. Female mictaceans bear oostegites either on thoracopods 2 to 6 or 3 to 7. Female isopods have oostegites on the second to fifth, second to sixth, or all eight thoracopods. One family forms a brood pouch from ventral extensions of the body wall rather than oostegite projections from the base of the legs. Several types of internal brooding are present in isopods, including paired invaginated pockets in the cuticle with narrow openings and a single median pocket extending into the thorax between the sixth and seventh segments. Two groups (the gnathiids and flabelligerans) brood young within the body cavity itself; in the flabelligerans, the enlarged oviducts serve as uteri and the oostegites are reduced to small flaps covering the gonopore. Female amphipods develop oostegites on the third to fifth, third to sixth, or fourth to fifth thoracopods; many amphipod taxa also have simple or elaborate gills that protrude from the base of the thoracopods; in some cases the oostegites are reduced and the gills provide a brood chamber. A few euphausiaceans and one species of dendrobranchiate shrimp brood embryos on their posterior thoracopods. Most other decapods, including the carideans, stenopodideans, thalassinideans, astacids, palinurans, anomurans, and brachyurans, brood embryos on their pleopods.

Crustacean embryos proceed through some form of spiral cleavage, often highly modified by the amount of yolk present, to form a nauplius larva, a triangular shaped organism (larger in the anterior) with a simple median eye (ocellus) and three pairs of biramous appendages (the future antennules, antennae, and mandibles). The embryo may either emerge as a planktonic nauplius when yolk reserves are depleted, or the nauplius may be laden with yolk, reducing the need to feed until a later stage. Larvae hatching as nauplii require long developmental times and are observed in many of the “less advanced” groups (some anostracans, some sotrostracans, some cladocerans, conchostracans [yolkey], some ostracods, mystacocarids [develop in interstitial sand], nonparasitic copepods [nonfeeding], some ascothoracid cirripeds, some acrothoracic cirripeds, some thoracican cirripeds, rhizocephalan cirripeds [nonfeeding], euphausiaceans, and dendrobranchiate shrimps). Evolution appears to have progressed toward hatching at a more advanced stage, either as a metanauplius (more segments but still only three appendages), a protozoa (where a metanaupliar stage undergoes a metamorphic molt within the egg membrane to a larval stage with sessile compound eyes, maxillules, and maxillae) or a zoa (a more advanced larval form with stalked eyes and three pairs of maxill-
peds; these swimming larvae eventually hatch into a postlarval form resembling an immature adult). The cephalocarids, fossil lipostracans, some anostracans, some natostracans, some cladocerans, some ostracods, some ascothoracidan cirripeds, and some acrothoraci-
can cirripeds hatch as meta nauplii. The reduction or loss of the free larval stage is accomplished by retaining the naupliar, metanaupliar, or protozoal stages within the egg membrane during a longer brooding period before hatching or, in some cases, by shortening the total development time by omitting the morphological development of some or all of these larval stages entirely.

Crustaceans have three types of development, including "direct," "anamorphic," and "metamorphic" development; the last two forms sometimes are called less and more extreme types of "mixed" development. In direct development, embryos hatch as juveniles that resemble miniature adults, either because larval stages are passed inside brooded eggs before hatching, because they have been suppressed or lost, or possibly because they never were present. Direct development probably occurs in remipeds and probably was present in the Cambrian canadaspid malacostracans, either because larvae were suppressed or never present. Also, most cladocerans, most ostracods, most branchiurans, leptostracans, anaspidaceans and bathyanellicans, and all of the peracarids show direct development. The young hatch as a juvenile in some freshwater and terrestrial astacids and branchiurans and in some taxa of very small body size, including some stomatopods and some of the very small caridean shrimps. Most of the instances of abbreviated larval phases probably are explained by passage of larval stages within brooded eggs, but some cladocerans, some ostracods, branchiurans, and the anaspidacean and bathyanellican syncarids do not brood. Either larval stages were suppressed or never were present in these groups.

In anamorphic development, embryos hatch as some form of nauplius, metanauplius, or later-stage larva, and the adult form is gradually achieved by a progression of morphological changes as the larva acquires more segments and appendages. Anamorphic development is found in many cephalocarids and allied fossil lipostracans, anostracans, natostracans, some cladocerans, conchostracans, some ostracods, mystacocarids, copepods, and one group of branchiurans. In the stomatopods, euphausiaceans and amphipodaceans, the morphology of the larvae generally resembles the form of the adult (thus not requiring a major metamorphosis) so these groups could be considered to have anamorphic development; however, these three groups usually are con-
sidered to have metamorphic development because they have protozoal and zoal larval forms.

In metamorphic development, the embryo hatches as some form of nauplius, metanauplius, or later-stage larva and develops with a body form that is radically different from that of the adult; thus, a major morphological reorganization (similar to that of holometabolous insects such as caterpillars and butterflies) is required before the adult stage can be attained. All of the euphausiids as well as the stomatopods usually are considered to have metamorphic development because they possess protozoal and zoal larval forms. Other groups that undergo a major metamorphosis between the larval and adult phases include the cirripeds, although they lack the typical protozoal and zoal stages. The cirripeds are unique in developing a morphologically distinct nonfeeding bivalved postlarval form called a cypride before transforming into the adult phase. In thoracicans, the cypride attaches to the substratum with its antennule and reorganizes its body into a sessile adult. In addition to a cypride stage, boring acrothoracicans metamorphose into a postcypride pupa stage in order to reorganize itself into the adult form, similar to holometabolous insects. Following four feeding nauplius stages and a nonfeeding cypride stage that attaches to the prospective host, the parasitic rhizopods form a hypodermic-like "kentrogon" stage through which the body, completely transformed, injects itself into the host's body. The parasitic tintalocarids and copepods also undergo metamorphoses in order to attain the highly modified adult parasitic form. The nauplius larval forms are suppressed in parasitic copepods, which hatch as copepodes (small juveniles).

The number of eggs produced by crustaceans generally correlates with body size, with larger species within lineages usually producing vastly more eggs than smaller species. This tracks only within lineages, however, and, if multiple lineages are combined, there is no trend. Certain crustacean groups are characterized by "weedy" life history characteristics while others are long lived and slow to reproduce. Anostracans and natostracans usually live only a few weeks and produce many eggs, including resistant eggs that are capable of surviving long periods of unfavorable conditions, similar to the seed banks of desert plants. Some copepods also produce dormant eggs. Cladocerans pass through several generations in one summer season. Amphipodaceans may be short-lived as adults (the female does not feed), and males in several taxa (gnathiid isopods, dwarf males in rhizocephalan cirripeds) are short-lived and do not feed. On the other hand, mysids reproduce infrequently and may require up to 7 years
to mature. Some Arctic cumaceans have delayed maturity and live many years, reproducing little. Some stomatopods, astacid and Palinurinae decapods, and some brachyuran crabs live many years.

The significance of the type of development and number of eggs produced in crustaceans lies in their relationship to dispersal. Species that produce more eggs and those that are released at early stages so that longer periods are spent in the plankton have the largest geographic ranges and are the most resistant to environmental change and extinction. The resistant eggs of anostracans and notostracans are dispersed by wind or on the feet of birds. Nonparasitic copepods develop through six naupliar stages and five copepodite stages. Stomatopods pass several protozoal stages (advanced naupliar or metanaupliar stages) in the egg while brooded by the female. The protozoal stage, termed an "antinozoa" or "pseudozoa," is followed by four to eight or nine zoal stages, called "eritozous" or "alma" larvae, in small reef and larger level bottom lineages, respectively. These modified zoae are pelagic for periods that range from 3 to 4 weeks to 9 months in small reef versus larger level bottom species, respectively. A minority of small species of stomatopods produce only a few, unusually large eggs relative to their body size and undergo their entire larval development within the brood of the female. Stenopodid shrimps, palinurian lobsters, and some anomurans are characterized by many larval stages and long pelagic phases. Most small species, most species that inhabit fresh water and terrestrial habitats, many species that inhabit high latitude or deep cold environments, and species with anamorphic or metamorphic development that hatch at advanced stages are likely to produce fewer, relatively larger eggs with shorter pelagic phases. These species, and those with direct development, are most likely to have extremely restricted, endemic distributions.

4. The Geographic Distributions of Crustaceans

Many of the geographic distributions for each of the major groups of crustaceans strongly reflect ancient conditions where the group originated. Dispersal (usually by long lived pelagic larvae) has not dominated the historical ecology of many groups of crustaceans. The importance of intermediate dispersal in speciation, however, where larvae travel far enough to establish new populations but not sufficiently frequently to swamp arising genetic variants, is well illustrated in the thoracian barnacles. Also, the long lived pelagic phases of some of the large level bottom stomatopods have been shown to dampen diversification within and among regions, while those lineages with short lived larvae and significantly more restricted geographic ranges rapidly speciate and also suffer higher rates of apparent extinction (Reaka 1980; Reaka and Manning, 1981, 1987). The distributions of fossil taxa are not extensively considered below because, except for the worldwide distributions of the phyllocarid malacostracans, many extinct taxa are recorded from relatively few regions, possibly as a result of preservation bias.

To understand patterns of distribution in contemporary crustaceans, data for all of the major taxa were divided into nested categories where the entire group (subclass, order, or suborder) had a wide (>half the circumference of the earth) or narrow (≤half the earth's circumference) global distribution, those with many (>100) or few species (≤100), and those in which most of the individual species had wide or narrow geographic ranges, producing a table of eight columns. This procedure revealed the somewhat surprising result that entire lineages of crustaceans with narrow distributions are rare.

Restricted geographic distributions for entire lineages are found only in four taxa (the nectiopodan remipeds, anaspidacean isopods, theromolanaean pancarids/peracarids, and speleaeophylicae peracarids). Each of these groups also contains few species (10, 15, 11, and 2, respectively) that all have highly restricted individual geographic ranges (one or a few localities). These four lineages likely represent the remnants of once more broadly distributed groups that have experienced extinction outside of the marginal refugia they currently inhabit. There are no lineages with narrow distributions and few but broadly distributed species, and no lineages with narrow distributions and many species with either wide or narrow species ranges.

Among groups where the entire taxon has a broad distribution, only six groups have low species diversity and contain species with restricted geographic ranges. Phyllocarids [10 species], tapiocarids [15 species], bathynellaceans [100 species], micraneans [3 species], procarid shrimps [3 species], and stenopodid shrimps [25 species]. Widely distributed lineages with few species but broad species ranges include another seven groups (notostacran [11 species], mystacocarids [10 species], acrothoracicans [most of the 50 species], leptostacran [most of the 13 species], lophogastriids [most of the 40 species], euphausiaceans [more than half of the 90 species], and the single, though variable, amphipodacean species).

Groups in which the entire taxon is widely distributed on a global scale and that contain many species...
with broad species ranges are not numerous. Only most
dendrobranchiate shrimps and most palinurans fit
clearly within this category, although whether the ma-
jority of copepods, cumaceans, anomurans, and brachy-
urans have wide or narrow individual species ranges is
not well documented.

By far the most common distribution pattern ob-
served in crustaceans includes lineages that are speciose
(>100 species), broadly distributed on a global scale,
and contain species with relatively restricted geographic
ranges (14 groups, 18 if copepods, cumaceans, anomur-
ans, and brachyurans, whose status is not certain, are
included). These lineages include most of the ostracods
(>8000 species), anostracans (>180 species), probably
most of the cladocerans (>450 species), most of the
conchostracans, (>180 species), most of the branchiur-
ans (150 species), most of the thysanurans (>800 spe-
cies), most of the stomatopods (>400 species), proba-
bly most of the tanaids (>850 species), most of the
isopods (>4000 species), most of the amphipods
(>6000 species), probably most of the carideans (2500
species), and probably most of the thalassinoideans and
astacideans.

In summary, the major lineages of crustaceans are
most commonly broadly distributed on a global scale
(29 of the 33 taxa documented). However, the dominant
pattern for species is to have relatively restricted geo-
graphic ranges. In at least 24 of the 33 groups docu-
mented, the majority of species have relatively small
graphic ranges, in some cases restricted to a single
locality. Lineages containing many species and those
with few species are about equally represented (16 and
17, respectively).

D. Vulnerability to Extinction

1. Extinction in the Fossil Record

Extinction in the fossil record tells us a considerable
amount about how prone modern crustaceans are to
extinction. Studies show that the breadth of geographic
distribution of a species or lineage, which is intimately
intertwined with their larval dispersal potential, is the
single most important predictor of extinction. Factors
that favor extinction or survival, however, depend on
whether one is considering organisms during normal
levels of background extinction or during cataclysmic
mass extinctions (Jablonski, 1991). During background
conditions, lineages with long-lived planktonic larvae
persist longer in the fossil record (i.e., have lower ex-
tinction rates) than those with abbreviated develop-
ment, individual species with broad geographic ranges
suffer lower extinction than those with narrow endemic
distributions, and lineages characterized by high species
richness survive longer than those with few species.
Species-rich lineages composed of broadly distributed
species have especially high survival. During mass ex-
tinction episodes, however, all of these traits are ineef-
cient, and the broad geographic deployment of an entire
lineage, regardless of the ranges of the individual species
or the number of component species, is the primary
characteristic that increases the likelihood that the lin-
eage will persist into the next geological era (Jablonski,

Jablonski (1991) designated nine major mass extinc-
tions (Late Ordovician, 61% and 85% of genera and
species lost; Late Devonian, 55% and 82% of genera
and species lost; Late Permian, 84% and 96% of genera
and species lost; Late Triassic, 47% and 76% of genera
and species lost; Middle Jurassic, 26% and 53% of gen-
era and species lost; End Jurassic, 21% and 45% of
genera and species lost; Middle Cretaceous 26% and
53% of genera and species lost; End Cretaceous 47%
and 76% of genera and species lost; Late Eocene, 15%
and 35% of genera and species lost). Of 15 crustacean
lineages that became extinct, 9 of them did not persist
beyond one of these biotic crises. Although this does
not prove that the mass extinctions were directly re-
sponsible for their demise, the failure of lineages to
survive into the next period, especially in taxa that had
persisted through several previous geological periods,
suggests that the conditions associated with the mass
extinction may have contributed to their disappearance.
In many cases, representation in the fossil record is
too spotty to determine whether the lineage and its
component species had broad geographic ranges, but
some useful information can be gleaned from what we
know about fossil crustaceans. The liposarcans (allied
to the cephalocarids) and the archaeostomatopod
hoplocarids are known only from the Devonian, and the
kazachthorans (allied to the nitostracans) are recorded
only from the Jurassic. Mass extinctions closed both of
these periods, and it appears that these lineages were not
widely distributed, being known only from Laurentia
(North America and Europe), Scotland, and Kazach-
stan, respectively. Also, only a few species are known in
each lineage. Known only from equatorial Laurentia
in the Cambrian and Ordovician, the hymenostracan
phyllocarid malacostracans did not survive beyond the
Late Ordovician mass extinction. The extraordinarily
successful archaeostacan phyllocarid malacostracans
persisted from the beginning until the end of the Paleo-
zoic, making it through both the Late Ordovician and
Late Devonian crises but failing to survive the greatest of mass extinctions in the Late Permian. This lineage had a worldwide distribution and included many taxa that occupied diverse habitats; these factors may have facilitated their persistence through several major crises but were not sufficient to allow survival through the End of Paleozoic debacle. Other primitive malacostracans, including the waterstonellideans, eocaridaceans, paleocaridacean syncarids, and pygocephalomorph mysidaceans, arose in the Carboniferous but also perished during the End of Paleozoic extinction. On the other hand, however, one order of the primitive remipedes (the enantiopods), the canadaspids and hoplostracan phyllocarid malacostracans, the aeschornectid hoplocarids, and the belotelsonid malacostracans, all known only from the Carboniferous, and the paleostomatopod hoplocarids, which originated in the Devonian and persisted until the Carboniferous, all perished in the absence of a major biotic crisis.

Tropical, especially reef, biotas have throughout history suffered disproportionately greater losses to mass extinctions than those at higher latitudes (Jablonski, 1991). Perusal of the habitats and distributions of the extinct fossil crustaceans suggests that most of those lost occurred in regions that were tropical at the time. The tropics also are the largest source of evolutionary novelties (due to speciation and development of major new adaptive types). This is possibly because the recurring extinction events open new evolutionary opportunities there, because high diversity in the tropics makes origination of new types more probabilistic simply because of the large numbers of lineages available for genetic experiments, or because something about the ecology of tropical environments causes greater divergence per speciation event (Jablonski, 1993).

Low density and small total population size also has been shown to render species susceptible to extinction in some taxa (Raup, 1986), and this characteristic has been of overriding concern to many conservation biologists in recent years. Evidence on the importance of population density as a predictor of extinction is equivocal in crustaceans. On one hand, the archaeostacan phyllocarid malacostracans were one of the most abundant benthic organisms, and they persisted from the Cambrian until the end of the Paleozoic, supporting the thesis that population density is important. On the other hand, however, the archaeostomatopod hoplocarids were abundant during the Devonian but did not survive into the next period. Aeschornectid hoplocarids dominated their benthic communities and belotelsonid malacostracans were abundant during the Carboniferous, but neither persisted beyond that period. Waterstonellidean malacostracans formed abundant aggregations, paleocaridacean syncarids were very numerous, and pygocephalomorph mysidaceans were among the most important components of their communities during the Late Paleozoic (Carboniferous to Permian); yet all succumbed to extinction at the end of the Permian.

Similarly, when a species occupies a unique or particularly restricted type of habitat, environmental changes are more likely to entirely eliminate these localized populations, and individuals may be unable to find or migrate to a similar specialized environment. This feature of organism-environmental interaction also has been a dominant concern in conservation biology. The evidence from fossil crustaceans on this topic, however, is not clear. An example of a lineage with a very specialized habitat that became extinct is the lipostracans, allied to the cephalocaris, that lived in Devonian hot springs. However, other crustaceans that became extinct were not obviously specialized for unique habitat types and inhabited apparently generalized benthic coastal environments. Enantiopodan remipedes; canadaspids and hoplostracan phyllocarid malacostracans; belotelsonid and eocaridacean malacostracans; archaeostomatopod, aeschornectid, and paleostomatopod hoplocarids; paleocaridacean syncarids, and pygocephalomorph mysidaceans all inhabited shallow benthic coastal environments. Of these, at least the abundant aeschornectids, pygocephalomorphs, and paleocaridaceans also co-occurred in brackish swamps and deltas; and at least the paleocaridaceans extended into fresh water. The latter four lineages possibly could be considered to exhibit some habitat specialization (fresh water), but most of the lineages lost to extinction were not habitat specialists.

Broadly compared across phyla, the fossil record shows that species high on the food pyramid (predators or parasites) have sometimes been at risk during environmental perturbations, particularly if they also had achieved large body size and had developed specialized morphology and high levels of activity (e.g., for predation). It is thought that their high metabolic requirements and need for large amounts of food imperils their survival during difficult periods compared to smaller species that usually are lower in the food web. A related concept, known from the 1800s as Cope's Rule, posits that most lineages originate as small generalized forms, evolve toward larger and more specialized morphologies, then become extinct (Stanley, 1973). Body sizes of extinct crustaceans are not synthesized in the literature in most cases, but the few available cast doubt on whether or not crustacean lineages of large body size
are particularly susceptible to extinction. Enantiopodan remipeds probably were relatively small, kazachthrans were small to moderately sized (<30 mm), and palaeocaridaceans and phygocephalomorphs probably were small to moderately sized for crustaceans. Yet all became extinct. In many crustaceans (Reaka and Manning, 1981, 1987; Reaka, 1980) and among invertebrates in general (Menge 1973; Strathmann and Strathmann, 1982), larger species produce many more small offspring that spend considerably longer periods in the plankton and have wider geographic ranges than smaller species. Larval dispersal and wide geographic ranges tend to slow down both extinction and speciation (Jablonski 1986; Jablonski and Lutz, 1983), suggesting that, in marine invertebrates, dispersal at large body sizes usually overpowers susceptibility to extinction. However, not all large-sized species produce large numbers of widely dispersing offspring (especially in brackish, freshwater, and terrestrial environments or in lineages where abbreviated development is inherited from past legacies), and available food may be sufficiently restricted to prevent reproduction (thus dispersal) during particularly stressful environmental conditions. Thus, although Cope’s Rule may apply more universally among fossil vertebrates (which were Cope’s original source of data for the pattern of greater extinction in species of large body size) than among invertebrates, large body size still may render lineages susceptible to extinction among invertebrates in some situations.

2. The Prospects of Extinction in Modern Crustaceans: Which Taxa Are Most Vulnerable, and Which Ones Represent the Most Urgent Conservation Targets?

Table II summarizes the major factors that are likely to affect the viability of contemporary crustacean groups during natural or human-induced environmental changes. Many scientists caution that drastic environmental shifts could occur as human populations increase exponentially (perhaps doubling in the next 100 years) and affect the atmosphere, soil, and coastal waters (Wilson, 1989). If such change impacts the biotic world in a manner similar to the mass extinctions of the past, the primary factor that would facilitate survival of the crustacean groups analyzed in this study would be wide distribution of entire groups. Fortunately, most (35 of 39) groups of crustaceans are distributed nearly worldwide, suggesting that at least some of these groups probably would survive. The remipedes, anaspidaceans, thermosbaenaceans, and spelaegraphaceans, however, would be at high risk.

If the deteriorating environment resembled background conditions rather than a mass extinction event, however, the number of lineages within the group, the size of the geographic ranges of species, whether or not most of the group inhabited tropical (especially reef) environments, and possibly whether the species were characterized by low population densities, occupied specialized or highly restricted types of habitat, were positioned high in the food web of their communities, or were large and active could influence the probability that they would succumb to extinction over the long run.

Of the 34 groups for which data were available on species numbers, 25 groups (74%) were considered to be at risk as a result of relatively few species within the lineage (14 at high risk, 11 at moderate risk). The remipeds, cephalocarids, nautostracans, mystacocarids, leptostracans, tannomocarids, anaspidaceans, thermosbaenaceans, and spelaegraphaceans all have fewer than 15 species and would be at very high risk. Even more notable, the miocarid and procarid lineages are represented by only 3 species each, and only 1 species is known in the entire Order Amphionidae. These phylogenetically unique groups could easily be lost.

Of 39 groups, 27 (70%) are composed of species with mostly restricted geographic ranges. Ten of these groups contain species with especially restricted distributions, considered at high risk, and 17 contain some species with restricted ranges that are judged to be at moderate risk. Of the 39 groups, 14 (36%) are considered to be at risk because they are primarily tropical or reef dwelling. Twenty-seven (74%) of the 39 groups are characterized by small population sizes. Eight of these have especially low densities or total populations and are considered to be at high risk, and 19 appear to be at moderate risk. Of the 39 groups, 29 (74%) occupy unique or highly specialized types of habitats; 11 and 18 of these are judged to be at high or moderate risk, respectively, because of their degree of habitat specialization. Although many crustaceans are predators or parasites, relatively few are at the apex of the food web (this position often being usurped by fishes), and relatively few are species specific in their choice of prey or hosts. Of the 39 groups, 23 (59%) are judged to be at moderate risk due to a relatively high position in the food chain. Also, relatively few entire groups of crustaceans are sufficiently large in body size that their persistence is likely to be imperiled. Large body size is unlikely to be associated with extensive extinction for the following reasons. Wherever the distributions of
<table>
<thead>
<tr>
<th>Taxon</th>
<th>Few living species (+&lt;c100, +&lt;500)</th>
<th>Entire group restricted geographic range (&lt;half circumference Earth)</th>
<th>Most species restricted geographic range</th>
<th>Group primarily tropical</th>
<th>Small population</th>
<th>Unique or restricted habitat type</th>
<th>High trophic level</th>
<th>Large body size (or long lived)</th>
<th>Old (=probably since mid-Paleozoic, + = probably since beginning Paleozoic) or phylogenetically unique lineage</th>
<th>Total score (sums of +, ±, − counted as 0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benthicola</td>
<td>++(10)</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>15</td>
</tr>
<tr>
<td>Cephalocarisca</td>
<td>++(10)</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>7</td>
</tr>
<tr>
<td>Amonita</td>
<td>+(?c200)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>4</td>
</tr>
<tr>
<td>Notocarisca</td>
<td>++(11)</td>
<td>+?</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>4.5</td>
</tr>
<tr>
<td>Cladocera</td>
<td>+(?c100)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>3.5</td>
</tr>
<tr>
<td>Ostracoda</td>
<td>(about 8000)</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>6.5</td>
</tr>
<tr>
<td>Myslocarida</td>
<td>+(?c100)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>5</td>
</tr>
<tr>
<td>Copepoda</td>
<td>(90000)</td>
<td>?</td>
<td>?</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>3.5</td>
</tr>
<tr>
<td>Branchiura</td>
<td>+(100)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>5</td>
</tr>
<tr>
<td>Tanaidocarida</td>
<td>++(5)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>5</td>
</tr>
<tr>
<td>Anomocarisca</td>
<td>+(?c200)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>4</td>
</tr>
<tr>
<td>Acanthocarisca</td>
<td>+(?c100)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>3.5</td>
</tr>
<tr>
<td>Thermocarisca</td>
<td>(a2000)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>5</td>
</tr>
<tr>
<td>Bithrocariosca</td>
<td>+(?c200)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>5</td>
</tr>
<tr>
<td>Leuconocarisca</td>
<td>++(?c100)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>5</td>
</tr>
<tr>
<td>Euphausiacea</td>
<td>+(?c100)</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>7.5</td>
</tr>
<tr>
<td>Thysanocarisca</td>
<td>(+100)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>4</td>
</tr>
<tr>
<td>Euphausiasca</td>
<td>(+40)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>3.5</td>
</tr>
<tr>
<td>Theriopodacea</td>
<td>(+11)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>9</td>
</tr>
<tr>
<td>Tanatidacea</td>
<td>(850)</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>2.5</td>
</tr>
<tr>
<td>Spelaeocarisca</td>
<td>++(?c200)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>10</td>
</tr>
<tr>
<td>Cymaciidae</td>
<td>(+c1000)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>3</td>
</tr>
<tr>
<td>Micrura</td>
<td>++(?c100)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>9</td>
</tr>
<tr>
<td>Isopoda</td>
<td>(+c1000)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>4</td>
</tr>
<tr>
<td>Amphipoda</td>
<td>(+c6000)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>2</td>
</tr>
<tr>
<td>Euphausiacea</td>
<td>++(90)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>3.5</td>
</tr>
<tr>
<td>Amphipoda</td>
<td>++(?c100)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>6</td>
</tr>
<tr>
<td>Dendrobranchiata</td>
<td>(+450)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>4</td>
</tr>
<tr>
<td>Cetacea</td>
<td>(+20000)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>5.5</td>
</tr>
<tr>
<td>Procyclopes</td>
<td>++(5)</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>4</td>
</tr>
<tr>
<td>Serranocerus</td>
<td>++(25)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>9</td>
</tr>
<tr>
<td>Thysanoocerus</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>10</td>
</tr>
<tr>
<td>Antocerus</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>5</td>
</tr>
<tr>
<td>Platytera</td>
<td>±(some)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>5.5</td>
</tr>
<tr>
<td>Anomura</td>
<td>±(some)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>2.5</td>
</tr>
<tr>
<td>Brachyura</td>
<td>±(some)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>±(some)</td>
<td>+</td>
<td>±(some)</td>
<td>±(some)</td>
<td>+ (fresh water)</td>
<td>6</td>
</tr>
</tbody>
</table>
body sizes among species of a group have been compared, there always are far more species of small than large body size (May 1988; Reaka-Kudla, 1991, 1997), and this almost certainly is true throughout the Crustacea. Also, in many crustacean groups, species of large body size produce long-lived pelagic larvae that confer protection against background extinction. The present analysis suggests that 12 of 39 groups (31%) are at possible risk due to their relatively large body sizes (if other life history characteristics do not outweigh the effect of body size).

One of the factors that should be included in prioritizing conservation efforts is the uniqueness of the lineage. Some crustacean groups represent a heritage that stems from half a billion years ago (the beginning of the Paleozoic). Among groups still alive today, the remipeds, ostracods, ascothoracian barnacles, and thoracican barnacles are known from Cambrian fossils, or their morphology suggests that they are derived from the most primitive lineages of crustaceans. Sixteen other groups are known from the Mid-Paleozoic (300–400 million years ago). These ancient groups are more likely than more recently derived groups to have totally novel characteristics, even in their biochemical and genetic constitution. As we enter a millennium in which biotechnology has only begun to offer human benefits through genetic manipulations, we can ill afford to lose these lineages. Some of these ancient groups are particularly vulnerable due to few constituent species, restricted geographic ranges, small populations, specialized habitat requirements, or other factors (especially the remipeds, cephalocarids, notostracans, ascothoracian and acrothoracian barnacles, leptostracans, anaspidaceans, thermosbaenaceans, speleogriphaceans, micraeans, procaridids, and especially the amphipodaceans). These groups merit special conservation attention.

Table II provides a tentative overall summation score for risk factors and intrinsic conservation value in each group. The overall mean risk/value score for the 39 groups was 5.9. Incorporating all of the different risk and value factors, eight groups (the remipeds, tantulocarids, anaspidaceans, speleogriphaceaens, micraeans, procaridids, stenopodideans, and thalassinideans) merit the closest conservation scrutiny overall (score of 9 or above). Of these, the remipeds are considered to be the highest priority. This represents only a beginning, however, at assessing which lineages are in most need of protection in order to ensure their survival in the next centuries. Identifying which lineages and where they are most at risk (as well as the most effective methods for conserving them) will be one of the most important tasks for at least the next decade.

See Also the Following Articles

AQUACULTURE • ENDANGERED MARINE INVERTEBRATES • INVERTEBRATES, MARINE, OVERVIEW • MARINE ECOSYSTEMS • MARINE ECOSYSTEMS, HUMAN IMPACT ON • MOLLUSCS

Bibliography


DARWIN, CHARLES (DARWINISM)

Michael T. Ghiselin
California Academy of Sciences

I. Introduction
II. Darwin's Life and Work
III. Darwin's Contributions to the Study of Biodiversity
IV. Darwinism after Darwin

GLOSSARY

biota Fauna and flora; the organisms in a given place.
coevolution Evolution where two species have a mutual influence on each other.
generic drift Change in gene frequencies due to random fluctuations in a finite population.
heterostyly Having two or more kinds of flowers with stamens and pistils of different lengths.
mimicry Defensive mechanism in which organisms of two species resemble each other, ordinarily because at least one of them is distasteful.
morphology The study of form.
phylogeny The history of a species or lineage.
plesiotropy Phenomenon of a gene affecting more than one trait.
sexual selection Kind of selection that Darwin distinguished from natural and artificial selection in which reproductive success depends on monopolizing mates.
taxon Group of organisms in a formal system of classification.

CHARLES DARWIN (1809–1882) DISCOVERED the principle of natural selection and laid the foundations for modern evolutionary biology. The term "Darwinian" is applied to his theory, and others like it, in which natural selection is considered the main, although not the only, mechanism. Terms such as "Lamarckian" suggest alternatives in which other mechanisms, such as use and disuse, are considered the main cause of change. "Darwinism" has come to mean both Darwin's version of evolutionary theory and those elaborated by his followers. Modified versions of Darwinism that emphasize selection are often called "neo-Darwinism." Among these are the neo-Darwinism of August Weismann and the Synthetic Theory that took shape in the middle of the twentieth century and still remains dominant. Darwinians have always acknowledged the existence of mechanisms other than natural selection. Darwin himself invoked Lamarckian and quasi-Lamarckian mechanisms that are no longer accepted by his followers. He also accepted auxiliary mechanisms that are still considered valid. These include sexual selection, plesiotropy, and developmental constraints. Evolution as a result of sampling error (generic drift) is "non-Darwinian" in the sense that Darwin did not know...
about it. The “neutral theory” that invokes sampling error as a case of essentially non-adaptive change at the molecular level is best considered part of mainstream neo-Darwinism.

I. INTRODUCTION

Although natural selection theory is generally considered Darwin’s most important and original contribution to knowledge, he also deserves credit for providing compelling evidence that evolution is a fact and for making its fundamental implications clear. Indeed, Darwin began an intellectual revolution of the first magnitude, one that has affected the whole scope of modern life and culture. The emphasis here will be on those aspects of Darwin’s contribution and of Darwinism in general that are particularly relevant to the study of biodiversity. Within biology it has been the sciences of systematics and ecology that have been most revolutionized by Darwinian thinking, because it enables us to understand both how and why diversity originates and is maintained. Some of the larger philosophical and social issues will also be considered, especially as they relate to the central theme. These include the rejection of teleology, the replacement of typological thinking with population thinking, the recognition of the importance of historical contingency, and a competitive view of the natural economy.

II. DARWIN’S LIFE AND WORK

A. Childhood and Education

Charles Darwin was born on February 12, 1809, in Shrewsbury, England. His father was a wealthy physician and his mother was a member of the Wedgwood family, famous for the manufacture of pottery. She died when he was eight years old. Young Darwin was not very successful in school and at the age of sixteen he was sent to study medicine at Edinburgh University. There he was able to cultivate his interest in natural history, largely on an extracurricular basis. A decision to abandon medicine in favor of preparation for a career as a clergyman led him to enroll in Cambridge University on October 15, 1827. During the years at Cambridge he continued his largely extracurricular studies of natural history and received much encouragement from both faculty and students. In those days there was nothing unusual about a clergyman also being a scientist, a university professor, or both. He passed the examinations for his B.A. degree on January 22, 1831, but remained there to complete his residency requirements. He took up the study of geology, and the following summer obtained some field experience in the company of Professor Adam Sedgwick (1785–1873).

B. The Beagle Voyage

On September 1, 1831, Darwin accepted an invitation to join a surveying expedition on H.M.S. Beagle, as an unofficial naturalist and gentlemanly companion for the captain of the ship, Robert FitzRoy (1805–1865). The Beagle left for South America on December 27, 1831, and after sailing around the world finally got back to England on October 2, 1836. During the voyage, Darwin was able to observe and collect on a vast scale, especially in South America, on the coast of which most of the survey work was done. He also visited Australia, the Cape of Good Hope, and many islands along the way. At first his work emphasized the marine invertebrates that had interested him since his days at Edinburgh. But he took increasing interest in geology. He read the first volume of the Principles of Geology by Charles Lyell (1797–1875) not long after the ship left port, and began to read the second volume the following November. Darwin became an enthusiastic supporter of Lyell’s uniformitarian methodology, and was also influenced by the discussions of biogeography and Lamarckism in the Principles. Darwin developed a theory of coral reef morphology while still working in South America. Later in the voyage he began to test it by means of observations on reefs. It was the coral reef theory that first established his high reputation as a scientist.
that the inhabitants had diversified more than he had assumed was possible, and furthermore that they had done so locally. They were related by community of descent. This was just the beginning of many of his biogeographical insights. Facts such as the lack of frogs and other amphibians on islands far distant from land could easily be explained in terms of evolution, but otherwise remained a mystery. Contrary to the impression that one sometimes gets from text books, the most compelling evidence for evolution comes from biogeography, not the study of fossils.

C. Natural Selection

The fossils that Darwin collected in South America turned out to be recently extinct members of groups of animals that were still living in the same area. Again this suggested the possibility of local diversification. Around March, 1837, he began to search for an evolutionary mechanism. After much reading and theorizing, he discovered natural selection. This happened toward the end of September, 1838, when he read the Essay on the Principle of Population by Thomas Robert Malthus (1766–1834), who was among other things the world’s first professor of economics. Darwin now understood how reproductive competition between organisms of the same species could produce evolutionary change.

The influence of Malthus on Darwin requires some elaboration here because it is so important for biodiversity studies. Malthus’s ideas (or what have been represented as his ideas) have been controversial and have often entered into discussions on population planning and conservation policy. His basic point was that the growth of population tends to outstrip the growth of the food supply. His notion that the growth of population is exponential whereas that of the food supply is linear is well known, although it never was more than just a crude approximation. Malthus based his arguments on empirical evidence. After plagues and other catastrophes have reduced the numbers of human population there are fewer mouths to feed and the amount of food available per capita increases. Consequently people can and do support larger families and population grows rapidly. As time passes, the amount of food per capita decreases and population growth levels off, as people marry later and have smaller families. The standard of living tends to drop as well. A comparable situation could be observed where new land had been brought into cultivation, as in the United States, following which a high living standard was accompanied by rapid population growth.

Malthus saw a problem in this scenario with respect to the future economic prosperity of society: population growth would tend to depress the standard of living. There was a serious question in his mind, as well as those of his critics, whether his model of population growth need have those particular consequences. “Prudential constraints” (as he called sexual abstinence) would have some mitigating effect. The growth of technology has of course at least delayed the sort of famine that might be envisaged. On the other hand, the Malthusian model applies quite well to non-human populations. It contributed to Darwin’s understanding of population dynamics, which he incorporated into his evolutionary and ecological theory. Darwin recognized that there was reproductive competition between individuals of the same species, and because these individuals varied, those that were able to utilize resources more effectively would have more offspring. Natural selection follows when it is realized that the properties of those individuals that have more offspring are differentially transmitted by inheritance to the next generation. Extrapolating further, it made sense that natural selection would operate differently depending on whether or not population levels had been lowered by the effects of weather or other density-independent factors.

Given this kind of population dynamics, it became clear to Darwin that variation was far more important than had previously been understood. Natural selection would not work without it. This shift in emphasis is often discussed under the rubric of “population thinking.” In contrast, the older “typological” approach, which conceived of groups of organisms in terms of stereotypes, viewed the differences between individuals as a kind of departure from a norm, and these were screened out as unimportant. To the extent that variability had been recognized, it was assumed that something held it within definite limits. It was believed that species might vary, but not so far as to give rise to new ones. At a deeper philosophical level, treating species as reproductive populations implied that they were something held it within definite limits. It was believed that species might vary, but not so far as to give rise to new ones. At a deeper philosophical level, treating species as reproductive populations implied that they were a kind of departure from a norm, and these were screened out as unimportant. To the extent that variability had been recognized, it was assumed that something held it within definite limits. It was believed that species might vary, but not so far as to give rise to new ones. At a deeper philosophical level, treating species as reproductive populations implied that they were different from the “natural kinds” of chemistry and other physical sciences. The new Darwinian thinking accepted that species were concrete, that they had a history, and that they could evolve. Treating species as something more than abstractions was a major conceptual innovation.

D. The Long Delay

Following this shift in his thinking, Darwin immediately began elaborating a comprehensive theory of evolution together with its many implications. However, he was busy with other things, and publication of his
theory was long delayed. One factor in the delay was the decline of his health, which, if not brought on by stress, was at least exacerbated by it. Ill health was a major consideration in his decision to leave London and live in the country. Darwin had married his cousin Emma Wedgwood on January 29, 1839, and the first of their many children was born on December 27th of that year. He moved to the village of Down, not far south of London, and resided there from September, 1842, until his death on April 19, 1882.

He had already published a semi-popular account of his travels in 1839. Usually called The Voyage of the Beagle, it was a great literary success. He also edited the scientific results of the voyage and published three books on the geology of the voyage, including one on coral reefs. He began to publish a few papers on zoology. One of these grew into a vast monograph on barnacles (Cirripedia). It was well over a thousand pages long, and consumed eight years on his working time.

E. The Origin of Species

At last, in 1854, the work on barnacles was completed and Darwin began to write a long book on evolution. It was to incorporate much original research as well as a detailed review of the literature. As it happened, Alfred Russel Wallace (1832–1913) was interested in the possibility of evolution too, and spent many years in the field collecting animals and plants and studying their geographical distribution, first in South America (1848–1852) and then in the region that is now mostly part of Indonesia (1853–1855). They entered into correspondence, and in a letter dated May 1, 1857, Darwin informed Wallace that he had an evolutionary theory but did not let on what it was. Wallace may have gained informed Wallace that he had an evolutionary theory and Darwin began to write a long book on evolution. This "abstract" became What Darwin's work on sexual selection and the evolution of social behavior as treated in The Descent of Man is

As a supporter, Darwin already had his friend and confidant the botanist Joseph Dalton Hooker (1817–1911), and others soon joined them. There were of course public debates, including the one at Oxford on June 9, 1860. Although the exchange there between Bishop Samuel Wilberforce (1805–1873) and Thomas Henry Huxley (1825–1895) is better known, Hooker was really more effective in support of Darwin's ideas. Darwin himself furthered his interests through behind the scenes negotiation and, more importantly, through research and publication. There were a total of six editions of the Origin of Species, the last of which was published in 1872.

F. Later Publications

Plans to complete and publish the big were abandoned, but part of it, somewhat revised, appeared in 1868 as The Variation of Animals and Plants under Domestication. Concurrently Darwin had an active empirical research program under way, much of which was conducted on plants. His book of 1862, entitled The Various Contrivances by which British and Foreign Orchids Are Fertilised by Insects, and on the Good Effects of Crossbreeding, dealt with the problems of adaptation. It was a seminal document of pollination ecology and the study of coevolution. Later works on floral biology, The Effects of Cross and Self-Fertilisation in the Vegetable Kingdom (1876) and The Different Forms of Flowers on Plants of the Same Species (1877) addressed mainly the effects of inbreeding and outbreeding.

Darwin was a first-rate plant physiologist, and his book The Movement and Habits of Climbing Plants (1875) are early treatises in what later came to be called ethology. They deal with behavior from a comparative and evolutionary point of view. Treating humans, other animals, and even plants as part of an evolutionary continuum was a major contribution to comparative psychology, as well as a serious challenge to traditional philosophical ideas such as mind–body dualism. There was also a lot of behavior in the Descent of Man, and Selection in Relation to Sex (1871) and its sequel, The Expression of the Emotions in Man and Animals (1872), and in the last of his books, The Formation of Vegetable Mould through the Action of Worms, with Observations on Their Habits (1881). Darwin's, work on sexual selection and the evolution of social behavior as treated in The Descent of Man is
strikingly modern in its approach. In retrospect, his efforts to solve the problems of heredity seem much less successful, though they stimulated others to work on them. It is unfortunate that his theory of pangenesis was appended to the 1868 book *Variation*, with the result that it is rarely read because that theory was superseded. Although pangenesis explained the inheritance of acquired characters, it was more a theory of development than of heredity. In fact the book presents a remarkably sophisticated treatment of the relationships between embryology and evolution.

### III. DARWIN'S CONTRIBUTIONS TO THE STUDY OF BIODIVERSITY

#### A. The Competitive Natural Economy

Darwin succeeded, where others had failed, in convincing the scientific community that evolution had in fact occurred, and that it was responsible for the patterns of diversity that may be observed in both fossil and living organisms. Biogeography, paleontology, and classification took on new significance, becoming profoundly historical in nature. He was less successful at gaining immediate support for his more fundamental (or at least original) contribution, which was the basic mechanism for evolution and the new way of thinking that went along with it. Evolution by natural selection made it possible to reject all sorts of older ideas about the world, especially the notion that organisms and the “natural economy” had somehow been designed, or produced by a supernatural agency. Darwin had demonstrated that there was much less “purpose” in the world than had been supposed, that there was nothing inherently “good” about the way things are, and that there was no necessary progression from lower to higher levels of organization.

Darwin’s theory was immediately recognized as having profound philosophical significance. From time immemorial it had seemed natural to interpret the living world and the universe in general as if they were the product of something like human intelligence. Such notions, and the anthropomorphic way of thinking that underlies them, are called “teleology.” The idea that somehow the world had been designed or was otherwise the product of supernatural action was of course fundamental to much religious thinking, even among those who had come to accept early versions of evolution. The remarkable adaptations of living organisms were taken as evidence for the existence of a supernatural order. Although many of the features of inanimate nature could be explained in terms of laws of nature and matter in motion, a more subtle purposefulness was widely attributed to the universe as a whole. Indeed, the great German philosopher Immanuel Kant (1724–1804) had proclaimed that there never could be a “Newton of the grass-blades” able to explain adaptation in strictly naturalistic terms. But that is precisely what Darwin accomplished.

Darwin also provided a way of thinking about adaptation that allows us to avoid the naïve anthropomorphic thought processes that lead to the illusion of design. Rather than ask what things are for, we ask how they might have evolved, and explain the facts in terms of what has happened and why. When this is done properly, it becomes clear that much ascription of adaptive significance has been mere guesswork and that there is far less “optimality” in the world than is sometimes supposed even today. On the other hand, research has repeatedly revealed that supposedly “non-adaptive” features turn out to be very important in the lives of organisms. Defensive metabolites are an excellent example.

Opponents of Darwinism of course include many persons who have been reluctant to abandon teleological interpretations of nature. It has not always been a matter of denying evolution or natural selection outright. Often evolution has been envisaged as a means of achieving some supernaturally ordained plan. And often the teleology is merely assumed or presupposed rather than being advocated explicitly. “Holistic” notions that treat the world as if it were a super-organism have a long history in myth and pseudoscience that goes back to antiquity. They retain their popularity among religious persons and advocates of occult metaphysics. Because such older ways of thinking are still widespread, it is important to avoid giving the appearance of endorsing them. Functional explanations should be called “teleonomic” (i.e., where the apparent purposefulness in organisms is derived from evolutionary adaptation) rather than “teleological,” and metaphors such as “design” should be avoided.

The idea of progress has rarely been disassociated from teleological thinking. For that reason, among others, Darwin’s ideas about progress have often been misinterpreted and but rarely understood. They are often confused with quite different ideas, especially those of philosophers, social scientists, and non-Darwinian evolutionists. In pre-evolutionary biology it was widely accepted that organisms could be placed in a single series from lower to higher, connecting to non-living matter at the bottom and to the various ranks of supernatural beings at the top. This great chain of being, or...
distinct gene pools. Furthermore, he did not fully ap-

preciate the difference between genotypic and pheno-
typic variability. Because his mechanism depended on
differences between individual organisms, he needed
to show that the necessary variability does in fact exist.
He was able to show that species are indeed far more
variable than his predecessors had believed. He also
discovered that sexual reproduction and outcrossing
are much more prevalent than even he had at first
suspected. Therefore, his contributions to the study of
population structure in nature were the foundation for
a great deal of work along such lines.

Variation within and between the component popu-
lations of a species is one important form of biodiversity
within species. Another is the sort of differences that
may be observed between the sexes, and here again
Darwin’s contribution was fundamental. His theory of
sexual selection helped to account for sexual dimor-
phism on the basis of males competing for the females,
either through fighting with one another or through
efforts to attract the females. He also discovered hetero-
ously in plants. This is a kind of polymorphism in which
hermaphroditic flowers of the same species have differ-
ent morphologies, furthering outcrossing and reduc-
ing inbreeding.

Darwin was also able to explain the global pattern of
diversity between species. Of course the splitting up of
species and their ability to undergo an indefinite
amount of change allows diversification through geo-
logical time. However, that by itself does not tell us
why or how the diversification takes place. To account
for that Darwin invoked what he called a “principle of
divergence of character.” According to this principle,
diversification allows the if a largely harmo-
nious and cooperative natural economy with that of a
fundamentally competitive one. The organisms within
populations reproduce differentially as a consequence of
how they utilize the resources in their environments, and
change occurs in the direction of more effective use of those resources. That accounts for adaptation
and for such progressive and regressive changes as may
in fact occur. But it greatly limits the kind of adaptation
that can evolve, because it is organisms, and to some
extent families, that compete as reproductive units. Dar-
win was profoundly aware of the point that his mecha-
nism would not produce adaptations at the species level
or at the level of the community or ecosystem. This
point, however, was often neglected by his contempo-
rarities and received due emphasis only in the latter half
of the twentieth century.

Darwin’s idea of what a species is lacked a clear
conception of reproductively isolated populations with
distinct gene pools. Furthermore, he did not fully ap-

Lyell and others had long denied that. Natural selection,
generous in plants. This is a kind of polymorphism in which
either through fighting with one another or through
Darwin's idea of what a species is lacked a clear
conception of reproductively isolated populations with
distinct gene pools. Furthermore, he did not fully ap-
that there is adaptive radiation within such groups as families, orders, and classes. The lack of intermediate forms among extant taxa is the consequence of those forms that were close to the ancestral ones having been driven to extinction. On a global scale, whole biotas would tend gradually to diversify through time. Darwin specified the conditions under which such diversification would lead to the evolution of competitively effective organisms. These included a large area and a diverse topography, coupled with a long period of competitive interaction among the inhabitants. He realized that the biotas of separate biotic provinces have not undergone the same amount or kind of modification, and that similar considerations apply to insular biotas. His insights were thus fundamental to our understanding of invasions, faunal interchanges, and many aspects of extinction.

Darwin contributed a great deal toward our understanding of the interdependencies among species by emphasizing how complicated such ecological relationships may be. He was particularly interested in pollination symbiosis, and his work in that area initiated the study of coevolution between animals and plants. Although his research was focused on showing the importance of outcrossing, it also revealed the ecological significance of pollination.

B. Classification

Systematics is often defined as the study of biodiversity, and Darwin made fundamental contributions to both the theory and the content of that science. Classification is part of every pattern of language and thinking, and is necessary for the organization of knowledge, whether it be the knowledge of the lay-person or the professional scientist. The objects of our experience can be referred to by more or less general terms, such as "food," "pasta," and "spaghetti." Because what is true of the more general groups is true of all the groups that fall under them, we have a powerful instrument for thinking about more than just single objects. If pasta is made of grain, and all spaghetti is pasta, then spaghetti is made of grain.

Scientific classification may be much more sophisticated than that of the lay person. Scientists attempt to discover groups of things that share properties that are of theoretical importance. A good example of a scientific classification system is the periodic table of the elements, in which chlorine and bromine are both halogens and undergo similar chemical reactions. The groups in question may be worked out on an empirical basis, and only later understood in theoretical terms. This was approximately the case with Darwin's work on biological systematics. Human beings could be assigned to more general groups, such as Mammalia, Vertebrata, and Animalia, but it was not obvious why. Sometimes it was attributed to God having some plan in mind when He created organisms, and sometimes unknown laws of nature were thought to be responsible.

Darwin's solution to this puzzle was straightforward and evolutionary. According to his biodiversity model, species have repeatedly undergone a kind of splitting, giving rise to separate lineages that evolve more and more differences as time since common ancestry becomes greater. The more general the group, the earlier the common ancestry. Once the theoretical reason for traditional classification was understood, a new approach could be developed in which the groups were explicitly recognized as lineages sharing a common ancestor. As Darwin put it: "Our classifications will come to be, so far as they can be so made, genealogies."

Putting that prognostication into practice was not easy and, generally speaking, classifications have been only rough approximations to genealogies. They also generally attempt to express the amount of difference that separates the various groups, and sometimes the criteria conflict. Be this as it may, Darwin was writing from experience. On October 1, 1846, he began to study the anatomy of barnacles in order to write a short paper on an aberrant species that he had discovered.

The project expanded to the point that he spent eight years revising and monographing the entire subclass, both living and fossil, and treating all aspects of their biology in great detail. Although he did not explain what he was doing at the time, it is clear that his classification is based on his understanding and thinking, and is necessary for the organization of knowledge, whether it be the knowledge of the lay-person or the professional scientist.

Darwin's solution to this puzzle was straightforward and evolutionary. According to his biodiversity model, species have repeatedly undergone a kind of splitting, giving rise to separate lineages that evolve more and more differences as time since common ancestry becomes greater. The more general the group, the earlier the common ancestry. Once the theoretical reason for traditional classification was understood, a new approach could be developed in which the groups were explicitly recognized as lineages sharing a common ancestor. As Darwin put it: "Our classifications will come to be, so far as they can be so made, genealogies."

Putting that prognostication into practice was not easy and, generally speaking, classifications have been only rough approximations to genealogies. They also generally attempt to express the amount of difference that separates the various groups, and sometimes the criteria conflict. Be this as it may, Darwin was writing from experience. On October 1, 1846, he began to study the anatomy of barnacles in order to write a short paper on an aberrant species that he had discovered. The project expanded to the point that he spent eight years revising and monographing the entire subclass, both living and fossil, and treating all aspects of their biology in great detail. Although he did not explain what he was doing at the time, it is clear that his classification is based on his understanding and thinking, and is necessary for the organization of knowledge, whether it be the knowledge of the lay-person or the professional scientist.

Darwin's solution to this puzzle was straightforward and evolutionary. According to his biodiversity model, species have repeatedly undergone a kind of splitting, giving rise to separate lineages that evolve more and more differences as time since common ancestry becomes greater. The more general the group, the earlier the common ancestry. Once the theoretical reason for traditional classification was understood, a new approach could be developed in which the groups were explicitly recognized as lineages sharing a common ancestor. As Darwin put it: "Our classifications will come to be, so far as they can be so made, genealogies."

Putting that prognostication into practice was not easy and, generally speaking, classifications have been only rough approximations to genealogies. They also generally attempt to express the amount of difference that separates the various groups, and sometimes the criteria conflict. Be this as it may, Darwin was writing from experience. On October 1, 1846, he began to study the anatomy of barnacles in order to write a short paper on an aberrant species that he had discovered. The project expanded to the point that he spent eight years revising and monographing the entire subclass, both living and fossil, and treating all aspects of their biology in great detail. Although he did not explain what he was doing at the time, it is clear that his classification is based on his understanding and thinking, and is necessary for the organization of knowledge, whether it be the knowledge of the lay-person or the professional scientist.

Darwin's solution to this puzzle was straightforward and evolutionary. According to his biodiversity model, species have repeatedly undergone a kind of splitting, giving rise to separate lineages that evolve more and more differences as time since common ancestry becomes greater. The more general the group, the earlier the common ancestry. Once the theoretical reason for traditional classification was understood, a new approach could be developed in which the groups were explicitly recognized as lineages sharing a common ancestor. As Darwin put it: "Our classifications will come to be, so far as they can be so made, genealogies."

Putting that prognostication into practice was not easy and, generally speaking, classifications have been only rough approximations to genealogies. They also generally attempt to express the amount of difference that separates the various groups, and sometimes the criteria conflict. Be this as it may, Darwin was writing from experience. On October 1, 1846, he began to study the anatomy of barnacles in order to write a short paper on an aberrant species that he had discovered. The project expanded to the point that he spent eight years revising and monographing the entire subclass, both living and fossil, and treating all aspects of their biology in great detail. Although he did not explain what he was doing at the time, it is clear that his classification is based on his understanding and thinking, and is necessary for the organization of knowledge, whether it be the knowledge of the lay-person or the professional scientist.
he treated as mere varieties are now classified as species. In many other respects, however, his approach was very advanced for his day. He worked with extensive series of specimens and went out of his way to document all kinds of variation.

IV. DARWINISM AFTER DARWIN

A. Alternatives to Darwinism

Darwin maintained that natural selection is the "main, but not exclusive" evolutionary mechanism. This seems about as good a criterion as any for distinguishing "Darwinians" from a wider range of scientists who accepted evolution but either rejected natural selection or treated it as an unimportant relative to other causes. Among these scientists, the followers of Jean-Baptiste Lamarck (1744–1829), who emphasized use and disuse, were particularly numerous. However, there were other alternatives. Étienne Geoffroy Saint-Hilaire (1772–1844), whose views are often confused with those of Lamarck, believed that the environment directly induces changes in structure. Indeed, Darwin was strongly influenced by Geoffroy, and accepted the direct action of the environment as a minor cause of change. There were also various versions of "orthogenesis" or evolution that is directed in a particular direction. Some versions of orthogenesis were based on the notion that evolution is like embryological development, and change has been built in from the start. In other versions, laws of nature, similar to those that determine the structure of minerals, were invoked. Sometimes the appeal was to unknown causes, perhaps of a supernatural character. Many scientists were, or at least claimed to be, agnostic with respect to what causes evolution.

B. Early Darwinians

Given such a range of alternatives and the small amount of research that had been done, it makes sense that from the outset there were few Darwinians other than Darwin himself. Because he was largely responsible for getting evolutionary thinking in general accepted by the intellectual community, a lot of evolutionists who did not accept natural selection nonetheless considered themselves his followers. The degree to which Darwinism, in the sense that we use that term here, was a minority position has sometimes been exaggerated. We can identify quite a number of important contemporaries of Darwin who established successful research programs based on the study of natural selection. Foremost among these of course was Alfred Russel Wallace (1832–1913), its co-discoverer. Also very distinguished was Wallace's traveling companion, Henry Walter Bates (1825–1892), the discoverer of Batesian mimicry. Second only to Darwin in his mastery of the theory was Fritz Müller (1822–1897). He is best remembered for his discovery of Müllerian mimicry, but he also was the first to propound the idea that developmental stages may recapitulate evolutionary ones. Both Fritz Müller and his brother Hermann (1829–1883) conducted magisterial research on pollination symbiosis under Darwin's influence. It is worth emphasizing that these scientists were outstanding for their performance as naturalists in the field. The kind of research that they did has been fundamental to our understanding of biodiversity because it documents how natural selection takes place in real environments.

Darwin also had important followers whose work was more focused in the museum and the laboratory. He had a close circle of followers in the botanist Joseph Dalton Hooker (1817–1911) and the zoologists George John Romanes (1848–1894) and John Lubbock (1834–1913). Many scientists were, or at least claimed to be, agnostic with respect to what causes evolution.

C. Genetics and the Modern Synthesis

Beginning in 1900, Mendelism, developed from the revolutionary research of Gregor Johann Mendel (1822–1884), began to supply the genetics support that Darwinism needed. However, it did not have the immediate effect of gaining adherents to Darwin's theory. Instead it gave rise to alternatives such as mutationism. Advocates of mutationism were over impressed by conspicuous mutants and erroneously concluded that their sudden appearance in a single generation could be extrapolated to the origin of new species and even higher taxa. However, the study of genetics soon moved away from mutationism and gradually undermined some of the other alternatives, especially Lamarckism and orthogenesis.

The reconciliation of Darwinism with genetics depended in part on the elimination of such alternatives, in part on the growth of theoretical population genetics,
and in part on the study of natural populations in the field. Studies of natural populations by systematists like Ernst Mayr (b. 1904) and geneticists like Theodosius Dobzhansky (1900–1975) were crucial to the “evolutionary synthesis” that began in the 1930s and matured by around 1950. A new and more sophisticated biological species concept emerged, and the richness of genetic diversity within populations became much better understood. The result was a theory very much like Darwin's, but without the minor elements of Lamarckism and Geoffroyism, and integrated with modern genetics. This modernized Darwinism continued to be expanded and refined. Regarding biodiversity, there were two important developments, both of which involved a return to Darwin's original theory. First, there was Darwin's ecology, which was based on individualistic competition and which rejected the notion that organisms do things “for the good of the species” or “in order to benefit the ecosystem.” Ecologists had commonly treated species and ecosystems as if they were superorganisms and had attributed a kind of adaptation to them that could only have resulted from selection of populations. The rejection of “group selection” in its unsophisticated and unscrutinized form was largely carried out in the 1960s and 1970s.

Second, there was a return to Darwin's idea that classification should be genealogical. Although both the extent and the manner to which this program should be carried out have been controversial, it is generally agreed that comparative biology is best conducted within the framework of phylogenetic relationships. Ill-defined “similarity” has proved too subjective and not rigorous enough for modern biodiversity research. This development has partly been the result of new methodologies (especially cladistic analysis). It has also been strengthened by new kinds of evidence, such as isozyme data and genomic sequencing, that supplement the more traditional data.

D. Contemporary Developments

From time to time Darwinism and neo-Darwinism are challenged by what are purported to be genuine alternatives. Some of these challenges, such as efforts to resurrect Lamarckism or orthogenesis, have simply not measured up. Others, such as developmental constraints, are components of Darwinism that may not have received as much attention as they deserve. Still others, such as the notion of punctuated equilibria, with its occasional periods of rapid change interspersed with long interludes of stasis, are by no means incompatible with the Darwinian tradition.

Darwinism has never been so monolithic or devoid of major unanswered questions as to preclude a broad range of possibilities for new developments. A present it is by no means clear to what degree processes at various levels need to be invoked to provide a satisfactory account of evolution. The “reductionist” view that would explain everything in terms of genes and essentially ignore organisms and species is obviously too simplistic. The models used in theoretical population genetics make all sorts of unrealistic assumptions, and more realistic ones might give surprising results. There is still no really satisfactory explanation for the prevalence and ecological patterns of sexuality. In spite of much recent progress, many of the traditional problems of phylogenetics remain highly debatable. Much work remains to be done in turning branching phylogenetic diagrams into explanatory historical narratives. The vast majority of extant species remain undescribed, and the fossil record consists largely of “roadkills” strung together with gaps. There are plenty of opportunities for important new discoveries within the Darwinian research tradition.

See Also the Following Articles

ADAPTATION • ADAPTIVE RADIATION • BIODIVERSITY, ORIGIN OF • COEVOLUTION • PHYLOGENY • SPECIATION, PROCESS OF • SPECIES, CONCEPT(S) OF • SYSTEMATICS, OVERVIEW

Bibliography

DEFENSES, ECOLOGY OF

Phyllis D. Coley* and John A. Barone†
*University of Utah and Mississippi State University

I. Diversity and Function of Chemical Defenses
II. Physical and Structural Defenses
III. Biotic Defenses and the Third Trophic Level
IV. Phenological Strategies
V. Plasticity in Defenses
VI. Human Uses of Plant Defenses
VII. Summary and Conclusions

GLOSSARY

chemical defenses Compounds used by plants to deter or poison herbivores and pathogens.
constitutive defenses Defenses that are manufactured and maintained by a plant, regardless of whether it has been attacked by an herbivore or pathogen.
endophyte Fungus or other organisms residing or growing within plant tissues.
herbivory Damage to plant tissues by herbivores or pathogens.
induced defenses Plant defenses, including both chemical and physical defenses, that are produced, at least in their final form, only after the plant has been damaged by herbivores or pathogens.
mutualism Interactions between organisms of different species that increase the fitness of both participants.
secondary compounds A synonym for chemical defenses in plants; contrasts with chemical compounds used in primary metabolism, such as photosynthesis and cellular respiration.

PLANT DEFENSES ARE ADAPTATIONS that reduce the damage and mortality caused by herbivores and pathogens. Here we describe the diverse array of features that have a defensive role in plants. Most prominent are chemical defenses that plants use to deter or poison their natural enemies. Also important are physical defenses, such as spines and trichomes, that prevent herbivores from feeding on plant tissues, and mutualisms, in which plants feed or house ants, other arthropods, or fungi in exchange for their defensive help. All of these defenses are presumed to cost plants in terms of energy and nutrients, and they only evolve when their benefits outweigh these costs. Selection, however, has favored different investments in defenses across habitats, reflecting both underlying costs of defenses and the potential for plant damage. We conclude with a discussion of how humans have used plant defenses, primarily chemical compounds, in their everyday lives, underscoring their importance in both traditional and Western societies.

1. DIVERSITY AND FUNCTION OF CHEMICAL DEFENSES

Plants contain a tremendous diversity of chemical compounds that have no function in any aspect of primary
DEFENSES, ECOLOGY OF

### TABLE I

<table>
<thead>
<tr>
<th>Class of compound</th>
<th>Activity</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alkaloids</td>
<td>Variety of mechanisms; often disrupt nervous system function</td>
<td>Caffeine, nicotine, morphine</td>
</tr>
<tr>
<td>Nonprotein amino acids</td>
<td>Disrupt proteins when incorporated in herbivores</td>
<td>Canavanine</td>
</tr>
<tr>
<td>Phenolics</td>
<td>May bind proteins, reducing the nitrogen available to herbivores</td>
<td>Hydrolyzable tannins and condensed tannins</td>
</tr>
<tr>
<td>Sapogenins</td>
<td>May reduce uptake of sterol from herbivore digestive system</td>
<td>Digitonin</td>
</tr>
<tr>
<td>Terpenes</td>
<td>Various; often toxic</td>
<td>Pyrethrins, hormone mimics</td>
</tr>
<tr>
<td>Toxic proteins</td>
<td>Variety of mechanisms; highly toxic compounds</td>
<td>Resin</td>
</tr>
</tbody>
</table>

metabolism, such as in cellular respiration or photosynthesis. Therefore, these compounds are frequently called “secondary compounds” and were originally thought to be either waste products or storage molecules. In the 1930s, however, it was realized that these compounds were in fact defenses against herbivores and pathogens. There is an impressive variety of forms of these compounds, and we begin with a brief description of the more common types and how they act. For additional details, see Rosenthal and Berenbaum (1991) and Harborne (1988) (Table I).

### A. Phenolic Compounds

Perhaps the most common chemical defenses in plants are composed of phenols. Simple phenols consist of a benzene ring with a hydroxyl group attached, along with varying functional groups. More than 200 different simple phenols have been described, and they have been found to deter feeding of both invertebrate and vertebrate herbivores. Several studies have shown that plants with higher concentrations of phenolics were avoided by species of herbivorous birds (Harborne, 1988). For example, the Canada goose selects plant species that are low in phenolic content, and in the subarctic phenolic resins deter the feeding of partridges and grouse on several plant species. In the latter case, the resins may be effective defenses because they have antimicrobial effects and may disrupt digestion by microbes in the birds’ caeca.

Compounds composed of multiple phenolic units are called tannins, most of which belong to either of two groups. Hydrolyzable tannins have relatively simple structures that are made up of phenolic acids, whereas condensed tannins (also called proanthocyanidins) have more complicated structures and are made from the condensation of hydroxyflavonol units. Condensed tannins are also more common across plant taxonomic groups, occurring in 54% of angiosperm and 74% of gymnosperm genera that have been examined. In contrast, hydrolysable tannins are found in only 13% of angiosperm genera (all dicots) and are absent from gymnosperms. Although the actual mechanism of tannin action is controversial, tannins may function by binding proteins. As a herbivore chews a leaf, the tannins stored in the leaf vacuole are released and mix with the proteins, binding them and rendering them indigestible. Since proteins are a primary source of nitrogen for herbivores, this chemical reaction makes leaves less nutritious, although it is unclear whether this reaction is effective in the digestive tract of herbivores. Nonetheless, evidence suggests that tannins do deter feeding and are even toxic to some herbivores.

### B. Alkaloids

Though a structurally and biochemically diverse group of secondary compounds, all alkaloids have a heterocyclic ring that either contains a nitrogen atom or has one attached to it (Roberts and Wink, 1998). Some of the best known include atropine, the main poison in deadly nightshade (Atropa belladonna), quinine, caffeine, cocaine, morphine, and nicotine. It has been estimated that between 15 and 20% of all vascular plants contain alkaloids, and that they are found in one-third of angiosperm families, primarily the dicots. The distribution of alkaloids in relation to climate or geographic distribution remains in dispute, but in general it appears that alkaloids are a more common form of defense in tropical habitats than in the temperate zone. About 16% of temperate species in one survey contained alkaloids versus 33% of tropical species. The mechanisms by which alkaloids affect herbivores (including humans) are highly varied and specific and often effective at small doses, making them effective poisons (see Section VII). Many that have been described affect the functioning
of the autonomic nervous system by competing with neurotransmitters. An alkaloid found in the genus
Erythrina (Fabaceae) blocks acetylcholine receptors in animals, whereas caffeine, an alkaloid found in Coffee
arabica (Rubiaceae), causes the release of calcium from neural receptors. Both types of disruption can lead to
the death of susceptible herbivores.

C. Terpenes
These are an extremely heterogeneous group of chemicals derived from a common metabolic pathway that
produces polymers of isoprene units, a five-carbon compound. The isoprene units may then be arranged in a
variety of ways, generating a diverse range of nonpolar compounds. Monoterpenes, which have 10 carbon
atoms, are often volatile and include pyrethroids from the genus Chrysanthemum that are highly toxic to in-
sects and used as commercial insecticides. These neurotoxins disrupt sodium channels in neurons, causing
uncoordinated movement and paralysis. Sesquiterpenes, diterpenes, and triterpenes are larger polymers
containing 15, 20, and 30 isoprene units, respectively. Terpenes are especially common in pine and fir trees,
which secrete the chemicals in resins that reduce insect attack, although monoterpenes can act as feeding at-
tractants for some species of beetles.

Terpenes can be combined with other molecules to form additional types of secondary compounds, for ex-
ample, sesquiterpene lactones. In this case, a lactone ring (composed of three carbon atoms, with an oxygen
and a carbonyl group) is attached to a sesquiterpene. These compounds are found predominantly in the fam-
ily Asteraceae. Studies have shown that they are toxic to a variety of herbivorous insects and mammals, partic-
ularly livestock. Currently, the cause of the toxicity of sesquiterpene lactones is not known, although they
cause tissue lesions in mammals and may disrupt hor-

D. Saponins
A type of triterpenoid glycoside, saponins are compli-
cated molecules composed of a polycyclic structure (with either 27 or 30 carbon atoms) attached to a carbo-
hydrate group. Consequently, these molecules have both hydrophobic and hydrophilic properties, making
them water soluble. Taxonomic surveys suggest that saponins are widespread, one study found that they are
present in 500 species from 80 plant families. Saponins have been shown to be toxic to many arthropods, in-
cluding mites, beetles, and lepidopterans. In insects, saponins may reduce the uptake of sterol from the diges-
tive system. Since sterol is necessary for insects to molt, saponins apparently interfere with insect development.
There is little evidence, however, that saponins are toxic to most mammals.

E. Nonprotein Amino Acids
Many plant species produce amino acids that are not
used in making proteins but instead are involved in
defenses against natural enemies. More than 600 of
these nonprotein amino acids have been identified from
a wide range of plant species. For example, canavanine,
a structural analog of the amino acid arginine, has been
found in 1500 species in the legume family. Fed to
sensitive insect herbivores, canavanine causes dramatic
and often fatal defects in development. This effect oc-
curs because canavanine becomes incorporated into
newly synthesized proteins in place of arginine, thus
altering their structure and function. A species that
produces canavanine, the woody vine Diosiea negra,
has seeds that are 8% canavanine by dry weight.
Not surprisingly, these seeds are largely free from attack
except by one specialized beetle (Rosenthal and Beren-
F. Toxic Proteins
Plants also make toxic defensive proteins that produce their effects through a variety of different mechanisms (Kiowa et al., 1997). One such protein, ricin, is produced by the castor bean (Ricinus communis) and functions by inhibiting ribosomes in herbivores. By weight, it is among the most toxic compounds known. In general, there is considerable interest in identifying genes that code for toxic proteins in wild plants. If such genes could then be used to modify agricultural crops, they would be protected from herbivores without the use of chemical insecticides.

II. PHYSICAL AND STRUCTURAL DEFENSES
A. Spines
In addition to chemical defenses, plants also employ a wide range of physical or structural features that either reduce the accessibility of plant tissue to herbivores or deter herbivore feeding. The most apparent of these are spines or thorns on plants. There is little doubt that the primary function of these structures is as a defense against herbivorous animals, particularly vertebrates, although in some cases spines may help plants to thermoregulate or, in the case of some vines, to climb.

B. Trichomes
In some ways resembling small spines, trichomes are structures found on the leaves and stems of many plant species that inhibit attack by herbivores and pathogens, either physically or chemically. The shapes and sizes of trichomes vary considerably across species from slender hairs to stout spikes, and many are armed with recurved barbs. Furthermore, glandular trichomes of many species produce or store chemicals that act as deterrents to herbivores. There are countless examples from plants in both natural and agricultural settings of how trichomes can kill or deter herbivores (Levin, 1973). Acting as physical barriers, trichomes can reduce egg laying by herbivores on some varieties of wheat and increase plant resistance to leaf hoppers and mites in some varieties of cotton. The chemicals secreted by many trichomes include sticky gums that immobilize small insects, as found in some varieties of wild potato and tomato. On other species trichomes secrete chemicals such as alkaloids and terpenes as well as waxes, fatty acids, and alcohols. These chemicals are frequently toxic to herbivores, for example, nicotine (an alkaloid) secreted by tobacco plants (Nicotiana in the Solanaceae) that kills aphids on contact. Thus, trichomes not only act as physical defenses but also provide a way for plants to use chemical defenses against their natural enemies without first having to suffer tissue damage.

C. Toughness
Leaves with thick cell walls, consisting of lignin and fiber, are tough. In one assessment of both physical and chemical defenses of mature leaves in tropical forests, leaf toughness and fiber content were the two most important factors in reducing herbivory. Toughness is an effective defense against insect herbivores, particularly small or immature ones, because they have a difficult time cutting or chewing the leaves. Furthermore, cell walls are largely indigestible; therefore, tough leaves are nutritionally poor.

III. BIOTIC DEFENSES AND THE THIRD TROPHIC LEVEL
In addition to using chemical and physical defenses against herbivores and pathogens, many plant species have evolved complex, mutualistic interactions with other groups of organisms that act as a type of defense against herbivores. In such interactions, the plants typically provide food, shelter, or both, whereas the other organisms defend the plant from its natural enemies.

A. Ants
The best studied defensive mutualisms in plants are those with ants (Huxley and Cutler, 1991). Given the incredible abundance of ants in most habitats and their propensity to forage on leaves and stems, it is not surprising that these mutualisms are so common. Ant–plant interactions vary from loose, facultative associations in which the plant offers rewards to any ants in its vicinity to more tightly coevolved relationships in which both partners display highly specialized traits. In facultative associations, a plant offers nectar, food bodies, and other rewards to lure ants that nest elsewhere to patrol its leaves and remove any herbivores they encounter. These associations are especially common in tropical forests.
There are also many obligate ant–plant mutualisms, in which one or both participants require the other to survive. The best known is the interaction between the tree *Acacia cornigera* (Fabaceae) and the ant *Pseudo- myrmex ferruginea* (Pseudomyrmecinae) in Central America. The plants in this pair have numerous adaptations for playing host to the ants, including swollen thorns on which the ants have their colonies, enlarged nectaries on the leaves that provide the ants with sugars, and modified leaflet tips on which the ants can feed. The young queen ant establishes a colony by landing on a young tree and setting up a nest in one of the swollen thorns. After the colony is established, the worker ants patrol the tree constantly, removing any other insects on the tree and cutting away any other plants that touch it. In one study, when the ants were experimentally removed the plants suffered high rates of defoliation, leading to a reduction in growth and a doubling of mortality during the 11 months of the study (55 vs 28% for control plants with ants). Thus, there is little doubt that ants can play an important role in protecting plants in these tightly obligate mutualisms.

The clear defensive effect of ants on their host plants has been demonstrated for other species as well. Less easily measured is the cost of the ants as a defense: Presumably the benefits outweigh the costs, but what are they? Recent studies show that *Cecropia* trees produce different types of food rewards for their ants in relation to the nutrients available—in effect producing rewards that use the less limiting nutrients. Circumstantial evidence also suggests that the ants are a costly defense since on islands on which native ants are rare or absent, few plant species have either extrafloral nectaries or food bodies.

### B. Domatia, Mites, and Other Predators

The mutualistic relationship between ants and plants is apparently paralleled by similar but looser associations between plants and small arthropods. Many plant species have structures on their leaves called domatia that can serve as homes for mites and small, predaceous insects. Although there are a variety of forms, most domatia are pits, pockets, or tufts of hairs on the undersides of leaves. Domatia are widespread taxonomically among dicots, occurring in 28% of 290 plant families, and they are quite common, at least in some forests. Recent work has supported the hypothesis that domatia provide shelter to arthropods, which in turn attack herbivores on the plant. For example, when artificial domatia were placed on the leaves of cotton plants, the density of predaceous bugs increased, the number of herbivores decreased, and the number of fruit produced increased by 30%. Although preliminary, this work suggests that domatia are a part of loose mutualisms used by plants to defend themselves.

### C. Endophytic Fungi

In addition to using animals as a means of defense against herbivores and pathogens, there is considerable evidence that plants can employ fungi in a similar manner. Most plants are infected asymptomatically by endophytic fungi—that is, fungi that live within the plant, usually between the plant cells. These endophytes are a potential defense since they produce chemical compounds not made by plants.

Although endophytes are common in an impressive diversity of plant species, most research has focused on their relationship with grasses. In fact, their defensive role in plants was only appreciated after it was discovered that cattle suffered toxic symptoms after eating tall fescue grass infected with a fungal endophyte. Since then, grasses infected with fungal endophytes have been shown to be toxic to other domestic and wild mammalian herbivores as well as herbivorous insects. At the community level, research has shown that pastures dominated by endophyte-infected grasses have lower populations of small mammals. Thus, it seems likely that plants can gain some protection against herbivores from fungal endophytes.

### D. Predators and Parasitoids as Plant Defenses

Plant defenses have evolved within ecological communities in which predators and parasitoids that attack their herbivores are often common. Evidence suggests that some plant species may directly or indirectly elicit the help of these predators and parasitoids as a form of defense. Recent research has found that some plants release volatile compounds when damaged, and that these are used as location cues by parasitoids and predators (Turlings and Bentley, 1998). Frequently, these volatile chemicals are emitted only when damage is accompanied by oral secretions of herbivores, meaning that the signals are specifically released when a herbivore is present and not when the plants incur other types of damage. By facilitating the discovery of herbivores by their predators and parasites, plants may enjoy reduced herbivory.
Other defenses, such as tannins and toughness, do not kill herbivores outright but reduce their growth and prolong the time until pupation. As a result, herbivores may actually consume more leaf tissue than if the plants lacked these defenses. This seems paradoxical, but by slowing growth these defenses increase the chance that herbivores will be preyed on while still in the earlier instars (Benrey and Denno, 1997). Since the majority of leaf damage occurs in the final instar, this could greatly reduce herbivory. Thus, the effectiveness of some defenses may depend critically on help from parasitoids and predators.

IV. PHENOLOGICAL STRATEGIES

New, expanding leaves are generally more vulnerable to attack by herbivores and pathogens than mature leaves because they are tender and nutritious. In addition to the defenses already described, many plant species reduce new leaf damage by altering the timing of new leaf production. These phenological strategies are of two general types. In the first, plants produce new leaves during times of the year when herbivores and pathogens are rare. In temperate forests, the spring flush of leaves occurs during the short window of opportunity between the end of winter and the recovery of herbivore populations in the late spring. Similarly, studies conducted in tropical forests in India, Ghana, and Panama have shown that when plants produce new leaves during the dry season, when herbivores are rare, herbivory is reduced. Obviously, this seasonal escape strategy would not work in nonseasonal climates, and because of coincident changes in light and water availability with season it is difficult to demonstrate conclusively that the timing of leaf production is mainly an adaptation to reduce herbivore and pathogen damage.

The second phenological strategy to decrease damage to new leaves is for the plants of a given species to produce new leaves synchronously. By flushing simultaneously, plants may be able to overwhelm their specialist herbivores with an abundance of new leaves so that the chance that any particular new leaf will be discovered and eaten is decreased. One test of this strategy involved following new leaf production and herbivory in a moist tropical forest for a year. The results showed that highly synchronous species experienced lowered herbivory than those that were continuous in production. This phenological strategy may be more important in less seasonal forests in which a seasonal escape strategy may be less effective.

V. PLANT INVESTMENTS IN DEFENSES

A. Assumptions

In their evaluations of plant defenses, ecologists typically make the following assumptions: (i) Herbivory and other tissue damage by natural enemies is bad for plants, reducing their fitness; (ii) defenses reduce damage by natural enemies; and (iii) defenses cost the plants in terms of energy and nutrients. Surprisingly, the first assumption about the negative effect of herbivory on plants has been controversial. Several studies have shown that some plant species, in certain circumstances, are able to "overcompensate" for tissue damage, achieving higher fitness than those that were never damaged. However, in general the evidence is quite clear that herbivory, for most plant species, reduces fitness. The billions of dollars spent annually throughout the world on pesticides to protect crops are a testament to the negative effects of herbivores and pathogens. Experimental studies have also demonstrated that damage reduces plant fitness. In a study by Marquis, for example, shrubs of the species *Piper aureum* (Piperaceae) were artificially defoliated and their growth and seed production followed for 2 years. The results showed that individuals that had lost at least 30% of their leaf area suffered a 30% decrease in growth and produced approximately half the seeds of control plants. Clearly, damage affects plant fitness.

B. Costs and Benefits of Defenses

Plant defenses may be costly to plants for several reasons (Fritz and Sims, 1992). First, in order to invest in defenses, plants must divert resources from other aspects of growth and reproduction. For example, nitrogen that a plant invests in alkaloids as a defense cannot be used in making proteins necessary for greater photosynthesis. A second reason why defenses may be costly to plants is that some chemical defenses, although effective deterrents against natural enemies, may also be toxic to the plant, so the plant has to expend additional energy protecting itself from its own defenses. Tannins must be sequestered in the vacuole to avoid cell damage, and many terpenes and resins are restricted to specialized canals in plants.

Third, a defense that reduces plant damage from one herbivore or pathogen species may inadvertently make the plant more attractive to another. For example, Eisner and coworkers showed that the trichomes on the leaves and stems of *Mentzelia pumila* (Loasaceae)
frequently ensnare a variety of insects, including many damaging herbivores. However, the trichomes do not affect a damaging aphid species, whereas they kill one of its predators. Thus, the trichome defense protects the plant from some herbivores but makes it more vulnerable to others.

Despite a variety of direct and indirect costs of defense, plants generally benefit from being defended from herbivores and pathogens. One instance was already discussed: When the ants were removed from ant-defended Acacias, herbivory and plant mortality increased. Other studies of chemical defenses have shown the benefits of chemical defenses as well as their costs. One such study was conducted by Berenbaum and associates, who examined whether wild parsnips benefited from being chemically defended in the presence of the herbivore parsnip webworms. They showed that plants with greater chemical defenses had higher fitness in the presence of the webworms, meaning that the defenses had a clear benefit. The chemicals were costly, however, because in the absence of herbivores in the greenhouse plants that were better defended chemically had lower seed production. Under natural conditions, however, the benefits of reduced herbivore damage outweighed the costs to these plants.

C. Induced Defenses

The chemical defenses described so far are usually constitutive—that is, they are produced and maintained regardless of whether herbivores or pathogens have damaged the plant. Many secondary compounds, however, are induced, with production (or at least the final stages of production) only occurring after the plant has been attacked (Karban and Baldwin, 1997). Usually, an aspect of damage, such as partially eaten cell walls, leads to a transduction process that causes the cell, tissue, or whole plant to begin synthesis of defensive compounds. This entire process can take place in less than 1 hr. In field and laboratory conditions, plants that have been induced suffer less herbivory and have higher fitness than controls. However, most species do not appear to have inducible defenses.

Plants should have induced defenses instead of constitutive ones if defenses are costly to plants, and energy and nutrients that are allocated for secondary compounds or other compounds cannot be used for growth or reproduction. Therefore, by producing defenses only when they are needed—during an attack by herbivores or pathogens—the plant is able to divert these resources to growth and reproduction. However, most species do not appear to have inducible defenses, suggesting that plants may frequently not be in the position to predict when it would be advantageous to induce defenses.

D. Theories of Plant Defense

Both across and within habitats, the amount of damage that plants suffer from herbivores and pathogens varies enormously. In a tropical forest, for instance, rates of herbivory to mature leaves can vary more than two orders of magnitude, from species that receive virtually no damage to those that lose more than 0.6% of their leaf area per day. This huge range implies that plants differ in both the types and amount they invest in defenses.

Much research has focused on trying to correlate defensive investment with other plant life history traits and habitat preferences. It is generally accepted that short-lived, fast-growing species, which typically are found in resource-rich habitats, are not as well defended as slower growing, longer lived woody species. However, the underlying reasons for these patterns are not understood, and numerous theories have been proposed. Two important elements appear in many of these theories: value and risk. First, selection should favor high investments in defense if the risk of herbivory is high. Second, effective defenses should be favored in tissues with a high value or replacement cost.

Apparenacy theory was the first influential attempt to provide a theoretical framework for plant defenses (Feeny, 1976). It emphasized differences in the risk of herbivory, arguing that defense investment should depend on how apparent plants are to herbivores. Species that are ephemeral, such as annuals, may not be readily apparent to herbivores, whereas long-lived trees and shrubs should be readily found. Thus, unapparent plants should be able to largely escape specialist herbivores and, as a result, should invest less in defenses and use defenses that are toxic at low concentrations to generalist herbivores. These “qualitative” defenses include low-molecular-weight compounds such as alkaloids, cyanogenic glycosides, and monoterprenes. In contrast, apparent plants are certain to be discovered by host-specialist herbivores and should produce leaves that are generally less palatable. The theory predicts that they should invest in “quantitative” defenses, such as tannins and toughness, that are not poisonous per se but have dosage-dependent effects, making leaves harder to digest. One test of this prediction found that herbaceous plants had a greater abundance of toxic
alkaloids than more apparent, woody plants. Several authors have also suggested that the risk of discovery by herbivores may depend on the frequency of other more palatable species in the community.

Other theories, although acknowledging that herbivore pressure is important, also suggest that the value of the plant tissue should influence the evolution of defense. For example, species adapted to resource-poor environments, such as those with low light or poor soils, have inherently slow growth rates and long leaf lifetimes. The resource availability hypothesis (Coley et al., 1985) argues that slower growing species should invest more in defenses than fast-growing ones because the cost of replacing damaged leaves is higher when resources are limited. Moreover, in slow growers, the decrement of lost growth for a given investment in defenses is small because the species are already growing slowly. However, in a fast growing species, the investment in defenses leads to a higher percentage decrease in plant growth. This means that the relative or opportunity cost of plant defenses is greater for fast-growing species than for slow growers. Thus, for a given abundance of natural enemies, plant species adapted for growth in high resource environments should maximize their growth by investing less in defense and suffering greater herbivory. The optimal level of defense for species adapted to low resource environments should be much higher.

The resource availability theory also predicts the general types of defenses that plants should use. Species with long-lived leaves should invest in defensive compounds that, despite high initial costs due to their high concentrations, have low turnover rates. These defenses include fibers, which increase leaf toughness, and tannins, both of which would be quantitative defenses in the plant apparency theory. In fast-growing species with short-lived leaves, these defenses would not be profitable because the leaves would be dropped before the plants would have an opportunity to recoup their investment. Instead, the hypothesis predicts that species with short-lived leaves should invest in compounds that are effective at low concentrations even though they may have high turnover rates. The combination of low concentration (or low incremental cost) and high turnover rates means that these defenses are relatively less costly to plants that have short-lived leaves, especially because the components of these defenses can be withdrawn from leaves before they senesce. These defenses would be very costly for species with long-lived leaves since they would have to be continually regenerating these defenses over a long period of time.

VI. PLASTICITY IN DEFENSES

The theories of plant defenses presented in the last section attempt to explain why selection would favor different amounts and types of defenses in different species. An underlying assumption is that defense strategies will be optimized for the predominant conditions experienced by each species. However, there is variation around this optimum, depending on environmental perturbations. In general, if a particular resource is available in excess of normal growth requirements, the extra is shunted into making defenses (Bryant et al., 1983). For example, if a plant is fertilized but retained in low light, it will shunt the extra nitrogen into nitrogen-containing defenses such as alkaloids. Of course, this occurs only in species that have the ability to make alkaloids or other nitrogen-based defenses. In contrast, if a plant is placed in a high light environment, the high rates of photosynthesis will lead to an excess of carbon in the form of starch. This extra carbon will be used to produce higher levels of carbon-based defenses such as tannins and terpenes. Note that these patterns are the opposite of the interspecific trends in which high resources select for lower defense (see Section V,D). Thus, shifts in the ratio of carbon and nitrogen available to an individual plant will cause phenotypic changes in defense allocation patterns that do not necessarily reflect the optimal trends seen across species.

VII. HUMAN USES OF PLANT DEFENSES

In their daily lives, humans have long made use of plant defenses, especially plant secondary compounds. This use has been one focus of ethnobotany, which is the general study of the relationship between humans and plants. Here, we briefly review a few examples of how these plant chemicals have been employed by people. More detailed information can be found in Johns (1996).

A. Hunting and Fishing

Undoubtedly, one of the first uses of plant secondary compounds by humans was as poisons on the tips of arrows. As described by Neuwinger (as cited in Roberts and Wink, 1998), such poisons were formerly used with arrows throughout the world (except perhaps in Australia and New Zealand) and are still used in some areas, both in hunting and in warfare. Poisons for hunt-
ing must be highly toxic in small quantities and either easy to process during hunting trips or persistent while on arrow tips so they do not have to be replaced often. For these reasons, cardiac glycosides and alkaloids are the most common arrow poisons.

Cardiac glycosides, which are secondary compounds common in the families Asclepiadaceae and Apocynaceae, were mostly used as arrow poisons in Africa. They were often the primary active ingredient in concoctions of several different plant extracts, with the other plants added to enhance the effectiveness of the poison, to help it adhere better to the arrows, or to fulfill a magical purpose. In northern South America curare poisons were used in blowgun darts. These were typically derived from plants in the genus Strychnos (Loganiaceae; also used in Africa) or from species in the Menispermacae. Chemically, curare poisons all contain quaternary alkaloids, albeit with different structures. Once in the bloodstream, these block the transmission of neural impulses to the skeletal muscles by competing for receptors with the neurotransmitter acetylcholine. Progressive paralysis of the skeletal muscles follows, eventually leading to respiratory failure in the animal (Roberts and Wink, 1998).

In many traditional societies, poisons from plant secondary compounds have also been used in fishing. These poisons are normally put into small ponds or slow-moving streams, in which the concentration of poison will be high enough to kill or at least stun the fish. A variety of compounds have been used as fish poisons, including isoflavonoids, saponins, cardiac glycosides, alkaloids, tannins, and cyanogenic compounds. Some of the better known are flavonoids derived from the tropical legumes in the genera Chondrodendron and Derris, Lonchocarpus, Mundulea, and Tephrosia.

B. Medicines

As has been noted many times, the difference between a poison and a medicine is often only a matter of dosage, and this is certainly true of products from plant secondary compounds. One example is a curare poison derived from the vine Chondrodendron tomentosum that contains the alkaloid tubocurarine. Although it works like other curare poisons in paralyzing muscles, it was introduced to Western medicine in 1939 and is now used as a muscle relaxant in anesthesia.

It would be difficult to exaggerate the importance of plant secondary compounds in the history of medicine, both traditional and Western. It has been estimated that 25% of all prescriptions written in North America are from plant-derived drugs, and the majority of medicines in traditional cultures are derived from plants. However, only about one-half of 1% of the approximately 265,000 flowering plant species have been comprehensively evaluated for biological active compounds, the first step in determining whether a chemical can be effective as a drug. Of course, most plants do not contain medically useful compounds, therefore, an important task is to determine how to screen species in a more effective manner than simply collecting and testing plants at random. One approach, pioneered by ethnobotanists, has been to document the plants used by healers in traditional cultures. In a preliminary study, plants used by a healer in Belize produced four times as many positive results in laboratory assays than did species collected at random. A second approach is to collect plants based on ecological information. For example, as noted previously, an important defense of mature leaves in tropical forests is toughness, resulting from increased lignin and cellulose in the leaves. However, new, expanding leaves cannot be defended by toughness and so are more likely to use chemical defenses. For this reason, one way to increase the effectiveness of plant screenings for drugs may be to focus on young leaves instead of mature ones.

Regardless of the screening process, plant secondary compounds have been an important source of drugs. In particular, alkaloids have provided many noteworthy medicines in both traditional and Western societies (Schmeller and Wink as cited in Roberts and Wink, 1998). These include atropine, codeine, colchicine, ephedrine, morphine, reserpine, taxol, and theobromine. Codeine, for example, is derived from the fruit of Papaver somniferum (Papaveraceae) and is a component of opium. Currently, more than 200 pharmaceuticals contain codeine, and it is used as a cough suppressant, a pain killer, and a sedative. In the West, ephedrine is used as a nasal decongestant in cold medicines and is a treatment for asthma in both Western and traditional societies. Its source is Ephedra sinica and E. shunnungiae (Ephedraceae), and it works in part by raising blood pressure and respiration while causing an opening of air passages. Taxol is a recently discovered diterpene alkaloid in the tree Taxus brevifolia (Taxaceae) that has therapeutic use as a treatment for breast and ovarian cancer. This compound disrupts microtubule assembly in cells, making it an effective weapon against rapidly dividing tumor cells.

C. Self-Medication in Animals

Mounting evidence suggests that humans are not the only animals that use plant chemical defenses as medi-
Naturalists have long observed that some animals occasionally eat or rub leaves or other plant parts on themselves in an apparent attempt to cure either diseases or kill off parasites (Clayton and Wolfe, 1993). Demonstrating that this behavior is truly self-medication requires showing that the behavior is deliberate, that the plant substance used kills parasites or disease-causing organisms, and that the self-medication leads to increased fitness in the animal. Since in many cases the behavior is rare, gathering evidence to meet these criteria is difficult, but some instances are suggestive. For example, kodiak bears chew the roots of the Ligusticum spp. and then rub their saliva through their fur. Since this species is also used by humans as a medicine, it is possible that the bears are also using the plant to cure or ward off infections. Other vertebrates, including chimpanzees, white-faced monkeys, and birds, also use leaves of particular plant species in unusual ways or only when they are sick, suggesting that they may have medicinal value. However, it remains difficult to demonstrate conclusively that animals are truly using plant defenses as medicines.

D. Spices in Foods

In addition to their uses in treatment of diseases, plant secondary compounds may also be used to prevent illness. Billing and Sherman (1998) argued that people use spices in food preparation to inhibit or kill pathogenic microorganisms. They marshaled several lines of evidence to support this "antimicrobial hypothesis." First, they noted that numerous studies have demonstrated that many spices, such as onions, garlic, pepper, and chili peppers, can inhibit the growth of or even kill many species of bacteria. Second, by comparing recipes of traditional cuisines, the correlations observed by Billing and Sherman also noted that spices are used in small quantities and thus the alternative explanation that they are used for nutritional reasons seems unlikely. Although other factors may come into play, including the greater diversity of plants in equatorial countries and the relatively slow incorporation of newly available spices into traditional cuisines, the correlations observed by Billing and Sherman strongly support their hypothesis.

VIII. SUMMARY AND CONCLUSIONS

Ultimately, the theme of this article, like that of this book, is biodiversity. Plants have evolved a stunning diversity of defenses and strategies to reduce damage by herbivores and pathogens. The most notable of these are chemical defenses, which plants use either to deter feeding (such as tannins) or to poison their natural enemies (such as nonprotein amino acids.) Plant defenses also include physical defenses such as thorns, and many species have evolved elaborate mutualistic relationships with ants, mites, and fungi as an additional means of defense. All defenses have some cost, and a defense will only be favored by natural selection if its benefits in reduced damage and higher fitness outweigh the costs of producing and maintaining the defense. Why particular species invest what they do in defenses, and why they invest in some defenses and not others, remains a subject of controversy, but there are patterns across habitats and successional series, making it clear that the evolution of plant defenses correlates with at least some abiotic factors. Finally, humans (and perhaps other animals) make use of plant defenses in their everyday lives. In food and with medicine, the lives of people have been greatly enhanced by the diverse wealth of plant defenses. Understanding how to make a fuller, more effective use of the diversity of plant defenses found in the natural world is an important, practical goal emerging from this type of research.

See Also the Following Articles

ADAPTATION • PARASITISM

Bibliography


DEFORESTATION AND LAND CLEARING

Jaboury Ghazoul and Julian Evans
Imperial College of Science Technology and Medicine, University of London

I. Introduction
II. Quantifying Global Deforestation
III. Causes of Deforestation
IV. Consequences of Deforestation
V. Alternatives for Sustainable Development
VI. Conclusion

GLOSSARY

afforestation The planting of previously unforested land with trees.
agroforestry The practice of combining agricultural crops or animal husbandry with the maintenance and cultivation of trees on the same patch of land.
criteria and indicators A monitoring system for the assessment of the economic, social, or ecological data about land management practices as an aid to improving sustainability.
deforestation The complete or almost complete removal of tree cover and conversion of forested land to other uses as a result of human activities.
desertification The development of desert conditions as a result of human activity, frequently by overuse of trees and overgrazing, or climate change.
forest degradation Damage to forest ecosystems through human activities that does not result in the total elimination of forest cover.
natural forest An area of land which supports a minimum of 20% tree cover that has arisen as a result of natural processes of establishment and succession.
plantation An area of land that supports planted forest, usually for commercial exploitation.
remote sensing The measurement or acquisition of information of landscape pattern by a recording device that is not in physical or intimate contact with the landscape under study.
secondary forest An area of previously forested land that was subsequently degraded or deforested through human or natural action but which now supports regenerating or mature natural forest.
shifting cultivation Any temporally or spatially cyclic agricultural system that involves the clearing of land followed by cultivation and fallow periods.

AT THE END OF THE TWENTIETH CENTURY, there were an estimated 1700 million ha of tropical forest and 1600 million ha of temperate forests worldwide. These figures represent about 60% of the original forest cover that is estimated to have existed approximately 8000 years ago. Much of this loss can be directly attributed to human impacts during the past three millennia with increased clearance in recent centuries, and even more recently in tropical regions. Forests provide important resources and a multitude of natural services, and their recent rapid destruction is causing increasing concern due to environmental, social, and economic
problems throughout the world. However, developing solutions is proving to be a highly complex task due to the variety of causes of deforestation and conflicting stakeholder interests.

1. INTRODUCTION

Deforestation is the complete or almost complete removal of tree cover and conversion of the land to other uses. Technically, deforestation may be defined as the semipermanent depletion of tree crown cover to less than 10%. In this respect, a distinction needs to be made between deforestation and forest degradation, which is the significant damage to forest ecosystems but without the total elimination of forest cover.

This article begins by describing historic changes in forest cover and proceeds to current deforestation trends and their measurement. Section II discusses the causes of deforestation in recent years and section III considers the consequences of deforestation for a variety of environmental parameters. Solutions to the deforestation problem are presented in section IV, and potential future trends are described in section V with a brief discussion of the impact of projected climate change.

A. Historical Deforestation and Land Clearance

At the advent of agriculture, approximately 8000 years ago, forests are thought to have covered approximately 40% of the world’s land area, or approximately 6000 million ha. Up to 1500 AD, the spread of agriculture across the globe resulted in the clearance of many forests, particularly those on the most accessible and fertile land. The areas most affected were the Middle East, the Mediterranean, South Asia, Europe, and the Far East.

Deforestation in the New World was underway long before the arrival of Europeans, but much of this clearance was patchy or localized. Settlers from Europe started to clear accessible forests of coastal Brazil and the Caribbean for sugar plantations. In North America, settlers cleared large areas of temperate forests for cattle ranches and food crops. Wood was used as fuel and for construction.

In the late eighteenth century the industrial revolution put tremendous pressure on the remaining forests to supply fuel for industrial development, and by the end of the nineteenth century many of Europe’s ancient woodlands had been cleared. However, it is also clear that demand for wood and charcoal also led to conservation of resources to ensure optimality of supply.

In the past 200 years, deforestation rates have increased greatly. Between 1850 and 1980, 15% of the world’s forests and woodlands were cleared (Rowe et al., 1992). The world forest area has shrunk to 3500 million ha as a consequence of human exploitation, most of which occurred in the latter half of the twentieth century [Food and Agriculture Organization (FAO), 1999].

B. Contemporary Deforestation

Deforestation and land clearance in the twentieth century have increased greatly, with the highest rates of clearance occurring since 1960. Most current deforestation occurs in the tropical regions, whereas in temperate countries there has been a net increase of forest cover by 0.1% due to reforestation and regeneration policies. In Canada, the area of land under tree cover increased by 1.4 million ha to 417.6 million ha in the late 1980s. FAO has estimated annual rates of forest clearance in developing countries at 15.5 million ha for the period 1980–1990 and 13.7 million ha for 1990–1995. Thus, the total area of forest cleared during this 15-year period is approximately 200 million ha.

The causes of deforestation vary among regions. In Africa, FAO reports that the major direct cause of deforestation is clearance by farmers driven by increasing population pressures. In Latin America, settlement and infrastructure projects in forested areas result in clearance of land for cattle ranching and permanent agriculture, often combined with financial incentives in the form of subsidies and favorable tax policies. In Asia, intensive timber harvesting and shifting cultivation as well as the expansion of large-scale agricultural projects and plantation crops, such as oil palm and rubber, and to a lesser extent transmigration projects, all contribute to deforestation. Forest land is often not suitable for sustainable agricultural development, and as soils become exhausted new areas of forest have to be cleared. For example, 80% of the Amazon basin is ill suited to sedentary farming. Desertification through unsustainable agricultural development has contributed to much deforestation in drier regions of Africa and Asia.

Most deforestation is concentrated in relatively few tropical countries. Fifty percent of global deforestation occurs in 10 developing countries (Table I). Brazil is typical of tropical countries in that deforestation rates were low until the 1970s. Since then, extensive spread of agriculture and ranching, encouraged by government subsidies, and clearance by landless farmers have re-
TABLE I
"Top 10" Deforesting Countries in Terms of Total Forest Loss (1995)*

<table>
<thead>
<tr>
<th>Country</th>
<th>Ranking</th>
<th>Annual loss (ha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>1</td>
<td>2,550,000</td>
</tr>
<tr>
<td>Indonesia</td>
<td>2</td>
<td>1,060,000</td>
</tr>
<tr>
<td>Congo</td>
<td>3</td>
<td>740,000</td>
</tr>
<tr>
<td>Bolivia</td>
<td>4</td>
<td>580,000</td>
</tr>
<tr>
<td>Mexico</td>
<td>5</td>
<td>510,000</td>
</tr>
<tr>
<td>Venezuela</td>
<td>6</td>
<td>500,000</td>
</tr>
<tr>
<td>Malaysia</td>
<td>7</td>
<td>400,000</td>
</tr>
<tr>
<td>Myanmar</td>
<td>8</td>
<td>390,000</td>
</tr>
<tr>
<td>Sudan</td>
<td>9</td>
<td>350,000</td>
</tr>
<tr>
<td>Thailand</td>
<td>10</td>
<td>330,000</td>
</tr>
</tbody>
</table>


resulted in very rapid deforestation (Table II). Large-scale development and industrial projects such as mining and hydroelectric plants have contributed to the high deforestation rates in the tropics. Thus, deforestation is largely a tropical issue. Because a large proportion of the world’s biodiversity is found within tropical forests, tropical deforestation also has very great relevance to global biodiversity.

In temperate countries there is no overall deforestation in terms of net area but a small increase in forest cover owing to policies of regeneration and afforestation. However, this statement disguises a steady and continuing transition from natural forest formations to managed and plantation forests. In this sense, there is some clearance of natural forest in temperate regions with the consequent loss of biodiversity. This trend is likely to decrease as conservation priorities assert themselves, except in eastern Europe and Russia, where the importance of forest resources as an accessible and tradeable commodity takes priority. Nevertheless, in Europe as a whole there is very little natural forest undisturbed by human intervention; the forest of Białowieża in Poland is one of the few examples of reasonably extant forest not dissimilar to the ancient “wild wood.”

II. QUANTIFYING GLOBAL DEFORESTATION

A. Improving Assessment Accuracy

Accurate measures of deforestation are increasingly needed for the development of appropriate forest management policies and for estimating global carbon pools and gas emissions. Conflicting estimates lead to confusion on the part of policymakers, the media, and the public, but obtaining accurate estimates is difficult due to the lack of reliable time-sequence maps, varying standards of forest and nonforest classification, and inadequate ground-truthing of satellite imagery. FAO, the United Nations Food and Agriculture Organisations, the body responsible for the collecting and interpretation of forest cover statistics, is largely dependent on the information supplied to it by national forest departments. These data are often based on out-of-date inventory records and frequently reflect political and institutional biases.

The objectivity of forest cover measurement has increased since the 1970s, when estimates varied considerably due to differences in the methods used. In 1976, FAO produced the first estimates of global deforestation rates extrapolated from information derived from 13 countries. Although these estimates had a large degree of uncertainty associated with them, they did serve to draw attention to the increasing problem of tropical deforestation. The Tropical Forest Resources Assessment Project carried out by FAO and UNEP, the United Nations Environment Programme, in the early 1980s was the first major step to measuring the extent of tropical forests and their deforestation. Seventy-six countries were surveyed, with estimates for 31 countries based on remote sensing data. Models were constructed to estimate deforestation rates through the late 1970s and to estimate projected deforestation trends through 1985. In 1989, Norman Myers attempted to determine global deforestation rates for the late 1980s using data from 37 countries accounting for 97% of tropical forests.
and relying primarily on nongovernment sources and remote sensing.

By 1993, FAO had produced the Tropical Forest Assessment for 1990, a computerized database of forest areas that was used to develop deforestation models to estimate forest cover in 1990 and rates of change in the previous decade. The database covers 90 tropical countries divided into 644 subnational units. Satellite data were available for 71 of these countries, whereas information for the remaining 19 was of generally poor quality and based on coarse resolution and often out-of-date vegetation maps and land-use surveys. Data were particularly poor for many countries in Africa. Multi-year observations are required to generate reliable estimates, but only 24 of the 90 countries satisfied this criterion. Estimates of deforestation for the remaining 66 countries are based almost entirely on models that relate forest cover change to population density, population growth, ecological zone, or initial forest cover. The 1993 assessment estimates the earth's total area of tropical rain forest and moist deciduous forest to have been 13.06 million km² in 1990 with an annual deforestation rate of 106,000 km² during 1981–1990 (i.e., 7.5% net loss of these forest types during this decade).

B. Remote Sensing Developments

The development of remote sensing technology in recent decades has vastly improved the capabilities for the measurement of forest cover and deforestation. Aerial photography, developed in the 1950s, provides high resolution and detail, and despite its relative costs it is still used for ground truthing of small areas. Airborne radar was used extensively in the 1960s and 1970s and has the advantage of being able to provide information in almost any weather conditions. Satellite imaging began in the 1970s with the launch of the first LANDSAT satellite. Of the six LANDSAT satellites successfully launched, three continue to operate, with each providing digital data "images" covering a ground area of about 185 km² with a resolution of either 80 m or 30 m. Radar satellite sensing capable of penetrating cloud cover is being developed and will improve assessment in cloud-covered tropical areas.

C. Conflicting Estimates

Despite the development and use of remote sensing for measuring deforestation, there remain large discrepancies for regional deforestation rates due to differences in the definitions used, geographical areas, types of satellite photos, and different methods of projecting trends. Definitions of deforestation are sometimes confused. "Deforestation" is confused with "degradation," which refers to significant damage to forests without actually removing forest cover entirely. Degraded forests may regenerate to secondary forest formations. Furthermore, given the continuous gradations between vegetation types, the classification of vegetation types into forest and non-forest categories may not be consistent across different studies. Open savanna forest, for example, may be included in some definitions of forest but excluded from others. Worse, the total land area often includes water bodies or open savannas, and the deforested area, which is expressed as a percentage of the total land area, is therefore underestimated.

Satellite data may be interpreted by computer or manually. Manual interpretation is less expensive and may give better results, but observers often differ in their interpretation of the data. Computer-aided interpretation of low-resolution images eliminates inconsistencies of judgment but may overestimate the area deforested if deforestation has occurred in a piecemeal manner. Low-resolution images have also tended to overestimate the amount of burning forest because even very small fires are capable of saturating a 1.1-km² pixel.

Variations in the models used for projection of future deforestation rates explain much of the variation among different projections. Exponential models based on rapidly accelerating deforestation in the 1970s have tended to give excessively high projected rates of deforestation for the 1980s. With improved models and more accurate interpretation of statistics on the ground, increasingly accurate estimates of deforestation are expected in the near future.

III. CAUSES OF DEFORESTATION

There is no single cause of deforestation but rather it is the result of the interaction of social, economic, political, and cultural forces with the environment. Several underlying socioeconomic causes create conditions that favor forest clearance by readily identifiable direct causes.

A. Underlying Causes of Deforestation

The underlying causes of deforestation are the factors that give rise to conditions in which forest clearance becomes a rational or necessary behavior. They may be local or national socioeconomic or political forces, or they may be external global forces such as the state of the global market economy. They are generally beyond
the control of an individual but strongly influence the decisions individuals make regarding the management and use of forests and forest resources.

1. Population Growth and Poverty

Population growth is one of the most publicized underlying factors for tropical deforestation. The global population currently grows by about 1000 million every decade, and most of this increase occurs in developing countries in which deforestation is greatest. Increasing populations place pressure on forests and the resources they supply for food, energy, water, wood, paper, and a variety of other products. The importance of population pressure as an underlying cause for deforestation can be misleading in that it is dependent on the ability of the land to support the population, the importance of forest for supplying goods and services to the people, and the level of control by national or local institutions.

An estimated 2.8 billion people live in rural areas and are dependent on agriculture to meet their basic needs. The rural poor have very few economic options and are often forced to seek short-term solutions to their economic problems. These solutions include clearing forested land to grow subsistence crops. Opportunities for improving livelihoods by other means are limited due to the lack of rural capital, low capacity of subsistence farming to generate income, and the lack of infrastructure and education.

2. Development Policies and Tax Incentives

Debt repayments constitute a large proportion of the national budget of many tropical countries, and structural adjustment programs introduced as a result often favor the maximization of foreign exchange through direct and unsustainable exploitation of forest capital and by conversion of forests to agriculture for export crops. Large-scale extensive agricultural development, frequently at the expense of small farmers as well as forest cover, is further encouraged through the provision of state subsidies for agriculture and livestock expansion, reduction in income and corporate taxes, and tax breaks on imports of equipment for new industries. Expansion of agricultural crops for export or to satisfy national demands destroys forest directly but also causes the displacement of subsistence farmers who are forced to relocate and clear new and often marginal lands elsewhere.

The privatization of public resources, advocated by the World Bank and other bilateral donor agencies, favors management strategies that maximize the short-term economic gain for the new owners, whereas non-monetary forest services, such as soil conservation and watershed protection, are not valued highly in a market-driven environment. Government incentives and subsidies have allowed some otherwise uneconomical industries to prosper at the expense of forest cover, whereas development projects often fail to account fully for the value of forest capital lost.

A lack of understanding of the real value of forests' goods and services results in poor policies. The institutional weakness of the national forest department or corruption within the government can lead to policy decisions that favor private interests at the expense of the benefits to society as a whole. In recent years there has been an improvement in the reformulation of forest policies of several tropical countries. Subsidies that promote cattle ranching have been withdrawn in Brazil, whereas Costa Rica is now beginning to account for the destruction of forest capital in its national economic accounts.

3. Resettlement Programs

Government-sponsored resettlement programs, such as the transmigration programs of Indonesia and Brazil, encourage landless farmers or the urban poor to clear areas of forest for subsistence agriculture or for the cultivation of cash crops for export. These schemes relieve the pressure of urban crowding and allow governments to avoid the difficult issue of land reform by providing new agricultural land from forest. However, many forested lands are unsuitable for permanent agriculture and resettlement programs have invariably failed as a result. Transmigrants are thus forced to revert to shifting cultivation, leading to further forest degradation and deforestation.

4. Tenurial Policies

Much of the agricultural land of tropical countries is owned by large landowners or corporations and is therefore not accessible to the majority of the farming population. Improved agricultural production is gained through the use of chemical fertilizers and pesticides which, together with mechanization of labor, is most efficient on large-scale agricultural systems. This favors large farmers who have the capital to invest in such innovations and the land area to benefit from economies of scale. As the large landowners become more economically successful and powerful, small farmers who may not have legal title to their land are frequently displaced or forced to sell their land because of increasing debts. These farmers often move to the forest frontier to clear a new plot of land, and it is usually politically easier for governments to ignore deforestation than to deal
with the difficult issues of land redistribution or job creation.

Legal ownership of land has a great effect on the attitude people have of the land. Without some legal land title there is little incentive to invest in the land to make it more productive. It becomes more economically logical to pursue short-term gain and to move on to clear new forest land once productivity declines. Because most tropical forest lands are owned by the state, clearance is often illegal and governments are unwilling to grant legal title to small farmers for land acquired in this way. Lack of ownership of the land excludes farmers from obtaining credit to purchase seed or fertilizers and pesticides and discourages any long-term investment.

In countries in which there is rapid growth of the rural population, individual farm plots that pass through inheritance to subsequent generations become smaller as they are divided among the inheritor, and eventually they become too small to be economically viable. Consequently, the growing rural population is forced to clear new land from the remaining reserves of unused forest.

In many Pacific rim countries, customary land ownership prevails. Precise boundaries are frequently unsurveyed, and local communities and groups know only from tradition what is their land. Although attachment to such land is deeply held, wholesale allocation of logging rights can be conveyed either by the local people with relatively little outside control or, conversely, imposed from outside by governments that fail to take account of local peoples’ interests.

In either case, the land tenure system is a weak instrument in preventing unplanned deforestation and land clearing.

5. Market Demands
As populations grow and become more affluent, the demand for forest products increases, particularly for industrial timber and pulpwood for making paper. However, although it can be readily demonstrated that some countries have significant exports of forest products (Table III), the extent to which international markets contribute to deforestation varies greatly from country to country. In the top 10 deforesting countries, it is the national demand for forest products that accounts for most industrial deforestation. Furthermore, there is no strong link between increasing international demand for forest products generally and deforestation. This is because almost all pulp and paper comes from temperate forests or specially established tropical plantations since paper manufactured from mixed tropical forest is generally expensive and has poor quality owing to the lack of wood uniformity. Second, much of the world’s industrial-grade timber is softwood (i.e., coniferous), and tropical forests are overwhelming broad-leaved. Supplies of industrial lumber in tropical countries increasingly come from tropical plantations, notably pines, cypress, eucalyptus, and teak.

International markets for forest-derived products are important in terms of volume only for Malaysia, which exports approximately 26% of its industrial forest products by volume. Indonesia and Brazil export large quantities of paper and paperboard, but much of this comes

<table>
<thead>
<tr>
<th>Country</th>
<th>Timber products</th>
<th>Paper and paperboard</th>
<th>Charcoal and fuelwood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Production</td>
<td>Export</td>
<td>%</td>
</tr>
<tr>
<td>Brazil</td>
<td>107,360</td>
<td>5,019</td>
<td>4.7</td>
</tr>
<tr>
<td>Indonesia</td>
<td>64,711</td>
<td>9,414</td>
<td>14.8</td>
</tr>
<tr>
<td>Congo, D.R.</td>
<td>3,554</td>
<td>227</td>
<td>6.4</td>
</tr>
<tr>
<td>Bolivia</td>
<td>1,070</td>
<td>150</td>
<td>14.0</td>
</tr>
<tr>
<td>Mexico</td>
<td>5,065</td>
<td>375</td>
<td>6.1</td>
</tr>
<tr>
<td>Venezuela</td>
<td>1,750</td>
<td>52</td>
<td>2.9</td>
</tr>
<tr>
<td>Malaysia</td>
<td>50,923</td>
<td>16,143</td>
<td>31.7</td>
</tr>
<tr>
<td>Myanmar</td>
<td>3,399</td>
<td>689</td>
<td>20.3</td>
</tr>
<tr>
<td>Sudan</td>
<td>2,321</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Thailand</td>
<td>3,636</td>
<td>614</td>
<td>16.9</td>
</tr>
<tr>
<td></td>
<td>12,110</td>
<td>3,396</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td>7,021</td>
<td>2,342</td>
<td>33.4</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3,558</td>
<td>328</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>50</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td>777</td>
<td>41</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2,744</td>
<td>335</td>
<td>12.2</td>
</tr>
<tr>
<td></td>
<td>36,894</td>
<td>53</td>
<td>0.1</td>
</tr>
</tbody>
</table>

*Note. All values in thousand m³. Source: FAO (1999).

*Includes industrial roundwood, sawnwood, and wood-based panels.

*Includes pulp for paper.
from plantations of fast-growing species that were established on nonforest land, previously cleared land, or logged forest. In some cases, market demand for wood-based products has led to clearance of natural forests and subsequent establishment of plantations, but most of this production has been to satisfy domestic rather than international markets. In countries in which deforestation is most severe, it is the national demand for forest products that is the more important cause of deforestation. The importance to tropical deforestation of the international market varies from region to region, but it is likely to diminish as populations in developing countries grow and become more affluent while those in developed countries demand products from demonstrably sustainable sources.

Clearance of forest for agricultural production also appears to be driven by increasing national demand for agricultural crops. All the important deforesting countries listed in Table 1 remain net importers of rice or maize except for Thailand, which exports approximately 40% of its rice production. Also, self-sufficiency in agricultural production has been a primary developmental goal of many of these countries and has led to policies that encourage conversion of forests to fields. Growth in the domestic market can far outstrip that in the export market. During the past decade, for example, the demand for palm oil in Indonesia has led to the widespread establishment of plantations even though palm oil exports have remained at between 6 and 8%. Similarly, rising production of beef in Central and South America to feed a growing domestic market has resulted in extensive deforestation by ranchers, farmers, and land speculators. Livestock exports from Brazil decreased from about 15 to 5% of production during the past two decades, and in a similar period Central American exports declined from 45% to approximately 20% of production.

Deforestation due to agricultural and livestock production is therefore largely a result of growth in the domestic markets and only partially attributable to markets in the developed world. Although the importance of export markets should not be underestimated, national market forces appear to be more important than international trade in determining the rates and extent of tropical deforestation. Consequently, it is likely that international trade offers only limited scope for reducing deforestation rates in most tropical countries.

6. Undervaluation of Forests and Forest Products

Where logging has preceded and been the first stage in wholesale land clearance leading to deforestation, it is often because the value placed on the timber is no more than the cost of extraction and marketing. In this sense, logging is not only unsustainable forest management but also becomes a mining activity in which timber value reflects the ease or otherwise of obtaining the raw material. The value does not reflect cost of replacement or cost of growing beyond what is often a nominal payment of royalty to the owner. If timber was valued to reflect its true cost of replacement, then growing trees to produce timber would become economically worthwhile and hence potentially a sustainable option. While it is not worthwhile, deforestation, especially in the tropics, is likely to continue because clearance and conversion is perceived as more profitable.

However, undervaluation has a further dimension. The non-timber benefits and services trees and forests provide are often far more important than their timber products. Environmental benefits, including soil protection, shelter, micro-climate amelioration, contribution to regional and global hydrological and carbon cycles, are rarely quantified and hardly ever incorporated in economic assessments. Only when massive downstream flooding is traced back to wanton deforestation in the catchment are such connections made.

A major underlying cause of deforestation is the widespread failure to value sufficiently both forest products and the many environmental benefits forests bring. Regrettably, this is largely because simple and widely accepted approaches to such valuations in economic terms do not exist. For example, stumpage, which is the charges that governments demand from loggers for state-owned timber, often undervalues the resource which encourages waste and makes other land uses more economically attractive.

7. Weak Government Institutions

Although almost all countries have explicit forest laws and policies designed to conserve forest, two features of forest conditions in tropical countries exacerbate the risk of deforestation. First, extensive forests by their very geography are likely to be remote from towns and cities and hence far from the rule of law. It is easy for illegal logging and clearance to continue unseen and unchecked. Second, forest services are frequently the "Cinderella" organization of government, being viewed as inferior to agriculture and even wildlife and tourism. Few resources are attracted and poorly paid staff often have difficulty both in ensuring that sustainable management practices are implemented and in imposing their authority on perhaps large private sector interests. Quite apart from the risk of corruption that these circumstances afford, many staff once trained simply dis-
like the remoteness of forest management and supervision and prefer the white-collar work of the city office.

**B. Direct Causes of Deforestation**

1. **Shifting Cultivation**

The contribution of small-scale shifting agriculture to tropical deforestation remains unresolved due to the widely variable agricultural practices that are encompassed by this term. Some types of small-scale agriculture undoubtedly cause deforestation, but Brown and Schrekenberg (1998) indicate that the inherent stability and long-term viability of many shifting cultivation systems are unlikely to result in long-term forest clearance. The least destructive form of shifting cultivation is when land cultivated for 2 or 3 years is then left for a long fallow period. This long fallow shifting cultivation occurs only under conditions of very low population density. Where land is scarce, traditional shifting cultivators using short fallow cultivation make decisions as to which crops to plant, what techniques to use, and how long to fallow the land. These management decisions have allowed this land use to continue in forest areas for centuries without undue forest loss.

Expanding populations, land scarcity, and government policies have also created shifting cultivators who are the typical slash-and-burn farmers of recent decades. Unlike the traditional farmers who have practiced shifting cultivation for decades, these shifted agriculturists have been forced by circumstances or government policy to cultivate habitats that are unfamiliar to them. Government resettlement and transmigration schemes attract migrants for whom forest cultivation is an unfamiliar means of generating a livelihood and income. Similarly, in the Amazon, migrant cultivators are attracted to the forest frontier, where they clear and cultivate land for a few years. This land quickly becomes exhausted due to unsuitable soils or farming techniques and the land is sold to cattle ranchers and the "farmers" move on to clear more land.

2. **Commercial Agriculture**

Large-scale commercial agriculture is most frequently practiced by large corporations or state enterprises. These large operations can dispossess local landowners and farmers of the best and most fertile agricultural land, indirectly leading to deforestation in areas in which the farmers relocate. The establishment of oil palm plantations in valleys of Honduras in the 1970s displaced thousands of farmers who were forced to clear forests from steep slopes to establish new farms.

Commercial agriculture often leads to direct conversion of large tracts of forest to plantation estates and rice fields. This has been particularly prevalent in Indonesia and other regions of Southeast Asia where oil palm, coconut, or rubber plantations have been established on cleared forest land. In Indonesia, oil palm plantations increased from approximately 4 million ha in 1980 to 5.8 million ha in 1995. Land clearance for agricultural development is often subsidized by governments, and because the owners of the agribusiness companies are politically well appointed there is little interest in forest protection. Areas for conversion are frequently burnt because this is the least expensive method of clearance, and natural events such as the 1997–1998 El Niño are used as an opportunity to do so.

3. **Cattle Ranching and Livestock Grazing**

Intensive clearing of forest land in South and Central America occurred as a result of the expansion of cattle ranching, which was economically attractive due to relatively low risk, little labor, well-established markets, and the availability of various government subsidies. Cattle ranching expanded greatly initially in response to the opening of large markets in North America but has been sustained by the development of the domestic markets for beef. Ranchers cleared forest land by purchasing it directly and employing laborers to clear the forest cover or by purchasing land or dispossessing it from slash-and-burn farmers that was then converted to grasslands. The shifting cultivators would move deeper into the forest to repeat the cycle. The area of pasture in Central America is estimated to have increased from 3.9 million ha in 1955 to 13.4 million ha in 1995 and this increase has largely occurred at the expense of the region's tropical forests. Thus, deforestation in Latin America due to ranching is also associated with slash-and-burn agriculture and land speculation.

Livestock grazing can be a serious cause of deforestation in Africa where livestock herds exceed the carrying capacity of the area. Such pressure is acute in the drier tropics such as the Sahel region of Africa and in the Middle East, where large flocks of sheep and goats are maintained. The history of deforestation around the Mediterranean is linked to grazing regimes, especially by goats, but past simplistic assumptions have given way to the recognition that climatic, sociological, and agrarian factors have also contributed to forest clearance.

4. **Infrastructural Development and Industrial Projects**

During the 1970s and early 1980s, development of the Amazon, largely illustrative of similar strategies
throughout the tropics, was actively encouraged by the Brazilian government through the building of roads, tax incentives and subsidies, massive resettlement programs, and large-scale development programs. The Trans-Amazonian highway opened up millions of square kilometers of previously inaccessible forest to colonization and allowed further expansion of the cattle industry. Such roads improve access to poorly developed areas and therefore tend to increase the adjacent land value for non-forest uses and encourage land speculation and deforestation. Recent slowing of deforestation is due to Brazil’s economic recession and has been aided by changes in government policies on tax incentives and subsidies and increased enforcement of environmental regulations. Logging roads in Asia also facilitate deforestation by allowing access to farmers, and illegal loggers who follow may deforest an area that is otherwise merely degraded through selective logging.

Mining and oil exploration have a localized deforestation impact that has been widely publicized. Large mines consume large amounts of wood for fuel, whereas oil companies clear forests to create lines for exploration which are later used by colonizers following the exploration teams. Other large-scale development projects include hydro-electric power plants, whose reservoirs flood forests and transmission line paths are cut through the forest causing permanent loss of forest cover.

5. Plantations

Much forest has been cleared for commercial plantation crops such as rubber, oil palm, and the beverage crops of cocoa, coffee, and tea. Huge areas of dipterocarp forest in peninsular Malaysia have been converted to oil palm or rubber plantations, and although such perennial woody crops offer some soil protection, the loss in biodiversity is enormous. Indeed, it can be argued that the great bulk of the world’s tropical plantation crops, approximately 26.5 million ha (FAO, 1999), are on former forest land.

Conversion of natural forest to forest plantations has also occurred, although currently such practice is deprecated and is less common (Evans, 1992). Indeed, a key principle of good forest stewardship is that forest plantations are only located on already cut-over, abandoned, or wasteland and in this way can actually help deflect pressure away from natural forest. However, the subject is not quite so simple since many forests have enjoyed enrichment by planting or have arisen through tree planting operations, such as many “natural” forests in France, Germany, and elsewhere.

Illustrative examples of significant conversion of natural forest formations to forest plantations include the following: conversion to exotic pines, Gmelina, and eucalypts at Jari in the drier east Amazon for pulp and paper production, conversion on the hills around Bulolo in Papua New Guinea—albeit with the species of Araucaria native to the area—for high-quality softwood veneer logs for making plywood, and clearance of miombo and savanna woodland for pine plantations in Zambia for industrial constructional material.

6. Fuelwood Collection and Charcoal Production

Fuelwood accounts for approximately 80% of all wood use in developing countries. Dependence on fuelwood is expected to decrease gradually with the introduction of electricity, kerosene, and propane, but heavy dependence on fuelwood by the rural poor is expected to continue well into the twenty-first century. Fuelwood collection as an agent of deforestation is particularly marked near urban centers and in villages in which continuous collection results in the gradual degradation and eventual deforestation of accessible areas. This is critical in the dry tropics, along with domestic use of wood for other uses such as construction and fencing material.

7. Logging

The most optimistic independent estimates of the amount of sustainably managed productive forest in the tropics are no more than 2% of the productive forest area. Most tropical logging consists of short-term exploitation of timber products with little or no concern for the future potential of the forest to provide additional resources and revenues. This is largely due to the insecurity of tenure and short concession periods. Although the intensity of logging in the tropics is usually low, removal of only 10% of the timber trees can result in damage to 55% or more of the remaining trees. Nevertheless, logging operations in the tropics usually result in degradation of the forest rather than its complete elimination. Deforestation does occur along logging roads, where forest is cleared for several meters on either side of the road, allowing the sun to dry the road surface. Poorly designed roads can result in severe erosion and landslides as well as facilitate movement into the area by illegal loggers and slash-and-burn cultivators. Illegal loggers may cause much greater damage to forest cover than the initial logging operations. Logging continues to be one of the most important causes of forest degradation but not deforestation, although intensive logging in Southeast Asia has resulted in the conversion of thousands of hectares of forest to alang-
along (Imperata cylindrica) grassland that excludes almost all other vegetation.

Following industrial extraction of timber, farmers, agribusinesses, ranchers, fuelwood collectors, and illegal loggers move in along logging roads to clear the land for other uses. Management plans and government policies oblige industrial foresters to prevent encroachment of this sort, but these rules are inadequately enforced due to lack of will or manpower. The failing of forestry departments and companies is the inability to maintain forest cover in a single area.

The length of concessions is very often short term, sometimes less than 10 years, and very rarely more than the length of time for a rotation of the crop. In the absence of a long-term commitment, the logging company has little incentive to invest in long-term forest management. Concessions are also granted for timber with little regard for the other resources provided by the forest and the impact of the logging on local people.

8. Fire

Serious losses in forest cover in Southeast Asia and South America have been reported as a result of forest fires in 1997 and 1998. The causes of these fires are new large-scale commercial agricultural projects (including plantations) and shifting cultivation. The fires were exacerbated by the dry, coarse, woody debris left after logging operations and the very dry climatic conditions caused by the El Niño phenomenon. The area of forest consumed by fires in 1997 and 1998 has not been accurately documented, but estimates vary from 170,000 to more than 2 million ha (Byron and Shepherd, 1998). Extensive fires in tropical moist forests have been previously associated with the El Niño phenomenon, as in 1982, but the underlying causes are clearance of forest to establish plantations of oil palm, pulp wood, and rice, and, in South America, cattle pastures and shifting cultivation.

IV. CONSEQUENCES OF DEFORESTATION

A. Biodiversity

Perhaps the most severe consequence of the loss of tropical forest cover is the associated loss of biodiversity. It is not known how many species become extinct through the loss of tropical forest habitats, but estimates range up to a yearly loss of 50,000 species. Small forest fragments that often remain after clearance usually support only a small fraction of the original forest biodiversity.

B. Ecosystem Services

The loss or disruption of ecosystem services following deforestation is well documented, particularly for regulation of water flow through watersheds and maintenance of soil structure and fertility. Watersheds become susceptible to extremes of water flow following removal of tree cover, with which erosion of topsoil is associated. Thus, forest clearance often results in erratic water flow with periodic droughts punctuated by flood events which put communities living downstream at risk. Loss of water purity and topsoil through erosion deprive rural communities of safe potable water and fertile soils. Tropical forest soils rapidly become marginal for farming following loss of vegetation cover due to compaction, loss of organic material, and leaching of soil nutrients. Sedimentation of coastal fisheries can deprive communities of their sources of food and income far from the sites of deforestation.

Undisturbed forested land exhibits low to negligible rates of soil erosion. The layers of forest canopy, composed of trees and shrubs, intercept rainfall and thereby reduce raindrop impact while the network of tree roots adds mechanical strength. The accumulated decaying debris branches, twigs, leaves, and other organic material on the forest floor—the litter layer—is effective at protecting the soil from direct water erosion and facilitates infiltration of water into the soil. Removal of this litter layer may contribute more to severe erosion than any other single factor.

Clearance of forests disrupts these soil-conserving roles and subsequent rapid soil erosion is frequently observed, as the muddy red, yellow, and brown streams and rivers of so much of the tropics testify. Typical erosion rates under different land uses are presented in Table IV.

Owing to higher evapotranspiration of forests than most other vegetation types, forests tend to consume more water and therefore lead to reduced runoff and diminished stream flow. In dry countries, this can have a negative impact where water is at a premium, but in general undisturbed forest or sensitively managed plantations reduce sediment loads and flooding and may, in some instances, prolong dry season flows owing to improved infiltration. Overall, less water flows from forested catchments, but it is usually of better quality with less sediment, and streams may show more persistent flow.

Land degradation through loss or overuse of forest
TABLE IV

<table>
<thead>
<tr>
<th>Land use</th>
<th>t ha⁻¹ yr⁻¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undisturbed natural forest</td>
<td>Negligible</td>
</tr>
<tr>
<td>Cut-over forest with litter and organic matter intact</td>
<td>1–5</td>
</tr>
<tr>
<td>Cut-over forest with litter and organic matter removed</td>
<td>30–100</td>
</tr>
<tr>
<td>Forest plantations with litter and undergrowth</td>
<td>0.5–3</td>
</tr>
<tr>
<td>Forest plantations with no undergrowth and litter removed</td>
<td>Up to 100</td>
</tr>
<tr>
<td>Undisturbed or lightly grazed grassland</td>
<td>2–10</td>
</tr>
<tr>
<td>Cultivated arable land (depending on slope, terracing, soil type, etc.)</td>
<td>20–400</td>
</tr>
</tbody>
</table>

*Source: Ambar (1996) as cited in Evans (1992).*

D. Economic Losses

Annual loss of forest capital is estimated at $45 billion (World Commission on Forests and Sustainable Development, 1999). Destruction of forest represents opportunity costs through the loss of future revenues and employment that might be derived from the sustainable management of the forest for timber and non-timber products as well as potential income from ecotourism and other such enterprises. Economic losses can be much greater if the costs of lost ecosystem services are factored into the equation. If the non-timber benefits and services tropical forests provide are even partially costed, then the value per hectare of extant forest increases enormously. This is now being quantified. For example, the carbon sequestration role of such forests is valued in ways simply ignored in the past.

E. Social Consequences

Deforestation leads to the loss of traditional lifestyles associated with forested regions. Individual and collective rights to forest land have frequently been ignored by developers, ranchers, transmigrants, and governments. The loss of traditional lifestyles, customs, and religious beliefs of indigenous communities in the Amazon and Borneo has paralleled the loss and degradation of forests in these regions. Approximately 200 million people in the developing world live in or at the forest margins and wholly depend on forests for their livelihoods. Rarely are their concerns heeded or their rights even acknowledged.

C. Global Climate Change

It is generally agreed that global climate is warming, with an increase in average global temperatures of approximately 0.3°C each decade. This is due to the emissions of carbon dioxide and other greenhouse gases from burning of fossil fuels mainly in industrialized countries. However, it is estimated that approximately 25% of total CO₂ emissions are a consequence of deforestation and forest fires as the carbon stocks in forest are liberated. The consequences of global warming are difficult to predict, but increased drought and desertification leading to crop failures and reduced productivity seem likely in many tropical regions. Higher temperatures and reduced rainfall coupled with more frequent and severe El Niño events are predicted to result in increased frequency and intensity of fires in even moist lowland tropical forests. Geographic and altitudinal shifts in the distribution of vegetation types can also be expected.

V. ALTERNATIVES FOR SUSTAINABLE DEVELOPMENT

A. Protection and Management of the Remaining Forests

1. Protected Area Systems

Protected area systems are needed to establish areas that conserve habitat and biodiversity from encroachment and poaching. The majority of countries have not placed 12% of their land area within a protected area system as endorsed by the United Nations Conference on the Environment and Development.

To be effective, protected areas need to be sufficiently large to conserve all the biodiversity they contain and to include a broad range of naturally occurring forest types and all stages of natural succession. Protected
areas need to be established in areas that harbor the best examples of the representative vegetation and in which biodiversity values are greatest or most at risk.

To be successful, protected area systems need to be supported by adequate funding and legislation and managed by strong institutional departments. However, forest protection must be developed with the cooperation of the local communities that use forest resources because alienation of the local communities has led to failure or inadequate protection. Environmentally sensitive development such as ecotourism, sustainable agriculture, or agroforestry should be permitted adjacent to, or even within, forest reserves, allowing local people to benefit directly from them. Protected area staff need to be adequately trained and equipped, and they should be sensitive to the needs of the local population. Environmental education programs should be promoted in and around protected forest reserves.

2. Joint Forest Management

New approaches to managing forest resources involve partnerships between local communities and local or national governments. These partnerships provide for sensitive management that acknowledges the needs of all stakeholders by providing a wide variety of benefits in a sustainable manner. Forest management decisions are made at a local level and are informed by state-supported science and developmental technologies. Providing local stakeholders with access to and benefits from forest resources and empowering them with the management of these resources encourages sustainable development and investment in forests.

3. Sustainable Timber Harvest Procedures

Timber harvest practices that minimize damage to forests are well-known but rarely implemented due to the perceived high cost of doing so. Indeed, over a short time scale reduced impact logging procedures are more costly, but over the longer term they are cheaper due to increased efficiency of extraction (by eliminating extraction of low-value trees and damage to high-value timber) and reduced damage to the remaining forest stand. National standards for logging operations, where they exist, are often flouted and enforcement is weak.

To encourage sustainable timber production, most timber-producing countries have been involved in developing criteria and indicator systems for sustainable forest management. These systems provide tools for assessing the state of forests which can be used to promote and inform sustainable forest management. The development of such systems has been hampered by the complexities of the definition of sustainable forest management and how to interpret the information generated by criteria and indicator systems.

Coupled with criteria and indicator systems is the idea of certification that aims to encourage producers to adopt sustainable practices which, after verification, may be certified as such. However, there are several difficulties associated with certification as a means to reduce forest degradation and deforestation, the principal one being the need to ensure that certified wood is kept separate from non-certified wood from the source through the chain of custody and to the retailer. The success of certification also depends on the market demand for certified forest products, which is currently greatest in Europe. The domestic markets of timber-producing tropical countries account for the majority of tropical timber consumption in these countries, in which the market demand for certified wood remains low.

B. Socioeconomic and Agricultural Development

1. Improving the Productivity of Subsistence Agriculture and Ranching

Agricultural expansion is the most significant cause of tropical deforestation. Large numbers of people are involved in poorly productive subsistence agriculture or extensive cattle ranching into which there is very little investment to improve production and efficiency. State-supported investment in deforested marginal lands, fair credit schemes, and educational development are all needed to provide extension services to improve the efficiency of land use. Greater productivity from and improved use of existing agricultural land will lessen the pressure for clearance of new forested lands and will promote private investment for currently occupied land which in turn encourages sustainable use.

Agroforestry, the practice of combining agricultural crops or animal husbandry with the maintenance and cultivation of trees on the same patch of land, has been demonstrated to increase crop production by as much as 25–200% through the use of multipurpose trees. These trees can reduce soil erosion, increase soil fertility, provide favorable microclimates for crops and livestock, as well as provide firewood or other tree products such as rubber or fruit.

An estimated 200 million ha of degraded lands exist in upland watersheds of tropical countries as a result of deforestation. Restoration of these lands into productive use through tree planting or clearing for agricultural production will further alleviate the pressure on forest
lands from agricultural expansion. Multiple-use species provide benefits directly from the products they provide, but they also serve to renew a watershed's ability to regulate groundwater and reduce soil erosion.

2. Tree Plantations
Tree and forest plantations are not a substitute for natural forest, but appropriately sited they can alleviate deforestation pressures. Plantations on already degraded land can be an excellent source of industrial-grade timber, pulpwood for paper making, and fuel—solid firewood and charcoal. Thus, they provide an alternative source of such products. In the tropics and subtropics, there are estimated to be approximately 70 million ha of industrial plantations, of which probably 55 million ha are reasonably well stocked. This area is slowly increasing, is often more productive than temperate plantations, and represents a major new wood resource. Considering the world as a whole, nearly half of all industrial timber is likely to be sourced from plantations by about 2020 (Fisher and de Fegely, 1999).

Small-scale plantings on farms, around villages, and in compounds help satisfy domestic demand for building and fencing materials, fuel, and local constructional uses. The estimated 20–30 million ha of such "woodland" further diverts pressure away from natural forest.

Profitable plantations risk undermining the perceived value of natural forest and may accelerate conversion of these to plantations. If the primary cause behind deforestation is the conversion of land for agricultural development, it is unlikely that plantations will reduce deforestation. However, tree plantations do offer an opportunity to reduce the pressure on natural forests if the latter are exploited primarily for wood products.

C. Policy and Institutional Reform
1. Forest Valuation and the Reformation of Government Policies and Institutions
National policies need to promote the necessary framework that encourages sustainable forest use before community-level projects can have an effect. However, under-valuation of forests by governments and people undermines forestry institutions nationally and internationally, making it difficult to enact the necessary changes. The lack of forest-user taxes and low stumpage fees discourage sustainable management of forests by undervaluing the resource. Subsidies to competing land uses should be removed and effective natural resource accounting should be established to reflect true forest values. Opportunity costs and externalities associated with conversion need to be factored into assessments of forest values. Thus, the value of the ecosystem services as well as alternative income-generating businesses, such as tourism, should be considered. Furthermore, forest land value can be greatly increased by more efficient use of a greater variety of forest resources and by improved marketing of these products.

Forestry institutions are weak not only because of the undervaluation of forests but also because they lack information about forest impacts and resources. The absence of clear policy guidelines and trends in forests and forest resources leads to a confusion of strategies among national and international organizations. Strengthening these institutions and developing mutually agreed frameworks that provide for more effective policy development and monitoring are needed if deforestation is to be arrested.

Finally, more collaboration, coordination, and communication between government forestry institutions, non-governmental organizations, and local stakeholders are needed to produce national and local development plans that are relevant to the needs of the society. Ecologically based land-use planning needs to include environmental and social impact assessments, and additional forest research is needed to document the impact of public policies on deforestation.

2. Provide Education
Educating the public and political and economic decision makers about environmental and socioeconomic issues related to forests is crucial if policies promoting forest preservation through sustainable use are to be heeded. Education must be based on thorough and demonstrable arguments about the economic, social, and biological benefits of using forests wisely and the costs associated with deforestation.

VI. CONCLUSION
Deforestation is largely a tropical issue. The quality of data on deforestation rates has improved but remains poor and the source of contention and debate. The causes of deforestation are complex and multi-faceted, although socioeconomic factors and trade are foremost. Deforestation impacts the environment through loss of biodiversity and disruption of ecosystem processes and the economy by affecting ecosystem services and because of inefficient squandering of resources. Perhaps the most tragic consequence is the loss of traditional beliefs and customs and the displacement of forest-dependent communities. Solving the problem of defor-
estation requires a suite of strategies that include establishing an effective and global protected area network to preserve forest biodiversity and the genuine implementation of good practice guidelines in the use of forest resources. Ultimately, a reformulation of policy and a change in the attitudes of decision makers are needed to ensure that forests are correctly valued economically, socially, and biologically.

See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • AGRICULTURE, SUSTAINABLE • BIODIVERSITY-RICH COUNTRIES • CATTLE, SHEEP, AND GOATS • DESERTIFICATION • FOREST ECOLOGY • LAND-USE PATTERNS, HISTORIC • REFORESTATION • TIMBER INDUSTRY

Bibliography


DESER T ECOSYSTEMS

James A. MacMahon
Utah State University

I. What is a Desert?
II. What Causes Deserts?
III. Desert Climates
IV. Desert Geomorphology
V. Desert Soils
VI. Distribution and Characteristics of World Deserts
VII. Desert Plants
VIII. Desert Animals
IX. Desert Ecosystem Processes
X. Humans and Deserts
XI. Disturbance in Deserts
XII. Conservation Issues and the Future of Deserts and their Inhabitants

GLOSSARY

alluvial fan Fine sediments deposited by water that form a conical fan when laid down by streams or rivers.
annuals A plant, sometimes called an ephemeral, that survives year to year as a seed rather than as a "plant."
convergence or convergent evolution Species develop similar behavioral, physiological, or morphological characteristics in response to living in similar environments in different places.
desertification Process of changing a nondesert area into what appears to be a desert, regardless of the climate. This process often has a negative connota-
tion relating it to overgrazing by domestic animals, or general human misuse; however, factors, not related to humans, may cause desertification.
Hadley cell A pattern of air circulation that causes deserts at about 30° N and 30° S of the equator. Sun heats the equator, air rises, pressure decreases, moist air moves north and south, cooling as it goes and falls onto the spinning earth, drying as it falls. The falling dry air increases pressure preventing the incursion of moist air.
life-form A system of classification of the plants in communities based on the position of their perennating structures relative to the ground.
mesic Moist, but not wet, environment.
perennating structures The tissue that carries a plant into a new growing season. For example, buds of trees, shrubs, etc., bulbs of some plants (e.g., tulips) or seeds of annuals where the "plant" dies each year.
physiognomy The general appearance of an area. For example, if you are in any desert in the world it has a certain recognizable "look," regardless of species composition.
saltation Sand particles jumping across the surface due to wind movement.
torpe r A condition, usually in birds and mammals, when the body temperature falls a few degrees below ambient for a short period of time. Differs from hibernation in (a) the amount the body temperatures fall and (b) the length of the resting state.
xeric A dry environment.
DEFINING “DESSERT” is difficult because the literature contains so many definitions, using quite different criteria. A widely used approach developed for UNESCO includes consideration of rainfall and temperature (Fig. 1). Others calculate aridity indices using a variety of data. One system recognizes extremely arid (<100 mm [<4 in] annual precipitation), arid (100–250 mm annual precipitation [<10 in]), semiarid (250–500 mm precipitation [up to 20 in]), and several more mesic subdivisions. While overly simple, this system is easy to understand. Other workers define arid lands by the ratio of precipitation to potential evapotranspiration. In this system arid = annual precipitation is 33% of the potential evapotranspiration and semiarid = 33–67%. Using this standard, Africa has the greatest absolute area of arid lands and Europe has the least. The largest single area of desert, the Sahara of North Africa, covers nearly $9 \times 10^6$ km$^2$, an area about the size of the United States.

I. WHAT IS A DESERT?

This essay discusses extremely arid, arid, and some drier semiarid areas (transitions), calling them, collectively, deserts. This area has been calculated as 19 to 23% (29–32 million km$^2$) of the earth's surface. The discussion will not include semiarid grasslands. Deserts range from areas devoid of any conspicuous vegetation to areas moderately well vegetated with shrubs and subshrubs, a scattering of grasses, and a variety of annuals and succulents. They have scath and unpredictable rainfall and poorly developed soils. Deserts represent a highly variable group of ecosystems found in areas ranging from temperate to tropical zones around the world. 1 exclude those cold, high-latitude areas termed polar deserts and those barren areas of high elevations termed mountain deserts.

II. WHAT CAUSES DESERTS?

Generally, deserts are caused by one of four phenomena. First is the rain shadow effect whereby moisture is lost from air as it moves inland over mountains. Moisture-laden air condenses to form rain as it ascends. As the air crosses the mountains and descends it becomes drier and creates arid conditions (rain shadow). The Mojave and Great Basin deserts of North America are predominantly rain shadow deserts, as are some deserts of Central Asia. A second source of drying air involves cooling and condensation of water in air over cool coastal cur-
ing both their inhabitants and their geomorphology. For detailed introductions to the specific deserts of the world, their causes, and biological and geological characteristics, refer to the publications in the bibliography.

III. DESERT CLIMATES

Typical desert climates include low and unpredictable precipitation, relatively cloudless days, and high summer and moderate to cold winter temperatures that vary greatly from day to night.

The clarity of desert air occurs because it has so little moisture. Thus, during the day the ground surface heats to very high temperatures because there is little water to intercept the radiation. The ground re-radiates heat upward causing an increase in the air temperature. At night, when there is no incoming radiation, the ground continues to radiate toward outer space, thus lowering the temperatures significantly. Representative climograms for a variety of deserts are presented in Figure 2.

A. Rain Shadow Deserts

The Sonoran, Mojave, and Great Basin deserts of North America, some areas of Australia, and much of the Patagonian Desert are rain shadow deserts. They tend to have highly variable temperatures and, depending on latitude, precipitation can occur in the form of snow or rain. There are one or two periods of annual precipitation (Fig. 3). Because these deserts cover a wide latitudinal range it is difficult to generalize about their climates. See specific deserts in Section VI.

B. Cool Coastal Deserts

Cool coastal deserts include the Nambih, the Atacama, the Baja California section of the Sonoran, and a small area in northwestern Australia. These deserts are dominated by fog that causes them to have relatively cool daytime temperatures ranging from 15 to 18°C. Condensation occurs at night, so that vegetation is wet many days of the year. Seasons are similar with winter being only 3 to 6°C below summer temperatures. Temperature extremes increase inland where summer maxima may be near 30°C. These tend to be extremely dry deserts, although in El Niño years the Atacama and Peruvian deserts may be subject to rare, devastating rainstorms.

C. Continental Deserts

Two large areas of continental deserts are the Great Basin of North America and the deserts of central Asia. Because they are far from the moderating influences of large bodies of water (usually oceans), temperature extremes are pronounced. Thus, continental deserts are extremely cold in the winter and extremely hot in the summer, where temperatures often exceed 38°C and may reach 44°C. Similarly, winter minima can be extremely low, regularly reaching −30°C and occasionally −50°C in the Caspian Lowlands. These deserts appear to have more water available because of the extended periods of snow cover, however, much of the water in the snow is lost by ablation (water goes from the solid state of snow directly to gas vapor without going through an intermediate liquid stage).

D. Subtropical Deserts

Subtropical deserts include the Sahara and Kalahari of Africa, several deserts of Australia, the Monte of Argentina, and the deserts of Arabia, the Middle East, India, west Pakistan, and the Sonoran Desert of North America. Freezing temperatures seldom occur with lows averaging above 5°C. Daytime temperatures reach 38°C on about one-third of the days each year and maximums may exceed 46°C. The Kalahari, because of its elevation, is an exception to the low temperature rule in that subfreezing temperatures regularly occur over large areas in the winter. Subtropical deserts also have very low humidity, with little dew, frost, or fog, except following some rainy periods. Precipitation may come from either summer storms with high intensity, low areal extent, and short duration or from winter storms of low intensity, high areal extent, and long duration. Some areas (e.g., the Sonoran Desert) may receive approximately equal proportions of both storm types.

Any one desert may be typified by one of the four climate patterns, however, one might also have influences of several patterns. For example, the Mojave Desert is a subtropical desert, in a rain shadow, with some continental influences.

IV. DESERT GEOMORPHOLOGY

The geological substrates of deserts are fully exposed because there are so few plants. Generally, desert topography falls into two categories: (a) shield and platform
deserts that are associated with the stable plains of areas such as the Sahara, Central Asia, and Australia and (b) basin and range deserts characterized by large basin areas dotted by mountain ranges that cover various proportions of the total area and that are frequently the product of tectonic activity. The southwestern United States, Iran, Afghanistan, Pakistan, and parts of central Asia have large areas of basin and range topography. In both desert topographies, one finds alluvial fans, pediments (impressive plains cut into bedrock), playas, dunes, and a variety of incisions caused by moving water.
A. Alluvial Fans

At their upper ends, alluvial fans have coarse soils, steep slopes and deep deposits. Fans gradually change down slope to areas of fine soils, gentle (generally 1% or less) slopes, and shallow deposits. Coalesced alluvial fans, forming bajadas, may occupy nearly 100% of valley floors in some areas. The soil particle size changes that occur down a bajada have significant influences on vegetation composition.

B. Playas

In places where watercourses drain to internal basins without outlets, ephemeral lakes, termed playas, are formed. These are characteristic of the world's deserts with nearly 50,000 playas worldwide (Fig. 4), most seldom exceeding 100 km² (Neal, 1969).

Playas often have commercially significant deposits of a variety of salts. They are common worldwide and are especially conspicuous in deserts of Australia, Africa (north and south), Argentina, Chile, and the United States.
Playas usually lack vegetation except on their perimeter. Many animals use playas and some (e.g., frogs of several genera) seem reasonably adjusted to the unpredictable filling of playas with water.

C. Dunes

Dunes are usually composed of relatively fine sands that are mobile or may have become stabilized either through geological processes or by vegetation. Some dunes are formed of gypsum and are so different from those of silica sand that they contain endemic plants.

Fine particles that form dunes are generally moved across the surface by either saltation or reptation (short jumps or creeping). This movement is caused by the shearing action of wind that is often restricted by sand grain size and shape, the slope of the ground surface, moisture content, and the surface roughness of the ground, especially that caused by vegetation. While there are many types of dunes, the various morphologies repeat themselves across dune fields of the world (Fig. 5). Dune form is controlled by a combination of the availability of sediments, the local wind regime, and vegetation.

D. Arroyos

Arroyos, stream channels formed by periodic water flow, characterize most world deserts, especially in the United States, Africa, and the Middle East (where they are termed wadis). Australia, and parts of South America. Arroyo vegetation is different from that of adjacent desert surfaces in species composition and stature. Many of these species are phreatophytes, drought avoiders that draw water from deep in the soil. Many animals use arroyos as “highways” and for hunting. Vegetation characteristics change as the arroyo goes from deeply incised with straight sides to wider, flat fans.

V. DESERT SOILS

A. General Characteristics

Soils of arid lands are generally low in organic matter and range from slightly acidic to strongly alkaline on the surface. In many places the accumulation of calcium carbonate in the upper parts of the profile may create
hardpans of cemented calcium carbonate, but also of silica or iron compounds. The calcium carbonate layers can be up to 90 m thick and in places are virtually impervious to water.

In most desert areas, with the exceptions of valley floors or gentle slopes, desert soils are thin and bedrock is exposed. Because soil development is relatively poor in desert areas, the parent materials tend to dominate the chemistry of the soil, except where significant plant activity has added organic matter and all of its constituent elements.

B. Pavements

Soil surfaces in deserts often have one of two unusual characteristics—pavements and crusts. A superficial layer of stones, termed desert pavement, gibber, hammada, or reg occurs in many deserts. Pavements can be caused by the removal of fine materials by wind or surface water or by the upward migration of stones caused by several physical processes. Most likely involved are cycles of freezing and thawing or of wetting and drying. Often the stones of pavements develop a thin patina referred to as desert varnish.

C. Desert Varnish

In some deserts (e.g., southwestern United States), 75% of the rock may be covered with desert varnish. The color of varnish ranges from red to black depending on manganese content. There is considerable debate about the process of varnish formation. Some workers believe only physicochemical processes are involved, while others promote a biological model involving lichens and bacteria. Both processes probably contribute to the variety of colors of varnishes. Surprisingly, desert varnish may form in as little as 13 yr.

D. Crusts

The second unusual soil surface in deserts is a thin crustlike layer. When viewed from the side, these crusts are full of vesicles caused by rainfall or perhaps the expansion of soil gases. Crusts that are formed in association with algae and lichens have increased infiltration rates, whereas others may be somewhat water repellent. Most biological crusts involve cyanobacteria, green algae, lichens, and mosses. In many areas of the world, disturbance of crusts decreases nitrogen and organic matter available to plants. Off-road vehicles, cattle, and human foot traffic can all disturb such crusts. In the Great Basin, the winter grazing of cattle had little impact on the soil crusts containing mosses.

VI. DISTRIBUTION AND CHARACTERISTICS OF WORLD DESERTS

The world’s deserts occur on five continents (Fig. 6). Any one continent may contain deserts caused by several factors. Similarly, a continent may contain both warm (precipitation comes as rain) and cool deserts (precipitation comes mainly as snow). The floristic affinities of world deserts are depicted in Figure 7.

A. North America

In North America, four desert types are recognized including one cool desert, the Great Basin Desert, and three warm deserts, the Mojave, Sonoran, and Chihuahuan deserts. These four deserts differ in climate and in the plant species and life forms that dominate them. The Great Basin Desert receives over 65% of its precipitation as snow and is dominated by low growing shrubs of a few genera (e.g., Atriplex, Artemisia, Tetradymia), some bunch grasses, few native annuals, but many alien species, and a few low-stature cacti.

The Mojave Desert, often considered a transitional desert between the Great Basin and the Sonoran, is characterized by winter precipitation in the form of rain, shrubby vegetation of low-form physiognomy, with the exception of the Joshua tree (Yucca brevifolia) and a few other yuccas and mesquite (Prosopis). The Mojave has an abundance of annual species, and in wet years they may number 1000/m².

The Sonoran Desert receives both summer and winter rain in approximately equal amounts and has the greatest variety of life forms, including subtribes, many species of shrubs, and a wide variety of cacti ranging in size from tall (over 12 m) columnar cacti such as saguaro (Carnegiea gigantea) (Fig. 8) down to a variety of small (a few cm high) species. Sonoran sites typically contain few grasses and have both summer and winter annuals, but only locally in the profusion of some Mojave Desert areas.

The Chihuahuan Desert has predominately summer rainfall and is dominated by low shrubs and species in the genera Agave and Yucca. The Chihuahuan has more dominant grasses than other North American deserts. Summer annuals and a number of cacti, but mostly of relatively low physiognomic form, occur.
All three of the warm deserts have a common plant, creosotebush (*Larrea tridentata*), that dominates their landscapes, especially on flats and on the lower slopes of bajadas with calcareous soils. Other species of *Larrea* occur in South American deserts.

**B. South America**

South America contains two general areas of deserts, a strip along the west coast of Chile and Peru from about 5° to 32°S that includes the Atacama and Peruvian deserts and a section east of the Andes on the Patagonian Plateau that includes both the Patagonian Desert 40° to 55°S, a cool semidesert, and the Monte of Argentina 28° to 35°S, a warm desert/semidesert.

The northern Peruvian Desert is a mosaic of vegetation and substrates ranging from sand dunes to rocky soils and from areas devoid of vegetation to those containing trees (*Prosopis*) and large cacti (*Neoraimondia*). Some extensive areas are covered with relatives of Spanish moss. These ground-dwelling species of *Tillandsia* absorb water from the humid air.

The Atacama is among the driest deserts in the world and in places like Iquique, the average annual rainfall is 3 mm. It is characterized by winter fogs that form over the cold Humboldt currents. The sparse vegetation of the Atacama consists of a number of succulents, low cushion-like shrubs and dwarf shrubs with some grasses.

Many basins have salt pans or pavements or, at the least, gravelly surfaces. Vegetation is least developed near the coast where the fog effect is limited. Many areas are dominated by blue-green algae and lichens. As the fog effect increases inland, grasses (*Pennisetum*) may dominate some sites. Some areas have a variety of cacti ranging from low barrel-like species to taller ones with some branching. This is especially so at somewhat higher elevations.

The Patagonian Plateau has vegetation bracketing the full range from somewhat developed grasslands to areas that are clearly arid and support desert vegetation. In those areas, vegetation ranges from some cushion-shaped shrubs in the genera *Brachyclados* and *Azorella*, among others, to normal shrubs including the genera *Mulinum*, *Nassauvia*, and *Verbena*. Many of the plants belong to the Asteraceae, Verbenaceae, and Rubiaceae.
 Significant grass grazing by sheep is common, an ongoing practice since at least 1885.

Moving from Chile to Argentina, we encounter the Monte Desert. The Monte is well known because of comparative studies between it and the Sonoran Desert of Arizona. This comparison was motivated by the similar life-forms and many shared genera of the two deserts. Dominant plant genera include *Acacia*, *Larrea*, *Opuntia*, and *Prosopis* that occur in the Sonoran Desert. Additionally, there are a number of genera that resemble those of North America, including suboases (*Buinsia*) and columnar cacti (*Trichocereus*). The Monte has soils that range from clay to quite rocky on upper slopes but does not include extensive sand dunes.

**C. Africa**

Africa has desert areas at both the north and south ends of the continent. The Sahara Desert covers nearly 7 million km², mostly covered by gravelly or pebbly surfaces or sand dunes. Rainfall ranges from near zero to 200 mm. Plants that occur in the Sahara often reoccur across the deserts of the Middle East in similar microhabitats. On some sand dunes, shrub genera such as *Retama* and *Ephedra* occur along with a variety of grasses. Stony deserts may be occupied by *Ziziphus* and some perennial grasses, as well as a few other shrubs. Saline areas contain several salt-tolerant succulents. Where more moisture is available, several acacias, and a variety of grasses predominate.

The deserts of southern Africa are unrelated floristically to those of northern Africa (see Fig. 7). These three deserts—the Kalahari, the Karoo, and the Namib—differ from each other. The Kalahari is an inland area that ranges in precipitation from 250 to 500 mm; it represents an extensive semidesert, although some areas of the Kalahari include well-developed sand dunes that are stabilized by the presence of plants. Where there are extensive surface depressions, hardpans of calcareous materials form. Plants dominating
the arid portions are shrubs and some scattered trees, along with a variety of grasses.

The Karoo is often divided into two sections. The Nama Karoo is an area covered by dwarf shrubs and a variety of low-growing grasses. This low-form vegetation occupies 600,000 km². The succulent Karoo is about one-fourth the size of the Nama Karoo and has a richness of plants in the Mesembryanthemaceae (Fig. 9). This family and several others contain a variety of succulent plants that offer a rich and unusual flora. Many of the species are sold worldwide to collectors of succulent plants because of their unusual forms, including the widely known “stone plants,” Lithops (Fig. 10). This desert has few tall shrubs or trees.

The Namib Desert is the most extreme of the desert areas, with precipitation ranging from 5 to about 80 mm. Much of the area is dominated by moving sand dunes up to 300 m in height. It is watered mainly by coastal fog that in some places occurs on half of the days each year. Annuals that respond to the sparse rainfall may blanket the area when rain occurs. Among the other plants in this landscape, some of the most famous are not considered true desert plants (e.g., Welwitschia). This unusual plant with large strap-shaped leaves, which is often pictured in desert books, resides in areas along water courses that have high amounts of available soil water.

Two other areas of Africa are sometimes identified as arid. The first is the Somali-Chalbi area that extends along the African coast from the Red Sea down the Indian Ocean to about the equator. In essence this is an extension of the Sahara. It includes zones with <100 mm precipitation supporting low shrubs and dwarf trees. During rainy periods there is good cover by annuals. There are very few areas with no vegetation. The second region is the southern and southwestern portion of Madagascar, which receives ~350 mm of rainfall/yr, but none in an obvious rainy season (i.e., precipitation occurs throughout the year). This area is dominated by a variety of woody plants that have such incredible forms that they are often pictured. These include the genera Pachypodium, Didierea, Alluaudia, and baobabs, Adansonia.

Vegetation of north Africa is similar to that stretching all the way to the Gobi of China and Mongolia. This whole area is essentially a desert broken by some semi-
arid patches or mountains. People award different names to the dissected desert areas, even though they are not dramatically different in many cases. The commonality of the Sahara and parts of Asia is exemplified by the recognition of the Saharo-Sindian floristic region that covers this zone (see also Fig. 7). A floristic region is one that is defined by an internally consistent group of plant species or genera.
D. Middle East

The Arabian deserts cover the Arabian Peninsula and Iran occupying some 2,600,000 km². This area has very close affinities to the Sahara and contains some areas of extreme aridity. The northern portion of the Arabian desert has winter precipitation, in contrast to the southern portion which has scant rainfall year round. The northern area is characterized by the presence of *Zygophyllum-Raeumuria* and *Artemisia-Gymnocarpus* communities and winter annuals. The southern region is characterized by a sparse vegetation with many annuals. The Rub' al Khali (empty quarter) is an extremely dry portion of the Arabian desert that contains the largest continuous body of sand in the world, covering 650,000 km².

E. Asia

The Iranian Desert, covering Iran, Afghanistan, and part of Pakistan, has a continental climate with winter precipitation. Vegetation types vary dramatically with specific substrates. Soils are generally coarse with some developing pavements. There are vast areas of saline plains that are virtually devoid of plants. Where plants occur, genera such as *Tamarix* with numerous species, *Artemisia, Atriplex, Zygophyllum, Calligonum, Ephedra,* and *Salsola* are common. These cool deserts reach their northern limit at about 30°N. The northern sections, dominated by *Artemisia,* appear similar to the Great Basin Desert of North America.

The Thar Desert of western India and Pakistan varies in climate from its northern and western sections where winter precipitation occurs, to the eastern portions where summer precipitation dominates. There are sandy areas broken by zones of rocky soils and plateaus. The soils differ between the north and the east with larger soil particles in the north and small gravel and sand in the east. Dominant genera include *Calligonum, Capparis, Ziziphus,* and *Prosopis.* Salt-laden playas occur throughout.

The Eurasian continent has a band of cool deserts that span 70° of longitude between approximately 40 and 50°N latitude. These are often divided into the Middle Asian deserts, mainly in the former Soviet Union, and the Central Asian deserts, those occurring in Mongolia and China. The Dzungaria Desert forms a transition between the two. The middle Asian deserts grade into the Iranian deserts to the south.

The Turkistan Desert in the western portion of middle Asia has a variety of subdivisions including the Kyzyl-Kum, the Kara-Kum, and Ust-Urt. This vast area has rainfall ranging from 75 to 200 mm/yr falling in the spring and autumn in the north but tending toward winter rains in the south. Summers reach a maximum of 50°C while winter reaches −26°C. The Kyzyl-Kum (red sands) and Kara-Kum (black sands) describe the nature of the substrate, vast sandy basins with a variety of plant genera that are widely distributed in Asia, for example, *Haloxylon* (Fig. 11), a genus often used as firewood, shade for domesticated animals, or even as fodder. Some areas across the Turkestan Desert have highly saline soils in the basins with virtually no vegetation. Less saline, nonsandy areas contain a variety of genera including *Salsola, Capparis, Artemisia, Ephedra, Ceratoide,* and *Calligonum.* The transitional area of Dzungaria has precipitation ranging from 96 to 247 mm/yr. The flora is more like middle Asia than central Asia, including genera such as *Haloxylon, Anabasis,* and *Nanophyton.*

The deserts of central Asia range east to west across China straddling the southern portion of Mongolia. The Takla Makan covers part of central, western China in the Tarim Basin, the most arid part of central Asia. This is a vast sand desert with many areas nearly devoid of vegetation. Where vegetation occurs, the genera *Ephedra, Nitraria, Tamarix, Zygophyllum,* and *Haloxylon* are locally abundant. The Tsaigam is southeast of the Takla Makan in China. Here July temperatures seldom reach 18°C and the mean annual temperature is near 3°C. These cool areas are dominated by *Artemisia-Salsola* communities with representatives of the genera *Ephedra* and *Tamarix* on a variety of soils. Some areas are covered by saline flats that are essentially vegetationless. The Ala Shan is just south of the western half of the Gobi Desert in China. Sand covers nearly 65% of the area with much of it shifting. Rainfall ranges from 39 to 220 mm/yr over a gradient from west to east toward the base of the Ala Shan Mountains. Common plant genera include *Nitraria, Lycium, Kalidium,* and *Haloxylon* (tree forms), especially in sandy washes. The Ordos lies south of the eastern half of the Gobi Desert in China and is a basin dominated by widespread genera such as *Artemisia* on sandy sites, and *Zygophyllum, Raeumuria,* and *Caragana* on a variety of soil types. The basins of the Ordos have saline soils and have either salt-adapted plants or they are vegetationless.

The famous Gobi Desert runs east-west straddling the border of northern China and southern Mongolia. Summer rains range from 100 to 150 mm, although there are years with no rain. In contrast to many of the other central Asian deserts, the soils are mostly rocky,
in many places forming pavements that are covered with varnish. Haloxylon occurs in sandy areas, but perhaps the most widespread and characteristic plant is baghr (Anabasis brevifolia). Other genera that are widespread in these central Asian deserts include Caragana, Raczumria, Kalidium, and Nanophyol. Two other areas that are often referred to as deserts occur on high elevation plateaus. The Tibetan Plateau receives less than 100 mm of precipitation per year and its flora is limited to about 50 species of low shrubs, grasses, and sedges. The Pamir, occurring in part of the former Soviet Union and Afghanistan, has a somewhat richer vegetation but has similar features. Neither of these is a typical desert.

F. Australia

The arid center of Australia covers more than 20% of the continent. If additional scattered arid areas are included the coverage is nearly 4.2 million km². The deserts of Australia are generally broken into five or six zones within the arid center. The Great Sandy Desert occupies an area in the northern part of Western Australia. South of there, the Gibson Desert occupies east central Western Australia and southwestern Northern Territory. Even further south, the Great Victoria Desert occupies southeastern Western Australia and west central South Australia. In the center of the Northern Territory the Tanami Desert is sometimes recognized. South of it, in the southeastern corner of the Northern Territory and northeastern South Australia, lies the Simpson Desert. Some authors include the Stuart Desert of east central south Australia.

Most areas of Australian desert have reasonably well-developed vegetation composed of scattered subshrubs, shrubs, and a variety of perennial grasses. Especially common are areas that contain hummock grasses and low shrublands dominated by members of the Chenopodiaceae. Australia’s deserts fall in the same 30°S latitude high pressure belt as the Kalahari of Africa. The lowest precipitation is near 20 mm/yr and the boundary of the desert is usually set at about 254 mm/yr. Some areas receive more than two times that on occasion. Summer (January) has temperatures ranging to 49°C, but more commonly they are 25 to 34°C. In winter (July) the lowest temperature is 4.4°C with the mean ranging from 10 to 19°C. Sandy areas that are generally
stabilized cover nearly 2 million km² and these support several types of vegetation. The main vegetation types occur in more than one of the deserts. Arid mallee covers 416,000 km². This vegetation contains eucalyptus trees and members of the genera Cassia, Atriplex, and Maireana as dominants. Hummock grass vegetation is the most widespread in the arid zone occupying ~1.6 million km². These areas are dominated by spinifex (Triodia spp.) (Fig. 12). Hummocks of these species are often widely scattered and after rainfalls are interspersed with a variety of annuals.

Areas dominated by Acacia shrublands occur throughout. Acacia shrublands contain several species of acacias, especially A. aneura. These communities are extremely complex, varying significantly from site to site, based on soil characteristics. Grasses form an important portion of the vegetation. Cassia occurs in some places as do other perennial grasses. This vegetation is especially common in the Gibson and Simpson Deserts.

A tussock grass vegetation that covers ~500,000 km² dominated by Astrebla represents a small but significant portion of the arid zone. Astrebla grasslands further east in Australia are of significant economic importance. The general distribution of arid tussock grasses is in the eastern third of the arid zone, while arid hummock grasses are in the central zone.

VII. DESERT PLANTS

Plants in deserts are exposed to extreme and unpredictable values of environmental variables, especially temperature, precipitation, and wind. Under these selection pressures, numerous morphological and physiological adaptations have developed. Current knowledge about these is changing our view of their functions compared to that postulated as recently as 20 years ago. Interestingly, the adaptations of desert plants may occur in other extreme environments such as those with saline soils, extreme cold, or at high elevation sites with high solar radiation and low temperatures. One common adaptation of plants in all ecosystems is to converge in their general structure (life-forms) in areas of similar climate. Table 1 compares the percentage of species in various life-forms in several deserts, a temperate forest.

![Figure 12](image_url) Ayers Rock in the central desert of Australia. The sandy soils are dominated by spinifex grass (Triodia spp.).
and a rain forest. It is obvious that species of annuals dominate deserts.

A common classification of plant responses to desert conditions includes (a) drought-escaping plants—those that bloom only during periods of high available moisture (e.g., annuals), (b) drought-evading plants—those nonstructural plants that conduct their physiological activities only when water is available, (c) drought-enduring plants—those adapted to high water stresses (e.g., many dominant shrubs), and (d) drought-resisting plants—species that store water in some manner to use for photosynthesis during periods of otherwise low water availability (e.g., cacti). These four strategies highlight the complex of adaptations that can occur.

### A. Morphology and Behavior

Desert plants have a variety of structural adaptations, for example, succulent stems or leaves, reduced leaf size, and a more linear leaf shape. It was thought that some of these structural characteristics prevented water loss. For example, having a low leaf area presents less total surface for transpiration. Recently, it has been argued that small leaf size reduces water loss by maximizing photosynthesis because such a leaf is cooler, not simply because the leaf is smaller and transpires less.

Belowground portions of plants are often equal to or exceed the biomass of aboveground portions. In a Great Basin Desert plant, *Atriplex confertifolia*, there are 133 g C/m² aboveground but 682 g C/m² belowground. This investment in roots provides sites for storage of energy as well as a high surface area for the uptake of water when it is available. Additionally, because many plants are exposed to wind, they are well anchored by extensive root systems. Some plants produce extremely long tap roots, sometimes exceeding 30 m, allowing plants to access the water table or water stored as a deep soil moisture.

### B. Physiological Processes

Desert plants need to obtain and retain water and produce photosynthetic pathways under extreme conditions. There are three photosynthetic pathways: C₃, C₄, and CAM. C₃ plants form a four-carbon molecule during photo-
synthesis, while C₄ plants form a three-carbon molecule in the early stages of photosynthesis. C₃ and C₄ plants differ morphologically in that the leaves of the C₄ species are often subdivided into leaflets or have leaves with serrated margins, while C₃ species more often have simple leaves with no incisions. There are also internal anatomical differences in the arrangements of cells of each type.

Most summer annuals are C₄ plants, while most winter annuals are C₃ plants. Compared to C₃ plants under similar conditions, C₄ plants are more effective at photosynthesizing at high temperatures and in high light intensities, and they use water more efficiently to produce photosynthesize. C₃ plants germinate during cool periods, sometimes forming rosettes that allow the plant to develop at the ground surface under temperature conditions not conducive for C₄ plants. Often there is a gradient in the percentages of the C₃ and C₄ plants in a community coinciding with measures of aridity, although this is an imperfect relationship.

The third kind of metabolism, CAM (short for crassulacean acid metabolism), derives its name from the Crassulaceae, a plant family widely distributed in some desert areas of the world but also occurring under cool, nondesert conditions and along the edges of oceans. CAM plants take up carbon dioxide at night but they cannot complete the light-mediated steps of photosynthesis. Instead they produce malic acid that is later used as a substrate to complete the C₃ pathway when light is available. CAM plants, under favorable conditions, can switch to using only the C₃ pathway.

VIII. DESERT ANIMALS

Animals in deserts are exposed to limited quantities of food, low and unpredictable water supplies, high and widely varying temperatures, loose substrates, high incident radiation, and lack of concealment because of scant vegetation.

A. Scant Food

There are several adaptations to the relative scarcity of food. One is to collect and store food when there are abundant supplies in order to have food available for long periods when food is scarce. Increasing body reserves is another form of storage. One obvious case of food caching is the storage of seeds by rodents and ants in deserts across the world. Both ants and rodents vary in the degree to which they rely on seeds. Some are specially adapted to handling seeds and may be nearly obligate granivores, (e.g., some ants in the genus Pogonomyrum in North American deserts). Other species use seeds according to availability as part of their diet. Beetles in the Namib store body fat very readily, as do a variety of lizards in several deserts across the world.

Another way to deal with food shortages, lack of water, or low temperatures is to become dormant. An extreme case is anhydrobiosis, where organisms lose all of their body water and metabolism halts (e.g., some nematodes). A variety of species of invertebrates, including collembolans in Australia and the eggs and immature stages of many insects in several deserts, exhibit this adaptation. A less dramatic approach is to simply lower metabolic rate, so the expenditure of energy is less. This can be done by changing behavior patterns or going into a state of rest, such as torpor, that occurs in heteromyid rodents weighing ~40 g. For example, Perognathus. Another avoidance strategy is simply to migrate away from the site. Many birds that cannot withstand the summer conditions of the deserts use them heavily in winter to obtain food and then migrate when food is less available.

B. Water

Water scarcity and unpredictability in deserts causes a host of adaptations. Amphibians become active only when water is available. It takes ~30 mm of rainfall to cause frogs to come out of the ground where many of them are protected by a cocoon of their shed skins that remains around their bodies. Similarly, birds may be able to breed throughout the year in some areas as long as there is sufficient water. Again, it takes ~30 mm of precipitation to assure that seeds or insects are available to feed their young. Some species are able to store water when it is available, for example, the genus Cyclocrana, an Australian frog that can increase its normal body weight by over 50% by acquiring additional water. Other species conserve body water. For example, many frogs store urea as a nitrogen waste product rather than producing urine which requires a lot of water. Many species withstand higher degrees of dehydration than relatives from more mesic areas. Birds have high mobility, and thus they can travel to water sources when many other animals cannot. In the Kalahari, sand grouse (Pterocles) may fly up to 80 km, one way, to obtain water. When birds drink water they can endure salt concentrations up to about 50% of that in seawater. Salt is also a problem for herbivorous lizards and mammals. Lizards often excrete salt through special nasal glands,
while some mammals use their highly adapted kidneys to concentrate their urine up to 17 to 25 times the concentration of their blood (e.g., sand rats [Psammomys]) in North Africa and hopping mice (Notomys) in Australia. Several small mammals in three families have a different solution to salt problems. These species, Dipodomys microps in North America, Psammomys obesus in North Africa, and Tympanoctomys barrerae in the Monte of Argentina, scrape the outer salty layers from leaves with specialized incisors and discard them, leaving less salty material to eat. T. barrerae has a unique bundle of bristles on the side of its mouth that assists the process. As is the case for many of the extremes in deserts, both behavioral and physiological adaptation solutions may be necessary.

C. Temperature

During periods of low temperatures, most animals simply retreat to burrows or other refugia and may go into torpor. This also occurs at high temperatures, however, some species remain active because they have higher critical thermal maxima than related species in cooler environments. Birds tolerate an increase in their body temperature of up to about 4°C above normal, as can a number of diurnal rodents (e.g., Ammospermophilus) and some large mammals like camels. Interestingly, birds have little other capacity to deal with temperature other than the normal range of cooling behaviors that birds have everywhere.

D. Loose Substrates

Loose substrates create problems for locomotion and for burrowing. Many adaptations for moving across sand surfaces occur. Some lizards in Asia and Australia and North America have enlarged scales on their toes that allow them to move across sand surfaces by dissipating the weight of their body over a larger area. Many species become bipedal on loose substrates so that the force is directed downward rather than laterally. Specialized forms of locomotion include the side-winding behavior of rattlesnakes in North America and sand-dwelling vipers in Africa. Many species swim through loose sand substrates, rather than burrow, and have the capacity to extract air from between the sand grains (Fig. 13). In loose substrates, soil particles could get into the eyes, ears, or other orifices of small animals, and frequently there are hairs, scales, or other protective means of preventing this.

E. High Radiation

Most species seem to lack techniques for dealing with the high radiation in deserts other than avoiding it by seeking shade or burrowing. The fur and feathers of birds and mammals insulates them to a degree. In contrast, most desert lizards have dark pigment covering their peritoneum that absorbs genetically harmful ultraviolet radiation. Interestingly, the genus Cnemidophorus

![Figure 13](image)

**Figure 13** A fringe-toed sand lizard (*Uma notata*) from the northern Sonoran Desert of Mexico. Note the enlarged scales on the hind foot. This helps these swift lizards run across loose sand.
in North America lacks such a peritoneum, but it does have dark pigment in its skin.

Another problem faced in deserts is the lack of cover, and thus, camouflage or cryptic coloration, is common. There has been some argument as to whether or not this is truly the reason that desert animals match their background so well, but most workers tend to accept the cryptic coloration hypothesis rather than some of its alternatives. Examples range from birds that match the color of the sand that they live on, with the same species having different colors in other areas, to grasshoppers and lizards responding in a similar manner.

Desert animals provide some dramatic examples of convergent evolution—that is, the evolution of similar body forms or physiological adaptations in unrelated organisms under the same environmental conditions worldwide. Bipedality in a variety of rodents—for example, springhares (Pedetes) in Africa, jerboas in Asia (Jaculus), and kangaroo rats (Dipodomys) in North America, is a good example of morphological adaptation in mammals. Similar adaptations occur in herbivorous lizards in Australia (Amphibolurus), North America (Dipsosaurus), Asia (Liopis), and Israel (Agama). One of the most striking cases of convergence is the two ant-eating lizards, the Australian thorny devil (Moloch) and the North American horned toads (Phrynosoma) that are unusually similar in body form and habits.

IX. DESERT ECOSYSTEM PROCESSES

In such extreme environments, it is no surprise that the functioning of desert systems is often at one end of a spectrum when compared to more mesic systems such as forests. Nonetheless, the processes that occur in deserts are categorized as they are elsewhere for other ecosystems; it is generally only the values of these processes that differ. It is not possible to cover all ecosystem processes, but it is informative to look at primary productivity, biogeochemical cycling of nutrients, the spatial pattern of resources in deserts, succession, the process of repair after disturbance, and, finally, the role of animals in controlling desert ecosystem processes.

A. Primary Productivity

For the entire world, the annual net terrestrial primary productivity is 56.4 Pg/yr (one petagram = 10^{15} g). Of that total productivity, 0.5 Pg/yr (about 1% of the total) occurs in deserts. This is unimpressive given that arid areas cover 22% of the terrestrial world. There is a strong correlation between annual net primary productivity and precipitation in deserts. Extreme deserts are 0 to 10 g/m²/yr; in well-developed shrubby deserts, this value is 10 to 250 g/m²/yr. These values are much lower than those of temperate grasslands (200–1500 g/m²/yr) and forests (600–1500 g/m²/yr). In North America, the Mojave Desert is the lowest producing area at 25 g/m²/yr. Chihuahuan Desert values are close to 120 g/m²/yr.

B. Nutrients

Nitrogen (N), phosphorus (P), and potassium (K) are important nutrients for plant growth. In arid areas, these nutrients are usually positively associated with shrubs. In a creosotebush community there were 0.253 mol/kg N beneath the canopy and 0.148 and 0.10 at the canopy edge and interspace, respectively. There may be several reasons for this association, the most general one relating to the concept of “islands of fertility.” Once established, a shrub sheds leaves (19–53 g/m²/yr) and portions of its belowground roots die adding organic matter to the soil. Additionally, the roots may extend into the barren area between shrubs and extract nutrients from those areas, concentrating them under shrubs. Rodents often make burrows at the bases of shrubs and the addition of their nesting material and digging activities can increase soil nutrients.

In desert areas, the amount of carbon (C) available is generally lower than in adjacent grasslands, e.g., in a creosotebush stand in Chihuahuan Desert, there were 102 g C/m² in aboveground biomass, and a total of 159 g C/m² both above and belowground in biomass. The total carbon for that area is 2,088 g C/m² including the biomass of plants (above and belowground) plus the stored carbon in the soil. In contrast, the same values for a grassland are 131, 190, and 2,302 g C/m², respectively.

Nitrogen is generally scarce in deserts. It enters the system through fixation by organisms in soil crusts, through fixation by the nodules on the roots of legumes, and through airfall deposition. In the past it was thought that low quantities of nitrogen limited primary productivity. More recently, it has been suggested that phosphorus may be more limiting because it is relatively unavailable in the alkaline soils of deserts. When nutrients exist in desert soils, they may not be available because decomposition occurs more slowly in the absence of water, and because the microfauna and microflora that usually affect decomposition are at much lower densities than in forest or grassland systems. None of these decomposers can continue activity when soils are very dry. In the Chihuahuan Desert, termites
are extremely important in decomposition, but seem less so in drier areas.

C. Succession

Ecosystem repair is referred to as succession. It is controversial as to whether succession occurs in deserts because there is usually not an obvious group of plant species occupying recently disturbed sites (pioneers) and that is ultimately replaced by a climax vegetation. Rather, in many cases following a disturbance, both climax and pioneering species may co-occur. In Israel, 16 of 20 species that were typical of mature communities germinated during the first year following a disturbance. In mesic areas, plants begin to occupy disturbed sites immediately following the disturbance. In arid areas, it may take much longer before establishment occurs. This is probably due to the episodic nature of the conditions necessary for establishment. One good year of rainfall is usually insufficient for establishment, and some species require the presence of other species for establishment. In one common example, some desert cacti must germinate and begin to develop in the shade of a shrub or subshrub. This "nurse-plant" phenomenon (Fig. 14) is known from North American deserts and the Karoo where mound-building members of the succulent Mesembryanthemaceae establish early and are later invaded by woody species. In other cases, it is merely a matter of a statistical probability as to what sequence of plants will occur following disturbance.

D. Role of Animals in Ecosystem Processes

Animals can alter the species composition as well as the environmental conditions of ecosystems. Observations in deserts around the world indicate that animals will enhance particular species and hinder others. When the local extinction of the plain Vizcacha (Lagostomus) occurred in Argentina, the number of forb species declined on colony sites as the cover of grasses increased. Five years later richness of grasses had changed again.

In almost all deserts of the world, one group of three types of organisms are influential. These include granivorous species of harvester ants, small mammals, and birds. These organisms disperse or consume seeds, turn over soil at very high rates, change the physical conditions of the soil through digging activities, incorporate organic matter, and, in the case of ants, remove vegetation from the sites of their nests. We can predict the area disturbed by digging rodents by knowing their longevity and applying the formula: longevity in years = 9.33 Area²⁷³.

Interestingly, the relative importance of these three groups differs around the world. In North America and Israel, small mammals are most important and remove the greatest number of seeds, followed by ants, and then by birds. In the Karoo, ants take more seeds than small mammals as they do in the Monte of Argentina. Birds, while generally taking fewer seeds than the other two groups, may be quite important seasonally. For some species of plants, nearly 100% of the seeds are removed while other species lose a very small portion. This differential feeding can clearly alter the composition of the plant community. Additionally, for species
like the harvester ants that carry seeds back to their nests and belowground, there are dramatic changes in soil properties where these nests exist. The plants that establish on mounds are often not the same as would establish in the disturbed area adjacent to mounds. Clearly, these animals, along with many others, have an important role in structuring desert ecosystems.

Another interesting facet of desert systems is that they seem to operate on what is called the "pulse-reserve paradigm." For long periods, when water is not available or temperatures are too high, plants and animals in deserts, to a great extent, shut down either physiologically or behaviorally. If temperature moderates and water is available, there is a sudden pulse of breeding activity by animals, primary production by plants, and nutrient cycling and decomposition mediated by a variety of soil organisms. The gains made during these favorable periods allow accumulation of material that will carry the organisms through the next period of extreme environmental factors. The adaptations discussed here help with this pulse-reserve approach to survivorship.

X. HUMANS AND DESERTS

A. Human Uses of Deserts

Deserts are home to about 15% of the world's human population. People use deserts and their transitions with grasslands for a wide variety of purposes. Sometimes, when vegetation is destroyed or significantly altered, the process of desertification begins (i.e., transitional grasslands are replaced by desert vegetation). While estimates vary, it is thought that nearly $10^6$ ha are subjected to desertification each year. One can imagine that anything that humans do might be exacerbated by natural disturbances (e.g., drought). Desertification does not have a single cause, rather it occurs because of people, climate, and environments of particular areas.

In addition to using desert areas for livestock grazing and browsing, firewood collection, habitation, and mining, humans also use deserts as sites for recreation. In some places human impact relates to unusual activities. For example, in North America, people with off-road vehicles frequently traverse desert areas at great speeds, tearing up vegetation, altering soil crusts, and changing the soil characteristics in other ways.

The collection of succulent plants around the world by hobbyists has become a serious threat for many species in at least five plant families. Many human activities can have indirect effects. For example, irrigated agriculture can change the vapor pressure deficit over relatively large areas, thus impacting the thermal and light environment of desert plants, often to their detriment.

An unusual use of deserts is to harvest water. In the Atacama large screens are erected to intercept and condense the fog so that it can be used by humans.

Interestingly, deserts seem to have the capacity to rebound from short duration, high-impact disturbances, even the testing of atomic bombs, and yet may be more negatively influenced by long duration, lower level assaults such as grazing and browsing.

B. Uses of Desert Organisms and Their Products

The list of desert animals and plants that are of direct value to human beings is as long as the uses to which they are put. A few examples should suffice: in North America guayule (Parthenium argentatum) was grown as a substitute for rubber for a period of time; jojoba (Simmondsia) oil appears in a variety of beauty products; in the Sudan Blepharis has wound healing properties; Agave is used in Mexico to produce an alcoholic beverage that is known worldwide; Sclius, from the Atacama Desert, has proven to have insecticidal activity, as does Calotropis in India.

Many species have turned out to be good forage plants, especially some in the genus Atriplex. Larrea, occurring in both the North and South American warm deserts, has great potential for commercialization because of its antifungal properties and nutritious forage. Similarly, Acacia is used for food and fuel but may be more important as a source of commercial gums and tannins. Other species worldwide have been used as sources of anti-diabetics, the basis for cosmetics, fibers for weaving and other uses, plant growth regulators, surfactants, waxes and many other uses.

Desert animals have been used in a variety of ways. Animals ranging from beetle larvae to large antelopes have been harvested for food, clothing, weapons, and medicines. Recently, as a greater knowledge of the chemistry of some compounds produced by animals has occurred, they have found even greater use. The venoms of cobras, rattlesnakes, and gila monsters all have medicinal properties under certain circumstances. Some desert animals (e.g., species of camels) have been domesticated in Africa, Asia, and South America.
XI. DISTURBANCE IN DESERTS

A. Natural

Natural disturbances include the effects of all environmental factors exceeding those that plants and animals are generally exposed to. Deserts are systems that occur in extreme environments, so periodic change in a natural factor that exceeds tolerance of desert organisms is highly likely. Because the time to recovery from such perturbations can be long, deserts are especially susceptible to disturbance and may, in fact, seldom be in equilibrium with their environment.

While this is not the place to discuss every environmental disturbance, a few will be informative. Temperature can vary dramatically in deserts (see Table 1). High temperatures have the capacity to kill both plants and animals if they are exposed for long periods. However, there are few data suggesting that established plants die because of high temperature. In contrast, young plants or seedlings may die upon short exposure to high temperatures. Low temperatures, however, are known to kill a variety of desert plants. Catastrophic freezes, where temperatures may reach $-8.4$ to $-5.6^\circ C$ and last for about a day, can wreak havoc as they did in the Mexican state of Sonora in 1937 when a variety of genera were killed following a rather short cold spell.

Temperature may control the life-forms of plants in deserts (see Table 1). It has been suggested that a major limiting factor for the succulent life history strategy is the inability of these life-forms to withstand low temperatures during the growing season. This might explain the relative lack of succulents in the deserts of central Asia, the Great Basin of North America, and the Patagonian Desert of Argentina.

Too little or too much water can always be a disturbance in deserts. In a very real sense, water is the environmental factor that drives desert ecosystems. Different attributes of water (e.g., availability, predictability, and flooding) have different influences on desert species. Deserts are already characterized by low annual values and high year-to-year variation in rainfall. In years of sequential drought, some species may be edited out of the community, although others have surprising resilience. The rebound of the Sahel following a long period of drought, and perceived devastation, is a good example of such resiliency. High rainfall amounts can cause flooding which, in turn, can mechanically remove plants or animals or cover them for long enough periods that respiration is not possible.

Wind is ever present in deserts. Many familiar land-forms (e.g., plants perched on pedestals) are caused, at least in part, by the downcutting of soil by wind in interspaces and deposition beneath plants. If wind carries a substantial load of geological materials, the abrasion can damage or kill plants and animals; the desiccating power of the air, partly due to wind, can bring organisms to critically low levels of tissue water. In the Patagonian Deserts, much of the dryness is due to the effect of wind rather than directly to the amount of rainfall.

Contrary to expectations, fire can be an important factor in deserts, especially in years of above average precipitation when a significant fuel load develops. In an interesting case in the Sonoran Desert, fuel load accumulated over a few years, and a single fire destroyed 68% of all the individuals of saguaro cactus (Carnegiea gigantea). In similar habitats, species like creosotebush (Larrea tridentata) were burnt to the ground but sprouted back a short time later.

In some systems, the activities of animals may form a disturbance for plants. This might take the form of plagues of grasshoppers in some African deserts or the soil turning activities of fossorial mammals, ants, or termites.

B. Anthropogenic Disturbance

Humans alter deserts dramatically. Semideserts hold an even larger portion of the world’s population and nearly every desert and semidesert in the world has been influenced by the grazing and browsing activities of domesticated ungulates.

There are good historical records of the uses of deserts for thousands of years, for example, the early Egyptian cultures in Africa and by the Seri Indians for at least 2000 years in the Sonoran Desert of Mexico. The result of human use is often to change areas that are transitional with grasslands into deserts, at least in part, because of the effect of grazing by domestic livestock or removal of vegetation for fuel. Similar changes occur because of climate variations. Vegetation removal may actually alter some climatic patterns in deserts because of changes in the albedo (reflectivity) of the earth’s surface.

Humans have introduced non-native plants, often to the detriment of deserts. In North America, the introduction of tamarisk (Tamarix) has created problems in riparian areas of deserts where these aliens alter water regimes and animal and plant community composition. Annuals and grasses have been moved around the world and often out-compete native plants. In North America,
the Great Basin Desert is dominated by non-native annuals that actually cause an increase in the fire frequency in those systems compared to natural systems that do not contain a large proportion of annuals. These fires kill native shrubs not adapted to fire (e.g., Artemisia) thus changing the entire community structure.

Obviously, building highways, large irrigation projects, and human settlement in what are often considered to be the favorable climates of deserts has dramatically diminished the area that can, by any sense of the imagination, be considered natural. The potential effect of different land uses is likely to differ dramatically in different desert areas (Table II).

XII. CONSERVATION ISSUES AND THE FUTURE OF DESERTS AND THEIR INHABITANTS

Misconceptions about deserts inhibit efforts to conserve them. First, because they are so severe, many believe that they are not much utilized by human beings. Second, because many animals and plants are inconspicuous it is thought that, compared to rain forests, there are very few species that require conservation efforts. Both of these are incorrect.

As mentioned earlier, the deserts are used, and at an increasing rate, for a variety of purposes by humans, often to their detriment. Second, the rich diversity of desert species are of interest in their own right and are potential sources for materials (food, medicine, clothing) of value to humans.

Even the United Nations, in its 1992 Conference on Environment and Development, called for a "Convention on Desertification" that would be similar to those that have been developed for rain forests.

Setting aside areas that currently represent sustainable desert communities is one form of conservation. Unfortunately, many areas have been degraded to the point where they require restoration, a process more challenging in deserts because of the extreme conditions than it is in more mesic areas. Additionally, certain species have to be restored or even reintro-
duced. For example, a program in Abu Dhabi was developed to restore the Houbara bustard (Chlamydotis undulata) that seems to have suffered from the effects of some range practices. In the Kyzylkum desert of Uzbekistan, studies are underway to reintroduce Przewalski’s horse. Animals that had been zoos reared seem to adapt well when turned out into semiwild, desert conditions.

We must remember that the deserts of the world have been populated by indigenous people who are highly adapted to a desert way of life. While such people have probably altered deserts by their presence, they have come into a sort of equilibrium with the altered deserts and their way of life must be protected, just as any rare plant or animal should be.

One problem with attempting to conserve desert areas is that vast tracts are often required to maintain viable breeding populations of rare species. This need is antithetical to the needs of ever-expanding populations that require spaces to live, as well as places to obtain resources to sustain themselves.

In cases where nature reserves have been established in deserts, many of them are suffering challenges to their boundaries. These often relate to the encroachment of human populations or to the movement of materials or alien species from populated areas into reserves. It has been suggested that large buffer areas are needed and care must be taken to prevent leakage of organisms and effects into or out of the reserve.

Clearly, deserts deserve as much attention as rain forests but they have not captured the imagination of the world’s peoples in the same way. They are not desolate lands that require no attention. Rather, they are amazing places housing plants and animals of indescribable beauty and complexity and humans with folkways that enrich the human spirit.

I have often said that the very starkness of the desert allows one to have an almost religious experience while in awe of the capacity of living organisms to adapt to such extreme and unpredictable environments. For that reason alone, I believe they should be preserved. Everyone deserves that experience.

Acknowledgments

John Mull, Weber State University, and an anonymous reviewer toiled over the manuscript to its benefit. Linda Finchum, as always, worked feverishly to keep me from making too many errors.

See Also the Following Articles

AFRICA, ECOSYSTEMS OF • AUSTRALIA, ECOSYSTEMS OF • C, PLANTS • DESERTIFICATION • NEAR EAST ECOSYSTEMS • SUCCESION, PHENOMENON OF

Bibliography


I. Introduction
II. What Are Drylands?
III. Regional Overviews
IV. "Combating" Desertification
V. Conceptual Model of Desertification
VI. Conclusions

GLOSSARY

combating desertification Activities specifically aimed at prevention and/or reduction of land degradation, rehabilitation of partly degraded land, and reclamation of desertified land.
desertification Land degradation in arid, semiarid, and dry subhumid areas resulting from various factors, including climatic variations and human activities.
dimensions of desertification The interactions and feedbacks of meteorological, ecological, and human components of land degradation.
drylands Arid and semiarid croplands, pastures, rangelands, and subhumid woodlands in which the index of aridity is less than 0.65; drylands cover about two-fifths of the land surface of the earth and are home to more than 20% of the human population.
index of aridity Ratio of mean annual precipitation to mean annual potential evapotranspiration.
land The terrestrial ecosystem that encompasses soils, vegetation, other biota, and the ecological, biogeochemical, and hydrological processes that operate therein.

land degradation Reduction or loss of the biological and economic productivity and complexity of terrestrial ecosystems, including soils, vegetation, other biota, and the ecological, biogeochemical, and hydrological processes that operate therein.

myths of desertification Controversy stemming from failure to consider all dimensions (meteorological, ecological, and human) of the problem, and alarmist tone connoted by the word desertification, which incorrectly suggests the action of deserts "moving" across the landscape, engulfing fertile lands and leaving starving people in their wake.

THE PHENOMENON OF DESERTIFICATION involves the loss of biological or economic productivity and biodiversity in arid and semiarid croplands, pastures, rangelands, and subhumid woodlands due mainly to nonsustainable human activities, such as overcultivation, fuel gathering, overgrazing by domestic animals, deforestation, and poor irrigation practices, and is often triggered or exacerbated by climate variability, mainly drought.

I. INTRODUCTION

A. Land Degradation in Global Drylands

Drylands cover about two-fifths of the land surface of the earth and are home to more than 1 billion people,
or 20% of the human population. The fate of rural people in these regions is ultimately dependent on the effective use of natural resources, e.g., water, soils, plants, and wildlife. However, rapid growth of populations in dryland regions, which have some of the highest birth rates in the world, has contributed to rapid land degradation. This degradation—which encompasses soils, vegetation, other biota, and the ecological, biogeochemical, and hydrological processes that operate therein—has occurred over large areas of drylands in Asia, the Mediterranean, Africa, Oceania, and the Americas. Natural vegetation has been eliminated or severely reduced, soils are eroding at accelerated rates, and the capacity of the land to support human populations, livestock, and wild herbivores has been substantially reduced (Figs. 1a–1f).

Can land degradation be reversed? What is the impact of deteriorated drylands on global climate systems? Do natural factors, such as climate variability, play a role in land degradation? Do ecologically sound land-use practices lie at the level of recognizing the rights and environmental knowledge of local communities? Such questions exemplify the suite of concerns underlying the general concept of land degradation in drylands, or desertification. Desertification involves the loss of biological or economic productivity and biodiversity in

![FIGURE 1](image-url) (a–f) Images of land degradation in arid and semiarid lands. (a) Cattle disturbance around a watering point in northern Australia (photo by J. Ludwig). (b) Soil erosion following logging in dry subhumid region in south China (photo by J. Wu). (c) Fence line contrast between inside and outside grazed rangeland on Santa Rita Experimental Range, southern Arizona (1992). (d) Woody plant encroachment in native grasslands of central Argentina (photo by R. Distel). (e) Demuded grassland in South Africa (photo by S. Milon, courtesy of Bioscience). (f) Extensive soil erosion due to cultivation of marginal lands in central Argentina (photo by R. Distel). (g–i) Rehabilitation efforts to curtail or reverse land degradation. In southern New Mexico, (g) mechanical shrub removal and root plowing, which is extremely expensive and leads to high soil erosion (photo by R. Gibbens) and (h) manure spreading to restore soil fertility (photo by J. Herrick). (i) Pile of mulga (Acacia aneura) branches used to trap soil and runoff on a bare slope in Australian rangeland, creating a new vegetation patch (photo by J. Ludwig).
arid and semiarid croplands, pastures, rangelands, and subhumid woodlands. It is due mainly to unsustainable human activities, such as overcultivation, fuel gathering, overgrazing by domestic animals, deforestation, and poor irrigation practices, and often is triggered or exacerbated by climate variability, mainly drought.

B. Reality vs Myths

The topic of desertification has evoked much controversy throughout the years in scientific, social, and political arenas (Table I). This controversy stems in part from the lack of scientific rigor in the preparation of many of the databases compiled in the 1970s, misinformation disseminated by various organizations, the lack of a consensus definition, and the alarmist tone connoted by the word itself, which suggests the action of deserts “moving” across the landscape, engulfing fertile lands and leaving starving people in their wake. These images have their roots in a series of papers written in the 1930s by the British forester Siebing, who used titles such as “The Encroaching Sahara” and “The Man-Made Desert in Africa.” Furthermore, a series of major droughts in sub-Saharan Africa in the 1960s and 1970s, combined with weak economies and civil strife, resulted in widespread famine and—along with reports that the Sahara desert was moving southward into the Sahel at a rapid rate (and which was referred to as “desertification”)—only served to reinforce this image. Subsequent evidence using satellite images of plant cover on the margins of the Sahara desert has clearly demonstrated that its boundary is actually dynamic (Fig. 2)—there are periods of both “expansion” and “retreat,” coinciding with the severity and duration of droughts. As findings such as these were reported in the scientific literature, news articles appeared with provocative titles such as “Threat of Encroaching Deserts May Be More Myth Than Fact” (New York Times, January 18, 1994) and “The Sahara is Not Marching Southward” (Science, July 31, 1998), illustrating that desertification remains a controversial topic.

There are several themes underlying misunderstandings of the concept, which have given rise to numerous myths throughout the years. Current estimates of the rates and the spatial extent of global desertification are based on limited data and thus are generally rough at best or (at worse) inaccurate (see Section IV, C). In some instances, gross exaggerations about the rates and extent of desertification have been made, e.g., areas that had been arid for centuries have been claimed to have become “desertified.” In other cases, observa-

---

**TABLE I**

Some of the Major Issues of the Desertification Debate

<table>
<thead>
<tr>
<th>Issue</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is land degradation?</td>
<td></td>
</tr>
<tr>
<td>What is the precise meaning of desertification?</td>
<td></td>
</tr>
<tr>
<td>Does desertification refer to the state of a system, or is it a process?</td>
<td></td>
</tr>
<tr>
<td>Can land degradation be quantitatively measured?</td>
<td></td>
</tr>
<tr>
<td>What is the amount of land affected or at risk?</td>
<td></td>
</tr>
<tr>
<td>How can a desert become “desertified”?</td>
<td></td>
</tr>
<tr>
<td>What are the causes of land degradation? For example, what is the relative importance of natural (climate-driven) vs man-made (e.g., due to overgrazing by domestic animals) processes?</td>
<td></td>
</tr>
<tr>
<td>Does climate variability cause desertification, or does desertification cause climate variability?</td>
<td></td>
</tr>
<tr>
<td>Are short-term ecosystem changes indicative of desertification? When are these changes considered “permanent”?</td>
<td></td>
</tr>
<tr>
<td>Does land degradation really matter?</td>
<td></td>
</tr>
<tr>
<td>Is land degradation reversible?</td>
<td></td>
</tr>
<tr>
<td>What are the relative roles of scientific vs grassroot efforts in abatement? In remediation?</td>
<td></td>
</tr>
</tbody>
</table>

6 Historically, desertification has been a contentious issue involving ecological, meteorological, and human dimensions but usually in isolation from each other.

---

**FIGURE 2** Extent of the Sahara Desert, shown in terms of annual deviations from average size defined by a 200-mm rainfall isohyet extending to the north of 25°N. Calculations are based on rainfall data collected at 141 locations throughout west Africa (Rain Obs), from METEOSAT satellite data (Rain Satellite), and from the normalized difference vegetation index (NDVI); a simple ratio between the red and near-infrared reflectance bands obtained from the NOAA polar-orbiting satellite. The NDVI correlates well with vegetation cover and biomass, and thus the interannual fluctuations of the desert boundary, as assessed from NDVI, generally mimic those of rainfall. This work by Sharon Nicholson and coworkers at Florida State University shows considerable agreement between estimates based on rainfall and those based on NDVI and demonstrates that there is no "march" of the desert over West Africa (plotted from data in Nicholson et al., 1998).
tions made on short-term ecosystem dynamics have been cited as evidence of desertification. With regard to short-term ecosystem dynamics, a key misunderstanding stems from the fact that drylands are fundamentally fragile ecosystems and, hence, highly susceptible to disturbance. Thus, the loss of vegetation cover (a short-term response to drought, e.g., that captured by satellite images in the Sahara) and soil degradation (a longer term response resulting from chronic and severe disturbances) are distinct, and not necessarily related, phenomena.

C. Dimensions of the Problem
The popular image of “creeping deserts moving like a series of waves of sand” continues to distract scientific and public attention from the real issue of desertification, which is indeed occurring globally, often far from the so-called true “deserts” such as the Sahara and Atacama. Desertification consists of three dimensions: (i) meteorological, (ii) ecological, and (iii) human (Fig. 3). Each of these components is complex, difficult to predict, and highly interdependent. A failure to recognize the simultaneous role of—and feedbacks between—these different components has led to many of the controversies and misconceptions alluded to previously. In this article, I provide a brief background to drylands, present regional examples of desertification, and discuss international efforts to “combat” desertification. Finally, I explore the concept of desertification from each of its critical dimensions.

II. WHAT ARE DRYLANDS?
A. Index of Aridity
As the term suggests, annual precipitation in drylands is quite low. Rainfall often consists of high-intensity events that occur over short periods of time, it is highly

---

**FIGURE 3** Desertification involves meteorological, ecological, and human dimensions. Past failures to recognize the interdependencies of these components has led to controversies and misconceptions about the concept.
variable throughout the year, and there is high overland runoff. Most, if not all, of the annual precipitation is lost from the system via evapotranspiration (evaporated from the soil or transpired by plants) rather than via drainage to ground water or surface runoff. This is due to the high evaporative demand of the atmosphere resulting from high air temperatures, low humidity, and abundant solar radiation, i.e., the potential for evapotranspiration is high. Therefore, drylands may be defined as regions where the ratio of mean annual precipitation (P) to mean annual potential evapotranspiration (PET), known as the index of aridity, is less than 0.65.

B. Global Extent and Classification

Drylands account for 47% of the global land mass (Table II). If we restrict the aridity index to the range of 0.05–0.65, drylands consist of arid, semiarid, and dry subhumid regions, which cover approximately 5.2 billion ha (Fig. 4). This definition of drylands omits 0.98 billion ha (~7.5% of the global land area) occupied by hyperarid regions (P/PET < 0.05), such as the Atacama, Arabian, and Sahara deserts. These hyperarid regions have very low rainfall and high rates of potential evaporation that restrict (or preclude) plant growth and settlement by human populations and are not included in this discussion of desertification.

The greatest concentrations of drylands are in Asia and Africa, which account for ~30% of the total land area of the world and 64% of all drylands. Hence, total dryland area elsewhere—when expressed on a global basis—appears relatively small (5% in Europe and 9% in South America). However, these numbers are deceptive. For example, although Australia contains only ~11% of the world’s drylands, they cover 75% of the Australian continent; similarly, both Europe and South America are about 30% drylands on a continental basis. In terms of a general land-use classification, arid (1569 million ha) and semiarid (2305 million ha) lands consist of rangelands and rain-fed farmlands, whereas the dry subhumid lands (1295 million ha) consist of woodlands, farmlands, and pasturelands.

![Aridity Zones](source: CRU/UEA, UNEP/GIRD)

**FIGURE 4** Global distribution of hyperarid, arid, semiarid, and dry subhumid areas (source: UNEP, 1997). See also color insert, Volume 1.
III. REGIONAL OVERVIEWS

The following brief overviews from select parts of the world illustrate the great diversity of issues and challenges presented by desertification. The causes and consequences vary from region to region, mainly in terms of localized intensity, as do programs to solve problems of land degradation, which vary regionally as a function of ecosystem characteristics, culture, economics, and political will. However, there exists some similarities. For example, land degradation resulting from human settlement in the rangelands of the United States, Argentina, and Australia share many of the same ecological consequences and social issues.

A. Africa: Extensive Problems

Desertification is perhaps most often associated with Africa, due largely to the great droughts and famines of the 1970s and 1980s. The drought in the Sudan–Sahel region in the early 1970s claimed more than 250,000 lives, left millions destitute, and resulted in mass migrations of rural people to urban areas in search of food, housing, and work. Similarly, in 1982–1986 the combined effects of severe drought and civil war killed more than 1 million Ethiopians. These high-profile events received much attention from the world media and served to cement the association of desertification with Africa. However, the true problems of desertification in Africa are much more complex than these tragic, but relatively short-lived, localized events.

More than one-third of the world's drylands are found in Africa (Table II) and one-third of Africa is considered to suffer from some degree of land degradation, including more than 70% of all agricultural lands. The African regions most affected by desertification include the northern portion of the continent, the Sudan–Sahelian (equatorial) region, and the Kalahari–Namib in the south. In the arid and semiarid zones of Algeria, Libya, Morocco, and Tunisia, overgrazing of rangelands has led to severe soil erosion and loss of vegetation. This loss of valuable topsoil, combined with a cycle of droughts, has resulted in declining production and intensified food insecurity. Rain-fed croplands in this region have been degraded by the use of inappropriate heavy machinery, salinization, and cultivation of dry, sandy soils. Woodlands have been cleared for fuel or by fires for agriculture and grazing. In the Sudan–Sahelian region the story is basically the same: fragile soils degraded through improper cultivation practices, deforestation, and overgrazing, leaving the land more vulnerable to drought and soil erosion. In southern Africa, overgrazing has led to severe land degradation and a decrease in productivity.

Michael Darkoh of the University of Botswana recently reviewed the large number of programs and organizations established in Africa during the past 30 years to help arrest and mitigate the effects of drought and

<table>
<thead>
<tr>
<th>Aridity Zone</th>
<th>All Drylands</th>
<th>% of World Drylands</th>
<th>% of Continental Land Area</th>
<th>% of Total Land Area of World</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperarid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semi-arid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dry sub-humid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>2,965.6</td>
<td>672.0</td>
<td>503.5</td>
<td>268.7</td>
</tr>
<tr>
<td>Asia</td>
<td>4,255.9</td>
<td>277.3</td>
<td>625.7</td>
<td>693.4</td>
</tr>
<tr>
<td>Australasia</td>
<td>882.2</td>
<td>0</td>
<td>303.0</td>
<td>309.0</td>
</tr>
<tr>
<td>Europe</td>
<td>950.5</td>
<td>0</td>
<td>11.0</td>
<td>105.2</td>
</tr>
<tr>
<td>North America</td>
<td>2,180.9</td>
<td>3.1</td>
<td>82.5</td>
<td>419.4</td>
</tr>
<tr>
<td>South America</td>
<td>1,767.5</td>
<td>25.7</td>
<td>44.5</td>
<td>264.5</td>
</tr>
<tr>
<td>Total</td>
<td>13,012.6</td>
<td>978.1</td>
<td>1569.2</td>
<td>2305.3</td>
</tr>
</tbody>
</table>

% of World Drylands: 16; % of Total Land Area of World: 7.5

*Based on UNEP (1997).

**P/PET ratio, where P is the mean annual precipitation and PET is the mean annual potential evapotranspiration.
desertification. The list is long, and these programs involve a plethora of action plans and activities and usually result in the formation of various types of organizations at the local, regional, and national levels (see Section IV). Unfortunately, many of these efforts have had limited success. The reasons for these failures vary from region to region but include such factors as lack of political will, shortage of funds, lack of participation of the people directly affected, inappropriate land-tenure policies, civil strife, misdirected research priorities, failure to include desertification control activities in national development plans, and a failure of governments to make desertification a priority. Poor countries have found it difficult to finance environmental programs aimed to protect natural resources, and often foreign financial and technical assistance has been used to support urban populations at the expense of rural ones. Of course, none of these problems and concerns are necessarily unique to Africa.

B. Mediterranean Region: Integrative Assessment

Land degradation is not a new issue in the Mediterranean. Humans have cut, grazed, and burned these lands for thousands of years, leaving virtually no natural vegetation and resulting in extensive soil erosion. However, rapidly changing social and economic conditions—along with the potential for climate change—pose many serious challenges to the Mediterranean region. In particular, the availability, distribution, and use of water underlines many of the contemporary themes of land degradation. Extensive forest fires throughout the Mediterranean in the early 1990s, which resulted in extensive soil erosion, runoff, and flooding, had a major impact on the public perception of land degradation. As a result, there has been a renewed emphasis on national programs of abatement in numerous Mediterranean countries.

The European Union (EU) recently established several major projects, such as Mediterranean Desertification and Land Use (MEDALUS). The aim of this project was to understand, predict, and mitigate the effects of desertification in Mediterranean countries of the EU, including the southern and eastern parts of the Iberian Peninsula of Spain, Mediterranean France, Italian Mezzogiorno, and almost all of Greece. This was motivated by many concerns, particularly increasing human populations and widespread land degradation (particularly soil erosion), air and water pollution, and water shortages. For example, as new EU economic policies are adopted, many pastoral and drylands farmers are being marginalized and moving to urban areas; social-economic conditions are rapidly changing, e.g., the rise of tourism, intensification of high-tech agriculture, and the shifting of populations to urban environments. Such changes have resulted in the abandonment of land for traditional agriculture and the increase in demand for water for urban expansion, tourism, and irrigation, resulting in increased land-use conflicts.

The Mediterranean, although small on a global scale, is a complex mosaic of landscapes, soils, different cultures, localized climates, and land-use histories. Although many of the problems grouped under the general heading of desertification are similar throughout the region (e.g., rainfall variability, growing urbanization, loss of farmland and agricultural potential, increased tourism and human population density, forest fires, soil erosion, and flooding), it is doubtful that any one approach for solving these problems will apply. However, programs such as MEDALUS have made significant advances in integrating desertification research across all levels of concern (Fig. 3), which may eventually provide general tools for linking science and policy in other regions of the world.

C. Australia: Landcare and Satellites

More than 70% of Australia is arid or semiarid. It is estimated that about 50% of these drylands are severely degraded or desertified to some extent. Most of this land has traditionally been used for pastoralism and the most obvious form of desertification is the encroachment of shrubs into grasslands and savannas. This encroachment is preceded by a loss of native perennial grasses and increased rates and extent of soil and wind erosion.

In contrast to the Mediterranean, which has experienced centuries of chronic land degradation, Europeans only began settling in Australia in the late 1880s. However, the landscape underwent rapid change over a very short period of time. Large herds of domestic livestock were put in the rangelands in order to maximize short-term economic profits but with little or no view toward long-term sustainability. These new pastoralists overstocked herds, suppressed fire (unaware of its importance in maintaining grasslands), and built fences that concentrated the impact of grazing animals. Exotic animals, such as the rabbit, were introduced with disastrous impact on native plant and animal communities. Furthermore, recurrent droughts tended to accelerate this chronic and systematic degradation of the land.

As in the Mediterranean, several instances of extreme drought, massive flooding, and soil erosion brought this
problem to the attention of the general public. Recent national programs of desertification management are relying on grassroots approaches to help resolve these problems, although this is complicated by the extreme isolation of many pastoralists. For example, federal funds are being used in Landcare Australia to encourage pastoralists and farmers to form local “landcare” groups and to employ a variety of proven techniques to restore degraded lands. The National Drought Alert Program employs satellite imagery to detect trends in greenness (similar to the NDVI index in Fig. 2) to forecast potential drought conditions in pastoral and farming regions; media alerts can then be issued to encourage appropriate strategies of land management, e.g., reduction of stocking rates.

D. China: Green Great Wall

China currently suffers from extensive land degradation, which has resulted in “colossal environmental detriment and economic loss” and threatens the lives of more than 400 million people (Zha and Gao, 1997). It is estimated that more than one-third of the total land area of China is desertified (at least to some extent), especially in the arid and semiarid northern areas of the country. Desertification has caused reduced soil fertility, degraded soil structure and massive erosion, and loss of native vegetation. Not surprisingly, China has experienced centuries of land degradation.

As in other areas of the world, the causes and consequences of desertification in China are varied but all can be attributed to various combinations of human and meteorological factors acting out in concert with localized ecological conditions (Fig. 3). There are some unique ecological settings in China that are conducive to desertification. Many northern regions of the country contain sandy, loose surficial sediment deposits that are readily transported. These regions are also characterized by the coincidence of droughts with the windy season. All of this combines to produce much wind erosion and transport of sand, particularly when the native vegetation is removed or decreased.

China has made an enormous effort to assess the causes and consequences of desertification, including programs to rehabilitate desertified land into productive uses. The Three North project (named after the three northern regions that suffer desertification) involves constructing a forested shelter belt that spans across 13 northern provinces. The hope is that this Green Great Wall, which is expected to be completed by 2025, will help stabilize sand dunes in the region, thus protecting cropland and pastures and abating soil erosion.

IV. “COMBATING” DESERTIFICATION

A. United Nations Efforts

Because of the global magnitude of desertification and its current and potential effects on human populations, the United Nations (UN) has taken a leading role in providing an international framework for studying and resolving this problem. Hence, much current desertification research (across a broad spectrum of disciplines) is conducted within the auspices of former and existing UN programs and related efforts. These programs are involved in educational and outreach efforts, basic research, and the coordination of disparate international and local projects. Although some of these programs have been subjected to criticism over the years, they provide a valuable context for addressing this problem. A brief overview is provided in the following sections.

1. Earth Summit

In June 1992, the UN convened the Earth Summit in Rio de Janeiro, Brazil, to address—among a variety of critical global issues—the problem of desertification. The Earth Summit was not the first UN attempt to tackle the problem of desertification. In the late 1970s, the term desertification was first brought to the public attention when the UN announced that this process was affecting more than 3500 million ha of land globally and that more than one-third of the terrestrial land surface of the earth was under similar risk. At the same time, the aforementioned great drought and famine of the Sudano-Sahel was in the public spotlight. The UN held a high-profile conference on desertification in Nairobi in 1977 at which a Plan of Action to Combat Desertification (PACD) was adopted “to curb this onslaught.” However, despite this and numerous other international efforts, by 1991 it was concluded that the PACD was not working and the problem of land degradation was actually intensifying, rather than abating, globally. Africa was identified as a particularly serious region of concern. As a result, desertification remained a major issue for the UN when it convened in Rio de Janeiro 15 years later.

The Earth Summit produced a significant, revamped program of desertification, including the establishment of a separate UN Convention on Desertification, an action framework designed to establish programs to counter the threat of dryland degradation throughout the world, particularly in arid, semiarid, and dry subhumid regions. The UN General Assembly was called on to establish an Intergovernmental Negotiating Com-
mittee to prepare a Convention to Combat Desertification (CCD) in "those countries experiencing serious drought and/or desertification, particularly in Africa." In December 1992, the general assembly adopted Resolution 47/188. The convention was adopted in Paris in June 1994 and on October 14, 1994, it was opened for signature; as of October 1999, the convention had been ratified or acceded by 159 countries (see http://www.unccd.de/ for updates). Meanwhile, many of the convention's provisions, particularly those relating to action programs in Africa, are being implemented.

2. Combating Desertification

What are the essential elements involved in combating desertification? The CCD adopted the following definition: Combating desertification includes activities that are part of the integrated development of land in arid, semiarid, and dry subhumid areas for sustainable development. These activities are specifically aimed at (i) prevention and/or reduction of land degradation, (ii) rehabilitation of partially degraded land, and (iii) reclamation of desertified land.

Although there are language and contextual differences, many of the goals and objectives of the CCD are not unlike those contained in the 1977 PACD. Hence, it is crucial to understand why the PACD failed. Three major reasons are most often cited: (i) Many processes and issues leading to land degradation (e.g., overcultivation of the land, deforestation, overgrazing by livestock, and poverty) have their roots in socioeconomic and political realms, and thus they are not amenable to technical or science "solutions" as proposed in the PACD; (ii) local populations were not included in drawing up solutions; and (iii) there was a failure to integrate the PACD into other existing development programs. Note that these problems reflect the failure to recognize the simultaneous importance of the three fundamental dimensions of desertification (Fig. 3).

Recognizing these shortcomings, the Rio conference supported a new, integrated approach to the desertification problem, emphasizing programs and strategies to promote sustainable development at the community level in an effort to overcome some of the failings of the previous PACD. Consequently, the CCD places considerable emphasis on the social dimensions of the problem and the role of local peoples and nongovernmental organizations in solving local problems. There is a conspicuous emphasis on the human dimensions (a "grassroot" effort), as evident in the prelude to the convention (Table III). This shift of emphasis from scientific solutions to social dimensions to solve prob-

| TABLE III |
| Excerpts from the Prologue of the 1994 United Nations Convention on Desertification, Classified as to Their Emphasis on Ecological, Meteorological, and Human Dimensions of the Problem |

<table>
<thead>
<tr>
<th>Parties to this convention</th>
<th>Ecological aspects</th>
<th>Meteorological aspects</th>
<th>Human dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afirm that human beings... are at the center of concern to combat desertification and mitigate the effects of drought.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are aware that arid, semiarid and dry subhumid areas together account for a significant proportion of the earth's land area and are the habitat and source of livelihood for a large segment of its population.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Note the high concentration of developing countries, notably the least developed countries... particularly... in Africa.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Note that desertification is caused by complex interactions among physical, biological, political, social, cultural, and economic factors.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are conscious that sustainable economic growth, social development, and poverty eradication are priorities of affected developing countries, particularly in Africa, and are essential to meeting sustainability objectives.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stress the important role played by women in regions affected by desertification and/or drought, particularly in rural areas of developing countries, and the importance of ensuring the full participation of both men and women at all levels in programs to combat desertification and mitigate the effects of drought.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Believe that strategies to combat desertification and mitigate the effects of drought will be most effective if they are based on sound systematic observation and rigorous scientific knowledge and if they are continuously reevaluated.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
lems of desertification is another reason it remains somewhat of a contentious issue (see Section I.B). There have also been criticisms that some developing countries are using the convention to assist in their economic development rather than combating desertification per se. Despite these concerns, the CCD is making significant progress and there is a clear recognition that regardless of emphasis, dryland degradation—given its ecological, climatic, economic, and social implications and complexities—will require contributions from a multitude of disciplines to understand and identify useful resolutions.

B. Defining Desertification

More than 100 formal definitions of desertification have been proposed in the literature. Although most restrict its usage for drylands per se (see Section II), these definitions cover a gamut of issues, involve a multitude of spatial and temporal scales of concerns, and often connote disparate meanings. As a result, some have called for eliminating the use of the term altogether. The main criticism is that it has been used in so many different ways throughout the years, and has been applied to such an extensive range of environmental problems, that it often lacks precise meaning. Does it refer to the state of a system, or is it a process? For example, is soil salinization a cause or a symptom? At what stage does land degradation constitute desertification? Issues such as these have contributed to the controversy surrounding the concept (see Table I). In general, the various definitions of desertification differ in their emphasis on the three distinct dimensions of the problem (Table IV).

<table>
<thead>
<tr>
<th>Definition</th>
<th>Ecological aspects</th>
<th>Meteorological aspects</th>
<th>Human dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The spread of desert-like conditions in arid or semiarid areas due to man's influence or to climate change.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Diminution or destruction of the biological potential of the land that can lead ultimately to desert-like conditions. It is an aspect of the widespread deterioration of ecosystems, and has diminished or destroyed the biological potential, i.e., plant and animal production, for multiple-use purposes at a time when increased productivity is needed to support growing populations in quest of development.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>An aspect of the widespread deterioration of ecosystems under the combined pressure of adverse and fluctuating climate and excessive exploitation.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>The process of environmental degradation in nonsandy areas where the fragile ecology is disturbed by excessive human activities.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>The impoverishment of terrestrial ecosystems under the impact of man … the process of deterioration … that can be measured by reduced productivity of desirable plants, undesirable alterations in the biomass and the diversity of the micro and macroflora and flora, accelerated soil deterioration, and increased hazards for human occupancy.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>The irreversible, sustained decline of the biological productivity of arid and semiarid land resulting from pressures caused both by people (e.g., increased population) and by abiotic factors (e.g., variable rainfall and long-term climate changes).</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>The spread of desert-like conditions of low biological productivity due to human impact under climatic variations.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>The expansion of desert-like conditions and landscapes to areas where they should not occur climatically.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Land degradation … resulting from adverse human impact.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Desertification, revealed by drought, is caused by human activities in which the carrying capacity of land is exceeded; it proceeds by exacerbated natural or man-induced mechanisms, and is made manifest by intricate steps of vegetation and soil deterioration which results, in human terms, in an irreversible decrease or destruction of the biological potential of the land and its ability to support population.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Land degradation in arid, semiarid, and dry subhumid areas resulting from various factors, including climatic variations and human activities.</td>
<td></td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>
The French forester Aubrèville is credited with being the first to coin the term desertification. Aubrèville's usage was in the context of observations he and colleagues made in 1949 of indiscriminate felling and burning of forest and woodlands in subhumid Africa and the accelerated soil erosion that resulted, i.e., the "creation of deserts." In 1977, the UN PACT defined desertification as the "diminution or destruction of the biological potential of the land [that] can lead ultimately to desert-like conditions." This definition, which identified human activities as the sole causal mechanism, was heavily criticized as being overly simplistic, alarmist, and ignoring the role of climatic factors. A more precise definition was needed—one that distinguished between true desertification and the shorter term phenomenon of cyclic oscillations of vegetation at desert fringes (i.e., desert "expansion and contraction" as observed by satellite data and which is related to natural climate fluctuations; Fig. 2).

The 1994 CCD addressed these concerns and adopted the following definition: land degradation in arid, semiarid, and dry subhumid areas resulting from various factors, including climatic variations and human activities. This definition distinguishes natural factors (i.e., climate) from anthropogenic ones, which is especially important since climate (both short and longer term) may play a substantial role in exacerbating the degradation of semiarid lands, particularly in marginal areas at the transitions between semiarid and arid ecosystems.

C. Global Extent of Desertification

1. UNEP Estimates

How widespread is land degradation in drylands? The most comprehensive study to date to estimate the global extent and distribution of desertification is that of the UN Environmental Programme (UNEP). Drylands (as in Table II) were subdivided into three dominant types of human land-use categories: irrigated agricultural cropland (146 million ha), rain-fed agricultural cropland (458 million ha), and rangelands (4556 million ha) (Fig. 5). Hence, from the point of view of agricultural land use, the overwhelming majority of drylands are rangelands (88%) and only 3% are irrigated croplands and 9% rain-fed cropland. It was estimated that 70% of all susceptible drylands suffered from some form of land degradation. By land-use category, this was 73% of all rangelands, 47% of rain-fed cropland, and 30% of irrigated cropland.

2. Limitations of Estimates

How accurate are these numbers? The UNEP numbers cited previously were compiled from a diversity of sources, including heavy reliance on the subjective

![Global Drylands](image)

![Irrigated Cropland](image)

![Rainfed Cropland](image)

![Rangelands](image)

**FIGURE 5** Estimates of dryland degradation (source: UNEP, 1992).
judgments of scientists and laypersons, and from surveys completed by local governments. Although they represent the most thorough estimates to date, they must be considered initial approximations. Although it is generally acknowledged that extensive areas of the earth have experienced varying degrees of chronic land degradation during the past century, there is a paucity of rigorous, consistent data. This is not surprising given that land degradation is a complex phenomenon (see Section V) and thus not amenable to practical and standardized methodologies for assessing, mapping, and quantifying on large scales. There are many perplexing challenges. For example, it is an exceptionally difficult task to quantify any single indicator from the suite of components that may reflect the rates and magnitudes of desertification, such as the degree of soil erosion, plant primary production, soil organic matter content, or socioeconomic conditions (Table V). This is true even at small spatial scales when each variable is analyzed in isolation, let alone consideration of key interactions with other variables. Many challenges remain. Most existing monitoring programs have been implemented for relatively short time frames, which makes it difficult to decipher short- vs long-term trends; meth-

<table>
<thead>
<tr>
<th>Indicator/process</th>
<th>Ecological aspects</th>
<th>Meteorological aspects</th>
<th>Human dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soils</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical degradation: salinization, contamination (heavy metals), acidification (aluminum, pH)</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical degradation: compaction, soil depth, infiltration, percolation, temperature, shear strength</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water erosion: runoff rates, soil loss, rill and gully density, sediment loading</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wind erosion: soil loss, dune formation</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soil organic matter: nutrient concentration, bulk density, soil texture</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depth to water table</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albedo (surface reflectance)</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio of soil carbon to nitrogen</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biological soil crusts: cyanobacteria, green algae, lichens, mosses</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Climate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precipitation: spatial distribution, amounts, intensity, frequency, variability, duration</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wind speed</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaporation; evapotranspiration</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vegetation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure: cover, density, spatial distribution, plant morphology, rooting depth</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agriculture: genetic diversity, species, harvest</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composition: species richness, diversity, endemism, exotic species, life forms</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Function: net primary productivity, standing biomass, leaf area index, growth rates, mortality, disease susceptibility</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Litter accumulation</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socioeconomics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human population density: changes in numbers, migrations</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income per capita</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Land-use changes: Ownership, clearing for firewood/cultivation</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water: availability, consumption, distribution, storage</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Grazing: pressure, wildlife restricted to game reserves, domestic cattle/sheep</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Tourism: game ranching, off-road vehicle use, road development</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Agroforestry practices: irrigation water quality, pesticides, fertilizer-use rates</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

* Indicators classified based on dominant driver (Fig. 3) of the processes, although all are highly interdependent.
ods and models that may work for one region may not be appropriate for another; and selecting appropriate indicators of desertification (e.g., to separate natural vs social factors) is difficult.

Current estimates of rates and extent of desertification are constantly being updated and improved as evident in the revised UN World Atlas of Desertification (UNEP, 1997). However, the process is difficult and slow. Although modern techniques of satellite-based remote sensing (combined with geographic information systems) provide some new opportunities for obtaining improved data—at least for some indicators of land degradation such as primary production and land-use changes—currently there are no integrative approaches or tools that can incorporate critical aspects of climate, ecology, and human dimensions into a single integrative framework.

V. CONCEPTUAL MODEL OF DESERTIFICATION

A. Ecological Aspects

1. Threshold of Response

Nearly all drylands are characterized by extreme year-to-year precipitation fluctuations; hence, it is often difficult to distinguish between short-term variability and long-term changes in ecosystem appearance as well as between temporary and permanent changes. Short-term variability in precipitation tends to affect the range and frequency of resource pulses, whereas long-term change alters the resource base: that is, the entire system moves beyond some threshold (dashed line in Fig. 3). Once this threshold has been crossed, it is unlikely that the system can be returned to its predisturbed state without expensive, and often extensive, remediation efforts (see Section V.D). All of this depends on the interactions of numerous climatic, edaphic, and biological factors in combination with the economic feasibility of rehabilitation.

I have stressed the inherent difficulties in teasing out underlying causal factors that give rise to short- vs long-term ecosystem dynamics in drylands. A recent case study from the Serengeti in Tanzania provides an excellent illustration (Morel, 1997). In the 1970s, there was much concern that the Serengeti was degrading and turning into a desert. The large bush and acacia forests that characterized the Serengeti were declining at an alarming rate. Elephants, which fed on these trees, were largely considered responsible and many believed that a culling program was the only way to save these “pristine” woodlands. In fact, these woodlands were not pristine but instead a relatively recent feature of the Serengeti reflecting a sequence of complex interactions between ecological factors (e.g., the life span of the acacias and the number of wildebeest), human dimensions (cattle grazing and fire setting), and meteorological factors (precipitation patterns) (Fig. 6).

Woodlands in the Serengeti started appearing after an outbreak of rinderpest disease in the 1890s that devastated wildebeest herds and local cattle populations and which led to mass human starvation. This decline in human populations meant a decrease in fires, which were set to create new pasture for cattle, and fewer fires meant higher establishment of acacia seedlings and the eventual establishment of a forest in the midst of grassland. By the 1920s and 1930s, the practice of setting fires was becoming reestablished as the human population recovered; however, the new fires were more frequent and hotter since conditions existed that favored grass growth—low herbivory (wildebeest herds were still low) and a relatively wet climate in the 1960s. These fires resulted in high mortality of acacias seedlings and, by the 1960s, the oldest trees (which have a life span of about 60–70 years) started to die off. However, concerns about the Serengeti turning into a desert were abated by the late 1990s as the Serengeti was experiencing a “return of the forest.” Heavy poaching of elephants, the recovery of the wildebeest populations, and fewer fires set by humans all favored high recruitment of tree seedlings.

This example illustrates several key elements of the conceptual model of desertification: (i) Short-term dynamics may not necessarily be indicative of long-term phenomena (i.e., system thresholds are extremely difficult to establish without long-term studies), (ii) natural ecosystems are complex and difficult to predict, and (iii) humans have major impacts on natural systems in often unknown and unpredictable ways. This serves to reinforce the view that land degradation is a complex phenomenon that frequently evades simple characterizations.

2. Ecological Dimensions of Land Degradation

Drylands typically have many distinguishing ecological characteristics that contribute to their susceptibility to disturbance and, ultimately, to desertification. The natural vegetation is composed of various mixtures of grasslands, shrublands, and savannas, with trees either scattered or concentrated along watercourses. Since vegetation cover is usually relatively sparse, much of
the soil is exposed directly to rain, overland flow, sunlight, and wind. Many dryland soils are sensitive to disturbances because they contain small amounts of organic matter and have low aggregate strength. Both tillage and grazing by domesticated animals can have profound effects—in a very short period of time—on these soils, including lowering their permeability to water (thus decreasing infiltration), disturbing their surface integrity (thus increasing susceptibility to erosion and sedimentation), and decreasing their quality (decreased nutrient status) for plant growth.

Although land degradation is often equated with soil degradation, it is a more general phenomenon that involves whole ecosystems. The UN CCD defines land degradation as the reduction or loss of the biological or economic productivity and complexity. This reduction in the complexity of the land is the major physical restructuring of a system that is symptomatic of land degradation. In drylands, this includes erosion and sedimentation by both water and wind, often resulting in a redistribution of topsoil, compacting of the soil, loss of soil silt fraction, dune formation, and arroyo cutting. There may be shifts in natural fire cycles with a disruption of biogeochemical cycling, including the redistribution of essential nutrients, decreased efficiency of nutrient cycling, and increased nutrient losses from the system. Both native perennial plants (cover and biomass) and many associated microbial and animal populations are reduced, whereas exotic species increase in dominance. All of this results in a decrease in the biological and hence economic potential of the land.

B. Meteorological Aspects

Williams and Balling (1996) concluded that substantial improvements have been made in our understanding of the causes of interannual variability in dryland climates, including the natural causes of droughts. For example, it is accepted that variations in annual precipitation levels are related to natural variations within global-scale climate systems. However, great uncertainties remain and the relationship between desertification and climate resembles the proverbial “chicken and egg” problem. To what extent does climate induce desertification? How do changes in land surface properties due to land degradation in turn affect climate? Since the two processes often work together, it is virtually impossible to separate the impact of drought from that of desertification.

Drylands are particularly vulnerable to climate variability, of which precipitation is the most important component. For example, a slight shift in seasonal pre-
cipitation and/or frequency of extreme rain events could potentially lead to the overexploitation of the meager resources of drylands and contribute to the further degradation of the very resource base on which human populations are so dependent. Preliminary studies with general circulation models (GCMs) in the 1980s projected that a doubling of atmospheric carbon dioxide (due to the rapidly expanding human population and associated activities) would result in lower precipitation, as well as shifts in the timing and frequency of rains, in the interior of large continents. Recent GCM studies also predict increases in rainfall intensity and longer dry periods in many dryland regions of the world. In the long run, global climate change may further exacerbate the already high natural variability of drylands, leading to permanent degradation of their productive potentials, particularly since there is a lack of "buffering" by large reserves of organic matter in the soils or in woody vegetation.

The array of impacts of climate on land and the implications of degraded land surface for the climate system are varied and complex (Fig. 3). Human activities impact surface characteristics (and hence albedo) and the atmospheric composition of dryland regions, including the breakdown of soil structure, reductions in soil moisture, increased surface runoff, changes in species composition and cover, increased dust in the atmosphere, and increases in aerosol and trace gas emissions from burning. In response to such human impacts, dryland climate is greatly influenced mainly via changes in energy balance. Changes in albedo affect the amount of solar radiation absorbed by the surface and changes in soil moisture levels affect the portion of energy used in evaporation and transpiration processes. Changes in surface roughness influence wind speeds and turbulence, which have a bearing on evapotranspiration. Finally, changes in atmospheric composition may affect atmospheric temperature profiles and thus influence the capacity to generate precipitation over land.

C. Human Dimensions

Rapidly changing social and economic conditions—along with the potential for climate change—pose serious challenges to many dryland regions of the world. Globally, there are differences in socioeconomic factors (e.g., human population growth rates) and biogeography (e.g., natural vegetation) that play a large role in the type of major human activities in any given area. Moreover, there are differences in how human interference is affecting biodiversity and ecosystem functioning in poor and richer countries. Taking these differences into account, there are approximately six major classes of human interference in drylands: (i) the loss of habitat; (ii) the fragmentation of crucial habitat; (iii) overexploitation (mainly overgrazing by domestic animals); (iv) the spread of exotic organisms (aliens); (v) air, soil, and water pollution; and (vi) climate change. Although it is difficult to generalize about the vulnerability of terrestrial ecosystems, some problems virtually exist everywhere (e.g., habitat loss, habitat fragmentation, and the direct and indirect effects of exotic species), whereas pollution and climate change tend to be more significant threats in richer, temperate-zone nations. Key ecosystem "goods and services" (e.g., food, construction materials, water purification, flood control, climate regulations, soil maintenance, carbon sequestration, nutrient recycling, wildlife habitat, erosion control, and tourism/recreation) are being seriously affected.

Given that all three of the developing regions of the world (Africa, Asia, and Latin America) have similar percentages of land degradation, and the highest levels of population growth, desertification is a growing problem with major ecological, economic, and social implications. However, to understand the human dimensions of desertification, it is important to compare and contrast different regions of the world to seek out generalities, particularly in richer vs poorer countries. As new economic policies are adopted, many pastoral and dryland farmers are marginalized and must move to urban areas; socioeconomic conditions are rapidly changing, e.g., the rise of tourism, intensification of high-tech agriculture, and the shifting of populations to urban environments. Often, such changes result in the abandonment of land for traditional agriculture and the increase in demand for water for urban expansion, tourism, and irrigation, resulting in increased land- and resource-use conflicts.

The human dimensions of desertification (Fig. 3) represent some of the most complicated and perplexing issues in identifying the causes and solutions to land degradation. For example, in large areas of Africa, civil strife and government policies are key factors influencing resources (land, water, and wildlife), land degradation, and food security. Civil strife displaces people, often to other marginal lands and in great numbers; their abandoned land is left unattended, indigenous management systems are lost, and they are forced to use methods of ranching and farming that are often not appropriate in these new areas. In some countries, government policies have encouraged planned human settlements in the wetter margins of arid and semiarid lands or nearer to water supplies. This has led to land-
use conflicts between agriculture, livestock, wildlife, and human settlements as a result of the intrusion of agriculture into lands traditionally used for domestic stock.

**D. Stepwise Degradation Model**

The conceptual model of desertification presented in Fig. 3 suggests that the process of chronic land degradation is directional—that is, it increases in severity until a threshold is eventually surpassed whereby the ecological state of the land (e.g., loss of soils and vegetation) is irreversible. An example is provided by a study of land degradation of rangelands in South Africa by Sue Milton and colleagues at the FitzPatrick Institute. Milton and colleagues showed that grazing-induced desertification is a stepwise phenomena and that the potential for recovery at any given step is related to the function of the affected component. Their model is summarized in Table VI, adapted to the scheme proposed in Fig. 3. In this particular case, the human dimensions component emphasizes management options and that part of the ecosystem specifically targeted for management.

At step zero, the biomass and composition of vegetation varies as a function of natural climatic cycles and stochastic events (e.g., fire, drought, and disease). An understanding of the relationships between these processes can be used effectively as a management tool. For example, livestock densities (secondary producers) can be manipulated depending on rangeland condi-

<table>
<thead>
<tr>
<th>Step No.</th>
<th>Ecological dimensions</th>
<th>Meteorological dimensions</th>
<th>Management option</th>
<th>Management level</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Biomass and composition of vegetation vary with climatic cycles and stochastic events</td>
<td>Natural climatic cycles and stochastic events</td>
<td>Adaptive management</td>
<td>Secondary producers</td>
</tr>
<tr>
<td>1</td>
<td>Herbivory reduces recruitment of palatable plants, allowing populations of unpalatable species to expand</td>
<td>Plant demography, increase in exotic or undesirable species</td>
<td>Strict grazing controls (rotation schemes, intensity, type of animals, browser: grazer ratios, etc.)</td>
<td>Secondary producers</td>
</tr>
<tr>
<td>2</td>
<td>Plant species that fail to recruit are lost, as are their specialized predators and symbionts</td>
<td>Loss of biodiversity (plants, animals, specialized predators, facilitators, symbionts), reduced primary and secondary productivity</td>
<td>Manage vegetation (e.g., seeding, plant removal), removal of livestock, culling abundant herbivores</td>
<td>Primary and secondary producers</td>
</tr>
<tr>
<td>3</td>
<td>Biomass and productivity of vegetation fluctuate as ephemerals benefit from loss of perennial cover</td>
<td>Perennial biomass reduced (short-lived plants and instability increase), resident birds decrease, mosaics increase</td>
<td>Manage soil cover (e.g., mulching, erosion barriers, roughen soil surface)</td>
<td>Physical environment</td>
</tr>
<tr>
<td>4</td>
<td>Denudation of all vegetation cover; changes in ecosystem structure and function</td>
<td>Bare ground; accelerated wind and water erosion, aridification, soil salinization</td>
<td><strong>?</strong></td>
<td><strong>?</strong></td>
</tr>
</tbody>
</table>

*Symptoms characterize the state of plant and animal assemblages, management options refer to actions to improve the condition of the range, and management level refers to the system (level of the food chain) in which management could be focused. The “system threshold” (Fig. 3) may be somewhere between steps 2 and 4 (based on conceptual model for stepwise degradation of arid and semiarid rangelands by Sue Milton and coworkers in South Africa).*
tions. The critical importance of this is illustrated by an example from Mexico. It was recently authorized that communally owned Mexican rangelands could be privatized in hopes of improving resource conditions and productivity. However, a follow-up analysis showed no differences between private and communal tenure systems, and instead the amount of annual precipitation was the most important factor explaining rangeland conditions (Coronado-Q., 1998).

The first step of degradation concerns changes in the age structure of the plant populations (Table VI). Those species most frequently defoliated by grazers fail to set seed, whereas toxic or distasteful plants, which are generally unaffected by grazers, are very successful. The type of management that might be used at this step is a function of the type of objective (e.g., game viewing and meat production) and vegetation type (annual or perennial grassland, shrubland, or savanna). Options to restore certain desirable forage species include shifting or varying the grazing season, stocking intensity, or animal type. The second step of degradation involves a decrease in biodiversity and both plant and animal productivity. There are fewer ecosystem “goods and services” (see Section V.C). Reversal of degradation at this stage is usually not cost-effective, particularly for arid regions in which it would involve removal of domestic livestock, culling of other abundant herbivores, and manipulating the vegetation (reseeding, herbicide treatments, bush cutting, etc.; see Figs. 1g and 1h). The third step involves processes associated with the reduction of perennial plant vegetation cover and an increase in ephemeral and weedy species. This includes accelerated wind and water erosion and various land surface impacts, such as increased albedo and surface temperatures, reduced soil moisture storage, and cloudiness. Rangelands in this condition are not profitable to ranchers and restoration must focus on the physical environment, e.g., reducing erosion, increasing water infiltration, protecting the soil surface from sun and frost, and creating microsites suitable for the establishment of perennial seedlings (Fig. 1). The final step in the degradation process is characterized by a complete loss of vegetation cover, accelerated erosion, and soil salinization—a true “human-made” desert. Such rangelands are usually abandoned due to the high costs of—and low probability of successful—restoration and rehabilitation.

VI. CONCLUSIONS

Desertification or land degradation is the loss of biological and economic productivity and biodiversity in arid and semiarid croplands, pastures, rangelands, and subhumid woodlands of the world. It is due mainly to nonsustainable human activities, such as overcultivation, overgrazing, deforestation, and poor irrigation practices, and is often triggered or exacerbated by climate variability, mainly drought. A new integrated assessment paradigm for rational decision making is needed to solve this problem. Such a paradigm must recognize the simultaneous roles of, and complex feedbacks between, the meteorological, ecological, and human dimensions of this problem. The failure to recognize and include the interdependencies of these dimensions has slowed progress in developing the necessary synthetic framework for tackling the enormous problem of dryland degradation. International researchers—representing a wide range of disciplines, such as ecology, atmospheric science, social sciences, policy, and integrated assessment—must work together to define new interdisciplinary approaches for addressing this pressing global problem.

Acknowledgments

The author thanks Roberto Dietz, Roberto Fernandez, Bob Gibbens, Jeff Herrick, John Ludwig, Klona Ogle, Sue Milten, Sharon Nicholson, Colin Saunders, John Tenhunen, and Jieqiu Wu for helpful input and acknowledges support from the Alexander von Humboldt-Stiftung during the preparation of the manuscript. This article is a contribution to the Carnegie Mellon Center for the Integrated Study of the Human Dimensions of Global Change (NSF SBR 95-21914).

See Also the Following Articles

AFRICA. ECOSYSTEMS OF • AUSTRALIA. ECOSYSTEMS OF • DEFORESTATION • DESERT ECOSYSTEMS • MEDITERRANEAN-CLIMATE ECOSYSTEMS

Bibliography


I. What Are Dormancy and Diapause?

II. Dormancy and Diapause as Adaptive Traits

III. Variation in Dormancy and Diapause

IV. Time Travelers: Dormancy and Diapause as "Migration from the Past"

V. The Storage Effect, Prolonged Dormancy and Diapause, and the Maintenance of Biodiversity

VI. Summary

GLOSSARY

bet-hedging strategy A trait of an organism, living in a variable environment, that leads to low variation in fitness. In general, such a trait provides an organism greater net fitness over a range of environmental conditions than would a trait specialized for any single environment. A bet-hedging trait is expected to evolve when the environment in which a species lives fluctuates over a fixed range of conditions that is sufficiently broad that fitness varies significantly, and when precisely which state the environment will take in the immediate future is unpredictable.

diapause A state of dormancy in some animals that is induced by a "token" environmental cue, such as day length. The token cue serves as a reliable indicator of a coming onset of harsh environmental conditions, but is not by itself harsh.

dormancy Any state of reduced metabolic activity of an organism. Typically, dormant organisms have associated characteristics such as cessation of development, the absence of reproduction, and enhanced resistance to harsh environmental conditions. Some disciplines have distinct meanings for dormancy that add further constraints to its meaning.

egg bank An accumulation of long-lived diapausing eggs (i.e., those eggs that persist in diapause for longer than a single growing season) of aquatic invertebrates in the sediments of marine or freshwater habitats.

seed bank An accumulation of dormant plant seeds that persist in dormancy for longer than a single growing season.

storage effect A general mechanism for the maintenance of biodiversity within a single habitat based on differences between competing species in their responses to environmental conditions. A resistant life-history stage can allow coexistence by the storage effect if each species reproduces successfully under the conditions favorable for that species and can survive through unfavorable periods (e.g., when a competing species dominates) in the resistant stage. Often, the resistant stage has prolonged dormancy. This mechanism can also serve to promote the maintenance of genetic diversity within a single population.

temporal dispersal The emergence of individuals from dormancy over a range of years (or other time interval), when those individuals entered dormancy in a
single year. Often, the years (or other time intervals) have different environmental qualities for growth and reproduction.

**Temporary Migration**
The avoidance of harsh environmental conditions in an environment by an individual organism that enters dormancy before conditions become harsh and emerges from dormancy when favorable conditions return.

**Dormancy Is Critical**
For surviving stressful environmental periods for a great many species. Some form of this trait is expressed by organisms ranging from bacteria to trees, from protozoa to vertebrates, and for a great diversity of organisms in between. In many cases, the period of dormancy lasts only as long, or slightly longer than, the duration of harsh conditions and simply determines the likelihood that a species can persist until the arrival of the next favorable period. For a variety of other species, however, dormancy can last for periods extending well beyond the typical duration of harsh environmental conditions. When this “prolonged dormancy” lasts long enough for multiple generations to have occurred in the active life history stage, powerful conditions exist for fostering the coexistence of multiple competing species within a community or genotypes within a species. Thus, both short-term and prolonged dormancy have important implications for understanding biodiversity.

**I. WHAT ARE DORMANCY AND DIAPAUSE?**

Dormancy is a very general term that encompasses a wide variety of different physiological states. The names applied to each of these states depend on the scientific traditions accompanying particular taxon-based disciplines. Only a small amount of this terminological diversity can be covered here, but it is important to recognize that the biological similarities (homologies) among the various physiological conditions called “dormancy” in different kinds of organisms are often questionable. In its most general sense, a dormant organism is simply one with a reduced metabolic rate. For bacteria, in which life cycle and cell cycle are synonymous, dormancy is defined as a temporary loss of the ability to reproduce (Henis, 1987), although many single-celled organisms also produce hardened cases at the time of reproduction (Henis, 1987), although many single-celled organisms also produce hardened cases at the time of dormancy (Whitton et al. and Bradbury in Henis, 1987). At the other end of the phylogenetic tree, vertebrate dormancy is most clearly expressed in “adaptive hypothermia” which includes both the daily torpor exhibited by a variety of small mammals and birds and the longer seasonal hibernation or aestivation seen in many taxa (Bartholomew, 1972). These latter states are often induced in part by “token” day-length cues that initiate physiological responses (e.g., fat storage), in advance of the onset of harsh conditions. In higher plants, dormancy is divided into “seasonal” (induced by external day-length cues that foreshadow a change in season) and “opportunistic” (imposed by direct exposure to harsh conditions) categories (Harper, 1977). Similarly, in arthropods, Tauber et al. (1986) distinguish between “aseasonal quiescence” (“a reversible state of suppressed metabolism,” similar to the “opportunistic dormancy” of plants) and “dormancy” (“a seasonally recurring period ... during which growth, development, and reproduction are suppressed”). By their definition, dormancy is then further divided as “diapause-mediated dormancy,” which is anticipatory, being induced by token cues (similar to seasonal dormancy of plants), and “non-diapause dormancy,” which is essentially seasonal quiescence. The common theme here is that organisms from bacteria to vertebrates exhibit dormancy as an adaptation for survival in temporally varying environments. Over a remarkable range of organismal complexity and phylogenetic origin, species have evolved broadly similar mechanisms for avoiding harsh conditions either by direct physiological responses to the imposition of harsh conditions or through distinct anticipatory responses to seasonally predictable changes. These mechanisms are generally lumped under the broad umbrella of dormancy. Depending on the organism and the environment in which it resides, the duration of dormancy or diapause can be brief or can last for extraordinary periods of time. For microbial cysts, dormant seeds, and diapausing eggs, there are reports of remarkable abilities to survive for centuries in dormancy. There are examples of bacteria becoming active after 200 years, of plant seeds germinating after thousands of years, and of crustacean eggs hatching after 300 years (Hairston et al., 1998). The diapausing eggs of Artemia (fairy shrimp from salt pans along San Francisco Bay) can survive more than 4 years of complete anoxia with no measurable amounts of stored carbohydrates used during this period and no detectable metabolic rate, even down to 1/50,000th of the aerobic respiration rate (Clegg and Jackson in Brendonck et al., 1998). Thus, at least some organisms can apparently survive as “living dead”—that...
is, in a state without detectable metabolism but still capable of becoming active when exposed to favorable environmental conditions.

Dormancy can occur at many points in an organism's life cycle, from embryo to juvenile and adult stages, depending on the species. Indeed, for some organisms, dormancy may occur at several different life history stages in a single population. Perennial plants may express both bud and seed dormancy (Harper, 1977). Some invertebrates, such as cyclopoid copepods, can enter diapause at several different stages ranging from immatures to adults (Elgmork in Alekseev and Fryer, 1996) and even as fertilized females (Alekseev and Fryer, 1996). Some fishes can aestivate as adults, whereas others such as the annual killifishes make diapausing eggs capable of surviving several years in desiccation. Some mammals can exhibit both adult hibernation and delayed implantation, a kind of embryonic diapause. In general, short-term dormancy (from a few weeks to a few years) is characteristic of species that are dormant as immature individuals or adults, whereas long-term dormancy (years to centuries) is only found in species that possess embryonic (i.e., seed or egg) dormancy. However, many seeds and eggs actually only remain dormant for brief time periods.

II. DORMANCY AND DIAPAUSE AS ADAPTIVE TRAITS

It is axiomatic that if individuals that possess dormancy within a population survive a harsh period, whereas others that lack dormancy do not survive, then genotypes expressing dormancy will be favored by natural selection. It is less obvious, however, precisely where the trade-off lies if the harsh environment is not fatal to all non-dormant individuals. Dormancy represents a reproductive delay (i.e., a lengthening of generation time) and therefore a reduction in potential rate of population increase. Cohen (1970), modeling insect diapause, showed that all individuals in a population should enter dormancy when the arithmetic mean of potential reproduction in the active stage drops below the harmonic mean of survival in the dormant stage. The ‘means’ in his analysis represent long-term probabilities of successful reproduction or survival. Likewise, Cohen (1966), this time taking plant seed dormancy as his inspiration, showed that the optimal fraction of seeds germinating in any given year depends on the probability that an individual will contribute to long-term fitness through reproduction as a growing plant compared with the contribution it would make by simply surviving another season as a dormant seed. His important contribution in these two papers was to show that dormancy is not simply a response to the certain extermination of active individuals in a highly seasonal environment but also that it is expected to evolve as a response to environmental uncertainty. It is because not all harsh periods are fatal, and not all growing seasons are bountiful, that variation in the expression of dormancy exists both within and among species.

III. VARIATION IN DORMANCY AND DIAPAUSE

There is often variation within a group of species living in a single environment either in whether any given species makes dormant eggs or seeds or in the time of year that these dormant stages are produced (see Harper (1977) for examples in plants and Tauber et al. (1986) for examples in insect diapause). This may be expected if different species react to the environment in different ways. One species may perceive a change as highly unfavorable, whereas another may be relatively immune to the change. There are many examples of populations in which only a fraction of individuals enter dormancy (Venable in Leck et al., 1989; Tauber et al., 1986). In these cases, it seems likely that a kind of “bet-hedging” strategy has evolved in response to variation through time, typically among years. In the models previously reviewed, dormancy is favored or not depending on the relative expectations of growth and reproduction in the active stage and survival in the dormant stage. If these expectations vary substantially over time, then the most successful phenotype may be one in which some individuals remain active (in case conditions remain favorable) and others enter dormancy (in case conditions become too harsh). For the same reason, there may be variation in the time of year that individuals within a population enter dormancy or diapause: In some years, the seasonal onset of harsh conditions occurs early, and in other years the onset occurs later. An example of this variation is a population of freshwater crustaceans (the calanoid copepod, *Daphnia sanguineus*, living in a small lake in Rhode Island) in which one subpopulation switches from making eggs that hatch immediately to making diapausing eggs significantly earlier in the season compared with the second subpopulation (Ellner et al. in Brendonck et al., 1998).
IV. TIME TRAVELERS: DORMANCY AND DIAPAUSE AS “MIGRATION FROM THE PAST”

Many organisms migrate seasonally away from environments that become uninhabitable at one time of year (usually winter). These include both vertebrates (birds, mammals, and fish) and invertebrates (e.g., some butterfly and lobster species). There is a large group of organisms, however, that are unable to undertake long-distance movement and therefore must survive the seasonally harsh environments in which they live. Microbes, fungi, plants, and a great many invertebrates have this life history constraint, as do some vertebrates that hibernate rather than migrate. In a very real sense, the organisms that cannot migrate spatially have evolved dormancy as a kind of temporal migration from one favorable season to the next. Migration is a directed movement between spatially separated habitats. Dormancy is especially analogous to migration when an organism not only begins dormancy before conditions become harsh but also emerges only when conditions become favorable again. For plant seeds, this is known as “predictive germination” (Venables in Leck et al., 1989), but many other types of dormancy termination are said to belong to a “seed bank” (Hastie et al., 1998). Organisms that have excellent spatial dispersal capabilities should not experience strong selection for long-term survival in dormancy, and vice versa, organisms with effective prolonged dormancy should not also evolve mechanisms for spatial dispersal. Consistent with this theory, there are significant negative relationships between spatial dispersal ability and seed dormancy in plants (Venables in Leck et al., 1989; Rees, 1993). Similar considerations may explain why winged insects, although often exhibiting single-season diapause (Tauber et al., 1986), very rarely possess diapause that extends over longer periods.

For a variety of species that make dormant propagules (i.e., either dormant seeds or diapausing eggs), there is evidence that these stages can facilitate both spatial and temporal dispersal. The resistant stages that permit dormant organisms to survive stressful periods in situ also make it possible for them to withstand unfavorable conditions during transport, whether by physical forces such as wind or water currents or as hitchhikers attached to more mobile animals. There is evidence, however, that trade-offs still exist in which function is most important for a given organism. Plant species with elaborations attached to their seeds that promote wind dispersal have shorter term dormancy than those that lack such structures (Rees, 1993). Limitations inherent in spatial dispersal for invertebrates living in lakes and other inland pools very likely explain why diapause and dormancy are much more prevalent among taxa living in this habitat than in related animals found in marine environments in which ocean currents provide for passive dispersal. In a survey of 167 species of crustaceans, Hairston and Cáceres (in Alekseev and Fryer, 1996) found that more than 53% of those living in inland water bodies possessed a long-lived diapausing stage, whereas less than 10% of marine species possessed this trait. More broadly, prolonged dormancy occurs much more frequently in invertebrate phyla with species that occur in nonmarine habitats than in invertebrate phyla that are exclusively marine.
Indeed, no exclusively marine phylum, and only 1 of 28 exclusively marine classes, is known to exhibit prolonged dormancy (Chesson, 1994). Several studies suggested that possession of prolonged dormancy or diapause may facilitate invasion of inland waters by aquatic invertebrates (Hairston and Caceres in Alekseev and Fryer, 1996; Caceres, 1997a; Hairston and Bohonak, 1998). Furthermore, those taxa that have persisted the longest in freshwater habitats, over geological time, are the ones with prolonged diapause (Alekseev and Starebogatov in Alekseev and Fryer, 1996).

V. THE STORAGE EFFECT, PROLONGED DORMANCY AND DIAPAUSE, AND THE MAINTENANCE OF BIODIVERSITY

Long-lived dormant stages that spread their germination or hatching over an extended period of time can play a major role in maintaining the coexistence of both species within a community and genotypes within a population. This is especially true when the environment varies through time with some species (or genotypes) favored under some environmental conditions and others favored under other conditions. For example, consider two competing plant species for which the environment varies through time so that each species has years in which it does well and others in which it does poorly. Also suppose that in those years when one species does well, the other does poorly. In the absence of seeds with prolonged dormancy, one species would be expected to be competitively superior on average, and thus to eliminate the other over time. However, if both species have long-lived seeds, then the one that is on average the poorer performer can persist by producing seeds when conditions are favorable and survive in dormancy through years of poor growth and reproduction. This scenario can be extended to communities of many species: If for each species there is an environmental condition in which it does better than its competitors, and if each has some seeds that germinate in each year, then all species can take advantage of years favorable to their own growth and reproduction while not suffering serious fitness loss in years that are unfavorable. Theoretical studies have shown that many species can coexist in this scenario, which is one example of a more general mechanism called the ‘storage effect’ (Chesson, 1994). Its importance in natural communities has been demonstrated for two species of competing freshwater crustaceans in the genus Daphnia that make long-lived diapausing eggs (Caceres, 1997b) and for desert annual plants with long-lived seeds (Philippi in Brudonck et al., 1998).

A similar process can maintain biodiversity of genotypes within a single species if individuals with different genotypes perform best in years of differing environmental quality (Hairston et al., 1996). The major difference between the theory for species competing within a community and that for genotypes competing within a population is evolution. In a community of competing species, we need only ask if those species present can coexist. With evolution, however, it is theoretically possible for natural selection to produce a single genotype that has a greater long-term fitness than any other genotype. The question now becomes one of whether other genotypes with lower long-term fitnesses can coexist in the presence of this most-fit type. The answer is ‘no’ if dormancy is absent or only short term but ‘yes’ if dormant stages are sufficiently long lived, and if natural selection fluctuates sufficiently to favor different genotypes at different times (Hairston et al., 1996). This has been shown to be the case for a population of freshwater crustacean copepods, Diaptomus sanguineus, with long-lived diapausing eggs that experience fluctuating natural selection due to year-to-year changes in fish predation pressure (Ellner et al. 1999).

The characteristics of organisms and their environments that make it likely that dormancy plays an important role in maintaining biodiversity are (i) an environment that fluctuates through conditions that favor a diversity of types of organisms, and (ii) dormancy or diapause of sufficient duration to span the time it takes for favorable conditions to recur for each organism type. A literature review on these topics by Hairston et al. (1996) shows that both of these conditions are quite common in nature for a wide variety of taxa and habitats.

VI. SUMMARY

Dormancy is a term that covers a variety of physiological states in a wide range of kinds of organisms. In each of its forms, dormancy plays a role in the ability of species to live where they do and is thus important in explaining what organisms are found in which environments. The presence of prolonged dormancy as a character in a taxonomic group of organisms can be critical to the ability of that group to invade and colonize new habitats. This is particularly true for habitats that vary greatly through time and for groups with restricted
abilities to disperse spatially. Finally, prolonged dormancy, when combined with a temporally varying environment, can be a significant factor in maintaining both genetic and community biodiversity within a habitat.

Acknowledgments

S. P. Ellner and C. E. Cáceres provided helpful comments. My research on diapause and dormancy is supported by grants from the U.S. National Science Foundation and the U.S. Environmental Protection Agency.

See Also the Following Articles

ADAPTATION • COMPETITION, INTERSPECIFIC

Bibliography


DifferEntiation

Nicholas H. Barton
Institute of Cell, Animal and Population Biology, University of Edinburgh, Scotland

I. Introduction
II. Neutral Evolution
III. Natural Selection
IV. Gene Flow
V. Ecological Differentiation
VI. Evolution Over Large Scales
VII. Conclusion

GLOSSARY

adaptive landscape A graph of average fitness against the state of the population, represented by allele frequencies or trait means.
allele A particular variant of a gene.
balancing selection Selection that acts to maintain two or more alleles in a stable equilibrium.
character displacement The displacement of characters caused by competition between species that live in the same place.
coalescent process The merging of lineages traced backward in time from a sample of genes.
fitness The total number of offspring produced over one generation.
frequency-dependent selection Selection in which fitnesses vary with the frequency of different genotypes.
gene flow The movement of genes from place to place.
genotype The particular combination of genes carried by an individual organism.
hybrid zone A narrow region in which genetically distinct populations meet, mate, and produce hybrids.
neutral Having no effect on fitness.
QTL Quantitative trait locus.
random genetic drift The random fluctuation in allele frequency caused by random variation in fitness between individuals.
stabilizing selection Selection on a continuously varying trait toward an intermediate optimum.
sympatric speciation The separation of a single population living in one place into two species.
transposable element A DNA sequence that can move from one place in the genome to another.

In Evolutionary Biology, the term Differentiation usually refers to the accumulation of genetic differences between populations or species. However, the term can be applied more broadly to the diversification of genes, organisms, and populations. The extraordinary conservation of the basic genetics and biochemistry across all living organisms implies that life evolved from one most recent common ancestor; the obstacles to the chemical evolution of the first replicator suggest that this ancestor itself evolved from a single origin—most likely a short self-replicating RNA sequence. Thus, we must explain how past and present biodiversity has differentiated from this simple beginning. We can ask how the first replicating molecules diversified to produce the many kinds of gene in pres-
ent-day organisms, how the first cells evolved into these diverse organisms, and how populations have come to contain so much genetic diversity. This chapter sets out to explain how elementary evolutionary and ecological processes lead to differentiation in both the narrow and broad senses.

I. INTRODUCTION

If we regard species as simply groups of organisms with similar morphology, then the differentiation of populations becomes equivalent to the process of speciation. However, under the biological species concept, speciation consists of the evolution of genetic differences that prevent successful interbreeding; under the phylogenetic species concept, speciation consists of the separation of lineages. Thus, at least in the sexually reproducing populations to which these ideas of species apply, population differentiation is distinct from speciation.

This chapter introduces successively more complex processes involved in differentiation. In the simplest kind of evolution, organisms and populations differentiate under mutation and random reproduction. If there is inherited variation in the rate of reproduction (i.e., in fitness), then natural selection will shape variation; populations in different environments will become differentiated in a straightforward way. This divergence may be opposed by the flow of genes from place to place, but will only be prevented over the smallest spatial scales. Just as diverse ecological niches are required if species are to coexist, so selection can maintain differentiation within a population only if genotypes exploit different limiting resources. Finally, the chapter examines how far these genetical and ecological processes can account for differentiation in the broadest sense and over the largest scales.

II. NEUTRAL EVOLUTION

Even if all copies of a gene are equivalent, some will by chance leave more descendants than others, leading to random drift of gene frequencies. As we trace their descendant lineages forward in time, some will become extinct, while others will grow in numbers. Eventually, it is inevitable that one gene will become the ancestor of the entire population (Fig. 1). Therefore, if we trace the ancestry of all present-day genes backward in time, the lineages will at some point coalesce in a single common ancestor. In an ideal population of $N$ diploid individuals, the chance that any two lineages coalesce is just

$$\frac{1}{2N}$$

per generation. A large sample of genes will take about $2N$ generations to coalesce into two ancestral lineages, and these two will then take, on average, another $2N$ generations to coalesce in one ancestor. Much attention has focused on identifying "mitochondrial Eve," the single female who (assuming strictly maternal inheritance) was the most recent common ancestor of all human mitochondria. The key issue here is when this ancestor lived: a recent Eve would imply a small effective size for the human population.

This coalescent process determines the simplest form of differentiation of genes and populations. Suppose that the gene in question consists of a long stretch of DNA, which is always passed on as a unit. Mutations will occur at a rate $\mu$, at random times and at random sites in the sequence; to a good approximation, mutations never occur at the same site twice (the infinite sites model). If we examine sequences that shared an ancestor a time $T$ in the past, they will on average differ by $2\mu T$ mutations. If these gene sequences come from two well-separated species, then we can date the divergence of those species by the steady ticking of a molecular clock. Indeed, it was the striking observation that any given protein accumulates mutations at a steady rate
that stimulated Kimura to develop the neutral theory of molecular evolution. Across a range of eukaryote lineages, sequences that have no effect on fitness mutate and diverge at a rate of $\sim 10^{-6}$ per base pair per year.

Variation within a population can be understood in a similar way. If two gene sequences are chosen at random, they will share an ancestor on average $T = 2N$ generations back; therefore, they will differ by, on average, $4N\mu$ mutations. This simple prediction is in qualitative agreement with observations on molecular genetic diversity within populations. However, variation in extremely large populations is not vastly greater than in much less abundant species; for example, about 5% of enzyme loci in humans are heterozygous, while about 25% are heterozygous in the bacterium Escherichia coli. There are two possible explanations for the weak relation between within-population diversity and census numbers. First, the effective population size may be much smaller than the actual census number, because of occasional bottlenecks and local extinctions or because of selection on linked genes. For example, in Drosophila melanogaster, $\sim 1\%$ of noncoding base pairs are heterozygous, which, with $\mu = 10^{-4}$ per base per year, implies that a random pair of bases share a common ancestor

$$T \sim \frac{0.01}{2\mu} = 5 \times 10^7$$

years back. Assuming 4 generations per year, this implies an effective population number of $\sim 10^6$ flies—many fewer than the census number. Second, variation in coding and regulatory regions of the DNA may be maintained by selection, regardless of population size (discussed later).

Neutral evolution does not lead to a homogeneous pattern of differentiation. Imagine that a process of random reproduction produces lineages as in Figure 1. (We might think of asexual organisms or of species originating and going extinct.) If morphology changes by random increments, then the variance among a group reflects the age of that group. Now, random geneologies typically have a hierarchical structure, with most family groups being closely related compared with the age of the whole. Morphology is therefore expected to fall into well-separated clusters (Fig. 2). Of course, the clustering that we see in the living world, which allows us to classify organisms into species, genera, and higher taxa, is not merely due to random reproduction. However, it is important to realize that what Darwin termed “descent with modification” tends naturally to produce clusters of distinct morphologies.

The recent availability of DNA-based markers, which for the most part can be presumed to have negligible effects on fitness, has transformed evolutionary biology. These provide a molecular clock against which to measure divergence times, an objective and quantitative means of reconstructing phylogeny, and a baseline level of differentiation within and between populations against which to measure other processes. Having established this straightforward benchmark, we can consider how selection, gene flow, and ecological interactions influence differentiation.

III. NATURAL SELECTION

A difference in relative fitness of $s$ (the selection coefficient) causes change over a time

$$\sim \frac{1}{s}$$

It dominates over random genetic drift if

$$s > \frac{1}{N}$$

and over mutation if $s > \mu$. Thus, even slight fitness differences of $s > 10^{-3}$ will be the predominant force
in a population of effective size larger than $10^5$
individuals and for genes consisting of, say, $10^6$ base
pairs, each with mutation rate $10^{-8}$ per generation.
Such minute fitness differences cannot be measured
directly. However, a multitude of studies show that
selection on morphological traits and on individual
genes can be strong ($s > 10\%$, say). Nevertheless,
we will see that this does not imply that differentiation
among populations is a simple consequence of select-
ion alone.

In this section, I consider the straightforward effects
of selection in the simplest case, where each genotype
(that is, each set of genes) has a fixed effect on fitness.
Variation within populations is then maintained by a
balance between mutation generating deleterious vari-
ants and selection eliminating them. (For the moment,
we leave aside cases in which heterozygotes are fitter
than either homozygote.) While mutation rates on indi-
vidual bases are extremely low, the net mutation rate
over the whole genome may be high, and so may sustain
substantial genetic variation. For example, Eyre-Walker
and Keightley (1999) used rates of DNA sequence diver-
gence among primates to show that the human lineage
has been subject to several deleterious mutations per
genome per generation. At the level of the whole organ-
ism, the variance of continuous traits such as growth
rate or skeletal proportions increases by roughly 0.1% every
generation as a result of new mutations; thus,
much of the response to long-term artificial selection
is due to mutation, rather than to standing variation.
The genetic basis of such mutations is beginning to be
understood. For example, Lai et al. (1994) found that
$\sim 10\%$ of the variance in bristle number in a natural
population of Drosophila melanogaster was due to the
insertion of transposable elements around the scabrous
gene; these insertions are clearly deleterious, since they
were all found to be of relatively recent origin, and
hence rapidly eliminated by selection.

There is a long-standing debate over the processes
that maintain genetic variation. While this is as yet
unresolved, it is at least plausible that most variation,
in both discrete genes and continuous traits, is due
to deleterious mutations. On this view, differentiation
within populations is merely a fluctuation around some
optimal state, while differentiation between populations
is a direct consequence of changes in that optimum. The
ability of artificial selection to rapidly change almost any
aspect of an organism was Darwin's strongest argument
for the efficiency of natural selection, and remains so
today. Almost all traits show heritable variation, and
therefore they respond to selection; what is remarkable
is that this response continues apparently without limit
(at least, over the span of human experiments). For
example, Weber (1992) selected on the angle between
two small veins on Drosophila wings. Lines selected in
opposite directions diverged steadily and differed by
20 standard deviations after 15 generations. We now
understand that at the genetic level, this divergence is
due to the establishment of individual alleles, either
present in the original base population or generated by
new mutations. Weber's lines, for example, differed by
at least 11 genes on one chromosome, each of small
effect.

There has been much dispute over whether adaptive
differences between populations and species are due to
the accumulation of minute differences or to a few genes
of large effect. Darwin emphasized the former view, on
the grounds that complex and finely adjusted adapta-
tions could only be built up by slight variations. In
contrast, the geneticists who rediscovered Mendelian
heredity emphasized the importance of major muta-
tions. After a century of research, the answer is only
beginning to become clear. On the one hand, sustained,
steady responses to selection must be based on multiple
genes of small effect, and genetic analysis of species
differences usually shows that many genes are involved
in maintaining their separation. On the other hand,
differentiation in particular traits is often based on one
or a few genes. Insecticide resistance usually evolves
through mutations in particular target genes (for exam-
ple, in the acetyl cholinesterase enzyme that is the target
of organophosphates). To take a natural example, the
striking differences between warning color patterns of
Heliconius butterflies are due to only a handful of genes.
Much effort is now being devoted to identification of the
quantitative trait loci (QTL) responsible for human
variation related to disease, for differences between vari-
esties of domesticated plants and animals, and for adap-
tive differences between species. Such methods do often
identify genes of major effect: indeed, the whole enter-
prise of genetic engineering rests on the possibility of
obtaining useful results by manipulating a few key
genes. However, there is a statistical bias toward finding
genes of large effect, and it is too early to know the
answer to this long-standing debate over the genetic
basis of adaptation.

Given that selection can rapidly adapt populations
to changing conditions, and given that selection is often
the strongest evolutionary force, one might suppose
that differentiation would simply reflect the optimal
genotype that is favored by the local environment. We
could then say rather little about what patterns of differ-
entiation to expect, since they would be the outcome
of the particular interactions between organism and
environment that determine reproductive success. However, matters are not so simple. There may be alternative combinations of genes that are locally optimal, and so populations may adapt to essentially the same conditions in different ways. Sewall Wright devised an influential way of thinking about this issue, the "adaptive landscape" (Fig. 3). This is a surface that represents the average fitness of a population, plotted against a set of axes that describe the state of a population (for example, gene frequencies or trait means). Under the simple assumptions of this section, natural selection pushes populations uphill on this landscape, toward a local "adaptive peak." In general, there will be many such peaks, and so populations will differentiate under selection even if they all experience the same environment.

Divergence may be observable at the phenotypic level or may be cryptic. For example, suppose that a trait is determined as the sum of effects of many genes, each carrying either + or - alleles. If selection favors a single intermediate optimum, any combination of genes that achieves that optimum can be established (+ - - - - - - - ...). Such cryptic divergence is revealed when crosses between populations produce progeny with reduced fitness and increased variance; over the largest scales, it is reflected by sibling species, which are morphologically indistinguishable and yet reproductively incompatible.

Divergence toward alternative "adaptive peaks" can occur in several ways. Random genetic drift in small populations may knock populations into the domain of attraction of different peaks. This process is especially likely if there are flat ridges on the adaptive landscape along which divergence is not opposed by natural selection. For example, if only the double recessive homozygote \( P_1P_1P_2P_2 \) is unfit, then populations initially containing \( Q_1 \) and \( Q_2 \) can fix either \( P_1 \) or \( P_2 \), with no effect on fitness. However, when crossed they have reduced fitness. Genetic analysis of hybrid inviability and sterility in Drosophila show that incompatibilities are largely recessive, which supports this model of quasi-neutral divergence. A different line of evidence comes from studies of functional RNA molecules (for example, the transfer RNAs that translate the genetic code). Many changes in individual bases cause no appreciable change in three-dimensional structure, so that extensive divergence can occur without disruption of function.

Random processes can drive divergence even in large populations. If adaptation depends on new mutations, then different mutations may arise first in different places. For example, humans have evolved resistance to malaria by a variety of different mutations in the \( \beta \)-hemoglobin molecule: the \( S \) allele in Africa, the \( C \) allele in Asia, and various thalassemias in the Mediterranean. These alternative alleles tend to exclude each other, so that chance differences in first occurrence of mutations leads to permanent divergence. Finally, divergence may be driven by fluctuations in selection. Suppose that the "adaptive landscape" fluctuates as conditions change but maintains overall distinct peaks. Then, populations may end up on different peaks, even if they experience the same mix of environments: the outcome depends on the particular sequence of environments that they encounter.

Although the "adaptive landscape" is a helpful metaphor, it gives a misleading picture of the geometry of evolution. Evolution can proceed in very many dimensions, rather than just the few that can be visualized by the human mind. This dimensionality is nicely illustrated by the model experiments of Eigen and coworkers, in which RNA molecules \( \sim 100 \) bases long are replicated by a polymerase derived from the QA virus. Under given conditions, some definite sequence (and its mutational variants) is picked out by selection from among \( 4^{100} \sim 10^{90} \) possibilities. This is possible because any sequence is connected to any other by no more than 100 mutational steps. At any stage, 300 variants are available from a single mutation, since each base can mutate to three alternatives. Provided that there is some sequence of steps that

---

**FIGURE 3** A simple "adaptive landscape" for two loci, each carrying two alleles. One allele at each locus is dominant, with frequency denoted by \( p \) and \( q \). The double dominant \( (P_1P_1P_2P_2) \) and double recessive \( (Q_1Q_2) \) genotypes are assumed to be most fit, with \( W_{PP} = 1 \), \( W_{QQ} = 0.8 \), \( W_{PQ} = 0.4 \) and \( W_{QQ} = 0.2 \).
leads uphill toward the optimal sequence, selection will effectively search from among the astronomical numbers of possibilities. This multidimensional view emphasizes that the outcome depends on which mutations are available: while evolution could take any of a large number of directions at any stage, this number is very much smaller than the total space of possibilities, and it sets a strong constraint.

IV. GENE FLOW

The most obvious impediment to differentiation is gene flow—that is, the mixing of genes from different populations caused by movement of the individuals that carry them. At least since the mid-19th century, there has been controversy over whether such mixing prevents differentiation. Darwin held that populations could adapt to gradually varying conditions across their range and that this could lead to the origin of new species. In contrast, Wagner emphasized the importance of barriers to migration as prerequisites for speciation. This view has been strongly propounded in more recent years by Ernst Mayr. However, both population genetic theory and observations of geographic divergence strongly support the Darwinian view, at least if one considers adaptations that can be built up by individual alleles or by continuous changes in quantitative traits.

Gene flow has straightforward and well-understood effects. The rate of gene flow into a single population is measured by m, the fraction of genes derived by immigration in each generation. Mixing then occurs over a timescale

\[ \frac{1}{m} \]

In a broadly continuous habitat, gene flow can be well approximated as a diffusion: in the same way that gas molecules diffuse as a result of random collisions, so genes diffuse through their habitat as a result of the random movement of the organisms that carry them. The rate of diffusion is measured by \( \sigma^2 \), the variance of distance between parent and offspring along any particular axis. Over a time \( T \), this diffusion causes genes to spread over a distance \( \sim \sqrt{2\sigma^2 T} \). For example, the annual grasshopper Chorthippus parallelus is divided into two subspecies C. p. parallelus and C. p. erythropus, which met in the Pyrenees after the last Ice Age, \( \sim 10^6 \) years ago. These grasshoppers move by about 30 meters in a year, so that \( \sigma^2 \sim 900 m^2 \) per year. Enzymes that are fixed for different alleles in the two subspecies mix over a region \( \sim 15 \) Km wide, which is of the same order as the distance predicted by simple diffusion, \( \sqrt{2\pi \sigma^2 T} \sim 7.5 \) Km (Hewitt, in Harrison, 1993).

Whether populations can adapt to local conditions, despite mixing, depends essentially on the relative rates of gene flow and selection. An allele with selective advantage \( s \) in the heterozygote can be established in a single population despite gene flow, provided that the rate of immigration of individuals carrying the alternative allele is lower than the rate of selection \( (m < s) \). If genes diffuse across a continuous habitat, then an allele can become established provided that it is favored in a region larger than some critical distance proportional to the characteristic scale \( \sigma/V \). New mutations have a chance of fixation approaching \( 2s \) if they arise within the appropriate region, but the probability falls to zero outside this region. Similar considerations apply to adaptations based on continuously varying traits. Suppose that fitness falls away in a Gaussian function with variance \( V_1 \) (i.e., stabilizing selection); the optimal trait value varies from place to place. The characteristic scale over which populations can respond to local conditions is now \( \sigma \sqrt{V/V_1} \), where \( V_1 \) is a measure of the strength of stabilizing selection, and \( V_2 \) is the additive genetic variance.

Because the dispersal range \( \sigma \) is typically much smaller than the species' range, these theoretical considerations imply that weak selection can allow a species to adapt to diverse local conditions. This can be seen directly. For example, the grass Agrostis tenuis has evolved tolerance to heavy metals on mines only a few meters across. Similarly, narrow clines \( \sim 10 \) Km wide separate races of Heliconius butterflies, which carry alternative warning color patterns (Mallet, in Harrison, 1993). Such examples leave little doubt that gene flow need not prevent divergence over very short scales.

If variation in selection can cause differentiation over local scales, why do species not fall apart into a myriad of separate forms? Mayr was led to argue that gene flow prevents differentiation of populations by his observation that species remain morphologically homogeneous over much of their contiguous range: it is this homogeneity that makes taxonomy practicable. How can we explain this apparent conflict with the potential power of selection to cause differentiation?

Much variation in protein or DNA sequence may have such small effects on fitness that selection is negligible relative to gene flow. Such neutral variation will differentiate as a result of random genetic drift. However, because drift is weak in even moderately abundant populations, weak gene flow homogenizes neutral vari-
ation. In a single population, the number of migrants, $Nm$, is crucial. The variance of allele frequency within the local population is

$$\frac{\bar{p}(1 - \bar{p})}{1 + 4Nm}$$

where $\bar{p}$ is the allele frequency in the source population; hence, if more than one migrant per generation enters, there will be little differentiation. This result carries over to populations spread over two dimensions, either in an array of subpopulations or in a continuous habitat. (In the latter case, the crucial parameter is $\rho \sigma^2$, where $\rho$ is the population density.) The effectiveness of gene flow in overcoming random drift has an important consequence: even if selection within local populations is ineffective relative to drift ($Ns << 1$), the cumulative effect of selection across the whole population of $N_e$ individuals dominates if $Nm > 1$. For many purposes, a population with $Nm > 1$ can be treated as a single gene pool.

Many surveys have been made of geographic variation in gene frequency, using both protein and DNA variation. The usual finding is that the variance in allele frequency between locations is small

$$\frac{\text{var}(\bar{p})}{\bar{p}(1 - \bar{p})} < 5\%.$$ 

This ratio is termed $F_{ST}$. This implies that if this variance is at an equilibrium between random drift and gene flow, the geographic pattern is large ($Nm > 1$). There are, of course, exceptions: for example, the salamander Ensatina eschscholtzii has exceptionally low dispersal, and correspondingly, $F_{ST} \sim 0.7$ (Wake and Yaney, 1986). However, the general conclusion is that gene flow is sufficient to homogenize differentiation generated by random drift.

Surveys of geographic variation often aim at estimating the number of migrants ($Nm$). However, even if large-scale differentiation is generated by random drift, its geographic pattern may be dominated by past history rather than by present-day rates of gene flow (see Whitlock and McCauley, 1998). If a species is distinguished from a large region (as happened in temperate regions during the Ice Ages) and then invades, the pattern of variation reflects the pattern of recolonization. In European human populations, Cavalli-Sforza and coworkers have reconstructed the history of population movements from the detailed genetic data that is available: many loci show traces of the introduction of distinct alleles as agriculture spread from the Near East.

When the genealogical relationships among DNA sequences can be inferred, historical events can be seen more directly. For example, Avise (1994) has identified in many species a cryptic subdivision that runs across the southwestern United States; this is a relict of a separation into different Pleistocene refugia. Such reconstructions should be based on several unlinked genetic loci; however, within a sexual species, different loci may have different histories.

Much molecular differentiation may have slight effects on fitness, allowing it to be used to reconstruct the history of gene movements. However, we cannot explain the apparent homogeneity of species simply by an absence of selection. There are many examples of geographic variation at enzyme loci that are maintained by selection for adaptation to different environments. Evidence of selection may come from direct measurement, from close correlation with environment, or from discordance in pattern between different loci. (For example, the geographic pattern at several enzymes in European humans differs significantly from the majority of genes, whose pattern reflects population history.) Morphological traits are usually strongly selected, and their geographic patterns therefore do not match those predicted from putatively neutral markers. (For a thorough comparison of morphological and molecular patterns of differentiation in Daphnia obtusa, see Lynch et al., 1999.) Because selection can cause rapid divergence, even in the face of gene flow, extinctions and recolonizations can only generate a transient homogeneity in selected traits. Thus, we expect to see morphological patterns that follow recent selection imposed by environmental features and cut across patterns at neutral markers that reflect population history. For example, Schneider et al. (1999) found that in the skinks of northeastern Australia, mitochondrial DNA and microsatellite markers tend to change along the Black Mountain Corridor, a biogeographic feature marking past range changes. In contrast, skeletal morphology changes in a quite different pattern, which parallels the transition from open to closed forest.

The evident power of selection to cause divergence remains hard to reconcile with the apparent similarity of species across large areas that Mayr and others have emphasized. The problem is close to that posed by examples of stasis in the fossil record, in which morphology remains similar over long times rather than large distances. One possibility is that in sexual species, different traits or genes can evolve independent spatial patterns, so that overall morphology is more stable than any one component and varies more smoothly. In any case, the supposed homogeneity may be exaggerated. Sharp transitions ('hybrid zones') do often occur,
prompting classification of the taxa on either side as species or subspecies. These may arise either when many traits respond to a sharp environmental discontinuity (as in the skink example used earlier) or when populations that have diverged to become partially incompatible are brought together by a range expansion. In the latter case, divergence remains permanently because the hybridizing populations are at two different "adaptive peaks."

V. ECOLOGICAL DIFFERENTIATION

A simple form of natural selection, in which each gene combination has a fixed fitness, can explain much of evolution: it accounts for differentiation between populations and species as a result of changing conditions, mutation, and random drift, and despite gene flow. However, at the very least absolute fitness (that is, the number of offspring) must decrease with density if the population is to remain bounded. We expect the relative fitnesses of different genotypes to change with their abundance in the population: selection is in general frequency dependent. If the reproductive success of an allele decreases as it becomes more common, then genetic diversity can be maintained. This is known as balancing selection.

Maintenance of genetic diversity by frequency-dependent selection arises from competition for different limiting resources, in just the same way as maintenance of species diversity. If the concept of a "limiting resource" is interpreted broadly, then all examples of balancing selection can be seen in this way. Evidence for a heterogeneity of resources, cited by Darwin and exploited by plant breeders, comes from the increased yield of a mixture of varieties compared with a pure stand. Here, the limiting resources may consist of different nutrients, though other mechanisms are possible. A classic example of frequency-dependent selection is Batesian mimicry, in which a palatable species avoids predation by mimicking a distasteful model species. As the mimic becomes common, predators fail to associate their pattern with unpalatability, and protection is lost. Batesian mimics such as swallowtail butterflies therefore evolve diverse mimetic patterns, each of which remains rare. The limiting resources here are the various different model species. In diploid species, variation can be maintained by heterozygote advantage. The classic example is the polymorphism for the 6-hemoglobin S allele, where the heterozygote gains resistance to malaria, while the homozygote suffers from sickle-cell anemia. The relative fitnesses of diploid individuals are fixed, but from the gene's point of view, fitness decreases with frequency: a common allele is more likely to find itself in a homozygote. The limiting resources here are the two kinds of allele with which each gene can find itself paired in a diploid.

The existence of heterogeneous limiting resources is, in a broad sense, necessary for maintaining diverse species and may be responsible for much genetic differentiation within species. However, we do not know how much genetic variation is maintained by balancing selection, as opposed to mutation and gene flow, and we do not know how far organisms respond to environmental heterogeneity by evolving diverse genotypes or diverse species. A tradeoff between variation within and between species is revealed by character displacement, where coexisting species tend to diverge in morphology and behavior. The Anolis lizards of the Caribbean provide a classic example. Where one species lives on an island, it has a broad distribution of body size; where two species share an island, they take up nonoverlapping distributions and so partition the available prey. A similar phenomenon can be seen where asexual clones compete with their sexually reproducing relatives. The freshwater fish Poeciliopsis monacha and P. lucida occasionally mate and produce a parthenogenetically reproducing clone. Vrijenhoek (1994) found that in streams with many different asexual clones, there tend to be fewer of the sexual progenitor species. This suggests that the space of available ecological niches can be filled either by a set of distinct asexual clones or by members of the genetically diverse sexual population.

If multiple forms, each exploiting a different resource, coexist in a sexual population, then matings between them may produce maladapted progeny. To make this point in another way, imagine that resource use depends on a set of continuously varying traits. If these traits are determined as the sum of effects of several genes, then sexual reproduction produces an approximately normal trait distribution. There is no reason why this should match the distribution of resources that are available. The problem can be avoided if the genetic system can somehow produce the appropriate distribution despite the random shuffling of genes that occurs with sexual reproduction. For example, swallowtail butterflies produce many distinct wing patterns, each mimicking a different distasteful model. These are determined by several genes, which are so tightly linked that they behave as a single genetic locus. If such genetic tricks do not evolve, then selection favors mating behaviors that cause like to mate with like. Ultimately, such selection could lead to complete reproductive isolation between the two forms. This splitting of a single population into two separate species in the absence of spatial separation is termed sympatric specia-
tion. It is almost impossible to demonstrate directly, since any present pattern of divergence might have originated with some spatial isolation. However, there are several plausible cases. For example, sticklebacks from the sea have colonized postglacial lakes within the past $\sim 10^5$ years and have repeatedly evolved distinct benthic and limnetic feeding morphs. These differ in morphology and also prefer to mate with their own kind; hybrids between them have reduced fitness in the wild (Hatfield and Schluter, 1999). A key question is how often ecological differentiation among genotypes evolves into differentiation between species in this way.

Darwin emphasized that natural selection is driven by competition for limited resources (the "struggle for existence"). Moreover, he argued that the lack of intermediate forms—the evident clustering of organisms that makes taxonomy possible—can be explained largely by the diversifying effects of competition. There is a "division of labor" among organisms, which (as with market economics) results from competition between individuals rather than any optimization of species or community productivity. Similar processes account for the differentiation of the $\sim 10^5$ different genes in complex organisms. Often, one gene may acquire multiple functions. For example, arginosuccinate lyase in birds acts as a lens crystallin as well as having a catalytic function. Similarly, most of the genes responsible for the establishment of segmentation in early Drosophila development are also involved much later in specifying various organs. If by chance a multifunctional gene duplicates, then each copy may specialize to become more efficient at one of the two original functions. The sequencing of complete genomes has made clear the extent of gene duplication and differentiation: most yeast genes are members of closely related families. There are close analogies between the division of labor among different species within an ecological community, different genotypes within a population, and different genes within an organism.

VI. EVOLUTION OVER LARGE SCALES

The processes of selection, mutation, and random drift that lead to differentiation are well understood, at least over the short time scales that are accessible to experiment and observation. How far do they account for the differentiation of genes and organisms over the whole span of evolution? First, consider the overall pattern. The first fossil evidence of multicellular organisms is found in the Ediacaran fauna, about 560 million years ago. An extraordinary diversity of animals with skeletons evolved soon after, at the beginning of the Cambrian, and included representatives of almost all present phyla. Molecular evidence suggests that the ancestral lineages of these phyla may trace back much further; nevertheless, there was a rapid diversification of body plans during this "Cambrian explosion." The subsequent fossil record shows an overall increase in diversity, as measured by numbers of species, but interspersed with occasional mass extinctions. Organismal complexity is hard to quantify, and trends in complexity are controversial: while the most complex organisms may become more complex as diversity increases, it is not clear that there is (or should be) any trend toward more elaborate organisms. Rates of morphological evolution can be quantified more readily, at least for hard parts. Morphology changes much more slowly over the long time span of the geological record than in current populations (Gingerich, 1983). Only weak selection pressures need be invoked to explain observed rates, and indeed it is surprising that morphology has often changed little over millions of years.

If we suppose that organisms differentiate primarily as a result of selection to fill different ecological niches (more precisely, to exploit different limiting resources), then we might imagine that these macroevolutionary patterns could be explained by diversification to fill a set number of preexisting niches. At equilibrium, there might be a balance between extinction and species formation, leaving many niches unfilled; nevertheless, diversity of species and their morphology would be expected to rise toward some definite value. This view is supported by statistical analysis of the fossil record, which suggests saturation of species numbers. Over shorter time scales, remote islands such as the Hawaiian archipelago are occupied by the descendants of a few colonizing species, which radiated to fill the roles normally occupied by other taxa. However, new organisms generate new niches. One host species is often plagued by several specialist parasites, which are restricted to that host by specific adaptations to evade its defenses. The evolution of novel features (multicellularity, bony skeletons, flight, etc.) are followed by a rapid diversification into novel ways of living. While the maintenance of diverse genotypes and species can be understood formally in terms of competition for distinct limiting resources, their number is not fixed and seems inherently unpredictable.

VII. CONCLUSION

The pattern of diversity that we see may not be entirely a matter of historical accident. Even though
there is no difficulty in explaining the rapid evolution of any one gene or trait, there are fundamental limits on the numbers of genes, the rates at which new genes are substituted for old, and the number of different genotypes that can be maintained within a population. The total mutation rate cannot greatly exceed one per generation, since most mutations are deleterious, and each deleterious mutation requires roughly one selective death to be eliminated. The human mutation rate approaches this upper bound, which may limit the number of functional genes that we carry. Each substitution of one allele by another requires that individuals carrying that allele leave more offspring than those that do not: Haldane (1957) showed that this implies a "cost of natural selection," which limits the rate at which new alleles are substituted. Again, this argument places a strong constraint on the number of adaptive changes that can have occurred during human evolution. Finally, the diversity that can be maintained within a finite population is limited: the maximum number of genes that vary independently can be no greater than the effective population size. These constraints give some hope that we can understand the overall features of organismal differentiation, even if the detailed pattern is a conglomeration of unique historical events.

See Also the Following Articles
ADAPTATION • ADAPTIVE RADIATION • DARWIN, CHARLES • GENETIC DIVERSITY • NUCLEIC ACID BIODIVERSITY • SPECIATION, PROCESS OF
I. Extinction as Counterpoint to Evolution
II. Misconceptions about Dinosaur Extinction
III. Current Theories
IV. Patterns of Vertebrate Extinction and Survival at the K/T Boundary
V. Testing K/T Extinction Theories Using the Vertebrate Fossil Record
VI. A Scenario for Dinosaur Extinction

Glossary

actinopterygia Ray-finned bony fishes.
amniota Taxon of vertebrates including mammals and reptiles (including birds).
campsosauridae Extinct taxon (Choristodera) of reptiles superficially resembling crocodilians
dinosauria Taxon of reptiles including Ornithischia (bird-hipped dinosaurs) and Saurischia (reptile-hipped dinosaurs, including birds).
elasmobranchii Sharks and their relatives, such as skates and rays.
K/T Abbreviation for Cretaceous/Tertiary, usually in reference to the K/T boundary.
multituberculata Extinct taxon of mammals superficially resembling rodents.
nonavian dinosaurs Dinosauria excluding birds.
squamata Lizards and snakes.

The loss of nonavian dinosaurs is the most widely recognized extinction either past or present. Although the dinosaurs that we commonly call birds survived and thrived, no nonavian dinosaurs are known beyond 65 mya (million years ago). Why all nonavian dinosaurs became extinct after some 145 my (million years) of evolutionary success as the world’s largest land vertebrates is a much debated topic. Here the most recent extinction theories are presented and tested against the known record of nonavian dinosaurs and their contemporaries.

I. Extinction as Counterpoint to Evolution

Extinction is commonly perceived as a rare, negative process. This is the case because much of the extinction that we recognize as occurring in historical times has been at the hands of humans, resulting in a loss to overall biodiversity. In this context, extinction certainly is a negative process, and unfortunately it is becoming more and more common.

Taking the much longer geological perspective, however, extinction is an integral part of the processes that have shaped the global biota ever since the appearance of life at least 3.5 billion years ago. Of all the species that have existed throughout Earth history, it is commonly estimated that over 90%, possibly over 99% of them have gone extinct. At first surprising, this figure is com-
prehensile when we realize that species have been constantly added through evolution and winnowed by extinction ever since life appeared. For the Phanerzoic era as a whole, there has been a general increase in species diversity. This trend can be further divided into several major increases in species diversity, along with at least one long interval when species diversity remained at a steady state. It is the balance between originations of new species and losses of existing species that maintains such a steady state of species numbers. Thus, extinction, considered a rare and negative process in human time frames and sensibilities, is an integral counterpoint to evolution. The vast majority of extant species, including Homo sapiens, would probably have never evolved had not many other species become extinct.

The vast majority of these extinctions, possibly as much as 95%, occurred as so-called normal or background extinctions. This is because rates and especially numbers of extinctions during most intervals of time do not stand out from extinctions in other such intervals. Although it is difficult to provide a specific figure because of vagaries of the fossil record, differences between environments, and differences between organisms, it can be safely said that normal extinction probably affects much less than 50% of existing species in a given time period. Similarly, there is no general agreement on the time frame over which these normal extinctions occur, but it is almost certainly for intervals of less than one million years. There are, however, at least five almost universally accepted intervals during the Phanerozoic when the numbers of extinctions far surpassed those of background extinctions. These are the so-called mass extinctions. During a mass extinction, well over 50%, in one case possibly reaching as high as 95%, of all extant species became extinct. This horrendously high level of extinction occurred at the end of the Permian period some 250 million years ago. All five known mass extinctions, from oldest to youngest, were in (or at the end of) the Late Ordovician, Late Devonian, Late Permian, Late Triassic, and Late Cretaceous. There is an unresolved debate as to whether these five mass extinctions represent a separate class of extinction from normal extinctions or instead form a continuum with normal extinctions in both rate and cause.

The most recent of the five mass extinctions near the end of the Late Cretaceous includes that of all the nonavian dinosaurs. The level of extinction for all species during the Cretaceous/Tertiary (or K/T) transition has been frequently estimated at about 75%, although no studies directly document this level. For backboned animals or vertebrates, the level of extinction hovers around 50% for species, but this is based on only one area, the Western Interior of North America.

II. MISCONCEPTIONS ABOUT DINOSAUR EXTINCTION

A. Quality of the Dinosaur Record

The belief that the nonavian dinosaurs disappeared from Earth instantaneously is not new, but the proposition in 1980 that an asteroid impact caused this very rapid decline and extinction gave new life to this idea. Some proponents of this theory have explicitly stated that these extinctions were essentially instantaneous around the world. Such explicit assertions about global records of nonavian dinosaurs are patently false.

All one need do is examine the fossil record of nonavian dinosaurs to see that this belief is unsupported. Figure 1 shows that for the Late Cretaceous as a whole, an interval of some 35 my, the record is reasonably good. Every continent, and even smaller areas such as New Zealand, produced Late Cretaceous nonavian dinosaurs. By comparison, Fig. 2 shows that the K/T record of nonavian dinosaurs is very incomplete. In only one region do we currently have a reasonably large sample of nonavian dinosaurs and contemporary vertebrates extending to or near the K/T boundary (and have a fossil record of similar quality above this boundary). This is in the Western Interior of North America,

FIGURE 1  Record of nonavian dinosaurs for the Late Cretaceous (modified from Archibald, 1996).
notably in the eastern part of Montana and into southern Canada. This region formed the eastern coast of a continental-sized landmass (Laramidia) that bordered the Pierre Seaway, a great inland sea that split North America in half in the Late Cretaceous. Although extensive, this region does not represent the entire globe regarding questions of extinction. A global record of latest Cretaceous nonavian dinosaurs may emerge in the next few years. Especially promising are new finds in several sedimentary basins in China, localities in central South America, and possibly sites on the Indian subcontinent. Until we have a more global record, arguments about the pace of dinosaur extinction on a global scale remain unsubstantiated speculation. We simply have no record of nonavian dinosaurs that permits us to show clearly whether dinosaur extinction was catastrophically fast or glacially slow. Rather, the data we do have are more regional in scope and only permit us to examine questions of the magnitude and selectivity of these extinctions, but nothing of its pace.

B. Ecological and Evolutionary Diversity among Dinosaurs

Although our perceptions of nonavian dinosaurs have changed during the past century, only now is a clearer picture emerging of them as living and breathing animals. Nevertheless, we must be cautious in assessing how these animals might have reacted to various stresses that could lead to extinction. Almost certainly the various groups of nonavian dinosaurs would not have reacted in monotonously similar ways to various ecological stresses. A short summary can provide a sense of this ecological and evolutionary diversity.

The two great clades of Dinosauria, Ornithischia and Saurischia, both appear in the Late Triassic, over 200 million years ago. Late Cretaceous Ornithischia ("bird-hipped" dinosaurs) includes the considerable diversity of ceratopsians ("horned" dinosaurs), hadrosaurians ("duck-billed" dinosaurs) and their relatives, the low-built ankylosaurs, and some lesser known clades. Saurischia most importantly includes sauropods (the largest dinosaurs) and Theropoda (all the medium- to large-sized carnivores, bipedal dinosaurs, plus Aves, or birds).

All of these dinosaurs were present in the Western Interior of North America leading up to the end of the Cretaceous. They ranged from small (about 1–2 kg) birds up to ponderous, many-tonned sauropods known from the southern areas of the continent. They also ranged from probable insectivores through carnivores, as well as a variety of herbivores. In short, although not great in numbers of species (e.g., compared to extant mammals), they were ecologically diverse. This means that any theory of extinction must explain how at least all the nonavian dinosaurs succumbed while many other vertebrates did not.

The cladogram in Fig. 3 shows the relationship of birds to other dinosaurs, as well as the relationships of other vertebrates known from the Late Cretaceous. The inclusion of birds within Dinosauria becomes stronger all the time with new discoveries, notably of feathered
nonavian dinosaurs, and ever more sophisticated phylogenetic analyses.

The most important concern regarding K/T boundary extinctions comes from workers that object to including birds within Dinosauroidea, even if they accept the very robust hypothesis of dinosaur–bird relationships as true. This objection comes from those who do not wish to accept the fact that dinosaurs survived the K/T boundary.

III. CURRENT THEORIES

It is estimated that more than 80 dinosaur extinction theories have been proposed, more than for any other group of animals. This list of theories ranges from the absurd (e.g., overhunting by aliens) to the highly reasonable. This section briefly outlines the three primary causes of extinction that are well enough articulated to be testable with the fossil record: marine regression (and habitat fragmentation), massive eruptions of flood basalts, and asteroid impact. In Section V these are tested using the vertebrate fossil record.

Marine regression refers to the draining of epicontinental seas. One of the greatest such regressions is recorded in rocks near the end of the Cretaceous period some 65 mya. Estimates suggest that 29 million km² of land were exposed during this interval (Fig. 4), more than twice the next largest such addition of land during the past 250 million years. This landmass is approximately the size of Africa. There were marked proximate effects of this regression—major loss of low coastal plain habitats, fragmentation of the remaining coastal plains, establishment of land bridges, extension of fresh-water systems, and climatic change with a general trend toward cooling on the newly emerged landmasses.

Massive eruptions of flood basalts called the Deccan Traps on the Indian subcontinent occurred over a longer interval than marine regression, perhaps as much as 4 million years surrounding the K/T boundary (Fig. 5). The volume of material estimated to have been erupted over this interval would cover both Alaska and Texas to a depth of 610 m. Proximate effects resulting from such massive volcanism have not been as well studied as for either marine regression or asteroid impact, but are argued to be similar to those from impact. Climatic changes caused by massive eruptions would have spanned a longer time than changes caused by impact.

The original asteroid impact theory argued that a 10-km-diameter asteroid struck Earth 65 mya, producing ejecta and a plume reaching far enough into the atmosphere to spread around the globe, blocking the Sun. The cessation of photosynthesis resulted in the death and extinction of many plants, the herbivores that fed upon them, and the carnivores that in turn fed upon the herbivores. The probable crater, named Chicxulub, has been located in the Gulf of Mexico near the tip of the Yucatan Peninsula (Fig. 6). At 170 km across, it is thought to be one of the largest such structures on Earth. Interestingly, the 100 km Popigai Crater in Siberia formed almost 36 mya at the same time as the 80-km-wide Chesapeake Bay impact crater. Thus we have 100- and 80-km-wide impact craters that formed at the same time (within present margins of error) that caused no identified increase in extinction levels. Further, there is as yet no clear evidence of an impact for the terminal Triassic and Permian mass extinctions. Nonetheless,

---

**FIGURE 4** Comparison of marine and nonmarine areas for the 250 million years. Note marked additions of nonmarine areas surrounding the Triassic/Jurassic and Cretaceous/Tertiary boundaries (modified from Smith et al., 1994).

**FIGURE 5** Geographic extent of latest Cretaceous and earliest Paleocene Deccan Traps (modified from Archibald, 1996).
two other important pieces of physical evidence supporting an impact are an increase in the element iridium at the K/T boundary and minerals, especially quartz grains, showing shocked lamellae in two directions. A high level of iridium at Earth’s surface and double lamellae are both more indicative of an impact than volcanism. Some of the more proximate effects of an asteroid impact that have been suggested are acid rain with a pH as low as battery acid, global wildfire consuming some 25% of aboveground biomass, sudden subfreezing temperatures lasting from 45 days to six months, tsunamis, and superhurricanes.

IV. PATTERNS OF VERTEBRATE EXTINCTION AND SURVIVAL AT THE K/T BOUNDARY

Although the vertebrate record of the K/T boundary is almost exclusively limited to the Western Interior, the uppermost Cretaceous Hell Creek Formation in eastern Montana has yielded a taxonomically rich sample of well over 100 vertebrate species, 107 of which are well enough known to be used in this analysis. There are 12 major vertebrate clades—5 species of elasmobranchs, 15 of actinopterygians, 8 of amphibians, 10 of multituberculates, 6 of placental mammals, 11 of marsupials, 17 of turdines, 10 of squamates, 1 of champsosaurs, 5 of crocodylians, 10 of ornithischians, and finally 9 species of saurischians (not including birds) (Table 1). Of these 107 vertebrate species, 49% (52 of 107) survived across the K/T boundary in the Western Interior. This is the minimum percentage survival of vertebrate species across the K/T boundary, because some of the very rare species may have survived undetected. Twenty of the 107 are quite rare species, represented by fewer than 50 identifiable specimens out of 150,000 specimens estimated to have been recovered from the Hell Creek Formation. Although any estimate is speculative, certainly some of these very rare species did survive. The extreme and very improbable scenario would be if all 20 survived, which would yield 67% (72 of 107 species) survival. This represents the extreme maximum percentage survival in the region. If I were to hazard an educated guess, I would say that no more than 60% of vertebrate species survived the K/T boundary.

When examined in greater detail, a very interesting pattern emerges within the record of survival and extinction for these Hell Creek Formation vertebrates. This is a pattern of differential species survival and extinction among the 12 major clades. Only five of the groups—elasmobranchs, marsupials, squamates, ornithischians, and saurischians—contributed to 75% of the extinctions. What do elasmobranchs, squamates, marsupials, and the two lineages of nonavian dinosaurs have in common in these faunas, other than that each suffered at least 70% or more species extinction at the K/T boundary in Western North America? If we are to understand causes of extinction at the K/T boundary we must explain this differential pattern of extinctions.

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Survival of Vertebrate Species Across the K/T Boundary</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Major clades from the Upper Cretaceous Hell Creek Formation, eastern Montana</th>
<th>Number and % survival</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELASMOBRANCHII</td>
<td>0/5 (0%)</td>
</tr>
<tr>
<td>ACTINOPTERYGH</td>
<td>9/15 (60%)</td>
</tr>
<tr>
<td>AMPHIBIA</td>
<td>8/8 (100%)</td>
</tr>
<tr>
<td>MAMMALIA</td>
<td>5/10 (50%)</td>
</tr>
<tr>
<td>LUCITUBERCULATA</td>
<td>6/6 (100%)</td>
</tr>
<tr>
<td>EUPHERIA (placental)</td>
<td>1/1 (100%)</td>
</tr>
<tr>
<td>METATHERIA</td>
<td>0/0 (0%)</td>
</tr>
<tr>
<td>REPTILIA</td>
<td>15/17 (88%)</td>
</tr>
<tr>
<td>TESTUDINES</td>
<td>3/10 (30%)</td>
</tr>
<tr>
<td>SQUAMATA</td>
<td>1/1 (100%)</td>
</tr>
<tr>
<td>CHORISTODERA (Champsosauridae)</td>
<td>4/5 (80%)</td>
</tr>
<tr>
<td>CROCODILIA</td>
<td>0/10 (0%)</td>
</tr>
<tr>
<td>ORNITHISCHIA</td>
<td>0/0 (0%)</td>
</tr>
<tr>
<td>SAURISCHIA</td>
<td>0/17 (49%)</td>
</tr>
</tbody>
</table>

| TOTAL NUMBER & % SURVIVAL | 22/107 (49%) |
Any theories about the cause(s) of extinction of dinosaurs and their contemporary vertebrates must be able to explain why elasmobranchs, squamates, marsupials, ornithischians, and saurischians suffered very high levels of or even total extinction, whereas actinopterygians, amphibians, multituberculates, placentals, turtles, champsosaurs, and crocodilians suffered 50% or often much less extinction.

There are at least four biological factors that may account for much of this pattern of differential species survival and extinction: freshwater vertebrates survived better than land-dwelling vertebrates, ectotherms did better than endotherms, small vertebrates did better than large species, and nonamniotes did better than amniotes. Except for possibly the last comparison, chi-square tests confirm the differences among these factors (Fig. 7). The biological significance of these differences is more difficult to demonstrate, but these comparisons emphasize that the K/T vertebrate pattern of extinction and survival was indeed differential.

V. TESTING K/T EXTINCTION THEORIES USING THE VERTEBRATE FOSSIL RECORD

Of the three testable hypotheses of K/T extinction—the impact, volcanism, and marine regression theories—only the impact and marine regression theories have been relatively thoroughly tested against the vertebrate fossil record. Proponents of the volcanism theory, however, suggest that many of the biotic responses to an impact would also be found with massive volcanism. One method to examine the efficacy of these theories is to examine them in the context of the known K/T vertebrate record, starting with the impact theory.

A. The Impact Theory

The original impact theory paper in 1980 still offers the basic mechanism of how such an impact might cause extinction among both animals and plants, including vertebrates. The impact would create a dust cloud enveloping the globe for a few months to a year, and darkness would shroud the world as long as the dust remained in the atmosphere. Photosynthesis in the sea and on land would cease. As the plants died or became dormant, herbivores would soon starve, followed by the carnivores. Anoxia would rise in the oceans. Some of the best physical evidence of such an impact is the discovery of anomalously high levels of the rare earth element iridium at the K/T boundary in strata around the world and the probable remains of an impact crater. Neither elevated levels of iridium nor an impact crater, however, is direct evidence for specific causes of extinction at the K/T boundary.

An incorrect assumption often made in testing the impact theory and its possible corollaries is that all major taxa show very high levels of extinction across the ecological spectrum on a global scale. As discussed earlier, for vertebrates, there is no K/T global record at the species level and there is a differential pattern of extinction. The impact-generated scenario of extremely
high levels of catastrophic extinction across most environments is so broad spectrum and tries to explain so much that it is difficult to test. The burden of proof for sweeping, catastrophic extinction scenarios rests with the proposers of the theory. The various corollaries of the impact theory, such as a sudden cold snap, highly acidic rain, or global wildfires, are more easily tested using the known K/T vertebrate record.

Although a short, sharp decrease in temperature was not emphasized in the originally proposed hypothesis, it soon became an important corollary of the impact theory. It is argued that if tremendous amounts of dust were injected into the atmosphere after a large impact, the darkness would not only suppress photosynthesis but would also produce extremely cold temperatures. This hypothesized condition has become known as “impact winter.” Accordingly, following a large impact, ocean temperatures would decrease only a few degrees because of the huge heat capacity of the oceans, but on the continents temperatures would be subfreezing from 45 days up to six months. The temperature would remain subfreezing for about twice the time of darkness caused by the dust.

If a suddenly induced, prolonged interval of subfreezing temperatures occurred in subtropical and tropical regions today, among vertebrates the ectothermic tetrapods would suffer most. Ectotherms, as the name suggests, heat or cool themselves using the environment. Endotherms such as mammals and birds generate their heat through metabolic activity. In endotherms, about 80% of the energy derived from food consumption goes toward thermoregulation. Fishes, which are by and large ectothermic, would be generally more tolerant of a severe temperature drop.

The northern limits of extant turtles and crocodilians today are controlled by temperature. These animals cannot tolerate freezing, and become sluggish or immobile at 10–15°C. Various amphibians and reptiles inhabit areas with low winter temperatures or severe drought, but they have evolved methods of torpor (estivation and hibernation) to survive. These are the exceptions, however, because species diversity for ectothermic tetrapods is far higher in warmer climates. More importantly, we should not assume that Late Cretaceous ectothermic tetrapods living in subtropical to tropical climates such as in eastern Montana were capable of extended torpor. Torpor is most often preceded by decreases in ambient temperature, changes in light regimes, and decreases in food supply. The ectotherms in eastern Montana could not have anticipated a short, sharp decrease in temperature. This is true even if the impact had occurred during a Northern Hemisphere winter, when temperatures would have been slightly lower. We must remember that this was a subtropical to tropical setting, and thus the extended, subfreezing temperatures advocated by proponents of this corollary would have been devastating to ectotherms even during a terminal Cretaceous winter in Montana.

Except for a 70% decline in lizards, the other ectothermic tetrapods (frogs, salamanders, turtles, amphibians, and crocodilians) did very well across the K/T boundary. The corollary of a sudden temperature decrease simply does not fit with the vertebrate data at the K/T boundary. A latest (but not terminal) Cretaceous vertebrate fauna from northern Alaska strengthens the evidence that a hypothesized sudden temperature drop was not a likely cause of K/T boundary extinctions. A comparison of Late Cretaceous vertebrate faunas from Alaska and eastern Montana reveals a striking difference. Although the Alaskan fauna is decidedly smaller and with fewer species than that from eastern Montana, both have euselachians, actinopterygians, dinosaurs, and mammals. The Alaskan fauna, however, completely lacks amphibians, turtles, lizards, chameleons, and crocodilians. These taxa comprise 41 of the 107 (38%) eastern Montana species. If even the fairly balmy temperature range of 2–5°C for Late Cretaceous Alaska was enough to exclude ectothermic tetrapods, a severe temperature drop to below subfreezing temperatures at the K/T boundary should have devastated the rich ectothermic tetrapod faunas at midlatitudes. Instead, these species flourished.

A second prominent corollary of the impact theory is the creation of highly acidic rain. The most commonly cited acids as products of an impact are nitric and sulfuric acid. It is argued that nitric acid would be produced by the combination of atmospheric nitrogen and oxygen as a result of the tremendous energy released by an impact. Sulfuric acid would be produced as large amounts of sulfur dioxide are vaporized from rock at the impact site. These acids would be precipitated in the form of rain.

Estimates of the pH of these acid rains vary, but go as low as 0.0–1.5. It is suggested that global effects could have caused the pH of near-surface marine water and freshwater to drop below 3. In today's environment, rain below a pH of 5.0 is considered unnaturally acidic. Rain as low as 2.4 has been recorded, but annual averages in areas affected by acid rain range from 3.8 to 4.4. Acid fog and clouds from 2.1 to 2.2 have been recorded in southern California and have been known to bathe spruce-fir forests in North Carolina. The biological consequences of such low pH values vary from one vertebrate group to another but are always detri-
mental. Aquatic species (fishes, amphibians, and some reptiles) are the first and most drastically affected, with those reproducing in water being the first to suffer. If pH drops lower than about 3.0, adults often die. The effects on aquatic vertebrates across the K/T boundary would have been very bad if a pH of 3.0 were reached and truly horrid if it hit 0.0 as suggested by some proponents. Some advocates of K/T acid rain argue that the surrounding soils or bedrock would have buffered the aquatic systems, even suggesting that limestone caves could have been important refugia for birds, mammals, amphibians, and small reptiles. The major problem with this scenario is that these kinds of buffering soils or bedrock and limestone caves did not exist in eastern Montana in the latest Cretaceous. From what we know of the present-day biota’s reaction to acid rain, aquatic animals should have been devastated by acid rain at the K/T boundary. Of all the aquatic species, only elasmobranchs show a drastic drop in eastern Montana. Thus, the likelihood of low pH rain is highly implausible.

A third corollary of the impact theory that receives various levels of support is global wildfire resulting from the aftermath of the impact. Soot and charcoal have been reported from several sites at the K/T boundary coincident with the enrichment of iridium noted earlier. It was argued that this pattern is unique and must come from the extremely rapid burning of vegetation equivalent to half of all present-day forests. Other scenarios posit that some 25% of the aboveground biomass burned at the end of the Cretaceous.

Such a global conflagration is really beyond our comprehension. To grasp the magnitude of this scenario, imagine one-quarter to half of all structures on the globe engulfed in flames within a matter of days or weeks. This still would be only a fraction of what is argued to have been burned at the K/T boundary. In such an apocalyptic global wildfire, much of the aboveground biomass all over the world would have been reduced to ashes. In freshwater, those plants and animals not boiled outright would have faced a rain of organic and inorganic matter unparalleled in human experience. These organisms would have literally choked on the debris or suffocated as oxygen was suddenly depleted with the tremendous influx of organic matter. The global wildfire scenario is so broad in its killing effects that it could not have been selective, yet as discussed earlier, the vertebrate pattern of extinction and survival is highly selective. Thus it is no surprise that this scenario of equal opportunity losers does not show any significant agreement with the pattern of vertebrate extinction and survival at the K/T boundary.

Not only is there almost no fossil evidence supporting global wildfire, but the physical basis for such an event is suspect. Proponents argue that a global charcoal and soot layer coincides with the K/T boundary, whose emplacement is measured in months. This also assumes that the sedimentary layer encasing the charcoal and soot was also deposited in only months. This is demonstrably not the case for at least one K/T section that continues to be cited in these studies—the Fish Clay of the Stevns Klint section on the coast of Denmark. The Fish Clay is a laterally discontinuous, complexly layered and burrowed clay reflecting the conditions at the time of its deposition. It is not the result of less than a year of deposition caused by an impact-induced global wildfire. Thus, carbon near the K/T boundary at Stevns Klint, as well as in other sections, is likely the result of much longer accumulation during normal sedimentation.

When all of the corollaries of an impact of an asteroid or comet are compared to the pattern of extinction and survival for vertebrates at the K/T boundary in eastern Montana, there is relatively poor agreement. Without special pleading, these corollaries as currently proposed are unlikely causes of vertebrate extinction. This does not mean that all corollaries of impact should be rejected. For example, the argument for a period of extended darkness and concomitant suppression of photosynthesis seems to be supported by the fossil record of lower and midlatitude plants. It is imperative, however, that researchers proposing the various corollaries of impact separate those that are supported by the vertebrate fossil record from those that are not.

B. The Volcanism Theory

Although some proponents of the impact theory do not agree, many advocates of the impact and volcanism theories feel that a number of the same physical events would have occurred at the K/T boundary if either extensive volcanism or an impact took place. These advocates also say that the biological results would be similar. In light of the foregoing discussion of how most of the corollaries of the impact theory do not test well against the vertebrate fossil record, the volcanism and impact theories are equally weak in their biological predictions. The major difference in these two theories lies in their timing. The impact theory measures most of the cataclysmic effects in months or years, with physical effects possibly lingering for a few hundred or a few thousand years, whereas the volcanism theory measures effects into the millions of years. The effects of many volcanic eruptions, such as that of Mt. St. Helens in
the United States, linger for a few months or a few years. Many other episodes of volcanism are very prolonged and these are flood basalt eruptions.

The best-known flood basalt eruption in the United States is the 16-million-year-old Columbia River flows in the Pacific Northwest. In the past 250 million years, arguably one of the biggest flood basalt eruptions occurred on the Indian subcontinent. This was occurring during (and is probably related to) the collision of the subcontinent with the remainder of Asia. The most obvious manifestation today is the tallest mountain range in the world, the Himalayas. As discussed earlier, these flood basalts, known as the Deccan Traps, cover an immense part of modern-day India and Pakistan. On the basis of radiometric dating, paleomagnetics, and vertebrate fossils, the bulk of the eruptions was centered around the K/T boundary during a reversal in Earth's magnetic poles known as 29R or 29 reversed. The number 29 represents the 29th reversal of the magnetic field counting backward from the present, which today has normal polarity by definition. The K/T boundary happens to fall in 29R.

What would the effect have been on the global biota if something of the magnitude of the Deccan Traps erupted for tens of thousands of years or longer? One of the greatest effects would have been to increase and maintain a much higher level of particulate matter in the atmosphere. Whether it would have caused warming through a greenhouse effect, cooling because of less light, or simply prettier sunsets is not certain. The amount of CO$_2$ pumped into the atmosphere by the eruptions may have been a boon for green plants that require CO$_2$ for photosynthesis, but a reduction in the light that reached Earth's surface because of particulate matter may have canceled the effects of increased CO$_2$. If I must choose, I would say that the effects of added particulate matter would prevail for no other reason than because they would linger longer after eruptions ceased, whereas the release of CO$_2$ would diminish much more rapidly after each eruption stopped.

If the latter scenario is correct, the longer-term effects over a million years or more would be to precipitate a global cooling. Most estimates suggest that the climate cooled regionally if not globally through the K/T transition. Because the time frame is moderately long, many species on land or in the sea, especially smaller ones, could have adapted to changes, whereas larger species such as dinosaurs may not have been fortunate. Although I doubt that the cooling across the K/T boundary can be said to be a cause of extinction for most species, it would have been an added stress.

A final possible long-term effect from eruption of the Deccan Traps is reduced hatching success for eggs of herbivorous dinosaurs. Volcanic activity can release elements such as selenium that are highly toxic to developing embryos. Increased levels of selenium in the eggshells of dinosaurs are reported from near the K/T boundary in southern France. The poisoning of eggs has also been suggested for dinosaur eggs near the K/T boundary in Nanxiong Basin in southeastern China.

C. The Marine Regression Theory

The final hypothesis that has been tested with the vertebrate fossil record is the marine regression/habitat fragmentation theory, or marine regression theory for short. Many areas of the modern terrestrial realm were repeatedly inundated by shallow epicontinental seas throughout geologic history. The term "epicontinental" refers to the occurrence of these very shallow seas on the continental shelves and platforms rather than in deep ocean basins. Epicontinental seas reached depths of only 1500 to 2000 ft, very shallow relative to most large present-day marine bodies. Epicontinental seas are almost nonexistent today, except for bodies of water such as Hudson Bay. It is known that during the Late Cretaceous, large areas of modern continents were submerged under warm, shallow, epicontinental seas.

It became clear only recently just how dramatic the regression of these seas was leading up to the K/T boundary. There is absolutely no mistaking that the K/T loss of shallow seas (or increase in nonmarine area) is greater than at any time in the past 250 million years. An equivalent of the land area of all of Africa, the second largest continent today, was added during this time. The second largest increase in continental area in the past 250 million years occurred across the Triassic/Jurassic boundary. Like the K/T transition, this is also during one of the five universally recognized mass extinctions during the Phanerozoic or last 550 million years.

Some of the most dramatic additions of nonmarine areas at or near the K/T boundary occurred in North America. Near the end of the Cretaceous, maximum marine transgression divided North America into two continents (Fig. 8). As regression continued until at or near the K/T boundary, coastal plains decreased in size and became fragmented; stream systems multiplied and lengthened; and as sea level fell, land connections were established or reestablished (Fig. 9).

The driving force for these repeated inundations or transgressions of the lower-lying portions of continents is still not fully understood. The general consensus is
Although these processes of marine transgression and regression were global in extent, a closer examination of North America is instructive because this is where we have the best vertebrate data at the K/T boundary. North America was split into two continents—a western continent, Laramidia, and an eastern continent, Appalachia—by the Pierre Seaway for almost 40 million years during the Late Cretaceous. Most of our latest Cretaceous vertebrate fossils come from the east coast of Laramidia. The west coast of Appalachia, as well as the eastern seaboard of Appalachia, have also produced some specimens.

In the last few million years of the Cretaceous the Pierre Seaway began to regress from both Laramidia and Appalachia. At or just shortly before the K/T boundary, the seaway reached its nadir. The positions of the receding coastlines both north and south have not been well established, but we know the southern coastline reached well into Texas. There is no question that there was a dramatic reduction in coastal plains. This is exactly the kind of environment from which we are sampling the last of the Late Cretaceous vertebrate community. A common refrain is that because the total amount of land increased with the regression, dinosaurs should have had more, not less, area and environments in which to live. We know with considerable certainty that dinosaurs did live in other environments such as the higher, drier Gobi Desert in Mongolia during part of the Late Cretaceous. At present, however, the only well-known vertebrate communities that preserve dinosaurs at the K/T boundary are coastal. Thus arguments about what dinosaurs and other vertebrates may or may not have done in other environments are moot. It is simply incorrect to say the dinosaurs and other vertebrates may have survived elsewhere, when we have little or no information about other environments. We must test our theories using the data we have, not what we wish we had.

The drastic reduction of coastal plains put tremendous pressure on some vertebrate species, especially large ones. Reduction of habitat, for example, in the Rift Valley System of modern-day East Africa, first affects larger vertebrates, especially mammals. In the shrinking coastal plains of latest Cretaceous North America, the equivalent large vertebrates first affected were the dinosaurs. An additional problem, whether in East Africa today or the coastal plains of latest Cretaceous North America, is the fragmenting of the remaining habitat. This present-day consequence as a result of human activity has become known as habitat fragmentation by ecologists. In larger, undisturbed habitats, animals (and plants) can move more freely from
one area to another. If the habitat is fragmented, although the amount of habitat may not have been greatly altered, it will reduce the flow of species from one fragment to another.

For some species, even seemingly small barriers such as two-lane roads can be insurmountable, and the consequences can be disastrous if viable populations cannot be maintained in the various fragments. This fragmentation can lead to extinctions. Barriers also arise in nature even among animals that would seem easily capable of dispersing. Although I have no doubt that such barriers often result in extinction, we usually see only what survives in the form of differences between closely related species. Small arboreal primates in the rain forests of both South America and Africa form small fragmented groups that are sometimes isolated from each other by rivers only tens of yards wide. Another example is the Kaibab squirrel on the North Rim of the Grand Canyon. Unlike its nearest relative, Abert’s squirrel, which is found on the south side of the Grand Canyon and in the western United States and Mexico, the Kaibab squirrel is restricted to an area of only 30 by 60 miles.

Fragmentation, in this case the development of the Grand Canyon, helped produce the differences, but the margin for survival for the Kaibab squirrel is slim.

The idea of habitat fragmentation extends not only to natural processes operating today, but to processes operating in the geological past. Although habitat fragmentation is not well understood among earth scientists, it is all too real for biologists studying the effects of human activity in present-day rain forests and in urban settings. Declines of bird and mammal populations have been well documented in the city of San Diego as urban development divides and isolates the remaining habitats in canyon areas. One would not expect that habitat fragmentation would be easily preserved in geological strata. Yet the forcing factor for fragmentation in the latest Cretaceous—marine regression—is a thoroughly documented fact during the waning years of the Late Cretaceous in North America. Globally, marine regression occurred within this same general time frame, although how close in time it occurred in various regions is a matter of debate.

Theory predicts that large species would be the most severely affected by habitat fragmentation, for the reasons discussed earlier. During the K/T transition in eastern Montana, only 8 of 30 large species survived and these were all or partially aquatic (two fishes, one turtle, one champsosaur, and four crocodilians), whereas all 22 large terrestrial species (and one aquatic species) became extinct (one turtle, one lizard, one crocodilian, and 19 dinosaurs). Thus, predictions from habitat fragmentation studies fit the observed data very well.

As noted in the preceding discussion, two other major physical events in addition to habitat fragmentation occurred with marine regression: stream systems multiplied and lengthened, and as sea level fell, land connections were established or reestablished. Only a few of the K/T boundary stream systems have been studied in detail in the Western Interior, and thus we do not know the exact drainage patterns for most latest Cretaceous and early Tertiary stream systems in the eastern part of Laramidia. Nevertheless, we are certain that as new land was added following marine regression in the early Tertiary, stream systems increased and lengthened. This process is another major corollary of marine regression. When freshwater habitats were increased following marine regression, most aquatic vertebrates did well except those with close marine ties—sharks and some bony fishes. Such fishes may need to spend at least a portion of their life in a marine environment, in some instances to reproduce. The major group most likely to suffer would have been the sharks and their relatives. In fact, all five species of sharks disappeared. It is not clear, however, whether these disappearances from the Western Interior were actually extinctions at the K/T boundary or whether they survived elsewhere in marine environments into the earliest Paleocene. The problem is that the definitively oldest marine sediments that postdate the K/T boundary in the Western Interior are at most late early Paleocene in age. This means that there is a gap in marine sedimentation in the Western Interior of possibly one million years or more. This pattern of disappearance and reappearance strongly suggests that as the Pierre Seaway regressed farther and farther away from eastern Montana, all sharks and relatives departed because connections to the sea became attenuated. New species of elasmobranchs did not occur in the area until a smaller marine transgression reached the Western Interior at or just before middle Paleocene times. This is known as the Cannonball Sea, which was a smaller seaway than the Pierre. The total disappearance of sharks and their relatives is the only prediction that can be made with any certainty as a result of the loss of marine connections and the lengthening of stream systems. The increase of stream systems helped to mitigate other stresses that may have been put on the freshwater system.

New land areas were exposed as the sea level lowered. In some cases this included the establishment or reestablishment of intercontinental connections. One such connection was the Bering Land Bridge joining
western North America and eastern Asia. At various times during the Late Cretaceous this bridge appeared and then disappeared. This is suggested by similarities in parts of the Late Cretaceous vertebrate faunas in Asia and North America, especially the better-studied turtles, dinosaurs, and mammals. Competition and extinction often result from biotic mixing, but predicting the fates of various taxonomic groups is usually not possible. An exception may have been the fate of marsupials in North America near the K/T boundary.

The oldest marsupials are known from approximately 100-million-year-old sites in western North America. By some 85 mya, we know of about 10 species of marsupial. This number rose to about 15 species and remained stable from around 75 mya until the K/T boundary some 65 mya, when it plummeted to a single species. These were all quite small mammals, from the size of a mouse up to a very well fed opossum or raccoon. Their teeth resembled those of modern opossums, with slicing crests and well-developed but relatively low cusps (compared to contemporary placental mammals) for poking holes in insect carapaces, seeds, or whatever they found. Most did not appear to be specialists on any particular food. With the reestablishment of the Bering Land bridge (or at least closer islands) near the K/T boundary, a new wave of placental mammals appeared in western North America. These mammals (traditionally known as condylarths) were the very early relatives of modern ungulates and whales. Their appearance in North America coincides with the very rapid decline of marsupials near the K/T boundary. Within a million years of the K/T boundary, 30 species of these archaic ungulates are known in North America, and their numbers kept on rising. Our best guess now is that the lineage that gave rise to these mammals first appeared in middle Asia between about 80–85 mya and reached North America near the K/T boundary. What is of interest is that the archaic ungulate invaders had dentitions very similar to those of the then contemporary marsupials and presumably are similar things. It seems more than coincidence that marsupials did well in North America for about 20 million years, only to nearly disappear with the appearance of the ungulate clade. It is ironic that both marsupials and ungulates were joint invaders of South America very soon after the K/T boundary. Their dentitions were already beginning to show differentiation, with the marsupials headed toward carnivory and ungulates headed toward herbivory. It shows what a little cooperation can do.

These various physical events accompanying marine regression fit very well the pattern of extinction and survival described earlier for the 107 vertebrate species from near the K/T boundary in eastern Montana. In fact, the patterns of extinction and survival for 11 of 12 of the major vertebrate groups agree quite well with predictions from the marine regression theory.

VI. A SCENARIO FOR DINOSAUR EXTINCTION

Some scenarios argue that dinosaur extinction was basically caused by one factor, such as the impact of an asteroid with its various related effects. The vertebrate fossil record, however, clearly points to multiple causes that may include volcanism, marine regression, and asteroid impact. Each cause might have been necessary for these extinctions to have occurred, but none was sufficient to cause the decidedly differential pattern of extinctions seen in the fossil record of vertebrates from western North America.

One must start at least 10 million years before the end of the Cretaceous in order to understand this biotic upheaval. Because the record of vertebrate change during this time is largely limited to the Western Interior, the scenario must be similarly limited in scope. How much more widely it can be applied remains a paleontological puzzle.

The scene some 75 million years ago is of an open plain with scattered trees on the eastern shore of Laramidia (Western Interior of North America). Rivers of moderate size meander across the landscape. In the distance, the plains are lapped gently by a shallow sea stretching to the eastern horizon. The scene is dominated by vast herds of several species of duck-billed and horned dinosaurs, much like the hoofed mammals that dominate Africa’s Serengeti Plain of today. Other ornithischian dinosaurs and the occasional meat-eating theropods cross the landscape. The streams are populated by numerous species of turtles, amphibians, crocodilians, and fish, including the occasional skate or shark swimming up from the nearby sea. It is daytime, mouse- and rat-sized mammals are asleep in their dens.

Fast forward to about 66 million years ago. The shallow seaway has begun to slip rapidly away to the south and east. As the exiting seaway reaches lower-lying, flatter terrain, the rate of exodus quickens, with the final stages of withdrawal occurring in at most tens of thousands of years. The great herds of duck-billed and horned dinosaurs have followed the retreating seas.
But as their remaining refugia of low coastal plains rapidly decrease, first one and then another of the species dwindles until the great herds are reduced to at most two or three remaining species, much like the herds of bison that once roamed North America. Dinosaurs, like large vertebrates everywhere and at every time, are the first to experience biotic stresses leading to their decline and disappearance. Thus, as with the bison, the slide to extinction of the few remaining species can occur quickly.

We do not know what was happening to vertebrates in more inland areas just before the K/T boundary, because few such areas are well preserved and none has been studied. The coastal plains dinosaurs certainly were capable of migrating from one shrinking coastal habitat to another, but finally even this could not stop further declines in population size—just as the relentless encroachments of increasing human populations are causing many biotas to shrink today. Other large vertebrates also suffered. The Komodo Dragon-sized lizards and the single exclusively terrestrial turtle, Boremys, experienced declines. Populations of smaller terrestrial vertebrates were also declining, but perhaps because they had shorter life spans and quicker turnover rates, they adapted more quickly to the environmental stresses caused by the loss and fragmentation of the coastal plains.

Marsupials had flourished for some 25 million years in North America. Newly emerging land bridges appeared as the seas retreated. Invaders appeared. In North America these were the diminutive archaic ungulates, probably from Asia. In the Western Interior, at least, they out-competed the marsupials for dwindling resources. In South America events were different. Both groups of mammals appeared in South America soon after the K/T boundary, but here they divided the guilds, with marsupials becoming the carnivores and the ungulates the herbivores. This coevolutionary arrangement lasted for almost 50 million years in South America, with only an infusion of rodents and primates from the outside world.

Unlike the terrestrial vertebrates, freshwater species faced far less stress, especially because the size of their habitat was at least holding its own as the lengthening streams followed the retreating seas. Yet not all aquatic vertebrates fared so well. With the loss of close ties to the seas in areas like eastern Montana, sharks and skates ventured into the rivers in the area less and less frequently, as the distance to the sea expanded from tens to thousands of miles, eventually reaching Texas to the south. Plants and nearshore species also showed added stresses as their respective habitats shrunk. Certainly, some species must have done fine as new habitats were formed as the seas regressed. As with vertebrates, however, we do not have any clear record of these environments away from the coastal areas.

Even before the seas began to retreat, the waxing and waning of the eruptions of the Deccan Traps added further stresses. One such stress was increase in particulate matter in the atmosphere that very slowly began to cool and dry some areas of the globe.

Suddenly, a literally Earth-shattering event magnified the differences between the have and have-not species. A 10-km-wide asteroid struck what today we call Yucatán. Material injected into the upper atmosphere formed a cover of darkness, blanketing the Sun to the point that photosynthesis ceased or diminished for many weeks, depending on where you were. The effects were especially acute at lower latitudes and closer to the impact, such as in North America. Plants unaccustomed to lower light regimes caused by seasonal changes in the Sun's position were especially hard hit. Higher-latitude plants accustomed to seasonally lower light regimes survived much better, as did the animals that fed upon them. The effects on higher-latitude plants and animals were tempered by which season they were experiencing when impact occurred. Extinction rates for coastal plants in North America soared because of the cumulative effects of continued habitat loss, drought, and loss of sunlight.

Except for the elasmobranchs, who had already departed or become extinct as the seas regressed, all ectothermic, aquatic vertebrate species (bony fishes, amphibians, turtles, champsosaurs, and crocodilians) weathered the impact quite well in their still flourishing freshwater habitats (80% survival, or 37 out of 46 species). Although some species can be resilient, they are also susceptible to certain kinds of shocks to their environment—just like the proverbial canary in the coal mine. If drastic temperature spikes, global wildfires, or battery acid rains had accompanied the Chixculub impact as is often argued, these species would have been decimated because they lacked the torpor mechanisms or places to hide from such hypothesized devastation.

With the added loss of more plant species, and the reduction of biomass that the impact brought in the already highly stressed terrestrial ecosystem, other vertebrate species rapidly succumbed. Most notable were the last of the large herbivorous nonavian dinosaurs. The remaining preaceous, nonavian dinosaurs followed very soon, with the larger species going first. In
some places on the globe the great creatures may have lingered a while longer, but finally, for the first time in more than 150 million years, no large land vertebrates graced Earth. The landscape was now open and waiting for evolution's next gambit—mammals.

See Also the Following Articles
EXTINCTION, CAUSES OF • MASS EXTINCTIONS, CONCEPT OF • MASS EXTINCTIONS, NOTABLE EXAMPLES OF

Bibliography
DISEASES, CONSERVATION AND

Sonia Altizer, Johannes Foufopoulos, and Andrea Gager
Princeton University

Parasites and Infectious Diseases can greatly influence the population dynamics and evolution of their hosts. Collectively, they comprise a major component of biodiversity (perhaps up to one-half of all living species) and are a ubiquitous component of natural ecosystems. Because of their effects on host abundance and diversity, diseases also pose an unusual problem for conservation biologists. From one perspective, they may have positive effects on biodiversity, facilitating species coexistence and maintaining genetic variation in natural populations. On the other hand, disease outbreaks can threaten biodiversity by catalyzing or accel-
erating species or population declines and extinctions, often in conjunction with human modifications of the environment or the accidental introduction of virulent pathogen strains. Therefore, tracking epidemics and measuring the effects of diseases in natural ecosystems are central to developing conservation strategies and managing biodiversity.

I. TYPES OF PARASITES AND CAUSES OF DISEASE

Parasites and pathogens obtain nutrients from one or a few individuals, usually to the detriment of their hosts. Disease refers to the signs and symptoms of infection, whereas parasites and pathogens are the disease-causing organisms. Infectious diseases can be caused either by microparasites (such as viruses, bacteria, protozoa, and fungi) or by larger macroparasites, including worms and arthropods. The distinction between macro- and microparasites is particularly useful to ecologists and epidemiologists because these groups differ in the degree of within-host replication, their ability to generate a lasting host immune response, and how they are quantified in natural populations (Table I).

Disease outbreaks require the presence of a susceptible host population, an infectious pathogen, and favorable environmental conditions; these three factors are often referred to as the disease triangle. Endemic parasites can often persist in host populations with few or no negative impacts, and infected hosts may never develop signs or symptoms of disease. Epidemics, on the other hand, are increases in the prevalence or intensity of infection and may result from the introduction of a new pathogen or from changing environmental conditions that favor pathogen transmission or increase host susceptibility. In some cases, these environmental modifications are human-induced changes that lead to host crowding, habitat degradation and host stress, or shifts in the geographic distribution of host, pathogen, or vector species. Diseases and parasites can also pose grave problems in captive populations in which plants and animals are maintained at high density, may be stressed, and are likely to be exposed to cross-species transmission.

II. BASIC EPIDEMIOLOGICAL PRINCIPLES

Since the pioneering work of Anderson and May in the late 1970s, ecologists have shown increasing interest in the effects of diseases at the population and community level. A general understanding of parasite ecology and epidemiology can illuminate problems in conservation biology ranging from detecting disease threats for endangered species to using parasites as control agents for invasive exotic species. Epidemiologists quantify and model disease spread over time and space to identify parameters that influence the prevalence and popula-

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>Characteristics, Examples, and Ecological Properties of Micro- and Macroparasites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representative taxa</td>
<td>Viruses, bacteria, protozoa, fungi, microsporidia</td>
</tr>
<tr>
<td>Helminths (e.g., nematodes, cestodes, and acanthocephalans), arthropods (e.g., mites, ticks, and lice)</td>
<td></td>
</tr>
<tr>
<td>Size and reproduction</td>
<td>Small, unicellular, short generation times within individual hosts</td>
</tr>
<tr>
<td>Large, multicellular, longer generation times, usually no direct replication within hosts</td>
<td></td>
</tr>
<tr>
<td>Transmission of infective stages</td>
<td>Transmission via direct contact (e.g., venereal and vertical), vectors, or contaminated in soil/water</td>
</tr>
<tr>
<td>Complex life cycles and intermediate hosts, vector transmission, or direct transmission (focal oral or host contact)</td>
<td></td>
</tr>
<tr>
<td>Effects on host immunity</td>
<td>Lasting host immunity develops quickly</td>
</tr>
<tr>
<td>Antigenic diversity of parasites usually too high for host to mount effective immune response</td>
<td></td>
</tr>
<tr>
<td>Effects on host fitness</td>
<td>Disease can be acute or chronic; may have strong effects on host survival or fecundity</td>
</tr>
<tr>
<td>Depends on the number of parasites within host (can affect mortality or fecundity, usually chronic infection with sublethal effects)</td>
<td></td>
</tr>
<tr>
<td>Quantification in host populations</td>
<td>Prevalence, seroprevalence, incidence</td>
</tr>
<tr>
<td>Prevalence, degree of aggregation in individual hosts</td>
<td></td>
</tr>
<tr>
<td>Frequency of epidemics</td>
<td>Common</td>
</tr>
<tr>
<td>Rare</td>
<td></td>
</tr>
</tbody>
</table>
tion-level effects of disease. Prevalence refers to the proportion of hosts harboring infection or expressing disease and allows for useful comparisons among geographic locations or over time. Parasite intensity refers to the average number of parasites within infected hosts and is important in quantifying the distribution and effects of macroparasites in host populations.

Models are used in epidemiology to detail how processes operating at the level of individuals (such as transmission, recovery, and death) translate into population-level phenomena. Mathematical models of host–parasite interactions differ from models of other antagonistic interactions (such as predators and prey) because diseases do not necessarily kill their hosts, and recovered hosts can develop a long-term immunity to reinfection. It is useful to address models designed for microvs macroparasites separately to simplify the analysis of disease dynamics and quantify their effects on host populations.

### A. Microparasites

Mathematical models for microparasites divide the host population into susceptible (S), infected (I), and recovered/immune (R) classes and track changes in the number of hosts within each category (Fig. 1). This type of compartment model (often called a SIR model) is commonly used in modeling directly transmitted microparasites, and has been developed and analyzed extensively through the work of Anderson and May (1991) and others from classical approaches of Kermack and McKendrick (1927). This model makes many assumptions, including that hosts are uninfected at birth, the disease does not affect host fecundity, and host

![SIR compartment model](image)

**Figure 1** SIR compartment model for directly transmitted microparasitic disease. This diagram depicts a population containing susceptible hosts (S), infected hosts that can transmit the parasite to others (I), and recovered or immune hosts that are no longer infected (R). Total host population size \( N = S + I + R \). Susceptible hosts arise from birth or immunity loss, at per capita rates \( a \) and \( y \), respectively. Individuals leave the susceptible class through natural mortality (rate \( b \)) or by acquiring the parasite at rate \( bI \) after encountering an infected host. Hosts leave the infected category through natural death or disease-induced mortality (rates \( b \) and \( a \), respectively) or through recovery (rate \( v \)) to an immune state.
populations are large enough that stochastic processes can be ignored. For diseases for which hosts do not acquire immunity to re-infection (e.g., many plant and insect pathogens and vertebrate diseases such as tuberculosis and brucellosis), the resistant class is eliminated and the equations simplify to a SI model. The simple SIR model in Fig. 1 is useful for parasites with density-dependent transmission, a mass-action process in which transmission increases directly with host population density.

Many complications can be added to the simple compartment model in Fig. 1. For example, a disease may reduce the fecundity of infected hosts or be associated with a long latent period. Age or social structure may complicate among-host contact rates and parasite transfer. In addition, the density-dependent mixing assumed by the equations in Fig. 1 is often inappropriate to describe the transmission dynamics of many pathogens. Other transmission modes, such as venereal, vector, or vertical transmission, can have profound effects on the invasion, persistence, and temporal dynamics of disease, and their consequences have been explored in many theoretical and comparative studies (Getz and Pickering, 1983; Thrall et al., 1993). Additional factors that increase the realism and complexity of host-parasite interactions lead to spatially explicit models to track disease epidemics and stochastic models that simulate disease dynamics in small populations.

Nevertheless, the basic SIR model gives rise to several key principles that characterize interactions between hosts and microparasites and have important consequences for diseases in wild populations. \( R_0 \) is the basic reproductive ratio of disease and specifies the conditions for which pathogens can increase in prevalence when the disease is initially rare. This is estimated by multiplying the expected number of new infections from a single infected host by the average duration of infectiousness. For the SIR model in Fig. 1,

\[
R_0 = \frac{\beta S}{\alpha + b + \nu}
\]  

This quantity must exceed 1.0 for the disease to invade. The form of \( R_0 \) suggests that diseases with high transmission rates (\( \beta \)), low virulence (\( \alpha \)), and low host recovery (\( \nu \)) are most likely to establish in host populations.

The establishment and persistence of many directly transmitted parasites can be related to a population density of susceptible hosts in which the parasite can just maintain itself, often called \( N_t \), or the disease’s threshold host population density. Assuming that the population is homogeneously mixed, this value is

\[
N_t = \frac{\alpha + b + \nu}{\beta}
\]

Pathogens that are highly virulent (high \( \alpha \)) or have lower transmission rates (low \( \beta \)) are likely to require much higher host densities for establishment than those that are highly transmissible and relatively benign. It is important to note that this host density threshold disappears if parasites are transmitted by a frequency-dependent rather than a density-dependent process (Getz and Pickering, 1983). Instead of ever-increasing with host density, frequency-dependent transmission remains constant or saturates at high host densities. This type of transmission characterizes vector- and sexually transmitted diseases, and in theory parasites with frequency-dependent transmission should be able to invade and persist at arbitrarily low host densities.

Microparasitic diseases will regulate or depress total host population size (\( N \)) through their effects on host survival or fecundity. If the intrinsic growth rate of uninfected hosts is \( r = a - b \), and the prevalence of disease is \( y = N/N_0 \) then the change in host population size can be written as

\[
\frac{dN}{dt} = (r - \alpha y)N
\]

Equation (3) suggests that the mechanism by which parasites regulate their hosts is through disease-induced mortality (\( \alpha \)) compensating for the host’s intrinsic growth rate. For example, with a SIR model modified to include density-dependent host regulation in the absence of disease, parasites can reduce host density below the disease-free carrying capacity if \( \beta \) is sufficiently high. If pathogens affect host mortality alone, those with intermediate virulence will depress host density to the greatest degree (Fig. 2) because extremely virulent parasites are likely to kill their hosts before new transmissions occur, and parasites with negligible effects on host survival usually have small population-level effects. In addition, disease prevalence at equilibrium (\( y = r/a \)) is inversely related to pathogenicity—that is, the more virulent a parasite, the lower its expected prevalence.

Analysis of simple microparasite models can generate important insights for managing and identifying disease risks in wild or captive populations. For example, they suggest that parasites that cause high host mortality in captive or laboratory populations are less...
likely to cause problems in low-density natural populations because infected hosts may die before transmitting the disease. The goal of vaccination and disease eradication programs is usually to drive the number of susceptible hosts in a population below threshold densities, because infected hosts may die before transmitting the disease. The basic reproductive ratio of macroparasites is the product of the mean number of new infections produced by a single adult parasite and the average life expectancy of adult and larval stages.

$$R_0 = \frac{\beta LH}{(\mu + b + a)(\gamma + BH)}$$
FIGURE 3. Schematic representation of host and parasite life cycle for macroparasitic infections. This diagram depicts a host population of size $H$, collectively harboring an adult parasite population of size $P$. As in the microparasite model, per capita host birth and death rates are denoted by $a$ and $b$, respectively. $\beta$ and $\alpha$ are the per capita host sterility and mortality rates induced by each individual parasite. Adult parasites give birth at rate $\lambda H$ and die at rates $\mu H$ (background mortality), $b H$ (host mortality), and $\alpha H$ (disease-induced mortality). Thus, the model assumes that when hosts die, so do their parasites. Free-living egg and larval stages die at rate $\gamma$, and transmission occurs when hosts eat these stages at rate $\delta$. This gives rise to new adult infections.

As with microparasites, Eq. (4) must exceed 1.0 for the parasite to establish when rare. Therefore, parasite invasion and persistence depend strongly on the rate of production of eggs or larval stages ($\lambda$), the rate at which parasites are consumed by hosts ($\delta$), and the survival of free-living infective stages ($1 - \mu$). The threshold host population necessary to sustain infection is

$$H_I = \frac{\gamma (\mu + b + \alpha)}{\delta (\lambda - (\mu + b + \alpha))}$$

Because the mortality rate of adult and larval parasites is likely to be low (larval parasites often have long-lived resistant stages and adult worms can live for years within their hosts), and the transmission rate of macroparasites is relatively high (especially when infective stages actively seek out their hosts), macroparasites will likely persist at lower host population densities than will directly transmitted microparasites.

For macroparasite models, the effects of parasite virulence on host thresholds and the ability of parasites to regulate hosts depend on the degree of parasite aggregation and the strength of within-host density dependence. In general, highly aggregated parasite distributions tend to stabilize host–macroparasite interactions, and random or regular parasite distributions tend to destabilize them. The dynamic properties of this model indicate that parasites regulate host density when they have strong negative effects on host survival, provided that within-host density-dependent effects are strong. Parasite-induced sterility, on the other hand, destabilizes the interaction and increases the probability of macroparasite-induced host population cycles.

Macroparasites can cause conservation problems for several reasons. Parasites adapted to low-density host populations often have high fecundity and long-lived infectious stages. When hosts are restricted in range or achieve high densities (as in zoos and small game parks), parasite burdens can rapidly increase. Macroparasites with negative effects on host fecundity and long-lived infectious stages are also likely to generate dramatic cycles in host abundance over time, causing problems for managing wild populations. Finally, these simple macro- and microparasite models can be used to quantify the impact of parasites and pathogens on host populations, provided that data are available on the effects of disease at the individual level and on patterns of infection in wild populations.

III. DISEASES AND THE MAINTENANCE OF BIODIVERSITY

A. Parasites and the Diversity of Ecological Communities

Although the risks they pose to endangered species are apparent, diseases and other natural enemies can play an important role in maintaining biodiversity and driving successional dynamics. The stability of ecological communities is thought to depend on the number of species they contain and the strengths of interactions between them (Elton, 1938; but see May, 1974). One way that diversity can be maintained at any given tro-
Fich level is through ecological tradeoffs between species with respect to factors such as competitive ability and resistance to pathogens and other natural enemies. Therefore, the effects of diseases on individuals or species may facilitate species coexistence and increase the diversity of ecological communities. For example, one explanation for why species diversity in the tropics is so high is because there are no seasonal climate changes to reduce pest populations, tropical species face greater pressure from parasites and diseases (Connell and Orias, 1964; Gilbert and Hubbell, 1996). Perpetual parasite populations may prevent any single species or group of species from dominating communities, allowing many species to coexist at relatively low densities. If more diverse and complex communities are more stable, then the goal of conservation should be not only to increase the abundance of rare or threatened species but also to maintain a complex and diverse array of interspecies interactions.

Clarifying the role of pathogens in community structure requires extending models of host–parasite dynamics to multi-species systems (Fig. 4). A range of mathematical models have been developed that explore the roles of pathogens in modifying interactions among species (Holt and Pickering, 1985). These systems fall into several classes: (i) a single pathogen species infecting multiple, potentially competing hosts; (ii) a host species with several natural enemies, at least one of which is a pathogen or parasite; or (iii) tri-trophic interactions in which the consumers (herbivores or predators) are susceptible to parasitism or disease. Theoretical studies of host–parasite interactions involving more than two species can yield complex dynamical outcomes and often support the proverb that "my enemy's enemy is also my friend" (Dobson and Crawley, 1994). In the case of host–host–pathogen systems, parasites can reverse the outcome of competition between species sharing the same resource if the dominant competitor is more susceptible to infection. Apparent competition is a related phenomenon whereby two or more hosts not directly competing for resources are affected by the same parasite to different degrees (Hudson and Greenman, 1998). In this case, generalist pathogens that are relatively benign to one host species may depress the density of other hosts for which they are more pathogenic. Diseases that are transmitted via vectors or long-lived infective stages are more likely to play a role in mediating host competition due to increased transmission opportunities among species. Although generalist pathogens can decrease diversity by threatening rare or endangered species, specialist diseases are more likely to facilitate coexistence by attacking common species that would otherwise dominate communities.

Several empirical observations illustrate the role of pathogens in determining plant and animal community structure and modifying ecosystems. Parasite-mediated competition has been shown to allow coexistence between Anolis lizard species on the Caribbean island of St. Maarten and in laboratory studies reversed the outcome of competition between two species of Tribolium beetles. Introduced pathogens can dramatically alter the species composition of ecological communities, particularly when they cause epidemics in dominant species. For example, fungal blights that removed chestnut trees from eastern deciduous forests of the United States and elm trees from much of western Europe favored less competitive species or those from earlier successional stages. Soil-borne diseases have been implicated in the rate and direction of succession in plant communities and in determining patterns of seedling recruitment in tropical rain forests. Fungal endophytes of grasses that limit seed production may also be responsible for enhancing the survival and dominance of tall fescue in open-field communities. Pathogens that attack herbivore species can in turn produce major changes in plant recruitment and abundance (Dobson and Crawley, 1994) and in the density of predators and other natural enemies (Dobson and Hudson, 1986). One example is furnished by the myxoma virus epidemic in rabbits in south England. Although a high abundance of rabbits in the mid-1950s prevented the regeneration of woody plants in grassland habitats, myxoma virus (introduced in the 1950s) led
to a scarcity of rabbits for the next 15 years. In areas where rabbit grazing had been so intense as to prevent tree establishment, a cohort of oak seedlings grew into forests following the initial epidemic (Dobson and Crawley, 1994). Similar cases can be found in east South Africa, where viral and bacterial pathogens of herbivores have caused changes in herbivore abundance and radically altered the structure of plant communities. Although these examples are cases in which diseases have generated striking changes in community structure, a vast majority of host–parasite interactions are likely to yield more subtle effects on host growth, survival, or fecundity that underlie the persistence of species and the assembly of ecological communities. When preserving intact and functioning ecosystems is the goal of conservation strategies, attention must focus on the structure of parasitic communities and the role they play in maintaining biodiversity.

B. Pathogens as Biological Control Agents

Another more optimistic side to parasites and pathogens is that they can be used as biological control agents to regulate the densities of introduced or pest species that pose major threats to endangered species or ecosystems. Alternate strategies for the eradication of exotic organisms usually involve trapping and culling, building exposures, or using chemical agents to attract, repel, or disrupt the reproductive biology of invasive species. However, these strategies can be time-consuming, expensive, and labor-intensive. Biological control strategies involving parasites and other pathogens have been proposed and successfully implemented for several vertebrate pest species and may offer long-term regulation with minimal input of labor or resources. For example, viral pathogens (myxomatosis and calicivirus) have been successfully employed against rabbit populations in Australia, and feral cats have been dramatically reduced on the sub-Antarctic Marion Island by a combination of feline parvovirus and culling. Microbial pathogens have also been proposed as biocontrol agents to regulate brown tree snakes on the Pacific island of Guam and fire ants in North America. Characteristics of promising biological control agents include a high specificity for the target host species, high R (determined by high transmission rates or low host recovery), and the ability to retain infectivity for long periods of time within hosts or in the external environment. Epidemiological models also suggest that pathogens with large effects on host fecundity (R) or moderate effects on host mortality (a) will provide the greatest degree of sustainable host regulation (Fig. 2). Clearly, the introduction of novel pathogens to island and other ecosystems requires careful feasibility studies to ensure the pathogens will not infect native species or escape to other areas and cause devastating epidemics. Therefore, this method of pest control should probably be restricted to oceanic islands to avoid the escape of novel pathogens into mainland ecosystems (Dobson, 1988).

C. Disease and Host Genetic Diversity

In many ways, the maintenance of genetic polymorphism within species parallels the coexistence of multiple species in ecological communities and will presumably enable hosts to evolve in response to changing environments and resist ecological invasions and perturbations. Coevolution between hosts and parasites is likely to generate host genetic diversity and may provide selective advantages for host recombination and sexual reproduction (Lively and Apanius, 1995). Although examples from wild animal populations are rare (Read et al., 1993), parasites infecting Soay sheep, snails, and insects have been implicated in the maintenance of allelic diversity or sexual recombination in their hosts. In contrast, one of the most distinctive features of natural plant populations is the staggering abundance of genetic polymorphism for resistance to fungal diseases.

Models of host–parasite coevolution demonstrate that advantages held by rare alleles can lead to the maintenance of genetic variation in host populations. These models are based on reproductive costs for hosts that acquire infection (or parasites that “attempt” to infect incompatible host genotypes), leading to cycles in both host and parasite allelic frequencies. In this way, frequency-dependent selection favors rare host genotypes that are resistant to diseases attacking the majority of individuals. Modern agriculture and animal husbandry provide examples of how parasites exploit monocultures or genetically homogeneous host populations. The phenomenon of parasites tracking common host genotypes is also critical for arguments concerning the role of parasites in generating advantages to genetic recombination through sexual reproduction (Lively and Apanius, 1995). Key insights from models of agonistic coevolution are that genetic diversity can be maintained even in the absence of stable equilibria, and that genetic polymorphism is more likely when fitness differences among host genotypes are small rather than large.

Parasites are likely to be powerful selective agents in natural populations, and host species that are continuously exposed to a diverse array of parasites should harbor a variety of resistance alleles or repertoire of inducible defenses. The major histocompatibility locus...
creased susceptibility to predation. For example, red mortality due to reduced competitive ability or in-
can happen through direct disease mortality or indirect threshold necessary for maintenance and growth. This manner, they can reduce population sizes below a populations in a variety of ways. In the most direct
some endangered species (Primack, 1998).

Parasitic organisms have been shown to impact small populations in a variety of ways. In the most direct manner, they can reduce population sizes below a threshold necessary for maintenance and growth. This can happen through direct disease mortality or indirect mortality due to reduced competitive ability or increased susceptibility to predation. For example, red

DISEASES AS THREATS TO BIOLOGICAL DIVERSITY

A. Introduced Pathogens and
Species Declines

Exotic diseases and parasites are increasingly recog-
nized as important factors driving population declines, geographic range contractions, or even terminal extinc-
tions in many plant and animal species. Although native parasites can play important ecological roles in natural ecosystems, a great number of introduced or exotic pathogens are directly implicated in catastrophic de-
clines of rare or endangered species. In such “virgin soil epidemics,” which occur after new pathogens are introduced into previously unexposed host populations, the disease progresses rapidly through immuno-
logically naive hosts and can cause mass mortalities. Because of the high rate of spread and the potential to devastate a host population, parasitic organisms are now considered the greatest threats to the survival of some endangered species (Primack, 1998).

Parasitic organisms have been shown to impact small populations in a variety of ways. In the most direct manner, they can reduce population sizes below a threshold necessary for maintenance and growth. This can happen through direct disease mortality or indirect mortality due to reduced competitive ability or increased susceptibility to predation. For example, red

(MHC) in vertebrates provides a notable example of extreme polymorphism and diversity of rare alleles (Nei and Hughes, 1993). Polymorphism at this locus deter-
mines the variety of foreign antigens the host immune system can recognize and attack, and it may be the ultimate response to selection in the face of unpredict-
able and frequent disease outbreaks. Gene-for-gene co-
evolution between plants and pathogens has also been demonstrated to promote a high diversity of resistance and virulence alleles. For example, long-term field stud-
ies of the interaction between wild flax and flax rust in natural populations in Australia indicate that many alleles can persist among metapopulations, and the dis-
tribution of genotypes can change rapidly during indi-
vidual epidemics (Burdon and Jarosz, 1991). From a broader perspective, coevolution between hosts and parasites may be a major force determining the earth’s biodiversity. Conservation strategies that fail to recog-
nize this potential and restrict disease spread may ulti-
mately rob host populations of genetic diversity needed to respond to future ecological changes.
generated complicated shifts in the communities of both plants and mammalian predators occurring in the same area. Furthermore, after the chestnut blight (caused by the introduced fungal pathogen *Endothia parasitica*) led to the virtual disappearance of the American chestnut, eight species of Lepidopterans feeding predominantly on this tree species also succumbed to extinction (Meffe and Carroll, 1997).

New and potentially dangerous pathogens can be introduced into wildlife populations in several ways. Shifts in the geographic distributions of species can result in two previously separated host species contacting each other and exchanging new diseases. Epidemics of introduced pathogens can also result from contact between domestic species and wildlife (Table II). Finally, new diseases may be introduced into the wild by captive-bred hosts that acquired an infection while in human care (Table III).

It is important to keep in mind that for most species, endangered populations are not likely to sustain any pathogens in the long run simply because they are too small. This is particularly true for either specialized or highly virulent parasites. In fact, it is often generalist (i.e., nonspecific) parasite species that are of greatest conservation concern because they can survive for long periods in alternate host species. Past experience suggests that it is mostly directly transmitted microparasites that cause most severe problems for endangered species. Nonetheless, more research is needed to understand the potential impacts of parasites with different life cycles or natural history requirements on small host populations.

## B. Endemic Diseases

In addition to introduced pathogens, conservation biologists regularly face outbreaks of parasites that are well

<table>
<thead>
<tr>
<th>Host species</th>
<th>Parasite/pathogen</th>
<th>Location</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>American chestnut (Castanea dentata)</td>
<td>Chestnut blight (<em>Endothia parasitica</em>)</td>
<td>Eastern North America</td>
<td>Pathogen introduced by ornamental trees</td>
</tr>
<tr>
<td>Flowering dogwood (Cornus florida)</td>
<td>Discula destructiva (fungus)</td>
<td>Eastern North America</td>
<td>Introduced pathogen</td>
</tr>
<tr>
<td>Several native plant species</td>
<td>Phytophthora cinnamomi (fungus)</td>
<td>Western Australia</td>
<td>Responsible for large-scale diebacks and permanent plant community shifts</td>
</tr>
<tr>
<td>American elm (Ulmus americana)</td>
<td>Dutch elm disease (<em>Ceratocystis ulmi</em>)</td>
<td>North America</td>
<td>Introduced pathogen</td>
</tr>
<tr>
<td>Invertebrates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long-spined sea urchin (Diadema antillarum)</td>
<td>Undetermined</td>
<td>Caribbean Sea</td>
<td>Epidemic spread across the entire Caribbean Sea</td>
</tr>
<tr>
<td>Fish</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rainbow trout, salmon (Salmo)</td>
<td>Myxobolus cerebralis (Protozoa; wholesale disease)</td>
<td>Montana</td>
<td>Introduced with stocked fish</td>
</tr>
<tr>
<td>Aral Sea sturgeon (Acipenser nudiventris)</td>
<td>Nitzschia sturionis (Monogenean)</td>
<td>Aral Sea, former USSR</td>
<td>Introduced with stocked Caspian sturgeon</td>
</tr>
<tr>
<td>Amphibians</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Various species of frogs and toads</td>
<td>Batrachochytrium dendrobatidis (Fungus)</td>
<td>Australia, Central America, Arizona, Sierra Nevada</td>
<td>Highly lethal pathogen; additional contributing stressors involved</td>
</tr>
<tr>
<td>Reptiles</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Desert tortoise (Gopherus agassizii)</td>
<td>Upper respiratory tract syndrome</td>
<td>Mojave Desert</td>
<td>Introduction through released pets</td>
</tr>
<tr>
<td>Birds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hawaiian honeycreepers</td>
<td>Plasmodium relictum capistranaeae</td>
<td>Hawaii</td>
<td>Implicated in the extinction of several Hawaiian bird species</td>
</tr>
<tr>
<td>Various species of ducks and geese</td>
<td>Duck plague (herpesvirus)</td>
<td>North America</td>
<td>Introduced with domestic waterfowl</td>
</tr>
</tbody>
</table>
TABLE III
Examples of Diseases-Related Problems in Captive Breeding Programs

<table>
<thead>
<tr>
<th>Species</th>
<th>Pathogen</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elephants (both African and Asian)</td>
<td>Herpesvirus</td>
<td>Cross-species infection between the two taxa</td>
</tr>
<tr>
<td>Black-footed ferrets (Mustela nigripes)</td>
<td>Canine distemper virus (morbillivirus)</td>
<td>High sensitivity to disease</td>
</tr>
<tr>
<td>Cranes (several species)</td>
<td>Herpesvirus</td>
<td>Outbreak at the International Crane Foundation facilities in Wisconsin</td>
</tr>
<tr>
<td>Mauritius pink pigeon (Columba mayeri)</td>
<td>Herpesvirus</td>
<td>Infected from domestic pigeons (foster parents)</td>
</tr>
<tr>
<td>Bali mynah</td>
<td>Isospora sp. (Coccidia)</td>
<td>Reduces reproduction in captivity</td>
</tr>
<tr>
<td>Various species of primates and large cats</td>
<td>Bovine tuberculosis (Mycobacterium bovis)</td>
<td>Pathogen infects wide range of mammals</td>
</tr>
<tr>
<td>Nene (Anser sandwichensis)</td>
<td>Mycobacterium</td>
<td>Captive-bred birds cannot be released into the wild because of infection</td>
</tr>
</tbody>
</table>

established in a regional wildlife population but may erupt because of unusual environmental conditions. As discussed earlier, the occurrence of disease depends not only on the characteristics of both the parasite and the host but also on the prevailing environmental circumstances. As a result, specific changes in environmental conditions can precipitate disease outbreaks in wildlife populations. This may occur either because the new conditions increase host population density above the threshold needed for rapid spread of the pathogen or because they directly favor parasite transmission. Hence, such conditions could either enhance the survival of infective stages between hosts or stress hosts to the point that they become particularly susceptible to infection. A good example in which environmental conditions caused increased host population size and eventually a disease epidemic is the outbreak of the Sin Nombre virus in the deer mouse populations in the American Southwest. In the early 1990s, El Niño–Southern Oscillation event (ENSO) associated rains led to an explosive increase in the rodent populations in Arizona and New Mexico. These enormously high population densities triggered an epidemic of the Sin Nombre virus in the deer mouse populations in the American West suffered from previously established pathogens such as Acrornonas hydrophila (tularemia disease) and the Saprolegnia fungus after some new environmental stressors (perhaps in the form of increased UV radiation) appeared in the area (Meffe and Carroll, 1997).

C. Concerns for Captive Breeding Programs

Conservation efforts rely increasingly on captive breeding programs to augment and restore free-living populations. Because captive animals are particularly susceptible to infections, disease has emerged as a serious threat to these programs, and pathogens have become one of the principal concerns of captive breeding managers. Disease outbreaks are now being taken seriously not only because of their unpredictable nature and the potentially catastrophic loss of breeding stock but also because of increased costs of screening and treatment, concerns for housing and animal care, and the erosion of public support in the event of outbreaks. In addition, diseases propagated in captive animals, when introduced into the wild with released hosts, can jeopardize the health and demography of the very same populations the programs are trying to save.

Captive animals may acquire novel infections from a wide range of infected hosts, including unrelated species kept in the same pen, wildlife living around outdoor cages, foster parents, or individuals from the same species but different subspecies. For example, captive African elephants kept in mixed collections have been infected with a lethal herpesvirus that occurs without disease symptoms in their Asian elephant pen mates. Furthermore, many captive-bred hatchlings of the endangered Mauritius pink pigeon contracted and succumbed to a herpesvirus infection that their foster parents (domestic rock doves) were carrying without ill
effects (Primack, 1998). Finally, humans pose a serious threat to many captive animals. Several human diseases, such as measles, tuberculosis, or influenza, are highly virulent for wildlife, because of the large numbers of visitors that pass through zoos every year, captive animals may become exposed to a very large pool of potentially infected humans. For example, captive-bred orangutans that were infected with tuberculosis from humans could not be released into the wild due to fear of introducing the pathogen into the wild population.

Many animals in captive breeding programs are often held close together, a practice that poses two disease-related risks. First, animals are likely to be stressed and hence more susceptible to infection (particularly those that are territorial or normally persist at low densities). Second, crowding in pens or cages can elevate host densities above the threshold necessary for virulent pathogens to invade and will also increase transmission rates (e.g., hosts may reinfect themselves by ingesting the eggs of their own parasitic nematodes released into their pen). Interestingly, parasites with complex life cycles may be less of a concern to captive breeding programs because of the likely absence of intermediate hosts (or vectors) that are necessary for transmission.

Finally, additional complications exacerbating disease problems in zoos stem from inbreeding depression, or the genetic impoverishment of a captive colony due to loss of diversity and the expression of deleterious recessive alleles. This loss of genetic variability leads to homogeneous captive populations that can be very susceptible to a variety of pathogens. Such loss of genetic heterozygosity has been implicated in the high mortality that captive cheetah populations experienced due to a feline infectious peritonitis virus (Soule, 1986). Hence, genetic and ecological problems can operate in synergy to reduce population size and diminish heterozygosity, leading populations toward increased disease susceptibility and possible extinction. In conclusion, as captive breeding programs expand, disease-related problems are likely to become even more prevalent than they are now. Because of their unpredictability and their potential to cause great harm, serious precautions are needed to prevent future disasters.

V. INTERACTIONS BETWEEN DISEASE, WILDLIFE, AND HUMAN ACTIVITIES

Wildlife disease outbreaks are usually directly or indirectly facilitated by human activities. In particular, economic development in conjunction with social processes can influence disease emergence through changes in land use or demographic patterns. These activities can lead to quantitative habitat changes (i.e., causing direct habitat destruction or fragmentation) or qualitative changes (such as pollution or other forms of habitat degradation).

A. Quantitative Changes in Natural Habitats

Humans destroy natural habitats in a plethora of ways, many of which have been summarized under the term habitat fragmentation. Habitat fragmentation is the process of reduction and subsequent isolation of a continuous natural habitat into smaller patches; today, it is an extremely widespread phenomenon affecting every major terrestrial ecosystem. Habitat fragmentation is of great conservation importance because it affects native communities both directly (by removing individuals and their resources) and indirectly (through changes in community composition and interactions). Fragmentation of a continuous habitat can affect the spread of parasitic organisms in a variety of ways, all of which depend on specific community characteristics or idiosyncrasies of individual host–parasite systems.

Habitat fragmentation results in both reduction and subdivision of a host population. Depending on the degree of isolation between the different habitat patches, parasites or pathogens may infect some but not all host subpopulations. As a result, many of these subpopulations escape a disease epidemic that can devastate other infected patches. Furthermore, if host subpopulations are relatively small they may fall below the critical host threshold required for disease persistence ($N_t$). Consequently, one possible result of habitat fragmentation may be the decline or extinction of a parasitic species.

Habitat fragmentation will also increase the contact between endangered native hosts and other domesticated and wild species living in adjacent habitats, facilitating the introduction of new diseases into previously unexposed populations. As human activities degrade continuous natural habitat into progressively smaller habitat fragments, the percentage of habitat edge—defined as the habitat close to the border with the disturbed matrix—steadily increases. As a result, increasingly more individuals of an endangered species living in this habitat are likely to encounter other host species living along the edge of this matrix. The increased likelihood of such chance encounters also raises the probability of cross-species infection. Indeed, there is evidence that in many cases host switching between domestic
and wild animals is facilitated by habitat fragmentation and penetration by exotic disease-carrying hosts. For example, such degradation appears to have preceded the spillover of rabies and canine distemper virus from domestic dogs to African wild dogs and Ethiopian wolf populations.

Fragmentation-caused disease problems are further exacerbated by wildlife crowding. As increasingly more habitat is destroyed, surviving native hosts will increasingly move into the last few remaining patches. Such crowding is particularly severe in waterfowl populations, which in many areas of the world are forced to overwinter or migrate through wetland areas greatly reduced by human activities. Such large numbers of birds not only degrade vegetation cover and water quality but also compete for reduced food resources and may hence become nutritionally stressed. High bird densities also increase the probabilities of disease transmission through either direct contact or fecal contamination and can further facilitate a disease epidemic. In line with this argument, wildlife managers have seen massive outbreaks of disease in crowded waterfowl populations during the past several years.

B. Qualitative Changes and Pollution

Degradation of natural habitats takes many forms and has frequently been associated with outbreaks of new diseases in both wildlife and humans. Human assaults on natural ecosystems as varied as logging, changes in vegetation cover, overexploitation of game and fisheries, erosion, and pollution can trigger pathogen epidemics. Direct injury due to logging and other anthropogenic factors can cause increased levels of disease in native plant communities (Gilbert and Hubbell, 1990). Many scientists suggest that the widespread 1988 phocine distemper epidemic that spread through the seal nation and can further facilitate a disease epidemic. In line with this argument, wildlife managers have seen massive outbreaks of disease in crowded waterfowl populations during the past several years.

B. Qualitative Changes and Pollution

Degradation of natural habitats takes many forms and has frequently been associated with outbreaks of new diseases in both wildlife and humans. Human assaults on natural ecosystems as varied as logging, changes in vegetation cover, overexploitation of game and fisheries, erosion, and pollution can trigger pathogen epidemics. Direct injury due to logging and other anthropogenic factors can cause increased levels of disease in native plant communities (Gilbert and Hubbell, 1990). Many scientists suggest that the widespread 1988 phocine distemper epidemic that spread through the seal nation and can further facilitate a disease epidemic. In line with this argument, wildlife managers have seen massive outbreaks of disease in crowded waterfowl populations during the past several years.

C. Global Climate Change and Disease

There is increasing evidence that the changes wrought by human activities on global climate are affecting the distribution, prevalence, and severity of pathogenic organisms in natural ecosystems. A growing body of research demonstrates that the infusion of anthropogenic greenhouse gases such as methane, CFCs, and CO$_2$ is changing both temperature and precipitation patterns throughout the planet. Because the distribution and prevalence of parasites often depends crucially on environmental conditions, it is not surprising that changing weather patterns affect the prevalence of wildlife diseases.

The best documented examples of such climate–disease interactions come from marine ecosystems. Increased water temperatures have been shown to trigger
coral bleaching (expulsion of symbiotic algae from the coral) and a suite of associated opportunistic infections that have led to massive coral die-offs (Harvell et al., 1999). Although such bleaching events can occur naturally in association with ENSO events, both their frequency and severity have increased during the past 20 years. Because such die-offs have profound community consequences in coral reef ecosystems (because affected reefs take many years to recover), this has become an issue of great conservation importance.

The link between global climate change and animal pathogens is less well documented for terrestrial systems. Although evidence exists that the expanding distribution of some human diseases (such as malaria and dengue) is tied to recent warming trends, much less is known about animal parasites. In the human transmission cycle, the distributions of both the Plasmodium parasites and their Anopheles mosquito vectors are limited by low environmental temperatures. As a result, malaria does not exist in high latitudes or at high elevations in the tropics. In recent years, however, malaria has been occurring at progressively higher elevations, potentially because of warmer ambient temperatures (Patz, 1997). Similar advances of animal diseases (such as the previously discussed *P. reticulatum*) could presumably spread into high-elevation refugia such as the mountains of Hawaii, leading to the extinction of many endangered bird species surviving there.

In addition to changes in temperature and precipitation regimes throughout the world, greenhouse gases also contribute to the depletion of the ozone layer and a concomitant increase in harmful UV radiation. Such radiation can affect negatively many species of vertebrates and may facilitate the spread of disease. There is indeed evidence suggesting that such increases impact amphibian populations by raising the susceptibility of egg masses to fungal infections (Kiesecker and Blaustein, 1995). In summary, there is accumulating evidence that global climate change will emerge as an important factor facilitating the spread of animal pathogens. Therefore, exploring the potential consequences of future changes in global climate should be an important future research goal for scientists and wildlife managers involved in conservation.

D. Accidental Introductions of Disease via Worldwide Commerce and Travel

Worldwide commerce and animal trafficking are probably the most important causes of new disease outbreaks in wildlife populations. This traffic, which has dramatically increased in the past few decades, takes a variety of forms. It includes international trade of live animal and plant stock for commercial breeding purposes, for zoos and animal parks, for the pet trade and hunting, and for laboratory research. The situation is further exacerbated by the staggering amounts of international transport of fruit, vegetables, and various animal parts, all of which facilitate the spread of pathogens. For example, most of the major pathogens that cause disease epidemics in wild birds in North America were probably introduced by humans. Both avian cholera (a highly virulent disease caused by the bacterium *Pasteurella multocida*) and duck plague (caused by a herpesvirus) appeared first in North American domestic birds and spread to wild waterfowl. Today these diseases are responsible for the deaths of tens of thousands of wild waterfowl each year in the United States alone. Similarly, avian malaria was introduced to Hawaii with imported birds, and brucellosis (caused by the bacterium *Brucella abortus*) was introduced to North America by domestic livestock (and has since spread to wild deer, elk, and buffalo populations). Although brucellosis is controlled in domestic livestock today, it is of great conservation concern for large ungulates in the greater Yellowstone area.

The spread of disease into wildlife populations is further facilitated by the continual breakdown of barriers between livestock and wild animals. For example, today there is a wide spectrum of waterfowl populations, all with varying degrees of dependency on humans (such as traditionally farmed animals, captive collections of exotic waterfowl, birds bred in game farms and released into the wild, feral birds, nonmigratory birds, and fully migratory wild birds). Many of these populations transfer pathogens as a result of their frequent contact with each other.

Finally, international travel and tourism has emerged as an important factor that facilitates the spread of pathogens and their vectors into new populations. The constant shuttling of infected humans between different continents can potentially allow parasites to spread throughout the world in a matter of days. Although this point is not lost to epidemiologists dealing with human infections, it probably needs to be taken more seriously by both ecologists and conservation biologists.

VI. MANAGING DISEASES AND BIODIVERSITY IN THE FUTURE

A. Preventing Disease Outbreaks

Pathogens that pose significant threats to conservation programs were discussed in detail earlier. In summary,
the sources of these diseases can be divided into two categories. The overwhelming majority of epidemics begin with the introduction of a disease into a naive population, either via captive-bred individuals carrying a pathogen or via the invasion of a host species into new areas. Also of concern are environmental changes that affect host physiology and shift relationships with endemic parasites. Anthropogenic effects on global and local environments (such as habitat loss and fragmentation) and predicted changes in global temperature and weather patterns will have a cascade of consequences for wild populations, including the threats from diseases.

To prevent future disease outbreaks it is imperative to monitor the sources of disease introduction. Disease threats to wild and agricultural populations are managed through screening imported animals, plants, and products made from their tissues. Frequently, wild and domestic animals must undergo quarantine before release into native populations. Furthermore, disease screening is increasingly included in captive breeding programs to prevent the inadvertent introduction of infected individuals into a susceptible population. Ultimately, the majority of disease-related threats to conservation arise from human-related disturbances, and it is difficult to find an example of a disease-induced wildlife crisis in which human activity is not the origin of the problem.

B. Monitoring Populations for Infection

Most diseases are discovered after epidemics have already spread through wild populations. A powerful tool in the management of diseases is thus to monitor threatened populations for signs of infection before overt mortality occurs. Several such tools exist to both monitor unaffected populations and track the spread of an epidemic. The efficacy of a monitoring program will increase with the number of host animals included in the screening.

Common sources of information that can be used to monitor parasite prevalence and intensity in animals are blood, tissue, and fecal samples. The presence of blood parasites (e.g., trypanosomes, malaria, and filarial nematodes), anemia, elevated leukocyte levels, and pathogen-specific antibodies can all be detected from relatively small volumes of blood. Antibody assays (e.g., enzyme-linked immunosorbent assay) can show evidence of both current and past infection. Many microbial pathogens can be cultured directly from tissue showing signs of infection or from swabs taken from the mouth, ears, eyes, nose, genitals, or anus. Scans of feces will also provide information on the diversity of species of intestinal parasites in an individual as well as the number of worms per host or intensity of infection. Hosts that are hunted or culled for other purposes are frequently examined for internal and external parasites. Parasitic arthropods (e.g., ticks, mites, fleas, and biting dipterans) are also monitored because they can decrease the condition of their hosts by drawing on resources. High parasite burdens can indicate a decrease in immune function due to other causes, and arthropods should be examined for internal microparasites because they often serve as vectors for other diseases. For both internal and external macroparasites, the distribution of parasites among hosts is particularly critical in assessing the impact of infection on individual hosts and at the population level.

Molecular techniques have recently been employed as methods for screening free-living populations with surprising results. For example, polymerase chain reaction tests can reveal the presence of infection in host blood or tissue by identifying pieces of pathogen genetic material. Such a study on native Hawaiian birds indicated that the prevalence of avian malaria was higher than original estimates based on microscopic examination of blood (Feldman et al., 1995).

C. Assessing Disease Threats

Screening programs will verify the presence or absence of a particular pathogen in a wild population. However, this information alone does not provide a definitive indication of severe population consequences resulting from infection. Epidemiological models discussed earlier indicate that this determination can only be made when information on both the prevalence of a disease (y) and its effects on individual hosts (a and b) have been collected. In estimating prevalence, care should be taken to sample a large number of hosts throughout their entire range. To determine the effects of disease on individual hosts, postmortem examinations are often performed. However, the presence of a pathogen in dead or dying animals does not necessarily indicate the population-level impact of the disease, nor does it offer a measure of disease effects on host life span or fecundity.

Ideally, captive or wild hosts should be monitored throughout the course of infection to compare survival and fecundity between infected and uninfected hosts. Experimental manipulation of parasite loads in natural populations remains the most direct way of assessing the effects of pathogens on host populations (McCallum and Dobson, 1995). Epidemiological models suggest that if parasites affect host mortality alone, then those of intermediate virulence are the most likely to pose problems for wild
populations (Fig. 2). These models also predict that a highly virulent pathogen may be maintained at a high prevalence in a host population if it infects multiple host species and is more benign in some hosts than in others. For example, in the late nineteenth century, Rinderpest devastated much of the native ungulate fauna of sub-Saharan Africa because it persisted in domestic cattle as a less virulent infection. Essentially, this notion explains most disease introductions into managed populations because the carrier host remains asymptomatic and undetected. Therefore, complete screening efforts and disease management plans should include the determination of the infection status and disease history of closely related taxa that overlap in range.

It is worth noting that correlations have been demonstrated between the degree of inbreeding in individuals (indicated by high levels of homozygosity or from known pedigrees) and the likelihood of contracting a serious infection. Given this association, measuring inbreeding coefficients as a part of population screening and avoiding inbreeding in captive populations may restrict future epidemics by increasing the genetic variation in managed populations.

D. Intervention Methods

Historically, diseases in wild populations have drawn the attention of wildlife managers only after an epidemic severely threatens the host population or when the disease is perceived at a threat to valuable agricultural crops or livestock. In general, the types of management regimes used to limit disease spread vary depending on the type of pathogen, the threat to the host population or to other neighboring species, available financial resources, and the existence of agricultural or veterinary information and tools. The use of veterinary and agricultural designed tools to treat domestic species is particularly important because very little information exists on disease treatment in native wildlife and plants. However, direct application of domestic animal medicine on wildlife can have disastrous effects, as illustrated by the mortality caused by a canine distemper vaccine given to endangered black-footed ferrets. Some of the more common intervention methods available to conservation biologists are discussed in the following sections.

1. Vaccination

Viral infections and, less frequently, bacterial infections are most effectively controlled by vaccinations. However, the cost of vaccinations may severely limit the success of population-wide disease control plans. Programs intended to eradicate or control the spread of a virus must consider basic epidemiological principles and models in determining where to concentrate efforts and what proportion of the population to target to limit or eradicate the disease. Both oral vaccines and intramuscular injections have been used in wild populations. Vaccinations have been administered to prevent the spread of rabies in European foxes, and vaccination of domesticated reservoir hosts (cattle and domestic dogs) has reduced the threats of morbillivirus outbreaks in wild ungulates and carnivores in east Africa.

2. Chemotherapy

Drugs are most frequently administered for bacterial, fungal, helminth, and ectoparasitic infections. This method of disease control least effectively addresses the ultimate cause of a disease and may be extremely costly for population-wide control measures (e.g., systemic fungicides to counter tree blights). Interestingly, treatment of helminth infections in red grouse in Scotland and in Soay sheep in St. Kilda reduces dramatic host population cycling. A more stable population may be easier to manage; therefore, the treatment of endemic macroparasite infections can have great conservation value.

3. Culling

The removal of infected individuals by lethal means is analogous to intensifying parasite-induced mortality (or, Fig. 1). This strategy can effectively decrease disease prevalence and lower R0, and it has been implemented to counter the spread of certain tree diseases (e.g., Dutch elm disease). Although less frequently applied to vertebrate populations, Cape buffalo in some South African national parks are screened for tuberculosis and killed if infected. The culling of individuals of a host species, regardless of infection status, has also been suggested as a means for reducing the numbers of susceptible hosts in population below the threshold required for parasite persistence.

4. Quarantine

The costly removal and care of infected individuals from a susceptible population has been carried out in extreme circumstances. For example, when host populations are dangerously low in numbers, then every member may be of great value and worth rehabilitating. This was the case during a distemper epidemic in Wyoming black-footed ferrets, and a similar situation may arise if rabies invades endangered Ethiopian wolf or African wild dog populations.
5. Biological Control
The use of one pathogen to attack another is certainly a controversial measure and should be included in management regimes with much caution. There are instances in which natural parasites of pathogenic organisms have been proposed as potential tools in slowing or controlling the spread of infection. For example, hypovirulence in the fungal pathogen responsible for chestnut blight, Cryphonectria (Endothia) parasitica, is caused by a virus that is transmitted via direct hyphal contact. This virus debilitates the pathogen and allows chestnut trees to counter and recover from fungal infections.

6. Hybridization
In cases in which variability in pathogen resistance in a host population is negligible, management plans may resort to crossing highly susceptible species with more resistant organisms. For example, crossing black-footed ferrets with domestic European ferrets has been proposed to decrease susceptibility to viral pathogens. Hybrids of American elm trees and Eurasian elm species (which are resistant to Dutch elm disease) are already available on the commercial market.

7. Habitat Alteration
Management regimes that address the ultimate causes of disease outbreaks have the greatest potential for removing disease threats but are also the most difficult to implement. Indeed, the ultimate causes of marine invertebrate diseases, namely, pollution and ocean temperature changes, are so global and diffuse in origin as to be impossible to confront in any single species management plan. However, alterations in terrestrial habitats have been suggested to alleviate disease problems in conservation. Two such examples are the proposed creation of a boid-free land zone around Yellowstone National Park to prevent contact between bison and cattle and the removal of feral pigs from the Hawaiian Islands (because their activities increase mosquito-breeding areas and elevate the transmission of avian malaria).

An understanding of disease ecology is also pertinent to the design of habitat reserves. For example, how large should reserves be to prevent host crowding? Do corridors between reserves increase the threat of pathogen transfer among locations, or does host dispersal among habitats facilitate the spread of resistance genes or aid in parasite avoidance? Maintaining species richness and genetic diversity within reserves is also critical to limiting threats from disease. For example, high species diversity may buffer natural communities from devastating epidemics, and habitats that are restored with genetically homogeneous stock may be much more susceptible to pathogen invasion.

Ultimately, no one method of intervention will ever be sufficient to conserve endangered species, eliminate disease threats, and satisfy local farmers, ranchers, and residents. Cost–benefit analyses of intervention methods along with the use of epidemiological models to predict effectiveness of any proposed measures are vital components to an effective disease management plan. The most expedient measures may not always address the true causes of an outbreak, such as habitat disturbance, invasive species encroachment, or crowding. Plans based in community ecology with a multispecies approach are more likely to address the presence and effects of reservoir species. Finally, it is important to note that parasites are natural elements of all communities and their presence alone may not always warrant intervention. In fact, parasites can act as selective forces in host populations, and maintaining variability in host resistance may be important for future population responses to disease.

See Also the Following Articles
CAPTIVE BREEDING AND REINTRODUCTION • CLIMATE CHANGE, SYNERGISM OF • COEVOLUTION • DIVERSITY, COMMUNITY/REGIONAL LEVEL • GENETIC DIVERSITY • MICROORGANISMS, ROLE OF • PARASITISM

Bibliography
Feldman, R. A., Freed, L. A., and Cann, R. L. (1995). A PCR test for hypovirulence in the fungal pathogen responsible for chestnut blight, Cryphonectria (Endothia) parasitica, is caused by a virus that is transmitted via direct hyphal contact. This virus debilitates the pathogen and allows chestnut trees to counter and recover from fungal infections.


DISPER SAL
 BIOGEOGRAPHY

Ran Nathan
Princeton University

I. Introduction
II. Perspectives on Dispersal
III. Determinants of Dispersal and Colonization Success
IV. Methodological Approaches and Case Studies
V. Biogeographic and Evolutionary Consequences of Dispersal
VI. Applied Dispersal Biogeography
VII. Synthesis

GLOSSARY

dispersal The movement of disseminules away from their parent source. Because of the variety of disciplines involved with the study of movements of organisms, the term is often misused and confused with a plethora of terms. The term dispersion refers to the fine-scale spatial distribution pattern of organisms, and migration is generally equivalent to dispersal, though it typically refers to a particular (directional and synchronized) type of dispersal. The term immigration is dispersal from the perspective of the recipient locality regardless of the source—the arrival of new individuals to an area previously not occupied by that species. It should be emphasized that dispersal does not deal with events that occur after a movement has been completed. The term colonization should be used for incidents of immigration that are followed by establishment of a viable population.
dispersal barrier The species-specific physical and biological restrictions a propagule must overcome to accomplish successful colonization.
dispersal biogeography Usually refers specifically to "a branch of historical biogeography that attempts to account for present-day distributions based on the assumption that they resulted from differences in the dispersal abilities of individual lineages" (Brown and Lomolino, 1998; p. 628). More generally, it is the study of the relationships between dispersal and the geographical distribution of organisms. I adopt the latter definition for this article.
dispersal route The particular set of physical and biological conditions that allow organisms to cross dispersal barriers.
disseminate The unit of dispersal; any part or stage in the life cycle of an organism that is used for dispersal.
propagule The unit of colonization; a disseminate, or group of disseminules, which has the full potential to establish a new population.
range expansion The spatiotemporal process in which successive colonization events increase the size of a species' geographical range. The terms invasion, migration (as used in the paleobotanical literature), and spatial spread are roughly synonyms for range expansion. Invasion, in particular, is often used to describe colonization outside the species range.
DISPERAL IS A PROCESS COMMON to all organisms, and it is highly variable in details among organisms and across space and time. At large spatial or temporal scales, this process has broad implications for biogeography through both long- and short-distance dispersal events. The study of the relationships between dispersal and the geographic distribution of organisms, advanced by various methodological approaches, is critical to our understanding of forces affecting the world’s biodiversity.

I. INTRODUCTION

In Chapters XII and XIII of *The Origin of Species*, Charles Darwin (1859) discussed how patterns of geographical distribution of species are associated with his ideas of natural selection. He argued that “all the grand leading facts of geographical distribution are explicable on the theory of migration, together with subsequent modification and the multiplication of new forms.” Darwin illustrated this argument through many examples, by explaining different biogeographic patterns, emphasizing the role of dispersal barriers (see glossary) and the various means of long-distance dispersal through which organisms are able to cross them. Although Carl Linnaeus (1707–1778) and Augustin-Pyramus de Candolle (1778–1841) had already expressed similar ideas about the importance of dispersal in determining species distributions, Darwin (1809–1882) and Alfred Wallace (1823–1913) are those most identified with the dispersalist school. This school asserts that disjunction and range expansion occur through rare long-distance dispersal events across preexisting barriers. Joseph Hooker (1817–1911), one of Darwin’s closest confidants, viewed long-distance dispersal across barriers as extremely unlikely and argued that range expansion was possible only across ancient land bridges that once connected currently disjoint continents. He thus argued that biogeographic disjunction can be explained by vicariant (splitting) events in which new barriers have subdivided a previously continuous range (see Vicariance Biogeography). Evidence for Hooker’s land bridges never emerged; however, the two basic opponent ideas of dispersal versus vicariant history were already brought into the front stage of biogeography. In *The Secular Ark*, Janet Browne (1983) describes the 19th century discussion on this issue involving Darwin, Wallace, Hooker, and others.

The 20th century has seen a further polarization of the vicariance-dispersal debate. Plant biogeographers such as Evgenii Wulf (1885–1941) and Stanley Cain (1902–1995) reiterated serious doubts about the ability of long-distance dispersal to account for present plant disjunctions, in light of the extreme rarity and lack of direct empirical evidence of these events. Others, like Philip Darlington (1904–1983), advocated Darwin’s dispersalist view, but without Darwin’s reliance on specific centers of origin. In particular, the equilibrium theory of island biogeography (MacArthur and Wilson, 1967) ascribes a critically important role to immigration (see glossary) in determining the number of species on islands. This landmark theory, which had a major impact on both ecology and biogeography, modeled the probability of immigration of a new species as roughly a negative exponential function of the number of species on an island, with the rate also decreasing with distance from the mainland [see Island Biogeography].

At approximately the same time that Robert MacArthur and Edward Wilson published their theory, the important theory of continental drift, more generally of plate tectonics, first proposed early in the 20th century, was finally accepted. This theory provided a vicariant mechanism to explain biogeographic disjunctions without long-distance dispersal events. In the 1970s and early 1980s, the concept of vicariance biogeography received much attention among biogeographers, as reflected in three books by Nelson and Rosen (1981), Nelson and Platnick (1981), and Wiley (1981). The main argument raised against dispersalsists’ explanations, following the similar criticism by Leon Croizat (1894–1982), was that long-distance dispersal is a unique, highly stochastic and unpredictable phenomenon, hence its relationships with species distributions are not testable. However, truth and testability remain distinct: obstacles to proving the truth of an idea do not imply its falsehood. The explanations of biogeographic patterns by dispersalsists were condemned as “storytelling” scenarios which “in explaining everything really explain nothing.” However, solutions of complex biogeographic patterns by vicariant events alone, if taken to their limits, are also speculative. There are, indeed, examples of extreme cases in which disjunctions are exclusively explained by dispersal (e.g., colonization of remote newly created islands) or by vicariance (e.g., postglacial relicts). However, none of these extreme arguments provides a general explanation; in fact, as we shall see, not only do both phenomena exist, but they also frequently co-occur.

The dispersal-vicariance debate reflects a more general controversy over the biogeographic relevance of dispersal: to what extent does dispersal influence species distribution? This debate concerns whether disjunct distributions have resulted from dispersal across a barrier
or from a barrier splitting a continuous range. But dispersal and vicariant events are not the sole explanations for disjunctions, which may also result from differential extinction within a species range. The major problem in limiting dispersal biogeography to its currently narrow definition is that such a definition reduces the entire phenomenon of dispersal to long-distance dispersal across barriers, thereby ignoring the various biogeographic patterns to focus solely on disjunct distributions. As this discussion will show, there are many other distributional patterns that cannot be explained by either vicariant or extinction events but only by long- and short-distance dispersal events. Rich evidence gathered from diverse independent sources indicates not only that long-distance dispersal exists, but also that it is the only possible explanation for the geographical distributions of many species, such as the Hawaiian biota (Carlijn, 1974). Further, long-distance dispersal is not necessarily as random or unpredictable as it is often considered. The relative importance of dispersal varies between species, biogeographic regions, and times. Thus, to assess the relative biogeographic importance of dispersal, we need to explore the nature of dispersal and its relationships with the geographical distribution of species. The broader view of dispersal biogeography (see glossary) propounded here aims to understand how dispersal influences, and is influenced by, the geographical distribution of organisms.

II. PERSPECTIVES ON DISPERSAL

A. The Ecological Perspective

As a general rule, most disseminules (see glossary) travel relatively short distances from their parent source. The basic description of dispersal, called the dispersal curve, is a graph plotting the distribution of disseminules against the distance from their parent source (Fig. 1). Dispersal curves, derived for a great number of organisms, show tremendous variability, but also a right-skewed leptokurtic distribution with overwhelming regularity. That is, they typically describe a rapid decline in disseminule density with distance from their parent source, with more disseminules both near the source and along the tail than a corresponding normal distribution (Section IV.C.1). The restricted spatial extent of dispersal has important implications for the dynamics of populations, metapopulations, and communities and for other topics addressed in ecological studies of dispersal (Nathan and Muller-Landau, 2000).

The ecological perspective of dispersal is concerned not only with small spatial scales, but also with relatively short temporal scales. Dispersal is often restricted to early stages during an organism's life cycle (seeds, eggs, larvae, and juvenile stages in animals); the term natal dispersal denotes dispersal of young animals while breeding dispersal indicates dispersal of reproductive adults. Disseminules of many plants, nematodes, rotifers, desert butterflies, copepods, and other invertebrates can persist for years—sometimes hundreds of years—as dormant seeds or eggs, providing the means for a temporal exploration for favorable establishment conditions. From another temporal perspective, dispersal is typically a seasonal phenomenon, and synchronization of dispersal in different individuals, popula-
tions, and species is fairly common in both animal and plant communities. These temporal patterns have important effects at the ecological scale, but they are unlikely to exert a significant effect at the biogeographic level.

B. The Evolutionary and Population Genetic Perspectives

The great variety of dispersal-determining traits presents clear evidence that dispersal is of important adaptive value for many life forms on earth. From the evolutionary perspective, the study of dispersal is concerned with the source and significance of this variation: the factors and processes that affect the advantages of dispersal. Typically dispersal is advantageous for escaping the high mortality near the source, reducing competition with close relatives, minimizing the fitness cost of inbreeding depression by reducing the chances of mating with close relatives (in sexual species), and for locating new resources for establishment, growth, and reproduction. Theoretical models (see Johnson and Gaines, 1990, for a review) show that even in environments in which the physical conditions are stable in time and space, surprisingly high proportions of dispersing progeny may be adaptive (Hamilton and May, 1977). Since mutations beneficial to successful dispersal are typically rare, evolutionary effects are observable only at relatively long time scales (but see Section III.C). However, there is no characteristic spatial scale for the evolutionary perspective of dispersal: adaptive advantages of dispersal are expressed at both small and large spatial scales.

Dispersal provides the means for gene flow within and among populations; thus studies of population genetics explore how dispersal, in conjunction with the breeding system and selection regime, affect the genetic structure of populations. In general, high rates of dispersal lead to mixing of genotypes and thus lower regional genetic diversity, while low rates of dispersal (isolation) typically result in low local genetic diversity, but high regional diversity. Although these effects may also occur at small spatial and temporal scales, they are notably more pronounced in large ones.

C. The Biogeographic Perspective

Dispersal becomes important at the biogeographic level only at large temporal or spatial scales, when and where propagules disperse outside the species range and successfully establish viable populations (Section III.B). The identification of locations as “within” or “outside” the species range (and obviously the definition of a species range itself) is problematic since sampling of any species is typically incomplete. Even if we assume complete information on the exact location of each individual at any particular time, the issue of species boundaries is difficult because each species exhibits a profound variation in its local density and, in fact, is entirely absent from most locations. In biogeography, this distinction is less problematic when dealing with larger spatial and temporal scales.

Spatial spread can occur at different rates and can generate different spatiotemporal patterns, from slow gradual expansion to abrupt long-distance jumps. Both rate and pattern depend on the two basic components of spatial progression—the movement itself (dispersal) and local population growth (reproduction)—and their interplay (Section IV.C). The biogeographic literature provides numerous examples showing a continuum between these extremes, and it is also clear that gradual spread and long jumps can follow each other. However, there are basic differences between these two modes: gradual range expansion is the outcome of multiple ordinary (i.e., short-distance) dispersal events over a relatively long time period; jumps are rare individual events that abruptly change the species’ range.

1. Gradual Range Expansion

Gradual range expansion is the continuous spread of populations through multiple ordinary (short-distance) dispersal events, usually over quite long time scales. Very slow expansion has been termed secular dispersal (or secular migration); more rapid (though still gradual) expansion is referred to as diffusion (Pielou, 1979). Examples of gradual range expansion are the spread of mollusks, worms, crabs, and fishes from the Red Sea to the Mediterranean (Section IV.A.3) and of muskrats in Europe (Section IV.A.5).

2. Jump Range Expansion (Jump Dispersal)

Although most dispersal is limited to relatively short distances and therefore covered appropriately by the ecological scale (Section II.A), rare long-distance events have crucial implications for biogeography. There is enormous variation in the ability to disperse disennuiles over large distances within and, more significantly, between species. Specific morphological characteristics are associated with long-distance dispersal ability (Box 1). When such an event ends in establish-
ment, the spatial spread is clearly discontinuous. The
frequently used term for this type of spatial spread,

*jump dispersal*, does not fully describe the process: spa-
tial spread is not restricted to the movement phase
(dispersal) but also necessitates establishment. Hence,
the term *jump dispersal* should be replaced with *jump
range expansion*. Note also that while the original term is
appropriate for describing long-distance dispersal over
barriers, the term *jump range expansion* is also applicable
to situations in which suitable habitats are as common
in the gap as in the recipient area but are unoccupied
(see Sections IV.A.4.b and IV.C.1).

**Box 1**

**Mechanisms and Adaptations for Dispersal**

Previous workers have suggested many methods of
classifying the various means of dispersal. The
most fundamental distinguishes between organ-
isms that are actively dispersed (by their own
power) and those that are passively dispersed (by
a physical agent or by other organisms). Organ-
isms unusually capable of *active dispersal* are said
to be *vagile*, while those unusually capable of
*pasitive dispersal* are called *pagile*. Another com-
mon classification, mostly used for plants, is based
on the primary dispersal agent, using the suffix
*chori* (to disperse) at different levels: for example,
*zoocory* is dispersal by animals, *ornithocory* by
birds, and *endozoochory* is within an animal. Nei-
ther classification should be taken as an absolute:
many vagile organisms also disperse passively,
and many pagile organisms can move actively
over short distances. Seeds of many plants are
actually dispersed by multiple agents, rather than
by a single one.

(a) Shy albatross (*Thalassarche cauta*)—a vagrant seen off the

(b) Squaring cucumber (*Echallium elatium*). Photo by Pietro
Pavone; drawing from Hegi (1915).

Relatively few animals have the capacity to
travel long distances under their own power. Most
of these are volant (flying) species such as birds
(a), bats, and butterflies. Although many migrate
seasonally [see Migration], this large-scale movement
usually does not affect the geographical dis-
tribution of the breeding range. However, ex-
tralimital sightings are well documented for many
taxa, and occasionally may result in the establish-
ment of a new colony outside the species range
(Section IV.A.4.a). Nonvolant animals cannot
vault dispersal barriers as do flying animals, ren-
dering their active dispersal generally less effec-
tive. Despite this, large-scale active dispersal of
large mammals, reptiles, and fishes is frequently
observed. In plants, active dispersal is very rare
and is restricted to the phenomenon of *ballistic
dispersal*, in which seeds are explosively dis-
charged from their base (b) and are dispersed at
most tens of meters.

Passive dispersal is the main, or sole, means of
movement for the vast majority of organisms. In
plants, morphological appendages such as wings
or hairs aid the passive wind dispersal of diaspores
(c); barbs or spines adhere to mobile animals;
fleshy fruits attract frugivores that may digest
the fruit and defecate the seeds (d). Seed dispersal
by volant vertebrates (especially birds and bats),
either external (epizoochory) or internal (endo-
zoocory), is considered an efficient adaptation
for long-distance dispersal.

Passive dispersal is also very frequent among
organisms other than plants. Many marine organ-
(c) Seeds of sticky ileabane (*Duirichia viscosa*). Photo by Avinoam Danin.

(e) Veliger larva of sea snail *Cymatium purtheneum*, a telepanic gastropod (*Rutelidae*). Photo by Rudolf S. Scheltema, used with permission.

(d) Fruits of thorny smilax (*Smilax aspera*). Photo by Avinoam Danin.

(f) Floating island of 14 m long, composed mainly of papyrus, in Lake Malawi, Africa. Photo by Kenneth R. McKay.

(g) Brown tree snake (*Boiga irregularis*) coiled in an airplane's wheelwell in Guam. Photo by Gad Perry.
isms have free-living juvenile stages in the form of cell, eggs, or larvae that are carried passively by ocean currents over large distances. Being concentrated near the water surface, these tiny planktonic propagules (e) are dispersed by the winds that drive the oceanic circulation. Microbes, spores, light seeds, and tiny insects, spiders, and mites constitute the so-called aerial plankton known to be an extremely effective source of colonization of remote islands. Some spiders, for example, climb to an eminence, release silk parachutes into the air, and are subsequently carried off by wind. Dispersal of larger disseminules by winds is usually very local, but strong updrafts and extreme storms can carry them over large distances. Another passive mechanism for dispersal of relatively large organisms is by floating islands or rafts (f), which may disperse entire trees, rodents, and reptiles. However, the most influential dispersal agent at the biogeographic scale—at least during the past 10,000 years—is man. Anthropogenic dispersal has moved many species across the globe (Section IV.A.5), either intentionally or not. The arrival of the brown tree snake to Guam is considered to be unintentional (Section V.B.); years later, a snake was found in an airplane’s wheelwell just before leaving Guam for Japan (g).

III. DETERMINANTS OF DISPERSAL AND COLONIZATION SUCCESS

A. Mechanisms and Adaptations for Dispersal

Most (if not all) organisms are capable of being moved or be transported at some stage during their life cycle, exhibiting a diverse array of dispersal mechanisms (Box 1). Henry Ridley (1930), Sherwin Carlquist (1974), Leendert van der Pijl (1982), and Hugh Dingle (1996) provide many examples of dispersal-determining traits in animals and plants. Overall, organisms vary greatly in their dispersal ability, the capacity for dispersal, which is determined by a complex set of interacting factors. High dispersal ability can be achieved by either active or passive dispersal (Box 1), and it is often associated with low competitive ability and low longevity (Tokeshi, 1999).

B. Components of Colonization Success

Successful colonization necessitates the physical presence of a propagule followed by establishment. Both dispersal and establishment require physiological, morphological, and behavioral preadaptations or adaptations enabling them to disperse far from the source, survive during dispersal, and become established in a remote location. In general, the larger the distance between any two locations, the more dissimilar the environmental conditions. Therefore, long-distance dispersal typically takes propagules to environments that are very different from their usual habitats. Thus, successful colonization in remote locations is generally extremely rare, because it is the product of the very low probabilities of these sequential processes. However, given enough time and multiple attempts, long-distance dispersal may result in long-distance colonization.

1. Preadaptations

Species that often combine high dispersal ability with broad ecological tolerance are those adapted to exploit disturbed or newly created environments and those that inhabit ephemeral or uncertain environments. Such species are thus preadapted for successful colonization outside their species range. MacArthur and Wilson (1967) called these species r-strategists, characterized by small size, early maturity, rapid growth, high fecundity, high dispersal ability, and broad niche. In contrast, K-strategists are better adapted for local competition.
through the efficient use of limited resources (Pianka, 1970). The dichotomous r/K scheme is obviously an oversimplification and the concept is rather controversial (see Williamson, 1996); yet empirical studies generally, but not always, do support the theoretical expectation that r-strategy preadaptations characterize effective colonizers (Brown and Lomolino, 1998; see Section IV.A.3).

2. Physical Access: Dispersal Barriers and Dispersal Routes

Physical access to a remote location depends not only on specific traits that enhance long-distance dispersal (Box 1), but also on the environmental conditions in the area between the source and the deposition location. Crucial to the problem of physical access is that of dispersal barriers, which propagules must overcome during jump range expansion (Section II.C.2). The manner of overcoming barriers defines the dispersal route (see glossary), the other key term in the context of physical access; some routes may be traversed with ease, while others rely upon chance. Examples of dispersal barriers and dispersal routes are given in Section IV.A.3 and in other sections that follow.

It is important to emphasize that dispersal barriers are species-specific; that is, conditions that may prevent dispersal of one species may not affect the dispersal of another. They may even be specific for a particular stage in the species’ life cycle. Dispersal barriers can be physiological, ecological, or behavioral, with organisms subjected to physiological stresses, ecological hazards, or behavioral difficulties in the environments they traverse.

Dispersal barriers become traversable via dispersal routes. George Simpson (1902–1984) made a distinction between corridors, filters, and sweeps routes, which roughly cover the entire range of probabilities of crossing a barrier from highly probable to highly improbable, respectively (Simpson, 1965). Corridor routes allow dispersal of many species from one region to another, conferring a considerable degree of similarity between the biotas of the two sides. Climatic changes and plate tectonics have repeatedly connected previously separated land or water masses, generating dispersal corridors such as those between Africa and Eurasia (Section IV.A.4.c, Fig. 2). The passage through filter routes is more selective than through corridors: some species pass rather easily, while others do not; as a result, each biota represents a biased subset of the other. The Suez Canal between the Red Sea and the Mediterranean Sea (Section IV.A.3) is an example of a filter route. Sweeps routes in which crossing occurs as a chance event at a very low frequency, such as by dispersal on floating islands (Box 1), represent the most improbable way to cross a barrier (Section IV.A.2).

3. Ecological Opportunity:

Postdispersal Establishment

Given that physical access has been attained, nevertheless the probability of successful colonization in remote areas is still rather low. As indicated above, both the physical access and the establishment of a viable population require physiological and morphological preadaptations; often each characteristic is restricted to either dispersal or establishment, and plays little role in the other. The importance of post-dispersal establishment for successful colonization is illustrated by the many examples in Section IV.

Establishment success is determined in part by the ecological conditions at the new site. For passively dispersed species, only the conditions in the particular arrival site matter, while actively dispersed propagules may exhibit habitat choice (also called habitat preference) due to their ability to reach preferred habitats among those locally available. Passively dispersed dormant propagules like seeds and invertebrate eggs may exhibit temporal habitat choice by becoming dormant until conditions favorable for establishment become available. Meeting the physical conditions for establishment may not be sufficient: colonizers may still be excluded by biotic interactions such as competition and predation. Thus establishment also requires traits specific to the recipient community, particularly the resident competitors and predators.

Successful colonization in remote areas requires that colonists be able to reproduce and establish a viable population. Given that long-distance dispersal is likely to bring only one or very few reproductive individuals close together in a remote area, the initial population size of a naturally colonizing species is typically very low. Very small populations are prone to extinction due to demographic, genetic, and environmental stochasticity—the influence of random events on population persistence—which are intensified as population size decreases. These risks may be reduced with repeated immigration to the site, by so-called rescue effects (Section V.B); hence dispersal ability itself is a factor that determines not only arrival but also, in part, establishment probability. Small, newly established isolated populations exhibit reproductive difficulties due to Allee effects such as difficulty in finding a mate. Such effects are obviously graver for obligate sexual reproducers (higher animals and plants) than for asexual or self-
compatible organisms that can reproduce without gametes from another individual. However, the severity of Allee effects is reduced in obligate sexual species by certain reproductive mechanisms, such as sperm storage and delayed implantation, which assure sufficient gametes of both types and by social behavior such as flocking.

C. The Geography of Dispersal

The main issue addressed so far is the effect of dispersal on the geographic distribution of organisms. The complementary issue of geography's effects on dispersal entered the discussion of dispersal barriers and dispersal routes (Section III.B.2). This topic is concerned
with geographic variation in dispersal explained by geographic variation in the external factors affecting dispersal; it typically neglects geographic effects on the dispersal traits themselves. Theoretical models predict that dispersal is favored during colonization but selected against when a population is strongly isolated. The degree of isolation is a relevant factor: (1) when isolation is low, colonization is easily achieved by dispersal. Thus, elevated levels of isolation initially select for greater dispersal abilities; (2) at higher levels of isolation colonization is unlikely, and dispersal is selected against. While the second prediction has been long recognized and is supported by overwhelming evidence, there have been only a few recent tests of the first.

The prediction that very high degrees of isolation select against dispersal is evident in the loss of dispersal ability on oceanic islands, as in many flightless birds, land snails, insects, and plants (see Carliquist, 1974). Reduced dispersal is also evident in organisms in unique isolated habitats as diverse as subterranean caves, mountaintops, deep-sea trenches, thermal vents, and hot springs. The observation of this phenomenon raises an intriguing paradox, which Darwin (1859) explained through countercacting selective forces acting before and after colonization of a remote island. Martin Cody and Jacob Overton (1996) demonstrated the strength of the selection for reduced dispersal potential in weedy, short-lived wind-dispersed plants of inshore islands in Canada. Over the course of just a few generations, island populations of two of three species with sufficiently large sample sizes showed marked reduction in dispersal potential relative to mainland populations or young island populations. Aphids, leafhoppers, crickets, and many other insects have the potential to develop into a normal or flightless adult, depending on the environmental conditions encountered during development (Harrison, 1980).

Habitat fragmentation has substantial effects on dispersal at mesoscales by creating dispersal barriers and disrupting dispersal routes. Recent empirical studies on British butterflies have taken the first step toward experimental investigation of the effect of landscape structure on dispersal-related traits (Van Dyck and Matthysen, 1999). These studies show that the relative investment in the thorax, which mainly contains flight muscles, is higher when patches are >40 km apart; similar selection appears to have occurred in a region where population expansion was most rapid. No differences were found in the relative investment in flight between sites <5 km apart. Since flight is associated with other functional needs of butterflies such as foraging and mate-location, further study is needed to test the relationship between morphological traits and dispersal ability.

IV. METHODOLOGICAL APPROACHES AND CASE STUDIES

A. Observing Dispersal and Colonization in Action

1. Observations of Dispersal

Direct observation is potentially a powerful method of quantifying dispersal. Peter Turchin (1998) provided an extensive review and explanation of the various methods used to track the movements of individuals and to estimate population redistribution in both animals and plants, including records of movement paths and mark-recapture methods. These methods are typically laborious and, from a biogeographic perspective, where large spatial or temporal scale are most important (Section II.C), nearly impossible. The indirect approach, in which dispersal is estimated from its outcome (recruitment), is more feasible. Therefore, all methodological approaches described here are based on indirect evidence. The greater feasibility of indirect methods is a result of the larger size and more conspicuous nature of organisms in relation to their propagules; moreover, it is usually easier to observe a pattern (recruitment) than a process (dispersal). As emphasized in Section III.B, dispersal is meaningless without subsequent recruitment; therefore, indirect evidence of dispersal is indeed important. Nevertheless, it does not tell us much about the dispersal process itself, since the distances dispersed, the number of dispersal events, and the mechanisms of unsuccessful recruitment are unknown. These are important parameters that should be incorporated into biogeographic models in order to enhance their generality and predictive ability.

Data on dispersal events are best obtained by following movements of individuals for which both source and end points are known. Unfortunately, for practical reasons, these kinds of data are relatively rare. In many cases only the end location is known and there are multiple potential sources. Although our knowledge of long-distance dispersal continues to increase, the paucity of direct observations of these dispersal events is one of the largest obstacles to the development of dispersal biogeography as a mature scientific discipline. One possible solution is the development of methods to track individual movements at both small and large spatial
scales. Recent advances in developing molecular and remote sensing techniques promise significant progress in the near future. Note that in genetic studies the term direct evidence is used for evidence of gene flow obtained by comparing the genotypes of recruits and their putative parents (and then applying methods such as paternity analysis) to identify the most probable parent; it would be considered as indirect evidence under the preceding definition. However, for the study of long-distance dispersal this approach is not very valuable; evaluating putative parents becomes impractical as the distance from the offspring increases. In contrast, the so-called indirect genetic approach, which compares the genetic structure between populations, has provided compelling evidence for jump range expansions (Section IV.A.4.b).

There are two types of methods for estimating the dispersal distances when there are multiple possible sources (reviewed in Nathan and Muller-Landau, 2000). The first makes a priori assumptions about the source location (the simplest assumes the nearest possible parent is the source), while the second uses inverse methods to fit a particular dispersal kernel (a functional form of the probability distribution of dispersal distances—see Section IV.C.1). Both methods have been applied to study dispersal at the ecological scale. Since short-distance dispersal governs gradual range expansion, it is also important from the biogeographic perspective. However, using these methods to estimate long-distance dispersal is problematic since the exact position along the dispersal kernel tail is typically very sensitive to small differences in probability—sampling errors are likely to obscure the actual probabilities. Furthermore, the number of endpoints in typically very small (in most cases there is only one), which prohibits inverse modeling techniques. With a better understanding of the process of long-distance dispersal itself, we might expect that mechanistic dispersal models (Section IV.C.1) would help surmount this difficulty.

2. Colonization of Virgin Habitats

Volcanic eruptions that eliminate an entire local biota provide unique opportunities to study patterns of dispersal and colonization. Although there are some well-studied cases of recolonization in terrestrial systems after volcanic eruptions, such as the 1880 eruption of Mt. St. Helens (western North America), the best case studies are on oceanic islands. A well-known example is the island of Surtsey near Iceland that emerged from the sea in 1963. The best example, however, has its roots 80 years earlier: on August 27, 1883, the island of Krakatau, located between Sumatra and Java (Fig. 3a), exploded with tremendous force, obliterating all traces of life. Less than 2 months after the eruption, a Dutch expedition found no evidence for any surviving life form; the first animal, a single tiny spider that had dispersed as aerial plankton (Box 1), and the first plant were detected 9 and 13 months after the eruption, respectively. Later biological surveys documented the rapid re-colonization of the island by various life forms, in a complex successional process, which began with fungi and lichens breaking down the volcanic rock to form soil, enabling seeds to germinate. The possible sources of immigrants were the mainland areas of Java and Sumatra, both approximately 40 km away, and two small islands, 12 and 20 km away (Fig. 3a). More precise identification of the sources for colonization in Krakatau by any current available method is highly speculative. Furthermore, most species were discovered after they had already established populations, with no data on their actual arrival; obviously, there are very few data on arrivals not followed by recruitment. Yet, even with these restrictions, the main insights from this unique long-term classical case study are extremely important; a recent book by Ian Thornton (1996) tells the story of the reassembly of this island ecosystem.

The immigration data collected on Krakatau, albeit incomplete, provides important information on the dispersal process itself. One of the most important modes of dispersal was by sea, whether passively (e.g., fruits and seeds) or actively (e.g., the water monitor Varanus salvator). There was a striking example of rafting (Box 1) in a 20-m² area floating island with 3 to 4 m high palms observed in the archipelago in 1986. Dispersal by air, either passively (spores, light seeds, and other aerial plankton) or actively (butterflies, birds, and bats), was another important mode. Birds and bats served as dispersal agents for plant seeds over short (within the archipelago) and long (from the mainland to the archipelago) distances.

One of the most remarkable patterns observed in the Krakatau system is that of unequal colonization. This pattern was largely clarified for plants when dispersal agents were classified into three primary types (wind, sea, and animals) and examined in three time phases since the eruption (Fig. 3b). Sea-dispersed species were the most important early colonizer but became the least important by the third phase; animal-dispersed species generally showed the opposing trend; the relative importance of wind-dispersed species remained fairly constant in all phases. The mode of dispersal also explains why several kinds of taxa were under-represented: species with large and heavy seeds dispersed either by wind (wings) or by bats and terrestrial mammals (Fig. 3c).
A

B

C

Phase 1

Sea
- Rapid colonization of shore line
- Simple succession

Wind
- Pioneers lose ground

Animal
- Delayed colonization of interior

Phase 2

Sea
- Colonization continues at a reduced pace

Animal
- Colonization rate decreases

Restricted Groups: colonization improbable

Phase 3

Sea
- Habitats restrictions only

Animal
- Large seeded seaborne or terrestrial mammal-spawned species

Animal-dispersed

Wind-dispersed

Sea-dispersed

Number of species

Year
Moreover, some species reached, but did not colonize, the archipelago. Over the period of eight years (1984–1992), 54 species of animals were recorded as new arrivals to Anak Krakatau (a small island in the center of the archipelago, which emerged from the sea in 1928: Fig. 3a): 14 were new to the archipelago, and more than 20 did not become established. The absence of several sea-dispersed plant species is attributed to post-dispersal conditions. Similarly, several species such as the estuarine crocodile (Crocodylus porosus) were seen in Krakatau but have not yet colonized, presumably due to the lack of suitable habitats. The role of establishment conditions is also reflected in the decrease of the colonization rate with species richness.

3. Removal of Barriers

Dissolution of barriers is typically accompanied by species invasion in both directions. Although the term biotic interchange might be interpreted as a balanced exchange, the pattern is often highly asymmetric (Vermeij, 1991). Climatic changes and tectonic plate movements in geological times have repeatedly created and destroyed dispersal barriers, and both kinds of events have had important biogeographic and evolutionary consequences (Fig. 2). Such events include the closure of the Tethyan seaway and the formation of a wide land bridge between Africa and Asia during the Miocene 18 million years ago (mya). Likewise, the reoccurring Bering land bridge and strait have intermittently served as dispersal route between the terrestrial biota of northeastern Asia and northwestern North America, and between the marine biota of the north Pacific and Arctic-Atlantic Basins. Another well-studied example is the current land bridge between South and North America which was formed 3.5 Mya; it provided a dispersal route for a massive passage—the Great American Interchange—of terrestrial organisms, and a dispersal barrier for the marine biota of the Atlantic and Pacific oceans. These examples demonstrate how dispersal and vicariant events can take place simultaneously.

Since the fossil record of mammals is richer than that of other vertebrates, the passage of mammals during the Great American Interchange has been the most thoroughly documented. It is clearly asymmetric, with many more North American taxa passing through the filter and diversifying in South America than doing the converse. In other terrestrial vertebrates, however, the interchange was either roughly balanced (e.g., birds) or biased in the opposite direction—that is, predominantly from south to north (reptiles, amphibians, and even freshwater fishes). Indeed, the differential response to the creation of the land bridge shows large variation also at lower taxonomic levels; in general, most taxa in both continents actually did not expand their ranges to the other continent.

Of the several episodes of biotic interchange that have occurred recently, perhaps the one most intensively studied is that through the Suez Canal, which connects the Mediterranean Sea to the Red Sea. Before the canal was opened for navigation in 1869, only a very few euryhaline species were able to cross the hypersaline lakes that have occasionally connected the two seas (Por, 1978). The creation of the Suez Canal led to an almost unidirectional colonization named Lessopian migration after Ferdinand de Lessop, the designer and chief constructor of the canal. It involves the colonization of the Mediterranean by at least 14 species of worms (Polychaeta), 91 mollusks, 15 crabs, and 44 fish species from the Red Sea. Two main explanations for the unidirectional colonization have been proposed. The first, that species from the Gulf of Suez, the saline arm of the Red Sea at the southern end of the canal, were better preadapted physiologically to cross the hypersaline canal and lakes than were species from the northern end of the Mediterranean. The second postulates that species from the more diverse biota of the Red Sea are superior in competition, defense, or reproduction to species of the impoverished biota of the eastern Mediterranean. A similar explanation has also been raised to explain the asymmetry in the passage of mammals during the Great American Interchange. Comparisons between species that colonize and closely related species that do not colonize show, in general, a tendency for r-selection attributes (Section III.B.1) among colonizers, especially higher dispersal ability and fast growth (Safril and Ritte, 1983).

**Figure 3** Colonization of plant species to the Krakatau archipelago between Sumatra and Java (a). The number of species colonizing the island of Rakata has increased over the period 1883–1989. Species with different primary dispersal mechanisms (sea, animal, and wind) exhibit dissimilar colonization dynamics over three phases of this period (b). A model of island colonization derived from these empirical data (c) incorporates the dispersal mechanisms and temporal phases with a distinction between strand-line (outer circle) and interior habitats, and emphasizes colonization constraints (the lower right figure). Arrow widths are proportional to the cumulative number of species in species year-1 (values are given by each arrow). From Bush and Whittaker (1991) with permission from Blackwell Science Ltd. and Whittaker and Jones (1994) with permission.
4. Natural Range Expansion

a. Recent Range Expansion

Well-documented major range expansions provide a unique opportunity to quantify and model the process of spatial spread at large scales. Charles Elton (1900–1991) and J. Skellam laid the foundations for the dynamic progress recently experienced in this field. Most range expansions at present are related to human activity (Section IV.A.5), but some that are presumably natural range expansion are of particular interest to explore how species expand their range without human-aided transportation (see Hengeveld, 1989; Shigesada and Kawasaki, 1997; Williamson, 1996; also see Plant Invasions).

Range expansion can be very abrupt and may be followed by a rapid contraction, a phenomenon called boom-and-bust (Williamson, 1996). The most remarkable examples are three African locusts (Schistocerca gregaria, Locusta migratoria, and Nomadacris septemfasciata), which usually inhabit limited areas, but which reproduce rapidly, change their morphology and behavior, and disperse in huge swarms to forage over much larger areas under particularly favorable weather and food conditions. Populations of several rodent species such as lemmings (Lemmus spp.) also exhibit marked (cyclic) fluctuations, but due to their lower dispersal capacities, the resulting geographical impact is much more limited. Typically, these outbreaks do not result in expanding the breeding distribution of these species, as occasionally occurs after outbreaks of some bird species such as crossbills (Loxia curvirostra) and Pallas’s sandgrouse (Syrrhaptes paradoxus).

The accidental occurrence of individuals far away from the breeding sites has been documented in many taxa, especially in volant animals. For example, in 1979 I observed an accidental arrival to Eilat (South Israel) of a female long-tailed duck (Clangula hyemalis) whose southernmost breeding and wintering regions in Europe are in Sweden and Northern France, respectively. Fifteen months later, at the very same salt pond, I collected an exhausted shy albatross (Box 1), a rare seabird that breeds on islands near Tasmania and New Zealand. Some accidental vagrants may establish viable populations (jump range expansion), as did the purple gallinule (Gallinula martinica), a colorful waterfowl, in the Tristan da Cunha Islands at the Mid-Atlantic Ridge after years as a vagrant visitor. This tropical species has been reported as a rare vagrant as far north as Nova Scotia in Canada.

Events of boom-and-bust and of vagrant occurrence illustrate three important points: first, that species differ in their dispersal ability (Section III.A, Box 1); second, that the current breeding distributions of these species is not limited by dispersal (Section III.B.2), and third, that postdispersal factors are important determinants of colonization success (Section III.B.3). However, these events are episodic and thus difficult to study. In contrast, some persistent cases of recent broad-scale range expansions that have been documented in detail are important contributions to the study of dispersal and population spread. As in the previous case, the best examples come from highly mobile animals. One well-studied case is the collared dove (Streptopelia decaocto), which since 1900 has expanded its range from Turkey through the Balkans to the rest of Europe. Another is the cattle egret (Bubulcus ibis), which expanded its range in all continents since the late 1800s. It naturally colonized South America from its breeding grounds in Africa by the mid-1930s (jump range expansion) and rapidly expanded throughout the New World. See Sections IV.A.5 and IV.C for a more detailed discussion of range expansion.

b. Holocene Postglacial Spread

The considerable fluctuations in the global climate during the Pleistocene (the geological epoch between 1.8 million years ago [Mya] and 11 thousand years ago [Kya]) have led to several glacial-interglacial cycles. The last glacial period reached its maximum about 18 Kya, when ice sheets covered extensive parts of Eurasia and North America. Species responded in various independent ways, depending on their physiology, life history strategies, and dispersal abilities. Many species were able to track the shift of their most favorable climate and habitats, experiencing range contraction or range shift southward. From about 15 Kya, the climate warmed, the ice retreated, and species expanded their range northward from refugia. Valuable data on this range expansion during the Holocene (the last 11,000 years) come from insect exoskeletons, plant macrofossils and pollen cores collected from many localities.

The postglacial expansion of trees in North America and Europe has been studied extensively based on fossil pollen cores from lake sediments across a wide geographical range, dated by radiocarbon (14C) analysis (Davis, 1976; Huntley and Birks, 1983; Delcourt and Delcourt, 1987; Birks, 1989). The results are often expressed as isochrone maps in which contours join sites where similar pollen-stratigraphical events occurred at the same time and at rates of range expansion during a given period (Table 1). These interpretations, however, should be made with care since pollen of many tree species are effectively dispersed over large distances;
pollen data also contains a nonlocal component. Estimation of this component is possible through data on other plant remains (seeds, fruits, leaves, twigs, wood), so-called plant macrofossils, that are less likely to be transported over large distances. The relatively few studies in which macrofossils have been well documented showed that the main conclusions based on pollen data alone were fairly robust. Genetic studies also strongly support these conclusions.

The paleoecological pollen record shows an individualistic response of tree species, in the timing, direction, and speed of their spread. It also shows that species were able to cross dispersal barriers as large as the North American Great Lakes, the English Channel, and the Baltic Sea. In particular, it provides compelling evidence that range expansion of trees can be very rapid with estimated average rates of spread of a few hundred meters per year (Table 1) and more. These estimated rates greatly exceed the rates expected from the observed dispersal rates of these species, collected at ecological and temporal scales (Section II. A). The discrepancy has been a long-standing puzzle for plant ecologists, and termed Reid's paradox after Clement Reid who described it for British oaks. Even more puzzling than the rate of expansion of oaks—whose acorns may occasionally be dispersed over large distances by birds—is the rapid rate of expansion of woodland herbs that mostly propagate clonally and possess no obvious mechanism for long-distance dispersal (Cain et al., 1998). The discussion of this paradox has recently been revived because of the need to predict biological responses to human-induced climate warming. In fact, evidence on population expansions and retractions from butterflies (Parmesan, 1996; Parmesan et al., 1999) and breeding birds (Thomas and Lennon, 1999) indicate that global climatic warming is already reflected in considerable poleward shift of species' distributions. Consequently, scientists have developed new mathematical models for seed dispersal that incorporate the long-distance component as well (Section IV.C).

Another recent advance in the study of postglacial range expansion follows the progress in molecular genetics that enables assessment of present genetic structure; that is, in turn, used to infer past colonization events (see Section IV.B). Godfrey Hewitt (1999) summarized the results of studies dealing with the postglacial colonization of the European biota (both animals and plants) from three main refugia at the eastern (Balkans), central (Italian), and western (Iberian) peninsulas of southern Europe. He proposed three main expansion patterns, which were termed, after their exemplars, “hedgehog,” “grasshopper,” and “bear” (Fig. 4c): the hedgehog (Erinaceus spp.) expanded with three genomes advancing north from the three refugia; the grasshopper (Chorthippus parallelus) colonized mostly from a eastern refugia, with other genomes blocked at the Pyrenees and Alps; the bear (Ursus arctos) colonized from the east and the west, and with its central refugium blocked at the Alps.

c. Spread of Early Hominids

Modern man (Homo sapiens sapiens) is currently the species with the most widespread geographical distribution, and other globally distributed species are mostly human commensals. The dispersal biogeography of humanity is a very intriguing subject; it is also an excellent case study that demonstrates many essential aspects of dispersal biogeography. The following discussion briefly summarizes the patterns of dispersal and range expansion of early hominids, mostly based on Roy Larick and Russell Ciochon (1996) and Ian Tattersall (1997); note that there still is some argument over this topic.

The ancestral source location of early hominids is considered to be the plains of the eastern Rift Valley in Africa (now Ethiopia, Kenya, and Tanzania) where fossils of the first, woodland-dwelling, hominids were found (Australopithecus anamensis, 4.2–3.9 Mya, and Australopithecus afarensis, 3.0–3.9 Mya). The cooling period during the Middle Pliocene (3.0–2.4 Mya) marks the emergence and dispersal of the genus Homo. The fossil record shows that during this period many mammal species shifted from forests to open landscapes, including the earliest Homo species, omnivorous scavengers who lived in open habitats between 2.5 and 2.0 Mya. One of these species, Homo ergaster (closely related to Homo erectus found in Java) had long torso and limbs, narrow hips, a large brain, reduced dentition, and elementary stone technology. The findings of many bone and stone refuse near lakes and streams suggest that early hominids congregated within catchment basins and utilized local resources until they were exhausted, and then moved to another catchment, and so on. All of these attributes served as morphological and behavioral preadaptations (Section III.B.1) for successful dispersal to and colonization of similar habitats outside Africa during the Pliocene (Fig. 5). At the same period some large bovids dispersed from Africa to Eurasia; perhaps the two groups emerged and dispersed together, with the bovids being hunted or scavenged by Homo.

The earliest fossil of Homo (species indeterminate) outside Africa, dated 1.9 Mya, was found in Longguo (China). Early Homo erectus skulls dated 1.8 and 1.7
### TABLE I

Observed Rates of Range Expansion, Arranged in Decreasing Order within Each Taxonomic Group

<table>
<thead>
<tr>
<th>Species</th>
<th>Period</th>
<th>Region</th>
<th>Average expansion rate (km y⁻¹)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mammals</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European rabbit</td>
<td>1870–1900</td>
<td>Australia</td>
<td>54</td>
<td>Myres, 1970</td>
</tr>
<tr>
<td>Oryctolagus cuniculus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muskrat</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ondatra zibethicus</td>
<td>1905–1947</td>
<td>Czechoslovakia to east-southeast</td>
<td>23</td>
<td>Andow et al., 1990</td>
</tr>
<tr>
<td></td>
<td>1905–1948</td>
<td>to north</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1932–1954</td>
<td>France</td>
<td>0.9–4.6</td>
<td>Andow et al., 1990</td>
</tr>
<tr>
<td></td>
<td>1923–1938</td>
<td>3 localities</td>
<td>5.5–6.7</td>
<td>Andow et al., 1990</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Finland 4 localities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grey squirrel</td>
<td>1965–1981</td>
<td>Britain</td>
<td>7.7</td>
<td>Okubo et al., 1989</td>
</tr>
<tr>
<td>Sciurus carolinensis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red deer</td>
<td>1861–1910</td>
<td>New Zealand</td>
<td>1–1.6</td>
<td>Shigesada and Kawasaki, 1997</td>
</tr>
<tr>
<td>Cervus elaphus</td>
<td>1910–1920</td>
<td>U.S.</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>California sea otter</td>
<td>1938–1972</td>
<td>U.S., California to south</td>
<td>3.1</td>
<td>Lubina and Levin, 1988</td>
</tr>
<tr>
<td>Enhydra lutris</td>
<td></td>
<td>to north</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>Himalayan uahr</td>
<td>1936–1966</td>
<td>New Zealand</td>
<td>0.7</td>
<td>Shigesada and Kawasaki, 1997</td>
</tr>
<tr>
<td>Hemivirusus jemlahicus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cattle egret</td>
<td>1950–1970</td>
<td>S. America to N. America</td>
<td>106</td>
<td>Van den Bosch et al., 1992</td>
</tr>
<tr>
<td>Budorcinus ibis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>House sparrow</td>
<td>1864–1888</td>
<td>U.S.</td>
<td>66</td>
<td>Okubo, 1988</td>
</tr>
<tr>
<td>Passer domesticus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European Starling</td>
<td>1900–1915</td>
<td>U.S.</td>
<td>11</td>
<td>Okubo, 1988</td>
</tr>
<tr>
<td>Stripes vulgaris</td>
<td>1915–1950</td>
<td>U.S.</td>
<td>51</td>
<td>Okubo, 1988</td>
</tr>
<tr>
<td>Collared dove</td>
<td>1928–1963</td>
<td>Europe</td>
<td>4+</td>
<td>Van den Bosch et al., 1992</td>
</tr>
<tr>
<td>Streptopella decanoto</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardopodius mexicanus</td>
<td>1962–1979</td>
<td></td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Insects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apis melifera scutellata</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rice water weevil</td>
<td>1976–1986</td>
<td>Japan to northeast to west</td>
<td>28–470</td>
<td>Andow et al., 1993</td>
</tr>
<tr>
<td>Lissorhoptrus oryzophilus</td>
<td></td>
<td></td>
<td>47–250</td>
<td></td>
</tr>
<tr>
<td>Cabbage white butterfly</td>
<td>1860–1883</td>
<td>Canada, Quebec</td>
<td>15–170</td>
<td>Andow et al., 1990</td>
</tr>
<tr>
<td>Pieris rapae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oulema melanopus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gypsy moth</td>
<td>1900–1916</td>
<td>N. America</td>
<td>9.5</td>
<td>Liebhold et al., 1992</td>
</tr>
<tr>
<td>Lymantria dispar</td>
<td>1916–1965</td>
<td></td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>Vascular plants*</td>
<td>1966–1990</td>
<td></td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Alder</td>
<td>8.5–5.5 Kya</td>
<td>British Isles</td>
<td>0.33</td>
<td>Birks, 1989</td>
</tr>
<tr>
<td>Alnus glutinosa</td>
<td>8.5–7 Kya</td>
<td></td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7–6 Kya</td>
<td></td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

*continues
Mya were found in Mojokerto and Sangiran (Java, Indonesia). Thus, the departure from Africa occurred roughly 2 Mya. During this period there were wide land bridges between Africa and Eurasia at the Sinai Peninsula, and at the straits of Bab-el-Mandab between the Red Sea and the Indian Ocean. Probably both served as dispersal routes for bovids and Homo; fossils and stone-tools dated 1.96 to 1.4 Mya were found along the possible paths to China and Indonesia from 'Ubeidiya and Erq-el-Ahmar (Israel), and through Dmanisi (Georgia), Riwar, and Pabbi hills (Pakistan). Before the findings from eastern Asia, the consensus about the earliest hominid "out-of-Africa" dispersal was that it occurred roughly 1 Mya later. Since this early departure from Africa 2 Mya, early hominids successively dispersed from Africa to Asia, and later to Europe, where the earliest Homo fossil found (in Atapuerca hills, Spain) is dated 0.78 Mya. Compelling evidence shows that Homo sapiens, our own species, also originated in Africa from where it eventually colonized a much broader range than any of its extinct predecessors.

5. The Spread of Alien Species

Humans can facilitate the physical access of organisms from one region to another by breaking dispersal barriers or by providing dispersal routes. Species that have colonized nonnative regions due to human activities are called alien species, and those that have been deliberately transferred are called introduced species. Many closely related terms, including adventive, exotic, and invasive species, have been used confusingly in the literature. Though examples are numerous, the spread of alien species represents merely a small, though growing, fraction of the total number of incidents in which humans have intentionally or accidentally facilitated physical access. Further, as a general rule—termed propagule pressure by Mark Williamson (1996), which is principally equivalent to the rescue effects (Section V.B)—single or very few propagules are unlikely to invade. Increasing the number of propagules (e.g., by repeated introduction attempts) increases the chance of an alien species to establish and spread. The literature provides a wide array of case studies of various taxa across the entire globe, from zebra mussels and starlings in North America to rabbits and myxoma virus in Australia (Table 1). I have chosen the classical case study of the muskrat (Ondatra zibethicus) to illustrate the implications from such events to dispersal biogeography; the chapter Plant Invasions provides more examples of the spread of alien plants.

The muskrat is a large semiaquatic rodent native to North America (Fig. 6). In 1905, five muskrats escaped from a fur-breeding farm near Prague in Czechoslovakia. Within 50 years, their descendants and others inhabited large parts of the European continent (Fig. 6), despite eradication programs initiated due to muskrat
FIGURE 4 Postglacial recolonization of European biota. The physical geography of Europe (a) is characterized by mountain ranges running from the Pyrenees eastward through the Alps to Caucasus, marked as black regions (>2000 m altitude). By the end of the Pleistocene ice age (b) extensive ice sheets (hatched) covered northern Europe, and a permafrost plain extended southward to the mountain ranges. The rapid postglacial expansion out of the three southern peninsular refugia (c), as deduced from molecular data and fossil evidence, followed three main patterns as illustrated by four examples in each column (see text for details). From Hewitt (1999).

damage to roads, dikes, and agriculture. Expansion of the muskrat appeared to be rapid and of constant rate (see Section IV.C.2 for quantification of this case study). At finer scales, however, expansion was fairly variable in space and, to a lesser extent, in time (Table I). Overall, the expansion was closely related to the geographical structure of the species’ favorable wet habitats. River valleys, one kind of favorable habitat, provided dispersal routes, while mountain ranges acted as dispersal barriers, forming a starlike expansion pattern. Interestingly, favorable habitats do not necessarily increase but may also decrease the rate of spread, as happened in regions with abundant wet habitats where muskrats settled and built up huge populations. Understanding the interplay between the processes of dispersal and of population growth is critical to the understanding of range expansion (Section IV.C). Temporal changes should also be taken into account, since expansion rates were lower in dry years when the species’ favorable habitats were more restricted.

B. Inference from Current Distributions

Early biogeographers have frequently compared species composition of different regions to evaluate their biogeographic affinities. This analysis, later scrutinized to include various quantitative methods, has revealed a common nonrandom pattern called nestedness, in which small biotas contain a nonrandom subset of the species in richer ones (Patterson and Atmar, 1986). Found in various taxonomic groups, nestedness is generally attributed to differential extinction and colonization,
or a combination of both, though other factors may also be responsible. Differential extinction is arguably the prevalent mechanism of nestedness, especially in fragmented biotas. However, several biotas assembled by colonization also exhibited nested patterns. Some studies suggested that nestedness reflects differential dispersal ability coupled with geographical isolation. This suggestion can be tested by ranking the subsets (e.g., islands) according to their distance from the presumed source (e.g., mainland), and by examining the correlation between individual species contributions to the observed nestedness and species dispersal ability. Ronen Kadmon (1995) used this approach for woody plant species in seven man-made islands in a reservoir in eastern North America and confirmed that strong nestedness was attributable to geographic isolation and was correlated with dispersal ability. Wind-dispersed species showed no evidence of nestedness and species without particular adaptations for long-distance dispersal showed the strongest nestedness. An equivalent pattern is seen in much broader spatial scales, where high dispersal abilities are associated with low levels of endemism, for example in marine benthic invertebrates with teleplastic larvae such as the gastropods shown in Box 1 (Scheltema, 1971). Similarly, the biotas of isolated islands such as Krakatau (Section IV.A.2), Galápagos, and Hawaii, are overrepresented by taxa having long-distance dispersal mechanisms (but see Section III.C), a characteristic termed disharmony to denote the lack of biotic elements that are conspicuously present on continents (Carlquist, 1974).

During the 1970s and early 1980s, the approach used in the study of historical biogeography shifted dramatically from explanations of the unique responses of individual groups to a quantitative and more rigorous analysis of general patterns. This progress was achieved mainly by the scholars of vicariance biogeography (Section I; Myres and Giller, 1988) who, based on Hennig's phylogenetic systematics and Croizat's panbiogeography, developed the area cladogram as a general technique for the analysis of biogeographic patterns (see Vicariance Biogeography). Area cladograms are branching diagrams representing history, similar to traditional phylogenetic trees, but with the geographic areas replacing taxa at the terminal tips. Correspondence between phylogeny and independently derived geological history is taken to imply vicariance; dispersal is generally invoked to explain lack of correspondence. Congruence among area cladograms of different taxonomic groups from the same region more strongly implies a vicariant event and may be used to infer geological history in the absence of independent data.
Juan Morrone and Jorge Crisci (1995) classified the historical biogeography methods into five groups: dispersalist, phylogenetic biogeography, panbiogeography, cladistic biogeography, and parsimony analysis of endemicty. The methods of cladistic biogeography (see Vicariance Biogeography) are the most diverse and most widely used; important ones include component analysis, Brooks parsimony analysis, and reconciled trees (Morrone and Crisci, 1995). These approaches construct an area cladogram assuming vicariance as a null hypothesis. An alternative interpretation is presented in the book edited by Warren Wagner and V. A. Funk (1995) where simple area cladograms were produced for 25 taxonomic groups, including insects, spiders, birds, and flowering plants, of the Hawaiian Islands. The cladistic analysis was undertaken using various attributes, including morphological and anatomical characters, banding patterns on chromosomes, and DNA sequences from chloroplast, mitochondria, and nuclear ribosomes. Area cladograms were compared with several hypothetical patterns that represent historical scenarios that putatively generated the observed patterns, based on the geological history of the islands. The most frequent pattern was of dispersal from older to younger islands; several patterns were interpreted either as recent colonization from the mainland or as vicariant events within the archipelago. A book on the historical biogeography of Southeast Asia edited by
Robert Hall and Jeremy Holloway (1998) provides another exemplary integration of ecological and evolutionary processes that account for the impact of dispersal and vicariance along with a profound analysis of the complex geological history of this region.

Marrone and Crisci (1995) classified Kärre Bremer's (1992) ancestral areas method as a dispersalist approach. This method is aimed at identifying the ancestral area of a group from the topological information of its cladogram by estimating the likelihood that different areas were part of the ancestral area for the group. More recently, Fredrik Ronquist (1997) developed the dispersal-vicariance analysis (DIVA), which reconstructs the ancestral distributions of the species in a given phylogeny without any prior assumption about the form of area relationships and allowing dispersal and extinction. The optimal ancestral distributions are those that minimize the number of dispersal and extinction events required to explain the pattern. The number of dispersal events is underestimated because similar distributions between species are interpreted as sharing vicariant history, though this is not necessarily the case. Nevertheless, this method disentangles some important methodological and conceptual restrictions of alternative methods and is so far the biogeographic method that most realistically incorporates the effects of dispersal. Gary Voelker (1999) has recently applied the two methods to reconstruct the ancestral distributions and the direction of dispersal events in the pipits (Anthus spp.), small songbirds taxonomically related to the wagtails that occur in all continents. The results show that 16 dispersal events are required to explain the biogeographic pattern of the 40 species analyzed. The reconstructed map of colonizations from the postulated ancestral area in central Asia (Fig. 7) shows broad intercontinental movements, suggesting that dispersal played an important role in the distribution of this group.

C. Modeling

Mathematical models are useful tools for exploring the role of dispersal in the process of range expansion and for predicting expansion dynamics. A comprehensive study from various fields such as genetics, epidemiology, and biological control has led to the development of highly diverse mathematical approaches of varying complexity to modeling movement and spatial popula-

---

**FIGURE 7** Major intercontinental and continent-to-island colonizations (arrows) by pipits (Anthus) form a postulated ancestral area in central Asia (large circle). The approximate time of each colonization event is given either in million years (numbers close to the arrowheads) or is defined as Recent (single species that have dispersed onto another landmass). From Voelker (1999) with permission.
tion dynamics. These models include dispersal models that focus on the movement alone and reproduction-dispersal models that incorporate dispersal with local population growth to model the spatial spread of populations. The rich literature on this subject was recently reviewed and summarized by Turchin (1998, p. 73), who concluded with the following prognosis: "the theory is well developed, and appears to be ready to incorporate whatever quantitative information about movement we can provide." Surprisingly, biogeographers have generally overlooked this progress, despite its clear relevance and potential contribution to the field. A brief overview of some important dispersal and expansion models follows; details can be found in Okubo (1980), Durrett and Levin (1994), Mollison (1995), Shigesada and Kawasaki (1997), and Turchin (1998).

1. Dispersal Models

Dispersal models include phenomenological (empirical curve fitting), diffusion, individual-based, and mechanistic (transport) models. The basic objective of these models is to estimate the probability distribution of dispersal distances, which is often given in a standardized functional form called the dispersal kernel (Nathan and Muller-Landau, 2000). A classic dispersal model used in biogeography is the stepping-stone model of MacArthur and Wilson (1967), which assumes that a disseminule moves in a constant direction from one island to another and has a constant probability \( m \) per unit time of landing (cessation of movement). Mathematically, these assumptions lead to a widely used phenomenological model—the negative exponential function—in which the probability to disperse to distance \( x \) decreases with \( x \) at a rate equal to \( \ln(m) \). Other widely used phenomenological dispersal models follow the Gaussian distribution and the inverse power law. The Gaussian models usually underestimate observed dispersal data, which almost universally exhibit a leptokurtic distribution (Section II.A). The inverse power law has the problem of singularity (infinite density) at zero and hence cannot be used as a dispersal kernel.

Another set of dispersal models that has been developed recently and is likely to prove beneficial to dispersal biogeography is that of mechanistic models, which simulate the process by a set of mathematical expressions that mimic the relationships and joint effects of the principal operative factors. For example, mechanistic models of seed dispersal by wind typically incorporate factors such as the height of release, the seed free-fall velocity, and wind velocities (Nathan et al., 2000). While complicated to develop and parameterize, these models have the decisive advantages of actually predicting, rather than merely fitting, changes in disseminate densities with distance, and of potential generalization to other systems. The task of applying and testing models of long-distance dispersal is inherently more complicated than for short-distance dispersal. However, recent technical developments with miniature transmitters, radar, and other remote sensing techniques, as well as molecular markers, are promising a breakthrough in the near future. Important attempts to model the mechanism of long-distance dispersal has already been made for the wind dispersal of aerial plankton (Westbrook and Isard, 1999) and tree seeds (Greene and Johnson, 1995), based on the knowledge of atmospheric dynamics. Similarly, knowledge of sea currents and the winds that drive them can be used to predict the trajectories of passively dispersed marine biota (Jokiel and Martinelli, 1992). Further, dispersal of seeds by birds may to a certain degree be predictable with knowledge of gut retention time and direction and velocity of bird flight (Sun et al., 1997).

From the biogeographic perspective, special attention should be paid to the tail of the distribution, since long-distance dispersal is disproportionately important in range expansions. Mathematically, it has been shown that kernels differing significantly in their tails yield dramatically different rates of population spread, even if they have the same mean. Recent recognition of the need to focus on long-distance dispersal has led to the development of several alternative "fat-tailed" phenomenological and mechanistic dispersal models, as emphasized in Shigesada and Kawasaki (1997), Turchin (1998), Clark et al. (1999), and Nathan and Muller-Landau (2000).

Turchin (1998) summarized the relationships between dispersal models and the types of population spread, focusing on the tail of the dispersal kernel. If the tail of the dispersal kernel drops off exponentially or faster with distance (for example, Gaussian tails that are predicted by simple diffusion), then we should observe linear spatial spread in which the range is expanded at a constant rate during each time step. If the tail drops off slower than the exponential tail, as in the so-called "fat-tailed" kernels, the type of spatial spread depends on whether the variance of the kernel is finite or infinite. In both cases the rate of spatial spread accelerates (i.e., increases with successive time steps), but the former reaches an asymptote of a constant expansion rate, while the later continues to increase. Thus, fat-tailed dispersal kernels lead to jump range expansion, while exponential and thinner tails lead to gradual range expansion (Section II.C). Since data on long-
distance dispersal is rare, the assumptions of the characteristics of the distribution's tail cannot yet be tested; at the same time it is clear that extrapolating beyond the observed range is questionable.

2. Reproduction-Dispersal Models
The dispersal models that have been discussed are useful tools to understand the general relationships between dispersal and spatial spread. Reproduction-dispersal models, however, are more realistic since they consider the effects of birth and death during the process of range expansion. The simplest reproduction-dispersal models are reaction-diffusion models; and the simplest of those, introduced to biologists by Skellam (1951), model dispersal as random diffusion and local population growth as exponential increase,

\[
\frac{dN}{dt} = rN + D \left[ \frac{\partial^2 N}{\partial x^2} \right]
\]

where \( N \) is the population size, \( t \) is time, and \( x \) is the distance from the source; the two parameters are \( D \), the diffusion coefficient, and \( r \), the intrinsic rate of population growth. Earlier, Ronald Fisher (1937) used a nonlinear variant of this equation in which the unbounded exponential growth \( (rN) \) is replaced by asymptotic logistic growth \( (rN(1 - N/K)) \), where \( K \) is the equilibrium population size (or carrying capacity). As shown by Fisher, this model predicts, as does Skellam's model, a constant asymptotic front velocity equal to \( 2 \sqrt{rD} \). As discussed in the previous section, this type of gradual range expansion does not incorporate the effect of long-distance dispersal events. Tests of reproduction-dispersal models against observed expansion patterns, starting from Skellam's (1951) pioneering study, have frequently showed that the simple reaction-diffusion approach cannot provide a general description for range expansion. For example, David Andow, Peter Kareiva, Simon Levin, and Akira Okubo (1993) found approximately linear rates of spread in two case studies and accelerating rates of spread in two others. Even for the cases that show linear rates, such as that of muskrats (Section IV.A.3), there was significant variation in space and time. Expansion rates of muskrats were lower in the first 2 to 3 years, possibly because of Allee effects (Section III.B.3), and the average rates range from 0.9 to nearly 80 km/yr. Thus, reaction-diffusion models are best used as a primary null analysis of expansion patterns, followed by more complex approaches, including stratified diffusion, interacting particle systems, and spatial contact models. These models incorporate both short- and long-distance dispersal and thus are able to predict patterns of jump range expansion as well.

V. BIOGEOGRAPHIC AND EVOLUTIONARY CONSEQUENCES OF DISPERAL

Dispersal, evolution (adaptation and speciation), and extinction are the fundamental processes in biogeography: together they determine the geographical distribution of organisms. Dispersal biogeography focuses on the effects of dispersal on evolution and on speciation, both within a lineage and for other lineages. The following section outlines some principal relationships; more details are given in related articles on speciation and extinction in this encyclopedia.

A. Speciation

Speciation is the process by which new independently evolving lineages (species) arise from a single ancestral lineage. In general, efficient dispersal between populations (patches) increases the rate of gene flow between them and hence tends to reduce genetic differentiation and to restrain speciation. In contrast, inefficient dispersal between existing populations leads to geographic isolation and tends to increase genetic differentiation and to promote speciation. This type of speciation, called allopatric speciation, is considered to be the most frequent one. Jump range expansion occurs with very low probability (Section II.C.2) and is thus unlikely to be followed by additional dispersal; this isolation thereby causes rare dispersal events (in this context also called founder events) to be relevant to allopatric speciation. The adaptive radiation of Darwin's finches in the Galápagos Islands (Grant, 1999) is one of the many examples of allopatric speciation that has resulted from dispersal events. However, as discussed in Section I, a dispersal barrier may divide a previously continuous distribution, hence allopatric speciation may also result from a vicariant event.

A second way in which dispersal affects speciation requires the opposite conditions. The breakdown of a dispersal barrier, and the following biotic interchange (Section IV.A.3), brings organisms to dissimilar environments and thus opens new chances for diversification. The colonization of North American mammals during the Great American Interchange (Section IV.A.3) was followed by extensive diversification in South America. In the same manner dispersal events
cause speciation: they affect the genetic structure within the species range, sometimes to the point that the populations are evolving independently, and can be considered as distinct species. Because there is a continuum of genetic structure and evolutionary independence, the final categorization of one or more species is never completely clear. Theoretical models and molecular studies have shown clear effects of long-distance dispersal on the spatial structure of the genetic variation within a species during the postglacial expansion of European and North American biota (Section IV.A.4.b).

As predicted by models, jump range expansion well ahead of the "normal" advancing wave generates large patches of low genetic variation, due to the so-called founder effects; consequently, genetic diversity decreases from south to north.

B. Extinction

Limited dispersal is associated not only with higher rates of speciation (Section V.A) but generally also with higher rates of extinction. This phenomenon has been observed at various spatial scales. For example, the fossil record shows that gastropods with planktonic larval dispersal, typically involved with long-distance dispersal (Box 1), tend to have larger geographical ranges and to persist for longer periods of time than those with nonplanktonic larvae, which are dispersed locally.

A similar phenomenon is observed at regional scales, where populations with higher rates of immigration persist longer due to increased population size and higher levels of genetic variability, which counteracts the deleterious effects of inbreeding. This so-called rescue effect has been observed in a variety of organisms. The very similar notion of source-sink dynamics attributes a longer persistence time to populations in unfavorable sink habitats (where death rates exceed birth rates) due to a rescue effect from populations in favorable source habitats (where birth rates exceed death rates). Another relevant term, mass effect, emphasizes the contribution of rescue effects to the breeding stock.

The regional framework of populations subjected to local extinction and recolonization is treated in meta-population models in which dispersal plays a critical role (Hanski, 1999) (see Metapopulations).

While dispersal tends to reduce the probability of extinction of the same species, dispersal, as immigration, may have the opposite effect on other species inhabiting the recipient site. Around 1950, the brown tree snake (Boiga irregularis) of Australia and Southeast Asia was accidentally introduced to the Pacific island of Guam (Rodda et al., 1999). This fairly large (up to 3 m) snake, which eats bird eggs from nests either on the ground or in trees, caused the extinction of most of the native birds and the decline of almost all other native terrestrial vertebrates. Despite this and other spectacular examples of massive extinction caused by colonization of a given species, most colonization events appear to have minor consequences. The consequences of colonization of one species to the persistence on other species depend on the species' biology and on postdispersal processes; however, dispersal is the ultimate cause of such extinctions, since it determines whether the prerequisite species is present. Though the particular dispersal mechanisms should not matter to the severity of such incidents, it is noteworthy that in most documented cases dispersal was related to human activity.

VI. APPLIED DISPERSAL BIOGEOGRAPHY

Biogeography in general, and dispersal biogeography in particular, has an important role in managing the world's biodiversity: the processes currently threatening the world's biota typically act on large scales, and the movement (or lack thereof) of organisms is a key process. Dispersal is strongly involved with issues such as invasions, introductions, fragmentation, climatic changes, design of nature reserves, and control of gene flow; it is hard to imagine any issue of applied biogeography that is not related to this process. Thorough discussions of these practical aspects of biogeography can be found in Williamson (1996), Brown and Lomolino (1998), and Spellerberg and Sawyer (1999); this section outlines some major principles of applied dispersal biogeography.

Applied dispersal biogeography deals with all aspects of dispersal that are involved with the management of the earth's biota and aims to predict the causes and the consequences of dispersal events. The mathematical models described above (Section IV.C) are efficient tools on small scales. Although our ability to predict long-distance dispersal has significantly improved recently, much still remains to be done, especially the testing of these predictions. Long-distance dispersal has a substantial stochastic component, which will always remain somewhat unpredictable. However, since most long-distance dispersal events with harmful consequences are related to human activity, our predictive ability of anthropogenic arrivals may be better than for natural arrivals; even unintentional dispersal as that of
the brown tree snake to Guam (Section V.B, Box 1) is to some degree predictable and controllable. As emphasized earlier, the variation among organisms in dispersal ability is enormous (Box 1); responses to dispersal barriers and dispersal routes also vary considerably (Section III.B.2). Mathematical models, especially mechanistic ones (Section IV.C.1), can be used to predict dispersal of particular species using knowledge of the species-specific dispersal traits (Box 1) and the effect of environmental factors on dispersal (Sections III.B.2 and III.C).

The invasion of muskrats to Europe (Section IV.A.5) illustrates these points. The muskrat’s arrival was related to human activity and the animal’s escape was controllable; expansion rates are predictable through simple mechanistic models based on dispersal and reproduction (Section IV.C.2); the variation in expansion rates was related to habitat-related factors (Sections III.C and IV.A.5) and was therefore also predictable. The large number of escapes and the initial slow rates of expansion demonstrate the importance of propagule pressure (Section IV.A.5) and probably of Allee effects (Section III.B.3) as well. Attempts to control this expansion—mostly extensive trapping campaigns—have failed in continental Europe, but have succeeded in Britain and Ireland. Mark Williamson (1996) suggested that the relatively early response in the British Isles, before the muskrats spread too far and built overly large populations, along with presumably inferior climatic matching that would have led to lower reproduction rates, are possible reasons.

VII. SYNTHESIS

The importance of dispersal in determining the geographic distribution of organisms has been emphasized by Darwin and other dispersalists. These views have been seriously questioned by their opponents, who argue that vicariant events are the principal determinant of distributions. Clearly, both processes occur and their effects vary between species, in space, and in time. The term dispersal biogeography is currently restricted to the arguments of extreme dispersalists concerning long-distance dispersal and disjunct distributions. The broader and more fruitful usage of this term adopted here would be the study of the relationship between dispersal the geographical distribution of organisms.

Dispersal is a widespread phenomenon occurring in nearly all organisms. There is a high variability of dispersal ability in organisms, due to differences in their physiological, morphological, and behavioral traits and responses to environmental conditions. Dispersal is a key biological process at the ecological, evolutionary, genetic, and biogeographic levels, which differ in spatial and temporal scales. Long-distance dispersal, disproportionately important at the biogeographic level, is typically rare and very difficult to measure; it contains a large stochastic component but in many cases is a result of a specific mechanism and hence can be predicted. If followed by successful establishment, long-distance dispersal can lead to abrupt long-distance colonization jumps. Short-distance dispersal is overwhelmingly important at the ecological level; at the biogeographic level, multiple short-distance dispersal events over a relatively long time, if followed by establishment, can lead to gradual range expansion. Patterns of speciation and extinction are strongly affected by dispersal, demonstrating the broad biogeographic consequences of this process, which itself is also affected by geography.

The greatest methodological challenge for dispersal biogeography today is to provide direct evidence, especially of long-distance dispersal. Indirect evidence is based on recruitment and therefore incorporates the obscuring effects of postdispersal processes. Yet there are various ways to infer from indirect evidence to dispersal biogeography, and the case studies described earlier have provided many valuable insights for this endeavor. These case studies include colonization of virgin habitats such as oceanic volcanic islands, observations of the biotic consequences of barrier removal and of natural and human-related range expansions, analyses of current distributions, and modeling of spatial spread. Models are extremely important for providing a better understanding of dispersal, therefore enabling prediction; they thus constitute a promising tool for implementing the principals of dispersal biogeography for maintaining biodiversity.

Acknowledgments

I am pleased to acknowledge the financial support of the National Science Foundation, the Andrew Mellon Foundation, and the Schroeter Foundation of the Hebrew University of Jerusalem. I am grateful to Simon Levin, Mark Williamson, Kai Chan, Helene Muller-Landau, Nechama Ben-Elihu, Gad Perry, Yossi Heller, and Itlana Tchernov for helpful comments that considerably improved the article.

See Also the Following Articles

BIOGEOGRAPHY, OVERVIEW • CONTINENTAL DRIFT, ECOCOLOGICAL EFFECTS OF • ISLAND BIOGEOGRAPHY • LIFE
HISTORY, EVOLUTION OF • MIGRATION - PLANT INVASIONS • SPECIATION, PROCESS OF • SPECIATION, THEORIES OF • DISPERAL BIOGEOGRAPHY

Bibliography


DISTURBANCE, MECHANISMS OF

Frank W. Davis* and Max Moritz†
*University of California at Santa Barbara and †California Polytechnic State University

I. Disturbance Ecology
II. Disturbance and Biodiversity
III. Mechanical Disturbances
IV. Physico-chemical Disturbances
V. Disturbance by Herbivores and Predators
VI. Interactions among Disturbance Mechanisms
VII. Humans as Agents of Disturbance

GLOSSARY

disturbance  Relatively discrete event in time that disrupts ecosystem, community, or population structure and changes resources, substrate availability, or the physical environment.
disturbance regime  The collective spatial, temporal, physical, and ecological characteristics of a disturbance process operating in an area.
resilience  Measure of a system's ability to recover to its original state when subjected to disturbance.
resistance  A measure of system's ability to remain unchanged when subjected to a disturbance.

DISTURBANCES in ecological systems promote characteristic patterns of environmental heterogeneity and regulate ecosystem processes, population dynamics, species interactions, and species diversity. Ecological disturbances are created by mechanical forces, extreme physico-chemical conditions, and biological consumers and pathogens. Each of these three classes of mechanisms has somewhat distinctive characteristics in the way it operates, the kinds of organisms affected, and the biological and physical legacies that remain. This article reviews general principles of disturbance ecology and describes and illustrates different disturbance mechanisms and their interactions, concluding with a brief discussion of the role of humans in altering disturbance regimes.

1. DISTURBANCE ECOLOGY

Ecological communities are subjected to sharp environmental perturbations, such as the passage of a fire, a storm wave, an avalanche, or a large animal that suddenly reduce standing biomass. Such events also alter ecosystem properties such as energy balance, nutrient fluxes, substrate texture, and chemistry. Generally referred to as "disturbances," these events are distinguished as a special form of environmental variability because they are relatively discrete in both time and space and cause unusual mortality or tissue loss in affected populations. Disturbances promote environmental heterogeneity and free up limiting resources, such as space, light, and nutrients, thereby triggering successional processes of community recovery.

Widespread agents of disturbance in terrestrial ecosystems include fire, wind, extreme temperature, desiccation, gravity (as a force on water, ice, rocks, and soil),
and organisms. With the exception of fire, one can find analogs in aquatic systems, in which the main agents are heat, solutes, currents, desiccation, waves, ice, sediments, and organisms. Disturbance agents operate by exerting mechanical force, altering physico-chemical conditions, or through biological consumption and disease. For example, practically any physical mass can mechanically disturb ecosystems at some scale provided it has sufficient velocity to dislodge organisms or kill tissues. In contrast, disturbance by fire involves chemical combustion of biomass. Disturbance of terrestrial communities by prolonged inundation is largely a chemical disturbance resulting from oxygen depletion. Freshets in estuarine and marine environments are also chemical disturbances. These mechanical and physicochemical disturbances contrast with biological disturbances such as herbivory or predation, which involve tissue removal and digestion by individual mobile consumers.

During the first half of the twentieth century, ecologists paid more attention to post-disturbance processes of species recovery and community succession than to specific disturbance mechanisms. Disturbance was generally treated as a temporary setback to communities that otherwise would tend to develop toward a relatively steady state or “climax community” whose structure and composition were determined by climate and other physical factors and were regulated by endogenous biological interactions. Recently, disturbance has been recognized as intrinsic to and ongoing in virtually all ecological systems. Closer attention has been paid to disturbance processes and their role in promoting characteristic scales of spatial and temporal environmental heterogeneity and in regulating ecosystem processes, population dynamics, species interactions, and species diversity (Paine and Levin, 1981; Sousa, 1984; Pickett and White, 1985). Disturbance ecology has become a familiar term in the lexicon of ecology and a recognized area of emphasis in ecological research.

In practice, it may not be easy to distinguish disturbances from other environmental variation. Most agents of disturbance operate over a continuum, and environmental perturbation is sudden and severe only relative to some set of reference conditions and from the perspective of the affected organisms. The mound of soil produced by a burrowing gopher is a significant disturbance to the underlying herbaceous plants and soil animals but likely to be of little consequence to a large tree a few meters away. A windstorm that topples trees in an open savanna may have no immediate effect on small herbaceous plants located a short distance outside of the canopy. Thus, it is important to bear in mind that disturbance is a relativistic concept and that disturbances can span a very broad range of spatial and temporal scales. Not surprisingly, the term disturbance has been applied somewhat indiscriminately in ecology. The most general definition of a disturbance—any process that causes a sudden decrease in standing live biomass and frees up ecological resources (Sousa, 1984)—is perhaps the most unambiguous.

A disturbance differs from a stress in that the latter is a more chronic condition inhibiting the growth or normal functioning of an organism (e.g., a lack of key nutrients or physical abrasion). A disturbance is termed a catastrophe if it causes extraordinary ecological impact.

Variables commonly used to describe a single disturbance event include timing, extent, and magnitude, where magnitude encompasses both intensity (e.g., energy per area per time) and severity (biological impact). These and other stochastic variables, such as event frequency or recurrence interval between events, have statistical properties that serve to define a disturbance regime. More broadly, a disturbance regime is the collective spatial, temporal, physical, and ecological characteristics of a disturbance process operating in an area. Predictability, which can be defined as the inverse of the variance in disturbance frequency, size, and magnitude (Christensen, 1988), is also an important consideration. In general, predictability increases as the spatio-temporal scale of analysis is expanded from local (the typical size of a disturbance event) to regional (the entire area over which the disturbance regime is manifested).

Because the magnitude of disturbance is defined relative to its ecological impact, it is practically tautological that disturbance regimes are dominated by events of relatively low magnitude and high frequency, whereas higher magnitude events are increasingly rare. However, the impact of a disturbance may not increase linearly with size, frequency, or duration. Romme et al. (1998) distinguish three classes of disturbance response: (i) threshold response, (ii) scale-independent response, and (iii) continuous response. Individuals and communities manifest threshold responses when there are discrete limits in their ability to resist a large disturbance (e.g., the wind speed at which a tree is uprooted). Disturbances can have their greatest ecological impact when one or more events follow close on the heels of another, preventing or disrupting normal community recovery (Paine et al., 1998).

Some mechanisms of disturbance, such as earthquakes or storm waves, are exogenous to the biological communities being impacted, whereas others such as treefall or fire could be considered endogenous. In the former, there is little or no feedback between the state of the ecosystem and the likelihood of a disturbance.
event so that the disturbance regime depends mainly on location and environmental context. In the latter, the likelihood of a disturbance depends on the state of the ecosystem as well as location. Although the categories of endogenous and exogenous disturbances are somewhat artificial, it is useful to examine the relative strength of coupling between disturbance processes and the biota. In many cases, disturbance processes and their effects are tightly coupled to the biological properties of individual organisms and communities. This coupling may promote the formation of specific scales of ecological pattern and reinforce certain ecological and evolutionary processes (Levin, 1992).

In thinking about patterns and processes, it is also useful to distinguish spatially propagating from non-propagating disturbances. Disturbances such as fire and flood spread from neighboring areas, and the spatial pattern of "susceptible" areas or organisms may have a constraining effect on disturbance dynamics, thereby linking spread to previous disturbance events.

II. DISTURBANCE AND BIODIVERSITY

The "intermediate disturbance hypothesis" (IDH; Connell, 1978) predicts that maximum levels of biodiversity should be observed under some intermediate disturbance frequency because few species are able to tolerate very intense disturbance regimes, and few are able to compete successfully in habitats that experience little or no disturbance. The IDH also implies that maximum diversity should be found at some intermediate span of time since the last disturbance. The IDH has been expanded to incorporate intermediate levels of disturbance intensity and extent, and it has been tested and supported in a wide variety of ecosystems.

The best experimental examples of the IDH come from the study of sessile species competing for space or some space-associated resource. Counter-examples come largely from the study of mobile consumers (e.g., freshwater invertebrates), for which rapid immigration may override local disturbance effects. Defining "intermediate" in the context of specific organisms and choosing the scale at which to measure diversity are also important issues in assessing the validity of the IDH, about which there is ongoing debate. In general, the IDH applies to small-scale disturbances and to plants and sessile filter feeders. The relationship between disturbance and diversity is more complicated at larger scales and may not apply when interactions among multiple trophic levels are considered.

At spatial scales much larger than the characteristic size of single disturbance events, disturbance regimes generate disturbance mosaics that maintain beta diversity in landscapes or regions by promoting co-existence of dispersal-limited competing species or prey species and by maintaining environmental heterogeneity and multiple seral stages (Pickett and White, 1985).

If they recur with sufficiently high frequency (e.g., on average at least once per generation), ecological disturbances can be a strong selective force operating on species' morphology, physiology, and/or behavior. Not surprisingly, many organisms are adapted to and/or depend on specific kinds of disturbances and disturbance regimes. Grime (1979) proposed that herbaceous plant life histories can be ordered along three fundamental axes: stress tolerance, competition, and disturbance. Somewhat analogously, animal species are often called r-strategists or k-strategists depending on whether they have high intrinsic rates of reproduction and tend to be favored by disturbances or whether they have lower reproductive rates but exert competitive dominance in the absence of disturbance. Because local abundance of many species is increased or maintained by disturbance, the long absence of specific kinds of disturbances (e.g., fire or floods) may have large negative impacts on biodiversity.

The life history strategies of some organisms may promote specific disturbance regimes (e.g., some fire-adapted shrub species possess canopy structure and foliar chemistry that promote fire spread). Organisms that recover quickly after a disturbance are said to be resilient to that disturbance, as opposed to those that show little response to disturbance and are considered resistant. These concepts are also applied to ecological communities, and the relationship between community diversity and community stability and resilience has long preoccupied ecologists (Holling, 1973). The subject has received renewed attention due to increasing concern over human-caused species extinctions and community impoverishment associated with habitat fragmentation. To date, the hypothesis that community stability to disturbance increases with community richness has met with mixed results in modeling and empirical studies, in part due to differences in spatial and temporal scale and in how stability is measured.

III. MECHANICAL DISTURBANCES

Most aquatic and terrestrial communities are constantly subjected to mechanical forces that reduce standing biomass at some scale. Avalanches, landslides, and debris flows remove soil, shear off and uproot plants, and bury plants and animals. High winds, snow, and ice uproot trees and break off branches and leaves that
then bury other plants and animals. Ice in the nearshore zone plucks and scour littoral communities. Floodwaters topple riparian vegetation and scour streambeds. Waves dislodge encrusting intertidal organisms and break down branching corals. Large animals break twigs and compact the soil as they move. Humans clear or clip vegetation and plow soils. Burrowing animals excavate plants and animals as they dig. In all of these cases, biomass is dislocated but not immediately chemically transformed or consumed. Most important, space occupied by living tissues is evacuated as those tissues are moved somewhere else.

Thus, mechanical disturbances leave conspicuous openings and persistent biological legacies, such as large woody debris embedded in landslides, downed trees next to treefall gaps, floating wracks of detached marine plants and animals, or rubble piles of coral. Materials are often transported to a location where they act to create another disturbance. Thus, the legacy of biological debris from mechanical disturbances not only affects the rate and pattern of community recovery but also may influence the timing and location of subsequent disturbance events.

The spatial distribution of mechanical forces in the environment is highly nonrandom. Features such as substrate, topography, and standing biomass create persistent biophysical pathways in which specific mechanical disturbances are concentrated, thereby compounding existing physical environmental heterogeneity. For example, the location of avalanche paths is strongly associated with topography, geology, and lithologic structure. The impacts of severe wind storms on forest communities varies systematically with topography and soil characteristics as well as vegetation composition and structure (which depend, in part, on disturbance history). Wave energy is concentrated on jutting shorelines. Animal burrowing is concentrated in specific soils and sediments. In other words, many mechanical disturbance processes can have somewhat predictable spatial distributions.

Most mechanical disturbances operate indiscriminately (i.e., they do not target specific organisms, as would selective herbivory, for example) and their impact depends on the ability of organisms to withstand the mechanical force. Thus, mechanical disturbances often winnow biological communities. For example, snow avalanches apply a bending stress on trees in their path that is a function of the snow’s density and the avalanche’s mass and drag. Susceptibility of woody species to breakage (vs bending) increases with tree size, leading to complex interactions between avalanche recurrence frequency and forest structure and composition in avalanche paths. Similarly, masses of attached barnacles and mussels in rocky intertidal environments become increasingly susceptible to detachment by waves as organism size and density increase. The same mechanical agent may create different kinds of disturbances depending on the size of the organism. In forests, for instance, woody stems may be most susceptible to breakage by wind when they are very small or very large, whereas intermediate-sized stems may be the most susceptible to uprooting by wind (Everham and Brokaw, 1996). When impacts of mechanical disturbances depend on the size or density of organisms, this produces feedbacks that influence population dynamics and create pronounced spatial patchiness at scales well below that of the disturbing process (Levin, 1992).

IV. PHYSICO-CHEMICAL DISTURBANCES

Physico-chemical disturbances, such as extreme cold, prolonged inundation of terrestrial organisms, hypoxic episodes in aquatic systems, freshets in coastal waters, and releases of toxic compounds (e.g., biological exotoxins such as those produced by the dinoflagellate *Pfiesteria piscicida*), reduce biomass by imposing lethal physiological stress. Physico-chemical disturbances do not involve mechanical force (except perhaps in the case of extreme cold when tissue damage is due to intracellular ice formation), and most physical and chemical disturbances leave dead tissues and organisms in situ. This contrasts with mechanical disturbances, in which space and resources are freed up through displacement of biological material. There are few studies comparing recovery processes under these two situations, but one might expect differences simply due to the presence or absence of residual detritus that could affect light, nutrient levels, and substrate quality (Reiners, 1983).

Like mechanical disturbances, physico-chemical disturbances have organism-specific impacts because of species-, age-, and size-specific physiological tolerances to environmental extremes. Thus, like mechanical disturbances, physico-chemical disturbances operate as an environmental filter of biological assemblages.

As mentioned previously, much of the emphasis in disturbance studies has been on sessile organisms such as plants and benthic invertebrates. Some large-scale physico-chemical disturbances may be somewhat distinctive in that they can kill large numbers of highly mobile organisms. For example, episodes of anoxia can
cause massive fish kills in estuaries and lakes. Severe cold weather can selectively kill large numbers of birds. In these instances, the suddenness and magnitude of the events prevent animals from escaping, a phenomenon which can also be observed during extreme wildfires.

Fire is obviously a special case of a physico-chemical disturbance because biomass is both transformed and removed in the combustion process. Fire is also unquestionably one of the most pervasive and obvious disturbances in terrestrial ecosystems, and an enormous scientific literature has been devoted to the subject of fire events and fire regimes. It is discussed here to illustrate in more detail the concepts of disturbance mechanisms and disturbance regimes.

The occurrence of fire requires an ignition, fuel combustion, and spread, each of which in turn depends on many physical and biological factors, notably climate, weather, fuel structure and chemistry, and topography. The most important effects of fire are partial or complete combustion of aboveground vegetation, mineralization and deposition of plant tissues as ash and charcoal, extreme heating of the local atmosphere and topsoil, and killing of selected plants, seeds, spores, and animals both above- and belowground.

Figure 1 presents a simple conceptualization of factors that influence fire dynamics at three distinctive spatial and temporal scales in Mediterranean-climate shrublands. The fire fundamentals triangle (Fig. 1a) captures essential elements of the actual combustion process, which occurs at minute scales and depends on oxygen, heat, and fuel. The timing, size, and severity of a wildfire (or controlled fire) event, which may last hours to weeks and extend over hectares to thousands of square kilometers, depends on local weather, topography, and fuel bed. Finally, the fire regime of a landscape depends on complex interactions among climate, vegetation, and ignition factors operating over decades to centuries.

Fire creates strong biological patterns over a broad range of spatial scales. At the scale of individual plants, biomass combustion, soil heating, and nutrient removal can vary widely, imposing very fine-grained variation in mortality and postfire conditions that is manifested in equally fine-grained patterns of biological recovery. A coarser pattern can be created at the scale of entire populations (vegetation stands). Fire severity and vegetation removal can vary with topography, soil, vegetation, and wind conditions, and distance from source populations for seeds and animal colonists can influence community recovery. At still broader scales, landscapes are composed of fire mosaics, a patchwork of stands with differing fire histories and biotic communities. This landscape mosaic is one of the factors aiding in the long-term persistence of species with different fire tolerances. The spatial patterns from individual plant distributions to landscape mosaics all affect

![Figure 1](image_url)

**FIGURE 1** Hierarchical model of fire controls at successively coarser scales of space and time. Connections between fire and individual controls indicate potential feedback loops (after Moritz, 1999, with permission).
the location and severity of future fires and their ecological impacts.

Fires are obvious, discrete events, and research on wildfire history has provided many useful models and methods for characterizing temporal and spatial properties of disturbance regimes (Johnson and Gutsell, 1994). Temporal properties such as time-since-fire (survivorship) and fire interval (mortality) are described using probability distribution functions, notably the Weibull and negative exponential distributions. Fire mosaics and their dynamics have been described using spatial Markov chain models, percolation theory, and spatial autocorrelation statistics. These quantitative characterizations are useful for describing changes in fire regime over time, for comparing different areas, and for estimating wildfire risk.

V. DISTURBANCE BY HERBIVORES AND PREDATORS

In many ways, consumers (and pathogens) have the same effects as other agents of disturbance in that biomass is removed and new opportunities for recruitment are created. There are some distinctive features, however, of disturbances by herbivores and carnivores. Most obviously, they operate at the relatively fine scale of individual consumers, and they leave a very different biological legacy than mechanical and physico-chemical disturbances because biomass is concentrated and converted into heat, new tissues, and discrete waste products. The waste products are frequently deposited away from the disturbed area. For example, frass from defoliating caterpillars in tree canopies falls to the soil surface. Shade-seeking ungulates in African savannas deposit a disproportionate amount of urine and feces under trees. Elk migrations lead to a net transport of nitrogen from summer ranges to winter ranges.

Foraging is also a far more selective form of disturbance than other disturbance mechanisms. Organisms (including humans) preferentially occupy specific sites and microsites and focus on specific plant resources, creating patchiness at many scales. Most plants are unpalatable to most herbivores, and selective herbivory by dominant grazers can exert strong, directional effects on community composition, structure, and ecosystem processes. These effects may be amplified by strong feedbacks between consumption, food quantity, and food quality. For example, moose in boreal ecosystems selectively browse nitrogen-rich deciduous species and avoid coniferous species. Over time this promotes nitrogen sequestration in conifer litter and accelerates succession of deciduous to coniferous plant communities (Hobbs, 1996). Disease spread from infected to susceptible hosts and the development of resistance in infected individuals is somewhat analogous in terms of the strong spatio-temporal feedbacks between the disturbance process and the state of the system.

Many consumers act as mechanical, chemical, and biological agents of disturbance. Perhaps the most sensational example is that of beavers, which selectively fell trees, dam streams, and inundate floodplains. Similarly, burrowing animals such as pocket gophers can operate as "physical ecosystem engineers" creating pervasive changes in the abiotic and biotic environment by redistributing soil and nutrients, altering soil hydrology and canopy light regimes, and altering vegetation composition and dynamics.

VI. INTERACTIONS AMONG DISTURBANCE MECHANISMS

Disturbances can interact in complex, nonlinear ways across space and time. One disturbance may promote another, such as occurs when dry-season fire in Mediterranean-climate shrublands promotes flooding and debris flows during the ensuing wet season. Conversely, one disturbance mechanism may reduce the likelihood of another. For example, grazing reduces the likelihood of fires in grasslands and the likelihood of crown fires in some forest ecosystems. Nearshore communities that are heavily impacted by wave disturbance may also experience lowered rates of predation. It is not difficult to imagine many such interactions operating across a wide range of spatial and temporal scales.

Disturbances often play an important role in invasion and spread of exotic species. In Hawaii, for instance, invasion of the nitrogen-fixing shrub, Myrica faya, was facilitated when native vegetation was thinned by heavy ash deposition from a volcanic eruption. There is a growing list of examples in which disturbance promotes the spread of an invasive species and then that species initiates ecosystem changes that introduce new disturbance processes. For example, in the western United States cattle grazing in shrub steppe promoted the spread of the exotic grass, Bromus tectorum, and the presence of a dry grass layer then increased wildfire size and frequency in that system, in which fires were historically uncommon.

As mentioned in Section 1, compounded disturbances, or disturbances of communities already stressed
by abiotic or biotic forces, can have large and persistent effects on ecological communities if they occur in such rapid succession that they disrupt normal recovery processes. Repeated burning over a series of 2 or more years is a common means of converting shrublands to grasslands in semi-arid ecosystems. In a particularly dramatic example, Hughes (1994) describes the decline of corals in Jamaica due to the combined effects of overfishing, two major hurricanes, and disease that caused mass mortality in sea urchins. These impacts appear to have collectively precipitated a massive phase shift from coral-dominated to algal-dominated reefs.

VII. HUMANS AS AGENTS OF DISTURBANCE

For many millennia humans have been manipulating ecosystem processes with varying impacts on ecosystem function and species distribution. Human activities that create disturbances at spatiotemporal scales to which organisms are adapted generally do not affect long-term changes in community composition or structure. For example, shifting agriculture in the humid tropical forests, like hurricanes, creates small, temporary forest clearings and results in landscape mosaics within which most species can persist. In contrast, commercial forest cutting and cropland and pasture development are large and/or persistent disturbances that can produce regional extinction of species and require much longer periods for forest recovery (Attiwill, 1994).

Another widespread human impact has been to regulate or arrest disturbances such as fires, floods, or animal activities that create habitats on which some species depend. Even when human ecosystem manipulations are, by design, modeled after a particular disturbance, it is difficult to incorporate all disturbance-related changes in the environment. For example, forest harvesting can occur at the spatiotemporal scale of a natural fire regime, but the post-fire nutrient pulse and legacy of dead snags are no longer factors in the recovering ecosystem. Streams may be regulated to mimic periodic flooding, but dams often inhibit fish migrations that would happen at that time (and agencies will understandably stop short of re-initiating very high-magnitude flood events).

As humans fragment terrestrial habitats through land-use conversion and aquatic systems through impoundments, diversions, dredging, trawling, and construction, permanent fragmentation of habitats is superposed on dynamic disturbance mosaics, causing changes in disturbance regimes. Such fragmentation can retard disturbance spread (e.g., the spread of fire or disease), alter disturbance regimes in remnant patches, and create edge environments with novel disturbance regimes. Local human activities can have large cumulative effects. For example, the building of roads alters the natural cycles of erosion and landslide initiation, whereas agricultural land use often leads to the concentrated nutrient inputs and changes storm flows in receiving water bodies. In fact, most human activities have the capacity for altering historical disturbance regimes in some way.

At the broadest scale, humans are changing global climate patterns, and nearly all mechanisms of disturbance and disturbance effects are ultimately tied to climate processes. Changes in disturbance regimes during global climate change are inextricably tied to other ecological responses to changing climate, affecting whether species can persist in their current ranges or can invade new areas of more suitable environmental conditions. The tight coupling between biotic processes and structures and most disturbance patterns and processes—a recurrent theme in this article—creates complex feedbacks that could amplify or moderate the effects of global climate change on disturbance regimes, species, and ecosystems.

See Also the Following Articles

DISEASES, CONSERVATION AND ECO SYSTEM FUNCTION, PRINCIPLES OF FIRES, ECOLOGICAL EFFECTS OF GRAZING, EFFECTS OF HERBACEOUS VEGETATION, SPECIES RICHNESS IN HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW

Bibliography


DIVERSITY, COMMUNITY/REGIONAL LEVEL

Howard V. Cornell
University of Delaware

GLOSSARY

between-habitat diversity Degree of dissimilarity in species composition between two habitats. The greater the number of species shared by two habitats, the lower the between-habitat diversity.

metacommunity An assemblage of species in a patchy landscape, each species comprising a metapopulation in that its distribution shifts among habitat patches because the distribution is determined by colonization and extinction dynamics.

metapopulation Group of subpopulations distributed among habitat patches in a landscape, each subpopulation being subject to periodic extinction and subsequent recolonization from other occupied patches.

richness Simplest measure of diversity, represented by a count of the number of species in some ecologically meaningful unit (community, metacommunity, region) without regard to variation in the number of individuals per species.

saturation Upper limit to species richness within a community, set by species interactions, and independent of the pool of colonists to which that community is accessible.

species pool Set of species that are able to colonize a habitat of interest.

type I community An assemblage of species where the richness is a constant proportion of the number of species occurring in a larger geographic unit in which the community is embedded.

type II community An assemblage of species where the richness is independent of the number of species in a larger geographic unit in which the community is embedded.

AN ECOLOGICAL COMMUNITY IS A COLLECTION OF SPECIES LIVING CONTEMPORANEously IN THE SAME PLACE and comprising populations of individuals that are spatially interspersed and among which direct and indirect interactions can potentially take place. Communities are nested within larger biogeographical regions that comprise many different habitats and that have unique evolutionary histories. Species can disperse among suitable habitats in the region and thus define the species pool for communities occupying those habitats. Species interactions, habitat differences, history, and dispersal have all been hypothesized to influence species diversity in communities, and thus community and regional species diversity are inextricably linked. This article presents two perspectives on
the control of community and regional diversity, the nature of the linkage between community and regional diversity, and methods for examining this linkage. Guidelines for synthesizing community and regional perspectives are also considered.

I. TWO PERSPECTIVES ON THE CONTROL OF COMMUNITY/REGIONAL DIVERSITY

A. Local Control of Species Richness

1. The Volterra–Gause Perspective

One of the most obvious and fundamental facts of ecology is that all species do not occur everywhere. Instead, each species has a unique distribution determined by its ecological context and evolutionary history. Nevertheless, species distributions overlap to varying degrees so that some places support more species than others do. Such variations in species richness have been a central issue in ecological and biogeographical studies at least since the turn of the century. Until recently, explanations focused mainly on processes operating within ecological communities over relatively short periods of time. This perspective arose out of a desire to add scientific rigor to ecological studies by focusing on mechanisms and ignoring history. A mechanistic approach made it possible to model ecological systems with simple equations and to test these models and their assumptions with carefully collected data.

One of the earliest truly influential models in community ecology was developed by Vito Volterra (and independently by Alfred Lotka) in the mid-1920’s. This model is a set of simple differential equations describing competition between two species in terms of reductions in per capita population growth rates by interactions among individuals. When the model is analyzed at equilibrium, the outcome of competition (coexistence or competitive exclusion) is easily predicted from the intensity of interactions within species relative to those among species. If populations limit their own growth more than the growth of other populations, the two species will coexist. If they limit other populations more than themselves, one will exclude the other. The model and its assumptions were later tested by G. F. Gause in laboratory systems comprising several species of protozoa. The tests lent strong support to the model, and as a result the so-called Volterra–Gause perspective has become a metaphor for the way in which ecologists think about control of species richness.

According to this perspective, species richness in a biogeographical region may indeed be determined in part by history. However, history has little relevance to the upper limits on local richness within communities. Instead local richness is limited primarily by species interactions (predation, parasitism, herbivory, competition, mutualism) in conjunction with abiotic conditions (water, nutrient, and energy availability, chemistry, temperature, disturbance). In other words, the outcome of interactions is determined in large part by the abiotic conditions in which they occur. For example, lower concentrations of a particular nutrient or lower water availability might favor one plant species in competition with another, but higher values of these factors might result in a reversal of competitive dominance. The number of species in the biogeographical region provides a species pool from which the local communities are drawn, and all species have the potential for colonizing all habitats in the region. Since habitats vary in their abiotic conditions and resident assemblages, only those species capable of coexisting in a given milieu will establish successfully. Differences in species richness thus result from differences in the abiotic template and species composition among habitats. In a recent series of publications from the early 1990s, James Drake has shown that the colonization order of species into a habitat can also have some effect, and communities can exhibit multiple stable points representing different levels of species richness.

A crucial element of this perspective that derives directly from the Volterra model is that communities exist at a dynamic equilibrium. For example, in the case of competition, population levels of the resident species are limited by resource availability and physical conditions. Thus, if sufficient time elapses and there are no perturbing disturbances to the community, the number of species will saturate at a level determined by the number of limiting resources and physical factors. At saturation, species are sufficiently similar in their resource use that finer subdivision of resources becomes increasingly difficult and new colonizations into the community must be accompanied by extinctions of current residents. If habitats throughout the region are all identical and saturated with species, then the total number of species in the region is limited to the total number coexisting within each identical habitat. That is, community and regional richness are identical and regional richness is limited by the same factors that limit local richness.

Since all habitats are not identical, species within the region have the option of partitioning habitats as well as resources and thus the same set of species will
not occur everywhere in the region. Nevertheless, local processes are still limiting both local and regional richness, but limits on regional richness now result from interplay between species interactions and spatial habitat variation. The greater the number of habitats, the greater the between-habitat diversity and the greater the number of species in the region relative to those within a single community. Historical processes such as massive extinctions might reduce the number of species in the region for a time, and subsequent evolution, speciation, and dispersal might generate a new biota that differs from the previous one. However, the essential feature of this viewpoint is that given sufficient time, the imprint of history on species richness will fade, communities will fill up and become saturated, and processes operating locally will also limit richness regionally.

Even strict adherence to the Volterra–Gause viewpoint still allows for other factors to affect local and regional richness. All else being equal, larger localities and regions will support more species than smaller ones as predicted by the species–area relationship. In addition, some species within a region might be ecologically equivalent and thus might be able to substitute for one another within a given community type. Chance might determine what species occupies what particular community, in which case ecologically equivalent species could increase regional richness without affecting local richness. Finally, spatial interspersion of habitat types might allow species adapted to one habitat type to persist in an adjacent, less favorable habitat type by means of propagule flow. This so-called source–sink effect proposed by Ron Pulliam in 1988 would increase local richness in the less favorable habitat above that which would be self-maintaining in the absence of immigration. Although these factors may have important impacts on species richness, they do not alter the premise that the primary direction of control on species richness is from the local to the regional level.

An ecological community at equilibrium and not subject to vagaries of history evokes a powerful image that continues to influence modern ideas about species diversity. Robert MacArthur’s (1972) book represents the pinnacle of this viewpoint, although he recognized that the equilibrium assumption and local control of richness did not apply to all communities. There are clear recent demonstrations that species interactions have strong influences on community structure and might exclude species from the community (see Lawton, 1999, for a brief review). In short, the Volterra–Gause perspective has been a useful guide to a limited understanding of the mechanisms by which particular communities are structured. However, this understanding has led to few if any general patterns or rules that apply to most systems (Lawton, 1999).

2. Species–Energy Theory

Another view of local control of species richness comes from the ecosystem perspective, and has been termed “species–energy theory.” The theory was developed mainly by David Wright, David Currie, and Brian Maurer (1993). Species–energy theory posits that local and, ultimately, regional richness are limited by available energy within the local environment. Available energy limits local richness by placing ceilings on the population densities of individual species. As energy decreases, densities decrease, random extinctions increase, and local richness is reduced. Available energy can be scaled up to whole regions by multiplying the energy per unit area by the total area of the region. Thus, population densities and species richness over the whole region are also controlled by energy availability.

Although the theory in this simplified version posits local control of local and regional richness, it differs from the Volterra–Gause viewpoint in that species interactions are of secondary importance in limiting richness. Energy can have direct and independent effects on population levels to increase richness rather than mediating interactions among species. Interactions can be included in the model, but they are not necessary for limits on richness to occur. Species–energy theory is very similar to island biogeography theory, which proposes that species richness is a balance between colonization and extinction rates in a local habitat. Smaller habitats have higher per species extinction rates and thus lower species richness as predicted by the species–area relationship. The main difference between the two theories is that species–energy theory substitutes energy availability for area as the predictor of species richness.

A recent literature review by David Wright and others demonstrated that several proxy variables for energy availability (precipitation, potential or actual evapotranspiration, productivity) are indeed strong predictors of species richness for various taxonomic groups (Table 1). However, it is not clear whether energy is the direct cause of richness variation, or whether it simply covaries with other possible causes such as area effects, habitat heterogeneity, species interactions, and historical effects. It is also not clear whether the assumptions of the theory are always met in particular studies. For example, it is assumed that energy limitation should operate at the local scale where individuals of each species seek sufficient energy to persist in the commu-
TABLE I

Energy-Related and Other Factors Which Correlate with Species Richness

<table>
<thead>
<tr>
<th>Factor cited</th>
<th>Significant</th>
<th>Not significant*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy-related factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean annual heat and/or humidity variables</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>Productivity (sometimes est. as standing crop)</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Nutrient, food availability</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Seasonality</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Other factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Habitat complexity, subdivision, microrelief</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Disturbance</td>
<td>5</td>
<td>5(2)</td>
</tr>
<tr>
<td>Environmental chemistry (pH, cations)</td>
<td>3</td>
<td>1(1)</td>
</tr>
<tr>
<td>Isolation, peninsula effects</td>
<td>2</td>
<td>1(1)</td>
</tr>
<tr>
<td>Diversity of food, prey</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Regional richness</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Time, historical factors</td>
<td>1</td>
<td>2(2)</td>
</tr>
<tr>
<td>Competition, predation</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Factors investigated as the principle determinant of species richness in 82 correlations from 53 studies.

*Numbers in parentheses indicate the number of correlations that were omitted from further analysis because no correlation coefficient was reported.


However, it is at the scale of the local community that the correlation between energy and richness is weakest and most variable; the correlation is strongest at larger spatial scales (104–108 km²). It is also assumed that average population size and species richness should increase together with available energy, but this is not always the case. In particular, David Tilman and Pacala have pointed out that plant systems often show an inverse relationship between per species population sizes and energy as predicted by the −3/2 power law, yet species richness commonly increases with energy in the expected way. Despite these limitations, energy can be a useful variable for estimating richness at larger scales because covariation is high regardless of the underlying causes, and because energy data (or proxies) are available for most geographic areas.

At more local scales, species interactions among resident populations may come into play, altering any potential increases in richness with energy availability. Such interactions might partially explain the more variable relationship between energy and richness within communities. Increased energy can lead to increased biomass of the total community, which in turn can lead to increases or decreases in average population density depending on the intensity of interindividual interactions and the size plasticity of the taxa involved. But it is the changes in the relative allocation of energy among species that are critical for predicting changes in richness. If allocation changes as energy availability increases, for example, caused by shifts in competitive superiority or ability to avoid predation, then the population densities of some species may increase while other species may be driven extinct. Richness can thus increase or decrease with increased energy depending on the circumstances. It is at this point that the species-energy theory converges with the Volterra-Gause viewpoint since the environment is now mediating interactions among species.

3. The Diversity–Productivity Relationship

The idea that richness can change with energy availability in unpredictable ways has had a long history under the guise of the diversity–productivity relationship. Productivity is a measure of energy availability to the
community and has been shown to affect local richness in contradictory ways; richness either increases, decreases, or shows no relationship with productivity depending on the assemblage. Recently it has been suggested that these contradictory relationships may just be incomplete segments of an overall hump-shaped, unimodal relationship over a broader range of productivity. In other words, richness peaks at intermediate productivity levels and declines toward the extremes. Different taxa can show richness peaks at different levels of productivity, resulting in unimodal curves that are displaced along the productivity spectrum. No ecological pattern is universal and, accordingly, contradictory relationships between richness and productivity need not always imply a truncated hump. Indeed, evidence for this possibility is currently limited at best (Watson et al., 1999). Monoecious relationships may be just as frequent as unimodal ones. Nevertheless, the unimodal curve is a useful synthetic construct and it is sufficiently frequent to invite explanation. Moreover, the same explanations for each arm of the hump may apply equally well to monotonically increasing or decreasing relationships over a particular range of productivity.

It is not clear why richness shows this strange relationship with productivity, and the pattern remains one of the most important unresolved mysteries in community ecology. The hypothesis that productivity differences are directly responsible for the unimodal relationship assumes that productivity is uniformly distributed among sites. In other words, low-, medium-, and high-productivity environments occur with equal frequency. If medium-productivity sites are more frequent than low- or high-productivity sites, then the unimodal curve might just be a reflection of simple species-area effects. There is some evidence to suggest that this is not the case (Rosenzweig, 1995), but the possibility has not been extensively tested. At the very least, diversity-productivity correlations should be corrected for unequal area effects before a causative link between these variables is seriously considered.

Unimodal curves (as well as monotonically increasing or decreasing ones) can occur at different spatial scales (local, landscape, regional, continental), but their frequency changes as spatial scale increases. Monotonically increasing patterns are more common at larger scales and unimodal patterns are more common at local scales (Fig. 1). Moreover, the mechanisms responsible for patterns at each scale probably differ (Watson et al., 1999). Since the within-community, among-community, and local scales are where ecological processes become important, theories that consider changes in species interactions and local environmental conditions along productivity gradients are the most relevant.

The decline to the left of the hump may derive from increased environmental stress or the direct effects of energy reduction on population levels as proposed by species-energy theory. Other mechanisms, such as increases in disturbance rate, poor recovery from disturbance due to slower population growth rates, declines in territoriality, and environmental heterogeneity have also been suggested to play a role. The decline to the right of the hump may involve some form of interaction mediation among species, such as increased intensity of competition and predation or increased population dynamic instability within food webs at high productivities. Such declines would then provide strong support for the idea that species interactions limit species richness in accordance with the Volterra-Gause viewpoint. However, it is important to point out that richness declines in highly productive areas may also result from regional effects. For example, if productive habitats are relatively rare, they are likely to be surrounded by larger, less productive areas and thus isolated from a large source pool of species adapted to productive conditions. As previously mentioned, area effects may also contribute to low richness. Regional effects must therefore always be taken into consideration when analyzing diversity-productivity relationships.

Tests that distinguish among these various theories are difficult to perform and have thus far been limited. Comparative approaches, although they require less disruption of the community, might not be able to distinguish among alternative theories unless natural "experiments" in which selected mechanisms are not operating can be found. Experimentation has better control of confounding variables, but care must be taken to provide enough time for all processes that contribute to changes in richness to ensure that they are allowed to respond to the manipulation. For example, in a 1999 review, Laura Gough and co-workers found that fertilization experiments consistently resulted in increases in inorganic nutrient concentrations, which in turn intensified species interactions in herbaceous plant communities. Fertilization increased productivity and reduced species richness across a broad range of community type, due primarily to increased dominance by one or a few species. These experiments certainly demonstrate the potential of interaction mediation to limit community richness. However, the relevance of such studies to richness declines in natural communities at highly productive sites is questionable. Fertilization reduces diversity even at poorly productive sites, where theory predicts that diversity should increase.
FIGURE 1 Breakdown of published plant and animal studies showing a unimodal, positive, negative, or no relationship between species diversity and productivity at four spatial scales. The number of studies in each histogram (n) is shown. (From R. B. Walde, M. R. Willig, C. F. Steiner, G. G. Mittelbach, L. Gough, S. I. Dodson, G. P. Juday, and R. Parmeter (1999). The relationship between productivity and species richness. Annu. Rev. Ecol. Systematics, 30, 257-300, with permission.)
This anomalous result may be a consequence of the short-term nature and small spatial scale of such experiments. Species adapted to the higher productivities of the fertilized sites might not be able to reach them if they are isolated in space or time. More appropriate experiments should therefore manipulate the species pool to simultaneously test for the effects of productivity and the dispersal effects on richness. If pool limitation is the cause of the anomalous results, then experimentally boosted recruitment to low-productivity sites by an appropriate species pool should generate higher species richness with higher fertilization rates. Naturally, the species selected for the boosted recruitment should be well-adapted to the increased nutrient levels at the fertilized sites. High-productivity sites, on the other hand, should remain depressed in diversity owing to interspecific interactions. Some combination of comparative and experimental approaches will probably be required to solve the mystery of the diversity-productivity relationship. Regardless of the methods used, it is important that it be solved, since it represents one of the clearest possibilities that local richness is locally controlled. It also provides a single framework that subsumes both species-energy theory and Volterra-Gause processes.

4. Theory Doesn't Always Predict

Local Control

Without question, the Volterra-Gause viewpoint has stimulated much research on the factors that control species richness. Nevertheless, the fundamental premise that local processes control local and ultimately regional richness may be incorrect for many systems. Even the restrictive Volterra-Gause models do not always predict limits on species richness due to local mechanisms. After appropriate modification to incorporate some realistic features of environments and species, such as spatial heterogeneity, environmental fluctuations, nonlinear population growth responses to environmental change, and trade-offs among species in their ability to compete in different environments, the models often predict that there are no upper limits on local richness (Tilman and Pacala, in Ricklefs and Schluter, 1993). Moreover, the Volterra-Gause assumption that population levels are at equilibrium, may not commonly be met in natural communities. In a classic example, periodic disturbance can drive populations below equilibrium, where the outcomes of interspecific interactions are never fully realized. More species can thus invade than would be allowed by equilibrium population levels.

If real communities have any of these features, it is not likely that local richness will reach saturation as predicted by the simplest models. Evidence from real communities suggests that lack of saturation is indeed common. Lack of saturation raises the possibility that factors external to the community may ultimately determine local richness. As a consequence, the factors that determine regional richness must also be sought elsewhere. In the Volterra-Gause view, regional richness is limited by a combination of the number of species within saturated habitats and the number of habitats within the region. If communities within habitats are not saturated, control on regional richness cannot be exerted at the within-habitat scale. Moreover, the number of habitats within a region is not fixed and may be more an effect of the number and kinds of species in the region than a cause of richness limitation at the between-habitat scale (Rosenzweig, 1995). In other words, the more species there are in the region, the more finely habitats can be partitioned and the more distinct habitat types that can be identified. How, then, do differences in richness among regions arise? And what relevance do these differences have for local richness? An alternative to the Volterra-Gause viewpoint that addresses these questions is thus needed.

B. Regional Control of Richness

1. Regional Processes

The alternative view rests on an idea that actually predates the Volterra-Gause view, namely, that regional richness is set to a large degree by historical biogeographical factors. The importance of history is implicit in the traditional definition of a biogeographical region that relies in part on the presence of a distinctive native biota. The presence of large groups of indigenous species implies that the region has undergone a period of isolation and independent evolution. Thus, the numbers and kinds of species that occur across the region are determined to some degree by its unique historical development. These different histories might result in widely divergent species richnesses among some components of the biota, which in turn may be reflected in local communities. Under this scenario, local richness is not limited by species interactions as predicted by simple Volterra-Gause models, but becomes some function of the number of species in the region that are able to colonize and persist in a given habitat. The richness of the community embedded within the region thus retains some of the region's historical signature.

The key prediction arising out of this scenario that distinguishes it from the Volterra-Gause model is that localities in richer regions can support more species
FIGURE 2  Phylogenetic and biogeographical patterns that suggest association by descent. Species sets (2, 7, 12) and (3, 6, 13) comprise communities that were found on opposite sides of a geographic barrier. The phylogeny (A) indicates that species sets (2, 3, 7, 8), and (12, 13) each had a common ancestor. These two patterns suggest that both communities descended from a common ancestral community that existed before the barrier was formed. (B) The species in each extant community are thus associated because their ancestors were associated. (From Brooks and McLennan, in Ricklefs and Schluter. (1993). Copyright © 1993 by University of Chicago Press.)

than those in poorer ones even though the localities are environmentally similar. This is simply because species co-occurrences within localities are determined to some extent by the same factors that determine regional co-occurrence. In their 1991 book *Phylogeny, Ecology, and Behavior*, Brooks and McLennan pointed out that regional co-occurrence can come about via two processes: association by descent and association by colonization. In association by descent, two species co-occur because their ancestors co-occurred. In other words, the community was assembled from ancestors in the distant past, and the species present today evolved *in situ* from members of the ancestral community, most likely by geographic isolation (Fig. 2). The community is thus rich or poor because the region in which it is embedded was historically rich or poor. In association by colonization, two species can co-occur because one evolved elsewhere and migrated into the region. Communities are open to colonization because (1) interactions are not limiting richness, (2) the colonizing species uses niche space that is not occupied by the *in situ* residents, or (3) after contact with the residents, evolution of niche partitioning occurs. In association by colonization, the community is rich or poor because its region has high or low barriers to colonization from elsewhere. It is likely that real communities are mixtures of colonists and *in situ* descendants, and Brooks and McLennan have pioneered the development of phylogenetic methods to distinguish between these two groups.

The influence of higher colonization rates into a region on regional and thus local richness is obvious, but the role of history is less so. Why are some regions historically richer than others? The answer to this question must involve region-specific differences in historical speciation rates, extinction rates, or both. Some processes might also act to increase the speciation rate relative to the extinction rate. For example, some taxa may have developed key innovations that allowed them to enter a new adaptive zone and undergo an adaptive radiation. The region may have been simultaneously dissected by colonization barriers, disrupting gene flow
among many populations and resulting in a burst of speciation.

Latham and Ricklefs present a possible example in the evolution of frost tolerance and proliferation into the temperate zone by forest trees. Asian temperate forests have three times more tree species than North American forests and six times more than European forests. Because Asia may be particularly diverse topographically and has no geographic barrier between the tropical and temperate zones, such proliferation may have occurred more frequently than in North America and Europe, where such barriers do exist (Ricklefs and Schluter, 1993). Alternatively, some taxa may undergo non-adaptive bursts of speciation resulting from inherent biological attributes, including genetic peculiarities (e.g., chromosomal anomalies or polyploidy), mate recognition and selection, or other factors. A dramatic example is the anomalous diversity of the Hawaiian Drosophila. The proliferation of this genus in Hawaii far exceeds that in any other region of equivalent area and has been attributed to a strong tendency for sexual selection in the Hawaiian taxa (see Cornell and Lawton, 1992). Adaptive and nonadaptive radiations might be distinguishable by differences in the relationship between speciation rate and richness. In adaptive radiation, speciation rate should at first exceed extinction rates and then come into balance with extinction as regional richness increases and the adaptive zone fills up. Nonadaptive radiations do not require ecological diversification, and thus speciation and extinction rates should be unaffected by regional richness. Adaptive and nonadaptive radiations are not mutually exclusive and both could be operating in the same clade.

Other processes might act to increase extinction rates relative to speciation rates in some regions. Increased probability of extinction is believed to be linked to low average population size, which in turn is correlated with geographic distribution. Thus, any event that reduces a species' range, such as habitat destruction or displacement by an invader, is likely to decrease its time to extinction. Returning to temperate forest trees, glaciation in conjunction with mountain barriers in southern Europe might have squeezed the geographic ranges of many temperate tree species below a critical threshold, driving them to extinction. The absence of mountain barriers in Asia and North America may have allowed temperate species to survive in southern refuges, accounting in part for the higher tree species richness in those regions.

The relative importance of colonization, speciation, and extinction rates in setting regional richness will vary depending on the taxon and the history of the region. Since regional and local richness are linked, an understanding of community assembly will require a new focus on these macroevolutionary processes. Fortunately, new phylogenetic methods involving sister group comparisons and other innovations have been developed by Mitter, Farrell, and others, which can distinguish among various macroevolutionary causes for regional and local richness variation. Such methods have given history and biogeography the consideration they deserve in some modern investigations of community structure (e.g., McPeek and Brown, 2000), but more work needs to be done. Community structure is undoubtedly influenced by contemporary processes as well, and distinguishing historical from present-day effects such as energy, local environment, and species interactions will be one of the biggest future challenges in this area.

2. Metacommunity Processes

Once species become part of the biota of a particular region, they can disperse among habitats, and as a result communities can be quite dynamic in space and time. The species within the region can thus be viewed as a species pool for a given local habitat. The species pool will be considerably larger than the number of species in any habitat for at least three reasons: (1) some species in the region will not be able to disperse to all sites because of colonization barriers or isolation-by-distance; (2) all habitats in the region are not favorable for all species; and (3) species can be driven locally extinct by species interactions, demographic and environmental stochasticity, disturbances, or other unfavorable changes in local conditions. It is of some interest to explore situations where species cut off by highly effective colonization barriers are not considered part of the pool for habitats from which they are isolated. The dynamic links between local and regional richness should thus be examined at a spatial scale where all species can reach all habitats over ecological time. Moreover, we should focus on a single habitat type, since looking at more than one habitat will confound regional effects with local environmental variation. This can generate artificial relationships between local and regional richness (Srivastava, 1999). The metacommunity within the landscape, in its simplest form, fulfills both of these criteria and thus represents an appropriate theoretical context in which to examine these links. The term metacommunity was first used by Gilpin and Hanski in 1991. The metacommunity can be thought of as a pool of species in a fragmented landscape of environmentally identical sites, each species behaving as a metapopulation in that its distribution among
sites in the landscape is determined by colonization and extinction. In this scenario, regional processes still determine the potential species pool, but now the effects of the pool on local richness are filtered through the metacommunity.

The simplest metacommunity model derives from Richard Levins' original 1969 equation for a single-species metapopulation:

\[ \frac{dp}{dt} = c_p(1-p) - m_p \]

Tilman and Lehman used this equation to construct a metacommunity model by writing independent equations for as many species as occur in the metacommunity (Tilman and Kareiva, 1997). The value of \( p_i \) is the proportion of local sites in the landscape occupied by the \( i \)th species, and \( c_i \) and \( m_i \) are the site-specific colonization and extinction rates, respectively. Thus, when \( \frac{dp}{dt} = 0 \), each \( p_i \) is set by a balance between colonization and extinction rates per site. The model assumes that all sites are identical. This assumption permits the examination of links between local richness and pool size without the confounding effects of local environmental variation. If \( p \) is defined as the mean of \( p_i \)'s for all species, then \( p \) becomes the average proportion of local sites occupied by a species. Alternatively, it is the average proportion of the metacommunity found at a particular site. Thus, local richness = \( p \) (landscape richness).

What this expression tells us is that if \( p \) is a constant, then local richness is a linear function of landscape richness, a pattern that has been called Type I or proportional sampling (Fig. 3). Proportional sampling predicts that matched habitats embedded in richer landscapes will always have richer communities, but that local richness will be less than landscape richness by a constant proportion. The slope of the relationship, \( p \), will be determined by the size of the pool balanced by the rate of species extinction at individual sites. In this simple model, species-specific colonization and extinction rates are assumed to be independent of pool size and the number of species at a site. Thus the model predicts the relationship between local and landscape richness in the absence of species interactions. Under these conditions, communities will be unsaturated and open to colonization from the species pool.

At least one recent study by Hugueny and Cornell (2000) has shown that the proportional sampling pattern in a real species assemblage can be predicted from colonization and extinction rates. Metacommunity models can be made more realistic by adding rescue effects, by adding provisions for variable population sizes among patches, by making habitats differentially suitable to different species, and by incorporating source–sink effects and habitat specialization. Such modifications may alter the details of the Type I curve, but they will not alter the conclusion that pool effects strongly influence community richness. Thus, for the purposes of conceptual distinction between saturated and unsaturated communities, the Levins model suffices.

The Levins model not only predicts the relationship between landscape and local richness; it also shows that the landscape filter constrains the species pool. Most obviously, landscapes that are smaller will have smaller species pools as predicted by the species–area relationship. More subtly, landscapes of a given area that are more highly fragmented will have smaller species pools because they suffer higher landscape-scale extinction rates. Fragmentation also reduces the average proportion of the pool \( (p) \) found at a local site by lowering per site colonization rates and increasing per site extinction. Thus any differences in landscape characteristics must be considered when estimating the species pool and evaluating the proportional sampling relationship.

If competitive interactions are added to the model, and it is assumed that competition leads to extinction rather than character displacement and speciation, then
increased competition can result in increased extinction rates and/or decreased colonization rates per species. As the species pool increases, competition could become sufficiently intense that colonization by a new species into the community would either fail or result in the rapid extinction of other species. Under this condition, further increases in the pool do not result in increases in local richness and the community becomes saturated with species. At saturation, the relationship between local and landscape richness levels off and becomes independent of the rate of colonization and the size of the species pool. This leveling relationship has been designated Type II (Cornell and Lawton, 1992). Species interactions are thus limiting local richness consistent with the Volterra–Gause perspective. Other dynamic models developed by Caswell and Cohen (in Ricklefs and Schluter, 1993) and Morton and Law (1997), which allow for dispersal, species interactions, and local extinction, also predict leveling off and saturation.

However, the spatial character of many dynamic models allows for realistic complexities that ameliorate the impact of species interactions on local richness. For example, if competition occurs only among close neighbors, and inferior competitors are better dispersers, then a potentially unlimited number of species can coexist locally (Lehman and Tilman, in Tilman and Kareiva, 1997). Other spatial models incorporating disturbance, limited dispersal, spatially local competition, and habitat heterogeneity also predict coexistence at high levels of richness in the face of competitive interactions (Cornell and Karlson, and Lehman and Tilman, in Tilman and Kareiva, 1997). Thus Type I relationships can still occur in the presence of intense competition.

It is important to point out that just because meta-community and other models predict proportional sampling or saturation under certain conditions, it does not mean that the Type I or Type II relationships need be explained exclusively by mechanisms operating at the landscape scale. The models simply represent theoretically plausible ways in which the pattern can be generated. Factors operating at other spatial scales can also produce Type I and Type II relationships. For example, suppose the scale of the analysis is expanded to the entire region. One reason to do this is that regions are easier to define than landscapes and thus more data on regional richness have been collected. Type I and Type II relationships are thus more conventionally represented as plots between local and regional richness (see Fig. 3). Under these circumstances, Type I relationships can result simply from dispersal limitation. That is, when regions of different species richness are compared, a constant proportion of the regional biota is unable to reach the habitat type of interest. Type II relationships, on the other hand, can be generated by historically stable communities in which all niche space is occupied and into which invasion is limited. In such situations, the community need not be particularly dynamic, need undergo little turnover in ecological time, and its species composition may bear the strong stamp of its historical development. Nevertheless, a Type I relationship would still be an indication that hard limits on local richness have not been reached even in the richest regions, whereas a Type II relationship would suggest saturation. The Type I and Type II relationships should thus be viewed as empirical patterns with multiple causes, but also as evidence that local richness is or is not limited by local factors.

II. TESTING FOR SATURATED COMMUNITIES

A. The Method

The preceding theoretical exploration suggests that one can test for saturation by examining the relationship between local richness and some reasonable estimator of the species pool. Regional richness has been the most commonly used estimator for the reasons stated earlier. To test for saturation, matched habitats are compared in different geographical regions of similar size that support different numbers of species. Habitats in richer regions are presumably exposed to richer species pools. Choosing regions of similar size should reduce any differential effects of colonization barriers that might generate differential dispersal limitation among regions (see Cornell and Lawton, 1992; Caley and Schluter, 1997). If the communities are unsaturated, local richness should increase steadily with increases in regional richness. The shape of the relationship can be linear (Type I) or curvilinear, but it cannot level off. Curvilinearity would suggest some resistance to invasion and thus a tendency toward saturation in richer regions, but that the limit to local richness has not been reached. If communities in the richest regions are saturated, then the curve will reach an asymptote and local richness becomes independent of regional richness (Type II).

The method assumes that matched habitats and species in different regions are similar in ways that are critical to species coexistence. This means that if there is the potential for saturation at some level of local richness determined by local environmental conditions and the characteristics of individual species, that level will be the same from region to region. However, even
if matched habitats are environmentally identical, some communities might exhibit multiple stable points and saturate at different levels, for example, due to differences in colonization order as shown by Drake. If the community is near saturation, multiple stable points might increase the variance around the putative asymptote but should not affect the average tendency for local richness to level off.

B. The Tests

Two recent surveys by Cornell and Karlson (1997) and Srivastava (1999) have evaluated the frequency of Type I and Type II relationships in a variety of species assemblages. The surveys were combined by Lawton (1999) to yield 38 papers. Some of the data were not analyzed strictly according to the protocol presented here, and in these cases interpretation of the pattern was conservative and based on common sense. By conservative estimate, there were at least 23 examples of Type I relationships (wood-boring beetles, tiger beetles, fig wasps, lizards, corals, plants, helminth parasites on introduced fishes, 6 examples of fish, 4 examples of plant-feeding insects, 3 examples of parasites, 2 examples of birds, and a variety of vertebrate taxa (birds, fish, mammals, reptiles) analyzed separately and in combination. By contrast, there were 15 examples of curvilinear relationships, and not all of these were Type II (saturating). The examples are classified as saturating (strong), weakly curved (weak), or uncertain (u) as follows: marine crustaceans (u), deep-sea gastropods (weak), tiger beetles (u), 2 examples of helminth parasites on fish (strong), helminth parasites on amphibians (strong), 2 examples of fish (u), 2 examples of birds (weak), 2 examples of mammals (strong), and 3 examples of higher plants (1 strong, 2 weak).

C. Caveats

Several methodological concerns about the saturation test have been raised by various authors. The most relevant of these to the present discussion are that various artifacts such as nonhomogeneous variance structures, insufficient sampling, and overestimates of the true species pool in richer regions sometimes exaggerate the curvilinearity of the relationship between local and regional richness. To the extent that such artifacts are present in some studies, true Type I relationships might appear curvilinear, and curvilinear but unsaturated relationships might appear to be Type II. The count of unsaturated patterns in the survey is thus a conservative and robust result, at least with respect to the presence of such artifacts.

D. Saturated or Unsaturated?

Given these caveats, at least it can be said that unsaturated patterns are common and widespread in ecological communities. Indeed, unsaturated patterns have been reported about twice as frequently as saturated ones. The survey thus confirms that local communities are often open to regional influences and have not reached any upper limit set by species interactions as predicted by the Volterra–Gause model. Those communities that have reached an asymptote might be saturated but have to be examined closely to rule out artifacts that might be responsible for the Type II shape. A Type II curve is not sufficient in and of itself to conclude that species interactions are limiting local richness. For example, helminth parasite communities on fish and amphibians appear to be saturated, but at least one expert opinion is that they are probably not (see Lawton, 1999).

The survey also supports the theoretical position that strong interspecific interactions are not always sufficient to generate a curvilinear or Type II relationship. Taxa that are known to experience strong interspecific interactions such as fish and birds show Type I or weakly curvilinear relationships that are characteristic of unsaturated communities. Although it is not certain that these particular assemblages were experiencing the strong interactions that are supposedly characteristic of these taxa, the results at least suggest that evidence for species interactions by itself might not permit generalizations about the effects of these interactions on broad patterns of community structure.

E. Experimental Support

Additional evidence from experimental introductions supports the survey results that communities are often unsaturated. Experimental manipulations of the species pool are only just beginning, but they are particularly useful because there is better control of confounding variables. In separate studies by Tilman (1997) and Robinson, Quinn, and Stanton (1993), seeds of native species were added to experimental plots containing different numbers of plant species and establishment was monitored. In the Robinson et al. study, California poppy was the test colonizer, and in the Tilman study, anywhere from 0 to 54 species were added to plots.
Neither study showed evidence of saturation; seeds became established and persisted in many plots. Unsaturation was dramatic in the Tilman study, as the number of species in the experimental plots continued to increase over the full range of species additions and the increases persisted for 4 years (Fig. 4). In other words, an asymptote characteristic of the Type II relationship was never reached. However, plots that were richer in resident species had lower establishment rates, suggesting some resistance to invasion.

The results of these experiments must be interpreted with caution, as they represent short-term responses to pulsed increases in the species pool. There are almost certainly time lags in the response of the community to species additions, and long-term monitoring will be required to confirm that the “invaders” have become established without driving “residents” extinct over ecological time. Nevertheless, pool manipulations offer a powerful probe for saturation and need to be expanded to other taxa and habitats to test the generality of these results.

III. MATTERS OF SPATIAL SCALE

A central issue in the exploration of saturation in communities is the spatial scale at which local richness should be measured. Ideally, a locality should be ecologically defined. That is, a scale should be chosen within which all species of interest are interspersed and can
potentially interact and within which the local environment is homogeneous. Since different species interact at different scales, and environments that are homogeneous for one species may be heterogeneous for others, the ideal locality will probably have indistinct boundaries. It may circumscribe the number of herbivore species associated with a host plant species, the number of insectivorous birds found in a patch of woods, or the number of flowering plant species associated with a serpentine outcrop. Local samples must be standardized in some way so that they are comparable among regions. Typically, the number of individuals or the unit area assessed is uniform, or local assemblages are sampled until the species/sampling curve levels off.

If regional richness is sufficiently high to engender saturation, it is not patently clear at which scale it should occur. Communities are hierarchically structured assemblages of organisms that may be saturated at one scale and not other scales because the intensity of dynamic coupling among species varies with the unit of study. Most comparisons of local and regional richness offer few clues about the relationship between hierarchical structure and saturation because local richness is evaluated at only one scale. Clearly, a more methodological exploration of hierarchical structure that involves multiscale sampling is required. A few studies have made limited attempts at a multi-scale analysis. Caley and Schluter (1997) measured local richness from distributional data of 10 different taxa at two local scales (1% and 10% of regional area) in various regions. All taxa fell along a Type 1 trajectory at both scales (Fig. 5). In a 1996 paper Kennedy and Guégan showed that parasite communities reached a hard upper limit of four species in individual cells whereas parasite richness in the cell population as a whole continued to increase steadily with regional richness. Finally, data on coral assemblages examined by Karlson and Cornell (2000) showed that local richness was less sensitive to regional richness in 1-m² quadrats than in 10-m line transects. In the latter example, the 1-m² quadrat is so small relative to the size of some coral colonies that local richness, particularly in very rich regions, could be limited simply by the number of individuals that can be packed into this small sample. In all three examples, only two local scales were examined, but for two of these examples, this was sufficient to show differential responses of local richness to changes in regional richness. Clearly, multiscale analyses will permit stronger inferences regarding saturation, or lack thereof, in communities. They will also be useful in defining the spatial arena within which local processes, which might in some cases limit local richness, should be sought.

IV. SYNTHESIS OF COMMUNITY AND REGIONAL PERSPECTIVES

A. A Thought Experiment

Since processes operating at scales larger than the local habitat (e.g., landscape or region) almost surely affect local richness, research emphasis must shift to these larger-scale processes. Only then can patterns of community assembly be completely understood. However, this shift does not imply that local processes should be ignored. Local- and larger-scale processes may both affect local richness and their relative influence must be determined. The advantages of this multi-scale perspective can be illustrated with a thought experiment.

Suppose we have two climatically similar regions with different regional richness due to historical factors. The Asian and European temperate forests discussed previously provide a good example. Suppose further that physically identical habitats of two types (e.g., sandy versus loamy soils) were transplanted into these regions and were allowed to achieve a steady state that showed no influence of initial colonization history. Once the steady state was reached, by how much would per locality tree species richness vary among habitat types and regions? The transplanted habitats would make it possible to partition the contribution to per-locality richness among processes operating at local and regional spatial scales. The design would also make it possible to evaluate interactions between regional and local factors and to examine mediation of species interactions by changes in the local environment.

B. Manipulations of Productivity and the Pool

Although the preceding experiment is possible in principle, it is likely to be difficult to carry out. Also, it is difficult to be sure that climatic conditions that may affect local processes are truly identical among regions. An alternative to habitat transplantation is a two-factorial manipulation of the species pool and local environmental conditions in the same region. For example, suppose some high-productivity lakes in a particular region had unusually low fish species richness. Is the low richness due to interaction mediation or pool limitation? Manipulations could be designed in which the entire fish species pool could be added to lakes of different productivity. In addition, productivity itself could be manipulated by adding fertilizer to particular lakes. Care would have to be taken that the fish were adapted
to live at the prevailing productivity levels. A full two-factorial design would make it possible to sort out pool effects, productivity effects, and pool X productivity interactions on local richness (see Section 1). If regional control of community structure is important, then pool effects or pool X productivity interactions should have important effects on local richness.

C. A Comparative Approach

In situations where experiments are intractable or when it is desirable to examine longer-term effects, the problem can be addressed using a statistical comparative approach. For example, Schluter and Ricklefs (in Ricklefs and Schluter, 1993) have suggested using multifactorial methods to partition the variance in local richness into components that can be explained by habitat differences within regions and species richness differences among regions. To the extent that local environments constrain local richness, species richness in similar habitats from different regions should converge relative to some hypothetical ancestral state of the community (Fig. 6). Detecting convergence is thus equivalent to measuring the variance component due to habi-
tat differences. Important regional effects on local richness will be detectable in the variance component due to differences in regional richness. Schluter and Ricklefs calculated variance components for 36 different studies and found strong convergence in many cases, consistent with significant local effects on local richness. Yet there were strong regional differences in richness within similar habitats as well. Using regression methods to study coral reef communities, Karlson and Cornell (1998) similarly found both local and regional effects to be important.

D. Conclusions

The results from experimental and statistical analyses confirm that both local and regional processes have important effects on local richness. Thus, not only must experimental work on local communities be continued, but investigations need to be expanded to encompass larger-scale processes as well. The Volterra–Gause image of an ecological system at equilibrium and independent of historical processes has undeniable appeal, and has engendered much productive research on community assembly. Nevertheless, integration of historical, biogeographical, and metacommunity processes with local interactions and environmental variation is essential for a complete understanding of richness variation in species assemblages.

See Also the Following Articles

BIogeography, overview • Diversity, molecular level • Diversity, organism level • Energy flow and ecosystems • Landscape diversity • Temperate forests

Bibliography

Diversity, Community/Regional Level


GLOSSARY

background selection  Selection that takes place at loci other than that of interest. Such events can strongly increase the intensity of genetic drift and hinder the action of selection at a locus of interest. These effects are normally modeled by reducing the effective population size applicable to the locus of interest.

divergence  A fixed genetic difference between two species or populations, or the process of evolving such a difference.

effective population size  The size that a random mating population must have to experience the same intensity of genetic drift as a population of interest. The size can vary across genes due to changes in background selection across the genome.

functional constraint  A limit to the kinds of nucleotides or amino acids which can appear in a gene or protein without compromising function; often called "selective" constraints.

genetic drift  Random changes in gene frequency that occur when a finite number of zygotes or propagules is formed by random sampling gametes or genotypes from the previous generation to create the next generation.

mutation  A change in the nucleotide sequence of a DNA or RNA molecule.

nearly neutral mutation  A mutation whose population genetical dynamics are influenced by both genetic drift and selection.

neutral mutation  A mutation whose population genetical dynamics are influenced only by genetic drift.

nonsynonymous mutation  A nucleotide change in a coding region that changes an encoded amino acid.

polymorphism  A genetic difference between two or more individuals of the same species; to exclude new mutations from the definition, the underlying mutations are required to segregate above an arbitrarily set minimal frequency.

replacement  Divergence due to a change of an amino acid in a protein, or the process to evolve such a change.

selection  Differential reproduction within a set of genes, individuals, groups, etc. caused by differential performance in fitness determining tasks. The appropriate unit of selection (gene, individual, or group) is determined by the level at which the causation of the fitness differences must be described (interactions or performances at the level of genes, individuals, groups, etc.).
substitution Divergence due to a change of base in a DNA sequence, or the process to evolve such a change.

synonymous mutation A nucleotide change in a coding region that does not change the encoded amino acids.

THIS ARTICLE DEALS WITH MOLECULAR DIVERSITY at the DNA and protein levels and will concentrate on diversity that is genetically heritable. Therefore, molecular diversity shown by other macromolecules and/or that is not heritable will not be discussed. However, this does not imply that the latter diversity is biologically insignificant. Indeed, much of molecular diversity is of this kind, with secondary metabolite profiles in plants and antibody diversity generated via mutation and genetic rearrangements in somatic cells of vertebrates being two outstanding examples. We discuss molecular diversity patterning and causation primarily at the level of single genes and genomic regions. Therefore, important aspects of molecular diversity that concern higher level genomic structuring and evolution will not be treated systematically (e.g., gene duplication, multigene family evolution, concerted evolution, genome size evolution, and mammalian isochore structure).

The article will first deal with DNA and protein diversity within populations and among populations of the same species (molecular polymorphism) and then with that caused by genetic differentiation between species and higher taxa (molecular divergence). Such order is justified because visualizing diversity in this way helps in understanding the evolutionary and population genetic forces and processes that generate it.

The information presented here is almost exclusively drawn from knowledge gathered by molecular evolutionary biologists and molecular population geneticists. Molecular evolutionary research aims at characterizing the degree and forms of molecular divergence, at clarifying how such differentiation takes place, and—by taking advantage of the former—at reconstructing the history of DNA sequences and proteins and the genealogical relationships among extant and sometimes extinct organisms. Molecular population genetics studies the form and extent of molecular polymorphism and tries to identify the forces that pattern it.

I. INTRODUCTION

Phenotypic differences between organisms are due to environmental and genetic differences, and the latter are the result of genetic divergence or polymorphism. When due to divergence, genetic differences often can tell us about the physical and geographical environment in which organisms evolved. When due to polymorphism, they can tell us about the ecological arena that is shaping the genetic constitution of a species (gene pool). Realizing that polymorphism is a common and striking phenomenon in the living world is not only important when trying to make sense of molecular diversity. Indeed, the transition from typological (essentialistic) to "variational-populational" thinking was a major contribution of the Darwinian revolution to biology.

Measuring genetic diversity within and between species is a long-standing interest of biologists. However, this only became possible in a systematic way in the 1960s when novel techniques from molecular biology (protein sequencing and enzyme electrophoresis) began to be applied to the study of evolutionary and population genetic questions. These techniques opened the possibility to study variation at the molecular level (allozyme variation and amino acid sequence divergence) and gave birth to the fields of molecular population genetics and molecular evolution. The availability of protein sequences from various taxa triggered sophisticated studies of the forces driving protein evolution and led to the first reconstructions based on protein sequence data of historical relationships among organisms (phylogeny). However, protein sequencing could not be used for quantitatively oriented studies of polymorphism because of its laboriousness. Electrophoretic studies of variation were carried out in hundreds of taxa and confirmed less conclusive previous observations that hinted at the presence of large amounts of genetic variation within populations. Electrophoresis, however, often failed to detect some amino acid variation so that it could not deliver data of sufficient quality to study the nature of the evolutionary forces maintaining variation and the importance of such variation for the process of adaptation.

The advent of DNA sequencing in the 1980s eliminated resolution problems and made it possible to describe exhaustively the genetic variation segregating within gene loci. Before DNA studies this was not possible, and in fact investigators developed and tested models that predicted patterns for electrophoretic variation segregating at many loci because only such data could be produced in statistically analyzable counts. The shift to DNA studies allowed researchers to quantify conclusively not only amino acid polymorphism but also coding region variation that does not change the encoded amino acids (synonymous variation) and variation in noncoding regions, both of which were found to occur
in copious amounts. This increased drastically the amount of data obtainable from a single locus and thus the number and the statistical quality of contrasts among classes of variation subject to arguably different functional constraints. New and possibly ultimate imagery was developed to describe the processes shaping DNA-level variation within and between populations and new, often more powerful modeling approaches were developed and/or given a more solid empirical grounding (e.g., coalescence and maximum likelihood estimation of genetic distances). However, despite all the technical and conceptual advances in describing and quantifying molecular variation, we still know very little about the genetic basis of adaptation and about the relationship between genetic and phenotypic differences among organisms.

II. THE GENERATION AND MAINTENANCE OF GENETIC DIVERSITY

There are three main sources of genetic variation in natural populations: mutation, recombination, and migration. Mutation is the only mechanism that can generate variation in a genetically homogeneous population since recombination and migration only shuffle and shuffle around preexisting variation, respectively. Mutation is therefore the original source of all genetic variation.

In order to understand the evolutionary and population genetic forces patterning genetic variation, it is useful to consider variation at two main levels: within and between species. The first level constitutes the polymorphism component of genetic diversity, which can also be considered at finer levels if gene flow is structured within the species under study (“geographic” structure). Within sexual populations this type of diversity is more or less freely combined each generation to form individual genotypes. The second level constitutes the divergence component of genetic diversity, which can also be considered at multiple levels depending on how closely related are the species under study. This type of diversity consists of fixed differences between species and is only exchanged in extraordinary circumstances (horizontal gene transfer and introgression).

The patterns of genetic variation within species arise from the combined action of two main sets of processes: (i) a molecular-level origination process in which a mutation in a DNA sequence arises and (ii) a population-level “fate” process that determines whether the mutation is quickly eliminated from the population or enters the pool of intraspecific variation becoming a polymorphism. The fate process also includes the possibility that the mutation becomes fixed in the population, i.e., part of the divergence component of diversity. If a new mutation is passed on to the next generation of individuals, its final fate depends on the stochastic error produced when creating a finite number of zygotes or propagules by sampling gametes or genotypes from the previous generation (random genetic drift), on demographic forces such as gene flow and changes in population size, as well as on the deterministic action of natural selection if the mutation is deleterious or advantageous.

III. DIVERSITY AS POLYMORPHISM

What is the quantity and the quality of genetic variation found in natural populations? What are the evolutionary forces responsible for the observed patterns of variation in natural populations? What is the meaning of that variation? What is the relationship between adaptive changes, phenotypic changes, and extant genetic variation? These are some of the questions asked by evolutionary geneticists. The following sections provide examples and summarize the current answers to the first two questions but do not elaborate on the last two questions given the un conclusiveness of the current knowledge on the subject.

A. Empirical Observations

Fruit flies (Drosophila sp.) are by far the best studied group of organisms with respect to their patterns of heritable molecular diversity. Two randomly chosen homologous sequences of Drosophila melanogaster show on average four nucleotide differences per kilobase (kb). A comparison of two full haploid genomes (i.e., including the X chromosome) would give a total of $7.2 \times 10^5$ nucleotide differences over the 180-Mb of a typical haploid genome on average, of this species. Drosophila simulans, a closely related species, has twice as much variation: eight differences per kilobase and an average of 1 million nucleotide differences between any two haploid genomes (the D. simulans haploid size is 130 Mb).

Studies of other taxa are limited to few genes or genomic regions but nevertheless allow interesting contrasts. Humans, for instance, have less variation than fruit flies (one difference per kilobase), but two average human haploid genomes could differ at more nucleotide sites (3.4 Mb) because the human haploid genome size is much larger (3400 Mb). On the other hand, the few
nuclear loci studied to date in the common and pygmy chimps (Pan troglodytes and P. paniscus) show about three times more nucleotide variation than do the homologous loci in humans. In plants, one of the genes encoding the enzyme alcohol dehydrogenase (Adh1) shows levels of polymorphism that range from 1.8 differences per kilobase in wild barley (Hordeum vulgare) to approximately 18 differences per kilobase in maize (Zea mays). In contrast, the homologous gene in D. melanogaster shows an average of 6.5 differences per kilobase. These quantities describe more or less the typical range of DNA polymorphism in natural populations.

The question about the quality of the variation (i.e., about the functional consequences of the observed nucleotide polymorphisms) is normally approached by extrapolating from a priori knowledge about the genetic code and from the conservation of certain nucleotides and nucleotide motifs in noncoding regions. For instance, amino acid changes are counted separately from synonymous changes, and changes violating the phylogenetically or functionally determined canonical motifs of regulatory sequences are distinguished from those that do not. With respect to identifying selective pressures affecting different variants, the frequency pattern of mutations (are variants rare or common?) can be very informative. For instance, an indication of deleteriousness is the occurrence in different samples of a certain kind of mutation only at low frequencies (i.e., in only one or few of the sampled individuals).

Drosophila data show that in noncoding regions (flanking regions and introns) there is more variation than in coding regions (2.5 and 2.4 times more variation in D. melanogaster and D. simulans, respectively) and that synonymous polymorphism is higher than noncoding polymorphism (1.2 and 1.6 times higher, respectively), and higher than amino acid polymorphism (2.7 and 7.6 times higher, respectively). Most amino acid variation tends to be found at low frequencies and so are silent changes away from the major (preferred) synonymous codon within a codon family (i.e., toward minor, preferred codons) in genes with biased usage of synonymous codons.

These observations about patterns of variation can be very informative when asking questions such as the following about the nature and sometimes the magnitude of the evolutionary forces generating the patterns: Why do some species have more variation than others? Is it due to differences in the rate of origin of mutations? To differences in the size of their populations? To recent evolutionary, demographic, and ecological events? To the species’ distributions over geographic space? To their breeding system? To the genomic context in which the mutations occur (e.g., the intensity of recombination and of deleterious mutation in and around the studied locus)? The latter possibility has been shown to apply in Drosophila by two recent studies that found a positive correlation between levels of variation and the rate of recombination.

B. Molecular Biological Considerations

The mutation process is known to be affected by physiological and life history characteristics of the organism which affect the rate and accuracy of DNA replication (generation time, number of divisions before gametogenesis, sex, and age) and by factors that increase damage to DNA (metabolic rate, age, transcription rate, and the frequency and degree of packing and unpacking of chromosomes for expression). DNA damage is caused by exogenous and endogenous agents. Exogenous agents include UV light, ionizing radiation from X-rays, heat, alkylating agents, and hydroxyl radicals originated by the interaction of ionizing radiation with water. Endogenous agents include oxidant by-products of normal metabolism. Changes generated by endogenous agents cause the highest fraction of DNA damage. In human cells, oxidants damage between 10⁴ and 10⁵ nucleotide sites every day. Such damage would be extremely deleterious were it not for the presence of enzymes involved in finding and repairing it. Of the five major types of DNA mutations—substitutions, deletions, insertions, inversions, and duplications—nucleotide substitutions and single base pair deletions or insertions (all called point mutations) account for 85% of the mutation events.

With the exception of macromolecular events such as chromosomal rearrangements, most mutations are the result of errors made by enzymes involved in DNA replication and repair. Replication errors are often due to the incorporation of noncomplementary nucleotides by the DNA polymerase and incorrect subsequent proofreading by the same enzyme. The latter process, however, is normally extremely effective at lowering the error rate by a million-fold relative to the intrinsic misincorporation rate of the polymerase. Table I shows the range of mutation rates observed in several organisms.

C. Cytological and Ontogenetic Considerations

In viruses, prokaryotes, and unicellular eukaryotes—with the notable exception of protists such as Parame-
Table 1

<table>
<thead>
<tr>
<th>Organism</th>
<th>Mutation rate per base pair</th>
<th>Mutation rate per genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteriophage Qb (RNA)</td>
<td>$1.5 \times 10^{-3}$</td>
<td>6.5</td>
</tr>
<tr>
<td>Influenza A virus (RNA)</td>
<td>$&gt;7.3 \times 10^{-3}$</td>
<td>1.0</td>
</tr>
<tr>
<td>Bacteriophage M13 (DNA)</td>
<td>$7.2 \times 10^{-7}$</td>
<td>0.0046</td>
</tr>
<tr>
<td>Bacteriophage T4 (DNA)</td>
<td>$2.4 \times 10^{-6}$</td>
<td>0.0040</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>$5.4 \times 10^{-10}$</td>
<td>0.0023</td>
</tr>
<tr>
<td>Neurospora crassa</td>
<td>$7.2 \times 10^{-14}$</td>
<td>0.0030</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>$2.3 \times 10^{-16}$</td>
<td>0.018</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>$3.4 \times 10^{-16}$</td>
<td>0.058</td>
</tr>
<tr>
<td>Mouse</td>
<td>$1.8 \times 10^{-16}$</td>
<td>0.49</td>
</tr>
<tr>
<td>Human</td>
<td>$3.0 \times 10^{-11}$</td>
<td>0.16</td>
</tr>
</tbody>
</table>

*After Drake et al. (1998).*

D. Factors Influencing the Fate of Mutations

Patterns of molecular diversity other than those created by a continuous flow of short-lived mutations which remain unique to the individuals in which they arise can be very striking. For instance, some of the loci encoding the antigen-presenting major histocompatibility complex (MHC) in humans show extremely high levels of amino acid and synonymous polymorphism as well as many alleles segregating at intermediate frequencies. On the other hand, loci in the human Y chromosome show little polymorphism and mainly low-frequency mutations.

In general, patterns of polymorphism are determined by mutation pressure, genetic drift, and natural selection. The effectiveness of the action of the last two forces depends on the effective population size (see Glossary) and on the fitness effect of the mutation involved. In particular, the transmission genetical sampling effect at the basis of genetic drift phenomena increases in strength with decreasing population size and this in turn hinders the deterministic action of selection.

The patterns of variation of selectively neutral mutations are totally determined by genetic drift, population size, and the rate of mutation to neutral variants. For selected mutations both drift and selection can affect the patterns. In the case of strongly deleterious mutations, the observed pattern is one of mutations unique to each individual and none segregating at high frequencies because such mutations remain in the population for only very few generations. When a steady state is reached between the number of new mutations being produced by mutation and those being removed by selection, one speaks of a mutation—selection balance. There should not be any observable polymorphism patterns for strongly favorable mutations because such mutations are normally fixed soon after being introduced in the population (if they are not immediately lost to drift) and practically do not participate in polymorphism. However, favorable mutations can be maintained at intermediate frequencies, e.g., when the heterozygote has higher fitness. In this case, one speaks of a balanced polymorphism (as in the MHC case previously mentioned). For weakly selected mutations, genetic drift and selection are important. Note that the previous considerations are mainly applicable to recombining populations because in asexual populations and in cases with very weak selection and low recombination intensity, situations can arise that require considering patterns and forces at the level of the whole genome or of long stretches of the genome.

E. Models of Molecular Evolution

The main models that make predictions about the patterns of polymorphism and divergence are the neutral
theory proposed in the late 1960s by M. Kimura, the nearly neutral theory proposed in the early 1970s by T. Ohta, and a series of models of natural selection developed by J. Gillespie during the past two decades. The first two models emphasize the importance of genetic drift and mutation, whereas the latter group of models stress the role of natural selection. All models agree that the vast majority of newly arisen mutations are deleterious, but they differ greatly in their assumptions about the relative proportions of neutral and favorable mutations that participate in polymorphism and divergence.

1. Neutral Theory

Kimura's neutral theory postulates that both the fixation of genetic differences between populations and the patterning of genetic variation within populations are the result of random changes in the frequency of neutral alleles driven by genetic drift. This model assumes that most new mutations are either deleterious or neutral and that advantageous mutations arise quite infrequently and are quickly fixed or lost to drift so that they are not important for polymorphism. Most of the theory's predictions for polymorphism patterns assume mutation–drift equilibrium, i.e., that the rate of allele loss to drift is balanced by the production of new alleles by mutation. The predictions therefore apply only when the population size and the mutation rate have been constant for a long period of time.

The theory predicts that the amount of polymorphism within a population is proportional to the effective population size \( N_e \) and the mutation rate \( \mu \). Nucleotide polymorphism is normally expressed in terms of the parameter \( \theta \) \( (= 4N_e \mu) \), which can be interpreted as the average number of differences per nucleotide site between two alleles randomly chosen from the population. This divergence is the result of nucleotide mutations that have accumulated since the two alleles last shared a common ancestor. Theory has shown that the expected time back to this ancestor for an autosomal locus in a diploid is \( 2N_e \) generations so that when each allele has been evolving independently for \( 2N_e \) generations, the expected number of nucleotide differences between them is \( 2 \times 2N_e \) times the mutation rate \( \mu \), i.e., \( 4N_e \mu \).

The neutral theory explains the fact that different genes and regions of the genome show different levels of polymorphism by assuming that the neutral mutation rate varies among such regions and genes because they are under different functional constraints. Recent simulation work has led to a restressing of the realization that the effective population size can be reduced by selection against deleterious mutations at linked and unlinked sites ("background selection"). This effect can compel neutral polymorphism to change drastically across regions that differ in recombination intensity and deleterious mutation production.

2. Nearly Neutral Theory

Ohta's model, also known as the slightly deleterious allele model and considered by many to be the "dominant paradigm for molecular evolution," was proposed in the early 1970s in order to account for two observations that could not be explained by the neutral model: the generation time effect for DNA divergence and the lower than expected levels of protein variation in natural populations. The model deviates from the neutral theory mainly because it assumes (i) that most observed molecular variation is slightly deleterious, (ii) that mutation rates are correlated with the generation time (g), and (iii) that the rate of molecular divergence for slightly deleterious alleles is inversely proportional to \( N \). The nearly neutral theory is biologically realistic because it proposes plausible correlations of population genetic parameters such as \( g \) and \( N \) with the origination and fate processes of molecular evolution, and because it avoids the unrealistic assumption of complete neutrality for most variation participating in polymorphism and divergence.

Slightly deleterious alleles, also called weakly selected alleles, are defined as those with selective coefficients close to the neutral boundary \( (s \approx 1/N) \) and are seen as major contributors to polymorphism and divergence in natural populations. The neutral theory stresses that the fate of neutral variants depends only on random drift, whereas the slightly deleterious model builds on the fact that the fate of slightly deleterious variants depends on the joint action of both drift and selection. Which of these two processes is more important depends on the effective size of the population because the efficacy of selection in removing slightly deleterious variants and fixing advantageous ones is higher in larger populations. In smaller populations slightly deleterious mutations can increase in frequency and become fixed by genetic drift, but in large populations selection prevails and eliminates them quickly. Therefore, the nearly neutral model predicts that if most of the observed variation in natural populations is slightly deleterious, polymorphism patterns should be dominated by rare variants.

The lack of correlation between genetic variation and population size (especially in large populations) was explained by this model with the argument that when divergence and variation involve mainly slightly
deleterious alleles, the amount of diversity in a population should be proportional to \( N \) but nevertheless reach a plateau when \( N \) is large enough for selection to impede that slightly deleterious variants increase in frequency. A main problem with this interpretation is that in organisms with very large population sizes one would not expect any variation if most polymorphism is due to slightly deleterious variants (at least no amino acid polymorphism), which is clearly not the case.

3. Selection Models

Gillespie proposed and developed a series of selection models that are less terse in their assumptions than the neutral and nearly neutral models but that can account for most of the empirical observations about molecular polymorphism and divergence. The models assume that balancing and episodic selection dominate the patterning of divergence and polymorphism at the molecular level. They predict that fluctuations of the environment can lead to the maintenance of polymorphism by balancing selection and, in the right circumstances (when the selective advantage of an allele is greater than the variance of the fitness effects of the alleles at that locus), to the fixation of favorable alleles. The models assume fitness effects that are underdominant, overdominant, or that change in a random environment (SAS-CFF model) or in a temporally fluctuating environment (TIM model). The addition of environmental variability is biologically realistic since the existence of fluctuations in the fitness value of segregating alleles due to changes in the environment over ecological timescales is quite plausible. These models predict patterns of polymorphism similar to those predicted by the neutral models, which makes hypothesis testing very problematic.

F. Statistical Tests on Molecular Polymorphism

Of all the models of molecular evolution, only the neutral theory makes simple quantitative predictions about expected levels and patterning of genetic polymorphism. This fact makes the neutral model a very popular null hypothesis against which to test the importance of other evolutionary forces in shaping patterns of polymorphism. There are two basic families of statistical tests for neutrality: (i) models that test neutrality using polymorphism data from a single locus and (ii) models that test neutrality using polymorphism and divergence data.

The data required for the single locus tests are homologous DNA sequences collected from individuals of a single species. This family of tests compares three estimates of the neutral parameter \( \theta \), which equals \( 4N\mu \) for an autosomal locus in a diploid. The three estimates are straightforwardly calculated from (i) \( \pi \), the average number of pairwise differences between the sampled alleles which should be equal to \( \theta \); (ii) \( S \), the number of polymorphic sites in the sample; and (iii) \( \eta \), the number of mutations present in only one of the sampled alleles (singletons). Under neutrality the values of these estimates should be the same. The tests therefore ascertain whether departures from zero of the differences between any two of these estimates are statistically significant for the sample at hand.

Tajima, for instance, proposed comparing the values of \( \pi \) and \( \theta \_ {\text{e}} \) (\( \theta \) estimated by \( S \)) using his statistic \( D \), which is the difference between \( \pi \) and \( \theta \_ {\text{e}} \), normalized by dividing the statistic by a factor proportional to the variance of the difference. Negative values of \( D \) are expected in the presence of purifying selection or following a purging of variation from the region by selection driving the quick fixation (1) of a new mutation or (2) a rapid increase in the frequency of an allele previously present at a lower frequency (selective "sweep"). Positive values of \( D \) are expected with balancing selection and population subdivision.

Although the patterns predicted by the neutral theory are always used as null hypothesis, the patterns produced by the two main kinds of selection regimes and/or events (sweeps and balanced polymorphism) have also been described thoroughly, and it has sometimes been striking how well they match the data. For instance, selective sweeps tend to reduce neutral variation in regions linked to the selected site. This "hitchhiking" effect on neutral variation (in this case, non-hitchhiking because most variation does not manage to get a "ride") is more pronounced in zones with low recombination. Balancing selection increases the levels of variation over the neutral expectation in the sites adjacent to the balanced site by making the times to the common ancestor at these linked sites more similar to that of the balanced site (i.e., to the time since the establishment of the polymorphism) than to the neutral time at unlinked sites. If the polymorphism is older than \( 2N \) generations, one should observe an increase in neutral variation in the region of the balanced site relative to the neutral expectation. Another phenomenon that can affect the level, but much less so the pattern, of neutral variation is "background selection." This effect and regular sweeps may account for the low levels of polymorphism observed in regions with low recombination.
IV. DIVERSITY AS PHYLOGENETIC DIVERGENCE

How could genes, proteins, and genomes become so different, and what limited how different they could become? Are there heterogeneities in the rates of divergence among different proteins, genes, and genomic regions? Do molecules from different organisms evolve at different rates? What are the mechanisms of nucleotide and amino acid substitution? Can we infer the genealogical relationships among different organisms from molecular data? Can we use molecular divergence data to date evolutionary events? These are some of the questions addressed by researchers in the field of molecular evolution. The following sections provide examples and summarize the current answers to the first three questions, i.e., those most relevant to understand the evolutionary dynamics of molecular diversity.

A. Some Empirical Observations

One of the main patterns of divergence that have been described by protein and DNA sequencing studies is that rate of amino acid replacement in different kinds of proteins and in different organisms are extremely variable (Table II). Histone H4, a protein involved in the packaging of DNA in chromosomes, is a striking case of an extremely conserved protein. Comparisons of H4 histones from plants and mammals show that only two amino acid replacements, over the 100 sites compared, have taken place since the two groups last shared a common ancestor about 1 billion years ago. On the other hand, viral envelope proteins, which are under constant challenge by the host immune system, change at extremely high rates—between 1 million to 10,000 times faster than the fastest mammalian protein (interferon-γ).

Table II also shows that rates of substitution at synonymous sites are higher than rates at nonsynonymous sites. This is expected because amino acid changes often disrupt the function of the protein, whereas synonymous changes are less likely to have comparable consequences. This explanation is corroborated by the observation that noncoding regions and pseudogenes (deactivated genes) evolve with rates similar to those of synonymous changes. However, in several proteins amino acid replacements occur more frequently than

<table>
<thead>
<tr>
<th>Organism</th>
<th>Protein</th>
<th>Nonsynonymous rate</th>
<th>Synonymous rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-rodent</td>
<td>Histone 4</td>
<td>0.00 × 10⁻⁴</td>
<td>3.94 × 10⁻⁹</td>
</tr>
<tr>
<td></td>
<td>Insulin 4</td>
<td>0.20 × 10⁻⁹</td>
<td>3.03 × 10⁻⁹</td>
</tr>
<tr>
<td></td>
<td>Hemoglobin α chain</td>
<td>0.56 × 10⁻⁴</td>
<td>4.38 × 10⁻⁹</td>
</tr>
<tr>
<td></td>
<td>Interleukin-1</td>
<td>1.90 × 10⁻⁹</td>
<td>3.27 × 10⁻⁹</td>
</tr>
<tr>
<td></td>
<td>Ig κ chain</td>
<td>2.03 × 10⁻⁴</td>
<td>5.56 × 10⁻⁹</td>
</tr>
<tr>
<td></td>
<td>Relaxin</td>
<td>2.59 × 10⁻⁹</td>
<td>6.29 × 10⁻⁹</td>
</tr>
<tr>
<td></td>
<td>γ-Interferon</td>
<td>3.06 × 10⁻⁹</td>
<td>5.30 × 10⁻⁹</td>
</tr>
<tr>
<td>Drosophila</td>
<td>Adh</td>
<td>0.90 × 10⁻⁹</td>
<td>9.50 × 10⁻⁹</td>
</tr>
<tr>
<td></td>
<td>Esterase 6</td>
<td>3.02 × 10⁻⁹</td>
<td>21.52 × 10⁻⁹</td>
</tr>
<tr>
<td>Plants</td>
<td>Grasses Adh1</td>
<td>0.32 × 10⁻⁹</td>
<td>7.00 × 10⁻⁹</td>
</tr>
<tr>
<td></td>
<td>Grasses Adh2</td>
<td>0.89 × 10⁻⁹</td>
<td>5.99 × 10⁻⁹</td>
</tr>
<tr>
<td></td>
<td>Palms AdhA</td>
<td>0.41 × 10⁻⁹</td>
<td>2.61 × 10⁻⁹</td>
</tr>
<tr>
<td>RNA viruses</td>
<td>Influenza A Hemagglutinin</td>
<td>3.59 × 10⁻³</td>
<td>13.10 × 10⁻⁵</td>
</tr>
<tr>
<td></td>
<td>HIV-1 gag</td>
<td>1.70 × 10⁻⁴</td>
<td>9.70 × 10⁻⁴</td>
</tr>
<tr>
<td>DNA viruses</td>
<td>Hepatitis B P</td>
<td>1.45 × 10⁻⁷</td>
<td>4.57 × 10⁻⁷</td>
</tr>
</tbody>
</table>

*Drosophila rates are from comparisons between the obscura and the melanogaster groups (after Li, 1997).*
synonymous substitutions, e.g., in the antigen recognition site of the HLA loci in humans, in two fertilization proteins of abalone, and in proteins involved in male reproduction (Acp26Aa) or associated with male sterility (OdsH) in Drosophila. In the latter locus, for instance, amino acid changes have occurred in less than 1 million years than during the previous 700 million years. The rapid evolution of this gene is likely due to a change in function.

Similarly, different parts of a protein (e.g., protein domains) can diverge more readily than others. A classic example is the hinge pocket and the surface regions in hemoglobin α and β, which show a 10-fold difference in their rates of amino acid divergence. A similar situation occurs in the mitochondrial cytochrome oxidase II, whose transmembrane domain evolves at a much faster rate than the rest of the protein.

Another major pattern evidenced by the data is that different taxa can evolve at quite different rates (lineage effects). For instance, a survey of substitution rates in 54 mammalian genes shows that rodents tend to have higher rates of synonymous divergence than primates, but their nonsynonymous rates are not significantly different. Similarly, sharks show a seven-fold reduction in rates of silent rRNA evolution with respect to primates. Finally, endosymbiotic bacteria associated with aphids show higher rates of amino acid replacement than do closely related free-living strains.

B. The Substitution Process

The fixation of new mutations gives rise to the divergence component of genetic diversity. This process describes how a new allele replaces the previously most common allele in the population, thereby becoming the common ancestor of all alleles at that locus. The dynamics of the substitution process are determined (i) by the rate of origination of new mutations in the population, and (ii) by the probability of fixation of such mutations. The main evolutionary forces affecting this process are mutation, genetic drift, and selection, with their relative importance differing for neutral, weakly selected, and strongly selected mutations.

Neutral autosomal mutants are produced in a diploid population with rate $2N\mu$ per generation, i.e., the number of new mutations produced each generation is equal to the product of the number of genes copies in the population that can mutate, times the mutation rate to neutral alleles per gene per generation $\mu$. The same result applies for selected mutations if we let $\mu$ become the rate of mutation to strongly selected or weakly selected alleles.

The probability of fixation of a mutation depends on the selective advantage or disadvantage of the mutation, on the population size, and on the initial frequency of the mutation. For a newly arisen autosomal mutation, the initial frequency is $1/2N$ in a diploid population, which for the neutral case equals the probability of fixation of the mutation (since any of the $2N$ alleles present at any given moment can become the common ancestor of all alleles in the population some time later). For a nonrecessive mutation with selective coefficient $s$ close to $1/N$, the probability of fixation is about $2s/(1 - e^{-2Ns})$. If the value of $s$ is positive (advantageous mutant) and the population size is large such that $Ns \gg 1$, the fixation probability is twice the selective advantage of the mutant ($= 2s$). Even in this case, however, genetic drift is the main determinant of the fate of new mutations (e.g., 80% of mutations with $s = 0.1$ are lost to drift). For a deleterious mutant (negative $s$) the fixation probability is lower than that in the neutral case and is practically zero for strongly deleterious mutations.

The rate of substitution equals the average number of new mutant alleles entering the population in each generation times the probability that one of them will fix. For a neutral allele, this equals $(1/2N) \times (2N\mu) = \mu$, i.e., the rate of substitution of neutral alleles equals the mutation rate. For selected mutations the rate of substitution equals $4N\mu$, and is therefore dependent on the mutation rate to favorable selected mutations, the population size, and the selection coefficient of the mutant. It is easy to see that for the neutral rate of substitution to remain constant across different populations it suffices that the neutral mutation rate does not change. For selected mutations, however, the substitution rate could vary much more since its constancy requires that three parameters do not change or that each varies in ways that balance changes in the others.

C. Models of Molecular Evolution

From their conception, most theoretical models of molecular evolution dealt with both divergence and polymorphism from a population genetically explicit perspective. In the following sections, we present the predictions for patterns of divergence for the most relevant models (for polymorphism, see Section III).

1. Neutral Theory

Kimura showed that genetic divergence at neutral loci should increase over time due to the fixation of neutral mutations by random drift at a rate that equals the neutral mutation rate $\mu$. This predicted constancy in the
rate of substitution matched the empirical observations showing constant rates of amino acid substitutions per year for different proteins in a wide range of vertebrate lineages that led Zuckerkandl and Pauling to propose the molecular clock hypothesis in the early 1960s. The match is often viewed as positive evidence for Kimura's neutral theory, but this is a misunderstanding since selection models can also produce clock-like divergence.

The observed differences in rates of evolution among different genes and the heterogeneity of divergence rates in different regions of a gene are explained by the theory as being the result of differences in the fraction of all mutants that are neutral for each given gene or region (i.e., different functional constraints).

2. Nearly Neutral Theory

Ohta's model assumes that most molecular diversity is slightly deleterious, and predicts that the rate of production and fixation of neutral mutations is negatively correlated with organismal generation time (generation time effect), and that the rate of divergence for slightly deleterious changes is inversely proportional to the population size. The negative correlation between generation time and mutation rate is expected when mutations occur mainly during DNA replication and the number of germinal cell replications per generation is the same across different organisms. The germinal cells of short-lived organisms should indeed experience more DNA replications per year than those of long-lived organisms so that more neutral mutations can be produced in short-lived organisms. This assumption corresponded well with observations from divergence studies from the early 1970s based on DNA hybridization that showed a strong generation time effect for noncoding DNA evolution. The generation time effect can be obscured in comparisons among distantly related organisms when they differ in developmental mode and germline ontology (see Section III).

An important contradiction that this theory tried to resolve was that protein evolution appears to be clock-like, whereas noncoding DNA evolution shows a generation time effect. The neutral model explained the former by asserting that most amino acid replacements are neutral. The nearly neutral model asserted that protein divergence is mainly due to the fixation of slightly deleterious alleles, so it had to resort to postulating that generation time is inversely proportional to population size in order to account for the constancy of amino acid replacement rates. It was argued that small organisms have higher mutation rates per year due to their shorter generation times but that due to their larger population sizes, they fix slightly deleterious amino acid changes less readily. On the other hand, organisms with longer generation times have lower mutation rates but smaller population sizes so that they fix slightly deleterious mutations more easily. Therefore, the fact that in actual populations generation time correlates negatively with population size lets the rate of amino acid divergence appear clock-like. Although this negative correlation initially failed to be corroborated, recent studies have shown it to hold across different taxonomic levels.

A major problem with this theory concerns divergence: slightly deleterious variants cannot accumulate indefinitely because they would decrease the fitness of the population, thereby giving new favorable mutations a major role in the process. Additionally, in large populations no divergence should occur because in them selection should be efficient enough to eliminate most slightly deleterious alleles.

3. Selection Models

Divergence data led Gillespie to propose his selection models. He showed that there is excessive variation in amino acid divergence rates in different mammalian proteins using the index of dispersion $R(t)$, which is the ratio of the variance in substitution rate to the mean substitution rate. The index ranged between 2.5 and 7.5, whereas the neutral theory predicted it to be approximately 1.0 (as expected for a simple Poisson process). These high values of $R(t)$, which were not confirmed in a recent Drosophila study, led Gillespie to propose that the process of divergence is characterized by bursts of substitutional activity during which alleles previously segregating at high frequencies become rapidly fixed by selection, followed by quiet periods with little substitution and long-lived intermediate frequency polymorphisms. This still allows for a molecular clock if the periods of bursting and stasis succeed each other more or less regularly. One speaks in this case of an episodic molecular clock. Nevertheless, neither these models nor the neutral and nearly neutral models can completely account for the excessive variance of the rate of amino acid replacement in mammals. Improvements of these models will include increasing the time frame over which the fitness effects of mutations change or fluctuate. However, the complexity of the models has already reached the point that they can fit almost any observation, which compromises their value for hypothesis testing. At this point, one will have to assess the plausibility of their assumptions rather than the accuracy of their predictions.
V. COMBINED PREDICTIONS FOR DIVERGENCE AND POLYMORPHISM

Two tests based on the mutual compatibility of divergence and polymorphism patterns have been developed and used with much success to test the neutral theory. When they reject neutrality, outside information can help in further identifying the forces that might have caused the rejection. The Hudson–Kreitman–Aguade (HKA) and the McDonald–Kreitman tests focus on the positive (1) correlation between the (2) amounts of polymorphism and divergence expected under neutrality (given the linear dependence of both patterns on the mutation rate).

The HKA test compares the correlation between polymorphism and divergence at two or more loci. Its rationale is that if loci have different neutral mutation rates, their rates of divergence and levels of polymorphism should be equally affected. Selection is therefore inferred when the ratio of polymorphism to divergence varies too much over loci, a conclusion that requires sophisticated corrections to deal with the fact that, due to stochastic reasons, different loci can have different levels of polymorphism if they are not totally linked.

The McDonald–Kreitman test uses a similar contrast but concentrates on “loci” that are interspersed with each other over the sequence, thereby avoiding the need for the previously discussed corrections and drastically increasing the statistical power. This test can be applied to contrast silent and replacement variation, or unpreferred vs preferred silent changes, but it has mainly been used to compare the former within and between species under the assumption that these two kinds of variation are homogeneously distributed along the coding region studied.

Rejections of neutrality with both tests can be due to departures at the divergence or at the polymorphism level. For instance, an excess of amino acid polymorphism relative to amino acid divergence should be observed if variation is slightly deleterious. However, this “excess” could very well be due to a deficiency in silent polymorphism and/or an excess in silent divergence. If the ratio of amino acid replacement to silent substitution is significantly greater than the ratio of replacement to silent polymorphism, one could infer that natural selection has been fixing amino acid mutants beyond the neutral expectation, but again this could be due to a deficiency in amino acid polymorphism and/or in silent divergence. This is why outside information must be used to identify the causes of a rejection of neutrality by these tests.

VI. OMITTED DIVERSITY PATTERNS

The patterns of molecular diversity presented previously were originally studied in order to test hypotheses about abstract evolutionary and population genetic dynamics. For this reason, they are not necessarily representative of the phenomenon of molecular diversity in a more comprehensive molecular biological, natural historical, and organism–evolutionary sense. This explains, for instance, the lack of information about the evolutionary dynamics of molecular diversity with developmental consequences. Although biologists are well aware that this kind of diversity may be extremely important in an organism–evolutionary sense, the current lack of knowledge about it makes it very unsuitable for evolutionary and population genetic research that focuses on hypothesis testing. The concentration on variation suitable for hypothesis testing also explains this article's lack of an extensive presentation of many important molecular biological and biochemical constraints on diversity patterns. For instance, the dinucleotide motif T/A ("T" is the codon boundary) is very under-represented in most coding regions of Drosophila, yeast, and humans, but no research has been done on how it constrains molecular diversity patterns. However, this constraint is quite likely to affect variation patterns in more coding regions than the high codon usage biases shown by most highly translated genes. It is an important task to increase our knowledge of molecular diversity with respect to molecular biological constraints and in the natural historical and organism–evolutionary sense.

VII. PROBLEMS OF THE MODELS PRESENTED

Most of the neutrality- and selection-based modeling of polymorphism and divergence is based on single-site approaches, and it was not until the 1980s that linkage effects began to be considered. In models with selection, single-site results which assume that no further concomitant mutants are segregating are the standard, and the few efforts to deal with many concomitant mutations assume multiplicative fitness effects of mutations, i.e., that fitness is \((1 - s)^i\) when one carries \(i\) mutations.
The latter assumption is made out of convenience and lack of clear alternatives but is not justified a priori nor has it been justified a posteriori. It is indeed not clear how selection acts on, for example, synonymous or insertion/deletion variation, and one cannot exclude that it acts at the level of the total individual genotype via a possibly complex epistatic relationship between the number of mutations and fitness. Similarly, it is very likely that the assumption that silent variation is weakly selected \( (s = 1/N) \) is incorrect because in fact too many species with too disparate values of \( N \) nevertheless manage to harbor intermediate levels of variation for synonymous changes and to avoid fixing the major synonymous codon or having codon usages solely determined by mutation pressures. All of this could not be possible under the slightly deleterious model if \( N \) changes by more than a factor of 3 up or down from the value \( 1/s \). This point is also stressed by the fact that genes on the fourth chromosome of Drosophila show clear signs of biased usage of synonymous codons despite the fact that the extremely low levels of polymorphism in this chromosome indicate that its local \( N \) should be reduced well beyond 10-fold, making the maintenance of such bias not possible. The \( s = 1/N \) assumption is thus clearly untenable, which does not necessarily imply that more realistic models will not allow nearly neutral dynamics to take place when genes are close, e.g., to an optimum codon usage pattern. Exciting new patterns of polymorphism and divergence will have to be mined in order to address this and many other important questions—an endeavor that will fundamentally bring the evolutionary and population genetical understanding of molecular diversity well beyond its current state.

**VIII. THE VALUE OF MOLECULAR DIVERSITY AND CONSERVATION BIOLOGY**

Diversity on Earth is endangered, and so is molecular diversity at both the polymorphism and the divergence levels. Much has been written about the value of the divergence component of diversity and about the need to preserve the biological information contained in adaptations evolved over millions of years in species and locally adapted gene pools all over the world. Conservation biologists are aware that the polymorphism component of genetic diversity may be important for the survival of a species and therefore try to preserve it. However, that they do it and the way in which they do it is important not only for evolutionary research but also for the other biological sciences. In fact, the polymorphism patterns of many species with large effective population sizes have been shaped during millions of generations by the action of very small selection intensities. Such patterns may take too long to recreate by expanding a population and allowing selection to act for a sufficient amount of time and/or due to effects too weak to study with standard laboratory techniques. The information in these patterns can indirectly guide researchers to subtle but important molecular biological, cytological, and physiological effects of mutations but will be lost if such populations disappear or their numbers are reduced drastically. Therefore, conservation biologists should also strive to conserve large “unendangered” populations.

**See Also the Following Articles**

DIVERSITY, COMMUNITY/REGIONAL LEVEL • DIVERSITY, ORGANISM LEVEL • DIVERSITY, TAXONOMIC VERSUS FUNCTIONAL • PHENOTYPE, A HISTORICAL PERSPECTIVE • POPULATION GENETICS

**Bibliography**


I. The Nature of the Organism

II. Diversity in the Real World: A Precis of Current Organismal Diversity

III. Summary

---

GLOSSARY

entropy An expression or measurement of the energy available for use by a system, including living systems; often described as the extent to which the system tends toward a state of disorder or randomness.

extremophile An organism that requires, or grows optimally in, extreme environmental conditions; e.g., extremes of temperature, pressure, or acidity.

heterotroph Literally, feeder on others; an organism dependent upon organic material from an external source to provide carbon for growth.

organism The smallest entity of life that can function as a whole, distinct from others of the same type; a single living being.

photoautotroph An organism that uses inorganic material as a source of carbon for growth, and light as an energy source; e.g., plants.

---

ORGANISMS ARE THE SMALLEST functional wholes of life, entities that impose themselves on their surroundings and are causally affected by them. Organisms are localized in space and time, connected with their surroundings, including other organisms, through ecological and behavioral interactions and with their evolutionary past through genealogical inheritance. Composed of smaller parts, they are themselves parts of larger wholes, such as demes, populations, species, guilds, communities, and ecosystems. These diverse connections provide opportunities and challenges for any organism-focused efforts to understand, manage, and preserve biodiversity.

I. THE NATURE OF THE ORGANISM

A. The Basic Units of Selection

Long before the advent of modern evolutionary thinking in the nineteenth century, organisms were recognized as a primary focus of observation and explanation of biological systems. Darwin extended this tradition by reserving a central role for organisms and organismal diversity in his theory of evolution by natural selection. He believed that evolutionary change resulted from the interaction of two factors, which he called "the nature of the organism" and the "nature of the conditions." Of these two, Darwin (1872) proposed that the nature of the organism...
as far as we can judge, dissimilar conditions; and, on the other hand, dissimilar variations arise under conditions which appear to be nearly uniform.

(p. 32)

Darwin thought that organisms were historically and developmentally cohesive wholes, and therefore it was in the nature of the organism to produce offspring that were all highly similar (but not identical) to each other and to their parents and other ancestors. He also postulated that reproduction produced variation without regard for environmental conditions and therefore it was in the nature of the organism to produce these offspring in numbers far exceeding the resources available for their support. When this inherent overproduction produced variety in critical characters, natural selection would preserve the versions that were functionally superior in that particular environmental context (adaptations). Whenever an environment changes, those organisms that already had the adaptations necessary to survive would do so, whereas those lacking appropriate adaptations would not. Selection did not create the adaptations—it only determined which ones, if any, would be favored for survival. The production of organismal diversity thus required that organisms be at once autonomous from, and sensitive to, the environment. Darwin's perspective contrasted sharply with Lamarck's proposal that adaptation was an immediate and directed response by organisms to their surroundings. Lamarck also believed that the nature of the organism was important in the production of diversity, but only because all organisms have the same ability to change according to their needs. Therefore, whereas Darwin postulated that the nature of the organism included autonomous, self-regulating properties, Lamarck believed that the nature of the organism was to be directly and completely connected to the environment.

The distinction between Lamarckian adaptationism and Darwinian selectionism became increasingly blurred in the second half of the twentieth century as biologists focused more attention on parts of organisms and less on organisms as wholes. This reductionist movement, driven first by the successes of population genetics and later by the development of molecular methods, may have been an unconscious response to developmental biologists' discoveries of the complexities underlying the transition from DNA sequences to complete organisms—complexities that threatened to swamp simplistic theories based on the mantra of one gene—one trait—one selection vector. Whatever the reasons, losing the perspective on whole organisms led to a loss of Darwin's panoramic view of biological diversity. In the last quarter of the twentieth century there have been many efforts to reemphasize the nature of the organism in evolutionary biology (Brooks and Wiley, 1988; Brooks et al., 1989; Brooks and McLennan, 1990; Depew and Weber, 1995; Kamin, 1991; Kaufman, 1993; Lewontin, 1982, 1983; Maynard Smith and Szathmáry, 1995; Odling-Smee et al., 1996; Raff, 1996; Salt, 1993; Van de Vlijver et al., 1998). These perspectives reassert that orderliness and organization in biological systems result from the interaction of selection processes with the inherent nature of the organism.

B. Organisms as Energy Flow Systems

Lotka (1913, 1925) was among the first twentieth-century authors to discuss biological systems, including organisms, in terms of energy flows and energy partitioning. He recognized that biological systems persist in space and time by transforming energy from one state to another in ways that generate and maintain organized structure. Maurer and Brooks (1991) recognized two classes of such energy transformations. Heat-generating transformations involve a net loss of energy from the system, usually in the form of heat. Conservative transformations involve changing free energy into stored states (e.g., structure). Because all conservative transformations in biological systems are coupled with heat-generating transformations, there is a heavy energetic cost to maintaining structure. Lotka (1913) suggested that the interplay between flow and partitioning of energy in biological systems acts to slow the rate at which energy stored by conservative transformations is degraded by heat-generating transformations.

The development of nonequilibrium thermodynamics in the past 50 years (Prigogine and Stéphane, 1946; Prigogine, 1980) has allowed us to generalize Lotka's view of the nature of the organism. Living systems are nonequilibrium thermodynamic systems; they exchange matter and energy irreversibly with their surroundings and they maintain themselves in far from equilibrium conditions. The basic features of nonequilibrium systems can be summarized heuristically as

\[
d\dot{S} = dS + dS, \quad dS > 0
\]

Total entropy changes \((dS)\) can be subdivided into two components: \(dS\), which measures exchanges between the system and its surroundings (changes in the surroundings), and \(dS\), which measures production by irreversible processes internal to the system (changes within the system). We can also call this heuristic equation the "cost of living" for organisms because all organi-
isms must take in high-grade energy and matter and dissipate lower grade energy to their surroundings in order to survive. Energy degraded in the uptake of raw materials from the surroundings into the system is dissipated into the surroundings \( dS \). These exchanges are accompanied by a great deal of waste; hence, \( dS \) is very large compared with \( dS \). Nonetheless, it is in the manifestations of internal production that we perceive organismal diversity.

Internal production \( dS \) includes (i) dissipation from the system, called the external dissipation function \( \gamma_e \), or heat-generating transformations). Heat-generating processes occur when energy and entropy flow in opposite directions, with entropy production tending to move the system toward disordered states; and (ii) dissipation within the system, called the bound dissipation function \( \gamma_b \), or conservative transformations. Conservative transformations are characterized by energy and entropy flowing in the same direction, with entropy production being retained within the system and tending to move the system towards more structured states. In biological systems, \( \gamma_b \) can be further subdivided into allocations for accumulating biomass \( \gamma_b^2 \) and allocations for accumulating genealogical information \( \gamma_b^3 \). Heuristically (Brooks and Wiley, 1988),

\[
dS = \gamma_e + \gamma_b^2 + \gamma_b^3
\]

Organisms thus have a dualistic nature. As open thermodynamic systems, they must simultaneously interact with their surroundings and perform critical functions internally. They maintain themselves in a viable state by exchanging matter and energy irreversibly with their surroundings, taking in relatively high-grade energy and using it to perform useful work within themselves. This requires sensing of, and causal engagement with, the surroundings mediated by a physical distinction ("phase separation", Prigogine, 1980) between the organism and its surroundings. That is, there must be an "inside" and an "outside" of the organism, delineated by a physical boundary. For all organisms, this boundary is provided by cell membranes, which are simultaneously physical barriers between the inside and outside of the organism and highly selective mechanisms for modulating the exchange of matter and energy between the organism and its surroundings. For multicellular organisms, this barrier is a complex of cell membranes.

C. Organisms as Information Systems

Elucidation of the genetic code made it possible to begin thinking of organisms as information systems. Information systems consist of a source of signals, a channel through which the signals are transmitted, and a receiver to translate the portion of the signals that made it through the channel into information. Gauvin (1972) argued that the genetic system is the source, reproduction and ontogeny are the channel, and the environment is the receiver. Genetic possibilities thus become phenotypic signals as a result of reproduction and ontogeny and become meaningful biological information as a result of causal interactions between the phenotype and the environment. Brooks and McLennan (1990) noted, however, that the environment is not a receiver in a physical sense because its only causal interaction with biological information is the possible elimination of some of it; it does not measure or interpret the information. Rather, the environment acts as interference in the channel, or as a "delete button" monitoring "incoming messages," eliminating relatively less fit organisms.

If biological information is a material part of biological systems, however, it is possible for biological systems to be their own sources and receivers. Current standard evolutionary theory stresses external causality for orderliness in biological systems, so the receiver has been construed as part of the surroundings, i.e., localized in space. Although it is true that biological systems are localized in space, they are also localized in time. In other words, the receiver can be a "time" (Brooks and McLennan, 1990). The source is a genetic system at time \( t_0 \), the channel is reproduction and ontogeny, and the receiver is the same genetic system at any given time \( t_1 \); thus, the receiver is temporally distinct from the source. If an information source precedes its receiver in time, it can produce the system that acts as receiver, and that system can then become a source. This perspective has been used by information theorists designing self-correcting computer programs, which can enhance their own abilities to store and transmit information efficiently. The same holds true for biological systems: DNA has significant self-repair capabilities.

Biological systems are physical information systems, a type of nonequilibrium thermodynamic system, open to exchanges of matter and energy but maintaining a closed information system internally which functions to reproduce the system—to perpetuate lineages through time. They are able to impose themselves and their functions on their surroundings and thus are self-stabilizing and self-organizing. They produce organized complexity cheaply \( dS \) is small compared to \( dS \), and the portion of \( dS \) allocated for the information system is small, in part because a small number of chemical templates are used to generate many organisms), variability (because even chemical templates are subject to
the statistical mechanical vagaries of the second law of thermodynamics), and functionally (because organisms must exchange matter and energy with their surroundings in order to maintain themselves), but without regard for details of the surroundings (because the information system is embodied in relatively autonomous internal chemical production, $dS$, of the system). As the source and receiver of organized information, they can be the embodiment of the organizing principles for that information. Biological systems thus transmit information through, not to, their surroundings. This supports Darwin's view that it is the (autonomous) nature of the organism that creates the necessary conditions for selection processes, especially natural (environmental) and sexual selection, to occur.

Treating biological systems as physical information systems provides a causal basis for the origin of selection processes consistent with their well-documented causal consequences. Selection processes originate as a result of the necessity that biological systems obtain matter and energy from their surroundings coupled with the relative autonomy of their information systems, which permits production of organisms regardless of the details of their surroundings. Without the constraints provided by this autonomy, there would be no selection; simultaneously, however, constraints provide systems with macroscopic properties that limit the ways in which and the extent to which the system will respond to selection. This means that biological systems should have their own macroscopic, or macroevolutionary, properties that are not reducible to microscopic dynamics. Some of the most important implications of reemphasizing the nature of the organism in the context of thinking of evolution as an informational process involve the manner in which evolutionary theory views major events in the history of life (Maynard Smith and Szathmary, 1995).

D. Putting It All Together

Biological systems, beginning with organisms, are functional wholes with respect to the way they engage their surroundings and with respect to their internal organization. A major component of internal organization is functional integration through the interdependence of parts, and this is most evident in the dynamics of ontogeny (Salthe, 1993; Raff, 1996) and in physiological processes. Evolutionary changes in biological systems do not occur at all at the same time; thus, when such changes occur, only part of the system changes. All changes, whether point mutations on a chromosome, genomic doubling, or alteration of part of a complex mating ritual, must integrate with the rest of the system, which has not changed, in such a way that viable organisms result. The functional necessity of developmental integration creates stability domains within bioinformational phase space (Brooks and Wilke, 1988; Kauffman, 1993; Niklas, 1999). Orderliness and organization in biological systems result from the interaction of selection processes with three aspects of the nature of the organism: (i) historical uniqueness, (ii) cohesive properties, and (iii) hierarchical organization.

Historical uniqueness manifests itself in historical contingency but, more importantly, in temporal irreversibility. Important organismal processes such as reproduction, development, aging, and death are inherently irreversible. Spontaneous irreversible behavior of this sort always involves growth and increasing complexity, and physical manifestations of at least some of the systems' history (Schnick and Wimsatt, 1988).

Cohesive properties range from membrane-bound nucleic acids to cell-cell adhesion and recognition, sexual reproduction and specific mate recognition systems, and common history of inheritance. Cohesion is especially important to evolutionary explanations; it is the "glue" of functional integration and hierarchical organization that are so characteristic of biological systems (Wake and Roth, 1989). Many biological processes that demonstrate irreversible behavior manifest such changes as a result of interactions among cohesive factors.

Hierarchical organization plays an important role in organismal diversity. Hierarchies provide stability, reinforce boundaries between organisms and their surroundings, allow increasing amounts of complexity without losing organizational coherence, and provide a way in which causation and control can be tied together (Salthe, 1993). Eldredge and Salt (1984; see also Eldredge, 1985, 1986; Salt, 1993) emphasized two major forms of hierarchically organized biological structure. The ecological hierarchy (interactors of Hull, 1988) is manifested by patterns of energy flow in ecosystems, and the genealogical hierarchy (replicators of Hull, 1988) is manifested by patterns of ancestral relationships among organisms and species. Environmental and genealogical phenomena are good starting points for investigating hierarchical interactions because they are intimately connected in biology. Prebiotic environmental conditions established the boundary conditions within which life could originate. Conversely, genealogical processes that characterize life are autonomous enough from environmental conditions to be capable of overrunning available resources and of changing the environmental conditions substantially. The longer life
exists on this planet, the more it shapes the environment of the planet. Today, much of the environment relevant to biological systems consists of other organisms (Maynard Smith, 1976; Brooks and Wiley, 1988). Organisms imposing themselves on their environments (creating niches) contributes greatly to the far from equilibrium nature of the environment in which evolution occurs (Odling-Schmee et al., 1996).

The intimate relationship between the two hierarchies can be illustrated with a sports metaphor: The ecological hierarchy establishes the dimensions of the playing field, whereas the genealogical hierarchy establishes the rules of the game being played. In other words, biological systems obey rules of self-organization transmitted genealogically (historically) and played out within environmentally defined boundaries. To complicate matters, however, the self-organizing rules of the game by which living systems evolve can produce changes in the dimensions of the playing field. To extend the metaphor, the game may redefine the dimensions of the playing field and may be subsequently constrained by those self-imposed changes. For example, the evolution of photosynthetic prokaryotes from anaerobic ancestors resulted in increased oxygen content in the atmosphere. This increase then altered the diversity and changed the distribution of anaerobic organisms, limiting them to relatively rare environments and paving the way for the evolution of a new array of species. Now that we have a consistent view on the nature of the organism in a physical sense that also accounts for their Darwinian nature, we can discuss observed organismal diversity.

II. DIVERSITY IN THE REAL WORLD:
A PRECIS OF CURRENT ORGANISMAL DIVERSITY

Organismal diversity is evolved diversity. This means that no matter how distantly related two organisms are, they still share at least some features in common, such as cell membranes made of lipid-bilayers and a common genetic code, linking them to a single phylogeny of life on this planet. Highlights of organismal diversity can thus be best recounted in a precis of organismal diversification—phylogeny. However, phylogeny is not a linear historical timeline. It is a hierarchically branching pattern of diversification and spread, producing organisms that are historically unique mosaics of ancient and recent traits, both functional and structural. Therefore, it is not possible to give even a precis of organismal diversity without twists and turns. The perspective on the major transitions in evolution presented by Maynard Smith and Szathmary (1995), however, provides some consistent reference points. Maynard Smith and Szathmary suggested that the major transitions in evolution occurred whenever life functions were compartmentalized and modularized in such a way that the reproductive flow of biological information was enhanced. This can occur either by enhancing the survivorship of organisms in particular environments, thus increasing the likelihood that they will reproduce, or by enhancing the mechanisms by which information is stored and transmitted, regardless of the particular environment in which it evolves. Maynard Smith and Szathmary highlighted what they considered to be the nine most important evolutionary transitions, all of which were examples of enhanced efficiency of information storage and transmission.

One observation that is relevant is that the smallest and the largest organisms known to have existed in the past exist today. Members of Thermoplasma are thermophilic and acidophilic prokaryotes living in coal deposits that possess only 1100 kilobase pairs of nucleotides in their genomes, whereas blue whales are oceanic monsters outstripping even the largest of the dinosaurs.

The oldest lineages of organisms on the planet are various prokaryotes, including bacteria and archaeans. These organisms are single celled, lack organelles or membrane-bound genomes, and exhibit only asexual or simple parasexual reproduction. Their cells form only rods, spherites, or spirals, and yet they have inhabited this planet continuously for almost 3 billion years and represent by far the largest number of differentiated lineages on the planet today, including the largest number of unnamed taxa. Representatives of the prokaryotes include chemoheterotrophs (thought to have been the lifestyle adopted by the first living things), chemoheterotrophs, phototrophic, and photoautotrophs. Photoautotrophic prokaryotes, primarily Cyanobacteria, were the first to demonstrate that organisms could change the environment of the planet, substantially affecting standing biodiversity. They did this by releasing large amounts of molecular oxygen into a reducing atmosphere, changing it in such a way that most prokaryotes became extinct or restricted to marginal habitats. This release of oxygen, as well as associated nitrogen fixation, by prokaryotic photoautotrophs created the conditions necessary for other life forms to evolve, so it is safe to say that prokaryote evolution played a major role in setting the stage within which all subsequent biological diversification occurred. Prokaryotes today include the so-called Archaea, or extremophiles,
which may represent a collection of distantly related remnants of the diversity affected so much by the evolution of those photoautotrophs. Among the extremophiles are members of the halophilic Halobacterium, which grow best at pH 11.5, and the acidophilic Sulfolobus, growing best at pH 2 or 3 but able to survive at pH 0.9. There are methanogens, producing most of the planet's methane, including a host of symbionts living in the rumens of cattle and their relatives in which they produce digestible carbohydrates for their hosts as well as methane. Another methanogen, Methanopyrus, which lives near ocean-bottom thermal vents, grows optimally at 98°C but is able to survive at 110°C and "dies of cold" at 84°C. Sulfolobus is also a hyperthermophile, living optimally at 70–75°C and dying at temperatures lower than 55°C. The extreme thermophile Thermus aquaticus, isolated from hot springs in Yellowstone National Park, was the original source of high-temperature DNA polymerase, the enzyme used to catalyze the polymerase chain reaction (PCR) technology that has become indispensable to molecular biology. Most other prokaryotes are bacteria, which include some of mankind's most valued and most loathed organisms. Bacteria living in our digestive tracts are essential for proper nutrition, and Escherichia coli is the workhorse of molecular biology and biotechnology, including the pharmaceutical industry. On the other hand, diseases such as syphilis, anthrax, diphtheria, tuberculosis, smallpox, tetanus, cholera, and plague and a variety of agents of food poisoning, including Salmonella and botulism, are all caused by bacteria. These same disease-causing prokaryotes have also shown a tremendous degree of adaptability in the face of substantial human efforts to eradicate them.

From within this great prokaryotic diversity evolved the first eukaryotes. Most species of eukaryotes, including the oldest surviving lineages, comprise single-celled organisms differing from prokaryotes by well-organized compartmentalization of functions within the cell. Cell energetics are handled by organelles called mitochondria and chloroplasts, which are thought to have originated as endosymbiotic prokaryotes. Digestion and excretion are handled by specialized vesicles within the cell that facilitate the movement of matter and energy in and out of the cell. Synthesis of amino acid building blocks for structural components takes place in specialized organelles called ribosomes. Finally, the storage, production, and transmission of information are compartmentalized inside a membrane-bound nucleus. Powered locomotion, first evolved by prokaryotes in a rudimentary manner, becomes highly effective in unicellular eukaryotes as a result of sophisticated cytoskeletal integration.

Most of the heterotrophic unicellular eukaryotes are grouped together in a nonmonophyletic collection of lineages called the Protista. Within this evolutionarily heterogeneous and diverse collection of lineages are the origins of the first sophisticated sexuality as well as alternation of sexual and asexual generations (diverse examples are found in ciliates, apicomplexans such as malaria, and various algae), the first sophisticated modifications of the cell membrane in ways that modulate organism–environment interactions (diatoms produce complex and beautifully sculpted exoskeletons), and the beginnings of true multi-cellularity, manifested by colonial aggregations of unicells that exhibit some degree of functional differentiation, such as the fruiting bodies of slime molds.

The protists are major underpinnings of the heterotrophic portions of the biosphere. Within the heterotrophic realm, protists represent saprophages and detritivores, predators, and a variety of symbionts, ranging from commensal ciliates living in the tests of sea urchins and the rectums of frogs to mutualist flagellates living in termite guts, providing digestible carbohydrates for the insect hosts in exchange for their own food and a safe haven, and parasites of global socioeconomic significance, including the causal agents of amebic dysentery, malaria, giardiasis, African sleeping sickness, Leishmaniasis, and Chaga's disease. Most show enhanced reproductive output and complex life cycles, but Giardia and Entamoeba show some secondary losses of structures.

The photoautotrophic protists, generally called Algae, comprise four distinct lineages: dinoflagellates, red algae, diatoms, and the green algae. The dinoflagellates are best known as the causal agents of red tide. Massive blooms of certain dinoflagellates produce large amounts of waste products that act as nerve toxins capable of killing large numbers of fish or of concentrating in shellfish and killing those who eat them. Red algae also include some basal species that are parasitic in more derived members of the same clades. Diatoms are ubiquitous phytoplankton that represent the primary source of photosynthesis (and oxygen production) in the ocean, although many are freshwater inhabitants. They are characterized by having bipartite siliceous shells (tests), exhibiting species-specific forms and beautiful surface sculpting. Because they are siliceous, these tests are highly resistant to decomposition. As a result, fossil deposits of diatoms are excellent indicators of past environmental conditions. Large deposits produce "diatomaceous earth," which has many industrial uses including insulation, filtration, and metal polishing. Diatomaceous earth has also been proposed as a "non-polluting" insecticide. Spreading diatomaceous earth on
fields may well clog the tracheae of pest insects but will also clog the tracheae of beneficial insects and may also affect the breathing structures of other animals, including those humans who apply it or work in fields on which it has been applied.

True multicellular lineages appear to have evolved four times within the protists: the brown algae and plants, which are photoautotrophs, and the fungi and animals, which are heterotrophs. According to Maynard Smith and Szathmary, multicellularity represents an advance over unicellularity because in multicellular organisms the partitioning of functions can occur at the level of entire cells, tissues, organs, or organ systems rather than having to take place within each cell. Complex reproductive systems can evolve because multicellular organisms can produce gametes without having to become gametes. Given the large numbers of unicellular eukaryote lineages and the small number of multicellular eukaryote lineages, we conclude that this is difficult to achieve, despite the evolutionary payoffs. Perhaps this is because it is difficult to coordinate a population of cells. There are no multicellular prokaryotes (there are bacteria that form strings, strands, and mats, but these are simply aggregations of individual organisms), perhaps because prokaryotes show no functional partitioning at the cellular level. If so, true multicellularity could only have arisen after unicellular eukaryotes evolved. In a complementary vein, there is only a single known example of secondary loss of the multicellular state; this occurred in the Microsporida, a group of obligate parasites of fishes and annelids which were thought to be an apicomplexan protist group until it was shown that they are actually highly modified cnidarians.

Each of the four cases of true multicellularity represents a very different evolutionary path. The least successful of these are the brown algae, sister group of the diatoms. Up to 60 m long in the case of Macrocytis, brown algae can be found on ocean shores with pounding surf, in relatively quiet near-shore meadows, or floating in the open ocean. The Sargasso Sea, for example, is an enormous mass of floating Sargassum. Although relatively species poor, brown algae can be abundant. They exhibit differentiated tissues and even organs, including holdfasts and filaments, which are analogs of roots and leaves that provide homes to a diverse array of animal life. Some have gas-filled bladders which may contain up to 5% carbon monoxide—enough to kill a human. Cellular cohesion in brown algae is provided by alginate acid in the cell walls. This is a gummy polymer of sugar acids that cement cells and filaments together and provides the glue for the holdfasts. Alginate acid is also an important commercial emulsifier in ice cream, cosmetics, and other products. Larger pieces of brown algae, especially some kelp species, are considered delicious and healthy food sources for humans.

The remaining three groups of multicellular organisms are relatively closely related to each other, albeit they diverged a very long time ago, and to two groups of protists—the photoautotrophic green algae and the heterotrophic choanozoa.

The second group of multicellular photoautotrophs are the plants, sister group of the green algae, or at least of some green algae. Unlike the brown algae, which originated in the sea and have remained there throughout their evolutionary history, plants are the descendants of an aquatic lineage that invaded and diversified on land. Two closely related living relatives of the true plants are Volvox and Micrasterias (sea lettuce), which form aggregates of single-celled organisms showing some degree of functional differentiation, including the production of large stationary eggs, which are highly differentiated from the very small, flagellated, and highly mobile male gametes, a condition also found in true plants. Cellular cohesion in true plants is provided by plasmodesmata, which are cytoplasmic strands connecting adjacent cells. Many plants exhibit a high degree of adaptability or permissiveness with respect to genome-level evolutionary modifications. Polyploidization is common among many groups, including both autopolyploidization and allopolyploidization as a result of interspecific hybridization.

The land represented an excellent substrate for photoautotrophs because there is more consistent sunlight on land than in the sea, where sunlight of sufficient quality for photosynthesis to occur does not penetrate deeper than approximately 50 m. However, being on land also posed problems, one of which was structural: Air provides little support compared to water. In the absence of inherent structural support, plants could only sprawl on the land—they could never rise above it. Cell walls made rigid by cellulose, a carbohydrate, provided the necessary structural support. Subsequent evolution of vascular tissue in a moss-like ancestor and then a sophisticated tubular transport system (tracheids) in the ancestor of the sister group of the mosses (those plants called, not surprisingly, Tracheophytes) enabled water and nutrients to be transported against gravity to considerable heights. Rigid cell walls and the tracheid system provided opportunities for tremendous vertical growth, represented by the convergent evolution of tree forms in many groups (including giants up to 110 m tall) and a tremendous flow rate of water and nutrients. A single maple tree only 15 m tall may support more than 175,000 leaves with a total surface area of more than 675 m², transporting up to 220 liters of
water per hour from the soil throughout the tree and eventually into the atmosphere through evaporation from the leaves.

Terrestrial life that is truly independent of aquatic environments must also be associated with suitable reproductive modes. When ancestral plants moved onto land, they brought with them the large stationary eggs characteristic of their closest relatives. In all land plants, when these eggs are fertilized, the resulting embryos are sporophytes contained within a protective covering produced by the parent. Those ancestral plants also brought with them aquatic flagellated sperm, maintaining a dependency on aquatic surroundings, at least for reproduction. Only when nonaquatic male gametes evolved in early angiosperms did land plants become completely terrestrial.

Once firmly established on land, major events in plant diversification were associated with increasing compartmentalization and modularization of the organism, all integrated and coordinated chemically and without a centralized nervous system. They elaborated functionally differentiated nonreproductive body parts, especially roots and leaves, that became specialized habitat for other terrestrial dwellers, creating many novel opportunities for diversification by these others and coevolutionary feedback that affected their own evolution as well. This included the evolution of parasitic plants such as Indian pipe, which has lost its chloroplasts and depends on the host plant for photosynthesis. Their root systems helped plants modify the soil in ways that enhanced the lives of the plants themselves but also affected the lives of plants and other types of organisms living in the soil. Some plants today produce toxins in their roots that diffuse in the soil, inhibiting the growth of other plants. Others, including common garden plants such as marigolds, secrete substances from their roots that deter plant-consuming soil nematodes and thus can be planted as a nonpolluting nematicide border in gardens. Aboveground vertical growth in tracheophytes led to a proliferation of types of habitats—plants, shrubs, bushes, vines, and trees. This vertical structuring allowed complex communities of plants to form, within which the various plants created their own ecological interactions with other plants by shading each other. Plants also differ in the relative amount of biomass allocated to below- and aboveground structure. This led to complex ecological interactions between root systems of different plants. In the tropics, for example, most plant biomass is aboveground, whereas in temperate areas in which there is more humus in the soil, a much larger proportion is belowground. In fact, a given plant may participate simultaneously in one set of aboveground and another set of belowground ecological interactions, just with other plants, depending on what species the plant belongs to and where it happens to be growing.

Plant organismal diversity is also associated with major modifications of the reproductive system, each of which has many variations. Primary among these was the evolution of seeds, which are embryos protected by a covering of nutrient tissue provided by the parent. These multicellular protective coats, containing cellular contributions from as many as three generations, are a source of protection and nutrition for the embryos which may allow them to withstand substantial periods of environmental harshness and still be capable of germinating and producing a plant. The earliest seed plants produced seeds exposed directly to the environment, called gymnosperms (naked seeds). Angiosperms (hidden seeds) evolved when the ovaries, containing the seeds, expanded into fruits, hiding the seeds and providing more protection. The evolution of double fertilization at more or less the same point in time, producing triploid endosperm, provided even more nutrition for the embryo. Angiosperm diversification produced some amazing organisms. The seeds of some plants need to be burned by forest fires in order to germinate, and others must pass through the intestinal tracts of herbivorous vertebrates before they germinate. Plant seeds today exhibit an immense array of structural modifications that enhance their dispersal from the parental plant. Those that pass through the intestinal tracts of vertebrate herbivores may be dispersed for many kilometers from their parental plants by the herbivore. Other seed modifications take advantage of wind—from the fluffy dandelions to the robust winged samaras of maples. Another major reproductive innovation was the evolution of flowers, which are specialized and self-contained reproductive modules. Flowers represented an enormous potential for reproductive cohesion and diversification.

The evolution of plants was followed closely by the evolution of herbivores, especially among the nematodes and insects. This in turn was followed by the evolution of a variety of antiherbivore adaptations. Defenses against herbivory include both diminution and elaboration of the part of the plant being eaten, the evolution of surface features such as thorns, and storage of toxic metabolic by-products in tissues that are targets of herbivory. Likewise, the diversification of flowering parts is also associated strongly with the diversification of particular groups of animal pollinators, again most notably insects. In fact, it appears that the evolutionary diversification of flowering plants and of phytophagous
insects has been facilitated by their close ecological associations.

Plants exist in many seemingly inhospitable habitats, particularly desert and boreal regions, where the evolution of resistant seeds as well as specialized reproductive, root, and stem and leaf systems allow them to survive and thrive. Cacti and other succulents represent convergent evolution of water-retaining and evaporation-resistant body forms that can survive in desert conditions. Many plants are able to flourish in nearly marine salinity conditions, although none is truly marine. Mangroves, for example, are able to excrete excess salt from the water they absorb in the tropical coastal estuaries in which they live. Plants also thrive in seasonally fluctuating environments. This is helped by the evolution of resistant seeds but is also enhanced by the ability to synchronize growth and reproduction with permissive times of the year. For example, the ability to shed leaves deciduously apparently arose as an adaptive response to seasonally dry climates but today functions in seasonally cold climates as well.

Tapping an almost unlimited energy source (sunlight), coupled with their sophisticated means of compartmentalization, has made plants extremely important to humans and to the biosphere. Terrestrial plants improve air quality by fixing carbon. Some also help improve the ability of soils to sustain life by fixing nitrogen mediated through mutualistic relationships with root-dwelling bacteria; the best known of these mutualists produce the root nodules called peanuts. Plants that have adopted life in freshwater habitats are often quite sensitive to changes in water quality and are thus useful as indicators of pollution. Those having extensive aquatic root systems may also be excellent biofiltration systems; water hyacinths, for example, can actually improve water quality. The cultivation of plants is thought to have been a major event in human evolution. Ironically, the cultivated production of large quantities of edible plant biomass is thought to have provided the conditions for the growth of human population and technology that now threaten our existence by threatening the biosphere. Different plants allocate substantial amounts of nutrients into root systems, especially tuberous ones, into stems and leaves, and into reproductive structures such as nuts, seeds, and fruits. Virtually any part of a plant may be highly edible for humans or their domestic animals. Simultaneously, substances functioning to deter herbivory or competition from other plants may also serve as sources of useful materials for humans, ranging from spices to pharmaceuticals. The American yew tree sequesters a compound called taxol, which has significant effects on certain forms of cancer.

a close relative, the European yew, produces a similar compound called taxotere, which also has anticancer properties.

Fungi represent one of the two major groups of multicellular heterotrophs. Among living groups, fungi are apparently the sister group of choanoflagellates and animals. Cellular cohesion in fungi is provided mostly by default: The vegetative body of a fungus, called a mycelium, comprises many hyphae, which are filamentous multicellular extensions maintaining cohesion among the cells by forming only incomplete cellular divisions, or septa, between them.

To many people, the presence of fungi indicates that something has gone wrong—one’s basement is too damp or one’s bread is too old. To others, fungi are something quite different: They are the major recyclers and decomposers of the biosphere, found most commonly in association with the refuse of the biosphere, including that of our own civilization. This is because most fungi are very adaptable saprobes, highly tolerant of great ranges in temperature and osmotic, especially hyperosmotic, conditions. A considerable number of fungi are parasitic and a few are even predators, attacking small animals such as soil nematodes. Fungi are very flexible reproductively, reproducing sexually or sexually with mating types not distinguishable as male or female and exhibiting an enormous diversity of variations on the general eukaryote theme of alternation of generations. Fungi are the primary source of cellulose and lignin breakdown in the environment. These features make them excellent agents for bioremediation. They can become established in highly degraded habitats and actually improve them, creating conditions which can support yet more life. Tropical forest soils, for example, lack much humus but have enormous mats of fungal hyphae forming mycorrhizal mutualisms on and in plant root systems, which maintain vital nutrients and structural cohesion in the small amount of soil that is present. Fossil evidence suggests that mycorrhizal associations are at least 300 million years old.

There are four major groups of fungi. The Chytridomycota, or water molds, are thought to be the sister group of the rest of the fungi. Chytrids have been implicated in recent massive die-offs of amphibians living above 1000 m elevation throughout the world. Many chytrids are known to be parasitic, with some exhibiting complex life cycles involving two different hosts. A few have been found that are predators of nematodes. Members of Arthrobotrys, Dactilaria, and Dactyella produce three-celled loops that entrap nematodes as they wind their way through the soil; once trapped, the
nematode cuticle is invaded by hyphae and ingested. The fungi most commonly known as "molds" are members of the Zygomycota. These fungi are also arguably the basis of the modern pharmaceutical industry—the mold that tells you your bread is too old is likely Penicillium, the original source of the world's first broad-spectrum antibiotic, penicillin. Two other members of the genus, P. camemberti and P. roqueforti, are required to produce fine cheeses. Some members of Aspergillus are essential to the production of soy sauce and sake, whereas others can cover edible products such as peanuts with carcinogenic aflatoxins. Probably the best known and respected fungi are members of the Basidiomycota. This group includes the edible fungi, called mushrooms, morels, or truffles, in ascending price. This group also includes rusts and smuts, which can cause serious diseases in agricultural plants. Basidiomycetes are also a major source of legal and illegal pharmaceuticals and some of history's most infamous deadly toxins. The sister group of the basidiomycetes are the members of the Ascomycota, which comprise the yeasts and sac fungi. Without members of this group, we would have no bread, yogurt, or many alcoholic beverages, including wine and beer.

On their own, fungi are ubiquitous and numerous throughout the world. In addition, some fungi form mutualistic associations with algae, called lichens, which are perhaps the champion extremophiles of the multicellular world. Extremes of heat and cold and dry and wet do not seem to deter lichens, who often flaunt their hardiness by living on rocks, slowly but inexorably digesting them into dust.

The second group of multicellular heterotrophic eukaryotes are the Metazoa, or true animals. Metazoa are the sister group of the choanoflagellates. This group of multicellular organisms maintains cellular cohesion by means of extracellular matrices of cell-adhesion proteins. Unlike plants, which invaded the land early and diversified there, most animal diversification occurred in the ocean, and various groups have invaded freshwater and terrestrial habitats convergently. Even today, most metazoan organismal diversity is in the sea. A great deal is known about many metazoans, probably because many of them are large enough to be seen with the naked eye and active enough to engage visually oriented organisms such as ourselves. Most metazoans, however, are microscopic. Highly complex and sophisticated adult organisms, such as some rotifers, may be as small as 50 μm.

The array of metazoan body plans provides clear evidence of modularization and compartmentalization. Metazoan bodies range from aggregates of specialized cells in sponges and placozoans to defined body forms composed of specialized tissues in cnidarians and ctenophores, with organ-level organization beginning with platyhelminthes, and complex organ systems in the majority of the other metazoans. Body cavities emerged very early in metazoan evolution as effective ways of physically compartmentalizing organ-level modules and laid the evolutionary groundwork for specialized compartments within body cavities, called segments, metamers, and tagmata. Metazoan bodies may be asymmetrical as in sponges, radially symmetrical as in cnidarians, bilaterally symmetrical as in the majority of metazoans, or even biradially symmetrical as in ctenophores and echinoderms. Developmental patterns among metazoans tend to be highly conserved within major groups, although secondary loss of structures is not uncommon: the platyhelminthes appear to be descended from an ancestor that lost its body cavity, and a large group of parasitic platyhelminthes, dominated by the true tapeworms, has secondarily lost its digestive tract as well.

Many metazoans are conspicuous because they tend to move around a lot. Modes of locomotion are varied, but most are variations on swimming, creeping, or crawling. Sophisticated locomotion in metazoans is highly correlated with enhanced cephalization and neural integration. Cephalopod mollusks have water-jet propulsion provided by specialized siphons. Complex appendages for swimming, walking, and running evolved in the arthropods and the chordates. It is within these two groups that organisms capable of true powered flight emerged.

All metazoans are heterotrophs, but they exhibit the entire range of heterotrophic functions—they are saprobes, detritivores, filter and suspension feeders, predators, herbivores, commensals, mutualists, parasites, and parasites. Herbivory is uncommon among marine metazoans but is a major attribute of the structure of terrestrial ecosystems. Modern herbivory diversity is concentrated in the insects, the annelid vertebrates, and the nematodes. Interestingly, no metazoans ever evolved the ability to digest cellulose. Thus, many metazoan herbivores dine on plant cells and tissues that contain nutrients in a form other than cellulose. Those that actually gain nutritional benefit from cellulose are hosts to a variety of prokaryotic and eukaryotic microbes which possess the enzymes necessary to break cellulose down into less complex carbohydrates, which can be absorbed by the metazoan host. Although this form of herbivory seems to be very successful when it evolves, it does not appear to have evolved often, presumably because it is difficult to acquire the proper
symbiotic microbes, provide a suitable home for them in the intestine, and provide a means of transmitting them to offspring.

Herbivory has a downside, however. As indicated previously, many plants allocate toxic compounds to the structures eaten by herbivores. The herbivores that have survived such evolutionary changes in their food have evolved ways of avoiding, sequestering, or detoxifying those substances. Metazoa not only graze on plants but also many graze on other metazoa. Parasitism has evolved in virtually every group of metazoa except the chordates. The phylum Acanthocephala is entirely parasitic, whereas the vast majority of phyla are parasites, including the well-known digeneans (trematodes or flukes) and eucestodes (tapeworms). Many nematodes and arthropods are also parasitic.

Reproductive modes are diversified in metazoa. Most metazoan organisms are either males or females, although hermaphroditism is not uncommon in members of some of the oldest lineages. Gonads may be transient specialized cells or tissues or permanent organs and organ systems. Reflecting their ancient origins and diversification in the ocean, fertilization is pleisomorphic in most groups. Internal fertilization, in which the female provides the aqueous medium within which the sperm can swim to the ova, has evolved multiple times. The ancient eukaryotic theme of alternation of generations is carried forward into metazoan evolution, although it drops out in several lineages convergently. In many cases, the asexual and sexual portions of the life cycle are ecologically partitioned; in most cnidarians these stages are characterized by distinct medusa and polyp forms which have distinct ecologies. However, even for metazoa that do not exhibit alternation of generations, it is common for the larvae and adults to have very different life-styles. Most larval frogs, for example, are herbivores, whereas virtually all adults are insectivores. Among ecdysozoans (nematodes, rotifers, tardigrades, kynorhynchs, gastrotrichs, acanthocephalans, and arthropods), haplodiploidy is a common theme, setting the stage for extreme sexual size dimorphisms and asymmetrical sex ratios biased heavily toward females. Many metazoa are parthenogenetic, forming unisexual female lineages. In some cases, these lineages are entirely self-sustaining, whereas in others, such as the gynogenetic poecilid fish called the Amazon molly, sperm from males of a parental species are required to trigger embryogenesis, even though no genetic information from the sperm is incorporated into the embryo.

Another major theme of metazoan organismal diver-
sto is temporal synchronization of reproductive cycles. Perhaps the most extreme case is that of the tropical Pacific polychaete called the Palolo worm, in which the entire species reproduces in a single night each year. A more familiar example is that of anadromous salmonid fishes, in which adults return to their natal streams from the ocean once a year, reproduce, and die. In herbivorous insects reproductive synchronization is tied to cycles of growth and reproduction in the different plants they eat.

One of the most arresting features of animals is complex behavior, especially with respect to reproduction. This can be manifested in sophisticated interactions among males, among females, between males and females, in parental care, or even in complex combinations of all possibilities. Complex mating rituals have been documented in tardigrades, in which males are known to apply ritualized Stokes to females to encourage them to reproduce. Some leeches are known to carry their offspring on their abdomens for a period of time after birth. Mollusks (especially cephalopods), arthropods (especially insects), and chordates (especially vertebrates) are the best examples of sophisticated behaviors among metazoa. All three groups have evolved complex control and sensory integration involving neurochemical systems and strong cephalization.

One outcome of integrated and sophisticated behaviors is sociality. Maynard Smith and Szathmary (1995) designated the evolution of social systems as one of the nine major transitions in biological evolution. Social systems permit functional compartmentalization at the level of different organisms within a social group. This level of modularization and compartmentalization is best seen in the social insects, notably bees, ants, and termites, whose colonies function very much like superorganisms. The fundamental feature of these societies is genetically stereotyped individual organisms, most of whom are sterile or sexually inactivated female workers and soldiers (anals of somatic tissues) along with a few reproductively active males and females (the anals of reproductive organs). In bees and their relatives, colonies consist of many sexually inactivated females called workers and a smaller number of reproductively active males called drones, whose only function is to reproduce with a single queen, the reproductively active female. The other major group of social hymenopterans, ants, differs from bees and their relatives by having males that die shortly after reproducing with a queen and by exhibiting a tremendous range of morphologically distinct, functionally stereotyped workers and soldiers. Ant societies fascinated Darwin, especially their
propensities to wage war, make slaves, use tools, construct complex domiciles, and undertake agriculture, farming fungi and milking aphids for nutrients. The other major group of social insects are termites, whose colonies comprise a single queen and king who do all the reproducing. Sexual inactivation of offspring is maintained from one generation to another by pheromonal spread through trophallaxis, which is the mutual feeding of nymphs. Workers and soldiers produce pheromonal secretions whose proportions in the colony act to maintain suitable proportions of each type of individual in succeeding generations.

Some vertebrates, including many primates, also form social systems. These tend to be much less stereotyped than those of the insects, with more flexibility in roles in which individuals perform multiple tasks and all are reproductively competent. Members of vertebrate societies never lose their individuality. Actual reproductive patterns are constrained by behavioral dominance hierarchies, both male and female, that can be quite complex and can be based in large part on kinship. None, other than humans, come close to the frenetic and compartmentalized social activity seen in the colonies of social insects.

The predominant metazoan groups are the chordates, nematodes, arthropods, mollusks, annelids, and echinoderms. The most organically diverse group is the Ecdysozoa, which includes the nematodes and the arthropods. Within the arthropods, the insects alone represent more than 1 million named species, with many more remaining undescribed. Nematodes may be as species rich as insects, but only 15,000 nematodes species have been named. Feats of relative strength and tolerance of extreme environmental conditions are legendary among the insects; among the ecdysozoa, however, tardigrades may be the most amazing extremophiles of all, being capable of entering a state of cryptobiosis. Under desiccating conditions, tardigrades can experience a loss of total water content from 85% of their body weight to 3%. Furthermore, while in this state, tardigrades can withstand, sometimes for years, extreme temperatures, including cold to near absolute zero, ionizing radiation, almost total lack of oxygen, and up to 1000 atm of pressure. When moisture levels return to normal, normal life functions resume.

Although they represent a relatively small percentage of the total species on this planet, metazoans have a major impact on the biosphere. Early in the evolutionary history of metazoans, reef-building sponges and corals evolved, creating biogenic alterations in ocean current patterns and novel forms of habitat for additional life forms. Benthic metazoans make use of, and help maintain, the ocean bottom while annelids, nematodes, rotifers, and a variety of arthropods perform the same function for terrestrial soil systems. Many metazoans serve as habitat for other metazoans: in the life cycle of an individual tapeworm, a copepod, a mollusk, a ray-finned fish, and an elasmobranch may all serve as habitat. The many metazoans that are filter or suspension feeders may act as bio-accumulators. In some cases, their activities may enhance local environmental quality, but this may be a mixed blessing. Oysters and other edible mollusks, for example, are excellent at extracting bacteria and other microbes from the water, but if those microbes include fecal coliform bacteria from human sewage the commercial value of the mollusks as food is severely compromised. The array of complex ecological interactions between plants and insects has permitted the classification of insects as "beneficial," meaning that they do not compete with us for plant biomass, "pest," meaning that they do compete with us for plant biomass, or "vector" when they transmit diseases to us or our livestock. We use parasitoid insects as agents of biocontrol by infecting pests and vectors with them.

III. SUMMARY

Organismal diversity is not just amazing, it is important. The extent to which parts of today's planetary array of organisms can be reduced without creating a cascade of extinctions is unknown but hotly debated. It is known that virtually every species has some value to humans, either directly or indirectly. Organisms provide us with food, shelter, raw materials, things of beauty, and pharmaceuticals; help us with our work; and maintain the ability of our air, soil, and water to sustain life, including our own. They are valuable to us indirectly as the building blocks of the biosphere within which we originally evolved and which we still require in order to survive.

There is also clear evidence that all organisms are intimately tied together in the structure of the biosphere because they are all simultaneously parts of larger genealogical and ecological wholes. The significance of the duality of organismal diversity is most apparent in the recognition that new types of organisms are derived from preexisting organisms, while at the same time almost all organisms make extensive use of the biodiversity that predated their origins. Newly evolved organisms always have an impact on preexisting ones. Because it is in the nature of the organism to be relatively autonomous from its surroundings, these interactions are not necessarily positive, often taking the form of
“conflicts of interest.” The evolutionary resolution of these conflicts of interest has produced an increasingly complex biosphere. As Maynard Smith and Szathmary (1995) observed, each major transition in evolution has been associated with the emergence of organisms, and by extension the entire biosphere, with enhanced abilities to produce, maintain, and transmit information cohesively and also associated with the emergence of novel forms of selection resulting from the evolution of those new organisms. In this way, each newly evolved form of organism becomes intimately involved with both local and global ecology, maintaining the biosphere as a relatively isolated system with its own windows of viability (Ulanowicz, 1997).

See Also the Following Articles

ADAPTATION • ARCHAEA, ORIGIN OF • BIODIVERSITY, ORIGIN OF • DARWIN, CHARLES • DIVERSITY, COMMUNITY/REGIONAL LEVEL • DIVERSITY, MOLECULAR LEVEL • EUKARYOTES, ORIGIN OF • THERMOPHILES

Bibliography

DIVERSITY, TAXONOMIC VERSUS FUNCTIONAL

John C. Moore
University of Northern Colorado

I. Species, Communities, and Ecosystems
II. Species, Function, and the Ecological Niche
III. Estimating Diversity
IV. Biodiversity and Systems Theory
V. Linking Taxonomic and Functional Diversity
VI. Natural Selection

GLOSSARY

functional diversity The different types of processes in a community that are important to its structure and dynamic stability.
functional group A group of species that utilize similar resources; synonymous with guild.
guild A group of species that utilize similar resources (usually food).
species evenness A measure of the relative abundance of species in an area.
species richness The number of species in an area.
species turnover The change in the composition of species in an area due to the extinction of some species and the replacement by a new species by colonization.
taxonomic diversity The number and the relative abundance of species in a community.

TAXONOMIC DIVERSITY refers to the number and the relative abundance of species in a community. Functional diversity refers to the different types of processes in a community that are important to its structure and dynamic stability. Taxonomic and functional diversity have been presented as being different, and as the title implies, competing approaches to the study of biodiversity. This article will provide specific examples of how taxonomic and functional diversity are interrelated and how information from both perspectives can be used to estimate biodiversity and propose mechanisms that govern biodiversity.

I. SPECIES, COMMUNITIES, AND ECOSYSTEMS

Linnaeus (1735) and Hutchinson (1959) provided a convenient starting point to address biodiversity in terms of taxonomic and functional diversity. Linnaeus formalized the two kingdom scheme of Plantae and Animalia and made the distinction between sessile photosynthetic plants and motile food-ingesting animals. Predating an understanding of photosynthesis or modern evolutionary theory, this work was descriptive and simply attempted to categorize and count organisms rather than posit mechanisms. Function was at the basis of the classification. In his address to the Society of American Naturalists, Hutchinson, like Linnaeus, made
a deliberate distinction between primary producers and heterotrophs by asking "why are there so many kinds of animals?" The unique feature of this address was that it studied biodiversity by invoking principles from ecology and evolution, formalizing niche theory, and including energetics. More important, it marks an early attempt to link species diversity to the functioning of an ecosystem.

The choice of taxonomic versus functional diversity illustrates a broader issue in science. In both cases, traditional taxonomies based on individual traits or ones based on the functional attributes of a species in a community or ecosystem are human constructions. Erhlich and Holm (1962) summarized the dilemma nicely:

> There seems to be no theoretical reason why there must be complete congruence among estimates of relationships based on characters from different developmental stages or on characters from different organ systems of the same stage, ... In a taxonomy based on ecological requirements, whales will be more closely related to sharks than to bears. Such a relationship is no more or less "true" than the classical one: It is merely based on different attributes.

At some point the question comes down to the type of information that is used, the way the information is used, and why the information is used.

### A. Different Approaches to Studying Biodiversity

There are different subdisciplines within ecology: autecology (individual/population ecology), community ecology, and ecosystem ecology. The focus and approach of each subdiscipline and the type of information used by each have influenced the degree to which taxonomic or functional diversity have been used (Table 1).

<table>
<thead>
<tr>
<th>Focus</th>
<th>Subdiscipline</th>
<th>Ecosystem ecology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individuals/populations</td>
<td>Community ecology</td>
<td>Forms of matter</td>
</tr>
<tr>
<td>Species</td>
<td>Species/guilds</td>
<td>Functional groupings</td>
</tr>
<tr>
<td>Units</td>
<td>Individuals</td>
<td>Biomass</td>
</tr>
<tr>
<td>Biodiversity</td>
<td>Taxonomic</td>
<td>Functional</td>
</tr>
</tbody>
</table>

### 3. The Ecosystem Approach

The ecosystem approach defines diversity in terms of processes and function within the community. The focus is not on species per se but rather on the products of species interactions, usually different forms of carbon or nitrogen. Underlying this approach is the belief that species interact in ways that reinforce stabilizing feedbacks that operate to maintain a persistent steady state. The feedbacks operate through the availability of nutrients in one pool of that critical to another pool.

A typical description of an ecosystem does not include many species (except for the dominant species) but rather aggregates of species that are similar in the ways in which they process matter relative to a key process. For example, the decomposition of pine needles on a forest floor is a complex process that is influenced by the nutrient quality of the pine needles, the temperature and moisture of the forest floor, and the organisms that are present. For the microbes alone, the process might involve more than 100 species of saprobic fungi and an equal number of strains of saprobic bacte-
ria. Depending on the level of resolution desired, the microbes could be aggregated into a single pool of "decomposers" or separated into pools of "fungi" and "bacteria." The justification for the seemingly casual treatment of the individual species is that the focus is not on how many species are present but rather on the general composition of the pool and the rate at which the pool processes matter.

II. SPECIES, FUNCTION, AND THE ECOLOGICAL NICHE

Taxonomic and functional diversity use similar information but in different ways. The species is central to taxonomic diversity. Species are arranged into phylogenetic relationships that are based on how they acquire energy, their cell types, the level of cellular organization, their embryology and life history, their physiology, and their genetic makeup (RNA and/or DNA). Different classification schemes have been proposed that use all or some of these criteria with different weights given to any one of the criteria. Ecosystem processes are central to functional diversity. Functional diversity has been based on ecological characteristics. Species are aggregated into functional units based on how they use resources. Various schemes have been proposed to aggregate species into functional groupings. Species are arranged into functional relationships based on how they acquire energy, where they acquire energy, the rates at which they process energy (physiology), and when they process energy (life history).

Linking taxonomic and functional diversity would require integrating information that describes the makeup of an organism (its taxonomic characteristics) along with the rates at which it processes and transforms matter (its functional characteristics). Niche theory offers a way to link taxonomic and functional diversity.

A. The Ecological Niche

The ecological niche of a species is defined by the resources it requires. The niche of a species can be viewed graphically as a hypervolume, in which each resource represents an independent axis (Fig. 1). The space that each species occupies within this hypervolume is defined by its resource requirements. An analysis of the resource requirements for several species from different taxa revealed that the principal niche axes are food, habitat use, and time.

B. Functional Groupings of Species

Both the community and the ecosystem perspectives have grouped species into larger units when studying interactions among species or processes mediated by species. The schemes used have ranged from ones of convenience, which are usually based on a broad taxonomic category, to ones that have adopted strict criteria that may include organisms from different taxonomic groupings. For this discussion, the groupings that have been based on the resource requirements of a species and its position in niche space relative to other species makes the most sense, given that function in an ecosystem context is directly related to how species utilize resources. Hence, a functional group is composed of species that utilize similar resources (Fig. 1). There is no standard number of resources (axes in niche space) used to establish functional groups, but many are based on the principal resources of food, habitat, and time. This has resulted in the groupings being given different names. For
example, a "guild" is defined as a group of species that feed on similar food in a similar way. Soil ecologists define a "functional group" as a group of species that feed on similar food in a similar way, occupy the same habitat, and possess similar life histories.

III. ESTIMATING DIVERSITY

Procedures to estimate species diversity originated from the fields of information theory and statistics. If presented a code with a set number of symbols (s) of known proportions (p), it is desirable to estimate its information content. Ecologists simply substituted the species (or aggregations of species) for the symbols. Species diversity of a community could be estimated in the same way as the information content of a code. Two aspects of species diversity are the number of species (species richness) and the relative abundance of species in a community (species evenness).

A. Species Richness

Species richness (S) is the number of species within a defined region. The species richness of a region is obtained through sampling or via a census. Because "region" is defined by the observer, species richness has been further categorized into three components to account for changes in spatial scale.

Alpha diversity, sometimes referred to as point diversity, is the species richness that occurs within a given area within a region that is smaller than the entire distribution of the species. Beta diversity is the rate at which species richness increases as one moves in a straight line across a region from one habitat to another habitat. In other words, it is the rate of change in species richness that occurs with a change in spatial scale. Gamma diversity is the species richness within an entire region. As the area being surveyed approaches that of the entire region, alpha diversity approaches gamma diversity and beta diversity approaches zero.

B. Species Evenness

Species evenness takes into account the number of species and the relative abundance of species in a community. Several indices have been proposed. Two of the commonly used measures of evenness are the Shannon index (H) and the Simpson index (D).

The Shannon index (H) is a measure of the information content of a community rather than of the particular species that is present. The index is as follows:

\[ H = -\sum_{i=1}^{s} p_i \log p_i \]

where s is the number of species in the community (species richness) and \( p_i \) is the relative proportion of species i. Hence, if two communities (A and B) with the same number of species (not necessarily the same species) present were compared, and the distributions of \( p_i \) were the same, then \( H_A = H_B \). On the other hand, if either the number of species or the proportions differed, then \( H_A \neq H_B \). In the latter case, the community with the greater number of species whose relative abundances were equal would possess the higher diversity.

The Simpson index (D) measures the dominance of a multi-species community and can be thought of as the probability that two individuals selected from a community will be of the same species. The Simpson index was originally proposed as follows:

\[ D = -\log \sum_{i=1}^{s} p_i^2 \]

where s is the species richness of a community and \( p_i \) is the relative proportion of species i. The index can be modified to \( 1 - D \) to give it the property of increasing as diversity increases (the dominance of a few species decreases).

C. Linking Species Richness and Species Evenness

Species richness and species evenness are special cases of dichotomous-type rarity measures. For dichotomous-type rarity measures for a pool of s species, the rarity of species i, whose function is denoted \( R(p_i) \), depends only on its relative abundance \( (p_i) \). The rarity measures \( \Delta \) take the following form:

\[ \Delta = \sum_{i=1}^{s} p_i R(p_i) \]

The relationships between species richness \( (S) \), the Shannon index \( (H) \), and the Simpson index \( (D) \) are
presented in Table II. Notice that the indices share three important properties: (i) $\Delta$ is maximized when $p_i = 1/s$ for all species; (ii) given that $p_i = 1/s$, the community with the higher species richness has the greater diversity; and (iii) $\Delta = 0$ if the community possesses a single species ($s = 1$).

### IV. BIODIVERSITY AND SYSTEMS THEORY

One of the enduring tenets of ecology in the early part of the twentieth century was that a diverse community was a stable community. The linkage between biodiversity and stability has an intuitive appeal. The more species there are in a community, the more likely they would be tightly networked, and hence the more stable the community would be. Interestingly, function plays an important role in this argument in that in a highly diverse community, there would be alternative species or assemblages of species to perform key tasks to sustain the cycling of matter and the current diversity. These ideas have been challenged and modified. Ecologists agree that biodiversity is important to the function and stability of an ecosystem. The current debates center on the nature of the relationship between biodiversity, function, and stability.

Although useful, the functions used to estimate biodiversity that were presented previously are descriptive measures and imply neither the mechanisms that shape biodiversity nor the importance of biodiversity to maintaining the structure and dynamic stability of a community. If the diversity of one community were higher than that of another, does this increase its likelihood of persisting in time or recovering from a disturbance? The answer depends on the species involved. For example, the colonization of an exotic species increases the species richness of a community but does not necessarily increase its stability—to the contrary, it often destabilizes the system. On the other hand, the addition of a predator can stabilize an ecosystem by tempering the oscillations in the population densities of its prey through time. Biodiversity is a property of complex dynamic biological systems. The steady state and stability are two concepts important to the study of complex dynamic systems.

#### A. The Steady State

A steady state is an equilibrium condition in a dynamic process in which the rate of input of a state variable is equal to the rate of output of that variable (Fig. 2). The dynamic process can involve any variable that changes in time. For taxonomic and functional diversity, the variables of interest are the number of species or functional groups, the densities and biomass of each species or functional group, and the mass of key elements within different pools within the ecosystem.

1. **The Number of Species**

The number of species in a community is at steady state when the rate of colonization equals the rate of extinction (Fig. 3). The classic model of this process, the theory of island biogeography (MacArthur and Wilson, 1967), assumes that the rate of colonization is affected by the distance that the habitat being colonized is from the source of species and by the size of the habitat being colonized. Colonization rates are higher for large habitats that are close to the source of species than for smaller distant habitats. The rate of extinction is affected by the size of the habitat. The rates of extinction for smaller habitats are higher than those for larger habitats.

2. **Species Composition**

In the model presented previously, although the number of species is constant at steady state, the species composition of the community does not remain constant. The model simply assumes that once the rates of colonization and extinction are equal, then for every new species that enters a habitat, one of the current resident species goes extinct. This point is important when trying to estimate the diversity of a habitat by sampling the habitat over time. If a habitat were sampled long enough, the estimate of the number of species within the habitat would equal the number of species in the habitat serving as the source, even though the source has a higher number of species at its steady state than the habitat being colonized.

3. **Population Densities**

The population density of an individual species within a habitat is at steady state when the number of births
and immigration equal the number of deaths and emigration. When modeling this process, ecologists typically assume that the habitat is closed to immigration and emigration and focus entirely on the rate of birth and the rate of death. More sophisticated approaches that involve multiple habitats will include the immigration and emigration of species among habitats.

4. Community Structure
A community is at steady state when the species composition of the community and the densities of the species remain constant. This implies that the species richness and evenness of the community do not change. This condition is rarely met since the species composition and the densities of the species fluctuate through time.

5. Ecosystem Processes
An ecosystem process is at steady state when the rate of input of matter or energy is equal to the rate of output. The input of matter and energy occurs via photosynthesis and from sources external to the ecosystem (e.g., leaves entering a stream). The export of matter and energy occurs through heterotrophic respiration and the physical movement of matter from the ecosystem (e.g., export of leaves from forest to a stream). The production of organic matter from inorganic matter (photosynthesis) and the transformation of organic matter to other organic forms and inorganic matter (heterotrophic respiration, hereafter referred to as respiration) are ecosystem processes that approach a steady state and illustrate the interplay between matter and energy. When the rate of photosynthesis equals the rate of respiration, the system is said to be at steady state.
not disagree that dynamic stability affects biodiversity, yield different results with regard to the nature of the relationship.

The concepts of local stability and the theory of island biogeography may seem at odds with one another, even though they rely on similar mathematics and assumptions and are modeling the same systems. An obvious question is, how can an ecosystem be locally stable if the species composition of the ecosystem is changing? The discrepancy can be clarified by revisiting the processes that are being modeled. When modeling populations within a community, local stability is in reference to the densities or biomass of the populations. The dynamics of the populations are governed by the difference between the rates of birth and death. When modeling the species diversity within the same community, local stability is in reference to the number of species that are present. The dynamics of the number of species is governed by the rates of colonization and extinction. To carry this a step further, the local stability of the carbon or nitrogen cycles within an ecosystem is in reference to the mass of the different forms carbon or nitrogen. The dynamics of the different forms is governed by the functional attributes of the species present and the rates at which the different forms are acted upon.

2. Resilience

Resilience is a measure of a stable system’s ability and the time required to recover from a minor disturbance. The principle of resilience can be applied to any state variable present in an ecosystem: population densities, the number of species, or the mass of nutrients. Resilience is a function of the rates of inputs of the state variables, the rate at which the state variables turn over, and the number of state variables.

The return time of a system is a measure of its resilience. More resilient ecosystems have shorter return times than less resilient ones. Return time is estimated as the time between the point when the disturbance altered the ecosystem and the point at which the ecosystem recovers. The persistence of a system is compromised if the frequency of disturbance is too great given the length of return time because a system may not have time to recover from any one incident.

3. Oscillations

The state variables of most ecosystems do not maintain a constant steady-state condition but rather they exhibit oscillations. Oscillations are fluctuations in the magnitude of a state variable through time that can be intrinsic to the dynamics of the system in the absence of a distur-
bance or can occur during the recovery phase of a system following a disturbance or both. The degree to which a system oscillates can affect its stability, persistence, and diversity. For example, if the amplitude of oscillations are too great or are such that the population size of a species approaches zero, then the likelihood of that species going extinct increases.

C. Modeling Species Interactions

Gardner and Ashby (1970) presented a simple mathematical exercise that illustrated how the number of interacting units (species) within a system and the number of interactions between these units (species interactions) related to the stability of the system. The exercise constructed simple systems of 4, 7, and 10 units assumed to be at a steady state. For a system of a given size, connections were made between the units in a way that ensured that all units would be connected to the system, creating several variations of the system in terms of the connections. Next, the strength of the connections was allowed to vary, and the systems were disturbed by moving the units away from their steady state. Two results emerged from their work. First, as the number of interacting units increased within a system the likelihood of the system returning to a steady state after a small disturbance (stability) decreased. Second, for a given number of interacting units, as the number of connections between the units increased, the stability of the system declined.

There is a clear connection between the simple mathematical exercise presented above and the diversity and stability of ecological communities (May, 1973). The species is substituted for the generic units modeled previously, and the interactions are defined in terms of predation, competition, or mutualism. Like the generic modeling exercise, high biodiversity does not necessarily lead to stability. Additional studies demonstrated that systems composed of species aggregated into blocks of highly interactive species that were connected by weaker interactions were more stable than systems with the same number and types of species and interactions arranged at random. Hence, the way in which the community is constructed affects its diversity and stability.

The types of interactions among species affect the stability of a community. Theory and observation agree in general terms on this point, if not on specifics. For example, mutualistic interactions, whereby the actions of each species benefit the growth and dynamics of one another, and omnivory, in which species derive energy from different trophic levels, have been shown to be destabilizing in simple models, but both are prolific in natural communities. All the dominant terrestrial ecosystems rely on a mutualism between the vascular plants and microbes, and nearly every described ecosystem possesses numerous examples of omnivory.

D. Modeling Species and Processes

Process-oriented models focus on the transformation of elements among various inorganic to organic states through ecosystems. Carbon, nitrogen, and phosphorus are the more common choices of elements to be modeled given their importance to plant and animal growth. Models of species interactions have been connected to models of processes. The general approach to linking these models is to use the structure and formulation adopted when modeling species with the currency used in models of processes. First, species are aggregated into functional groups using one of the schemes described previously, and the interactions among the functional groupings are established. Second, the biomass, expressed in terms of carbon, nitrogen, or phosphorus, of each functional group is estimated. Third, a series of differential equations is developed that establishes which functional groups interact with one another and the nature of the interactions. Parameters within the models include birth and death rates, feeding rates, and the efficiencies by which the functional groups convert consumed matter into new biomass. The models are used to estimate the flow of elements among the functional groups, the flow of elements out of the system, the stability of the ecosystem, and the resilience of the ecosystem.

Three results have emerged from these models. First, the models are sensitive to the changes in the parameters associated with the physiology and life histories of the species within the functional groups. For example, microbes possess high reproductive rates and high rates of energy conversion, and they process matter at a higher rate compared to a functional group dominated by arthropods or nematodes, which possess lower reproductive rates and rates of energy conversion. Second, functional groups are arranged into interactive blocks that form pathways of material flow that originate from a specific source and end with a top predator. The blocked arrangement of functional groups, the distribution of biomass within the ecosystem, and the distribution of the flow of elements through the ecosystem influence the stability and resilience of the system.
V. LINKING TAXONOMIC AND FUNCTIONAL DIVERSITY

The linkage between taxonomic diversity and functional diversity involves many facets. The morphology of an organism determines the type of habitat in which it resides and its ability to colonize new habitats. The physiology of a species influences its adaptiveness to a habitat and the rates and efficiency with which it transforms matter. The life history of an organism influences the rate at which it processes matter (births, deaths, and consumption). Linking taxonomic diversity to functional diversity requires finding measurable attributes that all organisms possess that have meaning to all the ecological subdisciplines, but it also requires flexibility in terms of how the information is used. Three case studies are presented in the following sections. The case studies illustrate how, when studying biodiversity, aspects of both taxonomic and functional diversity are used interchangeably.

A. Diversity of C₃ and C₄ Grasses

The first case study involves a comparison of the distribution of grass species that possess either the C₃ or C₄ photosynthetic pathways in North America. The study illustrates how the information of autecology can be used to find mechanisms to explain patterns in diversity. The study demonstrates how the functional aspects of an organism's physiology influence a species' adaptability to a habitat and ultimately the diversity of species with similar physiologies.

Plants that possess C₃ and C₄ photosynthetic pathways differ in terms of their morphology and biochemistry as they relate to the dark reaction of photosynthesis. For C₃ plants, photosynthesis occurs entirely within the outer mesophyll cells, in which CO₂ enters the Calvin-Benson cycle through the plants' stomata. The enzyme (RuBP carboxylase or Rubisco) that incorporates CO₂ into the Calvin-Benson cycle also has an affinity toward O₂. If O₂ were incorporated instead of CO₂ the plant would experience photorespiration—energy is expended, no carbon is fixed, and the plant loses organic carbon. For C₄ plants, there are two types of photosynthetic cells: the outer mesophyll cells and the inner bundle sheath cells. In C₄ plants, CO₂ enters the mesophyll through the stomata and is incorporated into a four-carbon compound before being exported to the bundle sheath cells in which it enters the Calvin-Benson cycle. The additional steps that occur within the C₄ plants prior to the Calvin-Benson cycle cost the plant additional energy and organic carbon. However, the enzyme (PEP carboxylase) that adds the CO₂ to the four-carbon molecule in the mesophyll cells is more efficient than Rubisco and less likely to initiate the loss of organic carbon to photorespiration. This trade-off places plants with the C₃ pathway at a disadvantage under cool and moist conditions compared to those with the C₄ pathway, but it affords them an advantage in warmer and dry conditions.

Within North America, the northern latitudes are dominated by species of C₄ grasses, whereas the southern latitudes are dominated by species of C₃ grasses. This pattern of diversity has been attributed to the adaptiveness of each pathway to patterns of temperature and light intensity. The lower latitudes possess higher temperature and higher light intensity than the upper latitudes given the orientation of the earth to the sun. Grasses that possess the C₃ pathway have been shown to be more productive than those with the C₄ pathway under conditions of cool temperature and low light intensity. At higher temperatures and light intensity, C₃ plants suffer greater water loss through their stomata and increased photorespiration than do C₄ plants.

B. A Test of the Theory of Island Biogeography

This case study illustrates how information from the community perspective has been used to study biodiversity. The experimental manipulations of several small mangrove islands off the Florida coast by Simberloff and Wilson (1969a, b) and the reanalysis of their work by Heatwole and Levins (1972) offer not only the best validation of the theory of island biogeography but also some insight into how taxonomic diversity and functional diversity are related.

Simberloff and Wilson (1969a, b) conducted a thorough survey of arthropods on four small mangrove islands and then fumigated the islands with methyl bromide to eliminate the arthropods. For 2 years following the disturbance, they sampled for arthropods that recolonized the islands. The results were consistent with those predicted by the model proposed for the theory of island biogeography. The colonization rates were higher and the steady states in species richness were achieved sooner on islands closer to the mainland than on the more distant island. Once a steady state in species richness was achieved, the species composition changed due to the predicted turnover in species.

Heatwole and Levins (1972) provided an explicit link between the recovery process, species diversity, and functional diversity. They aggregated the species
into guilds based on the feeding and foraging behaviors of the arthropods (predators, herbivores, decomposers, ants, etc.). One year following the fumigation, the taxonomic diversity of the islands was lower than it was prior to the fumigation (Table III). However, the proportion of species within each of the guilds was the same as it was prior to fumigation. These results suggest that the functional diversity of the islands recovered before the species diversity. In other words, the taxonomic diversity of the communities was in some way dependent on functional characteristics of the community.

C. Invasion of an Exotic Species

The final case study illustrates how elements of taxonomic and functional diversity and information from autecology, community ecology, and ecosystem ecology have been used to study biodiversity. Vitousek and colleagues (1987, 1989) studied the natural history of the invasive plant species Myrica faya and the effects that it has had on the nitrogen cycle and plant community development of recent lava flows in Hawaii. Since its introduction to Hawaii from the Canary Islands by Portuguese immigrants at the end of the nineteenth century, M. faya had colonized more than 34,365 ha. Myrica faya forms a symbiotic association with the soil nitrogen-fixing actinomycete, Frankia. The plant possesses two characteristics that separate it from the native nitrogen-fixing plants on the islands. Unlike the native plant nitrogen fixers (Acacia koa and Sophora chrysophylla), the seeds of M. faya are small, have an edible fleshy fruit, and pass through the digestive systems of birds intact. Moreover, M. faya is a more efficient nitrogen fixer than the native species. When combined, these factors make M. faya a superior colonizer of recent lava flows, pastures, and forests.

Once established, the M. faya and its soil symbiont, Frankia, alter the nitrogen status of the soils. For example, soil nitrogen inputs within open canopies of M. faya were four times more than those of similar sites without the exotic plant (23.5 and 5.5 kg per hectare per year, respectively). This added soil nitrogen becomes available to other plant species, facilitating invasions by other exotic species (e.g., strawberry guava, Psidium cattleianum) to the detriment of native plant species. The functional attributes of a single species and its symbiont affected the biodiversity of a community by altering the course of community development.

VI. NATURAL SELECTION

Species evolve and communities and ecosystems change, but do communities and ecosystems evolve? Does natural selection operate on communities and ecosystems even though they do not possess genes? Natural selection needs to be defined before these questions can be answered. Natural selection is the differential perpetuation of genes in successive generations caused by different degrees of adaptiveness to the environment.

One argument is that ecosystems do evolve because the composition of species within a community changes and natural selection operates on the gene pools of each species. In addition to the changes in the genetic makeup of the species, there are changes in key processes that feed back to reinforce stable configurations or that initiate change. A criticism of this analogy is that natural selection at the ecosystem level for the type of coordination among species that leads to a stable ecosystem is not the same as selection at the organismal level on genes. Selection on genes operates through differential reproductive success of individuals, whereas this may not be true at the ecosystem level since the boundaries of an ecosystem are not so clearly defined.

An alternative view posits that ecosystems do not evolve, but that the species composition of the stable ecosystem is a result of the selection operating on individuals and the coevolution of species interacting with one another. The outcome of selection on individuals

<table>
<thead>
<tr>
<th>TABLE III</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Functional Groups of Arthropods by Number and Relative Proportion (%) on Islands before and 1 Year after Fumigation</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>No. and relative proportion (%)</th>
<th>Herbivore</th>
<th>Scavenger</th>
<th>Detritus</th>
<th>Wood borer</th>
<th>Ant</th>
<th>Predator</th>
<th>Parasite</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>55 (34)</td>
<td>7 (4)</td>
<td>13 (8)</td>
<td>8 (5)</td>
<td>32 (20)</td>
<td>36 (22)</td>
<td>12 (7)</td>
<td>153</td>
</tr>
<tr>
<td>After</td>
<td>55 (40)</td>
<td>5 (4)</td>
<td>8 (6)</td>
<td>6 (4)</td>
<td>23 (17)</td>
<td>31 (23)</td>
<td>9 (6)</td>
<td>137</td>
</tr>
</tbody>
</table>

* Adapted from Heatwole and Levins (1972).
can influence function within a community, which in turn feeds back as a form of selection. Unique to this argument is that the adaptiveness of a species includes the degree to which its dynamics and function contribute to and are compatible with the stability of the ecosystem as a whole.

See Also the Following Articles

GUILDS • HABITAT AND NICHE, CONCEPT OF • ISLAND BIOGEOGRAPHY • MEASUREMENT AND ANALYSIS OF BIODIVERSITY • PLANT INVASIONS • SPECIES DIVERSITY, OVERVIEW • STABILITY, CONCEPT OF • TAXONOMY, METHODS OF

Bibliography


I. Introduction

The original focus of farming was on growing food crops, but this was soon extended to include plants essential for other utilities such as fiber sources, stimulants, medicinal aids, dye crops, forage, and timber production. In addition, the past two or three centuries witnessed another major development, namely a massive introduction of ornamental plant species into cultivation. (Today the number of ornamental species exceeds the number of food crops.) Most recently, plants with medicinal properties are getting new attention. The potentials of scores of such plant species (traditionally collected from the wild) have been evaluated. To ensure a steady supply, and for improving quality and yields, many of them are presently introduced into cultivation.

A parallel trend is apparent in forestry. Here, too, dozens of timber producing trees are being quickly domesticated. To date, thousands of plant species, native to the various phytogeographic regions of the World, have been introduced into cultivation. In the majority of these cases, the domestic derivatives have already been drastically altered by this move. So much so that the cultivated forms can no more survive in the wild and have become dependent on humans for their existence. Grain crops, for example, lost the wild type adaptation to disperse their seeds and depend on the harvesting, threshing, and sowing by the farmer. Vegetatively propagated tuber plants (under domestication) are commonly sterile. So are many seedless varieties of fruit crops (e.g., bananas). Moreover, as a result of (a) the ecological shift from the original wild habitats into the greatly different and varied anthropogenic environments and (b) repeated cycles of deliberate selection by humans, many domesticates have greatly diverged.
under domestication. In most major food crops (e.g., bread wheat, rice, maize, apple, or grapevine) or in many popular ornamentals (e.g., roses, rhododendrons, narcissi, tulips, irises), hundreds or even thousands of distinct cultivated varieties (“cultivars”) or ornamental forms are now recognized, fitting the wide range of conditions under which these plants are being maintained and the different and contrasting wishes of the growers who constantly try to improve their plants by selection. All in all, domestication provides us with most impressive examples of rapid evolution in plants. Vast arrays of morphological, anatomical, physiological, and chemical changes have evolved in crops and ornamentals, and this in a very short span of evolutionary time. Moreover, selection under domestication molded the gene pools of the crops (the “cultivated gene pools”) to become the backbone for the production of food and other essential utilities for humankind. Today, our very existence depends on this genetic diversity. Thus not only do the crops become dependent on humans; also humans became dependent on crops. Obviously, civilizations and domesticated plants have coevolved; and this coevolution continues today.

II. THE TRADITIONAL AGRICULTURAL SYSTEMS OF THE WORLD

In classical times (about 2500–2000 years ago), and to a large extent until Columbus’s time, agriculture was structured quite differently from what we find it today. At present, each country grows a mixture of crops introduced from the various parts of the world. But this fusion is very recent. Traditionally, farming was practiced in largely separated, independent regions, each occupying extensive territories and each possessing a characteristic and largely unique ensemble of native crops. When the Europeans discovered America, they found that the flourishing agricultural civilizations in the New World were based on growing crops that were totally different from what they were familiar with in Eurasia and in Africa. The two hemispheres are indeed very rich in cereals, pulses, tubers, fruits, vegetables, spices, and other types of cultivated plants. Yet the list of domesticated plants in each of them is unique. With only one or two exceptions, these two landmasses originally had no crops in common. This strongly indicates that domestication and agriculture in each of them evolved independently.

Also within each of the two hemispheres, one can detect similar patterns. Each landmass seems to contain several historical agricultural regions, previously separated by difficult-to-cross geographic and cultural barriers. Thus the crops of the traditional agriculture system in Southwest Asia, the Mediterranean basin, and temperate Europe were very different from those grown in East Asia. Moreover, the crop composition in each one of these two agricultural systems is very different from the crop assemblage in Africa south of the Sahara. Similarly in the New World, numerous crops grown in South America were not known in Mexico and middle America, and vice versa.

The main food crops in the traditional agricultural regions of the world are given in Appendix 1. A comparison between the various regional lists of crops reveals that each agricultural system usually contains dominant cereals, companion pulses, oil plants, fruit crops, vegetables, spices, tubers and corms, and stimulants. In other words, the composition of crop blends repeats itself. Only the plant species are different in each traditional region (for example, wheats and barley, lentil, and pea in Southwest Asia, the Mediterranean basin, and Europe; rice and soybean in East Asia; maize and Phaseolus beans in America; and sorghum and other millets and cowpea in Africa south of the Sahara).

III. THE EMERGENCE OF AGRICULTURE

The shift from hunting and gathering to farming started relatively very late in the history of Homo sapiens. The archaeological and botanical evidence already assembled on this development is not even. Some parts of the world (the Mediterranean basin, the Nile Valley, temperate Europe, Southwest Asia, North America) have already been extensively studied. Other parts of the world (East Asia, the Indian subcontinent, South America) are much less explored but at least provide us with some critical evidence on the beginning of agriculture in them. In still others (Africa south of the Sahara, most of the tropical parts of the world) archaeological and botanical information is still very poor or almost nonexistent. Even so, the available information clearly shows that farming was independently initiated in several geographically and culturally separated “nuclear areas,” both in the Old World and in the New World, between 10,000 and 5,000 years before present (a.r.) uncali-
brated radiocarbon time. In each such area cultivation seems to have been independently invented, and native wild plant species—largely unique to each area—were introduced and tested in cultivation. Some evolved to be the ‘founder crops’ that initiated farming in these independent places. Once this new way of life had been successfully established, these domesticates also formed the basis for a territorial expansion of the newly formed agricultural populations and for further development in food production.

A. The First Old World Territory

The oldest and the most extensively studied nuclear area is the ‘fertile crescent’ belt of the Near East. In this rather small geographic territory, a string of early Neolithic farming villages started to appear some 9500 years ago and further established themselves in the subsequent 1500 years. Plant remains retrieved from these early archaeological sites show that eight grain plants growing wild in this Near Eastern ‘arc’ were domesticated at these times. Most common in these pre-ceramic Neolithic Near Eastern sites were remains of three cereals: emmer wheat, einkorn wheat, and barley. They were accompanied by pulses (lentil, pea, chickpea, bitter vetch) and by flax. Soon later, by 8900–8600 n.r., signs of domestication of sheep, goat, cattle, and swine appeared in the sites as well, and the Near Eastern “package” for food production was formed. Once the package was assembled, and the early Neolithic farming villages were soundly established, farmers started to expand. They did so explosively. By 8000 n.r. this type of agriculture was already introduced into Greece, and by 7000 n.r. it had reached the Danubian basin, the Nile Valley, the Caucasus, and Turkmenistan, and soon afterward it was introduced in the Indus basin. By 6500 to 6000 n.r., grain crop agriculture (and cattle and caprine rearing) was already established all over major parts of temperate Europe—from Ukraine in the east to northern France in the west. More or less at the same time, this new technology also appeared in the middle and western parts of the Mediterranean basin as far as south Spain. All over these vast territories of Southwest Asia, the Nile Valley, the Mediterranean basin, and temperate Europe, agriculture was started by the introduction of the same Near Eastern crops (wheats and barley, and their companion pulses and flax). Only later were additional plants, native to other parts of this huge region, added (e.g., poppy Papaver somniferum in the west Mediterranean basin)

Some 3000 years after the start of Neolithic grain agriculture, fruit crop horticulture (based on the invention of vegetative propagation) appeared in this region. As with grain crops, the earliest convincing signs of fruit tree cultivation were found in the Near East. Olive, fig, grapevine, and date palm have been cultivated here at least since Chalcolithic times (5500 n.r.). Horticulture, too, spread quickly. By the late Bronze Age, olive, grapevine, and fig (as well as pomegranate and almond) were already used as principal elements of food production in the Aegean belt; and date palm groves flourished in the warm fringes of the Near East and the eastern Mediterranean basin (including Egypt) and extended eastward as far as the Indus Valley.

From the Bronze Age onward there are sound indications of cultivation of vegetables. Melon, watermelon, onion, garlic, leek, and lettuce were the first vegetable crops grown in Mesopotamia or in Egypt. Definite signs of their cultivation appear by 4500 to 3300 n.r. By 2800 to 2000 n.r., the list of Mediterranean and Southwestern Asian vegetable crops had grown considerably; and beet, turnip, cabbage, radish, carrot, parsnip, celery, parsley, and asparagus had also entered cultivation. More or less at the same period (2400–2000 n.r.), a second group of native fruit trees (those in which cultivation depends on grafting) was also added. Most conspicuous among them are apple, pear, plum, and cherries. Contrary to the earlier crops that were almost all taken into cultivation in the Near East “arc,” many of the vegetables and the late fruit trees were probably picked up for domestication, not in the Near Eastern nuclear area but in other parts of the already vast Mediterranean and Southwest Asian system of agriculture.

Thus, starting in the Neolithic and ending in classical times, an impressive assemblage of native crops were domesticated and diffused all over Southwest Asia, the Mediterranean basin, and temperate Europe. Most of them remained economically important until today.

B. The Second Old World Territory

A second Asian nuclear area—or more exactly two closely situated, but independent initiations of farming—have been discovered in East Asia. The first was uncovered in the loess soil belt in Honan province and adjacent areas, in the middle part of the Yellow River Valley in North China. The second in the Hupei basin.
in Central China, along the middle Yangtze River, and also in the wetlands of its delta. The evidence available on the start and on the subsequent development of agriculture in China is still fragmentary compared to that available on the Near East. Yet very ancient farming villages have been discovered both in the northern and in the central parts of this large country. They represent the oldest undisputed signs of plant domestication found yet in East and in South East Asia.

In the Yellow River Valley the oldest farming sites date from 8800 to 7000 B.P. and they contain very large amounts of foxtail millet, indicating that in temperate North China this cereal had been the principal founder crop of the local agriculture. Also in later times foxtail millet continues to be a major food crop in China. More or less in the same time, farming villages seem to have appeared in the middle of the Yangtze River Valley. The earliest sites discovered there date even somewhat older (8500–7000 B.P.) than those in north China. Significantly, in these sites the dominant crop is rice, which later emerges as the most important crop of the vast agricultural economies of East and Southeast Asia. Massive remains of this cereal were unearthed in the Hupei basin, and they continue to appear in the Yangtze River Valley and in later archaeological contexts. These are the oldest remains of rice found yet. But because they occur on the fringe of the wild rice distribution area, some workers suspect that even older rice sites will be discovered in the future further south.

C. The Third Old World Territory

A third territory in the Old World suspected to have had an independent initiation of agriculture is the Savanna belt and its forest margin in Africa south of the Sahara. The traditional agriculture in this region (which stretches from Senegal and Guinea to Sudan) is based, almost entirely, on a rich assemblage of native crops such as sorghum, pearl millet, African rice, fonio, cowpea, and bambara groundnut (Appendix 1). The highlands of Ethiopia border the Savanna belt in the east and also harbor an impressive list of local crops, such as teff, finger millet, coffee, and noog. But here the African plants grow side by side with wheat, barley, and many other Near Eastern elements that reached Ethiopia in the past. All in all, African agriculture is diverse and adapted to a wide range of situations. Apart from the Near Eastern elements in Ethiopia, its rich crop assemblage seems almost fully indigenous. This is borne by the fact that the wild relatives of the crops are restricted in their distribution to this part of the World. Most of them could have been introduced into cultivation only somewhere in this geographic belt. All this seems to suggest the independent initiation of cultivation. Whether this is indeed the case is yet hard to say. Archaeologically, Africa south of the Sahara is practically unexplored. The few data available on cultivation of pearl millet or other crops are no more than 3000 years old.

D. The American Territory

There is little doubt that farming was independently invented in the Americas. As already noted, in pre-Columbus times the Eastern hemisphere and the Western hemisphere shared almost no crop in common. On basis of this evidence it is hard to imagine agricultural contacts between the two landmasses before that time. Also in America itself archaeological research has revealed definite signs of early cultivation, not in a single area but in several distant locations.

Indications of early, possibly independent initiation of cultivation come from South America. Chemical comparisons have shown that the common bean and lima bean were independently domesticated both in the middle and in South America. Finds in the central and south Andes show that potato and Quinua cultivation in these elevated areas might be 5000 years old, although dating is yet problematic. Finally, sound evidence has been assembled that independent domestication of plants took place in eastern North America. Here goosefoot Chenopodium berlandieri, sunflower, and marsh elders were cultivated at least 4000 to 3500 years ago. This is some 2000 years before maize or any other outsider crop reached this area.

IV. REPRODUCTIVE SYSTEMS IN PLANTS UNDER DOMESTICATION

Cultivated plants differ markedly from wild plants in the way by which they are reproduced and maintained.
In the various families of the flowering plants, cross-pollination is the principle genetic system. Most species build out-breeding populations. Other reproductive systems such as self-pollination and vegetative propagation (including apomixis) occur as well and are even common in some genera and families. Yet compared to sexual reproduction by cross-pollination, their overall weight is small. In contrast, cross-pollination is relatively rare in plants under domestication. Most crops are maintained by one of the following two systems: (a) self-pollination and (b) vegetative propagation. In contrast to cross-pollination, both these reproductive systems are effective in bringing about immediate ‘fixation’ of desired genotypes.

A. Self-pollination

Self-pollination, or more exactly almost full self-pollination, is the principle mating system found in grain crops and in many vegetables. The majority of the 50–60 main grain crops of the world are predominantly self-pollinated. Only a few (such as maize, rye, pearl millet, buckwheat, or scarlet runner bean) are cross-pollinated. Now that the wild progenitors of the majority of the grain crops are already satisfactorily identified, we know that the wild ancestors of the self-pollinated crops are also self-pollinated. In other words, self-pollination in grain crops did not evolve under domestication. It is rather a ‘preadaptation’ of the wild ancestor, which considerably enhances its chance to perform successfully in cultivation. One major advantage of self-over cross-pollination is the fact that self-pollination isolates the crop reproducitively from its wild progenitor. It enables the farmer to grow a desired genotype in the same area in which the wild relatives abound without endangering the identity of the cultivar by genetic swamping. A second advantage of self-pollination lies in the genetic structure maintained within the crop. Self-pollination results in splitting the crop’s gene-pool into independent homozygous lines. Variation is thus structured in the form of numerous true breeding cultivars. Because they are automatically ‘fixed’ by the pollination system, they can be easily maintained by the farmer, even if they are planted together. In contrast, the preservation of varietal identity in cross-pollinated plants is much more problematic. It requires repeated selection towards the desired norms, constant care to avoid the mixing of types and the prevention of contamination from undesirable plants. It is therefore not surprising that early successes in plant domestication frequently involved selfers. In fact, all the eight Neolithic Near Eastern founder crops have this system of reproduction.

B. Vegetative Propagation

Vegetative propagation is the second widely adapted means to fix and maintain desired genotypes under domestication. This way of handling prevails in the fruit trees, in tuber and corn crops, and in numerous ornamentals. Here, domestication means first of all changing the reproductive biology of the plants involved by shifting from sexual reproduction (in the wild) to vegetative propagation (under cultivation). As a rule, cultivated varieties of these plants have been maintained as clones by cuttings, rooting of twigs, suckers, or by the more sophisticated technique of grafting (and recently also by meristem tissue culture propagation). This is in sharp contrast with their wild progenitors, which reproduce from seeds. Their wild growing populations are usually variable, maintain themselves through sexual reproduction, and are distinctly cross-pollinated. Consequently, seedlings raised from any mother plant segregates widely in numerous traits, including the size, shape, and palatability of the fruits or the tubers.

In the hands of the grower, vegetative propagation has been a powerful device to prevent genetic segregation and to ‘fix’ desired genotypes. By discarding sexual reproduction and inventing clonal propagation, the farmer was able to select, in a single act, exceptional individuals with desirable traits from among a large numbers of variable, inferior plants. Moreover, the farmer could duplicate the chosen types to obtain genetically identical saplings. This is no small achievement. Because the plants are cross-pollinated and widely heterozygous, most progeny obtained from them (even from ‘superior’ plants) are economically worthless. The change from seed planting to vegetative propagation has been a practical solution to assure a dependable supply of desired genotypes. In most fruit trees—and in many tuber crops and ornamentals—it made domestication possible. The history of the Mediterranean fruit trees illustrates this clearly. Olive, grapevine, fig, and date palm that can be relatively simply cloned (by twig rooting, by suckers) were the first to enter cultivation (already by 5500 B.C.). Other native fruit trees that do not lend themselves to easy rooting (e.g., apple, pear, plum, cherry) were domesticatecl much later (about 2400–2000 B.C.). Their incorporation into horticulture became possible only after the invention of a new clonning tool, namely scion grafting.
V. CONSCIOUS VERSUS UNCONSCIOUS SELECTION

Two main types of selection operate (and complement each other) in plants under domestication:

1. Conscious selection deliberately applied by the growers for traits of interest to them.
2. Unconscious selection brought about by the fact that the plants concerned were picked up and isolated from their original wild environment and placed in a new (and usually very different) human-made environment. This shift in the ecology led automatically to drastic changes in selection pressures. In response to the introduction of the plants into the anthropogenic environment, numerous adaptations vital for survival in the wild lost their fitness. New sets of traits were automatically selected for to fit the new conditions, resulting in the buildup of characteristic domestic adaptations and the appearance of "domestication syndromes"—each fitting the specific agricultural system provided by the domesticators.

Although the pressures and the effects of conscious selection in shaping domesticated plants are familiar to both plant breeders and crop plant evolutionists, unconscious selection has been less frequently evaluated. Because this type of selection shapes so many of the principal traits that characterize crops, several developments caused by the shift in the ecology are discussed in some detail in the following sections.

VI. METHODS OF MAINTENANCE OF CROPS AND THEIR IMPACT

As already noted, two main methods of maintenance are employed by the cultivators to maintain their crops: (a) planting of seeds and (b) vegetative propagation. The choice between these two agronomic practices is also the choice between two contrasting patterns of selection and modes of evolution under domestication.

With very few exceptions of apomixis (such as meiotic seed in citrus fruits and mango) planting of seeds can be equated with sexual reproduction. Cultivated plants maintained by seeds (the bulk of the grain crops, numerous vegetables and forest trees, some ornamentals) therefore undergo a recombination-and-selection cycle every planting. In other words, such crops had, under domestication, hundreds or even thousands of generations of selection. They have been continually molded either as (a) clusters of inbred lines (in predominantly self-pollinated crops) or as (b) distinct cultivated races (in cross-pollinated plants). In numerous sexually reproducing crops, the results of such repeated selection are indeed striking. Under domestication these crops have diverged considerably from their wild progenitors. They are now distinguished from them by numerous morphological, developmental, physiological, and chemical traits.

Crops maintained by vegetative propagation (most fruit trees and tuber crops, few vegetables, many ornamentals) have had an entirely different history of selection. Cultivars in these crops are not true races but only clonal replications of exceptional individuals, which are as a rule highly heterozygous. They had been picked up by the cultivator from among numerous sexually produced variable individuals and "fixed" by cloning. In terms of selection, the development of a vegetatively propagated cultivar is largely a single-step operation. With the exception of rare somatic mutations, selection is completed the moment the clone is picked up. In traditional fruit tree horticulture, the turnover of clones has been low. Appreciated genotypes were maintained for long periods of time. Thus, clonal crops underwent in cultivation only a few recombination-and-selection cycles. In sharp contrast to sexually reproducing cultivated plants, their cultivars do not represent true breeding races but only clones, which, as a rule, are highly heterozygous and segregate widely when progeny tested. Significantly, the majority of such segregating progeny are not only economically inferior or even worthless, but they often regress toward the means found in wild populations, showing striking resemblance to the wild forms.

VII. THE PURPOSE FOR WHICH THE PLANT IS BEING GROWN

In different crops, different parts of the plant body are being used. Some are raised for their vegetative parts (tubers, leaves, stems, etc.). In others the reproductive parts (inflorescences, flowers, fruits, seeds) constitute the used products. Also the choice of the desired plant parts introduces automatically contrasting selection pressures, particularly in regard to the reproductive system of the crops involved.

When crops are grown for their seeds (or at least when they are reproduced by seed), they are kept (like their wild relatives) under constant stabilizing selection
that keeps their reproductive system intact. Grain crops provide us with the most rigid cases of such normalizing selection. Yields in these crops depend decisively on the streamlined development of flowers and fruits, normal chromosome pairing in meiosis, and full fertility. Deviants are weeded out automatically and the reproductive system is kept in balance. It is no wonder that among cultivated plants, grain crops are the most conservative in this regard. They are characterized by strictly balanced chromosome systems and show very little chromosome "problems" under domestication. With very few exceptions (such as the formation of hexaploid bread wheat), the chromosome complements in cultivars of grain crops are identical to those found in their wild progenitors.

Drastic changes in seed fertility (as well as in the chromosome system) can be tolerated when the plants are grown for their fruits but are maintained by vegetative propagation. Crops in this group (the bulk of the fruit trees) do not depend on seed fertility for their maintenance. Yet they have to keep the basic reproductive growth functions to assure the development of fruits. Moreover, growers usually prefer types in which the size and the number of stones or pips had been reduced, or seedless fleshy fruits. Several solutions for reducing seed fertility without harming fruit set evolve automatically under cultivation. They include the establishment of polyploid clones, some of them meiotically unbalanced (e.g., triploidy in some pears and in bananas), or the incorporation of mutations inducing parthenocarpy—that is, the induction of fruit development without fertilization and without seed set (e.g., bananas, fig, some pears).

Crops grown for their vegetative parts and maintained by vegetative propagation exhibit the most drastic disruption of their reproductive system and the most variable and bizarre chromosomal situations among cultivated plants. Because such crops are clonally propagated, the conscious selection exerted on them by the growers to increase the output of the desired vegetative part is rarely counterbalanced by normalizing selection to retain sexual reproductive functions. Tropical root and tuber crops provide us with outstanding examples for this development under domestication. Cultivated clones of cassava, yam, sweet potato, or garlic frequently show drastic reduction in flowering. In some, flowering ceases altogether, or almost altogether. When flowers develop they are frequently sterile. Also chromosomally many of these crops are highly polyploid or variable aneuploid chromosome numbers. Clones may contain 3x, 5x, or even higher meiotically unbalanced chromosome complements. Thus in the yams, Dioscorea alata shows all chromosome levels between 3x to 8x, whereas in D. esculenta, 4x, 6x, 9x, and 10x forms are known. Sugarcanes confront us with even more complex chromosome pictures. Cultivated clones in this crop are all highly polyploid and frequently aneuploid. Modern cultivars range from 2n = 100 to 2n = 125 chromosomes. Older cultivars vary from 2n = 80 to 2n = 124.

VIII. THE IMPACT OF SOWING AND REAPING

Traditional grain agriculture is based on sowing the seeds of the crop in a tilled field, reaping the reproductive parts of the mature plants, and threshing out the grains. The introduction of grain plants into such a farming practice mutated, automatically, selection toward the following changes, setting them apart from their wild counterparts.

1. Most conspicuous is the unconscious selection for mutants in which the mature seed is retained on the mother plant (i.e., for the breakdown of the wild mode of seed dissemination). In cereals this implies a shift from shattering spikes or panicles (in wild forms) to nonshattering ones (in plants under domestication). In pulses it usually means the prevention of the pod from dehiscing and from shedding the seeds. In wild wheat, wild barley, or wild maize, the seed dispersal unit comprises a single internode of the ear. Full disarticulation of the ear, at each segment, is thus an essential element for wild type dispersal, and plants in wild populations are constantly selected for quick shattering of their mature spikes. In contrast, the introduction of planting and harvesting brings about automatic, unconscious selection in exactly the opposite direction. Under the new system a sizable proportion of seed produced by brittle plants will shatter and would not be included in the harvest, whereas all grains produced by nonshattering mutants "wait on the stalks" to be reaped by the grower. Under cultivation, nonshattering individuals have therefore a much better chance to contribute their seed to the subsequent sowing. To summarize, nonbrittle mutants that were totally disadvantaged under wild conditions became highly successful under the new system. Thus, when wild cereals are introduced into the system of sowing and reaping, one should expect selection for nonbrittle forms whether or not the cultivator is aware of this trait. Furthermore, the incorpora-
tion of nonbrittle mutants makes the crops fully dependent on humans, as nonshattering plants lose their seed dispersal ability and can no longer survive under wild conditions. For this reason, the presence of nonshattering remains of grain crops in archaeological contexts provide critical indication that these plants were under cultivation. Finally, both theoretical considerations and experimental evidence seem to suggest that at least in wheats and barley, the establishment of nonshattering mutants under the system of sowing and reaping was very likely a fast process. The shift could have been accomplished in the course of several scores of generations.

2. A second major outcome of introducing wild grain plants into this regime of cultivation is the breakdown of the wild mode of their seed germination. Most wild plants, especially annuals, depend for their survival on regulation of their germination in space and time. A common and vital adaptation is a delay of germination and its spread over two or more years. Again, under the system of sowing and reaping such inhibition of germination is automatically selected against. Most grain crops have lost their wild-type germination inhibition patterns. Their seeds germinate fully whenever the farmer plants them.

3. Several other traits characterize cultivated grain crops and are part of the domestication syndrome under sowing and reaping conditions. Because of a rather dense stand in the cultivated field there is a stress to develop forms with erect habit and to reduce the growth of side tillers or branches. Because of the mode of harvest there is a selection for synchronous ripening. In cases of biennial or perennial wild stocks, there is an automatic selection for the shift to annuality. Because the seeds stay protected in the granary, thick shells are selected against. Finally because tilling increases soil fertility there is also a pressure to increase the number of fertile flowers in the reproductive parts of the crops.

IX. APPENDIX I: CROPS OF VARIOUS REGIONS

This appendix lists the native crops of the various traditional agricultural regions of the world (see Section II). For illustrations and short notes on the various food crops, consult Vaughan and Geissler (1997). The important crops of the world and their evolution are surveyed in Smartt and Simmonds (1995).

SOUTHWEST ASIA, MEDITERRANEAN BASIN, AND TEMPERATE EUROPE

Cereals
- Emmer and durum-type wheats *Triticum turgidum*
- Einkorn wheat *Triticum monococcum*
- Bread wheat *Triticum aestivum*
- Barley *Hordeum vulgare*
- Rye *Secale cereale*
- Common oat *Avena sativa*

Pulses
- Lentil *Lens culinaris*
- Pea *Pisum sativum*
- Chickpea *Cicer arietinum*
- Faba bean (bread bean) *Vicia faba*
- Oil or Fiber
  - Flax *Linum usitatissimum*
- Oil seed turnip *Brassica rapa*

Fruits and Nuts
- Olive *Olea europaea*
- Grapevine *Vitis vinifera*
- Fig *Ficus carica*
- Date palm *Phoenix dactylifera*
- Apple *Malus pumila*
- Pear *Pyrus communis*
- European plum *Prunus domestica*
- Sweet cherry *Prunus avium*
- Almond *Amygdalus communis*
- Walnut *Juglans regia*

Vegetables and Spices
- Melon *Cucumis melo*
- Watermelon *Citrullus lanatus*
- Onion *Allium cepa*
- Garlic *Allium sativum*
- Leek *Allium porrum*
- Lettuce *Lactuca sativa*
- Beet *Beta vulgaris*
- Turnip *Brassica rapa*
- Cabbage *Brassica oleracea*
- Carrot *Daucus carota*
- Celery *Apium graveolens*
- Parsley *Petroselinum sativum*
- Asparagus *Asparagus officinalis*
- Cumin *Cuminum cyminum*
- Coriander *Coriandrum sativum*
- Fenugreek *Trigonella foenum-graecum*
EAST ASIAN AGRICULTURE

Cereals and Cereal-Like Grain Crops
Rice Oryza sativa
Foxtail millet Setaria italica
Broomcorn millet Panicum miliaceum
Buckwheat Fagopyrum esculentum

Pulses
Soybean Glycine max

Oil or Fiber
Hemp Cannabis sativa

Vegetables and Spices
Pe Tsai (Peking cabbage) Brassica pekinensis
Pak-choi (Chinese cabbage) Brassica chinensis
Rakkyaya Allium chinense
Chinese chives Allium tuberosum
Chinese pickling melon varieties Cucumis melo

Fruits and Nuts
Chinese chestnut Castanea henryi
Chinese white pear Pyrus bretschneideri
Chinese sand pear Pyrus pyrifolia
Loquat Eriobotrya japonica
Oriental persimmon Diospyros kaki
Litchi Litchi chinensis
Apricot Prunus armeniaca
Peach Prunus persica

Corms and Tubers
Chinese varieties of turnip Brassica rapa
Chinese yam Dioscorea esculenta
Lotus Nelumbium nucifera

Stimulants
Tea Camelia sinensis

SOUTHEAST ASIAN AGRICULTURE

Pulses
Sword bean Canavalia gladiata
Mung bean Vigna radiata
Black gram Vigna mungo
Rice bean Vigna caracata

Oil or Fiber
Sesame Sesamum indicum
Tree cotton Gossypium arboreum
Coconut Cocos nucifera
Jute Corchorus capsularis

FRUITS AND NUTS
Banana: cultivated derivatives of Musa acuminata and M. balbisiana

Vegetables and Spices
Cucumber Cucumis sativus
Eggplant Solanum melongena
Black pepper Piper nigrum
Cardamom Elettaria cardamomum
Turmeric Curcuma longa
Nutmeg Myristica fragrans
Cloves Syzygium aromaticum
Ginger Zingiber officinale

Corms and Tubers
Taro Colocasia esculenta
Greater yam Dioscorea alata
East Indian arrowroot Tacca leontopetaloides

SUGAR
Sugarcane Saccharum officinarum

SUB-SAHARAN AFRICAN AGRICULTURE

Cereals
Sorghum Sorghum bicolor
Pearl millet Pennisetum glaucum
Finger millet *Eleusine coracana*
Fonio *Digitaria exilis*
African rice *Oryza glaberrima*

**Pulses**
Cowpea *Vigna unguiculata*
Bambalata groundnut *Voandzeia subterranea*

**Oil or Fiber**
Oil palm *Elaeis guineensis*
Noog *Guizotia abyssinica*

**Vegetables and Spices**
Gherkin *Cucumis anguria*

**Corms and Tubers**
Yam *Dioscorea cayenensis*
Enset *Ensete ventricosa*

**Stimulants**
Coffee *Coffea arabica*

### American Agriculture

**Cereals and Pseudo-Cereal Crops**
Maize *Zea mays*
Amaranth *Amaranthus cruentus*
Amaranth *Amaranthus caudatus*
Quinoa *Chenopodium quinoa*
Goosefoot, *Chenopodium berlandieri*
Marsh elder *Iva annua*

**Pulses**
Common bean *Phaseolus vulgaris*
Lima bean *Phaseolus lunatus*
Scarlet runner bean *Phaseolus coccineus*
Tepary bean *Phaseolus acutifolius*
Jack bean *Canavalia ensiformis*
Peanut (groundnut) *Arachis hypogea*

**Oil or Fiber**
Sunflower *Helianthus annuus*
Island cotton *Gossypium hirsutum*
Upland cotton *Gossypium barbadense*

**Fruits and Nuts**
Avocado *Persea americana*
Prickly pears *Opuntia ficus-indica*
Sapodilla *Manilkara zapota*
Guava *Psidium guajava*
Passion fruit *Passiflora edulis*
Papaya *Carica papaya*
Pineapple *Ananas comosus*
Chirimoya *Anona cherimola*

**Vegetables and Spices**
Tomato *Lycopersicum esculentum*
Peppers *Capsicum annuum* and *C. frutescens*
Squashes and pumpkins *Cucurbita pepo*, *C. maxima*, *C. moschata*, *C. mixta* and *C. ficifolia*
Vanilla *Vanilla fragrans*

**Corms and Tubers**
Potato *Solanum tuberosum*
Sweet potato *Ipomoea batatas*
Oca *Oxalis tuberosa*
Yam *Dioscorea esculenta*
Cush-cush *Dioscorea trifida*

**Stimulants**
Cacao *Theobroma cacao*
Tobacco *Nicotiana tabacum*
Coca *Erythroxylon coca*

See Also the Following Articles

*AGRICULTURE, TRADITIONAL* • *BREEDING OF ANIMALS* • *BREEDING OF PLANTS* • *CROP IMPROVEMENT AND BIODIVERSITY* • *EDIBLE PLANTS* • *PLANT SOURCES OF DRUGS AND CHEMICALS*

### Bibliography


ECOLOGICAL FOOTPRINT, CONCEPT OF

William F. Rees
UBC School of Community and Regional Planning

I. The Conceptual Framework: Why Do We Need Ecofootprint Analysis?
II. Revisiting Human Carrying Capacity: The Roots of Ecofootprint Analysis
III. The Logic and Structure of Ecological Footprint Analysis
IV. Human Ecological Footprints: Implications for Global Sustainability
V. Human Ecological Footprints and the Biodiversity Crisis
VI. Assessing the Ecological Footprint Approach
VII. Epilogue

GLOSSARY

carrying capacity Usually defined as the average maximum number of individuals of a given species that can occupy a particular habitat without permanently impairing the productive capacity of that habitat (see also human load).

competitive exclusion The displacement of one species from its habitat or ecological niche by another. When humans appropriate other species' "ecological space," it often leads to the local or even global extinction of the nonhuman organism.

ecological deficit An ecological deficit exists when the "load" (see human load) imposed by a given human population on its own territory or habitat (e.g., region, country) exceeds the productive capacity of that habitat. In these circumstances, if it wishes to avoid permanent damage to its local ecosystems, the population must use some biophysical goods and services imported from elsewhere (or, alternatively, lower its material standards).

human load The total "human load" imposed on the "environment" by a specified population is the product of population size times average per capita resource consumption and waste production. The concept of "load" recognizes that human carrying capacity is a function not only of population size but also of aggregate material and energy throughput. Thus, the human carrying capacity of a defined habitat is its maximum sustainability supportive load.

overshoot A population is in overshoot when it exceeds available carrying capacity. It may survive temporarily but will eventually crash as it depletes vital natural capital (resource) stocks. A population in overshoot may permanently impair the long-term productive potential of its habitat, reducing future carrying capacity.

patch disturbance The measurable habitat and ecosystem modification caused by large animals, including humans, as they forage for food or other resources. Patch disturbance is most pronounced near the den site, temporary camp, or other central place within the overall home range of the individual or group.
sustainability gap. The global ecological deficit—that is, the difference between any excessive human load on the ecosphere and the long-term carrying (or load-bearing) capacity of the planet.

ECOLOGICAL FOOTPRINT ANALYSIS is a quantitative tool that represents the ecological load imposed on the earth by humans in spatial terms. Thus, the ecological footprint of a defined population is the total area of land and water ecosystems required to produce the resources that the population consumes, and to assimilate the wastes that the population generates, wherever on earth the land/water are located. Ecofootprinting can be used to assess the ecosystem area effectively "appropriated" in support of any specified human population or economic activity. We can then compare this to available productive area (i.e., load-bearing capacity). The size and nature of the human ecological footprint is relevant to biodiversity conservation because energy and material resources extracted from nature to serve human purposes are irreversibly unavailable to other species. The larger the human footprint, the less nonhuman biodiversity. This article describes the contemporary context for ecofootprint analysis and the theory on which it is based. It then discusses the implications of the analysis for sustainable development in general and for the future of biodiversity in particular.

I. THE CONCEPTUAL FRAMEWORK: WHY DO WE NEED ECOFOOTPRINT ANALYSIS?

A. Human Activity and Global Ecological Change

There is little disagreement among scientists that humanity has assumed a major role in changing the face of the earth. Up to one-half of the land has been directly transformed by human action; more than half of the planet's accessible fresh water is already being used by people; atmospheric carbon-dioxide has increased by 30% in the industrial era; more atmospheric nitrogen is fixed and injected into terrestrial ecosystems by humans than by all natural terrestrial processes combined; two-thirds of the world's major fisheries are fully- or overexploited; and biodiversity losses are accelerating (Lubchenco, 1998, Vitousek et al., 1997).

What is in dispute is whether any of this really matters to socioeconomic sustainability. The more pessimistic analysts see human domination of the earth and the deterioration of nature as ultimately fatal to civilization itself. Various mechanisms are invoked in coming to this conclusion. For example, because of our culture's near total dependence on liquid petroleum and the present lack of affordable substitutes, energy analyst Richard Duncan (1993) argues that the total life expectancy of western technindustrial society is less than 100 years—counting from the 1930s! Professor W. M. Hern attacks industrial culture's obsession with growth and compares humanity to a self-destructive cancer on the planet. The continuous expansion of the human enterprise apparently "exhibits all four major characteristics of a malignant process: rapid uncontrolled growth; invasion and destruction of adjacent tissues (ecosystems, in this case); metastasis (colonization and urbanization, in this case); and dedifferentiation (loss of distinctiveness in individual components") (Hern, 1997).

According to geographers J. W. Smith and G. Sauer-Thompson, environmental decay coupled with geopolitical stresses induced by resource scarcity, will inevitably cause industrial society to "self-destruct, producing massive ecological damage, social chaos, and mega-death" (Smith and Sauer-Thompson, 1998). From these perspectives, the human enterprise has already overshot global carrying capacity, is eroding future options, and will inevitably crash—there is virtually nothing we can do to prevent the collapse of global civilization.

At the other extreme, we can take comfort in the view of those scientists and economists who believe that human ingenuity will prevail. Their logic is simple. Stimulated by incipient scarcity, rising prices, and the opportunity for profit, scientists and entrepreneurs will find technological substitutes for exhaustible resources and natural processes. As energy economist R. L. Gordon explains, "the tendency to technical progress is viewed as the most critical economic law involved." And because human ingenuity has historically been remarkable at increasing living standards and warding off the pressures of resource depletion, "the immediate need for avoiding depletion is nil" (Gordon, 1994).

Perhaps the most extreme expression of this belief is revealed in the ebullient optimism of the late Julian Simon who wrote: "Technology exists now to produce in virtually inexhaustible quantities just about all the products made by nature" and "We have in our hands now ... the technology to feed, clothe, and supply energy to an ever-growing population for the next seven billion years" (Simon, 1995). Some conventional economists even argue the benefits of continuous population growth on grounds that "additional people can create more resources than they use up, thanks to technologi-
cal improvements" (Block, 1990). Clearly, from the techno-optimists' perspective, there are no practical limits at all on human carrying capacity—we don't need to do anything to sustain global civilization but stay our present, increasingly market-driven, global development course.

With such diametrically opposing assessments of the human prospect, both starting from present circumstances and held with equal conviction, it is little wonder that ordinary citizens and policy makers alike are confused about global ecological change and what—if anything—to do about it. This chapter assesses the overall sustainability conundrum using ecological footprint analysis and links the general problem to the specific issue of biodiversity conservation. First, however, we examine the origins of the ongoing perceptual conflict.

B. Dueling Paradigms

Many of the conflicting views about "sustainable development" can be traced to differing fundamental beliefs, values, and assumptions about the nature of reality, particularly humankind-environment relationships. Different societies and interest groups have different preanalytic visions (or worldviews or paradigms), which shape how they interpret any set of facts or circumstances.

Remarkably, people are often unaware that they possess a particular worldview—each person acquires his or her culturally specific paradigm unconsciously, simply by living and growing up in a particular sociocultural milieu. We may therefore be unconscious of the subtle ways in which the prevailing worldview shapes our approach to critical issues or that there may be more viable alternatives. When we think that "the world is flat" was once accepted as self-evident truth, it raises the unsettling possibility that much of our present cultural worldview consists largely of shared illusions.

The dominant social paradigm in western (and increasingly global) technocultural culture can be characterized as the "expansionist" or "cornucopian" worldview. As we have already seen, its adherents believe that humankind has achieved mastery over the natural world and that technology will be able to compensate for the depletion of natural resources and the loss of life-support services.

The economics of expansionism is the neoclassical (neoliberal) economics that has come to dominate geopolitics in the past 20 years. Neoclassical models generally represent the economy as a self-regulating and self-sustaining mechanical system whose productivity and growth are not seriously constrained by nature. In effect, conventional economics treats the economy as separate from, and more or less independent of, "the environment." People can therefore be treated solely as economic agents.

By contrast, ecologists, environmentalists, and many pessimists profess an ecological or steady-state worldview. From this perspective, the economy, far from being a separate system, is seen as an inextricably integrated, completely contained, and wholly dependent growing subsystem of a non-growing ecosystem. People are not only economic agents but also ecological agents.

The ecological perspective recognizes that the so-called environmental crisis is really a human ecological crisis. The distinction is by no means trivial. The former term externalizes the problem, effectively blaming it on a defective environment, which then needs to be fixed. By contrast, the latter term traces the problem to its source—the nature and behavior of people themselves—and suggests that it is the latter that need "fixing." This is the starting point for ecological footprint analysis.

C. Human Ecology/Economy and the "Second Law"

If humans are ecological entities then human activity is governed by natural laws. In particular, economic production requires continuous, irreversible energy and material transformations, and these transformations are ultimately governed by the second law of thermodynamics. This fundamental law is ignored by conventional economic models.

In its simplest form, the second law states that any isolated system will tend toward equilibrium; alternatively, the "entropy" of any isolated system always increases. Available energy spontaneously dissipates, concentrations disperse, gradients disappear. An isolated system thus becomes increasingly unstructured in a inexorable slide toward thermodynamic equilibrium. This is a state of maximum entropy in which there is no structure and nothing can happen.

The second law was originally formulated for simple isolated systems close to equilibrium. We now recognize, however, that even complex open systems are subject to the forces of entropic decay. Any differentiated far-from-equilibrium system has a natural tendency to erode and unravel.

But not all complex systems do disintegrate in this way. Many biophysical systems, from individual lizards to the entire ecosphere, actually gain in organizational
complexity and mass over time (i.e., they increase their distance from equilibrium). This seemingly paradoxical behavior can readily be reconciled with the second law. Biophysical systems exist in loose, nested hierarchies, each component system being contained by the next level up and itself comprising a chain of linked subsystems at lower levels. Living systems can therefore import available energy and material (essergy) from their host environments and use it to maintain their internal integrity against entropic decay and to grow. They also export the resultant waste (entropy) back into their hosts. In effect, modern formulations of the second law posit that all highly ordered systems develop and grow (increase their internal order) at the cost of destroying order at higher levels in the systems hierarchy. Because such systems maintain themselves by continuously degrading and dissipating available energy and matter, they are called "dissipative structures."

Now, the human economy is clearly one such highly ordered, complex, dissipative structure. Both our biological and our industrial metabolisms require enormous inputs of high-grade energy and material resources from the rest of the ecosphere for their maintenance and growth. Indeed, from the ecological perspective, humans are strictly secondary producers—all production by the human enterprise, from the increase in population to the accumulation of manufactured capital, requires the consumption of a much larger quantity of energy and material first produced by nature (although we can sometimes increase the amount of incoming solar energy captured by photosynthesis through irrigation, or by building terraces for agriculture, for example).

Moreover, recall that within the global hierarchy, the economy is a dependent subsystem of the ecosphere. In effect, the expanding human enterprise is positioned to consume the ecosphere from within. Beyond certain limits, continuous population and material economic growth is, therefore, inherently self-destructive. This reality suggests a necessary second law condition for sustainability: material consumption and waste production by the economy must be no greater than the resource production and waste assimilation capacity of the ecosphere.

D. Homo Sapiens: The Archetypal Patch Disturbance Species

Evidence to support the ecological view has accumulated throughout history. Even preagricultural humans had significant effects on local ecosystems' structure and function and on the biodiversity of their habitats. This was the inevitable consequence of the second law combined with two additional facts of human biology: human beings are large animals with correspondingly large individual energy and material requirements (i.e., we are big consumers), and humans are social beings who live in extended groups. Consequently, whenever human hunter-gatherers invaded a previously "stable" ecosystem, their material demands would significantly alter established energy and material pathways, benefiting some nonhuman species and harming others. People invariably perturb or disturb the systems of which they are a part.

Perhaps the most dramatic evidence is the permanent systemic changes that occur when humans first invade and settle a new habitat. Consider biodiversity loss. The recent paleoecological, anthropological, and archeological literature tells a convincing story of the extinctions of large mammals and birds that accompanied first contact and settlement of their habitats by human beings (Diamond, 1992; Flannery, 1994; Ponting, 1991). It seems that everywhere on earth that paleontologists have studied and that humans first reached within the past fifty thousand years, human arrival approximately coincided with massive prehistoric extinctions. In North America, South America, and Australia, about 72, 80, and 86%, respectively, of large mammal genera ultimately became extinct after human arrival. Scientists estimate that with only Stone Age technology, the Polynesians exterminated more than 2000 bird species, about 13% of the world total.

All this is to emphasize that humans are, by nature, a patch-disturbance species, a distinction we share with other large mammals ranging from beavers to elephants. Large animals, due to their size, longevity, and food and habitat requirements, tend to have substantial physical and systemic impacts on the ecosystems that sustain them. A patch-disturbance species may thus be defined as any organism that, usually by central place foraging, degrades a small "central place" greatly and disturbs a much larger area away from the central core to a lesser extent.

There is, of course, a major difference between human patch disturbance and that of other species. Because human knowledge and technology are uniquely cumulative, human patch disturbance has been intensifying since the Neolithic. It received a major boost with agriculture and became the dominant force in the ecosphere with the use of fossil fuels and the industrial revolution. (Cheap, plentiful fossil fuels have enabled humans to accelerate the exploitation of everything else.) Today, human patch disturbance is evident on a global scale in the form of such persistent negative
trends as greenhouse gas accumulation, increasing climatic variability, ozone depletion, landscape destruction, and accelerating biodiversity loss. Ultimately, the erosion of systems integrity may lead to the irreversible loss of basic life support functions.

Whatever joy one might take from the human symphony, the fact that economic growth now threatens critical biophysical systems sounds a discordant note. Global change implies that economic activity is pressing against the limits of human carrying capacity at the expense of other species. Regrettably, the concept of human carrying capacity has long been rejected by mainstream analysts, but for reasons explained below, it is the organizing principle of ecological footprinting.

II. REVISITING HUMAN CARRYING CAPACITY: THE ROOTS OF ECOFOOTPRINT ANALYSIS

Ecologists who study nonhuman species generally define carrying capacity in terms of the numbers of deer, elk, or other species that a particular habitat type can support indefinitely per unit area. However, because of such factors as climatic variability and community succession, instantaneous carrying capacity is constantly changing. Carrying capacity is therefore not a particularly rigorous concept, even when applied to nonhuman species with simple stable demands on their ecosystems. It is all the more problematic when applied to humans whose demands on the environment are anything but simple and stable.

Indeed, many economists and other cornucopians argue that interregional trade alone is virtuously enough to cancel concerns about human carrying capacity. Any region or country that can trade services or surpluses of resource A for needed supplies of resource B need not be limited in numbers or economic growth by the theoretical carrying capacity of its home territory. Furthermore, if trade fails, we can always rely on human ingenuity and technology to increase carrying capacity. These are seemingly powerful arguments, and they explain how such densely populated countries as the Netherlands and the United Kingdom can remain economically vigorous despite having long exceeded their domestic carrying capacities.

But let’s examine the assumptions here. The idea that trade acts to increase local carrying capacity treats each trading region as an isolated open system. Looked at in the aggregate the situation becomes cloudier. The world as a whole is a materially closed system. Resources imported to and consumed in region A are no longer available for consumption in the exporting region B, and visa versa. Hence, the exchange may result in a one-time increase in the population of each region taken separately, but it also increases global consumption and may deplete local stocks of natural capital, thereby limiting future options. Thus, although total human load increases, there is no unambiguous increase in total load-bearing capacity.

Indeed, in some circumstances unfettered trade can lead to a permanent loss of carrying capacity. Global trade exposes pockets of scarce resources everywhere to the largest possible market. This subjects them to ever-greater exploitation pressure, often to the point of depletion or collapse. (Such is the history of trade in fisheries products, for example.) In short, instead of increasing load-bearing capacity, trade often just shuffles it around. This ensures that all countries, their economies happily expanding through trade, hit the (now shrinking) limits to growth simultaneously. Stratospheric ozone depletion is a case in point.

What about human ingenuity? The general argument here is that cumulative knowledge and technological skills enable humans to squeeze more out the environment than nature-in-the-raw can provide. This is achieved mainly by continuously increasing resource productivity and by creating substitutes for any resources that are eventually depleted. In these ways, we are able to increase the capacity of any given habitat to support humans and their activities without practical limit.

Let’s take these points one at a time. The so-called green revolution, for example, is frequently cited as a miracle of human technological prowess. High-input production agriculture has, in fact, greatly increased the short-term productivity (carrying capacity) of arable land. This has seemingly banished the threat of Malthusian famine and certainly facilitated the modern population explosion. However, this miracle was achieved largely by supplementing renewable sun, soil, and rain with nonrenewable fossil fuel, fertilizer, pesticides, and depletable groundwater at a great cost in soil degradation, biodiversity loss, widespread pollution, and falling water tables. The net effect is a swollen human population, likely in overshoot, and increasingly dependent for survival on nonrenewable artificial inputs, even as pollution mounts and natural long term load-bearing capacity is steadily eroded.

Another potential fix is our demonstrated ability to greatly increase resource productivity, to do more with less. In theory, more efficient technologies should en-
able a constant energy and material supply to support a given population at a higher material standard, or a higher population at the same material standard, thereby seeming to increase carrying capacity.

Unfortunately, real-world practice has so far betrayed theoretical promise. The steady gains in material efficiency in the post World War II period have been accompanied by both an exploding population and steadily increasing consumption. A recent study of resource flows in a selection of the world's most technologically advanced and efficient countries found that although there has been some reduction in the ratio of resource inputs per unit gross domestic product since 1975, there has also been "in most, a gradual rise in *per capita* natural resource use." The study therefore concludes that "meaningful dematerialization, in the sense of an absolute reduction in natural resource use, is not yet taking place" (WRI, 1997, p. 2). In fact, efficiency gains may actually work against conservation by lowering prices and raising incomes. This gives people more money to spend on cheaper goods and services so total throughput—human load—keeps expanding.

Finally, overconfidence in resource substitution is also problematic. It is true that human ingenuity has been markedly successful at finding replacements for simple material inputs to production. For example, optical fiber is a product of mind, not of nature, and has displaced much of the demand for copper wire. However, as ecological economist Herman Daly (1991) has emphasized, there are many more circumstances in which technology and nature are complements, not substitutes. More fish boats cannot long substitute for fewer fish; more saws and carpenters are no substitute for less lumber. Indeed, natural capital (e.g., timber, steel, aluminum) is actually a prerequisite for manufactured capital (e.g., fish boat) since the latter is made from the former. More generally, we should emphasize that the growing need for technology-based substitutes (e.g., fertilizer) is a warning signal that natural load-bearing factors (e.g., productive soil) are being permanently eroded.

In summary, confidence that expanding trade and technical wizardry have infinitely expanded human carrying capacity and banished the Malthusian specter forever is premature. Carrying capacity is still very much applicable to humans, and given present trends it is likely to become a major preoccupation of global development policy in coming decades. However, as this discussion will show, we need new approaches to the analysis of carrying capacity, methods that account for uniquely human cultural attributes, including trade and technology.

III. THE LOGIC AND STRUCTURE OF ECOLOGICAL FOOTPRINT ANALYSIS

A. Basic Concepts

Ecological footprint analysis is one such analytic method. Ecofootprinting starts from the premise that, conscious of it or not, modern human beings are integral components of the ecosystems that support them and therefore still very much dependent on "the land" (Rees, 1996; Wackernagel and Rees 1996).

The method also explicitly recognizes (a) that whether one consumes locally-produced products or trade goods, the land connection remains intact, however far removed from the point of consumption some of that land may be, and (b) that no matter how sophisticated our technology, the production/consumption process requires some land- and water-based ecosystems services. Ecofootprint analysis thus incorporates the trade and technology factors simply by inverting the standard carrying capacity ratio: rather than asking what population can be supported by a given area, ecofootprinting estimates how much area is needed to support a given population, regardless of the location of the land or the efficiency of relevant technologies.

Ecofootprinting builds on traditional trophic ecology by constructing what is, in effect, an elaborate "food-web" for the study population. This requires quantifying the material and energy flows supporting the population and identifying corresponding significant sources of resources and sinks for wastes. As we have noted, the human food-web differs significantly from those of other species. In addition to the material and energy required to satisfy the metabolic requirements of our bodies, a human food-web must also account for the needs of our industrial metabolism.

Ecofootprinting is further based on the fact that many material and energy flows (resource consumption and waste production) can be converted into land and water-area equivalents. These are the ecosystem areas required to produce the biophysical goods and services used by the study population. Thus, the ecological footprint of a specified population is the area of land and water ecosystems required to produce the resources consumed and to assimilate the wastes generated by that population on a continuous basis, wherever on earth the land/water may be located. A complete ecofootprint analysis would therefore include both the area the population "appropriates" through commodity trade and the area it needs to provide its share of certain free land-
and water-based services of nature (e.g., the carbon sink function).

The area of a given population's ecofootprint actually depends on four factors: the size of the population, the people's average material standard of living, the productivity of the land/water base, and the technological efficiency of resource harvesting, processing, and use. Regardless of how these factors interact, ecofootprinting represents critical natural capital requirements of the study population in terms of corresponding productive land and water area. We can also think of the ecological footprint as representing the extended patch (productive habitat) occupied ecologically by the study population.

It is critical to recognize that ecofootprints represent ecologically exclusive areas. The productive capacity used by one human population is not available for use by another. Although two or more human groups may share in the output of the same exporting region, the total ecosystem area appropriated is the sum of the areas required by the individual populations. In the final analysis, all human populations are in competition for the available load-bearing capacity of the earth.

### B. The Method in Brief

Population ecological footprints are generally based on final demand for goods and materially intensive services by the referent population. Thus, the first step in calculating the ecological footprint of a study population is to compile, item by item, the total annualized consumption of each significant commodity or consumer good used by that population. Data are obtained from national production and trade statistics and other sources such as United Nations statistical publications. For accuracy, consumption data should be trade-corrected. Thus the population's consumption of wheat can be represented as follows:

\[
\text{consumption}_{\text{wheat}} = \text{production}_{\text{wheat}} + \text{imports}_{\text{wheat}} - \text{exports}_{\text{wheat}}.
\]

The second step is to convert consumption of each item into the land area required to produce that item by dividing total consumption by land productivity or yield. This actually gives us the ecological footprint of the individual item. In general:

\[
a_i = c_i / y_i,
\]

where: \(a_i\) is the ecofootprint of item \(i\) in hectares, \(c_i\) is total consumption of item \(i\) in kilograms, and \(y_i\) is the yield of item \(i\) in kilograms per hectare. Thus, for wheat:

\[
a_{\text{wheat}} = c_{\text{wheat}} / y_{\text{wheat}} = \text{kg}_{\text{wheat}} / (\text{kg}_{\text{wheat}} \times \text{ha}^{-1})
\]

Next, the aggregate ecological footprint of the population, \(F_p\), is determined by summing the footprints for the \(n\) individual items:

\[
F_p = \sum_{i=1}^{n} a_i
\]

Finally, the per capita ecological footprint, \(f_c\), is obtained by dividing the total population footprint by population size, \(N\):

\[
f_c = F_p / N
\]

For some wastes (such as carbon dioxide emissions) or nutrients (such as phosphates and nitrates), it is also possible to calculate the land (or aquatic) ecosystem area required for sustainable assimilation and recycling. In these cases, the assimilation rate per hectare and year is substituted for \(y\) (yield) noted earlier. Box 1 illustrates a basic ecofootprint calculation (aggregate grain consumption in Canada).

We calculate ecofootprint estimates using standard spreadsheet software. Spreadsheets for recent ecofootprint calculations feature up to 132 rows (consumption items) and 15 columns (yield, production, imports, exports, apparent consumption, etc.), with the last column giving the ecofootprint components attributable to each consumption item (see Wackernagel et al., 1999).

Double-counting is avoided whenever identified. For example, if in the analysis of a particular city's ecological

---

**Box 1**

Sample Calculation: The Ecological Footprint of Grain Consumption in Canada (1993 data)

- Total population (N): 28,817,000
- Average yield (y): 2744 kg × ha⁻¹ (2.74 tonnes × ha⁻¹)
- Domestic production (p): 51,416,000 tonnes (t)
- Imports (i): 499,100 t
- Exports (e): 26,428,400 t
- Domestic consumption (c = p + i - e): 25,486,700 t
- Ecofootprint of grain consumption by Canadians \((a = c \times y^{-1})\): 9,288,200 ha arable land
- Per capita ecofootprint attributable to grain consumption \((f = a \times N^{-1})\): .32 ha arable land.
footprint one wishes to account for food-related nutrient assimilation but finds that the city's domestic wastes are composted and spread on adjacent agricultural or commercial forest land, then only the agricultural and forest products footprint components should be compiled in the total footprint analysis. Similarly, some consumer products such as leather goods are the by-product of another industry (in this case, beef production). In such cases, one would generally count only the primary land requirements (the grazing and grain lands required for feeding cattle). (For some purposes, of course, it may be of interest to apportion the land among the various products associated with the primary use.)

To estimate typical population ecological footprints (e.g., for whole regions or countries), we usually start with world average land productivities or yields. This greatly simplifies calculations because we do not have to trace all the sources of trade goods and waste sinks or determine the productivity and assimilative capacities of the corresponding production/assimilation areas. As important, using a common base yield facilitates comparison among countries and comparisons of individual countries with global totals. This shortcut is also becoming increasingly realistic for many countries that are heavily dependent on trade flows of commodities from various sources. (Note that the global ecofootprint estimate produced by this method would be the same as the global total estimate that would result if actual yields were used for all items and countries.)

For some kinds of analyses, of course, it might be necessary or useful to base the ecofootprint calculation on actual land/water yields where sources are known and data are available. Also, to compare a country's or region's ecofootprint with the productive or load-bearing capacity of its own domestic territory, we must know the actual productivity of various domestic land categories (cropland, pasture land, forests, carbon sinks, etc.) in order to calculate that country's ecological deficit (or surplus).

For example, if a given country's cropland is three times as productive as world average cropland, we would multiply the area of that country's cropland by a yield factor of three before comparing it to (i.e., subtracting it from) the calculated cropland footprint, as each domestic hectare is the yield equivalent of three global average hectares. This gives a more accurate estimate of the de facto ecological deficit (or surplus) in cropland. Note, however, that if the higher productivity of the domestic land is the result of intensive fertilizer and pesticide use, the apparent reduction in the ecological deficit due to higher cropland productivity will be at least partially canceled out by the increase in the ecological footprint attributable to the use of agricultural chemicals (fossil fuels are an important feedstock in the manufacture of these inputs).

In general, we err on the side of caution in making ecofootprint estimates. For example, if there is dispute over, or several estimates of, land productivity, we use the higher estimate (reduces footprint size). Most ecological footprint calculations are therefore likely to under- rather than overestimate.

IV. HUMAN ECOLOGICAL FOOTPRINTS: IMPLICATIONS FOR GLOBAL SUSTAINABILITY

So, how big are our ecological footprints? Most citizens of industrialized high-income countries feel so distanced from nature that it has simply never occurred to them to ask, "How much of the earth's surface is dedicated to supporting just me in the style to which I have become accustomed?" They are generally shocked to learn that average residents of North America, Europe, Japan, and Australia require the biophysical output of 5 to 10 hectares (12-25 acres) of biophysically productive land and water each to support their consumer lifestyles (Rees, 1996; Wackernagel and Rees, 1996; Wackernagel et al., 1999). Such large ecofootprints are also exerted by members of wealthy urban elites throughout the developing world.

A. The Functional Footprints of Cities

These findings should alter our perceptions about many things. To begin, they should change how we think about cities and urban land. Almost 30 years ago, American ecologist Eugene Odum wrote, "Great cities are planned and grew without any regard for the fact that they are parasites on the countryside which must somehow supply food, water, air, and degrade huge quantities of waste" (Odum, 1971). Ecofootprinting enables us to quantify the extent of this urban "parasitism." For example, in 1996 Canada's largest city, metropolitan Toronto, had a population of approximately 2,385,000 living in an area of 630 km². Assuming they are typical Canadians, Toronto's citizens had an average ecological footprint of about 7.7 hectares (19 acres). Thus the ecological footprint of metro Toronto was 183,650 km², or about 290 times the size of its then political area.
Most of the city's supportive ecosystems are located at great distance from the people they sustain; indeed, they are scattered all over the planet.

This situation is characteristic of high-income cities. In a particularly comprehensive analysis, Folke et al. (1997) estimated that the 29 largest cities of Baluc Europe appropriate for their resource consumption and waste assimilation an area of forest, agricultural, marine, and wetland ecosystems 565 to 1130 times larger than the areas of the cities themselves. A study for the International Institute for Economy and Development in London shows that the biophysical demands of that city alone appropriate an area scattered around the world equivalent to all the ecologically productive land in the United Kingdom.

Such findings have important implications for both urban development and rural sustainability in the 21st century. Some people interpret the global migration of people from the countryside to the city as implying that modern humans are abandoning the countryside and becoming less dependent on the land. This is illusion. The reality is that productive croplands, pasture lands, and forests everywhere are being used more intensely than ever to sustain the world's burgeoning urban populations. The human occupants of cities may think of the latter as their principal habitat, but cities per se represent only a tiny fraction of the total human ecosystem.

There is, of course, a certain mutualism between the city and the countryside. Cities need the resources of rural areas and rural areas benefit from urban markets and technology transfers from cities. However, while rural areas could survive without cities, the dependence of cities on rural environments is absolute. In short, there can be no urban sustainability without rural sustainability. In this light, we might even want to reconsider what we think of as urban-dominated land—in a whole-systems ecological sense, the great plains of North America, one of the world's major breadbaskets, are an essential component of the increasingly urban global human ecosystem.

According to United Nations' projections, there will be 27 cities with populations exceeding 10 million by 2015 (up from only 1 in 1950). Forty-four more cities will have populations of 5 to 10 million by the same year. This should trigger at least a cautionary alarm. The world's megacities—all cities for that matter—are dependent on a vastly larger area of productive lands outside their boundaries and political control. But just how secure can any urban population be if the lands represented by its ecological footprint are under threat from ecological change or geopolitical uncertainty?

B. Ecofootprints, Global Development, and Social Equity

It is not just cities that overshoot the productive capacity of ecosystems within their political domain. Most high-income countries have an ecological footprint several times larger than their national territories. In effect, they are running massive ecological deficits with the rest of the world. As a result of continuous population growth, rising material demand, and expanding trade, wealthy consumers in particular have ecological footprints that treat the entire planet as their common patch.

In fact, several ecofootprint studies suggest that humanity is now running a global ecological deficit of up to 30% and is accumulating a nonpayable ecological debt. Even at current average levels of economic production and consumption, the human load already exceeds the long-term carrying capacity of the earth. The direct empirical evidence (e.g., ozone depletion, climate change, fisheries collapse) is becoming the stuff of daily newspaper headlines.

This means that contrary to the assumptions of prevailing international development models, so-called first-world material lifestyles are not sustainably extendible to the entire world population using prevailing technology. Although wealthy consumers require the goods and life support services of 5 to 10 hectares of productive land and water there are only about two hectares of such land per person on earth. Meanwhile, the great majority of humankind is understandably not satisfied with its present material lot and is determined to improve it. (Despite the recent industrial revolution in China, the ecological footprint of the average Chinese is less than 1.5 hectares and Bangladeshis get by on only half an hectare.) The problem is, to support all six billion members of the human family at the average North American material standard would require about three additional earthlike planets (and this does not account for the needs of the three billion more people to come by the middle of the 21st century).

If we are at or beyond the human load-bearing capacity of the planet, then the only way additional population or material growth can be sustained, without ravaging biodiversity and ultimately destroying the ecological basis of human life, is through massive cutbacks in resource consumption by the presently wealthy. In theory this could be achieved through a tightly managed new efficiency revolution (doing much more with less) in combination with an absolute reduction in material demand (i.e., a shift to simpler lifestyles).

Other carrying capacity studies have reached similar
conclusions. Using data on current waste discharge rates and the assimilative capacity of biophysical systems, the Sustainable Europe Campaign has estimated that sustainability requires an approximately 50% reduction in the material throughput of the global economy in coming decades. However, in order to create the necessary environmental space for developing countries to claim their fair share of global carrying capacity, developed countries must reduce their consumption of various nonrenewable resources by between 88 and 94% (Carley and Spapens, 1999).

This might seem to be an extreme conclusion and impossible goal. However, the United Nations Environment Program's “Global Environment Outlook 2000” report also argues the need for a tenfold reduction in resource consumption by high-income countries. Even the Geneva-based world Business Council for Sustainable Development acknowledged as early as 1993 that “industrialized world reductions in material throughput, energy use, and environmental degradation of over 90% will be required by 2040 to meet the needs of a growing world population fairly within the planet's ecological means” (BCSD, 1993).

If so-called postindustrial society fails to meet this challenge, the anticipated doubling and doubling again of the human material demand in the next half-century will be disastrous for conservation efforts. As explained next, the expanding human ecological footprint inevitably reduces the ecological space available for other species.

V. HUMAN ECOLOGICAL FOOTPRINTS AND THE BIODIVERSITY CRISIS

Human beings are uniquely successful among large mammals in mastering the variety of environments on earth and in expanding both numerically and spatially over the globe. Our capacity to extend continuously our ecological footprint is attributable to several species-specific qualities of which perhaps three stand out.

First, humans have a remarkably variable diet—we have wide-ranging omnivorous tastes and if we cannot consume something directly (such as grass), we domesticate an animal that will and then eat the animal. This means that humans may well possess the broadest food niche of any vertebrate on the planet. Second, humans are as behaviorally adaptable as they are catholic in their diets. Together, these two factors make virtually any terrestrial ecosystem on earth accessible to Homo sapiens. No other species has managed successfully to colonize and dominate virtually all major terrestrial ecosystem types, from grasslands and forests to deserts and tundra, on all the world’s significant land masses. Finally, we are creatures of language, culture, and cumulative learning. Technological advance has enabled humans continuously to increase the intensity of resource exploitation and to extend their ecological footprint into virtually all remaining productive habitats on the planet. For example, thanks to electronic fish-finding devices and ruthlessly effective fishing gear, humans have actually become the functionally dominant marine mammal. By the mid-1990s, the world's fishers were appropriating the equivalent of 25 to 35% of the product of net photosynthesis from the 10% of the oceans that produces 96% of the catchable fish.

We previously noted that human ascendance has historically been achieved at the expense of other species. People's material demands, culminating in today's global consumer society, have steadily appropriated an increasingly disproportionate share of the planet's finite productive and sink capacities. The recent human dominance of ocean ecosystems provides several examples of the resulting biodiversity loss.

To begin, some modern fish-harvesting techniques have devastating impacts on the physical and biological structure of marine habitats, thereby reducing the productivity, diversity, and abundance of both target and no-target species. Various surveys reveal that fishers now drag heavy trawls over all the oceans' continental shelves at least once every two years and some areas are hit several times in a season. In heavily fished regions of the North Sea, trawlers make up to seven passes annually with nets and chains weighing up to five tons. The total area affected by trawls is 150 times larger than the area of forest clear cut globally each year.

Many fish-stock depletions and the destruction of benthic communities are the direct result of the increasing intensity of commercial fishing. However, other systemic problems stem from the fact that expanding the human marine ecofootprint inevitably diverts an ever greater proportion of the finite energy budget of the sea away from other marine predators, including various fish, birds, and mammals.

A particularly telling example of the resultant domino effect can be traced to a tripling of the commercial catch of Alaska pollock since 1986. The reduced availability of pollock apparently precipitated a decline in populations of Steller's sea lions and seals, which feed on pollock. By the late 1990s, sea lion populations in the Gulf of Alaska had fallen by 80 to 90%, in turn depriving local orca (killer whale) populations of their normal primary food source. With the collapse of sea
lions and seals, the orcas turned to eating sea otters—smaller, less fatty animals—whose populations have also now also fallen by 80 to 90% since 1990. But the systemic shifts and biodiversity losses do not end there. Sea otters seem to be a keystone species in the shallow marine kelp forests that constitute their preferred habitat. Among the otters’ major prey are sea urchins, which, in turn, graze on kelp. As the otter population declined there was an explosion of sea urchins, leading to the destruction of the kelp beds by overgrazing. So it is that human overfishing for pollack precipitated the collapse of one of the most productive inshore-marine ecosystems in the North Pacific, as so-called sea urchin barrens spread along the coast of Alaska and the adjacent Aleutian Island chain. (Dramatic as this example might seem in its own right, it is only just the final chapter of a longer book. Direct human overharvesting of sea otters for their pelts, beginning in the 19th century, had already long since decimated the kelp forests, which previously characterized much of the Pacific coast of North America from California north.)

Some popular accounts of this phenomenon have emphasized the role of orcas in switching prey from seals to otters. Certainly this was an important link in the chain of events, but the structural change mainly responsible for precipitating the collapse of the near-shore ecosystem was the displacement by humans of sea lions, seals, and ultimately orcas from their ecological niches among the top carnivores in the offshore ecosystem. As humans replaced seals in the food-web, the killer whales were forced to switch to sea otters, a much inferior food source.

This example underscores how, whether conscious of it or not, Homo sapiens have become the ecologically dominant marine mammal. But the main point to take away is that uncontrolled increases in human appropriations of bioproducitivity inevitably result in the rerouting of energy and material flows through the affected ecosystems—marine or terrestrial—with consequent and potentially permanent changes in ecosystem structure and function.

We now recognize that several mechanisms are at work as humans extend their ecological footprints. People

1. Passively displace other species from their food niches or appropriate their habitats. (Commercial fishing displaced sea lions, seals, and orcas on the northwest Alaska coast; agriculture pushed bison from the Great Plains of North America; clearing for crops and grazing has destroyed thousands of species in former tropical forests.)

2. Actively eliminate nonhuman competitors—other species that compete with us for “our” food. (We shoot wolves that hunt ungulates, and seals that eat commercially valuable fish; we poison insects that would devour our crops.)

3. Deplete both self-producing and nonrenewable “natural capital” stocks. (Humans overexploit many wild prey populations from hios to fish; destroy whole ecosystems such as forests; and deplete vital assets such as groundwater, soils, and fossil fuels.)

These processes are all consumption-related. The first two are forms of competitive exclusion. Technological “man” is simply more effective than other organisms at appropriating nature’s bounty for his or her own use. Because energy and material flows used by people are irreversibly unavailable for other species, the latter decline, even to extinction, at least locally.

The third mechanism, stock depletion, is the product of many things, including confidence in technological substitution, blind ignorance, material greed, sheer desperation, and the relentless working of the so-called common property problem on an overcrowded planet. Sometimes it is even the result of willful disregard on the part of those who give no moral standing to other creatures or who simply don’t care about the state or fate of the world.

Economic globalization, the sanctioning of greed, the rise of consumerism, and the spread of energy-intensive technologies have intensified all these processes. Overharvesting and habitat destruction are driving what some analysts refer to as the “sixth extinction.” According to Biologist E. O. Wilson, the earth is now experiencing the greatest extinction episode since the natural catastrophes at the end of the Paleozoic and Mesozoic eras. The current species extinction rate is 100 to 1000 times prehuman levels inferred from the fossil and paleontological record. A quarter of bird species have already been extinguished by people and fully a quarter of the 4400 mammal species living today are on a path of decline that, if not reversed, is also likely to end in extinction.

The contemporary human ecofootprint is also characterized by massive increases in waste production. Not surprisingly, the resultant pollution is adding to the toll on biodiversity. A poignant example is the phenomenon of “coral bleaching,” which has been in the news since the mid-1980s. This effect is produced when the animal component of coral organisms—corals are a
The mutualistic union of animal polyps and plant (algae) cells—expel their algal cohabitants as a result of thermal or other stress. The loss of algae leaves the coral pale and washed out (bleached) and unable to grow or reproduce. In the exceptionally warm spring and summer of 1998, coral bleaching associated with a one Celsius degree increase in ocean temperature extended across the tropics, including, for the first time, the Indian Ocean. Reports indicate that 70 to 90% of corals in the Indian Ocean basin appear to have been affected.

Some scientists have referred to the bleaching and die-back of corals as an unprecedented ecological disaster. Coral reefs are among the most diverse and productive ecosystems on the planet. Numerous fisheries, particularly locally important food fisheries, are sustained by the productivity of corals. What economic and social costs would be incurred if the decline of corals and associated reef habitats continues unabated? How will we measure the ultimate consequences of the sheer loss of biodiversity? Most important, what does the apparent spread of the problem say about the state of the earth in general and the marine environment in particular?

These are important questions, but the main point is that to the extent that coral bleaching is caused by human-induced global warming, it is the indirect result of greenhouse gas accumulation in the atmosphere. (The earth’s carbon sinks are overwhelming.) The attendant biodiversity loss is therefore the distal product of the industrial world’s addictive dependence on fossil fuels and our ever-expanding industrial metabolism. Regrettably, if the climate change/global warming hypothesis proves to be correct, coral bleaching will not soon be reversed by any political action feasible under the prevailing global development model.

All these examples underscore the inverse relationship between growth of the human ecological footprint and biodiversity (the competitive exclusion principle at work). Unfortunately, at present, ecofootprint analysis does not directly track biodiversity losses. However, the negative correlation is most directly associated with the conversion of wildlands to agriculture, intensive (including plantation) forestry, and various forms of habitat fragmentation. Many of these conversions are measurable. It may therefore eventually be possible, using the historical record and principles of biogeography, to link some index of biodiversity loss to the growth of population ecofootprints.

Whether or not this relationship can be rigorously defined, the basic point stands. To the extent that further human population and material growth requires the clearing of forests and the intensification of agriculture, and is accompanied by urban sprawl and pollution, it necessarily means less ecological space to sustain nonhuman species and ecosystems. Biodiversity loss is thus virtually certain to accelerate under the prevailing global development scenario.

VI. ASSESSING THE ECOLOGICAL FOOTPRINT APPROACH

A. Conceptual and Methodological Strengths

Ecofootprint analysis brings several methodological strengths to bear in assessing the state of the world. First, it is compatible with various concepts developed by other analysts to address human ecological problems. For example, the ecological footprint concept

- incorporates and extends George Borgstrom’s 1960s notion of “ghost acreage,” referring to the extra-territorial food lands required to support densely populated regions and countries.
- corresponds closely to Paul Ehrlich’s and John Holdren’s 1970s definition of human impact on the environment: I = PAT, where I is impact, P is population, A is affluence, and T is technology. The population ecological footprint (E_P) corresponds to impact (I) in the latter formulation and is itself a function of population size and consumption. However, because consumption is a function of income (affluence) and the state of technology, I, like E_P, is a function of population and consumption. From this perspective, E_P is an area-based analogue of I.
- provides one measure of human load defined by Catton (1980) as a function of population size and average individual impact (consumption again). The larger the footprint, the greater the load. (Catton defined carrying capacity as the “maximum persistently sustainable load.”)
- is conceptually related to the embodied energy (energy) analyses of ecologist Howard Odum and recognizes the importance of the second law of thermodynamics to human affairs as stressed by ecological economists Nicholas Georgescu-Roegen and Herman Daly. Indeed, the area represented by the ecological footprint can be conceived as the photosynthetic surface (solar collector) needed to replace the free energy (essergy or negentropy) dissipated by humans and their industrial metabolism.
However, for all its technical compatibility, perhaps the major strength of the eco-footprint concept is conceptual simplicity and intuitive appeal. First, the ‘footprint’ metaphor seems particularly effective in communicating the idea that we each have an impact on the Earth for which we are responsible through consumption choices. Second, ecofootprinting consolidates real data on a variety of energy and material flows into a single concrete variable, land area. Land is a particularly powerful indicator because it too is readily understood by ordinary people.

Because the supply of land is finite at any scale, ecofootprint analysis provides a measure of human demand (load) for comparison with supply (load bearing capacity) at all geographic levels from local to global. Such comparisons suggest several secondary indices of (un)sustainability that can be used to stimulate discussion of sustainability options or in establishing measurable policy targets. For example, if we are running a local or regional ecological deficit, the question becomes: Is this a major problem and, if so, what must be done to reduce it? (Note that humanity’s total ecological deficit, the sustainability gap, may ultimately be more important than any fiscal deficit yet is totally ignored in most national accounting systems and international developmental planning.)

To be sure, it is unlikely ecofootprint analysis will ever cope satisfactorily with various toxic waste discharges into the ambient environment. The biological effects of chronic low-level chemical contamination (e.g., carcinogens), for example, do not readily translate into an appropriated land area. However, this may not be as problematic as it seems. Environmental and ecological economists already agree that society should have zero tolerance for highly toxic chemical wastes and radioactive substances for which the ecosphere has no measurable assimilative capacity. Such substances should simply be banned or phased out (as in the case of chlorinated pesticides and ozone-depleting CFCs).

In any event, no single index can be expected to represent all ecological impacts and this limitation does not invalidate what ecofootprinting does reveal. Rather, it suggests that ecofootprint calculations are almost certainly underestimates of actual ecosystem appropriations and that methodological improvements and extensions may well result in considerably larger ecofootprints. As it is, ecofootprinting already provides unambiguous results and suggests clear policy directions. Increasing the accuracy and scope of the analyses may add to our sense of urgency but will not likely shift the direction of needed policy change.

2. Is Ecofootprinting Too Simplistic?
Some commentators have argued that the ecofootprint concept oversimplifies both nature and society and has little predictive value. Both the ecosphere and the economy are dynamic systems, but ecofootprint analysis produces static estimates of a single aggregate variable.

Again, it is true that footprint analysis is not dynamic modeling and cannot directly project the effects of technological change or social adaptation. However, simulation and prediction were never the intent. Ecofootprinting provides a simple ecological camera—each analysis is a snapshot of current demands on nature, a portrait of how things stand right now under prevailing technology and social values.

This in itself is important information. Current “snapshots” indicate that humanity has already significantly exceeded carrying capacity using prevailing technology and that some people contribute significantly more to this ecological overshoot than do others. This is sufficient to suggest approximately how much we must reduce our consumption, improve our technology, redistribute wealth, or change our behavior to achieve sustainability.

Of course, once an initial baseline has been established, subsequent ecofootprint assessments can contribute to a time-series study—repeated snapshots over
years or decades—to monitor progress toward, for example, reducing national ecological deficits or closing the global sustainability gap. (After all, even a motion picture is a series of snapshots.) Such sequential analyses are fully capable of accounting for the actual effects of new technologies, or changes in consumer behavior or cultural values. We can also assess the potential impacts of alternative technologies or settlement patterns on the size of a population’s ecological footprint in static studies. This involves using assumed levels of consumption or assumed technologies (based on real-world data) in the standard calculations to address such questions as, how would a partial shift from fossil fuels to biomass fuels or solar energy affect a population’s ecofootprint?

To summarize this point, ecological footprint analysis is not intended to provide a dynamic window on the future but rather a snapshot in time. As such it can both help to assess current reality and to test alternative “what-if” scenarios on the road to sustainability.

On a related matter, some critics have stressed the impossibility of mapping the state of complex multidimensional processes or systems with a single indicator and that ecofootprint analysis ignores the implications of complexity theory. This comment, while true, misses two points. First, ecofootprinting is not intended to provide a complete picture of any complex system. It generates only one indicator of the state of humanity’s engagement with the rest of nature. To the extent that this index is reliable and has policy relevance to sustainability, it should be used in conjunction with economic, social, or any other indicators that bear on the issues at hand.

Second, although science has certainly gained important insights from complex systems theory—the behavior of dynamical, self-organizing, adaptive systems often confounds standard forecasts and details simplistic approaches to ecosystems management—the fact remains that many phenomena do have simple explanations. In these circumstances, Occam’s razor still applies: there is no need to derive a complex explanation where a simple one suffices.

For example, one can hardly imagine a better example of a complex human-dominated system than the world’s great cities. Cities are a marvel of biophysical, economic, political, social, and cultural subsystems that somehow come together to function as a complex, dynamic, integrated, self-organizing whole. Nevertheless, if any city as currently conceived were enclosed in an impermeable glass dome whose base coincided with its political or geographic boundaries, it would rapidly disintegrate.

The explanation is simple—the city would be starved of resources and suffocated on its own wastes at the same time. This is because most of the ecosystems and biophysical processes supporting the city lie outside our hypothetical bell jar. Ecofootprint analysis measures the city’s (region’s, country’s, etc.) dependence on nature’s services and estimates the ecosystem area needed to produce them. There is certainly room to debate the methods used to make these estimates. However, we don’t need to know anything about city governance or the complex relationships among its physical and socioeconomic subsystems to predict the collapse of our city-in-a-jar or to perform an ecofootprint analysis and assess its implications.

3. The Carbon Assimilation (Fossil Energy) Footprint Is Not a Real Land Use

For many high-income industrialized countries, the fossil energy component may make up half of the total ecofootprint as presently calculated. Because science cannot account for the global carbon budget and there is dispute over the relative roles of terrestrial and marine ecosystems in carbon sequestration, some critics see energy footprint calculations as unreliable or hypothetical. Others don’t regard carbon assimilation as a real land use.

This last criticism reflects our cultural bias that land use requires physical occupancy or the consumption of some tangible product of the land. We are accustomed to thinking of the provision of other (particularly nonmarket) biophysical services of nature, no matter how essential to sustainability, as effective “uses” of landscape. In terms of thermodynamic law and mass balance, it is this cultural bias that needs adjustment, not ecofootprint theory. All energy and material use is “throughput,” implying the generation of wastes, and to the extent that the assimilation and recycling of those wastes (including carbon dioxide) requires an exclusive dedication of land or water, it is legitimate component of ecofootprint calculations.

Returning to the first criticism of energy footprinting, while the details of the carbon budget may be in dispute, there is no question that carbon dioxide levels are increasing in the atmosphere and that these increases represent about half of current carbon emissions from fossil fuel and biomass combustion. This implies that available land and water carbon sinks are insufficient to sequester all anthropogenic carbon dioxide at current rates of emission. In ecofootprint terms, our current global energy footprint is
excessive. We are running a global carbon sink deficit in exactly the same sense that many countries run a food-land deficit.

In short, the notion of carbon sink land is not hypothetical even if the real thing is in short supply. Energy footprint calculations are therefore designed to estimate the area of dedicated carbon-sink forests that would be necessary to assimilate a study population's carbon emissions based on the average estimated carbon sequestration rates of the world's growing forests. In fact, various electric utilities are now planting dedicated carbon sink forests to offset the carbon emissions of their fossil fuel-burning generating plants. Similarly, some countries are contemplating developing carbon sink forests as an alternative to reducing their carbon emissions and to be part of a future system of tradable carbon emission rights. (For example, planting a carbon sink forest would release an equivalent quantity of emission rights for sale on the market.) Note that a dedicated carbon sink forest is an exclusive land use insofar as it cannot be harvested for any purpose such as pulp and paper production that would quickly release sequestered carbon back into the atmosphere.

Finally, we should recognize that any alternative to fossil fuels will also generate an energy ecofootprint. For example, the ethanol equivalent of fossil fuel, generated from biomass, would require a fuel-crop growing area considerably larger than the energy footprint based on carbon sinks. Even solar-based alternatives require that a large land area be appropriated for collector surfaces (though this land need not be biologically productive). The point is that all energy sources and related sinks have a corresponding land equivalent and the present method of calculating an energy footprint may actually be fairly conservative.

VII. EPILOGUE

Whatever its weaknesses, the ecological footprint concept is strong enough to have captured both the scientific and popular imaginations. It is now being applied as an assessment tool by nongovernmental organizations and government agencies in towns, cities, and countries around the world. A small industry of consulting firms has sprung up to develop and commercialize various products based on the ecofootprint concept.

Ecofootprinting seems to concretize the human ecological crisis in a way other energy and material flows studies do not. Knowing our ecological footprint not only makes us conscious of our personal contribution to global change, but also makes us responsible for doing something about it. Pluralistic democracies can work only if the public are well informed about key policy issues and have a keen sense of their capacity to make a difference in improving the human prospect and those of other species. To the extent that ecological footprint analysis helps to generate a more ecologically literate and politically active populace, it will have achieved its most important objective.

See Also the Following Articles

CARRYING CAPACITY, CONCEPT OF • ENERGY FLOW AND ECO SYSTEMS • ENERGY USE, HUMAN • HUMAN IMPACT ON BIODIVERSITY, OVERVIEW • LOSS OF BIODIVERSITY, OVERVIEW • SUSTAINABILITY, CONCEPT AND PRACTICE OF • URBAN/SUBURBAN ECOLOGY

Bibliography


ECOLOGICAL GENETICS

Beate Nürnberg
Ludwig-Maximilians-Universität München

1. Introduction
2. Genetic Variation in Ecological Traits
3. Ecological Variants in Spatially Structured Populations
4. Inferences from the Spatial Distribution of Neutral Markers
5. The Genetic Basis of Ecological Variants
6. Limitations to the Evolution of New Ecological Variants: Tradeoffs
7. Ecological Genetics of Species Interactions

GLOSSARY

allozyme Enzyme locus that produces more than one electrophoretic variant.
deme Smallest population unit in population genetic models.
frequency dependence Phenomenon that the dynamics of a given ecological or genetic type depend on its frequency in a population.
gene marker Genetically variable locus that produces distinct variants when analyzed by standard methods such as electrophoresis. It may be based on protein or DNA and may or may not be selectively neutral.
metapopulation Collection of populations that are connected through dispersal and are typically extinction-prone. A key feature is the distinction between local (transient) and regional (persistent) dynamics in one or several interacting species.

phenotypic plasticity Property of a given genotype to express different phenotypes as a function of ecological conditions.
polymorphism Existence of more than one discrete intraspecific type such as distinct phenotypes or genotypes at one locus or several loci.
population structure Subdivision of a given population according to features such as location, age, size, or social status.

ECOLOGICAL GENETICS is an interdisciplinary field of study with the aim to understand adaptation in response to natural selection at the intraspecific level. It applies the principles of neo-Darwinian theory to directly observable microevolutionary processes.

I. INTRODUCTION

Broadly speaking, there are two main objectives at the heart of ecological genetics research: (a) the analysis of the nature and strength of specific selection pressures within a given ecological setting and (b) the genetic basis of the adaptive response. Consequently, ecological genetics draws heavily on two bodies of theory. On the one hand, ecological theory yields predictions about the kinds of adaptations that a given selection pressures might produce. On the other hand, the theory of population genetics provides specific models about allele
frequency changes under different selective regimes and in a variety of population structures.

Hutchinson (1957) succinctly captured the focus of this research program in his famous metaphor of the ecological theater in which the evolutionary play is enacted. In this tradition, E. B. Ford (1964) published the first treatise on ecological genetics. Weary about the possible artifacts of evolutionary experiments that are conducted in the laboratory only, he advocated the study of natural populations as an indispensable complement. To him natural populations provided the ultimately relevant setting in which to investigate natural selection and adaptation. Within the past 30 years, the scope of ecological genetics has become even broader thanks to both theoretical and methodological advances.

The evolutionary analysis of genetically complex phenotypic traits rests on the theory of quantitative genetics. The incorporation of this theory into the conceptual framework of ecological genetics has greatly enlarged the range of phenotypic traits under study. Whereas early studies had mostly focused on the dynamics of visible polymorphisms with a relatively simple genetic basis, adaptations in continuously varying characters such as body size, disease resistance, or courtship coloration have since become amenable to ecological genetic analysis. Moreover, the development of a whole spectrum of genetic markers with varying evolutionary dynamics has provided tools for the study of population structure as well as for the spread of selectively favored genes. More recently, genetic markers have been used for the genetic dissection of complex phenotypic traits.

The ecological setting no longer functions as a mere backdrop to the evolutionary process, as Ford saw it, but provides a research focus in its own right. Ecological theory makes testable predictions about the kinds of strategies that are expected to evolve under a given set of environmental conditions. For example, life history theory defines the optimal schedules of growth, reproduction, and intrinsic mortality. Similarly, behavioral ecology makes predictions about such features as optimal breeding systems or diet breadth. For a broad range of traits, ecological theory establishes the link between a given adapted phenotype and the particular selective agent that prompted its formation.

Ubiquitous environmental heterogeneity produces phenotypic differentiation on a variety of spatial scales across a species range. For any one species, there might be broad differentiation with respect to one trait and at the same time a patchwork of local variants in another. Empirical studies have indeed demonstrated a large amount of ecological variation at the intraspecific level, which covers the range from life history traits over behavior and physiology to morphology. Quite often, these differences are hidden from the casual observer and become apparent only through detailed study that extends over more than one generation of the organism in question. Superficial morphological uniformity over large areas therefore does not imply ecological equivalence. The phenotypic changes may be small, and they need not accumulate to change the appearance of the species in the long term. In fact, they might be partly or entirely reversed when the local conditions change. Yet at any one time they can determine the ecological role that the organisms of a particular population play.

I will use the term 'ecological variants' to emphasize the context-dependence of these phenotypes. Ecological genetics is beginning to tell us not just about the effects of complex ecological settings on the process of evolution, but also about the role of genetic variants in shaping local ecological interactions.

II. GENETIC VARIATION IN ECOLOGICAL TRAITS

A. Genetic Variation within and between Populations

All phenotypic evolution depends on within-population genetic variance. It is this variance that natural selection can mold to produce either distinct locally coexisting variants or differentiation among populations. There is abundant evidence that this substrate of evolution exists for many phenotypic traits. There is genetic variation within populations of the swallowtail butterflies *Papilio <i>zeltocam</i> and *P. <i>oregus</i> for oviposition preference on a number of host plants. Individuals within two populations of the wood frog, *Rana <i>sylvatica</i>, differ genetically in their larval development time. Aphids collected from a single field vary in their resistance to a parasitoid wasp, and among the wasps collected from the same field there is genetic variation in their ability to parasitize the local aphids. The sand cricket, *Gryllus <i>firmus</i>*, shows large within population variation for the duration of the male mating call. Several different genetic types that confer resistance to a fungal pathogen exist within populations of the Australian flax, *Linum <i>marginala</i>*. This list could be extended to include many more examples, all of which underscore the fact that the raw material for evolution exists for a broad range of ecological traits.
The extent to which this variation has been transformed into genetic variation between populations is equally impressive. Garter snakes (*Thamnophis elegans*) in coastal populations of California where slugs are common have a much greater innate preference for eating slugs than those from populations further inland. Female guppies vary in their preference for male color patterns even among populations within the same stream. Sticklebacks (*Gasterosteus*) differ geographically in traits such as courtship behavior and cannibalistic tendency. One population of the milkweed bug, *Oncopeltus fasciatus*, shows strong migratory behavior whereas another more southerly population is nonmigratory. Populations of the spider *Agelenopsis aperia* have diverged genetically in number of traits including territorial behavior, predator avoidance, and diet breadth. The wood frog populations just mentioned differ from each other in larval development time and size at metamorphosis.

### B. Is the Observed Phenotypic Variation Heritable?

In all of the foregoing examples, the genetic basis of the observed variation has been demonstrated through careful experimentation. The easiest way to demonstrate between population genetic differences is to rear field-collected eggs or seeds from different localities under uniform conditions. Differences between the population means in the trait of interest suggest a genetic component. However, the effect could be purely environmental. More resources in one location could lead to better provisioned offspring that reach a larger body size, develop more rapidly, or have greater disease resistance. Such nongenetic effects can be eliminated or at least minimized if one rears two generations in the laboratory and measures the trait of interest only on the second generation.

Controlled breeding schemes are used in order to determine the within-population genetic variation in sexually reproducing species. If the trait in question has a simple genetic basis and gives rise to a small number of distinct phenotypes, it might be possible to infer the number of loci involved from the Mendelian segregations within crosses. However, such a detailed genetic dissection is impossible in the majority of cases in which phenotypic variation is continuous and a function of both polygenic inheritance and environmental influences. In that case, the resemblance between parents and offspring can be used to estimate the so-called heritability of the trait. Heritability is defined as the proportion of phenotypic variance that is due to heritable genetic variation (see Box 1). Care needs to be taken when one extrapolates from laboratory estimates to the situation in the field. Because heritability is a proportion of total phenotypic variance, *Vp*, it can be influenced by changes in any of the components of *Vp*. For example, uniform rearing conditions reduce the environmental

---

### Box 1: Heritability

Quantitative genetic theory is built on a model in which a given phenotype is a combination of genetic factors at a large number of loci, each of small effect, and a random component that reflects the environmental influence. There are several types of genetic effects. An additive component represents the sum of allelic effects. Per locus and allele, the allelic effect can be thought of as the average phenotypic effect that the allele would have if it were placed into a large number of genetic backgrounds within the population. In addition, there may be effects due to dominance. If the joint effect of two alleles in a heterozygote is not equal to the sum of the allelic effects, then the difference between additive and observed effect gives the dominance deviation. Similarly, if the effect of an allele depends on which alleles are present at one or more other locus, one talks of epistasis or of an interaction effect.

The phenotypic variance *Vp* in the population can be broken up into components in an analogous manner:

\[
V_p = V_A + V_D + V_I + V_E,
\]

where *V_A*, *V_D*, *V_I*, and *V_E* are the additive, dominance, interaction, and environmental components, respectively. Because sexually reproducing individuals pass on genes but not genotypes to their offspring, all nonadditive effects that result from particular gene combinations in the parent are not heritable. The resemblance between parents and offspring depends only the additive genetic variance. It is this resemblance that determines the evolutionary dynamics of a trait. Hence, heritability, *h^2*, is defined as the proportion of phenotypic variance that is due to additive genetic variance in the population:

\[
h^2 = V_A/V_p.
\]
variance component and therefore cause an increase in heritability. Nevertheless, laboratory estimates generally give a good indication whether or not the observed variation has a genetic basis.

An alternative method to demonstrate genetic variation within population in a given trait is to perform an artificial selection experiment. If the trait responds to selection, then the base population must have harbored genetic variation at the underlying loci. Experiments of this type are generally successful. Genetic variation for phenotypic traits appears to be ubiquitous, yet not all of it is necessarily available for adaptive change in nature (discussed later in this chapter).

The famous long-term studies of Darwin's finches on the Galapagos Islands have provided not just field-based estimates of heritabilities but also proof of short-term phenotypic change in response to natural selection. For example, the medium ground finch Geospiza fortis on the island Daphne Major shows heritable variation in body weight, wing length, bill size, and tarsus length. These variables are largely correlated, because they are all components of body size. After a period of exceptionally heavy rains, small soft seeds were abundant for the following two years. This food supply differed markedly from the predominantly large and hard seeds of the preceding period of drought. After the rains, the finch size distribution shifted within two years toward smaller birds that showed relatively better survival. Moreover, there was evidence of direct selection in favor of reduced bill width over and above the overall selection on body size.

C. The Alternative: Phenotypic Plasticity

Although genetic differentiation in ecological traits is common, by far not all phenotypic differences between populations need to have a genetic basis, even if they are obviously adaptive. In fact, a surprising amount of intraspecific phenotypic variation lacks a genetic basis. Such variation is referred to as phenotypic plasticity if a given genotype predictably produces different phenotypes under different environmental conditions. This environment-dependent phenotypic repertoire of a genotype is called its norm of reaction. Phenotypic plasticity is thus distinct from random environmental variance due to chance effects that impinge on the expression of traits within a given environment.

In the Texan desert, tadpoles of the spadefoot toad Scaphiopus couchii adjust their development time to variation in pond duration. They are able to accelerate their development in rapidly drying ponds and thus increase their chances to evade death from dessication. However, this apparently adaptive behavior comes at the price of a smaller size at metamorphosis and less well-developed hind limbs, factors which are likely to reduce juvenile survival. Nevertheless, this flexible larval strategy appears adaptive in an environment in which gravid females are seemingly unable to predict pond duration. A large group of examples of phenotypic plasticity involves predator avoidance. The snail Physella virgata delays sexual maturity in the presence as opposed to the absence of a predatory crayfish. Small freshwater cladocerans of the genus Daphnia develop a variety of defensive structures such as helmets, spines, and neck teeth when they receive chemical cues from predators. Some plants produce defensive chemicals once their leaves have been damaged by herbivores. The sun and shade leaves of many plants are another well-known example of phenotypic plasticity.

In heterogeneous environments, phenotypic plasticity is clearly a valuable strategy if it allows organisms to produce the locally optimal phenotype. The evolution of such adaptive plasticity depends on intricate control mechanisms: relevant environmental cues need to be obtained and processed such that the appropriate developmental switches are triggered in time. The ideal norm of reaction matches the local optima in all environments that an organism is likely to encounter. However, this might be difficult to achieve in reality. Environmental cues might provide only imperfect information about future conditions such that the phenotypic expression shows a time lag relative to the optimum. In addition, adaptive phenotypic expression across environments can be constrained by the underlying genetic machinery. For example, a mutation that increases egg size by a constant amount in two environments might move the population mean closer to the optimum in one environment but could cause a shift away from the optimum in the other. What would be preferable instead is a mutation that has an environment-specific beneficial effect on egg size. Moreover, the norm of reaction is more likely to be optimized for those environmental conditions that are frequently encountered rather than those that occur only rarely. The norm of reaction as a whole can be viewed as a trait in its own right. It has a genetic basis and is subject to the selection pressures that particular patterns of environmental heterogeneity produce. As such it is of concern to the ecological geneticist.

D. Are the Ecological Variants Adaptive?

The question of whether the observed phenotypic variation is adaptive arises similarly for genetic variants and
phenotypic plasticity. There are a number of different methods to demonstrate adaptation. For example, one can try to observe the fitness consequences of phenotypic variants directly. In the butterfly Colias meadii, male competition for female is determined partly by flight performance. Males that are capable of prolonged flight and effective courtship display have higher mating success than those that fly less well. A significant proportion of the variation in flight ability and mating success is explained by the males' genotype at two particular enzyme loci. Both phosphoglucone isomerase (PGI) and phosphoglucomutase (PGM) play key roles in the glucose metabolism and thus in the energy supply of the flight muscles. At each of these loci, heterozygous males had better mating success than homozygous males. It appears that heterozygote have a superior balance of enzyme kinetics and stability.

Adaptive differentiation in Trinidadian pupfish, Poecilia reticulata, has been subjected to a particularly detailed analysis. There is marked small-scale differentiation among populations in the intensity of the males' color, despite the fact that all females tend to prefer more brightly colored males. The observed differences were shown to represent a balance between sexual selection and varying predation pressures in different habitats. Differences between populations in schooling behavior matched a priori expectations based on predation levels. The action of natural selection could be directly observed in pupfish that were transplanted from a high to a low predation site in Trinidad in 1976. Within a few years, the males had become brighter in coloration. In addition there was genetic divergence in life history traits that continued to intensify over 11 years. A second experiment in a different river in Trinidad prompted similarly dramatic phenotypic evolution.

III. ECOLOGICAL VARIANTS IN SPATIALLY STRUCTURED POPULATIONS

A. Scale

Given abundant genetic variation within populations and large-scale environmental heterogeneity, one might expect widely separated populations within a species range to evolve different ecological characteristics. It is much less obvious how closely organisms can track small-scale environmental heterogeneity. Intuitively, one might expect that there is a minimal scale of spatial variation below which local adaptation is impossible. This idea is closely allied to the concept of environmental grain. If environmental conditions change on a smaller scale than that on which individual organisms operate, then the environment is fine grained and perceived in an averaged form. On the other hand, the environment is coarse grained if organisms are exposed to only one of several environmental conditions within their lifetimes. The relative scales of spatial variation in selection pressures and of individual dispersal should thus be critical to the question of local adaptation in heterogeneous environments. Furthermore, the rate of new favorable mutants is influenced by stochastic forces, which can slow the rate of adaptation. The strength of this effect depends in turn on the population structure of the organisms.

B. Selection versus Gene Flow

Consider first the case in which a large, randomly mating population exchanges migrants with the remainder of the species gene pool such that a proportion $m$ of the population is replaced with immigrants every generation. In this population, an allele $a$ at a given locus has arisen with a selective advantage $s$ over the wild-type allele $A$. This so-called selection coefficient $s$ applies to individuals with the heterozygous genotype $Aa$ relative to the wild-type homozygote $AA$. The other homozygote $aa$ will initially be too rare to affect the dynamics. For every offspring produced by an $AA$ individual, a heterozygote will on average produce $1 + s$ offspring. Under these conditions, $a$ can become established in the population as long as $s > m$; it will be lost due to the influx of $AA$ immigrants otherwise. In the latter case, migration outweighs selection.

This model is simple, because the only structuring element is the boundary of the population which coincides with the area in which the allele $a$ is favored. A much fuller analysis of spatial dynamics is possible if one assumes a uniform distribution of organisms in space. No populations are delineated. Instead the subset of organisms with which any one individual is likely to interact is determined by the mean dispersal distance, denoted as $\sigma$. More specifically, $\sigma$ represents the standard deviation of the distribution of distances between the organisms' places of birth and reproduction. The dispersal direction is assumed to be random. Thus individuals interact on a local scale that is determined by their movement range, but given enough time genes can in principle flow through the entire gene pool without encountering any physical barriers.

Assume further that the allele $a$ has a selective advantage $s$ inside an "environmental pocket" of specified size and a disadvantage $-s$ outside this region. The
minimum size of an environmental pocket in which a could become established is determined by the critical scale \( l_c = \frac{\sigma}{\sqrt{s}} \). As intuition suggests, this scale increases with the dispersal range of the organisms and decreases with the strength of selection. In two dimensions, the minimum size of the environmental pocket has to be greater than 1.66 \( l_c \). Consider for example a snail species with a mean dispersal distance of \( \sigma = 30 \) m. If a locally favored allele has a fitness advantage of 1%, then the minimum diameter of the pocket is roughly 500 m. In order to halve that diameter, selection would have to be five times stronger. It follows that the environmental grain associated with local adaptation is often much larger than the movement range of the organisms, even in the case of fairly strong selection.

Moreover, the establishment of the locally favored allele does not imply that it will reach fixation throughout the pocket. It might do so only in a small central portion of the range. At the edges, a cline in allele frequency should form whose steepness again depends on \( l_c \); weak selection or a wide dispersal range would create a shallow cline, which allows the alternate allele to occur at appreciable frequencies inside the pocket. The flux of alleles across its edge further implies that the mean fitness of organisms on either side of it will be reduced relative to areas where the locally favored allele is fixed. The distribution of genotypes thus cannot match the local optimum perfectly, unless the selection pressure is very strong indeed. This conclusion depends on the assumption of a sharp ecotone, where the selection coefficients reverse their sign. The effect on mean fitness would be less pronounced if there was a more gradual environmental change.

The same modeling framework can be used to study the dynamics of quantitative traits in heterogeneous environments. Instead of the change in allele frequency, one considers the change in mean phenotype across an ecotone. Phenotypic variance around that mean is created at any one location by a moderate to large number of segregating loci and by environmental effects on the trait. Natural selection tends to shift the phenotypic mean toward the local optimum. The strength of selection is expressed as the rate of decline in fitness on either side of that optimum. The response of a quantitative trait to spatially varying selection is again governed by the balance between this strength of selection and \( \sigma \). In addition, the effectiveness of selection depends on the heritability of the trait: a low heritability limits the influence that natural selection can have on the spatial distribution of phenotypes.

The swamping of local adaptation has been documented in a number of empirical studies. Blue tits (Parus caeruleus) tend to lay their eggs shortly after the bud burst of trees such that the peak abundance of caterpillars on fresh foliage coincides with the greatest food demand of the nestlings. Around the Mediterranean coast of France, bud burst occurs later in evergreen forest than in deciduous forest. Blue tit populations in deciduous forest on the mainland, where this forest type predominates, lay their eggs at the “correct” time. On Corsica, where mostly evergreen forest occurs, egg laying happens later, again in accordance with the local food supply. However, mainland populations in small patches of evergreen forest lay their eggs too early, which entails reduced fecundity. Genetic data indicate that this apparent maladaptation is due to gene flow from predominant deciduous forest.

Similarly, a small riparian population of the spider Agelenopsis aperta has diverged from a second population in dry grassland. In contrast to the grassland site, the riparian population experiences less thermal stress, has a better supply of food, but is subject to stronger predation pressure. There are a number of genetic differences between the two populations. The riparian spiders defend smaller territories and are less likely to escalate when an intruder appears. Their generally more fearful behavior also results in a lower prey capture rate. Despite the qualitative agreement with the local selection regime, the spiders still deviate significantly from the optimal strategy as predicted by detailed theoretical models. They still defend excessively large territories, and they show too much aggressive behavior and too little selectivity in their diet. However, a second and much more isolated riparian population does match the model predictions. Direct and indirect measures of gene flow show that the first riparian site cannot reach the local optimum due to massive influx of genes from the surrounding grassland.

On the other hand, evidence for very local adaptation due to strong selection comes from the classic studies of metal contaminated mine sites. Metal tolerant plant genotypes on the mine are separated from nonmetal tolerant genotypes on normal soil by only a few meters. Similarly, in the marine snail Littorina saxatilis, a cline in allele frequency exists at the aspartate aminotransferase (AAT) locus over only 7 to 10 m at the transition from the submerged surf zone to the splash zone of a rocky shore. After a toxic algal bloom in one year killed all snails in the surf zone, recolonization from the splash zone initially obliterated the allele frequency clime. But within two years, the original allelic distribution was reestablished. The selection coefficient necessary to produce the observed differentiation over such a small spatial scale was estimated to be 0.4.
C. Selection versus Genetic Drift

Chance effects are an integral part of the passage of genes from one generation to the next. Some individuals will always leave more offspring than others even if all have identical fitnesses. If by chance all bearers of a given allele leave more than the average number of offspring, the frequency of that allele in the population increases even though it may not confer a selective advantage. This process of random change in allele frequency is called genetic drift. It interferes with the establishment of favorable mutants and may limit adaptation by reducing the amount of genetic variance in phenotypic traits.

The standard model of drift in a finite population is due to Fisher and Wright (see Box 2). The key feature is the random variance in reproductive success, which poses the risk of imminent chance loss to any rare allele. Even a favorable mutant may disappear from one generation to the next, if it exists in only a few copies and if all its bearers happen to leave no offspring. The probability that an advantageous mutant becomes established in a Wright-Fisher population is \( 2s \), where \( s \) is the selection coefficient of the heterozygote. Thus most mutants that confer a fitness advantage of, say, a few percent will be lost from the population due to chance before they reach appreciable frequencies. In large populations, the deterministic effect of selection begins to dominate once an allele has reached a threshold copy number of roughly \( 1/s \). In small populations, drift influences the fate of common selected alleles. In fact, an allele will be effectively neutral irrespective of its frequency as long as \( s < 1/2N \), where \( N \) is the population size.

Although many local populations may be small, few of these will be completely isolated. It is therefore more realistic to consider the fate of a favorable mutant in a collection of local populations or demes that are connected through gene flow. Imagine a large number of Wright-Fisher populations that exchange individuals at a rate \( m \) through a migrant pool. Allele frequencies in each group of immigrants represent the average for the total population. Gene flow is thus an entirely deterministic force in this model. Moreover, there is no spatial structure, because all demes are equally connected. Taken together, these assumptions make up the so-called island model of population genetics. Remarkably, population subdivision here has no effect on the fixation probability of a favorable allele: fixation is equally likely in a set of \( n \) demes of size \( k \) as in an unstructured population of size \( nk \). The reason for this is that the probability of chance loss of a rare mutant depends on its copy number and not on the size of the population in which it arose.

Although population subdivision per se does not affect fixation probabilities, the addition of local population dynamics does. There is copious empirical evidence in many organisms for pronounced fluctuations in local abundance including the occurrence of local extinctions and (re-)colonizations. The idea that species can persist regionally despite locally transient populations is central to metapopulation theory. As long as changes in abundance are spatially uncorrelated and migration is sufficiently common to ensure the colonization of empty habitat, these local numerical fluctuations tend to average out on the larger scale of the metapopulation. In the population genetic context, fluctuations in population sizes increase the random variance in per capita reproductive success. The more demes within a metapopulation vary in allele frequency, the more likely it is that alleles are differentially affected by local changes in abundance. Drift is thus a stronger force than in the island model with constant local population sizes. Consequently, newly arisen beneficial mutations are more likely to be lost again while they are still rare. However, once they have spread to a few demes, their subsequent fate should be determined by selection.

At present, it is difficult to assess the importance of

---

**Box 2**

**Wright-Fisher Model**

Assume an isolated sexual, diploid population of constant size \( N \) with nonoverlapping generations. To create the next generation, all individuals contribute evenly to an effectively infinite gamete pool. From this pool \( 2N \) gametes are drawn that unite at random and represent the adults of the next generation. For a locus with two alleles, this sampling process implies that the allele frequency in generation \( t+1 \) follows a binomial distribution centered on the frequency in generation \( t \). Thus, there will be successive random changes in frequency through time. Eventually, this process will lead to the fixation of one or the other allele in the population. The probability that a given allele will go to fixation increases with its initial frequency. The average "stepsizes" of the random changes per generation increases with decreasing population size. Thus, the time to fixation decreases with population size.
random drift for the establishment of new ecological variants, because we do not know how much adaptation is limited by the supply of novel mutants. Artificial selection experiments suggest that there is typically sufficient standing genetic variance in natural populations to allow for a response to selection. If intraspecific phenotypic evolution depends mostly on shifts in the frequency of existing alleles, then the main role of random drift might be to slow the process of adaptation rather than to inhibit it altogether. This may be of little consequence, if there is only spatial variation. However, environmental heterogeneity typically has a temporal as well as a spatial component. If the local optimum changes through time, then a delay in the adaptive response due to drift might keep the population away from the optimum for most of the time.

IV. INFERENCES FROM THE SPATIAL DISTRIBUTION OF NEUTRAL MARKERS

A. Direct and Indirect Measures of Gene Flow

Gene flow clearly plays a central role in the dynamics of ecological variants in heterogeneous environments. Its role has been the matter of considerable controversy in the past. Does wide ranging gene flow impose limits on intraspecific differentiation or is gene flow on the contrary so limited that populations of a species behave as nearly independent evolutionary units? Based on the available evidence, the latter view is closer to the truth. Nevertheless, it is difficult to predict the potential for local adaptation and differentiation for any given species. Reliable estimates of gene flow are thus highly desirable. They might be obtained by monitoring the movement of marked individuals. However, such direct estimates typically give an underestimate of gene flow. Long distance dispersers will often be missed, yet they play an important role in the spread of genes. Mark-recapture studies provide a snapshot of dispersal whereas a longer-term average is required for evolutionary inferences. On the other hand, not all observed movement necessarily leads to gene flow. For example, immigrants might be less successful than residents in the competition for territories. Much effort has therefore been devoted to the developments of indirect measures of gene flow that can provide a suitably averaged estimate. Most of these methods are based on the spatial distribution of neutral genetic markers.

Consider again the island model: a large number of demes, each of constant size $N$, are connected by gene flow at a rate $m$ via a common pool of migrants. The overall frequency of a certain selectively neutral allele in the population as a whole, and consequently in the migrant pool, is $\hat{p}$. Genetic drift within demes produces variance in $p$ across demes as a function of local population size. Without gene flow, this process would inevitably lead to the random fixation of one or the other allele in each deme. With gene flow there is an equilibrium amount of differentiation: the divergence among demes due to drift is balanced by the homogenizing effect of gene flow such that the variance in $p$, $V_p$, is constant. For a locus with two alleles, Sewall Wright defined the standardized variance in allele frequency as $F_{st} = V_p / \hat{p} (1 - \hat{p})$ and showed for the island model that

$$F_{st} \approx \frac{1}{1 + 4Nnm}.$$ 

The degree of differentiation among demes in the island model thus depends on the number of migrants per deme per generation. Consider a given combination of $N$ and $m$. An increase in $N$ reduces random drift such that a smaller migration rate suffices to keep $V_p$ constant. The two forces exactly balance each other. The analysis of the island model shows that only a small amount of gene flow is enough to maintain neutral genetic variability within demes. The threshold number of migrants below which there is a tendency for demes to fix by chance one or the other allele is $Nnm = 0.5$, which is equivalent to one migrant every other generation or an $F_{st}$-value of 0.33.

B. Interpreting $F_{st}$

At equilibrium, $F_{st}$ expresses the balance between gene flow and drift. As such it can be estimated from the distribution of neutral alleles in any set of populations. However, in principle no inference can be drawn about the relative importance of the two component processes. A given $F_{st}$ estimate may stem from a set of large populations that are connected by weak gene flow or from small populations that exchange migrants at a correspondingly higher rate. Even the interpretation of $Nnm$ as the number of migrants is problematic, because real populations differ in important respects from the island model.

Consider populations that fluctuate in size around a certain mean abundance $\bar{N}$. Random drift is disproportionately affected by the smallest population sizes. Therefore, allele frequencies are subject to stronger stochastic forces in this fluctuating population than in a population of constant size $\bar{N}$. In many species, the variance in reproductive success is much larger in males than in females: few males may sire most of the offspring.
while others do not contribute to the next generation at all. The pool of paternally transmitted alleles is therefore much smaller than the census number of males would suggest. Again, drift forces are stronger as a result and correspond to those in a much smaller ideal population (cf. Box 2). The effective population size \( N_e \) of a natural population is equal to the size of an ideal population that experiences the same amount of genetic drift. In the previous two examples and in natural populations in general, \( N_e \) will be much smaller than the census population size. It follows that \( F_{st} \) can only be translated into the rather abstract quantity \( N_e m \). \( F_{st} \) describes a pattern of differentiation. Without additional information about the population structure at hand, it is difficult to infer a process from it.

Further complications arise when \( F_{st} \) estimates are low due to recent common descent of the populations rather than high levels of gene flow. Imagine a species that rapidly colonizes an area of patchy habitat from a large pool such that all local populations start out with nearly identical allele frequencies. If gene flow is subsequently reduced to a low level, differentiation among demes will increase with time until the migration-drift equilibrium is reached. During this time period, estimates of \( F_{st} \) reflect past history as well as current processes. The relevant time scale in generations is given in the island model by \( 1/m \) for \( N_e m < 1 \) and by \( N_e \), otherwise. These results imply that \( F_{st} \) equilibrates faster than any other measure of genetic differentiation.

C. Extrapolation to Selected Variants

Finally, it is difficult to extrapolate from the dynamics of neutral genes to those that govern selected variants. The fate of a novel mutant is in the first instance determined by the balance between selection and drift within a deme, which depends on \( s \) and \( N_e \), whereas its spread to other demes, once it has reached appreciable frequencies, depends on \( m \). Thus, predictions about the fate of such a mutant in structured populations require knowledge of \( s \), \( N_e \), and \( m \) separately. Different population structures that give rise to the same \( F_{st} \) could differ with respect to adaptive processes.

There is a large body of empirical literature on indirect measures of population structure that covers a broad range of taxa. These studies indicate that the balance between gene flow and genetic drift allows in most cases for no more than moderate differentiation in allele frequencies at neutral loci. Most \( F_{st} \) values lie between 0 and 0.1; only very few exceed Wright's threshold of 0.33, which implies that most demes tend to be near fixation at a given locus for some neutral allele. The typically low estimates of \( F_{st} \) indicate that local populations are unlikely to be so small and isolated that adaptive processes are severely compromised. The populations are either large and should therefore generate a reasonable supply of new mutations on their own or they are so well connected that they receive a steady stream of new alleles from elsewhere. Even if genetic markers indicate limited within-population variation, there may still be considerable genetic variance in phenotypic traits. In the latter, genetic variation can usually arise at a large number of loci, whereas the single locus mutation rate strictly determines variation in markers.

Limited differentiation at neutral loci does not imply uniformity in adaptive traits. This is illustrated by a number of studies. For example, Scots pine, *Pinus sylvestris*, has a circumpolar distribution occurring from Western Europe across Siberia to northern North America. Allozyme studies have detected only very slight genetic structure across this vast distribution range (\( F_{st} = 0.02 \)). However, direct assessments of the fitness of transplanted versus locally grown trees indicate important differentiation in frost hardiness, the onset of flowering as well as a number of other ecologically important traits. Similarly, the endemic perennial *Primula scotica* has a fixed genotype at 14 out of 15 allozyme loci. Yet common garden experiments reveal significant heritabilities in 6 out of 11 morphological characters. The number of seed capsules, a fitness-related trait, has a heritability of 0.25. Moreover, roughly one-third of all phenotypic variation is due to differences between populations. Clearly, the remaining 15 populations of this species in the far north of the British Isles have not lost their ability to respond to environmental change.

V. THE GENETIC BASIS OF ECOLOGICAL VARIANTS

A. Simple versus Complex Traits

Every phenotypic trait results from a complex interplay of many gene products that influence its mature expression over the course of ontogeny. A mutation at any one of the underlying loci can in principle alter a trait. Yet only those loci that actually are polymorphic at any one time contribute to the trait's heritable variation. It is the genetic basis of phenotypic variation and not of the phenotype per se that is relevant in the context of evolutionary analysis. Traditionally, traits have been referred to as simple if their heritable variation could be attributed to a small number of loci on the basis of crossing experiments. In contrast, a complex trait is typically based on many polymorphic loci as well as
on random environmental variation. Its evolutionary
dynamics are best analyzed statistically in the frame-
work of quantitative genetics. However, recent develop-
ments suggest that the dichotomy of simple versus com-
plex traits is somewhat misleading. Whether few or
many loci determine phenotypic variation appears to
be more due to the nature of selection than to an
inherent property of the trait. Moreover, modern ge-
netic techniques demonstrate that loci of large effect
are segregating for traits that were thus far believed to
fit the quantitative genetic model, which assumes a
large number of loci, each of which has a small effect
on the phenotype.

B. Are Major Mutations Important
for Adaptation?

Mutations of large effect have until recently been
thought to play only a limited role in phenotypic evolu-
tion. After all, when a large and random change is
applied to an integrated organism, it is much more
likely to cause harm than good. According to this view,
such major deleterious mutations are efficiently re-
moved from the gene pool by natural selection, and
most adaptive evolution occurs through the accumula-
tion of mutations of small effect. Recent theoretical
analyses (Orr, 1998), however, have cast a new light
on the genetic basis of adaptation. The contribution
of large mutations depends on their frequency of occur-
rence, their chance of being advantageous, and finally
on their chance of fixation given that they are advan-
tageous. The first two criteria indeed limit the import-
ance of large mutations: they are presumably rare and are
relatively unlikely to improve fitness. Nevertheless, the
small set of large advantageous mutations will be impor-
tant for evolution if selection is strong and if the current
mean phenotype in the population differs from the optim-
um by a large amount. As adaptation moves the mean
toward the optimum, mutations of decreasing effects
should become fixed. One should thus expect that espe-
cially adaptation to novel conditions involves both mu-
tations or large and of small effect, which tend to be-
come fixed early and late in the process, respectively.

C. The Strength of Selection Shapes
Genetic Architecture

The foregoing discussion suggests that the genetic basis
of ecological variants is intimately linked to the nature
of selection. In fact, the classic case studies of ecological
genetics such as the banding patterns of snail shells
and the color morphs of butterflies have often been
considered to be special cases because of the apparent
simplicity of the traits. Although they may well be end-
points of a spectrum, it appears that this spectrum is
defined not by some measure of trait complexity but
by the strength of selection involved. It is also not
surprising that many of these adaptations such as the
cryptic dark coloration of moths on sooty tree trunks
(industrial melanism) are responses to strong anthropo-
genic selection pressures.

The application of insecticide is an example for the
most extreme selection pressure imaginable: the aim
is to kill 100% of the insect population. All cases of
insecticide resistance that have been analyzed in detail
are based on one or two loci of major effect which
are able to transform a susceptible population into a
resistant one. They need to be either present in the base
population before the application of the pesticide or
arise soon thereafter. There are some indications that
mutations that allow for adaptation to extreme anthrop-
genic selection pressures might be in short supply.
For example, in the blowfly Lucilia cuprina, resistance
to each of two different insecticides produced resis-
tances at the same two loci in replicate selection lines.
Under less stringent artificial selection, polygenic resis-
tances were observed involving different loci in different
lines. These results imply that there were many avenues
to evolve weak resistance, but there was only one ge-
etic way to achieve resistance to a given insecticide
applied at field strength. Nevertheless, the rapid evolu-
tion of resistances in insects and weeds argues against
a strong limitation on available mutations.

Minor loci that contributed to a polygenic response
were similarly observed in plants that had evolved heavy
metal tolerance: there was heritable variation within
distinct resistance classes, each of which corresponds
to a major mutation. Similarly, in the case of industrial
melanism in Biston betularia, a trait that is largely attrib-
utable to a single dominant allele, one finds additional
alleles of relatively smaller effect. These observations
underscore the point that the genetic variation in any
of these traits is based on a spectrum of allelic effects.
Which of these alleles contributes to a given adaptive
response should depend on the selection pressures and
the difference between the population mean and the op-
timum.

In comparison to the application of insecticides, arti-
ficial selection experiments are more benign because
the experimenter needs to keep the selected lines going.
Here too, alleles of large effect contribute a large part
of the selection response, but typically the response is polygenic. The critical question is how often natural selection on ecological variants is so strong that major mutations are fixed. It seems likely that microevolutionary processes that are of interest to the ecological geneticist will often fall into that category simply because only the rapid response to strong selection is easily observable. Moreover, all of the preceding examples involved so-called directional selection in which the population mean is adjusted to match a new phenotypic optimum. If the population mean is at the optimum, then selection is stabilizing and mutations that cause large deviations from the mean are strongly selected against. The relative importance of stabilizing versus directional selection in natural populations depends on whether or not populations are typically in a state of selective equilibrium. This is at present an unanswered question of considerable interest.

D. QTL Mapping

Even those traits that so far have appeared to have a complex genetic basis have recently become amenable to detailed genetic dissection. The critical methodological advance has been the development of a nearly unlimited supply of genetic markers. These can be used to identify relatively small genomic regions that carry one or more loci affecting a given trait. These regions are referred to as quantitative trait loci or QTLs. The approach requires two or more populations that have diverged in some character of interest and ideally have fixed alternate alleles at each marker locus. Through controlled crosses, one initially establishes a so-called linkage map that represents the relative location of the markers to each other. With the help of further crosses, one searches for statistical associations between the expression of a given trait and a marker locus. Such an association is equivalent to finding a QTL for that trait in the vicinity of the marker. Studies of this type have uncovered a surprising number of loci of large effects that segregate in natural populations. Thus major loci may explain a large proportion of phenotypic variation even in traits that show continuous variation and conform phenotypically to the model of quantitative genetics.

E. Ecological Consequences of Genetic Structure

Whereas the genetic basis of a selection response might be in part determined by the selective regime under which it arose, a given genetic structure can subsequently influence ecological processes. A famous example in this context is the analysis of the mimetic colorations of Heliconius butterflies. Both H. melpomene and H. erato exist in a number of distinct color morphs or races that occur in nonoverlapping regions of their respective species range. Both taxa are unpalatable to avian predators. In any one location, birds learn to avoid the local wing pattern that is shared between co-occurring Heliconius species form a so-called mimicry ring. The maintenance of local color morphs is due to frequency dependent selection: the locally frequent wing morph is well protected against predation because birds have learned to avoid it, whereas any rare morph that might appear suffers predation from naive birds that have yet to learn about its distastefulness.

Through the analysis of numerous hybrid crosses between the local races, the total phenotypic variation of the different color morphs within each of the two species could be attributed to 22 and 17 loci respectively. Some of these have large effects on the formation of pattern components such as wing bands and spots, whereas others modify some detail of an existing pattern. There are complex dominance relationships as well as interactions among loci, such that the effect of a given allele at a particular locus depends on the individual genetic background in which the allele is placed (epistasis, cf. Box 1). For any one pair of geographically adjacent mimetic forms within a species, one finds genetic differences at no more than four major loci plus some small number of modifier loci. Nevertheless, this number of loci is sufficient in principle to produce a large number of recombinants upon interbreeding in areas where the racial distribution ranges adjoin. These animals with recombined wing patterns should suffer greatly increased predation, as any of the local birds would not recognize them as distasteful.

In contrast, there is another mimetic system that involves the swallowtail butterfly Papilio memnon. This butterfly is edible, but females gain protection from predation because they mimic other distasteful species. In this case, one finds several different female morphs within populations, which are due to allelic variation at a so-called supergene consisting of several tightly linked loci. As a consequence, only a limited number of phenotypes is produced, each of which may be mimicking different models. The high fitness of these morphs is in part a consequence of the underlying genetic structure. The existence of several unlinked loci as in the case of Heliconius would preclude intrapopulational variation in color morphs, because too many of the possible recombinants would not be adaptive.
VI. LIMITATIONS TO THE EVOLUTION OF NEW ECOLOGICAL VARIANTS: TRADEOFFS

A. Antagonistic Pleiotropy

Although the broad spectrum of genetically determined ecological variation within species underscores the efficacy of natural selection, organisms are not infinitely malleable. Even if a trait displays heritable genetic variation, it may not be able to evolve toward a new phenotypic optimum, because the necessary genetic changes could be deleterious with respect to some other trait. If a given set of polymorphic loci affects more than one phenotypic trait, these loci are said to have pleiotropic effects and the traits in question are genetically correlated. If an adaptive change in one trait causes a fitness reduction in a correlated trait, one speaks of antagonistic pleiotropy. In that case, the realized set of phenotypes in a given natural population can at best represent the optimal tradeoff between the conflicting selection pressures on correlated traits.

Direct evidence for such antagonistic pleiotropy comes from a number of laboratory experiments in which a given trait is altered by artificial directional selection for a number of generations. When the selection regime is then relaxed, one observes in the following generations a return of the trait mean toward the original value. This reversal reflects natural selection in the opposite direction due to some correlated trait.

B. Life History Tradeoffs

Strict negative genetic correlations for fitness are at the heart of life history theory. The schedules of growth, reproduction, and mortality together determine the life history of a species. Although one can imagine an arbitrarily fit organism that maximizes its fitness through independent optimization of each life history component, actual life histories are constrained through a web of genetic correlations. The most obvious examples are those that have to do with the allocation of energy. For any one bout of reproduction, for example, a given amount of energy can be allocated to few large or many small offspring. Within-population genetic variation in offspring size and number implies polymorphism at one or more allocation loci. More often, life history tradeoffs involve time delays. For example, deer hinds (Cervus elaphus) with fawn enter the winter season with smaller fat deposits and suffer a higher mortality than those that did not have a fawn that summer. Similarly, female collared flycatchers (Ficedula albicollis) that first reproduce at the age of 1 have lower subsequent reproductive success than those that delay first reproduction until they are 2 years old. Aside from tradeoffs between current and future reproduction, those between early reproduction and life span and between reproduction and growth have been demonstrated. In most cases, their physiological and genetic basis remains to be discovered.

Genetic variation in life history strategies allows for their adaptive adjustment. For example, an environment in which adults suffer high mortality from predation selects for earlier reproduction. This strategy brings a net gain in fitness, even if there is a correlated increase in intrinsic adult mortality. The latter is of no consequence in this environment. Due to their intimate relationship with fitness and strong dependence on ecological conditions, one expects and does in fact find abundant intraspecific variation in life history strategies. Moreover, by its very nature local life history adaptation creates genotypes that trade off their fitness in one habitat against that in another one and thus creates ecological specialization.

C. Threshold Traits

A remarkable illustration of a tradeoff is the wing dimorphism in a variety of different insect species. Even within populations, one finds individuals with large functional wings as well as those that are flightless. This is an example of a so-called threshold trait, in which a polygenic basis produces two distinct phenotypes through a developmental switch. Winged individuals are clearly better able to disperse when local conditions deteriorate, yet the production of flight muscles diverts energy away from other organismal functions. The relative proportion of the two types tends to correlate with the local ecological conditions. In the cricket Gryllus firmus, the production of a functional flight apparatus entails reduced fecundity in females and the production of a less attractive, shorter mating call in males. The tradeoff, however, goes beyond this simple dichotomy in energy allocation between the two morphs and involves an intricate genetic correlation: in families with a high proportion of long-winged offspring, even flightless sons produce a relatively shorter call than is seen in flightless male offspring from mostly short-winged families.
D. Are Genetic Correlations Constant?

Genetic correlations need not be unbreakable. Like heritabilities, they reflect the genetic architecture in a given population at a given time. For example, only in one of two populations of Rana sylvatica was there a strong and positive genetic correlation between development time and size at metamorphosis, while the other two traits were uncorrelated in a second location. With time, new mutations can change an existing genetic link between two traits. In the sheep blowfly, Lucilia cuprina, the newly evolved resistance to an insecticide initially entailed a fitness cost in the absence of the toxin. This cost was subsequently eliminated through the appearance of additional mutations. These so-called modifiers restored the fitness of the resistant genotype to the same level as that of susceptible wild type. Nevertheless, genetic correlations undoubtedly constrain phenotypic evolution especially over the relatively short time scales that are of interest to the ecological geneticist. They are a likely explanation for the fact that plant genotypes that confer resistance to herbivores typically do not reach fixation in natural populations. Even though genetic correlation may often not be strong enough to rule out an adaptive response, they can certainly slow the process so that a given population might not be able to track spatio-temporal environmental heterogeneity efficiently.

VII. ECOLOGICAL GENETICS OF SPECIES INTERACTIONS

The joint analysis of population genetics and population dynamics has long been on the agenda of ecological geneticists. In fact, the integration of these two large bodies of theory seems like a natural focus in a field of research that is by definition positioned at the interface between ecology and genetics. Yet progress has been hampered by the conceptual and practical complexity that any natural study system poses. Most progress has been made in the fields of host-pathogen and host-parasite interactions.

If several host genotypes interact with several pathogen genotypes, then the question arises how such genetic diversity can be maintained. Imagine a system in which a host population harbors a resistance alleles for every virulence allele in the pathogen population, but each host individual carries only one such allele. The host is resistant against the pathogen strain with matching allele, but it will be infected by all other pathogen strains. Frequency dependence in such a system can in principle lead to a stable equilibrium of multiple genotypes, but the approach to equilibrium in a single population is typically marked by pronounced fluctuations in allele frequency such that most variants could be lost in the process. In a single population of constant size, the system only yields abundant polymorphism when there is a large fitness cost to resistance and virulence. However, if one allows for numerical and genetic dynamics and moreover adds spatial structure, coexistence is possible even under much reduced costs. Genetic types that have been lost locally can be reintroduced through gene flow. Even if the disease dies out in a local population because only resistant host genotypes are left, a colonization event elsewhere may help to reset the system. Stochasticity in local encounters breaks the violent oscillations of frequency-dependent dynamics.

Metapopulation dynamics of this sort have been extensively documented in the interaction of the plant Silene alba and its pathogen Ustilago violacea. Heritable resistance levels within a host population ranged from 0 to 100%. However, little is known in this system about pathogen virulence types. A lack of close matching in the occurrence of host and pathogen strains was observed in the Australian flax Linum marginale and its fungal pathogen Melampsora lini. The local co-occurrence of genotypes of host and pathogen in a set of interconnected populations gave little indication that the latter tracked the frequencies of their hosts. Instead, the pattern appears rather haphazard. Gene flow together with local extinctions and colonizations are likely causes of this pattern.

One of the ecological consequences of diverse host and pathogen genotypes is that the chance of epidemics is small and leads to no more than local outbreaks. The metapopulation dynamics are also likely to affect the evolution of resistance and virulence. Consequently, the joint numerical and genetic dynamics are relevant to the process of coevolution. The so-called geographic mosaic theory of coevolution postulates that the evolutionary interactions among closely interacting species vary qualitatively from one population to another depending on the transient local mix of genotypes. In a few places at any one time, these local interactions may be strong and persistent enough to produce truly coevolutionary dynamics. This idea integrates within-population genetic and numerical dynamics with metapopulation dynamics and spatially varying selection. Complex dynamics of this type clearly are commonplace in nature. The development of this theory prom-
ises new insights into microevolutionary processes of species interactions.

See Also the Following Articles
DIFFERENTIATION • GENETIC DIVERSITY • INSECTICIDE RESISTANCE • PHENOTYPE, A HISTORICAL PERSPECTIVE

Bibliography
ECOLOGY, CONCEPTS AND THEORIES IN

Peter Kareiva* and Michelle Marvier
*National Marine Fisheries Service and 'Santa Clara University

I. Historical Development of Ecology’s Central Concepts and Theories
II. Population Dynamics
III. Theories of Species Interactions
IV. Theories of Biodiversity
V. Spatial Patterns
VI. Scaling Laws
VII. Ecosystem Theories
VIII. Evolutionary Theory in Ecology
IX. Synthesis

ECOLOGY IS THE STUDY OF THE DISTRIBUTION AND ABUNDANCE OF ORGANISMS. Ecologists attempt to understand the factors that both promote and limit biodiversity. In practice, the science of ecology blends observation of natural systems, manipulation of environmental and biotic factors in the field, more highly controlled laboratory or microcosm studies, and both qualitative and mathematical theory. The major concepts and theories of ecology concern the dynamics of populations as well as the patterns and processes of communities and ecosystems.

GLOSSARY

coevolution Process of reciprocal genetic changes in interacting species that result from the mutual selection pressures that strongly interacting species can exert on one another.

keystone species Species whose presence fundamentally alters the structure or function of an ecological community or ecosystem.

metapopulation Collection of subpopulations connected by dispersal. Subpopulations within a metapopulation may undergo local extinction and subsequent recolonization.

threshold theorem of epidemiology If the density of susceptible individuals in a population falls below some critical threshold, then a disease may not be able to maintain itself in the population.

1. HISTORICAL DEVELOPMENT OF ECOLGY’S CENTRAL CONCEPTS AND THEORIES

Ecological theory has three major origins: (1) attempts to manage fisheries, pests, or wildlife, (2) explorations of patterns in nature—especially spatial pattern and the apparent order of species assemblages, and (3) investigations of “the balance of nature” (and how that balance can be disturbed). To some extent, ecological theories developed separately in the sub-disciplines of plant and animal ecology. For example, quantitative plant ecologists tended to focus on patterns of plant dispersion and relationships between individual plants and their environments, whereas animal ecologists focused more on population dynamics. But the most important early
ecological models came from neither plant nor animal ecologists; instead these pioneering models were largely contributed by scientists from other fields, such as physics, mathematics, and human demography.

The acceptance of mathematical approaches into mainstream ecology has been slow (and some might argue that mathematical theory is still not accepted by many ecologists). The reluctance of ecologists to embrace mathematical models reflects the different ways that empiricists and theoreticians view the world. In particular, whereas field ecologists are typically concerned with the uniqueness of their study systems and pay attention to the role of history and “details,” mathematical ecologists often neglect the role of history and gloss over details to arrive at more general results. Fortunately, there has been increasing reconciliation between the theoretical and empirical approaches to ecology, largely because modern ecological theory has matured enough to allow examination of how details such as age structure, spatial heterogeneity, or environmental variation alter the generalities derived from more simple models. In addition, recent ecological theory relies less on “equilibrium analyses,” which tended to downplay historical effects.

Whereas many branches of natural science are grounded in some core body of theory and first principles, this is not the case for ecology. Instead, ecological theory is remarkably varied and discordant, with natural selection as the only fundamental “law.” Moreover, although one could argue that little in ecology makes sense without the perspective of natural selection, there is a great deal of useful ecological theory that never makes any contact with evolutionary principles. Thus, even though all ecologists accept evolutionary theory, different subfields of ecology do not share a common view of the most important theoretical questions (such as ecosystem analysis versus population biology). However, in spite of this notable lack of a single unified and ascendant ecological theory, current attention to conservation and resource management has stimulated the practical integration of models that range from evolving gene pools to ecosystem nutrient cycling.

II. POPULATION DYNAMICS

Births and deaths are the bottom line of ecology—this is because variation in birth and death rates determines the distribution, density, and dynamics of populations. However, most population models avoid tabulating particular sequences of births and deaths and instead keep track of average rates of change. In most cases, models of population change are not precise forecasting tools, but rather are conceptual frameworks with which to understand when a population is expected to stay constant as opposed to fluctuate wildly. Models are also valuable for asking how external factors such as harvesting and habitat destruction might alter a population’s long-term fate. Two fundamental features of population dynamics are inescapable:

1. Because of resource limitation, per capita rates of change for populations decline with increasing density.
2. If there is any time lag in the response of demographic rates to declining per capita resources, the population may overshoot their resources, then crash to low levels, then rebound, only to repeat the process again.

Formal mathematical representations of these principles have led to elegant models of nonlinear population dynamics. “Nonlinear” refers to the fact that the rate of change in populations varies nonlinearly with changes in population density. As a consequence, under a wide variety of circumstances populations are prone to fluctuate, sometimes so erratically that a graph of numbers through time would appear random, and such that any small changes in initial population size become amplified through time (an attribute called chaotic dynamics). This theoretical result is now widely appreciated by ecologists, but its illumination in the 1970s represented a major advance in the field. First, it emphasizes that populations may fluctuate without any environmental or climatic variation. Second, the phenomenon of chaos illustrates how complicated and unpredictable population dynamics can arise in the simplest systems, if nonlinear feedbacks are strong enough to cause the overshooting of equilibria followed by pronounced population crashes. Prior to the development of a theory of chaotic and nonlinear dynamics there was a tendency to always look to the external environment as the cause of population fluctuations, and to think that complicated patterns of fluctuation must have complicated explanations.

Applications of population theory have largely focused on three types of species: rare species that are of conservation concern, species that reach extraordinarily high population densities (“outbreaks”) that cause severe damage to crops or forests, and finally species that are harvested by humans, especially many fish species.

To assess threats to rare species, ecologists have developed a suite of analytical tools collectively known as population viability analysis (or PVA). PVAs are com-
monly used to estimate the probability of extinction under a range of alternative management strategies or to quantify the minimum number of individuals that will ensure survival of the species. PVA models may incorporate the effects of environmental and demographic stochasticity, catastrophes, and spatial structure and movement (see Section V). One approach that has been quite successful entails detailed demographic analyses that keep track of the birth and death rates for groups of similar individuals. When the groups are defined by age of the organisms, a life table can be developed. However, demographic matrices are flexible enough to incorporate groups based on size, stage, or sex, as well as age. The principal advantage of demographic matrices is that one can ask how sensitive the growth rate for the entire population is to changes in the survival or fecundity of a particular class of individuals. This approach is being used increasingly to compare the efficacy of alternative management scenarios. Analyses of this nature have revealed that adult survival is generally a very sensitive demographic parameter for long-lived species, whereas fecundity is often the most sensitive parameter for shorter-lived species. These findings have shaped a number of conservation decisions—for example, a demographic matrix applied to loggerhead sea turtles made it clear that even substantial improvements in the survival of baby hatching turtles would provide little benefit to the turtle population, whereas elevating the survival of adults could dramatically help this threatened long-lived species.

For outbreaking or pest species, population ecologists have asked what factors lead to population outbreaks and whether certain life-history traits are shared among those species with a propensity to outbreak. The goal of this research is to determine how best to control these species and prevent future outbreaks from occurring. One quantitative focus concerning outbreaking species has been on the role of time lags. Time lags can occur when, for example, a population does not respond immediately to changes in its own density, perhaps because the individuals store up energy reserves. Alternatively, a population may lag in its response to one of its natural enemies, such as a predator or a disease. Any such time lag, especially when combined with a high potential for population growth, can lead to dramatic oscillations in population abundance.

Finally, for harvested species there has been a focus on determining the maximum sustainable yield (MSY) that can be harvested from a population without placing it at risk of extinction. Clearly most natural populations do not grow without bound, but rather are regulated around some equilibrium density, called the carrying capacity of the environment. This assumption of density-dependent population growth (a reduction in the per capita rate of population growth as population density increases) is central to the calculation of MSY. In particular, theory predicts that the largest number of individuals can be sustainably harvested when a population is maintained at approximately half its carrying capacity. Although density dependence is a central tenet of population ecology and particularly of fisheries management, actually detecting density dependence from a time series of population abundance estimates remains difficult, and discussions about the appropriate methodology have been contentious. A large part of the difficulty in detecting pattern is the random variation inherent in estimates of population size. While part of the variation is due to observer error, a large fraction is often due to random fluctuations in the environment that change either the population growth rate or the equilibrium population size from one year to the next. It is straightforward to investigate in theory the interplay of density dependence and these external sources of variation. The bigger challenge lies in developing management models for real-world data sets in which the apportionment of variation among observer error and different effects of environmental variation is unknown. Fisheries biologists have made the greatest progress in this area, but the theory is still incomplete; it remains especially challenging to connect models of population dynamics to economic models that integrate the biological and economic aspects of resource management.

III. THEORIES OF SPECIES INTERACTIONS

The traditional core of mathematical ecological theory involves dynamical models of pairwise interactions between a predator and its prey, or between two competing species. Both of these types of interaction lead easily to the extinction of species—either because a predator imprudently consumes its entire food base or because a superior competitor usurps so much of the resources that other species cannot make a living. Not surprisingly then, the focus of species-interaction theory has been to ask what processes mitigate the tendency toward extinction in these antagonistic interactions.

Although the elimination of an inferior competitor seems inherent to competitive interactions, in nature we regularly find many species coexisting. Early on, theoretical ecologists suggested that coexistence gener-
ally occurs only when the populations of the two species are limited by different resources. In other words, coexisting species must occupy distinct ecological niches. The concept of the ecological niche motivated numerous attempts to quantify just how different two species must be in order to coexist. Measures of various traits related to resource use, such as beak size, body size, and habitat preferences, were often invoked as evidence for “niche partitioning” (the use of different resources by potentially competing species). Careful attention to such data indicated that this evidence was often not as clear-cut as initially assumed, and research was devoted to alternative explanations for the coexistence of competitors. A second generation of competition models made it clear that competing species may readily coexist without any niche partitioning, as long as some form of disturbance interrupts the exclusion of the inferior competitor. Such disturbance can take the form of predation, or abiotic factors such as storms or droughts. Implicit in disturbance-mediated coexistence is the concept of a trade-off between the ability to be a superior competitor and the ability to recover from a disturbance.

As with competition models, early models of predators and their prey suggested that stable coexistence would require certain special conditions. In the absence of these conditions, predators tend to drive their prey to extinction, and then, having exhausted their food supply, go extinct themselves. One answer to the dilemma of predator–prey coexistence was provided by a classic series of laboratory experiments using predatory mites and their mite prey that live on oranges. These experiments demonstrated that, in a simple "landscape" (a tray of oranges), the mite predator was able to quickly find and consume all of its prey, leading to the extinction of both species. However, more complicated laboratory microcosms that impeded the predator’s movement could prolong the coexistence of the predatory and prey mites by as much as an order of magnitude—slower predator movement allowed the prey to stay "one orange ahead" of predators and enjoy a temporary refuge from attack. Subsequent mathematical models have made it clear that this is a general principle: predator–prey interactions can be stabilized by any attribute that affords prey some refuge from attack. The refuge can take the form of patches of habitat, or age or size classes of individuals that enjoy low attack rates. With spatial refugia, there is still a tendency for predator–prey systems to fluctuate locally, but different patches will fluctuate out of phase so that the entire collection of patches never goes through a phase in which prey are everywhere absent.

In the last three decades predator–prey theory has been modified to model plant–herbivore and host–disease (or parasite) interactions. These models are similar to classic predator–prey theory in structure and in terms of the possible dynamics associated with the interactions. However, there are also important differences. First, the effects of herbivory are not as obvious as those of predation—individual plants are usually not killed outright by herbivory, and some plant species may actually perform better after being partially consumed. Also, studies of herbivory have typically had more of an evolutionary emphasis in focusing on how herbivores act as selective pressures on plant traits, especially plant structural and chemical defenses.

Host–disease interactions also have been a rich subject for mathematical modeling, with much of the initial theory originating from work on the epidemiology of human diseases. Epidemiological models reveal that disease spread depends on the transmission rate, virulence, latency, and infectious period of the disease, as well as the density of susceptible hosts. More specifically, a disease will be unable to persist if the density of susceptible hosts falls below a threshold value. Immunization programs represent an application of this result, sometimes called the “threshold theorem of epidemiology.” Immunization can lead to the elimination of a disease, because if a sufficiently large percentage of the population is vaccinated, an infected individual will, on average, fail to transmit the disease to a susceptible host before the infected host either recovers or dies. Epidemiological models further reveal that diseases can potentially maintain host populations at low levels or cause pronounced cycles in the host population.

Although much theoretical work has focused on interactions within species pairs, species do not, of course, occur only in pairs in nature. Most natural communities include scores or hundreds of different species, simultaneously interacting with one another to various degrees. Recently, increased attention has been paid to relatively complex interactions occurring among multiple species. The food web or, more generally, interaction web has proven to be a useful way to visualize the interactions among large groups of species with vertical links representing trophic interactions and cross-links representing competitive or mutualistic interactions (Fig. 1). Enormous effort has been focused on understanding the relationship between the patterns of connections among species and the traits of the community itself. However, interactions vary in strength, and it is clearly as important to measure the relative strength of an interaction as to merely document its presence.

The study of multiple species interacting simultane-
Figure 1. Interaction webs demonstrating three types of indirect interaction: (a) apparent competition; (b) predator-mediated competition; and (c) trophic cascade, as well as (d) a more realistic, species-rich interaction web depicting the major interactors associated with fireweed at Mt. St. Helens, in Washington.

Previously has led to the recognition that indirect effects can be important determinants of community structure, and several distinct types of indirect effects have been distinguished. For example, apparent competition occurs when two species appear to be competing because, as the population of one species grows, the abundance of the other declines. However, these changes in abundance are not caused by direct competition, but rather by a third species that consumes each prey species in proportion to its relative abundance (see Fig. 1a). A second type of indirect interaction occurs when a consumer species indirectly affects a species via its competitor. For example, a consumer that limits a competitively dominant species may indirectly benefit a competitively inferior species (see Fig. 1b). A third type of indirect effect thought to be common and important is a trophic cascade. An example of a trophic cascade is when a carnivore species limits the abundance of herbivores, resulting in luxuriant plant growth (see Fig. 1c). The presence of a trophic cascade is demonstrated when removal or exclusion of the carnivore results in a dramatic increase in the herbivore population and the subsequent decimation of plant biomass. Field studies are increasingly documenting the ubiquity and importance of indirect interactions. However, the difficulties encountered when studying the simplified food webs depicted in Figs. 1a–1c are greatly exacerbated when trying to understand the interactions occurring in more realistic, species-rich food webs (see Fig. 1d). The significance of this is that simple ecological experiments that focus on just one or two species may fail to reveal the important determinants of community structure.

In sum, theories of species interaction have largely focused on the problem of coexistence or, as it has also been called, the paradox of diversity. These theories address how so many species manage to coexist in nature, whereas simple lab experiments and mathematical models so often predict extinction. A major shift in thinking about this problem was the recognition that natural populations are subjected to environmental variability and other disturbances that can prevent the process of competitive exclusion. In addition, the complexities of multispecies interactions, including indirect interactions, can promote the coexistence of species.
Although the earliest models of species interactions seem abstract and disconnected from reality, ecology has a rich tradition of using experiments to test models. For example, the Russian biologist G. F. Gause pioneered the testing of species-interaction models using laboratory cultures of protozoans and yeast in which he could follow population dynamics on the time-scale of weeks, and thus could rapidly see whether or not a species could persist in his experimental systems. When examining competition between Paramecium species, Gause often observed that some species could drive others to extinction, a process termed competitive exclusion (and a result predicted in theory by the differential equations that depict continuous reproduction of two protozoans in competition for the same resource). Currently, laboratory microcosms involving protozoan communities are receiving fresh attention—nowadays these experiments are aimed at testing models of multispecies interactions and interaction webs.

IV. THEORIES OF BIODIVERSITY

Biodiversity is the variety of all living things, and most commonly refers to the number of different species in a particular area. Ecologists have documented striking patterns in the number of species occurring in different regions of the planet. For example, for most groups of species, there is a strong gradient of declining diversity from the equator to the poles, and the vast majority of species are therefore concentrated in tropical and subtropical regions.

Ecologists have proposed a variety of mechanisms that may be responsible for generating global patterns of biodiversity. For instance, areas with more varied landscapes and more topographic relief tend to support more diverse species assemblages than more uniform areas. Thus, spatial heterogeneity is likely to provide more ecological niches for species to occupy and therefore allow more species to coexist. As discussed earlier, classic ecological theory predicts that competitive interactions among species should lead to competitive exclusion and therefore reduced biodiversity. Several of the proposed mechanisms for global diversity patterns focus on forces that counter the process of competitive exclusion. Accordingly, it has been hypothesized that tropical regions harbor many species because their benign climate favors diseases and predators—these predators and diseases increase species diversity by attacking competitively superior species, thereby allowing weak competitors an escape from competitive exclusion. Disturbance is thought to similarly preclude competitive exclusion, although disturbance that is too intense or that occurs too frequently may result in reduced biodiversity. The idea that diversity should be reduced when disturbance is either too rare or too frequent has been termed the intermediate disturbance hypothesis.

Another robust pattern of biodiversity is the observation that islands generally have fewer species than mainland, and that larger islands tend to have more species than smaller islands. R. H. MacArthur and E. O. Wilson proposed the equilibrium theory of island biogeography to account for this pattern. According to this theory, there is a balance between colonizations of new species on islands and the subsequent extinctions of established species. On larger islands, colonizations are relatively more frequent and extinctions are less frequent, resulting in a higher equilibrium number of species on larger islands.

Human activities such as the direct harvesting of species, introductions of alien species, habitat destruction, and various forms of habitat degradation have greatly accelerated the loss of biodiversity. Consequently, current extinction rates are estimated to be 100 to 1000 times higher than pre-human extinction rates. This extinction crisis has spurred a great deal of scientific interest in the biological and ecological functions of biodiversity. Organisms are responsible for a variety of ecosystem functions, including maintenance of the gaseous composition of the atmosphere, regulation of the global climate, generation and maintenance of soils, and recycling of nutrients and waste products. However, it is not obvious whether these biological services require many species or only a handful of species. Much current research is focused on understanding how much (if any) diversity we can afford to lose yet still maintain the necessary ecosystem functions (see Section VII).

Some species play obviously important roles in ecosystems—the addition or deletion of these "keystone" species leads to dramatic changes in ecosystem functions such as productivity or nutrient uptake. However, most species probably do not exert such important effects on ecosystems. In other words, it may be possible to lose a number of species from an ecosystem without little overall impact on ecosystem function. This could be the case if several species that perform approximately the same function are present in the original ecosystem. The situation where multiple species play a similar role has been termed "species redundancy." If species redundancy is a common phenomenon, ecosystem function should be largely independent of species diversity, so long as major functional types are represented. Thus, when one species is lost from an ecosystem, some other
species with a similar function may become abundant and compensate for the lost species, leaving the ecosystem as a whole relatively unaffected. Indeed, ecosystem processes often do remain stable despite large fluctuations in the abundance of the various species involved. The term species redundancy may seem to imply that all species are not necessary for an ecosystem to function properly. However, species redundancy may in fact be an essential feature for the long-term health of ecosystems, and the importance of any particular species may only become evident during the occurrence of rare, drastic events.

V. SPATIAL PATTERNS

Spatial patterns and habitat patchiness or heterogeneity have spawned a host of concepts that try to both describe these spatial patterns and explain their impacts and sources. The two most important general results of spatial theory in ecology are:

1. The fact that species interact in spatially extensive and patchy worlds creates opportunities for persistence and coexistence of species that would otherwise not exist.

2. In uniform environments, the interplay of dispersal and nonlinear population dynamics can produce complex spatial patterning in population densities, and this patterning can itself promote coexistence in predator–prey or competitive interactions.

More generally, spatial theory emphasizes that it may be impossible to understand ecological processes and interactions if the focus and scale of a study is too small. This is an important caution, given the fact that many experimental ecology studies are conducted at scales of less than 1 m2.

One of the most important practical applications of spatial theory stems from the use of metapopulation models in conservation biology. Metapopulation theory describes a species as a collection of populations (or patches) in which there is turnover due to local extinction and subsequent recolonization by dispersing individuals, and in which the fate of a species can only be understood by tracking the fate of a collection of populations or patches, as opposed to a single local population or patch. Metapopulation models are of considerable importance in conservation applications because they identify critical thresholds for habitat destruction or for the creation of barriers to plant and animal dispersal, thresholds that, once crossed, imply doom for a species (because colonization becomes too infrequent to counterbalance local extinctions). It is obvious to anyone who has looked out an airplane window that one of humankind’s major impacts on the world is to destroy habitats and to create ever more fragmented landscapes dissected by highways, agricultural fields, and commercial or residential developments that surely hinder the dispersal of species. If this trend proceeds too far, metapopulation theory predicts the slow but inexorable loss of species. Spatially detailed models of human-modified landscapes are increasingly being used to identify strategies for land use and human development that will minimize the undesirable effects of habitat loss and habitat fragmentation. The two most common strategies involve clustering of habitats and the protection of dispersal corridors.

At a more theoretical level, spatially explicit models of species interactions indicate that the interactions themselves can result in pattern formation. For example, high rates of predation combined with limited prey dispersal can result in a dense patch of prey surrounded by a zone that is relatively devoid of prey due to intense predation. This “ring” of intense predation can limit the spread of prey populations. It is important to realize that these spatial patterns can emerge without any underlying environmental heterogeneity. Spatial patterning can allow more species to coexist because the species occupy different portions of the pattern. However, the opportunity for spatial segregation of species into different characteristic portions of spatial patterns requires sufficiently large habitats—if habitats are carved into patches that are too small, patterns cannot be generated, and opportunities for coexistence will be lost. Thus, theories of pattern formation reveal yet another mechanism by which biodiversity can be maintained, and conversely a route by which biodiversity can be degraded.

VI. SCALING LAWS

Because ecological processes operate across a huge range of scales, there have been numerous attempts in ecology to identify scaling laws—laws that allow us to study phenomena on one scale and predict those same phenomena at different scales. Just as important as identifying rules for extrapolating across scales is the idea that there are critical spatial scales at which scaling laws break down and processes fundamentally change in character.

An especially good example of successful extrapolation across scales concerns the prediction of the rate at
which exotic species invade new areas and expand their ranges. Simple invasion models extrapolate short-term observations of dispersal (over the scale of one year or less and involving the movements of single individuals) to the broad expanse of an entire species range after decades and over thousands of kilometers. For organisms as diverse as muskrats and cabbage butterflies, invasion models have successfully used small-scale and short-term data to predict long-term rates of range expansion.

To quantitatively examine the legitimacy of extrapolations across scales, it is often useful to graph the variance of a measured variable as a function of the spatial scale at which it was measured. For example, population density could be measured in anything from 1-m² quadrats to hectare plots, and the variance in density (per unit area) among plots is expected to decline with increasing plot size. The rate of this decline is a measure of spatial correlation. However, marked changes in the slope relating variance in population density to measurement scale indicate some fundamental shift in biological processes—such as a shift from patterns generated by the dispersal range of a plant’s seeds to patterns generated by large-scale climatic heterogeneity. Such nonlinear relationships between variance and scale indicate circumstances where it may not be appropriate to extrapolate across scales.

Most ecologists continue to conduct research at one and only one scale. Although theories regarding scaling laws are still in their infancy, they plainly draw attention to the fact that any extrapolation across scales is tenuous, and can only be justified if the processes have been studied or measured at more than one scale.

VII. ECOSYSTEM THEORIES

Ecosystem theories focus on patterns of energy flow, nutrient cycling, and the factors that control or limit the rates of these processes and the stability or predictability of these processes through time. To assess the flow of energy and nutrients, it is useful to recognize that organisms fill different trophic roles: autotrophs (usually plants) acquire energy, usually from the sun, and nutrients from nonliving materials; heterotrophs derive energy and nutrients by consuming living matter, and decomposers transform once-living materials back to their inorganic forms. Both nutrients and energy flow from autotrophs to heterotrophs and decomposers, and the connections among these groups, constitute a food chain or food web.

The trophic interactions among species affect the rates of ecosystem functions, including biogeochemical processes, the flow of nutrients, water, and atmospheric gases, and the processing of energy through ecosystems. Species are the key working parts of ecosystems, and therefore biodiversity must be related in some way to ecosystem function. Certainly ecosystems would not function if all species were lost, but it is unclear just how many species are necessary for an ecosystem to function properly. The current extinction crisis has prompted ecologists to assess the relationship between biodiversity and various ecosystem properties. Evidence regarding the importance of biodiversity for ecosystem function largely comes from comparing ecosystems that naturally differ in the number of species present. More recently, ecologists have undertaken more controlled manipulative experiments in which the number of species has been directly varied. Several studies have demonstrated that various measures of ecosystem function, such as production of biomass and nutrient uptake, increase as the number of species present increases. However, some studies also report no effect or even negative relationships between biodiversity and ecosystem processes.

Although some evidence suggests that biodiversity increases or improves the overall functioning of ecosystems, the underlying mechanisms remain unclear. For example, a positive relationship between species diversity and productivity could result simply because including more species increases the chance of including particularly productive, or fast-growing, species. Alternatively, a diverse group of species may use the available resources more efficiently, because each species has a slightly different strategy, resulting in higher overall growth.

A second commonly purported benefit of biodiversity is that more diverse ecosystems may be more stable compared to species-poor ecosystems. Stability can be defined at the community level as fewer invasions and fewer extinctions, meaning that a more stable community will contain a more stable composition of species. However, stability can also be defined at the population level as reduced fluctuations in population size, meaning that a more stable population will contain a more constant number of individuals.

The idea that biodiversity confers stability to ecosystems has enjoyed a long and controversial history. Ecologists initially hypothesized that diverse ecosystems were more stable than those with fewer species because, it was reasoned, with more species in an ecosystem, there are more paths through which energy and nutrients can flow. Consequently, in diverse ecosystems each species should be less affected by changes in the abun-
dance of other species, leading to higher overall stability. However, mathematical models exploring the relationship between biodiversity and population stability indicated that higher species diversity generally reduces the stability of population dynamics of individual species. The conflict between these modeling results and the original intuitions of ecologists remained unresolved for many years.

Recent experimental manipulations of species diversity have helped to resolve this long-standing controversy. In particular, several manipulations of plant and microbe diversity have been performed to examine the stability of various measures of ecosystem function over time. Generally, such studies have shown that, although the abundance of individual species fluctuates more dramatically in high-diversity ecosystems, the total abundance or productivity of all species combined is actually more stable. These empirical results are consistent with both views of diversity and stability. First, the fact that individual species fluctuate more in diverse systems supports the results of mathematical models regarding diversity and stability. On the other hand, the observation that aggregate attributes such as productivity become more stable with increased diversity supports the intuition that multiple pathways for energy (high diversity) foster stability.

Although the relationship between biodiversity and ecosystem stability is becoming increasingly clear, the mechanisms generating this pattern are not. In particular, diverse groups of species may be more stable because complementary species compensate for one another. Alternatively, variation in aggregate measures such as total productivity may increase with richness due to averaging of random fluctuations in the growth of each species. On the basis of simple probability theory, it is expected that as more independently varying species are summed together, the more stable will be the sum of their abundances. The strength of this averaging effect depends on correlations among the species' fluctuations, but a positive relationship between biodiversity and the stability of aggregate measures of ecosystem function should almost always be expected, simply due to averaging.

VIII. EVOLUTIONARY THEORY IN ECOLoGY

Evolutionary theory has entered ecology in two major areas: the evolution of species attributes and how they might vary in a predictable manner, and the coevolution of interacting species, such as the evolution of virulence and resistance in host–pathogen interactions.

There is often a remarkable matching between the traits of organisms and the local conditions in which they live. Ecologists and population geneticists alike have worked to identify the factors contributing to or preventing the adaptation of populations to various aspects of their local environments, including food sources, predators, and microclimatic conditions. Reciprocal transplant and common garden experiments have been the major tools used to test for local adaptation. In particular, there is interest in the balance between selection for locally adapted traits and the immigration of individuals with nonadapted traits into the population. This balance between differentiating selection and the homogenizing effects of migration can create a difficult conundrum for conservation biologists—for a rare species that exists in several small populations, it might be better to pool all the individuals together to bolster the size and viability of a single population. Alternatively, the subpopulations might be so genetically distinct and locally adapted to their environments that pooling the individuals might actually harm the species. Thus, one of the more difficult challenges in conservation is to balance the short-term viability of a species with preservation of its long-term evolutionary potential.

As a species evolves to become locally adapted to other species in its environment, it can exert selection pressure on the co-occurring species, leading to changes in the traits of these species with which it interacts. The reciprocal evolution of two species in response to one another is called coevolution. The term coevolution was originally coined to describe the presence of toxic compounds found in plants and the ability of butterfly species to detoxify these compounds. Indeed, many of the best-known examples of coevolution involve reciprocal changes between adversaries, a sort of ever-escalating "arms race" as might occur between a predator and its prey or between a parasite and its host.

Plant breeders try to directly manipulate the coevolutionary process that occurs between crop plants and the herbivores and pathogens that attack these plants. Selective breeding is used to maximize the defenses naturally present in crop species and, more recently, genetic engineering has allowed breeders to co-opt defenses of very distantly related organisms and introduce completely novel defense traits to crops. Unfortunately, when highly defended crops are planted over the enormous spatial scales of modern farms, herbivore and pathogen species tend to quickly evolve resistance to the plant defenses. Thus, an important application of
evolutionary ecology involves determining how we can best deploy plant defensive traits such that evolution on the part of plant pathogens and herbivores does not make the plant defenses obsolete. Theory suggests that alternating plantings of crops with distinct defensive traits may reduce the ability of herbivores to evolve resistance. In addition, incorporating combinations of defensive traits into single individuals should substantially reduce the likelihood that herbivores will become resistant. Thus, theory in evolutionary ecology has contributed concrete strategies that could greatly enhance the “longevity” of plant defenses, with enormous economic consequences.

IX. SYNTHESIS

Of the numerous specific concepts and theories discussed in this article, some general ideas emerge that relate pattern to process and dynamics to structure. First, complicated spatial or temporal patterns in population densities do NOT require environmental explanations or complicated underlying dynamics—rather, we should expect patterning, cycling, and complexity simply because of the feedbacks inherent in ecological systems. This is of more than theoretical interest, because it implies that any management of ecological systems must heed the implications of nonlinear dynamics. In other words, we must anticipate that rates of change in population growth or in number of species may change dramatically and suddenly, not just because of random fluctuations in the environment, but also as a result of deterministic processes. Second, several ecological models, including epidemiological and metapopulation models, point to critical thresholds across which systems change qualitatively (e.g., a disease can no longer persist, or a population dwindles toward extinction because colonizations fail to balance local extinctions). Third, opportunities for coexistence of species arise whenever the world in which organisms live becomes less uniform—spatial heterogeneity creates refugia, disturbances interrupt the process of competitive exclusion, and so on. In contrast, human actions tend to enforce regularity and control on the natural world, and ecological theory suggests that this homogenization of nature is contributing to the erosion of our planet’s biodiversity. The consequence of this reduction in biodiversity remains an open theoretical question, with the greatest challenges to ecological theory being the development of models that contend with rich webs of species interactions in complex ecosystems. The future of ecological theory lies in blending human activities with natural processes, so that we can better understand how to encourage a sustainable biosphere while accommodating human interests and demands.

See Also the Following Articles
COEVOLUTION • COMPLEXITY VERSUS DIVERSITY • KEYSTONE SPECIES • POPULATION DYNAMICS • POPULATION VIABILITY ANALYSIS (PVA) • RESOURCE PARTITIONING • SPECIES INTERACTIONS

Bibliography
I. Planned and Unplanned Diversity in Agroecosystems
II. Impact of Planned Diversity on Ecological Processes
III. Unplanned Diversity and Ecological Services
IV. Impacts of Agriculture in the Landscape
V. Future Directions

GLOSSARY

polyculture Growing two or more crops in the same field at the same time.
resilience The rate at which an ecosystem returns to original conditions after a perturbation.
resistance The degree to which an ecosystem changes in response to a perturbation.
transgenic crop Crop whose genome contains a gene(s) from a distinct species that has been inserted by genetic engineering.

THE POPULARITY OF AGRICULTURE as a focus of ecological study has waxed and waned during the past century. Agricultural systems have been touted as model systems for testing ecological theory because of their relatively well-characterized components, simple food webs, and manipulability. In the past, they have often served as effective model systems for plant ecology, plant-insect interactions, and predator–prey theory. In recent years, the growth of sustainable agriculture has resulted in increasing interest in understanding ecological processes in agriculture so that they can be manipulated more effectively for enhancing agricultural productivity and reducing negative environmental impacts of agricultural activities. A variety of types of ecological theory have been applied to agricultural systems, and thus the subdiscipline of “agroecology” encompasses a very broad range of approaches, including physiological ecology, population biology, community ecology, ecosystem ecology, and landscape ecology. All these approaches offer significant insights into the ecology of agricultural systems and provide useful guidelines for improving the functioning of these systems. Because the role of diversity in ecological systems has been a consistent element in much agroecological research, this article focuses on diversity in agricultural systems, with particular attention to areas of agroecology that appear to have the most potential to offer agricultural and environmental enhancement.

I. PLANNED AND UNPLANNED DIVERSITY IN AGROECOSYSTEMS

The biodiversity in agroecosystems can be classified as either "planned" or unplanned diversity. Planned diversity includes the spatial and temporal arrangement of domesticated plants and animals that farmers purposely include in the system. It may also include bene-
ficial organisms that are deliberately added to the agroecosystem, such as biological control agents or plant-associated nitrogen-fixing bacteria. Unplanned diversity includes all the other associated organisms that persist in the system after it has been converted to agriculture or colonize it from the surrounding landscape. This component of diversity is likely to include a variety of herbivores, predators, parasites, and microorganisms that make up the majority of species in any ecosystem, even a simplified one such as an agroecosystem. It has become increasingly clear that the two components of diversity are significantly linked: As planned diversity increases, the diversity of the associated biota also increases. Although the evidence for the plant-based food web is strongest, this relationship also appears to exist for the detritus-based food web.

Given this relationship between planned and unplanned diversity, how do both types of diversity influence agricultural productivity and other ecological processes that occur in and around agroecosystems? The impact of planned diversity is relatively well understood, but much remains to be elucidated about the role of unplanned diversity in agroecosystem function.

II. IMPACT OF PLANNED DIVERSITY ON ECOLOGICAL PROCESSES

Agroecosystems vary dramatically in their complexity and degree of planned diversity, and planned diversity may include genetic diversity, species diversity, structural diversity, and functional diversity (Fig. 1). The input-intensive monoculture systems that dominate commercial agriculture worldwide have low planned diversity and relatively simplified community assemblages. At the other extreme, traditional agricultural systems and home gardens of the tropics typically have extremely high planned diversity and complex assemblages of associated biota. This range of planned diversity in existing systems has provided rich opportunities for exploring the impact of planned diversity on a suite of ecological processes, including primary production, pest regulation, decomposition, and nutrient cycling.

A. Planned Diversity, Productivity, and Stability

The idea that diversity should impart stability to ecological systems, including agricultural systems, dates back at least to the 1950s and has stimulated much ecological theory, some empirical research, and extensive debate. Until recently, the best evidence for the idea that diverse systems should be more stable than simple ones came from agricultural ecosystems rather than natural ecosystems. Many agricultural studies have demonstrated increased productivity (i.e., significant yield increases) in diverse cropping systems compared to monocultures. However, the majority of these studies address systems of relatively low plant species richness, comparing monocultures to polycultures of two or three species. We expect that the relationship between productivity and plant species richness should resemble the asymptotic curve in Fig. 2, as it appears to do for natural grasslands (Tilman et al., 1996), but we have little data to indicate where the asymptote might be. Researchers have speculated that the increase in productivity with increasing species richness in agroecosystems may level out at approximately 10 plant species, although this

<table>
<thead>
<tr>
<th>Agroecosystem</th>
<th>Genetic</th>
<th>Species</th>
<th>Structural</th>
<th>Functional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herbaceous monoculture</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herbaceous varietal mixture</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herbaceous polyculture</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plantation crops</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agroforestry</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tropical home gardens</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 1  Degree of different types of planned diversity found in some typical agroecosystems. Intensity of shading indicates degree of diversity.
The tendency of diverse cropping systems to have higher productivity than would be expected on the basis of the productivity of their component crops grown in monoculture is often called "overyielding." Overyielding may result from a variety of mechanisms, such as more efficient use of resources (light, water, and nutrients) or reduced pest damage, and there have been numerous experimental studies examining these mechanisms. When interspecific competition for a limiting factor is less than intraspecific competition for that factor, overyielding is predicted. Facilitation occurs when one crop modifies the environment in a way that benefits a second crop. For example, by lowering the population of a critical herbivore or releasing nutrients that can be taken up by the second crop. Facilitation may result in overyielding even where direct competition between crops is substantial (Vandermeer, 1989).

There is evidence that diverse cropping systems exhibit greater yield stability and higher productivity, suggesting that resistance to environmental perturbation may be higher in diverse systems. Yield stability has been measured in at least three ways: by calculating coefficients of variation of yield, by computing regressions of yield against an environmental index, and by estimating the probability of crop failure. Polycultures exhibit greater yield stability according to all criteria. Polycultures tend to have lower coefficients of variation than crops grown as separate monocultures, the response of polycultures to environmental change tends to be as stable or more stable than the most stable component crop grown in monoculture, and polycultures tend to have a much lower probability of crop failure than the component crops grown in monoculture. The probability of crop failure is an estimate of risk and lower probabilities result from both the higher yields of polycultures and the putative yield stability. Overall, studies indicate that diverse cropping systems are more stable than monocultures, both agronomically and economically.

Several mechanisms may lead to greater yield stability in diverse systems. When one crop performs poorly because of drought or pest epidemic, for example, the other crop(s) can compensate, using the space and resources made available. Such compensation is obviously not possible if the crops are grown separately. If the yield advantages of polycultures are greater under stress conditions, then yield stability is higher. This polyculture advantage has been demonstrated for crops under nutrient stress and drought stress. Moreover, where polycultures lead to reduced pest attack, as they often do, then greater yield stability may result from the dampening of pest outbreaks and disease epidemics.

These processes may operate simultaneously in diverse systems and are consistent with the idea that species richness buffers productivity under conditions of environmental variability and that diversity imparts resistance to perturbation. The ability of polycultures to compensate for losses might also be considered to represent resilience. Whereas systems with greater resistance will be less impacted by a perturbation, systems with greater resilience return rapidly and reliably to original conditions.

As a result of this evidence that diverse cropping systems often have greater productivity and stability, there is a persistent conviction among many agroecologists that species-rich natural ecosystems should provide us with models for the design of sustainable agricultural systems. Jackson (1985) and colleagues at the Land Institute in Salina, Kansas, have promoted the idea that agriculture in the U.S. plains should mimic natural prairies by emphasizing polycultures of herbaceous perennials. Others have noted the structural and functional similarity between diverse home gardens in the tropics and humid tropical forests and have suggested that agricultural systems for the tropics should be designed as analogs of natural tropical forests.

Ecological research on the role of diversity in natural grassland systems by Tilman and colleagues (1996) indicates that diverse natural communities may be more productive than simple systems, at least partially as a result of increased nitrogen use efficiency in more diverse systems. These studies also suggest that more diverse plant communities are more resistant to disturbance and more resilient in the face of environmental perturbations such as drought. That is, the productivity of diverse communities appears to decline less during a drought and to return more quickly to predrought
levels than is the case for species-poor communities. It remains to be seen, however, whether the "natural systems mimicry" approach can really lead to useful improvements in the productivity and sustainability of agroecosystems.

B. Planned Diversity and Pest Regulation

The planned diversity of the agroecosystem has important effects on herbivorous insects and the microbial community attacking crops. Traditional agricultural systems often include substantial planned genetic and species diversity, and genetically diverse grain crops are used in many parts of the world to control pathogens. In contrast, the low planned diversity of most commercial monocultural systems often results in large crop losses from a pest complex that is less diverse but more abundant than that in more diverse systems. The trend for higher pest densities in monocultures compared to diverse cropping systems is especially strong for specialist insect herbivores with a narrow host range. As planned diversity increases, population densities of these specialist herbivores decrease.

In a classic paper, Root (1973) offered two hypotheses to explain higher densities of herbivorous insects in simple, monocultural systems: the resource concentration hypothesis and the enemies hypothesis. The first hypothesis predicts that herbivores, particularly specialists, in pure, dense host plant stands will be more likely to find their hosts and more likely to survive and reproduce. In contrast, herbivores in less dense or more host plant-diverse stands should be less likely to find their hosts and more likely to lose them. Although the details of herbivore-host interactions vary considerably, subsequent experimental tests of this hypothesis have generally supported it with respect to the effects of diversity, especially for specialist herbivores. Host-finding behavior and insect movement, both colonization and emigration, appear to play important roles in the response of herbivorous insects to agroecosystems. Densities of specialists may be lower in diverse systems because (i) they have difficulty locating hosts due to interference with olfactory or visual cues or (ii) they leave hosts more often due to lower plant quality and then have difficulty relocating them. These behaviors are significantly affected by the chemical, nutritional, and structural diversity that accompanies planned plant species diversity.

The enemies hypothesis predicts that diverse systems should have higher densities of herbivore natural enemies (predators and parasites) because they provide more resources for these natural enemies, such as alternate prey or hosts, nectar, pollen, and refugia. This hypothesis has also largely been supported in experimental studies. Compared to monocultures, diverse systems are likely to have higher predation rates, higher parasitism rates, and higher ratios of natural enemies to herbivores, all of which may contribute to lower pest densities. Natural enemies comprise one component of the unplanned diversity that increases as planned diversity increases in agroecosystems.

Microbial pathogens also respond to the planned diversity of agroecosystem, but their response is more variable than that of herbivorous insects. Crop diversification can modify the microclimatic conditions that play an important role in the development and severity of plant disease. Pathogen growth and reproduction may be either encouraged or inhibited in more diverse cropping systems, depending on the particular requirements of the organism. The effects of diversity depend on a variety of dispersal processes, infection efficiency, and the rate of disease progress. The majority of viruses, however, are transmitted by insects, and these tend to be found at lower incidence in diverse systems due to the effects of plant species diversity on their insect vectors.

The genetic diversity of crops can dramatically reduce pathogen impacts on crop productivity. Mixtures of genotypes of a single species, such as multiline cultivars and varietal mixtures, have been used effectively to retard the spread and evolution of fungal pathogens in small grains and other crops. There is evidence that they may also have lower densities of insect herbivores and a lower incidence of plant viruses. Typically, these mixtures include both resistant and susceptible crop genotypes, although they may be mixtures of several different resistant genotypes. The reduction in pathogen spread is greater than would be expected on the basis of the proportion of resistant genotypes in the mixtures, and therefore it appears to be due to the effects of diversity per se on the ability of pathogens to disperse.

High planned diversity that includes genetic diversity, species diversity, and structural diversity thus has a strong influence on populations of herbivorous insects and plant pathogens. The intentional manipulation of planned diversity for the purpose of pest control is common in traditional agricultural systems, and it offers real potential for increasing the productivity of contemporary agricultural systems without the negative environmental impacts of pesticides.

C. Planned Diversity and Soil Processes

Planned diversity may also have significant impacts on the soil community. Plant pathogens and their antagonists in the soil are well-known to respond to crop
diversity. Long-term, continuous monoculture can result in dramatic shifts in the competitive balance among microbial species and increase the aggressiveness of plant pathogens. Conversely, populations of plant parasitic fungi, bacteria, and nematodes may decline when monocultures of their host plants are replaced by diverse cropping systems. Rhizosphere microbial communities, including a variety of bacteria and fungi, respond to plant species composition of the cropping system, plant phenology, plant nutrient status, and even plant genotype. Plant mutualists such as mycorrhizal fungi or nitrogen-fixing bacteria may be strongly determined by crop composition and phenology. Some ecological processes such as nitrogen cycling may be substantially controlled by crop diversity and plant composition since the inclusion of legumes can increase rates of biological nitrogen fixation. Due to the effects of litter diversity, decomposition rates may also be quite responsive to planned diversity.

Despite an increasing awareness of the effects of planned crop diversity on the soil community, our ability to predict the impacts of manipulating planned diversity is limited because our understanding of the complex interactions among soil organisms and between plants and soil microbes is still rather weak. Greater attention to the ecological dynamics of these soil communities under different cropping systems will aid in determining the extent to which we can manage these communities to enhance agricultural productivity using planned diversity.

III. UNPLANNED DIVERSITY AND ECOLOGICAL SERVICES

The unplanned diversity that accompanies planned diversity in agricultural systems can provide many ecological services to agriculture. This means that the conservation of biodiversity can offer significant benefits to agriculture. Uncultivated species, including wild relatives of crops that occur in and around the agroecosystem, are an important source of germplasm for developing new crops and cultivars. Increasing planned crop diversity can augment the resources available to pollinators and to natural enemies such as parasitic wasps, resulting in higher populations of these beneficial organisms. Increasing planned diversity may also foster beneficial soil organisms and the conservation of functional processes, such as decomposition and nutrient cycling.

Overall, unplanned diversity within the agroecosystem affects plant and soil processes that can in turn affect crop productivity. For example, many traditional cacao and coffee systems in the tropics resemble forest ecosystems since they are well shaded and humid, with a diversity of shade trees, a diverse herbaceous understory, and a thick leaf litter layer. Thus, these traditional agroecosystems have high planned species diversity, high genetic diversity, and high structural diversity. Recent studies have shown that shaded cacao and coffee plantations support higher levels of biodiversity than full-sun plantations grown without shade trees, i.e., monocultures (Perfecto et al., 1996). The shaded systems may be traditional “rustic” systems in which the understory of native forests is cleared and the crop is planted under the natural forest overstory. They may also be improved systems in which a few carefully chosen tree species are planted interspersed among the crop. In either case, these shaded systems have high planned diversity accompanied by high unplanned diversity.

Shaded coffee and cacao have been demonstrated to be good habitats for migrant and resident forest birds in the neotropics, and these birds may be important consumers of pest insects. They also contain relatively high diversities of insects, lizards, and other fauna, some of which contribute to pest regulation. Pollination also appears to be enhanced in shaded systems. Since many of the commonly used shade trees are leguminous, the trees reduce the amount of nitrogen inputs required. In addition, the shade provides significant weed control and, at least in cacao, reduces the incidence of several important diseases. The high levels of unplanned diversity thus contribute significantly to agricultural productivity. Moreover, by serving as habitat for a relatively high diversity of vertebrate and invertebrate animals, such agroecosystems may enhance the conservation value of a agricultural landscape.

A. Unplanned Diversity and Pest Regulation

One of the ecological services provided by the unplanned diversity in agroecosystems is the regulation of herbivorous insects and microbial pathogens by competitors, predators, and parasites. As discussed previously, high planned diversity often results in higher densities of predators and parasites. Even in agroecosystems of low planned diversity, the diversity of natural enemies may be quite high as long as pesticides are not used. For example, paddy rice monocultures managed without pesticides can have surprisingly high diversities of herbivorous insects, predators, and parasitoids compared to similar monocultures in which pesticides are used. Recent pest management programs in Southeast
Asian paddy rice have taken deliberate advantage of this diversity and drastically reduced pesticide inputs without sacrificing yields. In traditionally managed rice fields, the predators are likely to include fish and amphibia, which contribute to pest regulation and also provide additional nutritional resources for farm families.

B. Unplanned Diversity and Soil Processes

Despite some significant advances in our understanding of soil processes, we know relatively little about the soil biota and their impact on agricultural productivity. In natural ecosystems, decomposition and soil nutrient cycling are regulated by a diverse community of invertebrates and microorganisms, such as termites, earthworms, nematodes, fungi, and bacteria. The composition, abundance, and activity levels of the soil biota in agricultural systems are markedly different from those in surrounding natural ecosystems. For example, the diversity and abundance of soil insects and earthworms in tropical agroecosystems are typically significantly reduced compared to those of a wide range of undisturbed tropical ecosystems. In cases in which abundance remains high in agricultural systems, the soil communities are often dominated by a single or small number of species highly adapted to the modified environment.

The changes in the soil community under agriculture result from a variety of perturbations to the soil environment. The initial conversion from undisturbed ecosystems typically involves the removal and/or burning of plant biomass followed by tillage—activities which have drastic impacts on soil structure and soil chemistry. The physical changes at the soil surface amplify diurnal and seasonal fluctuations of temperature and moisture. In addition, organic inputs to the soil are significantly reduced as a result of plant biomass removal, and the chemical composition of organic inputs is altered. These extreme modifications of the soil environment can result in the elimination of some soil organisms and, at a minimum, are likely to change the competitive balance among species. To the extent that agricultural systems minimize these perturbations (e.g., by reducing tillage or burning), the impacts on the soil community may be less severe.

Despite these well-known effects of agricultural conversion on the soil community, the link between loss of soil biodiversity and various ecological processes has not been extensively studied. Decomposition and nutrient mineralization, for example, are controlled by the activities of a diverse community of organisms. It is not clear that the loss of some species will result in significant changes to these functions, but the degree of functional redundancy among different species is still controversial. On the other hand, some processes such as nitrogen fixation are carried out by very specific organisms, whose loss might substantially affect nitrogen cycling. Given the importance of soil biota for decomposition and nutrient cycling, it is essential that the link between this component of unplanned diversity and function be explored more fully.

IV. IMPACTS OF AGRICULTURE IN THE LANDSCAPE

Ecological concepts have often been utilized to address the consequences of agricultural perturbations for natural ecosystems in agricultural landscapes. For example, the impacts of nutrient inputs and pesticides for downstream aquatic and terrestrial ecosystems have been well studied. Much of the pesticide applied to crops fails to reach the target pests and instead moves into adjacent ecosystems, where it can have significant impacts on the diversity and abundance of nontarget species. Studies have shown that pesticides have strong effects on trophic dynamics and a range of ecosystem processes in complex natural ecosystems, both aquatic and terrestrial. Similarly, movement of nitrogen from agroecosystems to natural ecosystems through leaching or deposition may lead to significant impacts at several trophic levels. Increased nitrogen loading may reduce plant species richness while increasing the biomass of dominant species. In response to these changes, both pathogens and herbivorous insects are likely to exhibit lower species richness but higher population abundance, just as they do in agricultural systems. Changes in pathogen and herbivore populations may in turn influence the productivity and fitness of their host plants. Even in the absence of changes in herbivore population pressure, research has shown that the effects of herbivory on plant growth and reproduction may be more severe at high levels of nutrient availability. As with the complex effects of pesticides, these impacts of inadvertent fertilization are likely to impact trophic dynamics and ecosystem processes.

A. Ecological Risks of Genetically Engineered Crops

In general, the impacts of agricultural technologies such as pesticides and chemical fertilizers were characterized many years or decades after the introduction of the technology and well after their use was widespread.
In contrast, ecologists began to call attention to the ecological consequences of genetically engineered crops before these crops were widely available commercially. In the past few years, however, these transgenic crops have rapidly become common in agricultural systems throughout the world, but the ecological impacts of these crops are still poorly understood. Many ecologists throughout the world are beginning to investigate these impacts in order to develop a realistic assessment of the potential risks to the ecology of both natural and agricultural ecosystems. Despite the relatively small number of studies undertaken to date, this research is having an important influence on how transgenic crops are being deployed.

Ecologists are investigating many different types of risks, including changes in the ecological competency of the crop, transgene movement from crops to wild relatives, the evolution of resistant pests, risks to non-target organisms, recombination between genetically engineered plants expressing virus resistance and natural viruses leading to new viral genotypes, and changes in crop management strategies and cropping systems. These different types of risk are receiving differing degrees of attention from ecologists.

There is evidence for changes in the ecological competency of a few crops. Through genetic engineering, the crop may acquire particular traits that allow it to survive better in natural ecosystems or as a volunteer in agricultural systems. Such traits might include seed survival, increase crop resistance to pests or diseases, or increase the competitiveness of the crop relative to wild plants. This process could lead to invasion of natural ecosystems and ecological disruption. In addition, genetically engineered traits such as herbicide resistance can make volunteer crops persistent weeds in subsequent cropping systems. This problem has already been detected with canola (rapeseed) that has transgenic herbicide resistance.

Genetically engineered traits may also move from crops to wild relatives of crops through natural processes of gene flow and hybridization. Many recent studies have examined the degree of hybridization between crops and wild relatives. Successful hybridization under natural conditions has been demonstrated for many crop–wild relative associations, and some studies have documented the persistence of crop genes in populations of wild relatives over several generations. Traits such as pest and disease resistance or herbicide tolerance may provide a selective advantage to wild relatives of crops and allow them to become weeds in agroecosystems or to increase in abundance in natural ecosystems and thereby disrupt natural patterns of plant distribution.

Increasingly, ecologists are exploring the possible effects of these transgenic crops on non-target organisms, including natural enemies and soil biota, and evidence for detrimental effects is accumulating. For example, lacewing predators fed on prey that consumed transgenic crops expressing Bacillus thuringiensis (Bt) toxins (bacterial insecticides) have higher mortality than those fed on prey that did not consume Bt toxins. Similarly, parasitic wasps that develop in Bt-treated host insects have higher mortality. Through the incorporation of crop residues into the soil, Bt toxins can accumulate in the soil and cause changes in the abundance and species composition of soil microorganisms, thus affecting decomposition and nutrient cycling.

There is much less work on the ecological implications of viral recombination, but recombination has been demonstrated several times between replicating natural viruses in host plants and transgenes that have been inserted into host genomes. It has been demonstrated that this recombination can lead to changes in viral competitiveness and pathogenicity to hosts. Recombination may also alter transmission characteristics and allow transmission by new species of vectors, which would augment the opportunities for virus establishment in new host species. Thus, the potential for ecological impacts of new, recombinant viruses arising from transgenic crops is significant and warrants more attention from ecologists.

V. FUTURE DIRECTIONS

Although the study of agricultural ecosystems has provided many insights into the role of diversity in agroecosystem function and the ecological impacts of agriculture, the subdiscipline of "agroecology" is still in its infancy. Two areas that urgently warrant more attention from ecologists are (i) the influence of planned and unplanned agroecosystem diversity on soil processes such as decomposition and nutrient cycling and (ii) the ecological consequences of widespread adoption of genetically engineered crops. A better understanding of the ecological processes operating in both areas could result in significant benefits to both agriculture and environmental protection.

See Also the Following Articles

AGRICULTURE. INDUSTRIALIZED • AGRICULTURE, SUSTAINABLE • BREEDING OF PLANTS • PARASITISM • PLANT CONSERVATION. OVERVIEW • PLANT-SOIL INTERACTIONS
Bibliography


I. Introduction
II. Net National Product
III. Environmental Kuznets Curves
IV. Technological Change and Growth Accounting

THIS ARTICLE DISCUSSES three questions: What is economic growth? Will economic growth deteriorate or improve the environment? and What is the role of environmental resources for economic growth?

I. INTRODUCTION

The title of this article has many meanings. Economic analysis of environment and growth can mean many different things. In this article, the focus is on the following issues:

• What is economic growth?
• Will economic growth deteriorate or improve the environment?
• What is the role of environmental resources for economic growth?

The first question is typically answered by growth in gross domestic product (GDP) or GDP per capita. However, seldom is it explained why this is an interesting definition of growth. Of course, behind this notion of growth is the notion that GDP or gross national product (GNP) is a measure of well-being in society. It is therefore necessary to discuss what we mean by well-being and possible indices for such a concept. This will, of course, immediately bring us to a discussion of green accounting. There is a huge literature on this subject, but in this article I will not try to survey those contributions but instead examine one particular formulation of the conceptual basis of green accounting in order to determine the difficulties in creating such an index.¹

The second question has during the past 10 years been discussed in terms of so-called Kuznets curves—that is, empirically estimated relations between environmental quality and GDP per capita. The general finding for many pollutants is that a country with a very low income does not have much pollution but when the scale of the economy grows, for example, as measured by GDP per capita, emissions of these pollutants will increase. However, when the income per capita is high enough, the economy will reach a turning point and pollution will decrease with further increases in per capita income. This has been taken by the World Bank to mean that growth implies a win–win strategy. By growing, a country may in the long run obtain both increased material well-being and a better environment. Of course, this kind of analysis has been criticized, from conceptual points of views and also from econometric and empirical points of view. Once again, the literature

¹ The discussion of these issues is based on Dasgupta and Måler (2000) and Måler (1999).
is huge, and no attempt to summarize this literature will be done. However, some central issues will be discussed.

Finally, the role of environmental resources as inputs for the growth process has very seldom been studied. When Solow (1957) studied economic growth in the United States, he found that most of the growth could not be explained by the two traditional factors of production (capital and labor), and he attributed the "unexplained" growth to the residual to technical progress. Later investigations by Dale Jorgenson, Zwi Griliches, and others reduced the residual considerably (e.g., by using better measurements of capital and by incorporating human capital), but it still remains quite high. This means that economists basically cannot account for all the productivity growth Western economies experienced during the past 100 years. To solve this problem, endogenous growth theory developed during the 1980s and the beginning of 1990s. The main idea is to incorporate the externalities that research and development activities generate because the benefits from these cannot be captured completely by the companies that carry out these activities. These externalities will therefore contribute to growth without being accounted for in the usual way that we analyze productivity growth. However, there are other externalities that also should be accounted for, such as networking externalities or social capital (Dasgupta, 1999). In this context, there is one particular externality that is of interest — environmental or resource externality. Because environmental resources or natural resources in general often lack well-defined property rights, they are not transacted on markets and therefore not accounted for in the standard national accounts (SNAs.) This omission of natural capital in growth accounting studies may explain at least part of the residual.

II. NET NATIONAL PRODUCT

Aggregate measures such as GNP or GDP were not constructed to measure welfare or well-being. They were aimed at measuring total monetary income in an economy and the total supply of goods produced in man-made facilities to provide economic planners with a database suitable for economic policies. However, despite warnings from many economists, GNP and GDP became used as indices of welfare. During approximately the past 25 years, economists have tried to define welfare and construct indices measuring welfare, and there is now a substantial literature on the subject. Instead of reviewing the literature, I will discuss some issues not very much discussed. Furthermore, most of the literature is based on the assumption that the economy is following an optimal growth path, which is rarely a realistic assumption. Instead, we will assume that we are able to make forecasts arbitrarily into the future.¹

The issues that will be discussed in this section are

- The notion of social well-being
- Social well-being and population growth
- The search for a linear index—net national product (NNP) as a marginal social cost–benefit criterion for elementary projects
- The search for a linear index—NNP as a measure of sustainable development
- Sustainable development and technical progress

A. Social Well-Being

In order to discuss social well-being, it is useful to imagine a social planner who is in charge of planning an economy from now (date 0) to infinity. Koopmans has convincingly shown (1960, 1972a, b) that the rational social planner should evaluate different policies with regard to how they affect the present value of the stream of future utilities, where the utility function is strictly concave and is a function of all the factors that influence the current well-being. In other words, the social welfare function is written

\[ W = \int_0^\infty e^{-\delta \tau} u(c(\tau)) d\tau \]  

(1)

where \( c(\tau) \) is the rate of consumption at moment \( \tau \). In principle, \( c(\tau) \) is a vector, the components of which represent all the factors that affect well-being at time \( \tau \). However, it will not change any conclusions if we restrict ourselves to the interpretation that \( c \) is the aggregate consumption good.

One of Koopmans’ condition is that \( \delta > 0 \); that is, the discount rate must be strictly positive. The discount rate is necessary in order to be able to define a social well ordering over all possible future paths of utilities. However, this is the rate for discounting utilities. The

¹ The basic reason behind the assumption of an optimal growth path in much of the literature is that predictions become very simple. It will be shown that nothing essential will change when we switch to arbitrary allocation mechanisms.

Jorgenson (1995a, b) provides masterful surveys of these issues.
rate for discounting consumption will in general be different and depends on the growth rate of consumption. The formula is

$$r = \delta + \eta g$$

(2)

where \(r\) is the consumption discount rate, \(\delta\) is the utility discount rate, \(\eta\) is the elasticity of marginal utility of consumption, and \(g\) is the growth rate of consumption. It is clear that if \(\delta\) is small and the growth rate is sufficiently negative, the social planner will want to discount future costs and benefits measured with consumption as a numeraire (i.e., the way we usually measure them in cost–benefit analysis) with a negative interest rate.

The strict concavity of the utility function guarantees an attitude in favor of equalization between generations. Holding everything else unchanged, a transfer from a rich generation to a poor generation will increase social welfare.

1. Mathematical Preludes

No in-depth mathematics will be used in this section and an elementary knowledge of calculus should be enough for the derivations made here. However, reading this section is not necessary for an understanding of later sections.

At time \(t\), the economy has inherited assets from the past. Denote the vector of these assets \(\mathbf{K}_t\). Assume that the social planner can predict the future flow of consumption, given \(\mathbf{K}_t\). In order to do this, the social planner must know how these assets are going to be used (now and in the future) for consumption and for accumulation of some of the assets and perhaps depletion of others (in particular, stocks of nonrenewable resources). Such predictions can be summarized by an allocation mechanism, which will be called \(\alpha\):

$$c(\tau) = \alpha(\tau, t, \mathbf{K}_t), \forall \tau > t$$

(3)

Thus, the mapping \(\alpha\) is the information the social planner uses for mapping forecasts.

Note that \(\alpha\) is a function of the initial stocks but also of calendar time \(t\). The reason for this is that there may be autonomous changing over time. If there is exogenously technical progress, for example, future predicted consumption will of course depend on this and \(\alpha\) must therefore depend on calendar time. Other reasons for autonomous changes in the allocation mechanism may be trends in trade and forecasted changes in government policies. I will discuss these issues later.

Given the allocation mechanism, the social welfare beginning at time \(t\) will be a function of the capital stocks at \(t\), the allocation mechanism, and the calendar time. Thus,

$$W = V(t, \alpha, \mathbf{K}_t)$$

(4)

\(V\) gives the social welfare achievable from the initial stocks \(\mathbf{K}_t\) with the allocation mechanism \(\alpha\).

A marginal change in one of the stocks at time \(t\) will change the welfare, and this change is basically the marginal value of that stock; that is,

$$p_i(t) = \frac{\partial V(t, \alpha, \mathbf{K}_t)}{\partial K_i}$$

(5)

where \(K_i\) is the stock of asset \(i\) at time \(t\). In most of the literature on green accounting, the accounting prices are those that support an optimal growth path. The prices defined here do not. However, they reflect the true values of the initial stocks at time \(t\).

It is practical to define the Hamiltonian \(H\) at time \(t\) as

$$H_t = u(c(t)) + \sum_{i=1}^{N} p_i \frac{dK_i}{dt}$$

(6)

One can now show that

$$\frac{dp_i}{dt} = \delta p_i - \frac{\partial H_t}{\partial K_i}, \quad i = 1, 2, \ldots, N$$

(7)

Although these equations are formally identical to the equations governing the costate variables in Pontryagin's maximum principle, they do not reflect any optimization except that the prices are the true measures of the marginal social values of the initial stocks.

B. Population and Welfare

So far, I have neglected population changes in the discussion of social welfare. However, we know that humans have a finite life and new generations will be born. How do we take this into account when we try to formulate a welfare function for a society?

The simplest way is perhaps to assume that population changes exogenously in a predictable manner. Then, at each moment in time we can represent the welfare of the cohort by the utility of an individual with average per capita consumption multiplied by the number of individuals. However, this is seriously flawed

* See Mäler (1999) for a proof.
in that fertility behavior is not exogenous. We now know that human reproduction behavior is governed by many social and economic factors. There are many studies showing that reproduction depends on the need for support during retirement, on the need for labor, on the need for risk diversification, and of course because children can be regarded as consumption goods—that is, they give immediate welfare to the parents. Dasgupta (1993, 1997) provides discussions of these issues and references to empirical material. Thus, we should treat fertility as endogenous, something that depends on the economic system in all its characteristics. However, this is exactly what the allocation mechanism introduced in the previous section represents. Thus, the population at any particular future point of time is determined by $\alpha$.

It is now easy to assume that the welfare function can be written

$$
\int_0^\alpha e^{-\beta L(\alpha(0), K_0)} u(c(t)) \, dt
$$

(8)

However, this formulation does not take into account the fundamental difference between a potential human being and an actual human. What is the value of having one more individual to that individual? There is no reasonable answer to this question [see Dasgupta (1997) for an extremely interesting discussion of these issues]. So far, we do not have an ethical base for assessing welfare to endogenous population changes.

C. A Marginal Cost–Benefit Rule

Define NNP with utility as the numeraire as the linearized Hamiltonian

$$
\text{NNP} = u'(c) c + \sum_{i=1}^n p_i \frac{dK_i}{dt}
$$

(9)

If, instead, we choose consumption at time 0 as the numeraire, NNP is defined by

$$
\text{NNP} = c + \sum_{i=1}^n \tilde{p}_i \frac{dK_i}{dt}
$$

(10)

where $\tilde{p}_i = p_i / u' (c(0))$.

Note that NNP or å is a linear index in all quantities, which makes it an interesting measure. A non-linear measure such as the Hamiltonian is hopeless in terms of empirical applications. We cannot estimate all the consumers' surpluses needed to get an empirical estimate of the Hamiltonian.

Second, note that $\epsilon$ is basically a proxy for all variables that affect current utility. Therefore, all these variables must be included in a generalized consumption set. Among these variables is the disutility of labor, which leads us to the conclusion that the wage bill, as long as the wage rate reflects the disutility of work, should be deducted. Immediate environmental damages should also be deducted. In all these cases the damages should be valued at a rate that reflects the marginal disutility of the environmental damage with consumption as the numeraire.

Third, note that the investment part of the NNP must include net changes in all assets that will affect future well-being. This means of course that the current narrow asset boundary in SINs must be substantially extended.

Also note that the asset prices $p_i$ must correctly reflect social the scarcity values captured by Eq. (7). These prices will not always coincide with market prices. However, surprisingly often, market prices will capture the social scarcity values of assets that are transacted on markets. However, for all the ecological assets, the values must be estimated by modeling the ecological systems and their connectedness. There is a growing literature on these procedures.

It should be noted that the NNP measure so defined deviates radically from the United Nations recommendations. The NNP measure we have defined has one important property, namely, that if there is a small perturbation (so that price changes are small) of the allocation mechanism during a small time interval, and if this small perturbation increases the social well-being, then it will also increase NNP and vice versa. This is what makes NNP interesting—a linear marginal cost—benefit criterion with which we can evaluate such perturbations (which henceforth will be called policy reforms).

Here, we consider discreet time, and the following example will indicate the use of the theorem stated previously. Assume that there is one allocation mechanism $\alpha$ and we consider a perturbation or policy reform $\Delta \alpha$ that for the first period presupposes the same allocation as $\alpha$ did for the second period. We compute the change in NNP when we make this policy reform and this change is positive. This can be interpreted as saying that the second period gives higher social welfare than the first period. By following the allocation mechanism.
$\alpha$, welfare increases over time. However, note that this depends on the assumption that the prices do not change much.

This theorem can now easily be generalized to situations in which we take atemporal equity into account, and even population changes can be accommodated. The trick is to find the appropriate shadow prices.

D. The Hamiltonian as Constant-Equivalent Utility

In the previous two sections it was shown that NNP can be used as an index for conducting social cost-benefit analysis of policy reforms. However, the theoretical literature on green NNP has been directed toward a quite different end (Weitzman, 1998). It has argued that NNP measures "constant-equivalent consumption."

It can be shown that the present value of a constant stream of utility equal to the Hamiltonian at time zero is the maximum feasible constant utility stream that an economy can afford. This was first shown by Weitzman in the case in which the utility function is linear. In this case, NNP and the Hamiltonian coincide, and NNP gives the maximum constant equivalent consumption.

However, a linear utility function is ethically flawed: It is insensitive to distributional issues. Furthermore, a large body of evidence concerning household saving behavior is at odds with linear utility functions.

E. Social Well-Being and the Concept of Sustainability

The World Commission (1987) defined "sustainable development" as an economic program in which, loosely speaking, the well-being of future generations is not jeopardized. There are many possible interpretations of this. Consider the following:

a. An economic development is sustainable if $dU/dt > 0$, where $U_0 = \lim U$, as $t \to -\infty$.

b. An economic development is sustainable if $dV/dt > 0$.

c. An economic development is sustainable if $dV/dt > 0$, where

\[ V_t = \int_0^\infty e^{\alpha U(C_t)} \, dt \]

It is clear that a lacks ethical foundation. For example, it may be desirable to reduce $U$ in the short run in order to accumulate assets so that the flow of $U$ is still higher in the future. In this sense, b offers greater flexibility in ethical reasoning: It permits initial sacrifices in the current standard of living, $U$ (a burden assumed by the generation engaged in the reasoning), but requires that no future generation should have to experience a decline in its standard of living.

In contrast to c, the focus of b. as a notion of sustainable development is social well-being, $V$. The criterion permits the first generation to make initial sacrifices in $V$ (relative to the past) but requires that social well-being should never decline in the future. Note that although b implies c, c does not imply b. In other words, c. is more general. In what follows, we adopt c. as our notion of sustainable development and develop criteria for judging whether a given economic program represents sustainable development.

One can now easily show that

\[ \frac{dV}{dt} = \frac{dV}{dt} + \sum p_i \frac{dK_i}{dt} \]

The first term on the right-hand side is the autonomous change in $V$, reflecting exogenous technical progress or exogenous changes in terms of trade. We will discuss this term in the last section. For now, we assume that there are no autonomous changes. The previous equation simply indicates that the rate of change in welfare is equal to the value of the net investment in the economy. Thus, an economic program represents sustainable development if, and only if, the program net investment in the economy's capital assets is always nonnegative.

The result has intuitive appeal. It says that social welfare is higher today than it was yesterday if the economy is wealthier today. Here, an economy's "wealth" is interpreted as the accounting value of all its capital assets, and wealth comparisons are made at constant prices. In a famous article, Samuelson (1961) argued in connection with national income accounting that welfare comparisons should deal with "wealth-like" entities. Proposition 4 formalizes this insight.

Note, however, that what we have obtained is an equivalence result: The theorem cannot on its own tell us if sustainable development is feasible. Whether the economy is capable of growing wealthier indefinitely depends on, among other things, the extent to which different assets are substitutable in production.

This idea of examining changes in wealth dates back to Samuelson (1961). The connection to sustainable development was first suggested by Pearce and Atkin-
son (1993), although they did not derive the theorem rigorously and they did not interpret it as an equivalence result. The idea was then investigated by the World Bank, which published many attempts to measure the value of net investment or genuine savings [see Expanding the Measure of Wealth (1997) and Kunte and Clemens (1998) for further details.]

Note that we cannot in general use NNP to determine whether the economy is on a sustainable path. Only if the prices are stationary is an increasing NNP equivalent to saying that the economy is on a sustainable path.

In conclusion, NNP can be used for cost benefit analysis in situations in which prices do not change very much. In particular, in such a situation it may give some guidance as to whether social welfare is increasing or not. Note that there is no conflict between growth in the NNP and the concern for the environment. These concerns have been incorporated into the linear index. If the economy is on a path in which environment is degraded, this will be captured by the NNP, and it can be expected that NNP will decrease if the degradation is sufficiently large.

NNP is not useful for determining whether the economy is on a sustainable path because the relative prices will change along the path. Here, concepts related to wealth become more interesting. In particular, the value of the net investments in a time period gives a correct criterion for judging whether the economy is on a sustainable path. However, it is only a necessary condition. The sufficient criterion is that in each future period, the value of net investments must be positive. A priori, we do not know whether this is feasible.

III. ENVIRONMENTAL KUZNETS CURVES

Since the early 1990s, economists and econometricians have been industrious in estimating empirical relations between GDP and environmental quality indicators. Typically, they have used cross-sectional data—country data for a particular year. For some quality indicators, they have found a relation that looks like an inverse U (Fig. 1).

The interpretation is simple. Very poor countries with low industrialization are not damaging the environment much and as a result the environmental quality indicator is good (in this case, a low sulfur emission). When the country starts the growth process, the environmental damages will increase. When income per capita increases, the demand for a cleaner environment will increase, environmental legislation will be enacted, and the environment will improve. Because the curve is similar (but inverse) to the curves Kuznets (Economic Growth and Income Inequality, 1955) estimated for the relationship between income inequality and GDP per capita, they have been known as the environmental Kuznets curves. For an overview and examples of environmental Kuznets curves, see Environment and Development, 2 (4), October 1997.

The breakthrough for this kind of analysis came with the World Development Report (1992), which was devoted to the environment. The existence of these curves was taken to be an indication of possible win-win situations. By stimulating growth, one would also in the long run stimulate environmental regulations. It is clear, however, that the analysis is quite muddy complex.

First, environmental Kuznets curves have only been established for a few environmental indicators. Second, at least earlier, the specifications of the relation between the quality measure and GDP per capita were quadratics, which almost certainly forced the curves to look like an inverse U. Third, the studies did not include environmental spillovers between countries. Fourth, they did not examine possible relations between indices of freedom and the environment. It has been shown that whenever people have had the right to voice concern about environmental degradation, the environment has improved, irrespective of the per capita income. Locally, in many cultures people have had a direct influence on, for example, local water pollution, which has improved even in very poor countries. It is therefore possible that the environmental Kuznets curves measure not the effect of income growth but growth in freedom.

However, most important is that the environmental Kuznets curves—if they exist in any meaningful way—
are not optimal in any sense of the concept. Typically, a country would do much better to develop institutional and regulatory reforms for better environmental management than just trust that economic growth will automatically solve environmental problems. In this sense, the belief in environmental Kuznets curves and the associated belief in win–win situations are very dangerous.

IV. TECHNOLOGICAL CHANGE AND GROWTH ACCOUNTING

Productivity growth as discussed previously is productivity growth in GDP in which the flow of services that are not transacted on markets are excluded (with of course some important exceptions, e.g., government output). Here, we are interested in productivity growth in NNP as defined previously. There is no reason why productivity growth in GDP would correspond to a similar growth of NNP. Furthermore, if, as Dale Jorgensen (1995a, Chap. 1) pointed out, changes in energy prices affected the observed productivity growth rate substantially, it should be clear that the utilization of other natural resources should also affect the growth rate of the total production. Thus, we should analyze the following issues:

• Growth in NNP
• The role of natural capital in the growth of NNP

The following calculations are very speculative, but I hope that some day someone will find it worthwhile to study the empirical consequences from the analysis. I use the simplest possible model to demonstrate the main ideas (which probably will be seen to be perhaps too obvious).

Let there be one aggregate good \( x \) which is produced with man-made capital \( K_e \) and one input representing natural capital services, denoted \( y \). Thus,

\[
x = f(K_e, y)
\]

Assume that the production function \( f \) is homogeneous to degree 1.

The total output is used for consumption \( c \), for investment in man-made capital \( I = K_e \), and for investment in man-made capital for exploiting the natural resources \( t = K_n \). Let the production function for \( y \) be

\[
y = g(K_n)
\]

Thus, \( y \) is an intermediary good in the system and \( x \) will correspond to GDP:

\[
x = c + K_e + K_n = c + K
\]

Assume that capital is freely mobile between the two uses so that we can reasonably talk about the total capital stock \( K = K_e + K_n \). Then we can define the productivity growth in GDP as

\[
\theta = \frac{1}{x} \frac{dx}{dt} - \frac{1}{K} \frac{dK}{dt}
\]

Now, let the dynamics of the natural or ecological capital \( K_e \) be

\[
\frac{dK_e}{dt} = \phi(K_e) - y
\]

Assume that there is an institutional failure in the economy so that there are no assigned property rights to natural capital and therefore no market transactions in \( K_e \). This is of course the reason why investments in natural capital are not included in GDP.

NNP according to the definition in equation is

\[
NNP = c + \frac{dK_e}{dt} + \frac{dK_n}{dt} = x + \frac{dK_e}{dt}
\]

It is clear from Eq. (16) that a positive productivity growth for GDP can be offset by depreciation of the stock of natural capital so that the productivity growth of NNP is zero or even negative. In fact, if we assume that \( K = 0 \), we have

\[
\frac{1}{NNP} \frac{dNNP}{dt} - \frac{1}{K_e} \frac{dK_e}{dt} = GDP \theta + \frac{K_n}{K_e} \left\{ \frac{p_t K_n}{K_e} \frac{K_n}{K_e} - 1 \right\}
\]

(17)

The last bracket in Eq. (17) will surely be negative because of the overexploitation of the resource as a result of the institutional failure. Thus, this bracket will compensate for the positive productivity growth of GDP.

In this case, because we know that the production func-
tions are stationary, we know for sure that the productivity growth rate of NNP (if the production functions have constant returns to scale in $K$, $L$, and $E$) must be zero. We can therefore conclude that even if measured GDP growth rates show a positive productivity growth, this does not necessarily imply that NNP will show growth irrespective of factor inputs.

This result also follows from the observation that conventional SNA regards the inputs of resources as intermediary goods (if there are costs in producing them and it neglects them completely if they are supplied free of any costs to the production processes), and the estimation of an aggregate production function with only labor and capital as inputs, necessarily will be biased because resources are the basis for material production. Thus, the total factor productivity of GDP should be measured as

$$\frac{1}{x} \frac{dx}{dt} = w_K \frac{1}{K} \frac{dK}{dt} + w_L \frac{1}{L} \frac{dL}{dt}$$

(18)

where $w_K$ is the value share of man-made capital and $w_L$ is the value share of the resource input. This will necessarily be zero if the production function is linearly homogeneous and stationary. Thus, it may be that the neglect of resource inputs in the production functions is an important explanatory variable of the measured positive factor productivity growth. However, the only way to find out whether this is important is to carry out new empirical measurements. In conclusion, it does not seem reasonable to assume autonomous productivity growth in models designed to create a new accounting system for welfare measurements. Therefore, the argument for a nonstationary value function is not necessarily true.

See Also the Following Articles

- Biodiversity as a commodity
- Economic Value of Biodiversity, Overview
- Economic Value of Biodiversity, Concept and Practice of Biodiversity, Measurements of Market Economy and Biodiversity, Sustainability, Concept and Practice of

Bibliography


ECONOMIC VALUE OF BIODIVERSITY, MEASUREMENTS OF

Robert Mendelsohn
Yale University

I. Introduction
II. Value to the Economy
III. Quality-of-Life

GLOSSARY

biodiversity Number of species on the planet.
local diversity Number of species or ecosystems in a small geographic area.
market value Benefit of wildlife or species as an input into the economy.
quality-of-life value Benefit of wildlife or species to people’s well-being or happiness.
williness-to-pay Expression by people of strength of preference in dollar terms.

BIODIVERSITY IS USED IN MANY DIFFERENT WAYS including the number of species on the planet, the number of species in one place, the number of individuals of desired species, the genetic diversity of a given species, and the ecological diversity of ecosystems. Each of these meanings implies a set of different values. This article reviews how we measure these different values and provides an overview of the values that have been measured to date. There are two broad types of biodiversity values. Market values reflect how biodiversity serves as an input to economic activities such as agriculture, construction, and medicine. Quality-of-life values reflect ways that biodiversity makes our lives richer simply by increasing our appreciation of life itself.

1. INTRODUCTION

Before one can describe the values that society places on biodiversity, it is first necessary to define what biodiversity means to people. Unfortunately, the term has come to mean many different things, which only complicates valuation. A narrow definition is that biodiversity represents the number of species on the planet. Hence the loss of a species diminishes biodiversity. This definition clearly links biodiversity with endangered species. Biodiversity has also been used to compare one place to another. Local diversity measures this type of biodiversity: places with more biodiversity have more species per hectare. This definition recognizes the value of complex ecosystems, as well as the value of a variety of ecosystems across a landscape. A third definition of biodiversity reflects the value of all living things: a place with more songbirds, more deer, more butterflies, and more trees is a “better” place. In this case, biodiversity is a pseudonym for the overall population. The term biodiversity means all of these things to society, making a precise measurement of the value of biodiversity difficult.

There is no single value of biodiversity since there is no single definition of what the word actually means. Biodiversity values are really only clear as one moves
from the general term to specific questions. Rather than
determining a single value of biodiversity, the literature
on valuation has attempted to measure how important
specific aspects of nature are to people. For example,
what would people pay to prevent a specific species
from going extinct? What would people sacrifice to
keep more complex ecosystems such as mixed-species
forests in place? How much income would people give
up to have a variety of ecosystems across the landscape
rather than having only one or two? What would people
pay to increase the population of particular creatures
in their local woods?

Biodiversity values have often been further divided
into two types: market and quality-of-life values. The
market values reflect the interaction between nature
and the economy, whereas the quality-of-life values
involve interactions directly between nature and people.
Methods to measure market values are well developed
and we consequently know more about market values.
Quality-of-life values are more difficult to measure.
There are many of such values we have just begun to
understand, much less measure. Although these quality-
of-life benefits are technically difficult to measure,
they cannot be dismissed because they are likely to
be the primary motivation for protecting species and
wildlife populations.

The remainder of this article is divided into a review
of biodiversity values for market and quality-of-life ser-
tices. In each case, we begin discussing the methods
and then review existing measurements. For both mar-
ket and quality-of-life, we examine species diversity,
local diversity, and natural populations. The article con-
cludes with an overview of what has and has not been
measured, discusses research priorities, and briefly
hints at policy implications.

II. VALUE TO THE ECONOMY

Biodiversity is one of many inputs into the economy.
From medicine to timber, biodiversity provides raw
materials for many economic processes. Although these
inputs are not always traded in markets, they nonethe-
less fuel market activities. By valuing the resulting mar-
ket activities, one can deduce the value that each input
contributed. Of course, one must be careful to value
all inputs to each economic process. One cannot assign
the market value of outputs to just one of the many
inputs required to make a good. For example, plant
biodiversity may be an input to pharmaceutical research
but it is only one of many inputs, including laboratories,
technicians, and clinical research. In this section, each
of the multiple definitions of biodiversity is examined to
determine what value they contribute to the economy.

A. Diversity

The diversity of species on the earth is of economic
value because the diversity offers many alternative ser-
tives. For example, the diversity of plants has been
used to find cures for diseases. Plants, in their constant
struggle for survival, have developed numerous de-
fenses to limit consumption by predators. Some of these
chemicals have important medical qualities. Many ex-
isting pharmaceuticals have come from plants. Some
47 drugs have come from tropical forest plants alone
(Soejarto and Farnsworth, 1989). Given that there are
250,000 plant species in the world, there is a strong
possibility that additional undiscovered drugs lie hid-
den in this diversity. Assuming that one out of a million
samples contains a hidden drug, there may be as many
as 325 useful drugs still hidden in tropical plants alone
(Balick and Mendelsohn, 1995).

What is the value of these hidden drugs? The drugs
themselves are worth billions of dollars. A successful
drug is worth about $94 million to a drug company
(OTA, 1993). Such a drug is worth about $450 million
to society when one includes the resulting health ben-
fits, taxes, and revenues beyond the patent period (Bal-
ick and Mendelsohn, 1995). Multiplying both figures
by the number of hidden drugs suggests that private
companies would pay $30 billion and society would
pay $145 billion for these drugs. This is the value of
the drugs, but not the value of the biological component
since many other inputs went into creating each drug.
For example, finding the samples, identifying active
chemicals, and developing a drug that can be marketed
costs over $100 million per drug. Pharmaceutical com-
panies should not be willing to pay much for samples
at all since the cost of developing new drugs from these
samples is likely to be greater than the expected reve-
uues. In contrast, society would still find the drugs
valuable since society obtains more benefits from the
drugs than companies can get. Society would be willing
to pay about $40 billion for these samples. When spread
over 125,000 plants, this amounts to about $320,000
per plant. Allocating this value to each of the 3 million
hectares of tropical forests suggests that each hectare
is worth an average of about $13. Of course, once the
plants on a particular hectare are investigated, they will
no longer have any more value in this regard. This
raises serious questions about how to collect these reve-
uues since many landowners will compete to be the
one and only collector of each plant.
Diversity may also be of value to crops and trees because it yields a broad gene pool. Although this gene pool may not be needed today, it could serve as a resource against future pathogens and problems. Because farmers and foresters rely on improved varieties, many scientists are concerned that the crop gene pool has become too narrow. Despite the success of modern farms and forests, these scientists argue that farmers need more diversity and thus it is imperative that we protect plants in the wild that are related to modern crops. Similar concerns are raised to protect old varieties that have fallen out of favor. It is possible that noncommercial plants related to valuable crops could be valuable in the future in order to create a broader gene pool to be used against future pathogens. However, it is hard to quantify the market value of this potential since most modern crop varieties appear to be doing well in relying on a more select gene pool that is already well protected. It would be useful to quantify the additional value of sustaining marginal crop varieties against some future date when they might be needed. The known cost of sustaining varieties that have little value today would have to be weighed against the uncertain benefit in the future when they might be needed.

### B. Local Diversity

Local diversity has often been championed as a defense against the ravages of nature. Whether one is considering hurricanes, ice storms, and other severe weather or whether one is concerned about insects, rust, or viral attack, ecologists argue that more diverse systems are more robust. For example, by planting a diversity of crops, a farmer can protect his or her income from weather, insects, and pathogens. Even if some crops are destroyed, the other crops will likely survive, thus assuring at least some income. Modern farms tend to plant a wide set of crops to protect against uncertainty from varying market prices as well as the adversities facing all farmers. Similar schemes can protect forests as well. By growing trees from a wide range of species and in a number of age groups, foresters can reduce the likelihood of catastrophic events.

Some efforts to combine forests and farms have also been practicing local diversity. Windbreaks across flat areas have been used as a way to protect crops from high winds. Mixed agroforestry methods have been used in the tropics to provide a portfolio of outputs (Padoch et al., 1985) as well as to sustain soils (Montagnini and Mendelsohn, 1997). These land use planning practices have been successful in reducing some of the damages that are more easily afflicted to a homogeneous landscape.

### C. Population

Sometimes plants and animals are a direct input to the economy. In this case, biodiversity represents the size of relevant populations. The most significant biological population to humans lies in the agricultural sector. Approximately 1.9 billion tons of grain, 0.6 billion tons of tubers, and 1.5 billion tons of sugar crops are grown each year. In addition, approximately 181 million tons of meat are harvested annually. Agriculture accounts for approximately 6% of the world’s gross domestic product, or about $1.2 trillion overall. This number reflects the marginal value of food. Of course, people would pay a lot more for food rather than be without it entirely. Assuming that people would pay twice as much for food if only half of the food was available, the all-or-nothing value of food would be $2.4 trillion.

Another economic sector that depends on raw materials from nature is forestry. The timber industry harvests trees from more or less natural settings and converts them into lumber and paper products. For the last 200 years, the industry has relied heavily on cutting natural stands of trees. Originally, this was done unsustainably as the industry harvested the vast stands of old growth around the world. As these stands have become rare, much of the industry has shifted to growing trees on plantations. The worldwide industry is approaching sustainability as regeneration now almost balances harvest rates. Once this balance is reached and maintained, trees can be harvested indefinitely from a given land base. The timber industry currently harvests 1600 million cubic meters every year (IPCC, 1996). Over half comes from temperate forests (890 million m³), another fourth from boreal forests (430 million m³), and the remainder from tropical forests (280 million m³). In contrast, most of the world’s land in forest is currently tropical (27.6 million km²) and boreal (13.8 million km²), with only 5.4 million km² in the temperate region (IPCC, 1996). The timber industry is most dependent on the productive lands in the temperate and subtropical parts of the globe.

This industry produces an annual output worth $418 billion (IPCC, 1996). However, a large fraction of this estimate is value added from production. The value of the trees themselves is more difficult to estimate. Assuming that the value of harvesting an additional tree, the stumpage value, is about $20/m³, the annual market value of the 1600 million m³ would be about $32 billion per year. This is the marginal value of the harvest.
The most abundant physical product from the forest is not industrial wood but rather fuelwood and charcoal. Each year approximately 1750 million m$^3$ of wood is used for energy. The vast majority of this use occurs in tropical regions, which rely on 1370 m$^3$ for energy. Temperate forests produce 20% of the wood energy (370 million m$^3$) and boreal forests produce very little (10 million m$^3$) (IPCC, 1996). Although the use of wood for energy consumes more forest biomass than any other use, it is not the most valuable activity. Wood is often used by communities that cannot afford more expensive sources of energy (such as fossil fuels). The average price of a cubic meter of wood for biomass is probably in the neighborhood of about $5/m^3$. The global market value of fuelwood is around $9 billion annually.

In addition to these foregoing products from trees, there is a host of non-timber forest products (NTFPs) that people enjoy around the world. The more diverse the forest, the more NTFPs that are present. Yet even in relatively simple northern forests, people enjoy wild rice, berries, mushrooms, boughs, maple sugar, and food resins. In the tropical zone, forests produce natural medicines, fodder, fruits, latex, rattan, roofing material, dyes, and other wood resins. It is difficult to quantify the global value of these products since most are traded only locally and so escape official enumeration. However, local studies of selected tropical forests suggest that these NTFPs are often as valuable as timber (Peters et al., 1989; Balick and Mendelsohn, 1992; Grimes et al., 1994; Houghton and Mendelsohn, 1996; Muntz-Miret et al., 1996). As a rough approximation, it is likely that the value of NTFPs is equal to the value of timber from tropical forests, or about $6 billion each year.

III. QUALITY-OF-LIFE

Although biodiversity makes a large contribution to the world’s economy, it is likely that the most important services that humans receive from nature are to our quality-of-life. Many aspects of biodiversity are not used by the economy and so provide no market value. Nonetheless, if we were to lose these species and populations, people would suffer a great loss. People clearly enjoy nature and obtain a great deal of pleasure from it.

A. Diversity

The earth contains an amazing diversity of plants and animals. Approximately 1.5 million living species have been identified. Almost 1.25 million of these species are arthropods, or insects. Scientists estimate that there are as many as 250,000 flowering plants, 80,000 fungi, and 100,000 mollusks. Only 30,000 species are vertebrates. Although there have been some major mass extinctions in the geological record, the number of higher-order species that have perished in industrialized nations is few. In fact, the vast preponderance of species lost today are the assumed losses of insects from the clearing of tropical forests. Scientists predict that each of these hectares likely contains a certain number of unique insects that are lost when the trees are removed. This is a predicted loss, since no one has actually identified the insects in these forests that have now disappeared.

Surveys of public opinion clearly indicate widespread and strong interest for species protection. Despite frequent attacks from land developers, business groups, and some industries, public support for endangered species and other wildlife continues to be strong. However, public preferences for which species deserve protection are complex. For U.S. society is not equally concerned with all species. People express the strongest preference for flagship or charismatic species (Kellert, 1984). These include North American animals such as the grizzly bear, black-footed ferret, Florida panther, and bald eagle. International examples include lions, elephants, tigers, rhinoceroses, seals, monkeys, and apes. In general, people appear to be willing to pay the most for these charismatic animals. For example, using a survey on the willingness to pay, Brookshire et al. (1983) found that people would pay between $10 and $26 to protect grizzly bears and Boyle and Bishop (1986) found that eagles were worth between $10 and $31 per person. Despite the broad enthusiasm for species protection, it is not clear that people are willing to sacrifice very much to protect lower-order animals and plants. For example, Boyle and Bishop (1986) found that people were only willing to pay between $4 and $5 to protect the striped shiner, a small fish. Valuation studies have not specifically determined a value for a broad class of lower-order animals. It is most likely the case that people would pay for their protection as part of a larger plan to protect entire ecosystems. However, the valuation of entire ecosystems has not yet been done.

B. Local Diversity

Scientists have determined that certain ecosystems have more species per hectare than others. For example, species diversity generally rises as one approaches the
equator from either pole. The highest number of endemic species per hectare tends to be in tropical locations. It is possible that people have preferences concerning whether they would like to have more species in the ecosystems that they live in. Unfortunately this has not been tested, so it is difficult to know. It is clear that certain “hot spot” locations with high numbers of endemic species have become desirable travel destinations. Madagascar, the Philippines, and the Galapagos Islands are all sites where ecotourists flock to see the high numbers of endemic species.

Local diversity can also mean that there are many ecosystems within a small region. Sometimes terrain can create small microsites with unique climates. For example, mountainous regions near the coast can have a wide variety of ecosystems in a relatively small area. This diversity gives people many landscapes to enjoy locally. Although most of this diversity is a natural phenomenon, it can be affected by land use patterns. For example, if people drain wetlands, the only land remaining may be an upland ecosystem. By draining the wetlands, people have simplified the landscape. The many recent laws to protect wetlands and other ecosystems suggest that this diversity has considerable value to people. Curiously, little valuation research has been done on the value of heterogeneous landscapes.

C. Populations

There is no question that people value local wildlife populations. A survey conducted by the U.S. Fish and Wildlife Service (1991) found that approximately one-half of Americans participated in some wildlife-related recreation. Of these, 76 million people enjoyed nonconsumptive activities such as bird watching, 35 million people fished, and 14 million hunted. Throughout America, people appear to derive great pleasure from having wildlife around them. Even with such common creatures as squirrels, sparrows, and pigeons, people are entertained by chance encounters with animals.

Economists have assigned values to wildlife interactions by using survey methods. For example, the techniques used to study endangered species protection (Brookshire et al., 1983; Boyle and Bishop, 1986) have also been applied to places and activities such as the Maine woods, duck hunting, and goose hunting.

Economists also measure values by observing the travel preferences of people. Because many people interact with wildlife by taking trips to outdoor sites, their choice of sites reveals how much they want to see different things. For example, with the simple travel cost method, visitation rates are compared with distances. The farther someone is willing to travel to see a site, the more valuable the site is. The relationship between visitation rates and distance reveals how much people are willing to pay to see the site. By calculating the area under this visitation function minus actual travel costs, one can determine the value of the site for a person. The total value of the site would be the sum of the values for all the visitors (Clawson and Knetisch, 1966; Brown et al., 1983). This model has also been extended to examine a set of sites (Burt and Brewer, 1987; Cichetti et al., 1992).

A weakness of the simple travel cost method, however, is that it values sites only in their entirety. One of the questions facing land managers is the valuation of changing site characteristics. For example, a manager might consider harvesting some old growth at a site or increasing deer or elk populations. To place a value on site characteristics, one must examine the set of sites that people choose from. Two techniques have been developed to value site characteristics: discrete choice and hedonic travel cost. The discrete choice method looks at all the available sites and determines a quality index that can explain why people choose particular places more often (Morey, 1981). The hedonic travel cost method examines the attributes of the places people choose and tries to estimate the marginal cost or value that people assign to each attribute. The hedonic travel cost then goes further and tries to estimate a demand function for each attribute (Brown and Mendelsohn, 1984).

The literature on valuing site attributes indicates that natural site qualities are important to people. For example, steelhead trout anglers care a lot about the number of fish in a stream (Brown and Mendelsohn, 1984). In Washington state, they would pay an average of $5 per trip for just 1 more fish per day. Hunters care a lot about the number of deer per acre. Urban hunters would pay an extra $4 for one more deer per square mile in Pennsylvania (Mendelsohn, 1984). Hikers care about old-growth forests, clear-cuts, and the size of open areas in wilderness hikes in Washington (Englin and Mendelsohn, 1991). Their consumer surplus per trip is about $70 for old growth, $14 to avoid clear-cuts, and $67 for large rock and ice areas. This literature also revealed that what people would pay for certain characteristics declined as they experienced more of them. Although people would pay $88 to avoid a 20% reduction in the steelhead population at an average site, they would pay very little for more steelhead at a site where there is already a large population. The quality-of-life value of wildlife populations consequently appears to depend on their abundance.
See Also the Following Articles

Aesthetic Factors • Economic Growth and the Environment • Endangered Mammals • Market Economy and Biodiversity • Plant Sources of Drugs and Chemicals • Timber Industry • Tourism, Role of

Bibliography


ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW

Partha Dasgupta  
*University of Cambridge and Beijer International Institute of Ecological Economics*

I. The Subject
II. Market Failure
III. Institutional Failures and Poverty: Global vs Local Environmental Problems
IV. Valuing Resources and Evaluating Projects
V. Biodiversity: Necessity or Luxury?
VI. Substitution Possibilities
VII. Ecosystem Bifurcations and Discontinuous Value Functions

---

GLOSSARY

**accounting price** In this context, a measure of the economic value of a resource in social terms, described as the increase in social well-being that would be enjoyed if one unit more of the resource were made available without cost, and expressed as the difference between the market price of the resource and the tax or subsidy that ought to be imposed on it.

**bifurcation point** A transitional boundary in the state of a dynamic system (such as an ecosystem), at which the stability or basic nature of the system changes due to alterations in its fundamental characteristics.

**intrinsic value** A measurement or description of the values inherent in a resource by virtue of its existence (e.g., a living organism), as opposed to its value for present or future use.

**option value** A variation on the concept of USE VALUE (see below); a measurement or description of the potential value of a resource for possible future use.

**net social benefit** A measure of the social benefit of a project, obtained by identifying the projects positives (favorable outputs) and its negatives (inputs, unfavorable outputs), then multiplying these by the accounting price of each commodity involved. The net result of comparing these positive and negative values will indicate the social benefit of the project.

**separatrix** Locus of points on the state space of a dynamical system separating different basins of attraction.

**shadow price** Another term for ACCOUNTING PRICE.

**substitution possibility** The possibility that the depletion of a given resource can be compensated for by one or more of various forms of substitution; e.g., by substituting a more plentiful natural resource for a scarce one, or by substituting a manufactured product for a natural one.

ANY ASSESSMENT OF THE VALUE of biodiversity should begin with an account of why we need to value it and the reasons market values would not be expected to suffice for the purpose. The first three sections discuss these matters in the wider context of valuing natural resources (biodiversity is but a special case). A transitional section (Section IV) shows how values can be translated into prescriptions for economic policy. Sections V–VII discuss the special problems that arise in valuing biodiversity and the techniques that are available for coping with these problems.
I. THE SUBJECT

Resource economists view the natural environment through the lens of population ecology. Since the latter’s focus is the dynamics of interacting populations, it has proven useful in resource economics to regard the functioning of the rest of the ecosystem as exogenously given. Well-known illustrations of this viewpoint include the use of the logistic function to chart the time path of the population of a single species of fish, the study of predator-prey interactions by means of variants of the Lotka-Volterra equations, and the estimation of growth in biomass of a species of trees at a given site. A prominent concern in resource economics has been to determine the rates at which a single resource would be harvested in different institutional settings. Thus, not only have socially optimum harvest rates been analyzed (Spence, 1974; Clark, 1976; Dasgupta and Heal, 1979; Hartwick, 1993) but also economists have determined harvest rates when harvesters have free access to the resource (Gordon, 1954; Dasgupta and Heal, 1979; Dasgupta, 1982). We now have an understanding of the effect on harvest rates of harvesting costs, the rates at which harvesters discount future costs and benefits, the productivity of the resource in situ, the “worth” of the harvest to harvesters, and the property rights regime in which the harvesting is done.

Environmental economists, in contrast, base their studies on ecosystem ecology. Here, the focus is on objects such as energy at different trophic levels and its rate of flow among them, the distribution and flows of biochemical substances in soils and bodies of water, and the flows of gases and particulates in the atmosphere. The motivation is to study the biotic and abiotic processes which underlie various ecosystem functions. Economic studies of global warming, eutrophication of lakes, the management of rangelands, purification of water in watersheds, and the pollution of estuaries are examples of such endeavor (Maler, 1974; Costanza, 1991; Nordhaus, 1994; Perrings and Walker, 1995; Chichilnisky and Heal, 1998). Such studies have provided valuable insights into the effects on ecosystems of the character of economic activities, as driven by technology, costs and revenues, discount rates, and the property rights regime that governs the ecosystem.

Formally, it is useful to interpret differences between resource and environmental economics in terms of the state variables that are taken to characterize complex systems (state variables are frequently called stock variables). In resource economics, state variables are “quantities” (biomass units for forests, cow dung, and crop residue; herd size, expressed in numbers, for animal populations; acre-feet for aquifers; and so on). In environmental economics state variables are frequently “quality” indices, such as those for air, soil, or water (although quality, for example, salinity, is typically inferred from quantity indices), with the intention to draw attention to ecosystem services. State variables are summary statistics, reflecting different forms of aggregation. Therein lie their virtue: They enable the analyst to study complex systems by means of a few strategically chosen variables.1

The viewpoint just discussed—that of distinguishing resource and environmental economics in terms of the state variables that summarize complex systems—has enabled economists to integrate problems of resource management with those of environmental pollution (Dasgupta, 1982). Insights from one field of study have been used for gaining an understanding of the others. The viewpoint also reminds us that environmental and resource economics is the study of renewable natural resource systems, when subject to human predation. Since it would be convenient to refer to environmental and resource economics by an overarching name, I do so in this article with the term ecological economics.2

The economics of biodiversity, still in its infancy (Perrings et al. (1994, 1995) are pioneering collections of studies on the subject), is all of the above, but it includes an additional complication in that it recognizes that aggregate statistics can mislead—for example, to devise schemes to manage a resource without understanding its function within the ecosystem of which it is a part can be a road to disaster because the ecosystem’s biodiversity can be a source of its productivity, including, for example, the system’s stability (Tilman and Downing, 1994; Tillman, 1997). In this article, I provide an account of valuation problems in ecological economics; they are, naturally, valuation problems as well in the economics of biodiversity. I then discuss some of the special problems that arise in the latter and techniques available for resolving them.

II. MARKET FAILURE

If you were to browse among such leading Western journals in ecological economics as the Journal of Envir-

---

1 This two-fold classification, based on the character of state variables, is implicit in contemporary ecology (Ehrlich et al., 1977; Roughgarden et al., 1989).

2 I am able to usurp the term from the literature because it appears to have no fixed meaning. "Ecological economics" seems to mean different things to different people.
ronmental Economics and Management, you would discover that a central concern in the field has been to devise ways by which it would be possible to ascertain the “value” of natural resources and the services they provide. A question that would inevitably occur to you is, why? Why should there be a special need to determine the worth of such resources? Why not rely on their market prices? More generally, we may ask why it would not do to rely on markets to guide decisions bearing on the natural resource base, be they global or local, in the way we do for so many other goods and services. In other words, we may ask why markets are not an adequate set of institutions for protecting the “environment.”

The answer is that for many natural resources, markets simply do not exist. In some cases they do not exist because the costs of negotiation and monitoring are too high. One class of examples is provided by economic activities which are affected by ecological interactions involving long geographical distances (e.g., the effects of uplands deforestation on downstream activities hundreds of miles away) and another by large temporal distances (e.g., the effect of carbon emission on climate in the distant future, in a world in which forward markets are nonexistent because future generations are not present today to negotiate with us). Then there are cases (e.g., the atmosphere, aquifers, and the open seas) in which the nature of the physical situation (viz., the migratory nature of the resource) makes private property rights impractical and therefore keeps markets from existing; in other cases, ill-specified or unprotected property rights prevent their existence or make markets function wrongly even when they do exist. In other words, environmental problems are often caused by market failure (but see Section III). Indeed, the phenomenon of externalities (i.e., exchanges among people which occur without their consent) looms large in what has traditionally been called environmental economics.3

Problems arising from an absence of forward markets for “transactions” between the present generation and those in the distant future are no doubt ameliorated by the fact that we care about our children’s well-being and know that they, in turn, will care for theirs in an intergenerational sequence. This means, by recursion, that even if we do not care directly about the well-being of our distant descendants, we do care about them indirectly. However, there is a distinct possibility that our implicit concern for the distant future via such recursion is inadequate due, for example, to institutional failure in other spheres of economic activity. This is why economists have argued that market rates of interest do not reflect socially desirable discount rates (Lind, 1982; Arrow et al., 1996; Portney and Weyant, 1999). In other words, market failure involves not only misallocation of resources in the present but also misallocation across time.

In each of these cases, the market prices of goods and services fail to reflect their social worth; typically, they are less than their social worth. In economics, the social worth of goods and services is called accounting price (or shadow price). The accounting price of a resource is the increase in social well-being which would be enjoyed if a unit more of the resource were made available costlessly. Therefore, a resource’s accounting price is the difference between its market price and the tax (or subsidy) that should be imposed on it. Needless to say, accounting prices reflect social objectives, ecological and technological constraints, and the extent to which resources are available.

It should be noted that externalities do not create market distortions; they are a form of market distortion. The presence of externalities leads to a wedge between market prices and accounting prices. Generally, laissez-faire economies are not much good at producing publicly observable signals of environmental scarcities. To illustrate, if there were free access to a resource base, the market price of the resource in situ would be zero. However, because it is in limited supply, its accounting price would be positive. Therefore, there is a directional bias in environmental externalities: Market failure typically results in an excessive use of the natural resource base, not an insufficient one.

One way to improve matters would be to impose regulations on resource users, for example, restrictions on effluent discharges, quotas on fish harvests, and bans on logging. Another would be to introduce a system of taxes, often called Pigovian taxes [in honor of Pigou (1920), who first discussed the difference between private and social costs in the context of environmental pollution]. Pollution charges, charges on the amount of fish harvested, and stumpage fees are examples. The idea underlying Pigovian taxes is to bring market prices (inclusive of taxes) in line with accounting prices. Each of the two schemes, quotas and taxes, has its advantages and disadvantages; some of the differences between the two become salient once we recognize not only that ecological processes are stochastic but also that resource users and government agencies do not

---

3 The early literature on ecological economics identified market failure as the underlying cause of environmental problems (Pigou, 1920; Lindahl, 1958; Arrow, 1971; Meade, 1973; Mäler, 1974; Baumol and Oates, 1975; Dasgupta and Heal, 1979).
have the same information about local ecology (e.g., the cost of waste disposal) (Meade, 1973; Weitzman, 1974; Baumol and Oates, 1975; Dasgupta, 1982). I will not discuss details here, but three points are worth noting. First, the two schemes are distributionally not equivalent: Under a quota, resource rents are captured by harvesters and polluters, whereas under a tax system they are collected by the tax authority. Second, the imposition of Pigouvian taxes provides greater incentives to resource users to explore resource-saving technological improvements. This is because if the users are taxed, they pay more for the resource than they would have if they had been issued quotas instead. Third, environmental taxes, when properly designed, remove market distortions. In addition, there is a presumption that tax revenues, thus collected, would enable the government to reduce distortingary taxes (e.g., taxes on earned income). There is thus a presumption that Pigouvian taxes yield a “double dividend” (Bovenberg and van der Ploeg, 1994; Goulder, 1995; Bovenberg and Goulder, 1996; but see Bohm, 1996), a rhetorical phrase that has been much used in recent years to persuade governments to impose “green” taxes. Matters of public finance have been a recurrent theme in ecological economics (Baumol and Oates, 1975; Cropper and Oates, 1992; Carraro and Siniscalco, 1996).4

III. INSTITUTIONAL FAILURES AND POVERTY: GLOBAL VS LOCAL ENVIRONMENTAL PROBLEMS

Thus far, market failure. Recently, however, certain patterns of environmental deterioration have been traced to government failure. For example, Binswanger (1991) argued that, in Brazil, the exemption from taxation of virtually all agricultural income (allied to the fact that logging is regarded as proof of land occupancy) has provided strong incentives to the rich to acquire forest land and to then deforest it. He argued that the subsidy the government has thereby provided to the private sector has been so large that a reduction in deforestation (via a removal of subsidies) is in Brazil’s interests and not merely in the interest of the rest of the world. This has implications for international negotiations. The current consensus appears to be that, as a country, Brazil has much to lose from reducing the rate of deforestation in which it is engaged. If this were true, there would be a case for the rest of the world to subsidize Brazil, providing compensation for losses it would sustain if it were to restrain itself. However, as Binswanger’s account suggests, it is not clear if the consensus is correct.5 Elsewhere (and by extension), one would imagine that the massive agricultural subsidies in the European Union considerably influence agricultural practices in ways that inflict substantial damage on the environment.

However, it is important to note that the causes of environmental problems are not limited to market and government failure; problems also occur because such microinstitutions as the household can function badly. In poor communities, for example, men typically have the bulk of the political voice. We should then expect public investment in, for example, resource regeneration to be guided by male preferences, not female needs. On matters of afforestation in the drylands, for instance, we should expect women to favor planting for fuelwood and men for fruit trees because it is the women and children who collect fuelwood, whereas men control cash income (and fruit can be sold in the market). This explains why, even as the sources of fuelwood continue to recede, fruit trees are often planted (Dasgupta, 1993).

That political instability (at the extreme, civil war) is a direct cause of resource degradation is obvious. What is not obvious is that it is a hidden cause as well. Political instability creates uncertainty in property rights. In its presence, people are reluctant to make investments that are necessary for environmental protection and improvement. The expected returns on such forms of investment are low. In a study comprising 120 countries, Deacon (1994) offered statistical evidence of a positive link between political instability and forest depletion.

Together, these examples reflect the environmental consequences of institutional failure. They have a wide reach, and in recent years they have often been discussed within the context of the thesis that environmental degradation, such as eroding soil, receding forests, and vanishing water supplies, is a cause of accentuated poverty among the rural poor in poor countries. There is truth in this. However, there is also accumulated

4 A hybrid policy instrument, which involves the government issuing a fixed number of transferable licenses, combines some of the features of quotas and Pigouvian taxes. For example, the scheme resembles quotas in that resource rents are not captured by government, and it resembles Pigouvian taxes in that the margin license holders pay the accounting price of the resource for its use. See Tietenberg (1980, 1990) for a discussion of transferable licenses, both in theory and in practice.

5 In a broader discussion of the conversion of forests into ranches in the Amazon basin, Schneider (1995) demonstrated that the construction of roads through the forests has also been a potent force.
evidence that poverty can be a cause of environmental
degradation (Dasgupta and Mäler, 1991, 1995; Das-
gupta, 1993a, 1999; Ehrlich et al., 1995). This reverse
causality occurs because some natural resources (e.g.,
ponds and rivers) are essential for survival in normal
times, whereas others (e.g., forest products) are also
a source of supplementary income in times of acute
economic stress. In changing circumstances (e.g., eco-
omic development in urban centers), social norms
which previously had maintained long-term economic
relationships among members of a community tend to
break down. Some (e.g., the able-bodied and mobile)
gain, whereas others (e.g., women, the old, and the very
young) lose and become poorer. In extreme cases, the
breakdown of social norms also means that local re-
sources which earlier were subject to communitarian
regulations become "open access," with all the at-
tended consequences.

These links between rural poverty and the state of
the local natural resource base in poor countries offer
a possible pathway along which poverty, resource degrada-
tion, and even high fertility feed on one another in
a synergistic manner over time (Dasgupta, 1993a,
1995, 2000). Recent experiences in sub-Saharan Africa
and Pakistan are not inconsistent with this (Cleaver and
Schreiber, 1994; Filmer and Pritchett, 1996). Indeed, an
erosion of the local natural resource base can make
certain categories of people destitute even while the
economy's gross national product increases. The belief
that entire populations can always be relied on to make
the shift from resource-based, subsistence existence to
a high-income, industrial one is belied both by evidence
and by theory.

These two causes of resource degradation (i.e., insti-
tutional failure and poverty) pull in different directions
and are together not unrelated to an intellectual tension
between the concerns people share about global warm-
ing and acid rains, which sweep across regions, nations,
and continents, and about those matters (e.g., the de-
cline in firewood or water sources) which are specific
to the needs and concerns of the poor in as small a
group as a village community. Environmental problems
present themselves differently to different people. In
part, it is a reflection of the tension previously discussed
and is a source of misunderstanding of people's atti-
dutes. Some people, for example, identify environmental
problems with poverty and unprecedented popula-
tion growth in the South, whereas others identify them
with wealth and unprecedented expenditure patterns
in the North (I use the geographical terms in their
current geopolitical sense). Even though debates be-
tween the two groups often become shrill, each vision
is in part correct. There is no single environmental
problem, and therefore no single valuation problem;
rather, there is a large collection of them (Dasgupta
and Mäler, 1995; Reardon and Vosti, 1995; Vincent et
al., 1997). Thus, growth in industrial wastes and re-
source use have been allied to increased economic activ-
ity; and in the former, Socialist block neither preventive
nor curative measures have kept pace with the produc-
tion of waste. Moreover, the scale of the human enter-
prise, both by virtue of unprecedented increases in the
size of the world's population and by the extent of
economic activity, has so stretched the capabilities of
ecosystems that humankind can today rightly be charac-
terized as Earth's dominant species (Vitousek et al.,
1997). These observations loom large not only in eco-
logical economics but also in the more general writings
of environmentalists and in the professional writings
of ecologists in the West.

On the other hand, economic growth has brought
with it improvements in the quality of many natural
resources. The large-scale availability of potable water
and the increased protection of human populations
against both water- and air-borne diseases in industrial
countries have in great measure occurred in the wake of
growth in national income these countries have enjoyed
during approximately the past 200 years. Moreover, the
physical environment inside the home has improved
beyond measure with economic growth. For example,
cooking in Southeast Asia continues to be a central
route to respiratory illnesses among women. Such pos-
tive links between economic growth and environmental
quality often go unnoted by environmentalists in the
North. I believe that this lacuna is another reflection
of the fact that it is easy to overlook the enormous
heterogeneity of Earth's natural resource base, ranging
from the atmosphere, oceans, and landscapes to water
holes, grazing fields, and sources of fuelwood. Both this
heterogeneity and the diversity of the human condition
throughout the world need constantly to be kept in
mind in discussions of the value of biodiversity in differ-
et locations.

IV. VALUING RESOURCES AND
EVALUATING PROJECTS

Since institutional failures abound in our dealings with
Earth, the commercial profitability of economic activi-
ties, for example, of investment projects (projects for
short), is frequently not an adequate measure of their
social worth. Therefore, recourse should be taken to
social cost–benefit analysis, the purpose of which is to estimate the impact of projects on human well-being, now and in the future. Note that, if undertaken, a project would be a perturbation to the economy. Therefore, a project consisting of the construction of a dam, for example, would be a perturbation to an economy without the dam. The economic forecast sans the project can be thought of as the status quo.

Analyzing the consequences of a project would involve estimating the need for labor, intermediate products, raw materials, and output as well as predicting the ecological effects of the project. These consequences need to be specified for each future period (see Section VII for a formalization). Since there is never sufficient knowledge to make precise estimates of the consequences, project evaluators should quantify estimates of the uncertainties, preferably in terms of probabilities. This means that, in general, project designers should model the integrated ecological and economic system. However, in practice this is infrequently done.6

In order to derive a good estimate of a project’s social benefits and costs, one should in principle value every commodity involved in it. The procedure devised by economists is to select some readily measurable bundle of goods ordinarily consumed and define the value of any other commodity as the amount of the bundle society would be willing to give up for it. This is a workable way for estimating the commodity’s accounting price. The net social benefit of a project in any given period of its life is obtained by multiplying the project’s inputs and outputs in that period by their corresponding accounting prices and adding them (outputs of “goods” are taken to be positive, and output of “bads” and inputs are taken to be negative). Using a suitable discount rate (often called the social discount rate; Arrow et al., 1996; Portney and Weyant, 1999), the net social benefits yielded by a project in each period are added. Projects which yield a positive present discounted value of net social benefits are recommended for acceptance, whereas those yielding a negative present discounted value of net social benefits are rejected. The theory of social cost–benefit analysis has been developed by economists during the past 50 years and is now, to all intents and purposes, complete (Dasgupta et al., 1972; Little and Mirrlees, 1974; Dasgupta and Maler, 2000).7

A prior exercise (i.e., prior to conducting social cost–benefit analysis) is to estimate accounting prices. A great deal of work in ecological economics has been directed at discovering methods for estimating accounting prices of natural resources.

It must be remembered that the kinds of resources I am considering here are sometimes of direct use in consumption (as with fisheries), sometimes of indirect use as inputs in production (as with plankton, which serves as food for fish), and sometimes both (as with drinking and irrigation water). The value may be utilitarian (e.g., as a source of food or a keystone species), it may be aesthetic (e.g., a shoreline), or it may be intrinsic; indeed, it may be all these things.

Economists have devised various methods for estimating accounting prices. As would be expected, the prices of some natural resources are easier to estimate than those of others. There are standard techniques for determining the accounting prices of irrigation water, fisheries, timber, and agricultural soil (Anderson, 1987; Repetto et al., 1989; Solorzano et al., 1991; Vincent et al., 1997). They involve estimating the resource’s use value. For example, the value of a piece of agricultural land, such as agricultural land, would be the present discounted value of the flow of net profits it is expected to generate from cultivation minus the environmental damage caused by the pesticides and herbicides to be used. Such an approach can also be used for estimating losses associated with water-logging and overgrazing. Reductions in air- or water-borne pollution can be valued in terms of improvements in health (e.g., reductions in the number of days people would be expected to be ill, World Bank, 1992). Other techniques have been devised for valuing “amenities” such as places of scenic beauty.8

Methods have also been devised for giving expression to the “precautionary principle,” which assumes a particularly subtle form when applied to resources such as genetic material in tropical forests. The subtlety arises from a combination of uncertainty in the future use values of these resources and irreversibility when they are lost. The twin presence of uncertainty and irrevers-

6 However, there are signs of change. See, for example, Perrings et al. (1994, 1995), Vincent et al. (1997), Chopra and Kadkoohi (1999), and various issues of the journal Environment and Development Economics.

7 Dally et al. (1999) provide a nontechnical account of the role of social cost–benefit analysis in environmental management.

8 One popular method involves asking people hypothetical questions concerning their willingness to pay for preserving the amenity (this is called the “contingent valuation method”); another involves estimating from sample surveys the distribution of costs that visitors from different locations have incurred to view the site (this is called the “travel cost method”). A third involves inferring how much people are willing to pay for enjoying the amenity (e.g., clean air) from the commercial value of land at sites which offer the amenity (this is called the “hedonic price” of land). See Maler and Wycka (1976), Mitchell and Carson (1989), and Freeman (1992).
ibility implies that preservation of the stock has a value in addition to its current use value, namely, the value of extending society's set of future options. Future options have an additional worth because, with the passage of time, more information is expected to be forthcoming about the resource's use value. This additional worth is often called an option value (Arrow and Fisher, 1974; Henry, 1974; Fisher and Hanemann, 1986). The accounting price of a resource, at the very least, is the sum of its direct use value and its option value.

These techniques enable us to estimate the use value of a given resource. The resource's accounting price may well exceed this. Why? Because there may be additional values "embodied" in a resource. An additional consideration is applicable to living resources: their intrinsic worth as living resources. It would be absurd to suppose that the value of a blue whale is embodied entirely in its flesh and oil, or that the value of game in Kenyan safari parks is simply the present discounted value of the flow of tourists' willingness to pay to view them. The idea of intrinsic worth of living things is inherent not only in traditional religious systems of ethics but also in modern ethical theories. Therefore, the question is not so much whether living creatures have intrinsic worth but rather regards the ways of assessing this worth. Because it is almost impossible to get a quantitative handle on intrinsic worth, the correct thing to do is to take note of it, keep an eye on it, and call attention to it in public debate if the stock is threatened with destruction.

We may conclude that the social worth of natural resources can be divided into three parts: their use value, their option value, and their intrinsic value. The components appear in different proportions, depending on the resource. For example, oil and natural gas would not be thought to have intrinsic value, nor perhaps an option value, but they have use value. On the other hand, primates would be thought to be intrinsically valuable.

It must also be emphasized that the purpose of estimating environmental accounting prices is not to value the entire environment; rather, it is to evaluate the benefits and costs associated with changes made to the environment due to human activities. Prices, whether actual or accounting, have significance only when there are potential exchanges from which choices have to be made (e.g., when one has to choose among alternative investment projects). Thus, the statement that a particular act of investment can be expected to degrade the environment by, for example, $1 million annually has meaning because it says, among other things, that if the investment were not to be undertaken, humanity would enjoy an additional $1 million of annual benefits in the form of environmental services. The statement also has operational significance: The estimate could (and should) be used for calculating the rate of return attributable to the investment in question.

Contrast such an estimate of the value of an incremental change in the natural resource base with the one which says that, worldwide, the flow of environmental services is currently worth, in total, $33 trillion annually (Costanza et al., 1997). The former is meaningful because it presumes that humanity will survive the incremental change and be there to experience and assess the change. The reason the latter should cause us to balk is that if environmental services were to cease, life would not exist. However, who would be there to receive $33 trillion of annual benefits if humanity wished to exchange its very existence for them? This is a case in which the value of an entire something has no meaning and therefore is of no use, even though the value of incremental changes to that same something not only has meaning but also has use.

V. BIODIVERSITY: NECESSITY OR LUXURY?

Biodiversity has been neglected in economics. Scratch an economist, and you are likely to find someone who regards the natural resource base as an "amenity." Thus, it is commonplace that, to quote a recent editorial in London's Independent (December 4, 1999), "[economic] growth is good for the environment because countries need to put poverty behind them in order to care" or that, to quote the Economist (December 4, 1999, p. 17), "trade improves the environment, because it raises incomes, and the richer people are, the more willing they are to devote resources to cleaning up their living space."

I quote these views not to question that poverty should be a phenomenon of the past, nor to suggest that arbitrary restrictions on trade do not cause much harm, but rather to show that natural resources are widely viewed as luxuries. This is, of course, an incorrect view: The natural resource base is not a luxury. Producing, as it does a multitude of ecosystem services, the natural resource base is a necessity. Indeed, the

---

* As stressed elsewhere in the encyclopedia, these services include maintaining a genetic library, preserving and regenerating soil, recycling nutrients, controlling floods, filtering pollutants, assimilating waste, pollinating crops, operating the hydrological cycle, and maintaining the gaseous composition of the atmosphere.
ECOLOGICAL ECONOMICS TO BE FOUND IN THE NORTH DOES NOT EXACTLY RESONATE AMONG THOSE WHO Worry ABOUT DEGRADATION OF THE NATURAL RESOURCE BASE IN THE SOUTH AND THE ADDITIONAL HARDSHIP THIS BRINGS TO THE MANY AMONG THE POOREST OF PEOPLE WHOSE LIVES DEPEND DIRECTLY ON THAT BASE. NEEDLESS TO SAY, IT DOES NOT RESONATE AMONG ECOLOGISTS, QUÀ ECOLOGISTS, EITHER.

This perspective of viewing natural resources as luxuries is expressed in the World Bank (1992) report, in which it was suggested that there is an empirical relationship between gross domestic product (GDP) per head and concentrations of industrial pollutants. Based on the historical experience of the Organization for Economic Cooperation and Development countries, it was argued in the document that, when GDP per head is low concentrations of atmospheric pollutants [e.g., sulfur dioxide (SO₂)] increase as GDP per head increases, but when GDP per head is high concentrations decrease as GDP per head increases further. In other words, it was found that the functional relationship between GDP per head and concentrations of industrial pollutants has an inverted U shape (Fig. 1). Among economists, this relationship has been christened the "environmental Kuznets curve".¹¹

Panayotou (1992) reported the inverted U shape in cross-country data on GDP per head and deforestation and emissions of SO₂, nitrogen oxides (NOₓ), and particulate matters. Sweden, for example, was found to lie on the downward part of the curve. Indeed, time series on timber stocks and sulfur and nitrogen emissions in Sweden, during the 1980s, are consistent with this: Timber stocks have increased, and the emission rates of sulfur and nitrogen oxides have declined.

Note that, excluding forests (but see Section VI), the previously mentioned findings concern mobile pollutants. Their mobility hides the fact that earlier emissions have had to find somewhere to lodge: Since matter is conserved, inputs and outputs of material must balance (Ayres and Kneese, 1969; d'Arge et al., 1970; Mäler, 1974). The point is that my emissions would affect not only me but also you if you lived downstream or downstream. However, even if we did not consider this point, the logic underlying the environmental Kuznets curve is that resource degradation is reversible: Degradation all you want now, you can always recover the stock later because Earth can be relied on to rejuvenate it. The science of biodiversity has shown this presumption to be false. The presence of ecological thresholds implies that damage to ecosystems can be irreversible. As an overarching metaphor for "trade-offs" between manufactured wealth and resource degradation, the environmental Kuznets curve has to be rejected.¹²

VI. SUBSTITUTION POSSIBILITIES¹³

The belief that constraints arising from resource depletion can be overcome as countries become wealthier in terms of their manufactured and human capital assets is frequently based on a subtler idea than the one that underlies the environmental Kuznets curve. The belief is based on possibilities of substitution.

Resource constraints facing an economy can be eased by four types of substitution. First, there can be substitution of one thing for another in consumption (nylon and rayon cloth substituting for cotton and wool, pulses substituting for meat, etc.). Second, manufactured capital can substitute for labor and natural resources in production (the wheel and double-glazing are two extreme examples). Third, novel production techniques can substitute for old ones. For example, the replacement of the piston with the steam turbine (i.e., converting from reciprocating to rotary motion) was introduced into power plants and ships more than 100 years ago. The innovation was an enormous energy saver in engines. Fourth, and most important, natural resources

¹¹ I have discussed these concerns in greater detail elsewhere (Dasgupta, 1982, 1993a, 1996, 1997).

¹² This is a misnomer. The original Kuznets curve, which was an inverted U, related income inequality to real national income per head on the basis of historical cross-country evidence.

¹³ For more extensive discussion of the environmental Kuznets curve, see Arrow et al. (1995) and the responses their article elicited in symposia built around the article in Ecological Economics 15 (1995), Ecological Applications 6 (1996), and Environment and Development Economics 1 (1996); also see the special issue of Environment and Development 2(4) (1997).

¹⁴ This and the following section are based on Dasgupta and Mäler (1999) and Dasgupta et al. (2000a,b).
can substitute for one another. This involves the belief that, as each resource (e.g., each species) is depleted, there are similar substitutes lying in wait, either at the same site or elsewhere. If this were true, then even as constraints increasingly tighten on any one resource base, humanity would be able move to other resource bases, either at the same site or elsewhere. The enormous additions to the sources of industrial energy that have been realized (successively, human and animal power, wind, timber, coal, oil, and natural gas and, recently, nuclear) are a prime historical illustration of this possibility.14

Humans have been "substituting" one thing for another since time immemorial. Even the conversion of forests into agricultural land in England in the Middle Ages was a form of substitution: Large ecosystems were transformed to produce more food. However, the pace and scale of substitution in recent centuries have been unprecedented. Landes (1998) argued that substitution created the industrial revolution in England in the eighteenth century. The extraordinary economic progress experienced in Western Europe and North America since then (during the past two centuries GDP per head in Western Europe has increased approximately 20-fold), and recently in East Asia, has also been a consequence of substitution. Spatial dispersion of ecosystems has enabled this to happen. The ecological transformation of rural England in the Middle Ages presumably reduced the nation's biodiversity, but it increased income without any direct effect on global productivity.

However, that was in the past, and we must deal with the present. A question currently much debated is whether it is possible for the scale of human activity to be increased substantially beyond what it is today without placing undue stress on the major ecosystems that remain. In any event, the cost of substituting manufactured capital for natural resources can be high. Low-cost substitutes could turn out to be not so low-cost if accounting prices were used in the costing, not market prices. Even when accounting prices are not used, degrading natural capital and substituting it with manufactured capital can be uneconomic. Chichilnisky and Heal (1998) compared the costs of restoring the ecological functioning of the Catskill watershed ecosystem in New York State to the costs of replacing the natural water purification and filtration services the ecosystem has provided in the past by building a water-purification plant costing $8 billion. They showed the overwhelming economic advantages of preservation compared to construction: Independent of the other services the Catskill watershed provides and ignoring the annual running costs of $300 million for a filtration plant, the capital costs alone showed a more than sixfold advantage for investing in the natural capital base.

Degradation of a natural resource base (e.g., destruction of native populations of flora and fauna) not only affects the volume and quality of ecosystem services the base provides but also challenges the system's resilience, which is its capacity to absorb disturbances, or perturbations, without undergoing fundamental changes in its functional characteristics. The way to interpret an ecosystem's loss of resilience is to view it as having moved to a new stability domain, which is another way of saying that the system, having crossed a threshold, has been captured by a different attractor (Levin et al., 1998; Levin, 1999; Brock et al., 1999). Sudden changes in the character of shallow lakes (e.g., from clear to eutrophied water), owing to increases in the input of nutrients (Schuster, 1997; Carpenter et al., 1998), and the transformation of grasslands into shrublands, consequent upon nonadaptive cattle management practices (Perring and Walker, 1995), are examples. Human societies have on occasion been unable to avoid suffering from unexpected changes in their local ecosystems. Fishermen on Lake Victoria and the nomads in the now shrublands of southern Africa are examples from recent years.

Biodiversity appears to be a key to ecosystem resilience. However, even today it is a popular belief that the utilitarian value of biodiversity is based mainly in the potential uses of genetic material (e.g., for pharmaceutical purposes) or, in other words, that its social worth is almost wholly an option value. Preservation of biodiversity is seen as a way to hold a diverse portfolio of assets with uncertain payoff. However, as other contributions to this encyclopedia make clear, biodiversity, appropriately conceived, is essential for the maintenance of a wide variety of services on which humans depend for survival. This has the important corollary that to invoke the idea of substitutability among natural resources in order to play down the use value of biodiversity, as people frequently do (Simon, 1981, 1994), is a wrong intellectual move. The point is that if an ecosystem's biodiversity is necessary for it to be able to continue providing us with its services, the importance of that same biodiversity cannot be downplayed by the mere hope that for every species there are substitute

14 However, these shifts have not been without unanticipated collective costs. Global warming, associated with the burning of fossil fuels (an "externality"), did not feature in economic computations in earlier decades. See Dasgupta (1993b) for a less coarse partition of substitution possibilities than that provided by the four-way classification.
species lying in wait within that same ecosystem. In other words, there is an inconsistency in this line of reasoning. Recall the famous analogy in Ehrlich and Ehrlich (1981) relating species in an ecosystem to rivets in an airplane. One by one, perhaps, species may disappear and not be missed. Eventually, however, the cumulative effect of loss of biodiversity will lead to the crash of ecosystem functioning, just as the cumulative loss of redundant rivets will lead to the crash of an airplane.

VII. ECOSYSTEM BIFURCATIONS AND DISCONTINUOUS VALUE FUNCTIONS

How do discontinuities in the social worth of ecosystems affect valuation exercises and social cost–benefit analysis? To answer this, it helps to formalize.

Consider an ecosystem describable by N state variables, indexed by i and j (i, j = 1, 2, …, N). For concreteness, we may think of each state variable as reflecting the population size of a particular species. (As noted in Section I, problems of environmental pollution can be formulated in a similar manner.) Denote time by t (≥0) and let $S_i$ be the population size of i at t. Time is taken to be a continuous variable. We imagine, therefore, that the dynamics of the ecosystem can be described by a system of (nonlinear) differential equations. For expositional ease, we assume that the system is deterministic.

Let the net reproduction rate of i at t be $F_i$. Since the ecosystem is coupled, $F_i$ is a function of the stocks at t. Thus I write as $F_i(S_{1i}, S_{2i}, \ldots, S_{Ni})$, for i = 1, 2, …, N. I assume that ecologists have estimated these functions. Next, assume that the ecosystem dynamics are autonomous. This means that $F_i$ is not an explicit function of t. Therefore, I drop the subscript t from $F_i$ and write the function as $F_i(S_{1i}, S_{2i}, \ldots, S_{Ni})$. In all the applications of this framework with which I am familiar, $F_i$ is taken to be a differentiable function. Let us assume this is true.

The analysis begins at t = 0 (the "present"). Denote by $X_i$ the rate at which species i is harvested at time t. We now imagine that economists have studied the human–ecosystem interactions in question. They have inquired into the structure of property rights, demand conditions, government policies, and so forth. On the basis of this they have concluded that harvests are based on an implicit policy, in that they are time autonomous and are functions solely of stocks. Therefore, we may write $X_i = X_i(S_{1i}, S_{2i}, \ldots, S_{Ni})$. Assume that $X_i$ is piecewise continuous and possesses right and left partial derivatives everywhere. This is a technical assumption and a good one. For example, optimal policy functions for those ecosystem management problems that have been studied have been found to possess this property (Skiba, 1978; Brock et al., 1999). Moreover, actual harvest rates have frequently been known to be approximately constant over time. Both sets of examples satisfy the assumption.

No doubt some of the $X_i$’s would be zero. For example, it could be that only one species in the ecosystem is ever harvested (e.g., because it is the only one that has economic worth). We should think of $X_i$ as a forecast. It should be stressed that $X_i(S_{1i}, S_{2i}, \ldots, S_{Ni})$ is not necessarily a socially optimal harvest policy function. It can be an actual policy function within an imperfect institution (e.g., the ecosystem could be one to which there is free access).

The rate of change of $S_i$ is the difference between $F_i$ and $X_i$. Therefore, given the economists' forecast for $X_i$, mathematicians would be able to forecast $S_i$ by solving the "coupled" system of differential equations:

$$\frac{dS_i}{dt} = F_i(S_{1i}, S_{2i}, \ldots, S_{Ni}) - X_i(S_{1i}, S_{2i}, \ldots, S_{Ni}), \text{ for all } i$$

(1)

For simplicity of exposition, assume that the social worth of the ecosystem is autonomous in time. We may then express that worth by a scalar V. Since V would be a function of the stocks, we may write it as $V(S_{1i}, S_{2i}, \ldots, S_{Ni})$. V is the value at t of the entire ecosystem. It is the maximum amount society should be "willing to pay" at t for the ecosystem's survival if the stocks of the N resources were $S_{1i}, S_{2i}, \ldots,$ and $S_{Ni}$, respectively. Any alternative use of the site (e.g., conversion into an urban center) would have to be worth at least V if the alternative were to be acceptable. The form of V would depend on the availability of substitutes for those species that are harvested. Again, to keep the mathematical notation from becoming too complex, I assume that there are no substitutes available at low cost from outside the ecosystem (e.g., because the community doing the harvesting is not near other sources of livelihood). Using Eq. (1), it is possible to use the forecast on harvest rates to determine forecasts on stocks. This in turn makes it possible to forecast the time path of V.

10 Actual harvest rates frequently display time trends, for example, because population and income grow. Time trends in $X_i$ would render the system of Eq. (1) below nonautonomous. In the text, I restrict the discussion to autonomous systems because I understand the mathematics of autonomous better than that of nonautonomous ones. However, experience with simple nonautonomous systems suggests that the arguments I offer later in the text apply to them as well.
At one level the valuation problem is now "solved": \( V(S_1, S_2, \ldots, S_n) \) would be the value of the ecosystem at \( t \). It would be the social worth of the ecosystem at \( t \). The problem is that \( V \) is typically a nonlinear function, which means that it is very difficult to estimate. The task of valuing ecosystems would be made much easier if recourse were taken to estimating accounting prices. The advantage would be that since accounting prices reflect the social worth of marginal units of the various populations, we could use such prices to construct a linear index of the ecosystem's value.

Assume that \( V \) is differentiable everywhere. Let \( P_i \) be the accounting price of \( i \) at \( t \). From the discussion in Section IV and from our assumption that no substitutive resources are near the human community in question, we know that

\[
P_i = \partial V / \partial S_i, \text{ for all } i \text{ and all } t
\]

(2)

At time \( t \) the value of species \( i \) would be \( P_i S_i \). It follows that the value of the ecosystem would be \( \sum P_i S_i \). Note that this is a linear function of stocks, with the weights being accounting prices.

In Section IV it was noted that a "project" can be thought of as a perturbation of the forecast \( X_n \). Therefore, a project can be denoted as \( (\Delta X_1, \Delta X_2, \ldots, \Delta X_n) \), for \( t \geq 0 \). (\( \Delta \) denotes an operator signifying "small difference.") Some of the \( \Delta X_i \)'s would be zero. Nevertheless, the project would be expected to perturb future stocks of all the resources since this is what a strongly coupled ecosystem would be expected to display.

Let \( r \) be the social rate of discount and let \( C_i \) be the unit cost of harvesting \( i \) at \( t \). \(^{18} \) It follows that the present discounted value of the flow of net social benefits from the project is

\[
o^t \sum e^{-rt} \left[ \sum (P_i - C_i) \Delta X_i \right] dt
\]

(3)

If Eq. (3) is positive, the project should be accepted; if it is negative, the project should be rejected.

It can be argued that projects, as I have defined them here, are merely "small" perturbations, whereas redirecting economic activity so as to avoid damaging an ecosystem irrevocably could involve drastic change. However, it should be noted that one way to conceptualize a "large" perturbation is to regard it as the sum of many small perturbations. A large perturbation (i.e., a large project) could then be evaluated by repeated use of Eq. (3).

However, if this route is not adopted, social cost-benefit analysis of large projects requires the project evaluator to estimate the large changes in \( V \) consequent on the adoption of large changes in economic policy. Accounting prices, reflecting the social worth of marginal units of the various resources (Eq. 2), would then not suffice. The evaluator would need to integrate over the marginal units so as to estimate "consumer surpluses," to use a term familiar in economics.

So far so good. However, there is a problem with the account: It is unreasonable to assume that \( V \) is differentiable, even continuous, everywhere. Ecosystems are nonlinear. So, even if it were reasonable to suppose that \( V \) is differentiable everywhere else, it would be wrong to suppose that it is even continuous at separatrixes and bifurcation points.\(^{19} \)

However, if the \( X_i \)'s are not optimal, \( V \) can be discontinuous at points on separatrixes and bifurcation points. This causes problems because accounting prices cannot be defined at such points. Now, we will study the implications of this for biodiversity valuation and social cost-benefit analysis.

Experience with nonlinear models of ecosystems tells us that, under the assumptions we have made, there could be at most a countable number of bifurcation points and separatrixes. This is fortunate because it means that points on the stock space that are "troublesome" are nongeneric. Therefore, let us assume this is so. In Fig. 2, the matter is illustrated in the context of an ecosystem comprising a single species. Figure 2 depicts the case in which the separatrix is a single point, \( S^* \), reflecting a threshold. For example, it could be that, under the harvesting policy \( X(S) \), the species would become extinct if its population were below \( S^* \) but would be harvested in a sustainable manner if the population were in excess of \( S^* \). Therefore, stocks to the

---

\(^{18} \) If substitutes were available, \( P_i \) would be the minimum of \( \partial V / \partial S_i \) and the accounting price of the substitute. I want to avoid such complications here.

\(^{17} \) Note that for all \( i \) and all \( t \), \( \Delta X_i = \sum (\partial X_i / \partial S_i) (dS_i/dt) \Delta t \).

\(^{16} \) \( C_i \) could depend on stock sizes at \( t \). For example, the unit cost of fishing depends not only on the technology available for fishing and the price of fishing equipment but also on the stock in the fishery. The larger the stock, the smaller the unit cost.

\(^{19} \) The location of bifurcation points depends on, among other things, the \( X_i \)'s. Therefore, they are endogenous to the economic-ecological system. In an important early contribution, Skiba (1978) showed via an example that if harvest functions are optimal, \( V \) is continuous even at points where harvests are discontinuous. This means that accounting prices are not uniquely specified at such points. However, \( V \) can be shown to possess right- and left-partial derivatives there. Therefore, accounting prices can be used for evaluation purposes, even though they are not uniquely given at every point on the space of resource stocks.
right and left of $S^*$ represent different basins of attraction. Figure 2 reasonably depicts a case in which the value of the species, $V(S)$, is an increasing function of the stock. It is assumed to be continuous (indeed, differentiable) everywhere except at $S^*$, where it jumps (an irreversibly dying population being a lot less valuable than a sustainable one). Of course, the location of $S^*$ depends on $X(S)$: Change the harvesting policy slightly, and $S^*$ will shift slightly. The influence of $X(S)$ on $S^*$ has to be estimated if ecologists and economists are to offer policy advice.

Now, excepting by fluke, the stock at $t = 0$ would be different from $S^*$. Therefore, let us assume it is different. If the project is sufficiently small, the account of social cost–benefit analysis given previously remains valid: The system would not cross into a different basin of attraction. However, a good theory should be extendable to fluke cases. Moreover, actual projects are frequently not “small” so that acceptance of a project or its rejection could mean that the ecosystem is eventually in one basin of attraction rather than in another. How do we extend the theory to handle the possibility that the ecosystem crosses into a different basin of attraction? In particular, is the repeated use of Eq. (3) a feasible means of evaluating projects?

It is also clear where the problem lies if we try using Eq. (3). The problem lies in that an accounting price cannot be defined at $S^*$. This means that a project which involves the stock passing through $S^*$ cannot be evaluated by means of a linear index of social profitability. The height of the jump would have to be estimated and put to use in social cost–benefit analysis. Ecologists and economists would have to combine their expertise to locate $S^*$ and identify the functional form of $V(S)$, both on the right and on the left of $S^*$. Estimating the height of the jump involves measuring consumer surpluses, a point noted earlier. In short, at least one small project in the series of small projects which add up to the large project in question would not be assessable by means of Eq. (3). This causes difficulties.

However, there is a way to avoid the problem. We have been studying deterministic systems. Introducing uncertainty about the location of $S^*$ can help matters by smoothing the value function. To see how, imagine that $V(S)$ represents the expected value of the resource's social worth at $S$. If the location of $S^*$ were a smooth probability distribution, $V(S)$ would be a continuous, even a differentiable, function. In this case, an accounting price of the resource would be definable at all $S$ (with $S^*$ being a smooth random variable). A linear index of the social profitability of projects could then be constructed. The methods of social cost–benefit analysis outlined earlier would remain valid.

It is not often that introducing realism simplifies analysis. Valuing biodiversity would seem to be an exception.

See Also the Following Articles

- BIODIVERSITY AS A COMMODITY
- ECONOMIC GROWTH
- AND THE ENVIRONMENT
- ECONOMIC VALUE OF
- BIODIVERSITY, MEASUREMENT OF
- HUMAN EFFECTS ON
- ECOSYSTEMS, OVERVIEW
- MARKET ECONOMY AND
- BIODIVERSITY

Bibliography


ECOSYSTEM, CONCEPT OF
Eugene P. Odum
University of Georgia

I. Introduction
II. The Ecosystem Concept
III. The Human-Dominated Techno-ecosystem

GLOSSARY

autotroph Literature, a self-feeder, an organism that is able to utilize inorganic carbon (carbon dioxide) as the sole carbon source for growth, for example, green plants and certain bacteria.

black box Entity that can be examined at the system level without specifying its internal contents.

heterotroph Literature, a feeder on others, an organism that is dependent on organic material from an external source to provide carbon for growth, for example, vertebrates.

industrialized agriculture Modern form of agriculture that differs from traditional agriculture in the use of elaborate and expensive machinery, the control of pests with toxic chemicals rather than biocontrols, fertilization by synthetic rather than organic products, excessive consumption of water, and farm ownership and management by corporations rather than individuals.

input environment Collective term for all energy and materials moving into a given system.

mega-city Modern city with a large, expanding population, characterized by high consumption levels of energy, water, and food from sources outside the city.

output environment Collective term for all energy and materials moving out of a given system.

 techno-ecosystem Technology-based ecosystem in the contemporary world that is fundamentally distinct from natural ecosystems in the use of energy sources other than sunlight (fossil fuels, nuclear power), an urbanized concentration of human population, and the generation of substantial amounts of air and water pollutants and waste materials.

LIVING ORGANISMS AND THEIR NONLIVING (ABIO- TIC) ENVIRONMENT are inseparably interrelated and interact with each other. An ecological system, or ecosystem, is any unit (a biosystem) that includes all the organisms (the biotic community) in a given area interacting with the physical environment so that a flow of energy leads to clearly defined biotic structures and cycling of materials between living and nonliving parts. An ecosystem is more than a geographical unit (or ecoregion); it is a functional system unit with inputs and outputs, and with boundaries that can be either natural or arbitrary.

I. INTRODUCTION

The ecosystem is the first unit in the molecule to ecosphere hierarchy (as shown in Fig. 1) that is complete,
that is, it has all the components, biological and physical, necessary for survival. Accordingly, it is the basic unit around which to organize both theory and practice in ecology. Furthermore, as the shortcomings of the "piecemeal" short-term technologic and economic approaches to dealing with complex problems become ever more evident with each passing year, management at this level, that is, ecosystem management, emerges as the challenge for the future.

Since ecosystems are functionally open systems, consideration of both inputs and outputs is an important part of the concept, as shown in Fig. 2. A diversity of species and genetic forms, together with a variety of functions and niches, are essential properties of natural ecosystems. Ecosystem diversity provides redundancy in times of environmental uncertainty.

II. THE ECOSYSTEM CONCEPT

The term "ecosystem" was first proposed in 1935 by the British ecologist A. G. Tansley, but of course the concept is by no means so recent. Allusions to the idea of the unity of organisms and the environment (as well as the oneness of humans and nature) can be found as far back in written history as one might care to look. Not until the late 1800s did formal statements begin to appear, interestingly enough, in a parallel manner in the American, European, and Russian ecological literature. Thus, in 1877 Karl Mobius wrote (in German) about the community of organisms in an oyster reef as a "biocoenosis," and in 1887 S. A. Forbes, an American, wrote his classic essay on the lake as a "microcosm." The pioneering Russian V. V. Dokuchaev and his chief disciple, G. F. Morozov (who specialized in forest ecology), emphasized the concept of the "biocoenosis," a term later expanded by Russian ecologists to "geobio-coenosis" (Sukachev, 1959).

In addition to biologists, physical scientists and social scientists began to consider the idea that both nature and human societies function as systems. In 1925, the physical chemist A. J. Lotka wrote in a book entitled *Elements of Physical Biology* that the organic and inorganic worlds function as a single system to such an extent that it is impossible to understand either part without understanding the whole. It is significant that a biologist (Tansley) and a physical scientist (Lotka) independently and at about the same time came up with the idea of the ecological system. Because Tansley coined the word "ecosystem" and it caught on, he gets most of the credit, which should be shared with Lotka.

In the 1930s, social scientists developed the holistic concept of regionalism, especially Howard W. Odum,
who used social indicators to compare the southern region of the United States with other regions (Odum, 1936; Odum and Moore, 1938). More recently, Machlis et al. (1997) and Force and Maddie (1997) have promoted the idea of the human ecosystem that combines biological ecology and social theories as a basis for practical ecosystem management. Accordingly, the concept of the ecosystem now brings together organisms, the physical environment, and humans.

As shown in Fig. 2, a graphic model of an ecosystem can consist of a box that we can label the system, which represents the area we are interested in, and two large funnels that we can label input environment and output environment. The boundary for the system can be arbitrary (whatever is convenient or of interest), delineating an area such as a block of forest or a section of beach, or it can be natural, such as the shore of a lake where the whole lake is to be the system, or ridges as boundaries of a watershed.

Energy is a necessary input. The sun is the ultimate energy source for the biosphere and directly supports most natural ecosystems within the biosphere. But there are other energy sources that may be important for many ecosystems, for example, wind, rain, water flow, or fuel (the major source for urban-industrial society). Energy also flows out of the system in the form of heat and in other transformed or processed forms such as organic matter (e.g., food and waste products) and pollutants. Water, air, and nutrients necessary for life, along with all kinds of other materials, constantly enter and leave the ecosystem. And, of course, organisms and their propagules (seeds and other reproductive stages) enter (immigrate) or leave (emigrate).

In Fig. 2 the system part of the ecosystem is shown as a "black box," which is defined by modelers as a unit whose general role or function can be evaluated without specifying its internal contents. Figure 3 is a graphic model of the solar-powered natural ecosystem showing internal system components and functions. The interactions of the three basic components, namely, (1) the community, (2) the flow of energy, and (3) the cycling of materials, are diagrammed in this simplified compartment model. Energy flow is one-way; some of the incoming solar energy is transformed and upgraded in

![FIGURE 3](image_url)
quality (i.e., converted into organic matter, a more upgraded form of energy than sunlight) by the community, but most of it is degraded and passes through and out of the system as low-quality heat energy (heat sink). Energy can be stored, then “fed back,” or exported, as shown in the diagram, but it cannot be reused. In contrast with energy, materials, including the nutrients necessary for life (carbon, nitrogen, phosphorus, and so on), and water can be used over and over again. The efficiency of recycling and the magnitude of imports and exports of nutrients vary widely with the type of ecosystem.

Each ‘box’ in Fig. 3 is given a distinctive shape that indicates its general function. Circles are renewable energy sources, bullets are autotrophs, hexagons are heterotrophs, and the round-bottomed shapes are storages (in this case of nutrients and organisms). The community is depicted as a ‘food web’ of autotrophs and heterotrophs linked together with appropriate energy flows, nutrient cycles, and storages.

Both graphic models (Figs. 2 and 3) emphasize that a conceptually complete ecosystem includes inputs and outputs along with the system as delimited, that is, an ecosystem = E + S + OE (input environment + system + output environment). This scheme solves the problem of where to draw lines around an entity that one wishes to consider, because it does not matter very much how the box portion of the ecosystem is delimited. Often, natural boundaries, such as a lakeshore or forest edge, or political ones, such as city limits, make convenient boundaries, but limits can just as well be arbitrary so long as they can be accurately designated in a geometric sense.

The box is not all there is to the ecosystem, because if the box were an impervious container, its living contents (lake or city) would die.

It is important to emphasize that it is the diversity of ecosystem functions including microbial recycling, inputs and outputs as well as habitats and human land uses that need to be maintained, not just the diversity of species or biodiversity in the narrow sense.

III. THE HUMAN-DOMINATED TECHNO-ECOSYSTEM

Current urban-industrial society not only impacts natural life-support ecosystems in negative and sometimes positive ways, but has created entirely new arrangements that we can call techno-ecosystems that are competitive with and parasitic on natural ecosystems. These human-made systems involve new, powerful energy sources, technology, money, and fuel-powered cities that have little or no parallel in nature. It is imperative that techno-ecosystems interface with natural life-support ecosystems in a more positive or mutualistic manner than is now the case if our rapidly growing urban-industrial society is to survive in a finite world.

Before the industrial revolution, humans were a part of—rather than apart from—natural ecosystems. In the ecosystem model of Fig. 3, humans functioned as top predators and omnivores (the terminal H box in the food web). Early agriculture, as is the case with traditional or preindustrial agriculture as still widely practiced in many parts of the world, was compatible with natural systems and often enriched the landscape in addition to providing food. But with the increasing use of fossil fuels and atomic fission—energy sources many times more powerful than sunlight—together with the mushrooming growth of cities and increasing use of money-based market economics as the basis for decision making, the model of Fig. 3 is no longer adequate. We need to create a new model for this techno-ecosystem, a term suggested by pioneer landscape ecologist Zev Naveh (1982).

Figure 4 is our graphic model for these new (in terms of human history) fuel-powered systems. It includes the four components listed earlier: powerful energy sources, technology, money, and cities. The model shows the inputs of the new fuel energy sources and natural resources, and the increasing outputs of air, water, and solid waste pollution that are very much larger and more toxic than anything that comes out of natural ecosystems. In Fig. 5 we add to the techno-ecosystem model some natural ecosystems that provide life-supporting goods and services (breathing, drinking, and eating!) and that maintain homeorhetic (i.e., pulsing) global balances in the atmosphere, soils, freshwater, and oceans. Note that money circulates as a two-way flow between society and human-made systems, but not natural systems, thereby creating a vast market failure when society fails to pay for ecosystem services.

A modern city,1 of course, is the major component of the fabricated techno-ecosystem. It is a very energetic hot spot that requires a large area of low-energy natural

---

1The term “city” is used synonymously with the geographers’ term “standard metropolitan district (SMD),” which includes industrial areas and residential suburbs that often extend far beyond official city limits.
FIGURE 4 A human-dominated techno-ecosystem.

FIGURE 5 A human-dominated techno-ecosystem and natural ecosystem.
and seminatural countryside to maintain it. Current
cities clean and recycle no air and or water (to the
point of being redrinkable), grow little or no food, and
generate a huge waste stream that impacts wide areas
of downstream rural landscapes and oceans. The city
does export money that pays for some natural resources,
and the city provides many desirable cultural instru-
cations, such as museums and symphonies, that are not
available in rural areas.

In summary, cities are essentially parasites on the
low-energy countryside. To call a city a parasite is not
to belittle it, but to be realistic. In undisturbed nature,
parasites and hosts tend to coevolve for coexistence;
otherwise, if the parasite takes too much from its host,
both die if the parasite has only one host. Currently
humans have only one habitable host—the earth.

Especially threatening to the global life-support eco-
systems is the explosive growth of mega-cities in the
less-developed nations, caused in part by the increasing
dominance of another techno-ecosystem, that of indus-
trialized agriculture, with its often excessive consump-
tion of water and use of toxic and enriching chemicals.
These systems produce more food products per unit of
space, but in turn are prodigious polluters and by their
economic might drive small farmers out of business
worldwide, forcing them into cities that are unable to
assimilate them. This current situation illustrates what
engineer and former president of MIT Paul Gray (1989)
has written: “A paradox of our time is the mixed blessing
of almost every technological development.” In other
words, technology has its destructive as well as benefi-
cial side. To bring the natural and technical ecosystems
into a mutualistic relationship will be society’s greatest
challenge in the twenty-first century.

See Also the Following Articles
ECOSYSTEM SERVICES, CONCEPT OF • ENERGY FLOW AND
ECOSYSTEMS • HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW

Bibliography

and Natural Resources 10, 369–382.
ology and the Environment, pp. 192–205. National Academy
Press, Washington, D.C.
Lotka, A. J. (1925). Elements of Physical Biology. Williams and Wil-
kins, Baltimore.
of North Carolina Press, Chapel Hill.
Holt, New York.
ecosystem as an organizing concept in ecosystem management.
Society and Natural Resources 10, 347–367.
Sukachev, V. N. (1959). The correlation between the concepts “forest
ecosystem” and “forest biogeocoenose” and their importance for
classification of forests. IX International Botanical Congress 2, 387.
Tansley, A. G. (1935). The use and abuse of vegetational concept
ECOSYSTEM FUNCTION MEASUREMENT, AQUATIC AND MARINE COMMUNITIES

John T. Lehman
University of Michigan

I. Conceptual Framework
II. Biological Comparisons between Freshwater and Marine Ecosystems
III. Primary Production
IV. Secondary Production
V. Biogeochemistry and Nutrient Cycling
VI. Stable Isotopes

GLOSSARY

aerobic In the presence of oxygen.
allochthonous Imported from outside the ecosystem.
biogeochemical Consisting of biological and abiotic transformations.
chemoautotrophy Use of energy-yielding chemical reactions as an energy source for synthesis of organic matter from inorganic precursors.
cohort A group of individuals of the same age that can be identified within a population.
compensation depth Depth where photosynthesis and respiration are in balance.
euphotic zone Water depth with sufficient light for photosynthesis.
photoautotrophy Use of light as an energy source for synthesis of organic matter from inorganic precursors.
plankton The community of organisms suspended in the water.
production Newly formed biomass of a population or trophic level, including the organic matter eliminated during the period of observation.
trophic level Position in a food chain, defined by the number of energy transfer steps to that level.
trophogenic region Region where net production of organic matter occurs by photoautotrophy or chemoautotrophy.
tropholytic region Region where respiration and decomposition of organic matter proceed in the absence of primary production.

INVESTIGATIONS OF ECOSYSTEM FUNCTION in fresh and saltwater have been dominated by two major themes. The themes originate in two different ways of thinking about biological production. Biological processes can be regarded fundamentally as conduits for the flow of mass and energy across trophic levels, or as the mechanisms by which individuals and populations maintain their existence. These processes include photosynthesis, mineral nutrient uptake, gathering of food, and extracting nutrition from it, as well as life history patterns, behavior, and survival strategies.

I. CONCEPTUAL FRAMEWORK

Investigation of biological mass flux and energy transfer relies on the first law of thermodynamics. Inquiries...
rooted in this theme require basic accounting for inputs and outputs balanced on carbon mass, nitrogen mass, or the chemically bound potential energy present in organic matter. The different bases for accounting are interlinked by the common stoichiometry of all protoplasm. As Alfred Redfield and his colleagues began to point out in the 1930s, there are consistent ratios among carbon, nitrogen, and phosphorus in living matter drawn from lakes and oceans. The ratios are not as rigid as the elemental composition of a crystalline mineral, but they are reliable within limits. They stem from the fact that organisms exist as biochemical aggregates of proteins, lipids, carbohydrates, and nucleic acids. The basic biochemical building blocks are universal, so the scope for differences among species is limited. If a class of organisms dominates the organic nitrogen pool within an ecosystem, it likely dominates in terms of organic carbon or calories as well. Thus the different measurement bases converge to common results.

The energy and mass-based approaches offer consistency of accounting, but they are intrinsically abiotic. They are indifferent to the forces that govern self-propagating genetic entities, for which the law of natural selection is as important as the laws of thermodynamics. Survival, persistence, and production of viable offspring are properties that define successful species, whether or not they dominate ecosystem biomass and material flux. An associated theme in measuring ecosystem function focuses on the properties of genetic entities rather than on the properties of conservation laws. This theme is keyed to the processes that add or subtract individuals within an ecosystem, processes such as birth, death, or migration, as well as the explicit match between individual age and the passage of time. Viewed through a lens that traces heritable lines of descent, mass and energy transformations are incidental consequences. Transformations of energy and material are the byproducts of life struggles whose object is indefinite persistence.

In 1961, G. E. Hutchinson coined the phrase “paradox of the plankton” to define the challenge to explanation presented by the species diversity of plankton communities. Nets towed from several hundred meters depth to the surface in the Pacific Ocean north of Hawaii, for example, routinely collect more than 300 species of zooplankton, and the richness of phytoplankton species in the surface waters is equally great. The oceanic pelagic region is the oldest continual habitat on the planet and its denizens are the product of ceaseless natural experiments. Near coastal regions and in lakes and rivers, the permanent plankton are joined by the larvae of benthic organisms. To Hutchinson, the con-temporaneous existence of so many distinct genetic lines in habitats that lack obvious structural complexity begged important questions about coexistence, interactions, competition, and resource use. Faced with a bewildering array of potential species interactions, investigations of the marine pelagic ecosystems have emphasized mass and energy flow rather than individual species dynamics. Compounding the problem of overwhelming diversity is the fact that marine plankton are notoriously difficult to census accurately, owing to the physical movements of water masses, to such an extent that some workers believe that time series data that are essential to population studies are nearly impossible to gather.

In an effort to span the breach between measurements of organic composition in units of carbon or nitrogen and the realities of the ways matter is packaged as individual organisms, there has been great interest in finding rules of general validity and broad predictive strength that blend the approaches. Special attention has been paid to relationships based on individual body size or biomass. These size-based or allometric models of ecosystem processes hold the promise of predicting metabolic rates and trophic interactions from simple measurements. Extensive tabulations have become available for physiological processes like respiration, motility, body growth, and feeding rates versus individual size. This activity has spurred the drive toward models of ecosystem dynamics that treat the size structure of organisms present in a system rather than the phylogenetic composition of the communities.

II. BIOLOGICAL COMPARISONS BETWEEN FRESHWATER AND MARINE ECOSYSTEMS

Freshwater ecosystems exhibit reduced species richness compared with marine systems. Many important invertebrate groups (e.g., Echinodermata, Ctenophora, Chaetognatha) have failed to colonize freshwater habitats, although in some cases their roles have been assumed by successful radiation of aquatic insects, particularly the Diptera, in lakes and streams. Much of the difference may be owed to the greater depth, antiquity, and continuity of oceanic plankton environments. The role of age alone is problematic, however. The great ancient lakes of the planet such as Baikal, Tanganyika, and Malawi, with basin ages measured in hundreds of thousands to a few millions of years, exhibit endemic species radiation of some groups but not all. The zo-
plankton faunas of the ancient lakes are notably undiversified, despite rich endemism among some fish, mollusks, or amphipods. Not only are there differences in overall species richness, but there are some differences in latitudinal trends as well. In general, the diversity of marine plankton is greater at low latitudes than at high latitudes. The latitudinal trend in species richness for freshwater plankton is the opposite. Tropical lakes have abbreviated zooplankton faunas compared with temperate sites. The tropical lakes are depauperate of large-bodied Cladocera and copepods, but they are poor in small-bodied pelagic rotifers as well.

III. PRIMARY PRODUCTION

A basic unit of measurement, and also a fundamental basis for comparisons among aquatic ecosystems, is the rate of primary biological production. Several methods are used, but they rely on the overall chemical reaction of carbon fixation in photosynthesis:

\[ 12 \text{H}_2\text{A} + 6 \text{CO}_2 \rightarrow \text{C}_6\text{H}_{12}\text{O}_6 + 12 \text{A} + 6 \text{H}_2\text{O} \quad (1) \]

where \( A \) represents an element that serves as electron donor for the photochemical oxidation-reduction reaction. In photoautotrophic reactions of algae and bacteria in lakes and oceans, oxygen, sulfur, or reduced organic compounds are used as the electron donors. Methods for measuring primary production differ depending on whether the focus is on the transformation of inorganic carbon into organic matter, the resulting release of oxidized product, or changes in the internal cellular redox system or photochemical state. The different methods are not equivalent, and the stoichiometry of carbon fixed to oxidation product released is not strictly 1 to 2 as indicated in Equation 1. Equation 1 is a simplification of the true synthetic reactions involved with cellular growth. Synthesis products are not only hexose sugars, but include all sorts of carbohydrates, as well as proteins, lipids, and nucleic acids. The photosynthetic quotient (PQ) of oxidized product released to carbon fixed varies according to the dominant synthesis products.

In practical application, measurement of primary production relies on one of two approaches. Either measurements are made on subsets of the natural community, which are enclosed and isolated for an experimental time duration, or the measurements are made directly on the natural community without isolation.

Both approaches have advantages and potential complications.

A. Oxygen-Based Methods

To avoid the complexity of diffusion or advection of dissolved substances and gases in or out of a water parcel, most measurements of primary production are conducted in enclosures. During the 1920s, oceanographers introduced one such method that is still in wide use: the light and dark bottle oxygen method. The method relies on making three measurements of oxygen concentration: (a) the initial concentration of dissolved oxygen in a water sample at the start of the experiment, (b) the final concentration of oxygen in a water sample that was enclosed in a transparent bottle called the Light bottle, and (c) the final concentration of oxygen in a water sample enclosed in an opaque bottle called the Dark bottle. From these three experimental measurements it is possible to deduce rates of respiration (\( R \)), gross primary production (GPP), and net primary production (NPP):

\[ R = \frac{([O_2]_{\text{light}} - [O_2]_{\text{dark}})}{\Delta t} \quad (2) \]
\[ \text{GPP} = \frac{([O_2]_{\text{light}} - [O_2]_{\text{dark}})}{\Delta t} \quad (3) \]
\[ \text{NPP} = \frac{([O_2]_{\text{light}} - [O_2]_{\text{dark}})}{\Delta t} \quad (4) \]

Where \( \Delta t \) is the duration of the incubation. The experimental bottles are either suspended at specified depths in the water column at the sampling site for duration of the experimental incubation, or they are placed in light and temperature conditions that simulate submersine conditions. The duration of experimental incubation \( \Delta t \) is invariably a compromise between the desire to obtain measurable changes in oxygen concentration and the desire to minimize artifacts resulting from prolonged exposure, such as growth of microorganisms on bottle surfaces or development of supersaturated oxygen concentrations inside the bottles. Typical incubation durations range from 4 to 24 hr.

B. Carbon-Based Methods

Methods that record fixation of inorganic carbon into organic matter rely either on measuring the removal of inorganic carbon or the appearance of new carbon in the organic matter. The pool of inorganic carbon available for photosynthesis in aqueous solution includes not only dissolved aqueous carbon dioxide, but also the reaction products of carbon dioxide with water molecules.
Carbon dioxide reacts with water to form carbonic acid:

$$\text{CO}_2 + \text{H}_2\text{O} = \text{H}_2\text{CO}_3$$  \hspace{1cm} (5)

Carbonic acid is a weak acid that dissociates to form its conjugate bases bicarbonate and carbonate:

$$\text{H}_2\text{CO}_3 = \text{HCO}_3^- + \text{H}^+$$  \hspace{1cm} (6)

$$\text{HCO}_3^- = \text{CO}_3^{2-} + \text{H}^+$$  \hspace{1cm} (7)

At equilibrium conditions the concentration ratio of carbon dioxide to carbonic acid is about 600 to 1. Proportions of carbon dioxide to bicarbonate to carbonate vary widely, but at the environmental pH of the ocean and many lakes, bicarbonate concentrations exceed all other molecular carbon species. The pool of dissolved inorganic carbon (DIC) available to photosynthesis is defined as

$$\text{DIC} = \text{CO}_2 + \text{HCO}_3^- + \text{CO}_3^{2-}$$  \hspace{1cm} (8)

The three dominant chemical species that comprise DIC exist in known stoichiometric relationships to each other as functions of pH, temperature, and ionic strength of the aqueous solution. There are several ways to measure or calculate DIC. An aqueous sample may be acidified, transforming all DIC to carbon dioxide, and the resulting gas can be measured. Alternatively, pH and CO$_2$ may be measured and then the other species are calculated. Or the concentrations of the conjugate bases can be measured by acid titration, and from their total plus measured pH, the DIC is calculated.

### C. C-14 Method

An alternative to measuring oxygen by light and dark bottle method is a method that uses radiocarbon as $^14$C-bicarbonate as a tracer. A small but known amount of radiocarbon is introduced to a water sample and the bottle is incubated for a time period in the same manner as with the oxygen method. At termination of the incubation period, the contents of the experimental bottles are filtered. Radiocarbon retained on the filters is considered to represent inorganic carbon fixed into particulate organic matter during the experiment. Any fixed carbon that leaked into the water from algal cells can also be assessed by acidifying the filtrate under vacuum to drive residual inorganic carbon out of solution. The radioactivity of residual organic release products can then be measured. The beta particles emitted from disintegrating C-14 nuclei are detected and counted by various methods, including gas or planchet Geiger-Muller (ionization) detector, liquid scintillation, or track autoradiography.

Calculation of total carbon fixation by the C-14 method requires knowledge of the specific activity of the inorganic carbon pool and of the isotopic fractionation differences between C-12 and C-14. Specific activity is the ratio of added radiisotope activity to the mass of stable isotope that the radiisotope is intended to trace. In this method the radiocarbon traces the pool of DIC. Uptake of inorganic carbon by photosynthetic organisms involves the kinetic process of diffusion and active transport across cell membranes. In such processes there are differences in the mobility of isotopes of the same element owing to differences in molecular mass. Empirical study has demonstrated that the ratio of C-12 uptake to C-14 uptake is approximately 1.06. Accordingly, the biological fixation of inorganic carbon is calculated as

$$\frac{\text{DIC fixed}}{\text{(total inorganic radiocarbon added})} = 1.06 \times \left(\frac{\text{DIC in the sample}}{\text{radiocarbon fixed}}\right)$$  \hspace{1cm} (9)

Both light and dark bottles are used in estimation of primary production by the C-14 method, but interpretation of results from the dark bottles are quite different from interpretations for the oxygen method. In the case of C-14, the dark bottle does not measure respiration, but rather it records the rate of nonphotosynthetic carboxylation reactions by algae and bacteria. In simultaneous comparisons between C-14 and oxygen methods, it appears that C-14 measures something intermediate between net and gross primary production. The reason for this is that some but not all of the carbon fixed during the experiment is preferentially respired during the incubation period.

### D. Fluorescence Methods

Indirect methods for assessing primary production rely on the fact that photosystem pigments alternate between a ground state in which they are able to absorb excitation photons of characteristic wavelengths and an energized state in which they either transfer the excitation energy to a chemical reaction or fluoresce a photon at longer wavelength and return to their ground state. The proportion of pigments in one state or the other varies with the rate of primary production expressed per unit mass of pigment. The usual reference
pigment is chlorophyll $a$, the antenna pigment present in all aerobic photoautotrophs. Application of indirect approaches requires calibration against direct measurements of primary production. Their attraction is their potential to estimate primary production in natural communities without isolation.

E. Interpretation and Analysis of Primary Production Results

Rates of primary production vary with depth according to light intensity, temperature, algal biomass, and physiological state of the algae. The photosynthetically active spectrum includes light from 400 to 700 nm. Near surface, light intensities are often sufficient to saturate the antenna pigments with photons. At high light, rates of photosynthesis are limited by the rates of biochemical dark reactions, specifically the carboxylation reactions, rather than by interception of light photons. Because carboxylation is enzyme dependent, the maximum rates of carbon fixation are temperature dependent and they vary with nutritional status as well.

Light intensities diminish with depth. At low light, the rate of primary production is limited by the rate at which photons are intercepted, and carbon fixation varies directly with light intensity. Temperature variations are less important at low light because in photochemical reactions limited by interception of photons, the role of temperature-dependent enzyme kinetics is minimized.

Respiration rates vary with temperature. Near the surface, as a daily average, gross primary production often exceeds respiration and net primary production is positive. At greater depth, however, gross primary production declines and a depth is reached at which net primary production equals zero. This depth is termed the compensation depth. For purposes of predicting the physiological state and growth response of the algae, it is important to recognize that organisms in aqueous suspension do not always maintain fixed vertical positions in the water. What is most important is the integral difference between the rates of gross primary production and of respiration during 24 hours:

$$ \int_0^{z_{mix}} \frac{\partial}{\partial t} \text{GPP}(z, t) - \text{R}(z, t) \, dz \, dt $$

where $z_{mix}$ is the maximum depth of water involved in vertical mixing during 24 hours. When the integral defined by Equation 10 is positive, net growth will occur. Seasonal variations in photoperiod, water transparency, temperature, and species composition influence the components of Equation 10, and so the critical mixing depth for net positive algal growth varies likewise.

IV. SECONDARY PRODUCTION

The principle of mass and energy balance used in primary production studies can be extended to heterotrophs. Among herbivores, predators, and detritivores, material and energy are considered commodities that continually turn over through biochemical means and by replacement of individuals within a population. Secondary production is the total growth increments over a given time interval as experienced by all individuals within a population that were alive at the beginning of the interval, whether or not they survive the complete interval. Unlike primary production, there is no distinction between gross and net production; secondary production ($P$) is inherently a net balance of several factors:

$$ P = I - E_g - E_x - R - M - G $$

where $I$ is ingestion, $E_g$ is egestion of unassimilated food, $E_x$ is excretion, $R$ is respiration, $M$ is molts or exuvia, and $G$ is gametes and reproductive products such as egg yolk and spermatophore capsules.

The efficiency of energy or material transfer during secondary production is expressed by several quotients. Assimilation efficiency is the ratio of assimilation ($I/E_g$) to ingestion ($I$). Net growth efficiency is the ratio of growth ($P$) to assimilation. Gross growth efficiency is the ratio of growth ($P$) to ingestion ($I$). Ecological efficiency is sometimes defined for predator-prey trophic interactions. Ecological efficiency is the ratio of predator production to prey production.

A. Bioenergetic Analysis

Bioenergetic analysis places its frame of reference on individual organisms of specified size, weight, or age. When Equation 11 is applied at the individual level, the result ($P$) is incremental growth. The units of accounting can be total body mass, content of carbon, nitrogen, or phosphorus, or calories. Analysis is facilitated if each of the terms on the right side of Equation 11 can be expressed with reference to body mass. For example, ingestion rate may be expressed as a function of food types and availability, temperature, and body mass of the consumer. Respiration may be expressed...
as a function of temperature, activity level, and body mass. In this way, the incremental growth of individual organisms is linked with organism size and a set of environmental conditions. The approach can then be extended to the population level by summing calculated growth increments across the empirical size distribution of the population.

**B. Biomass Accrual**

The standing stock biomass of a population is the sum of the weights of all individuals in the population. Changes in biomass occur owing to changes either in number \( (N) \), or mean weight \( (W) \):

\[
\Delta \text{ Biomass} = \Delta (N \times W) = N \Delta W + W \Delta N \tag{12}
\]

Equation 12 suggests a means to calculate secondary production for populations consisting of distinct cohorts. Presence of distinct cohorts is often the case among long-lived, seasonally reproducing populations. In such cases, a survivorship chart or life table for each cohort (decreasing \( N \) versus time) defines \( \Delta N \). A parallel chart or table reporting the average individual weights of cohort members versus age or time defines \( \Delta W \).

When cohorts cannot be distinguished, as happens with continually reproducing populations, secondary production can be calculated by summing the growth increments \( \Delta w_i \) of individual life history stages:

\[
P = N_1 \Delta w_i / T_i + N_1 \Delta w_j / T_j + N_i \Delta w_k / T_k + \ldots + N_i \Delta w_k / T_k
\]

where the subscripts denote individual stages (e.g., instars) or size classes and where \( T \) represents the mean duration of each stage. Abundance at each stage \( N_i \) is considered to be the mean stage-specific abundances during the period of interest.

**C. Birth and Death Rate Analysis**

Numerical changes in populations can be represented as a simple function of population size \( (N) \):

\[
\frac{dN}{dt} = r \times N \tag{14}
\]

where \( r \) is the net intrinsic population growth rate \( (\text{time}^{-1}) \) resulting from additions and removals:

\[
r = \text{birth} - \text{death} + \text{immigration} - \text{emigration} \tag{15}
\]

In cases where changes owing to migration are not an issue, net population growth amounts to the balance between birthrate \( (b) \) and death rate \( (d) \):

\[
r = b - d \tag{16}
\]

Over short time intervals an assumption is made that birth and death rates are stationary, so that the simple integral solution to Equation 14 is

\[
N(t + \Delta t) = N(t)e^{rt} \tag{17}
\]

and net intrinsic population growth rate \( r \) can be calculated from population sizes at the beginning of any time interval \( \Delta t \):

\[
r = \ln[N(t + \Delta t)/N(t)]/\Delta t \tag{18}
\]

To decompose the net rate \( r \) into its individual components, an independent estimate is needed for either birth or death rate.

**D. Egg Ratio Method**

The egg ratio method, pioneered by W. T. Edmondson, calculates birthrate from demographic properties of populations in nature. The method is particularly successful in application to parthenogenic populations, especially to species in which females produce and brood clutches of eggs, such as among rotifers and Cladocera.

The egg ratio, \( E \), is defined as the ratio of eggs and embryos to the sum of all postembryonic stages (juveniles and adults) in the population. From egg ratio and duration of development for the embryonic stage \( (D) \), the best estimate for population birthrate \( (b) \) is

\[
b = \ln(E + 1)/D \tag{19}
\]

Derivation of Equation 19 requires assumption of a stable age structure for the population, and further that all embryonic and postembryonic stages experience a common death rate \( (d) \), such that \( d \) can be calculated from \( r \) and \( b \) by difference:

\[
d = b - r \tag{20}
\]

These assumptions are most nearly applicable to short-lived species with abbreviated juvenile stages. Accuracy of the birthrate estimate is improved if the egg ratio is based on a census of egg stages that are in final phases of development and if the development time is corre-
spondingly restricted in range. Most of the populations to which this analysis is applied are ectotherms, so egg development times are strong reciprocal functions of water temperature.

E. Instar Analysis

For populations that exhibit distinct morphological stages, such as size classes or the developmental instars of arthropods, it is possible to estimate stage-specific mortality. This method requires knowledge of the durations ($T_i$) of each stage including their variation with time and water temperature. The numerical change in each stage is defined as

$$\frac{dN}{dt} = N_i \int_{T_i-1}^{T_iT_i} - \frac{dN}{dN}$$

(21)

The first term on the right side of Equation 21 represents recruitment of new individuals into the stage or age class from the preceding stage. The second term is recruitment out of the stage, and the third term is mortality within the stage. This approach assumes a uniform age distribution within each stage. It further assumes that the calculation time step is small compared to the stage durations $T_i$.

Instar analysis may be combined with birthrate estimation by the egg ratio method to estimate recruitment into the youngest instar category. Addition of mass or energy information about each instar or stage permits further calculation of secondary production according to Equation 13.

V. BIOGEOCHEMISTRY AND NUTRIENT CYCLING

Transformations of matter and energy in aquatic ecosystems are dependent on, and in turn influence, the availability of all the elemental constituents of protoplasm including both major components like carbon and nitrogen as well as trace constituents like iron and manganese. Even elements and compounds that are not essential constituents of biological material are affected by ecosystem processes. For example, redox reactions and pH changes caused by biological reactions can influence the solubility and reactivity of metals like mercury and cadmium, which can become toxic.

Alfred Redfield is credited with the conceptual theory of “biochemical circulation” within ocean ecosystems, and the theory is equally applicable to fresh waters. According to this view, processes of synthesis and decomposition may be separated in space and time, but they remain linked. Organic matter is generated at the ocean surface where radiant energy from sunlight permits positive net primary production. Carbon is assimilated from inorganic sources into organic form, together with nitrogen, phosphorus, silicon, iron, and all the other constituents of protoplasm. Over time, a portion of this synthesized organic matter will be transported actively or passively to deeper depths in perpetual darkness where photosynthesis is impossible. Once there, the allochthonous organic matter is used in secondary production, but its ultimate fate is to be recycled to inorganic mineral form or to be buried in sediments for long periods of geological time. Organic matter that is not transferred to the depths is recycled in situ, owing to the production inefficiencies of multiple secondary production pathways including herbivory, predation, and detritivory.

Physical processes of upwelling and thermohaline circulation permit some mineral nutrients dissolved in deep water to be returned to the surface for another round of synthesis reactions. Over geological time scales even the elements buried in the ocean sediments can become available again through weathering of the ancient deposits. This alternation between the inorganic state of an element and a state in which it is associated with organic processes at both biological and geological time scales is the basis for the term biogeochemistry.

For elements such as carbon and nitrogen, organic transformations entail changes in oxidation state and allied chemical properties. For others such as phosphorus and silicon, transformations affect not oxidation state but rather solubility.

Elemental nutrients that are essential for the production of organic matter can be traced through ecosystem transformations owing to the principle of conservation of matter. Quantitative illumination of the pathways has led to production of descriptive “cycles” for each element. Different aquatic ecosystems exhibit characteristic magnitudes of different pathways. Even water itself is involved in a hydrological cycle, in which water alternates among liquid, gas, and solid phases, and among ocean, atmosphere, and terrestrial environments. Water evaporates from the sea surface to the atmosphere, from where some of it falls as rain on the continents.

Surface and groundwater on the continents react with carbon dioxide from the air and from respiration of soil biota. The resulting solution thus includes carbonic acid. In some regions the carbonic acid is supplemented by strong acids introduced by anthropogenic activities. For example, combustion of fossil fuels introduces oxides of nitrogen and sulfur into the atmosphere, because
the elements were components of the ancient prote-
plasm that underwent diagenesis to form the fuel
sources. After reacting with atmospheric water vapor,
the oxides enter aqueous solution as nitric and sulfu-
ric acids. The resulting mix of weak and strong acid is the
primary agent of chemical weathering, a process that
dissolves the rocks of the continents. Chemical weather-
ing accounts for most of the ions in river water.

Elements that enter the ocean by runoff from the
continents or as wind-blown dust are subject to further
reactions and removal processes. Biological processes
are prominent among these. Microbiological processes
reduce sulfate to sulfur, followed by loss of hydrogen
sulfide gas or chemical precipitation of metal sulfides.
Inorganic carbon is transformed into particulate organic
matter or is precipitated from solution with calcium as
the skeletons of marine organisms. Silica is precipitated
likewise, as cell wall covering or skeletal material for a
variety of plants and animals.

Inputs of new material and removal processes are in
close balance, for the bulk chemical composition of the
ocean appears to be in steady state. The relationship
between material flux and pool sizes in biogeochemical
cycles is expressed through the concept of residence
time, defined as

\[
\text{Residence time (yr)} = \frac{\text{Amount of a given element in the ocean}}{\text{Rate of addition or removal per year}}
\]

Elements that exhibit great biological reactivity, such as
the nutrient elements N, P, and Si, have short residence
times and are usually present at low concentrations.
Elements with lower significance to biological reactions
like Na and Cl have long residence times, and corre-
spondingly account for most of the dissolved salt in
sea water.

### A. New and Regenerated Production

In the late 1960s, R. C. Dugdale and J. J. Goering intro-
duced a conceptual model of biological production in
the surface ocean euphotic zone. The model represented
the euphotic zone as a compartment or box with inputs
and outputs of nitrogen. The inputs were delivered
via upwelling, nitrogen-fixation of diazo-nitrogen, and
river runoff from the continents. Outputs were the sink-
ing material consisting of dead plankton and fecal pel-
lets. In addition to the inputs and outputs, they de-
scribed internal recycling processes within the euphotic
zone, through which nitrogen was regenerated by zoo-
 plankton and bacteria.

To maintain steady state, the outputs from the eu-
photic zone must equal the inputs, irrespective of the
magnitudes of regeneration. Dugdale and Goering rea-
sioned that primary biological production that was sus-
tained by the inputs, or “new” nitrogen, should be called
new production, and that the primary production sus-
tained by regenerated nitrogen should be called regener-
ated production. “New” nitrogen typically arrives in
the form of nitrate, a form that is thermodynamically
favored at the redox potential of surface sea water, and
is the form to which organic matter is decomposed
derunder aerobic conditions. “Regenerated” nitrogen is
typically ammonium, which is at the same reduced
oxidation state as the primary amino nitrogen that com-
prised proteins. Most aquatic animals are ammonotelic
in their excretion metabolism. Thus, the forms of nitro-
gen used to support primary production of organic mat-
ter are a clue to the processes that dominate production.

The concepts were further refined by R. W. Eppley
and B. J. Peterson into a modern ecosystem paradigm
of biological oceanography. New production is sus-
tained by imported nutrients. It is considered to be
equal to the sum of sinking flux of organic matter plus
transfers of organic matter to higher trophic levels,
which may be exported from the ecosystem by, for
example, fisheries yields. By empirical evidence, ecosys-
tems dominated by “new” production exhibit higher
primary production overall and are more capable of
sustaining exploitation such as harvest of fisheries
stocks. Ecosystems dominated by “regenerated” pro-
duction cannot endure prolonged exploitation of their
sustaining elements.

The concept of new production and its linkage with
export production has further implications for global
biogeochemical cycles and specifically for carbon diox-
ide levels in the global atmosphere. Organic matter that
is exported from the surface ocean to the deep sea
represents a sink for DIC from the surface ocean. Defi-
cuts of DIC are replenished in part by dissolution of
carbon dioxide from the atmosphere. As a consequence
of this linkage, the fraction of global oceanic primary
production that is new production represents an upper
limit to the amount of carbon that can be exported and
sequestered in the deep ocean.

Techniques for measuring the rates of nutrient re-
generation differ depending on whether the measure-
ments are conducted in trophogenic regions or in tro-
pholytic regions. In tropholytic regions, primary
production is minimal and thus regeneration can be
assessed by measuring the accumulation of mineral nu-
trients directly. In trophogenic regions, synthesis reac-
tions dominate over regeneration, and regenerated nu-
trients are quickly reassimilated. Techniques for measuring nutrient recycling under such conditions rely on assessment of uptake and regeneration simultaneously, such as by isotope dilution. In isotope dilution experiments, a deliberate tracer such as $^{15}$N-ammonium is introduced to a closed system in which $^{14}$N is the predominant natural isotope. Over time, the ratios of $^{15}$N to $^{14}$N in the ammonium pool become diluted as regenerated $^{14}$N-ammonium is added, even while uptake of both isotope forms proceeds.

VI. STABLE ISOTOPES

Stable isotopes of natural elements are ubiquitous. They have often been collected and used as deliberate tracers, such as in studies of new and regenerated production, but they can be measured and used at natural occurrence as well. Owing to the proportions in which different nuclear species survived the first instants of the universe, and the proportions that have been retained in the planetary composition of the Earth, many common elements exhibit more than one stable form. Hydrogen, carbon, nitrogen, oxygen, and sulfur, for example, all have two or more nuclear configurations that differ in numbers of neutrons and hence differ in atomic mass. Oxygen-16, O-17, and O-18 are stable (not radioactive), but they exist in different proportions in the environment. The different forms can be distinguished by mass spectrometry. The preferred form of analysis is by isotope ratio mass spectrometer, which compares the stable isotope ratios in a sample with ratios in a standard.

Measurement of stable isotope ratios in natural materials has been a technique used by geochemists for decades to help interpret complex processes. The relative proportions of the stable isotopes vary from one medium or substance to another. Mechanisms that can change the proportions of two stable isotopes of the same element are called fractionation processes. Fractionations result from differing mobilities and reactivities by the same chemical element owing to differences in atomic mass. Both kinetic and thermodynamic factors affect the magnitudes of fractionation. Five general rules apply:

1. Elements with high atomic mass generally exhibit smaller fractionation among their isotopes than do lightweight elements.
2. If multiple isotopes of an element exist, fractionation is greatest between isotopes that have the largest atomic mass difference.
3. In unidirectional reactions, the lighter isotope becomes enriched in the endpoint.
4. In reactions that reach equilibrium, the lighter isotope is enriched in the reactant compound or phase that has the weaker bond strength.
5. As temperature increases, fractionation decreases asymptotically to zero.

Changes in the fractionation ratios with temperature have been the basis for using stable isotopes in geochemical studies as "paleothermometers".

Atmosphere, terrestrial, and marine environments are heavily dominated by single isotope forms of common elements, and the alternate isotopes are usually very rare. For example, in the atmosphere, O-16, O-17, and O-18 co-occur in overall ratios by atoms as 1 : 0.0004 : 0.002. Similarly, C-12 and C-13 occur in atmospheric carbon dioxide in the mean ratio 1 : 0.011. Given such disparities of magnitude, isotope ratio differences between samples and standard are calculated in $\delta$ notation with units of parts per thousand, or per mil ($\delta$‰). For example,

$$\delta^{18}O = \left(\frac{^{18}O/^{16}O}_{\text{sample}}/^{18}O/^{16}O_{\text{standard}} - 1\right) \times 1000$$

Isotopes are useful in ecosystem studies because physical reactions and enzyme reactions cause reproducible discrimination between isotopes. Reaction differences lead to differences in the composition of biological material, skeletal products, and inorganic nutrient pools. Stable isotopes have been used to infer and trace the origins and pathways of materials in aquatic ecosystems. They are useful when different sources are isotopically unique and when the sources change in predictable ways. Successful applications require that the magnitudes of all isotope fractionations that result from physical, chemical, and biological processes be known.

Isotopes of carbon and nitrogen have been used to trace food-web relations in freshwater and marine environments. Phytoplankton discriminate against C-13 during photosynthetic carbon fixation, just as they discriminate against C-14. The magnitude for discrimination against C-13 is less than for C-14, or about $-20$ to $-30 \delta$‰ with respect to the source DIC. Organic matter that originates from terrestrial sources sometimes has very different carbon isotope ratios than material that is of aquatic origin. Terrestrial plants fractionate atmospheric CO$_2$ differentially during photosynthesis according to whether they use C3, C4, or CAM pathways for carbon fixation. Similarly, organic matter generated microbially from methane as a

...
substrate typically exhibits unique carbon isotope ratios. Stable isotopes may thus be used to identify chemosynthetic pathways as well as photosynthetic ones. Carbon isotopic ratios in animals resemble those of their diets within about 1/100. Slight enrichment of C-13 in an animal versus its diet is presumed to be caused by preferential loss of C-12 during respiration or by preferential assimilation of C-13 from the food. Synthesis of lipid rather than protein can also influence the isotope ratios because there are reported fractionation differences between biochemical constituents. As long as the isotope ratios in different potential food sources are well defined and consistently different, the relatively conservative levels of fractionation for carbon by trophic level can indicate whether an animal is eating its food from one source, another source, or a definable mixture of the two.

Nitrogen isotope ratios also reflect the composition of an animal’s diet, but the animal is usually enriched with N-15 by about +5/100 compared with its food. This increased retention of N-15 over N-14 seems to be favored by thermodynamic factors associated with transamination reactions, and by kinetic factors leading to preferential loss of N-14 as excreted ammonium and urea. Reproducible differences in Δ-15N with trophic level provide a diagnostic measure of relative trophic position among organisms that depend on the same primary source of organic matter. Carnivores are isotopically heavier than herbivores, which in turn are isotopically heavier than the algae; omnivores are isotopically intermediate between herbivores and carnivores. Interpretation of nitrogen isotope dynamics can be complicated by temporal and spatial variations in isotope fractionation and nutrient source materials. In laboratory experiments when nutrients are not limiting, isotope fractionations associated with assimilation of nitrate and of ammonium can be large. However, if a nutrient element becomes limiting such that all of it is consumed, there can be no isotopic fractionation. In nature, seasonal alternation between times of nutrient excess and nutrient limitation is common, with the result that isotope ratios can vary greatly in the source material and in the primary producers. Such variations provide useful information about nutrient dynamics, but simultaneously introduce complication to interpretations of trophic relations particularly for long lived organisms.

See Also the Following Articles

BIOGEOCHEMICAL CYCLES • ECOSYSTEM FUNCTION MEASUREMENT, PLANT COMMUNITIES • MARINE ECOSYSTEMS

Bibliography


I. Why Measure Ecosystem Functioning?

II. Ecosystem Resource Dynamics

III. Ecosystem Functioning under Disturbance: Resistance and Resilience

IV. Ecosystem Functioning and Functional Diversity

V. Measurement, Analysis, and Prediction of Ecosystem Functioning: Major Protocols and Obstacles

GLOSSARY

biodiversity Number and composition of genotypes, species, functional types, and/or landscape units present in a given system.

diversity Conceptual view of an assemblage of organisms and of physical and chemical components in their immediate environment, and the flow of materials and energy between them.

ecosystem functioning Flow of energy and materials through the arrangement of biotic and abiotic components of an ecosystem.

ecosystem stability Capacity of an ecosystem to persist in the same state. It has two components: Ecosystem resistance is the ability to stay in the same state in the face of perturbation. Ecosystem resilience is the ability to return to its former state following a perturbation.

resource dynamics Inputs, outputs, and internal cycling of key resources, such as carbon, water, and mineral nutrients, in an ecosystem.

trophic transfer The amount of biomass and/or energy which is transferred from the primary producers to the herbivore-based food chain rather than directly from plants to detritivores.

THIS ARTICLE DEALS WITH the functioning of terrestrial ecosystems with emphasis on primary producers and their relationship to biodiversity. Primary producers, in particular vascular plants, are the major sustainers of terrestrial life. They determine the amount of energy and materials available to terrestrial food webs. Biodiversity in this context involves not only the number and relative abundance of species (taxonomic diversity) but also the kind and relative abundance of traits they possess (functional diversity). These traits influence, and in turn are influenced by, major ecosystem processes. Ecosystem resource dynamics (carbon, nutrients, and water) and ecosystem stability in the face of disturbance (resistance and resilience) are discussed, including conceptual issues and major theoretical and methodological approaches to their study.

I. WHY MEASURE ECOSYSTEM FUNCTIONING?

Ecosystem functioning involves processes such as primary production, trophic transfer from plants to animals, nutrient cycling, water dynamics, and heat transfer. Traditionally, the term has referred to functioning
in equilibrium, namely, the amount (how much), the rate (how fast), and sometimes the seasonal variations of those processes. Some authors believe that the responses of ecosystem processes to perturbation, resilience and resistance, should also be incorporated into the term. This is the approach taken in this article (Fig. 1). Ecosystem processes provide essential goods and services to humankind, such as food; fiber; fodder; fuel; water provision, control, and detoxification; amelioration of weather; soil formation; retention of carbon that otherwise would be released into the atmosphere contributing to climate change; and medicinal, recreational, and cultural resources. The impact of changes in the quantity and quality of services provided by various terrestrial ecosystems to humankind has been gaining increased attention, and in some cases their costs are very high compared with their traditional market value. The measurement of how much seed, green biomass, and animal biomass an ecosystem can produce or how much water and nutrients can be retained in situ are key inputs for an evaluation of ecosystem services to humankind. Concepts, methods, and data retrieval on ecosystem functioning have progressed enormously in the past few decades, and today an impressive body of quantitative information on flows and stocks of materials and energy through a wide spectrum of ecosystems is available. Much less is known, however, of the effects of perturbations on these processes. Physiological processes underlying biomass production, water balance, and nutrient cycling are well understood. However, there are still many theoretical and methodological difficulties and information gaps in the prediction of processes at coarser scales of time (decades or centuries) and space (ecosystems, landscapes, and biomes). In order to determine how the biosphere will respond to the changes in climate, atmospheric composition, and land use projected for the future, accurate predictions at these levels are vital.

**FIGURE 1** Conceptual diagram of the relationships between ecosystem functioning, biodiversity, ecosystem services, and external structuring factors. Rectangular boxes and solid arrows represent the main focus of this article (partially based on Chapin et al., 1998).
II. ECOSYSTEM RESOURCE DYNAMICS

Three main processes will be analyzed here, which are linked to the acquisition and loss of chief resources for terrestrial ecosystems: light, carbon, water, and mineral nutrients. These processes are biomass production (primary productivity and trophic transfer), nutrients (especially nitrogen) cycling, and water dynamics. There is a strong association between water flow, carbon assimilation, and nutrient uptake by a plant. Since carbon dioxide (CO₂) input and water output occur through open stomata, and nutrient and water uptake occur through root hairs and fine roots, they are usually linked. Nutrient availability controls the increase in plant mass, and incorporation of mineral nutrients into biomass is not possible without carbon assimilation. The derived processes at the ecosystem level, namely, biomass production, water balance, and nutrient cycling, are therefore closely linked and basically regulated by net radiation, temperature and precipitation, soil composition and structure, and the plants that are present. Any consideration of carbon, nutrient, or water balance in isolation is probably artificial. The examples presented here are described in different sections for the sake of clarity, but they highlight the strong interdependence of these processes.

A. Approaches to Analysis and Measurement of Ecosystem Functioning

1. Whole-Ecosystem Approaches and Approaches Based on Community Structure and Composition

The measurement of ecosystem functioning focuses on the sizes of major pools of resources, such as water, carbon, and mineral nutrients, and on the rates of flows connecting them. The emphasis on one of these two aspects has varied with time and authors. Not until the 1970s did compartmental models become an important tool in “systems ecology” (illustrations of this view can be found in the classic textbook on ecology by E. P. Odum, 1971). In these models, boxes represent pools or stocks (usually major trophic levels), and arrows signify the flows between them. This represents a first and very useful approximation. However, pools are often treated as “black boxes” in which only the general size, and not the internal composition, is important. This often makes this approach too coarse for the management of real situations and of very little predictive value for novel situations. In most cases, the whole pool is not altered, but instead subtle intrapool changes occur until a threshold is reached or substances are chemically altered or sequestered. This is because organisms which are members of the same pool tend to respond in an individualistic way and have slightly different properties. As a consequence, other approaches have focused on the internal composition of pools (what taxa or functional types are present within the boxes and in what abundance) and how changes in that composition can alter flow rates between pools. These approaches are strongly based on community structure and composition and will be addressed in Section IV.

2. Measurement of Short-Term Resource Dynamics and Long-Term, Large-Scale Ecosystem Processes

Some methods of ecosystem analysis emphasize the measurement of short-term process rates (e.g., photosynthesis, nutrient uptake, and evapotranspiration), whereas others emphasize the measurement of the size of pools accumulated as a result of these processes, with or without consideration of their internal components (e.g., biomass production, nitrogen stocks in different compartments, and water use efficiency integrated over a whole season) (Table I). The choice depends mostly on the objectives pursued and on the scale of observation selected. However, as a general rule, methods based on short-term flows provide more mechanistic understanding, but they do not provide a clear picture of what happens at the time and space scales most relevant to ecosystem functioning (meters to thousands of kilometers and months to decades or centuries). Measurements of flow rates (see measurements in Table I) tend to be very precise but highly variable. They give a snapshot of ecosystem functioning, but they do not necessarily reflect processes over a growing season. For example, short-term (hourly or daily) variation in gas exchange per leaf area is not directly reflected in annual biomass production. Plant growth and vegetation productivity cannot be directly equated with leaf photosynthetic rates. The productivity of plant communities is determined more by the amount of photosynthetic tissue (which is in turn controlled by carbon allocation) than by photosynthetic rate. There is often only a slight relationship between these two parameters. Gas exchange measures the instantaneous plant performance, whereas growth necessarily integrates along time and reflects plant allocation patterns. At the community level, additional regulations occur; therefore, caution is necessary in generalizing the ecological meaning of leaf physiological processes for production. For example, a 10- to 50-fold difference between species in maximum rates of photosynthesis at the leaf level is reduced
TABLE I
Examples of Ecosystem-Level Measurements with Emphasis on Short-Term Flow Rates (F) and on Longer Term Pool Sizes (P)

<table>
<thead>
<tr>
<th>Parameter measured</th>
<th>Usual units</th>
<th>Techniques</th>
<th>Emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biomass production</td>
<td>µmol CO₂ cm⁻² s⁻¹ or mg CO₂ g⁻¹</td>
<td>Gas exchange systems (CO₂, O₂), or simultaneous H₂O and CO₂ exchange</td>
<td>F</td>
</tr>
<tr>
<td>Standing biomass</td>
<td>g m⁻² or t ha⁻¹</td>
<td>Harvest at the end of growing season</td>
<td>P</td>
</tr>
<tr>
<td>Productivity</td>
<td>g m⁻¹ year⁻¹ or t ha⁻¹ year⁻¹</td>
<td>Sequential harvest, equations including climatic and soil factors, satellite imagery</td>
<td>P</td>
</tr>
<tr>
<td>Cover or frequency</td>
<td>% ground area (cover); % total interception (frequency)</td>
<td>Estimation through nondestructive measurement, which can be converted into units of biomass using allometric equations developed on similar neighboring individuals that are destructively harvested</td>
<td>P</td>
</tr>
<tr>
<td>Trophic transfer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer biomass</td>
<td>g or kJ m⁻² or t ha⁻¹</td>
<td>Harvest (with or without replacement); equations including NPP</td>
<td>P</td>
</tr>
<tr>
<td>CO₂ efflux from respiration (estimates micromass biomass)</td>
<td>µmol CO₂ m⁻³ or mm⁻¹ s⁻¹</td>
<td>Gas exchange measurements</td>
<td>F</td>
</tr>
<tr>
<td>Secondary productivity</td>
<td>g or kJ m⁻² year⁻¹ or t ha⁻¹ year⁻¹</td>
<td>Sequential harvest (with or without replacement); animal countings, equations including NPP</td>
<td>P</td>
</tr>
<tr>
<td>Consumption</td>
<td>g or kJ m⁻² year⁻¹ or t ha⁻¹ year⁻¹; % plant tissue; % leaf surface</td>
<td>Plant harvest in herbivore-free and grazed situations; estimation of proportion of plant tissue removed; equations including NPP</td>
<td>P</td>
</tr>
<tr>
<td>Nutrient cycling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decomposition through biomass loss from soil or jars</td>
<td>% initial dw</td>
<td>Sequential weighing of litter samples</td>
<td>P</td>
</tr>
<tr>
<td>Decomposition through CO₂ efflux from soil or jars</td>
<td>µmol CO₂ m⁻³ s⁻¹ or µg CO₂ g dw substrate⁻¹ hr⁻¹</td>
<td>Gas exchange system (soil-only control samples or isotope labeling sometimes used in order to distinguish from root respiration)</td>
<td>F</td>
</tr>
<tr>
<td>Nutrient status</td>
<td>mg nutrient g dw⁻¹ or % nutrient dw</td>
<td>Harvest followed by chemical analysis; estimation through remote sensing</td>
<td>P</td>
</tr>
<tr>
<td>Nutrient use efficiency in litter (amount of biomass produced per unit of nutrient expended and lost; it reflects initial leaf status and efficiency of nutrient resorption)</td>
<td>kg dw mol N⁻¹ in litter</td>
<td>Harvest followed by chemical analysis</td>
<td>P</td>
</tr>
<tr>
<td>Nutrient uptake</td>
<td>mg nutrient g dw⁻¹ year⁻¹</td>
<td>Change in nutrient pool in live biomass over time interval; estimated through sequential harvest followed by chemical analysis of samples</td>
<td>P</td>
</tr>
<tr>
<td>Nutrient uptake potential</td>
<td>µg radiotopes g⁻¹ hr⁻¹; pg radiotopes mg⁻¹ min⁻¹</td>
<td>Soil labeling with radiotopes such as ¹⁵N, ³²P, and ¹³C followed by monitoring of isotope appearance in plant tissue</td>
<td>F</td>
</tr>
<tr>
<td>N transformations in the ecosystem (e.g., &quot;tightness&quot; or &quot;openness&quot; of nitrogen cycle)</td>
<td>δ¹⁵N (‰)</td>
<td>Ratio of stable isotopes ¹⁵N to ¹⁴N in plant material compared with standard (air) using mass spectrometer</td>
<td>P</td>
</tr>
<tr>
<td>Nitrogen in soil available to plants</td>
<td>NO₃⁻ + NH₄⁺ = total nitrogen (µg g⁻¹, %)</td>
<td>Soil sampling followed by chemical analysis, ion exchange resin procedures</td>
<td>P</td>
</tr>
</tbody>
</table>

(continues)
Continued

<table>
<thead>
<tr>
<th>Parameter measured</th>
<th>Usual units</th>
<th>Techniques</th>
<th>Emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mineralization rate</td>
<td>µg N g substrate dw⁻¹ day⁻¹</td>
<td>Tracing of ¹⁵N released from control samples and from samples treated with chloroform fumigation to kill microbial cells; ion exchange resin procedures</td>
<td>F</td>
</tr>
<tr>
<td>Nutrient retention or loss</td>
<td>mg nutrient L⁻¹ leachate, µg nutrient g⁻¹ soil or plant sample; ¹⁵N recovered (%)</td>
<td>Measured as nutrient concentration in leachates (the higher the concentration in leachates, the lower the retention by the ecosystem) or as ¹⁵N recovered in plants and soil organic matter</td>
<td>P</td>
</tr>
<tr>
<td>Water balance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water content</td>
<td>mg H₂O g dw⁻¹ or % H₂O dw</td>
<td>Harvest followed by measurement of fresh and dry weights</td>
<td>P</td>
</tr>
<tr>
<td>Water use efficiency (C assimilated per unit water transported, integrated over a long period, e.g., a whole season or longer)</td>
<td>µmol CO₂ mmol H₂O⁻¹ ³⁰C⁻¹</td>
<td>Ratio of stable isotopes ¹³C to ¹²C in plant material compared with standard (PeeDee Belemnite)</td>
<td>P</td>
</tr>
<tr>
<td>Evapotranspiration</td>
<td>mg H₂O m⁻² min⁻¹</td>
<td>H₂O exchange measurements</td>
<td>F</td>
</tr>
<tr>
<td>Water potential (the more negative the water potential, the more negative the balance between absorption and transpiration)</td>
<td>MPa</td>
<td>Pressure chamber measurements (pressure needed to expel a drop of sap out of the xylem)</td>
<td>F</td>
</tr>
</tbody>
</table>

* More details can be found in Pearcy et al. (1989).

to only a factor of two to four at the primary production level between different forest types.

Therefore, for integral ecosystem responses, methods which focus on changes of pools (P measurements in Table I) are usually more meaningful and/or convenient. This point is illustrated by Fig. 2 and by some examples of the use of different methods for assessing biomass production at different scales (see Section II.B). An example is provided in Fig. 3. A long-term study of the responses of tundra-dominant plants to different treatments showed that processes that are readily integrated at annual time steps (shoot growth, mortality, and allocation) were more useful than instantaneous physiological measurements in predicting decadal vegetation changes. Strong treatment effects on photosynthetic rate (Fig. 3a) and nutrient uptake were poorly related to longer term changes in production and nutrient concentration (Figs. 3b and 3c). This was probably due to the operation of buffering mechanisms (e.g., allocation, nutrient relations, altered phenology, interspecific interactions, and positive and negative feedback mechanisms) which compensate for immediate physiological responses to the environment.

B. Biomass Production

Net primary production (NPP) is strictly defined as the difference between the energy fixed by autotrophs and their respiration, and it is most commonly equated to increments in biomass per unit of land surface and time. Because the increment in biomass over a given time depends on the rate at which new biomass is produced and also on the initial amount of carbon-assimilating photosynthetic tissue, stands with a large standing biomass often show higher NPP than stands with lower biomass. Therefore, another useful concept is that of relative productivity rate, or the time needed by a vegetation stand to produce its standing biomass. For example, the estimated relative productivity rate for a dry tropical forest can be many years, whereas in an annual grassland it is less than 1 year.

The fate of assimilated carbon—that is, whether it is allocated to increase the pools of aboveground or belowground biomass, root exudates, litter, soil organic matter, grazers, symbionts, or parasites—varies strongly between ecosystems, depending on prevailing climatic conditions, disturbance regimes, and allocation patterns of dominant plant functional types (Fig. 4).

At the regional scale, net primary production can be largely accounted for by climatic factors. For example, precipitation, potential evapotranspiration, and radiation are enough to account for the aboveground net primary production (ANPP) of North American forests, deserts, and grasslands. In regions of the United States with up to 1400 mm of annual rainfall, annual precipita-
tion is enough to account for 90% of the variability in ANPP of grasslands (Fig. 5a). At higher precipitation, ANPP depends more on other factors, and equations based on annual rainfall lose part of their predictive power. At the site level, variability in production seems to be accounted for by annual precipitation and soil water-holding capacity (whc; Fig. 5b). Soil whc can have a positive or negative effect depending on the precipitation value. In dry regions, major losses of soil water occur via bare soil evaporation. However, where sandy soils occur, bare soil evaporation is lower than in loamy soils because water penetrates deeper into the soil. For the same reason, surface runoff is also lower in sandy soils than in loamy soils. In more humid regions, substantial water losses occur via deep percolation, which is reduced in soils with high whc. This is known as the inverse texture hypothesis, proposed by I. Noy-Meir in 1973.

At finer scales of analysis (e.g., paddocks and vegetation patches), more variables are needed to account for ANPP. Species composition and land-use regime become important factors, although drivers at a coarser scale are still in operation and constrain responses (e.g., irrespective of management or species composition, annual precipitation will set an upper boundary to ANPP). For instance, in Argentine montane and pampean natural grasslands, ANPP decreased between 50% and more than 300% when subjected to moderate to heavy grazing. Species composition is crucial at this level; for example, ANPP tends to be higher in legume-dominated pastures than in grass-dominated ones because legume growth is much less limited by soil nitrogen availability due to their capacity for symbiotic nitrogen fixing.

Biomass production from local to global scales can also be estimated by remote sensing. The normalized difference vegetation index, derived from the reflectance in the red and infrared bands measured by the meteorological satellites NOAA/AVHRR (National Oceanic and Atmospheric Administration/Advanced Very High Resolution Radiometer), shows strong correlation with vegetation processes such as photosynthesis and primary productivity and has been widely used to assess primary production (Fig. 6).

C. Trophic Transfer

Trophic transfer can be defined as the amount of biomass and/or energy which is transferred from the primary producers to the herbivore-based food chain rather than directly from plants to detritivores. In this article, only the transfer from plants to herbivores will be analyzed. Two concepts are relevant: consumption,
or the amount of plant biomass consumed by herbivores, and net secondary productivity (NSP), or the amount of biomass/energy at the herbivore level which is available to carnivores. Both depend on the amount of available ANPP, the quality of plant biomass, the kind of metabolism of the herbivores, and the linkages between herbivores and plants.

1. Effects of Nutrient Availability, Plant Production, and Plant Quality on Herbivore Performance

Plant nutrient quality can directly affect animal populations. Plants growing in low-nutrient sites tend to have elevated concentrations of carbon-based secondary compounds that deter consumption by making a higher proportion of plant biomass unavailable. In addition, low foliar concentrations of nutrients can result in either decreased or increased consumption by herbivores. In the second case, herbivores (most commonly insects) consume higher quantities of biomass in order to meet nutritional requirements, with or without reduced fitness as the final result. This effect, sometimes called the “nutrient dilution effect”, is particularly true in the case of generalist herbivores, whose opportunities for co-evolutionary adjustment to the chemistry of a particular plant species are slight.
Secondary chemicals in foliage seem to be more important as regulators of the types of herbivores prevailing in different ecosystems and their dietary habits, than as regulators of consumption at the ecosystem level. There is evidence that herbivore biomass, consumption, and productivity are closely correlated with plant productivity across a wide range of ecosystems, such as deserts, savannas, agricultural grasslands, tropical forests, and salt marshes (Fig. 7). Respiratory costs per unit production at the consumer trophic level are higher for homeotherms (such as ungulates) than for heterotherms (such as insects). In forests, in which most herbivores are heterotherms, most production is allocated to wood, whereas in grasslands, in which homeotherms predominate, much higher production is allocated to green tissue. This explains why net foliage productivity predicts consumption considerably better than net aboveground biomass.

Herbivore biomass and consumption increase as a power of net aboveground primary production, whereas net secondary production increases linearly (Fig. 7). This indicates that highly productive ecosystems sustain a larger level of herbivory per unit of net aboveground primary production than unproductive systems. This larger level of herbivory, however, is accompanied by a lower secondary production per unit of consumption. Because the foliage unconsumed by herbivores will flow into decomposer food webs, the relationship shown in Fig. 7 between consumption by herbivores and net aboveground primary production indicates that the relative importance of the direct flow to detritus decreases as ecosystem productivity increases. However, the relative amount of litter which is actually consumed is small, even in the most productive systems.
2. Effects of Herbivores on Nutrient Cycling and Primary Productivity

Animal activity can affect nutrient cycling directly and indirectly. The amount of ANPP consumed by herbivores can vary from less than 10% in tropical rain forests to more than 50% in meadows (Larcher, 1993). Herbivores can short-circuit the decomposer pathway since urine and feces are much easier to decompose than plant litter. They can also redistribute nutrients and create patches (ungulates, rodents, and ants), and they can promote secondary compound production in plants, which may further deter grazers and may retard decomposition. Mammals, ants, and termites can play important roles in spatial distribution of nutrients at different scales, with impacts on the dynamics of the whole food web.

In many systems primary productivity increases with light grazing, then decreases, and finally decreases more or less sharply as grazing becomes severe. This was at the heart of the herbivore optimization curve hypothesis proposed by S. J. McNaughton in the 1970s. Herbivore optimization models are still under debate, and to determine whether total (aboveground and belowground) biomass actually increases under herbivory is operationally very difficult. However, supporting evidence has been found in a wide range of terrestrial ecosystems, including not only grasslands and wild and domestic ungulates but also forbs and geese, trees and mosses, and crops and birds. It is generally accepted that the highest production rates of plants occur when grazing occurs but it is not too high.

3. Bottom-up and Top-down Controls of Food Chains

Bottom-up or resource control in food chains emphasizes the importance of resource availability for primary producers and the subsequent energy and nutrient flow through a series of trophic levels. Organisms at each trophic level are food limited. The rationale of the opposite, or top-down, view is that organisms at the top of the food chain are food limited but those at lower trophic levels are alternatively predator and food limited. Both mechanisms are recognized to occur in nature, with their importance varying from place to place. S. D. Fretwell and L. Oksanen proposed that the importance of top-down control increases with primary productivity. This idea has recently been empirically demonstrated for invertebrate herbivory and predation on limestone grasslands through a series of experiments summarized in Fig. 8 involving pesticide treatments, transplant of turves, and the use of bioassays (lettuce discs and blowfly maggots in order to evaluate the degree of herbivory and predation, respectively).

D. Nutrient Cycling

1. Nutrient Capture, Retention, and Release by Plants

A key, albeit controversial, concept in relation to nutrient cycling is that of resource use efficiency, or the
relationship between a limiting resource (light, nitrogen, and water) and a biological process (photosynthesis and primary production). Nutrient use efficiency was originally defined in the early 1980s by P. Vitousek as the total NPP (above- plus belowground) per unit nutrient absorbed annually. In practice, it has been usually measured as the ratio of dry mass to nutrient content in litter (Table I), which is a good index of the nutrient economy in a stand as a whole and is based on information reasonably easy to obtain. Recently, a distinction has been proposed by J. Pastor and S. D. Bridgham between nutrient use efficiency (production per unit nutrient uptake) and nutrient response efficiency (production per unit nutrient available).

2. Major Controls over Nutrient Cycling
The main factors underlying variations in nutrient cycling in different ecosystems are climatic factors, soil
FIGURE 8 (a) Estimation of herbivore and carnivore activity, measured as a percentage of lettuce discs and maggots consumed, respectively, and (b) illustration of Fretwell–Oksanen theory of interaction between trophic dynamics and primary productivity in British limestone grasslands with low (I), intermediate (II), and high (III) productivity. At very low productivity, the vegetation did not experience a detectable amount of herbivory. Vegetation of intermediate productivity supported a high level of herbivory and responded strongly to the removal of herbivores. Productivity in this situation is believed to be insufficient to sustain a high intensity of "top-down" control. In the case of highly productive vegetation, carnivory was strong and the intensity of the top-down control by carnivores on plant mass removal by herbivores was maximum (reproduced with permission from Fraser and Grime, 1997).

3. Open vs Tight Nutrient Cycles
A concept strongly linked with nutrient use is the degree of "tightness" or "openness" of nutrient cycling in different ecosystems. This refers to the relative importance of within-system nutrient cycling vs external inputs and outputs. Tropical forests are considered systems with tighter nutrient cycling than temperate forests because within-system nutrient recycling is more important than influx into or effluxes out of the system.

4. The Effects of Rainfall
Concentrations of available nutrients in the soil are relatively (although not absolutely) high in semiarid sites and decrease with increasing precipitation. Soil organic carbon tends to follow the opposite pattern, increasing with increasing rainfall. While total soil nutrient content may also increase, carbon: nutrient ratios in soil increase with higher mean annual precipitation. This suggests that, as rainfall increases, rate of carbon accumulation in soils is higher than that of total nutrient accumulation due to differences in mineralization.

Changes in nutrient cycling "tightness" with climatic factors are well illustrated by an analysis of soil and foliar nutrients in a rainfall gradient in Hawaii (Fig. 9). As rainfall increases with altitude, there is a shift from relatively high nutrient availability to relatively high carbon gain by producers, indicated by a decrease of leaf mass and leaf nitrogen concentration and an increase in lignin concentration with altitude (Fig. 9b). A progressively depleted 15N signature in both soils and vegetation in the wetter sites (Fig. 9a) suggests that N cycling shifts from more open at the drier sites (larger turnover) to tighter (smaller losses) as precipitation increases.

5. The Effects of Soil Type, Land Use, and Vegetation Structure
A comparison among different types of Amazonian rain forest ecosystem, differing in soil properties and topographic positions (Table II), illustrates how patterns of nutrient allocation depend on soil chemical properties and flooding regimes. In turn, soil properties influence nutrient supply, and flooding regimes affect nutrient uptake ability. Mixed and guaço forests, located in higher topographical positions, show relatively high nitrogen contents in both soil and vegetation. In the tall caatinga forest, the proportion of total nitrogen in living biomass is much higher than in the mixed forest...


and guaco forests, and decomposition is slower, associated with waterlogging and nitrogen limitation. Nitrogen circulates in larger amounts in mixed and guaco forest, which show more "open" nitrogen cycling than tall caatinga forest.

Changes in vegetation structure due to land use can lead to different nutrient relations in sites under similar climatic and original soil conditions. A comparison between a tropical dry deciduous forest and savannas derived from the same forest and now maintained by grazing and fire in northern India (Table III) shows that nutrient cycling is faster in the vegetation, litter, and soil of the savanna. The forest shows a higher nutrient use efficiency: Whereas the biomass and nutrient content are much higher, the annual net production and nutrient uptake are similar to those of the savanna. In the savanna there is smaller permanent nutrient storage, and significant nutrient leakage from the system, reflected in lower soil content, indicating more open nutrient cycling.

E. Decomposition: From Nutrient Organic to Inorganic Forms Available to Plants

The process of decomposition, or disintegration of plant and animal residues by the soil detrital food web, is a
key step in nutrient cycling since it makes nutrients already present in the system available for new plant growth. Some plants can incorporate organic forms of nitrogen with or without the intervention of mycorrhizal symbionts. However, these cases seem mostly restricted to very nutrient-poor systems. In most ecosystems, the majority of primary production is not consumed by herbivores but passes directly to detritus. A smaller fraction of primary production is incorporated into herbivores and carnivores which becomes detritus when these organisms die. The ultimate end product of organic matter breakdown is inorganic forms of carbon (CO$_2$) and nutrients (nitrates, ammonium, and phosphates). Decomposition depends on soil environment (water potential, temperature, and aeration), microsite characteristics (slope, texture, drainage, aspect, and cover type), substrate quality, and composition of decomposer community (size and specific composition of animal and microbe assemblages and synergistic or antagonistic relationships among them).

The control of decomposition by macroclimatic parameters is very strong. Whereas mean NPP increases by a factor of approximately 20 from tundra to tropical rain forests, mean residence times of dead organic mat-

### Table II
Patterns in Nutrient Flows and Stocks in Three Different Amazonian Forests under Similar Macroclimate

<table>
<thead>
<tr>
<th>Soil Geomorphicological features</th>
<th>Mixed</th>
<th>Guaco</th>
<th>Tall caatinga</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total biomass (t ha$^{-1}$)</td>
<td>310</td>
<td>465</td>
<td>400</td>
</tr>
<tr>
<td>Leaf fall (t ha$^{-1}$ year$^{-1}$)</td>
<td>6.3</td>
<td>7</td>
<td>5.5</td>
</tr>
<tr>
<td>Disappearance constant (k; year$^{-1}$)</td>
<td>$3.68\times10^{-4}$</td>
<td>---</td>
<td>$0.93\times10^{-4}$</td>
</tr>
<tr>
<td>Biomass nitrogen ratio (g; g)</td>
<td>$110$</td>
<td>$120$</td>
<td>$260$</td>
</tr>
<tr>
<td>Litter N content (kg ha$^{-1}$)</td>
<td>137</td>
<td>89.5</td>
<td>132</td>
</tr>
<tr>
<td>Soil N content (at 10 cm depth; kg ha$^{-1}$)</td>
<td>1479</td>
<td>2405</td>
<td>716</td>
</tr>
</tbody>
</table>

*Data from Medina and Cuevas (1989).*

### Table III
Nutrient Cycling in Two Tropical Dry Deciduous Vegetation Types under the Same Climatic and Soil Conditions, With Different Land-Use History in Northern India

<table>
<thead>
<tr>
<th>Dry deciduous forest</th>
<th>Savanna</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production (t ha$^{-1}$ year$^{-1}$)</td>
<td>13</td>
</tr>
<tr>
<td>Total biomass (t ha$^{-1}$)</td>
<td>93</td>
</tr>
<tr>
<td>Nutrient uptake per unit energy captured (mg 1000 kcal$^{-1}$)</td>
<td>3000</td>
</tr>
<tr>
<td>Vegetation nutrient content (kg ha$^{-1}$)</td>
<td>680</td>
</tr>
<tr>
<td>Litter nutrient content (kg ha$^{-1}$)</td>
<td>37</td>
</tr>
<tr>
<td>Soil nutrient content (0-10 cm depth; kg ha$^{-1}$)</td>
<td>2008</td>
</tr>
<tr>
<td>Total litter fall (kg ha$^{-1}$ year$^{-1}$)</td>
<td>80</td>
</tr>
<tr>
<td>Total nutrient release (leaf + root decomposition; kg ha$^{-1}$ year$^{-1}$)</td>
<td>104</td>
</tr>
<tr>
<td>Nutrient retention in vegetation (kg ha$^{-1}$ year$^{-1}$)</td>
<td>62</td>
</tr>
</tbody>
</table>

*Data from Singh (1989).*
ter decrease by a factor of approximately 200 primarily because soil temperatures limit decomposition more than air temperatures limit production.

At a more local scale, litter quality plays an extremely important role in determining decomposition rates. The relative velocity of decomposition of litter from different plant species tends to remain constant under different environmental conditions and to be strongly correlated with the palatability of living leaves. This suggests that the same compounds which determine palatability aboveground control litter decomposition by soil microbiota. The most widely used indexes to describe litter quality are the carbon : nitrogen and lignin : nitrogen ratios, with higher ratios being associated with lower decomposition rates (Table IV).

### Table IV

<table>
<thead>
<tr>
<th>Reference Plant material</th>
<th>( r )</th>
<th>( p )</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 English shrubs and trees</td>
<td>(-0.780)</td>
<td>(&lt;0.01)</td>
<td>12</td>
</tr>
<tr>
<td>2 Mediterranean shrubs and trees</td>
<td>(-0.720)</td>
<td>(&lt;0.01)</td>
<td>8</td>
</tr>
<tr>
<td>3 South American temperate and subtropical graminoids, forbs, succulents, shrubs, and trees</td>
<td>(-0.520)</td>
<td>(&lt;0.001)</td>
<td>52</td>
</tr>
</tbody>
</table>

* Reference: 1, Cornelissen (1996); 2, Gallardo and Merino (1993); 3, Pérez-Harguindeguy et al. (1999). \( r \), Spearman's correlation coefficient; \( p \), significance level; \( n \), number of species involved.

2. Litter Decomposition and Composition of Detrital Food Webs

The composition of the decomposer communities, which include macrofauna (earthworms and arthropods that carry out initial comminution, mixing, and dispersion of litter and microbial propagules), mesofauna (springtails, mites, and enchytraeid worms), and microbiota (fungi, bacteria, protozoa, nematodes, and actinomycetes), also determines the local decomposition rate. Grazing on the bacteria, actinomycetes, and fungi is important (often indispensable) for net mineralization to occur. Several experiments manipulating densities of bacteria, fungi, and their predators (e.g., protozoa and nematodes) and macrofauna (e.g., millipedes) have shown that the composition of the detrital food web is a strong factor determining the balance between net mineralization and immobilization. Belowground grazing on bacteria and fungi and aboveground grazing on plants have both been shown to increase microbial activity and mineralization.

### F. Water Dynamics

Water flow through the plant compartment of the soil–plant–atmosphere continuum is regulated at the plant–air interface, where the transition of liquid water to vapor occurs and the steepest gradient of water potential exists. At this interface, the shoots are exposed to the low water potential of the air and a flow of water through the plant is set in motion. In this way, and depending on xylem structure and cavitation, the plant component of the system bridges the steep water potential gradient between soil and air.

Vapor loss from land and surfaces is governed by leaf area, by stomatal and aerodynamic conductances of plant canopies, and by the contribution of evaporation from soils, and it affects numerous terrestrial processes ranging from the biogeochemical cycling of elements to the development of regional climate. The
ECOSYSTEM FUNCTION MEASUREMENT, TERRESTRIAL COMMUNITIES

III. ECOSYSTEM FUNCTIONING UNDER DISTURBANCE: RESISTANCE AND RESILIENCE

Ecosystem stability is often divided into two components: resistance and resilience. Ecosystem resistance is the ability of a system to avoid change—the capacity to stay in the same state in the face of perturbation (e.g., fire, unusual frost or drought, plowing, eutrophication, and pollutant input). Ecosystem resilience is the rate at which a system returns to its former state after being displaced from it by a perturbation. Not all aspects of ecosystem functioning are equally resistant or resilient. The most commonly measured ones (which are not necessarily the most important or sensitive ones) are species composition and biomass. Usually, a standard or ad hoc index is constructed that relates vegetation structure or composition before the perturbation to that after the perturbation. A higher dissimilarity means the ecosystem has a lower resistance. The longer amount of time needed to achieve maximum similarity between the predisturbance and postdisturbance situation, the lower the resilience. Some authors have used indicators other than community composition, such as nutrient loss rate, to estimate stability; the higher the relative nutrient loss rate following disturbance, the lower the resilience. Evidence has accumulated for the idea that both components of ecosystem stability are primarily determined by key traits of the dominant plant species. Highly productive communities, dominated by fast-growing plants, tend to have high resilience and low resistance, with the opposite being true for communities dominated by slow-growing plants. Productivity and seed production (especially persistent seeds) favor resilience, whereas preferential allocation to storage and...
defense favors resistance. These ideas were formalized in the early 1980s on the basis of studies of old-field successional communities (Fig. 11). Recently, new experimental support has been provided for the idea that vegetation resistance and resilience in the face of extreme events are a function of the nutrient stress tolerance of the component species. Figure 12 clearly illustrates that resistance to extreme events increases, and resilience decreases, with increasing nutrient stress tolerance in herbaceous communities.

FIGURE 11 Relationship between life history strategies of dominant plants and community resistance and resilience (reproduced with permission from Kluwer Academic Publishers, Lepp et al., Fig. 5 © 1982 with kind permission from Kluwer Academic Publishers.).
ECOSYSTEM FUNCTION MEASUREMENT, TERRESTRIAL COMMUNITIES

The size and the turnover rate of the detrital compartment also have implications for ecosystem stability. Ecosystems with high carbon:nutrient ratios in the soil are effective at immobilizing large amounts of nutrients.

This immobilization can substantially control the nutrient losses as long as a sufficient amount of plant residue is left on the site to sustain the microorganisms that immobilize the nutrients in their biomass. Therefore, soil organic mass is a major determinant of ecosystem resistance. On the other hand, systems with tight nutrient cycles have increased return time to equilibrium (lower resilience). When substantial amounts of biomass are removed from these systems, causing high nutrient loss, recovery may be very slow because there is little throughflow of nutrients coming from outside the system compared with the nutrient lost. As a general rule, higher nutrient mean residence time in an ecosystem increases its resistance and decreases its resilience.

IV. ECOSYSTEM FUNCTIONING AND FUNCTIONAL DIVERSITY

A. Taxonomic and Functional Diversity

The most common way of assessing biodiversity is the measurement of the number and relative abundance of species per area. Taxonomic diversity per se is an important parameter for conservation, but it provides little information on the magnitude and rate of ecosystem processes. On the other hand, as discussed in Section II,A, measurements of pools and flow rates at the whole ecosystem level provide little information about a system’s conservation value or about its likely response in the face of a perturbation. An intermediate approach is the one based on plant functional types. Functional types are sets of plants sharing similar responses to environmental conditions and similar effects on major ecosystem processes. This approach is as old as ecology itself, but it has gained renewed interest in the past few years because it bridges the gap between individual species and whole-ecosystem functioning, explains particular values of pools and flow rates, and improves predictions of how ecosystems can be modified by the introduction of new abiotic (e.g., fertilization) or biotic factors (e.g., invasion by alien species).

The practical value of distinguishing discrete “types” is obvious. However, it is important to bear in mind that many real plants represent transitions between, rather than typical examples of, different functional types.

B. Ecosystem Functioning and Biodiversity

Although the role of species richness per se remains controversial, and the evidence is sparse, the impor-
FIGURE 13 Proposed ecosystem effects of species in different positions along an idealized dominance–diversity curve. Transients are those species unable to regenerate and persist in situ; their sources are soil seed bank and seed rain from the surrounding landscape. Immediate effect, control over major processes of resource dynamics; filter effect, positive or negative influence on the regeneration of dominants following major perturbations; founder effect, reservoir of potential colonizing dominant and subordinate species in the event of ecosystem reassembly (adapted from Grime, 1998 with permission from Blackwell Science Ltd.).

C. Links between Plant Functional Traits and Ecosystem Functioning

The existence of suites of vegetative traits consistently associated across taxa and ecosystems (e.g., plants that "go for it" and plants that "sit and wait") has been repeatedly documented and is increasingly accepted. Central to the plant functional type approach is the idea that life history, allocation, and phenological, physiological, architectural, and reproductive traits of dominant plant species appear to be associated in a limited number of combinations and strongly determine vegetation structure and ecosystem functioning (Fig. 14). By being primary components of ecosystem-level fluxes of matter and energy, dominant plant functional types are indirect determinants of the biogeochemical cycles of carbon, water, and nutrients. Individual examples of strong links between plant traits and ecosystem processes have accumulated in the past few decades (Table V).

Major ecosystem processes appear more strongly and directly linked to recurrent suites of vegetative traits than to regeneration traits (Fig. 14). Plant regeneration traits appear to be under the control of selective forces different from those that operate on resource dynamics during the established phase. In situ plant regeneration by seed is an important aspect of resilience and also strongly influences migration across the landscape in the face of climatic changes. Seed production, germination, dispersion, and establishment are therefore key aspects in determining ecosystem functioning in the face of major changes of climate and land use, although their role has proved much more difficult to document than that of adults-phase traits.

D. Positive-Feedback Switches

The effect of the biotic component on ecosystem dynamics often takes the form of "switches" or positive-feedback processes in which members of a community
modify their environment, making it more suitable for themselves. This term was coined in 1992 by J. B. Wilson and A. D. Q. Agnew, who provided an extensive list of examples, some of which are shown in Fig. 15.

V. MEASUREMENT, ANALYSIS, AND PREDICTION OF ECOSYSTEM FUNCTIONING: MAJOR PROTOCOLS AND OBSTACLES

A. Monitoring, Field Manipulations, and Synthesized Ecosystems: A Gradient of Questions and Methods

The understanding of how ecosystems function can be achieved by a whole gradient of protocols, from simply documenting what is happening in real systems to synthesizing model ecosystems from scratch. These extreme approaches represent a trade-off between realism and precision and between the documentation of general patterns and that of the specific mechanisms underlying them. The monitoring of real ecosystems often needs to be performed by “soft” approaches. It is important at the initial heuristic stage and when testing whether results found in simpler experiments can be reasonably applied to the real world (synthesis stage). Experiments on synthesized ecosystems, on the other hand, are ideal for the stage of testing specific hypotheses on mechanisms which may account for the patterns observed (analytic stage). Field manipulations lay in the middle of this gradient (Fig. 16).

1. Monitoring Real Ecosystems: Approaches Based on Hard and Soft Traits

Processes in natural ecosystems can be recorded by applying the usual methods (hard approaches) of measuring pools and flows described previously (see Sec-
TABLE V Examples of Individual Plant Traits which Strongly Influence Processes of the Community/Ecosystems in which They Are Dominant

<table>
<thead>
<tr>
<th>Individual trait</th>
<th>Ecosystem/community process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative growth rate</td>
<td>Productivity, resilience, trophic transfer</td>
</tr>
<tr>
<td>Leaf turnover rate</td>
<td>Nutrient cycling, detritivore diversity and biomass</td>
</tr>
<tr>
<td>Nitrogen content</td>
<td>Nutrient cycling, trophic transfer, detritivore diversity and biomass</td>
</tr>
<tr>
<td>Biomass</td>
<td>Trophic transfer, carbon sequestration, flammability</td>
</tr>
<tr>
<td>Life span</td>
<td>Resistance</td>
</tr>
<tr>
<td>Canopy structure</td>
<td>Aerodynamic conductance, interception, water relations, runoff, roughness/albedo, temperature buffering, soil stability, consumer biodiversity</td>
</tr>
<tr>
<td>Secondary growth</td>
<td>Carbon sequestration, trophic transfer, nutrient cycling</td>
</tr>
<tr>
<td>Ramification</td>
<td>Structural complexity, consumer biodiversity, resistance (particularly drought), temperature buffering</td>
</tr>
<tr>
<td>Root architecture</td>
<td>Water uptake, soil stability</td>
</tr>
<tr>
<td>Reserve organs</td>
<td>Resilience</td>
</tr>
<tr>
<td>Pollination mode</td>
<td>Expansion over landscape</td>
</tr>
<tr>
<td>Persistent seed bank</td>
<td>Resilience</td>
</tr>
<tr>
<td>Seed number</td>
<td>Expansion over landscape</td>
</tr>
<tr>
<td>Dispersal mode</td>
<td>Expansion over landscape</td>
</tr>
<tr>
<td>The presence of root symbionts</td>
<td>Diversity, nutrient cycling, carbon sequestration, rate of succession</td>
</tr>
</tbody>
</table>

* Modified from Díaz et al. (1999).

The presence of root symbionts (A.1 and A.2). They are sometimes termed hard approaches because they are quantitative and usually represent a direct measurement of a process. Alternatively, ecosystem functioning can be inferred from the presence and abundance of plant traits which are easily measured but at the same time have clear implications at the ecosystem level. These are called soft traits because they are usually indirect indicators of ecosystem processes. Some examples of soft traits are given in Table V. The soft approaches have the advantage that they need a minimum investment in financial and technological resources and can be utilized to characterize extensive areas or high numbers of systems in a short period of time. The result is usually a comparative estimation of ecosystem functioning (Fig. 17), which then needs to be calibrated against a hard approach. This is achieved by testing how well the soft traits correlate with hard traits, which have more direct and well-documented relationships with ecosystem processes (e.g., the soft trait leaf tensile strength correlates well with the hard trait decomposition rate, and the soft traits seed mass and shape correlate well with the hard trait seed persistence in the soil; Fig. 18).

2. Experimental Manipulations Involving “Natural” vs Synthesized Ecosystems

The major disadvantage of field experiments is that the degree of control of independent and external variables is low compared to manipulating them. The amount of “noise” (unwanted variance) is usually high, and enough replicates in order to reduce it are usually unavailable or involve prohibitive costs. On the other hand, synthesized ecosystems, such as microcosms, never come close to the realism of experiments involving manipulations of real ecosystems, they are always a badly simplified version of nature, and they have severe size limitations (e.g., for ecosystems based on woody vegetation). Considering the realism-precision trade-off illustrated in Fig. 16, and taking into account
ECOSYSTEM FUNCTION MEASUREMENT, TERRESTRIAL COMMUNITIES

FIGURE 15 Switches of (a) fog precipitation on a hillside in the montane tropics, (b) termite mounds in tropical savanna, and (c) water (sediment entrainment)/salt with salt pans on a salt marsh (reproduced with permission from Wilson and Agnew, 1992).

FIGURE 16 General characteristics of different approaches to the study of ecosystems.
that some processes scale up poorly from single systems to complex ecosystems (see Section II,A,2); field experiments are indispensable. However, microcosms work has produced highly relevant insights into ecosystem functioning, including relationships between trophic levels, community roles of symbionts, feedbacks involved in soil fertility and climate manipulation, and some of the most controversial and inspiring experiments on the role of biodiversity in determining ecosystem functioning. Field manipulations, such as the application of treatments to real vegetation plots, or the selective removal of certain components of the community have some limitations. These include the difficulty of separating the effects of removal of one component from the effect of the disturbance caused during the removal process and the difficulty of minimizing spatial heterogeneity that usually masks treatment effects. However, they are an excellent intermediate step between laboratory experiments and the long-term monitoring of undisturbed systems. Additionally, they have produced important evidence of the difficulties and potentialities of scaling up from individual physiology to
ecosystem functioning (see Fig. 3), and they are likely to continue to provide fruitful insight in ecosystem ecology.

Acknowledgments
I am grateful to Diana Abal-Solís for drawing the figures and to two reviewers for fruitful comments.

See Also the Following Articles
Ecosystem, Concept of; Ecosystem Function Measurement, Aquatic and Marine Communities; Ecosystem Function, Principles of; Measurment and Analysis of Biodiversity; Terrestrial Ecosystems

Bibliography


I. DEVELOPMENT OF THE ECOSYSTEM CONCEPT

The concept of the ecosystem as a functioning unit in the natural world is a relatively recent one. The term ecosystem was coined by the British ecologist Tansley in 1935 and has since become a common word in science and with the public. An ecosystem encompasses all the organisms of a given area and their relationships with one another and the physical or abiotic environment. The ecosystem contains the linkages and dynamic interactions between life and the environment, many of which are essential to society. A focus on the ecosystem as the unit of study represents a shift from studying the ecology and behavior of individual organisms and species (natural history) to the study of processes and how they influence or are influenced by organisms and their interactions with the environment.

Dividing the complexity of nature into convenient units of study is required for scientific investigation but can present problems. Ecological systems can be...
ECOSYSTEM FUNCTION, PRINCIPLES OF

organized in a hierarchy of increasing levels of organization and complexity: individual, population, species, community, ecosystem, landscape, and biome. The size (scale) of an ecosystem is defined by the purposes of the study. Ecosystems may have distinct boundaries as in the case of a lake or a watershed. More often, the boundaries of one ecosystem (a forest) may grade gradually into another (a meadow) across an intermediate area called an ecotone. The ecotone is often a zone of higher diversity because it may be a suitable habitat for species from each of the adjoining ecosystems. At one extreme of scale, the earth is sometimes treated as an ecosystem. At the other extreme, the complex symbiotic community of organisms inhabiting the gut of a termite has all the functional properties of an ecosystem. The definition and delineation of an ecosystem has practical importance because ecosystems are increasingly seen as a functional unit for resource and conservation management purposes. It has become evident that the management of lands for sustained levels of ecosystem services and natural resources requires an understanding of how ecosystems function, how they respond to disturbance, and how the role of biodiversity is regulating their function and stability.

II. ECOSYSTEM FUNCTIONING AND ECOSYSTEM SERVICES

Society depends on the functioning of ecosystems for many essential ecosystem services on which we place economic and aesthetic value (Daily, 1997). Ecosystem functioning results from the collective activities of organisms and their life processes (production, consumption, and excretion) and the effects of these activities on the condition of the environment. These functions (services when they provide utility to humans) include production of food, fuel, and fiber, the cycling and purification of water, and the maintenance of organisms that have a role in ecosystem functioning or that provide products for human use (Table 1). Humans are rapidly changing the earth’s ecosystems and their services by altering land use or by harvesting biological resources (forest cutting and fisheries) (Vitousek et al., 1997). Approximately 40% of the earth’s primary production is diverted to human use. One consequence of these economic activities is an abrupt increase in the rate of change in biological diversity leading to species extinction, replacement of high-biodiversity ecosystems with less diverse managed systems, and invasions of natural ecosystems by exotic species. This pattern of ecosystem change has raised serious concern that the functioning and stability of our global ecosystem are threatened by the loss of biodiversity.

A. What Do Ecosystem Scientists Study?

Ecosystems share certain characteristics and functions that allow scientists to study ecosystem types (e.g., deciduous forest, temperate grassland, arctic tundra, coral reef, and deep-ocean hydrothermal vents) that vary greatly in structure, biodiversity, and spatial extent. For example, all ecosystems require inputs of energy (usually solar) and a supply of the mineral elements (nutrients) essential for life. These inputs support many ecological processes operating at multiple scales. For example, sunlight, carbon dioxide, and water are inputs for the process of photosynthesis, which can be measured and studied at the scale of individual cells, a leaf, the plant canopy, or an entire ecosystem. Photosynthesis acting with other processes such as mineral uptake by roots combine to create an ecosystem function—primary productivity.

Scientists can discover basic principles about the behavior of ecosystems by studying the functions that very different ecosystems, such as the polar desert of Antarctica and the rangelands of the southwestern United States, share in common (Virginia and Wall, 1999). The movement of energy and materials within and between ecosystems and the role of organisms in mediating these processes are the parameters used by
TABLE II
Examples of Ecosystem Services That Would Be Affected by a Decline in Ecosystem Function

<table>
<thead>
<tr>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pest control</td>
</tr>
<tr>
<td>Insect pollination</td>
</tr>
<tr>
<td>Fisheries</td>
</tr>
<tr>
<td>Climate regulation</td>
</tr>
<tr>
<td>Soil retention</td>
</tr>
<tr>
<td>Flood control</td>
</tr>
<tr>
<td>Soil formation and maintenance of soil fertility</td>
</tr>
<tr>
<td>Cycling of matter</td>
</tr>
<tr>
<td>Composition of the atmosphere</td>
</tr>
<tr>
<td>Maintenance of genetic diversity</td>
</tr>
</tbody>
</table>

* Based on Daily (1997).

ecosystem scientists to compare the functioning of ecosystems and their responses to disturbance. Some of the important processes and functions central to the integrity and sustained activity of an ecosystem are summarized in Table II. Ecosystem scientists study the rate at which ecosystems remove carbon from the atmosphere by photosynthesis, store it in the soil as organic matter, and then return the stored carbon to the atmosphere during decomposition. They study how nitrogen is cycled through ecosystems to sustain continued plant productivity. Our knowledge of how carbon and nitrogen move in the ecosystem helps us to understand when an ecosystem has been seriously altered by humans, for example, by adding nitrogen in the form of air pollution (acid rain) and fertilizers.

Many basic principles provide insight into the functioning of ecosystems and their response to human use and disturbance. Here, we will consider some of the essential functions of ecosystems and examine the principles that govern their operation, with an emphasis on the role of organisms (biodiversity) in determining ecosystem functioning.

III. IMPORTANT ECOSYSTEM FUNCTIONS

A. Ecosystem Productivity

A central process of most ecosystems is photosynthesis, the capture of solar radiation and its conversion to stored chemical forms (biomass). Plants require sunlight, water, and essential nutrients for the processes of photosynthesis. Photosynthesis is coupled with other plant processes that result in plant growth, i.e., the accumulation of biomass. Primary productivity, the change in plant biomass per unit area and time, is an important index of ecosystem function. Primary productivity (often referred to as ecosystem productivity) has been related to plant species diversity as well as the diversity of organisms (soil biota) that influence the availability of limiting resources. Humans depend on ecosystem productivity as the basis of our agriculture and forestry and fisheries. Thus, factors that alter ecosystem productivity (e.g., climate change and biodiversity loss) affect us directly.

Ecosystems with high rates of primary productivity have favorable amounts of the resources required for plant growth and optimal climate. These systems also tend to have higher diversity (Table III). The highest rates of terrestrial ecosystem productivity are seen in the tropics, where temperature and moisture are favorable for plant growth throughout the year. In contrast, water-limited hot and cold deserts have much lower productivity, averaging less than 10% of that of tropical systems.

1. Limits to Ecosystem Productivity

A basic principle invoked to explain variation among ecosystems in their productivity is Liebig's Law of Minimum. Justus Liebig formulated this concept during pioneering studies of the mineral nutrition of plants in the early 1800s. He found that addition of a single "limiting element" to a soil would increase plant growth. Once this element was in sufficient supply, another mineral element would have to be supplied in increased amounts to stimulate additional increases in plant growth. From these observations, he proposed that a limiting factor was responsible for limiting the growth or reproduction of an organism or population. This

TABLE III
Typical Values for the Net Primary Productivity of Major Ecosystems

<table>
<thead>
<tr>
<th>Ecosystem type</th>
<th>Net primary productivity (g C/m²/year)</th>
<th>Relative species diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical rain forest</td>
<td>900</td>
<td>Highest</td>
</tr>
<tr>
<td>Temperate forest</td>
<td>540</td>
<td>Intermediate</td>
</tr>
<tr>
<td>Grassland</td>
<td>315</td>
<td>Intermediate</td>
</tr>
<tr>
<td>Deserts</td>
<td>3.2</td>
<td>Low</td>
</tr>
<tr>
<td>Extreme desert</td>
<td>1.3</td>
<td>Lowest</td>
</tr>
</tbody>
</table>

* Ecosystem productivity and biodiversity are often positively related.
factor might be a chemical factor (a growth-stimulating nutrient such as nitrogen), a physical factor such as moisture, or a biological factor such as the presence of a competing species. Thus, any change in a limiting factor is expected to have large effects on ecosystem functioning.

There are many examples in which a change in a limiting factor alters ecosystem function. The large increase in the amount of nitrogen cycling in the environment from fertilizers and fossil fuel should have significant effects on rates of ecosystem functions since nitrogen frequently is the primary limiting element for plant growth in terrestrial ecosystems. Humans have doubled the rate of nitrogen inputs to ecosystems with increases in carbon storage and declines in biodiversity (Vitousek et al., 1997). In fact, the forests of the northeastern United States may have reached "saturation" in their ability to absorb and retain anthropogenic inputs of nitrogen.

Are plant species diversity and primary production related? Ecologists are accumulating evidence from experiments in controlled growth facilities and in the field that ecosystem primary productivity increases with increasing plant species diversity. The theoretical basis for the expectation that productivity and diversity should be related derives from an understanding of how limiting resources (water and nutrients) are distributed in ecosystems and an appreciation for the diversity of physiological or "functional" traits that organisms have evolved to capture and utilize these resources for growth. Differences between plant species in rooting depth, phenology (seasonality of growth), photosynthetic rates, and other physiological traits allow multispecies communities to more fully utilize the available resources.

The ability of diverse plant communities to obtain higher productivity than low-diversity systems is demonstrated in traditional (low-input) agriculture in which polycultures (multiple-species plantings) often have higher yields than single-species plantings (mono-cultures) (Gliessman, 1998). For example, corn (Zea mays) yields at comparable densities are higher when corn is grown in the presence of nitrogen-fixing beans (Vicia spp.). The bean crop forms a symbiotic association with bacteria that "fix" atmospheric nitrogen (N₂) to other inorganic forms (ammonia and NH₃) usable by plants. The nitrogen fixed by the bean crop improves the overall supply of this limiting element in the soil and increases the growth of the interplanted corn. The functioning provided by the diverse corn–bean–nitrogen-fixing bacteria association is often replaced in intensive agriculture by applying inorganic nitrogen fertilizers. With external inputs (fertilizers) the corn monoculture can produce higher yields than can the polyculture. Substituting an industrial source of nitrogen for a biological source has environmental costs resulting from the production and combustion of fossil fuels used to produce fertilizers. In addition, overapplication of fertilizers is a major source of water pollution in surface and groundwaters.

There are similar examples of diversity influencing productivity in natural ecosystems. In a California grassland ecosystem, Hooper and Vitousek (1997) manipulated the number of plant functional groups in a community (early vs late-season forbs, perennial grasses, and nitrogen-fixing plants) in combinations of one to four groups in a given plot. They found that the number of plant functional groups was not the main factor that determined productivity. Rather, certain functional characteristics of individual species within functional groups contributed more to ecosystem productivity than overall diversity of the plot. This study points to the complexity of trying to simply relate species diversity to function. As a general principle, ecologists recognize that some species play particularly important roles in regulating important ecosystem functions such as productivity and nutrient cycling.

B. Keystone Species

Certain species, termed keystone species, have a disproportionate influence (relative to their biomass) on ecosystem functioning. The loss of a keystone species will produce a cascade of effects on the diversity and function of the remainder of the ecosystem (Bond, 1993). Consequently, since keystone species can control ecosystem diversity and associated ecosystem functions, they and the habitats they live in often receive high priority in conservation management plans. There are many well-documented studies of keystone species and how they interact with ecosystem functioning, e.g., the North Pacific sea otter preys on sea urchins, which consume kelp. In the absence of the keystone predator, sea urchin populations increase and create areas devoid of kelp and, consequently, the myriad of fish and other species that depend on the kelp forest (Fig. 1). This is an example of a food web, the representation of trophic (feeding) relationships between species in an ecosystem.

There are many examples of keystone species in terrestrial ecosystems. A large change in African elephant numbers has dramatic effects on the diversity
and structure of the vegetation types (savanna woodlands and forests) they consume, altering ecosystem productivity, soil nutrient cycles, and plant community diversity. The much smaller tsetse fly shares the elephant’s habitat and also has the attributes of a keystone species. The tsetse fly is the vector for the human disease sleeping sickness (African trypanosomiasis). This biting fly also influences the behavior of large herbivores that tend to avoid heavily infested areas. Consequently, herbivore-related impacts on plant communities and associated ecosystem functions are altered in tsetse-occupied ecosystems. This small insect may control the biodiversity of large tracts of Africa through another mechanism. Diverse native ecosystems have been “protected” from agricultural development and species loss because humans avoid regions where the tsetse and therefore sleeping sickness are endemic.

C. Nutrient Cycling

The sustained functioning of any ecosystem requires a minimum number of species to develop the intricate relationships between producers, consumers, and decomposers that regulate the flow of energy and nutrients. The productivity of all ecosystems is dependent on the cycling of essential elements. The movement and biological transformations of organic matter and nutrients are mediated by biota, especially those found in soil and sediments (Wall and Virginia, 1999). Therefore, changes in the biodiversity of ecosystems can alter biogeochemical processes.

1. Succession

Scientists study the process of ecological succession (ecosystem change with time, often in response to disturbance) in part to untangle relationships between biodiversity and function. Although not all ecosystems follow a predictable pathway as they develop in time, examples of succession highlight the linkage between organisms and diversity and ecosystem function. They include the recovery of a forest after harvest or following damage by a hurricane, the reestablishment of grassland following fire, and the old-field succession of natural vegetation reclaiming abandoned agricultural land. During succession, ecosystems change in generally predictable ways as they accumulate species, increase in biomass, and gain structural complexity. Odum (1969) proposed a model of ecological succession (development) that relates ecosystem diversity, structure, and functioning as ecosystems redevelop and “mature” following disturbance (Table IV). Odum’s model related the stability (constancy) of function and the conservation of nutrients to increasing diversity—themes that are at the center of biodiversity and ecosystem research today.

![Figure 1: The influence of a keystone species on the biodiversity of an ecosystem](image)

**TABLE IV**

<table>
<thead>
<tr>
<th>Ecosystem trait</th>
<th>Developing</th>
<th>Mature</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Energetics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net primary production</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Food chains</td>
<td>Linear</td>
<td>Web-like</td>
</tr>
<tr>
<td><strong>Communities</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species diversity</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td><strong>Nutrient cycling</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mineral cycles</td>
<td>Open</td>
<td>Closed</td>
</tr>
<tr>
<td><strong>System dynamics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stability</td>
<td>Poor</td>
<td>Good</td>
</tr>
</tbody>
</table>

*Based on Odum (1969)*
The relationships represented in Odum's (1969) model between ecosystem function and diversity are elucidated in the Hubbard Brook watershed experiment (Likens and Bormann, 1995). One of the first long-term ecosystem studies, the Hubbard Brook project began in 1963 in the White Mountains of New Hampshire. The study was designed to understand the process of forest recovery following harvest with a focus on ecosystem functions related to production, nutrient cycling, and nutrient loss. Measurements of the mature intact deciduous forest showed that less than 0.1% of the nitrogen contained in living forest biomass and dead organic matter in the soil and litter left the site in stream flow.

A nutrient cycle in which outputs are low and internal recycling of nutrients is high (the loop from soil to vegetation and back to soil) is called a closed nutrient cycle. After the unperturbed patterns of growth and nutrient cycling were known, an entire Hubbard Brook watershed was clear-cut. What followed was a dramatic change in ecosystem functioning. Stream flow increased by approximately 40% because water use by plants had been nearly eliminated by the forest harvest. The previously ‘closed’ nutrient cycle of this forest became ‘open.’ After clear-cut the concentrations of nitrogen (nitrate) in the stream water draining the watershed increased approximately 60-fold. Concentrations of elements that are important to the biology of the ecosystem leaked into the streams and were exported from the ecosystem. Elements not essential to plant growth or required in very small amounts (e.g., sodium) were not lost to the same degree, indicating their cycling was not regulated by biotic activity of the forest. Odum (1969) predicted that nutrient losses would decline with increasing plant biomass and function. After the Hubbard Brook forest was allowed to regrow (undergo succession), nutrients resumed being absorbed by plants and nutrient losses to streams declined to near baseline levels. The Hubbard Brook ecosystem experiment informed forest management practices by providing a better understanding of how forest removal and regrowth affect the retention of soil nutrients and therefore the long-term productivity and diversity of the ecosystem.

**D. Ecosystem Stability**

Ecosystems are dynamic. They experience change in species composition and function in response to variations in climate and an array of disturbances. Fire, flood, drought, frost, and biological events such as the outbreak of pathogens and pests can ‘stress’ ecosystems and alter their condition. Ecosystems vary widely in their responses to disturbance. The ability of an ecosystem to withstand stress without a loss of function (resistance) or to recover rapidly from disturbance (resilience) is an important ecosystem trait. Some ecosystems, such as tropical forests, appear very stable (high resistance and resilience) and their functioning is little affected by variations in factors external to the system (e.g., weather). Ecosystems with high resilience are buffered against perturbation. Many ecosystems, however, show large decreases in productivity and biodiversity when disturbed. These ecosystems are ‘fragile’ and have low resistance.

The relationship between ecosystem stability and diversity has been the subject of many field studies and theoretical tests using mathematical modeling. Ecologists have hypothesized that ecosystems with high biodiversity are more resistant (will experience less change) in response to a given level of disturbance and will also exhibit resilience—a high rate of recovery to predisturbance functioning (Folke et al., 1996).

Does diversity influence the stability of ecosystem functioning? There is experimental evidence that it can do so (Chapin et al., 1997). Several mechanisms have been proposed and tested to varying degrees to examine this relationship (Chapin et al., 1997). Higher species diversity means that the trophic structure (feeding relationships among species) of the ecosystem is more complex, providing alternate pathways for energy flow within and between trophic levels (producers, consumers, and decomposers). Alternative pathways for energy transfers within the ecosystem could increase resistance to disturbance (species loss). Naeem and Li (1997) tested the hypothesis that redundancy (multiple species with similar functions in a food web) would stabilize ecosystem functioning by creating experimental microcosms with a varying number of species in each functional group. The simple systems contained producers (algae), decomposers (bacteria), and a primary and secondary consumer trophic level (protozoa)—the trophic structure of a typical aquatic ecosystem. Nutrient levels, light, and the number of species per trophic level were manipulated, and the biomass and density of the producers and decomposers were measured as an indicator of ecosystem functioning. As the number of species in a trophic level increased, the biomass and density of replicate communities were more consistent. Thus, the communities with more species were more predictable in function (biomass production) and had higher reliability, i.e., the probability that an ecosystem will pro-
provide a given level of performance over a specified period of time. Higher species diversity may ensure functioning by reducing the risk of invasion by species that have the capacity to alter the structure or function of the ecosystem. An example is the higher resistance of species-rich natural systems to pest outbreaks compared to low-diversity agricultural ecosystems growing under the same environmental conditions. The spatial arrangement of individuals in an ecosystem can affect their risk to disease, predation, or consumption. In higher diversity systems, the mean distance between individuals of the same species is on average greater than that of low-diversity systems. The wider spacing of individuals acts to slow the movement of pathogenic organisms, which should limit the occurrence of pest outbreaks that alter the performance of the ecosystem. These and other observations lead to the general expectation that diversity increases the resistance of ecosystems to disturbance.

The benefits of biodiversity to ecosystem functioning should be multiple since the processes of production and nutrient cycling are coupled by the biological interactions of organisms. The response of a Minnesota grassland to a severe drought (disturbance) illustrates this principle (Tilman et al., 1996). In 1987 and 1988, a drought decreased productivity of the grassland. The species diversity of experimental plots prior to the drought explained the degree of productivity loss. Diversity in productivity was about 50% decline in productivity, whereas productivity in the least diverse plots declined by more than 90%. The greater resistance of the higher diversity plots resulted from compensatory increases in productivity shown by drought-resistant species. The more diverse plots also had lower concentrations of nitrate in the rooting zone, indicating a more efficient use of this limiting resource.

This experiment demonstrates that species diversity has an effect on productivity and nutrient cycling and that declining species diversity influences these functions. However, we lack an understanding of the mechanisms producing these patterns of ecosystem response to disturbance and biodiversity change. Increasing diversity may increase the chance that a single drought-adapted and productive species will be present in the community, ensuring relatively high productivity. Alternatively, higher diversity may provide for a more efficient utilization of limiting resources, as suggested by the lower soil nitrate in more diverse plots. Before the basic relationships between biodiversity and ecosystem functioning can be more fully formalized, we need more detailed information on the critical levels (thresholds) of diversity associated with specific ecosystem functions and how environmental conditions operating over time alter their relationship (Folke et al., 1996).

IV. CONCLUSIONS

Humans have become major agents of environmental change and influence the biodiversity and structure of ecosystems in many ways. Air pollution, clearing of natural systems for agriculture, forestry and urban development, the spread of exotic species, changes in the composition of the atmosphere, and other anthropogenic influences are altering ecosystem functioning. By changing ecosystem biodiversity and altering the processes that biota mediate, we significantly decrease the ability of ecosystems to provide services and resources for our use. The management of ecosystems for sustained levels of services and the restoration of damaged ecosystems will require greater knowledge about the role that species play in ecosystems functions related to production and nutrient cycling. Although we cannot know with certainty the roles of most species in ecosystems, it is prudent to assume that all biodiversity is essential to ecosystem function and stability and should be valued and protected.

See Also the Following Articles

ECOSYSTEM, CONCEPT OF • ECOSYSTEM SERVICES, CONCEPT OF • ENERGY FLOW AND ECOSYSTEMS • KEYSTONE SPECIES • NITROGEN CYCLE

Bibliography


ECOSYSTEM SERVICES, CONCEPT OF

Gretchen Daily* and Shamik Dasgupta†
*Stanford University and †University College London

I. Overview
II. History
III. Biophysical Characterization
IV. The Ecosystem Services Framework
V. Safeguarding
VI. Conclusions

GLOSSARY

ecosystem services The wide array of conditions and processes through which ecosystems, and their biodiversity, confer benefits on humanity; these include the production of goods, life-support functions, life-fulfilling conditions, and preservation of options.
marginal value Economic value of the next incremental unit of something. In this context, marginal values are those associated with managing the next small unit of an ecosystem in a particular way (e.g., preserving, rather than clearing, the next unit of forest).

ECOSYSTEM SERVICES ARE ESSENTIAL TO HUMAN EXISTENCE, and yet their supply is seriously threatened by the intensification of human impacts on the environment. This article provides an overview of issues concerning the identification, biophysical and economic characterization, and safeguarding of ecosystem services.

I. OVERVIEW

Human societies derive many benefits from natural ecosystems. These include the production of a diversity of ecosystem goods, or extractive benefits, such as seafood, timber, biomass fuels, and precursors to many industrial and pharmaceutical products. The harvest and trade of these goods represent an important and familiar part of the economy. Ecosystem services also include non-extractive benefits—fundamental life-support processes including pollination, water purification, renewal of soil fertility, and climate regulation. Ecosystem services include life-fulfilling functions, encompassing aesthetic beauty and the cultural, intellectual, and spiritual values derived from nature. Finally, preservation of the option to use these (or new) services in the future is also an important service in itself.

One way to appreciate the nature and value of ecosystem services is to imagine trying to set up a happy existence on the moon. Assume for the sake of argument that the moon miraculously already had some of the basic conditions for supporting human life, such as an atmosphere and climate similar to those on Earth. After packing one’s prized possessions, the big question would be, Which of Earth’s millions of species would be required to sustain the lunar colony?

Tackling the problem systematically, one could first choose from among all the species exploited directly for food, drink, spice, fiber and timber, pharmaceuticals, industrial products (such as waxes, lac, rubber, and oils), and so on. Even being selective, this list could
amount to hundreds, or even several thousand, species. The space-ship would be filling up before even begin-
ing to add the species crucial to supporting those at
the top of one’s list. Which are these unsung heroes? No one knows which—nor even approximately how
many—species are required to sustain human life. This
means that rather than listing species directly, you
would have to list instead the life-support functions
required by your lunar colony; then you could guess
at the types and numbers of species required to perform
each. At a bare minimum, other companions on the
spaceship would have to include species capable of
supplying a whole suite of ecosystem services that
Earthlings take for granted. Table I provides a classifi-
cation of important ecosystem services.

<table>
<thead>
<tr>
<th>TABLE I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A Classification of Ecosystem Services with Illustrative Examples</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Production of Goods</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
</tr>
<tr>
<td>Terrestrial animal and plant products</td>
</tr>
<tr>
<td>Forage</td>
</tr>
<tr>
<td>Seafood</td>
</tr>
<tr>
<td>Spices</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Pharmaceuticals</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicinal products</td>
</tr>
<tr>
<td>Precursors to synthetic pharmaceuticals</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Durable materials</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural fiber</td>
</tr>
<tr>
<td>Timber</td>
</tr>
<tr>
<td>Energy</td>
</tr>
<tr>
<td>Biofuels</td>
</tr>
<tr>
<td>Low-sediment water for hydropower</td>
</tr>
<tr>
<td>Industrial products</td>
</tr>
<tr>
<td>Waxes, oils, fragrances, dyes, latex, rubber, etc.</td>
</tr>
<tr>
<td>Precursors to many synthetic products</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Genetic resources</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Which enhance the production of many of these goods</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Regeneration Processes</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycling and filtration processes</td>
</tr>
</tbody>
</table>
| Detoxification and decomposi-
tion of wastes                    |
| Generation and renewal of soil fertility |
| Purification of air              |
| Purification of water            |

<table>
<thead>
<tr>
<th><strong>Translocation processes</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dispersal of seeds necessary for revegetation</td>
</tr>
<tr>
<td>Pollination of crops and natural regeneration</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Stabilizing Processes</strong></th>
</tr>
</thead>
</table>
| Coastal and river channel sta-
bility                         |
| Compensation of our species for another under varying conditions |
| Control of the majority of potential pest species |
| Moderation of weather extremes (such as of temperature and wind) |
| Partial stabilization of climate |
| Regulation of hydrological cycles (mitigation of floods and droughts) |

<table>
<thead>
<tr>
<th><strong>Life-Fulfilling Functions</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aesthetic beauty</td>
</tr>
<tr>
<td>Cultural, intellectual, and spiritual inspiration</td>
</tr>
<tr>
<td>Existence value</td>
</tr>
<tr>
<td>Scientific discovery</td>
</tr>
<tr>
<td>Serenity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Preservation of Options</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintenance of the ecological components and systems needed for future supply of these goods and services and others awaiting discovery</td>
</tr>
</tbody>
</table>

Armed with this preliminary list of services, one
could begin to determine which types and numbers of
species are required to perform each. This is no simple
task. Consider soil fertility. Soil organisms play impor-
tant and often unique roles in the circulation of matter
in every ecosystem on Earth, they are crucial to the
chemical conversion and physical transfer of essential
nutrients to higher plants, and all larger organisms,
including humans, depend on them. The abundance of
soil organisms is tremendous: under a square yard of
pasture in Denmark, for instance, the soil was found
to be inhabited by roughly 50,000 small earthworms
and their relatives, 50,000 insects and mites, and nearly
12 million roundworms. And that is not all. A single
gram (a pinch) of soil has yielded an estimated 30,000
protozoa, 50,000 algae, 400,000 fungi, and billions of
individual bacteria (Overgaard-Nielsen, 1955). Which
to bring to the moon? Most of these species have never
been subjected to even cursory inspection. Yet the so-
berefit fact of the matter is, as Ed Wilson put it: they
don’t need us, but we need them.

In the early 1990s, the first Biosphere 2 “mission”
carried out this thought experiment to the greatest de-
gree possible on Earth. Eight people were enclosed in
a 3.15-acre closed ecosystem, featuring agricultural land
plus a wide array of natural habitats (desert, savanna,
tropical forest, wetland, and even a miniature ocean).
The aim was to demonstrate a (mostly) closed system
that could supply people with their material needs for
two years. Yet in spite of an investment of over $200
million in the design, construction, and operation of
this model Earth (including $1 million in annual energy
inputs), it proved impossible to do so and the experi-
ment was shut down early in failure. Numerous un-
pleasant and unexpected problems arose, including a
fall in atmospheric oxygen concentration to 14% (the
level normally found at an elevation of 17,500 feet);
ECOSYSTEM SERVICES, CONCEPT OF

High spikes in carbon dioxide concentrations; nitrous oxide concentrations high enough to impair brain function; rapid species extinctions (including 19 of 25 vertebrate species and all pollinators, therebydooming most of the plant species to eventual extinction as well); overgrowth of aggressive vines and algal mats; and, to top it all off, population explosions of crazy ants, cockroaches, and katydids. Even heroic personal efforts on the part of the Biospherians did not suffice to make the system viable and sustainable for humans nor many nonhuman species, illustrating the tremendous expense and difficulty of replicating many basic ecosystem services (Cohen and Tilman, 1996).

Ecosystem services are generated by a complex of natural cycles, driven by solar energy, that constitute the workings of the biosphere—the thin layer near Earth’s surface that contains all known life. The cycles operate on very different scales. Biogeochemical cycles, such as the movement of the element carbon through the living and physical environment, are truly global and reach from the top of the atmosphere to deep into soils and ocean-bottom sediments. Life cycles of bacteria, in contrast, may be completed in an area much smaller than the period at the end of this sentence. The cycles also operate at very different rates. The biogeochemical cycling of carbon, for instance, occurs at a rate that is orders of magnitude faster than that of phosphorus, just as the life cycles of microorganisms may be many orders of magnitude faster than those of trees.

All of these cycles are ancient, the product of billions of years of evolution, and have existed in forms very similar to those seen today for at least hundreds of millions of years. They are absolutely pervasive, but unnoticed by most human beings going about their daily lives. Who, for example, gives a thought to the part of the carbon cycle that connects him or her to Earth’s surface that contains all known life. The cycles operate on such a grand scale and in such intricate and little-explored ways that most could not be replaced by technology.

Human activities are already impairing the flow of ecosystem services on a large scale. And in such intricate and little-explored ways that most could not be replaced by technology.

In addressing these points in more detail, we shall first consider briefly the history of concern for ecosystem services. Then we shall explore the functioning of a small set of services in biophysical terms; the economic characterization of ecosystem services is discussed elsewhere in the Encyclopedia. Finally, we will examine the utility of the Ecosystem Services Framework, in the abstract and then in operation, reviewing recent developments in the safeguarding of ecosystem services.

II. HISTORY

Interestingly, the nature and value of Earth’s life-sustaining ecosystems have been illuminated primarily through their disruption and loss. Thus, for instance, deforestation has demonstrated the critical role of forests in the hydrological cycle—in particular, in mitigating floods, droughts, the erosive forces of wind and rain, and siltation of waters and irrigation canals. Release of toxic substances, whether accidental or deliberate, has revealed
the nature and value of physical and chemical processes, governed in part by a diversity of microorganisms, that disperse and break down hazardous materials. Thinning of the stratospheric ozone layer sharpened awareness of the value of its service in screening out harmful ultraviolet radiation.

A cognizance of ecosystem services, expressed in terms of their loss, dates back at least to Plato around 400 B.C. and probably much earlier:

What now remains of the formerly rich land is like the skeleton of a sick man with all the fat and soft earth having wasted away and only the bare framework remaining. Formerly, many of the mountains were arable. The plains that were full of rich soil are now marshes. Hills that were once covered with forests and produced abundant pasture now produce only food for bees. Once the land was enriched by yearly rains, which were not lost, as they are now, by flowing from the bare land into the sea. The soil was deep, it absorbed and kept the water . . . , and the water that soaked into the hills fed springs and running streams everywhere. Now the abandoned shrines at spots where formerly there were springs attest that our description of the land is true.

—PLATO


Mooney and Ehrlich (1997) trace modern concern for ecosystem services to George Perkins Marsh, a lawyer, politician, and scholar. Indeed, his 1864 book Man and Nature describes a wide array of services, again, often expressed in terms of their loss. Remarkably on the terrain of the former Roman Empire, he notes that it “is either deserted by civilized man and surrendered to hopeless desolation, or at least greatly reduced in both productiveness and population” (p. 9). He goes on to describe the reduction of hydrological services: “Vast forests have disappeared from mountain spurs and ridges, the vegetable earth . . . [is] washed away; meadows, once fertilized by irrigation, are waste and unproductive, because . . . the springs that fed them dried up; rivers famous in history and song have shrunk to humble brooklets” (p. 9). Marsh also made connections between deforestation and climate: “With the disappearance of the forest, all is changed. At one season, the earth parts with its warmth by radiation to an open sky—receives, at another, an immoderate heat from the unobstructed rays of the sun. Hence the climate becomes excessive, and the soil is alternately parched by the fervors of summer, and seared by the rigors of winter. Bleak winds sweep unrestrained over its surface, drift away the snow that sheltered it from the frost, and dry up its scanty moisture” (p. 186). Finally, he even wrote of decomposition services: “The carnivorous, and often the herbivorous insects render an important service to man by consuming dead and decaying animal and vegetable matter, the decomposition of which would otherwise fill the air with effluvia noxious to health.” (p. 98)

Following World War II, other eloquent writers on the environment emerged, including Fairfield Osborn (Our Plundered Planet, 1948), William Vogt (Road to Survival, 1948), and Aldo Leopold (A Sand County Manac and Sketches from Here and There, 1949). Each discusses ecosystem services without using the term explicitly. In The Population Bomb (1968), Paul Ehrlich describes anthropogenic disruption of ecosystems and the societal consequences of doing so, addressing the need to maintain important aspects of ecosystem functioning. Along these lines, the Study of Critical Environmental Problems (1970) presents a list of key “environmental services” that would decline with a decline in “ecosystem function,” including many in Table I. This list was expanded upon by Holdren and Ehrlich (1974). By the early 1980s, efforts were initiated to investigate two questions: the extent to which ecosystem function (and the delivery of services) depends on biodiversity, and the extent to which technological substitutes could replace ecosystem services. The first question is addressed elsewhere in this Encyclopedia, and the second question was tackled by Ehrlich and Mooney (1983). Work on these topics proliferated and, in 1997, a collective effort was made to synthesize the wealth of scientific information that had accumulated on the functioning of ecosystem services, with a preliminary exploration of their economic value, and of key issues meriting further work (Daily, 1997).
a sevenfold increase or more in some countries. It is difficult to forecast how these needs will be met. Whatever course is taken, it is reasonable to expect that growing human pressures will continue to alter ecosystems, probably very dramatically.

At the same time, food production is highly dependent on ecosystem services. In what ways? Let us consider four types of services that are key to agriculture: climate stability, services supplied by soil, pollination, and pest control. In each case, an overview of the functioning of the service will be given, along with a general assessment of its importance.

A. Climate Stability

Earth’s climate has fluctuated tremendously since humanity came into being, but it has been relatively stable since the Invention of agriculture around 10,000 years ago. At the peak of the last ice age 20,000 years ago, for example, much of Europe and North America were covered by mile-thick ice sheets. Even relatively recently, from 1530 to 1850, Europe was significantly cooler during a period known as the Little Ice Age. Many of these changes in climate are thought to be caused by alterations in Earth’s orbital rotation or in the energy output of the sun, or by sudden perturbations (such as violent volcanic eruptions and asteroid impacts) or more gradual tectonic events (such as the development of the Himalayas). Remarkably, climate has been buffered enough through all of these changes to sustain life for at least 3.5 billion years—although the increases in climate change is thought to have induced the past mass extinctions. Human activities now threaten to induce rapid climatic change, with wide-reaching consequences for society.

Climate plays a major role in the evolution and distribution of life over the planet. Most scientists would also agree that life itself is a principal factor in the regulation of global climate, helping to dampen the effects of episodic perturbations through negative feedback mechanisms that offset climate oscillations with changes in greenhouse gas concentrations. For instance, natural ecosystems may have helped to stabilize climate by removing CO₂ from the atmosphere. The relative influence of stabilizing and destabilizing feedbacks remains uncertain; what is clear is that climate and natural ecosystems are tightly coupled, and maintaining the stability of that coupled system is an important ecosystem service.

Ecosystems also help moderate regional and local weather. For instance, evapotranspiration of plants in the morning contributes moisture to the atmosphere that then falls in thunderstorms in the afternoon, damping both moisture loss and surface temperature rise. In the Amazon, for example, 50% of the mean annual precipitation is recycled via evapotranspiration by the forest itself. Amazon deforestation could so dramatically reduce precipitation that the forest might be unable to reestablish itself following large-scale destruction. Temperature extremes are also moderated by forests, which provide shade and surface cooling and also act as insulators, blocking searing winds and trapping warmth by acting as a local greenhouse agent.

B. Services Supplied by Soil

The ecosystem services supplied by soil are so tightly interrelated as to make any discrete classification arbitrary. We shall briefly touch on six services here: (i) physical support of plants; (ii) retention and delivery of nutrients to plants; (iv) disposal of wastes and dead organic matter; (v) regulation of major element cycles.

An enormous amount of water, about 119,000 km³, falls annually on Earth’s land surface—enough to cover the land to an average depth of 1 meter. Much of this water is soaked up by soils and is gradually metered out to plant roots and into aquifers and surface streams. Without soil, rainfall would rush off the land in flash floods. Plant foliage, roots, and residues shield the soil from the full, destructive force of raindrops, holding it in place. Rain on denuded landscapes compacts the surface and turns soil rapidly to mud (especially if it has been loosened by tillage), which clogs surface cavities in the soil, reduces infiltration, increases runoff, and further enhances clogging in a positive feedback. Detached soil particles are splashed downslope and carried off by running water. Soil also shelters seeds and provides physical support
as they sprout and mature into adult plants. The energetic costs to plants of packaging and storing seeds and of anchoring plant roots would be enormous without soil. Human-engineered hydroponic systems grow plants in the absence of soil and their cost provides a lower-bound index to the value of this service. The physical support trays and stands used in such operations amount to about U.S. $55,000 per hectare.

Third, soil retains and delivers nutrients to plants. Tiny soil particles, primarily bits of humus and clays, carry a surface electrical charge that is generally negative. This property holds positively charged nutrients (cations) near the surface, in proximity to plant roots, allowing them to be taken up gradually. Otherwise, they would quickly be leached away. Soil also acts as a buffer in the application of fertilizers, holding the fertilizer ions on soil exchange sites until required by plants.

Fourth, soil plays a paramount role in the decomposition of dead organic matter and wastes, in the process rendering harmless many potential human pathogens. People generate a tremendous amount of waste, including household garbage, industrial waste, crop and forestry residues, and sewage from their own populations and their billions of domesticated animals. A rough approximation of the amount of dead organic matter and waste (mostly agricultural residues) processed each year is 130 billion metric tons, about 30% of which is associated with human activities. Fortunately, there is a diverse array of decomposing organisms—ranging from vultures to tiny bacteria—that earn their living by extracting energy from the large, complex organic molecules found in many types of waste. The simple inorganic chemicals that result from natural decomposition are eventually returned to plants as nutrients. Thus the decomposition of wastes and the recycling of nutrients, the fifth service, are two aspects of the same process. The fertility of soils—that is, their ability to supply nutrients to plants—is largely the result of the activities of diverse species of bacteria, fungi, algae, crustacea, mites, termites, springtails, millipedes, and worms, all of which, as groups, play important roles. Some bacteria are responsible for “fixing” nitrogen (a key element in proteins), drawing it out of the atmosphere; the latter cause acid precipitation, eutrophication, and contamination of groundwater and surface drinking water sources by high nitrate-nitrogen levels.

C. Pollination

Animal pollination is required for the successful reproduction of most flowering plants, including both wild plants and about 70% of the agricultural crop species that feed the world. Over 100,000 different animal species (including bats, bees, beetles, birds, butterflies, and flies) are known to provide “free” pollination services, which maintain croplands, backyard gardens, meadows, forests, and rangelands. The availability of these pollinators depends on the existence of a wide variety of habitat types needed for their feeding, successful breeding, and completion of their life cycles.

Considering the proportions in which we consume different crops, about one-third of human food comes from animal-pollinated plants. Critical grain crops, such as rice and wheat, are not animal-pollinated. Without natural pollination services, yields of other important crops would decline precipitously and many wild plant species would become extinct. In the United States alone, billions of dollars of crops would be lost (annually) without the services of wild, native pollinators—those sustained by natural habitats adjacent to farmlands. Pollination by honeybees, originally imported from Europe, is extremely important as well, but they are presently in decline, making pollinators from natural ecosystems all the more important. Management of the honeybee in the New World is currently threatened by the movement of, and hybridization with, the aggressive African strain of honeybee that was accidentally released in Brazil in 1956. Diseases of honeybee colonies are also causing a marked decline in the number of managed colonies. Meanwhile, the diversity of pollinators available to both wild and domesticated plants is diminishing: more than 60 genera of pollinators include species now considered to be threatened, endangered, or extinct.
D. Natural Pest Control Services

Humanity’s competitors for food, timber, cotton, and other fiber are called pests; these include numerous herbivorous insects, rodents, fungi, snails, nematodes, and viruses. These pests destroy an estimated 25% (up to as much as 30%) of crops, before and after harvest. In addition, numerous weeds compete directly with crops for water, light, and soil nutrients, further limiting yields.

Chemical pesticides, and the strategies with which they are applied, can have harmful unintended consequences. First, resistance is now found in over 500 insect and mite pests, over 100 weeds, and in about 150 plant pathogens. This means that achieving a fairly constant level of pest control requires applying higher and higher doses of pesticides—or continuously developing new, more powerful chemicals to combat evolving pest defenses. Second, populations of natural pest enemies are decimated by heavy pesticide use. Not having the same evolutionary experience with plant chemicals that the pests themselves have had, and typically having much smaller population sizes (being higher on food chains), natural predators are often more susceptible to synthetic poisons than are the pests. Destruction of predator populations leads to explosions of their prey and the “promotion” of species previously not classified as pests to pest status. In California in the 1970s, for instance, 24 of the 25 most important agricultural pests had been “created” by the overuse of pesticides. Third, exposure to pesticides and herbicides may pose serious health risks to humans and many other organisms, the recently discovered declines in human sperm counts may be attributable in part to such exposure.

Fortunately, an estimated 99% of potential crop pests are controlled by natural enemies, including many birds, spiders, parasitic wasps and flies, ladybugs, fungi, viral diseases, and numerous other types of organisms. These natural biological control agents save farmers and society billions of dollars annually by protecting crops and reducing the need for chemical control.

IV. THE ECOSYSTEM SERVICES FRAMEWORK

All of these services are presently under threat. When human activities approach or exceed the environment’s capacity to sustain them, growth in those activities is rarely brought to an immediate halt. Rather, the people so engaged suddenly find themselves confronted with a set of trade-offs in the allocation of resources to competing uses and users. These trade-offs are becoming increasingly vexing and difficult to resolve, from both ethical and practical perspectives. They involve our most important ideals (such as ensuring a prosperous future for our children), our oldest tensions (such as between individual and societal interests), and sometimes our bloodiest tendencies.

At the local level, allocation of land or water to competing activities often involves a zero-sum game. This is apparent in the widespread loss of water and land from native habitat to farms, and increasingly to urban and industrial uses. On what basis should such allocations be decided? How can individual preferences for alternative allocations be aggregated fairly? How can the costs and benefits of alternative schemes be distributed fairly? And how can the parties with the most at stake—future generations—be represented at the bargaining table? At the international level, these questions are writ large. Consider efforts to allocate among nations permits to produce chlorofluorocarbons, to harvest certain marine fish stocks, or to use the global carbon dioxide sink. How these questions are decided will profoundly influence the willingness of nations and individual actors to make and comply with agreements.

The ecosystem services approach provides a conceptual framework for helping to resolve these trade-offs. The framework recognizes natural ecosystems and their biodiversity as capital assets that, if properly managed, will yield a stream of life-support goods and services over time. Relative to physical, human, or financial capital, renewable natural capital (embodied in ecosystems) is poorly understood, typically undervalued, scarcely monitored, and—in many important cases—undergoing rapid depletion. Up until now, there has been little incentive to measure or manage natural capital: it has been treated as essentially inexhaustible.

Measuring natural capital involves assessing both its stock and its importance to society. Neither is simple to do, but the latter is particularly difficult: valuation in economic or other (e.g., cultural, spiritual) terms involves resolving fundamental philosophical issues (such as the underlying bases for value), the establishment of context, and the defining of objectives and preferences, all of which are inherently subjective. Even after doing so, one is faced with formidable technical difficulties with interpreting information about the world and transforming it into a quantitative measure of value. Just as the full value of a human being cannot be captured in economic terms (on the basis of his or her wage-earning power, or the economic value of his or her constituent materials), there exists no absolute...
value of ecosystem services waiting to be discovered and revealed to the world by a member of the intellectual community. Other entries in the Encyclopedia discuss these issues and describe alternative empirical valuation techniques, their applicability to different types of ecosystem services, and the advantages and limitations of their use.

As a whole, ecosystem services have infinite value because human life could not be sustained without them. The evaluation of the trade-offs currently facing society, however, requires that we estimate the marginal value of ecosystem services (the value yielded by an additional unit of the service, all else held constant) to determine the costs of losing—or the benefits of preserving—a given amount or quality of services. The information needed to estimate marginal values is also difficult to obtain and is presently unavailable for many aspects of the services. Nonetheless, even imperfect measures of their value, if understood as such, may prove superior to ignoring ecosystem services altogether, as is generally done in decision making today.

V. SAFEGUARDING

How can the Ecosystem Services Framework be made operational to ensure the safeguarding of key services? Here success hinges on the institutional framework in which valuation is done and incorporated into policy. Valuation is not a solution in itself—it is merely one way of organizing information in the much larger politics of decision making. The actual safeguarding of ecosystem services will require the establishment (or strengthening) of institutions that reward such action. Institutions must be tailored to local ecological, economic, and cultural circumstances and provide a payoff for conservation in a relatively short and certain time frame to be effective (see Heal, in press).

The management by New York City of its drinking water quality offers a model example of how ecosystem services can be characterized (both biophysically and economically), monitored, and safeguarded. The city’s water originates in the Catskill Mountains, about a hundred miles to the north of the city, and was once considered so pure and salubrious that it was bottled and sold throughout the Northeast. In recent years, the natural purification system of the Catskills has been overwhelmed by sewage and agricultural runoff, resulting in a reduction in water quality below minimum drinking standards. The United States Environmental Protection Agency (EPA) required that the city remedy this problem.

The city administration first considered replacing the natural system with a filtration plant, but found that it would cost an estimated $6–8 billion in capital plus $300 million in annual operations. These high costs prompted investigation of an alternative solution, namely, restoring and safeguarding the natural purification services of the Catskills. This would involve the purchase of land in and around the watershed to protect it, as well as subsidizing several changes on privately owned land: the upgrading of sewage treatment plants, the improvement of farming practices, and “environmentally sound” economic development. The total cost of this option was estimated at about $1.5 billion.

Thus, New York City faced a choice of investing $6–8 billion in physical capital or $1.5 billion in natural capital. In spite of some political opposition, the latter option clearly appeared to be the more attractive, so the city floated an Environmental Bond Issue to fund its implementation. This financial mechanism captures the important economic and public health values of a natural asset (the watershed) and distributes them to those assuming stewardship responsibilities for the asset and its services.

The Catskills supply many other valuable services, such as flood control, carbon sequestration (locking up, in forests and the soils beneath them, carbon dioxide, an important greenhouse gas), biodiversity conservation, and—perhaps above all—beauty, serenity, and spiritual inspiration. Moreover, these services benefit others besides water consumers in New York City. It would be absurd to try to express the full ecosystem service value of the Catskills in a dollar figure. In this case, fortunately, there was no reason to try: even a lower-bound estimate of the value of the natural asset was sufficient to induce adoption of a conservation policy.

The challenge is to extend this model to other geographic locations and to other services. In other places, water quality is certainly a growing concern: the United Nations Environment Programme reports that most diseases in the less-developed world are caused by contaminated water, and that 30% of people in these countries suffer from one or more water-related diseases. An estimated 10–15% of the U.S. population currently consumes water from systems that violate EPA contaminant standards. The EPA forecasts that communities will have to spend $1–40 billion over the next 20 years to maintain drinking water quality at minimum required standards.

To evaluate the potential for extending the New York City model, Walter Reid posed this question: How much land could be protected with an economic justification, using water quality as a major goal? A first-order ap-
proximation suggests that the area would be significant: 10% of U.S. land area and 14% of global land area under current population sizes. Reid further points out that hydroelectricity generation is another important good whose supply, like that of drinking water, economically justifies watershed protection in many instances. Costa Rica, for example, derives 99% of its electricity from hydroelectric plants; most of the water generating this electricity flows from protected areas. In 1999, Costa Rica established a small tax on water and electricity use, the revenues from which are now recycled into managing the protected areas to limit erosion and sedimentation and to maintain high water flows. Numerous ancillary benefits accrue in this scheme, including protection of valuable flood control, biodiversity conservation, and ecotourism services.

To extend this model to other places and other services, appropriate educational, financial, and legal institutions, tailored to cultural and economic circumstances, will be required. Without these, statements from ecologists and economists that ecosystems are important and valuable assets will accomplish little. Promising new institutions for safeguarding ecosystem services have emerged in a wide array of cultures and economies (e.g., Australia, Madagascar, the United States, Vietnam); at a variety of scales, from local to international; and among governments, nongovernmental organizations, and private sectors. The services safeguarded by these emerging institutions include pollination; pest control; water supply for drinking, for irrigation, and for hydropower generation; maintenance of soil fertility; sustainable harvesting of tropical timber; provision of aesthetic beauty; and even decomposition (of orange peels produced by Del Oro, an orange juice company in Costa Rica, which are carefully distributed in a reserve area).

VI. CONCLUSIONS

An enormous payoff could result from further research in the characterization (biophysical and economic) of ecosystem services and in the development of institutions for their safeguarding. A series of basic questions, spanning a wide array of disciplines, beg addressing in this area. With emphasis here on biophysical characterization, they include:

• Which ecosystems supply what services? What is the scale of delivery, transport, and consumption of the services?
• What are the relationships between the quantity or quality of services and the condition (e.g., relatively pristine versus heavily modified) or areal extent of the ecosystem supplying them? Where do critical thresholds lie?
• To what extent do the services depend on biodiversity?
• To what extent, and over what time scale, are the services amenable to repair?
• How interdependent are the services? How does safeguarding or damaging one influence the functioning of others?
• What indicators could be used to monitor accurately and efficiently the changes in the supply or quality of ecosystem services?
• How effectively, and on how large a scale, can existing or foreseeable human technology substitute for ecosystem services?
• What are the main sources of uncertainty regarding ecosystem services, and how important are they? How can the uncertainty best be quantified and incorporated into policy?
• How can economic principles and tools best be brought to bear on the management of natural capital?
• Given that many values of ecosystems lie mostly in the future (and will always lie mostly in the future), how should future benefits be valued, in economic, cultural, or other terms?
• What financial, legal, and other social institutions are needed to safeguard critical ecosystem services? How can their development be catalyzed?

While a great deal is known about the functioning of ecosystems and the supply of services in general, abstract terms, there is a paucity of information on particular, local ecosystems and economies. Very little is known about marginal values (the net benefit or cost associated with protecting or destroying the next unit of an ecosystem) or about the nonlinearities in ecosystem responses to human impact. Often this information is not acquired until after it is too late to reverse harm done (e.g., after heavy flooding). And we still have relatively little experience with institutional mechanisms for safeguarding ecosystem services: these will vary with the different ecological and economic characteristics of the services (such as whether they are public or private goods, the time period and spatial scale over which benefits are realized, and so on).

Further development of case studies addressing these issues would be most helpful. Such work would define the envelope of opportunities and limitations in applying this conceptual framework; it would illumi-
nate how general are the findings from specific locali-
ties; and it would serve as a guide to policy develop-
ment. In the New York City case, for instance, officials
are purchasing land and changing agricultural and mu-
nicipal practices in the hopes of restoring the natural
water purification services of the Catskills—all with
quite limited scientific information. In this particular
case, and generally, success in the policy arena hinges
on whether the scientific underpinnings of policies
are sound.

See Also the Following Articles

BIODIVERSITY AS A COMMODITY • ECONOMIC VALUE OF
BIODIVERSITY, OVERVIEW • ECOSYSTEM, CONCEPT OF •
ECOSYSTEM FUNCTION, PRINCIPLES OF • HISTORICAL
AWARENESS OF BIODIVERSITY

Bibliography


lessons so far. Science 274, 1130–1131.

Natural Ecosystems. Island Press, Washington, D.C.

Daily, G. C., Alexander, S., Ehrlich, P. R., Gosdler, I. H., Lubchenco,
J., Matson, P. A., Mooney, H. A., Postel, S., Schneider, S. H.,
Tilman, D., and Woodwell, G. M. (1997). Ecosystem services:
Benefits supplied to human societies by natural ecosystems. Issues
in Ecology 2, 1–16.

Daily, G. C., Soederqvist, T., Arrow, K., Dasgupta, P., Ehrlich, P.,
Folke, C., Jansson, A. M., Jansson, B.-O., Levin, S., Lubchenco,
J., Maler, K.-G., Starrett, D., Tilman, D., and Walker, B. (In re-
view). The value of nature and the nature of value.

Ehrlich, P. R., and Mooney, H. A. (1983). Extinction, substitution,

ton, D.C.

Holdren, J. P., and Ehrlich, P. R. (1974). Human population and the

Jansson, A. M., Hammer, M., Folke, C., and Costanza, R. (eds.).
Approach to Sustainability. Island Press, Washington, D.C.

Fragmentary history. In Nature’s Services: Societal Dependence on
Natural Ecosystems (G. C. Daily, ed.), pp. 11–19. Island Press,
Washington, D.C.


Reid, W. A. (In press). A business plan for ecosystem services: Ex-
tending the New York City watershed model to other geographic
regions and other ecosystem services.

Vitousek, P. M., Mooney, H. A., Lubchenco, J., and Molino, J. M.
(1997). Human domination of Earth’s ecosystems. Science 277,
494–499.
ECOTOXICOLOGY CAN BE DEFINED as “the study of harmful effects of chemicals upon ecosystems.” This implies that ecotoxicology is not concerned with the detection of chemicals per se, but that it deals with biological effects of (toxic) chemicals that contaminate (or have contaminated) the environment. Biological effects induced by chemicals can be anything from a molecular effect (e.g., genetic change) in an individual organism to effects on the biosphere as a whole.

I. INTRODUCTION

The manifest principles of ecotoxicology involve the application of the principles of toxicology to the environment: focused on human activity leading to the release of molecules such as polynuclear aromatic hydrocarbons (PAH), polychlorinated biphenyls (PCBs), xenoestrogens (XEs), chlorofluorocarbons (CFCs), pes-
...ticides, heavy metals, radionuclides, greenhouse gases (CO_2, CH_4), sulfur dioxide (SO_2) and oxides of nitrogen (NO_x) into the environment. Biomonitoring is undertaken sometimes as the "prelude to remediation." Remediation aims to remove potential toxic substances from a contaminated site, thus restoring ecosystem function as far as is reasonably possible following the removal of the ecotoxicity by biological, physical, or chemical means.

Molecular toxicology is the branch of toxicology (the study of the effects of manifestly poisonous substances on individuals) that adopts the biochemical approach to the understanding of the detrimental threats to life, often recognized by a range of morbid conditions in a variety of fauna and flora. Such pathology can be understood at a molecular level in terms of biomolecular damage undergone by such macromolecules as DNA (deoxyribonucleic acid: the genetic determinant mainly residing in the nucleus of the living cell), RNA (ribonucleic acid: responsible for transfer of the genetic message inherent in DNA structure sequence to the cytoplasm of the cell), and diverse proteins constructed from a choice of sequence of residues of 20 amino acids (obtained mainly from dietary protein hydrolysis during digestion).

It is the increased understanding of chemistry and its associated biochemistry that allows a meaningful prediction of toxicity of chemicals in all forms of life (ranging from bacteria and higher fungi to plants and animals including Homo sapiens). Although no meaningful distinction can be made between natural and manmade chemicals released into the environment, detoxifying enzymes such as the ubiquitous oxygen-requiring enzyme family, cytochromes P-450, present in most forms of life, are useful biomarkers for constitutive or acquired tolerance to chemicals such as polycyclic aromatic hydrocarbons (PAHs), polycyclic halogenated biphenyls (PCBs), and xenoestrogens (XE). In the course of millions of years of biological evolution of species, survivors now display a range of such biochemical defenses against toxins including the more recent pollutants released through industrial activity in processing and manufacturing industries. Particular concern has been displayed recently for the contamination of the environment with endocrine disruptors (environmental hormones). These mimic the biochemistry of natural estrogens (such as 17β-estradiol) and these have been associated with feminization of males, leading to displayed hermaphroditism in a few cases, in polluted environments. Some of these endocrine disruptors are derived from the release of industrial chemicals and are therefore referred to as xenestrogens (foreign-compound estrogens). This is using the accepted nomenclature of xenobiotics as "foreign" or "manmade" chemicals, as contrasted with natural chemicals of which the environment, the diet, and all life is composed.

Several important concepts have been incorporated into the principles of ecotoxicology. For example, "hazardous chemicals" have the potential for causing harm to a variety of organisms; but the assessment of the risk to the biosphere posed by these chemicals must be based on consideration of the exposure (dose) to individuals, populations and communities as well as direct toxicity to the different species involved. Environmental exposure is not only determined by the quantity of a chemical released but depends also on the characteristics of a particular environment and the distribution of the biota in that environment. Soil that contains large quantities of clay minerals and humic substances might lead to immobilization of hazardous chemicals through absorption, while chemicals released in aquatic environments might be dispersed rapidly. In both cases, toxicity to the biota living in these environments might be limited as exposure (bioavailability) is small. Also the distribution of different species in the environment is not random, but populations of a particular species are concentrated at particular sites (niche). For example, in soil some species are found in the litter that lays on the soil surface (epigeic and hemigeic species), while others frequent the deeper mineral soil layers (eudaphic species). It is clear that epigeic and hemigeic species are more likely to be affected by toxic chemicals that enter the soil environment from above than eudaphic species. Besides distribution, exposure will be determined by feeding patterns. Organisms that feed on prey that is likely to have accumulated the substance will be exposed to larger quantities of the toxicant than those that feed on substances that are relatively uncontaminated. Especially recalcitrant, low water soluble chemicals such as PCBs and organochlorines (DDT, for example) can be biomagnified in the environment to lethal concentrations higher up the food chain. Furthermore, toxicity of each hazardous chemical should be related to the metabolism of each major species present, as affected by the toxicant. Therefore, signs of toxicity need to be sought by a battery of methods before meaningful conclusions can be drawn on the environmental risk of a potential chemical released in the environment; environmental pollutants at a specified dose may cause changes in finely balanced ecosystems due to differential toxicity to different species and differential exposure of different species that inhabit the environment.

Much of ecotoxicology relates to the toxic effects of...
chemicals in natural ecosystems, but it must be remem-
bered that man is an integral part of these ecosystems
and effects on the other biotic components of the eco-
systems of which he is part can affect him directly or
indirectly. For example, in relation to bioaccumulation
of toxic chemicals, substances like PCBs, chlorinated
organopesticides, and heavy metals, are likely to accu-
mulate (and cause toxic effects) in the human body
because of the simple fact that man is high up in the food
chain. Indirectly, man is dependent on the biosphere for
the maintenance of the climate, purification of water
and air, and the provision of a sustained food supply.
Disruption of biological activity that impairs the capac-
ity of the living world to fulfill these functions will
ultimately affect humankind’s well-being and survival.
In this respect, therefore, human toxicology and ecotox-
icology are of necessity closely interlinked.

II. APPLICATION OF DIFFERENT
SCIENTIFIC DISCIPLINES
TO ECOTOXICOLOGY

A. Chemistry
Many different xenotoxicants of many chemical struc-
tures are now ubiquitous in aquatic, terrestrial, and
atmospheric environments. Most, however, are present
at extremely low concentration. For example, there may
be 60,000 molecular-species of “hormone-disrupters”
such as the xenoestrogens that may be detected (albeit
by remarkably sensitive detector systems that can re-
spond to even a few hundred molecules for a positive
identification). Nevertheless, most of these xenoestro-
gens are extremely weak molecular-mimics of human
estrogens, and many of these can, paradoxically, stop
the estrogenic response of particular target tissues.
Proposed chemical approaches to ecotoxicology are
often analytical in nature and seek to quantify the effect
of each chemical species that can be identified and
assayed so as to define the concentration-dependence
of any hazardous substance and its consequent environ-
mental or human risk. Quick tests for biohazards are
essential, such as that achieved in the famous Ames
Test for mutagenicity (and related carcinogenicity) in
selected bacteria, for example, by the polycyclic hy-
drocarbon, benzo(a)pyrene. Such mutagenicity (novel
growth-behavior) due to genetic (DNA) damage may
manifest as cancer in mammalian organs (because of
loss of growth-control in particular tissues, due to subtle
alterations in regulation of cellular growth by particu-
lar genes).

B. Biochemistry
Biochemistry studies the chemical processes that take
place in living organisms. The main pertinent principle
of biochemistry that can be applied to ecotoxicology is
the modern concept of enzyme (biocatalytic proteins)
regulation by molecular-intermediates of tissue metabo-
lism in cells. Many xenobiotics (foreign chemicals un-
usual in nature) can interfere with the finely balanced
biochemical reactions of living cells by perturbation of
the web of molecular interactions necessary for life.
Additional problems are due to the molecular dam-
age to DNA (deoxyribonucleic acid) and its related RNA
(ribonucleic acid) along with cellular membranes
(photophosphides that are readily destroyed by reactive
oxygen species (ROS), generated by reactions of these
chemicals with atmospheric molecular-oxygen). The
ecotoxicity of ROS may, however, be ameliorated by
antioxidants in the environment (and by dietary antiox-
idants such as vitamins C, E, A, and D as well as phytoest-
rogen in the diet).

C. Microbiology
The activity of microorganisms is at the basis for most
functions on which life on this planet depends. Due to
their almost limitless metabolic capacities, microbial
communities are responsible for the transformation and
recycling of organic and inorganic molecules in the
environment. This activity results in the maintenance
of nutrient cycles, the maintenance of soil fertility, and
the detoxification of toxic substances in the environ-
ment. Furthermore, many microorganisms are inti-
mately associated with every conceivable higher life
form on this earth. They perform functions without
which higher organisms could not function. Examples
of this kind are organelles that have a microbial origin,
including mitochondria, which are responsible for the
energy generation in eukaryotic organisms (organisms
with a true nucleus), and chlorophyll that allows algae
and higher plants to convert carbon dioxide into sugars
via photosynthesis. Other symbiotic interactions be-
tween microorganisms and plants that are of crucial
importance for plant growth and nutrition are dinitro-
gen fixing microorganisms and mycorrhizae, which
have close associations with the roots of many plant
species. In fact plant roots have adapted specially to
accommodate for these microorganisms. Furthermore,
microorganisms associated with the gut of animals pro-
vide essential vitamins that animals cannot synthesize
themselves. On the negative side, microorganisms are
responsible for causing a wide variety of diseases in
plants and animals. Although this is negative from the viewpoint of the organism (or population) that is affected, disease causing organisms are important in ensuring that excessive population growth of particular populations is curtailed, thus ensuring the maintenance of an environment that contains a rich diversity of species. The interdependence between microorganisms and higher life forms means that toxic effects that affect either group of organisms can have important consequences for the other.

D. Biotechnology

Biotechnology is an applied science that aims to harness different life forms for the benefit of man. Agriculture, antibiotic production, and bioremediation are examples of such applications. Recent advances in molecular biology have given biotechnologists new tools to change the nature of life by genetic modification. This technology has resulted in the creation of organisms that would very unlikely have arisen in nature via normal mating and exchange of genetic material. It is now possible for humans to combine the genes of widely different organisms into one organism, giving this recombinant organism the means to express novel characteristics. Such “novel characteristics” include resistance to diseases and pests, resistance to pesticides, or the means to produce products that are of commercial or medical value. Whereas these organisms are of potential value to society, it is necessary to evaluate their environmental impact before they can be safely released into the environment. Therefore, with the advance of biological science comes the responsibility for biotechnologists to ensure that the integrity of the environment is maintained by ensuring that the biological functions that allow life on this planet to thrive are not impaired. Interestingly, recombinant technology has also opened the way to construct organisms that can be used to monitor the environmental impact of toxicants more sensitively. An important interface of ecotoxicology with biotechnology arises in the use of such genetically modified organisms as biomonitoring tools. However, the detection (and assay) of ecotoxics can be achieved by a variety of methods. An interesting example is the detection of carbon monoxide. This can be achieved using a caged canary (that responds sensitively to carbon monoxide in coal mines) or using an enzyme-based detector electrode that functions as a biosensor for this gas. A detailed account of biotechnological techniques for biomonitoring and bioremediation in relation to ecotoxicology can be found in the book by Lynch and Wiseman (1998).

E. Ecology

Ecology is the study of the interactions between organisms and their environment that determine the distribution and abundance of organisms. These interactions can be studied at different levels of biological integration starting from individuals, which are part of populations, which are part of communities, which are part of ecosystems, which are part of the earth’s biosphere. To gain a living from nature, humans have to understand the interactions that determine the abundance and distributions of organisms on which human life depends. This is not only important for the harvest of species from natural ecosystems by hunting, fishing, and or gathering of plant products, but also where the aim is to modify the natural environment to yield predominantly products that are of economical value as is the case in agriculture and forestry. Not only do we need to understand how we can most efficiently use the environment, but also increasingly we need to be aware of the impact of human activities on ecosystem function in general. Pollution of the environment by human activities (manufacturing, mining, waste disposal, transport, energy use, pesticide use, etc.) can have a large impact on ecosystem function via effects at all levels of biological integration. Especially since the scale of human activities has increased so dramatically over the past century there is a real danger that human activity interferes significantly with the basic biological processes of sustained life, including our own. Therefore, ecological knowledge applied to ecotoxicology aims to evaluate the consequences of human activity to ecosystem function, especially in relation to pollution caused by human activity.

F. Toxicology

Important principles of toxicology that relate to ecotoxicology include the concepts of extent of exposure, persistence, and distribution of chemicals in the environment. Subsequently predictions can be made on the toxicity of such chemicals to individual organisms or populations. The starting point of such analysis is often the chemical structure of each toxin, which will allow some prediction of the behavior of the chemical in the environment to be made. One of the best examples of this kind is the environmental impact assessment of the insecticide DDT. DDT is almost completely insoluble in water but readily soluble in fat. Furthermore, DDT and its degradation product DDE are highly persistent. Its low solubility means that it is easily dispersed in aquatic environments (in the case of DDT it is found at low
concentrations all over the globe) and only accumulates in places with a high fat content (i.e., living organisms). Once it has entered living organisms, it persists and accumulates in the fat tissues of organisms that are higher up the food chain resulting in toxicity. For this reason, populations of both fish eating birds and bird eating raptors such as peregrines and hawks were badly affected by DDT, even in cases where they inhabited pristine environments.

Another aspect of toxicology is risk analysis. This quantitative topic is most easily studied when death rates within a population can be quoted. Whereas the rates can be determined experimentally for animals, plants, and microorganisms, the determination of the effects on the human population is inevitably more difficult to determine, and one of the few situations where effects are well known is the mutagenic effects of radionucleotides, resulting from contamination caused by accidents at nuclear power installations or nuclear warfare.

III. ENTRY, MOVEMENT, AND FATE OF POLLUTANTS IN ECOSYSTEMS

Organic and metallic (including radionucleide) toxicants that enter the environment are regarded as pollutants of air, soil, and water. Their movement and transport is through air and water depending on their volatility and solubility, respectively. Soil and the detritus of sediments provide important solid supports for adsorption to regulate the movement and flow of the pollutants through terrestrial ecosystems, as well as influencing the localization and persistence of the pollutants in the environment. In this respect clay and humus particles that are electrostatically charged have an important role to play. Some organisms will take up the pollutant and concentrate them in their cells, a phenomenon known as bioaccumulation or biomagnification. This is particularly serious where the toxicant enters the food chain leading to progressive accumulation of toxic molecules higher up the chain, including humans. This is often the case with molecules that are relatively biologically inert such as heavy metals, PCBs, and organochlorine insecticides. Accumulation of such molecules in different species can lead to toxicity expressed as reduced growth, reduced fecundity, changes in behavior, susceptibility to diseases, or increased mortality. However, on the positive side, where there is storage of the pollutant in the cell, as is normally the situation with (heavy) metals, it is possible that the pollutant might be harvested and therefore removed from the environment. This especially applies to plants that hyper accumulate metals in their tissues. The application of this process for the cleanup of contaminated land is known of phytoremediation. Also, many pollutants are biologically degradable and can therefore be detoxified. Both plants and mammals have a range of enzymes that are involved in detoxification of molecules that are potentially harmful to them (cytochromes P-450, for example), whereas the almost limitless metabolic capacity of a wide variety of microorganisms allows degradation of pollutants, especially hydrocarbons, into nontoxic molecules such as water and carbon dioxide. The stimulation of microbial degradation and metabolism of pollutants is known as bioremediation and is currently used to clean up contaminated land while microbial degradation of pollutants present in sewage and water is used in a variety of water purification systems. The fate of pollutants in the environment is therefore not only dependent on the molecular characteristics of the pollutant but also on a range of biotic and abiotic factors.

IV. BIOMONITORING

The ultimate goal of biomonitoring is to use biological effects resulting from chemical exposure for making predictions and deductions about the quality of the environment for life in general and human life in particular. Biomonitoring activities can be conducted on various levels of biological integration, over a period of time. Several overlapping terms are commonly used in this process. It is worth noting that some of the tools are, strictly speaking, alternatives for chemical/physical methods to measure bioavailability of chemicals in the environment. In other words, they do not provide information on ecological effects per se.

• The term bioreporter is used for a molecular tool, often genetically modified cells, to transform the presence of a chemical into an easy measurable signal (e.g., luminescence).
• The measurement of biochemical and physiological variables in individuals or in their excretion products, providing information on exposure or damage, is indicated by the term biomarker.
• A biosensor or bioprobe is a physical device that allows the detection of a chemical as an electrical signal derived from a biocatalyst, such as an enzyme or an antibody.
• A bioassay is a toxicological test system in which
TABLE I
Some Tests of Chemicals That Affect Soil Function

<table>
<thead>
<tr>
<th>Retention capacity of soil and endangering of groundwater</th>
<th>Living space for plant production</th>
<th>Living space for soil communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth (algae) Daphnia</td>
<td>Dehydrogenase (Bacillus cereus)</td>
<td>Biomass (microorganisms)</td>
</tr>
<tr>
<td>Immobilization (Daphnia)</td>
<td>Nodulation (Rhizobium)</td>
<td>Enzyme activity (microorganisms)</td>
</tr>
<tr>
<td>Light emission (Photobacterium phosphoreum)</td>
<td>Shoot or root growth (various plants)</td>
<td>Nitrification (bacteria)</td>
</tr>
<tr>
<td>Mortality (fish, nematodes)</td>
<td>Yield (various crops)</td>
<td>Mortality (earthworms)</td>
</tr>
<tr>
<td>Mutagenecy (Ames, Umu)</td>
<td></td>
<td>Reproduction (collembola, earthworms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Respiration (microorganisms)</td>
</tr>
</tbody>
</table>

the activity of a chemical is measured as an adverse effect on a population of a specific test species. A bioindicator is used to detect the effects of chemicals at the individual, population, community, or even ecosystem level.

Important in this respect is that when we move up to the higher levels of biological integration (populations, communities, ecosystems), effects are increasingly difficult to link to a specific chemical, even though the effects measured are of more importance. Furthermore, it takes an increasingly long time for higher levels of integration to respond to a chemical insult. For example, DDT took years to bioaccumulate up the food chain resulting in negative effects at the population/community level. On the other hand, chemicals that are apparently nontoxic at the lower levels of biological integration (carbon dioxide, chlorofluorocarbons (CFCs), sulfur dioxide, etc.) can have profound effects on all levels of biological integration. These chemicals might not directly affect biological processes but affect the chemical balance of the earth's biosphere (air, water, and soil). For example, greenhouse gases (carbon dioxide, methane, and water vapor) trap the sun's long wave radiation augmenting global warming, which could lead to profound changes of the world's ecosystems. CFCs contribute to the destruction of the earth's protective ozone layer, leading to an increase in harmful UV radiation that reaches the earth, resulting in increased lethal mutation rates and cancer. Sulfur dioxide and NOx emissions lead to acid rain and acidification of soil and water, which in turn leads to the release of toxic quantities of metal ions in soil and water that kill both vegetation and animal life. An interesting example of this kind is the acidification of lakes in Scandinavian countries as a result of acid rain. The resulting lower pH of the lake water means that the water can contain a higher concentration of aluminum ions. However, the pH around the gills of fish is slightly alkaline, resulting in the precipitation of aluminum on the gills. This results in the impairment of gill function and suffocation of the fish. Even though Scandinavian lakes affected in this way look perfectly clean, they are devoid of fish.

A variety of specific tests have been employed to investigate the effects of chemicals on the environment. Generally the bionic effects used for biomonitoring deploy biota from the environment and expose them to different concentrations of the test chemical. Mortality of fishes, for example, has been used to assess toxicity of chemicals that are used in the soil environment (Table I). In the soil environments a variety of effects can be investigated as to whether the retention capacity of soil would affect the toxicity of a chemical leading to leaching into groundwater or the living space for plant production and soil communities. In this respect, the assays relate to the effects on populations, gene products (enzymes and metabolites), or impacts (plant bioassays) (Fig. 1). However, the exciting development of...
recent years is the recognition that these effects are all controlled by genes and another method of biomonitoring is to use gene (DNA) probes (Fig. 2). This option is very much in its early stages and the concept is to link a gene, which responds to a specific chemical or general effect, to a promoter DNA sequence and a reporter DNA sequence that either encodes for an enzymatic effect or any other specific signal. It is therefore possible to use genes, which respond to mutagenic substances with a light signal that is proportional to its mutagenic potential in the environment. Whereas this approach offers much promise, the reality is that it is complementary to the other methods currently deployed and should be used in conjunction with them.

A particular useful activity, which has been monitored in aquatic environments, is cytochrome P450. (Table II). This is the terminal component of the mixed-function oxygenase, which catalyses exopoxidation, hydroxylation, dealkylation, and desulfuration, all critical in detoxification of toxic molecules in invertebrates and vertebrates (including humans). To date, more than 750 cytochrome P450 genes have been identified (Nelson, 1998). The cytochrome P450/1A (CYP1A) isozyme is a biotransformation enzyme involved in a wide range of xenobiotic metabolisms (chemicals foreign to the natural environment) including metabolism of polycyclic aromatic hydrocarbons (PAHs). Usually CYP1A has been used as a biomarker in fish and mollusks. The linking of this enzymatic function with its genetic control offers a great deal of potential for the biomonitoring of toxicology of both marine and freshwater ecosystems.

### Table II

<table>
<thead>
<tr>
<th>Fish species</th>
<th>Point sources of pollution</th>
<th>Diffuse areas of pollution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic tomcod (Microgadus tomcod), cod (Gadus morhua), Large-mouth bass (Micropterus salmoides), Rainbow trout (Oncorhynchus mykiss), Whitefish (Coregonus lavaratus)</td>
<td>Chemical plants, incineration plants, industrial complexes, landfill sites, oil industry, oil spills, natural oil seeps, pulp mills, sewage</td>
<td>Bays, bights, estuaries, fjords, gilfs, harbors, lakes, offshore areas, river systems</td>
</tr>
</tbody>
</table>

### V. BIOLOGICAL PROCESSES

The question can be raised why it is necessary to analyze complicated biological processes if the chemical in question can be measured directly using chemical or physical methodologies. There are at least three situations where biomonitoring provides answers that cannot be arrived at using direct chemical measurements. The first use for biomonitoring is in situations where one wants to trace the chemical history of an environment. For example, the pH history of lakes can be deduced from the percentage acidophilous and alkalophilous species within the diatom communities preserved in sediment cores (Renberg and Hellberg, 1982). Similarly, pollen records preserved in soil provide information on past vegetation from which the past environmental conditions can be deduced. The second use of biomonitoring is a situation where the activity of a chemical is short-lived but the effects are persistent. For example, pesticides that disperse rapidly or are broken down rapidly cannot be measured easily, but their biological effects might persist for long periods. The insecticide deltamethrin is active only for a few hours after spraying, while effects on sensitive ground spiders and beetle fauna remain visible for weeks after spraying (Evers, et al., 1989; Jagers op Akkerhuis, 1993). The third reason, and probably the most important one, for using biomonitoring is that effects of chemicals are dependent on the interactions that take place within the environment itself after their release. These interactions can be of a physical/chemical nature where...
the chemical is immobilized on to soil particles and is therefore not bioavailable. For example, the toxicity of metals, such as copper and zinc, in soil is determined by their bioavailability in soil. Both copper and zinc ions bind to soil minerals (mainly clay particles) and humic substances and as such are unavailable to the biota present in soil. Soils contain large quantities of these potential binding sites and therefore toxic effects of heavy metals are often buffered, even at relatively high concentrations. Because of this soil-buffering capacity, it usually takes some time of sustained metal inputs before metal ions become bioavailable at concentrations that are toxic. Both the mineral and organic fractions of soil interact with metal ions, providing negatively charged surfaces on to which cations can bind. Such adsorption displaces the cation that was previously counter-balancing the colloid's negative charge, hence the ability of colloids to adsorb ions in this way is known as the soil's cation exchange capacity (CEC). Heavy metals may also specifically adsorb on to hydrated oxides of aluminium, iron, and manganese by the formation of partly covalent bonds and can be chelated by solid-phase humic substances or by other organic ligands. The adsorption of metal ions is also affected by the soil pH. Acidification decreases the CEC of humic substances when carboxyl and ammonium groups will become protonated, reducing their capacity to bind metal ions. The increase in hydrogen ions will also increase the competition for negative binding sites, leading to the displacement of metal ions. Therefore, the size distribution of mineral particles, the amount of organic matter and the soil pH will to a large degree determine the metal buffering capacity of the soil (Gupta, 1991).

Besides physical/chemical interactions, chemicals interact with the biota present in the soil. For example, sustained use of specific pesticides leads to the development of microbial communities that degrade the pesticide, reducing their half-life time and therefore their environmental impact. On the other hand, chemicals such as PCBs and organochlorides are likely to accumulate up the food chain resulting in high concentrations of residues distant from the point of release, thus increasing their environmental impact. As biological systems are not just collections of independent organisms but are strongly dependent on each other for their survival, it follows that chemical effects on one species can have important consequences for species that depend on the affected species. An example of this type of indirect effect is the decline of the grey partridge (Perdix perdix) in West Sussex, England. It could be shown that 48% of chick mortality was explained by the density of preferred insects. Given that reduced availability of insects is the key to the partridge decline, it was postulated that this might be due to increased pesticide usage, especially since there was a strong correlation between the partridge decline and increased herbicide use. The effect of the herbicides was shown to be indirect (they are relatively nontoxic to partridges), removing the host plants on which the insect larvae fed that were the food source of the partridge chicks, resulting in less insects and therefore less food for the partridge chicks. In conclusion, because biological effects are dependent on the behavior (movement, persistence, bioavailability) of the chemical in the environment, the sensitivity of biota and the role these biota play in ecosystem function, biomonitoring is often the only available way for monitoring chemical effects.

VI. EARLY WARNING SYSTEMS

Life as a whole can be seen as an hierarchical systems with distinct levels of biological integration. Going from a low level of organization to a high level of organization we can recognize cells that make up tissues, that make up organs, that are part of an organism; these organisms are arranged in populations that form part of communities, that are ultimately part of ecosystems. In general, it can be said that biological systems are at each level of organization buffered to resist change in overall performance by adjusting the components that make up that level of integration. The result of these buffering effects is that effects at lower levels of integration are dampened so that they do not affect higher levels of integration. Only when the buffering capacity at a certain level of integration is exceeded will the next level of integration be affected. This implies that toxic effects can be measured at the lower levels of biological integration before they result in effects at higher levels of integration. Because the aim of biomonitoring is “to provide a means by which (impending) changes in ecosystem function can be detected,” compensatory effects at lower levels of integration can provide an early warning system of adverse chemical effects. A good example of such an early warning system is the thinning of bird eggshells as a result of DDE (one of the metabolites of DDT). DDE affects Ca deposition in eggshells (physiological change). However, only when eggshell thinning is in excess of 16% to 18% will this lead to a decrease in breeding success of the birds affected. Eggshell thinning is therefore an early warning system for population change. Not all bird populations are likely to be affected as DDE accumulates up the food chain. Only birds that...
ALA dehydrogenase is only induced by lead, while in-
specific stimulus. For example, inhibition of the enzyme
is meant that a biological effect can be related to a
forms of (chemical) stress. With the term
ent degrees of specificity and sensitivity to different
Within each level of biological integration the biological
are classified shows that we have still a long way to go.
Simple fact that only a fraction of the species that make up “biological life”
are at the top of the food chain (raptors and fish-eating
birds) will be affected in the first instance. However it
can be assumed that if DDT use had continued, the
decline in raptors and fish-eating birds would eventu-
ally have resulted in community changes.

What is important in this context is to recognize
which factors determine the buffering capacity at each
level of biological integration. These factors can be used
profitably in biomonitoring. It is, however, important
to recognize that if we are interested in impending
changes at the community level, it is probably most
sensible to look for changes that are occurring at the
level below (i.e. the population level). Similarly, im-
pending population effects are most profitably investi-
gated using effects on individuals. However, changes
at the organism level will be too sensitive for making
reliable predictions at community level and above. This
simple fact is often overlooked in biomonitoring where
current molecular technology has allowed precise inves-
tigation at the lowest levels of integration. These tech-
niques allow the development of sensing systems that
pick up changes that occur at the molecular and physio-
logical level. Whereas these approaches are valuable for
medical purposes in which we are interested in warning
systems that help prevent problems at the individual
level, they are inappropriate for monitoring ecosystem
changes, in that they are far too sensitive. In choosing
the most appropriate monitoring tool, one has to evalu-
ate its relevance, reliability, robustness, responsiveness,
and reproducibility. These factors, known as the 5 Rs,
will differ according to level of biological integration
and the composition of biological units that make up
that level. It is unlikely that an environmental monitor-
ing system will be “a freeze dried, talking bug on a
stick” (van Straalen, 1998). As interactions that govern
processes at each level of biological integration become
apparent, new and more appropriate biomonitoring
tools will become available. The simple fact that only
VII. SPECIFICITY AND RESOLUTION
Within each level of biological integration the biological
units that make up the level of integration display differ-
ent degrees of specificity and sensitivity to different
forms of (chemical) stress. With the term specificity it
is meant that a biological effect can be related to a
specific stimulus. For example, inhibition of the enzyme
ALA dehydrogenase is only induced by lead, while in-
duction of an immune response might be triggered by
a wide variety of causes. The term resolution is used to
indicate the ability of a bioindicator to respond to small
environmental changes. For example, a specialized
predator at the top of a food chain is a more sensitive
indicator of bioaccumulation of persistent pesticides
than omnivores with a varied diet. Although it seems
that sensitive indicators with a great resolving power
hold the greatest promise for biomonitoring, these
properties are likely to result in overcaution and false
alarms. In general, it can be said that indicators at lower
levels of biological integration are more sensitive than
characteristics that relate to higher levels of biological
integration. Therefore, indicator species are more sensi-
tive than diversity indices, for example. What makes a
“good” bioindicator will depend on the information that
is required. Looking at a “soil ecosystem” for example,
the biological components that make up the system
consist of plants, vertebrates, invertebrates (insects,
earthworms, nematodes, mites and springtails), proto-
zoa, bacteria, and fungi, all of which perform crucial
roles in the maintenance of soil ecosystem function.
When the system is exposed to a potentially toxic chem-
ical the ability of each population to cope with this
exposure will depend on genotypic plasticity within
the population, phenotypic plasticity, mutation rate,
metabolic capacity, and level of exposure. Genotypic
plasticity will be high in microorganisms that multiply
fast and can acquire resistance genes via mutations or
by acquisition of plasmids carrying resistance genes.
As a result, adaptation of microbial communities to
chemical stress is likely to be high. Furthermore, be-
cause of the almost limitless metabolic diversity dis-
played by microorganisms, populations that contribute
to the detoxification of a toxic chemical might increase
rapidly (comparable with the activation of the immune
system as a response to foreign molecules/organisms
that enter a higher organism). Other organisms (plants,
vertebrates, and invertebrates) that have a large pheno-
typic plasticity might respond to a chemical insult by
avoiding those areas that are most contaminated. Fur-
thermore, spatial distribution of the biota in the soil
environment and their mobility will both determine
exposure and therefore toxicity. Only when there is a
clear understanding of “normal” distribution and diver-
sity of the biota in a given ecosystem is it possible to
relate biological patterns to the effects of a toxicant in
a meaningful way.

Because of the biological complexity of ecosystems,
efforts have been made to classify organisms not ac-
cording to species but according to similarities in life
strategy. Biological life can be roughly classified by the
way organisms spend resources on their offspring. Or-

ECOTOXICOLOGY
371
organisms that invest in large numbers of offspring (giving each limited resources) are termed r-strategists, while those species that invest in few offspring (each of which is provided with ample resources) are termed K-strategists. Looking at the spectrum of life, there is a general trend that large organisms are K-strategists, while small ones are r-strategists. Although this is true to a large extent, within a certain body class size there are organisms that have adopted r- or K-type survival strategies. For example, mice are typical r-strategists (they have many large litters each year), while bats produce every two years a single offspring, making them typical K-strategists. Not surprising, K-strategists are characteristic for stable, climax environments while r-strategists are characteristic for unstable pioneer environments. In terms of “stress,” K-strategists are adapted to cope with biological stress (competition, predation, etc.), while r-strategists are adapted to cope with physical/chemical stress but are not good at coping with biological stress. In general, it can therefore be stated that r-strategists are insensitive to chemical stress, while K-strategists are sensitive. When an ecosystem is challenged by toxins (pesticides, for example), the r-strategists (the pest organisms) are the first to develop pesticide resistance, while the K-strategists (nonpest species) will disappear. The reason for this might be found in the fact that large numbers of offspring provide large genetic variability leading to rapid selection of the best-adapted individuals. Chemical/physical stress on an ecosystem will therefore lead to a shift in the balance between r- and K-strategists in favor of r-strategists. This principle has been used successfully to monitor environmental pollution and disturbance using colonizer/persister (c-p) indices applied to nematodes (Bongers, 1990) or perturbations of rhizosphere microbial communities (De Leij et al., 1993).

Another useful way of looking at communities and ecosystems is by classifying the organisms that inhabit that ecosystem by their “trophic function.” Using this approach, organisms can be classified (aboveground) as primary producers (plants and algae), herbivores, predators, and top predators, while organisms that feed on a variety of food sources are classified as omnivores. In general, food chains have no more than four trophic levels because less and less energy is available higher up the food chain. Furthermore, dependence of predators and top predators on exclusive feeding sources make them vulnerable to perturbations lower down the food chain. Using the concept of trophic levels, it can be argued that specialized feeders are more sensitive to environmental stress (when this stress affects their food source) than organisms that are less specialized. Not surprisingly, when recalcitrant chemicals enter the food chain it is the specialized feeders (for example, peregrines and sparrow hawks that feed exclusively on birds) that are the first to be affected by toxins. Beside the fact that they are exposed to relatively high levels of a pollutant, these organisms are often very skillful specialized hunters. Any impairment of these skills due to a toxin will result in starvation. Therefore, these species can be sensitive bioindicators of toxins, even at sublethal concentrations.

Another important ecological concept that can be used to simplify ecosystems in a meaningful way is the concept of keystone species. The idea is that even though there is dependency between all species that are involved in a certain food web, some species are more important than others. It is important in this context to make a distinction between dominant species that derive their importance to the ecosystem in terms of biomass or energy flow and keystone species. Keystone species are often relatively rare but have a major impact because of their key regulatory function in the system as a whole. This is because these species are involved in more links or interact with parts of the ecosystem that maintain living conditions for a wide variety of other species. In other words, the species diversity of an ecosystem is determined in large part by a few keystone species on which many others depend. A detrimental effect on these species will result in the collapse of the whole ecosystem (or at least a large part thereof). Several different categories of keystone species have been recognized, such as keystone predators, keystone prey, and keystone habitat modifiers. One of the best-studied examples of the importance of keystone predators is the starfish (Pisaster) and predatory whelks (Nacella) in pools along the rocky intertidal zone. Removal of these predators led to the disappearance of 80% of the prey species and the nearly complete dominance of the prey community by mussels (Mytilus). Sea otters (Enhydra lutris) have also been labeled as keystone species as they limit the density of sea urchins, which, in turn, eat kelp and other macro algae that provide the habitat for a large variety of species. Without sea otters the sea urchin population explodes resulting in the destruction of the kelp forests leading to the loss of habitat for those species that depend for their survival on the kelp forests for reproduction, shelter, and food. The way keystone species are affected by toxins is an important consideration in ecotoxicology, because they represent the Achilles’ heel of an ecosystem. Not surprisingly, keystone species such as earthworms and honeybees are often used in toxicity testing because they are important for ecosystem function in general.
Clearly a negative effect on ecosystem function is often due to alterations of complex interactions within the ecosystem itself. In most cases, toxic effects are buffered due to the compensating capacity of an ecosystem. However, occasionally toxic effects can affect species that are responsible for essential functions in that ecosystem. Further insight in the way that different species contribute to ecosystem function is an essential requirement for the development of the most appropriate biomonitoring tools for assessing the ecotoxicological effects of chemicals.

VIII. CONCLUSION

Currently there is growing concern internationally about environmental toxicants affecting biodiversity and human health. Biodiversity has come to the fore by the United Nations meetings; the first of which was in Rio de Janeiro in 1992. The medical profession is increasingly concerned about a range of conditions that might be induced by chemicals in the environment. This is resulting in moves to increase legislation. The study of ecotoxicology seems therefore certain to grow in importance, as well as there being a stimulus to prevent problems occurring in the first place via the development of clean technology and stimulation of sustainable land use. On the other hand, ecotoxicology can help to identify existing pollution problems, which in many cases can be cleaned up using a variety of remediation strategies.

See Also the Following Articles

ACID RAIN AND DEPOSITION • AIR POLLUTION • ECOLOGICAL FOOTPRINT, CONCEPT AND MEASUREMENT OF • GREENHOUSE EFFECT • KEYSTONE SPECIES • SOIL BIOTA, SYSTEMS AND PROCESSES

Bibliography

EDIBLE PLANTS

I. Food Habits
II. The Diversity of Food Plants
III. Geographic Patterns of Food Plants
IV. Are Edible Weeds, Weeds?

GLOSSARY

cultivar Cultivated variety or genetic strain of a domesticated food plant.
domesticate Plant that has been selected by humans and adapted for use as a food crop, nutrient, fiber, or other purpose.
ethnobotany Study of the variety, natural history, and characteristics of the plants used by human cultures.

THE VARIETY OF PLANTS NORMALLY CONSUMED BY HUMANS represents a tiny proportion of what nature supplies. Of the estimated 270,000 plant species recognized in the world, about 12,500 species from over 3000 genera are considered edible to humans.

I. FOOD HABITS

Animals, being terrestrial or marine, have limitations with respect to the variety of species they consume. They depend on what nature offers in the place where each animal lives. Since the geographic ranges of species are relatively widespread, the individuals of a given animal species may vary their food resources in different sites of the range, according to their opportunities. For this reason, the variety of plants ingested by a given species is always greater than the variety of plants ingested by an individual or by an entire population.

Domestic as well as wild herbivores may eat a wide variety of food items, but they have clear preferences for particular plant species. Some ungulates are capable of selecting the most nutritious individual plants among those of the same species. They select forage mainly by smell, and secondarily by taste. Volatile substances in the plants may either inhibit or attract foragers, and are largely responsible for their palatability. According to Klein (1970), the nutritive value of plants available to wild ruminants depends on the stage of maturity of the vegetation, with highest nutritive quality coinciding with the initiation of growth, as well as with soil type and climate. Rapid growth in plants is correlated with high nutritive quality. Similar conclusions were reported by Gardarsson and Moss (1970) in a study of food selection by the Icelandic ptarmigan (Lagopus mutus). This bird consumed leaves and flowers of 8, 11, 10, and 11 species, respectively, in summer, autumn, winter, and spring.

Similarly, the European hare (Lepus europaeus) in northwestern Patagonia makes use of 17 to 21 species of plants in a given season out of a total of 26 species that it consumes over the year. Table I provides a short sample of the variety of food consumed by different
TABLE I
The Number of Food Plants Consumed Annually by Different Mammal Species

<table>
<thead>
<tr>
<th>Mammal species</th>
<th>Common name</th>
<th>No. of plant species</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alouatta fusca</td>
<td>Brown howler monkey, SE Brazil</td>
<td>52</td>
<td>Galetti et al. (1994)</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>Cattle, NW Patagonia</td>
<td>23</td>
<td>Relva (1998)</td>
</tr>
<tr>
<td>Capra hircus</td>
<td>Goat, Mendoza, Argentina</td>
<td>76</td>
<td>Dalmasso et al. (1995)</td>
</tr>
<tr>
<td>Cebus apella</td>
<td>Capuchin monkey, SE Brazil</td>
<td>73</td>
<td>Galetti and Pedroni (1994)</td>
</tr>
<tr>
<td>Cervus elaphus</td>
<td>European red deer, NW Patagonia</td>
<td>34</td>
<td>Brez (1981)</td>
</tr>
<tr>
<td>Chironomus satanas</td>
<td>Monkey, Venezuela</td>
<td>29</td>
<td>Kimey and Norstrom (1993)</td>
</tr>
<tr>
<td>Conomys mendoza</td>
<td>Tucuo-tucuo, Argentina</td>
<td>28</td>
<td>Madors (1993)</td>
</tr>
<tr>
<td>Lagidium viscacia</td>
<td>Vizcacha de la sierra, Argentina</td>
<td>21</td>
<td>Galdene and Grigera (1999)</td>
</tr>
<tr>
<td>Lama guanicoe</td>
<td>Guanaco, Mendoza, Argentina</td>
<td>47</td>
<td>Camila and Dalmasso (1993)</td>
</tr>
<tr>
<td>Lepus europaeus</td>
<td>European hare, NW Patagonia</td>
<td>28</td>
<td>Galdene and Grigera (1999)</td>
</tr>
<tr>
<td>Pithusa pithus</td>
<td>Monkey, Venezuela</td>
<td>25</td>
<td>Kimey and Norstrom (1993)</td>
</tr>
</tbody>
</table>

herbivorous and omnivorous mammals. The case of the Soay sheep was included to show how a domestic herbivore may restrict its diet in a species-poor environment. Of course, the range of foods ingested by other animals may vary widely. Monophagous insects restrict themselves to only one plant species, soil amoebae (Acanthamoeba) normally ingest five species of microscopic algae (Heal and Felton, 1970), and polyphagous, pest arthropods may feed on more than 300 species of crops and wild plants. The number of food plants eaten by humans is not far from the figures shown in Table I.

Domestic and wild ungulates may show a copious range of food species but they focus on a few, preferred plants. A number of less palatable species are used only in time of food scarcity. Studies performed in western Argentina by M. B. Kufner and S. Monge showed that the rodent Lagostomus maximus increases the variety of its food sources in degraded habitats and during droughts.

A. Standards of Consumption

On an individual basis, people use a small number of plant items per day, perhaps between 10 and 20 species or products. A normal diet includes common vegetables, fruits, seeds (in the form of flour or oil), roots, sugar, beverages (beer, wine, colas), condiments, teas, and herbas. The number of plant species normally used by an individual over a year, however, is about 100, although this is limited by the number of edible plant items commonly offered by popular markets and supermarkets. Taking into account all the variety of greens, vegetables, fruits, grains, nuts, and condiments in an exceptionally well-provisioned supermarket, the figure (including different varieties and brands) may rise to 600 according to Duke (1992).

The stomach contents of two mummies that were found well-preserved in Danish bogs provided interesting information about the gastronomic habits of people during the Iron Age. Their last meals contained 66 different plant taxa (Godwin, 1960; King, 1966), many of which are nowadays considered as cosmopolitan weeds. Before the invention of agriculture, in the Paleolithic Age, humans were hunter-gatherers, and probably had a better knowledge of the variety of edible wild plants than modern people. This knowledge, however, has slowly been lost since the Neolithic and, in present times, is still lost after one or two generations of acculturation in aboriginal communities (Plotkin, 1993). The process of “civilization” goes hand in hand with the loss of knowledge, as well as with the abandonment of traditional crop varieties and the habit of gathering wild plants. But at the same time, new cultivars, coming from distant countries, are constantly increasing the variety of foods. A recent case is that of the kiwi fruit (Actinidia delicosa), which originated in China and then was renamed, cultivated, and popularized in New Zealand. The roots of the ahipa (Pachyrhizus ahipa) and jicama (P. erosus), of Central American origin, have become increasingly popular in the United States and Southeast Asia, especially the latter species, which has a similar texture and flavor to the bamboo shoots and is used by Asian food restaurants in Western countries (National Research Council, 1989; Brücher, 1989). Rice
(Oryza sativa), of Asian origin, is at present the most popular staple in warm countries of South America, while the South American potato (Solanum tuberosum) has become "mandatory" in the European cuisine.

II. THE DIVERSITY OF FOOD PLANTS

No one has compiled a complete record of edible plants for the entire world. The Food and Agriculture Organization (FAO), part of the United Nations, publishes an annual report of the production of the commercially most important foods. This list includes about a hundred species of plants. In The Oxford Book of Food Plants (Nicholson et al., 1969), the number increases to 389 species distributed among 81 plant families. These are both locally and widely known cultivated plants. Duke (1992) estimated that North American Indians ate 1112 plant species. This figure is set at 1886 species according to Moerman (1998). More than 3000 edible species are carefully listed and commented on in the voluminous book Cornucopia, compiled by Facciola (1990), but in its preface Noel D. Vietmeyer suggests that there are about 20,000 edible species across the world. Probably, the most complete inventory is Kunkel's (1984) book, which lists roughly 12,560 species from 3100 genera belonging to about 400 families of flowering plants and ferns. This list, however, is being constantly enriched by the contribution of many ethnobotanical studies.

The proportion of edible plant species in slightly disturbed communities is variable. In the Sonoran Desert it is about 15%. Oma Indians from Tierra del Fuego made use of at least 6% of this island flora, whereas the Cháechebo Indians, in the Bolivian Amazon, use 21% of their surrounding flora. Medium to highly disturbed communities may contain similar or higher proportions of edible species for human consumption. For example, in western Uruguay the proportion is 17%, in southwestern Córdoba province (Argentina) it is 19%, in the outskirts of Havana (Cuba) it is 33%, in swidden (slash-and-burn) fields of northern India it is 43%, and in experimental fields in Saskatchewan (Canada) it may reach 63% of all wild plants. Yet studies like these do not necessarily reveal the actual possibilities offered by nature, but rather the knowledge of informants and/or the perspicacity, experience, and field-work time employed by the investigators. According to the estimate that about 10% of any flora represents food resources, then 10%, or 27,000, of the 270,000 species of plants already recognized by world botanists should be edible.

Since historical times, because of written testimony, Europe has conserved people's knowledge of gastronomic matters. From the botanical point of view, the United Kingdom is probably the best-known country in the world. If we compare the floristic list compiled by Martin in 1976 with Kunkel's list of food plants, and discard the exotic species, hybrids, and other sub-specific taxa, as well as plants used only during famine times, we can verify that out of 1503 species considered, 350 are edible. In other words, 23% of the British flora is edible. Thus, we have two estimates of the possible richness of edible vascular plants—10 and 23%—and they represent between 27,000 and 62,000 species, respectively, based on the 270,000 known at present. Because the description of the world flora has not been completed yet, the final list of comestibles will probably increase in the future.

By comparison, less than 2% of the Central American flora is edible, based on the list prepared by Duke (1992). Possible explanations for this remarkable difference are: (1) greater taxonomic ignorance and/or less exploration of the natural resources—because the flora of Central America is much richer than the flora of Britain, humans may have concentrated on fewer, more abundant and profitable plants, and disregarded the less useful ones; and (2) widespread loss of cultural heritage and environmental knowledge following the conquest and colonization by European countries.

A. The Most Prolific Taxa

A first, rough estimation at higher taxonomic ranks indicates that the proportions of edible species are quite similar to the proportions of "common" (edible and nonedible species) species present in the plant kingdom. The right-hand column in Table II is based on a random sample of 1790 food plants appearing in Kunkel (1984). If the property of being edible or palatable were

<table>
<thead>
<tr>
<th>Taxonomic group</th>
<th>Common species (n = 270,000)</th>
<th>Edible species (n = 1790)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pteridophyta</td>
<td>3.9</td>
<td>1.9</td>
</tr>
<tr>
<td>Gymnospermae</td>
<td>0.3</td>
<td>1.0</td>
</tr>
<tr>
<td>Dicotyledoneae</td>
<td>69.9</td>
<td>75.5</td>
</tr>
<tr>
<td>Monocotyledoneae</td>
<td>25.9</td>
<td>21.6</td>
</tr>
</tbody>
</table>
randomly distributed among the different taxa, then it would be predictable that the most numerous families of plants would contain a higher number of edible species than from common species. Among families with the lowest ratios are the Asclepiadaceae (8.0), Cruciferae (8.6), and Rutaceae (10.0). Families that are richer in species per genus are the Begoniaceae (295.0), Aizoaceae (208.3), and Eriocaulaceae (92.3), yet they show no evidence of having experienced a process of selection, that is, of "proclivity" or "rejection" by humans, that would have led them to speciate toward palatability or distastefulness.

Of the 389 more frequently cultivated species considered by Nicholson et al. (1969), the ranking goes as follows: Rosaceae (13.3% of total species), Leguminosae (8.5%), Compositae (5.9%), Umbelliferae (5.4%), and Cruciferae (5.1%). They are followed by Palmae, Cucurbitaceae, Rutaceae, Alliaceae, Chenopodiaceae, Dioscoreaceae, Annonaceae, Ericaceae, Grossulariaceae, and 66 less prolific families. Kunkel (1984) states that the Rosaceae is the richest family among food plants. The analysis of a sample of 6222 items from his list confirms this assessment. Rosaceae appears at the top of the list, comprising 5.8% of the cases. The first 30 families is shown in Table IV.

### Edible Parts

Some plant genera are extremely abundant in edible species and may show particular tendencies toward a given kind of food (Table V). For example, all of the 205 species of *Rubus* appearing in Kunkel’s list provide edible fruits. Among them, there are three species whose leaves are also used as tea. Similarly, the 80 or more...
The Most Prolific Genera of Food Plants

<table>
<thead>
<tr>
<th>Genus</th>
<th>Family</th>
<th>No. of food species</th>
<th>Total No. of species</th>
<th>% of food species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Rubus</td>
<td>Rosaceae</td>
<td>205</td>
<td>2500–3000</td>
<td>7–8</td>
</tr>
<tr>
<td>2. Ficus</td>
<td>Moraceae</td>
<td>137</td>
<td>700</td>
<td>20</td>
</tr>
<tr>
<td>3. Dioscorea</td>
<td>Dioscoreaceae</td>
<td>130</td>
<td>600</td>
<td>18</td>
</tr>
<tr>
<td>4. Solanum</td>
<td>Solanaceae</td>
<td>100</td>
<td>1400–1700</td>
<td>6–7</td>
</tr>
<tr>
<td>5. Azadirachta</td>
<td>Meliaceae</td>
<td>80</td>
<td>800</td>
<td>10</td>
</tr>
<tr>
<td>6. Eugenia</td>
<td>Myrtaceae</td>
<td>79</td>
<td>800</td>
<td>10</td>
</tr>
<tr>
<td>7. Diospyros</td>
<td>Ebenaceae</td>
<td>69</td>
<td>200</td>
<td>35</td>
</tr>
<tr>
<td>8. Garcinia</td>
<td>Guttiferae</td>
<td>68</td>
<td>400</td>
<td>17</td>
</tr>
<tr>
<td>9. Quercus</td>
<td>Fagaceae</td>
<td>67</td>
<td>470–1000</td>
<td>7–14</td>
</tr>
<tr>
<td>10. Vaccinium</td>
<td>Ericaceae</td>
<td>66</td>
<td>300–400</td>
<td>17–22</td>
</tr>
<tr>
<td>11. Passiflora</td>
<td>Passifloraceae</td>
<td>58</td>
<td>300</td>
<td>12</td>
</tr>
<tr>
<td>12. Opatia</td>
<td>Cactaceae</td>
<td>52</td>
<td>230</td>
<td>21</td>
</tr>
</tbody>
</table>

species of Prunus provide edible fruits, as also occurs with Rosa spp. and Ribes spp. The majority of Piper species are used as black or white pepper or as a spice for curries. Rumex provides 44 species with leaves used as vegetables and 3 species with edible roots. Of the 100 edible species of Solanum, 59 are used only for their fruits, 20 species only for their tubers, 14 for both fruits and leaves, 6 species only for their leaves, and a single species exclusively for its seeds.

An analysis of the data compiled by Duke (1972) for Central America indicates that 85% of the plant species are used for a single purpose, 10% have two uses, 4% have three uses, and 1% have four uses (leaves, flowers, fruits, and seeds). A similar analysis of the 350 edible species of the native British flora yields slightly different proportions: 78%, 19%, <3%, and <1% for one, two, three, and four purposes, respectively. One of the exceptional cases of variability of uses is that of

---

**TABLE VI**

An Estimate of Food Usage (as Percentages of Regional Totals) from Different Sources

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>21.8</td>
<td>28.9</td>
<td>20.6</td>
<td>46.8</td>
<td>7.6</td>
<td>22.8</td>
</tr>
<tr>
<td>Fruits</td>
<td>10.6</td>
<td>30.9</td>
<td>37.4</td>
<td>9.5</td>
<td>6.1</td>
<td>26.0</td>
</tr>
<tr>
<td>Seeds</td>
<td>15.9</td>
<td>13.2</td>
<td>17.3</td>
<td>10.1</td>
<td>10.6</td>
<td>17.0</td>
</tr>
<tr>
<td>Condiments, flavorings</td>
<td>11.5</td>
<td>6.2</td>
<td>3.2</td>
<td>0.9</td>
<td>6.1</td>
<td>3.3</td>
</tr>
<tr>
<td>Tea, herbs</td>
<td>9.5</td>
<td>3.0</td>
<td>3.6</td>
<td>7.6</td>
<td>0.0</td>
<td>3.1</td>
</tr>
<tr>
<td>Beverages</td>
<td>7.0</td>
<td>0.8</td>
<td>6.0</td>
<td>0.0</td>
<td>2.3</td>
<td>1.7</td>
</tr>
<tr>
<td>Flowers, capsers</td>
<td>6.7</td>
<td>4.0</td>
<td>4.1</td>
<td>4.8</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Roots</td>
<td>5.6</td>
<td>8.7</td>
<td>6.3</td>
<td>12.6</td>
<td>12.1</td>
<td>7.5</td>
</tr>
<tr>
<td>Sap(s)</td>
<td>4.2</td>
<td>3.9</td>
<td>2.2</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Bark(s)</td>
<td>0.2</td>
<td>0.9</td>
<td>1.0</td>
<td>1.1</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>No. of species</td>
<td>3220</td>
<td>1710</td>
<td>182</td>
<td>350</td>
<td>96</td>
<td>186</td>
</tr>
<tr>
<td>No. of items</td>
<td>6311</td>
<td>2442</td>
<td>536</td>
<td>478</td>
<td>132</td>
<td>294</td>
</tr>
</tbody>
</table>

* A more detailed analysis of food usages appears in Table VIII.
* Including stems, sprouts, and meristems.
* Including bulbs and rhizomes.
* Including gums, latex, sugars, and masticatories.
hops (*Humulus lupulus*). Its leaves, roots, flowers, and bark may be used as food and as condiment. The pumpkin (*Cucurbita moschata*) provides fruits, seeds, flowers, young leaves, and shoots for human consumption.

It is clear from Table VI that some kinds of food are scarcely used, such as bark, flowers, sap, and liquorice. In contrast, leaves (including stems, sprouts, seedlings, and shoots) and fruits seem to be the most preferred food.

### III. GEOGRAPHIC PATTERNS OF FOOD PLANTS

On the basis of a sample of 1790 species from Kunkel's list of food plants, it appears that edible herbs are more numerous than trees, shrubs, and vines. "Vines" include all the climbing, creeping, and epiphytic plants. Table VII shows that in decreasing order of species richness, the Indomalayan regions appear first, followed by the Neotropical, Palearctic, Ethiopian, Nearctic, and Australian (or Australasian) regions. By means of a chi-square test, at a significance level $P < 0.05$, the Nearctic and Australian regions show greater, and the Palearctic lower, proportions of food trees than expected. The proportion of edible shrubs does not vary significantly among regions, although at a $P < 0.1$ the Neotropics seem to have a higher proportion than the rest of the world. The Palearctic region shows a higher and the Neotropics a lower proportion of herbs, while the latter region is richer in vines ($P < 0.05$). In contrast, the Northern Hemispheres (Paelearctic and Nearctic regions) shows a significantly lower proportion of vines than the Southern Hemisphere.

According to biogeographic regions (Table VIII), the Palearctic shows a significantly greater proportion of species that provide edible leaves, stems, and sprouts than the other regions. The Australian region is characterized by a greater proportion of edible seeds, and the Neotropics by its abundance of fruit species. Of course, these differences were derived from the sample analyzed by us. For instance, the appearance of zero values for beverages from the Australian and Palearctic regions does not mean that there are no species of this kind in their florae; they were simply not registered in our sample of 1790 species. On the contrary, the high proportion of edible fruits in the Neotropical region is repeatedly verified in ethnobotanical studies performed in different countries of Central and South America. These contrasts may be attributed to differential characteristics of seed dispersal evolution within their plant communities, as well as to prevailing cultural trends in the use of natural resources. The possibility of ethnobotanists showing biased attention toward particular kinds of food should not be discarded.

### IV. ARE EDIBLE WEEDS, WEEDS?

The habit of gathering wild food plants has not been totally lost. E. R. Spencer maintained that "any plant is a weed if it insists upon growing where the husbandman wants another plant to grow. It is a plant out of place in the eye of man; in the nice eye of nature it is very much in place." Many of our dearest crops, however, have originated from weeds. In the course of time, humans learned how to profit from them. Oats (*Avena sativa*), foxtail millet (*Setaria italica* derived from *S. viridis*), chicory (*Cichorium intybus*), pak choi (*Brassica rapa*), spinach beet (*Beta vulgaris*), and many more
TABLE VIII

A Sample of 1790 Species used in 2442 ways, Classified by Types of Food in Different Regions (Figures Are Percentages of Regional Totals)

<table>
<thead>
<tr>
<th></th>
<th>Nearctic</th>
<th>Neotropical</th>
<th>Indo-Malayan</th>
<th>Australian</th>
<th>Ethiopian</th>
<th>Palaearctic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>24.3</td>
<td>18.3</td>
<td>30.0</td>
<td>25.0</td>
<td>27.3</td>
<td>48.0</td>
</tr>
<tr>
<td>Fruits</td>
<td>28.3</td>
<td>32.3</td>
<td>31.9</td>
<td>14.1</td>
<td>13.7</td>
<td>14.4</td>
</tr>
<tr>
<td>Seeds</td>
<td>12.8</td>
<td>11.7</td>
<td>13.7</td>
<td>28.9</td>
<td>13.3</td>
<td>10.0</td>
</tr>
<tr>
<td>Condiments, flavorings</td>
<td>3.0</td>
<td>3.9</td>
<td>8.2</td>
<td>2.3</td>
<td>4.7</td>
<td>2.2</td>
</tr>
<tr>
<td>Tea, herbs</td>
<td>6.9</td>
<td>2.3</td>
<td>2.2</td>
<td>4.5</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Beverages</td>
<td>1.0</td>
<td>2.8</td>
<td>0.3</td>
<td>0.4</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Flowers, capsers</td>
<td>3.3</td>
<td>2.7</td>
<td>1.7</td>
<td>4.6</td>
<td>5.4</td>
<td>4.2</td>
</tr>
<tr>
<td>Roots</td>
<td>15.8</td>
<td>7.6</td>
<td>3.8</td>
<td>8.6</td>
<td>9.0</td>
<td>9.1</td>
</tr>
<tr>
<td>Saps</td>
<td>4.6</td>
<td>2.0</td>
<td>2.0</td>
<td>13.3</td>
<td>8.1</td>
<td>1.7</td>
</tr>
<tr>
<td>Barks</td>
<td>2.0</td>
<td>0.4</td>
<td>1.1</td>
<td>1.6</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>No. of items</td>
<td>304</td>
<td>487</td>
<td>643</td>
<td>128</td>
<td>407</td>
<td>471</td>
</tr>
</tbody>
</table>

* Including stems, sprouts or shoots, and meristems.
† Including bulbs and rhizomes.
‡ Including gums, latex, sugars, and masticatories.

species appeared originally as invaders in ancient cultivated fields. Furthermore, several cosmopolitan “weeds,” such as dandelion (*Taraxacum officinale*), garden rocket (*Eruca vesicaria*), common purslane (*Portulaca oleracea*), and shepherd’s purse (*Capsella bursa-pastoris*), are nowadays cultivated for the specialty food market.

After analyzing a set of 22,521 species of plants inhabiting natural ecosystems of North and South America, we concluded that 11.3% were edible. In another sample of 1269 species from seminatural communities in the temperate Neotropical region, edibles were 20.3% of the total. But if only weeds are considered (2455 widely spread species), the proportion mounts to 35.6%. This demonstrates that the degree of environmental disturbance correlates with the proportion of weeds, which in turn increases the proportion of edible food resources. This should not be surprising given that 16 of the world’s 18 most aggressive weeds provide parts for human consumption. Because weeds are so numerous (more than 10,000 species catalogued) and so abundant (averaging 1.3 and 2.1 tons/ha in a temperate and a tropical area of northwestern Patagonia and eastern Mexico, respectively. Diaz-Betancourt et al., 1999, Rapoport et al., 1998), they stand ambiguously as both enemies and potential benefactors of humanity.

The prospects for the future of food plant diversity appear to be auspicious. Further research should be conducted to manage and profit from these varied and abundant natural resources. At present, more than 15,000 species of food plants are recorded and this figure is constantly growing. Yet most of the world is fed with about 20 crops. As has been pointed out by Facciola (1990), 8000 cultivars of apples have been developed by humans, but only a handful are available in supermarkets.

See Also the Following Articles

AGRICULTURE, TRADITIONAL • DOMESTICATION OF CROP PLANTS • PLANT BIODIVERSITY, OVERVIEW

Bibliography


EDUCATION AND BIODIVERSITY

Shirley M. Malcom
The American Association for the Advancement of Science

I. Introduction
II. Learning about Biodiversity: K–12 Education
III. Environmental Education
IV. Learning in the Informal Education Sector
V. Public Awareness
VI. Colleges and Universities
VII. Systematics Research
VIII. Conclusion

GLOSSARY

formal education Education that takes place in a school environment.
informal education Education that takes place in non-school settings, such as museums, zoos, parks, or through the media.
out of school experiences Things that individuals experience that may support or reinforce learning. These might include participation in youth-serving organizations or hobbies such as bird watching.

THIS CHAPTER REFLECTS a view from the science education and scientific communities of what all high school graduates would know and understand about biodiversity. This is a statement of learning goals and does not necessarily represent the actual understandings that an average student would attain through study of science through grade 12 in the typical American high school.

1. INTRODUCTION

The Convention on Biological Diversity articulates a case for nations of the world to come together to undertake activities to improve conservation of biodiversity and sustainable use of biological resources. As of June 1997 more than 170 nations had ratified the convention. In addition to calls for better management, more research, and study and international and regional cooperation, there was also recognition of the role of education, public participation, public information, and the development of a cadre of professionals to support the goals of the convention.

A recent report to President Clinton from the President's Committee of Advisors on Science and Technology (PCAST) makes recommendations to strengthen "the understanding and management of biological resources" (PCAST, 1998). Among the recommendations in Teaming with Life: Investing in Science to Understand and Use America's Living Capital are calls for increased opportunities for formal and informal education centered on biodiversity and ecosystems, for interactions between scientists and students, and for continuing professional education for K–12 teachers.

This chapter will outline aspects of formal and informal education focused on biodiversity. Specifically, it will outline learning goals around biodiversity for K–12.
education, recommended school-based experiences that can lead to an attainment of these goals, and evidence as to the extent to which these understandings are being achieved by students. The chapter will review school-based environmental education initiatives and describe how these compare to and differ from education about biodiversity. The chapter will then continue with a discussion of informal learning opportunities in "places of science" as well as those available through youth-serving organizations, tourism, and field experiences. The section on general education will conclude with information on public interest in and awareness of biodiversity.

A brief section will discuss biodiversity and tertiary education: biodiversity themes and courses as part of liberal education as well as issues in the education of professionals and specialists who work in biodiversity research, management, and conservation.

II. LEARNING ABOUT BIODIVERSITY: K–12 EDUCATION

How do students learn about biodiversity? What specific concepts must they learn and what ideas must they acquire to support that understanding? What class work, materials, curriculum, set of courses, and experiences would provide an adequate background so that they come to an understanding of this concept?

In 1985 the American Association for the Advancement of Science (AAAS) began Project 2061, a long-term initiative to reform K–12 education in natural and social sciences, mathematics, and technology. *Science for All Americans*, published in 1989, was conceived as a statement of learning goals for science, mathematics, and technology education, defining what all students should know and be able to do by the time they conclude secondary education. *Science for All Americans* includes learning goals related to biodiversity (see Box 1).

<table>
<thead>
<tr>
<th>Box 1</th>
<th>Biodiversity: What Would a Science Literate Adult Understand?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity of Life</td>
<td>Millions of different types of individual organisms inhabit the earth at any one time—some very similar to each other, some very different. Biologists classify organisms into a hierarchy of groups and subgroups on the basis of similarities and differences in their structure and behavior. One of the most general distinctions among organisms is between plants, which get their energy directly from sunlight, and animals, which consume the energy-rich foods initially synthesized by plants. But not all organisms are clearly one or the other. For example, there are single-celled organisms without organized nuclei (bacteria) that are classified as a distinct group. Animals and plants have a great variety of body plans, with different overall structures and arrangements of internal parts to perform the basic operations of making or finding food, deriving energy and materials from it, synthesizing new materials, and reproducing. When scientists classify organisms, they consider details of anatomy to be more relevant than behavior or general appearance. For example, because of such features as milk-producing glands and brain structure, whales and bats are classified as being more nearly alike than are whales and fish or bats and birds. At different degrees of relatedness, dogs are classified with fish as having backbones, with cows as having hair, and with cats as being meat eaters. For sexually reproducing organisms, a species comprises all organisms that can mate with one another to produce fertile offspring. The definition of species is not precise, however; at the boundaries it may be difficult to decide on the exact classification of a particular organism. Indeed, classification systems are not part of nature. Rather, they are frameworks created by biologists for describing the vast diversity of organisms, suggesting relationships among living things, and framing research questions. The variety of the earth's life forms is apparent not only from the study of anatomical and behavioral similarities and differences among organisms but also from the study of similarities and differences among their molecules. The most complex molecules built up in living organisms are chains of smaller molecules. The various kinds of small</td>
</tr>
</tbody>
</table>
molecules are much the same in all life forms, but the specific sequences of components that make up the very complex molecules are characteristic of a given species. For example, DNA molecules are long chains linking just four kinds of smaller molecules, whose precise sequence encodes genetic information. The closeness or remoteness of the relationship between organisms can be inferred from the extent to which their DNA sequences are similar. The relatedness of organisms inferred from similarity in their molecular structure closely matches the classification based on anatomical similarities.

The preservation of a diversity of species is important to human beings. We depend on two food webs to obtain the energy and materials necessary for life. One starts with microscopic ocean plants and seaweed and includes animals that feed on them and animals that feed on those animals. The other one begins with land plants and includes animals that feed on them, and so forth. The elaborate interdependencies among species stabilize these food webs. Minor disruptions in a particular location tend to lead to changes that eventually restore the system. But large disturbances of living populations or their environments may result in irreversible changes in the food webs. Maintaining diversity increases the likelihood that some varieties will have characteristics suitable to survival under changed conditions. As noted in Science for All Americans, Project 2061, from the American Association for the Advancement of Science:

Our planet’s essential goods and services depend on the variety and variability of genes, species, populations and ecosystems. Biological resources feed and clothe us and provide housing, medicines and spiritual nourishment. The natural ecosystems of forests, savannahs, pastures and rangelands, deserts, tundras, rivers, lakes and seas contain most of the Earth’s biodiversity. Farmers’ fields and gardens are also of great importance as repositories, while gene banks, botanical gardens, zoos and other germplasm repositories make a small but significant contribution. The current decline in biodiversity is largely the result of human activity and represents a serious threat to human development. (1989, pp. 60–61).

• Some animals and plants are alike in the way they look and in the things they do, and others are very different from one another.
• Plants and animals have features that help them live in different environments.
• Stories sometimes give plants and animals attributes they really do not have.

Recommended learning activities include providing students the opportunity to observe a variety of plants and animals in the classroom; on the school grounds; in the community; at home in parks, streams, and gardens; and at the zoo. The observations would prompt students to pursue questions about how the organisms live, where they are found, or how they interact with other organisms.

B. Grades 3–5

For children in later primary education, Benchmarks recommends that students learn the following:

• A great variety of kinds of living things can be sorted into groups in many ways using various features to decide which things belong to which group.
• Features used for grouping depend on the purpose of the grouping.

Recommended learning activities include providing students the opportunity to learn about an increasing variety of living organisms and offering them a chance to invent schemes for classification. Students would be encouraged to develop different classification schemes (without being introduced to the Linnean classification system) and shown how their usefulness varies depending on the purpose of the classification. The purpose of the work would be to help students develop a deeper understanding about the relatedness of organisms.

C. Grades 6–8

For children in upper primary and lower secondary education Benchmarks argues that science should provide students with opportunities to enrich their growing knowledge of the diversity of life on the planet and to begin to connect that knowledge to what they are learning in geography. That is, whenever students study a particular region in the world, they should learn about the plants and
animals found there and how they are like or unlike those found elsewhere.

Food patterns of development and external and internal structures would all be used to illustrate interrelationships, interdependence, similarities, and differences. Students would be introduced to the features that biologists use in classification systems and would be taught why these classifications are made.

Benchmarks suggests that students in this group should know the following:

• One of the most general distinctions among organisms is between plants, which use sunlight to make their own food, and animals, which consume energy-rich foods. Some kinds of organisms, many of them microscopic, cannot be neatly classified as either plants or animals.

• Animals and plants have a great variety of body plans and internal structures that contribute to their being able to make or find food and reproduce.

• Similarities among organisms are found in internal anatomical features, which can be used to infer the degree of relatedness among organisms. In classifying organisms, biologists consider details of internal and external structures to be more important than behavior or general appearance.

• For sexually reproducing organisms, a species comprises all organisms that can mate with one another to produce fertile offspring.

• All organisms, including the human species, are part of and depend on two main interconnected global food webs. One includes microscopic ocean plants, the animals that feed on them, and finally the animals that feed on those animals. The other web includes land plants, the animals that feed on them, and so forth. The cycles continue indefinitely because organisms decompose after death to return food material to the environment.

D. Grades 9–12

For students at the secondary level, curricular objectives lead to understanding diversity within and among species by looking at ‘same and different’ features at a molecular level. Students would learn the following:

• The variation of organisms within a species increases the likelihood that at least some members of the species will survive under changed environmental conditions, and a great diversity of species increases the chance that at least some living things will survive in the face of large changes in the environment.

• The degrees of kinship between organisms or species can be estimated from the similarity of their DNA sequences, which often closely matches their classification based on anatomical similarities.

Understanding built up over this period of study would lead students to comprehend the diversity of ecosystems, diversity of species, and the genetic diversity within species.

While Benchmarks sets out a recommended sequence of learning goals to help students come to an understanding of biodiversity as a complex idea, it is not clear that most students have access to the education, ideas, concepts, and learning experiences needed to achieve such understandings. Therefore it would be necessary to explore what students are taught or expected to learn over time during their schooling.

E. Biodiversity and School Science

Formal education in science is an important contributor to students’ fundamental understandings about science. While self-directed study—books, articles, the Internet, museum visits, and field experiences—augment science learning for many students, the quality of the curriculum, textbooks, and other instructional materials, the preparation of teachers, the school-mediated experiences provided to students both inside and outside of the classroom all interact to shape what students take away from school science. Understandings of biodiversity would be based on accumulated experiences and knowledge. These would include the early school focus on ‘natural history’ and ‘nature study’ and development of an ‘intuitive’ understanding of biological diversity and the relationships among living organisms (National Research Council, 1990). Students’ out-of-school experiences, where such are available, would reinforce school learning. The curriculum focus shifts in lower and upper secondary levels (grades 6–12) to more formal, taxonomic instruction.

This pattern of topic coverage for life sciences concepts is present in the curriculum of other countries around the world and was prevalent among the majority of the 50 educational systems of countries that participated in the Third Mathematics and Science Study (or TIMSS). Data were collected in 1993 and results presented in 1996.

In the United States, according to statistics provided by the U. S. Department of Education, biology is the most frequently taken high school science course, with over 93% of 1994 graduates of public high schools
reporting having completed such a class. Over 97% of graduates from nonpublic high schools reported completing biology. Analysis of content core is not available to determine the extent to which the ideas critical to a student’s understanding of biodiversity are actually taught. It is also not clear if earlier foundational ideas are provided to students as a part of instruction at primary and lower secondary levels.

In the United States there has been considerable discussion about the adequacy, pacing and structure of curriculum and, especially, of textbooks in middle and high school biology. Researchers have criticized their ‘encyclopedic’ nature, with too many ideas covered too superficially, and too much focus on vocabulary at the expense of big ideas. A recent study that evaluated the science textbooks used in middle grades against three concepts from Benchmarks and Standards, one each from the earth, physical, and life sciences, led AAAS Project 2061 to conclude that all nine titles examined were inadequate to help students achieve understanding of the ideas that they were attempting to explain.

Efforts to actually measure what students know about science are undertaken through the regularly administered National Assessment of Educational Progress and recently have been internationally benchmarked through TIMSS. At lower and upper secondary levels, the results suggest that U. S. students gain little real understanding of the big ideas of science, including life sciences, in spite of the fact that they are most likely to have had formal classes in this area than in other areas of science. It has largely been in the study of biology that the basic concepts of biodiversity have been advanced; diversity within species, across species and ecosystems; relatedness of species; interdependence among living things; threats to this diversity due to population pressures; loss of habitat; changes in species over time; natural extinctions; and human induced extinctions. While biodiversity and biodiversity education are more recent concepts, nature study, conservation, and environmental education have older roots, in a more general emphasis on environmental improvement and appreciation. The following section will explore some of the history of environmental education and consider how it might relate to biodiversity education.

### III. ENVIRONMENTAL EDUCATION

In 1977 the world’s first intergovernmental conference on environmental education was organized by the United Nations Environment Programme and held in Tbilisi, Georgia (Soviet Union). The Tbilisi Declaration, adopted at the conference, stressed the importance of environmental education in the preservation and improvement of the world’s environment. Agenda 21, Chapter 36, outlines issues related to formal education, public awareness and training to promote sustainable development and emphasizes environmental and development education as an essential aspect of all education. There were also calls for linking to environmental education in recommendations from the World Conference on Education for All (held in Jomtien, Thailand in 1990) that urged a move toward universal access to basic education for girls and boys.

The recommendations include calls for better informal education, promotion of environmentally sound leisure and tourism activities, programs to involve young people and children, as well as respect for and support of efforts to promote dissemination of traditional and socially learned knowledge through mechanisms based in local cultures.

Environmental education was called “nature study” when it got its start in the 1920s with Junior Audubon Clubs teaching children to appreciate nature. According to Karen Schmidt in a December 13, 1996, article in Science, the movement was transformed into conservation education in the 1930s when the Dust Bowl environmental tragedy led to incorporation of ideas into some schools about the management of natural resources.

With the initiation of Earth Day in 1970 and passage of the National Environmental Education Act of 1970, teachers received supplemental training in environmental education. Many states enacted their own environmental education laws, and schools began incorporating these topics into science classes. On the 20th anniversary of Earth Day, President Bush signed the National Environmental Education Act of 1990, which created an Office of Environmental Education at the Environmental Protection Agency and supported curriculum development and teacher training in the states.

In the United States concern has been expressed about the place of environmental education, as currently configured into the curriculum, and especially about the quality of programs sometimes offered in lower secondary schools. This has included concerns about an advocacy orientation in instruction, about the need for and balance in materials used and instruction provided. The North American Association for Environmental Education (NAAEE), a professional group that includes college faculty and K–12 educators among its members, developed “Guidelines for Excellence” to help guide teachers and others in the selection of quality, balanced materials.
Teaming with Life makes a strong recommendation that environmental education have a stronger base in science, using scientifically grounded curricula. Innovative programs such as GLOBE (see Box 2) that depend on student-scientist partnerships and collection of real data may point the way to science education based around environmental and biodiversity concerns.

In 1996 a national conference held in Washington, D.C., spotlighted projects that depended on student-scientist partnerships, including a number that focused on biodiversity as learning themes. Included among these were “Classroom Feederwatch,” a program of authentic research for grades 5–8 developed by Cornell Laboratory of Ornithology and TERC, with funding from the National Science Foundation. In this program, students (and their teachers) learn to identify birds, to ask scientific questions and to design experiments to answer them. Students analyze and display data to answer their questions and collect and share data in a research database used by professional ornithologists in their studies of bird populations. Students thus become part of a larger team of those contributing to biodiversity studies. (For more information see http://birdsource.cornell.edu/cfw/watiscf.htm.)

IV. LEARNING IN THE INFORMAL EDUCATION SECTOR

A. Overview

A wide range of informal education experiences are available for adults and children to extend their knowledge about biodiversity. These include organizations that incorporate biodiversity education and exhibition within their missions, such as the following:

- Zoos
- Botanical gardens
- Aquariums
- Museums
- National parks

Depending on their size, these “places of science” might also incorporate research, collections, conservation, or other functions important to biodiversity. Signage and docent-led and audio tours provide additional information to visitors. Classes, lectures, and workshops (including those for teachers) are also often provided to

Box 2

Global Learning and Observation to Benefit the Environment (GLOBE)

GLOBE, a hands-on, school-based international environmental science program, was introduced by U.S. Vice President Al Gore in April 1994 and began operation on Earth Day in April 1995. In 1999 there were more than 6000 participating schools in more than 70 countries. GLOBE brings together students, teachers, and scientists from around the world to enhance environmental awareness of individuals worldwide, increase scientific understanding of the earth, and support improved student achievement in science and mathematics.

Students make environmental observations, or take environmental measurements near their school site, report their observations via the Internet, receive and use GLOBE images created from the combined data, and study the environment by relating their observations to larger environmental topics. GLOBE educational materials were developed by environmental educators and curriculum development specialists working with scientists. Materials are used in schools under the guidance of teachers who have received training using GLOBE materials. Teachers include GLOBE activities as appropriate within their local curricula. GLOBE materials are translated into the six United Nations’ languages (Arabic, Chinese, English, French, Russian, and Spanish) and are also available in Czech, Estonian, German, Greek, Hebrew, and Japanese. GLOBE international partners sign bilateral agreements with the United States and manage participation of schools in their countries. In the United States, GLOBE is administered by an interagency team that includes the National Oceanic and Atmospheric Administration, the National Aeronautics and Space Administration, the National Science Foundation, the Environmental Protection Agency, the Departments of State and Education, and others.

Observations range from basic weather parameters (temperature, atmospheric pressure, and precipitation) to measurements such as water chemistry, biodiversity, and biomass assessment. The focus to date has been more heavily directed toward physical systems measurements. Potential is great to increase the biodiversity and life systems aspects of the project.
extend the learning experience. Increasingly, materials and web sites incorporate aspects of a visit, bringing resources to audiences at a distance to make some part of the visitor experience remotely available.

Informal education also includes more intensive immersion experiences such as where it is incorporated into visits to natural preserves such as parks (including those within national park systems) and forests, or that provided through ecotourism.

Interpretive programs using volunteers, staff, and written and video materials provide enhanced learning experiences by bringing the science, the issues, and the concerns into sharp focus as a part of the overall environmental experience. Programs such as Earthwatch have biodiversity-focused visits that involve the participant in the research as data collector. Other informal learning opportunities are available through television and IMAX programs, web sites, and books.

Youth-serving groups provide a broad range of activities and experiences that can support education around concepts of biodiversity (see Box 3). Many may involve long-term projects of environmental monitoring, animal and plant breeding, habitat restoration, and other activities undertaken individually or in groups.

Box 3
Selected Youth Serving Groups Providing Informal Science/Environmental Education

American Camping Association (ACA)
Founded in 1910 ACA provides an accrediting mechanism for camps. Of the more than 2200 accredited camps listed, 284 provided nature/environmental study as part of the camp experience.

Boy Scouts of America (BSA)
Enrollment: 5.6 million
Program: Incorporated in 1910 BSA provides programs for boys that include outdoor skills, nature study, and conservation activities through an elaborated badge structure and group activities. BSA is a charter member of the World Scout Conference.

Camp Fire Boys and Girls
Enrollment: 667,000 young people (birth–age 21)
Program: Camping and environmental education programs offering children an appreciation and commitment to the natural environment.

Girls Incorporated
Enrollment: 350,000 girls (ages 6–18) at more than 1000 sites nationwide
Program: Activities include experiences in mathematics and science education through Operation SMART

Girl Scouts of the U.S.A. (GSUSA)
Enrollment: 3.5 million
Program: Activities include out-of-doors, nature study. Badge structure that includes environmental issues. GSUSA is member of the World Association of Girl Guides and Girl Scouts.

World Association of Girl Guides and Girl Scouts undertakes world projects including building world citizenship. Environment is one of the themes of this program.

National 4-H Clubs
Enrollment: 6.0 million in more than 76,500 clubs
Program: Part of the U.S. Department of Agriculture’s Cooperative Extension Service established in 1914. Project areas include agricultural and natural sciences, and technology, Cornell Nest Box network.
B. Zoos and Other “Places of Science”

Animal parks were established by and for rulers. Maier and Page, in their volume Zoo: The Modern Ark (1990), describe how animals were kept by royalty for entertainment and as a show of wealth. The third dynasty ruler of the Sumerian city of Ur had a park that dated around 2300 B.C. A millennium later as civilization spread in the Near East and Asia, rulers and pharaohs exchanged “exotic” animals for their zoos. Emperor Wu Wang of the Chou dynasty laid out a zoological garden called the Park of Intelligence. Animal collections were found around the globe in early civilizations such as in Egypt some 3500 years ago.

Alexander the Great, perhaps influenced by Aristotle’s private menagerie, installed what was perhaps the first public zoo in Alexandria, Egypt. With the coming of the “Dark Ages” of Europe, monasteries became the keepers of menageries and game parks. When Cortes arrived in the Aztec capital of Tenochtitlan, he found a large zoo behind Emperor Montezuma’s palace. Zoos in India were established by Akbar toward the end of the 16th century. He, like the Aztecs, employed people specially trained to care for and medically tend to animals.

The zoo at Vienna was reinvigorated by Maria Theresa and her husband as the Imperial Menagerie at Schönbrun for the convenience and entertainment of the nobility. The zoo remains today as likely the oldest in continuous operation, dating from the 1750s. Democratization of Europe and establishment of urban centers that accompanied industrialization led to the “modern zoo” as a repository of exotic specimens of life that were to be studied as a way of understanding “flora and fauna” of the world. Public monies (rather than private patrons or royal largesse) were available to begin systematic scholarly study. Maier and Page date the modern zoo to 1826 when the Zoological Society of London founded the zoological gardens at Regent’s Park for the purpose of understanding the natural history of the animals inhabiting the reaches of the British Empire.

Since zoos as public institutions had to raise funds and attract money (independent of their research and conservation goals), they had to become popular attractions. Zoo organizers also had to learn to manage space and figure out and meet animals’ requirements, such as for social interaction. Zoo’s role in conservation became educational as they raised visitor awareness about endangered species and loss of habitat. Where larger zoos also developed significant breeding herds, they established breeding farms. In San Diego, for example, this “wild animal park” has become an additional attraction.

The National Zoological Park (National Zoo), associated with the Smithsonian Institution, established a “biopark,” Amazonia, to emphasize the relationships among soil, plant, invertebrate, and other animal forms and the need to preserve the habitats of the world. Zoos, aquariums, and game parks are being seen as tools to affect public attitudes regarding the variety of life on earth. As these “places of science” intentionally blend education and entertainment they are increasingly adding materials from museum collections and incorporating interactive exhibits from science—technology centers to reinforce conservation messages, concern about loss of species numbers, and diversity and loss of habitat.

1. Botanical Gardens

In 1989 the World Resources Institute estimated that 150 million persons visited some 1500 botanical gardens around the world. In addition to visits and guided tours, gardens offered continuing education for adults, workshops and hands-on experiences for children and families, and professional education courses and seminars for K–12 teachers. The New York Botanical Garden and Missouri Botanical Garden are examples of two of 21 member gardens of the American Association of Botanical Gardens and Arboreta offering graduate studies programs, usually in collaboration with universities in their area.

2. Museums

Through collections, education programs, exhibitions, and graduate-level research, museums have been very active in promoting biodiversity in both the formal and informal sectors. The American Museum of Natural History in New York (AMNH) provides an interesting example of an institution with current involvement in all these areas:

- **Exhibition.** The 11,000-square-foot Hall of Biodiversity is the newest permanent exhibit of AMNH and uses collections, interactive technologies, and an immersive environmental replica of a portion of the rain forest of the Central African Republic—complete with sound, smell, movement, and running water—to provide a unique visitor experience.
- **Graduate and continuing education.** The Center for Biodiversity and Conservation collaborates within and outside the museum in the development of courses and programs. AMNH is home to the oldest and largest doctoral and postdoctoral training...
program of any scientific museum in the world, collaborating with Yale, Columbia, Cornell, and City University of New York.

- **Education.** The National Center for Science Literacy, Education and Technology supported by the National Aeronautics and Space Administration (NASA) has developed a number of projects related to the theme of biodiversity, including **Biodiversity Counts: A Student Inventory Project**, a program for middle school students across the United States to inventory plant and animal life in their communities and to share their findings through publications and on-line field journals.

### C. Biodiversity "Experiences" and Resources

**Overview**

For most adults, biodiversity education will take place in the informal sector as they read books; visit zoos, museums, and national parks; listen to lectures; and watch programs on the increasing number of science- and nature-based cable channels, public television, or the increased coverage of science on the news or news magazines. Others will visit the World Wide Web, where an increasing number of excellent sites developed by universities, museums, federal agencies, and non-profit organizations provide high-quality information. Several examples of resources for biodiversity education available to the adult public follow.

1. **Earthwatch**

   Earthwatch Institute is an international nonprofit organization founded in 1971 that supports scientific field research worldwide. Volunteers participate in actual field research, assisting scientists in gathering data. Since its beginnings it has "mobilized 150 projects around the globe, resulting in the discovery of 2000 species, the establishment of 12 national parks, and the founding of eight museums." The Earthwatch web site lists active projects that volunteers can join in seven topical areas including Endangered Ecosystems and Biodiversity. The site also includes virtual field trips. Visit [http://www.earthwatch.org](http://www.earthwatch.org).

2. **National Biological Information Infrastructure (NBII)**

   The NBII attempts to organize the disparate sources of information available through agencies, departments, museums, and other organizations, providing a source of links to sites. A "Biodiversity, Systematics and Collections" section connects to other web sites, many of which have education or "for kids" sections. Many federal agencies such as the U.S. Geological Survey, the Environmental Protection Agency, NASA, and others have relevant sites. Visit [http://www.nbii.gov/biodiversity](http://www.nbii.gov/biodiversity).

3. **Nonprofit Environmental and Biodiversity Groups**

   A number of organizations produce materials to support education about environmental and biodiversity issues. These include groups such as the World Wildlife Fund, the Sierra Club, and the Audubon Society. These groups develop a wide variety of public information and educational materials.

   While most mainstream advocacy groups are conscious of the need to "get the science right" and to present balanced viewpoints concerns are sometimes expressed about school use of materials that emanate from an "advocacy position." Guidelines have been developed by NAAEE to assist educators in assessing the scientific accuracy of such materials.

### V. PUBLIC AWARENESS

How much does the public understand about environmental issues in general and biodiversity in particular, and what are the attitudes toward these issues? Surveys from a number of sources indicate that there is strong public interest in and support for issues related to the environment. The National Science Board's *Science and Engineering Indicators* (1998) suggested strong interest and "informedness" of the public around environment and health topics, especially when compared with other science and technology areas, and interest and support were stronger among women than men. The National Environmental Report card, an attitudinal and knowledge survey of American adults conducted by the National Environmental Education and Training Foundation and Roper Starch Worldwide, concluded that there was "an alarming lack of knowledge about some of our most critical environmental problems." With regard to biodiversity, however, 73% of adults surveyed correctly responded about the direct relationship between species loss and habitat destruction.

In 1995 the Consultative Group on Biological Diversity initiated the Biodiversity Project, a public education effort by a nonprofit grant makers' forum to "assess public opinion on biodiversity, to develop collaborative strategies to increase public awareness and engagement, and lay the groundwork to implement those strategies."
In June 1998 a “summit” on biodiversity and environmental education was convened at the American Museum of Natural History as a Biodiversity Educators Summit. The summit was supported by evidence from focus groups and surveys and co-convened by The Project, AMNH, and World Wildlife Fund. The 1996 Biodiversity Poll, conducted by the public opinion research firms Belden & Russonello and R/S/M and reported at the Summit, revealed the following about the environment and biodiversity:

- People care about the environment, but it isn’t in the top tier of public concerns.
- Of environmental concerns, the public considers the most serious problems to be toxic waste, destruction of the rain forest, loss of places in nature, and air and water quality.
- Extinction is a concern, but it is not high on the list.
- People understand that nature is connected and interdependent, but most people do not recognize or use the word biodiversity. Only 2 in 10 said they had heard about the “loss of biological diversity.”
- The public understands that species are declining and that human activity is largely responsible. But the public does not understand much about specific reasons or about the seriousness of the rate of loss.
- Public support for biodiversity conservation (once biodiversity is explained) is wide—87%. But this support is shallow.
- Countervailing pressures (values) can peel support away from biodiversity protection. These include concerns about jobs, individual property rights, comfort and convenience, and preservation of “unattractive” species. However, 91% of Americans agree that the world would suffer if such “unattractive species” (e.g., mosquitoes) are eliminated.

The gap may relate to the following:

- A general lack of attention to biodiversity and its consequences by the media
- The way that the public message about biodiversity is conveyed
- The biological understandings that members of the public bring to the discussion

Interestingly, the spottiness of adult knowledge and concepts (some individual ideas understood, but not the overall concepts nor the consequences that flow from them) tracks with observations made about the K–12 student understanding of life sciences ideas. For most members of the public, high school is the last time for a formal course in the life sciences. After that, they generally rely on the informal sector for additional information and updates to their knowledge.

VI. COLLEGES AND UNIVERSITIES

According to Science and Engineering Indicators, persons taking college-level science courses are more likely to be informed about, supportive of, and interested in science and technology topics. College level courses in biology influence knowledge and attitudes of the public. Beyond Biology 101, produced by the Howard Hughes Medical Institute, describes efforts to transform college-level courses in biology for both majors and nonmajors, moving away from vocabulary-driven courses to ones that are more integrative and that include meaningful laboratory and field experiences. A number of programs are described at http://www.hhmi.org/BeyondBio101.

These include an innovative program in human biology developed and in place since the 1960s at Stanford University. The program, taught by faculty from biology, education, anthropology, psychology, and other disciplines, focuses on the relationship between human biology and human behavior including human interactions with environment. It is a major course of interdis-
The program for majors involves students in a range of activities designed to get them to think like scientists. Biodiversity and human impacts on the environment are explicit foci of instruction, and in the late 1980s and 1990s the college program was “translated” into a middle grades life sciences project (HumBio). Biodiversity concepts and activities are explicitly included among the curriculum materials in the ecology theme. While most four-year institutions offer majors in biology or related specialties (including ecology), biodiversity was found as an area of major concentration at only a few U.S. colleges and universities. A search of the site, Academic Programs in Conservation Biology (http://www.conbio.rice.edu/programs), yielded programs at Columbia University, Illinois State University, San Diego State University, State University of New York at Albany, University of California at Riverside, University of Massachusetts at Amherst, University of Southwestern Louisiana, University of Wisconsin at Madison, and Yale University. Searches for “sustainable development” added to this list programs at Cornell University, Harvard University, University of Georgia, University of Maryland, University of Miami, and University of Minnesota. Distance learning classes were also available in these fields and listed at http://eelink.net.

Other countries that have evidence of strong interest in collegiate and university studies in biodiversity include Canada and Australia. The United Nations University, Institute for Natural Resources in Africa (UNU/INRA) has a number of relevant programs in the following categories:

- Soil and water management: Soil fertility restoration and maintenance.
- Conservation of biodiversity: Genetic improvement and increased utilization of Africa's indigenous food crops and useful plants
- Conservation and management of mineral resources.

Education and training are key areas of interest to UNU/INRA. In cooperation with other agencies of the UN, the program develops curriculum and contributes to training in areas such as ecological economics, natural resource economics and environmental accounting, germplasm and biodiversity conservation, wildlife management, and taxonomy. Gender and Natural Resources is a major cross-cutting theme in the work of the institute.

VII. SYSTEMATICS RESEARCH AND TRAINING

While the demand for expertise in conservation, biodiversity, and systematics has been increasing, concern is being expressed about the human resources, especially in developing countries, available to manage and inform natural resources utilization around the globe. A 1995 workshop on Priorities in Systematics Research and Training organized by the United Kingdom Systematics Forum and held at the Linnean Society of London raised issues about the adequacy of support for systematic biology, the declining interest in systematics among students, and the decline in the teaching of systematics in many universities. A search for university departments worldwide that provide training in systematics and taxonomy revealed 24 institutions, 10 in the United States. This may be deceptive, however, in that a number of institutions provide graduate training in partnership with research-oriented museums and botanical gardens. New emphases such as work in molecular systematics may exist in cellular, molecular, microbiology or biochemistry programs. BIO NET-INTERNATIONAL is a global network of people and institutions that develop biosystematics capacity in developing countries. Training at all levels (in service, short courses, distance courses, and joint graduate programs) is a major focus of the network’s activities.

VIII. CONCLUSION

A combination of school-based learning and out of school experiences combine to provide young people with knowledge of and attitudes about biodiversity that they then take into adulthood. For those who pursue higher education, college-level courses are available in some institutions that integrate biodiversity education into larger biological, environmental, or human impacts courses. Other adults must depend on the informal education sector, with experiences provided by a variety of different institutions and media. Biodiversity education may also, in some cultures, rely on community transmission of locally held knowledge of plants and animals of a region. Whatever the process for developing understanding, education and public awareness have been seen as crucial precursors to building support for biodiversity.
Additional Information


Center for Biodiversity and Conservation. Please visit http://research.amnh.org/biodiversity/ or contact the Center for Biodiversity and Conservation, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024; telephone: (212) 769 3742; fax (212) 769 3262; or via e-mail at biodiversity@amnh.org.

Consultative Group on Biological Diversity. Please contact Lincoln and Torney, P.O. Box 1014, 555 New Jersey Avenue, Washington, D.C. 20001; telephone: (202) 884-4949; or visit http://www.bio-diversityproject.org/more.htm#cgbd for additional information.

NAEPE. Please contact NAAEE Headquarters, 1825 Connecticut Avenue, NW, Suite 800, Washington, D.C. 20009-5708; telephone: (202) 884-8912; fax: (202) 884-8455; or visit http://www.naaee.org/html/contact.html.

NAEP. Please contact Bob Clemons, National Center for Education Statistics, 333 New Jersey Avenue, NW, Washington, DC 20202; telephone: (202) 401-8000; or visit http://nces.ed.gov/nationsreportcard/site/contact.asp.


TERC. Please contact TERC, 2067 Massachusetts Avenue, Cambridge, MA 02140; telephone: (617) 547-0430; fax: (617) 349-3535; or visit http://www.terc.edu/.

TIMSS. Please contact the U.S. TIMSS National Research Center, University of Pennsylvania, 415 Ericsson Hall, East Lansing, MI 48824-1034; telephone: (517) 353-7798; fax: (517) 353-7788; or visit http://www.nusm.edu/timss/.

UNESCO. Please visit http://www.unesco.org/.


See Also the Following Articles

Biodiversity, Definition of; Conservation Efforts, Contemporary; Government Legislation and Regulation; Historical Awareness of Biodiversity; Human Impact on Biodiversity, Overview

Bibliography


I. INTRODUCTION

Endangerment is the condition in which a species or subspecies of animal or plant is at risk of extinction. The nature of the impending extinction may be local, national, or global. There is no vocabulary to differentiate between these conditions; as a consequence, there is often confusion over appropriate priority levels. A trend to use extirpation to indicate nonglobal extinction brings its own problems, since the active verb extirpate is synonymous with eradicate; there is also the problem that disappearing subspecies of species may be both extinct and extirpated (extinct as taxonomically distinct forms, extirpated as representatives of a higher taxonomic unit). This article concentrates primarily on bird species at risk of global extinction, but nationally and locally endangered forms, including subspecies, are also considered; so extirpated is here used to mean locally extinct. The terms endangered, threatened, and at risk
are commonly used interchangeably; endangered is prevalent in U.S. usage, but IUCN/The World Conservation Union status categories use threatened as a generic term and Endangered (with a capital E) for a particular status. In this article, endangered is preferred, except in contexts involving the IUCN threat categories.

II. THE IDENTIFICATION OF ENDANGERED BIRDS

A. Taxonomy and the Identification of Target Units

In general, decision makers look to science in two ways for help in establishing priorities in species conservation: the first is the degree of a taxon's risk of extinction, and the second is the degree of its evolutionary differentiation. Many people are surprised that these aspects, particularly the second, remain hard to assess. Ever since the passing of the United States Endangered Species Act, the term “endangered species” has become widely used in popular and indeed jocular English parlance; so it seems extraordinary that debate still rages over what a species is. Nonetheless, its definition is a pervasive problem, and the profile of this issue will increase as more morphologically distinct, local forms, currently considered subspecies, come under pressure from human development activities.

When considering closely related taxa in which there is no geographical contact, and therefore no test for reproductive isolation (the key criterion under the prevalent Biological Species Concept or BSC), taxonomists are compelled to rely on the degree of their morphological similarity to judge, subjectively, whether they are conspecific or not. Among birds there are thousands of such cases, in part because flight has allowed them to colonize so many offshore and oceanic islands, where they have evolved features that distinguish them at some level from continental or other island stock.

Some influential museum-based ornithologists have recently advocated a narrow phylogenetic species concept (PSC), which bestows species-level identity on any population whose members fully share distinct characters, irrespective of hybrid zones. Many of these populations currently have taxonomic standing as subspecies, and proponents of the PSC expect its adoption to cause a doubling of the number of avian species, to about 20,000 from the roughly 9500 in current usage. However, BSC supporters contend that, rather than clarifying the status of disjunct populations, the PSC shifts the difficulty to a yet more complex and subjective level, where very minor differences (sometimes biochemical only, and certainly not used to recognize even subspecies) may or may not be regarded as sufficiently distinct and consistent to admit species-level status, so that no final number of avian species can be even approximated.

This debate is important, not least because biological diversity is directly at stake. Some years ago the BirdLife Red Data Book program eliminated subspecies from its concern, on the basis that there were simply too many to document and that species had to take priority. This decision, made with the knowledge that many threatened subspecies will benefit from site management for threatened species with which they are sympatric, still tends to expose some subspecies, especially those with small ranges and therefore with relatively constrained populations, to the vagaries of deteriorating global conditions. Unless a country is as wealthy as the United States, whose legislation embraces subspecies and independent populations, or has little or no other biological heritage in which to invest (for birds, Barbados is one such), the chances of intervention on behalf of these forms are relatively low.

Nevertheless, subspecies are far less stable as taxonomic entities than are species, and are easily erected and just as easily subsumed (prior to 1950 at least two museum ornithologists, Oberholser and Koelz, each established over 100 subspecies in the course of a single paper!). Because of this, they are generally unwelcome in law, which requires widespread long-term agreement on taxonomy to be able to function. Consequently, species rather than subspecies are the units of concern in most national and international legislation. Partly perhaps as a consequence of this, many ornithologists seem willing to countenance the steady, consistent “unlumping” of many forms, on the basis of multiple character differences extending beyond morphology to voice, behavior, and even habitat. For conservation purposes this process needs to be expedited to ensure that specific identity is not bestowed too late for intervention. However, the elevation of many weakly distinct forms to species level may actually make the identification of conservation targets harder, for if resources are insufficient to support all the resulting endangered species, many deserving cases (in terms of their evolutionary distinctiveness) might be lost amid the competing claims of virtual look-alikes.

B. Scale and the Identification of Target Units

Birds can be endangered at the global, broad regional (e.g., subcontinental), national, narrow regional (e.g.,
Global endangerment is the most important priority level, since total extinction is a far worse risk than any other form. In the past 35 years or so, global endangerment of birds has been registered principally through the Red Data Book program of the International Council for Bird Preservation (now BirdLife International). This program, accepted by IUCN/The World Conservation Union as the official source of globally threatened bird listings, has developed into a long-term, continent-oriented project to create detailed profiles of every endangered species (Collar and Stuart, 1985; Collar et al., 1992; Collar et al., 1999), based on the rationale that all information relevant to a species’s conservation should be included. This in turn has led to the need for abbreviated global listings (Collar et al., 1994).

The species that have found their way into these full and abbreviated Red Data Books are normally found on national “red lists,” but not always at the expected priority level, owing to the various algorithms that non-global assessment tends to involve. Thus an apparently priority level, owing to the various algorithms that non-abbreviated Red Data Books are normally found on these parameters as well as on decline rate by which to measure eligibility for and degree of threatened status. In IUCN terminology, the word threatened means what has hitherto been called endangered, whereas “Endangered” denotes a specific conservation status.

In crude form, the criteria stipulate that, to qualify as threatened, a species must possess a total population (A) declining at a rate (projected or past) of 20% over 10 years or three generations, or (B) within a range of less than 20,000 km² and declining, or (C) of less than 10,000 mature individuals and declining, or (D1) of less than 1000 mature individuals, or (D2) within a range of less than 100 km². Species meeting any one of these criteria qualify as threatened with the category Vulnerable; nested thresholds qualify species for the categories Endangered and Critically Endangered. Sub-species can be subjected to the same criteria, but this may result in the curious circumstance where all races of a species qualify as at risk (e.g., all five races of an island species with declining populations of under 10,000) but the species itself does not (total population still above 40,000).

The general experience with birds, almost certainly as with all animals and plants, is that populations tend to be significantly underestimated: for example, one observer on the New Caledonian island of Uvea judged there to be 70–90 Uvea (Horned) Parakeets (Eunymphicus cornutus) in 1993, but more intensive fieldwork later that year yielded a formal estimate of 617 ± 379. Similarly, a CAMP (Conservation Assessment and Management Plan) run by the Captive Breeding Specialist Group of IUCN in 1992 suggested a total
population of some 5000 Tanimbar Corellas (Cacatua goffini), whereas analysis of quantified data from fieldwork that same year produced an estimate of 300,000–400,000 birds. Consequently, lists of threatened species cannot be expected to remain stable: while some will be added over time as their situation deteriorates or their taxonomy is revised, others will be removed as their true status is revealed.

In 1988, 1030 bird species were identified as at risk of extinction. In 1994, this figure, based on the new IUCN criteria, rose to 1111. It was not, however, the case that 81 species were added to the 1988 complement. In fact only 816 species were common to both lists. The 214 disappearances from and 295 additions to the 1994 list largely resulted from new “pioneering” knowledge (involving new areas or new identification insights) rather than from “monitoring” updates that disclosed a clear trend. Exploration is thus still the strongest biological data source, and “rare” species (for which see Kunin and Gaston, 1998) sometimes prove to be relatively common in some part of their range, or in some previously uninvestigated habitat. Consequently, attempts to predict future extinction rates using changes in red lists have been premature: the changes in question are not real-world events (Crosby et al., 1996).

A degree of red list stability derives from some species being destined to remain endangered in perpetuity. This is because of their irremediably small ranges or populations (a circumstance that has made the criteria unpopular in some quarters, since no active threat need exist to trigger the listing). These are species—avian examples including the Lava Gull (Larus fuliginosus), with 300–400 pairs maximum, and the Tinian Monarch (Monarcha takatsukasae), on an island of less than 100 km² despite its estimated 40,000 individuals—for which the price of survival is eternal vigilance.

Appropriate criteria for use at the national level are still under development by IUCN; meanwhile, a good model is that of Avery et al. (in Coulson and Crockett, 1995).

III. TRENDS AND FACTORS IN THE ENDANGERMENT OF BIRDS

A. Globally Threatened Birds in 1994: Numbers, Criteria, and Extinction Rate Predictions

The 1111 bird species judged to be at risk of extinction in 1994 represented 11% of the world’s avifauna. Moreover, a further 66 (1%) species were then listed as Data Deficient and 875 (9%) as Near-Threatened, so that altogether over 20% of all bird species were identified as being of some global conservation concern. The majority of threatened species were classified as Vulnerable (704, 63%), with 233 (21%) Endangered and 168 (13%) Critically Endangered.

The commonest criterion triggered by threatened birds was C, which combines small population (<10,000) with significant decline. As many as 764 species (approaching 70%) of all threatened birds were judged (or, under the precautionary principle, thought likely) to fulfill this criterion; thus 8% of all bird species are known or suspected to have dangerously low populations. The other four criteria proved to be rather evenly distributed: A (rapid decline) and B (small range with significant decline) were triggered by around 400 species each, with somewhat smaller numbers triggering D1 (very small population) and D2 (very small range).

The new IUCN criteria attach hypothetical probabilities of extinction to the different categories of threat, on which basis 400 species may be expected to become extinct, without remedial action, in the coming century, however, it will take 1750 years for 90% of the 1111 listed threatened species to disappear. Comparison with the 1988 listing allowed three different listing recruitment rates to be tested on two models, indicating that with current trends between 400 and 1200 species of bird may die out within the next 100 years, with a time to extinction for half the planet’s avifauna (ca. 8500 species) of 800–2800 years (Crosby et al., 1996). Although these figures are less pessimistic than other recent estimates derived from less robust data, human pressures on the environment will only increase in the foreseeable future, so these extinction rates will probably prove much too conservative.

B. Globally Threatened Birds: Regions, Countries, Habitats

The majority of threatened bird species occur in Asia and the New World, with relatively few in Africa. The top ten countries for the highest numbers of threatened species are Indonesia (104), Brazil (103), Philippines and China (both 86), India (72), Colombia (62), Peru (60), Ecuador (50), United States (46), and Vietnam (45). Asian countries predominate in this list; the United States ranks high because of its Pacific territories. The highest African countries, ranking 21st and 22nd, are Tanzania (30) and Madagascar (28).

Priority countries might be selected on this basis,
but several further filters can be applied. Because species in the higher categories of threat are likely to become extinct sooner, a reranking involving just those 403 species that are either Critically Endangered or Endangered yields a top ten of Brazil (47), Philippines (45), Colombia (31), United States (25), Indonesia and Mexico (20 each), Peru (18), and Vietnam, Ecuador, and Argentina (16 each). Using this category, the emphasis shifts dramatically to the New World.

A further filter involves only those species from the preceding analysis that are nationally endemic, thereby indicating the degree of “ultimate responsibility” that falls to these countries as their most urgent bird conservation tasks. The Philippines (40) emerge far ahead of Brazil (32), itself far ahead of Colombia (24), United States (17), Mexico (13), Indonesia and New Zealand (12 each), Australia (11), Madagascar (10), and Peru (9). Although it was fairly obvious from the first analysis, given their small land area, that the Philippines would be a priority area, this refined analysis offers startling evidence of the critical importance of the country in terms of avian biodiversity and its impending loss.

The catastrophic erasure of forests from the planet in the course of the twentieth century means that most threatened birds are (mostly tropical) forest dwellers; the only surprising thing is that the figure is as low as 65%. Wetland species account for 9%, scrubland for another 9%, and grassland for 6%. The relatively low forest representation can be explained in part by the fact that the largest tracts of forest, in Amazonia, the Congo basin, and Borneo, are—despite the destruction visited upon them—still too extensive, and the species they contain too widespread, to have resulted in more than a handful of listings from these areas.

C. Causes of Endangerment in Birds

Endangered birds suffer from a range of different threats. Some 52% of them are affected by habitat loss and degradation (although this figure is almost certainly higher, indeed, low-level and hence unreported habitat loss could probably safely be indicated for the other 48%); this theme is explored in the following paragraphs. The next most important threat is simply restriction of range or population, involving 23% of all threatened birds. Hunting afflicts 8%, introduced species 6%, and trade 3%.

Until recently the notion that habitat loss was important in the demise of the Passenger Pigeon (Ectopistes migratorius) had not been entertained, but Bucher (1992) showed how the species was a specialist on seeds produced in masting events whose scale and geographic location varied from year to year, that is, that were patchy in both space and time. Human settlement of the east and center of the North American continent fragmented the native forests to the point where the lapse in both time and distance between masting events simply became too great. Despite the settlers' prodigious slaughter of birds, which has always been blamed for their disappearance, it appears that it may have been their axes, not their guns, that caused the loss of their quarry; the last wild birds very possibly starved to death.

An equally celebrated North American species, perhaps still extant, is the Eskimo Curlew (Numenius borealis), which, like the Passenger Pigeon, used to be hunted in phenomenal numbers in the nineteenth century and likewise never recovered after the slaughter finally abated. It seems likely, however, that the loss of its Argentine grassland wintering grounds and North American prairie spring stopover sites was to blame.

Loss of stopover habitat is beginning to supplant hunting (never a convincing case) as the best explanation for the virtual disappearance of the closely related Slender-billed Curlew (N. tenuirostris), which breeds in western Siberia and migrates southwest to the Mediterranean basin, almost certainly using the once extensive east–west Russian steppes along the way.

Other notable examples of this space/time vulnerability include the Thick-billed Parrot (Rhynchopsitta pachyrhyncha), Purple-winged Ground-dove (Claravis go-dejuna), Andean (Phoenicoparrus andinus), and Puna Flamingos (P. jamesi), Lesser Florican (Sypheotides inalica), and Resplendent Quetzal (Pharomachrus mocinno). The parrot shows the same trait as the Passenger Pigeon, since it is dependent on pine seed, a notoriously unpredictable resource. The species is nomadic, but as its native pine forests in Mexico's Sierra Madre are further fragmented, there is a serious danger that a cone-crop failure will leave the last populations “stranded” too far from food for any to survive. The ground-dove specializes on the seeds of forest bamboo in southeast Brazil. So much forest has been destroyed within its range, and bamboo seedset is so temporally patchy, that the species has become one of the rarest in the country.

The flamingos move between lakes in search of an unpredictable resource. The species is nomadic, but as its native pine forests in Mexico’s Sierra Madre are further fragmented, there is a serious danger that a cone-crop failure will leave the last populations “stranded” too far from food for any to survive. The ground-dove specializes on the seeds of forest bamboo in southeast Brazil. So much forest has been destroyed within its range, and bamboo seedset is so temporally patchy, that the species has become one of the rarest in the country.

The flamingos move between lakes in search of appropriate conditions (which shift over time) and are therefore exposed to the possibility that human damage to even a small number of sites may one day leave the species with nowhere to go. The florican selects different grassland sites from year to year in western India, depending on the effects of local rainfall. Grassland is under enormous human pressure in India, and conserving tracts that may be empty of birds for several years at a time is not a simple proposition. Post-breeding
quetzals are now known to make complex short-distance movements to several different areas, so that many more tracts of forest than one or two may be needed to ensure the long-term survival of viable populations of this species.

In Australia, many birds have been affected by human alteration of the natural fire regime. The Paradise Parrot (Psephotus pulcherimus) almost certainly became extinct (it may conceivably survive somewhere) owing to new burning patterns that suited livestock but not the grasses on whose seeds the parrot subsisted. Similar problems affect the food supply of the Golden-shouldered Parrot (P. chrysopsyrargus) and the habitat of the Noisy Scrub-bird (Arrethornis clamosus).

The reason why restriction of range ranks so highly in the list of threats is because any cause of decline is likely to affect the entire species too quickly for human intervention to help. Such species are often restricted to islands, and when those islands are oceanic, the birds have usually evolved in the absence of continental pressures from mammalian predators. As a consequence they are behaviorally and physically adapted in ways that leave them highly vulnerable when continental predators become established within their ranges, through either the direct or indirect agency of humans. These behavioral adaptations are not degenerative.

The Dodo (Raphus cucullatus) was given its name from the Portuguese slang for “stupid,” and indeed is the possessors dearly if less shy creatures have more time to exploit whatever resource is at stake. Nesting in inaccessible places is needlessly expensive if there is no risk in nesting on the ground in the open. Moreover, the relatively stable conditions on tropical islands tend to promote marked K-selected traits (e.g., slow reproductive rates). So the very things that render island birds so vulnerable to aggressive, fast-breeding, continental animals that are entirely tame, or that nest in what to human eyes are ludicrously undefended places, or that have lost the function of their wings are not evolutionary failures. On the contrary, these seemingly disadvantageous attributes are the result of continuing evolutionary pressures. Wings cost energy to carry and maintain, and in the absence of predators they offer no return on the investment in such energy. Shyness costs its possessors dearly if less shy creatures have more time to exploit whatever resource is at stake. Nesting in inaccessible places is needlessly expensive if there is no risk in nesting on the ground in the open. Moreover, the relatively stable conditions on tropical islands tend to promote marked K-selected traits (e.g., slow reproductive rates). So the very things that render island birds so vulnerable to aggressive, fast-breeding, continental animals have actually been selected for through narrower, often intraspecific competition.

Most avian extinctions since 1600 have been on islands (King, in Moors, 1985), and a significant number of threatened birds today are island species. The impact of rats has been and remains massive (see Atkinson, in Moors, 1985): Magenta Petrel (Pterodroma magenta), Zino’s Petrel (P. madrii), Tuamotu Sandpiper (Prosobonia cancellata), Polynesian Ground-dove (Galliminuva erythroptera), Seychelles Paradise-flycatcher (Terpsiphone curvina), and Barosanga Monarch (Pomarea dimidiata) are a few of the Critically Endangered birds whose fate is directly linked to the invasion of their islands by rats. If the cats can also have been introduced, it is not clear which predator is the greater culprit, and these affected species often live on larger islands where very little can be done to help. The New Caledonian Rail (Gallirallus lafresnayi), Cuba’s Zapata Rail (Cyanelmimus crivertii), the Samoan Moorhen (Gallinula sylvestris), and virtually the entire endemic avifauna of Hawaii (the few that it that survives, but notably the honeycreeper, Drepanididae) are good examples of this uncertainty and impotence. Cats alone are responsible for the plight of some species, such as Townsend’s Shearwater (Puffinus auricularis), Socorro Dove (Zenaida graysoni), and the Marquesan Ground-dove (Gallilcomma rubescens). Mongooses, which were often released to devour rats or control snakes, are major threats to species such as Hawaiian Duck (Anas wylieiii) and St. Lucia’s Semper’s Warbler (Leucoptera semperi).

Ancient lakes, like oceanic islands, often harbor endemic faunas and floras that are highly susceptible to exotic introductions. Among the birds, the grebe family Podicepitidae has been particularly hard hit, with the Alaotra Grebe (Tachyhyphus rufolavatus) of Madagascar and Junin Grebe (Podiceps taczanowski) of Peru close to extinction; the Attilan Grebe (Podilymbus gigas) of Guatemala and Colombian Grebe (Podiceps andinus) have already vanished. Linear water bodies may similarly expose certain species to extinction risks from a single event or series of events: Scaly-sided Merger (Mergus squamatus), Brazilian Merger (M. octosetaceus), Wrybill (Anarhynchus frontalis), Rufous-throated Dipper (Cinclus schulzi), and Luzon Water-redstart (Rhynchorus bicolor) are all vulnerable in this way. Moreover, species that concentrate in a small area for even part of their life cycle may be abnormally exposed to danger. For example, the entire world population of Ascension Frigatebird (Fregata aquila) breeds on a single stack smaller than a municipal parking lot. Even birds that are briefly drawn to individual fruiting trees are liable to suffer: as many as 40 Visayan Wreathed Hornbills (Aceros wal- deni) were shot in a single tree over the course of a single day in October, 1997, an event that quite possibly killed 30% of the population.

The foregoing examples represent “spatial” threats. Other threats are better characterized as “temporal,” involving a seemingly innocent event whose conse-
quences cannot be remedied by the time they become apparent. New Zealand’s Kaka (Nestor meridionalis) and Yellowhead (Mohoua ochrocephala) face enormous difficulties now that introduced wasps compete with them for honeydew, an extremely important foodstuff (breeding success in the Kaka is directly correlated with honeydew intake in the previous autumn). The White-headed Duck (Oxyura leucocephala) faces long-term extinction through hybridization with its New World counterpart, the Rudy Duck (O. jamaicensis), which became feral in Britain in the 1960s and is now spreading into its range. In both cases the costs and logistics of eradication are too great to consider.

Sometimes threats come not from exotics but from natives expanding their ranges, often owing to human modifications of habitat. The spread of cowbirds (Molothrus) through the Americas and the Caribbean is particularly worrying. Birds such as Kirtland’s Warbler (Dendroica kirtlandii) and Black-capped Vireo (Vireo atricapillus) need constant-effort programs to reduce cowbird brood-parasitism to tolerable levels. Similarly the spread through the Caribbean of the Pearly-eyed Thrasher (Margarops fuscatus), a nest-hole competitor, has been viewed with alarm by the Puerto Rican Amaz- (Amazona vitata) recovery teams. Analysis of Neotropical data (Collar et al., 1997) shows that the avian families with significantly high numbers of threatened species suffer from particular threats in combination. Thus the parrots (Poicephalidae) have the greatest proportion of threatened birds in any family (28%), closely followed by the curassows and guans (Cracidae, 26%), and then the tinamous (Tinamidae, 13%), still almost double the 8% rate of endangerment in the New World avifauna as a whole. All three families are sensitive to habitat loss; but the parrots also experience intensive trapping for trade purposes, and the tinamous are no less intensively exploited for food. Possession of a distinct economic value within a beleaguered habitat type, particularly one with some (bio-)geographic restriction, confers a strong likelihood of endangerment.

However, there are many endangered birds whose rarity has abidingly obscure causes. Among these are the White-winged Duck (Cairina scutulata, Southeast Asia), Giant Ibis (Pseudibis gigantea, Indochina), Himalayan Mountain-queal (Oxyyysa supersiliosa, India), Boronean Peackock-phantail (Polyplectron schisternacheri, Borneo), Negros Fruit-dove (Ptilinopus arcaus, Ne-gros, Philippines), Blue-headed Lorikeet (Charmosyna tsonopi, Buer, Indonesia), New Caledonian Lorikeet (C. diadema), Forest Owlet (Athene biewitti, India), Liberian Greenbul (Phyllostrepthus leucoplexus), Cone-billed Tana-ger (Conothraupis mevoluca, Brazil), Cherry-throated Tanager (Nemosia rourei, Brazil), Bidadan Maltme (Mul-imbus bidamanesis, southern Nigeria), and Isabella Oriole (Oriolus isabellarius, Luzon, Philippines). All are biffingly rarer—some are known only by a single museum speci-men—than might be inferred from the habitat appar-ently available to them. Unidentified factors must afflict them all, and clearly sometimes natural causes may be in play, particularly unseen ones such as diseases and infestations (see May, in Coulson and Crockford, 1993), which are known to afflict the Iphis Monarch (Pomarea iphis) and Gouldian Finch (Erythura gouldiae). The safest assumption in cases of unexplicable rarity, at least on continents, is that habitat degradation or loss is in some way involved.

IV. APPROACHES TO THE CONSERVATION OF THREATENED BIRDS

A. Research and Synthesis

The primary conservation need of an endangered spe-cies is information. There is a common behavioral trait among academics of ignoring literature much older than 10 years, presumably partly reflecting the assumption that the data contained in such literature are incorpo-rated into more recent work. Often, in fact, the informa-tion has been ignored; sometimes, however, it has been used but, on proper reconsideration, proves to be faulty. There is therefore considerable virtue in seeking out, assembling, and critically evaluating all information re-lating to an endangered species before deciding on the most appropriate remedial or merely investigative ac-tion. Despite the need for speed in cases of species at risk, precipitate intervention can waste hundreds of thousands of dollars and even prejudice attitudes against a species when its real needs are finally recog-nized. For example, captive breeding management for the Philippine Eagle (Pithecophaga jefferyi), a costly long-term, but to date unsuccessful program, appears to have come into being in response to somewhat over-cautious estimates of the bird’s population size.

Recent BirdLife Red Data Books have sought to pre-sent detailed syntheses of relevant data, extending to translations from languages with which biologists may not be familiar, so that a clear picture of a situation can rapidly be considered and the options objectively assessed. However, there are many other examples of the careful construction of evidence in endangered spe-cies management. One of the most notable is the ex-
The conservation of sites at which endangered birds occur, and of the habitat they are known to occupy, is the primary management technique for ensuring their survival. The key tool is the protected area, which for larger sites is usually reflected in law as a national park, indicating the clear public interest of setting aside a major proportion of a country for noneconomic reasons. Smaller areas are often designated as nature reserves, and are frequently considered more as refuges or scientific laboratories for research purposes than as sites with a broader public service, public access can be more difficult than in national parks. However, the size of many national parks is important for conserving viable populations of larger, low-density species. Naturally it is appropriate to seek to save these species where they are sympatric, thereby maximizing the efficiency of the expense. Evaluations such as those by Wege and Long (1995), Stotz et al. (1996), and Stattersfield et al. (1998) provide clear rationales for the targeting of conservation resources in such a way as to secure not just individual species but the key representatives of biogeographic regions.

In general, endangered birds on continents require larger-scale habitat conservation, whereas those on islands need intensive multi-faceted management (aspects of which are treated in Section III,D). The cardinal element in successful site conservation is local support, backed of course by national government. In its work since 1983 to conserve the montane forests of western Cameroon (to which 25 species of bird are endemic, 9 of them threatened), BirdLife International has implemented a major ICDP (integrated conservation and development project) at Mt. Kilum-Ijim to aid local human communities around the mountains, and this has led to a widespread appreciation of the biological value of the forest without generating the kind of hostility that goes with attempting to create strict exclusion zones. The same kind of program operates at Arabuko-Sokoke Forest in Kenya, home to six endangered birds. Other conservation organizations have been using the same fundamental formula—that local people must be made part of the solution, not demonized as part of the problem.

Campaigns to promote interest in and support for species conservation programs are vital elements of those programs. These may take the form of extension work related to ICDP promotion of sustainable use of local resources, or more direct appeals to people to appreciate the unique value of the wildlife in their neighborhood.

B. Site and Habitat Conservation

The conservation of sites at which endangered birds occur, and of the habitat they are known to occupy, is the primary management technique for ensuring their survival. The key tool is the protected area, which for larger sites is usually reflected in law as a national park, indicating the clear public interest of setting aside a major proportion of a country for noneconomic reasons. Smaller areas are often designated as nature or biological reserves, and are frequently considered more as refuges or scientific laboratories for research purposes than as sites with a broader public service, public access can be more difficult than in national parks. However, the size of many national parks is important for conserving viable populations of larger, low-density species. Naturally it is appropriate to seek to save these species where they are sympatric, thereby maximizing the efficiency of the expense. Evaluations such as those by Wege and Long (1995), Stotz et al. (1996), and Stattersfield et al. (1998) provide clear rationales for the targeting of conservation resources in such a way as to secure not just individual species but the key representatives of biogeographic regions.

In general, endangered birds on continents require larger-scale habitat conservation, whereas those on islands need intensive multi-faceted management (aspects of which are treated in Section III,D). The cardinal element in successful site conservation is local support, backed of course by national government. In its work since 1983 to conserve the montane forests of western Cameroon (to which 25 species of bird are endemic, 9 of them threatened), BirdLife International has implemented a major ICDP (integrated conservation and development project) at Mt. Kilum-Ijim to aid local human communities around the mountains, and this has led to a widespread appreciation of the biological value of the forest without generating the kind of hostility that goes with attempting to create strict exclusion zones. The same kind of program operates at Arabuko-Sokoke Forest in Kenya, home to six endangered birds. Other conservation organizations have been using the same fundamental formula—that local people must be made part of the solution, not demonized as part of the problem.

Campaigns to promote interest in and support for species conservation programs are vital elements of those programs. These may take the form of extension work related to ICDP promotion of sustainable use of local resources, or more direct appeals to people to appreciate the unique value of the wildlife in their neighborhood.

C. Trade Controls and International Legislation

Although trade is not a strong factor in the endangerment of birds in general, it is important for a few groups of species, most notably the parrots. The major international instrument for the control of trade is the Washington Convention, universally known as CITES (Convention on International Trade in Endangered Species), which has been in operation since the 1970s. Animals and plants may be registered in three ways: on Appendix I, which essentially prohibits all movement of the species in question; on Appendix II, which prohibits all commercial trade except under license; and on Appendix III, which allows a particular nation to prohibit trade across its borders irrespective of a species’ status elsewhere.

Because of the immense volume of traffic in parrots, and the problem of identification in so diverse a family (around 350 species), all but three species were placed on Appendix II in 1981. This move had the intention if not the effect of giving protection to the more endangered parrots—naturally their rarity increased their desirability among bird-fanciers—which, if listed alone on the Appendix, could easily have been traded indiscriminately under other names without customs officials necessarily being able to identify them. Appendix II species have quotas set by exporting countries, in theory based on data that show the exploitation to be sustainable. Where the evidence suggests that it is not, movement to Appendix I is supposed to bestow immunity.
Curiously—but as a measure of the power that some trade interests can exert—listing on Appendix I can actually stimulate trade (a) while the species is still at the proposal stage for upgrading but also (b) following listing, on the basis of its enhanced rarity value. This happened to the Hyacinth Macaw (Anodorhynchus hyacinthinus), whose rarity in the wild today is primarily a consequence of a trapping blitz in the 1980s. In general, CITES trade controls have some effect, and allow for very useful monitoring of changes in trading fashions over time. However, for truly prized species such as Lear’s Macaw (A. leari) and Spix’s Macaw (Cyanopsitta spixii), both from Brazil, there is little that any control system can do to eliminate smuggling; the financial inducements are simply too strong. The same tends to be true in Indonesia for exquisite songsters like the Straw-headed Bulbul (Pycnonotus sinensis) and good-lookers like the Bali Starling (Leucopar s Rothschildi).

Other important international instruments for endangered birds are the Ramsar Convention (for important wetlands), Bonn Convention (for migratory species), and Berne Convention (for European species). The Convention on Biological Diversity, already mentioned, ought to be the cornerstone of endangered species conservation across the planet, but considerable effort by nongovernmental organizations is needed to ensure the inclusion of global priorities in the national conservation strategies that each party to the convention is obliged to produce and implement.

D. Management Techniques for “Critically Endangered” Birds

1. Habitat Restoration

Major conservation projects almost invariably involve some habitat restoration work. However, for species at the brink of extinction through habitat loss, the emphasis falls more directly on the rapid replanting of food-plants and land areas. Lear’s Macaw appears to be constrained by the availability of licurı palm (Syagrus coronata), many stands of which show no sign of regeneration owing to cattle-grazing; programs have long been planned to establish many new groves within the species’ range. The Cebu Flowerpecker (Dicaeum quadricolor) survives in an area of heavily degraded forest that consists of a mere 3 km²; efforts are under way to reforest adjacent areas within the next 30 years. There is now an emerging discipline of restoration ecology, and it is likely that a great deal of conservation energy in the twenty-first century will be channeled into reconfiguring habitats that were ruined during the twentieth.

2. Control and/or Restriction of Aliens and Natives

Eradication of exotic predators and pests from islands has been pioneered in New Zealand, where several small offshore islands have been rid of various mammals in order both to preserve resident breeding species and to translocate stricken native species from the main islands (Clout and Craig, in Coulson and Crockford, 1995). This painstakingly systematic restoration of islands—mainly involving the elimination of cats, rats, and mustelids—has been essential to the survival of birds such as the Black Petrel (Procellaria parkinsoni) and Kakapo (Strigops habroptilus), and the expertise generated is now being exported to other islands in the Pacific and Indian Oceans.

Researchers on Mauritius recently discovered why the native, Critically Endangered Pink Pigeon (Nesoenas mayeri) and Mauritius Fody (Foudia rubra) have managed to persist in the face of rat predation. These birds now nest almost exclusively in a grove of exotic Cryptomeria japonica, whose bark produces a sticky gum and whose leaves consist of spiny needles, both of which discourage rats from getting at nests. Curiously, then, in rare cases it appears that exotic vegetation can help rather than hinder native species in their struggle against exotic predators; in this instance the planting of more Cryptomeria is clearly called for.

It is not always the case that exotic predators are the problem. In New Zealand, the native Weka (Gallirallus australis) has proved to be a significant influence in depressing the numbers of Little Spotted Kiwi (Apteryx oweni) marooned on Kapiti Island, of Cook’s Petrel (Pterodroma cookii) on Codfish Island (until removed), and of Chatham Oystercatcher (Haematopus charthemensis) on various Chatham islands. On Bermuda, White-tailed Tropicbirds (Phaethon lepturus) outcompete Cahows (Pterodroma caucho) for nests and have to be controlled. As already mentioned, brood-parasitism by cowbirds requires intensive local control efforts to prevent the suppression of breeding success in Kirtland’s Warbler and Black-capped Vireo.

3. Captive Breeding

The role of captive breeding in endangered birds has long been controversial. The prevailing view among conservation biologists is that captive breeding is not a major management tool for endangered birds, and that indeed it can positively distract attention and resources from serious problems that affect the species. The Philippine Eagle has already been mentioned; similarly, cranes and parrots have in the past been launched into
ex situ programs in the mistaken belief that major benefits will result (both species imprint heavily, and parents of which learn survival techniques through observation of their parents, are particularly disadvantaged for return to the wild). Captive breeding has many other drawbacks, notably the particular dangers of disease transmission to wild birds from captive stock (Snyder et al., 1996).

Nevertheless, captive breeding has achieved several outstanding successes. The Northern Bald Ibis (Geronticus eremita) and California Condor (Gymnogyps californianus) both flourish in captivity while efforts to improve environmental conditions in the wild continue. The Lord Howe Rail (Gallirallus sylvicollis) and Guam Rail (G. owstoni), members of a notoriously vulnerable family (since so many rails reached islands in the past and proceeded to lose the power of flight), probably only persist thanks to ex situ regimes. The Socorro Dove (Zenaida graysoni) survives only because a few Californian bird-fanciers took and bred specimens at a time before its native island was overrun by exotic predators. But zoos throughout the world have bred the species (the wild-caught birds flew back to Mexico; the captive-bred ones flew nowhere at all).

There is also a role for veterinarians in intensive in situ management programs, through their expertise in providing appropriate advice on toxicity of nestbox materials, control of nest parasites, composition and secure provision of supplementary foods, disease screening, and minimization in cross-fostering and translocation exercises.

4. Reintroduction and Translocation

The extirpation of a species at a discrete site may have been caused by a short-term or remediable factor, so that restocking with individuals from elsewhere can be undertaken. This is usually relevant or worthwhile only when the species is globally endangered, but whatever the urgency, the endeavor requires careful planning (Black, 1991). A series of feasibility assessments should determine the site’s continuing ecological suitability (the original constraining factors must no longer operate), the threats it faces, the availability of appropriate stock, the socioeconomic implications, and local, national, and international awareness needs. The birds must be in optimal condition, and if captive-bred they should have been reared in disease-free conditions and in such a way as to be behaviorally and genetically fit for independence in nature. Monitoring of the project is crucial, and it should be documented for circulation to other biologists contemplating such efforts.

Bird species that have benefited from well-designed reintroduction programs include the Nene or Hawaiian Goose (Branta sandvicensis), California Condor, Peregrine (Falco peregrinus), and Chatham Islands Snipe (Coenocorypha puillla). Those that have not include Cheer Pheasant (Catras waliwicht) in Margalla Hills, Pakistan (fox predation), Shore Plover on Mangere Island, New Zealand (the birds flew back to South East Island), and Thick-billed Parrots in the United States (the wild-caught birds flew back to Mexico; the captive-bred ones flew nowhere at all).

Translocation to previously unoccupied sites has become a greatly valued technique, if only as a temporary measure while efforts are concentrated on habitat restoration or predator eradication at the native site. Benefits of this approach include Niuafoou Megapode (Megalopelia pachyura), Vint Iorkeets, Kakapo, Guam Rails following captive breeding, Seychelles Magpie-robin (Copsychus sechellarum), and Seychelles Warbler (Acrocephalus seychellenesis). The work done in preparation for the Seychelles Warbler translocations showed that target islands had many times the insect abundance of the host island, Cousin. This was clearly related to the absence of predation pressure on the insect fauna, and—unsurprisingly but still very strikingly—when birds were released onto the target islands they began breeding almost immediately, and very rapidly expanded their numbers. On Cousin the birds had been so packed that they bred only very slowly, with one offspring tending to stay on territory and help at the nest, waiting for a parent to die (Komdeur, 1997). This observation underscores the value of the food resource base in managing endangered birds, and leads to the next point.

5. Supplementary Feeding

The value of increasing food availability for species whose populations need rapid growth seems to have been perceived only relatively recently, perhaps beginning with work on Peregrines (see Temple, 1978). The technique has been adopted for endangered birds on Mauritius, principally with the Mauritius Kestrel (Falco punctatus), for which extra food clearly enhanced reproductive output (Jones et al., in Coulson and Crockford, 1995). New Zealand workers supporting efforts on Mauritius have found the same with the kakapo: normally the species breeds only in response to major masting events, which may occur once every five years, but with dietary supplements it appears to be capable
of breeding every year. On Hokkaido, Red-crowned Cranes (Grus japonensis) have recovered from near-extinction through a combination of better protection and extensive food provision. Clearly, as a relatively short-term measure supplementary feeding can be crucial in producing an unnaturally rapid reproductive output, a very desirable effect in critically low populations where the genetic value of every individual needs to be maximized.

6. Nest-Site Provision or Enhancement
Where nest sites are limiting (often the case with hole-nesting species), the provision of nestboxes or the enhancement of natural cavities may be appropriate. Puerto Rican Amazons ignored the former but benefited from the creation of nesting ledges. The provision of grilles at entrances has helped prevent White-tailed Tropicbirds from appropriating Cahow burrows. In the Galapagos, a new, more secure colony of Dark-rumped Petrels (P. phaeopygia) was established by digging nest burrows and playing calls among them at night.

7. Cross-Fostering and Cross-Breeding
Many large raptors and most cranes lay two eggs but normally rear only one young, so biologists attempting to increase productivity of endangered forms of such birds have long sought to make use of the expendable second eggs by taking them for hatching and captive breeding. In the case of the Whooping Crane (Grus americana), an attempt was made to establish a second population by placing such eggs in the nests of Sandhill Cranes (G. canadensis), but this did not result in a breeding population, possibly as a result of imprinting. The most famous instance of cross-fostering involved the Black Robin (Petronia traversi), whose population fell to five individuals in 1980 but recovered by the placing of some eggs in the nests of Chatham Island Tuft (P. macrocephala).

Cross-breeding with another subspecies was regarded as the last hope of the dusky seaside sparrow (Ammospiza maritimus nigrescens), but the last stock died out before the endeavor could begin. However, it has been used on the endemic Norfolk Island race undulata of the New Zealand boobook (Ninox novaeseelandiae), which by 1986 had been reduced to a single female. After nestboxes were erected to overcome an immediate shortage, two male nominate male boobooks were introduced to the island; one of them paired with the female and in due course produced a string of hybrid offspring. The taxonomic (and legal) implications of this technique may not have been considered, but it clearly represents a means of preserving genetic diversity.

8. The Role of the Concerned Citizen
Finally, it is worth stressing that much of the advocacy for endangered birds, and many of the insights into their plight and salvation, comes from national and international conservation organizations. One of the most valuable things that any sympathetic individual can do to help endangered birds is simply to join or support such organizations. Their conservation effectiveness depends not only on the financial security derived from a broad membership base, but also on being recognized as the representative mouthpieces of a constituency consisting of millions of concerned citizens.

See Also the Following Articles

- BIRDS, BIODIVERSITY OF
- CONTEMPORARY ENDANGERED MAMMALS
- ENDANGERED REPTILES AND AMPHIBIANS
- EXTINCTIONS, MODERN INSTANCES OF

Bibliography


IUCN Species Survival Commission. (1990). IUCN Red List categories, as approved by the 40th meeting of the IUCN Council, Gland,
ENDANGERED ECOSYSTEMS

Raymond C. Nias
World Wide Fund for Nature Australia

I. Introduction
II. Major Ecosystems at Risk
III. Conservation of Ecosystems

GLOSSARY

biological diversity The variety of genes, species, and ecosystems in the living world.
biome Group of ecosystems with similar characteristics, usually a major vegetation type such as tropical rain forest or grassland.
community All of the populations of organisms that live within a given area.
conservation status Relative likelihood of extinction of a species or community.
deforestation Physical removal of trees and consequent reduction in forest cover.
decoregion Geographic region of similar ecological processes and ecosystems.
ecosystem Community of species and their environment.
endangered Very likely to become extinct within the near future.
edemism Degree to which species are unique to a specific location.

THE SCALE OF HUMAN IMPACT ON THE NATURAL ENVIRONMENT has resulted in the decline and loss of many ecosystems. All of the world’s major biomes are to some degree adversely affected by human activities, with forests, scrub and grasslands, freshwater and coastal ecosystems being the most threatened. Specific ecosystems at risk include dry tropical forests, especially on islands, temperate forests, Mediterranean shrublands, temperate grasslands, and coral reefs. Overall, the large marine and oceanic ecosystems, polar regions, and dry deserts are generally less threatened.

I. INTRODUCTION

In terms of area, there seems little doubt that the recent expansion of mechanized agriculture has been the main factor in a rapid conversion of suitable ecosystems into agricultural production. Large areas of forest, woodland, and grassland have been destroyed following the expansion of agriculture, starting in southwestern Asia, China, the Mediterranean, and Europe.

Many temperate ecosystems, such as forests, woodlands, grasslands, and wetlands, are endangered as a result. Similar trends are now obvious in the tropical regions of the world, exacerbated by poor forestry practices and the replacement of forests with cash crops. Even where natural ecosystems have not been subject...
to deliberate modification, human impacts are detectable as a result of exploitation, pollution, and the presence of foreign species.

Ecosystems can be considered endangered when their extent, species composition, or the natural processes that sustain life become sufficiently disrupted or degraded. The difficulty in describing the exact extent and status of an ecosystem, however, makes the assessment of their conservation status much more difficult than it is for a species of organism. In particular, the conservation status of an ecosystem depends greatly on the scale at which the ecosystem is considered. From the major global biomes such as tropical forests, or deserts, to the regional and local scale, various threats operate that affect the likely persistence of that ecosystem, or its components. The ecoregional approach offers good potential for classifying, assessing, and managing ecosystems at the regional scale—combining environmental and geographic attributes to identify specific regions of high biodiversity.

Efforts to conserve ecosystems are being made throughout the world, but these are often hampered by conflicting land-use issues and a general lack of understanding of ecosystem processes. At the global level, international conventions and agreements have established a legal framework for ecosystem conservation. Several countries have begun the task of classifying ecosystems and determining threats. The conservation of natural ecosystems will require coordinated action at a variety of levels, from international to local, and will require a fundamental change in the patterns of resource exploitation and consumption.

A. Definition of an Ecosystem

Nearly sixty years after the development of the concept of an ecosystem, it is now apparent that many have been lost through human activities, or so significantly altered that they may be considered endangered. Our understanding of ecosystems has changed profoundly, with much greater recognition now given to their dynamic and complex nature and the potential for human activities to change their biological and physical components.

The International Convention on Biological Diversity, developed prior to the United Nations Conference on Environment and Development held in 1992, defines an ecosystem as “a dynamic complex of plant, animal and micro-organism communities and their non-living environment interacting as a functional unit.” By itself, however, the term ‘ecosystem’ does not connote any specific dimensions and the boundaries of ecosystems are often indistinct. Even in aquatic ecosystems, where the presence of water helps to identify the lateral boundaries of lakes and rivers, the extent of an aquatic ecosystem will fluctuate with changing water levels, floods, and tides. In the marine environment, ecosystems may be defined by particular attributes of water temperature or chemistry, or by the characteristics of the substrate, in addition to specific plant communities (e.g., seagrass beds or kelp forests). Other ecosystems may be defined by specific geographic or special physical characteristics, such as caves or sea-mounts.

The uncertainty in defining the exact extent of an ecosystem has hampered attempts to quantify their conservation status. Ecosystems can be increasingly finely described and there is no consensus on which level is the most appropriate for conservation. At the broadest level, ecosystems cross international boundaries and are subject to a wide range of threats and human influences. These major ecosystem types can be sub-divided, as they vary enormously in their biotic composition, environment, and functioning, often as a result of major geographic features such as mountain ranges, bays, and river catchments. At a fine level, an ecosystem may comprise a small group of underwater cave systems, sand dunes, an isolated island and fringing coral reef, or a mossy high-altitude swamp that covers no more than a few hectares.

The concept of an endangered ecosystem is relatively new compared to the widely accepted concept of an endangered species. The World Conservation Union (IUCN) has maintained an internationally recognized system of “Red Data Books” (e.g., Walter and Gillett, 1998) for defining and listing threatened species for many years, but no universally applied system has been developed for ecosystems. Similarly, there is no recognized system for measuring the relative conservation status of an ecosystem and it is often not clear in what sense the whole of the ecosystem can be seen to be threatened, especially if many of the resident and visible animals and plants remain reasonably abundant. While it is clear in most cases at what point a species has become extinct, it is much less clear at what point a specific ecosystem ceases to exist.

A paucity of information on many ecosystems, as well as the lack of a standard approach, has hampered attempts at national or global assessments of conservation status. The European Environment Agency assessment (Stanners and Bourdeau, 1993), for example, found that incomplete and fragmented data made it impossible to prepare a complete and systematic assessment of Europe’s ecosystems. Information about nature conservation at the international level was found to be
focused largely on protected areas only and frequently did not cover parameters such as habitat types, human pressures, or type of management that might be important in determining their status.

B. Overviews of Endangered Ecosystems

1. General Reviews

Despite the various problems and general lack of knowledge, there have been several attempts to provide an overview of the conservation status of the world’s ecosystems. The most notable of these overviews have been compiled by the World Conservation Monitoring Centre (e.g., WCMC, 1992) and the United Nations Environment Programme (UNEP) (e.g., Heywood, 1995). Regional assessments of varying thoroughness have been completed for Europe (Stanners and Bourdeau, 1995), United States (Noss and Peters, 1995), and the Asia and Pacific regions (UNESCAP/Asian Development Bank, 1995). Reviews of particular ecosystem types have been completed by a number of agencies. The United Nations Food and Agriculture Organisation, for example, produces a regular update of forest ecosystems (e.g., FAO, 1997), and some coverage of grasslands and other drier ecosystems is provided by UNEP (e.g., Middleton and Thomas, 1997). Marine ecosystems subject to review include mangroves and coral reefs (World Resources Institute, 1996) and oceans (IUCN, 1991). A number of regional freshwater ecosystem reviews have been completed (e.g., Dugan, 1993), and wetland conservation databases are maintained by Wetlands International, World Resources Institute, and other agencies.

2. Endangered Ecosystems in the United States

As one of the most extensively cleared and developed of the world’s large countries, the United States has experienced a dramatic decline in the extent of many ecosystems in the past few hundred years. Using the decline in extent of major vegetation associations, Grossman et al. (1994) described some 371 globally rare terrestrial and wetland plant communities in the United States. Another 482 communities required further definition before they could be mapped and ranked with certainty. Noss and Peters (1995) used this system to identify hundreds of threatened ecosystems. They based their assessment on four factors—if the ecosystems have been greatly reduced since European settlement, if they are currently very small in extent, if they have a number of threatened species, and/or if the continued threat to their existence is high. The greatest reduction in extent has occurred in the prairies and other grasslands, savannas (such as the oak savannas in the Midwest), and some forest and wetland ecosystems. Twenty-seven ecosystem types have lost more than 98% of their original extent, including spruce–fir forest in the southwest Appalachians, pine rockland habitat in south Florida, and mesic coastal prairies in Louisiana, sedge meadows in Wisconsin, and Palouse prairies in the Pacific Northwest (Table 1).

3. Endangered Ecosystems of Australia

In Australia, several studies have also attempted to identify endangered ecosystems, particularly in relation to forests subject to commercial logging operations. The most extensive of these inventories has been carried out as part of the Regional Forest Agreement (RFA) process, which is a national effort to create a system of protected areas within regional forest management plans across Australia’s major commercial forest areas. The RFA process defined an endangered forest ecosystem as one “where its distribution has contracted to less than 10% of its former range or the total area has contracted to less than 10% of its former area, or where

<table>
<thead>
<tr>
<th>Endangered Ecosystems of the United States*</th>
</tr>
</thead>
<tbody>
<tr>
<td>South Florida landscape</td>
</tr>
<tr>
<td>Longleaf pine forest and savanna</td>
</tr>
<tr>
<td>Eastern grassland, savanna, and barrens</td>
</tr>
<tr>
<td>Northwestern grassland and savanna</td>
</tr>
<tr>
<td>California native grassland</td>
</tr>
<tr>
<td>Coastal communities in lower 48 states and Hawaii</td>
</tr>
<tr>
<td>Southwestern riparian forest</td>
</tr>
<tr>
<td>Southern California coastal sage scrub</td>
</tr>
<tr>
<td>Hawaiian dry forest</td>
</tr>
<tr>
<td>Large streams and rivers in lower 48 states and Hawaii</td>
</tr>
<tr>
<td>Cave and karst systems</td>
</tr>
<tr>
<td>Tallgrass prairie</td>
</tr>
<tr>
<td>California riparian forest and wetlands</td>
</tr>
<tr>
<td>Florida scrub</td>
</tr>
<tr>
<td>Ancient eastern deciduous forest</td>
</tr>
<tr>
<td>Ancient forest of Pacific Northwest</td>
</tr>
<tr>
<td>Ancient red and white pine forest, Great Lakes states</td>
</tr>
<tr>
<td>Ancient ponderosa pine forest</td>
</tr>
<tr>
<td>Midwestern wetland</td>
</tr>
<tr>
<td>Southern forested wetland</td>
</tr>
</tbody>
</table>

90% of its area is in small patches which are subject to threatening processes and unlikely to persist.” These studies have resulted in the identification and subsequent protection of many endangered forest ecosystems. Although these efforts have not resolved all outstanding issues relating to forest conservation, or adequately protected all forest types, it has been clear that a rigorous definition and assessment process is a prerequisite for successful conservation of these ecosystems.

A major study has also been undertaken to identify threatened ecosystems in the south-west corner of Western Australia—an area of extremely high plant diversity. Based on a modification of the Conservation Status categories adopted by the IUCN/World Conservation Union (e.g., Walter and Gillett, 1998), the study is one of the first attempts to carry out a scientifically rigorous and repeatable assessment of the status of ecosystems, referred to in the report as “ecological communities.” Criteria are established for each of the IUCN categories: Presumed Extinct (modified to Presumed Totally Destroyed), Critically Endangered, Endangered, Vulnerable, Data Deficient, and Lower Risk. As with the species categories, each ecosystem is assessed against a set of modified criteria relating to the degree to which the original extent of the ecosystem has been reduced, the number of occurrences of the ecosystem and the extent to which it is restricted or isolated, and the degree to which the ecosystem has been modified and could be rehabilitated or restored. A total of 110 ecological communities were identified as being possibly threatened. Sixteen communities were found to be Critically Endangered, 7 as Endangered, 10 as Vulnerable, and 3 as Data Deficient. The most severely threatened ecosystems tended to be forest, woodland, and shrubland communities restricted to specific soil types, and the wetlands associated with these communities.

As with the Regional Forest Agreement process, this approach allows for a more meaningful assessment and, importantly, allows conservation targets to be established as part of the management of the ecosystem. Efforts to define and identify endangered ecosystems have also been advanced by the adoption of national endangered species conservation legislation in Australia (the Endangered Species Protection Act of 1992), which includes a legal definition for endangered ecosystems. Under this legislation, specific ecosystems can be identified and listed on appropriate schedules and guidelines have been established for the assessment of status. Importantly, the legislation recognizes that ecosystems are defined not only by their boundaries and extent, but also by the identity and number of species (structure) and the interactions that link species and their environment (processes).

Box 1


1. Community may already be extinct.
2. Community is subject to current and continuing threats likely to lead to extinction as demonstrated by one or more of:
   (a) Marked decrease in geographic distribution.
   (b) Marked alteration of community structure.
   (c) Loss or decline of native species that are believed to play a major role in the community.
   (d) Restricted geographic distribution such that the community could be lost rapidly by the action of a threatening process.
   (e) Community processes being altered to the extent that a marked alteration of community structure will occur.

C. The Ecoregion Approach

In recognition of the difficulty involved in describing and measuring the status of ecosystems, an approach based on the combination of environmentally and geographically similar features into “ecoregions” has been developed. Olson and Dinerstein (1998) used this approach to identify some 200 of the world’s most significant ecoregions by assessing the number of species and levels of endemism they contained. The advantage of such an approach is that it allows a reasonably precise area to be defined, some level of risk to be determined in relation to the various biological components, and a reasonable basis for comparison, at least for terrestrial ecosystems. Given the dynamic nature of ecosystems, and the fact that they are as much a result of interactions between species as they are of a particular locality, it probably makes more sense to use the ecoregion concept as a basis for comparing conservation status. Much
of the overview of endangered ecosystems in this article is therefore based on the ecoregion concept.

II. MAJOR ECOSYSTEMS AT RISK

A. Forests

1. General Status

It is variously estimated that between 1700 and 1980, the total area of all land remaining in a natural state decreased by more than one-third, from about 95% to about 69%, mostly as a result of expanding human populations and agriculture. In densely populated subregions, such as the western part of Europe and southern China, this level of conversion has been much greater, with less than 30% now remaining. The large-scale loss of ecosystems resulting from human influence can be seen most readily with the conversion of fertile temperate forests for cropland and pasture. Considerable decline in forest area had already taken place by 1700, especially in the Mediterranean Basin and the Indus Valley (areas occupied by ancient Egyptian, Indian, Greek, and Roman civilizations) and in northern and northwestern China. Large declines also occurred in northwestern Europe during the Middle Ages. Primary or old-growth forest (more than 200 years old) is now only a small part of the world’s total forest area. Primary forest cover has been greatly reduced in most industrial countries and is rapidly decreasing in less-developed ones.

In contrast to the Mediterranean region, development in North America did not lead to the almost complete loss of forest. Over some 130 years, an initial rapid conversion of forested land was followed by a slowing down in clearance rates, and eventually by a stabilization. In Australia, some 47% of all forests were cleared over the last 200 years, but mostly during the period of rapid economic expansion after the end of the Second World War. A similar pattern of forest loss is now occurring in the developing (mostly tropical) world. These patterns are again fueled by expanding human populations and agriculture, with the added impetus of industrial crops such as rubber and oil palm.

2. Major Threats to Forests

Deforestation is a serious problem throughout the world’s major forest areas, although its causes and magnitude vary by region. In Africa, for example, the major cause of deforestation is forest clearance for agriculture (particularly commercial farming and to some extent shifting cultivation) and the harvesting of fuelwood. Commercial logging is limited, but settlement and agriculture around roads built to transport timber have resulted in additional clearing of forest areas. The total area of tropical forests in Africa in 1990 was estimated to be approximately 330 million hectares (ha), compared with 369 million ha in 1980, a loss of nearly 7%. The remaining forests in humid West Africa are disappearing at the rate of about 2% a year, and exceeding 5% in the extreme case of Côte d’Ivoire. The relic blocks of forests left at Gola in Sierra Leone, at Sapo in Liberia, and at Tai in Côte d’Ivoire are now of global importance as the last significant remains of the structurally complex and species-rich forests of the upper Guinea zone. Some areas, such as Fouta Djallon, Mount Nimba, and Loma at the head of major water-sheds in western Africa (the Niger, Senegal, and Gambia Rivers), harbor exceptional biodiversity.

Similarly in Latin America, deforestation is due mainly to the expansion of agriculture. This process is driven by a number of factors, the most important being the displacement of peasant farmers from traditional farming areas, large-scale settlement programs such as in Rondonia in Brazil, and the deliberate promotion of livestock-raising activities in the Amazon through financial incentives. In addition, commercial logging, collection of firewood for household use, and road construction have all led to the unsustainable exploitation of the region’s forests. Deforestation has been most severe on the Pacific coast of Central America, where less than 2% of the original 530,000 square kilometers remains, and within the Atlantic forests of Brazil, which now has only 4% of its original 1 million km² as pristine forests. The rapid loss of highly diverse native forests is of particular concern, because they are often replaced by species-poor pastures or by monocultures of exotic timber species. Most of the endangered tropical plants in Brazil (65%) are found in this highly endangered tropical forest ecosystem.

In the Asia and Pacific regions, three countries—Australia, Indonesia, and China—account for 52% of the remaining forest cover. As a result of industrialization, agricultural expansion, and poor forestry practices in many countries, deforestation is one of the major environmental issues in the region. Various agencies, such as the FAO, report that deforestation in the region increased from 2 million ha per year during 1976–1981 to 3.9 million ha per year in 1981–1990, with Bangladesh, Pakistan, the Philippines, and Thailand experiencing the fastest rate of forest clearance. The Pacific subregion has the lowest rate of deforestation (around 130,000 ha per year), most of which occurs in Papua.
New Guinea. Rapid population growth has contributed to depletion of forests not only through land clearing for cultivation but also through over-harvesting of forests for fuelwood, roundwood, and fodder. At the current rate of harvesting, the remaining timber reserves in Asia may not last for more than 40 years.

Although the extent of forest cover in Europe may be significant and stable, or even increasing in parts, many of the elements of the forest ecosystem have been severely modified or lost and for this reason are considered to be threatened. Mature natural forest ecosystems are very rare. In northern and central Europe this rarity has been caused primarily by intensive logging, resulting in a significant decrease in the integrity and natural condition of forests. Forests have been modified significantly to increase their commercial value at the expense of natural ecosystem complexity and structure. They are more likely to contain uniform, relatively even-aged stands, with little dead timber, and with a significantly reduced complement of animal species. In addition, there has been extensive use of nonnative species such as Sitka spruce (Picea sitchensis) in the north and eucalyptus (Eucalyptus spp.) in the south.

The Mediterranean forests have lost much of their natural condition as a result of excessive logging, fire, and overgrazing by livestock, especially goats. Forests are especially threatened at the boundaries of their natural distribution, such as the forest-steppe regions of southern Ukraine and Russia or in the Mediterranean Basin. Most of central Europe’s alluvial forests along all major river courses have virtually disappeared. Airborne pollution (mostly in the form of acid deposition and photochemical smog) is also a major cause of damage to forest ecosystems in Europe, particularly in central and eastern Europe. First reported in the Black and Bavarian forests of Germany in the early 1970s, damage from air pollution may now be detected in a quarter of all European trees.

One of the growing threats to particular forest ecosystems stems from the vast increases in tourism to remote and spectacular regions such as the Himalayas. The fragile high-altitude ecosystems of the Annapurna Himal, Nepal’s most densely populated and ethnically diverse region, have been badly affected by a boom in tourist numbers. To provide trekkers with hot food and showers, the lodges in a single village are obliged to cut one hectare of virgin rhododendron forest each year. Throughout Nepal, about 400,000 ha of forest are cleared each year—an annual loss of 3% of its forests. Between 1975 and 1980, 13% of Nepal’s forests were destroyed. Moreover, each hectare of logged forest results in the erosion of anywhere from 30 to 75 metric tons of soil annually.

3. Endangered Forest Ecosystems

Table II shows examples of some of the most endangered and biologically diverse forest ecosystems in the world. It is clear from this table that many tropical moist forests are endangered, particularly forests in western Africa, the lowland forests of Southeast Asia, and forests on islands such as Madagascar and the Philippines. Rapid development of these regions for agriculture and large-scale commercial logging have been primarily responsible for this extensive forest loss. Large-scale fire now poses a major problem in some of the previously contiguous forests of Indonesia and Malaysia, especially during periods of prolonged drought.

Similar problems face the seasonally dry forests, especially the monsoon forests and conifer forests in the tropical and subtropical regions of the world, especially because these are often more restricted in their original distribution. The forests of Madagascar are of particular concern, as are the dry tropical forests of Central and South America, both of which are biologically distinct and now much reduced. The remaining dry forests of Madagascar represent some of the richest dry forest ecosystems in the world in terms of their natural diversity. Around 165 million years ago, Madagascar drifted away from Africa, which allowed its native species to evolve in isolation from the mainland continent. Among the many unique species are the lemurs—members of the primate family found only on Madagascar and the neighboring Comoros Islands—including the recently discovered mouse lemur (Microcebus myoxinus) and golden-crowned sifaka (Propithecus satraxis), both of which appear to be restricted to this ecosystem. Temperate forests have also been extensively cleared for hundreds of years and few large intact examples remain. Forest loss has been most severe in western Europe, and most examples of this forest type are now restricted to central and eastern Europe. Clearing has also proceeded apace in the temperate forests of China and around the Himalayan region in the past few decades. Most examples of temperate rain forests, including the southern temperate beech forests (Nothofagus spp.) of Chile and New Zealand, are now much reduced in extent.

The Caucasus ecoregion harbors some of the most diverse and distinctive temperate conifer and broadleaf forests in Eurasia, as well as rich woodlands, steppe, and grasslands. One of the world’s seven temperate rain forests formerly occurred here, but this habitat has been
TABLE II

<table>
<thead>
<tr>
<th>Region</th>
<th>Tropical and subtropical moist broadleaf forests</th>
<th>Tropical and subtropical dry and monsoon broadleaf forests</th>
<th>Conifer and temperate broadleaf forests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>Madagascar moist forests</td>
<td>Madagascar dry forests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Guinean moist forests (West Africa)</td>
<td>Maputaland–Pondoland dry forests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eastern Arc monsoon forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>East African coastal forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Albertine Rift highland forests (central Africa)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Seychelles and Mascarene Islands forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gulf of Guinea island forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Macroneum (Atlantic islands) forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Congoleum coastal forests (West Africa)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>Western Ghats moist forests (India)</td>
<td>Western Himalayan temperate forests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sri Lankan moist forests</td>
<td>Central China temperate forests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peninsular Malaysia lowland and montane forests</td>
<td>Eastern Himalayan broadleaf and conifer forests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sumatra–Nicobar Islands forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Northern Borneo–Palawan moist forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Philippines moist forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Southeastern China subtropical forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hainan Island forests (China)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nansei Shoto Archipelago forests (Japan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australasia and Pacific</td>
<td>New Caledonia moist forests</td>
<td>New Caledonia dry forests</td>
<td></td>
</tr>
<tr>
<td>Europe and Russia</td>
<td>New Zealand subtropical forests</td>
<td>Hawaiian dry forests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lord Howe and Norfolk Island forests (Australia)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hawaiian moist forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>South Pacific island forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Brazilian Atlantic forests</td>
<td>Bolivian lowland dry forests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Northern Andean montane forests</td>
<td>Tumbesian and North Inter-Andean Valleys dry forests</td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>Coastal Venezuela montane forests</td>
<td>Southern Mexican dry forests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Greater Antilles moist forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vareta Flooded forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* After Olson and Dinerstein (1998)

virtually destroyed. Lowland forests are dominated by oak, chestnut, and lime, while higher regions are covered by beech, fir, and spruce. Although the ecoregion covers a relatively small area, its varied range of landscapes and climates, and its geographic position at the edge of Europe and Asia, has given the ecoregion a high level of biological diversity, with many European and Asian species represented. In the Caucasus Mountains alone, more than 6000 species of vascular plants have been recorded (of which 20% are endemic), as well as...
B. Scrub, Heath, and Grasslands

1. General Status

Despite their often uniform appearance, low-stature vegetation communities of the type referred to as scrub, heaths, and grasslands are complex and varied, with many regional variations and subtle differences resulting from soil type and climate. They are usually present where soil nutrients, water availability, or climate are not conducive to the growth of forests. They are well suited, however, to grazing by domestic animals and for conversion to cultivation, especially cereal crops. As a result of their suitability for agricultural development, temperate scrub, heaths, and grasslands are among the most modified of the world’s major biomes.

Large areas of current and former temperate scrub and grasslands occur throughout continental Europe and Asia and North America. Significant areas also occur in the Southern Hemisphere, in southern Australia, southern Africa, and temperate South America. In northern polar regions, these communities extend into the Arctic Circle, forming a vast patchwork of semi-frozen and permanently frozen swamps and low vegetation. This latter region appears to be relatively intact, although the future impacts of global warming might considerably alter the distribution and extent of these ecosystems.

European communities of this type range from the Arctic tundra to the Mediterranean shrublands, including many different types of woody shrub communities (including maquis and dwarf shrub heaths), tall herb stands, and many types of grassland. They occur at all altitudes and can range from very sparse to very dense cover, and from a few centimeters to 2 or 3 m in height. Human activities have modified most European scrub and grassland plant communities other than those found on mountain tops or on very poor soils. In some cases, particularly where wetland drainage or forest clearance for timber and farming has taken place, new scrub and grassland ecosystems have been created on lands that were previously wetland or forest.

Closer to the tropics, vast areas of tropical savannas and grasslands can still be found in Africa, Australia, and South America. Large parts of these savannas are used for pastoralism, and the impact of large numbers of grazing animals, often poor soils, changed fire regimes, and human settlements have modified and degraded many areas. As native vegetation cover declines, soil erosion proceeds and turns these semi-arid ecosystems into deserts.

In both hot and cold regions, lack of water typically creates heathlike low scrub, sparse grassland, and desert communities (with characteristic species such as cacti and other succulents), many with a high level of endemism and species richness. Large areas have been cleared for cultivation in North America, Australia, and Europe in regions where rainfall is reasonably predictable.

2. Endangered Scrub, Heath, and Grassland Ecosystems

Table III shows some of the most endangered and biologically diverse ecosystems of this type in the world. Many temperate ecosystems have been lost or so extensively modified that it is difficult to determine their original nature. In Australia, for example, less than 1% of the original temperate lowland grasslands remain, much of this restricted to road sides, railway lines, cemeteries, and lightly grazed unimproved pastures. Some of the world’s most floristically diverse regions, such as the mixed shrublands of the Mediterranean, South Africa, and southwestern Australia are now highly fragmented and suffering from impacts such as dryland salinity and more frequent or intense fires.

Given their suitability for grazing animals and agriculture, it is not surprising that large areas of grasslands have been converted from their natural state. In eastern

8300 species of ferns and 1200 types of lichens and mosses.

The endangered temperate forests found in southern China are also among the world’s most biodiverse temperate regions. Subtropical evergreen broadleaf forest is the dominant vegetation of this ecoregion, which also includes alpine ecosystems, steppe grasslands, and wetlands. Alpine vegetation of the Qinghai–Tibet plateau predominates in the west and the higher regions of the Hengduan Mountains of the southwest. Deciduous broad-leaved forests of the warm temperate zone occur north of the province. Though the biological richness of China and Sichuan is well recognized, international attention has largely focused on a single species that has become a symbol for conservation efforts worldwide—the giant panda (Ailuropoda melanoleuca). Sichuan province is home to the vast majority of the world’s wild pandas. Loss of vegetation cover is one of the main threats to wildlife in the province and half of its panda’s habitat (10,000 km²) was lost between 1974 and 1989 alone. Logging and small-scale agriculture have been converted from their natural state. In eastern

In both hot and cold regions, lack of water typically creates heathlike low scrub, sparse grassland, and desert communities (with characteristic species such as cacti and other succulents), many with a high level of endemism and species richness. Large areas have been cleared for cultivation in North America, Australia, and Europe in regions where rainfall is reasonably predictable.

In both hot and cold regions, lack of water typically creates heathlike low scrub, sparse grassland, and desert communities (with characteristic species such as cacti and other succulents), many with a high level of endemism and species richness. Large areas have been cleared for cultivation in North America, Australia, and Europe in regions where rainfall is reasonably predictable.

In both hot and cold regions, lack of water typically creates heathlike low scrub, sparse grassland, and desert communities (with characteristic species such as cacti and other succulents), many with a high level of endemism and species richness. Large areas have been cleared for cultivation in North America, Australia, and Europe in regions where rainfall is reasonably predictable.
Europe the steppe ecosystems have become extremely rare and large areas of undisturbed steppes are now found only in the south, where they border semidesert regions (north and west of the Caspian Sea, i.e., the Volga Delta and Terek region). The principal cause for the loss of scrub and grassland habitats over the last 50 years has been the advent of widespread agricultural mechanization, which has resulted in the cultivation of large areas of natural and semi-natural grassland. Remaining grasslands are isolated and highly fragmented, and usually restricted to steep slopes and ground with thin soils. Increased use of fertilizers and biocides has further threatened many species. The vast grassy plains of North America have been drastically reduced. The tallgrass prairies of the American Midwest and Great Plains are a unique set of plant communities that once covered vast areas of the continent and supported huge herds of grazing animals such as the bison. Originally thought to be poor agricultural lands, because of the scarcity of trees, the prairie ecosystem was discovered to be immensely fertile, and during the period from the early 1800s to 1930, about 90% of this ecosystem was converted to farmland.

One of the most biologically diverse and endangered of all the world's ecosystems are the fynbos shrublands of the southwestern and southern Cape of South Africa. Although relatively small in extent, the 470 km² of the Cape Peninsula, including Table Mountain, are home to 2236 different plant species, including one of the highest concentrations of endangered plant species. The fynbos is under serious threat from increasing urban expansion (especially in the Cape Flats area) and invasion by exotic weed species, including a number of Australian acacias. Similar shrubland communities can be found along the coastlines of the Mediterranean, southwestern Australia, southern California, and Chile, and in each case they are considered to be threatened.

C. Wetlands

1. General Status

The term "wetland" covers a variety of inland, coastal, and marine habitats that are temporarily or permanently under water. The convention for wetlands of international importance (the Ramsar Convention) defines wetlands as "areas of marsh, fen, peatland or water whether natural or artificial, permanent or temporary, with water that is static or flowing, fresh, brackish or salt, including areas of marine water the depth of which at low tide does not exceed six metres." With such a wide definition and range of ecosystems classified as...
wetlands, and the dynamic nature of floods, tides, and rainfall, it is difficult to accurately gauge the total extent of wetland ecosystems or to precisely define their location. Nevertheless, there have been a number of attempts to classify the world’s wetlands and to provide an overview of their extent and status at the regional level (e.g., Scott and Poole, 1989; WCMC, 1992). In 1971, the Ramsar Convention was signed in the small Iranian town of Ramsar to protect wetlands and promote their “wise use.” Acceptance of the convention has since grown and most countries are now signatories. The existence of the International Wetlands Convention (as it is also known) has resulted in the development of a catalog of many of the world’s most significant wetlands. In addition, the Convention allows the recording of “sites where changes in ecological character have occurred, are occurring or are likely to occur”—the Montreux Record. By 1998, 61 sites were listed on the Montreux Record, including the Florida Everglades (United States), Chilika Lake (India), and the 400,000-ha Shadegan Marshes and mudflats of Khor al Amaya and Khor Musa in Iran.

2. European Wetlands

Given the long agricultural and industrial history of Europe and the Mediterranean region, it is not surprising that most of its wetlands have been significantly modified or lost. Only in the extreme north do large wetland ecosystems remain relatively intact. Of the more than 300 European and Mediterranean wetlands of international importance designated in the Ramsar Convention, 80% are considered to be threatened to some extent. Of major concern has been the continued loss and degradation of freshwater habitats in the Mediterranean Basin, including many inland lakes and coastal lagoons and estuaries. In Spain alone, more than 60% of all inland freshwater wetlands have disappeared during the last 25 years.

The major threats to wetland ecosystems in this region result from human activities, most notably the diversion of water supplies for industry and agriculture, drainage for urban expansion and grazing, and more recently pollution from industrial sources. The use of lake and river water for industrial purposes often affects water quantity (when abstracting water) and water quality (when reintroducing it as wastewater, sometimes polluted or of higher temperature). Many industries, such as pulp and paper production and mining, wash large quantities of particulate matter into lakes and rivers.

The harmful increase in nutrient levels (eutrophication) occurs widely where the catchments of lakes and other wetlands are intensively farmed or densely populated. Fish-farming operations and sewage effluent discharge may also result in the eutrophication of water bodies. As a result of increased nutrient levels (most commonly nitrogen- and phosphate-based fertilizers or animal and human wastes), there is an increase in phytoplankton production and microbial decay with consequent deoxygenation of the water and production of toxins.

The Aral Sea of southwestern Asia is one of the most threatened wetland ecosystems in the world, and will require major international effort and cooperation to restore it to a reasonable state of health. The Aral Sea is fed by the Amu Darya and Syr Darya Rivers, flowing from the mountains of Tian Shan and Pamirs. These waters are the main source of water for Uzbekistan, Kyrgyzstan, Tajikistan, Turkmenistan, and the southern part of Kazakhstan. Traditionally, about half of the water was used for irrigation and the other half flowed into the Aral Sea. Intensive development of cotton and other irrigated crops over the last few decades has reduced the inflow of these rivers to about 3% of former levels. Since 1960 the Aral Sea has lost two-thirds of its volume and its salinity level is now approaching that of seawater. Almost all the native organisms have died out. The 3.3 million ha of exposed seabed have become a source of wind-borne salt and agricultural residues that now spread over surrounding areas. The five republics developed a water distribution agreement in 1992 and an Aral Sea Programme in 1995, which aims to stabilize the region and improve the management of the waters in the basin.

In some cases, the source of wetland pollution may be distant, with contaminants being transported through the air in the form of industrial emissions. One of the best-known examples of this threat to wetland ecosystems was the acidification of European rivers and lakes, especially in the southern parts of Scandinavia. During the 1950s and 1960s, at the peak of sulfur deposition, the rate of acidification was several hundred times that of the natural process and thousands of Scandinavian lakes became too acidic to support their original fish populations. Despite a decline in atmospheric pollution over the last two decades and partial mitigation through liming activities, acidification continues to have detrimental effects on these lake ecosystems.

Marshes, bogs, and fens are a group of vegetated wetlands that were once widespread in Europe. Since early medieval times, however, these habitats have been drained or mined for peat. Though some extensive areas of peatlands and similar ecosystems still exist in the
waters of Florida Bay with silt, which now threatens to blooms have contaminated much of the previously clear a half-century ago. Seagrass die-offs and massive algal glades have shrunk to a tiny fraction of what they were in wading bird populations by some 90%, and the Everglades have been managed successfully and there has been a deal to restore environmental quality. Nutrient levels have been managed successfully and there has been a decline in the concentrations of some toxic chemicals such as persistent organic pollutants derived from pesticides and industrial processes. The flora and fauna of Lake Superior have recovered and are again in good condition. The system remains in a fragile balance, however, and concentrations of many toxic chemicals remain high. Aquatic habitats and wetlands are still in a poor state overall, and the aquatic ecosystems of Lakes Michigan, Ontario, and eastern Erie remain highly modified.

Equally famous are the Everglades of the U.S. Southeast—the largest example of a rare flooded grassland ecosystem. For the past 30 years, south Florida has experienced an accelerated loss of this large and unusual wetland complex. From the Kissimme River, a series of marshy oxbow lakes flow into Lake Okeechobee, one of America's largest freshwater lakes. Okeechobee in turn feeds slowly into the Everglades itself, known as the fabled “River of Grass.” To the south, the shallow estuary of Florida Bay once contained a very rich marine life, and an extensive coral reef system thrived off the Florida Keys. Excessive consumption of water for agriculture and residential populations has led to a decline in wading bird populations by some 90%, and the Everglades have shrunk to a tiny fraction of what they were a half-century ago. Seagrass die-offs and massive algal blooms have contaminated much of the previously clear waters of Florida Bay with silt, which now threatens to

6. Wetlands of Asia

Some of the most thorough reviews of threatened wetlands have been carried out by nongovernmental organizations in East Asia, such as Wetlands International (e.g., Scott and Poole, 1989). Over half of the wetlands of international importance in Asia show a moderate to high degree of threat, with the largest number of threatened sites occurring in China, India, Indonesia, Malaysia, Pakistan, the Philippines, and Sri Lanka.
Hunting and human disturbance are significant threats to wetlands in Asia, as are human settlements, drainage for agriculture, pollution, fishing, and logging.

The rivers, lakes, and other freshwater habitats of the lowland Sundas region (western Indonesia, Malaysia, and Brunei) contain the most significant freshwater diversity of any Asian wetland ecosystem. Forest loss is by far the most serious threat to these freshwater ecosystems as it disrupts the flow of nutrients, leads to water temperature rise because of decreased shading, and lowers concentrations of dissolved oxygen, resulting in conditions that are unsuitable for many fishes and freshwater invertebrates. Road construction, logging, and shifting agriculture, particularly in mountainous regions, result in soil erosion and thus increased sediment loads in rivers, which further reduces water quality. Hunting also represents a major threat to some aquatic species in this ecoregion, such as crocodiles and freshwater turtles. Water pollution is a constant and increasing pressure on many water systems. Part of this problem stems from increased motor traffic on the larger rivers, but chemical runoff from agriculture and the dumping of untreated waste into lakes and rivers are also major concerns. Tourism will also contribute to pollution if it is allowed to develop in an uncoordinated manner and scenic areas such as Lake Toba in northern Sumatra are particularly vulnerable to pollution.

D. Coastal and Marine Ecosystems

1. General Status

Given that up to 60% of the world's human population lives along the coast, it seems inevitable that many highly diverse and productive coastal ecosystems should be particularly threatened. Early human settlements were common along coastlines as they offered a combination of marine and land-based sources of food, navigable rivers, and access to maritime trading routes. With the advent of industrial development, coasts also offered a superficially simple solution to the growing human populations and their wastes, namely, dumping in large coastal rivers. With few areas of the globe now out of reach by modern transportation, even the most remote coastlines are now subject to tourism impacts as travelers seek ever newer opportunities to enjoy unspoiled beaches.

The impacts of human activities on marine and coastal areas fall into three main categories: (1) habitat destruction and degradation through direct activities such as reclamation, coastal constructions, dredging, and damaging fishing techniques; (2) deterioration in quality of habitat and of the water through pollution; and (3) damage to biological communities and living resources through unsustainable levels of harvesting. The World Resources Institute (1996) reported that more than one-third of the world's coastlines are under a high degree of threat from development-related activities. More than 70% of the European coastline is considered to be under a high degree of threat.

2. Major Threats to Coastal and Marine Ecosystems

Threats to coastal and marine ecosystems arise from a wide variety of human impacts and vary considerably in different parts of the world. In the temperate and industrialized regions, areas such as the Wadden Zee along the Netherlands coast show the effects of many centuries of human impact. Much of the biological richness of the Wadden Zee derives from its mudflats (wadens), which appear to be decreasing as a result of the construction of dykes and barriers that affect not only the immediate vicinity but also the flow and direction of water currents and sedimentation patterns. Pollution and eutrophication, caused by synthetic organic substances, heavy metals, oil, and a surplus of nutrients, have been identified as a major threat to the area's wildlife populations. The use of tributyl tin (TBT) as an antifouling agent in paint has resulted in a sharp decline of two species of whelk. Shorebirds are threatened by oil spillages in the Wadden Zee, and pollutants such as chlorinated hydrocarbons have caused reduced breeding success among common tern (Sterna hirundo) populations. High levels of pollutants may also interfere with the immune systems of certain species, and in 1998 a large number of harbor seals died from a viral outbreak. Many of the region's natural resources have been overexploited, and as early as the end of the nineteenth century native oysters began to disappear. Natural mussel beds have also dramatically declined, and in large parts of the Wadden Zee mechanized cockle fisheries continue to disturb the mud surface.

In comparison to the temperate regions, coastlines in much of the tropics are only beginning to be affected by industrial development. More typically, the major threats are rapidly expanding human settlements, the release of untreated sewage, and depletion of natural resources such as mangroves or fish. Coastal marine ecosystems in eastern Africa are biologically diverse, but they are also increasingly threatened by expansion of coastal populations and development. Corals form an almost continuous fringing reef along the entire coastline of eastern Africa. The western coasts of the
islands of Zanzibar, Pemba, and Mafia are characterized by patch reefs, whereas well-developed fringing reefs predominate on the eastern sides. One of the most biologically diverse areas—the Bazaruto Archipelago in Mozambique—supports a range of marine habitats, including deep-sea areas, coral reefs, rocky intertidal areas, sandy beaches, tidal sand flats, seagrass meadows, and mangrove communities. Marine turtles and many species of marine mammals are found in these coastal waters, including the highly threatened dugong (Dugong dugon). The mangroves are spawning and nursery habitats for many fish and crustacean species, and provide an invaluable ecological service by filtering riverine sediments that can damage nearby coral reefs. Environmentally harmful fishing practices, such as dynamite fishing, as well as overfishing, pollution, and uncontrolled tourism development, threaten the survival of these reefs and other coastal ecosystems throughout the region.

3. Endangered Coastal and Marine Ecosystems

a. Sand Dunes

Sand dunes are unique ecosystems occurring at the margins of the land and sea, founded upon the sands that are washed up by the movements of the sea. Sand dunes may stabilize and become forested or covered in shrubs and grasses. Alternatively the dunes may move landward, or be dispersed by subsequent tides. Over time, substantial areas, such as Fraser Island off the eastern coast of Australia, can be formed by this process. In Europe there are less than 428,000 ha of dunelands remaining on Atlantic coasts and widespread afforestation with exotic species has changed their character. The situation is more acute along the southern European coastline, where sandy beaches are the main attraction for many of the 100 million tourists who visit the Mediterranean each year. Construction of coastal towns and resorts has resulted in an estimated 71% loss in dune areas since 1900.

b. Mangroves

Under the right conditions, extensive mangrove forests can be found throughout the tropics in shallow bays and inlets and along rivers. These biologically rich communities have been subjected to enormous pressure from human settlement. By 1996, it was estimated that half of the world's mangroves had been destroyed. The main reasons for this loss in the past have been the felling of firewood or the production of charcoal. More recently, the rapidly expanding shrimp aquaculture industry represents the greatest threat to the world's remaining mangroves. The clearing of mangrove forests to make way for shrimp ponds is responsible for about half of all the mangrove loss in recent years, and has greatly reduced the extent of mangroves in Southeast Asia, South Asia, and Latin America. Thailand has lost more than half of its mangrove forests since 1960 as a result of the booming shrimp aquaculture industry. In Ecuador, estimates of mangrove loss range from 20% to nearly one-half of Ecuador's original 362,000 ha of mangrove forest. Typically, the shrimp ponds are abandoned after a few years and new areas have to be cleared.

c. Coral Reefs

Although coral reefs occupy a small fraction (less than 0.2%) of the total area of the world's oceans, they are among the most biologically diverse of all ecosystems. Formed by the accumulation of calcium carbonate deposited by certain corals and algae over thousands of years, coral reefs cover some 600,000 km² and are the largest structures formed by living creatures. The destruction of coral reefs is caused by many human activities, ranging from coastal development and destructive fishing practices to overexploitation of resources, marine pollution, and runoff from agricultural activities and deforestation. A number of reefs of particular biological interest are also under the most serious threat, including almost all the reefs of the Philippines and coral communities in coastal Indonesia, Tanzania, the Comoros, and the Lesser Antilles in the Caribbean.

An analysis of the conservation status of the world's coral reefs (Bryant et al., 1998) found that 58% were potentially threatened by human activity. Coral reefs in Southeast Asia are the most threatened, where an estimated 80% are at risk as a result of coastal development and fishing-related pressures. The study also concluded that coral reef ecosystems are very poorly conserved, with at least 40 countries lacking any marine protected areas for the conservation of coral reefs. A widespread phenomenon—coral bleaching—became evident during the extreme El Niño weather pattern of 1982-1983. Corals became stressed by high temperatures and expelled the symbiotic zooxanthellae that provide their nutrients, leading to coral death. Large areas of coral bleaching were recorded during this period, as well as in the subsequent El Niño event in the mid-1990s. There is considerable speculation that the severity of these events may be linked to global warming caused by the accumulation of carbon dioxide and other "greenhouse" gases in the atmosphere.
The Nansei Shoto Islands, a chain of 200 islands off southwestern Japan, contain some of the most extensive and biologically diverse coral reefs in the western Pacific. The relative isolation of the islands' marine and terrestrial ecosystems has produced unusual patterns of endemism. On Shiraho reef there are at least 120 species of coral, including the oldest and biggest community of blue coral (Heliofungia actiniformis) in the Northern Hemisphere and extensive colonies of the massive porous coral (Porites australiensis), Madracis coral (genus Madracis), and staghorn coral (family Acroporidae), as well as more than 300 species of fish. These coral reef ecosystems have become seriously threatened since 1972, with up to 90% lost as a result of silt runoff from construction, farming, and logging activities.

d. Island Ecosystems

In many ways islands are among the most vulnerable of ecosystems. Islands are particularly susceptible to invasion by exotic species or rapid depletion of their resources following human colonization. Some island ecosystems have evolved in long isolation from nearby landmasses and may contain many unique plant and animal species. They are also the breeding ground for many marine species, such as turtles, seals, and seabirds, that may be vulnerable to predators on the larger landmasses. In some cases, birds have become flightless as a result of the absence of predators, making them especially vulnerable should humans or exotic species arrive. It is not surprising then that most animal and plant extinctions that have occurred in historic times have been island species (WCMC, 1992). Islands that are of particular concern for both their significant biological richness and vulnerability to human impact are those of New Caledonia, Hawaii, Madagascar, and the Galápagos.

The Galápagos Islands represent one of the most outstanding examples of the evolutionary processes that influence isolated island ecosystems and their susceptibility to human impacts. Across the many individual islands can be found arid lowlands of open cactus forest, transitional subtropical forest, moist dense forest at higher elevations, and treeless upland areas covered with ferns and grasses. Habitat destruction and degradation from grazing by introduced goats, sheep, and cattle, as well as uncontrolled fires, are a major threat. Introduced pest species such as rats and cats prey on native species. Poaching of rare and threatened species is increasing, as is the overexploitation of many marine species such as sea cucumbers (Ischnopoma fuscus) and sharks. Tourism and settlement from the mainland pose additional pressures on an already stressed ecosystem.

4. Oceans

Although the boundaries between marine ecosystems do not appear as distinct as those on land, marine ecosystems may be identified as distinct and moving water masses, often with identifiable and characteristic planktonic assemblages. The sea bottom (benthic systems) can be defined on the basis of sediment type, again associated with characteristic faunal and floral groups, though in some areas other physical and chemical features such as turbidity, light, or salinity help define the ecosystem.

Long considered a virtually infinite resource, the oceans are beginning to show signs of detrimental human impact, including algal blooms, disease outbreaks, and dramatic changes in species composition. In contrast to most terrestrial ecosystems, human impacts on open-water systems and deeper benthic systems are often not specific to any particular ecosystem. Almost no marine ecosystem can be considered immune to the effects of pollutants owing to their dispersal in the marine environment or through the food web. Endangered marine ecosystems are most obvious in those parts of the world subject to extensive industrialization, such as the North Atlantic Ocean and the Baltic, Caspian, Mediterranean, and other seas adjacent to Europe. Other ocean systems that would appear to be at a high risk include the Sea of Japan, South China Sea, Caribbean Sea, and Gulf of Mexico.

Studies of European marine systems show that land-based pollution and diffuse sources of pollution are the major causes in a general deterioration in water quality, from polychlorinated biphenyls (PCBs) in the Baltic Sea, DDT and pesticides in the Caspian, Organic wastes, including sewage, by-products from industries (pulp and paper mills or tanneries), and fertilizers contained in runoff from agriculture, are common throughout European marine ecosystems. Three-quarters of the region's pollution originates from only three countries—France, Spain, and Italy—and causes the contamination of seafood and eutrophication of enclosed bays. Throughout the region, 80% of municipal sewage is discharged untreated from coastal cities, and up to 600,000 million tons of crude oil are released annually from oil-related operations and shipping. Agricultural runoff in the form of pesticides, including persistent organochlorinated compounds, contaminate the food chain, and nitrate and phosphate run-off cause eutrophication and increased episodes of algal growth, including the poisonous “red tides.” Chemical pollution,
bottom-trawling fishing practices, and the introduction of exotic species have contributed to the depletion of indigenous marine life, such as Posidonia (seagrass) meadows.

Increasingly, overfishing is being implicated as a threat to marine ecosystems, particularly in areas such as the North Sea, which are subject to intensive exploitation. Depletion of commercial fish populations may cause changes in the species composition and adversely affect populations of birds, seals, and cetaceans that feed on these species. In addition to depletion of fish stocks, some fishing techniques have a direct impact on nontarget populations. Trawling or dragging of fishing gear, for example, can have devastating local impacts on benthic ecosystems such as the rare Norwegian coral communities (Paragorgia arboresa) or the diverse communities associated with the horse mussel (Modiolus modiolus) at Strangford Loch in Scotland. In many parts of the world, other fishing techniques have also severely damaged marine ecosystems. Drift-netting is particularly notorious for the high by-catch of species such as sea turtles, cetaceans, and seals, whereas long-line fishing has been responsible for the deaths of tens of thousands of albatross and other seabirds in southern oceans. The trawling of seamounts has the potential to destroy these unique ecosystems and their numerous unusual and restricted species.

III. CONSERVATION OF ECOSYSTEMS

There is an urgent need to improve our knowledge of the distribution and conditions of the world's ecosystems. Current national and international initiatives for developing habitat classifications and evaluation methodologies need further coordination. The information that is available points to the inescapable fact that the extent and quality of remaining natural ecosystems are in decline and in certain cases this decline is accelerating. Many ecosystems are at the point of disappearing, or being irreparably modified. Table IV gives a summary of threats to major ecosystem types. It is clear that the major threats arise as the result of (1) the need to continually increase

### TABLE IV

<table>
<thead>
<tr>
<th>Biome</th>
<th>Main threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical moist forests</td>
<td>Clearing for agriculture and plantations, large-scale forestry operations,</td>
</tr>
<tr>
<td></td>
<td>development of roads, towns, and other urban infrastructure, fire.</td>
</tr>
<tr>
<td>Temperate forests</td>
<td>Clearing for agriculture and plantations, large-scale forestry operations,</td>
</tr>
<tr>
<td></td>
<td>development of roads, towns, and other urban infrastructure, fire.</td>
</tr>
<tr>
<td>Boreal forests</td>
<td>Large-scale forestry operations, fire, development of roads, towns, and other</td>
</tr>
<tr>
<td></td>
<td>urban infrastructure.</td>
</tr>
<tr>
<td>Tropical woodlands and</td>
<td>Clearing for agriculture and plantations, irrigated crops, pastoral</td>
</tr>
<tr>
<td>savannas</td>
<td>development and grazing by domestic stock, larval animals (e.g., grubs),</td>
</tr>
<tr>
<td></td>
<td>development of roads, towns, and other urban infrastructure, fire,</td>
</tr>
<tr>
<td></td>
<td>introduced exotic species.</td>
</tr>
<tr>
<td>Mediterranean woodlands</td>
<td>Clearing for agriculture and plantations, irrigated crops, pastoral</td>
</tr>
<tr>
<td>and savannas</td>
<td>development and grazing by domestic stock, larval animals (e.g., grubs),</td>
</tr>
<tr>
<td></td>
<td>development of roads, towns, and other urban infrastructure, fire,</td>
</tr>
<tr>
<td></td>
<td>introduced exotic species.</td>
</tr>
<tr>
<td>Scrubs and grasslands</td>
<td>Clearing for agriculture and plantations, irrigated crops, pastoral</td>
</tr>
<tr>
<td></td>
<td>development and grazing by domestic stock, larval animals (e.g., grubs),</td>
</tr>
<tr>
<td></td>
<td>development of roads, towns, and other urban infrastructure, fire,</td>
</tr>
<tr>
<td></td>
<td>introduced exotic species.</td>
</tr>
<tr>
<td>Deserts</td>
<td>Large-scale irrigation developments, changes in water regimes (e.g., water</td>
</tr>
<tr>
<td></td>
<td>extraction for irrigation), changes in fire regimes, introduced plant and</td>
</tr>
<tr>
<td></td>
<td>animal species, urban encroachment.</td>
</tr>
<tr>
<td>Tundra and ice sheets</td>
<td>Pollution (land and marine), climate change.</td>
</tr>
<tr>
<td>Freshwater wetlands</td>
<td>Water extraction, pollution (especially eutrophication, persistent organic</td>
</tr>
<tr>
<td></td>
<td>pollutants, and heavy metals), drainage and altered flows, large-scale</td>
</tr>
<tr>
<td></td>
<td>dams and weirs, urban development, introduced species (especially nonnative</td>
</tr>
<tr>
<td></td>
<td>fish, invertebrates, and water plants).</td>
</tr>
<tr>
<td>Coasts and shallow seas</td>
<td>Direct impacts from dredging and reclamation, coastal development and urban</td>
</tr>
<tr>
<td></td>
<td>infrastructure (including cities, ports, and harbors), marine and land-based</td>
</tr>
<tr>
<td></td>
<td>sources of pollution, introduced marine organisms, climate change (especially</td>
</tr>
<tr>
<td></td>
<td>for coral reefs and coastal wetlands), overfishing and destructive fishing</td>
</tr>
<tr>
<td></td>
<td>practices (e.g., dynamite and cyanide fishing, bottom trawling).</td>
</tr>
<tr>
<td>Oceans</td>
<td>Marine pollution (including oil pollution and dumping of toxic chemicals),</td>
</tr>
<tr>
<td></td>
<td>overfishing and destructive fishing practices.</td>
</tr>
</tbody>
</table>
human food production, (2) the production of waste from human settlements and industry, (3) the overexploitation of natural resources such as timber and fish, and (4) the accidental or deliberate release of exotic organisms into the environment.

The conservation of ecosystems depends largely on the management of human-induced impacts. Factors that threaten ecosystems can be seen as falling into three categories: factors that reduce the extent of an ecosystem, factors that alter the species composition of an ecosystem, and factors that disturb ecosystem processes.

The clearing of land for agriculture, as well as the draining of wetlands, often results in the almost complete removal of the original ecosystem. Trees and larger plants are felled and removed, and the remaining vegetation is often burned. Such clearing, or deforestation, may be exacerbated by the commercial exploitation of forests and other factors. Wetlands are drained by cutting channels or by damming the supply of water upstream. The dry land that is left behind may be cropped or used for housing and other developments. The almost complete loss of many ecosystems has been a feature of agricultural development in the temperate regions of the world. Evidence of widespread ecosystem disturbances can be seen from the beginnings of modern agriculture in Europe, the Middle East, and North Africa. During this period, large areas of forest were cleared and many wetlands were drained as agriculture and human settlement expanded. Similar patterns can be observed in China and South Asia. During the 1800s and early 1900s, large areas of North America and Australia also became subject to widespread agricultural development and, in the past 50 years, this expansion has spread rapidly into tropical regions. Deforestation in the humid tropics is now of major concern and many tropical forest and wetland ecosystems in central and west Africa, Latin America, and Southeast Asia are endangered. In many cases, the extent of removal is so complete that the original ecosystem must be considered completely destroyed or, at best, confined to small viable remnants. In such circumstances there are few options for conserving the remnants, and areas that have been set aside, such as nature reserves, may provide the only tenuously viable examples of the original ecosystems.

Apart from complete destruction, most of the world's remaining natural ecosystems show some effects of human intervention. Commonly, these are factors that alter the distribution or abundance of species within an ecosystem. Although species populations constantly change, disruptions to these natural cycles can have widespread effects. Certain species, for example, are referred to as "keystone" species because of their vital role in maintaining the populations of other species (e.g., they may pollinate flowers or be large predators). When such changes affect so-called "keystone" species, then larger impacts may be observed in the rest of the ecosystem. Extensive hunting, the harvesting of timber, grazing of livestock, commercial exploitation of fisheries, or other natural resources often affect populations of keystone species and thus alter natural ecosystems, sometimes to the point where they can no longer sustain their original plants and animals. The deliberate or accidental spread of exotic animals and plants also disrupts the ecosystem as these new predators or competitors flourish at the expense of the original species. The pollution of land, water, and air may affect ecosystems by introducing additional nutrients or toxic substances into food chains, causing deaths or reduced reproductive success and consequent changes in species composition. The cumulative impact of these factors can often be observed at some distance or point in time from the original impact. The modification of a river high in the mountains, for example, may eventually affect the water quantity and quality reaching a coastal marine ecosystem, changing that ecosystem and in turn affecting the viability of coastal communities. In such cases it may be possible to manage these uses sustainably, and even to restore and rehabilitate many of the attributes of the original ecosystem.

In extreme cases, the species composition of an ecosystem or its natural processes may be so disturbed from the original that they do not appear to be capable of recovery. Such disturbance may occur, for example, with the loss of nutrients, soils, or water supply. A good example of changes in ecosystems brought about by disruptions in natural processes can be seen in the case of fire. Many ecosystem types (particularly some open forest, shrubland, and grassland communities) are adapted to frequent fires and their species may exhibit a range of responses to fire that ensure their survival and continuation. On the other hand, ecosystems such as temperate and tropical rain forests may be extremely susceptible to fires and the subsequent loss of nutrients that follows heavy rains and leaching.

The effects of climate change will pose new threats to natural ecosystems, particularly those at the edge of their climatic tolerance. Species composition might change rapidly as specialized species are replaced by more generalist species able to take advantage of changing conditions. Of course change is a natural part of evolution and ecosystems should not be considered as
The maintenance, protection, and restoration of degraded ecosystems will depend to a significant extent on the capacity and willingness of people and their governments to reduce and control the impact of human development. There are few proven methods to achieve this outcome, and indeed the trend of humanity over the past millennium has been in the opposite direction—increasing populations, greater levels of per capita resource consumption, and the rapid transformation of natural ecosystems to provide the economic resources necessary for this expansion.

Legal protection and management are clearly necessary in many cases to halt the rapid loss of ecosystem types. Although the first stated goal of the U.S. Endangered Species Act is "to provide the means whereby the ecosystems upon which endangered species and threatened species depend may be conserved," it is clear that even this relatively powerful legislation has only limited ability to achieve the aim of ecosystem conservation. The situation is obviously much more difficult in many poor and developing countries. In many cases the intent of environmental law vastly exceeds the willingness or capacity of the country to effectively implement its provisions. In such cases, international agencies, organizations, and funding bodies are required to help build this capacity and to provide training and institutional strengthening. International treaties and conventions, such as the Convention on Biological Diversity, established as a result of the 1992 Earth Summit, provide an important framework and guidelines for such international cooperation.

The major pressures on ecosystems and their biodiversity arise directly and indirectly from the amount of resources being consumed by a growing human population. Many social and economic factors contribute to this impact. Overall, however, it is the demand for food, water, housing, and material goods, and the necessary economic infrastructure required to produce these goods, that places stress on natural ecosystems. Eventually the expansion of human impact on the environment will need to be modified dramatically if more than a small collection of protected ecosystems are to survive in the future. Controlling our accelerating consumption of natural resources will require fundamental changes in the pattern of human development. There are few proven methods to achieve this outcome, and indeed the trend of humanity over the past millennium has been in the opposite direction—increasing populations, greater levels of per capita resource consumption, and the rapid transformation of natural ecosystems to provide the economic resources necessary for this expansion.

Legal protection and management are clearly necessary in many cases to halt the rapid loss of ecosystem types. Although the first stated goal of the U.S. Endangered Species Act is "to provide the means whereby the ecosystems upon which endangered species and threatened species depend may be conserved," it is clear that even this relatively powerful legislation has only limited ability to achieve the aim of ecosystem conservation. The situation is obviously much more difficult in many poor and developing countries. In many cases the intent of environmental law vastly exceeds the willingness or capacity of the country to effectively implement its provisions. In such cases, international agencies, organizations, and funding bodies are required to help build this capacity and to provide training and institutional strengthening. International treaties and conventions, such as the Convention on Biological Diversity, established as a result of the 1992 Earth Summit, provide an important framework and guidelines for such international cooperation.

The major pressures on ecosystems and their biodiversity arise directly and indirectly from the amount of resources being consumed by a growing human population. Many social and economic factors contribute to this impact. Overall, however, it is the demand for food, water, housing, and material goods, and the necessary economic infrastructure required to produce these goods, that places stress on natural ecosystems. Eventually the expansion of human impact on the environment will need to be modified dramatically if more than a small collection of protected ecosystems are to survive in the future. Controlling our accelerating consumption of natural resources will require fundamental changes in the pattern of human development. There are few proven methods to achieve this outcome, and indeed the trend of humanity over the past millennium has been in the opposite direction—increasing populations, greater levels of per capita resource consumption, and the rapid transformation of natural ecosystems to provide the economic resources necessary for this expansion.

Legal protection and management are clearly necessary in many cases to halt the rapid loss of ecosystem types. Although the first stated goal of the U.S. Endangered Species Act is "to provide the means whereby the ecosystems upon which endangered species and threatened species depend may be conserved," it is clear that even this relatively powerful legislation has only limited ability to achieve the aim of ecosystem conservation. The situation is obviously much more difficult in many poor and developing countries. In many cases the intent of environmental law vastly exceeds the willingness or capacity of the country to effectively implement its provisions. In such cases, international agencies, organizations, and funding bodies are required to help build this capacity and to provide training and institutional strengthening. International treaties and conventions, such as the Convention on Biological Diversity, established as a result of the 1992 Earth Summit, provide an important framework and guidelines for such international cooperation.

The major pressures on ecosystems and their biodiversity arise directly and indirectly from the amount of resources being consumed by a growing human population. Many social and economic factors contribute to this impact. Overall, however, it is the demand for food, water, housing, and material goods, and the necessary economic infrastructure required to produce these goods, that places stress on natural ecosystems. Eventually the expansion of human impact on the environment will need to be modified dramatically if more than a small collection of protected ecosystems are to survive in the future. Controlling our accelerating consumption of natural resources will require fundamental changes in the pattern of human development. There are few proven methods to achieve this outcome, and indeed the trend of humanity over the past millennium has been in the opposite direction—increasing populations, greater levels of per capita resource consumption, and the rapid transformation of natural ecosystems to provide the economic resources necessary for this expansion.

Legal protection and management are clearly necessary in many cases to halt the rapid loss of ecosystem types. Although the first stated goal of the U.S. Endangered Species Act is "to provide the means whereby the ecosystems upon which endangered species and threatened species depend may be conserved," it is clear that even this relatively powerful legislation has only limited ability to achieve the aim of ecosystem conservation. The situation is obviously much more difficult in many poor and developing countries. In many cases the intent of environmental law vastly exceeds the willingness or capacity of the country to effectively implement its provisions. In such cases, international agencies, organizations, and funding bodies are required to help build this capacity and to provide training and institutional strengthening. International treaties and conventions, such as the Convention on Biological Diversity, established as a result of the 1992 Earth Summit, provide an important framework and guidelines for such international cooperation.

The major pressures on ecosystems and their biodiversity arise directly and indirectly from the amount of resources being consumed by a growing human population. Many social and economic factors contribute to this impact. Overall, however, it is the demand for food, water, housing, and material goods, and the necessary economic infrastructure required to produce these goods, that places stress on natural ecosystems. Eventually the expansion of human impact on the environment will need to be modified dramatically if more than a small collection of protected ecosystems are to survive in the future. Controlling our accelerating consumption of natural resources will require fundamental changes in the pattern of human development. There are few proven methods to achieve this outcome, and indeed the trend of humanity over the past millennium has been in the opposite direction—increasing populations, greater levels of per capita resource consumption, and the rapid transformation of natural ecosystems to provide the economic resources necessary for this expansion.
See Also the Following Articles

DEFORESTATION • DIVERSITY, COMMUNITY/REGIONAL LEVEL • ECOSYSTEM, CONCEPT OF • ECOSYSTEM SERVICES, CONCEPT OF • ENDEMISM

Bibliography


ENDANGERED FRESHWATER INVERTEBRATES

David L. Strayer
Institute of Ecosystem Studies

1. Introduction
2. Vulnerability of Freshwater Invertebrates
3. Pressure of Human Activities
4. Number and Distribution of Endangered Freshwater Invertebrates
5. Protection of Endangered Freshwater Invertebrates

GLOSSARY

alien species Species that has been moved and established outside of its native range as a result of human activities; also called exotic species, introduced species, nonindigenous species.
aquifer Geological formation that contains and allows movement of groundwater.
endangered species Species that is at substantial risk of extinction as a result of human activities.
endemic species Species that occurs only over a limited geographical range.
eutrophication Process of increasing the productivity of an ecosystem by enriching it with nutrients.
groundwater Water that occurs in saturated soils and geological formations beneath Earth's surface.

THE EARTH'S FRESHWATERS CONTAIN A BEWILDERING DIVERSITY OF INVERTEBRATE LIFE. More than 70,000 species have already been described, and tens of thousands of species remain to be discovered and described by scientists. This diversity is not spread evenly over the surface of the globe, but is concentrated in local "hot spots," usually geologically ancient lakes, streams, or groundwaters. These hot spots often contain dozens to hundreds of species of freshwater invertebrates that are found nowhere else in the world. Because freshwaters are such an important resource for people, and have been used intensively for water supply, power, irrigation, fisheries, navigation, waste disposal, and as sites for cities, environmental conditions in many of the world's freshwaters have been altered greatly from their original states. Especially where hot spots of diversity coincide with areas of intensive human development, many freshwater invertebrates have disappeared from their native habitats. Some invertebrate species have already become extinct, and thousands of others are in danger of disappearing from the earth. Careful management of
freshwaters, especially in hot spots of high biological diversity, is needed to prevent catastrophic extinctions of freshwater invertebrates in the future.

I. INTRODUCTION

A. Freshwater Invertebrates of the World

Over 70,000 species of freshwater invertebrates have been described, representing about 570 families and 16 phyla. The world's freshwater invertebrate fauna is actually much larger than this, probably 10,000–100,000 species await discovery, and new genera and families are discovered regularly. We are particularly ignorant about what lives in groundwaters, what lives outside of Europe and parts of North America, and small, soft-bodied invertebrates. Especially widespread and species-rich groups of freshwater invertebrates include insects, crustaceans, mollusks, mites, nematodes, and rotifers. Invertebrates live in nearly all kinds of freshwater habitats; lakes, rivers, brooks, ephemeral ponds, wetlands, caves, alluvial groundwaters, and even hot springs each contain a rich and characteristic invertebrate community. Among important freshwater habitats, perhaps only the deepest groundwaters usually lack invertebrates. A typical lake or stream contains a few hundred species of invertebrates representing several dozen families and 8–12 phyla. By comparison with the better-known vertebrates, freshwater invertebrates possess a wide range of biological traits. Life spans range from days to more than a century. Many invertebrates reproduce sexually, others reproduce asexually by budding or parthenogenesis, and still others change their sexuality or mode of reproduction depending on environmental conditions. Some invertebrates produce eggs or other reproductive bodies that remain viable for years to centuries. Freshwater invertebrates include herbivores, bacteriovores, fungivores, predators, and parasites, and exhibit a wide range of specialized morphologies and behaviors to aid in food gathering. Some even use symbiotic algae to photosynthesize!

B. What Is “Endangered”?

Various terms such as “endangered,” “threatened,” “imperiled,” and “at risk” have been used to describe species that are in danger of extinction through human activities. Conservation organizations and governments typically have tried to develop a graded series of carefully defined terms, running from species only remotely threatened with extinction to those on the verge of extinction. Table I shows an example of such a system. In this article, “endangered” is used loosely to mean a species or population that is at substantial risk of becoming extinct over the next few decades as a result of human activities.

C. Causes of Endangerment

The specific causes of endangerment of freshwater invertebrates are highly varied from case to case. It would be impractical (and probably not very illuminating) to discuss all the known cases of endangerment of freshwater invertebrates. Further, because information about the world's freshwater invertebrates is still so incomplete, a catalog of known cases of endangerment many be misleading. Instead, the focus here is on the problem of endangerment in a more general way, and selected case studies are used to illustrate major points.

Endangerment is a product of three factors: the preexisting vulnerability of a species, the pressure of human activities, and the sensitivity of the species to specific human activities (Fig. 1). Thus, a species may become endangered if it already was vulnerable to extinction prior to human involvement, if human activity TABLE I

The Nature Conservancy's System for Ranking the Global Conservation Status of Species

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GX</td>
<td>Presumed Extinct: believed to be extinct throughout its range. Not located despite intensive searches and virtually no likelihood that it will be rediscovered.</td>
</tr>
<tr>
<td>GH</td>
<td>Possibly Extinct: known only from historical occurrences. Still some hope of rediscovery.</td>
</tr>
<tr>
<td>G1</td>
<td>Critically Imperiled: critically imperiled globally because of extreme rarity or because of some factor(s) making it especially vulnerable to extinction. Typically 5 or fewer occurrences or very few remaining individuals (&lt;1000).</td>
</tr>
<tr>
<td>G2</td>
<td>Imperiled: imperiled globally because of extreme rarity or because of some factor(s) making it especially vulnerable to extinction. Typically 6 to 20 occurrences or few remaining individuals (1000 to 3000).</td>
</tr>
<tr>
<td>G3</td>
<td>Vulnerable: vulnerable globally either because very rare and local throughout its range, found only in a restricted range (even if abundant at some locations), or because of other factors making it vulnerable to extinction. Typically 21 to 100 occurrences or between 3000 and 10,000 individuals.</td>
</tr>
<tr>
<td>G4</td>
<td>Apparently Secure: uncommon but not rare, and usually widespread. Possibly cause for long-term concern. Typically more than 100 occurrences globally or more than 10,000 individuals.</td>
</tr>
<tr>
<td>G5</td>
<td>Secure: common, typically widespread and abundant.</td>
</tr>
</tbody>
</table>
II. VULNERABILITY OF FRESHWATER INVERTEBRATES

A. Small Ranges

Many species of freshwater invertebrates had small ranges even before human intervention. Species with small ranges are called narrowly endemic species. For example, half of the 281 North American pearly mussel species were found in only one to three states, even before human intervention. Such species may have had elevated probabilities of extinction through natural catastrophes, and certainly are especially vulnerable to human activities (Fig. 2). Small natural ranges often arise through a small number of understandale processes. Because these processes are focused in certain regions and on species with characteristic biological traits, narrowly endemic species often are clustered together into small regions and concentrated in certain taxonomic or ecological groups.

1. Causes of Small Ranges

A primary cause of small ranges in freshwater invertebrates is the limited dispersal abilities of these animals. All freshwater habitats are islands in a sea of terrestrial habitats, and are more or less isolated from other similar habitats. Although streams are connected into drainage networks, the streams of one drainage network are isolated from those in other drainage networks. The dispersal abilities of freshwater invertebrates, and thus the perceived isolation of freshwater habitats, vary widely.
For animals like dragonflies, whose long-lived aerial adults are strong fliers, or ectoproct bryozoans, whose tough resting stages ("statoblasts") are readily dispersed by migratory waterfowl, the separation of freshwaters probably does not present an important barrier to dispersal or gene flow. For other animals, such as fragile groundwater crustaceans that are poor swimmers, avoid the light, and lack tough dispersal stages, adjacent streams or aquifers may be nearly as remote as distant continents, and even small barriers may prevent migration and gene flow. For instance, the present-day distribution of microparasellid isopods nearly follows the pattern of marine beaches from over 20 million years ago (Fig. 3), where these species presumably arose and from which they subsequently apparently have been unable to disperse.

The isolation of freshwater habitats may produce small ranges in two ways. First, endemic species with small ranges may evolve in place following infrequent crossing of dispersal barriers, resulting in a group of more or less closely related species whose ranges are separated by barriers to dispersal. Second, a formerly widespread species may be eliminated from most of its former range, for instance by a changing climate or the arrival of a competitor, stranded in small refuges, and be unable subsequently to disperse out of the refuges. This second mechanism may become especially important for the freshwater biota if humans cause large changes in regional or global climate, especially because habitat alterations and pollution have eliminated many of the natural dispersal corridors between freshwater habitats.

Finally, a species may have a small range because it requires an unusual habitat, which is itself rare. For example, the thermosbaenacean crustacean Thermosbaena mirabilis was described from ancient Roman warm baths and is known from only a few thermal springs in Tunisia. Its small range presumably derives from its unusual habitat requirements as well as its limited dispersal abilities.

2. "Hot Spots" of High Endemism

Because of processes of speciation, extinction, and dispersal do not occur uniformly over the earth, species richness and endemism vary greatly across the world's freshwaters. Some bodies of water contain more than 1000 invertebrate species, many of them unique to that single body of water. At the other extreme, some bodies...
of water support fewer than 100 invertebrate species, all of them widely distributed. We might expect sites of high richness and endemism to be habitats of great age, habitats where dispersal is limited, either by geographic isolation or by characteristics of the habitat, or habitats that harbor animal groups that are prone to speciate. Thus, many ancient lakes (Baikal in Siberia, Tanganyika and Malawi in Africa) and river systems (the Tennessee in the United States, the Mekong in Southeast Asia) that have not been recently disturbed by glaciation, marine submergence, or desiccation support unique assemblages of invertebrates. Many aquifers seem to contain a high proportion of endemic species, probably because both the characteristics of aquifers (slow water flow, tortuous passageways within aquifers, and barriers between aquifers) and the characteristics of their inhabitants (e.g., fragile bodies, strong thigmotaxis) discourage long-range dispersal. Conversely, glacial lakes and temporary ponds rarely support locally endemic species. Although we know many hot spots of freshwater invertebrate diversity, it probably is not yet possible to produce a reliable global map that shows all major hot spots.

3. Species with High Endemism

Groups of animals vary in their tendency to form new species with small ranges. As already suggested, narrow dispersal probably allows the development of local species, whereas broad dispersal probably provides so much gene flow across populations that speciation is unlikely to occur. Other traits that have been suggested to encourage local speciation include a requirement for outcrossing (as opposed to selfing hermaphroditism or parthenogenesis), production of large young, live-bearing (as opposed to egg-laying), and narrow habitat requirements. The importance of each of these (and other) factors is unclear, but it is clear that groups of freshwater invertebrates do differ widely in their degree of endemism.

Figure 4 shows two examples. Dragonflies, most of which are strong fliers and easily cross drainage divides, are much less likely to have small ranges than pearly mussels, which do not readily cross drainage divides. Even the same group of animals may have dramatically different degrees of endemism depending on the habitat occupied. Thus, groundwater cyclopoid copepods have very much smaller ranges than their relatives in surface waters. This difference presumably arises because groundwater animals have distinctive behaviors and especially because dispersal between aquifers is more difficult than dispersal between lakes or streams.

4. Traits of Narrowly Endemic Species

Narrowly endemic species may possess traits (other than small range size) that influence their vulnerability to endangerment. For example, the limited dispersal abilities and specialized habitat requirements of many narrowly endemic species may make them particularly sensitive to and slow to recover from catastrophes, whether natural or human-caused. Further, some environments that contain endemic species may encourage the development of traits that influence species vulnerability. Thus, in the food-poor groundwater environment, many animals have sparse populations, delayed maturity, and low reproductive rates, all of which probably add to their sensitivity to human impacts.

B. Sparse Populations

Species may be vulnerable to endangerment because their populations are sparse. Because population densities of freshwater invertebrates are much less well known than their geographic ranges, relatively little is
Rivers in the Southeast United States

The U.S. Southeast (extending roughly from the Ohio River south to the coastal plain of Alabama and Georgia, plus the highlands of Arkansas, Missouri, and Oklahoma) contains ancient river systems with an extraordinarily rich biota. This region was not covered by Pleistocene glaciers, nor was it covered by the sea or desiccated for hundreds of millions of years, so river systems like the Tennessee, Cumberland, and Alabama (and their associated aquifers) are very old. The rivers and groundwaters of the Southeast are examples of biologically rich ecosystems that have suffered badly from human activities.

The freshwater invertebrate fauna of the Southeast contains hundreds of species of mollusks, crustaceans, insects, mites, and other animals (Fig. 7) that are found nowhere else in the world. Dozens of genera and two families of invertebrates [Parvidrilidae (Oligochaeta) and Neoplanorbidae (Gastropoda), the latter now probably extinct] are known only from this region. Many of these species have small ranges within the Southeast and may have occurred in only one stream. Despite a long history of scientific study in the Southeast, new species and genera of freshwater invertebrates are discovered regularly in the region.

As is the case for many river systems, the streams and rivers of the Southeast have been profoundly affected by human activities. Impoundments have been especially damaging to the invertebrate fauna. All of the large rivers in the region have been extensively impounded for flood control, hydroelectric power, and navigation, to the point that some of the large rivers have been converted into a continuous series of reservoirs. These reservoir systems differ from natural rivers in their hydrology, temperature, chemistry, sediments, and so on, and often are unsuitable for the native riverine biota. Other physical alterations of stream channels, such as channelization, dredging, diking, and instream gravel mining, also have severely damaged the freshwater biota in parts of this region. Further, as in most developed countries, these river systems have been badly polluted by toxins, nutrients, and sediments from industries, farms, and cities. Coal is mined in parts of the Southeast, which brings acid mine drainage and fine sediments into streams. Finally, invasive species, particularly the zebra mussel and biotic exchanges through the Tennessee–Tombigbee Waterway, may affect the southeastern freshwater fauna.

As a result of these massive changes to southeastern rivers, much of the freshwater invertebrate fauna is extinct or imperiled. Among the mollusks, the only group for which reasonably complete data are available, about 60 species and 4 genera from the Southeast are now extinct. Literally hundreds of additional mollusk species, representing over half of the native fauna, are threatened or endangered. Hundreds of southeastern crayfish and aquatic insects are likewise rare or endangered, and additional species of small, poorly known animals like copepods, isopods, amphipods, and oligochaetes are doubtlessly extant or at risk of extinction.

The southeastern fauna is now receiving some protection from the U.S. Endangered Species Act and parallel state laws. Nevertheless, unless the continuing damaging effects of human activities like impoundments are reversed or remediated, it is difficult to be optimistic about the long-term prospects for the southeastern freshwater biota.

Sources: Benz and Collins, 1997; Lydeard and Mayden, 1995.
Much of the arid interior of eastern Australia is underlain by a large aquifer called the Great Artesian Basin. Freshwater and brackish-water springs occur along the margins of this aquifer in Queensland, New South Wales, and South Australia. These springs range in size from small, moist seeps to large (>100 liters/sec), flowing springs, and some of them have built up large (>10 m high) hills of sand and mineral deposits, and so are locally called “mound springs.” Although perhaps connected to one another in the past, when the Australian climate was wetter, the springs are now separated from one another by a few meters to many kilometers of desert, and are not connected by streams or rivers.

Like many springs in arid regions, the springs of the Great Artesian Basin support animals that live nowhere else in the world. Many of these species are found in only one or a few neighboring springs. Only the fish and the snails of these Australian springs have received serious study. About 25 species and 3 genera of snails have so far been found to be endemic to the springs (Fig. 8). All of the endemic snails belong to the Hydrobiidae, a widespread family that has produced flocks of endemic species in springs, caves, and groundwaterers in the Balkans, the arid Southwest of the United States and Mexico, and elsewhere. Although these snails may be very abundant in the Australian springs (>1,000,000/m²), some species are restricted to one or a few springs, and all are highly vulnerable to human impacts. The endemic fauna is thought to have originated by speciation in the more or less isolated springs, perhaps after a more widespread fauna was stranded in the springs by an increasingly arid Australian climate in the Pleistocene.

The chief threat to the spring fauna is from development of wells in the Great Artesian Basin. Because this is an arid region, there is great demand for water for humans and livestock. When new wells are brought into production, the groundwater level drops, causing springs to dry up. Additional threats include the trampling of springs by livestock, which has badly degraded many springs, conversion of springs into pools by excavation or damming, and introduction of alien species (so far, the mosquitofish, Gambusia affinis, seems to be the only potentially damaging alien in the springs). Over the past two decades, the springs of the Great Artesian Basin have come to be recognized as important habitats for conservation, and steps are being taken to limit at least local impacts from grazing and habitat alterations. Nevertheless, many springs, especially in New South Wales, have dried up as a result of groundwater extraction, and many have been badly altered by livestock or people. It seems likely that at least some of the unique invertebrates of the Great Artesian Basin have gone extinct, and the remaining fauna is at risk of loss.

In arid regions around the world, extreme isolation of aquatic habitats has promoted speciation and development of endemic invertebrate faunas. As in the Great Artesian Basin, water in arid regions is a critically important resource that has been exploited heavily by people. Consequently, freshwater invertebrates of arid regions around the world are endangered by forces similar to those at work in this basin.

**Box 2**

**Mound Springs of the Great Artesian Basin in Australia**

Much of the arid interior of eastern Australia is underlain by a large aquifer called the Great Artesian Basin. Freshwater and brackish-water springs occur along the margins of this aquifer in Queensland, New South Wales, and South Australia. These springs range in size from small, moist seeps to large (>100 liters/sec), flowing springs, and some of them have built up large (>10 m high) hills of sand and mineral deposits, and so are locally called “mound springs.” Although perhaps connected to one another in the past, when the Australian climate was wetter, the springs are now separated from one another by a few meters to many kilometers of desert, and are not connected by streams or rivers. Like many springs in arid regions, the springs of the Great Artesian Basin support animals that live nowhere else in the world. Many of these species are found in only one or a few neighboring springs. Only the fish and the snails of these Australian springs have received serious study. About 25 species and 3 genera of snails have so far been found to be endemic to the springs (Fig. 8). All of the endemic snails belong to the Hydrobiidae, a widespread family that has produced flocks of endemic species in springs, caves, and groundwaterers in the Balkans, the arid Southwest of the United States and Mexico, and elsewhere. Although these snails may be very abundant in the Australian springs (>1,000,000/m²), some species are restricted to one or a few springs, and all are highly vulnerable to human impacts. The endemic fauna is thought to have originated by speciation in the more or less isolated springs, perhaps after a more widespread fauna was stranded in the springs by an increasingly arid Australian climate in the Pleistocene.

The chief threat to the spring fauna is from development of wells in the Great Artesian Basin. Because this is an arid region, there is great demand for water for humans and livestock. When new wells are brought into production, the groundwater level drops, causing springs to dry up. Additional threats include the trampling of springs by livestock, which has badly degraded many springs, conversion of springs into pools by excavation or damming, and introduction of alien species (so far, the mosquitofish, Gambusia affinis, seems to be the only potentially damaging alien in the springs). Over the past two decades, the springs of the Great Artesian Basin have come to be recognized as important habitats for conservation, and steps are being taken to limit at least local impacts from grazing and habitat alterations. Nevertheless, many springs, especially in New South Wales, have dried up as a result of groundwater extraction, and many have been badly altered by livestock or people. It seems likely that at least some of the unique invertebrates of the Great Artesian Basin have gone extinct, and the remaining fauna is at risk of loss.

In arid regions around the world, extreme isolation of aquatic habitats has promoted speciation and development of endemic invertebrate faunas. As in the Great Artesian Basin, water in arid regions is a critically important resource that has been exploited heavily by people. Consequently, freshwater invertebrates of arid regions around the world are endangered by forces similar to those at work in this basin.

**Sources:** Knott and Jasinska, 1998; Ponder, 1986, 1995; Ponder and Clark, 1990; Ponder et al., 1989.

---

**A. Habitat Destruction and Degradation**

Humans have massively altered the physical characteristics of many freshwater ecosystems, usually without consideration for consequent effects on the biota. These physical alterations probably are the chief cause of endangerment of freshwater invertebrates. Dams have been especially damaging (e.g., Fig. 5). Above the dam, running-water habitats are converted into an artificial pool that is usually unsuited to the native invertebrate species. Below the dam, the water temperature and flows of water and sediment often are so altered that downstream reaches support a highly artificial biota as well. Finally, the dam itself is a barrier that blocks normal migrations and movements of the riverine biota. Thus, even a single dam may endanger the riverine biota for hundreds of kilometers, and many river systems are now dismembered by dozens or even hundreds of dams.
shallow-water marginal habitats have been destroyed by dredging, channelization, or filling, or separated from the main channel by dikes and levees. Further, humans often simplify shoreline habitats in both lakes and streams by straightening or filling shorelines and removing trees along and in the water. Because marginal habitats often are important for the feeding and spawning of the freshwater biota, the loss or isolation of these habitats may have grave consequences for the native biota. When water is removed from a river for irrigation or held up in a reservoir for hydroelectric generation or flood control, downstream reaches may dry up or lose critically important floods. Several of the world’s major rivers (e.g., the Colorado and the Ganges) no longer flow to the sea during dry periods. Likewise, the drawdown of many of the world’s aquifers from overuse of groundwater presumably has major effects on the freshwater biota, although these effects have scarcely been studied. Finally, mining of underwater deposits (for gravel or gold, for instance) may have devastating effects at the site of mining, as well as far downstream through sediment transport and far upstream through headcutting of the streambed.

B. Pollution

Water pollution is another widespread activity that has had severe effects on freshwater invertebrates. Rivers and lakes often have been used for waste disposal. These wastes include sewage and other organic matter, the decomposition of which may reduce concentrations of dissolved oxygen to levels too low to support most species of invertebrates. Other wastes include substances (e.g., mercury used in gold mining, acid precipitation from power plants and automobiles) that are directly toxic to freshwater invertebrates. Particularly in industrialized regions, long reaches of streams and rivers have been nearly sterilized of invertebrates as a result of severe, chronic pollution (e.g., Fig. 6).

In addition to pollution caused by deliberate waste disposal, pollution may arise from a wide range of human activities in the watershed. Thus, conversion of forests or native grasslands to agricultural fields or development typically greatly increases loadings of sediments, nutrients, and toxins that are washed in from the altered watershed. These “non-point-source” pollutants are more difficult to track down and control than point loadings of pollutants from factories, yet may have equally serious effects on freshwater ecosystems. While pollution has come under partial control in many developed parts of the world, residual pollution from past releases and inadvertent spills still damage the freshwater biota. Spectacular recent examples include a large spill of pesticides into the River Rhine following a fire at a Sandoz chemical plant, which killed fish and invertebrates for hundreds of kilometers, and the overturning of a truck that spilled a rubber accelerator into the Clinch River, Virginia, which killed most aquatic animals in a 10-km reach, including hundreds of endangered mussels. This was the largest “take” of endangered species in the United States since the Endangered Species Act was passed in 1973. Of course, in many less-developed parts of the world, water pollution is still poorly controlled.
Lake Baikal, part of a rift system in southeastern Siberia, is the oldest (>25 million years old) and deepest (>1600 m deep) lake in the world, and covers 31,500 km². It is the only great rift lake that is oxygenated to its bottom, allowing colonization of the entire lake by a wide range of invertebrates. So far, more than 1400 species of invertebrates have been found in the lake, about 60% of which live nowhere else in the world. Many genera and four families of invertebrates [Labsomirskiidae (Porifera), Baicalarciidae (Turbellaria), Baicaliidae (Gastropoda), Benedictiidae (Gastropoda)] are endemic to Baikal. Probably the most remarkable group of endemic species in Baikal is the huge flock (46 genera and >250 species) of endemic gammarid amphipod crustaceans (Fig. 9). These amphipods, which constitute more than one-third of all gammarid species in the world, have diversified into a wide range of morphologies and behaviors and occupy a range of ecological niches, including planktonic and benthic herbivores, detritivores, predators, and semiparasites on sponges. New species of invertebrates are found regularly in Baikal, so it is clear that the true diversity in the lake is even higher than these figures imply. A distinctive feature of the Baikal fauna is that, though nearshore areas contain a mixture of endemic and widespread species, the open water and abyssal sediments are inhabited chiefly by Baikalian endemics.

In many ancient lakes (e.g., Victoria in Africa, Biwa in Japan, Lanao in the Philippines), pollution, habitat destruction, overfishing, and introductions of alien species have extinguished many endemic species. Baikal has been protected by its remoteness and vast size, and its fauna has so far been relatively unaffected by human activities. Nevertheless, industrial and domestic waste and siltation arising from deforestation of the catchment have polluted nearshore areas. Although lakewide water quality seems not to have suffered yet, pollution is a concern in Baikal because the long residence time of water in the lake means that contaminants entering Baikal may remain in the lake for a very long time. In addition, alien species (e.g., including several fish and the aquatic plant Elodea canadensis) have the potential to affect the nearshore fauna. Lake Baikal is a remarkable example of an ancient lake with a richly endemic invertebrate fauna that may yet be preserved through careful management.


C. Direct Harvest

Some freshwater invertebrates are harvested for human use, which may contribute to their endangerment. Freshwater mussels have been fished for their shells, pearls, and meat since prehistoric times. Especially in the nineteenth and twentieth centuries, harvest rates became so high that many populations were locally depleted in Europe and North America. For example, over 13 million kg of shells from living unionids were taken from Illinois streams and rivers in a single year during peak harvests. Another heavily harvested freshwater invertebrate is the medicinal leech Hirudo medicinalis, which was collected in large numbers in Europe when blood-letting was widely practiced in the eighteenth and nineteenth centuries. As a result of this collecting and widespread pollution and habitat destruction, this animal is now threatened. Harvests of Australian crayfish from the wild are now hundreds of tons per year and have contributed to the endangerment of some species. Other invertebrates are collected for bait or the pet trade, which may contribute to local depletion of populations. Invertebrates sometimes are protected by harvest regulations (e.g., closed seasons, size regulations, bag limits), but such regulations may be inadequately conceived and poorly enforced. Fortunately, for economic reasons, harvests usually (but not always) concentrate on common species rather than rare ones.

D. Alien Species

Humans often move species outside of their native ranges. Such introductions may be deliberate, such as the stocking of trout throughout much of the temperate world, or inadvertent, such as the widespread movement of species in ships’ ballast water. Whatever the cause, these alien species often have strong ecological impacts, and sometimes are responsible for the endangerment of freshwater invertebrates. A spectacular example is the loss of native unionid mussel populations throughout much of northeastern North America as a result of competition with the introduced zebra mussel (Dreissena polymorpha). It is projected that the zebra mussel will be the final blow that will drive several
species of unionids into global extinction. Alien species may also serve as important predators of freshwater invertebrates, as in the case of the brown trout (Salmo trutta), which when introduced to Tasmania apparently preyed on and reduced the range of the unusual and endemic anaspidacean crustacean Anaspides tasmaniae. Because the effects of alien species tend to be cumulative and difficult to reverse, this is a difficult and growing problem in invertebrate conservation.

E. Global Climate Change

Humans have changed the chemistry of the earth’s atmosphere so much that significant changes in global climate are expected in the twenty-first century. At this point, it is difficult to make precise predictions about how these changes will affect specific bodies of freshwater. In many bodies of water the changes may be varied and large, involving such diverse characteristics as temperature, hydrology, water level, rising sea level, stratification, the nature and severity of disturbances, increases in damaging ultraviolet light, water chemistry, riparian vegetation, and food quality. In regions where freshwater becomes scarcer while human demands for water continue to grow, human destruction and degradation of freshwater habitats probably will become more severe. Even though we cannot yet specify the details of global climate change, it is almost certain that this change will endanger or extinguish many freshwater invertebrate species. A rapid, large change in climate will make habitats unsuitable for some of their native species. To survive, such a species will have to disperse to a body of water with suitable ecological conditions. As we have seen, though, the dispersal rates of many freshwater invertebrates are slow, almost surely too slow to keep up with the pace of climate change that current models predict. Further, human modifications to waterways (e.g., impoundments) probably have made long-distance dispersal more difficult.

IV. NUMBER AND DISTRIBUTION OF ENDANGERED FRESHWATER INVERTEBRATES

How many of the world’s freshwater invertebrates are endangered, and where do they live? The most comprehensive list of endangered animals available is the IUCN Red List (compiled by the World Conservation Union), which includes 1151 species of freshwater invertebrates (Table II). Although an enormous amount of work by experts went into compiling the IUCN list, it is clearly incomplete. The list is dominated by large, conspicuous, and attractive animals (mollusks, decapods, dragonflies, and damselflies). There is no reason to believe that smaller and less conspicuous animals are less endangered, but there is simply insufficient information on the status and trends of their populations to identify many endangered species. Likewise, almost 80% of IUCN-listed species are from North and Central America, Australia, or Europe, which probably reflects the geographical distribution of conservation biologists as much as the actual distribution of endangered freshwater invertebrates. Another way to assess global en-

<table>
<thead>
<tr>
<th>Taxonomic Group</th>
<th>Number</th>
<th>Continent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbellaria (flatworms)</td>
<td>1</td>
<td>Total</td>
</tr>
<tr>
<td>Hirudinidae (leeches)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Gastropoda (snails, limpets)</td>
<td>375</td>
<td>North and Central America</td>
</tr>
<tr>
<td>Bivalvia (clams, mussels)</td>
<td>192</td>
<td>South America</td>
</tr>
<tr>
<td>Amphipoda (scuds)</td>
<td>73</td>
<td>Africa</td>
</tr>
<tr>
<td>Anaspidae (anaspidaceans, bathynellaceans)</td>
<td>4</td>
<td>Europe</td>
</tr>
<tr>
<td>Cladocera (water fleas)</td>
<td>8</td>
<td>Asia</td>
</tr>
<tr>
<td>Ostracoda (seed shrimps)</td>
<td>9</td>
<td>Australia</td>
</tr>
<tr>
<td>Copepoda (copepods)</td>
<td>78</td>
<td>Oceanic islands</td>
</tr>
<tr>
<td>Decapoda (crayfish, crabs, prawns)</td>
<td>169</td>
<td>Total</td>
</tr>
<tr>
<td>Isopoda ( sow bugs)</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Ostracoda (seed shrimps)</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Ephemeroptera (mayflies)</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Placopera (nonellus)</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Diaphana (dragonflies, damselflies)</td>
<td>143</td>
<td></td>
</tr>
<tr>
<td>Caliopepida (beetles)</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Trichoptera (caddilflies)</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Diplura (true flies)</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>North and Central America</td>
<td>601</td>
<td></td>
</tr>
<tr>
<td>South America</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>143</td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>147</td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>363</td>
<td></td>
</tr>
<tr>
<td>Oceanic islands</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1151</td>
<td></td>
</tr>
</tbody>
</table>

TABLE II
Numbers of Species of Freshwater Invertebrates Included on the 1996 IUCN Red List of Threatened Animals, by Taxonomic Group and Continent
dangerment of freshwater invertebrates is to examine where damaging human activities coincide with areas of high endemism (Table III). Areas where freshwater invertebrates are especially likely to be endangered include river systems throughout much of the unglaciated world, groundwaters and springs in arid and semiarid regions, and many industrialized areas.

Except in or near areas covered by Pleistocene ice, recently emerged from the sea, or desiccated, river systems often support species that are endemic to that drainage basin. Most river systems have been very highly modified through impoundment and other physical modifications, water withdrawals, and pollution. Thus, we could project that most old river systems probably contain endemic invertebrates, and that many of these invertebrates probably are endangered as a result of human activities (see Box 1 on the U.S. Southeast, for example). Only species that are good dispersers and thus live in many drainage basins or are habitat generalists and thus occur in nonriverine environments are likely to escape endangerment. Human impacts on rivers are almost global, and even river systems that have not yet been heavily impounded and modified (e.g., some basins in Southeast Asia and central Africa) are facing impoundment and other large modifications in the coming decades.

A second area where we might expect to find many endangered invertebrates are groundwaters and springs in arid and semiarid regions (e.g., northern Africa, the American Great Plains and Southwest, and Australia).

These environments typically support highly endemic invertebrate faunas that are probably very sensitive to human impacts. Throughout much of the world, humans living in arid and semiarid regions are pumping water out of aquifers faster than it can be replenished, resulting in large, rapid drops in the water table, which in turn dries up springs and aquifers. We know that desert spring communities are endangered (see Box 2 on Australian springs), and it is possible that aridland aquifers are experiencing large but unseen losses in biodiversity.

In addition to these current threats to freshwater invertebrates, we can expect increasing problems in any regions of rapid human population growth or economic development (e.g., China and Southeast Asia) from the wide range of impacts that typically accompany human populations. Further, any rapid changes in climate probably will cause endangerment and extinction of freshwater invertebrates. These climate changes are projected to be most severe in middle to high latitudes. Because of glaciation, endemism of freshwater invertebrates is higher in midlatitudes than at high latitudes, so impacts of climate change may be most severe at midlatitudes, especially where the freshwater fauna and environments already have been damaged by human activities. Again, the poorly dispersing species will probably be most severely affected.

As to the question of how many freshwater invertebrates actually are endangered (or extinct) globally, there is no certain answer at present. Only for a few

---

**TABLE III**

Expected Global Patterns of Endangerment of Freshwater Invertebrates as a Function of Human Activities

<table>
<thead>
<tr>
<th>Activity</th>
<th>Geographic distribution</th>
<th>Groups of animals affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impoundment</td>
<td>Global, especially North America, China, India, and arid regions</td>
<td>Many, especially migratory species or those that depend on flooding or turbid water</td>
</tr>
<tr>
<td>Physical alterations</td>
<td>Global, especially highly developed regions</td>
<td>Many, especially those that depend on marginal habitats (shallows, floodplains)</td>
</tr>
<tr>
<td>(diking, channel- modification)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water withdrawal</td>
<td>Global in arid and semiarid regions</td>
<td>Many, perhaps especially species living in groundwater and springs</td>
</tr>
<tr>
<td>Toxic pollution</td>
<td>Global, especially industrialized regions</td>
<td>Most species</td>
</tr>
<tr>
<td>Eutrophication</td>
<td>Global, especially densely populated or farmed regions</td>
<td>Species of lake profundal sediments or plant beds</td>
</tr>
<tr>
<td>Harvest</td>
<td>Locally important throughout the world</td>
<td>Brine shrimp</td>
</tr>
<tr>
<td>Alien species</td>
<td>Global in surface waters, perhaps rare in groundwaters</td>
<td>Many</td>
</tr>
<tr>
<td>Climate changes</td>
<td>Global, especially in high latitudes</td>
<td>Many, perhaps especially species that disperse poorly</td>
</tr>
</tbody>
</table>
conspicuous and well-studied invertebrates (e.g., unionoid mussels, odonates) are the IUCN estimates likely to be an accurate reflection of actual endangerments. A reasonable guess might be that 3000–10,000 of the world’s freshwater invertebrate species are extinct or endangered as a result of human activities.

V. PROTECTION OF ENDANGERED FRESHWATER INVERTEBRATES

Although freshwater invertebrates are protected by international, national, and local regulations, this protection often is inadequate, for several reasons. First, lists of protected species usually underlist invertebrates and include only the largest, most conspicuous species. For instance, in the United States, only 111 domestic species of freshwater invertebrates are protected under the Endangered Species Act (compared with 318 species of vertebrates), all but 25 of them mollusks. Thus, many species of endangered freshwater invertebrates, especially small and inconspicuous animals, are not being protected by existing regulations. Second, simply being listed as a protected species may not provide enough help to endangered invertebrates. In many countries, resources for managing endangered species are insufficient, and attention naturally goes to the larger, more charismatic species. In a recent year in the United States, half of all money spent on endangered species was directed to just seven species, all of them vertebrates. As a result, so little money and attention may be spent on invertebrates that plans for recovery of invertebrate species often are general and not pursued aggressively. Third, current approaches to species protection may be inadequate to protect species over the long term. Often, legal protection focuses on trying to prevent further losses from the remaining small populations of an endangered species, without adequate attention to removing the threats that endangered the species in the first place. Consequently, legal protection may slow the rate at which an invertebrate species approaches extinction without reversing its downward trajectory.

How might we more effectively reduce extinction rates of freshwater invertebrates? First, we need to devise ways to protect species without formally listing them. Human activities will endanger or extinguish many species of freshwater invertebrates before we ever
FIGURE 9  Endemic gammarid amphipods from Lake Baikal, showing some of the wide diversity of body forms. Body
lengths are 1.5–6 mm. [From Salemaa, H., and R. Kamaltynov. (1994). The chromosome number of endemic Amphipoda
and Isopoda—An evolutionary paradox in the ancient lakes Ohrid and Baikal. Arch Hydrobiol. Ergb. Limnol. 44, 247–256,
with permission.]
gather enough information on their status to satisfy requirements for legal listing, and before we have enough information on their biology to develop effective species-specific plans to protect them. One way to do this is to take advantage of the fact that many species of endemic freshwater invertebrates co-occur in hot spots by protecting such hot spots from the most damaging of human activities (e.g., dams, excessive water withdrawals, toxic pollution). Such a program of hot spot protection will require better identification, recognition, and protection of hot spots. Nonetheless, it may require less research and provide more effective protection to the world’s freshwater invertebrates than existing species-based programs. Of course, not all endangered species occur in hot spots, so species-based research and protection will necessarily have to accompany any program of hot spot protection. Further, conservation biologists and policymakers need to be more aggressive in identifying and removing threats that endanger species rather than just trying to protect the few populations that have somehow escaped threats. This will require creative thinking about how to preserve or restore essential features of habitat without making unrealistic demands on humans.

A second class of possible solutions could be focused on alleviating the dispersal limitations that are so acute for many freshwater invertebrates by actively establishing new populations of endangered species. This class of solutions is motivated by two main concerns. First, simple protection of existing populations of endangered species may fail to assure long-term survival because natural or human-caused catastrophes (e.g., the Sandefjord spill) or normal population fluctuations ultimately may drive many isolated populations into extinction. Second, it seems likely that global climate change may occur faster than the abilities of some freshwater invertebrates to disperse into suitable habitats. To preserve species under these conditions it may be necessary to deliberately establish populations of endangered species in new locations where suitable habitat exists. Several serious problems attend this approach. First, we currently cannot reliably identify “suitable habitat” for most freshwater invertebrates. Second, we do not have good protocols for reintroductions for most species. Finally, many biologists feel that it is unethical to introduce species outside of their known historical ranges (of course, for many invertebrate species, the known historical range is much smaller than the actual and unknowable historical range).

Despite these problems, it may be necessary to confront the problem of species reestablishments, especially if climate change in the twenty-first century is substantial. Clearly, we will need much better information on how fast species are able to disperse in response to a changing climate (to identify which species, if any, will perish without intervention), practical information on how to establish populations of freshwater invertebrates, and reliable models to predict whether the species we move will have undesirable effects in their new homes.

See Also the Following Articles

ENDANGERED MARINE INVERTEBRATES • ENDANGERED TERRESTRIAL INVERTEBRATES • ENDEMISM • EUTROPHICATION/OLIGOTROPHICATION • HOTSPOTS • INTRODUCED SPECIES • INVERTEBRATES, FRESHWATER, OVERVIEW • LAKE AND FLOOD ECOSYSTEMS

Bibliography


Mammals constitute only about 10% of all vertebrates and less than 0.3% of all known species of organisms, yet endangered mammals (mammals facing an imminent threat of extinction) have received a disproportionate amount of attention and conservation efforts. To a large extent this is the result of the higher visibility and appeal of mammals compared to members of other taxa. This stems in part from our own relationship to this group, both evolutionarily and in the roles that other mammals play in a wide range of human activities from subsistence hunting to sport and commercial activities to domestication for food, work, and companionship.

I. MAMMALIAN DIVERSITY

Mammals can be found from the arctic to the tropics, on every continent, on many of the smaller islands, and in all of the oceans. There are wholly aquatic mammals (Cetacea), semiaquatic mammals, terrestrial mammals, arboreal mammals, fossorial mammals, and even aerial mammals (Chiroptera). Mammals fall into three main groups: the monotremes, who lay shell-covered eggs and have a number of physiological structures in common with reptiles, including a single urogenital opening; the marsupials, noted for a suite of physiological traits including an incomplete placenta, a female bifid...
reproductive tract, and usually an abdominal pouch for the almost-embryonic young, and the eutherians, who have a choriovallantoic placenta and give birth to relatively precocial young. Mammals may live for over a hundred years, as suspected in some cetaceans, or as short as one year, as in the “annual” males of the marsupial Antechinus. Mammalian social systems vary from solitary individuals through a wide range of social organizations to the termite-like sociality of the naked mole rat (Heterocephalus glaber). Mammals can have individual home ranges as small as a few square meters to systems where groups of tens of thousands migrate together (e.g., wildebeest, caribou) and migrations may cover thousands of kilometers (e.g., gray whales). Mammals have evolved to feed on virtually anything that might qualify as edible, from invertebrates, to other mammals, to fungi, grasses, leaves, bark, and even conifer needles. The variety of niches utilized by mammals and their behavioral and physiological adaptations are enormous—it is difficult to think of a niche not already occupied by a mammal, from aquatic grazer (the Sirenia) to aerial piscivore (Noctilio and two other genera of bats).

Not only can mammals be found in most places on earth, many species can have dramatic effects on their ecosystems. Classic examples include the alteration of drainage systems and vegetation patterns caused by beaver dams; the alteration of bush and forest to grassland caused by elephant foraging; the alteration of forest pattern and structure caused by sciurid seed predation, caused by elephant foraging; the alteration of forest pattern and structure caused by grazing from both large and small rodents such as the prairie dogs (Cynomys) and viscachas (Lagostomus maximus). Many of these effects have been shown to increase biodiversity of both plants and animals within the region of modification.

Despite the high visibility and appeal of mammals and their significance to ecosystems, we still know astonishingly little about most species. This is especially true of the smaller, more cryptic, and nocturnal groups, such as the rodents, insectivores, and bats, which make up the bulk of mammalian species diversity. Medellin and Soberon (1999) point out that 459 new species of mammals were described between 1983 and 1993 (partly due to laboratory genetic investigations), with over half weighing under 100 g. For many of these and other species, there is not even basic information on population size, geographic range, or even whether the species is still extant. This also holds true for larger, more conspicuous species. For example, there is almost no information regarding population numbers, specific locations, or even existence for the dhole (Cuon alpinus), a large, pack-living canid, throughout most of its putative range in Asia.

II. A BRIEF HISTORY

A. Early Mammalian History

The history of mammals has been traced back approximately 250 million years to offshoots of the therapsid reptiles in the Triassic. According to limited fossil evidence, the first major radiation of mammals occurred in the Jurassic, about 100 million years ago. Although the diversity of mammals was increasing at this time, with marsupials and placentals in evidence by the early Cretaceous, these early mammals tended to have small and fairly uniform body sizes. A number of hypotheses have been generated to explain both the increase in mammalian diversity and the small body size. The breakup and then reattachment of the continents allowed for isolation, speciation, and then faunal change, and the development of angiosperm plants also undoubtedly resulted in coevolutionary adaptations and speciation by a wide range of mammals. Small body size perhaps was influenced by the dominance of reptiles, which may have forced mammals into a secretive, nocturnal lifestyle where a small body size would be an advantage. An explosive mammalian radiation occurred in the Paleocene, again partly from repeated separations, isolation, and reintroductions as continents continued their movements, and perhaps further spurred by the loss of a wide range of large reptilian competitors during the great extinction event of 65 million years ago.

B. The Pleistocene Extinctions

Between 10,000 and 10,000 years ago a number of sudden and major die-offs occurred, almost entirely among large terrestrial mammals. Approximately 11,000 years ago in North America, a megalaulan extinction event claimed up to 33 genera of large mammals within a span of only 1000 to 2000 years, including mastodons, mammoths, ground sloths, camels, horses, and various large predators such as saber-toothed cats and dire wolves. A similar extinction event involving 46 genera of mammals occurred in South America between 13,000 and 10,000 years ago. In Australia, 13 genera of large mammals disappeared about 30,000 to 20,000 years ago. Two hypotheses have been forwarded to ex-
plain these extinction events. The first is climate change, as in the end of the Ice Age in North America. The second hypothesis relates to the arrival and subsequent hunting of large herbivores by humans, as it appears that all three extinction events occurred at approximately the same time that humans colonized and spread across these continents.

C. Recent Mammalian Extinctions

Many island mammal fauna suffered extinction events similar to those that occurred on the continents during the Pleistocene, although most island extinctions occurred in more recent times. On islands in the Mediterranean, 13 endemic genera vanished, including small goats and even dwarf elephants; most disappeared about 4000 years ago. A dwarf form of the mammoth disappeared from Wrangell Island at about the same time. In Madagascar, eight genera of lemurs, two genera of pygmy hippos, and what is now considered a distinct order, the Bibymalagasia (comprising two species that are variously aligned with the aardvarks and the ungulates), became extinct about 1000 years ago. In each of these cases human activity, either hunting, habitat alteration, or a combination, is thought to be the main cause. Island extinction events have continued into modern times. In the Caribbean Islands, 5 insectivores, 12 rodents, a raccoon, and a seal have all become extinct since 1600—almost a quarter of all recorded mammal extinctions during that period (Fig. 1).

Although the total number of mammalian extinctions in recent times does not appear to be large, the current rate of extinction, about 0.1% per year, is anywhere from 100 to 1000 times greater than what would be predicted from the fossil record. This rate will very likely increase in the near future. The status of 4355 species of mammals was reviewed in the 1996 IUCN Red List of Threatened Animals. Of those reviewed, approximately 40% were considered at risk, from critically endangered to near threatened. This percentage is undoubtedly a low estimate, for as mentioned the status of many mammal species is still unknown.

III. CAUSES AND THREATS

Many of the dangers that threaten other taxa also threaten mammals. These dangers include habitat loss, exploitation, disease, and exotic introductions, but there are some differences in the manner and level of these threats for mammals as compared to other taxa. For example, human persecution and harvesting pressure is probably greater for mammals than for most other taxa, with the probable exception of fish. True mammalian specialists are rare compared to taxa such as insects, where many species are entirely dependent on only one species of plant and thus are vulnerable to habitat changes that alter community structure. However, many mammals are relatively large, and their habitat needs in terms of area are greater than for most other taxa, putting them at high risk from habitat loss and fragmentation.

A number of historical mammalian extinctions have occurred due to human exploitation or introductions rather than habitat loss. Of 18 mammal species that have become extinct since 1600 and whose cause of extinction is known, eight became extinct due to direct human persecution, eight became extinct as the result of introduced predators and competitors, and only two became extinct because of habitat destruction. However, it is apparent that many mammalian species are now under a growing threat from habitat changes. The World Conservation Monitoring Centre (1992) estimated that habitat modification or loss is the major danger for more than three-quarters of all threatened mammals, and this number is likely to grow as the human population continues to increase and expand.

A. Mammalian Physiology and its Relationship to Threats

Mammals constitute a class of organisms that span an enormous size range of approximately eight orders of magnitude, from 2 g in some bats and shrews to 190,000 kg in the blue whale. However, the majority of mammals tend toward large sizes compared to other taxa, and they tend to have a correspondingly long maturation and slow reproductive rate. This puts them at risk from exploitation and other threats, as breeding rates may not be able to keep up with losses in a population. An extreme example is the blue whale (Balaenoptera musculus), whose population was lowered to 1% of its original numbers in less than a century and has not yet shown any appreciable increase, despite complete protection for more than 30 years. Large body size often means large habitat needs, which also puts mammals at risk. Estimates for self-sustaining populations of carnivores such as brown bears (Ursus arctos) and African wild dogs (Lycaon pictus) range from 1000 to more than 10,000 km² of suitable habitat, which means that many protected areas are too small to sustain long-term populations of these species.
B. Habitat Loss, Degradation, and Encroachment

Habitat loss and fragmentation is the single greatest threat to biodiversity worldwide, and this certainly holds true for mammals today. Conversion of habitats by humans into other land uses can fragment and separate mammal populations and increase the likelihood of local population extinctions and eventual species extinction. Rapid deforestation of tropical areas is a growing threat to a number of mammalian species, including many large, wide-ranging, or specialist species of primates, cats, and forest ungulates, as well as numerous small species with restricted ranges such as rodents, insectivores, and marsupials. Most of these species cannot adapt to a highly fragmented or altered landscape, and the few that do adapt often come into conflict with humans by feeding on crops or livestock.

The example of the giant panda (Ailuropoda melanoleuca) shows some of the complexities related to fragmentation and habitat loss. Pandas feed primarily on bamboo that may live for decades but then tends to flower, seed, and die en masse within certain areas. When this happens pandas must switch to other bamboo species, often having to move to new locations to find these alternative food sources. The increase in human population within the panda’s range in China has limited most populations of pandas to very small islands of habitat. A recent seeding and die-off of three species of bamboo resulted in the starvation of over 10% of the world’s remaining wild panda population.

Widely separated and very small populations of pandas may not be viable over the long term, even without the problems faced from the fluctuations in their food source. Because mammals are often relatively poor dispersers, the creation of corridors linking habitats has been suggested as a way to help some species, especially large or wide-ranging (including nomadic or migratory) ones. But for many mammals the necessary size and structure of corridors is unknown, and few management plans have yet to put this idea into practice.

C. Genetic Loss

It is generally assumed that a population’s long-term survival is at least partly dependent on sufficient genetic variation for individual fitness and population adaptability. Loss of genetic diversity and reduced fitness from inbreeding depression and the chance fixation of detrimental alleles has been presumed to reduce adaptive potential and increase the probability of extinction.
However, this has been notoriously difficult to quantify and has not yet been proven to have directly caused a decline in a mammal’s population in the wild. Other events, such as disease or predation, often complicate analysis, especially as lowered fitness may increase an individual’s susceptibility to these factors. There are some examples for which reduced genetic fitness may well be a cause for declines in mammal populations. The genetic bottleneck that occurred in cheetahs (*Acinonyx jubatus*) in the Pliocene has led to extremely low genetic variability among populations in sub-Saharan Africa and is believed to be the cause of sperm aberrations and low sperm counts. Isolation and reduction in numbers leading to inbreeding of Florida panthers (*Puma concolor coryi*) is considered to be the cause of low sperm counts, malformed spermatozoa, and cryptorchidism (undescended testicles) as well as an increase in heart defects within this population.

Another potential genetic threat to endangered mammals is hybridization. Documented cases of hybridizations involving endangered species include red wolves (*Canis rufus*) with coyotes (*C. latrans*), Ethiopian wolves (*C. simensis*) with domestic dogs (*C. familiaris*), and Asian wild ass (*Equus hemionus*) with domestic horses (*E. caballus*). Hybridization with domesticated forms or relatives is also seen as a threat to wild yaks (*Bos grunniens*), Batractan camels (*Camelus bactrianus*), Mongolian wild horses (*Equus ferus przewalskii*), Bactrian camels (*Camelus bactrianus*), and European wild can (*Felis silvestris*), among others.

Another genetic issue is the loss of native species and their replacement by common generalist species characteristic of human landscapes. This can result in a local numerical increase in species diversity but a worldwide loss of genetic diversity.

### D. Livestock and Disease

Domestic livestock and their attendant parasites and diseases have had dramatic effects on some wild mammal populations. Land clearing and overgrazing from domestic cattle, sheep, and goats have altered and fragmented landscapes, increased desertification, and reduced forage for other herbivores. The African rinderpest epizootics of the late 1800s, introduced by cattle, devastated a wide range of wild ungulate populations across the continent. More recently, African horse sickness decimated Indian wild asses, canine distemper killed one-quarter of the lions (*Panthera leo*) in the Serengeti National Park, and all eight packs of wild dogs (*Lycaon pictus*) involved in a study in the Serengeti region disappeared in 1991 with rabies as the suspected or confirmed cause. Disease can spread in the opposite direction, with equally problematic results for wildlife; for example, bison have been shot when they wandered outside of protected areas in Yellowstone National Park in an attempt to control the spread of brucellosis to cattle.

Another factor related to livestock and disease has been the erection of fences to separate wildlife and domestic animals. The results have sometimes been dramatic, such as the case in Botswana, where enormous fences were built in an attempt to control the transmission of disease. These fences blocked the natural migration routes of wild ungulates during the dry season and resulted in the deaths of tens of thousands of red hartebeest (*Alcelaphus buselaphus caama*), blue wildebeest (*Connochaetes taurinus*), and other antelope.

### E. Exploitation

Exploitation of mammals has occurred for many different reasons. Subsistence hunting is a growing pressure due to the explosion of the human population, especially in developing countries where much of the populations’ dietary protein may come from wildlife. Along with the increase in the human population has come an increase in the number of accurate firearms, including automatic weapons, that are now available to subsistence hunters. Subsistence hunting especially threatens ungulates, such as various deer and tapir species in Asia and South America—a third of all deer species and all four tapir species are now considered threatened or near threatened. Subsistence hunting also threatens aquatic mammals whose ranges have already been limited from other causes or who, like stellars and river dolphins, can primarily be found in coastal or riverine areas. Steller’s sea cow (*Hydrodamalis gigas*), the largest and only cold-water sirenian, was hunted to extinction by hungry sailors and other visitors to the north Pacific within 27 years of the first scientific description of this monotypic genus.

Commercial meat hunting has also threatened some species of mammals. The saiga antelope (*Saiga tatarica*), a migratory species of the central Asian steppes that once numbered in the millions, was hunted almost to extinction for its meat, hides, and horns, with the population dropping to a low of about 1000 individuals in the early part of the 1900s.

Until recently the fur trade was an important part of many countries’ economies and cultures. In North America this trade led to the depletion of many fur-bearing species by the early to middle 1800s, including beaver, marten, fisher, otter, and a number of species of seals. Fur hunting for sea otter (*Enhydra lutris*)
caused the population of this species to drop from an estimated 300,000 to under 2,000 by 1911, and hunting for skins was the main cause of depletion in numbers for the now-extinct Caribbean monk seal (Monachus tropicalis). More recently the market for fur resulted in heavy losses for most wild cat species, including large cats such as tigers and jaguars and the smaller spotted cats of South America and Asia. However, it should be noted that other mammal species, especially wide-ranging, abundant, and fast-breeding species such as the muskrat, raccoon, and coyote, have maintained their numbers despite heavy trapping efforts that continue to this day.

A second result of the fur trade was the introduction of fur-bearing animals outside of their ranges, with often unintended results. The accidental release of American mink (Mustela vison) into Europe now threatens the European mink (M. lutreola), as the American species breeds earlier and appears to be outcompeting its European counterpart.

Commercial hunting is not limited to fur. The hunting of whales for oil as well as baleen led to severe depletions of populations of many species in both hemispheres, with some populations driven to extinction and others lowered to below 5% of their original estimated numbers. Although bans have been in place for most species of whales for some time, many populations have yet to show signs of recovery.

Exploitation for body parts used in traditional medicine has also led to the decline in a number of mammal species. Almost every part of a tiger (Panthera tigris) is used for medicinal purposes, including the feet, fat, bones, blood, testes, penis, bile, whiskers, claws, and tail. The Asiatic black bear (Ursus thibetanus) is under great pressure from poaching for its gallbladder, which is used in traditional remedies, and the poaching of rhinos for their horns, which are used both for knife handles and traditional medicines, is the major cause of their precipitous decline. Although habitat loss, trophy hunting, and persecution played an important role in the initial decrease in tigers, bears, and rhinos, poaching for the medicinal market is now the greatest threat to their continued survival.

F. Persecution

Persecution usually takes the form of purposeful attempts to control or extirpate a species. Examples include persecution of potential livestock predators such as wolves and tigers, potential livestock disease threats such as antelope species in Africa and bison in Yellowstone National Park, potential livestock competitors such as prairie dogs in the United States, and, in the case of the American bison in the 1800s, an effort to control and eradicate the indigenous human culture that once depended on this large ungulate. In some cases, control efforts have driven nontarget species to endangered status, such as poison control efforts aimed at coyotes that eradicated the swift fox (Vulpes velox) throughout much of its range, and the prairie dog eradication efforts that eliminated the rodents over 98% of their range and drove the black-footed ferret, a predator specialist on prairie dogs, to the edge of extinction.

G. Interference

Interference can occur in a number of ways. As mentioned, fences have interfered with migratory ungulates in Africa and elsewhere. Dams have negatively affected Ganges and Indus river dolphins (Platanista gangetica and P. minor) by fragmenting populations, increasing siltation, and reducing fish prey. For the critically endangered North Atlantic right whale (Eubalaena glacialis) injuries from boat collisions and entanglement in fishing and lobster gear have accounted for as much as a third of all mortalities in a year, while for manatees in Florida boat collisions have accounted for up to a quarter of all mortality in a year. Death from net entanglement is thought to exceed reproductive rate in Hector’s dolphins (Cephalorhynchus hectori), a small coastal species from New Zealand. Spinner dolphin (Stenella longirostris) populations were significantly reduced due to drowning by tuna seiners who targeted the species because of their association with tuna schools. All three species of spinner are now listed as conservation dependent by IUCN. In Florida, the greatest cause of mortality for the endangered subspecies of key deer (Odocoileus virginianus clavium) and Florida panther is from being struck by cars. Even tourism may pose a threat. For example, tourist vehicle chases in desert environments have led to death from exhaustion for threatened African wild asses (Equus asinus) and addax (Addax nasomaculatus) in northern Africa. The expansion of the whale watching industry has led to concerns that the increasing number of boats, motor noise, and chases may negatively effect whale behavior, including disturbing migratory patterns, breeding efforts, and even separating mothers and young.

War is an extreme example of interference, and it can have an equally dramatic effect on already-rare mammal populations. Although there are arguments that historical tribal warfare may have created source buffer zones for large mammal populations (Martin and Szuter, 1999), today’s warfare results in numerous negative
effects including habitat destruction from deforestation and fighting efforts and the increase in killing of wildlife due to improved hunting efficiency from the availability of modern firearms. This combination is thought to have had an effect on most wild mammals in Vietnam, including the threatened Douc langur (Pygathrix nemaeus). Soldiers and guards may also negatively affect wildlife. For example, the disappearance of most of the remaining population of Marco Polo sheep (Ovis ammon polii) along the China-Pakistan border is attributed to target practice and trophy hunting from armed border forces. War may also cause local increases in subsistence hunting and habitat destruction from desperate indigenous and displaced people, especially in and around formerly protected areas. This appears to be the case in Ethiopia and Somalia, where rare antelope and wild ass are heavily hunted, and in western Africa, where mountain gorillas (Gorilla gorilla beringei) are under mounting pressure from human refugees seeking an escape from fighting as well as food and firewood in protected forests.

**H. Pollution**

Most documented cases of pollution affecting wildlife relate to taxa other than mammals, such as DDT and peregrine falcons or acidification and brook trout. There are cases where pollution has had an effect on populations of threatened or endangered mammals. For example, bats in southwestern North America suffered from DDT spraying, and a number of European bats have been affected by wood preservatives used in buildings and manatees (Trichechus inunguis) have been shown to exist in harbor seals (Phoca vitulina) from the Baltic. Pollution is thought to have led to infections from a calcivirus in California sea lions (Zalophus californianus) that caused abortion and premature parturition. Dall porpoises (Phocoenoides dallii) have been found to have an inverse correlation between serum testosterone level and DDE concentrations in their blubber. Pollution may also have more indirect effects, such as increased mortality of fish prey for river dolphins in the Ganges and Indus.

**I. Introductions**

Introductions of exotics, both purposeful and accidental, have had very strong impacts on some native species of mammals. These impacts may involve predation, competition, and the concurrent introduction of diseases or parasites. Most documented introduction problems have involved mammals as the pest species and have taken place on islands. Cases include the introduction of goats, sheep, and other herbivores to islands where endemic animals lack defenses against herbivory, and the introduction of omnivores such as rats, pigs, and macaques or carnivores such as mongooses and stoats to islands where endemic animals lack defenses against predators. The introduction of rabbits (Oryctolagus cuniculus) to Laysan appears to have driven three species of birds to extinction from grazing that destroyed avian food sources, both plants and insects dependent on the plants. The results from the explosive population growth of rabbits introduced into Australia is well documented, although predation by nonnative red foxes probably has had a greater negative effect on indigenous marsupials. Both pet and feral cats and dogs also pose a threat to wildlife. Feral cats in Australia are estimated to kill more than 400 million native mammals, birds, and reptiles each year, while in Argentina predation by dogs is the biggest threat to the endangered pink fairy armadillo (Chlamyphorus truncatus). Introduced aoudad (Ammotragus lervia) and feral burros in the southwest United States are considered to be potential competitors and disease vectors for the desert bighorn sheep (Ovis canadensis nelsoni). Introduced deer species have had major effects on vegetation in New Zealand, and in South America they may be serious competitors of the endangered Andean huemel (Hippocamelus brasiliensis).

**J. Multiple Threats**

For many species, a series or combination of the factors listed have led to precipitous population declines and endangerment. For the tiger, trophy hunting, eradic-
tion programs, market exploitation for fur and traditional medicines, and loss of habitat all played a role in the extinction of three races and a decline from perhaps over 100,000 tigers worldwide in 1920 to as few as 3000 today.

Even for mammals that have recovered from previous threats, new factors may arise to threaten the species. The saiga antelope recovered from commercial hunting once a ban and then regulations were put in place, rebounding from a low of only a thousand to more than 2 million in just 30 years. However, overgrazing, desertification, migratory barriers, and poaching for medicinal products now threaten the species, and the population has dropped to around 1 million animals in four widely separated regions.

IV. Locations for Endangered Mammals

A. Geographic Analysis of Endangered Mammals

As with most other taxa, mammalian species diversity tends to increase with decreasing latitude. The top eleven countries for mammal diversity are Mexico (450), Indonesia (436), United States (428), Zaire (415), Brazil and China (394), Colombia and Kenya (359), Peru (344), Tanzania (322), and India (316). Only the United States is not found entirely or primarily within the tropical zone. The number of threatened mammals is shown in Table I, with Indonesia leading the list with 128 and China and India tied for second with 75 each. Most of the listed countries also have some of the highest human densities in the world. However, some of these countries, such as Australia, have high numbers of threatened species yet relatively low human densities. In most of these cases a second factor, endemism, can explain much of the data.

B. Endemism: The Case of Australian and Island Mammals

Oceanic islands tend to have depauperate mammalian faunas, primarily because terrestrial mammals make poor oceanic dispersers. Many islands also do not have enough habitat to maintain viable populations of terrestrial mammal species. However, on islands that terrestrial mammals were able to colonize and that were large enough to sustain populations, isolation often led to speciation. Some examples of island endemism are truly spectacular, and often these locations also contain some of the most threatened species in the world. Indonesia, an archipelago consisting of more than 13,000 islands, has the second-greatest species richness in the world and also has the greatest number of endemic mammal species with 201. Indonesia also has the greatest number of threatened mammals with 128 (29% of its mammalian fauna). In Australia, out of a total of 252 species of mammals, 198 species are considered endemic. Twenty species are recorded as extinct since 1600 and, of the remaining species, 58 (23%) are considered threatened with extinction. Madagascar is the fourth largest island in the world and has been separated from mainland Africa since the Cretaceous. Because of its long isolation, all but one of the terrestrial species are endemic forms (the exception is a pig that may have been introduced). There are five families and 13 genera of living endemic primates (the lemurs), an endemic family of insectivores (the Tenrecidae), three endemic subfamilies of the carnivore family Viverridae, at least eight endemic genera of rodents, and a family of bats (the Myzopodidae). Of the endemic species, 46 (49%) are currently threatened, many from the recent loss of 90% of the island’s original forest cover. Papua New Guinea has 214 species of mammals, of which 57 are endemic; 57 species are also considered threatened. In Cuba almost half (15 out of 31) of the mammal species are endemic, and 9 of these are considered threatened. As mentioned, many of these islands have already suffered extinction spasms: of a total of 88 mammal species recorded as extinct since 1600, 53 are from islands, with 20 more extinctions occurring in Australia.

V. Which Mammal Taxa Are Most Threatened?

A. IUCN Red Data Book

Lists such as the Red Data Books are useful in drawing attention to the plight of certain species and in developing conservation and management efforts, including the creation of protected areas. The 1996 IUCN Red List of Threatened Animals attempted to assess the conservation status of every mammal species in the world. Each species was placed within one of eight categories. The first three are categories of threat: critically endangered, endangered, and vulnerable. The next three are lower risk categories: conservation dependent, near threatened, and least concern. The last two categories are data deficient and not evaluated. As well as these categories, two more exist: extinct and extinct in the
TABLE I
Countries with the Most Species of Threatened (CR, EN, VU) Mammals

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of threatened mammals</th>
<th>Total number of mammals in country</th>
<th>Percent threatened</th>
<th>Total number of endemics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indonesia</td>
<td>128</td>
<td>436</td>
<td>29</td>
<td>201</td>
</tr>
<tr>
<td>China</td>
<td>75</td>
<td>394</td>
<td>19</td>
<td>77</td>
</tr>
<tr>
<td>India</td>
<td>75</td>
<td>316</td>
<td>18</td>
<td>96</td>
</tr>
<tr>
<td>Brazil</td>
<td>71</td>
<td>394</td>
<td>14</td>
<td>140</td>
</tr>
<tr>
<td>Mexico</td>
<td>64</td>
<td>450</td>
<td>18</td>
<td>140</td>
</tr>
<tr>
<td>Australia</td>
<td>58</td>
<td>252</td>
<td>23</td>
<td>108</td>
</tr>
<tr>
<td>Papua New Guinea</td>
<td>57</td>
<td>214</td>
<td>27</td>
<td>97</td>
</tr>
<tr>
<td>Philippines</td>
<td>49</td>
<td>153</td>
<td>14</td>
<td>77</td>
</tr>
<tr>
<td>Madagascar</td>
<td>46</td>
<td>101</td>
<td>44</td>
<td>77</td>
</tr>
<tr>
<td>Peru</td>
<td>46</td>
<td>344</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>Kenya</td>
<td>43</td>
<td>339</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>Malaysia</td>
<td>42</td>
<td>286</td>
<td>13</td>
<td>27</td>
</tr>
<tr>
<td>Zaire</td>
<td>38</td>
<td>413</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Viet Nam</td>
<td>38</td>
<td>213</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>35</td>
<td>253</td>
<td>14</td>
<td>31</td>
</tr>
<tr>
<td>USA</td>
<td>35</td>
<td>428</td>
<td>8</td>
<td>101</td>
</tr>
<tr>
<td>Colombia</td>
<td>35</td>
<td>399</td>
<td>10</td>
<td>38</td>
</tr>
<tr>
<td>Thailand</td>
<td>34</td>
<td>269</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>South Africa</td>
<td>33</td>
<td>247</td>
<td>13</td>
<td>27</td>
</tr>
<tr>
<td>Tanzania</td>
<td>33</td>
<td>322</td>
<td>10</td>
<td>34</td>
</tr>
</tbody>
</table>

From IUCN Red List of Threatened Animals (IUCN Gland, Switzerland, and Cambridge, UK, 1996) with permission.

ENDANGERED MAMMALS

An analysis of endangered mammals by taxa, using the IUCN Red List of Threatened Animals (1996) shows that among the larger orders of mammals primates are the most endangered, with 41% of the species under threat (critically endangered, endangered, or vulnerable). All primate species are either on Appendix I or II of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). Primate threats are many and varied, but the principal threat is habitat destruction. Because most primates have arboreal habits, tropical forest loss and fragmentation have had strong negative effects. Captures for the pet trade,
zoos, and medical research had a major impact on a number of species until recent legislation limited wild captures. Other species have been affected by subsistence hunting or accidental take from snares set for other species.

Insectivores are the next most threatened order, with almost 41% of species under threat. Many of these species are poor dispersers found in one or a few populations that are especially vulnerable to habitat alteration and fragmentation. Artiodactyls, or even-toed hoofed mammals, have 31% of the species under threat. Roughly 28% of all marsupial species are threatened, with 33% of species from neotropical marsupial orders and 26% from Australasian orders threatened. Approximately 25% of Chiroptera are threatened. Although much attention has been given to threatened members of the order Carnivora, “only” about one-quarter (24%) of the species are threatened. Rodents have about 16% of species listed as threatened, but because of the relatively large number of rodent species, this order has the highest number of species under threat with a total of 330. Among smaller orders, almost 58% of Perissodactyls (odd-toed hoofed mammals, such as horses, tapirs and rhinos) are threatened, with 4 of the 18 species critically endangered. Almost half (7 out of 15) of the species of Macroscelidea (elephant shrews) are considered threatened. Four out of five Sirenia (manatees and dugongs) are listed as vulnerable, whereas both Proboscidea (elephants) are listed as endangered.

VI. CONSERVATION

A. Conservation Legislation

An enormous number of treaties and laws govern protection for wildlife or their habitat. These measures include prohibitions or controls on taking, collection, possession, and trade of specific species, control of exotic species, and the protection of habitat, either directly through the creation of protected areas or through indirect measures such as tax incentives, permits, and zoning.

Three levels of conservation legislation exist. The first is worldwide treaties, usually involving conservation of habitat as in RAMSAR’s wetland protection, pollution issues, commercial exploitation and trade, or combinations of these issues, as in the Convention on
Biological Diversity of 1992. The second is regional treaties, such as the ASEAN Agreement covering South- east Asia or the Berne Convention for Europe. The last involves individual species or species groups, such as the creation of the International Whaling Commission or the 1973 Agreement on the Conservation of Polar Bears.

CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora), is an example of a worldwide treaty that is especially important for conservation of endangered mammals. CITES was established in 1973 to deal with the enormous level of international trade in wildlife, with estimated earnings as high as US$20 billion. CITES has three appendices, with Appendix I listing species threatened with extinction, Appendix II listing species that may become threatened with extinction unless trade is regulated, and Appendix III listing species protected under national law. Unfortunately, a treaty is only as good as the on-ground enforcement of its provisions, and while CITES has gone a long way toward stemming the tide of illegal trade in wildlife, black markets commercial exploitation continues to threaten many species throughout the world.

B. Species Survival Commissions

The International Union for the Conservation of Nature and Natural Resources (IUCN) is an international union of states, agencies, and organizations dedicated to the conservation of biodiversity, primarily through monitoring and advisory roles. IUCN now has more than 100 specialist groups within the Species Survival Commission (SSC), each focused on a particular taxa. More than 30 of the specialist groups work on mammals, ranging from entire orders such as the Rodent Specialist Group to individual species such as the Asian Elephant Specialist Group. SSC specialist groups consist of volunteer scientists and other experts who provide information on their taxa and even lobby governments for conservation and research efforts. One method of providing this information is through Status Surveys and Conservation Action Plans that describe trends, threats, and conservation options, which most specialist groups have now published for their particular taxa.

C. Protected Areas and Preserve Sizes

With human population increasing and unmodified habitat rapidly dwindling, habitat protection and maintenance is certainly the single most important conservation method for the preservation of biodiversity, including mammals. Large mammals often live at low densities and over large individual areas. A few protected meadows with the appropriate plant species may be enough habitat to maintain a population of butterflies, but an enormous tract of wilderness is necessary to sustain a viable population of rhinos or rhinos. Sadly, many large mammals, such as rhinos and tigers, are now entirely or mostly found within park boundaries, and it is uncertain whether populations are large enough to maintain themselves in the face of stochastic events and anthropogenic changes. Mammals within protected areas often come into conflict with people sharing the land or living on the edges of the preserves, and poaching may continue to cause declines in mammal populations within protected areas for a variety of social or economic reasons. A new focus on off-reserve or nonprotected area conservation of threatened mammals faces daunting challenges from multiple use, social, cultural, and economic growth issues. However, off-reserve conservation may be the only hope for some mammals, as the money and political will to continue creating protected areas is limited.

D. Captive Breeding and Reintroduction

Captive breeding has become a well-accepted way of managing and increasing populations of critically endangered species. New techniques have led to breakthroughs in captive breeding with some species, including the use of extra-specific surrogate mothers and embryo manipulation, including transfer, cryo-preservation, and microsurgical division. However, captive breeding has not worked in every case. For example, the critically endangered Sumatran rhino (Dicerorhinus sumatrensis) has not been bred in captivity despite a number of attempts, and a third of the 27 captive animals have died.

A number of captive breeding programs have led to reintroduction efforts for mammals that were extinct in the wild, including wisent (Bison bonasus) in Europe, Arabian oryx (Oryx leucoryx) in Oman, and Per David's deer (Elaphurus davidianus) in China. Other reintroduction programs have attempted to supplement decreasing wild populations; for example, 23% of wild golden lion tamarins (Leontopithecus rosalia) are now from captive-bred stock. Perhaps the best-known case of captive breeding and reintroduction of a mammal involves the black-footed ferret (Mustela nigripes). This small carnivore was once common across the North American plains, where it specialized in hunting large, colonial ground squirrels called prairie dogs (Cynomys). Prairie dogs were considered to be pests and competitors with
cattle, so an intensive eradication program eventually reduced historic prairie dog range by over 95%. The black-footed ferret subsequently declined and was thought to be extinct in the wild in the 1970s, although a small number of ferrets still existed in captivity. This captive colony suffered from physical problems, perhaps related to inbreeding as well as disease, and the colony died out in 1979, leading to fears that the species had truly gone extinct. However, in 1981 a new colony of ferrets was discovered in Wyoming. The first six ferrets to be captured for breeding died of canine distemper. The remaining wild ferrets were captured and successful captive breeding resulted in their numbers reaching about 300, with some released back to the wild. However, the program has not been without criticism: politically motivated decisions and arguments between state, federal, and private organizations have caused numerous problems and may have even jeopardized the success of the project at times. Finding suitable prairie dog colonies (both in terms of finding colonies of adequate size and the political difficulty of maintaining large numbers of what many consider to be a "pest" species) for continued reintroduction is yet another roadblock to ferret recovery.

Another successful and controversial example has been the reintroduction of gray wolves (Canis lupus) into the Rocky Mountain region of the United States, beginning in 1995 and including most recently Yellowstone National Park. The wolf population in Yellowstone, numbering more than 100, has already had an obvious effect on the ecosystem of the area, causing a decrease in some species (i.e., coyotes, through competition and killing) and an increase in other species (due at least in part to an increase in carrion from wolf kills). Wolves also appear to be altering the behavior of their ungulate prey, which in turn is having other effects on the system. This introduction has not been without controversy, partly stemming from wolf predation on livestock. During a 3-year period wolves in the Rockies killed more than 250 sheep and cattle, but a compensation fund for livestock lost to wolves has helped mitigate some of the antiwolf feelings from ranchers. Despite these problems, the success of the program has led to discussions regarding possible reintroduction of wolves into the northeastern United States.

Other predator reintroduction attempts have not fared so well, however. Recent efforts to reintroduce the Mexican subspecies of the wolf to the southwestern United States resulted in the shooting death of 5 of the first 11 released wolves. Attempts to reintroduce lynx (Lynx lynx) into parts of the Swiss Alps and the Adirondacks of New York have failed. Another lynx reintroduction at the southern limit of the species’ range in the Colorado Rocky Mountains has also had problems, with 4 of the first 13 animals starving to death. The success of many reintroductions has also been threatened by legal challenges from individuals or organizations opposed to land use restrictions or worried about potential direct conflicts with humans or livestock. Reintroductions do not always involve captive breeding. A number of African antelope and Eurasian caprines have been successfully translocated from still-viable populations into areas where they had been eradicated. One issue regarding this technique is whether there was a distinct local genotype. This is a serious problem if translocations are being considered as a method to bolster an existing but declining or no-longer-viable population, in which case mixing genotypes may result in the swamping of possible adaptive differences. There have also been introductions into areas that were not part of the original range of a species, with occasional unintended results of competition with indigenous species and habitat destruction. In one example, 11 Baha- mian hutias (Geocapromys ingrahami), an endangered rodent, were introduced to a nearby island in a successful attempt to establish a second population of the species. However, it appeared that the hutia had not lived on this second island at least within historic time, and after 16 years the population had expanded to the point where seven plant species had vanished from the island overgrazing.

Reintroductions are a complex undertaking, involving a solid understanding of the biology of the species, from feeding and breeding to individual and population genetics. The animals must be trained to survive in the wild, from finding food to socializing and, in the case of arboreal primates, even learning to move through the trees properly. The location for reintroduction should be part of the original range, be large enough to sustain a viable population, contain the correct habitat specifications, and no longer have whatever pressures resulted in the original loss of the species in the wild. Introduced animals must be closely monitored, and there must strong political and local support and extended funding for the project. Thus captive breeding and reintroduction may only be appropriate in specific cases where this kind of information and support is available.

E. Economic Incentives

Because reserves only cover approximately 5% of the earth’s surface, and captive breeding and reintrodu-
tions are only feasible for certain well-studied and well-funded species, other methods will be necessary to ensure the survival of a number of mammal species. One method is economic utilization, which applies the fact that mammals have always been heavily exploited. It also has the important potential advantage of paying for itself. Because much of the present extinction crisis of mammals has an economic incentive, this idea has won wide if cautious acceptance, although controversy still exists on issues such as lack of data on population numbers and dynamics, whether overuse is inevitable, and concerns over animal welfare.

Sports and trophy hunting has been linked with historic declines in a number of mammal species, often in conjunction with other factors such as habitat loss. However, today sports and trophy hunting is being considered as a management tool. For some species, especially big game animals, it is suggested that conservation can “pay for itself” when fees from hunters are redirected back into management. This can be most effective for polygynous species where males are not a limiting resource and are also considered desirable as trophies. Licensing and management for white-tailed deer in North America has resulted in a population explosion of that species to the point where it is now necessary to actively control deer as pests in many locations. Trophy hunting fees may also be directed back to local communities to help replace losses from subsistence hunting or nuisance animals and encourage conservation awareness and behavior. Wild goat trophy hunting programs in the mountains of northern Pakistan have resulted in locally effective conservation efforts on behalf of populations of ibex (Capra ibex) and markhor (Capra falconeri), although the long-term success of these projects is still in doubt. In Zimbabwe, the Communal Areas Management Programme for Indigenous Resources (CAMPFIRE) has had success with trophy hunting as an economic incentive for local people to manage their wildlife in a sustainable manner. However, sports hunting as a management tool is only possible when populations of the target species are large enough to support regular culling, monitoring of the target species is constant, regulations are capable of being altered if necessary, and where the political situation enables funds generated from hunting fees to be returned either directly or through other forms of management to local people to encourage their support. From the biological side, potential problems associated with the removal of dominant males include skewed sex and age ratios, lowered genetic variability, altered behavior of other group members and even the destruction of group cohesion in some targeted species, such as big-horn sheep (Ovis canadensis) and elephants (Loxodonta africana). Tourism is another method of raising funds so that the organism under protection pays for itself. This has worked well for a few charismatic species, such as mountain gorillas, but the vast majority of threatened mammal species are not capable of generating a level of public interest that will result in tourism revenues. However, it has been argued that certain charismatic species may act as “umbrellas” to create protection that will cover other species.

F. Umbrella Species: The Case for and against Charismatic Megafauna

The umbrella principle of conservation involves the protection of a large or wide-ranging species in the hopes that this will result in the protection of a number of other, smaller species. This concept has numerous advantages: it is easier to mobilize public interest toward large and charismatic species, it is easier to get funding for conservation work, and efforts to conserve large species are likely to result in conservation of large numbers of smaller species, due to the large habitat needs of the large species overlapping or including those of smaller ones. The tiger is often used as an example of a species where conservation success will likely mean protection for a number of other species. However, the 15 reserves set up for tigers in India have not even adequately protected tigers, and poaching and encroachment threatens the ability of these reserves to protect many other species found within them. Another problem associated with the concept of mammalian umbrella species is that areas of species richness or endemism for other taxa, such as invertebrates, may not be within the designated core area of protection for a large mammal. A third issue of concern is that management for a single species, such as antipoaching efforts for tigers, may not satisfactorily protect other species whose main threats may involve other factors such as fire, fragmentation, or the spread of exotic species.

VII. PRESENT AND FUTURE TRENDS

Although mammals get a disproportionate amount of attention and funding for conservation efforts, the future for many species still looks grim. As of 1996, one-quarter of all mammal species were considered to be threatened. For a number of these mammals, a long-
term commitment will be needed scientifically, financially, and politically if they are to survive. However, species-based conservation efforts for mammals can have enormous costs. It has been estimated that conservation of the northern subspecies of the white rhino (Ceratotherium simum cottoni) runs about US$10,000 a year per rhino, while the cost over seven years for reintroduction of the golden lion tamarin was estimated at more than US$1.5 million. The cost of creating and maintaining a preserve that will protect populations of many mammal species (as well as other taxa) is only a fraction of that needed for species-based management.

Unfortunately, for many endangered mammals there is a desperate need for immediate and focused management efforts. There also is a tremendous need for more protected areas and better protection within the reserves themselves, and a need for directing conservation efforts toward nonprotected land, whether it be through zoning of public and private lands or the use of private ranches or communal lands. For this to succeed, efforts must also be directed toward educating local people about conservation issues, minimizing wild mammal–people conflicts and developing compensation schemes, and managing the threat of disease. In all cases, research and monitoring of wildlife populations and threats is critical, especially in situations where sustainable harvesting is being attempted.

See Also the Following Articles

CAPTIVE BREEDING AND REINTRODUCTION • ENDANGERED BIRDS • ENDANGERED REPTILES AND AMPHIBIANS • EXTINCTIONS, MODERN Instances OF • MAMMALS, BIODIVERSITY OF • MAMMALS, CONSERVATION EFFORTS FOR • MARINE MAMMALS, EXTINCTIONS OF

Bibliography

ENDANGERED MARINE INVERTEBRATES

James T. Carlton
Williams College—Mystic Seaport

I. Endangered Marine Organisms—Considerations and Definitions
II. Why Is It Hard to Know If a Marine Invertebrate Is Endangered?
III. What Causes the Endangerment of Marine Invertebrates?
IV. What We Know: Examples of Endangered Marine Invertebrates
V. What We Do Not Know: The Potential Scale of Endangerment

GLOSSARY

critically endangered: A species facing an extremely high risk of extinction in the wild in the immediate future.
dangerous: A species not critically endangered but facing a very high risk of extinction in the wild in the near future.
euryhaline: Able to live over a wide range of salinities, from brackish to fully marine waters.
planktotrophic: Larval forms of invertebrates that feed on plankton to survive and grow.
stenohaline: Able to live only in fully marine waters.
threatened: Vulnerable, endangered, and critically endangered species.
vulnerable: A species not critically endangered or endangered but facing a high risk of extinction in the wild in the medium-term future.

ALTHOUGH ONLY A FEW MARINE INVERTEBRATE SPECIES are known to be endangered, there may be hundreds to thousands of species actually at risk of extinction. The ability to determine the conservation status of such species faces considerable challenges. Habitat destruction, fisheries activities, and the introduction of exotic species are among the primary causes leading to the endangerment of marine invertebrates.

I. ENDANGERED MARINE ORGANISMS—CONSIDERATIONS AND DEFINITIONS

More marine organisms are threatened by increasing human pressures in the 21st century than ever before. Of the many enduring challenges in marine conservation science, one of the greatest is securing an objective understanding of extinction risk of individual species. With increasing attention paid in the world’s oceans to the broad concept of “threatened and endangered species”—attention cascading from similar concerns in terrestrial ecosystems—the question of perceived versus actual vulnerability of many marine animals and plants is a pressing one for scientists and conservationists.

In the following treatment one group of marine organisms—the invertebrates—is considered relative to endangerment and potential extinction. In turn, given their broad range of reproductive strategies, habitats, and susceptibility to human-induced extirpation, inver-
Endangered marine invertebrates may serve as model systems to understand endangerment in the broader array of marine life, including algae (seaweeds), seagrasses, vertebrates, bacteria, protists, and fungi.

What constitutes endangerment in marine invertebrates? The IUCN Red List (Baillie and Groombridge, 1996) considers three categories: critically endangered, endangered, and vulnerable species, all grouped under the broad aegis of threatened. Assignment of a species to one of these three categories is based on quantitative assessments (Table I). Information required includes temporal data (for example, rate of population decline, population fluctuations), spatial data (for example, extent of occurrence, number of populations), and population data (for example, size of population, number of adults). A species is critically endangered when “it is facing an extremely high risk of extinction in the wild in the immediate future”; a species is endangered when “it is not critically endangered but is facing a very high risk of extinction in the wild in the near future”; and a species is vulnerable when “it is not critically endangered or endangered but is facing a high risk of extinction in the wild in the medium-term future.”

There are other scales of endangerment as well, and these are parallel to scales of extinction. Local endangerment occurs when a species faces extinction in a small area or habitat. Regional endangerment occurs when a species faces extinction in a broad geographic region. Global endangerment occurs when a species faces extinction everywhere. Functional endangerment occurs when

**Table I**

<table>
<thead>
<tr>
<th>IUCN Categories and Criteria of Extinction Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Y</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>Critically endangered</td>
</tr>
</tbody>
</table>

**A. Declining Populations**
- Population decline rate at least
  - 80% in 10Y or 3G
  - 50% in 10Y or 3G
  - 20% in 10Y or 3G

Using either
  1. population reduction observed, estimated, inferred, or suspected in the past or
  2. population decline projected or suspected in the future based on
    a. direct observation
    b. an index of abundance appropriate for the taxon
    c. a decline in area of occupancy, extent of occurrence, or quality of habitat
    d. actual or potential levels of exploitation
    e. the effects of introduced species, hybridization, pathogens, pollutants, competitors or parasites

**B. Small Distribution and Decline or Fluctuation**
- Either
  - extent of occurrence
    - <100 km²
    - <500 km²
    - <2,000 km²
  - area of occupancy
    - <10 km²
    - <500 km²
    - <2,000 km²

and two of the following three:
  1. either severely fragmented (isolated subpopulations) or known to exist at a number of locations
  2. continuing decline in any of the following:
    a. extent of occurrence
    b. area of occupancy

continues
a species faces loss of its role in structuring a community (that is, influencing the diversity, distribution, or abundance of other species) or mediating energy flow in an ecosystem.

In practice, these categories may be expressed as a dichotomy between (1) species that have been demonstrably reduced to one or a few populations (whether adjacent or widespread) everywhere and thus face global extinction, and (2) species that have been demonstrably reduced in some locations but still maintain populations over a large area and thus face “only” local or regional extinction.

A final category of endangerment involves international transport and trade. Under Appendices I and II of the 1997 Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), in terms of marine invertebrates, all species of giant clams (family Tridacnidae), the queen (pink) conch snail Strombus gigas, all corals (Scleractinia), as well as other Anthozoa (Coenothecalia, Stolonifera, and Antipatharia) and certain hydrozoan corals (Milleporina and Stylasterina) are listed. Appendix I species are considered “the most endangered” and include "all species threatened with extinction which are or may be affected by trade.” Appendix II includes “other species at serious risk . . . which although not necessarily currently threatened with extinction may become so unless trade is subject to strict regulation.” These CITES trade protected species may thus, in popular or legal writing, become endangered species, although many would not be so listed (as either critically endangered, endangered, or vulnerable) under

<table>
<thead>
<tr>
<th></th>
<th>Critically endangered</th>
<th>Endangered</th>
<th>Vulnerable</th>
</tr>
</thead>
<tbody>
<tr>
<td>c. area, extent, or quality of habitat</td>
<td>&gt;1 o/m</td>
<td>&gt;1 o/m</td>
<td>&gt;1 o/m</td>
</tr>
<tr>
<td>d. number of locations or subpopulations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. number of adults</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. fluctuating in any of the following:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. extent of occurrence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. area of occupancy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. number of locations or subpopulations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. number of adults</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Small Population Size and Decline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of adults</td>
<td>&lt;250</td>
<td>&lt;2,500</td>
<td>&lt;10,000</td>
</tr>
<tr>
<td>and one of the following two:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. rapid decline rate</td>
<td>25% in 3Y or 1G</td>
<td>20% in 1Y or 2G</td>
<td>10% in 1Y or 3G</td>
</tr>
<tr>
<td>2. continuing decline</td>
<td>Any rate</td>
<td>Any rate</td>
<td>Any rate</td>
</tr>
<tr>
<td>and either</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. fragmented</td>
<td>All subpops</td>
<td>All subpops</td>
<td>All subpops</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>250</td>
<td>1,000</td>
</tr>
<tr>
<td>or</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. all individuals in a single subpopulation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. Very Small or Restricted</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Either</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. number of adults</td>
<td>&lt;50</td>
<td>&lt;250</td>
<td>&lt;1,000</td>
</tr>
<tr>
<td>or</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. population is susceptible</td>
<td>N/A</td>
<td>N/A</td>
<td>Area of occupancy &lt;100 km² or number of locations &lt;3</td>
</tr>
<tr>
<td>E. Quantitative Analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indicating the probability of extinction in the wild to be at least</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>80% in 10Y or 3G</td>
<td>20% in 20Y or 5G</td>
<td>10% in 10Y</td>
</tr>
</tbody>
</table>
IUCN quantitative guidelines. As an example the North Atlantic coral *Lophelia pertusa*, which ranges to 1500 m depth from Scandinavia to the Mediterranean (and which can occur on the Norwegian shelf in deep water (270–310 m) in the form of coral reefs over 13 km in length, 10 to 35 m in height and up to 300 m in width) was reported in the popular press in 1999 as “endangered,” citing CITES listing of all corals, after being found on North Sea oil platforms. Awarding species protected status may provide certain benefits, such as regulated fisheries, regulated trade and transport, habitat protection, and so forth, even if the species is not close to extinction per se. Alternatively, very broad application of the concept of vulnerability to species that are not immediately threatened may decrease sensitivity to the issue of potential extinction and thus increase the difficulty in affording truly endangered species adequate protection.

II. WHY IS IT HARD TO KNOW IF A MARINE INVERTEBRATE IS ENDANGERED?

Knowledge of marine biodiversity is poor and getting poorer. The names, distribution, and ecology of most marine organisms are not well known. There are fewer professional systematists and biogeographers with every passing generation. It is thus difficult to assess the status of now uncommon and rare species—those that are becoming or have become endangered. And just as in terrestrial ecosystems, many species in the oceans may go extinct or become endangered before they are described. In familiar and relatively easily accessible marine environments in Pacific, Atlantic, and Gulf of Mexico seas, often more than half of the species encountered in certain groups are not known to science (Table II). Indeed, most of the species in the oceans are undescribed, even though the majority of the largest common species in shallow waters have been given Latin names. More generally, the last half of the 20th century, in parallel with increased interest in the scale of modern human-mediated change in the oceans, has been marked by decreased documentation of changes in the status, range, and natural history of marine invertebrates. Carlton et al. (1991) reported that a common species of intertidal snail along the New England coast became extinct about 1930, but the species remains listed as “common” in many modern seashell books. Gathering the precise data required under the IUCN guidelines (Table I) thus may often pose considerable challenges.

As a result, Carlton et al. (1999) were able to document only four species of marine snails that have become extinct in historical time. Roberts and Hawkins (1999) identified no further candidates among marine invertebrates for clear evidence of extinction. Turgeon et al. (1998) list no North American marine mollusks as endangered or threatened. Does this situation mean that marine invertebrates are generally “safe” or “immune” from extinction, or does this reflect our level of

<table>
<thead>
<tr>
<th>Location and general habitat</th>
<th>Invertebrate group</th>
<th>Common name</th>
<th>Number of undescribed species out of total collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gulf of Mexico:</td>
<td>Copepoda</td>
<td><em>Lophelia</em></td>
<td>Up to 27 of 29</td>
</tr>
<tr>
<td>Shallow shelf sediments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Caledon:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shallow lagoon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philippines:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shallow waters of one island</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Georges Bank:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shallow shelf sediments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hawaiian Islands:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coral reef sediment on one island</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Great Barrier Reef</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shallow waters of two islands</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Modified from National Research Council (1995).
knowledge of the conservation status of marine life? Modern authors have concluded that the latter is now the case, and that a ‘tip of the iceberg’ understanding of marine endangerment now exists.

III. WHAT CAUSES THE ENDANGERMENT OF MARINE INVERTEBRATES?

Both natural and human-mediated changes to the environment can lead to the endangerment and extinction of marine organisms. Examples of natural changes include volcanic eruptions destroying endemic island marine communities, tropical storm activity (such as hurricanes and monsoons) causing destruction of reef communities, and other severe weather systems, such as sustained winter freezing impacting temperate shore communities. While many cases of these natural impacts have been documented, none have been linked to the extinction or endangerment of any marine invertebrate species (although such may have occurred).

An example of the ocean-wide demise of a marine invertebrate by what may have been a natural series of events is the case of the eelgrass limpet Lottia alveus. This snail, which occurred from Labrador to New York, lived solely on the blades of the eelgrass Zostera marina. Between 1930 and 1933, 90% of the eelgrass in this region died due to a disease caused by the slime mold Labyrinthula zosterae. The last known population of the limpet occurred on Mt. Desert Island, in the Gulf of Maine, where in 1929 it occurred by the thousands. Lottia was a stenohaline species; however, its host, the eelgrass (with a broader physiological repertoire than the limpet) survived in brackish-water refugia below the salt tolerance of the slime mold. A distinct subspecies—or sibling species—of Lottia alveus survives in the North Pacific Ocean, from Alaska to British Columbia. It may be noted that the possibility remains that Labyrinthula was introduced to North America in some manner.

In an important lesson in understanding the relative susceptibility of marine organisms to extinction, another North American Atlantic eelgrass specialist, the sea slug Elysia catalaou, did not become extinct, because it, like the eelgrass, is euryhaline. Disease events that impact the host of specialized species could, in theory, lead to species endangerment as well, and it may be that for a period of time Elysia would have been regarded as a threatened species until the reexpansion and recolonization of eelgrass over the following decades.

Human-induced pressures on marine communities include fishing, chemical pollution, destruction of habitat, invasions of nonindigenous (exotic) species, and atmospheric chemical alterations leading to global climate change. We know a great deal about many of these pressures. However, our record of changes in the distribution and abundance of marine life in the 19th and 20th centuries is inverse to our record of the rapid increase of these anthropogenic activities.

Overall, habitat destruction, fisheries (both overexploitation and the secondary impacts of fishing) and the invasion of nonindigenous (exotic) species are held to be the primary causes of endangerment to marine organisms. The global extinction of three species of marine mollusks, the southern California rocky shore limpet Calliellula edmitchelli, the Chinese periwinkle Littoraria fluminea, and the southern California mudflat hornsnail Cerithidea fascata, have all been linked to habitat destruction. While overfishing has not yet been linked to the endangerment of a marine invertebrate, vast declines in marine invertebrate populations around the world due to fishing and fishing activities are increasingly recorded. The North Atlantic whelk Buccinum undatum is now absent in areas of the southern and central North Sea and the Dutch Wadden Sea where it was abundant in the 1970s. Overfishing, combined with a disease of unknown origin, have led to such vast declines in commercial sponges in the Mediterranean Sea that some workers consider certain species to be on the ‘brink of extinction.’ Here again regional extinction needs to be distinguished from global extinction, although unrelenting fisheries within an enclosed basin could lead these extinction states to synonymy.

Trawling on the sea floor (dragging nets, chains, bags, scoops, and a plethora of similar devices on the bottom) represents a combination of both habitat destruction and fisheries and has extended the reach of human perturbation well onto the continental shelves of the world, far from the marine habitats that have been classically viewed as those able to be destroyed.

While trawling is particularly a concern of the 20th and 21st centuries, the roots are deep, as they are with many human-induced pressures on the oceans. As early as 1370 the Commons petitioned the King of England ‘that the great and long iron of the wondyrchoun [a net] runs so heavily and hardly over the ground when fishing that it destroys the flowers of the land below water there’ (from the Records of the Parliament in the Reign of Edward III). It seems plausible, given some six and more centuries of trawling in selected regions of the world, that certain invertebrate species would have been heavily impacted. However, no reports of
endangerment or extinction are in hand, a situation resulting, as discussed earlier, from a lack of knowledge of the distribution of threatened species in benthic habitats on the sea floor (and noting again the number of undescribed species found in shelf habitats [Table II]).

Underscoring the probability that marine benthic invertebrates may be endangered by consistent habitat destruction (and emphasizing the difficulty of recognizing such species) is the discovery that a large fish, the barndoor skate Raja laevis, formerly common and widespread in the Northwest Atlantic Ocean, virtually disappeared in the 1990s from benthic trawls, when it used to be common as trawl bycatch. Given that the demise of Raja, a 1.5 m long fish, was not recognized until 1999, the chances of detecting the loss of much smaller organisms such as most invertebrates is all the more unlikely.

IV. WHAT WE KNOW: EXAMPLES OF ENDANGERED MARINE INVERTEBRATES

The IUCN ‘Red List’ lists 13 marine invertebrate species as high-profile threatened species: one species is considered endangered, two critically endangered, and ten vulnerable. Evaluating whether these species (10 mollusks, 1 crab, 1 sea anemone, and 1 seafan) are globally threatened is often a challenge, as no published data are cited for any listing. Often species are suggested for global listing based on local or regional extinctions of populations or based on the rate at which local or regional populations are discovered to be in decline. While such observations may be a harbinger for what may eventually become global concerns, the data may be lacking to understand the full extent of endangerment. As examples, the status of 4 of these 13 species is as follows:

- The tiny Northeastern Pacific crab Parapinnixa affinis is listed by IUCN as ‘endangered’. It lives in bays in shallow, perhaps largely intertidal, mud bottoms, as a commensal in the tubes of terebellid polychaete worms. It was first collected in 1895 in what was to become Los Angeles–Long Beach Harbor, and as late as 1932 it was collected at several locations between Los Angeles and San Diego. No further material from southern California has surfaced. Given the extraordinary modifications of these embayments in the 20th century, with little or no semblance of original habitat often remaining, the demise of this (and many other) species would not be unexpected. However, it may remain common in similar habitats of Baja California, the Gulf of California, and further south, sites that remain largely unexplored. P. affinis was ‘rediscovered’ in one such location, in Tortugas Bay, southwestern Baja California, in 1987. It thus may be an example of local and regional extinction.

- The salt marsh sea anemone Nematostella vectensis is listed by IUCN as ‘vulnerable,’ also a high-profile endangerment category. This species occurs, often in large numbers, in salt marsh pools in the Northwest Atlantic Ocean from Nova Scotia to the Gulf of Mexico and in the Northeast Pacific Ocean from Washington to California. It also occurs in England, its type locality (but not apparently otherwise in Europe); given its widespread distribution in North America, it may be introduced in the Eastern Atlantic Ocean. In England it is considered rare (based on localized populations and declining habitat, although where found it may be abundant) and is given protected status. The habitat of this anemone has, throughout portions of its range, been fundamentally altered (if not completely removed), and thus the abundance of this anemone has inevitably been reduced, but it remains common in many areas. It too would appear to be an example of local extinction.

- The giant clam Tridacna derasa is listed (along with other species of Tridacna) as vulnerable. These clams characterized many shallow-water reef environments up to the mid-20th century, before commercial removal caused their widespread disappearance and population declines. There is thus no doubt that both the range and abundance of a number of Tridacna species have been severely modified, and indeed certain populations are extinct. However, T. derasa, as an example, remains widespread throughout a vast area of Australasia and the Indo-Pacific and is widely used in mariculture.

- The sea fan Eunicea verrucosa is listed as vulnerable. It lives at sublittoral depths of 10 m and deeper on sheltered rocky bottoms. The species is widely distributed from northern Europe to Africa, and it remains common in a number of sites.

All of these species have suffered at the hands of habitat destruction (Parapinnixa, Nematostella) or human exploitation (Tridacna, Eunicea). Because of knowledge lacunae and provincial perspectives, they illustrate the difficulty in ranking threat status and in attempting to assess whether local and regional extinctions imply impending global extinction. There are many

Underscoring the probability that marine benthic invertebrates may be endangered by consistent habitat destruction (and emphasizing the difficulty of recognizing such species) is the discovery that a large fish, the barndoor skate Raja laevis, formerly common and widespread in the Northwest Atlantic Ocean, virtually disappeared in the 1990s from benthic trawls, when it used to be common as trawl bycatch. Given that the demise of Raja, a 1.5 m long fish, was not recognized until 1999, the chances of detecting the loss of much smaller organisms such as most invertebrates is all the more unlikely.

The IUCN ‘Red List’ lists 13 marine invertebrate species as high-profile threatened species: one species is considered endangered, two critically endangered, and ten vulnerable. Evaluating whether these species (10 mollusks, 1 crab, 1 sea anemone, and 1 seafan) are globally threatened is often a challenge, as no published data are cited for any listing. Often species are suggested for global listing based on local or regional extinctions of populations or based on the rate at which local or regional populations are discovered to be in decline. While such observations may be a harbinger for what may eventually become global concerns, the data may be lacking to understand the full extent of endangerment. As examples, the status of 4 of these 13 species is as follows:

- The tiny Northeastern Pacific crab Parapinnixa affinis is listed by IUCN as ‘endangered’. It lives in bays in shallow, perhaps largely intertidal, mud bottoms, as a commensal in the tubes of terebellid polychaete worms. It was first collected in 1895 in what was to become Los Angeles–Long Beach Harbor, and as late as 1932 it was collected at several locations between Los Angeles and San Diego. No further material from southern California has surfaced. Given the extraordinary modifications of these embayments in the 20th century, with little or no semblance of original habitat often remaining, the demise of this (and many other) species would not be unexpected. However, it may remain common in similar habitats of Baja California, the Gulf of California, and further south, sites that remain largely unexplored. P. affinis was ‘rediscovered’ in one such location, in Tortugas Bay, southwestern Baja California, in 1987. It thus may be an example of local and regional extinction.

- The salt marsh sea anemone Nematostella vectensis is listed by IUCN as ‘vulnerable,’ also a high-profile endangerment category. This species occurs, often in large numbers, in salt marsh pools in the Northwest Atlantic Ocean from Nova Scotia to the Gulf of Mexico and in the Northeast Pacific Ocean from Washington to California. It also occurs in England, its type locality (but not apparently otherwise in Europe); given its widespread distribution in North America, it may be introduced in the Eastern Atlantic Ocean. In England it is considered rare (based on localized populations and declining habitat, although where found it may be abundant) and is given protected status. The habitat of this anemone has, throughout portions of its range, been fundamentally altered (if not completely removed), and thus the abundance of this anemone has inevitably been reduced, but it remains common in many areas. It too would appear to be an example of local extinction.

- The giant clam Tridacna derasa is listed (along with other species of Tridacna) as vulnerable. These clams characterized many shallow-water reef environments up to the mid-20th century, before commercial removal caused their widespread disappearance and population declines. There is thus no doubt that both the range and abundance of a number of Tridacna species have been severely modified, and indeed certain populations are extinct. However, T. derasa, as an example, remains widespread throughout a vast area of Australasia and the Indo-Pacific and is widely used in mariculture.

- The sea fan Eunicea verrucosa is listed as vulnerable. It lives at sublittoral depths of 10 m and deeper on sheltered rocky bottoms. The species is widely distributed from northern Europe to Africa, and it remains common in a number of sites.

All of these species have suffered at the hands of habitat destruction (Parapinnixa, Nematostella) or human exploitation (Tridacna, Eunicea). Because of knowledge lacunae and provincial perspectives, they illustrate the difficulty in ranking threat status and in attempting to assess whether local and regional extinctions imply impending global extinction.
species, however, that may be certainly endangered. We discuss examples here (Tables III and IV).

The invertebrates—and all species—that live in marine habitats most influenced by human activity have sustained the most damage. It is in these habitats that endangered species are especially to be sought. These include salt marshes, mangroves, estuaries, lagoons, beaches, dunes, the supralittoral (maritime or strand line) zone, and seagrass beds. Especially vulnerable are habitats within regions that are now fragmented remnants of their previous extent. It is estimated that approximately half of the world's salt marshes and mangroves have been destroyed and along with them, of course, an untold number of species. Seagrasses—particularly the estuarine species, such as the eelgrass *Zostera*, turtle grass *Thalassia*, shoal and star grasses *Halodule*, manatee grass *Syringodium*, and others that support unique species of snails, chitons, and other invertebrates precisely adapted to their blades—serve as particularly compelling harbingers of threat to coastal marine biodiversity.

Several examples are available of endangered species living where expanding human populations impinge on these fragile shoreline habitats. Two small sea slugs, the anaspidean sea hare *Phyllaplysia smaragda* and the sacoglossan slug *Stiliger vossi*, known only from Florida lagoons, have been missing for many years. *Phyllaplysia* was first described in 1977 and last found in 1981. It was only known from the Indian River Lagoon of eastern Florida. It ate epiphytic algae growing on the manatee grass *Syringodium*, a plant subsequently obliterated from the sea slug's type locality (although the plant remains widespread elsewhere in southern waters, including Florida and the Caribbean, and in many of which locations *Phyllaplysia* has been sought unsuccessfully). *Stiliger*, another herbivore, was first and last collected in 1980 in Biscayne Bay, southeast Florida, despite years of subsequent searching. While subtropical Caribbean lagoons and bays remain to be thoroughly explored, there appears to be little doubt that these slugs have been and remain at risk because of living in habitats proximal to rapidly expanding human populations.

The false limpet *Siphonaria compressa* is represented by what is probably one last remaining population in the Langebaan Lagoon on the west coast of South Africa. It is listed by IUCN as critically endangered. Little is known of its life history (although it may have non-planktonic development), population dynamics, or environmental tolerances. It is restricted, however, to living on the eelgrass *Zostera capensis*, the width of whose narrow blades it matches (thus the name of the limpet). In this regard, it is an ecological equivalent of the northern hemisphere true limpet *Lottia alveus*, which similarly was, in the North Atlantic Ocean, restricted to the eelgrass *Zostera marina*. Although Langebaan Lagoon is part of a national park and is a site identified under the Ramsar Convention on Wetlands of International Importance, regional harbor development leading to pollution, the introduction of nonindigenous species, severe weather events, disease, or other phenomena could threaten this last stand of *Siphonaria*.

An unusual example of an endangered taxon in a coastal habitat is the small (1.0 cm) talitrid amphipod crustacean (*beach hopper*) *Transorchestia enigmatica*. It is an intertidal species, eating decaying plant material, known only from a single beach approximately 0.8 km in length on the shores of an estuarine lagoon (Lake Merritt) in the middle of a densely urbanized region (Oakland) of San Francisco Bay. It is closely related to the South Pacific Ocean species *Transorchestia chilliensis*, which is known from New Zealand and Chile. *Transorchestia enigmatica* is presumed to have been carried into San Francisco Bay, perhaps in the 19th century, in the dry (beach) ballast of ships from the southern hemisphere. However, no populations of *T.
TABLE IV
Which Marine Invertebrates May Be at Risk of Being Endangered?

<table>
<thead>
<tr>
<th>Risk group</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species with restricted distributions</td>
<td>Sea hare <em>Phyllaplysia smaragda</em>&lt;br&gt;Sealug <em>Stiliger vossi</em>&lt;br&gt;Flatworm <em>Siphonaria compressa</em>&lt;br&gt;Many other species in coastal habitats, including seagrass&lt;br&gt;stenotypic endemics.</td>
</tr>
<tr>
<td>Species with limited distributions under heavy extraction pressure or under pressure from related extraction activities (such as bottom trawling)</td>
<td>White abalone <em>Haliotis sorenseni</em>&lt;br&gt;Selected benthic invertebrates in areas of heavy trawling; endemic coral reef species.</td>
</tr>
<tr>
<td>Short-range endemics with nonplanktrophic development, especially those under extraction pressure, fisheries activities, or at risk from coastal development and pollution</td>
<td>Many invertebrates species in numerous phyla, including short-range cone snails (<em>Conus</em> spp.) listed by IUCN; endemic coral reef species (all invertebrates); species restricted to islands, offshore banks, and sea mounts.</td>
</tr>
<tr>
<td>Species restricted to an extinguishable habitat</td>
<td>Eelgrass limpet <em>Lottia alveus</em> (extinct in the Atlantic Ocean)</td>
</tr>
<tr>
<td>Species impacted by exotic species</td>
<td>Blue mussel <em>Mytilus trossulus</em> in southern California (a regional and local extinction)</td>
</tr>
<tr>
<td>Species associated with other endangered species</td>
<td>Species associated with&lt;br&gt;* Mammal: <em>Phocoena sinus</em> (Gulf of California vaquita)&lt;br&gt;* Fish: <em>Syngnathus affinis</em> (Texas pipefish)&lt;br&gt;* Fish: <em>Raja laevis</em> (barndoor skate)&lt;br&gt;* Bird: <em>Oceanodroma macrodactyla</em> (Guadalupe storm petrel)&lt;br&gt;* Other endangered vertebrates&lt;br&gt;* Endangered invertebrates&lt;br&gt;* Endangered plants</td>
</tr>
</tbody>
</table>

* If not already extinct.
Another risk group are those species with relatively limited distributions (even if such may be an ocean basin!) under heavy, inexorable extraction pressure. Under this category we find one of the most compelling examples of endangerment among invertebrates in the oceans.

The white abalone, Haliotis sorenseni, is a large gastropod (up to 22 cm long) which formerly ranged from Point Conception in southern California to the islands of central Baja California, on low-relief rocky reefs in sublittoral waters to depths of about 65 m. Its modern postfisheries range, especially its northern limits, is not known. A number of species of abalone, a widely sought after mollusk, occur along the Pacific coast of North America and have since the 1850s attracted sport and commercial fisheries both for the shell and the meat. Marked depletion of certain species on the California coast was noted by the turn of the 20th century and regulatory laws were passed in 1903 to set size limits. Keep noted in 1904 that the “persistent warfare” against the red abalone Haliotis rufescens had rendered large specimens rare compared to about 1884. Edwards (1913) reported that the green abalone Haliotis fulgens and the pink abalone Haliotis corrugata, which about 1893 occurred piled four and five deep on rocks near Avalon, Santa Catalina Island (southern California), were by 1913 entirely gone.

The white abalone H. sorenseni was not described as a distinct species until 1940. It is a deeper water species that entered the commercial fishery only later, after shallow water abalone were exhausted. Thus there are no comparable early records of depletion. However, in remarkable parallel with the reports of Edwards on the green and pink abalones are the much later findings of Davis et al. (1998) who consider white abalone to be on the brink of extinction. Davis and colleagues searched more than 107,000 square m of white abalone habitat on the Channel Islands, the species’ historical center of abundance. By the early 1990s at depths of 25 to 42 m (where in the 1970s mean densities were 2000 to 10,000 abalone per hectare) mean white abalone densities had decreased to 1.6 ± 0.5 per hectare. In 1996 and 1997, at depths of 27 to 67 m, densities had further dropped to extremely low numbers, 1.0 ± 0.4 per hectare. Commercial harvests virtually ceased after the 1970s, but no fishery-independent assessment of white abalone populations was made until 1992 to 1993, in the meantime the fishery remained open until 1996, based on a minimum harvest size of 15.3 cm and a closed season during spawning. All surviving white abalone known in the wild are adults and dying of old age. Davis et al. suggest that while spontaneous recovery is highly unlikely, even in the absence of any further harvesting, a mariculture program may be able to save the species from global extinction.

White abalone produce planktotrophic larvae. However, as the last known recruitment event occurred in the late 1960s or 1970s, and with population densities now far below 1 per square m, successful fertilization may now be impossible, and complete reproductive failure of the species appears to be at hand. Unless unknown populations exist in refugia in deeper water (abalones are rarely if ever seen below 67 m) or populations of sufficient size exist to the south in Mexican waters, the total size of the white abalone population may now be fewer than 1000 individuals.

The white abalone offers the following lessons about the pathway to endangerment: a combination of restricted habitat (a narrow sublittoral zone), the ability to discover and remove most individuals by advanced diving technology, a severe disruption of the life history of the species such that sporadic recruitment (that likely relied on sufficient adult densities) was interrupted, and the lack of a fishery management framework have all led to the demise of this once common open ocean species. No juvenile abalone have been found in many years. An intensive, focused fishery on the white abalone began only about 1965 but vanished 35 years later, perhaps with the mollusk itself.

There are several other guilds of at-risk marine organisms as well. Mentioned earlier were species restricted to an extinguishable (but not necessarily itself endangered) living host, which could be impacted by a disease or other agent. The demise of the eelgrass Zostera marina under this category. Endangered species would also include, of course, those invertebrates that are commensals, parasites, and symbionts of other endangered marine organisms (Table IV). Finally, the invasion of exotic species may lead to at least local or regional extinctions (Geller, 1999). The native mussel Mytilus trossulus, once abundant in southern California both historically and in native American shell middens, was rendered virtually regionally extinct by the invasion in the 20th century of the Mediterranean look-alike mussel Mytilus galloprovincialis.

V. WHAT WE DO NOT KNOW: THE POTENTIAL SCALE OF ENDANGERMENT

Given the extent of the modification, destruction, devastation, and extirpation that has characterized coastal
marine environments over the past 1000 and more years, hundreds or thousands of species of marine invertebrates may in fact be endangered, as opposed to the few cases reviewed here or listed under IUCN. For example, it has been estimated that given that a minimum of 5% of the world’s coral reefs have been degraded to a nonreef state, and based on estimates of coral reef biodiversity per unit area, as few as 1,000 species of coral reef invertebrates may have already become extinct in modern times (although, as with many tropical forest ecosystems, we have no records of these species, many of which may not have been described before going extinct). It follows that many thousands of coral reef invertebrates may be endangered as well, given that coral reef destruction continues largely unabated in many parts of the world.

More generally, the demise of the white abalone offers a striking counterpoint to the ancient assumption of the invulnerability of marine life to exhaustion by human extraction. Lamarck in 1809 wrote,

Animals living in the waters, especially the sea waters . . . are protected from the destruction of their species by man. Their multiplication is so rapid and their means of evading pursuit or traps are so great, that there is no likelihood of his being able to destroy the entire species of any of these animals.

These words were echoed by Byron in 1818 in a similar fashion,

Man marks the earth with ruin, his control Stops with the shore.

These turn-of-the-19th century views of the oceans were clearly incorrect and at the turn of the 21st century the reverse now appears to be true.

See Also the Following Articles

ENDANGERED FRESHWATER INVERTEBRATES • ENDANGERED TERRESTRIAL INVERTEBRATES

Bibliography


THE GROWING PUBLIC CONCERN about endangered plants is well founded. It is estimated that there are between 235,000 (Raven et al., 1986) and 270,000 (IUCN, 1998) species of vascular plants (including ferns and fern allies, gymnosperms, and flowering plants) in the world today; two-thirds of these species are found in the tropics (Raven et al., 1986). The World Conservation Union (IUCN, 1998) currently lists 33,798 species (12.5% of the world’s flora) as threatened, and 6,522 species of plants (2.4% of the world’s flora) as endangered. Threatened plants are found among 369 plant families in 200 countries. The U.S. Fish and Wildlife Service reports 540 endangered species and 130 threatened species of flowering plants in the United States (U.S.) as of November 30, 1998.

I. INTRODUCTION

Extinction has always played an important role in nature: 95 to 99% of all species that ever existed are now extinct. It is the current rapid rate of extinction that has many plant ecologists worried. Worldwide, as many as 654 species of plants have gone extinct since A.D. 1600 (Heywood, 1995). In this time period, about 110 plant species may have gone extinct in Hawaii according to records of the U.S. Fish and Wildlife Service and The Nature Conservancy. Estimates of plant extinctions in the continental United States range from below 25 (IUCN, 1998) to above 90 (Davis et al., 1986; Flather
et al., 1994). The discrepancy in estimates is due to incomplete systematic surveys, limited monitoring activities, and poor information on viable seed banks. Estimates of extinctions in other countries include 71 plant species in Australia, 53 in South Africa, 47 in Mauritius, and 23 in Cuba. Three endemic species of ebony ( Diospyros spp.) went extinct in the Mascarene Islands before they were described. Nuttall's mudwort (Micranthemum micranthomodes) had been recorded in Delaware, Maryland, New Jersey, New York, Pennsylvania, Virginia, and the District of Columbia, but, despite extensive searches it has not been seen since 1941. The single individual of Dichotera dohmi, which clings to life in western Ecuador (Gentry, 1986), may be the next victim of accelerated extinction.

Estimates of future plant species extinctions vary widely. Raven et al. (1986) estimated that about 40,000 tropical plant species may go extinct in the wild within the next several decades. The New York Botanical Garden suggests that about 700 of the more than 2000 species of threatened and endangered plant species in the United States may be extinct in the next 10 years. The Center for Plant Conservation reported that 680 U.S. plant species were critically endangered, with approximately 253 species estimated to become extinct in 5 years and 427 species to become extinct in 10 years (based on unpublished data in 1988). Today, more than 10 years later, there is little evidence of these mass extinctions, but the concern may still be well founded in several areas.

Hawaii provides a good example of endangered plants. Up to 47% of the Hawaiian flora may meet the criteria for official listing by the U.S. Fish and Wildlife Service, with as many as 16% immediately threatened with extinction (see the Missouri Botanical Garden website listed in the Bibliography). The genus Hibiscadelphus in Hawaii includes only six species (with two extinct species) with a total of 14 live individuals, and one species with only one individual (Cody, 1986). The Hawaii Endangered Plant Task Force, which includes many federal, state, and non-government partners, now tracks 997 threatened, endangered, and rare plant species.

There is a growing backlog of candidates for listing as endangered species. Between July, 1976, and August, 1992, about 231 species per year were added to the U.S. endangered species list. In the second half of that 16-year period, 33 species per year were added to the list (Flather et al., 1994). More than 200 plant species are being petitioned for future listing in the United States. It has been said that “all species are rare somewhere” (Orians, 1997), because most plant species have larger populations in well-suited habitats and smaller populations in marginal or sub-optimal habitats. Many threatened vascular plants in Finland were found to be rare largely because the species were in marginal rather than suitable habitat (Cropper, 1993). Along with the increasing threats of rapid environmental change, habitat loss, contaminants, and invasive exotic species, we may face an uphill battle in protecting our treasured botanical resources in these sub-optimal habitats.

Maintaining biodiversity requires a considerable understanding of rarity and the processes and stresses that promote the endangerment of certain plant species. This article reviews the kinds of rarity, patterns of endangerment, causes of endangerment, consequences of rarity, and legal mandates to protect endangered plant species. It then provides selected examples of endangered plants and concludes with the management and social implications of protecting endangered plants.

II. KINDS OF RARITY

There are many published definitions of rarity. In an extreme example, DuMond (1973) stated nine criteria for rarity, including species that are: (1) found out of its expected context; (2) particularly subject to extinction or severe reduction in total population size by human activities that have already caused a significant population reduction; (3) found only in a very specific habitat of limited occurrence; (4) thought to be a relic of a no-longer extant vegetation association; (5) an indicator of a unique extant vegetation association; (6) recognized as an example of a wide, disjunction pattern; (7) at its natural distribution limits within the area in question; (8) known to be introduced and has become naturalized only on a very small scale; or (9) does not consistently occur as a member of any particular natural plant community. Gaston (1997) described rarity in another way, noting that rare species can be delimited on the basis of one, two, or at most a few of the following variables: abundance, range size, habitat specificity (habitat occupancy), temporal persistence (e.g., taxon age), threat (probability of, or time to, extinction), gene flow, genetic diversity, endemism, and taxonomic distinctness. Rare species in the IUCN Red Book (1998) are described in terms of population size rather than an assessment of extinction risk, which is reserved for threatened and endangered status. The Nature Conservancy and associated network of Natural Heritage Programs rate species endangerment based on five levels of global rarity (i.e., G1, G2, G3, etc.), national rarity (N1, N2, etc.), and state rarity (S1, S2, etc.).
The kinds of rarity that concern conservation biologists relate specifically to the potential vulnerability to extinction. Commonly recognized kinds of rarity (Rabinowitz, 1981; Cropper, 1993) include species with:

- Small populations, large geographic range, and occurrences in several habitats. For example, American chaffseed (Schwallrea americana) had a historical distribution from Mississippi to Massachusetts. It resides now in 20 relatively small populations in five states, with the greatest populations in South Carolina. It can survive in several vegetation types as long as there is enough light in the understory. Fire suppression, which allows for succession and canopy closure, threatens several Schwallrea populations. Another example is Hypochoris maralana, which has small populations over a wide range in Great Britain.

- Large populations, large geographic range, but occurrences in specific habitats. For example, several grassland species of Calochortus (lilies) in California have large populations following wildfires. They were thought to be far more widespread prior to a century of fire suppression and encroachment of shrubs into grassland habitats. Sparse populations in a large geographic range but in specific habitats. For example, Pulstum madsoni grows only on rocky outcrops and always in sparse populations, but it occurs in Australia, New Zealand, and Easter Island (Cropper, 1993).

- Small geographic range, but locally abundant in specific habitats. For example, the 48 endemic species of Bignoniaceae in Amazonia have a total home range of only a few thousand square kilometers, but some of these species are locally abundant in restricted habitats (Gentry, 1986). A specific example in Scotland is Primula scotica, which has a tiny range with a few large populations.

- Small geographic range, and sparse in specific habitats. This may be the most critical type of rarity, for small populations restricted to small and specific habitats are highly vulnerable to extinction. For example, the scouring of a new stone quarry in South Africa ended life in the wild for Moraea loubservi, a small, sparsely population iris (Sterner, 1993).

- Small populations of concern in any region despite populations elsewhere. For example, many floristically poor northern European countries typically have 200 or more plant species listed as threatened or endangered. However, many of these species invaded broadly since the last Ice Age, and they are buffered from extinction with widely scattered populations in several countries (Davis et al., 1986).

- Small populations of new species. For example, a newly discovered species of Calyptranthes from El Yunque, Puerto Rico, has a population of four trees, and Auerodendron pauciflorum from Quebradillas and Calyptranthes liquitissimus from the Luquillo Mountains have only five individuals each (Cody, 1986).

Generally, species may be ranked for protection based on overall rarity, magnitude of potential threats to populations, and the immediacy of the threats. Thus, a monotypic genus with high-magnitude and imminent threats might receive a higher priority ranking than a subspecies with moderate or non-imminent potential threat.

III. ENDANGERMENT PATTERNS

Slightly over 90% of threatened plants are single-country endemics. Species with restricted ranges face the greatest threat of extinction. It follows that endemic plant species on islands appear to be the most highly endangered. About 93% of the plant species on the Canary Islands are endemic, and 50% are considered endangered (Cody, 1986). Senecio heritieri, for example, is restricted to a small area of rocky slopes on the south coast of the island of Tenerife, one of the Canary Islands. On Crete, two-thirds of the 135 endemic species are endangered (Cody, 1986). Compare these numbers to those of the entire United States, where there are about 20,000 plant species of which 2,050 species are rare and threatened.

About half the plant species in Mediterranean climate areas (parts of California, South Africa, Australia, and the Mediterranean basin) are narrow endemics that dominate the threatened and endangered species lists in their various countries (Fig. 1) (Davis et al., 1986). California, for example, contains 869 endemic species of the 2,050 species on the U.S. rare and threatened list (Davis et al., 1986). Chile has 30 forest tree species; 47 of these species are endemic and 38 species are listed as endangered, vulnerable, or rare.

Over half of the rare and endangered plants of the continental United States grow within the borders of 12 western states (Arizona, California, Colorado, Idaho, Montana, Nevada, New Mexico, Texas, Utah, Oregon, Washington, and Wyoming; Fig. 2) (Flather et al., 1994). However, among U.S. states Hawaii is the hot
FIGURE 1  Distribution of rare plants in the world. Data from the IUCN Red List for 1997 (IUCN, 1998).

FIGURE 2  Distribution of rare plants in the continental United States. The “top 10” continental states for rare plants include six western states, and California clearly dominates. [From C. H. Flather et al., BioScience 48(5), pp. 365–376. ©1998 American Institute of Biological Sciences.]
Thus, Eriogonum shrubs of the species-rich genera prone to extinction: 72 of 268 native species are very (IUCN, 1998). Cacti in the United States are particularly less adapted to today’s rapidly changing environment. Sporums, containing relatively ancient species, may be of rare species than do species-poor genera. Gymnosperms, containing relatively ancient species, may be less adapted to today’s rapidly changing environment (IUCN, 1998). Cacti in the United States are particularly prone to extinction: 72 of 268 native species are very rare (Benson, 1982). In California, shrubs and subshrubs of the species-rich genera Arctostaphylos, Ceanothus, Eriogonum, and Ribes have a higher proportion of threatened and endangered plants relative to their proportion in the flora (Cody, 1986). Curiously, some rare plant habitats may have human origins. Some moorlands, or blanket bogs, in western Europe were created around 7,700 years ago by human activities of burning and grazing. These heathlands and grasslands today support many endangered plant species.

IV. CAUSES OF ENDANGERMENT

There are several natural causes of rarity that lead to endangerment. Ancient taxa, like cycads, are thought to be prone to extinction as they cling to increasingly shrinking relic habitats. New taxa, resulting from recent speciation, are thought to be susceptible to extinction because they have not had time to spread in distribution. Rare species generally lack an ability to rapidly colonize areas and they are often poor competitors. Many rare species are edaphically restricted to specific soils or geology. However, based on recent rates of extinction, natural causes of rarity (individually or combined) are no match for human-related causes of endangerment. Species usually become endangered because of multiple, human-related causes. The overwhelming cause of plant species endangerment is habitat loss, directly linked to agriculture, forestry, and urbanization. Remaining small populations have a higher risk of extinction (local extinction) or extinction. Coinciding with habitat loss is reduced habitat quality related to invasive species, grazing, and other land-use changes (Flather et al., 1994). There are many examples of habitat loss to choose from. In the wetter forested Sierra Madres of Mexico, Guatemalan fir (Abies guatemalensis) or Pinus bete trees are considered endangered from over-exploitation of old-growth forests, land-use change, and livestock grazing (Burton, 1991).

Persea theobromifolia, once an important timber species in western Ecuador, has been reduced to fewer than 12 trees at Rio Palenque (Centry, 1986). In the United States Burton (1991) reported that only 1% of the 1,036,000 km² (400,000 mi²) of tallgrass prairie remains, now existing in isolated patches and small nature reserves. Over half of all continental U.S. wetland and aquatic habitats have been destroyed. Fire suppression and succession, invasive exotic plants, and habitat fragmentation will make it increasingly difficult for rare species to persist. Causes of rarity vary by region. Though habitat loss was generalized as an overall problem, agricultural development was specifically noted in the southern Appalachian region, whereas urban development and forest clearing were specifically noted in Florida. In the Gulf Coast areas, shoreline modification and development were specifically noted in endangerment patterns (Flather et al., 1994).
Commercial exploitation is also a problem. Cacti in the southwestern United States are a prime example. Legislation regulating the collection of cacti in Arizona and California date from 1929. Yet, in a single year, over half a million small cacti were illegally collected for sale in the United States, Japan, and Europe (Burton, 1991). One of the two colonies of the 2.5-cm-diameter Nellie Cory cactus (Coryphantha minima) was completely eliminated by thieves in the 1960s. The rarest species demand the highest prices from collectors. Rare orchids are plucked one by one by various “collectors.” Some threatened orchids in India, such as Paphiopedilum druri, Dendrobium pauciflorum, D. nobile, and Diplomeris hirsuta, are used for medicinal purposes. As such, they are facing increasing pressure from growing and aging human populations.

Natural environmental change can both cause and maintain rarity. Cropper (1993) reported that some rare species, such as Leptothyrsa gazesi in Australia, are observed only following natural fire. A riparian herb, Collomia rawsoniana, required periodic high and low streamflows to persist. Likewise, disruption of the natural hydrology of Lake Okeechobee, the largest lake in Florida, has led to the endangered listing of Caesalpinia okeechobensis. Land management decisions can obviously upset required disturbance patterns. Fire suppression activities and flood control will likely increase the rarity of such species.

The introduction of exotic grazers such as goats, pigs, rabbits, and sheep often leads to devastated native floras. For example, 87 of the 49 native plant species on St. Helena Island in the south Atlantic Ocean are now rare or threatened. The island was once dominated by beautiful forests of St. Helena ebony (Trichetta melanoxylon). Goats were introduced on the island in 1513, and the goat population skyrocketed by 1588. Goats ate ebony seedlings and humans used mature trees as fuel, so by 1810 the forests were destroyed and the ebony was extinct. About 260 naturalized exotic plant species were also introduced to the island (Davis et al., 1986).

Introduced rabbits have greatly threatened Acacia corte in arid western New South Wales, Australia (Cropper, 1993), and nearly decimated Dudleya truxillae (ironically called the Santa Barbara live-forever) on Santa Barbara Island off California (Benseler, 1987). Hawaii’s beautiful Haleakala silversword (Argyrocephalum sandwicense ssp. macrocephalum), with a flower stalk up to 2 m tall, almost went extinct in the 1920s due to vandalism and grazing by introduced cattle and goats. Like many endangered plant species, several additional threats must be held at bay. A major effort in Hawaii is under way to protect the habitat of the Haleakala silversword from invasive exotic plants (Verbascum thapsus, mullein; and Pennisetum setaceum, fountain grass) and the Argentine ant (Iridomyrmex humilis). The Argentine ant poses a significant threat to native pollinators of the silversword.

Competition from invasive exotic plant species may be a major contributor in the future to native plant endangerment. Large, nearly pure stands of exotic purple loosestrife (Lythrum salicaria) have directly influenced the endangered small spikerush (Eleocharis parvula) in New York and Long’s bulrush (Scirpus longiz) in Massachusetts.

V. CONSEQUENCES OF RARITY

The most obvious consequence of rarity is extinction. Charles Darwin, like many naturalists of his time and since, recognized that rarity often preceded extinction. Small populations of sessile organisms are vulnerable to catastrophes. Landslides, fire, flooding, hurricanes, and other disturbances can simply wipe out populations. Small populations of vascular plants are also vulnerable to breeding problems from higher variability in breeding success caused by inbreeding. Inbreeding has been documented in several localized endemics such as Limnanthes bakeri, a vernal pool species in Mendocino County, California. In contrast, a common congener Limnanthes douglassii reproduces almost exclusively by cross-fertilization. Stephanomeria malheurensis, a plant species confined to one small hilltop in Burns, Oregon, is auto-fertile, whereas widespread congeners are not. Reproductive failure is not uncommon in sparse species. Lower genetic variation in small populations may also make them more vulnerable to rapid environmental change.

Not all small populations march rapidly towards extinction. Many rare plant populations can persist for centuries and millennia. Sparse populations can often avoid pathogens and herbivory: Pinus ponderosa, which may have been restricted to small refugia in Arizona and New Mexico at the end of the last Ice Age, has become the most widespread pine in the western United States; it now occurs from Mexico to Canada, and from California to Nebraska. Other small populations may speciate (form new species). For example, two species of Ranunculus in alpine areas of the North Island of New Zealand (R. verticillata and R. insignis) have given rise to R. nivicola.
VI. LEGAL MANDATES TO PROTECT ENDANGERED PLANTS

A. Policy and Legal Mandates throughout the World

The International Plant Protection Convention held in Rome in 1951 set forth recommendations for the protection and promotion of plant life throughout the world. Since then, the International Union for Conservation of Nature and Natural Resources (IUCN; now known as the World Conservation Union) has taken center stage in the protection of the world’s flora. The IUCN Plant Red Book strongly defines “endangered species” as a species in danger of extinction and whose survival is unlikely if the causal factors (e.g., over-exploitation, extensive habitat destruction) continue operating, “including taxa whose numbers have been reduced to a critical level or whose habitats have been so drastically reduced that they are deemed in immediate danger of extinction” (IUCN, 1998). It also identifies and tracks “vulnerable species” as those “believed likely to move into the endangered category in the near future if the causal factors continue operating.” Finally, it tracks “rare species”—taxa with small world populations that are not presently endangered or vulnerable, but are at risk.” These IUCN classifications are determined by scientists and government officials around the world and classified species are not necessarily afforded legal protection after designation (Stermer, 1995).

Several countries have policies or legislation that protect endangered plant species. Following the Convention on International Trade in Endangered Species (CITES) of Wild Flora in 1973, 113 countries have agreed not to trade certain threatened species. Many countries augment these agreements with additional legislation. For example, the Mauritius National Plant Protection Legislation (the Plants Act of 1976) and the Forest and Reserves Act (1983) legally protect endangered species and habitats in the territories of Mauritius. Enforcement of endangered species laws and policies in many countries is generally considered to be well-intentioned but weak.

B. The United States Endangered Species Act

The intent of the U.S. Endangered Species Act of 1973 (16 USC1531-1543) is to prevent further decline and help restore endangered and threatened species and the habitats upon which such species depend and to “provide a means whereby the ecosystems upon which endangered species and threatened species depend may be conserved” (Greenwalt and Gehenger, 1975). Thus, the Act recognizes the inseparable link between protecting a species, its habitat, and the surrounding ecosystem. The Act also provides broad-ranging protection for all species threatened with extinction in the “foreseeable future.”

The Act makes the “taking” of endangered species anywhere within the United States a federal offense, requires federal agencies to use their existing authorities to conserve listed species, prohibits federal agencies from taking actions that may jeopardize a species’ existence, provides a formal structure for listing endangered species, and provides a means for citizens to bring suit against any federal agency for failure to meet its obligations under the Act (Flather et al., 1994). Ayensu and DeFilipps (1978) noted that a species may be rare at the edge of its range, but not endangered or threatened as a whole. In determining national endangered, threatened, and extinction status, the total range and abundance of the species must be considered. However, states may further protect a species threatened with extirpation. Still, the cost of protecting individual species and habitats against multiple stresses is high, and the reality is that enforcement on public and private lands is generally weak.

VII. SELECTED EXAMPLES OF ENDANGERED PLANTS

Coleus forskohlii (Willdenow)

Coleus forskohlii, a 40-cm-tall, rare herb, is found in the Yunnan Province of China, Bhutan, India, Nepal, Sri Lanka, and Africa at about 2300 m on steep slopes (Fig. 3). Compounds extracted from the roots have long been a Hindu and Ayurvedic traditional medicine. Rampant collection has increased the rarity of this species. A German pharmaceutical company holds at least six U.S. patents for use of the plant as treatments for high blood pressure, cardiovascular disease, colic, respiratory problems, insomnia, painful urination, and convulsions. This is sometimes referred to as “biopiracy,” when the intellectual property of indigenous peoples is appropriated and used by foreign companies to develop and patent commercial products.

Fitzroya cupressoides

Fitzroya cupressoides is an ancient tree species in southern Chile (Fig. 4). Only 9% of the world’s temperate...
forests are in the Southern Hemisphere, and one-third of the threatened temperate forests occur in Chile. The highest biodiversity in any temperate forest is also found in Chile. These forests are remarkably productive, with some of the world’s largest concentrations of biomass. Old-growth Fitzroya cupressoides trees often reach 4 to 5 m in diameter and may live for 6000 years. One-third of Chile’s forests were burned or cleared by 1955. Some Fitzroya populations are protected in nature reserves, but most of the forests containing this species are privately owned.

*Sarracenia rubra* ssp. *alabamensis*

The Alabama canebrake pitcher-plant is a carnivorous plant with maroon flowers on 0.6-m stalks originating from rhizomes (underground roots; Fig. 5). It is found in sandy and gravelly bogs, seeps, springs, and swamps, and flowers from late April to early June. *Sarracenia rubra* ssp. *alabamensis* is restricted to only 12 localized sites in a three-county area in central Alabama. Four of the 12 sites have 70 to 300 plants each, and half the populations have 2 to 20 plants each. Much of the original habitat (16 other sites) has been modified or destroyed by agriculture and construction of farm ponds in boggy areas. Fire exclusion, gravel mining, and invasive plants pose additional threats. Several populations have also been lost or degraded by plant collectors.

*Encephalartos longifolius*

South African cycads are ancient gymnosperms and the most primitive living seed-bearing plants on Earth (Fig. 6). They flourished 50 to 60 million years ago and provided forage for dinosaurs before then. In South Africa, all 40 cycad species are endangered, and some species are extinct in the wild. The thick-trunked plants
with rigid spiked leaves grow very slowly; some species take 100 years to grow 1 m. Habitat loss and competition from invasive exotic plant species are often cited as causes of rarity, but illegal collecting is also a major problem. South Africa has some of the world’s strictest laws controlling cycad theft, but the thefts continue. Some cycads now have implanted microchips for identification and tracking purposes in an effort to curb poaching.

_Pterostylis truncata_

Brittle greenwood is a ground-dwelling orchid that emerges in the fall with large, squat flowers (Fig. 7). It is a clonal plant that can regenerate vegetatively, and its flowers can be pollinated by a small flies. _Pterostylis truncata_ is found in only three locations in south-central New South Wales, Australia. One of the populations is threatened by introduced European rabbits (_Oryctolagus cuniculus_), feral goats (_Capra hircus_), and eastern gray kangaroos (_Macropus giganteus_). An additional recent threat comes from the highly invasive weed _Chrysanthemoides monilifera_. The second population is on private land, where it is potentially threatened by kangaroos and rabbits. The third population occurs in a forest reserve; here it is threatened from trampling by orchid enthusiasts, weed invasions, and bird predation.

**VIII. IMPLICATIONS**

Reducing the rate of habitat loss around the world will be difficult. Between 1980 and 1990, species-rich tropi-
cal forests were cleared at the rate of 6.3 million ha/yr (15.4 million acres/yr), or 0.8% of the forest per year (Heywood, 1995). Deforestation causes habitat loss, habitat fragmentation, and edge effects at the boundaries, so the effects are greater than indicated by the actual deforested area. Large areas will be lost to urbanization and agriculture. Many dry grasslands in Germany have been converted to range and arable land in recent decades, making it more difficult to protect the nation’s 164 threatened plant species (Heywood, 1995). Thus, saving critical habitats throughout the world is seen as increasingly important and extremely urgent.

A species-by-species approach to rare plant conservation is expensive and difficult. However, the unstated assumption of the habitat preservation approach is that species are inseparably linked to habitats, which in turn are stable and predictable. Yet because habitats may be neither stable nor predictable (Flather et al., 1994),...
both species and habitats must be monitored to protect endangered plants and animals. In Australia, management techniques to protect critical habitats and species include: burning or slashing overgrown vegetation, removal of weeds, removal of grazing animals, hand pollination of selected species, propagation and seed storage, reintroduction into restored habitats, and quarantine to reduce the threat of introduced pathogens (Cropper, 1993).

Less than 10% of named plant species have been analyzed for medicinal or nutritional properties (Stermer, 1995). Nonetheless, 25 to 40% of all drug prescriptions in the United States contain plant ingredients (Durant and Saito, 1985), and many of these ingredients cannot be synthetically made. About 80% of people in developing countries use traditional medicines. As more and more plant species go extinct, so may our chances to find the next heart medicine (e.g., *Digitalis*; foxglove) or treatment for childhood leukemia (from *Catharanthus roseus*; the rosy periwinkle).

To improve the effectiveness of biodiversity conservation, increased emphasis is needed on systematic surveys and monitoring. Many species may be classified as rare owing to poor surveys, as evidenced by the rate of species discoveries. Since 1970, for example, one botanist has discovered 55 new plant species in Utah, and over 80 species have been named since 1970 (Durant and Saito, 1985). Botanists searching the well-studied Rocky Mountain National Park in Colorado added over 100 species to the Park’s plant checklist between 1987 and 1992 (Stohlgren et al., 1997). On average, one or two native plant species are added to the flora of New York each year. The Nature Conservancy and Natural Heritage Programs have found that plant surveys often show that many plant species are more common than previously believed. For example, several populations of rare orchids have been found in Australia after intensive searches. Systematic surveys of plants and rare habitats are badly needed.

The preservation of intact ecosystems may be the most promising way to protect endangered plants. It is now widely understood that maintaining natural disturbance regimes, such as fire and flooding, is important for many rare species (Cropper, 1993), and that preserv-
ing the habitat around rare plant locations is vitally important. However, subtle assaults to endangered species may include the use of fertilizers, herbicides, and insecticides that may reduce plant pollinators, invasive plant species that may out-compete the endangered plants, and introduced foreign diseases, insects, or domesticated/feral animals that may further endanger plants. Where habitat loss and degradation have been significant, habitat restoration efforts and species reintroductions are needed.

Cultivation, the last resort for endangered plants, may become necessary in some cases to preserve genetic variation. Franklinia (Franklinia alatamaha) is a familiar example. Franklinia is a small tree that was restricted naturally to the Altamaha River basin in southeastern Georgia. Now extirpated in the wild, the species survives as an ornamental throughout the eastern United States. In Australia, 1053 of 3329 rare and threatened plant species are found only in botanical gardens, and 515 plant species are held in only one collection (Cropp, 1993). Several botanical gardens in the United States have active programs to study, collect, and grow rare and endangered plants from all over the world. However, duplicate and more complete collections are needed in most countries, and such a system of collections should not be viewed as a substitute for preserving native habitats and ecosystems.

Some plants have appeared to bounce back from extinction. Mountain golden heather (Hudsonia montana) was described by Thomas Nuttall in 1818 in North Carolina, and was thought to be extinct in the 1960’s. About 2000 individuals were found in the Blue Ridge Mountains around 1990 (Burton, 1991). Running buffalo clover (Trifolium stoloniferum) was also believed to be extinct, but it was rediscovered in 1983 in West Virginia and later discovered in four other states. Also in 1983, Lomatium peckianum was rediscovered in Oregon after “disappearing” for over 50 years. In 1985, one species of clover (Trifolium microcephalum) that had not been seen since the Lewis and Clark expedition in 1805–1806 was rediscovered. Such stories are not uncommon. Eleven plant species in New York have been “rediscovered” in the past 10 years. However, expecting miraculous rediscoveries for many long-lost species is probably unrealistic.

This article has focused on endangered vascular plants. However, many non-vascular plants are similarly threatened. For example, >50% of the mushrooms in Europe are listed as endangered or threatened in at least one country. Species that were once common,
such as Hydnum repandum, have been extirpated in some countries. Over-collection of the lichen Gymnoderma lineare in the Great Smoky Mountains National Park has led to its recommendation for listing. Air pollution and acid deposition are known to adversely affect many non-vascular plant species. The assessment of endangered plants cannot be done without assessing the interacting species (e.g., pollinators and symbiotic fungi) and the ecological processes that affect complexes of rare and common species.

Understanding the causes and consequences of rarity also requires a comprehensive knowledge of biology, evolutionary and recent history, and species demography. Often, the causes of population decline may be elusive. For example, Torreyra taxifolia, the Florida torreyra, is a narrowly restricted endemic conifer that has suffered catastrophic declines since the 1930s. Pathologists and ecologists have studied the problem relentlessly and found no obvious cause for the decline. There are now fewer than 1500 trees in the wild, with no solonion in sight. In many cases, careful field and laboratory experiments may be necessary to isolate the causes of rarity.

It is equally important to increase public awareness about the ecology of rarity and the economic and social consequences of losing our endangered plants. In the end, protecting endangered species and biodiversity is a question of ethics and values. Endangered plants are best viewed as valuable resources and inherently valuable species with which we share the Earth.

Acknowledgements
April Owen and Lisa Schell assisted with the literature review and graphics. Corin Flather provided information and Fig. 2. Geneva April Owen and Lisa Schell assisted with the literature review and graphics. Corin Flather provided information and Fig. 2. Geneva

See Also the Following Articles
ENDANGERED ECOSYSTEMS • ENDEMISM • EXTINCTION, RATES OF • PLANT BIODIVERSITY, OVERVIEW • PLANT INvasions


ENDANGERED REPTILES AND AMPHIBIANS

Tim Halliday
The Open University

I. Reptiles and Amphibians
II. The Conservation Status of Reptiles and Amphibians
III. Threats to Reptiles and Amphibians
IV. Declining Amphibian Populations

GLOSSARY

amphibian Member of a class of vertebrates (the Amphibia), comprising frogs and toads (order Anura), newts and salamanders (order Caudata), and caecilians (order Gymnophiona), which typically return to water to breed and pass through an aquatic larval stage with gills. Amphibians have a moist skin without scales, which is permeable to water and gases.

reptile Member of a class of vertebrates (the Reptilia), comprising turtles and tortoises (order Testudinata), lizards, snakes, and worm-lizards (order Squamata), the tuatara (order Rhynchocephalia), and crocodiles and alligators (order Crocodylia), which typically lay eggs with a leathery, impermeable shell. Reptiles have a dry, horny skin with scales, plates, or scutes.

THE EARTH CURRENTLY CONTAINS approximately 7150 species of reptiles and 4600 species of amphibians. These data are underestimates because new species are still being discovered, partly as a result of exploration of previously unknown parts of the world and partly because new genetic techniques have revealed that what were thought to be single species are in fact several species. Many reptile and amphibian species are currently listed as threatened or endangered, and several have recently become extinct. Current estimates of the number of species at risk of extinction do not reflect the true extent of the threat faced by these two groups. Both groups are threatened by a variety of environmental factors.

I. REPTILES AND AMPHIBIANS

The reptiles and amphibians are two groups of vertebrate animals that are commonly grouped together, and the study of the two groups is called herpetology. In reality, however, the biological differences between reptiles and amphibians are more numerous and extensive than are their similarities. Amphibians are descended from fishes and, like fishes, both amphibians and reptiles are ectothermic, meaning that the heat that they require to maintain physiological processes is derived externally, directly or indirectly from the sun. Reptiles are the ancestors of the birds and the mammals, which are both endothermic, meaning that body heat is primarily generated within their bodies. Their dependence on external heat sources limits both amphibians and reptiles to tropical and temperate regions of the earth and both groups are especially numerous in the tropics;
80% of the world’s amphibian and reptile species live in tropical habitats. There are two major differences between amphibians and reptiles. First, amphibians have a thin, permeable skin that severely limits their ability to retain water within the body. In contrast, reptiles are covered in impermeable scales and are able to retain water very effectively. Second, the eggs of amphibians are bound by only a thin membrane and can only survive if they are kept wet. The eggs of reptiles are covered by a hard shell, providing a sealed environment for the developing embryo, which does not depend on external sources of water. These two differences mean that, whereas reptiles have successfully colonized very dry habitats, including deserts, amphibians are restricted to those habitats in which water is available for all or most of the year.

Neither reptiles nor amphibians are as well-known, in terms of documentation of species, as either the birds or the mammals. Neither group has been studied as intensively as birds and mammals, both are rather secretive in their habits, and both are very numerous in parts of the world, notably tropical forests, that have not been fully explored. Therefore, new species of amphibians and reptiles are being discovered and described at a much higher rate than is the case for birds or mammals. The number of recognized species of amphibians is approximately 4600 and of reptiles approximately 7150, but both these figures are continuously being revised upwards as new species are described.

There are three major groups of amphibians: the frogs and toads (approximately 4500 species), the salamanders and newts (approximately 415 species), and the caecilians (approximately 165 species). Reptiles are divided into four major groups: the lizards, snakes, and worm-lizards (approximately 6850 species); the turtles and tortoises (approximately 2200 species); the alligators and crocodiles (approximately 22 species); and the tuataras (1 species).

II. THE CONSERVATION STATUS OF REPTILES AND AMPHIBIANS

An accurate estimate of the number of reptile and amphibian species that are currently threatened with extinction, or which have recently become extinct, is not currently available. Table I presents two recent estimates of the numbers of endangered and threatened reptiles and amphibians. The disparity between the two sets of data, and in the total numbers of species that they each assume, reflects the current incompleteness of our knowledge of these two groups and of their conservation status. Currently, major efforts are being made, for example, by the Species Survival Commission of the International Union for the Conservation of Nature (IUCN), to correct this situation by compiling accurate lists of reptile and amphibian species and making assessments of their current status.

<table>
<thead>
<tr>
<th>Taxonomic group</th>
<th>Number of species</th>
<th>Threatened A</th>
<th>Threatened B</th>
<th>Endangered A</th>
<th>Endangered B</th>
<th>Approximate number of described species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turtles and tortoises</td>
<td>78</td>
<td>8</td>
<td>11</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sphenodon</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lizards</td>
<td>43</td>
<td>14</td>
<td>9</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snakes</td>
<td>35</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crocodiles and alligators</td>
<td>15</td>
<td>2</td>
<td>11</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>REPTILES</strong></td>
<td><strong>170</strong></td>
<td><strong>31</strong></td>
<td><strong>38</strong></td>
<td><strong>81</strong></td>
<td></td>
<td><strong>4771</strong></td>
</tr>
<tr>
<td>Salamanders and newts</td>
<td>25</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frogs and toads</td>
<td>32</td>
<td>2</td>
<td>6</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AMPHIBIANS</strong></td>
<td><strong>57</strong></td>
<td><strong>3</strong></td>
<td><strong>8</strong></td>
<td><strong>18</strong></td>
<td></td>
<td><strong>4014</strong></td>
</tr>
</tbody>
</table>

III. THREATS TO REPTILES AND AMPHIBIANS

Many environmental factors have been identified that are detrimental to the continued survival of reptile and amphibian species. Some of these are common to the two groups of animals; others are more relevant to one group than the other.

A. Habitat Destruction

The single most important factor that adversely affects reptiles and amphibians is habitat destruction and modification. Many habitats that are essential for the two groups are shrinking or disappearing at an accelerating rate as a result of human population growth and economic development. Tropical forests provide some of the most species-rich habitats in the world and are particularly vulnerable to destruction by humans, with the land being turned over to agriculture or to provide residential areas for people. In 1991, the United Nations Food and Agriculture Organization reported that the world’s tropical forests were being destroyed at a rate that was 50% faster than a decade previously. Tropical forests support a high diversity of reptile and amphibian species, many of them still to be described. At the current rate of deforestation, within 30 years all extensive areas of tropical forest, together with their reptile and amphibian fauna, will have disappeared.

Amphibians are particularly dependent on freshwater habitats, especially for breeding. The World Wildlife Fund (1998) published a report which suggests that freshwater habitats are the most seriously threatened on Earth. Between 1970 and 1995, the diversity of freshwater species decreased by 50%, a faster rate of decline than has been detected in any other component of the global ecosystem. Although there is legal protection for some of the world’s larger lakes and rivers, or at least parts of them, largely because of their importance in providing fish to support the human population, the smaller streams, ponds, and swamps that are essential habitat for amphibians are generally not protected. The draining of wetlands to make way for agriculture and housing has had a particularly serious negative effect on many amphibian populations.

In much of western Europe, traditional methods of agriculture previously provided good habitat for amphibians in the form of small woodlands, hedgerows, and numerous ponds created to water livestock. In the past 30 years, however, agricultural practices have changed and all these landscape features have been destroyed over very large areas. In parts of Britain, for example, the number of ponds suitable for amphibians to breed in has declined by 90% in the past 50 years. Habitat destruction in a given area is not always total; often, small pockets of forest, heathland, or wetland are set aside for conservation purposes. This results in the fragmentation of previous areas of habitat and there is increasing evidence that habitat fragmentation is a serious threat to the continued survival of species that it was assumed were afforded some degree of protection. Fragmentation of habitat leads to the isolation of small populations of reptile or amphibian species. Inbreeding in such populations reduces their genetic diversity, and their isolation prevents interchange of individuals and thus of genetic variation with other populations. As a result, isolated populations tend to decline slowly and eventually die out, even though they are protected. Many reptiles and amphibians have only limited powers of dispersal, and even a road built through an area of otherwise suitable habitat will reduce dispersal.

B. Climate Change

There is increasing evidence that the earth’s climate is undergoing major changes as a result of human activities such as the destruction of forests. Most notably, average temperatures are steadily increasing in many parts of the world and there are major, long-term changes in rainfall patterns. The long-term effects of such changes on reptiles and amphibians are largely a matter of speculation, but there is evidence that they have had an impact on some species. For reptiles, temperature is crucial because many species rely on basking in the sun to raise their body temperature to a level at which they can carry out essential activities such as feeding and reproduction. For amphibians, rainfall patterns are critical because they determine whether, and for how long, the small streams and ponds in which they breed contain water.

The apparent extinction of the golden toad (Bufo periglenes), along with several other frog species, in the montane forest of Costa Rica appears to be due, at least in part, to climate change (Pounds et al., 1999). A critical factor in the habitat of this species is low-lying cloud, which provides the water that maintains water flow in the small streams in which many of the native frogs breed. The extent of such cloud has decreased in recent years, leading to extensive reductions in the amount of available water and thus a general reduction in stream habitat. In Britain, winters are becoming increasingly less severe and there is evidence that several
native amphibians are now breeding earlier in the spring than they were 20 years ago.

C. Ultraviolet Radiation
The steady erosion of the ozone layer in the earth's stratosphere has led to an increase in the amount of ultraviolet (UV) radiation reaching the earth's surface. Such radiation, especially UV-B, is harmful to living organisms and there is accumulating evidence that it may be a factor in the decline of some amphibian species, especially those living at high altitudes, where incident levels of UV-B are highest. Increased UV-B causes the genetic material DNA to mutate, leading to the abnormal development and eventual death of embryos. Frog and toad species that lay their eggs close to the surface of water, which filters out UV-B, may be particularly susceptible to increased UV-B. Although there is considerable experimental evidence that ambient levels of UV-B radiation are harmful to the early life stages of amphibians (Blaxter et al., 1993), it is not clear to what extent elevated UV-B has been a factor in the decline of amphibian populations in nature (Alford and Richards, 1999).

Amphibian species vary in their susceptibility to elevated UV-B. Some species, such as the Pacific tree frog (Hyla regilla), produce high levels of the enzyme photolyase, which repairs DNA damage of the kind caused by UV-B. This species is one of those that has not declined in the Pacific Northwest of North America, unlike species with lower photolyase levels, such as the Cascades frog (Rana cascadensis) and the western toad (Bufo boreas) (Blaxter et al., 1994). If elevated UV-B does affect natural populations of amphibians, it is likely that it acts synergistically with other adverse environmental factors, such as acidification and pathogens. There is experimental evidence that both these factors have a more detrimental effect on the survival of amphibian embryos when they are combined with elevated UV-B.

D. Pollution
Environmental pollution has caused the decline or extinction of some local populations of amphibians, and it is likely that it also has a widespread harmful effect. For example, declines of amphibians in Yosemite National Park in California appear to be due to chemical pollution that has drifted on the wind from agricultural areas many miles away. Amphibians may provide sensitive biological indicators of pollution because their highly permeable skin rapidly absorbs toxic substances.

Examples of pollutants include fertilizers, herbicides, pesticides, heavy metals, and poisoning resulting from logging and mining operations. A major form of pollution that can affect very large areas is atmospheric acid deposition or acid rain. Much of the rain that falls in regions downwind from major industrial areas, such as in the eastern United States, Scandinavia, and western Europe, is markedly acidic, with a pH of approximately 4.3; unpolluted rain has a pH of approximately 5.6. Acid rain lowers the pH of natural water bodies below the tolerance level of many species of amphibians. The toxic effects of low pH on amphibian development are well-documented. Acidic conditions reduce the mobility of sperm and may cause them to disintegrate, with a consequent reduction in the fertilization success of eggs. Eggs that are fertilized may develop abnormally, if they hatch, they produce deformed tadpoles that soon die. Acid precipitation has been implicated in the declines of tiger salamanders (Ambystoma tigrinum) in the Rocky Mountains of Colorado and of the natterjack toad (Bufo calamita) in lowland heaths in Britain.

Pollution in the form of solid plastic waste is harmful to some sea turtles that may eat it. Green turtles (Chelonia mydas) eat plastic bags while feeding on aquatic plant food and leatherback turtles (Dermochelys coriacea) mistake plastic bags for their jellyfish prey. Half of the sea turtles examined in some localities have plastic debris in their intestines. This may interfere with their digestion, respiration, and buoyancy, and some plastics are toxic.

Some chemical compounds of human origin are readily absorbed by animals and interfere with their endocrine systems; these are known as endocrine disrupters and include estrogen mimics, which are compounds that disrupt the reproductive development of both sexes (Stebbins and Cohen, 1995). Males may become feminized to varying degrees, for example, suffering lowered sperm counts. Reptiles are especially susceptible to such effects because of their sex-determination mechanisms which, in many species, are very dependent on environmental factors. Of particular concern are polychlorinated biphenyls (PCBs), industrial chemicals such as those used in fire retardants and adhesives, which persist for a very long time and accumulate in the environment. Some PCBs have a molecular structure so similar to estrogen that they mimic its effects when they enter an animal's body. PCBs can turn male red-eared slider turtles (Trachemys scripta) into females, drastically reducing the reproductive success of populations.

Many pesticides, such as DDT, are not only poison-
During the 1990s, mass mortalities occurred among amphibians in many parts of the world have occurred, and some of these have had a major impact on the populations of many species. These were caused by ranaviruses similar to pathogens that have caused mass mortalities among tiger salamanders (Ambystoma tigrinum) at several localities in the United States and Canada during the past 5 years (Daszak et al., 1999). Recently, a disease called chytridiomycosis has been identified in Australia and Central America, where it has wiped out many populations of frogs, and among captive amphibians in the United States (Daszak et al., 1999). The disease is caused by a chytrid fungus, which is an organism that invades the skin of amphibians. Whether it kills them by blocking water and oxygen transport across the skin or by producing a lethal toxin is not known. These sudden, major outbreaks of the same, previously unknown disease in widely distant parts of the world raises several questions that are urgently being addressed: Is the chytrid fungus a new organism that has just evolved? Is it a long-established organism that has recently found its way to these localities and, if so, how? One possibility is that the fungus is not new and has long been distributed worldwide, and that what has changed is the susceptibility of amphibians to it. It has been suggested that, as a result of one or more kinds of environmental stress, the immune system of amphibians has been compromised so that they are no longer able to survive infection by the fungus.

Another kind of pollution which may affect amphibians in agricultural habitats results from nitrates derived from fertilizers. Recent research has shown that frog tadpoles reared in water containing nitrates at levels low enough to be considered safe for human consumption suffered physical abnormalities, paralysis, and death (Macne and Blaustein, 1999). Nitrate fertilizers are widely used throughout the world and may constitute a serious threat to many amphibian populations.

### E. Disease

Diseases are a natural cause of morbidity and mortality among animals, but very little is known about their role as a determinant of population size among reptiles and amphibians. In the past few years, however, major outbreaks of disease among amphibians in many parts of the world have occurred, and some of these have had a major impact on the populations of many species. During the 1990s, mass mortalities occurred among populations of the common frog (Rana temporaria) in southeast Britain. These were caused by ranaviruses similar to pathogens that have caused mass mortalities among tiger salamanders (Ambystoma tigrinum) at several localities in the United States and Canada during the past 5 years (Daszak et al., 1999). Recently, a disease called chytridiomycosis has been identified in Australia and Central America, where it has wiped out many populations of frogs, and among captive amphibians in the United States (Daszak et al., 1999). The disease is caused by a chytrid fungus, which is an organism that invades the skin of amphibians. Whether it kills them by blocking water and oxygen transport across the skin or by producing a lethal toxin is not known. These sudden, major outbreaks of the same, previously unknown disease in widely distant parts of the world raises several questions that are urgently being addressed: Is the chytrid fungus a new organism that has just evolved? Is it a long-established organism that has recently found its way to these localities and, if so, how? One possibility is that the fungus is not new and has long been distributed worldwide, and that what has changed is the susceptibility of amphibians to it. It has been suggested that, as a result of one or more kinds of environmental stress, the immune system of amphibians has been compromised so that they are no longer able to survive infection by the fungus.

Another kind of pollution which may affect amphibians in agricultural habitats results from nitrates derived from fertilizers. Recent research has shown that frog tadpoles reared in water containing nitrates at levels low enough to be considered safe for human consumption suffered physical abnormalities, paralysis, and death (Macne and Blaustein, 1999). Nitrate fertilizers are widely used throughout the world and may constitute a serious threat to many amphibian populations.

### F. Commercial Exploitation

Many people in many parts of the world have long eaten amphibians and reptiles because they are a good and readily available source of protein. Until recently, such exploitation rarely had a serious impact on natural populations because it was localized, seasonal, and of low intensity. Unfortunately, modern commercialization of amphibians and reptiles for the world’s luxury food market is generally done with little regard for the long-term protection of natural populations. Most of the frogs that are killed for human consumption are not a vital component of the diet of local people in the developing world but are a luxury item in the diet of people in developed countries.

The scale of trade in frog legs has long been substantial, but it is beginning to be controlled and reduced. In 1976, 2.5 million kg of frog legs was imported into the United States, mostly from India and Japan. In France, the annual consumption is estimated to be 3000 or 4000 tons, imported mostly from Bangladesh and Indonesia. Until recently, 200 million pairs of legs were exported annually from Asia to the United States, Europe, and Australia; however, since 1987, India has banned this trade because declines in natural frog populations had caused dramatic increases in the densities of insect pests.

There is also a substantial trade in reptile meat. Between 1979 and 1987, the hunting of alligators in Louisiana yielded 45,000 kg of alligator meat each year. Many turtles are heavily exploited as a food source, both by indigenous people and to supply luxury food markets in developed countries. The green iguana (Iguana iguana) has declined severely as the result of being hunted for food. Eaten by humans in Central America for centuries, destruction of its habitat and the expansion of the human population have led to it being hunted much more intensely in the past 30 years. In the late 1960s, as many as 150,000 iguanas were eaten each year in Nicaragua alone. Not only is iguana flesh regarded as a great delicacy but also their fat and their eggs are used for a variety of medicinal purposes, including a cure for impotence.

The highly durable skin of reptiles makes it an excellent alternative to leather, and reptile skins have long been used...
have considerable potential as medicines. For example, list of compounds, found in the skin of frogs, that their continued existence or ensure their survival, declarations is doubtful, but the future exploitation of amphibians and reptiles could either seriously threaten relations with them, deliberately or accidentally, a variety of organisms that have been harmful to indigenous wildlife. Populations of many reptile species that are endemic to islands have been devastated by the introduction of alien rats, cats, dogs, and other animals that feed on them or their eggs, and introduced pigs and goats have destroyed the ground cover that supports their food supply and which they need to escape from predators. The tuataras (Sphenodon) have become extinct on the two major islands and on many of the smaller islands of New Zealand and are now confined to a very few small islands that remain free of alien immigrants. Introduced rats have caused the extinction of several frog species of the genus Leiopelma on New Zealand; only three such species still exist, reduced to very small populations on remote, rat-free islands.

The eggs and larvae of amphibians make easy prey for fishes and other freshwater predators, and many amphibian species are dependent on water bodies that are free of fish to breed successfully. In many parts of the world, exotic fishes have been introduced by humans to the detriment of native amphibians. In California, trout have been introduced, to provide sport fishing, to high-altitude lakes and have caused the decline of several amphibian species; in many parts of the world, mosquito fish (Gambusia affinis) have been introduced to control mosquitoes and other insect pests, with equally disastrous results for native amphibians.

Some reptiles and amphibians have been implicated in the declines of other species where they have been introduced to parts of the world in which they did not previously occur. The most notable example is the marine toad (Bufo marinus), a native of Central America, which was deliberately introduced to Hawaii, Australia, and many other places to control insect pests in sugar-cane plantations. In Australia, where it is known as the cane toad, it has spread relentlessly outwards from the Queensland coast to the detriment of much of Australia’s native fauna, including many frog species. The American bullfrog (Rana catesbeiana), a native of eastern North America, has been introduced to many parts of the world to be farmed to supply the trade in frog legs, and it has had a harmful effect on many native frog species both because it is a greatly superior competitor and because it is a predator of smaller frogs. Brown tree snakes (Boiga irregularis), accidentally introduced onto the island of Guam, have seriously reduced or extirpated not only populations of endemic birds but also several native reptiles.

### G. Introduced Species

As the human population has expanded and people have colonized new parts of the world, they have taken with them, deliberately or accidentally, a variety of organisms that have been harmful to indigenous wildlife. Populations of many reptile species that are endemic to islands have been devastated by the introduction of alien rats, cats, dogs, and other animals that feed on them or their eggs, and introduced pigs and goats have destroyed the ground cover that supports their food supply and which they need to escape from predators. The tuataras (Sphenodon) have become extinct on the two major islands and on many of the smaller islands of New Zealand and are now confined to a very few small islands that remain free of alien immigrants. Introduced rats have caused the extinction of several frog species of the genus Leiopelma on New Zealand; only three such species still exist, reduced to very small populations on remote, rat-free islands.

The eggs and larvae of amphibians make easy prey for fishes and other freshwater predators, and many amphibian species are dependent on water bodies that are free of fish to breed successfully. In many parts of the world, exotic fishes have been introduced by humans to the detriment of native amphibians. In California, trout have been introduced, to provide sport fishing, to high-altitude lakes and have caused the decline of several amphibian species; in many parts of the world, mosquito fish (Gambusia affinis) have been introduced to control mosquitoes and other insect pests, with equally disastrous results for native amphibians.

Some reptiles and amphibians have been implicated in the declines of other species where they have been introduced to parts of the world in which they did not previously occur. The most notable example is the marine toad (Bufo marinus), a native of Central America, which was deliberately introduced to Hawaii, Australia, and many other places to control insect pests in sugar-cane plantations. In Australia, where it is known as the cane toad, it has spread relentlessly outwards from the Queensland coast to the detriment of much of Australia’s native fauna, including many frog species. The American bullfrog (Rana catesbeiana), a native of eastern North America, has been introduced to many parts of the world to be farmed to supply the trade in frog legs, and it has had a harmful effect on many native frog species both because it is a greatly superior competitor and because it is a predator of smaller frogs. Brown tree snakes (Boiga irregularis), accidentally introduced onto the island of Guam, have seriously reduced or extirpated not only populations of endemic birds but also several native reptiles.

### H. Species at High Risk of Extinction

The fact that introduced species, such as cane toads and the American bullfrog, can multiply and spread to the point at which they become a serious threat to other species indicates that population declines among amphibians are not due to a single factor that affects all species. Although many species are adversely affected by current environmental conditions, at least some can thrive. This raises the question of whether there are particular characteristics of certain reptiles and amphibians that make them particularly susceptible to environmental changes.

There are many reptile and amphibian species, notably many turtles, that live for a long time but have
delayed sexual maturity and low fecundity. Such life history characteristics limit the capacity of a population that has been reduced, for example, by human exploitation, to build up its numbers. Some species have very low reproductive rates because they do not breed every year. This is a feature of many amphibian species that live in near-desert habitats in which sufficient rain to fill their breeding ponds falls only occasionally.

Many amphibians move only small distances during the course of their lives. Such species have very poor dispersal abilities, with the result that, if their habitat is destroyed or modified, they do not have the option of colonizing suitable habitat elsewhere. There are many species that have very specialized habitat requirements and very restricted distributions. Such species are vulnerable to any very small change in their environment. Examples of such species are the golden toad (B. periglenes) of Costa Rica and two species of gastric-brooding frog (Rheobatrachus), restricted to a very few streams in southeast Queensland, Australia. All three species appear to have become extinct in the past 10 years.

Animals of all kinds that live on small oceanic islands are especially vulnerable to extinction, and reptiles and amphibians are no exception. Island species are often very vulnerable to introduced predators, such as cats, rats, dogs, and mongooses, having evolved in an environment in which such predators were previously absent. Small islands are also highly susceptible to habitat destruction and excessive hunting. Among endangered reptiles that live on islands are giant tortoises such as those found on Aldabra and the Galapagos.

Some reptiles and amphibians are vulnerable to exploitation by humans and to natural predation because they gather to breed in large numbers. A marine turtle, Kemp's ridley (Lepidochelys kempi), was reduced to near extinction in the 1950s by the commercial exploitation of breeding females and eggs, gathered at their coastal breeding sites in Mexico and Texas. This species is now internationally protected and its numbers are recovering.

Migratory species, such as some sea turtles, are very vulnerable both because their conservation requires that they be protected in two, often distant, parts of the world and because they have to make long journeys across possibly hazardous environments. The green turtle (Chelonia mydas) is protected in Australia but is exploited in Indonesia. Many amphibians migrate to specific breeding sites that are used year after year. High mortality can occur when these sites are near roads. In some places in the United States and Canada, roads that cross frog or salamander migration paths are closed in the spring, and in several places in Europe special tunnels have been built to allow frogs and toads to pass safely under roads.

In the Monteverde cloud forest reserve in Costa Rica, the formation of large mating aggregations of the spectacularly colorful golden toad (B. periglenes) used to be an annual event. In 1989, however, this species failed to appear and it has not been seen since (Pounds and Crump, 1994). The golden toad has become the icon of the declining amphibian phenomenon, but it is only one of many species throughout the world that have disappeared because of this deeply disturbing process. The Monteverde reserve was established to protect biodiversity and yet, since 1990, 40% of its native frog and toad species have disappeared (Pounds et al., 1997). Similar catastrophic declines of amphibian populations in supposedly pristine, protected habitats have occurred in Queensland, Australia, the Atlantic forests of Brazil, and the Pacific Northwest of North America. In a section of California's Sierra Nevada that includes Yosemite National Park, 5 of 7 amphibian species have seriously declined in recent years. In protected remnants of tropical rain forest in eastern Queensland, 14 species of stream-dwelling frogs have drastically declined or totally disappeared. These events carry a disturbing message for conservationists: As one scientist said at a recent workshop on amphibian declines, “locking up nature just isn’t working.”

Although it is clear that the majority of the world's amphibians that have declined or become extinct in the past 50 years have done so as a result of habitat change or destruction, these declines in protected areas suggest that some other kind of process is adversely affecting amphibian populations. Because of their dual life history, spent partly in water and partly on land, and because none of their life stages (egg, larva, or adult) have the kind of protective covering possessed by animals such as insects, reptiles, birds, and mammals, it has been argued that amphibians are especially sensitive to environmental insults such as chemical pollution (Stebbins and Cohen, 1995). According to this argument, amphibian declines may be the prelude to an environmental catastrophe that could affect many forms of life. Like the coal miner’s canary, frogs may be providing an early warning to all biodiversity.

It is becoming increasingly clear that this kind of argument is not applicable to all amphibians. Indeed,
a feature of all the declines that have occurred recently in supposedly protected habitats, such as Monteverde, is that although some species have declined or vanished, others have been quite unaffected. It is not clear, however, what character or characters differentiate amphibian species that have declined in pristine areas from those that have not. Amphibians are a very diverse group of animals in terms of their habits, their life histories, and their physiology. For example, although some species do have a highly permeable skin and are extremely sensitive to pollutants, for others the skin is highly effective as a protective covering. The American bullfrog (R. catesbeiana), for example, seems to be remarkably resistant to the effects of pesticides.

The current intense interest in amphibian population declines began in 1989 at the first World Congress of Herpetology. In 1991, the Declining Amphibian Populations Task Force (DAPTF) was set up under the aegis of the Species Survival Commission of IUCN, and by 1993 more than 500 populations of amphibians on five continents had been listed as declining or of serious conservation concern (Vial and Saylor, 1993). The DAPTF is continuing its work and plans to publish major overviews of the current status of amphibians throughout the world in 2001. Currently, it is clear from the information already gathered by the DAPTF that (i) amphibian population declines are widespread throughout the world and (ii) that they are not due to a single, global cause. Rather, they are the result of many factors, with more than one factor usually being implicated in any particular instance.

A. A Possible Link with Deformities?

Since a group of children found some severely deformed frogs in Minnesota in 1991, there has been much media attention devoted to this phenomenon in the United States, often making a link with amphibian declines there and elsewhere. These deformities reflect abnormalities in development and include missing eyes, digits, or entire limbs as well as extra limbs, sometimes growing from unexpected parts of the body. There are several reasons for being cautious about a possible link between deformities and the global decline phenomenon, however. First, such deformities are not new but rather have been reported periodically for more than 200 years, especially in the mid-northern United States. Second, although reports have become more frequent recently, there is no reason to believe that deformed frogs have become more common; there are simply more people looking for them. Third, developmental abnormalities are caused by a variety of natural causes, including parasites and injury following attack by predators, and by anthropogenic factors, such as chemical pollution and increased UV-B radiation.

Deformed amphibians are a cause for concern because they are often symptomatic of local environmental degradation. Before we can conclude anything of wider significance from them, however, we have to remember that they are both naturally occurring and typical of amphibians and that they do not represent a new phenomenon. Only additional research into the role of various pollutants and of UV-B into the development of deformities can reveal to what extent they are another facet of processes that are adversely affecting amphibians on a global scale.

See Also the Following Articles

AMPHIBIANS, BIODIVERSITY OF • ENDEGARED BIRDS • ENDEGARED MAMMALS • REPTILES, BIODIVERSITY OF

Bibliography


I. Evidence of Multitudes of Endangered Terrestrial Invertebrates

II. Awareness and Understanding of Endangered Terrestrial Invertebrates

III. The Special Perils of Specialists

IV. The Significance of Endangered Terrestrial Invertebrates

V. No Second Chances

GLOSSARY

alien invertebrates Invertebrates intentionally or accidentally imported by humans into new geographic areas.

coevolution Long-term evolutionary adaptation of species to each other (e.g., mutually beneficial relationships between bees and flowering plants).

endangered species Species that are likely to become extinct in the near future because of normal human activities. Examples of such activities are land clearing for agriculture or housing and accidental importation of invasive species through commerce.

endemic species Species confined to areas where they evolved (e.g., flightless crickets in the Hawaiian Islands).

habitat specialists Species found only in a specific habitat (e.g., species found only in forests of Sequoia trees).

terrestrial animals Animals that live on land for their entire lives (e.g., spiders) as opposed to animals that live in water for their entire lives (e.g., lobsters). There are also many amphibious animals that spend part of their lives in both places (e.g., dragonflies).

terrestrial invertebrates Animals that are not vertebrates (such as fish, reptiles, amphibians, birds, and mammals) that live on land for their entire lives. Examples are all insects except for those with aquatic larvae; nonaquatic mites and nematodes; and all spiders, millipedes, centipedes, and scorpions.

SIMPLE STATEMENTS CAN RAISE the most complicated questions. There are two simple facts about endangered terrestrial invertebrates. The first is that there are great numbers of these endangered animals—many thousands of species. The second is that nobody knows enough to make a comprehensive list of these species. Without such a list, how is it possible to say whether this list would be long or short and whether it would really contain many thousands of species? What is the evidence that there are huge numbers of endangered terrestrial invertebrates that are missing from lists of endangered species? Why aren’t biologists moving more quickly to add the missing names to the list? Are terrestrial invertebrates particularly vulnerable to ongoing rapid changes caused by our own species? If there are such large numbers, what does this mean for the forests, fields, deserts, and other habitats that are threatened...
with the loss of these species? What does it mean for our own species?

I. EVIDENCE OF MULTITUDES OF ENDANGERED TERRESTRIAL INVERTEBRATES

The belief that there are very large numbers of endangered terrestrial invertebrates is based on a combination of knowledge and logic. The great majority of all animals are terrestrial invertebrates, in which are included at least three-fourths of the insects, most mites, terrestrial mollusks, a large percentage of the nematodes, and all the spiders, millipedes, and centipedes. The last three groups total more than all the vertebrates. If the factors that endanger other animals also affect terrestrial invertebrates, the number of endangered terrestrial invertebrates must be large indeed. These factors are primarily habitat destruction and the introduction of nonnative species into new areas, and it is known that they affect terrestrial invertebrates.

Hawaii, showcase of beleaguered biota, is a revealing example of these factors at work on terrestrial invertebrates. Of the total number of animals that live on the Hawaiian Islands and in the surrounding sea, about three-fourths are terrestrial invertebrates, 99% of which are found only on the Hawaiian Islands. The islands are so strongly affected by habitat destruction that whole habitat types, containing all their habitat-specific invertebrates, are considered endangered. Invasions of alien plants and animals are so severe that whole groups of invertebrates have been devastated, such as the Hawaiian land snails, which were originally believed to number approximately 1000 species but now number approximately 500, with most of the remaining species endangered. The Hawaiian Islands may provide the best opportunity to document relatively easily the scale of the threat to terrestrial invertebrates in especially vulnerable places, such as oceanic islands (Fig. 1).

A. Logistical Problems

A major logistical problem is an insufficiency of experts. There are so many species of terrestrial invertebrates that it is easy to find groups with hundreds or thousands of species that can only be identified by two or three people. Who can identify the dark-winged fungus gnats, Indonesian bark beetles, the egg parasites of tropical sac spiders, the mites associated with millipedes, or any one of innumerable large groups of small organisms? The few people who do study these groups are generally overwhelmed with species that are undescribed and with species whose biology is almost completely unknown. Recognizing any of these species as endangered requires documentation of distribution and abundance—levels of information that will probably not be available for most terrestrial invertebrates for approximately another century (Fig. 2).

A second logistical problem is that specialized knowledge is needed not only for identifying most terrestrial invertebrates but also for finding and counting them. Species that are rare in museum collections are not necessarily rare in nature; they are frequently species that are difficult to find because their habits are unknown. Certain flightless pygmy mole crickets, for example, are very common in ancient dune areas in Florida, but until recently there were no specimens in museums because nobody knew to look for these insects just under the surface of the sand just after a heavy rain. Many grasshoppers have species-specific songs which biologists must learn to study the abundance of the
FIGURE 2 With regard to terrestrial invertebrates, there may be few people with the specialized knowledge to identify even large and easily characterized species. Many species could become rare or extinct without triggering alarms. The giant lacewing, Polystoechotes punctatus, apparently a common and widespread insect a century ago, is now so rare that few entomologists have seen a living individual. Nobody is writing distraught letters asking about the disappearance of this animal, and it does not appear on any list of endangered species. Nobody is doing any methodical surveys for this species; nobody knows how one would go about making such a survey because the ecology of the insect is poorly known. This species, which is relatively large (7-cm wingspread), distinctive, and of special interest because it is so primitive, has gone unrecognized as an endangered species. The great number of smaller and less conspicuous invertebrate species have that have become rare or extinct have even less of a chance of attracting attention.

Given all these logistical problems, however, there are still ways of quickly identifying large numbers of genuinely endangered invertebrates. There are whole groups of invertebrates that include many habitat specialists and have poor dispersal abilities, and therefore they are likely to be endangered if their habitat is reduced to small fragments. Examples of these unfortunate animals are the land snails, millipedes, flightless beetles, flightless grasshoppers and crickets, and several groups of primitive flies. It would be relatively easy to compile a long list of endangered invertebrates by surveying any old and distinctive habitat that is rapidly disappearing. There are some ecological groups of invertebrates that tend to have specific, coevolved relationships with other organisms and are endangered by anything that threatens their host. Examples of these excessively specialized invertebrates are most leaf-mining and gall-making insects and mites, seed predator insects, insects and mites associated with plant genera that contain only one species, pollen feeders that visit one or a few species of flowers, and the fleas, lice, feather mites, and den inhabitants of vertebrates. Since most species of endangered plants and vertebrate animals are likely to have at least one host-specific invertebrate, it would be relatively easy to compile a long list of endangered invertebrates by studying the invertebrate adult insects; this has recently become a tool for studying endangered grasshoppers. Estimating the abundance of a species also requires specific knowledge of life cycles and population dynamics. Many terrestrial invertebrates have a dormant period in the egg or pupal stage, during which they are almost impossible to find. Invertebrates often produce large numbers of offspring per female, and the percentage of survival of these offspring may be strongly dependent on variable factors, such as weather or the population levels of certain predators. This means that populations may go through frequent fluctuations of abundance and rarity, neither of which may be good predictors of the long-term survival of a population (Fig. 3).

FIGURE 3 The red widow spider (Latrodectus bishopi) is rare in the sense that it has a very small geographic range (a few areas of the Florida peninsula) and a strict habitat requirement (Florida scrub). Within its remnant patches of habitat, the red widow is sometimes very common, whereas at other times it is extremely rare. This is probably due to predation by natural enemies such as spider wasps and egg sac parasites. If one took a single survey in an area, one might conclude that this species was endangered or not at all endangered, depending on the population level at the time of the survey. Is this species endangered? Nobody knows. It depends on whether there are factors that could wipe out populations when they are at their low point, since the habitat for the species is now in smaller patches that are farther apart.
associates of larger and better known endangered species.

In summary, there are currently insurmountable difficulties that prevent the recognition of more than a small percentage of the probable number of endangered terrestrial invertebrates. Even this small percentage of easily recognized species, however, is not receiving the intensive attention that one might expect, which suggests that there are additional problems with recognizing endangered invertebrates.

B. Strategic Considerations

There are two species of bird lice (Franciscoa thompsoni and Neopsitticonirmus emersoni) that live only on the threatened Philippine cockatoo. These are examples of endangered invertebrates that might go unrecognized for strategic reasons. First, lice have no general appeal, and to say that this particular parrot is infested with unique and interesting species of lice is most unlikely to further spur efforts to save the bird. Simultaneously, there seems to be no strategic need to recognize the endangered species of lice because their survival depends on that of their host, whose status is already recognized. In the long run, however, it is important to understand that each species of organism represents a complex of ecological relationships, and the presence of a host-specific parasite is indicative of ancient sequences of adaptation which add to the significance of the host species. Moreover, host animals may be molded by their parasites in ways that are just beginning to be acknowledged by biologists. Therefore, for example, mutual grooming behavior may have evolved through the presence of body parasites on places that are difficult to reach, and this mutual grooming may have a major role in social bonding between individuals. Some biologists also believe that strenuous courtship performances and extravagant male ornamentation may evolve to demonstrate vigor and resilience in the face of the parasite load that is borne by almost all wild animals (Fig. 4).

At a more general level of strategy, conservationists are caught in the same contradiction that afflicts all biological educators. On the one hand, simple messages are the most effective. On the other hand, life is irreducibly complex. The story of the Florida scrub jay, for example, is complicated enough without considering all the grasshoppers, beetles, millipedes, and other invertebrates that are also restricted to Florida scrub habitat. The choice is often made to focus on a few species whose protection through habitat preservation will automatically protect a large number of other species, such as terrestrial invertebrates. It may seem that this is an unscientific approach, one that panders to a perceived bias against small, crawling creatures. In reality, it may be more a question of available information—both scientific information and information that is part of our human heritage. The biology of the Florida scrub

FIGURE 4  The sucking louse, Phthirpediculus propitheci, is a parasite on the body of the endangered sifaka lemur, Propithecus diadema. This louse, which is indubitably endangered because of the status of its host, seems an ideal candidate for Gilbert and Sullivan’s “little list of society offenders who might well be underground, and never would be missed!” Maybe not, however, because sifakas, like most other primates, do much of their socializing during episodes of mutual grooming, and the removal of the stimulus for this interaction could change the intimate and supportive behavior of the species.
jay is known in great detail, including the complex of factors that threaten its existence, whereas the biology of the Florida scrub millipede is almost completely unknown. Likewise, almost everybody has an intuitive understanding of the significance of habitat loss to the Florida scrub jay because this bird, like our own species, is relatively intelligent, home-loving, hierarchical, territorial, and lives in nuclear families. Accurate and detailed communication about the world of the Florida scrub millipede is not currently possible. If one were to make the choice between studying a few endangered species in detail or studying many endangered species more superficially (a choice that few biologists actually make), it might make sense to study a few species in-depth (Fig. 5).

In the long run, however, the study of a great variety of endangered species provides a wealth of fine-grain information on management and biogeography that is omitted when pandas or tigers are used as "umbrella species" whose protection helps to protect thousands of species that share the same habitat. A portion of the earth’s endangered biota is contained in sites that do not support any endangered vertebrate species. The endangered blue butterflies of the genus *Maculinea* often occur in small sites in which there are no endangered vertebrates. Certain small islands near the larger islands of New Zealand lack remaining endangered vertebrates but have retained some endangered invertebrates. Perhaps this millipede also has some special microhabitat requirements.

III. THE SPECIAL PERILS OF SPECIALISTS

The very factors that have made the terrestrial invertebrates such a huge, successful group have guaranteed that large numbers of their species will become endangered as humans change and destroy natural habitats. The small size of terrestrial invertebrates has allowed them to specialize on miniature resources. Huge numbers of species, for example, feed on only one part of a single species or genus of plant. Many other species are internal parasites in a narrow range of insect hosts. The great advantage to this specialization is that it has allowed spectacular efficiency in finding and exploiting resources. The small bark beetle, *Cactopinus hubbardi,*
which raises its larvae in the injured tissue lining the cavities made by woodpeckers in saguaro cactus, only needs to deal with the nutrients and defensive compounds of one kind of plant. It is probably able to zero in on its breeding sites, which are unlikely to be numerous in any one area, by following an odor plume emanating from the injured cactus.

Since terrestrial invertebrates are able to rest in a dormant state with minimal energy expenditure, they can evolve life cycles synchronized with the availability of their resources, and this also increases the ability to specialize. The hundreds of species of solitary bees in arid habitats of Mexico and the southwestern United States can synchronize their emergence with the seasonal changes or the periodic rains that stimulate blooming in their plant hosts. This, combined with coevolution between the mouthparts of bees and the architecture of flowers, has led to many species-specific bee and flower relationships. The evolution of many specializations is driven by the benefits of greater efficiency and less competition accrued by specialists. When humans disturb natural habitats, however, extreme specialization is a liability for many inhabitants because even temporary loss of a resource can eliminate the species that depended on that resource (Fig. 6).

The small size of terrestrial invertebrates allows them to maintain thriving populations of habitat specialists in a small area, such as an isolated mountaintop with some alpine habitat that was colonized at the end of a glacial period. Evolutionary biologists believe that speciation is most likely to occur in such isolated populations peripheral to much larger populations. The isolated subspecies (really species in the making) of butterflies known as arctics (Oenis) and alpines (Erebia) show that this process can occur over a relatively short time. In areas in which there has been long-term isolation of habitat fragments, it is often possible to find isolated populations that have diverged so much that they are clearly distinct species that could never merge, even if they were brought back together. Some groups of flightless Orthoptera provide good examples of this process of speciation at work. Many of these distinctive forms and species of terrestrial invertebrates could persist indefinitely in the small areas of habitat where they now occur, but these species and forms can be considered endangered because they could easily be eliminated by an episode of habitat destruction that would be small by current standards (Fig. 7).

Terrestrial invertebrates as a group are not at risk because there are many species that are generalists or are widely distributed. A significant proportion of the total diversity of invertebrates, however, is composed of species that are highly specialized or restricted to one or a few small patches of habitat.

**IV. THE SIGNIFICANCE OF ENDANGERED TERRESTRIAL INVERTEBRATES**

A sparrow falls, Thou art mindful; A spider is gone, art Thou vexed?

Why should anybody be concerned about endangered terrestrial invertebrates? Some people view this as a question of ethics: They believe that Homo sapiens is not the only species with a right to exist, or they believe that it is wrong to rob all future generations of the rich biological heritage that was passed down to us. It is true that there have been previous waves of extinction...
FIGURE 7 In Florida a series of isolated ancient dune fields have their own species of grasshoppers that are unable to live in the surrounding lowlands. Since these upland sites escape the flooding that is common in lower areas, they are well suited for housing and agriculture, and these grasshoppers have much reduced habitat and at least three appear to be endangered. Distinctive endemic species such as these grasshoppers also serve as biogeographic indicators, showing that the areas they inhabit have been isolated for a long time and might have additional endemic plants and animals. The number of examples of endangered Florida invertebrates appearing in this article is due to my familiarity with these species; there are many examples of endangered invertebrates almost everywhere.

of invertebrates caused by such events as ice ages or perhaps the impact of asteroids. Ethics, however, denies to us the innocence of an ice sheet or shooting star. A foundation stone of ethics is that we must endeavor to understand and take responsibility for the consequences of our actions. Since our species is strongly guided by ethics, ethical considerations may be as real and important as the dictates of materialism. The origin and nature of ethical attitudes toward other species have been considered by several scientists, especially Edward Wilson. Even from a materialistic standpoint, however, it makes sense to be concerned about endangered invertebrates for a variety of reasons.

First, endangered terrestrial invertebrates may be viewed as repositories of information. Currently, we have neither the time nor the skill to interpret this information. Some endangered invertebrates might produce useful chemicals. They might have innovative defenses against fungi or bacteria. They might have peculiar genetic systems or developmental pathways that are easy to analyze. Endangered species might inspire new areas of microengineering. They might display a variety of physiological mechanisms to deal with extreme environmental conditions. They may present more detailed evidence of evolutionary trends. They may be convenient indicators of environmental change. They might have novel types of mutualistic relationships with other animals or with plants. They may be remnant populations of formerly abundant species with important roles in restored ecosystems. Introduced pests might be controlled by invertebrates that are rare or endangered in their homeland.

In addition to these potential material benefits, non-material attributes of endangered invertebrates could have their own materialistic spin. Humans are willing to spend prodigious effort and money on nonmaterial things, such as entertainment, aesthetics, or the opportunity to make new discoveries. Many endangered species of terrestrial invertebrates might achieve their highest value as examples of the beauty and intricacy of life; considering the psychological needs of our species, this is an offering that can never come too often or in too many guises. Already, the conservation of several species of butterflies is pushed by the market value of aesthetically pleasing specimens or of live specimens for butterfly houses that charge admission.

With regard to terrestrial invertebrates, we are currently in an inspiring but frustrating state of ignorance in which any useful quality that we can imagine may well be represented in the group, but our imaginations are clearly insufficient to the task.

Second, the things that we care about are often dependent on small, unconsidered details: the linchpin that holds the wheel on the axle or the knot at the end of the thread anchoring the stitching. From this truth, it is easy to postulate that there are ecological systems that are tenuously held together by endangered invertebrates. Good examples might be found in old, isolated systems with a very limited number of species, such as a cave or an oceanic island. The situations in which a single species has this linchpin role are likely to be few and exceptional. Most large-scale ecological systems, such as a large area of rain forest in New Guinea or desert in Arizona, must have impressive built-in versatility or they would not have survived the many natural changes affecting the planet before the advent of our species. On the other hand, biological systems of all kinds can only withstand a certain degree of stress. It is difficult to recognize the approach of a breaking point,
much less identify the relatively small events that could serve as triggers.

Third, although a single rare and endangered species of invertebrate is unlikely to determine the fate of a large ecological system, rare invertebrates may be important in the aggregate (Fig. 8). In the remaining natural habitats of the world, especially those in warm and warm-temperate climates, most of the total diversity of invertebrates is composed of relatively rare species. Many of these rare species could quickly become endangered species if their habitat changes in some major way. This change could be general habitat disturbance, or it could be habitat fragmentation. It could also be caused by extensive contamination by pesticides to control agricultural pests. Aerial application of pesticides to control biting insects is a special problem because a reduction in the incidence of human disease can be used to excuse even the most destructive of practices. Most insect-borne diseases, however, are cycling in heavily settled areas; they seldom sweep out of large tracts of natural habitat. In the future, invertebrates might be endangered by genetically altered plants or pathogens that carry self-replicating pesticides from agricultural areas into other habitats. One of the greatest threats to rare invertebrates today is invasion by exotic invertebrates. All these kinds of threats, which often occur in combinations, have a single general effect on ecosystems: They replace many specialized invertebrate species with a few generalists. The cumulative effect of this loss of invertebrate diversity may be to threaten populations of large, noticeable organisms, such as certain large vertebrates or plants. At a more basic level, however, much of the efficiency and precision of energy flow through ecosystems relies on the rarer and more specialized species of invertebrates, just as the efficiency and precision in our communication relies on the availability of a huge number of rarely used and specialized words.

FIGURE 8 General ecosystem functions, such as recycling of nutrients, are most efficiently done by many specialists rather than by a few generalists. The plates that cover the shells of tortoises are made of keratin—a protein that, because of its strong disulfide bonds, few scavengers can digest. However, there is a moth, Ceratophaga vicinella, whose caterpillar appears to have a specialized diet of the shells of dead gopher tortoises. There are innumerable specialized invertebrates that feed on particular kinds of wood, bark, dung, and carrion. No single one of these species is likely to be important in an ecosystem, but as a group they have a large impact.

V. NO SECOND CHANCES

Each species of endangered invertebrate is an old and irreplaceable entity. This seems like an obvious point, but one so often hears of “new” species of invertebrates that it is easy to unconsciously absorb the impression that invertebrates are evolving at a prodigious rate. These new species are always old species that are newly discovered. It is true that in special circumstances invertebrate populations can diverge relatively rapidly, evolving behavioral and physiological differences specific to the different conditions in different sites. Most invertebrate species, however, are distinguished by morphological specializations that probably took a long time to evolve. This can be seen by examining the fauna of recently formed islands, such as the Bahamas. The Bahamas were submerged by the sea about 135,000 years ago and emerged again approximately 100,000 years ago. Although 100,000 years is a long time by human standards, it is evidently short by evolutionary standards—insufficient for the evolution of a major group of Bahamian species of terrestrial invertebrates or terrestrial plants.

Recently, systematists (specialists who study the evolutionary relatedness of animals and plants) have put forward the argument that special attention should be given to endangered species that are the last remnants of evolutionary lines that largely disappeared millions of years ago. It is noted that these relics may have unusual kinds of adaptations and also provide glimpses
of ancient life on our planet. Certain endangered terrestrial invertebrates would be high on the list of species that would benefit from special consideration of relics: Onychophora, giant mites, redwood Thysanura, and many primitive lineages known from small areas of the tropics or south temperate areas. Although there is some logic to priority for the most ancient and conservative of the endangered terrestrial invertebrates, unless systematists are willing to work closely with ecologists, the identification of relict lineages is not very useful. Moreover, although ancient relics may be important for historical reasons and because they embody peculiar adaptations, they may have less relevance to modern systems than species that evolved from widespread lineages only a few million years ago. For example, useful natural enemies of pests of solanaceous crops (such as potatoes, peppers, and tomatoes) are most likely to be found among the large numbers of wasps and flies that attack insects on wild Solanaceae in some remnant habitat in the Andes Mountains.

In summary, terrestrial invertebrate species, once lost, cannot "reevolve," and the generation of new species that might restore lost diversity occurs on a time-scale too long to be relevant to our own species. Every year there are more species of terrestrial invertebrates than there ever will be again, even should human civilization endure 100,000 years.

See Also the Following Articles
ENDANGERED FRESHWATER INVERTEBRATES • ENDANGERED MARINE INVERTEBRATES • INVERTEBRATES, TERRESTRIAL, OVERVIEW • TERRESTRIAL ECOSYSTEMS

Bibliography


Enemism

R. M. Cowling
University of Cape Town

GLOSSARY

biotope Region that is distinguished by particular environmental conditions (climate, soil, altitude, etc.) and therefore a characteristic assemblage of organisms.

stenotopic Referring to taxa with restricted habitat requirements (i.e., confined to a single biotope) and hence restricted distributions.

Endemic taxa are those restricted to a specified geographical area. Therefore, the concept is a relative one; the patterns, correlates, and causes of endemism will vary according to the size and location of the geographical area, as well as the taxonomy and phylogenetic relatedness of the assemblage under consideration. At a global scale, all taxa are endemic and there is relatively little to say on the topic. Most research has focused on species that are endemic to relatively small areas. In this sense, endemism is best viewed as a form of rarity, that is, range-restricted rarity. This article presents biogeographical, evolutionary, ecological, and conservation perspectives on endemism and discusses generalizations regarding the patterns, correlates, and causes of species-level endemism in relatively small areas.

I. CATEGORIES

Endemics may be categorized according to their spatial distribution, inferred evolutionary age, affinities, and local abundance.

A. Spatial Distribution

Endemics are loosely and commonly categorized in four contexts of spatial distribution: site or restricted area; biotope; biogeographical region; and political area.

B. Evolutionary Age and Affinity

Categorization of endemics according to evolutionary age and affinity is summarized in Box 1. These schemes have been widely used by botanists but rarely by zoologists. Problems associated with the schemes are that (1) age is regarded as a categorical rather than continuous variable; (2) the establishment of relationships among...
Box 1  
Categorization of Endemics According to Evolutionary Age and Affinities

A. Engler’s scheme, published in 1882.
1. Neoendemics: comprising clusters of closely related species and subspecies that have evolved relatively recently.
2. Palaeoendemics: comprising phylogenetically high-ranking taxa, usually monotypic sections, subgenera, or genera that may be regarded as evolutionary relicts.

R. C. Favarger and J. Constandriopoulos’s scheme, published in 1961. This scheme uses cytological data to provide a more rigorous basis for assessing the age and affinities of endemics.
1. Palaeoendemics: ancient isolated taxa with a high ploidy level, whose diploid ancestors are extinct or unknown.
2. Schizoendemics: vicariant species of equal ploidy level, resulting from either gradual or rapid divergence.
3. Patroendemics: restricted diploid species that have spawned younger, widespread polyploid species.
4. Apoendemics: polyploid endemics that are derived from widespread species of a lower ploidy level.

Schizo-, patro-, and apoendemics are further subdivisions of Engler’s neoendemics.

taxa lacks rigor; and (3) many palaeoendemics are diploid.

Phylogenetic methods, which consider the distribution of characters among taxa in a cladistic context, provide a rigorous categorization of endemics in terms of relative age and propinquity of descent. In this context, low-ranking taxa correspond to neoendemics and high-ranking taxa to palaeoendemics (cf. Box 1). An absolute estimate of the age of endemics can be given when congruent phylogenetic relationships correlate with identifiable historical events.

C. Local Abundance

The classical, biogeographical perspective on endemism has tended not to consider the local abundance of species. However, in the more recently developed ecological and conservation perspectives, in which endemism is conceived as a category of rarity, population abundance is invariably explicitly considered. Thus, geographical range size as a categorical variable (wide/narrow) has been used as one of the factors in defining seven forms of rarity recognized for plants. Endemics (narrow range) may belong to any four categories of rarity according to biotope specificity (broad/restricted) and local population size (somewhere large/everywhere small).

II. PERSPECTIVES

The concept of endemism has a long history in biology, dating back to A. P. De Candolle’s treatise, published in 1820. Most research on the topic has been in the field of descriptive biogeography, where distribution patterns of taxa have been used to define centers of endemism at various spatial scales. This approach provides a static perspective of endemism.

Over the past few decades, historical biogeographers have evaluated areas of endemism for monophyletic lineages in a phylogenetic context. This approach provides a dynamic perspective of endemism, especially when endemic taxa show congruent phylogenetic relationships that can be correlated with historical events. Evolutionary biologists, studying both fossil and extant biotas, have explored the role of range restriction as a cause and consequence of speciation. Recently, several statistical techniques have been employed—collectively termed the comparative method—to exploit the phylogenetic relationships among species to extract independent information on the evolutionary correlates of endemism. These techniques acknowledge that related species may have similar range sizes, that is, range size cannot be assumed to be independent among species. However, in at least some cases, variance in range sizes seems to be partitioned mostly at the species level.

Community ecologists have conceptualized endemism as one of several forms of rarity, namely, range-restricted rarity, and have explored its role as an explanatory variable for taxon-specific ecological traits, such as local population size, body size, reproductive fitness, and dispersal distance. Increasingly, they are using comparative methods to correct for phylogenetic relatedness among biotas. However, for every cause–effect relationship documented, there are numerous exceptions.

Conservation biologists view range-restricted rarity
as an attribute that predisposes a taxon to extinction. They seek to understand the abiotic and biotic correlates of this form of rarity as a basis for management guidelines that will reduce rates of extinction. A distinction is often, although not always, made between naturally rare species that may have some adaptation to rarity and those that have previously been widespread and are now restricted. Conservation planners often use patterns of endemism to identify reserve systems that are representative of a region's biodiversity. Many reserve selection algorithms have been formulated to select sites that have a unique or endemic complement of species.

III. MEASUREMENT

In quantifying patterns of endemism, the units of measurement (spatial scale and taxonomic entity), the mode of reporting of the data (percentages or counts), and a number of biases all influence the interpretation of the results. Of great importance is the relative nature of endemism: evaluation is always dependent on the spatial context and biological assemblage under consideration. This section provides a clarification of the problems and approaches associated with the measurement of endemism.

A. Units of Measurement

A variety of methods have been used to measure the range sizes of taxa. A useful distinction can be made between measures that attempt to estimate the extent of occurrence of a taxon—the distance between the outermost limits of a species' occurrence—and the area of occupancy—the area over which the species is actually found. The latter measure is particularly relevant for ecological studies that seek correlations between range size and environmental tolerances, as well as for conservation planning research; extent of occurrence is widely used in biogeographic studies.

Measures of endemism invariably seek to identify a subset of taxa within an assemblage that can be classified as having a lower than average range size. Within the biotas of larger-scale regions—biogeographic areas or countries—many researchers have recognized "local" endemics as a distinct category. However, the range size, or extent of occurrence, for defining this category is often arbitrarily set, varying between 50,000 km² (for Neotropical birds and plants, as well as for birds globally) to 2000 km² or less (for plants in the Cape Floristic Region). Endemics with extremely small range sizes—<5 km²—are regarded as point endemics. An approach that is increasingly being used is to evaluate endemism as a continuous variable, calculated as the sum of the inverse range sizes of all taxa in each quadrat (cell grid or map unit).

From both the biogeographical and ecological perspectives, patterns of endemism are best studied in relation to ecologically homogeneous, biogeographical regions. However, conservation planners often use political regions or property boundaries when evaluating endemism, since these may be the most effective decision-making unit for the preservation of endemics.

The taxonomic or phylogenetic scales employed also influence patterns of endemism. Centers of endemism identified on the basis of patterns among low-ranking taxa (sub-species or closely related species) often differ from those where the units are high-ranking members of the same lineage. Similarly, the spatial scale for defining endemism will vary among different taxa of the same rank.

B. Percentage versus Counts

Endemism may be expressed as a percentage of all extant taxa present, or as the absolute number of endemics in an area. Depicting plant endemics in biogeographic regions as percentages or counts, and using area and latitude as explanatory variables, results in different patterns with different significance (Fig. 1). Some species-poor areas, such as oceanic islands and arid regions, although low in actual numbers of endemics, may support a high percentage of endemic taxa. Others, such as Madagascar, the Cape Floristic Region, and parts of the Neotropics combine high richness and high endemism for some taxa. Ideally, both measures of endemism should be considered when explaining patterns, but seldom are.

C. Biases

Endemism is influenced by taxonomic interpretation, sampling error, and human perceptions of rarity. Of particular importance is the fact that limited geographical exploration, as well as variation in the application of taxonomic concepts, introduces biases in the identification of endemics and the significance of their status. Pseuendoendemics are widespread species incorrectly classified as endemics, whereas nonapparent endemics are endemic species that are incorrectly classified as widespread. The fact that widespread species are usually more thoroughly researched than those with smaller
range sizes introduces biases in studies that explore the correlates of range size.

IV. PATTERNS

There are very clear global and regional patterns of endemism for a wide range of taxa: endemics are not randomly distributed across the globe. However, these patterns are constrained by poor taxonomic knowledge and distributional data in key areas (e.g., the tropics) and for some taxa (e.g., most invertebrate groups).

A. Latitudinal Gradients

The incidence of endemism for whole assemblages in biogeographic zones increases with decreasing latitude (see Fig. 1). Range sizes, as measured by latitudinal extent, increase for a wide range of organisms above a latitude of approximately 40°–50°N, but the same patterns are not evident in the Southern Hemisphere. There are many patterns that are not consistent with the generalization—termed Rapoport’s Rule—that range sizes of taxa decrease with decreasing latitude, as a consequence of greater ecological specialization in less seasonal environments. For example, very high
endemism for terrestrial taxa is recorded in the mid-latitudes of the Southern Hemisphere, particularly in and adjacent to Mediterranean-climate regions. Marine teleost fishes have smaller range sizes at higher than at lower latitudes, and endemism for marine algae peaks in mid-latitude areas. These patterns are probably more a product of speciation and extinction processes than contemporary ecological conditions. Thus, widespread glaciation during the Pleistocene at high latitudes in the Northern Hemisphere resulted in the extinction of less tolerant terrestrial taxa. In mid-latitude Mediterranean-climate regions that escaped glaciation, rates of speciation, at least for plants, have overwhelmed extinction rates, resulting in an accumulation of habitat-specialist, range-restricted species.

B. Centers
Many centers of endemism—areas of higher than average concentrations of range-restricted taxa—have been recognized globally and regionally, principally for higher plants and large-bodied terrestrial vertebrate faunas. Generally, many groups of organisms show a concentration of centers at lower latitudes (Fig. 2). Following from the previous section, it is no surprise that the high-latitude areas of the Northern Hemisphere support few centers. However, this is not always the case in the Southern Hemisphere, where large numbers of range-restricted taxa occur in middle- to high-latitude landmasses that were never glaciated during the Pleistocene.

C. Congruence
Overlapping or congruent areas of endemism for different taxa have been used extensively by biogeographers to reconstruct historical events. Patterns of congruence of endemism are also important for identifying reserve systems that maximize the preservation of different biotas.

Although strong patterns of congruence have been recognized for some taxa at the global scale—for example, swallowtail butterflies and tiger beetles, amphibians, birds, and mammals—higher plant centers often do not coincide with faunal centers. Nonetheless, on the basis of congruent patterns of endemic species diversity for mammals, reptiles, amphibians, and higher plants, it has been possible to identify 17 “megadiversity countries” (Table I), that is, political units of high conservation value.

At a finer scale, patterns are highly variable among different taxa and in different regions, and no generalizations have emerged. This lack of strong congruence underlines the fact that endemism is an expression of many different causes, both ecological and historical.

---

**FIGURE 2** The distribution of Endemic Bird Areas of the globe, as recognized by Birdlife International. These centers are identified on the basis of the distributions of 2609 bird species that have had in historical times a global breeding range of less than 30,000 km². (Reprinted with permission from Bibby et al., 1992.)
TABLE I
Vertebrate and Higher Plant Endemism in the World’s 17 Megadiversity Countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Area (km$^2 \times 10^3$)</th>
<th>Mammals</th>
<th>Birds</th>
<th>Reptiles</th>
<th>Amphibians</th>
<th>Plants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>8512.0</td>
<td>131</td>
<td>172</td>
<td>294</td>
<td>17,500</td>
<td>ca. 17,500</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1916.6</td>
<td>201</td>
<td>130</td>
<td>100</td>
<td>ca. 10,000</td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td>1141.7</td>
<td>28</td>
<td>97</td>
<td>367</td>
<td>10,000</td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td>1072.5</td>
<td>123</td>
<td>368</td>
<td>169</td>
<td>ca. 12,500</td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>7606.8</td>
<td>210</td>
<td>616</td>
<td>169</td>
<td>14,498</td>
<td></td>
</tr>
<tr>
<td>Madagascar</td>
<td>587.0</td>
<td>77</td>
<td>133</td>
<td>174</td>
<td>ca. 10,000</td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>9361.0</td>
<td>77</td>
<td>133</td>
<td>174</td>
<td>ca. 10,000</td>
<td></td>
</tr>
<tr>
<td>Philippines</td>
<td>300.8</td>
<td>116</td>
<td>131</td>
<td>44</td>
<td>ca. 3000</td>
<td></td>
</tr>
<tr>
<td>India</td>
<td>3287.0</td>
<td>44</td>
<td>187</td>
<td>110</td>
<td>ca. 7500</td>
<td></td>
</tr>
<tr>
<td>Peru</td>
<td>1285.2</td>
<td>46</td>
<td>98</td>
<td>109</td>
<td>5316</td>
<td></td>
</tr>
<tr>
<td>PNG Papua New Guinea</td>
<td>475.4</td>
<td>37</td>
<td>79</td>
<td>134</td>
<td>ca. 13,000</td>
<td></td>
</tr>
<tr>
<td>Ecuador</td>
<td>285.6</td>
<td>21</td>
<td>114</td>
<td>130</td>
<td>ca. 4900</td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>9372.1</td>
<td>101</td>
<td>90</td>
<td>126</td>
<td>4036</td>
<td></td>
</tr>
<tr>
<td>Venezuela</td>
<td>912.1</td>
<td>11</td>
<td>57</td>
<td>76</td>
<td>ca. 6000</td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td>320.7</td>
<td>27</td>
<td>68</td>
<td>97</td>
<td>ca. 7250</td>
<td></td>
</tr>
<tr>
<td>South Africa</td>
<td>1221.0</td>
<td>27</td>
<td>76</td>
<td>36</td>
<td>16,300</td>
<td></td>
</tr>
<tr>
<td>Democratic Republic of Congo</td>
<td>2544.0</td>
<td>26</td>
<td>23</td>
<td>93</td>
<td>3200</td>
<td></td>
</tr>
</tbody>
</table>

*Figures in parentheses are rankings for the number of endemic species among the top 12 countries.

V. CORRELATES AND CAUSES

Range size or degree of endemism shows some clear relationships with a wide array of abiotic and biotic factors. These correlations are very useful in conservation biology since they may be used to identify factors that predispose endemic species to extinction. However, correlates may be either a cause or a consequence of endemism. To identify the causes of endemism in an evolutionary context, comparative methods that exploit phylogenetic relationships must be employed. The causes of endemism are complex and numerous, and include intolerance of widespread habitats, niche specialization, isolation in marginal habitats owing to climate change, phylogenetic predisposition to narrow habitat selection, competition from alien species, and recent speciation of isolates in marginal habitats. Therefore, historical processes, contemporary ecological factors, and inherent biological properties of lineages are involved. In many cases, historical factors may be overriding, resulting in a poor relationship between measures of endemism and explanatory variables reflecting the contemporary environment.

Establishing correlates is a useful step in explaining patterns and causes of endemism. Most pertinent studies have addressed the following question: When compared to more widespread taxa, are endemics, however defined, a random subset of the biota with regard to abiotic and biotic factors? Developing these profiles, however, has been complicated by different definitions of endemism, multiple interactions between different traits, and a failure to consider phylogenetic relatedness.
This section provides a brief review of the abiotic and biotic interspecific correlates of narrow range size, and concludes with an assessment of the role of endemism in speciation.

A. Regional Species Richness
There is often a positive relationship between the incidence of endemism and regional-scale richness. This results from the importance of high habitat-related and geographical compositional turnover (beta and gamma diversity, respectively) in producing regional richness. Habitat specialists (or stenotopic species) and geographical vicariants often have narrow range sizes. However, there are also many cases where patterns of endemism and diversity are largely noncoincident. Examples include plants in the Neotropics, birds in the Andes, dragonflies and terrestrial vertebrates in southern Africa, and the biotas of many oceanic islands and deserts.

B. Area
As a generalization, proportionate and absolute measures of endemism increase with increasing area (see Fig. 1), irrespective of the taxonomic level. However, the relationship between number of endemic species (counts) and area is not as tight as that for the more widely studied species–area relationship. This results from the lack of congruence between endemism and richness in many areas (e.g., and lands and oceanic islands).

C. Abiotic Environmental Factors
Levels of endemism may vary in a predictable way along gradients of rainfall, temperature, productivity, and habitat heterogeneity. Models that accurately predict levels of endemism on the basis of easily measurable environmental variables have been used for the rapid identification of endemic-rich areas.

For higher plants, levels of endemism increase with increasing productivity, with increasing elevation (reflecting increased habitat heterogeneity and isolation in high-altitude areas), and with higher rainfall in low- and middle-latitude areas, although many exceptions to these patterns exist. In the Mediterranean-climate regions of the Cape and southwestern Australia, there is a negative relationship between local endemism and soil fertility. In the California Floristic Province, palaeoendemics are clustered in the wettest and driest areas, whereas neoendemics occur in transitional rainfall areas where rates of speciation are highest. Similar patterns exist for Afrotropical birds and Neotropical butterflies. For a wide range of marine taxa, endemism is more pronounced in exposed and variable nearshore environments than in the more stable distant-shore habitats.

D. Biotope
Geographically isolated areas and biotopes, such as certain islands, mountain peaks, ancient lakes, caves, thermal vents, hot springs, vernal pools, the abyssal zone, and chemically imbalanced substrata, support a disproportionately high number of stenontopic endemics.

Most studies have focused on endemism on islands, mountains, and unusual substrata. Generally, larger continental islands such as Madagascar, New Caledonia, and New Zealand support the greatest number and proportion of endemic taxa, especially of higher plants. Elevational range explains the incidence of plant endemism on the Canary Islands and bird endemism in Indian Ocean archipelagos, suggesting the importance of topographical diversity. Continental islands are typically rich in palaeoendemics, whereas some taxa have undergone extensive and unusual adaptive radiation on oceanic islands such as the Canaries and Hawaii.

Mountains are also often rich in endemics, in both tropical and temperate regions, but not in recently glaciated, high-latitude areas of the Northern Hemisphere. Many desert inselbergs (granitic outcrops) act as mesic refugia that support endemics, this is particularly pronounced for plants in middle Asia. As on islands, endemism on mountains results from both historical (e.g., isolation) and ecological (e.g., heterogeneity) factors.

The restriction of endemic plant species to nutritionally imbalanced substrata, especially when these occur in an islandlike configuration, is widespread in Mediterranean-climate and humid tropical regions. These sites provide both a strong selective force for the evolution of neoendemics and a refuge from competition for palaeoendemics. The restriction of animal taxa to unusual substrata has not been studied in any detail, but is likely to be a response to habitat effects on vegetation structure rather than nutritional peculiarities per se.

E. Biology
Very few studies have addressed the relationships between restricted range size and biological factors such as body size, growth form, life-history traits, population size, and genetic architecture. Of these studies, few have
considered multiple trait interactions or phylogenetic relatedness.

As a generalization, there is a positive, albeit weak relationship between a species’ range size and its local population abundance for a wide range of taxa (Fig. 3). However, not all endemics have low local abundances; indeed, many narrow plant endemics are extremely abundant locally. There are a number of hypotheses to explain the positive relationship between range size and local abundance. These are based principally on artefacts (e.g., sampling effects), resource use, metapopulation dynamics, and spatially independent rates of population growth. This pattern and its causes are currently attracting considerable attention.

There is a broad positive relationship between geographical range size and body size for animal species (Fig. 4). There is also a trend, both within regional floras and specific taxa, for low-stature growth forms to be overrepresented among plant endemics. This is especially true of the South African Mediterranean-climate region, where low shrubs (Fig. 5) dominate the endemic flora. Among Neotropical forest plants, endemics tend to be herbs, shrubs, or epiphytes rather than forest trees, whereas in the rain forests of Sri Lanka, endemics are overrepresented among long-lived, late-successional trees.

Gigantism is a common feature among some plant groups endemic to alpine habitats at low latitudes. Among animals, gigantism, dwarfism, and flightlessness are widespread among island endemics, as well as among some continental endemics associated with insular biotopes.

The reproductive correlates of endemism have been more extensively studied than other biological attributes. There are a number of pertinent generalizations, although exceptions exist for all of them. Range-
FIGURE 5  Percentage of endemic and nonendemic species in the Langeberg mountain flora (Cape Floristic Region, South Africa) in (a) seven growth form classes (G = geophyte, HG = graminoid, FO = forb, T = tree, LSH = low shrub, MSH = mid high shrub, TSH = tall shrub); (b) two postfire regeneration classes; and (c) four dispersal mode classes. Chi-square analyses were performed on untransformed data. (Reprinted from Biological Conservation 72, D. J. McDonald and R. M. Cowling. Towards a profile of an endemic mountain fynbos flora: Implications for conservation, pp. 1–12. Copyright 1995, with permission from Elsevier Science.)

restricted species differ from common ones in that they:

- tend to be self-compatible or rely on asexual reproduction;
- tend not to be wind-pollinated or have other inefficient forms of pollen transfer;
- invest less in reproduction;
- have poorer dispersal abilities;
- have shorter generation times.

The last two attributes are shown in Fig. 5, where short-distance ant dispersal and fire sensitivity (rapid generation time) are overrepresented among plant endemics in a mountain region of the Cape Floristic Region in South Africa. However, these and other reproductive traits, such as seed size, seed number, and reproductive investment, all interact in complex ways. Furthermore, these traits are not independently distributed among species.

Many studies indicate that plant and animal endemics have lower levels of genetic variation in comparison with widespread congeners. This may be due to several factors, including adaptations to narrow ecological conditions, small population size, and self-incompatibility in plants. However, there are also cases of little difference in genetic diversity between closely related endemic and widespread plant species.

F. Taxonomy and Phylogeny

Many biotas that are endemic to biogeographic regions are not a random phylogenetic assemblage. Some plant families are significantly overrepresented among the endemic floras in many parts of the globe, especially in Mediterranean-climate regions. The same is true of certain dragonfly families in southern Africa. Among plants, Cyperaceae and Poaceae are underrepresented as endemics in many floras throughout the world. In many cases, these patterns can be attributed to taxon-specific biological attributes that predispose a lineage to endemism. Thus, the existence of discernible phylogenetic correlates of endemism implies that range size may be an evolutionarily stable character of a lineage. Hence there is a need for the comparative approach to assess the role of phylogenetic relatedness in explaining patterns. However, it is important to establish the taxonomic level at which these relationships are manifested. For example, for several data sets, the majority of variation in range size is explained at the level of species within genera.
G. Endemism and Speciation

At face value, the relationship between range size and speciation appears to be quite simple: a reduction in range size will always accompany a speciation event, and a species nearing extinction—in an advanced stage of the taxon cycle—will occupy a limited range size. The deeper issue of the extent to which range size is a cause or consequence of speciation is a question of considerable interest.

There has been a long-standing and as yet unresolved debate regarding the causal relationship between range size and speciation. The arguments assume positive relationships between range size, population size, and dispersal ability. One viewpoint suggests that owing to extensive gene flow and reduced extinction rates, widespread taxa should have lower rates of speciation than range-restricted taxa. An alternative hypothesis is that owing to greater genetic variability and a higher frequency of founder effects, species that comprise large and well-dispersed populations that occupy large range sizes are prone to vicariant speciation.

Many studies of fossil and extant lineages suggest that turnover (speciation and extinction) is associated with relatively low local population abundance, poor dispersal, and narrow range size. Clearly, at extremely low values for these variables, extinction rates will overwhelm rates of speciation. Elevated speciation and extinction rates are also associated with increased specialization, reduced body size, and increased generation times; all of these are correlates of narrow endemism. Thus, endemism and its correlates are responsible not only for enhanced rates of speciation, but also rapid rates of extinction. In E. S. Vrba's parlance, these processes are flip sides of the same coin.

The alternative view, that speciation is associated with large, centrally located and widely-ranging populations, and that peripheral isolates are relictual taxa, also has support. Ultimately, aspects of this debate will be resolved by studies that assess range size and its correlates in a phylogenetic context.

VI. CONSERVATION

Because of their restricted geographical range size, high habitat specificity, and generally low population abundance, endemics are more vulnerable to extinction than are widespread and common species, as a result of both deterministic (habitat transformation) and stochastic (small population effects) factors. Therefore, considerable attention has been given to the conservation of local endemics. Attempts have been made to use the correlates of local endemism to devise management plans that will reduce anthropogenic extinctions.

Recent advances in systematic conservation planning have identified priorities for conservation on the basis of complementarity of biotas (representation), but also for the retention of biodiversity in the face of threatening processes. This approach involves the assessment of the irreplaceability of an area—a measure of the likelihood that the area will be needed to achieve a conservation goal—and its vulnerability to biodiversity loss as a result of current or impending threatening processes. Endemic-rich areas inevitably emerge as priorities since they combine high irreplaceability, owing to their unique biota, and high vulnerability, since endemics are prone to extinction. However, some endemics, particularly plants, may be preadapted to persist in small populations and could be effectively preserved in small, fragmented areas.

VII. CONCLUSIONS

There are few generalizations regarding geographical patterns and correlations of endemism. This is understandable, given that definitions of endemism are mostly study-specific, and that endemism is partly a consequence of regional-specific historical events acting on phylogenetically distinct biotas. Furthermore, species with similar range size often have different local abundances that are likely to be manifested in very different biological attributes. Finally, within-region analyses invariably lump together palaeoendemics and neoenemics, groups with different origins and phylogenetic relationships, and often, different biologies. The recent trend to correct for phylogenetic relatedness holds much promise for understanding the ecological and evolutionary correlates of endemism.

The most active fields of research currently are studies on the correlates of range size, particularly local population abundance, body size, and reproductive traits; the role of endemism in reserve selection, especially as a measure of irreplaceability and surrogate measure of vulnerability; and historical reconstructions using congruent areas of endemism in phylogenetic studies. Much less classical biogeographic research is being carried out on the identification of centers of endemism, despite the fact that reliable distribution data are lacking for many areas and taxonomic groups. This lack of data has serious consequences for the iden-
tification of endemic-rich areas for conservation purposes.

See Also the Following Articles

Biodiversity-Rich Countries • Biogeography, Overview • Diversity, Community/Regional Level • Extinction, Causes of • Island Biogeography

Bibliography


ENERGY FLOW
AND ECOSYSTEMS

Alan P. Covich
Colorado State University

I. Ecosystem Boundaries: Inputs, Outputs, and Transformations of Energy
II. Multiple Energy Pathways
III. Internal and External Nutrient Cycling
IV. Biodiversity Effects on Energy Flow
V. Why Species Matter
VI. Future Studies

GLOSSARY

chemoautotrophs These microbes use inorganic compounds as a source of carbon and energy and function as primary producers.

decomposition The biotic breakdown of dead organic matter (detritus) by bacteria and fungi that releases carbon dioxide and nutrients for recycling.

ecosystems Composed of species assemblages (producers and consumers) that interact with each other and their associated abiotic environment within well-defined natural or conceptual boundaries.

food chains Composed of species that are connected by the flow of energy and material from producers to consumers.

food webs “Flow maps” that depict connections among multiple food chains.

functional groups Aggregations of species that perform similar ecosystem processes, such as grazers, suspension or filter feeders, leaf shredders, predators, and decomposers.

photoautotrophs These, such as green plants and some bacteria, use solar energy and inorganic compounds to synthesize organic matter as primary producers.

primary productivity The rate of synthesis of organic matter by plants (biomass per unit area of habitat per unit time or, in some cases, biomass per unit volume per unit time).

secondary productivity The rate of assimilation and growth by animals (biomass per unit area of habitat per unit time or, in some cases, biomass per unit volume per unit time).

trophic cascades These occur when changes in the presence or absence (or shifts in abundance) of a top predator alter the production at several lower trophic levels; primary and secondary production at lower levels are alternately constrained or unconstrained by the feeding activities of consumers at upper levels.

trophic levels Groups of individuals classified as primary producers or primary or secondary consumers within food webs; individuals feeding both as primary and secondary consumers are omnivores. A single species may be represented on more than one trophic level.

ECOSYSTEMS ARE THERMODYNAMICALLY OPEN, hierarchically organized communities of producers, consumers, and decomposers together with the abiotic...
factors that influence species growth, reproduction, and dispersal. These abiotic factors include the flow of energy and the circulation of materials together with the geological, hydrological, and atmospheric forces that influence habitat quality, species distributions, and species abundances. Energy flows through many species, and the way in which this flow affects the persistence of ecosystems is influenced by land-use changes, precipitation, soil erosion, and other physical constraints such as geomorphology.

Energy flow through ecosystems is essential for nutrients to cycle through food webs. These food webs are often subwebs of more complex species assemblages and may be only partial descriptions of more complex hierarchies of energy flows. The hierarchy of species' interactions in natural food webs typically results in some important feedback loops and recycling of nutrients and materials within the conceptually defined boundaries of an ecosystem. Many species of producers and consumers are usually interconnected and some may be interdependent. Food webs are diagrams that can function as "flow maps" to document which species interact with other species, either directly or indirectly, as energy flows through the community and determines the movement of nutrients and other materials.

Different species have important functional values, such as for organic matter production (plant and animal growth) or organic matter breakdown (decomposition), oxygen production, nitrogen fixation, and nutrient cycling. These species have intrinsic values as the unique end products of evolution, and native species are likely to have adapted specific ways to respond to local or regional environmental disturbances. Conceptually, the loss of even a single native species, or the introduction of a nonnative species, could alter how the other remaining native species continue to perform different ecosystem functions. Disruptions of ecosystem processes are known to have occurred after certain well-adapted, native species were lost through local extinction following intense (pulse) or prolonged (press) disturbances. However, predicting which species are essential to ecosystem functions has generally remained difficult because information is lacking on many species interactions as well as on life history, adaptations to different disturbances, and dispersal abilities among key species. Ecosystem studies can provide a broad perspective regarding species relationships and recycling of essential nutrients. Species' shifts in patterns of abundance (or local extinctions) following natural and anthropogenic disturbances illustrate how some key species regulate nutrient cycling and other ecosystem functions. Field testing of many concepts related to understanding the importance of key species in ecosystem functioning is just beginning. Results of these long-term ecosystem studies can provide guidelines for the stewardship of biodiversity.

I. ECOSYSTEM BOUNDARIES: INPUTS, OUTPUTS, AND TRANSFORMATIONS OF ENERGY

An ecosystem approach can be used to address many different questions spanning scales from the global biosphere to small ponds or patches of habitat. As the questions change, so do the boundaries and the complexities of species interactions within and among different compartments or across trophic levels. Ecosystem boundaries are often defined to include natural species assemblages and to analyze inputs and outputs of energy and materials for cross-site comparisons of efficiencies in energy transfers and studies of changing conditions. These analyses often take the form of mathematical models such as computer simulations or individual-based models of species and their specific functions within the biotic assemblage and environmental conditions under study. One ecosystem may export nutrients and organic matter (stored energy) to other ecosystems so that cross-site linkages often become important. For example, in studies of nutrient cycling in terrestrial ecosystems, the definition of boundaries would likely have some compartments of organic matter production by living plants and their relationships with herbivores and carnivores. Macro- and microminutrient inputs would likely be derived from the atmosphere through dry deposition of particulates—nitrogen gas being taken up by some nitrogen-fixing species of microbes. Other sources of nutrients, especially phosphorus, would come from weathering and erosion of soil and bedrock deposits, with movement among other compartments derived from actions of wind and water. This combination of interactions illustrates the importance of defining clear boundaries for subsystems within the complete ecosystem so that measurements of movements (fluxes) among compartments can be measured accurately. Generally, ecosystems and their boundaries are abstractions that can only be useful and insightful when combined with sufficient knowledge regarding the natural history and general ecology of communities and their physical environment. There is wide recognition that a combina-
tion of direct field observation, experimentation, and modeling is essential when conducting ecosystem studies. Currently, fundamental questions dealing with relationships between energy flow and the species-specific roles of organisms are attracting increased attention. For example, can results of controlled small-scaled (fine-grained) experimental studies of productivity be used to predict responses of other natural assemblages at larger scales (coarse-grained) of ecosystem dynamics? Does energy flow through an ecosystem increase, decrease, or remain the same if one species goes locally extinct but the abundances of other “similar species” change rapidly to compensate for the lost species? Under what environmental conditions do species substitute for one another and compensate functionally for the lost species? These and many other questions are beginning to be answered, but studies related to biodiversity and the persistence of species assemblages remain incomplete. A series of symposia during the 1990s dealt with the relationships between biodiversity and ecosystem functions (Fig. 1) and stimulated many new ideas (Schulze and Mooney, 1994; Orians et al., 1996; Lawton, 1997; Palmer et al., 1997; Naeem, 1998; Wall, 1999).

A. A Historical Perspective: The Ecosystem Concept

How energy moves from one group of species to another has been an active area of study at least since Charles Darwin and Alfred Russel Wallace first wrote in the 1850s about the interconnections among species. They were intrigued with the general proportions of population abundances that were thought to exist among different groups of large predators and their prey, and they emphasized competition and predation as important factors for regulating species interactions. These early observers lacked a conceptual approach to what was later viewed as ecosystem-level dynamics. Darwin’s studies of earthworms and their roles in soil development and his views on the roles of multiple species in the “tangled bank” metaphor stimulated others to consider how these many interactions could be viewed holistically. In 1887, Stephen Alfred Forbes described material cycles within lakes and used a table of predator-prey data to examine which fish species consumed similar or different prey species in Illinois rivers and lakes. Forbes drew his conceptual boundaries and predation matrix to coincide with the shoreline and emphasized putting the many pieces together in a type of system homeostasis in his new metaphor of the lake as a microcosm. This holistic view was taken up in a different way by E. A. Birge in 1915 with his work on heat budgets of lakes in Wisconsin. Birge measured inflows and outflows of energy, and in so doing he set the stage for viewing ecosystems in terms of their general physical attributes rather than numerous component parts. This “black box” approach allowed for observations at larger scales without full analysis of the controlling variables within the process under study. The British ecologist, Charles Elton, recognized that food chains existed in the context of energy flow and nitrogen cycling from his early work with V. S. Summerhayes on Spitzbergen and Bear Islands. The sum of these chains formed a food web which Elton called “food cycles” and included in a diagram combining nutrient cycling and energy flows over landscapes scales. Elton built on earlier concepts of Victor Shelford and introduced the idea of a “pyramid of numbers” and a “pyramid of energy” that reflected general patterns of community organization. More individual plants and plant material (biomass) can usually be harvested in a plot than herbivores or the relatively rare carnivores.

![Figure 1](image-url)
showing a rough resemblance to the growth curve of rate of production may be expressed as a sigmoid curve. This natural aging process was thought to reach a climax stage and undergo senescence before developing into a bog forest. Some of his thinking on patterns of productivity change through succession.

Lindeman emphasized that groups of species both "co-act" and "re-act" in their relationships with each other and with their abiotic environment. He further emphasized that these patterns of productivity change as the biotic and abiotic community develop over time through succession.

Based on his own detailed studies at Cedar Bog Lake, Minnesota (and other studies by Chauncey Juday on Lake Mendota, Wisconsin, who was actively studying lake productivity with E. A. Birge), Lindeman proposed that lake productivity increased during succession as nutrients accumulated in the basin during eutrophication. This natural aging process was thought to reach a climax stage and undergo senescence before developing into a bog forest. Some of his thinking on patterns of change in productivity over time was influenced by earlier studies by G. E. Hutchinson and Anne Wollack on Linsley Pond. Lindeman (1942) noted that their work suggested that "these generalized changes in the rate of production may be expressed as a sigmoid curve showing a rough resemblance to the growth curve of an organism or of a homogenous population" (p. 409). Such analogies between ontogeny of individuals, populations, communities, and ecosystems were commonly used during that time period. However, as more accurate data on rates of sedimentation became available and isotopic dating improved, the generality did not hold up and multiple patterns of lake eutrophication were later documented.

The concept of trophic dynamics and its focus on transfers of energy between trophic levels was not widely accepted until a new postwar influx of investigators began detailed studies of energetics. Several questions have persisted for decades: How can energy flow regulate the number of trophic levels within an ecosystem? How does one group of consumers regulate the number of individuals and energy flow in other trophic levels? These questions and many others were rapidly taken up by ecologists such as G. Evelyn Hutchinson, Eugene Odum, and Howard Odum and their students in the 1940s and into the 1970s.

After considerable debate and continued developments of the concept, many ecologists today rely heavily on models and field experiments using modifications of Lindeman's approach (Harlan and Harstan, 1997). Some still have concerns about fundamental issues regarding how trophic levels are defined relative to the complexities of natural food webs (DeRuiter et al., 1996). As a result, there are various definitions of what constitutes a trophic level. Because most descriptive field-based studies of food webs really study subwebs and therefore are incomplete, a thorough test of predictions regarding food chain length or long-term stability relationships derived from trophic models is usually not feasible.

Conceptually, the predictions of how energy flow regulates trophic dynamics are relatively straightforward. First, some of the initial energy entering the first trophic level is lost by reflectance from the plants, lost as heat, expended in metabolism and evaportranspiration, or lost because of a less than complete coverage of foliage (leaf area) or algal volume. Thus, some warming occurs by energy absorption by the physical habitat (e.g., soil, rock, or water). Then, from constraints imposed by the second law of thermodynamics, energy is lost at each step in the flow of energy from the first trophic level to successively higher levels within food webs. The use of solar energy or chemical energy by primary producers and the consumption of plants and animals at higher trophic levels are relatively inefficient because some energy is lost to metabolism and as heat at each transfer across trophic levels. The total energy flow through the plant trophic level is termed gross...
Empirical studies demonstrate that most food webs contain fewer than four trophic levels (DeRuitter et al., 1996). However, the number of trophic levels is not a consistent measure because of the complexity of feeding relationships over time and space, the mobility of consumer species, and the movement of food resources across ecosystem boundaries. Many species vary in how they obtain their energy and how efficient they are at different stages of their life histories and under different conditions. Among consumer species, many rapidly growing juveniles or reproductive adults require high-quality, nutrient-rich foods. These same individuals typically feed on lower protein foods when they become nonreproductive adults. Numerous species are omnivores and feed on plants and animals from different trophic levels. Because of these complexities, there has not been complete agreement on how to operationally define trophic levels.

The transformation of inorganic elements into organic matter requires energy to be converted into biomass by species of algae, green plants, and a few types of bacteria. These micro- and macroautotrophs are often represented by many species. The degree of similarity (niche overlap) in their abilities to produce and to store organic matter is important in predicting the consequences of any losses of species. Many species have evolved into persistent assemblages that store carbon and nutrients such as nitrogen and phosphorus. Energy stored in the form of plant-produced organic matter is later passed on directly to grazing species and then indirectly to predators within food webs.

Efficiency of energy transfer from one trophic level to the next higher level is of fundamental importance in understanding conceptually how different ecosystems function. Measures of efficiency, however, are only a part of the explanation for why some ecosystems have longer food chain lengths than others. Relatively "inefficient" food webs with few trophic levels appear to be adapted to certain types of frequent disturbances. Ecologists realize that a single explanation or mechanism is unlikely to account for all the various complexities that exist in determining how ecosystems are organized in terms of energy flow. However, comparisons among well-studied ecosystems and their numbers of trophic levels (food chain length) can provide a useful basis for predicting vulnerability of food webs to major disturbances (Orians et al., 1996) and movement of toxins such as mercury and other heavy metals.

Analysis of guilds and that of functional groups are different approaches used to study shifts in feeding behavior and to complement trophic-level analysis. Guilds are defined as assemblages of species or individual age classes that share a common source of energy at any given time (e.g., nectar-feeding birds and bees and insect-feeding birds and spiders). Functional groups are defined as being similar in their mode of feeding (e.g., filter feeders, shredders, and pursuit predators), but individuals may use a variety of different resources as resource availability shifts. Thus, individuals of a single species may be distributed over several trophic levels and belong to different guilds and functional groups during each individual's life span and reproductive period. As nutritional requirements change and the availabilities of different types of food resources also change, individuals can often adapt to find different available sources of energy. Analyses of similarities in these adaptations and the degree of overlapping functionality are used to understand the degree to which producer and consumer species are interdependent. Numerous complex linkages (e.g., herbivory, predation, decomposition, parasitism, and mutualism) imply that few species are likely to be complete substitutes for other species (Vest et al., 1995; Naeem, 1998; Covich et al., 1999; Crowl et al., 2000). As discussed later, some species may interact positively, negatively, or neutrally in association with other species. Such complex relationships among species, especially under changing environmental conditions, complicate field experiments and make predictive models difficult to test fully.
D. Controls of Energy Flow in Food Webs

A small increase in species richness can have a large effect on how energy flows through food webs. The main features are the number of linkages among species and, especially, the type and strength of those linkages (Paine, 1969). For example, a simple trophic structure would be a linear series of three species (A–C) in a food chain with one species in each trophic level (Fig. 2). Thus, a single species of plant is consumed by a single species of herbivore, which is consumed by a single predator species. Although analysis is relatively definitive in these types of communities with few species, this simple food chain structure may preclude consideration of some questions of general concern, such as resiliency of the assemblage following a disturbance and species loss.

If one more plant species is added (D) to a simple community, then this slightly more species-rich food web provides some important additional dynamics in terms of alternative pathways for energy to flow. Moreover, the herbivore (B) can switch from one food resource to another and this additional complexity increases generality and realism incrementally. With two herbivore species (B and C) the food web is more complex and the predator (A) has a choice of prey resources. Even with the same number of species, much more realism is added by considering the predator (A) to be an omnivore, and even more is added if the predator and one of the herbivores (C) are also cannibalistic. These simple diagrams show how quickly the types and numbers of linkages (connectance) within food webs can change the dynamics of energy flow even with only a few species. This last example is typical of some low-diversity stream food webs on isolated tropical islands currently under study (Covich et al., 1999; Crowl et al., 2000) and discussed later.

E. Ecosystem Analysis

General rules regarding the relationships between energy flow and the control of food web complexity are currently incomplete. In some habitats a complex relationship apparently does exist among the total annual amount (and seasonal distribution) of energy, the nutrient inputs to ecosystems, and the number of different species in a habitat. In other habitats there is no evidence for a cause-and-effect relationship among the rate of energy flow, species growth and productivity, and the number of species in an ecosystem. Other likely variables include evolutionary time and biogeographical distributions as well as the frequency and intensity of disturbances. The particular species composition of a food web may also alter productivity. Empirical evidence for predicting the importance of species-specific relationships is increasing, but methods for establishing which species regulate ecosystem functions remain controversial. Currently, only a few studies have focused on the species-specific roles to determine which species have unique roles and how these roles shift as environmental conditions change.

Although recent experiments have examined some fundamental relationships, we do not have a full understanding of the effects of varied energy inputs on the
richness of consumer species that naturally coexist in ecosystems. For example, in most tropical forests and coral reefs there is generally a high number of species, but the cause and effect of this species diversity are open to different interpretations other than the potential importance of relatively high and continuous inputs of energy (Waide et al., 1999).

At the earth’s surface the annual, seasonal, elevational, and latitudinal distribution of solar energy provides varied inputs of energy to deserts, grasslands, forests, wetlands, lakes, rivers, and oceans. Generally, energy always flows through ecosystems but does so at different rates under different global geographical locations and local conditions of slope and aspect. Depending on latitude, ecosystems generally receive a seasonally pulsed or a continuous annual supply of solar energy for primary producers. The rate of energy flow and associated biological productivity are dependent not only on the availability of energy but also on water, on combinations of different macronutrients (nitrogen and phosphorus) and micronutrients (trace elements such as iron, manganese, and silica), and on the presence of an assemblage of interactive plant and animal species. Some natural food webs in extreme environments, such as hot springs, saline lakes, caves, or certain deep-sea thermal vents, have relatively simple linear food chains and have species adapted for specific habitats. For example, dark, closed caves that are deep underground only receive indirect sources of detrital energy from sunlit surfaces aboveground and are typically characterized by a relatively low number of endemic species not found on the surface. Deep-sea thermal vents in the oceans rely solely on chemical energy derived from microbial breakdown of gases such as hydrogen sulfide and are characterized by sulfur bacteria and unique consumer species. These simple ecosystems continue to provide an opportunity to test some fundamental concepts regarding food webs and energy flow relationships.

II. MULTIPLE ENERGY PATHWAYS

It has been evident since Lindeman’s work that energy travels along different pathways and includes microbial species and macrospecies in various interconnected relationships. Experimental assemblages are now being widely used to provide some insights into which mechanisms control ecosystem dynamics. There is evidence for biotic control mechanisms (interspecific competition for resources, predation, parasitism, and mutualism), abiotic controls (nutrient limitation and frequent and/or intense disturbances), and combinations of controls in different ecosystems. These studies have also provided important insights regarding two main energy pathways. The distinction between direct, solar-driven, photosynthetically based food webs and indirect transfers of stored energy in the form of detritus (that can be wind driven or washed into habitats) has sorted energy flows into two main classes. The earliest work on ecosystems recognized this bimodal classification and it remains an important organizing framework in linking aspects of terrestrial and aquatic ecology (Polis et al., 1997; Covitch et al., 1999; Wall, 1999).

The importance of organic detritus as a means for storage of energy was recognized by studies of Jerry S. Olson in the 1960s at Oak Ridge National Laboratory on the carbon cycle. Analysis of time lags requires an understanding of how rapidly organic detritus accumulates and then breaks down to recycle carbon, nitrogen, phosphorus, and other materials. These insights are critical in current discussions regarding carbon dioxide accumulation in the atmosphere as fossil fuels are burned (i.e., coal, oil, and natural gas taken from storage that accumulated over geological timescales and are now being rapidly cycled back into the atmosphere following combustion). Debates regarding global warming, the greenhouse effect, and where the “missing” carbon is in the present-day ecosystem require a thorough understanding of the entire biosphere and the carbon cycle as it relates to other nutrient cycles.

Recent studies of deep-sea vents and hot mineral springs define a third distinct class of ecosystems that is not solar driven but depends on chemical energy sources used by chemosynthetic microbes. Geologic sources of hydrogen sulfide and other gases provide examples of chemical energy pathways that may well have been the first modes of ecosystem formation by the earliest microbial species on Earth before the evolution of photosynthetic species. Various lines of evidence, such as the banded iron formations in pre-Cambrian rock strata, indicate that the earliest atmosphere lacked oxygen, suggesting that chemosulfate microorganisms dominated the first phases of evolution. Once oxygen-producing photosynthetic organisms evolved and dominated the oceans and lakes (and later developed terrestrial forms of green plants), their high levels of primary productivity resulted in an accumulation of oxygen in the atmosphere and a decrease in carbon dioxide (possibly through carbon uptake and storage by plants and deposition of sedimentary limestones). This early shift into a phototrophically based ecosystem apparently put the chemosynthetic organisms at a competitive disadvantage in
A nutrient-rich lake ecosystem and the flow of energy and materials through major compartments. The inputs of dissolved nutrients and particulates (leaf litter and suspended organics) move through the lake ecosystem at a rate determined primarily by the amount of solar energy entering from the surface of the lake, the stored energy in the form of organic matter inputs (for bacteria, fungi, and detritivores), and throughflow of water. Productive lakes can be either sources or sinks of nutrients relative to downstream river and lake ecosystems, depending on their uptake and storage of energy and nutrients. Shallow lakes have a higher surface to volume ratio than deep lakes and are usually more productive because a larger proportion of their volume receives solar energy inputs. Winds mix nutrients from bottom waters and often circulate limited concentrations of dissolved phosphorus that flux from the deeper sediments. Energy transformations from solar input to green plants to herbivores and carnivores are essential for nutrients to move through the food web and be recycled (reproduced with permission from Covich et al., 1999).

A. Lake Ecosystems

To illustrate the flow of energy through ecosystems it is useful to consider some examples derived from lake studies. These convenient habitats have been used for comparative ecosystem studies because distinct boundaries provide clear definitions of inputs and outputs (Fig. 3). The main boundaries include any inflowing and outflowing rivers as well as the lake surface–atmosphere and the sediment–water interfaces and also shorelines and topographic ridges (that delimit drainage basins). Water temperatures, nutrient inflows and outflows, and mixing and transport processes all influence species distributions and abundances in generally predictable ways.

Solar energy transformed into organic matter through the process of photosynthesis is the main source of energy for most ecosystems, especially in large lakes. Different sizes and types of plants in lakes vary greatly in their rates of productivity. For example, in shallow-water ecosystems solar energy can be used by microphotoautotrophs (attached algae or suspended phytoplankton) and macrophotoautotrophs (pond weeds such as cattails and water lilies). The ratio of the biomass of organisms relative to their rate of production is termed “biomass turnover time.” Biomass is generally measured by multiplying the number of individuals in an oxygen-rich environment. These remnants of the earliest species of microbes now dominate only in deep-sea vents and hot mineral springs, in which they are still well adapted to compete.
a population by each individual's weight. Turnover time is related to how rapidly organisms increase in biomass during their life span. Rapid turnover is associated with high rates of productivity by small, fast-growing individuals of plants and animals. Small species with rapid turnover usually exploit resources at relatively fine spatial scales.

B. Importance of Depth and Spatial Heterogeneity

Different wavelengths of light energy penetrate into waters of different depths. Shade-tolerant species of plants living in deep water (or in deep shade of the canopy trees in terrestrial ecosystems) have distinct physiological adaptations that allow them to obtain sufficient energy to grow and reproduce even at low light intensities. The range of wavelengths that is used by plants is light visible to the human eye. This range of visible light (between approximately 400 and 700 nm) is termed photosynthetically active radiation. Short wavelength (ultraviolet light) and long wavelength (infrared light) are not used in photosynthesis but are important in regulating floating and emergent aquatic plants (or terrestrial plants in other ecosystems) because ultraviolet radiation degrades organic molecules and infrared increases leaf temperatures. Daytime warming from the sun is essential for some species to survive, especially insects and cold-blooded vertebrates (poikilotherms). Other transfers of solar energy are essential to various lake ecosystem processes (e.g., sediment and nutrient transport by wind and water). Work done by solar-generated winds mixes the lake and thereby influences nutrient cycling. Similarly, wind-driven currents disperse planktonic larvae and aereate deeper waters. Winds also increase salinity through evaporation.

Consumer species are limited by the availability of plant-derived organic materials (or microbially derived organic materials from chemosynthetic microorganisms). The well-mixed, brightly lit, open water (pelagic zone) is dominated by small species of zooplankton that are well adapted to feed on suspended phytoplankton. Microbial breakdown of dead organic matter (detritus or seston) and grazing on algae (by numerous invertebrate and vertebrate species) both function to recycle nutrients in the open waters. In other distinct habitats such as the shallow littoral zone near the shoreline, a vegetated zone is dominated by invertebrate species that feed on larger, rooted plants that have a slower turnover rate compared to the small, suspended algae. Some species of fish live in the littoral zone while young and move in and out of the pelagic open waters as they grow larger. As these predators increase in size, they switch from feeding on invertebrates and small fishes to consuming larger fish and crayfish prey.

III. INTERNAL AND EXTERNAL NUTRIENT CYCLING

Ecosystems are generally "open" with regard to external additions of nutrients from the atmosphere and from catchment basins. Deserts, grasslands, forests, and wetlands as well as most lakes and rivers continue to receive new supplies of nutrients from precipitation (rain and snow), from dust and eroded soils that are carried to the ecosystem by wind and water, and from human-derived fertilizers, sewage effluents, air pollutants, and other agricultural and industrial wastes. Decomposers recycle essential elements for continued productivity by other species. As a result of rapid decomposition, some ecosystems can be relatively "closed" in that nutrients may be rapidly recycled internally so that these nutrients remain within the ecosystem's boundaries and are not transported or lost to other ecosystems. Nutrients can recycle internally from temporary storage in living tissue (biomass), detritus, or storage in soils and sediments.

Different species of plants and consumers within the food web and specific environmental conditions (such as warm temperatures, low concentrations of nutrients, and high pH) influence rates of internal cycling. Land-use practices also have a large effect on vegetation and on how nutrients recycle within a catchment or move into streams and lakes. For example, by measuring the concentrations of nutrients in a stream draining forested catchments, a team of ecologists (led by Gene Likens and Herbert Bormann) identified different pathways for nutrients such as gaseous nitrogen and erosional phosphorus in Hubbard Brook, New Hampshire, from a series of long-term studies on the forest and its components (streams, lakes, and catchments). The research at Hubbard Brook demonstrated the importance of large-scale experiments (such as removal of forest vegetation) to determine how various components are related. The increased outputs of some major nutrients (nitrogen, phosphorus, and calcium) following the experimental manipulation of the forest cover provided information about the role of trees in taking up and storing water and nutrients. By tracking the movement of phosphorus out of the forest and into the stream, Judy Meyer concluded that pulses of stream flow were the primary...
ciliated protozoans. DOM is also a food resource for other microbial species, such as bacteria, fungi, and of microbes and is a source of energy and nutrients for ter (FPOM) is composed of small fragments of detritus (from leaves, fine roots, dead algae, and other small groups of specialized species that shred organic detritus internal nutrient cycling by using organic detritus as a source of energy. Their small size and fast turnover make them highly important. Bacteria and fungi rapidly break down dead organic matter before it accumulates within terrestrial soil and aquatic sediment (benthic) ecosystems. In many soil, stream, and estuarine ecosystems the amount of energy and nutrients cycled by microbes is relatively larger than that of the photosynthetic pathway. The soil and sediment biota have similar groups of specialized species that shred organic detritus (from leaves, fine roots, dead algae, and other small invertebrate consumers). Fine particulate organic matter (FPOM) is composed of small fragments of detritus and aggregates of dissolved organic matter (DOM) that form from breakdown products and from cell exudates and leachates in aquatic ecosystems (such as groundwaters, streams, lakes, and estuaries). The release of DOM and inorganic nutrients is accelerated by some species of microbes and is a source of energy and nutrients for other microbial species, such as bacteria, fungi, and ciliated protozoans. DOM is also a food resource for larger invertebrate consumers and provides a parallel set of pathways for energy flow. Invertebrate detritivores break down dead organic materials to obtain their energy. Many detritivores are dependent on microbes to condition the detritus before it can be consumed by invertebrates or vertebrates. Nutrient cycling allows for continued energy flow through the ecosystem. The continued input and flow of energy through the food web is likewise required for nutrient uptake and recycling. Numerous species of bacteria and fungi form a ‘microbial loop’ that provides for high rates of energy flow by breaking down organic matter. Excretion of wastes by consumers provides one source of nutrients for species in the microbial loop. The microbial loop is especially important in marine and freshwater pelagic (open-water) ecosystems. For example, a crater lake (e.g., Crater Lake, Oregon) with a limited surface area and positioned in a volcanic depression usually has a very small drainage basin compared to large lakes with extensive drainage runoff. A lake with a small surface area will also receive very little atmospheric inputs of nutrients. Primary production remains low and mostly dependent on rapid, internal recycling of nutrients from fine, suspended detritus (dead plankton) as a result of breakdown by microbial species. Nutrients may only slowly accumulate in the sediments of the deep crater lake and not be mixed by currents back into the upper layers of well-lit waters (photic zone). Whenever nutrients are low (as in midsummer in some more productive lakes following the spring bloom of algae) the role of microbes and invertebrate species in cycling nutrients is important to maintaining energy flow in the ecosystem. Similarly, nutrient concentrations are relatively low in the open ocean, far away from continental sources of nutrient runoff and from coastal upwellings of deeper, nutrient-rich waters. Microbes attach to dead plankton and use this organic detritus as a source of energy. In the process of breaking down this fine detritus, the microbes release nutrients for further growth by photoplankton (floating algae). Nutrients are also released in the photic zone by leakages from algal cells and by excretion of herbivores. The small size of the detritus particles slows their rate of sinking through the well-lit photic zone and allows the microbes to release nutrients where photoplanktonic photosynthesis is not light limited. Larger pieces of detritus fall more rapidly and accumulate in deeper, darker waters where short-term seasonal storage occurs. The nutrients in these deep, dark waters are not available for continued photosynthesis until this entire layer of nutrient-rich water is mixed vertically by wind- and gravity-driven currents.
back into the photic zone and taken out of temporary, seasonal storage. In very deep lakes, the wind energy is insufficient to mix the entire volume and nutrients accumulate over many years in these deep waters.

There are also some important horizontal linkages in that nutrients and organisms (especially differently sized fishes and wind-driven currents carrying drifting zooplankton) move from the shallow, nearshore waters (littoral zones) of lakes to the open waters (pelagic zones). In these surface waters, nutrients are also recycled by the grazing zooplankton (feeding on phytoplanktonic algae) and by consumption of zooplankton and phytoplankton by fishes. These nutrients are returned to the well-illuminated surface waters, where they are again available for continued algal growth.

This series of vertical and horizontal transformations of organic detritus and primary production of algae in the pelagic zones has some similar analogs in streams, in which the dynamics occur mostly horizontally, but to some extent vertically, along the network of stream channels. The combined vertical and horizontal currents in stream channels form a spiral in the downstream flow of water that carries nutrients and organisms various distances. Dissolved nutrients move in and out of solution as they are briefly taken out of solution by adsorption on sediments and by active uptake by microbes and attached algae (growing along the bottom of the channel), and are later consumed and released back into the water by cell leakage, excretion from grazers, and predatory fishes. This spiral pattern of nutrients being transported downstream while moving in and out of sediments and organisms and then taken up again by other organisms or adsorbed onto sediment particles farther downstream was first examined by ecologists at the Oak Ridge National Laboratory using isotopes to trace pathways and is now widely studied, for example, with stable isotopes of nitrogen (N\textsuperscript{15}). The distance that a particular nutrient element such as nitrogen moves downstream is termed the “spiral length.” If the nutrient is taken from solution and tightly held in different species’ biomass, then the spiral length is relatively short and the role of microbes and other organisms along the channel is relatively important compared to that for streams where biotic interactions are less significant and during other periods when the spiral length is longer.

Use of functional groups such as shredders and filter feeders (or scrapers, burrowers, and predators) allows for analysis of groups of different species in terms of particular attributes (e.g., how each species processes different sizes of detritus). In both vertically and horizontally structured interactions, the processing rate of organic detritus by microbes is accelerated by the presence of those invertebrate species that convert large organic fragments into finer fragments. If one species (shredders) uses larger sizes of suspended particles and breaks them down into smaller particles as a result of its feeding, then the downstream supply of FPOM is increased for use by other species (filter feeders). However, if the species that shreds detritus is lost, the filter-feeding species may not function effectively. Different species may form “processing chains” that require particular combinations of linked-species to complete certain ecosystem processes efficiently. These processing chains are one way that increased species diversity can increase efficiency of processing detrital resources, especially if detritus is available at low levels of abundance or if the detritus quality is low. Similar linkages are known to occur among burrowing organisms in soils and sediments where strong interactions are generally important in understanding how different ecosystems function (Wall, 1999).

IV. BIODIVERSITY EFFECTS ON ENERGY FLOW

Productivity and related ecosystem processes such as nutrient cycling and decomposition are generally known to be influenced by particular species. Species attributes, including length of life span, rates of dispersal, and tolerance of frequent and intense disturbances under various environmental conditions, have important effects on ecosystem processes. Key processes (such as nitrogen fixation) performed by only a limited number of species or particular modes of feeding characteristic of only a few species are recognized as important characteristics associated with biodiversity (Palmer et al., 1997; Covich et al., 1999). The importance of a single species in biogeochemical cycling was demonstrated by Peter Vitousek and his students in studies of the invasion of some habitats in Hawaii by the nitrogen-fixing plant Myrica faya and by determining the long-term consequences for nitrogen cycling and impacts on the ecosystem.

David Tilman and others conducted several experiments to test the hypothesis that primary productivity in prairie grasslands is related to the number of species of plants grown in the same plot. These experiments and others have the advantage of considerable replication and careful controls but were conducted at relatively small scales. Although the relationship between species richness and energy flow in ecosystems has
attracted considerable study, the scale of these experiments and the composition of the species assemblages studied have resulted in different interpretations of general relationships. Results of field tests suggest that the number of species per se is apparently not as important as the particular attributes of different species that relate to their efficiency of nutrient uptake and retention, as well as growth. The larger the number of species included in a study, the more likely some species will be included that are well adapted for the conditions and function effectively. Thus, although ecologists have conducted field tests to determine how different species alter ecosystem processes, there is no complete consensus on how energy flow by itself influences species richness or vice versa (Tilman et al., 1998; Waid et al., 1999). Part of this lack of consensus is a result of using different scales and methods in field studies of many different types of ecosystems (e.g., boreal and tropical forests, grasslands, deserts, deep and shallow lakes, and large and small rivers). Theoretical and conceptual developments are being actively developed and are stimulating additional field testing of these relationships between biodiversity and productivity.

A. Top-Down and Bottom-Up Regulation of Energy Flow

In the 1980s, some of the first experiments designed to identify the importance of different species and key attributes included the removal of predators. Because top predators are relatively few in numbers, their removal can be operationally feasible over some relatively large areas. Robert Paine removed predatory starfish from the intertidal zone and observed a shift in the relative and absolute abundances of some of the mussels prey species. His results demonstrated that competing species can be held in check by selective predation, especially if the predators consume more individuals of the more numerous prey species. Work by James Estes in the 1970s on the effects of declines in abundances of sea otters and increased abundances of sea urchins and other prey (as well as the shifts of algal regrowth in kelp beds) along the west coast of North America also demonstrated that predators had large effects on entire food webs. These “keystone species” were viewed as important regulators of energy flow in natural food webs in that they had a disproportionate effect despite their relatively small numbers or biomass.

Recognizing the importance of keystone species led some ecologists to use this idea for managing certain ecosystems. Researchers introduced the concept of bio-manipulation to study different combinations and abundances of algal species and grazing species in order to regulate nutrient cycling in aquatic ecosystems. In a series of lake studies by several ecologists (especially Joseph Shapiro, J. Hrbacek, John Brooks, and Stanley Dodson), they controlled the presence or absence of top fish predators. From these studies it became apparent that consumers in upper trophic levels could regulate populations of prey, and this regulation in turn had consequences for lower trophic levels as well as nutrient cycling. Manipulating consumer species as a means of removing algae or altering the uptake and storage of nutrients became an area of active study in order to manage lakes and improve water quality. These types of food web studies were also recognized as a means to monitor and to understand movements of toxic compounds such as DDT and mercury in lake ecosystems. Studies of introduced predators by Paine and Zaret further emphasized that nonnative predatory species disrupted food webs. However, some ecologists argued that native predators rarely altered prey populations or shifted species composition. Steve Carpenter and James Kitchell demonstrated that these species changes did occur in open-water food webs of temperate lakes. Certain fishes could selectively remove large-sized zooplankton grazers (as Brooks and Dodson had earlier demonstrated), resulting in increased biomass of phytoplankton and decreased levels of dissolved nutrients in a trophic cascade (Carpenter and Kitchell, 1993). In other lakes additions of certain predatory fish species had major effects on phytoplankton growth and nutrient cycling through the predators’ effects on grazers. Carpenter and Kitchell performed a series of whole-lake experiments to show that fish predation altered zooplankton species composition and size spectra. Shifts in sizes and types of zooplankton herbivores (and other invertebrate predatory species) in turn altered the phytoplankton community and its productivity. These trophic cascades have mostly been observed in aquatic ecosystems (Carpenter and Kitchell, 1993; Power et al., 1996). Recent studies, however, in a tropical forest demonstrate top-down effects through four trophic levels.

Many studies have demonstrated that bottom-up control by nutrients also influences the rate of energy flow in food webs. For many years, additions of nutrients were known to increase phytoplankton biomass and to alter species compositions of algae. Zooplankton abundances and species richness, in turn, often decline when the quality of their algal food resources is reduced. Thus, the distinct pathways of energy flow are altered by the amounts and proportions of nutrients available.
to algae and bacteria. High phosphorus concentrations lead to phytoplankton communities dominated by blue-green algae (cyanobacteria) because some species of blue-greens can fix nitrogen (incorporate it into their cells) from nitrogen gas in the atmosphere. Their capacity to use high amounts of both phosphorus and nitrogen allows them to outcompete species of green algae and diatoms that cannot use atmospheric sources of nitrogen. Blue-greens can also take up and store nutrients internally beyond their metabolic needs for nutrients and deprive other species of recycled nutrients. Many species of blue-greens have distinct features (slime sheaths, toxins, and large sizes of colonies) that make them relatively unpalatable to many grazers. Some species of blue-greens produce toxic substances. By avoiding grazers, blue-green algal species can grow rapidly in the upper waters of lakes and ponds, thereby shading out the competing species. Blue-green algae often form long, chain-like colonies of cells that are less readily filtered by certain species of large zooplankton. Thus, zooplankton production often declines in waters dominated by blue-greens, and with fewer large zooplankton available as prey the production by fish predators also declines.

In many ecosystems a combination of top-down and bottom-up control can be expected. The roles of specific species in determining how nutrients are used and how rapidly energy flows through different food webs are being studied. The generality of these complex relationships and the predictability of patterns of food web responses to additions of nutrients or the additions or deletions of predators are not completely established. There are some traits that appear to increase the likelihood of certain patterns, and these are under intensive study. For example, as plants and animals die and sink into deep, density-stratified lakes the breakdown of organic matter releases nutrients that can accumulate and dissolved oxygen can be depleted. Deoxygenation sets up chemical conditions in the sediments that release previously bound nutrients that are added to bottom waters. During mid- or late summer nutrients will be scarce in uppermost most brightly lit waters (photic zone) in temperature-density-stratified lakes because essential nutrients have become increasingly concentrated (during the spring and early summer growing season) in the lowest layers of water below the photic zone. With these accumulated nutrients stored seasonally in deep waters in which light is limited and nutrients cannot be taken up by plants, the roles of grazers and predators are less likely to control nutrient dynamics. Biomass manipulation of consumer species may be more effective in shallow, frequently wind-mixed lakes and ponds.

B. Stressful Environments as Testing Grounds: Ecosystem Services and Biodiversity

In some ecosystems the number of producer and consumer species is relatively low because they are adapted to severe environments. Similar situations can occur on isolated islands or in deep caves or hot springs. Only a few species can tolerate the physiological stresses associated with very high or low temperatures or highly variable or extreme salinities, acidities, or nutrient concentrations. Such sites are useful for conducting field experiments because the low number of species can be studied in detail and manipulated over relatively brief periods.

Because of their distance from mainland sources of colonizing species, headwater streams on tropical islands contain a few abundant species of freshwater decapods (shrimp and crabs) and only a few species of other detritivores. Recent studies in streams on Caribbean islands illustrate how species differ in their effectiveness as leaf shredders. Furthermore, these studies show that different species form processing chains and interact to transform suspended organic detritus into benthic biomass that is retained within headwater food webs rather than being washed downstream or accumulating in deep pools.

Whole-pool experiments (Covich et al., 1999; Crowl et al., 2000) in the Luquillo Experimental Forest on the island of Puerto Rico were used to study rates of leaf litter processing. Leaves of Cecropia schreberiana were placed into pools that were cleared of other detritivores. To start the experiments, either Xiphocaris elongata or Atya lampes were placed into the treatment pools in which other detritivorous shrimp had been removed. Controls were pools with Cecropia but with no shrimp. Downstream concentrations of suspended fine, medium, and coarse particulate organic matter, dissolved organic carbon, total dissolved nitrogen, nitrate, and sulfate were measured during 23 days. Results of the experiments illustrate how differently these two species function as detritivores. Both species of shrimp accelerated leaf breakdown rates relative to controls where only microbial breakdown occurred. Xiphocaris shredded Cecropia leaves much faster than Atya. Xiphocaris rapidly shredded the large, intact Cecropia leaves and converted them into fine suspended particulates. Xiphocaris increased the rate of downstream export of particulate organic matter and concentrations of total dissolved nitrogen and dissolved organic carbon relative to controls. These differences in
processing rates also affected downstream distributions of suspended particulate organics and nutrients. Atya increased rates of leaf breakdown less than did Xiphocaris and apparently filtered out fine organic particulates resulting in less downstream export. Atya are especially well adapted to shred conditioned leaves and to scrape microbes from leaf surfaces at low flows. Atya are also well adapted with highly modified cheliped fans to filter suspended organic particulates from the water when flow rates are higher than 20 cm/sec. Although both of these shrimp are detritivores, they are not complete substitutes for each other. Their co-location and relative abundances affect rates of detrital processing when they form detrital processing chains.

V. WHY SPECIES MATTER

Recent studies have demonstrated that some ecosystem processes apparently do change as the number of species increases (Tilman et al., 1998). One of the most cited examples is the relationship between increased numbers of species in mixtures of annual plants that are grown in experimental plots. The total area of the plot covered by plant growth (an indirect measure of primary productivity) increased as the number of species increased from 1 to 24. The mechanisms for this relationship are not clear, but progress is being made in interpreting the effects of species composition on these and similar replicated experiments dealing with primary productivity (Hector, 1998).

Linking species to different ecosystem processes (such as productivity and decomposition) highlights the importance of how organisms interact. In some cases, these interactions may facilitate how rapidly processes occur. For example, decomposition of organic matter prevents buildup of organic detritus, which could lead to increased or reduced growth of macrophytes depending on the amounts and types of litter. In other cases the role of different benthic species in breaking down litter can prevent deoxygenation of lakes and streams and maintain supplies of clean water (Covich, 1993).

VI. FUTURE STUDIES

Given the rapid and accelerated loss of species and the irreversibility of global extinction, it is imperative that experimental and conceptual studies do more to examine multiple levels of ecological organization, from populations and communities to ecosystems and landscapes, before more species are lost. Unfortunately, we lack sufficient information about how the loss of different species can disrupt natural ecosystem services. These losses represent a type of “warning light” that should draw more attention to analyzing the consequences of losing native species and introducing nonnative species that may disrupt ecosystem dynamics.

Improved techniques using stable isotopes, enhanced computer models, and many other developments have expanded the means of addressing ecosystem questions. However, only recently have ecologists evaluated the importance of species-specific attributes in ecosystem processes. Part of the reason for this slowness is that complete food webs are complex and dynamic. Furthermore, the specific functions of a single species are difficult to isolate from the functions of other species in most natural communities. Probably the most important factor has been the development of different scientific perspectives by population ecologists, community ecologists, and ecosystem ecologists. The recent emphasis on integration across disciplines has created a new perspective on the importance of different species’ roles in performing different ecosystem “services” or functions, such as nutrient cycling and productivity.

See Also the Following Articles

CARBON CYCLE • ECO SYSTEM, CONCEPT OF • ENERGY USE, HUMAN • FOOD WEBS • LAKE AND POND ECOSYSTEMS • TROPHIC LEVELS

Bibliography


I. Patterns and Scale of Human Energy Use
II. Implications of the Laws of Thermodynamics
III. Biodiversity Impacts of Industrial Energy
IV. Biodiversity Impacts of Traditional Energy
V. Future Energy Paths

GLOSSARY

energy The capacity to perform work. Potential energy is this capacity stored as position (e.g., in a gravitational or electromagnetic field) or as structure (e.g., chemical or nuclear bonds). Kinetic energy is this capacity as manifested by the motion of matter. The joule (J) is the common SI unit of energy, where 1 J equals the amount of energy required to increase by one Kelvin the temperature of one gram of water. Other units include kilocalories (kcal), kilowatt-hours (kWh), and British thermal units (BTU).

energy, industrial Forms of energy generally transformed in bulk at centralized facilities by means of complex technology. The major forms of industrial energy are oil, coal, natural gas, nuclear, and hydroelectric. In addition to hydroelectric, industrial energy also includes other technologically complex methods of harnessing renewable energy, including photovoltaics, electricity-generating wind turbines, and geothermal turbines.

energy, nonrenewable Forms of energy whose transformation consumes the energy source. The major forms include oil, coal, natural gas, and nuclear.

energy, renewable Forms of energy whose transformation does not consume the ultimate source of the energy, harnessing instead solar radiation, wind, the motion of water, or geologic heat. The major forms of renewable energy are solar, biomass, wind, hydropower, and geothermal. The forms of renewable energy that depend on complex technology are forms of industrial energy. The simpler renewable systems are forms of traditional energy.

energy, traditional Forms of energy generally dispersed in nature, renewable, utilized in small quantities by rural populations, and often not counted in government statistics. The principal forms of traditional energy are firewood, charcoal, crop residues, dung, and small wind and water mills.

energy efficiency A measure of the performance of an energy system. First law efficiency, the most commonly used measure, equals the ratio of desired energy output to the energy input. Second law efficiency equals the ratio of the heat or work usefully transferred by a system to the maximum possible heat or work usefully transferable by any system using the same energy input.

entropy A measure of disorder or randomness at the microscopic level. The entropy of a completely ordered system (e.g., a system at a temperature of absolute zero) is zero.

fossil fuels Forms of stored energy produced by the action of pressure and temperature on organic matter.
buried over geologic time. The major types of fossil fuels are oil, natural gas, and coal.

law of thermodynamics, first Physical principle that energy is neither created nor destroyed, only converted between different forms. Energy is therefore conserved. In thermodynamic terms, the change in energy of a system equals the difference of the heat absorbed by the system and the work performed by the system on its surroundings.

law of thermodynamics, second Physical principle that any system will tend to change toward a condition of increasing disorder and randomness. In thermodynamic terms, entropy must increase for spontaneous change to occur in an isolated system.

power The rate of energy transformation over time. The watt (W) is the common SI unit of power, where 1 W equals the power expended by the transformation of one joule in one second.

HUMAN ENERGY USE is the extraction, collection, harnessing, and conversion of energy into forms that available technologies can utilize. Our energy use directly alters patterns of biodiversity through changes in land use and through industrial pollution. Indirectly, human energy use is changing global biodiversity through the emission of greenhouse gases that cause global climate change and through other broad environmental effects of industrialization. Whereas the direct effects cause acute damage, the indirect effects generally induce chronic harm. Because human energy use is equivalent to the product of population, per capita economic production, and energy use per unit of economic production, each of these factors can exert an equivalent indirect impact on biodiversity. Several other chapters in the Encyclopedia of Biodiversity cover important topics closely related to human energy use. Consequently, this chapter focuses on issues most unique to human energy use. Related entries include Acid Rain and Depositions; Air Pollution; Economic Growth and the Environment; Greenhouse Effect; Pollution, Overview.

I. PATTERNS AND SCALE OF HUMAN ENERGY USE

We use energy both to meet our subsistence needs and to satisfy our wants. In a subsistence society, a farmer's wife will burn wood to cook the day's meals. In an industrial society, a couple will jump in the car on Saturday night to go to a movie. Yet the forms of energy involved in these activities—wood, gasoline, electricity—constitute just the means to desired end-uses—cooking, driving, operating a theater—that ultimately provide unique services—food, transportation, entertainment.

As used by humans, energy falls into two broad categories: industrial and traditional. Industrial energy includes those forms of energy generally transformed in bulk at centralized facilities by means of complex technology. In general, these forms fuel the technology developed in the two-and-a-half centuries that have passed since the Industrial Revolution. The major forms of industrial energy are oil, coal, natural gas, nuclear, and hydroelectric. Industrial energy also includes other technologically complex methods of harnessing solar radiation, wind, and heat, including photovoltaics, electricity-generating wind turbines, and geothermal turbines.

Traditional energy includes those forms generally dispersed in nature and utilized in small quantities by rural people. The principal forms are firewood, charcoal, crop residues, dung, and small wind and water mills. Humans most depended on these forms of energy in the early stages of the development of the species. Because traditional energy sources occur widely and because their transformation does not rely on complex technology, they constitute the most important sources today for rural people in the less industrialized parts of the world. In most cases, a rural household will harvest its own traditional energy sources for its own needs. Because no commercial transaction occurs in these situations, and because most governments do not regulate the use of traditional sources, official statistics do not closely track traditional energy use.

Traditional energy is one form of renewable energy, which includes those forms of energy whose transformation does not consume the ultimate source of the energy. Renewable energy harnesses solar radiation, wind, the motion of water, or geologic heat. The major forms of renewable energy are solar, biomass, wind, hydropower, and geothermal. Conversely, the nonrenewable energy systems consume the very source of the energy, most notably, oil, coal, natural gas, and nuclear fuel.

Besides traditional energy and industrial hydroelectric energy, renewables include a host of recently developed, sometimes technologically complex, methods of harnessing sunlight, wind, water, or heat. These other renewable energy forms include photovoltaics,
TABLE I

<table>
<thead>
<tr>
<th>Region</th>
<th>Natural gas</th>
<th>Oil</th>
<th>Coal</th>
<th>Nuclear</th>
<th>Hydroelectric</th>
<th>Traditional</th>
<th>Total of total percent of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>Asia and Oceania</td>
<td>1.6</td>
<td>0.3</td>
<td>1.3</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Europe</td>
<td>1.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Latin America</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>United States and Canada</td>
<td>1.3</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>2.8</td>
</tr>
<tr>
<td>Former Soviet Union</td>
<td>0.3</td>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>&lt;0.05</td>
<td>1.4</td>
</tr>
<tr>
<td>World</td>
<td>4.8</td>
<td>2.8</td>
<td>3.3</td>
<td>0.9</td>
<td>1.0</td>
<td>1.8</td>
<td>14.6</td>
</tr>
</tbody>
</table>

Percent of total 33% 19% 22% 6% 7% 12% 100%

The world uses renewable energy sources for only one-fifth of its energy use. The main renewables and their approximate rates of use are firewood and charcoal (0.7–1.1 TW), large hydroelectric (1 TW), agricultural crop residues (50 GW), biomass electric (23 GW), small hydroelectric (20 GW), wind electric (8 GW), geothermal (7 GW), urban waste (1 GW), biomass methane (1 GW), energy crops (500 MW), and photovoltaics (400 MW).

Figure 1 shows the tremendous increase in world energy use over time. In the 20th century alone, energy use has increased by a factor of 12. While total biomass use has remained constant, the world has witnessed an explosion in the use of fossil fuels. Figure 2 shows the share of the United States in world population, economic production, and industrial energy use in 1997. Although the United States hosts
FIGURE 2. Share of the United States in world population, economic production, and industrial energy use in 1995 (data from the World Bank and IEA 1997a, 1997b). (a) Population. (b) Gross national product (GNP) adjusted for purchasing power parity (PPP). (c) Industrial energy use. (This total of 11.8 TW counts only the energy output of hydroelectric generators. Counting the equivalent input if the electricity were produced by nonrenewables, the method used in Table I, would increase the total to 12.2 TW.)

only 5% of the world’s population, it generates 21% of the world’s economic production and uses 25% of the world’s energy. The average 1995 industrial energy use per person in the United States of 11,200 W cap1 greatly exceeded the world average of 2000 W cap1, as well as the industrial energy use in other industrial countries, such as the United Kingdom at 5400 W cap1. On average, each American uses 10 times the amount of energy as each person in the People’s Republic of China (3000 W cap1) and 30 times the amount of energy of each citizen of India (370 W cap1). Figures 3a and Figure 3b show the 10 countries with the highest and the 10 countries with the lowest industrial energy use per person.

One measure of energy efficiency is energy intensity, the amount of energy used per unit of economic production, generally per dollar of gross national product, adjusted for purchasing power parity. The 1995 industrial energy intensity of the United States, 0.62 W S$1, exceeded the world average of 0.35 W S$1. Figures 3c and Figures 3d show the 10 countries with the highest and 10 ten countries with the lowest industrial energy intensity.

Concerning energy end use, detailed data on a global scale are not gathered. In the United States, however, the Department of Energy does regularly survey energy end use. Americans use approximately 40% of total energy for industrial processes and agriculture. Approximately 33% of energy use goes to cooling, heating, lighting, and maintaining commercial and residential buildings. The remaining 23%, almost all from oil, goes to transportation. Passengers vehicles use half of all transportation energy. The high energy per unit volume and the flexibility of a liquid render petroleum products extremely convenient for powering vehicles.

Globally, a third of energy use goes to electricity generation, mainly from coal, hydroelectric, and nuclear. Power plants release two-thirds of that as waste heat (see the next section). The remaining third mainly goes into the end use of industrial processing with the balance going to cooling, heating, and lighting. Households generally use the traditional energy sources of firewood and charcoal for the end uses of cooking and heating. Generally, cooking a joule of food requires 2 J of firewood wood or 8 J of wood converted to charcoal. Consequently, rural people use 1 to 2 kg wood cap−1 d−1 for a rate of energy use of 250 to 500 W cap−1. Actually, a total of only 20 to 40 W cap−1 actually enters the cooked food and warmed people.
Open fires will diffuse the rest as waste heat (see the next section). In urban areas of nonindustrial countries, people often rely on charcoal for energy. Even though the conversion of wood to charcoal releases waste heat, the end product has higher energy per unit mass than firewood. This makes charcoal easier to store and transport than firewood. Urban people use 100 to 150 kg charcoal cap$^{-1}$ y$^{-1}$, requiring 800 to 1200 kg wood cap$^{-1}$ y$^{-1}$. The ultimate end-use energy requirement is 30 to 45 W cap$^{-1}$.

### II. IMPLICATIONS OF THE LAWS OF THERMODYNAMICS

The first law of thermodynamics states that energy is neither created nor destroyed, only converted between different forms. This is the principle of conservation of energy. The first law means whatever energy a process does not convert into useful forms must still go somewhere. The nonuseful energy does not just disappear. Humans use the environment as the sink for this waste energy.

The second law of thermodynamics states that any system will tend to change toward a condition of increasing disorder and randomness. This is the principle of increasing entropy. The second law means that no energy transformation can convert 100% of one energy form completely into a useful form. The process will always release amounts of energy wasted in forms that are unrecoverable due to the disorderliness or randomness of the waste energy forms. The fewer energy transformations that a system contains, the fewer chances for random second law energy losses.

For example, the objective of an automobile’s internal combustion engine is the conversion of chemical energy in the covalent bonds of hydrocarbons in gasoline to heat energy of an expanding fuel-air mixture in the piston, to kinetic energy of the drive shaft, to kinetic energy of the main axle. No matter how efficient the engine and automobile technology is, the conversion process will always waste energy as heat in the friction of engine parts, sound in the banging of vehicle compo-
neutrons, heat in the friction of tires on the road, kinetic energy of the wind displaced by the vehicle, and countless other unrecoverable losses.

As another example, the objective of a coal-fired electric power plant is the transformation of the chemical energy in the coal to heat energy in the boiler, to heat energy in steam, to kinetic energy of a turbine fan, to electromagnetic energy in the generator coil. Along the way, the conversion processes lose energy as the light and sound of the boiler fire, the vibration of turbine parts, the heat of power plant components, and, most significantly, the waste heat carried by the power plant cooling water.

Theoretically, the maximum efficiency across a heat gradient is the Carnot efficiency:

$$\eta = 1 - \frac{\text{Temperature of heat sink}}{\text{Temperature of transformation}}$$

with temperatures in Kelvin.

For a coal-fired power plant, materials limit boiler temperatures to 1000 to 1200 K. At an ambient environmental temperature of 293 K, the maximum efficiency will be 70 to 73%. Typically, coal plants only achieve 30 to 35%, releasing two-thirds of the total as waste heat.

Table II gives various formulations of the first and second laws of thermodynamics. The inevitability of entropy losses makes the colloquial interpretation of the second law “You can’t even break even.”

III. BIODIVERSITY IMPACTS OF INDUSTRIAL ENERGY

A. Oil

The major impacts of oil on biodiversity derive from a fuel and use cycle that ranges over vast areas of terrestrial and marine habitat. Exploration, drilling, crude oil transport, refining, and utilization in vehicles change land use and introduce industrial pollution to land and sea.

Petroleum, or oil, consists of a complex mixture of hydrocarbons formed over geologic time from organic matter compressed under anoxic conditions. The most important chemical constituents are alkanes such as octane and methane and aromatics such as benzene and toluene.

The majority of oil deposits derive from aquatic plants and bacteria deposited in inland seas and coastal basins during the Cretaceous Period 100 million years ago. In the early stages of formation, bacteria initiated the anoxic reduction of the organic matter. Over time, pressure and temperature replaced microbial activity as the main agent of transformation. Eventually, these forces drove off most of the water, oxygen, and nitrogen from the condensate, leaving carbon and hydrogen compounds. Dispersed between sediment granules, the oil eventually migrated to low pressure geologic traps at depths of 1 to 7 km. Today, oil fields occur at an average depth of 1.5 km. On average, the stoichiometric composition of crude oil is CH₁.₅, with a very small amount of sulfur.

Petroleum exploration entails geologic surveys over extensive areas often with low human populations and relatively undisturbed natural communities. Exploratory surveys generate vehicle traffic and temporary dwellings that bring localized disturbances, but the most serious impacts occur with seismic detection. This method involves controlled detonations along lines or at points so that seismometers can extrapolate the layout of subsurface formations. These activities destroy areas of vegetation, disturb certain animals, especially ground-nesting birds, and fragment habitat. If such activities disturb animal behavior during breeding times, the impact can last over many growth periods.

Edwin L. Drake drilled the world’s first commercial oil well in Titusville, Pennsylvania, in 1859. All oil wells require access roads, and high-volume wells require buildings and electric and water lines. This infrastructure destroys vegetation and takes land away from animal habitat. The more extensive an exploited oil field, the wider the habitat impacts extend. Infrastructure at the Prudhoe Bay field, opened for drilling in 1968, now extends over 1700 km² of Arctic tundra. This has noticeably displaced calving of Rangifer tarandus (caribou) from the field. Likewise, proposed exploitation of Area 1002 in the Arctic National Wildlife Refuge would disrupt the migration routes of the porcupine caribou herd to its calving grounds.
Drilling operations produce water and air pollution. Serious water pollution comes from the vast amount of used drilling muds, which are lubricating substances pumped down to the drilling bit to carry away rock cuttings, to keep the bit from overheating, and to protect the drilling shaft from surrounding rock. Drilling muds consist of water mixed with light molecular weight oils. Used muds contain bits of metal from drilling components and any trace metals mobilized out of the drilled rock.

All stages of the oil production system from drilling operations to end use spill oil into surface and ground waters. Globally, oil spills into surface waters total more than 3 million tons each year. Half of these spills come from oil production, 40% come from nonpoint urban runoff, and the remaining come from natural seeps. At the start, spills occur at well blowouts when equipment fails to contain naturally high fluid pressures in oil-bearing strata. Spills also occur along the significant lengths of pipeline from the wellhead to tank farms to supertanker ports to refineries to gas stations. Pipes, valves, and tanks leak from fatigue and from human error.

These and other problems roused concern when oil companies first proposed construction of the Trans-Alaska pipeline to carry crude oil 1300 km from the North Slope to the Gulf of Alaska. When it eventually started operations on June 20, 1977, the Trans-Alaska pipeline integrated a set of environmental protection features. To prevent thawing of permafrost areas, brackets elevate 700 km of pipeline to heights of 3 m. Heat pipes at the bracket legs dissipate heat generated by the friction of oil passing through the pipe.

The elevated sections serve as underpasses for caribou. Over buried sections in certain permafrost areas, construction engineers designed refrigerated overpasses for caribou. Bridges carry the pipeline over 800 streams. Zigzags along the pipeline translate longitudinal movement of pipes expanding under heat to lateral movement, reducing the risk of leakage. Oil companies revegetated areas denuded by construction activities.

An access road now open to the public parallels the entire length of the pipeline. This road has opened up a strip of habitat to human contact, possibly changing behaviors among caribou and other mammals.

On March 24, 1989, the supertanker Exxon Valdez ran aground on Bligh Reef in Prince William Sound in the Gulf of Alaska, ruptured, and poured out 41 million liters of crude oil, the largest oil spill ever in U.S. waters. The spill caused acute damage to birds, marine mammals, and intertidal communities. The spill also caused chronic damage to fish species and intertidal and subtidal communities. The progression of the spill demonstrates a pattern repeated in smaller spills that occur frequently in the world’s shipping lanes.

Oil floats on top of water. Gravity and wind will spread a floating slick out to a thickness of 0.3 to 10 μm. Patches 0.1 to 5 mm thick can cover just 10% of the total slick area yet contain 90% of the total slick volume. Some oil dissolves and emulsifies into the water column, forming emulsions containing 80% H2O. Oil generally will not sink to depths below 20 m. In the Exxon Valdez spill, recovery teams deployed an array of countermeasures that included booms, skimmers, sorbents, pumps, burning, and surfactants for chemical dispersion.

Exposure to sunlight initiates photolysis of hydrocarbons into lighter molecular weight compounds. Heterotrophic bacteria will also oxidize hydrocarbons to smaller compounds, CO2, and water. The lightest hydrocarbons, as well as aromatic compounds such as benzene, volatilize. Loss of the lighter fraction leaves the remaining slick more viscous over time. This thick oil forms tar balls and pancake-like forms. Oil from the Exxon Valdez eventually spread across hundreds of kilometers of beaches, penetrating deeply into cobbled stretches and mussel beds. Today, oil still persists beneath the surface layer of rocks in many areas.

Three years after the Exxon Valdez spill, photolysis degraded 70% of the original oil. Bacteria then eliminated photolysis products amounting to 50% of the original crude oil; the other 20% evaporated. Work crews recovered 14% of the spill. Thirteen percent of the original oil sank into subtidal sediments. Beaches absorbed 2%, leaving 1% still suspended in the water column.

The spill occurred in early spring, just before the young of many species emerged to rejuvenate marine animal populations. C. pallasi (Pacific herring) were spawning inshore. Millions of O. hermaphroditus (pink salmon) fry were soon to be washed from gravel spawning beds into the spring plankton bloom offshore. P. vitulina (harbor seal) and Enhydra lutris (sea otter) pups were testing the frigid waters. Seabirds were beginning to converge on breeding colonies in the gulf. Consequently, the oil devastated populations of birds, marine mammals, and fish.

Thousands of birds can die in even moderate spills, but the Exxon Valdez spill eventually killed more than a quarter of a million birds of more than 90 species, the greatest demonstrated mortality of birds from any oil spill. Workers physically recovered 36,000 carcasses. Of these, 8000 were Brachyramphus marmoratus (mar-
bled murrelets) and 150 were Haliaeetus leucocephalus (bald eagles). The spill killed individuals of two Fratercula spp. (puffins) and four Uria spp. (loons). For the following three years, fewer breeding Uria aalge (common murre) showed up at spring colonies.

Oil coats feathers, matting and waterlogging them. The water repellency, buoyancy, and insulating properties of plumage derive from a precise, orderly arrangement of feather barbules and barbicelles. Contact with oil disrupts these arrangements. Soaked birds can die of hypothermia and drowning. Those that survive risk chronic exposure to toxic organic compounds through ingestion, inhalation of fumes, or absorption. Moreover, eggs are highly sensitive to contact with oil.

Many of the aromatics in petroleum, including benzene, toluene, xylene, and phenols, are lethal to animals on contact and carcinogenic under chronic exposure. Moreover, polycyclic aromatic hydrocarbons bond to lipophilic sites, an affinity that magnifies these compounds up the food chain. Even when not deadly, sublethal disruption of physiology or behavior activities can reduce resistance to infection and cause generalized stress.

In the wake of the Exxon Valdez spill, 300 of the 2200 harbor seals in Prince William Sound died, as well as 3500 to 5000 out of 10,000 sea otters. Like birds, oil obliterates the insulating properties of marine mammal pelage, leaving them to die of hypothermia. In addition, oil can clog the nostrils of seals, causing them to suffocate. Whales, insulated not by hair, but by layers of oily blubber, resist the effects of oil, although a well oiled whale can clog its nostrils for hours. In Prince William Sound, harbor seals may experience chronic problems because oil is accumulating in their bile and fatty tissues. Since the spill, sea otters have continued to experience elevated mortality.

An unfortunate coincidence has resulted in the geographic juxtaposition of important commercial fisheries and high yield offshore oil fields on the continental shelves. Not only do oil spills invariably cause fish kills, but chronic effects also reduce fish fitness years after initial exposures.

Oil at the air-water interface acts as a physical barrier interfering with gas exchange. In fact, oil has been a traditional line of defense used for mosquito larvae control. Under a thick slick, fish larvae can suffocate. Fish eggs, which often float at the sea surface, and fish larvae, which are often distributed in the upper water column, both occur in the areas of highest oil concentration. Hydrocarbons and aromatics damage eggs on contact. Oil concentrations will quickly exceed the LC₅₀ of 1 to 10 ppm for fish larvae. The early life stages of intertidally spawning fish are especially susceptible.

Exposure to toxics from the Exxon Valdez spill has caused chronic problems in Clupea pallasi (Pacific herring), Oncorhynchus clarki (cutthroat trout), Oncorhynchus gorbuscha (pink salmon), and Salvelinus malma (Dolly Varden). Fish species have shown elevated egg, larvae, and adult mortality, larval deformities, and poor adult growth rates, even in situations of constant food supply. Fish tissues in some species contain elevated concentrations of toxics.

Over time, tidal action spreads and coats the shore of the intertidal zone in a band of oil. This oil ring smother intertidal invertebrates, crustaceans, mussels, barnacles, limpets, and algae. Oil coatings will asphyxiate filter feeders. Recovery crews use hot water washes to clean oil coated shores, an effective method, yet destructive to intertidal organisms.

Oil spills also damage phytoplankton and other marine plants. Oil absorbs photosynthetically active radiation, so direct coating hinders plant growth and increases plant tissue temperatures. Aromatics may disrupt the orderly arrangement of grana in chloroplasts. An increase in ruderals characterizes the changes in plant species diversity. Blue-green algae blooms will increase eutrophic conditions.

Oil refining focuses on the catalytic cracking of carbon-carbon bonds of long-chain alkanes for the production of lower molecular weight hydrocarbons. Refiners try to recover every possible useful organic compound, from the light products methane, benzene, toluene, and kerosene, to medium-weight products like gasoline and diesel fuel, to heavy tars and asphalt. These processes, as well as sulfur recovery, inevitably generate water pollution.

Most constituents of petroleum and refined oil products volatilize easily. Consequently, each step of the petroleum fuel cycle generates air pollution. Methane, ethane, benzene, toluene, and other compounds will evaporate from crude oil exposed to air. The major emissions from oil refineries include CH₄, CO, CO₂, H₂S, NOₓ, and SO₂.

This section has concentrated on the impacts from the core stages of the petroleum fuel cycle: exploration, extraction, transport, and refining. Nevertheless, manufacture of the infrastructure and materials needed for these end uses generates industrial pollution and requires land. Moreover, armed conflicts caused, in part, by efforts to control access to oil fields and refineries take human life and directly disrupt ecosystems.

The combustion of refined oil products for transpor-
B. Natural Gas

Natural gas is a mixture of light hydrocarbons that exists at a gaseous state at standard temperature and pressure. Methane (CH₄) is the main constituent, but the presence of higher molecular weight alkanes, including ethane, propane, and butane, changes the average stoichiometric composition for natural gas with the water vapor removed to 0.79 CH₃.62. Formed by the same processes that formed oil, natural gas is often found at the top of oil deposits. The most voluminous natural gas reservoirs occur in Cretaceous strata. The land use changes brought by the exploration and extraction of natural gas produce the same biodiversity impacts as described for oil.

In the nineteenth century, companies had not yet erected natural gas pipelines or processing facilities. Moreover, industry had not yet developed much technology for using natural gas. Because companies found natural gas uneconomical to exploit, they just burned it off to reduce the risk of fire and explosion. The entire history of natural gas production has flared the equivalent of 8 years worth of U.S. energy use. Today, U.S. companies generally flare only small amounts at refineries, but companies from other countries flare enough that the total amount flared amounts to 5% of global natural gas production.

Gas companies generally pump natural gas straight from the well to a processing plant, eliminating the need for storage facilities at the wellhead and thus reducing the potential for leakage. Gas companies generally divide natural gas into three fractions: natural gas liquids (NGL), liquefied petroleum gas (LPG), and liquefied natural gas (LNG). NGL consists of the higher molecular weight fraction of natural gas that often settles out by gravity. Processing of natural gas from oil wells produces liquefied petroleum gas (LPG). Finally, pressurization of natural gas produces liquefied natural gas (LNG), a product that is expensive because of the special containers required for transport.

The major end uses of natural gas, cooking and heating, burn the fuel directly with no further transformation. Electricity generation from natural gas uses a gas turbine, which directly uses the hot gas products of combustion to turn the turbine fan, eliminating the intermediate step of steam generation used in oil and coal-fired plants. Cogeneration plants increase the energy efficiency of gas turbine systems by utilizing the waste heat of gas turbines for space heating or industrial processes.

The extraction and combustion of natural gas pollute much less than the extraction and combustion of oil. Because it exists in the gaseous state for much of the fuel cycle, natural gas exploitation does not produce significant amounts of water pollution. However, methane itself is a greenhouse gas, and its combustion produces the main greenhouse gas, CO₂.

C. Coal

Coal consists of hard carbonaceous material formed by the compression and transformation of terrestrial plant matter rich in cellulose buried at the bottom of ancient freshwater swamps and bogs. The richest coal-bearing strata date from the Cretaceous period 100 to 200 million years ago and from the Permian period 250 million years ago. Similar to the process of petroleum formation, the deposited plant matter undergoes incomplete decay in anoxic conditions.

In geologic time, the pressure of overlying rock and the temperatures generated therein drive off oxygen and hydrogen, leaving thick seams of reduced carbonaceous rock containing much more organic than mineral matter. The average stoichiometric equation of coal is 0.75CH₄0.8, but elemental sulfur also contaminates most coal deposits. The four major types of coal, in order of decreasing carbon content and increasing sulfur, are anthracite, bituminous, subbituminous, and lignite. Bituminous coal is the most physically abundant type worldwide. Peat, the partially oxidized, moist, organic soil that forms in marshes and bogs, is the very early precursor to coal. In certain areas, people burn peat for heating, cooking, and light.

The coal fuel cycle extends from extraction at the mine to conversion at a power plant to distribution across the electric grid to end uses in lighting, heating, and all the uses of electricity.

Coal mines generally fall into three types: deep, open pit, and strip. Deep mines extend down to a depth of around 1 km. Open pit mines reach down to 300 m. Strip mining generally removes the upper 30 m of land surface. Coal mines consume land, not just for areas actually excavated and areas used to dump unwanted extracted rock, but also for the support infrastructure of buildings, roads, and rail lines.

Deep and open pit mines remove huge amounts of rock, termed overburden, lying over the coal. The land over deep mines will sink, a process termed subsidence, drastically changing the topography, hydrology, and
Miners dump the huge amounts of unwanted extracted rock, termed mine tailings, in abandoned parts of active mines or on the surface. Pyrite (FeS₂) usually comprises a significant fraction of the tailings. In addition to being poisonous to plant and animal life, sulfurous acid mobilizes other toxic substances. The leaching of acids, trace metals, dissolved solids, and toxic organics produces a brew known as acid mine drainage that can devastate surface waters. Selenium and cadmium often occur in high concentrations in tailings, so acid mine drainage can initiate the bioaccumulation and bioconcentration of these trace metals in the surviving sections of the food chain.

Surface mining consumes vast tracts of land. Heavy machinery remove the upper layer of a landscape to expose relatively shallow coal seams, completely destroying the mined area. Although coal companies generally fill back the overburden into the mined area and replant it, strip-mined land never recovers its original characteristics. Replanting even creates opportunities for ruderals to expand where perennial plant species may have dominated. Rodents and other animals that adapt readily to human disturbance also take advantage of reclaimed areas.

Coal mines often need to impound surface streams to satisfy the significant water needs of mine operations. These needs include water cannon drilling, transport by slurry, fugitive dust spraying, coal washing, and size sorting. Mines crush and screen coal for uniform sizing, then wash and dry the coal for open air storage. The fugitive emissions from these processes consist of particulates that coat any exposed surface, blocking photosynthetically active radiation from plants, contaminating food and water sources for animals, and acidiying affected soil. Leaching of toxic substances from coal storage piles can also add to the pollution of surface waters. Rail transport provides the most cost-effective means of moving the bulky commodity of coal. Fugitive emissions from unit trains increase the particulate load in rail corridors. To save money on rail transport, many utilities will site electric power plants next to the mine to wire out the electricity. In certain regions, this shifts the pollutant load from urban areas to less polluted rural areas.

Most coal worldwide goes to electricity generation. A conventional power plant burns coal in a boiler to boil water that circulates through a closed loop system of pipes. The steam from the boiler enters a steam turbine to turn huge fans that power an electric generator that converts kinetic energy to electric energy. As a principal of physics, the movement of a conductor across a magnetic field creates electric current in the conductor. In the coal-fired power plant electric generator, the conductor consists of stationary coils of wire surrounding a magnet on a shaft rotated by the turbine fan. Much of the steam that moves through the fan transfers its heat energy to the kinetic energy of the fan, causing the steam to condense back to water. A condenser will then allow heat to transfer from any steam that continues past the turbine to an external supply of cold water. The water in the internal loop from the condenser returns back to the boiler to enter the steam cycle again.

Coal combustion releases CO₂, CO₂, SO₂, NOₓ, particulates, fly ash, arsenic, cadmium, chromium, mercury, and selenium. Nearly 40% of anthropogenic CO₂ emissions come from burning coal, whereas coal burning produces 80% of human SO₂ emissions. Consequently, greenhouse gases and acid precipitation may constitute the agents of coal's most extensive environmental effects. The slag remaining from coal burned in the boiler contains high amounts of trace metals, especially cadmium and mercury. In addition, the sludge from flue gas desulfurization units, the pollution control devices known as scrubbers, contains trace metals and toxic organics. The disposal of this sludge presents problems for land use and water quality.

Internal steam turbine water is the working fluid circulating from the boiler to the turbine to the condenser and back to the boiler. Cooling water is the medium that draws heat from the internal steam turbine water. In most conventional coal-fired power plants, the internal steam turbine water remains separate from power plant cooling water. A typical condenser consists of copper coils, carrying cooling water, that pass through larger structures carrying the internal steam turbine water. As a physical principle, heat passes from the steam turbine water through the walls of the copper coils into the cooling water.

A 1 GW coal-fired power plant typically requires 4 million m³ per day for all operations, mostly for cooling. These water needs dictate the necessity to locate a plant next to a natural water body. Power plants mainly use fresh water because of the corrosive effects of salt water. Water withdrawals change the hydrology of a watershed, changing water levels, surface area of mudflats,
surface area of wetlands, and other important habitat characteristics that can strand hydrophilic plant species such as Salix spp. (willows) and harm fish and shorebird populations. Impingement on intake screens kills significant numbers of fish and other aquatic species. Organisms that get through the screens undergo entrainment through the condenser, causing even greater mortality. The stress that any surviving organisms undergo reduces their fitness considerably.

All power plants, including coal, oil, and nuclear, generate three-quarters of the waste heat dumped into U.S. surface waters and into the atmosphere above the United States. Once-through systems dump the waste heat directly into local waters. Cooling towers dump waste heat into the atmosphere, condensing steam from the air. Cooling ponds provide a buffer for releasing some of the heat from cooling water into the atmosphere, reducing the temperature of the cooling water before it enters surface waters.

Thermal discharges into freshwater and coastal zones cause a host of negative effects on aquatic species:

1. Direct lethality to fish and crustaceans at water temperatures >35°C.
2. Decrease in dissolved oxygen.
3. Increase in metabolic rates and nutrition needs for fish and changes in nutrition requirements for other taxa.
4. Displacement of diatoms by green and blue-green algae.
5. Inhibition of vertical migratory behavior by zooplankton.
6. Thermal plume blockage of migratory fish movement.
7. Avoidance of warm areas by migratory waterfowl.
8. Early emergence of aquatic insect adult life stages into inhospitable environmental conditions.
9. Copper contamination from condenser coils.

Long-range transmission of electricity occurs across high-voltage lines strung on metal towers up to 30 m tall. The 115 kV network in the United States stretches across 200,000 km and occupies 2 million ha. The clear cutting of corridors 30 to 60 m wide for transmission easements directly changes the vegetation and plant life in cut areas. Periodic clearing maintains and intensifies the original changes. The areas that remain favor ruderals and animal species that adapt readily to human disturbance, such as Odocoileus virginianus (white-tailed deer). Herbicides used for periodic clearing can harm insect and bird species. Transmission line corridors fragment habitat and increase the area of habitat susceptible to edge effects. The cleared areas can also block migrating land animals.

Short-range electricity transmission occurs across low-voltage lines strung on wood, metal, or concrete poles generally 5 m tall. Harvesting wood poles can produce all the potential biodiversity impacts of commercial logging, monospecic plantations, and milling. In many countries, utilities treat the wood with creosote to guard against the action of insects and weather. Creosote, a by-product of crude oil refining, contains significant amounts of toxic organics that can leach and contaminate surface waters.

The material and energy needs for building the massive infrastructure of the coal fuel cycle produce wide-ranging environmental effects. Because most coal goes to electricity generation, the end uses of coal produce the environmental effects associated with climate control, lighting, commercial production machinery, residential appliances, and other electric devices.

D. Nuclear Fission

Nuclear fission is the splitting of high molecular weight elements to release energy held among protons and neutrons in the nucleus of the atom. Uranium and plutonium are the elements that provide the most effective yield from fission at current levels of technology. A fission reaction produces energy in the form of light, heat, motion of the split pieces, and radiation. Radiation consists of kinetic energy of small molecules and atomic particles and electromagnetic energy of photons traveling at certain frequencies. When radiation passes through living tissue, the particles or photons impart their energy to atoms and molecules in the tissue, disrupting molecular and atomic structures.

The fission products themselves will continue to emit radiation until they reach a stable atomic state. Strontium-90 and Cesium-137 have half-lives of decades, but Plutonium-239 decays with a half-life of 25,000 y and Iodine-129 will halve in mass only after 17 million years.

Nuclear fission plants require highly processed uranium fuel. A 1 GW fission plant requires 150,000 Mt UO₂-containing ore to fabricate enough fuel for one year. Milling this removes 150 Mt UO₂. In order to concentrate Uranium-235, a conversion plant converts UO₂ to 188 Mt of UF₆ gas. Differential diffusion of the UF₆ separates 31 Mt UF₆ enriched in Uranium-235. A fuel fabrication plant then produces 30 Mt of UO₂ pellets.

The mining and milling of uranium ore creates most of the same environmental problems already described
for deep mining for coal and for coal processing. Conversion, enrichment, and fuel fabrication require fluo-
rescent gas, which is lethal on contact to animals, damages vegetation, and reacts to form toxic by-products.

According to the International Atomic Energy Agency (IAEA), at the end of 1998, 334 nuclear fusion
plants were operating in 33 countries around the world; 104 nuclear fission plants were operational in the
United States. These plants possessed a combined rated capacity of 349 GW and generated 2300 TWy of elec-
tricity, 16% of the world total. Plant operating experi-
ence reached 9000 plant years.

Nuclear plants generate electricity in a steam cycle very close to that employed in coal plants, except that
nuclear fission provides heat to the boiler. The higher operating temperatures require more cooling water than
a coal-fired plant of the same electric generation capac-
ity. A 3 GW nuclear fission plant requires 6 million m³
of cooling water each day, so the effects of water intake
and thermal discharge described in the previous coal
section are all more serious for nuclear plants.

Because nuclear plants involve combustion only in
construction and in support vehicles for operations,
they produce few air emissions. Nuclear plants do, how-
ever, produce long-lived radioactive wastes. Low-level
wastes include reactor containment water, worker

clothing, exposed tools, and plant fixtures irradiated
and in a steam explosion. The explosion destroyed the
instantaneous catastrophic increase of thermal power
RBMK graphite moderated reactor resulted in a virtually
principal effects of hydroelectric plants come from the
total physical and hydrologic alteration and partial in-
undation of a watershed. Besides the forced removal of
people and inundation of homes, hydroelectric plants
also cause significant ecological changes.

More than 40,000 large dams now straddle rivers
around the world, creating reservoirs that inundate
more than 400,000 km². The Akasambo Dam on the
Volta River in Ghana created the largest impoundment
in the world, covering 8500 km². The Three Gorges
through the year 2080. The Soviet government evacu-
ated all people from a zone of 30 km radius and con-
structed a cement sarcophagus to contain the remains
of the reactor core.

Lethal radiation killed many conifers and small
mammals within 10 km of the accident in the first few
weeks, but populations have since mostly recovered.
By 1998, radioactive decay had diminished the amount
of radioactive materials in the immediate area to 1% of
their original amount, mainly as Cesium-137 in topsoil.
Trees have accumulated Cesium-137 in growth rings.
Grass, mushrooms, and berries also continue to incor-
porate the isotope, perpetuating a source of exposure
for species that feed on contaminated plants. Aquatic
ecosystems have generally tolerated the radioactivity
concentrating in sediments, although fish may be accu-
mutating radionuclides.

Ecologists have still not determined the long-term
genetic effects of the fallout from Chernobyl. Appar-
ently, the accident did not eliminate any plant or animal
species, except where cleanup activities involved soil
removal. Indeed, as a result of the evacuation, some
plant and animal populations have thrived.

From 1961 to 1976, ecologists, led by George M.
Woodwell, examined the chronic effects of irradiating a
forest at Brookhaven, New York. Gamma radiation
from Cesium-137 caused sensitive species to die, allowing resistant species and ruderals to invade. Spe-
cies richness in 2 m square plots fell by half.

E. Hydroelectric

Hydroelectric systems harness the potential energy rep-
resented by an elevated mass. The potential energy of
water at elevation will convert to increased kinetic en-
ergy of the water when the water runs to a lower eleva-
tion. A dam concentrates the difference in elevation,
termed hydraulic head, in a spillway equipped with a
turbine and electric generator. The electricity produced
immediately enters the electric grid. In this manner, a
hydroelectric plant will generate electricity with few
direct air emissions and little thermal discharge. The
principal effects of hydroelectric plants come from the
total physical and hydrologic alteration and partial in-
undation of a watershed. Besides the forced removal of
people and inundation of homes, hydroelectric plants
also cause significant ecological changes.

More than 40,000 large dams now straddle rivers
around the world, creating reservoirs that inundate
Dam under construction in the People's Republic of China will be the hydroelectric plant with the highest generation capacity in the world, 18.2 GW. The project, under construction in the period 1993–2009, will flood 1100 km² along 600 km of the world’s third longest river, the Yangtze, and displace 1.2 million people. The dam will require 26 million m³ of concrete.

The inundation of formerly dry land submerges vegetation and immediately decreases the area of animal habitat. Lost forests represent ecosystem services and biomass wasted to decomposition. In Brazil, a country that depends on hydroelectric for 20% of its industrial energy, the land requirement for hydroelectric reservoirs averages 450 km² GW⁻¹ with a range of 17 to 10,000 km² GW⁻¹.

A dam blocks nutrient-rich sediment that a river system otherwise would have deposited in floodplains, wetlands, and at the outlet delta. Not only does the sediment buildup fill in a reservoir and eventually impair electricity generation, but the blocked sediment also represents a source of organic carbon and other nutrients wasted at the bottom of the reservoir. As the delta, bay and estuary topography changes, mudflat areas decrease, and nutrient-rich upwellings can decline. The Aswan High Dam blocks 98% of the 120 million tons of sediment that the Nile had carried to the sea each year, formerly depositing 10 million tons on the floodplain and delta. Consequently, soil depth has thinned and agricultural production has declined in the Nile Valley. Blockage of sediment and fresh water by the Akosombo Dam in Ghana has caused the decline of clam populations in the Volta estuary, and populations of Sphyraena barracuda (barracuda) offshore in the Gulf of Guinea. In addition, coasts become more susceptible to erosion.

Utilities start and stop the flow of water based on electricity and operational requirements. One operational objective is to smooth out natural extremes in the flood regime. This will usually change the meandering response and other channeling processes of a river. Ever since the Glen Canyon Dam removed spring floods in the near downstream section of the Colorado River, sandbar erosion has increased because the river does not flow fast or deep enough to move the amount of silt required for extensive sandbar formation. The resulting disappearance of some riparian tree species has led to the decline of Empidonax trailli (Southwestern willow flycatcher) and other birds. To mitigate the problem, the U.S. Department of the Interior staged a controlled flood in 1996.

For some dammed rivers, the flow of water unburdened by silt can deepen the riverbed. This lowers the water table of surrounding land. Also, the depletion of riverbed gravel can harm any species of fish, insect, mollusk, or crustacean that requires gravel bottoms to spawn. Many insect, amphibian, and fish species also use gravel areas for habitat or for protection.

The depth of a reservoir will often keep water at a temperature lower than that in the native river. For example, the Glen Canyon Dam changed the water temperature in the near downstream section of the Colorado River from a range of 0 to 27°C to a relatively constant 8°C. This has been a major factor in the extinction of Ptychocheius huctus (Colorado squawfish), Gila robusta (roundtail chub), and Gila elegans (bonytail chub) and in the endangerment of five other fish species. Whereas the release water is clear, reservoir water often becomes slightly eutrophic and turbid. This degraded water quality can harm certain species.

The impacts of dams on anadromous fish relate to the migratory behavior and timing of the life cycles of these unique species. Dams render hazardous the downstream migration of young fish and block the upstream migration of adults. Moreover, salinity and temperature adaptations occur on a precise schedule, making long delays lethal. Disoriented and fatigued fish more easily fall prey to predation. Despite the deployment of extraordinary means in contemporary times to facilitate fish migration, including fish ladders, elevators, and trap and haul trucking, dams have eliminated anadromous species from many rivers. Runs of Salmo salar (Atlantic Salmon) and Alosa sapidissima (American shad) have disappeared from many rivers in the Northeast United States. In the Columbia River Basin in the Northwest United States, overfishing, pesticide runoff, and hydroelectric plants have endangered populations of Oncorhynchus nerka (Snake River sockeye salmon) and Oncorhynchus tschawytscha (Snake River chinook salmon). The physical barrier formed by a dam can even divide populations of aquatic species, altering patterns of gene flow and genetic variation.

Exotic fish species adapted to human disturbance and introduced into reservoirs for sport fishing will often outcompete native species. In the 1350 km² reservoir straddling the Brazil–Paraguay border behind the Itaipu Dam—with a generation capacity of 12.6 GW that currently ranks it the highest in the world—a non-native species, Plagioscion squamosissimus (curvina) has become the second most numerous species. PCBs released from circuit breakers and oil leaking from machinery constitute the worst direct industrial pollution from dams. These and toxic organics build up in sediments and magnify up through the food chain. Impingement of aquatic organisms on intake screens...
The material and energy needs for building the massive infrastructure of the hydroelectric energy cycle produces wide-ranging environmental effects. The end uses of hydroelectricity will produce the environmental effects associated with climate control, lighting, commercial production machinery, residential appliances, and other electric devices. Because smelting aluminum from bauxite ore requires a large amount of electricity, aluminum smelting comprises an end use closely tied to the hydroelectric option. The air emissions from smelters include CO, CO₂, particulates, NOₓ, and trace metals. Major water emissions include trace metals and sulfates.

F. Renewable Energy Technologies

Renewable energy includes those forms of energy whose transformation does not necessarily consume the ultimate source of the energy, harnessing instead solar radiation, wind, the motion of water, or geologic heat. This section covers renewable energy technologies, including solar heating, solar thermal electric, solar photovoltaics, electric wind turbines, biomass-to-electricity conversion, biomass-to-alcohol fuels, and geothermal electric. These are renewable energy systems that depend on complex technology, so they are forms of industrial energy. The following section on traditional energy covers the simpler forms of renewable energy—firewood and charcoal.

Renewable energy forms share the physical characteristics of site specificity, variable availability, diffuse flow, and low or no fuel costs. Except for biomass-to-electricity conversion, renewable energy technologies do not involve combustion, so they do not directly produce much air pollution. The major environmental impacts derive from the fabrication, installation, and maintenance of renewable energy devices.

Solar energy systems fall into the categories of passive and active. Passive solar technologies consist of architectural forms that more effectively follow the diurnal and seasonal patterns of sunlight for the efficient heating and cooling of a building. Passive systems use the natural phenomena of radiation and convection. On the other hand, active systems use moving devices to achieve heat transfer. The simplest active systems use pipes or other collectors to heat water for residential or commercial use. For the most part, solar heating is environmentally benign.

Solar thermal uses arrays of reflective collectors to focus sunlight on a water boiler for the turbine production of electricity. These systems require significant amounts of land for the parabolic or trough collectors. Because solar is generally economically feasible only in hot sunny areas, sites are generally arid and water is scarce. Water withdrawals for the turbine and for washing the collectors can damage aquatic ecosystems. The bright arrays can also harm birds. Photovoltaics are solid-state devices in which photons stimulate the emission of electrons and semiconductor materials channel the electrons for collection. In this way, photovoltaics directly convert sunlight to electricity with no moving parts, except for devices that move photovoltaics to track the sun, and no water, except for water to occasionally wash photovoltaic surfaces.

The fabrication of photovoltaic (PV) cells produces noxious environmental impacts. The first step is mining the quartz that constitutes the base material of a PV cell, so this produces many of the impacts on aquatic and terrestrial biodiversity described for coal mining. Then the production of metallurgical grade silicon requires the refining of quartz to 99% purity at 3000°C in an electric arc furnace. The production of semiconductor grade silicon occurs through a fluidized bed reaction of the silicon with hydrochloric acid. Then the production of semiconductor grade polycrystalline silicon occurs by electrically heating at 1000°C the semiconductor-grade silicon for vapor deposition on a silicone substrate. Remelting the polycrystalline silicon produces a form that can grow into crystals. These crystals are sawed into wafers 0.5 mm thick, wired, and encapsulated in glass 3 mm thick.

Trace metals are used to dope the semiconductor for the principal types of photovoltaic cells, including gallium arsenide, copper indium diselenide, cadmium telluride, indium phosphide, and cadmium sulfide. The trace metals, together with chlorinated organic solvents and phosgene gas, produce hazardous air, water, and solid wastes that can be lethal on contact or carcinogenic in small doses. The conversion of biomass into electricity involves burning specially grown wood or crops in low pressure boilers to power steam turbines or the gasification of organic matter into methane to power gas turbines. The United States currently possesses a biomass electricity-generating capacity of 7.6 GW. Biomass-for-energy plants in the United States often employ cogeneration to provide process heat for an adjacent industrial facility. The principal species used include short rotation trees Populus spp. (poplars, aspen, cottonwoods), Plat-anus spp. (sycamore), and Acer saccharinum (silver maple) grown at densities of 1600 to 5000 trees ha⁻¹. Herbaceous energy crops include Panicum virgatum
(switchgrass) and Andropogon gerardii (big bluestem). Brazil generates electricity and cogenerates heat from the organic wastes, or bagasse, left from the processing of sugarcane and orange juice.

The environmental impact of an energy crop depends on the previous land use as well as the cultivation techniques of both the new crop and any previous old crops. An energy crop can generate negative effects on biodiversity if it is grown in monoculture, if it is grown using pesticides, and if vegetation is clear-cut to prepare for the energy crop. Still, if previous land-uses were less environmentally sound than the energy crop, then the energy crop constitutes a mitigating practice. Previous sections detail the negative biodiversity effects of electric turbines and condensers.

The conversion of biomass into alcohol fuels also requires the dedicated growing of energy crops. Fermentation of cellulose and other complex carbohydrates produces ethanol, which certain engines can burn straight or mixed with gasoline. The United States now produces 4 billion liters of ethanol annually, mainly from corn. Brazil produces enough ethanol from sugarcane to provide for 10% of the country’s energy use.

Otto Cycle engines burn neat ethanol, a mixture of 96% ethanol and 4% water. Modified conventional automobile engines can burn gasohol, a mixture of 78% gasoline and 22% ethanol. At one time, half of the automobiles in Brazil ran on gasohol, the other half on ethanol. Because the combustion of ethanol mainly produces CO₂ and water, with much smaller amounts of hydrocarbons and NOₓ than gasoline combustion, ethanol used for transportation mitigates the most harmful direct effects of petroleum.

For centuries, society has captured wind for moving sailing ships, pumping water, and milling grain. Contemporary wind turbines also power electric generators. Rated at 100 to 300 kW per wind turbine, the steel machines reach heights of 10 m with pinwheel diameters up to 7 m. Arranged in arrays of up to hundreds of turbines, wind “farms” occupy considerable land areas. The greatest arrays cover unique areas in the Altamont Pass in the San Francisco Bay Area in California, the Tehachapi Pass in Southern California, in the Netherlands, and Denmark. Wind farms fragment terrestrial habitats and access road networks cause soil erosion. Spinning turbines can also kill birds.

Geothermal energy captures the heat of hot geologic formations, generating more than 7 GW of electricity worldwide. Geothermal plants sink pipes down to either capture deep hot water or to inject water for it to boil on contact with hot rocks. This process mobilizes trace metals contained in certain geologic strata and releases H₂S gas associated with geothermal deposits.

### IV. BIODIVERSITY IMPACTS OF TRADITIONAL ENERGY

The most important sources of traditional energy are firewood and charcoal, which is produced from firewood. Local people harvest firewood either by coppicing, or cutting at the base, moderately sized shrubs, by lopping branches off mature trees, or only rarely by felling whole trees and splitting the logs.

In semiarid areas of Africa, women prefer the straight, moderately sized branches that only coppiced shrubs produce. Each year, women and, sometimes, their husbands or fathers go out in the dry season and cut at the base shrubs mainly in the family Combretaceae, carry the branches back to the village, and let them dry out. Just before the first rains, men and women cut a store of firewood for the rainy season. This serves, first, to avoid cutting wood that is wet and difficult to burn and, second, to get a time-consuming and strenuous chore out of the way before the exhausting and rushed rainy season. Coppiced shrubs will resprout in the rainy season and, in a year, regrow a full set of branches. When shrubs become scarce, women begin to pull down branches from adult trees, sometimes using long-handled hooks. This harms the growth potential of a tree by removing shoot apical meristem tissues and only provides difficult, thorny branches. When branches are exhausted, women fall back on noxious, dead stalks of spurges, family Euphorbiaceae. The last resort is animal dung. Only rarely will people cut down an adult tree for their own firewood needs. Men cut down trees for firewood for community events, large baptisms, weddings, or funerals, but even then, men prefer trees that have already died because these yield dry, more combustible wood.

Although women carry firewood for rural use, rural people load beasts of burden and carts to transport wood for sale in urban areas. So a town or city can produce land use changes far beyond its borders.

The low energy density of wood makes its transport onerous relative to the energy gained. Conversion of firewood to charcoal creates a product with double the energy per unit mass, but emits as waste heat up to two-thirds of the energy contained in the original wood. Charcoal makers cut down live and dead trees, particularly prizing sturdy tree trunks. In the field, they pile the wood, cover it with soil to form a kiln 1 to 3 m in
height and 1 to 3 m in diameter, and ignite a slow burn. Over 3 to 6 days, the wood converts to charcoal by partially anaerobic pyrolysis. Firewood harvesting can superimpose locally severe changes in biological diversity over wider alterations caused by long-term climate change. Global warming, the El Niño Southern Oscillation, and desertification all lead to systemic ecological changes at a regional scale. Embedded within these regions, firewood harvesting can reduce vegetative cover in less extensive areas where people depend on wood for their energy use.

The impacts of desertification in the West African Sahel clearly illustrate this complex situation. In Senegal, anthropogenic and climate factors caused a decline in forest species richness of one-third in the last half of the twentieth century (Gonzalez 1997). Rainfall in the Sahel has shown a persistent downward trend in the past four decades, with the rainfall average of all years since 1919 falling at Louga, Senegal (15°37’ N, 16°14’ W) from 470 mm in 1953 to 400 mm in 1993. Serious droughts have hit in the periods 1910–1914, 1942–1949, and 1968–1973. An increase in human population has coincided with the decline in rainfall. The population of Senegal doubled in the period 1945–1988, growing at a rate of 0.025 y⁻¹.

In Northwest Senegal, the average forest species richness of areas of 400 ha fell from 64 ± 2 species ca. 1945 to 43 ± 2 species in 1993. Moreover, densities of trees of height ≥ 3 m declined from 10 ± 0.3 trees ha⁻¹ in 1954 to 7.8 ± 0.3 trees ha⁻¹ in 1989. Both the fall in species richness of 33 ± 5% and the decrease in tree densities of 23 ± 5% translate to a rate of −0.8% per year.

In West Africa, rainfall increases and evapotranspiration decreases toward the equator, creating a gradient that differentiates species into three broad bands of increasingly mesic vegetation: the vegetation zones of the Sahel, the Sudan, and Guinea. In Senegal, arid Sahel species (e.g., Family Mimosaceae) expanded in the north, tracking a concomitant retraction of mesic Sudan (e.g., Family Caesalpiniaceae) and Guinean species (e.g., Family Bombacaceae) to the south. Vegetation zones shifted southwest 25 to 30 km in the period ca. 1945–1993 (Fig. 4), a rate of 500 to 600 m y⁻¹, foreshadowing the magnitude of projected shifts driven by CO₂-induced climate change. The historical change acted through a higher mortality among mesic species, leaving drought-resistant species to dominate the remaining tree cover. The most notable species that have experienced local extinctions include Dalbergia melanoxylon (Senegal ebony), Prosopis africana (ironwood), Sterculia setigera (mbep), and Tamarindus indica (tamarind).

Out of 215 ecological and socioeconomic variables, multivariate statistical analyses identifies rainfall and temperature as the most significant factors explaining the distribution and densities of trees and shrubs in Northwest Senegal. Rainfall and temperature override local anthropogenic factors.

Examination of dead trees along the coast supports a predominance of climatic over local anthropogenic factors. The sparsely populated coast offers a view of the state of the countryside before cultivation. Natural stands of Euphorbia balsamifera still occur there. In contrast, elsewhere in the Senegal Sahel, farmers have cut all natural stands of this species and replanted it...
along field boundaries. In the collective memory of local people, vast areas along the coast have not been cultivated. Dead trees still stand along the coast, but they show no ax marks or any sign that humans directly caused their death.

On a subcontinental scale, however, human activities may have caused the decline in rainfall. Deforestation of tropical rain forests in the Congo vegetation zone from the Republic of Guinea to Côte d’Ivoire may have reduced the evapotranspiration inputs essential to the maintenance of the Southwest Monsoon. Reduced rainfall over an extended period would reduce the vegetation cover in the Guinea zone. Thus in turn would decrease rainfall and vegetation in the Sudan, eventually reducing rainfall and vegetation in the Sahel. Thus, human activities in the distant rain forests may initiate a concatenation of climatic links that ultimately touch the Sahel.

Nevertheless, population growth has undoubtedly placed increasingly inordinate pressures on the area’s vegetative cover. In Northwest Senegal, rural firewood use exceeds firewood production from shrubs over 90% of the land area, affecting 95% of the rural population. The rural population density of 45 people km⁻² exceeded the 1993 carrying capacity of firewood from shrubs of 15 people km⁻² (range 1–21 people km⁻²). The rural population density has exceeded carrying capacity since 1996.

The standing biomass of trees across the research area decreased from 14 t ha⁻¹ in 1956 to 12 t ha⁻¹ in 1993, matching a cumulative firewood deficit in the same period of 2 t ha⁻¹. The reduction in standing biomass released carbon into the atmosphere at a rate of 60 kg C cap⁻¹ y⁻¹, somewhat less than the 100 kg C cap⁻¹ y⁻¹ released from the burning of fossil fuels, mainly by the urban industrial and transport sectors.

Not only do the quantitative uses of firewood and charcoal exceed the area’s wood production, but the fall in species richness has also reduced people’s options qualitatively. For example, rural women depend on two particular shrub species for firewood because of the size of the branches, high wood density, and ease of collection. Beyond that, few fallback species remain. The fraction of women that reported shrub species as most prevalent in firewood use fell from 87% ca. 1945 to 50% in 1993. With respect to traditional medicine, 25 useful species have diminished significantly. Furthermore, eight species that provided fruit, leaves, and gum in past droughts have disappeared from as much as 53% of their range. If a grave famine hit the area in its current condition, people would not be able to find the emergency foods that saved others in past episodes.

In the Sahel, the natural regeneration of local species could halt the declines in biodiversity and forest biomass. Natural regeneration is a traditional practice in which farmers and herders protect and promote the growth of young native trees. Traditionally, local people protect small trees that have germinated naturally or regrown from roots, prune them to promote growth of the apical meristem, and, if necessary, set a stake to straighten the small tree.

Natural regeneration has expanded Acacia albida from an original restricted range along rivers in Southern Africa over thousands of km² up through the Sahel and the Sudan. In Senegal, the Sereer have protected dense parks of Acacia albida and Adansonia digitata in wide areas south of the research area. On the Mossi Plateau in Burkina Faso, farmers have similarly protected expanses of Butyrospermum parkii for the valuable oil from the tree’s seeds. Across the Sahel, leather workers protect Acacia nilotica adansoni for the tannin enriched bark.

Natural regeneration requires no external inputs. It concerns species well known and appreciated by villagers. It focuses on young trees that have demonstrated their hardiness by surviving with no human caretaker, no watering, and no special treatment. Furthermore, natural regeneration not only augments the supply of wood, poles, fruit, medicine, and other products, it puts trees where farmers and herders really need them: in fields to maintain soil fertility and in pastures to provide forage.

Although photosynthetic activity in semiarid lands is an inefficient conversion of the total available solar radiation, the inefficiency of human tools renders end uses even more inefficient in the final conversion into heat and light. Table III shows this energy chain from sunshine to wood end use in the West African Sahel.

### TABLE III

<table>
<thead>
<tr>
<th>Source of Energy</th>
<th>West African Sahel (W ha⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunlight</td>
<td>2,400,000</td>
</tr>
<tr>
<td>NPP</td>
<td>1,720</td>
</tr>
<tr>
<td>Total wood production</td>
<td>130</td>
</tr>
<tr>
<td>Human wood energy use</td>
<td>213</td>
</tr>
<tr>
<td>Imported fossil fuels</td>
<td>93</td>
</tr>
<tr>
<td>Food consumption</td>
<td>53</td>
</tr>
<tr>
<td>Human wood energy end-use</td>
<td>11</td>
</tr>
</tbody>
</table>

Gonzalez 1997.
Therefore, another practice that can serve to conserve vegetative cover in rural areas dependent on firewood is the use of improved efficiency cook stoves. In many areas, women customarily cook with a kettle over an open fire. International development agencies have worked to develop and introduce stoves such as the ban ak suuf in Senegal, a horseshoe-shaped hearth constructed from clay and cow dung that provides a more enclosed combustion space to more effectively channel heat to the cooking vessel. The lorena in Guatemala is another earthen stove. The jiko in Kenya and sakkanal in Senegal are enclosed metal or ceramic charcoal stoves that more effectively contain heat than do traditional open charcoal burners.

V. FUTURE ENERGY PATHS

Human energy use directly alters patterns of biodiversity through changes in land use and through industrial pollution. Indirectly, human energy use is changing global biodiversity through the emission of greenhouse gases that cause global climate change and through other broad environmental effects of industrialization. Not only does the direct processing of energy generate environmental impacts, but the end uses that convenient energy forms make possible produce impacts locally and globally. Table IV summarizes the major environmental impacts of human energy use on biodiversity. Almost every source requires land (Table V), a requirement that leads to habitat fragmentation and destruction. Globally, the climate change caused by CO2 emissions constitutes the major impact of fossil fuels, but nonfossil fuel sources also produce air and water pollution. No energy transformation system operates without negative environmental effects, yet renewable sources generally restrict harmful effects to the capital formation stage and do not produce much ongoing pollution.

Holdren and Ehrlich (1974) proposed that environmental impact is equivalent to the multiplicative effect...
of population, affluence, and technology.

Environmental Impact = Population \times \frac{Resource use}{Person} \times \frac{Environmental Impact}{Resource use}

People now call this identity the IPAT equation (impact = population \times affluence \times technology). Because the environmental impact of human energy use is proportional to the rate of energy use, and energy use is proportional to economic production, then the IPAT equation for energy becomes

Environmental impact \propto \frac{energy use}{GNP \times Person} \times \frac{GNP}{energy use}

Here, economic production per person indicates the level of material affluence, while energy use per unit of economic production indicates the level of technological efficiency. This highlights the great leverage that both energy conservation and efficiency wield to reduce the environmental impact of energy use. Indeed, improvements in energy efficiency reduced the energy intensity of economic activity in the United States by nearly one-third between 1973 and 1993.

For a 2300 AD global population of 10 million people to stay within the environmental limits of the earth, Holdren (1991) suggests that industrial countries im-

prove their energy efficiency to allow for an increase in economic activity in the nonindustrial countries so that everyone converges on an average use of 3 kW per person. This would increase total world energy use to 30 TW, more than double today’s total. Imagine a world with twice as many nuclear power plants, coal mines, automobiles, and other energy infrastructure as today. Because many environmental impacts increase exponentially, the total impact will be much more than double.

Yet, the historical path of industrialization has left the world with only costly and environmentally disruptive energy alternatives. The earliest exploitation of fossil fuels depleted the most convenient oil and gas deposits. This is why fossil fuel production over time follows the bell-shaped Hubbert Curve. Not only has this left the current generation with deposits that are farther in polar and desert regions, deeper underground, and dispersed, but the low-cost energy has also shaped the expectations of people around the world for inexpensive on-demand energy services. Societies even subsidize the provision of convenient energy through infrastructure support to energy industries, tax breaks to oil drillers, preferential treatment to automobile companies, and other schemes.

The depletion of nonrenewable resources (Table VI) and other serious environmental and social constraints hobble most energy options for the future: recoverable oil and gas reserves will last only another 30 years; coal burning releases the principal agent of global warming, CO₂; biomass energy requires vast amounts of land; the small number of exploitable sites limits the potential

<table>
<thead>
<tr>
<th>TABLE V</th>
<th>Land Requirements and Major Air Emissions for Electric Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land req. (ha MW⁻¹ h⁻¹)</td>
<td>CO₂</td>
</tr>
<tr>
<td>Geothermal</td>
<td>0.1–0.3</td>
</tr>
<tr>
<td>Natural gas turbine</td>
<td>0.3–0.6</td>
</tr>
<tr>
<td>Wind electric</td>
<td>0.4–1.7</td>
</tr>
<tr>
<td>Nuclear</td>
<td>0.8–1.0</td>
</tr>
<tr>
<td>Coal</td>
<td>0.8–6.0</td>
</tr>
<tr>
<td>Solar thermal electric</td>
<td>1–4</td>
</tr>
<tr>
<td>Hydroelectric</td>
<td>2–1000</td>
</tr>
<tr>
<td>Photovoltaics</td>
<td>3–7</td>
</tr>
<tr>
<td>Biomass</td>
<td>150–1000</td>
</tr>
</tbody>
</table>

OTA 1995.

<table>
<thead>
<tr>
<th>TABLE VI</th>
<th>Estimates of World Energy Resources at Current Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonrenewable stocks</td>
<td>TWy</td>
</tr>
<tr>
<td>Petroleum</td>
<td>600</td>
</tr>
<tr>
<td>Natural gas (conventional)</td>
<td>400</td>
</tr>
<tr>
<td>Coal</td>
<td>3000</td>
</tr>
<tr>
<td>Heavy oil, tar sands, unconventional gas</td>
<td>1000</td>
</tr>
<tr>
<td>Uranium</td>
<td>3000</td>
</tr>
<tr>
<td>Renewable flows</td>
<td>TW</td>
</tr>
<tr>
<td>Solar electric</td>
<td>32</td>
</tr>
<tr>
<td>Biomass</td>
<td>26</td>
</tr>
<tr>
<td>Hydroelectric</td>
<td>1.2</td>
</tr>
<tr>
<td>Wind electric</td>
<td>1.3</td>
</tr>
</tbody>
</table>

for hydroelectric and wind power; and health and safety concerns prevent expansion of nuclear energy.

As a response to these constraints, governments, industry, and academia have placed enormous effort into the development of technologies such as electric vehicles, hydrogen cars, fuel cells, and nuclear fusion. Still, the future of human energy use may not lie with exotic devices. Instead, renewable energy sources, conservation, and efficiency of end use could form the future energy path of the world. Amory Lovins (1977) called this the “soft path.” This would occur if societies set as their goal the provision of services, not just the acquisition of energy stocks and devices. In effect, people don’t require light bulbs, they need illumination.

Acknowledgment
The author wishes to thank John P. Holdren, from whose work, both published and unpublished, much of the material in this entry derives.

See Also the Following Articles

DEFORESTATION • DESERTIFICATION • ECOLOGICAL FOOTPRINT, CONCEPT OF • ENERGY FLOW AND ECOSYSTEMS • GREENHOUSE EFFECT • HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW

Bibliography


ENVIRONMENTAL ETHICS

Richard B. Primack* and Philip J. Cafaro†
*Boston University and †Colorado State University

I. Duties to Protect Biodiversity: Intrinsic Value
   Arguments
II. Duties to Protect Biodiversity: Obligations
   Human and Divine
III. Biodiversity and Human Flourishing
IV. Environmental Philosophies

GLOSSARY

anthropocentrism Position that only human beings
   have moral worth or intrinsic value.
biocentrism Position that all living beings have moral
   worth or intrinsic value.
deep ecology Activist philosophy that advocates radical
   personal and political change to protect wild
   nature.
ecofeminism Liberation philosophy that draws con-
   nections between preserving nature and promoting
   women’s rights.
environmental ethics Philosophical discipline that
   specifies proper human relationships to the natural
   world.
ethical holism Position that complex aggregates such
   as species, ecosystems, or human societies have intrin-
   sic value.
instrumental value Value of something relative to hu-
   man interests or desires.
intrinsic value Value of something independent of its
   value to people.

ETICS IS THE BRANCH OF PHILOSOPHY that seeks
knowledge of human flourishing and right conduct to-
ward others, so that we may act upon it. Modern philo-
osophers have tended to limit their ethical concern to
human beings, but throughout history people have also
attempted to cultivate proper relationships to the rest
of nature. Recently philosophers have turned to this
topic, largely in response to environmental degradation
and the loss of biodiversity, and have created a new
discipline: environmental ethics. Environmental ethi-
icists attempt to specify appropriate human relation-
ships to the nonhuman, natural world. In the course
of their work they have developed strong ethical argu-
ments for preserving biodiversity. They have also chal-
enged conventional views of happiness and human
welfare and the materialistic values at the base of much
modern life. While environmental ethics treats the full
range of environmental issues, from air pollution to
nuclear risk assessment, this article focuses on ethical
issues directly related to the preservation of biodi-
versity.
I. DUTIES TO PROTECT BIODIVERSITY: INTRINSIC VALUE ARGUMENTS

A. From Economics to Ethics

Biodiversity has great economic value, both direct and indirect. Direct economic values are provided by wild game, seafood, fuelwood and timber products, and indigenous and high-technology medicines. Indirect economic values accrue from biodiversity's roles in waste disposal, climate regulation, protecting soil and water resources, recreation and ecotourism, and much more. Both directly and indirectly biodiversity safeguards or adds to human wealth, often justifying its protection in purely economic terms.

Thus economic arguments by themselves provide a basis for valuing and protecting species, especially when we expand our concept of economic value from short-term, next-quarter profits to include longer-term and indirect benefits. But economic arguments can also provide grounds for extinguishing species or for saving one species rather than another. In conventional economic terms, low value will be given to species having small populations or limited geographical range, small physical size or unattractive appearance, no immediate use to people, or no relationship to other species of economic importance. Such qualities may characterize a substantial proportion of the world's species, particularly insects and other invertebrates, fungi, nonflowering plants, bacteria, and protists. Halting profitable developments or making costly attempts to preserve these species may not have any obvious economic justification. In some circumstances, economic justification could exist for destroying an endangered species, particularly organisms that cause disease or attack crop plants. Still, many people believe that such destruction is morally wrong even if it is economically profitable.

To say that an object has economic value is to say that it is useful to human beings or that they desire to possess it. Put another way, it has an instrumental value: someone can use it as an instrument for his or her purposes. On the other hand, we recognize that at least some entities, such as human beings, have value regardless of whether anyone else values or uses them. These entities have an intrinsic value: a value that is grounded not in their usefulness to others, but in what they are themselves.

Human beings, we usually think, have both instrumental and intrinsic value. Because we have intrinsic value we possess certain rights that no one can legitimately infringe, even if it is in their self-interest to do so. Conversely we have certain duties towards other people that specify how we should or should not treat them in various situations. We cannot help but look at the people we interact with each day in terms of their usefulness to us, but if we look at them solely in these terms we disrespect their (and our) humanity and if we treat them solely as a means to our own ends we are likely to behave immorally.

Many conservationists argue that similar duties restrict the morally acceptable treatment of wild nature. They consider it wrong to destroy a rare woodland or cause a species to go extinct, even if this action is in an individual's or corporation's self-interest. Not only have we no right to destroy any species, we have a moral responsibility to actively protect species from extinction as the result of our activities. Opponents counter that this position is illegitimate because only human beings have intrinsic value. Unless our actions affect other people, directly or indirectly, any treatment of the natural world is morally acceptable.

B. Extensionist Arguments for Intrinsic Value

Proponents provide both extensionist and nonextensionist arguments for the intrinsic value of wild nature. Extensionist arguments ask what qualities give intrinsic value to human beings, then assert that some other beings possess these same qualities. Therefore, they conclude, we should recognize the intrinsic value of these other beings and extend proper treatment to them.

One common justification for valuing human beings is our ability to reason. But some of the so-called "higher animals" also seem to possess the rudiments of reason. Chimps and gorillas have been taught sign language involving several hundred words; wolves have an elaborate social life and the ability to coordinate long hunts; dolphins, whales, and other cetaceans send complex signals that we are just beginning to understand. Many argue that because of these factors, we should not hunt these animals for food or sport, use them in research, or in general treat them solely as means to our own ends.

Other philosophers believe that sentience—the ability to perceive the surrounding world and feel pleasure and pain—demands moral consideration. The extensionist argument here states that because pain is bad we should avoid inflicting pain unnecessarily on others who can feel it—even if those others have fur, feathers, scales, or numerous legs. Immoral actions against human beings are wrong, at base, because they cause unnecessary pain; actions that cause other animals unnec-
essential pain are likewise immoral. Millions of people around the world act on such beliefs by abstaining from eating animals or using products that can only be procured by killing or harming them. Even most meat-eaters accept that inflicting gratuitous pain on animals is wrong, suggesting a widespread belief that where there is sentience there exists some moral responsibility.

One possible problem with the extensionist arguments just discussed, from the point of view of conservationists, is that they only encompass certain organisms. Basing intrinsic value on some rudimentary form of reasoning or complex mental experience would appear to rule out most animals, for example, almost all invertebrate species. Basing moral consideration on the ability to feel pain includes a wider class of animals, but again many simpler animals and all plants, fungi, and single-celled organisms are ruled out.

However, extensionist arguments can be extended further by recognizing that all organisms have a drive to stay alive and reproduce. In the same way that we acknowledge the rights of people to live, have children, and satisfy basic needs, we may extend these rights to individuals of all species. In his autobiography, Albert Schweitzer maintained that “a man is ethical only when life, as such, is sacred to him, that of plants and animals as that of his fellow men, and when he devotes himself helpfully to all life that is in need of help.”

This position may lead to a very demanding code of conduct! Believers in strict biocentric equality assert that it is always wrong to kill individuals of any species (because of their intrinsic value) unless we need to do so to survive. Others believe that using nature to provide necessities and some measure of comfort is morally acceptable, but not to provide luxury goods. In this view, cutting down trees for firewood or to build a modest house is morally acceptable, particularly if this is done sustainably, whereas harvesting mahogany trees from rain forests to make expensive furniture for wealthy individuals half a world away is morally unacceptable. Whatever the particular judgments made, recognition of an intrinsically valuable organic world leads to distinctions between essential and inessential human uses and to a more limited use of natural resources.

C. Nonextensionist Arguments for Intrinsic Value

In addition to extensionist arguments, which point up similarities between wild nature and intrinsically valuable humans, there are arguments that find value in nature without referring to such similarities. Some believe it is a mistake to value other beings only for the ways in which they resemble humans. Writing in The Monist in 1992, Robert Elliot suggested that natural organisms may have the following properties that give them intrinsic value: “diversity, stability, complexity, beauty, harmony, creativity, organization, intricacy, elegance, and richness.” These are qualities of natural organisms that we can appreciate—and that may call forth responses of personal restraint and active protection.

All species represent unique biological solutions to the problem of survival. They have solved the challenges placed before them by their environments and thrived. Some people see a value in this creativity and time-tested uniqueness. Others appreciate the beauty and elegance of the natural forms created by this process. Still others value the complexity and ingenious structures that science and close observation have revealed. People who know and value this uniqueness, beauty, and complexity feel a special horror at its permanent disappearance. After all, if an individual dies it may be replaced by another individual more or less the same, but take away the last passenger pigeon or giant moa and their like will never return. Nonextensionist arguments thus support preservation of species as well as protection of individual organisms (see Section 1.E).

Species interact in complex ways in natural communities. The loss of one species may have far-reaching consequences for other members of the community: other species may become extinct or the entire community may become destabilized. As we learn more about global processes we are finding out that many chemical and physical characteristics of the atmosphere, the climate, and the ocean are linked to biological processes in a self-regulating manner. More diverse biological communities may be better able to deal with environmental disturbances such as drought and global climate change. For these reasons, if we value some species, we should arguably protect all species.

D. Anthropocentric Denials of Intrinsic Value

Skeptics reply that even though some people do value non-human organisms and species, we are not obligated to do so, because only human beings have intrinsic value or genuine rights. Humans have a value beyond all other beings because only we are fully conscious and rational. Unless our actions affect other people, directly or indirectly, any treatment of the natural world is morally acceptable.

Such a viewpoint is anthropocentric—locating value solely in humans—and to many it seems the most obvious common sense, while departures from it seem irr
tional or overly sentimental. The appellation "tree hugger" expresses this view, suggesting inappropriate sentiments toward trees, leading to inappropriate actions. It also suggests a callous disregard for the interests of people who cut trees for a living, whom we really should care about (consider the bumper sticker "Hug a Logger, Not a Tree"). People who value nature counter that anthropocentrism is selfish and that speciesism—the privileged treatment of one species over another—is no more justified than racism or sexism.

Although the debate between anthropocentrists and biocentrists has tended to incorporate all the uncertainties attending ethical justification in general, some clarification may still result. Charges of irrationalism to the contrary, it is possible to love and value nonhuman nature and act on this view. The challenge to biocentrists is to fashion fulfilling lives that limit their negative environmental impacts and help preserve and celebrate nature. Anthropocentrism also remains a rationally defensible position, which may be consistently acted upon. Anyone who values humanity based on qualities that we share with other species, however, cannot consistently deny intrinsic value to those other species. Furthermore, anthropocentrists who value humanity primarily for our ability to reason may consider the many arguments in favor of lessening pollution and preserving wild nature that appeal to our rational self-interest. Anthropocentrists are more likely than biocentrists to accept some amount of pollution in rivers or the extinction of certain species, but they also acknowledge human needs for clean air and drinking water, the enjoyment we get from fishing, swimming, and canoeing, and the value of biodiversity to science, art, and business. When it comes to particular environmental policies, anthropocentrists and biocentrists may find considerable common ground.

E. Ethical Holism

The preservation of biodiversity seems to demand that the needs of endangered species take precedence over the needs of individual organisms. For example, the U.S. National Park Service killed hundreds of introduced rabbits on Santa Barbara Island off the California coast to protect a few plants of the endangered Santa Barbara live-forever (Dudleya traskiae) (Fig. 1). In this case, one endangered species was judged to be more valuable than hundreds of individual animals of a common species. Similarly, conservation biologists would not find it acceptable to destroy the last remnant of a rare biological community even if every species living there could be maintained in captivity; the ecological interactions and evolutionary processes of the community would be lost if the species only lived in captivity. These examples illustrate that most conservationists are holists, finding value in larger groupings, such as species and biological communities. They are thus sometimes willing to sacrifice the interests of individuals to preserve species and communities.

Many writers, especially animal welfare advocates, have difficulty assigning rights to species. Peter Singer and Tom Regan, two prominent philosophical animal welfare advocates, argue that species are not conscious entities and so do not have interests. In their view, to sacrifice the genuine interests of an individual animal, who can suffer or possess rights, to the imagined interests of a species, which cannot, is mistaken. Many animal welfare advocates also reject conservationists' spe-
cial concern for native species over exotics: an animal is an animal, with a greater or lesser ability to suffer, and this ability should determine our actions, rather than its point of origin.

On both biological and ethical grounds, however, most conservation biologists argue that species, rather than individual organisms, are the appropriate targets of conservation efforts. All individuals eventually die; it is the species that continues, evolves, and sometimes forms new species. In a sense, individuals are temporary 
representatives of species. Whether or not we allow them rights, species carry great value as the repositories of the accumulated experience and history of millions of previous life-forms through their continuous, evolutionary adaptation to changing environments. The premature extinction of a species due to human activities destroys this natural process and renders its history. It can be regarded as a "superkilling," because it kills future generations of the species and eliminates whole lines from the processes of evolution and speciation.

Furthermore, conservationists typically argue that species should be prevented from spreading beyond their natural ranges as a result of direct or indirect human activity. For example, the zebra mussel (Dreisena polymorpha) is native to the Caspian Sea, but it has recently become an aggressive invader of North American aquatic habitats (Fig. 2). Arguably this species should be destroyed whenever possible in North America, for two reasons. First, exotics often displace native species, sometimes contributing to their extinction. To prevent this loss of biological diversity we destroy individual exotics, on the assumption that the exotic species will continue to thrive in its native habitats. Second, and more controversially, many conservationists deny intrinsic value to individual plants and animals that have spread beyond their natural range due to human activities. Part of what gives a species value is its unique evolutionary history and its ecological roles in native habitats, both of which are tied to particular places and biological communities. When these species instead invade new natural areas and destabilize or radically change their species composition, these justifications of their value no longer hold. Exotic species contribute to biodiversity in their native habitats but often diminish biodiversity when they become established and common in new locales.

II. DUTIES TO PROTECT BIODIVERSITY: OBLIGATIONS HUMAN AND DIVINE

A. Duties to Humans

That human beings have direct moral duties to other species remains controversial; that we have direct moral duties to other human beings is not (although great disagreement remains regarding their scope). Our duties to other human beings may support protection of the environment and preservation of biodiversity, based on their instrumental value to intrinsically valuable people.

We have responsibilities to our neighbors and fellow citizens. Because of this we must arguably minimize damage to the natural environment, since such damage harms not only other species but people as well. Increasingly, connections are being made between environmental pollution and high levels of human disease. For example, massive environmental pollution in the former Soviet Union has greatly increased cancer, birth defects, lead poisoning, and lung disease among the population. Much pollution and environmental degradation is unnecessary and could be minimized with better planning. Often pollution occurs because corporate leaders are unwilling to spend money to prevent it, despite the resulting ill effects on the health, wealth, or happiness of other people. Citizens must recognize the biological and social costs of environmental damage and force corporations to be good neighbors and governments to enact and enforce strong environmental laws, for the common good.
We also have responsibilities to future generations. Economic decision making tends to focus on the short term and it is this economic system that is driving environmental change and degradation. In response, many ethicists have emphasized the importance of intergenerational justice. Humanity's unprecedented numbers and technological power mean that one generation can now radically remake the earth that the next will inherit, for better or worse. If in our daily living we degrade the natural resources of the earth and cause species to become extinct, future generations will pay the price in terms of a lower quality of life.

This truth gives us arguments not just for keeping ecosystems safe for present human health, but also for preserving biodiversity for future human use, enjoyment, and development. For example, if species and wild places are lost, children will be deprived of one of their most exciting experiences in growing up—the wonder of seeing "new" animals and plants in the wild. These concerns are a powerful motivating force for the members of organizations such as the Sierra Club and The Nature Conservancy, who see themselves as land stewards preserving biodiversity for future generations.

Writing in the *Encyclopedia of Environmental Biology* in 1995, Holmes Rolston predicted: "It is safe to say that in the decades ahead, the quality of life will decline in proportion to the loss of biotic diversity, though it is often thought that one must sacrifice biotic diversity to improve human life." Of course, this contention is debatable. Many argue that job creation and increased wealth are more important to future generations than the preservation of biodiversity. Debate on this issue, while inconclusive, is essential, because it forces us to specify the actual, long-term benefits of development projects and the sorts of societies we want to create for our children. Given our unprecedented power and the tendency to put personal interests above the common good, even inconclusive debate may be a force for creating a better future for all.

At a minimum, our duties to posterity seem to require us to live sustainably—that is, limiting our consumption so as not to degrade essential life-support systems or deplete natural resources that future generations will need. The world's governments have formally recognized this through treaties governing air pollution, whale hunting, and disposal of wastes in the ocean. One important treaty, the Montreal Protocol of 1987, reduced the permitted use of ozone-depleting chlorofluorocarbons, leading to a phase-out of production three years later. The grave dangers these chemicals posed to future generations, with their potential to thin the earth's protective ozone layer and cause millions of new cases of deadly skin cancer, clearly outweighed the interests of producers and users of CFCs, especially since many of those uses were not essential to human survival (e.g., in luxury items such as hairsprays and air conditioners) and alternatives were available or quickly developed for most uses.

Similar appeals to the good of future generations have so far proved less successful in convincing world governments to reduce fossil fuel use to slow global climate change. This is partly due to strenuous lobbying by oil and automotive corporations and oil-producing nations, whose leaders have implicitly placed short-term profits above the health and well-being of future generations. Partly though it is because the risks of increased fossil fuel use, though grave, are less certain than those associated with continued use of CFCs, while the economic costs of decreasing fossil fuel use are far greater.

Human maturity leads naturally to self-restraint and a respect for others. Many conservationists agree with Arne Naess (1989), who writes that the further maturation of the human species will involve an "identification with all life forms" and "the acknowledgment of the intrinsic value of these forms" in an expanding circle of moral obligations. Moving outward from oneself, the circle would include duties to our family and relatives, our local community, our country, all humanity, mammals (Save the Whales!), all animals (Save the Snail Darter!), all species (Save the Yellow Lady Slipper Orchid!), ecosystems (Save the Rain Forest!), and ultimately the whole earth (Fig. 3). Such an expansion of ethical concern involves new limitations on acceptable actions, but also new opportunities for personal growth and flourishing.

Although some worry that recognizing intrinsic value in nature devalues human beings, many environmental philosophers argue that respect for human life and human diversity is compatible with a respect for nonhuman nature. Citizens of all countries will be more likely to accept their responsibility for protecting biological diversity, they assert, when they have full political rights, a secure livelihood, and an awareness of environmental issues. Some of the most exciting developments in conservation biology involve supporting the economic development of disadvantaged rural people in ways that are linked to the protection of biological diversity. Helping poor women establish sustainable plots of cash crops and achieve a degree of economic independence may reduce the need to overhunt wild species. Working with indigenous people to establish legal title to their land may give them the means to protect the biological communities in which they live.
Actions taken to protect species and biological communities should whenever possible benefit people as well.

B. Duties to God: Religious Stewardship

Various preindustrial cultures successfully coexisted with a rich local flora and fauna for hundreds of years, in part because their religions and societal ethics encouraged personal responsibility and thoughtful use of resources. People in these societies respected wild animals and plants even as they harvested them or borrowed their habitat for human purposes. For example, the Cherokee Indians of the southeastern United States spoke special prayers to the deer that they killed, telling the deer spirit that they indeed needed the meat, that they would not waste it, and that they would bury the bones with due solemnity. Traditional peoples often treated rivers, mountains, and other ecosystems as sacred places to be approached with reverence and an appreciation for what they were, rather than for what human beings could make of them.

Many modern religious adherents abhor the destruction of species, because they are God's creation. If God created the world, then presumably the species God created have value. Within the Jewish and Christian traditions, human responsibility for protecting animal species is explicitly described in the Bible as part of a covenant with God. The Book of Genesis describes the creation of the earth's biological diversity as a divine act, after which "God saw that it was good" and "blessed them." In the story of Noah's Ark, God commanded Noah to save two of all species—not just the ones human beings found useful. God provided detailed instructions for building the ark, an early species rescue project, saying "Keep them alive with you." After the flood subsided, the animals were released to repopulate the earth. This story, versions of which were told throughout the ancient Near East, can be interpreted as an early awareness of the importance of biological diversity.

The prophet Mohammed, founder of Islam, continued this theme of human responsibility, saying "The world is green and beautiful and God has appointed you as His stewards over it. He sees how you acquit yourselves." Many other religious traditions also support the preservation of nature. For example, Hinduism locates divinity in certain animals and recognizes a basic kinship between humans and other beings, including the transmigration of souls from one species to another. A primary ethical concept in Hinduism and other Indian religions, such as Jainism and Buddhism, is ahimsa—avoiding unnecessary harm to life. In attempting to live this ideal, many religious people become vegetarians and live as simply as possible.

III. BIODIVERSITY AND HUMAN FLOURISHING

Economic arguments stress that we should preserve biological diversity because it is in our material self-interest. Ethical arguments based on the intrinsic value of wild nature, our duties to other human beings, or our duties to God stress that we should sometimes act altruistically; that is, we should set aside our personal interests to preserve biological diversity. A second kind of ethical argument appeals to a fuller understanding of our self-interest, arguing that preserving biodiversity and developing our knowledge of it will make us better and happier people. The following are the main arguments for preserving biodiversity in our own enlightened self-interest.

A. Health and Wealth

It cannot be repeated too often that biological diversity preserves our basic life-support systems of food production, water supply, oxygen replenishment, waste disposal, soil conservation, and more. People are healthier and more productive in clean, intact environments. We
depend on this and should value it. Similarly, biodiversity allows us to create tremendous economic wealth, directly and indirectly. An article published in Nature in 1997 by Robert Costanza and others, “The Value of the World’s Ecosystem Services and Natural Capital,” estimated that the world’s ecosystems produce $32 trillion per year of value to people, substantially higher than the $18 trillion per year of goods and services those people produce themselves. In other words, human society could not exist without what the natural world provides us for free, nor could we afford to pay for substitutes even if they existed.

B. Aesthetic and Recreational Enjoyment

Nearly everyone enjoys wildlife and landscapes at an aesthetic level and this is part of a good life. The beauty of a field of wildflowers in Glacier National Park or a migrating warbler on a spring morning in a city park enriches the lives of those who appreciate them. For many people, high quality of life involves experiencing nature first-hand. Simply reading about species or seeing them in museums, gardens, zoos, or videos will not suffice. Hiking, canoeing, nature photography, and bird watching are physically, intellectually, and emotionally satisfying.

Hundreds of millions of people spend tens of billions of dollars annually in these pursuits, proof enough of their value. As the world becomes more crowded, it becomes ever clearer that these activities are in competition with other human uses. We are creating a world with ever-diminishing opportunities for aesthetic and recreational encounters with wild nature (Fig. 4). If species and ecosystems are not to disappear altogether, they must be consciously preserved.

C. Artistic Expression and Scientific Knowledge

Throughout history, poets, writers, painters, sculptors, and musicians of all cultures have drawn inspiration from wild nature. Nature provides countless forms and symbols for visual artists to render and interpret, while poets have often found their greatest inspiration in either wild nature or pastoral countrysides (Fig. 5). Preserving biological diversity preserves possibilities for all artists. It also allows those of us who appreciate such creative acts access to the sources and experiences that inspired them. A loss in biological diversity could very well limit the creative energies of people in the future and thus restrict the development of human culture. For example, if many species of whales, butterflies, and

![Figure 4](https://example.com/fig4.png)

**Figure 4** Reaching out. Most people find interacting with other species to be an educational and uplifting experience. Here people greet a minke whale that is being rescued after it became entangled in a trawler’s gill net; the fish behind the whale was attached to the net to keep the whale at the surface so it could breathe. Later, rescuers were able to release the whale from the netting. (Photograph by Scott Kimis, New England Aquarium.)
If *Homo sapiens*’ closest living relatives, the great apes, disappear from the wild, we will lose important information regarding human physical and social evolution.

**D. Historical Understanding and Religious Inspiration**

Knowing nature, both scientifically and through personal experience, is a key to self-knowledge and an understanding of human history. In walking the landscapes that our ancestors walked, we gain insight into how they experienced the world, at a slower pace and without mechanized aids. People often forget just how recently humankind has moved to ultrafast transportation, fully illuminated cities that shut out the night, and other aspects of modern life. Preserving natural areas allows us to develop our historical imaginations.

Many religions have traditions of “wandering in the wilderness” in order to commune with God or with spirits. In the Western tradition, Moses, Isaiah, Jesus, and St. Francis of Assisi all sought out the solitude of wilderness. So did the Chinese sage Lao-tzu, the Japanese Zen poet Basho, and generations of Lakota vision-seekers. Being in nature allows us to clear and focus our minds and sometimes experience the transcendent. When we are surrounded by the artifacts of civilization, our minds stay fully focused on human purposes and our everyday lives. Religion probably would not disappear from a totally tamed human environment, but perhaps it would become diminished for many.

In brief, while the preservation of biodiversity sets limits to some human activities, it is a prerequisite for our continued enjoyment of others. There are good reasons to believe that preserving and exploring biodiversity can make us better, happier people. Many conservationists are convinced that a better understanding of our true self-interest would lead to greater efforts to protect biological diversity.

**IV. ENVIRONMENTAL PHILOSOPHIES**

**A. Deep Ecology**

Paul Sears, recognizing that increased knowledge would lead to the questioning of destructive practices common in modern society, and often taken for granted, called ecology a “subversive science.” During the twentieth century, ecologists, nature writers, and practicing environmentalists have increasingly articulated an appreciation of nature and spoken of the need for changes in human lifestyles to protect it. “Green” political parties and activist conservation organizations such as
Greenpeace, EarthFirst!, and India’s Chipko movement have appeared throughout the world.

One well-developed environmental philosophy that supports this activism is known as deep ecology. Deep ecology builds on two basic premises: biocentric equality and self-realization. Biocentric equality expresses the intuition... that all things in the biosphere have an equal right to live and blossom and to reach their own individual forms of unfolding” (Devall and Sessions, 1985). Humans have a right to live and thrive, but so do the other organisms with whom we share the planet.

Self-realization describes a striving to grow, develop, and find fulfillment. Human beings are assumed to share this goal with the rest of nature. In the case of humans, deep ecologists believe that true self-realization involves spiritual growth and an expansion of knowledge and concern, rather than an increase in material wealth. Individual human self-realization should lead to a concern with all of nature: the preservation and development of the larger wholes to which we belong, including both human and natural communities. Deep ecologists articulate this position in opposition to a “dominant worldview” that makes human concerns paramount and views human happiness in materialistic terms (see the following table). Deep ecology’s idealism and its call to action make it an appealing philosophy for people concerned with protecting biodiversity.

B. Ecofeminism

Ecofeminism represents another recent, important development within environmental ethics. Ecofeminists see a connection between the domination of women and the domination of nature. They attempt to specify links between these two forms of domination, and advocate both environmental protection and full, equal rights for women.

Like deep ecologists, ecofeminists argue that how we conceptualize wild and human nature has important ethical consequences. Karen Warren argues that some conceptual frameworks are inherently oppressive, because they split the world up into “value dualisms, i.e., disjunctive pairs in which the disjuncts are seen as oppositional (rather than as complementary) and exclusive (rather than as inclusive), and which place higher value on one disjunct rather than the other” (Armstrong and Bozler, 1998). Thus men are privileged over women, human culture over nature, animals over plants, and reason over emotion. The inferior disjuncts tend to be lumped together: women may be seen as closer to nature, more emotional, or less rational. Such conceptual frameworks legitimize discriminatory treatment of those “others” who fall on the wrong side of these value dualisms.

Ecofeminists launch a three-pronged attack on such a “logic of domination.” First, they may deny that certain differences exist, that they are hard and fast differences, or that they are as extreme as they are portrayed. For example, most ecofeminists simply deny that women are less rational or more emotional than men. In like fashion, they may point up similarities between humans and the other animals and downplay our differences. Second, ecofeminists may deny that actual differences make some beings morally superior to others. Human beings are superior reasoners compared to frogs, but that does not mean that we are therefore morally superior, in the sense that our interests should always trump theirs. Third, ecofeminists may deny that moral superiority underwrites domination. Even if humans are morally superior to frogs, it may still be wrong to kill them for food or destroy the last members of a rare frog species to create a new subdivision. Perhaps the proper response to those inferior to us in abilities or moral value is care and restraint.

C. Humanism: Our Default Mode

Like deep ecologists and ecofeminists, conventional ethical philosophers, including Kantians, utilitarians, and theologically grounded ethicists, have begun to

<table>
<thead>
<tr>
<th>Dominant worldview</th>
<th>Deep ecology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans dominant over nature</td>
<td>Humans living in harmony with nature</td>
</tr>
<tr>
<td>Natural environment and species are resources for humans</td>
<td>All nature has intrinsic worth, regardless of human needs</td>
</tr>
<tr>
<td>A growing human population with a rising standard of living</td>
<td>A stable human population living simply</td>
</tr>
<tr>
<td>Earth’s resources are unlimited</td>
<td>Earth’s resources are limited and must be used carefully</td>
</tr>
<tr>
<td>Ever higher technology brings progress and solutions</td>
<td>Appropriate technology must be used with respect for the earth</td>
</tr>
<tr>
<td>Emphasis on material progress</td>
<td>Emphasis on spiritual and ethical progress</td>
</tr>
<tr>
<td>Strong central government</td>
<td>Local control, organized according to watersheds, bioregions, or other natural units</td>
</tr>
</tbody>
</table>
address environmental issues. Most philosophers who have considered these matters have found strong reasons to support environmental protection. As previously noted, biocentrists and anthropocentrists may agree on a wide range of measures to protect the environment, despite great philosophical differences.

Still, some philosophers and many members of the general public remain unconvinced of any moral imperative to protect biological diversity. Their position may be characterized as humanism, a philosophy committed to the following propositions:

1. Biological diversity exists for humans, has no value apart from humans, and need not exist apart from humans.
2. The transformation of wild nature into natural resources adds value to nature, since nature possesses value only in human use. Indeed, the ever more thorough transformation of wild nature allows increased human numbers to lead longer, happier, and better lives.
3. The creation of just societies filled with flourishing individuals is the highest achievement of which humans are capable. We should judge ourselves based on our technological, scientific, artistic, and social progress—not on whether we preserve nature.

Like the philosophies discussed earlier, humanism may be developed in a variety of ways. A humanist’s ideal society may be more or less egalitarian, wealthy, or stable, and more or less racially and culturally diverse. But humanists share a belief in the centrality of humans. They generally applaud increased human numbers, wealth, and technological power and the development of new arts and activities that flourish in highly artificial environments. For these reasons they see little to lament in the loss of biodiversity.

Humanism may be called humanity’s default mode, since current trends are moving us more and more in this direction, whether it is desirable or not. We are creating a world with much less biological diversity and this artificial world has come to seem normal, indeed natural, for many people. It is not, however, inevitable. Human beings can reconnect to nature and curb our activities that threaten it. Environmental ethics reminds us that we have many reasons to value and protect earth’s remaining biological diversity.

See Also the Following Articles

- Economic Value of Biodiversity: Overview
- Ethical Issues in Biodiversity Protection
- Literary Perspectives on Biodiversity
- Religious Traditions and Biodiversity
- Social and Cultural Factors
- Stewardship: Concept of
- Traditional Conservation Practices

Bibliography


ENVIRONMENTAL IMPACT, CONCEPT AND MEASUREMENT OF
Ellen W. Chu* and James R. Karr†
*Northwest Environment Watch and †University of Washington

I. Natural and Human Disturbance
II. History of Human Impacts
III. Biotic Impoverishment
IV. Root Causes of Human Impact
V. Measuring Environmental Impacts
VI. Recognizing and Managing Environmental Impacts
VII. Reclaiming Lost Connections

GLOSSARY

biological integrity Wholeness of a living system, including the capacity to sustain the full range of organisms and processes having evolved in a region.
biota Living things, in particular, the flora and fauna of a region.
biotic impoverishment Systematic reduction in the earth's capacity to support life.
environment Surroundings; the complex of physical, chemical, and biotic factors acting upon a living system and influencing its form and survival.
health A flourishing condition, well-being; capacity for self-renewal.
impact A forceful contact; a major effect of one thing on another.

ALL LIVING THINGS influence their environment. For the first time in the earth's history, however, the environmental impact of one species, Homo sapiens, is the principal agent of global change. Humans' most harmful impact is worldwide degradation of living systems—an impact that threatens humans' own life-support network. The 21st century's greatest challenge for scientists, decision makers, and citizens worldwide will be to understand and control human environmental impact and to protect the health and integrity of the biosphere.

I. NATURAL AND HUMAN DISTURBANCE

On the outermost rocks of scenic Big Sur's windswept headlands live rubbery palm tree look-alikes called sea palms (Postelsia palmaeformis). Graceful forests of these brown algae extend along the Pacific coast from central California northward to Alaska, always where the waves hit hardest. For organisms living from the high-tide line to the lower surf zone, wave action is life renewing, bringing in oxygen and nutrients and carrying away reproductive products to colonize other shores. Sea palms are able to colonize rocky open places that other intertidal organisms cannot; they thrive on the violence of the waves.

A few miles north of Big Sur, fields of artichokes,
brussels sprouts, and strawberries stretch to the seaward edge of the coastal plain. These boulderfields have been farmed for two centuries by successive waves of European settlers. The area's land vegetation is dominated by nonnative plants brought in with agriculture. Countless streams and rivulets carry seaward a legacy of eroded soil, fertilizers, and pesticides. Many small creeks are choked with algae, which flourish because of excess nitrogen and phosphorus from fertilizers and sewer outfalls; many plants and animals die or are otherwise harmed by the eroded soils and seeping chemicals. Even before farming and European settlement, however, humans were shaping the landscape: for thousands of years, native Costanoan people harvested shellfish from the intertidal zones and set fires to burn off woody shrubs and trees in favor of the grassland they preferred. Along this coast and others around the world, relentless wave action and human occupation are both driving forces of landscape change; both can be said to "disturb" the environments they influence. Both forces have an environmental impact, but the effects of the natural disturbance differ fundamentally from the effects of human-driven change. Over millennia, organisms evolve to contend with routine disturbances in their natural environments. Those that do not adapt go extinct. Those that survive are molded by evolution and biogeography to succeed in their environments, developing ways to coexist. 

II. HISTORY OF HUMAN IMPACTS

All organisms change their environment as they live, grow, and reproduce. Some organisms, like beavers or elephants, change their surroundings so dramatically that they have been called "ecosystem engineers." Beaver dams alter the flow of rivers, increase dissolved oxygen in downstream waters, create wetlands, and modify streamside zones. African elephants convert vast areas of wooded savannah to open grassland by toppling trees as they browse. Change brought about by living things, including the first humans, has been slow and incremental in evolutionary terms. Even ecosystem engineers evolve along with other inhabitants of their ecosystems, developing ways to coexist.

The human evolutionary line began about 7 million years ago in Africa. It took some 3 or 6 million years for protohumans to spread from there to Asia and then Europe. These early humans lived a hunting-gathering life, using a variety of stone tools to feed on easy-to-kill prey; they differed little from other primates trying to survive. But by 13,000 years ago, modern humans had become unlike any other species, having spread to all the continents and many islands across the globe.

Still, most human-environment interactions at this stage consisted of threats to human health and well-being that came from the environment, including injuries in fights or wars, periodic famine, vector-borne diseases, and accidents (Table I). Then, some 10,000 years ago, people began to domesticate plants and animals. Instead of searching for food, they began to produce food. Food production changed the course of human and environmental history. Domestication of plants and animals enabled people to adopt a sedentary lifestyle. People gathered food rather than searching for it; they began to live in crowded conditions, inadequate sanitation compounded the effects. From agriculture, civilization followed; and with it, cities, writing, engineering, and political empires.

In just 10,000 years, these developments led to 6 billion people on Earth, industrial societies, and a global economy founded on complicated technologies. Now the predominant human-environment interaction consists of threats to all four of the earth's environmental "spheres":

• Geosphere (lithosphere), the earth's crust and upper mantle, containing nonrenewable fossil fuels, minerals, and nutrients plants require. The activities of plants, animals, and microorganisms weather mineral soils and rocks, create organic soils, and alter erosion and sedimentation rates. Humans mine minerals, metals, and
TABLE I
Challenges to Human Health and Welfare over the Past 200,000 Years

<table>
<thead>
<tr>
<th>Years before 2000 A.D.</th>
<th>Major events</th>
<th>Human population</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>200,000</td>
<td>Appearance of anatomically modern humans</td>
<td>7 million</td>
<td>Accidents, Combat, Vector-borne infection, Periodic famine, Nutritional deficiencies, Contagious diseases, Local ecological disruptions</td>
</tr>
<tr>
<td>10,000</td>
<td>Emergence of agriculture, Rise of towns and cities</td>
<td>8 million</td>
<td>Chemical scars, Overnutrition, Global ecological disruptions</td>
</tr>
<tr>
<td>150</td>
<td>Industrial revolution</td>
<td>1.3 billion</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Modern affluence</td>
<td>3 billion</td>
<td></td>
</tr>
</tbody>
</table>


New challenges are listed in the era they began; all continue through the present.

• Atmosphere, the thin envelope of gases that encircles the planet. Living systems maintain the earth’s atmosphere, its temperature, and the amount of water it contains by continually generating oxygen through photosynthesis, holding down atmospheric carbon dioxide, and changing the amount and forms of other gases. Humans release toxic chemicals into the air and alter the climate by raising the atmospheric concentration of greenhouse gases such as carbon dioxide and methane through the burning of fossil fuels in motor vehicles, electricity-generating plants, and airplanes.

• Hydrosphere, the earth’s liquid surface and underground water; its polar ice caps, oceanic icebergs, and terrestrial permafrost; and its atmospheric water vapor. Living systems alter the water cycle by modifying Earth’s temperature and the amount of water plants send into the atmosphere in a process called evapotranspiration. Humans build dams, irrigation canals, drinking-water delivery mechanisms, and wastewater treatment systems. They use water to generate electricity; they mine groundwater for farming and drinking; they alter the flows of surface waters for everything from transportation to gold mining; they drain wetlands to gain land area and abate waterborne diseases. Modern human interference in global climate is likely to alter the entire planetary water cycle.

• Biosphere, the earth’s living systems, which occupy a thin layer from the deepest oceans to the top of the highest mountains. Life on Earth emerged 3.9 billion years ago and has sustained itself through changes in form, diversity, and detail since then. No other known planet supports life. Like other predators, humans affect the populations of their prey, on a small scale millions of years ago and on much larger scales since. Thousands of years ago, humans hunted the large mammals of Australia and the Americas to the point where many became extinct. Humans have domesticated plants through agriculture and massively redistributed vegetation through cutting of forests, burning of grasslands, or land uses creating vast areas of desert. Human actions today have precipitated a spasm of extinctions rivaling five previous mass extinctions caused by astronomical or geological forces; each of these eliminated more than 70% of species then existing.

Yet despite today’s world-altering technologies and computer-generated virtual reality, humans are as dependent on their environments as other organisms are. History, not just ecology, is very clear on this point. From the Old Kingdom of Egypt more than 4000 years ago to the culture that created the huge stone monoliths on Easter Island between 500 and 1550 A.D. to the 1930s dust bowl of North America, civilizations or ways of life have prospered and failed by using and (mostly unwittingly) abusing natural resources.

In Old Egypt, the resource was the valley of the Nile,
The first step in dealing with the present impact of human activity is to correctly identify the nature of humanity's relationship with the environment and how human actions affect that relationship. Many people still see the environment as something people must overcome, or they regard environmental "needs" as something that ought to be balanced against human needs (for example, jobs versus the environment). Most people still see humans' relationship with the environment as a one-way street: the environment provides us with goods and services, and our job is to keep them这位us and services coming.

Thus, when asked to name humanity's principal environmental challenges, people usually think of running out of nonrenewable raw materials and energy, or about water and air pollution. Our environmental research and development institutions focus on ways technology can help solve each problem, such as fuel cells to provide clean, potentially renewable energy or scrubbers to curb smogstack pollution. Even when people worry about "biodiversity loss," they are primarily concerned with stopping species extinctions rather than with understanding the underlying losses leading up to species extinctions or the broader biological crisis that extinctions signal.

This perspective misses a crucial point: humanity has a two-way relationship with the biosphere. The reason pollution, energy use, extinction, and dozens of other human impacts are important is their larger impact on the biosphere. At this point in the planet's history, one species is compromising the earth's ability to support the living systems that evolved here over millions of years.

The systematic reduction in the earth's capacity to support life, termed biotic impoverishment, is thus the most important human-caused environmental impact. At best, the ethics of this impact are questionable; at worst, we are jeopardizing our own survival.

The connection between extinction and biotic impoverishment is intuitively obvious. By overharvesting fish, overcutting forests, overgrazing grasslands, or paving over land for cities, humans are clearly killing other organisms outright or eliminating their habitats, thereby driving species to extinction. But biotic impoverishment takes many forms besides extinction. It encompasses three categories of human impacts on the biosphere: (a) indirect depletion of living systems through alterations in physical and chemical environ-
The Many Faces of Biotic Impoverishment

1. Degradation of water (redistributed flows, depletion of surface and ground water, wetland drainage, organic enrichment, destruction and alteration of aquatic life.)
2. Soil depletion (destruction of soil structure, erosion, salinization, desertification, nutrient leaching, destruction and alteration of soil biota)
3. Chemical contamination (land, air, and water pollution from pesticides, herbicides, heavy metals, and toxic synthetic chemicals and atmospheric ozone depletion; kills, extinctions, biodiversity loss, bioaccumulation, hormone disruption, immunological deficiencies, reproductive and developmental anomalies, respiratory disease, intergenerational effects)
4. Altered biogeochemical cycles (alteration of the water cycle, nutrients enrichment, acid rain, fossil fuel emissions, particulate pollution, degradation of land and water biota and outbreaks of pests, pathogens, and red tides)
5. Global climate change (rising greenhouse gases, altered precipitation and airflow patterns, rising temperatures, effects on individual and community health and shifts among and within global ecosystems)

Direct depletion of nonhuman life

1. Overharvest of renewable resources such as fish and timber (depleted populations, extinctions, altered food webs)
2. Habitat fragmentation and loss (extinctions, biotic homogenization, emerging and reemerging pests and pathogens)
3. Biotic homogenization (extinctions and invasions)
4. Genetic engineering (homogenization of crops, antibiotic resistance, potential extinctions and invasions if genes escape, other unknown ecological effects)

Direct degradation of human life

1. Epidemics and emerging diseases (occupational hazards, asthma and other respiratory ills, pandemics, plagues, HIV/AIDS, malaria, tuberculosis, Lyme disease, antibiotic resistance, diseases of overnutrition, higher human death rates)
2. Loss of cultural diversity (genocide, ethnic cleansing, loss of cultural and linguistic diversity, loss of knowledge)
3. Reduced quality of life (environmental refugees, malnutrition and starvation, failure to thrive, poverty)
4. Environmental injustice (environmental discrimination and racism; economic exploitation; growing gaps between rich and poor individuals, segments of society, and nations; gender inequities; trampling of the environmental and economic rights of future generations)
5. Political instability (resource wars, civil violence especially under intransigent regimes, international terrorism, increased number of environmental refugees)
6. Cumulative effects (environmental surprises, increased frequency of “natural” catastrophes, “boom and bust” cycles, collapse of civilizations)

dry up wetlands, lakes, and inland seas. Our demands are outrunning supplies of this nonrenewable resource, and the scale of our transformations risks altering the planet’s water cycle.

Physical alterations to the planet’s waters, combined with massive industrial and residential pollution, have taken a heavy toll on nonhuman aquatic life. Ten percent of the world oceans’ coral reefs have died from pollution and other causes. Globally, the number of so-called oceanic dead zones, where there is little or no dissolved oxygen, tripled in the last 30 years of the 20th century. The biota of freshwater systems fares no better. A 4-year survey of the freshwater fishes inhabiting Malaysian rivers in the late 1980s found only 46% of the 266 known Malaysian species. Nearly 40% of North America’s freshwater fishes are at risk of extinction; two-thirds of freshwater mussels and crayfishes and one-third of amphibians that depend on aquatic habitats in the United States are rare or imperiled. Wetlands worldwide continue to disappear; the lower 48 United States lost 53% of its wetlands between the 1700s and mid-1980s. Such losses destroy major fish and shellfish nurseries, natural flood and pollution control, and habitat for countless plants and animals.

Humans now use 54% of the accessible water runoff on Earth, a figure that is likely to grow to 70% by 2025. By then, more than a third of the world’s population could suffer shortages of fresh water for drinking and irrigation. Groundwater aquifers in many of the world’s most important crop-producing regions are being drained faster than they can be replenished. Natural flood regimes, as in the Nile River basin, no longer spread nutrient-rich silt across floodplains to nourish agriculture; indeed, the High Dam at Aswan traps so much silt behind it that the Nile delta, essential to Egypt’s modern economy, is falling into the sea. In the past 30 years, the number of large dams on the world’s rivers grew more than seven times, to some 38,000 today. The huge Three Gorges Dam across China’s Yangtze River promises to create a 690-km serpentine lake behind it and displace 1.2 million Chinese citizens before its scheduled completion in 2009; the impact on the river’s fertile delta farming region and downstream estuaries has just begun to be assessed. Whole inland seas, such as the Aral Sea in Uzbekistan, are drying up because the streams feeding them have so little water. In addition to eliminating habitat for resident organisms, the seas’ drying is bringing diseases to surrounding human populations. Diseases caused by waterborne pathogens are making a comeback even in industrialized nations.

2. Soil Depletion

Hardly just “dirt,” soil is a living underground system that makes it possible for raw elements from air, water, and bedrock to be physically and chemically assembled, and reassembled with the aid of living macro- and microorganisms into the thin green shell of life above ground. Accumulated over thousands of years, soil cannot be renewed in any time frame useful to humans alive today, or even to their great-grandchildren. Humans degrade soils when they compact it, erode it, disrupt its organic and inorganic structure, turn it too salty for life, and cause desertification. Urbanization, logging, mining, overgrazing, altering soil moisture, air pollution, fires, chemical pollution, and leaching out of minerals all damage soils. Thanks to removal of vegetative cover, mining, agriculture, and other activities, the world’s topsoils are eroded by wind and water ten to hundreds of times faster than they are renewed (at roughly 1 ton per hectare per year). Soils constitute the foundation of human agriculture, yet agriculture, including livestock raising, is the worst culprit in degrading soils. Agricultural practices have eroded or degraded an area equal to 38% of present cropland. Besides topsoil erosion, the damage includes salting and water-logging of poorly managed irrigated lands, compaction by heavy machinery and the hooves of livestock, and pollution from excessive fertilizers, animal wastes, and pesticides.

Living, dead, and decomposing organic matter is the key to soil structure and fertility. Soil depleted of organic matter is less permeable to water and air and thus less able to support either aboveground plants or oxygen-dependent microbes. It’s the linkages between soil’s inorganic components and the soil biota (naturalist E. O. Wilson’s “little things that run the world”) that give soil its life-sustaining capacity. A clear-cut forest patch whose soil biota has been damaged beyond recovery can no longer sustain trees, no matter how many are planted, another clear-cut patch whose soil community is still functioning will support new tree growth. Destroying soil biota exerts a whole series of impoverishing biotic effects both below and above ground.

3. Chemical Contamination

In 1962 Rachel Carson’s landmark book Silent Spring alerted the world to the pervasiveness of synthetic chemicals produced since World War II. Some 75,000 synthetic chemicals are in use today. True to one company’s slogan, many of these have brought “better living
through chemistry,” providing new fabrics and lighter manufacturing materials, antibiotics, and life-saving drugs. But industrial nations have carelessly pumped chemicals into every medium. Chemicals directly poison organisms; they accumulate in physical surroundings and are passed up the food chain. Many chemicals travel the ocean and atmospheric currents to sites far from their source: sulfur emissions from the U.S. Midwest, for example, fall to earth again as acid rain in Europe, killing forests and so acidifying streams and lakes that they too effectively die. China’s burning of soft coal sends air pollution all the way to northwestern North America; the heavy haze hanging over China’s chief farming regions may be cutting agricultural production by as much as a third. Chlorofluorocarbons (CFCs), once widely used as refrigerants, have damaged the atmospheric ozone layer, which moderates how much ultraviolet radiation reaches the earth, and opened ozone “holes” over the Arctic and Antarctic. Chemicals ranging from prescription drugs flowing out of sewage plants to pesticides to heavy metals to cancer-causing by-products of countless manufacturing processes now lace the world’s water, soil, and air and the bodies of all living things, including humans. Carson’s book revealed the real danger of chemical pollution: we have not simply altered the chemistry of water, soil, and air but harmed the biosphere as well.

Although many of the most obviously deadly compounds were banned in the 1970s, they continue to impoverish the biosphere. Polychlorinated biphenyls (PCBs)—stable, nonflammable compounds once used in electrical transformers and many other industrial and household applications—accumulate and persist in soils and the food web. They are found in polar bears and arctic villages; they are implicated in reproductive disorders, particularly in such animals as marine mammals, whose long lives, thick fat layers where chemicals concentrate, and position as top predators make them especially vulnerable. The agricultural pesticide DDT, sprayed with abandon in the 1940s and 1950s, even directly on children, severely thinned wild birds’ eggshells. Populations of birds such as the brown pelican and bald eagle dropped precipitously by the 1970s but began to recover in the last decade of the 20th century. Chemicals cause cancer, disrupt hormonal systems, provoke asthma, and impair the functioning of immune systems. They have intergenerational effects, such as intellectual impairment in children whose mothers have eaten contaminated fish. What’s more, a half century of pesticide and antibiotic overuse has bred resistance to these chemicals among insects, plants, and microbes, prompting fears of new and reemerging scourges. The litany of chemicals’ effects on living things is so long that chemical pollution is humans’ environmental impact in most people’s minds; yet it is just one form of biotic impoverishment.

4. Altered Biogeochemical Cycles

All the substances found in living things, such as water, carbon, nitrogen, phosphorus, and sulfur, cycle through ecosystems in biogeochemical cycles. Human activities alter or have the potential to alter all these cycles. Sometimes the results stem from changing the amount or the precise chemistry of the cycled substance; in other cases, humans alter biogeochemical cycles by changing the biota itself.

Freshwater use, dams, and other engineering feats change the amount and rate of river flow to the oceans and increase evaporation rates, directly affecting the water cycle and indirectly impoverishing aquatic life. Direct human modifications of living systems also perturb the water cycle. In the tall “cloud forests” of redwoods in northern California or Douglas firs in the Pacific Northwest, the trees gather more moisture directly from the clouds than falls as rain. Among the effects of logging in such forests is a depletion of stream flows, which shifts the water cycle. In South Africa, European settlers supplemented the treeless native scrub, or fynbos, with such trees as pines and Australian acacias from similar Mediterranean climates. Because these trees are larger and thirstier than the native scrub, regional water tables have fallen sharply. Human activity has altered the global nitrogen cycle by greatly increasing the amount of nitrogen fixed from the atmosphere (combined into compounds usable by living things). The increase comes mostly from deliberate addition of nitrogen to soils as fertilizer but also as a by-product of the burning of fossil fuels. Agriculture, livestock raising, and individual yard maintenance chronically add tons of excess nutrients, including nitrogen and phosphorus, to soils and water. The additions are often invisible; their biological impacts are often dramatic. Increased nutrients in coastal waters, for example, trigger blooms of toxic dinoflagellates, the algae that cause red tides, fish kills, and tumors and other diseases in varied sea creatures. When huge blooms of algae die, they fall to the seafloor, where their decomposition so robs the water of oxygen that fish and other marine organisms can no longer live there. With nitrogen concentrations in the Mississippi River two to three times as high as they were 30 years ago, a gigantic dead zone forms in the Gulf of Mexico.
every summer; it covered a record 20,020 square kilometers in the summer of 1999. The burning of fossil fuels is radically altering the carbon cycle, primarily by greatly increasing the atmospheric concentration of carbon dioxide. With other greenhouse gases, such as methane and oxides of nitrogen, carbon dioxide helps keep the earth's surface at a livable temperature and drives plant photosynthesis, but since the industrial revolution, atmospheric carbon dioxide concentrations have risen 30% and are now widely thought to be disrupting the planet’s climate.

5. Global Climate Change

The Intergovernmental Panel on Climate Change (IPCC), comprising 2300 of the world’s preeminent atmospheric scientists, had concluded by 1995 that humans have caused at least part of the atmospheric warming observed globally since the beginning of the 20th century. The concentrations of heat-trapping gases in the atmosphere are at their highest level in more than 200,000 years. The 20th century is the Northern Hemisphere has been described as the warmest of the past millennium; the 1990s ranked as that century’s warmest decade.

Higher global temperatures precipitate a whole series of effects, making the study of climate change, and of humans’ role in it, complex and controversial. Spring now arrives one week earlier in the Northern Hemisphere than in the past. Each year’s growing season starts earlier, and seasonal temperature patterns are also shifting, making the weather seem more unpredictable than ever. Unusually warm or cold winters, massive hurricanes such as those that devastated North Carolina in late summer 1999, and weather-related damage to human life and property are all predicted to increase with global warming. Weather-related damage totaled $92 billion in 1998, exceeding the total for the entire decade of the 1980s, even after adjusting for inflation.

Rising carbon dioxide concentrations may enhance plant photosynthesis and growth where other nutrients are not limiting. Rising temperatures may shift the ranges of many plants and animals, shifting the composition and distribution of the world’s biomes. The resulting disruptions to habitats will have far-reaching implications not only for the displaced plants and animals but also for the goods and services humans depend on from living systems.

B. Direct Depletion of Nonhuman Life

From their beginnings as hunter-gatherers, humans have become highly efficient, machine-aided ecosystem engineers and predators. We transform the land so it produces what we need or want; we harvest the oceans in addition to reaping our own fields; we cover the land, even agricultural land, with sprawling cities. All these activities directly affect the ability of other life-forms to survive and reproduce. We deplete nonhuman life by eliminating some forms and favoring others; the result is a loss of genetic, population, and species diversity. We are irreversibly homogenizing life on Earth, in effect exercising an ‘unnatural selection’ that is erasing the diversity generated by millions of years of evolution by natural selection. One species is now determining which other species will survive, reproduce, and thereby contribute the raw material for future evolution.

1. Overharvest of Renewable Resources

In the 1930s, so many sardines were scooped from the waters off Monterey’s Cannery Row that the population collapsed, taking other sea creatures and human livelihoods with it; the species has still not recovered fully. According to the U.S. National Marine Fisheries Service, nearly 80% of commercially valuable fish of known status were overfished or ‘fished to their full potential’ by 1993. Atlantic commercial fish species at their lowest levels in history include runas, marlins, cod, and swordfish. Overfishing not only depletes the target species but changes the whole structure of marine food webs.

Marine mammals, including whales, seals, sea lions, manatees, and sea otters, were so badly depleted by human hunters that one species, Steller’s sea cow (Hydrodamalis gigas), went extinct; many other species almost disappeared. In the 19th century, Russian fur traders wiped out sea otters (Enhydra lutris) along the central California coast; with the otters gone, their principal prey, purple sea urchins (Stronglyocentrotus purpuratus) overran the offshore forests of giant kelp (Macrocystis pyrifera), decimating the kelp fronds and the habitat they provided for countless other marine creatures, including commercially harvested fishes. Thanks to three decades of protection, most marine mammal populations are slowly rebounding—only to face food shortages following disruptions in regional marine food webs due to fishing, changing oceanic conditions, and contamination.

Timber harvest has stripped vegetation from the Amazon rain forests to mountainsides on all continents, diminishing and fragmenting habitat for innumerable
forest and stream organisms, eroding soils, worsening floods, and contributing significantly to global carbon dioxide emissions. In the Northern Hemisphere, only 10% remains of old-growth temperate rain forests. The uniform stands of trees usually replanted after logging do not replace the diversity lost with the native forest, any more than monocultures of corn replace the diversity within native tallgrass prairies.

2. Habitat Fragmentation and Loss

Human land use inevitably damages or destroys habitat for some living thing. But satellite-mounted remote-sensing instruments over the past two decades reveal transformations of a scale unimaginable in centuries past. At one extreme, row-crop agriculture or urban-industrial areas occupy only 10 to 15% of the earth's land surface; pastureland occupies another 6 to 8%. At the other extreme, every alteration of the land alters the physical, chemical, and biological surroundings of anything living there. Estimates of the share of land wholly transformed or degraded by humans fall at around 50%. Landscapes that have not been entirely converted to human use have been cut into fragments. In Song of the Dodo, writer David Quammen likens our actions to starting with a fine Persian carpet and then slicing it nearly into 56 equal pieces; we still have the same square footage but not 36 nice Persian rugs, only ragged, nonfunctional fragments.

Our roads, farms, cities, feedlots, and ranches either destroy or fragment the habitats of most large carnivorous mammals. Mining and oil drilling damage the soil, remove vegetation, and pollute marine areas. Grazing compacts soil and sends silt and manure into streams, remove vegetation, and pollute marine areas. Grazing compacts soil and sends silt and manure into streams, where they harm stream life. Habitat damage may lead directly to extinction, or it may isolate organisms in ways that make them extremely vulnerable to natural disturbances, climate change, or further human disturbance.

3. Biotic Homogenization

"The one process now going on that will take millions of years to correct," writes E. O. Wilson, "is the loss of genetic and species diversity by the destruction of natural habitats. This is the folly our descendants are least likely to forgive us." Both deliberately and unwittingly, humans are rearranging the earth's living components, reducing diversity and homogenizing biotas around the world. The present, continuing loss of genetic diversity, of populations, and of species vastly exceeds background rates. At the same time, our global economy is transporting species worldwide at unprecedented scales.

The globe is now experiencing its sixth mass extinction, the largest since the fall of the dinosaurs 65 million years ago; present extinction rates are thought to be on the order of 100 to 1000 times those before people dominated the earth. Many of the world's bird species have been driven to extinction by humans, especially on oceanic islands. According to a survey by the World Conservation Union, at least one of every eight plant species is threatened with extinction. Although mammals and birds receive the most attention, massive extinctions of plants, which form the basis of the biosphere's food webs, undermine life-support foundations. Some scientists are calling extinction "the worst biological tragedy," but extinction too is but another symptom of global biotic impoverishment.

Ever since they began to spread over the globe, people have transported other organisms with them, sometimes for food, sometimes for aesthetic reasons, and most often inadvertently. With the mobility of modern societies and today's especially speedy globalization of trade, the introduction of alien species has reached epidemic proportions, causing some scientists to label it "biological pollution." Aliens—zebra mussels (Dreissena polymorpha) and tamarisks, or saltcedar (Tamarix spp.), in North America; the Red Sea jellyfish Rhopilema nomadica and the common aquarium alga Caulerpa taxifolia now choking the Mediterranean Sea; and Leidy's comb jelly (Mnemiopsis leidyi) of northeastern America in the Black Sea, to name just a few—are present everywhere, and they usually thrive and spread at the expense of native species. On many islands, for example, more than half the plant species are not native, and in many continental areas, the figure reaches 20% or more. Such alien invasions cause extinctions and, when added to other extinctions and the deliberate monocultures of agricultural crops, worsen biotic homogenization. Introduced species are fast catching up with habitat loss and fragmentation as the major engines of ecological deterioration.

4. Genetic Engineering

Humans have been manipulating their crop plants and domesticated animals for 10,000 years or so—selecting seeds or individuals, breeding, and cross-breeding. The goal was something better, bigger, tastier, harder, or all of the above; success was sometimes elusive, but the result was crop homogenization. Of the myriad strains of potatoes domesticated by South American cultures, for example, only one was accepted and cultivated when potatoes first made it to Europe. The new crop made it possible to feed more people from an equivalent area of land and initially staved off malnutrition. But the...
strain succumbed to a fungal potato blight in the 1880s; had more than one strain been cultivated, the tragic Irish potato famines might have been averted.

In the last few decades of the 20th century, people began to manipulate genes directly using the tools of molecular biotechnology, even cloning sheep and cows from adult body cells. U.S. farmers routinely plant their fields with corn whose genetic material incorporates a bacterial gene resistant to certain pathogens. More than 40 genetically altered crops have been approved for sale to U.S. farmers since 1992, with genes borrowed from bacteria, viruses, and insects; worldwide, upward of 40 million hectares are planted with genetically modified crops. Biotechnologists see this new-millennium green revolution as finally solving world hunger. But other scientists worry about unknown human and ecological health risks; these concerns have stirred a deep scientific debate, akin to the debate over pesticides in Rachel Carson’s time.

One worrisome practice is plant genetic engineers’ technique of attaching the genes they want to introduce into plants to an antibiotic-resistance gene. They can then easily select those plants that have acquired the desired genes by treating them with the antibiotic, which kills any nonresistant plants. Critics worry that the antibiotic-resistant genes could spread to human pathogens and worsen an already growing antibiotic-resistance problem. Another threat comes from allergies humans might have or develop in response to genetically modified foods.

Supporters of genetic engineering believe that genetically altered crops pose few ecological risks. But studies in the late 1990s indicated that pollen from genetically engineered “Bt corn” can kill monarch butterfly caterpillars. It is a strain of bacteria that has been used since the 1980s as a pesticidal spray; its genes have also been inserted directly into corn and other crops. Studies in radishes, sorghum, canola, and sunflowers found that genes from an engineered plant can jump to wild relatives through interbreeding. The fear is that a gene confering insect or herbicide resistance might spread through wild plants, creating invasive superweeds, which could potentially lower crop yields and further disrupt natural ecosystems.

Great economic and environmental benefits are claimed for the products of biotechnology: higher yields, lower synthetic pesticide and herbicide use, increased farmer profits. But so far the benefits have not always materialized, and growing consumer and scientific concern may slow the spread of genetically modified crops, at least in the developed world.

C. Direct Degradation of Human Life

Human biotic impacts are not confined to other species; human cultures themselves have suffered from the cascading indirect and direct impacts humans exert on the rest of nature. Over the past hundred years, human technology has been a two-edged sword with regard to public health. Wonder drugs controlled common pathogens at the same time that natural selection strengthened those pathogens’ ability to resist the drugs. Reservoirs in the tropics made water supplies more reliable for humans but also created ideal environments for human parasites. Industrialization exposed human society to a remarkable array of toxic substances.

Though “man’s inhumanity to man” has been both fact and the subject of discourse for thousands of years, the discussions have mostly been removed from any environmental context. Few people today regard social ills as “environmental impacts” or humans as part of a “biota.” But diminished societal well-being—whether manifest in high death rates or poor quality of life—shares many of its roots with diminished nonhuman life as a form of biotic impoverishment.

1. Epidemics and Emerging Diseases

The interface between the environment and human health is the core of the discipline known as environmental health. Among the environmental challenges to public health are the direct effects of toxic chemicals; occupational health threats, including exposures to hazardous materials on the job; and sanitation and hazardous waste disposal. Exploitation of nonrenewable natural resources—including coal mining (causing black lung disease), rock quarrying or other mining operations (causing silicosis), and petroleum extraction and refining—often chronically impairs workers’ health and shortens their lives. Farmworkers around the world suffer long-term ills from high exposures to pesticides and herbicides. Partly because of increased air pollution, asthma rates are rising, particularly in big cities. Synthetic volatile solvents are used in products from shoes to semiconductors, producing lung diseases and toxic wastes. Nuclear weapons production starring in World War II, and the associated contamination, have been linked to a variety of diseases.

Infectious diseases have challenged human populations throughout history, playing a significant role in their evolution and cultural development. The 20th century brought major successes in eradicating such infectious diseases as smallpox, polio, and many waterborne illnesses. But toward the century’s end, emerging
and reemerging diseases were again reaching pandemic proportions. Human migrations have been the main source of epidemics throughout history, but environmental and social changes resulting from human activities are accelerating pathogen traffic. Infectious diseases thought to be on the wane—including tuberculosis, malaria, cholera, diphtheria, leptospirosis, encephalitis, and dengue fever—began a resurgence. In addition, seemingly new scourges—Ebola virus, hantavirus, HIV/AIDS, legionnaires’ disease, and Lyme disease—are also spreading, often, it appears, from animal hosts to humans as people encroach further upon previously undisturbed regions.

Especially in highly developed countries such as the United States, diseases of affluence and overconsumption are also taking a toll. Heart disease is the number one cause of death in the United States; overnutrition and obesity due to sedentary, technology-driven lifestyles, particularly among children, is chronic and rising. One estimate put the share of U.S. children considered overweight or obese at one in four.

2. Loss of Cultural Diversity

Although not conventionally regarded as elements of biodiversity, human languages, customs, agricultural systems, technologies, and political systems have evolved out of specific regional environments. Like other organisms’ adaptive traits and behaviors, these elements of human culture constitute unique natural histories adapted, like any natural history, to the biogeographical context in which they arose. Yet unique indigenous human cultures, including knowledge of local medicines and geographically specialized economies, are disappearing even more rapidly than the natural systems that nurtured them.

Modern technology, transportation, and trade are moving the world toward a globalized culture, thereby reducing human biological and cultural diversity. Linguists, for example, are predicting that at least half of the modern world’s 6000 languages will become extinct in the twenty-first century. Indigenous knowledge and practices are dwindling fast with the spread of Euro-American culture. This loss of human biodiversity is in every way as troubling as the loss of nonhuman biodiversity.

3. Reduced Quality of Life

Human activities that degrade environmental conditions threaten not only the biosphere but also humans’ own quality of life. Many sectors of society now recognize this link. In 1990 the United Steelworkers of America endorsed a report saying “We cannot protect steelworkers jobs by ignoring environmental problems” and “the greatest threat to our children’s future may lie in the destruction of their environment.” Independent of race or economic class, declining quality of life is manifest in both the increased incidence of asthma in the United States caused by environmental contaminants and the high disease rates in the former Soviet Bloc after decades of unregulated pollution. Even with explicit legal requirements that industries release information on their toxic emissions, many people throughout the world still lack both information and the decision-making power that would give them any control over the quality of their lives.

Food availability, which depends on environmental conditions, is a basic determinant of quality of life. Yet according to the World Health Organization, roughly half the population in all nations suffers from two forms of poor nutrition: undernutrition and overnutrition. A swollen abdomen is now a symptom shared by malnourished children, who lack calories and protein, and overweight residents of the developed world, who suffer clogged arteries and heart disease because of eating too much food.

Deeply concerned about the degraded environment and resulting quality of life in his homeland, Ogoni activist Ken Saro-Wiwa stated, shortly before he was executed by the Nigerian government in 1995, “The environment is man’s first right. Without a safe environment, man cannot exist to claim other rights, be they political, social, or economic.” Environmental degradation and desperate living conditions have spawned a new kind of refugee: the number of international environmental refugees exceeded the number of political refugees around the world for the first time in 1999. Environmental refugees—the homeland of which is lost forever—were flooded from dam building, extraction of mineral resources, desertification, and unjust policies of national and international institutions. Such degradation preempts many fundamental human rights, including the rights to health, livelihood, culture, privacy, and property.

4. Environmental Injustice

Making a living from nature’s wealth has consistently opened gaps between haves and have-nots, between those who bear the brunt of environmental damage to their home places and those who do not, and between the rights of people alive now and those of future generations; these disparities too are part of biotic impoverishment. Inequitable access to “man’s first right”—a
healthy local environment—has come to be known as environmental injustice.

Environmental injustices, such as institutional racism, occur in industrial and nonindustrial nations. Injustice can be overt, as when land-use planning sites landfill, incinerators, and hazardous waste facilities in minority communities, or when environmental agencies levy lower fines for hazardous waste violations that are lower in minority communities than in white communities. Less overt, but no less unjust, is the harm done to one community when unsound environmental practices benefit another, as when clear-cut logging in the highlands of northwestern North America benefits logging communities while damaging the livelihoods of lowland fishing communities touched by debris flows, sedimentation, and downstream flooding.

The plight of the working poor and the disparities between rich and poor are also examples of biotic impoverishment within the human community. According to the United Nations Research Institute for Social Development, the collective wealth of the world’s 358 billionaires equaled the combined income of the poorest 2.4 billion people in 1994. In the United States during the last decade of the twentieth century, the incomes of poor and middle-class families stagnated or fell, despite a booming stock market. The Center on Budget and Policy Priorities and the Economic Policy Institute pointed out that, between 1988 and 1998, earnings of the poorest fifth of American families rose less than 1%, while earnings of the richest fifth jumped 13%.

Despite great advances in the welfare of women and children over the past century, poverty still plagues both groups. Children from impoverished communities, even in affluent nations, suffer from the lethargy and impaired physical and intellectual development known as failure to thrive. Poverty forces many children to work the land or in industrial sweatshops; lack of education prevents them from attaining their intellectual potential. This impoverishment in the lives of women and children is as much a symptom of biotic impoverishment as are deforestation, invasive alien organisms, or species extinctions.

The creation of a “third world” to supply raw materials and labor to the dominant European civilization after 1300—and the resulting schism between today’s developed and developing nations—is perhaps the grossest example of human and environmental domination leading to continued injustice. Exploitation of poor countries and their citizens by richer, consumer nations (as well as by the governments of developing nations themselves) persists today in agriculture, wild materials harvesting, and textile and other manufacturing sweatshops. In the mid-1990s, industrial countries consumed 86% of the globe’s aluminum, 81% of its paper, 80% of its iron and steel, 75% of its energy, and 61% of its meat; they are thus responsible for most of the environmental degradation associated with producing these goods. Yet most of the actual degradation takes place in poorer, developing countries.

Ironically, local environmental injustice often arises in the name of national development. Locals resist projects that seem socially disruptive or ecologically unsustainable while governments push them as potential boons to the national economy. China’s Three Gorges Dam presents one example; another is the governments-backed exploitation, fought by Saro-Wiwa, of Ogoniland’s oil reserves in southern Nigeria by the Shell Petroleum Development Corporation. After Saro-Wiwa’s execution, the Ogoni were left, without a voice, to deal with a scarred and oil-polluted homeland.

Governments of developing nations are starting to demand a share of profits made from biological resources within their borders, including pharmaceuticals from plants and animals or new strains bred from local crop varieties. Nature’s “free wealth” is now commonly viewed as a “genetic resource,” the raw material for the biotechnology age like fossil fuels or minerals for the industrial age. Still, profits tend to flow to the elite, wealthy classes of both developing and developed nations while indigenous groups remain marginalized.

Little by little, so-called community-based conservation and development initiatives are being mounted by local citizens worldwide to combat these injustices. And one might contend that all efforts to protect the biosphere constitute a fight for the rights of future generations to an environment that can support them.

5. Political Instability

It was only during the last two decades of the 20th century that environmental issues found a place on international diplomatic agendas, as governments began to see threats to political security in an environmentally destabilized world. The 1987 Montreal Protocol on Substances That Deplete the Ozone Layer, which forced countries to curb their use of CFCs and other ozone-destroying chemicals, was an example of governments, nongovernmental organizations, and industry successfully working together to safeguard part of the environmental commons. More often, diplomacy stalled in conflicts over natural resources: arguments over water rights held up Israeli-Palestinian peace agreements; fights over fish erupted between Canada and the United States, Spain, and Portugal.
Scholars, including Norman Myers and Thomas Homer-Dixon, began calling attention to the links among growing populations, environmental scarcity and degradation; environmental injustice; and civil violence, including urban unrest. Myers noted that we cannot dispatch the military to turn back the advancing deserts, rising seas, or the greenhouse effect. Homer-Dixon and others pointed out that nations with high population growth, rising urbanization and unemployment, and depleted resources are becoming increasingly hard to govern except by tyrannical means. Perversely, the laudable goals of democracy and increased human rights can lead to civil riots in such nations as India, Pakistan, Indonesia, and China. Environmental and economic injustices worldwide leave no nation immune to this type of threat.

6. Cumulative Effects
If scientists have learned anything about the factors leading to biotic impoverishment, they have learned that the factors' cumulative effects can take on surprising dimensions. The multiple stresses of global climatic cycles such as El Niño–Southern Oscillation, natural disasters like droughts or floods, biotic impoverishment, and political instability have changed the course of history. Civilizations as far-flung as ancient Egypt, Peru, Easter Island, and the American Southwest prospered and collapsed because of unwise management of their environments. The city of Ubar, built on desert sands, literally disappeared into the sinkhole created by drawing too much water out of its great well. In modern Sahelian Africa, a combination of well digging and improved medical care and sanitation led to a threefold population increase; sedentary ways, heavy taxes imposed by a colonial government, and an impoverished people took the place of a nomadic culture evolved within the desert's realities.

During the last decade of the 20th century, numerous natural disasters befell nations around the world: massive flooding along the Mississippi River; devastating hurricanes in the Caribbean and southeastern United States; catastrophic landslides and floods in Honduras, China, and Venezuela; and destructive earthquakes in Turkey and Taiwan. Neither the rains nor the earthquakes were caused principally by human activity, but the cumulative effects of human land uses and management practices—from dikes separating the Mississippi from its floodplain to hog and poultry farms in North Carolina to clear-cut logging in Honduras to crowded cities in Turkey and Taiwan—made the disasters much worse in loss of human life and property.

IV. ROOT CAUSES OF HUMAN IMPACT
The ultimate cause of humans' massive environmental impact is our reproductive and consumptive behavior, which has given us spectacular success as a species. The very things that enabled humans to thrive in nearly every environment have magnified our impacts on those environments, and the technological and political steps we take to "mitigate" our impacts often worsen them. There are simply too many of us, we take too much from the natural world, and we ask it to absorb too much waste.

A. Fragmented Worldviews, Fragmented Worlds
For most of human history, people remained tied to their natural surroundings. Even as agriculture, writing, and technology advanced, barriers of geography, language, and culture kept humans a diverse lot, each group depending on mostly local and regional resources for survival. Their worldviews, and resulting economies, reflected this dependency. In northwestern North America, for example, a native economy centered on the abundance of salmon began to develop about 3000 years ago. At its core was the concept of the gift and a belief system that treated all parts of the earth—animate and inanimate—as equal members of a community. In this and other ancient gift economies, a gift was not a possession that could be owned; rather, it had to be passed on, creating a cycle of obligatory returns. Individuals or tribes gained prestige through the size of their gifts, not the amount of wealth they accumulated.

This system coevolved with the migratory habits of the salmon, which moved en masse upriver to spawn each year. Because the Indians viewed salmon as equals to themselves, killing salmon represented a gift of food from salmon to people. Fishers were obligated to treat salmon with respect or risk losing this vital gift. The exchange of gifts between salmon and humans—food for respectful treatment—minimized waste and the chance of overharvesting and ensured a continuous supply of food. Further, the perennial trading of gifts among the people effectively redistributed the natural wealth brought each year by fluctuating populations of migrating fish, leveling out the boom-and-bust cycles that usually accompany reliance on an uncertain resource.

In modern times, the gift economy has gone entirely, along with the egalitarian worldview that accompanied...
it. We now have a redistributive economy tied not to an exchange of gifts with nature but to the exploitation of nature and to the technologies that enhance that exploitation. Nature became a resource for humans rather than an equal to humans. In economic terms, natural resources fell under the heading of ‘land’ in an economic trinity comprising three factors of production: land, labor, and capital. Land and resources, including crops, became commodities, expendable or easily substitutable forms of capital, whose value was determined solely by their value in the human marketplace.

In 1776 Adam Smith published his famous Inquiry into the Nature and Causes of the Wealth of Nations, in which he argued that society is merely the sum of its individuals, that the social good is the sum of individual wants, and that markets (the “invisible hand”) automatically guide individual behavior to the common good. Crucial to his theories was division of labor and the idea that all the factors of production were freely mobile; his mechanistic views created an economic rationale for no longer regarding individuals as members of a community linked by moral, social, and ecological bonds.

About the same time, fueling and fueled by the beginnings of the industrial revolution, the study of the natural world was transforming itself into modern physics, chemistry, geology, and biology. Before the mid-19th century, those who studied the natural world—early 19th-century German biogeographer Baron Alexander von Humboldt and his disciple Charles Darwin among them—took an integrated view of science and nature, including humans. For von Humboldt, understanding interdependence was the “noblest and most important result” of scientific inquiry. But this integrated natural philosophy was soon supplanted by more atomistic views, which fit better with industrialization.

Mass production of new machines relied on division of labor and interchangeable parts. Like automobiles on an assembly line, natural phenomena too were broken down into their supposed component parts in a reductionism that has dominated science ever since. Rushing to gain in-depth, specialized knowledge, science and society lost sight of the need to tie the knowledge together. Disciplinary specialization replaced integrative scholarship.

Neoclassical economics, which arose around 1870, ushered in the economic worldview that rules today. A good’s value was no longer tied to the labor required to make it but derived instead from its scarcity. A good’s price was determined only by the interaction of supply and demand. As part of “land,” natural resources therefore became part of the human economy, rather than the material foundation that makes the human economy possible. Because of its doctrine of infinite substitutability, neoclassical economics rejects any limits on growth; forgotten are the classical economic thinkers and contemporaries of von Humboldt, including Thomas Malthus and John Stuart Mill, who saw limits to the growth of human population and material well-being.

B. Too Many and Too Much

In October 1999, the global human population reached 6 billion. It doubled from 3 billion to 6 billion in about 40 years—before most post–World War II baby boomers even reached retirement age. From the appearance of Homo sapiens about half a million years ago, it took humans until 1804 to reach their first billion, 123 years to double to 2 billion, and 33 years to achieve 3 billion. With fertility rates declining in developed countries, China, and some developing countries where women are gaining education and economic power, and with pandemics like AIDS claiming more lives, the United Nations predicts that world population will reach 8.9 billion by 2030. Lined up one by one, the world’s 6 billion people would stretch nearly to the moon; jammed together in a crowd, they would just fit into the state of Texas.

But these 6 billion people consume far more resources than can be produced by a piece of earth the size of Texas. As already noted, humans appropriate about 40% of global plant production, 54% of the earth’s freshwater runoff, and enough of the ocean’s bounty to deplete or nearly extinguish two-thirds of the world’s major marine fisheries. In energy terms, a human’s food consumption is about 2500 to 3000 calories a day, about the same as that of a common dolphin. But with all the other energy and materials humans use, the rise in global per capita energy and material consumption has soared even faster than population growth in the past 40 years. Mostly in the form of fossil fuels, North Americans consume six times as much energy as the average human. Now, instead of coevolving with a natural economy, global society is coevolving with fossil fuels and the American Dream.

V. MEASURING ENVIRONMENTAL IMPACTS

For most of the 20th century, environmental indicators tracked primarily two classes of information: counts of
administrative activities in support of environmental protection and the supply of products to people. Counting bureaucratic achievements—such as the numbers of environmental laws passed, permits issued, enforcement actions taken, or treatment plants constructed—focuses on actions rather than on the results of those actions. Ultimately such counts provide only limited information about real environmental status and trends. Resource managers concentrate on resource supply. Water managers, for example, measure “water quantity,” allocating water to domestic, industrial, and agricultural uses and rarely reserving any to sustain aquatic life or to protect scenic and recreational values. Foresters, farmers, and fishermen count board-feet of timber, bushels of grain, and tons and thousands of fish harvested. All these measures keep track of commodity production, not broader environmental quality. Even biologists’ counts of threatened and endangered species—which would seem to measure biotic impoverishment directly—still focus narrowly on biological parts, not ecological wholes. Enumerating threatened and endangered species is just like counting any other commodity, reinforcing humanity’s view that we know which parts of the biota are important or valuable.

Society needs to use to better effect the environmental measures it has had (Table III). It particularly needs objective measures more directly tied to the condition, or health, of the environment so that people can judge whether their actions are compromising that condition. Such measures should be quantitative yet easy to understand and communicate; they should be cost-effective and applicable in many circumstances. Unlike narrow criteria tracking only commodity or extinction numbers, they should provide reliable signals about status and trends in ecological systems. Ideally, effective indicators should describe the present condition of a place, document change, diagnose its cause, and predict what is likely to happen. They should reveal not only risks from present activities but also potential benefits from alternative management decisions.

Most important, these indicators should, either singly or in combination, give information explicitly about living systems. Measurements of physical or chemical factors can sometimes act as surrogates for direct biological measurements, but only when the connection between those measures and living systems is clearly understood. Too often we make assumptions—when water managers assume that chemically clean water equals a healthy aquatic biota, for example—that turn out to be wrong and fail to protect living systems.

As environmental concerns grow more urgent, several new measurements aim to direct the public’s attention to the value of living things. Using the value system best understood by the people and institutions driving the global economic engine, some researchers assign standard monetary amounts to the goods and services humans derive from nature. A calculation by Robert Costanza and his colleagues puts the average total monetary value of natural capital and ecological services in 16 biomes at US$33,000 billion per year—an amount exceeding the gross global product. Another study, by ecologist David Pimentel and his colleagues, calculated separate values for specific biological services, such as soil formation, crop breeding, or pollination; by summing these figures, these researchers estimate the total economic benefits of biodiversity for the United States at $319 billion, for the world at $2928 billion—in other words, 5 and 11% of the gross national and gross global products, respectively. Still another of Pimentel’s analyses reports that the approximately 50,000 nonnative species in the United States cause major environmental damage and reparation costs amounting to $137 billion a year. Though this monetary approach does not create a comprehensive indicator of environmental quality, it certainly points out that ecological values ignored by the global economy are very high.

Another accounting approach, pioneered by Canadian landscape planners Mathis Wackernagel and William Rees, translates humans’ impact on nature, particularly resource consumption, into a metaphorical “ecological footprint.” Researchers estimate the amount of land and water area required by a city, town, nation, or other human community to produce consumed resources and absorb the generated wastes; they then compare the physical area occupied by a city or country with the area required to supply that city or country’s needs. The 29 largest cities of Baltic Europe, for example, appropriate areas of forest, agricultural, marine, and wetland ecosystems that are at least 565 to 1130 times larger than the areas of the cities themselves. National ecological footprints range from a high of 10.3 hectares per person for the United States to 0.5 hectares per person for Bangladesh. Thirty-four of 52 nations are now operating under ecological deficits; that is, their material standards are subsidized by exploiting other nations’ natural capital.

Raising all 6 billion people on Earth at the start of the 21st century to living standards, and their ecological footprints, equal to those in the United States would require an additional two planets. Clearly, humans are consuming more resources, and discarding more waste, than the earth’s living systems can produce or absorb in a given time period. This gap is the global sustainability gap the world now faces.
### Indirect depletion of living systems through alterations in physical and chemical environments

1. Degradation of water (chemical contaminant concentrations, river flows, rainfall, runoff)
2. Soil depletion (erosion rates, desertification rates, salt accumulation in soils)
3. Chemical contamination (pollutant and toxic emissions, pollutant and toxic concentrations in air, water, soil, and living organisms)
4. Altered biogeochemical cycles (river flows and lake levels, amount of nutrients going into water bodies, or nutrient loading, nutrient concentrations in water bodies, chlorophyll concentrations reflecting nutrient-triggered algal blooms, oxygen depletion in water bodies, trophic status of lakes, changes in air and soil chemistry, atmospheric greenhouse gas concentrations)
5. Global climate change (atmospheric greenhouse gas concentrations, change in atmospheric temperatures, distribution and intensity of severe storms or droughts)

### Direct depletion of nonhuman life

1. Overharvest of renewable resources such as fish and timber (tons of fish harvested, fish population "escapement," hatchery fish released and recovered, board-feet of timber harvested, forest regrowth rates, "stumpage," or standing timber, ecological footprints)
2. Habitat fragmentation and loss (area of forest or wetland habitats remaining, landscape connectivity, rates of habitat destruction)
3. Biotic homogenization (number of extinct, threatened, and endangered taxonomic groups, spread of nonindigenous species, local or regional diversity, damage and repair costs of invasions or extinctions, major shifts in species distributions)
4. Genetic engineering (diversity among cultivated crop strains, genetic diversity within strains, escape of genetically engineered organisms or traits to wild populations)

### Direct degradation of human life

1. Epidemics and emerging diseases (death and disease infection rates, or mortality and morbidity, geographic spread of diseases, recovery rates, frequency and spread of antibiotic and other drug resistance)
2. Loss of cultural diversity (extinction of languages, disappearance of cultures)
3. Reduced quality of life (population size and growth, changes in death rates or average life spans, infant mortality rates, teen pregnancy rates, number of chronically malnourished individuals, starvation rates, literacy rates, rates of stress and other diseases of affluence, length of work week, child or other forced labor, employment shifts by economic sector)
4. Environmental injustice (string of toxic waste dumps or waste emissions relative to residents communities, economic exploitation of certain groups, worker strikes, wage and income gaps, unemployment rates for different economic sectors)
5. Political instability (frequency of domestic and international strife, environmental terrorism rates, number of environmental refugees, ethnic "cleansing")
6. Cumulative effects (frequency of "rare" natural disasters, costs of weather-related property damage, human death tolls, government subsidies of environmentally destructive activities such as fishery overcapitalization, below-cost timber sales, water projects, and agricultural supports, replacement costs for ecological services, pricing that reflects environmental costs, "green" taxes, use of polyvalent practices, number of organic farms)

---

<table>
<thead>
<tr>
<th>Plausible Indicators of Environmental Quality*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Degradation of water (chemical contaminant concentrations, river flows, rainfall, runoff)</td>
</tr>
<tr>
<td>2. Soil depletion (erosion rates, desertification rates, salt accumulation in soils)</td>
</tr>
<tr>
<td>3. Chemical contamination (pollutant and toxic emissions, pollutant and toxic concentrations in air, water, soil, and living organisms)</td>
</tr>
<tr>
<td>4. Altered biogeochemical cycles (river flows and lake levels, amount of nutrients going into water bodies, or nutrient loading, nutrient concentrations in water bodies, chlorophyll concentrations reflecting nutrient-triggered algal blooms, oxygen depletion in water bodies, trophic status of lakes, changes in air and soil chemistry, atmospheric greenhouse gas concentrations)</td>
</tr>
<tr>
<td>5. Global climate change (atmospheric greenhouse gas concentrations, change in atmospheric temperatures, distribution and intensity of severe storms or droughts)</td>
</tr>
</tbody>
</table>

---

*These indicators have been or could be used to monitor status and trends in environmental quality, including dimensions of biotic impoverishment. Without a full spectrum of indicators, however, and without coupling them to direct measures of biological condition, only a partial or segmented view of environmental quality (the degree of biotic impoverishment) will emerge.

A few indexes have integrated economic, environmental, and human community indicators into a general measure of sustainability. The index of environmental trends for nine industrialized countries incorporates ratings of air, land, and water quality; chemical and waste generation; and energy use since 1970. By its rankings, environmental quality in the United States has gone down by 22% while Denmark has declined by 11%. Social scientists Herman Daly and John Cobb developed an index of sustainable economic welfare that adjusts the United States' gross national product by factoring in environmental good things and factoring out environmental bad things. Public expenditures on education, for example, are weighted as "goods" while costs of pollution cleanup, depletion of natural resources, and treating environmentally related illnesses are counted as "bads." Instead of showing continual growth as gross national product does, the index of sustainable economic welfare has remained nearly unchanged over three decades.
Such approaches offer important insights into the dangers of local and global resource consumption and consumerism and paint a clearer picture than ever before of humans’ domination of the earth and the biosphere. But the accounting systems are still human centered. They still do not measure the condition of the biosphere itself. We may know that biodiversity’s services are worth huge sums of money and that our hometown’s ecological footprint is much bigger than our town, but how do we know whether specific actions damage living systems or that other actions benefit them? How do we know if aggregate human activity is diminishing life on Earth? To answer this question, we need direct measures of the condition of living systems.

More comprehensive than monetary or footprint analyses, biological assessment directly measures the attributes of living systems to determine the condition of a landscape. The very presence of living systems—sea palms on the California coast, salmon in Pacific Northwest waters, monk seals in the Mediterranean Sea—says that the conditions those organisms need to survive are also present. A biota is thus the most direct and integrative indicator of local, regional, or global biological condition. Biological assessments give us a way to evaluate whether our monetary valuations and ecological footprints are telling the truth about human impact on the biosphere. Biological assessments permit a new level of integration because living systems, including human cultures, register the accumulated effects of all forms of degradation caused by human actions.

Direct, comprehensive biological assessment has been done for many aquatic systems; measures are less developed for terrestrial systems. The index of biological integrity (IBI), for example, was developed in 1981 to assess the health of streams in the U.S. Midwest and has since helped scientists, resource managers, and citizen volunteers to understand, protect, and restore rivers worldwide. The index borrows a page from well-known composite indexes of economic performance—the concept of multiple indicators—and applies it to animals and plants in bodies of water. The specific measurements (Table IV) are sensitive to a broad range of human effects in waterways, such as sedimentation, nutrient enrichment, toxic chemicals, physical habitat destruction, and altered flows. The resulting index combines the responses to human actions of both biological parts (such as species) and processes (such as food web dynamics).

Indexes of biological integrity have been developed for a number of aquatic and terrestrial environments; the widest-used indexes for assessing rivers examine fishes and benthic (bottom-dwelling) invertebrates. These groups are abundant and easily sampled, and the species living in virtually any water body represent a diversity of anatomical, ecological, and behavioral adaptations. As humans alter watersheds and water bodies, shifts occur in taxonomic richness (biodiversity), species composition (which species are present), individual health, and feeding and reproductive relationships.

Sampling the inhabitants of a stream can tell us much about a stream and its landscape. Biological diversity is higher upstream of wastewater treatment plants than downstream, for example, at the same location, year-to-year variation is low (Fig. 1). Biological sampling also reveals differences between urban and rural streams. For instance, samples of invertebrates from one of the best streams in rural King County, Washington (United States) contain 27 kinds, or taxa, of invertebrates; similar samples from an urban stream in Seattle contain only 7. The rural stream has 18 taxa of mayflies, stoneflies, and caddisflies, the urban stream only 2 or 3. When these and other metrics are combined in an index based on invertebrates, the resulting benthic IBI (B-IBI) ranks the condition, or health, of a stream numerically (Table V). The B-IBI for the rural stream in King County was 46 (the 10-metric index maximum is 50); that for the urban stream, 12 (the index minimum is 10).

A benthic IBI can also be used to compare sites in different regions. Nearly pristine areas in Wyoming’s Grand Teton National Park have near-maximum B-IBIs. Streams with moderate recreation taking place in their

<table>
<thead>
<tr>
<th>TABLE IV</th>
<th>Biological Attributes in Two Indexes of Biological Integrity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benthic invertebrates</td>
<td>Fish</td>
</tr>
<tr>
<td>Total number of taxa</td>
<td>Number of native fish species</td>
</tr>
<tr>
<td>Number of mayfly taxa</td>
<td>Number of riffle-benthic insectivore species</td>
</tr>
<tr>
<td>Number of stonefly taxa</td>
<td>Number of water-column insectivore species</td>
</tr>
<tr>
<td>Number of caddisfly taxa</td>
<td>Number of pool-benthic insectivore species</td>
</tr>
<tr>
<td>Number of inselreut taxa</td>
<td>Number of inselreut species</td>
</tr>
<tr>
<td>Number of long-lived taxa</td>
<td>Relative abundance of omnivores</td>
</tr>
<tr>
<td>Number of clinger taxa</td>
<td>Relative abundance of invertebrates</td>
</tr>
<tr>
<td>Relative abundance of tolerant taxa</td>
<td>Relative abundance of tolerant taxa</td>
</tr>
<tr>
<td>Relative abundance of predators</td>
<td>Relative abundance of top carnivores</td>
</tr>
<tr>
<td>Number of caddisfly taxa</td>
<td>Number of pool-benthic insectivore species</td>
</tr>
<tr>
<td>Number of mayfly taxa</td>
<td>Number of riffle-benthic insectivore species</td>
</tr>
<tr>
<td>Number of stonefly taxa</td>
<td>Number of water-column insectivore species</td>
</tr>
<tr>
<td>Total number of taxa</td>
<td>Number of native fish species</td>
</tr>
</tbody>
</table>
washes have B-IBIs that are not significantly lower than those with no human presence, but places where recreation is heavy are clearly damaged. Urban streams in the nearby town of Jackson are even more degraded, yet not as bad as urban streams in Seattle.

The core message embodied in biological assessment is that preventing harmful environmental impacts goes beyond narrow protection of clean water or clear skies, even beyond protecting single desired species. Certain species may be valuable for commerce or sport, but these species do not exist in isolation. We cannot predict which organisms are vital for the survival of commercial species or species we want for other reasons. Failing to protect phytoplankton, zooplankton, insects, higher plants, bacteria, or fungi ignores the key contributions of these groups to healthy biotic communities.

No matter how important a particular species is to humans, it cannot persist outside the biological context that sustains it. Direct biological assessment objectively measures this context.

VI. RECOGNIZING AND MANAGING ENVIRONMENTAL IMPACTS

Every animal is alert to dangers in its environment. A microscopic protozoan gliding through water responds to light, temperature, and chemicals in its path; it turns or retreats at the first sign of something noxious. A bird looking for food must decide when to pursue prey and when not, because pursuit might expose the bird to predators. The bird might risk pursuit when it is hungry but not when it has young to protect. Animals that assess risks properly and adjust their behavior are more likely to survive; in nature, flawed risk assessment often means death or the end of a genetic line.

Humans too are natural risk assessors. Each person chooses whether to smoke or drink, to drive a car or ride a motorcycle and at what speeds, to fly or take the train, to engage in "extreme" sports or go for a woodland stroll. Each decision is the result of a partially objective, partially subjective internal calculus that weighs benefits and risks against one another.

Risk is a combination of two factors: the numerical probability that an adverse event will occur and the consequences of the adverse event. People may not always have the right signals about these two components, however, and so base their risk calculus on the wrong clues. Urban dwellers in the United States generally feel that it is safer to drive home on a Saturday night than to fly in a jetliner, for example. Even though the numerical odds of an accident are much higher on the highway than in the air, people fear more the consequences of an airliner falling out of the sky.

Human society also strives to reduce its collective exposure to risks, primarily through government agencies responsible for protecting the public's interests. Governments do not hesitate to use military power and international agreements to protect their people from external aggression. They have, albeit more reluctantly, also used their regulatory power to reduce workplace risks and risks associated with consumer products like automobiles. But people and their governments have been much less successful in defining and reducing a broad range of ecological risks, largely because they have denied that the threats are real.

Society has long behaved as if its activities did not
![Image of document content]

**Table V**

<table>
<thead>
<tr>
<th>Region</th>
<th>Land use</th>
<th>B-IBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>King County, Washington</td>
<td>Rural</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Urban Seattle</td>
<td>12</td>
</tr>
<tr>
<td>Grand Teton region, Wyoming</td>
<td>Little or no human activity</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Light to moderate recreation</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Heavy recreation</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Urban Jackson Hole</td>
<td>21</td>
</tr>
<tr>
<td>Clackamas County, Oregon</td>
<td>Upstream of wastewater treatment plant</td>
<td>40, 42</td>
</tr>
<tr>
<td></td>
<td>Tickle Creek up (1997, 1998)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Foster Creek</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Downstream of wastewater treatment plant</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tickle Creek down (1997, 1998)</td>
<td>14, 16</td>
</tr>
<tr>
<td></td>
<td>North Fork Deep Creek</td>
<td>10</td>
</tr>
</tbody>
</table>

* Benthic index of biological integrity: the highest possible score is 50, the lowest is 10.

* See Fig. 1 for graphs of selected B-IBI metrics at these sites.

Entail any risks to nonhuman living systems, that is, any ecological risks. The plans generated by economists, technologists, engineers, and even ecologists typically assume that the lost and damaged components of living systems are unimportant or can be repaired or replaced. Widespread ecological degradation has resulted directly from the failure of modern society to properly assess the ecological risks it faces. Like ancient Egypt's or Easter Island's fate, our civilization's future depends on our ability to recognize this deficiency and correct it.

Risk assessment as formally practiced by various government agencies began as a way to evaluate the effects on human health of toxic substances, usually the effects of single substances, such as pollutants or drugs, from single sources, such as a chemical plant. During the 1990s, the focus widened to encompass mixtures of substances and also ecological risks. Ecological risk assessment by the U.S. Environmental Protection Agency asks five questions: Is there a problem? What is the nature of the problem? What are the exposure and ecological effects? (A hazard to which no one or nothing is exposed is not considered to pose any risk.) How can we summarize and explain the problem to 'stakeholders' (both at-risk populations and those whose activities would be curtailed)? How can we manage the risks?

Even though these are good questions, ecological risk management has not made any visible headway in stemming biotic impoverishment. Its central failing comes from an inability to correctly answer the second question, What is the nature of the problem? Our present political, social, and economic systems simply do not give us the right clues about what is at risk. None of society's most familiar indicators—housing starts, gross national product, index of leading economic indicators, educational testing scores, air quality alerts, even number of threatened and endangered species—measure the consequences, or risks, of losing living systems.

Moreover, ecological risk assessment is still an assessment of the risks of business as usual, not an evaluation of potential benefits from alternatives to business as usual. When deciding what level of grazing to allow, for example, ecological risk assessors still examine the risks to, or impacts on, existing resources such as soil, water, and vegetation. They do not ask what the benefits of different grazing regimes would be; neither do they consider the benefits of no grazing at all, even though the no-grazing alternative might offer the greatest ecological, and ultimately societal, benefits.

If biotic impoverishment is the problem, then it makes more sense to direct environmental policy toward protecting the integrity of biotic systems. Aldo Leopold, in *A Sand County Almanac*, was the first to invoke the concept of integrity in an ecological sense: 'A thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends to do otherwise.' Integrity implies a wholeness or unimpaired condition. In present biological usage, integrity refers to the condition at sites with little or no influence from human activity; the organisms there are the products of natural evolutionary and
biogeographic processes in the absence of humans. Ty-
ing the concept of integrity to an evolutionary frame-
work provides a benchmark against which to evaluate
sites that humans have altered.

Directing policy toward protecting biological integ-
ruity—as called for in the United States’ Clean Water Act,
Canada’s National Park Act, and water policy directives
being established by the European Union—does not,
however, mean that humans must cease all activity that
disrupts a pristine earthly biota. The demands of feed-
ing, clothing, and housing billions of people mean that
few places on Earth will maintain a biota with evolution-
ary and biogeographic integrity. Rather, measurements
founded on the evolutionary idea of integrity allow us
to directly assess biotic condition and to compare that
condition with what might be expected in a place with
little or no human influence. At least then we can make
an informed choice: continue with activities that de-
grade biotic condition or think of an alternative.

Ethical arguments aside, humans’ own dependence
on living systems makes it in our interest to manage
our activities so they do not compromise a place’s ability
to support those activities in the future; that ability can
be called ecological health. Ecological health describes
the preferred state of sites heavily used for human pur-
poses: croplands, tree farms, water bodies stocked for
fish, and urban areas. Integrity in an evolutionary sense
cannot be a goal at these places, but we should avoid
practices that so damage places that we cannot continue
to use them. Agricultural practices that leave soils
salted, lower regional water tables, and erode fertile
topsoil faster than it can be renewed destroy prospects
for future agriculture; such practices are unhealthy.

In contrast to risk assessment, and more akin to
ecological benefits assessment, striving to protect bio-
logical integrity is more likely to lead away from technolo-
gical fixes for environmental problems and toward
practices that prevent ecological degradation and en-
courage restoration. Biological integrity as a policy goal
turns our focus away from maximizing the goods and
services provided for the human economy and toward
ways to manage human affairs within the bounds set
by the natural economy. It begins to turn our attention
away from “How much stress can landscapes and eco-
systems absorb?” to “How can responsible human ac-
tions protect and restore ecosystems?” It could even
help avoid contentious fiscal allocation issues that arise
when massive emergency operations—like rescuing en-
dangered Pacific salmon in the Columbia River—are
needed to pull species back from the brink of extinction.

To be sure, managing for biological integrity requires
a deep commitment to self-imposed limits on human
population size, curbs on consumerism, less-selfish atti-
itudes toward land stewardship, and the realization that
the biosphere matters. Instead of calling on human
technical and spiritual井springs to manage resources,
we have to call on them for managing human affairs.
We have to find and use appropriate measurements for
all the factors contributing to biotic impoverishment,
be they climate change, overharvesting, agriculture, or
environmental injustice.

Managing our affairs to prevent ecological risks re-
quires integrating ecological systems and human social
and political systems. The ecological world is a com-
plex, variable system. We cannot predict with certainty
the intensity of the next El Niño drought or rainstorm.
We cannot know with assurance whether the cumula-
tive effects in a place will turn the next earthquake into
da disaster or an inconvenience. We cannot calculate
the absolute risks (or benefits) of pumping (or not
pumping) a given tonnage of carbon dioxide into the
air. Purely quantitative risk assessment only works in
the absence of such uncertainties.

Instead, when managing for ecological risks, people
and their governments need to expect the unexpected
and develop formal, yet flexible means of coping with
environmental surprises. Rather than plunge ahead
with projects entailing ecological risks because they can
be done, decision makers should follow the precaution-
ary principle, which holds that regulators should act
to prevent potential environmental harm even in the
absence of certainty. It acknowledges the existence of
uncertainty rather than denying it, and it includes
mechanisms to safeguard against potentially harmful
effects.

Though inappropriate ecological risk assessment and
management is more often the norm today, modern
institutions can recognize ecological threats correctly
and respond to them in time. The Montreal Protocol
is a prime example. Just over a decade after its adoption,
satellite measurements in the stratosphere indicated
that harmful chlorine pollution was in fact on the
decline.

VII. RECLAIMING LOST CONNECTIONS

Early in the 20th century, two sciences of “home main-
tenance” began to flourish: the young science of ecology
(from the Greek oikos, meaning home) and a maturing
neoclassical economics (also from oikos). Ecology arose
to document and understand the interactions between
organisms and their living and nonliving surroundings—in essence, how organisms make a living in the natural economy. In fact, Ernst Haeckel, who coined the term in the 1860s, defined ecology [in an 1870 article] as "the body of knowledge concerning the economy of nature." Neoclassical economics reinforced humans' self-appointed dominion over nature's free wealth and brought unparalleled gains in societal welfare in some places, but it also divorced the human economy from the natural one on which it stands.

Monitored by both those sciences, human actions and their effects have reached scales unprecedented in the history of life. We have altered the earth's physical and chemical environment, changed the planet's water and nutrient cycles, and shifted its climate. We have unleashed the greatest mass extinction in 65 million years and disrupted the human and human communities worldwide. In trying to make our own living, we have contributed not only to the global loss of individual lives and loss of species but, worse, to the loss of life itself—the removal of life from the biosphere. In the end, life—and earth's capacity to sustain life—is at risk.

Early in the 21st century, we need a new science and art of home maintenance, one that helps us understand and interpret the consequences of human-driven change. We need a new story, a different worldview, to guide our behavior, one that is in harmony with nature's economy. We must seek a balance between our modern industrial economies and our homelands' natural economies. As individuals and as societies, we need to understand the consequences for the present and future biosphere of what we do.

To reclaim our connections to the natural world, those of us insulated within industrial societies have to start small, with connections we can understand—like what plants and animals live in our backyards, which chemicals are used to grow our coffee, which laborers glued the soles on our walking shoes, where the pulp in our newspapers came from. We have to pay attention to how living systems respond to our behavior. In order to live, we have to let live. In his 1995 contribution to A New Century for Natural Resources Management, environmental ethicist Holmes Rolston Ill writes, "The next millennium is, some say, the epoch of the end of nature. But another hope is that we can launch a millennium of culture in harmony with nature."

See Also the Following Articles

BIOGEOCHEMICAL CYCLES • ECOCOLOGY • ENERGY USE, HUMAN • ENVIRONMENTAL ETHICS • HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW

Bibliography

I. Introduction
II. Definition and Classification
III. Estuarine Biodiversity
IV. Ecological Function of Biodiversity
V. A Case Study: The Chesapeake Bay
VI. Future Challenges

GLOSSARY

biological diversity (biodiversity) The collection of genomes, species, and ecosystems occurring in a geographically defined region (NRC, 1995).
coastal zone Zone whose terrestrial boundary is defined by (a) the inland extent of astronomical tidal influence or (b) the inland limit of penetration of marine aerosols within the atmospheric boundary layer and including both salts and suspended liquids, whichever is greater; the seaward limit is defined by (a) the outer extent of the continental shelf (approximately 200 m depth) or (b) the limits of territorial waters, whichever is greater (Hayden et al., 1984).
estuary Semi-enclosed coastal body of water that has a free connection with the open sea and within which seawater is measurably diluted with freshwater derived from land drainage (Pritchard, 1967).

functional diversity Variety of different responses to environmental change, especially the diverse time and space scales with which organisms react to each other and to the environment (Steele, 1991).
metapopulation An abstraction of the population to a higher level at which individuals frequently move from one place (population) to another, typically across habitat types that are not suitable for their feeding and breeding activities, and often with substantial risk of failing to locate another suitable habitat patch in which to settle (Hanski and Gilpin, 1991).

ESTUARIES ARE AMONG THE MOST IMPORTANT INTERCONNECTIONS between land and sea. They are situated in the coastal zone, which accounts for a disproportionate amount of global ecological functions. For example, the coastal zone (modified from Pernetta and Milliman, 1995):

• occupies only 18% of the surface of the globe, 8% of the ocean surface, and 0.3% of ocean volume;
• but provides for up to 50% of global denitrification, 80% of global organic matter burial, 90% of global...
sedimentary mineralization, 75–90% of the global sink of suspended river load and its associated elements/pollutants, and in excess of 50% of present-day global carbonate deposition;

- also supplies approximately a quarter of global primary production, around 14% of global ocean production, and 90% of the world fish catch.

It follows that estuaries, as major pathways of aquatic exchange between land and sea, are major influences on a large proportion of these functions.

I. INTRODUCTION

Some of the steepest environmental gradients on planet Earth occur in the coastal zone, where land, sea, and atmosphere uniquely interact to exchange energy and materials. Also, the dynamic linkages among biological, physical, and chemical systems are exceptionally strong in estuaries, and are characterized by cyclic changes that occur at different frequencies—such as for tides, salinity cycles, freshwater inputs, light, and temperature stratification. Estuaries also bear the brunt of extreme events, such as flooding, storms, hurricanes, and seasonal sea ice. All of these are of importance for organisms, which have evolved suites of adaptive mechanisms to cope.

Estuaries have usually been considered as transitional areas between freshwater and saltwater environments. However, relatively few species are totally confined to estuarine conditions, even though various stages of many species’ life cycles are estuary-dependent. This raises questions about whether estuaries can be considered as transitional or as more-or-less autonomous ecosystems in their own right. The distribution of biodiversity provides important information toward the resolution of this apparent dichotomy, which needs to be clarified. These questions require both ultimate, historical-evolutionary explanations and proximate, functional-ecological explanations.

Despite the location of estuaries in the critical portion of Earth called the “coastal zone,” the Global Biodiversity Assessment (Heywood and Watson, 1995) contains no sections specifically devoted to them; the term “estuary” does not even appear in the index! Nevertheless, this volume does characterize biodiversity as comprising three disciplines, which also apply to estuaries:

1. taxonomy: provides the reference system and depicts the pattern or tree of diversity for all organisms;
2. genetics: gives a direct knowledge of the gene variations found within and between species; and
3. ecology: provides knowledge of the varied ecological systems in which taxonomic and genetic diversity are located, and it also provides the functional components.

Evolutionary biology brings these together, as it provides explanations of how biodiversity arose, and the processes, such as speciation and extinction, by which it continues to change.

The third aspect of biodiversity, namely, the functional-ecological aspect, is the focus of this article. In this respect, it is worthwhile to note that, even today, estuarine science continues to be organized along disciplinary lines. Although the study of land–seascape ecology of estuaries remains in its infancy, there are extensive publications on geomorphology, land–sea interactions, coastal zone management, and other disciplinary lines. Although the study of land–seascape ecology of estuaries remains in its infancy, there are extensive publications on geomorphology, land–sea interactions, coastal zone management, and other disciplinary lines. Although the study of land–seascape ecology of estuaries remains in its infancy, there are extensive publications on geomorphology, land–sea interactions, coastal zone management, and other disciplinary lines.

The distribution of estuaries corresponds to regional and coastal characteristics; that is, they tend to be extensive, large, and numerous where coastal plains are wide and flat, but are relatively small where coastal plains are steep and narrow. Particularly in the former, estuaries and associated lagoons constitute a much higher percentage of the coasts than is generally recognized. In fact, many of the world’s largest cities (London, New York, Karachi, Amsterdam, Alexandria, Tokyo, etc.) have been built on or near drained marshes or filled land adjacent to estuaries. In the United States, 80–90% of the Atlantic and Gulf Coasts and 10–20% of the
Estuaries are best understood in the context of the coastal zone, definitions of which vary. Ketchum (1972) was among the first to take a functional perspective, that the coastal zone "is the broad interface between land and water where production, consumption, and exchange processes occur at high rates of intensity." NERC (1992), on the other hand, defined the coastal zone as: "An indefinite zone of land and sea that straddles the shoreline; includes all land that is the product of, and/or at risk from (Holocene) marine processes, and extends seaward from the shoreline to water depths of about 30 m." The key element is "marine processes" and, from that point of view, it seems best to adopt Ketchum's broader view. Accordingly, Hayden and, from that perspective, it is most reasonable to assume that the communities of estuarine biota that exist today are as young and equally subject to change.

Many estuaries around the world have been studied in some detail. The North American bias in this article reflects the considerable body of research that has been conducted on North American estuaries during the past few decades, motivated unfortunately by the depleted, over-enriched, polluted, and over-populated states of many of them, some aspects of which will be examined in the Chesapeake Bay case study in Section V.

II. ESTUARINE CLASSIFICATION

Classification is essential as a comparative reference system, for otherwise data and information cannot be made comparable among estuaries. Various classifications, or typologies, of estuaries have been attempted, but these are mostly physical; no typology is directed specifically to biodiversity, even though the distributions of estuarine species have resulted in various classification schemes.

To my knowledge, the first classification was the so-called 'Venice system' (Anonymous, 1999), in which estuaries were divided into salinity zones. This was modified later by Bulger et al. (1993) on the basis of species' salinity tolerances. These two schemes align rather closely and may be compared as follows (Anonymous, 1999 = V; Bulger et al., 1993 = B; ppt = parts per thousand):

- Limnetic: freshwater, 0.5 ppt (V); freshwater, 4 ppt (B)
- Oligohaline: 0.5–5 ppt (V), 2–14 ppt (B)
- Mesohaline: 5–18 ppt (V), 11–18 ppt (B)
- Polyhaline: 18–30 ppt (V), 16–27 ppt (B)
- Euhaline: 30 ppt–full marine (V), 24–ppt marine (B)

The reason for the differences in salinity ranges between the Venice system and Bulger et al. is that the former was derived from salinity, whereas the latter was derived analytically from species' salinity tolerances, in which the zones would be expected to overlap. In both cases, however, the compartments are over-simplified, as estuaries exhibit many characteristics that influence biotic distribution and the distinction of estuarine zones, variably identified as "upper reaches," "upper-middle reaches," "lower reaches," and so forth. Nor do salinity-derived systems distinguish zones according to
variations in bottom type, water movement, volume of flow, and other attributes important to the biota. Another classification concerns basin geomorphology, which is of obvious importance for circulation patterns. Classification on this basis appears in many texts and may be summarized as:

- coastal plain estuary (drowned river valley): Usually confined to areas with a wide coastal plain where seawater has invaded existing rivers because of sea level rise since the Pleistocene Ice Age. Generally the up-estuary limit is where chlorinity is about 0.06% (salinity about 0.1%); above this point there may be a portion of tidal freshwater.
- fjord: Generally U-shaped in cross section, in which the sides are steep and have been glaciated. May be fed by a river, have a deep basin, and a shallow sill may be present near the mouth.
- bar-built: Occurs in flat, low-lying areas, where sand tends to be deposited in bars lying parallel to the coast. Usually shallow and wind-mixed. Can be a composite of drowned river valleys and embayments, and occurs when offshore sand barriers are built between headlands into a chain to enclose the body of water. May be fed by multiple rivers, but the total drainage area is usually not large.
- tectonic: A miscellaneous category including estuaries formed from faults or folding of Earth’s crust. Often have an excess of freshwater flow.

The interchange of freshwater and seawater provides yet another classification. The inlet (mouth) must be of sufficient dimension to allow mixing of seawater and freshwater, and the dilution of seawater provides the density gradients that drive characteristic circulation patterns. In terms of this interchange, the general classification is:

- salt wedge: Wherein a layer of relatively fresh water flows out at the surface.
- partially mixed (moderately stratified): Wherein tidal flow, turbulence, and mixing are increased, tending to erase the salt wedge.
- vertically homogeneous: Wherein tidal flow is strong, river runoff is weak, and all stratification is broken down.

Combinations of these typologies are possible; that is, it may be possible to find a stratified or a mixed bar-built estuary, or a fjord with a salt wedge or not. Furthermore, the extents of salinity zones can vary considerably for all categories. Such combinations of structure and hydrologic process result in highly varied conditions in the distributions of, for example, sediment, phytoplankton, submerged aquatic vegetation, and fishes and invertebrates. Additionally, variations in freshwater inputs, circulation, turbulence, and mixing can modify the typology.

A final classification concerns estuarine evolution, such as that of Roy (1984) for estuaries of New South Wales, Australia. There, estuaries are of three successional types: drowned river valleys, barrier estuaries, and saline coastal lakes. All are characterized by infilling during relatively short time spans. This affects their size, configuration, the invasion of mangroves and other aquatic vegetation, and fish communities. Biodiversity maxima are reached in the intermediate stages, because faunal population densities and species diversity increase with ecological complexity. However, as infilling becomes more advanced, the estuary becomes simplified and biological diversity declines. Therefore, estuarine geology, hydrology, and biology form a hierarchical succession.

### III. ESTUARINE BIODIVERSITY

From the foregoing discussion, the impression may be gained that estuaries are simply transitional and, therefore, not biologically diverse. Indeed, Sanders (1968) found that estuaries are relatively non-diverse biologically, but also noted: “What is significant is that each environment seems to have its own characteristic rate of species increment.” This is to say that salinity, for example, is an important determinant of the distribution of the biota, but also that estuaries exhibit high habitat and land–seascape diversity, a consequence of which is high variability among the biota and a high degree of biotic interaction. Thus, estuarine biotic communities would be expected to be especially varied and complex, contrary to earlier impressions of estuarine biological and ecological simplicity. Additionally, their biota have evolved resiliency to disturbance, both natural and human-caused. This is expressed at species, community, and ecosystem levels, leading to the impression that estuarine species are facilitative with respect to estuaries as preferred environments. These characteristics have resulted in a tendency to describe any species that enters estuaries, or those that tolerate brackish waters, as “estuarine,” which can be misleading. Nevertheless, some species seem to be restricted to estuarine and near-shore environments, at least at some life-history stage. A notable example concerns temperate oysters, which build extensive reefs in...
and the invertebrates—aquatic, xanthid mud crabs, improvisus, cea mamatorium, Rhizophora mangle, alterniflora, Zostera marina, Ruppia maritima, Cymodoceae. Carriker concentrated mainly on benthic invertebrates, but concluded that an “estuarine biocenose” may be justified as a discrete functional aggregation of interdependent, regularly recurring, dominant, benthic populations that are strongly represented numerically. He acknowledged that much needs to be learned of ecology and life histories to justify this, but that the estuarine biotope appears to be more than “just a simple overlapping of factors (an ecotone) extending from the sea and the land, but is characterized by a unique set of its own factors arising from within the estuary from the materials and forces contributed by its bounding environments” (Carriker, 1967).

Some of the dominant, or “true,” macroscopic biota of estuaries that he named are the plants—Sparganium alterniflorum, Zostera marina, Ruppia maritima, Cymodocea mamatorium, Rhizophora mangle, and Avicennia nitida, and the invertebrates—Nereis diversicolor, Balanus improvisus, xanthid mud crabs, Uca pugnax, Callinectes sapidus, Mya arenaria, Mytilus edulis, Modiolus demissus, and Crassostrea virginica. Additionally, he noted that characteristic estuarine habitats include tidal marshes, mangrove swamps, seagrasses, oyster reefs, soft clam–worm flats, and others. Finally, Carriker stated that: “Little is known of the sum of these effects on community structure, but they do emphasize the need to consider benthic organisms in the context of the total ecosystem rather than as an independent benthic biocenose.” This statement, made a third of a century ago, has yet to be fully realized.

Fishes are the best known of aquatic groups in a general sense, mostly due to their commercial value. Therefore, insights into “estuarine dependency” may be best revealed through their study. One reason for this is their mobility in which various life-history stages inhabit quite different environments. Winemiller (1995) reviewed fish ecology and made the following points. First, fishes are by far the most diverse vertebrates, and they inhabit an incredibly wide range of aquatic habitats from pole to pole. Second, fishes are ecologically diverse, with a wide variety of food habits, behaviors, reproductive habits, physiologies, and morphologies. Third, fishes exhibit a range of life-history strategies that result from trade-offs among various attributes, including clutch and egg size; these strategies can be classified as opportunistic, periodic, and equilibrium; but a range of intermediate strategies also exist. Finally, fishes and their diversity in ecosystems can be used as “indicators” of environmental conditions.

Recently, much attention has been directed toward the early life histories of fishes, as this is closely related to recruitment and, therefore, of much interest to fish ecologists and to fisheries. Houde (1997) provided a review of the selection factors that are of special importance in this regard. Able and Fahay (1998) extended studies on juvenile stages of fishes to “estuarine dependence” and determined that the numbers of permanent estuarine residents is relatively low, at least in part because estuaries exhibit extremes in environmental conditions. Also, the fish diversity of estuaries is augmented by transients, such as freshwater species that occasionally occur in estuaries and marine species that spawn at sea but whose young use estuaries as nurseries. Therefore, the estuarine fish fauna includes both residents and transients and a wide range of sizes, ages, and adaptations. In addition, those species that have successfully invaded estuaries usually inhabit only a small number of broad niches, implying that larger estuaries have larger numbers of species owing to increased habitat and niche complexity.

Able and Fahay found that, of the species for which good information is available, 60% are transients, 28% are residents (uncannily close to the “educated guess” of C. R. Robins and myself that 27% are “obligate” on estuaries; see Section V), 6% are infrequent, and 6% are unclassifiable. Furthermore, they have suggested the following adaptive groups for juveniles:

**Group I. Facilitative estuarine breeders**:
- Species whose nurseries are either in estuaries or on the inner shelf (e.g., *Menidia menidia, Menidia cantis*).

**Group II. Seasonal residents**:
- Species whose adults migrate into estuaries to spawn in spring or summer (e.g., *Centropristes striata, Brevoortia tyrannus*).

**Group III. Anadromous species**:
- Species whose adults migrate through estuaries in order to spawn in freshwater (e.g., *Morone saxatilis, Allosa spp.*).

**Groups IV–VI. Early users, delayed users, and distant spawners**:
- Species that spawn exclusively in the ocean, but the location, timing, and manner of use
of estuaries by young-of-the-year juveniles vary (e.g., Pollachius virens, Prionotus carolinus, Mugil cephalus).

Group VII. Expatriates: species whose estuarine larvae come from distant spawning (e.g., Chrysopteris ocellata, Monacanthus hispidus).

Group VIII. Summer spawners: the largest group, represented by shallow-water spawners whose larvae develop in the immediate vicinity of spawning sites (e.g., Cyprinodon variegatus, Fundulus heteroclitus).

Group IX. Winter–spring spawners: a few species that spawn in the winter or spring (e.g., Pseudopleuronectes americanus).

Group X. Migrating spawners: species that undergo spawning migrations within the estuary (e.g., Morone americana).

Group XI. Species difficult to classify: species for which some populations appear to be estuarine and other populations do not (e.g., Tautogolabrus adspersus).

Able and Fahay (1998) caution that, for fishes at least, “estuarine dependence” depends on the resolution of three areas of research: (1) the need to sample well-defined areas thoroughly for habitat evaluation; (2) assessment of the effects of habitat loss; and (3) more detail on temporal and spatial use of habitats where early stages are collected. In short, a coherent understanding of the life-history factors that control the early life histories of fishes remains to be accomplished. The same no doubt holds for invertebrates. For macroscopic plants, the situation is perhaps less uncertain, as their life histories are simpler and assessments are more easily accomplished.

In sum, most truly estuarine species are typically resistant to environmental variations due to the extreme conditions of estuaries, and/or take advantage of favorable situations; consequently, they do not appear to have strong habitat associations. This makes difficult the strict establishment of a definition of “estuarine dependency.” Also, the seaward boundary of an “estuary” is often blurred, so that the definition of “dependency” is hampered by lack of comparative, quantitative data from offshore habitats. The easiest distinctions are for those species for which at least one stage is shown to be physiologically or behaviorally obligate, but good natural history and experimental data are required for this. Therefore, the question “What is an estuarine species?” remains elusive. In addition, the oft-made contention that estuaries with similar habitats may support similar species assemblages seems reasonable, but may be misleading if assumptions of estuarine dependency are based on occurrence rather than in an adaptive-evolutionary sense.

IV. ECOLOGICAL FUNCTION OF BIODIVERSITY

In addition to genome, species, and ecosystem aspects of biodiversity, a fourth category must be considered, namely, “functional diversity” (Steele, 1991; see Glossary), which concerns ecological functions with respect to environmental maintenance and change. Ecological functions within the coastal zone and its estuaries are complex and variable, and they must be understood before we can interpret the composition and patterns of biodiversity. Holligan and Reiners (1992) listed a number of factors that underlie the biological diversity of the coastal zone and its estuaries, first for natural processes:

Exchanges of Materials  Riverine and atmospheric export and import, groundwater exchange, and ocean–land material transport operate at various levels, but are presently poorly understood. [Recent information on anadromous fishes is shedding light on organic-matter transport; e.g., Hesslein et al. (1991); Bilby et al. (1996); Garman and Macko (1998).]

Physico-chemical Properties  The coastal zone is a region of high energy exchange due to interactive oceanic and atmospheric forcing associated with topographical discontinuities, density gradients caused by freshwater inflows, and seasonal heat exchanges. Deltas, estuaries, and lagoons are the major sites for transformation and accumulation of organic matter and sediment, and all are highly variable spatially and temporally, so that their average conditions are not good indicators of net fluxes. Estuaries, in particular, are “sites of complex interactions, related to salinity gradients, phase transformation involving particle–water reactions, and to biological processes that cause biogeochemical transformations” (Holligan and Reiners, 1992).

Biological Properties  Favorable conditions of light and nutrients in the coastal zone maintain high rates of primary productivity that are several times greater than for the open ocean, and even greater than for certain coastal upwelling areas; some coastal systems, such as salt marshes, mangrove swamps, mudflats, beds of aquatic vegetation, and coral reefs, exhibit even
higher productivity. Some areas act as sources, others as sinks, and the nature of the coupling of primary productivity to the bottom or to open waters may determine community structure and function.

**Biogeochemical Processes** Organic matter is readily reoxidized in coastal waters, but some poorly drained areas may become anaerobic. This is especially apparent in the bottom water of estuaries in summer, when temperatures are high.

Many present-day human activities influence both ecological functions and biological diversity:

**Altered Delivery of Freshwater** Freshwater impoundment by damming has decreased total discharge into estuaries and coastal seas by about 15% since the 1950s, an amount equivalent to a change in sea level of ~0.7 mm/yr. Seasonal flows have also been altered; alteration in the residence time of water in estuaries may have far-reaching effects on chemical processes.

**Changes in the Transport and Fate of Suspended Matter** Coastal subsidence, sediment starvation and consolidation, and nutrient levels have all been altered by human interventions. Land clearing especially on steep slopes, has increased sedimentation.

**Chemical Modification** Nutrients, eutrophy, and blooms have become widespread and their frequency seems to be increasing. Contaminants that are of most concern include heavy metals, synthetic organic compounds, radionuclides, and hydrocarbons.

**Ecosystem Modification** This takes many forms, from physical change, to habitat loss, to depletion of resources. The worst-affected areas are those with high human population densities, such as Southeast Asia, and along temperate coasts that have significant sources of pollutants, such as the Baltic Sea.

Longer-term processes that influence biodiversity are the effects of climate change, especially in response to global warming, should that continue to occur:

**Natural Variations in Climate** Many climate-change studies describe possible variations in the altered distributions of biota. However, rather subtle changes in climatic conditions can induce large ecological changes that reflect the sensitive nature of marine food chains to climate and to climate-dependent factors such as nutrient levels and salinity. The direct effects of climate are difficult to distinguish from those incurred by humans.

**Temperature** The largest climate changes are expected in the higher latitudes. Thus, the poleward extension of climate-sensitive species is to be expected in case of global warming. Temperature changes can also affect behavior and physiology (e.g., reproduction, feeding and food availability, predation, migration), so that predictions are destined to be speculative.

**Wind** Wind strongly influences upwelling and stratification, thus affecting productivity through nutrient and light availability. According to most climate change scenarios, wind intensity is expected to increase.

**Extreme Events** Short time-scale events are also expected to increase with climate warming, and these may induce dramatic, long-term changes. A single storm lasting <5 days can result in sand transport equivalent to two-thirds of the total for an average year. Tsunamis have had the greatest effects recorded to date.

**Changes in Sea Level** Presently, sea level is rising faster than the rate during the late Holocene due to a combination of thermal expansion of seawater and melting of ice as the climate warms. Severe impacts of sea level rise on deltas and estuaries are already apparent, partly because they are low-lying, strongly perturbed by humans, and exhibit enhanced erosion and subsidence. Natural communities of plants and animals play a crucial role in determining the response of the coastal zone to changes in sea level.

This array of effects requires the development of research programs to address hypotheses that are relevant to the ecological function of estuarine biodiversity. Among many possibilities, the following seem essential (slightly modified from Solbrig, 1991):

- For species: no aspect of life history has any influence on extinction probability.
- For communities: keystone species are essential for maintaining species richness in communities under all environmental conditions.
- For ecosystems: removal or addition of functional or structural groups that produce changes in temporal or spatial configuration of landscape elements will have no significant effect on ecosystem properties over a range of time and space scales.

These hypotheses can be clarified by means of a case-
By-case examination (see the Chesapeake Bay case study). For example, some species seem very alike in their life histories. However, redundancy in species function may mean that diversity and function are somewhat independent of one another. Many species of benthis, infauna and epifauna are extremely abundant and ecologically important in estuaries. Many feed on sediments, and those with complete alimentary canals can consolodate organic residues into often long-lived, sculptured pellets. The question is: Many species have similar ecological requirements and, therefore, are species replaceable?

With respect to physical structure, Roy (1984) stated that the ecology of an estuary depends on the geological stage it has reached in its evolutionary progression, and that the rate and direction of natural change provide a yardstick to assess impacts induced by humans. However, as Roy emphasized, factors influencing estuary development include (1) inherited factors, mainly of a geological nature, that control the size and shape of the basin and the nature of the sediment supply, and (2) contemporary factors of a process nature (such as tides, river discharge, waves, etc.) that influence modes of sedimentation, hydrodynamics, and the biota. This prompts the question: To what extent are structure and biodiversity related?

Mann (1982) observed that, in general, estuaries are more productive than adjacent shelf systems, bringing up the question of nutrient flushing. That is, estuaries tend to act as nutrient traps. Many are enriched by pollution; the Hudson is a spectacular example of enrichment of a large shelf area well beyond its mouth. Within 600 km² of sea at the apex of the New York Bight, phytoplankton production amounted to about 310 g C/m²/day, compared with only 100 g C/m²/yr at the edge of the shelf. Mann and Lazier (1991) also noted that the dynamics of coastal waters, including estuaries, are made complex by: (1) shallowness, resulting in relatively mixed water that may extend to the bottom, and dead biological material that may accumulate to release nutrients that are carried rapidly to surface waters; (2) tidal currents that create turbulent mixing, which has especially marked effects on food particles, fertilization of planktonic eggs, and larval dispersal; and (3) barriers to convection imposed by coastlines, meaning that wind drives surface water away from the coast, and upwelling is the only way for it to be replaced, bringing nutrients to the surface. The question here is: To what extent are enrichment and/or pollution and circulation related to biodiversity?

Turning to larval transport, a variety of organisms have adapted to the seaward flow of low-salinity water and a compensatory landward flow of bottom water in estuaries. Organisms can make vertical migrations to maintain themselves in the estuary, or to enter or leave it seasonally. For example, estuarine larval transport and retention mechanisms are evident on two scales: circulation patterns on a large, regional scale and small-scale, local water motion. There is evidence that oyster larvae (Crassostrea virginica) rise into the water column to be carried upstream, and that this is cued by increasing salinity associated with increasing upstream flow; larval of the blue crab (Callinectes sapidus), on the other hand, occur in maximum numbers in surface waters at the mouth of Chesapeake Bay at night as the salinity falls on the ebb tide (Boicourt, 1982). From this and other evidence, it has been concluded that the crab larvae develop offshore, then reinvaed as megalopa larvae or juveniles. Fishes have also been shown to vary their depths, some rising into surface waters during flood to remain in the estuary, and others doing the opposite to be taken out to sea. Thus, many invertebrates and fishes utilize the two-layered estuarine structure for dispersal, and this may not be entirely passive, as has often been assumed. Despite some improved knowledge, Boicourt's conclusion is still pertinent, that the larval transport and retention problem 'stands at the state of the art in both physical and biological fields.' The question is: Does recruitment depend on return or retention (in the strict sense) as the operative process, and to what extent do larvae determine their own fates?

As another example of the importance of functional diversity, juveniles of the five species of Pacific salmons (Onchorhynchus spp.) vary in time spent in estuaries, but for all of them a high proportion of their prey tends to be detritus feeders (Healey, 1982). This means that the configuration of the estuary and the efficiency of entrainment of detrital matter are important for juvenile salmon habitat. Retention of detritus is enhanced by restricted exchange with the ocean and low bed-load transport. Marshes and submerged aquatic vegetation are efficient detritus traps, and these habitats also shelter salmon from predation. Thus, it may be hypothesized that the complex of intertidal marshes, tidal creeks and secondary river channels, lower intertidal and subtidal weed beds, and basin morphology all contribute to the carrying capacity of the estuary for young salmon, and that the appropriate configurations must be conserved if salmon production is to be maintained. The question here concerns how the complexity of the land–seascape enhances biodiversity, and how this may operate differently for closely related species.

From these examples, it is apparent that, insofar as ecosystem functioning is concerned, the addition or
Crassostrea virginicus

that of the oyster reef, formed by the eastern Oyster, and mud. Chesapeake Bay’s one major hard feature is erosional and whereby the bottom is largely soft sand wherein the dominant processes are sedimentary and case for many estuaries associated with coastal plains, origin is that of a drowned river valley. This is the Chesapeake Bay is one of Earth’s largest estuaries. Its regimes, autonomy may be incorrectly assumed. ever, the simple fact is that under many management domain of a given size is not a simple endeavor. How-

storms). Obviously, the elucidation of autonomy for a

dynamics of the contributing watershed and adjacent shelf (e.g., tides, currents, flushing, river inputs, storms). Obviously, the elucidation of autonomy for a
domain of a given size is not a simple endeavor. However, the simple fact is that under many management regimes, autonomy may be incorrectly assumed.

V. A CASE STUDY: THE CHESAPEAKE BAY

Chesapeake Bay is one of Earth’s largest estuaries. Its origin is that of a drowned river valley. Thus is the case for many estuaries associated with coastal plains, wherein the dominant processes are sedimentary and erosional and whereby the bottom is largely soft sand and mud. Chesapeake Bay’s one major hard feature is that of the oyster reef, formed by the eastern Oyster, Crassostrea virginicus.

Many scientists have observed the drastic decline of oysters and of oyster reefs during the past hundred years and more, and the associated ecosystem effects. From a structural point of view, oyster reefs represent a unique and dominant biogenic structure of the Bay. Their distribution and ecological importance during the mid-1800s were analyzed by McCormick-Ray (1998). Their loss would be expected to have extensive repercussions on biological, hydrological, erosional, and sedimentary patterns and processes, all of which can have major influences on biological diversity. Indeed, history has borne out this conclusion. For example, a review by Rothschild et al. (1994) stated that “considerable concern is voiced regarding Chesapeake Bay water quality and the effects of disease on oysters” and that “the effects of a diminished oyster population abundance cer-
tainly must have changed the ‘ecology’ of Chesapeake Bay, and these effects must have become evident at the time of maximum stock decline (1884 to 1910).”

To understand the ecosystem effects of the oyster and oyster reefs better, one must begin at the regional scale, wherein the coastal zone is conceived as a nested hierarchical system (Ray et al., 1997). The regional scale is that of biogeographic and physiographic provinces. The mesoscale is represented by major regional subdivisions, such as watersheds, estuaries, coastal islands, lagoons, and coastal–ocean fronts that separate major marine regimes. The smallest scale is that of the inter-

tacting mosaics of land–seascapes, for example, wet-

lands, hard and soft bottoms, and water masses that are distinguished by salinity, temperature, and density. The oyster reef represents this latter scale.

This hierarchy is illustrated in Fig. 1, which indicates top-down “controls” and bottom-up “feedbacks” and which places estuaries in a central role. First, the bio-

geographic province (and/or “region”) is an area whose limits are defined by the relative homogeneity of the biota. For example, the traditionally accepted bound-

aries for the Virginian Province are Cape Cod, Massa-

chusetts, to Cape Hatteras, North Carolina. These capes are significant points of deflection for major ocean cur-

rents, principally the warm, north-flowing Gulf Stream and the cold, south-flowing Labrador Current. At these capes, dramatic changes in coastal characteristics, such as water temperatures and circulation patterns, occur and these physical features play major roles in de-

termining the ranges of the biota. One major feature of the Virginian Province is the presence of very large estuaries, such as the Chesapeake and Delaware Bays.

Species’ ranges respond to these large-scale attrib-

utes, as well as to species’ physiological and behavioral adaptations. Fishes are a case in point. Of the almost 1100 East Coast fish species, 536 species presently occur in the Virginian–Carolinian region (Ray, 1997; Ray et al., 1997). Estuary-dependent species are drawn from this species pool. As discussed earlier, “estuary-

dependent” has usually been interpreted very broadly. C. R. Robins and I re-examined this matter and con-

cluded that occurrence and even abundance of fishes in estuaries do not necessarily infer “dependence.” Rather, we determined that a species must be truly “obligate” in an evolutionary, adaptive sense for this definition to apply; that is, if estuaries were removed,
FIGURE 1 A hierarchical model of coastal zone relationships, showing top-down “controls” and bottom-up “feedbacks” of coastal zone interactions, involving levels from biogeographic provinces, to estuaries, to the oyster reef. The biogeographic province provides the species pool from which estuaries may draw “estuary-dependent” representatives. This biota is influenced by the morphometrics of individual estuaries, leading to different species communities among the estuaries in a biogeographic region. The oyster is a “keystone” species both biologically and ecologically, as the reefs it builds influence the morphometrics of the estuaries in which it occurs. Over-harvesting of oysters in the Chesapeake Bay, and elsewhere, has had major effects on estuarine function, structure, and probably biodiversity as well. (From Ray et al., 1997.)

“dependent” species would be at risk of significant depletion, even to the point of local or regional extirpation. According to this definition, we determined that 151 species (27% of 556 species) qualify as “estuary dependent,” less than has been assumed in the past, but still a significant part of the total. This figure is remarkably consistent with the results of Able and Fahay (1998: see Section III). A principal components analysis of the ranges of these species resulted in four assemblages. Figure 2 shows these assemblages and demonstrates that so-called “faunal breaks” between provinces must be viewed as gradients, and not as “boundaries” in a rigid sense.

FIGURE 2 A principal components analysis of the ranges of 151 Carolinian and Virginian estuary-dependent species revealed the following assemblages: Component I = Virginian, Component II = Carolinian, Component III = tropical, and Component IV = boreal. These four assemblages overlap, as would be expected. (From Ray et al., 1997.)
This finding brings up the following question: How might changes in estuaries, human-caused or not, influence the composition of these fish assemblages? For insight into an answer, we must examine the dynamics of estuaries themselves. Many factors interact to characterize an estuary. Among these are drainage area, tides and mixing, estuary area, depth, dimension, water column stratification, floods, habitat types, and many others. A principal components analysis (Ray et al., 1997) revealed five components that may influence biological diversity: estuarine dimensions, dominance of marine processes, co-dominance of marine and freshwater processes, fjord-like attributes, and surface area. The interplay of these factors may be used to classify estuaries into the following types: (1) those that are long and wide with extensive catchment areas; (2) large, embayed, well-stratified estuaries with extensive seawater zones; (3) marine-dominated, deep, and well-stratified estuaries; (4) long and narrow, fjord-like estuaries, with large tidal prisms; and (5) estuaries with large surface areas. Chesapeake Bay falls somewhere between the first and second categories.

It seems reasonable, from what we know of the natural histories of the biota, that these estuarine types would be expected to host different communities of species, and further that different disturbance regimes would be expected to affect these estuarine types and their species’ communities differently. The conclusion seems obvious that biotic communities will differ among estuaries and that seasonal or weather-related changes in salinity and other factors will be reflected in the variability of biotic patterns. Furthermore, because the great majority of estuarine fishes, in particular, also occur over the continental shelf, fluctuations of estuarine fish communities would also be reflected, up-scale, by shelf-fish communities.

This approach offers a series of environmental top-down “controls” over biodiversity and ecosystem function. But this can not totally explain what might be the consequence of bottom-up environmental alterations. That is, the prediction of biodiversity and faunal dynamics requires that the response of the organism to the environment at different scales and the modifications the organism may make to the environment both be made explicit. For example, Fig. 1 indicates that the decline or removal of a species or a local structure, in this case oyster reefs, will influence the total biological diversity of the system by influencing environmental conditions through environmental feedbacks.

For the Chesapeake Bay, and many other Virginian–Carolinian estuaries, oysters are especially critical because they form reefs, which influence biodiversity at many levels (McCormick-Ray, 1998). The location of these reefs is not accidental. Their formation depends on the geometry of the estuarine basin, tidal stream channels and meanders, and other factors. Furthermore, oyster reefs influence estuarine development, sedimentation, and water clarity, and thus the formation of habitats (e.g., submerged aquatic vegetation, marshes, soft bottoms, and hard bottoms) for a host of organisms. In sum, the eastern oyster appears to be a classic example of a “keystone” species at the level of the ecosystem. Structurally and functionally, individual oysters and the reefs they build strongly influence species diversity and productivity. Additionally, the distribution of oyster reefs may be of fundamental importance to development of the estuarine land–seascape.

Another type of feedback concerns the fact that most species exist as a number of separate populations that mix together as one or more “metapopulations.” For example, an estuary-dependent species, such as menhaden (Brevoortia tyrannus), forms populations in individual estuaries, and these populations assemble over the shelf to form one or more metapopulations. Furthermore, these metapopulations join those of other species and become part of the shelf “metacommunity,” as illustrated in Figure 3. It follows that fluctuations of any one metapopulation within any one estuary will affect the total “metacommunity” to a greater or lesser extent (Ray, 1997). This form of biodiversity concerns community composition, not necessarily the presence or absence of individual species, and is strongly affected by functional alterations of estuaries. The conclusion is that at the scale of the large, regional ecosystem, each estuary may be conceived in terms of the sum total of individual oysters and the reefs they build strongly influence species diversity and productivity. Additionally, the distribution of oyster reefs may be of fundamental importance to development of the estuarine land–seascape.

The concepts presented in the case of the Chesapeake Bay suggest controls and feedbacks among organisms and the environment at several scales, in which one fundamental factor seems clear. East Coast estuaries have been perturbed in many ways, but one of the most dramatic for the Chesapeake Bay has been the depletion of oyster reefs and the practical eradication of their functional ecosystem role. Although data are lacking that would explain beyond doubt what changes have been perpetrated by the oyster’s decline in Chesapeake Bay, it seems apparent that, at the very least, the oyster reef’s demise has had a marked effect on the distributions of estuarine species, not necessarily because the
FIGURE 3 The concept of estuarine metapopulations and shelf meta-communities. Oyster reef metapopulations influence estuarine morphometrics and biodiversity. Consequently, the fish biota of various estuaries influence the fish metacommunity of the shelf.

reef is required habitat, but because of its functional importance to the Bay as a whole. It is possible that these effects may have cascaded up-scale to the adjacent continental shelf.

VI. FUTURE CHALLENGES

I make three points in conclusion. The first concerns the need for greatly increased attention to the natural histories of estuarine and shelf species. The natural histories of these organisms underlie both theory and management practice. The minimal requirements for informed conservation and management are descriptions of species’ life histories in the context of their environmental relationships.

Second, many estuarine organisms range widely and form metapopulations over the shelf, as components of estuary–shelf communities. Thus, the minimal scale for sustainability of biodiversity becomes that of the biogeographic region. Quantitative, landscape-level descriptions of the regional coastal zone, including estuarine habitats, are a necessary prerequisite for conservation and management.

Third, it has become a truism in ecology that no one scale adequately describes ecosystem phenomena. Rather, the interaction among phenomena on different scales must become the centerpiece of research and management. This strongly suggests that explanations for fluctuations in biodiversity, including those within biotic communities and at regional scales, will continue to be obscure until multiscale ecosystem functions are better understood. Ecosystem management is the logical outcome of interdisciplinary, multiscale knowledge. This recognizes that understanding the ecology and diversity of coastal zone biota depends in large part on understanding land–sea and estuarine interactions, and also on the joint application of metapopulation and land–seascape theory and methods.

The National Research Council (NRC, 1995) stated that a major future research objective is “to understand the patterns, processes and consequences of changing marine biological diversity by focusing on critical environmental issues and their threshold effects, and to address these effects at spatial scales from local to regional.” This objective cannot be met absent a specific consideration of estuaries as major, scale-dependent pathways of biotic and abiotic interchanges. Estuarine biodiversity, structure, and function have been severely modified by humans around the globe. Nevertheless, many estuaries remain either good candidates for restoration or relatively rich, productive, and resilient. Documentation of impacts is severely hampered by lack of long-term baseline information, inadequate assessment of biodiversity, lack of trained taxonomists, and difficulty in sampling.

Nevertheless, an extensive estuarine literature is now available, and it illustrates that control of pollution, development, excessive natural resource extractions, and changes in ecosystem function urgently need to be addressed. Problems may not be eliminated, only ameliorated, but increased understanding is essential for the future sustainability of estuaries. Carricker (1967) put the matter boldly three decades ago: “There is consequently an urgency to study estuaries before unenlightened defacement obliterates them and before it becomes expedient to investigate them primarily as outdoor pollution laboratories.”
See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • FRESHWATER ECOSYSTEMS • INTERTIDAL ECOSYSTEMS • LAKE AND POND ECOSYSTEMS • MARINE ECOSYSTEMS • RIVER ECOSYSTEMS • WETLANDS ECOSYSTEMS

Bibliography


I. Conservationists’ Ethical Consensus

There is widespread disagreement in modern societies regarding the proper human relationship to the rest of the natural world. Conservationists and conservation biologists, however, mostly agree on the following ethical principles: the diversity of organisms is good; ecological complexity and natural evolution are good; the untimely extinction of populations, species and biological communities is bad; biological diversity has great value both to people and in its own right; and human beings have both strong altruistic and strong self-interested reasons for preserving biodiversity.
Conservationists justify these principles in very different ways. One may value biodiversity because it is God's creation and testifies to God's glory; another because it is the natural culmination of evolution and global forces that have occurred over hundreds of millions of years; a third because of its value to science, perhaps humanity's noblest pursuit; a fourth because of its beauty and the enjoyment that current and future generations may take in it; a fifth because of its intrinsic value grounds a strong duty to protect it, regardless of human interests; a sixth because preserving biodiversity helps preserve human life-support systems. Such justifications may be more or less anthropocentric—centered on human interests. However different their ultimate justifications, conservationists tend to accept these ethical principles and work with others to preserve biodiversity.

Difficult ethical and practical issues emerge when we ask two further kinds of questions. First, what do these general principles entail? Given a high valuation of nature and a sense of personal responsibility to help protect it, what exactly is required of us as scientists and citizens? Second, how does our concern to protect biodiversity mesh with our obligations to humans? What if they clash? Such questions come up in numerous practical contexts and must be answered, although we cannot wait for certainty or full consensus in order to act.

II. CONSERVATION GOALS: HEALTH, INTEGRITY, SUSTAINABILITY

For much of the twentieth century, public and private land management goals have been primarily economic. It has become increasingly clear, however, that an exclusive focus on economic productivity leads to environmental degradation: pollution, the extinction of species, and the creation of a progressively simplified landscape. In response, laws have been passed directing government land managers to preserve the health and integrity of ecosystems. A focus on short-term economic returns also tends to undermine the long-term economic productivity of biological systems, such as forests and fisheries. Recognizing this, governments have begun to promote sustainability as an ideal in the use of natural resources. Conservationists hope that health, integrity, and sustainability specify goals concrete enough to guide us in our actions while commending themselves as basic values that a wide variety of people will share.

A. Health

Responding to widespread criticism of forestry practices, Canada's Provincial Forest Ministers agreed in 1992 that "our goal is to maintain and enhance the long-term health of our forest ecosystems, for the benefit of all living things both nationally and globally" (Westra and Lemons, 1993). Government agencies around the world have similarly embraced the notion of land health. Participants in a major symposium on ecological health defined it as follows:

An ecological system is healthy and free from "distress syndrome" if it is stable and sustainable—that is, if it is active and maintains its organization and autonomy over time and is resilient to stress. (Costanza et al., 1992)

Just as a healthy human being is free from disease and able to perform his or her characteristic functions well, a healthy forest or stream is free from air or water pollution, siltation, or invasions of exotic organisms. Such "stresses" impede or radically alter the systems' natural functions (such as photosynthesis or net primary productivity) and structure (often eliminating sensitive species and simplifying biological communities)(Table I).

Healthy ecosystems provide both natural biodiversity and a wide array of products and services valued by humans. For this reason, economic productivity (as measured in board feet, fish caught, or crops yielded) can sometimes stand as one measure of overall ecosystem health. Rapidly declining fish catches in Lake Victoria, due to the introduction of the Nile perch, signaled a decrease in the system's health. However, economic productivity may sometimes be increased by a radical simplification of an ecosystem, as when diverse, natural forests are converted to single-species, even-aged pine plantations, or when natural prairies are converted to cattle ranches (Fig. 1). Such simplified ecosystems are not necessarily economically productive in the long run, but if they are, then whether they are seen as unhealthy depends on whether preservation of native biodiversity is part of our definition of land health.

If simplified ecosystems are managed well, they may stay healthy for humans and a reduced biota, remain stable and resilient to stress, not pollute surrounding ecosystems (e.g., not dump excessive silt or fertilizer in streams), and continue to perform valuable ecosystem services. Because some relatively intensive land use is necessary for human survival and this can be done better or worse, it makes sense to apply concepts of
TABLE I

<table>
<thead>
<tr>
<th>Types of stress</th>
<th>Nutrient pool</th>
<th>Primary productivity</th>
<th>Size distribution</th>
<th>Species diversity</th>
<th>System retrogression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvesting of renewable resources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Aquatic</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Terrestrial</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>Pollutant discharges</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Aquatic</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Terrestrial</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>Physical restructuring</td>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>Aquatic</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Terrestrial</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>Introduction of exotics</td>
<td></td>
<td></td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Aquatic</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Terrestrial</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Extreme natural events</td>
<td></td>
<td></td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Aquatic</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Terrestrial</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>−</td>
</tr>
</tbody>
</table>

Note: Signs (+ or −) indicate direction of change compared with normal functioning of relatively unstressed systems. An asterisk indicates that a characteristic response was not sufficiently determined. Rapport et al. (1985).

land health to simplified ecosystems. At the same time, we must recognize that preserving as much biodiversity as possible in managed systems (croplands, lakes, forests) is desirable and that excessive simplification across a large landscape leads to extinction of species and loss of characteristic natural communities. Ecosystem health must be supplemented with the ideal of ecosystem integrity—at least in some areas (see the next section).

We should also remember an important difference between human health and ecosystem health. Individual human beings grow old, decay, and die, and while we fight this we also accept it. In contrast, we want healthy ecosystems in perpetuity, for their own good and the good of our descendants. Thus ecosystem health must also be supplemented with the ideal of ecological sustainability (see Section II,C).

B. Integrity

Amendments to the U.S. Clean Water Act of 1972 first set ecological integrity as a management goal by calling for the restoration and maintenance of "the chemical, physical, and biological integrity of the nation's waters" (Costanza et al., 1992). Rolston (1994) noted that "both integrity and health are combined fact-value words. Both convey the idea of wholeness and of unbroken functioning." But ecosystem integrity implies a greater measure of freedom from past and current human manipulation and a closer approximation to natural structure and functioning than does ecosystem health. Karr and Dudley (1981) defined ecological integrity as "the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having a species composition, diversity, and functional organization comparable to that of natural habitats of the region."

Integrity is a key environmental ideal because it encompasses the full preservation of biological diversity, including individual organisms, species, natural communities, and the ecosystem processes that have created and sustained them. It has taken a long time for this ideal to be acknowledged, even in supposedly fully preserved areas. For example, wolves, mountain lions, and other predators were routinely shot, trapped, and poisoned in U.S. national parks up until the 1960s. It is only in the last ten years that wolves, once extirpated, have been reintroduced into Yellowstone National Park. For most of this century, fires were suppressed throughout the national parks; they were perceived as dangerous to both people and forests. Over time, however, fire ecologists documented the historical role of fires in creating the landscape and preserving fire-dependent species of plants and animals. In response, land managers have introduced controlled burning into parks and allowed some lightning-ignited fires to burn. Improved ecological understanding and an attempt to distinguish between what is good for the ecosystem and what is
comfortable or familiar for people have shown that wolves and fire belong in the parks.

The fact that ecosystems are not as strongly integrated, clearly bounded, or stable as individual organisms complicates attempts to specify ecological integrity in particular cases. It is sometimes difficult to stipulate which outside influences represent assaults on integrity rather than mere changes. Recent study suggests that many ecosystems are relatively loose assemblages of species; that these assemblages may be relatively recent and transient creations; and that even left to themselves many natural communities will not necessarily reach a
particular, invariant climatic state but may instead reach any one of a number of more or less stable states, or no stable state whatsoever. If species assemblages are always changing naturally, why distinguish species that are extraported or introduced by humans, as when conservationists reintroduce wolves in the Rockies or eradicate wild boars in the Great Smoky Mountains? If there is no one natural end point to succession, why assume that human-created early-successional stages (fields of shrubs after clear-cuts) have less integrity than naturally occurring late-successional stages (old-growth forests)?

Still, ecosystems typically go through characteristic successional stages and support characteristic (if not invariant or exact) species assemblages. Though natural species assemblages change, this usually occurs on a time-scale that allows for much stability, the development of detailed interactions between organisms, and the increase of biological diversity at the landscape level. In contrast, human-induced changes typically decrease biodiversity and always lead to a landscape that is partly our creation. This loss of independence arguably marks a qualitative change in an ecosystem’s natural history and a corresponding loss of ecosystem integrity. Applying this complex concept to particular ecosystems thus involves attention to scale, knowledge of an ecosystem’s particular history, and comparison with the structures and functioning of similar ecosystems. Maximizing ecological integrity involves balancing the sometimes conflicting goals of freedom from human interference and preservation of historical natural communities.

Despite these complications, the desire to restore degraded ecosystems demands robust conceptions of ecosystem integrity and health. Ecologists attempting to restore degraded pasture-lands to native prairie in Wisconsin, worked-out strip mines to forests in Appalachia, or drained lands back to wetlands in Florida must set specific objectives (Fig. 2). As with land managers seeking to limit their effects on relatively pristine areas, restorationists have taken health and integrity, defined in relation to natural baselines, as goals for restored landscapes.

C. Sustainability

“Live Sustainably!” has become a rallying cry in the environmental movement. Like health and integrity, sustainability is a term that blends facts and values, helping us carry out our environmental convictions. At a minimum, sustainability involves preserving resources for future generations. But this minimum will not satisfy conservation biologists, who insist that part of what must be sustained is the full complement of biological diversity, for its own sake and for the benefit of humans.

Wide differences exist here. For instance, in 1987 the United Nations World Commission on Environment and Development (the Brundtland Commission) defined sustainable development as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs.” The focus here is wholly on human beings. In this definition the mass extinction of species is sustainable, provided future generations of people can meet their self-defined needs. R. Noss summarized the shortcomings of such narrow, anthropocentric accounts of sustainability as follows:

A failure of those who promote sustainability to consider environmental and social limits to growth, an unwillingness to address the unsustainability of the current human population, much less its expected growth; a reluctance to confront the implications of the lifestyles of average citizens of the more affluent societies . . . a failure to recognize the claims of other species to their share of the planet’s resources. (Westra and Lemons, 1995)

Contrast the Brundtland definition with Barbier’s more generous definition of sustainable development: “to maximise simultaneously the biological system goals (generic diversity, resilience, biological productivity), economic system goals (satisfaction of basic needs, enhancement of equity, increasing useful goods and services), and social system goals (cultural diversity, institutional sustainability, social justice, participation)” (Munda, 1997). Such a definition implies a different conception of human development and a more restrained treatment of the non-human world than has prevailed up to this point in human history. Satisfying the basic needs of all people and creating just and flourishing societies are part of this goal; limitless wealth creation and unbounded consumerism are not.

Conservationists are particularly concerned to keep the definition of sustainability based in ethics and biology and not solely in traditional economics. Economical based definitions often define sustainability in terms of indefinite economic growth, but in the crowded, fossil-fuel-propelled world of the twenty-first century, continued emphasis on economic growth will inevitably undermine ecosystem health and integrity worldwide. We cannot sustain increased consumption
III. MANAGEMENT AND NONMANAGEMENT

Setting aside parks or wilderness areas where human beings are not permitted to use resources has been conservationists’ most effective tool in conserving biological diversity. Here the results from conservation biology are clear. Wild areas preserve greater numbers of native species than do areas managed for agriculture or forestry. Large protected areas are more likely than smaller ones to preserve full complements of native species. Some species need large areas of habitat to preserve minimum viable populations. Some species tolerate little human disturbance. Larger areas are less likely to suffer “edge effects” that render otherwise good habitat unusable by certain species.

A dilemma arises, however, because in order to preserve certain wild species or communities, we may have to actively manage wild or semi-wild areas. Rare, threatened plant populations may have to be fenced off from
ETHICAL ISSUES IN BIODIVERSITY PROTECTION

high populations of browsers or new populations created from seeds or cuttings. Small, widely scattered populations of an endangered mammal may have to be gathered together to preserve genetic variety or create viable populations. Controlled burning may have to be carefully introduced into a forest or field to preserve fire-dependent species (Fig. 3).

Most conservationists accept reintroductions of indigenous species into wild areas and other manipulations that have as their goal the preservation or re-creation of indigenous flora and fauna. But they sharply distinguish actions that have this goal from actions that support further resource extraction or tourist comfort (road building, tree harvesting, construction of lodges and restaurants), which they continue to oppose in wild areas. Further, they see value in the absence of manipulation itself and tend to view the need for heroic measures in pristine areas as an indication of what has already been lost.

As a practical matter, hands-on management will often be necessary to preserve biodiversity. Even our largest natural areas are increasingly "islands" of diversity surrounded by development. These islands will likely lose species over time, as predicted by conservation biology theory and confirmed by empirical study; more native biodiversity may be preserved by augmenting populations and managing habitat. Other semi-wild areas will continue to be utilized by humans; these areas support wild species that may also need management in order to survive. Some manipulation will be needed, in perpetuity, if we are going to preserve as much biodiversity as possible.

It should be noted, however, that heroic measures and intrusive management often fail to preserve species. Skeptics also warn that a casual acceptance of manipulation may lead to a loss of biodiversity, if all sorts of human purposes are allowed to vie with nature preservation on our park and forest lands. Too often, management of nature becomes a substitute for management of ourselves. Rather than close popular trails or campgrounds in Yellowstone National Park, for example, state officials sanctioned the culling of hundreds of deer in the Quabbin Preserve in central Massachusetts to prevent overpopulation, overdevelopment, and overconsumption that make such biodiversity management necessary. By managing ourselves more wisely, we can limit the need to manage wild nature.

IV. BIODIVERSITY PROTECTION AND ANIMAL RIGHTS

Conservationists typically argue that preserving species and whole biological communities should take precedence over preserving individual animals. In practice, this means that exotic animals may be sacrificed when necessary to preserve native species or overpopulous native or exotic animals may be sacrificed to preserve land health. These positions have provoked arguments with animal rights advocates, who might seem to be conservationists' natural allies.

For example, U.S. government agencies judged the continued existence of the endangered plant Santa Barbara live-forever (Dudleya traskiae) to be more valuable than the common rabbits that had been introduced on its island home. The rabbits, which fed on the plant's fleshy leaves, were killed to stop the destruction of this fragile plant species. In another example, state officials sanctioned the culling of hundreds of deer in the Quabbin Preserve in central Massachusetts to prevent overbrowsing as a way to allow tree regeneration and protect water quality. Protection of the overall health of the forest and continuation of basic ecosystem functioning (forest regeneration, water purification) outweighed the interests of individual deer.
FIGURE 3 Conservation management: intervention versus leave-it-alone. (A) Heathland in protected areas of Cape Cod, Massachusetts, is regularly burned to maintain open habitat and protect wildflowers and other rare species. (Photograph by P. Dunwiddie.) (B) This old-growth stand in the Olympic National Forest in Washington is the result of many years of keeping human disturbance to an absolute minimum. (Photograph by Thomas Kitchin/Tom Stack & Associates.)
Such practices seem mere common sense to many conservationists, but many animal rights advocates argue that they are misguided, for two reasons. First, animals (at least some higher animals) have rights or interests based on their consciousness or sentience (their ability to feel pleasure or pain). This means that we cannot sacrifice them casually for other goals, such as a healthy forest. Second, species and forests do not have rights or interests, because these complex aggregates are too loosely organized (see earlier discussion). To sacrifice the real interests of a higher animal to the bogus interests of a species or biological community is thus ethically mistaken.

Conservationists answer that as we learn more about biodiversity, we realize that its greatest value lies not in individuals but in the wholes that those individuals help constitute. It is the species that persists and evolves rather than the individuals, which come and go. It is the natural community that sustains and generates new forms of life. While an ethic based on individual rights best specifies appropriate human interactions, a less individualistic ethic is needed to properly value the natural world.

This ethical shift seems to be necessary for effective conservation. However, animal rights advocates have given strong arguments for respecting individual higher animals such as dogs, cats, rabbits, and wild horses. Conservationists, who usually value species over individuals, may concede that individual animals should not be sacrificed casually, but only if it truly is necessary to preserve species or the health of biological communities. They might add that individual animals should always be culled humanely, that is, with a minimum of suffering. This would rule out most trapping and much sport hunting, although these activities are routinely defended on conservation grounds.

For example, Quabbin Preserve managers might carefully monitor deer populations and tree regeneration and authorize hunting where appropriate. All deer kills would have to be performed humanely, and in some years no hunts may be authorized. Whereas a traditional wildlife management position would authorize hunting game birds and trapping beavers, mink, and muskrats as long as this was done sustainably, a position responsive to animal welfare might not allow any of these activities, unless they could be justified with reference to the good of the forest or particular species.

Conservation and animal rights ethics both typically lead beyond anthropocentrism—the belief that only human interests matter—and they are usually complementary. For example, a concern for individual animals leads logically to vegetarianism, which prevents much animal suffering and bloodshed. But since ‘eating lower on the food chain’ is less energy and land-use intensive, it also helps conserve resources and preserve biodiversity. Limiting consumption and distinguishing between needs, comforts, and luxuries are precisely what are needed to carry out both ethics. Nothing prevents individuals from living highly consistent lives that incorporate both ideals. We may act to limit animal suffering and preserve wild nature so that the entire realm of biodiversity can flourish.

V. BIODIVERSITY PROTECTION AND HUMAN RIGHTS

Conservationists often argue that individual organisms, species, and biological communities have intrinsic value—that is, value in addition to their usefulness to human beings—that obliges us to preserve them. This means human interests should sometimes be sacrificed to preserve biodiversity. For example, a developer should not build a housing project that destroys critical habitat for an endangered songbird species—whether or not it is legal to do so. The continued existence of this species is more important than the developer’s profits. His less-than-vital interests should be foregone or pursued in ways that do not sacrifice something infinitely more valuable: a natural species. In many cases, however, preserving biodiversity and furthering human interests are complementary goals, especially if we take a broad view of human interests. Present and future generations have strong material, aesthetic, scientific, recreational, and spiritual interests in preserving biodiversity.

A. Owls and People

The coniferous forests of the U.S. Pacific Northwest are some of the most majestic on earth, with thousand-year-old trees towering over crystal clear streams filled with salmon. As commercial timber harvesting escalated after World War II, the big trees fell and species dependent on old-growth forests such as the northern spotted owl (Strix occidentalis caurina) (Fig. 4), the marbled murrelet (Brachyramphus marmoratus), and various salmon species (Oncorhynchus spp.) declined dramatically. In response, environmentalists staged logging blockades and tree sittings during the 1980s and suc-
Figure 4. (A) The northern spotted owl is an indicator species for old-growth forests of the Pacific Northwest, a habitat coveted for its rich timber stands. (B) Demonstrators protest government and corporate policies that destroy forests and threaten species extinctions.
The political battle was often billed as “owls versus people,” but such a simplistic formulation was misleading. Reduced harvests resulted in the loss of thousands of logging jobs, but some reduction was inevitable, given unsustainably high harvest levels based on a one-time cutting of thousand-year-old trees. Logging restrictions helped preserve and create jobs in fishing and tourism and enhanced watershed maintenance, flood control, and other objectives. A calculation based on the full spectrum of long-term human interests would likely have resulted in substantially decreased timber harvesting. Still, old-growth forests do have great economic value as timber, and preserving more in reserves rather than cutting the big trees and moving to high-yield forest plantations may indeed lead to lower profits for industry and fewer overall jobs.

Are such trade-offs ever justifiable to prevent the extinction of species or important biological communities? Forestry industry executives and loggers’ unions said no. The interests of a logger in supporting his family should supersede the interests of animals or trees. Logging restrictions to protect other species are wrong because they lead to real human hardship for loggers and their communities. Environmentalists insisted that more logging jobs and higher corporate profits did not outweigh the permanent loss of species and the old-growth temperate rainforest itself. They also noted that industry and government had made no efforts to preserve logging and mill jobs when technological changes, log exports, or overharvesting had led to job loss in the past (see Fig. 4).

These considerations suggest a solution to the dilemma: harness the economic system to maximize both human interests and biodiversity protection. Whether we focus on intrinsic or instrumental value, the human communities of the Pacific Northwest have an interest in preserving their remaining old-growth forests. Fortunately, the regional economy generates enough wealth to provide laid-off workers with guaranteed health care, generous severance pay, and job retraining for those who want it. The big trees could be preserved and displaced workers supported by their fellow citizens. Such a course would not reconcile all conflicting interests or solve all problems: economic “winners” would have to be taxed to support economic “losers,” some loggers would have to give up a valued livelihood, and some businesses would have to forego profits or even fold. But many of the costs of environmental protection would be borne by society as a whole, which benefits from the preservation of biodiversity.

### B. Property Rights

When environmentalists won suits to limit logging in the Pacific Northwest, the resulting bans affected both private and public lands. To some observers, the restrictions on private lands were unjustified. In their view, the right to own and control property is essential. Dictating to individuals or corporations what they can or cannot do on their own land is seen as an intolerable abridgment of that freedom. In devising a plan to continue tree harvesting in the region, the U.S. government focused preservation efforts on public lands, allowing maximum freedom to private landowners. Does private property ownership give owners unlimited control over their land, even if their actions contribute to the extinction of a rare or endangered species? Historically it often has, but morally it does not. Both the intrinsic value of species and duties to our fellow citizens to preserve a common biological heritage argue against such actions, regardless of how much profit must be foregone. Land ownership confers both rights and responsibilities.

What then of a small landowner who needs to cut an old-growth stand in order to send a child to college or keep the land in the family? Clearly as the need becomes greater, the ethical justification increases, even if some loss of biodiversity occurs. Hopefully, though, this small owner is not so poor that he must view his land solely as a money-generator. The ideal is some compromise in which the land is used in a responsible way that also preserves biological diversity.

Finally, it is important to distinguish between individual small owners and the large timber corporations that rival some national governments in the size of their landholdings. Arguments for preserving the freedom of small forest owners do not justify laissez-faire policies for large corporate landowners. Corporate managers’ personal freedoms are not at stake in the same way, and their management decisions have vastly greater effects on natural and human communities. No government that cares about preserving its citizens’ biological heritage will fail to regulate the environmental impacts of large corporate landowners.

### C. Wilderness in Less-Developed Countries

Recently, some writers have argued that biodiversity preservation is a specifically North American or West-
ern preoccupation, whose promotion in less-developed nations amounts to cultural imperialism. According to Ramachandra Guha, for example, wilderness preservation is inappropriate and unnecessary in these countries, whose peoples face more pressing environmental issues centered on meeting basic human needs. He argues that:

the setting aside of wilderness areas has resulted in a direct transfer of resources from the poor to the rich [in the less-developed nations]. Thus, Project Tiger, a network of parks hailed by the international conservation community as an outstanding success, sharply posits the interests of the tiger against those of poor peasants living in and around the reserve. The designation of tiger reserves was made possible only by the physical displacement of existing villages and their inhabitants; their management requires the continuing exclusion of peasants and livestock... transplant[ing] the American system of national parks onto Indian soil. In no case have the needs of the local population been taken into account, and as in many parts of Africa, the designated wildlands are managed primarily for the benefit of rich tourists. (Desjardins, 1999)

There is some truth in Guha's claims. Several of Project Tiger's reserves were built around old hunting preserves from which the poor had long been excluded, while others were set up on former state or communally owned lands and displaced numerous villages, causing real hardship to thousands. Grazing and tree cutting in the core areas of the reserves were prohibited, both to limit human-tiger conflicts and to preserve a more complete flora and fauna.

On the other hand, India's wild tigers were clearly headed for extinction when Project Tiger was undertaken: between 1900 and 1972, the year it was initiated, the tiger population fell from tens of thousands to just 1800 individuals. Strictly limiting tiger hunting had not halted this decline. Habitat conservation was clearly necessary to protect the tigers over the long term and, given the requirements of tigers and their inevitable conflicts with humans, this had to involve some displacement and restrictions on local inhabitants.

Should Project Tiger have been attempted? Guha suggests not: he believes it represented an unjust appropriation of the resources of poor people. At the same time, a focus on tigers went hand in hand with ignoring environmental problems that impinge far more directly on the lives of the poor—e.g., fuel, fodder, water shortages, soil erosion, and air and water pollution... [which are] far more pressing environmental problems within the Third World (Desjardins, 1999). Others disagree, arguing that habitat preservation and restrictions on human economic use of this habitat were necessary and appropriate responses to the imminent extinction of the Bengal tiger. This would have been a great cultural loss to the Indian people and an unjust destruction of an intrinsically valuable species. Many conservationists do not accept the extinction of species as a 'less pressing' problem than those mentioned by Guha, because they do not accept his exclusive focus on human interests, nor his narrow definition of those interests (Fig. 5).

Recently, wildlife managers in the less-developed countries have been experimenting with ways to give local people a stake in the success of wildlife and park conservation, with encouraging results. In Africa, these efforts have included disbursing a percentage...
of park revenues directly to local villages, increasing efforts to hire locals as guides and forest wardens, compensating for crop damage done by wildlife straying outside of parks, and giving local people input into protected areas management. In India, beginning in the late 1980s, the Ramnabinhore Foundation and other groups began work to promote sustainable development in the villages on the periphery of the Project Tiger reserves. This has included replanting denuded forests and improving fodder—to decrease pressure to graze and cut firewood within the reserves—as well as efforts to educate local children in natural history and conservation principles.

Justice demands that local people be treated with respect and their interests considered in all programs to protect wildlife and wildlands. Such respect and consideration will not solve all problems, however, particularly in a country as overpopulated as India, whose human population has increased from 280 million to 910 million in this century. In some instances, if we want to preserve wildlife, we will have to sacrifice human interests.

Overall, securing human rights and furthering essential human interests should strengthen efforts to preserve biodiversity. The United Nations General Assembly has affirmed that all human beings have fundamental rights to clean air, clean water, pure food, and a healthy environment generally. The measures needed to secure these human rights would also benefit wildlife and natural systems. Environmentalists have argued that future generations have a right to an undiminished natural heritage. Respect for this right would halt many environmentally destructive development projects. Natural areas are often threatened by remote governments and corporate managers, while their strongest defenders tend to be local inhabitants who know and love these areas. Securing political rights and economic justice for local inhabitants should help preserve these areas.

VI. THE IDEAL OF THE CONSERVATION BIOLOGIST

Four hundred years ago, modern science's great pioneer and prophet Francis Bacon located the primary value of science in its creation of a powerful technology for controlling nature. "The End of our Foundation is the knowledge of Causes, and secret motions of things," Bacon wrote in The New Atlantis, "and the enlarging of the bounds of Human Empire, to the effecting of all things possible." Without denying science's utilitarian benefits, conservation biologists attempt to move in a different direction, developing the knowledge needed to preserve nature rather than to change, control, or exploit it.

Both historical and contemporary examples attest that science can be carried out in this spirit. Aldo Leopold and Rachel Carson, Paul Ehrlich and E. O. Wilson, and many others from a wide variety of disciplines and backgrounds have combined contributions to science, literature, education, public policy, and conservation. The challenge for the scientist is to personally cultivate a loving knowledge of nature and to speak about nature in a way that promotes its celebration and protection (Fig. 6).

A. Professional Decisions

If conservation biologists are to help successfully preserve biodiversity, they must take on several active roles in addition to their scientific pursuits. For a start, they must become more effective educators in public forums. Conservation biologists often teach college students and
write technical papers addressing environmental issues, but they need to reach a wider range of people by also speaking in villages, elementary and secondary schools, parks, and neighborhood gatherings. They must spend more time writing articles and editorials for newspapers and magazines, as well as speaking on radio, television, and other mass media. Because tenure, promotion, and professional prestige often accrue for technical work and peer-reviewed science but not for more popular and disseminative work, conservation biologists may face difficult choices when deciding whether to devote time to the latter.

Similarly, research projects describing habitat requirements or methods of transplating an endangered species may have important conservation implications, but little potential to make basic advances in ecological or biological theory. Again, hard choices between professional glory and conservation usefulness may be necessary. As conservation biology becomes more established as a discipline, the system of rewards in place for practical conservation work will hopefully improve. Tenure and promotion committees at universities should acknowledge the value in community outreach and research that addresses local conservation issues.

Conservation biologists must also become politically active. Involvement in the political process allows conservation biologists to influence the passage of new laws to support the preservation of biological diversity or to argue against harmful legislation. Though much of the political process is time-consuming and tedious, it is often the only way to accomplish major conservation goals, such as acquiring new parks and reserves. Conservation biologists need to master the language and methods of the legal process and form effective alliances with environmental lawyers, citizen groups, and politicians. With their detailed knowledge of specific organisms and ecosystems, biologists are well placed to alert the public to the threats to biodiversity. They may also present this natural heritage in an appealing and inspiring way that leads to its preservation.

Such outreach and activism take time away from the pursuit of pure science. But many believe that at this point in human history, we no longer have the luxury to pursue knowledge for its own sake. To learn about biological diversity today is to learn about the threats facing it. These threats demand an active response.

Fortunately, there is room for many types of work within conservation biology. Field and laboratory biologists who perform detailed natural history studies and genetic analyses on endangered species are necessary; so are hands-on managers who put together species recovery programs and coordinate interagency partnerships to facilitate them; so are teachers and writers who popularize wild nature; and so are political firebrands who alert the public to biodiversity losses and corporate environmental crime.

B. Personal Choices

What about conservation biologists’ private lives? Should a conservation biologist own a sports utility vehicle, for example? It gets terrible gas mileage, yet it might greatly facilitate her field work. Should a conservation biologist take a spectacular eco-tourist vacation to Amazonia? The trip uses prodigious resources, yet he might gather information and take slides to educate audiences back home about important environmental issues. Then again, such reasoning may hide our real motivations: to drive in maximum comfort, to travel and enjoy ourselves. There is nothing wrong with comfort and enjoyment per se, but such high-consumption activities take a great toll on world biodiversity. Perhaps a conservation biologist should set a more restrained example.

Integrity is a key human virtue. It means living in conformity with our values and resolutely striving to toward worthwhile goals. As spokespeople for biodiversity protection, conservation biologists will make more converts if they practice conservation in their own lives. Above all, this means limiting their personal consumption of resources through owning fewer things, recycling, using public transportation, and taking other appropriate actions. However, it is not necessary to put on a hair shirt or engage in a quest for absolute purity. Showing that a life devoted to conservation can be stimulating and enjoyable is itself an important conservation message.

Should a conservation biologist dive into political debates? Give up time in the lab to meet with a group of concerned citizens? Drive a car to work? Eat meat? Have children? These difficult professional and personal questions must be faced by conservation biologists and other committed conservationists. Responses may legitimately differ, but we should not forget that the fate of world biodiversity depends, in part, on our answers.

See Also the Following Articles

Biodiversity as a commodity • Conservation biology, discipline of • Conservation efforts, contemporary • Environmental ethics • Human effects on ecosystems, overview • Property rights and biodiversity • Sustainability, concept and practice of • Wildlife management
Bibliography


I. Historical Development of Ethnobiology

II. Current Trends in Ethnobiology

III. The New Synthesis

GLOSSARY

analysis, emic and etic  Concepts derived from the linguistic terms “phonetics” (representing speech sounds by precise and unique symbols and by technical descriptions of articulation, as practiced by trained linguists) and “phonemics” (characterization of speech through a minimal number of symbols, typically recognized by the speakers of a language). By extension, etic refers to the external explanation of cultural knowledge and practice (such as the use of Linnean taxonomy or scientific nomenclature to describe local useful plants), whereas emic denotes the internal perspective of local people (e.g., ethnobiological categories and nomenclature).

economic botany  As originally conceived, a branch of applied botany that arose during the colonial period to identify and characterize economically important plants and the products derived from them. Currently, it is a scientific endeavor that seeks to document the properties of useful plants through agronomic, archaeological, ecological, ethnobotanical, genetic, historical, phytochemical, and other empirical approaches. It overlaps broadly with ethnobiology because both fields have witnessed a similar development in theory and methodology in recent years.

ecosystems, anthropogenic and natural  Communities of organisms and their environment formed either through human action or through natural processes. In practice, it is difficult to establish the extent to which an ecosystem is anthropogenic or natural, reflecting the current and historical impact of people on the environment.

ethnobiology  A term coined in 1935 that has been defined as the study of the reciprocal interactions between people and the biological organisms in their local environment and, recently, as the study of biological sciences as practiced in the present and the past by local people throughout the world. Many researchers consider that ethnobiology comprises numerous subfields, such as ethnobotany, ethnoecology, ethnoscience, and ethnozoology, but there is no consensus on this point.

ethnobotany  and ethnozoology  Approaches to studying the reciprocal interactions between people and the plants and animals in their local environment. This definition has been criticized as broad and open-ended, but it captures the common goals of analyzing traditional biological knowledge and assessing human impact on the environment. These approaches include subfields such as paleoethnobotany and paleoethnozoology, which evaluate archeological evidence on the past interactions between people, plants, and animals.
ethnoecology. Typically defined as the study of local knowledge and management of ecological interactions. Recently, some researchers have proposed an alternate definition, considering ethnoecology as an emerging field that focuses on local peoples’ perception and management of complex and coevolved relationships between the cultural, ecological, and economic components of anthropogenic and natural ecosystems. It is concerned with the interaction between knowledge, practice, and production, and it is oriented toward applied research on conservation and community development.

ethnoscience. Arose as a minor subfield of ethnography concerned with recording in great detail local peoples’ knowledge of biological organisms and the physical environment. Later, the term came to be used in a more restricted sense by cognitive and linguistic anthropologists to refer to local classification systems (as an object of study) and their semantic analysis (as a methodological approach). In France, the term is used to refer to ethnobiological studies in general.

indigenous, local, and traditional. Adjectives used by anthropologists, ethnobiologists, and other academics to describe people, practices, and knowledge. Indigenous denotes people (and their cultural practices and knowledge) who claim to be the original or long-term inhabitants of a particular place, in contrast to more recent colonizers. Traditional refers to established lifestyles, practices, and beliefs that guide cultural continuity and innovation—a definition that recognizes that traditions are always in a process of adaptation and change. Local, preferred by many researchers because it is the broadest and least value-laden term, indicates cultures that are found in a specific part of the world. It is commonly used to refer to people, whether long-term residents or recent arrivals (rural or urban), who make a living from the land and are knowledgeable about the biological resources in their environment.

ALTHOUGH THE. 100th anniversary of ethnobotany (coined in 1896), the golden anniversary of ethnobiology (first used in 1935), and the silver anniversary of ethnoecology (appearing in 1954) have passed, there is no consensus on the precise definition of these fields. This is explained in part because of their relatively recent origin and the current surge in their theoretical, conceptual, and methodological refinement. Disagreement over definitions is typical of multidisciplinary fields; in the words of Brent Berlin, ethnobiology “combines the intuitions, skills, and biases of both the anthropologist and the biologist, often in quite unequal mixtures.”

In one sense, ethnobotany, ethnobiology, and ethnoecology are new terms for old practices. People have been exploiting the usefulness of diverse plants, animals, and ecosystems since the dawn of humanity. Documentation of local people’s perception of the environment emerged slowly over thousands of years as scholars from many cultural traditions recorded local ways of classifying and using plants and animals. The onset of European colonization of Africa, Asia, the Pacific, and the New World gave added impetus to the study of local knowledge of tropical and temperate organisms and ecosystems.

Toward the end of the nineteenth century, academics began to use the prefix ethno- to refer to the way that local people view the natural world, in contrast to the perspective of natural scientists trained in universities. They coined terms such as “ethnobotany” (first used in print by Harshberger in 1896) and “ethnozoology” to describe these emerging fields of study that crossed the boundaries of natural and social sciences. Interest in traditional environmental knowledge continued apace in the early twentieth century, and in 1935 Castetter coined the term “ethnobiology,” setting as its agenda the systematic analysis of data collected by ethnobotanists and ethnozoologists to achieve a deeper understanding of local peoples’ knowledge and lifestyles. Economic botany gained importance as a parallel field focused on useful plants and the products derived from them. In 1954, Harold Conklin proposed the term “ethnoecology,” originally conceived as a holistic and integrated approach to understanding local ecological knowledge and practice on their own terms, even while drawing on the concepts and methods of diverse scientific disciplines. A focus on classification systems and the linguistic and anthropological methods used to analyze them gave high visibility to an approach called ethnoscience.

In the 1980s and 1990s, further development of these various lines of research gave rise to new definitions, innovative theoretical orientations, and sophisticated qualitative and quantitative methodological approaches applied to local knowledge of the environment. In addition, ethnobiology expanded beyond its original geographical borders as the field gained importance in countries such as China, India, and Mexico. There is currently a new synthesis emerging—as yet without consensus—that defines ethnobiology as the study of biological sciences as practiced by local people through-
out the world, comprising both empirical knowledge (savoir) and technical know-how (savoir-faire), and inclusive of subfields such as economic botany, ethnobotany, ethnology, and ethnozoology.

I. HISTORICAL DEVELOPMENT OF ETHNOBIOLOGY

Despite the insights provided by archaeological and historical linguistic studies, setting an even approximate date of the emergence of local biological knowledge is a matter of opinion. A detailed understanding of the natural world was key to the independent emergence of plant and animal domestication over a period ranging from 8500 B.C. in southwest Asia to 2500 B.C. in the eastern United States. However, environmental knowledge reaches even further back into history, when hunting and gathering dominated subsistence activities.

Some researchers would place the beginning of human ecological knowledge at the dawn of humanity, approximately 7 million years ago. Early human ancestors, who lived on the African continent 2.5 million years ago, apparently fashioned stone tools for harvesting and processing food, probably allowing them to adapt to new environmental conditions. It is widely assumed that humans have been observing natural phenomena, distinguishing between biological organisms and discovering their uses ever since the emergence of Homo sapiens approximately 500,000 years ago. The archeological record reveals that by 500,000 years ago, Cro-Magnons had developed technologies for construction, fishing, gathering, and hunting that were dependent on a detailed understanding of plants, animals, and other elements of the natural environment.

Just as no one knows exactly when ecological knowledge appeared on the cultural landscape, there is no clue when the original precursors of ethnobiologists came on the scene. The first critical observations of other peoples’ ways of perceiving nature are probably as old as culture contact itself. Because these observations went unrecorded, the origin of the study of traditional knowledge reaches even further back into history, when hunting and gathering dominated subsistence activities.

Some researchers would place the beginning of human ecological knowledge at the dawn of humanity, approximately 7 million years ago. Early human ancestors, who lived on the African continent 2.5 million years ago, apparently fashioned stone tools for harvesting and processing food, probably allowing them to adapt to new environmental conditions. It is widely assumed that humans have been observing natural phenomena, distinguishing between biological organisms and discovering their uses ever since the emergence of Homo sapiens approximately 500,000 years ago. The archeological record reveals that by 500,000 years ago, Cro-Magnons had developed technologies for construction, fishing, gathering, and hunting that were dependent on a detailed understanding of plants, animals, and other elements of the natural environment.

An overview of these early texts and later works reveals that ethnobiology and natural history have evolved—much like the biological species and ecosystems that are their focus—through a process of punctuated equilibrium. Certain historical periods are marked by an intensive effort to expand empirical knowledge of natural phenomena, often by incorporating local lore, whereas other epochs are characterized by an unquestioned acceptance of published works.

A. Early Scholars in Europe

Academics often trace their intellectual history to the era of Greek philosophers, who lived more than 2300 years before our time. This is when classical botany and zoology were brought into existence by scholars such as Aristotle, who sought to summarize all current knowledge about plants and animals in encyclopedic works. It was as part of this endeavor that scholars in Europe first made a systematic study of what local people knew about the environment. In part, Aristotle and other early naturalists such as Theophrastus—who, as author of De Historia Plantarum and other works, is considered the father of botany—rejected many local supernatural beliefs in their quest to understand the natural world. Simultaneously, they drew on commonsense explanations and empirical knowledge of local people when describing the classification and use of plants and animals.

Among the people who followed in the footsteps of these early naturalists is Dioscorides, a military physician born in Asia Minor in the first century B.C. He wrote De Materia Medica, a treatise on medicinal plants which was the standard reference of botanists, medical doctors, and other scholars in Europe for 1500 years.

Apart from drawing on previous herbals, Dioscorides learned much about herbal remedies by interacting with local people he encountered during his wide-ranging travels with the Roman army in the Mediterranean region. Pliny the Elder, a Roman scholar who was one of Dioscorides’ contemporaries, recorded extensive plant lore in his 37-volume encyclopedia called Historiarum mundi or Natural History. He devoted 9 volumes to medicinal plants, making frequent reference to traditional practices and knowledge.

B. The Doldrums of the Middle Ages

The documentation of local knowledge that marked the origin of biological thought in Greek and Roman Antiquity was much less evident in the Middle Ages in Europe. The decline of the Roman Empire virtually
halted scholarly research on natural history and resulted in the destruction of much existing literature of the epoch. Throughout the Middle Ages, Europeans based their studies of medicinal plants almost entirely on the works of Theophrastus, Pliny, Dioscorides, and other early naturalists. Physicians from across the continent relied heavily on De Materia Medica, often trying unsuccessfully to match the local flora to the approximately 600 Mediterranean species described by Dioscorides instead of documenting the popular knowledge of their own region. As anthropologist Scott Atran (1990) summarized,

After Aristotle, the practice of copying descriptions and illustrations of living kinds from previous sources superseded actual field experience in the schools of late antiquity. Well into the Renaissance, scholastic “naturalists” took it for granted that the local flora and fauna of northern and central Europe could be fully categorized under the Mediterranean plant and animal types found in ancient works. Herbs and bestiaries of the time were far removed from any empirical base.

C. The Golden Age of the Moors

Despite this stagnation on the part of European scholars, general knowledge of medicinal plants was enriched by the flow of information coming from the Arab world, particularly through Spain (Andalucia). Sicily, and North Africa. Although dedicated in part to translating the works of Aristotle, Theophrastus, Dioscorides, and other classical writers, Moorish and other medieval scholars in these regions pursued empirical research that they applied to practical ends in agriculture, astronomy, botany, mathematics, medicine, and other fields. Although there are records of Arabic writings on botany dating to the ninth century, it was particularly in the twelfth and thirteenth centuries that scholars became prolific in recording precise original observations on plant and animal biology, conducting experiments on agricultural crops, and attempting to classify plants systematically. One of the key early scholars was Abū-l-Kheyr Al-Ichābi, apparently the “anonymous botanist of Seville” of the latter half of the twelfth century who wrote the Umdat at-tabīb, a botanical treatise that contains detailed descriptions of the anatomy, habitat, and local names of plants. Maimonides, a Jewish scholar resident in southern Spain during the twelfth century, contributed works such as Book Explaining Medicinal Drugs and Treatise on Poisons. The thirteenth century brought the works of Rachid-ed-din Ibn Es-Sirri in the Maghreb (western Arabic region) by Abū-l-Abbas En-Nahāti—who documented local plants and their uses during his extensive travels in what is now Spain, North Africa, Syria, and Iraq—and later scholars such as his student Ibn Al-Bayan, author of the Treatise on Simples or Jami’ al-mufradat, which contains information on the synonymy (including Berber and other local names), description, properties, and uses of approximately 1400 species. Throughout this period, descriptions of specific plants were drawn from multicultural sources, including Berber plant knowledge and the traditional practices of Jewish pharmacists who lived throughout Europe and North Africa, passing their profession from one generation to the next. As the Moors were forced from Andalucia and other parts of Europe, Arab science fell into decline. Later botanical treatises, including the Ḥadīqat al-azhār of the sixteenth century Moroccan scholar and medical doctor Al-Wazir Al-Ghassani, were largely based on the Umdat at-tabīb and other early works.

D. Scholars from Other Cultural Traditions

A similar pattern of initial empirical discovery and cross-cultural learning mixed with centuries of uncritical acceptance of written works is evident in other cultural traditions. In China, the first scholarly studies of traditional biological knowledge are thought to date at least to the fifth century B.C., approximately 200 years before early Greek philosophers began recording their ideas about botany and medicine. During this epoch, the Chinese philosopher Confucius is said to have encouraged his students to study ancient knowledge, including traditional names of plants and animals. In the first or second century A.D., about the time that Dioscorides was alive in the Mediterranean, Chinese scholars published the first of many materia medica, referred to generally as bencao (from a combined term for tree and grass) in Chinese, which contain information on minerals, plants, and animals used traditionally in the treatment of illness. Medical doctors of the fifth
or Gangmu began in India during the sixth century B.C. and spread by traveling to the countryside to talk with people. He obtained not only by studying ancient texts but also information on more than 10,000 herbal remedies that many centuries before. His major work, the Bencao Gangmu or Compendium of Materia Medica, contains information on more than 10,000 herbal remedies that he obtained not only by studying ancient texts but also by traveling to the countryside to talk with people.

Ayurveda, a system of medicine which putatively began in India during the sixth century B.C. and spread to Sri Lanka by the third century B.C. and into Tibet by the seventh century A.D., was partially based on traditional knowledge. According to folklore, shepherds and forest dwellers familiar with the types and properties of medicinal plants first discovered the remedies used in this oriental medical practice. Their knowledge was discussed in various literary religious works called vedas (from the Sanskrit word for knowledge), which were apparently written in India approximately 3200 years ago, after millennia of oral transmission. Ayurvedic scholars later compiled additional empirical observations in a series of books referred to as the Nigh- ants, or Vedic glossaries. During the subsequent period of foreign domination and internal conflict that brought innovation and documentation of local knowledge to a standstill, these standard texts of ayurveda remained unchanged.

Other ancient written sources that document local biological knowledge were in part the product of culture contact and changes in political and economic dominance. In the New World, for example, the Aztecs broadened their own sophisticated knowledge of medicine and agriculture as they sought tribute and learned of new useful plants from the different Mesoamerican cultures they conquered. The Aztecs cultivated many newly discovered species in extensive highland botanical gardens tended by people from various geographical regions of Mesoamerica. The depth and richness of the preconquest indigenous knowledge of the natural world are demonstrated by scholarly works, including the Biduaras Manuscript, an illustrated herbal written in 1552 by two Aztecs who had been educated by Catholic missionaries. One author, Martín de la Cruz, was an indigenous physician who had acquired his medical knowledge empirically. The Mayas and Incas had similar literate traditions and they doubtlessly recorded some aspects of the ecological knowledge of the various ethnic groups they dominated at the height of their political power and cultural development. Many of these New World written sources of local knowledge were victims of the conquest, destroyed by overzealous missionaries and conquerors who wished to impose European culture, languages, and religion on the people of the New World.

E. The Renaissance and Exploration

During the sixteenth and seventeenth centuries, Renaissance botanists began to emulate the methods that Dioscorides had applied approximately 1500 years previously, bringing an end to the intellectual stagnation that characterized the Middle Ages in Europe. They carefully observed plants in the field and inquired about their local names and uses in Germany, Holland, Italy, and other parts of Europe. This experience served them well when faced with the influx of exotic species from areas of the world discovered and colonized by Europeans during this period. The diversity of biological organisms discovered by explorers stimulated Linnaeus, Darwin, and other natural scientists to formulate many of the concepts that are the building blocks of modern-day systematics and evolutionary studies. Although Linnaeus left some notes and sketches on the use of plants by local people, his greatest contribution to the future field of ethnobotany was the incorporation of notions of folk biology and nomenclature, including the concept of morphological affinity as a criterion for defining taxa, in the scientific classification of plants.

This Renaissance was the golden age of the European herbals. The shift from manuscripts (produced by hand) to wood-cut and metal-engraved herbals published in large numbers allowed new botanical knowledge to be disseminated widely. The quest to exploit local knowledge and economically important species which went along with coloniza-
tion inspired adventurers, missionaries, and natural historians to record their observations on traditional biological knowledge in many parts of both the New World and the Old World. As ethnobotanist Richard Ford (1978) described,

A rapid progression of expeditions came to North America to discover and to colonize, and the chronicles of adventure are a record of the utilitarian value of an unfamiliar landscape and the use the indigenous people made of it. Its economic potential certainly had priority to any interest in
attitudes about the land. The observations... provided the first natural history of North America and the bases for the beginning of ethnobotany.

From the sixteenth century onwards, researchers began to focus increasing attention on the biological wealth of tropical countries and the benefits it promised for Europeans. To this end, scholars drew on the knowledge of local people, who continually experiment with cultivated and managed species in anthropogenic ecosystems and wild plants harvested in natural ecosystems. Scholars consulted both written sources, such as Ayurvedic works and Chinese pharmacopoeias, and oral history to produce extensive encyclopedias of useful plants from around the world, ranging from the *Coloquios dos simples e drogas da India*, written by the Portuguese explorer Garcia ab Orta in 1563, to the 12-volume *Historia Malabarica* of Van Rheede published in the late 1700s and *A Dictionary of Economic Products from the Malay Peninsula* produced in 2 volumes by Burkhill in 1935.

Another notable example is the herbal of Rumphius, a seventeenth-century natural historian from Germany who spent nearly 50 years in Asia working for the Dutch East-Indies Company. Increasingly released from his administrative duties but still under the employ of the company, he focused his attention on studying useful plants, animals, and minerals in various regions that today constitute Indonesia. He provided descriptions of more than 700 medicinal or toxic plants, published posthumously in the six volumes of the *Herbarium Amboinense*.

**F. Creation of New Fields**

Although Rumphius and his counterparts approached natural history as a holistic phenomenon, scholarly activity in later centuries began to reveal a fragmentation of research into distinct disciplines, marking the beginning of reductionism in the sciences. Subdivisions of science became more clearly defined, and professional practitioners began to specialize in specific subfields. Theoretical science was increasingly applied to technological innovation, stimulating the growth of educational institutions and commercial enterprises. Theoretical science, including botany, zoology, pharmacognosy, and other fields, began to mature into distinct disciplines, each with separate methods and goals. Scott Atran (1990) characterized this as the “breakaway of science,” a time when natural historians began to leave behind commonsense descriptions of natural phenomena—drawn in part from local peoples’ perception and classification of nature—in order to embrace rigorous experimental methods. Social scientists began to focus on separate aspects of human society and culture, with
the consequent emergence of fields such as anthropology, linguistics, and sociology.

2. Economic Botany

Nineteenth-century botanists, who focused primarily on the utility of plants and only secondarily on local culture, began to refer to their approach as economic botany. The goal of their research has been to document local uses of plants and to organize the resulting data according to the global system of plant classification. They have produced detailed works on plants employed by local people for food, medicine, textiles, utensils, and many other purposes (Schultes and Raffauf, 1990).

Research on the commercial value and utility of plants expanded as botanists from the United States and Europe explored the New and Old World tropics in search of products that would increase the wealth of developed countries and the well-being of people in general. Today, economic botanists continue to search for marketable products in tropical forests and elsewhere, but they are increasingly interested in how the commercialization of these resources can contribute to resolving the poverty, malnutrition, and diminished social status of local people as well as spurring economic development in developing countries. An increasingly important offshoot of this enterprise is bioprospecting (a term derived from “biodiversity prospecting”), the search for useful and novel products (including chemical constituents of medicinal value) from plants, animals, fungi, and other biological organisms. Other economic botanists are concerned with developing theoretical and methodological approaches to understanding the subsistence and commercial value of plant resources, now and in the past.

3. Ethnoscience

While botanists were establishing economic botany, anthropologists and other social scientists were developing a different perspective. In the tradition of ethnography developed by anthropologist Franz Boas, ethnoscience emerged as a minor subfield dedicated to recording in minute detail local peoples’ knowledge of biological organisms and the physical environment. The subfield underwent a further transition in the 1950s and 1960s, when cognitive and linguistic anthropologists began to focus on the empirical categories, social rules, symbolic systems, and modes of behavior that reflect how local people perceive the natural world. These early ethnological studies formed the foundations for a new ethnoscientific approach that advocated rigorous analyses of ethnobiological knowledge, with particular emphasis on systems of ethnobiological classification. Thus, whereas economic botany emerged as a utilitarian practice firmly rooted in commerce and development (and later developed theoretical frameworks), ethnoscience arose as an intellectual endeavor oriented toward a deeper understanding of human culture and cognition.

4. Ethnoecology

Even though he is most associated with the development of the ethnoscientific approach, Conklin is credited with coining the term ethnoecology in 1954. Given the precedent set by terms such as ethnobotany and ethnoecology, it would be natural to assume that ethnoecology would refer to the study of local perceptions of ecological processes, such as nutrient cycling, vegetational succession, or the interactions between plants and animals. An increasing number of researchers propose a different definition, using the term to refer to local peoples’ perception and management of the complex and coevolved relationships between cultural, ecological, and economic components of anthropogenic and natural ecosystems. This emerging subfield, much as the broader field of ethnobiology, is concerned with the interaction between knowledge, practice, and production and is oriented toward applied research on conservation and community development. Mexican ecologist Victor Toledo stated that the aim of ethnoecology should be the ecological evaluation of the intellectual and practical activities that people carry out during their appropriation of natural resources.

Although the definition of ethnobiology includes a reference to knowledge and know-how (savoir and savoir-faire), for ethnoecologists the distinction is between an ethnobiological corpus, local peoples’ repertoires of concepts, perceptions, and symbolic representations of nature, and praxis, the art, science, and skill of appropriating nature and biological resources. The interrelationship between knowledge and practice is manifested in production, as people apply their intellectual understanding of nature to the everyday tasks of farming, gathering, and hunting for subsistence and commercial purposes. In order to understand these complex interactions, ethnoecologists seek to elucidate how the management of anthropogenic and natural ecosystems—and the biological organisms they harbor—has arisen through a process of coevolution between the environment, knowledge, technology, social organization, and values of local peoples.

Although attractive conceptually, the development of this conception of ethnoecology has been limited
by the lack of a unifying theoretical framework and a practical methodology. This distinguishes it from ethnobiology, which is developing a central organizing theory, an orientation toward hypothesis testing, and an increasingly elaborate set of qualitative and quantitative methods, drawn in part from ethnobotany, ethnozoology, and economic botany.

5. Ethnobiology

Clement (1998) proposes that the starting point for ethnobiology—as the field which integrates related approaches such as economic botany, ethnobotany, ethnozoology, and ethnoscience—is the 1860s, when the first designations for the field began to be used by American and European scientists. In a historical sketch that spans a period of more than 130 years, he discusses the origins, key theories, and methodological approaches of the main trends of ethnobiology. Although no such historical framework of a scientific discipline is without controversy and potential modification, Clement’s synopsis is a serious effort to provide a detailed historical analysis of ethnobiology.

Clement divides the development of the discipline into three eras and seven periods (Table I). The preclassical period, from 1860 to 1953, is dedicated to gathering empirical data on the uses of plants and animals from an etic perspective and to the first syntheses that begin to define the scope of the discipline. During the classical period (1954–1980), there is a shift to studies carried out from an emic perspective and a particular focus on ethnobiological classification. An increase in collaborative work between academic specialists and local people and the formation of professional associations of ethnobiology characterize the postclassical period, from 1981 to the present. Later in the period, there is an increased focus on the appropriation and management of plant and animal resources and a concern for application of research results to the resolution of environmental and social problems. This historical review provides an appropriate starting point for considering the current trends in basic and applied ethnobiological research.

II. CURRENT TRENDS IN ETHNOBIOLOGY

Recent studies in ethnobiology can be classified in three general, interrelated areas. Documentation and analysis of uses of plants and animals is the oldest aspect of ethnobiology. A focus on knowledge, which became especially popular in the 1950s and 1960s, characterizes the study of perception and classification of the natural world. In the 1980s and 1990s, attention shifted to local management of biological resources and the environment, often drawing on ecological concepts and methods. These areas are combined in various degrees in the following major trends in ethnobiological research.

A. Cognitive Mapping

Ethnobiologists have documented local peoples’ spatial conceptualization, including their ability to locate biological resources, discern landscape features, and identify different types of vegetation. This geographical literacy is linked to an aptitude for assessing the potential productivity of the environment, often through the recognition of plant species that indicate fertility or sterility of soils. This expertise derives in part from local peoples’ sophisticated perception of how the various elements of the ecosystem (organisms, soils, climate, topography, etc.) form an interdependent whole. In addition, it is derived from their classification of key geographical landmarks that are labeled by specific toponyms, or geographical place names. Local views of the landscape are being integrated into geographical information systems, which in turn are useful in creating biological resource maps and management plans for specific areas. When united with ethnobiological inventories and studies of local categorization of ecological succession, community mapping allows researchers to test hypotheses on whether the highest number of useful plants and animals come from primary or secondary forest, or anthropogenic versus natural ecosystems.

Local peoples’ mastery in selecting plant populations that yield the best fruit, the most potent medicine, or the best materials for construction is attributed in part to their grasp of the landscape. These mental maps also explain their adeptness at selecting the best place to cultivate the earth, create human settlements, or leave natural areas that maintain soil fertility, water purity, and other environmental benefits. Researchers have demonstrated that these skills can degenerate when local people find themselves pressured by economic needs to overexploit the resources available to them or are forced onto lands for which they are maladapted culturally.

The ability of some local people to integrate and recall complex information on the local environment often gives them special proficiency as stewards of community reserves and other protected areas. Their ability to assess the quality of useful biological resources and to locate areas where these organisms are found in
TABLE I
Features of Various Historical Periods and Stages in Ethnobiology*

<table>
<thead>
<tr>
<th>Period</th>
<th>Stage</th>
<th>Dates</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preclassic</td>
<td>Economic uses</td>
<td>1860–1899</td>
<td>Studies of biological resources and their utility carried out by researchers affiliated with major museums and universities, general lack of appreciation of the sophistication of local knowledge and subsistence systems from an emic perspective</td>
</tr>
<tr>
<td></td>
<td>Information</td>
<td>1900–1931</td>
<td>Greater empirical depth in research but continued emphasis on economic uses of plants and animals, better appreciation of complexity of local knowledge and use of plants and animals, especially as reflected in systematic attempts to record local terminologies, myths and beliefs, and knowledge of anatomy and behavior; emergence of comparative studies and standard methods</td>
</tr>
<tr>
<td></td>
<td>The first</td>
<td>1932–1953</td>
<td>Emergence of ethnobiology as a distinct field of inquiry and appearance of the first syntheses that delimit its scope; increasing distinction between economic botany and ethnobotany, with the latter emphasizing the systematic documentation of local knowledge and management of plants; continued lack of recognition of scientific aspects of traditional biological knowledge</td>
</tr>
<tr>
<td>Classic</td>
<td>Emic knowledge</td>
<td>1954–1980</td>
<td>Emergence of ethnoscience, leading to a focus on the organization of knowledge systems from the local perspective, with insights from linguistics and empirical anthropological methods; relegation of the study of plant and animal resources to secondary importance; beginning of interest in ethnobiological classification and appreciation of the scientific basis of traditional knowledge</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td>1981–1992</td>
<td>Focus on ethnobiological classification, including principles of categorization and nomenclature, and the analysis of correspondence between scientific and local classifications; accumulation of evidence for the scientific basis of local biological knowledge; increasing interest in ethnobiology beyond the United States and Europe, especially in Latin America and the Pacific</td>
</tr>
<tr>
<td>Postclassic</td>
<td>Associations</td>
<td>1981 to present</td>
<td>Production of major empirical works based on close collaboration between academic and local researchers, development of theoretical approaches beyond classification, including gender relations in resource use, cultural significance of plants, and historical reconstruction of ethnobiological knowledge systems, emergence of academic societies and specialized journals of ethnobiology, especially in developing countries</td>
</tr>
<tr>
<td></td>
<td>Resource</td>
<td>1993 to present</td>
<td>Publication of standard methods manuals, quantitative techniques, and innovative empirical studies; emergence of concern about applying ethnobiology to conservation and development; renewed interest in economic botany, including nutritional and medicinal benefits of plants, but incorporating novel theoretical and methodological approaches and informed participation by local people.</td>
</tr>
</tbody>
</table>

* Adapted from Clément (1998).
about the classification of specific cultural domains, practices, and production, ethnobiologists assess the value of anthropogenic and natural ecosystems and the rationality of resource harvesting decisions made by small-scale farmers and gatherers of forest products. Much of this research is carried out in 1-ha plots, agricultural fields, home gardens, and other measured study sites, resulting in quantitative assessments of sustainability and value. These methods allow researchers to evaluate the hidden costs of tropical forest destruction, the economic benefits derived from both subsistence and commercial use of wild species, and the environmental advantages of maintaining forest cover as a way of buffering local climate and preserving the purity of local air, water, and soil. Through these perspectives, conservation biologists, development specialists, and communities monitor the sustainability of current productive practices, propose new methods of managing fields and forests, and select new biological species that can be domesticated, cultivated, or gathered locally. Studies of local systems of resource management can also enrich the work of ecologists, who seek to restore the diversity and value of forest ecosystems damaged by mismanagement or natural catastrophes. In addition, evidence of long-term management of biological organisms and ecosystems reinforces local peoples’ claims for traditional resource rights, including just compensation when novel biological resources are commercialized.

C. Scientific Covalidation

A key activity of ethnobiologists is to understand the rationale behind the way local people interact with the natural environment. In research laboratories, scientists carry out a broad array of analyses that seek to corroborate the efficacy of local uses of plants and animals, ranging from the identification of active compounds in medicinal plants to appraisal of the tensile strength of natural fibers and assessment of the nutrient content of wild foods. In the field, ecologists assess how the yields of agroecosystems compare with those achieved by large-scale monocultures and to what extent traditional methods of wildlands management are sustainable. Ethnobiologists compare ethnobiological categories with scientific taxa, judging the extent to which local biological classifications correspond to biosystematics. This range of activities is aimed at revealing the complex interrelated sets of categories for plants, animals, soils, climates, vegetation, illnesses, food, and other cultural domains and natural phenomena.

Much attention has focused on describing universal similarities in the ways in which local people perceive the natural world. Many generalizations on the categorization, naming, and identification of plants and animals are now widely accepted, whereas others continue to provoke controversy, especially among anthropologists who place emphasis on cultural relativity or the uniqueness of each ethnic group. In particular, there is disagreement over why people are motivated to classify and name plants and animals, soils, and other natural features. Those who follow an intellectualist line of reasoning argue that there is a universal human tendency to categorize plants and animals according to their overall appearance or symbolic role. These researchers note that local people tend to group organisms with a similar morphology or behavior, regardless of their cultural utility. In a similar vein, some colleagues propose an ecological rationale, noting that some aspects of ethnobiological classification can be elucidated by reference to the role and interaction
of plants and animals in diverse ecosystems. Although there is an emerging consensus that classification is motivated by a combination of these factors, there continues to be much debate about which is the most important dimension.

Studies of ethnobiological classification contribute to understanding—in part by reference to the global system of plant systematics—how local people perceive and manage natural resources. When carried out in a participatory way, ethnobiological inventories provide an opportunity for local people and researchers to work together to document the distribution, management, and use of biological resources locally and globally. They produce the baseline data needed to produce bilingual and bicultural manuals that compare and contrast different ways of classifying, managing, and using biological organisms. These ethnobiological manuals contribute to applied programs of conservation and development by highlighting culturally significant species that local people are harvesting from the wild, managing in anthropogenic landscapes, or cultivating in gardens and fields. Often, these species are selected for use in initiatives that promote reforestation, sustainable harvesting of minor forest products, or cultivation of useful plants that contribute to the well-being and income of local people.

E. Knowledge Variation

There are significant differences in the way local people perceive and use biological resources. Ethnobiological studies reveal that some plants and animals are known by a majority of the population, but others are the domain of curers, the elderly, women, or members of another social group. There are significant differences in the biological knowledge of people living in separate communities, belonging to distinct ethnic groups, speaking different languages, or subsisting in diverse ecological zones. These differences can be explained in part by the fact that each person’s knowledge is correlated with sociological characteristics such as age, gender, occupation, education, social status, and zone of residence. Perception and management of plants and animals are also affected by each individual’s life experiences—for example, if he or she has suffered from a major illness, migrated to other communities, or worked as an apprentice to a plant specialist.

Awareness of these differences is essential when assessing the depth and breadth of traditional knowledge and the diversity of ecological practices in a particular area. It is now recognized that researchers run the risk of obtaining a biased perspective of local biological knowledge if they work with few informants. In contrast, consultation with a representative cross-section of local people can provide a relatively accurate portrayal of perception of the natural world. Ethnobiologists are analyzing patterns of agreement among different individuals, producing a quantitative method of identifying the most culturally significant plant and animal species, soil types, forest zones, and other elements of the landscape. These results ensure accuracy when preparing natural resource management plans and popular manuals produced with local communities.

They can also play a role in understanding the knowledge held by specialist user groups and how they are transmitting it from one generation to the next.

F. Mechanisms of Change

Another principle that has emerged in recent studies is that knowledge about the natural environment and ways of managing biological resources are not static. They change as people move from one region to another, as youth reinterpret what they have learned from elders, and as cultures come into contact with each other. Ethnobiologists are analyzing the dynamics of these cultural changes and assessing how they can enrich or impoverish local knowledge systems.

Archaeology, linguistic reconstruction, archival research, and oral history open a window to the past, allowing researchers to discover how ecological knowledge and resource management have evolved during decades and centuries of political, demographic, and economic change. Studies of historical ecology are revealing which species, ethnobiological categories, and ecological practices have persisted, changed, or disappeared over time, including those that have become widespread through borrowing across cultural boundaries. Research that focuses on world events in recent centuries allows historians to assess the impact of colonization and other forms of culture contact on local ways of perceiving and managing the natural world.

Many regions of the world are currently in an intense period of change characterized by not only the destruction of wildlands and loss of biological diversity but also the transformation of traditional biological knowledge. In some regions, young people are not learning what their elders know about the environment, particularly as traditional ways of using biological organisms and managing natural areas fade away. Specialized knowledge held by only a few curers or spiritual leaders is lost when no apprentice is found to carry on traditional medical or religious practices. This loss of knowledge is often linked to increasing contact with national
and international cultures and is often exacerbated as local people incorporate formal education, major religions, and migration into their lifestyles. Ethnobiologists are playing a practical role in assessing the extent of cultural transformation and searching for ways of promoting the survival of local ways of classifying, using, and managing natural resources. These actions are based on the assumption that participation in the joint management of ecosystems and resources stimulates local people to retain and build on empirical knowledge and practices acquired during the tens, hundreds, or thousands of years that they and their ancestors have resided in the region. Studying the evolution of ethnobiological knowledge elucidates mechanisms of cultural resistance, allowing us to understand why local ecological knowledge persists and how we can reinforce the mechanisms of resistance. These studies also highlight the impact that local people have had on the natural environment over time, often providing evidence that they are responsible for maintaining biological diversity and stimulating innovative ideas for conservation in the future.

G. Ritual, Religion, and Symbolism

Plants and animals play an important role in rituals and spiritual practices and are a recurring element in myths, legends, and stories. Some researchers assert that local people, because of their cosmology or understanding of the universe, relate to the natural environment in ways fundamentally different from that found in, for example, European or American culture. A central element in this conception is that traditional lifestyles link people to nature in a way that provides them with a special understanding of nature. A common way for outsiders to gain access to the world of ritual, magic, and religion of local peoples is through apprenticeship with spiritual leaders, often accompanied by the use of psychoactive plants.

Whether or not these generalized notions of differences between global knowledge systems and traditional knowledge are valid in all cases, it is undeniable that spiritual beliefs about the forest and associated taboos on the use of natural resources are important elements of conservation and sustainable use of natural resources in many cultures. When people begin to abandon these beliefs, traditional controls on resource exploitation often disappear, potentially leading to devastation of formerly protected areas such as sacred groves. Ethnobiologists are increasingly aware of the need to record these belief systems and to verify empirically what impact they have on conservation and use of natural resources. In areas in which traditional cosmology is still a viable element of local culture, conservation biologists are exploring ways of integrating it with the management of protected areas.

H. The Internationalization of Ethnobiology

Although ethnobiology as an academic discipline originated in Europe and the United States, it has now been embraced by researchers in many developing countries who have subsequently adapted the techniques and concepts to their own goals and local conditions. The emergence of professional societies of ethnobotanists in developing countries, ranging from the Indian Society of Ethnobotanists in 1980 to the Asociación Mexicana de Etnobiología in 1993 and the Sociedade Brasileira de Etnobiologia e Etnoecologia in 1997, is evidence of this trend. The internationalization of ethnobiological research and training has resulted in new directions in theory and application, enriching the field. In India, the tradition of conducting ethnobotanical inventories in various tribal areas has continued, but it is now supplemented by innovative approaches to studying the harvest of non-timber forest products in joint forest management schemes and practical strategies to create community biodiversity registers. Researchers in China have contributed studies on ecological succession in swidden fields, marketing of useful plants, and analysis of agroforestry practices. Equally impressive are developments in Mexico, where ethnobiologists have focused on the management of anthropogenic and natural ecosystems as well as the process of domestication of botanical resources.

III. THE NEW SYNTHESIS

The proliferation of labels and orientations for the study of local biological knowledge and practice is likely to continue. Consensus on precise definitions will be difficult to achieve because researchers are approaching the field from a variety of academic disciplines, bringing with them a wealth of new concepts and methods. Despite this dynamic development, there is some semblance of an agreement on a typology for the discipline. Distinct approaches to gathering empirical data on the reciprocal interactions between people and biological organisms will continue to be referred to by terms such as economic botany, ethnobotany, ethnoecology,
...and ethnobotany. Ethnobotany is becoming the preferred term for an integrative discipline that draws on all these approaches to analyze traditional biological knowledge and practices throughout the world. It is unified by a central theory that local peoples’ systematic knowledge and management of biological organisms and ecosystems can be classified as biological sciences, covalidated by qualitative and quantitative research methods. Perceived in this way, ethnobotany blends conventional studies carried out by economic botanists, ethnobotanists, ethnoecologists, and ethnoscientists that present a limited vision of local people’s interaction with the natural environment. This provides an opportunity for reintegration of various disciplines of natural science, counteracting the impact of the reductionism of global scientific knowledge.

Ethnobotany seeks not only to integrate these various lines of scientific research but also to focus them on supporting community development and biodiversity conservation, which are clearly multidisciplinary endeavors. American ethnobotanist Darrell Posey argued for this style of “advocacy” or “applied ethnobotany,” whose goal is to reform the economic, environmental, and social policies that are at the root of many problems which affect people in rural and urban settings. This approach has stimulated a new generation of researchers to blend scientific research with an awareness of the loss of biotic and genetic resources, indigenous struggles for land and resource rights, and negative aspects of globalization. The ultimate goal is sustainable development, as defined by cultural, ecological, and economic parameters.

The yen for integration goes beyond creating an interdisciplinary, applied field of study. Although ethnobiologists have tended to concentrate on the empirical side of local biological knowledge, there is renewed interest in symbolic and other interpretive approaches that could give a broader view of how people perceive their natural surroundings. Many ethnobiologists advocate adopting a participatory approach through which the entire study, from research design to application of the results, is conceived as a collaborative effort between local people and researchers. These developments represent a significant achievement for ethnobotany, a field that has always sought to blend perspectives from many cultural traditions.

See Also the Following Articles

AGRICULTURE, TRADITIONAL • HISTORICAL AWARENESS OF BIODIVERSITY • HUNTER-GATHERER SOCIETIES • ECOLOGICAL IMPACT OF • INDIGENOUS PEOPLES • BIODIVERSITY AND • SOCIAL AND CULTURAL FACTORS

Bibliography


I. Overview

This article discusses the origin of eukaryotes, the only life-forms on Earth that are not bacteria. Eukaryotes, the one- or many-celled beings whose cells contain nuclei, can be classified unambiguously into one of four groups (or "kingdoms"): plants, animals, fungi, and protoctists. Bacteria (kingdom Monera), by contrast, are made of prokaryotic cells, whether or not their bodies are composed of one or many cells. Protoctists (kingdom Protoctista), which may be an unfamiliar term, include all those nucleated organisms that are not plants, animals, or fungi. Among the live present-day protoctists are free-living nucleated organisms such as single-celled amoebae and Paramaecium and huge multicellular seaweeds such as the giant kelp. The smaller members of the kingdom, the protozoa, are made of cells that resemble the tissue cells typical of plants and animals. Protoctista includes composite organisms—we call them the "water neithers" because they are all aquatic but are neither plant nor animal (nor fungus). All green, brown, and red seaweeds are protoc-
tists, but so are water molds, slime molds, slime nets, diatoms, ciliates, and many other even more obscure members of this kingdom. The aquatic photosynthesizers—seaweeds, diatoms, golden yellow algae, and the like—differ from the plants, members of the kingdom Plantae, in that, unlike plants, no protist develops from embryos. There is evolutionary continuity between large familiar organisms made of nucleated (eukaryotic) cells and the nucleated free-living single-celled organisms: the protists (this informal term includes all single- and few-celled protists). For over 100 years, and even today, unfortunately, these microorganisms are known as "protozoans." Protozoa, a word derived from Greek roots meaning "first" and "animal," is misleading because "Protozoa" are in no way animals and indeed the early uni- and multi-cellular eukaryotes—the protists—evolved not only into animals but into modern protocists, plants, and fungi as well.

The greatest morphological difference in all of modern life is between prokaryotic organisms—bacteria—and eukaryotes—all the rest of life. From unicellular pond ciliates to blue whales and redwood trees, the eukaryotes show fundamental similarities to each other. At a cell level, the difference between a nucleated amoeba cell and a bacterial cell is far greater than the difference between that of an amoeba, a human epidermal cell, and a sperm cell from the male sex organ of a ginkgo tree. The difference between prokaryotes and eukaryotes is profound: it is comparable to the gap between the Kitty Hawk flying machine of the Wright brothers and the Concorde jet liner. Unlike the evolution of humans from apelike predecessors, which is bridged by the fossil bones of ape–human intermediates (australopiths and Homo species like Australopithecus afarensis, Homo erectus, Homo ergaster, and H. sapiens neanderthalensis), the evolution of eukaryotes from protocysts appears disjoint.

It is almost miraculous that organisms composed of the larger, more complex cells appear in the fossil record. The earliest eukaryotes are fossils known as acritarchs, dated by radioactive igneous intrusives and stratigraphic correlation to be 1,200 million years old. The only evidence for any life prior to the body and trace fossils of extinct, presumed eukaryotes is for bacteria—both as single microfossils and as microbial communities that left fossil remains. Bacterial life is ancient. Evidence for the earliest bacteria places them roughly contemporaneous with Earth's formation of a solid crust, almost four billion years ago. So the big question is: How, from bacteria or protocysts, did eukaryotes, nucleated cells, evolve? The answer seems to be: by symbiosis, or the living together of different kinds of bacteria. The evolutionary process of symbiogenesis—the sequence of events that occur when symbiotic partners merge to form new cells, tissues, organs, or organisms—was crucial in bridging this evolutionary gap. The union of different types of bacteria to form amoebalike cells that became ancestral to the rest of the prokaryotes, fungi, plants, and animals is described here. This event spawned a revolution in biodiversity.

II. DIFFERENCES BETWEEN PROKARYOTES AND EUKARYOTES

One model of the origin of nucleated cells is depicted in Fig. 1. Three distinct symbiotic mergers (blue triangles) are hypothesized.

The first microbes to be discovered were protocysts, today called eukaryotic microorganisms. The pond water microbes studied by Dutch draper turned microscopist Antoni van Leeuwenhoek included Mesodinium rubrum and green algae. Most protocystic cells, evolutionary predecessors to protists, are up to a thousand times smaller than nucleated cells such as those of an amoeba or an animal tissue cell. The term eukaryote literally means "true kernel," from Greek roots; protokaryote comes from words meaning "before kernel." The kernel in question is the nucleus, which contains chromosomes of varying numbers. These numbers are not necessarily correlated to complexity. A dandelion, for example, has many more chromosomes than the forty-six found in all human cells except sperm and egg, which have twenty-three chromosomes each, and blood cells. Mammalian blood cells even lose their nuclei in the process of development and of course eventually die, being unable to reproduce.

As an individual understands the world on the basis of his or her own experience, so have biologists traditionally understood life on the basis of large, familiar organisms. Close inspection of the putative prokaryotic ancestors of all eukaryotes from single-celled yeasts to giant sequoias reveals significant and great differences. Prekaryotes lack true chromosomes. They never have membrane-bounded nuclei. The DNA of bacteria floats free in strands that have little or no protein and often the cells bear DNA in tiny rings called plasmids. Bacterial ("chromonemal") DNA, unlike that of eukaryotes, is not packaged into the winding protein-coated chromosome structures. In eukaryotic cells, DNA complexes with protein and capable of staining a deep red color lines up and separates in an intricate pattern (called mitosis) as the cell divides.
FIGURE 1  Origin of nucleated cells.
Although standard bacteriology texts refer to species of bacteria and assign them binomials (e.g., *Escherichia coli*, *Bacillus anthracis*, *Chromatium pfennigii*), the life-habits of bacteria defy any definition of species, especially the zoological. If a species is defined as a group of organisms whose "members can breed with each other," only one species of bacterium lives on Earth today! In eukaryotes, at least plants and animals, species are maintained by sexual reproduction. Sex cells reminiscent of ancestral protist eukaryotes recognize each other, come together, and fuse to form new cells now with twice the number of chromosomes as the sex cells. These new cells, after fertilization in plants and animals, grow by mitotic cell division (reproduction) to form an embryo. Embryos continue to grow and differentiate into adult plants or tissues and organs of adult animals. Since a bacterium can infect from a few to virtually all of its genes into any other entirely different bacterium, the standard species concept applied to them has no meaning. A bacterium after receipt of genes from another bacterium is a recombinant. This new and different bacterium can then reproduce in its recombined state indefinitely. Sexually-preserved species borders are an imposed fiction in the case of bacteria, because every minute recombination is rampant. Sorin Sonea and Maurice Panisset note that if bacteria had been discovered on Mars instead of on Earth, their astounding natural history would have led scientists to classify them as radically different from all other life-forms.

Other common features of eukaryotes are also lacking in bacteria. Eukaryotes harbor small membrane-bounded bodies called organelles inside their cells. A most important set of eukaryotic organelles, in addition to the ever-present nucleus, are the mitochondria. These are oxygen-using inclusions, respiring entities that provide the cell with energy. All algal and plant cells have another set of organelles that contain chlorophyll: plastids. These are called chloroplasts if they are green. Plastids are named rhodoplasts or phaeoplasts if found inside certain red or brown protoctists (e.g., in coralline red algae or brown kelp). Plastids contain many different pigments, but these always include chlorophyll a, one of the chemical requisites of oxygenic photosynthesis. Chloroplasts mediate the conversion of solar energy into intracellular fuel in algae and plants. They are, of course, absent in animals, whose metabolic fuel is derived from energetic compounds in the environment—food. Another general difference between prokaryotes and eukaryotes is that, with a couple of significant exceptions, eukaryotic cells are aerobes: they depend on oxygen for cell respiration.

The fossil record of oxidized minerals in Earth's crust clearly tells us that free oxygen gas was not present in significant quantities until well after life's origins. Many modern forms of bacteria—such as the methane-generating archaea—coexist with cellulose-degrading bacteria in the special grass-digesting "stomachs" (the rumens) of bison, cows, and water buffaloes—are anaerobes. They are poisoned by oxygen. Today's eukaryotes somehow evolved from earlier life-forms that were intolerant of oxygen. Indeed, modern-day eukaryotes that lack mitochondria, such as pelomyxids, retortamonads, pyrsonymphids, trichomonads, calonymphids, and hypermastigotes, can be interpreted as "living fossils." These swimming protozoa are legacies of the missing links between the tiny gene-trading prokaryotes and the sexual species-forming eukaryotes.

III. SYMBIOSIS AS AN EXPLANATORY MECHANISM

Symbiosis was defined in 1873 by the German botanist Anton de Bary as the "living together of differently named organisms." To the biologist it refers specifically to organisms of different species that remain in more than casual contact. Bacteria of different kinds are certainly recognizable by their morphological and physiological traits rather than by their genetic and sexual behaviors. They need not be confused by inappropriate comparison to the familiar, sexually reproductive plants and animals who can easily be assigned to species proper. Biologists distinguish epibiosis, the symbiotic contact of skin or outer coating, from endosymbiosis, a kind of Jonah-in-the-whale association in which one organism enters, and sometimes fully merges with, the entered organism. Symbiogenesis is a kind of permanent mating. One might be tempted to accuse nature of bestiality were it not for the fact that the "unions" in question far transcend fleeting sexual "mistakes." The Russian naturalist Konstantin Merezhkovsky (1855–1921) was among those of the last century to suggest that symbiosis leads not only to fleeting exploitative (parasitism) or mutually food-providing (symbiotrophic) relationships, but also to the formation of entirely new organs and organisms. He coined the term "symbiogenesis" and described many of the most conspicuous examples. The origin of eukaryotic cells, which represents such a "jump" in both the fossil record and the composition of extant life, is best explained by Merezhkovsky's symbiogenesis. Today, genetic, comparative morphological,
and other lines of evidence suggest that this kind of “innovation by association” was the modus vivendi that transformed gene-trading bacteria into cells with nuclei, mitochondria, and, in the case of algae and plants, plastids. In life’s long history, natural selection acting on symbiogenesis created new species, not only of eukaryotic cells but also of their descendants: sexually reproducing lineages of multicellular bodies. We think that only eukaryotic organisms with their multiple integrated genetic systems form true species.

But how can mere living together lead to the formation of new species? A variety of modern-day organismic associations help to explain this. Epithons, such as the plovers that feed atop crocodiles by cleaning their teeth, are examples of symbiotic partners. They are compensated, if one considers their location and delicious potential edibility, for their services. Episymbiosis often evolves because organisms, which can be viewed as open thermodynamic systems, “leak”—that is, they incompletely make use of food and energy reserves. This leakage allows other organisms to move in. Human skin, even the surface of the human eye and its lashes, serve as a breeding ground for mites and microorganisms. Such normal human accoutrements as bacteria and protists lead to health problems if their populations grow too much. Yet their removal is also dangerous, in part because they keep each other in check. The fungus Candida, for example, which we associate with yeast infections, is normally kept in check by bacteria, which is why yogurt, which contains Acidophilus, is sometimes prescribed as a culinary prophylactic for the ailment.

Endosymbiosis seems disturbingly intrusive, but it leads to radical transformation. An example of endosymbiosis would be the human gut bacteria that metabolize vitamin B12, a vitamin that humans, unaided by symbiosis, would be the human gut bacteria that metabolize vitamin B12, a vitamin that humans, unaided by symbiosis, would be unable to synthesize. A striking example of symbiosis is the seafloor worm, Convoluta roscoffensis, which dwells on the beaches of northern France, southwestern England, and southwest Spain. To the casual observer, this organism is a seaweed: it looks like green scum. During rough seas or stormy weather—or even when a biologist stomps the surrounding ground—the "seaweed" uses its muscles to burrow into the sand. The "seaweed" is really composed of tiny worms whose mouths are closed. They live by direct internal absorption of nutrients provided by the gardens of algae within their cells. The endosymbiotic feeling, as often is the case, is mutual: the host organisms, the worms, receive free nutrients, whereas the resident, eaten but undigested, receive a protective enclosure. The living translucent "greenhouse" of a mobile being results: the green animal eludes environmental danger at a moment’s notice.

Endosymbiosis seems to be a species-creating phenomenon. The freshwater coelenterate Hydra, for example, is transparent, brown, or green, depending on whether its cells contain photosynthetic inclusions full of the green alga Chlorella. Symbiotic associations between other animal hosts abound. Green, food-producing algae or cyanobacteria live inside giant clams, snails, and even in the hair shafts of polar bears. These minute tubes provide a warm, snug locale for Arctic photosynthesizers.

Photosynthetic alliances grow and evolve from relatively casual epibiosis to permanent endosymbiotic mergers. The photosynthesizers that dwell on sun-illuminated surfaces of other living beings are ingested, but in their struggle to resist digestion they continue to metabolize. Their continued metabolism provides the would-be predator with additional food—the evolutionary equivalent of a free lunch.

A somewhat different mode for the establishment of endosymbiosis is that of infection. Microbes in the millions are normally and nonpathologically present on and in plant and animal (including human) bodies. Intimate associations among distinct organisms are not necessarily detrimental. But of course relationships degrade. When one rapid grower enters and exploits another, it naturally produces wastes and may upset a delicate internal balance. Infection can evolve into tolerable truce and permanent productive association if the rapidly growing invader moderates its behavior and growth in response to the "objections" of the would-be invaded.

A striking example of endosymbiosis-by-infection was documented by Korean-American biologist Profesor Kwang Jeon at the University of Tennessee, Knoxville. Jeon’s collection of amoebae were tainted by a sample sent to him by colleagues that contained amoebae infected with bacteria. Most of the new amoebae died. All of the survivors of this natural selection scourgce eventually transformed into what was described as a new species of amoeba. The transformed amoebae retained an internal collection of relatively slow-growing endosymbiotic bacteria. Some 40,000 bacteria resided in each amoeba! Jeon showed that the new amoebae, now an amoeba–bacterial complex, depended on what had once been a fatal infection. To test this hypothesis, he replaced the nuclei from bacteria-containing amoebae with "normal" nuclei from amoebae that were never infected by the newly acquired bacteria. He found that members of the "new" amoeba species invariably died unless they were "rescued" at about three days
postoperative by a reimplantation of the formerly pathogenic bacteria. The nuclei, now accustomed to the presence of hordes of cytoplasmic bacteria, required their continued presence for existence. A chronic deleterious association had become a condition for life; the bacterial "infection" was now an intracellular requirement, and by definition the pathogen had become an organelle.

IV. EUKARYOTE ORIGINS

RNA sequence comparisons of molecules that contain some 1400 nucleotide base pairs sharply divide bacteria into two groups. Professor Carl Woese (at the University of Illinois, Urbana) christened these two groups of organisms—both prokaryotes—"archaebacteria" and "eubacteria." A comparison between Woese's phylogeny that ignores symbiosis and is based on molecular criteria and ours based on symbiogenesis is shown in Fig. 2. Both are of course highly simplified.

As their name suggests, the archaebacteria are putatively a more ancient lineage. Their metabolism appears to have remained relatively unchanged from the early days of an Earth that lacked atmospheric oxygen. The archaebacteria include three great subgroups: methane-producing, thermacidophilic, and salt-tolerant prokaryotes. The first two derive their energy in environments producing, thermacidophilic, and salt-tolerant prokaryobacteria include three great subgroups: methane-producing, thermacidophilic, and salt-tolerant prokaryotes. Both are of course highly simplified.

As their name suggests, the archaebacteria are putatively a more ancient lineage. Their metabolism appears to have remained relatively unchanged from the early days of an Earth that lacked atmospheric oxygen. The archaebacteria include three great subgroups: methane-producing, thermacidophilic, and salt-tolerant prokaryotes. The first two derive their energy in environments producing, thermacidophilic, and salt-tolerant prokaryobacteria include three great subgroups: methane-producing, thermacidophilic, and salt-tolerant prokaryotes. Both are of course highly simplified.

As their name suggests, the archaebacteria are putatively a more ancient lineage. Their metabolism appears to have remained relatively unchanged from the early days of an Earth that lacked atmospheric oxygen. The archaebacteria include three great subgroups: methane-producing, thermacidophilic, and salt-tolerant prokaryotes. The first two derive their energy in environments producing, thermacidophilic, and salt-tolerant prokaryobacteria include three great subgroups: methane-producing, thermacidophilic, and salt-tolerant prokaryotes. Both are of course highly simplified.
V. SERIAL ENOSYMBIOSIS THEORY

More than one explanation has been proffered to cover the evolutionary gap from prokaryotes to eukaryotes. Clearly, however, symbiosis—and not just a one-time symbiosis—appears to be involved. All organisms are open thermodynamic systems exchanging matter, energy, and information with their environment. The notion of a totally isolated, independent organism may work as a mathematical assumption or a Platonic idealization, but in the real world of responsive biological activity both genetic exchange and metabolic evolution present untold opportunities for organismic interaction. One organism sheltering another, predator–prey relations, and genetic exchanges lead to alliances of sexuality and symbiosis. Nonetheless, in obeisance to Darwin, traditional evolutionary theory permits only a series of minute and gradual changes that are touted to generate all propagating mutant descendants. Symbiogenesis, the sudden appearance of new organisms by permanent association of different kinds of partners, is not usually what jumps to mind when one considers the evolution of new species. It should.

Serial endosymbiosis theory posits a succession of up to four evolutionary associations that we think led to modern-day cells. The eukaryotic cells of plants are the result of at least two and perhaps three separate evolutionary partnerships among at least three or perhaps four kinds of bacteria. One was the photosynthetic bacteria that conferred photosynthetic abilities upon the ancestors to plants—latter-day cyanobacteria that today exist as the plastid organelles of plant cells. The nucleated cells of animals are the result of at least one, perhaps two separate evolutionary partnerships among at least two or perhaps three kinds of bacteria.

To understand the evolutionary origins of nucleated cells, it helps to return to Earth in its geological infancy. In the Lower Archean eon, about four billion years ago, no protoctists, plants, animals, or fungi existed, only...
the metabolically diverse bacteria. We can provisionally trace the evolution of metabolic diversity by comparison with pathways in modern-day organisms. The first bacteria may have been fermenters because such organisms would need only complex organic compounds preexisting in the environment, rather than other organisms, on which to feed in order to grow. Among the earliest life-forms were photosynthetic cells, green and purple bacteria that much later yielded most of the real estate of the sunlit Earth’s surface to algae and plants. Comparative analyses suggest that these ancient purple bacteria, like their cousins today, used hydrogen sulfide \((H_2S)\) rather than water \((H_2O)\) to make their organic (carbon–hydrogen) bodies. The hydrogen sulfide, spewed from volcanoes, would have been more plentiful on the early, more tectonically active Earth. But as Earth cooled and tectonic activity subsided, sulfide became less available than that ubiquitous solvent water. Mutations appear to have allowed some bacteria to alter the chemical reactions of their metabolism such that they could now, using the energy of sunlight, break the hydrogen–oxygen bonds of water to make their bodies. Life of course existed in water since its inception. But the use of water as a metabolic resource led necessarily to an entropic waste—oxygen. Powered by sunlight, early life discarded oxygen, which reacts strongly with the carbon–hydrogen compounds of living bodies, into the atmosphere. The rock record of rust, oxidized iron, and uranium oxides demonstrates that enormous quantities of oxygen did not begin to build up in Earth’s atmosphere until about two billion years ago. 

Life’s difference from the universe around it, and its tendency to evolve toward greater levels of complexity, reflects its status as an open, entropy-generating system. The use of energy and the inevitable production of waste as a result lead to changes in life as it metabolically stays the same and reproduces. Reproduction of course is an extension of metabolism. But life also changes because of the toxic effects that its growth has on itself. The cyanobacteria that first mutated to use the hydrogen in water for their electron donor were also the first to be poisoned by oxygen toxicity. Among the prolific anaerobic organisms living at the surface were also rapidly swimming bacteria including spirochetes, many kinds of which dwell today in symbiotic alliance with other organisms in the hindguts (swollen inenterines) of wood-eating termites. Just as the Industrial Revolution changed human civilization in part due to new systems of railway, steamboat, and eventually airplane transportation, fast-moving bacteria are posited in Serial Endosymbiotic Theory (SET) to have conferred major changes upon early cells.

The most provisional part of the evolutionary story of the symbiotic eubacterial origin of eukaryotic cells is the part that explains the origin of intracellular movement. The association of fast-moving eubacteria, perhaps spirochetes (speedy corkscrew-shaped burrowing feeders upon and even permanently attached to other organisms), with larger archaeabacteria is postulated to be the most far-reaching but integral part of the story. Intracellular mobility—the “cytoskeletal system”—of all eukaryotic cells is remarkable for its extent, complexity, and ubiquity. Cells with nuclei not only undergo the elaborate alignment and separation of reproduced chromosomes during mitotic cell division, they also have characteristic motility structures such as the mitotic spindle with its microtubules. Microtubules, 24 nm in diameter and varying in length, are not found in prokaryotes. Among the many microtubule-based structures such as sperm tails, cilia, and moving spines are the mitotic spindles themselves. The mitotic spindle, crucial for eukaryotic cell reproduction, seems to appear out of nothing to become superbly organized and to separate chromosomes during cell division. Eukaryotic cells are famous for the incessant streaming of their cytoplasm. The intracellular activities are reminiscent of a crowded train station filmed in time-lapse photography. Many kinds of nucleated cells (of plant sperm, animal tissue, and myriad protists) bear special motility structures that display a ninefold symmetry in cross section. These, whether cilia, sperm tail, or sensory hair, are called “undulipodia.” The whipping tail used in cell propulsion, sensory detection, or propulsion of eggs along the fallopian tubes of mammals are all examples of undulipodia. In cross section the shafts of these motile cell structures have a characteristic “\(9(2)+2\)“ pattern. As seen with the electron microscope, nine pairs of tubes surround one central pair. This cross section structure, 230 nm in diameter, is found to be nearly universal in eukaryotic organisms, from the whipping tails of the sperm of ferns, mosses, cycads, and trees (yes, some trees do produce sperm) to the cilia that propel gastropod and insect cilia. In all prokaryotes whether archaeabacteria or eubacteria, undulipodia are the most far-reaching but integral part of the evolutionary story of the symbiotic eubacterial origin of eukaryotic cells. As has often been witnessed in modern-day symbiotic partnerships, the rapid swimmer would have been attracted to the food leaking out of the larger
bacterium and the partnership would eventually have been selected in the evolution of permanent motile associations. Ciliates today often differ only in the placement of their undulipodia. Moreover, the microsurgical grafting of patterned patches of undulipodia from one ciliate to another leads to altered patterns in which the ciliate reproduces without benefit of a sexual partner. Although DNA is absent in the motility organelle of waving undulipodia itself, many other facts suggest a possible symbiotic origin. Grafted patterns of altered undulipodia “breed true.” The sheer resemblance of whipping undulipodia—not easily differentiable in some organisms, such as Mixotricha paradoxus, from actual attached spirochetes—to fast-moving, free-living bacteria provides a clue.

Early evolution may have featured a great proliferation of mastigote (undulipodiated) cells resulting from a successful union of anaerobic archaebacteria and spirochetes or spirochete-like organisms in which the DNA of the motile form entered the newly evolved nucleus as it formed. The quick-swimming partnership complex would have conferred a greater ability to acquire food on the merger. The greater access to new food by the larger, more quickly moving moving archaebacterial–eubacterial complex portended a great future. As often occurs in evolution, association led to new opportunity and power. Eventually, the greatest, most integrated partnerships formed unified genetic entities. These genetic entities went on to add new members in the evolution of new cells, protocist lineages ancestral to plants, fungi, and animals.

Much more and better-quality evidence exists for later symbioses. Whether or not swimming eubacteria merged with archaebacteria to form mastigotes such as ciliates, these undulipodiated cells were very likely invaded or infected, some two billion years ago, by oxygen-using bacteria. These oxygen-using or respiring bacteria, similar to free-living forms today—some of which are not only oxygen-respiring but also invaders of other cells (e.g., Bdellovibrio)—proliferated in the aftermath of the evolution of oxygenic photosynthesis. The gaseous waste from using water as a source for hydrogen was the production, eventually at a planetary level, of the highly reactive oxygen gas (O2). This gas naturally, and sometimes fatally, burns cell tissue. Many organisms were forced to retreat to the muds, out of harm’s way. Some hung on until animals such as the ancestors to insects evolved. Hiding like soldiers in a tank or submarine, the oxygen-poisoned anaerobes today prolifere only in biological enclosures, such as sulfurous muds, the termite hindgut, the cow rumen, or other natural protective settings. The respirers, however, turned the metabolic negative of volatile oxygen into the metabolic positive of a new, more efficient source of energy. Bulging about, funneling potentially dangerous redox reaction into their own growth, the ancestors to mitochondria kept the harmful gas away from others by using it themselves. This quality led to the greatest burst of biodiversity of all: the fabulous protocist proliferation of respirers, both alone and in associations with other organisms. The symbiotically evolved amoeba of Jeon, mentioned earlier, were infected with respirers. Merged with mitochondriate, anaerobic mastigotes, the respiring forerunners to modern mitochondria powered the metabolism of the cells of all the major macroscopic lineages: protocists, fungi, plants, and animals.

The final major symbioses—for which, as in the case of mitochondria, there is good genetic evidence—occurred in the ancestors to algae and plants. Taxonomically, the major difference between algae and plants is that plants form from a fertilized nucleus that grows into an embryo that is retained in maternal tissue. By definition the embryo is a multicellular eukaryote that develops from a fertilization event, the fusion of two nuclei from complementarily gendered parents. Algal growth is more beltier-skeleton, amorphous, and never embryonic. The archaebacteria, perhaps with symbiotic spirochete offspring providing them varied means of mobility, merged with respiring bacteria. Today mitochondria, although they cannot be grown on their own, appear sometimes to revert back to their earlier, freely reproducing lifestyle; in cancer cells, for example, mitochondria reproduce rampantly. The nucleus itself does not appear to be the direct result of cell symbioses. The nucleus may, however, represent the intracellular equivalent of a “government” that evolved to deal with the first merger of separate cell lineages shown as the lowest triangle in Fig. 1. After the original archaebacterium–mitochondria module was established, the new nucleated cells had to feed. Many fed on photosynthetic bacteria. Bacteria do not have immune systems, and “eaten” food, like Jonah in the whale, continued to live and sometimes even reproduced. Although perhaps highly perilous, for starving mastigotes with symbiotically derived mitochondria it might also have been a godsend. With all the competition to find food, those who produced their own, or merged with photosynthetic organisms capable of production of their own food, would have been at a distinct advantage under the usual conditions of starvation. The algae and its derivative plant lineage are traceable to eukaryotes that symbiotically acquired bacterial photosynthesizers.
Today all familiar visible organisms are eukaryotes with nucleated, mitochondriate cells. Whether at the zoo, the forest, or the green prairie, we witness not isolated individuals but an evolutionary fact of life: hundreds of millions of years of symbiotic interliving. The simplest amoeba is no independent cell, rather it represents the outcome of an evolutionary flirtation with death and microbial diversity. A large animal, such as a hedgehog or woman, is a highly organized, differentiated clone of amoeba-like protist cells. An overall classification of life based on this view of the symbiogenetic origin of eukaryotes is shown in Fig. 3. Biodiversity is obvious, widespread, and crucial to our global environment. The existing major phyla, classes, genera, and species resulted not only from mutation accumulation but also by symbiotic alliance and permanent merger. Microbial proliferation and diversification behind the production of new species and large evolutionary transi-
EUKARYOTES, ORIGIN OF

Eukaryotes, the cells with a clearly defined nucleus and organelles, are the origin of complex life on Earth. The processes in the past are still with us as the understory of life on Earth.

See Also the Following Articles

ARCHAEA, ORIGIN OF • BACTERIAL GENETICS • BIODIVERSITY, EVOLUTION AND • DIVERSITY, MOLECULAR LEVEL • MICROBIAL BIODIVERSITY • NUCLEIC ACID BIODIVERSITY • ORIGIN OF LIFE, THEORIES OF

Bibliography


EUROPE, ECOSYSTEMS OF

Ladislav Mucina
University of Stellenbosch

I. Introduction
II. Ecosystem Diversity: Concepts and Approaches
III. Biodiversity of Habitats and Vegetation
IV. Challenges for the Future

GLOSSARY

biome A large-scale complex of ecosystems sharing similar climate and vegetation structure.
ecological biodiversity Variety of biotic communities (plant, animal, and microbial communities) and their complexes (ecosystems, landscapes, and biomes).
formation Large-scale (subcontinental or continental) vegetation complex defined primarily on the basis of a combination of dominating life-forms (hence, vegetation structure).
habitat A place of dwelling of a biotic community (or a complex thereof) showing particular combination of ecological factors occupying certain area in a certain period of time.
oroborne Mountain range characterized by particular climatic pattern and characteristic sequence of vegetation zones.
syntaxon (plural syntaxa) A category of vegetation typology based on the floristic-sociological approach (known also as the Braun-Blanquet approach); the basic syntaxon rank is “association,” which further groups into “alliances,” which group into “orders,” and orders group into “classes.”
vegetation megazone Large-scale vegetation complex characteristic for a zonobiome.
vegetation survey Product of research activity aimed at the description and classification of vegetation cover on various levels of complexity in a certain geographic area using various field and data-evaluation methods.
zonobiome Broad ecological topographical unit characterized by a certain climatic pattern.

THE PROBLEM OF biodiversity of European ecosystems is approached from the point of view of biomes and zonobiomes, flora-based vegetation typology (phytosociological tradition to classification of vegetation), and habitat classifications. The concepts of biome (based on a combination of dominating life-form and climate) and zonobiomes (by H. Walter) are compared for Europe. The most effective classification of ecosystems in Europe is based on the floristic-sociological approach, which has deep roots in European tradition and is widely used throughout Europe. This approach has yielded a hierarchical system of vegetation units which allow insight into the diversity of vegetation types of this continent. Basic features of diversity of vegetation types (syntaxa) are briefly outlined, and the importance of vegetation classification for delimitation...
and calibration of habitat classification systems in Europe is discussed. The tasks for further research and socioeconomic implications of the diversity of ecosystems in Europe are outlined.

I. INTRODUCTION

Hardly any place in the world supports in a relatively small area such a variety of countries, peoples, cultures, histories, languages, political views, cuisine, and types of cheese and wine as does Europe. There might be more languages and people in India, higher species diversity of plants in the tropical rain forest of Columbia, or more diverse cuisine in China, but there is only one Europe.

Europe's nature is showing many faces, both pristine ones in the form of tracks of tundra and forests and created (or mis-created) ones—those carrying the signs of the hand of man. Europe is a patchwork of ecosystems connected by an intricate net of gradients of ecological factors ranging from those controlling continental patterns of vegetation megazones and biomes to local gradients creating small-scale mosaics of biotic communities.

This article discusses the variability of European ecosystems and a major framework of biodiversity patterns at various scales of complexity.

Europe is a continent, a large chunk of land; therefore, I discuss this subject from the viewpoint of diversity of large-scale ecological units such as vegetation megazones and biomes. The terrestrial habitats (and their complexes) will dominate this discourse and I emphasize the diversity of vegetation assemblages, which are traditionally considered the core of structure and functioning of all terrestrial ecosystems.

Europe is approximately 10 million km$^2$ and spans 35–81°N of latitude and 60°E–10°W of longitude. Although Europe is the second smallest continent (7% of the world's land surface), the basic classification of ecosystem diversity can be discussed only at a very large scale of biome. Not only the limited extent of this article but also the extraordinary wealth of scientific knowledge collected over centuries and the diversity of the subject impose constraints on the level of detail and accuracy of this discourse.

II. ECOSYSTEM DIVERSITY: CONCEPTS AND APPROACHES

Ecosystem, often defined as the union of biotic community and its environment, presents an obvious difficult scaling problem. Ecosystems as real entities occur in space and time. Depending on conditions, one can recognize natural or artificial borders allowing classification of ecosystems, an exercise meant to set a framework for simplifying the complexity of ecosystems and featuring them in a synoptic way.

A. Classification of Vegetation

Most European ecosystems are obviously terrestrial ecosystems accompanied by semiterrestrial and aquatic ones. Vegetation (defined as the unity of plants occupying certain area in a certain time, Palmer and White, 1994) is the major biotic element of terrestrial ecosystems. Hence, the problem of classifying terrestrial ecosystems is a problem of classification of vegetation.

The approaches to the classification of vegetation are manifold and largely depend on criteria (and their weighting), aims, and means (Mucina, 1997a). Several basic approaches to classification of vegetation have emerged during the past century (Whittaker, 1978). Among these, the floristic–sociologic approach (Westhoff and van der Maarel, 1978; Dierschke, 1994) became a standard communication tool among European (and other) vegetation scientists involved in or harvesting fruits of vegetation classification exercises.

1. Floristic–Sociologic Approach to Vegetation Classification

The basis of the floristic–sociologic approach to classification of vegetation is the notion of the total floristic composition and the notion that some species indicate the environmental conditions better than others. These are called diagnostic species (Braun-Blanquet, 1964). Vegetation stands are selected by following various criteria, among which the so-called representativity of the studied vegetation type is the leading one. This approach is aimed at recognition of plant communities—plant assemblies recurrent in space and in time and showing distinct floristic composition which reflects certain combinations of current and past habitat conditions. The basic vegetation unit of this approach is the so-called association, a theoretical construction that is a result of the abstraction classification process involving many vegetation stands. A hierarchy of vegetation units based on groups of species having similarity between plant communities and their groups is also a vital part of the floristic–sociologic approach. Floristically and ecologically similar associations are grouped into alliances, the alliances are grouped into orders, and the orders are grouped into classes. The vegetation classes can be less formally grouped into divisions. Particular
associations, alliances, orders, classes, and divisions are termed syntaxa (Westhoff and van der Maarel, 1978) of various ranks and build the syntaxonomic hierarchy. An elaborate, highly formalized system of rules has been introduced to govern the formation of names of the syntaxa.

2. Dominance Approach to Classification of Vegetation

Another approach to classification of vegetation emphasizes the role of the dominant species or dominant growth forms, marking a departure from using plant taxonomy as the currency. Without discussing this approach in great detail (Whittaker, 1978), the notion of formation is the core tool of the dominance approach. The formation is the vegetation component (thus, in terrestrial ecosystems it is the leading element) of the concept of biome. Although the floristic–sociological approach is well suited for classification and description of vegetation on small geographic scales, formation and biome are, for the same purpose, well suited on large scales (subcontinental and continental).

3. Classification of Habitats

Despite much critical challenging of some subjective points regarding criteria for selection of stands and criteria of weighting of species in the classification process and sometimes awkward nomenclature, the floristic–sociological approach is aimed at enhancing effectiveness of communication when addressing units of vegetation cover. It is therefore not surprising that it was the terminology and concepts of this approach that were adopted as the basis of the majority of the units of the habitat systems developed for various purposes by the former European Communities (now European Union) authorities.

The classification of habitats passed several stages of development, spanning CORINE habitat classification (Commission of European Communities, 1991), Palearctic habitat classification (Devillers and Devillers-Terschuren, 1996), and EUNIS habitat classification (Davies and Moss, 1998). Obviously, the purpose of habitat classification is a practical one. It should serve (and serves) important purposes by delimitation and evaluation of land-use characteristics, projection and management of nature reserves of other areas of special nature conservancy interests (e.g., the Natura 2000 network of Special Areas of Conservation), etc. Not surprisingly, such a habitat classification scheme is a legal standard within the borders of the European Union.

III. BIODIVERSITY OF HABITATS AND VEGETATION

A. Sources

The current vegetation landscape of Europe is a result of manifold forces forming and reforming the habitat networks during the past few thousands of years; undoubtedly, some features of this habitat and vegetation patchwork are very old.

One can hardly avoid using oversimplification when talking about the sources of community (including vegetation and animal communities) and ecosystem diversity of present-day Europe. I distinguish four major prerequisites to the community and ecosystem diversity (Table I):

1. Taxon diversity: Plant and animal communities are composed of individuals belonging to several (often many) taxa. Speciation and associated processes of within-taxon diversification in addition to migrations of taxa—either ancient, natural ones or current ones largely driven by man’s activities—are the major sources of taxon diversity.

2. Habitat diversity: Habitats are “homes” of biotic communities and the greater the variety of habitats, the greater the variety of biotic communities populating them. The character of habitats is determined primarily by ecological factors—their character and dynamics. Regardless of whether one considers spatial or temporal scales, one can recognize geological, geomorphologic,
and soil-formation processes accompanied by hydrological dynamics and climate changes as the major natural sources of the habitat diversity.

3. Diversity of biotic interactions: The individuals representing different taxa interact within the biotic communities in many ways. They may compete for resources, facilitate each other's growth and reproduction by a plethora of positive interactions, and may behave indifferently. Undoubtedly, the biotic interactions form the face of each community and thus contribute to ecological diversity.

4. Diversity of human interference: Man has made himself the center of the universe and from this point of view has also become the dominant source of the disturbance (up to complete removal) but also creation of new habitats and new biotic communities through manipulating ecological factors and facilitation of species migrations.

B. Patterns

1. Biomes of Europe

Europe has six (of the world's nine) zonobiomes as defined by Heinrich Walter (Walter and Breckle, 1991, pp. 22–25). These include, from south to north, Mediterranean zonobiome (IV), warm-temperate zonobiome (V), nemoral zonobiome (VI), continental zonobiome (VII), boreal zonobiome (VIII), and polar zonobiome (IX). Furthermore, many zonoecotones mediate between the particular zonobiomes and those are designated as IV–V, V–VI, VI–VII, VII–VII, VII–VIII, VIII–VII, VIII–VI, and IX–VIII. The zonoecotone III–IV, mediating between subtropical zonobiome (not represented on the continent) and Mediterranean zonobiome, is found on the Canary Islands.

The Walter's biome scheme, largely based on a combination of climatic characters, is currently under revision in light of more modern definitions of the biome concept involving the combination of climate and functional types (including traditional life-forms) using modeling to predict vegetation patterns (Prentice et al., 1992; Haxeltine and Prentice, 1996) or to reconstruct the patterns of the past (Prentice et al., 1996).

Following the combined approach to define biomes using climate and vegetation structure, one can tentatively recognize 10 biomes in Europe and the Canary Islands (Table II).

Surprisingly little attention has been devoted to the definition of extant European biomes in comparison with other parts of the world. This can to a great extent be attributed to traditions in classification of ecosystems in Europe. The dominance-based approach to classification of vegetation leading directly to definition of biomes when exercised on a large scale is the predominant approach in parts of the world in which there is a lack of advanced knowledge on flora. Europe is undoubtedly a region in which the flora-based approach has a long and firmly rooted tradition.

2. Diversity of Vegetation Types

To list all syntaxa described in Europe is impossible here. Recent account of the high-rank syntaxa (Mucina, 1997b; Rodwell et al., 1998) revealed that the European vegetation can be classified into 75 classes (Table III).
### TABLE III
Survey of European Phytosociological Classes and Their Correspondence with the Biome and Orobiome Classification Scheme

<table>
<thead>
<tr>
<th>Class</th>
<th>Or</th>
<th>AI</th>
<th>Characteristics</th>
<th>Biome/zonality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aquatic vegetation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lemnietea</td>
<td>1</td>
<td>2</td>
<td>Communities of floating cormophytes (pleustophytes)</td>
<td>All but T</td>
</tr>
<tr>
<td>Characeae fragilis</td>
<td>3</td>
<td>5</td>
<td>Communities of submerged, ramified macroalgae</td>
<td>All but T</td>
</tr>
<tr>
<td>Potametea</td>
<td>3</td>
<td>7</td>
<td>Communities of freshwater, rooted, submerged, and floating macrophytes</td>
<td>All</td>
</tr>
<tr>
<td>Ruppietea maritimae</td>
<td>1</td>
<td>2</td>
<td>Communities of rooted, submerged macrophytes of brackish waters</td>
<td>B, N, M, CSB</td>
</tr>
<tr>
<td>Zosteretea</td>
<td>3</td>
<td>3</td>
<td>Communities of rooted, submerged macrophytes of shallow seas</td>
<td>Shallow sea</td>
</tr>
<tr>
<td>Vegetation of freshwater marshes and fens</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isoeta-Litorelletea</td>
<td>2</td>
<td>8</td>
<td>Dwarf amphibious vegetation of oligotrophic water bodies, and dystrophic lakes and bog hollows</td>
<td>T, B, N, M, OA</td>
</tr>
<tr>
<td>Isoeto-Nanojuncetea</td>
<td>2</td>
<td>10</td>
<td>Dwarf amphibious vegetation of banks and bottoms of mesotrophic and eutrophic temporary water bodies</td>
<td>T, B, N, M, OA</td>
</tr>
<tr>
<td>Montio-Cardaminetea</td>
<td>1</td>
<td>8</td>
<td>Moss- and herb-rich vegetation of water springs and edges of fast-running mountain rapids</td>
<td>T, B, N, M, OM</td>
</tr>
<tr>
<td>Phragmiti-Magnocaricetea</td>
<td>3</td>
<td>12</td>
<td>Vegetation of reeds and sedge-dominated fresh-water and brackish swamps</td>
<td>T, B, N, OA</td>
</tr>
<tr>
<td>Schoenitio-Caricetea</td>
<td>2</td>
<td>6</td>
<td>Vegetation of transitional mires, fens and bog hollows</td>
<td>T, B, N, OA, OM</td>
</tr>
<tr>
<td>Oxyccocio-Sphagnetea</td>
<td>3</td>
<td>9</td>
<td>Vegetation of ombrogenic raised bogs and wooded boreal bogs</td>
<td>T, B, N, OA</td>
</tr>
<tr>
<td>Coastal vegetation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calidromea maritimae</td>
<td>3</td>
<td>7</td>
<td>Strandline dwelling, short-lived nitrophilous vegetation</td>
<td>B, N, M, OA</td>
</tr>
<tr>
<td>Aegionietea maritimae</td>
<td>2</td>
<td>5</td>
<td>Dwarf pioneer vegetation populating closed and sandy soils in habitats under salt-spray influence</td>
<td>B, N, M, Intrazonal</td>
</tr>
<tr>
<td>Agropyro-ponganietea</td>
<td>1</td>
<td>1</td>
<td>Coastal, salt-sprayed grasslands on soft rocks and related vegetation of inland saline heavy loamy soils</td>
<td>M</td>
</tr>
<tr>
<td>Celmiso-Salicoreteae</td>
<td>3</td>
<td>12</td>
<td>Chasmophytic coastal vegetation under salt-spray influence</td>
<td>B, N, M</td>
</tr>
<tr>
<td>Thero-Salicoreteae</td>
<td>2</td>
<td>7</td>
<td>Vegetation of maritime salt marshes dominated by annual succulents</td>
<td>B, N, M</td>
</tr>
<tr>
<td>Spartito-Salicoreteae</td>
<td>1</td>
<td>1</td>
<td>Pioneer salt marsh grasslandways dominated by Spartina</td>
<td>B, N, M</td>
</tr>
<tr>
<td>Salicornio-Frisonietea</td>
<td>2</td>
<td>11</td>
<td>Mediterranean and thermo-atlantic perennial salt marsh scrub</td>
<td>M</td>
</tr>
<tr>
<td>Juncetea maritimi</td>
<td>3</td>
<td>13</td>
<td>Perennial maritime meadows and related herb-rich salt marshes</td>
<td>B, N, M</td>
</tr>
<tr>
<td>Hedeo-Elymo-Elymasio</td>
<td>2</td>
<td>4</td>
<td>Vegetation of embryonic shifting dunes of Atlantic, North Sea, and Baltic coasts</td>
<td>B, N</td>
</tr>
<tr>
<td>Ammophila maritimi</td>
<td>2</td>
<td>8</td>
<td>Vegetation of white and gray coastal sand dunes of Atlantic–Mediterranean distribution</td>
<td>B, N, M</td>
</tr>
<tr>
<td>Chasmophytic vegetation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amphiromo trichomatus</td>
<td>18</td>
<td>73</td>
<td>Chasmophytic vegetation of rock faces, lichen, and lichens</td>
<td>B, N, M, OA, OM</td>
</tr>
</tbody>
</table>

continues
<table>
<thead>
<tr>
<th>Class Or A1</th>
<th>Characteristics</th>
<th>Biome/zoality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adiantetea</td>
<td>1 2</td>
<td>Chasmophytic, fern- and moss-rich communities of water-rilled habitats of the Mediterranean</td>
</tr>
<tr>
<td>Thallicion remutellifolii</td>
<td>13 49</td>
<td>Vegetation of scree, gravel streptanthum, and related habitats</td>
</tr>
</tbody>
</table>

**Arctic and alpine vegetation**

| Loiseleurio-Vaccinietea | 1 4 | Dwarf shrub alpine and subalpine heathland of boreal and arctic Europe | T: Zonal |
| Salicion herbaeae | 1 9 | Snow-bed communities on siliceous substrates | OA: intrazonal |
| Juncetum trifidi | 7 23 | Siliceous alpine and boreal grasslands | T: OA Intrazonal |
| Carici-Rhododendron | 3 3 | Circumpolar cold steppe and field vegetation and alpine wind-erect Elymus grasslands | T: Zonal |
| Elyno-Seslerietea | 5 19 | Alpine and subalpine calcareous grasslands | OA: intrazonal |
| Melido-Acinetum | 3 15 | Subalpine–alpine to alpine tall-grass and tall-herb grasslands and related krummholz | B, N, OA, Animal |

**Synanthropic vegetation**

| Oryzetea sativae | 1 1 | Weed communities of rice fields | N, M |
| Bidentetea tripartiti | 1 2 | Annual ruderal communities of periodically flooded, nutrient-rich riparian and related synanthropic habitats | Azonal |
| Polygono-Poetea annuae | 1 4 | Short-lived, trampled, ruderal plant communities | B, N, M, OA, Animal |
| Stellario-Bronneria | 6 49 | Annual, herb-rich ruderal and agrostal communities | B, N, M, OA, Animal |
| Artemisietea vulgaris | 6 18 | Perennial thistle-rich (sub)xerophilous ruderal communities of temperate and Mediterranean regions | B, N, M, (CSD) Animal |
| Galio-Urticetea | 5 18 | Nitrophilous synanthropic herb-rich communities of woodland and riparian fringes | B, N, M, Animal |
| Epilobietea angustifolii | 1 4 | Tall-herb and tall-grass communities of woodland clearings and related shrubbery | B, N, (M) Animal |

**Temperate heathlands and grasslands**

| Calluno-Ulicetea | 3 16 | Temperate and boreal heathlands and related grasslands on nutrient-poor soils | N: Intrazonal |
| Koelerio-Corynephoretea | 6 17 | Graslands and related short-lived, herb-rich plant communities of sandly, nutrient-poor soils of temperate and boreal Europe | R: Zonal |
| Melinito-Artemishoffneria | 9 39 | Nutrient-rich meadow (pastures, hay meadows, and back) and wet grasslands | B, N, (M), S, OA Animal |
| Trifolio-Geranietea | 2 5 | Herb-rich fringed communities of temperate woodlands | B, N, M, Animal |
| Festucetum bromi | 12 64 | Steppes, rocky steppes, and sandy grasslands of the temperate and subboreal regions | S: Zonal |
| Puccinellio-Salicentra | 8 37 | Continental (inland) salt marshes, salt pans, and salt marshes | N: intrazonal |

**Temperate and boreal woodlands and shrubs**

<p>| Betulino-Pontica | 3 17 | Shrub mantle communities in regions of temperate deciduous woods | B, N, M, Animal |
| Salicion purpureae | 1 7 | Willow and poplar riparian woods and shrubs of temperate Europe | B, N, M, Animal |</p>
<table>
<thead>
<tr>
<th>Class Or A1</th>
<th>Characteristics</th>
<th>Biome/taxonality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alnetea glutinosae</td>
<td>3 5 Alder and willow carrs</td>
<td>B, N, (M) Annual</td>
</tr>
<tr>
<td>Querceta-Fageta</td>
<td>5 22 Deciduous mixed and suberophilius woods of temperate and subboreal Europe</td>
<td>(B), (N), (M) Zonal</td>
</tr>
<tr>
<td>Erica-Festucetea</td>
<td>1 7 Bell heath woods of cool/mesothermal and montane temperate distribution over limestone</td>
<td>(N), OA, (OM) Intrazonal</td>
</tr>
<tr>
<td>Juniperus-Pinetea</td>
<td>2 7 Mediterranean dry juniper—pine woods and inner-alpine steppic pine woods and related scrub</td>
<td>OM Intrazonal</td>
</tr>
<tr>
<td>Vaccinio-Piceetea</td>
<td>2 9 Boreal needle woods</td>
<td>B, (N) Zonal</td>
</tr>
</tbody>
</table>

**Oro-mediterranean grassland and scrub**

<table>
<thead>
<tr>
<th>Class</th>
<th>Or A1</th>
<th>Characteristics</th>
<th>Biome/taxonality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Festuceta indigestae</td>
<td>1 3 Oro-thermic siliceous Festuceta indigestae grasslands</td>
<td>OM Zonal</td>
<td></td>
</tr>
<tr>
<td>Asperulo-Poletea</td>
<td>1 2 Cynio-Sarcochion siliceous and herblands</td>
<td>OM Zonal</td>
<td></td>
</tr>
<tr>
<td>Festuceto-Compositae</td>
<td>2 9 Dry basiphilous pastures at high altitudes of the Submediterranean and western Mediterranean</td>
<td>OM, OM Zonal</td>
<td></td>
</tr>
<tr>
<td>Castaneo-Quercetea</td>
<td>1 3 Cynio-Sarcochion calciphilous grasslands and phrygana</td>
<td>OM Zonal</td>
<td></td>
</tr>
<tr>
<td>Rumino-Ammophilo</td>
<td>1 1 Sicilian oromediterranean scrub and related grasslands</td>
<td>OM Zonal</td>
<td></td>
</tr>
<tr>
<td>Daphnie-Festucetea</td>
<td>2 7 Greek and Aegean oromediterranean calciphilous grasslands and phrygana</td>
<td>OM Zonal</td>
<td></td>
</tr>
</tbody>
</table>

**Mediterranean vegetation**

<table>
<thead>
<tr>
<th>Class</th>
<th>Or A1</th>
<th>Characteristics</th>
<th>Biome/taxonality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thero-Brachypodio</td>
<td>4 15 Mediterranean terrestrial plant communities dominated by annual low-grown herbs and grasses</td>
<td>M Intrazonal</td>
<td></td>
</tr>
<tr>
<td>Lygeo-Stipetalia</td>
<td>3 9 Mediterranean pseudosteppe and related perennial grasslands</td>
<td>M Intrazonal</td>
<td></td>
</tr>
<tr>
<td>Cynio-Levanduletalia</td>
<td>1 7 Thermo-Mediterranean broom heathlands (retamal)</td>
<td>M Intrazonal</td>
<td></td>
</tr>
<tr>
<td>Cent-Compositae</td>
<td>2 5 Low-grown Mediterranean scrub (macchia, garrigue, tocot, and phrygana) on siliceous and edaphically variable substrates</td>
<td>M Zonal</td>
<td></td>
</tr>
<tr>
<td>Rosmarino-officinalis</td>
<td>6 16 Low-grown calciphilous Mediterranean scrub (maquis, matorral, garigue, tocot, and phrygana)</td>
<td>M Zonal</td>
<td></td>
</tr>
<tr>
<td>Nerio-Tamaricetea</td>
<td>3 8 Mediterranean riparian scrub</td>
<td>M Intrazonal</td>
<td></td>
</tr>
<tr>
<td>Quercetum Silesiacum</td>
<td>2 10 Mediterranean broad-leaved forests and maquis</td>
<td>M Intrazonal</td>
<td></td>
</tr>
</tbody>
</table>

**Semidesert vegetation**

<table>
<thead>
<tr>
<th>Class</th>
<th>Or A1</th>
<th>Characteristics</th>
<th>Biome/taxonality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organo-Salicion</td>
<td>3 10 Thermomediterranean and Macaronesian halophilius semidesert scrub</td>
<td>M Intrazonal</td>
<td></td>
</tr>
<tr>
<td>Artemisietum herbae</td>
<td>2 2 Aralo-Caspian subhumid semidesert</td>
<td>OM Zonal</td>
<td></td>
</tr>
</tbody>
</table>

**Mediterranean vegetation**

<table>
<thead>
<tr>
<th>Class</th>
<th>Or A1</th>
<th>Characteristics</th>
<th>Biome/taxonality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typhulo-Polygalietum</td>
<td>1 2 Macaronesian halophilous coastal dune scrub</td>
<td>SSD Intrazonal</td>
<td></td>
</tr>
<tr>
<td>Eleo-Euphorbio</td>
<td>1 1 Macaronesian succulent scrub on lava beds (tabaibal and cardonal)</td>
<td>SSD Zonal</td>
<td></td>
</tr>
<tr>
<td>Ophi-Thamnietum</td>
<td>2 2 Macaronesian maquis and related scrub</td>
<td>M Zonal</td>
<td></td>
</tr>
</tbody>
</table>

---

*continues*
The definition of Europe includes Europe proper, Iceland, Madeira, Azores, and the Canary Islands. For a full nomenclature, major species groups, and literature sources, see Mucina (1997a). Or, number of phytosociological orders; Al, number of phytosociological alliances. Biomes and orobiomes: T, Tundra Biome; B, Boreal Evergreen Forest Biome; N, Temperate Deciduous Forest Biome; M, Mediterranean Biome; CSD, Continental Semidesert Biome; SSD, Subtropical Semidesert Biome; S, Temperate Grassland (Steppe) Biome; L, Temperate Laurisilva Forest Biome; R, Temperate Rain Forest Biome; OA, Alpine Orobiome; OM, Mediterranean Orobiome. The abbreviations in parentheses indicate “in part” statements.

provides the major features of these classes), 223 orders, and 844 alliances. The number of associations is unknown, although one estimate suggests there might be as many as 2300–3000. For further details on the diversity of syntaxa occurring in various European countries, refer to the many national vegetation surveys (and transnational vegetation accounts) listed in Table IV.

Several interesting biodiversity patterns can be observed within the groups of syntaxa, including classes, orders, and alliances (Table III):

1. The Mediterranean region (supporting the Mediterranean biome and Mediterranean orobiome) shows a high concentration of classes and subordinated syntaxa. This can be partly ascribed to high alpha, beta, and gamma diversities (Cowling et al., 1996) and partly to the insular character of many types of Mediterranean habitats (and their complexes). Not only are numerous true marine islands very abundant within the Mediterranean basin but also the mountain summits form an archipelago of their own, demonstrating their own rates of evolution and combination of ecological and biogeographical factors leading to specific plant assemblages.

2. The high-latitude regions (arctic and boreal zones) show lower diversity of vegetation types partly as a result of low extant alpha diversity that diminished dramatically by the last ice age, partly because of the large-scale occurrence of uniform habitat complexes, and partly due to adversity of climate.

3. A group of azonal (and intrazonal) rather than broadly distributed zonal classes shows the highest diversity of orders and alliances (reflecting high diversity of ecological–biogeographic patterns). Among those with a particular position are Asplenietea trichomanis and Thlaspietea rotundifoli (classes comprising vegetation of special habitats such as rock faces, fissures, and screes), on the one hand, and classes of Festuco-Brometea, Koelerio-Corynephoretea, and Puccinellio-Salicornietea (comprising both primary and secondary dry grasslands, on the other hand. The former two classes have insular character of distribution pattern: The communities are small scale, occupying ecologically isolated habitats and housing most European endemic species. The group of dry grassland classes have relatively large-scale distribution and comprise Europe’s most species-rich plant communities. Russian steppes, for instance, can support more than 60 species per square meter. Large-scale distribution and rich species pools are prerequisites to high beta diversities manifested in a high number of vegetation types.

4. The synanthropic (man-made and man-controlled) vegetation of Europe is also very diverse in terms of the number of syntaxa. The classes of Festuco-Brometea, Koelerio-Corynephoretea, and Puccinellio-Salicornietea (comprising both primary and secondary dry grasslands) on the other hand. The former two classes have insular character of distribution pattern: The communities are small scale, occupying ecologically isolated habitats and housing most European endemic species. The group of dry grassland classes have relatively large-scale distribution and comprise Europe’s most species-rich plant communities. Russian steppes, for instance, can support more than 60 species per square meter. Large-scale distribution and rich species pools are prerequisites to high beta diversities manifested in a high number of vegetation types.

5. Mediterranean classes such as Thero-Brachypodietea, Rosmarinetea officinalis, Quercetea ilicis, and, to some extent, Cisto-Micromeretea also show high diversity with regard to the number of alliances. Still
### Major Monographs and Summary Accounts of Ecosystem, Habitat, and Vegetation Diversity in Europe

<table>
<thead>
<tr>
<th>Source</th>
<th>Country</th>
<th>Type</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goodall (1977–1999)</td>
<td>Society</td>
<td>SM</td>
<td>Accounts on major ecosystems of the world (including Europe)</td>
</tr>
<tr>
<td>Braun-Blanquet (1935–1943)</td>
<td>Europe</td>
<td>SM</td>
<td>Unfinished series of syntaxonomic monographs</td>
</tr>
<tr>
<td>Dieckhofer (1980)</td>
<td>Northern Europe</td>
<td>SM</td>
<td>Detailed accounts on vegetation of northern Europe</td>
</tr>
<tr>
<td>Ellenberg (1990)</td>
<td>Central Europe</td>
<td>SM</td>
<td>Last editions of a classic vegetation monograph</td>
</tr>
</tbody>
</table>
| Eiten (1976) | Europe | SM | Syntaxonomic survey of the Violina calaminariae
| Hartmann and John (1967) | Central Europe | SM | Woodland vegetation |
| Hartmann (1974) | Central Europe | SM | Woodland vegetation |
| Horvat et al. (1974) | Southeastern Europe | SM | Detailed account of vegetation of southeastern Europe and the Balkans |
| Mayer (1986) | Europe | SM | Woodland vegetation |
| Ozenda (1988) | Europe | SM | Vegetation monograph, translated into English |
| Pohlenz and Wallis (1985) | Europe | SM | Popular account of major forests and ecosystems of Europe |
| von Hobbelt (1980) | Central Europe | SM | Survey of moss communities |
| Communion of European Communities (1991) | Europe | R | CORINE habitat classification system |
| Davies and Snow (1990) | Europe | R | EUNIS habitat classification |
| Devilliers and Devilliers-Terschuren (1990) | Europe | R | Palearctic habitat classification |
| Pálhoss (1990) | Northern Europe | R | Nordic habitat classification system |
| Redfæl et al. (1990) | Europe | R | Syntaxonomic characterization of the EUHIS units |
| Khila and Hadić (1994) | Central Europe | C | Revised checklist of syna of Central Europe |
| Mucina et al. (1997) | Europe | C | Checklist of vegetation communities |
| Lohmeyer et al. (1982) | Europe | L | List of high-ranked syna of Europe |

#### National surveys

<table>
<thead>
<tr>
<th>Source</th>
<th>Country</th>
<th>Type</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mucina et al. (1993)</td>
<td>Austria</td>
<td>SM</td>
<td>Completed series in three volumes</td>
</tr>
<tr>
<td>Redfæl (1931–1998)</td>
<td>United Kingdom</td>
<td>SM</td>
<td>Completed series in five volumes</td>
</tr>
<tr>
<td>Václavík (1980 et seq.)</td>
<td>Slovak Republic</td>
<td>SM</td>
<td>Two volumes of the Slovak Vegetation Survey</td>
</tr>
<tr>
<td>Various editors (1990 et seq.)</td>
<td>Various countries</td>
<td>SM</td>
<td>Four volumes of the German Vegetation Survey</td>
</tr>
<tr>
<td>Čelina (1999)</td>
<td>Romania</td>
<td>SM</td>
<td>First volume of national survey</td>
</tr>
<tr>
<td>Pretzsch (1997)</td>
<td>Norway</td>
<td>SM</td>
<td>Monographic survey of major vegetation types</td>
</tr>
<tr>
<td>Moravec (1988)</td>
<td>Czech Republic</td>
<td>SM</td>
<td>First volume of national survey</td>
</tr>
<tr>
<td>Prindado Lora and Rivas-Martínez (1993)</td>
<td>Spain</td>
<td>SM</td>
<td>Account of vegetation of regions of Spain</td>
</tr>
<tr>
<td>Rahmonovics (1997)</td>
<td>Lithuania</td>
<td>SM</td>
<td>First volume of national survey (meadows)</td>
</tr>
<tr>
<td>Stadler (1950)</td>
<td>Poland</td>
<td>SM</td>
<td>Monographic survey of vegetation in two volumes</td>
</tr>
<tr>
<td>Tamásy (1980)</td>
<td>United Kingdom</td>
<td>SM</td>
<td>Monographic survey of vegetation in two volumes</td>
</tr>
<tr>
<td>Weerdt and Devoldere (1980)</td>
<td>Netherlands</td>
<td>M</td>
<td>Brief, syntaxonomic descriptive account of vegetation units</td>
</tr>
<tr>
<td>Apostolova and Slavova (1997)</td>
<td>Bulgaria</td>
<td>C</td>
<td>Account of vegetation units described between 1891 and 1995</td>
</tr>
<tr>
<td>Berkeš (1996)</td>
<td>Hungary</td>
<td>C</td>
<td>Account of Hungarian vegetation units</td>
</tr>
<tr>
<td>Braun-Blanquet et al. (1942)</td>
<td>France</td>
<td>C</td>
<td>Account of syna of the Mediterranean France</td>
</tr>
<tr>
<td>Juhn (1993)</td>
<td>France</td>
<td>C</td>
<td>Revised checklist of high-rank vegetation units of France</td>
</tr>
<tr>
<td>Korolev et al. (1991)</td>
<td>USSR</td>
<td>C</td>
<td>Revised checklist of vegetation units of the former USSR</td>
</tr>
<tr>
<td>Matschkevitsch (1983)</td>
<td>Poland</td>
<td>C</td>
<td>Key to identification of vegetation units</td>
</tr>
<tr>
<td>Moravec (1983)</td>
<td>Czech Republic</td>
<td>C</td>
<td>Revised checklist of vegetation units</td>
</tr>
<tr>
<td>Schmukha (1986)</td>
<td>Ukraine</td>
<td>C</td>
<td>Revised account of vegetation units</td>
</tr>
<tr>
<td>Lakić et al. (1987)</td>
<td>Bosnia</td>
<td>L</td>
<td>List of syna of Bosnia and Herzegovina</td>
</tr>
<tr>
<td>Rivas-Martínez et al. (1998)</td>
<td>Spain and Portugal</td>
<td>L</td>
<td>List of syna of the former Spain</td>
</tr>
</tbody>
</table>

#### *A bibliography of the featured surveys and checklists can be requested from the author.*

1 M, single monographic treatment; SM, series of monographic treatments; R, report serving a government body; C, conspectus (short account of vegetation units including limited descriptive treatment); L, list of vegetation units.
higher diversity becomes obvious on the level of associations reflecting the relict character and diversity of disturbance regimes showing a fine-grain pattern in the Mediterranean.

6. The lowest diversity of vegetation types is encountered in the classes typifying vegetation of the freshwater, marine, and coastal vegetation. Often, only a few orders or alliances are found within classes such as Lemnetae minor, Potametea, Spartinetea, Hockenyo-Elymetea, and Crithmo-Staticetea. However, on the whole, the nonterrestrial complex of habitats is very diverse, which is reflected in the high number of phytosociological classes.

3. Diversity of Habitat Types

Although vegetation forms, at least in the terrestrial ecosystems, the major component of these ecosystems, the habitat typology cannot reflect and is not reflecting only vegetation typology. For instance, as found by Rodwell et al. (1998), 60% of 277 units of the EUNIS3 habitat classification were characterized in phytosociological (syntaxonomic) terms, whereas the other units are largely abiotic and carry geomorphologic and hydrological features or they are associated with fauna or nonvascular plants. Table V shows that this is the case especially in EUNIS habitat classes (groups of similar habitats) such as class A (marine habitats) and class J (constructed, industrial and other artificial habitats).

The judgment on diversity of habitats appears more difficult than that based on classification of vegetation. First, there are many incommensurable variables to be considered by definition of a habitat type. Second, the spatial and temporal scaling of the habitat types

| TABLE V | Simplified System of EUNIS3 Habitat Units in Relation to Syntaxonomic Units1 |
|-----------------------------------------------|-------------------|-------------------|
| EUNIS3 habitat class/subclass | CI | Or | Al |
| **Marine habitats** | | | |
| Littoral rocks | — | — | — |
| Littoral sediments | 2 | 2 | 2 |
| Coastal salt marshes and saline habitats | 8 | 10 | 17 |
| Infralittoral (shallow subtidal or nontidal) rocks | — | — | — |
| Infralittoral (shallow subtidal or nontidal) sediments | 3 | 5 | 5 |
| Circalittoral (deep subtidal or nontidal) rocks | — | — | — |
| Circalittoral (deep subtidal or nontidal) sediments | — | — | — |
| Deep seabed (more than 200 m depth) | — | — | — |
| Pelagic water body | — | — | — |
| Anoxic marine habitats | — | — | — |
| Tidal or permanently flooded sea caves | — | — | — |
| **Coastal habitats** | | | |
| Coastal dune and sand habitats | 22 | 25 | 33 |
| Coastal shingle habitats | 2 | 4 | 4 |
| Rock cliffs, ledges, and shores, including the supralittoral | 3 | 3 | 3 |
| **Inland surface water habitats** | | | |
| Surface standing waters | 8 | 12 | 16 |
| Surface running waters | 2 | 2 | 3 |
| Littoral zone of inland surface water bodies | 9 | 9 | 9 |
| **Mire, bog, and fen habitats** | | | |
| Raised and blanket bogs | 2 | 4 | 7 |
| Valley bogs, poor fens, and transition mires | 4 | 5 | 11 |
| Aapa, palus, and polygon mires | — | — | — |
| Base-rich fens | 3 | 3 | 3 |
| Sedge and reedbeds, normally without freestanding water | 3 | 5 | 5 |
| Inland saline and brackish marshes and reedbeds | 2 | 2 | 4 |

continues
EUROPE, ECOSYSTEMS OF

<table>
<thead>
<tr>
<th>EUNIS3 habitat class/subclass</th>
<th>Cl</th>
<th>Or</th>
<th>Al</th>
</tr>
</thead>
</table>

**Grassland habitats**

| Dry grasslands                                    | 17 | 43 | 75 |
| Mesic grasslands                                  | 5  | 9  | 21 |
| Seasonally wet and wet grasslands                | 6  | 10 | 20 |
| Alpine and subalpine grasslands, tall herbs, and limes | 9  | 18 | 35 |
| Moss- and lichen-dominated habitats               | 3  | 3  | 3  |
| Inland saline grasslands                          | 3  | 7  | 7  |

**Heatlands and scrub habitats**

| Arctic and temperate scrub habitats               | 8  | 9  | 10 |
| Maquis, matorral, and submediterranean deciduous thickets | 11 | 13 | 34 |
| Arctic, alpine, and subalpine dwarf shrub habitats | 6  | 9  | 13 |
| Temperate heathland                               | 3  | 4  | 9  |
| Garrigues                                         | 10 | 12 | 33 |
| Spiny Mediterranean heath (phrygana and hedgehog heaths) | 12 | 15 | 34 |
| Thermo-Atlantic xerophytic habitats               | 3  | 3  | 4  |
| Shrub plantations                                 | —  | —  | —  |

**Woodland and forest habitats and other wooded land**

| Broad-leaved deciduous and mixed woodland         | 13 | 22 | 50 |
| Broad-leaved evergreen woodland                   | 7  | 8  | 16 |
| Coniferous woodland                               | 13 | 18 | 33 |
| Lines of trees, sparsely wooded land, and coppice | —  | —  | —  |

**Inland sparsely vegetated and unvegetated habitats**

| Terrestrial underground case, cave systems, and water bodies | — | — | — |
| Screes                                                | 1  | 9  | 35 |
| Inland cliffs and exposed rock habitats               | 7  | 19 | 22 |
| Frost or ice-dominated habitats                       | —  | —  | —  |
| Inland solstitial and organic habitats with very sparse or no vegetation | 1  | 1  | 1  |
| Volcanic features                                    | 3  | 3  | 3  |

**Regularly or recently cultivated habitats and gardens**

| Arable land and market gardens                      | 2  | 2  | 2  |
| Gardens                                              | 3  | 2  | 2  |

**Constructed, industrial, and other artificial habitats**

| Buildings of cities, towns, and villages            | 2  | 2  | 2  |
| Low-density buildings                               | —  | —  | —  |
| Extractive industrial sites                         | —  | —  | —  |
| Transport networks                                  | —  | —  | —  |
| Highly artificial man-made waters and associated structures | —  | —  | —  |
| Waste deposits                                      | —  | —  | —  |

*From Davies and Moss (1997) and Rodwell et al. (1998). Cl, Or, and Al: number of phytosociological classes, orders, and alliances, respectively.*

is very complex. This complexity is reflected in all classification systems of habitats known until today. In order to cope with the complexity, the habitat classification systems are hierarchical, but often on the same level of hierarchy, thus mixing small-scale habitats with habitat complexes recognizable on the landscape level. Obviously, the calibration of the habitat units and the classification system leaves much room for improvement (Rodwell et al., 1998; Waterton et al., 1998).
IV. CHALLENGES FOR THE FUTURE

There are many perspectives on ecological diversity. I present one that is not the favorite of granting agencies in Europe (and perhaps also elsewhere): the perspective of the diversity of ecosystem types themselves.

The legitimate status of the concept of ecological diversity as a virtual part of biodiversity has been recognized since “biodiversity,” a brushed-up concept, started to sweep research agendas throughout the world. This happened largely due to the catalyzing effect of the U.S. Strategy Conference on Biological Diversity (Ghilarov, 1996) and later “beatification” of biodiversity by the Rio World Summit and resulting documents of the convention. Ecological diversity (including diversity of communities, ecosystems, landscapes, and biomes) was recognized by international and national bodies as a subject worth studying (and supporting) in the name of developing technological tools for its protection and sustainable use.

In Europe, as elsewhere, taxonomic diversity is the leading topic of biodiversity research and conservation. However, due to deeply rooted research traditions and the awareness of national governments, surveys on higher levels of biological complexity have been given much attention in the past. Almost every European nation has a national vegetation map, and many have land-use and landscape-unit maps. Modern national vegetation surveys have been recently completed or are under way in many crucial European countries. [see Table IV for a survey; see Mucina et al. (1994) and Rodwell et al. (1993) for progress reports].

With Europe growing together as a union of nations sharing political agendas and an economic future, the large-scale tools for ecological diversity conservation and sustainable use are emerging. These include pan-European habitat classification scheme EUNIS, a vegetation map of Europe (a long-term cooperative project headed by Dr. U. Bohn, Bonn, Germany), and the European Vegetation Survey (a Working Group of the International Association for Vegetation Science). The fall of the Iron Curtain in 1989 and 1990 posed a new challenge for research agendas in ecological diversity—unification of scientific standards and application tools along the West–East gradient within Europe. There is still much progress to be made.

Also, because we need to know the flora and fauna (as well as their intraecosystem variability down to the genetic structures), we need to survey biotic communities, habitats, ecosystems, and biomes. This need is not only fed by scientific curiosity but also by practical, technological needs to ensure effective conservation and use of biological diversity as a whole.

Acknowledgments
I am indebted to Milam Chitty, Harriet Dutschke, Julian Dring, Sandro Pignatti, John Rodwell, and Joop Schaminee for long and fruitful discussions on the diversity of European vegetation—our common passion. Roberto Camillo carefully read the penultimate manuscript version.

See Also the Following Articles
AFRICA, ECOSYSTEMS OF • ASIA, ECOSYSTEMS OF • AUSTRALIA, ECOSYSTEMS OF • NORTH AMERICA, PATTERNS OF BIODIVERSITY IN • SOUTH AMERICA, ECOSYSTEMS OF

Bibliography
Prentice, I. C., Guost, J., Humbley, B., Jolly, D., and Chadali, R.
(1996): Reconstructing biomes from palaeocological data: A general method and its application to European pollen data at 0 and 6 ka. *Climate Dynamics* 12, 185.


EUTROPHICATION AND OLGOTROPHICATION

JoAnn M. Burkholder
North Carolina State University

I. Introduction
II. Changes in Flora and Fauna
III. Oligotrophication: Reversing the Impacts of Eutrophication

GLOSSARY

algae Primitive plantlike organisms that photosynthesize. These organisms can be unicellular, filamentous, or colonial microscopic forms (microalgae), or they can be macroscopic (macroalgae), consisting of a primitive plant body (thallus) that lacks vascular tissue in most species.
anoxia Condition in which there is almost no dissolved oxygen in the water (<0.1 mg of dissolved oxygen (DO)/L).
bloom Proliferation of algae in river, lake, estuarine, or marine waters. Older literature referred to a bloom as 3000 or more algal cells per liter (L) (or 5 cells per ml), although this density generally is too low to discolor the water. Algal 'blooms' range in cell density from eukaryote blooms (e.g., dinoflagellates at $10^3$ to $10^4$ cells/ml, to cyanobacteria at $10^8$ to $10^9$ cells/ml).
eutrophic Trophic status of an aquatic ecosystem that is characterized by relatively low phytoplankton species diversity but high phytoplankton production (biomass as mean chlorophyll a ca. 15–40 μg/L), with the phytoplankton often dominated by cyanobacteria in lakes, and by dinoflagellates or other flagellates in estuaries; high nutrient concentrations and loadings (for example, in lakes, mean inorganic N ca. 1900 μg/L, mean total P ca. 80 μg/L), high decomposition in the bottom water and surface sediments (with abundant organic materials available for this process); and bottom-water dissolved oxygen deficits, sometimes with occasional to frequent fish kills. Eutrophic lakes typically are shallow with well-developed littoral zones (area where light penetration is sufficient to support growth of rooted plants), sometimes extending across most of the bottom area.
harmful algae Algae that are undesirable to humans because (a) they become too abundant in response to nutrient overenrichment and then, at night, use most or all of the oxygen in the water for their respiration, so that fish and other organisms suffocate or become seriously physiologically stressed; (b) they become too abundant in response to nutrient enrichment, and overgrow beds of desirable rooted vegetation so that the beneficial plants cannot receive enough light to survive; (c) they cause or promote disease in other plants or animals, or (d) they produce toxins that hurt or kill finfish, shellfish, or other higher trophic levels including humans. "Harmful algae" include prokaryotic cyanobacteria or blue-green algae. More recently, the term has been used to include organisms that are not photosynthetic, primitive plantlike organisms—for example, certain nontoxic animal-like dinoflagellates, which cause fish disease (e.g., Amyloodinium ocellatum), and toxic animal-like dinoflagellates (e.g., the toxic Pfiesteria complex), which do not have their own chloroplasts
mesotrophic
Trophic status of an aquatic ecosystem that is characterized by moderate phytoplankton production and moderate nutrient concentrations and loadings.

mixotrophy
Form of nutrition involving both autotrophic (photosynthetic) and heterotrophic carbon acquisition.

oligotrophic
Trophic status of an aquatic ecosystem that is characterized by relatively high phytoplankton species diversity but low phytoplankton production (chlorophyll $a \leq 10 \mu g/L$); low nutrient concentrations and loadings (for example, mean inorganic N $< 200 \mu g/L$, mean total P $< 10 \mu g/L$); low decomposition (with little organic material available to decompose); and plentiful oxygen throughout the water column.

salt wedge
Water from the ocean, with higher salt content, that moves into an estuary along the bottom of the water column, beneath less dense fresh water that has moved into the same area from a river. The salt content makes the ocean water heavier than the fresh (riverine) water, so that under calm conditions, the estuarine water becomes density-stratified. This “salt wedge” of bottom water can become somewhat isolated from the overlying fresh or less brackish water. The longer the period in which the total water column is not mixed by winds or storms, the more distinct the two water layers or strata become. Salt wedges most often develop in warm seasons when plantlike phytoplankton production is high in the surface waters and respiration by heterotrophs (bacteria, fungi, animals) is also high, especially in the lower water column and sediments. Nutrients typically are higher within the salt wedge than in the overlying water, because of decomposition processes. At the same time, the bottom-water salt wedge can become hypoxic or anoxic, underlying waters can be saturated or supersaturated with oxygen from phytoplankton photosynthesis.

trophic status
Ranking system for aquatic ecosystems, based on the amount of organic production and nutrient (N,P) levels. The major component that is usually considered in assigning trophic status is phytoplankton production, but this can be misleading. For example, some lakes are classified as oligotrophic because water-column nutrients and phytoplankton production are low, despite the fact that benthic plant production (e.g., of rooted angiosperms) is high.

EUTROPHICATION IS the natural aging process of aquatic ecosystems, formerly used mostly in reference to the natural aging of lakes wherein a large, deep, nutrient-poor lake eventually becomes more nutrient-rich, more productive with plant and animal life, and slowly fills in to become a pond, then a marsh. More recently, this term has been used in abbreviated reference to “cultural” or accelerated eutrophication of lakes, rivers, estuaries, and coastal marine environments, wherein the natural eutrophication process is accelerated (often by hundreds or thousands of years) by human activities that add nutrients to the aquatic system. Oligotrophication is a reversal of the eutrophication process, in which waters become less nutrient-enriched and/or support less plant and animal production. This process can occur naturally, for example, when the inflow to a moderately productive lake is severed so that the lake slowly becomes an ombrotrophic, nutrient-poor system or bog. Oligotrophication can also occur from human intervention, for example, following diversion of sewage from a small lake and dredging/ removal of nutrient-rich bottom sediments.

### I. INTRODUCTION

#### A. Overview across Aquatic Habitats

From freshwater lakes to estuaries and marine coastal waters, human-derived sources of nutrient pollution have rapidly changed water quality and aquatic community structure within the past 200 years of industrialization and rapid population growth. The predominant theme throughout most of the world is increased nutrient enrichment or cultural eutrophication, rather than nutrient decreases or oligotrophication; thus, eutrophication will be emphasized here although both phenomena will be addressed. Despite advances in treatment of human sewage in some countries during the late 20th century, despite bans on use of phosphorous in certain domestic or industrial practices, despite declines in agricultural fertilizers in some geographic regions,
EUTROPHICATION AND Oligotrophication

and despite modest improvements in environmental education in localized areas, the massive recent increase in global human population growth has increased nutrient loadings to aquatic ecosystems (Fig. 1). These increases have been greatest in estuarine and coastal marine areas where population growth has been highest, and where nearly two-thirds of the people of the world now reside. As a reflection of this trend, estuaries have been reported to receive more nutrient inputs per unit surface area than any other type of aquatic ecosystem. Nutrients are essential for primary production by phytoplankton, benthic micro- and macroalgae, and aquatic angiosperms, which directly or indirectly support aquatic food webs. In freshwaters, phosphorus is the least abundant among the nutrients needed in large quantity (macronutrients) by biological organisms. Thus, it is the first element that becomes limiting to biological productivity in many freshwater systems. Phosphorus is also the nutrient that limits plant growth in many tropical coastal marine waters. In temperate and polar coastal marine environments, nitrogen is the most important nutrient that first limits primary production; and at the estuarine interface between marine and freshwater habitats, both N and P can "colimit" plant production, especially in late winter-spring seasons of high precipitation and accompanying high inorganic N inputs. Other nutrients, notably silica and iron, can also sometimes significantly influence the outcome of species dominance and the structure and abundance of phytoplankton communities under cultural eutrophication.

In moderation, nutrients promote beneficial increases in phytoplankton and benthic algal production and, in turn, higher production of zooplankton, macroinvertebrates, fish, and shellfish that utilize the plant production directly or indirectly for food. But when added in excess, nutrient pollution can cause overgrowth of macro- and macroalgae, leading to oxygen depletion in bottom waters and sometimes throughout the water column. Although fish kills can be an obvious sign of acute impacts from cultural eutrophication, associated subtle, chronic impacts may be more damaging to aquatic communities over the long term. Sustained loss in biodiversity is foremost among these chronic impacts, resulting from underlying mechanisms that scientists are only beginning to understand.

Cultural eutrophication promotes major shifts in the structure of both plant and animal communities, generally affecting dominant components of every trophic level from microbial decomposers to macrofauna. There is clear, compelling evidence of altered aquatic community structure and significantly reduced biodiversity from cultural eutrophication in many freshwater, estuarine, and marine ecosystems. Surface waters across the earth are now sustaining such impacts; even the open oceans are no longer sufficiently isolated to avoid nutrient pollution from atmospheric deposition. Similar trends have been demonstrated from the microfossil records of lakes and of estuaries, showing a dramatic increase in nutrients and associated organic carbon deposits and a sharp, sustained decrease in the diversity of aquatic species.

In moderation, nutrients promote beneficial increases in phytoplankton and benthic algal production and, in turn, higher production of zooplankton, macroinvertebrates, fish, and shellfish that utilize the plant production directly or indirectly for food. But when added in excess, nutrient pollution can cause overgrowth of macro- and macroalgae, leading to oxygen depletion in bottom waters and sometimes throughout the water column. Although fish kills can be an obvious sign of acute impacts from cultural eutrophication, associated subtle, chronic impacts may be more damaging to aquatic communities over the long term. Sustained loss in biodiversity is foremost among these chronic impacts, resulting from underlying mechanisms that scientists are only beginning to understand.

Cultural eutrophication promotes major shifts in the structure of both plant and animal communities, generally affecting dominant components of every trophic level from microbial decomposers to macrofauna. There is clear, compelling evidence of altered aquatic community structure and significantly reduced biodiversity from cultural eutrophication in many freshwater, estuarine, and marine ecosystems. Surface waters across the earth are now sustaining such impacts; even the open oceans are no longer sufficiently isolated to avoid nutrient pollution from atmospheric deposition. Similar trends have been demonstrated from the microfossil records of lakes and of estuaries, showing a dramatic increase in nutrients and associated organic carbon deposits and a sharp, sustained decrease in the diversity of aquatic species.

**FIGURE 1** Export of total nitrogen from watersheds surrounding the North Atlantic Ocean, as a function of net anthropogenic inputs of nitrogen to their watersheds. Net anthropogenic inputs are defined as industrial N fertilizer + N fixation by legume crops + atmospheric inputs of oxidized N = net imports of N in food and feedstock. Reprinted from Vitousek et al. (1997), originally from Kluwer Academic Publishers, Biogeochemistry 35, 181–226, Howarth et al., Fig. 3a (1996) with kind permission from Kluwer Academic Publishers.
B. Factors Influencing Impacts from Nutrient Enrichment

Among the most important factors that influence the extent of adverse impacts from nutrient enrichment is water exchange rate, or the time required to completely replace the water in the system with new water. Waters that can “self-cleanse” or rapidly flush, such as fast-flowing rivers or run-of-river impoundments, are less sensitive to elevated nutrient loading than more contained or enclosed waters such as lakes and estuarine lagoons. Similarly, waters along open, wave-swept marine coasts are less sensitive to nutrient loading than quiet, sheltered upper embayments. The occurrence of other pollutants along with the nutrients—for example, suspended solids that can reduce light for growth of aquatic plants or toxic substances and harmful bacteria that can stress or kill some aquatic species—can exacerbate impacts from nutrient enrichment. The form of the nutrients can also be important. For example, swine wastes are much richer than untreated human sewage in organic nutrients (N, P, and C). Certain undesirable algae prefer organic nutrient forms over inorganic sources.

The timing and frequency of the source inputs are important, as well. During high-precipitation periods in winter, sewage is more effectively diluted and cold temperatures discourage the growth of many nuisance algae. Also, in some waterways, agriculture is the major source of nutrients annually, but during low-flow periods in warmer seasons, sewage can contribute half or more of the river volume and the loading of nutrients—at a time when the readily available nutrients stimulate noxious algal overgrowth of the system.

The initial aquatic community structure also influences the overall impacts of nutrient loads to aquatic ecosystems. Systems with low nutrient enrichment rely more on recycled nutrients than on introduction of new nutrient sources, and this may be a driving evolutionary force leading to greater specialization and diversity, as suggested by Howarth and others. Diverse communities, characteristic of oligotrophic or nutrient-depauperate waters, tend to be dominated by sensitive species that are limited to a narrow range of environmental conditions. Oligotrophic ecosystems are more sensitive (or less resistant) overall, to stress from nutrient enrichment. The most predictable outcomes are loss of sensitive species and increased abundance of generalist or opportunistic species that are more resistant to the undesirable water quality changes and other stresses affected by nutrient pollution. An ecosystem becomes more eutrophic (that is, increasingly stressed by eutrophication), it becomes more resistant to further change.

Increasingly eutrophic lakes, lower rivers, and estuaries are commonly characterized by seasonal low oxygen stress. Hypoxia and anoxia develop when oxygen consumption exceeds supply, as a result of two major factors. The algae—often in densities of millions to billions of cells/ml of water—often are the first food web component to be stimulated by nutrient enrichment. They are net producers of oxygen through photosynthesis during the day, but at night they consume oxygen for respiration. Dense populations of algae in the upper and mid-depths of the water column can consume most of the available dissolved oxygen. Low-oxygen stress in the system also occurs because of increased decomposition. As plants and animals die and settle out over the growing season, their remains are decomposed by bacteria and fungi in oxygen-demanding processes that can rapidly deplete the oxygen from the lower water column, sometimes extending to mid-depths or shallower. Oxygen solubility decreases with increasing temperature. Thus, low-oxygen stress tends to be most pronounced in warmer seasons when algal biomass, decomposition rates, and respiratory rates (and oxygen requirements) of fish and other animals generally are high.

Low-oxygen conditions are exacerbated when bottom waters become somewhat isolated from the upper water column that receives oxygen (i) during the day from phytoplankton growth, and (ii) throughout the diel cycle from the overlying air as it diffuses into the water, and from phytoplankton growth (during the day). The isolating effect occurs when the water column becomes thermally (density) stratified, as in the bottom water layer (hypolimnion) of a stratified lake, or salinity (density) stratified, as in the lower water-column salt wedge in an estuary. As a result, sessile bottom-dwelling animals become physiologically stressed and may suffocate unless they can slow their metabolism until oxygen is replenished. Mobile animals that rely on the bottom waters as critical nursery areas for their young, or as a refuge area to escape predation, can physically avoid the low-oxygen areas but may nonetheless suffer population declines because of critical habitat loss.

Organic matter in poorly treated sewage and animal waste contributes to low-oxygen stress in aquatic systems, as another source of material for decomposition or “biochemical oxygen demand.” Chemical oxygen demand may further deplete oxygen, through oxidation of high levels of ammonia and other inorganic reduced compounds in the wastes. These oxygen-demanding processes often are relatively localized problems oc-
EUTROPHICATION AND OLIGOTROPHICATION

653

 occurring, for example, near sewage outfalls or animal waste spills. In contrast, certain oxidized nutrients such as nitrate can be highly soluble in water and, therefore, can be transported considerable distances as shown, for example, by Mallin et al.’s work. This transport creates a time lag between the introduction of nutrients into one area and adverse impacts at some distance from the source. Oxygen consumption from decomposition of excess phytoplankton production thus can occur on a large scale that is difficult to track or to relate to one specific, original source.

II. CHANGES IN FLORA AND FAUNA

A. Microalgae

1. Species Shifts across Nutrient Gradients

Phytoplankton dominate the flora of oligotrophic systems (Fig. 2). They respond quickly to nutrient inputs because the tiny plantlike organisms are immersed in the enriched medium, in contact with it on all surfaces. Phytoplankton with optimal growth at elevated nutrient concentrations are especially stimulated or “selected for” by nutrient enrichment, and they eventually overgrow and replace species that grow best at lower nutrient levels. Along a nutrient enrichment gradient from oligotrophic to highly nutrient-enriched, the phytoplankton community structure gradually shifts from low abundance of many species and dominance by small flagellates and picoplankton (algae ≤ 2 μm in diameter), with energy flow channeled through a microbial loop of bacteria and small flagellates rather than directly up the food chain to herbivorous zooplankton, to high abundance of relatively few species consisting mainly of large cells or large colonies as well as seasonally abundant flagellates. As lakes, lower rivers, and estuaries become more eutrophic, diatom species with higher N and P optima predominate in colder periods, certain colonial green algae with high N optima are abundant in early summer, and dinoflagellates (Dinophyceae: Gymnodinium spp., Peridinium spp.) and filamentous and colonial cyanobacteria with high P optima dominate in late summer. If dissolved silica, needed by diatoms to make their cell walls, is limiting in colder seasons, flagellates (as examples, dinoflagellates, cryptomonads, euglenoids) become more abundant. In shallow freshwater lakes, estuaries and lagoons where open-water habitat is limited, high phytoplankton densities occasionally occur in response to nutrient enrichment (for example, high flagellate densities that historically occurred in response to excrement from duck farms in western Long Island Sound). More commonly, benthic algae known as epiphytes (growing on the leaves of submerged aquatic vegetation or SAV), or floating “drift” macroalgae (in sheltered estuarine and marine coastal embayments), are rapidly stimulated by the nutrient increases and restrict light for underlying plants (Figs. 2 and 3). Increased nutrient enrichment generally promotes a decrease in the species diversity of benthic, photosynthetic algae, and a shift in dominance from diatoms to filamentous green algae or coccoid cyanobacteria. The combination of low light and rich organic substrates, characteristic of planktonic and (more so) benthic habitats in such systems, also selects for mixotrophic microalgae, which become abundant and diverse under moderately eutrophic conditions.

Many regions of the world have landscapes domi-
EUTROPHICATION AND OLIGOTROPHICATION

FIGURE 3
Generalized shift in biomass of major plant groups with increasing nutrient inputs to shallow marine ecosystems. Occasionally the phytoplankton dominate, but usually macroalgae dramatically increase, while submerged rooted plants decline through competition for light and/or nitrogen. Reprinted from Harlin (1993), with permission.

...nated by run-of-river impoundment or reservoirs, rather than natural lakes. Similar trends in impacts from nutrient loading on aquatic communities can occur over time in reservoirs as in lakes, with the exception of depauperate rooted macrophyte populations (discussed later) in reservoirs with variable depth imposed by controlled drawdown involved in potable water supply management or flood control. Impacts of cultural eutrophication in reservoirs can be mitigated by rapid water exchange (weeks to months, rather than years in most natural lakes). Moreover, the high turbidity from sediment loading/resuspension, characteristic of many reservoirs and estuaries as well as some natural lakes, often makes available light—rather than nutrients—the primary resource limiting the productivity of phytoplankton and other flora in the systems. Nutrients act as secondary factors controlling plant growth under such conditions. Such systems often can sustain higher nutrient (especially P) loading than clear, natural lakes while supporting less phytoplankton biomass because of their features of more rapid water exchange and light limitation from high suspended sediments.

"Bottom-up" control by nutrients interacts with grazing pressure from zooplankton and other fauna in "top-down" trophic-level effects. Such effects are well documented in freshwater lakes and streams, and can also be operative in estuaries and coastal marine waters. Many field and laboratory studies have demonstrated that proliferation of phytoplankton and benthic microalgae under nutrient enrichment (for example, seagrass epiphytes as described by Neckles and colleagues) can be significantly reduced during periods when grazers are abundant. Thus, herbivores can sometimes alleviate eutrophication impacts by holding algal production in check. Decreased grazing pressure can allow higher algal biomass to develop in response to excess nutrients, but decreases in grazing can help promote algal blooms only where nutrient inputs are sufficiently high to support such blooms. Therefore, grazing pressure is regarded as a secondary factor controlling phytoplankton production under cultural eutrophication.

2. Long-Term Human Influence
Long-term data sets have provided two lines of compelling evidence in support of major impacts from cultural eutrophication on phytoplankton assemblage structure in aquatic ecosystems, with potentially serious ecological and economic ramifications. First, the geological record in sediment cores taken from many freshwater lakes and estuaries clearly shows that long-term major shifts have occurred under increased nutrient enrichment—progressing from a balance between species-rich planktonic (mostly centric) and benthic (mostly pennate) diatom assemblages, to dominance by plank-
tropic diatoms with low species diversity (Fig. 4). Planktonic diatoms are selected for because they are the early recipients of water-column nutrient sources. Benthic diatoms encounter decreased light because of the overlying planktonic growth, eventually leading to decreased production.

The second line of evidence concerns living phytoplankton assemblages with species shifts that have been related to changes in supply ratios of the two most limiting nutrients. Under nutrient-limiting conditions, species with similar optima in other environmental factors (for example, temperature or light) that maintain faster growth compete more successfully for the available nutrient resource (Figs. 5–7). Elegant work by Tilman and colleagues examined the response of freshwater diatom species to shifts in Si : P ratios. Rhee and colleagues extended these concepts to controlling influences of N : P ratios. Low molecular N : P ratios (7–15 in those studies; up to ca. 29 : 1 by weight in Smith’s work) favored tested blue-green and diatom species, whereas higher N : P ratios favored green algae. Many green algae grow optimally at high N, concentrations, whereas many blue-greens and dinoflagellates have a high P requirement. Diatoms, unlike the other algae, require major supplies of silica as well as N and P. When Si : N and Si : P ratios are high, the available Si favors growth of diatoms that can effectively compete for N and P resources. However, as Si : N and Si : P ratios decrease, silica becomes limiting for diatom growth, and more N and P remain available for growth of flagellates and other algae that do not require silica. These concepts were extended to natural lakes by Tilman and colleagues (for diatoms and other algae), Smith (for blue-greens), and others.

Silica is only slowly (years) made available for new diatom growth, through dissolution of dead diatoms and other natural silica sources. In contrast, N and P cycles are rapidly affected by anthropogenic inputs. As N and P enrichment increase, the Si : N and Si : P supply ratios are depressed. Eutrophication can effect a decrease in dissolved Si abundance by initially stimulating high growth of diatoms, to the point that they deplete the dissolved silica pool needed by developing diatom populations in subsequent seasons. Long-term data sets on estuarine, coastal, and freshwater phytoplankton communities indicate that shifts to dominance by flagellated algae or blue-greens—including some harmful bloom-forming species—have coincided with decreased abundance of diatoms and decreased Si : N and Si : P ratios. Such trends have been documented by Smayda and others in temperate and sub-Arctic waters such as the Great Lakes, the Gulf of Mexico near the mouth of the Mississippi River, New England coastal waters, Chesapeake Bay estuarine waters, the Black Sea, and the coasts of northern Europe.

In oligotrophic tropical marine waters, mixotrophic dinoflagellates tend to predominate and symbiotic interactions are common in both planktonic and benthic communities. Little is known about impacts of eutrophication on phytoplankton community structure in such systems, but it is hypothesized that increased nutrient enrichment would shift the community structure to higher proportions of flagellated photosynthetic algae with less reliance on heterotrophy or symbiosis. Trace metals such as iron have also been shown to be limiting to phytoplankton growth in some estuarine and marine waters, where N and P are at levels that would otherwise be expected to support more algal production.

As eutrophication progresses, the previously described shift in temperate-zone phytoplankton community structure from certain diatoms to other diatom species, flagellates, and cyanobacteria causes subtle, but important, undesirable changes for the food web that can adversely affect secondary production. For example, as reviewed by Kilham and colleagues, some diatom species produce high quantities of certain lipids that are essential for zooplankton reproduction. Algal species that may replace these diatoms under increasingly eutrophic conditions do not produce these lipids, or produce much fewer of them. Analogous phenomena occur in fresh waters, estuaries, and coastal environments. For example, Starr and colleagues reported that spawning of green sea urchins and blue mussels apparently is triggered by a heat-stable metabolite that is released in high abundance by certain species of phytoplankton, especially certain diatoms such as Skeletonema costatum (Bacillariophyceae). This substance is not produced, or is produced in much lower quantities, by flagellated algae that replace these diatoms under cultural eutrophication.

3. Harmful Algal Blooms and Anthropogenic Nutrient Enrichment

Among the algal species favored by nutrient enrichment are noxious forms such as cyanobacteria that are toxic to zooplankton, fish and wildlife in fresh waters and certain estuaries (Baltic Sea and Australia), and certain dinoflagellates that are toxic to finfish and shellfish in estuarine and marine coastal areas worldwide (Table I). For example, the two known toxic Pfiesteria species have been most active in waters degraded by poorly treated sewage, swine effluent spills, and other excessive nutrient inputs, and these organisms have been experimentally stimulated by nutrient enrichment. However,
The geological record of cultural eutrophication in Chesapeake Bay, showing major changes since European settlement. The data from four sediment cores are graphed by the average date assigned to each sediment sample (depth layer) according to radiocarbon and pollen methods. (a) Total organic carbon (TOC, indicating total system productivity) preserved over time (historic record from A.D. 150 to A.D. 1990). Modified from Cooper and Brush (1991), with permission. (b) Diatom cell numbers per year (A.D. 400 to A.D. 1990). (c) Diatom community diversity calculated as Shannon's H (A.D. 400 to A.D. 1990). (d) Centric/pennate diatom ratios (A.D. 400 to A.D. 1990). TOC, diatom numbers, and the centric/pennate diatom ratios all showed a significant and abrupt increase following the time of European settlement in the 1700s. The total diatom community diversity, in contrast, significantly decreased post-1700s, relative to pre-1700s diversity. Graphs b–d were modified with permission from Cooper © (1999) American Association for the Advancement of Science.
EUTROPHICATION AND OLIGOTROPHICATION

FIGURE 5 The outcome of competition between diatoms for limiting nutrients (silica or phosphate) across a gradient of Si : P ratio comparing two freshwater planktonic diatoms, Cyclotella meneghiniana and Asterionella formosa, under growth-limiting conditions (temperature, light, and other conditions held constant). These chemostat culture experiments indicate conditions in which A. formosa is dominant because it can outgrow C. meneghiniana at very low P (as micromolar concentrations or micromoles per liter of phosphate, μM P3), conditions wherein the two species can coexist (circles) because one is limited by P, and the other is limited by Si (as silicon dioxide concentrations, μM SiO2); and conditions wherein C. meneghiniana can dominate because it can outgrow A. formosa at lower Si. Reprinted from Tilman (1982), with permission.

Other species of harmful algae are found in oligotrophic waters and do not appear to be favored by elevated nutrients. The planktonic Gulf Coast red tide dinoflagellate, Gymnodinium breve, and certain benthic toxic dinoflagellates that inhabit coral reefs (Gambierdiscus toxicus, Prorocentrum lima) are among various harmful species that have been reported to achieve optimal growth in low-nutrient conditions.

Despite their economic impacts worldwide, remarkably little is known about the nutritional ecology of most species of estuarine and marine toxic, parasitic, and other harmful algae. Overgeneralizations should be avoided that prematurely negate the potential for stimulation of various other harmful/toxic estuarine and marine algae by nutrient enrichment. As additional species are thoroughly evaluated, it will be important to consider that the concentration of a nutrient at any given point in time may not be correlated with its actual bioavailability, and that phytoplankton can grow for long periods on internally stored (luxury-consumed) nutrients.

FIGURE 6 The outcome of competition among diatoms, green algae, and blue-green algae for limiting nutrients (nitrogen or phosphorus) across a gradient of N : P ratios under silica-replete (nonlimiting) conditions, showing additional influence of temperature. The comparison includes representative species of freshwater lake diatoms (circles), green algae (triangles), and blue-greens (cyanobacteria; diamonds) in chemostat culture under growth-limiting conditions for N and P (but not for Si). Note that at the coldest temperature indicative of conditions during spring and fall seasons in this north temperate lake, the cold-optimal diatoms outcompeted the greens and blue-greens across all N : P ratios. That is, temperature, rather than N or P, was the most important condition limiting the growth of greens and blue-greens relative to diatoms. The midrange temperature represented a transitional area where some growth could occur for species representing each group. The highest temperature favored warm-optimal blue-greens (that tend to have high P optima), especially at low N : P ratios when P was more abundant. Reprinted from Harper (1992), with permission; originally modified from Tilman et al. (1986).
FIGURE 7  The outcome of competition among diatoms, green algae, and blue-green algae for limiting nutrients (phosphorus or silica) across a gradient of P : Si, showing the influence of temperature. The outcome of competition is indicated among freshwater lake diatoms (circles), green algae (triangles), and blue-greens (cyanobacteria; diamonds) in chemostat culture under growth-limiting conditions for Si (needed only by the cold-optimal diatoms) and P at two temperatures. At the lower temperature, cold-optimal diatoms dominated across all Si : P ratios. However, at the higher temperature they were able to dominate over green algae only at high Si : P ratios, when their Si requirements were alleviated by the relatively high Si. The blue-green species were able to outcompete the other groups only at one transitional Si :P ratio even at the higher temperature, which was still suboptimal for these organisms. Reprinted from Harper (1992), with permission; originally modified from Tilman et al. (1986).

nutrient pools. The luxury consumption and subsequent use of inorganic N, P, and C; the cell budgets for these nutrients; and the role of heterotrophy (including indirect stimulation by nutrient enrichment, through increased growth of microbial prey) in supplying carbon to be examined for most harmful estuarine and marine species.

B. Macroalgae

1. Freshwater Assemblages

The major flora of certain aquatic habitats are macroalgae, and they include some notorious aquatic “weeds” that respond to cultural eutrophication. Temperate-zone freshwater, oligotrophic hard-water lakes may contain abundant benthic charophytes and, rarely, a few species of small brown macroalgae (Phaeophyceae). Oligotrophic soft-water lakes may have populations of small, benthic red macroalgae (Rhodophyceae), especially near spring-fed areas with bubbling carbon dioxide. As eutrophication progresses, populations of certain filamentous green algae and cyanobacteria that would otherwise be considered as microalgae become visually common, especially under high P enrichment. For example, Cladophora glomerata (Chlorophyceae) forms masses of long hairlike growth (dominant during the 1960s in the littoral zone of the west basin of Lake Erie); and cyanobacteria (or blue-green algae, Cyanophyceae; for example, Anabaena, Aphanizomenon, Microcystis) form tufts or amorphous masses of greenish, bluish-green, or reddish gray slime.

Macroalgae are the dominant autotrophs of midorder streams where hard substrata are available for colonization and light is moderate to high, especially after leaf fall in colder seasons. Stream-inhabiting macroalgae are often much more diverse and abundant than lake macroalgal floras. Nutrient-poor (often soft-water) stream segments are colonized by diverse, abundant cold-optimal chrysophytes (e.g., Eunotia pectinalis [Bacillariophyceae], Tetrasporopsis [Chrysophyceae]), green algae (Spirogyra, Tetraspora, Oedogonium), red algae (Batrachospermum, Paralemanea [Rhodophyceae]), or cyanobacterial mat formers (Phormidium). As eutrophication progresses, soft waters may develop larger populations of filamentous green algae such as certain species of Oedogonium and Mougeotia or dense growth of colonial Hydrodictyon and Stigeoclonium (also known to tolerate high metal concentrations that may be found in poorly treated sewage). Hard-water streams under high P enrichment may become choked with massive growth of Cladophora spp.

2. Estuarine and Marine Assemblages

Macroalgae also dominate the flora of many shallow estuaries, lagoons, and upper embayments, coral reefs, and rocky intertidal/subtidal habitats (Fig. 3). Nutrient enrichment leads to reduction in the diversity of macroalgae and associated fauna. Opportunistic species of green algae within the genera Enteromorpha (also found in saltwater lakes such as the Great Salt Lake of Utah), Ulva and Cladophora, and within the brown algal genus Ecklonia have rapid growth rates and proliferate because they can more quickly take advantage of the ele
### Table I

**Harmful Estuarine and Coastal Marine Microalgae That Have Been Linked to Anthropogenic Nutrient Enrichment**

<table>
<thead>
<tr>
<th>Harmful species</th>
<th>Link to cultural eutrophication</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Chattonella antiqua</em></td>
<td>Bloomed under cumulative high loading of poorly treated sewage and other wastes, coinciding with human population growth (Japan; fish kills, toxic; Lam and Ho, 1989).</td>
</tr>
<tr>
<td><em>Chrysochromulina polylepis</em></td>
<td>Bloomed under cumulative high loading of poorly treated sewage and other wastes, coinciding with human population growth (Europe; fish kills, toxic; Kaas, et al., 1991; Maestrini and Granéli, 1991).</td>
</tr>
<tr>
<td><em>Gymnodinium microtum</em></td>
<td>Bloomed under cumulative high loading of poorly treated sewage and other wastes, coinciding with human population growth (Japan, as <em>G. nagasakiense</em>; fish kills, PSP; Lam and Ho, 1989).</td>
</tr>
<tr>
<td><em>Gonyaulax polygramma</em></td>
<td>Bloomed under cumulative high loading of poorly treated sewage and other wastes, coinciding with human population growth (Japan; fish kills from oxygen depletion; Lam and Ho, 1989).</td>
</tr>
<tr>
<td><em>Noctiluca scintillans</em></td>
<td>Bloomed under cumulative high loading of poorly treated sewage and other wastes, coinciding with human population growth (Japan; fish kills from oxygen depletion; Lam and Ho, 1989).</td>
</tr>
<tr>
<td><em>Nodularia spumigena</em></td>
<td>Bloomed following cumulative high loading of poorly treated sewage (Europe; fish—<em>Phaeocystis pouchetii</em> [Hariot] Lagerheim); blooms were correlated with altered N/P ratios from cumulative increased nutrient loading (<em>P. pouchetii</em>; Hallegraeff, 1993; Riegman et al., 1993); bloomed 1 week after a major swine effluent lagoon rupture into a eutrophic estuary (<em>Phaeocystis globosa</em> Scherffel, along with <em>Pfiesteria</em> spp.; southeastern United States; Burkholder et al., 1996).</td>
</tr>
<tr>
<td><em>Pfiesteria piscicida, P. shumwayae</em></td>
<td>Most kills (with highest cell densities) have occurred in P- and N-enriched estuaries (e.g., near phosphate mining, sewage inputs, or animal waste spills); between kill events can prey upon flagellated algae that are stimulated by inorganic nutrients; bloomed 1 week after a major swine effluent lagoon rupture (with extremely high phosphorus and ammonium) into an estuary, in a location where high abundance of these dinoflagellates had not been documented for at least the previous 1.5 yr, highly correlated with phytoplankton biomass in other eutrophic estuaries (mid-Atlantic and southeastern United States; fish kills, epizootics; Burkholder et al., 1995a; Glasgow et al., 1995; Burkholder and Glasgow, 1997).</td>
</tr>
<tr>
<td><em>Phaeocystis</em> spp.</td>
<td>Bloomed following cumulative high loading of poorly treated sewage (Europe, fish—<em>Phaeocystis pucheta</em> Harrold; blooms were correlated with altered N/P ratios from cumulative increased nutrient loading (<em>P. pucheta</em>); Hallegraeff, 1993; Bigman et al., 1993); bloomed 1 week after a major swine effluent lagoon rupture into a eutrophic estuary (<em>Phaeocystis globosa</em> Scherffel, along with <em>Pfiesteria</em> spp.; southeastern United States; Burkholder et al., 1996).</td>
</tr>
<tr>
<td><em>Prymnesium parvum</em></td>
<td>Toxic outbreaks usually have occurred under eutrophic conditions (fish kills, Edwardsen and Paasche, 1997).</td>
</tr>
<tr>
<td><em>Pseudo-nitzschia multiseries, other Pseudo-nitzschia</em> spp.</td>
<td>Have occurred with sewage and other wastes (Canada; ASP) (Smith et al., 1990); consistent seasonal blooms in the Mississippi and Ashknabula River plume areas, associated with hypernutrophication conditions (Q. Dortch et al., unpublished data) and in Prince Edward Island, Canada following anthropogenic nutrient loading and drought (Smith et al., 1990).</td>
</tr>
</tbody>
</table>

Note that many of the known harmful estuarine and marine microalgae and heterotrophic or animal-like dinoflagellates also have been shown to be stimulated by N or P enrichment in culture, expected since they are photosynthetic. Also note that blooms of the toxic *Pseudo-nitzschia* complex have not been associated with cultural eutrophication in the northwestern United States. Reprinted from Burkholder (1998), with permission.

\[ \text{Ammonium is more readily used (less energetically costly)} \text{ than nitrate, but either N form has significantly stimulated growth of opportunistic macroalgae in field research and mesocosm experiments. Phosphate can stimulate macroalgal growth when N is abundant.} \]

For example, in Boston Harbor, Massachusetts, sea lettuce (*Ulva*) formed dense populations for many years near outfalls of poorly treated sewage. The massive seaweed growth reduced water flow, smothered shellfish, and affected shoreline areas with an intolerable...
stench of hydrogen sulfide from decomposition of senescent/dead material. In Peel/Harvey Estuary, Aus-
tralia, dense mats of Cladophora, other green macroal-
gae, and cyanobacteria developed in response to P load-
ing. In seagrass meadows worldwide, such overgrowth has been documented to reduce light and promote de-
clines in various seagrass species. Along the intertidal area of the Baltic Sea, sewage inputs have been related to overgrowth of formerly dominant brown seaweeds (Fucus spp.) by opportunistic macroalgae.

In subtidal outfalls, nutrient enrichment from sewage and other wastes has stimulated overgrowth by red algae such as certain species of Polysiphonia. The exotic green macroalga Caulerpa taxifolia is a subtidal, sub-
tropical “weed” that has invaded colder waters of the Mediterranean, and sewage appears to further stimulate its robust growth. This organism produces metabolites that discourage predation; sea urchins typically avoid consuming C. taxifolia and starve to death if it is the only available food. Subtidal forests of giant kelp (Mac-
rocystis pyrifera) died back and failed to reproduce dur-
ing the 1960s and 1970s near outfalls of sewage dis-
charged from Los Angeles. The kelp bed loss apparently resulted from toxic substances in the poorly treated sewage.

Sensitive oligotrophic coastal marine coral reefs have been overgrown and smothered by macroalgae after nutrient input from sewage. In Kaneohe Bay, Hawaiis, for example, phytoplankton near sewage outfalls first removed the N and were transported to the central area of the coral reef where they decomposed and released the N for stimulation of the opportunistic green macroalga, Dictyosphaeria cavernosa. Inorganic nutrient concentrations are frequently below detection in natural coral reef waters. Lapointe’s elegant work indicated that even minor sewage-related increases in N (≥1 μM) and P (0.1–0.2 μM soluble reactive phosphate) off the coast of Jamaica, together with decreased herbivory from natural and fishing-related disturbance, were suffi-
cient to stimulate blooms of the green “eutrophic indi-
cator” macroalga, Chaetomorpha linum and cyanobacte-
ria that have overgrown coral reefs in that area.

Reduced light from such excessive nutrient-stimu-
lated macroalgal (and sometimes phytoplankton) growth affects coral growth by decreasing the produc-
tivity of zooxanthellae, the symbiotic photosynthetic dinoflagellates inside the coral tissue that provide much of the corals’ nutrition. Nutrient overenrichment also can shift species dominance within the coral commu-
nity. As the corals are outcompeted for space, filter-
feeding taxa such as sponges may be stimulated by increased phytoplankton food resources, and they can replace corals as dominant species. Sedimentation of decomposing phytoplankton, and other disturbances such as destructive techniques for fish harvesting, can exacerbate the impacts of cultural eutrophication on coral growth and survival.

C. Aquatic Macrophytes

1. Freshwater Communities

Floating, submersed, and emergent aquatic vascular plants, commonly called macrophytes, respond to nutrient enrichment in different ways because the compet-
titive forces that they encounter vary with plant growth habit. Rooted plants—including those with specialized floating leaves, submersed growth, or emergent growth—generally do not compete with phytoplankton and other algae for nutrients, because they obtain most nutrients (except carbon, taken from the water) from the nutrient-rich sediment, with leaf uptake of nutrients from the less enriched overlying water as a secondary nutrient source. In contrast, floating plants at the water surface, such as the Lemnaceae (duckweed family) and the exotic weed Eichhornia crassipes (water hyacinth), must compete with suspended algae for N and P, but rely upon the overlying air for their carbon. Floating submersed plants must compete with suspended algae for all major nutrients. They tend to have well-devel-
oped root systems to aid in nutrient acquisition, and in shallow waters with sparse open-water habitat, they can outcompete phytoplankton for light by forming dense surface populations that reduce or eliminate available light in the underlying water. However, sub-
mersed rooted aquatic vegetation (SAV) is usually elimi-
nated in highly eutrophic systems, primarily through light reduction by phytoplankton and other algal over-
growth, as described by Phillips and colleagues, Wetzel, and others (Fig. 2).

Like stream-inhabiting macroalgae, rooted macro-
phytes in river systems reach maximal abundance and species diversity in midorder segments. But, in contrast to lake-inhabiting macroalgae, submersed freshwater macrophyte communities attain highest spe-
cies richness and abundance in mesotrophic lakes, often with dominance by Potamogeton spp. Light plays a ma-
jor role in progressive SAV decline as lakes become increasingly eutrophic. Nutrient enrichment can stimu-
late macrophyte growth through luxury uptake and tight recycling, sometimes without an increase in phyto-
plankton. Over time, however, epiphytic algae take ad-
vantage of the greatly increased surface area for coloni-
zation afforded by the macrophytes, and they can both
2. Estuarine and Marine Macrophyte Communities

Brackish waters are colonized by rooted, mostly fresh-water species with moderate salt tolerance (as examples, certain Potamochiron spp., Valvularia americana, Zanichellia spp.). In such habitats, light is often the primary resource limiting growth. Species with broader salt tolerance such as Ruppia maritima also can be abundant. Only about 50 species of angiosperms, mostly close relatives of freshwater Potamogeton spp., have the salt tolerance needed to thrive in marine habitats. Nearly all of these “seagrasses” grow in muddy substrata of shallow coastal lagoons and quiet embayments. Estuarine and marine macrophytes tend to be highly sensitive to light reduction and, thus, susceptible to eutrophication-related turbidity from algal (phytoplankton, epiphyte, and macroalgae) overgrowth and sediment loading/resuspension. Such shading causes gradual dieback and loss of most SAV, promoting dramatic declines, in turn, in the diversity and abundance of many plant and animal species that depend on the habitat provided by these plants.

More sensitive SAV species are replaced by others that are tolerant of eutrophic conditions. For example, among subtropical seagrasses, field observations and limited experiments have indicated that turtle grass (Thalassia testudinum) is more sensitive to eutrophication than shoal grass (Halodule wrightii) and manatee grass (Syringodium filiforme). Other regions may not have additional seagrass species available to replace more sensitive species and, where such species are present, they typically offer less desirable fish nursery habitat than the former dominant. Thus, in seagrass meadows under nutrient over-enrichment, more oligotrophic seagrass species are replaced by less sensitive species when available. As eutrophication progresses the seagrasses are eliminated, and rapidly growing macroalgae or phytoplankton become the dominant flora (Fig. 3).

Although light reduction is considered the major mechanism for seagrass decline under cultural eutrophication, excessive nutrients can act independently of light to promote seagrass loss (Fig. 8). The dominant north temperate species, elgergrass (Zostera marina), apparently lacks a physiological mechanism to inhibit nitrate uptake through its leaves, as indicated by the research of Pregnall, Burkholder, and colleagues. Most plants take up nitrate during the day with energy from photosynthesis. In contrast, Z. marina takes up water-column nitrate day or night if it becomes available, as shown by Touchette and colleagues (Fig. 9). This species probably evolved in N-poor coastal waters, and sustained nitrate uptake under temporary enrichment may have developed as a once highly advantageous competitive strategy. However, as coastal waters have become more eutrophic from sewage, septic effluent...
FIGURE 8 The effects of water-column nitrate enrichment and light reduction on shoot production of the seagrass, *Zostera marina*. From author's outdoor mesocosm experiments, indicated as the percent decrease from shoot production of control plants that did not receive water-column nitrate additions or light reduction (except that plants in controls and treatments all received an additional 30% light reduction for 3 hr at 0900, 1200, and 1500 hr on a 3-day rotation using neutral density screens to simulate conditions during high tide). Treatments were imposed for 10 weeks during the fall growing season for *Z. marina*. Controls were maintained at ambient natural light (except during simulated high tide) and nitrate (50 μg NO₃⁻/L). Treatments included low N (at 50 μg NO₃⁻/L, added daily as a pulse of enrichment) and high N (at 100 μg NO₃⁻/L) at each of three imposed light levels as 30, 50, or 70% reduction of ambient surface light (*Io*, accomplished using neutral density shades, with additional shading at simulated high tide as noted). *Z. marina* in all treatments with water-column nitrate enrichment declined in shoot production relative to shoot production of control plants, and the nitrate inhibition effect was exacerbated by light reduction (means ± 1 standard error; *P* < 0.05, *n* = 3). These effects were not caused by algal overgrowth, which was maintained at low levels in controls and all treatments throughout the experiment.

leachate, and other anthropogenic sources, sustained uptake of water-column nitrate likely has become a disadvantage. Nitrate enrichment to the sediments, under control by an abundant microbial consortium, does not cause a similar effect and, instead, can be mildly stimulatory.

Nitrate uptake is a metabolically expensive process, requiring high cellular energy. Research by Touchette and colleagues indicated that sustained water-column nitrate uptake by *Z. marina* can promote severe internal carbon imbalances, apparently from the need to shunt C skeletons from photosynthesis for use in high amino acid synthesis to prevent internal accumulation of toxic products such as ammonia. The physiological mechanism of an internal “carbon drain” from sustained nitrate uptake has been documented (for algae) by Turpin and colleagues. A common trait of *Z. marina* shoots under excessive water-column nitrate enrichment is structurally weakened growing regions, perhaps analogous to the above-mentioned loss of stem strength that has been reported in certain freshwater emergent macrophytes under nitrate enrichment. Excessive N enrichment has also promoted seagrass attack by pathogens (for example, the slime mold Labyrinthula zostericola), hypothesized to occur because N and C are internally shunted to amino acid production rather than to production of alkyloids and other antimicrobial compounds.

Another seagrass that has been examined for the nitrate inhibition phenomenon, *Halodule wrightii*, and certain macroalgae (for example, *Ulva lactuca*) have shown depressed growth in response to nitrate enrichment, although at much higher N levels (ca. 1.4 mg NO₃⁻/L, pulsed daily for 4 to 5 weeks) than for *Z. marina* (50–110 μg NO₃⁻/L, pulsed daily for 3 to 8 weeks). *Ruppia maritima* is stimulated by high water-column nitrate but inhibited by elevated N, as ammonia;
and Z. marina has been experimentally inhibited by high ammonia levels as well. In mesocosm experiments, light reduction has been shown to exacerbate the inhibitory effects of water-column nitrate enrichment on shoot production in Z. marina. Warm temperatures also exacerbate water-column nitrate enrichment impacts on root growth of this seagrass, suggesting that warming trends in climate change may be expected to interact with eutrophication and adversely affect this beneficial habitat species.

D. Microfauna

1. Freshwater Communities

Whereas plants and mixotrophic algae respond directly to nutrient enrichment, animals are generally indirect recipients of eutrophication impacts. Nonetheless, all are significantly affected by nutrient control of the quality (including size, taste, and ease of filtering) and quantity of their algal, bacterial or detrital food. Freshwater zooplankton consist of three main groups: (i) protozoa (mainly ciliates and flagellates) and (ii) rotifers form the "microzooplankton" (maximum dimension ca. 45 μm to less than 70 μm); and (iii) crustaceans form the "macrozooplankton" (mainly cladocera, and calanoid and cyclopoid copepods, ca. 70 μm to less than 500 μm). Protozoa diversity and abundance increase with the amount of available organic matter that becomes directly or indirectly available through eutrophication until the fauna are restricted by low-oxygen conditions. Oligotrophic protozoan plankton can be dominated in number by small-bodied ciliates, whereas large ciliates generally contribute most of the biomass (especially members of the Oligotrichida such as Phagotrypa nasuta and Paramecium rictum). Mesotrophic and eutrophic lakes are more commonly dominated in both number and biomass by small-bodied species (e.g., member of the Scuticociliatida). Protozoans are relatively sparse in oligotrophic lakes in comparison to the species-rich and abundant fauna that develops in moderately eutrophic waters. Aggregates of bacteria and detritus form in the water column as organic materials decompose, and these microhabitats can support protozoa that rival shallow benthic communities in diversity. Subtropical lakes show similar trends, but have been reported to support higher ciliate abundance and diversity than shallow lakes at a given trophic status.

Heterotrophic flagellates in the water column are bacterivores, whereas larger ciliates and heliozoa commonly consume mixed populations of heterotrophic and autotrophic flagellates and other algae. In low-oxygen waters, heliozoan amoebae and ciliates such as Coleps and Euplotes retain endosymbiotic green algae (zoochlorophyceae) that photosynthesize and generate oxygen for their hosts. As nutrient enrichment increases, bottom waters and surface sediments become more organically enriched; but the physiological demands on protozoans and other microfauna such as certain rotifers are also greater. Survival generally depends on tolerance of low-oxygen and often co-occurring acidic conditions. In organically overloaded, hypoxic sediments, protozoa species diversity is very low relative to that in well-aerated sites. Bacteria and dissolved organic matter form the major food sources, and are consumed by species such as Paramecium, Chilomastus, and Astasia. Anoxic sediments become colonized by specialized protozoa such as pelobionts and diplomonads, and by certain ciliates and amoebae.

Rotifers include many algal herbivores as well as detritivores and a few carnivorous species. They generally attain highest species diversity and abundance, along with increasing small-celled phytoplankton, in moderately eutrophic lakes and lower rivers. These microzooplankton tend to be favored in abundance over macrozooplankton under increasing eutrophication. However, rotifer species diversity declines as eutrophication progresses to phytoplankton dominance by cyanobacteria, especially large mucilaginous colonial species that clog the filtering apparatus of the animals. Rotifers are adversely affected as well by the toxins from certain cyanobacteria that are seasonally characteristic of eutrophic lakes.

Among macrozooplankton, cyclopoid copepods (raptorial feeders) generally feed most efficiently on larger "particles" (for example, algae with biovolume ≤1,000 μm³), whereas calanoid copepods consume particles ≤100 μm³, and cladocerans eat small particles at ca. 10 μm³. Changes in biodiversity under cultural eutrophication depend, to a large extent, on differences among these species in efficiencies of feeding at certain particle (algal or detrital) concentrations and size ranges, and on the responses of life history stages to food limitation. Nutrient enrichment alters food particle size and abundance which, in turn, leads to competitive species displacement. Larger-bodied individuals are often more efficient filter feeders than smaller fauna. However, larger species may not be competitively superior because their juveniles tend to be more vulnerable to food limitation than adults of smaller species. Smaller species withstand food depletion by reducing metabolism, growth, and egg production. Moreover, their specific food ingestion rate (milligrams food ingested per milligram of zooplankton biomass) is usually higher than that of larger zooplankton.
Calanoid copepods (for example, *Eudiaptomus*) often dominate the microfauna of oligotrophic and mesotrophic lakes. These organisms feed efficiently on larger algal cells in those systems, with higher ingestion efficiencies at low food density. As algal prey densities increase under eutrophication, larger cladoceran species become abundant under low to moderate predation by fish, whereas high predation tends to select for smaller cladocerans such as *Bosmina* or *Ceriodaphnia* over *Daphnia*. The larger cladoceran species generally have higher population growth rates than calanoid copepods. Their population cycles during warmer months have been shown to be mainly controlled by the relative proportions of edible and inedible algae, temperature, and predation. Like rotifers in the presence of abundant food, cladocerans can reproduce parthenogenetically, thus allowing for rapid growth rates. They are more efficient filter feeders at moderate to high algal densities because they ingest more food for the same amount of energy expended, relative to ingestion under food-limited conditions. Macrozooplankton biomass increases under eutrophication in both temperate and tropical lakes, with increasing dominance by small-bodied species. High-efficiency bacterial feeders are selected for as bacterial abundances increase under accelerated eutrophication. However, abundant co-occurring cyanobacteria with copepods may clog the filtering apparatus of large-bodied microzooplankton. Moreover, some cyanobacteria are directly toxic to these fauna. In highly eutrophic lakes, the more selective feeding of calanoid copepods, and the seizing behavior of herbivorous cyclopoid copepods, may afford one or both groups competitive advantage. Smaller cladocerans, such as the littoral-zone chydrorid *Chydorus*, may increase under eutrophication, larger cladoceran species generally have higher population growth rates than calanoid copepods. Their population cycles during warmer months have been shown to be mainly controlled by the relative proportions of edible and inedible algae, temperature, and predation. Like rotifers in the presence of abundant food, cladocerans can reproduce parthenogenetically, thus allowing for rapid growth rates. They are more efficient filter feeders at moderate to high algal densities because they ingest more food for the same amount of energy expended, relative to ingestion under food-limited conditions. Macrozooplankton biomass increases under eutrophication in both temperate and tropical lakes, with increasing dominance by small-bodied species. High-efficiency bacterial feeders are selected for as bacterial abundances increase under accelerated eutrophication. However, abundant co-occurring cyanobacteria with copepods may clog the filtering apparatus of large-bodied microzooplankton. Moreover, some cyanobacteria are directly toxic to these fauna. In highly eutrophic lakes, the more selective feeding of calanoid copepods, and the seizing behavior of herbivorous cyclopoid copepods, may afford one or both groups competitive advantage. Smaller cladocerans, such as the littoral-zone chydrorid *Chydorus*, may become abundant during dense cyanobacteria blooms because of their high efficiency in filtering the extremely small, solitary bacteria and blue-green prokaryote particles associated with the blooms while avoiding filter apparatus clogging by large, mucilaginous cyanobacteria colonies. Eutrophication can cause other impacts on zooplankton species, apart from changes in available food. For example, as bottom waters become increasingly hypoxic, survival is depressed for zooplankton eggs that sink to the bottom as part of the life cycle. Thus, nutrient enrichment can impair zooplankton recruitment as an indirect impact.

2. Estuarine and Marine Communities

Impacts of anthropogenic nutrient loading have, understandably, been more difficult to generalize in the complex flow/water exchange environments of estuaries and from large-scale marine environments with high physical and biological variability. Analyses by Michell and others have indicated that nutrients generally enhance phytoplankton biomass and carnivores depress herbivore biomass, but that the couplings between trophic levels (for example, phytoplankton and zooplankton) are weaker than in more "closed" systems (lakes and coastal lagoons). Brackish and marine habitats differ from fresh waters in having fewer rotifers or cladocera, more extensive representation by protozoans (especially ciliates and foraminifers), and often-abundant planktonic nauplii (young life history stages) of sessile adult fauna ranging from mollusks and malacostracan crustaceans to vertebrates. Holoplankton spend their entire lives in the water column. Among these are microzooplankton including larval forms of certain macrozooplankton, as well as tintinnid and non loricate ciliated protozoans, heterotrophic flagellates, and amoebozooe. In benthic habitats, the “meiofauna” (similar in size to microzooplankton) include nematodes, harpacticoid copepods, many turbellarians, and several minor phyla with diverse feeding habits and lifestyles, mostly acting as consumers of other microbes as prey. Macrozooplankton include copepods, especially of the genus *Acartia*, other cyclopoid copepods and planktonic harpacticoids, noncopepod crustaceans (especially carideans and mysids), and chaetognaths (arrow worms). Meroplankton, which spend only part of their lives in the plankton as larval stages, may include immature forms of benthic invertebrates and tunicates; eggs, larvae, and juveniles of shrimp, crabs, and fish; and sexual stages of hydroidozoans and scyphozoan ctenoidea (jellyfishes). The microzooplankton are important phytoplankton herbivores among estuarine and marine zooplankton. They are strongly influenced by changing phytoplankton food quality and abundance under eutrophication, with trends that are somewhat analogous to those described for freshwater microzooplankton. Zooplankton species diversity is highest in moderately nutrient-enriched waters, but significantly declines as nutrient enrichment becomes more excessive, with accompanying shifts to dominance of unpalatable algal species and pronounced bottom-water oxygen deficits. For example, in the Peel-Harvey Estuarine System of Australia, large populations of calanoid copepods *Haliclymenia* and *Calanoides* grazed winter diatom blooms but were rapidly eliminated when the noxious (toxic) filamentous cyanobacterium, *Nodularia spumigena*, became abundant. Predation of carnivorous macrozooplankton by planktivorous fish such as menhaden (*Brevortia tyrannus*)...
Eutrophication and Oligotrophication

E. Invertebrate Macrofauna

The freshwater zoobenthos, or bottom macrofauna (>0.5 mm, or 500 μm), consists mostly of insect larvae (with terrestrial adult stages), crustacea, worms, and mollusks. Eutrophication in the littoral zone of lakes and rivers, with accompanying organic pollution, causes similar impacts on the invertebrate macrofauna of these systems. In early stages of eutrophication, oligochaetes, chironomids, gastropods, and sphaerids increase, and mayfly nymphs such as Hexagenia decrease. As eutrophication continues, the major change that coincides with progressive hypoxia/anoxia in overlying waters and surface sediments is the decline and then disappearance of additional oxygen-sensitive species such as stonefly (Diasca spp.) and certain other taxa of mayfly nymphs (e.g., Baetis, Rhiithrogena) as well as certain caddisfly larvae (e.g., Rhyacophila, Hydropsyche) and bivalve mollusks.

Eutrophication increases organic matter and bacterial decay and depresses oxygen concentrations while increasing settlement of organic detritus to benthic organisms such as certain flatworms (e.g., Polycelis) that use the detritus for food. In the profundal sediments underlying deeper waters of lakes, the macroinvertebrate biomass increases but is comprised of low oxygen-tolerant species such as certain chironomid larvae and oligochaete worms (e.g., the oligochaete Tubifex tubifex, which, in one study, survived, grew, and reproduced under continuous anoxia for 10 months). The species diversity of tubificid oligochaetes decreases with advanced eutrophication and organic enrichment, and associated oxygen deficits. However, if oxygen is periodically available, the rich food supply in combination with the lack of more oxygen-sensitive competitors allows robust growth. Declines in chironomid communities occur most rapidly in the change between nutrient-poor to moderately nutrient-enriched (mesotrophic) waters; and numbers of oligochaetes (relative to chironomids) increase as organic enrichment increases. In the intermediate zone between the littoral and profundal, the increased supply of "fresh" littoral detritus is consumed by large-bodied detritivores such as the bivalve mollusk Dreissena.

Estuarine and marine coastal invertebrate macrofauna, including mollusks, polychaetes, decapods, and other crustacea, and nemertean, spend their adult lives buried beneath the sediment surface. They are highly diverse with food acquisition as filter feeders, nonselective deposit feeders, selective deposit feeders, or raptorial/other predators, making generalizations difficult. Increasing organic matter from accompanying nutrient enrichment tends to cause similar impacts as in freshwater systems, namely, an increase in macroinvertebrate abundance under moderately eutrophic conditions, and a decline in species diversity and abundance as eutrophication and associated hypoxia/anoxia progress. Organisms that burrow into anaerobic sediments must be able to gain access to oxygen in the overlying water. Increasing eutrophication leads to elimination of burrowing organisms as the anaerobic zone moves closer to the sediment surface. In highly overenriched areas, only worms such as Capitella may survive. Oxygen does not need to be completely absent for damage to occur—hypoxic waters with 3.0 to 4.3 mg DO/L have been related to mortality of some benthic invertebrate species and to loss of habitat for shellfish species such as lobster that require higher oxygen availability.

Changes in food quality/quantity also can reduce the species diversity and abundance of estuarine and coastal marine invertebrate macrofauna under highly eutrophic conditions. For example, an extensive early study by Filice in San Francisco Bay demonstrated that in domestic sewage outfall areas, few species survived. Species diversity was higher but still depressed in surrounding areas that received dilute sewage. However, in those areas, some species (e.g., the clams Gemma gemma, Mya arenaria, and Macoma balthica) apparently took advantage of new energy and material resources, and became highly abundant relative to abundance in control areas without sewage influence. In the Great South Bay complex of
Long Island, New York, previously mentioned duck farms along the bay tributaries fertilized the water with nutrients and organic wastes, and stimulated dense blooms of the small algae, *Nannochloris* sp. and *Stichococcus* sp. (densities greater than $10^6$ cells/ml). These algae were very different from the previous phytoplankton community in the area, which consisted of mixed species that are needed to support oysters. Following these changes, oyster populations significantly declined, apparently because they were unable to thrive on a diet consisting only of these small algae.

**F. Vertebrate Macrofauna**

Eutrophication initially reduces and then eliminates sensitive lake fishes (e.g., salmonids and coregonids), by eliminating oxygen-replete, colder bottom-water habitat and well-oxygenated spawning areas. Fish can avoid low-oxygen waters, but the cold-optimal species encounter warmer waters as they are forced to move from deep areas into the shallows. Most temperate-zone fishes breed in the littoral. Their eggs are more vulnerable to short-term (e.g., nightly) oxygen deficits and to the low-oxygen microenvironment of increased detritus. Thus, the critical habitat of these sensitive species is destroyed by nutrient overenrichment, and spawning and recruitment are depressed.

Analogous impacts occur in rivers and estuaries. For example, in parts of the Baltic Sea, cod eggs laid in well-oxygenated surface waters die when they sink to anoxic bottom waters. Oxygen levels in the bottom waters of the Baltic’s deep basins are negatively correlated with juvenile codfish abundance. Hypoxia in estuaries has been linked to depressed survival of larval fish, mortality of certain benthic invertebrates used as fish prey, and loss of habitat for mobile species of fish such as cod that require high oxygen availability. Hypoxia and anoxia represent a growing problem for many estuaries and coastal marine waters, such as the Chesapeake Bay, the Baltic Sea, the Black Sea, the Pamlico Estuary, Long Island Sound, the North Sea, and the Gulf of Mexico by the mouth of the Mississippi River.

Overall population sizes and biomass of fish usually increase with nutrient overenrichment, which is sometimes regarded as a beneficial effect in early to midstages of eutrophication. However, dominance shifts from species such as lake trout to cyprinids, bullheads (*Ictalurus* spp.), and other coarse fish that can tolerate low oxygen concentrations (Fig. 10). In lakes with well-developed littoral zones of rooted aquatic vegetation, high predation pressure from piscivorous fish such as largemouth bass (*Micropterus salmoides*) or pike (e.g., *Esox lucius*) maintains low biomass of cyprinid fish and, thus, low predation pressure on large-bodied zooplankton. These large zooplankton, in turn, control phytoplankton biomass through strong grazing pressure, so that there is high visibility for visually feeding piscivorous fish and abundant light for macrophyte growth.

As eutrophication progresses, the littoral-zone macrophytes disappear and the carrying capacity of the lake for piscivorous fish is reduced. Thus, (200-) planktivorous fish are freed from high predation. Their larvae tend to selectively feed on the largest herbivorous zooplankton species (based on the size-efficiency hypothesis of Brooks and Dodson), causing an overall reduction in herbivore size. Smaller herbivores cannot exert enough grazing pressure to control blooms of cyanobacteria and other noxious algae, which proliferate in eutrophic lakes and lower rivers in the absence of the large-bodied zooplankton. Turbidity increases from higher cyprinid feeding activities and from loss of the littoral-zone "filtering" effect of the macrophytes, further impairing visual predation by remaining piscivorous fish. Photosynthetic activity from the high algal
biorichness can elevate the pH to levels that adversely affect gill function in sensitive fish species.

Similar impacts on fish populations from loss of desirable littoral zone species have been documented in estuarine and coastal waters. For example, along the Swedish coast of the Baltic Sea, excessive nutrient enrichment has been related to increases in nuisance cyanobacteria and other filamentous species, as mentioned, and a decrease in formerly dominant Fucus vesiculosus (Phaeophyceae), probably because of reduced light availability. This seaweed had provided critical habitat for herring spawns and the shift to dominance by filamentous macroalgae led to decreased egg hatch in the herring populations.

Although aquatic birds and other vertebrates can move among aquatic environments with relative independence, some general effects of eutrophication on these macrofauna also have been described. Increased macrophyte vegetation under moderate nutrient enrichment in mesotrophic systems has been linked to increased numbers of herbivorous waterfowl such as moorhen (Gallinula chloropus) and whooper swan (Cygnus cygnus). Some piscivorous birds such as grebes and herons have also increased. As eutrophication progresses, however, bird species may decline following undesirable changes in food and habitat. For example, pochards (Aythya ferina) decreased on Lake Constance as their main food supply, the macroalgae Chara, declined under nutrient enrichment. General macrophyte habitat loss under nutrient overenrichment has also been linked to the gradual disappearance of many waterfowl such as black and mute swans (Cygnus atratus, C. olor strepera), Canada geese (Branta canadensis), coots (Fulica atra), teal (Anas crecca), and gadwells (Anas strepera). Sensitive life history stages of amphibians (for example, frog eggs) have been killed in increasing hypoxic events within littoral zones of eutrophic lakes.

III. OLTROPHICATION—REVERSING THE IMPACTS OF EUTROPHICATION

It is possible to reverse nutrient loading impacts, at least to some extent, by reducing the nutrient inputs. This phenomenon can occur naturally when, for example, a major storm such as a hurricane causes high flooding or sedimentation that effectively re-routes a nutrient-rich tributary away from a receiving lake or estuary (such as an oxbow lake that is cut off from the main river); when natural flooding destroys a dam and eliminates a run-of-river impoundment; or when a raised bog system becomes isolated from nutrient-rich stream and groundwater sources. As nutrient loading declines, species diversity generally increases while the system production decreases. In embrotophic bog systems, there is also a shift over time to more acid-tolerant species, as the system's sole source of new water is low-pH precipitation.

Human influences are more pervasive than the natural forces that can contribute to oligotrophication. Cultural eutrophication is an acceleration of a long-term, natural process. Resilience, defined as the rate of recovery to the predisturbance state, depends on the initial status of the system in the natural eutrophication process and on the degree of nutrient loading sustained. Ecosystem resilience is generally considered to increase with increased nutrient loading rate and to decrease with increasing food chain length. Recent analysis by Cottingham and Carpenter suggested a modification of this relationship for pelagic food webs of north temperate lakes. In that analysis, planktivore-dominated (short-length) food webs were more resilient at baseline P loading rates of 0.1 to 1.0 μg/L/day. Piscivore-dominated (long-length) food webs were more resilient at high baseline P loading rates (2.0 μg/L/day), apparently because the additional nutrients were incorporated into the biota more rapidly.

Eutrophic systems, dominated by "generalist or opportunistic" species that are insensitive to the adverse impacts from nutrient overenrichment, are considered to be relatively resistant to further change or stress as mentioned. Moreover, the adverse impacts of cultural eutrophication in such systems can be partially reversed in a relatively short period of time (several years). Less nutrient enriched systems (for example, early phases of moderate nutrient enrichment or mesotrophy) have more specialized species, more opportunity for biotic adjustment of elemental cycles, and tighter coupling among element cycles. The diverse communities of such systems, comprised of more sensitive species, are relatively sensitive to further stress imposed by nutrient enrichment under accelerated eutrophication; and reversal or "recovery" to a "pristine" oligotrophic state is difficult to accomplish.

The degree of "restoration" success also depends on certain physical features of the system, especially mean depth and flushing rate. Aquatic ecosystems tend to retain most nutrients from loading events in the bottom sediments. In deep lakes, most of the nutrient inputs
from the previous years are unavailable to phytoplankton of the next growing season because the lower water column remains permanently stratified and, thus, isolated from the upper water column where most of the viable phytoplankton occur. In shallow systems, wind and internal currents mix the water column and superficial sediments completely and frequently over an annual cycle, so that some portion of the nutrients from the relatively rich bottom waters and sediments repeatedly becomes available for phytoplankton growth. Reservoirs, rivers, well-flushed estuaries and wave-swept marine coastal waters are usually less sensitive to nutrient loading and more easily ‘reversed’ in the eutrophication process than lakes or poorly flushed coastal lagoons with long water exchange times (on the order of months to years) because the latter systems cannot flush the nutrient-laden water through and ‘self-cleanse.’

As other, more practical considerations, the extent to which the acceleration of the eutrophication process can be reversed would be expected to depend on the feature of the aquatic ecosystem that is targeted for improvement. Nutrient reductions would exert the most direct effects on plant communities such as freshwater/estuarine phytoplankton or estuarine/coastal macroalgae. Thus, a goal of reduced incidence of algal blooms may be more rapidly achieved than that of improved fish communities with growth that is indirectly rather than directly affected by the nutrient inputs. The economic feasibility of controls that can be exerted on anthropogenic nutrient sources is also important. Pragmatically, the highest prognosis for success would be expected for natural waters that are affected mostly by sewage and other point (pipe) source nutrient dischargers, because they are much easier to control than nonpoint (diffuse) sources.

One of the most famous early reports of successful reversal of cultural eutrophication involved removing sewage discharges from Lake Washington within metropolitan Seattle, Washington, in the United States. This large, deep lake (128 km², maximum depth 59 m, mean depth 18 m) historically had shown water quality degradation in response to sewage inputs. In 1922 a diversion was created to carry the raw sewage from 30 outfalls away from the lake (into nearby Puget Sound). Algal blooms and fish kills soon abated. However, in 1930 sewage effluents began to be discharged into the lake from treatment plants in outlying communities, and noxious cyanobacteria blooms and hypoxia again increased. About 76 million liters of sewage without inorganic nutrient removal were discharged daily into the lake by 1962. Nearly a decade of effort by a courageous limnologist, Dr. T. Edmonson, led the city to support zero discharge of sewage into the lake by 1968. By 1970 the lake’s phytoplankton growth had decreased to levels that had not been seen since the early 1950s, and the improvements in water quality and aesthetics were heralded by the citizenry as a ‘rapid and remarkable’ recovery. This recovery was aided by the fact that the lake is relatively deep and that the major nutrient sources to be controlled were sewage pipes. It is unlikely that the degree of recovery reversed the water quality and aquatic communities to a circa 1930, more pristine status. However, the partial recovery (partial oligotrophication) represented a major improvement.

Other partial reversals of cultural eutrophication have been reported worldwide. In freshwaters they mostly have been achieved by targeting P reductions. In estuaries, both P and N have been reduced or ‘co-managed’ for best results. Studies in north temperate estuaries by Fisher, Chesterikoff, and colleagues indicate that if only P, but not N, is removed from sewage inputs to upstream fresh waters, the P removal can actually exacerbate N-related eutrophication problems in downstream estuaries. The freshwater P reductions decrease riverine algal growth that, if present, would also have removed a substantial portion of the N, in the sewage before it reached the estuary. Soluble nitrate, in particular, is transported downstream at higher concentrations than if freshwater algal blooms had been available to consume it, and the increased N is thus available to stimulate higher phytoplankton growth in the receiving estuary.

In large lakes, partial reversals can achieve highly desirable results. For example, the degradation of fish communities in Lake Erie reached its most extreme level in the 1960s from a combination of eutrophication, overexploitation of fishery resources, extensive habitat modification, and other pollution. Beginning in the 1970s, fishery management strategies and pollution abatement programs contributed to a dramatic reversal. Lake Erie walleye fisheries rebounded to world-class status, and point-source P loading significantly declined, especially after sewage treatment was improved at the major point source discharger to the west basin of the Lake (the Detroit metropolitan wastewater treatment plant) and after mandated use of detergents without phosphorus. The P reductions effected a dramatic decrease in the abundance of nuisance phytoplankton species and of zooplankton biomass, as well as a decline in the abundance of pollution-tolerant zooplankton and an overall shift in macroinvertebrates to more pollution-intolerant taxa. Similarly, in the Bay of Quinte on Lake Ontario, P loading reductions after 1977 led to a decline...
in the abundance and biomass of oligochaete worms, sphaeroid mollusks, isoped crustacea, and some chironomids. Dominance in both chironomid and oligochaete conditions shifted to species less tolerant of eutrophic conditions.

Nutrient reductions have also led to success stories in estuaries and coastal lagoons. For example, industrial point source nutrient reductions to the Seto Inland Sea in Japan during the 1980s promoted a noticeable decline in the frequency and magnitude of toxic dinoflagellate blooms. Nutrient loadings to Cockburn Sound, Australia, from industrial point sources, and reduction in nutrient loadings to Chesapeake Bay from sewage point sources, decreased phytoplankton blooms and promoted an increase in seagrass meadow habitat. Sewage discharges to the shallow estuary, Mumford Cove, in Connecticut (United States) were rerouted to another waterway in the late 1980s, and within two years massive nuisance blooms of the macroalgae, Ulva lactuca, from industrial point sources, decreased phytoplankton blooms and promoted an increase in seagrass meadow habitat. Sewage discharges to the shallow estuary, Mumford Cove, in Connecticut (United States) were rerouted to another waterway in the late 1980s, and within two years massive nuisance blooms of the macroalgae, Ulva lactuca, were eliminated. Thus, partial "oligotrophication" from removal of sewage can have fairly rapid, positive results even in shallow systems where the sediments could provide some degree of nutrient replenishment to the overlying water.

From an overall ecosystem standpoint, the extent to which aquatic habitats can be "fully reversed" in the cultural eutrophication process depends on a more complex "endpoint" than variables such as reductions in nuisance algal growth, recovery of beneficial macrophyte beds, or increases in desirable fish species. The full range of chronic impacts from variables that frequently accompany nutrient enrichment are poorly understood. For example, scientists only recently determined that low-level water-column nitrate enrichment could inhibit the growth of certain sensitive seagrasses as a direct, subtle, and potentially serious physiological effect. Scientists only recently reported, as well, that nutrients can act to indirectly stimulate toxic, animal-like organisms such as Pfiesteria, mediated through stimulation of algal prey that are consumed by Pfiesteria species. The National Research Council has reported that there are more than 100 different enteric pathogens in sewage including viruses, bacteria, and parasites that cause waterborne diseases in humans. The fate of many of these organisms and their impacts on aquatic life have not been examined, but available evidence indicates that some of them can survive for months to years in the bottom sediments of lakes, rivers, and estuaries. Thus, many challenges of cultural eutrophication remain to be resolved before the full extent of its impacts on aquatic ecosystems, and the success of reversing those impacts, can be evaluated.

See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • ESTUARINE ECOSYSTEMS • LAKE AND POND ECOSYSTEMS • PLANKTON, STATUS AND ROLE OF • RIVER ECOSYSTEMS • SEAGRASSES

Bibliography


* Asterisk indicates comprehensive secondary references on aspects of this topic; other references were cited in the text, as well.
EUTROPHICATION AND OLIGOTROPHICATION


I. A Brief History of Evolutionary Thought
II. The Universal “Tree” of Life
III. Structure of Evolutionary Biology and How It Is Studied

GLOSSARY

**evolution** Descent, with modification from one (or at most, a few) original ancestors
**evolutionary theory** A body of statements about the general laws, principles, or causes of evolution
**lateral transfer** Genetic information passed between organisms through means other than inter-breeding
**phylogenetic tree** A hypothesis for describing the history and relationships among living species.

**EVOLUTION, DESCENT WITH MODIFICATION**, is a scientific fact. Evolutionary theory is a coherent body of interconnected statements, based on reasoning and evidence, that describes the processes of branching of lineages and changes within lineages (including extinction and gene transfer). Nothing in biology makes sense unless it is studied in an evolutionary context (Dobzhansky, 1973), and all fields of biology provide insight into the processes of evolution and mechanisms of evolutionary change. Therefore, the goal of evolutionary biology is to determine patterns of ancestor-descendant relationships among organisms in time and space, understand the processes that gave rise to them, and discern how the specific attributes of organisms originated and changed through time.

I. A BRIEF HISTORY OF EVOLUTIONARY THOUGHT

The development of evolutionary theory to date can be characterized by four major periods of expansion that reflect increasingly sophisticated means of documenting and quantifying natural variation at the molecular, cellular, and organismal level. The most significant development in evolutionary theory occurred between 1859 and 1930, the Darwinian period. The most significant insight of this time was recognition of the common ancestry of organisms and the force of natural selection. The next period of expansion in evolutionary thought occurred between 1930 and 1960 when Darwinian theory was reconciled with modern genetics, systematics, and paleontology. The current period of expansion in evolutionary biology, 1960 to the present, began when Lewontin and Hubby discovered the enormous amount of natural variation at the protein and nucleotide level. Currently, the molecular revolution has refocused evolutionary studies on decoding the genetic ciphers that explain evolution. The most fundamental concept in evolutionary biology, the concept
that all organisms can be divided into discrete species, is being challenged as are long-held notions about adaptation and the relationships among species. The current tree of life divides organisms into three kingdoms, the Eukarya, the Archaea, and the Bacteria.

This chapter is organized into two sections. The first section outlines the periods of expansion of evolutionary thought; the second section outlines the structure of evolutionary biology, explains how evolution is studied, and discusses recent empirical works that highlight new findings.

A. Darwinian Evolution: A Paradigm Shift in Evolutionary Thought

The word evolution is derived from the Latin word evolutio, or "unrolling." Prior to Darwin, classical thought argued that God created all species according to a chain of being or Scala Naturae. The chain of being was an expression of God's design, all objects were perfect, permanent, and unchanging; no new forms of life could have arisen or become extinct since the time of Creation. Therefore, the role of natural science was to catalog God's creations to make manifest His wisdom. Based on this view, Linnaeus established a universal framework for classifying similar species into similar genera that were thought to reflect God's design but that did not imply any genealogical connectedness. These ideas culminated in the most significant pre-Darwinian theory of evolution, put forth by Jean Baptiste Lamarck (1744–1829). Lamarck proposed that organisms evolved through adaptation to the physical environment. He introduced the theory of "organic progression" in which species originated continually via independent events of spontaneous generation and subsequently evolved "up" the scale of nature (Fig. 1a). Lamarck envisioned two hierarchical chains of life, one for plants and one for animals, and a world where extinction was not possible. Lamarck viewed changes in organisms to result from the use and disuse of organs and that these acquired characteristics were then passed to descendants. Extinction was not possible and, therefore, species identified from fossils were still present, although modified, through adaptation.

Charles Darwin (1809–1882) founded the current field of evolution. His most important contribution to the theory of evolution was the concept that all species diverged from a common ancestor and that natural selection is the mechanism driving speciation. Prior to Darwin, natural scientists accepted the idea that species struggle against one another for existence. Darwin's experience as a naturalist and the writings of Malthus, "Essay of Population Growth" (1798), helped him see that individuals, not species, struggle for life. Malthus's essay argued that the rate of human population growth was greater than the rate of the increase of food. Therefore, if reproduction is untethered, famine would drive animal populations to extinction. The realization that individuals with superior characteristics would live to reproduce, while individuals of the same species with inferior characteristics would not, triggered Darwin's belief in evolution. The struggle for existence, that Darwin believed resulted in the differential reproduction of individuals, is defined as natural selection (Fig. 1b).

B. The Synthetic Theory of Evolution

By the 1870s most scientists accepted the theory of evolution by common descent but, for at least 60 years after the publication of The Origin of Species, there was no consensus that natural selection was the mechanism by which evolution was achieved. The next period of expansion in evolutionary thought occurred in the 1930s and 1940s, when contributions of geneticists, systematists, and paleontologists reconciled Darwinian theory with the facts of genetics (Mayr and Provine, 1980). A modern understanding of genetics, together with the mathematical theory of population genetics by R. A. Fisher (1890–1962), J. B. S. Haldane (1892–1964), and Sewall Wright (1889–1988), demonstrated that mutation and natural selection were causes of adaptive evolution. The synthesis of studies in taxonomy and genetic principles built by E. Mayr (b. 1904) and G. L. Stebbins (b. 1906) provided evidence that evolution occurred gradually. The integration of paleontology with population genetics by G. G. Simpson (1902–1984) demonstrated that the synthetic theory of evolution was fully consistent with the fossil record. Alternative theories of the mechanisms of evolutionary change that had been formulated since Darwin but prior to the evolutionary synthesis, such as Neolamarckism and the view that biological systems were creative and purposeful, were demonstrated to be inconsistent with the fossil record. Therefore, the major achievement of the evolutionary synthesis was to fully integrate genetics and Darwinian evolution and to argue that the major features of evolution could be accounted for by within species processes, such as mutation, recombination, and natural selection (Futuyma, 1998).
FIGURE 1 Proposed trees of ancestor-descendant relationships visualized by Lamark (A) and Darwin (B). (A) Jean Baptiste Lamark proposed that organisms originated spontaneously, continuously and evolved "up" a chain of natural being. (B) Darwin proposed that all species evolved from a common ancestor through the mechanism of natural selection (from Futuyma, 1998).

C. The Discovery of Protein and Nucleotide Diversity

Two schools of thought emerged during the synthetic period: the selectionist school from genetics and the mathematical theory of population genetics, and the adaptionist school from paleontology and systematics. The selectionist view is that there is no limit to an organism's variability and hence ability to evolve. Furthermore, all aspects and characteristics of organisms are variable and able to change rapidly. No qualitatively new phenomena results from increased organismal complexity and adaptations reflect the sum of infinitely variable genotypes. In contrast, the adaptionist school put forth the concept of "Bauplan," or the view that groups of organisms represent broad schemes of organization and that variations among groups reflect some degree of functional and developmental constraint. In its most extreme, the adaptionist school views organisms as relatively static and only able to change over long periods of time after populations of species have been isolated.

In the early 1960s, evolutionary population genetics became the central discipline of the study evolutionary processes. The most important contribution made during that period was the discovery of the vast amount of unexplained nucleotide and protein diversity and the subsequent explosion of mathematical theory to explain the role of mutation, drift and natural selection in the evolution of populations. The resulting, deepened understanding of the molecular basis of life and genetics revealed those questions that needed to be answered to reconcile the theory of evolutionary population genetics with concepts of speciation that are based on shared patterns of organismal organization and design.

D. The Molecular Revolution and Current Evolutionary Thought

The foundation of evolution, descent with modification from a common ancestor, as stated by Darwin, remains unchanged. The molecular revolution, however, continues to transform our perspective on evolutionary thought by providing molecular tools to dissect the pathways by which characters and organisms are modified and molecular tools to probe the mechanisms by which molecules, genes, and genomic architecture interact. The unfolding complexity of these genetic pathways continues to reveal an increasing number of insights that are both exciting as well as disturbing to some. One of these is the deep genetic similarity between what are otherwise very different organisms, for example, houseflies and humans. We now know that a substantial part of genetic and biological innovation is the result of the simple mechanism of gene duplication in which genes are subsequently coopted for new functions. Furthermore, the chimeric aspect of genomes shows that organisms from the archaean lineages (the methanogens or methane-producing bacteria and bacteria that live in extreme environments), bacterial lineages (all remaining single cell organisms), and eukaryote lineages (all multicellular organisms) all contain genes
from multiple sources that are likely to have been obtained by the process of lateral transfer, not interbreeding (vertical transfer) or even de novo (Fig. 2). In fact, the lateral transfer of genetic material is so extensive that some investigators no longer consider the Archeae and Eubacteria discrete or closed systems of organization but recognize them as fluid and reisulated biological systems (Doolittle, 1999).

**FIGURE 2** (a) The three kingdoms of living organisms. The origin of the Eukaryotes, all multicellular organisms, is thought to lie somewhere between the Archeabacteria (methane-producing bacteria and bacteria living in extreme environments) and the Eubacteria (all remaining single-celled organisms). Lateral transfer (b) of genetic information among all three lineages may be extensive and its implications a matter of active debate (from BioEssays 21.2, W. Martin, © 1999. Reproduced by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.).
One effect of this more pluralistic view of organisms and the complexity of genetic systems revealed through the molecular revolution has been to refocus evolutionary thinking from the inheritance of fixed characters to the inheritance of systems of organization and control. Evidence for this perspective comes from integrative studies of the fossil record and molecular developmental biology. For example, current evidence from paleontology shows that by the end of the Ediacaran, 520 Ma, all of the major clades of animals, or “stem” groups of animals, had diverged, and since that time, only one of group of organisms, the Vendobionta (preserved as impressions, casts and molds; the status of this group is controversial), has gone extinct (Fig. 3): Divergence of the Bilateria (organisms with bilateral symmetry, i.e., the Chordata and Echinodermata; the Mollusca, Annelida, Platyhelminthes, and Brachiopoda; the Arthropoda, Nematoda, and Priapula) from the sponges, cnidarians, and ctenophores resulted in all of the animal taxa recognized today. Molecular evidence shows that early-diverging Bilateria taxa display both ancestral and derived characters that include similar gene content, development, and morphology (Knoll and Carroll, 1999). All bilaterians are characterized by a modern body plan (rostral-caudal body axis) and an extensive cluster of homeobox genes (Hox genes encode a sequence of 60 amino acids that bind to DNA and control its expression) that control interacting networks of developmental regulators and key structural genes (Knoll and Carroll, 1999). The different degrees and types of organization that characterize the modern descendants of these groups are the result of a mosaic of genetic and morphological phenotypes whose components evolve at different rates.

Understanding the developmental basis of organismal variation provides insight into the potential for evolutionary change, as well as its pattern. For example, the two major groups of bilaterians—the Protosomes (illustrated primarily in Drosophila melanogaster) and the Deuterostomes (illustrated primarily by vertebrates)—share a common genetic regulatory repertoire. This means that the common ancestor of the arthropods and chordates would have had all the genes that they share, as well as all of the morphological characters that the shared genes regulate: photoreception organs, appendages, a heart, and the propensity for body seg-

---

**FIGURE 3** Animal diversity across the Phanerozoic–Cambrian transition. By 520 Ma, all major clades of animals had diverged. All bilaterians are characterized by a rostral-caudal body plan and extensive cluster of homeobox genes. The descendants of these groups are mosaics of genetic and morphological phenotypes whose components evolve at different rates (Reprinted with permission from Knoll and Carroll, Early Animal Evolution © 1999, American Association for the Advancement of Science).
EVOLUTION, THEORY OF

mentation (Knoll and Carroll, 1999). Through the replication and multiplication of regulatory networks, such as the Hox gene complex, taxa diverged as the complexity and the hierarchy of gene interactions increased. All of the proteins that cause the differentiation of body organization, however, did not and hence were present over 530 million years ago (Marin and Baker, 1998). Therefore, periods of great diversification in evolution, such as the Cambrian explosion, record the differentiation of crown (recently evolved) groups within clades that diverged as a result of an earlier radiation of genetic regulatory systems among stem (ancestral) taxa (Knoll and Carroll, 1999).

II. THE UNIVERSAL “TREE” OF LIFE

The current phylogenetic tree, based on small subunit RNA as well as whole genome-based phylogenetic analysis, distributes species among three groups of primary relatedness or domains: the Bacteria, the Eukarya, and the Archaea. The root of the Eukarya is between the Bacteria and Archaea (Fitz-Gibbon and House, 1999). Exploring the relationships among genes, regulatory processes, and the species that are contained in them has led to the concept of a phylogeny (the evolutionary history of an organism) as a diffuse cloud of gene histories that may make it difficult to organize organisms into Linnaean systems. Therefore, one current perspective of phylogenies or the histories of lineages are as models of the change in the interbreeding probabilities between organisms and models of genetic potentialities of organisms through time (Maddison, 1997).

III. STRUCTURE OF EVOLUTIONARY BIOLOGY AND HOW IT IS STUDIED

The are two central questions in evolutionary biology: The first question asks, what is the history of life, as defined by the history of the species, their origins, and extinction, and what are the steps through organisational features evolved? The second question asks, what are the causes of the history life? While the major principles of evolutionary biology and major causes have been established, new questions are continually posed and many long-standing questions remain to be fully answered. The processes of evolution are both random and deterministic: Studies in nine subdisciplines, as well as syntheses among these, contribute to understanding evolution and its mechanisms: evolutionary ecology, behavioral evolution, evolutionary paleontology, evolutionary developmental biology, evolutionary physiology and morphology, evolutionary systematic biology, evolutionary genetics, molecular evolution, and human evolution. Research in all of these fields has accelerated in the past 20 years as the focus of evolutionary studies continues to shift.

A. Evolutionary Ecology

The most important questions in evolutionary ecology address the role the environment plays in speciation and extinction. During the synthetic period of evolution, the dominant view of speciation was that geographic barriers would develop between two populations prohibiting gene flow between them. Even though the populations might live in identical environments, they would diverge gradually as a result of the accumulation of random mutations. In contrast, however, some researchers have proposed that the barriers spawning species can also be ecological. In this case, selection, as opposed to isolation, is the factor that drives speciation. This implies that speciation is the product of an event, not a process. Both modes have received support in recent studies.

Peterson et al. defined and compared the ecological niches of birds (21 pairs), mammals (11 pairs), and butterflies (5 pairs) in an area of active speciation and population differentiation, the Isthmus of Tehuantepec. They tested the degree to which the ecological characteristics of one taxon were able to predict the geographic distribution of its putative sister taxon and vice versa. They then computed a genetic algorithm to produce a set of decision rules in ecological space (a model of the fundamental niche) that were projected onto maps to predict potential geographic distributions of the taxa. Across the 37 species pairs they found that a taxon on one side of the Isthmus could more accurately predict the distribution of its sister taxon on the opposite side of the Isthmus of Tehuantepec than the distribution of any of the other species. They concluded that because the species on each side of the Isthmus had been isolated for several million years, the sister species that evolved on each side of the barrier were the result of random genetic events but not natural selection (Peterson, Soborin, and Sanchez-Cordero, 1999).

An alternative view of the speciation process is that ecological variation, such as climate variation or differences in food resources can act as barriers, and can select for adaptation. This may result in individuals that are morphologically different and are unwilling to mate with each other even though they have never been
physically separated. The most famous example of adaptive radiation and how ecological niche can drive the evolution within a population is seen in studies of Darwin’s finches. While journeying on the Beagle, Darwin collected 14 different species of finches from the Galápagos Islands. The different species of finches showed a degree of variation in beak size, beak shape, and body size that usually characterizes differences among families of birds. Nevertheless, the entire radiation of the finches is believed to have occurred in less than 3 million years. Using microsatellites (multilocus genetic markers with high mutation rates), Petren, Grant, and Grant (1999) found that 13 of the species of Darwin’s finch that they studied were monophyletic (derived from the same common ancestor). Their many differences in phenotypic traits, including beak size and shape, body size and plumage, are derived from an ancestor with a relatively long pointed beak that is associated with an insectivorous diet. From this ancestor, ground finches evolved blunted beaks that are efficient for crushing seeds while tree finches evolved beaks that allowed greater biting strength at the tip. Furthermore, once a novel beak evolved, body size and beak size changed rapidly and allometrically (Petren et al., 1999). They concluded that divergence among the finches reflects strong selection for ecological type.

B. Behavioral Evolution

Behavior is studied at two levels: (a) the neural, hormonal, and developmental mechanisms that underlie the adaptive differences and the historical pathways leading to a current behavior and (b) the selective processes shaping behavioral evolution. When put in an evolutionary context, the importance of a behavior is measured through its effects on animal reproductive success. One particularly clear example is when the status or dominance rank of an animal that lives in a group or community correlates with the animal’s reproductive success. For example, a recent analysis of 30 years of data gathered on the chimpanzee community at the Gombe Stream Research Center, Tanzania, has shown that female chimpanzees, who are solitary foragers, do not display an obvious linear hierarchy. Female status, however, affects her reproductive performance in several ways: offspring of high-ranking females survive longer, the age at which offspring reach sexual maturity is sooner when mothers are high-ranking, and average female life span is longer among high-ranking females (Fig. 4). Furthermore, a female’s dominance correlates with age until she reaches 21 where the rank she has achieved strongly predicts her rank for the next 10 years. Understanding the effects of this kind...
of reproductive skew is important to the genetic diversity of species. When populations become small and isolated, genetic diversity of the population is likely to be reduced by the successful reproduction of a few dominant individuals (Pusey, Williams, and Goodall, 1997).

C. Evolutionary Paleontology

Paleontology addresses broadscale evolutionary patterns by tracing origins and fates of lineages and major groups, changes in characteristics and relationships of evolving lineages, and temporal variations in species diversity through the fossil record. Paleontologists, however, can only assess these changes through shifts in animal size. Hence paleontology cannot address the diversity or effects of the evolution of organismal physiology, only its pattern.

Research in evolutionary paleontology has become increasingly interdisciplinary and as a result four key, interrelated research questions have emerged: (a) What are the rules that govern biodiversity dynamics and do they apply at all temporal and spatial scales? (b) Why are major evolutionary innovations unevenly distributed in space and time? (c) How does the biosphere respond to environmental perturbations at global and region scales? (d) How have biological systems influenced the physical and chemical nature of the earth's surface, and vice versa (Jablonski, 1999)? For example, mass extinctions are not only important for the species that they eliminate, but for the evolutionary diversification and ecological restructuring that occur in their aftermath (Jablonski, 1998). Five major mass extinctions in the Mesozoan fossil record were followed by rapid recovery of global taxonomic diversity and the radiation of new or previously minor groups. In a recent study, Jablonski examined the end-Cretaceous biotic recovery patterns among marine mollusks in four geographic provinces: the U.S. Gulf Coast, northern Europe, Northern Africa, and Pakistan-northern India. Although extinction intensities and selective effects were similar in the molluskan faunas of all four areas, the regions differed in the dynamics of diversifying clades and in the proportion of local taxa and invaders in the post-extinction biotas (Jablonski, 1998). In particular, three of the four regions lacked a rapid expansion and decline of "bloom taxa," as well as proportionally fewer invaders early in the recovery phase. Applying this result—contrary to the existing paradigm where global compendia of taxa are amassed to provide broad taxonomic analyses—Jablonski showed that distinct biogeographic regions are semiautonomous in their response to environmental crises. While some investigators have proposed that the differential success of individual clades was due to intrinsic properties of the clade, Jablonski's studies show that biogeographic, environmental, and paleoecologic context are critical to any understanding how diversity changes.

D. Evolutionary Developmental Biology

Developmental, evolutionary biology reveals how changes in genetic information at the DNA level or genotype are translated into changes in organismal morphology or phenotype. For example, the molecular phylogenies of 18S ribosomal sequences show that the bilatera should be reorganized into three different clades: the Deuterostoma (chordates and echinoderms), the Lophotrochozoa (molluscs, annelids, platyhelminths, and banchiopods), and the Ecdysozoa (chelicerates, crustacea, myriapods, insects, and onychophora). The result of this reordering, however, is that early evolving groups in derived clades must have displayed both ancestral and derived characters in combination (Knoll and Carroll, 1999).

Developmental biology shows that the apparent discrepancies between molecular and morphological data are resolved when the mechanisms that regulate gene expression are understood. For example, a marked trend in arthropod evolution has been an increase in body segment and appendage diversity. Hox genes play three major roles in the evolution of arthropods: (a) the entire arthropod clade contains the same set of Hox genes that are responsible for the recent diversity in Cambrian animals, (b) the increase in segment diversity is correlated with changes in the relative domains of Hox genes, and (c) changes in the morphology of homologous appendages are correlated with changes in the array of genes that are regulated by the same Hox gene (Knoll and Carroll, 1999). Therefore, most arthropod body plans are the result of very similar complexes of Hox genes. In addition, the correlation between diversification of Hox gene expression and patterns of evolution suggest that the diversification of bilaterian body plans is primarily due to the evolution of developmental regulatory systems, not changes in the genes themselves (Knoll and Carroll, 1999).

E. Evolutionary Physiology and Morphology

Evolutionary physiology and morphology determine how biochemical, physiological, and anatomical aspects
of organisms affect their ability to adapt to new environments and ways of life. Contemporary studies focus on asking how form and function relate to each other during evolution, why some species are tolerant of broad physical ranges and others are not, and if there is more than one way for an animal to adapt to its environment. One of the most interesting discoveries of this field is that even when the fossils of ancestors are similar to their descendants, the descendant’s biochemistries may be diverse and able to adapt to almost any physiological need.

For example, the macroevolutionary importance of plant chemistry on herbivore host shifts is critical to understanding the evolution of insect-plant interactions (Becerra, 1997). Members of the New World, monophyletic genus of the beetle, *Blepharida* (Chrysomelidae: Alticinae), feed mainly on *Bursera* (Burseraceae). While plants in the *Bursera* genus produce an array of terpenes that are toxic or repellent to most herbivores, they only decrease *Blepharida* survival and growth rate. To determine the importance of plant chemistry to the phylogenetic diversification of the beetles and their feeding behaviors, Becerra (1997) constructed a dendrogram of *Bursera* species based on their chemical similarity. She found that most clades of *Bursera* include plants that are in different chemical groups suggesting that their chemical similarity is partially independent of plant phylogeny. She also found that *Blepharida* rarely shifted between chemically dissimilar plants and, in fact, subclades of *Blepharida* appear to have colonized species of plants that belonged to only one chemical group even if they shifted between hosts belonging to different subclades several times (Fig. 5). Therefore, comparison of plant phylogeny and plant chemical variation indicate a greater influence of host plant chemistry than host plant phylogeny in the evolution of the *Blepharida* and *Bursera* interaction.

**F. Evolutionary Systematic Biology**

Evolutionary biology is based on the principle that all organisms share a common history. Evolutionary systematics attempts to organize history by cataloging species, determining their genealogical or phylogenetic relationships, and classifying them into inclusive or hierarchical groups. Systematic studies prior to 1980 were largely based on morphological comparisons among taxa. Recent research, however, has reanalyzed and broadened past work using molecular tools. DNA analyses provide independent estimates of the age of phylogenetic groups while extending and revising our understanding of organismal relationships. For example, corals are Cnidarians, the sister taxa of the Bilateria, and evolved at least 600 million years ago.
ago. The scleractinian corals (reef-building corals) have a continuous fossil record from the mid-Triassic (about 160 million years ago) and support some of the world’s most diverse marine communities. Despite this detail, their skeletal variability and morphology have made it difficult to understand the relationships among coral families and suborders (Romano and Palumbi, 1996). Molecular phylogenetic analysis of mitochondrial 16S ribosomal RNA, however, has helped to clarify coral evolution and revealed that scleractinians originated 80 million years earlier than they appear in the fossil record. The mitochondrial data show a deep split in the coral lineages that predates skeleton formation and thus is invisible in the fossil record. Using a combined molecular and traditional analyses, Roman and Palumbi suggest that the scleractinian skeleton evolved repeatedly and converged during evolution. This is important because if the selective events that led to the multiple convergences in the scleractinians can be identified, we will be able to understand the selective events that led to the complex, marine ecosystems that the corals support (Romano and Palumbi, 1996).

G. Human Evolution

Evolutionary biology is studied from two different viewpoints: a conceptual view, such as theoretical population genetics, or an organismal view, such as Homo. Anthropologists and biologists studying human evolution use the same principles, concepts, and tools as other biologists. Some investigators focus on the genetic processes that affect contemporary human populations, such as human genetics and medical genetics. Other investigators draw information from evolutionary systematics, paleontology, genetics, ecology, and animal behavior. One example of an organismal study that taps information from many fields of anthropological research explores the hypothesis that cooking has played a critical role in the evolution of human social systems. Using evidence from the fossil record, Wrangham et al. (1999) propose that cooking first appeared among hominoids 1.9 million years ago. They propose that the observed increase in female body mass was the effect of increased food energy that is released when foods are cooked, and that smaller tooth size reflected the reduced digestive effort that results when foods are partially broken down before eating. Furthermore, foods that were cooked and made more digestible resulted in increased nutrient accessibility. Food that was gathered, accumulated and stored until cooked became a valuable and defensible resource that changed the economics hominoid foraging. Females fitness, limited by access to energetic resources, could be enhanced if the females formed bonds with food-guards or males willing to defend them. Females could have competed for the best food-guards through extended sexual attractiveness. Wrangham et al. (1999) suggest that cooking was responsible for the evolution of human social systems where pair-bonds are embedded within multimale and multifemale communities and are supported by strong, mutual, and frequently competing sexual conflicts (Wrangham, Jones, Laden, Pilbeam, and Conklin-Brittain, 1999).

H. Evolutionary Genetics

Evolutionary genetics, which includes population genetics, uses both molecular and classical genetic methods to understand the origin and effects of mutation and recombination. It employs empirical studies and theoretical methods to uncover the roles of genetic drift, gene flow, and natural selection to predict and interpret evolutionary change. Perhaps the most important finding in this field is the large amount of genetic variation in protein phenotypes that seem to have no immediate selective value. Furthermore, recombinant DNA technology has revealed that there are large amounts of previously undetected polymorphism at the nucleotide level that do not affect changes in amino acid sequence. Constancy in protein sequence in the face of nucleotide polymorphism shows that selection is acting directly at the DNA level preventing the evolution of deleterious proteins (Kreitman, 1983).

I. Molecular Evolution

Springing directly from the molecular revolution, molecular evolution focuses on causes of evolutionary change at the level of the gene, protein, and genome. It takes the position that evolution occurs on diverse scales of time. For example, antibodies recognize and distinguish among specific molecular patterns of antigens. Because antigens are structurally diverse, the repertoire of antibodies must be large enough to protect organisms from a wide range of pathogens and toxic agents. This diversity is achieved through three, germ-line, gene segments, which combine to generate antibody types. These are further altered through somatic mutation (not inherited) that increases the antigen’s affinity and specificity as the immune response proceeds (French, Laskov, and Scharff, 1989; Tonegawa, 1983).

Wedemayer et al. (1997) compared the structure of the germline Fab fragments, and its complex with hapten, to the corresponding crystal structure of the affinity
matured antibody 48G7. Combinatorial association between CDR1 and CDR2 regions in the primary antibody repertoire allows antibody combining sites to be more diverse than the flanking, germine-encoded sites. However, rather than a few large changes occurring at the active site, antibody maturation depends on small additive changes, many of which result from somatic mutations, to reconfigure the active site. By mapping structural changes due to nine amino acids, Wedemayer et al. (1997) showed that somatic mutations resulted in 30,000 times higher affinity for hapten. None of these, however, directly contacted the hapten; only two were within 5.5 Å of the hapten binding site and all others were at least 10 Å away. Instead, the mutations resulted in reorganized hydrogen bonding networks on the molecule’s surface, shifting the backbone conformation of the antibody protein.

See Also the Following Articles

BIODIVERSITY, EVOLUTION AND • DARWIN, CHARLES • GENES, DESCRIPTION OF • PHYLOGENY

Bibliography


I. Introduction
II. Why Does Biodiversity Need Conservation?
III. What Threatens Biodiversity?
IV. What Is Biodiversity Conservation?
V. Methods of Conservation
VI. Sustainable and Integrated Biodiversity Conservation

GLOSSARY

complementary conservation Application of a range of conservation techniques (including ex situ and in situ) to conserve the target taxon, one technique acting as a backup to another. The degree of emphasis placed on each technique depends on the conservation aims, the type of species being conserved, the resources available, and whether the species has utilization potential.

conservation Maintenance of the diversity of living organisms, their habitats, and the interrelationships among organisms and their environment.

econeography Analysis of a species’ ecological, geographical, and taxonomic characteristics to assist in the formulation of collection and conservation priorities.

effective population size (Ne) Number of conserved individuals that would undergo the same amount of random genetic drift as the actual population.

ex situ conservation Conservation of the components of biological diversity outside their natural habitats.

gene pool The total genetic diversity found within an individual species or species group.

genetic erosion Loss of genetic diversity from a species, often caused by anthropogenic factors.

in situ conservation Conservation of ecosystems and natural habitats and the maintenance and recovery of viable populations of species in their natural surroundings and, in the case of domesticates or cultivated species, in the surroundings where they have developed their distinctive properties.

keystone species Usually the dominant species within a habitat that tend to define it physiognomically and ecologically, for example, by determining nutrient and water cycling.

target taxon Species or species group that the conservation action is focused upon.

EX SITU AND IN SITU CONSERVATION approaches focus on the protection of species outside their natural habitats or in their natural surroundings, respectively. This article outlines the basic concepts of biodiversity conservation by discussing why conservation is needed, the major threats to animal and plant species, and how we can set about ensuring that the diversity of living organisms is maintained at the ecosystem, species, and genetic levels. The importance to humankind of the
EX SITU, IN SITU CONSERVATION

The Convention on Biological Diversity is also considered. A simple model for biodiversity conservation is proposed and discussed. This involves the selection of target taxa for conservation, gene pool concepts, ecogeographic surveys, preliminary survey missions, clarification of conservation objectives and field exploration, the two basic conservation strategies (ex situ and in situ) and the range of conservation techniques, and ways that conservation is often linked to some form of utilization. Finally, the critical need of adopting a sustainable and integrated approach to conservation is emphasized, so that we can endeavor to pass to future generations the levels of biodiversity that we inherited from our forebears.

I. INTRODUCTION

The conservation of biodiversity is of critical importance now, because that very diversity is under threat of extinction and erosion, but also because it can be of direct and indirect benefit to humankind. Biodiversity benefits humans through the exploitation of animals and plants in agriculture and horticulture, the development of medicinal drugs, and the pivotal roles played by species in the functioning of all natural ecosystems. Biodiversity is also valuable for ethical, aesthetic, and recreational reasons. The fundamental importance of these issues to humankind was paramount at the United Nations Conference on the Environment and Development (UNCED) held in Rio de Janeiro, Brazil, in 1992. The Convention on Biological Diversity (CBD), which has subsequently been ratified by 175 countries (as of January, 1999), has as its objectives the conservation of biological diversity, the sustainable use of its components and the fair and equitable sharing of the benefits arising out of the utilisation of genetic resources. (Article 1: Objectives, Convention on Biological Diversity, 1992)

II. WHY DOES BIODIVERSITY NEED CONSERVATION?

Estimates of the total number of species in the world vary, and it is difficult to estimate rates of species extinction, but the consensus view summarized by Lugo (1988) was that 15–20% of all species could become extinct by the turn of the century. It is even more difficult, if not impossible, to estimate precise levels of genetic erosion, that is, the loss of genetic diversity from extant species. However, the loss of genetic diversity must always be faster than the loss of species because there will be some genetic erosion from the species that remain extant. Loss of any genetic diversity means that the affected plants or animals will not be able to adapt to changing conditions quite so readily; for example, potato cultivars with a narrow genetic base were unable to withstand the infection of late potato blight (Phytophthora infestans) in the late 1860s in Ireland, the crop was devastated, and millions were forced to emigrate or starve. Although genetic erosion cannot be quantified accurately, it seems likely that virtually all species are currently suffering loss of genetic variation to varying degrees. Therefore, using Lugo’s figures as a starting point, it may be estimated that 25–35% of plant and animal genetic diversity could possibly be lost over the 12 year period leading up to the year 2000 (Maxted et al., 1997a).

Of course loss of biodiversity also occurs at other levels than species, for habitats and ecological communities can also be degraded or destroyed. For example, virtually all of the natural grasslands in the United States have been lost since 1942 (Spellerberg, 1996), over 90% of natural wetlands in New Zealand have been lost since European settlement, and Fernside (1990) estimated that world forest loss was proceeding at 20,000 km² per year. In a similar study, FAO and UNEP (1991) found that annual rates of forest loss had increased from 113,000 km² to 169,000 km² per year between 1980 and 1990 in 76 countries.

It is important to realize, however, that both species extinction and genetic erosion can be natural events, just as species and genetic evolution are natural, nature is and it seems has always been dynamic. Yet the current levels of species extinction and genetic erosion are dramatically higher than the so-called background levels that existed hundreds, thousands, and millions of years ago. Humankind now has the capacity to drastically alter the world environment in ways that were not previously possible, and these anthropogenic changes are undoubtedly increasing the rates of species and genetic extinction.

III. WHAT THREATENS BIODIVERSITY?

The threat to biodiversity as a result of anthropogenic changes is not universal for all species. Some species...
are in greater danger of genetic erosion (or even of complete extinction) than others. These dangers must be evaluated carefully, so that those exposed to the highest risk can be given higher priority for conservation. However, it must also be borne in mind that levels of threat often change rapidly and unexpectedly. Thus, an area may suddenly come under the threat of industrial development, road-building, or logging. Diamond (1989), Gomez-Campo et al. (1992), WCMC (1992), and WRI et al. (1992) discuss in detail the kind of events that may lead to genetic erosion. These may be broadly grouped under the general headings of

- destruction, degradation, and fragmentation of natural habitats;
- overexploitation and incidental take;
- introduction of exotic species that compete with, prey on, or hybridize with native species;
- human socioeconomic change and upheaval;
- changes in agricultural practices and land use; and
- calamities, both natural and man-made.

It is valuable to establish a system by which the relative threat of genetic erosion can be assessed objectively. Guarino (1995) proposed a model for estimating the threat of genetic erosion that a taxon (wild or cultivated) faces in a particular region. The model may be used without having to visit the region involved, providing some background data are available on the taxon and the area. The model is based on scoring a number of parameters, such as the abundance of the taxon, the level of agricultural development, and the proximity and intensity of various types of human activity to the populations being studied. The higher the score, the greater the risk of genetic erosion and therefore the higher priority for in situ or ex situ conservation.

The International Union for Conservation of Nature and Natural Resources (IUCN, now known as the World Conservation Union) has developed a system of categories of conservation status, which is based on detailed knowledge of the population dynamics and genetics of the species concerned, the so-called IUCN Red Data List Categories (IUCN, 1994) (Fig. 1). Table 1 shows how some of these categories are determined. Using these categories, for example, a “critical” species would be assigned higher conservation priority than a “vulnerable” species. However, IUCN has tended to focus their attention almost exclusively on species extinction rather than genetic erosion within individual gene pools, and the latter may be of equal importance in terms of loss of biodiversity.
### IV. WHAT IS BIODIVERSITY CONSERVATION?

Spellerberg and Hardes (1992) stated that biodiversity conservation aims to "maintain the diversity of living organisms, their habitats and the interrelationships between organisms and their environment." But how is this to be achieved?

Conservationists must clearly define and understand the processes involved, and then attempt to develop practical techniques to achieve this objective. When undertaking a particular conservation exercise, a conservationist must use his or her knowledge of genetics, ecology, geography, taxonomy, and many other disciplines to understand and manage the biodiversity that is being conserved.

It is important to stress that conservation is not just about individual plant and animal species, but includes all aspects of biodiversity from ecosystems (a community of organisms and its abiotic environment), through communities (a collection of species found in a common environment or habitat), species, and populations, to genetic diversity within species. In recent years there has been a differentiation between conservation at the ecosystem and at the genetic levels, and these may be referred to as ecological and genetic conservation, respectively.

Ecological conservation focuses on the conservation of whole communities, although the survival of individuals and the extinction of particular species are a major concern, both are viewed in the larger context of overall community health. This form of whole-community conservation is exemplified by the 'Man and the Biosphere' program (UNESCO, 1996), which established a network of biosphere reserves representing distinct biomes and ecosystems throughout the world. The clear emphasis was on conservation of ecosystems: this program believed that individual species should be conserved as a component of ecosystems rather than on a species-by-species basis.

Genetic conservation focuses more explicitly on particular taxa and attempts to conserve the full range of genetic (allelic) variation within those taxa. The aim of this form of conservation is often utilitarian, for genetic diversity conservation is often linked intimately to human utilization. However, Maxted et al. (1997b) stressed the following points: first, species usefulness does not have to be defined in the strictest utilitarian sense, because plants and animals considered of aesthetic value are equally worthy of receiving human value as those of immediate use to plant or animal breeders; and second, in many cases individual species cannot be conserved without conserving the communities in which they naturally occur. So the distinction between the two basic forms of conservation is in practice blurred and may be viewed as artificial and of limited semantic importance, because the conservation of ecosystems and species are intimately linked. Just as it is difficult to focus conservation effort on the generality of the entire ecosystem, in practice conservationists need to focus on something more tangible that can be monitored and managed, even if only as an exemplar for the ecosystem as a whole.

So to undertake effective conservation, species inter-
actions must be understood as far as possible. Even if the conservation target is a population of a single species, no populations survive in isolation. They exist within a community or ecosystem and interact with other species and the abiotic environment. Obvious examples of interactions include pollinators, seed dispersers, microbial symbionts, herbivores (whether natural or introduced by humans), predators, and pathogens. Thus, even when applying genetic in situ conservation, the maintenance of genetic diversity will have to be considered within the context of whole-ecosystem conservation.

The so-called keystone species are important in this context, for these species contribute significantly to the structure of a community or its processes; they are the dominant species. The removal of a keystone species renders other members of a community vulnerable to extinction. Tropical trees that produce a rich food resource in the form of fruit or seeds, for example, can be considered keystone species, as they provide a vital food source for a diverse array of mammals and birds. Generally keystone species play an important part in interactions between different trophic levels, whether they are predators, herbivores, mutualists such as pollinators, or decomposers. So when considering the conservation of any particular species within an ecosystem, one must identify the inherent interactions and ensure their maintenance if the conservation project is to be successful and sustainable.

V. METHODS OF CONSERVATION

There is a need to develop appropriate methodologies for biodiversity conservation, particularly in the tropics. The tropical regions of the world have the highest levels of biodiversity, but their fauna and flora are the least well known and are most under threat. Also, tropical nations have few conservationists and often they are insufficiently trained; furthermore, the resources available for conservation activities are relatively limited. To address these issues, the CBD asks nations to:

Promote and encourage research which contributes to the conservation and sustainable use of biological diversity, particularly in developing countries. (Article 12; CBD, 1992)

Improve capability for determining scientific research priorities via transparent, “user friendly” risk evaluation methodologies. (Chapter 35 of Agenda 21; CBD, 1992)

Maxted et al. (1997b) responded to this requirement—to clarify and enhance the methodologies and research programs that currently enable scientists to classify, conserve, manage, and utilize biodiversity—by proposing a model for plant and animal genetic diversity conservation (Fig. 2). The raw materials of genetic conservation are genes within gene pools, which represent the total diversity of genetic material of the particular taxon being conserved. The product of the gene pool (seeds, ovules, etc.) is either preserved or utilized as genetic diversity. The processes that link the raw matter and the utilized gene pool represent genetic conservation.

A. Selection of Target Taxa

Conservation activities will always be limited by the financial, temporal, and technical resources available. Conservation of ecosystems or species has a cost and
the effort expended is directly related to how much society values that species and is therefore willing to pay. It is impossible to actively conserve or monitor all species, so it is important to make the most efficient and effective selection of species on which to focus conservation efforts. This choice should be objective and based on logical, scientific, and economic principles related to the perceived value of the species. Maxted et al. (1997a) discussed the sort of factors that provide a species with "value": current conservation status, potential economic use, threat of genetic erosion, genetic distinction, ecogeographic distinction, national or conservation agency priorities, biological importance, cultural importance, relative cost of conservation, conservation sustainability, and ethical and aesthetic considerations.

Rarely will one of the these factors on its own lead to a taxon being given conservation priority. More commonly, all or a range of these factors will be assessed for a particular taxon and then it will be given a certain level of national, regional, or world conservation priority. If the overall score passes a threshold level or is higher than those of competing taxa, the taxon will be conserved; it will then be either collected and conserved ex situ, an appropriate reserve will be established, or on-farm conservation will be proposed. Having listed the factors that affect the selection of target taxa in terms of value to society, those related to potential economic use will commonly be given higher comparative value, especially in economically poorer economies where income generation is of the highest priority. This anthropocentric and utilitarian view in the selection of conservation priorities may offend some conservationists, but when financial resources for conservation are, and are likely to remain, limited, and when men, women, and children are still suffering from malnutrition in many parts of the world, there appears to be no practical or ethical alternative to giving those species of most direct use to people the highest value and thus the highest conservation priority.

B. Project Commission

In practice, once taxa are selected for conservation, the actual conservation activities are necessarily preceded by some form of commission statement. This is likely to establish the objectives of the conservation, specify the target taxa and target areas, state how the material is to be utilized and where the conserved material is to be safely duplicated, and perhaps indicate which conservation techniques are to be employed. A clear, concise commission statement will help to focus subsequent conservation activities.

Who writes the commission? The commission statement may be written by those contracting the conservation or those who actually undertake the conservation work. The commission may vary in taxonomic and geographic coverage from a systematic collection program for a single species throughout its geographic range to a range of target taxa from a restricted location, for example, onion (Allium) species of Central Asia, large cat species worldwide, or chickpeas (Cicer) from the Western Tien Shen. In each case, however, a particular group of taxa from a defined geographical area must be considered to be insufficiently conserved (either in situ or ex situ), of sufficient actual or potential use, and/or endangered to warrant active conservation.

C. Ecogeographic Survey and Preliminary Survey Mission

Once the target taxon or group of taxa have been selected and delimited, the conservationist begins to amass and synthesize fundamental biological data to help formulate an appropriate conservation strategy. The synthesis and analysis of these data enables the conservationist to make vital decisions concerning, for example, which taxa to be included in the target group, where to find these taxa, which combination of ex situ and in situ conservation to use, what sampling strategy to adopt, where to store the germplasm and site the reserve or what captive breeding program would be most successful. If the basic biological data for a particular species, for example, the close lentil relative Lens orientalis, indicate that the species has been previously found on stony slopes at the edge of the Fergana valley in Uzbekistan, then further material of this species is likely to be currently found under similar constraints and is less likely to be found in different habitat types or in far distant regions.

The process of collating and analyzing geographical, ecological, and taxonomic data for use in designing conservation strategies is referred to as ecogeography and was defined by Maxted et al. (1995) as an ecological, geographical and taxonomic information gathering and synthesis process for a particular taxon. The results are predictive and can be used to assist in the formulation of collection and conservation priorities. Ecogeographic studies involve the use of large and
complex data sets obtained from the literature and from the compilation of passport data associated with herbarium specimens and germplasm accessions. These data are synthesized to produce three basic products: the database, which contains the raw data for each taxon; the conspectus, which summarizes the data for each taxon; and the report, which discusses the contents of the database and conspectus, as well as proposing appropriate conservation strategies. Ecogeographic techniques enhance the efficiency of crop relative and wild species conservation because they enable the conservationist to identify clearly the geophysical regions and ecological niches that the taxon inhabits, and so not only identify areas with high numbers of target taxa, but also areas that contain high taxonomic or genotypic diversity of taxa, uniqueness of habitat, economic or breeding importance, and so on.

If the available ecogeographic data for the target taxon are limited, the conservationist will not have sufficient background biological knowledge to formulate an effective conservation strategy. In this case it would be necessary to undertake an initial survey mission to gather the novel ecogeographic data required on which to base the actual strategy. The survey mission may be in the form of “coarse grid sampling,” which involves traveling throughout a likely target region and sampling sites at relatively wide intervals over the whole region. The precise size of the interval between sites depends on the level of environmental diversity across the region, but it may involve sampling every 1–30 km. The population samples and data collected during this mission can then be used to formulate further conservation priorities and to develop an appropriate strategy, thus providing the same result as the ecogeographic survey for groups that are better biologically understood.

D. Conservation Objectives

The products of the ecogeographic survey or survey mission provide a basis for the conservationist to formulate future conservation priorities and strategies for the target taxon. Within the target area, zones of particular interest may be identified, for example, areas with high concentrations of diverse taxa, low or very high rainfall, or high frequency of saline soils or extremes of altitude or exposure. In general, areas with very distinctive characteristics are likely to contain plants with distinct genes or genotypes. If a taxon is found throughout a particular region, then the researcher can use the ecogeographic data to positively select a series of diverse habitats to designate as reserves. If a taxon has been found at one location but not at another with similar ecogeographic conditions, then the ecogeographer may suggest that these similar locations should be searched. Within the target taxon, specific variants may be identified that warrant conservation priority, for example, species that have previously unrecognized utilization potential, populations that are particularly in danger of genetic erosion, or those that had not previously been conserved.

The conservationist must set out a clear, concise statement of the proposed conservation strategy for the target taxon and, if appropriate, prioritize actions. These may have been established in the project commission, but if not the conservationist should undertake the task. This should answer questions such as: Which populations require conservation? Can local farmers play a part in conservation activities? Do population levels require close monitoring? Should a national or international collecting team be directed to collect the priority target taxa? What conservation strategy or strategies are appropriate? What combination of conservation techniques is appropriate or is a more detailed study required before any of these questions can be answered?

E. Field Exploration

Once the conservation objectives have been clarified, whichever conservation strategy is to be applied, the ecogeographic information is used to locate and identify the general locality of the animal or plant populations that are to be conserved. The ecogeographic data will rarely be sufficiently comprehensive to precisely locate actual populations. Therefore, the preparatory element of conservation activities will be followed by field exploration, during which actual populations are located. Ideally, populations of the target taxon that contain the maximum amount of genetic diversity in the minimum number of populations will be identified, but how is this goal to be achieved? Commonly there will be too much diversity in both crops and wild species to conserve all their alleles, even if these were known then or at some future time. Thus the conservationist must attempt to conserve the range of diversity that best reflects the total genetic diversity of the species. How many animals or plants must be sampled, which specimens and what pattern of sampling is appropriate? To answer these specific questions the conservationist should know the amount of genetic variation within and between populations, local population structure, the breeding system, taxonomy and ecogeographic requirements of the target taxon, and many other biologi-
cal details. Some of this information will be supplied following the ecogeographic survey, but some will remain unavailable. Therefore, the practice of field exploration will be modified depending on the biological information on the target taxon and target area that is available.

For a botanical project, the field botanist should select populations if they are found on the periphery of the target taxon’s distribution or those that contain morphological or ecological variants. Atypical populations or those growing under atypical conditions may possess genes or alleles that are unknown or extremely rare in the target taxon’s center of diversity, and this material possibly contains genetic variation that is of special use to breeders (e.g., disease or pest resistance or adaptation to soil or climate that is unknown in the crop itself).

F. Conservation Strategies

There are two basic conservation strategies, each composed of various techniques, that the conservationist can adopt to conserve genetic diversity once it has been located. The two strategies are ex situ and in situ conservation. Article 2 of the Convention on Biological Diversity (CBD, 1992) provides the following definition of these categories:

**Ex situ** conservation means the conservation of components of biological diversity outside their natural habitats.

**In situ** conservation means the conservation of ecosystems and natural habitats and the maintenance and recovery of viable populations of species in their natural surroundings and, in the case of domesticates or cultivated species, in the surroundings where they have developed their distinctive properties.

There is an obvious fundamental difference between these two strategies: **ex situ** conservation involves the sampling, transfer, and storage of target taxa from the target area, whereas **in situ** conservation involves the designation, management, and monitoring of target taxa where they are encountered. Because of this fundamental difference, there is little overlap between the two strategies. The two basic conservation strategies may be further subdivided into the following specific techniques, which are discussed in Sections V.G and V.H:

**Ex Situ**
- Seed/embryo storage
- In vitro storage
- Semen/ovule/pollen/DNA storage
- Field gene bank/livestock parks
- Botanic/zoological garden

**In Situ**
- Genetic reserve
- On-farm

G. Ex Situ Techniques

In ex situ conservation, genetic variation is maintained away from its original location and samples of a species, subspecies, or variety are taken and conserved either as living collections of plants or animals in field gene banks, botanic or zoological gardens, and arboreta, or as samples of seed, semen, ovules, tubers, tissue explants, pollen, or DNA maintained under special artificial conditions.

1. Seed/Embryo Storage Conservation

**Ex situ** seed/embryo collection and storage is the most convenient and widely used method of genetic conservation (Fig. 3). Seeds and embryos are the natural dispersal, storage, or generative organs for the majority of species. This technique involves collecting samples from individuals or populations and then transferring them to a gene bank for storage, usually at sub-zero temperatures. The procedure used for the bulk of orthodox-seeded plant species is to dry the seeds or embryos to a suitable moisture content (5–6%) before freezing at −20°C, but this method is only suitable for species that can be dried and stored at low temperature without losing viability. The advantages of this technique are that it is efficient and reproducible, and feasible for short-, medium-, and long-term secure storage. However, the disadvantages are that there are problems in storing recalcitrant-seeded plant species. The latter species cannot be dried and frozen in the way used for orthodox seeds, because they rarely produce seed or are normally clonally propagated.

![Figure 3](image-url) Collecting seed from Tajikistan for ex situ conservation.
2. Botanical/Zoological Garden Conservation

Historically, botanical or zoological gardens were often associated with physic or medicinal gardens or displays of single specimens of zoological curiosities, and as such they did not attempt to reflect the genetic diversity of the species. These gardens now hold living collections of species that were collected in a particular location and moved to the garden to be conserved. The advantage of this method is that gardens do not have the same constraints as many other conservation agencies; they have the freedom to focus on wild species that may otherwise not be given sufficient priority for conservation. Yet there are two disadvantages to this technique. The first is that the number of species that can be genetically conserved in a botanical or zoological garden will always be limited because of the available space. The majority of these gardens are located in urban areas in temperate countries, and at their present sites most expansion would be prohibitively expensive. The majority of botanical and animal diversity is located in tropical climates, yet because most botanical and zoological gardens are in temperate countries, the collections must be kept in expensive greenhouses or other facilities, which also limits the space available. The second disadvantage is related to the first, namely, very few individuals of each species can be held, and this severely restricts the range of genetic diversity found in the wild that is protected. However, if the target species is very near extinction and only one or two specimens remain extant, this objection of course does not hold.

3. In Vitro Conservation

In vitro conservation involves the maintenance of ex-plants in a sterile, pathogen-free environment, and it is widely used for vegetatively propagated and recalcitrant-seeded species. This method offers an alternative to field gene banks. It involves the establishment of tissue cultures of accessions on nutrient agar and their storage under controlled conditions of either slow or suspended growth. The main advantage is that it offers a solution to the long-term conservation problems of recalcitrant, sterile, or clonally propagated species. The main disadvantages are the risk of somaclonal variation, the need to develop individual maintenance protocols for the majority of species, and the relatively high-level technology and high cost required. The best answer for cheap, long-term in vitro conservation in the future may be cryopreservation (Hoyt, 1988), that is, the storage of frozen tissue cultures at very low temperatures, for example, in liquid nitrogen at −196°C. If this technique can be perfected to reduce the damage caused by freezing and thawing, it may be possible to preserve materials indefinitely.

4. Field Gene Bank/Livestock Park Conservation

The conservation of germplasm in field gene banks or livestock parks involves the collecting of plant or animal specimens from one location and the transfer and conservation at a second site. It has traditionally been the method for recalcitrant- (whose seeds cannot be dried and frozen without loss of viability) or sterile-seeded plant species or for those species for which it is preferable to store clonal material. Field gene banks are commonly used for species such as cocoa, rubber, coconut, mango, coffee, banana, cassava, sweet potato, and yam. Livestock parks or rare breed centers, as distinct from zoos, emphasize captive breeding programs and therefore genetic conservation. The advantages of field gene banks and livestock parks are that the species are easily accessible for utilization and evaluation can be undertaken while the material is being conserved. The disadvantages are that the material is restricted in terms of genetic diversity, is susceptible to pests, disease, and vandalism, and may require large areas of land.

5. Pollen/Semen/Ovule/DNA Conservation

The storage of pollen grains is possible under appropriate conditions that allow their subsequent use for crossing with living plant material. It may also be possible in the future to regenerate haploid plants routinely from pollen cultures, but no generalized protocols have been developed yet. The development of artificial insemination techniques in recent years has made semen and ovule storage routine, especially for domesticated animals. The storage of DNA under prescribed conditions can easily and inexpensively be achieved given the appropriate level of technology, but the regeneration of entire organisms from DNA cannot be envisaged at present, although single or small numbers of genes could subsequently be utilized. The advantage of pollen storage is that it is low cost and simple, but the disadvantage is that only paternal material would be conserved, and with DNA storage there are problems with subsequent gene isolation, cloning, and transfer.

H. In Situ Techniques

In situ techniques involve the maintenance of genetic variation at the location where it is encountered, either in the wild or in traditional farming systems. The majority of existing nature reserves and natural parks were
established to conserve animals or to protect aestheti-

cally beautiful landscapes, but even today few have

plant conservation as their primary goal, let alone the

genetic conservation of species (Hoyt, 1988).

1. Genetic Reserve Conservation

Conservation of wild species in a genetic reserve in-
volves the location, designation, management, and
monitoring of genetic diversity in a particular natural
location (Fig. 4). This technique is the most appropriate

for the bulk of wild species, because it can, when the
management regime is minimal, be relatively inexpen-
sive. Whether dealing with plants or animals, the objec-
tive is to contain the minimum number of individuals

that can maintain genetic diversity within the species.

If too few individuals are protected, genetic diversity
will decline over time, and if too many are protected,
resources may be wasted in managing the large popula-
tion. To guide such efforts, conservationists will need
to estimate the effective population size (\(N_e\)), that is,

the number of conserved individuals that would un-
dergo the same amount of random genetic drift as the

actual population. Genetic reserves are appropriate for
animals as well as for orthodox and non-orthodox
seeded plant species, because numerous taxa can be
protected in a single reserve that allows the continued
evolution of species. However, the disadvantages are

that the conserved material is not immediately available
for human exploitation and, if the management regime
is minimal, little characterization or evaluation data
may be available. In the latter case, the reserve manager
may even be unaware of the complete specific composi-
tion of the reserve that he or she is managing.

2. On-Farm Conservation

Farmer-based conservation involves the maintenance
of traditional crop or animal breeds or cultivation sys-
tems by farmers within traditional agricultural systems.
On traditional farms, what are generally known as
‘land-races’ of plants are sown and harvested, and each
season the farmers keep a proportion of harvested seed
for re-sowing. Traditional breeds of domestic animal
are maintained by inter-breeding within and between
local village stocks. Thus the land-race or breed is highly
adapted to the local environment and is likely to contain
locally adapted alleles that may prove useful for specific
breeding programs. Home garden plant conservation is
a closely related variant of on-farm conservation of land-
races but on a smaller scale. It involves the cultivation
of more species-diverse material in home, kitchen,
backyard, or door-yard gardens. These home gardens
focus on medicinal, flavoring, and vegetable species
(e.g., tomatoes, peppers, digitalis, mint, thyme, pars-
ley). The overall advantage of the on-farm technique is
that it ensures the maintenance of highly adapted land-
races and breeds and those wild species that traditional
agriculture often depends on. However, these land-
races or traditional breeds may yield less than their
modern counterparts, and so traditional farmers may
require some subsidy and possibly monitoring to ensure
continued farming. It should be noted that contempo-
rary economic forces tend to act against the continued
farming of ancient land-races and breeds, which are
currently suffering rapid genetic erosion; many face
imminent extinction. A back-up system of ex situ con-
servation is therefore essential, as discussed in Sec-
tion VI.

I. Community-Based Conservation

When applying all of the conservation techniques dis-
cussed here, professional conservationists have often
failed to appreciate the role that local communities have
successfully played in conserving animal and plant di-
versity within their local environment. It is now gener-
ally accepted that the present-day wealth of domesti-
cated and non-domesticated biodiversity would not
exist were it not for the conscious effort of local commu-
nities over millennia to conserve biodiversity in all its
forms. For example, indigenous farmers in the Andes
maintain a gene pool of over 3000 varieties of potatoes
representing eight cultivated species, and in Papua New
Guinea approximately 5000 varieties of sweet potato
are cultivated, with a single farmer growing up to 20
varieties in one garden (McNeely et al., 1995). For wild
species, Prance et al. (1995) showed that four groups

FIGURE 4 Surveying plant populations for an in situ reserve in Turkey.
of Amazonian Indians use up to 79% of the tree species in their home ranges, and Milliken et al. (1992) in a similar study found that 81% of tree and vine species were utilized; this number rises to 86% when other categories of plants are added from literature sources. So local communities have had and continue to have an essential role in biodiversity conservation. They not only continue to conserve by using traditional practices for their own future direct and indirect benefit, but also increasingly work in collaboration with professional conservationists to conserve broad-based biodiversity for the benefit of their host countries and humankind as a whole. Specifically, collaboration involving conservationists and local communities increases the overall efficiency of "professional" conservation, because local communities have a broader local knowledge base concerning the animal and plant species found in their area. Local communities are therefore able to assist in the development of a more practical, focused, and hopefully efficient approach to locally targeted conservation. The employment of a collaborative approach also empowers local people and engenders increased pride in native biodiversity and its conservation. In this way, rather than deferring responsibility to outside science-based experts, they can retain environmental responsibility and take greater pride in maintaining their own environment.

J. Conservation Products

The products of conservation activities are primarily conserved germplasm (seed, embryos, semen, and ovules), live plants and animals, dried plants, cultures, and conservation data. Ex situ conserved orthodox seed or animal semen and ovules are commonly held in gene or semen banks at sub-zero temperatures and, for seed, low moisture content to prolong their life. Live plants or animals are conserved in genetic reserves, field gene banks, botanical or zoological gardens, or parks and research laboratories. Germplasm that is stored in a suspended form, such as tissue, pollen, or DNA, is kept as cultures in specialist laboratory facilities. Dried voucher plant specimens are held in herbaria and linked to specific samples of germplasm, and are as much as possible made representative of the conserved populations. Conserved material is ideally associated with a range of passport data, which detail the taxonomic, geographical, and ecological provenance of the material. All passport data should be entered into a database and made available for the management of the material, the formulation of future conservation priorities and strategies, and any exploitation. The various conservation products, where they are stored, and where they should be duplicated are presented in Table II.

K. Conserved Product Dissemination

The conservation products are either maintained in their original environment or deposited in a range of ex situ storage facilities. Whether the germplasm, voucher specimens, or passport data are conserved in situ or ex situ, to ensure its safety it should ideally be duplicated in more than one location. The distribution of duplicate

---

### Table II: Conservation Products and Their Storage and Duplication Sites

<table>
<thead>
<tr>
<th>Conservation product</th>
<th>Storage site</th>
<th>Duplication site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant germplasm (seed, vegetative organs, etc.)</td>
<td>Gene bank</td>
<td>National, regional, and international gene banks, duplication with other conservation techniques</td>
</tr>
<tr>
<td>Animal germplasm (semen, ovules, eggs, embryos, etc.)</td>
<td>Gene bank</td>
<td>National, regional, and international gene banks, duplication with other conservation techniques</td>
</tr>
<tr>
<td>Live plants</td>
<td>Field gene bank, botanical garden, genetic reserve, on-farm</td>
<td>Duplication with other conservation techniques, e.g., gene bank storage of seed</td>
</tr>
<tr>
<td>Live animals</td>
<td>Zoological garden, genetic reserve</td>
<td>Duplication with other conservation techniques, e.g., storage of germplasm</td>
</tr>
<tr>
<td>Dried plants or preserved animals</td>
<td>Herbarium or museum</td>
<td>National, regional, and international herbaria or museums</td>
</tr>
<tr>
<td>Explants or plantlets</td>
<td>Tissue culture</td>
<td>Duplication with other conservation techniques, e.g., gene bank storage of seed</td>
</tr>
<tr>
<td>DNA and pollen</td>
<td>Various cultures</td>
<td>Duplication with other conservation techniques, e.g., gene bank storage of seed</td>
</tr>
<tr>
<td>Conservation data</td>
<td>Conservation database</td>
<td>Duplication with other national, regional, and international conservation agencies</td>
</tr>
</tbody>
</table>
sets of material avoids accidental loss of the material due to fire, economic or political difficulties, warfare, or other unforeseen circumstances. Duplication of the data is relatively easy from the conservation database, and copies should be held by the commissioning agency, relevant host country institutes, and other interested parties.

L. Biodiversity Utilization

As discussed earlier, there should be an intimate linkage between conservation and utilization. The products of conservation, whether they be “living” or “suspended,” should be made available for use by humankind. Conservation can be seen as the safe keeping of preserved material, so that the material is available at a future date. In certain cases the material can be used directly, say in the selection of forage accessions or local domesticated animal breeds, where little breeding is undertaken. The conserved material may also be used in reintroduction programs where the traditional breed or land-race has been lost locally owing to civil unrest or the application of perverse government incentives that encourage the alteration of traditional practices.

More commonly, the first stage of utilization will involve the recording of genetically controlled characteristics (characterization) and the plant material may be grown out under diverse environmental conditions to evaluate and screen for drought or salt tolerance, or be deliberately infected with diseases or pests to screen for particular biotic resistance (evaluation). The biotechnologist will be screening for single genes, which when located may be transferred into a host organism to generate more rapid growth, for example. The biochemist (bioprospector) will be screening for particular chemical products that may be of use to the pharmaceutical industry. The products of utilization are therefore numerous, including new varieties, new crops, improved breeds, and pharmaceuticals as well as more nebulous but equally valuable products such as a beautiful or more diverse environment for human recreational activities.

VI. SUSTAINABLE AND INTEGRATED BIODIVERSITY CONSERVATION

Having discussed in detail what constitutes biodiversity and how that biodiversity can be conserved, the point should be made that any biodiversity conservation program should be sustainable and integrated. Each conservation technique has its advantages and disadvantages. The two strategies of ex situ and in situ conservation should not be seen as alternatives or in opposition to one another, but rather as being complementary, as stated in Article 9 of the Convention on Biological Diversity (CBD, 1992). One conservation strategy or technique can act as a backup to another, the degree of emphasis placed on each depending on the conservation aims; the type of species being conserved; the resources available, and whether the species has utilization potential. The efficacy of adopting an integrated approach to conservation, or as some have called it a “holistic” approach (Withers, 1993), is now well established as the only sustainable option. Therefore, when formulating an overall conservation strategy for a species, conservationists should think in terms of applying a combination of the different techniques available, including both in situ as well as ex situ techniques, where the different methodologies complement each other. It may be helpful to think of the various techniques as pieces in a jigsaw puzzle that will complete the overall conservation strategy and thus ensure the maintenance of plant or animal genetic diversity.

The adoption of an integrated approach requires the conservationist to consider the characteristics and needs of the particular gene pool being conserved, and then to assess which of the strategies or combination of techniques offers the most appropriate option to maintain genetic diversity within that taxon. To formulate the conservation strategy, the conservationist may also need to address not only biological questions but also the practical and political ones: What are the species’ storage characteristics? What do we know about its breeding success in captivity? Do we want to store the germplasm over the short, medium, or long term? How important is the species? Where is the species located and how accessible is it/does it need to be? Are there legal issues relating to access? How good is the infrastructure of the established reserves? What backup is necessary and/or desirable? How does the species conservation strategy fit within the local community development program? Given answers to these questions, the appropriate combination of techniques to conserve the gene pool can be applied in a pragmatic and balanced manner. The integration of conservation and community development is an important point to stress. We cannot expect local communities to altruistically forgo development for the benefit of a more abstract greater good, and so the practical application of the conservation strategy may need to be a compromise between scientific protocols and meeting the needs and
desires of local people. Therefore the actual combination of techniques will be formulated afresh for each species or group of species, demonstrating the flexibility of the integrated approach.

Sustainability in the sense of conformance is a fundamental concept for conservation. Whether seed, semen, or embryos are collected for ex situ conservation in a gene bank, animals are incorporated into a captive breeding program, or a habitat is designated as a reserve, each option has a financial cost and it would be a waste of limited conservation funds not to ensure that the conservation project is sustainable at least in the medium term. If the species or genetic material is lost from a reserve, the resources expended on establishing the reserve would have been wasted and the cost of rehabilitating populations using materials stored ex situ would have to be considered. The latter option is commonly expensive and may require extensive research to ensure that the reintroduced animals or plants do not likewise go extinct. Unfortunately, many conservation projects are funded on a short-term basis, so it is essential that an effective project exit strategy is developed so the conservation program itself is sustainable.

Not only is it necessary to integrate the different conservation strategies and techniques, and to involve the local community in a sustainable conservation project, but it is also important to integrate the different potential agencies involved. This is particularly true for large in situ reserve projects where the project may naturally span national borders and professional disciplines. In these cases, the project team must ensure that the local, provincial, national, regional, and international conservation agencies, as well as professionals from the different disciplines involved such as environmentalists, foresters, agriculturalists, and politicians, work together to promote the success of the conservation project.

See Also the Following Articles

**ENDANGERED ECOSYSTEMS • ENVIRONMENTAL ETHICS • GENE BANKS • GENETIC DIVERSITY • KEYSTONE SPECIES • SPECIES INTERACTIONS**

**Bibliography**


EXTINCTION, CAUSES OF

Richard B. Primack
Boston University

I. An Expanding Human Population
II. Habitat Destruction
III. Habitat Fragmentation
IV. Habitat Degradation and Pollution
V. Global Climate Change
VI. Overexploitation
VII. Exotic Species
VIII. Disease
IX. Multiple Factors

GLOSSARY

air pollution Lowering of air quality due to release of toxic materials by factories, automobiles, fires, and other human activities.
disease Infections by parasitic organisms that can cause weakness, decreased reproduction, and death.
exotic species Species that occurs outside of its natural range owing directly or indirectly to human activity.
global climate change Current and predicted changes in global temperature, rainfall, and other aspects of weather due to increased human production of carbon dioxide and other greenhouse gases.
habitat fragmentation Process by which a continuous area of habitat is divided into two or more fragments by roads, farms, fences, logging, and other human activities.

overexploitation Harvesting of a natural resource, such as fish or timber, at a rate more rapidly than it can be naturally replenished.
water pollution Lowering of water quality due to input of sewage, pesticides, agricultural run-off, and industrial wastes that can result in harm to aquatic plants and animals.

IF SPECIES AND NATURAL COMMUNITIES ARE ADAPTED TO LOCAL ENVIRONMENTAL CONDITIONS, why should they be faced with extinction? Shouldn’t species and communities be able to persist in the same places that they have for thousands of years? Why are species going extinct now? The answers to these questions have become clear in recent decades: massive disturbances caused by people have altered, degraded, and destroyed the natural landscape on a vast scale, driving species and even communities to the point of extinction. Current rates of human-induced species extinctions are around 1000 times greater than past natural rates of extinction. The balance between natural rates of speciation and extinction have been drastically upset by human activities. The process of evolution will eventually create new species, but it will take thousands, if not millions, of years for these species to develop. And numerous unique species, such as pandas, elephants, and cheetahs, will be gone forever.
I. AN EXPANDING HUMAN POPULATION

The major threats to biological diversity that result from human activity are habitat destruction, habitat fragmentation, habitat degradation (including air and water pollution), the overexploitation of species for human use, the introduction of exotic species, and the increased spread of disease (Table I). Most threatened species face at least two or more of these threats, speeding their way toward extinction and hindering efforts to protect them. Typically, these threats develop so rapidly and on such a large scale that species are not able to adapt genetically to the changes or to disperse to a more hospitable location. These threats will continue to increase in the coming decades as the human population increases, as development and overexploitation continue, as the remaining natural habitats disappear, and as the global climate continues to change.

These threats to biological diversity are all caused by an ever-increasing use of the world’s natural resources by an expanding human population. In Europe, only 15% of the land area remains unmodified by human activities, and the amount of many specific habitat types remaining is below 10%. The greatest destruction of biological communities has occurred during the last 150 years, during which the human population grew from 1 billion in 1850, to 2 billion in 1930, to 6 billion in 1999. World population will reach an estimated 10 billion by the year 2050 (Fig. 1). Human numbers have increased because birth rates have remained high while mortality rates have declined, particularly during the last century, as a result of both modern medical discoveries (specifically the control of disease) and the presence of more reliable food supplies. Population growth has slowed in the industrialized countries of the world, but it is still high in many areas of tropical Africa, Latin America, and Asia, where the greatest biological diversity is also found.

People use natural resources, such as fuelwood, wild meat, and wild plants, and convert vast amounts of natural habitat for agricultural and residential purposes. Because some degree of resource use is inevitable, population growth is partially responsible for the loss of biological diversity. All else being equal, more people equals less biodiversity. Some scientists have argued strongly that controlling the size of the human population is the key to protecting biological diversity. However, population growth is not the only cause of species extinction and habitat destruction: overconsumption of resources is also responsible. The rise of industrial capitalism and materialistic modern societies has greatly accelerated demands for natural resources, par-

TABLE I
Factors Responsible for Some Extinctions and Threatened Extinctions

<table>
<thead>
<tr>
<th>Group</th>
<th>Habitat loss</th>
<th>Overexploitation</th>
<th>Species introduction</th>
<th>Predators</th>
<th>Other</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extinctions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mammals</td>
<td>19</td>
<td>23</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>Birds</td>
<td>20</td>
<td>11</td>
<td>22</td>
<td>0</td>
<td>2</td>
<td>37</td>
</tr>
<tr>
<td>Reptiles</td>
<td>5</td>
<td>32</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Fishes</td>
<td>35</td>
<td>4</td>
<td>30</td>
<td>0</td>
<td>4</td>
<td>48</td>
</tr>
<tr>
<td>Threatened Extinctions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mammals</td>
<td>68</td>
<td>54</td>
<td>6</td>
<td>8</td>
<td>12</td>
<td>—</td>
</tr>
<tr>
<td>Birds</td>
<td>58</td>
<td>30</td>
<td>28</td>
<td>2</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>Reptiles</td>
<td>53</td>
<td>63</td>
<td>17</td>
<td>3</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td>Amphibians</td>
<td>77</td>
<td>29</td>
<td>14</td>
<td>—</td>
<td>3</td>
<td>—</td>
</tr>
<tr>
<td>Fishes</td>
<td>78</td>
<td>12</td>
<td>28</td>
<td>—</td>
<td>2</td>
<td>—</td>
</tr>
</tbody>
</table>

* From Reid and Miller (1989), based on data from various sources.
* These values represent the percentage of species that are influenced by the given factor. Some species may be influenced by more than one factor, thus the rows may exceed 100%.
* Overexploitation includes commercial, sport, and subsistence hunting, as well as live animal capture for any purpose.
* Threatened species and subspecies include those given in the IUCN categories critically endangered, endangered, and vulnerable.
Extinction, Causes of

A. Unequal Use of Natural Resources

In many countries there is extreme inequality in the distribution of wealth, with a small percentage of the population controlling, owning, and consuming much of the wealth and natural resources such as good farmland, livestock, and timber resources. As a result, poor, rural people with no land or resources of their own are forced by necessity to destroy biological communities and hunt endangered species to extinction. Large landowners and business interests that force local farmers off their land is a common pattern in many countries of the developing world, a pattern often backed up by the government, the police, and the army. Political instability, lawlessness, and war also displace farmers into remote, undeveloped areas where they feel safer; most practice shifting cultivation, a form of agriculture involving cutting down forest, burning the plant material, and planting crops in the nutrient-rich ash, because it is the simplest way to make a living when they may be obliged to move again within a short time. The landless farmers and their families often exploit natural resources in their surroundings just to stay alive; often these resources are components of species-rich biological communities.

The responsibility for the destruction of biological diversity in species-rich tropical areas also lies in the unequal use of natural resources worldwide (Fig. 2). People in industrialized countries (and the wealthy minority in the developing countries) consume a disproportionate share of the world's energy, minerals, wood products, and food. Each year the average U.S. citizen uses 43 times more petroleum products, 34 times more aluminum, and 386 times more paper products than the average citizen of India. Wealthy countries and individuals leave a widespread "environmental footprint" in which their excessive patterns of consumption affect a wide area of the world that must supply their needs. This excessive consumption of resources is not sustainable in the long run. If this pattern is adopted by the expanding middle class in the developing world, it will cause massive environmental disruption.

B. Large-Scale Development Projects

In many cases, the factors causing habitat destruction, particularly in the developing world, are the large industrial and commercial activities associated with a global economy—such as mining, cattle ranching, commercial fishing, forestry, plantation agriculture, manufacturing, and dam construction—and initiated with the goal of making a profit. Many of these projects are funded by
national governments and international development banks and are touted as sources of jobs, commodities, and tax revenues. Others are initiated and funded by large multinational corporations. However, this exploitation of natural resources often is neither efficient nor cost-effective because the emphasis in these industries is on short-term gain, often at the expense of the long-term sustainability of the natural resources, and generally with little regard for the local people who depend on the resources.

II. HABITAT DESTRUCTION

Increasing human populations and their activities use even greater proportions of the world’s terrestrial and marine environments and associated natural resources, resulting in the inevitable destruction of species, genetic variation, habitats, and ecosystem processes.

A. Habitat Loss

Habitat loss is the primary threat to the majority of vertebrate species currently facing extinction, a generalization that is certain to be true for threatened invertebrates, plants, and fungi as well. In many countries of the world, particularly on islands and in locations where human population density is high, most of the original habitat has been destroyed. More than 50% of the wildlife habitat has been destroyed in 49 of 61 Old World tropical countries. In tropical Asia, fully 65% of the primary forest habitat has been lost, with particularly high rates of destruction reported for Bangladesh (96%), Sri Lanka (86%), Vietnam (76%), and India (78%). Sub-Saharan Africa has similarly lost a total of about 65% of its forests, with losses being most severe in Rwanda (80%), Gambia (89%), and Ghana (82%). Two biologically rich nations, Zimbabwe and the Democratic Republic of Congo (formerly Zaire), are relatively better off, still having about half of their forests, although it is too soon to say how the recent civil war in the latter country has harmed its wildlife population. Present rates of deforestation vary considerably among countries, with particularly high annual rates of 1.5–2% for tropical countries such as Vietnam, Paraguay, Mexico, Cote d’Ivoire, and Costa Rica. In the Mediterranean region, which has been densely populated by people for thousands of years, only 10% of the original forest cover remains.

For many important wildlife species, the majority of habitat in their original range has been destroyed, and very little of the remaining habitat is protected. For
certain Asian primates, such as the Javan gibbon, more than 95% of the original habitat has been destroyed, and some of these species are protected on less than 2% of their original range. The orangutan, a great ape that lives in Sumatra and Borneo, has lost 63% of its range and is protected in only 2% of its range. Such losses of habitat inevitably lead to the loss of species.

B. Rain Forest Loss

The destruction of tropical rain forests has come to be synonymous with the loss of species. Tropical rain forests occupy 7% of the Earth’s land surface, but they are estimated to contain over 50% of its species. Therefore, the loss of these forests will result in the extinction of vast numbers of species. These evergreen to partly evergreen forests occur in frost-free areas below about 1800 m in altitude and have at least 100 mm (4 inches) of rain per month in most years. These forests are characterized by a great richness of species and a complexity of species interaction and specialization unparalleled in any other community. The original extent of tropical rain forests and related moist forests has been estimated at 16 million km², based on current patterns of rainfall and temperature. Less than half of this expanse was still present as of 1990, with the rate of destruction accelerating in the last decade. On a global scale, most rain forest destruction results from small-scale cultivation of crops and collection of firewood by poor farmers, most of whom have moved to forest areas to practice shifting cultivation out of desperation and poverty.

Other major causes include commercial logging in clear-cutting and selective logging operations, clearing for cattle ranches, clearing for cash-crop plantations (oil palm, cocoa, rubber, etc.) plus road building, and mining. At the current rate of destruction (around 140,000 km² per year), there will be no large blocks of tropical forest left after the year 2040, except in the relatively small national parks and protected areas, and a few remote areas of the Brazilian Amazon, central Africa, and the islands of Borneo and Papua New Guinea (Fig. 3). The situation is actually more grim than these projections indicate because the world’s population is still increasing, and poverty is on the rise in many developing tropical countries, putting ever greater demands on the dwindling supply of rain forest.

C. Other Threatened Habitats

The plight of the tropical rain forests is perhaps the most widely publicized case of habitat destruction, but other habitats are also in grave danger.

1. Tropical Dry Forests

Tropical deciduous forests contain a large number of species, which in some places rival the diversity of the tropical rain forest. The land occupied by tropical deciduous forests is more suitable for agriculture and cattle ranching than the land occupied by tropical rain forests. Moderate seasonal rainfall, in the range of 250 to 2000 mm per year, allows mineral nutrients to be retained in the soil where they can be taken up by plants. Consequently, human population density is five times greater in dry forest areas of Central America than in adjacent rain forests. Today, the Pacific Coast of Central America has less than 2% of its original extent of deciduous dry forest remaining, and many species of this community have been eliminated or are threatened with extinction.

2. Grasslands

Temperate grasslands are another habitat type that has been almost completely destroyed by human activity with a consequent loss of species. It is relatively easy to convert large areas of grassland to farmland and cattle ranches. Illinois and Indiana, for example, originally contained 15 million ha (37 million acres) of tall-grass prairie, but now only 1400 ha (3500 acres) of this habitat—one ten-thousandth of the original area—remain undisturbed; the rest has been converted to farmland. This remaining area of prairie is fragmented and widely scattered across the landscape. Widespread efforts are being made to restore prairies in many areas of the world. Though such efforts are to be encouraged, it will not be possible to bring back the species that have already been lost.

3. Wetlands and Aquatic Habitats

Wetlands are critical habitats for fish, aquatic invertebrates, aquatic plants, and birds. They are also a resource for flood control, drinking water, and power production. Many aquatic species in lakes and streams have limited distributions that make them especially vulnerable to extinction. Wetlands are often filled in or drained for development, or they are altered by channelization of watercourses and dams. Many wetlands have been degraded by chemical pollution and siltation. When this happens, the aquatic species are fated with extinction. All of these factors are currently affecting the Florida Everglades, one of the premiere wildlife refuges in the United States, which is now on the verge of ecological collapse. During the last 200 years, over half of the wetlands in the United States have been destroyed, resulting in 40–50% of the freshwater snail species in the southeastern United
Tropical rain forests are found predominantly in wet, equatorial regions of the Americas, Africa, and Asia. Eight thousand years ago, tropical forests covered the entire shaded area, but human activities have resulted in the loss of a great deal of forest cover, shown here in the darkest shade. In the lighter-shaded area, forests remain, but they are secondary forests that have grown back following cutting; plantation forest such as rubber and teak; or forests degraded by logging and fuelwood collection. Only in the regions shown in black are there still blocks of intact natural tropical forest large enough to support all of their resident biodiversity. (After Bryant et al., 1997.)
States becoming either extinct or endangered. More than 97% of the vernal pools in California’s San Diego County have been destroyed; these unusual wetlands fill up with water in the winter and dry out in the summer, and support a unique endemic biota. When the habitat is damaged, many sensitive aquatic species are not able to survive.

4. Mangroves
Mangrove forests are among the most important wetland communities in tropical areas. Composed of species that are among the few woody plants able to tolerate salt water, mangrove forests occupy coastal areas with saline or brackish water, typically where there are muddy bottoms. Such habitats are similar to those occupied by salt marshes in the temperate zone. Mangroves are extremely important breeding grounds and feeding areas for shrimp and fish. They also play an important role in reducing storm damage. Despite their great economic value, mangroves are often harvested for timber and charcoal production and cleared for coastal development. In recent years, mangroves have been increasingly cleared for rice cultivation and commercial shrimp hatcheries, particularly in Southeast Asia, where as much as 15% of the mangrove area has been removed for aquaculture. The loss of mangroves is extensive in some parts of South and Southeast Asia; the percentage of mangroves lost is particularly high for India (85%), Thailand (87%), Pakistan (78%), and Bangladesh (73%).

5. Coral Reefs
Tropical coral reefs contain an estimated one-third of the ocean’s fish species in only 0.2% of its surface area. Already 10% of all coral reefs have been destroyed, and as many as 30% more could be destroyed in the next few decades. The most severe destruction is taking place in the Philippines, where a staggering 90% of the reefs are dead or dying. The main culprits are pollution, which either kills the coral directly or allows excessive growth of algae; sedimentation following the removal of forests; overharvesting of fish, clams, and other animals; and, finally, fishermen blasting with dynamite and releasing cyanide and other poisons to collect the few remaining living creatures.

Extensive loss of coral reefs is expected within the next 40 years in tropical East Asia, around Madagascar and East Africa, and throughout the Caribbean (Fig. 4). In the Caribbean, a combination of overfishing, hurricane damage, water pollution, and disease is responsible for a dramatic decline of a large proportion of the coral reefs and their replacement by fleshy macroalgae. Elkhorn and staghorn corals, which were formerly common and gave structure to the community, have already become rare in many locations.

D. Desertification
Many biological communities in seasonally dry climates are degraded into man-made deserts by human activities, a process known as desertification. These commu-
nities include tropical grasslands, scrub, and deciduous forest, as well as temperate shrublands, such as those found in the Mediterranean region, southwestern Australia, South Africa, Chile, and southern California. While these areas initially may support agriculture, repeated cultivation, especially during dry and windy years, often leads to soil erosion and loss of water-holding capacity in the soil. Land may also be chronically overgrazed by domestic livestock, such as cattle, sheep, and goats, and woody plants may be cut down for fuel. The result is a progressive and largely irreversible degradation of the biological community and the loss of soil cover. Ultimately, the region takes on the appearance of a desert, and the original species of the area are lost.

III. HABITAT FRAGMENTATION

In addition to outright destruction, habitats that formerly occupied wide unbroken areas are now often divided into pieces by roads, fields, farms, houses, industries, fences, powerlines, and a broad range of other human activities. Habitat fragmentation is the process whereby a large, continuous area of habitat is both reduced in area and divided into two or more fragments. When habitat is destroyed, a patchwork of habitat fragments may be left behind. These fragments are often isolated from one another by a highly modified or degraded landscape (Fig. 5). Fragmentation almost always occurs during a severe reduction in habitat area, but it can also occur when area is reduced to only a minor degree if the original habitat is divided by roads, railroads, canals, power-lines, fences, oil pipelines, fire lanes, or other barriers to the free movement of species. The island model of biogeography is applicable to this situation: the fragments may be considered habitat islands in an inhospitable human-dominated sea. Habitat fragments differ from the original habitat in two important ways: (1) fragments have a greater amount of edge for the area of habitat and (2) the center of each habitat fragment is closer to an edge.

A. Barriers to Dispersal

Fragmentation may limit a species’ potential for dispersal and colonization. Many bird, mammal, and insect species of the forest interior will not cross even very short distances of open area. In many species this is due to the high risk of predation in edge and open habitats. Habitat fragmentation creates barriers to the normal dispersal and colonization processes. In an undisturbed environment, seeds, spores, and animals move passively and actively across the landscape. When they arrive in a suitable but unoccupied area, new populations begin to develop at that site. Over time, populations of a species may build up and go extinct on a local scale as the species disperses from one suitable site to another and the biological community undergoes succession. Habitat fragmentation limits the ability of species to disperse to new habitats, and consequently the species may gradually die out.

Habitat fragmentation also reduces the foraging ability of individual animals. Many animal species, either as individuals or social groups, need to move freely across the landscape to feed on widely scattered resources. However, fences and other barriers may prevent the natural migration of animals, such as wildebeest or bison, forcing them to overgraze an unsuitable habitat, eventually leading to starvation of the animals and degradation of the habitat. Barriers to dispersal can restrict the ability of widely scattered species to find mates, leading to a loss of reproductive potential for many animal species. Plants also may have reduced seed production if butterflies and bees are less able to migrate among habitat fragments to pollinate flowers.

B. Edge Effects

Habitat fragmentation often changes the microenvironment at the fragment edge, resulting in increased light levels, higher daytime temperatures, higher wind speeds, and lower humidity. Each of these edge effects can have a significant impact on the vitality and composition of the species in the fragment. Species sensitive to humidity such as amphibians, many insects, and herbaceous plants will be eliminated from the forest fragments. Also, increased wind, lower humidity, and higher daytime temperatures make fires more likely in forest fragments. Fires may spread into habitat fragments from nearby agricultural fields that are being burned regularly, as in sugarcane harvesting, or from the irregular activities of farmers practicing shifting cultivation. In the process, many species will be eliminated.

C. Interspecific Interactions

Habitat fragmentation also increases the vulnerability of the fragment to invasion by exotic and native pest species. Omnivorous native animals, such as raccoons, skunks, and blue jays, and introduced animals, such as
FIGURE 5 The forested areas of Warwickshire, England (shown in black), were fragmented and reduced in area over the centuries from A.D. 400 to 1960.

rats, may increase in population size along forest edges, where they can eat foods found in both undisturbed and disturbed habitats. These aggressive feeders may seek out the nests of interior forest birds, often preventing successful reproduction for many bird species hundreds of meters from the nearest forest edge. Weedy plant species and exotic herbivores can eliminate native plant species along the edges of fragments, and disease can similarly spread into the interior of habitat fragments.

IV. HABITAT DEGRADATION AND POLLUTION

Even when a habitat is unaffected by overt destruction or fragmentation, the communities and species in that habitat can be profoundly affected by human activities. Biological communities can be damaged and species locally extirpated or driven to extinction by external factors that do not change the structure of dominant
plants in the community, so that the damage is not immediately apparent. For example, in temperate deciduous forests, physical degradation of a habitat might be caused by frequent, uncontrolled ground fires; these fires might not kill the mature trees, but the rich perennial wildflower community and insect fauna on the forest floor would gradually become impoverished. Keeping too many cattle in grassland communities gradually changes the biological community, often eliminating many native species and favoring exotic species that can tolerate grazing. Frequent boating and diving among coral reefs degrade the community, as fragile species are crushed by divers' flippers, boat hulls, and anchors. The most subtle form of environmental degradation is pollution, commonly caused by pesticides, sewage, fertilizer run-off from agricultural fields, industrial chemicals and wastes, emissions from factories and automobiles, and sediment deposits from eroded hillsides. The general effects of pollution on water quality, air quality, and even the global climate are cause for great concern, not only because of the threats to biological diversity, but also for their effects on human health.

A. Pesticides

The dangers of pesticides were brought to the world's attention in 1962 by Rachel Carson's influential book *Silent Spring*. Carson described a process known as biomagnification through which DDT (dichlorodiphenyltrichloroethane) and other organochlorine pesticides become concentrated as they ascend the food chain. These pesticides, at the time widely used on crop plants to kill insects and sprayed on water bodies to kill mosquito larvae, were harming wildlife populations, especially birds that ate large amounts of insects, fish, or other animals exposed to DDT and its by-products. Birds with high levels of concentrated pesticides in their tissues, particularly raptors such as hawks and eagles, became weak and tended to lay eggs with abnormally thin shells that cracked during incubation. As a result of failure to raise young and the outright death of many adults, populations of these birds showed dramatic declines throughout the world.

Recognition of this situation in the 1970s led many industrialized countries to ban the use of DDT and other stable pesticides. The ban eventually allowed the partial recovery of many bird populations, most notably peregrine falcons, ospreys, and bald eagles. Nevertheless, the continuing use of these classes of chemicals in other countries is still cause for concern, not only for endangered animal species, but for the potential long-term effects on people, particularly the workers who handle these chemicals in the field and the consumers of agricultural products treated with these chemicals.

B. Water Pollution

Water pollution has serious consequences for human populations: it destroys important food sources and contaminates drinking water with chemicals that can cause immediate and long-term harm to human health. In the broader picture, water pollution often severely damages aquatic communities. Rivers, lakes, and oceans are used as open sewers for industrial wastes and residential sewage. Pesticides, herbicides, oil products, heavy metals (such as mercury, lead, and zinc), detergents, and industrial wastes directly kill organisms living in aquatic environments. Even if the organisms are not killed outright, these chemicals can make the aquatic environment so inhospitable that species can no longer thrive. In contrast to a dump in the terrestrial environment, which has primarily local effects, toxic wastes in aquatic environments diffuse over a wide area. Toxic chemicals, even at very low levels, can be concentrated to lethal levels by aquatic organisms. Many aquatic environments are naturally low in essential minerals, such as nitrates and phosphates, and aquatic species have adapted to the natural absence of minerals by developing the ability to process large volumes of water and to concentrate these minerals. When these species process polluted water, they concentrate toxic chemicals along with the essential minerals, which eventually poison the plant or animal. Species that feed on these aquatic species then ingest these concentrations of toxic chemicals.

Essential minerals that are beneficial to plant and animal life can become harmful pollutants at high levels. Human sewage, agricultural fertilizers, detergents, and industrial processes often release large amounts of nitrates and phosphates into aquatic systems, initiating the process of cultural eutrophication. Although small amounts of these nutrients can stimulate plant and animal growth, high concentrations often result in thick “blooms” of algae at the surface on ponds, lakes, and coastal areas. These algal blooms may be so dense that they outcompete other plankton species and shade out bottom-dwelling plant species. As the algal mat becomes thicker, its lower layers sink to the bottom and die. The bacteria and fungi that decompose the dying algae grow in response to this added sustenance and consequently absorb all of the oxygen in the water. Without oxygen, much of
the remaining animal life dies off, sometimes visibly in the form of masses of dead fish floating on the water’s surface. The result is a greatly impoverished and simplified community consisting of only those species tolerant of polluted water and low oxygen levels. The spreading “dead zone” where the Mississippi River enters the Gulf of Mexico is an example of the dire consequences of water pollution.

C. Air Pollution

In the past, people assumed that the atmosphere was so vast that materials released into the air would be widely dispersed and their effects would be minimal. But today several types of air pollution are so widespread that they damage whole ecosystems.

1. Acid Rain

Acid rain is created when nitrates and sulfates released into the air by the burning of fossil fuels combine with atmospheric water to form acids that fall as rain. Acid rain lowers the pH of soil moisture and water bodies such as ponds and lakes. Increased acidity alone damages many plant and animal species; acid rain has been blamed for the death of large numbers of trees in Europe and North America. As the acidity of water bodies increases, many fish either fail to spawn or die outright (Fig. 6). Both increased acidity and water pollution are two likely factors behind the dramatic decline in amphibian populations throughout the world.

2. Ozone and Smog

Automobiles, power plants, and other industrial activities release hydrocarbons and nitrogen oxides as waste products. In the presence of sunlight, these chemicals react with the atmosphere to produce ozone and other secondary chemicals, collectively called photochemical smog. Although ozone in the upper atmosphere is important in filtering out ultraviolet radiation, high concentrations of ozone at ground level damage plant tissues and make them brittle, harming biological communities and reducing agricultural productivity. Ozone and smog are detrimental to both people and animals when inhaled, so controlling air pollution benefits both people and biological diversity.

3. Effects on Lichens

Even when communities are not destroyed by air pollution, species composition may be altered as more susceptible species are eliminated. Lichens, symbiotic organisms composed of fungi and algae that can survive in some of the harshest natural environments, are particularly susceptible to air pollution. Because each lichen species has distinct levels of tolerance to air pollution, the composition of the lichen community can be used as a biological indicator of the level of air pollution.

4. Ozone Depletion and Ultraviolet Radiation

As a result of human use of chlorofluorocarbons (CFCs) and other ozone-depleting chemicals, the atmospheric ozone layer has been significantly reduced. Ozone plays an important role in filtering out harmful ultraviolet radiation in sunlight. With less atmospheric ozone, more solar ultraviolet radiation reaches the Earth’s surface. In humans, exposure to this UV radiation increases the risk of skin cancer. This UV radiation will possibly have a significant, negative impact on animals and plants exposed to direct sunlight, for example, amphibian eggs at the water surface.
V. GLOBAL CLIMATE CHANGE

Scientists are now intensively studying atmospheric carbon dioxide, methane, and other “greenhouse” gases that are transparent to light but that absorb heat. During the past 100 years, global levels of carbon dioxide (CO$_2$), methane, and other trace gases have been steadily increasing, primarily as a result of burning coal, oil, and natural gas. Clearing forests to create farmland and burning firewood for heating and cooking also contribute to rising concentrations of CO$_2$. Carbon dioxide concentration in the atmosphere has increased from 290 parts per million (ppm) to 350 ppm over the last 100 years, and it is projected to double somewhere in the latter half of the twenty-first century. Increased concentration of these gases in the atmosphere will almost certainly cause a rise in global temperatures.

A. Prediction of Climate Change

Many scientists believe that these increased levels of greenhouse gases have affected the world’s climate already, and that these effects will increase in the future. The best evidence seems to suggest that world climate has warmed by between 0.3° and 0.6° Celsius (°C) over the last 100 years. Predicting future weather patterns is extremely complex and difficult, even with all of the available weather data, simulation models, and supercomputers. However, the consensus among leading meteorologists is that the world climate will increase in temperature by an additional 1° to 3.5°C over the next century as a result of increased levels of carbon dioxide and other gases (Fig. 7). The increase could be even greater if carbon dioxide levels rise faster than predicted; it could be slightly less if all countries agreed to reduce their emissions of greenhouse gases. The increase in temperature will be greatest at high latitudes and over large continents. Many scientists also predict an increase in extreme weather events, such as hurricanes, flooding, and regional drought, associated with this warming.

B. Extinctions and Climate Change

In the northern temperature zone, species will need to migrate 200–300 km northward over the next century to remain in an area with the same climate conditions. It seems likely that many species will be unable to disperse rapidly enough to track the changing climate. Habitat fragmentation caused by human activities may further slow or prevent many species from migrating to new sites where suitable habitat exists. Many species of limited distribution and/or poor dispersal ability will undoubtedly go extinct, with widely distributed, easily dispersed species being favored in the new communities. Endemic mammals that are restricted to isolated

![Figure 7](image-url) Complex computer models of global climate predict that temperatures will increase significantly when CO$_2$ levels double, which is projected to occur in the middle- to late twenty-first century. Predicted temperature increases, shown in °C, are greatest over continents and at high latitudes (i.e., closer to the poles). (After Intergovernmental Panel on Global Climate Change (IPPC). 1996)
mountain peaks or fish species found in a single lake are examples of species that will not easily be able to cross inhospitable terrain to reach a new, suitable location. The best hope for many species will be to migrate higher on mountain slopes or to disperse along valleys, rivers, and coastlines that are aligned north to south.

Concerns about global climate change, as important as they are, should not, however, divert our attention away from the massive habitat destruction that is the principal current cause of species extinction. The preservation of intact communities and the restoration of degraded communities are the most important and immediate priorities for conservation.

VI. OVEREXPLOITATION

A. Exploitation in Traditional Societies

People have always hunted and harvested the food and other resources they need in order to survive. As long as human populations were small and the methods of collection simple, people could sustainably harvest and hunt the plants and animals in their environment. In traditional societies, restrictions were often imposed to prevent overexploitation of natural resources. For example, the rights to specific harvesting territories were rigidly controlled; hunting in certain areas was banned; there were often prohibitions against taking females, juveniles, and undersized individuals; certain seasons of the year and times of the day were closed for harvesting; and certain efficient methods of harvesting were not allowed. These restrictions, which allowed traditional societies to exploit communal resources on a long-term, sustainable basis, are almost identical to the rigid fishing restrictions imposed on and proposed for many fisheries in industrialized nations. Among the most highly developed restrictions were those of the traditional or artisan societies of Micronesia and Polynesia. However, there are also numerous cases of large bird and mammal species being hunted to extinction by traditional people using simple methods of hunting.

B. Exploitation in Modern Societies

As human populations have increased, their use of the environment has escalated and their methods of harvesting have become dramatically more efficient, leading to an almost complete depletion of large animals from many biological communities, leaving strangely “empty” habitats. Guns are now used instead of blow-pipes, spears, or arrows for hunting in the tropical rain forests and savannas. Powerful motorized fishing boats and enormous “factory ships” catch fish from the world’s oceans. Small-scale local fishermen now have outboard motors on their canoes and boats, allowing them to harvest wider areas more rapidly.

In much of the world today resources are exploited opportunistically. If a market exists for a product, local people will search their environment to find and sell it. Whether people are poor and hungry or rich and greedy, they will use whatever methods are available to secure that product. Sometimes traditional groups will sell the rights to a resource, such as a forest or mining area, for cash to buy desired goods. In rural areas, the traditional controls that regulate the extraction of natural products have generally weakened. Where there has been substantial human migration, civil unrest, or war, controls may no longer exist. As a result, species are exploited to the point of extinction.

1. Trade in Wildlife

The legal and illegal trade in wildlife is responsible for the decline of many species. Worldwide trade in wildlife is valued at over $10 billion per year, not including timber and edible fish. One of the most pervasive examples of this is the international trade in furs, in which hunted species, such as the chinchilla (Chinchilla spp.), vicuña (Vicugna vicugna), giant otter (Pteronura brasiliensis), and numerous cat species, have been reduced to low numbers. Overharvesting of butterflies by insect collectors, of orchids, cacti, and other plants by horticulturists, of marine mollusks by shell collectors, and of tropical fish for aquarium hobbyists are further examples of whole biological communities being targeted to supply an enormous international demand (Table II). It has been estimated that 500 to 600 million tropical fish are sold worldwide for the aquarium market, and many times that number are killed during collection and shipping.

Besides a surprisingly large legal trade, billions of dollars are involved in the illegal trade of wildlife. A black market links poor local people, smugglers, corrupt customs officials, rogue dealers, and wealthy buyers who don’t question the sources that they buy from. This trade has many of the same characteristics, the same practices, and sometimes the same criminal players as the illegal trade in drugs and weapons. Confronting these illegal activities has become a job for international law enforcement agencies. Clearly, people involved in the illegal trade of wildlife do not worry about species going extinct, unless it affects their profits.
2. Overfishing
In the North Atlantic, one species after another has been overfished to the point of diminishing return. The Atlantic bluefin tuna, for example, has experienced a 90% population decline over the past ten years. Similar grim scenarios can be recounted for other large fish prized for their flesh and for sport, such as the swordfish (Xiphias gladius). One of the most dramatic cases of overexploitation in recent years involves sharks. Over the last ten years, the fishing industry has been exploiting shark fisheries of the North Atlantic at a rate approximately 60% higher than the sharks can sustain in order to supply a growing international market. The boom in demand for shark meat and shark fins comes at a time when populations of many commercial fish species are severely depleted, so shark fishing has become a lucrative alternative for fishermen. But most shark species have a relatively slow reproductive cycle. As a result, the populations of many sharks are declining dramatically, and some species may soon go extinct.

Another striking example is the enormous increase in demand for seahorses (Hippocampus sp.) in China, which is tied to the nation's economic development. The Chinese use dried seahorses in their traditional medicine because it resembles a dragon and is believed to have a variety of healing powers. Around 20 tons of seahorses are consumed in China per year—roughly 6 million animals. Seahorse populations throughout the world are being decimated to supply this ever-increasing demand.

VII. EXOTIC SPECIES
Humans have radically altered patterns of species distribution by deliberately or accidentally transporting species throughout the world. The extent of this modern movement of human-transported species is unprecedented on a geological scale and has been described by Elton (1958) as ‘one of the great historical convulsions of the world’s flora and fauna.’ Many areas of the world are strongly affected by exotic species. The United States currently has over 70 species of exotic fish, 80 species of exotic mollusks, 2000 species of exotic plants, and 2000 species of exotic insects. Exotic perennials completely dominate many North American wetlands: purple loosestrife (Lythrum salicaria) from Europe dominates marshes in eastern North America, while Japanese honeysuckle (Lonicera japonica) forms dense tangles in bottomlands of the southeastern United States.

The effects of these exotic insects on the native insect fauna can be devastating. At some localities in the southeastern United States, the diversity of insect species has declined by 40% following the invasion of exotic fire ants. The fire ants either directly attack and consume other insect species or outcompete them. Many bird species have shown dramatic declines once fire ants enter their habitat, again because of direct attack as well as loss of insect prey.

The great majority of exotic species do not become established or dominant because the new environment is not suitable to their needs. However, a certain percentage of species do establish themselves in their new
homes, and many of these become abundant at the expense of native species. Exotic species are often able to thrive because their populations are not held in check by any of the local parasites or predators. These exotic species may displace native species through competition for limited resources, they may kill and eat native species to the point of extinction, or they may alter the habitat so that many natives are no longer able to persist. Exotic plant species frequently displace native species because they are better suited to the new conditions created by people, such as increased fire and introduced grazing animals.

Exotic species are considered to be the most serious threat facing the biota of the United States national park system. While the effects of habitat degradation, fragmentation, and pollution can potentially be corrected and reversed in a matter of years or decades as long as the original species are present, well-established exotic species may be impossible to remove from communities. They may have built up such large numbers and become so widely dispersed and so thoroughly integrated in the community that eliminating them may be extraordinarily difficult and expensive.

A. Exotic Species on Islands

Island species are particularly vulnerable to exotic species, because they have evolved in the absence of mainland herbivores and predators. The introduction of one exotic species to an island may cause the local extinction of numerous native species. Two examples illustrate the effects of introduced species on the biota of islands.

On Santa Catalina Island off the coast of California, 48 native plant species have been eliminated, primarily due to grazing by introduced goats, pigs, and deer. One-third of the plant species currently found on the island are exotics. Removal of goats from part of the island has led to the reappearance of several native plant species.

Birds of the Pacific islands are especially vulnerable to the effects of exotic species. For example, the brown tree snake (Boiga irregularis; Fig. 8) has been introduced onto a number of Pacific islands where it is devastating endemic bird populations. The snake eats eggs, nestlings, and adult birds; on Guam alone, the brown tree snake has driven ten endemic bird species to the point of extinction. Recent visitors have remarked on the absence of birdsong: “between the silence and the cobwebs, the rain forests of Guam have taken on the aura of a tomb” (Jaffe, 1994).

B. Exotic Species in Aquatic Habitats

Exotic species can have severe effects on vulnerable freshwater communities, in particular lakes and isolated stream systems. There has been a long history of introducing exotic commercial and sport fish species into lakes. Many introductions have been accidental. Often these exotic fish are larger and more aggressive than the native fish fauna, and they may eventually drive the local fish to extinction. Aquatic plants, invertebrates, and disease organisms can also become aggressive exotics outside their normal range.

Two examples illustrate the effects of exotic species on aquatic species. The freshwater fish fauna of the island nation of Madagascar has extremely high levels of endemism, with 14 of its 23 genera found nowhere else. Recent surveys of freshwater habitats were able to locate only 5 of the known native freshwater fish species of the island. Introduced fish dominate all of the freshwater habitats, and have completely replaced native species on the central plateau.

One of the most alarming recent invasions in North America was the arrival in 1988 of the Eurasian zebra mussel (Dreissena polymorpha) in the Great Lakes. Within two years zebra mussels had reached densities of 700,000 individuals per square meter in parts of Lake Erie, choking out native mussel species in the process. Zebra mussels have been found in the Detroit, Cumberland, and Tennessee Rivers; as it spreads south, this exotic species is causing enormous economic damage to fisheries, dams, power plants, water treatment facilities,
and boats, as well as devastating the aquatic communities it encounters.

VIII. DISEASE

Disease caused by internal parasites is a natural control mechanism that reduces populations when they reach high densities. However, levels of disease can often increase in populations as a result of human activity. When animals are confined to habitat fragments at abnormally high densities, disease may spread more easily among individuals. Also, animals under stress or weakened by living in a degraded or polluted environment may be more susceptible to disease. Furthermore, as areas are fragmented by human activities, disease can spread more easily from domestic animals into wild populations. At Tanzania’s Serengeti National Park, at least 25% of the lions have recently been killed by canine distemper, a viral disease apparently contracted from one or more of the 30,000 domestic dogs living near the park. For endangered species, such outbreaks can do phenomenal harm: the last population of black-footed ferrets known to occur in the wild was destroyed by canine distemper virus.

Diseases transported by people to new parts of the world can decimate species: North American chestnut trees (*Castanea dentata*), once common throughout the eastern United States, have been virtually obliterated by an ascomycete fungus carried on Chinese chestnut trees imported to New York City. Introduced fungal diseases are also eliminating elm trees (*Ulmus americana*) and flowering dogwoods (*Cornus florida*) from these forests (Fig. 9). Introduced diseases have particularly powerful adverse effects on endemic island species. Many endemic Hawaiian birds have been decimated and even driven to extinction by introduced avian malaria protozoans spreading from introduced bird species by introduced mosquitoes.

IX. MULTIPLE FACTORS

A combination of factors acting simultaneously or sequentially can overwhelm a species, as illustrated by the case of the large freshwater mussel *Margaritifera auricularia*. This species was formerly known from Western Europe to Morocco, but now it occurs in only one river and its adjoining canals in Catalonia, Spain. Its attractive shell and pearls have been used as ornaments by humans as far back as the Neolithic Age. The main reason for its decline, overcollecting, originally led to its disappearance from rivers in Central Europe in the fifteenth and sixteenth centuries, while pollution, destruction of freshwater habitats, and overcollecting continued to reduce its range in recent times. The mus-

![FIGURE 9](image)

Populations of flowering dogwood (*Cornus florida*) are declining in eastern North American forests because of anthracnose disease caused by the introduced fungus *Discula destructiva*. (Photograph by Jonathan P. Evans.)
sel is also affected by the loss of other species, since its larval stage needs to attach to certain species of fish to complete its life cycle. Unless strict conservation measures are implemented to prevent overcollecting, control water quality, maintain fish stocks, and protect the habitat, this culturally important species will soon be extinct. Such comprehensive conservation strategies are often needed to deal with the multiple threats to species.

Threats to biological diversity come from a number of different directions, but their underlying cause is the same: the magnitude of destructive human activity. It is often easy to blame a group of poor, rural people or a certain industry for the destruction of biological diversity, but the real challenge is to understand the local, national, and international linkages that promote the destruction and to find viable alternatives. These alternatives must include stabilizing the size of the human population, finding a livelihood for rural people that does not damage the environment, providing incentives and penalties that will convince industries to value the environment, and restricting trade in products that are obtained by damaging the environment. Yet an equally important part of the solution is to increase the willingness of wealthy and middle-class people in both developed and less-developed countries to reduce their consumption of the world’s resources and to pay fair prices for products that are produced in a sustainable, nondestructive manner.

See Also the Following Articles

DEFORESTATION AND LAND CLEARING • DESERTIFICATION • EXTINCTION, RATES OF • EXTINCTIONS, MODERN INSTANCES OF • INTRODUCED SPECIES, EFFECT AND DISTRIBUTION • MASS EXTINCTIONS • PESTICIDES, USE AND EFFECTS OF • POLLUTION, OVERVIEW

Bibliography

I. Measuring Extinction in the Fossil Record

II. Mass Extinction

III. Background Extinction and Turnover

GLOSSARY

**background extinction** A distinctly lower rate of extinction, more typical of most of the fossil record.

**extinction rate** The number or proportion of taxa becoming extinct per unit time or after an important geological temporal boundary.

**mass extinction** An extinction occurring over a short period of time that is of large magnitude, wide biogeographic impact, and involves the extinction of many taxonomically and ecologically distant groups.

**MANY SPECIES ARE THREATENED** by impending extinction and attempts have been made to assess population declines and to enact policies of recognizing endangerment by means of simple rules of thumb, such as the International Conservation Union’s rule of three successive years of 80% decline. It is difficult to develop a measure of extinction rates of entire floras or faunas, if only because we usually have scant knowledge of the species pool before the impact. This is particularly a problem in species-rich tropical habitats, where cryptic species abounds yet have not been identified completely. Surveys of especially rich faunas in tropical wet forests are only being organized now, and the decline of the areal extent of these forests in recent years makes it nearly impossible to measure extinction rates, except by means of indirect estimates of species—area relationships. The same applies to species-rich marine communities such as coral reefs.

I. MEASURING EXTINCTION IN THE FOSSIL RECORD

A. Why the Fossil Record?

Extinction is very much the domain of the paleontologist. We believe that we are now possibly living through a mass extinction caused by human disturbance of high diversity tropical habitats. But we know very little about the extinction of species, except by fairly obvious mechanisms such as hunting. Can knowing that the dodo or the passenger pigeon was hunted to extinction help us very much with understanding climatically induced changes in key structural groups such as forest trees and reef corals, and their dependent species? Could such spotty knowledge be used to extrapolate to the broad sweep of geological time? Paleontological data has the advantage of large banks of “before and after” data on biodiversity. Its weakness, however, is in associating extinctions with unique causes, as we shall see. Invasions have caused extensive extinctions on oceanic islands, particularly when alien predators overwhelmed small populations of endemic species in a matter of decades. Extinction on larger time scales, even
over hundreds of years, is much more difficult to track. Unfortunately, the time scale for larger-scale changes over 100,000 years or more is probably unapproachable by the neontologist, who can only observe “normal” extinction, and we don’t have much understanding of what occurs normally. It may be that fine-scale studies of the fossil record may eventually give us more insight into species-level extinction than neontological studies ever will. After all, the durations of animal species’ life spans range from the order of $10^5$ (land vertebrates) to $10^6$ to $10^7$ years (marine species). Even with millions of living species we are not likely to be able to document many cases of typical extinction of living animal species. The fossil record is probably our only hope of a model for study of extinction rates, especially on the scale of ocean basins and continents.

While its coverage of the total potential living biota is poor, the fossil record affords us a more complete glimpse of extinction rates of a number of readily fossilizable groups, both marine and terrestrial. We have a reasonably complete database that has stabilized over the years and one can readily trace extinctions across geological time horizons. The fossil record moreover gives us a deeper insight into what extinction really means. After all, we would like to produce a prospectus of the biological future of living communities following an extinction. Does the loss of a species have a disproportional importance, resulting in the extinction of many associated species? Following an extinction event, is there enough redundancy for the surviving species to evolve a new diverse fauna? Does the extinction of certain species cause the snowballing of a larger extinction event? With some judicious reasoning we can infer the answers to some of these questions with the use of the fossil record.

B. Measures and Types of Extinction Rates

Many extinctions in the fossil record appear to be precipitous and occur over short time periods of hundreds of thousands to a few million years. Impacts of extraterrestrial objects may have caused changes in a year or less. Such lengths of time are short when you consider the length of the record of the Phanerozoic era ($\sim 545$ million years).

It is possible to quantify the extent of the extinction with the following data:

1. The total pool of taxa before the extinction
2. The number of taxa that became extinct
3. The time span over which the extinction occurs

We could calculate the number of taxa that become extinct or the percent of the former pool of taxa that became extinct. Time could be measured in years, but often we only have segments of relative geological time units such as geological stages.1 In many parts of the fossil record, the absolute time represented by a stage is not accurately known, and different geological stages are often of great difference in temporal extent. Charles Lyell developed an ingenious technique to estimate extinction rate by charting the gradual diminution of living species as one went back in geological time. This type of analysis can give longevity and extinction rates. Such Lyellian curves demonstrate, for example, that the diminution of bivalve species on the Pacific coast of the United States is at a steady pace whereas a more precipitous extinction occurred in the Atlantic.

C. Problems in Measuring Extinction Rates

1. Taxon-Level Bias

We typically think of extinction rate as a measure of the loss of species. To create a database for paleontology, the species level is very difficult to trust with any degree of confidence; most paleontologists tend to trust the genus and higher taxonomic levels in identifications. In recent years more and more effort has been directed toward accounting for the ranges of all named species in the fossil record, but most analyses have been done at the family or genus level. The large-scale database we now employ owes its existence to the dedicated work of David Raup and especially the late Jack Sepkoski, who continuously sought to produce a more and more complete database of the geological ranges of all fossil groups. Initially, the compilation was at the level of taxonomic order but subsequent analyses have moved down the taxonomic hierarchy to the family and generic levels.

Can extinctions of higher-level taxa be used to estimate species-level extinctions? To estimate species richness using numbers of higher-level taxa (e.g., orders), we assume that taxonomic diversity at higher taxonomic levels is correlated with species richness, but

---

1 A set of stages comprises a geological series, and a set of series comprises a geological period (e.g., Cambrian, Cretaceous).
they are not necessarily correlated in this manner. This can be seen clearly where changes in ratios of one taxonomic level to another occur over broad spans of time. If the ratios change then higher taxonomic units might be flawed estimators of changes in species diversity. For example, the ratio of taxonomic orders to families decreased significantly from the Paleozoic to the Mesozoic era.

Over short periods of time the number of taxa at a higher taxonomic level (e.g., level of family) might have a regular relationship with a lower taxonomic level, such as species. In order to estimate species-level extinction from family-level extinction, David M. Raup used a rarefaction technique based on the sampling curve that relates the number of species collected at random to the number of families recovered.

The rarefaction approach is the best we have so far. Nevertheless, we must be careful in applying it. The biggest problem is the potential change in the relationship over geological time. For example, the ratio of families to species decreases by a factor of two from the Mesozoic to the Cenozoic, and other cases are known of changing ratios of species to genera. Selective extinction can also bias our conclusions. For example, certain families may be much more prone to extinction, owing to their presence in a particularly vulnerable habitat (e.g., coral reefs during a cooling event). This might overestimate total extinction, if these are added to a larger species list. It is also difficult to get sufficient data to calculate good rarefaction curves for all but the most abundant fossil groups.

2. Biased Preservation and Convergence

Estimates of extinction rates may be biased by preservation and abundance at the time of extinction. Preservation of appropriate habitats during an extinction may be greatly reduced. Thus a species might have survived, but there are no opportunities to see it because its usual facies of occurrence has not been preserved.

A common change in probability of preservation takes place when a systematic change in rock preservation occurs, as in the reduction of deposition during a regression phase of the sea, as at the end of the Permian and just before the end of the Cretaceous. Suppose the ranges of a group of species all ended at the very terminus of the Cretaceous. A gradual reduction of deposition would, by sampling error alone, give the impression of a gradual disappearance of the fossil species. Even if deposition does not decline, previously rarer species would be difficult to sample for presence during a general decline in abundance during extinction, just because we would be unlikely to find them. These biases have come to be known as the "Signor-Lipps effect," or "backward smearing," because a sharp extinction might appear to be gradual from fossil sampling. Only abundant forms would be sufficiently "findable" that we could assess their total geological range with confidence, especially up to the time of their extinction.

3. Accurate Estimate of Fossil Ranges

To estimate extinction rates, one must have an accurate accounting of the geological ranges of species. Then at any time horizon one would be able to account for the number of taxa that disappear from below to above the horizon. Preservation and rock distributions may strongly bias our perception of geological ranges of taxa. All geological ranges of fossil groups are incomplete, owing to lack of appropriate preserved habitats and poor preservation. As incompleteness (or gaps) between fossil occurrences in a vertical section increases, it stands to reason that the actual temporal range of a taxon is greater than the record would indicate. The number of gaps in preservation may also be combined with a fossil recovery potential curve, which might correct for a change in the probability of preservation during the history of the taxa in question. On the grand scale, the volume of rock correlates positively with the total number of fossil taxa recovered in both marine and terrestrial environments; this suggests that incompleteness of preservation of environments may give us a false impression of true diversity.

4. Extinction Must Be Compared with Origination Rates

If extinction occurs over a very short period of time, one can count the number of species before and after an event and assume that the decline can be explained by extinction alone. But speciation may be occurring continuously, which means that a decline in species richness may just as easily stem from a drop in speciation rate as an increase in extinction. Alternatively, speciation might keep pace with extinction, resulting in no loss of biodiversity. When the speciation and extinction rates are equal, an equilibrium exists, meaning the number of taxa remains constant. At the end of the Devonian, for example, large speciation rates balanced high extinction rates. But speciation rates collapsed during the latest Frasnian (Upper Devonian), which precipitated a severe reduction of marine species diversity. A dramatic extinction of mollusk species oc-

---

1. Regression refers to a lowering of sea level in a given area; transgression refers to a rise in sea level.
curred in the subtropics of the western north Atlantic at the end of the Pliocene (ca. 3 million years ago), but this loss was more than compensated by the origin of new species.

5. Pseudoextinction
In many cases, paleontologists have followed lineages through a geological column and have named successions of species, which are recognized by a variety of character transformations. Thus, even though a lineage may not become extinct, the morphological changes result in an arbitrary extinction or pseudoextinction. Pseudoextinctions are a significant fraction of the total disappearances of taxonomic names from one geological horizon to the next.

6. Other Biases
Paleontologists are accustomed to dealing with a wide range of problems in preservation. Occasionally, an exquisitely preserved fossil biota, such as the Middle Cambrian Burgess Shale, demonstrates that most of the remaining fossil record has not preserved a wide variety of soft-bodied species and even a number of skeletonized taxa. Such unevenness of preservation also works at smaller scales and therefore preservation strongly biases our estimates of diversity. Monographic studies of the fossil record are also uneven and descriptions of species are often strongly correlated with the intensity of study by specialists, either between fossil groups or between time horizons. Recent studies have attempted to correct for these problems by normalizing diversity estimates by the number of monographs produced for a given group at a given time.

II. MASS EXTINCTION
A. Definitions and Identification of Mass Extinction
Strong temporal changes in taxon turnover were quantified first by paleontologist Norman Newell, who found peaks of activity in the Ordovician, Carboniferous, and Jurassic. Declines in standing taxon richness were simultaneous and relatively rapid among distantly related taxa, although increases were not so obviously coordinated. Figure 1 shows five conspicuous and precipitous drops in diversity, the most dramatic occurring at the end of the Permian. David M. Raup and John J. Sepkoski, Jr., analyzed overall extinction rate of marine taxa at the family level and found that four events fell outside of a one-sided 99% confidence interval from the mean extinction rate trend: Ashgillean (Upper Ordovician), Frasnian (Late Devonian), Guadalupian-Dzhulfian (Late Permian), and Maastrichtian (Late Cretaceous). The Norian (Upper Triassic) fails this test but its widespread occurrence forces us to include it in the "big five" (Table I). A recent compilation of stratigraphic ranges of a wide variety of taxa produced by a wide variety of specialists yielded a similar overall pattern (Fig. 2). Inspection of extinction rates revealed peaks in the same times identified by Sepkoski, and, indeed, by paleontologists traditionally.

The big drops, mass extinctions, are to be distinguished from background extinction, which refers to the remainder and overwhelming majority of extinctions. While the big five are conspicuous, other mass extinctions have been recognized. About half of the marine genera disappeared in the Lower Cambrian and archaeocyathid reefs were decimated, perhaps owing to widespread marine anoxia. Another possible anoxic event caused a major extinction at the Cenomanian-Turonian boundary (Upper Cretaceous), although lowered productivity and global cooling may have contributed. Statistical tests of mass extinction do not inspire confidence, because they combine many taxonomic groups of complicated taxonomic structure, reify them to independent data points, and usually analyze them using the assumptions of parametric statistics. Because the groups are enmeshed in a phylogenetic tree structure, it is not easy perform such analyses. Distributions of extinction events are usually skewed toward many events of extinction rates of a few percent. There may be as many as 12 mass extinctions.
It may seem inappropriate to fix on mass extinctions, which could be atypical end members, but if we cannot characterize these events, will we be able to explain the smaller extinctions that were far more common in the history of life? If mass extinctions are more or less larger-scale or even global versions of what might happen on a more local scale (extinctions stemming from, e.g., local tectonism, anoxia, regional sea level change), then maybe we can extrapolate what we learn about them to smaller scales, and vice versa. If we focused on times of heightened turnover in taxon richness that are confined to basins we might see mass extinctions writ small. There is some reason, however, to believe that the big five were distinctive and the effects of extinctions during these times transcended those of more mundane times.

If an extinction event is a statistical outlier, then how could we justify a separate category, requiring perhaps a set of extinction mechanisms that differ qualitatively from background extinction? This question opens up a can of worms. There has been a good deal of debate about what a statistical outlier really is and whether mass extinctions are really different from lesser periods of extinction. If we assembled extinction rates

### Table I

<table>
<thead>
<tr>
<th>Mass extinction</th>
<th>Families</th>
<th>Genera</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed extinction</td>
<td>Calculated species-level extinction</td>
</tr>
<tr>
<td>End-Ordovician 439 Ma</td>
<td>26</td>
<td>84</td>
</tr>
<tr>
<td>Late Devonian 367 Ma</td>
<td>22</td>
<td>79</td>
</tr>
<tr>
<td>End-Permian 243 Ma</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>End-Triassic (Norian) 208 Ma</td>
<td>22</td>
<td>79</td>
</tr>
<tr>
<td>End-Cretaceous 65 Ma</td>
<td>16</td>
<td>70</td>
</tr>
</tbody>
</table>

* Extinctions/standing taxon richness x 100.

*FIGURE 2* A recent analysis of extinction rate of combined marine and fossil taxonomic families, based on recent compilations of a broad range of specialists. From Benton (1993).
into a frequency distribution curve we might argue that the mass extinctions sit squarely on the tails of some expected probability distribution, or a kill curve. The distribution of risk of marine genera consists of groups with mainly low risk, with some of much higher risk. With no mechanistic model in mind, it is not clear whether the somewhat bumpy distribution of extinction rates of marine genera is smooth or discontinuous, and we have no idea of the distribution for species. If we have to wait 100 million years for a mass extinction, are we waiting for an intense version of the same stuff or a truly distinctive event?

In the context of the time that they occur, mass extinctions are clear and major drops in taxon richness, distinct from extinction rates in the time periods before and after. The Permian appears as a sharp trough after an early Paleozoic time of expansion and then stabilization of numbers of taxa. The same can be said for the end-Cretaceous extinction. But numbers really tell only part of the story. A criterion based on a high extinction rate alone would stretch the confidence we have in our statistical assessments too far, and should only be a means of screening for candidates. To qualify, mass extinction events must have the following features:

1. The number of taxa becoming extinct is significantly greater than times of other extinctions.
2. The decline is concentrated in a small fraction of the Phanerozoic, for example, less than a geological series, or at most a few million years.
3. The extinction is broad-based taxonomically, affecting many distantly related taxa that have not arisen in the same time period as the decline.
4. The extinction affects many different biomes, perhaps not equally. For example, a mass extinction may be swamped during mass extinctions. David Jablonski found that during ordinary extinctions species with greater geographic ranges were less prone to become extinct.
5. The extinction is geographically widespread, most likely global, in extent.
6. Mass extinction may be caused by mechanisms qualitatively different from background extinctions, but it may be caused merely by “much more of the same” (e.g., anoxia, climate change, unusually large extraterrestrial impacts).
7. Mass extinction affects taxa differently than during other extinctions; small-scale influences on the degree of extinction, such as geographic range are swamped during mass extinctions. The analysis reveals a number of extinction peaks not easily seen in a plot of diversity alone. Most notable are strong extinctions in the Carboniferous, Jurassic, and mid-Cretaceous periods.
8. The recovery period following the mass extinction marks the rise of either new taxonomic groups, the expansion of formerly rare groups, or complete reorganizations of ecosystem structure.

Table I demonstrates two important issues in quantifying the degree of extinction. First, the extinction event is characterized as a loss. But the nature of loss is unclear, as it may result from declining speciation rates, increasing extinction rates, or both. In the case of mass extinctions, speciation usually declines and extinction rate increases precipitously. Table I shows the startling results: About 99% of the marine (readily fossilizable) species became extinct at the end of the Permian, and the others of the “big five” took similarly big hits. The Permian was also bad for families (ca. 50% loss), but the others hovered around 20%. The lower loss of families in many extinctions suggests that there might be survivors of many families that could recover and proliferate following the extinction.

A compilation of major changes in both terrestrial and marine fossil groups, coordinated by Michael Benton, reveals some important features of the fossil record (Fig. 2). For one thing, extinction rate can be very high, with not much overall effect on total diversity, since origins may be high or even higher. Thus if one considers extinction rate separately, the Cambrian must be added to our roster of mass extinctions. In terms of percentage extinction, it looms over the rest of the fossil record, even the Permian. There is a possibility, however, that this extinction is more apparent than real. Some have argued that poor Upper Cambrian preservation biases our perception of Cambrian diversity. The analysis reveals a number of extinction peaks not easily seen in a plot of diversity alone. Most notable are strong extinctions in the Carboniferous, Jurassic, and mid-Cretaceous periods.

The quantitative aspects of mass extinction should not obscure some of the major qualitative effects, causing irreversible changes in the world’s biota and therefore major reorganizations of the structure of life. At the end of the Permian, marine communities were reorganized completely owing to the end of dominance by brachiopods and the extinction of long-abundant forms such as trilobites. The end of the Mesozoic witnessed the demise of the long abundant carnivorous ammonites, leaving the modern world with a pitiful representation of the former glory of externally shelled cephalopods. Bivalves such as the inoceramids and rudist dominated Cretaceous shallow-water seas but they disappeared. Of course the dinosaurs also became extinct.
at this time, which presaged the evolution and diversification of the modern orders of mammals.

It is often difficult to estimate the time span over which extinction occurred, owing to strong uncertainty in the time span of geological stages, which is often the crucial level at which extinction is assessed. This is particularly a problem in the early and middle part of the Paleozoic, where time estimates of stage lengths are difficult to estimate with certainty. Within stages, rates are especially difficult because of uneven rates of sedimentation, which makes it invalid to assume a linear relationship between meters of geological section and time. For example, during a final flood stage of a transgression of the sea, sedimentation rates are often very low, and much time may be compressed in very little geological section.

The estimate of extinction is plagued by other factors as well. Worst of all, the incompleteness of the fossil record imposes biases that cloud an assessment of the tempo and degree of a mass extinction. During the Permian, an enormous regression of the sea resulted in the deposition of few marine deposits and correspondingly few marine fossils. As a result, a number of taxa appeared to become extinct, but they reappear in the Triassic, much as Lazarus was raised from the dead. Lazarus taxa may be explained by poor preservation in all facies, resulting in poor sampling of species, even those that really existed during the low point at the end of the extinction. Even if appropriate sedimentary rocks are widespread, reductions in population size may make any species more resistant to successful sampling. Alternatively, there may be localized havens in which such taxa may survive, but these refuge environments might escape preservation. In any case, extinction is overestimated. The problem is compounded by so-called Elvis species, which evolve after a mass extinction and converge by means of natural selection to resemble premass extinction morphotypes, much as Elvis impersonators now cluster the landscape, at least if you frequent Las Vegas. In Triassic reefs, sponges may be mistaken for Permian taxa, but they are unrelated.

Preservation and rock distributions may strongly bias our perception of geological ranges of taxa. A common change in probability of preservation occurs when a systematic change in rock preservation occurs, as in the reduction of deposition during a regression phase, as at the end of the Permian and just before the end of the Cretaceous. Suppose the ranges of a group of species all ended at the very terminus of the Cretaceous. A gradual reduction of deposition would, by sampling error alone, give the impression of a gradual disappearance of rarer fossil species. Thus a sudden ending of many taxa can be made to appear gradual merely by a gradual reduction of percent preservation.

Patterns of taxon survivorship of mollusks indicate that mass extinction may or may not be a qualitatively different phenomenon from background extinction. David Jablonski found that during normal periods in the Cretaceous period, extinctions of mollusks were correlated with planktotrophic larval development and geographic range, while clade survivorship was positively correlated with species richness. During the end-Cretaceous event, however, none of these held, and clade survival was correlated only with the geographic extent of the clade. After the event, the correlations found previously again obtained. In the end-Cretaceous extinction of planktonic diatoms, however, a different pattern emerged, as diatom species with benthic resting stages survived far better than those with no resting stages. Foraminifera species with specialized morphologies and larger size were eliminated and simpler morphologies were favored. Sea urchins suffered extensive extinctions across the K-T boundary, but bulk sediment processors and shallow water herbivorous species suffered more extinction than omnivores or selective deposit feeders, which suggests a relationship between high extinction and starvation. Here, properties that normally would be related to survival of individuals can be extrapolated to taxon survival. During the Permian mass extinction, gastropod success did not differ especially from periods of more subdued extinction, but groups with planktotrophic larval and geographically restricted groups suffered more than average (Erwin, 1993). Thus, as extinction intensity increases, some qualitative changes may emerge for some taxa and biogeographic/dispersal properties, but not for all.

B. Causes of Mass Extinctions

Mass extinctions are associated in time with major environmental changes. The problem, of course, is, that other times of no mass extinction also mark times of environmental change, and it is fair to say that we could not easily predict all mass extinctions with nonfossil data alone. If environmental forcing, which transcends the abilities of species to survive or adapt, is a major cause of mass extinction, what are the factors? We can list them, but finding smoking guns is often another matter.

1. Impact or a series of impacts of extraterrestrially derived objects
2. Volcanism
3. Climate change
4. Lowering of sea level, which reduces available habitats for marine species
5. Anoxia, especially transgressive spread of deep-anoxic waters onto the continental shelves

These causes stem more from associations in time between inferred geological events and extinctions, and not from a solid model linking environmental change to extinction. The best example of the latter is the Permian mass extinction. Figure 3 illustrates a scenario of environmental change that may have triggered the extinction. The vast marine regression may have been the driving force behind a variety of environmental changes, including a rise in carbon dioxide, which led to increased temperature and oceanic anoxia. At the end of the Permian, sea level dropped, perhaps about 200 m, which was followed by a transgressive rise of sea level in the Lower Triassic of similar magnitude in just 2 my. Seasonality and reduction of habitat complexity during the regression may also have begotten environmental instability, beyond the adaptive ranges of a number of specialized groups. Volcanism may be a minor contribution to climate change at the end of the Permian, because calculations preclude much of a change in the large 13C deviations at this time, owing to outgassing. Douglas Erwin likened this multicomponent explanation to Murder on the Orient Express by Agatha Christie, where twelve culprits are ultimately found to have conspired to murder the victim. Great for murder mysteries but maddening for science. Even this cast of characters ignores the hypothesis of global cooling triggered by glaciation, but this may be discounted as glacial evidence can be dated much before the extinction begins. Paleontologists Norman Newell and Anthony Hallam have implicated sea level change in a number of extinctions throughout the Mesozoic, but they too are often combined with other events, such as bolide impacts, anoxia, and temperature change.

C. The Pace of Mass Extinctions

The end of the Cretaceous is not the most dramatic mass extinction in the Phanerzoic (Figs. 1 and 2). At the time, however, both major terrestrial and marine elements were lost, the fauna was sufficiently modern to be understood ecologically, and some of our favorites,
such as dinosaurs and ammonites, hit the dust. Luis and Walter Alvarez and colleagues set off a debate that has yet to flag by suggesting that a massive asteroid impact caused the extinctions by blanketing the earth with dust spread along ballistic trajectories outside the atmosphere. Such catastrophes had been suggested before by paleontologists, but here was the first tangible evidence.

The fact of an end-cretaceous impact is supported by a worldwide anomaly of high concentrations of the element iridium in rocks just at the end-Cretaceous boundary (K-T boundary). Although there still is some controversy about this, extraordinarily high iridium concentrations indicate an extraterrestrial origin for some of the material in the rock. Shock structures on quartz crystals suggests that an enormous crater should be present on a continent. A possible piece of ejecta has been found in a core in the Pacific at the K-T boundary, which indicates that the bolide was likely a typical metal- and sulphide-rich carbonaceous chondrite rather than deriving from cometary materials.

The site of impact has probably been located in the megacrat at Chixulub in the Yucatan of Mexico. The crater harbors an armory of smoking guns, including shocked breccia clasts similar to shocked rock fragments found worldwide, tektite-like glasses, a pronounced iridium anomaly and a radiometrically estimated geological age of 65 2 Ma, which match ages of worldwide K-T boundary samples with tektites. The crater suggests a bolide of some 10 km in diameter. If the impact were at an angle, presumably more material would be sputtered into the atmosphere, but it is clear from the worldwide iridium anomaly that winds could have spread the calamity throughout the earth.

We at present can only speculate the possible biological consequences. The dust cloud would exist for a time sufficient to severely disrupt climate by cutting off all light, and temperature might have been expected to drop precipitously. A stable oxygen isotope anomaly at the boundary gives evidence for a sudden temperature change. The impact should therefore have affected all organisms dependent on light and warm temperatures. Deep-water forms not so dependent on light or warm temperature, such as nuculid bivalves, would be expected to survive. Alternatively, a hot plume emanating from the impact site could have accelerated the production of nitric acid, causing a worldwide shower of acid rain that might have poisoned the upper ocean.

Can extraterrestrial impacts be used to explain other major extinctions? The results are mixed. Positive and negative evidence for an iridium anomaly has been found for the Frasnian-Famennian (Late Devonian) mass extinction, but George McGhee and colleagues found that the extinction itself was spread over at least 7 my and climatic effects are evident. Solid evidence for impacts unfortunately postdate the Frasinian-Famennian event. There seems to be no iridium anomaly associated with the terminal Ordovician extinction. A similar iridium anomaly in sediments of 34 my of age occurs simultaneously with the disappearance of five dominant radiolarian species, and at the general time of a mammalian extinction, but the larger picture of biotic change across this boundary is gradual, with no suggestion of a catastrophe. We are therefore left with the end-Cretaceous extinction to consider.

The Alvarez theory has one strong and other weaker predictions. (a) Extinctions must follow or coincide with the impact. (b) One might also expect many groups to die off instantaneously, but a less catastrophic change in temperature and light might have a prolonged effect. (c) Finally, groups more prone to light stress or temperature increase would be more vulnerable (e.g., phytoplankton versus deep-water deposit-feeding benthos), because an impact might likely spew dust into the atmosphere, lowering world temperature. The response of the sensitive groups should be geologically instantaneous.

As in most other mass extinctions the end of the Cretaceous was preferential as to organisms affected. Groups associated strongly (Foraminifera, coccolithophorids) or weakly (ammonites) with the water column suffered the most strongly, while benthic forms (e.g., bivalve mollusks) generally suffered less. Members of food webs less dependent on plant material (marine deposit feeders, scavengers, stream inhabitants, and small invertebrate mammals) suffered less than strict herbivores. The relative success of sediment-feeding invertebrates relative to suspension feeders may be due more to their occurrence in deeper waters.

There is also an apparent thermal bias in extinction at the K-T boundary: S. M. Stanley found that mollusks and Foraminifera in the tropical Tethyan sea suffered large-scale extinction, and were replaced by higher-latitude contemporaries. The question of timing is more confusing. Coccolithophores and nonglobigerinoid Foraminifera disappeared so precipitously (and simultaneously with the iridium anomaly) that chalks give way to clastic sediments in a knife-edge contact in several sections. In the chalk of Denmark, the Maastrichtian fauna, dominated by brachiospods, disappears abruptly, with no prior warning in terms of reduced diversity or early extinction of specialized forms. The sediments above the chalk are clayey and indicative of anoxic conditions. They also have a spike of iridium.
Turbidity, loss of an appropriate sediment, and anoxia may all have contributed to the abrupt extinction. Radiolarian and haptoparian radulae but the dust during a period of flourishing radiation. The current evidence suggests that Cretaceous vertebrates also hit the dust at the boundary, but preservation is probably too spotty to tell whether it is sudden in any respect.

Unfortunately, the larger story is not nearly so simple. Some fossil groups, including land plants, inoceramid bivalves, and ammonites, experienced major extinctions or reduced speciation rates several million years before the impact occurred. The rudistids, cone-shaped bivalves that often formed Cretaceous reefs died relatively suddenly, but well before the very end of the Cretaceous. Most embarrassing for the impact theory, the freshwater biotas seem to have emerged unscathed. One would have thought that organisms such as turtles and crocodiles would be most vulnerable to a major extraterrestrial impact.

The dinosaurs represent an interesting case. The dinosaur fauna of the late Maastrichtian included fewer than 20 species in 15 genera and 10 families, chiefly in the North American western interior. There is no good evidence, however, that the dinosaurs were declining steadily toward this low number in the last 9 million years of the Cretaceous. It may well be that, while the dinosaur fauna was but a remnant at the end of the Cretaceous, their demise was nevertheless caused by the impact.

The biogeography of extinction in the end-Cretaceous provides some insight. Maastrichtian planktonic foraminifera disappeared suddenly at the K-T boundary in middle and low latitudes. At high latitudes, however, a number of groups survive unscathed into the Danian, the beginning of the Paleocene epoch. The rudistid bivalves, associated largely with tropical and subtropical waters became extinct toward the end of the Cretaceous, but otherwise there are no differences with latitude in bivalve mollusk extinctions.

The sharpness of the boundary for any group is clouded by the imperfections of preservation. If the ranges of certain groups fail to extend all the way to the K-T boundary, then it is possible that perfect preservation would have given us a far different picture. Many western Tethyan ammonites appear to become extinct below the boundary, but Charles Marshall and Peter Ward demonstrated that the confidence limits of a number of lineages allow the possibilities that poor preservation is the reason why some fossil ranges fail to continue right to the K-T boundary. Of course, this does not necessarily prove that the species became extinct at this time, they could have become extinct before or after.

The analysis does exclude being sure that the geological range should be read literally. N. MacLeod performed a similar analysis with Upper Cretaceous foraminifera and found out the ranges are compatible with a sudden extinction. They also, however, are compatible with many other possible scenarios. Some groups penetrate the boundary and it appears that there is good evidence that the fossils were not reworked up into the Danian by erosion and bioturbation.\footnote{1 The Danian is the earliest part of the Paleocene epoch, which followed the K-T boundary.}

In a mass extinction we would love to have a single cause to explain extinctions, but, like the Permian, there may be several interacting and a succession of climatic changes that caused a range of extinctions at the end of the Cretaceous.

1. In the late Maastrichtian, sea level decreased by 150 to 200 m, making a hypothesis of increased terrestrial seasonality compatible with the ultimate disappearance. Just below the K-T boundary sea level suddenly rises.
2. Temperature dropped.
3. The Deccan volcanics in India, enormous in scope, probably spewed a variety of substances that strongly affected the atmosphere and might produce effects resembling those of a bolide impact, but the activity probably predated the K-T boundary.

The evidence suggests that there may be more than one process at work in causing major extinctions. The evidence from land plants, dinosaurs, and many mollusks points to a change in conditions well before the boundary and it appears that there is good evidence that the fossils were not reworked up into the Danian by erosion and bioturbation.\footnote{1 The Danian is the earliest part of the Paleocene epoch, which followed the K-T boundary.}

Complexity also characterizes the Permian extinction. It makes sense that sea-level drop was an important
factor, operating by the effect of reduced habitable area and reduced environmental heterogeneity, but we have to reconcile the precipitous change in sea-level coverage in the last stage of the Permian, with the pattern of extinction, which was initiated earlier than the sea-level drop. Also, the extinction seemed to concentrate on certain ecological groups, particularly tropical forms at the end, but it affected high-latitude groups earlier in the Permian. Most discouraging of all, what if the extra-terrestrial influence occurred in a series of impacts, as has been argued to be possible in the Frasnian-Famen- nian (end-Devonian) extinction, rather than as one big bang? Without a series of signals (e.g., a series of definitive iridium spikes of extraterrestrial origin), such a hypothesis is very speculative.

In conclusion, there is credible evidence for the role of an extraterrestrial object in an extinction in the end-Cretaceous. While other extinction factors may have been at work in the Late Cretaceous, extinctions occurring before the time corresponding to an extraterres- trial iridium anomaly do not falsify the impact hypothe- sis. At any time in the record, some groups must be declining. Even if we accept the impact as a source of some extinction, the Cretaceous still appears to be a compound event, as environmental change and the pace of extinction both accelerate before the K-T boundary. Isn't it bad luck that a large asteroid happened to smash into the earth just as sea level was changing as much as it did in the whole of the Phanerozoic? Several other mass extinctions also seem to be complicated, and were probably associated with changes in climate and sea level that are probably interrelated. Anoxia may also be an important cause of mass extinctions, as witnessed by the extensive development of black shales in certain periods. What is lacking at present is a credible evalua- tion of the relative effects of these factors.

D. Biogeographic Aspects of Extinction

Some of the Phanerozoic changes in taxon richness can be related to a large degree to changes in the degree of provinciality. If we take the species area effect as given and constant, it is easy to calculate that, as one large province is divided in two, reductions of area are more than compensated by the increased total species richness as long as dispersal between areas is limited. Thus a temporal increase in provinciality results in an in- crease of total biodiversity. A reduction results in a decrease. Major worldwide deterioration of climate (e.g., worldwide cooling, increase of seasonality) might be an example of a time when provinciality might decrease.

The Silurian period, for example, was one of extreme cosmopolitanism, with one province of approximately 90 articulate brachiopod genera in the North Silurian Realm. In the Ludlow (Upper Silurian) two provinces can be delineated, with about 90 genera in each prov- ince. In the Devonian, Arthur Boucot found that the number of provinces increased to six; the total numbers of articulate brachiopod genera increased to about 350 on average. During the Frasnian (Middle Devonian), this provinciality decreased relatively suddenly, and ge- neric richness returned to 93.

The onset of the Permian extinction was also marked by a decrease in numbers of provinces, and the Early Triassic marks a nadir of provinciality in the Phaner- zoic. During the end of the Paleozoic, geographically restricted hynve genera succumbed before more wide- spread genera, suggesting that the overall environmen- tal change was filtering out those forms that define provinciality in the first place. Norman Newell argued that the extinction was related to the major fall in sea level. Shallow marine seas were reduced from a coverage of 40% of their possible extent in the Early Permian to less than 15% in the latest Permian and then expanded to 34% in the Early Triassic. James Valentine and Eldredge Moores speculated that reduced rates of sea-floor spreading may have been responsible for a lowering of ridge activity, depression of deep-sea bottoms, and the consequential large-scale marine regression. The sig- nificant reduction in area, coupled with continental assembly of Pangaea at the end of the Permian, may have increased extinction rates and would have homogenized the fauna due to the possible presence of more intern shelf dispersal possibilities. In contrast, the Pleistocene re- duction of area covered by the sea was far lower, and on the basis of area alone the modest marine extinctions are therefore not surprising from this point of view. Area reduction itself might not be a potent agent of extinction. Sea-level drops would hardly affect the shal- low water habitat distribution of oceanic islands, where most modern families are widely distributed. Sea-level drop may just be a correlate of another change.

The changing spatial relationships generated by con- tinental drift and sea level fluctuations must have had important influences on climate. James Valentine's the- ory of climate change generated by continental assem- bly and fragmentation attempted to relate climate and sea level to sea-floor spreading. Periods of continental assembly were envisioned as times when interior conti- nental climates were severe, affecting the continental shelf fauna. In contrast, times of fragmentation were times when the continents' climate was more moderate due to ameliorating marine conditions; this permitted...
the buildup of shallow water diversity. While the post-
Permian expansion may fit this pattern, evidence from
the Paleozoic does not seem to show an increase in
continental fragmentation during the early-mid Pale-
ozoic. Indeed, the continents were maximally fragmented
and arrayed along the equator during the Cambrian.
Continental drift and arrangement nevertheless has had
profound effects on climate and probably extinction.
During the Ordovician and Silurian periods, Gondwana
drifted southward from its Cambrian position at the
equator and came to rest on the geographic south pole.
This coincides with the Late Ordovician glacial tilities
that have been found in North Africa, and a large reduc-
tion in the degree of marine provinciality relative to
the early Ordovician. In the Cenozoic, the spatial ar-
rangements of the continents about the Pacific and At-
lantic Ocean made for a quite different climatic history.
The North Atlantic was a more enclosed basin and was
far more severely affected by the late Cenozoic polar
cooling. The Pleistocene initiated severe enough cli-
mates to cause a major molluskan extinction in the
southeastern United States Shelf, while Pacific Ameri-
can faunas showed no increased extinction.
The effects of increasing access between biogeo-
graphic realms can be illustrated by the large-scale inter-
change of mammals between North and South America
after the Pliocene establishment of the Isthmus of Pan-
amo, following the disappearance of the Bolivar Trough
marine barrier. Before the interchange there was long-
term stability in numbers of mammalian families. As a
probable result of North America’s initial higher taxon
richness, more taxa moved from north to south than in
the reverse direction. In South America, where taxon
richness now exceeded previous “steady-state” levels by
more than 50%, there was about a 70% increase in
extinction rates. Descendants of the North American
invaders participated in an evolutionary radiation, re-
sulting ultimately in an overall richness higher than
previous levels. Mammalian diversity is now higher in
South America, in contrast to the situation previous to
the exchange. This suggests that area does have an effect
on regulating diversity, but evolutionary changes can
impose a significant overprint on diversity.

E. Periodicity in Extinction, or Just Ups
and Downs?
Periodicity of extinction or climatic change predicted
by astronomical or geophysical theories would be the
most convincing way to establish a terrestrial or extra-
terrestrial cause of extinction. If extinctions are measur-
ably periodic, it may be that only one credible cyclic
theory would fit the available pattern. The precedent
for such an approach lies with the longstanding theories
of the periodicity of Pleistocene glaciations. The Yugo-
slav astronomer Milankovitch theorized that Pleisto-
cene glacial advances and retreats might be regulated
by changes in high latitude insolation, caused by cyclic
changes in the earth’s orbital eccentricity, tilt, and time
of perihelion. A power spectrum analysis of temporal
changes of abundance of Pleistocene planktonic fossils
in oceanic cores corresponded well to climate changes
estimated by stable oxygen isotopes and to periodicity
peaks predicted by the Milankovitch theory.
A number of studies in recent years have taken up
this theme and related these cycles to sedimentary cy-
cles, including some of the classic midcontinent alterna-
tions of carbonate and mudstone. Many of these cycles
occurred during times when there was no significant
amount of continental glaciation, and represent trans-
gressive-regressive cycles. For example, sedimentary
cycles in the lacustrine Early Mesozoic supergroup cor-
respond to periodicities of approximately 25,000,
44,000, 100,000, 130,000, and 400,000 years. These
periodicities, in turn, correspond to those expected
from celestial processes, such as the precession of the
equinoxes, the obliquity cycle, and the eccentricity cy-
cle. Cyclical processes such as the precession of the equi-
oxes may have driven continental heating cycles that
rearranged wind and climate.
Milankovitch climatic rhythms also appear in mid-
Cretaceous black shale sedimentary cycles. These cycles
consist of alternations of carbonate and shale, with in-
tervals of highly oxidized (red) and highly reduced
(black) strata. They are particularly interesting, as they
occur in marine sequences and must have reflected
periods of ocean bottom anoxia, alternating with vigor-
ous bottom mixing and high productivity in the water
column. On an even smaller scale, El Niño-La Niña
cycles and the North Atlantic Oscillation, whose forcing
mechanisms of periodicity are not well understood, are
known to cause cycles of benthic abundance in coastal
communities and in small bays and fjords, such as the
Swedish Gullmar Fjord.
The earth’s history has been dominated by large-
scale changes in climate, arrangement of continents,
volcanism, and sea level. Alfred G. Fischer developed
a theory connecting physical conditions with the overall
pattern of Phanerozoic life. Global sea level was rela-
tively high in both the mid-Paleozoic and Mesozoic.

4 Rises and falls of sea.
Periods of continental breakup, when dispersed and thinner continents resulted in smaller ocean basins, would be associated with higher sea levels. Periods of continental aggregation, when continental crust was bunched up due to collisions and ocean basins, were therefore more commodious, which resulted in lower stands of sea level. The temporal variation in granite emplacement matches the sea-level curve. This suggests a causal link between active continental fragmentation, volcanism, and sea level, an environmental condition of obvious importance to the world marine biota.

Fischer speculated on the presence of a causal connection between changes in terrestrial volcanism and global climate through the greenhouse effect (Fig. 4). Increased volcanism may have liberated carbon dioxide into the atmosphere. As these periods were of higher sea-level stand, erosion would have been minimal, and loss of CO₂ in weathering would be suppressed. During times of low sea level, low volcanism would reduce the liberation of CO₂, and increased weathering would consume CO₂. Thus, the mid-Paleozoic amelioration was associated with high CO₂, which, in turn, caused a greenhouse effect and an increase of surface temperature. The end of the Paleozoic witnessed the termination of such conditions, and an “icehouse effect” resulted in a deterioration of climate mainly at high latitudes. It is not clear whether these cited fluctuations are irregular temporal changes or regular oscillations.

In a time of extraordinary paleontological excitement, that, regrettably, has passed, David M. Raup and John J. Sepkoski reported a periodicity of about 26 my in the occurrence of extinction peaks of taxonomic families. Analyses of genera produced an even better periodic signal. To consider an extinction important, Raup and Sepkoski used a threshold level of 2%; minor variations on this criterion change the periodicity to an average time of as much as 30 million years between peaks, but the time between specific peaks varies substantially.

A number of celestial cycles have been suggested to explain the cyclicity. Of course, the most interesting ones are those that would cause rains of extraterrestrial objects on the earth or major changes in climate. Thus far, no theory works very well and it has also been suggested that a number of random models can explain the presence of cycles. This issue has not been settled yet, and if the cycles are real there is great hope that they can be related to an extraterrestrial source. Right now, the variation in extent of extinction and the average time between extinctions are not definitive enough to corroborate any models, which usually involve impacts of extraterrestrial objects such as asteroids or comets. An independent survey of extinctions supervised by Michael Benton failed to corroborate the presence of extinction periodicity.

III. BACKGROUND EXTINCTION AND TURNOVER

A. Normal Extinction?

The great spans of geological time between mass extinctions also witnessed significant appearances and extinctions, but at lower frequencies. The temporal pattern of such extinctions is not clear. Arthur Boucot suggested that periods of several million years are often dominated by a set of ecologically distinct species, whose coordinated extinction might be followed by the invasion or evolution of a new set of ecologically similar forms. In recent years, a number of studies have shown that turnover (including extinction) is common at this tem-
extinctions by high rates of a wide range of distantly related taxa, which are correlated with local environmental change such as basinal sea level change. Other studies however show rather high extinction and appearance rates with no punctuations in extinction. This is a field that needs to be explored with far more data collection and studies of environmental reconstruction before we can conclude anything.

At this smaller scale of extinction some of the same mechanisms as mass extinction may be in effect. Sea-level rise and fall, climatic change, and other factors can operate on a smaller scale to cause extinction. At this scale, however, biological factors may be of great importance. Biological factors in extinction might include the following:

1. Competitive displacement by an invader
2. Elimination by an overwhelming predator or herbivore
3. Spread of disease

Some of these factors may have been involved in a worldwide extinction of the so-called megafauna, a group of mammals and large flightless birds that disappeared at the end of the last glacial advance. The extinction involved large marsupial mammals and large flightless birds, the renowned Irish elk, large elephantine forms, saber-toothed marsupial, and placental cats, among others. It is possible that human hunting is the cause of these extinctions, but mobile human populations may also have brought novel diseases as they spread around the planet.

B. Declining Background Extinction

Sepkoski's extraction from the fossil database of statistically entities known as evolutionary faunas produced the fascinating result that the so-called evolutionary faunas (EF) are less and less prone to extinction as we approach the present. In the Ashgillean and Frasnian extinctions, for example, the more ancient Cambrian EF suffers more than the Paleozoic EF. In the Permian and Norian extinctions, the relatively older Paleozoic EF suffers more than the Modern EF. The successive evolutionary faunas also have progressively lower turnover (appearance plus extinction), which may make for increasing stability.

A class of distinctly lower family-level extinction rates decline during a long Paleozoic period of fairly constant taxon richness. In order to keep a steady state, a decline in extinction must be matched by an overall decline of originations, which has also been found. Why should family-level extinction and origination rates decline over geological time? There is no simple solution, although it is tempting to believe that taxa over time have evolved more and more resistance to extinction, including reduced competition with other groups. Given the vagaries of extinction and the fact that extinction is usually an overwhelming process, driven by habitat loss, widespread marine anoxia, and other factors, this idea appears to be far fetched (Fig. 5).

There may an explanation to declining extinction that is a bit more mundane. Extinction may have declined as a result of the ratio of species numbers per families, which has been increasing steadily since the Mesozoic. If a family's representation in the world biota increases in numbers of species and its consequent ecological and geographic coverage, then the probability of family-level extinction may decline.

This explanation still does not provide a satisfactory answer to the decline in originations, which also decline over the same time period. Such a decline implies a long-term reduction in the production of novelties sufficient to define taxonomic families. In other words, the rate of origin of morphological diversity has decelerated over time. Two concomitant processes may have contributed to this decline in origins of basic morphological diversity. A general filling in of resource space may have made it difficult for wholly new forms to take root later and spread. Our world may very well be the tangled bank conceived by Darwin.

To summarize, extinction in the fossil record reveals the following main points:

1. Mass extinctions caused worldwide precipitous losses of species over a wide variety of taxonomic groups and habitats.

FIGURE 5 Total numbers of marine animal family originations and extinctions per geologic stage. From Hoffman and Gillooly (1985).
2. Environmental change in mass extinctions overwhelm the ability of a species to survive by profound changes in the environment over the whole species' range.

3. The major mass extinctions eliminated a very large majority of the species on the planet and caused major reorganizations of the world's biota.

4. In many cases, extinction appears to be selective with regard to ecological characters of the groups that survive, but in others one cannot identify any traits that make one group more resistant to extinction than others.

5. A distinctly lower level of extinction can be extracted from an analysis of the fossil data. This type of extinction might be a smaller version of mass extinctions, with relatively low rates of extinction punctuated by major regional environmental change driven by sea level changes. Lower-level extinctions rates appear to decline toward the present day. The explanation for this is unclear, but it may just be an artifact of the way taxonomic groups are classified into species. More recent taxonomic families have more species and more species-level extinctions would therefore be required to see the loss of a fossil family that occurred closer to the present day.

See Also the Following Articles
BIODIVERSITY, ORIGIN OF • EXTINCTION, CAUSES OF • EXTINCTIONS, MODERN EXAMPLES OF • FOSSIL RECORD • MASS EXTINCTIONS, CONCEPT OF • MASS EXTINCTIONS, NOTABLE EXAMPLES OF

Bibliography
EXTINCTIONS, MODERN EXAMPLES OF

Gábor L. Lovel

Danish Institute of Agricultural Sciences

GLOSSARY

extinction Disappearance of the last living individual of a species. Extinction can be "local" if it concerns a definite population or location; we speak of "extinction in the wild" when the only individuals alive of a species are in captivity and of "global extinction" when no living individual remains of a species.

extinction cascade A chain of extinctions triggered by the extinction of a particular species on which many others depend. Species affected by other species are directly (parasites that live only on that species) or indirectly (predators that rely heavily on the species for food) linked with the extinct species through ecological links. Most species support other ones: a number of specialist herbivores can depend on a plant species for food or many parasites are host specific (can only parasitize one species). When these supporting species die out, the dependent species also go extinct. This can trigger a chain of extinctions, termed an "extinction cascade." For example, when the passenger pigeon died, at least two feather lice parasites followed them into extinction. When more than one species dies out at the same time (this by definition must happen when the host of an obligate parasite dies out), the term "coextinction" is also used.

first-contact extinctions A wave of extinction of species native to a continent or island, following the first arrival of humans to that area.

living dead A term coined by the American tropical biologist Daniel Janzen, denoting the last living individuals of a species destined to extinction. By definition, extinction happens when the last individual of a species dies. In reality, however, extinction of a species can be certain even earlier. Most species need both male and female to reproduce. If there are no fertile individuals of one sex, the species is doomed even if several individuals are still alive. Similarly, below a certain population size, a species cannot form a self-sustaining population, and its numbers dwindle. The decline may take many years but its course cannot be easily altered.

metapopulation A series of populations belonging to the same species that are connected via regular migration to each other's habitat patches. An important recent realization is that most species exist as meta-populations and that this is probably the original, "natural" state of all species.
pseudo-extinction Extinction of local populations is sometimes erroneously termed as "pseudo-extinction." This is misleading because global extinction proceeds through the stepping stones of extinctions of local populations. There is no fundamental distinction between the extinction of a local population and the extinction of a species other than the species becomes extinct when the last local population dies out.

proximate cause(s) of extinction The actual immediate agent(s) that cause(s) a species to become extinct.

recolonization The reappearance of a species in an area where it has earlier been present, then went extinct.

describes life span The time between the first record of a species in the fossil record to its disappearance. This time span is typically in the range of millions of years.

ultimate cause of extinction Being rare (few in numbers) and of limited distribution are precursors to extinction. The causes leading to rarity are the ultimate causes of extinction.

NO SPECIES LIVES FOREVER, AND EXTINCTION IS THE ULTIMATE FATE OF ALL LIVING SPECIES. The fossil record indicates that a recent extinction wave affecting terrestrial vertebrates was parallel with the arrival of modern humans to areas formerly uninhabited by them. These modern instances of extinction started at around 40,000 years ago. On continents, large mammals (especially those >50 kg body mass) were affected, while on islands, the impacts were mainly felt by birds. The causes of these extinctions are not well known but hunting, habitat alteration, and the introduction of nonnative species have caused extinctions. Our knowledge about extinctions is very incomplete, due to bias in research by taxonomy (vertebrate groups are better studied), geography (northern areas have received more attention), habitat (terrestrial habitats are better known than marine ones), biological reasons (certain groups do not fossilize), and methodological problems (methods of excavation and identification). Consequently, we can only crudely estimate the current rate of extinction. Even so it is evident that humans generated a new mass extinction, affecting all species in all habitats, and by the time it has run its course, it will potentially surpass the previous five mass extinction events in the history of earth. This article only deals with examples of extinction in the Quaternary period (from the final period of the last ice age, 10,000 years ago).

I. SPECIES LIFE SPANS

The life span of a species can vary widely but no species lives forever. Fossil records indicate that the average life span of an invertebrate species is about 11 million years, while mammal species live for about 1 million years (Table I). As a consequence, species existing today form only a small fraction of species that have ever lived. If we assume that the average life span of a species is 3 to 10 million years, and multicellular organisms have been on earth for a period of 600 million years, the plant and animal species currently living are not more than 1 to 2% of all those that have ever lived. For marine invertebrates, an estimated 93% of the species that had ever existed are today extinct. Extinction is thus the natural fate of all living species.

Extinction can and does happen at any time, and one can say that extinction is occurring continuously. Most of these extinctions are of local populations. For many species, a landscape contains several suitable habitat patches but not all patches are occupied at all times. Species constantly recolonize unoccupied patches and go extinct in others. The local populations in these patches form a kind of network called a "metapopulation." There is constant migration among the habitat patches, some of them (source patches) producing surplus individuals that colonize other patches; others are not so productive (sink patches). When a metapopulation cannot produce enough individuals to compensate for mortality, the species becomes regionally extinct.

### Table I

<table>
<thead>
<tr>
<th>Group</th>
<th>Estimated life span, million years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dinoflagellates</td>
<td>13</td>
</tr>
<tr>
<td>All invertebrates</td>
<td>11</td>
</tr>
<tr>
<td>Cenozoic bivalves</td>
<td>10</td>
</tr>
<tr>
<td>Dinosaurs</td>
<td>8</td>
</tr>
<tr>
<td>Planktonic foraminifera</td>
<td>7</td>
</tr>
<tr>
<td>Echinoderms</td>
<td>6</td>
</tr>
<tr>
<td>Marine invertebrates</td>
<td>5–10</td>
</tr>
<tr>
<td>Marine animals</td>
<td>4–5</td>
</tr>
<tr>
<td>Cenozoic mammals</td>
<td>1–2</td>
</tr>
<tr>
<td>Mammals</td>
<td>1</td>
</tr>
<tr>
<td>All fossil groups</td>
<td>0.5–5</td>
</tr>
</tbody>
</table>

From May et al. (1993).
This results in a range contraction as regular migration between metapopulations does not occur.

Species may become globally rare and subsequently go extinct at any time. The “background rate of extinction” is estimated to be in the magnitude of 1 to 10 species/year through the geological periods. However, significant extinctions in earth’s history occurred in clusters. During the last 300 million years, there were five such “mass extinctions,” wiping out large proportions of the then-living species. The fossil record related to these has been intensively studied and hotly debated, but without producing an accepted interpretation about the causes of these mass extinctions. Extinctions affecting a more restricted group of species have also occurred on a smaller scale. Within those groups the extinctions were significant.

The most recent of such events commenced during the late Quaternary, about 100,000 years before present (yBP), and started to intensify about 40,000 yBP. Since then, on different continents and also on islands, at different times, several hundreds of land vertebrates, mostly large species (>50 kg body mass) have gone extinct. This extinction wave has not yet ended.

II. EXAMPLES, POSTGLACIAL

A. North American Extinctions

In late glacial North America (called the Wisconsin glaciation period, ending about 10,000 years ago), 71% of midlatitude mammal genera were lost, while in Alaska, the same loss was 56%. This is the opposite that would be expected from environmental conditions—we expect that if climate is the cause of these extinctions, more northerly species would be more severely affected. According to their trophic position, 71% of the herbivores, 67% of the bears, and 50% of the dogs and cats became extinct. Many of these have lived through cycles of glacial and interglacial periods, and extinction was not biased toward either older or newer genera. Environmental changes are therefore thought unlikely to have caused these extinctions. On the contrary, general conditions were at their worst during the period preceding the extinctions, 20,000 to 18,000 yBP. Conditions for large mammals have improved afterward, notably between 18,000 and 7000 yBP, when most extinctions occurred.

The postglacial extinctions are generally connected to the appearance of humans in the regions affected. The North American continent has suffered numerous avian extinctions during the end of the last glacial: 19 genera of birds became extinct during this period. In spite of taxonomic problems as well as scarce records (ten of these birds are only known from the single area of the Rancho La Brea tar pits in metropolitan Los Angeles in the United States), we can generalize that most of these extinct birds were large to very large by avian standards, and the loss of most or all of them can be attributed to ecological dependency on large mammals that also went extinct during the same time. The largest group of these extinct birds were raptors: condors, eagles, accipitriform vultures, and caracaras. Extant hunting birds, including eagles, feel on carrion as well as live prey that they themselves captured, so it is safe to assume the same way of life for these birds. The disappearance of large mammals must have resulted in a significant reduction of the available food base. As a consequence, many of them became extinct. A similar extinction cascade can be observed in the only remaining continent with diverse large ungulate fauna, Africa: where game becomes scarce or disappears, vultures and eagles also disappear. This points to the ecological plausibility of this hypothesis. Two other birds, Panandris and Pycrorhamphus, related to the North American icterids (Icteridae) of today, are thought to have been in a commensalist association with large herbivores—the “cowbirds” of the Pleistocene—and followed their hosts into extinctions. In Africa, there are several further groups of songbirds associated with large mammals, such as oxpeckers and drongos. It is likely that a variety of commensalist relationships also existed in the New World, and these must have been lost with the disappearance of most large mammals in North America.

B. Australian Extinctions

Australia, until the end of the last glacial, had a fauna of monotremes and marsupials that was as diverse as the placental faunas of other continents. In contrast to those, however, the Australian fauna was rich at the species if not the genus level, and it was not subjected to any significant intercontinental faunal exchange. This lead to a homogeneous fauna that seems to have been unable to withstand ecological stress. During the late Pleistocene, many species went extinct. This loss was comparable, in numbers of species, to extinctions on other continents. For example, while there existed only 15 genera versus the 32 in North America, the number of extinct species is about 60 in Australia and 51 in North America. All 19 species of marsupials heavier than 100 kg, and 22 of the 38 species that are 10 to 100 kg have become extinct. Three reptiles and the
ostrich-sized Genyornis newtoni have met the same fate.

A few of the extinct animals are depicted on Figure 1. The largest reptile was the varanid lizard Megalania prisca, which, at 7 m long, was probably a top predator. Among the extinct monotremes were large echidnas, such as Zaglossus hacketti, which was 1 m in length and 0.5 m in height. This seems to have been a proportionately large version of the small living echidna. Among the marsupials, there were large carnivores: a large morph of the tiger cat (present on one island until European contact), or the leopard-sized Thylacoleo carnifex, a marsupial lion named “giant killer possum.” One species of the koalas, Phascolarctos stirtoni, which was about 30% larger than the living koala, also survived until the very late Pleistocene. The living koala is the only survivor of a diverse family that had its peak in the Tertiary.

The large and varied superfamily of Diprodontoidea has totally disappeared during the late Pleistocene, completing a longer sequence of decline. By the last glacial, only two families were represented. Some species of the Palorchestidae were beasts that resembled a giant kangaroo but had a tapir-like trunk and huge, curved claws. The cowlike Zygomaturus tribohicus had a 2 m long body, a huge, broad head, and a narrow, upturned snout. Judging from the frequent fossil remains, it was widely distributed in coastal and mountain Australia. Another browser was the large, slow Diprotodon optatum that had a feeding apparatus suggesting that it was browsing tough, succulents and shrubs. The kangaroos (family Macropodidae) today are the largest group of marsupials still living in Australia, although their diversity, too, was seriously reduced by the beginning of the Holocene. From the Macropus genus itself, at least eight species died out. Some of these were small, like today’s wallabies, but Macropus titan and M. ferragus were real...
individuals in their prime age were found as well as relative, these are about 25% or more smaller.

Although Australia has been inhabited by humans since at least 30,000 yBP, the extinctions were generally not believed to be linked to their presence. Recent evidence based on more exact radiocarbon dating of bird and egg remains casts doubt on this belief, and human predation seems the probable cause also in Australian extinctions.

C. African Extinctions

When examples of modern extinctions are discussed, Africa is often ignored. Some have called Africa "the living Pleistocene" because this is the only continent where a diverse and abundant fauna remained that bears any resemblance to the Ice Age. However, Africa has had its postglacial extinctions and as this is the continent with the longest period of human occupation, the analysis of these is potentially very important.

The most impressive examples of such extinctions come from northern Africa. This region, after the dry, hyperarid period between 40,000 and 12,000 yBP, experienced a moist period during which the fauna included, among other species, the African elephant, white rhinoceros, a zebra, warthog, giraffe, blue wildebeest, hartebeest, eland, roan antelope, and a species of red-buck. Between 5000 and 4000 yBP, this moist period changed again, and this fauna disappeared from most of the Sahara but survived in the Maghreb area of North Africa. Some species were certainly lost, though: the appearance of the long-horned buffalo, giant hartebeest, eland, roan antelope, and a species of red-buck. Between 5000 and 4000 yBP, this moist period changed again, and this fauna disappeared from most of the Sahara but survived in the Maghreb area of North Africa. Some species were certainly lost, though: the Atlantic gazelle (Gazella atlantica), Thomas' camel (Camelus thomasi), the giant North African deer (Megaloceros giganteus), and the long-horned North African buffalo (Pelorus antiquus).

A similar cycle can be observed in southern Africa, although well dated records are missing from most of this region. In the Cape zone, however, the total disappearance of the long-horned buffalo, giant hartebeest (Megalotragus priscus), the giant Cape horse (Equus capensis), and the southern springbok (Antidorcas australis) happened around 12,000 to 9000 yBP. Their extinction in southern Africa was, at least partially, related to climate-driven environmental change.

During this period, however, there was a dramatic change in artifacts throughout the continent, indicating a very significant shift in human cultures, and with this, probably of hunting techniques and efficiency. Analyses of archaeological sites support the hypothesis of increased hunting proficiency: more remains of individuals in their prime age were found as well as relatively more bones of "dangerous game." In North Africa, the appearance of domesticated animals may also have contributed to the decline.

The pace and extent of late Pleistocene and Holocene extinctions in Africa parallel that of Eurasia, where there was no sudden and massive extinction wave such as in North America. The only significant difference between these continents and the Americas is the length of human occupation. After Africa, Eurasia has the longest period of human presence, about 700,000 y. North America did not have a previous history of human habitation, and the pattern of extinctions is entirely different. This continent was swept by waves of extinction during the late Pleistocene/early Holocene, coinciding with the migration through the Bering Strait and then southward of anatomically modern humans. During this period, a very large proportion of the extant fauna disappeared in what is geologically and evolutionarily a very short period of time.

III. EXAMPLES, FIRST-CONTACT EXTINCTIONS

On small islands all over the world, many species, especially birds, went extinct during the past 10,000 years (note that islands have an impoverished mammal fauna to start with, due to dispersal problems). The other common feature of these extinctions was that there was no taxonomic replacement of lost species. These extinctions are so tightly correlated with the arrival of humans that they were termed "first contact extinctions" (FCEs). In the Americas, FCEs occurred between 12,000 and 10,500 yBP, on the West Indies between 7000 and 5300 yBP, and on Madagascar between 2000 and 500 yBP (MacPhee and Marx, 1997).

These FCEs can take as little as 1 year on small islands and up to 1300 years on large islands and continents. On the Commander Islands, east of the Kamchatka Peninsula in the northern Pacific Ocean, humans arrived in 1741. Steller's sea cow (Hydrodamalis gigas) was extinct by 1768. On the Mascarenes Islands, humans arrived A.D. 1600, and the major extinctions terminated around 1900. In New Zealand, the first human colonists arrived at around A.D. 1200, and the first major episode of extinctions terminated by A.D. 1300. In the Mediterranean, humans colonized all the major islands between 10,000 and 6000 yBP, and this also coincides with the extinctions of several endemic species, such as pigmy elephants, rhinoceroses, and hippopotamuses.
A. Madagascar

The extinction of large mammals and birds on the island of Madagascar during the Holocene was of similar magnitude than the earlier, Quaternary extinctions in North America, Australia, and New Zealand. Today’s Madagascar fauna is a pale shadow of a once-diverse assemblage of spectacular species, a “magnificent bestiary” (R. E. Dewar). Humans colonized Madagascar only in historical times. The earliest dated archaeological site is from about 300 A.D. The extinctions started to happen not much later. It is generally agreed that these extinctions were caused by human activities, opinion only differs in what type of activity this was.

Seven of the 17 primate genera have disappeared completely, and two more lost its largest species. The extinct lemurs were all large, and probably diurnal. The largest of these, Megaladapis edwardsi, had males with a body mass between 50 and 100 kg. Members of the smallest extinct genus (Mesopropithecus) was about as large as the largest living species, the indri (Indri indri). Several of these species had ways of locomotion that are unknown among today’s living primates: walking on the ground on four legs (Hadropithecus), arm-swinging (Palaeopropithecus), and vertical climbing similar to that of the koala bear (Megadapis).

The other group that was severely affected is the large, flightless birds, ratites, commonly known as elephant birds. They are classified into two genera and 6 to 12 species. The largest of them (Aepyornis maximus) had a height of nearly 3 m and resembled a massive ostrich. The smallest, Mullerornis betsilei, was about half this size. These species are thought to have been terrestrial grazers—this ecological group is otherwise only represented by the pigmy hippo (Hippopotamus lemerlei). This species, together with a large viverrid (Cryptoprocta spelea) and an endemic aardvark (Plesioryctops madagascariensis), is also extinct. C. spelea was the largest known carnivore in Madagascar and resembled a short-legged puma so much that earlier it was classified into the cat family (Felidae). The only reptiles that went extinct were giant land tortoises. The two species had carapace lengths of 80 cm and 120 cm, respectively, and were important consumers of ground vegetation.

B. The Pacific Islands

Humans have gradually colonized the world, and have relatively recently arrived to several oceanic islands. The colonization history of the Pacific Ocean islands is relatively well studied. The human expansion across the Pacific, starting from Southeast Asia, was accompa-
were inhabiting forests, not grasslands. No species of the moas are left, and this is one of the best-known examples of large-scale, human-caused extinctions. Polynesians have successfully colonized New Zealand about 1000 to 800 yBP and although we have no reliable record of moa densities before or after this period, nor do we know about moa evolutionary history in earlier ages, it is well documented that the cause of their demise was that the Maori have intensively hunted all species. Archaeological sites with large amount of moa bones are found all over New Zealand, some covering many hectares. The most detailed research on them was conducted on the eastern side of the South Island, and these convincingly demonstrate that man was a voracious hunter of moa: their nests were robbed, and their carcasses were probably utilized in a wasteful way. Dogs and rats introduced by man have probably also played a role in the extermination of moa, especially the smaller species. Moa hunting became intensive about one century after the arrival of Maori, coinciding with a rapid growth of the human population. Within a few centuries, hunting and forest burning accelerated the decline so that by about 400 yBP moas had become too scarce that they were no longer systematically hunted. Continued habitat destruction, sporadic hunting, and probably predation by feral dogs continued to destroy birds, and none were left by the time of European settlement.

The extinct bird species that have never been seen by Europeans include not only the moas but about 20 other bird species. These were often flightless (79% of all extinct species), ground nesting (99%), diurnal (96%), and larger than the closest surviving relative (71%). Fifteen of these were endemic to New Zealand, and five were very similar to living Australian relatives. No less than four of the fifteen were rails, thus echoing the extinction patterns of the Pacific islands (see earlier). Other birds lost include a flightless goose (Cnemiornis calciatrum), a giant rail (Aptornis otidiformis), a swan (Cygnus sumnerensis), and several flying birds. A coexistence with the moa was the extinction of the giant eagle, Harpagornis moorei, that was the largest known flying bird, probably preying on moa. After the extinction of its prey, or possibly even earlier, when the prey became rare, the predator disappeared.

IV. EXAMPLES, HISTORICAL EXTINCTIONS (1600–)

Since A.D 1500, during the “modern era,” extinctions were closely correlated with the European expansion, starting with the discovery of America in 1492. The time span of resulting extinctions differ by species and the area affected, but it gradually expanded to include all areas and habitats of the earth.

A. New Zealand

During the European period of occupation in New Zealand (although “discovered” by the Dutch seafarer Abel Tasman in 1642, colonization of New Zealand did not start until about 1840), at least five further bird species have become extinct. There is no doubt that the environmental changes brought by Europeans in about 200 years exceeds those caused by the Polynesian occupants during the preceding centuries. This difference, however, is not due to intent but due to the difference in technology. The impact of the initial colonization in terms of extinctions is larger and more obvious because the Polynesians arrived to predator-free islands.

One of the recently exterminated species is the Stephens Island wren (Xenicus lyalli), the only known flightless songbird. Stephens Island is a small island in the Cook Straight, between the North and the South Islands of New Zealand. The first specimen of this bird was brought to the lightkeeper’s house by his cat. Described as a new species to science, it was exterminated by the same cat within one year (1894). No person has ever seen a live specimen.

The catastrophic impact of predator invasion is exemplified by another New Zealand story, the rat invasion of Big South Cape Island. Big South Cape Island lies south of the South Island, and was known to harbour several endangered species when in 1964, ship rats (Rattus rattus) got on shore from a shipwreck. In two years’ time, the rats reached very high densities, and four species of birds endemic to New Zealand, one native bat species (greater short-tailed bat, Mystisciua tuberculata robusta), and numerous invertebrates became extinct. Other species were removed from the island, and thus, for example, the South Island saddleback (Philesturnus carunculatus), a thrush-sized bird, survived.

Many more species of birds, reptiles, amphibians, sea mammals, and invertebrates have also suffered a reduction of their former range. Typically, they became extinct on the main islands, surviving only on offshore ones, that were frequently but accidentally free of introduced mammals. For example, the tuatara, Sphenodon punctatus (with its sister species S. guntheri), the only living relative of the dinosaurs, has been found in early archaeological sites on the main islands. Today it only survives on a few offshore islands. It did not survive
on islands where Polynesian rats (Rattus exulans) are present, but can be common on rat-free islands. Another example of on-islands-only species is the little spotted kiwi (Apteryx oweni), which had only one self-sustaining population on Kapiti Island near Wellington, and the recently discovered, undescribed tusked weta (a relative of grasshoppers).

B. Hawaiian Islands

The Hawaiian Islands are a group of volcanic islands, in distant isolation from any other land mass, in the middle of the Pacific Ocean. They were reached by Polynesian settlers at around 500 A.D. These islands have had a very diverse and unique fauna and flora, and as elsewhere, especially the vertebrates were seriously decimated. The best documented examples are again the birds. The extinct species include flightless geese, ibises, rails, a long-legged owl, a sea eagle, honeycreepers, and crows. Further, there is a group of species that have living populations on one island or another, but not on the one where they were found as subfossils.

The patterns of extinction are strikingly similar to New Zealand, except that there are no large numbers of songbirds reported from New Zealand. Man-induced changes in Hawaii may have been more extreme, or New Zealand originally did not have many songbirds.

The prehistorically extinct birds of the Hawaiian Islands include 1 species of petrel, at least 10, mostly flightless species of geese, 3 flightless ibises, 8 of rails, 3 of long-legged owls. 1 Accipiter, 2 large crows, 1 large mephagid, and 15 species of Hawaiian honeycreepers, relatives of finches.

In the early 1980s, 82 endemic bird species were known from the Hawaiian Islands. Fifty-three of these became extinct prehistorically (before 1778 when Captain Cook discovered the islands). Area and elevation show significant positive correlation with the number of fossil and historically recorded bird species. On Molokai, the smallest of the 5 largest islands, with 676 km sq. of area and 1315 m a.s.l., there are 21 fossil and 9 historically known species. On Hawaii, the largest and highest (10646 km sq., 4206 m a.s.l.), 3 fossil and 23 historic species are known—although more fossil species are expected after more excavations are done.

Significant paleoecological findings are accumulating and it is difficult to draw a reliable and comprehensive picture about the original fauna of the Hawaiian Islands as well as a proper assessment of the extent and nature of extinctions. However, what we know now indicates that the effect of the human as exterminator, direct or indirect, of the fauna of this island archipelago is much more significant than earlier thought. Authori-

C. Extinction Paradoxes

Interestingly, current extinction rates seem to be lowest in areas with a long history of human habitation. Plant extinction rates in areas with Mediterranean climate are low, ranging from 1% of all species in Western Australia to 0.1% in the Mediterranean itself. Current threats to plants are one order magnitude larger: 10.2 to 13.2% of species are considered threatened. The suspected cause of the current low extinction rate is a "recording error": many of the extinctions occurring in the "prebotanical age." Indeed, the current extinction rates are lowest where agricultural cultivation has been the longest, 8000 to 6000 yBP. This is consistent with the view that most vulnerable species will have been lost by the time when botanical investigations started.

Similarly, the proportion of bird fauna extinct in the Pacific islands is inversely proportional to the length of human habitation of these islands: 80% of Hawaii's bird fauna is recently extinct or endangered against 10% in Vanuatu. Hawaii has been inhabited for about 1300 years and Vanuatu for 4000 years. Pimm and coworkers (in Lawton and May 1995) argue that the sensitive species have been eliminated by first colonists before record keeping began, and thus we have no direct evidence of first-contact extinctions in the Pacific.

V. EXAMPLES, EXTINCTION CASCADES

A species' "ecological environment" almost always includes other organisms that are essential for the species' survival. Species are connected through trophic links—they eat each other. Other vital ecological links include pollination, dispersal of seeds, and providing habitat. For example, bees, birds, and bats pollinate flowers, birds, and mammals disperse seeds, and trees provide nesting holes for birds. The extinction of a species can have reverberating consequences, affecting other species that are, directly (such as obligate parasites) or indirectly (such as shared predators), linked with the extinct species through such ecological links. Most species support other ones: a number of specialist herbivores can depend on a plant species for food, or many parasites are host specific (can only parasitize one species). When these supporting species die out, the depen-
dent species also go extinct. This can trigger a chain of extinctions, termed an “extinction cascade.”

With the death of the last passenger pigeon, a female named Martha in the zoo in Cincinnati, Ohio, in 1914, at least two species of feather lice, which were obligate parasites of this species, must also have perished, although there is no mention of this in any list of extinct species.

All moa, a group of 12 species of ratites of different size, went extinct not long after Polynesians settled in New Zealand. The largest known raptor, the giant eagle (Harpagornis moorei) also followed them into extinction. As there are large middens with thousands of moa bones at several sites in New Zealand but these do not contain bones of the eagle, it is thought that the eagle was not a victim of persecution or hunting, but became extinct after its food base, the formerly very common moa disappeared.

Likewise, several bird species that went extinct in North America at the end of the last ice age are suspected to have died out in an extinction cascade (see earlier).

VI. PROBLEMS IN OUR UNDERSTANDING OF EXTINCTIONS

We are aware that our knowledge of the actual extent of even recent extinctions is very fragmented and incomplete. There is, in other words, a huge difference between documented and real extinctions. This is due to a series of reasons. Some of these can be overcome with the development of science, but several of them results from the organisms’ biology.

A. How Many Species Are There?

We do not know, even to an order of magnitude, how many species we share the earth with. Estimates of global species richness range from 3 to 80 million species. This has an obvious consequence for the estimation of extinction rates: 1000 species is a different relative share of a global total of 3 versus 80 million species.

B. Record Keeping Is Insufficient/Inadequate

The documentation of extinction is also uneven, both geographically (mostly from islands and northern temperate region) and taxonomically (higher organisms better reported). Since 1600, only 485 animal and 584 plant species are listed as extinct (Table II). We strongly suspect that even among vertebrate groups, documented extinctions are serious underestimates. For example, on the Solomon Islands, where 164 bird species have been recorded, 12 have not been seen this century.

<table>
<thead>
<tr>
<th>Total number of species</th>
<th>Extinct since 1600</th>
<th>Listed as Threatened</th>
<th>Described (thousands)</th>
<th>% Extinct</th>
<th>% Threatened</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molluscs</td>
<td>101</td>
<td>394</td>
<td>100</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>Crustaceans</td>
<td>4</td>
<td>126</td>
<td>40</td>
<td>0.01</td>
<td>0.3</td>
</tr>
<tr>
<td>Insects</td>
<td>81</td>
<td>873</td>
<td>4000</td>
<td>0.006</td>
<td>0.09</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>229</td>
<td>2,212</td>
<td>47</td>
<td>0.3</td>
<td>3</td>
</tr>
<tr>
<td>Fishes</td>
<td>20</td>
<td>452</td>
<td>24</td>
<td>0.1</td>
<td>2</td>
</tr>
<tr>
<td>Amphibians</td>
<td>2</td>
<td>39</td>
<td>3</td>
<td>0.1</td>
<td>2</td>
</tr>
<tr>
<td>Reptiles</td>
<td>23</td>
<td>187</td>
<td>6</td>
<td>0.4</td>
<td>3</td>
</tr>
<tr>
<td>Birds</td>
<td>116</td>
<td>1,229</td>
<td>95</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Mammals</td>
<td>99</td>
<td>309</td>
<td>43</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>465</td>
<td>3,363</td>
<td>1,800</td>
<td>0.04</td>
<td>0.3</td>
</tr>
<tr>
<td>Plants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gymnosperms</td>
<td>2</td>
<td>242</td>
<td>0.8</td>
<td>0.3</td>
<td>30</td>
</tr>
<tr>
<td>Dicotyledons</td>
<td>120</td>
<td>17,474</td>
<td>190</td>
<td>0.006</td>
<td>9</td>
</tr>
<tr>
<td>Monocotyledons</td>
<td>462</td>
<td>4,421</td>
<td>52</td>
<td>0.9</td>
<td>9</td>
</tr>
<tr>
<td>Palms</td>
<td>4</td>
<td>923</td>
<td>28</td>
<td>0.1</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td>364</td>
<td>22,137</td>
<td>240</td>
<td>0.2</td>
<td>9</td>
</tr>
</tbody>
</table>

but only 1 is listed as extinct. In Malaysia, a 4-year search for 266 freshwater fish species reported in the last century found only 122, yet few are recorded as extinct.

The current method of documenting extinct species is not entirely biologically valid. A species becomes officially extinct with the death of the last living individual. A species may be destined to extinction long before this happens. If mortality surpasses reproductive success, a species may get onto an “extinction trajectory”—numbers will continuously decrease without reversal, but it will take many years or decades until all individuals perish. Likewise, if there are no reproductively successful pairs remaining, the species has no hope of surviving, even though not all individuals are dead yet. These species are termed the “living dead.”

Sometimes species can go through a “genetic bottleneck” when populations become so small that genetic variability practically disappears. Two such cases are known: the cheetah (*Acinonyx jubatus*), a fast-running predator in the cat family, must have gone through such a population crisis some thousand years ago. Today all living cheetahs are genetically virtually identical. The black robin (*Petroica traversi*), a small, endemic songbird on the Chatham Islands of New Zealand, had only one fertile female in the 1980s. In one of the success stories of today’s conservation, this species was brought back from the brink of extinction—but genetic variability of the species is much reduced. Without human intervention, this species would have become extinct.

We also have to consider that so many species are recorded from only one location (for example, up to 40% of beetle species described in the Natural History Museum collection, London, United Kingdom) that it is difficult to assess anything but local extinctions.

### C. Uneven Recording Effort

Documented insect extinctions are 100 times less than among vertebrate fauna. The difference is even larger if we consider that the number of insects is certainly much larger than we know today. There are 100 times more vertebrate than invertebrate taxonomists, and they are also 10 times more than the number of taxonomists of flowering plants. This uneven attention by humankind to different groups is also evident from the rate of describing new species. This uneven attention by humankind to different groups is also evident from the rate of describing new species. This is only 0.03 to 0.05% new species/year for birds. In tropical areas, 1 of every 100 plant specimens is new for science, for insects, fungi, and marine macrofauna, this can reach 20 to 80%.

### D. Taxonomic and Habitat Bias

Our knowledge is also very biased by habitats and taxonomic relationships. We know much more about forests than seas, and while most of the mammal or bird species of the world are known, this cannot be said of other important and species-rich groups of organisms like fungi, nematodes, or arthropods.

The great taxonomic bias in our records is well exemplified in the 61 extinct insect species: 33 of these are butterflies and moths. These groups do not constitute more than half of all insect species (they are more likely to be about 25% maximum), their prevalence merely reflects that they are much better studied than other insect groups. It is perhaps real that 51 of these are island species but not that 42 of them are from Hawaii. Similarly, of the 10 continental extinct species, 9 are from North America. This indicates the distribution of researchers, not the real distribution of threatened or extinct species.

### E. Geographical Bias

Our knowledge is particularly scant in areas of the earth with the highest biological diversity, the tropics, and thus any changes are much better documented in the northern temperate regions where only a minority of the global biodiversity can be found.

The geographical bias and variability reported in the literature include patterns that are real, while others are imaginary. Fifty-one percent of all recorded animal extinctions are from islands—this is probably a real pattern. The numerical preponderance of the Pacific Islands is due to both their large numbers and recent human colonization. However, in the pattern that two-thirds of recent animal extinctions are from North America and the Caribbean, 20% from Australia is certainly an artifact. Similarly, all 43 plant extinctions in Africa are from the Cape flora, and two-thirds of continental plant extinctions are from North America and Australia. The rarity of such records from South America, Asia, and Africa is surely an artifact.

### F. Methodological Obstacles

Some of these artifacts are historical and irreparable. Further inaccuracies result from the fact that much of extinction information is gathered by paleozoology. We have never witnessed these extinctions, and only remains of these extinct organisms are found. There are special difficulties in studying and interpreting fossil or subfossil material. Just to mention one, the screens...
used for sieving soil when excavating animal bones have been, until recently, too coarse to retain bones of small bird species. As a natural consequence, our knowledge of the true extent of bird extinction is grossly biased by this because there are many more small than large species of birds (just as in other groups of organisms).

G. Inherent, Biological Problems

Most of our fossils are from marine organisms, because they often have calcareous body parts that fossilize well. Fossilization on dry land is different: some groups and some climatic regions (insects in tropical climates) are simply not amenable to fossilization.

Birds are better known than other organisms because their skeletons fossilize better and their taxonomy, generally, is better known. In contrast, the original vegetation of the Hawaiian Island lowlands is a matter of conjecture as they were largely destroyed before botanists arrived to collect there. Entomologists can only speculate what the effect of this deforestation could have been for the arthropods as very few insects are preserved under Pacific island conditions.

VII. THEORETICAL ASPECTS OF EXTINCTION

The first step in the extinction process for a species is to become rare. It is conceptually useful to distinguish between ultimate causes of extinction (what causes species to be rare and thus vulnerable to extinction in the first case) and proximate causes of extinction (what is the actual cause of extinction). These latter generally include demographic and environmental stochasticity (random, large fluctuations in density and environmental conditions), genetic deterioration, and social dysfunction, although their respective importance is not well understood. Ultimate causes include hunting, habitat destruction, invasion by introduced species, and pollution.

Two general tendencies are relevant for the study of extinctions:

1. Species that are widespread tend to be abundant as well, but the causes of this positive correlation are not well understood. This also means that species most at risk from extinction (those that are sensitive to proximate causes) have small geographic ranges, because they will also be locally rare. This double jeopardy may be serious when populations and ranges are artificially reduced by ultimate causes of extinction.

2. The distribution of tropical species is generally more restricted than that of temperate species. Smaller ranges have been documented for tropical than temperate-region trees, mollusks, crustacea (crabs and relatives), fish, amphibians, reptiles, and mammals. A related trend is that average population densities of individual species increase from the equator toward the poles (proven for invertebrates, mammals, and birds). This fits with the first trend listed and is also consistent with a decline in range sizes toward the tropics.

As a consequence, disproportionately more tropical than temperate species are threatened with extinction. Of the 1029 threatened bird species, 442 live in tropical forests, more than twice the number of species living in wetlands, the next most threatened habitat category.

Most (direct as well as circumstantial) evidence indicates that most of the recent extinctions were caused by humans. Climate change has been invoked in some cases but the evidence for this is not strong. The actual form of human impact can be overhunting, habitat destruction, or introduction (voluntary or accidental) of nonnative species, mostly of predators (cats, dogs, rats) or browsing herbivores (pigs, goats, sheep). It was also suggested that humans have spread an extremely virulent pathogen, causing a "hyperdisease." In mollusks, birds, and mammals that went extinct since 1600 and have a known cause, 23% was due to hunting, 36% to habitat destruction, and 39% due to the introduction of exotic organisms. Once again, our knowledge is rather sketchy: in mammals that became extinct since 1600, only 30% have an established proximate cause of extinction.

A. Introductions as a Threat to Species

Introduced species have often been implicated in the extinction of native species. Many introduced species, however, have had no detectable effect on species in their new environments. However, the massive spread of organisms by humans to other areas of the globe may increase local diversity, but will result in large losses in global biodiversity. In order to understand the danger that pan-mixing of the earth's fauna and flora signify, let us consider a thought experiment in island biogeography. Species richness of an island is largely determined by its area; the larger the area, the more species the island contains. The same applies for continents. For example, mammal species richness is related to the size of the individual continents. The resulting correlation allows to extrapolate the global species rich-
ness. A supercontinent, with an area equal to the total dry land on earth would support about 2000 mammal species. Currently, there are about 4200 mammal species. Therefore geographical isolation allowed evolution to generate nearly twice the biodiversity that could otherwise, on the basis of habitat area alone, be expected. As today human-assisted invasion is becoming a more and more prevalent biogeographic phenomenon, it is inevitable that more extinctions are predicted, with possibly catastrophic consequences for biodiversity.

Often there are more than one cause of extinctions. For example, the kokako (Callaeas cinerea), an endemic wattlebird in New Zealand, became extinct in most of its former distribution range (and is on the brink of global extinction) due to a combination of factors. These include the contraction and fragmentation of its original forest habitat plus the effects of introduced predators, mainly the European stoat (Mustela erminea) and the Australian brush-tail possum (Trichosurus vulpecula).

B. Insights from Population Dynamics

A further difficulty to understand extinction is that the actual process of extinction is also very imperfectly documented. Only a few documented examples exist that link population decline to changes in species distributions. The stepping stones of global extinction are local extinctions, so it is logical to assume that as a species becomes more restricted and rare, its distribution range will become fragmented and gradually smaller. The European fir tree (Abies sp.) decline was indeed accompanied by population range fragmentation. The skipper butterfly (Hesperia comma) in Britain has crashed in this century. This process left scattered and highly fragmented populations by the 1930s. The same happened with many bird species in New Zealand (kokako, kaka Nestor meridionalis, a large parrot or the weka Gallirallus australis, a flightless rail).

The sensitivity of fragmented populations is underlined by a trend seen in the success of reintroduction attempts. Seventy-six percent of 133 documented reintroductions into former “core” areas, while only 48% of 54 translocations to periphery or beyond succeeded. However, not all species show a similar range dynamics in the process of becoming rare. The Kirtland’s warbler (Dendroica kirtlandii), a small insectivorous bird living in North American forests, withdrew into the historical center of its range during a recent 60% population collapse.

C. Special Traits Related to Density

While population densities typically fluctuate widely, some species are naturally rare. The study of rarity holds promise to understand processes related to extinction, although only vague clues are available today. It would be important to know, for example, if naturally and anthropogenically rare species are equally sensitive to proximate causes of extinction.

In plants, locally rare and geographically restricted species have lower levels of self-incompatibility and poorer dispersal abilities. Rare plants are overrepresented in certain families (Scrophulariaceae, Lamiales) and underrepresented in others (Rosaceae), at least in North America. This may indicate that there are some biological traits and adaptations that are shared by rare species. Populations of large-bodied species fluctuate less than smaller-bodied taxa (although the measurement of population variability is not as easy as the concept suggests), yet body size is not a useful predictor of risk to extinction. In birds, body size was not a useful predictor of rates of population increase or decrease in a global sample of threatened species from 12 families at various trophic levels.

One important but counterintuitive fact is that trophic position has no consistent effect on extinction. It is difficult to detect a consistent tendency for more frequent extinction of species at higher trophic levels, fossil or extant. This is complicated by the difficulty in separating body size and trophic position (species at higher trophic levels are mostly large). Top predators, in other words, are not more prone to extinction than consumers at other levels.

It seems that large-bodied species are vulnerable to ultimate causes of extinction (hunting, habitat destruction) but less so to proximate causes (their populations fluctuate less).

D. Time Factor

The preceding important determinants are thought to vary in ecological time, 10 to 1000 years. However, as no species lives forever, there may be processes that are operating in evolutionary time. If range and abundance are also species-specific characteristics, some species will be more extinction prone (i.e., naturally rare and restricted in distribution) than others.

Among songbirds on West Indian islands, older taxa occur on fewer islands, have more restricted habitat distributions and have reduced population densities. However, body size: abundance plots within tribes of
birds have more positive relationships than expected in taxonomically ancient tribes; large-bodied species in old tribes are more common than small-bodied ones. This could be the product of differential extinctions of large-bodied rare species over time.

Bivalves and gastropods in Cretaceous fossils achieved characteristic range sizes early in their history, and this changed little thereafter. In this group, locally endemic species have much higher chances of extinction than more cosmopolitan genera.

VIII. THE PRESENT: A FULL-FLEDERED MASS EXTINCTION

There are very few documented cases of extinction of lower organisms. This is an inevitable consequence of our ignorance of the degree of biodiversity of those groups. Given this, it is not surprising that the current rate of extinction can only be roughly guessed. It is extremely probable that the rate of recent extinctions is several magnitudes higher than the “background extinction rates” and probably surpasses any similar mass extinction events in the earth’s history.

Given the above deficiencies, we can only estimate current rates of extinction. Among the comparatively well-studied birds and mammals, the documented extinctions this century numbers about 100 species. There are a total of 14,000 species of these classes, so the documented extinction rate this century is about 1%. This translates to an expected average life span of a bird or mammal species of around 10,000 years. This is 100 to 1000 times shorter than the average life span calculated from the fossil record.

Three different methods for predicting impending extinction rates suggest future life spans of birds and mammals of 200 to 400 years if current trends continue. These impending extinction rates are at least 10,000 times higher than background rates in the fossil record.

All evidence suggests that a sixth mass extinction event in the history of earth is underway. While the total effect cannot yet be guessed, we know that the sixth mass extinction will be unique in the earth’s history. It will be the first resulting not from environmental changes but the extraordinary population growth and the activities of one species. Our species now uses an estimated 25 to 50% of terrestrial net primary productivity. This is without precedent, and will make the coming extinction qualitatively different from all previous mass extinctions. We know enough to realize the gravity of the problem. We need a more elaborate understanding of the phenomenon, how it affects different groups and geographic locations, as our conservation actions will become more and more critical for the future of life on earth.

See Also the Following Articles

EXTINCTION, CAUSES OF • EXTINCTION, RATES OF • LATEX TIENTIONS ("LIVING DEAD") • MAMMALS, LATE QUATERNARY, EXTINCTIONS OF • MARINE MAMMALS, EXTINCTIONS OF • MASS EXTINCTIONS, CONCEPT OF

Bibliography


FIRES, ECOLOGICAL EFFECTS OF

William Bond
University of Cape Town

I. Fire in Earth History
II. World Biomes and Fire Incidence and Influence
III. Species Response to Burning
IV. Determinants of Fire
V. Ecological Effects of Fire
VI. Interactions between Fire and Other Ecological Processes
VII. Managing Fires
VIII. Fire and Global Change

GLOSSARY

fire regime The type of fire, mean and variance in fire frequency, intensity, severity, season, and areal extent of a burn in an ecosystem.
general circulation model (GCM) Computer models developed to simulate global climate and widely used for global climate change predictions.
serotiny Seeds stored on the plant with dispersal triggered by fire.

FIRE IS BOTH A NATURAL and anthropogenic disturbance influencing the distribution, structure and functioning of terrestrial ecosystems around the world. Many plants, and some animals, depend on fire for their continued existence. Others, such as rainforest species, are extremely intolerant of burning and need protection from fire. The properties of fire are changing as natural landscapes become fragmented and human influence becomes all pervasive. A sound understanding of its effects is an essential prerequisite for effectively managing this widespread ecological process.

I. FIRE IN EARTH HISTORY

Fire is an enormously influential disturbance over very large areas of land. Vegetation burns because the earth's atmosphere contains sufficient oxygen (>13%) to support combustion. Should oxygen levels rise above 30%, fires would be so frequent that dense forest vegetation, even in persistently wet climates, would be incinerated. There is an almost continuous record of fossil charcoal during the past 350 million years indicating that the atmosphere supported combustion for most of terrestrial plant evolution. Oxygen levels reached maxima in the Upper Carboniferous, 300 million years ago, when abundant fossil charcoal indicates frequent fires. At this and other times, frequent fires may have played a significant part in the ecology of palaeo-ecosystems.

Contemporary ecosystems contain a mix of ancient and modern plants with varied response to burning. Ancient fire-dependent elements include conifers such as the giant sequoias of California, whose association with fire has roots in the days of the dinosaurs. Many of the least fire-tolerant elements occur in angiosperm-dominated temperate and tropical forests of early Tertiary origin. Grasslands are the most flammable vegetation that has existed on the planet. Tropical (C4) grasses occupy one-fifth of the land surface but first appeared...
The current extent of fire-prone ecosystems has been drastically altered by humans. Hominids have used fire for perhaps as long as 1 million years. Regular use of fire, indicated by the frequency of domestic hearths in archeological sites, only became common in the past 100,000 years associated with the evolution of modern humans. Fire was widely used to promote the growth of edible plants in hunter–gatherer communities and to attract animals to hunting grounds. Fire was (and is) a tool used by farmers to clear new lands and to prepare sites for swidden-type farming. Increased intensity of human use of fire by crop farmers has led to the expansion of fire-prone systems in many parts of the world during the past few thousand years.

II. WORLD BIOMES AND FIRE

Incidences and Influence

Fires are rare only at the extremes of the climatic continuum. Neither the most humid tropical and temperate forests nor the driest deserts experienced fire as a major factor. However, between these two extremes, fire has influenced the extent and composition of a great diversity of ecosystems, including boreal forests, dry conifer forests, many grasslands (especially those dominated by tall grasses), temperate woodlands, tropical savannas, Mediterranean-type shrublands, heathlands, and eucalypt woodlands. Forests with mast-flowering bamboo understories are also prone to burning after the bamboo flower and die, creating massive fuel loads. Humans have changed landscape patterns of burning. Even humid tropical forests are beginning to burn as a result of logging. All these biomes experience fires of widely differing frequency and severity which help shape ecosystem structure and function.

Given the wide geographic extent of ecosystems that burn, fire influences the distribution and abundance of many species. Some ecosystems are dominated by species that depend on fire to complete their life cycles. Others are dominated by species that tolerate burning but have no direct dependence on fire. Ecosystems that seldom or never burn, except when disturbed by human activity, contain mixtures of species that fortuitously tolerate burning and species extremely intolerant of fire. The impact of burning on biodiversity varies greatly among these different types of ecosystems and species response patterns.

III. SPECIES RESPONSE TO BURNING

A. Plants

Fire has influenced some types of vegetation for so long that it has not only affected distribution of species but also led to the evolution of fire-dependent life histories. Burning triggers different stages in plant life cycles, including flowering, seed dispersal, and seed germination in fire-dependent plants. Perennial grasses and herbs, including orchids, lilies, and other bulb plants, flower prolifically after they have been burnt, often as a facilitative response to higher light, water, and nutrient availability. The African fire lilies, Cyrtanthus spp., flower only after fire. In these and other species, flowering is stimulated by constituents of smoke such as ethylene (e.g., Xanthorrhoea, the Australian “grass-trees”). Burning stimulates seed release from species with serotinous cone-like structures which store seeds on the plant for years between fires. Serotiny is common in conifers of the boreal forests and Mediterranean-climate regions and also among diverse groups of flowering plants in Australia and South Africa. Some species are polymorphic for the trait, with serotinous forms increasing in populations that regularly experience large severe burns. Burning also stimulates seed germination in plants that store dormant seeds in the soil. Heat-stimulated seed germination is common in many legumes and other groups with hard seed coats (e.g., members of Rhamnaceae). Thick seed coats prevent imbibition of water until cracked by the heat of a fire. Smoke-stimulated seed germination has been reported in fire-prone shrublands of South Africa, Australia, and California and also in some grassland species. Nitrogen dioxide, released in large quantities in smoke, cues seed germination in Emmenanthe penduliflora, a charapatral annual. Species with fire-stimulated flowering, seed release, or seed germination depend on fires for populations to expand.

Vegetative features of plants affect tolerance to burning. Thick bark and the ability to resprout allow many plants to survive burning. Neither feature is necessarily a fire adaptation. Sprouting is a common fire survival mechanism, either from the root stock or from branches above the ground. Some species posses large swollen burls or lignotubers which are thought to act as bud banks or storage reserves. Paradoxically, many woody plants in fire-dependent vegetation cannot resprout and are killed by fire. These nonsprouting plants often have higher seed production and higher seedling growth than related sprouting species. In some lineages, sprouting is the ancestral feature and loss of sprouting is viewed
Themeda triandra
Chionochloa

as an adaptive response to fire. Non-sprouting shrubs are particularly common in chaparral and similar shrublands and require fire to release seeds from serotinous cones or to stimulate germination. Among trees in fire-prone forests, many conifers do not sprout and a few eucalypts are also killed by fire. Non-sprouting species are particularly prone to local extinction if recruitment fails after burning.

Grasses are among the most fire resistant of all plant growth forms. The buds of new shoots are insulated by either layers of leaf sheaths or the soil where species have underground rhizomes. Grasslands recover from burning more rapidly than woody plants and can carry very frequent fires (1–3 years) in productive sites. Although many grasslands burn readily, few species have an obligate dependence on burning. Fire-stimulated flowering is rare but has been reported in many mostly temperate tussock grass species, including species of Chionochloa in New Zealand. Several very widespread warm-climate (C4) grasses (e.g., Themeda triandra and Andropogon gerardi) decline rapidly in the absence of burning. These species become locally extinct if fires are suppressed for more than a decade since litter accumulates, shades, and kills the grass plants.

Some woody plants have developed a remarkable ability to tolerate very frequent grassland fires. These species are particularly hostile environment for juvenile stages. Tree seedlings survive by rapidly acquiring the ability to sprout, and juveniles slowly develop food reserves in swollen roots and eventually produce bolting stems that place foliage above flame height. This peculiar life history occurs in several pine species (e.g., Pinus halepensis) growing in grassy habitats and also in many savanna trees. A tropical Asian species, Pinus merkusii, produces persistent sprouting juveniles in frequently burnt populations but nonsprouting forms in infrequently burnt island populations. Juvenile stages of savanna trees can tolerate repeated burning, suffering repeated killing of the stem parts, for decades before they die or escape the flame zone to become adults.

The occurrence of species with flammable morphologies in fire-prone environments has led to the suggestion that flammability has evolved to promote burning. In most cases, features that promote flammability have probably evolved for alternative functions. However, some woody plants accumulate highly flammable fuel by retaining dead branches and require fire for recruitment. In western North America, serotinous pine species retain dead branches and recruit seedlings only after crown fires have destroyed the population. These species appear to both promote the spread of crown fires and benefit from fire-stimulated reproduction. This strategy contrasts strongly with self-pruning, thick-barked pines (e.g., P. ponderosa) which tolerate frequent surface fires but are restricted to productive sites which support rapid juvenile growth.

B. Animals

The direct effects of fire on wildlife are often surprisingly small. Agile animals flee to unburnt refugia, often moving across the fireline to places of safety. Soil is a very effective insulator so that many animals survive in crevices and cracks or in burrows in the soil. Mortalities of large mobile vertebrates, including humans, do occur but only in the most severe fires. Reptiles and slow-moving invertebrates can suffer higher mortalities and their carcasses provide a food source to scavenging birds and other creatures in the first few days after a burn. The threatened bald ibis of South Africa makes extensive use of recently burnt grasslands, as does the endangered whooping crane in its Texan winter feeding grounds.

The indirect effects of burning are generally far more important, especially changes in habitat attributes and the successional mosaic initiated by forest fires. A large crown fire in a forest causes drastic structural change and local extinction of all faunal elements that depend on unburnt forest habitat. Post-burn stages are colonized by a new suite of species. Different successional stages support different suites of animals. Even frequently burnt grasslands, such as those of the South African highveld, have distinct bird assemblages which turn over with successive years of regrowth after burning.

The pattern of fires across a landscape imposes a mosaic of patches of different successional ages. The size and configuration of patches influence rates of local extinction and patch re-colonization (meta-population structure). For example, nectar-feeding birds in shrublands of Australia and South Africa lose their food source, shrubby members of the Protea family, after a burn and have to seek unburnt stands for food. The landscape configuration of old stands with flowering proteas, and young stands with immature shrubs, produces a highly mobile bird assemblage. Changes in the spatial pattern of fires may change extinction risks in different faunal elements. Some species of Australian honeyeaters are threatened with extinction because changes in the fire regime no longer produce the right mix of mature and immature populations of nectar plants. The decline of many bird species in Australian savannas has been attributed to the development of homogeneous landscapes through the systematic burn-
such as those produced by El Niño conditions. In con-
tems are generally associated with rare drought events,
conditions are needed for fire to burn pristine humid
sustain a fire, whereas months of extreme hot and dry
weather are sufficient to dry tall grasslands enough to
greatly on vegetation properties. A few days of hot, dry
of a warm, dry period needed for ignition depends
marked influences on the timing of burns. The length
of vegetation, climate and weather conditions exert
thus dry and heat adjacent material.
cause they are released from leaves, burn fiercely, and
burn readily. Volatile substances enhance burning be-
large amounts of oils, fats, waxes, and terpenes also
decomposing litter are highly flammable. Leaves with
and burn readily. Ecosystems that accumulate slow-
decomposing litter are highly flammable. Leaves with
fire and pre-heat flammable material are also key
factors. The moisture content of plant matter is critical.
Dead matter has the lowest moisture content, whereas
live leaves will burn more easily if their moisture con-
tent is low. The shape, size, and arrangement of plant
parts influence moisture content and flammability.
Plants with narrow leaves or thin branches dry rapidly
and burn readily. Ecosystems that accumulate slow-
decomposing litter are highly flammable. Leaves with
large amounts of oils, fats, waxes, and terpenes also
burn readily. Volatile substances enhance burning be-
cause they are released from leaves, burn fiercely, and
thus dry and heat adjacent material.

Because fire depends very much on moisture content
of vegetation, climate and weather conditions exert
marked influences on the timing of burns. The length
of a warm, dry period needed for ignition depends
greatly on vegetation properties. A few days of hot, dry
weather are sufficient to dry tall grasslands enough to
sustain a fire, whereas months of extreme hot and dry
conditions are needed for fire to burn pristine humid
tropical forests. Therefore, large fires in woody ecosys-
tems are generally associated with rare drought events,
such as those produced by El Niño conditions. In con-
trast, large fires in and grassy ecosystems are limited
by the availability of fuel and are more common after
high rainfall years. Fires generate their own heat that,
under hot, dry, and windy conditions, creates a positive
feedback, increasing the area burnt and the severity of
the fire by drying the vegetation before the fire front.
A source of ignition is needed to start fires. Today,
most fires are ignited by humans, except in sparsely
populated regions. Lightning fires are still common in
many landscapes. Lightning often accounts for large
numbers of fires but small areas burnt because many
are dowsed by subsequent rain or controlled by fire
crews attuned to the risk of thunderstorm fires. Rock-
falls also ignite fires, accounting for the spate of burns
often triggered by earthquakes. Volcanic activity is of
local significance, especially on islands in which light-
ning is unusual.

The contingent requirements of ignition, suitable
weather conditions, and contiguous flammable vegeta-
tion make for a high degree of uncertainty in predicting
landscape fire patterns.

IV. DETERMINANTS OF FIRE

The ecological effects of fire depend on the fire regime
rather than the occurrence of a single fire. The fire
regime is produced by the combined effects of climate,
fuel properties of vegetation, and ignition frequency. It
is characterized by the type of fire, mean and variance
in fire frequency, intensity, severity, season, and areal
extent of a burn. Types of fire include ground fires that
burn the organic layers of the soil, surface fires that
burn just above the ground, and crown fires that burn
in the canopies of trees. Ground fires occur only in
peaty soils in which they can be extremely damaging,
destroying roots and completely altering soil properties.
Crown fires predominate in short vegetation, including
grasslands, shrublands, and low conifer forests in boreal
and Mediterranean climate regions. Surface fires are
common in most woodlands and forests. When crown
fires do occur in forests, they produce massive stand-
replacing fires. Fire exclusion in national parks, such
as the giant sequoia forests of California, has led to an
increase in young trees which now act as bridging fuels
turning surface to crown fires with extremely damag-
ing consequences.

Fire frequency is estimated from maps of fires, re-
cords of fire scars, or patterns of charcoal deposition
in sediments. The mean and variance of fire return
interval are important descriptors of historical distur-
bance patterns. Changes in fire frequency are an impor-
tant cause of changes in ecosystem structure and func-
tion. Fire intensity is measured as energy released per
meter of fire front. It is widely used in fire fighting
operations. Fire severity is a distinct but related concept
defined as the ecological impact of a single fire and
usually estimated from the amount of plant biomass

A. The Fire Regime

The ecological effects of fire depend on the fire regime

consumed. A fast-moving fire that consumes little bio-
mass and a slow-moving fire that consumes more can
have the same fire-line intensity but different severity.
Fire severity is highly variable, literally depending on
the weather during a burn, wind conditions, and the
pre-burn condition of the vegetation. Fire season is
largely dictated by the moisture content of flammable
biomass. Where the vegetation dries out quickly, fires
can burn in almost any season. Seasonal timing of burns
can cause significant changes in species composition
and ecosystem structure. Continuity of flammable vege-
tation, especially at the landscape scale, strongly influ-
ences the spread of fires. Habitat fragmentation can lead
to a reduction in fire frequency of isolated fire-prone
ecosystems or an increase in fire-excluding forests sur-
rrounded by flammable vegetation. Land abandonment
in some countries has led to successional changes pro-
ducing large, contiguous, highly flammable vegetation.
In Spain, reduction of pastoral activities has led to the
conversion of grasslands to highly flammable shrub-
lands. This process has contributed to an increase in
the area burnt annually from a few thousand hectares
in the 1960s to hundreds of thousands of hectares in
the 1990s. The increase in fires in southern European
countries during this period has not been matched in
southern Mediterranean countries in which rural de-
population is not as advanced.

V. ECOLOGICAL EFFECTS OF FIRE

A. Ecosystem Structure

The consequences of increasing fire frequency and se-
verity for ecosystem structure are

to reduce vegetation height (tall forests to shorter ones
and woodlands to shrublands);
to replace woody vegetation by grasslands;
to promote flammable species or communities (low lit-
ter decomposition rates, more xeromorphic leaves,
and finer twigs/branches); and
to reduce biomass.

There are no reliable global estimates of the percent-
age reduction of plant biomass attributable to burning.
Both tropical and temperate landscapes contain mix-
tures of fire-prone grassland or shrubland communities
and closed forests which tend to exclude fire. Success-
sional replacement of the fire-prone communities by
fire-excluding forest elements frequently occurs when
fires are suppressed. These changes can be very rapid.
In southern Africa, forests have replaced both grass-
lands and savanna woodlands after 10–30 years of fire
suppression. However, counterexamples exist in which
fire suppression does not lead to changes in savanna/
forest boundaries. Stable boundaries often coincide with
different soil types, with forests occurring on the better
drained soils.

Changes from fire-excluding to fire-promoting eco-
system structures may also be rapid. In the Brazilian
Amazon, fires in closed-canopy forests spread as a “thin,
slowly creeping ribbon of flames a few tens of centime-
ters in height.” Despite the low severity of an initial
fire, structural changes causing an opening of the can-
opys dries out the understory and contributes to an
increase in flammable understory biomass, increasing
the risk of a second fire. Weedy vines and grasses
quickly colonize twice-burned forests, further adding
to the flammable biomass. Positive feedback of this kind
is estimated to reduce a forest to vegetation resembling
recently abandoned farmland in 20–30 years.

B. Ecosystem Function

The immediate effect of fire is gaseous loss of carbon
and nitrogen from material that burns. More biomass
burns in more severe fires and the nutrient losses are
greater. Strong winds accompanying fire lead to losses
of phosphorus and cations blown away in ash. Cation
nutrients in ash tend to be mobile and in a plant-avail-
able form. Their presence leads to increases in soil
pH—large increases in acid forest soils and smaller
increases in neutral or alkaline soils in grasslands or
savannas. Increased solar radiation, decreased evapora-
tion, and higher pH lead to increased microbial activity,
increased rates of mineralization, and increased avail-
ability of nutrients after a burn. After a chaparral burn,
for example, nitrate increased more than 20-fold rela-
tive to unburnt controls. Short-term increases in nutri-
ent availability can be offset by long-term decreases
where fire frequencies are high and inputs to the system
between fires are not high enough to replace losses.
Severe fires can lead to nutrient shortages. Many ecosys-
tems have nitrogen-fixing organisms as major compo-
nents of postburn vegetation which replace nitrogen
losses in a few years.

Fire can lead to changes in ecosystem processes at
landscape scales. The reduction in biomass caused by
burning and changes in soil properties lead to tempo-
rary hydrological changes in patterns of stream flow.
Severe fires can lead to changes in soil erosion. The
Yellowstone fire of 1988 led to significant increases in
sediment loads and altered the geomorphology of river
systems. Debris spread over a distance of 12 km in one
valley bottom.
C. Species and Populations
At the local scale, and within fire-prone communities, species respond to differences in fire frequency, season, and intensity. Variation in the fire interval is an important determinant of population status. In shrublands, the effect of fire on population growth depends on key demographic attributes of the species. Population size of non-sprouting species fluctuates more than that of sprouting species, and local extinction is not uncommon after a single fire. Species that are slow to mature are particularly vulnerable where populations are burnt before they have first flowered and set seed. Populations are also negatively affected where intervals between fires exceed the life span of a species or its seedbank. Many grasses are also sensitive to variation in fire frequency. In Australia, South Africa, and Southeast Asia, Themeda triandra dominates many frequently burnt grasslands. Populations may decline to local extinction if fires are excluded for longer than a decade. Manipulation of the fire interval is a key tool for influencing biodiversity of vegetation stands. Information on the reproductive status of plants, especially the size of the viable seedbank, at different postburn stages has been extensively used to help determine optimum fire frequencies to maintain particular species in fire-prone woody ecosystems.

Recovery of plant populations also depends on the unique combination of circumstances on the day a fire burns. These “event-dependent” effects can be as important as fire frequency in influencing biodiversity in some ecosystems. Fire season has a marked influence on recruitment of serotinous members of the Proteaceae in Western Australia and South African fynbos. Spring burns can reduce protea populations to less than one-tenth of their preburn density, whereas autumn burns can result in a 10-fold or greater increase in plant densities. Fire season also influences recovery of sprouting plants where the size of root reserves varies seasonally, affecting the vigor of resprouting. Manipulation of fire season is sometimes the only effective tool for managing densities of sprouting shrubs. Seed density of clonal species of hazel (Corylus spp.) in the understorey of temperate forests increased 5-fold with four successive spring burns but was halved by successive summer burns. In Zambian woodlands, annual burns in the early dry season caused a 10-fold increase in tree seedlings and halved the adult mortality rate over a decade. Grassland composition can also be very sensitive to fire season. In a long-term burning experiment in the Kansas prairies, late-spring burns caused a halving of Andropogon scoparius biomass relative to fires burnt a few weeks earlier. The effects of fire season on species and ecosystem recovery are poorly known in many ecosystems. They may be unimportant where the fire season is short because of climate constraints.

By definition, severe fires cause the most extreme biomass losses in ecosystems. In eucalypt and conifer forests, intense crown fires kill all aboveground plant growth. Where trees are incapable of sprouting, these fires cause complete replacement of canopy trees. Sprouting plants, especially shallow-rooted species, can be killed by intense burns. Fire intensity is an important factor in savanna ecology. Where grass growth is sufficient to carry fires at frequent intervals, burning kills the aboveground parts of trees. The amount of dieback depends on the intensity of the burn. In mesic savannas, fires are so frequent and so intense that juvenile trees may spend decades trapped in the grass layer. The frequency and intensity of fire are important determinants of tree biomass (and habitat structure) in these ecosystems.

The effects of fire severity on recruitment from seeds vary among species. Legumes and other plants with hard, dormant seeds in fire-prone shrublands do not germinate unless a burn heats the soil sufficiently. For example, the Australian shrub, Acacia suaveolens, will not germinate unless soils are heated to more than 50°C. Variation in fire intensity can directly affect species composition in fire-prone shrublands, incinerating some seeds and stimulating germination in others. Key species, such as legumes, may fail to germinate after low-intensity fires that are applied for safety reasons.

The relative sensitivity of plants to fire season and fire severity varies among species. This makes general prediction of population trends under different fire cycles difficult without species-specific information. In Mediterranean shrublands, species with diverse responses to frequency, season, and intensity of burn occur in the same community suggesting a history of fires that vary in these factors. It is a considerable conservation challenge to incorporate variability into fire regimes to maintain the full diversity of species.

VI. INTERACTIONS BETWEEN FIRE AND OTHER ECOLOGICAL PROCESSES
A. Fire and Herbivory
Fire interacts with, and is influenced by, other agents of disturbance that can also influence ecosystem structure on a large scale. Herbivores influence the distribution and biomass of plant parts and therefore the attri-
butes of fire regimes. Heavily grazed savanna grasslands
do not burn. Persistent heavy grazing often leads to an
increase in tree densities because of the reduction in
fire frequency. In Africa, elephants open up woodlands,
engendering grass growth which promotes more frequent
severe fires. The combination of elephants and grass
fires can cause a marked reduction in tree densities. In
miombo woodlands of Zimbabwe, changes in woodland
structure under the combined influence of elephants
and fire markedly reduced bird diversity and led to
local extinction of four endemic woodland bird species.

Insect herbivory also influences fire regimes, especi-
cially in northern ecosystems. In Balsam fir (Abies balsas-
mea) and red spruce (Picea rubens) forests of eastern
North America, fires are rare. Large-scale tree mortality
is caused by spruce budworm outbreaks which inhibit
the spread of fire by changing vegetation structure and
fuel properties. In general, however, plants that provide
fuel for burning make poor food for herbivores and
vice versa. Fires burn readily where decomposition is
slow, causing dead matter to accumulate. Slow decom-
position is associated with high C:N ratios, high fiber
contents, and high leaf-specific weight, all of which
inhibit herbivore food intake. Thus, the most fire-prone
vegetation tends to be least edible and vice versa.

B. Fire and Landscape Fragmentation

The pattern of fire in a landscape is sensitive to the
contiguity of flammable vegetation. Landscape frag-
mentation can have major impacts on the fire regime,
in turn affecting survival of species within fragments.
Species losses from North American prairie fragments
during a 30-year period were greater among plants of
early postburn environments. Similar patterns of local
extinction occur in fynbos fragments surrounded by
nonflammable forests in South Africa. The causes of
species loss in both prairies and fynbos is attributed to
infrequent burning due to isolation of the fragments.
Small forest patches in extensive areas of fire-prone
shrublands or grasslands are also more likely to suffer
local extinction of species intolerant of burning. In the
tropical forests of Guyana, fire-tolerant forest tree spe-
cies with thick, fissured bark and small seeds are unusu-
ally common near savanna boundaries and human set-
ttlements.

C. Fire and Invasives

The interplay between fire and invasive species can
cause alarming ecosystem transformation. Direct effects
of invasive plants can be minor relative to the indirect
effects on fuel properties and fire regimes. Grass inva-
sion into woody ecosystems has particularly damaging
consequences. In South America, invasion of tall non-native
grasses has transformed the unique forests to grasslands
entirely alien to the island ecosystem by fueling fre-
quent fires. In South America, invasion of logged-over
tropical forests by fire-promoting grasses can cause
elimination of the forest ecosystem and its rapid replace-
ment by the alien grass. In southwestern Australia, spe-
cies-rich heathlands are fire dependent but they have
also been invaded by non-native grasses which burn so
frequently that the heathlands are transformed into a
species-poor savanna with scattered relictual trees. The
reverse pattern, that of invasion of fire-prone grasslands
by plants that do not burn easily, may also be a problem.
For example, Lantana camara is invading fire-depen-
dent grasslands in South Africa but burns much less
readily than the native vegetation.

VII. MANAGING FIRES

Because wildfires are such a widespread feature of world
vegetation, managing fires for particular objectives, in-
cluding conservation of biodiversity, is a major concern.
Attitudes and actions to wildfire are never neutral. Fires
are actively suppressed in some ecosystems but pur-
posely ignited in others. Wildfires commonly arouse
public and media concern. The spread of the urban
interface as more people flock to cities has increased
threats of wildfire to people and properties. Smoke man-
agement has become a health and safety issue. Enor-
mous effort and expense are often expended on fire
management. In the former USSR, fires in protected
areas were suppressed regardless of the ignition source
using hundreds of planes and more than 8000 air-borne
fire-fighters. Appropriate fire management policies con-
tinue to be a central management headache in protected
areas as different as Yellowstone National Park in a
cinereous ecosystem and Kruger National Park in Afri-
can savannas. The expense of maintaining fire fighting
teams consumes a significant fraction of conservation
agency resources. Litigation, when fires move beyond
the borders of protected areas, can also strain conserva-
tion budgets. Arson burns are not uncommon, some-
times as a protest against state authority. Fire manage-
ment consumes a great deal of time and is a major
expense in fire-prone ecosystems.

A. Fire Management Policies

There is no consensus on how to manage fires in pro-
tected areas or outside them. For the first half of the
Second century, complete fire suppression was a common policy, and it is still in many parks. Suppression policies have slowly changed, partly because of their cost, partly because they do not work, and partly because of changes in ecological thinking. Fire suppression leads to the build-up of dead biomass in fire-prone ecosystems which produce very severe fires when they do burn. Suppression policies were also relaxed following the discovery of fire-dependent features of plants. In the Cape fynbos, the beautiful marsh rose (Oрошammos zeyheri) declined to a handful of plants before managers realized that the species had an obligate dependence on fire to stimulate germination of its seeds. There has also been increasing recognition of disturbance as a natural process in ecosystems. Outside of rain forests, complete fire suppression is rarely the aim of fire management. However, decades of suppression have led to pronounced changes in ecosystem structure that now pose major challenges regarding how to re-introduce fires without causing more problems. Many conifer forests were maintained as open parklands by frequent surface fires. Fire suppression has allowed numerous young trees to establish, creating “bridging fuels” and a switch from surface to crown fires which destroy the whole forest structure. In some savannas, fire suppression led to invasion of grasslands by forest species in a process that is very difficult to reverse except by costly manual clearing of trees to restore large mammal habitat.

Prescription burns are fires intentionally lit for management purposes. Safety for the fire crew, and for property, is always an important consideration. Therefore, prescribed burns often cause significant changes to the fire regime, especially to season and intensity but also to fire frequency. Changes to the fire regime can lead to significant population declines in sensitive species such as legumes, which depend on intense heat for seed germination.

Prescription burning calls for clear management objectives. Because fire is so influential a force on community structure and composition, decisions have to be made regarding the desirable objective and what kind of burning pattern should be used to achieve it. In the South African savanna parks, fire policy changed from complete fire exclusion to prescribed burning at fixed intervals (promoting large mammals) to fire regimes that “maximize biodiversity.” The intention of the latter is to create and sustain a landscape mosaic of different successional ages or habitat types that can maintain viable populations of most species. What to do in practice is more difficult.

Another commonly considered policy is to re-create “natural” burning regimes. This policy permits fires that have been ignited by lightning, rockfalls, or other non-human agents but suppresses fires of human origin. It is difficult to apply in practice because vegetation has been fragmented by roads, buildings, croplands, and the like so that fires can no longer spread unhindered across the landscape. Re-creating natural fire regimes has also been criticized for not including aboriginal influences on fire regimes which may have significantly influenced landscape patterns over millennia. A variation on natural burning policies is therefore to re-create aboriginal burning practices to try to reproduce pre-agricultural landscapes. Implementation of these policies is constrained by lack of biological knowledge, techniques for fire management, and safety-related considerations.

In practice, fire management increasingly involves a mix of fire suppression, prescription burns, and controlled natural fires depending on safety concerns and conservation objectives in different parts of the landscape.

VIII. FIRE AND GLOBAL CHANGE

A. Effects of Global Change on Fire Ecology

The indirect effects of climate change on disturbances such as fire and herbivory, and thereby on ecosystems, is likely to be as great or greater than direct influences of climate change in many fire-prone ecosystems. The pace of vegetation change in the western Canadian Arctic, an area subject to particularly large temperature change, appears to depend on fires. The treeline is moving northward with each successive fire because trees are limited more by suitable seed beds for regeneration than by climate constraints on growth. Ground fires that remove peat layers provide the necessary seed bed to cause trees to advance.

It is difficult to predict global change impacts on disturbance regimes. The multi-causal origins of fire create considerable uncertainty. Even in large tracts of undisturbed land such as the boreal forests, it is still difficult to balance the effects of changes in vegetation type, ignition (more convective storms), length of the fire season, and fire weather to predict how fire regimes may change. Models that predict climate change impacts on fire regime combine physical models for fire spread, GCM predictions of future climate parameters, and models for future plant growth. Predictions suggest a higher incidence of fire, for example, in chaparral,
with larger and more intense burns because of higher temperatures and higher fuel accumulation rates. However, instantaneous fire weather effects and extreme events drive contemporary fire patterns and are still poorly predicted by GCMS. In boreal forests, despite nearly a century of temperature increase, fires have declined in areas remote from human intervention. Boreal forests are predicted to burn more frequently and more intensely under global warming and elevated CO$_2$. The difference between predicted and observed patterns is partly due to an increase in precipitation associated with increasing temperatures.

B. Fire as a Source of Greenhouse Gases

Vegetation fires may contribute significantly to global climate change. Annual gross amounts of carbon released into the atmosphere from global savanna and forest fires (1990s) are estimated to be 1.7–4.1 Pg compared with release of carbon from fossil fuel burning of 3 or 6 Pg per year. If biomes remain stable in their distribution, most of the carbon released from burning will be taken up again in new plant growth. However, there is increasing concern that increasing fires will release stored carbon more permanently into the atmosphere because of more permanent ecological changes. Increased fires, and a change from surface to ground fires, in boreal forests could release 86–94 Pg carbon into the atmosphere according to some predictions. The increasing frequency of fire in humid tropical forests is also of concern. Tropical forests are estimated to store one-fifth of the world’s carbon. If burning of tropical forests continues, very large areas could be converted to flammable secondary scrub or grassland releasing this carbon into the atmosphere.

Fire is also important as a source of aerosols. Aerosols decrease regional and global irradiation through backscattering of incoming solar radiation. It has been estimated that the net effect of aerosols of pyrogenic origin is to reduce global irradiation by 2 W m$^{-2}$, causing a net global cooling effect of approximately 2°C. The possible magnitude of these effects has led to intensive efforts to quantify the frequency of vegetation fires using dedicated satellites and measurements of gaseous and aerosol emissions from fires. Public concern regarding atmospheric impacts of burning is leading to public pressure regarding the use of fire for conservation purposes. This may have positive effects if it leads to adoption of logging practices that reduce fire hazard in humid forests. Public pressure to suppress fires could have negative impacts on fire-prone ecosystem and their fire-dependent species.

See Also the Following Articles

**CARRIER CYCLE • CLIMATE CHANGE AND ECOLOGY, SYNERGISM OF • DISTURBANCE, MECHANISMS OF • FARMLAND AND RANGE ECOLOGY • GREENHOUSE EFFECT**

Bibliography


FISH, BIODIVERSITY OF

Gene S. Helfman
University of Georgia

I. Taxonomic Diversity
II. Geographic Diversity

GLOSSARY

adipose fin A small, fleshy fin without supporting spines or rays, set far back on the dorsal surface of many catfishes, characins, salmons, and other groups.
ancestral The taxon from which descendant species are derived, often synonymized as primitive or generalized. Ancestral traits or conditions are those which appear in an ancestor.
depauperate Of low diversity, lacking in species; opposite of speciose.
derived Later-appearing taxa within a lineage, often synonymized as advanced or specialized. Derived traits are those which appear in a descendant species and are changed from the ancestral condition.
dependent Restricted or native to a geographically defined area.
estant Living; opposite of extinct.
extincted Locally extinct.
endemic A large supercontinent in what is now the Southern Hemisphere that separated during the Mesozoic, forming the modern continents of Australia, Antarctica, South America, and Africa.
Gondwana A large group of closely related, endemic species in a small area; all are descended from a single ancestor.

A FISH IS AN AQUATIC VERTEBRATE that (usually) has scales, fins, and gills (Box 1 and Fig. 1). Fishes constitute more than 23,000 of the known 48,000 species of living vertebrates and are divided taxonomically into three major groupings: jawless fishes (agnathans), cartilaginous fishes (chondrichthyans), and bony fishes (osteichthyans). Fishes are found almost everywhere on Earth where water of reasonable integrity exists. Fish habitats include the deep sea to depths of 8000 m, high mountain streams and lakes to 5000 m altitude, and just about every aquatic habitat in between. Marine fishes make up 58% of all species, freshwater species make up 41%, and 1% of fishes move regularly between the ocean and fresh water. Tropical areas have the highest diversity. Fish biodiversity is threatened by a wide range of human activities, but habitat modification, overharvest, and introduced species are particularly injurious.

I. TAXONOMIC DIVERSITY

A. Overview

Approximately 23,000 known fish species inhabit the earth’s oceans, estuaries, rivers, lakes, and streams. This
Box 1

What Is a Fish and What Are Fishes?

Generally defined, a fish is a cold-blooded, aquatic-living chordate with fin-like appendages, a body covered with scales, and that breathes using gills (Fig. 1). Exceptions to and variations in all traits are common. Some sharks, tunas, and billfishes are warm-blooded. Fins can be unsegmented and spiny or segmented and soft rayed (and soft rays can be hardened). Tail types include (i) the heterocercal tail of sharks and sturgeons, in which the notochord or vertebral axis extends considerably into the upper lobe; (ii) the abbreviated heterocercal tail of gars and bowfin, in which the vertebral axis extends only slightly into the upper lobe; (iii) the leptocercal or diphycercal tail of lungfishes, the coelacanth, and rattails, in which vertebrae or tail rays extend through the middle of the tail, forming a pointed tail; and (iv) the homocercal tail of most advanced bony fishes, in which the vertebral column ends at the tail base (the urostyle) and fin rays form a symmetrical, two-lobed tail. Scales also vary in terms of the number of layers of bony material that constitute them and the extent of spiny projections that cover their surface; more primitive fishes generally have heavier scales, and more advanced fishes have lighter scales, often with more projections. Scale types include the placoid scales of sharks, the ganoid scales of gars and bichirs (and sturgeons), the cycloid scales of lower teleosts, and the ctenoid scales of higher teleosts.

To ichthyologists, “fish” refers to one or more individuals of a single species, whereas “fishes” refers to more than one species, regardless of how many individuals are involved. Hence, this article is about fishes.

Incredible diversity exceeds that of all other vertebrate groups combined. It forms a wealth of biological wonder for ichthyologists (fish scientists), but such large numbers can be overwhelming to someone unfamiliar with the many taxonomic groups and their names. However, the different taxonomic groups are logically arranged according to well-studied evolutionary relationships, with those more closely related to groups that evolved earlier in geologic time (so-called ancestral or primitive taxa) placed earlier in lists and those groups that evolved relatively recently (derived or advanced taxa) placed later in lists.

The most primitive of the living fishes are the jawless fishes. These arguably include 22 species of marine lancelets, and less arguably about 85 species of marine hagfishes and freshwater or anadromous (migratory between freshwater and marine) lampreys. Cartilaginous sharks, skates, rays, and chimaeras include approximately 10 orders, 43 families, and 850 species of almost entirely marine, relatively large-bodied predators. Skates and rays are more diverse than sharks, constituting about 55% of all cartilaginous fishes. Chimaeras are cartilaginous fishes distantly related to sharks that consist of 3 families and 31 species.

Bony fishes make up the vast majority of living fish species, exceeding all other groups in species, habitat, reproductive, and feeding diversity. Bony fishes vary in length from the 8-mm pygmy gobies to the 12-m long oarfish, 900-kg marlin, and 1000-kg-plus ocean sunfish. There are perhaps 45 orders and 435 families among the approximately 24,000 species of bony fishes, ranging from the relatively primitive lungfishes, coelacanth, sturgeons, bichirs, gars, and bowfin to the more advanced teleostean (higher bony fish) groups that include bonytongues, eels and tarpons, herring-like fishes and the so-called true teleostean groups of minnows, salmons, various deep-sea taxa, and cods. The spiny-rayed teleosts are the most evolutionarily advanced of the fishes and include mullets, silversides, scorpion fishes, perch-like fishes, tunas, flatfishes, and triggerfishes.

FIGURE 1 Anatomical features of a hypothetical fish, including some common features that are measured during fish identification. 1, premaxilla; 2, maxilla; 3, dentary; 4, barbel; 5, snout; 6, nostril; 7, preopercle; 8, branchiostegals; 9, opercle; 10, opercular spine; 11, pectoral fin; 12, lateral line; 13, first part of (or spinous) dorsal fin; 14, second part of (or soft) dorsal fin; 15, adipose fin; 16, caudal fin; 17, caudal peduncle; 18, lateral scutes; 19, anal fin; 20, pelvic or ventral fin (reproduced with permission from Greenfield and Thomerson, 1997).
Phylum Chordata

Subphylum Cephalochordata; order Amphioxiformes
(lancelets, two families, 22 species, marine, tropical and temperate)

Subphylum Vertebrata

Superclass Agnatha

Class Myxini; order Myxiniformes (living hagfishes; one family, 43 species, temperate marine)

Class Cephalaspidomorphi; order Petromyzontiformes
(living lampreys; one family, 41 species, temperate fresh water and anadromous)

B. Cephalochordates

Lancelets may not be fishes because they lack scales, fins, and gills. However, their evolutionary and anatomical affinities are similar to what most workers believe characterized the ancestors of fishes, and lancelets are studied primarily by ichthyologists. Lancelets possess a dorsal nerve tube, in common with all vertebrates. They also have a notochord, which is a cartilaginous rod that runs the length of the body and is shared with all embryonic vertebrates as well as with the adults of many primitive living fishes. Lancelets are small (to 3 cm), slender organisms that as adults occupy sandy, usually shallow bottoms in all major tropical and temperate oceans. A commercial fishery for lancelets involving bottom dredging exists in southern China. The fishery is in apparent decline.

C. Agnathans

The first fishes lacked jaws. Modern jawless fishes—hagfishes and lampreys—look approximately similar, with slippery, eel-like bodies and jawless heads. The fossil record, however, indicates that they have been separated evolutionarily for hundreds of millions of years. Hence, most similarities are due to convergent or parallel evolution.

1. Hagfishes

Hagfishes, known also as slime eels or slime hags, produce copious mucus from many pairs of slime glands. A disturbed 2-ft-long hagfish can fill a 5-gallon bucket with slime. However, a hagfish covered in its own slime will suffocate. To rid itself of slime, a hagfish ties a knot in its tail and passes the knot forward until the slime is pushed off. Hagfishes are nocturnal predators on small invertebrates but are better known for their scavenging behavior, which involves burrowing into a dead or dying fish and consuming the prey from the inside.

Hagfishes occur almost worldwide in temperate and cold-temperate oceans, usually in water deeper than 30 m. Hagfishes can reach high densities, upwards of 0.35 m⁻², on soft-bottom marine areas, which is the most abundant habitat type in the world. Hence, hagfishes could be ecologically important as predators and scavengers. They are also common in the diets of seals and sea lions. Hagfishes are commercially important because their hides are the popular "eelskin" of wallets, purses, and briefcases. Overfishing has depleted hagfish stocks in Korea and Japan, and new fisheries are being exploited, and probably overexploited, in the eastern Pacific and western Atlantic. This unfortunate chain of events has characterized marine fisheries worldwide. Drastic reductions in hagfish populations brought on by overfishing could potentially disrupt a widespread ecosystem, not enough is known about hagfish ecology to predict these impacts.

2. Lampreys

Lampreys also have a notochord rather than vertebrae. Lampreys go through a long-lived larval phase; the free-living, blind, toothless larva lives in silty stream-beds in which it filters microscopic organisms from the water for up to 7 years before transforming into an adult. In brook lampreys, larvae transform into non-feeding adults, live for 6 months, spawn, and die. Other species transform into feeding adults and live for 1–3 years as parasites on other fishes. They rasp holes in their host's skin and live off its body fluids. Accidental introduction of the parasitic marine lamprey into the Great Lakes of North America has contributed to the decline of lake trout, whitefishes, and blue pike.

Lampreys are cool-water species (30° north and south latitude or higher). Most lampreys live in fresh water, but some parasitic species are anadromous. Brook lampreys typically live in headwater streams, an ecosystem type frequently disrupted by human activities. Hence, several U.S. brook lampreys are imperiled. North America's smallest lamprey, the Miller Lake lamprey, was poisoned into extinction because it parasitized introduced trout in its only habitat—Miller Lake, Oregon (Fig. 2).

D. Cartilaginous Fishes

Elasmobranchs (sharks, skates, and rays) include about 800 species that live in all the world's oceans; a few live in fresh water. Elasmobranchs are characterized by a cartilaginous skeleton hardened by calcium deposits and usually five (sometimes six or seven) gill slits. They lack lungs or gas bladders but instead have large, oil-
filled livers which may aid in buoyancy. Their teeth and pedestal-like placoid scales develop from the same embryonic structures. Teeth are continually lost and replaced, and a shark may produce as many as 30,000 teeth during its lifetime. All elasmobranchs have internal fertilization, and many bear live young that are nourished by the mother via a complex umbilicus and placental structure analogous to that found in mammals. Slow growth, late maturation, and low reproductive output make many elasmobranchs exceptionally vulnerable to human exploitation.

Class Chondrichthyes: cartilaginous fishes
Subclass Elasmobranchii: shark-like fishes
Superorder Euselachii: modern sharks and rays
Order Heterodontiformes: 1 family, 8 species, bullhead and horn sharks
Order Orectolobiformes: 7 families, 31 species, including wobbegongs, nurse, and whale sharks
Order Carcharhiniformes: 8 families, 210 species, including catsharks, requiem, and hammerhead sharks
Order Lamniformes: 7 families, 16 species, including sand tiger, megamouth, thresher, basking, and mackerel sharks
Order Hexanchiformes: 2 families, 5 species, frill and cow sharks
Order Squaliformes: 4 families, 74 species, including sleeper and dogfish sharks
Order Squatiniformes: 1 family, 12 species, angel sharks
Order Pristiophoriformes: 1 family, 5 species, saw-sharks
Order Rajiformes: 13 families, 456 species, including sawfishes, electric rays, guitarfishes, skates, sting-rays, eagle rays, and manta rays

1. Sharks

Approximately 350 shark-like species are alive today (Fig. 3A). Sharks are generally large (>1 m), predatory fishes. The diverse requiem or ground sharks (carcharhiniforms) include the tiger, grey reef, bull, blue, lemon, and hammerhead sharks. Lamniform mackerel sharks are primarily offshore, pelagic inhabitants. The squalliform dogfishes, the second largest shark order, are most successful and abundant in the North Atlantic, North Pacific, and deep-sea regions.

Sharks range in size from the 15-g, 16- to 20-cm dwarf dogshark to the 12,000+ kg, 12+ m long whale shark, the largest fish in the world. White sharks as large as 6 m and 3324 kg are known; larger individuals are suspected. Other large sharks include basking sharks (9.8 m), great hammerheads (5.9 m), Greenland sharks (6.4 m), tiger sharks (5.9 m), and megamouth (5.4 m).

Sharks inhabit all oceans except the Antarctic. Depth records for sharks are held by the Portuguese shark at 3690 m and an unidentified dogfish at 4050 m. A few carcharhinid sharks enter fresh water; bull sharks have been captured 4200 km up the Amazon River and 1200 km up the Mississippi River. Large pelagic sharks may cross entire ocean basins; blue sharks have been tracked across the North Atlantic Ocean and back, a distance of 16,000 km.

Most sharks are predatory on large prey, but three of the largest sharks—the basking, megamouth, and whale sharks—feed on zooplankton. A small, 40-cm-long, midwater species, the cookie-cutter shark, is an ectoparasite on the sides of tunas, dolphins, whales, an occasional megamouth shark, and even rubber sonar domes of nuclear submarines. Some sharks use structures other than jaw teeth to capture prey. Thresher sharks use the long upper lobe of their tails to stun schooling prey. Sawsharks (and rajiform sawfishes) have elongate, blade-like snouts studded with lateral teeth which they slash laterally to disable prey. Hammerheads use their broadened hammer-shaped head to pin stingrays against the bottom before biting chunks out of the rays’ wings.

Sharks are sensitive to chemicals, able to detect 1 part fish extract per 10 billion parts seawater. Sharks have good vision, although they tend to be slightly myopic (farsighted). Sharks are also highly sensitive to sounds, including infrasonic sound below 10 Hz, and can localize the direction from which underwater
sounds originate. Sharks can also locate prey by detecting the weak electric fields prey emit. This electro-sensitivity may also allow them to navigate using the earth’s geomagnetic fields. Many sharks have relatively large brains, with brain to body weight ratios comparable to those of some birds and mammals.

Sharks grow slowly and live long: spiny dogfish live 70–100 years and lemon sharks 30–60 years. Sharks also reproduce slowly. Lemon sharks may not mature until they are 24 years old, and spiny dogfish may not mature until they are 35 years old. Sharks produce relatively few, large young with a long gestation period. Clutch size varies from 1 or 2 live young (sand tigers, threshers, and makos) to 300 in the whale shark. Gesta-
tion periods average 9–12 months but may be 2 years in spiny dogfish and 3.5 years in basking sharks. Bullhead and nurse sharks lay eggs, but most sharks bear live young. The requiem sharks have the most advanced developmental pattern, in which an umbilical cord connects mother and embryo, transporting nutrients and oxygen to the embryo and carrying metabolic wastes to the mother.

Because many sharks mature late, reproduce at long intervals, and have low reproductive output, shark populations are easily and frequently overfished. Thresher sharks, school sharks, spiny dogfish, porbeagles, basking sharks, bull sharks, and soupfin sharks are all examples of shark stocks that have been overexploited. White sharks are protected in Australia and South Africa; white sharks and whale sharks are listed as endangered by the International Union for the Conservation of Nature.

2. Skates and Rays

The rajiform skates and rays are 450 species of mostly benthic (bottom-living), mostly marine forms (Fig. 3B). In skates and rays, the pectoral fins are fused to the sides of the head and the five gill slits are under the head. Skates are most diverse in deep water and at high latitudes, whereas stingrays are most diverse in tropical, inshore waters. Some rajiforms live much or all of their lives in fresh water. Largetooth sawfish frequently swim up rivers in Central and South America. Two stingray families contain entirely freshwater species—the river stingrays of South America and several species in the large stingray family Dasyatidae. The latter inhabit African, Southeast Asian, and New Guinea rivers.

Skates and rays feed mostly on benthic invertebrates, except for the huge (up to 6-m wide) manta rays, which capture small crustaceans and fishes in the water column. Torpedo rays stun prey with powerful electrical discharges (50 V and 50 A = 1 kW output). Rajiforms reproduce by either laying eggs (skates) or bearing live young (rays). Embryonic skates develop inside the “mermaid purse” egg cases for as much as 15 weeks.

Skates in some locales are actually increasing in number because of overexploitation of competing bony fishes, such as cod in the North Atlantic. However, the giant barndoor skate of the northwest Atlantic and its relative, the common skate of the northeast Atlantic, are caught incidentally to bottom trawling for bony fishes; they have been seriously depleted and may face extinction. Largetooth sawfish in Nicaraguan lakes have been drastically overfished.

3. Chimaeras

Class Chondrichthyes
Subclass Holocephali
Order Chimaeriformes: 3 families, 31 species, chimaeras

Chimaeras, also known as rat- or rabbitfishes, share a cartilaginous skeleton and other features with elasmobranchs. They differ by having (i) the upper jaw permanently attached to the braincase, (ii) continually growing tooth plates in the jaws instead of replaceable teeth, (ii) a single gill flap instead of five or more gill slits, and (iv) no scales. Chimaeras swim by flapping their pectoral fins and by undulating their bodies. All chimaeras are egg-layers, the egg being protected by a horny shell. Adult chimaeras range in size from 60 to 200 cm. Chimaeras are cool-water, marine fishes that live in shallow to moderate depths between 80 and 2600 m, where they usually swim just above the bottom. Chimaeras eat predominantly hard-bodied benthic invertebrates, which they crush with their tooth plates. Surprisingly little is known about their general biology and natural history.

E. Bony Fishes

Modern bony fishes, often referred to as Ostechthyes (literally “bony fishes”), consist of seven major taxonomic groups. The first six, primitive taxa belong to groups that were much more diverse during Paleozoic and Mesozoic eras. Several of these primitive but modern fishes are classified as “bony” even though they have cartilaginous skeletons. Their skeletal condition is actually an advanced, specialized trait, their immediate ancestors were bony.

Grade Teleostomi (more commonly Ostechthyes)
Class Sarcopetergi
Sub- (or infra-) class Dipnoi
Order Ceratodontiformes: 1 family and 1 species, the Australian lungfish
Order Lepidosireniformes: 2 families, 5 species, South American and African lungfishes
Subclass Coelacanthimorpha
Order Coelacanthiformes: 1 family and 2 species, the coelacanths

Class Actinopterygii
Subclass Chondrostei
Order Acipenseriformes: 2 families, 26 species, sturgeons and paddlefishes
1. Lungfishes

Lungfishes are anadromous (they swim in fresh water) actinopterygians (ray-finned) fishes. They have paired lungs; have larvae with gills; and lack scales. The most primitive actinopterygian (ray-finned) fishes is the bowfin (3.8 m, 630 kg). The world’s largest freshwater fish is the beluga sturgeon of eastern Europe and Asia, Huso huso, at 8.6 m and 1300 kg. The most primitive actinopterygian (ray-finned) fishes are the chondrostean sturgeons and paddlefishes. All are the order Acipenseriformes (= Actinopterygii). All are fished mercilessly. A spawning female can be worth thousands of dollars for her caviar alone, and many sturgeon stocks have been reduced 99% from historical levels. The shortnose sturgeon of Atlantic coastal rivers is listed as endangered, its young lack external gills. It can, but does not have to, breathe atmospheric oxygen. Its young lack external gills. Neoceratodus populations have declined dramatically and species recovery efforts include transplantation into several Queensland reservoirs and rivers. The one South American and four African lungfishes have eel-like bodies; slender, almost-filamentous paired fins; lack scales; have paired lungs; have larvae with external gills; and must breathe air to survive. The four African species occur across central and south Africa, often in swampy areas that frequently experience drought. When a swamp dries up, African lungfishes dig a burrow and can wait 4 years for rains to return. The South American species occurs in swampy regions of the Amazon and Para river basins. Comparatively little is known about its biology.

2. The Coelacanths

Coelacanths were thought to have gone extinct at the end of the Cretaceous, 65 million years ago, until a live one was trawled up in 1938 off South Africa (Fig. 4B). Today, a small, endangered population of 200–600 coelacanths lives at 100- to 500-m depths off two small volcanic islands in the Comore Archipelago, between Madagascar and Mozambique. In 1998, another species was discovered at similar depths in northern Indonesia. The living coelacanths have fleshy pectoral, pelvic, anal, and second dorsal fins (= the lobed fins that define the class Sarcopterygii); a symmetrical, three-lobed tail with a central extension; hollow neural spines (hence “coelacanth” or “hollow spines”); a unique unconstricted notochord; a joint in the dorsal braincase that aids jaw opening; relatively large, thick, bony scales, and live young. Coelacanths are large (to 180 cm, 95 kg), old (probably 40–300 years!), and produce relatively few, live young (5–26 young per clutch). The gestation period is about 13 months. Replacement rate in the population is therefore slow. Coelacanths are captured primarily as by-catch in the hook-and-line fishery for oilfish, and it is unlikely they can sustain even the current by-catch rate of 3–10 animals per year. Counts from small submarine trawling indicate the Comoran species is declining. The Comoran government has outlawed its capture, and trade in coelacanths is outlawed by the Convention on the International Trade in Endangered Species (CITES).

3. Sturgeons and Paddlefishes

The most primitive actinopterygian (ray-finned) fishes are the chondrostean sturgeons and paddlefishes. All 24 species of sturgeons live in the Northern Hemisphere. All spawn in fresh water, although some species move seasonally between marine and fresh water. North American freshwater species include the lake sturgeon and three river sturgeons. Anadromous species include the Atlantic and white sturgeons, the latter being an occupant of west coast bays and rivers. White sturgeons attain the largest size of any North American freshwater fish (3.8 m, 630 kg). The world’s largest freshwater fish is the beluga sturgeon of eastern Europe and Asia, Huso huso, at 8.6 m and 1300 kg. Sturgeons have four barbels ahead of a ventrally located mouth, five rows of large bony shields on an otherwise scaleless body, and a heterocercal tail (Fig. 4C). They are exceptionally long-lived (118 years for beluga and 70–80 years for white sturgeon), mature slowly (as late as 30 years old), and spawn infrequently (every 3–5 years). They migrate up rivers to spawn in clean sand and gravel areas; hence, dam building and siltation of rivers both impede their reproduction. Coelacanths are captured primarily as by-catch in the hook-and-line fishery for oilfish, and it is unlikely they can sustain even the current by-catch rate of 3–10 animals per year. Counts from small submarine trawling indicate the Comoran species is declining. The Comoran government has outlawed its capture, and trade in coelacanths is outlawed by the Convention on the International Trade in Endangered Species (CITES).
Chinese paddlefish. Paddlefishes are water column swimmers that feed on zooplankton and fishes. The paddle of the North American species may help direct food into its mouth, but abundant electroreceptor cells on the paddle suggest an additional, unknown function. North American paddlefish may live 30 years and grow to be 2.2 m in length and 83 kg in weight. Late maturation (at 10 years of age), infrequent spawning (every...
2–5 years), lack of clean gravel spawning habitat, and overfishing have all contributed to population reductions and range contraction. The exceedingly rare and poorly studied Chinese paddlefish is larger, reaching more than 3 m in length. This paddlefish may be the most endangered fish in China because of overfishing, habitat destruction, pollution, and dam construction that blocks spawning migrations in the Yangtze River basin where it lives.

4. Bichirs
The polypteriform bichirs and ropefish of west and central tropical Africa are difficult to assign to any particular taxonomic group. Their gilled larvae, lobe-like fins, thick ganoid scales, and modified heterocercal tail suggest affinities with several living groups, particularly the chondrostean sturgeons and paddlefishes. However, their unique dorsal, caudal, and paired fins and unusual chromosomes place them apart from all extant groups. Taxonomists usually place them either in the Chondrostei or in their own subclass, the Brachiopterygii.

Ten species of polypteriforms are called bichirs (Polypterus) and an 11th, elongate species is the reedfish, Erpetoichthyes. Bichirs reach 120 cm in length and ropefish 90 cm. All are predatory and inhabit shallow, swampy regions, in which they breathe atmospheric oxygen with their paired lungs. Bichirs, also called "flagfins," have unique dorsal and pectoral fins. Each dorsal finlet consists of a vertical spine with attached horizontal rays, looking like a flagpole with streaming banners. In more usual ray-finned fishes, dorsal fin rays emerge vertically from the body of the fish. The pectoral fin has an internal arrangement that involves a wishbone-shaped, flattened plate, again unlike that in any other fish. Bichirs are not particularly well studied; no evidence exists to suggest they are imperiled.

5. Gars
The lepisosteiform gars are seven species of elongate, predaceous fishes that occur in eastern North America and Central America (Fig. 4D). They typically inhabit backwater areas of lakes and rivers, such as oxbows and bayous, and breathe atmospheric oxygen using a highly vascularized gas bladder. Gars are bony skeletons, but their vertebral centra are unique, being convex anteriorly and concave posteriorly. In most fishes, the vertebrae are concave on both surfaces. Gars have an abbreviated heterocercal tail and hinged, diamond-shaped, interlocking ganoid (heavy and bony) scales. Gars are the only freshwater fish in North America with poisonous eggs. Alligator gars can be 3 m long and weigh 140 kg. In recent years, alligator gars have come under intense commercial fishing, and concern for their well-being is increasing.

6. The Bowfin
The bowfin, Amia calva, is the only living member of its genus, family, and order. The bowfin has the abbreviated heterocercal tail and spiral valve intestine of the gars but also has teleoste-like biconcave vertebrae as well as cycloid scales, a relatively high scale type also possessed by many teletosts. The bowfin's head is exceptionally bony and the throat is covered by a distinctive large bone, the gular plate. Bowfin swim slowly forwards or backwards by passing undulations back and forth along their long dorsal fin. Bowfin occur throughout much of the eastern North America in backwater, often swampy areas; they also have a highly vascularized gas bladder which functions as a lung. They are relatively large and robust (to 1 m and 9 kg) and predatory on anything that moves. Bowfin males guard the young vigorously until they are relatively large (10 cm).

7. Teleosts
The division Teleostei (‘perfect bone’) contains most living fishes. Teleosts are not only taxonomically diverse but also ecologically diverse, occupying every aquatic habitat type and niche imaginable. The 23,600 living teleostean species are placed in 4064 genera, 426 families, and 38 orders. This incredible diversity is generally organized into four taxonomic subdivisions that reflect patterns of evolution that date back to the Mesozoic. These four main subdivisions are the osteoglossomorphs (bony tongues), elopomorphs (eels and tarpons), clupeomorphs (herrings), and the euteleosts, with the latter group containing the vast majority of modern bony fishes.

a. Class Actinopterygii
Subclass Neopterygii
Division Teleostei
Subdivision Osteoglossomorpha
Order Osteoglossiformes: 6 families, 217 species, including bonytongues, mooneyes, African knifefishes, and elephant fishes.
Osteoglossiforms derive their name “bonytongue” from the teeth on their tongue that forms part of their bite. These freshwater fishes occur on all major continents except Europe. The arapaima of South America is one of the world’s largest freshwater fishes, reaching a length of 2.5 m. The African mormyrid elephant fishes produce and detect weak electric fields, have large cerebellums, and have a brain size: body weight ratio comparable to that of humans (Fig. 5A). Two species, the moomeye and goldeye, occur in major river systems of northern North America.

b. Subdivision Elopomorpha

Order Elopiformes: 2 families, 8 species, ladyfishes and tarpons
Order Albuliformes: 3 families, 29 species, including bonefishes and spiny eels
Order Anguilliformes: 15 families, 738 species, including freshwater, moray, cutthroat, and conger eels
Order Saccopharyngiformes: 26 species, including swallow and gulper eels

Elopomorphs all have ribbon-shaped “leptocephalus” larvae. The Atlantic tarpon is a highly prized gamefish that reaches a length of 2.5 m and a mass of 150 kg (Fig. 5B). Albuliform bonefishes are also popular gamefishes that occupy sandy flats in shallow tropical waters. The 15 families of anguilliform “true” eels are distinguished from the approximately 45 other families of “eel-like” fishes that have independently evolved an elongate body. Anguillid eels are catadromous, spawning at sea but spending most of their lives in fresh water. Muraenid moray eels and their relatives are marine, tropical and warm temperate, predatory species. Synaphobranchid cutthroat eels include an endoparasitic species, the snubnose parasitic eel, which has been found in the heart of a mako shark. The saccopharyngiform deep-sea gulper and swallow eels have giant mouths but lack many head bones, scales, and fins found in most other fishes.

c. Subdivision Clupeomorpha

Order Clupeiformes: 5 families, 357 species, including anchovies and herrings

Clupeomorphs are small, schooling, silvery, pelagic marine and occasionally freshwater feeders on zooplankton and phytoplankton. Herrings, round herrings, shads, alewives, sprats, sardines, pilchards, and menhaden are extremely important commercial species. Anchovies range in size from a 2-cm Brazilian species to a piscivorous, riverine, 37-cm New Guinea anchovy. The largest clupeids are the Indo-Pacific chirocentrid wolf herrings, which reach a length of 1 m and have
FISH, BIODIVERSITY OF

fang-like jaw teeth. Anadromous shads, alewives, and herrings occasionally establish landlocked populations in rivers, lakes, and reservoirs.

d. Subdivision Euteleostei

Approximately 95% of the teleosts are placed in the Euteleostei or “true” teleosts. This subdivision contains 391 families, 3795 genera, and 22,262 species—all in nine superorders.

i. Superorder Ostariophysi

Order Gonorynchiformes: 4 families, 35 species, including milkfish
Order Cypriniformes: 5 families, 2660 species, including minnows, barbs, algae eaters, suckers, and loaches
Order Characiformes: 10 families, 1340 species, including freshwater hatchet fishes, tetrar, and characins
Order Siluriformes: 34 families, 2405 species, including North American freshwater, airbreathing, electric, sea, upside-down, parasitic, callichthyid armored, and suckermouth armored catfishes
Order Gymnotiformes: 6 families, 62 species, including glass, ghost, and naked-back knifefishes and electric eel.

Rivers, lakes, and streams worldwide are dominated numerically and ecologically by members of the superorder Ostariophysi. Ostariophysans include milkfish, minnows, carps, barbs, suckers, loaches, pranathas, tetras, catfishes, and electric eels. Two distinctive traits unite this otherwise disparate assemblage: (i) the Weberian apparatus, which is a series of modified anterio vertebral that link the gas bladder to the inner ear and aid in hearing, and (ii) production and reaction to chemical alarm substances that are released when a fish is injured and lead to a stereotyped escape response in school members.

The gonorynchiform milkfish, Chanos chanos, is an important food fish in the Indo-Pacific region and is often cultured in brackish fishponds. The Cypriniformes make up the largest order in the superorder. The Cyprinidae is the largest family of freshwater fishes and contains 2000 species of minnows, shiners, carps, barbs, barbels, gudgeons, chubs, dace, squawfishes, tench, rudd, bitterlings, bream, southeast Asian “sharks” (redtail black shark and bala shark), goldfish, koi (domesticated common carp), danios, and rasboras. Cyprinids are most diverse in Southeast Asia, followed by Africa, North America (270 species), and Europe, but are missing from South America and Australia. The world’s smallest freshwater fish is a Burmese cyprinid, Danionella translucida, which matures at 10 mm. The largest minnow in North America is the endangered, piscivorous Colorado squawfish, Ptychocheilus lucius.

The third largest family, the Cyprinidae, contains 391 species, including many popular aquarium fishes such as the kuhli, clown, and skunk loaches, the weatherfishes, and the golden dojo. Weatherfishes (Misgurnus) become restless when barometric pressure decreases preceding a storm.

ii. Superorder Euteleostei

This subdivision contains some 391 families, 3795 genera, and 22,262 species—all in nine superorders.

Order Gymnotiformes: 6 families, 62 species, including electric eel.

The gymnotiform South American knifefishes are the flathead and blue catfishes at about 1.5 m and 50–68 kg. Some small catfishes are notable, such as the parasitic catfishes (Trichomycteridae) of South America, which normally parasitize the gills of fishes but are known to swim up the urethra of bathers and lodge there, necessitating surgical removal.

The gonorhynchiform milkfish, Chanos chanos, is an important food fish in the Indo-Pacific region and is often cultured in brackish fishponds. The Cypriniformes make up the largest order in the superorder. The Cyprinidae is the largest family of freshwater fishes and contains 2000 species of minnows, shiners, carps, barbs, barbels, gudgeons, chubs, dace, squawfishes, tench, rudd, bitterlings, bream, southeast Asian “sharks” (redtail black shark and bala shark), goldfish, koi (domesticated common carp), danios, and rasboras. Cyprinids are most diverse in Southeast Asia, followed by Africa, North America (270 species), and Europe, but are missing from South America and Australia. The world’s smallest freshwater fish is a Burmese cyprinid, Danionella translucida, which matures at 10 mm. The largest minnow in North America is the endangered, piscivorous Colorado squawfish, Ptychocheilus lucius.

Gymnotochelid algae eaters scrape algae off rocks in swift, flowing waters. The catostomid suckers (i.e., bul-faloes, quillback, carpsuckers, blue sucker, redhorse, jumprocks, and the extinct barelip sucker) include 70 species of North American fishes, with 1 species in eastern China. Loaches (Cobitidae) are 110 species of predominantly Eurasian stream fishes, including popular aquarium fishes such as the kuhli, clown, and skunk loaches, the weatherfishes, and the golden dojo. Weatherfishes (Misgurnus) become restless when barometric pressure decreases preceding a storm.

The characiforms are a speciose group of primarily tropical ostariophysans characterized (usually) by a rayless adipose fin and mouths armed with replacement dentition (e.g., piranhas). Body size ranges from very small (13 mm) tetras to large (1.5-m-long) tiger-fishes. Numerous aquarium fishes are included (headstanders, freshwater hatchet fishes, blind characins, pencil fishes, tetras, and silver dollars), as well as important food fishes (Prochilodus, Colossoma, and Brycon). Most characins (ca. 1130 species) are South American, about 200 are African, a few live in Central America, and 1 species, the Mexican tetra Astyanax mexicanus, extends naturally into southwestern Texas.

Catfishes (Siluriformes) are surprisingly diverse, with 34 families and more than 2400 species (Fig. 6A). Catfishes usually have barbels (“whiskers”) and sometimes toxic spiny fins, and they are almost entirely freshwater, nocturnal, and benthic. Catfishes are most diverse in South America (e.g., loricariid suckermouth armored catfishes, 330 species; pimelodid long-whiskered catfishes, 300 species). Large species include the European wels (5 m, 330 kg), the Asian Mekong catfish (3 m, 300 kg), and a 3-m-long whiskered pimelodid of South America. The largest catfishes in North America are the flathead and blue catfishes at about 1.5 m and 50–68 kg. Some small catfishes are notable, such as the parasitic catfishes (Trichomycteridae) of South America, which normally parasitize the gills of fishes but are known to swim up the urethra of bathers and lodge there, necessitating surgical removal.

The gymniform South American knifefishes are unusual ostariophysans that produce and receive weak electric impulses. They have elongate, compressed bodies; an extremely long anal fin; and electrogenic tissue usually derived from modified muscle cells. Their electrical output is constant at high frequencies, whereas the osteoglossomorph mormyrids produce a pulsed,

Euteleostei or “true” teleosts. This subdivision contains 391 families, 3795 genera, and 22,262 species—all in nine superorders.

Order Gymnotiformes: 6 families, 62 species, including electric eel.

The gymnotiform South American knifefishes are the flathead and blue catfishes at about 1.5 m and 50–68 kg. Some small catfishes are notable, such as the parasitic catfishes (Trichomycteridae) of South America, which normally parasitize the gills of fishes but are known to swim up the urethra of bathers and lodge there, necessitating surgical removal.

The gonorhynchiform milkfish, Chanos chanos, is an important food fish in the Indo-Pacific region and is often cultured in brackish fishponds. The Cypriniformes make up the largest order in the superorder. The Cyprinidae is the largest family of freshwater fishes and contains 2000 species of minnows, shiners, carps, barbs, barbels, gudgeons, chubs, dace, squawfishes, tench, rudd, bitterlings, bream, southeast Asian “sharks” (redtail black shark and bala shark), goldfish, koi (domesticated common carp), danios, and rasboras. Cyprinids are most diverse in Southeast Asia, followed by Africa, North America (270 species), and Europe, but are missing from South America and Australia. The world’s smallest freshwater fish is a Burmese cyprinid, Danionella translucida, which matures at 10 mm. The largest minnow in North America is the endangered, piscivorous Colorado squawfish, Ptychocheilus lucius.

Gymnotochelid algae eaters scrape algae off rocks in swift, flowing waters. The catostomid suckers (i.e., bul-faloes, quillback, carpsuckers, blue sucker, redhorse, jumprocks, and the extinct barelip sucker) include 70 species of North American fishes, with 1 species in eastern China. Loaches (Cobitidae) are 110 species of predominantly Eurasian stream fishes, including popular aquarium fishes such as the kuhli, clown, and skunk loaches, the weatherfishes, and the golden dojo. Weatherfishes (Misgurnus) become restless when barometric pressure decreases preceding a storm.

The characiforms are a speciose group of primarily tropical ostariophysans characterized (usually) by a rayless adipose fin and mouths armed with replacement dentition (e.g., piranhas). Body size ranges from very small (13 mm) tetras to large (1.5-m-long) tiger-fishes. Numerous aquarium fishes are included (headstanders, freshwater hatchet fishes, blind characins, pencil fishes, tetras, and silver dollars), as well as important food fishes (Prochilodus, Colossoma, and Brycon). Most characins (ca. 1130 species) are South American, about 200 are African, a few live in Central America, and 1 species, the Mexican tetra Astyanax mexicanus, extends naturally into southwestern Texas.

Catfishes (Siluriformes) are surprisingly diverse, with 34 families and more than 2400 species (Fig. 6A). Catfishes usually have barbels (“whiskers”) and sometimes toxic spiny fins, and they are almost entirely freshwater, nocturnal, and benthic. Catfishes are most diverse in South America (e.g., loricariid suckermouth armored catfishes, 330 species; pimelodid long-whiskered catfishes, 300 species). Large species include the European wels (5 m, 330 kg), the Asian Mekong catfish (3 m, 300 kg), and a 3-m-long whiskered pimelodid of South America. The largest catfishes in North America are the flathead and blue catfishes at about 1.5 m and 50–68 kg. Some small catfishes are notable, such as the parasitic catfishes (Trichomycteridae) of South America, which normally parasitize the gills of fishes but are known to swim up the urethra of bathers and lodge there, necessitating surgical removal.

The gymniform South American knifefishes are unusual ostariophysans that produce and receive weak electric impulses. They have elongate, compressed bodies; an extremely long anal fin; and electrogenic tissue usually derived from modified muscle cells. Their electrical output is constant at high frequencies, whereas the osteoglossomorph mormyrids produce a pulsed,
FIGURE 6  (A) Callichthyid armored catfish (reproduced with permission from Burgess, 1989); (B) chinook salmon (female above and male below) (reproduced with permission from Scott and Crossman, 1973); (C) tripodfish (reproduced with permission from Helfman et al., 1997); (D) lantern fish (reproduced with permission from Nafpaktitis et al., 1977).
low-frequency output. Both groups detect objects that disrupt their electric fields. The 2-m-long electric eel (Electrophorus) produces a weak field for electrolocation and strong pulses for stunning prey or deterring predators.

ii. Superorder Proteocephalopterygii

Order Esociformes: 2 families, 10 species, pikes and mudminnows
Order Osmeriformes: 13 families, 236 species, including barbeleyes, smelts, salmonades, and galaxiids
Order Salmoniformes: 1 family, 66 species, whitefishes, graylings, char, trout, and salmons

Proteocephalopterygians are a mixed agglomeration of marine, freshwater, and diadromous fishes. Esociform pikes, pickerels, and mudminnows are Northern Hemisphere predators; the northern pike has the largest geographical distribution of any Northern Hemisphere fish, occurring across the northern portions of North America, Europe, and Asia. Mudminnows can survive winters in high-latitude lakes by breathing from air bubbles trapped under the ice. Osmeriforms are small, silvery, elongate, water-column dwelling fishes. Osmerids include commercially important species such as capelin, eulachons, Asian ayu, and smelts. The order also includes the Southern Hemisphere lepidogalaxiid salamander fish and the galaxiids. Salmonades inhabit seasonal ponds of southwestern Australia, burying in drying mud and reemerging with the next rains. Galaxiids have suffered numerous extinctions and extinctions as a result of the stocking of nonnative trouts.

Salmoniforms are important commercially, ecologically, and aesthetically. Whitefishes and ciscoes are relatively large-scaled, zooplanktivorous salmonids of high-latitude North American and Eurasian lakes. Several North American species have been decimated due to introduced predators, competitors, and parasitic lampreys. Graylings are riverine fishes with a flowing dorsal fin. At least one species, the Michigan grayling, is extinct. The subfamily Salmoninae contains seven Eurasian and North American genera. The Siberian taimen, Hucho taimen, is the world’s largest salmonid at 2 m and 70 kg. North American Salmoniniae include the char (lake, brook, and bull trout, Arctic char, and dolly varden). Arctic char live farther north than any other freshwater fish. The remaining salmonines are the Atlantic basin salmon and trout (e.g., Atlantic salmon and European brown trout), and the 11 species of Pacific basin trout and salmon in the genus Oncorhyncus, 2 of which are endemic to Japan (Fig. 6B). Pacific trout and salmons include golden, cutthroat, and gila trouts and the spectacularly anadromous coho, chinook, chum, pink, and sockeye salmon, some of which undergo oceanic migrations of thousands of kilometers before returning to their birth river to spawn and die. The actual number of genetically distinct races of Pacific salmonids is unknown because many stocks are reproductively isolated in small river systems. Evidence suggests that as many as 1000 stocks exist, 100 of which have gone extinct and an additional 314 of which are imperiled.

iii. Superorder Stenopterygii

Order Stomiiformes: 4 families, 321 species, including bristlemouths, marine hatchet fishes, and barbeled dragonfishes
Order Ateleopodiformes: 1 family, 12 species, jellynose fishes

Stenopterygians are deep-sea fishes, often with long teeth and large mouths. Gonostomatid bristlemouths may be the most abundant and widely distributed vertebrates on Earth. Idiacanthine black dragonfishes have a larva with eyes at the ends of elongate stalks.

iv. Superorder Cyclosquamata

Order Aulopiformes: 13 families, 219 species, including telescope fishes, tripod fishes, lizard fishes, and lancet fishes

Cyclosquamates are also deep-sea forms, including the bizarre giganturid telescope fishes with large tubular eyes, a huge mouth, flexible teeth, and an expandable stomach. Deep-sea tripod fishes have long pectoral, pelvic, and caudal rays that they use for resting on soft sediments of the deep ocean floor (Fig. 6C). Shallow representatives are the synodontid lizard fishes, which are common benthic predators on coral reefs worldwide. Alepisaurid lancet fishes are large (to 2 m) mesopelagic predators with a sail-like dorsal fin of unknown function.

v. Superorder Scopelomorpha

Order Myctophiformes: 2 families, 240 species, including lantern fishes

Scopelomorphs include the abundant, commercially important lantern fishes, which are identified based on species-specific photophore (light organ) patterns. Lantern
tern fishes occur at middle depths from the Arctic to the Antarctic. They are important in the diets of many fishes as well as of marine mammals (Fig. 6D).

vi. Superorder Lampridiomorpha

Order Lampridiformes: 7 families, 19 species, including opahs, tube-eye, ribbonfishes, and oarfishes

Lampridiforms are generally open-water, oceanic fishes. Opahs are relatively large (1.8 m, 70 kg), oval-shaped, pelagic predators on squids and other fishes. The 30-cm-long tube-eye (*Stylephorus*) can increase the volume of its mouth 40-fold during feeding—a record among vertebrates. The elongate oarfish, *Regalecus*, may attain 12-m length and is the longest living teleost. It has a bluish-silvery body, scarlet head crest, and deep red fins. It is thought to be responsible for many “sea serpent” sightings.

vii. Superorder Polymixiomorpha

Order Polymixiiformes: 1 family, 5 species, beardfishes

viii. Superorder Paracanthopterygii

Order Percopsiformes: 3 families, 9 species, trout-perches, pirate perch, and cavefishes

Order Ophidiiformes: 5 families, 355 species, including pearlfishes, cusk-eels, and viviparous brotulas

Order Gadiformes: 12 families, 482 species, including rattails, hakes, and cods

Order Batrachoidiformes: 1 family, 69 species, toadfishes

Order Lophiiformes: 16 families, 297 species, including goosefishes, frogfishes, handfishes, batfishes, and deep-sea anglerfishes

Paracanthopterygians are primarily benthic, marine, nocturnally active fishes; many live in the deep sea or in caves. Percopsiforms are small (≤20 cm), freshwater fishes, most of which live in eastern North America. The anus of the swamp-dwelling aphrododerid pirate perch is located in the throat region of adults for functionally mysterious reasons. Amblystopid cavefishes are often blind and scaleless forms highly adapted for cave life.

Ophidiiforms often live in holes or even inside other animals. Carapid pearlfishes live inside the body cavities of starfishes, sea cucumbers, clams, and sea squirts; some feed on the internal organs of their hosts. Ophidiod and bythiid cusk eels and brotulas include blind cave species in freshwater systems of Caribbean and Galapagos Islands as well as coral reef species that hide deep within crevices. The neobythitine cusk eel, *Abyssobrotula galatheae*, holds the depth record for a fish at 8370 m in the Puerto Rico Trench.

The gadiforms include the cods, haddocks, hakes, pollocks, and whittings, which are some of the world’s most important commercial fishes. True cods (*Gadinae*) have three dorsal fins and two anal fins. Many species have chin barbels. The burbot, *Lota lota*, of high-latitude, Northern Hemisphere lakes is the only freshwater species in the group. The commercially important Atlantic cod is the largest species (1.8 m, 90 kg), but fish more than 10 kg are rare due to drastic overfishing (Fig. 7A). One of the world’s largest food fisheries is for North Pacific walleye pollock.

Batrachoidiforms include the midshipmen, which have hundreds of photophores, an unusual trait for a shallow dweller. Many batrachoids produce sounds by vibrating their gas bladders. Venomous toadfishes have dorsal and opercular spines which can inject a powerful toxin. Three South American toadfishes are restricted to fresh water.

Lophiiforms are a diverse and often bizarre-looking order of marine fishes that include benthic, shallow-water forms as well as highly modified, open-water, deep-sea forms. Many use a modified first dorsal spine as a lure for catching smaller fish. The meter-long western North Atlantic goosefish, *Lophius americanus*, has a huge mouth with long, recurved teeth which it uses to catch fishes and even diving seabirds. Ammianruid frogfishes also rest on the bottom or walk over it with their pectoral and pelvic fins (Fig. 7B). The ogcocephalid batfishes walk on their pectorals, but they can also swim via jet propulsion by shooting water out their round, backward-facing opercular openings. The ceratioid anglerfishes include 11 families of strange-appearing bathypelagic predators, many of which have very small males that fuse to and become parasitic on the larger females. The endemic Australian handfishes include a Tasmanian species, the spotted handfish, that was once common but is now critically endangered due possibly to egg predation by an introduced starfish.

e. Superorder Acanthopterygii

Most bony fishes belong to a single superorder, the Acanthopterygii, which contains about 13,500 species in 251 families. Two small and one large taxonomic groupings, called series, are recognized, with the vast majority in the third series, the Perciformes.
FIGURE 7  (A) Atlantic cod (reproduced with permission from Norman, 1905); (B) bloody frogfish (reproduced with permission from Heller and Snodgrass, 1903); (C) atherinomorph four-eyed fish (reproduced with permission from Norman, 1905); (D) lined seahorse (reproduced with permission from Norman, 1905).
i. Series Mugilomorpha

Order Mugiliformes: 1 family, 66 species, mullets

The mullets are a family of nearshore, marine and freshwater fishes of considerable economic importance. Many mullets feed on organic silt and minute plants, an unusual food type among fishes.

ii. Series Atherinomorpha

Order Atheriniformes: 8 families, 285 species, including rainbow fishes and silversides
Order Beloniformes: 5 families, 191 species, including needlefishes, flying fishes, and halfbeaks
Order Cyprinodontiformes: 8 families, 807 species, including topminnows, killifishes, livebearers, and pupfishes

Atherinomorphs are shallow-water, marine or freshwater fishes that live near the surface. Many atherinomorphs bear live young. Atheriniforms include the melanotaeniid rainbow fishes of Australia and New Guinea, in which males have brighter colors and longer fins than females, traits that make them popular aquarium species. Silversides are widespread, freshwater and marine schooling fishes and include the grunions of southern and Baja California, which ride waves up beaches on dark nights to spawn in wet sand biweekly during the summer. Beloniforms are predominantly silvery, marine fishes active at and sometimes above the surface of the water. The lower lobe of the tail in flying fishes is relatively long and is used to scull rapidly during takeoff. Many cyprinodontiforms, although basically freshwater fishes, tolerate considerable salinity and hence occur in streams on isolated oceanic islands (Fig. 7C). The rivulines of South America and Africa live only 1 year, laying eggs which survive in the dried bottoms of pools and which hatch with the next season’s rains. Rivulus marmoratus of south Florida and the West Indies is the only truly hermaphroditic fish, fertilizing its own eggs. The livebearers include the mollies, platys, guppies, and swordtails of the aquarium trade. Some species are all female. Many cyprinodontid pupfishes are tolerant of extreme water conditions and consequently can live in saltmarsh and desert conditions. However, they cannot tolerate total desiccation, which has endangered many desert species that have to compete with humans for water. The Devil’s Hole pupfish has the smallest known range of any fish species—one shallow shelf in a single spring in Death Valley, California. Other pupfish relatives inhabit Lake Titicaca, which at an elevation of 4570 m in the Andes Mountains is the highest natural lake with fishes.

iii. Series Percomorpha

Order Stephanocypriniformes: 9 families, 86 species, including whale fishes
Order Beryciformes: 7 families, 123 species, including flashlight fishes, roughies, and squirrelfishes
Order Zeiformes: 6 families, 39 species, including dories and boarfishes
Order Gasterosteiformes: 11 families, 257 species, including sticklebacks, pipefishes, seahorses, trumpet fishes, and shrimp fishes
Order Synbranchiformes: 3 families, 87 species, including scorpion fishes, rockfishes, sea robins, sabletfishes, greenlings, sculpins, tilaikal odesfishes, and lumpfishes
Order Perciformes (9300 species)
Suborder Percoidae: 71 families, 2860 species, including snools, temperate basses, sea basses, centrarchid sunfishes, black basses, darters, perchies, cardinal fishes, bluelines, remoras, dolphin fishes, jacks, pompanos, snappers, grunts, croakers, drums, goat-fishes, archerfishes, butterfly fishes, and angelfishes
Suborder Elassomatoidei: 1 family, 6 species, pygmy sunfishes
Suborder Labroidei: 6 families, 2200 species, including cichlids, surfperches, damselfishes, wrasses, and parrotfishes
Suborder Zoarcoidei: 9 families, 318 species, including eel pouts, gunnels, and wolfishes
Suborder Nototomisti: 5 families, 122 species, including icefishes and Antarctic dragonfishes
Suborder Trachinoidei: 13 families, 212 species, including sand lances, weeverfishes, and stargazers
Suborder Blennioidei: 6 families, 732 species, including clinids and blennies
Suborder Icosteoidei: 1 family and species, ragfishes
Suborder Gobiesocidi: 1 family, 120 species, clingfishes
Suborder Callionymoidei: 2 families, 137 species, dragonets
Suborder Gobioidae: 8 families, 2120 species, including sleepers and gobies
Suborder Scophthrichidae: 1 family, 2 species, nurseryfishes
Suborder Scophthrichidae: 6 families, 125 species, including spadefishes, scats, rabbitfishes, moomish idol, and surgenfishes
Suborder Scrombrolabracoidae: 1 family and species, the scrombrolabracid
Suborder Scorpaenidae: 136 species, including barracudas, mackerels, tunas, swordfish, and billfishes
Suborder Stomatoidei: 6 families, 349 species, including driftfishes and butterfishes
Suborder Anabantoidei: 5 families, 81 species, gouramis
Suborder Channidae: 1 family, 21 species, snakeheads
Order Pleuronectiformes: 11 families, 370 species, including flounders and soles
Order Tetraodontiformes: 9 families, 339 species, including triggerfishes, boxfishes, trunkfishes, cowfishes, puffers, porcupine fishes, burrfishes, and ocean sunfishes

The percomorphs constitute by far the largest taxonomic group of fishes, far too many to deal with in any detail. What follows is a very brief overview of some of the more interesting orders, suborders, and families.

Beryciforms are shallow- to moderate-depth, often red, almost always nocturnal fishes, including the reef-dwelling squirrelfishes. Also included is the commercially important orange roughy, Hoplostethus atlanticus, of high-latitude, southern ocean regions. Orange roughies are being overexploited because they are slow growing and long-lived, taking more than 20 years to mature and reaching ages more than 100 years. Zeiforms include commercial species such as the European John Dory, Zeus faber.

Gasterosteiforms are small marine, estuarine, and freshwater fishes with dermal armor plating. Sticklebacks are well-studied fishes that frequently form distinct, isolated populations characterized by unusual spines, plates, and behavior. The suborder Syngnathoidei includes the bizarre pegasus seamounts and syngnathid pipefishes, sea dragons, and seahorses. The suborder Syngnathidae includes the bizarre pegasus seamounts and syngnathid pipefishes, sea dragons, and seahorses. Syngnathoidei includes the bizarre pegasus seamounts and syngnathid pipefishes, sea dragons, and seahorses.

The Centrarchidae contains the sunfishes, crappies, rockbasses, and black basses of North America. The sea bass family Serranidae is one of the largest families (450 species) that range in size from small (<5 cm) anthiines and hamlets to giant groupers and jewfish (3 m long, 400 kg). Sea basses also include commercially important hinds, coney, gage, and scamp (Fig. 8B). Many serranids are hermaphroditic, usually starting as female and then later becoming male, although some hamlets are both male and female simultaneously.

The Centrarchidae contains the sunfishes, crappies, rockbasses, and black basses of North America. The sea bass family Serranidae is one of the largest families (450 species) that range in size from small (<5 cm) anthiines and hamlets to giant groupers and jewfish (3 m long, 400 kg). Sea basses also include commercially important hinds, coney, gage, and scamp (Fig. 8B). Many serranids are hermaphroditic, usually starting as female and then later becoming male, although some hamlets are both male and female simultaneously.

The Centrarchidae contains the sunfishes, crappies, rockbasses, and black basses of North America. The sea bass family Serranidae is one of the largest families (450 species) that range in size from small (<5 cm) anthiines and hamlets to giant groupers and jewfish (3 m long, 400 kg). Sea basses also include commercially important hinds, coney, gage, and scamp (Fig. 8B). Many serranids are hermaphroditic, usually starting as female and then later becoming male, although some hamlets are both male and female simultaneously.

The Centrarchidae contains the sunfishes, crappies, rockbasses, and black basses of North America. The sea bass family Serranidae is one of the largest families (450 species) that range in size from small (<5 cm) anthiines and hamlets to giant groupers and jewfish (3 m long, 400 kg). Sea basses also include commercially important hinds, coney, gage, and scamp (Fig. 8B). Many serranids are hermaphroditic, usually starting as female and then later becoming male, although some hamlets are both male and female simultaneously.
reef species. The voracious, schooling pomatomid bluefish, *Pomatomus saltatrix*, occurs in most temperate and semitropical oceans except the eastern Pacific. In echeneid remoras or sharksuckers, the modified first dorsal fin forms a suction disk that is used to cling to various hosts. The coryphaenid dolphin fishes or mahi mahis are two species of open-water, pelagic predators that often associate with floating structure. They are also one of the few marine pelagic fishes that is successfully aquacultured. Carangid scads, jacks, pompanos, and amberjacks are a large family (140 species) of tropical nearshore predators.

Lutjanid snappers include 125 species of generally carnivorous marine fishes. Snappers are usually near-bottom dwellers (i.e., gray snapper, red snapper, and mangrove snapper), but some are water-column zooplanktivores (vermilion snapper).

Haemulidae grunts are moderate-sized coral reef fishes and are most diverse in the New World tropics. They are seen most often in their daytime resting schools around coral heads; at night they disperse to feed in surrounding reef and grass areas. Sciaenid croakers (270 species) have chin barbels and a muscularized gas bladder used for sound production. Sciaenids are a widespread family that is particularly diverse in the southeastern United States, but representatives occur widely in tropical marine and freshwater habitats. Species include red drum (spot tail bass), black drum, croakers, weakfish, sea trout, kingfishes, white sea bass, corbinas, and the endangered Mexican totoaba, one of the few unquestionably imperiled marine fishes. The range of the freshwater drum, *Aplodinotus grunniens*, includes much of eastern North America into Central America. Another bottom-oriented family is the tropical reef-dwelling mulloid goatfishes, which have movable, muscularized chin barbels. Toxotid archerfishes are Indo-Pacific, brackish-water fishes well-known for their ability to shoot droplets of water that knock insects out of overhanging vegetation.

Two closely related, colorful reef families are the chaetodontid butterfly fishes (114 species) and the po-macanthid angelfishes (74 species). Both families are most diverse in the Indo-Pacific region. Butterfly fishes feed on coral polyps, small invertebrates, tube worms, or zooplankton, whereas angelfishes eat attached invertebrates such as sponges, tunicates, and anthozoans (Fig. 8D).

Elassomatoid pygmy sunfishes are a suborder of minia-ture (20- to 45-mm), colorful freshwater swamp dwellers of the southeastern United States. Several very speciose families belong to the suborder Labroides. Tropical marine families include the po-macanthid damselfishes (313 species), which are small, colorful, usually herbivorous, territorial reef dwellers. Some are zooplanktivores (*Chromis* and anemonefishes). Several occupy temperate regions (e.g., the garibaldi of California). The largest labroid family is the mostly tropical, reef-dwelling wrasses, Labridae.
FISH, BIODIVERSITY OF

(500 species). Wrasses range in size from the 5-cm-long cleaner wrasses, slippery dicks, and blueheads to the 2.3-m-long Maori wrasse of the Indo-Pacific, which is hunted unmercifully for the live-fish restaurant trade. Cool temperate species include the California sheephead (Fig. 9A) and senorita, western Atlantic tautog and cunner, and the eastern Atlantic cuckoo wrasse. Many wrasses change sex from female to a more colorful male. The scarid parrotfishes (83 species) are almost exclusively coral reef dwellers, best known for their fused parrot-like teeth that are used for biting off algal and coral pieces, which are then crushed in the massive pharyngeal (throat) jaws. The embiotocid surfperches are 24 species of deep-bodied, temperate (mostly eastern Pacific) fishes associated with kelp beds and rocky reefs. They are live-bearers, feeding on zooplankton or small invertebrates. The largest labroid family is the freshwater cichlids, with more than 1300 species. Cichlids are chiefly tropical, South American and African fishes, with a few species that occur further north (the Rio Grande cichlid is found in south Texas). Central and South American species include freshwater angelfishes, discus, oscars, convict cichlids, and peacock bass (Fig. 9B). Most cichlids occur in Africa, where they are particularly speciose in the African Great Lakes and are threatened by introduced predators such as Nile perch. African tilapias and other cichlids have been deliberately or accidentally introduced into Florida, California, and Hawaii as a by-product of aquaculture activities.

![Figure 9](A) California sheephead (reproduced with permission from Norman, 1905); (B) blue-eye cichlid (reproduced with permission from Greenfield and Thomerson, 1997); (C) Japanese goby (reproduced with permission from Norman, 1903); (D) four-spot flounder (reproduced with permission from Bigelow and Schroeder, 1953); (E) gray triggerfish (reproduced with permission from Bigelow and Schroeder, 1953).
Fishes of the suborder Zoarcoidae are all eel-like, bottom-living, marine, cool-to-cold-water species. They range in size from the small, intertidal pricklebacks and gunnels to the live-bearing eel pouts, some of which live 3000 m below the surface. The large (to 2.5 m long) anarhichadid wolfishes and wolf eels of shallow North Pacific and Atlantic waters are anatomically and ecologically similar to moray eels.

Icefishes (suborder Notothenioidei) are mostly Antarctic, mostly benthic fishes that live under the ice and have antifreeze compounds in their blood. The crocodile icefishes lack red blood cells, hemoglobin, and myoglobin, and hence have colorless blood and flesh.

Trachinoids are marine, generally benthic fishes with a tendency to bury themselves in sand. Ammodoid sand lances are small, elongate, and abundant zooplankton feeders that spend their nights buried. Trachinid weeverfishes occur in the eastern Atlantic and Mediterranean and have venomous opercular and dorsal spines. Uranoscopid stargazers emit strong pulses of electricity (up to 30 V) from highly modified eye muscles.

Blennioids are small marine fishes that usually associate with structure. Chaenopsid pikeblennies and tubeblennies often live inside corals and worm tubes. Combsnout blennies are diverse (345 species) small fishes in tropical and subtropical waters; they scrape algae with their comb-like teeth.

The suborder Icosteoidei consists of the peculiar North Pacific ragfish, Icosteus aenigmaticus, which has a largely uncalcified, cartilaginous skeleton and is a preferred food of sperm whales. The small, marine and freshwater gobioid clingfishes are shallow-water species that are convergent in many ways with some butterfly fishes, a description that also applies to the moorish idol (Zanclidae). The 72 species of usually herbivorous acanthurid surgeonfishes, unicornfishes, and tangs have a knifeblade on the caudal peduncle.

The suborder Scombroidei contains some of the largest and most spectacular marine fishes. Twenty species of barracudas inhabit tropical and subtropical oceans almost worldwide. The gempylid snake mackerels (23 species) are pelagic and deep-water predators, including the cosmopolitan oilfish, Ruvettus pretiosus, a large (1.8 m, 45 kg) predator of moderate depths. An active fishery for oilfish in the Comoro Islands captures endangered coelacanths as by-catch. The scombrid mackerels and tunas are quintessential open-sea predators, with streamlined bodies and a physiology geared to a high-speed lifestyle. They range from relatively small, 50-cm mackerels to giant bluefin tuna (4 m, 500 kg). Most are schooling fishes of tremendous commercial importance. The temperate and warm-temperate xiphiid swordfish and the more tropical istiophorid sailfishes, spearfishes, and marlins have an elongate upper jaw bone that forms the bill. It is used as a spear, a cutlass, or a billy. Swordfish grow to 530 kg, whereas blue and black marlin grow to 900 kg. Swordfish have been heavily overfished, particularly in the Atlantic.

Labyrinth fishes (Suborder Anabantoidae) have an auxiliary breathing structure in the gill chamber for aerial respiration. Anabantid climbing gouramis are African and Asian freshwater fishes that can move across wet ground and reportedly up wet tree trunks. The kising gourami is the sole member of the family Helostomatidae. The belontiid gouramis, fighting fishes (betta-tas), and paradise fishes have elongate pelvic fin rays that serve as feelers. Bettas (Siamese fighting fish) have been bred to battle like fighting cocks, placing them among the few fishes that have been cultured for purposes other than food, appearance, or research.

Pleuronectiform flatfishes are distinctive, compressed, benthic fishes that have both eyes on the same side of the head (Fig. 9D). Many flatfishes are important commercially (e.g., dab, flounders, halibuts, plaice, sole, tonguefishes, turbots, and whiff). Paralichthyids include the summer flounder and California halibut, the latter reaching 1.5 m and 30 kg. The pleuronectid right-eye flounders include the Atlantic and Pacific halibuts. Pacific halibuts may live 40 years and attain lengths of 3 m and masses of 200 kg. The fishery for Pacific halibut in the North Pacific is a well-regulated, sustainable enterprise.
The most advanced bony fishes are in the order Tetraodontiformes, an almost entirely marine order of medium-sized fishes with thick, leathery skin and with scales often modified into spines or bony plates. In balistoid triggerfishes and filefishes, the long, rigid first dorsal spine is locked erect by an interaction with the shorter, second spine (Fig. 9E). Ostraciid boxfishes are encased in a triangular or rectangular bony box, with just the fins and caudal peduncle emerging. Puffers and ocean sunfishes lack true teeth. Instead, the jaw bone has a cutting edge that looks like separated teeth or is fused into a parrot-like beak. Diodontid porcupinefishes inflate their body by filling the stomach with water, a process that also helps erect and interlock their body spines. Tetraodontid puffers concentrate a powerful and potentially fatal toxin, tetrodotoxin, in their viscera, which adds to the allure of eating puffers in licensed Fugu restaurants in Japan. The ocean sunfish, Mola mola, is one of the world’s heaviest fishes at 1000–2,000 kg, producing as many as 300 million eggs. All this biomass is supported on a diet of jellyfishes.

II. GEOGRAPHIC DIVERSITY

A. Overview

Fishes occur just about everywhere water occurs as long as water is in its liquid state, is available through most of the year, and remains below 40°C. A major zoogeographic distinction can be made between marine and freshwater fishes, with substantial overlap occurring where intermediate salinities occur. Many fishes are restricted to pure fresh water (little or no salinity), many are restricted to normal oceanic salinity (about 35 parts per 1000 salt in water), some occur in both habitats at different times of their lives or of the year, and some occur and are even restricted to areas of intermediate salinity, such as estuaries.

In terms of numbers, about 58% of all fishes are marine and 41% live in fresh water, with the remaining 1% moving regularly between the two salinity designations (Fig. 10 and Box 2). Among the 10,250 freshwater species, 80% are primary or obligatory freshwater fishes and are intolerant of even moderate salinities. The remaining 20% can tolerate some salinity and hence inhabit upper estuarine areas or can cross through nearshore ocean regions to move from one river basin to another. Among the 14,500 marine fishes, the vast majority (69%) live in shallow, warm areas such as coral reefs. The remaining marine species are divided fairly evenly among shallow, cold, deep, open-ocean and deep-bottom areas (about 10% each). About 2% of marine fishes live in near-surface, open-sea (pelagic) habitats. The approximately 160 diadromous species that live in different salinity regions at different times of their lives are divided among three groups. Anadromous fishes (54%) live most of their lives in the ocean but then migrate to fresh water to spawn; this group includes lampreys, sturgeons, herrings, and salmons. Catadrom-
Box 2

Why Are There so Many Freshwater Fishes?
The high global diversity of freshwater fishes (Fig. 10) is at first surprising. Fresh waters comprise only about 0.009% of the earth’s water, which means that almost half of all fish species live in less than 1% of the world’s water. This 7500-fold discrepancy in biodiversity per unit volume is probably best explained by the relative productivity and isolation of freshwater bodies. Most freshwater habitats are relatively shallow and receive ample sunlight and nutrients running off from adjacent land. Hence, freshwater habitats are relatively productive and capable of sustaining abundant life. Most of the ocean, in contrast, is deep, dark, and nutrient poor. Given that 81% of marine diversity occurs in shallow regions, a relationship between water depth and diversity in fresh waters is not surprising. In addition to the influence of available food is the comparative isolation of most freshwater habitats. Lakes are often created and affected by climatic and geologic forces (e.g., drought, floods, landslides, earthquakes, and uplifts) that separate them from other systems. Every lake can be relatively isolated from other lakes, which means that genetically distinct populations can evolve into new species and that little genetic mixing occurs between lakes. Small streams are separated from each other by larger rivers, which are barriers to the movement of small fishes, and large rivers are separated by oceans. As a result, freshwater habitats are perfect for the speciation process. Oceans, in contrast, are largely continuous habitats that are connected by currents, and ocean fishes typically produce larvae that float for several weeks or months on these currents. Hence, genetic exchange is common and opportunities for speciation are not as great. Oceanic basins have relatively distinct faunas, but connectedness within basins discourages the kind of genetic isolation needed for speciation of the sort seen in lakes and rivers.

Anadromous fishes (25%) spend most of their lives in freshwater and migrate to the sea to spawn; included here are river eels, mullets, and temperate basses. Amphidromous fishes (21%) move between fresh and salt water, but migration to the spawning habitat occurs long before the fishes actually spawn; examples include gobies, sleepers, and galaxiids.

B. Freshwater Diversity

The world’s freshwater habitats occur in six major zoogeographic regions or realms that correspond approximately to continental distributions, with important exceptions (Fig. 11). Each region has a fairly distinct fish fauna (again with some exceptions and shared elements).

1. The Nearctic Region

The Nearctic region consists of subtropical, tropical, temperate, and arctic North America. The region stretches from the Mexican Plateau to northern Canada and Alaska. The Nearctic contains 19 families of primary freshwater fishes, with about 950 species. The most diverse families are minnows, suckers, North American catfishes, perches (and darters), and sunfishes. Other important families include the lampreys, gars, salmon (many of which are anadromous), and whitefishes; sculpins, which are freshwater species in a primarily marine family (= “marine derivatives”); pickerels and mudminnows; killifishes; and livebearers. The Nearctic is further subdivided into three subregions: the Arctic-Atlantic (with six provinces), the Pacific (with seven provinces), and the Mexican Transition subdivision. Eleven major river systems drain the region; major lakes are abundant, the largest being the five Laurentian Great Lakes (Ontario, Erie, Huron, Michigan, and Superior).

2. The Neotropical Region

The Neotropical region contains South America and Middle America. It is the most speciose region of the world in terms of freshwater fishes, with 32 families and more than 2500 species. Particularly diverse groups include the colorful characiforms (1200 species of tetras, piranhas, characins, and freshwater hatchets), 13 families and 1300 species of catfishes, 6 families and 62 species of gymnotiform South American electric knifefishes, and 150 species of cichlids. Several secondary freshwater and marine derivative groups are included: freshwater stingrays, herrings, silversides, needlefishes, killifishes, and croakers. Many species remain to be discovered and described, particularly in South America. The Neotropical region has been further divided into eight subdivisions with fairly distinctive faunas. Eight major river systems drain the region; major lakes include Lake Titicaca, the world’s highest fish-containing lake.
3. The Palearctic Region

The Palearctic region encompasses Eurasia, including Europe, northern Africa, and Asia north of the Oriental region. Twenty-seven families and about 550 species of temperate freshwater fishes occur in the region, dominated by minnows and loaches but also perches, pickerels, sturgeons, salmonids, and sculpins (including the Lake Baikal endemics), and 10 species of catfishes in four families. Diversity is greater in the southeastern part of the region and also increases to the south, as is also the case in the Nearctic region. The Palearctic and Nearctic regions share numerous families and genera (sturgeons, paddlefishes, minnows, smelts, salmonids, pikes, mudminnows, and perches) but only a few species (i.e., northern pike, longnose sucker, burbot, threespine stickleback, and fourhorn sculpin). The region is sometimes subdivided into six subregions based on faunal groupings. Ten major river systems drain the region; major lakes include the Black and Caspian Seas and Lake Baikal, the world’s oldest and deepest lake.

4. The African or Ethiopian Region

The African or Ethiopian region is second to the Neotropics in freshwater fish diversity, with 47 families and more than 2000 species of primary and secondary freshwater fishes. The African region includes all the African continent south of the Sahara Desert, plus the large island of Madagascar with its endemic fauna. Half of the fishes are in the superorder Ostariophysi, including 300 minnows, 190 characiforms, and 360 catfishes in six families. Other diverse groups include killifishes and topminnows, elephant fishes and other osteoglossiforms, and cichlids. As many as 1000 cichlid fishes may occur in the three African Great Lakes of Lake Victoria, Lake Tanganyika, and Lake Malawi, with more cichlids in smaller surrounding lakes and rivers. Four lungfishes and all 11 polypteriform bichirs occur in Africa. Ten to 12 zoogeographic provinces are recognized, with six major river drainages, and numerous lakes including the African Great Lakes.

5. The Oriental Region

The Oriental region includes eastern Iran, India and Sri Lanka, China south of the Yangtze River, Southeast Asia, and the large island regions of Taiwan, the Philippines, and the East Indies/Indo-Malayan Archipelago. The Oriental region contains 43 families of primary and secondary freshwater fishes. Most diverse are the minnows, loaches, and 12 families of catfishes; clariid walking catfishes and bagrid catfishes are particularly diverse. Other important groups include algae eaters, river loaches, snakeheads, spiny eels, labyrinth fishes and gouramis, a few cichlids, and archerfishes. The Oriental region shares many families with the Palearctic to the north and the Ethiopian to the west but few with the Australian region to the southeast. The Oriental is
often subdivided into two major subregions: Peninsular India with more than 700 species and Southeast Asia with more than 1000 species. Each subregion has two major river drainage systems; large lakes are uncommon. Southeast Asia is sometimes divided further into five zoogeographic regions.

6. The Australian Region

The Australian region (New Guinea, Australia, New Zealand, and Oceania) is relatively depauperate in true freshwater fishes, and in fact all but three of the freshwater fishes in the region are members of families obviously derived from marine groups. The northwestern border of the region, and the practical limit of primary freshwater fishes, is dramatically delineated by an ocean boundary that lies southeast of Java, Borneo, and Sulawesi and is known as Wallace’s or Weber’s Line. Nineteen families and about 210 species occur primarily in fresh water in the region, but only the Australian lungfish and 2 species of bonytongue saratogas are true freshwater fishes (another 33 families and 150 species of marine fishes frequently enter fresh waters in Australia). Other important families, many with species endemic to specific regions, include lampreys, river eels, herrings, two families of catfishes, southern smelts and graylings, salamander fish, galaxiids, silversides, rainbowfishes, baramundi, grunters, gourami, silurids, perchids, sleeper and gobies, and torrentfishes (in New Zealand). One major river system occurs on New Guinea (the Fly) and two on Australia (Darling and Murray); permanent, large lakes are rare.

C. Marine Diversity

Delimiting zoogeographic regions in the world’s oceans is complicated by depth, currents, and geographic locales; different faunal breaks occur depending on nearshore, pelagic, or deep-sea environments. The greatest fish diversity and the greatest geographic differentiation occur in nearshore, continental shelf (to about 100-m depth) regions. These regions are separated by continents, by large expanses of open ocean, and by currents that differ in temperature from that of the region in question. Temperature zones divide the seas into tropical, temperate, boreal, and polar regions (Fig. 12). In addition, different faunal groupings apply to pelagic fishes and to fishes of the deep sea.

1. The Indo-West Pacific Region

The Indo-West Pacific region includes shallow tropical seas that extend from South Africa and the Red Sea eastwards through the Indo-Malayan area and Australia to Hawaii and Easter Island; it also includes Micronesia, Melanesia, and Polynesia. The Indo-West Pacific is by far the most species-rich marine area, containing 3000 to 4000 tropical fish species, and is similarly diverse in sea snakes and many invertebrate taxa such as reef-building and soft corals, mollusks, tube worms, and echinoderms (Box 3). It is considered the center of evolution for many of the common coral reef fish families that occur in other tropical regions. Only a few families are endemic to the Indo-West Pacific (e.g., sillaginid whirings and rabbitfishes). Common families include moray eels, squirrelfishes, sea basses, grunts,
The Deep Reef

Coral reef fishes—those associated with tropical, shallow waters in which coral reefs grow—are the most diverse group of fishes on the planet. Fish diversity is generally correlated with coral species diversity and coral coverage; 2000 fish species, and many other taxa, occur on the Great Barrier Reef of Australia, on which 500 species of coral also occur. However, most estimates of diversity of coral reef fishes are based on species found only in shallow water less than 50 m deep, where reef-building corals are abundant and where ichthyologists equipped with scuba gear can collect. An underappreciated and barely explored component of the diversity of reef fishes involves species that live in the deeper (50–150 m) “twilight zone” portion of the reef face, which has only recently been made accessible by the use of specialized, mixed-gas rebreathing equipment. The fish fauna of this region consists of species that seldom occur above 75 m. Many new species and even a few new genera have been discovered recently from the deep reef environment in families that are common in shallower water (e.g., wrasses, sea basses, damselfishes, angelfishes, and gobies). Most of the recent collecting using advanced diving techniques has occurred in deep reef areas of only Indo-Polynesia (e.g., Rarotonga, Palau, and New Guinea), but nearly 100 new species have already been found (Fig. 13). If we can extrapolate from the few deep reef areas that have been surveyed, our estimates of the number of fishes inhabiting “coral reefs” will have to be increased considerably. Also, fishes are just part of this barely explored, intermediate depth zone. How much of marine biodiversity are we missing?

Two. The Eastern Pacific Region

The Eastern Pacific region, with approximately 800 fish species, runs from southern Baja California to Ecuador, its northern and southern limits defined by the cold-water California and Peru currents. Despite its location in the Pacific Ocean, the Eastern Pacific is faunistically more similar to the tropical Atlantic, containing many species that are almost indistinguishable from Atlantic forms. The two oceans mixed before the Panamanian isthmus formed and the two areas still share 12 species, despite 3 million years of physical separation. Most families are less diverse here than in the Western Atlantic, with the exceptions of sea catfishes, croakers, and herrings. Dactyloscopid sand stargazers occur here and in the Atlantic, but not in the Indo-West Pacific. Sixty-two Indo-West Pacific species have managed to cross the Eastern Pacific Barrier. Three provinces—Mexican, Panamanian, and Galápagos—are recognized.

Three. The Western Atlantic Region

The Western Atlantic region is the second most diverse oceanic area, containing 1200 fish species. It includes Bermuda (which, although at 32°N, sits in the tropical Gulf Stream), southern Florida, the Bahamas Bank, the Caribbean Sea, and tropical and temperate portions of South America. Most of the families that occur in the Indo-West Pacific also occur in the Western Atlantic; a few families are more diverse here, such as grunts and toadfishes. Strong currents of warm water separate the Western Atlantic fauna from colder waters along much of its boundaries. It is subdivided into Caribbean, Brazilian, and West Indian provinces.

Four. The Eastern Atlantic Region

The Eastern Atlantic region is a relatively small region along the west coast of Africa from Senegal to Angola and extending out to oceanic islands such as Ascension.
and St. Helena. Tropical marine fishes here are limited by cool-water currents impinging from both the north and the south as well as by substantial freshwater runoff and sediments from several major west African rivers, all factors which discourage coral reef growth. The region contains “only” 300 shore fishes; most coral reef families occur but are represented by only a few species. Porgies are particularly diverse. No subdivisions are recognized, with the possible exception of the warm-temperate Mediterranean. The Mediterranean Sea contains 540 species, with many species in the same families as those in the Eastern Atlantic. The Mediterranean has the dubious distinction of being the most heavily invaded tropical marine area in the world, with at least 52 alien fish species having moved in from the Red Sea (an Indo-West Pacific subregion) via the human-made Suez Canal.

5. The Arctic Region
The Arctic region encompasses high-latitude (above 60°N) waters of both the Pacific and Atlantic Oceans. It reaches from Nunivak Island, Alaska, northward and across the polar region to Newfoundland and Norway in the northern Atlantic. Of the two polar areas, the Arctic is more diverse. Successful groups include skates, herrings, salmons, smelts, cods, eelpouts, greenlings, sculpins, poachers, snailfishes, pricklebacks, wolfishes, gunnels, and right-eyed flounders. Diversity within many of these groups is greater in the Pacific than in the Atlantic portions of the region. A total of 415 species occur here. Distribution of many of these families appears to be limited by temperature, with warmer waters and currents to the south determining species’ boundaries.

6. The Antarctic Region
The Antarctic region (above 60°S) has its own distinctive fauna that is restricted to Antarctic waters and the surrounding Southern Ocean, including the cold waters of Australia, New Zealand, and nearby oceanic islands. Forty-nine families and 274 species occur here, 13 families and 174 species of which are identified as Antarctic continent species. A particularly successful group is the notothenioid icefishes and relatives, which account for 55% of Antarctic species. Families include bovichthyids, cod and crocodile icefishes, plunderfishes, and dragonfishes. Non-notothenioid include skates, greenlings, eelpouts, lantern fishes, eel cods, deep-sea cods, and southern flounders.

7. Temperate Regions
To the north of the Indo-West Pacific lie cooler temperate waters with their own characteristic fish faunas. This area can be divided into four fairly distinct regions according to location and temperature: Japanese warm temperate and Californian warm temperate regions and Eastern and Western boreal regions. The warm temperate areas (from about Hong Kong to Tokyo in the west and from lower Baja California to central California in the east) contain a fauna that fluctuates seasonally, as tropical species move north in the summer and boreal species move south in the winter. Notable families to the west include lizard-fishes, flying fishes, mullets, jacks, sea basses, and croakers and to the east include endemic silversides, sea basses, croakers, damselfish, wrasses, and flattfishes. The more northerly boreal regions (approximately north from central California in the east and Korea in the west) contain similar families but different species. Important families include migratory salmonids, sculpins, rockfishes, snailfishes, greenlings, gunnels, pricklebacks, and right-eyed flounders.

In the southern Pacific, cold water currents create at least three temperate faunal regions, with centers of distribution around Australia, New Zealand, and South America.

In the Atlantic, three northern temperate areas occur: the western and eastern Atlantic boreal regions and the Atlantic warm temperate or Carolinian region. The boreal regions (Newfoundland to Cape Hatteras in the west and British Isles to northern Europe and Scandinavia via the east) share a fauna of salmonids, cods, sticklebacks, poachers, sculpins, wolf-fishes, and right-eyed flounders, with occasional strays from more southerly waters during warm months. The Carolinian extends from Cape Hatteras south to Florida and also to the Gulf of Mexico, with southern Florida housing tropical species. Common groups are clupeoids, sea robins, pipefishes, silversides, needlefishes, killifishes, croakers, left-eyed flounders, and puffers. Temperate faunas also occur in the southern Atlantic, but their areas and diversity are less than those of the southern temperate faunas of the Pacific Ocean. Two recognized regions are the eastern Southern Hemisphere and southern African warm temperate regions. The former region’s fauna includes sea catfishes, croakers, herrings, gobies, scorpion fishes, and sea basses; the latter area has many colder water members of Indo-West Pacific families.

8. Pelagic Regions
The 350 species of pelagic fishes occur in ocean surface waters to a depth of 200 m. This habitat type can be divided into 10 different regions based on faunal differences, with more joining of Southern Hemisphere areas because of the relative lack of large land masses. These regions are Arctic and Antarctic polar, North Pacific cold temperate, North Pacific warm temperate, tropical
Indo-Pacific, North Atlantic boreal, North Atlantic warm temperate, Atlantic tropical, southern warm temperate, and southern cold temperate. Many of the world’s most important fisheries species occur in pelagic regions, including numerous sharks, sardines and herrings, salmon, codfishes, pollock, hakes, haddock, sauries, mackerels, and tunas. In addition, about 100 species of mostly pelagic fishes have a worldwide distribution. This group includes pelagic sharks (white, whale, tiger, and perhaps megamouth), swordfish, and ocean sunfish.

9. The Deep Sea

Waters deeper than 200 m are as much a zoogeographic as a habitat entity. The deep sea is generally divided into three major regions based on depth: the open-water mesopelagic (200–1000 m) and bathypelagic (1000–4000 m) regions and the bottom-associated benthal (200–1000 m) region. Benthal fishes are further divided into bentipelagic fishes that hover just above the bottom and benthic fishes that rest in contact with the bottom. Deep-sea fishes are most common in these regions between 40°S and 40°N latitude, approximately between San Francisco and Melbourne in the Pacific and between New York City and the Cape of Good Hope in the Atlantic. Abyssal and hadal (trench) regions deeper than 4000 m are relatively depauperate.

Each region has a characteristic and relatively diverse bony fish fauna consisting of fishes from many different taxonomic groups. The mesopelagic region worldwide contains about 750 species in seven different superorders and nine orders. Despite their lack of relatedness, deep-sea fishes share many anatomical and physiological features, suggesting independent, convergent evolution of adaptations to deep-water existence. Mesopelagic fishes typically have photophores (light organs) on their silvery bodies; have relatively large, often-tubular eyes; undergo daily migrations to surface waters to feed at night; and have large mouths and long teeth. Common names of mesopelagic fishes reflect these traits: barreleyes, bristlemouths, dragonfishes, sabertooth fishes, lantern fishes, tube-eyes, and swallowers.

The bathypelagic region is the largest habitat space on the earth, accounting for 88% of oceanic volume. The five superorders, nine orders, and 200 species in the cold, dark bathypelagic region share some traits with mesopelagic fishes but possess them in the extreme. Photophores are concentrated on lures used to attract prey; eyes are often small; mouths are extremely large and teeth very long; stomachs are expandable; bodies are black; and body musculature, bones, and scales are greatly reduced. These traits reflect greater habitat space and increasingly rare feeding opportunities with increasing depth, which select for an increasing need to conserve energy and to be able to take advantage of feeding opportunities. Again, bathypelagic fishes have names indicative of their adaptations: sawtooth eels, gulp eels, swallow eels, dragonfishes, anglerfishes, sea devils, and langousts.

Benthal fishes include about 1000 bony fish species in four superorders and nine orders, plus chimaeras and squaloid sharks. Different families inhabit bottom compared to open-water regions. Benthal fishes include greeneyes, tripod fishes, hakes, grenadiers, cusk-eels, batfishes, snailfishes, and eelpouts. Although diversity decreases below 1000 m, grenadiers and rattails live between 1000 and 4000 m, tripod fish have been found to 6000 m and snailfishes to 7000 m, and some cusk-eels have been found as deep as 8000 m.

Although some differences in species composition occur in different ocean basins or in association with different water masses, deep-sea species are relatively cosmopolitan, occurring in several different oceans. One trend is for fishes to occur deeper at lower latitudes, such that species that are bathypelagic near the equator may be mesopelagic at middle latitudes and even epipelagic at the poles.

Websites
www.fishbase.org (a remarkable site with information on fishes and fisheries worldwide)
www.fisheries.org/ (website of the American Fisheries Society)
www.fishlinkcentral.com (a clearinghouse for the aquarium trade)
www.fws.gov/9Endopp/endopp.html (U.S. Fish and Wildlife Service, Endangered Species homepage)
www.utexas.edu/depts/ash/ (website of the American Society of Ichthyologists and Herpetologists)
www.wcmc.org.uk (World Conservation Monitoring Centre; lists information on imperiled species worldwide)

See Also the Following Articles
FISH CONSERVATION • FISH STOCKS • LAKE AND POND ECOSYSTEMS • OCEAN ECOSYSTEMS • PELAGIC ECOSYSTEMS

Bibliography
Bigelow and Schroeder (1953). Fishes of the Gulf of Maine. Fish Bull. Fish Wildlife Serv. 33 (Fish Bull. 74).
FISH CONSERVATION

Carl Safina
National Audubon Society

1. Introduction
2. The World Ocean Fishing Situation
3. Other Factors Affecting Fish Conservation
4. Marine Fish Conservation in a Biodiversity Context
5. The Course Ahead

GLOSSARY

aquaculture Commercial farming of aquatic organisms, including seaweeds, raising captive-bred fish, and raising wild-born fish in captivity. Mariculture refers specifically to marine (saltwater) aquaculture.

ballast Water taken in by ships to balance them after they have unloaded their cargo.

bycatch Any living thing caught unintentionally in fishing gear; sometimes called bykill because so many such creatures are discarded dead. About one-fourth of the total world catch is bykill.

depletion In fisheries, reduction to population levels low enough to reduce or threaten future productivity.

ecological integrity The naturally evolved numerical and functional relationships among species and their environment in a given area.

fishery A collective effort to gather, collect, or catch wild aquatic wildlife or plants for recreational or commercial purposes. Fisheries extract large numbers of wild fish, sea urchins, corals, seaweeds, shrimp, snails, clams, scallops, squid, turtles, whales, and other creatures.

harvest To gather a crop. “Harvest” is an appropriate word for farming operations, including fish farming, but not for catching or collecting wild animals or plants. This term is widely misused in industry public relations to make the extraction of wild fish, natural stands of trees, and other wild organisms seem like agriculture, though nothing is planted or nurtured and these things are merely taken for profit. For wildlife, including wild fish, appropriate words include, among others: catch, fish for, take, extract, land, gather, and collect.

high seas Parts of the ocean outside national boundaries, usually beyond 200 miles of any nation’s coast.

keystone species Species whose removal causes a chain or “cascade” of ecological effects among other species.

marine reserves Designated areas where no fishing, mining, or other consumptive use is allowed, usually for purposes of replenishing nearby fishing grounds or maintaining normal evolution, growth, and fecundity.

overfishing Extracting marine organisms faster than they can reproduce.

pollution Introduction of substances in quantities that are threatening to living resources, biological processes, and human health and activities.
OVERTHINKING AND HABITAT DEGRADATION have driven many fish populations to historic lows. Poor fisheries management, increasing human population pressures, and habitat deterioration from several factors have caused this situation. But many fish are resilient and can recover within a decade or two if given viable habitat and a respite from overfishing.

I. INTRODUCTION

An ocean of water covers more than 70% of Earth, and travelers from another world might more logically assume that this planet would be named Ocean. Ninety-seven percent of Earth's water is in the ocean (2% is locked up as ice, and 1% is in surface freshwater or groundwater). Moreover, 99% of the living space on the planet in which life can exist—the "biosphere"—is in the sea. The basis of most life in the sea are the single-celled plants, or "phytoplankton," that create food from sunlight (through photosynthesis) and drift in the upper 1% of the ocean's volume. Plants attached to the bottom can only live in shallow coastal areas where they can get enough sunlight. Most life in the other 99% of the sea relies on food coming from that thin upper layer. Of 33 living animal phyla (the category that reflects the different basic body plans of living things), 32 are found in the ocean, 15 exclusively so. Only one is exclusively found on land (the Onycophora, or velvet worms). This article focuses on the world's ocean fishes. (For a synopsis of biodiversity issues facing freshwater fish, see Box 1.)

II. THE WORLD OCEAN FISHING SITUATION

Many factors affect life in the sea, but fishing has caused the largest changes and is the major current agent of ecological disturbance in the ocean. The United Nations' Food and Agriculture Organization has stated that modern fishing is "globally nonsustainable."

In the twentieth century, annual landings of wild ocean fish increased 23-fold, from 3 million metric tons to a plateau of over 80 million metric tons (Fig. 1). In the 1950s and 1960s, fishing technologies exploded as fishing fleets adapted war technologies such as radar, sonar, and LORAN to peaceful efforts of food gathering. Radar allowed boats to keep fishing in total fog, and sonar could find schools of fish deep beneath the sea's surface. LORAN turned the trackless sea into a grid and allowed boats to find and return to precise spots where fish gather. Satellite-generated maps faxed directly to boats in midocean now track movements of water temperature fronts, showing where to find the fish.

Overfishing, first recognized in the early 1900s, is now prevalent in most major fishing areas. Not all of the catch is used for food. About a third of the world's catch becomes fertilizer, animal feeds, and industrial

**Box 1**

Synopsis of Freshwater Biodiversity Issues

Worldwide, lakes and rivers contain at least 8400 fish species, roughly 40% of known fish species on Earth, and almost 20% of all vertebrates (freshwaters support almost one-quarter of the planet's known biodiversity, in only 0.01% of the planet's water; however, the oceans hold vastly larger populations). Lakes are isolated habitats, leading to a high rate of evolution of species of fish and other animals. For instance, three-quarters of the 2000 plant and animal species in Russia's Lake Baikal are found nowhere else. The lakes of Africa's Rift Valley have produced explosive speciation—99% of the 300 cichlid fish in Lake Malawi, for instance, live nowhere else. And Lake Tanganyika, the least-species-rich lake in the Rift Valley lake chain, has 23% more species of freshwater fishes than all of Europe.

Freshwaters are being degraded and species eliminated at a rate probably comparable to that occurring in tropical rain forests. Habitat loss and introduced species are the two greatest problems for freshwater biodiversity. Introduced predatory fish have already wiped out nearly 70% of Lake Victoria's cichlid fishes, and threaten to soon reduce its unique fish biota by 90%. In the United States half the rivers and streams are significantly polluted, and 90% of U.S. rivers (outside Alaska) are blocked by dams. Consequently, 20% of fish species and more than half the mussels in U.S. freshwaters are endangered or have become extinct, in contrast with only 7% of the mammals and birds in the United States.

Of the world's estimated 9000 freshwater fish species, about 1800 (20%) are in serious decline or have gone extinct. Ecologist Norman Myers has called the freshwater fish situation "the greatest extinction spasm of vertebrates in recent times."
products, and about a quarter of the catch is unwanted and shoveled overboard. The annual rate of increase of world fish landings is now approaching zero. But many fisheries reached the limit decades earlier than the peak in global landings. Indeed, most regions in the Atlantic, Mediterranean, Pacific, and even Antarctic have declining catches. In some regions where catches peaked as long ago as the early 1970s, catches have declined by half. Only in the Indian Ocean has the total catch been barely increasing, but there is little room for more growth there. Declines in some individual fisheries have been catastrophic. Newfoundland cod, which supported one of the world’s largest fisheries for over 400 years, crashed 99% between the early 1980s and early 1990s. Many other fish populations have declined by 80% or more. Some major fishing grounds are now closed in hopes that they will recover from exhaustion. Most important marine resources are considered fully exploited or overfished, and no major untapped fishery resources remain in the world. Humanity’s former faith in the sea’s inexhaustibility was wrong.

A. Social and Economic Concerns

Worldwide, about 200 million people depend on fishing for their livelihoods. Marine fishing supplies almost 20% of all animal protein consumed, and in Asia more than one billion people rely on fish as their main source of animal protein. In the last few years, many thousands of fishers around the world have experienced severe decline or loss of their source of income. Depletion caused primarily by industrial overfishing and exacerbated by coastal habitat degradation threatens tens of millions of jobs, as well as major food sources. As preferred species are fished out, and less-preferred fish are targeted, prices rise for these less-preferred fish. Increases in price eventually removes the species from the tables of the poor. Distant markets and increasingly globalized trade networks allow and encourage regional overfishing or irresponsible aquaculture that exhausts local resources, then move on to tap new source areas. This results in serial depletion and habitat degradation along faraway, out-of-sight-out-of-mind shores. For example: bluefin tuna are depleted in the Atlantic and off Australia and New Zealand because of demand in Tokyo, where they become the world’s most expensive sushi; appetites in the Northern Hemisphere support destructive shrimp farms in the tropics; and shark populations are declining in many places around the world because of demand for their fins, primarily in China.

B. Management Problems

In most of the world, intense political lobbying by the fishing industry usually causes fishery managers to disregard scientific assessments and scientific advice to bring catches within sustainable limits.
While no group of fishes is exempt, Atlantic bottom fishes such as cod, haddock, pollock, flounders, and Atlantic halibut have been among the worst managed. Various species of groupers and wild salmon are depleted throughout most of their ranges. And the world’s giant fishes—the tunas, billfishes (like swordfish and marlins), and sharks—among the most magnificent of the ocean’s wildlife, are also among the most mistreated. In much of the world, management in these fisheries is lacking or ineffective. Sharks, subject to intense overfishing in many regions, are easily depleted because they are generally slow growing with low reproductive rates. Fishing annually kills roughly 30 million sharks worldwide, mostly for just their fins, which are prized as a thickener in “shark fin soup.” Only 3 of 26 major shark-fishing countries have management and research programs. Overfishing has driven virtually all large fishes in the Atlantic to all-time lows. The Atlantic tuna commission’s mismanagement has resulted in systematic overfishing of swordfish, white marlin, blue marlin, bigeye tuna, and bluefin tuna, those populations have declined 60–90%.

In many other cases, poor monitoring or outright cheating enlarge the gap between fishery science and fishing activities. In one 7-year period, foreign fleets fishing the Grand Banks removed sixteen times the quotas that had been set for cod, flounders, and redfish by the Northwest Atlantic Fisheries Organization. A seized Spanish boat kept two log books—one showing actual catches and one for reporting to authorities. Illegal fishing is rampant in the Southern Ocean (around Antarctica), particularly for toothfish, which are marketed as Chilean sea bass (this fishery kills large numbers of seabirds as well.)

No country can be called generally successful in fisheries management. Most are hardly trying. International cooperation on migratory populations and fishing outside national waters—on the “high seas”—is particularly difficult, and in some cases a country unhappy with the restrictions of an agreement simply ignored them outright or through lack of enforcement.

C. Fish Fights: International Fishing Conflicts

Fish conflicts, many of them violent, have erupted around the world. Norway, Iceland, the United States, Canada, Indonesia, Taiwan, Nicaragua, Russia, China, the Philippines, Japan, France, Britain, Spain, and many others have been involved in international fishing disputes. In disputed waters off the Philippines, Chinese vessels were seized and scores of crew members were jailed. In the Galapagos, troops were deployed after fishers seized a biological station that was trying to stop rampant illegal fishing inside a marine reserve. In France, fishers protesting fish imports rioted, causing $4.5 million in damage. NATO allies have pointed guns at each other in disputes over cod, and a Russian Navy boat fatally shot Chinese poachers. More trouble is likely as hungry boats scour the oceans for dwindling resources.

D. Fish as Commodities/Fish as Wildlife

Marine creatures are the only wild animals still hunted commercially on a large scale. And we usually treat fish as mere commodities, forgetting that fish are wildlife. For instance, we speak of a population of fish as “stock,” as in “the New England groundfish stock.” We would not speak of the “Serengeti mammal stock” or the neighborhood “woodpecker stock.” By preventing us from seeing fish as wildlife in natural ecological communities, the language used helps facilitate their overexploitation.

Probably the most abused word throughout fisheries is “harvest.” “Harvest,” applied to wild populations, is an industry public-relations word—the intent is to make activities that take wild living things from natural communities seem like farming. Until recently, humanity has distinguished between hunting and gathering wild things that we do not grow and agriculture, where we raise and harvest crops that we do grow. Nowadays, many people try to blur this distinction by using “harvest” to describe hunting and gathering operations as a way to put a better face on taking large quantities of individuals from wild populations, usually for profit. But harvest is an agricultural word—it means to reap a plant crop that has been sown and nurtured. We usually don’t even say we’ve “harvested” agricultural animals like cows, pigs, and chickens when we slaughter them for meat, even though we raise them for those purposes. So why should we say that we “harvest” wild animals like cod or sharks? Industry people even speak of “harvesting” whales to avoid using the obviously correct word, “hunting,” for the pursuing and killing of wild mammals, in this case the largest animals that have ever lived. When a wild fish or whale or a naturally growing tree is simply taken from a wild community, “harvest” is the wrong word. Appropriate words include: catch, fish for, take, extract, land, gather, collect (e.g., oysters), boat (i.e., they boated 2000 salmon), cut, or hunt.
E. Intensity and Limits of Fisheries
Humans exert tremendous pressure on the seas, using about a third of all “primary production”—the food energy made by plants from sunlight—in coastal areas. The coasts, where nutrients continually enter the marine environment primarily from the land, are the most productive parts of the oceans by far. The coastal areas, representing only 10% of the oceans’ total area, produce 90% of the world catch.

Many industrialized fishing ships deploy large-scale fishing gear such as long-lines that are 50 miles long with thousands of baited hooks, drift nets 40 miles long (now subject to a United Nations ban), and bag-shaped trawl nets sometimes large enough to engulf twelve Boeing 747 jetliners.

Between 1970 and 1990, the world’s industrialized fishing fleet increased at twice the rate of the catch, doubling in number and tonnage. The fleets of the world now wield twice the fishing power needed to catch what the oceans can produce. If the fleets had not grown at all since 1970, they would be able to catch the same number of fish. Fishing has either reached or exceeded natural limits virtually everywhere in the ocean.

F. Kinds of Overfishing
Overfishing is the main reason fish populations have declined. Overfishing means catching marine creatures faster than they can reproduce. “Fisheries” extract large numbers of wild fish, shrimp, snails, clams, scallops, squids, sea urchins, corals, seaweeds, turtles, whales, and other kinds of creatures. Different kinds of overfishing have been identified (see Box 2), and various factors can influence a species’ vulnerability to overfishing (see Box 3).

Overfishing causes: (1) massive depletion of many species; (2) loss of spawners, thus fewer young produced and increased risk of reproductive failure in times of poor environmental conditions (e.g., unusual ocean temperatures); (3) declines in average sizes of fish and other marine creatures; (4) loss of genetic diversity; (5) genetic change toward less desirable characteristics like smaller size potential; (6) disruption of natural communities, and (7) disruption of human communities. Simultaneous overfishing of many species leads to functional loss of species or species groups. “Keystone” species are those whose removal causes a chain or “cascade” of ecological effects. For example, removal of algae-eating fishes can cause coral reefs to be killed by algae overgrowth.

Until recently, one of the main assumptions in fishery management was that the number of young produced is generally not related to the number of spawners, and many practicing managers who were taught this in school still believe it. The reasoning was that environmental factors cause a different percentage of young to survive from year to year regardless of the number of eggs laid. But because any surviving fraction (say, 1%) of a large number of eggs is a higher number of survivors than the same percentage of a small number of eggs, the assumption that numbers of young are not related to numbers of parents is illogical. In 1996, Ransom Myers and Nicholas Barrowman proved this assumption wrong. They studied nearly 400 data sets from different species and asked the following questions: (1) Are the largest groups of young fish entering a fishery produced by the highest populations of spawners? (2) Are the smallest groups of young fish entering a fishery produced by the lowest populations of spawners? (3) Are above-average groups of young fish entering a fishery produced by above-average populations of spawners? Their findings: “The answer to all three questions is almost always ‘yes.’”

<table>
<thead>
<tr>
<th>Box 2</th>
<th>Different Kinds of Overfishing Affect Many Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Growth overfishing” is when too many fish are caught while very small. In species like groupers that change sex with age, a critical shortage of one sex can result.</td>
<td></td>
</tr>
<tr>
<td>“Recruitment overfishing” is reproductive failure due to depletion of breeders.</td>
<td></td>
</tr>
<tr>
<td>“Demographic overfishing” turns a population with many age groups into a population with only one or two significant age classes doing most of the breeding, making the population vulnerable to years when natural fluctuations cause poor survival of young.</td>
<td></td>
</tr>
<tr>
<td>“Genetic overfishing” is when intense fishing changes a population’s gene pool.</td>
<td></td>
</tr>
<tr>
<td>“Serial overfishing” is the depletion of one targeted species after another.</td>
<td></td>
</tr>
<tr>
<td>“Ecosystem overfishing” causes great changes in species composition and functional loss of key species, and can result in long-term community changes.</td>
<td></td>
</tr>
<tr>
<td>Overfishing related directly to the increasing human population is sometimes called “Malthusian overfishing.”</td>
<td></td>
</tr>
</tbody>
</table>
Consideration of These Factors Can Help Us Assess the Vulnerability of a Species to Overfishing

1. Inherent vulnerability: Does the species have low growth rate, low fecundity, high catchability (especially while immature), vulnerable behavior (e.g., spawns in groups), vulnerability to environmental changes, increasing vulnerability due to changes in fishing technology, or a poorly understood life cycle?

2. Human-related environmental risk: Does the species suffer destruction of key habitats, widespread effects of pollution, or conflicts with introduced species?

3. Is the population large or small, lightly or heavily exploited?

4. Is there management? Does it involve long-term conservation and sustainability as a goal? Does it benefit from independent and objective scientific advice? Does it have effective mechanisms for monitoring and enforcement and data collection?

When fishing is reduced, many depleted populations can recover. When fishing ceased in the North Sea during the world wars, fish populations increased significantly. On the East Coast of North America, a spectacular recovery of striped bass followed strict limits on fishing. Subsequent limits on fishing for several other species also resulted in increases in those populations. Protecting fish until they reach average spawning age is one of the simplest, fastest, most effective ways to allow recovery.

G. Bycatch

Every fishery catches unintended, unwanted creatures, called bycatch or bykill. Bycatch includes nontarget and juvenile fishes, seabirds, marine mammals, and any other creature that the fishers are not trying to catch. Bykill currently threatens several species of dolphins and albatrosses with extinction.

By weight, about 20% of the animals caught worldwide are discarded. By number, a much higher percentage are thrown away, because usually it is smaller fish in the catch that are discarded. In many fisheries, bycatch exceeds target catch. Shrimp trawling incurs more bykill than any other fishery. Discarded bykill outnumbers shrimp catch by anywhere from 2-to-1 up to 8-to-1 or even more. Shrimp trawls are the largest human source of mortality to adult sea turtles. The Gulf of Mexico shrimp fishery annually kills and discards 12 million juvenile snappers and 2800 metric tonnes of mostly juvenile sharks. But waste is not the only issue. Using everything caught would not fix the biological effects of killing millions of young fishes, as well as breeding-age seabirds, turtles, dolphins, and other mammals.

Lost fishing gear also catches, kills, and wastes sea life (this is called "ghostfishing"). Gill nets are frequently lost. For instance, the drift gill net (drift net) fishery of the North Pacific, which set approximately 30,000 miles of netting per night before the United Nations banned it, lost roughly 6000 miles of netting annually. A Norwegian study concluded that lost gill nets continued killing fish for at least 7 years; a New England study found them still catching fish 3 years after they were lost. Lost fishing gear frequently tangles seabirds and marine mammals even in areas as seemingly "remote" as the Antarctic. (See Boxes 4 and 5.)

H. Subsidies and Economics

Because there are too many boats, fisheries are said to be "overcapitalized," and are generally unprofitable. Worldwide, to catch $70 billion dollars worth of fish, fisheries incur total costs of $124 billion per year. Government subsidies, variously estimated at $22 to $54 billion...
Bykill from commercial fishing operations is the most serious threat facing albatross populations. Each year over one hundred million hooks are set on longlines in the Southern Ocean. Each longline is up to 80 miles long and can bear 3000 baited hooks. Albatrosses often congregate in the same areas where boats fish. As the line is being played out from the moving vessel, albatrosses sometimes pick up the bait just before the weight of the sinking line pulls them underwater. Work by New Zealand ornithologist Dr. Sandy Bartle and colleagues shows population declines of 50 to 80% in some albatross colonies in the last several decades.

III. OTHER FACTORS AFFECTING FISH CONSERVATION

As discussed so far, fishing is currently the major agent of change in the oceans, but a fuller compilation of the major threats to marine biodiversity includes:

- fishing operations, which cause depletion, catch large numbers of nontarget species, degrade seafloor and coral reef habitats, and cause major changes in living communities throughout the world’s oceans;
- distant markets, which exert tremendous pull on resources from around the world, yet are relatively buffered from the economic, social, and ecological effects of local depletion because they can switch to new sources;
- pollution from those chemicals that cause toxic effects and hormonal disruption;
- pollution from excessive nutrients, which increase the frequency and severity of harmful algal blooms such as red tides, brown tides, and the single-celled fish-killing organism called Pfiesteria, and cause oxygen depletion in the waters;
- physical habitat alterations, including development, wetland alteration, and on-land deforestation, which sends clouds of sediments into rivers and coral reefs;
- introductions of species to areas where they are not native, causing competition, predation, and spread of disease;
- aquaculture or “fish farming,” which usually entails extensive habitat damage, pollution, and introductions of alien species and diseases;
- debris, which tangles, traps, or is ingested by large numbers of marine fishes, mammals, birds, and other creatures;
- dams in rivers, which prevent fishes such as certain salmon, sturgeons, herring, and others from reaching spawning areas, or kill their young on their way to the sea;
- atmospheric warming, which stresses corals, lowers overall ocean productivity, melts polar ice, threatens to raise sea level and inundate coastal wetlands and low islands, and may change ocean current patterns, resulting in changes in abundance and distributions of marine wildlife; and
- ozone thinning, which lowers ocean productivity by killing significant amounts of plankton.

We should bear in mind that gravity takes things from land to sea. Almost everything that enters a storm drain goes directly into the ocean, from used oil to litter. The less chemicals people use on their lawns and farms, the less we will swim in and eat in our seafood. More oil enters the ocean from urban runoff, including automobile leaks, than from large tanker spills. About three-quarters of ocean pollution comes from the land (Table I).
TABLE I
Sources of Marine Pollution

<table>
<thead>
<tr>
<th>Source</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offshore oil production</td>
<td>1%</td>
</tr>
<tr>
<td>Maritime transportation (shipping)</td>
<td>12%</td>
</tr>
<tr>
<td>Dumping</td>
<td>10%</td>
</tr>
<tr>
<td>Subtotal of ocean-based sources</td>
<td>23%</td>
</tr>
<tr>
<td>Runoff and land-based discharges</td>
<td>44%</td>
</tr>
<tr>
<td>Atmospheric deposition into oceans</td>
<td>33%</td>
</tr>
<tr>
<td>Subtotal of land-based sources</td>
<td>77%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

A. Human Population Growth

By 2050, 7–10 billion people will be joining you for dinner, and the seas will feel their impact directly. Each year the world adds 100 million new human appetites. To keep pace for just the next decade, we will have to find 20 million more tons of seafood. A third of the world catch is now used for animal feed and fertilizer, but even if it were simply eaten by people, it would maintain present consumption levels per person for only 20 years. The widespread recoveries of wild fish in the face of increasing demand seem unlikely, and improved conservation would not be able to keep pace with population growth anyway. We have hit the limit of the oceans’ natural capacity to feed us.

Coastal habitat disappears in proportion to population growth, and growth in coastal areas is four times the U.S. national average. Sixty-five percent of major cities are located along the coasts, and half to three-quarters of the human race will soon live within 50 miles of the sea. Diminished habitat means diminished productivity.

B. Habitat Issues

Human alteration of aquatic habitats is a major cause of fishes’ decline—the major reason in freshwater, and the second-largest factor (after fishing) in estuarine and continental shelf habitats. Marine habitat is physically altered by development, aquaculture, mining, and fishing activities. Development often causes: extensive filling and diking of critical wetland breeding and nursery areas; dredging to deepen shallow estuaries for shipping; shoreline stabilization; and dams that alter normal patterns of sedimentation, erosion, and water flow. Dams also block migrations to and from spawning areas, and divert to agriculture, industries, and homes the freshwater flows needed by wetland and estuarine creatures. Aquaculture often destroys local wetlands and mangrove areas. Mining removes coral and minerals, often adding toxic chemicals and sediments that destroy seafloor creatures. Agriculture and deforestation put large amounts of sediments into coastal waters, smothering spawning beds, seagrass flats, and coral reefs. Trawl nets and dredges dragged for fish, clams, and other species extensively damage seafloor structures and living communities. To catch fish from coral reefs, where nets are impractical, fishers often use explosives or cyanide that fragment or poison corals.

Coastal habitat losses entail major costs. In the Chesapeake estuary since 1970, hickory shad has declined 90%, American shad 66%, alewife and blueback herring 92%, and oysters 90% as a result of habitat alteration. Destruction of half the world’s mangroves has eliminated an estimated 5 million tons of annual catch, which is about 3% of total world landings. In the Pacific North-west of the United States and Canada, intensive deforestation and water diversions destroyed thousands of river miles of salmon spawning and nursery habitat, eliminating hundreds of salmon runs and thousands of related jobs.

Coral reefs support Earth’s most diverse fish communities. Coral mining (for construction), fishing with explosives and poisons, destruction from boat anchors and ships running aground, and silt from farms and clear-cut logging cause major damage to coral-dependent communities. Sediment flowing off deforested land is a major threat to coral reefs, killing corals by clogging them, blocking sunlight, and preventing larval settlement.

C. Introduced Species

Aquaculture, the pet trade, sport fishing, and shipping intentionally and accidentally introduce species to new parts of the world. When ships empty their cargo, they often rebalance by taking on water. In the water are larvae, eggs, living cells, and small animals. When the ships arrive and discharge this “ballast” water, these hitchhiking species—it is estimated that more than 3000 species are in motion in ships on any given day—may be introduced to a part of the world where they have never existed. Often they arrive without the predators, competitors, or pathogens that limit them at home. Sometimes, organisms are introduced to places where the native species cannot escape or compete with them. Invasive species greatly alter ecological communities. In San Francisco Bay, a tiny Chinese clam displaced 95% of the bottom community, and now consumes so much plankton that native plankton-eating and, in turn, predatory fish may be affected. In eastern Europe’s Black
Sea, introduction of an American jellyfish reduced plankton biomass by 90% in less than a decade, causing small fishes like anchovies to crash. Reef fishes intentionally introduced into Hawaiian waters for sport fishing have affected local fish communities. The list of examples is long, and future invasions are inevitable. In one study, 5 of 80 ships entering Australia carried toxic phytoplankton, which can produce red tides and cause shellfish to become poisonous.

Even transporting salmon relatively short distances from one river system to the next for aquaculture or into hatcheries, as often happens, can move diseases into new areas, or expose fish to habitats where they are doomed by diseases for which they have no immunity.

D. Aquaculture

Many are counting on aquaculture—farming aquatic organisms—to fill the gap between nature's bounty and the hunger of people. But it is by no means clear that aquaculture actually produces more food.

Aquaculture is exploding. For example, shrimp farming increased sevenfold from the mid-1980s to the mid-1990s, with global value now exceeding $6 billion annually. Aquaculture now produces a fifth of humanity's fish. In freshwater, aquaculture now exceeds wild catch. In the oceans it produces about half the salmon interbreeding with wild salmon cause genetic degradation, because wild salmon are genetically adapted to specific characteristics in the rivers where they spawn. Moreover, ponds and other open systems attract wild fish-eating birds such as cormorants and herons, often bringing them in range of farmers' guns. Most importantly, most aquaculture consumes vast quantities of wild fish as feed—and is thus a net loss of food from the oceans. Most believe that aquaculture contributes to the supply of human food, and this is true for the few herbivorous species. But fish are not cabbages; they don't grow on sunlight. Most fish and shrimp are carnivorous, and feeding them uses two to four times as much wild fish as the farms produce. Farming such species actually contributes to the ocean's depletion.

Aquaculture does not take pressure off the seas. Wild larvae of species that do not breed in captivity are sometimes intensively caught for pen-rearing. Increased shrimp and salmon farming has not lessened fishing intensity even for those species, and has brought trouble to wild shrimp and salmon through diseases and habitat destruction.

Aquaculture can be done responsibly but as usually done now it destroys naturally productive habitats that support ocean fisheries and biodiverse wild communities, and introduces alien species, parasites, and diseases that threaten local biodiversity. Responsible aquaculture enterprises should not be sited where natural habitat is affected, its wastewater should be treated, and it should focus on herbivorous species in closed systems to prevent escape.

E. Pollutants

Pollution is the introduction of substances in quantities that are threatening to living resources and human health and activities. Chemical pollution can cause spectacular mass mortalities or cause subtle changes in population composition, impaired sexual development and reproductive success, impaired growth rates, outbreaks of harmful algal blooms such as red tides, *Pfiesteria*, or normally innocuous algae that overgrow corals or deplete waters of oxygen. Chemical pollutants tend to concentrate in surface waters, where larvae and eggs also concentrate. Chemicals enter the seas from sewage, industrial outfall, agricultural runoff, ocean dumping, aquaculture, accidental spills, and from the air (acid rain has significantly damaged freshwater fishes in North America and Europe). Major chemical pollutants include insecticides and herbicides, deter-
gents, PCBs, elements such as chlorine and heavy metals, petroleum products (nearly half the gasoline/oil mixture run through 2-cycle outboard engines exits unburned), mining wastes, fuel ash, radioactive materials, and excessive nutrients from sewage, farm animals, and fertilizers.

When algal blooms caused by excessive nutrients use up those nutrients and die, their subsequent decomposition can rob the water of oxygen, suffocating marine creatures. The surprising fact that fertilizer or other nutrients in excessive amounts can kill aquatic organisms is sometimes called “the paradox of enrichment.”

About three-quarters of marine pollution comes from the land, creating an asymmetry whereby the people causing the pollution do not feel its effects directly or immediately, and the people feeling it (e.g., fishers) cannot directly affect its origin. That is one reason why laws are necessary. In the United States clean water legislation has helped keep some fisheries economically viable by controlling pollutants.

F. Global Atmospheric Change and Fish Conservation

Major atmospheric changes, including ozone thinning and global warming, have significant implications for marine life. Earth’s climate has changed dramatically as ice ages came and went, but these changes took thousands of years, allowing life-forms long periods in which to adapt. Human-caused climate change is happening much faster, and some habitats and species will probably not change or move fast enough to survive.

The burning of fossil fuels has increased atmospheric levels of carbon dioxide, methane, and other heat-trapping gases, and most scientists agree that this is intensifying the “greenhouse effect,” the warming of Earth’s atmosphere. Oceans slow the buildup of greenhouse gases by absorbing about a third of the carbon produced by burning, but substantial climate change is likely over the next few decades. The effects of such change are not straightforward, but planetary warming is melting polar ice caps, causing population declines in ice-dep-endent krill (the shrimp-like key prey species for Ant-arctic fishes, marine mammals, penguins, and other seabirds), and raising sea levels throughout the world. This rise is expected to cause some flooding and loss to coastal marshes, mangrove areas, low-lying islands, and their critical associated fish-nursery habitats, as well as to some cities.

Warming is likely to alter ocean currents. Such changes may intensify the recurrent El Niño phenomenon, which changes sea temperatures and exerts a major influence on the survival and distribution of oceanic populations. In parts of the California Current, planktonic animals declined 70% over the last 40 years, a decline many scientists think may be linked to global warming. Other possible changes include alteration of currents such as the Gulf Stream and Kuroshio Current, and weakening of major upwellings off South America and Africa. Such changes would affect the production, distribution, and survival of fishes and other creatures in the world’s oceans, and alter the fish availability among countries.

Certain human-generated compounds such as chlorofluorocarbons rise into the stratosphere and destroy the ozone that shields Earth from the sun’s ultraviolet radiation (UV). The biologically damaging UV is called UV-B, and it penetrates many meters below the sea surface. It has increased under the Antarctic “ozone hole” and elsewhere. Also, recent volcanic eruptions (such as that of Mt. Pinatubo in the Philippines) have reduced total atmospheric ozone by as much as 10%, leading to increased UV-B of up to 20% in latitudes as far from the poles as Florida and the Bahamas. This has caused significant damage to plankton (small drifting animals and plants, including fish eggs and fish larvae), corals, and bottom-living organisms.

IV. MARINE FISH CONSERVATION IN A BIODIVERSITY CONTEXT

Biodiversity is the diversity of living things, and can be thought of as occurring on three levels: genetic diversity, which is the genetic variability among members of the same species; species diversity, the variety of species found in a community or ecosystem; and community diversity, the varieties of biological communities on Earth. We will use these categories to examine the fish conservation issues introduced above.

A. Marine Fish Biodiversity

The greater the diversity of habitats, and the more isolated habitats are from each other, the greater the diversity of species adapted to those habitats. (Though fresh-water covers only 1% of Earth and ocean fish populations are vastly larger, 40% of the world’s fish species live in freshwaters because they are so isolated from each other.) Because most ocean habitat diversity is coastal, so is most ocean fish diversity—there may be 10 times as many coastal fish species (roughly 13,000) as truly open-ocean fishes (about 1200 species).
B. Genetic Diversity within Species

Conservation of biodiversity is virtually never an official management goal in fisheries. Far from it; if we wanted to exterminate the fish, we could hardly have gone about it differently than we have gone about fishing. Fishing can induce genetic changes through unintentional selective breeding because of intensive, non-random fishing mortality. In intentional artificial selection (as for farm animals) we choose and breed individuals with traits that we desire. But in fishing, we remove from the breeding population those individuals with the traits we like. Fishing activities can cause evolution. Evolution is a change in the relative frequency of genes. It need not result in “progress” or new species; it merely shapes species for better survival under prevailing, often changing, conditions (if change is extensive, new species can result). Fishing is so intense in so many places that many fishes have a higher chance of dying by being caught by people than by natural causes (for instance, in some populations 80–90% of adult cod or salmon are killed by fishing; in such situations the probability of survival becomes vanishingly small).

Fishing can alter inherited characteristics of a population in two ways: (1) by applying a selective pressure that kills individuals based on certain traits, such as size or age, and (2) by applying a random pressure so intense that the population is depleted low enough to lose genes from the pool. Note that fishing can do these things, and it does both of them sometimes, but not all fishing results in genetic changes; whether it does is usually determined by fishing intensity.

Genes related to size, growth rate, and age of sexual maturity are most likely to be affected by fishing. “Growth overfishing” prevents fish from reaching large size, thus incidentally selecting for genetically pre-programmed to mature and breed younger and at smaller size than normal. Fish that reproduce pass their genes on. The genes of fish that get killed before breeding will begin dying out. Fish genetically programmed to breed at slightly younger age than average are also less likely to have reduced growth potential. Thus, intense fishing that allows few individuals to live long will allow early-maturing fish to make a disproportionately large contribution to the next generation, and allow normally maturing fish to make a disproportionally small contribution. In a study modeling the contribution of three age groupings of cod to overall reproduction, the late-maturing subgroup made virtually no contribution to reproduction after only 7 years of intense fishing. Such a scenario indicates the likelihood of artificially driven evolution in wild populations. So fishing can unintentionally breed a population of younger-reproducing, shorter-lived, smaller fish.

Probably most intensive fisheries create conditions necessary for genetic change, but genetic changes can be difficult to confirm. The reduced average sizes and first-breeding ages caused by killing most fish before they live long do not necessarily reflect accompanying genetic change. Whether genetic or not, examples abound of population-level effects of overfishing. In a rockfish (Sebastes alutus), lightly exploited populations had a modal age of 30 years, and 73% of the fish were over age 20; heavily exploited populations had a modal age of 12 years, with only 7% of the fish over age 20.

Average sizes of many U.S. Southeast species declined about 75% between the early 1970s and the mid-1980s: red porgies went from 2.6 pounds to 1.3 pounds, red snapper from 18.0 to 4.4 pounds, snowy grouper from 17.6 to 4.4 pounds, speckled hind from 19.1 to 6.6 pounds, scamp from 10.1 to 3.3 pounds, and gag grouper from 18.0 to 4.4 pounds. For species such as groupers, which change sex in older age, intense removal of larger individuals can dramatically affect sex ratios and mate supply, greatly lowering the population’s reproduction. Jim Bohnsack of the National Marine Fisheries Service says that if almost all the big ones are taken and the little ones left, “The result is a race of miniatures.”

Overfishing reduced the California sardine population’s age structure from five spawning age classes to only two, and two consecutive years of poor oceanic conditions led to spawning failure and collapse. However, even when vast populations of naturally small, short-lived fishes such as sardines, anchovies, and herring collapsed to one-thousandth of their former numbers, high enough numbers have remained to maintain biodiversity. Yet in some animal populations, older or larger individuals are more likely to carry more “heterozygosity”—differing forms of particular genes. Larger individuals often spend more time on breeding areas than younger ones, so fisheries targeting spawning groups, such as those for groupers, orange roughy, and many others, can subject the most genetically diverse individuals to the most intense mortality. This kind of genetic diversity loss is believed by some to be a common consequence of heavy exploitation, even without reducing populations to near-extinction levels.

Theory suggests that genetic change from fishing is likely to be common, and several studies have found convincing evidence. Perhaps the species best-studied before and after commercial exploitation is a long-lived fish called the orange roughy. Off New Zealand, only...
6 years of heavy commercial exploitation reduced populations 70% and significantly reduced genetic diversity within those populations, probably because older individuals in this very long-lived species were more heterozygous.

In one North Sea cod population, intensive fishing reduced the chances of a young cod surviving to breeding age by 93%. Over several decades, average age of sexual maturation declined from about 10 to 7–8 years, apparently through genetic change. Some populations of flounder and haddock also show convincing evidence of genetic change. In a population of Atlantic salmon, average age at maturity had dropped after two decades of heavy fishing. In a study of chinook salmon, average spawning age fell by 2 years. Other studied salmon showed other changes.

In sum, some studies found no apparent changes attributable to fishing, some showed changes that were inconclusively genetic, and other studies showed genetic changes that were driven by intense fishing.

C. Species Diversity

Human activities significantly change species composition and abundance, and predator–prey and competitive relationships. Naturally evolved numerical and functional relationships among species are sometimes referred to as the community’s “ecological integrity.”

One indication that many regions’ ecological integrity is greatly diminished is the changing world catch. Since 1950, fisheries, forced to work lower on the food chain as they deplete large fish, have shown a gradual transition from catching mostly large, long-lived, fish-eating bottomfish such as cod toward catching small, plankton-eating, open-ocean fishes like herring, as well as short-lived, low-on-the-food-chain invertebrates like squid. Daniel Pauly, the eminent ecologist who helped discover this trend, remarked, “If things go unchecked, we might end up with a marine junkyard dominated by plankton.”

Various marine communities have been changed by overfishing. A common pattern—particularly acute in coral reef systems but also seen in kelp communities and elsewhere—is selective removal first of the largest species, then of large individuals of smaller species. People removed most large groupers and sea basses from many coral reefs and kelp forests, and manatees, dugongs, and sea turtles from seagrass communities, before scientists ever studied these habitats. Though no one really understands how those communities are evolved to function, they were certainly very different.

Since Columbus landed in the Bahamas, for instance, Caribbean sea turtle populations have declined 99%. Depletion of fish populations changes patterns of abundance, distribution, and competition among plants, urchins, corals, sponges, tunicates, and other creatures. Alarming declines in North Pacific seabirds, Steller’s sea lions, and other wildlife may be caused by heavy fishing. When fishing depleted Barents Sea herring and capelin, cod failed for lack of food. Squid sometimes increase following the collapse of their overfished predators. Off New England and maritime Canada, overfishing shifted communities dominated by cod, haddock, and flounder to domination by spiny dogfish (a small shark) and skates (a type of ray).

But not even skates are safe. Fishing has driven the commonest skate in the northeast Atlantic, Raja batis, to local extinction in the Irish Sea. The barrudo skate was one of the Northwest Atlantic’s largest and most numerous skates in the 1950s, appearing in 10% of research samples. Once estimated to number more than half a million off Newfoundland alone, not a single barrudo has been caught in that area since the 1970s, and if trends continue this skate could become the first well-documented extinction in a fully marine fish.

1. Vulnerability of Marine Fishes to Extinction

Few fully marine creatures, and perhaps no fully marine fish species, have gone extinct because of human activities. Many marine fishes have large distributions and a greater chance for replacement by neighboring populations if they are locally wiped out.

But the important thing is to prevent, not document, endangerment and extinctions. The time to save a species is when it is still common. Waiting until biodiversity declines, until populations falter, and until more species names get added to various endangered lists will only ensure three things: (1) more species will get into serious trouble, (2) the value to humanity from extremely important natural resources such as fishes will be seriously diminished, creating more of the economic, social, and nutritional problems already suffered by people in too many communities throughout the world, and (3) fixing the problems will be much harder, more expensive, and prone to failure.

Numerous fish have already been listed on the U.S. Endangered Species List, on the U.S. Overfished Species List, and under the world Convention on International Trade in Endangered Species (CITES) and the World Conservation Union’s “Red List.” The Red List contains over 100 marine fish that have suffered major and rapid
population declines or significant local extinctions (ex-
tirpations).
Most vulnerable are the anomalous species—those
that breed in freshwater but mature in salt water—such
as salmon and sturgeons, because they suffer overfish-
ing plus the effects of habitat disruption in rivers that
are heavily dammed, subject to intensive agricultural
water withdrawals, or subject to massive siltation from
destructive logging practices. Hundreds of North Amer-
ican salmon runs are already extinct, and the trends
for sturgeons and totoaba (a large fish of the croaker
family that is dependent on Colorado River flow into
the Gulf of California) suggest that biological extinction
may become a real possibility.
Groupers are also vulnerable; they are generally
slow-growing fishes whose spawning groups are easily
targeted by fishers. Long-lived, slowly reproducing spe-
cies such as large sharks and sawfishes are particularly
vulnerable, and many such species are now depleted
and threatened. Several large sharks, including the great
white, are now totally protected in some regions, an
acknowledgment of their threatened status.
No doubt endangerment to fishes is underappreci-
ated simply because their populations are not well stud-
ied. Most of the 100-plus species listed on the World
Conservation Union's "Red List" got there simply be-
cause the London Zoological Society and World Wide
Fund for Nature convened a workshop to compare
known information about fish population changes
against the Union's listing criteria. Similarly, the harm-
door skate became endangered without anyone realiz-
ing it, simply because no one had analyzed decades of
already-existing data.
Not just fish are affected. Wild abalone populations
off California have been greatly reduced by fishing, and
the white abalone is almost certain to go extinct in the
wild because probably too few individuals are left to
breed effectively. Who would have thought that fishing
would be intense enough to drive a snail to extinction?
2. Kinds of Extinction
Overfishing and habitat degradation can lead to a four-
step process of extinction, in which people suffer the
major effects long before the last animals vanish. From
social, economic, and genetic standpoints, extinction
is a process rather than an event; fishing economies
usually go extinct and gene pools lose diversity long
before the last fish dies.
The first stage of the extinction process is depletion,
wherein the fish population is reduced below its most
productive level; fish are smaller and fewer than they
could be, and less likely to spawn a strong year class
of young fish. Consequently, the population's ability to
support fishing is reduced, and profit margins decline.
Other members of the ecosystem may experience food
shortages or an unnatural relaxation of predation pres-
sure or competition.
The next stage is ecological extinction, where the
animal's population drops so low that the species no
longer effectively functions as prey, predator, or com-
petitor in the community. If it is an important "keystone
species" affecting many others, entire marine communi-
ties may undergo profound shifts in numerical and
functional relationships, and species less valued by peo-
ples may increase. Animals that are ecologically ex-
tinct—essentially removed—from the southern Cali-
ifornia kelp community, for example, include sea otters,
black sea bass, white sea bass, large groupers, all the
large kelp bass, sheephead, rays, flatfish, rock fish, lob-
sters, abalones, sea urchins, sea cucumbers, and others.
Another stage, commercial extinction, occurs when
the species becomes too rare to be fished profitably.
During this phase, fishery managers sometimes close
large areas to fishing, as off New England and New-
foundland, or in the salmon fisheries off the Pacific
U.S. coast.
If commercial and ecological extinction are not re-
versed, total extinction may become a possibility. But
by the time total extinction becomes an issue, all the
other practical effects of the creature's disappearance
from commerce and from the ecosystem have already
been suffered by people. One cannot ignore the commu-
nity effects of overfishing on the top predators—
humans. Through dislocation, malnutrition in remote
local villages, job loss, and loss of social identity, hu-
mans suffer the major effects of overfishing long before
the animals themselves completely vanish.
To date, total extinctions are rare in the oceans, but
this may not always remain so. We may unfortunately
see more fish on endangered species lists in the future.
If we do, it may not be too late to save them. But it
will likely be too late to save the fishing jobs and coastal
communities that once depended on them. A recovery
vision should always include regaining naturally
evolved numerical and functional relationships
among species.
D. Ecosystem Diversity and Fishes
Though the ocean may seem like one big bathtub, to
the creatures in it and the people making a living from
it, the sea is a mosaic of habitat types. Water circulation
patterns can create discrete habitats because tempera-
ture alone can be a boundary for sea life. Currents also
cause drifting seaweeds to accumulate in places like the
Sargasso 'Sea,' an area in the middle of the Atlantic
characterized by vast mats of sargassum weed used as
habitat by many creatures. The sediments of the seafloor
vary in texture, from boulders to fine silt, all creating
different habitats. The area between high and low
tides—the intertidal zone—likewise contains particu-
lar habitats for specially adapted creatures that live on
rocks, in sand, or in tide pools. Along coasts, shallow
"estuaries," where freshwater from rivers mixes with
salty seawater, have their own suite of habitats: seagrass
flats, salt marshes, bays between outer beaches and
mainlands, and tidal inlets. Tropical and subtropical
shores often support mangrove forests growing in salt
or brackish water. Near warm shorelines, coral reefs
support the most spectacular diversity of fishes and
other creatures in the sea.

The distribution pattern of these major habitat types
is what we speak of with the term "ecosystem diversity." Just as fishery management has generally ignored
genetics, it has also generally ignored ecosystem consider-
ations. Fishing activities, aquaculture, coastal devel-
oment (for instance, damming rivers and filling wet-
lands), and atmospheric warming have resulted in
changes in the distribution and functioning of habitats
upon which marine communities rely, that is, changes
in ecosystem diversity.

Aquaculture has significantly altered coastal and es-
tuarine habitats in many parts of the world. The conver-
sion of mangrove habitats into fish and shrimp farms
over extensive areas is a form of ecosystem depletion.
Coastal development has also destroyed habitat in eco-
logically significant areas. For instance, about half of
the wetlands in the United States have been lost to
practices like filling, which has led to a significant shift
in the distribution of aquatic habitats. Atmospheric
warming is both melting Antarctic ice and stressing
coral reefs. This is causing change in these habitats on
a global scale.

Coral reefs and continental shelves are two major
kinds of habitats experiencing profound changes due
to human activities. Small increases in temperatures in
tropical sea have been implicated in coral "bleaching,"
whereby corals lose the helpful algal cells living inside
them, turn white, and may die. Widespread death of
corals, of course, has potentially severe implications for
coral-dependent fish communities.

Coral reefs are the oldest and largest structures built
by living things. When scientists first surveyed them
in the 1930s and 1940s, most coral reefs were poised
on the edge of profound change in the composition and
abundance of species and the functioning of coral reef
communities. For example, Caribbean overfishing had
already depleted herbivorous fishes to very low levels
when algae-eating sea urchins crashed from a disease.
Without herbivores, algae overgrew corals, killing them
over large areas, and altering the community. Con-
versely, heavy fishing off East Africa depleted fish that
prey on coral-eating sea urchins. With their predators
largely fished-out, the urchin populations exploded to
densities that were 100 times normal, and proceeded
to inflict significant damage to reefs. The depletion of
predator fishes has also been implicated in destructive
outbreaks of the coral-eating sea star known as the
crown-of-thorns. Such ecological changes affect large
communities of organisms.

Fishing often kills corals directly when people use
explosives to kill fish, or pound the coral with rocks
on lines to scare fish toward nets, or use cyanide to
stun fishes. All of these methods cause extensive de-
struction of coral reefs, through either breakage or, in
the case of cyanide, death to corals by poisoning.

Coral reef systems are perhaps more sensitive to
environmental change than most other marine ecosys-
tems. For various reasons, species of algae may overtake
a damaged reef prevent corals from recovering or recol-
oring. Many coral reefs are now degraded by human
activities, their fish communities modified, and the liv-
ing corals replaced by algae growing on dead coral
heads. Some areas of the Caribbean lost 50–90% of
their live corals in the decade from the late 1980s to
the late 1990s, because of overfishing, diseases, algal
overgrowth, sewage overfertilization, and smothering
silt running off deforested land areas. Some scientists
believe these changes may be very long-lasting.

Aside from coral reef areas, continental shelves
nearly everywhere are affected by bottom trawling—the
towing of large bag-shaped nets or shellfish dredges
over the seafloor. Trawling is the fishing method that
takes half of the world catch. The vast majority of the
world's seabed is encrusted or honeycombed by living
things, and trawling causes enough incidental damage	of nontarget invertebrates, plants, shellfishes, and fishes
to cause major community changes over large areas.
Trawling crushes, kills, exposes, and removes these
living sources of nourishment and hiding places, mak-
ing life dangerous for young fishes and almost certainly
lowering the habitat's capacity to produce abundant
fish populations. In one study of a scallop fishery, only
12% of the scallops in an area were actually caught,
but many of those not caught were crushed by scal-
lop dredges.

Until just a few years ago, trawlers were unable to
work on high-relief, bouldery, or rubble-strewn bottom
habitats without risking hanging up and losing their nets and gear. But nowadays, new kinds of trawls make every kind of seabed—whether silt, sand, clay, gravel, cobble, boulder, rock reef, worm reef, mussel bed, sea-grass flat, sponge bottom, or coral reef—vulnerable to trawling. Furthermore, trawls hit many areas repeatedly each year. The seafloor structures that juvenile fishes rely on are often easily destroyed by bottom trawls, which kill or disperse the organisms that create them. Our focus on extracting edible fishes at all costs has effectively blinded fishery managers to the essential food and shelter that these fish themselves require.

Trawling is not uniformly bad for all species or all bottom habitats, and in fact some species do better in trawled than in undisturbed habitats (just as a few species do better in logging clear-cut areas). But most species are not helped by trawling, and some communities are seriously damaged. Recovery of certain bottom communities could require up to a century even if trawling was stopped today. Writing in the journal Nature, Daniel Pauly and Villy Christensen observed that, “Contrary to some terrestrial systems such as rainforests, of which large undisturbed tracts still exist, the overwhelming bulk of the world’s trawlable shelves are impacted by fishing, leaving few sanctuaries where biomasses and biodiversity remain high.”

V. THE COURSE AHEAD

A. Marine Protected Areas

In times past, fisheries benefited from de facto refuges: undiscovered locations or places too deep or remote to probe. But now, because of new technologies, when fish run they can’t hide. Facing this reality, some scientists have proposed that simply closing some areas to fishing and resource extraction, to allow them to function naturally, while leaving adjacent areas open, is the best way to manage fishing. These closed areas have been suggested especially where precise estimates on population sizes and sustainable catch levels are lacking and where gathering lots of data or intensive management and monitoring are impractical.

Such “marine reserves” allow fish to go through their normal patterns of growth, abundance, reproduction, genetic diversity, and community structure. Marine reserves can (1) protect critical spawning adults, (2) maintain natural size and age structure, (3) maintain genetic diversity, (4) prevent “serial overfishing,” where one species and then the next and the next are depleted, (5) maintain natural communities, while allowing fishing to continue nearby, and (6) provide scientific areas for the study of more naturally functioning systems. Few such reserves exist; indeed, fishing is even allowed in all U.S. National Marine “Sanctuaries.” Marine reserves have been established in several parts of the world, and they can increase fish yields in adjacent areas as populations recover, fish reach their normally evolved maximum size, and their offspring begin wandering outside the reserve. In the Philippines, catches adjacent to one reserve tripled within two years of the reserve’s designation. In a small Caribbean reserve, overall biomass of commercially important species increased 60% and snappers increased 220% in two years; but groupers did not increase, probably because severe depletion eliminated any source of larvae.

Eggs and larvae of fishes and other marine creatures may drift many miles from their source, so reserves must either be self-sustaining sources of fish or areas where incoming juveniles can grow and reproduce. Ideally, reserve networks should be designed to maintain genetic and community diversity over large regional areas. Large adults are disproportionately fecund—a single female red snapper 24 inches (61 cm) long produces as many eggs (about 9 million) as a 200 16-inch (41 cm) females. One study estimated that if 20% of the red snapper’s Gulf of Mexico habitat was protected, total egg production would be 1200% greater than under current fishing pressure. Total fish catch would increase even though a fifth of the area was off-limits to fishing.

For more migratory fish, seasonal and area closures of areas such as spawning and nursery grounds could be helpful. For example, longlines catch mostly immature swordfish in certain areas. Those areas should be closed to long-lining during times where breeders or juvenile fish concentrate. Reserves are generally opposed by fishers at first, but in New Zealand they worked so well that 10 years after their establishment nearly 80% of fishers wanted more reserves.

B. Other Solutions, and Reasons for Optimism

The current situation is poor, but not bleak. There are reasons for optimism. For one, we know enough about many human-induced problems, especially overfishing and habitat alteration, to fix them. Clearly one of the most important things that could be done for overfishing and bycatch is to remove the artificial tax breaks, supports, and all the other subsidies that are propping up fisheries incapable of existing off the resources. And many fishes (though not sharks) have high reproductive
Smuggling Whales

Though whales are mammals, not fishes, whaling is sometimes considered a "fishery" and is similar to fishing in many aspects.

"Are the whale products available today exclusively from species hunted or traded in accordance with international treaties? A recent spot check of Japanese retail markets shows that they are not and that the existence of legal whaling serves as a cover for the sale of illegal whale products." That was the question and the answer reported by two marine biologists in the journal Science in 1994.

Japan legally kills 300 minke whales in Antarctic waters each year, ostensibly for "scientific" purposes. But the "samples" go to food markets, and the markets serve as continued incentive to hunt whales illegally. Iceland recently quit the International Whaling Commission to evade the ban on whaling, and Norway simply resumed whaling for minke in nearby waters.

Sometimes the meat labeled "minke whale" is really from endangered species and from parts of the world where the great whales are supposed to be protected after being hunted to near extinction. While only minke whale meat can be legally sold in Japan, DNA tests of 14 samples purchased in commercial markets in 1993 indicated that "minke" meat included minke, humpback, and finback whales from the North Pacific, North Atlantic, Australian region, and Antarctic, as well as dolphins. "We were surprised by the large number of species and the fact that they came from all over the world," wrote Steve Palumbi, now at Harvard University.

Smuggling whales is not difficult. Recently 260 tons of whale meat labeled as Norwegian shrimp was seized in Russia en route to Japan. This is why the arguments about sustainable hunting of minke whales are flawed and sometimes fraudulent, and why any whaling may prevent recovery of endangered whales in the vast areas of the world's oceans where they are still vanishingly rare.

The Atlantic gray whale was apparently hunted to extinction, but whales stand a good chance of recovery when effectively protected. This has been proven by increases of several great whale species in the Northwest Atlantic, humpbacks that breed in Hawaii, blue whales off California, and the recent full recovery of the Pacific gray whale— one of the great successes of the U.S. Endangered Species Act.
increasingly crowded coasts. Of course, human overpopulation and overconsumption drive or worsen each of the problems discussed here. Anyone interested in conservation can join organizations working in this area, write letters to agency and government officials, speak at meetings and hearings, and write opinion pieces and letters to editors of various publications. Making your voice heard is the most important thing.

The solutions to overfishing and marine habitat destruction depend on political will, but politics reflects public opinion, and that means that politics can be changed. Already, because of changing political “seascapes,” several formerly unlikely long-shots have been achieved—including the global ban on drift netting, the strengthening of U.S. fisheries law toward ending overfishing and hastening the rebuilding of depleted populations, several new international fishing treaties and agreements, and increasing acceptance of the “precautionary principle” that says we should act conservatively, consider future needs, and so avoid irreversible change. Only by increasing our political involvement and effectiveness will we be able to improve the ecological health and protect the biodiversity of the world’s oceans. Conservation itself is political activity in the service of living nature.

See Also the Following Articles

AQUACULTURE • COASTAL BEACH ECOSYSTEMS • ESTUARINE ECOSYSTEMS • FISH, BIODIVERSITY OF • FISH STOCKS • GREENHOUSE EFFECT • KEYSTONE SPECIES • MARINE ECOSYSTEMS, HUMAN IMPACT ON • REEF ECOSYSTEMS • SALMON

Bibliography


I. Major Adaptations of Fishes
II. Respiratory Constraints to Growth and Related Processes
III. Distribution of Exploited Fish Stocks
IV. Ecosystem Impacts of Fisheries
V. Managing Fish Biodiversity Information
VI. Preserving Fish Biodiversity

GLOSSARY

biomass    Collective weight or mass of all the members of a given population or stock at a given time, or, on the average, over a certain time period.
biocides   Occurrence record of organisms, serving as key units for biodiversity research and consisting of four elements (species names, location, time, and source).
catches    The fish (or other aquatic organisms) of a given stock killed during a certain period by the operation of fishing gear(s). This definition implies that fish not landed, that is, discarded at sea, or killed by lost gear (“ghost fishing”), should be counted as part of the catch of a fishery.
ecosystem  Area where a set of species interact in characteristic fashion, and generate among them biomass flows that are stronger than those linking that area to adjacent ones.
recruitment Entry of juvenile fish into the (adult) stock. Recruitment is distinguished from reproduction, because the eggs and larvae that result from fish spawning usually suffer tremendous and largely unpredictable mortalities, thus uncoupling spawning from recruitment.
trophic level A number indicating the position of a species within an ecosystem though the number of steps linking it to the plants. By definition, plants are TL = 1, herbivores are TL = 2, and so on. Note that trophic levels do not need to be whole numbers; intermediate values occur among omnivorous consumers.

FISH STOCKS ARE POPULATIONS OF “FISH,” THAT IS, VERTEBRATES WITH GILLS AND FINS, SUBJECTED TO EXPLOITATION BY HUMANS. Populations are components of species, inhabiting part of their overall range, and usually having little genetic exchange with adjacent populations. The major adaptations determining the spatial distribution of fish stock biomass pertain to the anatomy, reproductive biology, and respiratory physiology of the species to which the stocks belong. Also, fishing has become increasingly important to the biodiversity of fish, either through its direct impacts (changes of stock size and age structure, and overall biomass reductions, down to extirpation of populations), or by modifying the ecosystems in which they are embedded. Research devoted to monitoring the biodiversity of fish (or other organisms) must be able to handle large amounts of suitably formatted distributional information, here defined as consisting of
“bioquads.” Management regimes aiming at preserving fish biodiversity will have to include much stricter regulation of fishing and the establishment of no-take areas.

I. MAJOR ADAPTATIONS OF FISHES

A. Anatomy and Physiology

With about 25,000 recognized species in over 500 families, fish are the most diverse vertebrate group. However, their watery habitat, while failing to protect them from modern fishing gear, makes it difficult to fully appreciate this diversity, and the extent to which it is now threatened. It is even more difficult for us, as air breathers, to perceive the constraints under which fish, as air breathers, were forced to evolve.

Water is an extremely dense medium, 775 times heavier and 55 times more viscous than air. Also, water contains 30 times less oxygen than air, and this oxygen diffuses 300,000 times more slowly than in air. These physical constraints, which shaped all early life-forms, including the jawless predecessors of the fish, the agnathans, are best visualized by describing the major evolutionary trends leading from agnathans to modern fish (Fig. 1A).

The first of these trends was the evolution of jaws from the first upper and lower gill arches of agnathans. This built on the intimate connection, in the most primitive vertebrates, between the feeding apparatus (i.e.,

![Figure 1: Major evolutionary trends from agnathans to extant fishes. (Note that no direct ancestor-descendant relationships are implied among the groups depicted.) (A) Trends toward larger gills; (B) trends toward efficient jaws; (C) trends toward effective paired and unpaired fins. (Note the aspect ratio of the caudal fin, defined by \( A = \frac{h}{s} \), where \( h \) is the height and \( s \) the surface (in black) of the caudal fin, and of which high values define fast, large-gilled continuous swimmers, and conversely for low values.)

\[ \text{Thunnus obesus, } A = 7.5 \]
\[ \text{Cyprinus carpio, } A = 3.2 \]
\[ \text{Pomatochistus microps, } A = 0.6 \]
\[ \text{Pomatochistus microps, } A = 0.6 \]
\[ \text{Cyprinus carpio, } A = 3.2 \]
\[ \text{Pomatochistus microps, } A = 0.6 \]
the mouth) and the respiratory apparatus (i.e., the gills adjacent to slits on both sides of the anterior part of the alimentary canal). Water-breathing invertebrates lack this close connection between feeding and breathing, one reason why even the largest among them (giant squids) cannot reach the mass of the largest fish (20 metric tons, for the whale shark *Rhincodon typus*).

The reorganization of the head of early fish allowed larger gills to evolve, which allowed the higher metabolic rates required for swimming in open waters. This transition was assisted by the gradual loss of the heavy armor protecting the slow, bottom-scurrying agnathans. The fine “teeth” covering the bodies of sharks are vestiges of this armor.

Fast swimming in open water required better fins, both for propulsion and for steering. Propulsion is provided in most fish by oscillations of a caudal fin whose aspect ratio (Fig. 1C) gradually increased toward tunas and other derived, fast-swimming groups with very large gills. Steering, on the other hand, is provided by dorsal, pectoral, and anal fins. These fins are stiffened for precise action by hard, bony rays in the most derived fish, the teleosts, whose evolutionary success was further enhanced by a complexly built protrusile mouth that enables capture of a wide range of food items (Fig. 1B).

Subtle anatomical changes in fish can thus create more niches for increasing the numbers of specialists, which then occupy increasing numbers of closely packed niches. Ecosystems in which these changes have run for long periods, undisturbed by physical changes, therefore contain very large numbers of fish species. Their numbers are even larger in areas such as the Great Lakes of Africa and the tropical Indo-Pacific, where changes of water levels have repeatedly isolated basins and subpopulations, thereby accelerating species differentiation (Fig. 2).

### B. Reproduction and Recruitment

Though many ancient fishes such as sharks and rays or the coelacanth *Latimeria chalumnae* practice internal fertilization and produce few large eggs or live offspring, most recently evolved fishes produce numerous small eggs that are fertilized externally and develop as part of the plankton, without parental care. The larvae that emerge from those eggs, after less than one day in warm tropical waters and up to two weeks (and more for larger eggs) in cold temperate waters, are usually elongated, as befit small, finless zooplankton feeders.

The average zooplankton concentrations that these larvae encounter, even during spawning seasons attuned with zooplankton production cycles, are usually far too low to allow survival of fish larvae, and the overwhelming majority of such larvae perish. Those that tend to survive usually happened to have hatched within plankton-rich water layers. These layers are usually a few centimeters thick and last for only a few days of calm, between wind-driven or other mixing events, such as storms or upwelling pulses, that enrich surface waters with nutrients from deeper waters. This implies that large biomasses of fish can build up only when and where the local oceanographic conditions take the form of “triads” defined by (1) nutrient enrichment, such as generated by wind-driven mixing, (2) high plankton concentration, such as generated by various mechanisms including fronts, and (3) retention of larvae, required to prevent these weak swimmers from drifting away from suitable habitat. In pelagic fishes that build high biomass, for example, the anchovies and sardines in coastal upwelling systems off northwestern and southwestern Africa, Peru, and California, these triads occur only when the coastal winds range from 4 to 6 m per second. Weaker winds do not generate enough enrichment, and stronger winds disperse the larvae offshore.
Fish have developed several strategies to deal with the uncertain recruitment that results from the triad requirements. One is being small, short-lived, and capable of quickly building up large biomass under favorable environmental conditions. The other is being large, long-lived, and capable of weathering long series of recruitment failures through repeated spawning by old, large, and highly fecund adults. An example of the former strategy is provided by the Peruvian anchovy Engraulis ringens, whereas the northern cod, Gadus morhua, provides an example of the latter. Yet another strategy is to reduce the dependence on environmental conditions by various forms of parental care, such as nesting and guarding (e.g., in catfishes, family Clariidae), mouth-brooding (e.g., in cardinal fishes, family Apogonidae), and live-bearing (e.g., in ocean perches, genus Sebastes).

Another important feature of fish stocks is that, contrary to earlier assumptions of homogeneity, most appear to consist of well-differentiated individuals, each tending to reproduce at the very place where it was hatched. Or, put differently: most migratory fish tend to “home.” This behavior, well documented only in Pacific and Atlantic salmon (Oncorhynchus and Salmo, respectively), implies that individual fish, when reproducing, do not seek “optimal” sites, but rather spawn as close as possible to the site at which they hatched, and to which they are imprinted. This tendency to either stay in or return to a certain area makes it difficult for fish stocks to rebuild once they have been decimated by local overfishing or pollution.

II. RESPIRATORY CONSTRAINTS TO GROWTH AND RELATED PROCESSES

A. Basic Geometrical Constraints

Fish growth, as in other animals, requires both food and oxygen, the latter being required to synthesize the substance (adenosine triphosphate or ATP) that serves as fuel to all organisms. For oxygen to be metabolically available, it must be inside the fish body, that is, it must have passed through its gills. Thus, since oxygen cannot be stored inside the fish body (contrary to food, which can be stored as gut contents and as fat), the metabolic and growth rate of fish are largely proportional to the surface area of their gills. So fish that quickly reach large sizes have gills with large surface areas (as in tunas), and conversely in slow-growing fishes (like groupers). Moreover, gill area per unit of body mass declines with size, because the two-dimensional gill area cannot keep up with the three-dimensional increase of body mass. Hence larger fish dispose of relatively less oxygen to supply their metabolism, the reason why they ultimately stop growing. Also, environmental factors that tend to increase metabolic rate—especially elevated temperatures, but also including other form of stress—have the effect of reducing the maximum size that the fish of a given population can reach (Figs. 3A and 3B). This is why tropical fish tend to be smaller than their respective cold-water relatives. A similar mechanism explains the nearly constant relationship in fish between size at first maturity and maximum size (Figs. 3C and 3D).

B. Adaptation to Respiratory Constraints

Fish have evolved various strategies and tactics to overcome respiratory constraints. One strategy, illustrated in Fig. 1B, is to evolve large gills, a route taken by numerous open-water (“pelagic”) species, culminating in tunas (Fig. 4).

Another strategy is the evolution of life cycles in which the juveniles migrate to deeper, cooler waters as they grow and then, upon maturing, produce eggs that quickly float up to the warmer surface layers, out of reach of the often cannibalistic adults. Such typical cycles are completed by an onshore drift of the larvae to coastal areas, and productive shallow nurseries for the early, voracious juveniles, which again migrate into deeper waters as they grow.

A tactic to accommodate metabolic stress, which is particularly useful in areas with strong seasonal temperature oscillations, is for the feeding adults to store fat during the warmer part of the season (late summer to early fall). Fat requires far less oxygen for maintenance than protein of muscle and other tissues. As temperature declines, the accumulated fat is converted into other tissues, notably gonads, whose contents are shed in spring, thus reducing body mass when temperatures again start to increase. These cycles, which use fat as protection against respiratory stress, are the reason why temperate fish tend to contain more muscle and visceral fat than tropical species, where temperatures, although high, do not fluctuate much in the course of a year. Another tactic that delays respiratory stress is associated with ontogenetic shifts in diet composition. Here, the young fish feed on a diffuse, small prey (e.g., invertebrate zooplankton), while the adults, via their sheer size, can capture energy-rich prey such as other fish, which are acquired at lesser cost by the predator.
C. Relationships between Growth and Mortality

Whichever strategy and tactic fish use to grow, more time will be needed in large species than in small fish for the size at first reproduction to be reached. Large sizes thus imply, other things being equal, more time during which the growing fish may become the prey of some predator. Hence the evolution of large fish was coupled with a reduction of their relative vulnerability to various predators, mainly by their ability to grow quickly through “small-size” stages in which mortality is highest. Fish capable of reaching large size and that have a high longevity also have low rates of natural mortality (Fig. 5). Hence fishing tends to have a stronger impact on species with low natural mortality, such as sharks or rockfishes. Because these are often the top predators, their reduction tends to disrupt the food webs in which they are embedded.
III. DISTRIBUTION OF EXPLOITED FISH STOCKS

A. Overall Distribution Ranges

Although mostly confined to water, fish occur in a wider range of habitats than any other vertebrate or invertebrate group. Thus, fish range from the upper reaches of streams in high mountain ranges (e.g., many river loaches, Balitoridae) to the mouths of temperate and tropical rivers (e.g., many gray mullets, Mugilidae). In the marine realm, fish range from the intertidal to the ocean’s abyss, both as predators in their desert-like expanses (e.g., skipjack tuna, Katsuwonus pelamis) or as components of the rich, newly discovered deep-sea vent ecosystems (e.g., some live-bearing brotulas, Bythitidae). Environmental adaptations include the ability to deal with an enormous range of pressures (from...
about one to hundreds of atmospheres), temperatures (from \(-1.8\)°C in polar waters to about 40°C in hot springs, tolerated by some tilapias), and salinities (from close to distilled water preferred by the discus fish, Symphysodon discus, of Amazonia to about 10%, e.g., in West African hypersaline coastal lagoons inhabited by the blackchin tilapia, Sarotherodon melanotheron), to list only three environmental factors. No single fish species or family, however, spans more than small fractions of these ranges. Rather, these various adaptations are exhibited by a bewildering variety of forms, ranging from minute gobies that are fully grown at close to 1 cm (e.g., Mystichthys laevis) to the 15 m reached by whale sharks (Rhincodon typus). These two species, incidentally, are exploited for food in the Philippines. The former, despite its turnover rate, is in danger of extinction in the small lake where it is endemic because of overfishing and pollution. The latter will be exterminated if the new directed, export-oriented fishery for tuna species or family. However, spans more than small fractions of these ranges. Rather, these various adaptations are exhibited by a bewildering variety of forms, ranging from minute gobies that are fully grown at close to 1 cm (e.g., Mystichthys laevis) to the 15 m reached by whale sharks (Rhincodon typus). These two species, incidentally, are exploited for food in the Philippines. The former, despite its turnover rate, is in danger of extinction in the small lake where it is endemic because of overfishing and pollution. The latter will be exterminated if the new directed, export-oriented fishery for this slow-growing fish continues.

B. Adaptations to Open-Ocean Habitats

Fish have different strategies to deal with the low productivity of the oceans. Tuna have adopted a high-energy strategy, wherein their tightly packed schools quickly move from one food patch to the other, essentially hopping from one “oasis” to the next and minimizing the time spent in the intervening desert-like expanses. Others, notably the lantern fishes (Myctophidae), occur in scattered populations that, at dawn, migrate from 1/00 m down to the surface waters, and back again at dusk. These different strategies imply very different biomasses: tens of millions of metric tons for the major tuna species (prior to their recent depletion by various longline, purse seine, and other fisheries) against an estimated global biomass of one billion metric tons for the lantern fish and associated communities. The latter number is often viewed as a promising figure, from which various estimates of potential yields have been derived. Most of these estimates, however, do not consider the extremely dilute nature of this biomass (usually less than 1 g per metric ton of water).

C. Shelf Communities

1. Definition of Neritic Stocks

Most fish stocks are neritic, that is, occur above the continental shelves, the productive areas of shallow waters (down to 200 m) around the continents, from which about 90% of the world marine fisheries catches are extracted. Shelves may have rocky or soft (sandy or muddy) substrates, and usually support two weakly connected fish communities, one species-rich and consisting of bottom or “demersal” fishes, the other consisting of fewer species of open-water or “pelagic” fishes. The fish of demersal communities are those exhibiting the specialized fins and mouths mentioned earlier, enabling utilization of distinctive food sources, particularly on reefs in both temperate and tropical regions.

On coral reefs, this fine partitioning of resources culminates in hundreds of fish species sharing a single reef, with dozens of specialists for each of its food resource types, from the filamentous algae consumed, for example, by damselfishes (Pomacentridae), the encrusting algae consumed by parrot fishes (Scaridae), the coral themselves, consumed by butterfly fishes (Chaetodontidae), to the small invertebrates consumed by, for example, wrasses (Labridae). A vast array of predators such as groupers (Serranidae) and sharks (Carcharhinidae) regulate the number of these smaller fishes. Hard-bottom shelves and, in tropical areas, the coral reefs that occur down to 30 m are also exploited wherever they occur. The fishing gear used over hard bottoms are mainly traps and handlines (the latter both sport and commercial), which are rather selective gears that would have relatively minor impacts were it not for their excessive numbers.

2. Demersal Fish Stocks

The demersal fish living in, on, or just above shelf soft bottoms consist of specialized flatfishes and rays and numerous generalized teleosts feeding on bottom invertebrates (the zoobenthos) and smaller fishes. The complex communities thus formed can reach very high biomass, at shallow depth in the tropics (20–50 m) and deeper in colder waters. In the warm waters of the tropics, bacteria induce a quick remineralization of the dead organic matter (detritus) falling out of the lighted part of the water column. This allows very little detritus to become available for consumption by the zoobenthi. In cold water, on the other hand, the short but intensive burst of algal production occurring in the spring is consumed only partly by the zooplankton of the upper water layers. Most of the remainder is consumed as detritus after falling down to the sea bottom as “marine snow.” Thus, cold-water soft-bottom communities can occur in very deep waters, down to the shelf slopes (200–300 m) and well beyond. Indeed, the latest trend in fisheries “development” is the exploitation of deep-sea stocks of cod-like fish (order Gadiformes), orange roughy (Hoplostethus atlanticus), and other fish, down to depths of 1000 m or more, through...
ventures that even in principle could never be managed so as to achieve sustainability.

Wherever they occur, soft-bottom shelves are nowadays invariably subjected to bottom trawling, a very unselective fishing method that is environmentally damaging. This involves dragging a heavy, chain-studded net over the sea bottom and “catching,” that is, removing all that it encounters. Not surprisingly, this procedure has often been compared to harvesting crops with a bulldozer. Trawler catches thus consist of target species (usually shrimps in the tropics and subtropics) plus a vast number of nontarget species, often the juveniles of demersals with large adult sizes, and literally parts of the habitat of bottom-fishes, notably sessile invertebrates and chunks of reefs lifted from the sea bottom. Nontarget species and debris are then discarded, and it is therefore trawlers that contribute most to the global discarding problem. Presently, about 30 million metric tons of various fish species are discarded; this is a very high discard rate when compared to the 90 million metric tons that appear in global landing statistics.

The contribution of trawlers to habitat destruction, including conversion of richly structured bottom habitats into featureless expanses of mud, is well recognized, and can only be compared in terms of scale with global deforestation and the ensuing trend toward desertification. Only recently has the impact on biodiversity of this mode of fishing begun to be evaluated in systematic fashion. The information so far available indicates high impacts and a tendency for small generalized fish and invertebrates to replace larger specialized fish, a trend that amplifies the food web effects to be described later.

3. Pelagic Fish Stocks

The pelagic communities over most shelf areas previously consisted of both major and minor stocks and stocklets of herrings, sardines (Clupeidae), anchovies (Engraulidae), and their relatives, and of their predators, notably mackerels and tunas (Scombridae) and various jacks (Carangidae). In many parts of the world, pelagic fisheries have eliminated the minor stocks and stocklets, and now depend wholly on annual recruitment to the remaining major stocks. The overfishing of old, highly fecund adults in these remaining stocks explains much of their volatility. Indeed, the present emphasis of much fisheries research on “variability” is thus devoted largely to a secondary phenomenon created by the fishery itself. It is true, however, that pelagic stocks, feeding lower in the food web, often closely track environmental changes, such as the decline of the Peruvian anchovy Engraulis ringens during El Niño events, and their subsequent rebuilding, mainly from recruits produced off northern Chile.

Pelagic fish tend to form tightly structured, dense schools, which protects them from predators and facilitates detection and herding of scattered food patches. The fisheries rely on this behavior when deploying purse seine, which can surround and catch such schools in one go, often with associated predators such as dolphins. Large pelagics such as billfish (Sphyridae and Istiophoridae) are caught by arrays of longlines, set by the thousands along shelf edges, which also capture, besides the target species, large amounts of by-catch (notably sharks). These sharks were previously left on the spot, but are now finned before the carcasses are discarded. Longlines are indeed as unselective as the now banned giant drift nets that, in the 1980s, erected “walls of death” that were hundreds of kilometers long across the migratory routes of fish in the North Pacific and the Atlantic.

4. Overall Status of Neritic Stocks

When combined, the demersal and pelagic fisheries of shelves and adjacent waters represent major threats to fish biodiversity. Particularly endangered are groupers and other slow-growing bottomfish, and pelagics such as bluefin tuna and various species of sharks and billfish. Besides the fisheries, one factor contributing to this endangerment is the traditional separation of research devoted to fisheries management (“stock assessments”) from that devoted to conservation and to ecosystem research. Both lines of research are separated institutionally, in terms of their methods and publication outlets, and in terms of what they perceive as their mandates. Overcoming this separation is crucial if fish biodiversity is to be maintained in the face of the onslaught by fisheries. Key needs are the development of tools and concepts for integrating information on fish biodiversity and ecosystem function with the knowledge gained through a century of applied, single-species fisheries research. Before considering these, however, evidence for fisheries impacts on ecosystems will be presented.

IV. ECOSYSTEM IMPACTS OF FISHERIES

A. Historical Trends

The earliest fishing gear so far identified by archologists are bone harpoons that were recovered, along with other evidence of systematic fishing, from a site 90,000
When subjected to less fishing, was required for the another demonstration of stocks rebuilding themselves of depleted fish stocks. Yet the Second World War, and first marine protected areas), would lead to a recovery fishing grounds by underwater mines (thus creating the vessels into the war effort, and the spiking of major fishing effort, as caused by the drafting of fishers and grams. It also established that a strong reduction of The First World War put an end to the stocking pro-

tries (notably Norway and the United States) initiated costly programs wherein juvenile cod and other fish were raised in hatcheries and then thrown into the sea, therein soon followed, after the introduction of bottom trawling, by coastal stocks of demersal fishes. The practical response to this was the introduction of bigger boats with bigger engines, fishing farther offshore. Another response was the creation of research bodies (such as the International Council for the Exploration of the Sea, founded in 1902) to assess the reason why the resources were declining. Also, several countries (notably Norway and the United States) initiated costly programs wherein juvenile cod and other fish were raised in hatcheries and then thrown into the sea, in the vain hope that they would replenish the stocks rather than be eaten by happy predators (which they were).

B. Emergence of the Sustainability Concept

The First World War put an end to the stocking programs. It also established that a strong reduction of fishing effort, as caused by the drafting of fishers and vessels into the war effort, and the sipping of major fishing grounds by underwater mines (thus creating the first marine protected areas), would lead to a recovery of depleted fish stocks. Yet the Second World War, and another demonstration of stocks rebuilding themselves when subjected to less fishing, was required for the notion of sustainable fishing to establish itself. This notion implies that some appropriate level of fishing effort (number of vessels or gear, mesh size) exists such that catches (or "yield") can be maintained at high levels—hence the concept of "maximum sustainable yield" or MSY. This led to the emergence of "fish population dynamics" and "stock assessments," wherein mathematical models of single-species fish stocks and of their response to targeted fishing became the mainstay of fisheries research. R. J. Beverton, S. J. Holt, and J. A. Gulland in England, W. E. Ricker in Canada, and W. E. Schaefer in the United States proposed most of these still-used models during an extremely creative period lasting from the early 1950s to the mid-1970s. Yet in spite of these advances, the fisheries never became sustainable. One obvious reason was that, given a resource to which access was essentially open, the fisheries never could limit their collective effort at the level supposed to generate MSY. Rather, effort levels increased well beyond that, permitting some fleet owners to increase their stakes even as the aggregate "rent" from the fisheries declined. Recent trends toward subdivision of offshore and distant water fleets, driven by international competition, have aggravated these economic issues, enabling commercial profits to be gained even from strongly overexploited stocks. These developments are so widespread that they have rendered obvious the impacts which fisheries have on ecosystems.

C. Fishing Down Marine Food Webs

The ecosystem impacts of fisheries are due mainly to the fact that the targeted fish function as part of food webs, both as consumers and as prey. Within food webs, the fish of different species occupy distinct trophic levels (TL), each defining a step away from plants, which have a definitional TL of 1. Thus, fish feeding on planktonic algae have TL = 2, fish feeding on herbivorous zooplankton have TL = 3, and so on. It is important here to recognize that most fish tend to have intermediate TL values (2.7, 3.5, 4.1, etc.), reflecting the catholic nature of their diet. Fisheries, by removing biomass from one of several fish stocks, necessarily modify food webs, thus forcing predators of the targeted species to shift toward available alternative prey, if any. Such adjustments were previously not distinguishable from natural fluctuations. They have gradually become highly visible, however, because they change the mean trophic level of the landings extracted from different stocks. Moreover, the changes induced by fishing are not of a random nature,
with decreases in one area matched by increases in another. Rather, they are directed, with a clear downward trend (Fig. 6A), due to the link between growth and natural mortality mentioned in Section II. Thus, in large fish, even a low level of fishing mortality generated by a well-managed fishery will quickly exceed the low level of total mortality (i.e., natural + fishing mortality) that can be accommodated by the stock. By-catch species are even more endangered because the fishing will not stop as their numbers dwindle until they are eradicated, as has happened with rays in the Irish Sea. The trend of mean trophic level resulting from this (see Fig. 6A), reflecting a phenomenon now known as "fishing down marine food webs," provides a clear indication that, globally, fisheries generate levels of effort well past those required for sustainability, however defined. Indeed, other indices can be used to indicate that global changes have occurred in the composition of global fisheries landings, and in the structure of the ecosystems from which these landings are extracted (Fig. 6B).

Fisheries-induced modification of the structure of marine and freshwater ecosystems has strong indirect impacts on fish biodiversity, in addition to the direct impacts of reducing the biomass of the target and associated stocks by a factor of 10 or more, as is usually the case. Incorporating these indirect effects in fisheries stock assessments has proven to be difficult so far. This is true for objective reasons (ecosystems are complex, and their behavior under exploitation, due to the large number of stocks to be considered, is difficult to simulate) and for subjective reasons (notably a perceived lack of suitable field data on these many stocks).

The recent development of robust ecosystem simulation tools should allow the first of these issues to be addressed. Overcoming the second not only involves pointing out the existence of suitable data, often lost in the "gray literature," but in making such data available in suitable format to all who are aware of the need for a transition from single-species to ecosystem-based fisheries assessments. This brings us to the issues related to the standardization, dissemination, and uses of biodiversity information.

V. MANAGING FISH BIODIVERSITY INFORMATION

A. Biodiversity as a Conceptual Challenge

There is a widespread perception that the main obstacle to the conservation of fish stocks and of fish biodiversity is "lack of data," a notion strengthened by public statements of biologists worried about the lack of funding for relevant research. However, simple lack of data cannot be the problem, not after the 250 years since Lin-
naeus created the taxonomic standards required for biodiversity research, 100 years of applied fisheries research, and at least 30 years of advances in ecosystem research. Rather, the problem here is the fragmentation of the database collected so far. Indeed, many studies conducted in recent years on the status of various stocks fail to consider previous knowledge on their relative abundance and distribution, and thus contribute to shifting baselines, wherein only the most recent and usually low estimates are used as reference for conservation or rebuilding efforts.

One reason for this reluctance of biologists to consolidate existing data into comprehensive, global databases may be due in part to the perception that biological data are too difficult to standardize, or are useless once standardized. Addressing these issues will be a key task of biodiversity research, and we now present a few ideas related to this.

There is consensus that the objects of biodiversity research are genes, populations, species, and ecosystems. However, there is little consensus as to what distinguishes biodiversity from the existing disciplines of fisheries biology, ecology, biogeography, population genetics, or taxonomy. As a result, the array of data being claimed to be essential for biodiversity studies reads like a composite list of the data traditionally used in the older disciplines, with few attempts at integration or prioritization. Such integration and prioritization are possible, however, by giving emphasis, in biodiversity studies, to data that are: (1) relevant to current research issues (e.g., richness, rarity, distinctiveness, representativeness, threat, function, and utility of species); (2) part of the data traditionally collected in taxonomy, biogeography, population genetics, and ecology; (3) widely available, in sufficient quantity; (4) pertinent to past, present, and most likely future trends; (5) easy to collect; (6) easy to standardize; (7) easy to verify; and (8) suggestive of new lines of research.

**B. Bioquads as Key Biodiversity Data Sets**

A minimum core of biodiversity information that fulfills these eight criteria is provided by “bioquads” (from “quads,” short for quadrads), consisting of: (a) the scientific name of a taxon, usually a biological species or other evolutionarily significant unit; (b) the locality where a specimen of this taxon has been encountered; (c) the date (time) of the encounter; and (d) the authority or source reporting (a)–(c).

Of the research items mentioned under criterion (1), richness (number of species encountered) is derived directly from the bioquads from a given area. Distinctiveness (how much the species encountered differ from each other) is derived from the classification of these species into higher taxa such as families, orders, and classes. Representativeness (how closely an area represents a predefined ecosystem type) is derived by comparing observed species composition with the typical composition of the ecosystem type under study. The utility of species to humans can be derived from published, or local knowledge, or from catches in the case of fish. Status of threat can be derived from trends in the distribution area defined by bioquads. Rarity can be estimated from the number of bioquads available for a species in a given area, standardized by sampling effort.

Taxonomists have made a conscious effort to systematically compile data of this sort in specimen collections, and to publish them in original species descriptions and revisions. As a result, bioquad-type data are readily available in enormous numbers (about 10 million for fish alone) in museum collections, survey reports, historical photos and films, and other forms (criterion 3). While museum collections go back over 200 years, some literature contains verifiable records that date back to antiquity (criterion 4). Also, archeological data reach back to the dawn of modern humanity (see the earlier record pertaining to giant carfish).

Numerous scientific surveys and projects also continuously collect contemporary bioquads. Other sources are the commercial fisheries and the many laypersons whose hobby is to observe and sometimes to collect fish and other wildlife. These activities are most likely to continue in the foreseeable future (criterion 5). An increasing number of the preceding data sources are available in computer-readable form (criteria 3, 5, and 6).

Efforts do exist to standardize the elements of the bioquad (criterion 6). For example, the Species 2000 Initiative has embarked on the task of providing a standard reference list of the valid names of the known 1.75 million species sharing Earth with humans (see the website www.sp2000.org). Geographical coordinates and the international date and time format are obvious standards for items (2) and (3), although there remains a need for a global gazetteer to deal efficiently with localities reported without coordinates, and there is a need for standards to deal with date and time ranges. On the other hand, standards exist for sources such as printed publications, databases, photos, films, and personal communications. Many of these were considered when developing FishBase, a computerized data-
base on the biology, ecology, and uses of fish containing a vast number of bioquads (see the following).

The necessary verification (criterion 7) of millions of data points can only be done automatically. Basically, a computer can verify a scientific name against a standard list, compare the indicated locality and date against the established range of a species, and judge the reliability of a source, for example, by the number of outliers it has reported previously. Procedures will have to be established, however, on how to deal with the different types of outliers, some of which may represent valid new information.

An important consideration is how fast a research agenda based on bioquads will be exhausted (criterion 8). Important here is the ability of well-structured relational databases to interlink independently developed data sets. Thus, the scientific name links to all available information on a species, including taxonomy, systematics, genetics, biology, ecology, and human uses. The locality connects to all available information on surrounding environments, including province, country, continent, habitat, ecosystem, and tectonic plate. The combination of species, locality, and date points to a population or stock. Date and time in connection with the locality can be used to infer a wide range of environmental conditions, from local temperatures to current fisheries legislation. The source relates to the human dimension, such as persons and institutions working on certain species groups or in a certain area, representing the scientific interface between humans and the other species (Fig. 7).

C. Databases as Tools for Management of Biodiversity Information

Two major initiatives presently exist to assemble and make widely available, for research on fish biodiversity, the information presently held by various institutions (notably museums). One is NEODAT, which makes accessible on the Internet about 400,000 bioquad records pertaining to freshwater fish of the Neotropics (NEODAT; www.fowler.acnatsci.org). The other is FishBase, an ongoing international collaborative project dedicated to assembling the estimated 10 million existing fish bioquads and to combining them with other, standardized biological information on fish. The intention here is to provide a global relational database, addressing head-on the data fragmentation issue mentioned earlier (see www.fishbase.org).

Figure 8 shows the geographic distribution of Nile tilapia, Oreochromis niloticus, through dots representing bioquads as defined previously. Important here is that a new original of this graph is generated on the fly, from currently available bioquads, every time the relevant routine of FishBase is evoked, and that each of its “dots” can be clicked to verify the four elements of the underlying bioquad.

VI. PRESERVING FISH BIODIVERSITY

A. Traditional Approaches to Stock Management

None of the foregoing considerations will help, however, if fisheries are allowed to continue undermining their resource base, which they will if fisheries management continues to rely on the panoply of approaches so far deployed. These traditional approaches include, among other things: (1) mesh size restriction; (2) restriction on the amount and/or species of fish that may be legally landed; (3) effort limitation, for example, through caps on the vessel tonnage that may deployed; and (4) seasonal closures.

Besides being extremely hard to enforce, these approaches—which are invariably conceived in the context of single-species assessments—fail to address the ecosystem effects mentioned earlier. Thus, mesh sizes
above a certain limit, meant to protect the young of a given species, do not prevent associated species from being caught. Indeed, when combined with restrictions on total allowable catch (TAC), and on the landing of bycatch (as is often the case), mesh size restrictions become the very reason for discarding both the young of targeted species and the nontarget species. Limits on nominal fishing effort are subverted by technological developments, such as improved gears and navigation instruments (e.g., GPS), which increase the catching power of fishing vessels. Thus, government-run vessel retirement schemes often end up subsidizing the modernization of fishing fleets. Finally, seasonal closure of various areas usually has negligible ecological impacts, because the fishing effort expended during the open season is sufficient for the sea bottom to be scraped up numerous times by trawlers, and for the stocks of long-lived fishes to be severely impacted.

B. Marine Protected Areas
There is an emerging consensus among fisheries scientists and conservationists that the only fisheries management tool that will allow the recovery of damaged stock and ecosystems is the establishment of Marine Protected Areas (MPAs), including permanent No-Take zones as their core. Such core zones are easy to enforce—at least relative to the task of enforcing mesh sizes or TACs. Also, technology-driven increases of fishing effort can be ignored, and there is assurance that the long-lived organisms of seafloors and their associated fish communities can gradually return to a semblance of their original configurations. However, much research will have to be devoted to identifying the optimal size and location of MPAs, particularly for migratory stocks.

Still, traditional fisheries management, aimed at lim-
iting effective fishing effort, will have to continue around MPAs, lest they become marine larders or fish-attracting rather than fish-producing zones from which resources are drained by fisheries operating at their very periphery.

Finally, the social context of fisheries will have to change: fisheries do not harvest crops they have sown. Rather, they exploit the natural productivity of wildlife; thus there are inherent limits to global fish catches, and future fisheries will not meet the demand of an ever-increasing human population. Indeed, the massive ecosystem changes already described indicate that these limits have been reached in most parts of the world, and that sustainable fisheries must be embedded in some form of ecosystem management.

See Also the Following Articles
ADAPTATION • FISH, BIODIVERSITY OF • FISH CONSERVATION • MARINE ECOSYSTEMS

Bibliography
Froese, R., and D. Pauly (eds.). (1998). FishBase 98: Concepts, Design and Data Sources. ICLARM, Manila. [Distributed with two CD-ROMs; also see the website www.fishbase.org.]
FLIES, GNATS, AND MOSQUITOES

Brian V. Brown
Natural History Museum of Los Angeles County

I. Introduction
II. Major Subdivisions
III. Life History Diversity
IV. Conservation Biology of Diptera

GLOSSARY

Diptera Group of insects to which the flies, including gnats and mosquitoes, belong.
larva (pl. larvae) Immature stage of flies, often called maggots.

THE DIPTERA, or true flies, are a well-established, monophyletic group of insects with more than 124,000 extant species. They are found worldwide and interact with the environment at almost all trophic levels as scavengers, filter-feeders, herbivores, predators, parasitoids, and parasites.

I. INTRODUCTION

The insect Order Diptera, with more than 124,000 currently described, extant species, ranks as one of the worlds largest groups of organisms. Along with the other insect megadiversity groups—Coleoptera (beetles; 350,000 species), Lepidoptera (butterflies and moths; 120,000 species), and Hymenoptera (ants, bees, wasps, and sawflies; 130,000 species)—they form the largest aggregation of species on the planet. Each of these megadiversity groups has more species assigned to it than to any other group of organisms except plants (300,000 species). Like other insects, most species of Diptera are still undescribed, and the actual number could range as high as 1 million or more species.

Diptera are found on every continent, including Antarctica. The relative percentage of Diptera within the insect fauna rises with latitude (and elevation) as other, less cold-adapted taxa are lost; for instance, over one-half of all insect species recorded from the Canadian high arctic are Diptera. The extant species of all biogeographical regions have been catalogued (Table I), as have the fossil Diptera. Some catalogs are badly outdated and the number of species described for each region is often more indicative of the amount of taxonomic activity directed at a given fauna than the true diversity. For instance, the Palearctic Region is the best studied and has more described species than the relatively poorly known, but presumably more diverse, Neotropical Region. Therefore, conclusions about dipteran biogeography based on these numbers will be highly inaccurate.

Diptera are a well-established monophyletic group, with the most obvious defining character being the reduction of the hind wings to small, club-shaped organs called halteres. The insect orders considered to be most closely related to Diptera are Siphonaptera (fleas) and Mecoptera (scorpionflies), although recent molecular
research indicates that the highly aberrant Strepsiptera (twisted-wing parasites) may be their closest relatives. Fossil Diptera are common in the amber faunas of the world, which however extend back in time only to the early Cretaceous. The oldest dipteran fossil is believed to be from at least the mid-Triassic. Like the other megadiversity groups, the Diptera are holometabolous insects, with a separate egg, larva, pupa, and adult stage. Generally, most feeding takes place in the larval stage, whereas the adult is usually specialized for reproduction and dispersal. Some adult Diptera are voracious feeders, however, requiring substantial supplemental feeding to mature their eggs or power their flight.

Flies are common in natural, disturbed, and urban habitats. Larvae are found on land and in freshwater; there are relatively few marine or brackish water species. Some species are synanthropic and have been transported around the world with human activities. Diptera are the most important vectors of human and animal disease, and a few plant-feeding species have become agricultural pests. The life history of most species of flies, however, is unknown. The group is so large, and there are so many undescribed species, that the science of dipterology is still in its relative infancy.

TABLE I
Biogeographical Regions and the Number of Described Diptera Species

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearctic</td>
<td>25,000</td>
</tr>
<tr>
<td>Neotropical</td>
<td>20,000</td>
</tr>
<tr>
<td>Palearctic</td>
<td>29,000</td>
</tr>
<tr>
<td>Afrotropical</td>
<td>16,318</td>
</tr>
<tr>
<td>Australasian</td>
<td>15,764</td>
</tr>
<tr>
<td>Oriental</td>
<td>15,964</td>
</tr>
<tr>
<td>Antarctica</td>
<td>60</td>
</tr>
</tbody>
</table>

Compiled from the various catalogs.

II. MAJOR SUBDIVISIONS

Table II presents a general list of the families of Diptera and some of the higher taxa that contain them. This list is based on only one of several possible classifications, however, and cannot be considered the final word in dipteran groupings. Common names, where well established, are given in Table II. Many families lack common names, including the second largest family in the order (Tachinidae). That a group with more described species than the mammals lacks a common name is a good indication of the lack of general appreciation for the importance and ubiquity of the Diptera. The number of described species in each family is given, but in many instances these numbers are badly out of date and should be considered a bare minimum. Certainly, in most families of Diptera there are a large number of undescribed species awaiting the attention of specialists. It is estimated that we have described only 10% or fewer of the species in many families.

The Diptera are traditionally organized into three suborders: the Nematocera, the orthorrhaphous Brachycera, and the cyclorrhaphous Brachycera. Of these, it is likely that only the Cyclorrhapha is a monophyletic, or natural, group. Although a consensus on the higher relationships within the Diptera is not yet available, it is generally agreed that the “Nematocera” is a paraphyletic assemblage of relatively primitive dipteran families and that some subgroup of this assemblage is more closely related to the Brachycera than to other nematocerans. Similarly, within the Brachycera, a group widely considered to be monophyletic, the orthorrhaphous families are now believed to be a paraphyletic assemblage relative to the Cyclorrhapha. Finally, within the Cyclorrhapha is a group of flies traditionally called the Aschiza (here called the lower Cyclorrhapha), which is probably paraphyletic with respect to the monophyletic Schizophora. The details of the phylogeny of the Diptera has been reviewed extensively by Yeates and Wiegmann and their discussion of various groups is highly recommended for understanding the current status of dipteran phylogeny.

Within the Diptera, higher taxa show repeated patterns of relatively primitive, nondiverse grades of organization with relatively highly derived, speciose sister taxa. Thus the lower Diptera are collectively much less diverse than the Brachycera, with the notable exception of the incredibly large family Tipulidae. Within the Brachycera, the Muscomorpha is by far the largest infraorder and the Schizophora has four times more species than the lower Cyclorrhapha. The numbers are less disparate in the Schizophora, with about 26,000 acalypters and 19,000 calypters, but the monophyly of the acalypters is highly contentious, making such a comparison questionable.

The distribution of species among the families of Diptera is extremely divergent, with the largest number found in the Tipulidae and the Tachinidae. Together,
<table>
<thead>
<tr>
<th>Lower Diptera (=Nematocerous groups)</th>
<th>Number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>### Infracrorder Psychopteromorpha</td>
<td></td>
</tr>
<tr>
<td>Ptychopteridae</td>
<td>61</td>
</tr>
<tr>
<td>Tanyderidae</td>
<td>42</td>
</tr>
<tr>
<td>### Infracrorder Calcomorpha</td>
<td></td>
</tr>
<tr>
<td>Superfamily Culicinida</td>
<td></td>
</tr>
<tr>
<td>Culicidae</td>
<td>3,000</td>
</tr>
<tr>
<td>Drosidae</td>
<td>175</td>
</tr>
<tr>
<td>Corethrellae</td>
<td>61</td>
</tr>
<tr>
<td>Chioboridae</td>
<td>90</td>
</tr>
<tr>
<td>Superfamily Chironomidae</td>
<td></td>
</tr>
<tr>
<td>Ceratopogonidae</td>
<td>5,360</td>
</tr>
<tr>
<td>Chironomidae</td>
<td>5,000</td>
</tr>
<tr>
<td>Simuliidae</td>
<td>1,475</td>
</tr>
<tr>
<td>### Infracrorder Blephariceromorpha</td>
<td></td>
</tr>
<tr>
<td>Blepharicerida</td>
<td>300</td>
</tr>
<tr>
<td>Deuterophlebiida</td>
<td>14</td>
</tr>
<tr>
<td>Symphyomyidae</td>
<td>7</td>
</tr>
<tr>
<td>### Infracrorder Bibionomorpha</td>
<td></td>
</tr>
<tr>
<td>Axymyiida</td>
<td>5</td>
</tr>
<tr>
<td>Bibionidae</td>
<td>875</td>
</tr>
<tr>
<td>Corethomidae</td>
<td>4,800</td>
</tr>
<tr>
<td>Mycetophilidae</td>
<td>3,000</td>
</tr>
<tr>
<td>Pachytmidae</td>
<td>4</td>
</tr>
<tr>
<td>Seriidae</td>
<td>1,000</td>
</tr>
<tr>
<td>### Infracrorder Tipulomorpha</td>
<td></td>
</tr>
<tr>
<td>Tipulidae</td>
<td>14,000</td>
</tr>
<tr>
<td>Trichoceridae</td>
<td>110</td>
</tr>
<tr>
<td>### Infracrorder Psychodomorpha</td>
<td></td>
</tr>
<tr>
<td>Anisopyidae</td>
<td>320</td>
</tr>
<tr>
<td>Perissomidae</td>
<td>5</td>
</tr>
<tr>
<td>Psychidae</td>
<td>2,700</td>
</tr>
<tr>
<td>Scatopidae</td>
<td>235</td>
</tr>
<tr>
<td>Symmeidae</td>
<td>4</td>
</tr>
<tr>
<td>### Brachycera (an unranked taxon)</td>
<td></td>
</tr>
<tr>
<td>### Infracrorder Xylophagomorpha</td>
<td></td>
</tr>
<tr>
<td>Xylophagidae</td>
<td>111</td>
</tr>
<tr>
<td>### Infracrorder Tabanomorpha</td>
<td></td>
</tr>
<tr>
<td>Atheridae</td>
<td>90</td>
</tr>
<tr>
<td>Pelcophrynidae</td>
<td>40</td>
</tr>
<tr>
<td>Rhagionidae</td>
<td>520</td>
</tr>
<tr>
<td>Tabanidae</td>
<td>3,000</td>
</tr>
<tr>
<td>Vermilomidae</td>
<td>31</td>
</tr>
</tbody>
</table>

*TABLE II*

Simplified List of Major Subgroups and Families of Diptera

<table>
<thead>
<tr>
<th>Common name of species</th>
<th>Number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phantom crane flies</td>
<td>61</td>
</tr>
<tr>
<td>Meniscus midges</td>
<td>175</td>
</tr>
<tr>
<td>Phantom midges</td>
<td>90</td>
</tr>
<tr>
<td>Biting midges</td>
<td>5,360</td>
</tr>
<tr>
<td>Midges</td>
<td>5,000</td>
</tr>
<tr>
<td>Black flies</td>
<td>1,475</td>
</tr>
<tr>
<td>Net-winged midges</td>
<td>300</td>
</tr>
<tr>
<td>Mountain midges</td>
<td>14</td>
</tr>
<tr>
<td>Fungus gnats</td>
<td>3,000</td>
</tr>
<tr>
<td>Dark-winged fungus gnats</td>
<td>1,000</td>
</tr>
<tr>
<td>Crane flies</td>
<td>14,000</td>
</tr>
<tr>
<td>Winter crane flies</td>
<td>110</td>
</tr>
<tr>
<td>Wood gnats</td>
<td>320</td>
</tr>
<tr>
<td>Sand flies, meth flies</td>
<td>2,700</td>
</tr>
<tr>
<td>Scatopidae</td>
<td>235</td>
</tr>
<tr>
<td>Ant luses</td>
<td>31</td>
</tr>
</tbody>
</table>

*continues*
<table>
<thead>
<tr>
<th>Infraorder</th>
<th>Common name of species</th>
<th>Number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infraorder Stratiomyomorpha</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pantophthalmidae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stratiomyidae</td>
<td>Soldier flies</td>
<td>2,500</td>
</tr>
<tr>
<td>Xylomyidae</td>
<td></td>
<td>107</td>
</tr>
<tr>
<td>Infraorder Muscomorpha</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superfamily Nematostomida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acroceridae</td>
<td>Small-headed flies</td>
<td>500</td>
</tr>
<tr>
<td>Nematostomidae</td>
<td>Tangle-tail flies</td>
<td>230</td>
</tr>
<tr>
<td>Superfamily Asilida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apoceridae</td>
<td>Flower-loving flies</td>
<td>149</td>
</tr>
<tr>
<td>Ailixidae</td>
<td>Rubber flies</td>
<td>3,600</td>
</tr>
<tr>
<td>Bombylidae</td>
<td>Bee flies</td>
<td>4,800</td>
</tr>
<tr>
<td>Mylabae</td>
<td>Mylas flies</td>
<td>593</td>
</tr>
<tr>
<td>Scenopinidae</td>
<td>Window flies</td>
<td>330</td>
</tr>
<tr>
<td>Theretraidae</td>
<td>Smile flies</td>
<td>850</td>
</tr>
<tr>
<td>Superfamily Empidida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dolichopodidae</td>
<td>Long-legged flies</td>
<td>3,300</td>
</tr>
<tr>
<td>Empididae</td>
<td>Dance flies</td>
<td>3,300</td>
</tr>
<tr>
<td>Cyclorrhapha (an unranked taxon of Muscomorpha)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acropliidae</td>
<td>Hover flies, flower flies</td>
<td>3,800</td>
</tr>
<tr>
<td>Schizophora</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cypselosomatidae</td>
<td>Sedge flies</td>
<td>30</td>
</tr>
<tr>
<td>Micropezidae</td>
<td>Sibyl-legged flies</td>
<td>300</td>
</tr>
<tr>
<td>Neriidae</td>
<td>Neriine flies</td>
<td>310</td>
</tr>
<tr>
<td>Superfamily Diopsidae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diopsidae</td>
<td>Stalk-eyed flies</td>
<td>113</td>
</tr>
<tr>
<td>Megamerinae</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>Nothylidae</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Pseliidae</td>
<td>Rust flies</td>
<td>200</td>
</tr>
<tr>
<td>Somatidae</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Strongylomyiidae</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>Syringogastidae</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Tanytarsidae</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>Superfamily Conopoidea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conopidae</td>
<td>Thick-headed flies</td>
<td>800</td>
</tr>
<tr>
<td>Superfamily Tephritoida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lonchopteridae</td>
<td>Spear-winged flies</td>
<td>35</td>
</tr>
<tr>
<td>Opetiidae</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Phoridae</td>
<td>Humpbacked flies, scuttle flies</td>
<td>3,200</td>
</tr>
<tr>
<td>Pipunculidae</td>
<td>Big-headed flies</td>
<td>660</td>
</tr>
<tr>
<td>Platypezidae</td>
<td>Flat-footed flies</td>
<td>215</td>
</tr>
<tr>
<td>Scatoceridae</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Syrphidae</td>
<td>Hover flies, flower flies</td>
<td>3,800</td>
</tr>
<tr>
<td>continued</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

818  FLIES, GNATS, AND MOSQUITOES

`continued`
<table>
<thead>
<tr>
<th>Family</th>
<th>Common name of species</th>
<th>Number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pallopteridae</td>
<td></td>
<td>52</td>
</tr>
<tr>
<td>Psychidae</td>
<td>Skipper flies</td>
<td>71</td>
</tr>
<tr>
<td>Platypecomatidae</td>
<td></td>
<td>1,000</td>
</tr>
<tr>
<td>Pyrgotidae</td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Richardidae</td>
<td></td>
<td>170</td>
</tr>
<tr>
<td>Tachinidae</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Tephritidae</td>
<td>Fruit flies</td>
<td>4,000</td>
</tr>
<tr>
<td>Superfamily Lauxanioida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coleophoridae</td>
<td>Breelie flies</td>
<td>90</td>
</tr>
<tr>
<td>Chamaemyiidae</td>
<td>Aphid flies, silver flies</td>
<td>230</td>
</tr>
<tr>
<td>Eurychoromyiidae</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Lauxaniidae</td>
<td></td>
<td>1,550</td>
</tr>
<tr>
<td>Superfamily Sciomyzoida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coelopidae</td>
<td>Seaweed flies</td>
<td>30</td>
</tr>
<tr>
<td>Dryomyzidae</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Helosciomyiidae</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Sciomyzidae</td>
<td>Marsh flies</td>
<td>515</td>
</tr>
<tr>
<td>Ropalomeridae</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Sepididae</td>
<td>Black scavenger flies</td>
<td>240</td>
</tr>
<tr>
<td>Superfamily Oponomyzoida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acarorhynchitidae</td>
<td>Leafminer flies</td>
<td>3,300</td>
</tr>
<tr>
<td>Clusiidae</td>
<td></td>
<td>217</td>
</tr>
<tr>
<td>Superfamily Agromyzoida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agromyzidae</td>
<td>Leafminer flies</td>
<td>3,300</td>
</tr>
<tr>
<td>Fergusoninidae</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Olotridae</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Superfamily Oponomyzoida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anthomyzidae</td>
<td></td>
<td>49</td>
</tr>
<tr>
<td>Margravidae</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Opomyzidae</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>Superfamily Antecoida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aestidae</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Aulacigastriidae</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Neurochaetidae</td>
<td>Upside-down flies</td>
<td>5</td>
</tr>
<tr>
<td>Perisciellidae</td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>Trichomyzidae</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Xenostridae</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Superfamily Carnoidea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australomyzidae</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Brachidae</td>
<td>Bee lice</td>
<td>7</td>
</tr>
<tr>
<td>Canacidae</td>
<td>Beach flies</td>
<td>113</td>
</tr>
<tr>
<td>Carnidae</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>Chloromyzidae</td>
<td>Fric flies</td>
<td>2,000</td>
</tr>
<tr>
<td>Cryptomyzidae</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Mychidae</td>
<td></td>
<td>190</td>
</tr>
<tr>
<td>Tephritidae</td>
<td></td>
<td>126</td>
</tr>
</tbody>
</table>

...continues
these families account for almost 20% of the species in the order.

The Tipulidae, or crane flies, are elongate, long-legged, somewhat fragile flies that are found nearly everywhere on earth. The adults are sometimes found at lights but are most often seen resting on vegetation. The larvae have a variety of habitats, including terrestrial—in soil, mosses, and decaying wood—freshwater, and intertidal. They are scavengers, herbivores, or predators.

In contrast to the eclectic habits of the Tipulidae, larvae of Tachinidae have a single way of life: parasitism. All known species are internal parasites of other arthropods, mostly other insects and, within the insects, mostly larval Lepidoptera. Eggs are either laid on or in the host or are broadcast in suitable areas. Larvae hatch from broadcast eggs and wait in ambush for hosts. Some tachinids produce microtype eggs that are designed to be ingested by the hosts; these eggs hatch inside the host and penetrate the gut wall to enter the body cavity. Adult tachinids are stout, bristly, housefly-like flies that are seen frequently on vegetation or flowers.

There are another 21 families of Diptera that can be considered large, possessing about 2000 or more described species. Together, these 23 largest families (including tipulids and tachinids) comprise about
100,000 species, or approximately 80% of the Order Diptera. In order of decreasing number of species, these other 21 large families are as follows:

- **Syrphidae**: hover flies, flower flies (3800 spp.). Adults of this family often are brightly colored mimics of Hymenoptera (bees and wasps). The larvae are saprophagous, predatory, or herbivorous. Some saprophagous species are called rat-tailed maggots because of their elongate posterior breathing tubes. Many of the predatory species live exposed on plants, feeding on aphids.
- **Asilidae**: robber flies (5600 spp.). The adults are voracious predators that usually attack their prey while in flight, stabbing them with their heavily sclerotized, swordlike proboscis and injecting digestive fluids. The larvae are also predatory.
- **Ceratopogonidae**: biting midges (3500 spp.). The tiny adult females of this family require blood meals to mature their eggs. Some species bite vertebrates, other feed on large insects, and still others are predatory on small insects. The larvae are predacious, living in damp terrestrial or freshwater habitats.
- **Dolichopodidae**: long-legged flies (5100 spp.). These flies are often metallic green in color and are found commonly on undergrowth or on tree trunks. The larvae and adults are predatory.
- **Chironomidae**: midges (5000 spp.). These flies are among the most abundant benthic invertebrates in freshwater environments; some are also terrestrial or intertidal. The adult males often form enormous mating swarms.
- **Bombyliidae** (including Mycetomyiidae): bee flies (4800 spp.). Species of this family are most diverse in dry areas, including deserts, and are nearly absent from tropical rain forests. The larvae are parasitoids of the immature stages of various other insects or predatory on grasshopper egg pods.
- **Cecidomyiidae**: gall midges (4600 spp.). Cecidomyiids are generally considered one of the largest, yet most poorly known groups of Diptera. Larvae often form galls on various plants, but there are non-gall-forming species and many mycophagous forms as well.
- **Tephritidae**: fruit flies (4000 spp.). Larval tephritids are phytophagous, attacking a wide variety of plants and sometimes forming galls. The adults often have color patterns on their wings.
- **Muscidae**: houseflies and relatives (3880 spp.). A few extremely well-known muscid species are synanthropic, especially *Musca domestica* and *Stomoxys calcitrans*. Larvae of most species live in decaying organic material, where they are either saprophagous or predatory. Some species are saprophagous in early larval instars, becoming carnivorous later.
- **Agromyzidae**: leaf-miner flies (3300 spp.). Along with some other insects, larvae of many species of agromyzids feed within plant leaves, excavating the distinctive, light-colored tunnels called mines. The adults are small, usually dark-colored flies; some species are marked with yellow.
- **Empididae**: dance flies (3300 spp.). Adults of this family are found mostly in damp terrestrial habitats; often near water. Common and diverse in temperate regions, they are less prevalent in tropical lowland forests. Larvae and adults of both sexes are predatory; larvae of a few species are parasitoids of caddisfly larvae.
- **Phoridae**: humpbacked flies, scuttle flies (3200 spp.). Phorids are common and diverse nearly everywhere except Antarctica. Larvae can be predators, parasitoids, true parasites, herbivores, or scavengers. Many species are associated with social insects.
- **Tabanidae**: horseflies, deer flies (3000 spp.). Adult female tabanids are well-known blood feeders, although some species feed only on nectar. The larvae are predacious, usually found near water.
- **Culicidae**: mosquitoes (3000 spp.). Adult females bite vertebrates to obtain blood meals, and they often transmit diseases. Larvae are aquatic.
- **Mycetophilidae**: fungus gnats (3000 spp.). Adults are found mostly in humid, forested areas. The larvae feed mainly on fungi, although some species spin webs to capture insect prey and a few species are parasitoids of flatworms.
- **Drosophilidae**: vinegar flies, pomace flies (2900 spp.). This family is best known for *Drosophila melanogaster*, the ubiquitous model organism for genetic research. Although adults are commonly found around overripe fruit or on mushrooms, the larvae of this family have a variety of lifestyles, from saprophagy to parasitism and predation.
- **Psychodidae**: moth flies (2700 spp.). The scavenging larvae of these flies are usually found in moist conditions, in soil, rotting wood, or other decaying vegetation. Adults of some subfamilies are blood feeders and transmit the disease leishmaniasis to humans.
- **Sarcophagidae**: flesh flies (2500 spp.). Sarcophagid larvae are saprophagous, parasitoids, predators or commensals in the nests of solitary Hymenoptera. The adults of some species are associated with filth.
- **Sphaeroceridae**: lesser dung flies (2300 spp.). Larvae and adults of this family are commonly found on dung, carrion, and decaying organic material.
- **Stratiomyidae**: soldier flies (2300 spp.). The adults of this family are often brightly colored and conspicuous...
FLIES, GNATS, AND MOSQUITOES

on flowers. The larvae are often found near water, in decaying organic material, or under bark. They are saprophagous, herbivorous, or predatory.

- Chloropidae: frit flies (2000 spp.). Larvae of chloropids have nearly every conceivable way of life, from scavenging to predation, parasitism and herbivory, although there are many more plant feeders in this family than the similarly diverse Phoridae.

A. Nematocerous Families

The nematocerous families are the relatively primitive members of the Diptera, characterized by long, unconsolidated antennae consisting of many segments. There are 26 families in this group and approximately 40,000 species included. Many are associated with aquatic habitats and some, such as mosquitoes, black flies, and biting midges, are voracious blood feeders.

B. Brachycera

Most Diptera belong to the Brachycera, a group characterized by the reduction or fusion of antennal segments to eight or fewer and by modifications to the larval head and mouthparts. With about 80,000 described species, this group contains many of the best known flies, such as houseflies and fruit flies.

The lower Brachycera includes several lineages constituting approximately 30,000 species. The larvae of most species are predatory, although there are a few parasitoid groups as well. The most familiar are the large families Asilidae, Bombyliidae, Tabanidae, Dolichopodidae, and Empididae.

The Cyclorrhapha, with about 50,000 species, includes a few primitive lineages and the Schizophora. The primitive groups are relatively small, with the exceptions of the large families Phoridae and Syrphidae. Within the Schizophora, there are a plethora of smaller acalypterate families that are rare, but many, such as the Tephritidae, Agromyzidae, Drosophilidae, Sphaeroceridae, and Empididae. The Cyclorrhapha, with about 50,000 species, includes several lineages constituting approximately 30,000 species. The larvae of most species are predatory, although there are a few parasitoid groups as well. The most familiar are the large families Asilidae, Bombyliidae, Tabanidae, Dolichopodidae, and Empididae.

1. Scavengers

This is the lifestyle most commonly associated with Diptera larvae, especially those of the Brachycera. More than half of the 128 families recognized herein have larvae that feed on decaying organic material or organic detritus. Most receive their nutrition from bacteria and other microorganisms of decay, not from the main substance on which they are found. They concentrate these organisms and other suspended particles with a sievelike pharyngeal filter.

Among the most obvious terrestrial scavengers are the larvae of blow flies (Calliphoridae) that are found on newly dead animal carcasses. Early stages of decay are characterized by large numbers of calliphorid and muscid larvae, followed later by drosophilids, latriniids, phorids, piophilids, sepsids, sphaerocerids, and others as decay proceeds. The fauna of buried carrion is different, with the calliphorids largely excluded. Instead, the muscid genus Muscina, various phorids, and sphaeroce-

3. Herbivores

Diptera that feed on living plants (including algae and fungi) are found in 37 of the 128 families recognized herein. Some, such as Agromyzidae, Anthomyiidae, Bibionidae, Cecidomyiidae, Chloropidae, Phoridae, Psilidae, and Sepsidae, have larvae that feed on a variety of plant materials, including roots, stems, leaves, flowers, and fruit. In aquatic environments, many larval Diptera feed on small organic particles in the water or on the substrate. Larvae of Culicidae and Simuliidae filter particles from the water with their brushlike labral fans.
Flies, gnats, and mosquitoes

dae, Sciuridae, Tephritidae, and Tipulidae, include species considered to be pests to human agriculture. Fly larvae attack all parts of plants, including fruits, flowers, stems, leaves (as leaf-miners), and roots.

3. Predators
Predators are organisms that kill more than a single host organism for their feeding. There are 35 families with this way of life, including nearly all of the non-cyclorrhaphan Brachycera. Most have extremely active larvae that attack other invertebrates as their major food source. The larvae of tabanids have been known to kill frogs, an interesting reversal of the usual chain of events. Many predatory dipteran larvae are beneficial to humans in controlling insects considered to be pests. Examples include syrphids attacking aphids and predatory muscid larvae that kill larvae of other muscids, such as houseflies.

4. Parasitoids
Parasitoids develop on and kill a single host. Twenty-two families of Diptera have this way of life, which was reviewed recently by Feener and Brown. All species of the second largest family of Diptera, the Tachinidae, are parasitoids. All dipteran parasitoids attack other invertebrates, usually other insects, but unusual hosts include terrestrial flatworms, mollusks, earthworms, millipedes, spiders, and scorpions.

5. Parasites
Parasites feed on a single host, but do not normally kill it. True parasites include the bot flies, whose larvae live under the skin or in the nasal cavities of various mammals, including humans. Some bot fly larvae live in the stomachs and alimentary tracts of horses, elephants, zebras, and rhinos. Other parasites of vertebrates are calliphorids, chloropids, piophilids, and muscids. True parasites of invertebrates are less well known but occur in a few families, such as Phoridae.

6. No Free-Living Larvae
The larvae of the four families of the Hippoboscoidea—Glossinidae, Hippoboscidae, Nycteribiidae and Streblidae—are retained in the female abdomen and nourished by secretions of the accessory glands. They are deposited by females as fully mature, third-instar larvae, which quickly pupariate.

B. Adults
Most adult Diptera receive the majority of their nutrition as larvae and do not feed extensively. Many need carbohydrates to power their flight, however, and feed on the nectar in flowers, or on honeydew, the sweet secretions of Homoptera. Some flies require nitrogen for nourishing their eggs and are thus avid flower visitors in search of pollen. Flower foraging for nectar and pollen makes Diptera adults important pollinators of plants.

At least some parasitoids feed on the hemolymph of their hosts after oviposition, and many dipteran species feed on dead insects, carrion, dung, or rotting vegetation. A few adult Diptera, including Deuterophlebiidae and some Oestridae, have vestigial mouthparts and do not feed.

Adults of a number of families have species that are well-known blood feeders, especially most or all species of Ceratopogonidae, Culicidae, Glossinidae, Hippoboscidae, Nycteribiidae, Simuliidae, Streblidae, and Tabanidae. Other families with fewer blood-feeding species are Muscidae, Rhagionidae, and Athericidae.

Some families of Diptera have species that are predatory as adults. This lifestyle is especially well developed in the Asilidae, Dolichopodidae, and Empididae but also occurs in some other families, such as Muscidae and Phoridae.

C. Special Associations

1. Aquatic Diptera
The larvae of many nematocerous families, as well as some Brachycera, are found in freshwater habitats. Among the most consistently aquatic forms are those families in the Culicomorpha and Blephariceromorpha, including such well-known families as Culicidae (mosquitoes), Simuliidae (black flies), and Chironomidae (midges). Chironomidae in particular can be exceedingly abundant, and various species assemblages are often used for assessment of water quality. The larvae of Simuliidae and Blephariceromorpha are found almost exclusively in clean, running water, with the Deuterophlebiidae and Blephariceridae being especially adapted to fast-flowing streams. Larval Culicidae are often found in ephemeral ponds that appear after snowmelt in temperate regions and after heavy rains elsewhere.

2. Phytotelmata
Phytotelmata are structures of plants that allow accumulations of water. They occur in various parts of plants, including leaves, leaf axils (especially of bromeliads and bananas), stems (especially of bamboos), fruits, and specialized structures (such as pitchers in pitcher plants). Among the approximately 20 families of Diptera that utilize phytotelmata, the Culicidae are the most prominent, being the most regular and most
Social insects are the ants, bees, wasps, and termites that have organized societies. They have one or a few females responsible for all the egg laying, while other members of the colony (usually sterile females) gather food and do other tasks. Diptera are associated with social insects as scavengers, predators, parasites, and parasitoids. In the following treatment, only those species intimately associated with social insects are discussed, whereas opportunistic, generalized predators, such as robber flies (Asilidae), will be ignored.

Among the Brachycera associated with social insects, the most diverse are the Phoridae. These small flies are commonly found in association with ants, especially army ants (both New World and Old World), termites (especially in Africa and Southeast Asia), and, to a lesser extent, social bees and wasps. Many are parasitoids, laying their eggs inside the bodies of their adult hosts, usually ants, termites, and stingless bees; some are also known to parasitize the immature stages of ants. Many are scavengers, living in the refuse piles of the large colonies of army ants and leaf-cutter ants, where they are joined by scavenging larvae of other families. Some phorid larvae are predatory on ant brood, as are the larvae of some syrphids. The females of phorid species living in social insect nests are often remarkably modified, with reduced wings, eyes, and body sclerotization. One remarkable genus from Southeast Asia has adult females that mimic the larvae of their army ant hosts. Other females are heavily armored and have a rounded, teardrop shaped body form. This limuloid body form allows the females to escape damage when accosted by aggressive host ants, bees, or termites.

Other dipteran larvae found in social insect nests belong to the families Brauliiidae, Calliphoridae, Fanniidae, Sarcophagidae, Sphaeroceridae, and Syrphidae. Most are scavengers, although some are predatory.

Army ant raids are spectacular tropical phenomena that provide a number of opportunities for dipteran associates. Tachinids and conopids hover or perch near the raid front, darting down to parasitize the crickets, cockroaches, and other insects flushed by the foraging ants. Closer to the ground, parasitic phorids dart at their respective hosts: either the army ants themselves or other ants that are victims of the ant raids. Sarcophagids swarm on the leaves of nearby undergrowth, feeding on the droppings of birds attracted to the insects flushed out by the ant raids. Often, the raid front is best identified by listening for the loudest buzzing from the activity of hundreds of flies.

4. Kleptoparasites

Many Diptera exploit other insects or invertibrates that sequester or reserve food for long periods of time. By waiting before feeding, or by feeding slowly, these hosts provide a window of opportunity for larvae or adults of flies. The relationship between the host and kleptoparasite is often developed to such an extent that the flies live permanently associated with their food provider. Phoresy (transportation of the kleptoparasite by the host) often occurs in these associations.

A commonly observed example of kleptoparasitism and phoresy in Diptera is the association of large spiders and flies. Spider webs provide a continuous supply of food, and often a source of stored insect carcasses for later feeding. This warehouse is exploited by a number of fly families, including Ceratopogonidae, Cecidomyiidae, Micropezidae, Chloropidae, Lonchaeidae, Phoridae, and Empididae. In many instances, the flies perch on the bodies of the spider, waiting for their next meal to arrive.

Another, much more restricted association occurs between dung beetles of the family Scarabaeidae and flies of the family Sphaeroceridae. The flies ride on the bodies of the dung beetles, waiting until the beetles find and bury feces for their own larvae to feed upon. The flies briefly hop off the beetles, lay their own eggs on the dung, and then rejoin their food-providing host.
Kleptoparasites also occur in the nests of social Hymenoptera, where fly larvae are fed by deceived workers or attack the provisions left for a developing bee, wasp, or ant larva. A full review of kleptoparasitism is given by Sivinski et al.

5. Swarming

Swarms of adult Diptera are a common sight in most environments. Usually these are aggregations of males, allowing females to easily find a mate. Often, swarms are seen in the canopy of forests, under overhanging branches, in sunlight “pools” in forests, near fire towers emerging above tree level, or at the summit of tall hills or mountains (“hill-topping”).

Most nematocerous families engage in swarming, as do about 15 families of Brachycera. Flies that engage in this behavior often have associated structural modifications, including a well-developed anal lobe of the wing and enlarged compound eyes. The development of the anal lobe probably allows for better maneuvering and hovering within the swarm, whereas the increased size of the eye allows male flies to more precisely place themselves, and assess their place, within the swarm. As females choose dominant males within the swarm, based on their relative position, assessment of position is critical for male mating success.

Species of dance flies (subfamily Empidinae) have elaborated on this basic pattern. Males of many species catch prey (usually smaller flies and other soft-bodied insects) and carry them in the swarm. They offer the prey to the nonhunting females as “nuptial gifts” to be fed upon during mating. A few species have reversed this behavior often have associated structural modifications, including a well-developed anal lobe of the wing and enlarged compound eyes. The development of the anal lobe probably allows for better maneuvering and hovering within the swarm, whereas the increased size of the eye allows male flies to more precisely place themselves, and assess their place, within the swarm. As females choose dominant males within the swarm, based on their relative position, assessment of position is critical for male mating success.

Species of dance flies (subfamily Empidinae) have elaborated on this basic pattern. Males of many species catch prey (usually smaller flies and other soft-bodied insects) and carry them in the swarm. They offer the prey to the nonhunting females as “nuptial gifts” to be fed upon during mating. A few species have reversed this behavior.

IV. CONSERVATION BIOLOGY OF DIPTERA

Because of the perception of almost all Diptera as disease-carrying filth flies, there is little public sympathy or interest in their conservation. Also, most Diptera are extremely poorly known, and the study of many families is still contingent on the recognition and description of the many undescribed species.

An exception to these statements is the Syrphidae, a group that is popular among amateur collectors, especially in Europe. There, Red Lists of endangered species exist for many regions, including Britain and parts of the mainland. When species that are at the edge of their distribution are eliminated from consideration, it appears that most endangered syrphids are either saproxylic or associated with wetlands.

Saproxylic insects of all orders are considered to be among the most endangered insects in Europe. These are species that are obligately associated with rotting wood, a habitat that is largely absent from young, even-aged tree plantations or from older forests that are sanitized or managed by removal of dead trees and wood. Most of the saproxylic habitats are afforded by old-growth forests with adequate numbers of injured, bleeding, mortibund, or dead trees available. Different species of insects are associated with different types of decay, including whether or not the trees are still standing. Other variables include whether the trees are large or small, whether they are exposed to light and low humidity or are shaded throughout the day, and how long the process of decay has proceeded. Rot holes (some of which contain water), sap runs, dead branches on otherwise healthy trees, and loosened bark are all microhabitats that have specialized insect (including Diptera) faunas. The saproxylic faunas of other regions have not been assessed, but as deforestation proceeds throughout the world, Diptera associated with this habitat will likely be threatened.

Wetlands have been greatly reduced in many areas of the world, in many instances for the express purpose of eliminating biting fly vectors of disease. The loss of such habitats, however, also eliminates populations of other aquatic Diptera that have important ecological roles in the environment.

Another group of Diptera whose conservation needs are relatively well known is the endangered fauna of Hawaii. There are several species of Drosophila that have been proposed for listing as endangered species. Because of their poor public image, Diptera are usually not considered flagship species for conservation projects. In the southwestern United States, however, an endangered species of myiid fly has been used to spearhead efforts to halt the final destruction of an endangered habitat in the Los Angeles area. The Delhi Sands Giant Flower-Loving Fly (Rhaphiomidas terminata abdominalis) is the largest and most identifiable of a number of threatened taxa that live in this habitat,
which is reduced to just a few acres in extent. The other subspecies of this taxon, *R. t. terminatus*, previously went extinct when its habitat in the coastal dunes near the Los Angeles International Airport was almost completely destroyed by urbanization. Other species of *Rhaphiomidas* are also considered endangered, as are other Diptera living in small, isolated, sandy habitats. In general, the conservation status of Diptera on a worldwide scale is unknown. Undoubtedly many species are lost to deforestation, but only those in a few well-known groups (like syrphids) or in habitats that are of interest to some people (wetlands, sand dunes) have been studied. The situation is unlikely to change until we know much more about the systematics of flies.

See Also the Following Articles

BEETLES • BUTTERFLIES • HYMENOPTERA • INSECTS, OVERVIEW • MOTHS

Bibliography


I. Introduction
II. Types of Food Webs
III. Omnivory and the Structure of Food Webs
IV. Patterns of Biomass and Energy in Food Webs
V. Current Topics/Trends in Food Web Studies

GLOSSARY

community The most practical definition is a set of species that interact at a given location.
connectivity web This type of food web illustrates only feeding links without reference to strength of interaction or energy flow.
detrital shunts Energy and nutrients from the saprovores web reenter the plant herbivore predator food web when detritivores are eaten by predators that also eat plants, herbivores, or other predators.
donor control Consumer population growth is affected by their resources but consumers do not affect the renewal rate of these resources and hence cannot depress their resources.
ecosystem A set of one or more communities and their abiotic environment.
energetic web This type of food web quantifies the amount of energy (or material) that flows across links joining species.
food or biomass pyramid A graphic representation of the energy or biomass relationships of a community, in which the total amount of biomass, or total amount of energy available, at each successive trophic level is proportional to the width of the pyramid at the appropriate height.
food chain A representation of the links between consumers and their resources, for example nutrients $\rightarrow$ plant $\rightarrow$ herbivore $\rightarrow$ carnivore. In these representations, energy or material flows up the chain in a linear fashion. In addition, a food chain can be a linear set of species within a food web.
food web A representation of feeding relationships in a community that includes all the links revealed by dietary analysis.
functional or interaction web This type of food web quantifies the strength of interaction between species linked using data from manipulative experiments.
recipient control Consumers substantially depress populations of their resources.
spatial subsidies Input from other habitats of organic carbon, nutrients, and prey or the movement of consumers. These resources can influence greatly the energy, carbon, and nutrient budget of recipient habitats. In general, nutrient inputs (nitrogen, phosphorus, and trace elements) increase primary productivity; detrital and prey inputs produce numerical responses in their consumers.
trophic level An abstract classification to describe subsets of species that acquire energetic resources in a similar way on a subset of species (e.g., top carnivores feed on primary carnivores which feed on herbivores which feed on primary producers). In natural systems, most species do not feed strictly on the
FOOD WEBS

KNOWLEDGE OF FOOD WEB structure and dynamics is central to our understanding of almost all aspects of population and community ecology. By their very nature of representing feeding relationships between species, food webs have the capacity to embody the rich complexity of natural systems. In fact, most important interactions (e.g., competition, predation, and mutualism) cannot be isolated from a food web context.

I. INTRODUCTION

Food webs occupy a central position in community ecology. Charles Darwin introduced the concept of an entangled bank in which he envisioned many kinds of species interdependent on each other in a complex manner governed by “laws acting around us.” In the simplest context, food webs incorporate the two factors that, a priori, one would consider most fundamental to the success of any one species: resources and enemies. All species must acquire resources (food or nutrients) and suffer energy losses or mortality from predators (Fig. 1). The abundance and success of any species is thus a product of these feeding interactions. This inclusion of such “bottom-up” (productivity and resources) with “top-down” (consumption) factors largely determines the distribution and abundance of almost every species on the planet. In particular, freshwater ecologists have enjoyed notable success by concurrently studying the interaction between these variable factors on the regulation of plant and animal abundance and thus the structure of freshwater communities. This research shows the rich dynamical outcomes that can occur when predation and productivity vary and interact within a food web (Fig. 2).

Many important advances have arisen from analyses that concurrently incorporate more than one interaction in a food web: keystone predation and herbivory, the intermediate predation and disturbance hypotheses, the size-efficiency hypothesis, trophic cascades, intraguild predation, apparent competition, and the recognition of the importance of indirect effects. The outcome of virtually all interactions within a community can be modified, directly and indirectly, by other members of the food web. This insight penetrates to all areas of community ecology. For example, the results of experiments must be interpreted carefully for at least two reasons. First, indirect effects, moderated by other species in the web, may exert large and sometimes contradictory effects to the direct effects of the manipulation. Thus, under some food web configurations, removal of a predator may directly increase the level of its prey or may actually cause the prey to decrease because of indirect interactions. Second, changes in species dynamics putatively caused by one factor may actually be a product of a second process.

II. TYPES OF FOOD WEBS

Food web research has grown at a tremendous rate and taken a diversity of forms. Not surprisingly, ecologists have diverged in their methods, emphases, and approaches. Nevertheless, trophic relationships in communities can be delineated in three basic ways. Paine (1980) and Polis (1991) distinguished three types of food webs that evolved from ecological studies (Fig. 3).

The first is the classic food web, a schematic description of connectivity specifying feeding links. Such connectivity webs simply demonstrate feeding relationships. Examples of these are the early food webs of Forbes and Summerhayes and Elton (Fig. 3). The second web type is also descriptive, quantifying the flow of energy (and/or materials) between trophically connected species. Examples of this type of food web include intertidal communities in Torch Bay, Alaska, and Cape Flattery, Washington (Paine, 1980) and Polis (1991) distinguished three types of food webs that evolved from ecological studies (Fig. 3).

The third type use experiments to dissect communities to identify strong links and dynamically important species. Such interaction or functional webs demonstrate the most important connections in an ecosystem (Fig. 3). These food webs depict the importance of species in maintaining the integrity and stability of a community as reflected in its influence on the growth rates of other species. They require experimental manipulations of the community (e.g., by removal or addition of particular species). In the following sections, we discuss the strengths and weaknesses of each approach.
Of the three, only the last two have contributed substantially to our understanding of natural systems.

### A. Connectivity Webs

Connectivity webs are representations of "who eats whom" without inference to the strength or type of interaction and energy flow (Fig. 3). Early food webs were constructed for essentially two reasons: (i) to depict the interconnectivity of natural systems and (ii) to examine issues of "the balance of nature," i.e., to analyze how harmony is maintained through complex predatory and competitive interactions within communities (Forbes, 1887). Such an approach was applied to agricultural systems to examine pests and possible food web manipulations to control pests. As early as the 1880s, beetles were introduced into the United States to control agricultural pests. Such control then benefited crop plants via an indirect interaction (predator prey crop) (following the success of Vedalia, a coccinellid beetle, in controlling cottony-cushion scale..."
in California in 1888, about 50 more coccinellids were introduced in the 1890s.

The knowledge required to construct connectivity webs is straightforward: An approximate, qualitative knowledge of who eats whom is all that is necessary to produce a simple food web, whereas experimental manipulations or quantitative measurements are necessary to construct webs of interaction or energy flow. Consequently, connectivity webs most frequently represent trophic interactions in communities and have received the most attention. Hundreds of such webs slowly accumulated over a century. They were useful to illustrate, in a totally nonquantitative manner, the feeding interactions within a specific community. Different scientists constructed webs of different diversity, complexity, and resolution, depending on their knowledge of the system and bias or understanding of particular groups. For example, some may emphasize birds and lump all insects as one group. Others will divide the insects into scores of groups and represent one or two bird species.

In the 1970s and 1980s, many theoretical and statistical studies were performed on connectivity webs cataloged from the literature to determine similarities and natural patterns among them. Empirical generalizations were abstracted from data of published connectivity webs. These "natural patterns" largely agreed with predictions made by early food web models. These models showed that food webs were constrained to be quite simple: Each species ate few species and had few predators; the total length of the number of links in a typical food chain was short, usually two or three; omnivory was very rare; and there were a few other patterns. Early modelers argued that the congruence of patterns from the cataloged webs validated the predictions of their models. They thus claimed that their Lotka–Volterra models were heuristic and represented processes that structure real communities. For example, the addition of omnivory to model food webs causes webs to be unstable dynamically and exhibit relative low persistence (time before species are lost). Thus, these models make the prediction that omnivory should be relatively rare in those webs that persist in nature. Comparison of omnivory in cataloged webs relative to its frequency based on chance shows that omnivory is statistically rare in real webs, as predicted by models. The same
FIGURE 4: Food web showing aggregation within some trophic levels but not others. (A) The dynamics of omnivory; (B) spatial subsidy; (C) detrital shunts.

general approach was used to validate other predictions of model webs, e.g., short chain lengths. Thus, modelers soon "explained" these empirically derived patterns. Although these studies, and the connectivity approach, make good food web diagrams, they are flawed to such a great a degree that today such analyses are viewed as providing little understanding of natural communities. There are many reasons why this is so, of which only a few are mentioned here:

1. Most vastly under-represent the species diversity in natural communities. Most communities have hundreds to thousands of species, but these webs would represent <10–30 species on the average. As a consequence, most connectivity webs have severe problems with "lumping" species and taxonomic biases. Some trophic levels are distinguished by species (e.g., birds or fish), whereas other groups suffer a high degree of aggregation, e.g., all species of insect or annual plants are represented as one super-species—"insects" or "plants" (Fig. 4).

2. Most species are highly omnivorous, feeding on many resources and prey that each have a distinct trophic history and are often at different trophic levels. Because diet is very difficult to delineate, most connectivity webs greatly underrepresent the true nature of omnivory. This poses several fundamental problems.

3. Connectivity webs typically only offer a static view of the world and webs are usually idealized representations that show all linkages that occur over large spatial and temporal scales. Therefore, much of the important variability and changes due to local environmental conditions are lost. However, studies that compare changes in connectivity over time and space and across environmental gradients (such as those by Mary Power and her group on the Eel River) can provide important insight into community structure and dynamics. One can view connectivity webs as a first step in examining the interactions in communities (i.e., performing "natural history" studies), to be followed by quantification of the fluxes of energy and nutrients (as in energetic webs).

B. Energetic Webs

Starting with the classic studies of Elton, Summerhayes, and Lindeman, food web studies turned toward quantifying flows of energy and nutrients in ecosystems and the biological processes that regulate these flows. This approach is an alternative to connectivity webs to describe trophic connectedness within communities. This "process-functional" approach explicitly incorporates producers, consumers, detritus, abiotic factors, flow out of a system, and the biogeochemical recycling of nutrients. It views food webs as dynamic systems in time and space. Such an approach necessitated analyzing energy and material fluxes in order to understand the behavior of ecosystems. Thus, a typical analysis would quantify the amount of energy or matter as it travels along different pathways (e.g., plants → consumers → detritus → decomposers → soil). For example, the tracking of energy and DDT through a food web in a Long Island estuary enabled researchers to study bioaccumulation effects on top predators.
The use of energetic webs has provided a rich understanding of the natural world and allowed us to understand much about ecosystems. Several important processes are included in energetic webs. First, they quantify energy and material pathways and key species or processes that facilitate or impede such flows. Second, they include an explicit recognition of the great importance of detritus, a subject virtually ignored in connectivity webs. (10 to >90% of all primary productivity from different habitats immediately becomes ‘dead’ organic detritus rather than being eaten by herbivores). Third, this approach recognized that a great amount of energy, nutrients, and prey originated outside the focal habitat, which is a key insight to understand natural communities. Thus, energetic webs show how ecosystems function and which species dominate biomass and energy.

Beginning with Lindeman, researchers began to examine the efficiency of transfer from prey species to predator species. It was found that energy transfer is generally inefficient with only about 3–15% of the energy of prey species being converted to energy of predators. Peter Yodzis used this information to suggest that the length of food chains within a community would be set by the amount of energy entering into the base of the chain. This argument was in opposition to Pimm and Lawton’s suggestion that food chain length is set by the resilience of the chain. By resilience, Pimm and Lawton using Lotka–Volterra models, meant the estimated time for model food chains to recover from some disturbance. They argued that frequent disturbances (relative to growth rates of species) would result in shorter food chain lengths. Furthermore, early studies examining the influence of primary productivity (thus, the amount of energy entering a food chain) did not support the hypothesis that food chain length was governed by energy transfer efficiency. However, recent reexaminations of Pimm and Lawton’s work suggest that two factors influenced their results—density-dependent regulation of the basal trophic level and food chain structure (the lack of omnivory in their models). Moreover, recent studies of the role of energy efficiency have found that decreases in productivity result in shorter maximum food chains. Thus, the relative role of resilience versus energy transfer in regulating the length of food chains is still debated.

One outcome of the argument for the role of energy transfer as the main governing factor of food chain length is a body of work that examines differences in energy efficiency among organisms. For example, carnivores are found to have greater efficiency than herbivores. Additionally, invertebrate ectotherms have greater efficiencies than vertebrate ectotherms, which in turn are more efficient than endotherms. Yodzis and Inness used this information (and relative body sizes) to parameterize nonlinear predator–prey models.

In summary, the analysis of energy and matter flow is necessary and central to understanding the dynamics of populations and communities. The success of a population is always strongly related to the energy and biomass available to it. Consequently, it is difficult or impossible to understand the dynamics and structure of food webs and interacting populations without incorporating energy flow from below. However, this energetic approach per se, although necessary, is not sufficient by itself to understand the dynamics of communities because energy flow and biomass production are functions of interactions among populations within the food web. The transfer of energy and matter becomes complicated as they pass through the many consumers that populate community food webs. For example, increasing the amount of nutrients to plants may increase the biomass of each consumer in the web or may just increase the biomass of a subset of consumers (e.g., only the plants, plants and herbivores, or only the herbivores), depending on the relationship between consumers and their resources. Because of these considerations, pathways must be placed in the context of ‘functional’ food webs to understand the dynamics of energy and material transfer.

C. Functional or Interaction Webs

Functional or interaction webs use experiments to determine the dynamics within a community. Starting with Connell and Paine, empiricists began to use experiments to examine communities and food webs to discover which species or interaction most influenced population and community dynamics. They manipulated species that natural history or energetic analyses suggested were important. They used either ‘press’ (continual) or ‘pulse’ (singular) experiments to manipulate populations of single species and then followed the response of other species within the food web. The philosophy of these studies was to simplify the complexity of natural systems with the assumption that many species and links between species were unimportant to dynamics. Paine tested this assumption and found that indeed many links between species were weak (essentially zero).

Experimental analyses of food webs are designed to identify species and feeding links that most influence population and community dynamics. These alone are placed into an ‘interaction web’ that, in theory, encom-
passes all the elements that most influence the distribution and abundance of member species. However, unlike connectivity webs, key species are identified through experiments rather than diet frequency or energy transfer. The initial process of choosing certain species and interactions for experiments and excluding others is subjective, optimally based on strong intuition and a rich understanding of natural history. As the researcher learns more, some elements are discarded and others are subject to further experimentation. Eventually, the community is distilled into an interaction web, a subset including only species that dominate biomass and/or regulate the flow of energy and matter.

This approach has been used by experimental and theoretical ecologists to produce a rich understanding of the processes that most influence their communities. They have been remarkably fruitful and have introduced many food web paradigms that go to the center of ecology, e.g. keystones species, the intermediate disturbance hypothesis, top-down and bottom-up control, trophic cascades, and apparent competition.

However, this approach is not without limitations. Three major problems stand out. First, many statistical shortcomings can beset experimental manipulation of food webs. For example, replications are commonly difficult (time-consuming and expensive) and therefore experiments often lack the statistical power necessary to avoid type II statistical errors (significant biological differences exist among treatments but low sample size precludes their detection statistically). Second, the number of possible experiments is almost infinite. Which ones should be conducted, and which species should be manipulated?

The third and perhaps most troublesome problem is that experiments isolate a subset of species and links from the community food web, largely ignoring how manipulations interact with the remainder of the community. Thus, unobserved indirect or higher order interactions may exert important effects on the dynamics of experimental species and, in theory, may change the outcome of experiments indeterminate. For example, predators are thought typically to suppress their prey. However, if a predator is omnivorous, not only eating the prey but also consuming a more efficient predator on the same prey (i.e., it is an “intraguild predator”), it may actually relax the predation load on their shared prey, thus increasing the shared prey’s abundance. For example, guilds of biological control agents must be carefully structured because some species eat not only the host but also other predators/parasitoids and thus their presence decreases the number of control agents and increases target pest populations. Many other cases exist in which consumers, via such intraguild predation, may indirectly facilitate its prey while concurrently exploiting it via direct consumption. Another example of indirect effects mediated by other than studied “focal species” is shown by the interaction between Australian bell miners and their homopteran food (“lerp”). After these birds were removed experimentally, the insects first increased greatly in number and then vanished when other bird species invaded the now undefended miner territories. Thus, the apparent effect of leaf miners on lerp insects (here, suppression or facilitation) depends on when the insects were surveyed. Such complications have undoubtedly interfered with clear interpretations of many experiments. The caveat is clear: Experiments can be indeterminate, producing contradictory, counterintuitive, or no results, depending on the relative strengths of the direct and indirect effects.

These problems can be anticipated and partially negated with the application of good intuition of the natural history of the system and important mechanisms. Such intuition is a product of intimate empirical knowledge gained through observation and guided by a conceptual awareness of which interactions are potentially important. Initially, this process is essential to design the appropriate experiments and identify which species and trophic links may be dynamically important. At the end, experiments must be interpreted in a food web context to assess possible indirect and higher order effects. Experimental results must be complemented with good descriptive, mechanistic, and comparative data to produce a deep understanding of the system. This is one role for energetic and dietary data. Experiments in the absence of natural history often do not succeed and may mislead. The important messages from this section are that the complex food webs of natural communities can be simplified and understood by isolating key species and links into “interaction webs,” experiments are absolutely necessary for this process, and experiments must be designed and interpreted with sound intuition based on natural history and theory.

III. OMNIVORY AND THE STRUCTURE OF FOOD WEBS

It is necessary to discuss feeding connections in more detail. Empirical research and logic have shown that the vast majority of consumers on this planet are very
omnivorous, feeding on many types of food throughout the entire food web. This is not to say that all species are so catholic in their diets. Specialists abound, e.g., many herbivores or parasites consume only specific plants or hosts. However, these form a minority of consumers. The ubiquity of omnivory carries many implications for our efforts to produce theory and models to understand how food webs operate in and shape natural systems.

Omnivory occurs ubiquitously when consumers eat prey from general classes of prey, such as arthropods, plankton, soil fauna, benthos, or fish. The existence of multiple trophic types within these classes causes consumers to feed on species from many trophic levels. For example, “arthropodivores” eat whatever properly sized arthropods are available (e.g., predaceous spiders and insects and insect parasitoids, herbivores, and detritivores) without pausing to discriminate among their prey according to trophic status. For example, in the Coachella Valley desert delineated by Polis (1991) over 10 years of study, predaceous and parasitoid arthropods formed 41% of the diet of vertebrate and 31.3% of invertibrate arthropodivores, with the remainder of the diet being herbivorous and detritivorous prey. Similarly, inspection of diet data of planktivores, piscivores, “insectivores,” carnivores, or bentthic feeders reveals that such different channel omnivory is almost universal with the exception of those few taxa that specialize on a few species of prey.

Another important type of omnivory occurs when consumers eat whatever resources are available or abundant at a particular time or place, regardless of their trophic history. When analyzed, the diet of a single species usually shows great differences through time (e.g., seasonally) and space (patches or habitats). Prey exhibit three general phenologies: pulsed (population eruptions lasting a few days or weeks), seasonal (present for 2–4 months), and annual (available throughout the year). Feeding on prey from all three phenologies produces diet changes over time for almost all non-specialist consumers. Furthermore, many (most?) vertebrates opportunistically switch from plant to animal foods with season. For example, grainivorous birds, rodents, and ants primarily eat seeds but normally feed on the abundant “arthropods” (= insects from all trophic levels and spiders) that appear during spring. Alternatively, many omnivorous, arthropodivorous, and carnivorous species consume significant quantities of seed or fruit. In the Coachella Valley, 79% of 24 primary carnivores eat arthropods and/or plants; for example, coyotes eat mammals (herbivorous rabbits, rodents, and gophers; arthropodivorous antelope and ground squirrels; carnivorous kit foxes and other coyotes), birds (including eggs and nestlings, e.g., carnivorous roadrunners; herbivorous doves and quails), snakes, lizards, and young tortoises as well as scorpions, insects, and fruit. In New South Wales, 15 of 27 ant species are “unspecialized omnivores” eating nectar, seeds, plant parts, and a broad range of living and dead insects, worms, and crustacea. Overall, it appears that most consumers eat whatever is available and whatever they can catch.

“Life history” omnivory describes the great range of foods eaten during growth and ontogeny by most species (the “age structure component” of dietary niche breadth). Such omnivory includes abrupt diet changes in species undergoing metamorphosis (e.g., many marine invertebrates, amphibians, and holometabolic insects) and gradual diet changes in “slowly growing species” (e.g., reptiles, fish, arachnids, and hemimetabolic insects). Changes at metamorphosis can be great; for example, 22% of the insect families in the Coachella Valley desert community undergo radical change in diet—larvae are predators or parasitoids and adults are herbivores. Although not as dramatic, significant changes characterize slowly growing species so that differences in body size and resource use among age classes are often equivalent to or greater than differences among most biological species. Life history omnivory expands the diet of species throughout the entire animal kingdom with the exception of taxa that use the same food species throughout their lives (e.g., some herbivores) and those with exceptional parental investment (e.g., birds and mammals) so the young do not forage for themselves.

“Incidental omnivory” occurs when consumers eat foods in which other consumers live. Thus, scavengers and detritivores not only eat carrion or organic matter but also the trophically complex array of microbes and macroorganisms that live within these foods. Frugivores and granivores commonly eat insects associated with fruits and seeds. Predators eat not only their prey but also the array of parasites living within the prey. In each case, consumers automatically feed on at least two trophic levels.

These types of omnivory are widespread and common. Their ubiquity poses many questions. First, how does omnivory affect food web structure? Most obviously, it increases complexity and connectivity. Second, can we ignore omnivory in the analyses of food webs? By its very nature, omnivory causes consumers to have a great number of links, each of which may be numerically unimportant in the diet. For many reasons delineated later, we cannot arbitrarily ignore apparently minor diet links if we hope to understand dynamics.
IV. PATTERNS OF BIOMASS AND ENERGY IN FOOD WEBS

Primary productivity is among the most fundamental biological processes on the planet, transferring the energy locked in light and various inorganic molecules into forms useful to sustain producers and the diversity of consumers. What factors control primary productivity and regulate its distribution among plants, animals, and microbes? How do changes in primary productivity work their way through a food web to alter the abundance and biomass of herbivores to predators and detritivores? As discussed later, such key questions are best assessed using a food web approach. However, considerable controversy exists regarding the exact way that food web structure influences community and ecosystem dynamics.

A. Trophic Levels, Green Worlds, and Exploitative Ecosystems

Ecological research has amply demonstrated that food webs in nature contain hundreds to thousands of species, reticulately connected via multiple links of various strength to species in the autotroph and saprophagous channels and in the same and different habitats; omnivorous, age-structured consumers are common. Nevertheless, much food web theory still relies on the idealization of trophic levels connected in a single linear chain (plant herbivore carnivore). Here, we evaluate this simplification and some of its implications. In particular, we focus on two grand theories whereby food webs are considered to be central to community organization. The trophic level ideal in a simple linear food chain has had great appeal. Trophodynamics sought to explain the height of the trophic pyramid by reference to a progressive attenuation of energy passing up trophic levels, envisioned as distinct and functionally homogeneous sets of green plants, herbivores, primary carnivores, and, sometimes, secondary carnivores. This is a bottom-up community theory based on the thermodynamics of energy transfer. In counterpoint, Hairston, Smith, and Slobodkin’s green world hypothesis (GWH; Hairston et al., 1960) is primarily a top-down theory, with abundance at each level set, directly or indirectly, by consumers at the top of the chain. Thus, carnivores suppress herbivores, which releases green plants to flourish. These and earlier theoretical studies attempted to simplify food webs greatly to find generalities among them. GWH reduced complex webs to food chains in which species were pigeonholed into specific trophic levels. This allowed for predictions on how higher trophic levels (e.g., predators) influenced the dynamics of lower trophic levels (e.g., primary producers).

Oksanen et al.’s (1981) exploitation ecosystem hypothesis (EEH) generalizes GWH to fewer or more than three trophic levels. Trophic cascades are examples of food chains that behave approximately according to EEH. Trophodynamics and EEH each rely on the integrity of trophic levels and the existence of a single, albeit different, overwhelming mechanism that imposes structure on ecosystems. EEH proposes a conceptual framework of “exploitation ecosystems” in which strong consumption leads to alternation of high and low biomass between successive levels. Even numbers of “effective” trophic levels (two or four levels) produce a low-standing crop of plants because the herbivore population (level 2) flourishes. Odd numbers (one or three levels) result in the opposite effect: Herbivores are suppressed and plants do well. Proponents of EEH differ on subsidiary points, the first being the role of bottom-up effects in which primary productivity sets the number of effective levels. The most productive systems support secondary carnivores and therefore have four levels and low-standing crops of plants. Low-productivity systems (e.g., tundra) support only one effective level—plants. More productive habitats (e.g., forests) have three. Productivity is never high enough to support more than three effective levels on land or four in water. Other studies argue that physical differences between habitats, by affecting plant competition and consumer foraging, cause three levels on land and four in water. EEH definitions of trophic levels are distinctive and adopt the convention that trophic levels occur only if consumers significantly control the dynamics or biomass of their food species. Without top-down control, consumers do not comprise an effective trophic level regardless of biomass or number of species involved. Supporters of EEH have noted that only when grazers regulate plants are grazers counted (as a trophic level), and only when predators regulate grazers are they fully counted. Thus, considerations of food chain dynamics do not become stranded in the immense complexity of real food webs. On the other hand, GWH trophic levels are based on energy deriving from primary productivity. Thus, “trophic level interactions . . . weight particular links in the food web for their energetic significance.” A trophic level is “a group of organisms acquiring a considerable majority of its energy from the adjacent level nearer the abiotic source.” Despite these differences, both EEH and GWH theory argue that variability
in the number of trophic levels exerts profound consequences on community structure and dynamics. Considerable controversy exists as to the validity of GWH and EEH. The consensus has swung against these grand theories. Numerous arguments and empirical observations suggest that such processes operate occasionally in water but never on land. Basically, the complexity observed in natural systems does not conform to the reality of simple trophic levels. It appears that the notion that species clearly aggregate into discrete, homogeneous trophic levels is a fiction, arising from the need of the human mind to categorize. Especially in speciose systems, groups of species with diets of similar species do not occur. Omnivory, ontogenetic and environmentally induced diet shifts, and geographical and temporal diet heterogeneity all obscure discrete trophic levels. Even plants do not easily form a single level; higher plants have diverse crucial trophic and symbiotic connections with heterotrophs and many phytoplankton are mixotrophic, obtaining energy via photosynthesis, absorption of organic molecules, and ingestion of particles and bacteria. With increasing diversity and reticulation in webs, trophic levels blur into a trophic spectrum rather than a level. These species-individualistic and continuous “trophic spectra” are a reasonable alternative to the simplistic construct of homogeneous trophic levels.

B. Complex Food Webs, Multichannel Omnivory, and Community Structure

Polis and Strong (1996) offered a framework in the context of functioning community webs as an alternative to theories based on discrete trophic levels. Substantial evidence indicates that most webs are reticulate and species are highly interconnected, most consumers are omnivorous on foods (frequently on both plants and animals) across the trophic spectrum during their life history, most resources are eaten by many species across the trophic spectrum, plants are linked to a variety of species via trophic murturalism, most primary productivity becomes detritus directly, detrital biomass re-enters the autotroph channel of the web when detritivores and/or their predators are eaten by consumers that also eat species in the herbivore channel, and species are often subsidized by food from other habitats. They proposed that such trophic complexity pervades and generally underlies web dynamics. High connectance diffuses the direct effects of consumption and productivity throughout the trophic spectrum. Thus, consumer and resource dynamics affect and are affected by species at multiple positions along the trophic spectrum rather than interacting only with particular trophic levels. Consumer density is elevated and they often persist by eating resources whose abundance they do not influence (i.e., the interaction is “donor controlled”). Such dynamics are illustrated by focusing on top-down interactions. Some consumers exert “recipient” control on some resources and, occasionally, produce trophic cascades. Polis and Strong (1996) suggest that such control is often enabled by omnivorous feeding and various consumer subsidies that are usually donor controlled. Here, the transfer of energy and nutrition affects dynamics; numerical increases in consumer abundance occur from eating diverse resources across the trophic spectrum in the autotroph channel, from detritivores and detritus from the saprovore channel, from other habitats, and across their life history. Consumers, so augmented, exert recipient control to depress particular resources below levels set by the nutrition traveling through any particular consumer-resource link (analogous to the effects of apparent competition). Top-down effects arising from such donor-controlled, “multichannel” omnivory are depicted in Figs. 2 and 4. Strong consumer-mediated dynamics occur precisely because webs are reticulate and groups of species do not form homogeneous, discrete entities. Multichannel omnivory has two essential effects on the dynamics of consumers, resources, food webs, and communities. First, it diffuses the effects of consumption and productivity across the trophic spectrum rather than focusing them at particular trophic levels: It increases web connectance, shunts the flow of energy away from adjacent trophic compartments, alters predator–prey dynamics in ways contra to EEH assumptions, and thus disrupts or dampens the ecosystem control envisioned by EEH. For example, Lodge showed that omnivorous crayfish can depress both herbivorous snails (consistent with GWH and EEH) and macrophytes (inconsistent).

Second, omnivory can affect dynamics in a way analogous to apparent competition. Feeding on “nonnormal” prey can increase the size of consumer populations (or sustain them during poor periods), thus promoting top-down control and depression of “normal” prey. Frugivory, herbivory, granivory, detritivory, and even coprophagy form common subsidies for many predators. Vertebrate carnivores consume amply from the lower web without markedly depleting these resources. Does energy from fruit help carnivores depress vertebrate prey (e.g., herbivores)? Arthropodivory by seed-eating birds is the norm during breeding, with insect
C. Trophic Cascades or Trike

One prediction of GWH and EEH is that communities are structured by trophic cascades. Trophic experiments to test cascades use two methods: a bottom-up approach by increasing a resource (e.g., nitrogen or phosphorus) or a top-down approach that adds a top predator to a system. In the former, trophic cascades lead through a set of intermediate steps to increase densities of particular species or trophic groups higher in the web. In the latter, the top predator suppresses the trophic level below leading to increased densities two levels below. Thus, the expected responses should follow GWH/EEH predictions where alternating trophic levels are arranged with opposite densities (common—rare—common). For example, in a tritrophic (three-level) food chain, an increase in nutrients results in increases in the primary producer (plant) trophic level, decreases in the primary consumer (herbivore) level, and an increase in the top consumer level.

Proponents GWH and EEH suggest that strong trophic cascades occur in numerous food webs whereby entire trophic levels alternate in abundance via cascading food web interactions. However, empirical evidence shows that such cascades rarely or never occur on land and are apparently only present in a few aquatic communities. What determines whether a strong trophic cascade occurs or food web interactions weaken to become a trophic “trickle”? One major consideration is the efficiency of energy and resource transfer up the food chain. Highly efficient transfers lead to large numbers of top predators/consumers that would affect top-down control and strong cascades. Any factors that decrease the efficiency of energy/resource transfer would lessen the top-down control. In accordance with Polis and Strong’s (1996) multichannel omnivory, an increasing list of factors have been examined to explain the differences between GWH/EEH expectations and experimental results and observations of natural communities that generally show weak or no trophic cascades. These factors include omnivory, ontogenetic shifts, edibility, food quality, ecological stoichiometry, cannibalism, disease, body size refuges (for prey), allochthonous resources, seasonality, life history characteristics, predator avoidance behavior, and spatial and temporal heterogeneity in the availability of resources.

V. CURRENT TOPICS/TRENDS IN FOOD WEB STUDIES

Here, relatively under-studied aspects of food webs perceived to be central to understanding populations, communities, and ecosystems are identified. Some of the topics are now focal points for food web research, both empirical and theoretical.

A. Food Webs as Open Systems

Recent methods of tracing stable isotopes through a food web can provide much information on feeding relationships and on the sources of productivity that drive communities. For example, using stable isotopes or diet data, one can determine whether a community utilizes resources that originate in the benthic or pelagic zones of lakes or both. Virtually all natural systems are open and can exhibit tremendous spatial heterogeneity. Great spatial heterogeneity exists and nutrients and organisms ubiquitously move among habitats to exert substantial effects. However, food web studies have tended to focus on communities at a given site without regard to potential interactions with the surrounding habitat. Thus, little attention has been given to the fact that food web structure and dynamics are influenced by the movement of resources and organisms across habitat boundaries. Trophic linkage between habitats depends on the degree of differentiation in habitat structure and species composition. Systems that are moderately different tend to have broader transition zones and greatly overlap in species composition; these include grassland–forest, littoral–sublittoral, and benthic–pelagic zones. Habitats that have significant and abrupt changes in structure and species composition occur at the land–water interface. Moving resources (energetic or nutrients) can be utilized by different trophic types and the organisms that move across boundaries may also differ trophically (e.g., predators and prey). Studies of communities on island systems have shown that most of the allochthonous inputs (i.e., input from other habitats) from the ocean are available to detritivores, predators, and scavengers. Such movement of nutrients, detritus, food, prey, and predators is absolutely ubiquitous, occurring in virtually all communities and across all habitats. Some systems heavily dependent on allochthonous inputs include caves, mountaintops, snowfields, recent volcanic areas, deserts, marine filter-feeding communities in currents; soil communities; the riparian, coastal areas; and lakes, rivers, and headwater streams that...
receive watershed inputs. However, all systems depend on allochthonous inputs. For example, recent work shows that plant productivity in both the Hawaiian Islands and the Amazon forest is dependent on phosphorus input from thousands of miles away (China and Africa, respectively). The migrations (e.g., songbirds or geese) and movement of herbivores (e.g., wildebeest or hippopotamuses) can also result in large energetic flows across habitats.

Allochthonous inputs into the top level include carrion or carcasses, the movement of prey species into the habitat, and movement of predators across habitats. For example, the Allen paradox describes cases in which secondary production within streams is insufficient to support levels of fish production in them. Similarly, studies of coyote populations along the coast in Baja California demonstrate that they are highly subsidized by inputs from the ocean (about half of their diet) and are able to maintain a 3 to more than 10 times higher density than in adjacent inland areas. Predators moving along the interface between ecosystems (i.e., shorelines, riverbanks, and benthic and pelagic systems) can utilize resources across habitat. The river continuum concept argues that allochthonous resources entering into small headwater streams provide much of the productivity for organisms downstream in larger order streams. These allochthonous resources include prey, dissolved and particulate organic matter, and litter fall. Such inputs also power estuarine systems in which rivers carry allochthonous inputs into estuaries. Similarly, runoff from terrestrial systems into aquatic systems (and vice versa) provides litter, dissolved and particulate organic matter, and prey.

Spatial coupling can be key to dynamics. For instance, arboreal anole populations, subsidized by insects imported from light gaps, increase so as to suppress some predators and herbivores. Abundant detrital kelp from the sublittoral zone promotes dense intertidal limpets and urchin populations that then graze non-coraline algae to low cover. Allochthonous subsidies commonly influence stream systems: Leaf fall subsidizes herbivores, which in turn depress algae. Spiders that live along the coasts of streams, rivers, lakes, or the ocean are often very dense because they feed on aquatic insects. These spiders can then depress herbivores and thus increase the success of plants on which they live. Such spatial subsidies appear to be the foundations of most of the well-known trophic cascades. All these interactions are donor-controlled: Consumers do not affect the rate of import, availability, or dynamics of the allochthonous resources. However, subsidies allow consumers to be more abundant than if supported solely by in situ resources, with consequent suppression of in situ resources decoupled from in situ productivity. A common thread that has begun to link most thinking on food webs is that they are dynamical systems that vary over space and time. This approach has been liberating to ecologists, both empirical and theoretical. Recent empirical studies have found that communities and food webs contain multiple pathways that allow them to respond to environmental change and disturbance.

B. Detritus

Little of the energy fixed by plants passes directly into the grazing food chain—herbivores eating plants and then eaten by carnivores. Most of this primary productivity is eaten by herbivores (median >80% on land, −30% in water). What happens to this dominant chunk of the world’s productivity? Is the detrital web a self-contained sink internally recycling energy and nutrients or a link that affects the population dynamics of the larger species?

Uneaten plants (and animals) enter the detrital web, in which they are processed by microbes, fungi, and some animals. Although some ecosystems are net accumulators of undigested biomass (e.g., carboniferous bogs and forests that supply today’s oil and gasoline), most ecosystems do not accumulate plant biomass. Rather, it is soon digested by detritivores, with nutrients and energy passing through “functional compartments” composed of diverse microbes and animals. Several factors regulate the flow and availability of detritus to detritivores and then onto other consumers. A major question is whether the detrital community is a sink that metabolizes most of this energy or a link that passes this energy up the food chain. An unknown fraction of detrital energy and nutrients re-enter grazing food chains when some detritivores are eaten by predators that also eat herbivores (e.g., a robin eats an earthworm). Such “detrital shunts” are common, interweaving energetics and dynamics of biophages and saprophages. Bypassing herbivores, this linkage can affect herbivore regulation in a manner analogous to the spatial subsidies to consumers discussed previously. Predator populations, subsidized by detritivorous prey, can increase and suppress other predators or herbivores. The exact effect of detrital shunts depends on the relative benefits for each species and where detritus reenters (to producers, herbivores, and intermediate or higher consumers). For example, nutrients from detritus greatly influence plant productivity; models show
that a 10% reduction in detritus can cause a 50% reduction of plant biomass. The dynamics of consumer control within the detrital web and those produced by infusion of detritivores into the grazing web are undoubtedly crucial to community structure and dynamics. For example, detrital shunts to predators in the grazing chain can create the appearance of a simple linear trophic cascade, but with the difference that nutrition from detritivores sustains or elevates predators to levels sufficient to suppress herbivores.

C. Age Structure Effects in Food Webs

Almost all species display complex life cycles, marked by moderate to radical changes in diet and habitat; such life histories fundamentally must affect every species with which they interact. However, our understanding of how age- and stage-structured processes affect food webs and communities is embryonic.

Life history omnivory describes shifts in diet during development; often, they are accompanied by ontogenetic changes in habitat. Diet can change substantially either discontinuously (e.g., at metamorphosis) or slowly with growth. Such life histories are widespread; an estimated 80% of all animal species undergo metamorphosis. Changes in resource use can be dramatic (e.g., predaceous juveniles, plant-feeding adults in parasauroids and many other insects, and herbivorous tadpoles and predaceous frogs and toads), with prey size variation as great as three or four orders of magnitude. Even among nonmetamorphic species, diets change greatly with age, with diet differences among age classes often more distinct than those among most species.

Overall, complex life histories and age structure omnivory can exert diverse and profound effects on the dynamics of populations and food webs. For example, they can either impede consumer control or amplify resource suppression via dynamics similar to those of spatial subsidy or detrital shunts.

D. The Roles of Nutrients and Stoichiometry

Animals require both energy and a variety of "nutritional requisites" to grow, complete their life cycle, and reproduce. Important nutrients include nitrogen, phosphorus, some trace elements, fatty acids, and vitamins. Nitrogen is an integral component of many essential compounds: it is a major part of amino acids, the building blocks of protein, including the enzymes that control virtually all cellular processes. Other nitrogen compounds include nucleic acids and chlorophyll. Phosphorus is used for adenosine triphosphate (ATP, the energy currency of all cells), nucleic acids (DNA and RNA), and phospholipids, particularly in cell membranes.

The availability of nutritional requisites constrains growth and reproduction in virtually every species. Nitrogen and phosphorus are particularly important. The ratio of carbon to nitrogen (C : N) in plants ranges from 10:1 to 30:1 in legumes and young green leaves to as high as 600:1 in some wood. The C : N ratios in animals and microbes are much lower, ordinarily between 5:1 and 10:1. Such differences in C : N ratios between plants and their consumers lower the rate of decomposition by microbes. There is ample evidence that heterotrophs chronically lack adequate nitrogen to grow or reproduce optimally. The importance of nutritional restriction is reinforced by the foraging literature that clearly shows that herbivores choose their foods based on nutrient as well as energy content.

In many cases, phosphorus availability constrains herbivore success. The Redfield ratio describes the approximate stoichiometric mix (110 C : 250 H : 75 O : 16 N : 1 P) of elements found in marine systems. In particular, the N : P ratio crucially determines productivity and species composition. Thus, energy (C–C bonds) and nitrogen could be abundant, but neither individuals nor populations grow maximally because phosphorus is insufficient. Because phosphorus is essential to cell division (and thus reproduction), a high N : P ratio especially limits the growth of organisms that have high potential rates, such as most herbivores and detritivores. These organisms are key to the potential regulation of plant biomass (and "detritus"). Evidence suggests that high N : P ratios can impede trophic cascades. For example, Daphnia, a key to many lake cascades, respond sufficiently rapidly to phytoplankton productivity to depress plant biomass. In lakes with inadequate phosphorus, slower growing copepods replace Daphnia; these copepods do not have the reproductive capacity to depress phytoplankton biomass.

Ecologists are beginning to understand how stoichiometry and nutritional balance affect population and food web dynamics. Nevertheless, it is extremely likely that herbivore growth is often less than maximal solely because their environment does not provide sufficient quantities of all key nutritional requisites. In fact, the greatest disparity in biochemical, elemental, and stoichiometric composition in the entire food web occurs at the link where herbivores convert plant material into animal tissue. The implication is clear: Even in a world full of green energy, many or most herbivores cannot...
obtain enough requisite resources to grow, survive, or reproduce at high rates. Nutritional shortages regulate herbivore numbers and often limit their effects on plant biomass.

Recent theoretical studies of the role of food quality in terms of edibility and nutrient content show that low food quality can greatly influence consumer resource interactions. This has two important consequences. First, low food quality reduces the growth rate of the consumer, making that interaction more stable. Second, in systems in which multiple resources could be limiting, the addition of large amounts of a single resource (such as nitrogen or phosphate) may increase that resource to a level at which it is no longer limiting; however, a second resource would become limiting and so on. This sequential limiting of resources means that the addition of a single resource would not push the system into highly unstable dynamics, reducing the probability that the “paradox of enrichment” occurs. Rosenzweig introduced the concept of the paradox of enrichment to explain the addition of a resource leading to the collapse of a consumer–resource interaction. This happens because the addition of the resource drives the population of the consumer to a higher level that results in overcompensation by the consumer (predator) driving the resource (prey) extinct. However, most systems have several potentially limiting resources. For example, Leibold’s study of ponds found that nitrogen additions do not lead to strong trophic cascades or the paradox of enrichment because light becomes limiting with relatively modest nitrogen additions.

E. Interaction Strength

One goal of functional webs is the quantification of interaction strengths within food webs. Various definitions have been used for “interaction strengths.” In Lotka–Volterra models, interaction strengths are due solely to the direct interactions between species pairs and are measured on a per capita basis. Estimations of the strength of these direct interactions are fraught with difficulties. Measurements in artificial systems may not allow for behavioral responses. For example, Sih has shown that prey species have different escape mechanisms or routes depending on the species of predator. Thus, when in the presence of two predators, the response of a prey may result in its increased susceptibility to one or the other predator due to a behavior that is not evidenced when only the one predator is present.

Measurements in natural systems are also problematic because they may not account for indirect interactions. Many studies have elucidated the interaction strength among pairs of species. However, indirect effects may play a strong role in determining the realized interaction strength. Thus, Paine has argued that interaction strengths should always be measured in the field with the full complement of natural species present and that these measurements should incorporate all indirect effects. The realized interaction strength accounts for all direct and indirect interactions. For example, predator–prey interactions are functionally negative due to the direct effect. However, the indirect effect of a predator may reduce the number of competitors of the prey species, thus resulting in an overall positive interaction strength (direct + indirect effects). Therefore, potentially strong indirect effects can make mechanistic interpretation of experimental results among species difficult.

Path analysis, a new statistical method, has been used to evaluate causal hypotheses concerning the strengths of interactions in many systems. Path analysis is essentially a multiple regression on each species in which specific causal relationships (e.g., alternative food web configurations), specific experimental treatments, and other interactions are diagrammed in a community interaction web. The community interaction is essentially a food web to which nonconsumptive interactions, such as pollination, competition, and mutualisms, are added. Hypotheses for the causal relationships between pairs of species not directly linked can become quite complicated. However, path analysis can test different hypothesized community web structures by accounting for both direct and indirect relationships. Then, experimental manipulations (e.g., species removals or additions) can test predictions of the path analysis.

F. Can Energetic Webs Provide Insight into Population and Community Dynamics?

A problem in food web studies is how to connect the great amount of quantitative information in energetic webs to population and community dynamics described by functional webs. Much progress would occur if we could determine the dynamical importance of a particular species or feeding link from an inspection of the magnitude of energy transfer or diet composition. Unfortunately, no clear answer is forthcoming. In fact, it appears that even highly quantified information such as the number of calories passed along a certain pathway or the frequency of prey in the diet of a consumer conveys little information about the dynamics of inter-
acting populations because these descriptive parameters do not correlate with interaction strength. There is no clear rationale to argue that food web dynamics and energetics are necessarily correlated; indeed, logic and evidence suggest dynamics often cannot be predicted from data on diet or energy flow. The degree of resource suppression is not a function of energy transfer. Consumer regulation of populations need involve little energy transfer and few feeding interactions. For example, removing predatory rats from New Zealand islands increased lizard abundance 3–30 times although lizards formed <3% of rat’s diet. Key regulatory factors may produce much less overall mortality than other factors. Brief, intense predation episodes may net little energy for the predator but may be central to prey dynamics. The consumption of young stages (seeds, eggs, and larvae) may provide trivial energy to a consumer but can greatly depress prey abundance. Pathogens and parasites form an extreme example: They take little energy, even when they decimate their host populations. In a well-studied food web of the marine benthic community in the Antarctic, Dayton showed that the species apparently exerting the strongest effects on the structure and dynamics of this community would be deemed unimportant from analyses of diet, energy transfer, or biomass. Such discoveries have stimulated many to argue that, without experimentation, one cannot a priori decide which are strong or weak links. An apparently weak link (in terms of diet or energy transfer) can be a key link dynamically, and an important energetic link may affect dynamics little. No necessary concordance of dynamics with either dietary or energetic measures exists. This insight counters the use of energetics to recognize strong interaction links.

G. Modeling Food Webs

To many ecologists, early food webs of Forbes, Sumnerhayes, and Elton and those of Lindeman emphasized the overwhelming complexity of natural systems and the need to simplify them into distinct trophic groups. This perspective was culminated in the green-world hypothesis of Hairston et al. (1960). Oksanen et al.’s (1981) EEH expanded this view for ecosystems that had fewer or more than three trophic levels and for which the exact number of trophic levels was set by productivity. The top level would then regulate the one below it and this would release the one below it, etc. In this sense, both GWH and EEH suggested that all ecosystems are essentially regulated from the top-down by predation. Lindeman envisioned the food web (or as he called it, the ‘food-cycle’) as a dynamic system in which energy and nutrients are transferred from one trophic level to the next and recycled. This was an important departure from simply determining feeding connectedness (and from the GWH) in that ecosystems could be regulated from the bottom up by the flow of energy and materials from the level below. However, much more information and data are required to quantify the transfer of energy (and material) through food webs, but this view allows for a more analytical approach.

MacArthur focused the attention of ecologists on the trophic–dynamic approach with his hypothesis that increasing complexity of community organization leads to increasing dynamic stability. The reasoning was simple: When predators have alternative prey, their own numbers rely less on fluctuations in numbers of a particular species. Where energy can take more routes through a system, disruption of one pathway merely shunts more energy through another, and the overall flow continues uninterrupted. MacArthur’s analytical approach linked community stability to species diversity and food web complexity and it stimulated a flurry of theoretical, comparative, and experimental work. This work may be divided into two contemporary approaches that use food webs to study community structure. The first approach involves the study of the properties of food web diagrams with the goal of uncovering general patterns that suggest mechanisms of community stability. This is done both by comparing food webs from natural communities and by the use of simulation and mathematical modeling to study hypothetical food webs. This research has yielded much of the terminology now associated with food webs and generated a body of food web theory that includes many hypotheses about community structure. The second approach, which grew from early theoretical and experimental community studies, involves the dynamical analysis of food webs to determine not only the pattern of interactions among the populations in the community but also the relative strengths of those interactions. Dynamic food web analysis also seeks to reveal interactions that are not obvious from simple food web diagrams, so-called indirect interactions. This approach requires the careful merging of experimental and theoretical approaches. The simplicity of the GWH enabled it to be a reasonable starting point to examine the dynamics of food webs. In general, dynamical models are rooted in a tradition based on the application of Lotka–Volterra equations to communities and advocated by May (1973). One of the major conclusions from these phe-
nomenclatural models is that complexity (e.g., omnivory and long chains) causes instability in model systems. This conclusion was viewed with skepticism by empiricists because observations from field studies (such as work by MacArthur) suggested that increased complexity should result in increased stability. Recent theoretical investigations into the relationship between stability and complexity have found that assumptions and structure of earlier models may have biased them toward decreased stability with increasing complexity.

Early theoretical studies of interactions and consequences of these interactions in food webs were based on equilibrium dynamics of Lotka–Volterra models. The assumption that ecological systems or species populations have some ‘equilibrium’ around which they fluctuate is totally unrealistic. Furthermore, these early models ignored the central belief of many empiricists that most interactions between species were weak. The outcome of many of these theoretical studies went against common sense intuition and the findings of empirical studies, including that omnivory was destabilizing and therefore rare and that complexity (greater diversity) was also destabilizing. Recent studies that incorporated the findings of mostly weak interactions and nonequilibrium dynamics have found that omnivory and complexity may actually stabilize food webs. This agrees with both the intuition and the current arguments of empiricists who find that many weak interactions occur within food webs and these promote stability.

Recent theoretical studies suggested three factors as important to reduce stability in earlier models: (i) linear Lotka–Volterra equations, (ii) using equilibrium solutions to these equations, and (iii) the distribution of interaction strengths overly estimated the number of strong links. Many studies have shown that many predator–prey relationships are not linear, but instead predators exhibit saturation such as described by a Holling’s type II functional response. Current models take advantage of this and use energetic uptake rates that saturate and therefore rare and that complexity (greater diversity) was also destabilizing. Recent studies that incorporated the findings of mostly weak interactions and nonequilibrium dynamics have found that omnivory and complexity may actually stabilize food webs. This agrees with both the intuition and the current arguments of empiricists who find that many weak interactions occur within food webs and these promote stability.

Recent theoretical studies suggested three factors as important to reduce stability in earlier models: (i) linear Lotka–Volterra equations, (ii) using equilibrium solutions to these equations, and (iii) the distribution of interaction strengths overly estimated the number of strong links. Many studies have shown that many predator–prey relationships are not linear, but instead predators exhibit saturation such as described by a Holling’s type II functional response. Current models take advantage of this and use energetic uptake rates that saturate and therefore rare and that complexity (greater diversity) was also destabilizing. Recent studies that incorporated the findings of mostly weak interactions and nonequilibrium dynamics have found that omnivory and complexity may actually stabilize food webs. This agrees with both the intuition and the current arguments of empiricists who find that many weak interactions occur within food webs and these promote stability.

In modeling food webs, a key consideration is the functional relationship between a consumer and its resource. As noted previously, Lotka–Volterra consumer–resource relationships are linear (type I). This assumes that the predators do not become saturated and can consume all available prey. Holling introduced nonlinear consumer–resource functional relationships with his disk (now called type II) functional response. This functional response assumes that the capture and consumption/digestion time of prey by the predator limits the amount of prey taken by a predator in a given amount of time. Holling also introduced a third type of functional response (type III) to simulate a predator switching capture of prey when a target prey species becomes rare to a more abundant prey species. Various other functional responses have been introduced, including Ginzburg and Arditi’s ratio-dependent functional response. Ratio dependence assumes that the growth rate of the predator is dependent on the ratio of prey and predator densities, whereas in types I–III predator growth rates are dependent only on the prey densities (prey dependent). Ratio-dependent models predict that all trophic levels increase proportionately, whereas prey-dependent models predict the alternating pattern of GWH and EEH. The arguments against ratio dependence arise from the lack of a mechanistic basis for the model. Proponents of ratio dependence argue that the formulation is simpler and can account for essential dynamics of food webs without added complexity. Detractors argue that using more mechanistic, albeit more complex, models that can account for realistic interactions is the correct way to proceed. Regardless of these arguments, using intermediate levels of complexity based on realistic mechanisms is the current trend in food web theory.

H. Intermediate Levels of Complexity

Community ecology has focused on interactions (mainly competition and predation) between pairs of species that are fundamentally important in food webs.
However, these interactions, taken out of context of the larger web, may result in misleading information (due to indirect effects). Highly complex food webs, however, are unwieldy and intrinsically difficult to study in model systems. Thus, Holt and others suggested investigating the dynamics of intermediate (between species pairs and whole food webs) levels of complexity—so-called “community modules”—that are defined as small subsets of species that are characterized by strong interactions. These modules are also more representative of the levels of complexity (i.e., number of species) examined in experimental studies. Recent theoretical studies have taken advantage of intermediate complexity by focusing on food web interactions that typify interactions found in real food webs but are common to many food webs. This allows one to examine how indirect effects interact with direct effects to structure food webs. Common types of interactions among sets of species and their resources (modules) are apparent competition, intraguild predation, omnivory, cannibalism, and spatial subsidies.

This modular approach has allowed for various theoretical studies to examine stability of food webs using mathematical approaches. For example, McCann, Hastings, and Huxel found that adding relatively weak interactions among species could enhance food web stability. They found this to be true for apparent competition, intraguild predation, omnivory, cannibalism, and spatial subsidies. Are weak interactions typical of food webs? The answer, from the few studies that have specifically examined this question, is yes. For example, in a study on intertidal food webs, Paine found that most interaction strengths are weak. Moreover, knowing that most predators eat tens to 100 species of prey suggests that most of these interactions are weak.

One may then ask, what about strong interactions? The answer goes to the heart of one major problem with earlier food web studies. Strong interactions may occur and be a regular component of food webs. However, in almost every case, they appear to be enabled by “multichannel omnivory” (i.e., feeding on many weak links, Polis and Strong, 1996) or are restricted temporally and/or spatially because they are inherently unstable. However, time and effort constraints and tradition have caused the vast majority of food web studies to ignore these weak interactions and the spatial and temporal aspects that characterize all systems.

In another theoretical study of food web processes that took advantage of the modular view, Huxel and McCann examined the flow of the allochthonous energetic resources. They found that allochthonous resources may spread evenly throughout the community or may become compartmentalized. High levels of allochthonous resources decreased stability, whereas low levels increased stability. Thus, again a weak link tended to increase stability.

See Also the Following Articles

DIVERSITY, COMMUNITY, REGIONAL LEVEL • ECOSYSTEM, CONCEPT OF • ENERGY FLOW AND ECOSYSTEMS • KEYSTONE SPECIES • PREDATORS, ECOLOGICAL ROLE OF • SPECIES INTERACTIONS • TROPHIC LEVELS

Bibliography


I. Canopy Architecture, Animal Substrate
II. Exploring the Last Biotic Frontier
III. Results of Studies
IV. Conclusions

GLOSSARY

*arbicolous* Living on the trees, or at least off the ground in shrubs and/or on tree trunks.
*emergent* A very tall tree that emerges above the general level of the forest canopy.
*epiphytic material* Live and dead canopy vascular and nonvascular plants, associated detritus, microbes, invertebrates, fungi, and crown humus.
*hectare* Metric equivalent of 2.47 acres.
*microhabitat* A small self-contained environmental unit occupied by a specific subset of interacting species of the forest (or any other community).
*scansorial* Using both the forest floor and canopy for movement and seeking resources.
*terra firme forest* Continuous hardwood forest of the nonflooded or upland parts of the Amazon rain forest.

*The forest canopy* is arguably the most species-rich environment on the planet and hence was termed the "last biotic frontier," mainly because until very recently it had been studied less than any place else, with the exception of the deep ocean floor and outer space. The reason for lack of study of the canopy was accessibility, and the evidence of the incredible species richness, mainly of tropical forests, is primarily the abundance of insects and their allies. This hyperdiverse and globally dominant group has adapted to every conceivable niche in the fine-grained physical and chemical architecture of the tree crowns. In less than three decades, canopy biology has become a mixed scientific discipline in its own right that is gradually gaining sophistication of both approach and access.

Tropical arbicolous (tree-living) arthropods were observed in the early 1800s in the "great forests near the equator in South America" and later that century were described by Henry Walter Bates. Even though Bates observed, described, and commented on the canopy fauna (as viewed from the ground and in recently felled trees), more than a century passed before Collie designed an insecticide application technique that allowed a rigorous sampling regime for canopy arthropods. William Beebe and collaborators early in the twentieth century recognized that the canopy held biological treasures, but "gravitation and tree-trunks swarming with terrible ants" kept them at bay. Frank Chapman, a canopy pioneer (of sorts), viewed the treetops from his "tropical air castle" in Panama in the 1920s, but his interest was vertebrate oriented, his perch was a small tower, and his observations of insects and their relatives were casual. By the mid-1960s and early 1970s, a few workers in both basic and applied science were seriously...
investigating canopy faunas of temperate and tropical forests in both the Western and Eastern Hemispheres. From the early 1980s until now, many workers have been improving methods of access and other techniques used to register, sample, and study the fauna (see reviews by Basset, Erwin, Malcolm, Moffett and Lowman, Munz and Loiselle, and Winchester in Lowman and Nadkarni, 1999; Moffett, 1993; Mitchell, 1987). Some of these workers have found that arthropods by far make up the fauna of the canopy (Erwin, 1982, 1988). Visiting and nesting birds, mammals, reptiles, and amphibians represent a mere 1% or less of the species and even less in the abundance of individuals in these groups (Robinson, 1986). There are no adequate measures of canopy nematodes, mollusks, or other nonarthropod microfauna groups.

What is meant by the forest canopy? Generally, the canopy, or tree crown, is thought of as that part of the tree including and above its first major lateral branches. The canopy of a single tree includes the crown rim (the leaves and small twigs that face the main insolation from the sun) and the crown interior (the main trunk and branches that gives a tree its characteristic shape). The canopy fauna is that component of animal life that inhabits the tree canopy and uses resources found there, such as food, nesting sites, transit routes, or hiding places. Hence, the forest canopy is collectively all the trees of all the trees in an area. The canopy is often thought of as being stratified into emergents, one to three regular canopy strata, and an understory of smaller trees living in the shade of a more or less continuous overstory. All types of forests have their own describable characteristics, from the spruce forests of the Northwest Territories of Canada to the pine forest of Honduras, the dry forests of Costa Rica and Bolivia, and the Rinorea and Mauritia forests of the upper Manu River in Peru. It is through "whose eyes" one views the community, habitat, or microhabitat that determines the scale of investigation and subsequent contribution to the understanding of the environment—the beetles, the rats, the birds, the ocelots, the investigators, or perhaps even the trees.

I. CANOPY ARCHITECTURE, ANIMAL SUBSTRATE

A temperate forest is composed of both broad-leaved and coniferous trees, with one or the other sometimes occurring in near pure stands depending on the latitude and/or altitude and also on soil and drainage conditions. Normally, there are few canopy vines or epiphytes and perhaps some wild grape or poison ivy vines. Soil and organic debris caches are few or absent in the tree crowns, except for tree holes which provide homes to numerous arthropod groups but few vertebrates. Temperate forests are subjected to cold and hot seasonal climate regimes as well as wet and dry periods. Great expanses of forest lose their leaves in the winter months, sap ceases its flow, and the forest "metabolism" comes to a slow resting state. The temperate forest seemingly provides a great variety of substrates for the canopy fauna, but faunas are depauperate compared to those in tropical forests. Virtually no mammals are restricted to temperate forest canopies—only a few frogs and lizards. However, many bird species are restricted to the canopies, as they are in tropical forests. Among insects, for example, the beetle family Carabidae has 9% of its species living arboreculously in Maryland, 49% in Panama, and 60% or more at the equator in South America. Tropical forests, on the other hand, have few if any coniferous trees; only forests at higher elevations and/or or located closer to subtropical zones have coniferous trees. Tropical canopies are often (but not always) replete with vines and epiphytes, tree holes, and tank bromeliads, and there are soil mats among the roots of orchids, bromeliads, and aroid plants. In the early 1990s, Nadkarni and Longino demonstrated that epiphytic material is fraught with macroinvertebrates, and Coxson and Nadkarni later showed that epiphytic material is important in the acquisition, storage, and release of nutrients. Lowland tropical forests are subjected to mild temperatures, without frost, but have both wet (sometimes severe) and dry seasons. Individual species of trees may be deciduous, but in general tropical forests are always green and there is a perpetual growing season. Substrates are constantly available for the fauna. Often, some microhabitats with their substrates are temporary in the sense that they remain in place for a season or two, but then their architectural structure collapses into a jumbled pile of organic detritus on the forest floor. Such microhabitats (e.g., a suspended fallen branch with its withering leaves) provide a home resource to thousands of arthropods in hundreds of species, many found only in this setting. Eventually, such a branch loses its dried leaves and crashes to the forest floor. However, a short distance away, another branch breaks from a standing tree and the process begins again. The arthropods of the old, disintegrating branch move to the new one. The microhabitat and its substrates are forever present across the forest; each individual branch
is ephemeral. The faunal members occupying such microhabitats are good at short-range dispersal.

II. EXPLORING THE LAST BIOTIC FRONTIER

Until recently, the forest canopy was impossible to study well. Getting there was the limiting factor, and even after getting there (e.g., via ropes) it was difficult to find the target organisms. Modern devises such as aerial walkways (e.g., ACEER, Tiputini Biodiversity Station, Fig. 1), one- or two-person gondolas maneuvered along crane booms (e.g., in Panama at STRI), and web-ropeing techniques (see review by Moffett and Lowman in Lowman and Nadkarni, 1995) now allow real-time observations, sampling, and experiments anywhere in the canopy. Inflatable rafts that suspend mesh platforms resting on the upper crown rims of several trees have provided access from above, although this technique seems more suited to botanical work or leaf-mining insects, especially epiphytes and lianas. Insecticidal fogging techniques allow passive sampling of all arthropods resting on the surfaces of canopy plants (Erwin, 1995), and suspended window/malaise traps collect the active aerial fauna. Many of these techniques have been used during the past two decades; however, often they were simply used as collecting devises to garner specimens for museums and/or for taxonomic studies, and for this purpose they are excellent. In some cases, ecological studies were desired, but the techniques were not properly applied and the results disappointing. It is important to first ask the questions and then design the experiments; in some cases, current canopy techniques can be powerful tools for answering questions. Unfortunately, although sampling is relatively easy, sample processing is time-consuming and laborious. For canopy fogging studies, after the sampling effort an average of 5 years was required before published products were achieved (Erwin, 1995). The main reason for this is a lack of funding for processing the results of fieldwork, even though the field studies were readily funded. Without processing, the data inherent for each specimen are unavailable for taxonomy or ecology studies. This is an historical funding problem and one of the reasons most studies examine but a few species from few samples.

III. RESULTS OF STUDIES

A. Invertebrates

Recent findings by Adis in the central Amazon Basin and by Erwin in the western part of the basin demonstrated that there are as many as $6.4 \times 10^{12}$ terrestrial arthropods per hectare. A recent 3-year study of virgin
terra firme forest near Yasuni National Park in Ecuador by Erwin found an estimated 60,000 species per hectare in the canopy alone. This figure was determined by counting the actual species in the samples with their known described taxonomic diversity. The predatory beetle genus, *Agra* (Fig. 2), has more than 2000 species found only in Neotropical forest canopies and scattered remnants of subtropical forest canopies in southern Texas and northern Argentina. The herbivorous weevil genus, *Aption*, likely has more than 10,000 species. In only 100 9-m² samples of canopy column from 1 ha of virgin terra firme forest near Yasuni National Park in Ecuador, there are more than 700 species of the homopteran family, Membracidae, which were found along with 308 species of the beetle family, Carabidae, and 178 species of the spider family, Theridiidae.

“Biodiversity” by any other name is “Terrestrial arthropods”—that is, insects, spiders, mites, centipedes, millipedes, and their lesser known allies. Forest canopy studies of terrestrial arthropods are few (Erwin, 1995). Many of these studies currently concentrate on host specificity as a herbivore or parasite that eats only one other species of plant or animal. However, there is another class of specificity that is very important in understanding biodiversity that has received almost no study: “where” species hide and rest. This is not random but rather species specific (T. L. Erwin, unpublished data).

Terrestrial arthropods are found in “hotels” and “restaurants” or “in transit” between the two (Fig. 3). Often, insects and their allies eat, mate, and oviposit in the restaurant or at the food source, for example, on fungi or in suspended dry palm fronds. These insects may hide during the day under debris or under bark near the fungus or on the palm debris, but they never roam far from the vicinity of the food source, except to locate new food sources when the old one is depleted. Members of other species eat in one place and then move to cover for a resting period, i.e., the hotel. An example of this is the subfamily Alleculinae of the beetle family Tenebrionidae. These beetles feed on lichens and moss on tree trunks at night and spend the day (hiding, resting, and possibly sleeping) in suspended dry leaves elsewhere in the forest. Many species found in the forest canopy during the day (utilizing leaves, fruits, and/or flowers) hide and rest at night in the understory (e.g., various pollen-feeding beetles and the larger butterflies).

Insects particularly, and some of their allies, have adapted to nearly every physical feature of the planet, and the canopy is no exception. Many beetles have special feet for walking on leaves; some even have modified setae on their feet to slow them down upon landing from rapid flight (Fig. 4). Because they are in an environment with rapturous birds, lizards, and frogs, many insect species have evolved camouflage coloration.

Climate is the main constraint on terrestrial invertebrates. In the temperate zones, it is the winter cold and
dryness; in the equatorial tropics, it is the dry season for some and the rainy season for others, with the temperature far less of an influence than it is in the far north or south. Many herbivores must contend with plants that produce toxic chemicals or other defensive systems. All insects must also deal with other insects that predate, parasitize, or carry bacteria, fungi, or other insect diseases. Hammond, Stork, and others, in their studies of insects in the Sulawesi dipterocarp forests, and Miller, Basset, and others in New Guinea found much less insect diversity and richness than Erwin and his teams in the Neotropical forests. Hammond also found in southwest Asia that the canopy fauna was not as delimited from the understory fauna as it is in the Amazon Basin. Unfortunately, all these teams used different methodology; hence, much of their results are not comparable. It is certain, however, that the Old World tropical forests are not as biodiverse as those in the New World, nor are the forests of Costa Rica and Panama as diverse as those of the Amazon Basin. Disparate regional richness is one of the main problems in estimating the number of species on the planet. Another is the incredible richness of terrestrial arthropod species and the fact that scientists likely know less than 3–5% of them if published estimates of 30–50 million extant species are close to reality. Stork (1988) has even gone so far as to suggest that there could be 80 million species on the planet.

B. Vertebrates

Availability of food year-round constrains vertebrates from living strictly in canopies (see reviews by Emmons and Malcolm in Lowman and Nadkarni, 1995). Only in evergreen rain forests is there a continuous supply of food (albeit somewhat dispersed and sporadic) for phytophagous and insectivorous vertebrates. In deciduous forests, most species also forage on the ground or hibernate when food supplies are short. Almost all canopy mammals live in evergreen tropical forests, but even there most are scansorial. Timing and distribution of food resources are the critical controlling factors. Among all nonflying vertebrates, anurans and lizards and to a lesser extent snakes are the most important truly canopy creatures. Birds and bats are also exceedingly important components. All these groups except snakes account for vertebrate predator-driven evolution on the far more dominant invertebrates of the canopy. For example (as Blake, Karr, Robinson, Servat, Terborgh, and others have shown), throughout the tropics approximately 50% of birds are strictly insectivores, whereas another 8% take insects and nectar.

Morphological adaptations that allow canopy life include feet that can firmly grip the finely architectured substrate of twigs, leaves, and scaly bark. Emmons, in her many articles on Neotropical mammals, demonstrated that among these animals, those with the ability to “jump” avoided wasting energy and time by descending and climbing new trees to find resources; hence, more true canopy species have this ability. This is certainly true also of frogs and lizards. However, it is the flying forms—birds and, to a lesser extent, bats—that account for most of the treetop vertebrate fauna. Physiological adaptations that allow vertebrate canopy life include the ability to subsist on diets of fruit, flowers, leaves, or insects and their allies. Among mammals, fruit eaters are dominant.

As shown by Duellman, Dial, and others, among canopy anurans and lizards, nearly all are primarily insect predators. Birds are overwhelming insectivorous in the canopy fauna, with approximately 40% in the upper Amazonian and 48% at Costa Rica’s La Selva.
Biological Station. Malcolm, in summarizing the few articles on the subject, estimates that 15% of mammal species are arboreal/scansorial in temperate woodlands, whereas between 45 and 61% exhibit this behavior in tropical forests. In Duellman's 1980 list of anurans and reptiles from Neotropical forest, 36% are strictly arborescent, whereas 8% are scansorial. Among birds, Blake and others found that scansorial species using the understory and ground were more numerous than strictly canopy species (31 and 42%, respectively), at their site in Costa Rica.

In summary, although canopy vertebrates are important in driving part of invertebrate evolution in the forest canopy, they have not overwhelmingly radiated into or made use of the canopy, as have the invertebrates. For example, the total vertebrate fauna known at Cocha Cashu, Peru, is approximately 800 species (approximately 45% of which are arborescent or scansorial), whereas at a nearby location there are nearly 900 species of the beetle family Carabidae, of which more than 50% are strictly arborescent. In Ecuador, near Yasuni National Park, there are in excess of 600 species of the homopteran family Membracidae in a single hectare, 100% of which are strictly arborescent.

IV. CONCLUSIONS

Although animals may use the air for dispersal, they live on substrate. Here, they eat, mate, hide, and walk. Forest canopies are rich in species because they offer a three-dimensional array of varying substrates that directly receive the sun's energy with little filtering.

Although much has been and is being accomplished by faunal studies of the forest canopy, there is still much to do. There are missing data links between vertebrates and invertebrates and between both of these and the plant food and plant architecture on which they depend, and data is also missing on the influence of the canopy physical features on the fauna such as microclimates (see Parker’s review in Lowman and Nadkarni, 1995). Each subsystem is receiving at least some attention, but the new discipline of canopy biology is in its infancy. Is it too late? The forests and their species-rich canopies are rapidly disappearing (World Resources Institute, 1993).

Topics of current investigation include canopy insect β diversity and measures of host specificity; the latter particularly in leaf-feeding beetles. Both areas of study were driven by earlier, somewhat naïve estimates of millions of species extant on the planet (Erwin, 1982; Stork, 1988; May, 1990; Casson and Hodkinson, 1991; Gaston, 1991). Although some of these studies may have been internally consistent within the parameters set for the estimations, no one had really gotten a handle on the true meaning of “host” specificity, biocomplexity of tropical forests, the influence of tropical biotope mosaics, β diversity or what is known as species turnover in space and/or time, or the disparities of richness among continents or even the disparity among regions within continents.

Even so, our current rudimentary knowledge indicates that we are losing hundreds, even thousands, of invertebrate species with “scorched earth” programs such as that in Rondonia, Brazil, clear-cutting of Borneo and other southern Asian forests, and other losses in Haiti, Puerto Rico, Hawaii, the western Amazon Basin, Madagascar, and so on.

Conservation strategies are currently dominated by data on vertebrates (Kremen et al., 1993; Samways, 1994); however, invertebrates are rapidly becoming sufficiently known to include them in analyses that are directed toward preservation of forest communities; to this end, the collective human conscience will soon be dealing with real extinction processes equivalent to those in the past, from the Permian to the Cretaceous. We are living at the beginning of the so-called “sixth extinction crisis” sensu Niles Eldridge of the American Museum of Natural History. Amelioration of the impact of this crisis rests on a better knowledge of the natural world around us and the development of conservation strategies that consider what we, Earth’s managers (whether we like it or not), want future evolution to look like, as so well described by David Quammen (1998).

See Also the Following Articles

AMAZON ECOSYSTEMS • ARTHROPODS, AMAZONIAN • BEETLES • FOREST CANOPIES, PLANT DIVERSITY • FOREST ECOLOGY • INVERTEBRATES, TERRESTRIAL OVERVIEW • TROPICAL ECOSYSTEMS

Bibliography


FOREST CANOPIES, PLANT DIVERSITY

Nalini M. Nadkarni,* Mark C. Merwin,* and Jurgen Nieder†
*The Evergreen State College and †Botanisches Institut der Universität Bonn

I. Introduction
II. Categories of Canopy Plants
III. Canopy Plant Taxa Diversity
IV. Diversity of Habitats of Canopy Plants
V. Diversity of Growth Habits in Canopy Plants
VI. Canopy Plant Biodiversity and Conservation Biology
VII. Areas for Further Study

GLOSSARY

accidental epiphyte Plant that normally grows terrestrially but that occasionally grows to maturity in a tree crown, usually in terrestriallike microsites such as the crotches of branches.
bryophyte Nonvascular plant of the division Bryophyta (a moss, liverwort, or hornwort).
cryptogam Plant that reproduces by spores or gametes rather than seeds; includes bryophytes and lichens.
epiphyll (flocculent) Plant that grows on the leaf surface of another plant.
epiphyte Nonparasitic plant that uses another plant as mechanical support but does not derive nutrients or water from its host.
facultative epiphyte Plant or lichen that commonly grows epiphytically and terrestrially, usually exhibiting preference for one or the other habit in a particular habitat.
homoiohydry Ability to maintain a constant internal water balance independent of fluctuating environmental conditions.
lichen Composite organism consisting of a fungus (the mycobiont) and an alga and/or a cyanobacteria (the phycobiont) that live in a symbiotic relationship.
mistletoe Woody parasite that taps the xylem of a tree, but is capable of photosynthesis.
oblige epiphyte Plant that always grows on another plant for structural support, but derives no nutrients from the host.
parasite Woody or nonwoody plant that taps into the vascular system of a host plant and derives energy and/or nutrients from it, often to the detriment of the host.
poikilohydry Condition of internal water balance varying with changes in ambient humidity.
primary hemiepiphyte Plant that begins its life cycle anchored in a tree crown and ultimately becomes rooted in the ground (e.g., strangler fig).
secondary hemiepiphyte Plant that begins its life cycle as a terrestrial seedling, ascends a tree, and can later lose root connections with the ground, including (a) lianas, woody climbing plants with relatively thick stems that generally grow in mature habitats, and (b) vines, herbaceous climbing plants that regularly grow in disturbed habitats or forest edges.

THE IMPORTANCE OF THE PLANTS THAT DWELL IN FOREST CANOPIES is becoming increasingly recognized in relation to understanding biodiversity. The
upper tree canopy of many forest ecosystems fosters extremely diverse plant communities, which include vascular and nonvascular epiphytes, hemiepiphytes, and parasites. Canopy-dwelling plants contribute substantially to overall forest biodiversity and biocomplexity by providing resources for arboreal vertebrates, invertebrates, and microbes, and by participating in nutrient and water cycling and gas and energy exchanges.

1. INTRODUCTION

A. Definition of the Forest Canopy

The forest canopy has been called “the last biotic frontier” (Erwin, 1988). It presents a habitat conducive to the evolution of literally thousands—perhaps millions—of species of plants, microorganisms, insects, birds, and mammals that are rarely or never encountered on the forest floor. Although forest canopies have been among the most poorly understood regions of our planet, their mysteries are being explored by increasing numbers of biologists. Canopy communities are now believed to be important in maintaining the diversity, resiliency, and functioning of the forests they inhabit.

The forest canopy is a structurally complex and ecologically important subsystem of the forest. It is defined as “the aggregate of all crowns in a stand of vegetation, which is the combination of all foliage, twigs, fine branches, epiphytes as well as the interstices (air) in a forest” (Parker, 1995). The forest canopy is the primary site of gas exchange between the atmosphere and vegetation and fosters many ecosystem processes that are crucial to the maintenance and diversity of the forest as a whole (Lowman and Nadkarni, 1995).

B. Scope of This Article

Trees are the most obvious structural component of forest canopies. Their trunks, branches, and leaves constitute the infrastructure of the canopy and provide mechanical support for thousands of species of arboreal plants and animals. Tree species diversity is discussed elsewhere.

Although much has been published on canopy plants, the question of the global importance of epiphytes for the biodiversity of tropical forests in general and the canopy in particular has not been explicitly addressed. Previous reviews of canopy biodiversity have primarily dealt with arthropods (Erwin, 1988), or with specific subgroups of canopy plants (Kress, 1986; Rhoades, 1995). Here, we describe the diversity of many types of canopy-dwelling plants. Included in this review are epiphytic vascular plants, epiphytic cryptogams (nonvascular plants that include lichens and bryophytes), primary and secondary hemiepiphytic vascular plants (lianas and vines), and arboreal parasitic vascular plants. Arboreal fungi and free-living algae are so poorly known that there is little to review. We place greatest emphasis on obligate and facultative epiphytes, and exclude “accidental epiphytes” from this review.

The term “diversity” in the following will be based on species as the unit of biological diversity, since assessment of other aspects of biodiversity is virtually nonexistent in the case of canopy plants. We first review the systematic distribution of canopy taxa and provide species counts based on the state of current knowledge. We discuss gradients of canopy plant diversity of microsites within the canopy at various spatial scales, spanning a single microsite within a tree (e.g., twig, branch bifurcation) to regional and global levels. Biogeographical analyses of canopy-dwelling taxa are then considered, as well as some of the major evolutionary elements that have influenced their distribution and abundance. “Habit diversity” (the diversity of morphological and physiological features) of arboreal taxa will then be described. Finally, we discuss conservation efforts that involve canopy plants and suggest future research possibilities.

C. Historical Roots and Sources of Information

In 1832, Charles Darwin first described what he termed the great diversity and profusion of “parasitical plants” (that we now understand to have been epiphytes), which he encountered in abundance in the coastal forests of Brazil. In the late nineteenth century, the German botanist A. F. W. Schimper first described epiphytes and outlined their importance to tropical botany.

Historically, canopy studies have been dominated by people who sought the thrill of climbing and followed the lure of discovering new species. Early European explorers hired climbers and trained monkeys to collect specimens of “exotic” air-plants that grew out of reach. Pioneering work in old-growth forests of the Pacific Northwest contributed to the application of mountain-climbing techniques for safe and reliable access to the canopies of tall trees. Since 1980, the innovation of high-strength and low-cost canopy access equipment has made canopy study more viable as an option for scientific research. There are now a wide variety of access tools from which to choose, depending on the
questions being addressed and the available budget (Lowman and Nadkarni, 1995). With the development of effective technological climbing methods such as the "canopy raft" and the canopy crane, and of ground-based methods such as insecticidal fogging, researchers now spend less time working on how to prudently work in the treetops and more time pondering the difficulties in recording meaningful canopy data, analyzing it, and interpreting the results.

A remarkable burgeoning of scientific interest in the canopy has occurred within the last decade. This is related to increasing concerns with such conservation issues as biodiversity, global atmospheric change, and management of tropical rain forests. The number of scientific publications on canopy structure has grown at a disproportionately rapid pace relative to the general field of biology (Nadkarni, 1994). Aspects of the canopy have been the focus of many recent symposia, scientific books, and popular articles and media.

Reviews of vascular epiphyte, hemiepiphyte, and parasite diversity have been compiled (Madison, 1977; Calder and Bernhardt, 1983; Kress, 1986; Benzing, 1990; Putz and Mooney, 1991; Williams-Linna and Lawton, 1995; Lowman and Nadkarni, 1995). The biodiversity of nonvascular plants and lichens has received less attention; only a single (but extensive) review of canopy cryptogams has been published (Rhoades, 1995). The biodiversity of canopy nonlichen fungi has not been well documented (D. Reynolds, pers. comm.).

For this review, we compiled the foregoing sources and searched the primary literature for additions and modifications. We also consulted on-line databases and communicated with numerous taxonomists and specialists to ensure that the information presented is current and accurate. To place canopy plant diversity into the context of biodiversity in a given study area, we compared inventories of epiphytes and nonepiphytes. The epiphyte quotient (sometimes called "epiphytic index") is defined as the percentage of epiphytes out of the total number of vascular plants in an area.

II. CATEGORIES OF CANOPY PLANTS

A. Vascular Epiphytes

Forest canopies support extensive flora that include over 24,000 species, or about 10% of all of the tracheophytes (Kress, 1986). Vascular epiphytes differ greatly in structure, function, and fidelity to their degree of dependence on canopy versus terrestrial habitats. Ecologists recognize their important roles in nutrient cycling and in providing arboreal and terrestrial animals with food, water, and nesting materials (Nadkarni, 1994). Ecophysiologists recognize the varied structures and mechanisms that protect vascular epiphytes from drought (Benzing, 1990).

Some of the characteristics for regular occurrence on bark and associated aerial substrates are obvious (e.g., holdfast roots and wind-dispersed propagules), but others are more subtle. In an extensive review of vascular epiphytism, Benzing (1990) outlined a variety of characteristics that are exhibited by vascular epiphytes (Table I).

Vascular epiphytes are mainly restricted to the low latitudes and within the tropics. They reach their greatest abundance and diversity at low to mid-montane elevations (Madison, 1977; Benzing, 1990). Ferns occur in higher latitudes along the margins of the Pacific, and a few hardy bromeliads and orchids occur in the mild north and south temperate zones (e.g., *Epidendrum rigidum*, *Polypodium polypridoides*, *Tillandsia usneoides*). The most extensively colonized temperate forests are those of southeastern Australia, New Zealand, and Chile, where a variety of vascular epiphytes grow in areas protected from frost by nearby warm ocean currents.

B. Nonvascular Epiphytes

In a recent review by Rhoades (1995), nonvascular (or cryptogamic) epiphytes were categorized into three groups: lichens, bryophytes, and free-living algae. Although the phylogeny and composition of the two plant groups considered here are very different—lichens are symbiotic fungi and algae, and bryophytes are plants—they occupy similar habitats and are often studied together. They have been the focus of little research in the upper canopy (relative to vascular epiphytes), except for a few studies in northwestern North America (Rhoades, 1995), the eastern deciduous forest of North America, and the boreal forest of Canada. Otherwise, most nonvascular epiphyte studies have been restricted to the lower trunks of trees and understory plants.

Bryophytes (phylum Bryophyta) are plants that lack true vascular tissues and organs. In canopy habitats they include the mosses (with about 10,000 species worldwide) and leafy liverworts (leafy hepatics, 7,200 species). Thalloid (strap- or fan-shaped) liverworts and hornworts are usually restricted to moist, lower trunks. (Rhoades, 1995).

Lichens are important components of canopy biodiversity and of ecosystem processes (e.g., nutrient cycling, providing food for wildlife). Lichens are not a
TABLE I

Characteristics of Vascular Epiphytes

1. Reproduction
   A. Pollination                   Exclusively zoophilous, flowers tend to be showy, pollinators highly mobile (Benzing, 1990, Chapter 5)
   B. Breeding systems             Little studied, although many orchids appear to be allogamous
   C. Population structure        Little studied
   D. Seed dispersal               Most families endozoochorous, most species anemochorous (because of the dominance of Orchidaceae)
   E. Life history                 Almost all iteroparous, long-lived perennials

2. Vegetative
   A. Follicle                     Usually evergreen, often succulent, and xeromorphic generally
   B. Habit                        Woody (wet forests) to herbaceous (wet and dry forests)
   C. Shoot architecture           Various
   D. Roots                        Adventitious, specialized for holdfast, often reduced
   E. Special features*            Impounding shoots (e.g., Bromeliaceae) and root masses (e.g., ferns), velamentous roots and absorptive foliar trichomes to prolong contact with precipitation and canopy washes, often lack capacity to grow in earth soil

3. Mineral nutrition
   A. Mycorrhizas                  Possibly significant in Orchidaceae and Ericaceae, probably relatively unimportant elsewhere compared to terrestrial flora
   B. Myrmecotrophy*               Nearly exclusive to epiphytes
   C. Carnivory                    Underrepresented in aboreal flora
   D. Saprotrophy*                 Phytotelm and trash-basket types
   E. Special features*            Tolerance for low pH (wet forests), effective nutrient scavengers (dry forests), frequent reliance on organic substrates for nutrient ions

4. Photosynthesis/water balance
   A. Photosynthetic pathways      CAM overrepresented, no typical C₃ types, much interesting detail probably remains undescribed
   B. Water economy                Often very high
   C. Moisture requirements        Various
   D. Other*                       Much flexibility, e.g., facultative CAM, CAM-C₃ intermediates

* Modified from Benzing (1990) with the permission of Cambridge University Press.
* These characteristics distinguish arborescent from terrestrial flora more than the others.

single taxonomically distinct category, but rather are symbiotic organisms, the association of a fungus (the mycobiont) and a photosynthetic partner (the photobiont), which are usually members of the Chlorophyta and Cyanobacteria. The degree of photobiont specificity varies among lichens. Generally, the mycobiont gives a lichen its overall form and provides the bulk of the biomass, outer protective layer, and a thinner, inner layer that functions in physical absorption and storage of water and nutrients. The photobiont is usually restricted to the layer just below the protective covering of the mycobiont.

These cryptogams are polikydric, that is, they depend on an atmospheric supply of water and inorganic nutrients from precipitation, dew, or fog interception. In general, they absorb water rapidly and lack the water-resistant coverings or cuticles of vascular plants. Bryophyte growth forms have been described and discussed to understand their relationship to water use and conservation (Schofield, 1992; Duration, 1979). The gametophytes (vegetative bodies) of some species form tight cushions or spherical balls that expose a reduced surface area to retain water; others have a pendant, creeping habit that exposes them to maximal amounts of bark surface water. The sporophytes of many bryo-

C. Hemiepiphytes

Hemiepiphytes have been defined as plants that have, at some point in their lives, an "umbilical" connection to the ground. Whether roots or stems, these connections buffer hemiepiphytes from problems of water and nutrient supply that are faced by obligate epiphytes (Williams-Linera and Lawton, 1995). Hemiepiphytes begin their life cycle either as epiphytes and eventually send roots and/or shoots to the ground (primary hemiepiphytes) or as terrestrially established seedlings that sec-
ondarily become epiphytic by severing all connections with the ground (secondary hemiepiphtyes) (Kress, 1986).

Hemiepiphtyes exhibit a tremendous variety in growth form, impact on their hosts, and degree of dependence on hosts. They range from being erect and treelike in form to species that grow in scandroid, clambering heaps. Their impacts on hosts range from lethal (e.g., strangler figs) to benign (e.g., shrubby Ericaceae in tropical cloud forests) (Williams-Linera and Lawton, 1995).

D. Parasites

The mistletoes, which are woody shrubby parasites, are an ecologically distinctive group of canopy-dwelling plants. They have received a great deal of attention from botanists because of their ability to tap into the vascular system of their hosts, as well as from foresters, who have been concerned with reducing the damage they wreak through timber loss and mortality of desirable trees and shrubs (Calder and Bernhardt, 1983). Parasitic mistletoes tend to show a greater tendency for host specificity than do the epiphytes.

III. CANOPY PLANT TAXA DIVERSITY

A. Vascular Epiphytes

Although the global species richness of plants is probably in the region of 270,000, neither their exact number nor their global diversity pattern is known. It has been estimated that possibly as many 24,000 vascular plant species are epiphytes (Kress, 1986), so they constitute a major part of the global biodiversity in the forest canopy.

Vascular epiphytes account for 10% of the total vascular plant diversity. Most extant epiphytes are angiosperms, representing about 9% of all angiosperm species (Table II). Many vascular plant families (84) have adapted to life in the canopy, but relatively few taxa have radiated successfully. Within the angiosperms, approximately 31% of the monocots are epiphytic, whereas only 3% of the dicotyledons occupy the epiphytic niche. The Orchidaceae constitute approximately two-thirds of all epiphyte species (Kress, 1986). Other important monocotyledon families are Bromeliaceae and Araceae. The important canopy-dwelling dicotyledon families are Cactaceae, Ericaceae, Gesneriaceae, Melastomataceae, Moraceae, Piperaceae, and Rubiaceae (Table II; Kress, 1986). There are some large taxonomic groups of plants that contain no epiphytes or very few epiphyte species, for example, the Asteraceae, Leguminaceae, and Poaceae (Benzing, 1987). Less than 1% of the gymnosperms are known to be epiphytic. The Pteridophytes (ferns) are another important group of epiphytic plants, of which 29% are epiphytes (Kress, 1986).

The epiphyte quotient (proportion of an entire flora that is epiphytic) varies widely both geographically and among forest types. This ratio has been measured directly in only a few study sites (Table III). Calculated epiphyte quotients based on published floristic studies in the Neotropics are known for Panama (12%), Peru (10%), Ecuador (22%), Costa Rica (26%), Venezuela (50%), and Florida (3%). In the Paleotropics, epiphyte quotients have been calculated from sites in Java (12%), West Malaysia (9%), Sri Lanka (4%), and Japan (0.5%). African forests seem to be much poorer in relative epi-

<table>
<thead>
<tr>
<th>Major group</th>
<th>Taxonomic category</th>
<th>Number of taxa containing epiphytes in each category</th>
<th>Percentage of taxa containing epiphytes in each category</th>
</tr>
</thead>
<tbody>
<tr>
<td>All vascular plants</td>
<td>Classes</td>
<td>6</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Orders</td>
<td>44</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Families</td>
<td>64</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Genera</td>
<td>876</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Species</td>
<td>23,456</td>
<td>10</td>
</tr>
<tr>
<td>Ferns and allies</td>
<td>Classes</td>
<td>2</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Orders</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Families</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Genera</td>
<td>92</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>Species</td>
<td>2593</td>
<td>29</td>
</tr>
<tr>
<td>Gymnosperms</td>
<td>Classes</td>
<td>2</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Orders</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Families</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Genera</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Species</td>
<td>5</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Angiosperms (dicots)</td>
<td>Subclasses</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Orders</td>
<td>28</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Families</td>
<td>52</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Genera</td>
<td>262</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Species</td>
<td>4251</td>
<td>3</td>
</tr>
<tr>
<td>Angiosperms (monocots)</td>
<td>Subclasses</td>
<td>4</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Orders</td>
<td>9</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Families</td>
<td>17</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Genera</td>
<td>520</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Species</td>
<td>16,608</td>
<td>31</td>
</tr>
</tbody>
</table>

* Modified from Benzing (1990) with the permission of Cambridge University Press.
phyte species richness. In Ghana, a typical epiphyte quotient in forest plots is 8%; one direct measurement in central Africa (Rwanda and Zaire) was 3%.

It is generally regarded that the New World supports greater vascular plant diversity than the Old World. The number of vascular families containing at least one epiphyte species is very similar in the Paleotropics (43) and Neotropics (42). Within the Paleotropics, the representative families do not exhibit a homogeneous distribution. All 43 of the families occur in Australasia, but only 15 are found in Africa and Madagascar. Vascular epiphytes are most diverse in the Neotropics, and less so in tropical Asia and Africa. There is approximately a twofold increase in species diversity in the Neotropics compared with Australasia, and a sixfold increase compared with Africa (Madison, 1977; Gentry and Dodson, 1987). Vascular epiphytes exhibit their greatest diversity in the montane cloud forests of Latin America. The temperate regions support considerably fewer species than tropical areas. Likewise, the temperate regions generally support more vascular epiphytes than do boreal areas.

### B. Nonvascular Epiphytes

In general, bryophytes account for 9–10% of the total species diversity of the plant kingdom. However, no one has calculated how many species of nonvascular plants are obligate epiphytes, as "the idea of an obligate epiphyte is a slippery concept" (D. Griffin, pers. comm.). Rhoades (1995) has ably summarized the results of regional floristic studies (Table IV).

The standard growth forms of lichens are arbitrary, but have often been used to describe functional groups in canopy habitats. “Foliose” refers to leaflike, “fruticose” refers to thalli without distinctive dorsoventral arrangements, and “crustose” refers to thalli firmly cemented to a substrate. According to the International Code of Botanical Nomenclature, lichen species are given the name of their mycobiont; photobiont names are subsidiary. Morphology (sexual structures, asexual structures, and vegetative surface characters) and thallus chemistry are important species characters. Only a few studies have focused on the worldwide biogeography of bryophytes (Schofield, 1992) or lichens (Rhoades, 1995). For many inventories, the crustose lichens have been lacking or incomplete, which is unfortunate as they are the dominant cryptogamic form in outer canopies. The proportion of lichens that grow arboreally is unknown.

The bryophytes are a very old group of plants, perhaps dating as far back in the fossil record as the Devonian period when the first land plants are known to have existed. The combination of a long history and small airborne spores has allowed several bryophyte families to show wide geographic ranges. The tropical regions of Australia and Asia generally have more endemic genera of mosses, whereas in the Neotropics endemic liverwort genera are richer (Schofield, 1992). The cosmopolitan families are not restricted by latitude, but may show local altitudinal variation in some parts of their range. Representative moss families in the tropics include Bryaceae, Dicranaceae, Fissidentaceae, Funariaceae, and Hypnaceae. Pantropical moss families include Calypogeiaceae, Pterorygiaceae, Racopilaceae, and Rhizogoniaceae. Important temperate families are Aulacoma niaceae, Encalyptaceae, Grimmiaceae, and Polytrichaceae. Representative tropical liverwort families include Frullaniaceae, Lejeuneaceae, Lophocoleaceae, Plagiochiaceae, and Radulaeaceae. Species-rich liverwort fami-

<table>
<thead>
<tr>
<th>Study site</th>
<th>Elevation</th>
<th>Precipitation</th>
<th>Number of epiphyte species</th>
<th>Epiphyte quotient (area of reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schoencas, Bolivia</td>
<td>2100–2300 m</td>
<td>5000 mm/year</td>
<td>230 spp.</td>
<td>37% (0.1 ha)</td>
</tr>
<tr>
<td>Otonga, Ecuador</td>
<td>1700–2200 m</td>
<td>2500 mm/year</td>
<td>196 spp.</td>
<td>—</td>
</tr>
<tr>
<td>Rio Guajalto, Ecuador</td>
<td>1800–2200 m</td>
<td>2700 mm/year</td>
<td>166 spp.</td>
<td>28% (400 ha)</td>
</tr>
<tr>
<td>Carbonera, Venezuela</td>
<td>2100–2300 m</td>
<td>1500 mm/year</td>
<td>192 spp.</td>
<td>45% (350 ha)</td>
</tr>
<tr>
<td>Surumoni, Venezuela</td>
<td>100 m</td>
<td>2800 mm/year</td>
<td>53 spp.</td>
<td>6 (fern plot only); 112 spp.</td>
</tr>
<tr>
<td>Location/forest type</td>
<td>Latitude ('N)</td>
<td>Number of trees sampled</td>
<td>Mosses</td>
<td>Liverworts</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>---------------</td>
<td>--------------------------</td>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>Guyana; dry evergreen Eperua spp.</td>
<td>57</td>
<td>11</td>
<td>28</td>
<td>53</td>
</tr>
<tr>
<td>French Guyana; mixed lowland rain forest</td>
<td>5</td>
<td>4</td>
<td>43</td>
<td>61</td>
</tr>
<tr>
<td>Colombia; montane rain forest, 1500 m</td>
<td>5</td>
<td>4</td>
<td>22</td>
<td>36</td>
</tr>
<tr>
<td>Colombia; montane rain forest, 2250 m</td>
<td>5</td>
<td>4</td>
<td>33</td>
<td>102</td>
</tr>
<tr>
<td>Colombia; montane rain forest, 3510 m</td>
<td>5</td>
<td>4</td>
<td>19</td>
<td>63</td>
</tr>
<tr>
<td>Guyana; mixed lowland rain forest</td>
<td>77</td>
<td>5</td>
<td>28</td>
<td>60</td>
</tr>
<tr>
<td>Oregon, United States; low, mixed coniferous forest</td>
<td>44</td>
<td>11</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Wisconsin, United States; mixed conifers and hardwoods</td>
<td>46</td>
<td>Many</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Montana, United States; old-growth Abies</td>
<td>48</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Montana, United States; managed, second-growth Abies</td>
<td>48</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Washington, United States; low-elevation fir forest on lava flow</td>
<td>49</td>
<td>Many</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Sweden; deciduous forest</td>
<td>56</td>
<td>Many</td>
<td>78</td>
<td>17</td>
</tr>
</tbody>
</table>


C. Hemiepiphytes

The phylogenetic distribution of hemiepiphytes suggests that this habit has evolved independently a number of times (Putz and Mooney, 1991; Williams-Linera and Lawton, 1995). Twenty-five families and 59 genera contain hemiepiphytes (Table V), with more than 820 species of primary hemiepiphytes and 650 species of secondary hemiepiphytes. These make up 1% of the total vascular plant species diversity, and 1% of the total canopy-dwelling vascular plant species. This is probably an underestimate, especially for woody hemiepiphytes ("lianas"), which are the most undercollected major canopy plant group. The stranglers most commonly occur in Moraceae and Clusiaceae, but are also found in Araleaceae, Rubiaceae, and Myrtaceae. The hemiepiphytic habit may have arisen from plants growing on rocks.

All of the hemiepiphytic monocotyledonous plants are secondary hemiepiphytes in the families Araceae and Cyclanthaceae. Secondary hemiepiphytes also occur in the dicotyledonous family Marantaceae. Primary hemiepiphytes are represented by 20 families of dicotyledons. The majority of primary hemiepiphyte species are found in the families Araceae, Clusiaceae, and Moraceae. The Moraceae contain the most species of hemiepiphytes, with approximately 500 species in the genus Ficus. Primary hemiepiphytes (whose aerial roots eventually reach the ground) represent about 0.8% of all epiphytes with almost 2000 species (Gentry and Dodson, 1987).

As with vascular epiphytes, the contribution of hemiepiphytes to the diversity of the tropical forest canopy varies among forests. The percentage of trees colonized by hemiepiphytes has been reported for study sites in Venezuela (10% and 13%), Zimbabwe (13%), French Guiana (17%), and the Ivory Coast (21%). In neotropical lowland forests, stranglers and large hemiepiphytes can occur on 10-15% of the trees. Stranglers can occur in much higher densities in some neotropical palm savannas. A rich hemiepiphytic flora is typical of mountain forest and cloud forest sites, but wet lowland forest can also show high percentages of these species (25% in the case of La Selva Biological Station in lowland Costa Rica). In dry forests, hemiepiphytes are usually not present (Williams-Linera and Lawton, 1995).

Woody lianas are distinct features of tropical forests,
### SECONDARY HEMIEPIPHYTES

**Monocotyledonae**

1. Araceae
   - *Amaryllid Schott.*, 4/4 Malaysia
   - *Anthurium Schott.*, 200/550 Neotropics
   - *Caladiopsis Engl.*, 2/2 South America
   - *Eugenia Schott.*, 15/15 Indonesia
   - *Monstera Adans.*, 24/25 Neotropics
   - *Philepallium Schott.*, 1/1 Borneo
   - *Pholidodron Schott.*, 133/275 Neotropics
   - *Pothos Schott.*, 3/3 Neotropics
   - *Rhaphidophora Foster.*, 100/100 Indonesia and Pacific
   - *Syngonium Schott.*, 18/18 Neotropics

2. Cyclanthaceae
   - *Asplundia Harling.*, 20/92 Neotropics
   - *Cardiophora Ruiz & Pav.*, 1/3 Central America
   - *Ludovia Brongn.*, 2/2 South America
   - *Spathenia Harling.*, 7/8 Neotropics
   - *Thoraquercus Harling.*, 1/1 South America

**Dicotyledonae**

3. Marantaceae
   - *Catha Schott.*, 2/2 Venezuela
   - *Marantoid Schott.*, 50/55 Neotropics
   - *Noratea Aubl.*, 20/35 Neotropics
   - *Saurousa Aubl.*, 20/25 Neotropics
   - *Ryschia Jacq.*, 2/10 Neotropics

### PRIMARY HEMIEPIPHYTES

**4. Araliaceae**
   - *Dicyomopanax Decne. & Planch.*, Neotropics
   - *Oreopanax Decne. & Planch.*, Neotropics
   - *Pentapanax Schott.*, 2/15 Java to Formosa
   - *Polyscias J.B. & G. Forst.*, 5/80 Malaysia to New Zealand
   - *Schefflera J.B. & G. Forst.*, 60/200 Neotropics
   - *Sciadophyllum P. Br.*, 5/50 South America and West Indies
   - *Tupinambis Hook.f. & Thoms.*, 1/1 Indonesia

**5. Bignoniaceae**
   - *Schlegelia*

**6. Burseraceae**
   - *Bursera*, 1 Costa Rica

**7. Celastraceae**
   - *Fouquieria L.*, 2/175 Himalayas

**8. Clusiaceae**
   - *Clusia L.*, 85/145 Africa, Madagascar, Neotropics
   - *Clusia Planch. & Triana.*, 3/7 South America
   - *Clusia Planch. & Triana.*, 3/7 South America
   - *Odontoporus Planch. & Triana.*, 1/10 South America
   - *Quaupio Aubl.*, 1/1 South America
   - *Renggeria Meisn.*, 1/1 Brazil

---


rarely or never occurring in temperate forests. Lianas account for approximately 10% of the tropical flora worldwide. They occur in greatest density in Madagascar and Africa, and less so in neotropical and Australian forests. At some sites in Madagascar, an average of 122 lianas with >2.5 cm dbh per 0.1 ha is reported, whereas lianas average only five individuals per 0.1 ha in northeastern temperate forests.

Lianas are more abundant in south temperate forests than in north temperate forests. The important families in north temperate forests are Anacardiaceae, Araliaceae, and Vitaceae. South temperate forests support more than twice as many families. Families like Bignoniaceae, Gesneriaceae, Gramineae, Lardizabalaceae, Saxifragaceae, and Vitaceae account for the majority of the climbing species (Putz and Mooney, 1991).

D. Parasites

About 1400 species of mistletoe occur in forests, woodlands, and shrublands on every continent except Antarctica, with most species in the tropics. Less than 1% of the total vascular plant species are mistletoes, and this group accounts for less than 1% of the total canopy-dwelling vascular plant species.

Mistletoes occur in two plant families. The Loranthaceae contain approximately 900 species in 65 genera, and the Viscaceae contain 400 species in 7 genera. The most species-rich genera in the Viscaceae are Phoradendron (170) and Viscum (100). The Viscaceae contain four genera restricted to the Old World, two genera that occur only in the New World, and one predomi-nately New World genus is also widespread through Eurasia and Africa. In the Viscaceae, the New World genera Dendrophthoe and Phoradendron contain about half of the 397 species of the family.

IV. DIVERSITY OF HABITATS OF CANOPY PLANTS

A. Canopy Microclimate

The values of light intensity and quality, temperature, wind, moisture content, and concentrations of various gases and aerosols are strongly modified by canopy structure in several ways. Canopy surfaces act as passive bodies for the absorption of wind energy, the dissipation of turbulence, and the sorption of heat and radiation. They also actively participate in exchanges of biologically important compounds, such as CO₂ and water vapor, which in turn may have an impact on regional, and even global, climate. Canopies also act as “filters” that remove small-scale turbulence, but allow large eddies to penetrate (Parker, 1995). Canopy structure therefore has a direct effect on the climate surrounding individual leaves, on the modification of microclimate through the layers of the forest, and on the large-scale environment of forest regions.

Canopy conditions are generally typified by more intense sunlight, greater extremes of relative humidity, higher water stress, and a smaller, more pulse-supplied pool of nutrients than on the forest floor. Sunlight attenuation can be as great as 98% between the tops of emergent trees and the levels reaching the forest floor. Rates of evaporation in the canopy have been recorded that are comparable to those occurring in open savannas. Relative humidity can range from nearly 100% at night to less than 30% during midday in the dry season. Differences in canopy versus forest floor wind speeds can also be extreme, especially in tropical cloud forests. In one Costa Rican ridge cloud forest, wind speeds within the canopy (10 m) were clocked at 11.3 m s⁻¹, while forest floor (2 m) speeds were only 4.0 m s⁻¹ (Williams-Linera and Lawton, 1995).

B. Spatial Scales of Canopy Plant Diversity

The forest canopy is a three-dimensional subsystem of the forest itself. Canopy plants need relatively little space in order to develop a striking diversity. In an Ecuadorian montane forest, for example, 109 epiphyte species occurred on just 20 m² of branch surface, compared to only 67 terrestrial plant species on a 100-m² ground plot of elongated shape in the immediate vicinity of the phorophyte. The amazing concentration of epiphytes on single trees has often attracted the attention of naturalists. For example, 66 epiphyte species were found on one specimen of Decussocarpus rospiglio-sit in the Carbonera Forest in Venezuela.

The canopy offers its occupants a wide variation in water, light, and nutrient regimes compared to the understory and the forest floor, and this variety undoubtedly contributes to arboreal plant diversity. For example, soil-like deposits and litter in the canopy function as a medium for canopy-dwelling plants that have well-developed root systems, such as vascular epiphytic shrubs. This material has a high organic content and is derived principally from leaf litter, feces, and other faunal remains. Also present is a small mineral component derived from fine particles carried there by wind, fog, and rain.

In tropical America, epiphytic bromeliads increase
the volume of arboreal soil and litter by creating water-filled tanks ("phytotelmata") in which litter accumulates and soil forms. Although these arboreal epiphytes and their associated soils are patchily distributed, they are linked by climbing vegetation, by percolating rainwater, and probably by the movement of animals, and so they can be likened to a three-dimensional matrix of interconnected islands.

Within a forest, microsite differences exist at many different spatial scales: within a single branch; between branches at different heights of the tree; between trees of different architecture; and within stands of differing topography and aspects. Some studies have described variation in the distribution of canopy plants within single crowns. In a West African rain forest, for example, more than 75% of the orchid species grow on the inner branches, 48% in a middle zone, and only 4% in the outer canopy. Some research has linked differences in microsite water regimes and levels of sunlight input to differential distribution of certain canopy plants between outer and inner crowns; other arboreal species display a "generalist" distribution.

C. Host Tree Specificity

Studies of the mechanisms influencing host tree specificity are scant. It has been suggested that bark texture and pH influence patterns of colonization. For example, the orchid Cymbidium pardalinum obligately grows exclusively on the fern Platycerium madagascarensis, whereas other epiphytes are found on numerous host tree species. This phenomenon warrants further investigation.

V. DIVERSITY OF GROWTH HABITS IN CANOPY PLANTS

Canopy-dwelling plants exhibit a great diversity of ecological adaptations, which is most likely a result of their diverse phylogenetic origins and the possibilities for adaptive specialization in the canopy habitat. The one unifying feature of these mechanically dependent plants is their occurrence in the canopy (Benzing, 1987). For vascular epiphytes, the different "habitats" can be classified by several criteria, including degree of dependence (obligate vs. facultative), nutritional dependence (parasitic vs. commensalistic), degree of light demand (heliofilous vs. sciophytes), architecture (tank vs. atmospheric), substratum (e.g., ant-gardens, humiphiles), or carbon fixation pathway (Crassulacean acid metabolism [CAM] vs. C₃) (Benzing, 1990). Similarly, the great diversity of habits within the hemiepiphytic species ranges from stranglers, which ultimately become freestanding trees, to epiphytes that have only one root connecting with the ground.

A. Diversity of Physiology

The C₃ photosynthetic pathway is more typical of epiphytes inhabiting the canopies of cloud forests and cool, shaded, humid microsites. Forest canopies with more arid conditions favor CAM plants. The occurrence of a C₄ epiphyte has not been documented. Some families containing C₄ plants such as Asteraceae, Cyperaceae, Orchidaceae, and Poaceae also contain epiphytic species, but none is known to exhibit C₄ photosynthesis.

B. Diversity of Modes of Resource Acquisition and Retention

Nutrient acquisition in canopy plants occurs though many modes, including rainwater, bark and leaf leachate, nitrogen-fixing cyanobacteria, airborne particles, carbon nests, crown humus, and decomposition of the host.

1. Humus Epiphytes

Ecologically, humiphily is the most common form of nutrient acquisition and supports the greatest diversity of epiphytes. An overwhelming majority of families containing at least one epiphytic species have at least one humiphile species, and most families contain only humiphile species (Benzing, 1987). Humus-rooted epiphytes include Ericaceae, Gesneriaceae, Melastomataceae, Piperaceae, and Rubiaceae, as well as the hemiepiphytes in Moraceae and Araliaceae. Many of the humus epiphytes are facultative. The obligate species exhibit a diversity of xeromorphic adaptations, including leaf succulence, a flattened pendulous growth form, and poorly developed root systems with strong mycorrhizal associations.

2. Tank and "Trash-Basket" Epiphytes

Some bromeliads, ferns, and orchids collect water, airborne particles, and leaf litter in the rosette created by overlapping fronds or leaves. Ferns in the genus Asplenium, for example, have roots that grow in the form of a trash basket to gain access to nutrients.
3. Ant-Associated Epiphytes
Myrmecophily is a common feature of vascular epiphytes and may be strictly a canopy phenomenon. Nest garden species include Araceae, Bromeliaceae, Cactaceae, Gesneriaceae, Marsdeniaceae, Orchidaceae, Piperaeaceae, and Rubiaceae. In Australia, Asclepiadaceae, Melastomataceae, and Rubiaceae occur in less-studied ant-garden symbioses. Asclepiadaceae, Bromeliaceae, Melastomataceae, Orchidaceae, Polypodiaceae, and Rubiaceae contain specialized ant-garden species. Rubiaceae appear to be the most specialized ant-garden species. Some ferns and orchids have hollow rhizomes (Polypodiaceae), hollow tubers (Solanopteris spp.), and hollow pseudobulbs (Schomburghia and Laelia) that provide domatia for ants.

4. Bark Epiphytes
To inhabit the bark substrate, epiphytic plants must cope with very low levels of water and nutrient availability. Therefore, many of the bark epiphytes are obligate epiphytes, including many specialized orchids.

5. Atmospheric Epiphytes
Some bromeliads (e.g., Tillandsia) have special hairs (trichomes) that allow them to absorb water from the atmosphere over the entire surface of their leaves.

VI. CANOPY PLANT BIODIVERSITY AND CONSERVATION BIOLOGY

A. General Considerations of Canopy Plant Conservation
Because of their small size, high degree of endemism, and frequent microsite specificity, epiphytes may be more vulnerable to human-induced disturbance than terrestrial plants. Methods to conserve existing epiphyte populations and floras have been discussed (e.g., Lowman and Nadkarni, 1995). Studies have shown the value of older trees in forests as habitats for certain sensitive species.

B. Effects of Forest Fragmentation and Habitat Conversion
The effects of forest fragmentation and habitat conversion on canopy plant diversity are poorly documented, especially in the tropics. It is generally accepted by researchers that secondary bryophyte and lichen communities are very different from those in primary forests (Gradstein et al., 1989). Most studies indicate a decrease in species richness between secondary habitats and primary forests, and even disturbances at small spatial scales (within a branch) are reported to result in a decrease in diversity.

Shade epiphytes growing in the understory are more affected by habitat conversion than the sun epiphytes of the canopy, but not all sun epiphytes are able to recolonize following disturbance. The available data from investigations of the regeneration rates of temperate and subtropical canopy plants indicate that many species are slow to recover. The rates for bryophytes range from 25 years in Australia to 80–100 years in California. In Britain, it is estimated that lichens may require 500 years to successfully regenerate.

C. Effects of Global Environmental Change
Water stress is a major limiting factor for plants inhabiting the crowns of trees. A rise in global temperatures may have an impact locally on the relative humidity of some forest canopies. Preliminary experimental work along an altitudinal gradient indicates that the species composition of canopy plant communities may be altered by such changes in temperature and humidity.

D. Vulnerability of Canopy Plants to Extinction and Invasion
There are no records of a specific canopy plant extinction in modern times. However, numerous endemic species are endangered or threatened by habitat conversion (Gradstein et al., 1989). A great deal of information indicating that lichens are very susceptible to air pollution and metal ion deposition has accumulated over the past several decades (Rhoades, 1995). Gradstein et al. (1989) suggest that relatively small reserves containing a diversity of life zones should suffice to conserve cryptogam biodiversity if the reserve is large enough to maintain a viable population of host trees. However, they warn that these recommendations are based on very preliminary data and more inventory data and taxonomic work are needed to better define species ranges and to determine which species are locally rare or endemic. Mistletoes are generally very susceptible to environmental changes (Calder and Bernhardt, 1983).
VII. AREAS FOR FURTHER STUDY

A. General Considerations

Habitat loss and climate change are growing threats to plant communities. Arboreal plants provide many opportunities and challenges for biologists from many disciplines, and because there plants have no access or sporadic access to terrestrial soil, they make excellent experimental subjects to study physiology and stress. Canopy plants warrant attention for the roles they play in forest dynamics, which affect biodiversity, productivity, and nutrient cycling. A list of research questions was created for vascular epiphytes (Table VI); these questions can also be related to the study of other types of canopy plants.

B. Monographs and Inventories

There is a pressing need for extensive and intensive work on plants that live in the canopy. However, the lack of resident tropical taxonomists is a serious concern. There are many more taxonomists in the more developed countries where the resources and infrastructure exist to train students, but there are relatively few specialists in less developed countries where many of the biological resources exist.

However, efforts to create monographs and inventories of canopy-dwelling plants have been increasing. For example, botanists at the Missouri Botanical Gardens and their collaborators have compiled inventories of regional floras. To date, Peru, Panama, Venezuela, and the Guianas have received a great deal of attention, and the study of other floras (e.g., Nicaragua and China) is planned. To fully represent regional biodiversity, it is crucial that botanists collect plants in the canopy. Likewise, to fully understand global biodiversity, the generally undercollected groups must be collected in the canopy (e.g., lianas and cryptogams). In terms of cryptogams, work needs to continue on broad regional inventories of all tropical species and of crustose lichens worldwide. The bryoflora of Australia is particularly poorly known.

C. Herbaria and Databases

Certain herbaria have significant canopy plant collections (Madison, 1977; Kress, 1986). Herbarium studies have been conducted at the Harvard University Herbaria, the Marie Selby Botanical Gardens, the Huntington Botanical Gardens, the Herbario Nacional Colombiano, the State University of Utrecht herbarium, and the University of Florida.

Several major botanical gardens have produced useful databases. For example, The Missouri Botanical Gardens has developed an on-line database (TROPICOS) for the floras of several Neotropical countries. Such databases can be of great use to systematic botanists and conservationists by providing the most up-to-date information available. With the increasing attention being paid to canopy plant ecology (e.g., Lowman and Nadkarni, 1995), perhaps it is time to initiate a canopy plant biodiversity database.

D. Experimental Fieldwork

Experimental field studies to investigate the potential effects of forest harvesting on plant community composition and species richness should be conducted, especially in tropical regions, so as to include biodiversity objectives in forest management practices. Humus epiphyte communities growing in bryophyte mats, for example, are ideal for experimental fieldwork because entire moss mat communities are easily transplanted with minimal disturbance to the rooting medium. Transplanting epiphytes along an altitudinal gradient is useful in helping to predict the effects of environmental change. The experimental removal of bryophyte mats is also useful in monitoring epiphyte succession and recolonization.

E. Ethnobotany of Canopy Plants

Ethnobotanical knowledge and usage of canopy plants is widespread in cultures around the world (Lowman and Nadkarni, 1995). However, the pharmaceutical potential of canopy plants has only begun to be investigated. In particular, the bryophytes and lichens have not been rigorously explored in this regard. The potential importance of canopy plants for human use may spark resources needed to learn more about both applied and basic aspects of these diverse organisms.

See Also the Following Articles

FOREST CANOPIES, ANIMAL DIVERSITY • PLANT BIODIVERSITY, OVERVIEW • TROPICAL ECOSYSTEMS

Acknowledgments

This summary of canopy plant diversity was supported by the National Science Foundation. We thank the Monteverde Cloud Forest Reserve and the Tropical Science Center in San José, Costa Rica for main-
<table>
<thead>
<tr>
<th>Subject</th>
<th>Obvious</th>
<th>Questions remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Fidelity to canopy versus other substrates</td>
<td>Occurrence on trees ranges from accidental to obligate.</td>
<td>What factors differentiate canopy from terrestrial substrates for the obligate epiphyte? How has specialization for arboreal life compromised capacity to survive on the ground?</td>
</tr>
<tr>
<td>2. Requirements for specific types of arboreal substrates</td>
<td>Specific epiphytes typically colonise only subsets of the many types of substrates present in occupied tree crowns.</td>
<td>What plant characteristics determine microsite requirements for twig, bark, buds, ant-nest garden, etc., epiphytes?</td>
</tr>
<tr>
<td>3. Plant adjustments to the often transitory and relatively unpredictable supplies of moisture in forest canopies</td>
<td>Broadly occurring accommodations to drought (e.g., CAM, seromorphy) are particularly well developed among the epiphytes.</td>
<td>What is the nature of the moisture supply in forest canopies and how are mechanisms such as photosynthetic pathways, osmotic balance, and stomatal behavior fine-tuned to reduce risk and maximize effective use of available moisture?</td>
</tr>
<tr>
<td>4. Plant adjustments to the absence of mineral soil</td>
<td>A variety of organic substrates, including the products of mutualistic bions, serve in lieu of earth soil as sources of nutritive ions.</td>
<td>How is impounded litter processed for phytodet epiphytes? How substantially do ant mutualists contribute to the nutrient budgets of associated epiphytes? How are the more oligotrophic epiphytes (e.g., atmospheric bionslads) equipped to scavenge scarce ions and use them economically?</td>
</tr>
<tr>
<td>5. Impacts of arboreal ants</td>
<td>Some epiphytes require ants for dispersal and to provide rooting media.</td>
<td>How much arboreal flora beyond the obvious ant-nest garden and myrmecotrophic species are dependent on ants for dispersal, substrates, and defense?</td>
</tr>
<tr>
<td>6. Epiphytic vegetation as a resource for canopy fauna</td>
<td>Much arboreal fauna, particularly invertebrates, use epiphytes as resources.</td>
<td>What is the full extent of this dependence and what are the broader consequences of these dependencies for the forest community?</td>
</tr>
<tr>
<td>7. Epiphyte involvements in nutrient cycles</td>
<td>Nutritional piracy exists. Epiphyte biomass sometimes contains much of the nutrient capital present in a forest ecosystem.</td>
<td>To what degree and under what conditions does the presence of an epiphyte load have an impact on the nutritional status of a phorophyte?</td>
</tr>
<tr>
<td>8. Impacts on community productivity and patterns of resource use</td>
<td>Resources present in epiphyte biomass (e.g., N and P) at least sometimes yield photosynthetic returns at different rates than those of supporting soil-rooted vegetation.</td>
<td>How does the presence of substantial epiphyte biomass affect aggregate forest productivity and help determine overall resource-use efficiency?</td>
</tr>
<tr>
<td>9. Conservation</td>
<td>Because many epiphytes occupy narrow ranges (especially orchids), often in regions of rapid development, endangered status is correspondingly common.</td>
<td>What conservation strategies are likely to preserve the greatest diversity of epiphytes?</td>
</tr>
<tr>
<td>10. Indicators of habitat quality and global change</td>
<td>Some epiphytes possess characteristics that impart extraordinary utility as air quality monitors.</td>
<td>How can epiphytic vegetation be more effectively used to monitor changing conditions in the troposphere?</td>
</tr>
<tr>
<td>11. Succession</td>
<td>Presumed seed stages identified.</td>
<td>Do species displace one another on bark? If so, by what mechanisms?</td>
</tr>
<tr>
<td>12. Community organization</td>
<td>Species often co-occur in predictable assemblages, but often distribution and spacing among individuals are random.</td>
<td>Are the factors responsible for the distributions and combinations of species on bark primarily density dependent or density independent?</td>
</tr>
</tbody>
</table>

*Modified from Benzing (1990).*
taining and protecting the Research Area. Funds were provided by
Evergreen State College Faculty Sponsored Research, the Na-
tional Science Foundation [Ecology Program (BSR 86-14925 and
89-18006), Long-Term Studies Program (DEB 90-15341), Data-
base Activities Program (BIR 96-30316)], the Whitewall Foun-
dation, the National Geographic Society, and the Evergreen
Foundation. Many plant tax-
omonomists provided information on epiphytic floristics. We thank
A. S. Taylor (Zamiaceae), T. Croat (Araceae), R. Faden (Commelin-
aceae), H. Luther and B. Hols (Bromeliaceae), H. Kennedy (Marant-
aceae), L. Thein (Winteraceae), P. Berry and P. Hoch (Onagraceae),
P. Lowry and D. Frodin (Araliaceae), M. Endress (Apocynaceae), and
W. J. Kress (for many families).

Bibliography

and adaptive diversity. Ann. Missouri Botanical Garden 74,
183–204.

Press, Cambridge, United Kingdom.

Academic Press, Sydney, Australia.

review. Lindbergia 5, 2–18.

diversity. In Biodiversity (E. O. Wilson, ed.), pp. 123–129. Stan-
ford University Press, Stanford, California.

74, 205–223.

richness and origin of the bryophyte flor of the Colombian Andes.

lichens in the Vancouver Island temperate rain forest. Vegeta-
ta 84, 99–112.

Kelley, D. L. (1985). Epiphytes and climbers of a Jamaican rain forest:

Kress, W. J. (1986). The systematic distribution of vascular epiphytes:
An update. Selbyana 9, 2–22.

Lowman, M. D., and N. M. Nadkarni. (1995). Forest Canopies. Aca-
demic Press, San Diego, California.


Nadkarni, N. M. (1994). Diversity of species and interactions in the
upper tree canopy of forest ecosystems. Amer. Zool. 34, 70–78.

In Forest Canopies (M. D. Lowman and N. M. Nadkarni, eds.),
Academic Press, San Diego, California.

University Press, Cambridge, United Kingdom.

Rhoades, F. M. (1995). Nonvascular epiphytes in forest canopies:
Worldwide distribution, abundance, and ecological roles. In For-
est Canopies (M. D. Lowman and N. M. Nadkarni, eds.), pp.

Schofield, W. B. (1992). Bryophyte distribution patterns. In Bryo-
phytes and Lichens in a Changing Environment (J. W. Bates and
A. M. Farmer, eds.). Oxford University Press, Oxford, United
Kingdom.

epiphytes in forest canopies. In Forest Canopies (M. D. Lowman
Diego, California.
I. Introduction

II. Biodiversity in Forest Ecosystems

III. Human Activity and Forest Biodiversity

IV. Biodiversity and Forest Ecosystem Function

GLOSSARY

disturbance  An event that results in a radical change in environment, usually as a consequence of death of the dominant plants.

food web  A set of species that live together and a specification of which species “eat” which other species.

forest stand  A relatively homogeneous forest landscape unit that can be distinguished from neighboring units by forest age or composition.

forest structure  The arrangement of all the parts of the forest stand—stems, branches, leaves, roots, and so on.

keystone species  A species that performs some crucial activity in the ecosystem and for which there is little or no redundancy.

succession  Sequential change in the relative dominance of various species in the biotic community at a location.

A forest is an ecological system (or biotic community) dominated by trees. Ecology is the study of ecological systems and the mutual interactions between organisms and the environment. The primary objective of forest ecology is to understand what controls the patterns of distribution and abundance of different organisms in forests of the world. The relevance of forest ecology to understanding biodiversity lies in the observations that forests dominate the natural landscape over much of the world and that forests harbor a large proportion of the world’s species. Moreover, because forests are valuable to humans for the products and services that they provide—especially wood—many forest ecosystems are extensively manipulated and modified by human societies, often disrupting the natural ecological patterns. In addition, many forests occur where climate and soil are suitable for intensive agricultural production and conversion of forests to farms results in radical changes in the composition of the biota. Hence, detailed knowledge of forest ecology is necessary to predict and ameliorate the effects of forest utilization and conversion on biodiversity.

I. INTRODUCTION

Forests occur naturally wherever environmental factors permit the upright woody growth form of trees to be a successful strategy. The primary advantage that trees realize in such settings is the ability to gain a competitive edge over any smaller neighbors in gaining access to solar radiation for photosynthesis. Thus, forests dominate the natural landscape in all regions except where the upright woody growth form is not possible or where other mitigating environmental factors eliminate any
Ecologists have classified the forests and other vegetation of the world into biological regions, or biomes, on the basis of the physiognomy (outward form) and species composition of the dominant plants. The biome classification represents ecological patterns at a very large scale. However, the patterns that ecologists endeavor to explain occur at a wide range of spatial scales, from global to local. Global scale maps of the world’s biomes have been constructed to illustrate the broad distributional patterns of the biota. Ecological patterns at smaller scales are represented by subdividing the various biomes to provide more detailed classifications and maps. In essence, the forest ecologist regards the global ecosystem as consisting of a nested series of ecological associations, at smaller and smaller spatial scales, down to the level of the relatively uniform forest stand.

The focus of ecological classification systems on the dominant plants is not just a convenience; the dominant plants in terrestrial ecosystems (e.g., the trees in the forest) play a crucial role in setting the stage for all the other organisms because they convert solar energy into food for the food web and they form the three-dimensional structure that constitutes the habitat of most associated biota. Hence, it is of primary importance in studying the ecology of forests to understand what controls the distribution and abundance of trees. However, this is not to say that the other organisms are not important to the functioning of the ecosystem; in fact, many of the less prominent organisms play important roles in regulating the distribution and abundance of the trees.

Underlying these ecological classification systems and maps at the broad scale is the principle that climate is the environmental feature that exerts primary control over the distribution of organisms on the earth’s surface. This principle was realized long ago by European geographers who observed during their explorations the coincidence between the broad patterns of climate and vegetation physiognomy on earth. The key components of climate, temperature and precipitation, exert this control through their effects on the growth of plants. Different plant traits, expressed in part as whole plant physiognomy, prove to be most suitable for growth and survival under different combinations of temperature and precipitation. Hence, climatic patterns largely coincide at the broad scale with the patterns of distribution of forest biomes. Expressed most simply, the effect of climate on vegetation distribution can be plotted in terms of mean annual temperature and precipitation (Fig. 1); forest vegetation is restricted to relatively moist climates. However, annual averages cannot account for the important effects of climatic seasonality—variations in temperature and precipitation through the year—in determining biome distributions. More complicated systems of expressing the influence of climate account for the important effects of seasonal drought and of subfreezing temperatures on plant growth and activity.

The principle that climate exerts primary control over the distribution of organisms fails to account for the observation that some equivalent climates support vegetation with differing composition and physiognomy. For example, the temperate rain forests of the Pacific Northwest of North America are dominated by evergreen needleleaf trees whereas in some equivalent climates of the southern hemisphere (New Zealand and South America) evergreen broadleaf forests may dominate. These differences reflect the fact that the flora and fauna that have developed by biological evolution in the various regions of the world are quite distinct because they have been geographically isolated from one another. Several biogeographic zones have been identified by botanists and zoologists, and these zones reflect geologic history, especially continental drift. For exam-
ple, plant geographers typically recognize four plant domains whose floras are distinct from one another: (a) paleotropical (Old World Tropics), (b) neotropical (New World Tropics), (c) north temperate, and (d) south temperate. To expand on our earlier principle, then, climate together with the available flora and fauna exert primary control over the distribution and abundance of organisms on earth. This principle points toward the likelihood that human effects on climate and the introduction of exotic species across the globe will have profound consequences for ecology and biodiversity.

At a more localized scale, species distribution and abundance are also affected by secondary environmental factors, topography and soils. Topography exerts its influence in part by locally modifying climate; for example, in the northern hemisphere south-facing slopes are warmer and drier than north-facing slopes. Also, topography influences the environment through the action of gravity moving matter downhill (e.g., water, soil particles). Soil supplies the essential environmental resources, water and mineral nutrients, for tree growth. Soil properties reflect the combined effects of geologic, climatic, and biologic forces, emphasizing the complex web of cause and effect that underlies ecological patterns and processes.

Geologists have divided the earth’s landscape into physiographic provinces that reflect the effects of geological processes on the earth’s surface features. Within each physiographic province, relatively orderly and recurring patterns of topography and soils are observed that differ fundamentally from those of neighboring provinces. For example, in the coastal plain province of the eastern United States, sandy sediments have emerged from the receding ocean to leave a gently rolling landscape of porous, sandy soils. Forest patterns reflect subtle variations in topography and drainage. In contrast, directly to the west, the ridge-and-valley province is marked by recurring combinations of bold narrow ridges of resistant rocks and intervening valleys underlain by softer substrates like limestone and shale. The forest patterns in this province directly mirror the bedrock geology of the province. An understanding of the ecology of these two regions begins with the recognition of the underlying differences in geological forces that have shaped their physiography.

One additional factor that strongly regulates the composition of the biotic community is the legacy of disturbance events. Disturbance is a natural phenomenon in all ecosystems and is defined as any event that results in a change in environmental conditions and resource availability, usually as a consequence of death of the dominant plants. Among the most prominent natural disturbance agents in forest ecosystems are fire, windstorms, and irruptions of pests and pathogens. When disturbances occur at a spatial scale that is much larger than the area occupied by individual dominant plants (i.e., large-scale disturbances), they so profoundly alter the environment that the suite of plants that subsequently colonizes the disturbed site may be quite different from the original community. These large-scale disturbances initiate the ecological process of succession—that is, successive changes in the composition of the biotic community occurs as a result of progressive changes in environmental factors. Hence, the actual composition of the biotic community at any particular time and place depends not only on climate, flora and fauna, topography and soils, but also on the time interval since the last large-scale disturbance that initiated successional change—as well as the nature and intensity of that disturbance event. That disturbance has played a major role in shaping ecological patterns and processes is evident from the many traits of the flora and fauna that reflect the selective force associated with disturbance.

The natural disturbance regime that characterizes any particular ecosystem depends on exogenous factors that act as disturbance agents (e.g., the combination of drought, lightning, and wind that favors the occurrence of fire; or frequent exposure to hurricanes or tornadoes in certain geographic regions) as well as endogenous factors, such as the traits of the plants themselves, that influence the frequency or intensity of the disturbance. For example, pine forests are more prone to fire disturbance than deciduous broadleaf forests in part because the fuels produced by pine trees are more flammable. And forests on sandy soils are more prone to fire because coarse soils dry out more rapidly than fine-textured soils. Thus, the influences of disturbances, environment, and biota may be mutually reinforcing in shaping ecological patterns and processes.

The prevalence of disturbance and succession in natural forest ecosystems complicates the task of classifying and mapping forest distributions, particularly at smaller scales, because forest composition is continually changing. Recognizing this problem, ecologists have defined the climax forest association as the assemblage of species that would persist under any particular combination of environmental factors in the absence of large-scale disturbance. In many regions, however, the recurrence of large-scale disturbances is naturally so frequent that climax forest associations rarely develop,
and the practical value of such a strict climax concept is somewhat limited. However, in some regions where large-scale disturbances are infrequent, particularly under humid climates (where fires are rare) and areas not often exposed to intensive windstorms (i.e., the humid tropics and some temperate areas), climax forest associations probably were common prior to the advent of anthropogenic influences. Natural disturbance regimes in these forests consisted of small-scale events resulting from the death of individuals or small groups of trees in a patchwork mosaic. Although the composition and structure of various patches would differ, at a larger scale the average composition of the forest ecosystem would remain relatively steady (i.e., it would exhibit a shifting-mosaic steady-state). These observations illustrate the importance of the scale of observation to our understanding of ecological patterns and processes.

II. BIODIVERSITY IN FOREST ECOSYSTEMS

A. General Patterns

The number of species observed in any particular forest varies markedly across the earth's forest biomes. Biodiversity can be conceived as consisting of three distinct elements, termed the gamma, alpha, and beta diversity. The total diversity in a large area, the gamma diversity, can be partitioned into two components, the local (alpha) diversity in a single habitat or forest stand, and the turnover of species between stands, the beta diversity. High gamma diversity could be associated either with high alpha diversity, high beta diversity, or both. Ideally, considerations of the patterns of biodiversity across forest regions need to account for the contributions of these different elements of biodiversity. Unfortunately, studies of forest biodiversity patterns generally have not provided samples of sufficient detail to resolve these elements, and more systematic approaches are needed. Nevertheless, many valuable insights into the patterns and causes of biodiversity variation across forest regions have emerged from research to date.

The most striking pattern of forest biodiversity is the latitudinal gradient: much higher biodiversity is observed both locally and regionally in tropical than in temperate forests. Because observations are most comprehensive for tree species, the following description focuses on these taxa.

The general relationship between latitude and the alpha diversity of trees is illustrated by a plot of the number of species in 0.1 ha samples taken from around the world (Fig. 2). At one extreme are lowland tropical rain forests in the upper Amazon basin where on average every second tree in a sample belongs to a different species; at the other extreme, a variety of natural, monospecific forests are common in many temperate and boreal regions. Within the tropical region considerable variation in alpha diversity has been observed (Fig. 2). The two most prominent features correlated with this variation are annual precipitation and biogeographic province or region. For example, the average alpha diversity (number of tree species in 0.1 ha samples) of neotropical lowland wet and moist forests is 152 species, whereas for seasonally dry forests the average is 65. And the alpha diversity of trees in Africa is generally much lower than for eastern Asia and America (Fig. 3). The extremely high alpha diversity of lowland tropical rain forest is not paired with equally high beta diversity. For example, the 307 tree species identified on a single 1-ha plot in the Ecuadorian Amazon constituted about 10% of the total tree flora (trees ≥ 5 cm diameter) of Amazonian Ecuador and a single 50-ha plot at Pasoh, Malaysia contained 830 species, 20 to 30% of the total tree flora of this country.

By comparison, tree species diversity in the temperate latitudes is much lower: the enormous area of temperate zone forests harbors only 1166 tree species, not much more than the 50 ha plot at Pasoh! As for the tropics, however, striking geographic differences in the diversity of tree taxa are observed in the temperate zone. The highest gamma diversity is in east-central Asia,

FIGURE 2 Speciee richness of 0.1 ha samples of lowland (<1000 m) forest as a function of latitude. Dashed line separates dry forest (bottom) from moist and wet forest (top) with intermediate sites (most forest physiognomy despite relatively strong dry season) indicated by alternate lines. X, the anomalous Coloso and Lima de los Colorados sites in northern Colombia. Reprinted, with permission, from Gentry (1995).
A comprehensive explanation of these global patterns in tree species diversity remains elusive, but a growing consensus on the role of historical or biogeographic factors and local physical habitat factors is emerging. The overall lower diversity of temperate than tropical regions probably is explained in part by the physiological constraints on colonization imposed by the need to tolerate subfreezing conditions and by the smaller contiguous area in temperate regions. Regional differences between biogeographic provinces in the temperate zone appear to owe in large part to higher extinction rates in Europe and North America during the Pleistocene glacial epochs as well as greater access of the east Asian region to dispersal routes from the tropics. Similarly, the combination of geographic area and climatic differences probably explains the contrasts in diversity between tropical Africa versus America and east Asia and between wet and dry forests. As noted by Latham and Ricklefs (1993), "further resolution of the causes of diversity patterns will require new paleontological, biogeographical and taxonomic data and synthesis." (p. 310)

**B. Forest Structure and Pattern and Disturbance Regimes**

The local biodiversity in particular forests depends on the complex suite of factors that characterize the habitats of individual species. These factors include such components as the species composition, phenological timing, structural complexity, and horizontal patterning of the vegetation, which in turn depend on environment and the legacy of disturbances. The dominant plants (i.e., trees in forests) play a pivotal role in defining the habitats for associated organisms by providing food and shelter and by regulating the local microenvironment. For some organisms (e.g., many insects) host-specific interactions with particular dominant plants result in strong correlations in their distributions, so that forest composition is the key factor influencing the composition of associated species. For other organisms, the structural and horizontal patterning of the dominant vegetation may be more important than composition alone.

The importance of the three-dimensional spatial arrangement of the branches and leaves of the plants in defining animal habitats was realized by pioneers in the study of ecology and evolution. MacArthur and MacArthur (1961) demonstrated that the diversity of birds in forests of the eastern United States could be...
TABLE I
Summary by Taxonomic Level and Region of Moist Temperate Forest Trees in the Northern Hemisphere

<table>
<thead>
<tr>
<th>Taxonomic level</th>
<th>Subclasses</th>
<th>Orders</th>
<th>Families</th>
<th>Genera</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern, central, and eastern Europe</td>
<td>5</td>
<td>16</td>
<td>21</td>
<td>43</td>
<td>124</td>
</tr>
<tr>
<td>Asia</td>
<td>9</td>
<td>37</td>
<td>67</td>
<td>177</td>
<td>729</td>
</tr>
<tr>
<td>North America</td>
<td>6</td>
<td>14</td>
<td>19</td>
<td>37</td>
<td>88</td>
</tr>
<tr>
<td>Eastern North America</td>
<td>6</td>
<td>26</td>
<td>46</td>
<td>90</td>
<td>253</td>
</tr>
<tr>
<td>Northern Hemisphere (total)</td>
<td>10</td>
<td>39</td>
<td>74</td>
<td>213</td>
<td>1,166</td>
</tr>
</tbody>
</table>

Families excluding those of predominantly tropical distribution (% of total)
- Europe: 18 (86%)
- Asia: 37 (55%)
- North America: 14 (95%)
- Eastern North America: 29 (63%)
- Northern Hemisphere (total): 41 (55%)

Genera excluding those of predominantly tropical distribution (% of total)
- Europe: 41 (95%)
- Asia: 121 (68%)
- North America: 35 (95%)
- Eastern North America: 77 (86%)
- Northern Hemisphere (total): 149 (70%)

Species excluding of predominantly tropical genera (% of total)
- Europe: 122 (98%)
- Asia: 570 (78%)
- North America: 66 (97%)
- Eastern North America: 236 (93%)
- Northern Hemisphere (total): 987 (85%)

Data from Latham and Ricklefs, 1993.

predicted by the structural complexity of the vegetation (Fig. 4): in habitats with high foliage-height diversity (FHD; defined by the formula $FHD = \% \cdot p_i \ln p_i$, where $p_i$ is the proportion of total foliage area in the $i$th layer), bird diversity was much higher than in habitats with low FHD. Moreover, this relationship largely transcended differences in plant species diversity. Similarly, for ground-dwelling organisms, the complexity of habitat at the soil surface—including vegetation cover, litter, rocks, fallen logs, and moisture—is strongly correlated with local biodiversity. By extension the belowground structural complexity might influence diversity of subterranean organisms, but little work on this topic has been accomplished.

The horizontal pattern of forest vegetation, particularly the arrangement of relatively uniform forest stand units across the landscape, influences biodiversity of the region and is the subject of a newly emerging branch of ecology—landscape ecology. The edges or ecotones (zones of rapid change in plant species composition) between forest stand units or forest associations often provide qualitatively different habitat than the interiors of those units; hence, the size, shape, and spatial arrangement of forest stands in the landscape influence the populations of associated species. Edges and sharp ecotones between stands may arise because of environmental discontinuities (e.g., topographic depressions where soil water accumulates) or because of the legacy of past disturbances. Thus, the local biodiversity of forests reflects both the compositional and structural diversity of the plants as well as the arrangement of units of relatively uniform composition and structure that we define as forest stands.

The structure of any particular forest stand traditionally has been defined as the distribution among age or size classes of the trees in a forest stand. Even-aged stands arise as the result of large-scale disturbances in which all or most of the large trees in an area are killed by a natural disturbance agent (e.g., crown fire, hurricane, fungal pathogen) or by human activity. The
forest that arises on such a site consists of trees of roughly the same age, from new seedlings that colonize the site or from advanced regeneration (previous seedlings and saplings that escaped the disturbance). The process of stand development following large-scale disturbance results in gradual changes in the structure of the forest, as the growth, mortality, and recruitment of new individuals proceeds. These changes in forest structure include not only age structure, but also the spatial arrangement of the stems, branches, foliage, and roots of the plants that define the habitat of other biota in the forest. Moreover, changes in species composition of the forest usually accompany stand development, as species capable of growing in the shaded understory replace the pioneer species that colonize the disturbed site.

A typical sequence of forest stand development following large-scale disturbance has been characterized by Oliver and Larson (1990) as consisting of four stages: (stage 1) stand initiation, (stage 2) stem exclusion, (stage 3) understory reinitiation, and (stage 4) old growth. In the stand initiation stage, trees colonize the disturbed area. The time interval of this stage varies markedly depending on the severity of the disturbance, environmental factors at the site (e.g., climate and soils), and often herbivory. This stage concludes when the forest canopy becomes closed or when some soil resource (often water) becomes limiting to further increases in the leaf area of the forest. During the stem exclusion stage resources like light and soil water are so limited that suppressed understory trees die and regeneration is severely restricted. In this stage there is usually a continual reduction in the density (number of stems/area) of the original cohort of trees. The structure of the canopy is exceptionally simple at this stage as the individual trees grow in height to co-opt the light resource from neighbors.

In the understory reinitiation stage the overstory begins to break up as canopy trees die and the differential height growth of various species or individuals results in more complex arrangements of the foliage. Increased light reaching the understory favors the establishment and growth of new cohorts of species most capable of surviving in the highly competitive understory environment. The old growth or late successional forest stage is attained as overstory trees age and the canopy develops even greater complexity of structure. Gaps form in the canopy as a result of injury or death of the large, mature individuals, and the previously suppressed individuals are released from severe competition and grow in height. Decaying, coarse, woody debris accumulates on the ground and dead tree snags also provide new habitats for animals. Obviously, this idealized model of forest stand dynamics exhibits myriad local variations depending on the nature of the disturbance, environmental factors, and the tree species that dominate the area.

The changes in forest structure (e.g., foliage-height distributions) that accompany stand development following large-scale disturbance result in consequent gradual shifts in the quality of the habitat for different animals and plants. In the example cited earlier, MacArthur and MacArthur (1961) observed that maintenance of high numbers of bird species in eastern deciduous forests of the United States depended on the adequate provision of three layers of foliage, corresponding roughly to ground vegetation (0–2 ft.), shrubs and small trees (2–25 ft.), and overstory trees (>25 ft.). If nearly all the foliage of the forest is in just one layer, as in the stand initiation and stem exclusion stages of stand development, bird diversity is much lower than in the old-growth stage, when canopy stratification becomes prominent.

The nature and degree of vertical stratification differ among forests. For example, temperate deciduous forests in the mature stage often exhibit the three strata just identified: an overstory stratum occupied by the canopy trees, an intermediate stratum represented by the crowns of saplings and understory species like dogwoods and hornbeams, and a ground stratum of low shrubs and forbs. In contrast, in the lowland tropical rain forest, very tall ‘emergent’ trees with broad shallow crowns overtop the main canopy, which may be subdivided into two or more additional strata above the understory layers. And in the boreal forest the low stature of the crowns of the conifers often preclude the formation of strong vertical stratification. These structural differences contribute to the contrasts in the number of distinct habitats provided for associated biota.

As noted earlier, many forest regions are only rarely affected by large-scale disturbances because they are both too moist to carry fires and not subjected to catastrophe. In these regions the steady-state forest is characterized as a shifting-mosaic landscape of small patches of different sizes and shapes, each patch reflecting the legacy of disturbance caused by the death of individuals or small groups of overstory trees. Each of the patches may follow a sequence of development analogous to that outlined for large-scale disturbances, but the overall structure of the forest is dependent on the arrangement of the tapestry of patches that comprise the larger forest stand. The edges between these patches and the vertical distribution of structural elements represent important dimensions of the habitat variability that permits species coexistence in forests.

Recognizing the importance of forest structure and
pattern for biodiversity, ecologists are developing new, more sophisticated approaches for quantifying these parameters. Because the structural features that are important in determining animal and plant habitats differ among taxonomic groups and forest types, it is unlikely that any single approach will provide a universal standard by which forest structure and biodiversity can be related. Current efforts are utilizing new tools in the areas of spatial statistics, computer modeling, and remote sensing to provide suitable protocols for evaluating the connections between forest structure and pattern, management activities, and biodiversity of various groups of biota. These efforts will provide a better basis for understanding how forest ecology and biodiversity are related.

C. Food Webs and Community Organization in Forests

A food web is a set of species that live together and a specification of which species "eat" which other species. Plants provide the base of the food web by converting solar energy into biomass; herbivores and detritivores utilize living and dead biomass, respectively, to build their own tissues; and these organisms are in turn consumed by predatory species. Although a connection between the structure of food webs and biodiversity seems axiomatic, the exact nature of this connection is extremely complex. The biodiversity of a particular community is expressed in three different elements of food web structure: (a) the food chain length—the number of trophic links in the food web (producer-1\textsuperscript{consumer}-2\textsuperscript{consumer}-3\textsuperscript{consumer} etc.); (b) the number of distinct trophospecies—the set of all species that share some particular set of predators and prey—within a trophic level, and (c) the diversity of species constituting each trophospecies.

In forests, as distinct from most other biomes, the first link in the food web is dominated energetically by the detritivores rather than the herbivores. That is, most of the biomass in forests is consumed after the plant tissues die and are added to the soil as detritus. Thus, the detrital food web dominates the energetics of forest ecosystems and much of the complexity of the detrital food web in forests remains to be explored. These energetic considerations are not translated in a simple way to biodiversity in the respective food webs. The variety of different food sources available to herbivores in forests (leaves, stems, fruits, seeds, flowers, and roots of different plant species) provide numerous niches for their diversification, whereas most of the biomass energy available to detritivores is in the form of woody tissues which are structurally and biochemically so similar from species to species of trees that the diversification of wood decay organisms is somewhat limited. For example, one taxonomic group of insects, the higher termites (family Termitidae), overwhelmingly dominates in the comminution and decomposition of plant biomass in many tropical and warm temperate forest biomes. Nevertheless, in comparison with herbivore-dominated biomes (e.g., grasslands, aquatic ecosystems), the role of detritivores in forest biodiversity is probably relatively high.

The interactions of species in the forest community are not entirely competitive and predator-prey in nature. Some of the most fascinating interactions provide selective benefits to both of the interacting individuals or species populations—mutualistic and symbiotic relationships. Mutualism refers generally to a relationship in which two interacting species enhance their survival, growth, or reproduction, while symbiosis refers more specifically to two such organisms living together in close association. The most important symbiotic mutualism in forests is the mycorrhiza, an association between the mycelia of fungi and the roots of trees. The tree roots provide a supply of food to the fungus, which in turn increases the capability of the plant to acquire soil nutrients and water. All forest trees are mycorrhizal, and each tree species may harbor dozens of different fungal species in this mutually beneficial relationship. Two distinct types of mycorrhizae are common in forest trees—the ectomycorrhizae and endomycorrhizae. These types differ taxonomically, anatomically, and physiologically. The possible role of these mycorrhizal associations in regulating the diversity of forests was pointed out by Connell and Lowman (1989). They observed that pockets of low-diversity forest are found within the matrix of high-diversity tropical rain forest in all tropical regions. These low-diversity forests are composed of trees with ectomycorrhizal associations, whereas most of trees in the high-diversity forest are endomycorrhizal. Functional differences between the mycorrhizal types in soil nutrient acquisition or transfer could maintain or reinforce competitive interactions between individuals in these distinct forest types.

Most mutualisms between plants and animals have developed around the successful completion of the reproductive cycle of the plant—pollination and seed dispersal. Although many trees simply disperse their pollen to the wind, this method of pollination is unreliable when individuals are widely scattered in the forest, as in the species-rich tropical rain forest. Insects, nectarivorous birds, and bats visit the plants to exploit them as a source of food and in the process carry pollen from
one flower to another. Similarly, plants with seeds too heavy to be dispersed by the wind rely on animals to carry the seeds away from the mother plant. Although in most cases these mutualistic interactions are nonobligate and facultative (at least on one side), many remarkable examples of highly intricate, obligatory interactions have evolved, especially in the tropics. These interactions promote specialization and increased biodiversity.

III. HUMAN ACTIVITY AND FOREST BIODIVERSITY

Few forests have escaped the effects of human activity. Ancient civilizations decimated forests locally as a source of fuel and fiber and as sites for intensive agricultural production. In the modern era the pervasive influence of industrial civilization on forests has expanded to the regional and global scale through the additional effects of species introductions, air pollution, and likely climatic change. Insights from forest ecology provide a basis to evaluate the implications of these human influences on biodiversity.

A. Introduction of Alien Species

Alien insects and pathogens typically wreak havoc on host trees because these hosts have not developed adequate defenses through the process of evolution. The result is widespread decline of the host trees throughout their range even to near the point of extinction. The consequences of great reductions in the abundance of such declining species for the wider biotic community are not well understood and undoubtedly vary depending on the characteristics of the declining species (discussed later). Similarly, introductions of other species in different taxonomic groups or at other positions in the food web will have consequences for forest biodiversity that depend on their particular role in the community. Invasive trees may displace congeneric species from the forest community; and introduced herbivores that lack natural controls on their populations may decimate populations of their favored food plant species. Because the pace of species introductions has increased very rapidly in recent years, the ultimate consequence for forest ecosystems and biodiversity will be played out over the coming century.

B. Forest Harvest

The consequences of tree harvesting for forest biodiversity depend on particular features of the forest and the methods of harvest. Because all forests are regularly subjected to disturbance, if forest harvest practices mimic the natural disturbance regime, then consequences for biodiversity should be minimal. However, the exigencies of the financial bottom line result in harvest practices that do not mimic natural disturbances, and the consequences of actual forest harvest practices for biodiversity may be substantial. Most serious are (a) logging practices that result in the failure of the cut-over site to regenerate (e.g., because of severe damage to soils); (b) the coincident harvest of extensive areas, so that most of the landscape is in a single stage of forest stand development; and (c) recurring harvest on short rotation intervals. Also, forest harvest differs fundamentally from natural disturbance in that wood products are removed from the site, any species that depends on decaying wood for its habitat will be harmed by harvest practices that do not recognize this dependency. Finally, some species are believed to be old-growth obligates (i.e., they depend on old forests to complete their life cycles). These species are threatened when the great majority of natural, old forests in a region enters the harvest-regrowth system of industrial forestry, leaving little old-growth habitat.

C. Forest Conversion and Fragmentation

Permanent or semipermanent conversion of forested areas to other land uses has more severe consequences for biodiversity than forest harvest. Many forests occur where climate and soils are suitable for permanent agriculture and where the expansion of urban communities gobbles up native vegetation. Maintenance of biodiversity in such regions depends on having protected forest areas large enough to harbor the native flora and fauna. However, general rules to guide forest preserve planning for biodiversity protection are complicated by variations in the habitat requirements of different species. Forests in most agricultural regions occur as small fragments dispersed across the landscape, and how effectively these fragments can maintain viable populations of forest species is a topic of great concern (Schelhas and Greenberg, 1996).

D. Pollution

Local declines of forests has been associated conclusively with point-source releases of air pollutants, especially sulfur dioxide, fluoride, and toxic metals from smelters. Broad-scale, regional effects of air pollution on forests have been more difficult to demonstrate. Regional pollution—by ozone smog in the southwest-
ern United States and by acidic deposition in the eastern United States and Europe—probably has contributed to documented forest declines. Although improvements in emission controls and regulations in these regions are likely to reduce the chances of further damage, rapid industrialization without adequate emission controls in other regions of the world threatens forest health and biodiversity.

E. Rapid Climatic Change

The Pleistocene epoch was marked by dramatic climatic shifts that profoundly affected forests and biodiversity. The rapid rise in greenhouse gas concentrations is likely to bring about similar climatic shifts in coming decades or centuries. Of course, the consequences of rapid climatic change for forests and biodiversity will depend on a combination of species' natural responses (e.g., dispersal, colonization, natural selection) and human mitigation efforts. In many forest regions certain species, like the dominant trees and particular wildlife populations, are likely to be controlled by management efforts because of their relatively high value to humans. For relatively low-valued species and forest regions, the maintenance of biodiversity may depend on natural mechanisms or heroic human efforts that recognize nonmarket values of species. The current level of understanding of the physiological and population ecology of many forest-dwelling taxa is insufficient to predict the effects of rapid climatic change, but species with limited capacity for dispersal and colonization (e.g., soil invertebrates, perennial herbs) may be most sensitive. The implication of the loss of these species from forest communities will vary depending on the role they play in forest ecosystem function.

IV. BIODIVERSITY AND FOREST ECOSYSTEM FUNCTION

Forests regulate energy flow and cycling of materials in the landscape, collectively known as ecosystem functions. The effects of biodiversity, expressed in terms of species richness, on forest ecosystem functions are not yet clear and apparently not very straightforward. There is great interest in these possible effects as ecologists probe the implications of loss of species diversity for the integrity of ecosystems. Will species extinctions result in destabilization of ecosystem functions and possible feedbacks in the form of undesirable shifts in dominant vegetation types?

Whereas the role of species richness per se in regulating forest ecosystem functions remains unclear, it is well known that loss of particular species from the biota of a community can have important ramifications for energy flow, material cycling, and maintenance of stable biotic composition. That is, all species in the ecosystem are not equal in terms of their quantitative influence on ecosystem function. In particular, for some species there appears to be little or no redundancy with respect to their role in the ecosystem; if such a species performs some crucial activity, its loss from a forest can create havoc for the normal functioning of the ecosystem. These species are known as keystones.

In forests, keystone species are represented among many different taxonomic groups or food web positions. For example, at the primary producer level, a nitrogen-fixing tree like Alnus rubra in N-poor conifer forests is a keystone species; elephants appear to be keystone herbivores in semiarid Africa; beavers are key "ecosystem engineers" in northern forests; and jaguars that prey on seed predators in neotropical forests may be keystone carnivores. The loss of these species has consequences for ecosystem structure, function, and composition that are out of proportion from their individual abundances. Important efforts to conserve forest ecosystem functions in the face of biodiversity loss are focused on the identification of keystone species and ways of maintaining stable populations of keystone species.

Although the broader effects of overall reductions in biodiversity on forest ecosystem function remain more obscure, a specific example will illustrate that this is also a cause for concern. During the past several decades, excessive inputs of nitrogen (from air pollution) to forests in northern Europe have resulted in striking reductions in the abundance of mushroom species. These mushrooms are the fruiting bodies of ectomycorrhizal fungi discussed earlier. Lilleskov et al. (2000) have shown that a single host tree, white spruce (Picea glauca), maintains associations with about 90 different mycorrhizal fungi in natural, N-poor forests in Alaska, whereas in adjacent N-polluted forests only about five fungal associates are found. The high diversity of this mycflora in the natural forest certainly represents some degree of functional redundancy. However, if dozens of species are lost from the mycorrhizal fungal flora in temporarily N-polluted regions, some long-term effects on forest ecosystem function are likely. The fungi in the N-rich forests appear to utilize only the mineralized nitrogen sources (NH₃ and NO₂⁻), which are abundant there; with a return to normal, low–mineral N availability, in the absence of
mycorrhiza that access organic N forms, the productivity and nutrient cycling in the forests could be altered profoundly. Analogous situations probably apply in other aspects of both the detrital and grazing food webs of forest ecosystems around the world.

See Also the Following Articles

- Boreal Forest Ecosystems
- Deforestation and Land Clearing
- Disturbance, Mechanisms of Fires
- Ecosystemal Effects of Food Webs
- Key Taxonomic Species
- Logged Forests
- Rainforest Loss and Change
- Reforestation
- Succession, Phenomenon of

Bibliography


OBTAINING RELIABLE ESTIMATES OF biological diversity in the fossil record is critical to addressing key questions about changes in biological diversity over long timescales and the causes and consequences of those patterns. On what scales is the history of life characterized by periods of stasis punctuated by relatively brief periods of rapid evolutionary change? On what scales is it a more gradual, continuous process? What are the causes of these changes? How often are major changes in biodiversity associated with major changes in the physical environment, such as extraterrestrial impacts? Unfortunately, the fossil record provides an incomplete picture of the history of biological diversity. This chapter reviews some important features of the fossil record that could cause the apparent history of biodiversity to differ quite markedly from its true history. It then introduces several methods that have been applied in an attempt to account for these features, it discusses advantages and disadvantages associated with each of them, and it reviews how their application in specific cases has altered earlier views about changes in diversity through time. Throughout this chapter, the terms “biodiversity” and “biological diversity” will refer to taxonomic richness: the number of taxa present at a particular time or place. Although alternative metrics of diversity have been proposed, most paleobiologists...
have focused on taxonomic richness, at least when addressing regional or global patterns in biodiversity.

I. SAMPLING EFFECTS IN THE FOSSIL RECORD

A. The Fossil Record Is Incomplete

One obvious feature of the fossil record is that it is incomplete: only a small fraction of individuals are fossilized; of those, very few are collected and identified. As a result, the biological diversity recorded in the fossil record is less than total diversity over the region and time interval from which it is a sample. An additional consequence of this fact is that it makes stratigraphic ranges of taxa correspond to shorter periods of time than their true life spans. This is because the first appearance of a taxon in the fossil record occurs sometime after it actually originated, unless its very first representative was fossilized and subsequently sampled (a highly unlikely prospect). Similarly, the last appearance of a taxon occurs sometime before it actually went extinct. Biased estimates of stratigraphic ranges are important for the study of trends in biodiversity because they are often used to estimate biodiversity. Thus, rather than estimating diversity as the number of taxa that actually occur in samples from a particular stratigraphic interval, diversity is estimated as the number of taxa whose stratigraphic ranges encompass that interval. This will always be equal to or greater than the number of sampled taxa.

B. The Quality of Preservation Varies

If incomplete sampling were the only problem with the fossil record, it would still be possible to accept relative trends through time at face value—that is, an increase in diversity in the fossil record would indicate a real increase in biological diversity, even if the true number of extant taxa could not be determined. Unfortunately, however, the degree to which the fossil record is incomplete varies in both space and time. One source of this variation is characteristics of organisms themselves. For instance, organisms with hard parts are much more likely to be preserved in the fossil record than those without them. Claims, for instance, have a better fossil record than nematodes. One consequence of this is obvious: the difference between apparent and true diversity tends to be greater for the latter group than for the former. However, estimates of diversity trends through time can be affected as well, if the proportion of taxa with body parts that are readily fossilized does not remain constant. Indeed, some have argued that the “Cambrian explosion” represents, not an explosion of multicellular life, but a rapid and extensive proliferation of hard parts. Recently, this hypothesis has been fueled by molecular phylogenetic studies predicting that the major animal phyla diverged long before the early Cambrian. This hypothesis remains highly controversial; nevertheless, the very fact that it has received considerable attention illustrates just how profoundly sampling effects are believed to influence the fossil record of major events in the history of life.

Differences in abundances among taxa can also affect their preservation in the fossil record. It is individuals (or parts of individuals) that are fossilized; thus, more abundant taxa are, on average, likely to have more complete fossil records than will rare taxa. Further, fossil diversity is likely to be higher for intervals during which abundances were higher, on average, than during other intervals. This, too, has important consequences for inferences about diversity trends. For instance, several major episodes of diversification, as recorded in the fossil record, coincide with geophysical changes that probably increased rates of nutrient supply to the biosphere. Some workers (e.g., Vermeij, 1995) have argued that these geophysical changes were important causes of the coincident biological diversification. However, if increases in rates of nutrient supply also allowed taxa to sustain higher abundances, then increased probabilities of sampling during those sequences could be contributing to the increased diversity of the fossil record. Thus, inferences about the causes of trends in biodiversity, in addition to inferences about the trends themselves, can be influenced by sampling effects.

Differences in abundances can also affect estimates of stratigraphic ranges. One of these effects is illustrated by the previous example: first appearances of new taxa in the fossil record will be higher between two stratigraphic intervals when large increases in abundance occur between them as well, even if there were, in fact, an equal number of origination during the time periods that correspond to those intervals. Assessing the severity of this effect is complicated by the fact that there are sound biological reasons for increases in abundance to facilitate originations of new taxa; to the evolutionary biologist, a correlation between abundance and rate of speciation may be precisely what is expected.

Yet another characteristic of organisms that can affect inferences about biodiversity trends is habitat. In particular, individuals in some habitats are more likely to become fossils than individuals in others. For instance, marine soft-bottom habitats are likely to provide
more complete records of their inhabitants than will rocky shores. In the former case, individuals will much more readily be covered in sediment shortly after (or even before) they die. In the latter, wave action may render the remains of individuals unidentifiable before currents carry them to a location where they might be buried by sediment and preserved. Indeed, the fossil record of rocky shore communities is among the poorest in the marine realm. Again, the difficulties associated with these effects has led to disagreements about key events in the history of life. For instance, fossils record an explosive diversification of mammals in the early Tertiary, following the mass extinction that ended the Cretaceous period (best known for catastrophic extinctions among dinosaurs). Most workers believe that this reflects the true pattern, at least qualitatively. However, others, using molecular phylogenetics, have argued that much of this diversification occurred in the Cretaceous, prior to this mass extinction. One explanation for this discrepancy has been that these early Cretaceous mammals occupied habitats (such as forest interiors) for which their probability of being preserved was much lower than for their Tertiary descendants, who expanded into new habitats to occupy niches vacated by the extinction of dinosaurs. Whether this discrepancy between molecular and fossil data is primarily due to a poor fossil record for Cretaceous mammals or to shortcomings in the molecular methods remains highly controversial. Attempts to assess the severity of these shortcomings, such as those of Foote and coworkers (1999), should ultimately lead to a resolution.

As this discussion of the effects of ecology suggests, the probability of an individual being preserved in the fossil record depends, in part, on the sediments themselves. This, too, can have profound consequences on inferences about diversity patterns. For instance, in a classic paper, Raup (1976) noted a systematic increase in the volume of sedimentary rock through time. The implication was that the probability of individuals being preserved in the fossil record becomes progressively greater through time. Indeed, he presented a graph of sedimentary rock volume through time that looked strikingly similar to a graph showing biological diversity in the fossil record through time. After removing the effect of rock volume on diversity, he found no evidence for a long-term increase in species diversity through time. Although a key paper by Sepkoski et al. (1981), along with subsequent work, has convinced most workers that the apparent long-term diversity increase is real; the extent to which systematic increases in rock volume exaggerate the trend remains unknown.

This long-term increase in sedimentary rock volume has another potential effect on large-scale diversity patterns. If the probability of preservation and sampling increases through time, then the difference between the time of origination of a taxon and the time of its first appearance in the fossil record should shrink. Similarly, the difference between the time of extinction and the time of last appearance should shrink. This effect would lead to a progressive decrease in apparent extinction rates through time. Indeed, a long-term trend of decreasing extinction rates has been noted in work in the 1980s by Sepkoski and Raup and by Van Valen. Knowing the extent to which this trend reflects a real decrease in extinction rates (rather than a sampling effect as just described), is important, because Sepkoski (1984) has noted that such a trend, if real, could explain long-term changes in the relative prevalence of different taxa in the fossil record. As a result, paleobiologists (e.g., Pease, 1988, 1992) have proposed methods for assessing the effects of sampling biases on these rates. There appears to be an emerging consensus that the declines in evolutionary rates are real, but that they may be exaggerated by coincident trends in the quality of preservation.

The extent and quality of the fossil record varies over smaller timescales as well. Geophysical transitions can dramatically affect the probability that individuals will be preserved as fossils. When these transitions are global in scope, or occur in regions that receive disproportionately large amounts of attention from paleontologists, then apparent changes in diversity can result. Just such a possibility has complicated analysis of what was probably the most extensive mass extinction in the history of life, an event that marks the end of the Permian and the beginning of the Triassic. This event coincided with a major loss of marine benthic habitat (due to decreasing sea level). While it is certainly reasonable to suspect that biodiversity would decrease as available habitat decrease, it would also reduce the probability of sampling taxa that remained, as noted by Siglpor and Lipps in an influential paper published in 1992. This and other sampling effects associated with characterizing this extinction event and its aftermath are extensively reviewed by Erwin (1993).

C. The Extent of Sampling Varies

Decisions paleontologists make can also introduce, or exaggerate, differences in the quality of the fossil record. For instance, a disproportionately large amount of effort has been devoted to sampling in the Mesozoic and Cenozoic eras, because paleontological sampling tends to be associated with petroleum exploration, and rocks
that date from these areas have been subject to more extensive exploration than older rocks. Similarly, North America has been more extensively sampled than many other regions of the world. This reflects, in part, the fact that many paleontologists live in North America. The logistical problems associated with fieldwork are generally less severe, and the cost of travel lower, when fieldwork is conducted near a home institution. In addition, fieldwork tends to be easier in regions with a well-developed infrastructure. As a result, more of the preserved biodiversity is actually sampled and recorded, and thus the apparent diversity of well-sampled regions is higher than that of poorly sampled regions.

Similarly, the questions that one can ask of paleontological data, and the degree of confidence that can be assigned to answers, depends in large measure on the quality and quantity of data that can be collected. As a result, disproportionately large numbers of paleontologists focus their efforts on taxa, stratigraphic intervals, and geographical regions where fossils are abundant and well-preserved. This tends to exaggerate the effects of differences in preservation: where preservation is good (something which, alone, would tend to make biodiversity appear greater), a larger proportion of that record has been sampled. One effect of this tendency is that more workers focus on the most recent era of time, the Cenozoic (0–65 million years ago), than on the previous era, the Mesozoic (65–230 million years ago), despite the fact that the latter era was longer in duration. Likewise, more paleontologists study the Cenozoic than the earlier Paleozoic (250–600 million years ago).

### D. Time-Averaging

Yet another factor affecting estimates of diversity patterns is time averaging. When fossils are collected, their location in time is generally reported according to the particular stratum in which they were found. This corresponds to an interval, rather than a specific point, in time. Some of these strata represent longer periods of time than others. Thus, the number of taxa found in a particular stratum may be greater than the number of those taxa that were actually extant at a particular point in time during that interval (assuming that extinctions or origins occurred during the interval). The longer the interval lasts, and the higher the rates of origination and extinction, the more apparent diversity is inflated.

The issue is further complicated by uncertainties associated with the points in time that correspond to the boundaries of stratigraphic intervals. Indeed, the ages of these boundaries are continually revised. Thus, it may not always be clear which intervals were longer than others, and it is never clear just exactly how long those intervals were. Unfortunately, this problem becomes more severe as intervals are more finely divided. This makes minimizing the time-averaging problem more complicated. That is, the time-averaging problem should become smaller as the length of time spanned by an interval decreases. Since stratigraphic intervals are classified hierarchically (with some types of intervals nested within larger ones), one could, in principle, minimize the time-averaging problem by using a low level in this hierarchy. Unfortunately, however, the shorter the interval, the greater the proportional uncertainty associated with the estimated ages of the intervals’ boundaries. For instance, overestimating the age of the lower boundary of a 10 million year interval by 1 million years inflates the estimated duration of that interval by 10%. However, an overestimate of the same magnitude on the duration of a 3 million year subunit of that interval represents a 33% inflation.

In addition, it can be difficult to determine how stratigraphic intervals in one location correspond to intervals in other locations. If one is interested in estimating diversity or macroevolutionary rates for a region that includes more than one sampling location, then this can introduce further uncertainties. One source of this uncertainty is a simple extension of the problem of uncertainty in dating the boundaries of particular strata: if there are errors in the estimates of the ages of these boundaries at two different locations, then two intervals may be assumed to be substantially coincident when, in fact, they are not. This problem, like that discussed in the previous paragraph, becomes more severe as the duration of stratigraphic intervals decreases.

A second problem associated with this process of chronostratigraphic correlation results from the fact that direct estimates of ages are not available for all stratigraphic boundaries in all locations. Often, correlations are based on the presence or absence of particular indicator taxa. That is, the stratum in which a particular taxon first appears at, say, location B is assumed to correspond to the stratum in which it appears at location A. If the ages of the boundaries of the stratum at location A have been estimated, but those of location B have not, then the lower boundary of the stratum at location B is assumed to fall within that range of ages. However, since new taxa originate in particular regions, then expand their ranges gradually into new regions, part of the uncertainty associated with this effect is based on this rate at which indicator taxa increase their
geographical ranges. In addition, however, as noted earlier, the difference between the time at which a taxon was first present at a location and the time of its first appearance in the fossil record for that location varies depending on factors specific to particular locations, such as quality of preservation and local abundance. This further contributes to uncertainties in stratigraphic correlation. In practice, paleontologists attempt to minimize this problem by using multiple indicator taxa. They also emphasize those taxa that are likely to have high probabilities of preservation in the fossil record and whose spread is likely to have been rapid, such as planktonic foraminifera.

II. ESTIMATING DIVERSITY TRENDS

A. Rarefaction

One way to control for these sampling effects in making inferences about diversity trajectories is to compare taxonomic richness among locations or strata with samples that are equivalent in extent. This technique, called rarefaction, was developed by Sanders (1968)—and amended by Simberloff (1972)—to compare the diversities of different habitats in present-day ecosystems. The approach is as follows: a sample is collected from each of a set of habitats (for instance, a certain volume of sediment is obtained) using an identical sampling scheme. All of the individuals in each sample are identified, and their taxonomic identity is recorded. The next steps can be visualized by imagining placing all of the records for a particular sample in a bowl, stirring them, and then randomly selecting records from that bowl until none are left. Each time a record is picked, the taxonomic identity of that sample is noted. From this sequence of random draws, one constructs a graph with the number of records on the horizontal axis and the number of distinct taxa on the vertical axis. What is plotted on that graph are the results of that sequence of record selection. This curve must intersect the points (0, 0) and (1, 1). That is, before the experiment begins, 0 records have been picked from the bowl, so 0 distinct taxa have been selected. When one record is chosen, exactly one distinct taxon has been found. When two records have been chosen, either one or two distinct taxa have been found. [If the second record has the same taxonomic identity as the first, the next point is (2, 1); if it has a different identity, the next point is (2, 2).] Once all of the points have been plotted, a curve is fit to them: this is the rarefaction curve. This process is repeated for each sample collected, so that each sample has a rarefaction curve associated with it. Biodiversity in the different samples is then compared by examining the number of distinct taxa encountered for a given number of records selected. That number must equal to or less than the number of records in the sample with the fewest records. This comparison is illustrated in Figure 1.

Figure 1 illustrates something else: rarefaction curves sometimes cross. This means that the rank order of diversities of a set of samples can change, depending on sample size. One might therefore ask what rarefied diversity means, and, further, what it means to say that one site has a greater rarefied diversity than another. Hurlburt (1971) ventured an answer to this question, proposing that rarefied diversity measured the number of distinct taxa encountered, on average, by an individual organism in a particular habitat over the course of a particular number of encounters with other individuals (provided that individuals are not encountered multiple times). The fact that one habitat has a greater rarefied diversity than another for a given number of occurrences of individuals does not mean that that habitat actually has a greater taxonomic richness (i.e., contains more taxa) than another.

Although rarefaction can be applied to the fossil record in the manner just described, it is often applied somewhat differently. When comparing diversities over very large scales (global diversity during two strati-
graphic intervals, for instance), a sample enumerating all of the individuals in those intervals will not be available. Rather, data will consist of a number of distinct samples for each of the regions or intervals being compared. Further, while a list of the taxa found in each sample may be available, information on the abundances of those taxa may not. Such data have been rarefied in two different ways. In one approach (hereafter called rarefaction by occurrence), a rarefaction curve is constructed for a habitat by randomly selecting occurrences of taxa from its associated samples. The rarefaction curve is then a plot of the number of distinct taxa sampled against the number of occurrences of taxa drawn from the available samples. This is illustrated in Figure 2. Diversities are compared by constructing these rarefaction curves for different regions, taxa, or stratigraphic intervals. An alternative approach (hereafter called rarefaction by list) involves randomly selecting the entire list of taxa present in a given sample as a unit, rather than selecting occurrences of taxa within those samples. In this case, the rarefaction curve for a particular habitat, region, or stratum consists of the number of distinct taxa detected plotted as a function of the number of samples drawn. Note that in this instance, the curve is not constrained to pass through the point (1, 1), since several taxa may appear in a single sample. However, it still must pass through the point (0, 0). This method is illustrated in Figure 3.

Rarefaction has been applied to paleontological data in several instances. Raup (1975) pioneered its use in paleobiology. He examined an apparent increase in diversity of echinoid families since the Paleozoic. Through rarefaction, he demonstrated that this increase was qualitatively unchanged after accounting for increased sampling in younger stratigraphic intervals. In the 1990s, Miller and coworkers applied rarefaction by occurrence to explore taxonomic and regional patterns in the diversification of benthic marine invertebrates during the Ordovician Radiation. Unlike Raup, they found that broad-scale patterns suggested by raw data were changed by rarefaction. For instance, global rarefied diversities increased only through the mid-Ordovician, when they stabilized. This contrasts with the trend apparent from the fossil record as a whole: increasing diversity through the late Ordovician. Similarly, large peaks in diversity at the scale of individual paleocontinents were often reduced or eliminated by rarefaction. By contrast, Alroy (1999) applied rarefaction by list to explore the extent to which the radiation of mammals preceded the end-Cretaceous mass extinction (see Section I.B, "The Quality of Preservation Variables"). His results agreed qualitatively with the trend apparent from the "uncorrected" fossil record: much of the taxonomic diversification of mammals appeared to occur after this mass extinction.

Paleontologists have typically been very cautious in their interpretation of rarefied diversities. There is a reason for this caution: the biological meaning of a rarefied diversity measurement is unclear in the paleontological context. The interpretation offered by Hurlburt (described earlier) is appropriate only when (a) the sampling protocol described in the first paragraph of this section has been followed or (b) each occurrence
of a taxon is a truly independent, random sample from the habitat about which one wishes to make inferences. Generally, neither is true for large-scale paleobiological diversity estimates, because sampling effort has not been distributed randomly. For instance, when comparing the rarefied diversities of two stratigraphic intervals, a disproportionately large number of samples may be available for one habitat or region during one interval, but not for the other interval (but see methods developed by Alroy [e.g., 1998] for minimizing these biases). In addition, for rarefaction by occurrence, occurrences of taxa are not independent, because some occurrences come from the same sampling locations and others from different sampling locations. A biological interpretation of rarefied diversity given these sampling effects remains elusive. Nevertheless, rarefaction can provide important information. In particular, discrepancies between rarefied and overall diversity patterns indicate where marked increases or decreases in apparent diversity are likely to be artifacts of sampling.

B. Capture-Recapture Estimates

Although rarefaction has probably been the most widely used tool to account for the effects of variation in sampling effort, other approaches have been tried as well. For instance, in the mid-1980s, Nichols and coworkers proposed that models used to estimate abundances from capture-recapture data in population biology could be adapted, by analogy, to estimate taxonomic diversity in the fossil record. When one conducts a capture-recapture study, individual organisms are captured during discrete sampling occasions, which can occur at different locations, different times, or both. Each individual captured is given a unique mark, so that its capture history can be constructed. That is, if one constructs a matrix, the columns of which represent different sampling occasions and the rows of which correspond to each individual captured at least once, then each element of the matrix will be either 1 or 0, indicating whether individual i was captured on occasion j. Similarly, fossil data are collected on discrete sampling occasions, and one can simply list the taxa sampled on each occasion. In this case, the matrix elements are either 1 or 0 according to whether a taxon is found in sample j. In this context, taxonomic diversity (i.e., the total number of distinct taxa) is analogous to population size (the total number of distinct individuals). Unlike rarefaction, these approaches estimate taxonomic diversity (rather than sampled diversity given a particular sample size) when the models’ assumptions are met. Unfortunately, however, these assumptions are rarely fully met by fossil data. It is not always clear which results are robust to violations of assumptions and which are not. Therefore, as with rarefaction, workers have been cautious when interpreting these estimates.

Two different types of capture-recapture models can be used to estimate diversity, depending on the sampling scheme. When sampling occasions occur at different times and the duration of a particular sampling occasion is short relative to the time between sampling occasions, open-population models can be used. These models are designed to account for “births” and “deaths” of taxa (i.e., originations and extinctions) that may occur between sampling occasions. In the paleobiological context, stratigraphic intervals have been treated as sampling occasions. Thus the taxa sampled in an interval are considered as having been “captured” at some point during that interval (usually its midpoint). By contrast, when sampling occasions are sufficiently close together that originations and extinctions are few relative to the total number of taxa extant, then closed-population models can be used. These models have been applied when multiple samples have been taken within a stratigraphic interval. Although these models assume that there are no originations or extinctions between the specific points in time represented by different samples, they tend to be more robust to other types of sampling problems (discussed later). The particular models that have been applied to paleontological data include the Jolly-Seber open-population model and the closed-population models of Burnham and Overton (1979) and Chao (1987).

1. Jolly-Seber Model

The Jolly-Seber model makes four key assumptions:

1. The strata that represent different sampling occasions are short in duration relative to the time between those strata.
2. All taxa have the same probability of being sampled within a particular stratum.
3. All taxa have the same probability of going extinct between strata i and j + 1.
4. If a taxon goes extinct in the region being sampled, it does not subsequently reinvade that region from elsewhere.

Given these assumptions, diversity (D) can be estimated for any stratum i from the following data (notation follows that used by Nichols and Pollock, 1983): the number of taxa sampled during i (ni), the number of n, that were also sampled during at least one earlier interval (ni), the number of n, that were also sampled during at least one later interval (ni), and the number
of taxa sampled at least once before \( i \), at least once after \( i \), but not during \( i \) itself (\( i_i \)). The first step in estimating diversity involves estimating the number of taxa extant, but not sampled, during \( i \) (\( M \)). Note that \( r_i/ n_i \) is the proportion of taxa sampled during \( i \) that were sampled again during a later interval. Similarly, \( z_i(M_i - m_i) \) is the proportion extant but not sampled during \( i \) that were sampled again subsequently. If the model's assumptions hold, these two fractions will be equal, and diversity can be estimated. The second step involves noting that \( m_i/n_i \) is the proportion of taxa sampled during \( i \) that were sampled during earlier intervals. Similarly, \( M_i/D_i \) is the proportion of taxa extant during \( i \) that were sampled during earlier intervals. Again, if the model's assumptions hold, these two fractions will be equal, and diversity can be estimated.

As discussed earlier, taxa differ in many characteristics that affect their probability of being sampled as fossils (e.g., presence of hard parts, habitat, abundance). This violates assumption (2) and causes estimated diversity to tend to be lower than true diversity. Perhaps not surprisingly, then, when Nichols and Pollock (1983) applied this model to late Eocene mammals from the Big Horn Basin, Wyoming, goodness of fit statistics indicated rejection of the model. When applied to molluscan diversity in the Middle Miocene of South Jutland, the model provided an adequate fit. However, diversity estimates were lower than those obtained with a method that does not assume equal sampling probabilities (discussed later).

2. Closed-Population Models

Several alternative models can be used when multiple samples are available for a particular time interval. Those that have been applied to the fossil record share one important feature in common: they are designed to allow for the possibility that some taxa are more likely to be sampled than others. All of these methods utilize the frequency distribution of occurrences of the taxa sampled. That is, the raw material for the diversity estimate is the number of taxa occurring in only one sample, \( f_1 \), the number occurring in two samples, \( f_2 \), and so on, as well as the total number of occurrences in all samples.

Burnham and Overton (1979) utilized a statistical approach known as the jackknife to estimate biodiversity. The mathematics of the derivations are too complex to review here, but with this approach they obtained a series of possible estimators. The simplest of these utilizes only the number of samples and the number of taxa occurring in only one sample:

\[
\hat{D}_1 = D_m + f_1 \frac{k - 1}{k}
\]

where \( D \) is the first-order jackknife estimate of diversity, \( D_m \) is the number of distinct taxa appearing in the sample, \( f_1 \) is the number of taxa occurring in exactly one sample, and \( k \) is the number of samples. They developed additional estimates by incorporating the number of taxa occurring in more than one sample (for instance, their second-order jackknife uses the number of taxa appearing in exactly two samples, as well as the number appearing in just one). Nichols and Pollock (1983) applied these models to same molluskan data on which they used the Jolly-Seber method described earlier. They found that, even under a relatively intensive sampling regime, sampled diversity was as much as 30% lower than estimated diversity.

An alternative model, proposed by Chao (1987), uses the number of taxa occurring in either exactly one or exactly two samples:

\[
S = D_m + \frac{f_1 f_2}{2f_1}
\]

Wing and DiMichele (1992) used this estimate, usually termed "Chao-2," to compare regional vegetation diversity in the late Paleozoic and early Cenozoic for North American river and delta floodplains. Somewhat surprisingly, they found similar biodiversity levels during the two periods for these regions, despite the markedly higher apparent diversities for the latter interval at the global level.

Like the other methods discussed, these models make several assumptions that limit their applicability. The jackknife and Chao estimates assume that each taxon has an equal probability of occurring in each sample. That is, taxon \( A \) may have a different probability of being present in a sample than taxon \( B \), but that taxon-specific probability is the same for every sample within each region. This assumption may be violated if samples differ in extent or quality. However, even if great care is taken to minimize this problem, the assumption may still be violated. For instance, if abundances of particular taxa differed among sampling locations, then the associated probabilities of sampling may differ accordingly.

C. Generalized Inverse Gaussian-Poisson Distribution

Models other than those based on capture-recapture theory utilize frequency distributions of occurrences to
estimate diversity. One of these methods is known as the Generalized Inverse Gaussian-Poisson (GIGP) distribution. This method involves fitting a statistical distribution of specified form to the observed frequency distribution, then extrapolating from this statistical distribution to estimate the number of species that have been sampled zero times, $f_0$. Total biodiversity, then, is simply this number plus the number of species that were sampled at least once. Anderson and coworkers (1996) applied this method to plant and insect data for late Triassic braided-river plains in South Africa. They used these estimates to argue that plant and insect diversity in the sampled habitats was comparable to those of the present day, again in contrast to apparent global diversity patterns, which record increasing biodiversity levels through time.

The primary limitation of this approach is that it assumes that the underlying frequency distribution of occurrences follows a particular statistical distribution. That is, it is a parametric method. In this respect, it differs from the closed-population capture-recapture methods discussed earlier, which are nonparametric. As a general rule, parametric methods are more precise than nonparametric analogs when the assumptions about the underlying distribution are met. When they are not, however, the estimates can be very inaccurate. Another limitation, shared by the capture-recapture estimators, is that the uncertainty associated with the estimated diversity increases as the proportion of unsampled taxa increases. That is, when the probability of a taxon appearing in a sample is low on average (or when there are few samples), the estimates are especially prone to error. For this reason, the fact that a method works well as an estimate of present-day diversity does not necessarily mean that it will work well when applied to the fossil record, except perhaps where that record is unusually complete.

D. Other Nonparametric Methods

While most biodiversity estimates other than the closed-population capture-recapture models are parametric, some are nonparametric. Two of these models have been applied to paleontological data. One of these, “Chao-1,” was formulated by Chao (1984) and is actually mathematically equivalent to the Chao-2 estimate discussed previously. In this case, however, the number of individuals representing each taxon in a single sample is used rather than the frequency of occurrences in a set of multiple samples. That is, $f_i$ is the number of taxa represented by only one individual in a sample, $f_{i1}$ is the number represented by two individuals, and so on. Wing and DiMichele (1994) used this approach to examine local vegetation biodiversities in the Paleozoic and Cenozoic floodplain data set discussed previously. They found that, on average, local diversity in the late Paleozoic was similar to local diversity in the early Cenozoic. This was consistent with their findings for regional diversity using the Chao-2 estimator. However, they did find greater variation in diversity levels among sites in the Cenozoic, in particular, the most species-rich Cenozoic sites were much more diverse than the most species-rich Paleozoic sites.

A second model, developed by Chao and Lee (1992), utilizes the entire frequency distribution of occurrences in a set of samples, rather than just the number occurring in only one or a few samples. Anderson and coworkers (1996) estimated diversities with this method for the late Triassic plant and insect data mentioned in the previous section. These estimates ranged from about 45% lower (for insects) to 55% higher (for plants) than estimates obtained with the GIGP model discussed previously. These models share the limitations of the closed-population capture-recapture models discussed previously. In particular, the Chao-1 estimator assumes that probabilities of sampling are identical among samples for a particular taxon. The Chao and Lee estimator makes a similar assumption: each taxon has a constant probability of occurring in any given sample. When this assumption is reasonable (for instance, when studying groups of closely related taxa preserved in very similar sediments), the resulting estimates should be more precise than the methods discussed previously, which generally have greater uncertainties associated with them. Other estimates involve fitting observed distributions of occurrences to particular statistical distributions (the

III. UNEXPLORED DIVERSITY ESTIMATES

To this author’s knowledge, no other diversity estimation methods have been applied to the fossil record in published studies. However, several other methods of estimating biodiversity exist, and some may be applied to fossil data in the future. These include those capture-recapture estimates that assume all taxa have an equal probability of occurring in any given sample. When this assumption is reasonable (for instance, when studying groups of closely related taxa preserved in very similar sediments), the resulting estimates should be more precise than the methods discussed previously, which generally have greater uncertainties associated with them.
Several approaches have been used to account for these sampling effects in the estimation of diversity or (in the case of rarefaction) to minimize the effects of variation in sample size on estimates of diversity. All of these approaches were originally designed to estimate local diversity; thus, their application to regional and global patterns is problematic, and workers have interpreted results cautiously. Nevertheless, these results can sometimes be used to eliminate some explanations for particular biodiversity trends. For instance, Raup’s use of rarefaction confirmed that the increased diversity of echinoids through time was not solely due to the greater number of paleontological samples available for younger strata. When the results of these approaches differ from one another, or from diversity trends apparent from the fossil record in its entirety, however, things become complicated. Do the estimates indeed account for sampling effects and reveal true diversity trends, or does violation of model assumptions render the estimated trends even less reliable than the uncorrected diversity trends they are intended to improve?

The future of diversity estimation will no doubt involve considerable effort on several fronts. One promising approach is to apply several methods, then identify biodiversity trends that are robust to these alternative methods. Another is to investigate directly how different estimates are biased when particular assumptions are violated, then devise means of minimizing these biases. Diversity estimation methods that have not yet been applied to the fossil record may be incorporated into the paleobiological research program, and new estimates will undoubtedly be forthcoming as well. Non-parametric methods based on the frequency distribution of occurrences have been identified as a promising area for further progress by many biostatisticians. Finally, much recent work focuses on assessing the completeness of the fossil record of particular taxa or strata. In the future, these tools will undoubtedly be brought to bear on the problem of estimating diversity. For the past two decades, quantitative approaches have been rapidly growing in popularity and sophistication among paleontologists. This movement is still in its infancy, and its future is likely to produce an increasingly clear picture of the history of biodiversity.

See Also the Following Articles

**Biodiversity, Evolution and Biodiversity, Origin of Extinction, Rates of Mass Extinctions, Notable Examples of Measurement and Analysis of Biodiversity, Paleobotany, Species-Area Relationships**

GIGP method discussed previously is an example of this. These methods have been applied to estimate present-day diversity for particular taxonomic groups and particular regions, sometimes with good results. Many of these methods and applications were reviewed by Colwell and Coddington (1994) and by Bunge and Fitzpatrick (1993). The major barrier to their application to the fossil record is that differences among taxa in their probability of entering the fossil record, and subsequent loss of fossiliferous rock (at rates that may vary among regions and through time), may make distributions of fossil occurrences very different from distributions of occurrences of living organisms.

**IV. Conclusions**

As we have seen, many features of the fossil record make assessing diversity trends difficult. Some of these difficulties can be eased, at least in principle. For instance, unequal distribution of sampling effort by paleontologists can be reduced by emphasizing undersampled regions and strata in future fieldwork. Similarly, the expanding palette of tools for dating fossils and their surrounding sediments, their increasing precision, and the development of more robust statistical methods for chronostratigraphic correlation should progressively improve the fossil record’s temporal accuracy and precision. However, some problems are less tractable. For instance, fossiliferous rock is progressively lost as it ages. Thus, the fossil records of large regions (and thus the record of biodiversity in the associated habitats) may simply not exist. This problem is particularly acute for older time intervals. Similarly, organisms in some habitats are simply less likely to be preserved than organisms in others, and those habitats will thus have poorer records of their biodiversity history.

Several approaches have been used to account for these sampling effects in the estimation of diversity or (in the case of rarefaction) to minimize the effects of variation in sample size on estimates of diversity. All of these approaches were originally designed to estimate local diversity; thus, their application to regional and global patterns is problematic, and workers have interpreted results cautiously. Nevertheless, these results can sometimes be used to eliminate some explanations for particular biodiversity trends. For instance, Raup's use of rarefaction confirmed that the increased diversity of echinoids through time was not solely due to the greater number of paleontological samples available for younger strata. When the results of these approaches differ from one another, or from diversity trends apparent from the fossil record in its entirety, however, things become complicated. Do the estimates indeed account for sampling effects and reveal true diversity trends, or does violation of model assumptions render the estimated trends even less reliable than the uncorrected diversity trends they are intended to improve?

The future of diversity estimation will no doubt involve considerable effort on several fronts. One promising approach is to apply several methods, then identify biodiversity trends that are robust to these alternative methods. Another is to investigate directly how different estimates are biased when particular assumptions are violated, then devise means of minimizing these biases. Diversity estimation methods that have not yet been applied to the fossil record may be incorporated into the paleobiological research program, and new estimates will undoubtedly be forthcoming as well. Non-parametric methods based on the frequency distribution of occurrences have been identified as a promising area for further progress by many biostatisticians. Finally, much recent work focuses on assessing the completeness of the fossil record of particular taxa or strata. In the future, these tools will undoubtedly be brought to bear on the problem of estimating diversity. For the past two decades, quantitative approaches have been rapidly growing in popularity and sophistication among paleontologists. This movement is still in its infancy, and its future is likely to produce an increasingly clear picture of the history of biodiversity.

See Also the Following Articles

**Biodiversity, Evolution and Biodiversity, Origin of Extinction, Mass Extinctions, Notable Examples of Measurement and Analysis of Biodiversity, Paleobotany, Species-Area Relationships**

GIGP method discussed previously is an example of this. These methods have been applied to estimate present-day diversity for particular taxonomic groups and particular regions, sometimes with good results. Many of these methods and applications were reviewed by Colwell and Coddington (1994) and by Bunge and Fitzpatrick (1993). The major barrier to their application to the fossil record is that differences among taxa in their probability of entering the fossil record, and subsequent loss of fossiliferous rock (at rates that may vary among regions and through time), may make distributions of fossil occurrences very different from distributions of occurrences of living organisms.

**IV. Conclusions**

As we have seen, many features of the fossil record make assessing diversity trends difficult. Some of these difficulties can be eased, at least in principle. For instance, unequal distribution of sampling effort by paleontologists can be reduced by emphasizing undersampled regions and strata in future fieldwork. Similarly, the expanding palette of tools for dating fossils and their surrounding sediments, their increasing precision, and the development of more robust statistical methods for chronostratigraphic correlation should progressively improve the fossil record’s temporal accuracy and precision. However, some problems are less tractable. For instance, fossiliferous rock is progressively lost as it ages. Thus, the fossil records of large regions (and thus the record of biodiversity in the associated habitats) may simply not exist. This problem is particularly acute for older time intervals. Similarly, organisms in some habitats are simply less likely to be preserved than organisms in others, and those habitats will thus have poorer records of their biodiversity history.

Several approaches have been used to account for these sampling effects in the estimation of diversity or (in the case of rarefaction) to minimize the effects of variation in sample size on estimates of diversity. All of these approaches were originally designed to estimate local diversity; thus, their application to regional and global patterns is problematic, and workers have interpreted results cautiously. Nevertheless, these results can sometimes be used to eliminate some explanations for particular biodiversity trends. For instance, Raup’s use of rarefaction confirmed that the increased diversity of echinoids through time was not solely due to the greater number of paleontological samples available for younger strata. When the results of these approaches differ from one another, or from diversity trends apparent from the fossil record in its entirety, however, things become complicated. Do the estimates indeed account for sampling effects and reveal true diversity trends, or does violation of model assumptions render the estimated trends even less reliable than the uncorrected diversity trends they are intended to improve?

The future of diversity estimation will no doubt involve considerable effort on several fronts. One promising approach is to apply several methods, then identify biodiversity trends that are robust to these alternative methods. Another is to investigate directly how different estimates are biased when particular assumptions are violated, then devise means of minimizing these biases. Diversity estimation methods that have not yet been applied to the fossil record may be incorporated into the paleobiological research program, and new estimates will undoubtedly be forthcoming as well. Non-parametric methods based on the frequency distribution of occurrences have been identified as a promising area for further progress by many biostatisticians. Finally, much recent work focuses on assessing the completeness of the fossil record of particular taxa or strata. In the future, these tools will undoubtedly be brought to bear on the problem of estimating diversity. For the past two decades, quantitative approaches have been rapidly growing in popularity and sophistication among paleontologists. This movement is still in its infancy, and its future is likely to produce an increasingly clear picture of the history of biodiversity.

See Also the Following Articles

**Biodiversity, Evolution and Biodiversity, Origin of Extinction, Mass Extinctions, Notable Examples of Measurement and Analysis of Biodiversity, Paleobotany, Species-Area Relationships**


FRAMEWORK FOR ASSESSMENT AND MONITORING OF BIODIVERSITY

James A. Comiskey, Francisco Dallmeier, and Alfonso Alonso
Smithsonian Institution/Monitoring and Assessment of Biodiversity Program (SI/MAB)
National Museum of Natural History

I. Introduction
II. Adaptive Management
III. Keys to Success

GLOSSARY

adaptive management Adaptive management is a systematic, cyclical process for continually improving management policies and practices based on lessons learned from operational programs.
biodiversity assessment A biodiversity assessment entails the identification and classification of species, habitats, and communities within a given area or region. The overall purpose is to provide information needed to evaluate whether management is necessary to conserve biological diversity. Assessments also provide data and information that can be applied to monitoring programs or for providing basic information for scientific inquiry.

biological indicators Indicators are species or communities that enable an evaluation of environmental conditions and detect changes. Indicator species are normally surrogates of other species in the area of interest and are usually sensitive to environmental change. Environmental variables may also be used as indicators.

management Management in relation to biodiversity conservation involves decisions that have consequences for biological resources. Management can be designed to protect or restore biological resources, especially under conditions where the lack of intervention would lead to an irreversible or undesired change.

monitoring Monitoring involves the repeated collection and analysis of observations and measurements to evaluate changes in populations of species and environmental conditions. Monitoring also helps in assessing progress toward meeting a management objective. Monitoring can serve as a warning system, alerting managers that changes in biodiversity may require changes in biodiversity management regimes to ensure protection of biological resources.

uncertainty Uncertainty describes the condition whereby managers have a lack of knowledge of biodiversity that prevents them from defining the best course of management action. Uncertainty can be overcome through adaptive management.

AS PRESSURES INCREASE on earth’s natural resources, there is a need for managing those resources to protect biodiversity and ensure its sustainable use. A growing network of protected areas aims to ensure the survival of representative components and processes of ecosystems, especially in temperate zones. In tropical areas around the world, however, regional biodiversity is less well represented in protected areas, despite the fact that scientific records indicate biodiversity is high-
est in the tropics. Many of these unprotected areas, important for both biodiversity and natural resource extraction, are disappearing. Well-planned monitoring and management programs will promote the long-term maintenance of biodiversity. Some such actions require the protection of native or endangered species, and the elimination or restriction of non-native invasive species.

1. INTRODUCTION

What is the most effective management approach? As you will read in this chapter, this question requires a basic understanding of ecosystem function and the participation of all interested parties from the start of the management plan. As we increase our knowledge of the ecosystem, we can make management decisions based on solid scientific knowledge. Since management and conservation of biodiversity is our ultimate goal, we must monitor the effectiveness of our conservation strategies and approaches. Thus monitoring is an integral part of biodiversity management, enabling us to learn from the outcomes of our management “experiments,” a term known as adaptive management. Management practices that enhance biodiversity conservation can be continued or promoted, while those that threaten biodiversity can be evaluated and modified or discontinued.

The simplest definition of monitoring is to warn. It involves the collection and analysis of repeated measurements to evaluate the progress toward meeting a management objective. Monitoring can only be initiated if there are management objectives and alternative courses of action that can be implemented. The repeated collection of long-term data alone does not automatically imply a monitoring activity. Monitoring must be programmed within a management framework to avoid just gathering baseline information. Biodiversity changes have many causes including, among others, natural population cycles, climatic effects, and the direct impact of humans. The effects of these relationships are evaluated by research-related questions that may require multiple controls and replicates. This is an additional step in the monitoring process. This chapter examines assessment and monitoring in an adaptive management framework and identifies the key steps for a successful biodiversity monitoring project.

II. ADAPTIVE MANAGEMENT

Adaptive management is a systematic, cyclical process for continually improving management policies and practices based on lessons learned from operational activities. It is used to generate feedback for managers. The underlying assumption is that we are dealing with uncertainty and no one can be certain which management practices are best in many situations. Regular monitoring to check on the impacts of management practices will provide the information needed to determine if those practices must be changed to achieve the desired outcome. In the case of biological resources, the desired outcome is to maintain biodiversity in an optimally functioning state. Human activities can have drastic impacts on biodiversity, impacts that often are irreversible or require long and costly periods of recovery. Managing for biodiversity conservation through the adaptive management process can help avoid or mitigate those impacts.

Adaptive management can be described as a cycle of activities, where each step builds on the learning experiences of previous steps as the calibration of the goals and objectives are conducted through the monitoring feedback. The four primary steps in the adaptive management process are (a) design management and monitoring objectives, (b) implementing management, (c) assessment and monitoring, and (d) evaluation and decision making (Fig. 1). Each step is calibrated periodically to assure that the appropriate information feeds the next level. Different managers are usually responsible for each component. Thus, the cyclical nature of the process is very important in validating the results of the separate steps.

There are three types of adaptive management: reactive, passive, and active. The reactive approach is driven by factors such as public pressure and policies that are outside the management area. For example, a new law requiring protection of a certain species may lead to new management practices in a given area. Under such a scenario, the main issue may well be whether the new practices will cause conflicts among various users of the area. Passive adaptive management monitors a single

![Figure 1: Adaptive management cycle.](image-url)
management practice, which is viewed as the only alternative to achieve the management objective. Maintaining a healthy riparian community of herbaceous plants in a given area, for example, is dependent on the presence of canopy trees; thus, logging is discouraged. Water quality, which may have an effect on the herbaceous plant community, is not monitored under this approach. Therefore managers may fail to detect opportunities for improving management practices. This chapter focuses on active adaptive management, which compares alternative practices in anticipation of potential pressures from outside the management area, and taking advantage of management opportunities that may be lost under passive adaptive management.

The following section details the steps in the adaptive management process within the context of devising a biodiversity monitoring project.

A. Plan and Design

The primary task of planning and design is to identify an issue or challenge that requires implementation of a management practice to achieve a solution. The first question is what to monitor (e.g., regions, habitats, species, or guilds), followed by a determination of the scale and intensity of the monitoring activity. Establishing these parameters avoids needless investments of money, time, and other resources on collecting data that do not help evaluate whether the management is effective. This first step of the adaptive management process helps managers focus their attention on issues of concern and finding viable solutions. The following discussion describes critical elements of planning and design (Fig. 2).

1. Stakeholder Participation

Project definition, development, and implementation requires continuous stakeholder involvement and feedback. Stakeholders are individuals representing local, regional, national and international communities, agencies, and nongovernmental organizations (NGOs); industry, and any other organization with an interest in the project or responsibility within the area of the project’s influence. Consulting with stakeholders throughout different stages of the project is very important. Workshops that focus on setting priorities are an excellent avenue for reviewing the various concerns of stakeholders and promoting consensus.

2. Biological Inventory and Assessment

Most traditional approaches to biodiversity assessments are oriented toward increased commercial harvesting prospects such as intense logging in temperate forests or wildlife management for waterfowl and fisheries to benefit commercial and recreational industries. More recently, biodiversity assessments have encompassed habitats and the population status of rare, endangered, and exotic or invasive species to determine which areas are important to conserve and restore. The assessment process includes literature reviews, field surveys, and inventories to gather data and information relevant to the site-specific monitoring program and to identify knowledge gaps and target indicator variables.

a. Management Plans and Literature

Priority species or habitats for monitoring may well have been identified in previous management plans or research papers. These documents should be reviewed thoroughly in the planning process to ensure that the management issues and species in question are still relevant to the current activities. Similar monitoring activities may have been conducted elsewhere and there may be valuable lessons to be learned concerning the techniques used and the responses of the species to management. At this stage, it is also possible to identify
additional stakeholders who can provide valuable advice in the development of the monitoring project. Documenting the existing information will help refine the planning process and provide a foundation for the interpretation and reporting on future results.

b. Field Assessments

Field assessments might be needed to obtain up-to-date information on the present condition of the populations or habitats in question prior to establishing specific monitoring objectives. If detailed information on species and habitats is not available, it may be necessary to conduct studies that will help identify which species are present, their abundance and frequency, and to determine which should be targeted for monitoring.

Habitat descriptions and vegetation and land-use maps can greatly enhance the process of selecting target areas for management and monitoring activities. If these are not available, the field assessment should include collection of this type of information. Sampling at a landscape scale requires this level of detail to select strata for random selection of sampling points.

c. Problem Assessment

Once available information is reviewed and any necessary field assessments are completed, planners are positioned to define the issues of management concern. The best approach often involves a workshop where specialists from various disciplines: (a) address the scope of the challenge, (b) synthesize existing knowledge, (c) identify areas of uncertainty, and (d) create a model for management. It is particularly important to develop models because they enable a clearer definition of the challenge and a more accurate evaluation of ecological understanding of the species and habitats concerned. Experts frequently help identify issues and methods for finding effective management solutions. Where necessary, strategically focused research can facilitate the development of effective monitoring designs.

d. Indicators

Sampling all components of biodiversity in a given area can be an impractical and costly task, but biodiversity surrogates enable monitoring of ecosystem functions. As examples, one might choose an oak tree species affected by gypsy moth defoliation or a species whose recruitment and mortality is impacted by extended droughts. During the planning stage, it is possible to identify potential indicator species that can be used in monitoring. Important criteria in selecting indicator species include the following:

1. The species, populations, or observed physical or chemical phenomena should be good measures of one or more questions that the monitoring program was designed to answer.
2. The indicators should be able to detect a condition in advance to assist in solving the problem or else they may have a limited role in achieving the monitoring goals.
3. It should be possible to monitor the indicators within realistic budgets.
4. Whenever possible, indicators should be selected for which there are experimental controls (e.g., populations under different management intensities).
5. The species should be selected based on their potential for impacting management decisions (charismatic species, for example, are more likely to facilitate management changes than less well-known species).

Several considerations play a role in meeting the previous criteria. First, in tropical zones where the indicator species generally are not as well known as in temperate zones, birds, large mammals, and vascular plants are frequently used as indicators for biodiversity monitoring.

Second, keystone species are those upon which many other species may rely at some point in time during their life cycles. In tropical regions, for example, nectar-feeding bats are considered keystone species because the reproductive success of many plant species depends on them through pollination. Key common species are those that are wide-ranging, easily observed and studied, long-lived, and generally occurring at high population densities (e.g., oak species in the eastern United States). The sudden increase, decline, or absence of these species in certain habitats serves as a warning that may require management attention.

Next, species or taxonomic groups that have sensitive life histories may be good indicators for biodiversity monitoring. Amphibians depend on water for reproduction and will be affected by drought or water pollution. Information on amphibian numbers, diversity, sex ratio, age, and size structure will indicate changes that may affect other components of biodiversity. And, monitoring the abundance and diversity of selected tropical frog populations can provide valuable data on the health of those populations, and hence the habitat, from one year to the next.

Finally, key habitats may also be important indicators. For example, the aspen and wet meadow habitats of the Colorado Rocky Mountains in the United States
are good indicator habitats because of the species that depend on them.

3. Identifying Resource Needs
Management support of monitoring activities is essential to ensure a successful monitoring project. Essential resources include funding, people to conduct the monitoring, and technology.

Ensuring sufficient funding is crucial to long-term success. One of the most important considerations is for equipment costs, which may include not only the equipment used in the field but also computer and data-archiving facilities and associated software as well as resources for preparation and publication of the monitoring findings.

It is also important to evaluate the skills required of those who plan and implement the project, including the need for ecologists familiar with the ecosystem types under consideration, a statistician for sampling design, skilled field technicians for gathering and identifying field samples, data managers, and geographic information systems (GIS) specialists for compiling the information. Trained volunteers can be an enormous help to monitoring programs, especially in gathering data in the field, and they have been used extensively in a variety of different monitoring projects around the world.

Evaluating the time required for various phases of monitoring is appropriate during planning and design and should include time for planning, implementation, data management, and reporting. Later stages of the adaptive management cycle must receive equal attention in this regard. They frequently are not well represented in the enthusiasm of getting the project underway.

4. Defining Spatial and Temporal Scales
Biodiversity assessment and monitoring at local and regional scales can provide decision makers with high-quality data and cost-effective choices. Generally, the scale for assessment and monitoring is based on the geographical boundaries of the protected area or conservation unit or on subsets of such areas. Identifying the scale is essential at the planning stage because it has important implications in defining effective sampling design and the resources needed for the project. Monitoring species populations at a landscape scale may require sampling under different land management conditions and may call for special permits. At the local scale, monitoring usually concentrates on specific communities chosen because of the degree of threat or because of their location, and the results may not be representative of the landscape-level biodiversity.

Additional issues should be addressed when considering temporal scales or the time frame over which monitoring should be conducted. Such issues may involve not only sampling frequency but also the time taken to gather the information from each sampling point. This helps evaluate the need for qualitative (presence/absence, population estimates, or condition assessments) versus quantitative (censuses, samples, or demographics) measurements in the field.

5. Setting the Objectives
As has already been mentioned, the concept of monitoring within an adaptive management framework is very much dependent on the establishment of specific management objectives. Clearly stated objectives need to include measurable standards, desired states, threshold values, and the range and amount of acceptable change for all components of biodiversity that will be measured. The objectives should specify the limits of change before management action is taken, and they should be realistic, specific, and measurable. Several considerations assist in defining the monitoring objectives, including species and indicators to be monitored, location, attributes to be measured, action of management, the measurable quantity and degree of change, and the time frame for monitoring. These components, which should be included in the monitoring objective statement, are addressed in the following sections.

a. Location for Monitoring
Clear guidelines identifying the location of management concerns provide limits within which the management and monitoring results will be applied. These guidelines may vary depending on the management responsibilities and may include a portion of the population that resides within a protected area. The location may also be determined by the purposes of the management objective and may not cover the entire range of the species within the protected area. They may, for example, be restricted to a particular area such as riparian habitats.

b. Attributes to Be Measured
For most components of biodiversity, measurable attributes comprise quantitative measures such as population abundance in the area of interest. Nevertheless, qualitative parameters may also be measured. Cover, for example, may be measured for vegetation at ground level, including the percentage cover of a given herbaceous species, or it may be measured as the proportion of forest cover determined from aerial photography.
Demography or population dynamics may be used to evaluate mortality and recruitment rates.

There are several considerations when selecting parameters to measure. The attribute should be sufficiently sensitive in detecting the desired level of change and capable of distinguishing between natural fluctuations and human induced change. The change registered by the attribute should be biologically meaningful and lead to a logical management response. The cost of measuring the variable must be within budget, and it is necessary to identify the expertise and technical ability needed to measure the attribute. In addition, the variability in observer error must be minimal. Some of these issues may be addressed by conducting pilot projects.

c. Actions and Thresholds
Populations of the selected species can increase, decrease, or remain unchanged. These parameters are important to quantify, as they will be the measures of success in a biodiversity monitoring program. Under certain conditions such as with invasive species, it may be desirable to achieve a decline in the indicator species. When the management objective is to achieve an improvement in species or habitat factors, then the increase in abundance may be the focus for management. When current conditions are favorable such as when the population of the selected species is at the desired level, the management objective should be to maintain that population density.

There may also be threshold levels beyond which the species parameters should not pass. These can be divided into biological thresholds and management thresholds. A biological threshold is the point beyond which an irreversible change in the population may occur. This may be the minimum viable population density of the species, which, if exceeded, may lead to a population crash, loss of genetic viability, or extinction. In general, biological thresholds for many species are not known. This makes it imperative to choose an indicator species with a known threshold.

A management threshold is the point beyond which management should be changed to avoid an unacceptable risk to biodiversity. Management thresholds should be maintained above biological thresholds, especially if the latter are poorly known, and management thresholds need to provide a suitable margin of error to ensure that the biological threshold is protected. The monitoring activity itself must focus on management threshold levels; for example, a population is determined to require a minimum of 100 individuals to ensure genetic diversity and prevent the local extinction of the species (biological threshold). The management threshold should be maintained above this level, say at 120 individuals, to ensure the long-term survival of the species.

d. Degree of Change
Monitoring is frequently associated with measurements, and the task of assigning a measurable range of change can be challenging because of limited knowledge of the ecology of many species. This may lead to development of objectives that are measurable but not necessarily the most appropriate in an ecological context. Several authors have suggested that objectives do not have to be quantified, but they need to be as precisely defined as possible so that management success can be evaluated. Whatever the approach taken, the objective should consider the natural fluctuations in the population, the size necessary to ensure a viable population (thresholds), the amount of change that is biologically meaningful, the intensity of management, and the cost of measuring the specified level of change.

e. Time Frame for Monitoring
The time frame depends on the biology of the species (short-lived species will respond more quickly), the intensity of management (intense management produces rapid changes), and the level of specified change (the smaller the change, the sooner it will be detected). Short-term responses may benefit from more frequent evaluation of the management objectives, and they may be less susceptible to funding cuts. Nevertheless, sufficient time must be allocated to detect changes such as the impacts that selective clear-cutting of the forest has on water quality, which may not be detected until as much as a decade after the initial activity. Conclusions showing no problems reached after only a few years could be deceptive.

f. Statement of Management Objective
The previous steps lead to the development of a statement outlining the management objectives. An example of a management objective for monitoring could include maintaining the population of a particular species in a given area within 10% of the originally measured population density over the next 10 years. This type of objective is sometimes referred to as a target, or threshold, management objective because it measures the population in relation to a desired condition. Another management objective might be to increase the mean density of a species in a particular area by 15% over the next 5 years. This objective is referred to as a change, or trend, management objective because it measures a change in density over time. In both cases, the objective includes information on the location, the species in
question, the degree of change, and a timeframe for the monitoring activity.

6. Design Management

At this stage, it is important to address which options are available for management. Based on the objectives, the first decision will involve whether management should be active or passive. In most situations, active management is more appropriate as it provides several alternative options and will generally lead to quicker learning. Passive management may be suitable when it is not possible to design management with several alternatives, the costs for implementation are excessive and previous experience supports the use of one alternative as the best course of action.

The monitoring protocols to be used should be designed based on the management objectives, which address issues such as sampling design, data management, and analysis, interpretation of results, and reporting mechanisms. These issues are more fully described in the assessment and monitoring section that follows.

7. Management Response

Before monitoring begins, it is useful to identify alternative management responses based on different monitoring results. It is important for stakeholders to be aware of the implications of the monitoring results. The different management options must be evaluated to ensure that they are both economically and politically feasible. The management changes are more likely to be implemented if they are addressed in the planning stage and consensus is achieved.

8. Reassess the Objectives

It is helpful at this stage to pose a series of questions concerning the management objectives that will help identify any potential challenges or issues to be addressed. Have the objectives been clearly stated and are they realistic? What monitoring protocols are required to achieve the objectives? What is the timeline for accomplishing the objectives? Will the information that is gathered assist managers in making informed decisions? Can the results of the management decisions be statistically analyzed? Has a cost/benefit analysis been completed? What is the scale of the monitoring program (protected area, entire forest, selected number of habitats)? Is a consortium of organizations needed to achieve the objectives?

B. Management Implementation

New management should now be implemented or current management continued or modified. In most situations even if management has been in place for a number of years, the preparation and definition of management objectives will identify alternative options. Different scenarios may also be used in different locations to implement an active adaptive management approach. Where new management is identified, a detailed outline will ensure that the management activity is implemented effectively. Implementation monitoring will help guarantee the successful completion of this phase (see the next section).

C. Assessment and Monitoring

Assessment and monitoring protocols provide an essential tool for monitoring the status of biodiversity and changes in that status over time. In recent years, there has been an increased emphasis on standardizing monitoring protocols to facilitate comparisons among different projects. These long-term data are helpful in detecting the magnitude and duration of changes, how related taxa are changing, and "early warning" indicators of ecosystem health. They also serve as the basis for formulating additional research hypotheses, and, most important, the data are used to guide management decisions related to biodiversity conservation. The following section describes the different stages involved (Fig. 3).

1. Monitoring Programs

Four monitoring processes can be distinguished: baseline monitoring, implementation monitoring, effectiveness monitoring, and validation monitoring. All share the common theme of periodic measurements related to management objectives.

a. Baseline Monitoring

Baseline monitoring provides a standard against which future change may be evaluated. It is often the first set of measurements in a monitoring program and, hence, the beginning of effectiveness monitoring. Under certain conditions, baseline monitoring may be the collection of data prior to implementation of management, providing a before-and-after comparison. It may also be conducted in monitoring situations where the establishment of controls and treatments is possible.

Where a large number of variables are measured on a regular schedule with no specific management objective in mind, the activity is referred to as a long-term ecological study or a baseline study (e.g., measuring the composition, structure, and dynamics of trees in a permanent plot over a number of years). These are very important for monitoring as they provide information on a wide range of trends in biological variables.
b. Implementation Monitoring
This type of monitoring addresses the question, "Is the monitoring program doing what it is supposed to do?" It provides quantitative feedback on whether the management has been carried out as planned. Implementation monitoring acts as a form of quality control that audits the degree of compliance with previously established standards and guidelines. This type of monitoring may be conducted by external, independent parties to ensure that the implementation team is following established specifications.

c. Effectiveness Monitoring
Effectiveness monitoring assesses the outcome of the management: "Did it work?" It provides an evaluation of how effective the management was in meeting the results, needs, and expectations for which it was established. If the effects of the management do not reach the desired objective, then the management was not successful. For example, an exclusion fence for cattle or deer may be placed at the right location (implementation), but the rare plant population may still experience a decline (effectiveness). Thus, the management approach was not successful.

d. Validation Monitoring
This type of monitoring addresses the validity of predefined assumptions: "Are the models used for developing the management plan correct?" The difference between this type of monitoring and research is that research is specifically geared toward addressing cause-and-effect questions. Even though research may not fulfill the objectives of monitoring, it does provide critical information for implementing management. If monitoring requires the establishment of cause and effect, the logistical and cost implications must be evaluated carefully.

2. Defining Sampling Objectives
Sampling objectives should be defined based on the management objectives, which will specify the parameters for sampling, including the level of precision necessary. Precision is a measure of repeatability, or how close two repeated measurements are to each other, and is frequently obtained as a standard deviation of the estimated mean from which confidence limits are calculated. Increasing the number and size of samples increases confidence in the estimated measures but may be constrained by logistic and cost factors.

Sampling objectives also identify the level of change that must be detected and the risk of missing a real change or, alternatively, of detecting a false change. These factors are related to the variability of the population and the number and size of the samples used, all of which will determine the power of the sampling approach. When dealing with biological communities and attempting to maintain their parameters above the biological threshold levels, it is always best to err on the side of caution. Failing to detect a real change will result in failure of the monitoring program.

3. Sampling and Data-Gathering Designs
Sound sampling designs should provide guidelines for the most cost-efficient and effective way to gather and analyze data while maintaining high quality-control standards. The design of biodiversity monitoring programs should be based on site-specific management objectives, which help determine the sampling intensity to allow comparisons among habitat types and the assessment of long-term changes within habitats and the effects of management.

The sampling design should address the following questions:

1. What do researchers need to know about the site (e.g., data on forest composition, structure, and di-
versity as well as site-specific information such as the effects of acid deposition, drought, typhoon, El Niño, fires, and other disturbances)?

2. How will data from the monitoring program be used (managerial and scientific uses)?

3. Is the site representative of the selected habitat?

4. Will the design be sensitive enough to detect changes?

5. What limits of change are expected and are important to detect (e.g., mortality rates of forest stands should not exceed 3 to 5% per year)?

6. What is the degree of confidence expected from the results?

7. Will the sampling design produce results that will be representative of what is happening in the entire study area?

The management objectives will help determine the type of monitoring needed—qualitative or quantitative monitoring.

4. Qualitative and Quantitative Monitoring

Qualitative monitoring (e.g., habitat condition) is generally quicker and less expensive to implement than quantitative monitoring (e.g., estimates of the frequency and abundance of a species). With qualitative techniques, variability among different observers may be high. It is essential to reduce this variability because it has important implications in distinguishing real changes from those brought on by sampling variability. This may require testing several techniques to determine which one produces consistent field staff results.

Quantitative methods require more precise measurements for estimating population parameters, thus increasing the time and cost investment. Observer variability is less of an issue because direct measurements, not estimates, are recorded in the field.

5. Site-Specific Sampling

Within a selected habitat, sampling may be carried out in a systematic, random, or stratified fashion. Systematic sampling requires establishment of a grid within which a series of points are sampled. For random sampling, all points have an equal probability of being sampled. An area may also be stratified according to the needs of the monitoring program.

The number and size of the samples will be determined by the precision required for the monitoring. It will also be necessary to identify whether the samples should be permanent or temporary. Permanent sampling points will increase the level of repeatability and the cost and time needed to establish and permanently demarcate the sampling point.

Cost analysis of the sampling design is advisable so that objectives and the precision level can be adjusted, if necessary, to budget limitations. The desired precision level can be calibrated by changing the sampling plot unit and design as well as the confidence intervals and level.

6. Baseline Information and Pilot Projects

Monitoring requires a base or expected norm against which all future change may be compared. Baseline information may come from existing publications or initial surveys, assessments, or baseline monitoring. Unfortunately, initial surveys do not give any indication of how representative the results are of the selected habitat, and the literature may be sparse, especially in tropical regions. It is therefore suggested that surveillance or a pilot test be conducted before establishing a baseline against which future comparisons will be made. As an example, when establishing a vegetation plot, the first census results in a list of species and a description of how they are assembled and distributed at the site. Repeated measurements indicate changes in these variables over time and identify trends on which the baselines may be constructed.

Pilot projects are extremely useful and will likely identify issues that need to be addressed at an early stage. As a starting point, pilot projects provide guidance for the monitoring project design, the feasibility of the field techniques used, whether the sample size and number are sufficient to detect significant change, and whether the protocols used are logistically and economically feasible. Pilot projects may also identify areas where further resources need to be obtained, where training of field staff is needed, or where additional assistance is appropriate. At this stage, it is possible to modify the sampling to ensure that the management objectives can be assessed.

7. Data Collection

Data collection entails the measurement and assimilation of information in the field. A central issue in data collection is the need to gather data under consistent standards. Data collectors should be trained specifically for the task. Standard protocols should be used to facilitate cross-site comparisons and evaluation in multitaxa monitoring. For example, in past efforts related to forest management, dead and dying trees were viewed as less important than live trees because they had little value for wood products. More recent efforts related to biodiversity recognize the importance of dead trees as
elements of the ecological community that support other taxa.

8. Data Management and Analysis
A primary tenet of data management is to assure that data are accurately transferred from field data sheets and securely stored, making them available for future analysis. Managing data sets from a monitoring project should include data entry, verification, validation, archiving, and documentation.

Data entry refers to the process of accurately transferring data from the field data forms to the computer. Validity checks should be incorporated to assist in the detection of errors. Data verification, which follows data entry, is the process of assuring that the computerized records match the information from the original data sheets. Data validation is performed on verified computerized data for range and logic errors. For this task, data managers must be knowledgeable about the data they are handling.

Because monitoring may take place over long periods of time, data must be safely archived. Data documentation should accompany the archive, ensuring that all the information required for understanding the data is stored for future reference. Data documentation incorporates a number of descriptors such as the data set name, title, files and their structure, location, investigator, and dates of monitoring along with any other information pertinent to the data. A data dictionary describing the structure of the database should be stored with the archived data. When planning a monitoring program, researchers must include the costs involved in data management in long-term budgets. It is money well invested.

Data analysis is essential in relating data to the established goals and objectives of the monitoring program. Results from data analysis, recorded in the form of syntheses or reports, need to be available to other researchers and to managers and decision makers. Data analysis should: (a) ensure that the monitoring objectives are effectively addressed, (b) emphasize the importance of early detection of critical changes to allow lower-cost solutions, (c) recognize that results from the analysis are meant to contribute to knowledge about the biodiversity component being monitored, and (d) emphasize that the results are intended to provide biodiversity managers with a scientific rationale for setting appropriate standards.

While data analysis should synthesize large quantities of information to make it meaningful to the user, sometimes the tendency is to overanalyze it. This often produces an overwhelming amount of analytical results and interpretations that can obscure meaningful findings. It is important to maintain the focus of the monitoring program by returning to the specific questions that need to be answered. The results from the data analysis should be presented in a format that enables others to make their own interpretations without difficulty. This may involve the preparation of reports detailing how the findings affect the management of the protected area and presenting recommendations.

D. Evaluation and Decision Making
Evaluation elicits answers to the questions underlying the project's objectives, thus allowing for generation of management recommendations and calibration of the monitoring program. In addition, it allows an assessment of the monitoring process. It is often useful to pose the following questions: Are the results of each specific monitoring technique well integrated with the overall monitoring program? Do methods used ensure reliable, timely, and effective data analysis? Are the collected data subject to the appropriate techniques for data management and analysis? Could the data gathered be coupled with new technologies for analysis and management? What mechanisms exist or can be developed to allow timely transfer of data and information to managers and decision makers?

Management approaches should be viewed as hypothetical means to reach operational goals. Through evaluation of the monitoring data, managers receive timely feedback as these hypothesis are tested. Thus, evaluations are the tools for improving management by checking on management actions and providing guidelines for improvement. For example, when a predetermined degree of change is detected, appropriate action is taken and the results must be evaluated.

All the preceding steps lead to decision making regarding the need or lack thereof to adjust the management practices and monitoring program. If the findings determine that biodiversity trends are within the expected values, monitoring will continue without substantial alterations. If significant changes in the trends are observed, managers or decision makers need to design the most appropriate response. The reasons for monitoring can be evaluated at this stage. Is the monitoring still required, and, if so, do the objectives still remain the same?

Inconclusive results require adjustment of the objectives and sampling approaches to increase the degree of precision. However, careful planning and design of the monitoring program can reduce the risk of inconclusive information.
III. KEYS TO SUCCESS

One of the most critical and challenging issues in the success of monitoring programs is the need to gain long-term support. Project personnel must elicit the endorsement of high-level managers and the commitment of resource experts and technicians in the field. One method of achieving this goal is to incorporate monitoring as a routine duty of appropriate personnel. Another is to implement a reward system for personnel who detect a situation that needs to be fixed and take steps to devise a management-oriented solution. Institutional commitment is essential to ensure the long-term success of any monitoring activity.

A common practice is to develop partnerships among several organizations to assist in fund raising, standardizing protocols, providing site-specific documentation, and supporting training for high-quality research and monitoring. This encourages sound data and information management and reduced duplication. Such partnerships have also aided in disseminating information about a monitoring program to inform the public of its importance and benefits. Greater public understanding may lead to increased acceptance and support for the program.

Creating an interdisciplinary team with an integrated vision of the overall project beyond individual interests greatly facilitates the monitoring success. This type of integration is frequently achieved through workshops or through training of staff and managers in leadership and persuasive communications skills that promote teamwork in dealing with conflict resolution that is common to interdisciplinary teams.

It is also clear from the framework outlined here that planning is important to monitoring success. Revisiting the management and monitoring objectives will ensure that the activity proceeds smoothly. By using the principles of adaptive management, the results of the monitoring will help determine the success of the management and refine future activities to ensure the conservation of biological diversity.

Acknowledgments

We thank L. Alonso, P. Campbell, G. Davis, D. C. Maciver, C. Drysdale, and D. Kloepfer for providing valuable comments on the manuscript. We also thank numerous other people who have helped in advancing these concepts over the years.

See Also the Following Articles

- ECOSYSTEM, CONCEPT OF • ECOSYSTEM FUNCTION
- MEASUREMENT, AQUATIC AND MARINE COMMUNITIES • ECOLOGY FUNCTION MEASUREMENT, PLANT COMMUNITIES • ECOSYSTEM FUNCTION, PRINCIPLES OF • POPULATION VIABILITY ANALYSIS (PVA)

Bibliography


I. Fresh Waters: Physical Ecosystem Structure
II. Diversity of Habitats
III. Ecosystem Functions: Interactive Regulatory Mechanisms
IV. Biodiversity within Freshwater Ecosystems

Glossary

**benthos** Nonplanktonic animals associated with substrata within sediments or closely above the sediment–water interface.

**detritus** Nonliving organic matter in both soluble and particulate forms.

**littoral zone** Region of a lake or river between the land and the open water (pelagic zone) that is colonized by emergent, floating-leaved, and submersed aquatic plants and their attendant sessile microbiota (periphyton).

**pelagic zone** Open-water portion of a lake or reservoir beyond the littoral zone.

**periphyton** Bacteria, fungi, algae, and sessile microfauna growing attached to substrata (sediments, rock, plants, animals, sand).

**phytoplankton** Small photosynthetic plankton, largely algae and cyanobacteria.

**plankton** Small organisms with no or limited powers of locomotion that are suspended in the water and largely dispersed by turbulence and other water movements.

**production** Amount of new organic biomass formed over a period of time, and includes any losses from respiration, excretion, secretion, injury, death, and predation.

**zooplankton** Small animal plankton, usually denser than water, that sink by gravity to greater depths.

**FRESHWATER ECOSYSTEMS ARE INTERACTIVE SYSTEMS** within which biotic species and their growth and adaptation, and associated biological productivity, nutrient cycling, and energy flows among inland aquatic microbial, plant, and animal communities, are integrated with their environment. These inland waters include lakes, reservoirs, rivers, streams, and wetlands. The subject discipline, termed limnology, evaluates how these interactive relationships are regulated by natural and modified variations in their physical, chemical, and biotic environments. The standing water (lentic) lake or reservoir ecosystem is a system that is intimately coupled with the land surrounding it in its drainage area via chemical components transported to the lake by groundwater or running surface waters (lotic) of streams and rivers.

**I. FRESH WATERS: PHYSICAL ECOSYSTEM STRUCTURE**

Most water on the Earth’s surface is saline in oceans (97.61%), and most of the remainder is polar ice
(2.08%) and ground water (0.29%). The remaining water (only 0.009%) exists in freshwater ecosystems as temporary storage in lakes and reservoirs on land. These surface waters have much shorter renewal times than those of the oceans (Table 1). The volume of water flowing from these storage sites to the sea in streams and rivers is very small (0.00009% of total water), with a mean residence time of about two weeks. During high-precipitation periods, surface waters of rivers increase and often recharge adjacent groundwater aquifers. Return flows from ground water to rivers occur during periods of low flow or drought, and usually maintain a base flow in river channels.

The storage of water in lakes and retention times are altered by shifts in the balance between inputs from all sources and water losses. Lakes receive water from precipitation directly on the surface, and from surface in- flows from the drainage basin, and from subsurface groundwater seepage. Lakes lose water by flow from a usually single outlet (drainage lakes), by seepage through the basin walls into the ground water (seepage lakes), evaporation, and evapotranspiration from higher aquatic plants. Saline lakes, which constitute nearly half of inland surface waters (see Table 1), occur in closed basins with no outflow except by evaporation. That evaporation results in a gradual marked increase in residual concentration of dissolved salts that were imported with influent water.

Natural lakes are concentrated in the subarctic and temperate regions of the Northern Hemisphere, whereas reservoirs are constructed predominately in the subtropical and subtropical regions. Some 40% of the total volume of surface freshwater is contained in seven great lake basins of Siberia (Lake Baikal), North America (the Laurentian Great Lakes: Superior, Michigan, Huron, Erie, and Ontario), and eastern Africa (Lake Tanganyika). Most of the millions of lakes and reservoirs, however, are very much smaller and relatively shallow, usually <15 m in depth (Fig. 1). Mean depths are even more shallow (<5 m). As a result, light commonly penetrates to over half of the sediments within the basins of a large percentage of lakes. Photosynthesis can thus occur not only by algae that are suspended in the water column but can extend to higher aquatic plants and attached algae. Major flowing-water ecosystems occur in low-gradient tropical and subtropical regions and are posited to support high photosynthetic productivity in extensive floodplain and land–water interface regions.

The productivity and internal metabolism of aquatic ecosystems are driven and controlled by energy from solar radiation acquired by photosynthesis. Inland wa-

<table>
<thead>
<tr>
<th>Water in the Biosphere</th>
<th>Volume (thousands of km$^3$)</th>
<th>Percentage of total</th>
<th>Renewal time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oceans</td>
<td>1,370,000</td>
<td>97.61</td>
<td>3100 years$^a$</td>
</tr>
<tr>
<td>Polar ice, glaciers</td>
<td>29,000</td>
<td>2.08</td>
<td>16,000 years</td>
</tr>
<tr>
<td>Groundwater (actively exchanged)$^b$</td>
<td>4000</td>
<td>0.295</td>
<td>300 years</td>
</tr>
<tr>
<td>Freshwater lakes</td>
<td>125</td>
<td>0.009</td>
<td>1–100 years$^d$</td>
</tr>
<tr>
<td>Saline lakes</td>
<td>104</td>
<td>0.008</td>
<td>10–1000 years$^c$</td>
</tr>
<tr>
<td>Soil and subsoil moisture</td>
<td>67</td>
<td>0.005</td>
<td>280 days</td>
</tr>
<tr>
<td>Rivers</td>
<td>1.2</td>
<td>0.00009</td>
<td>12–20 days$^c$</td>
</tr>
<tr>
<td>Atmospheric water vapor</td>
<td>14</td>
<td>0.00009</td>
<td>9 days</td>
</tr>
</tbody>
</table>

$^a$ Modified from Wettzel (1983, 2000) after Kalinin and Bykov. Slightly different values are given by I. A. Shiklomanov, but ratios are similar.

$^b$ Based on net evaporation from the oceans.

$^c$ Kalinin and Bykov estimated that the total groundwater to a depth of 5 km in the earth’s crust amounts to $60 \times 10^6$ km$^3$. This is much greater than the estimate by the U.S. Geological Survey of $8.3 \times 10^6$ km$^3$ to a depth of 4 km. Only the volume of the upper, actively exchanged groundwater is included here.

$^d$ Renewal times for lakes vary directly with volume and mean depth, and inversely with rate of discharge. The absolute range for saline lakes is from days to thousands of years.

$^e$ Twelve days for rivers with relatively small catchment areas of less than 100,000 km$^2$, 20 days for major rivers that drain directly to the sea.
ters receive organic products of photosynthesis directly from their aquatic flora and indirectly from their drainage basins as particulate and dissolved organic matter from terrestrial and wetland plants imported by stream water, storm runoff, ground water, and the atmosphere.

Light is attenuated exponentially with increasing depth in water. As the light is attenuated, ultraviolet wavelengths are absorbed strongly by dissolved organic compounds in the water. The blue portion (500–600 nm) of the visible spectrum penetrates most deeply in relatively clear waters. Infrared wavelengths are absorbed rapidly, largely as a result of the molecular structure of water molecules, and much of this energy is dissipated as heat. Because water becomes less dense as heat content increases above 4°C, the less dense warmer water floats upon the more dense cooler water. This density stratification results in separation of major strata within lakes, where a warmer, less dense stratum (epilimnion) overlies a cooler, denser zone (hypolimnion). The interface zone (metalimnion) is a region of rapid thermal discontinuity, often as much as several degrees change per meter, between the epilimnion and the hypolimnion.

The resulting density stratification affects not only the thermal structure and water mass stratification, but also the hydrodynamics of lakes and reservoirs. Heat is distributed and altered by the physical work of wind energy, currents and other water movements, basin morphometry, and water losses. Patterns of density-induced stratification influence physical and chemical properties and cycles both spatially within the lake and seasonally. These characteristics structure the aquatic habitats and have marked attendant effects on all chemical cycles, metabolic rates, and the population dynamics of organisms and their productivities.

The total salinity of inland waters nearly always consists of eight most abundant ionic species, of usually four major cations (Ca²⁺, Mg²⁺, Na⁺, K⁺) and four anions (HCO₃⁻, CO₃⁻, SO₄²⁻, Cl⁻). The salinity of fresh waters has a world average concentration of about 120 mg/L, but varies among continents and with the lithology of landmasses (Table II).

Of the major constituents of the cellular protoplasm of organisms (C, H, O, N, P, and S), the biogeochemical cycles control the physical and metabolic availability of phosphorus, nitrogen, and several minor nutrients. These elements often limit biotic development, particularly photosynthetic generation of organic matter. Phosphorus, in comparison to other macronutrients required by biota, is least abundant geochemically in a large portion of the global landforms and drainage. As a result, phosphorus is commonly the first element to limit biological productivity, especially of photosynthetic producers and heterotrophic microbiota. When phosphorus is in adequate supply, nitrogen availability invariably limits productivity. Standing inland waters
have been categorized into various trophic scales on the basis of ranges of major nutrients and algal productivity (Table III).

Functionally similar organisms of biological communities can be grouped into *trophic levels* based on similarities in patterns of organic matter production and consumption. Energy is transferred and nutrients are cycled within an overall ecosystem trophic structure. The productivity of each trophic level is the rate at which energy enters the trophic level from the next lower level. Because organisms expend considerable energy for maintenance and since death of an organism routes much energy and nutrients into the detrital pool of dead organic matter, only a portion of the energy of one trophic level is available for transfer and use by higher trophic levels. Available energy decreases progressively at higher trophic levels, so that rarely can more than five or six trophic levels be supported. The efficiency of energy transfer from one level to the next is low (5–15%), and often decreases as trophic level increases.

The organic matter of freshwater ecosystems is primarily (>90%) dead organic matter, and most (usually 80–90%) of this organic detritus occurs as soluble or colloidal organic compounds. Oxygen, however, is very insoluble in water, and most diffuses to the atmosphere. When organic matter in lakes is metabolized and organic compounds respired by heterotrophic organisms, the amount of organic matter to be oxidized can be much larger than the dissolved oxygen available for oxidation. This limitation can become acute when lakes are stratified by density differences, and exposure of deep strata to oxygen of the atmosphere may be restricted for long periods of time (months).

In large lakes, the volume of water with sufficient light to support photosynthesis (*euphotic zone*) is small relative to the total volume of water containing stored dissolved oxygen that can be used in heterotrophic respiration. In small and relatively shallow lakes, however, the proportion of water supporting photosynthesis is large relative to total lake volume, and respiratory demands can exhaust oxygen dissolved in the lower strata of the lake. The process of eutrophication exacerbates the exhaustion of dissolved oxygen in large portions of lake ecosystems. *Eutrophication* is the increased loading of organic matter to an aquatic ecosystem. Often in lakes, eutrophication results from increased photosynthetic production of organic matter in response to excessive loading of nutrients, particularly phosphorus, to the water from external sources.

<table>
<thead>
<tr>
<th>Cations (μeq l⁻¹)</th>
<th>Anions</th>
<th>North America</th>
<th>Mg²⁺</th>
<th>Na⁺</th>
<th>K⁺</th>
<th>SO₄²⁻</th>
<th>Cl⁻</th>
<th>NO₃⁻</th>
<th>Fe (as Fe₂O₃)</th>
<th>SiO₂</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.0</td>
<td>5</td>
<td>9</td>
<td>1.4</td>
<td>68</td>
<td>20</td>
<td>8</td>
<td>1</td>
<td>0.16</td>
<td>9</td>
<td>142</td>
<td></td>
</tr>
<tr>
<td>7.2</td>
<td>1.5</td>
<td>4</td>
<td>2</td>
<td>31</td>
<td>4.8</td>
<td>4.9</td>
<td>0.7</td>
<td>1.4</td>
<td>11.9</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>31.1</td>
<td>5.6</td>
<td>5.4</td>
<td>1.7</td>
<td>95</td>
<td>24</td>
<td>6.9</td>
<td>3.7</td>
<td>0.8</td>
<td>7.5</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td>18.4</td>
<td>5.6</td>
<td>5.5</td>
<td>3.8</td>
<td>79</td>
<td>8.4</td>
<td>8.7</td>
<td>0.7</td>
<td>0.01</td>
<td>11.7</td>
<td>142</td>
<td></td>
</tr>
<tr>
<td>12.5</td>
<td>3.8</td>
<td>11</td>
<td>—</td>
<td>43</td>
<td>13.5</td>
<td>12.1</td>
<td>0.8</td>
<td>1.3</td>
<td>23.2</td>
<td>121</td>
<td></td>
</tr>
<tr>
<td>3.9</td>
<td>2.7</td>
<td>2.9</td>
<td>1.4</td>
<td>31.6</td>
<td>2.6</td>
<td>10</td>
<td>0.05</td>
<td>0.3</td>
<td>3.9</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>4.1</td>
<td>6.3</td>
<td>2.3</td>
<td>58.4</td>
<td>11.2</td>
<td>7.8</td>
<td>1</td>
<td>0.67</td>
<td>13.1</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>Cations (μeq l⁻¹)</td>
<td>750</td>
<td>342</td>
<td>274</td>
<td>59</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1425</td>
<td></td>
</tr>
<tr>
<td>Anions</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>958</td>
<td>233</td>
<td>220</td>
<td>17</td>
<td>—</td>
<td>1428</td>
<td></td>
</tr>
</tbody>
</table>

*From Wetzel (1983, 2000), after data from several sources.*

| Values of calcium are likely less, on the average, than Na and Mg in Australian surface waters.*
<table>
<thead>
<tr>
<th>Type</th>
<th>Mean primary productivity (mg C m⁻² day⁻¹)</th>
<th>Phytoplankton density (cm³ m⁻³)</th>
<th>Phytoplankton biomass (mg C m⁻³)</th>
<th>Chlorophyll a (mg m⁻³)</th>
<th>Dominant phytoplankton</th>
<th>Light extinction coefficients (η m⁻¹)</th>
<th>Total organic carbon (mg L⁻¹)</th>
<th>Total P (µg L⁻¹)</th>
<th>Total N (µg L⁻¹)</th>
<th>Total inorganic C (mg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitropic</td>
<td>&lt;50</td>
<td>&lt;1</td>
<td>&lt;50</td>
<td>0.01–0.5</td>
<td>Chrysophyceae, Cryptophyceae,</td>
<td>0.03–0.8</td>
<td>&lt;1–5</td>
<td>&lt;1–250</td>
<td>2–20</td>
<td></td>
</tr>
<tr>
<td>Nitropic</td>
<td>50–300</td>
<td>20–100</td>
<td>0.3–3</td>
<td></td>
<td>Dinophyceae, Bacillariophyceae</td>
<td>0.05–1.0</td>
<td>&lt;1–3</td>
<td></td>
<td></td>
<td>10–25</td>
</tr>
<tr>
<td>Meso</td>
<td>1–3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meso</td>
<td>250–1000</td>
<td>100–300</td>
<td>2–15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10–30</td>
</tr>
<tr>
<td>Oligo</td>
<td>3–5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oligo</td>
<td>&gt;1000</td>
<td>&gt;300</td>
<td>10–500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50–1100</td>
</tr>
<tr>
<td>Eutrophic</td>
<td>&gt;10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100–400</td>
</tr>
<tr>
<td>Eutrophic</td>
<td>&lt;50–500</td>
<td>&lt;50–200</td>
<td>0.1–10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>400–2000</td>
</tr>
</tbody>
</table>

Wetzel (1983, 2000), after many authors and sources. According to approximately net primary productivity, such as measured by the ¹⁴C method.
depending on the environmental conditions. When these electron acceptors are essentially depleted, organic matter can be anaerobically degraded by fermentation to methane and CO₂ in two stages. Facultative and obligate anaerobic bacteria convert proteins, carbohydrates, and fats primarily to fatty acids by hydrolysis and fermentation. Obligatory anaerobic methanogenic bacteria then convert the organic acids to methane and CO₂.

II. DIVERSITY OF HABITATS

Freshwater ecosystems consist of entire drainage basins as water moves from land and in groundwater runoff to stream and river channels, and to recipient lakes or reservoirs. The nutrient and organic matter content of drainage water from the catchment area is modified in each of the terrestrial soil, stream, and wetland–littoral components as water moves downgradient to and within the lake or reservoir itself (Fig. 2). Photosynthetic productivity of organic matter is generally low to intermediate in the terrestrial components, highest in the wetland–littoral interface regions between the land and water, and lowest in the open water (pelagic) zone. The same productivity profile emerges in the gradient from land to river channels, where the greatest productivity occurs in the marginal floodplain regions. Autotrophic productivity in river channels is generally low, as is also the case in the pelagic regions of lakes. Most of the organic matter utilized by heterotrophic communities in running water is imported from floodplain and terrestrial sources as particulate and especially dissolved and colloidal organic compounds.

The interface region between the land and water is always the most productive per unit area along the gradient from land to the open water of lakes, reservoirs, and streams. Because most aquatic ecosystems occur in geomorphologically mature terrain of gentle slopes and are small and shallow, the wetland–littoral components usually dominate in both productivity and the synthesis and loading of organic matter to the systems. The region of greatest productivity is the emergent macrophyte zone. Emergent aquatic plants have a number of structural and physiological adaptations that not only tolerate the hostile reducing anaerobic conditions of saturated sediments but also exploit the high nutrient conditions and water availability of this habitat. Nutrients entering the zone of emergent macrophytes from external sources tend to be assimilated by bacterial and algal microflora of the sediments and detrital organic particles, and are then recycled to the emergent macrophytes. Dissolved organic compounds released from decomposition of plant detrital materials dominate the export of organic matter from the emergent plant zone.

Submersed macrophytes are limited physiologically by slow rates of diffusion of gases and nutrients in water within the boundary layers surrounding the leaves and by reduced availability of light underwater. Internal recycling of resources, particularly of gases (CO₂, O₂) of metabolism and of critical nutrients, is important to the abilities of submersed plants to function and grow as well as they do in underwater conditions of chronic light and gas limitations. Despite these adaptive mechanisms, growth and productivity of submersed plants are less than those of emergent and floating macrophytes.

The second most productive component of the wetland–littoral community is the microflora attached to aquatic plants epiphytically and to other surfaces, both living and dead. The surfaces provided by aquatic plants in lakes and rivers can be very large, often exceeding 25 m² per square meter of bottom sediments. High sustained growth of attached microflora results from their recycling of essential gases (CO₂, O₂) and dissolved nutrients within the attached communities. Nutrient uptake from the surrounding water is directed primarily to the high net growth of attached microflora and is responsible for the high capacity of wetland–littoral areas to improve the quality of water passing through these communities.

The wetland–littoral complex of higher plant and microbial communities produces the major sources of organic matter and energy of many freshwater ecosystems, including the marginal floodplains of many rivers. Most of the particulate organic matter is decomposed within these interface regions. Organic matter is exported predominantly from these marginal regions as dissolved organic matter to the recipient lake or river (Fig. 3).

The deep-water pelagic zone of lakes is the least productive along the gradient from land to water (see Fig. 2), regardless of nutrient availability. Growth of phytoplanktonic algae of the pelagic zone is limited by sparse distribution in a dilute environment where efficient nutrient recycling is restricted by the sinking of senescent phytoplankton below the depth of photosynthesis. When nutrient recycling and availability are increased, greater phytoplankton cell densities attenuate underwater light and reduce the volume of water in which photosynthesis occurs. Despite low productivity per unit area, pelagic productivity can be collectively important in large lakes and for higher trophic levels that depend on this source of organic matter.

Higher trophic levels of communities in freshwater
ecosystems consist of zooplankton (dominated by four major groups of animals: protozoa/protista, rotifers, and the crustaceans cladocera and copepoda) and benthic invertebrates. In the pelagic zone, small fishes, fry of larger fishes, and predatory zooplankton, which collectively comprise a third trophic level (primary carnivores), consume a portion of these generally herbivorous organisms. A fourth trophic level may consist of medium-sized piscivorous fishes, and the fifth level of large predatory piscivorous fishes. Higher trophic levels are rare in freshwater ecosystems.

The species composition of the higher trophic levels affects the pathways of energy utilization from lower trophic levels. Environmental factors that selectively influence the populations of the communities can alter the pathways and strengths of energy fluxes from subordinate trophic levels. For example, efficiency of consumption of primary production by zooplankton is often appreciably greater in the absence of zooplankton-feeding fishes than in their presence. The population structure of the phytoplankton community responds variably to grazing impacts in concert with their available resources (light, nutrients, and organic constituents). The phytoplankton community may or may not be able to compensate for grazing losses in overall primary production, but generally is able to shift quite quickly to an alternative, less vulnerable species composition.

III. ECOSYSTEM FUNCTIONS: INTERACTIVE REGULATORY MECHANISMS

Most of the organic carbon of aquatic ecosystems, and hence much of the energy potentially available for me-
tabolism of heterotrophic biota of the ecosystem, exists as dead organic matter (detritus), in both dissolved and particulate forms. Dissolved organic matter and much of the particulate organic matter, usually expressed as organic carbon, move with the water. Much of the particulate organic matter that is not decomposed ultimately settles by gravity and is deposited at the bottom of static water bodies. Dissolved organic compounds can also sediment if adsorbed to inorganic or organic particulate matter or if polymerization and aggregation occur into particulate form. Because much of the highest photosynthetic production of organic matter occurs in shallow regions of lakes and streams, and because most lakes are small and shallow, most heterotrophic decomposition occurs in benthic regions as organic matter. Therefore, because much of the organic matter is displaced to the sediments, most of the total heterotrophic decomposition is also largely displaced from sites of production to the sediments.

The specific composition of organic matter varies greatly. The organic material of pelagic phytoplankton, which can dominate in-lake production in very large, deep lakes, is relatively labile. Much of the phytoplanktonic algal and cyanobacterial organic matter is decomposed by microbiota either in the water column within the water strata of production or below as it sediments to the bottom. Some highly variable portion, usually much less than 50% on an annual basis, of the phytoplanktonic organic matter is ingested and partially digested by zooplankton. An appreciable portion, usually >50%, of that particulate organic matter ingested by zooplankton is egested as dissolved and as particulate organic matter, some of which also settles as feces to the sediments.

The composition of organic matter of the wetland and littoral regions of lakes and streams tends to be considerably more recalcitrant to rapid microbial degradation than that of phytoplankton because of the littoral dominance of lignocellulose-based organic compounds of structural tissues of higher plants. Because of the much higher rates of production and slower rates of decomposition of both the particulate and dissolved organic matter from these shallow-water sources, and because of the shorter distances and times for particulate organic matter to be displaced to the sediments, organic matter accumulates on and in the sediments at greater rates in shallow regions than in those of the pelagic region. The deposition of organic matter in the sediments encourages intensive bacterial metabolism. As a result, intrusion of dissolved oxygen from the overlying water is consumed rapidly; nearly all sediments in standing freshwater ecosystems are anaerobic and highly reducing chemically.

Although the specific composition of organic matter varies greatly, detritus inevitably carries most of the energy of the ecosystem from its points of photosynthetic origins to places of transformation by heterotrophic organisms. Most of that heterotrophic transformation occurs by microbes in areas where organic matter is concentrated. Even though greater than 70% of the
very large pool of dissolved organic matter of the pelagic region of lakes and rivers consists of relatively recalcitrant humic and fulvic acids that originated from higher plants, these compounds are slowly metabolized by bacteria, often being catalyzed by partial photolytic degradation by ultraviolet light (both UV-A and UV-B). In contrast, much of the particulate organic matter of the productive littoral areas accumulates on the bottom, largely under anaerobic conditions particularly in sediments and in detrital mats.

Transformation of particulate organic matter of the pelagic zone also occurs by digestion in predatory metazoan animals as well as by pelagic microbes. Productivity of the biomass is maximized within the resources available from dead organic matter and nutrients. Microbial heterotrophic utilization is highly dynamic and changes with sufficient rapidity to allow many generation turnovers before predators with much slower turnover rates can respond productively to changes in availability. Viral mortality of microbes, as well as ingestion of bacteria by protists (the “microbial loop”), diverts much organic matter from animal trophic levels. Bacterial, viral, and much of protistan heterotrophic metabolism as respiration represents a major output of carbon, largely as CO$_2$, from the aquatic ecosystem. Although this output is a loss of carbon from the ecosystem, it is not an energetic loss from the ecosystem.

That productivity that is not ingested by metazoan animals is of major importance and dominates the energy flows within the pelagic region. Nonpredatory death and metabolism by prokaryotic and protistan heterotrophs usually dominate energy flows within the pelagic region and completely dominate energy flows of the composite wetland–littoral, benthic, and pelagic components that collectively constitute freshwater ecosystems. Even ignoring the benthic portions of lakes where most of the organic carbon metabolism occurs and considering only the pelagic metabolism of lakes, for example, most of the organic carbon entering the system does not reach higher trophic levels.

The microbial heterotrophy is of primary importance to higher trophic levels by means of feedback processes, both positively (e.g., nutrient recycling and utilization by primary producers) and negatively (e.g., oxygen consumption and production of toxic fermentative metabolic end products). The abundance, distribution, and microbial decomposition of dissolved and particulate detritus both regulate and stabilize energy metabolism and nutrient availability in aquatic ecosystems. Because of the very large magnitudes and relative chemical recalcitrance of these detrital sources, the large but slow metabolism of detritus provides an inherent ecosystem stability that energetically damps the ephemeral, volatile fluctuations of numbers and biomass of the biota of higher trophic levels.

IV. BIODIVERSITY WITHIN FRESHWATER ECOSYSTEMS

The biodiversity of most microbial, plant, and animal groups of freshwater stream, lake, and wetland ecosystems is very poorly known on a global basis. Moreover, it is likely erroneous to believe that biodiversity of many taxa in temperate ecosystems is appreciably less than that in tropical freshwater ecosystems. For example, the rivers and streams of Alabama, a “hot spot” within the United States, contain 43% of all gill-breathing snails, 52% of freshwater aquatic or semiaquatic turtles, 60% of mussels, and 38% of freshwater fishes of North America (Lydeard and Mayden, 1995). Over half of the species in some of the groups mentioned are either threatened or endangered under the U.S. Endangered Species Act of 1973.

The fresh waters of the world are collectively experiencing markedly accelerating rates of degradation. Major sources of disturbance impact the biodiversity of freshwaters in many ways. Direct chemical toxicants released into surface waters and ground waters are common. Many forms of heavy metals, inorganic reducing agents, and organic compounds enter the environment and eventually fresh waters. Although some are inactivated, such as by chemical precipitation, or oxidized, many have long residue resident times. Despite dispersion and dilution in the aquatic environment, biocentration of both metals and organic compounds is common, by which toxicity can be increased exponentially. The biological effects of many of these compounds are unknown. Elemental and compound toxicity is increased markedly when the pollutant substances are radioactive or acidic. Increased acidity in poorly buffered fresh waters can increase concentrations and potential biotic availability and toxicity of metals, such as aluminum, that are not normally abundant in soluble, reactive states.

Nutrients, particularly phosphorus and nitrogen, can lead to well-understood enhancements of plant and other organic productivity in fresh waters. This eutrophication process often results in enhanced rates of decomposition and in chemical conditions that greatly reduce or eliminate suitable habitat for many species of plants and animals. Similar excessive loading of organic
matter and enhanced rates of degradation result from organic sewage from human populations, industry, and agriculture. A further common pollutant that markedly reduces habitat availability is the suspension of finely divided organic and inert inorganic matter. Erosion and transport of such suspensions are increasing as large areas of forest and former wetland interface zones between land and water are eliminated, primarily for agricultural expansion. As a result, flow patterns of surface waters are often altered and benthic habitats of surface waters can be obliterated by sedimentation.

Changes in biodiversity in freshwater ecosystems can arise from many other disturbances. Introduction of certain competitively superior species can result in marked losses of biodiversity. The infamous example of introduction of the Nile perch into Lake Victoria of East Africa resulted in the extinction of over 200 species of its endemic cichlid fish taxa in two decades. There are many other examples of introductions of exotic plant and animal species that resulted in either direct destruction of prey or inferiorly competitive species or indirect alteration of habitats required by many species. Dense, floating macrophyte communities and other eutrophication-associated excessive plant productivity often result in deoxygenation and reduction in habitat and elimination of many plant and animal species.

Largely based on terrestrial studies, the Eltonian diversity–stability hypothesis suggested that, because of the many different traits of multiple species, ecosystems with more diverse habitats would likely have species that will survive and expand during and following an environmental disturbance and compensate for those species that are reduced by the disturbance. Therefore, a more species diverse freshwater ecosystem should be more resilient to disturbances than a less biodiverse system.

Because genetically based physiological differences and tolerances among species can be small, the individual interactive strengths of some species in a freshwater ecosystem can become saturating at high biodiversity. A point can be reached where increasing species may be functionally redundant and have reduced individual impact on the ecosystem processes. On the basis of both theoretical and experimental grounds, only a small fraction of species manipulations have strong influences on food web structure. Species redundancy implies that an appreciable functional resiliency exists in which the ecosystem can compensate in its collective metabolism and biogeochemical cycling when disturbed. Although the population dynamics become progressively less stable as the biodiversity and the number of competing species increases, biodiversity can enhance the resiliency of many community and ecosystem processes in the rate that the system metabolism returns to equilibrium states following a disturbance.

There is very little storage capacity for organic carbon within the higher trophic levels. Low residence times among the higher trophic levels results in rapid cycling of carbon and nutrients of food web components. Such rapid cycling and recycling result in a reduction in the resiliency of the higher trophic levels. Most of the storage of organic carbon occurs in the dissolved organic carbon compartment in the open water and in the particulate organic carbon deposited in the sediments. In both of these compartments, the soluble organic carbon of the pelagic areas of lakes or running water of streams and the organic carbon of the sediments, the cycling of carbon is slowed. That rate of cycling is slowed in the pelagic by the recalcitrant chemical composition of the dissolved organic carbon emanating largely from higher plants. In the sediments, cycling is further impeded by the anoxic conditions that prevail almost universally among aquatic sediments. The reduced rates of cycling and recycling result in an inherent increase in resilience stability of the ecosystem.

Any factor that influences the rates of nutrient and carbon cycling in freshwater ecosystems will influence the resilience of the ecosystem and its biodiversity to disturbances. Changing sources of organic matter, as discussed in the following, and hence bacterial metabolism and nutrient cycling thus change resilience and biotic stability.

A wealth of limnological data from a spectrum of hundreds of lake ecosystems of differing productivity suggests that with a shift in nutrient loadings, concomitant shifts occur in the development of photosynthetic producers and loadings of organic matter. During the common sequential development of lake ecosystems over long periods of time (centuries, millennia), shifts in the ratios of higher vegetation versus algal dominance can occur. Increased relative organic loading from higher vegetation results in proportionally greater loading of recalcitrant dissolved organic carbon, which can suppress nutrient cycling and increase the resilience of the ecosystem.

In addition, the development of higher vegetation in littoral and wetland combinations increases the habitat heterogeneity enormously, often by a factor of 10 or more, in comparison to lakes with limited littoral development. Species diversity nearly always increases under these circumstances by at least an order of magnitude among nearly all major groups of organisms, particularly among the lower phyla.
Greater biodiversity may have a greater collective effect by improving the capacity of ecosystems to recover from large disturbances. Reduced biodiversity increases vulnerability by reducing the total collective physiological tolerances of the community to large habitat changes. Recovery after a major or catastrophic disturbance would be slower with a reduced aggregation of residual physiological ranges within the remaining species. Recovery then must depend to a greater extent on slower fortuitous methods, such as importation of species, rather than generation from residual surviving species, or slow recolonization processes such as from remnants in resting stages or seed banks. In some cases, such as in many ponds, streams, and reservoirs in clay-rich regions where high turbidity often occurs with successive rain events, photosynthetic productivity within the water is intermittently but repeatedly suppressed. Biodiversity is likely also suppressed under these conditions or restricted to species with high reproductive potential that utilize improved conditions in periods between turbidity events.

As indicated earlier, disturbance to freshwater ecosystems can occur in many forms and to different extents. Certain perturbations can be catastrophic, such as overwhelming a lake or stream with an organic or inorganic poison in which most of the biota are eliminated. Many disturbances, however, are more gradual over long periods of time (months, year), such as nutrient enrichment, or irregularly episodic and often of short duration (days, weeks), such as severe flooding and the scouring of a section of river. Biodiversity is coupled to ecosystem stability and the type and extent of disturbances.

A model of the responses of organic productivity of lake ecosystems to changes in nutrient loading from the drainage basins has been abundantly verified by nutrient and comparative primary productivity data of phytoplankton, attached algae, and macrophytes from hundreds of lakes in different stages of ontogeny (Fig. 4). The differences in plant productivity result in very different amounts and types of chemical composition of organic matter loaded to lakes. Because of the large amounts and relative chemical recalcitrance of dissolved and particulate organic detrital sources from higher plants, decomposition of this pool is slowed. The resulting large but slow metabolism of organic detritus provides an inherent ecosystem stability that energetically buffers the organic matter reserves from rapid exploitation and cycling of nutrients contained in the organic matter. Forested watersheds, undisturbed riparian floodplains, and wetland–littoral zones are therefore important to the metabolic stability within the lakes and streams as a result of both the chemical quality and the large quantity of their inputs of organic matter. The organic carbon couplings between the drainage basin and land–water interface zones are metabolically as important to lake and stream energetic and carbon utilization stability as are the traditionally studied nutrient (phosphorus, nitrogen, silica) loading relationships to food web productivity. As the sources of organic matter change, the chemical recalcitrance to degradation can change. Alterations of rates of utilization, nutrient recycling, and energetic resiliency can ensue as a result. These changes directly affect biodiversity by altering the chemical and physical habitat characteristics.

A common disturbance to lake and reservoir ecosystems involves progressive increases in loading of nutrients. These loadings, in excess of losses to sites of temporary or permanent inactivation such as to the sediments, result in enhanced nutrient availability for phytoplankton and other autotrophs and increased

![Relative Primary Productivity](image)

**FIGURE 4** Relative changes in the primary productivity of phytoplankton, macrophytes, and attached microflora, and habitat diversity, species diversity, and ecosystem stability along a gradient of nutrient loading to lake ecosystems. (From Wetzel, 1999.)
rates of growth and productivity. If nutrient loading increases to an oligotrophic lake, increased productivity is rapid. Similarly, if the disturbance is brief (i.e., the duration of increased nutrient loading is relatively short), nutrient cycling is rapid, the ecosystem will recover rapidly, and productivity would be reduced proportionally to the load reduction. The return time of such an oligotrophic ecosystem is high, but the resiliency is low. Species diversity of the plankton tends to be high in oligotrophic lakes, but because of limited physical habitats, particularly associated with littoral areas and surfaces, composite biodiversity of the lake ecosystem would be low (see Fig. 4).

As nutrient loading increases, particularly among shallow lakes that predominate globally, a marked shift in the productivity occurs from the pelagic to attached surfaces associated with living aquatic plants and the particulate detritus of senescing macrophyte biomass (see Fig. 4). Under these conditions, total primary productivity and biomass increase greatly. Habitat diversity among the massively dissected surfaces of submersed aquatic plants increases exponentially, and biodiversity among attached biota would increase by at least an order of magnitude. It is likely that among all autotrophic and heterotrophic microbial groups, as well as for most of the smaller metazoans, over 90% of the total freshwater ecosystem species are sessile in association with surfaces.

As nutrient loading increases further, phytoplanktonic productivity per unit volume increases, but self-shading by the suspended algae restricts the depth of light penetration and hence the depth of the photogenic zone. Phytoplanktonic productivity per unit area declines precipitously under these eutrophic conditions of light limitation and is usually accompanied by a marked decrease in planktonic biodiversity as well (see Fig. 4). Submersed light limitations also reduce or eliminate submersed macrophyte growth and attendant surfaces for microbial growth. The elimination of photosynthesis of attached microbiota decreases collective productivity markedly. Biodiversity of the attached biota would decline precipitously as a result. The losses of attached microbial communities and their interdependent metabolism cause a massive reduction in the capacities of the lake ecosystems to retain loaded nutrients and dissolved organic compounds. Because of the great accumulation and recycling of nutrients in the sediments from a predominantly planktonic eutrophic ecosystem, the time for recovery of the water quality of a lake from a reduction of nutrient loading would be long and not proportional to load reductions. Under these conditions after the nutrient loading perturba-

tions end, the internal nutrient loading and recycling could be greater than external nutrient loading.

Continual high nutrient loading and resulting hyper-eutrophic phytoplanktonic conditions generate large areas of anaerobic reducing conditions and slower rates of decomposition of organic matter. Production exceeding decomposition leads to rapid sedimentation and generation of increased shallow habitat conducive to colonization by emergent macrophytes. The high productivity of emergent macrophytes increases the proportion of lignocellulose supporting tissues that are relatively recalcitrant to rapid decomposition, particularly under reducing conditions. The collective result is markedly increased productivity and habitat diversity. However, although habitat diversity is high among very shallow wetland-dominated waters, environmental fluctuations are also more extreme than in continually submersed habitats. Quite different community structures occur and biodiversity is likely tempered as a result, though comparative data to support this statement are few.

Many descriptive and experimental studies suggest that greater biodiversity results in a higher efficiency of utilization of nutrients. Greater biodiversity likely entails more physiological diversity that can better cope with natural vagaries in environmental parameters. Competition for resources, nutrients in this specific case, is thus intensified. Efficiency of nutrient retention in an ecosystem would be maximized under conditions of greatest microbial community diversity. That microbial diversity (algae, fungi, protists, and bacteria) is maximized in the attached communities where habitat diversity and microenvironment differences are the greatest. Physical constraints of boundary layers and mucopolysaccharide matrices in which the attached microbiota live mean that movements of ions and gases to organisms living within the attached communities occur predominantly by diffusion. Nutrients, once acquired, are intensively recycled among the attached biota and conserved. Resources from external sources can then be utilized largely for new growth and reproduction. Nutrient retention is very high within the microcommunities and collectively within the attached habitats.

In certain productive lakes, piscivore consumption of planktivorous fishes can lead to enhanced development of cladoceran zooplankton. The high grazing rates of cladoceran zooplankton can result in selective reduction of larger algae for brief to moderate intervals of time. The effects of these changes among the larger forms are, however, poorly translated to smaller microbiota, and only minor changes are seen at the mi-
croalgal and bacterial levels. For example, removal of portions of the larger algae can decrease competition for nutrient and light resources. Often the smaller forms with shorter generation times increase in productivity. Nutrient recycling likely also increases, particularly as decomposition of these algae by bacteria is accelerated by protistan microconsumers (heteroflagellates, ciliates, and related organisms) and tightly retained among the microbiota. Little is known concerning changes in biodiversity under these conditions, but it appears that biodiversity among the microbiota would increase and likely compensate for the losses of larger forms by the crustacean grazing. A shift in energy fluxes occurs, here from larger planktonic forms to smaller forms with much higher turnover rates and increased rates of resource turnover. Compensatory mechanisms likely also appear in the composite biodiversity of the smaller biota. Resiliency to disturbances within the pelagic food web likely declines as a result.

See Also the Following Articles
FRESHWATER ECOSYSTEMS, HUMAN IMPACT ON • INVERTEBRATES, FRESHWATER, OVERVIEW • LAKE AND

POND ECOSYSTEMS • PELAGIC ECOSYSTEMS • WETLANDS ECOSYSTEMS

Bibliography
FRESHWATER ECOSYSTEMS, HUMAN IMPACT ON

Kaj Sand-Jensen
University of Copenhagen

I. Freshwaters: Opportunities and Conflicts
II. The Small, but Numerous Freshwater Habitats
III. Species Evolution and Richness in Freshwater
IV. Human Impacts on Freshwater Ecosystems
V. Biological Quality of Freshwater Ecosystems
VI. Conclusions

GLOSSARY

denitrification Bacterial nitrate respiration of organic matter to elemental nitrogen under oxygen-free conditions in soils and aquatic sediments.
detritus Dead organic matter.
oxic and anoxic conditions Environmental conditions with and without oxygen, respectively.
Red Lists Compilations of recently extinct, threatened, and vulnerable species for a country or a larger region.
Secchi-depth Measurement of the transparency of water by determining the depth at which a white Secchi-disc (20-cm diameter) lowered from the surface disappears out of the sight of the observer.

FRESHWATERS ARE INHABITED BY A REMARKABLE VARIETY OF SPECIES from most taxonomic groups evolved on Earth, but they are also a scene of human opportunities, requirements, and conflicts. Strong human impacts have restricted the number and variety of freshwaters and polluted and disturbed many others. This article discusses the range of human impacts and their consequences for species richness and functional diversity of individual lakes and streams. Special emphasis is placed on freshwaters in Europe and North America and their intensely cultivated regions. Biological invasions are discussed briefly; climatic changes are not dealt with because of a lack of general predictions.

I. FRESHWATERS: OPPORTUNITIES AND CONFLICTS

Why are we so interested in freshwater and natural freshwater ecosystems? The answer is obvious—we cannot live without it! Freshwater is essential to our personal daily life and the production of vital food and industrial products.

We use freshwater to drink and to prepare and cook our food. We use freshwater to wash ourselves and our clothes and to remove waste products. Securing clean tap water and good sanitary conditions through installation of water closets and sewer pipelines has been important in reducing infectious diseases and infant mortality. Daily life also includes the enjoyment of natural lakes and streams, which provide variation in the landscape and opportunities for bathing, fishing, sailing, and observations of animal and plant life. Therefore, it is of high priority to all human beings that clean water...
and unspoiled freshwater habitats are available in great amounts.

Freshwater is also required for irrigation of agricultural crops, and it is used in many industrial processes and to transport waste products. Because of the magnitude of these uses, the pressure on groundwater and surface waters is immense in regions with dense populations and intense agriculture and industry. Unfortunately, many of the uses deteriorate water quality, thereby affecting water use for other purposes. The growing number of uses and users generates conflicts among neighboring countries and different user interests, and further exacerbates the dilemma of rising demand and the need and desire for high-quality aquatic habitats. Overall, freshwater ecosystems have been more restricted, manipulated, and polluted than any other ecosystem on Earth, and conflicts will intensify in the future in the light of growing human populations and contemporary environmental trends.

II. THE SMALL, BUT NUMEROUS FRESHWATER HABITATS

Earth has much water. A deep, continuous ocean covers 70% of Earth's surface and contains 97% of all water. Only 0.63% of all water is found as freshwater on the continents, and most of this is groundwater (0.62%). Even smaller proportions are found in streams (0.0001%) and lakes (0.017%).

Lakes and streams cover a variable proportion of the land surface. In dry regions, surface waters cover less than 0.1% of the land surface. In wet environments such as the tundra, the boreal forest, and the rain forest, plenty of water is found in shallow pools, lakes, and streams, which may occupy 5–10% of the land surface. A map of Great Britain reveals that surface waters cover approximately 1 km² of land area and on average 1 km² of land area is in contact with 3.6 km shoreline of freshwater lakes and streams and 0.2 km of the sea, according to estimates from 1,250,000 maps. Before agriculture removed numerous shallow lakes, pools, and streams, the freshwater contact was probably two or three times longer.

The greater freshwater than marine contact to land is more striking for the world, because land areas are joined in large continents. As land-masses grow in size, the contact zone to the sea increases only with the periphery and, thus, with the square root of the land area, whereas the contact zone with inland waters approximately increases in proportion to the land area. For the continents, the estimated contact zone is 100–1000 times longer between land and freshwater than between land and sea.

The transition between land and freshwater is gradual, gentle, and suitable for organisms, because physical forces and disturbance are weaker than in the transition from land to sea. The transition zone from small lakes and streams to land has a closed vegetation cover; that of the oceanic coasts faces strong winds, waves, and moving sand with only scattered vegetation. It has been much easier for organisms to cross between land and freshwaters than between land and the sea. The many new freshwater bodies that are formed during glaciations or appear transiently during wet periods also offer opportunities for the emigration and development of new species without the intense competition and predation caused by well-established species, as is typical of the sea. On the other hand, shallow lakes disappear after only a few thousand years, because they are filled with particles eroded from the land and organic matter produced in the lake. In essence, freshwaters provide many opportunities, which constantly come and go, but they lack the long stability of the sea that has lasted for billions of years.

This scenario should have stimulated the selection of species with rapid evolution and efficient means of dispersal allowing them to colonize new freshwaters as their original habitats disappear. Indeed, apart from species of fish, molluscs, and crustaceans associated mainly with large ancient lakes (e.g., Lakes Baikal, Titicaca, Victoria, and Tanganyika) and large ancient river systems (e.g., the Danube, rivers of southeastern North America, the Amazonas, and other tropical rivers), most freshwater species are widespread within and even among continents. Freshwater microorganisms among the bacteria, algae, and protozoans are generally both locally abundant and cosmopolitan. The same species of cyanobacteria and microalgae live in extreme and
rare freshwater habitats such as hot springs or highly acidic pools in distant locations on different continents and oceanic islands. Most aquatic rooted plants are also widespread, though a few species are endemic because underwater dispersal restricts their spread.

III. SPECIES EVOLUTION AND RICHNESS IN FRESHWATER

As a result of the intimate contact, the long coastline, and the suitable transition zone, there has been a lively exchange of species from freshwaters to terrestrial environments and back again. Plants have evolved from a special group of freshwater green algae (Charophyceae, Coleochaetales) and diversified under the highly variable conditions on land. Insects have evolved from groups of arthropods in the transition between freshwater and land. Amphibians and reptiles have evolved from special fish in freshwater and brackish wetlands. Many evolutionary lines have been followed at different times and places.

Freshwaters have formed a corridor for the two-way dispersal of organisms between land and sea. Freshwaters also share many taxonomic groups of algae, plants, and animals with the terrestrial environment and the sea, while divergences are stronger between the land and the sea. Freshwater environments are surrounded by large surface areas of terrestrial and oceanic environments, which promotes the emigration of marine and terrestrial species and their adaptation to freshwater environments.

Among major groups of land plants, several have secondarily returned to freshwaters, though few have reached the sea. Among liverworts, true mosses, bog mosses, horsetails, and ferns there are many freshwater species, but no marine representatives. Among flowering plants, 1000–1400 truly aquatic freshwater species have been described from many plant families, illustrating that the secondary return from land to freshwater has occurred independently and repeatedly. The return process continues today, with more than 4000 species living a double-life in the transition between land and freshwater. The sea includes only about 60 species of flowering plants with a restricted taxonomic diversity.

TABLE I

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Freshwater species</th>
<th>All species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyanophyta (blue-green algae)</td>
<td>1500</td>
<td>2000</td>
</tr>
<tr>
<td>Rhodophyta (red algae)</td>
<td>150</td>
<td>5000**</td>
</tr>
<tr>
<td>Chrysophyceae (gold algae)</td>
<td>1900*</td>
<td>2000</td>
</tr>
<tr>
<td>Xanthophyceae (yellow-green algae)</td>
<td>550*</td>
<td>600</td>
</tr>
<tr>
<td>Bacillariophyceae (diatoms)</td>
<td>5000</td>
<td>10,000</td>
</tr>
<tr>
<td>Phaeophyceae (brown algae)</td>
<td>Few</td>
<td>2000**</td>
</tr>
<tr>
<td>Cryptophyta</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Dinophyta (dinoflagellates)</td>
<td>200</td>
<td>2000**</td>
</tr>
<tr>
<td>Euglenophyta</td>
<td>700*</td>
<td>800</td>
</tr>
<tr>
<td>Chlorophyta (green algae)</td>
<td>7000*</td>
<td>8000</td>
</tr>
<tr>
<td>Zygmenophyceae (e.g., desmids)</td>
<td>6000*</td>
<td>6000</td>
</tr>
<tr>
<td>Charophyceae (stoneworts)</td>
<td>80</td>
<td>80</td>
</tr>
</tbody>
</table>

* Values are presented for major taxonomic groups. Mainly freshwater groups (*) and marine groups (**) are marked.
variability among individually confined water bodies and the suitable conditions for emigration and establishment of species from the large, adjacent terrestrial and marine environments.

Taking the vertebrates as an example, there are about 10,000 named species of freshwater fish and 15,000 species of marine fish. Among birds and mammals, virtually all depend on freshwater bodies for drinking, but a large proportion of the species also are dependent on freshwater lakes, streams, and wetlands for breeding and feeding. In Europe, about 23% of the bird species and 11% of the mammal species live their entire life or part of it in freshwaters.

A. Species in Freshwater Sediments

A global overview of organisms associated with freshwater sediments yields ca. 175,000 described species (Table II), but the true number of species is much higher than this. The most speciose groups in freshwater sediments are the invertebrates and especially the insects, nematodes, and crustaceans. Among nematodes and rotifers living between the particles in freshwater sediments, there are probably many thousands of undescribed species.

Compilation of local species richness and taxonomic diversity (see Table II) yields equally high, or often even higher, values in freshwater environments than in marine or terrestrial environments. There are no systematic differences in local species richness between lakes and streams, but upstream sections of streams lack the variety of phytoplankton, zooplankton, and fish that are typical of open waters, and instead support greater species richness in the surface sediments and in the deep sediments below the streambed. The downstream parts of streams have a greater resemblance to lakes, so the similarity of species composition and richness will depend on whether entire stream systems or just certain stream sections are compared with the lakes. The biota in porous groundwaters is much deprived in species due to lack of light, degradable organic matter, and dissolved oxygen. Specialized species of protozoans and small invertebrates live here in local species numbers that are typically 10 to 20-fold lower than those encountered in lakes and streams (see Table II).

### IV. HUMAN IMPACTS ON FRESHWATER ECOSYSTEMS

When evaluating the human impacts on freshwaters, we tend to concentrate on the numerous examples of water pollution. Over the last 150 years, the environmental issues have gradually changed as new problems have appeared and become recognized. Since the mid-1800s, organic pollution of streams and lakes with organic wastes from households and domestic animals has been of major concern in Europe and North America. Since 1945, the focus has been on cultural eutrophication of inland and coastal waters with nitrogen and phosphorus from agriculture, towns, and industries. After 1970, acidification of inland waters came onto the agenda due to increasing concentrations of sulfuric and nitric acids in the precipitation and changes in land use. The latest chemical concerns include trace metals and an enormous range of synthetic organic compounds of largely unknown behavior and ecological effect.

Many physical changes in the catchments are, however, much more important to the existence, environmental quality, and biodiversity of surface waters than the direct water pollution. The most significant influence on terrestrial and freshwater environments is the removal of natural vegetation and the cultivation of land, which lead to immediate, profound alterations of the hydrology and nutrient cycling. These alterations are grossly enhanced when soils are drained and streams are canalized. The intimate linkage between natural wetlands and streams, which has been important for the evolution and contemporary diversity of plant and animal life, is also disrupted when surplus water directly flows to the stream through drain pipes rather than slowly percolating through the wet, sponge-like organic soils.

### Table II

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Global</th>
<th>Lake/stream</th>
<th>Groundwater</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td>&gt;10,000</td>
<td>&gt;1,000</td>
<td>&gt;100</td>
</tr>
<tr>
<td>Fungi</td>
<td>600</td>
<td>30–300</td>
<td>0–10</td>
</tr>
<tr>
<td>Algae</td>
<td>14,000</td>
<td>0–1,000</td>
<td>0</td>
</tr>
<tr>
<td>Plants</td>
<td>1,000</td>
<td>0–100</td>
<td>0</td>
</tr>
<tr>
<td>Protozoans</td>
<td>&lt;10,000</td>
<td>100–400</td>
<td>0–20</td>
</tr>
<tr>
<td>Crustaceans</td>
<td>8,000</td>
<td>25–130</td>
<td>3–60</td>
</tr>
<tr>
<td>Insects</td>
<td>43,000</td>
<td>30–300</td>
<td>0–10</td>
</tr>
<tr>
<td>Molluscs</td>
<td>4,000</td>
<td>0–30</td>
<td>0–10</td>
</tr>
<tr>
<td>Other invertebrates</td>
<td>&gt;12,000</td>
<td>30–70</td>
<td>5–70</td>
</tr>
</tbody>
</table>

* Data compiled by Palmer et al. (1997).
The impact of biological invasions is viewed as being much more critical in North America than in Europe, probably because North America has been disturbed more recently by the agricultural settlement of European colonists, and because it includes a greater variety of endemic fish and large invertebrates that survived the latest glaciations in refugia in the U.S. Southeast and Mexico. Allan and Flecker (1993) discussed the serious threats of biological invasions to biodiversity in temperate (e.g., New Zealand) and tropical countries.

Most tropical countries now face the unregulated impact from all pollution sources. The most widespread problem appears to be the rising organic pollution of streams and lakes by untreated domestic sewage. However, in some regions heavy pollution takes place from (1) extensive application of pesticides and nitrogen in plantations of bananas, cocoa, and oil palms, (2) large oil spills (e.g., Ecuador, Venezuela, Nigeria), (3) acidification and pollution with heavy metals (e.g., copper, mercury) from mining areas and tanneries, and (4) outlet of phenolics from the timber industry. A fundamental problem is the erosion of soils deriving from the massive deforestation of tropical rain forests. Many streams have become chocolate colored from the heavy load of solids eroded by the strong rainfall. Erosion is a severe problem in agriculture, which loses the fine, fertile top-soils. It also causes environmental problems as hydroelectric dams are quickly filled with sediments, and when suspended loads in streams prevent light penetration and thereby the growth of algae, which are food for many invertebrates and fish. Furthermore, stream bottoms get clogged with fine sediments that destroy the habitats of invertebrates and the spawning banks of fish.

In many tropical countries, physical alterations of streams are still relatively few. In Ecuador, for example, streams still run in natural streambeds with meanders and rapids surrounded by strips of riparian vegetation even in densely populated areas. The hydrological cycle has been altered, however, since water is removed for irrigation, leading to artificially low discharge and current velocity during the dry season. As a consequence, the stream biota is impoverished due to oxygen depletion and smothering of the streambed following organic pollution.

A. Restriction of Area and Variability of Freshwater Habitats

The scene of physical changes and areal reductions of freshwater ecosystems changes among countries and with time. Changes of freshwater ecosystems have been greatest over the last 100–200 years in densely populated countries and in regions with intense agriculture and industry. Thus, the combination of population density, resource use, and powers of technology is a suitable measure of human impact on the biosphere in general and the freshwaters in particular.

Pets (1984) defined four phases in the recent era of river modification in Europe. Phase 1, from 1750 to 1900, includes ambitious regulations of the major rivers for the purpose of navigation, flood control, and cultivation of the river valley. Phase 2, from 1900 to 1950, marks the first major technological period, during which large dams and power plants were built across major rivers. In many European lowlands, extensive drainage of wetlands and shallow lakes and channelization of streams took place, and continuous management of the dimensions of stream channels and cutting of aquatic plants were initiated. Management intensified in phase 3 from 1950 to 1980 by the use of specialized machines. During the recent phase 4, from 1980 onward, the intensity of regulation works and dam building has gone down, because most watercourses have already been exploited and public resistance has increased because of rising environmental concerns. However, most countries of the world are still in the most exploitive phase 3.

An overview of the 69 major rivers in Europe and Russia shows that most of them are strongly (80) or moderately (10) affected by dams and regulations, whereas only 19 rivers located in Arctic and northern boreal regions have remained relatively unaffected. River regulation has been undertaken to the greatest extent in western and southern Europe. In Belgium, Denmark, England, and Wales, the percentage of river reaches that are still in a natural state is less than 20%.

In southern Canada and the United States, rivers are heavily regulated; those in northern Canada and Alaska have remained relatively pristine. The history of Willamette River in Oregon is an example of how expansion of agriculture and construction of 11 dams have transformed a complex multi-channeled river into a simple one- or two-channeled river. Over a 25-km-long stretch in the floodplain, the length of the shoreline has declined from 230 km in 1854 to only 64 km in 1967. In addition to the shortening of rivers and streams, watercourses have become more uniform and physically disturbed. When streams are channelized, the natural variability in depth, width, current velocity, and sediment composition disappears between straight reaches and meanders and between riffles and pools. The natu-
eral dynamics of the flow channel, characterized by spatial variations of erosion and sedimentation, formation of new meanders, and the cut-off of oxbow lakes, also vanishes. Drainage of the floodplain reduces its storage capacity of water and the ability to buffer storm surge in the rivers. Moreover, drainage reduces baseflow during dry periods. Peak discharge during rainy periods introduces high physical stress and erosion in the stream channel, and low baseflow during drought increases the risk of high water temperatures, oxygen deficiency, and insufficient currents for the well-being of invertebrates and fish. All three major changes—(1) smaller area and length of the streams, (2) lower habitat diversity, and (3) greatly fluctuating flow and higher physical, oxygen, and temperature stresses—will drastically reduce species number and population density of plants and animals.

Numerous shallow lakes and ponds in European farmland have disappeared from drainage and filling-in. Many ponds were originally dug to extract calcareous sediments for the liming of acidic fields, clay for bricks, and sand and gravel for construction works and to provide watering places for cattle. In some regions more than half of the ponds that were present 150 years ago are now gone. As a result, several species of amphibians and large insects that preferred the warm, shallow pond appearance due to lack of water, filling in, or overgrowth by littoral vegetation. New ponds have been created in farmlands with the purpose of regaining the patchwork of small, suitable biotopes for aquatic plants, insects, and amphibians. Populations of several threatened species of frogs and newts have since increased thanks to the reestablishment of suitable ponds.

C. Changes of the Hydrological Cycle

It is often over-looked how much the hydrological cycle has been changed by humans and the environmental implications of these alterations. Archeology, paleoclimatology, and historic studies of old maps and written records in Europe reveal that the groundwater level was much higher and there used to be many more open waters in the past compared to today. Previously, water moved slowly and diffusively through the large water reservoirs in the soil and the wetlands before reaching streams and lakes. Following cultivation of the soil and establishment of ditches and drains, the water has taken a more direct and rapid course from the fields to the streams, resulting in huge variations in discharge. Water for irrigation of crops and domestic supply to metropolitan cities is abstracted from groundwaters and surface waters, and this water demand has markedly increased the risk of wetlands drying up and critically low flows developing in the streams during dry periods.

Breaking the long contact between water and soil...
has dramatically reduced the natural purification of water by preventing the removal of surplus nitrogen and phosphorus through binding to soil particles, incorporation in new organic material, precipitation as phosphate minerals, and denitrification. In addition to the increased cultivation of crops, application of fertilizers, and discharge of sewage, changes in the hydrological cycle have resulted in the explosive eutrophication of inland and coastal waters. A few countries have tried to reestablish some of the wetlands that have been lost so extensively, thereby reinstalling the cleaning sponge between cultivated soils and streams. In most other countries, however, wetlands continue to disappear at a high rate. Spain, for example, has lost two-thirds of its wetlands since 1965 with the support of European Union subsidies to farmers.

D. Organic Pollution of Inland Waters

The discharge of organic wastes from towns and from livestock operations creates the classic pollution problem of inland waters. Human wastes are derived from bathing, cooking, laundry, and the flushing of feces and urine in lavatories, and discharges from agriculture include manure from the animals and food spills. Intense organic pollution can also occur from breweries, dairy factories, slaughterhouses, sugar refineries, and countless other industrial sources. In regions where water quality otherwise is high, the farming of fish, prawn, and crayfish can be common. The trout farms of northern Europe mainly use cold spring water of superb quality, but these farms in turn pass on substantial organic pollution from food spills and fish excreta.

With the increase in domesticated animals, human populations, and the installation of water closets and sewers, there has been a profound increase of organic loading of inland waters in Europe and North America from 1850 to 1950. In several cases, the external loading of lakes has increased more than 10-fold (Fig. 1). Substantial purification of domestic sewage has been established over the last 50 years. Meanwhile, wet slurry from burgeoning animal farms has presented a new source of pollution. In poor countries of the world, organic pollution has continued to grow and few attempts have been made to ameliorate the problem through the use of natural wetlands with self-purification capacities or the construction of costly sewage treatment plants.

The organic pollution of streams increases with the density of human beings and animal livestock and the consumption of oxygen by organic matter in the water. An acceptable water quality can be reached by extensive purification, as shown by a 10-fold reduction of the biological oxygen demand at the outlet from some sewage treatment plants and a 3-fold reduction in the stream water in Denmark from 1975 to 1995. Otherwise, the European inland waters with the best water quality in terms of low oxygen demands and high oxygen concentrations are found in the water-rich regions in the north, where population densities are low and large areas remain uncultivated. In contrast, streams with the poorest water quality are located in intensely cultivated regions of middle and southern Europe, where purification is insufficient and water is in short
supply during the summer. To illustrate the differences in environmental state, oxygen conditions are regarded as good in 97–99% of the streams in Scotland and Iceland and 64–77% of the streams in Wales, but in less than 20% of the streams in Belgium, Poland, and countries in the Balkans. Most European countries have experienced an overall decline of organic pollution in the 1980s and 1990s.

E. Agricultural and Industrial Chemicals

An immense variety of chemical products are being manufactured, used, and released by agricultural and industrial activity. Some inorganic compounds include acids, alkalis, ammonia, chlorine, radio nuclides, and heavy metals (e.g., cadmium, copper, iron, mercury, and zinc). Organic compounds are grouped under different names such as chlorinated hydrocarbons, hydrocarbons, pesticides, phthalates, and phenolic compounds. Perhaps 20,000–50,000 substances are manufactured or applied within industrial countries, and a few thousand are added each year. In most European countries, between 120 and 530 active pesticides are approved for agricultural use today. The annual usage is usually between 1 and 14 kg per hectare (ha) of arable land. Substances that are applied within a country are also detected (or their degradation products) when tested for in surface waters and groundwaters. The most common pesticides in groundwaters (i.e., atrazine, deethylatrazin, lindane, and simazine) are often found in concentrations exceeding the maximum allowable threshold (0.1 µg L⁻¹ in the European Union).

A major obstacle to pollution control is the need to first recognize the presence of potentially harmful substances in potentially harmful concentrations in relevant ecological situations. This requires a lot of money, appropriate chemical skills, and advanced analytical methods to conduct an adequate survey of the distribution and concentrations of just a small number of these trace organic chemicals. It is then even harder to evaluate the biological consequences under natural conditions. As more traditional and obvious pollution problems in developed countries are stabilized or reduced, these organic substances may perhaps become the key pollution problem in many water bodies. The pollution effects may become more apparent as the application of these new organic compounds spreads and they accumulate over time in groundwaters. However, because their biological effects are subtle, chronic, and extremely costly to verify in the complex blend of numerous compounds, living organisms, and environments, the only effective solution to the pollution problem is not to discharge the pollutants at all, or reduce their toxicity, magnitude, and rate of application as far as possible. In less-developed countries, it will be impossible to monitor the release and biological effects of these trace organic compounds. Some frightening pollution events have been reported from the heavy misuse of pesticides in tropical crop production.

F. Cultural Eutrophication

Cultural eutrophication—predominantly due to increasing loads of nitrogen and phosphorus—leads to profound changes in the composition, biomass, and productivity of algae and plants. Lake eutrophication results in phytoplankton blooms, untransparent water, and oxygen deficiency. Eutrophication spoils the quality of bathing water and threatens the survival of bottom animals and fish. Algal blooms can include toxic algae and thereby harm animal life and become a public health risk.

Eutrophication of streams can also enhance the growth of attached macroalgae and flowering plants. Most European lowland streams have long passed the threshold at which nutrient concentrations limit plant growth. Streams in sparsely populated regions of northern Europe and Canada can still hold such low concentrations that plant growth is enhanced at sites of elevated nutrient input. Stimulation of algal growth by eutrophication is also very important in most tropical streams. Nutrient input to watercourses has increased dramatically during the last 150 years, and it has intensified over the last decades. The sources of nitrogen and phosphorus input include (1) towns and industries, (2) scattered settlements, (3) agriculture, and (4) a background input deriving from precipitation and runoff from uncultivated areas. Input from towns and agriculture usually dominates the overall nutrient budget, but all four sources have increased because of anthropogenic impacts.

The classic examples of lake eutrophication have been documented in the vicinity of metropolitan towns. In the case of Lake Fure close to Copenhagen, annual phosphorus input has increased 30-fold from 1900 to 1969 due to an eightfold increase of the population density in the catchment, the installation of sewers, and the use of phosphorus-rich detergents (see Fig. 1). Agriculture is the other major source of nutrient loading. Paleolimnological studies document the increase of accumulation rates of mineral particles, phosphorus, and organic matter in lake sediments due to increased erosion and runoff following the early cultivation of watersheds. Though nutrient input with domestic sewage has recently declined in some industrial countries thanks to tertiary sewage treatment, agricul-
FRESHWATER ECOSYSTEMS, HUMAN IMPACT ON

FIGURE 2 Rise of nitrate concentrations in selected English rivers and groundwater reservoirs between 1920 and 1990. (Data compiled by Heathwaite, Johnes, and Peters, 1996, and adapted from Moss, 1998.)

Natural sources of nitrogen have either leveled off or continued to increase because of the heavy application of fertilizers. A large proportion (50–80%) of added nitrogen is washed out as nitrate from the soil or released as gaseous ammonia to the atmosphere. Measurements in England between 1920 and 1980 show a 2- to 6-fold increase in nitrate concentrations in a series of streams and groundwater reservoirs over periods of variable duration (Fig. 2).

An assessment of European streams reveals that nitrate concentrations have increased from 1970 to 1990 because of the strong agricultural input, while phosphate concentrations have declined thanks to better cleaning of domestic and industrial sewage (Kristensen and Hansen, 1994). With unregulated application of phosphorus fertilizers, phosphorus pools in agricultural soils have increased, and phosphorus will eventually get lost to inland waters by erosion of soil particles and by leaching as the binding capacity to soil minerals is surpassed. This development represents a new risk to the control of lake eutrophication in countries attempting to control phosphorus through tertiary treatment of domestic sewage. In streams exposed to strong human impact, median concentrations of nitrogen and phosphorus are typically 10-fold higher than in the few pristine streams in mountain ranges and forest regions. However, even oligotrophic waters in uncultivated areas have experienced a 2- to 6-fold enrichment of nitrogen through atmospheric deposition since 1950 in most regions of Europe and North America.

G. Anthropogenic Acidification

Anthropogenic acidification of precipitation and surface waters commenced with the Industrial Revolution, but accelerated in Europe and North America after the 1950s. Burning of fossil fuels releases sulfur oxides, which are converted to sulfuric acid in the atmosphere. High combustion temperatures in car engines and power plants release nitrogen oxides, which are converted to nitric acid. Although the use of cleaner fuel gases reduced sulfur emission in the 1980s and 1990s, emission of nitrogen oxides from car traffic has continued to increase.
of physical and chemical influences, whereas major Streams have presumably experienced the entire range cycle, physical disturbance, and water pollution. to loss of lakes and streams, changes in the hydrological Anthropogenic impacts on freshwater ecosystems lead had significant local effects.

Acidification is substantial in the northeastern United States and southeastern Canada. Acidified regions of Europe include large parts of Finland, Norway, and Sweden, mountains in Great Britain, Poland, and the Czech Republic, and carbonate-poor, well-leached sandy soils in Belgium, Denmark, and the Netherlands. In Norway and Sweden, acidification represents a significant national problem as thousands of inland waters have been acidified and lost their resident invertebrate species and fish. Vast forested areas have also been acidified to the extent that species diversity and elemental cycling have been grossly disturbed.

The biological effect of acidification is in part due to the direct reduction of pH, which is intolerable to many organisms. Associated effects appear from an altered ionic balance, high metal concentrations, and immobilization of vital nutrients. Acidification has also affected a large proportion of oligotrophic softwaters in uncultivated, remote areas. Many freshwater habitats that have escaped drainage and eutrophication have instead been acidified.

The best way to reduce acidification is to lower the emissions of sulfur dioxide and nitrogen oxides through national and international agreements rather than trying to reverse the problem locally. Thus, the reduction of sulfur emissions appears to have reversed acidification in many regions in the 1990s. Also, a remarkable national scheme in Sweden to lime sensitive lakes has substantially reduced emission of sulfur oxides and ammonia in the plant canopies and the strength of acidifying or buffering processes by chemical weathering and biological transformations.

Acidification is substantial in the northeastern United States and southeastern Canada. Acidified regions of Europe include large parts of Finland, Norway, and Sweden, mountains in Great Britain, Poland, and the Czech Republic, and carbonate-poor, well-leached sandy soils in Belgium, Denmark, and the Netherlands. Acidification has also affected a large proportion of oligotrophic softwaters in uncultivated, remote areas. Many freshwater habitats that have escaped drainage and eutrophication have instead been acidified.

The best way to reduce acidification is to lower the emissions of sulfur dioxide and nitrogen oxides through national and international agreements rather than trying to reverse the problem locally. Thus, the reduction of sulfur emissions appears to have reversed acidification in many regions in the 1990s. Also, a remarkable national scheme in Sweden to lime sensitive lakes has substantially reduced emission of sulfur oxides and ammonia in the plant canopies and the strength of acidifying or buffering processes by chemical weathering and biological transformations.

Acidification is substantial in the northeastern United States and southeastern Canada. Acidified regions of Europe include large parts of Finland, Norway, and Sweden, mountains in Great Britain, Poland, and the Czech Republic, and carbonate-poor, well-leached sandy soils in Belgium, Denmark, and the Netherlands. Acidification has also affected a large proportion of oligotrophic softwaters in uncultivated, remote areas. Many freshwater habitats that have escaped drainage and eutrophication have instead been acidified.

The best way to reduce acidification is to lower the emissions of sulfur dioxide and nitrogen oxides through national and international agreements rather than trying to reverse the problem locally. Thus, the reduction of sulfur emissions appears to have reversed acidification in many regions in the 1990s. Also, a remarkable national scheme in Sweden to lime sensitive lakes has substantially reduced emission of sulfur oxides and ammonia in the plant canopies and the strength of acidifying or buffering processes by chemical weathering and biological transformations.

Acidification is substantial in the northeastern United States and southeastern Canada. Acidified regions of Europe include large parts of Finland, Norway, and Sweden, mountains in Great Britain, Poland, and the Czech Republic, and carbonate-poor, well-leached sandy soils in Belgium, Denmark, and the Netherlands. Acidification has also affected a large proportion of oligotrophic softwaters in uncultivated, remote areas. Many freshwater habitats that have escaped drainage and eutrophication have instead been acidified.

The best way to reduce acidification is to lower the emissions of sulfur dioxide and nitrogen oxides through national and international agreements rather than trying to reverse the problem locally. Thus, the reduction of sulfur emissions appears to have reversed acidification in many regions in the 1990s. Also, a remarkable national scheme in Sweden to lime sensitive lakes has substantially reduced emission of sulfur oxides and ammonia in the plant canopies and the strength of acidifying or buffering processes by chemical weathering and biological transformations.

Acidification is substantial in the northeastern United States and southeastern Canada. Acidified regions of Europe include large parts of Finland, Norway, and Sweden, mountains in Great Britain, Poland, and the Czech Republic, and carbonate-poor, well-leached sandy soils in Belgium, Denmark, and the Netherlands. Acidification has also affected a large proportion of oligotrophic softwaters in uncultivated, remote areas. Many freshwater habitats that have escaped drainage and eutrophication have instead been acidified.

The best way to reduce acidification is to lower the emissions of sulfur dioxide and nitrogen oxides through national and international agreements rather than trying to reverse the problem locally. Thus, the reduction of sulfur emissions appears to have reversed acidification in many regions in the 1990s. Also, a remarkable national scheme in Sweden to lime sensitive lakes has substantially reduced emission of sulfur oxides and ammonia in the plant canopies and the strength of acidifying or buffering processes by chemical weathering and biological transformations.

Acidification is substantial in the northeastern United States and southeastern Canada. Acidified regions of Europe include large parts of Finland, Norway, and Sweden, mountains in Great Britain, Poland, and the Czech Republic, and carbonate-poor, well-leached sandy soils in Belgium, Denmark, and the Netherlands. Acidification has also affected a large proportion of oligotrophic softwaters in uncultivated, remote areas. Many freshwater habitats that have escaped drainage and eutrophication have instead been acidified.

The best way to reduce acidification is to lower the emissions of sulfur dioxide and nitrogen oxides through national and international agreements rather than trying to reverse the problem locally. Thus, the reduction of sulfur emissions appears to have reversed acidification in many regions in the 1990s. Also, a remarkable national scheme in Sweden to lime sensitive lakes has substantially reduced emission of sulfur oxides and ammonia in the plant canopies and the strength of acidifying or buffering processes by chemical weathering and biological transformations.
affected by human impact. In North America, 11–14% of the mainly terrestrial birds, mammals, and reptiles are among the extinct, threatened, and rare species (Table III). Among fresh-water amphibians, fish, crayfish, and mussels, the percentages are particularly high (28–73%). No less than 103 species of North American freshwater fish are classified as endangered, 114 are threatened, and 147 deserve special attention. These species represent about one-third of all fish species. Twenty-seven species have gone extinct over the last 100 years from habitat loss, chemical pollution, introduction of exotic species, hybridization, and overfishing (Allan, 1995).

A similar assessment is not available for Europe. However, the European status of freshwater species is probably worse because of a long history of strong human impact. Europe is relatively poor in species of freshwater fish (ca. 250) compared with North America (ca. 850), Africa (1800), and the Amazon region (2000 species). Species numbers are particularly low in northern Europe (e.g., 38 species in Denmark), while they are much higher in middle and southern Europe. The Volga and the Danube include 60–70 species each, or 25% of the entire European fish fauna. In European countries about one-third of all fish species are on the IUCN Red List, much like in North America. In the global status, 20 European species are threatened, susceptible, or rare.

On both continents, many genetically isolated stocks of salmon and trout confined to certain stream systems are threatened by eradication, or they have already been lost from a variety of threats such as river regulation and construction of dams that prevent upstream migration and that destroy spawning grounds. Discharge of wastewater from agriculture, industry, and towns has also contributed to the loss, as have acidification in areas with poorly buffered waters, inter-breeding with hatchery-reared individuals, and over-fishing in the ocean.

The Atlantic salmon once was very common along its range of distribution from Iceland to Portugal. It has now disappeared from many major rivers on the continent and the annual catch has dropped profoundly. The large salmon population in the Rhine supported an annual marketed catch of more than 100,000 individuals in the late 1800s. The species declined during the 1900s and went extinct in 1957, but was later reintroduced from an artificially reared stock. Atlantic salmon was also lost from all Danish rivers during the 1900s, apart from a small population (ca. 50 individuals) surviving in the Skjern River.

DNA studies of preserved salmon scales from the now extinct populations have shown that the individual

### B. Red Lists and Historical Development in Streams of Denmark and the United States

Red Lists focus on threatened, vulnerable, and rare species and usually include those species that have disappeared recently. Among the five studied groups of freshwater insects, the percentage of red-listed species of all national species in Denmark ranges from 32% for caddisflies to 50% for mayflies (Table III). Freshwater fish have 39% and amphibia have 36% of all national

<table>
<thead>
<tr>
<th>Group</th>
<th>North America</th>
<th>Denmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshwater habitats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amphibians</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>Fish</td>
<td>34</td>
<td>39</td>
</tr>
<tr>
<td>Crayfish</td>
<td>65</td>
<td>—</td>
</tr>
<tr>
<td>Unionid mussels</td>
<td>73</td>
<td>—</td>
</tr>
<tr>
<td>Caddisflies</td>
<td>—</td>
<td>32</td>
</tr>
<tr>
<td>Dragonflies</td>
<td>—</td>
<td>62</td>
</tr>
<tr>
<td>Mayflies</td>
<td>—</td>
<td>80</td>
</tr>
<tr>
<td>Stone flies</td>
<td>—</td>
<td>40</td>
</tr>
<tr>
<td>Terrestrial habitats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birds</td>
<td>11</td>
<td>37</td>
</tr>
<tr>
<td>Mammals</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Reptiles</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>Butterflies</td>
<td>—</td>
<td>26</td>
</tr>
<tr>
<td>Moths</td>
<td>—</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>16</td>
</tr>
</tbody>
</table>

*Species on the Red List are recently extinct, threatened, vulnerable, or rare. Red Lists focus on particular vulnerable groups often highly valued by humans.
species on the Red List. Among terrestrial insects, the percentages vary from 16% for moths to 49% for butterflies. Birds and mammals have 37 and 30%, respectively.

Overall, the percentage of species on Red Lists is high for three reasons. First, the most vulnerable groups are more often included than the least vulnerable groups. Second, many species are naturally rare. Third, many species are present in very small numbers, because natural habitats have experienced profound areal restriction and deterioration. Thus, although the percentage of red-listed species tends to be higher for freshwater than for terrestrial species, this tendency disappears if more common groups of freshwater invertebrates such as dipterans, oligochaetes, and polychaetes were included.

Consequently, Red Lists have several weaknesses because they are selective and qualitative, and the intensity of search for rare species has increased over time. Red Lists do not give quantitative data on the abundance of species, and they can only describe the temporal development in crude ways provided evaluations are repeated at suitable intervals.

Many evaluations of species development, however, suffer from the lack of suitable historical description of species distribution and abundance. If such studies indeed exist, they did not use exactly the same methods and survey intensity and are therefore open to critique, even though differences may be very profound and without reasonable doubt are real. An analysis of Potamogeton (a large, aquatic plant genus) in 13 localities in Danish lowland streams, for example, revealed the existence of 6.0 species per locality and 16 species altogether 100 years ago; today the mean number is 2.8 species per locality and only 7 species grow in the same 13 localities. Several of those species that have disappeared from the 13 localities have become extremely rare throughout the country. Overall, the vegetation has become poorer in species and ecological types. Mainly oligotrophic or large, slow-growing species have disappeared, and the few survivors are species of high dispersal capacity and tolerance to disturbance and eutrophication (e.g., P. crispus and P. pectinatus); these species have now established a more profound dominance. The same type of development has taken place in other lowland regions of northwestern Europe such that some species (e.g., P. acutifolius, P. filiformis, P. zosterifolius) have become rare or threatened over wide areas.

It is noteworthy that plant species in streams display an overall positive relationship between geographical range size and local abundance, resembling the pattern more thoroughly described for terrestrial herbs and animals. There is an overall transition from species of low geographic range size and low local abundance to species of high range size and high local abundance. Those first-mentioned species face a double jeopardy of extinction because they grow in just a few places and they are infrequent at sites where they do occur. Stochastic loss of some habitats and degradation of others should therefore have a strong impact on these species, because their few and small populations make them susceptible to further losses and reduce their ability to disperse to new suitable places that may arise. In contrast, widespread species of high abundance have a double security, because they have a higher probability of surviving stochastic changes and spreading to new habitats.

Freshwater insects have undergone a similar decline of species richness, number of ecological types, and number of occupied habitats. About 20 species of the 283 Danish species of caddisflies, dragonflies, mayflies, and stone flies have gone extinct during the last 100 years. An even larger number of 76 species are either threatened, vulnerable, or rare. The red-listed species often require high water quality and have long life cycles that are sensitive to disturbance. The historic development has, therefore, led to a more stereotypic composition of both plant and macroinvertebrate communities with the same few robust species dominating at most stream sites. Species holding refugia within the stream systems are likely to recover within a few years following environmental improvement via redistribution by upstream and downstream migration. However, recovery is slow for species that have been lost from entire stream systems. Recovery may take from decades to centuries, or not occur at all, if species have been lost from islands or from geographically isolated regions remote from possible founder populations.

There are very few similar historical evaluations of species distribution and abundance. What comes closest are studies of agricultural areas of the U.S. Midwest. These demonstrate profound reduction of habitat quality, species richness, and abundance of stream invertebrates and fish over the last 50 years as a result of removal of most of the natural riparian forest vegetation, stream regulation, and alteration of water chemistry. Indices of stream quality decline with the intensity of cultivation in the catchment, and this is accompanied by a predictable decline in species richness and feeding types of invertebrates and fish. Predominantly robust, generalist species have survived in the stream communities. Fish specialized on insects have disappeared, while omnivorous fish have survived. The historical
development of plants and animals, therefore, has many similarities in the lowland streams of Europe and the United States, because all macroscopic organisms have been strongly influenced by cultivation and agricultural practices.

C. Biological Consequences of Organic Pollution and Agricultural and Industrial Chemicals

Organic pollutants give rise to deoxygenation and release of ammonia and other mineral nutrients in the water due to high microbial activity. Organic pollution leads to unhealthy aquatic conditions, high nutrient levels, untransparent water, and smothering of sediments. These conditions are detrimental to the survival of invertebrates and fish, and highly oxygen-sensitive groups of insects such as beetles, caddisflies, mayflies, and stone flies, as well as salmonid fish, may disappear almost entirely. Also, the diversity of feeding groups of macroinvertebrates declines as stony substrata for algal grazers are smothered, coarse terrestrial detritus for shredders becomes insignificant relative to the accumulation of organic mud for deposit feeders, and large predators with long life cycles find it hard to sustain episodes of oxygen depletion. Thus organic pollution is accompanied by a general loss of species richness, taxonomic diversity, and functional diversity of macroinvertebrates, in addition to the appearance and dominance of particularly robust species (e.g., tubificid polychaetes, Chironomus and Psychoda) in exchange for the loss of many more sensitive species (Fig. 3).

The biological changes are principally the same in temperate and tropical streams. The macroinvertebrate indices originally developed to describe organic pollution in temperate streams work well in tropical streams. The obvious reason is that the taxonomic composition of orders and families of macroinvertebrates does not differ greatly between temperate and tropical streams. Taxonomic similarity is particularly high between temperate lowlands and tropical highlands. The warm tropical lowland streams may have a high sensitivity to organic pollution because of low oxygen solubility, fast microbial degradation, and high invertebrate metabolism. Tropical highland streams also may be more sensitive to organic pollution than temperate streams because of the reduced partial pressure of oxygen at higher elevations, while temperatures resemble those in temperate lowland streams.

Domestic wastewater contains many chemical stressors in addition to the oxygen-consuming organic material. Sensitivity of organisms to low oxygen often goes hand in hand with the sensitivity to industrial chemicals. Similarly, sensitivity to acid waters usually changes in parallel with the sensitivity to heavy metals. Chemical stressors are lethal at high concentrations, and sublethal disturbance of body form, reproduction, neural function, and behavior occurs at lower concentrations. Among Chironomus, for example, the proportion of individuals with deformed mouthparts increases with the level of pollution. Among net-spinning caddis larvae, the proportion of abnormal, incomplete nets increases with the concentration of toxic chemicals such as aluminum. These sublethal effects will eventually lead to a decline of populations due to reduced growth and survival compared with more resistant species. However, under very severe oxygen deficiency or high concentrations of toxic chemicals, no species of higher life-forms among fish, invertebrates, and plants are capable of surviving, and only microorganisms will remain. Chemically polluted streams and ponds without any higher life include habitats with high concentrations of cadmium, copper, zinc, and ferrous iron.

The immune and hormone systems of animals can
be disturbed by exposure to organochlorine compounds (e.g., DDT, PCBs), nonylphenols and octylphenols of industrial detergent, and an estradiol component from degradation of contraceptive pills. In water bodies with a high proportion of sewage effluents, such compounds can apparently lead to the development of hermaphrodite individuals ofroach and rainbow trout, abnormally small testicles in alligators, and feminization of populations of terms and gulls feeding on aquatic animals. Estrogenic effects have received great attention because human populations have been showing increasing frequencies of testicular cancer and decreasing numbers of viable sperm cells over the last 50 years in Scandinavia.

The great variety of pesticides now being found in surface waters are also potentially dangerous because they have been manufactured to be deadly poisonous in trace amounts to fungi, insects, and plants. Insecticides target the most species-rich group of animals on land and in freshwater. For the time being, however, the broad-scale environmental significance of pesticides is largely unknown. There are suggestions that intense application of pesticides in some regions can account for an impoverished insect fauna in streams, where other pollution sources have largely been eliminated.

D. Biological Consequences of Lake Eutrophication

Lake eutrophication influences ecosystem structure and functioning as well as the distribution and abundance of most species. This fundamental influence is the result of the primary role of nutrients in boosting the biomass and productivity of phytoplankton, setting the scene for many structural and functional regulations. The attitude toward cultural eutrophication changes with its magnitude and the public interests in different regions. In lowland regions of Europe and North America, substantial eutrophication has shifted the natural wide range of lakes to a higher nutrient level that encompasses conditions from mesotrophy to hyper-eutrophy. Water purification aims at improving hyper-eutrophic lakes and moving them to eutrophy, such that most lakes may occupy the narrow range from mesotrophy to eutrophy with no lake being unpolluted and no lake being extremely polluted. As a consequence, species assemblages have become more stereotypic as the susceptible species of oligotrophic habitats, and in some cases also the extremely robust species of hypereutrophic habitats, have become very rare. Full control of pollution and the ecological disturbance associated with agriculture, fish farming, forestry, and recreational use, which may be needed to maintain pristine, oligotrophic, clear-water lakes, has rarely been implemented.

Social and environmental priorities are different in other societies. In Southeast Asia, for example, the focus is more on fish production and less on aesthetics and the protection of rare species. A small lake that is heavily fertilized by village sewage may be a valuable source of fish that feed on the thick algal soup and that are tolerant to periods of deoxygenation (Moss, 1998).

1. The Case of Lake Fure and Other Hardwater Lakes

The 100-year development of Lake Fure in Denmark can serve as an example of the widespread, accelerated nutrient loading of temperate lakes followed by recent reduction of loading rates (see Fig. 1). The long-term evaluation for this lake matches the long water residence time (15 years) and the slow numerical response of macrophyte communities in this naturally mesotrophic, hardwater lake.

Annual external phosphorus loading of the lake increased from 1.3 ton of P in 1900, to over 10 tons in 1950, to 37 tons in 1969. Until the late 1940s, the lake had managed to tolerate the increased loading relatively well because of the efficient sediment binding of phosphorus to calcium carbonates and iron oxy-hydroxides. The bottom waters had remained oxic and iron-bound phosphorus had managed to tolerate the increased loading relatively well because of the efficient sediment binding of phosphorus. Following diversion and tertiary treatment of the sewage in 1969, the external phosphorus loading declined to 3–4 tons P, but the internal concentrations have remained high owing to sediment release of about 15 tons P year−1 from the large P-pools that have accumulated.

The mean Secchi-depth during summer was about 5–6 m from 1900 to 1940, but it has fluctuated between 1.5 and 3.2 m since the early 1970s. The number of phytoplankton species has remained approximately constant between 70 and 85. However, species within particularly nutrient-demanding groups typical of eutrophic lakes (e.g., blue-green and chlorococcalean green algae) have increased from 17 to 39, and species typical of oligotrophic lakes (e.g., chrysophytes, desmids, and certain dinoflagellates) have declined from 32 to 13. Certain species of blue-greens, chrysophytes, and dinoflagellates, which were able to live in the deeper and colder waters within or below the thermocline at 10–16 m depth, have disappeared with the restriction of the illuminated zone to the upper few meters of the water column.
Species richness of submerged macrophytes in Lake Fure before the recent cultural eutrophication commenced was 35 species, including 18 rooted flowering plants, 10 characeans, and 5 mosses. Tall flowering plants grew from shallow water to the depth limits at 8 m, but dominated the vegetation down to 5 m depth. Small flowering plants, characeans, and mosses formed a mixed carpet of vegetation below the canopy of large flowering plants, while characeans dominated from 5 to 8 m. Between 1983 and 1993 only 10 of the original 35 species remained, whereas 3 new species of pollution-tolerant green macroalgae had appeared. Small flowering plants and all mosses and characeans had vanished. Eight large species of canopy-forming flowering plants have survived in the turbid lake, probably because they can compensate for the poor light conditions by exposing the leaves close to the water surface. Depth distribution of all original submerged macrophyte species has been restricted from 8 m to 2–4 m because of the reduced light penetration.

The development reveals a reduction of species richness, taxonomic diversity, and life-forms of submerged macrophytes. In addition, there is now a greater dominance among the few surviving species in the present turbid phase of the lake. Certain species have disappeared because of the restriction of suitable growth habitats, and other species have been outcompeted by the canopy-forming plants and the fast-growing macroalgae that proliferate at higher nutrient availability in water and sediments. The plants increasing in rank among the surviving species in Lake Fure are the same eutrophic species (e.g., Potamogeton crispus and P. pectinatus) that have become more abundant in lowland lakes and streams throughout northwestern Europe. The steep decline of aquatic mosses and characeans in Lake Fure is symptomatic of the broad-scale changes in eutrophicated temperate lakes. Characeans include about 80 species globally. In countries experiencing profound lake eutrophication, the characean group has become rare and many species have been added to the Red Lists. Formerly oligotrophic, hard-water lakes and brackish lagoons dominated by species-rich communities of characeans now represent a threatened habitat that is included in the European Community Habitat Directive.

2. Broad-Scale Comparisons and Whole-Lake Experiments

Other individual lake studies basically tell the same history of eutrophication. In the large lakes, however, the role of the littoral zone and bottom plants is small, and emphasis is greater on the increase of phytoplankton biomass, loss of bottom fauna, and alteration of fish communities. In most lakes, phosphorus is the main nutrient limiting phytoplankton biomass, but in some lakes the importance of phosphorus and nitrogen as limiting nutrients alternates with the seasons, and in other lakes nitrogen plays the major role (e.g., Lake Tahoe on the California–Nevada border).

In lakes with a small magnitude or a short period of cultural eutrophication, rapid recovery is observed upon reduction of nutrient loading, because only small nutrient pools have accumulated in the lake bottom. Recovery also is fast and more complete when water renewal is high, and when large proportions of unpolluted water are available from undeveloped mountains and forest areas such that nutrients can be flushed from the lake (e.g., Lake Washington in the state of Washington). Case studies of individual lakes, comparisons among many lakes, and controlled whole-lake fertilization experiments all show strong positive relationships among nutrient availability, phytoplankton biomass and productivity, zooplankton and fish production, light attenuation in the water column, and risk of anoxia in bottom waters. Predictions of phytoplankton biomass based on external phosphorus loading and water renewal are often excellent for lakes within a region of the same climate and soil geology and for lakes where nutrient inputs are precisely known. These results suggest that biotic differences do not play a major role for the biomass of phytoplankton. It has therefore been argued that the lower accuracy of predictions observed among lakes from different regions and studies may result from methodological differences and uncertainties rather than from the influence of biological differences in the food webs. In shallow lakes, however, wide shifts sometimes take place from transparent lakes of low phytoplankton biomass and high macrophyte cover, even though the rate of phosphorus loading remains approximately the same. In shallow lakes, rooted macrophytes and attached animals have the potential to exert considerable control on the communities in the open waters—in part through internal alterations of nutrient cycling. The same intense interactions between organisms and processes at the bottom and in the open water are not possible in large, deep lakes.

3. Resource and Predatory Control in Food Webs

In many lakes, phosphorus stripping and the diversion of sewage have reduced nutrient input, but improvement of lake quality has been minimal because of high
internal nutrient circulation. To reduce sediment release and internal water concentrations of phosphorus, several methods have been attempted. One principle is to remove P-enriched cultural sediments by pumping them up. Another principle is to increase the P-binding capacity in the sediment by adding iron and ensuring oxidized conditions conducive to efficient phosphorus binding to iron oxy-hydroxides by aerating the bottom waters or injecting nitrate.

In shallow lakes, biomanipulation has had the objective of reducing the number of fish, which eat the large zooplankton (especially daphnids), and thereby hopefully reduce the biomass of phytoplankton through enhanced zooplankton grazing. The additional goal is to increase the cover of submerged vegetation, which can directly reduce sediment release of nutrients, impede phytoplankton development by shading, and provide refuge to daphnids.

The success of attempts to reduce phytoplankton blooms and improve water clarity by internal manipulation of sediments and food webs is variable, and some positive short-term effects have not been sustainable in the long term. Opinions on the statistics of successes and failures of biomanipulation often vary along with the subjective view on the relative importance of resources and predatory control in the food webs. The view that grazing control on phytoplankton is high in lakes with an even number of trophic levels in the food web and low in lakes with an odd number of levels has received some support, but also attracted growing opposition. Among reasons for the opposition are (1) predictions based on the concept are poor, (2) it is difficult to define trophic levels and numbers of links, and (3) many species are omnivorous, show ontogenetic changes in food preference, and cover more than one trophic level. Moreover, there is a strong positive relationship from nutrients and resource levels to biomass and productivity of phytoplankton, zooplankton, and fish because of the need to transfer energy and matter from lower to higher trophic levels.

Many studies clearly demonstrate a strong role of predators on size composition and species abundance of prey organisms, and they show behavioral alterations of predators and prey to increase their own growth and survival. These flexible biological responses, however, do not imply that the main control of the combined biomass and productivity of all the species at each trophic level is exerted by predation. Overall, the predatory fish have a small cascading trophic effect through the entire food web and a minor influence on phytoplankton biomass and water transparency. As a consequence, there is no easy technical or biological fix to ameliorate lake eutrophication, but various techniques may assist the effect of sustained reduction of external nutrient input with the aim of reducing unwanted phytoplankton blooms.

4. Arctic and Tropical Lakes
Consequences of eutrophication are also strong in Arctic and tropical lakes. Although Arctic lakes are ice-covered for most of the year and have very low temperatures, nutrient limitation is a main determinant of the variability in phytoplankton biomass and productivity. Arctic lakes, being surrounded by frozen catchments with low rates of mineralization, tend to be more nutrient-poor and have lower rates of phytoplankton photosynthesis per unit of chlorophyll pigment than Arctic marine waters. Temperatures close to zero reduce enzyme activities, and an efficient way for algae to compensate is to achieve higher enzyme concentrations by greater nitrogen investment.

Arctic lakes are often shallow and covered by ice and snow for extended periods. They are, therefore, highly sensitive to oxygen depletion following degradation of a nutrient-stimulated phytoplankton productivity. No replenishment of oxygen is possible at the ice-covered surface, no oxygen is produced by photosynthesis owing to lack of light for many months, and the initial oxygen pool is limited by shallow depths. The presence of fish in Arctic lakes is critically dependent on water depths and the availability of oxygen throughout the year.

Heavy use of fertilizers is spreading in many tropical regions with the growth in intensive farming, and expanding tropical cities are releasing increasing amounts of untreated sewage. Lake Victoria in East Africa illustrates the catastrophic increase in this problem. This huge, ancient lake used to contain many unique species of haplochromine fish, and the fishery supported the protein needs of many people. From 1960 to 1990, the phytoplankton community showed a 10-fold increase in biomass and a shift from the dominance of diatoms to troublesome blue-green algae. Earlier, oxygen depletion was restricted to limited areas just above the deepest sediments, but nowadays oxygen-free conditions and fish-kills are widespread in the bottom waters. The most likely explanation for these changes is the clearing of woodlands, with the subsequent increase in erosion and inflow of sediment, and the fast rise in human population density, resulting in the release of larger volumes of sewage. The introduction of Nile perch has completely altered the food webs, reduced the important stocks of tilapias, and apparently driven many of the small haplochromine fish species to extinction.

In the 1990s, there has been a prolific spread of water hyacinth along the coasts of Lake Victoria, leading
to severe problems for the fishery and boat traffic. The mass growth of floating-leaved plants in lakes, rivers, and canals is a spreading problem in subtropical and tropical regions, and it is enhanced by eutrophication and introduction of exotic pest species such as Azolla, Salvinia, and water hyacinth, which lack their natural competitors, herbivores, and pathogens in these new habitats.

5. Species Richness and Lake Trophy

Species richness increases with habitat area and heterogeneity. The increase of species richness (S) with habitat area (A) is often predicted by the equation $S = constant \times A^z$. The z-value depends on the organisms, the spatial range of the study, and the habitats, and there is no single mechanism that can account for the variability of z. However, low z-values are consistent with high immigration rates, low extinction rates, and a low rate of increase in additional habitat with increasing area. As a consequence, very small z-values are predicted for bacteria, microalgae, protozoans, and zooplankton. Indeed, the z-values predicting crustacean zooplankton species richness are only 0.054 for European and 0.094 for North American lakes. A low z-value of 0.10 for submerged macrophytes in Scandinavian lakes is also consistent with their widespread occurrence.

In contrast, species richness of fish in lakes of four different regions in Canada and the United States shows much higher z-values (0.16, 0.22, 0.36, and 0.37) consistent with the more restricted geographic range size, smaller local abundance, and greater risk of extinction of fish following disturbance or restriction of natural habitats. As a consequence, risks of local, regional, and global extinction are high among fish and low among microorganisms and plankton organisms, with aquatic plants and insects probably holding an intermediate position. Some large crustaceans and molluscs presumably resemble the fish by having a low rate of dispersal and including several species of restricted geographical distribution (e.g., the freshwater crabs and prawns on tropical oceanic islands).

Changes in ionic composition and nutrient concentrations of freshwaters are superimposed on changes in habitat area and heterogeneity. Water chemistry is not independent of habitat area and heterogeneity, since small lakes include a higher proportion of low-pH, low-calcium, and nutrient-poor waters than do large lakes. In many studies of the influence of pH and lake trophy on species richness, it is not possible to fully compensate for the influence of habitat area and heterogeneity.

The most diverse conditions and highest species richness often are found under mesotrophic conditions, where many species can co-exist at different sites and depths within the lakes. Thus, high habitat heterogeneity is probably important for the common peak of species richness under mesotrophic conditions. For example, many free-living and attached species adapted to different conditions of light, temperature, and exposure can replace each other along depth gradients from shallow to deep waters. Under very oligotrophic conditions, many algae and plant species are nutrient-limited, and low primary production places an energetic restriction on the diversity and density of invertebrates and fish. In contrast, under hypereutrophy, many species are restricted by the lack of light and oxygen in deep waters, and special adaptations are required for survival in very muddy sediments.

Two other conditions contribute to the decline of species diversity under hypereutrophic conditions. First, the natural local variability between unfertilized and fertilized sites disappears with the overall nutrient enrichment at all sites accompanying eutrophication. Second, hypereutrophy is exceptional for most species, which have evolved over many millions of years of no or weak human impact. Species richness is expected to be highest in those common habitat types that have had the most widespread and long-term natural occurrence, because there has been time and room for speciation. Hypereutrophic lakes, which are common today, have been exceedingly rare during most of the development of freshwater species.

It is noteworthy that lake ecosystems are much more sensitive to environmental deterioration and catastrophic declines of organisms by eutrophication than terrestrial ecosystems for at least four reasons. First, nutrients tend to stay in circulation within the lake boundaries. Second, phytoplankton responds by a steep increase in biomass and productivity. Third, photosynthesis and growth of algae and plants are restricted to the uppermost surface waters because of the impoverished light conditions. Fourth, the risk of oxygen depletion is very high because of low oxygen solubility in the water. Therefore, species richness within many groups of organisms declines in hypereutrophic lakes because of reduced habitat heterogeneity, restriction to the distribution of organisms, and development of stressful and highly variable environmental conditions with respect to light, oxygen, pH, and sulfate.

E. Biological Consequences of Acidification

Acidification initiates complex chemical and biological changes in surface waters and catchments with poorly weathered rocks and thin soils. The acidification of
Freshwater habitats is also accompanied by acidification of terrestrial ecosystems in the catchments, leading to a widespread regional decline of biodiversity.

Whole-lake experiments involving acidification to pH 5 demonstrate that damage to ecosystem functioning is of secondary concern compared to impoverishment of the biotic communities. Although many phytoplankton algae disappear, the remaining algae can maintain biomass and primary production at the same level, largely regulated by the input of phosphorus and water renewal. The rate of decomposition is governed by primary production and not by pollutants. However, the complete disappearance of large invertebrate detritivores upon acidification may lead to the accumulation of detritus and over the long term reduce plant production based on recirculated nutrients. Overall, nutrient cycling was not disrupted in whole-lake experiments, but nitrification declined at pH values below 5.4.

Acidification of the watershed changes ionic composition and may reduce availability of phosphorus in soils and surface waters. Soil acidification increases the leaching of base cations, and it has specific effects on the solubility of aluminum and phosphate. Chemical weathering of aluminum minerals is enhanced under acid conditions. However, soluble forms of aluminum ions may precipitate with phosphate at the proper pH, leading to "oligotrophication." As a result, lakes in acidified watersheds may become very clear from the enhanced phosphorus limitation of phytoplankton growth, and certain acid-tolerant aquatic mosses (e.g., Sphagnum spp.) and filamentous green algae (e.g., Mougeotia spp.) may spread across the lake bottom to greater depths. Only a few vascular plant species can tolerate highly acidic conditions (e.g., Juncus bulbosus).

Species diversity and abundance decline within most groups of organisms upon acidification, though some groups are much more sensitive than others and distinct differences exist among species in their tolerance to acidification (Fig. 4). In acidified regions of Europe and the U.S. Northeast, many lakes have lost 30% or more of the species in some taxonomic groups. Among planktonic and benthic microalgae, desmids—a group of green algae—prefer buffer-poor waters of low pH and low calcium content, and they are often accompanied by a few acid-tolerant species of blue-green algae, dinoflagellates, diatoms (e.g., Eunotia exigua), green algae (Chlamydomonas acidophila), and euglenophytes (Euglena mutabilis). In contrast, most species of diatoms, chlorococcalean green algae, and blue-green algae prefer well-buffered waters of neutral pH. Among planktonic rotifers, copepods, and daphnids, only a few species tolerate acid waters (<4–5), so the planktonic food webs become simple. The tolerants species can utilize the vacant ecological niches left by others, thus maintaining a relatively constant zooplankton biomass despite the decline of species richness.

Many snails, mussels, and crustaceans (e.g., crayfish) are highly sensitive to acidification due to problems in maintaining the ionic regulation of sodium, chloride, and potassium and in forming the exoskeleton of calcium carbonate or calcium-chitin when pH and calcium concentrations are low. Higher calcium concentrations can help in maintaining membrane integrity, reducing gill permeability to salt loss, and ensuring the construction of shells. Several metal toxics such as aluminum mobilized under acidic conditions have the opposite influence.

Among insects, the most acid-tolerant species include surface-breeding beetles with a hard exoskeleton and certain stone flies and caddisflies with a low metab-
olism and mostly lacking thin-walled body parts. In contrast, mayflies with their large, thin gills are often highly sensitive to acidification. The decline of species richness of insects is not as pronounced when acidic conditions occur in waters rich in humic substances, which can bind and detoxify the metal ions mobilized at low soil pH.

A common result of acidification is the dramatic loss of fish. Acid-stressed regions of Scandinavia, the Adirondack Mountains and Wisconsin (both in the United States) and La Cloche Mountains (Canada) reveal a steep decline of fish species as lake pH declines from 8 to 4, and many lakes with pH less than 5 are fishless. The decline of fish species richness with reduced pH is much less pronounced in subtropical Florida lakes and in tropical lakes. The presence of more acid-tolerant fish species in warm lakes may be due to a much slower and more gradual decline of pH over time and, accordingly, better opportunities of genetic adaptations to low-pH waters. Also, Florida lakes have lower concentrations of aluminum and free mineral acids, longer growing seasons, and no episodic input of acids resembling the massive acid input of north temperate lakes during snowmelt.

Fish studies in north temperate lakes show that eggs, embryos, and larval stages of fish are more sensitive to acidification than the adult stages. As a consequence, young generations may vanish upon acidification, whereas old fish may remain for a while. Populations will eventually die out unless young fish are continuously stocked. Salmonids and cyprinids with a high metabolism and a large gill surface appear to be more susceptible to acidification than fish with a low metabolism and a small gill surface (e.g., eels). A critical phase for the well-being of fish and invertebrates is the maintenance of efficient oxygen uptake and ion regulation across the gill surfaces.

1. Ecosystem Functions and Explanations

Ecosystem functions have proven to be more resistant to acidification than species changes. Species assemblages are often plastic enough to maintain ecosystem functions under quite extreme stresses, although only a few species survive. For example, plant productivity continued in a dense monospecific vegetation of Juncus bulbosas growing in a highly acidic, temperate stream (pH 2–3) from a lignite mining area with intense pyrite oxidation, while neighboring alkaline streams (pH 6–7.5) contained a mixed assemblage of about 10 rooted plant species. Similarly, phytoplankton biomass and production were unaffected in lakes polluted with strong acids and trace metals from smelters. Phytoplankton and other microorganisms are also locally abundant and are easily dispersed, such that highly pH-tolerant species will rapidly show up and multiply in recently acidified localities.

Some ecosystem processes, however, may decline or disappear entirely under very acidic conditions. At pH values below 5.4–5.7, ammonium begins to accumulate due to the cessation of nitrification. Therefore, the loss of internally produced nitrate through denitrification will stop. However, the enhanced atmospheric deposition of nitric acid and sulfuric acid will permit the microbial activity of denitrifiers and sulfate reducers in oxygen-free sediments and, thereby, generate alkalinity in the water bodies.

The biotic communities of lakes and streams are strongly damaged and impoverished by acidification. The most vulnerable species appear to be organisms with year-long life cycles and poor dispersal, such as large invertebrates and fish. These groups are also the less speciose, particularly in northern water bodies, which means that there is little redundancy in the food webs to prevent pH-sensitive species of large detritivores, plankton-eating fish, and predatory fish from being replaced by pH-tolerant species. Fish and crayfish, which are highly valued by humans, are therefore easily damaged. Even bird populations such as dippers, osprey, and pied flycatchers may vanish from the lack of food or the formation of thin-shelled eggs while feeding on calcium-deficient and aluminum-rich prey.

Overall, surface waters of very low pH represent a special environment to which very few species have become adapted as a result of (1) the direct stress at low pH, (2) the unsuitable ionic composition, (3) the toxic metals, and (4) the rarity of these habitats during the evolution of aquatic organisms, particularly in temperate and subarctic regions.

VI. CONCLUSIONS

Freshwaters of high quality are essential to humans for domestic and agricultural uses. Freshwater ecosystems provide many other highly valued and critical ecological services, and they support a very high diversity of species and taxonomic groups considering their relatively small volumes and surface areas. Human impact has caused numerous stream reaches and shallow lakes to disappear, disturbed and polluted many others, and disrupted the close contact between the floodplain and the streams. Consequently, the diversity and abundance of numerous freshwater species
have declined due to the profound restriction of areal cover and habitat diversity of freshwater environments. Moreover, the systematic loss of pristine, unpolluted freshwaters has threatened or eradicated those species that require such environments.

The overall disruption of the linkage between freshwater and terrestrial environments has far-reaching consequences for global hydrology, elemental cycling, and biodiversity. Among those consequences are enhanced risks of flooding and drought, greater risks of land erosion and marine siltation, and reduced opportunities for the countless plants and animals who live a natural double-life between freshwater and land to survive in the future.

Accompanying all of this disturbance and pollution has been the accidental or deliberate introduction of nonnative organisms such as crayfish, plants, sport fish, and zebra mussels, which may monopolize freshwater habitats and drive the native, endemic species in formerly undisturbed lakes and river systems to extinction. Several species of freshwater fish have gone extinct during the last few centuries, and many additional species of plants, large invertebrates, fish, and frogs now exist in extremely small and geographically isolated populations. These species face a great risk of extinction, and they can no longer provide their accustomed ecosystem functions. Nor can these species be enjoyed within most of their native area. Unfortunately, they will find it exceedingly difficult to recolonize newly restored aquatic habitat and contribute to future specification.

See Also the Following Articles

- AGRICULTURE, INDUSTRIAL
- ENDANGERED FRESHWATER INVERTEBRATES
- EUTROPHICATION/OLIGOTROPHICATION
- FRESHWATER ECOSYSTEMS
- HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW
- POLLUTION, OVERVIEW
- WETLANDS RESTORATION

Bibliography


I. Measurement of Functional Diversity
II. Explanation of Concepts and Terminology
III. Early Work on Functional Diversity and Ecosystem Processes
IV. The Effects of Functional Diversity
V. Conclusions

GLOSSARY

diversity-productivity hypothesis The proposal that greater diversity would lead, on average, to greater total biomass or productivity.
diversity-stability hypothesis The proposal that ecosystems containing more species would be more stable.
ecosystem composition The list of species or functional groups that are present in a given ecosystem.
ecosystem functioning The rate, level, or temporal dynamics of one or more ecosystem processes such as primary production, total plant biomass, or nutrient gain, loss, or concentration.
functional diversity The range and value of those species and organismal traits that influence ecosystem functioning.
functional group A set of species that have similar traits and that thus are likely to be similar in their effects on ecosystem functioning.
niche differentiation Differences in the morphology, physiology, or behavior of species that can influence their abundances, dynamics, and interactions with other species, including the ability of various competing species to coexist.
sampling effect The hypothesis that diversity might influence an ecosystem process because of the greater chance that a given species trait would be present at higher diversity, and the effect of its presence on ecosystem functioning.

FUNCTIONAL DIVERSITY refers to those components of biodiversity that influence how an ecosystem operates or functions. The biological diversity, or biodiversity, of a habitat is much broader and includes all the species living in a site, all of the genotypic and phenotypic variation within each species, and all the spatial and temporal variability in the communities and ecosystems that these species form. Functional diversity, which is a subset of this, is measured by the values and range in the values, for the species present in an ecosystem, of those organismal traits that influence one or more aspects of the functioning of an ecosystem. Functional diversity is of ecological importance because it, by definition, is the component of diversity that influences ecosystem dynamics, stability, productivity, nutrient balance, and other aspects of ecosystem functioning.
I. MEASUREMENT OF FUNCTIONAL DIVERSITY

Because of the large number of traits that each species possesses, the large number of different species that exist in most habitats, and the incomplete knowledge of which species traits influence various ecosystem processes, there is, as yet, no simple way to measure functional diversity. Rather, items that are more easily measured than functional diversity are used as indices or correlates of functional diversity. The most common of these indices is the number of species present in a habitat, which is called the species richness or species number of the habitat. All else being equal, habitats with greater species richness should also have greater functional diversity. This occurs because species differ in their traits. Sites that contain more species should thus also contain, on average, a greater range of species traits, which is greater functional diversity. Species diversity indices, such as the Shannon diversity index, are similarly used as indirect measures of functional diversity. Another commonly used index of functional diversity is the number of different functional groups (defined later) that exist within a given community or ecosystem. This is also called functional group diversity. Assuming that organisms can be categorized as belonging to groups that differ in traits relevant to ecosystem functioning, greater functional group diversity should correlate with greater functional diversity. However, variations among species within a given group could also contribute to functional diversity. Observational, experimental, and theoretical studies indicate that functional diversity, as measured by any of these three means, is one of several important factors that determine ecosystem functioning. Because there is, as yet, no clear way to measure functional diversity, one or more of these three indices will be used as a proxy for functional diversity in this chapter. Before reviewing the research linking functional diversity to ecosystem processes, which is the focus of the remainder of this chapter, it is important to introduce and define some terms.

II. EXPLANATION OF CONCEPTS AND TERMINOLOGY

A. Functioning

As they are used by ecologists, the words function, functional, and functioning are not meant to imply that an ecosystem process has any underlying goal or purpose. Indeed, to try to minimize any such implications, it has become standard practice to refer to “ecosystem functioning” or “ecosystem process” rather than the “function of an ecosystem.” The latter might be misinterpreted as meaning that an ecosystem exists to perform a given function, which is inconsistent with our knowledge of the process of evolution. Rather, functioning refers solely to the way in which an ecosystem operates.

B. Ecosystem Processes

Ecologists study many different aspects of the functioning of communities and ecosystems. The three most frequently considered ecosystem processes are productivity, stability, and resource dynamics. Productivity refers to the rate of production of biomass within a given trophic level. The production of plant biomass is called primary production, the production of biomass of herbivores is called secondary production, and that of predators is called tertiary production. Stability has a wide range of definitions, including the degree to which an item is resistant to change when experiencing a single perturbation, the degree to which an item fluctuates in response to an ongoing suite of small-scale perturbations, and the dynamics of return to its prior state after a single perturbation. Stability can be measured at the level of populations, communities, or ecosystems. The resource dynamics of an ecosystem are measured by the rates of supply and loss of limiting nutrients, by the efficiency with which organisms use limiting resources, and by the proportion of limiting resources that the organisms living in an ecosystem are able to capture.

C. Functional Groups

Each species has a large number of morphological, physiological, and behavioral traits, many of which might influence the abundance of species and ecosystem functioning. One way to deal with such complexity has been to identify traits that seem more likely to influence ecosystem processes. Chapin et al. (1997) suggested that the species traits with the greatest effects on ecosystem functioning were those that (a) controlled the acquisition, use, and availability of limiting resources; (b) modified the feeding structure of food webs; and (c) affected the occurrence and magnitude of distur-
FUNCTIONAL DIVERSITY

bances. Such traits can be used to classify organisms into different functional groups. For instance, species can be divided, first, into functional groups based on their position in a food web: photosynthetic plants, herbivores, predators, parasites, parasitoids, decomposes, and so on. Organisms within each of these groups can be further subdivided based on their acquisition and use of their limiting resources. For grassland plants, for instance, this might be based on the time, within the growing season, when each plant was maximally active (cool-season versus warm-season plants), and on its carbon (C-3 or C-4 photosynthetic pathway) and nitrogen physiology (high nitrogen use efficiency, low nitrogen use efficiency, ability to fix atmospheric nitrogen). Such considerations might lead, for instance, to the classification of grassland plants into six functional groups: C-3 grasses, C-4 grasses, C-3 forbs, C-4 forbs, legumes, and woody plants. The assumption inherent in making such a classification is that species within a class are highly similar, and those in different functional groups differ markedly from one another.

D. Diversity versus Composition

It has long been recognized that the functioning of an ecosystem depends on which species the ecosystem contains (i.e., on its species composition). Interest in species diversity as an alternative or additional explanation for ecosystem functioning means that it is necessary to define species diversity, especially functional diversity, in a way that distinguishes diversity from species composition. This requires a definition that is more restricted than that traditionally used. In particular, effects should be attributed to diversity only once there has been simultaneous control for effects of composition, and effects should be attributed to composition only once there has been simultaneous control for effects of diversity. To achieve this in an experimental, theoretical, or observational study, it is necessary (a) to hold composition constant via randomization (numerous communities with randomly-chosen compositions) while changing composition, (b) to hold diversity constant while changing composition, (c) to simultaneously vary both in an appropriately randomized and replicated design, or (d) to control for each statistically, such as via multiple regression, which is most appropriate for observational studies.

III. EARLY WORK ON FUNCTIONAL DIVERSITY AND ECOSYSTEM PROCESSES

Effects of diversity on ecosystem processes were first recognized by Darwin in *The Origin of Species*. Darwin noted that it was well-known that increased plant diversity led to greater primary productivity in pastures. The British ecologist, Charles Elton, hypothesized in his 1958 book titled *The Ecology of Invasion by Animals and Plants* that diversity would impact many aspects of ecosystem functioning. In particular, he suggested that greater diversity would lead to greater ecosystem stability, an idea that was further developed by the leading ecologists of that era, including Robert MacArthur, Gene Odum, and Ramon Margalef. Elton also suggested that greater diversity would decrease the susceptibility of an ecosystem to invasion by other species and would decrease the incidence of outbreaks by diseases and pests. Elton's diversity-stability hypothesis was called into question, though, by the mathematical theory of May (1972), which predicted that the linear stability of communities of competing species would, in general, decrease as the diversity of the communities increased. The general consensus reached after publication of May's book was that other factors were likely to be more important than diversity as determinants of ecosystem processes. This view led ecologists to focus more of their attention on other issues, with much of that effort dedicated to better understanding the mechanisms of species interactions and the effects of species composition on ecosystem processes.

Recent explorations of the potential effects of diversity on ecosystem processes were inspired, to a great extent, by the publication of *Biodiversity and Ecosystem Functioning* (Schultz and Mooney, 1993). In a chapter in that book, Vitousek and Hooper hypothesized that many ecosystem processes, like primary productivity, should increase as diversity increased, and they stressed that the most important component of diversity might be functional group diversity. Agricultural studies were reviewed in a chapter by Swift and Anderson, who noted that mixed crops, especially those containing a legume and a grass, were often more productive than either crop species growing alone, supporting the diversity-productivity hypothesis. A chapter by McNaughton reviewed and evaluated a large number of observational and small-scale experimental studies in which stability was greater for ecosystems containing more species and highlighted data supporting Darwin's diversity-productiv-
in a community, diversity would depend on the number of species drawn, the relevant species trait (diversity in this model because the range in the values of term average biomass of a plant community, sampling effect yield a simple equation that relates the long-
term average biomass of a plant community, \( R_{obs} \), to its original plant species diversity, \( N \):

\[
R_{obs} = aQ \left( S - \frac{R_{obs}^* - R_{max}^*}{N + 1} \right)
\]

Here, \( a \) is the rate of resource mineralization, \( Q \) the coefficient of resource conversion into biomass, and \( S \) is the rate of resource supply in the habitat.

For this treatment, let \( R^* \) be the level to which a limiting resource is reduced by a species when growing alone. As shown both theoretically and in numerous competition experiments (Groover 1997), the best competitor would be the species with the lowest \( R^* \). The \( R^* \) value of the species can be used to rank them from good to poor competitive ability (i.e., from the lowest to the highest \( R^* \) value). Assume that the species composition of a community is determined by random draws (sampling) from the infinite pool of species with all possible \( R^* \) values between a minimum \( (R_{min}^*) \) and maximum \( (R_{max}^*) \). On average, the functional diversity of a community would depend on the number of species drawn, \( N \), which is the initial diversity. The number of species in a community, \( N \), is a good measure of functional diversity in this model because the range in the values of the relevant species trait \( (R^*) \) is higher in communities containing more species. These assumptions of sampling effect yield a simple equation that relates the long-term average biomass of a plant community, \( R_{obs} \), to its original plant species diversity, \( N \):

The sampling effect model predicts that total community biomass, a measure of primary productivity, increases with plant diversity, as shown in Fig. 1a. The trend predicted is one in which added diversity leads to large increases in productivity when diversity is low, but has progressively smaller impacts when diversity is higher. This simple model demonstrates that the magnitude of the effect of functional diversity, as measured by \( N \), on ecosystem functioning depends on the range of interspecific differences in the species pool—that is, on the term \((R_{species} - R_{max})\) in Equation 1. This gives basis to the intuitive concept that diversity effects ecosystem processes because ecosystems with greater diversity have a greater range in those species traits that influence functioning.

The sampling effect model also predicts that the average quantity of unconsumed resource should decrease as diversity increases (Fig. 1b). Indeed, in the sampling effect model, the increased biomass at higher diversity is caused solely by the more complete utilization of the limiting resource that occurs, on average, at higher diversity.

The model also illustrates the importance of species composition. Each point in the two graphs of Figure 1 represents the response of a community with a different randomly determined species composition. Thus, the variability among plots with the same diversity measures the impact of composition, and the variability among diversity levels represents the impact of diversity. Both diversity and composition are strong determinants of productivity and resource levels in the sampling effect model.

The other major type of models that have been proposed to relate productivity to diversity are niche differentiation models. In essence, such models assume that a habitat is spatially or temporally heterogeneous, that species differ in the traits that determine their response to this heterogeneity, and that each species is a superior competitor, and thus is more productive, for some subset of the heterogeneous habitat conditions. These assumptions can allow a large number of species to coexist and assure that ecosystem productivity increases, on average, as diversity increases. For instance, two factors, such as soil pH and temperature, might limit plant abundance. Each species could have some combination of these factors at which it performed best. Such niche
FIGURE 1. (A) The sampling effect model predicts that productivity should be greater at greater functional diversity, here measured by the number of species present. The variation within a given level of species richness is caused by different species compositions. (B) Productivity is higher in plots with greater functional diversity because of greater capture of the limiting resource. The concentration of unutilized resource is predicted to decline as diversity increases.
differentiation would mean that each species did best in a part of the habitat, but that no species could fully exploit the entire range of conditions.

The essence of such niche models can be captured by making the simple assumptions that each species has a circular area of radius \( r \) in which it can live and be a good competitor (Fig. 2a), that all species attain comparable abundances per unit habitat occupied, and that competition similarly reduces abundances of all overlapping species. If the values for one limiting factor range from \( 0 \) to \( a \) \( r \) and the other from \( 0 \) to \( b \) \( r \), where \( a \) and \( b \) measure habitat heterogeneity for factors 1 and 2, and if species are drawn at random from all those that could live at some point in the habitat, then total community biomass (i.e., the proportion of environmental conditions "covered" by one or more species) would be

\[
\frac{\text{community biomass}}{\text{environmental conditions covered by one or more species}} = 1 - \left(1 - \frac{\pi}{ab + 2(a + b) + \pi}\right)^{\frac{a}{b}}
\]

Here \( N \) is species diversity. As for the sampling effect model, the variance within a given level of diversity is caused by differences in species composition, and differences between diversity levels is caused by diversity.

In addition, the niche model predicts that greater habitat heterogeneity (i.e., greater values of \( a \) and \( b \)) requires greater diversity in order to achieve a given level of productivity. In general, heterogeneity should increase with habitat size, leading to the prediction that greater biodiversity is required to attain a given level of productivity in larger habitats. For instance, for small, relatively homogeneous habitats (\( a = b = 1 \)), only six species are needed to attain 93% of maximal productivity. However for spatially heterogeneous habitats (\( a = b = 10 \)), a diversity of 135 plant species is needed to achieve this level.

A comparison of the sampling effect model with the niche differentiation model reveals a major difference in the expected pattern of the dependence of productivity on diversity. For the sampling effect model, there are no higher diversity plots that are more productive than the most productive monoculture. In contrast, for the niche model, there are two-species plots that are more productive than the most productive monoculture, three-species plots that are better than the best two-species plot, and so on. For ecosystems that meet the assumptions of the sampling effect model, which might occur for highly productive agricultural fields, there might be situations in which judicious choice of the right species and variety could lead to as great productivity from a monoculture as would be possible for a highly diverse mixture of species. In contrast, for habitats with spatial or temporal heterogeneity, which should occur for almost all natural ecosystems and for all but the most intensively managed ecosystems, niche differentiation models are more likely to hold. In such cases, increased diversity is expected to lead to greater productivity and to more complete use of limiting resources.

Although these models, and the models of Michel Loreau, have predicted that greater diversity can lead to greater ecosystem productivity, this need not always be the case. For instance, if the assumptions of the sampling effect model were modified to have progressively better competitors be progressively less productive, productivity would be a decreasing function of diversity. This suggests a more general principle: if species differ in their competitive abilities, and if higher competitive ability is correlated with some other traits, then these traits will, on average, be better represented in more diverse communities, thus biasing the functioning of these communities in the direction determined by these correlated traits.

2. Experimental Studies

Darwin suggested that it was common knowledge among farmers that a greater diversity of pasture plants would lead to a greater production of herbage in pastures. In his 1993 chapter, McNaughton cited this and presented more recent examples in which greater plant diversity led to greater productivity, as did Swift and Anderson. Indeed, earlier work reviewed in Harper’s 1977 book showed that pairs of coexisting species often yield more than either species did when living by itself. As reviewed in the 1993 chapter by Vitousek and Hooper, some of the first evidence linking higher plant diversity to greater retention of soil nutrients came from a field experiment in Costa Rica by Ewel as collaborators. They found that communities planted to many tropical species generally retained more soil fertility than those planted to monocultures.

The first published direct test of the diversity-productivity hypothesis came from a greenhouse experiment by Naeem et al. (1995). By growing various randomly chosen combinations of 16 plant species 1, 2, 4, 8, or 16 at a time in a greenhouse, they found that community biomass was greater at higher plant diversity (Fig. 3a). This team performed another experiment in a series of growth chambers and also had results...
A graphical illustration of a niche differentiation model. Here each circle represents the range of environmental conditions in which a given species can live, and the full rectangle shows the range of environmental conditions that occur in a given habitat. This model and similar niche differentiation models predict that productivity should be an increasing function of diversity. (B) The predicted effects of diversity on productivity for the model illustrated in part (A).
suggesting that greater diversity leads to higher productivity (Naeem et al., 1994). Next came results from a large-scale field experiment begun in Minnesota in 1993 (Fig. 4). Its 147 plots, each 3 m × 3 m, were planted to contain 1, 2, 4, 6, 8, 12, or 24 plant species randomly and independently chosen from a set of 24 prairie-grassland species (reviewed in Tilman, 1999). It found highly significant effects of plant diversity on both productivity (Fig. 3b) and on the soil concentration of the limiting resource, nitrate (Fig. 3c). By the fifth year of this experiment, its results supported niche differentiation models more than the sampling effect model as the major cause of the effects of diversity on the measured ecosystem processes. Indeed, the most productive plot in 1998 was a 24-species plot that had 65% greater total biomass than the most productive monoculture. A second experiment, adjacent to this Minnesota exper-

FIGURE 3 (A) The observed effect of plant diversity on the productivity of plant communities in the greenhouse experiment of Naeem and collaborators. (B) Effects of diversity on productivity for the Minnesota field experiment in which grassland diversity was experimentally controlled in 147 plots. (C) Effects of diversity on the concentration of unutilized soil nitrate for the Minnesota experiment.

FIGURE 4 The smaller of the Minnesota biodiversity experiments, shown here, has demonstrated that plant diversity has a strong effect on ecosystem productivity and nutrient dynamics. The experiment has 147 plots, each being 3 m × 3 m (about 10 feet by 10 feet) in size. See also color insert, Volume 1.
iment (reviewed by Tilman, 1999), controlled for both species diversity and functional group diversity (Fig. 5). Its results were similar to those of the first experiment and showed highly significant effects of species diversity, functional group diversity, and functional group composition on primary productivity and nutrient dynamics. In both of the Minnesota grassland diversity experiments, the vast majority of species coexisted in all plots to which they had been added, further supporting niche differentiation models.

Knops et al. (2000) recorded the number of non-planted species that invaded the Minnesota diversity experiment plots, and their biomass at the time when they were removed from the plots. They found that significantly fewer species invaded higher diversity plots and that the total biomass of invading species was lower in higher diversity plots. Further analyses suggested that the effect of diversity on invasions was caused by the lower levels of soil nitrate in higher diversity plots. This provides one simple mechanism whereby diversity may influence the extent to which an ecosystem is invaded by other species and suggests that levels of unconsumed limiting resources may, in general, be an important determinant of the success of an invading species.

For native, undisturbed grasslands close to the two Minnesota biodiversity experiments, plant abundances were greater and soil nitrate was lower in more diverse plots (see Tilman, 1999), which is consistent with the experimental results and with the predictions of theory. However, correlational patterns must be interpreted carefully because they could be confounded by other correlated variables. Michel Loreau used a model that linked environmental factors, biodiversity, and ecosystem functioning to explore this point. The model illustrated that correlational field data could be misinterpreted easily because of a confusion of cause-and-effect relationships. Just such issues cloud the interpretation of the possible effects of island diversity on ecosystem processes for a study of 50 Swedish islands. In an intriguing study that showed links between island size and the frequency of wildfire, David Wardle and collaborators found that a suite of ecosystem traits were correlated with both island size and plant diversity. However, it is unclear if diversity caused the observed differences in ecosystem processes or if both these processes and diversity were controlled by fire frequency.

Hooper and Vitousek (1998) performed a field experiment, planted in 1992, in which they controlled plant functional group diversity and composition using plants common to California grasslands. After a year of growth, they found that functional group composition had a much greater effect on plant community biomass than functional group diversity, but that the utilization of soil nutrients increased significantly as diversity increased.
In a 4-month greenhouse experiment, Symstad et al. (1998) found that total plant biomass was significantly higher at higher diversity and that most of this effect was attributable to the presence of legumes. They also determined the effects of the deletion of individual species on total biomass and found that the strength and direction of these effects depended on which species were present and which was deleted.

In an experiment that was replicated at eight different sites across Europe, ranging from Scotland and Ireland to Portugal and Greece, Hector et al. (1999) found that greater plant diversity led to greater primary productivity. An important finding of this unique experiment was that the quantitative effect of diversity on primary productivity was the same across all eight sites. In combination with the other field and laboratory experiments, the European experiment suggests that there is a general, repeatable effect of grassland diversity on primary productivity.

In total, these studies show that plant productivity is greater at greater diversity and that this also corresponds with greater utilization of limiting soil resources. In general, short-term experiments showed weaker effects of diversity on productivity and soil nutrients than longer-term experiments. This is expected because diversity should impact ecosystem processes via changes in plant abundances mediated by competition, and such interactions can require several years to occur. Further work is needed on other trophic levels and in other communities to determine the extent to which the patterns observed to date apply to other communities (e.g., marine fisheries, forest ecosystems, coral reefs).

B. Functional Diversity and Stability

1. Theory and Concepts

A large number of authors, including Charles Elton, Robert May, Stuart Pimm, and Sam McNaughton, have contributed considerable insights into the effects of diversity on stability. May (1972), for instance, showed that the abundances of individual species become progressively less stable as the diversity of the community in which they live increases. Several recent papers have explored the effects of diversity on the stability of communities of competing species (Doak et al., 1998; Ives et al., 1999; Tilman, 1999). The first two of these papers showed that the temporal variability of an ecosystem process, such as ecosystem productivity, is expected to be lower when the ecosystems contain more species. This can occur for the same reason that a portfolio composed of many different types of stock tends to be more stable than one containing stock of a single company. An additional factor that can cause ecosystem functioning to be more stable for more diverse ecosystems is competition. When some disturbance harms one species, the species with which it interacts experiences less competition. This allows these competitors to increase in abundance. Their greater abundance partially compensates for the decreased abundance of the first species, thus stabilizing the functioning of the ecosystem. Ives et al. (1999) showed that increased diversity only led to increased stability when the species differed in their responses to habitat fluctuations and disturbances. Because such differences are a direct measure of functional diversity, the work of Ives et al. (1999) showed that increases in functional diversity lead to greater stability.

For a thorough treatment of theory relating diversity and stability, see “Stability, Concept of.”

2. Experimental and Observational Studies

The evidence that led Elton to propose the diversity-stability hypothesis was anecdotal. In his 1993 chapter, and in earlier papers, McNaughton defended the diversity-stability hypothesis by citing several observations and experiments in which greater diversity was associated with greater stability. A variety of other studies, summarized in Tilman (1999), also have found effects of diversity on stability. For instance, a study by Frank and McNaughton of eight grassland sites within Yellowstone National Park found that those with greater plant species diversity had smaller shifts in plant community compositions during a severe drought. Two British ecologists, Taylor and Woiwod, performed a long-term project in which they monitored the abundances of hundreds of insect species at a large number of sites. The data they collected provide evidence that supports the hypothesis that more diverse insect communities should be more stable. The greater stability is expected because of the statistical averaging (or portfolio) effect pointed out by Doak et al. (1998). Specifically, because the temporal variances in the abundances of individual species in this community scales as their abundance to a power of about 1.6, the portfolio effect should cause more diverse insect communities to have lower temporal variability.

Several authors have found that greater oak tree diversity stabilizes the population density of an animal, the acorn woodpecker, that feeds on the seeds of the acorn.
Acorn woodpeckers are highly dependent on acorns as a source of food, but oaks produce acorns as a mast seed crop. Masting means that there is great year-to-year variability in the rate of acorn production. There is a striking decrease in the year-to-year variability of acorn woodpecker abundances for woodpeckers living in habitats containing a greater diversity of oaks. Thus, greater oak diversity led to more stable acorn woodpecker populations. Moreover, acorn woodpecker densities were much lower for areas with a single oak species than for those with several.

A long-term experiment in Minnesota provides additional evidence suggesting that greater plant diversity leads to greater stability (reviewed in Tilman, 1999). In a series of 207 plots annually monitored from 1982 to 1999, total plant community biomass was found to be more stable in plots containing more species. Both in response to a major disturbance, a severe drought (Fig. 6), and in response to normal year-to-year variation in climate (Tilman, 1999), plots with greater diversity had lower year-to-year variability in their total plant biomass. In particular, the severe drought caused plant biomass to fall to half of its predrought level in plots with about 15 or more species, but caused it to fall to 1/8 to 1/12 of its predrought levels in plots containing one or two plant species (Fig. 6). Similarly, year-to-year variation in total biomass fluctuated about twice as much in low diversity as in high diversity plots (Tilman, 1999). Although total community biomass was more stable at higher diversity, analyses of the stability of individual species showed that these declined slightly but detectably, at higher diversity. Thus, diversity stabilized total community biomass at the same time that it destabilized the abundances of individual plant species. Plant diversity and composition were confounded in this experiment because both changed in response to nitrogen addition. Multiple regression, used to control for this confounding, found highly significant effects of diversity on stability for both cases. These analyses also showed that species composition and functional group composition also had significant effects on stability.

McGrady-Steed, Harris, and Morin (1997) found, in a laboratory study of the effects of diversity in microbial communities, that the temporal variability was significantly smaller at higher diversity. Indeed, a four-fold increase in diversity led to about a three-fold decrease in the temporal variability of whole-community net respiration, a measure of ecosystem activity. The rate of microbial decomposition of particulate organic matter also increased with diversity in this study. Finally, they found that greater diversity led to lower susceptibility to invasion by another species, but that invader success was highly dependent on community composition. Naem and Li (1997) similarly found that greater diversity led to greater reliability, which was measured as the lower variability in total community biomass among communities of identical diversity. This effect was also apparent in the greenhouse experiment that Naem and collaborators had performed earlier.

In total, these studies provide strong evidence that communities with greater diversity are more stable and suggest that individual species in such communities may be less stable. Theory, experiment, and observation are in general agreement, but this topic merits additional exploration.

V. CONCLUSIONS

The research performed to date illustrates that a variety of different ecosystem processes are impacted by the number and kinds of species living in the ecosystem.
This work illustrates that species differ in traits that influence ecosystem functioning and suggests that ecosystem processes depend on the range in those traits represented in the ecosystem. However, there are, as yet, no clear demonstrations of the specific traits that are relevant to particular ecosystem processes and no simple ways to directly measure functional diversity. Rather, correlates of functional diversity, such as species richness or functional group richness, remain the best, albeit indirect, way to measure functional diversity.

See Also the Following Articles

C. PLANTS • ECOSYSTEM FUNCTION, PRINCIPLES OF • FUNCTIONAL GROUPS • HABITAT AND NICHE, CONCEPT OF • STABILITY, CONCEPT OF

Bibliography


I. Functional Groupings and Their Utility
II. Decoupling Relatedness from Functional Similarities
III. Functional Changes in Assemblages over Evolutionary Time
IV. Functional Convergences
V. The Structure and Functioning of Natural Communities and Ecosystems
VI. Ecosystem Diversity, Stability, and Redundancies: Examples from Coral Reefs
VII. Conclusions

GLOSSARY

convergent evolution Distant or unrelated organisms evolve the same anatomical, morphological body plan characteristics, or ecological function.
ecosystem The combined plant and animal communities plus their physical environment.
ecosystem function The energy flow, productivity, element cycling, and resilience of ecosystem structure; synonymous with “behavior of ecological systems” and “ecological processes.”
ecosystem structure The organisms, their communities, biodiversity, and habitats that comprise an ecosystem.
functional groups Polyplyetic suites of species that share ecological characteristics and play equivalent roles in natural communities and ecosystems. Commonly, organisms with convergent anatomical, mor-

EARTH’S BIODIVERSITY IS DISTRIBUTED among a surprisingly few functionally different organisms. These “functional groups” are suites of species that play equivalent roles in natural communities and ecosystems. They result from convergent evolution which is channeled by phyletic constraints that limit the variations possible on a given body plan relative to the biomechanical limits imposed on organisms by their environment. The diversity and ecological function of these groups are independent of regional species diversity. Large-scale changes in functional groups can alter the structure and functioning of ecosystems. Such alterations have occurred over evolutionary time and are occurring today in ecosystems under strong human influences. In highly diverse ecosystems there are redundant species within functional groups. These redundancies provide a buffer against ecosystem collapse should individual species within the group become rare or extinct. Describing communities and ecosystems at the functional group level conveys information that transcends taxonomic and evolutionary boundaries. Ultimately, functional groups may provide a low-resolution tool for accurately predicting ecosystem change.
I. FUNCTIONAL GROUPINGS AND THEIR UTILITY

More than 40 years ago, ecologist and evolutionary biologist G. E. Hutchinson wondered why there were so many kinds of species. This question engaged many of the most powerful minds in ecology and the natural sciences. It launched decades of research, brought us the concept of ecological niche and considerable ecological and evolutionary theory, and is the focus of many articles in this encyclopedia. Much less has been written about an equally compelling question: Why are there so few kinds of morphologies or physiologies, and what might the ecological consequences of such limits be? It also raises new questions, such as the following: Are there limits to the possible number of variations on any given body plan? Are there similar limits to the functional attributes of species and suites of species? Can functional attributes of species be abstracted such that their ecological roles can be inferred? Can several species fulfill the same functional roles? In other words, are species replaceable such that there are ecologically equivalent or redundant species? The answers to these questions have profound implications to how we view the natural world and possibly how we can conserve it. For example, does the extinction or extirpation of any given species degrade natural ecosystems in tangible ways?

Biodiversity includes the diversity of organic life on Earth and how it functions. How organisms “make their living,” how they interact and contribute to the ecosystem, relates to the functional properties of biodiversity. Functional groupings are polyphyletic suites of species that share ecological characteristics. Commonly, organisms with convergent anatomical, morphological, physiological, behavioral, biochemical, or trophic characteristics are grouped together. Functional properties may correspond to different ecological characteristics depending on the question being asked. They can include the habitats in which they live, the biomass or structure they impart to a community, the rates at which they grow (their productivity), the rates at which they lose biomass (i.e., disturbance), the organisms they consume, the organisms that consume them, and their numerical or mass-specific importance to their communities or ecosystems. A functional group analysis can be applied more broadly in space for making biogeographic comparisons and in time for reconstructing paleocommunities than is possible at the level of species or of related taxa.

I explore functional groupings first from the perspec-

tive of the evolution of functional similarities, and then I discuss what functional groupings might mean to the structure and function of communities and ecosystems. I draw heavily, but not exclusively, on examples from the marine realm, in which there is the greatest phyletic diversity, longest evolutionary history, and clearest fossil record of how organisms and assemblages have evolved. Other extensive works on the topic of functional groups, such as that by Schluze and Mooney (1994a), are exclusively terrestrial in their approach.

Functional distinctions can be based on any characteristic shared among organisms, such as morphology, physiology, or behavior. These are not independent features of organisms. An organism's morphology can influence its behavior and its physiology. Functional groupings would be unnecessary if organisms' taxonomic relationships corresponded well with their functional roles. However, functional characteristics are often decoupled from even closely related species, and thus the relationship between taxonomic relatedness and functional groupings warrants consideration.

II. DECOUPLING RELATEDNESS FROM FUNCTIONAL SIMILARITIES

Species that are closely related evolutionarily are not necessarily morphologically or functionally similar. Conversely, just because organisms are distantly related does not mean they are morphologically or ecologically different. There are examples of unrelated, or distantly related, organisms having similar morphologies, anatomies, behaviors, physiologies, and functional roles in natural communities. Such convergent functional properties are well-known and operate at many levels. Obviously, bats are flying mammals that exploit some of the same habitats and foods as some birds and moths. Squid are swimming mollusks that exploit some of the same habitats and foods as fish. Although bats are not perfect substitutes for birds, nor are squid identical to fish, their phyletic relationships tell us little about the ecological roles or functional characteristics they posses.

Although most closely related organisms share traits with the clade from which they evolved, and thus will tend to be functionally similar, this is not always the case. For example, hermit crabs are relatively small, anomuran decapods with an exposed, asymmetric abdomen that requires protection from unoccupied snail shells. They differ morphologically from true crabs, the brachyuran that have abdomens protected by a shield-like cephalothorax or carapace that also covers their
walking legs (Fig. 1). True crabs can attain large size. However, hermit crabs are very closely related to the morphologically distinct anomuran king crab (Fig. 1). King crabs convergently evolved many of the characteristics of true brachyuran crabs; they are large and their abdomen is protected by a broad shield-like cephalothorax. Therefore, phylogenetically related king and hermit crabs are morphologically different, whereas phylogenetically distant king and true crabs are morphologically similar.

Morphological similarities can translate to functional similarities. In shallow coastal habitats of the Gulf of Maine, there are four species of large, clawed, decapods that live together and feed on the edible blue mussel, *Mytilus edulis* (Fig. 2). The two most closely related species are *Cancer irroratus* and *Cancer borealis*. Both are brachyuran crabs in the family Cancridae. The third brachyuran crab, *Carcinus maenas*, is in the Portunidae family. However, the two distantly related crabs, *Carcinus maenas* and *C. irroratus*, share functional characteristics of small body and chelae ("claws") size while being behaviorally flexible, quick, and dexterous. They characteristically chip the perimeter of mussel shells at many locations to gain access to the food. Conversely, *C. borealis* and the lobster *Homarus americanus* share functional characteristics of large body and chelae size while being behaviorally constrained, slow, and strong (Fig. 2). These two decapods characteristically crush mussel shells in the middle of the shell by applying brute force to the center of the shell. Therefore, the two most closely related crabs are functionally dissimilar and the two sets of most distantly related crabs (*Carcinus maenas* and *C. irroratus*) and other decapod groups (*C. borealis* and *H. americanus*) are both functionally similar.

Even within species, large functional differences exist. Amphibians such as frogs begin life as tadpoles with diets and habitats different from those of adults. Ferns and algae have ecologically distinct haploid and diploid phases. The heteromorphic range in some marine algae is striking. For example, the brown (Phaeophyta) algal kelp *Macrocystis pyrifera* can exceed 40 m in length and forms forests in its sporophyte (2n) phase but is a microscopic filament in its gametophyte (1n) phase. Some red algae (Rhodophyta) alternate between encrusting and erect phases; some green algae (Chlorophyta) alternate between endolithic microscopic and filamentous phases. Some habitats only support one phase. Each of these different ontogenetic and ploidy phases has strikingly different ecological properties, and differences within species can be greater than some differences among species. Therefore, knowing the species does not always provide insight into how the organism functions ecologically.

There are distinct limits to variety in morphology or important ecological function that are independent of biodiversity. Some characteristics, such as body size, are biomechanically limited; therefore, depending on the biomaterials available to an organism (i.e., cellulose, chitin, calcium carbonate, or bone), their morphology and the environment in which they live may limit their ecological functions. These relationships are well described in works such as Thompson's (1966) *On Growth and Form*, Peter's (1983) *The Ecological Implications of Body Size*, or Niklas's (1994) *Plant Allometry*. Arguably, size and shape are prime determinants of many functional groupings of both plants and animals. Predators generally scale with, or are larger than, their prey. Thus,
organisms constrained to small size are likely to be more vulnerable to predators and rarely if ever are "apex predators" in an ecosystem or community. Insects are a good example. They also illustrate the disconnect between species diversity and morphological variety. Insects account for most of Earth's biodiversity, but they are all relatively small. In contrast, most vertebrates are much larger (especially mammals and dinosaurs) and their biomass distribution does not overlap that of insects. Despite their long evolutionary history, there are no cow-sized insects, nor are there insect-sized cows. Body size constraints among insects occur because, among other reasons, their respiration is relatively inefficient and limited by their open circulatory system and trachea. Other factors, such as the composition of their exoskeleton, further constrain morphological evolution and the habitats into which they can and have radiated. Although insects have radiated impressively into terrestrial systems, they are virtually nonexistent in the marine realm. Thus, three-fourths of all species are insects, but three-fourths of the habitable space is devoid of them. There are morphological and functional limits to diversity.

III. FUNCTIONAL CHANGES IN ASSEMBLAGES OVER EVOLUTIONARY TIME

Ultimately, the environment controls the distribution, abundance, and body plans that live in it. Evolution at several levels generates diversity of form and function, but the environment filters it. Constraints to diversification are many and relate to the environment as it changes and how organisms deal with it.

The history of life on Earth is punctuated with significant functionally novel additions to biotic assemblages. An excellent example of a functional group effect on global-scale ecosystem function is the cyanobacteria that began oxygenating the atmosphere more than 3 billion years ago. Biologically available oxygen began increasing approximately 1.4 billion years ago and is coincident with (and possibly necessary for) the evolution of the first metazoans. Still higher levels of oxygen may have been necessary for the Cambrian "explosion" of small shelly fauna. The sudden phyletic and global appearance of hard skeletal material (chitin and calcium carbonate) allowed for the evolution of large body sizes, efficient mobility, teeth, and defenses against predators. The structure and functioning of life on Earth changed at this time. Organisms functionally similar to those alive today first appeared. Since then, during the past 600 million years, there have been several biotic revolutions.

Decoupling phyletic evolution from functional group evolution is important. All but one of the major phyla may have evolved at or near the Precambrian–Cambrian boundary in as few as 25 million years. Alternatively, some argue that phyla evolution may have occurred much earlier (1 billion years ago) as minute and embryonic or larval-like organisms and thus not recognizable by today's criteria. If this latter theory is correct, it would be a clear case of phyletic evolution being disconnected from the functional/morphological (phenetic) evolution that occurred 400 million years later. Despite the Cambrian (or earlier) origin of most major phyla, functional evolution and the filling of unexploited habitats ("ecospace" sensu Bambach, 1985) took considerably more time. Throughout the Paleozoic, morphological evolution and the filling of unoccupied habitats occurred episodically and polyphtyletically, beginning with small surface deposit feeders and low-profile suspension feeders, progressing to greater depths into sediment and new epifaunal heights. The progressive utilization of three-dimensional space called "tiering" is thought to have contributed substantially to the global diversification in the early Paleozoic (Fig. 3). The diversifying continued into the pelagic realm and into the sediment habitats (infaunalization) and recently included the radiation of large-bodied predators.

This pattern of diversification is composed of three relatively distinct faunas, each characterized by its own characteristic dominant taxa (Fig. 3A), initial rates of diversification, and "equilibrium" diversity. Each fauna was functionally different. Surface deposit feeders such as trilobites and monoplacophorans dominated the Cambrian fauna, whereas the Paleozoic fauna radiated into and relied heavily on the pelagic realm. For example, the taller suspension-feeding crinoids of the Paleozoic fauna joined near-benthos suspension-feeding brachiopods, anthozoans, and stenolomate bryozoans and pelagic groups such as cephalopods and graptolites as dominants of this fauna. There were no known macroalgal grazers in the Paleozoic, and large stands of macroalgae were probably common. Sand habitats were virtually devoid of infauna. In contrast, the modern fauna included the rise to dominance of large invertebrate and vertebrate predators and herbivores. These included predatory neogastropod snails, malacostracan crab and lobsters, and at least four groups of swimming vertebrate predators among the bony fishes (Ostechthyes), dinosaurs (Reptilia), sharks (Chondrichthyes), and cetaceans (Mammalia, i.e., dolphins and whales). Most modern groups of fish (e.g., most modern reef
fish) evolved at this time. Explanations for why large-scale changes occurred at the family level are complex. Organisms such as small deposit feeders may have suffered higher extinction rates as more mobile and more pelagic organisms with larger gene pools replaced them. A global crisis in the water column contributed to the most severe extinction event at the end of the Permian. Therefore, suspension feeders, pelagic organisms, and organisms with pelagic larvae suffered the highest extinction rates. This extinction caused a sharp decline in several of the dominant families of the Paleozoic fauna and marked the transition to the dominance of the modern fauna (Fig. 3B). Prior to the dominance of the modern flora there were virtually no shell-crushing carnivores. Several groups of shell crushers, such as true brachyuran crabs, evolved in the Cretaceous and became important predators. Deep-grazing herbivores also evolved among the modern fauna. They created a uniquely intense level of grazing pressure which may have denuded many shallow marine habitats of macroalgae.

A. The Sudden Origin of Functionally Distinct Organisms and Then Functional Stasis

Functional evolution in the marine realm is striking. The transitions from surface-dwelling deposit feeders to a range of suspension feeders, to organisms using...
vertical space, and to big predators and herbivores represent significant functional changes over evolutionary time. There were also terrestrial revolutions in organism function. Angiosperms displaced functionally similar conifers and evolved functionally unique weeds. Pollinating insects and angiosperms coevolved and diversified. Obviously, the history of life is punctuated with the sudden appearance of functionally unique organisms that often change the biotic world.

Often, functionally unique groups change little over evolutionary time. Among marine invertebrate herbivores, there are only a few capable of biting into limestone or calcium carbonate. These herbivores leave a grazing trace fossil that allows for functional interpretations to be made over evolutionary time. Three groups known to excavate calcium carbonate are regular echinoids (sea urchins), true limpets (Patello gastropoda), and chitons (Polyplacophora). Trace fossils from the Cretaceous (approximately 70 million years ago) reveal that each of these groups grazes distinctively with the same grazing marks and to about the same depth into calcium carbonate as their modern counterparts (Fig. 4). Although considerable variation exists within groups of herbivores, the differences among groups are much greater. Echinoid sea urchins have a characteristic intensity of grazing that is much deeper than that of chitons and limpets which are very distantly related but have similar shallow bites into calcium carbonate. In all cases, the first trace fossils known show these invertebrates biting the surface of calcium carbonate with the pattern and the depth into the substrate as they produce today. There appears to be little or no functional change in how they graze after the group first evolved. There are other examples. Predatory gastropods that drill their shelled prey have not changed their mode of attack since they first evolved in the Triassic more than 200 million years ago. Many limb and running characteristics between terrestrial vertebrate predator-prey interactions have also been shown to be remarkably stable over evolutionary time. This suggests that among these consumers and within established clades, there is functional stasis over evolutionary time.

IV. FUNCTIONAL CONVERGENCES

Convergent evolution results in distantly related organisms converging on the same body plan or ecological function. Charles Darwin noted in his Origin of Species (1859) that

for animals belonging to two most distant lines may have become adapted to similar conditions, and thus have assumed a close external resemblance; but such resemblances will not reveal—they rather tend to conceal—their blood relationship (page 463).

Relatively few convergently evolved anatomical, morphological, or physiological characteristics are the cornerstones of functional groupings.

A. Groupings among Mobile Organisms

Convergent evolution is well-known and documented in the terrestrial realm. Marsupial and placental mammals have converged to similar morphologies and ecological function (Fig. 5). The limited variations on the mammalian body plan are evident in the wolf and cat-like carnivores, the arboreal gliders, fossorial herbivores, anteaters, and subterranean insectivores that evolved independently in Australia for the marsupials and on the other continents for the placental.

Convergent functions can be found among dissimilar-looking organisms. Distantly related marine mollusk herbivores, chitons and limpets, provide an excellent example. Chitons (class Polyplacophora) and snails (class Gastropoda) evolved in the Cambrian at the beginning of the mollusk diversification. The true limpets, Patello gastropoda, evolved from the Archaeogastropoda, much later in the Triassic. However, the radula (the teeth) of these two groups are functionally similar (Fig. 6). Both groups have relatively few teeth that contact the substrate, and those that do are hardened.

FIGURE 4 Comparison of modern and ancient limpet grazing marks on calcium carbonate substrates. (A) Modern limpet, Tectura testudinalis, grazing calcareous algae (scale bar = 0.5 mm). (B) Cretaceous (more than 65 million years ago) limpet, Radulae tus sp. (scale is the same as in A) [Reproduced with permission from (A) Stezlock (1982) and (B) Crimes and Harper (1977)].
by mineralization of iron or silica compounds (note the black teeth in Fig. 6). Both groups have strong buccal musculature for applying downward forces. Within the molluscan body plan, only chitons and limpets have such a large foot area-to-mass relationship and an excavating-type radula. These morphological and anatomical characteristics, along with their size and mobility, allow species of these groups to specialize on large and
expansive macrophytes such as sea grasses, kelp, and encrusting coralline algae. Although species diversity is much higher in other groups of mollusks such as nonlimpet gastropods, it is the functional characteristics of these two groups that make them capable of consuming and even trophically specializing on the tough or limestone-imbedded cells of kelp or coralline algae, respectively.

Convergent functions are numerous among terrestrial organisms. For example, a diversity of flowering plants and their pollinators possess similar morphological and anatomical characteristics despite significant phyletic separation among the plants and the pollinators. The geometry of the flowers, such as the length and width of the floral tube, as well as the placement of nectaries are common features among the plants. The flying characteristics and mouthparts are convergent among the pollinating insects, butterflies, moths, bats, and birds. It is surprisingly easy to find functional similarities among distantly related organisms.

B. Groupings among Sessile Organisms

Convergent morphology and anatomy is common among sessile organisms on land and in the sea. I briefly consider terrestrial plant communities that have many of the same convergences described previously for animals. Families of cactus-like plants in the Old World (Euphorbiaceae) and New World (Cactaceae) dominate arid environments but are unrelated. Much has been written about the remarkably convergent forest and biomes that develop under similar environmental conditions (Huston, 1994). There appear to be relatively few adaptive solutions to common environmental conditions. Often, the suite of characteristics that improve ecological function under specific conditions define functional groupings.

Functional groupings among sessile organisms were studied by terrestrial ecologists, who grouped plants according to their size and the location of their growth and regenerative structures. Raunkiaer (1934) showed that particular functional groups (called “life-forms”) persisted and dominated the vegetation under similar environmental conditions. Using this scheme, Arctic regions were clearly distinguished from arid or tropical regions. Similar floras, defined at the functional group level, dominated under similar conditions in different regions even though there were often no species in common. This early focus on the size and placement of growth and regenerative structures was based on the realization that gradients in environmental stress and productivity relative to disturbances are critical to the success of specific life-forms. Plants with protected meristems and perennating structures persisted in stressful or highly disturbed habitats. Where environmental limitations were relaxed, plant size and competitive ability were more important to the flora.

Recently, the field has become popular among marine ecologists and paleontologists who study sessile marine organisms. Again, morphological and functional
convergences are viewed more as the rule than the exception. Several growth forms have evolved convergently among widely disparate groups, such as marine algae, Cnidaria, and Bryozoa (Fig. 7). Most extant and extinct species within those higher taxa can be placed within the groups illustrated in Fig. 7. The growth forms represent a morphological progression. The simplest growth form is solitary and can be a single cell as is the case of a diatom or a solitary organism. This group is often composed of small organisms. The next step in a morphological progression of growth forms simply involves a linear series of the single cells or modules, creating simple filaments. These can grow prostrate along the substrate as recumbent algal filaments or as runners or vines among sessile invertebrates. Similar morphologies growing vertically escape the benthic boundary layer. They experience greater water flow, and most of the cells of these groups are in contact with their aqueous environment. The more massive multiserial or multiseriate organisms allow three-di-
mensional space to be exploited. Several ecological processes, such as productivity and competition, vary predictably as sessile organisms become larger and more massive. With increasing mass, fewer cells directly contact the marine environment. Fewer still are in contact with the environment among the encrusting forms because half of their surface area is affixed to primary substrate. The growth forms in Fig. 7 exhibit progressively increasing size and generally decreasing surface area to volume ratio.

For benthic marine algae of all three major divisions (= phyla), the progression from single-celled microalgae to filamentous, foliose, larger erect forms and finally to crustose algae (Fig. 7) corresponds to a reduction in productivity per unit biomass (Fig. 8A; Steneck and Dethier, 1994). In general, microalgae and filamentous forms with a high surface area to volume ratio are capable of the highest mass-specific productivity. This decreases progressively from filamentous forms to the more massive erect forms including kelp and is lowest

<table>
<thead>
<tr>
<th>Growth Forms</th>
<th>Construction &amp; Characteristics</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solitary individual or small size</td>
<td>Microalgal dense</td>
<td>Solitary</td>
</tr>
<tr>
<td>Uniserarial colony or uniseriate prostrate filament</td>
<td>Recumbent filaments (seafloor)</td>
<td>&quot;Runners &amp; Vines&quot;</td>
</tr>
<tr>
<td>Uniserarial colony or uniseriate or oligoserial erect filament</td>
<td>Cladophora, Sphacelaria, Polyphylia</td>
<td>Madrepora</td>
</tr>
<tr>
<td>Multiserarial colony or multiserial, corticated erect, large size (exploits vertical space)</td>
<td>Corticated and Leathery macroalgae Gracilaria, Sargassum kelp (laminariae)</td>
<td>Branching</td>
</tr>
<tr>
<td>Multiserial or multiseriate encrusting (Lowest SA/V), expansive but low profile (exploits horizontal space)</td>
<td>Crustose Lithothamnion Petractis</td>
<td>Montipora</td>
</tr>
</tbody>
</table>

**FIGURE 7** Growth forms of sessile marine organisms. Examples include several divisions of marine algae and Cnidaria and Bryozoa phyla. Other groups having some of the same growth forms include the extinct tabulate and rugose corals and modern sponges (after Coates and Jackson, 1985).
among the crustose forms. Thallus longevity (Fig. 8B) increases along the same continuum of growth forms. Canopy height, however, shows somewhat of a bell-shaped curve, with the tallest algal forms such as kelp and fucoid algae having intermediate levels of productivity and longevity (Fig. 8C). These emergent ecological properties were shown for cold-water floras of the Atlantic and Pacific as well as for tropical floras of the Caribbean (Fig. 8).

Morphologically and functionally similar organisms have evolved similar ways to overcome similar problems. For example, as organisms become more massive, fewer cells are in direct contact with the marine environment. That is, they have a plumbing problem. Sessile organisms such as plants, algae, and invertebrates evolved physiological integration or a degree of independence and cooperation among cells within an organism or modules within a colony. Often this involves sharing cytoplasm or photoassimilates within the organisms and among modules of a colony.

Physiologically integrated colonial organisms have an advantage over physiologically isolated asexual organisms by sharing cytoplasm among members of the colony. Sessile organisms can respond from a stimulus by growing at a place distant from the stimulus. Wound healing can be facilitated and mortality of the genet reduced by physiological integration. Physiological integration evolved iteratively among sessile organisms on land and in the sea. This allows land plants such as vines to colonize the inhospitable habitats of sand dunes or tree canopies by being physiologically integrated. The most abundant marine algae, the red encrusting coralline algae, are integrated and this is thought to have contributed to their ecological and evolutionary success over unintegrated or poorly integrated algae. Arguably, physiological integration in algae is essential for their morphological diversity. The least physiologically integrated group, Chlorophyta or green algae, have the lowest morphological diversity despite having the same long period of evolutionary history dating back to the Precambrian.

Physiological integration may be a prerequisite for
the development of coral reef ecosystems. It evolved independently among cnidarian reef builders: the tabulate, rugose, and finally the scleractinian corals. In each case, the proportion of species with a high degree of physiological integration increased during times of significant reef-building activity. Coral reef ecosystems require high rates of coral growth to keep up or catch up with rising sea level. Reef-building corals today are physiologically plants, with endosymbiotic algae producing carbohydrates necessary for growth. Transport of photosynthetic products throughout the coral colony requires a high degree of physiological integration. Without high rates of coral growth, the coral reef ecosystems would stand little chance of remaining in the shallow turbulent zone in which they thrive. In fact, it is likely that many coral reef ecosystems that have dominated shallow seas for much of the past 600 million years require the framework-building corals to be physiologically integrated.

Physiological integration must be functionally important to sessile organisms because it evolved independently many times. It is necessary for the similar functional groups found among the corals, algae, and bryozoans (Fig 7). Colonial integration, or the degree of interdependence and cooperation among modules within a colony, is important for all these groups because it increases colony function. It also allows the development of nonfeeding defensive or reproductive functions within the colony or clone, and it allows greater structural integrity and morphological diversity. The important functional advantages of physiological integration may account for why it has persisted and become ubiquitous among so many unrelated sessile, colonial, and clonal organisms.

V. THE STRUCTURE AND FUNCTIONING OF NATURAL COMMUNITIES AND ECOSYSTEMS

There exists a broad spectrum of ways to describe patterns. Historically, the emphasis in ecology has been on species. As Naem (1998) pointed out, species are often viewed as being phylogenetically "singular" and likely to be "ecologically singular." This was the basis of the important works by Hutchinson and MacArthur (1959), who stressed the uniqueness of species. That no two species can occupy the same ecological niche was the foundation of decades of studies on the structure and function of natural communities and ecosystems. Functional group approaches are different. They stress similarities among unrelated species that share critical organismal features. Convergent evolution in many cases results in phylogenetically distant organisms sharing ecological properties. The utility of using shared species characteristics for describing communities rests on the idea that in any given environment there are relatively few species attributes important to the structure and functioning of natural communities and ecosystems. These attributes are largely independent of the biodiversity of the ecosystem.

The importance of any species or group can relate to its abundance (measured as population densities, biomass, areal cover, or structural height) or its impact on other species. Whittaker (1965) argued that species' importance should be measured by its productivity or effect on productivity. Obviously, the production of biomass must be approximately matched to its degradation or "disturbance" (sensu Grime, 1981). In some systems, species importance depends more on the role of disturbance (e.g., predation) than on productivity. For both disturbance and productivity, properties of the organisms and properties of the environment must be considered simultaneously and independently. Holdridge's (1947) classic "life zone system" for classifying terrestrial vegetation considered only precipitation and temperature as driving productivity and thus creating discrete vegetative communities and ecosystems. Others have shown how disturbances such as storms and fires additionally shape terrestrial plant communities. Although debates continue regarding the primacy of bottom-up or productivity-driven vs top-down or disturbance-driven structure of natural communities, in most systems both components are viewed as important. Therefore, the debate is largely over relative importance.

Productivity and disturbance are the two variables most often identified as environmental structuring agents of natural communities and ecosystems. Huston (1994) reviewed the work of others to develop a model that compared the frequency of disturbance (defined as patch or gap formation) as a function of "potential plant productivity." Relative to these two axes, he showed consistent patterns in the terrestrial realm in species diversity, patch formation, dominance, and competitive ability. General models of community dominants have been advanced by Grime (1981), Southwood (1988), and Steneck and Dethier (1994) (Fig. 9). Specifically, Grime (1981) suggested that, in environments with a high productivity potential (low stress) and low disturbance potential, large, canopy-forming trees will dominate the system (called a "competitive strategy"). Under conditions that favor high productiv-
ity with increasing productivity, perennial herbs and finally weedy annual herbs will come to dominate the system (called a "ruderal strategy"). Under low disturbance and low productivity potential environments, stress-resistant groups such as lichens will dominate the system (Fig. 9A).

Several studies applied Grime's (1981) model to the marine realm (Fig. 9C). Strong similarities between the two realms surfaced, including that the highest biomass and largest canopy-forming organisms (trees and kelp) grow under high productivity and low disturbance potentials (Fig. 9C). Furthermore, low biomass and can-

**Figure 9** Generalized model of community dominants that compares Grime's (1981) model for terrestrial plants (A) with that of Steneck and Dethier (1994) for marine algae (B). (C) A synthesis of A and B. Grime's primary "strategies" or groupings of "competitive," "ruderal," and "stress tolerance" are denoted by the letters C, R, and S, respectively.
opy height dominated by encrusting growth forms are characteristic of stress-tolerant assemblages in both realms. A difference might occur among the organisms that dominate under conditions of chronically high levels of disturbance. Grime (1981) assumed that it is impossible to withstand such conditions and thus weedy, "ruderal strategies" would persist, whereas Steneck and Dethier (1994) argued that an additional group of organisms tolerant of such disturbances (the coralline algae) predictably dominate such systems. Although it is common to find disturbance-resistant plants under conditions of high stress in the terrestrial realm, they can be found under conditions of low stress (high productivity potential) in the marine realm. It is unclear whether this functional difference between the marine and terrestrial realms is real or perceived. Nevertheless, one value to a functional group approach is that this dialog can occur at all.

VI. ECOSYSTEM DIVERSITY, STABILITY, AND REDUNDANCY: EXAMPLES FROM CORAL REEFS

Much has been written on this topic, but most of the literature draws from terrestrial studies such as O'Neil et al. (1986), Schulze and Mooney (1994a), and Huston (1994). However, rather than reiterating these studies and their myriad examples, I will use coral reef ecosystems as examples in this section. Coral reefs provide excellent, albeit often overlooked, examples. They are spatially discrete and smaller in vertical and horizontal structure than most of their terrestrial counterparts. Most of their functional components are readily visible (there are few below-surface functions), and major alterations and resilience can be measured in years and decades rather than centuries for observing ecosystem changes. Biogeographically, they vary widely in species diversity and thus are ideal for comparing ecosystem structure and function relative to their biodiversity.

If relatively few environmental factors can structure communities and ecosystems at the functional group level, then what role does species diversity play? Here, I consider biodiversity and its role in the stability, structure, and functioning of coral reef ecosystems. If functional groups are composed of ecologically equivalent species, then perhaps diversity within these groups contributes to the redundancy of important functions and thus may be ecosystem insurance for stability.

A. Diversity and Stability

Ecosystem and community stability was once thought to correspond directly with species diversity. MacArthur (1955) theorized that the stability of a community will increase as the number of links in its food web increases. It followed that with the loss of each species "the integrity of the biosphere will degrade in a small but significant way" (Lawton and Brown, 1994, page 255). Alternatively, species richness may be relatively unimportant as long as structuring processes controlled by primary producers, consumers, and decomposers are maintained and function well with very few species. In fact, it is now widely believed that "ecosystem processes often have considerable redundancy built into them" (Lawton and Brown, 1994, page 266). Functional groups can be composed of the redundant species that perform essential ecosystem processes.

There is good empirical evidence that diverse ecosystems are not stable. Coral reefs are among the most diverse and most productive ecosystems in the world, but they are not characteristically stable. In the Caribbean and Indo-Pacific reef coral cover, abundance and dominance suddenly changed as the result of changes in abundance of a single echinoderm species in each of the two regions. In 1983, the mass mortality of the Caribbean sea urchin, Diadema antillarum, caused a rapid "phase shift" that increased macroalgal biomass and contributed to the decline of live coral cover in Curacao, St. John, Jamaica, and St. Croix. In all cases, algal assemblage dominance shifted from turf algae (Fig. 8A, Nos. 1–3) to macroalgae (Fig. 8A, Nos. 4–6), and this caused a decline in mass-specific primary productivity. It also precipitated changes in the abundance of other herbivores. The abundance of all reef-building organisms (corals, hydroids, and coralline algae) declined sharply in shallow zones where carbonate production is normally highest.

In the Indo-Pacific, periodic, localized outbreaks of the predatory sea star Acanthaster planci caused coral cover declines of all species of between 85 and 94% as reviewed by Connell (1997). Other disturbances, such as storms, sedimentation, toxins, and fishing, have also been shown to cause significant reductions in coral cover. In the cases of very rapid ("acute or pulse") mortality from sea star predation, rapid recovery at decadal timescales was observed. Other longer lasting ("chronic or press") mortality events had slower rates of recovery. However, in the scale of years to decades these highly diverse ecosystems were demonstrably unstable.

Ecosystem resilience is sometimes considered a cor-
relate of stability. Rapid recovery from disturbances was more commonly observed in reefs of the Indo-Pacific than in reefs of the Caribbean, where no recoveries have been documented. Reef recovery relates more to the nature and duration of the disturbances than to the biodiversity of the system. One of the best examples of recovery from collapse is in some reefs of Hawaii, although the species diversity there was as low as that found throughout the Caribbean (Table 1).

**B. Diversity and Ecosystem Structure**

The basic structure of coral reef ecosystems does not vary with biodiversity. Coral reefs throughout the Indo-Pacific, Caribbean, and at Clipperton Atoll (eastern Pacific) have similar morphological zonation of corals despite a two orders of magnitude difference in coral species richness (Table 1). Branching corals dominate shallow zones, mound-shaped forms at greater depths, and platy corals at still greater depth. At Clipperton Atoll, one of the most species-depauperate reefs in the world, Glynn and coworkers (1996) measured vertical growth rates of coral similar to those found for similar morphologies at similar depths elsewhere in the Indo-Pacific. Steneck (1988) found similar patterns of zonation, distribution, and abundance among marine algae and herbivores at the functional group level in most studied reef systems of the Indo-Pacific and Caribbean. Others showed that functional characteristics of reef fish, such as size, activity rhythms, and relative abundances among ecological and trophic categories, were similar regardless of the species richness of the fish fauna.

**C. Diversity and Ecosystem Function**

Species diversity is always low at high trophic levels. Often a single, "apex" predator is critical to the structure of some systems. However, at lower trophic levels, diversity increases, and it is at these levels (among herbivores and primary producers) that the relationship between ecosystem diversity and function is best examined at the level of functional groups.

1. Herbivores

Reef-dwelling herbivores have been grouped according to their grazing abilities on algal communities (Table II). Deep grazing (i.e., carbonate scraping herbivores), parrotfishes, and certain sea urchins (e.g., Diadema) have the greatest impact on algal abundance. Other fishes, such as surgeonfish (Acanthurus), certain damselfishes and blennies, and some gastropods, are "denuding" herbivores in that they can significantly reduce fleshy algal biomass when at sufficiently high densities. This group does not bite into limestone structures of reefs. They cannot feed on crustose corallines and have a limited ability to consume articulated algae and large leathery macroalgae. Nondenuding herbivores have little or no ability to reduce algal biomass and include territorial damselfishes and smaller herbivores, such as amphipods and polychaetes. Each of these herbivore functional groups performs distinctly different ecological roles in the ecosystem, and all were found on the species-depauperate Clipperton Atoll. The abundance of species within a functional group compensates for low diversity within the group. For example, high rates of scraping herbivory were evident at Clipperton, but all such grazing resulted from a single parrotfish species and several species of diadematid urchins. The abundant crustose coralline algae, which are often found on reefs experiencing high rates of grazing, were covered with graze marks from scraping herbivores (Table II). High rates of grazing, low fleshy algal biomass, and a high abundance of both corals and coralline algae are characteristics of relatively pristine reefs. That these conditions were found in as species depauperate a system as Clipperton Atoll suggests that ecological roles

<table>
<thead>
<tr>
<th>Region</th>
<th>Coral species</th>
<th>Reference</th>
<th>Fish species</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clipperton (Eastern Pacific)</td>
<td>7</td>
<td>Glynn et al. (1996)</td>
<td>101</td>
<td>Robertson and Allen (1996)</td>
</tr>
</tbody>
</table>

*Species richness of scleractinian corals and reef fishes.*
of functional groups can be independent of species diversity.

2. Primary Producers

Primary productivity is a significant component of ecosystem functioning. A functional approach to ecosystem study often "implies that energy flow and nutrient cycling are somehow more important or more fundamental than the biotic entities performing the function" (O'Neill et al., 1986, page 10). Although there may be no species in common, the same functional groups of marine algae are found on virtually all coral reefs. Since the productive components of reefs (algal turfs, corals, and macroalgae) all occur in approximately similar proportions, ecosystem-level productivity is fairly consistent. Thus, Caribbean reefs with only a fraction of the species richness of the western Indo-Pacific reefs have similar rates of gross primary productivity. Other factors, such as the geological history of the reefs and their proximity to land, have much greater impacts on ecosystem production than do reef biogeography or species diversity per se. Note that these patterns of consistent rates of productivity among unstrressed reefs do not negate the well-known pattern of low diversity in stressed habitats.

D. Equivalent Species and Ecosystem Redundancy

Species within a function group, by definition, are ecologically equivalent and as such provide a degree of redundancy to the ecosystem. One test of functionally equivalent species is if their abundance compensates as other species within the functional group change. Morphological and functional equivalency is evident within turf algae on coral reefs. Many genera of different divisions look similar. They often grow as prostrate branching filaments, with relatively few laterally divided "pericentral" cells. Many species among these genera (Fig. 10A) have about the same high rates of productivity (Fig. 8A) and they respond similarly to injuries. In the fore reef of St. Croix, dominance and abundance of algal species changed dramatically from season to season at the species level but very little at the functional group level (Fig. 10B). That algal community biomass remained relatively constant throughout the year, as species composition fluctuated greatly, suggests that species within functional groups compensate for species fluctuations.

It is necessary to sort out functionally important changes from redundant species "noise." On the fore reef of St. Croix (Fig. 10), seasonal changes among species were so great that it was impossible to describe it as having a characteristic dominant species. However, at the functional group level, a clear pattern of filamentous and microalgal (i.e., turf algal assemblage) dominance was evident. These algae readily coexist with reef-building corals. In contrast, macroalgae comprised a minor component of this flora and appeared only in winter and spring. The functional distinction between macroalgae and the dominant micro- and filamentous algae is important. The algal turf assemblage was composed of species which have the highest mass-specific productivity (i.e., algal functional group Nos. 1 and 2 in Fig. 8A). Trophically, these algal communities support very high levels of grazing. The herbivorous fish and urchins cannot distinguish species within the diverse algal turfs, and thus no species preferences are possible.
and there are no known chemical deterrents within this assemblage. Any of the filamentous algal species could have gone locally extinct (many did seasonally) without functionally changing the system. As such, many of the species within these functional groups are redundant.

Compensation is also evident among ecologically equivalent functional groups of herbivores. In shallow reef zones, the grazing effects from parrotfish and sea urchins are functionally similar (Table II). They are both capable of grazing deeply into limestone substrates and can control the structure of algal communities on reefs. Several studies have reported that deep-grazing sea urchins and fish compete and show compensatory changes in relative abundance when one group is excluded or removed. Hay (1984) found that reefs where fishing was important were dominated by sea urchin grazing, whereas with little fishing pressure reefs were dominated by herbivorous fishes (e.g., scraping and denuding fishes in Table II). Subsequent studies at several locations in the Caribbean and Indian Ocean found that when urchins were removed experimentally or naturally due to a disease-induced mass mortality, fish abundance and grazing increased as urchin abundance declined. In regions of the Caribbean traditionally having a relatively low abundance of *Diadema antillarum,*
another urchin, *Echinometra viridis*, was abundant and filled *Diadema*’s functional role.

**E. Ecosystem Insurance**

Species diversity within functional groups buffers ecosystems against species loss, but often there is “a threshold of change that will overwhelm the damping effect of biodiversity, with an associated break point of ecosystem function to quite different levels” (Schulze and Mooney, 1994b, page 501). Unfished reefs contain high within-functional group diversity at several levels. There are several species of large carnivores (e.g., groupers and snappers), denuding herbivores (e.g., surgeonfish), and scraping herbivores (e.g., parrotfish) that, when abundant on reefs, control the reef’s overall structure and functioning. Fishing affects reefs by first removing large-bodied carnivores and then herbivores. Reefs under particularly intense fishing, such as in Jamaica, are left with few or very small fishes. Compensatory population increases in the sea urchin *D. antillarum* probably occurred (Fig. 11), causing it to become hyper-abundant and the prime grazer of the reefs of Jamaica and elsewhere in the Caribbean. Levitan (1992) showed that *Diadema*’s increasing role as a grazer throughout the Caribbean corresponded with increasing human populations and associated fishing pressure. On many Caribbean reefs this herbivore became the primary grazer within the functional group of scraping herbivores (Table 11). Arguably, the diversity within this functional group had decreased to one species (Fig. 11). When *D. antillarum* suffered a mass mortality in 1983, the “break point” had been exceeded and reef ecosystems throughout the Caribbean collapsed.

**F. Keystone Species to Functional Groups: Some Species Are More Equal Than Others**

Some species have a disproportionately great impact on ecosystems. Hurlbert (1997) noted, “The general functional importance of a species [can be] defined as the sum, over all species, of the changes (sign ignored) in productivity which would occur on removal of the particular species from the [system]” (page 369). Most communities and ecosystems are dominated by a few species that are both abundant and important to that system (Fig. 12). The few tree species that dominate given forests, corals that dominate zones of reefs, and specific grasses that comprise specific prairies need not be functionally grouped to be easily understood. The loss of community or ecosystem dominants has a large impact on natural systems. Diseases can eliminate domi-

---

**Figure 11** Temporal changes in Jamaica’s coral reef ecosystem in coral to macroalgal abundance relative to herbivorous fish and sea urchin, *Diadema antillarum*. After herbivorous fishes were extirpated by overfishing, grazing remained high due to compensatory increases in *Diadema* abundance. The reef ecosystem collapsed soon after the mass mortality of the urchin (see text) (adapted with permission from Done, Ecological criteria for evaluating coral reefs and their implications for managers and researchers, 183–192. © 1995 Springer-Verlag).
FUNCTIONAL GROUPS

**FIGURE 12** Functional importance and abundance of species. The relatively low diversity of dominant/foundation and keystone species as well as the relatively low ecosystem importance of rare organisms make these groups poorer candidates for functional groupings (modified from Hurlbert, 1997).

In our era of rapid environmental change, the importance of functional groups is increasingly recognized. The relative importance of species within ecosystems can vary significantly due to the ecological and evolutionary processes. The stability of ecosystems is directly related to the diversity and function of species, and the loss of any significant functional group can have profound effects on the ecosystem's overall health.

For example, the American chestnut, a dominant organism, was severely impacted by chestnut blight, which drastically reduced its population. Similarly, the humpback whale, a keystone species, plays a crucial role in maintaining the health of the marine ecosystem. Its absence due to hunting and habitat loss can have cascading effects on other species and ecosystem processes.

**VII. CONCLUSIONS**

Despite Earth's remarkable biodiversity, the range of functionally different organisms is surprisingly low. Phylogenetic constraints on evolution relative to biomechanical limits of what works in nature limit the number of functional groups in any system to a management level. In highly diverse ecosystems there are many redundant species within functional groups. These redundancies provide a buffer against ecosystem collapse should individual species within the group become rare or extinct. Ultimately, functional groups may provide a low-resolution tool for accurately predicting ecosystem change. Research to improve how we identify and define functional groups is likely to increase our understanding of the structure and functioning of natural communities and ecosystems.

**Acknowledgments**

Drafts were read by E. Amin, A. Leland, J. Vawrinec, and S. Zimmen, with additional help from F. Yund and L. Wadling. Greg Weik helped with manuscript preparation. To all, I am grateful.

**See Also the Following Articles**
- BIODIVERSITY GENERATION, OVERVIEW • CLADOGENESIS • COEVOLUTION • GUILDS • MAMMALS, BIODIVERSITY OF • REEF ECOSYSTEMS

**Bibliography**

Fungi

Thomas J. Volk

University of Wisconsin–La Crosse

I. Introduction
II. What Are Fungi?
III. How Fungi Grow
IV. Reproductive Biodiversity of Fungi
V. Fungi in the Environment
VI. Industrial Uses for Fungi
VII. Threatened or Endangered Fungi

GLOSSARY

ectomycorrhiza (pl. ectomycorrhizae; adj., ectomycorrhizal) Part of a mutualistic relationship between a fungus (usually a basidiomycete, but sometimes an ascomycete) and a host plant in which hyphae aggregate as an extra surface around roots of the plant, aiding nutrient transfer between the plant and fungus.

endomycorrhiza (pl. endomycorrhizae; adj. endomycorrhizal) Part of a mutualistic relationship between a fungus (a zygomycete) and a host plant in which no sheath is formed. Nutrient transfer occurs with the aid of highly branched arbuscules.
exoenzymes Digestive enzymes excreted by fungi into the environment to digest materials externally.
fruiting body Sexual reproductive structure of a fungus.
fungus A member of the kingdom Fungi. Members of this kingdom are heterotrophic (requiring a preformed organic source, i.e., not able to make their own food), eukaryotic, have walls of chitin, and reproduce by means of spores.
karyogamy Fusion of nuclei.
mushroom Vernacular word for a large, fleshy fruiting body consisting of a cap with gills or pores on the undersurface (or sometimes flat or with teeth or folds), usually on a stalk, and producing sexual spores. The term is usually reserved for members of the Basidiomycota.
mutilalism Two organisms living together for the mutual benefit of both.
mycelium (pl. mycelia) Vegetative filamentous body of a fungus; mass of hyphae from a single individual.
parasite Organism that obtains its nutrition from a living host, and in so doing harms the host.
plasmogamy Fusion of the cytoplasm of two cells.
spore One- to several-celled propagule of totipotent cytoplasm with cell walls, produced by cell division with concomitant meiosis or mitosis, that may serve for dispersal or overseeding, but that does not contain an embryo.
substrate Material or host from which a fungus derives its nutrition.
toadstool Vernacular name for a poisonous mushroom.

1 www.wisc.edu/botany/fungi/volkmyco.html

WHEN SOMEONE MENTIONS “FUNGI” you may think immediately of mushrooms on pizza or maybe
moldy food in your refrigerator. But in fact fungi are everywhere and affect our lives every day, from edible mushrooms to industrially important products to plant helpers, plant pathogens, and human diseases.

I. INTRODUCTION

Many fungi are detrimental in causing a large number of plant diseases that result in the loss of billions of dollars worth of economic crops each year. They also cause a number of animal diseases, including many human maladies. On the other hand, there is a long and rapidly growing list of useful fungi. Fungi have been used in the preparation of foods and beverages for thousands of years, and there are many mushrooms that are edible. Industry has used other fungi in the manufacture of many valuable organic compounds, including organic acids, vitamins, antibiotics, and hormones. They have been used in the research laboratory to study metabolic pathways, mineral nutrition, genetics, and a variety of other problems. But perhaps their greatest contribution has been, and continues to be, their role in recycling carbon and other essential elements in the ecosystem. Because all fungi are heterotrophic (see Section V), they rely on organic material, either living or dead, as a source of energy. Thus, many are excellent scavengers in nature, breaking down dead animal and vegetable material into simpler compounds that become available to other members of the ecosystem.

“Mycology” comes from the Greek words mykos, which means “fungi,” and -logy,” which means “the study of.” Mycologists study many aspects of the biology of fungi, usually starting with their systematics, taxonomy, and classification, and continuing on to their physiology, ecology, pathology, evolution, genetics, and molecular biology. There are quite a few aspects of applied mycology, such as plant pathology, human pathology, fermentation, mushroom cultivation, and many other fields.

II. WHAT ARE FUNGI?

A. Characteristics of Fungi That Separate Them from the Other Kingdoms

The most significant characteristics of fungi that place them apart from the plant, animal, and other kingdoms are the following:

- Fungi are eukaryotic.
- They are nonvascular organisms, meaning they have no specialized transport tissues.
- Most reproduce by means of spores, usually wind-disseminated, and occasionally by insects in some species.
- Both sexual (meiotic) and asexual (mitotic) spores may be produced, depending on the species and conditions.
- Fungi are typically not motile, although a few (e.g., chytrids) have a motile phase.
- Like plants, sexually reproducing fungi have an alternation of generations, although the generations may be different from those of plants.
- The fungal vegetative body may be unicellular (yeasts) or, more often, composed of microscopic threads called hyphae.
- Fungal cell walls are similar in structure to those of plants but differ in chemical composition—fungal cell walls are composed mostly of chitin, which are $\beta-1,4$ linkages of $n$-acetylglucosamine. In contrast, plant cell walls are composed mostly of cellulose, which are $\beta-1,4$ linkages of glucose. Many plants also contain lignin in their secondary walls.
- Fungal cytoplasmic ultrastructure is broadly similar to that of plant cells, but differs significantly in the kinds of organelles and their structures.
- Fungi are heterotrophic (i.e., “other feeding” in that they must feed on preformed organic material), not autotrophic (i.e., “self-feeding” in that they make their own food by photosynthesis, like most plants and algae).
- Unlike animals (which are also heterotrophic), which ingest and then digest, fungi first digest and then ingest. Fungi produce exoenzymes to accomplish this.
- Most fungi store their food as glycogen (like animals)—plants store food as starch.
- Fungal cell membranes have a unique sterol, ergosterol, which replaces the cholesterol found in mammalian cell membranes.
- The lysine biosynthesis pathway in fungi is different.
- The microtubule protein formed during nuclear division is different from that of all other organisms.
- Most fungi have very small nuclei, with little repetitive DNA, usually with few chromosomes.
- Mitosis and meiosis are generally accomplished without dissolution of the nuclear envelope.

There are about 70,000 named species of fungi and this is believed to be about 5% of the total number of species that exist in nature. If this is the case, 95% of
all fungal species are unknown to science and do not yet have names!

In the outdated two-kingdom system, fungi were included in the plant kingdom, as were almost all walled organisms that are not motile. In almost all systematic schemes today, the heterotrophic and eukaryotic fungi are placed in their own kingdom, cleverly called the kingdom Fungi.

The eukaryotes, organisms with nuclei and membrane-bound organelles, can be divided into five kingdoms: Fungi, Plantae, Animalia, Protista, and Strameno-pila. Recent DNA evidence suggests that fungi are more closely related to the animals than to plants. Prokaryotes, without nuclei and membrane-bound organelles, include the Bacteria and the Archaea.

As with all taxonomy, the names of various taxa of fungi each have a specific ending that refers to their taxonomic level. The fungal kingdom has its own endings for many taxa; these include the suffix -myc. It is of considerable advantage to be familiar with these endings so that one knows immediately what taxonomic level is being referred to when confronted with a particular name. The names of fungi are “regulated” by the International Code of Botanical Nomenclature, and thus endings for the order and family levels are the same as those of plants (Table I).

B. Fungal Roles in the Ecosystem

Lack of chlorophyll profoundly affects the lifestyle of fungi. Since they are not dependent on light, they can occupy dark habitats and can grow in any direction. The exoenzymes allow them to invade the interior of a substrate with absorptive filaments. Fungi may gain their nutrition from dead organisms, in which case they are called saprophytes. Some fungi derive their nutrition from living organisms; these are called symbionts. Symbionts can be further divided on the basis of whether or not they harm their host. Parasites cause harm to the host, whereas mutualists engage in a reciprocally beneficial association with their host. These categories are further described in Section V.

C. Importance of Studying Fungal Biodiversity

Fungi affect human lives in many and varied ways, so it is important to know something about the fungi to be able to control or exploit them for our own purposes. For example, more than 90% of known fungal species have never been screened for antibiotics or other useful compounds. However, even more important is the role that fungi play in the ecosystem. They are a vital part of the links in the food web as decomposers and pathogens and are important in grassland and forest ecosystems alike. Fungi have many different kinds of associations with other organisms, both living and dead. To learn more about the impact that fungi have on our lives, we must learn a lot more about them.

Many fungi are harmful to human interests. They can cause human disease, either directly or through their toxins, including mycotoxins and mushroom poisons. They can also cause diseases of plants and animals (e.g., crops, fruit trees, farm animals). Very often fungi cause rot and contamination of foods—most of us probably have something green and moldy in the back of our refrigerator right now. They can destroy almost every kind of manufactured good, with the exception of some plastics and some pesticides.

On the other hand, many fungi are very useful to humans. Of course, there are many edible mushrooms. Yeasts have been used for baking and brewing for many millennia. Antibiotics such as penicillin and cephalosporin are produced by fungi. The immunosuppressive antirejection transplant drug cyclosporin is produced by the ascomycete Tolypocladium inflatum. Steroids and hormones—and even birth control pills—are commercially produced by various fungi. Many organic acids are also commercially produced with fungi, for example, citric acid in cola and other soda pop products is produced by an Aspergillus species. Some gourmet cheeses such as Roquefort and other blue cheeses, brie, and camembert are fermented with certain Penicillium species. Stone-washed jeans, strange as it sounds, are softened by Trichoderma species. There are likely many more potential uses that have not yet been explored.

Fungi are also important experimental organisms. They are easily cultured, occupy little space, multiply rapidly, and have a short life cycle. Since they are eukaryotes and more closely related to animals, their study is more applicable to human problems than is the study

<table>
<thead>
<tr>
<th>Taxonomic level</th>
<th>Taxonomic ending</th>
<th>White button</th>
<th>pizza mushroom</th>
<th>The morel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Eukarya</td>
<td>Eukarya</td>
<td>Eukarya</td>
<td>Eukarya</td>
</tr>
<tr>
<td>Kingdom</td>
<td>Fungi</td>
<td>Fungi</td>
<td>Fungi</td>
<td>Fungi</td>
</tr>
<tr>
<td>Phylum</td>
<td>-mycota</td>
<td>Basidiomycota</td>
<td>Ascomycota</td>
<td>Ascomycota</td>
</tr>
<tr>
<td>Class</td>
<td>-mycetes</td>
<td>Hymenomycetes</td>
<td>Discomycetes</td>
<td></td>
</tr>
<tr>
<td>Order</td>
<td>-ales</td>
<td>Agaricales</td>
<td>Pezziales</td>
<td></td>
</tr>
<tr>
<td>Family</td>
<td>-aceae</td>
<td>Agaricaeace</td>
<td>Morchellaceae</td>
<td></td>
</tr>
<tr>
<td>Genus</td>
<td>—</td>
<td>Agaricus</td>
<td>Morchella</td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>—</td>
<td>Agaricus bisporus</td>
<td>Morchella esculenta</td>
<td></td>
</tr>
</tbody>
</table>
of bacteria. Fungi are used to study metabolite pathways, to study growth, development, and differentiation, for determining mechanisms of cell division and development, and for microbial assays of vitamins and amino acids. Fungi are also important genetic tools; the “one gene one enzyme” theory in *Neurospora* won George W. Beadle and Edward L. Tatum the Nobel Prize for Physiology or Medicine in 1958. The first eukaryotic genome to have its DNA sequenced was that of the bakers’ and brewers’ yeast, *Saccharomyces cerevisiae*.

**III. HOW FUNGI GROW**

**A. Biology of Hyphae and Yeast Forms**

A fungus is more than just the visible mushroom structure. In fact the mushroom, more properly called a fruiting body, is a very small portion of the individual life cycle and is mainly used for reproduction. The major portion of the life cycle, or the vegetative growth form, in the great majority of fungi consists of a system of threadlike, walled, more or less cylindrical hyphae (singular, hypha) making up what is called a mycelium (plural, mycelia) (Fig. 1). The Ascomycota and Basidiomycota have crosswalls called septa (singular, septum) separating compartments of the mycelium. An exceptional group is the yeasts, which consist of about 800 species that have a single-celled vegetative form. Note that yeast is a morphological term and has no taxonomic significance; yeasts and yeastlike forms can be found in all of the fungal phyla.

**B. Exoenzymes and the Heterotrophic Lifestyle**

Exoenzymes are the most important reason why fungi are so successful. Fungi excrete exoenzymes at the tips of the growing hyphae into their surrounding environment, where they play a major role in breaking down the substrate. Simpler molecules can then move into the hyphae by diffusion.

**IV. REPRODUCTIVE BIODIVERSITY OF FUNGI**

**A. Fungal Life Cycles**

The major events of any sexual life cycle are plasmogamy (cell and protoplast fusion), karyogamy (nuclear fusion), and meiosis (Fig. 2). In most other familiar types of organisms, such as plants and animals, plasmogamy and karyogamy occur in rapid succession and are usually referred to as the single event of fertilization. In the fungi, however, plasmogamy and karyogamy may be separated in time by several minutes, several hours, several days, several years, or even several centuries! Thus the dikaryon, the n+n stage, is a major component of the life cycle of fungi, especially in the Basidiomycota and Ascomycota. Nuclear cycles of all the members of the various phyla can be placed within this generalized nuclear cycle, differing mainly in the amount of time spent in each of the phases.

**FIGURE 1** Hyphae with dark-stained nuclei.
Generalized Nuclear Cycle of Fungi

Besides this sexual cycle, many fungi, commonly called molds, can also reproduce asexually (mitotically) in the absence of meiosis and karyogamy. Many of them produce specialized structures that bear the asexual spores. As in much of biology, there is some “compet- ing” terminology here—the asexual state is also known as the anamorph or mitosporic state. Asexual reproduction can take place at any point in the life cycle (haploid, diploid, or dikaryon), depending on the species and conditions. The sexual state is also known as the tel-eomorph or meio sporic state.

B. Phyla of Fungi

Based primarily on variation in their sexual reproductive structures, the kingdom Fungi is usually divided into four major phyla.

• Chytridiomycota—sexual and asexual spores are motile, with posterior flagella.
• Zygomycota—sexual spores are thick-walled rest- ing spores called zygospor es, and asexual spores called sporangiospores (when present) are borne internally in structures called sporangia.

FIGURE 3  *Rhizopus* sporangia.
• Ascomycota—sexual spores (ascospores) are borne internally in a sac called an ascus, and asexual spores called conidia (when present) are borne externally on structures called conidiophores.
• Basidiomycota—sexual spores (basidiospores) are borne externally on a club-shaped structure called a basidium; most species do not have asexual spores, but when present they often take the form of conidia.

There is an additional group, the deuteromycetes, or “Fungi Imperfecti,” for which there is no known sexual state. Its members have affinities to members of at least three of the phyla, especially the Ascomycota.

1. Chytridiomycota
The Chytridiomycota, commonly called the chytrids, are a group of mostly water-inhabiting organisms, although some are plant pathogens. In aquatic environments they mostly form scanty filaments with sporangia. Some examples of the Chytridiomycota are Allomyces, a water mold, Synchytrium endobioticum, a pathogen of potato, and Neocallimastix, a chytrid that lives symbiotically in the gut of herbivores, such as cattle. Batrachochytrium dendrobatidis and possibly other chytrids have been implicated as an infection associated with the worldwide decline in frog populations.

2. Zygomycota
Commonly called the bread molds, the Zygomycota are terrestrial fungi whose fruiting bodies are mostly microscopic in nature, although their asexually produced sporangia can reach greater than 5 cm tall in some species (Fig. 3). Under certain conditions they may sexually produce thick-walled resting spores called zygospores (Fig. 4). Some, such as Rhizopus, Mucor, and Phycomyces, can grow on a wide variety of substrates, and a few can act as human pathogens. Most importantly, members of one order, the Glomales, are responsible for forming mutualistic associations called endomycorrhizae with the roots of about 70% of the world’s plants. Ectomycorrhizae (from Basidiomycota and Ascomycota) form with another 20% of plant species. See Section V,D,1 for further discussion of mycorrhizae.

3. Ascomycota
The Ascomycota bear their sexual spores (ascospores) in sacs called asci, which are usually cylindrical. Many
members also form conidia as asexual spores. Familiar members of this phylum include the morels and other cup and saddle fungi, powdery mildews, the industrial yeast *Saccharomyces cerevisiae*, the incitant of chestnut blight (*Cryphonectria parasitica*), the cause of Dutch elm disease (*Ophiostoma ulmi*), and a variety of other plant pathogens (Figs. 5 and 6).

4. Basidiomycota

The Basidiomycota bear their sexual spores externally on a usually club-shaped structure called a basidium, which is often borne on or in a fruiting body called a basidiocarp or basidiome (Fig. 7). This phylum includes the well-known mushrooms, both edible and poisonous, as well as puffballs, shelf fungi, jelly fungi, and coral fungi (Fig. 8). These species that produce fruiting bodies exhibit various methods of increasing their surface area, as discussed in Section IV,C.

The Basidiomycota also contain perhaps the most important plant pathogens, the rusts and the smuts. These fungi do not produce macroscopic fruiting bodies, but instead bear their spores on the stems, leaves, and flowers of host plants. However, remember that the mycelium is internal and “sucks” the nutrients out of the plant. Effects on the plant range from a reduced yield to death. Rusts in particular have very complicated life cycles, often requiring two unrelated host species to complete their growth stages.

5. “Deuteromycetes,” the Fungi Imperfecti

The deuteromycetes, commonly called molds, are “second-class” fungi that have no known sexual state in their life cycle, and thus reproduce only by producing spores via mitosis. About 90% of these have affinities to the Ascomycota. Most food spoilage and fungal human diseases are caused by members of this group (Figs. 9 and 10). They are also known as the fungi imperfecti, because of their “imperfect” lack of sex. When the “perfect state” of one of these organisms is discovered, as happens every year, the fungus is more properly classified with the teleomorph name. Notice that this group is not classified as one of the phyla. It is just a loose assemblage of organisms that we are not sure where to place accurately in the taxonomic order.
FIGURE 8  (A) Armillaria mabsnona, (B) Tremella reticulata, (C) Trametes versicolor, (D) Pulcherricum caeruleum. See also color insert, Volume I.
now placed in the kingdom Stramenopila (or Protista according to some authors). Ecologically many of its members act like fungi, especially the plant pathogens. *Phytophthora infestans* causes a disease called late blight of potato, which was the cause of the Irish potato famine in the 1840s, in which more than a million Irish people perished and another million emigrated. The downy mildews *Peronospora parasitica* and *Plasmopara viticola* cause diseases of members of the cabbage family and of grapes, respectively. *Pythium* species cause damping-off disease of seedlings in agricultural practice. However, many other species are innocuous saprophytes that decompose debris in water. A few of these so-called water molds (some species of *Saprolegnia* and *Achlya*) are opportunistic fish parasites, especially in aquariums and fish hatcheries (Figs. 11 and 12). They are a particular problem in trout and salmon hatcheries.

b. **Myxomycota: True Slime Molds**

The Myxomycota are the true slime molds, also known as the plasmodial slime molds, and are considered members of the kingdom Protista. They exist in nature as a plasmodium—a blob of protoplasm without cell walls and only a cell membrane to keep everything in (Fig. 13). It is really nothing but a large amoeba and feeds much the same way, by engulfing its food (mostly bacteria) with pseudopodia, in a process called phagocytosis. So the slime mold ingests its food, then digests it. True fungi have a cell wall and digest their food with exoenzymes before ingesting it. When the plasmodium runs out of food it can form fruiting bodies. Most slime mold fruiting bodies are quite small, 1–4 mm in height, but
FIGURE 11 Saprolegnia oogonium containing oospores.

FIGURE 12 Saprolegnia sporangium containing small zoospores.

FIGURE 13 Plasmodium of a slime mold. See also color insert, Volume 1.
some can be up to 25 cm in diameter (Fig. 14). In such cases a large plasmodium may be seen crawling along the ground, inspiring science fiction movies such as “The Blob.” These slime molds have been traditionally studied by mycologists because their small, delicate fruiting bodies tend to be fungal in appearance. Most slime mold fruiting bodies are quite beautiful.

c. Dictyosteliomycota: Cellular Slime Molds
The Dictyosteliomycota are the cellular slime molds or “social amoebae” and are among the most bizarre of microorganisms. These members of the Protista are free-living amoebae with no cell walls, indistinguishable from garden-variety amoebae until they begin to run out of food. At that point they signal to one another using cyclic AMP (a small nucleotide molecule) and begin to aggregate to form a “slug” or pseudoplasmodium. This slug, composed of hundreds of amoebae, acts as a single organism and can actually migrate along a light or temperature gradient. Eventually the slug stops migrating, rounds up, and forms a sorus, a kind of sporangium containing spores on a stalk. Not all of the cells become spores; some of them “sacrifice” themselves to become stalk cells to raise the spores up into the air for a better position for wind dispersal. So formerly free-living organisms act like a single organism for the good of the species. Very strange indeed!

d. Plasmodiophoromycota: Endoparasitic Slime Molds
The Plasmodiophoromycota are an odd group of endoparasites that live almost their entire life cycle inside a cell of another organism. They lack cell walls in the assimilative state and reproduce by means of swimming spores, the only part of their life cycle that does not occur inside a cell. *Plasmodiophora brassicae* causes club root in crucifers, and *Spongospora subterranea* causes powdery scab of potatoes. Scientists really do not know where to place these organisms taxonomically; they have been allied with the Oomycota, the Myxomycota, or various other protists, but in reality they are not closely related to any other known group of organisms. No one is even sure what kingdom to place them in!

C. Surface Area and Reproduction

Many fungi have very specialized habitats. For example, the basidiomycete *Suillus americanus* can grow only in association with the eastern white pine, *Pinus strobus*. To infect a new white pine, the immobile mycelium of the fungus must produce spores to move to a new host. These spores are carried by the wind, not by some specific insect or other animal vector. Thus the fungus must produce enormous numbers of spores so that a few of them will land on the “correct” substrate. The underside of the *Suillus* mushroom has small pores that are lined with microscopic spores. This increases the surface area for bearing spores by more than 100 times. More commonly, mushrooms have gills underneath the cap to increase surface area. Other fungi increase their surface area by forming upright coral-like branches, whereas others form downward-pointing teeth or spines.

![Figure 14](image-url)

**FIGURE 14** (A) *Lycogala epidendrum*, (B) *Comatricha typhoides*, (C) *Badhamia utricularia*, (D) *Dictydiom cancellatum*. 
FIGURE 15  Different ways of increasing the surface area for bearing spores. Clockwise from upper left: pores, upright branches (coral), teeth, and gills.
(Fig. 15). Others (not shown) have blunt ridges and some have folds, and some increase the time of sporulation rather than the spore-bearing physical space. Analogous surface area modifications are made by members of other phyla of fungi.

V. FUNGI IN THE ENVIRONMENT

Fungi occupy many different niches in the environment, although any one species usually occupies only a single niche. Many niches have not yet been explored for fungi, which is the major reason why mycologists believe there are high number of fungi yet to be discovered. Fungi can be divided into groups based on their nutritional status and the nature of their relationship with their host. Saprophytes use nonliving organic material and are important scavengers in ecosystems. Along with bacteria, fungi are important in recycling carbon, nitrogen, and essential mineral nutrients. Parasites use organic material from living organisms, harming them in some way. Their hosts range from single-celled diatoms to other fungi, plants, animals, and humans. Fungi are the major parasites (pathogens) of plants. Mutualists are fungi that have a reciprocally beneficial relationship with other living organisms, in which both organisms benefit. The two main types of mutualistic associations involving fungi are mycorrhizae, which are associations of fungi with plants roots, and lichens, which are associations of fungi with algae or cyanobacteria. There are also a few commensal fungi that use other organisms as merely a place to live; these fungi derive no nutrition from their host.

A. Fungi as Saprophytes

Along with bacteria, fungi are the major decomposers and recyclers in the environment. For every sort of dead material present, there is usually at least one fungus that can degrade that material. A few exceptions include some pesticides and some types of plastics; no fungi have yet developed exoenzymes capable of digesting these synthetic materials. Fungi are important in breaking down carbon- and nitrogen-containing compounds into components that they and other organisms can use. Fungi are especially important in the breakdown of the wood components cellulose and lignin (discussed in Section VI,F).

B. Fungi as Plant Parasites

About 90% of plant diseases are caused by fungi, resulting in billions of dollars in crop and forest losses each year. A parasite is referred to as a pathogen if it elicits a recognizable and persistent disease. The most severe pathogens are those that are imported accidentally from other continents and that cause severe problems with the native populations and cultivated plants. The severity of any plant disease is determined by three factors, known as the disease triangle, consisting of the interaction of the host (conditions favoring susceptibility), pathogens (conditions affecting virulence), and the environment (conditions favoring disease) (Fig. 16). All three factors must work in conjunction to produce a disease and determine its harshness.

1. Fungi in Crop Pathology

The science of plant pathology deals mostly with organisms, especially fungi, that cause plant disease. Plant diseases are generally classified on the basis of what kinds of symptoms occur in which part(s) of the plant (Table II). Some important fungal pathogens of crop plants include Puccinia graminis (black stem rust of wheat), Erysiphe graminis (powdery mildew of grasses), Claviceps purpurea (ergot), and Ustilago maydis (corn smut).

2. Fungi in Forest Pathology

There have been a number of fungal diseases of forest trees that have caused great problems in North America (Table III). Forest pathogens often need a longer time to kill their larger, perennial hosts, sometimes living inside the tree for several decades before the host finally succumbs. Crinipellis parasitica, Ophiostoma ulmi, and Discosphaera destructiva are introduced pathogens that have dev-

FIGURE 16 The disease triangle, showing the three factors that determine the disease severity.
TABLE II
Types of Plant Diseases

<table>
<thead>
<tr>
<th>Type of disease</th>
<th>Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blight</td>
<td>Extensive necrosis and rapid death</td>
</tr>
<tr>
<td>Dieback</td>
<td>Dead branches protruding from apparently healthy tissue</td>
</tr>
<tr>
<td>Lesion</td>
<td>Localized cell death</td>
</tr>
<tr>
<td>Canker</td>
<td>Lesion on a woody stem</td>
</tr>
<tr>
<td>Rot</td>
<td>Extensive decay of tissue</td>
</tr>
<tr>
<td>Vascular wilt</td>
<td>Blockage of the vascular transport system, resulting in starvation or disiccation</td>
</tr>
<tr>
<td>Gall</td>
<td>Localized tumorlike growth</td>
</tr>
<tr>
<td>Stunting</td>
<td>Reduction in overall growth size</td>
</tr>
<tr>
<td>Rusts</td>
<td>Rust-colored lesions on various parts of the plant</td>
</tr>
<tr>
<td>Smuts</td>
<td>Drastic malformations, usually of reproductive parts</td>
</tr>
</tbody>
</table>

astated their hosts in North America. In their native Asia, these pathogens coevolved with their hosts and caused them only minor annoyance. However, when they were accidentally introduced into North America, either on live plants or on logs for sawing, the fungus escaped and subsequently devastated the North American tree species because there was not enough time for the host to develop resistance. *Cronartium ribicola*, *Ceratocystis fagacearum*, and *Armillaria* species are native American fungi, but they cause severe diseases nonetheless. A good pathogen does not kill its host right away, but keeps it alive as long as possible to continue deriving nutrients from it.

C. Fungi as Animal Parasites

1. Fungi Affect Humans

Fungal effects on humans are classified by whether they can grow on the body (mycoses) or whether they cause problems by their ingestion. Mycetismus refers to the eating of poisonous mushrooms, and mycotoxicosis refers to the ingestion of toxins produced by the fungus, not involving eating the fungus itself. Mycoallergies may be a separate category.

a. Mycosis: Fungi Growing Directly on Human Tissues

There are a number of fungi that can grow directly on human tissue. Most of them require some debilitation of host defenses, especially the immune system. Fungal diseases have gained in importance over the last couple of decades because of increases in the numbers of such patients, including those with AIDS, people on corticosteroid therapy or immunosuppressive drugs (such as the transplant drug cyclosporin), diabetics, and people undergoing chemotherapy treatment for cancers.

About 175 human pathogens are recognized among the approximately 70,000 known species of fungi. Around 20 are regularly isolated from cutaneous infections (dermatophytes and yeasts), a dozen are associated with severe subcutaneous localized disease, and

TABLE III
Some Important Forest Pathogens

<table>
<thead>
<tr>
<th>Pathogen</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Cryphonectria parasitica</em></td>
<td>Chestnut blight</td>
</tr>
<tr>
<td><em>Ophiostoma alni</em></td>
<td>Dutch elm disease</td>
</tr>
<tr>
<td><em>Discula destructiva</em></td>
<td>Dogwood anthracnose</td>
</tr>
<tr>
<td><em>Cronartium ribicola</em></td>
<td>White pine blister rust</td>
</tr>
<tr>
<td><em>Ceratocystis fagacearum</em></td>
<td>Oak wilt</td>
</tr>
<tr>
<td><em>Armillaria</em> spp., especially <em>A. mellea</em> and <em>A. ostoyae</em></td>
<td>Armillaria root rot</td>
</tr>
</tbody>
</table>
about 20 may cause systemic infections (Table IV). In addition, there are many opportunistic pathogens that cause disease in debilitated or immunosuppressed patients.

Probably the most common fungal infection in North America is yeast infection, caused by *Candida albicans*. This yeast can grow on outside portions of the skin or in any area of the body that is moist and warm. It is especially prevalent in the genital area of women. Oddly enough *Candida* is part of the normal flora of the body and can be isolated from almost everyone at any time. Only when conditions get out of balance does the yeast flourish and cause disease. However, the systemic true pathogens are the most feared, since they do not require a debilitation of the host defenses to become pathogenic. These dimorphic fungi get around host defenses by changing their form from a mycelium to a yeast upon change in temperature, thus evading even the healthiest of immune systems. Fortunately they are all geographically restricted. (see also Rippon, 1988.)

b. Mycetismus: Eating Poisonous Mushrooms

Of the 70,000 species of fungi, there are about 250 species of delicious edible mushrooms and about 250 species that will kill you. Most of the other species are innocuous or unpleasant tasting. Some people call poisonous mushrooms “toadstools,” a word probably derived from the old German *Todstuhl* for death chair. But how can you tell the difference? There is no easy way to tell if a mushroom is poisonous. Old wives’ or old husbands’ tales about silver spoons or whether it grows on wood or whether animals eat the mushrooms do not work! The only way to be sure if a mushroom is edible is to identify it to species by consulting books or knowledgeable persons. The old saying is “there are old mushroom hunters, and there are bold mushrooms hunters, but there are no old, bold mushroom hunters.” The best way to learn about edible mushrooms is to join a local mushroom club. Do an internet web search for the North American Mycological Association for a list of these clubs. (NOTE: The pictures and descriptions in this article are NOT adequate to identify mushrooms for eating. We have a report of a woman who nearly died from eating wild mushrooms that she identified only from an encyclopedia drawing. Don’t let this be you! Be absolutely sure of your identifications!)

As shown in Table V, unrelated species in various genera may contain the same toxin, and one or more toxin groups may be found in the same genus. To complicate things further, many genera contain both poisonous and nonpoisonous species, although several genera are more or less homogeneous with respect to poisoning ability. Most species of fungi have not yet been analyzed for toxins. Evidence suggests that mushroom toxins have not evolved for protection. For example, what good does it do a mushroom such as *Amanita virosa*, known as the death angel, if an animal eats the mushroom and then dies two days later? The toxins

<table>
<thead>
<tr>
<th>Table IV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mycoses Can Be Characterized by Their Location on the Body</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of mycosis</th>
<th>Location</th>
<th>Some diseases</th>
<th>Some fungi causing disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superficial mycoses</td>
<td>Infections of the hair shaft or dead outer layer of skin (stratum corneum)</td>
<td>Pityriasis versicolor, tinea nigra palmaris, piedra</td>
<td>Malassezia furfur, Trichosporon beigelli, Cladosporium, Piedraia hortae</td>
</tr>
<tr>
<td>Cutaneous mycoses</td>
<td>Dermatophytes—infections of skin, hair, and nails</td>
<td>Ringworm, jock itch, athlete’s foot, tinea corporis, tinea capitis, and many others</td>
<td>Microsporum, Trichophyton, and Epidermophyton spp., occasionally a Candida sp.</td>
</tr>
<tr>
<td>Subcutaneous mycoses</td>
<td>Chronic localized infections of the skin and subcutaneous tissues</td>
<td>Sporotrichosis, chromoblastomycosis, phaeohyphomycosis</td>
<td>Sporothrix schenckii, Phialophora, Cladosporium</td>
</tr>
<tr>
<td>Systemic mycoses (deep mycoses), true human pathogens</td>
<td>Fungal infections of the body caused by dimorphic fungal pathogens, usually entering through the lungs</td>
<td>Histoplasmosis, Blastomycosis, Coccidioidomyositis, Paracoccidioidomycosis</td>
<td>Histoplasma capsulatum, Blastomyces dermatitidis, Coccidiodes immitis, Paracoccidioides brasiliensis</td>
</tr>
<tr>
<td>Systemic mycoses, opportunistic pathogens</td>
<td>Fungal infections of the body caused by common fungi becoming pathogenic owing to patient debilitation</td>
<td>Aspergillosis, Cryptococcosis, Candidiasis, Zygomycosis, Pneumocystis pneumonia, Penicilliosis</td>
<td>Aspergillus spp., Cryptococcus neoformans, Candida spp., Mucor, Rhizopus, Pneumocystis carinii, Penicillium marneffei</td>
</tr>
</tbody>
</table>
appear to be merely waste products, usually from nitrogen metabolism, that happen to be poisonous to animals or humans.

c. Mycotoxicosis: Eating Toxins Produced by Fungi, not Necessarily from Eating the Fungi Themselves

Mycotoxins are usually produced by molds growing on foods. As the molds grow, they metabolize the food product and excrete their waste products back into the substrate. Some of these waste products happen to be highly toxic, and when a person eats the toxin (even after cutting the mold off), there is usually some harm to the person. The common contaminants Aspergillus flavus and A. parasiticus can produce aflatoxins, especially in peanuts. If you have ever eaten a whole peanut that tastes so bitter that you spit it out, it was probably contaminated with aflatoxin. This compound is highly carcinogenic at about 10–20 parts per billion, and kills very rapidly at higher concentrations. There are legal limits for the amount of aflatoxin allowed in peanut butter sold in the United States.

2. Fungi Affect Other Animals

Nonhuman animals are affected by fungi in much the same way that humans are, with some differences in certain species. For example, dogs are susceptible to cyclopeptide poisoning while cats apparently are not. There are numerous other examples of this discrepancy—so do not pick and eat a mushroom just because you see an animal eat it. Another important point to consider is that you do not know what happened to that squirrel after it ate that mushroom.

Animals are also affected by many of the same fungal diseases as are humans. For example, dogs are particularly susceptible to blastomycosis and are often used as a warning sentinel, like a canary in a mine, for alerting humans to possible risk.

D. Fungi as Mutualists with Other Organisms

Rather than being harmful, some fungi benefit their host in some way while receiving nutrients from them.
The two most common mutualistic associations are mycorrhizae, an association between a fungus and the roots of a plant, and lichens, an association between a fungus and either an alga or a cyanobacterium or both.

1. Mycorrhizae

According to Harley and Smith (1983), a mycorrhiza is defined as “an association between a fungus and a host plant in which destructive disintegration of the host does not occur and which is a prevalent and usual condition of the host plant in natural habitats and as such is very common and widespread.” Myco, of course, means “fungus,” and rhiza is “root,” so mycorrhiza literally means “fungus root.” More than 90% of plants in nature have a mycorrhizal symbiont. The only groups of plants that regularly lack mycorrhizae are some crucifers, sedges, and some legumes.

There are several types of mycorrhizae. In ectomycorrhizae, the fungus, usually a Basidiomycota or sometimes an Ascomycota, forms a sheath outside the root (Fig. 17). Exchange of nutrients takes place in transfer cells called a Hartig net, which penetrates between the cells of the root cortex, but does not penetrate the cells themselves. In endomycorrhizae, also called vesicular-arbuscular mycorrhizae (VAM) or sometimes simply arbuscular mycorrhizae, no sheath is formed. This fungus is always a member of the Zygomyctota. Nutrient exchange takes place in highly branched hyphae called arbuscules, which penetrate into the cortical cells, but do not penetrate the cell membrane. Orchid mycorrhizae and ericoid mycorrhizae are special types that are found with plants in the Orchidaceae and the Ericaceae, respectively, and they differ significantly in their structure and life strategies.

In all types of mycorrhizae, hyphae extend from the root into the surrounding soil, greatly increasing the surface area for absorption of nutrients, particularly phosphate, nitrogen, and potassium. In return for shunting some of these nutrients into the plant, the fungus receives some sugars from plant photosynthesis. Thus both organisms benefit. Mycorrhizal fungi are abundantly represented in fossils from the Devonian and later periods and apparently coevolved with their hosts. It has been hypothesized that these fungi were necessary for the movement of water plants onto land.

As an interesting sidelight, there are also several hundred species of nonphotosynthetic plants (such as Monotropa uniflora, the Indian pipe) that get their energy as parasites of fungi that are mycorrhizal with photosynthetic plants. Radioactive carbon has been used to trace nutrient flow from the host plant through the mycorrhizal fungus and into the achorophyllous plant.

2. Lichens

A lichen is a dual organism that consists of a mutualistic relationship between a fungus (the mycobiont) and an alga or cyanobacterium (the photobiont). Usually neither can survive on its own. Most of the fungi involved are Ascomycota, though a few are Basidiomycota. There are about 16,000 species of lichens, many of which can grow in very inhospitable environments—on rocks, sides and branches of trees, and gravestones, from the tropics to deserts to the Arctic (Fig. 18). Lichens are

FIGURE 17 Ectomycorrhizae on pine roots.
very sensitive to air pollution, especially sulfur and nitrogen, and so they are natural indicators of air quality. However, even under optimal conditions, lichens grow extremely slowly, usually 1–2 mm per year. The main ecological importance of lichens is their capacity to break down rocks into soil. They are an important food source for caribou and reindeer on the tundra. One of the first indications to the outside world that there had been a nuclear accident at Chernobyl, then in the Soviet Union, was the accumulation of radioactivity first in the lichens in Scandinavia, then in the milk given by the reindeer. There is significant evidence that the manna referred to in the Bible was actually a lichen, *Lecanora esculenta*, a loosely attached lichen that was blown down from the mountains into the lowlands. Some lichens have been used as natural dyes, such as tweed. In less enlightened times, a bright yellow lichen called *Letharia vulpina* was used by “pioneers” as a wolf poison.

**VI. INDUSTRIAL USES FOR FUNGI**

**A. Mushroom Cultivation**

The most widely available mushroom produced in the United States is the white button mushroom, *Agaricus bisporus* (Table VI). Commonly used on pizza and at salad bars, this mushroom is a secondary decomposer that is grown on composted cow or horse manure. Some brown forms of *A. bisporus* are currently being cultivated; brown buttons are sold as crimini and opened mushrooms are sold as portabella mushroom.

There are a number of steps in the commercial production of *Agaricus bisporus:*

- Manure is placed in large concrete “runways.”
- Composting (breakdown of the substrate into simpler components) occurs with the bacteria and fungi naturally present in the manure.
- Large machines turn the compost weekly, otherwise the center of the pile gets too hot because of metabolic heat and kills the composting bacteria and fungi. This also ensures even composting. The different species present naturally shift as composting progresses.
- After a couple weeks, the odorless compost is ready and is placed into large trays, 6 ft. × 6 ft. × 2 ft.
- Mycelium of *Agaricus* is inoculated into the composted substrate and allowed to grow for a few weeks.
- When the substrate is colonized, a sterile layer of
nutrient-poor casing soil is placed over the top of the substrate.

A few weeks later, the mycelia send up rhizomorphs (hyphal aggregations) through the casing layer and form mushrooms. Contrary to popular belief, most mushrooms require light to initiate fruiting body formation. Light is often used as a signal to the mycelium that it is outside the substrate and that fruiting bodies can be formed. Some fungi form fruiting bodies in response to reaching outside air, where the concentration of carbon dioxide is lower. A few fungi fruit only when they run out of available nutrients. _Agaricus bisporus_ has been bred so that it requires no light for fruiting.

There are a number of other specialty mushrooms being grown throughout the world that are just becoming available in North American supermarkets. Almost all the others are primary decomposers of wood or other cellulose-containing substrates. These have gained in popularity in the past 10 to 20 years as consumers discover that these specialty mushrooms have great flavor and other interesting qualities. In Japan, China, and Korea, very little _Agaricus_ is being grown; growers concentrate on mushrooms with more robust flavor and interesting texture, such as _shiitake_ (Fig. 19). Most of these specialty mushrooms are now grown on artificial sawdust logs by the following method:

- Place sawdust or wood chips with supplementary bran and millet and the appropriate amount of water into clear polypropylene bags. Autoclave or sterilize the filled bags, then allow to cool.
- Inoculate spawn (usually grain or sawdust with mycelium of the fungus growing on it) into medium. Mix thoroughly by hand or mechanically.
- Place bags in growth room. Allow spawn to grow rapidly through the substrate. Fungus colonizes bag in 30–60 days; sawdust is easier than solid wood to colonize because of increased surface area-to-volume ratio, abundance of air spaces, and uniform distribution of nutrients.
- During this period the loose medium is joined together into a coherent synthetic log.
- Depending on the species, at this time the plastic is removed, and the synthetic log can be handled like a natural log. Fruiting usually occurs within 90–120 days after inoculation.

### B. Antibiotics and Other Drugs

Penicillin is the first antibiotic that was discovered to fight the bacteria that cause human disease. It is naturally produced by _Penicillium chrysogenum_ and related species as a way of killing bacterial competitors in their environment. Cephalosporins are another class of antibiotics produced by _Acremonium_ and related species. For a fungal infection of the fingernails and toenails, one prescribed drug is griseofulvin, produced by _Penicillium griseofulvin_. Ergotamine, produced by _Claviceps purpurea_, is used to facilitate the delivery of babies and can also be used to relieve migraine headaches. Another chemical found in _Claviceps_ is a precursor to the hallucinogen LSD (lysergic acid diethylamide) that has the same effects as that illegal drug. The steroids in birth control pills are produced industrially by the fungus _Rhizopus nigricans_, as are the steroids cortisone and...
prednisone. People who have had an organ transplant usually take the antirejection drug cyclosporin, which is produced by the fungus Tolypocladium inflatum. The pharmaceutical industry is constantly searching for new antibiotics and drugs to counteract microbes that become resistant to frequently used medications.

C. Wine and Beer Making

Yeast, especially Saccharomyces cerevisiae, have played an important role in the development of human culture. Since brewing is an ancient art going back for millennia, yeasts have been our allies for many years. Virtually anything that contains simple sugars is fermentable by yeasts. In wine making, the grapes are pressed and yeast is added. Fermentation breaks the sugars down into carbon dioxide and ethyl alcohol. Since barley is the main carbohydrate source in beer making, there is an added step before fermentation, that of allowing the barley to germinate to produce sugar from its starch. Various other ingredients, such as hops, oatmeal, and flavorings, are added to make the wide variety of beers that are available. The aging process in alcoholic beverages also adds to their distinct flavors.

D. Bread Baking

Bread making also uses Saccharomyces cerevisiae to ferment sugars into carbon dioxide and ethyl alcohol. However, bakers are only interested in the carbon dioxide, which makes the bread rise. The alcohol evaporates rapidly on baking.

E. Fermentation Products for Food Use

Fungi are often used in the large-scale fermentation of liquid or solid substrates. Fermentation vats can sometimes be several hundred thousand liters in size. For example, citric acid in cola drinks is produced by large-scale vat fermentation of the deuteromyete Aspergillus niger. Yeasts are sometimes grown in large fermenters and used directly as food supplements. Vitamin B6 (riboflavin) in enriched flour is produced by the ascomycete Ashbya gossypii. Cocoa beans are processed by a “fermentation” (sensu food scientists) of Candida krusei and Geotrichum. Authentic soy sauce is fermented in a three-step process with the fungi Aspergillus oryzae and Zygosaccharomyces rouxii, as well as the bacterium Pediococcus halophilus. Tempeh, a soybean product popular in Indonesia, is partially fermented with a species of Rhizopus. Many good cheeses, such as blue cheese, camembert, and brie, are ripened through the action of fungi to obtain their distinctive flavors. Blue cheeses such as Roquefort, Gorgonzola, and Stilton are ripened by Penicillium roquefortii—the blue color is caused by sporulation of the fungus! The white crust on the outside of brie and camembert is the mycelium of Penicillium camembertii.
F. Biopulping and Bioremediation
Several important industrial processes, which have the potential to be great boons to ecosystem health, are in the pilot stage: biopulping and bioremediation. The lignin-degrading enzyme system of *Phanerochaete chrysosporium* is special for these two uses. One of the biggest energy and pollution expenditures in paper making comes from removal of the brown lignin from wood so that the white cellulose is all that is left to make paper. What if paper companies could use the enzymes of a fungus to remove the lignin? This could result in a savings in both energy and time and avoid the polluting wastes that are commonly dumped out of the mills. This process is known as biopulping. There are several products in the pilot stage, but no large-scale biopulping is yet being done.

To understand this system, you must know that wood consists primarily of cellulose, which is white, and lignin, which is brown. *Phanerochaete* species cause a white rot of wood. That is, the fungus decays the lignin and leaves the cellulose behind. There are also fungi that cause a brown rot, digesting the cellulose and leaving the lignin behind. Many kinds of fungi cause a white rot, but *P. chrysosporium* has several features that might make it very useful. First of all, unlike

![Image](image.png)

**FIGURE 20** The author with *Brongeopus nobilissimus* on a large host tree of *Abies procera*. 
some white rotters, it leaves the cellulose of the wood virtually untouched. Second, it has a high optimum temperature (about 40°C), which means it can grow on wood chips in compost piles, which attain a high temperature. These characteristics point to some possible roles for this fungus in biotechnology applications.

Some of the lignin-degrading enzymes of *P. chrysosporium* will also degrade toxic wastes, such as PCBs (polychlorinated biphenyls), PCPs (phenycyclodienes), and TNT (trinitrotoluene). The structure of these chemicals is similar to that of lignin, and the ligninase enzymes will work on them. The fungus performs well on the laboratory bench, but as with many industrial bioprocesses, there are difficulties in scaling up the process. Nonetheless, this procedure has the potential to clean up some industrial and toxic waste sites.

**VII. THREATENED OR ENDANGERED FUNGI**

Like the proverbial canary in the coal mine, fungi may be the first indicators of things going wrong in ecosystems. Remember that fungi are the thread that tie the whole food web together, since they are the primary decomposers and aid most plants as mycorrhizae in the absorption of minerals and water. Thus if something goes wrong with the fungi—if they disappear—there may be dire consequences for any plants and animals that are dependent on them. Lichens have already been proven to be accurate indicators of air quality, in both their quantity and diversity.

Several European countries maintain “red lists” of threatened or endangered fungi. One fungus that may be a candidate for the endangered species list in the United States is *Bridgeoporus nobilissimus*, a polypore fungus with a very large, perennial fruiting body (Fig. 20). For a long time this fungus was in the *Guinness Book of World Records* as the largest known fruiting body of a fungus, at over 160 kg (300 pounds)! There are just six known sites in Washington and Oregon at which *B. nobilissimus* is now known to occur. It is considered by many to be a rare and probably endangered fungus. The main reason for this designation is that it is restricted to very large specimens of noble fir (*Abies procera*) and occasionally Pacific silver fir (*Abies amabilis*) with a diameter at breast height (dbh) of 1–2 meters. Trees of this diameter are not very common. This fungus is considered endangered because its habitat is endangered. Unfortunately, it is not clear whether the U.S. Endangered Species Act applies to fungi. In this zoo-centric world, most people are more interested in the “charismatic megafauna” than in some lowly fungus. However, the fungi have many important roles to play in the ecosystem and should not be ignored. No ecosystem could exist for long without fungi!

Two states have thus far recognized the importance of fungi by naming a state mushroom. Minnesota was the first, naming the morel (*Morchella species*) as their state mushroom. Some ten years later, Oregon declared the Pacific golden chanterelle (*Cantharellus formosus*) as their state mushroom. Surprisingly, some forests yield more income annually from wild mushroom harvesting than from lumber harvesting. Moreover, intact forests continue to produce edible mushrooms yearly, whereas lumber harvesting can occur only once every 50 years or so. Mushrooms for human food is only one of the many contributions that fungi make to our lives, for their fundamental decomposer role supports almost all ecosystems and so helps to provide the essential ecological services that we take for granted, as well as the recreational opportunities of enjoying nature. Fungi have contributed a great deal to our standard of living by making most of the living world possible.

See Also the Following Articles

EUKARYOTES, ORIGIN OF • PARASITISM

**Bibliography**


Volk, T. J. *Tom Volk’s Fungi Webpage*: www.wisc.edu/botany/fungi/volkmyco.html

Volk, T. J. Mycological Society of America web page. www.erin.utoronto.ca/~w3msa


GENE BANKS
Simon H. Linington and Hugh W. Pritchard
Royal Botanic Gardens

I. Role within Conservation Strategies
II. Seed Banks
III. Other Types of Gene Bank

GLOSSARY
accession Usually a sample (e.g., seed lot) but may be a set of genetically related samples.
collection One or more accessions.
genetic pool The genetic diversity contained within a population, species, or crop. The primary gene pool of a crop represents the biological species, the secondary gene pool includes species that can be crossed with it, allowing at least some transfer of genes, and the tertiary gene pool includes related wild species where such gene transfer involves specialized techniques.
genotype The genetic makeup of an individual.
seed lot A sample of seeds with a common harvest and post-harvest history.
wild or nondomesticated species Those species that have not been brought into regular cultivation. Many may, however, have well-known uses.

A GENE BANK is an ex situ (or off site) collection of genetic material held for long-term security or for ease of access. The genetic material is usually in the form of live propagules or gametes though, increasingly, pure DNA is held. A bank can also be a collection of full-grown plants representing the diversity of certain species (a field gene bank) or a set of cultures (cell, tissue, embryo, or microorganism). The widest definition might also embrace collections of plants grown primarily for purposes of education and horticulture in botanic gardens and animals represented in zoological gardens or parks.

I. ROLE WITHIN CONSERVATION STRATEGIES
Ex situ (or off site) conservation underwrites conservation measures in situ (on site or in the natural state) and provides centers from which genetic material (germplasm) can be accessed for research, breeding, or (re-) introduction. Nearly any ex situ collection might be included under the loose definition of a "gene bank." As with a financial bank, deposits are made with the intention of later withdrawal. Furthermore, while in the gene bank, safety and integrity will be paramount and value will be added through study. However, the emphasis placed on withdrawal varies considerably between gene banks. For some collections, short-term utilization is the sole motivation, though inevitably conservation will be served to some extent. In others, long-term insurance against loss in situ is the stronger driving force. Given the increasing risks of genetic erosion (i.e., a narrowing of the gene pool leading to species being...
II. SEED BANKS

A. Seed Banks Are an Effective Form of ex Situ Conservation

What makes seed banks such an effective ex situ conservation technique is that the methodology can be applied to a wide range of species in a universal and straightforward way and that large amounts of intraspecific diversity can be conserved for long periods of time without intervention.

The fecundity of most plants means that a small seed sample can be removed from the annual seed rain with limited effect either on that population’s survival or on the seed yield. Once carefully sampled, can hold the majority of alleles found within that population. Moreover, if carefully germinated and then grown up under the same conditions, a seed lot has the potential to recreate the original population. In the meantime it can be held within a relatively small volume. For instance, a million tobacco seeds, each one with the potential to grow into a genetically distinct plant, occupy about a fifth of a liter.

Additionally, germinating seeds to obtain fully grown plants is relatively simple compared with obtaining plants from in vitro stored material. Plants recovered from banked seeds can also be compared with natural populations from which the material was harvested years before and which may have been subjected subsequently to environmental change (e.g., as a result of global warming).

B. The Historical Context

Seed banks might be seen as an invention of the 20th century, though the exploitation of crop seeds’ ability to store in a dry state has almost certainly played a significant role in early human civilization. Early forms of seed storage probably included burying the seeds in ash or sealing them in adobe huts for the following season. By comparison, modern seed banks use controlled drying facilities and deep freeze stores and aim for storage lives of decades, centuries, or more. Modern seed banks have a lineage built on plant exploration and the development of botanical gardens that dates back nearly 5 millennia.

A collection of medicinal plants was established in 1749 in the garden of a Dominican monastery near the coastal town of Assisi in northern Italy. This represents the first botanical gardens specimen collections.

A second botanical gardens specimen collection was established in 1752 in the garden of the Botanisches Institut der Universität in Jena, Germany. These collections are currently part of the Jena Botanical Garden, which is one of the oldest botanical gardens in the world. The history of these collections is closely tied to the development of botanical gardens as well as the historical context of seed banking.

While in situ conservation should be seen as the primary goal of all conservationists, having a second line of defense is a prudent measure particularly if it assists in other ways. For example, by acting as a supply source, ex situ collections reduce the pressure of repeated sampling of wild populations. Importantly, through the use of such collections, access to national genetic resources can also be controlled. Furthermore, because research is facilitated by such collections, knowledge about both the conservation biology and socioeconomic value of wild populations is improved, thereby increasing the likelihood of survival in situ. It is sometimes asserted that ex situ conservation of certain species reduces political will to conserve the habitats in which they occur. Should this be true, it would seem essential to ensure against the loss of, at the very least, flagship species within the habitats so that they can be reintroduced. Additionally, it should be apparent that habitat complexity may limit reconstruction opportunities from ex situ collections. In summary and as often stated, ex situ conservation is complementary to that carried out in situ (see also the entry for “Ex situ, in situ conservation”).

The main types of animal ex situ collections are zoo logical gardens and parks, sperm and ova banks, and DNA banks. Due to controlled breeding programs, genetic resource databanks also play a more prominent role in the management of such collections than is perhaps the case with those for plants and particularly those in botanical garden collections. Microbe collections are mainly cultures though the possibility exists for fungal spore banks.

When considering plant species, the main types of ex situ collections are botanic gardens specimen collections (each normally comprising many species and few individuals per species); field gene banks (usually few species, many individuals); seed, pollen, and spore gene banks (potentially many species, many individuals); in vitro cultures (relatively few species, many individuals); and DNA banks (potentially many species, many individuals—though the latter are rarely held separately). The collections vary in the degree of technical input required, their ability to effectively conserve inter- and intraspecific variation, their longevity, and the ease with which gene products can be obtained. Seed gene banks (in this entry abbreviated to ‘seed banks’) nicely balance these factors and offer a very effective form of ex situ conservation for many plant species. For this reason and the fact that most references to gene banks relate to seed banks, this chapter will firstly consider this type of gene banking and in particular the technology.
China in 2800 B.C. Three hundred years later, the Sue- 

Europe than it is now. However, the rate of erosion has 

carrot color was infinitely more diverse in Medieval 

genetic base of crops is not new. For instance cultivated 

number had reduced to a thousand. Erosion of the 

regions (FAO, 1996). In 1949 there were about 10 

corning relates to diversity found in the arid and semiarid 

of genetic erosion are considered to be replacement of 

local varieties, land clearance (e.g., fuel-wood collec-

and wild species. Currently, the most important causes 

of genetic erosion are considered to be replacement of 

of local varieties, land clearance (e.g., fuel-wood collec-

and overexploitation; some of the greatest con-

ern relates to diversity found in the arid and semiarid 

(McIntyre, 1987). Not surprisingly, many botanic 

gardens now hold important collections of both en-

and exotic nondomesticated species including 

those known to be used locally for medicines and timber 

plus many ornamentals. 

changed agricultural practice and increased pressure 

on the natural world during the 20th century has led 

to increased erosion of genetic diversity both in crops 

and wild species. Currently, the most important causes 

of genetic erosion are considered to be replacement of 

of local varieties, land clearance (e.g., fuel-wood collec-

and overexploitation; some of the greatest con-

ern relates to diversity found in the arid and semiarid 

regions (FAO, 1996). In 1949 there were about 10 

thousand wheat varieties in China; in the 1970s the 

number had reduced to a thousand. Erosion of the 

genetic base of crops is not new. For instance cultivated 

carrot color was infinitely more diverse in Medieval 

Europe than it is now. However, the rate of erosion has 

accelerated in the 20th century with the drive toward 

greater genetic uniformity of crop varieties. Some highly 

uniform varieties have been sown across huge areas. In 

1983 a single variety was sown across two-thirds of the 

wheat fields in Bangladesh and the following year across 

nearly a third of all those in India. This uniformity has 

productivity advantages within agricultural systems but 

is not without risk. For instance, there was a severe 

shortfall in winter wheat production in the Soviet Union 

when a variety grown over 15 million ha did not have 

sufficient cold tolerance to survive the winter of 1972. 

Less dramatic was the abandonment in 1975 of all of 

the United Kingdom’s white clover varieties due to sus-

cceptibility to the disease Sclerotinia trifoliorum (FAO, 

1996). 

The more modern uniform varieties have been se-

lected from genetically heterogeneous old varieties and 

landraces (primitive varieties) that they have then dis-

placed. Therein lies the conundrum. For these new 

varieties to be developed further they need to draw on 

the genetic diversity that they have displaced. Half of 

the production increases of the Green Revolution were 

brought about by the use of plant genetic resources for 

plant breeding. 

C. The Current Status of Seed Banks for 

Food and Agriculture 

Although the outstanding Russian genetic resource sci-

entist Nikolai Vavilov had assembled a collection of 

50,000 cereal and pulses during the 1920s and 1930s, 

which were maintained by an annual grow-out, the first 

impetus to develop modern seed banks occurred in the 

United States. Here it was noted in the late 1940s that 

less than 10% of 160,000 plant accessions in the na-

tional system since 1898 could still be found (Plucknett 

et al., 1987). Following the establishment of U.S. re-

gional genetic resource facilities to address this prob-

lem, the National Seed Storage Laboratory (NSSL) was 

established at Fort Collins, Colorado, in 1958. With 

about 360,000 accessions, it and the Institute of Crop 

Germplasm in Beijing with about 316,000 accessions 

are the world’s largest seed banks 

A significant proportion of the accessions held in 

crop genetic resource collections (including seed 

banks) date from the 1960s onward when the extent 

of loss of diverse landraces was becoming apparent. 

Catalyzed by the Food and Agriculture Organization of 

the United Nations (FAO) and subsequently by the 

International Board for Plant Genetic Resources 

(IBPGR, now the International Plant Genetic Resources 

Institute, IPGRI), a great amount of crop genetic diver-
The Centro Internacional de Agricultura Tropical (CIAT) in Colombia (70,940)

The International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in India (110,478)

The International Centre for Agricultural Research in the Dry Areas (ICARDA) in Syria (109,029)

The International Rice Research Institute (IRRI) in the Philippines (80,646)

The Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) in Mexico (136,637)

The International Rice Research Institute (IRRI) in the Philippines (80,646)

The Centro Internacional de Agricultura Tropical (CIAT) in Colombia (70,940)

Just under half of the 6 million accessions are held in 12 national collections. To some degree this is a function of the early establishment of their genetic resource collections. The collections include those in Russia and the United States (noted earlier), Japan (established in 1966), Germany (1970), Canada (1970), and Brazil (1974). Of the 1308 national or regional collections currently noted by FAO (1996), only 397 within 75 countries are held in long-term seed banks. Medium-term storage might be assumed to be in the order of 10 or more years.

Whilst 496 collections are within Europe and 293 are within Asia, some areas of the world have very few such facilities. FAO (1996) noted that there is only one long-term store within the Caribbean and this is in Cuba. Relatively few of the collections covered by the FAO survey are held by private companies, though many breeders hold smaller working collections.

The remits for crop banks are very varied. Some cover wide ranges of the crops that are important at the national or regional level, others cover very specific material. For instance, facilities at Horticulture Research International in Wellesbourne (United Kingdom) and the Asian Vegetable Research and Development Centre (Taiwan) have remits for vegetable germplasm.

While the science and technology underpinning these seed banks is generally well founded (discussed later), problems such as unreliable electricity supply mean that only the nine CGIAR banks, those in 35 countries and four regional centers, meet international standards as established by FAO and IPGRI (1994). To some extent this problem is obviated by the high level of duplication within the world’s 6 million accessions. It is estimated that perhaps 1 to 2 million of these are genetically unique.

The way in which seed bank and associated genetic resource work is organized both at a regional and national level varies considerably. In Europe, the European Cooperative Programme for Crop Genetic Resources (ECP/GR) works through crop networks to coordinate effort. In Southern Africa, work is coordinated through a network set up through the Southern Africa Development Corporation (SADC). At a national level, in India, genetic resources activities are formally centered around the National Bureau for Plant Genetic Resources. Within the United Kingdom, genetic resources activities are less centralized with responsibility residing with a number of institutes and their respective ministries. Advice to government is provided through the U.K. Plant Genetic Resources Group, which draws members from all of the relevant organizations.
D. Seed Banks for Wild Species

Increasingly, botanic gardens are adopting seed banking technology. Although FAO (1996) note that 150 botanic gardens (about 10% of the world total) have seed banks, relatively few of these have embraced international standards (discussed later). Those that have include the ones at the Berry Botanic Garden, Oregon (United States), the National Botanic Garden of Belgium, the Royal Botanic Gardens, Kew (Wakehurst Place), and the network of seed banks within the Spanish botanic gardens (notably Cordoba). Most botanic garden seed banks were born out of a need to more effectively manage the annual seed exchange. However, many now concentrate on the conservation of undomesticated species and interspecific variation. Additionally, several banks not in botanic gardens concentrate on wild species. These include the Threatened Flora Seed Centre established by the Department of Conservation and Land Management in Western Australia, the forest tree seed banks (e.g., the National Tree Seed Programme in Tanzania), and the seed bank of the Universidad Politecnica de Madrid. The latter bank was established in 1966 to conserve Crucifers and species from around the Mediterranean. Together with the Royal Botanic Gardens, Kew Seed Bank that has concentrated on wild species from the world’s drylands, it has helped pioneer the application of seed bank technology for the conservation of wild species. Faced with the acceleration in wild plant genetic erosion and extinction, the Kew Seed Bank has taken on the ambitious objective of facilitating the conservation of seed from 10% of the higher plant species by the year 2010. Termed the Millennium Seed Bank Project, the work is a collaborative exercise involving botanical organizations from around the world. One of the aims of the Project is to maximize the efficiency of the banking work through an active seed research and technology transfer program. Other than the NSSL at Fort Collins, the project will have one of the largest seed research groups associated directly with a seed bank. Certain other botanic gardens, such as that in Rio de Janeiro, also carry out seed research work applicable to seed banking.

E. Community Seed Banks

A valuable role in seed banking is played by nongovernmental organizations (NGOs) at the community level. There is an increasing use of community seed banks within the developing world that involve donation, storage (usually on a short-term basis), and use of seeds at the local level. Examples include a set of community seed banks set up in North Eastern Brazil. The Plant Genetic Research Centre in Ethiopia is an example of a national program integrating with a community-based system (Heywood, 1995). Seed Savers in the United States and the Henry Doubleday Research Association in the United Kingdom are examples of NGOs active at the national level in the conservation of traditional vegetable varieties through their heritage seed libraries.

F. Collection Value

A financial value can be placed on biodiversity in one of two ways (see FAO, 1996):

- **Direct value**, or a willingness of the market to pay for it
- **Indirect value**, including its value as an insurance policy, the value to a product derived from the input of specific genetic material or the travel costs people would accept to collect it or to visit it for tourism

Obviously, there are other ways of considering the value of biodiversity such as its “service” value to biosystems and its value as a heritage to each human generation. However, debate has not surprisingly concentrated on its economic value and consequently the issue of access to the genetic wealth contained within seed banks is an increasingly controversial subject. Behind this debate is an attempt to reconcile the financial contribution of samples collected in one country to the economy of other countries. Fundamental concerns that is often expressed is the degree of control over agricultural systems, particularly within the developing world, by multinational agrochemical/seed companies. Concern revolves around the use of the following:

- **Uniform F1 hybrid varieties** (where the seed, if harvested from the crop by the farmer, subsequently produces a heterogeneous crop of lower yield)
- **Varieties that work best within systems when treated with certain agrochemicals**
- **The use of genetically modified organisms (GMOs)**
- **Plant breeders' rights and, increasingly, patents**

Set against a backdrop of a $15 billion estimated annual turnover for the seed industry in Organization for Economic Cooperation and Development (OECD) countries, the debate has polarized on an “Industrial
North–Developing South’ divide. This has led to a useful reassessment of the sharing of benefits, though inevitably there has been some over expectation of the value of individual germplasm collections. Undoubtedly, there are examples of large financial contributions to the commercial world from plant germplasm. The value of genes for high sugar content from *Lycopersicon chilense* have been estimated to contribute $3 to 8 million annually to the tomato industry (see FAO, 1996). However, it is worth bearing in mind that these are the exceptions to the rule. Relatively few advances in plant breeding have been spectacular and few have resulted from the incorporation of single genes. Most have occurred from the gradual improvement of characteristics governed by many genes (polygenic characters). Of course, advanced genetic manipulation techniques may change this picture and, indeed, may have an impact on the perceived value of traditional crop genetic resources.

In contrast to the seed industry, the pharmaceutical sector is estimated to be valued at $235 billion (FAO, 1996). Included within this turnover are plant-derived drugs, many of which relate back to plants collected from developing countries. However, large financial returns from plant-derived drugs are unusual because major breakthroughs in the identification of new bioactive compounds from plants are infrequent. It is true that 34 out of the top 150 prescription drugs in the United States have a plant-based origin. This aside, despite the estimated tens of thousands of species investigated in bioprospecting programs, only 90 plant species have yielded 119 compounds considered to be important drugs (see ten Kate and Laird, 1999).

The debate has driven the establishment of agreements covering the access to and the sharing of benefits derived from plant genetic resources. Conse- quently, seed bank curation needs to accommodate not only national and international legislation but compliance with increasingly strict conditions attached to the material by the country of origin of the material. This aside, despite the estimated tens of thousands of species investigated in bioprospecting programs, only 90 plant species have yielded 119 compounds considered to be important drugs (see ten Kate and Laird, 1999). The Convention on Biological Diversity (CBD) came into force on December 29, 1993, following its opening for signature at the Rio Earth Summit in 1992. It has provided party governments with a framework within which access to genetic resources and the sharing of benefits arising from their utilization can be negotiated. Many countries have either developed, or are preparing national legislation to implement, access and benefit-sharing provisions of the CBD. Fundamental concepts that are often incorporated into national access legislation are as follows:

- The need to obtain prior informed consent for acquisition of genetic resources from the appropriate national authority.
- The need to reach mutually agreed terms, which implies a negotiation leading to a form of bilateral material transfer agreement specifying the terms and conditions under which material is transferred.

In addition to the provisions of the CBD and implementing legislation at the national or regional level, certain seed bank collections may fall within the ambit of the International Undertaking (IU) on Plant Genetic Resources for Food and Agriculture (PGRFA). This is currently being revised in order to establish a multilateral access and benefit-sharing system for PGRFA that is compatible with regimes established under the CBD. It also recognizes the contribution made by indigenous farmers, in particular, to the development of plant varieties.

Seed banks, particularly the larger ones such as those within the CGIAR system (discussed previously), have responded to the emerging legal framework on access to genetic resources and benefit-sharing by reviewing the terms under which they acquire genetic resources and supply them to third parties. In many cases, seed banks have adopted the use of standard material transfer agreements that entitle the providing organization (or the country of origin of the material) to a share of any benefits arising from the utilization of the material. They may also restrict the subsequent transfer and use of the material. A pioneering worldwide project by 16 botanic gardens, which like the agricultural seed banks have traditionally supplied genetic resources to a wide range of users, has also attempted to find broadly standard terms under which material may be transferred to third parties.

Particular challenges faced by seed banks include the following:

- The acquisition of new material on terms that enable its subsequent distribution and use, so far as possible, on a common basis.
- The development of a supply policy flexible enough to ensure that terms and conditions of any bilateral agreement (e.g., with the country of origin) are met, while enabling plant genetic resources to be also made available for food and agriculture under any multilateral system to be adopted in the revised IU.
- The development of procedures that enable curators to clearly identify whether a proposal to acquire or supply material should be dealt with on a
multilateral or bilateral basis. For instance, due to the single source nature of many successful pharmaceutical "hits," such work may fall into the latter category.

Finally, it is important to note that the benefit of the annual investment in PGFRA conservation and utilization compared with its estimated total annual cost of $1 billion (FAO, 1996) is considered to be very high. For this reason it is essential that the future operations of seed banks are not threatened by financial insecurity.

G. The Scientific Principles Underlying Seed Banking

Viable seeds of many species when maintained in a dry and cold state are capable of being germinated many years later. This capability means that the long-term ex situ conservation of many higher plants is a realistic possibility.

1. Seed Storage Conditions

The science of seed storage is not a new one and dates back, at least, to China in the 6th century. Advances in the quantification of seed longevity under different storage conditions were made in the second half of the 20th century through the work of Harrington in the United States and Ellis and Roberts in the United Kingdom (see review in Hong, Jenkins, et al., 1998). Critical factors that determine seed longevity are the seed's moisture content, temperature, and gaseous environment; its initial viability; and its genetic background.

With respect to the latter, differences between species would appear to be much greater than those within. Genetics particularly influences the relationship between seed longevity and seed moisture content. Most species produce seeds that can be dried to low moisture contents (e.g., where less than 5% of the seeds' fresh weight is water) without loss of viability. The seeds of such species are termed "orthodox." Furthermore, orthodox seed longevity increases in a predictable fashion as the moisture content is reduced. Within limits, there is a straight-line relationship of negative slope between longevity and seed moisture content (up to about 13 to 25%) when both are expressed on a logarithmic scale. More simply, Harrington's rule of thumb states that a 1% reduction in moisture content roughly doubles seed longevity. Ellis et al. in 1989 found that drying seeds below a certain moisture content (equating to that in equilibrium with air at about 11% relative humidity at room temperature) did not increase longevity in most species studied. However, in a few species, a decrease in longevity was noted.

Temperature also has a predictable effect on seed longevity over the range −13°C to 90°C (Dickie et al., 1990) such that there is a quadratic relationship between longevity on a logarithmic scale and temperature—although an Arrhenius relationship also fitted the data well. Importantly, the relative effect of temperature on the seed longevity of eight species was identical. In its simplest form, Harrington's rule of thumb states that a 5°C reduction in temperature doubles seed longevity.

In 1980, Ellis and Roberts developed a predictive model for orthodox seed storage that incorporated the determinants of longevity (except gaseous environment that has minimal effect at the low moisture contents over which the model is applicable). Based on acceleration of the aging process in seed lots by increasing their moisture content and temperature, the model allows extrapolation of longevity at cooler and drier conditions. For instance, rice seeds at 5% moisture content and ~20°C (fairly typical seed bank conditions) have a theoretical potential to survive at least 1000 years (see Hong et al., 1998). Such data lend weight to occasional press reports of the germination of seeds of great antiquity, many of which cannot be confirmed due to the lack of supporting archaeological evidence. Nonetheless, there is, for example, evidence of longevity greater than 100 years at room temperature for dry wheat seeds (Steiner and Buckenbauer, 1995). The seed storage model indicates the independence of temperature and moisture content on longevity over a wide range of conditions (Ellis et al., 1995). It has been suggested that this may not necessarily be the case, particularly at very low moisture contents (Vermint and Rees, 1990). The model also appears to be influenced by developmental stage (Hay et al., 1997). These deliberations do not, however, alter the current seed bank storage recommendations (details of which are given later).

Even if these longevity predictions prove to be unattainable for all species, they do imply that, for some at least, storage in seed banks may be sufficiently long to carry the current heritage of genetic material to new eras when other technologies will be available. Predicted longevities of this magnitude also imply that the amount of regeneration of seed stocks inherent in many gene bank programs currently may be more a function of preferred working practice (e.g., regular study of collections), or poor application of the technology, rather than a failure of the technology itself.

In a study of the seed storage behavior of 6700 species, Hong et al. (1998), found that 91% had orthodox
(desiccation tolerant) seed. The remainder had either
desiccation susceptible or “recalcitrant” seed or had
characteristics intermediate between the recalcitrant
and orthodox types. Also certain genera (e.g., Acer)
contain both orthodox and recalcitrant species. There is
little evidence, however, that individual species produce
both orthodox and recalcitrant mature seeds. It should,
though, be noted that orthodox seed passes through a
desiccation susceptible phase during development. The
true percentage of species with orthodox seed within
the plant kingdom is probably significantly less than
91% as the sample was biased toward species in trade.
Although recalcitrant seed behavior is loosely associ-
ated with long-lived perennial species producing large
fleshy fruits, there is great uncertainty of the number
of wet tropical forest species producing such seeds.
Seed storage behavior is usually assessed by quantifying
the survival of seed lots of a species after drying to
different moisture contents. A more rapid biochemical
diagnosis of potential survival of drying is clearly of
interest. Most recalcitrant seeds, including oaks and
rubber, die below a value of about 40% moisture con-
tent. Because they need to be kept moist (and aerated)
they have short life spans, generally of a few months,
and are not suitable candidates for seed banking. Other
forms of genetic conservation must be employed. The
same is true for species, such as coffee, which produce
“intermediate” seeds that cannot be dried much below
about 10% moisture content. They also tend to be sus-
ceptible to storage at cold temperatures (−20 and 0°C).
Work on other species, such as orchids, suggests the
use of other subzero temperatures may be appropriate
for the conservation of species sometimes included in
this storage category (Pritchard et al., 1999).

2. Genetic Considerations
The genetics of seed bank storage is an important issue.
A criticism occasionally leveled at seed banks is that
there is selection in storage. Selection can occur during
the seed lot is in storage. It might be argued that those species with short genera-
tion times will quickly adapt to all but the most ex-
tremely changed environments (i.e., the gene frequen-
cies in the seed lot will quickly be selected to meet the
new needs). Of course, maintaining the original genes
sampled from within the population may have longer-
term benefits of fitness for the species if pathogen attack
has altered the natural population in the intervening
period. An additional consideration is the prospect of
genetic contamination of natural populations by chance
hybridization with exotic material including genetically
modified organisms. This concern is also partially allevi-
ated by the existence of long-term stored germplasm
in its original state.

H. Seed Bank Management

1. Procedure
The basic elements of the seed banking procedure (more
or less in order) are as follows:
• Collection planning and permission seeking
• Seed (and pressed specimen) collecting and field
data recording
• Shipment of the seeds
• Creation of a data record about the accession
• Seed cleaning (sometimes preceded by initial dry-
ing and sometimes accompanied by X-ray analysis
and quantity determination)
• Main drying
• Seed moisture determination
• Initial germination test (sometimes left until after
banking)
• Packaging and banking and security duplication
• Characterization (including verification of identity
in the case of wild species) and evaluation (where
appropriate)
• Distribution of stocks to users (through time)
• Germination retests (through time)
• Regeneration/multiplication (as required)

Although many of the well-established seed banks
broadly follow these procedures, most modify them to
meet their own specific needs. Some banks carry out
very little field collecting and some multiply the seed
sample on arrival. Costs for each stage vary considerably
between seed banks. The RBG Kew Seed Bank, which
has an international collecting program, estimated, in 1997, that the ratio of costs between collecting, processing, and annual maintenance for a collection was in the order of 100:50:1.

2. Seed Collecting
Set against the background of the CBD and the IU, collecting should only be carried out with the permission of the national and local authorities. Foreign collectors should work collaboratively with local scientists and clear agreements on benefit sharing should be in place. One immediately tangible benefit is the sharing of collections and the information relating to them. Other international regulations need to be adhered to. These include the Convention on International Trade in Endangered Species (CITES) and national quarantine laws.

Seed collecting methodology and genetic resource exploration have been thoroughly covered by Guarino et al. (1995). In most instances, random and even sampling of wild plant or crop populations is recommended, including careful note taking of the sample method (in often less than perfect conditions). Objective data recording is essential as is accurate recording of location. This latter aspect is now facilitated by the use of Global Positioning System receivers that help fix latitude, longitude, and even altitude using satellites.

Harvesting seed that is close to maturity and keeping the seed alive in the field should be paramount. Also it is essential that harvested seed is returned to the seed bank facility as securely and rapidly as possible. Prestorage conditions influence seed viability at the start of storage. Because the longevity of a species stored under specific seed bank conditions is substantially fixed, the only other key longevity factor that can be varied is initial seed storage viability. Delay in drying the seed properly or unduly harsh cleaning methods can severely reduce initial viability; in turn, this can dramatically reduce both seed longevity and seed bank efficiency. Where space is not limiting within a cold store and the longer seed lots can be kept alive, the less the unit fuel cost per species.

It is worth noting that seed lots of the same species should not be mixed either within or between years. Such action could diminish both the unique genetic makeup and reduce the longevity (through lower initial seed quality) of a seed lot.

3. Seed Cleaning
Seed lots harvested from some populations of wild species in particular contain large numbers of "empty" (or aborted) and insect-infested seeds. They may be outwardly similar to the "competent" seed and often are thus not removed during the cleaning process. Their presence can be detected by X-ray analysis or a simple cut-test applied to a subsample. Once detected, such incompetent seeds need to be allowed for when distributing seed to users and when carrying out germination tests. Although the insects may not be removed at this stage, nearly all adult and most larval forms of insects are killed by standard seed bank practice. However, insect eggs along with fungal and viral contaminants can survive seed banking. Thus, appropriate quarantine procedures need to be in place when the seed is removed from store.

4. Seed Bank Storage Standards
No two banks are the same. Traditionally, seed banks have been classified into the following categories:

- **Base collections** that are for the long-term storage of seed lots and from which seeds are not normally sent to users (though this is not always the case).
- **Active collections** from which seeds are made available to users. Often such stores are maintained under less optimal storage conditions compared to those holding base collections though this need not be the case. FAO/IPGRI (1994) recommends that storage lives of 10 to 20 years might be appropriate.

Base and active collections can occur at the same or different genetic resources centers or, as is the case in the SADC genetic resources network, the active facilities in different countries are linked to a regional base facility. In theory, the collections held within the active store need occasionally to be refreshed by the same samples held in the base store. Because active stores are likely to more frequently regenerate (discussed later) their samples, selection and genetic drift could cause a genetic divergence from the same samples held in the base store.

The FAO/IPGRI standards (1994) for base collection storage requires fresh seed to be dried within the range of 3 to 7% moisture content, packaged in a moisture-proof way, and placed at subzero (and preferably −18°C) temperatures. Although many of the world's seed banks aspire to such conditions, as has been noted, relatively few are able to emulate them. Nonetheless, the robust nature of seed bank technology means that substantial longevity is achievable even with partial fulfillment of the standards that have been discussed and using a low technology approach.

The standards set out in 1994 are important as they
set the benchmark for standard operating procedures by seed banks. Consolidation of these standards and increased use of a common format for data would facilitate exchange of information between facilities. Such unification of data standards could ultimately embrace all specimen banks including those specializing in the storage of samples for environmental monitoring.

5. Seed Drying

Drying is perhaps the key to success in seed banking of orthodox-seeded species. Not only is storage prolonged in orthodox-seeded species but germination is prevented, the risks of insect and mite attack are reduced, and the seeds are protected from freezing damage. FAO (1996) found good drying facilities to be a limiting factor in a number of seed banks. Drying involves the manipulation of the water potential gradient between the inside and the outside of the seed. To all intents, the water potential is the difference between the activity of water molecules in any system compared with pure water. When the water potential of the air is lower (more negative) than that of the seed, there is a net movement of water out of the seed and drying occurs. As the water potential gradient reduces, drying slows down. When no gradient exists, the seed moisture content is said to be in equilibrium with the surrounding air conditions. Consequently, factors that influence the water potential of the air surrounding the seed are important to drying. Lowering the relative humidity or increasing the temperature of the air lowers its water potential. By keeping air moving over the seed, low humidity can be maintained by preventing moist air from accumulating around the surface of the seed. Seed size and seed depth affect rate of drying. Migration of moisture from the center of a large seed to the outside will take longer than from the center of a small seed. The same principles apply to large and small seed sacks. Shape, affecting surface to volume ratios, and seed structure will also have an effect on drying rate as can seed maturity.

It is worth noting that at any particular water potential, the seed moisture content will depend on its chemical composition. Seeds with a higher oil content will have a lower moisture content. Most seed banks are located where the ambient relative humidity is not sufficiently low to allow the seeds to dry to the moisture content levels set out in the previous standards. High temperature gives rapid drying. However, although shade and oven drying can be used for drying, there are seed aging dangers of placing wet seeds at high temperatures and of leaving dried seeds too long under such regimes. Similarly, sun drying has risks associated with radiant heat gain. A degree of caution needs to be exercised if these methods are used. Consequently, seeds are usually placed in a drying environment where the relative humidity has been artificially lowered. Most often, this is achieved by sorption or occasionally by refrigeration drying systems. At their simplest, sorption systems can consist of a closed container into which the wet seeds are placed and the air dried by a quantity of silica gel or dried rice. Such methods require a degree of experimentation to achieve the desired results. Greater control over the extent of drying can be achieved by use of a suitable saturated salt solution (e.g., that of lithium chloride) that will maintain a set relative humidity (about 11% at room temperature) within a closed container. However, many of the larger seed banks have now adopted controlled drying rooms that allow for large samples to be dried in thin layers. Air from the chamber is dried using dryers containing lithium chloride or silica gel, cooled, and then returned. In effect, this is a closed system through which most facilities allow for some fresh air intake. Ducting the air to and from the dryer on opposite sides of the chamber encourages air movement. Conditions within the chamber are those necessary to achieve the desired moisture contents at equilibrium for long-term storage. Such conditions may be 10 to 15% relative humidity at 10 to 25°C, and equilibrium is usually achieved in about one month.

6. Packaging

Maximizing orthodox seed longevity depends on drying and freezing. Cooling ambient air increases its relative humidity and hence its water potential. Placing the dried seeds directly inside a refrigerator will cause the seeds to absorb moisture. Therefore seed banks need to carefully package the dried samples. In 1996, IPGRI reviewed the types of container used in conventional long-term seed stores. Types of container used by banks include laminated aluminum foil bags, sealed steel cans, sealed glass tubes, screw-top glass bottles, and lever-top fruit preserving jars. All containers have some limitations but, carefully managed, risks of moisture ingress can be minimized especially if some monitoring system is in place. The type of container chosen will partly depend on the frequency of access required. The main concern with all is that the seal though ingress of moisture through the fabric of the material over long periods is a risk where foil is punctured or cans have poor seams. Glass offers the advantage that the contents are visible but is of course at risk from breakage. A number of banks double-pack for added security and others add a desiccant. A number of facilities dry the air in the
cold store and store the collections in paper or cloth bags. Consequently, these facilities must have adequate generator backup. This is less of a problem where seeds are held in moisture-proof containers and there is a loss of electrical power. In such circumstances there is little evidence to suggest that more than a few days storage life is lost per disruption.

7. Storage Temperature

Many long-term seed banks store seed under deep freeze conditions using either purpose-built prefabricated cold rooms or domestic deep freezers. To reduce staff time at subzero temperatures, a few seed bank cold rooms, such as one at the National Institute of Agrobiological Resources (NIAF), Japan, have mechanized banking/retrieval systems. The use of such systems has implications to energy consumption by the bank.

Use of permafrost has been considered for long-term duplicate storage of seed in places such as Svalbard. Although the dependence on electricity is cut, such stores are usually unable to match the lowering of temperatures possible in conventional base storage conditions. In 1997 the Japanese-based Biological and Environmental Specimen Time (BEST) Capsule 2001 Project discussed the possibilities of long-term storage of flagship samples under Antarctic ice at −38°C (which incidentally is not sufficiently low for animal tissue preservation) or even on the dark side of the moon at −230°C.

More usually, seed storage at ultra-low temperatures (cryopreservation) is achieved using liquid nitrogen. Seed samples are normally held in polypropylene (or similar) screw-cap containers placed in the vapor phase above liquid nitrogen (about −160°C). Preferably samples of dry seed should be cooled and rewarmed at a relatively slow rate (about 10°C min⁻¹) to reduce problems of rapid thermal contraction and expansion that can cause physical injury to the seeds, such as cracking of the embryo tissue. The largest cryogenic seed bank is operated by the National Seed Storage Laboratory (United States) where there are more than 37,000 accessions stored over liquid nitrogen. More than 11,000 of these are also stored under conventional seed bank conditions at −18°C. The National Bureau for Plant Genetic Resources in New Delhi also conserves seeds under conventional and cryogenic storage conditions. Moreover, seeds of a significant number of North American and Australian wild species have been tested successfully for their tolerance of cryopreservation (Pence, 1991; Touchell and Dixon, 1993). It is worth bearing in mind when considering setting up a cryogenic facility for seed material that such low temperatures are not necessary to achieve practical periods of long-term storage (Pritchard, 1995), that the setup costs are relatively expensive, and the storage volume is less efficient than a conventional seed bank. However, these additional costs may be acceptable when creating the ultimate “base collection” for material that is in short supply, is inherently short-lived, or is of particular commercial value (unique genotypes).

8. Monitoring Seed Lot Viability

Perhaps one of the most important parameters of seed bank effectiveness is the result of germination monitoring. Germination is the preferred test for seed lot viability. Providing such information to those using the seed is helpful. Additionally, other viability tests such as vital staining using tetrazolium solution have a greater element of subjectivity about them. This staining test is, however, sometimes used to help distinguish between dead and dormant seeds among those that did not germinate under a given test regime. Two problems relate to the germination monitoring of seed bank accessions. First, because seeds are tested soon after arrival at the bank and then at regular intervals (often every 5 to 10 years) during their storage life, the tests need to be repeatable and operator independent. Second, in order to recover as many genotypes as possible represented within a seed lot, it is necessary to break seed dormancy. This can be a particular problem in seed of wild species and where the seed is freshly harvested. Key techniques include scarification of hard seed coats to facilitate water or oxygen permeation, imbibed chilling at 5 to 10°C, and later incubation at diurnal alternating temperatures with fluorescent light (i.e., rich in red light) provided only in the higher temperature phase. Tests have to be seed lot specific as most seed dormancy is not strongly genetically inherited and the form it takes depends on the conditions under which the seed matured. Once determined, the same treatments can be used during the monitoring of that seed lot through time.

Seed banks use a variety of media for germinating the seeds, such as filter paper and sterilized sand. However, 1% (w/v) plain water agar is increasingly popular especially as it reduces the risk of imbibition damage to dry seeds. Obviously through time these monitoring tests can consume a significant proportion of each seed lot. Ideally, when the seed viability has fallen to a level where regeneration needs to take place, sufficient seeds should remain in the collection to make several attempts at growing out the collection. Because base collections may have a projected storage life of as much as 200 years, collection size needs to be large. The international
9. Duplication

One of the main advantages of seed banks is that they centralize collections of genetic material making them more easily accessed and studied. Indeed, some seed banks might be seen as some of the world’s greatest plant diversity hot spots with more individuals and, in some banks, more species per square meter than anywhere else on the planet. This centralization poses a risk to all but the most carefully located and constructed facilities. Potential catastrophic loss, which of course threatens plants conserved both ex situ and in situ, mean that duplication of collections and their associated data is an important element of seed bank safety. The FAO report (1996) indicates that the level of security duplication of plant genetic resources for food and agriculture still needs to be improved and is at best uncertain.

10. Characterization and Evaluation

Characterization can vary from accurate naming of the species or subspecies represented by the collection through to more detailed recording of characters governed by genes that are little modified by environmental factors (major genes). Such information, published in the form of descriptor lists, is of great value to plant breeders wishing to narrow their choice of material from, often, vast collections. Similarly, the concept of core collections has been established to facilitate use by breeders. A core collection genetically represents a limited set of accessions of a crop gene pool with the minimum of repetition.

Increasingly, characterization is taking the form of more detailed molecular techniques such as screening by Amplified Fragment Length Polymorphism. By contrast to characterization, evaluation records data on traits such as yield that are strongly influenced by the environment in which the plants are grown. Such data are thus site and year specific and are perhaps of less use to breeders.

11. Distribution to Users

A very important element of seed bank work is to make the seed available wherever possible. In evidence of the scale of such dispatch, ten Kate and Laird (1999) quote an annual distribution of nearly 120,000 samples by the U.S. National Plant Germplasm System of which 65% are sent abroad, many requested by reference to the Germplasm Resources Information Network (GRIN) available on the Internet. Furthermore, the usage rate of crop banks by plant breeders is probably less than the rate of request from banks holding broader plant diversity collections where uses include a wide array of pure and applied research in addition to field trials. During 1994–1996, there was a 50% request rate for seeds offered through an extensive list offered by the Kew Seed Bank.

Several elements need to be considered concerning the distribution of seed to users. The recipient should be provided with accurate information about the collection and how to germinate the seeds. It should also be remembered that certain species require a symbiont for effective growth (e.g., legumes and Rhizobium), and that the user may need to draw on germplasm for both plant and symbiont. The seed sample must be dispatched having fulfilled all necessary plant health and, where appropriate, CITES requirements. Finally, to meet the needs of the CBD and to clarify the conditions under which the material can be used, all germplasm samples are increasingly dispatched under material supply agreements.

12. Regeneration

Seed bank accessions are grown out for the purposes of regeneration of seed stock (either when seed numbers are low or when viability has reduced), for characterization and for evaluation. Many banks have a regeneration standard below which the germination of a seed lot should not fall. This is usually set at 85%. This high value limits the risk of accumulated genetic damage that is associated with seed aging. Even though falling levels of seed germination are correlated with falling levels of field establishment, many banks have adopted lower standards. This may in part be due to the backlog of regeneration work that in some national facilities highlighted by FAO (1996) is nearly 100% of the collection. By collecting high-quality seed lots in good quantity, other banks have reduced the necessity for regeneration that can be time and labor consuming and that can have adverse effects on the genetics of the collection. Samples regenerated under conditions different from where they originated can experience selection. If too small a sample is regenerated, genetic drift may occur in which rarer alleles are lost through chance. Under some circumstances, recollection, if possible, may be the more desirable option.
13. Seed Bank Design
Having considered the aspects of seed bank management, a brief consideration of seed bank design is appropriate (also see Cromarty et al., 1985). The location of the bank is important from political, practical, and security aspects. Potential risks have to be considered be they earthquake, flooding, or radiation fallout. Some facilities are placed underground such as the seed bank at Krasnodar in Russia and the Millennium Seed Bank in the United Kingdom. Others such as the NSL are located on the first floor to limit possible impact from structures above resultant from seismic activity. The size of most banks should be dictated by peak annual intake (seed drying and cleaning facilities), projected capacity before a rebuild is practical (seed storage), and annual collection maintenance (germination, field, and greenhouse facilities). Cold storage facilities vary from a few domestic deep-freezers up to large rooms such as one of 140 m² (with capacity for 130,000 samples) at NIAR in Japan.

III. OTHER TYPES OF GENE BANK
This section provides summaries of the current status of nonseed gene banks, starting with dry propagules (pollen and spores), which can be stored under conditions similar to those used for seeds, and covering normally hydrated tissues that can also be preserved under a different set of controlled conditions. Finally, field gene banks are covered and the role of botanical and zoological gardens is briefly mentioned.

A. Pollen
There are many practical reasons for storing pollen: to support work on allergenic responses, plant hybridization, and fertility; haploid plant production; and genetic transformation systems with isolated pollen (or gametes). Optimal storage conditions for pollen are similar to those used for seeds, and covering normally hydrated tissues that can also be preserved under a different set of controlled conditions. Finally, field gene banks are covered and the role of botanical and zoological gardens is briefly mentioned.

By comparison, tricellular pollen, as found in Graminaceae and Compositae, is relatively short-lived and is much more sensitive to desiccation. Most work on the long-term storage of pollen has focused on fruit tree or forestry species, for which 10 years storage at conventional gene bank temperatures (−20°C) is easily attainable. Pollen of at least 30 species are known to survive liquid nitrogen temperatures. Although pollen banking is evidently possible for many species, there does not appear to be any large-scale gene bank operation using such material.

B. Spores
Spores of many species of both pteridophytes and bryophytes are stable for months or years when dried, and this time can be extended with storage at cold or freezing temperatures. The longevity of short-lived (chlorophyllous) spores of some species can be extended significantly by drying and freezing in liquid nitrogen (Pence, 2000). Although there are data on fungal spore storage (e.g., work by Hong et al., in 1998), the majority of fungal germplasm appears to be conserved in culture or through cryopreservation of hyphae (see later).

C. Somatic and Zygotic Embryos of Plants with Nonbankable Seeds
Nonbankable seeds can nonetheless be stored using alternative approaches. Usually, these revolve around the use of rapid, partial desiccation of embryos or embryonic axes to about 20% moisture content and subsequent transferal to liquid nitrogen temperature or the use of other subzero storage temperatures. Recovery levels may be improved by pre-treatment of embryos with cryoprotectants, encapsulation of the material in alginate beads, or careful manipulation of the in vitro recovery environment. The embryos of more than 30 species have been successfully cryopreserved. Other parts of the plant can also be used to establish ex situ gene banks for species with nonbankable seeds, as described in the following section.

D. Vegetative Parts of Plants
Vegetative tissues of both pteridophytes and bryophytes can be banked for germplasm preservation (Pence, 2000). Gametophytes of many bryophyte species are naturally adapted to desiccation stress and can be cryopreserved after sufficient drying. Gametophytes of pteri-
dophytes and some desiccation-insolvent bryophytes can also be frozen when provided with some cryoprotec-
tion, such as encapsulation in alginate beads followed by dehydration or the use of abscisic acid and the amino acid, proline, as a pre-treatment. Shoot tip freezing of sporophytes of pteridophytes has also been demon-
strated. It is estimated that fewer than 200 taxa of bryo-
phytes and pteridophytes combined are currently banked worldwide using vegetative tissues, primarily at the Cincinnati Zoo and Botanical Garden and the University of Kansas in the United States, but there is significant potential for increasing this number.

The National Seed Storage Laboratory of the U.S. Department of Agriculture (USDA) also cryopreserves about 1700 apple lines using dormant scion sections, which are retrieved by grafting (i.e., no culture of meri-
stim). The lines are mainly from Malus x domestica, plus 10 to 15 other apple species. Some pear and cherry species (Towill and Forsline, 1999) are also banked in this way.

Apical shoot tips and other meristems/buds are the most popular vegetative materials for cryopreservation. Lines from about 50 species are now routinely cryopre-
served. Initial studies used shoot tips from cold hardy, temperate zone species (apple, pear), but cryopreserva-
tion methods have been extended to tropical zone spe-
cies (banana, pineapple). Successful cryopreservation depends on defining the physiological adaptation of the stock plant, the explant size and type, and its water content, the steps in the preservation process (cryopro-
tectant concentrations and rates of addition/removal; cooling/warming rates), and the recovery system. Two-
step cooling procedures have been useful for some spe-
cies, but vitrification procedures (solution-based sys-
tems and encapsulation/dehydration systems) are more favored because of the technical simplicity (Sakai, 1993). All methods are designed to reduce ice crystal growth in the specimen. Other vegetative material that has been cryopreserved using similar methodological approaches include cell suspensions and callus (more than 40 species), protoplasts (more than 10 species), and root cultures (5 species).

It should be noted that cryopreservation of vegetative germplasm overcomes the problem of genetic instability during storage as all cellular divisions and metabolic processes are stopped. In contrast, two other methods of in vitro preservation, normal and slow growth tech-
niques, run the risk of genetic changes (somaclonal variation) in the conserved germplasm that may result in the loss of distinct genotypes. Species are stored under normal growth conditions (e.g., Coffea at 27°C) for short-term purposes only. The explant (usually meristem or nodal cutting) is fre-
ently transferred (subcultured) to fresh nutrient me-
dium with the risk of microbial contamination, or loss through human error. To retard growth and hence ex-
tend the subculture interval, temperature and light in-
tensity are reduced. For example, 0 to 5°C and 1000 Lux are generally used for cold tolerant species, and 15 to 22°C and reduced light intensity for tropical species. Alternatively, growth can be slowed down by the addi-
tion to the medium of chemicals to induce mild osmotic stress (e.g., mannitol) or hormonal retardants (paclobu-
trazol, abscisic acid). Also, maintenance of tissue under reduced oxygen (e.g., under mineral oil or liquid me-
dium) slows growth. Under the appropriate conditions subculture intervals can be extended to one year or longer. The slow growth technique is now routinely used for the medium-term conservation of a number of species such as banana, potato, yam, cassava, and strawberry.

Although in vitro culture without cryopreservation poses considerable threat of genetic drift, the propaga-
tion of plant material in an aseptic environment ensures the production of disease-free stock material, which is readily accessible internationally because it satisfies most country quarantine requirements. Undoubtedly, in vitro culture is a valuable complementary approach to field conservation and is particularly useful when applied to species that are predominantly propagated vegetatively (banana, potato, and pear). produce non-
bankable or highly heterozygous seeds, and have a par-
ticular gene combination (i.e., elite genotypes; see Ash-
more, 1997). The impact of these positive features of the technique is such that FAO estimates that 37,600 accesses of plant material (vegetative and embryos) are conserved in vitro (including cryopreservation) worldwide.

### E. Animal Germplasm Samples

A majority of ex situ animal germplasm is maintained in zoological gardens and institutes around the world. The NIAR in Japan holds 621 accesses of animal germplasm, including silk worms in the living state. More than 100 of these accesses, mostly sperm, are cryopreserved (for a general methodology, see the dis-
cussion presented later). Similarly, the main gene bank methodology for ova is cryopreservation. At present, however, procedures for sperm and ova preservation are not well developed for wild species, even though a number of reported successes with artificial insemina-
tion and frozen semen can be found in the literature, especially for ungulates such as deer and antelopes. The
concept of gamete rescue from tissues has considerable value for spermatozoa, where epididymal spermatozoa are readily obtainable post mortem and can be frozen using glycerol as a reasonably standard cryoprotectant. Oocyte cryopreservation has only been achieved in the hamster, rat, rabbit, and cow and is therefore not a practical proposition at present. Interest in freezing ovarian tissue, and then culturing follicles and oocytes by various methods after thawing, has recently been resurrected and progress has included the birth of a lamb originating from ovarian tissue autotransplanted into the donor-recipient after freezing and thawing. In another recent study, isolated rat spermatogenic cells were transferred to a mouse testis, where they displayed the ability to develop into spermatozoa. The testicular cells were frozen-thawed prior to transplantation and development, and thus there may be some merit in exploring the cryopreservation of testicular cell suspensions as an alternative or adjunct to the preservation of spermatozoa. Cell suspension from genetically important animals could be used to populate the testes of common species, thus permitting the eventual harvest of spermatozoa. It should be stressed that this technique is still only in its infancy.

Unlike animal species in which reproductive cells and tissues are stored to conserve the gene pool, in humans they are stored for the use of the couple/woman electing to have them cryopreserved. Some embryos are subsequently donated for the treatment of others. Storage is limited by law in some countries.

The storage of embryos is common practice in in vitro fertilization (IVF) clinics worldwide, with about 70 offering the service in the United Kingdom alone. Oocyte cryopreservation is poorly developed and few clinics store them other than for research. An increasing number are offering tissue storage. Despite sporadic reports of cryopreservation using nonequilibrium rate cooling, the vast majority of clinics prefer conventional slow cooling rate procedures with the samples loaded into “straws” (embryos and oocytes) or plastic vials (ovary). The first step involves the addition of a cryoprotectant [about 1.5 M, 1–2 propylene glycol ≥ sucrose (pronucleate and early cleavage stage embryos), dimethyl sulfoxide (4–8 cell embryos, and blastocyst stage embryos)], ice formation is then induced at −5 to −7°C and the samples are cooled further at rates of about 0.3 to 0.5°C per minute to various subzero temperatures before storage in liquid nitrogen. After warming at appropriate rates (of about 20 to more than 300°C per minute), the samples are returned to isotonic conditions stepwise, with or without the addition of sucrose to the diluent.

F. Microorganisms

The most widely applicable preservation method for the preservation of microorganisms that retains viability and stability is cryopreservation. However, for convenience and ease of transport freeze-drying is preferred for most bacteria, viruses, and sporulating fungi (see Hunter-Cevera and Belt, 1996).

There are currently 497 collections from 60 countries maintaining microorganisms registered with the World Data Centre for Microorganisms in Japan (see Sugawara and Miyazaki, 1999). Their on-line database lists species held and the expertise and services provided by the collections along with contact addresses and links to collection websites. There are about 11,300 species held. Over 23% of the strains are held by 5 of the 497 collections (see Table I).

G. DNA Banks

DNA banks have been established in several places worldwide (Adams, 1997), the largest (more than 140,000 clones) being for plant and animal material at the NIAR in Japan. This collection is mainly constituted of rice clones (about 30,000 comprising Random Fragment Length Polymorphisms, cDNA, YAC) and pig clones (about 106,000 comprising cosmids, BAC, cDNA). The most diverse DNA bank for plants is at the Royal Botanic Gardens (RBG) Kew, which currently holds more than 10,000 DNA samples from a wide range of species. Standards of quality of preservation differ between banks and some, such as those at the Missouri Botanical Gardens (St. Louis, Missouri, in the United States) and several zoos, store only frozen tissue. Others extract DNA and purify it to varying degrees. Purposes for these banks differ as well, with some established to hold samples of a particular country or region with the intent of using these in conservation genetic studies, whereas others focus more on taxonomic and systematic studies, such as that at RBG Kew. Most banks are prepared to consider sharing aliquots of DNA or small samples of tissue with researchers at other institutions.

H. Field Gene Banks

Field gene banks are ex situ collections of mainly agricultural or forest species. They should be contrasted with what might be termed “farm gene banks” where crop germplasm is, in effect, conserved in situ by the farmer. Field gene banks normally comprise considerably more individuals per accession than is the case in
botanic gardens. Their particular use is for the conservation and utilization of species with the following traits:

- Have nonbankable seeds
- Have long life cycles where growing up material for regular study from a seed collection is impractical
- Are normally vegetatively propagated

Essentially, they are not a new idea as the Kayapo people of Brazil maintain germplasm collections of tuberous plants in hillside gardens protected from flood (see Plucknett et al., 1987). Most countries have at least one field gene bank and FAO (1996) estimate some 527,000 accessions are conserved in this way worldwide. Examples of field gene banks include the National Fruit Collection in the United Kingdom, one for cassava at CIAT, Colombia, one for sugarcane at the Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia (CENARGEN), Brazil, and the potato collection at the Centro Internacional de la Papa (CIP) in Peru. Such facilities are considered by FAO to be particularly important in small island developing states. It should also be noted that the Nordic Gene Bank, which operates on a regional basis, unlike many seed banks, includes within its remit the *in situ* conservation of wild crop relatives.

While they offer the opportunity for characterization and evaluation, such collections are labor intensive and are susceptible to catastrophic events. For example, a field collection of yams was lost in St. Lucia during 1994 as the result of cyclone damage. This is one of the reasons that a number of field gene bank collections are now backed up *in vitro*.

### Table I

Collections Maintaining Microorganisms

<table>
<thead>
<tr>
<th>Collection</th>
<th>Location</th>
<th>Number of strains</th>
<th>Type of material</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural Research Service Culture Collection, USDA</td>
<td>Peoria, IL (US)</td>
<td>78,010</td>
<td>Algae, bacteria, fungi, yeasts, actinomycetes</td>
</tr>
<tr>
<td>American Type Culture Collection (ATCC)</td>
<td>Rockville, MA (US)</td>
<td>53,615</td>
<td>Algae, bacteria, fungi, yeasts, prototaxa, cell lines, hybridomas, viruses, vectors, plasmids, phages</td>
</tr>
<tr>
<td>CABI Bioscience UK Centre (formerly IMI)</td>
<td>Egham, UK</td>
<td>21,000</td>
<td>Fungi, bacteria, yeasts</td>
</tr>
<tr>
<td>Centraalbureau voor Schimmelcultures (CBS)</td>
<td>Baarn, The Netherlands</td>
<td>41,100</td>
<td>Fungi, yeasts, lichens, plasmids</td>
</tr>
<tr>
<td>Culture Collection, University of Goteborg (CCUG)</td>
<td>Goteborg, Sweden</td>
<td>28,100</td>
<td>Bacteria, fungi, yeasts</td>
</tr>
</tbody>
</table>

I. Botanical and Zoological Gardens

Botanical and zoological gardens may be seen as types of gene bank with relatively few individuals per accession. While zoos have embraced careful breeding programs that help maximize the genetic value of the limited collections across the world, this is much less true of collections in botanical gardens. Here species may be represented within the "botanical gardens flora" by a single genotype such as the clonal exchange of material. See also "The Role of Botanical Gardens" and "Zoos and Zoological Parks."

Acknowledgments

The authors would particularly wish to acknowledge the following for their assistance in providing information: Mr. M. Ambrose (John Innes Centre, United Kingdom); Dr. D. Astley (HRI, United Kingdom); Dr. M. Fay, Mr. R. Smith and Mr. M. Way (RBG Kew, United Kingdom); Dr. W. Holt (Institute of Zoology, London, United Kingdom); Dr. S. Miyazaki and Mr. A. Yamamoto (NIAR, Japan); Dr. B. Panis (Leuven, Belgium); Dr. V. C. Pence (CREW, Cincinnati, United States); Dr. D. Smith (CABI Bioscience, United Kingdom); Dr. S-H Tan (CBS, Netherlands); Dr. L. Towill and Dr. C. Walters (NSSL, United States); Dr. M. Wood (London, United Kingdom); and Dr. X-Y. Yang (XTBG, Peoples Republic of China).

See Also the Following Articles

- BREEDING OF ANIMALS
- BREEDING OF PLANTS
- CAPTIVE BREEDING AND REINTRODUCTION
- CROP IMPROVEMENT AND BIODIVERSITY
- EX SITU, IN SITU CONSERVATION
- GENETIC DIVERSITY
- ZOOS AND ZOOLOGICAL PARKS
Bibliography


Hong, T. D., Linnington, S., and Ellis, R. H. (1998). Compendium of Information on Seed Storage Behaviour. 2 volumes. Royal Botanic Gardens, Kew, United Kingdom in collaboration with the University of Reading and IPGRI.


GENES,
DESCRIPTION OF

Michael F. Antolin and William C. Black IV
Colorado State University

I. Introduction
II. Structure of DNA and RNA
III. Genes
IV. Prokaryotic Genomes
V. Eukaryotic Genomes—Chromosomes
VI. Organelles and Their DNA
VII. Mutation

GLOSSARY

allele - Variant of a gene that can vary at the nucleotide level with or without affecting phenotypic expression.

chromatin - Complex of DNA and associated proteins that make up chromosomes of eukaryotes.

chromosome - Threadlike structure that includes DNA and proteins (containing genes arranged in a linear sequence along the thread), which can be visualized when condensed during cell division.

DNA - Deoxyribonucleic acid, the molecule of inheritance that stores genetic information that is passed from one cell to another and from one generation to succeeding generations; it is composed of four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T).

eukaryotes - Organisms with genetic material organized into chromosomes that are contained within a membrane-bound nucleus in the cell; eukaryotic cells undergo mitosis and meiosis during cell division, ensuring the equal division of chromosomes among daughter cells.

exons - Set of segments of interrupted genes that remain after cutting and splicing of messenger RNA, and that include the parts of the gene that are translated into proteins.

genetic code - Language that specifies how DNA will be translated into protein sequences by means of three-nucleotide "words" (codons) that specify the 20 amino acids and regulators of transcription (start and stop codons).

heterochromatin - Regions of chromosomes that do not include coding DNA, generally make up the structure of chromosomes, and always remain condensed during a cell's life cycle.

introns - Segments of interrupted genes that are removed after transcription and before translation of messenger RNA to proteins.

locus - A precise location in the genome, whether a gene is found there or not; formerly this term was used interchangeably with gene, but the definition has become more specific in the era of molecular genetics.

nucleotide - Subunit of DNA and RNA composed of a ringed five-carbon sugar, a ringed nitrogen-rich base, and phosphates; nucleotides are often referred to as base pairs because individual types form complementary hydrogen bonds (G-C and A-T in DNA, G-C and A-U in RNA) to make double-stranded DNA and RNA molecules.
organelles Membrane-bound structures within eukaryotic cells that carry out specific functions and that may contain their own DNA; typical organelles are mitochondria and chloroplasts.

post-transcriptional modification Cutting and splicing of mRNA in eukaryotic cells, in some cases to produce alternative proteins with different structural or regulatory properties (e.g., sex determination in Drosophila).

prokaryotes Single-celled organisms defined by having their DNA arranged as a circular molecule not contained within a nucleus, and which reproduce by simple fission.

repetitive DNA Regions of DNA that include the same DNA sequence repeated up to several hundred or thousand times; regions with repeated segments that involve only 2–5 base pairs of DNA are called microsatellites.

RNA Ribonucleic acid, composed of nucleotides like DNA, but differing from DNA in that the base uracil (U) in RNA replaces thymine (T) and single-stranded RNA molecules form important structural and regulatory parts of cells.

transposable elements Fragments of DNA containing genes that provide the ability for the DNA fragment to change its location in the genome.

UNDERSTANDING THE MECHANISMS OF HEREDITY IS A KEY TO THE STUDY OF BIODIVERSITY FOR AT LEAST THREE REASONS. First, heritable genetic variation provides the material for evolution—descent with modification—including adaptation within populations, diversification of lineages, and the formation of new species. Second, standing genetic variation within populations may reflect the recent history of populations and indicate prospects for change in the future. Third, the variation in molecular genetic markers can be used to track the relationships among living groups of organisms and the taxonomic status of individual populations. Genes are defined as sequences of DNA that specify cell structure, including proteins and several types of RNA. Genes also code for enzymes and proteins that bind to DNA or RNA in order to control gene expression and other physiological processes in cells. In former times, the term “locus” was used interchangeably with “gene.” Today, locus simply means a particular place in the genome, whether the DNA at that location codes for a gene product or not. Locus may still be used to describe neutral molecular markers; allelic variation at the DNA level is used to study population genetics, relatedness within populations, parentage, and genome mapping.

1. INTRODUCTION

Genes and genomes are both complex and continually modified. Organisms share a common biochemical mechanism of heredity, but at the same time display characteristic differences that allow us to make hypotheses about the history of life. Genetic variation is pervasive in Nature and continually arises by the inevitable and inexorable process of mutation, both on the level of individual nucleotides within DNA molecules and on the level of chromosomal gene rearrangements. One of the most significant (and initially unappreciated) discoveries in genetics in this century is that genomes can be altered by transposable elements. The original discovery of transposable elements was in maize by Barbara McClintock, who in the 1930s correlated regular breaks in chromosomes with the deactivation of purple pigment genes in corn seeds. Reactivation of pigment genes was also associated with the same breaks appearing at other places in the genome. We now know these as “jumping genes,” which carry with them the genes that control their ability to duplicate and reinsert themselves into multiple places within genomes, sometimes carrying other genes with them, sometimes disrupting genes at their insertion sites. Dozens of transposable elements are now known, and more continue to be discovered in an ever-widening sample of organisms (Li, 1997). As a consequence, it is no longer possible to view genomes as static entities in which genes are aligned in precise locations that remain stable over long periods of time.

This article considers genes in the context of the entire genomes of diverse organisms. In prokaryotic organisms (bacteria, blue-green algae), genomes are organized as naked circular DNA molecules within the cytoplasm. In eukaryotic organisms, most genetic material is arranged into chromosomes—long, linear arrays of DNA bound with proteins—found within a membrane-enveloped nucleus. Eukaryotic organisms also have genes in separate prokaryotelike genomes inside cell organelles such as mitochondria (plants and animals) and chloroplasts (the photosynthetic organelles of plants). All genomes contain regions of noncoding single-copy DNA and repetitive DNA. Population genetic and phylogenetic studies depend on a variety of molecular markers derived from noncoding parts of genomes that are presumably neutral (i.e., not under selection), and Avise (1994) provides an in-depth over-
view of the use of molecular markers in the study of evolution.

II. STRUCTURE OF DNA AND RNA
What follows should serve as a review for biologists who do not regularly converse with the world of molecular genetics. A full description of genes is not possible in the space of a few pages; those seeking more detail should consult other sources (e.g., Hartl and Jones, 1998; Lewin, 1997). Molecular genetics is a mushrooming field, but the salient features of genes are universal, and genes and the genetic code can be understood on an intuitive level without extensive knowledge of biochemistry. In general, genes are carriers of information working in a nested series from individual nucleotides, to relatively large RNA molecules and enzymes composed of hundreds of nucleotides or amino acids, to enormous cell-structure proteins comprising thousands of amino acids folded into complex three-dimensional shapes.

First and foremost in this hierarchy of information are nucleic acids, DNA and RNA, which are the molecules of heredity for most life-forms (one notable exception being prions, which cause the famed "mad-cow disease" and a number of neurological disorders in humans: see the review by Prusiner and Scott, 1997). Both DNA and RNA are made up of four nucleotide subunits linked together to form long chains (Fig. 1). Nucleotides themselves are made up of three parts: ringed sugars, phosphates, and ringed nitrogen-and-carbon bases. Nucleotides are joined to each other to form chains by bonds between phosphates and sugars. Nucleotides are complementary to each other in terms of the numbers of hydrogen bonds they form between nitrogenous bases: A (adenine) and T (thymine) make two bonds, G (guanine) and C (cytosine) make three bonds. Two complementary chains intertwine to form double-stranded DNA or RNA by hydrogen bonding between bases. Double-stranded DNA molecules take on the classic "double-helix" shape, where two chains of nucleotides joined by hydrogen bonds spontaneously form twists of $-360^{\circ}$ so that $-10$ base pairs make a complete turn of $360^{\circ}$. DNA can take on a number of

![DNA Structure Diagram](image.png)

**FIGURE 1** DNA (and RNA) are composed of four nucleotide subunits—A, C, G, and T (U replaces T in RNA)—each covalently bonded to a sugar, with sugars linked to each other by a phosphate bond (circled P) between the 3’ and 5’ carbon atoms of the adjacent sugars. Complementary strands of DNA are held together by hydrogen bonds (shown as broken lines). Of the two strands, the lower is in the 5’ to 3’ direction, showing how nucleic acid sequences are usually written (GTCA).
other forms, such as supercoiling structures, depending on conditions such as temperature, pH, and the exact sequence of nucleotides within the molecule. RNA also can form a double helix, but one with a broader diameter. An additional feature of RNA is that single-stranded molecules include complementary sequences that allow the single-stranded RNAs to fold back on themselves, forming double-stranded helical stems and other three-dimensional grooves, knobs, and loops (Gutell, 1996). Three-dimensional structure is a vital characteristic of the long ribosomal RNA (rRNA) that make up subunits of ribosomes and the shorter transfer RNA (tRNA) molecules that are part of the mechanisms of translation of messenger RNA (mRNA) into protein.

III. GENES

The number of genes within organisms varies tremendously among viruses, bacteria, and multicellular eukaryotic organisms. For instance, it is estimated that humans have approximately 50,000 genes, the fruit fly Drosophila melanogaster has 13,600, the nematode worm Caenorhabditis elegans has 19,000, the mustard Arabidopsis thaliana has 25,000, the yeast Saccharomyces cerevisiae has 6000, the bacterium Escherichia coli has 4286, and the HIV (human immunodeficiency virus) RNA retrovirus has only 9. For the organisms whose entire genomes and all nucleotide sequences have been characterized (yeast, worm, E. coli, HIV), the number of genes is known with accuracy. For others, the numbers represent estimates based on incomplete mapping studies. For example, some estimates of the number of genes in humans are as high as 140,000 genes, but the upper limit is usually given as 80,000 genes.

The flow of information within a cell generally follows a pathway, with genes coded in DNA transcribed into three types of RNA, which then combine to read the genetic code and translate the information into proteins (or enzymes). This process of information transfer, which has been called the “central dogma of molecular biology,” is depicted in Fig. 2. Two kinds of genes result from the transfer of information from DNA to proteins. First, genes may encode enzymes such as DNA polymerases and protein transcription factors that control the physiological and biochemical processes within a cell. The second type of gene encodes the physical structures of the cell, including cytoskeletal proteins like tubulin and actin that affect cell shape, and subunits of ribosomal RNA that make up the machinery of protein translation. The majority of genes are expressed as proteins and enzymes, and most of these are involved in cellular regulation rather than cell structure.

Protein and enzyme coding is specified by the genetic code (Table I), which contains the language of protein synthesis. The genetic code has a series of three-base codons that specify 20 amino acids as well as codons for “start” and “stop.” This code is almost universal, although there are a number of differences between nuclear and mitochondrial DNA in eukaryotes and there are unique codons in a few bacteria and Protista (Li, 1997). The code functions by specifying three-base-pair codons of mRNA that match three-base-pair anticodons found on a loop of the transfer RNA. Each tRNA is specific to a particular amino acid. The complex dynamic molecular structure involved in translation of the genetic code contains mRNA, ribosomes (rRNA plus proteins), and tRNAs.

Because DNA contains specific information, transcription and translation must proceed in only one di-
TABLE I

The Genetic Code That Shows How Twenty Amino Acids Are Specified by Three-Base-Pair Codons

<table>
<thead>
<tr>
<th>First Position (5’ end)</th>
<th>Second position</th>
<th>Third position (3’ end)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU</td>
<td>UAU</td>
</tr>
<tr>
<td></td>
<td>UUC PHE</td>
<td>UAC TYR</td>
</tr>
<tr>
<td></td>
<td>UUA SER</td>
<td>UAA</td>
</tr>
<tr>
<td></td>
<td>UUG LEU</td>
<td>UAG STOP</td>
</tr>
<tr>
<td></td>
<td>CLU</td>
<td>CAU</td>
</tr>
<tr>
<td></td>
<td>CUC</td>
<td>CAC HIS</td>
</tr>
<tr>
<td></td>
<td>CLA LEU</td>
<td>CAA</td>
</tr>
<tr>
<td></td>
<td>CUG</td>
<td>CAG GLN</td>
</tr>
<tr>
<td></td>
<td>AUG MET</td>
<td>AAG LYS</td>
</tr>
<tr>
<td></td>
<td>AUU</td>
<td>AAA</td>
</tr>
<tr>
<td></td>
<td>AUC ILE</td>
<td>AAC ASN</td>
</tr>
<tr>
<td></td>
<td>AUA</td>
<td>AAA</td>
</tr>
<tr>
<td></td>
<td>AUG MET</td>
<td>AAG LYS</td>
</tr>
<tr>
<td></td>
<td>GGU</td>
<td>GGU</td>
</tr>
<tr>
<td></td>
<td>GUC</td>
<td>GAC ASP</td>
</tr>
<tr>
<td></td>
<td>GUA VAL</td>
<td>GAA</td>
</tr>
<tr>
<td></td>
<td>GUG</td>
<td>GAG GLU</td>
</tr>
</tbody>
</table>

The three stop codons (UAA, UAG, UGA) and the start condon for methionine (AUG) are shown in bold. Most amino acids are specified by more than a single codon, with several (arginine, leucine, serine) specified by six codons. There is a weakly positive general relationship between the frequency of incorporation of an amino acid into proteins and the number of codons specifying the amino acid. In most cases, a change in the third position in a codon does not result in a change in the amino acid because most tRNA anticodons will still bind to mRNA codons if the first two positions on the mRNA match the tRNA. This is the so-called wobble effect and redundancy in the genetic code.

Correction for the information to be communicated accurately. This is analogous to the convention in most written Indo-European languages, in which words form sentences that are read on a page from left to right. The direction along any strand of DNA or RNA can be identified by which carbon atoms of the ringed-sugar-and-phosphate backbone are bonded together. Carbons in the ribose sugar are numbered 1 to 5, and bonds occur between a phosphate attached to the 5’ carbon of one sugar with a hydroxyl group attached to the 3’ carbon of the next. Consequently, a directionality can be specified along each nucleic acid molecule, and the genetic code must follow this direction for the language of transcription and translation to be faithfully transmitted. For instance, RNA polymerases transcribe by adding new nucleotides to the 3’ end of a growing RNA molecule, and so create the RNA in the 3’ to 5’ direction. However, the transcription is from a complementary DNA chain that is read in the 3’ to 5’ direction. In general, reading frames of DNA are written by placing the 5’ end at the left, and protein sequences are written in the same left-to-right direction specified by the nucleotide sequence (Table II).

IV. PROKARYOTIC GENOMES

The genomes of most prokaryotes (bacteria and blue-green algae) are composed of one double-stranded circular DNA molecule attached to a central core of proteins in a series of supercoiled loops emanating from the center like the cotton fibers of a dust mop. Although the nucleoid “chromosome” is not enclosed within a nuclear membrane, it tends to be found in a particular part of the cytoplasm. An additional feature of prokaryotes is their ability to acquire smaller circular DNA molecules, called plasmids, in their cytoplasm. Plasmids may contain numerous genes, including genes found in the nucleoid. Transfer of plasmids into and out of bacterial cells is the major feature that makes them vital.
TABLE II
Types of Chromosomal and Gene Mutations

**A. Chromosomal Mutations**

<table>
<thead>
<tr>
<th>DNA sequence (of a chromosome)</th>
<th>Type of mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGTTCGTACCTGATCGACG</td>
<td>Deletion of AGGGT</td>
</tr>
<tr>
<td>AGTTCGTAGGGTGATCGACG</td>
<td>Insertion of CATGAG</td>
</tr>
<tr>
<td>GAGGGCGAGGTACCTGATCGACG</td>
<td>Translation of GTGTC and GAG</td>
</tr>
<tr>
<td>AGTTCGTAGGGTGACCTGACG</td>
<td>Duplication of CCTGA</td>
</tr>
</tbody>
</table>

**B. Gene Mutations**

<table>
<thead>
<tr>
<th>Three-base-pair codons (top line), and the proteins specified (lower line)</th>
<th>Type of mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUG UGG UGU GAG UAC AUU CGA GAG AAG UAG</td>
<td>Synonymous (no amino acid change)</td>
</tr>
<tr>
<td>AUG UGG UGU GAG UAC AUU CGA GAG AAG UAG</td>
<td>Missense (amino acid change)</td>
</tr>
<tr>
<td>AUG UGG UGU GAG UAC AUU CGA GAG AAG UAG</td>
<td>Non-sense (mutation to a stop codon)</td>
</tr>
<tr>
<td>AUG UGG UGU GAG UGU GAG AAG UAG</td>
<td>Non-sense (insertion and shift of reading frame)</td>
</tr>
<tr>
<td>AUG UGG UGU GAG UGU GAG AAG UAG</td>
<td>Non-sense (deletion and shift of reading frame)</td>
</tr>
</tbody>
</table>

*Very small fragments are illustrated; chromosomal mutation usually affect large fragments. After Li (1997).*

to DNA-based biotechnology and is responsible for the rapid transfer of genes for traits like antibiotic resistance between bacterial strains.

Genes in prokaryotes are collinear and uninterrupted, in that one to several genes are expressed in the order in which they follow a promoter. Transcription of mRNA leads directly to translation with little modification (Fig. 3). This arrangement of several genes being expressed as a group is sometimes referred to as polycistronic, and it is restricted to prokaryotic and related genomes (e.g., organelles; see Section VI). The control of gene expression in prokaryotes is through either positive or negative regulation. Under positive control, a transcription factor must bind with the regulatory site to initiate transcription. This is part of the lac operon of *E. coli*, one of the earliest metabolic pathways described at the molecular level by Jacques Monod and colleagues at the University of Paris. The lac operon is mainly under negative control of a repressor that is removed when lactose is present in the cell. However, transcription also depends on a second positive control on the promoter, where a second protein must bind for transcription of RNA. This double control allows the

![FIGURE 3 General structure of a prokaryotic gene, in which a repressor, promoter, and operator act together to control expression of three genes. Repressors, operators, and promoters are binding sites on the DNA where enzymes attach to begin or prevent transcription.](image)

FIGURE 3 General structure of a prokaryotic gene, in which a repressor, promoter, and operator act together to control expression of three genes. Repressors, operators, and promoters are binding sites on the DNA where enzymes attach to begin or prevent transcription.
bacterial cells to turn off the lac operon when glucose, a preferred food source to lactose, is present in the cell.

V. EUKARYOTIC GENOMES—CHROMOSOMES

The genomes of eukaryotic organisms are characterized by DNA in linear molecules with associated proteins in a complex structure known as chromatin, which is contained within a membrane-bound nucleus. Chromatin is located in a diffused state during interphase of the cell cycle when transcription and translation occur. When the chromatin condenses during cell division and becomes visible, the heavily staining structures are called chromosomes. The number of chromosomes found in different eukaryotic groups vary considerably, but each group displays a characteristic range of chromosome numbers. For example, humans and great apes have haploid chromosome numbers of \( N = 23 \) and \( N = 25 \), respectively, whereas different groups of parasitic Hymenoptera (wasps) have \( N = 5 \) and \( N = 10 \).

Eukaryotic genes contrast sharply with prokaryotic genes in both structure and regulation of gene expression (Fig. 4). First, whereas several prokaryotic genes are grouped together under control of a single promoter, each eukaryotic gene is expressed individually under control of a promoter that is specific to that gene. Second, eukaryotic genes are often interrupted, with non-translated introns interspersed between the exons that are translated into proteins. The result is that after mRNA is transcribed in eukaryotes, introns are excised from the mRNA, which is then spliced back together as a mature mRNA for translation into proteins. One consequence of this post-transcriptional modification is that the same gene can be spliced into several different mature mRNAs by the inclusion or exclusion of different exons. A classic example of how alternative splicing controls development in animals is through the cascades of gene expression that determine sex and control sexual differentiation in animals (Hodgkin, 1990; Marin and Baker, 1998). In Drosophila melanogaster, for instance, the male and female forms of the gene Doublesex both have their first three exons in common, but differ in which 3’ exons are included in the transcript after splicing.

Because most eukaryotes are multicellular, they face three other challenges that do not apply to prokaryotes. First, development from a single-celled egg to an adult requires regulated and coordinated changes in gene expression throughout the life span of an organism. Second, after tissues differentiate, they have unique suites of genes expressed within them. Think of the differences between a heart muscle cell and a secretory cell in the pancreas. Third, cells must communicate with each other, so that much of the control of expression of genes within a cell may come from hormones or other signaling molecules that diffuse or are actively taken up into the cell.

Most regulation of gene expression in eukaryotes is through positive control. A transcription factor must bind with the promoter to initiate transcription of the message by RNA polymerase. However, the ability of the transcription factor to bind to the promoter will be influenced by other proteins that bind to both the enhancer sequences on the DNA and the transcription factor. Enhancer sites are usually sequences of DNA less than 20 nucleotides long. Promoter binding regions will often include a specific sequence recognized by a transcription factor. A common promoter site consists of a TATA nucleotide sequence, thus the name “tata box.”

Another common feature of eukaryotes is gene duplication. Some genes that are important for basic housekeeping, like rRNA, may have hundreds or thousands of copies of the gene within the genome, which results in functional redundancy. Another form of duplication is represented by genes coding for blood proteins (globins) in vertebrates. These gene families in-
include several kinds of duplications, where gene clusters code for a diversified set of proteins that are part of fetal blood versus adult blood on the one hand, and a series of nonfunctional pseudogenes on the other hand.

**A. Noncoding Repetitive DNA Sequences**

A large proportion of the genomes of eukaryotes and some prokaryotes consists of repetitive DNA, which is completely untranscribed and consists of tandemly arranged sequences with identical nucleotide composition that can be repeated up to several million times within the genome. Some repetitive DNA is tied up in the heterochromatic parts of the genome that make up structural elements of chromosomelike centromeres (where sister chromatids bind during cell division) and telomeres (the ends of chromosomes). Otherwise, repetitive DNA may be interspersed throughout the genome. Repetitive elements can be very large (>10,000 nucleotides) or of moderate size (100–800 nucleotides), but the majority of highly repetitive sequences consist of repeats smaller than 100 nucleotides. Generally speaking, repetitive DNA makes up more than half of eukaryotic genomes, and much of the repetitive DNA is now thought to arise as a by-product of transposable elements moving within genomes (Li, 1997).

Repetitive DNAs are highly variable among individuals in a population and are generally classified according to the size of the repeat units, the number of repeat units per array, and the genomic location of the tandem arrays. There is presently no consensus in the classification of simple tandem repeats and the terminology used varies among authors. The term “microsatellite” denotes any tandem array of repeats with a unit length of 2–5 nucleotides. Microsatellites are usually found at places in the genome where genes are large duplications or amplifications, where the number of copies may range from a few to a few million, forming a series of nonfunctional pseudogenes on the other hand.

A. Noncoding Repetitive DNA Sequences

A large proportion of the genomes of eukaryotes and some prokaryotes consists of repetitive DNA, which is completely untranscribed and consists of tandemly arranged sequences with identical nucleotide composition that can be repeated up to several million times within the genome. Some repetitive DNA is tied up in the heterochromatic parts of the genome that make up structural elements of chromosomelike centromeres (where sister chromatids bind during cell division) and telomeres (the ends of chromosomes). Otherwise, repetitive DNA may be interspersed throughout the genome. Repetitive elements can be very large (>10,000 nucleotides) or of moderate size (100–800 nucleotides), but the majority of highly repetitive sequences consist of repeats smaller than 100 nucleotides. Generally speaking, repetitive DNA makes up more than half of eukaryotic genomes, and much of the repetitive DNA is now thought to arise as a by-product of transposable elements moving within genomes (Li, 1997).

Repetitive DNAs are highly variable among individuals in a population and are generally classified according to the size of the repeat unit, the number of repeat units per array, and the genomic location of the tandem arrays. There is presently no consensus in the classification of simple tandem repeats and the terminology used varies among authors. The term “microsatellite” denotes any tandem array of repeats with a unit length of 2–5 nucleotides. Microsatellites are usually found at places in the genome where genes are large duplications or amplifications, where the number of copies may range from a few to a few million, forming a series of nonfunctional pseudogenes on the other hand.

A. Noncoding Repetitive DNA Sequences

A large proportion of the genomes of eukaryotes and some prokaryotes consists of repetitive DNA, which is completely untranscribed and consists of tandemly arranged sequences with identical nucleotide composition that can be repeated up to several million times within the genome. Some repetitive DNA is tied up in the heterochromatic parts of the genome that make up structural elements of chromosomelike centromeres (where sister chromatids bind during cell division) and telomeres (the ends of chromosomes). Otherwise, repetitive DNA may be interspersed throughout the genome. Repetitive elements can be very large (>10,000 nucleotides) or of moderate size (100–800 nucleotides), but the majority of highly repetitive sequences consist of repeats smaller than 100 nucleotides. Generally speaking, repetitive DNA makes up more than half of eukaryotic genomes, and much of the repetitive DNA is now thought to arise as a by-product of transposable elements moving within genomes (Li, 1997).

Repetitive DNAs are highly variable among individuals in a population and are generally classified according to the size of the repeat unit, the number of repeat units per array, and the genomic location of the tandem arrays. There is presently no consensus in the classification of simple tandem repeats and the terminology used varies among authors. The term “microsatellite” denotes any tandem array of repeats with a unit length of 2–5 nucleotides. Microsatellites are usually found at places in the genome where genes are large duplications or amplifications, where the number of copies may range from a few to a few million, forming a series of nonfunctional pseudogenes on the other hand.

**VI. ORGANELLES AND THEIR DNA**

The most likely origin of organelles within the cells of eukaryotic organisms is ancient bacteria or blue-green algae taken up as endosymbionts into the cells of early eukaryotes. The strongest line of evidence for an endosymbiotic origin is that organelle genomes most closely resemble those of prokaryotes, both in general characteristics and in the details of DNA sequences within genes. Like prokaryotic genomes, organelle genomes are circular double-stranded molecules with very small amounts of repetitive DNA. In addition, organelle genes do not contain introns, are transcribed polycistronically, and have relatively simple prokaryote-like promoters (Mayfield et al., 1999). Although organelles carry out their own transcription and translation, gene expression and regulation in organelles depend on regulatory proteins that originate from nuclear genes.

**A. Mitochondrial DNA in Animals**

The animal mitochondrial genome (mtDNA) is a single circular, double-stranded DNA molecule that ranges in size from 14,000 to 42,000 nucleotides. With few exceptions the animal mitochondrial genome contains 37 genes that encode 13 protein subunits, 22 tRNAs, and a small and a large ribosomal subunit RNA. Several compilations of primers that are useful for polymerase chain reaction (PCR) amplification of segments of the mitochondrial genomes of animals are currently available (Avise, 1994; Simons et al., 1994; Palumbi, 1996; Roehrlanz and Degriggellier, 1998).

The analysis of mitochondrial DNA has become one of the most powerful tools for studying animal popula-
The mitochondrial genome is primarily maternally inherited as a haploid genome. It does not recombine and it mutates at a faster rate than the nuclear genome in most animal groups. For these reasons mitochondrial DNA can be used to estimate phylogenetic relationships among maternal lineages. Sequence differences arising from mutations in mtDNA haplotypes record the phylogenetic histories of female lineages within and among species. Variation in the frequencies of different mitochondrial haplotypes can be used to estimate effective migration rates among populations and genetic diversity within and among populations. This level of variation is also used for estimating phylogenetic relationships among recently evolved taxa. There are a number of unique perspectives on evolution that have been prompted by genetic findings on animal mitochondrial DNA. These include the usual pattern of within-individual homoplasy (predominance of a single mtDNA sequence) despite between-individual sequence differences, and the rapid pace of nucleotide substitution.

B. Mitochondrial DNA in Plants

Plant mtDNA is extremely different from animal mtDNA, and this fact has prompted several investigators to suggest independent symbiotic origins. The mtDNA genome in plants ranges from 200,000 to 2,400,000 nucleotides in circumference, and typically exists as a collection of different-sized circles arising from extensive recombinational processes within individuals that convert between a “master” molecule and subgenomic circles. Inheritance is usually (but not always) maternal. Although plant and animal mtDNAs are similar with regard to gene content and general function, their evolutionary patterns are diametrically different. Plant mtDNA appears to evolve rapidly with respect to gene order, but slowly in nucleotide sequence (perhaps 100-fold slower than animal mtDNA). Reasons for the slow accumulation of point mutations are not understood, but plant mtDNA may possess relatively error-free DNA replication systems, or perhaps highly efficient enzymes for repair of DNA damage. The lack of nucleotide-level variation and regular recombination have limited the utility of plant mtDNA in population genetics.

C. Chloroplast DNA (cpDNA)

The cpDNA molecule varies in size from about 120,000 to 247,000 nucleotides in circumference in photosynthetic land plants, due largely to reiteration of a large inverted repeat that includes genes for the ribosomal RNA subunits. There are from 22 to 900 copies of the genome in each chloroplast and each encodes 123 genes. These include 4 genes that encode rRNA, 20 genes for the ribosomal proteins, 30 genes for the tRNAs, many proteins that are important for photosynthesis, 6 of the 9 genes for the synthesis of ATPase, and chloroplast RNA polymerase. Unlike mitochondrial DNA, 15 of the cpDNA genes contain introns. Most chloroplast proteins are encoded by the nucleus. For example, chloroplast ribosomes consist of 52 proteins, only 19 of which are encoded by the plastid genome. Chloroplast DNA is transmitted maternally in most flowering plants, biparentally in a few, and paternally in gymnosperms. These cpDNA genes have been shown to transpose to the nucleus and there is good evidence that mtDNA, cpDNA, and nuclear genomes exchange genes. The rate of cpDNA evolution generally appears to be slow in terms of both primary nucleotide sequence and gene rearrangement. Because of the large size of the cpDNA genome, most systematic treatments have involved restriction site or sequence determinations for particular genes or have monitored the taxonomic distributions of unique cpDNA structural features across higher-level plant taxa. Nonetheless, some studies have uncovered considerable intraspecific cpDNA variation as well, and this suggests that portions of the cpDNA should be useful in population genetics. Several primers for amplifying cpDNA genes in a number of plant species are available (e.g., ndhF: Olmstead and Sweere, 1994; rbcL: Palumbi, 1996).

VII. MUTATION

Mutation is the ultimate source of genetic variation and is generally defined as any change in genetic material, whether phenotypic effects caused by the mutation are large, small, or nondetectable. The causes of mutation include direct insults from environmental effects like ultraviolet light, chemical solvents, X rays, and other atomic particles, but many mutations arise from errors in DNA replication and DNA repair during cell division. One class of mutations arises from gross changes at the level of chromosomes, where chromosomes break and are then modified by deletions, insertions, inversions, translocations, and duplications (see Table II). Although the term “chromosome” is used here, similar mutations affect the genomes of prokaryotes as well. Chromosomal mutations can result in changes in the total amount of DNA found in different species and, in the case of gene duplications or deletions, mutations
can change the number of genes. Mutations like these can have drastic phenotypic effects. Individual genes can be disrupted if insertions occur within them, or if parts of a gene are lost in a deletion. Transposable elements cause deleterious mutations by inserting themselves into genes.

Inversions of chromosomal fragments, where a fragment is flipped 180° from its original orientation and reinserted at the same place, will cause ballooning of the chromosomes when they pair during meiosis and will prevent recombination. Inversions are common in some organisms, like the Duperia (especially Drosophila). The lack of crossing-over and recombination means that genes within an inversion will be inherited intact as a “supergene.”

Translocation of pieces of chromosomes from one place to another within the genome will result in change in the size or number of chromosomes. The effects of translocations can be dramatic, often resulting in sterility because of the inability of the altered chromosomes to pair during meiosis. Thus, translocations are a potential mechanism of isolation of a single species into incompatible breeding groups, leading eventually to the formation of new species.

The magnitude of phenotypic effects of duplications depends on the size of the affected chromosomal fragment, and thus the number of genes that are duplicated. An example is trisomy 21, which results in Down syndrome (mental retardation, flat facial features, heart defects). Trisomy results from a failure of proper segregation of chromosomes during meiosis so that one gamete contributes two complete copies of chromosome 21, resulting in three copies in the zygote. On the other hand, duplications are an integral part of the evolution of multigene families, like the various globin genes and nonfunctional pseudogenes in vertebrates. Duplications are the main mechanism by which entirely new genes arise (Li, 1997).

Mutations can also occur at the level of single-nucleotide changes, including some that affect the reading of the genetic code (see Table II). Single-nucleotide substitutions are classified as transitions if they occur between similar nucleotides, that is, between purines (A, G) or pyrimidines (C, T), and as transversions if mutations substitute purines for pyrimidines and vice versa (A, G → T, C). This distinction is important because transitions are much more common than transversions, sometimes occurring 20 times more frequently. Further examination of the genetic code (see Table I) reveals that where a substitution occurs within a codon will determine the severity of the mutation. Most mutations in first nucleotide position will result in an amino acid change (missense) or a mutation to a stop codon (non-sense). Mutations at the second position always result in a change (missense and non-sense). Third-position mutations are usually synonymous (“silent”), resulting in no change of amino acid. Nucleotide insertions and deletions will result in a shift of the “reading frame,” so that all amino acids after the mutation are changed.

Mutations occur at very low rates, and are usually measured as the number per meiosis in eukaryotes, or as the number per cell division for prokaryotes. The numbers of new mutations are commonly less than 1 in 10,000 (10^{-5}) for nucleotide substitutions and are as low as 1 in 100,000,000 (10^{-8}) for some visible mutations like coat colors. In an evolutionary sense, mutation is considered a very slow process. However, mutation still can produce considerable variation within populations because the numbers of new mutations produced in each generation in a population will be a function of population size. In addition, if the number of genes in a genome is taken into account, even low mutation rates can generate high numbers of new mutations in each individual within a population.

See also the following articles:

ADAPTATION • BACTERIAL GENETICS • DIVERSITY, MOLECULAR LEVEL • ECOLOGICAL GENETICS • SPECIATION, PROCESS OF

Bibliography


GENETIC DIVERSITY

Eviatar Nevo
University of Haifa

I. The Problem
II. Methodologies
III. Evidence: Genetic Diversity within and among Species
IV. Theory

GLOSSARY

electrophoresis Separating charged molecules (such as polypeptides or polynucleotides) between the two poles of an electric field.
geno\linebreak1me A complete single set of genes of an organism or organelle; also the basic haploid chromosome set.
genomics The study of the molecular organization of genomic DNA and physical mapping.
heterozygosity The average number of different heterozygotes across loci divided by all loci studied.
polymorphism Different alleles at a gene locus within a population.

GENETIC DIVERSITY (i.e., molecular hereditary differences within or between populations) is the basis of evolutionary change. The nature of genetic diversity among organisms has always been the basic problem of evolutionary genetics (Darwin, 1859; Lewontin, 1974; Kimura, 1983) as well as of domestication, agriculture, and medicine. However, despite its cardinal role in evolutionary theory and application, the maintenance of genetic diversity remains largely enigmatic, notwithstanding the dramatic discoveries of molecular biology, which revealed abundant genetic diversity in nature.

I. THE PROBLEM

A. Historical Background

The era of molecular evolution was ushered in by three major discoveries in the late 1950s and 1960s: (i) recognition by Markert and Moller in 1959 of molecular diversity of enzymes and their importance in genetics, physiology, development, and evolution; (ii) identification by Zuckerkandl and Pauling in 1965 of protein sequence variation between species and the resulting postulate of the molecular clock; and (iii) evaluation in 1966 of enzyme variation (isozymes and allozymes) in Drosophila by Lewontin and Hubby and Johnson et al. and in humans by Harris. Molecular biology permitted the characterization of genetic diversity among individuals, populations, and species—the three cornerstones of evolution. It did so first by unraveling relationships between genes and proteins (Lewontin, 1974). Second, it did so by elucidating, at the extranuclear and nuclear coding and noncoding DNA regions, the structure, expression, function, mechanism, and evolution of genes, intergenic spacers, and multigene families by employing recombinant DNA methodologies (Watson et al., 1987; Avise, 1994).

Advancements in molecular evolution have followed
the introduction of new laboratory techniques (Avise, 1994). Among the most influential methods have been protein electrophoresis in the late 1960s and 1970s, restriction fragment length polymorphism (RFLP) analyses of nuclear and mtDNA in the late 1970s and 1980s, DNA fingerprinting in the mid to late 1980s, and polymerase chain reaction (PCR)-mediated DNA sequencing in the 1990s. Comparative genomics and genome prospecting (Science 286, October 15, 1999) promise to highlight the genetic basis of evolutionary change and the nature, expression, meaning, transfer, and regulation of information in biological systems, thus unraveling the blueprint and evolutionary forces driving life.

B. The Enigma of the Maintenance of Genetic Diversity

Early and later molecular–genetic studies revealed massive levels of genomic diversity, surpassing any prediction. In addition, a comprehensive compilation of dense genetic maps was published of viruses, bacteria, plants, animals, and humans. In 1998, Deloukas et al. physically mapped 30,000 encoded human genes for proteins of known functions. The ongoing human genome project (HGP), the largest biological project ever (Bodmer and McKie, 1994), ushered in a new era in the life sciences in which complete genomes of prokaryotes (Casjens, 1998) and eukaryotes became available. Genetic maps, current complete genome sequences and those expected in the near future, and tens to hundreds of thousands of protein and DNA sequences have contributed dramatically to elucidating genome diversity, organization, expression, dynamics, and evolution (Kärlin et al., 1998). They form the basis for future studies of genetic diversity, when more genomes within species will be analyzed.

However, despite the ease of measuring and deciphering genetic diversity and the dramatic advances in comparative genomics, the evolutionary forces that generate and maintain segregating genetic diversity, preventing allele fixation or random elimination in nature, remain elusive.

C. Theories of Molecular Evolution: Selection versus Neutrality

Theoretically, balancing selection could account for protein polymorphism (Gillespie, 1991). In contrast, the neutral theory of molecular evolution (Kimura, 1983) suggests that most of the molecular–genetic diversity within and between species is neutral (i.e., non-selective) or "non-Darwinian." The neutralist–selectionist debate has been one of the major controversies in evolutionary biology since the late 1960s. How much of the genetic diversity at single and multilocus structures is adaptive, processed by natural selection and contributing to differences in fitness? The problem of distinguishing between deterministic and stochastic forces in evolution has pervaded evolutionary biology at all levels, genotypic and phenotypic, and is now focused on DNA polymorphisms. I recognize the contribution of the neutral and nearly neutral theories of molecular evolution, primarily by representing a null hypothesis to selection. Nevertheless, by ignoring the ecological heterogeneity and stress in evolution, neutral and nearly neutral theories have stripped genetic diversity from nature. I believe that in-depth understanding of genetic diversity in nature is intimately linked to the interface between ecology and genetics; hence, to ecological genomics and now to ecological genomics. I submit that only this essential interface can meaningfully highlight the dynamic evolution of genetic diversity in nature.

D. The Structure of This Review

This review comprises three unequal perspectives of molecular–genetic diversity of protein and DNA in nature: methodology, evidence, and theory. The main focus is on the huge amount of protein and DNA evidence from nature in diverse groups of organisms across phylogeny from bacteria to humans and across diminishing geographical scales (global, regional, and local). The evidence naturally leans heavily on results of genetic diversity in plants and animals obtained during the past 25 years at the Institute of Evolution, University of Haifa, Israel, that demonstrate the pervasiveness of ecological determinants in genetic differentiation. In theory too, I focus on theoretical results obtained during the past decade at the Institute of Evolution that show how selection in cyclical environments can maintain genetic polymorphism in nature, thus preventing drift. I mainly refer to books which cite the primary literature.

II. METHODOLOGIES

The methodologies analyzing protein and DNA diversity have been extensively described (Lewontin, 1974; Avise, 1994; M Itten, 1997), so they are only briefly reviewed here.
A. Protein Polymorphism

1. Technical Innovation

New techniques at the protein level permit an increasingly subtle resolution of allelic variation at a locus. These techniques include isoelectric focusing in various media: high-resolution, two-dimensional electrophoresis; thermal and urea denaturation analysis; and sequential gel electrophoresis (SAGE), in which electromorphic classes were retested with other pHs, buffer systems, and gel pore sizes (Ramshaw et al., 1979).

2. Hidden Variation

"Alleles" detected by routine gel electrophoresis are essentially phenotypes or "electromorphs" (i.e., internally heterogeneous, genetically involving "hidden variation"). For example, xanthine dehydrogenase has many more alleles than are visible routinely (Ramshaw et al., 1979). Recent pulsed-field gel electrophoresis for separating large DNA fragments, recombinant DNA techniques, methods of genomic "hopping," and new DNA sequencing strategies and genomic analysis reinforce the powerful field of molecular evolution, climaxing in the HGP (Bodmer and McKie, 1994). Extensive comparative analyses across genomes of model organisms (Karlin et al., 1998) became the focus of genomic and proteomic studies, highlighting diversity in nature (Avise, 1994). These new horizons partly unraveled the molecular structure, function, and evolution of life. Bioinformatics became increasingly important as many genes and genomes became analyzable. Comparative genomics, using dense genetic maps based on coding genes and on microsatellite and single nucleotide polymorphisms (SNPs), permit precise gene homolog alignment across taxa, the unraveling of gene function and regulation, highlighting of genome organization and evolution, and the determination of the genetic basis of speciation and adaptation.

B. DNA Polymorphism

1. RFLP, PCR, and Mini- and Microsatellites

Many assays reveal DNA diversity (Avise, 1994). Before the revolutionary PCR was established, DNA fingerprinting technology had been revolutionized by RFLP analysis, a very efficient technique but quite laborious and not suitable for high-throughput applications. Hundreds of restriction enzymes, originating in bacteria, cleave duplex DNA at particular oligonucleotide sequen-...
peat motifs), of which SNPs are the most common; (ii) variant detection in coding (cSNPs and noncoding DNA); and (iii) combined detection and genotyping using a single method.

5. Summary

DNA techniques revealing DNA diversity include RFLP, RAPD and AFLP, mini- and microsatellites, SNP and sequencing; all are relevant for analyzing DNA diversity. PCR is currently the method of choice for amplifying DNA segments for detecting polymorphism. Generally, DNA diversity is higher than protein diversity (see Fig. 4).

C. Comparative DNA Analysis across Diverse Genomes

The era of comparative genomics is dramatically advancing (Casjens, 1998; Karlin et al., 1998; Science, October 15, 1999), including the analysis of complete genomes for various organisms and imminent completion of the HGP. Genes, individuals, and species are becoming comparable. Sequence polymorphism will highlight the generation, maintenance, and function of genetic diversity responsible for controlling normal development, physiologic homeostasis, and disease processes. This information explosion will revolutionize molecular evolutionary studies. Comparative analysis of complete genomes includes assessments of genomic compositional contrasts based on di-, tri-, and tetranucleotides relative abundance values; identification of rare and frequent oligonucleotides; evaluations and interpretations of codon biases in several large prokaryotic genomes; and characterization of compositional asymmetry between the two DNA strands in certain bacterial genomes. Comparative analysis also allows identification of alien (e.g., laterally transferred) genes and detection of potential specialization islands in bacterial genomes and the assessment of DNA curvature. It can compare genomes within species and across life and cluster organisms according to linguistics based on many words in large DNA stretches.

D. Microarrays: Biotechnology’s Discovery Platform for Functional Genomics

Advances in microarray technology make possible massive parallel mining of biological data, based on PCR, with biological chips providing hybridization-based expression, monitoring, polymorphism detection, and genotyping on a genomic scale, as reported by Schena et al. (1998). Microarrays containing sequence represenative of all the genes of an organism may soon permit the expression analysis of the entire human genome in a single reaction. These “genome chips” will provide unprecedented access to genomic diversity of many thousands of genes to large-scale gene discovery as well as polymorphism screening and mapping of genomic DNA clones on a massive scale—critical for science and application in agriculture and medicine. Oligonucleotide microarray (DNA chip)-based hybridization analysis is a promising new technology which potentially allows rapid and cost-effective screens for all possible mutations and sequence variations in genomic DNA. Currently, it is performed in humans, mice, and Arabidopsis, but use in additional species has been reported. The increasing use of relatively inexpensive microarrays is expected to revolutionize genomic, proteomic, and other biological research projects.

E. Encyclopedia of Genes

Extensive ongoing programs in plants and animals are generating a large database of expressed sequence tags (ESTs) that can provide rapid access to numerous genes and their diversity. The development of a comprehensive database of ESTs for Arabidopsis, corn, soybean, rice, barley, wheat, mouse, and human has been reported. As of October 23, 1998, 352,040 sequences had been generated in the mouse and annotated and deposited in dbEST, in which they comprised, according to Marra et al. (1999), 93% of the total ESTs available for the mouse. EST data are versatile and have been applied to gene identification, comparative sequence analysis, comparative gene mapping and candidate disease gene identification, genome sequence annotation, microarray development, and the development of gene-based map resources. Large-scale exploration of the genetic diversity of the genome is becoming possible through the UniGene database (http://www.ncbi.nlm.nih.gov/Unigene).

III. EVIDENCE: GENETIC DIVERSITY WITHIN AND AMONG SPECIES

A. Protein Diversity

Electrophoretic results have revealed large amounts of genic polymorphism in natural populations (Lewontin, 1974; Nevo, 1978, 1988, 1998; Hamrick et al., 1979; Nevo et al., 1984; Avise, 1994; Mitton, 1997). This is true for both eukaryotes and prokaryotes. Currently, theory appears to lag behind evidence. The review and reanalysis of approximately 1100 plant and animal
species (Nevo et al., 1984) revealed an average heterozygosity, $H$, of 0.073 (SD 0.076) and an average proportion of loci polymorphic, $P$, of 0.284 (SD 0.197). Across species, the coefficient of correlation between $H$ and $P$ was $r = 0.793$, $p < 0.001$. These estimates are based on the average of 23 electromorphically detectable gene loci (minimum 14 loci) and an average of 199 individuals per species (minimum 10 individuals) (Fig. 1). This set of data was assembled up to 1983, and many hundreds of species have since been tested (see Fig. 2.2 in Avise, 1994, that summarizes heterozygosity estimates of 1803 species: 648 vertebrate, 370 invertebrates, and 783 plant species), all demonstrating diverse levels of polymorphisms (Mitton, 1997). Genetic diversity varies dramatically among species. Fruit flies, marine mussels, and conifers have much genetic diversity, whereas large vertebrates have much less (Mitton, 1997).

The aforementioned genetic indices appear to be lower than the real ones since routine horizontal gel electrophoresis underestimates diversity and protein coding genes represent only a small portion of the genome. Regarding levels of diversity, protein diversity does not represent the noncoding genome, which amounts to 90–98%, involving invariant and variant portions whose general diversity levels in natural populations appear to be high. Studies of “hidden variation” indicate that most protein diversity is detectable (Ramsaw et al., 1979), thus increasing $H$ of enzymatic loci. Moreover, at the DNA sequence level of individuals, most loci may prove heterozygous. Nevertheless, there are significant merits to the estimates of $H$ and $P$, derived from routine protein electrophoresis. First, although relative, low, and limited genomically, they are commensurate estimates of genetic diversity among numerous populations and species living in varied ecologies. Second, $H$ and $P$ are highly correlated in global (Nevo et al., 1984) and regional (Nevo, 1988, 1998) analyses. Thus, although hidden variation is regretfully lacking in studies of routine protein electrophoresis, in most species the high correlation of $H$ and $P$ makes for a profitable analysis since $P$ is largely independent of the high resolution achievable by SAGE and other new techniques. Third, the relatively large number of electromorphic protein loci currently available in *Drosophila* and the many species analyzed permit a sound statistical analysis of isozyme phenotypes in an ecological context rather than only in genetic terms (Nevo et al., 1984; Nevo, 1978, 1988, 1998).

1. The Adaptive Evolution of Enzyme Kinetic Diversity

Kinetic studies of hemoglobin, haptoglobin, and transferrin proteins and at least a dozen enzyme polymorphisms typically reveal biochemical kinetic differences among the gene products of alternative genotypes at a locus (Mitton, 1997). These include single gene effects such as lactate dehydrogenase in killifish, leucine aminopeptidase in blue mussel, and phosphoglucose isomerase (PGI) in *Collia* butterflies. In the latter, Watt

**FIGURE 1** Levels of genetic (allozymic) diversity among higher taxa based on a mean of 23 gene loci per species, in 1111 species. *Hatched region indicates parthenogenetic species (reproduced with permission from Nevo et al., 1984).*
and colleagues discovered in the 80s that PGI heterozygotes fly over a greater range of temperatures and produce 33% more eggs. Additional examples are alcohol dehydrogenase (ADH) in fruit flies, salamanders, and barley; glutamate pyruvate transaminase in copepods; and esterase, glucose-6-phosphate, 6-phosphogluconate dehydrogenase, and superoxide dismutase in fruit flies. In all these cases, biochemical–kinetic studies reveal differences among phenotypes, either within or between species, that have measurable effects of alternative genotypes on the physiology of whole individuals. They result in fitness differences distinguished by natural selection in accord with their alternative spatiotemporal environments (Mitton, 1997). Usually, the heterozygotes are intermediate between homozygotes, but sometimes molecular overdominance for enzyme kinetics is found.

2. Patterns of Variation among Loci
Proteins vary genetically. Regulatory enzymes appear to be more variable than nonregulatory enzymes, and enzymes that work on many substrates appear to be more variable than enzymes that utilize a single substrate. Genetic diversity tends to decrease from monomer to dimer to tetramer as the steric restrictions on the molecule increase. The number of alleles at a locus and the heterozygosity at a locus tend to increase with the subunit size of the protein.

3. Genetic Population Structure
Geographical clines may reflect the action of natural selection on genetic polymorphism at local, regional, and global scales. In Drosophila melanogaster, several latitudinal clines occur for many characters such as allozymes, inversions, and quantitative traits. The identical nature of these clines on the various continents, in both the Northern and Southern Hemispheres, strongly suggests adaptation to specific stress factors, primarily climatic selection. Polymorphism for stress-resistance genes abounds in natural populations. The ADH polymorphism shows high frequencies of the S allele in tropical regions and this declines with latitude. The reason for this cline are difficult to determine because of the entanglement with other polymorphisms varying with latitude. In 1977, Van Delden and Kamming reviewed the tentative connections with other polymorphisms such as α-GPDH, in(2L) inversion, body size, and development time with respect to the possible environmental stress factors involved. They concluded from these results, and also from recent experiments, that the (2L) inversion plays a dominant role in resistance to high temperature and is partly responsible for the ADH cline. They are currently studying the specific selective forces acting on ADH, focusing on the physiological and life history aspects. Many plant and animal species distributed in several climatic zones—tropical, temperate, and arctic—display a decline in gene diversity, both allozyme and DNA, towards the Arctic (Nevo et al., 1984). This is highlighted by postglacial colonizing populations. Regional and local clines also abound in nature.

4. Partition of Genetic Diversity
An important analysis is that of genetic partition within and between populations. In most analyzed outcrossers, genetic diversity is chiefly within populations (e.g., 85% in humans). In contrast, in some selfers (e.g., wild emmer wheat) only 40% of genetic diversity is within and 60% is between populations. This partition is shaped by both external (gene flow and selection at the population level) and internal (genomic and genic) evolutionary forces.

B. DNA Diversity
DNA technologies revolutionized population genetics by providing an unprecedented amount of genetic diversity for critical analysis and hypothesis testing. I briefly review these dramatic developments, from molecular (extranuclear, mitochondriol, chloroplast, and nuclear) DNA markers to sequence polymorphism.

1. MtDNA Diversity
MtDNA is largely maternally nonrecombiningly inherited and demonstrates a rapid pace of evolution and an extensive high level of intraspecific restriction. PCR and sequence polymorphism derive primarily from base substitution with a preponderance of transitions, length variation, sequence rearrangements, and duplications of coding sequences (Avise, 1994). Hence, its relevance is not only to large-scale phylogenies but also to microgeographic divergence, i.e., to intraspecific phylogeography, the recently established bridge between population genetics and systematics. The amount of mtDNA polymorphism within species and sequence heterogeneity is striking. Remarkably, a different scenario emerged for plant mtDNA from that in animals. Whereas animal mtDNA ranges in size from 14 to 26 kb in length, that of plants (e.g., maize) is approximately 30 times larger. Most polymorphism in plant mtDNA is attributable to major reorganizations of sequence. The level of mtDNA polymorphism in higher animals is several-fold higher than that of single-copy nuclear DNA. The maternal inheritance and high polymorphism of mtDNA and the assay of whole mtDNA by long PCR provide unique opportunities for population evolutionary studies of
animals and plants on both micro- and macrogeographic scales. They contribute to natural history, population structure and signature, gene flow, hybridization, biogeography, phylogeny, and biological conservation.

Mitochondrial genomes are increasingly being used to study ancient divergences among animal groups. Recent studies by Curd and Koder (1999) of complete mitochondrial DNA sequences and reached somewhat heretical conclusions, raising questions about the use of mitochondrial gene sequences for studying the relationships among highly divergent lineages. Other studies have documented convergent evolution of mitochondrial gene order, casting doubt on the use of these characters for phylogenetic analysis. The use of mitochondrial genomes for studying such deep divergences is coming under increased scrutiny, and these novel results need to be confirmed with data from nuclear genes.

2. Comparative Summary of Genetic Distances in the Vertebrates from the Mitochondrial Cytochrome b Gene

Mitochondrial cytochrome b is among the most extensively sequenced genes to date across the vertebrates. Johns and Avise (1998) employed approximately 2000 cytochrome b gene sequences from GenBank to calculate and compare levels of genetic distance between sister species, congeneric species, and con-specific genera within and across the major vertebrate taxonomic classes. The results of these analyses parallel and reinforce some of the principal trends in genetic distance estimates derived from multilocus allozymes. In particular, surveyed avian taxa on average show significantly less genetic divergence than do same-rank taxa surveyed in other vertebrate groups, notably amphibians and reptiles.

3. Chloroplastid Genome Comparisons

Chloroplast genome (cpDNA) evolution is also very dynamic, indicating gene transfer to the nucleus. Among the 210 different protein-coding genes contained in the completely sequenced chloroplast genomes from a glaucocystophyte, a rhodophyte, a dia-tom, an euglenophyte, and five land plants, Martin and colleagues identified in 1998 the set of 45 genes common to each and to a cyanobacterial outgroup genome. Phylogenetic influence, with an alignment of 11,039 amino acid positions per genome, surprisingly indicates that independent parallel gene losses in multiple lineages outnumber phylogenetically unique losses by more than four to one. They identified homologs of 44 different plastid-encoded proteins as functional nuclear genes of plastid origin, providing evidence for endosymbiotic gene transfer to the nucleus in plants. cpDNA diversity in lodgepole pine was high within populations, with little (>5%) differentiation among populations, in contrast to the mtDNA pattern as shown by Hamrick and Godt in 1990 (as cited in Mitton, 1997, p. 71). Extensive intraspecific cpDNA diversity may sometimes exceed interspecific diversity, reflecting dynamic ecologies and seriously affecting phylogenetic conclusions, as in the pine Draba species.

4. Simple Sequence Repeats (SSRs)

SSRs consist of tandem repeats of relatively short nucleotide motifs, such as TCCTCCTCCTCC. Microsatellite repeat number can range from two (T2), or three (GA3), to a few dozen (GCAA11), whereas minisatellites often consist of many dozens or even hundreds of repeated motifs. Significantly, SSRs experience mutation at notably higher rates than do nonrepetitive sequences: 10^-4 to 10^-5 per locus, per gamete, per generation, which leads to their high polymorphism. Replication slippage, sister chromatid exchange, unequal crossing-over, and gene conversion may cause microsatellite diversity. Replication slippage seems to play a major role in producing new alleles at microsatellite loci (Goldstein and Schlotterer, 1999). SSRs are characterized by high sitespecific and reversible rates of gain or loss in the number of tandem repetitions of a short DNA motif. Importantly, recent critical population genetics and specific chromosome data derived from Li and colleagues in 1999 at our laboratory indicate massive genomic nonrandom chromosomal and environmental distributions of microsatellites, suggesting that they are subjected to natural selection, generating adaptive complexes in contrast to neutral theory expectations.

The functional properties of SSRs suggest evolution's effect on mutability, as recently concluded by King and Soller (as cited in Wasser, 1999). Many SSRs are functionally integrated into the genome and exert a quantitative regulatory effect on gene transcription activity affecting phenotype and fitness. Genes associated with SSRs may be favored by indirect selection whenever quantitative variation in the affected traits can provide a population with genetic resilience for adaptation, especially in stressful, fluctuating, or heterogeneous environments. Such "adjustable genes" may provide a prolific and evolutionarily significant source of quantitative genetic variation.

Nucleotide diversity of the human nuclear genome has been estimated to be approximately 0.1%. For two randomly selected sequences, this number translates into one polymorphic site for 1000 nucleotides (or,
in a large sequence sample, one polymorphic site is expected for every 200–500). Since the human nuclear genome contains approximately 3 billion nucleotides, several million polymorphic sites are expected to exist.

5. Single Nucleotide Polymorphism
SNPs are the most frequently found DNA sequence variations in animal and plant genomes, usually followed by SSRs, RAPDs, and allozymes. Recent SNP surveys in humans reported different rates of polymorphism among classes of sites within genes (noncoding, degenerate, and nondegenerate) as well as between genes. Of all coding SNPs, 54% lead to predicted change in the protein sequence. As expected, the coding SNPs that alter amino acid sequence of the encoded protein are found at a lower rate and with lower allele frequencies than silent substitutions. This was interpreted as a reflection of selection against deleterious alleles during human evolution. Determination of ancestral alleles from human SNP polymorphisms became available using high-density oligonucleotide arrays. A densely packed map of human SNP sites could efficiently identify disease-associated genes by linkage disequilibrium between sets of adjacent markers and highlight human history. In plants, SNP polymorphism has been associated with transcript efficiency, nonconsensus splice sites, and gene expression. The following sections discuss the results of genetic diversity and divergence at global, regional, and local scales based primarily on results derived at the Institute of Evolution since 1979.

C. Global Analysis of Genetic Polymorphisms

1. Global Allozyme Diversity across Phylogeny
We used the entire globe as a large-scale ecological genetic laboratory. We analyzed the correlates of biotic factors involving ecological, demographic, and life history variables with the level of genetic diversity in natural populations of animals and plants (Nevo et al., 1984). This review involved 1111 species studied for allozymic variation with an average of 23 gene loci in each species and a biotic profile characterized by 21 variables (7 ecological, 5 demographic, and 9 life history and other biological characteristics). We then (i) estimated the levels of genetic diversity, indexed by heterozygosity and polymorphism, for all species, three major taxa (vertebrates, invertebrates, and plants), 10 different higher taxa, and the categorized 21 biotic factors; (ii) correlated the levels of genetic diversity with the biotic factors; and (iii) matched some of the evidence obtained with theoretical predictions.

The following results were obtained:

1. The levels of genetic diversity vary nonrandomly among populations, species, and higher taxa and also among ecological parameters (life zone, geographical range, habitat type and range, and climatic region) (Fig. 2), demographic parameters (species size and population structure, gene flow, and sociality), and life history characteristics (longevity, generation length, fecundity, origin, and parameters related to the mating system and mode of reproduction).

2. Generally, genetic diversity is higher (i) in species living in broader environmental spectra; (ii) in large species with a patchy population structure and limited migration as well as in solitary or social species; (iii) in species with small body size, annuals or long-lived perennials, older in time, with smaller diploid chromosome numbers; and (iv) in plant species primarily outcrossed, reproducing sexually, and pollinated by wind. Diversity among species has also been summarized by Mitton (1997). Higher genetic diversity characterizes species with high fecundity, large populations, broad geographic ranges, and a high speciation rate. In contrast, low genetic diversity characterizes endemic species with low fecundity and low speciation rate.

3. Genetic diversity is partly correlated and predictable by a three- or four-variable combination of ecological, demographic, and life history variables, largely in that order. Ecological factors account for the highest proportion of the 20% explained genetic variance of all species compared with demographic and life history factors (90, 39, and 3.5%, respectively). Within individual higher taxa, the explained portion of genetic diversity increases considerably (mean of 44% and maximum of 74% in mollusks). Neutrality theory would be substantiated if demographic variables, rather than ecological variables, could better explain the variation in genetic diversity. The opposite, however, was often found, substantiating selection as an evolutionary driving force of genetic diversity in nature. A more detailed global analysis of genetic differentiation in amphibians was performed in 1991 by Nevo and Beales and in small mammals across the globe (Nevo, 1999). In both cases, genetic diversity, and in amphibians also genome size, is predictable by ecological factors: The higher the environmental heterogeneity and stress, the higher the heterozygosity. Similar conclusions were drawn for 480 species of plants in 1990 by Hamrick and Godt and for
diverse animal species (Mitton, 1997). As expected, theoretically the global evidence exhibits a positive relationship between environmental diversity and genetic diversity, as was also shown in cage experiments. Clearly, environmental diversity enhances genetic diversity. Levins (1968) predicted that environmental grain determines the level of polymorphism. Fine-grained species (highly mobile) would evolve a monomorphic strategy, whereas coarse-grained species (highly sedentary) would maintain polymorphism. This theoretical prediction is largely supported by evidence (Nevo et al., 1984).

The patterns and correlates of genetic diversity revealed in the global analysis for many unrelated species, subdivided into different abiotic and biotic regimes, strongly implicate selection in the genetic differentiation of species. Natural selection in several forms, but most likely through the mechanisms of spatiotemporally varying environments at the various life cycle stages of organisms, appears to be an important evolutionary force causing change at the molecular level. Other evolutionary forces, including mutation, migration, and genetic drift, certainly interact with natural selection, either directly or indirectly, and thereby con-

![Figure 2](image)

**FIGURE 2** Global analysis of genetic diversity: levels of heterozygosity of ecological factors. Numbers in the bars indicate species. Significance ns = $p > 0.05$; * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$ (reproduced with permission from Nevo et al., 1984).
tribute differentially, according to circumstances, to population genetic differentiation at the molecular level. However, the final orientation of the evolutionary process is determined by natural selection (Bell, 1997).

D. Regional Analysis of Genetic Polymorphisms

1. Regional: Allozyme Diversity across Phylogeny in Israel and the Near East

a. Israel as an Ecological Genetic Laboratory of Increasing Aridity Southwards

Our regional analyses of genetic diversity extended over Israel and the Near East. We used Israel, with its remarkable physical and biotic diversities, as a medium
to large-scale ecological genetic laboratory (Nevo, 1988, 1998). In 1988, Nevo and Beiles conducted an ecological test of protein polymorphism in 13 unrelated taxa of plants, invertebrates, and vertebrates involving 21 species, 142 populations, and 5474 individuals (Fig. 3), following the extensive studies of allozyme diversity in 38 species in Israel.

Each individual, population, and species were tested, on average, for 27 enzymatic gene loci. These species varied in population size and structure, life histories, and biogeographical origins, but they largely shared geographically short and ecologically stressful gradients of increasing aridity in Israel, both eastward (70 km) toward the Syrian and Jordanian deserts and (mainly) southward (260 km) towards the Negev desert. We found genetic parallelism across most taxa and most loci. Observed average heterozygosity, \( H \), and gene di-

![Figure 3 - Parallel genetic patterns in the level of gene diversity, \( H_i \), of 13 enzymatic systems averaged across all taxa studied here. The average regression line is \( H = -0.0556 + 0.4603 \cdot RV \). The only change is that Gryllotalpa gryllotalpa has been divided into two species, G. tali \( (2n = 19) \) and G. marismortui \( (2n = 23) \). See Broza et al., 1998, cited in Nevo, 1999. The regression line represents G. tali (reproduced with permission from Nevo and Beiles 1988).](image-url)
versity, \( H \), were positively and overall significantly correlated with rainfall variation, which is lowest in the northwest and highest in southern deserts (i.e., with a wider, fluctuating, climatic temporal niche in the desert) (Fig. 3). Similar trends were found for several DNA systems in subterranean mole rats in Israel (Nevo, 1999). This result corroborates the environmental theory of genetic diversity, which regards ecological heterogeneity and stress as major determinants enhancing genetic divergence.

Notably, in the regional Israeli studies, heterozygosity, \( N \); is negatively correlated with effective population size, \( N_e \); \( H \) increases, whereas \( N \) declines drastically toward the deserts. This contradicts a basic postulate of neutrality theory, which predicts positive correlation between \( H \) and \( N \), (Kimura, 1983). Our results are clearly inconsistent with the genetic drift postulated in 1931 by Sewall Wright or the neutral theory of molecular evolution, even in its milder form of near neutrality (Ohta and Gillespie, 1996). Our results suggest that natural selection, through environmental range and stress, in space or time or both appears to be an important genetic differentiating evolutionary force at the protein and DNA levels (i.e., driving molecular evolution) (Nevo, 1998). Similar results were obtained in extensive ecological genetic studies in wild wheat in 1989 by Nevo and Beiles and in subterranean mole rats across the Near East and Asia Minor (Nevo, 1999). Genetic diversity either in the inbreeding wild cereals above ground or in outbreeding mole rats underground is primarily determined by climatic selection. Similar conclusions—that is, that natural selection on allozyme polymorphism may be intense with selection coefficients in the range of 0.1–0.8—have also been reached for many model organisms (Mitton, 1997). Remarkably, the abundant genetic diversity found in natural populations of plants and animals provides immense genetic resources for crop improvement in agriculture, biotechnology, and gene therapy in medicine.

b. Genetic Parallelism and Ecological Stress in Subterranean Mammals

Parallel genetic patterns of increasing diversities (primarily southward and secondarily eastward) were shown in 1996 by Nevo and colleagues in Israeli subterranean mole rats for diploid chromosome numbers, nuclear allozyme, mRNA, RAPDs, minisatellites, chiasma frequency, and organismal traits and also for approximately 200 genes by AFLP (in preparation).

Clearly, genomic molecular evolution of nuclear and extranuclear diversities of proteins and DNA, coding and noncoding sequences, in subterranean mole rats, generally assumed to be neutral by proponents of the neutral theory of molecular evolution, is nonrandom and correlated positively with increasing aridity stress and climatic unpredictability. The increase in heterozygosity is primarily toward the southern and eastern deserts, where temporal stressful climatic fluctuations and unpredictability climax. Adaptive climatic selection appears to be the prime driving force in both molecular and organismal evolution of mole rats. Natural selection through fluctuating environmental stress (i.e., wider climatic niche) appears to be a major cause of molecular evolution and genome organization in subterranean mole rats, similar to the pattern found in aboveground organisms facing stressful unpredictable climates, across the same transect of increasing aridity stress (Nevo, 1999).

E. Local Population Genetics Studies across Sharp Ecological Contrasts

1. Past Microgeographic Studies in Plants and Animals

Microgeographical studies in nature, particularly those involving sharp contrasts (climatic, thermal, edaphic, geologic, topographic, and chemical), provide remarkable long-term experiments, representing small-scale natural laboratories, for analyzing ecologic genetic population dynamics. Since 1975, we have conducted several microgeographic studies in nature (Nevo, 1998). We employed thermal (high vs low temperature) and chemical (polluted vs clean areas) stresses in marine balanids and conducted an extensive research program on inorganic and organic pollution and its effect on allozyme diversity in marine organisms. Likewise, we examined edaphic (terra rossa vs basalt soil and rock vs deep soil) and microclimatic (sun vs shade and high vs low solar radiation) stresses in wild barley (*Hordeum spontaneum*), wild emmer wheat (*Triticum dicoccoides*), and *Aegilops peregrina*. The conclusions of these studies all point inferentially to natural selection as a major differentiating factor of qualitative and quantitative patterns of genetic diversity at single loci, but primarily at multilocus structures and genome organization (Nevo, 1998, 1999).

The following is the summary of our extensive studies in wild cereals (wild barley and wild emmer wheat) that have been studied spatiotemporally since 1975 at the Institute of Evolution at the protein and DNA levels on micro- and macrogeographical scales. We found significant climatic and soil divergence at four microsites as well as in our macrogeographic studies in Israel and the Near East (Nevo, 1998). All the foregoing studies illustrate massive genetic nonrandom divergence at sin-
In 1992, we embarked upon a long-term multidisciplinary project at the "Evolution Canyon" microsite. It aims to elucidate the causes of biodiversity and genome evolution at the molecular and organismal levels in diverse groups of organisms across phylogeny from cyanobacteria to mammals, representing a microcosm of life.

2. Evolution in Action across Phylogeny
Caused by Climatic Stresses:
"Evolution Canyon"

Biodiversity and genome evolution, and the relative importance of forces driving evolution, are critically tested at the "Evolution Canyon" microsite (Lower Nahal Oren, Mount Carmel), a dynamic ongoing microscale research program (Nevo, 1997). Our aim is to draw generalizations across life in genotypes and phenotypes and to highlight controversial and unresolved problems of biological evolution. The opposed slopes of "Evolution Canyon" display dramatic physical and...
FIGURE 5 Discriminant analysis between two soil types, basalt and terra rossa of wild emmer wheat (*Triticum dicoccoides*), across three microsites (Yehudiyya, Tabigha, and Ammiad) north of the Lake of Galilee, Israel. This analysis is based on 2, 5, 10 and microsatellite loci, classifying correctly 21, 83, 88, 90, and 91% of plants into their original soil types (based on data from Li et al., 2000b).
biotic contrasts on a microscale (Fig. 6). Higher solar radiation (up to 600% more) on the south-facing slope (SFS) makes it warmer, drier, and spatiotemporally more heterogeneous and fluctuating than the north-facing slope (NFS). The two slopes are separated by only 100 m (at bottom) and 400 m (at top). The SFS represents an Afroasian savanna park forest, whereas the NFS represents a dense Euroasian live oak maquis forest.

The tropical Afroasian warm and xeric SFS savanna displays wider ecological heterogeneity and higher stress for temperate terrestrial organisms. As predicted, genetic diversity was higher on the ecologically more heterogeneous (wider niche) and stressful SFS in 9 of 11 tested temperate model species (lichen, wild barley, 2 species of land snails, earthworm, diplopod, 3 beetles, and 2 rodents) (Fig. 7). In 6 species, heterozygosity was negatively correlated with population size, i.e., population abundance was lower, whereas genetic diversity was higher on the SFS. This negates the positive correlation expected between $H$ and $N$, by neutrality theory (Kimura, 1983). We have shown in wild barley, $H. spontaneum$, that RAPD and sequence-tagged site PCR analysis mirror allozyme interslope patterns.

Remarkably, heritable mutation and recombination rates were several-fold higher in the soil fungus $Sordaria fimicola$ and in the fruit fly $D. melanogaster$ on the more stressful SFS. Importantly, in terrestrial taxa species

**FIGURE 6** "Evolution Canyon," Lower Nahal Oren, Mount Carmel, Israel. Note the plant formation on opposed slopes. The lush, green "European" live oak maquis forest on the temperate, cool-mesic, north-facing slope sharply contrasts with the open park forest of warm-xeric, tropical "Afroasian" savanna on the south-facing slope (SFS). (Top) Cross section; (bottom) air view with the seven experimental stations—three on the SFS (1–3), one at the valley bottom (4), and three on the NFS (5–7) (reproduced with permission from Nevo, 1997).
G. Evolutionary Significance of Molecular Polymorphisms: Summary and Evidence

The foregoing macro- and microscale field and laboratory evidence across phylogeny displays massive genetic parallelisms to environmental heterogeneity and diverse stresses, physical (thermal, climatic, and edaphic) and biotic (pathogens, competitors, vegetation, and habitats). In general, higher levels of genetic polymorphisms, at both the protein and the DNA levels, occur under more stressful and variable environments. This is highlighted by the enormous diversity recorded in the vertebrate major histocompatibility complex (Bodmer and Bodmer as cited in Wasser, 1999) and immunoglob-
ulin gene families, resistance genes in plants, against the immense pathogen diversity, and in the regulation of clock genes. Linkage disequilibria and genome organization (Nevo, 1998) are correlated with higher environmental heterogeneity (also called the niche width variation hypothesis by Van Valen, which predicts positive correlation between niche width and morphologic variation) and stress (Figs. 3 and 7). Similarly, mutation and recombination rates in the soil fungus Sordaria fimicola and in male Drosophila melanogaster at “Evolution Canyon” were both higher on the ecologically more heterogeneous and stressful SFS. Ecological heterogeneity and stress appear to be major determinants of the level of genetic diversity in nature, as was demonstrated in numerous organisms (Mitton, 1997). Physical and biotic ecological variables are better predictors of genetic diversity than demographic variables of population and species size. In fact, the levels of genetic diversity are often negatively correlated with effective population size, negating neutrality theory predictions. Generally, genetic polymorphism is positively correlated with niche width, or the level of ecological heterogeneity and stress, as emerged from our global, regional, and local analyses and from critical laboratory experiments. A brief theoretical discussion follows.

IV. THEORY

A. Mechanisms of Generating Genetic Diversity

It is beyond the scope of this article to describe the mutation mechanisms that generate diversity. Suffice it to say that a host of mechanisms provide a permanent input of diversity into natural populations, including point mutations, deletions, additions, recombination, gene conversion, and mismatch repair. Mutation rates also vary dramatically among organisms and genes and under diverse ecological stresses. Evidence from bacterial genetics (Casjens, 1998) suggests that organisms can resist stress through mutational changes or acquisition of pre-evolved functions via horizontal transfer. Activating mutagenic response and inhibiting antimutagenic activities (e.g., mismatch repair) could be important in recruiting genetic diversity as an adaptation to stress. Stress-induced increases in mutation rates enhance genetic polymorphism, primarily in genomic “hot spots.”

B. Significance of Genetic Diversity

The functional value of genetic diversity within populations has been confirmed in a series of studies. Clearly, this rich pool of diversity provides the resource for continuous selection of adapted genotypes. Reduced genetic diversity may not only compromise the capacity of an impacted population for genetic adaptation in the face of further environmental challenge but also may result in increased energy requirements, lower production efficiency, and reduced homeostasis and reproductive output. These metabolic consequences of reduced genetic polymorphism would further lower that population’s potential for survival under lethal conditions of contaminant exposure and also affect the genetic makeup of populations through differential reproduction under conditions of sublethal stress. In laboratory studies with population cages, higher levels of allozyme and additive genetic diversity are generally maintained in cages with greater heterogeneity, as is also generally true in nature (Mitton, 1997). Genetic diversity makes it possible to establish genetic distances within and between species, identify species and strains by unique and specific genetic profiles, establish phylogenetic trees, reinforce conservation biology and management programs, and have profound implications for plant and animal husbandry in agriculture and medicine.

C. Maintenance of Genetic Diversity in Nature

Explaining the maintenance of genetic diversity in natural populations has been a central problem of evolutionary genetics since the discovery of abundant protein polymorphisms in nature (Lewontin, 1974; Nevo, 1978, 1988, 1998; Kimura, 1983; Gillespie, 1991; Avise, 1994; Mitton, 1997). It is now clear that stochastic and bottleneck explanations, relating heterozygosity primarily to population size effects, are not realistic. Populations are rarely at equilibrium because environmental conditions fluctuate with abiotic (e.g., climate and seasons) and a host of biotic (e.g., parasites, pathogens, and competitors) factors. Correspondingly, genetic diversity can be gained or lost rapidly in dynamically changing natural populations due to ecological stresses. The major models explaining genetic diversity in nature relate primarily to the levels of ecological heterogeneity (niche breadth) and stress, or multilocus, and to environmental grain (Levins, 1968) rather than to effective population size or gene flow. Environmental variability enhances genetic diversity. Natural selection appears to be a major driving evolutionary force maintaining genetic diversity in nature (Bell, 1997). However, heterosis (i.e., overdominance or heterozygote advantage) alone is not a mechanism for maintaining many alleles segregating at a locus, as argued in 1978 by Lewontin, Ginzburg,
and Tulipurkar; hence the resort to spatiotemporal driving forces at multilocus structures.

**D. Natural Selection**

Theoretically, spatial and temporal variations of selection ("diversifying selection") could maintain and enhance genetic polymorphisms, although the conditions of their applicability are strongly limited [see cited papers by Levene, Haldane, Karlin, Lande, Feldstein, Hedrick, Hoekstra, Maynard Smith in Ewens (1979) and Nevo (1998)]. Spatial variation appeared more effective than temporal variation, although their joint action could reinforce the maintenance of polymorphism. Most results related to selection of variation in time were derived from the one-locus case. Polymorphism maintenance may be reinforced in the case of two-locus or multilocus structures (Gillespie, 1991; Kirzhner et al., 1998). Fitness components (viability, growth rate, fecundity, mating success, and developmental stability) increase with heterozygosity, suggesting that selection balances diversity at protein polymorphisms (Mitton, 1997). Some of the correlations between heterozygosity and fitness components can be attributed to allozyme loci or loci in strong linkage disequilibrium with them. This positive relationship between heterozygosity and fitness is not only expected theoretically but also abundantly documented empirically (Mitton, 1997). The selective mechanism is much more effective in promoting genetic diversity if carriers of the alternative alleles are able to select the niche in which their fitness is greatest as argued in 1977 by Taylor and Powell and by Nevo and colleagues in 2000 for mole crickets. For selection theories of molecular evolution and polymorphism based on the combined forces of natural selection, genetic drift, and mutation, see Gillespie (1991). For a discussion on the development of neutral and nearly neutral theories, see Ohta and Gillespie (1996). A series of articles dealing with natural selection and Darwinian fitness in nature appear in Wasser (1999).

In a series of articles and books during the period 1931–1978, Sewall Wright developed his "adaptive landscape" theory. In this model, populations are imagined to spend most of their time on selective peaks, with genetic drift providing the push when a population jumps from one peak to another. In contrast, Fisher's (1958) model on the nature of adaptation assumes that a population is never exactly at its optimal phenotype. Likewise, in Gillespie's (1991, pp. 291–305) random environment selection theories, there are no analogs to adaptive peaks because the adaptive landscape is changing faster than the genetic system: "The population is always running uphill, but the peak is always two steps ahead. All the population ever sees, in effect, is the side of the mountain. Should it stop evolving, it will face extinction." This view is related to Fisher's model and to the red queen hypothesis of Van Valen (1973), which assumes that if a population does not continue to adapt at the same rate as its competitors or environmental deterioration it may become extinct.

**E. Stabilizing Selection in Cyclical Environments**

Stabilizing selection for an intermediate optimum is generally considered to rapidly deplete genetic diversity in quantitative traits, with increased number of loci. In contrast to previous conclusions, we found in both haploids and diploids, in the case of an additive two-locus model, that stabilizing selection with cyclical moving optimum may be an efficient factor in protecting polymorphisms for linked loci additively affecting the selected trait (Korol et al., 1994; Kirzhner et al., 1998). We proved that within the same class of fitness functions, nonequal gene action and/or dominance effect for one or both loci may lead to local polymorphism stability with substantial polymorphism-attracting domain. A higher intensity of selection could result in two forms of polymorphic limiting behavior: (i) the usually expected forced cycles with a period equal to that of environmental changes and (ii) "supercycles," which are nondamping autooscillations with a period composed of hundreds of forced oscillations.

We have demonstrated (Kirzhner et al., 1998) that a multilocus system subjected to stabilizing selection with cyclical moving optimum can generate ubiquitous complex limiting behavior, including supercycles, T cycles, and chaotic-like phenomena. This mode of multilocus dynamics far exceeds the potential of complex dynamics attainable under ordinary selection models resulting in simple behavior. It may represent a novel evolutionary mechanism increasing genetic polymorphism over long-term periods (Kirzhner et al., 1998).

**F. Selection versus Genetic Drift in Small Isolated Populations**

Remarkably, our findings of high-level heterozygosity in small (several dozen or 100 individuals) desert isolates of the subterranean mole rats, *Spalax ehrenbergi* superspecies (2n = 60), in the northern Negev (Nevo,
contradict both the Wrightian notion of genetic drift in small populations that was elaborated in 1931 and the extreme version of genetic drift, i.e., the neutral theory of molecular evolution (Kimura, 1983). The latter predicts positive correlation between effective population size \( N_e \) and heterozygosity \( H \), whereas our finding demonstrated the opposite (i.e., the highest level of \( H \) in the smallest populations, or a negative correlation between \( H \) and \( N_e \)). A negative correlation between population size and heterozygosity was also found in several species tested at the “Evolution Canyon” microsite (Nevo, 1997). Current theoretical models predict fast gene fixation in small panmictic populations without selection, mutation, or gene inflow. Using simple multilocus models, we demonstrated that moderate stabilizing selection (with stable or fluctuating optimum) for traits controlled by additive genes could oppose random fixation in such isolates during thousands of generations (Nevo et al. 1997a, as cited in Nevo, 1999).

We also showed that in selection-free models polymorphism persists only for a few hundred generations even under high mutation rates. Our multichromosome models challenge the hitchhiking hypothesis of polymorphism maintenance for many neutral loci due to close linkage with a few selected loci.

G. Genetic Ecological Diversity and Stress

The foregoing discussion suggests several explanations for the maintenance of genetic diversity subjected to ecological diversity and environmental stress. Spatial and temporal ecological variation, which predominates in nature, is of prime importance in maintaining and enhancing genetic diversity in natural populations. This may be true because different genotypes display varying fitnesses in variable environments and stresses. Recombination frequencies and mutation rates tend to increase under stressful conditions (Hoffmann and Parsons, 1991; Korol et al., 1994; Nevo, 1997). Rates of evolutionary change are therefore enhanced in adverse environments, as was demonstrated under controlled laboratory experiments in the case of mercury pollution (Baker et al. as cited in Nevo, 1998), under regional aridity stress across the physically stressful Israeli environment, and locally at “Evolution Canyon” because of high solar radiation on the SFS (Nevo, 1997).

Developmental instability may be enhanced under both environmental stress and genomic stress. This could increase genetic diversity under stressful conditions (Hoffmann and Parsons, 1991). Heterozygote advantage tends to increase diversity with stress up to extreme points. However, heterosis alone is not a mechanism for maintaining many alleles segregating at a locus. It is much more likely that stable equilibria for multiple alleles will be best explained by multiple-niche selection. Models of sexual reproduction as an adaptation to resist parasites proposed by Hamilton and colleagues in 1990 may also contribute to sex evolution, recombination, and polymorphism. Finally, the simple model advanced by Kirzhner and colleagues in 1999 of genetic interaction between multiple species on a trait for trait basis governed by abiotic and biotic selection for multilocus quantitative traits opens wide horizons for the evolution of genetic diversity due to species dynamic interactions in nature.

Ecological heterogeneity and stress appear to enhance genetic polymorphisms, particularly in dynamically cycling environments (which can generate complex dynamic-like supercycles and chaotic-like behavior). This mode of multilocus dynamics far exceeds the potential for maintaining genetic polymorphism attainable in ordinary selection models. It may represent a novel evolutionary mechanism increasing genetic polymorphism over long-term periods. This novel mechanism could contribute to the observation that biological diversity has increased over geological time despite the well-known massive extinctions, providing ever-increasing genetic diversity and thus enhancing the evolution of biodiversity.

Acknowledgments

I thank Abraham Korol and Avigdor Belles for critically reading and commenting on the manuscript. This study was supported by the Israel Discount Bank Chair of Evolutionary Biology and the Ancell-Teicher Research Foundation for Genetics and Molecular Evolution.

See Also the Following Articles

ECOLOGICAL GENETICS • EVOLUTION, THEORY OF • GENES, DESCRIPTION OF • NUCLEIC ACID BIODIVERSITY • PHENOTYPE, A HISTORICAL PERSPECTIVE

Bibliography


GEOLGIC TIME, HISTORY OF BIODIVERSITY IN

James W. Valentine
University of California, Berkeley

I. Biodiversity in the Fossil Record
II. Evolution and Biodiversity
III. Capturing Fossil Biodiversities
IV. Microbial Diversity of the Archaean and Proterozoic Eras
V. Phanerzoic Marine Biodiversity
VI. Phanerzoic Terrestrial Biodiversity
VII. Principal Biodiversity Factors

GLOSSARY

alpha diversity Diversity of species within a single habitat.
beta diversity Measure of the rate and extent of change in species along a given habitat or physiographic gradient.
disparity Range of diverse morphological architectures present in higher taxa such as classes, phyla, and kingdoms; it is used to explain the origins and maintenance of this diversity of life-forms and body plans.
taxonomic richness Number of species, genera, and families from a given time period or fossil excavation.

LEARNING THE PATTERNS OF PAST BIODIVERSITY CHANGE is fundamental to understanding the causes and consequences of processes that generate and de-

complete the diversity of life on earth. Although the fossil record is incomplete and biased, methods have been developed to correct for these problems in large measure, thus permitting the study of biodiversities of the geologic past.

I. BIODIVERSITY IN THE FOSSIL RECORD

There are two distinctive though related aspects to paleobiodiversity studies. One approach focuses on the number of taxa present during a certain geologic time, or over a certain geologic time span. The taxonomic categories involved are usually species, genera, or families, and this sort of diversity is commonly termed taxonomic richness. The ultimate aim of most such studies is to understand and explain the origin and maintenance of patterns of species richness in space and time; when genera or families are used it is usually as proxies for species. The evolution of species richness has been studied since Darwin’s time and continues to be a highly productive area of research. The second aspect deals with diversity, not of numbers of species, but of different types of organisms. The taxonomic categories involved are the higher ones, such as kingdoms, phyla, and classes, and the aim is to understand and explain the origins and maintenance of disparity—the range of diverse morphological architectures that is present. The major patterns of disparity over geologic time have been
described, chiefly using taxonomic categories, as proxies for morphological distance. However, the evolution of that disparity has received much less attention than the evolution of richness, partly because it was thought that disparity could be explained by extrapolation from the processes that create richness. With the rise of molecular developmental studies, however, the genetic basis for disparity is becoming understood, and it has become possible to decouple studies of the origin of disparity from those of richness.

The ecological regulation of local biodiversity is not well understood, but empirical patterns of species distributions and richnesses indicate some of the factors governing species accommodation on regional and global scales. These large scales are of particular importance to tracing biodiversity through time. A basic measure of species richness is alpha diversity, the number of species found at a given locality or in a single sample. A second sort of species richness, beta diversity, is a measure of the additional species that are found at a second locality or in a second sample. Alpha diversity was taken to indicate the amount of species packing (reflecting ecological niche partitioning) at a locality, and beta diversity to indicate the extent of habitat differences between localities. In practice, however, different species are found at a given locality at different times, varying with season, weather, or other factors that affect successful reproduction or colonization at the locality. Alpha diversity can therefore vary significantly from time to time, in which case beta diversity is affected as well.

The global environment is quite heterogeneous, a mosaic of habitats separated by gentle gradients in some places and by sharp boundaries in others and further broken into disjunct regions, such as different continental masses or remote islands. Beta diversity will vary according to whether two samples are from adjoining areas within the same general habitat type (i.e., represent the same biotic community), or are from distinctive habitat types (represent different communities), or are from distinctive climatic or physically disjunct regions (represent different biotic provinces). Though beta diversity is sensitive to differences among samples within a community, it is not as useful a measure between communities; rather, a more appropriate measure is the number of species added as entire communities are combined—a sort of super beta diversity. To capture global biodiversity patterns, a hierarchy of beta-type measures is required; differences between bioprovinces can be evaluated by the numbers of species added when combining one entire bioprovince with the next, or similar evaluations can be made between entire realms. The pattern of environmental heterogeneity in the world has varied continuously over geologic time, thus affecting the pattern of accommodation of biodiversity.

II. EVOLUTION AND BIODIVERSITY

A. Richness

Speciation is produced by isolation of populations, severing the flow of genes between them and permitting the evolution of differences in their features, underlain by differences in genes, gene frequencies, and gene organizations that arise between the populations. Once sister population crosses are unable to produce fertile offspring, speciation is complete. Environments favoring more isolation among populations are therefore more likely to produce rich biotas, other things being equal. A world broken into many continents favors global species richness, with a nearly separate biota on each continent. A continental landscape consisting of a great variety of climates and habitat types favors more species richness than does a monotonous landscape, for species tend to become well adapted to particular habitats, and the more habitat types available, the more species. A landscape in which any given habitat type is scattered among other types, which therefore provide barriers to gene flow, favors more species richness than a landscape where each habitat type occurs in a large tract. In short, environmental heterogeneity clearly promotes isolation and therefore species richness through speciation. Other factors also promote speciation, but they are not so clear nor so well understood. For example, low-latitude, tropical regions hold many more species than midlatitude temperate regions, and these latter hold more than high-latitude arctic/antarctic regions, but the mechanisms that link the obvious climatic differences to this latitudinal gradient in species richness are not yet understood. It is clear that the reduction of environmental heterogeneity by human activities is lowering the capacity of the globe to support species.

B. Disparity

The genetic differences producing morphological disparity are not based so much on differences in genes and gene frequencies as on gene organization. Individuals belonging to different higher taxa, such as phyla, differ from their very early stages, producing different adult
body plans as they develop. In animals, the developmental regulatory genes that are responsible for organizing the many disparate body plans are surprisingly similar. The animal phyla originated over half a billion years ago, so similarities in those developmental control genes have been conserved since that time. Even members of other kingdoms, such as fungi and plants, share some similarities with animals in their developmental control genes, which therefore have been conserved for over a billion years. Some key developmental genes have been found in all phyla that have been investigated, though they may play a variety of roles. These genes are responsible for regulating still other genes. The regulatory genes govern cascades of gene expression that mediate cellular differentiation and organogenesis, the hallmarks of multicellular body plans, and control the final body architecture. The evolution of disparity has involved the evolution of these basic developmental control systems. Thus, although the genes are similar among phyla, many of the gene expression pathways are different within each phylum and class and so produce distinctive body plans.

In animals, the use of genes that mediate the development of numbers of cell types and morphologies seems to have originated in a spongelike organism. Animal genomes contain two to three times as many genes as most unicellular protists, suggesting that one or more early, massive gene duplication event(s) provided the genetic resources that were necessary to regulate the development of complex morphologies. Vertebrates have approximately four times as many genes as an average invertebrate, suggesting that some massive gene duplication events in early chordate lineages provided a genome of the requisite size to permit the evolution of the very complex vertebrate bodies. Regulatory genetic systems have evolved in land plants as well, quite independently, mediating the disparity found in plant anatomy, but as yet these systems are not as well known as those of animals.

III. CAPTURING FOSSIL BIODIVERSITIES

A. Problems of Sampling Fossil Biodiversity

It is not easy to become a fossil; the remains of only an exceedingly small fraction of all the organisms that have lived have survived to provide information on the nature of life through geologic time. Therefore we must treat the fossil record of biodiversity as a huge sampling problem. One approach would be to estimate how many samples from how many places it would take to adequately represent present biodiversity, and then to compare this ideal sampling pattern with the sorts of fossil samples with which nature has presented us. To estimate today's biodiversity we would need to sample each community in each province on each continent in each major realm, such as terrestrial plains and moutains, continental shelves, and each major oceanic water mass and deep-sea region. We would not have to find every last species, but we would have to understand the patterns of species' distributions well enough to calculate biodiversity from our samples. Although valiant attempts have been made, the problem is so large that estimates of modern biodiversity are not closely constrained. Assessing fossil samples is not easy either. We need to know how densely the biota has been sampled regionally and ecologically by nature in presenting us with the fossil assemblages. And for a given fossil assemblage, we must estimate how completely the living biota was sampled, without knowing how diverse it was in the first place. We also need to know how different ancient biotas were from region to region, for it is not uncommon for large regions to lack fossils for given time periods, and their diversities must then be estimated from assumptions about global patterns. Finally, we must be able to estimate the temporal density of the fossil record and make allowances for intervals of time that lack records.

B. Completeness of the Rock Record

The rocks from which fossils have been recovered are nearly all exposed on land and have therefore been subjected to erosion, and many fossiliferous formations have been stripped away and lost. The first question to ask when reconstructing fossil biodiversities is how much of the original rock record remains. The completeness of a sequence of rocks is independent of its age, relating rather to its depositional history. In general, sedimentary sequences contain many small gaps but increasingly fewer gaps of longer durations. For a given sedimentary sequence, then, fossil taxa with very short geologic ranges may have been present only during a gap, whereas longer-ranging forms can "jump" most gaps and may be captured as fossils. Fortunately, the gaps in one region are commonly represented by sediments elsewhere, though a few periods of time are poorly represented in most regions. For example, few marine sediments are preserved on continents from times when sea levels were particularly low, such as at
the end of the Permian and beginning of the Triassic periods.

Terrestrial sedimentary rocks, deposited as they are above sea level, are subjected to very active processes of erosion, which remove much of the sediment in rivers or in winds and deliver it to the sea. Preservation of terrestrial formations is thus restricted chiefly to downwarped epicontinental basins or interior continental platforms, where there is some protection from the main erosional forces. Such settings tend to be more localized and geologically shorter-lived than nearshore marine sites. Furthermore, terrestrial animal remains are also exposed to decay and erosion in the subaerial environment and tend to be destroyed more easily than the shells of marine forms. Thus, terrestrial rock sequences are less complete than marine ones, and terrestrial animals are not as well sampled by the fossil record as marine animals.

C. Incompleteness of the Fossil Record

Organisms differ greatly in their potential for preservation as fossils. In general, entirely soft-bodied, or small, or rare, or geographically restricted species of short durations have fewer opportunities for preservation than their opposites: a large, abundant, widespread species that lived for many millions of years and has a very durable skeleton is most likely to be represented by fossils. Some phyla, such as Mollusca (clams and snails, etc.) are well represented in marine sediments, whereas others, such as Platyhelminthes (flatworms), though at least as old as the Mollusca and quite rich in species, have left no recognized body fossils. The biodiversity of the living biotas from which fossil assemblages are recruited must usually have been considerably greater than the numbers of fossil taxa that are found. However, it seems that most species with durable skeletons are in fact originally captured in sediments. For example, a study of fossilization of living species of marine bivalves (Mollusca) of the Californian bioprovince found that about 80% of the species, 84% of the genera, and 90% of the families were represented by fossils in sediments of Pleistocene age (the last 1.6 million years). As expected, the missing species were small, thin-shelled, and rare forms, and the missing genera and families were made up of such species.

In richness studies of taxa with high preservation potentials, such as shelled mollusks, fossil assemblages actually have an advantage over samples of the living biota. The fossils are usually time-averaged, that is, they represent the accumulation at a given locality of many generations and may include individuals of species that have occupied the region for only some fraction of the time represented. Thus, a fossil assemblage can sample the species richness that was found in a given area far better than a sample of alpha diversity in a living community. The fossils represent a sampling through time, often over hundreds to thousands of years.

Foote and Sepkoski (1999) devised a method that estimates the probability of preservation of fossil higher taxa from a sort of frequency distribution of the observed geologic range lengths of their members. When such an estimate is made for marine bivalves (clams), the probability of preservation of a genus within a stratigraphic interval representing about 5 million years is around 50% (Fig. 1). The bivalve faunas preserved in the sediments have been significantly degraded. Although some elements of the fossil faunas may have been destroyed by dissolution within the sediments, it seems likely that most of the missing genera were originally present in sediments that were then either removed during intervals of erosion or are not now exposed for collection and study. In other words, although fossils commonly represent a good sample of living associations of easily preservable taxa, not enough such associations are preserved to directly count ancient diversities over lengthy intervals of geologic time.

The completeness of the record of a number of other higher taxa is given in Fig. 1, where the probability of their generic preservation is plotted against the proportion of the living families that have a fossil record. There are two groups that deviate from the general trend. The Cephalopoda (Mollusca) fall well above the trend because they are represented by rich fossil assemblages of extinct higher taxa and have few living representatives. By contrast, Chordichthytes (Chordata) fall well below the trend because their living families have a good fossil record, chiefly of teeth found in relatively young sediments, but the earlier fossil genera, represented by bodies, are mostly known only from single localities and thus have low preservation potentials. Thus the two outliers are readily explained, and the correlation otherwise evident between these totally independent data sets suggests that they are reasonably accurate. Because the marine sedimentary rocks of continental shelves combine high completeness and availability, the marine invertebrates with the highest preservation potentials provide the longest, most complete fossil record of any major groups of organisms.

In establishing the geologic range of a fossil taxon, it is almost certain that we have not found the earliest members to evolve or the last individuals to become
extinct, because the fossil record is so incomplete. It is possible, however, to judge the probable extent of a taxon’s range beyond its actual record by how commonly it is found between its first and last occurrences: if it occurs rather continuously within its known range, then when it disappears it is likely to be absent; if it occurs rarely, then it may have existed well before we see it and continued to exist well after our last record. The number of a taxon’s occurrences in a given time interval thus permits a statistical estimate of the likely extent of its true geologic range. By taking account the observed preservation potentials of a taxon, estimates of the numbers of that taxon present in a given interval can be augmented by granting range extensions beyond the fossil occurrences. Corrections of this sort will produce better approximations of the richness within geologic intervals than the raw numbers of fossil taxa in collections.

D. Partial Remedies for the Spotty Fossil Biodiversity Record

The taxonomic hierarchy, based on Linnaean principles, produces many paraphyletic taxa, but it is useful for purposes of biodiversity reconstruction. Even within taxa with the highest preservation potentials, a large proportion have been lost to erosion and other factors. Genera, on the other hand, contain on the average several species, so that a genus is usually more widespread, occurs in more environments, and contains many more individuals than any one of its species. As a consequence, the fossil representation of genera is more com-
plete than that of species. Families usually include several genera and so are even more completely represented as fossils. Thus by compiling biodiversity data on the generic level, one produces a significantly more complete and somewhat more accurate record than is possible at the species level, and the record at the family level is even better.

There is a price for this taxonomic remedy. The lower levels of taxonomic biodiversity are the more volatile through time, being more sensitive to the environmental changes that affect biodiversity. For example, a wave of extinction may decimate a biota at the species level, but to be recorded in a compilation of family diversity an extinction must extirpate every last species of a family. A small extinction wave might not even be noticed at the family level, considering that fossil data are rather noisy anyway. Of course the time of an extinction, whether recorded at the family, generic, or species level, will be registered in precisely the same place on the geologic timescale. Because species contain reproducing populations, they are in direct ecological contact with environmental parameters, and the volatility of their richness is the most sensitive measure of environmental changes. Higher taxa record the outcome of environmental change more indirectly, that is, when their species first appear, or when all of them disappear from the record. Nevertheless, their behavior in the face of environmental change presumably reflects the broad similarities in physiological and behavioral responses of their species, inherited from common ancestral species. While generic and family diversities can make quite useful proxies for species diversity in the fossil record, they are in fact measuring different features, and this should be taken into account in interpretations of fossil richness.

Most individual fossil samples represent a particular ecological community, perhaps somewhat intermixed with rare species that are chiefly found in other communities, where they are common. In some cases, though, the samples represent mixtures of common faunal elements that represent different communities but that lived in close proximity. It is sometimes possible to establish biodiversities for metacommunities by taking many samples of similar, approximately contemporaneous fossil assemblages at different localities throughout a region, using the rule that species that occur together most often as fossils are most likely to have lived together. Because two fossil localities are likely to vary in composition even if drawn from the same parental biota, an increasingly full picture of regional biodiversity is gained when numbers of localities are pooled. Such studies can lead to understanding the gross diversity patterns among biotic provinces. Finally, sampling of approximately contemporaneous localities can be made at the global level, in this way providing a sample of global biodiversity for the realm(s) involved.

Because most fossil localities are not directly dated by the more accurate radiometric techniques, correlation between them can only be approximate, particularly when the fossil assemblages represent different communities or provinces. This difficulty can be mitigated by using fairly coarse time units, within which most fossil localities can be placed with some confidence. Global biodiversity studies using families and genera usually employ temporal bins of from 5 to 8 million years in duration. Some studies have used geologically based time subdivisions, such as Stages; in these cases the diversity must be normalized to the differing Stage lengths.

The various approaches to help in overcoming the incompleteness of the fossil record yield binned samples of adequate size and quality to establish general levels and trends of biodiversity within well-skeletonized groups on global scales over long periods of time. The times of major diversifications or extinctions and periods of high and low biodiversities are captured by such data. Changes in the relative dominance of different major clades are also clearly demonstrated. However, the price for binning is that fine details of diversity changes are lost. Yet preliminary studies indicate that the generic record captures the same events as the family record and indeed is more sensitive to diversity changes. Furthermore, increasingly accurate dating techniques are permitting the use of ever-shorter time intervals. These ongoing refinements have confirmed the general trends described from the coarser data, while permitting the recovery of more detailed diversity information that reflects more closely the history of species richness.

IV. MICROBIAL DIVERSITY OF THE ARCHAEOAN AND PROTEROZOIC ERAS

The earliest rocks known on earth date to about 4.6 billion years ago. The earliest eras in earth history, long thought to be barren of indications of life, were named the Archaean (about 4.6–2.5 billion years ago) and the Proterozoic (about 2.5 billion–543 million years ago). The most recent era was named the Phanerozoic (visible life) because it contained a fossil record. However, microbial fossils have now been found to date well back
into the Archaean. The earliest are filamentous prokaryotic microbes resembling cyanobacteria ("blue-green algae"), recovered from chert beds that are over 3.5 billion years old. There are a few other Archaean records of microbial fossils that may be from other bacterial groups, representing both benthic and planktonic realms. However, bacteria and archaee have such extremely low preservation potentials that it is impossible to reconstruct their history of richness from fossil evidence. According to molecular phylogenies, a number of major bacterial branches, equivalent at least to phyla in a hierarchical system, arose before the cyanobacteria. It appears likely that there was a rapid early diversification within the bacterial and, probably, archaean domains that produced a highly disparate microbiota well over 3 billion years ago. Considering the way in which lineages branch during evolutionary diversifications, it is likely that the microbiota was quite rich during much of Archaean and Proterozoic times. Recorded fossil diversity is much greater in the Proterozoic, but sampling is much better in rocks of that age and the recorded fossil richness simply tends to correlate with the number of formations and samples studied. One exception is in the latest Proterozoic, when sampling is highest but when the richness of fossil microorganisms declines, suggesting a real diversity drop at that time. The fossil record of richness of the Proterozoic microbiota is given in Fig. 2.

Unicellular fossils belonging to the Eukarya are difficult to separate from those of other domains. Relatively large fossil algal ribbons (1 m long, 2 mm in diameter) about 2.1 billion years of age are likely to be the first known eukaryotes. It is possible that Eukarya evolved during the Archaean. The best Proterozoic records of unicellular Eukarya are of spheroidal plankters with organic walls, sufficiently tough to be preserved in some quantity. They increase in number, richness, and morphological complexity throughout the Proterozoic. As noted earlier, great bacterial disparity is likely to have been achieved in the Archaean, and the major richness trend may be attributed chiefly to an improving fossil record rather than to increased biodiversity.

V. PHANEROZOIC MARINE BIODIVERSITY

A. Neoproterozoic Animals

The record of metazoan fossils dates from at least 570 and possibly 600 million years ago, in the late Proterozoic (the "Neoproterozoic"), but it has proven

![Diagram of fossil richness over time](image)

**FIGURE 2** Species richness of several groups of unicellular microfossils as represented in Proterozoic sediments. The morphologic categories to the left of the heavy dashed line are presumptive Bacteria; those to the right are Eukarya. The record is unlikely to mirror the history of either richness or disparity, although the decline late in Proterozoic time may be real. (From Schopf and Klein, 1992, p. 535. Reprinted with the permission of Cambridge University Press.)
to be very difficult to relate those early fossils to living groups, which first appear near the beginning of the Phanerozoic. There are two chief modes of fossilization: trace fossils, which are trails, burrows, and similar structures left in sediments by animal activities; and body fossils, which register the gross morphology of the animals. Most Proterozoic body fossils are restricted to body impressions, skeletons being absent until latest Neoproterozoic time. The early trace fossils are simple curved trails, and though traces become increasingly complex and diverse as time passes, they remain small throughout the Neoproterozoic, mostly 1 mm or less in width, though a few range to 5 mm. The early body fossil impressions (the “Vendian” or “Ediacaran” fauna) are chiefly of frondose and discoidal organisms that are large (some reach 1 m) and bear a general resemblance to cnidarians, although they differ from living forms in important structural details. Ediacaran faunas also include rare sponges and a few forms that resemble bilateral animals but that cannot be assigned to any living group. Estimating the living diversity of Neoproterozoic times from these fossils is difficult; body fossils of the trace-makers, for example, are not known, and many kinds of organisms can leave a given type of trace fossil. When body fossils appear during the Cambrian (see the following section), traces become larger (some are centimeters wide), much more varied, and more common. Therefore the best guess for Neoproterozoic animal diversity is that it was lower than that of the Cambrian, beginning sometime before 570 million years ago (Ma), possibly tens to hundreds of millions of years earlier. This guess is likely to be correct as far as body plan disparity is concerned, but it remains possible that Neoproterozoic seas supported a rather high richness of species that had such low preservation potential that they have not yet come to light.

B. The Cambrian Explosion Establishes Metazoan Disparity

About 543 Ma there was a notable increase in the complexity, sizes, and abundances of trace fossils, one of which (Tripichus) marks the basal Cambrian boundary by international agreement. Rare, minute, mineralized skeletons of uncertain affinities had appeared just prior to the first appearance of Tripichus, and continued to appear in increasing kinds and numbers, along with an increasing trace diversity, in younger rocks. At about 530 Ma the earliest undisputed skeletons of living animal groups appeared and, during the next 10 million years or so, 11 animal phyla made their first appearances, a period known as the “Cambrian explosion” because of the geologically sudden appearance of diverse, abundant skeletal remains. Invertebrate skeletons are then continuously present for the remainder of geological time. Among the body plans to appear during the explosion are complex representatives of all major animal clades, including annelids, arthropods, and chordates (Fig. 3). Judging from the phylogenetic tree of animals, all living phyla had originated by the close of this interval, although some living phyla with low preservation potentials are still completely unknown as fossils.

Many of the explosion animals include features that are not present in their living allies, and thus extend the known morphological ranges of their clades. Some of these forms have such distinctive body plans that they are considered to be phyla of their own, whereas others can be nested within living phyla but not within living classes. These forms indicate that the disparity among the body plans of Early Cambrian animals was at least as great as it is today. For example, the relatively few Cambrian arthropod taxa are just as morphologically disparate as are living arthropods, which have had the benefit of over half a billion years of subsequent evolution and consist of million of species today. Thus disparity and richness are evolutionarily decoupled. The Early Cambrian burst of disparity in the fossil record suggests that there was relatively rapid evolution of the developmental regulatory gene systems at that time.

C. Phanerozoic Diversity Patterns in the Marine Environment

1. Standing Diversity of Marine Invertebrates

The combination of completeness, disparity, richness, and duration of the marine fossil record is superior to that of other realms, especially for marine invertebrates, and thus is best suited as an introduction to major diversity patterns. Figures 4 and 5 depict the overall pattern of family richness for marine animals, almost entirely shallow-water (continental shelf depth) forms. It is likely that the range of body plan disparity declined somewhat after the Cambrian, and probably showed a significant drop and then a recovery following the Permian–Triassic extinction, but there are no data that incorporate the entire marine fauna. However, Jack Sepkoski has compiled an excellent data set on family richness, which has certainly varied significantly.
<table>
<thead>
<tr>
<th>PreC</th>
<th>Paleozoic</th>
<th>Mesozoic</th>
<th>Ceno</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendian</td>
<td>C</td>
<td>O</td>
<td>S</td>
</tr>
<tr>
<td>Cambrian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordovician</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silurian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Devonian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carboniferous</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triassic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jurassic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cretaceous</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Nematoda</th>
<th>Nematrinia</th>
<th>Entoprotea</th>
<th>Pediformes</th>
<th>Nematostomes</th>
<th>Placozoa</th>
<th>&quot;Mesozoa&quot;</th>
<th>Phyla Helecoretes</th>
<th>Gnathostomulida</th>
<th>Gastrotricha</th>
<th>Acanthocephala</th>
<th>Longifera</th>
<th>Kinorhyncha</th>
<th>Sipuncula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bryozoa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 3 The earliest appearance of body fossils of living phyla. The earliest records of small, soft-bodied forms with low fossilization potentials are scattered through time, being found at localities with unusual preservation, such as in amber, or having no fossil records at all. The only phylum with well-mineralized skeletons that does not appear in the Cambrian is the Bryozoa, which, however, has numbers of nonmineralized forms today and may have had no mineralized representatives during the Cambrian. C., Cambrian; O., Ordovician; S., Silurian; D., Devonian; C., Carboniferous; P., Permian; Tr., Triassic; Jr., Jurassic; and K., Cretaceous. (Modified from Valentine et al., 1999.)

During the Cambrian, family richness rose abruptly and seemed to be reaching a plateau as the period ended. However, the Ordovician witnessed a dramatic increase in family richness, which reached a plateau near 400 families by the end of the period (see Fig. 4). This richness plateau lasted for about 300 million years, during which there were two major extinctions that lowered family richness for a relatively brief time, followed by recoveries to the plateau level until the close of the Permian period. At that time, the plateau was abruptly terminated at about 250 Ma by the greatest extinction on record, near the Permo-Triassic (P/T) boundary. About half of the marine families were lost. From that low point during the depths of the extinction, family richness has climbed at a rate quite similar to the Cambro-Ordovician rate, and for a far longer time, peaking at the present. This last richness rise is interrupted twice by major extinctions, the last of which, near 65 Ma, is the famous Cretaceous–Tertiary (K/T) extinction that eliminated the dinosaurs. Despite these interruptions, family diversity today far exceeds that of the long Paleozoic plateau.

Perhaps as few as one-third of living marine families have durable skeletons; the rest have very poor preservabilities. Nevertheless, there is evidence to suggest that the major trends in the fossil record are real, and that the ratio of easily to poorly preservable forms has been similar enough, at least following the Cambrian explosion, that the general magnitudes of family diversities can be reconstructed from the preserved fossils. For example, the diversity patterns of easily and poorly preservable taxa are similar throughout the shallow ocean today, so far as we can tell. Further, the ratio of poorly preserved to easily preserved families that are found (see Fig. 5) holds reasonably steady through the Phanerozoic, increasing...
slightly toward the present, suggesting that the record is sampling a fairly consistent fraction of the poorly preserved clades, and the ratio improves slightly in younger sediments as would be expected. Also, the presence of a long richness plateau during the Paleozoic is unlikely to be a matter of chance, and suggests some stability in sampling success.

2. Standing Diversity of Marine Chordates
The phylum Chordata includes an invertebrate class, the Cephalochordata, that is known from the Early Cambrian and is represented today by amphioxus, (Branchiostoma spp.), but this class is not known to ever have been diverse. There are primitive groups within the class Vertebrata, such as lampreys and hagfish, that in fact lack vertebral columns. A group of marine vertebrates at about this level of organization, known as the Conodonta from toothlike feeding structures, appeared during the Middle Cambrian and left a significant Paleozoic fossil record. Conodonta became extinct during the Triassic. Nonconodont jawless fishes (Agnatha) appeared in the late Upper Cambrian and radiated into four distinctive, disparate clades during the Silurian; many agnathans had bony armor that has provided most of their fossil record, but the agnathans seem not to have been rich in species. They were joined by jawed fishes by Devonian time. The jawed forms radiated to produce several disparate types—arthrodires and placoderms, with bony head shields and plated armor, and several chondrichthian groups (sharks and rays), which were the richest of fish clades in Late Paleozoic seas. Also arising at least by Early Devonian time was a clade, the osteichthyes or bony fishes, which was of secondary importance during the Paleozoic but came to be the most richly diverse vertebrates in the sea during the Mesozoic and quite dominate marine vertebrate faunas today. Thus the history of marine fishes can be summed up as beginning modestly in the Cambrian and finally achieving significant disparity during Silurian and perhaps Early Devonian radiations that established most major groups, which then waxed and waned in richness, with bony fishes finally coming to dominance. Reptiles entered the sea during the earliest Mesozoic and mammals early in

![Graph showing marine animal family richness](image)

**FIGURE 4** Marine animal family richness known from the fossil record. Geologic period abbreviations are as in Fig. 3. Major faunas are labeled C, (Cambrian Fauna), Pz (Paleozoic Fauna), and Md (Modern Fauna). Shaded area at top of curve indicates animals that lack durable skeletons. (From J. J. Sepkoski and A. Miller, in Valentine, 1985. Copyright © 1985 by PUP. Reprinted by Permission of Princeton University Press.)
FIGURE 5  Phanerozoic diversity in the marine fossil record, showing animal family richness by phylum or class. Geologic periods are as in Fig. 3. (From J. J. Sepkoski and M. Hulver, in Valentine, 1985. Copyright © 1985 by PUP. Reprinted by permission of Princeton University Press.)
the Cenozoic, but neither group became very rich in marine species.

3. Taxonomic Turnover
Relative clade richnesses vary considerably across Phanerozoic time. In Fig. 4, the changing assemblages of dominant clades are classed into three major faunas. The Cambrian Fauna (labeled €), composed dominantly of trilobites (Arthropoda), does not diversify during the Ordovician and is overtaken by the expanding clades of the Paleozoic Fauna (labeled Pz), dominated by articulate (Brachiopoda) and crinoids (Echinodermata). The Paleozoic Fauna is decimated by the P/T extinction and does not increase significantly thereafter, and is overtaken by the diversifying Modern Fauna (labeled Md), which is dominated by mollusks and fishes. Elements of all of the faunas are present throughout the Phanerozoic.

Figure 5 depicts the great variety of family richness histories recorded for marine classes and phyla; the dominance of the characterizing clades during each of the three faunas is clearly evident. There are some general features that link the diversity patterns to evolutionary rates that seem to be inherent in the clades. The dominant clades of the Cambrian Fauna have very high evolutionary turnover rates—high rates of both diversification and extinction. The dominant clades of the Paleozoic Fauna have moderately high turnover rates, and those of the Modern Fauna have the lowest turnover rates of the dominant clades. As a result, the extinction rate for all families combined, excluding the major extinction events, declines steadily throughout the Phanerozoic. Evidently, a strategy of being resistant to extinction is superior to having a high speciation rate, perhaps because slight increments in environmental fluctuations can force a high-turnover lineage too close to extinction to recover. Note that the presently dominant bivalve and gastropod mollusks, slow-turnover clades, increase their relative representations inexorably throughout the Phanerozoic, and are not greatly affected even by the P/T extinction. A few clades that have fair preservation potentials can be shown to have survived for hundreds of million of years despite having very low diversities; the inarticulate brachiopods and the scaphopod mollusks are classic examples (see Fig. 5). These are low-turnover clades; if they had high turnovers they would presumably have been swept away by even a small extinction event. Both speciation and extinction rates are probably tied to population parameters that of course evolved without respect to clade longevity or dominance, which in a sense are merely side effects, though they profoundly influence the composition of the world’s biotas.

VI. PHANEROZOIC TERRESTRIAL BIODIVERSITY

A. Plant Diversity
“Land plants” include two major groupings: forms such as mosses and liverworts, collectively termed bryophytes, which lack a water-transporting system; and the tracheophytes, which possess vascular tissues that conduct water vertically and permit significant upward growth. It seems likely that tracheophytes arose from among the mosses, perhaps more than once. The earliest fossil record of plants on land may be furnished by spores of Middle Ordovician age that are possible bryophytes; tracheophytes probably date from Early Silurian, but their vegetative elements first appear in the Middle Silurian. From then through the Devonian, a plexus of early plant types diversified, and from this plexus the modern groups of lycopods, horsetails, ferns, and gymnosperms evolved by latest Devonian time; there were forests before the close of the Devonian. This early burst of disparity is similar to that found among animals, but the richest of living plant groups, the angiosperms or flowering plants, appear much later, arising perhaps in the Late Jurassic but spreading during the Late Cretaceous and dominating Cenozoic plant associations. There is no animal analog for the late origination of a dominant major taxon, except for the class Mammalia, also terrestrial of course. However, the disparity indicated in the pollen record of angiosperms has been investigated, and it shows an early burst of disparity increase in the Early Cretaceous that gradually tapers off to a plateau by late in the Cretaceous. Evidently angiosperm “body” plans were evolved early in the history of the group and the rise of species richness followed later, the same pattern that is found within many marine invertebrate groups and in numbers of vertebrate groups. Figure 6 indicates the relative richness of major land plant groups through time as judged from their fossil records.

B. Animal Diversity
1. Invertebrates
Animals came ashore with the earliest plants, so far as can be told. It is quite possible that early terrestrial animals fed on bacteria, fungi, or protists and were ashore before tracheophytes; early arthropod burrows are reported in Ordovician soils. However, the earliest undisputed terrestrial animal remains, of arthropods, are from the Late Silurian. Because these fossils are predators (primitive centipedes and an arachnid), there
FIGURE 6  Phanerzoic diversity in the terrestrial fossil record, showing plant and animal family richness by phylum, class, or other higher taxon. Geologic periods are as in Fig. 3. (From J. J. Sepkoski and M. Hulver, in Valentine, 1983. Copyright © 1983 by PUP. Reprinted by permission of Princeton University Press.)
must have been by that time an extensive fauna of primary consumers and a fairly complex ecosystem, evolved in previous epochs. There is little evidence to judge how quickly and to what levels terrestrial invertebrate diversity rose during those times.

Insects evidently arose in the Early Devonian, and they have a relatively rich fossil record for a terrestrial group, being richer than the vertebrate record. The same species proportions found among living orders are also found among Tertiary fossils, suggesting that the insect record, while certainly incomplete, is not heavily biased taxonomically, and that at least their major fossil trends probably reflect actual paleobiological trends. During their earliest, Devonian history, insects are sporadically represented; most of them are predators, whereas the primary consumers seem to have been chiefly detritus feeders; little herbivory is found. Insects underwent a significant expansion in the Carboniferous, and species richness rose to a high just before the major Permo–Triassic extinction, which affected insects significantly (see Fig. 6). However, the insects rebounded, rising steadily to their present overwhelming richness. Probably the rise was more evenly distributed in time than indicated in Fig. 6, for there are some spectacularly rich insect faunal localities in the Late Tertiary, while equivalent faunas are not yet known from earlier deposits. As with marine invertebrates, significant morphological disparity was reached very early in insect history, preceding the greatest rises in richness. Though insect richness seems closely tied with the richness of the angiosperm flora, insect disparity, even including the evolution of the array of mouthparts and feeding structures present today, was largely accomplished well before the origin of angiosperms.

2. Chordates

All the primitive groups of jawed fishes are found in terrestrial aquatic deposits during the Middle Paleozoic, but on land as in the sea it was the bony fishes (Osteichthyes) that radiated to produce rich faunas that dominated in fresh waters, from the Devonian onward. The richness of terrestrial fossil fish faunas declined significantly in the Mesozoic, perhaps partly owing to a poor aquatic fossil record from that era.

Tetrapods first appear in the Devonian, possibly Early Devonian, in effect as part of the general radiation of fishes, but with limbs that eventually permitted locomotion on land and led to the evolution of the Amphibia. The fossil record of this group is very sparse, but it is clear that the amphibians diversified extensively to reach a level of family richness comparable to today’s by the close of Carboniferous time. The early amphibians were rather different from living forms, however, and included large animals, which were eventually replaced ecologically by reptiles. Three important groups arose from amphibian stocks, one leading to reptiles, one to mammals, and one to dinosaurs and birds. The early histories of these groups are so poorly known that their patterns of diversification cannot be accurately reconstructed. It is possible that both reptiles and dinosaurs gradually increased in richness and diversity through most of the Mesozoic era. The richness of these clades was reduced drastically during the extinction at the close of the Cretaceous.

While mammals must have diversified as well during the Mesozoic, it is in the early Cenozoic that their fossil record is marked by a burst of new appearances and, by about 10 million years after the start of the Cenozoic, all of the 17 modern orders of placental mammals had appeared. Only two of these orders are known from the Cretaceous, and while it is likely that most trace their roots to a few tens of millions of years earlier, mammals were clearly a minor though probably not unimportant part of the tetrapod fauna before the Cretaceous–Tertiary (K/T) extinction. It is usually assumed that mammals replaced dinosaurs, which became extinct at the K/T boundary, in ecological roles that permitted large body sizes. The only dinosaur descendants today are birds, which have a poor fossil record. Modern bird orders are unknown before the Cenozoic, suggesting that there was a significant K/T extinction of birds, leading to the radiation that produced modern types. There must indeed have been a significant diversification of birds during the Cenozoic, but as the roots of the modern lineages are uncertain and the bird record is poor, the pattern of bird diversity increase has not been established.

VII. PRINCIPAL BIODIVERSITY FACTORS

A. Plate Tectonics and Global Heterogeneity

The geographic pattern of land and sea is constantly changing due to the processes of plate tectonics, which create and destroy the crust of the earth. At present the crust is divided into about 6 major and 14 minor crustal plates. New crust is added at one plate margin, a rift marked by deep-sea ridges, from sources of molten rock in the earth’s interior. At the opposite plate margin, the crust plunges back into the interior, its descent marked by deep-sea trenches or “subduction zones”; this crust
then remelts. Continents or islands on a crustal plate therefore move across the earth’s surface, from the constructional margin toward the destructive margin, as if on a conveyor belt. Continents separated by a rift will diverge as they are carried apart on their respective plates, and continents will fragment if a rift cuts across them. On the other hand, continents separated by a subduction zone will converge, and if they meet they can weld together. The rates of movement of the plates are only several centimeters per year. Nonetheless, because plate tectonic processes have been incessant since well before the origins of animals and plants, the continents and oceans have changed their numbers, sizes, and geographic configurations over the many millions of years since those organisms first evolved.

As continents drift across the earth’s surface they enter new climatic zones, and as oceans enlarge or shrink or change their shapes, the oceanic circulation changes, with warm or cool currents directed to different regions. The biota of any given area will evolve to adjust to the changing conditions, and may become enriched, perhaps if moving into the tropics, or depauperate, perhaps if moving into high latitudes. At the same time, migration routes are opened or closed by the shifting geography, permitting invasions of biotas into some regions while isolating the biotas of other regions. The biodiversity patterns both on land and in the sea reflect the environmental patterns created by plate tectonic processes.

In addition to its effects on rearranging regional biodiversity patterns, changes in global geography can significantly affect global biodiversity levels. The most impressive example is in the rise of biodiversity during the last 200 million years or so, from the Middle Mesozoic era through the Cenozoic era. At the beginning of the Mesozoic, the continents were welded together into a supercontinent, called Pangaea. Although Pangaea was certainly environmentally heterogeneous, with large climatic and topographic variations, the climatic zones tended to be occupied by widespread biotas that could spread across what are now separate continents. In the seas, the shelf encircled the single supercontinent in an essentially uninterrupted band. While the distribution of species was limited by climatic zonation and perhaps regional habitat peculiarities, even shallow-water species tended to be relatively widespread and of course were largely confined to the one shelf, though perhaps some were present on oceanic islands, of which little is known. As the Mesozoic passed, however, the continents were traversed by rifts and began to break up into smaller landmasses, between which were arms of the sea that finally widened into oceanic expanses. By the close of Early Cenozoic time, all of the present continents that are now separated had broken from each other.

Species that could once range widely across Pangaea became isolated on separate continents. As the continents dispersed into different climatic zones and developed different topographies, their terrestrial biotas each evolved distinctively in response to local conditions and events. For example, North America, Eurasia, and Australia each developed distinctive temperate biotas, and the tropics of the Americas became distinct from the Old World tropics and those of the western Pacific. In the sea, the growing geographic differentiation was even more profound, as new shelves appeared on each side of the new oceans, significantly raising the number of isolated shelf segments. As on land, each continent evolved a distinctive biota in response to the unique set of environmental conditions and events that developed in each region. The heterogeneity of world environments thus increased greatly during this period, on a global scale. These plate tectonic effects are responsible for a significant part of the important rise in standing global marine diversities during the Mesozoic and Cenozoic as represented at the family level in Fig. 4; an accompanying rise occurred in the continental biota.

B. Climate Change

Climate is another major factor governing global biodiversity. For example, earth underwent a general cooling trend during the Cenozoic, although there were shorter warming and cooling events superimposed on the generally falling temperature curve. The temperature change was greatest in high latitudes, thus increasing the temperature differential between the poles and the equator. The margins of the tropics were shifted toward the equator, the tropics became compressed, and cooler climates came into existence in high latitudes, shifting equatorward themselves as cooling continued, to be replaced poleward by still cooler climatic regimes.

In the sea, the increasing temperature gradient made for a markedly increased provincialism. Tropical species whose biology was related to climate became progressively more restricted to lower latitudes, with some occasional, relatively short-lived reversals in which species tolerant of lower temperatures arose and diversified in the more poleward regions. Because the major rifting that produced the Cenozoic continental pattern was primarily in a north–south direction, most major continental shelves trend north–south in the Cenozoic, as they do today, and so there was a rise in species richness on each shelf as the latitudinal thermal gradient in-
creased, multiplying the effect on global marine biodiversity by the number of isolated shelves. The Cenozoic global marine rise in family richness shown in Fig. 4 is compounded partly of the increase in the number of isolated shelves following continental rifting and partly of the increased provincialism on each shelf created by an increasing thermal gradient; together they have formed a powerful engine of global species enrichment.

In terrestrial environments, family richness also increased during those Cenozoic events, and for the same reasons. The continental expanses provided for a three-dimensional array of habitats (including high mountains and basins) unlike the rather two-dimensional array along the narrow, shallow continental shelves, and therefore the biotic response has been correspondingly more complicated in terrestrial than in marine environments. Not only did high-latitude climate zones open up, but high-altitude regions became cooler, increasing the contrasts between mountain and lowland environments. A lowering of rainfall in some continental interiors produced semi-arid plains and steppes. The biota responded to these increases in environmental heterogeneity by producing, for example, alpine and grassland plants and animals, thus enriching both the flora and fauna. The species richness of angiosperms among plants and of insects among animals profited the most from these events.

During the Late Cenozoic, as cooling has continued in high latitudes, massive ice accumulations on a subcontinental scale have produced a series of glacial ages interrupted by warmer interglacial periods, most pronounced during the Pleistocene, which began about 1.6 million years ago. Some regional extinctions occurred during the early onset of cold periods, but in general there has been little extinction, and thus no significant lowering of global biodiversity, associated with these climatic swings. Instead, the biotas of both land and sea have tended to migrate with their climatic zones, southward during glaciations and northward during interglacial times. Thus regional diversities have changed as species have migrated, but overall diversity has not been significantly affected. Indeed, the Recent biosphere may be the richest in species during the entire history of animals and plants, at least prior to the deleterious influences of human activities on biodiversity.

C. Extinction Events
Changing environments have often provided opportunities for the origin of species, some of which have clearly led to major morphological novelties and enhanced morphological disparity. At the same time, changing conditions have led to the extinction of species, which has occurred more or less continually over Phanerozoic time. As noted earlier, extinction rates vary among taxa, whose histories tend to reflect their rates of turnover. The species richness of a taxon results from an interaction between its rates of speciation and extinction. Thus when there has been some unusually profound disturbance to the global environment, and extinction rates have become unusually high, species richness has fallen dramatically in events called mass extinctions. The most drastic such event, recorded in marine fossils, was the Permian–Triassic (P/T) extinction about 250 million years ago, at the close of the Paleozoic era. As many as 90% of marine species may have become extinct across the P/T boundary (see Figs. 5 and 6). This extinction was highly differential; the groups that did best in surviving the extinction and in rediversifying were those that dominate the marine fauna today. Disparity as well as richness was affected; some entire classes and orders of marine invertebrates disappeared. Brachiopods, crinoids, and other groups that were dominant before the extinction were reduced to minor roles. The causes of this extinction, and the extent of its effects on the terrestrial biota, are not yet clear.

There have been a number of other mass extinctions (five major events in all) and numerous regional extinctions that were severe locally. The most famous of the mass extinctions, the K/T, occurred about 65 million years ago at the close of the Mesozoic era, although it did not have as profound an effect on the marine environment as the P/T event. The K/T extinction strongly affected terrestrial as well as marine life. It probably resulted from the impact of an extraterrestrial object, such as a meteor, with the earth.

The “background” extinctions that account for the general turnover of taxa are accompanied by “background” speciations, so that species richness levels do not change much during normal turnover and tend to be significantly affected by background effects only during trends that last tens of millions of years or more. However, the more abrupt, severe regional and mass extinctions are not immediately compensated by correspondingly massive speciations, and indeed the recovery of species richness usually requires millions of years following a mass extinction. For phyla, classes, or orders that are lost to extinction, there is no recovery at all, and the unique gene regulatory systems involved in their architectures are gone forever.

Clearly, biodiversity data from the fossil record support the notion that species richness is largely associated with environmental heterogeneity, and thus the preser-
vation of species diversity can be promoted by preserving habitat diversity. There are entire animal phyla that are represented today by very few species: the phylum Phoronida by only 12 or so, and the phylum Priapulida by perhaps 17, for example. These phyla harbor uniquely organized genomes that have been present for over half a billion years of earth history, and that are largely unstudied. Their preservation should be a priority, for they represent major elements in the disparity of the present biosphere.

See Also the Following Articles
ARCHAEA, ORIGIN OF • BACTERIAL BIODIVERSITY • BIODIVERSITY, ORIGIN OF • BIOGEOGRAPHY, OVERVIEW • CLADOGENESIS • EUKARYOTES, ORIGIN OF • GENES, DESCRIPTION OF • MICROBIAL BIODIVERSITY • TAXONOMY, METHODS OF

Bibliography
GLOSSARY

case law Interpretations of the U.S. Constitution, statutes, and regulations provided by the judicial branches of U.S. federal or state government. Such interpretations are provided when two or more parties disagree as to the meaning of a law in a specific context and bring it to the courts to decide.

executive order A directive from the U.S. president, or a state governor, specifying actions by government officials and agencies. State and federal executive orders often describe how to administer a provision of a statute, treaty, or the constitution and usually have the force of law.

regulations Rules and administrative codes required by statutes and issued by local, state, and federal government agencies. Regulations have the force of law since they are adopted under the authority granted by statutes.

statutes or legislation Laws passed by the legislative branch of U.S. federal or state governments. The U.S. Congress passes federal laws and state legislatures pass state laws. Local laws are usually called municipal ordinances.

BIOLOGICAL DIVERSITY, at the genetic, species, and ecosystem levels, is influenced by legislation and regulation that is initiated and enforced at local, national, and international levels. Although a great many laws and regulations governing human activities have indirect impacts on biodiversity, this article will summarize only major laws and regulations that directly affect biodiversity by influencing how genes, species, and ecosystems are used and conserved. Although the legal measures taken by each nation have an impact on global biodiversity, for brevity this article will focus primarily on U.S. laws.

I. INTRODUCTION

For centuries, customary or traditional laws have governed the use of biological resources throughout the world. The systems of customary law in traditional cultures often dictate limits on how much of a specific biological resource can be used and in what seasons. In many societies an individual's rights to biodiversity
are not always linked to land tenure. As traditional societies are becoming more integrated into national and global markets and political systems, traditional knowledge and laws are being lost. As human population levels increase in all societies, voluntary actions governed by cultural practices are sometimes insufficient to protect biodiversity, leading to an increasing reliance on the creation and enforcement of government legislation and regulation to protect biodiversity.

A variety of motivations have led to creation of legislation intended to help conserve biodiversity. Early legislation was primarily aimed at protecting biodiversity as an economic resource for humans (e.g., for hunting or agriculture). Recent laws, and amendments to earlier laws, have included other reasons, such as the value of biological diversity to overall ecosystem health and the intrinsic value of species. Although U.S. national parks were originally created primarily for scenic and recreational values, the 1916 National Park Act identified conservation of wildlife as an important goal of parks.

Laws and regulations can aim to influence actions undertaken by a range of actors (government agencies, private corporations, and individual citizens) and to affect a range of resource types (animals, plants, water, land, and specific habitats) with differing ownership and resource control (public and private). The success and evolution of legislation and regulation in the United States are in part due to the actions of nongovernmental organizations that lobby for new or amended legislation, oversee how agencies implement legislation, and use the judicial system to help ensure that laws are enforced and interpreted appropriately.

All laws and regulations in the United States are based on rights found in the U.S. Constitution. The federal government derives its authority to regulate wildlife from its constitutional authority to regulate commerce, to protect its own property (federal lands), and to make treaties. When federal laws conflict with state laws, the federal laws generally take precedence.

Many advocates of the right to private property, also supported by the U.S. Constitution, view legislation to protect biodiversity as sometimes conflicting with their right to private property. The Fifth Amendment to the Constitution requires that when private property is taken for public use (whether to benefit public safety, create roads, protect wildlife, or for any other purpose) there must be "just compensation." Most state and federal case law indicates that restrictions on killing wildlife are not a sufficient taking of private property to require compensation by the government (Bean and Rowland, 1997).

Despite the large number of environmental laws and regulations that now exist, and the many positive impacts that they have had, these positive steps have been insufficient to prevent the loss of biodiversity. Some argue that federal and state governments give greater weight to protecting current economic growth and private property rights than to protecting biodiversity because of a lack of constitutional authority to protect future generations. Some conservationists advocate an amendment to the U.S. Constitution that obligates the government to protect biodiversity for future generations. Many other nations and at least 12 of the United States require conservation of nature in their constitutions. For example, the Constitution of the Commonwealth of Pennsylvania (Art 1, Sec 27) states that people have a right to preservation of natural values and that "natural resources are the common property of all people, including generations yet to come. As trustee of these resources, the Commonwealth shall conserve and maintain them for the benefit of all the people."

One of the controversial issues confronting the agencies and courts in implementing all legislation described in this article is determining how the laws apply to American Indian reservations, which are covered under separate treaties, and to traditional subsistence use of biodiversity by Alaskan natives. Balancing the needs of people who depend on biological resources for food and shelter with the need to protect wildlife is an issue with both legal and ethical challenges.

This article will summarize only legislation and regulations that aim to directly conserve biodiversity. Table I presents a summary of selected laws intended to help conserve biodiversity. Many other environmental laws have indirect positive impacts on biodiversity, for example, by controlling air pollution (Clean Air Act), water pollution (Clean Water Act), hazardous waste (Resource Conservation and Recovery Act), global climate (Global Climate Change Convention), trade (General Agreement on Tariffs and Trade) or by creating conservation incentives through tax policies (Internal Revenue Code). This article will also address the federal, state, and local programs (often resulting from legislation) that help to conserve biodiversity.

## II. INTERNATIONAL TREATIES

International agreements are important for conserving biodiversity because many species have very large natural ranges or migrate between countries, most habitats do not follow national boundaries, and the earth is increasingly globalized in trade and movements of people. Under international treaties, the ratifying countries
### TABLE 1

Selected Legislation Influencing Biological Diversity

<table>
<thead>
<tr>
<th>Law (year)*</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multilateral international agreements</strong></td>
<td></td>
</tr>
<tr>
<td>International Convention for the Regulation of Whaling (1946)</td>
<td>Designates protected whale species and hunting limits</td>
</tr>
<tr>
<td>Convention on International Trade in Endangered Species of Wild Fauna and Flora (1973)</td>
<td>Prohibits and regulates international trade of specified plants and animals</td>
</tr>
<tr>
<td>Convention on Biological Diversity (CBD, 1993)</td>
<td>Encourages nations to conserve, use, and share the benefits from conservation of genetic, species, and ecosystem levels of biodiversity; the United States is not yet a party to CBD</td>
</tr>
<tr>
<td><strong>U.S. federal laws</strong></td>
<td></td>
</tr>
<tr>
<td>Lacey Act (1990)</td>
<td>Prohibits commerce in wildlife injurious to agriculture and in wild animals and some plants taken in violation of U.S. and foreign laws</td>
</tr>
<tr>
<td>National Park Service Act (1916)</td>
<td>Requires conservation of scenery, natural and historic objects, and wildlife in national parks and other protected areas for public benefit</td>
</tr>
<tr>
<td>Migratory Bird Treaty Act (1918)</td>
<td>Implements treaties with Canada, Mexico, Japan, and the Soviet Union to establish hunting seasons and protect habitat for migratory birds</td>
</tr>
<tr>
<td>Bald Eagle Protection Act (1940)</td>
<td>Prohibits killing or possessing bald and golden eagles, their nests, or parts</td>
</tr>
<tr>
<td>U.S. Agricultural Research and Marketing Act (1946)</td>
<td>Requires collecting, preserving, and disseminating genetic material important to U.S. agriculture</td>
</tr>
<tr>
<td>Foreign Assistance Act (1962)</td>
<td>Requires that U.S. foreign aid to developing nations helps to protect biodiversity and tropical forests, and that development projects generally avoid adverse impacts on biodiversity</td>
</tr>
<tr>
<td>Marine Mammal Protection Act (1972)</td>
<td>Creates a moratorium on importing marine mammals and limits taking of U.S. marine mammals</td>
</tr>
<tr>
<td>Endangered Species Act (1973)</td>
<td>Prohibits or restricts taking of endangered or threatened species, requires federal agencies to protect their critical habitat, and regulates their trade; requires evaluation of the impact of federal projects on endangered and threatened species</td>
</tr>
<tr>
<td>Fishery Conservation and Management Act (1976)</td>
<td>Requires plans from state/federal councils to manage fisheries in a 200-mile offshore exclusive economic zone to ensure a maximum sustainable yield</td>
</tr>
<tr>
<td>National Forest Management Act (1976)</td>
<td>Requires management plans for all national forests, including conservation of a diversity of plant and animal communities</td>
</tr>
<tr>
<td>Wild Bird Conservation Act (1992)</td>
<td>Limits imports of exotic birds to protect wild populations in country of origin and reduce inhumane treatment of birds</td>
</tr>
<tr>
<td>National Wildlife Refuge System Improvement Act (1997)</td>
<td>Requires that National Wildlife Refuge System conserve fish, wildlife, plants, and their habitat while maintaining the biological integrity, diversity, and environmental health of the system</td>
</tr>
</tbody>
</table>

*This is the year legislation was first passed. All laws shown have been amended subsequent to passage. The stated purpose may include objectives added in the subsequent amendments and/or may not include all the purposes of the law.*
agree to take specific actions at the national level. Although the earliest international treaties reflected concerns that one nation might overexploit a species that many nations found economically important, recent treaties have acknowledged the importance of conserving all levels of biodiversity, whether or not there is an immediate economic value. Ratification of a treaty by each country is voluntary, and enforcement is often difficult. The United States is signatory to many international treaties relating to biodiversity, some dating back to the early 1900s.

A. International Convention for the Regulation of Whaling (1946)

The major nations then involved in whaling (including the United States, Japan, and the Soviet Union) were party to the 1946 Convention to Regulate Whaling, which prohibits killing calves, immature whales, or females accompanied by calves and encourages using as much of the killed whales as possible. The convention created the International Whaling Commission and gave it the authority to designate protected species and to establish hunting methods and seasons, size limits, and take limits. The convention had the conflicting goals of both conserving whale stocks and further developing the whaling industry. Most species of whales continued to decline dramatically until a 10-year moratorium on commercial whaling was agreed to by signatory countries in 1986. By that time, only Japan, the Soviet Union, and Norway were still whaling. The moratorium is still in effect, although it is not adhered to by Norway, and many countries continue to take whales for the “research purposes” allowed under the convention.


The Convention on International Trade in Endangered Species (CITES), which was introduced in 1973 and entered into force in 1975, prohibits international trade in endangered species by assigning each protected species to one of three lists. Appendix I lists “all species threatened with extinction which are or may be affected by trade.” They cannot be traded primarily for commercial purposes, and trade in the species cannot be detrimental to the species survival. Trade in Appendix I species requires both import and export permits approved by the “management authority and scientific authority” of the nations involved (the secretary of interior in the case of the United States). The permit has to certify that the specimen was obtained legally, that it can be shipped without risk of cruel treatment or harming the health of the specimen, and that the trade is not detrimental to survival of the species.

Species listed in Appendix II are not necessarily currently threatened, but unregulated trade could jeopardize their survival. Two-thirds of the parties to CITES must agree in order to add or remove a species from Appendix I and II lists. Any nation can unilaterally add or remove a species from Appendix III if it believes that cooperation from other nations is needed to control trade in a vulnerable species under their jurisdiction. Only export permits are required for Appendix II and III species. By 1999, more than 140 countries had ratified the treaty.

C. Convention on Biological Diversity (1993)

The International Union for the Conservation of Nature began encouraging the creation of a global biodiversity treaty in the early 1980s. The United Nations Environment Program led the process of creating the convention, which was agreed to in 1992 and entered into force in 1993. By 1998, 174 countries had ratified the treaty. The United States signed the Convention on Biological Diversity (CBD) in 1993, but it has not yet been ratified by two-thirds of the Senate. Until ratification, federal executive agencies are expected to conform with the treaty to the extent possible, but it is not recognized as law in federal and state courts.

The CBD outlines objectives for the conservation, sustainable use, and equitable sharing of the benefits from biodiversity. Signatory countries are called on to develop national strategies for protection of biodiversity, establish a system of protected areas, conduct environmental assessments of projects in order to prevent adverse impacts on biodiversity, enact laws regulating conditions of access to genetic resources, and support a financial mechanism for developing countries to obtain grants to assist in meeting the terms of the treaty. The convention supports conservation of agrobiodiversity (the plants and animals that contribute to food security for humans) as well as wild biodiversity. The importance of habitat conservation and of controlling alien species is acknowledged in the treaty.

Some of the most discussed provisions of the treaty (and the basis for the U.S. reluctance to ratify it as of 1999) are those concerning access to genetic resources and the “intellectual property rights” resulting from
commercial use of genetic resources. Countries with large biotechnology industries (e.g., pharmaceutical and agribusiness) supported continued open access to genetic resources, whereas countries supplying biodiversity to industry wanted to share the profits to help ensure continued protection of biodiversity. The treaty both acknowledges each nation’s sovereign control of its biodiversity and encourages access to genetic resources by other contracting parties.

III. NATIONAL LEGISLATION

Most national laws discussed here are primarily concerned with actions influencing biodiversity in the United States. A few, including the Foreign Assistance Act, legislate how actions of the U.S. government relate to conservation in other countries. The U.S. states were solely responsible for conserving wildlife until states urged the federal government to pass the Lacey Act in 1900. States still retain primary responsibility for biodiversity. The role of the federal government is limited to managing federal lands; conserving economically important genetic resources, protecting migratory waterfowl, marine species, birds of prey, and endangered and threatened species; and enforcing international treaties. The following discussion describes U.S. laws supporting conservation at the genetic, species, and ecosystem levels according to their primary purpose. Such divisions are arbitrary to a degree since genes are best conserved in the organisms in which they exist, and species depend on their habitats for survival.

A. Conservation of Genetic Resources

Federal laws regarding genetic resources are primarily concerned with conservation of varieties of economically important domesticated plants and animals, with some attention to “wild genetic resources,” e.g., wild relatives of plants or animals known to be economically important. For wild species, legislation is more concerned with access to genetic resources and sharing of benefits if those resources end up being economically valuable. In 1990, an amendment to the 1946 U.S. Agricultural Research and Marketing Act modified the National Genetic Resources Program in the Agricultural Research Service in the Department of Agriculture. The purpose of this program is to collect, preserve, and disseminate genetic material that is important to American agricultural production. This amendment expanded the types of genetic resources in the program to include silvicultural species, animals, and aquatic and microbiological organisms. Congress authorized creation of the National Seed Storage Laboratory in Colorado by the U.S. Department of Agriculture. Although this gene bank is world renowned for the variety of agricultural species it contains, it has been criticized for inadequately maintaining seed from wild relatives of important crops.

B. Conservation of Species

The earliest federal wildlife legislation in the United States was concerned with species, both migratory and nonmigratory, with ranges that went beyond state boundaries. Most early legislation emphasized regulation of consumptive uses of species, such as hunting and fishing. By the 1970s, Congress was aware habitat degradation is a significant threat to biodiversity; this is reflected in the habitat conservation provisions of the Endangered Species Act and in the increase in laws concerning habitats described in Section III.C.

1. Lacey Act of 1900

The Lacey Act was created to strengthen the ability of states to protect wildlife by regulating interstate commerce in game birds and wild birds. The Lacey Act now also prohibits importing all wildlife and some plants that were taken contrary to the laws of another country. The original act also aimed to protect U.S. agricultural and horticultural interests by prohibiting the import of injurious foreign wildlife, including mongooses, English sparrows, and other birds and animals. Later amendments broadened the definition of injurious wildlife to include wild birds, wild mammals, fish, amphibians, and reptiles and their eggs, and the protected interests were also expanded to include humans, forestry, and U.S. wildlife resources. The 1981 amendments also increased the maximum penalties and lowered the standard of proof for a violation, making the Lacey Act one of the laws most frequently used by federal law enforcement officials to protect wildlife (Bean and Rowland, 1997).

2. Laws Controlling the Introduction of Nonnative Species

One of the major threats to biodiversity is the introduction of nonnative species to new habitats, intentionally or by accident. The federal authority to regulate interstate commerce led to creation of the Federal Noxious Weed Act of 1974 and the Nonindigenous Aquatic Nuisance Prevention and Control Act of 1990. The Noxious Weed Act gives the secretary of agriculture authority to regulate both foreign imports and interstate com-
merce in potentially harmful plants, defined as plants of foreign origin that can directly or indirectly injure agriculture, natural fish and wildlife resources, or public health. Amendments in 1990 give authority to the Fish and Wildlife Service, the National Park Service, and the Bureau of Land Management to control undesirable plants on their lands.

The accidental introduction of the economically destructive zebra mussel into U.S. waters when a ship's ballast water was released led Congress to create the Nonindigenous Aquatic Nuisance Prevention and Control Act in 1990. This act covers any plant, animal, or other "viable biological material," such as a virus, that disperses to an aquatic ecosystem in which it is not historically found. In contrast to the Lacey Act provisions, a nonindigenous species does not have to be from a foreign country. To be termed a nuisance, a nonindigenous species must threaten the abundance or diversity of native species or the ecological stability or commercial productivity of the infested waters. The act mandates creation of a task force to implement a program to prevent the introduction and dispersal of aquatic nuisance species, but it does not specify who determines which species are a nuisance (Bean and Rowland, 1997).

3. Birds

a. Migratory Bird Treaty Act of 1918

This act was passed to implement a treaty signed in 1916 between the United States and Great Britain with the aim of conserving birds that migrate between the United States and Canada. The treaty establishes closed seasons for hunting and prohibits taking of nests or eggs (except for scientific purposes) for three groups of migratory birds (game, insectivorous, and nongame—a classification that includes almost all birds). The act also implements treaties signed with Mexico in 1936, Japan in 1972, and the Soviet Union in 1976. The latter two treaties include provisions for protecting bird habitat, for example, from pollution. Numerous cases have resulted in judicial interpretations that "taking" under the treaty act does not include habitat destruction. The act was successful in supporting the 1991 requirement that migratory waterfowl not be hunted with lead shot. The courts have rejected landowner claims that the hunting prohibitions under this act are unlawful "takings" of their private property (Bean and Rowland, 1997).

b. Bald Eagle Protection Act of 1940

The Bald Eagle Protection Act prohibits killing or possessing bald eagles, their nests, or any part of the eagle. The secretary of interior could approve exceptions for scientific and educational purposes. A 1962 amendment added the same protections for golden eagles and authorized the secretary to permit taking of bald and golden eagles by Indian tribes only for religious purposes. A 1972 amendment of the act expanded the definition of prohibited taking to include poisoning.


The United States is the largest importer of wild birds in the world. It is estimated that as many as 50% of captured exotic wild birds die before they even leave the country of origin, and another 15% die in transit to the United States (Bean and Rowland, 1997). The goal of the Wild Bird Conservation Act is to promote conservation of birds not indigenous to the United States by (i) limiting or prohibiting imports of nonindigenous birds in order to reduce depletion of wild populations and reduce inhumane treatment of birds in transit, (ii) ensuring that trade is biologically sustainable, and (iii) assisting in wild bird conservation programs in the country of origin. Congress found that CITES was not effective in decreasing the rate of loss of wild bird populations because many signatory exporting countries were unable to implement it adequately. The act allows the secretary of interior to declare a moratorium on importing any exotic bird listed on any CITES appendix and to permit exceptions for scientific purposes.


The Marine Mammal Protection Act (MMPA) replaced individual state authorities and programs concerning marine mammals and created a moratorium on importing marine mammals to the United States and on taking of U.S. marine mammals. Taking was defined to include attempts to harass, hunt, capture, kill, and (after later amendments) feed marine mammals in the wild. Responsibility for implementing the act rests with the secretary of commerce for (i) incidental take of mammals during commercial fishing, (ii) all Cetacea (whales and porpoises), and (iii) all Pinnipedia (seals) except for walruses. The secretary of interior implements the MMPA for all other marine mammals, including manatees, polar bears, sea otters, and walruses. Like most environmental legislation, the MMPA reflects a compromise between many interest groups: commercial interests wanting to protect an economic product, scientists who valued marine species and their role in marine ecosystems, and animal welfare groups wanting to protect marine mammals for their individual and intrinsic value.

The MMPA outlines management principles that in-
clude maintaining the health and stability of marine ecosystems and the "optimum sustainable population" for all marine mammals as well as the "maximum sustainable yield" for commercially exploited species. The taking prohibitions allow many exceptions; one that received the most attention was the controversial allowance for "incidental taking" during commercial fishing operations. At the time the act was passed, more than 5 million dolphins had died accidentally by drowning in purse seining nets used to catch tuna (Bean and Rowland, 1997). Beginning in 1976, a series of regulations set quotas on how many dolphins could be killed incidentally by U.S. tuna fishing boats. In 1984, amendments sought to ensure that countries exporting tuna to the United States had kept incidental dolphin kills as low as those by the U.S. fleet; otherwise, a moratorium could be imposed on importing tuna from that country.

The 1992 International Dolphin Conservation Act required no import to the United States after 1994 of any tuna caught in association with incidental dolphin mortality (Bean and Rowland, 1997). This and similar laws have been challenged by other nations for inhibiting the free trade agreed to under the General Agreement on Tariffs and Trade.

5. Endangered Species Act of 1973

The 1973 Endangered Species Act (ESA) has a broad mandate to restrict the taking of species that are at risk of becoming extinct, protect and acquire the habitat necessary for those species to survive, regulate their trade, and force all federal agencies to evaluate the impacts of their activities on endangered species and to avoid jeopardizing a species' continued existence. This act was intended to enable the recovery of the populations of species it protects.

An endangered species is defined by the act as one that is in danger of extinction throughout all or a significant portion of its range. A threatened species is considered likely to become endangered in the foreseeable future throughout all or a significant portion of its range. The act outlines detailed procedures for the secretary of interior or commerce to follow in listing or delisting a species as threatened or endangered. The decision must be based on objective, scientific information on the biological status of the species and not on its economic or other value. Until a species is listed, it does not receive most of the protections of the act and the Department of Interior has a large backlog of species in the listing process.

The ESA outlines many duties and prohibitions that apply to all people subject to U.S. laws (in Section 9) and other duties that apply only to federal agencies (in Section 7). It is illegal to take any endangered or threatened animal species; take means to harm, harass, pursue, hunt, shoot, wound, kill, trap, or capture. Regulations further define harm to include habitat degradation to the point that wildlife is injured or killed. Although the original act did not prohibit taking of plants, later amendments prohibited taking plants on federal lands and prohibited removing or destroying plants on nonfederal lands where such action violated other laws.

Federal agencies are prohibited by Section 7 from approving, funding, or directly undertaking actions that are likely to jeopardize the continued existence of an endangered or threatened species or result in the destruction or adverse modification of its designated critical habitat. These agencies must consult with the Department of Interior or Commerce to determine the potential for jeopardy to the species or harm to its critical habitat and reasonable ways to avoid such harm. There is no direct legal obligation under ESA for nonfederal agencies or persons to avoid adverse modification of critical habitat. After the Supreme Court ruled that the Tellico Dam could not be completed due to the threat to the endangered snail darter (a fish), Congress amended ESA to allow a process for seeking exemption from the act's restriction on federal agencies.

Section 7 also requires designation of critical habitat for species listed as endangered or threatened. Critical habitat is defined as an area critical to the survival of the species; this could include a portion of the area a listed species occupies or an area beyond currently occupied habitat. Although the requirement to protect critical habitat has been applauded by many, it has been difficult to implement and designations of critical habitat by the Department of Interior or Commerce have been rare. The requirement has resulted in a great deal of litigation brought both by those challenging critical habitat designations and by those challenging the agencies for failing to make them.

The ESA can be enforced by the government through civil and criminal penalties and by citizens through citizen suits provided for in the act. Those who knowingly violate the ESA for an endangered species can be fined up to $50,000 and put in prison for up to 1 year. For a threatened species, a violation can result in a $25,000 fine and prison for 6 months. Additional civil penalties can be imposed. States can participate in enforcement of ESA by signing cooperative agreements with the secretary of interior and are also eligible for federal financial assistance for enforcement.

An international mandate is also included in the ESA. The act directs the secretary of interior to encour-
age other countries to protect endangered species and authorizes financial support and assistance from U.S. wildlife officers for such programs. Implementation of CITES is also authorized by ESA.

The ESA is one of the most controversial federal environmental laws passed during the past century. Its very specific, substantive, and legally binding provisions make it a target for continuous challenge from a variety of perspectives. Challenges have come from those who argue about whether ESA really helps protect species, inhibits use of private land, or applies to specific federal and private actions. The act forces attention to solving the problem of threatened and endangered species, with its wider attention to habitat conservation. This focus on solutions has forced discussion of how humans develop and use broad landscapes for economic production in areas with important biodiversity. More procedurally oriented laws are able to avoid controversy by not requiring that threats to biodiversity be addressed.

6. Fishery Conservation and Management Act (Magnuson Act) of 1976

Beginning in the early 1900s, international agreements between nations fishing in the same waters attempted to establish control of exploitation of ocean fish; by the 1970s the United States was a party to approximately 20 such agreements (Bean and Rowland, 1997). While waiting for agreement on a single international convention on fishing, the United States passed the Fishery Conservation and Management Act in 1976. The act established a 3-mile-wide territorial sea off the coast of the United States and a 200-mile-wide exclusive economic zone (EEZ). Within the EEZ the United States claimed exclusive management authority over fish and all other forms of marine animal and plant life except birds, mammals, and tuna. Tuna was added to this authority in 1990 amendments. The United States also claimed authority to manage certain sedentary species found in places where the continental shelf extends beyond 200 miles from shore.

The act required eight Regional Fishery Management Councils, which include state and federal officials, to develop comprehensive plans for conserving and managing the fisheries off their coasts. These plans have to comply with standards that include preventing overfishing and ensuring a maximum sustainable yield. The 1996 reauthorization of the act created new standards for minimizing “bycatch” (the accidental harvest of fish species not intended for harvest) and required that management plans try to minimize adverse impacts on fish habitats. Implementation authority for the act lies with the National Marine Fisheries Service in the Commerce Department. An important and controversial issue is the extent to which international trade agreements, such as the General Agreement on Tariffs and Trade or the North American Free Trade Agreement, have the potential to conflict with, and undermine, national laws aimed at protecting biodiversity.

C. Conservation of Habitats and Ecosystems

During the past 30 years, legislators, administrators and the general public have come to understand that species are not conserved without their habitats and ecosystems. The ESA reflected this by requiring conservation of critical habitat. Implementation of ESA has increasingly focused on habitat conservation plans as a way to conserve species. Many recent laws have focused on conserving lands and ecosystems in their entirety. These laws are not meant to replace species-oriented laws but rather to complement them. One of the challenges in dealing with ecosystem conservation is that ecosystems are naturally unstable and change continually; it can be difficult to separate natural background rates of change from changes that result from human disturbance.

One important and threatened habitat type, wetlands, received protection through the Clean Water Act of 1972. The dredge and fill permit program in Section 404 of the act is a significant force in protecting aquatic ecosystems by regulating the physical alteration of “waters of the U.S.,” including wetlands, estuaries, and streams.

1. Federal Lands

Legislation and regulations addressing conservation of habitats and ecosystems in most countries focus primarily on creation of protected areas on publicly owned lands. On average, less than 5% of a country becomes legally designated as protected, and enforcement of that limited area of protection is often weak. Many protected areas allow some consumptive use of plant and animal species. An increasing number of countries have legislation that requires special permits to clear native vegetation or special habitat types outside of protected areas.

In the United States, approximately 30% of the land area (more than 700 million acres) is owned by the federal government and managed primarily by four agencies. Congress has enacted separate laws outlining how each agency should manage its lands. Historically, public land management agencies have given high priority to production of a commodity, such as timber, grazing land, or waterfowl production, and put less
emphasis on maintaining the health of an ecosystem and its biodiversity. The 1964 Wilderness Act authorized Congress to designate portions of any federal land as wilderness areas, in which new commercial enterprises and permanent roads are prohibited and natural ecological processes are supposed to be allowed to occur without interference. During the early 1990s, federal agencies were directed to begin using “ecosystem management” to administer their lands. A level of production of a particular biological resource is to be considered in light of the health of the entire ecosystem.

Only the National Wildlife Refuge System (created in 1966) and the National Marine Sanctuary System (created in 1972) have as primary goals the conservation of wildlife and marine resources, respectively. The National Marine Sanctuary Act of 1972 authorizes the secretary of commerce to designate sanctuary areas to “maintain, restore, and enhance living resources” by providing critical habitat for endangered species and areas where ecosystem structure and function are maintained. The National Wildlife Refuge System Improvement Act of 1997 consolidated refuge-related authorities and stated the mission of the National Wildlife Refuge System. The system’s mission is to administer a national network of land and water “for the conservation, management, and restoration of fish, wildlife, and plant resources and their habitats” and to maintain the system’s biological integrity and diversity.

a. National Park Service Act (1916)

The National Park Service Act outlines the goals and administrative guidelines for conserving the scenery, natural and historic objects, and wildlife in the national park system (which includes parks, monuments, preserves, wild rivers, and lakeshores) in a manner that does not impair those resources for future generations. There have been many instances in which the National Park Service has had to balance the (occasionally) conflicting goals of wildlife management and human recreation. Amendments to the act in 1978 provided the secretary of interior with some authority to protect park resources from outside threats, such as logging on private lands adjacent to parks. Courts have also upheld National Park Service authority to prohibit or regulate fishing, trapping, and hunting within park lands.

b. National Forest Management Act of 1976

National forest lands began to be set aside in the late 1800s, with the purpose of protecting water supplies and timber production. The 1960 Multiple-Use Sustained-Yield Act indicated that the purpose of national forests is to provide for outdoor recreation, grazing, timber, watersheds, fish, and wildlife. The 1976 National Forest Management Act (NFMA) required the creation of management plans for each national forest. The plans are required to integrate physical, biological, and economic issues in order to provide for a diversity of plant and animal communities, including viable populations of native vertebrates. The plans are supposed to outline how timber harvests will mitigate impacts on biological diversity.

One of the most famous ecosystem management plans resulting from this legislation was for the old-growth forest in the Pacific Northwest. The Forest Service, after several court cases, was forced to develop a plan for old-growth forest that managed for a diversity of plant and animal species, including the spotted owl. Once the spotted owl was listed under the ESA, the management plan had to comply with both ESA and NMFA.

2. Private Land

Individuals and institutions privately own approximately 60% of the U.S. land area. Many of the previously discussed laws, such as the ESA and Section 404 of the Clean Water Act, aim to regulate actions with adverse impacts on biodiversity on private lands. Other federal laws aim to create positive financial incentives for conservation on private lands.

The 1985 Food Security Act included creation of a Conservation Reserve Program with the goal of encouraging farmers to voluntarily remove marginal lands from intensive agricultural production and plant them in perennial grass or tree cover for at least 10 years. A farmer enters into a contract to keep land in the reserve program for 10–15 years in exchange for cash or commodities. In addition, the farmer might receive up to 50% of the cost of certain conservation practices, such as creating wildlife corridors.

Biodiversity on private lands can also be conserved through donation or purchase of conservation easements. An easement is an interest in land that restricts the owner’s use of the property in some way. The easement holder, usually a private interest or a government, has the right to enforce the restriction in perpetuity or for an agreed on time period. Since the 1970s, the Fish and Wildlife Service has been negotiating easements on wetlands as waterfowl production areas. The 1990 Food, Agriculture, Conservation and Trade Act created a Wetlands Reserve Program through which the government purchases easements on wetlands from farmers, who then restore and protect the functional values of the wetland. The farmer is eligible to receive a share of the costs of implementing the land management plan
for the easement. A similar Environmental Easement Program was created at the same time. Under this program the government purchases easements on areas critical for wildlife on farms and ranches. The Forest Legacy Program, created in 1990, allows the secretary of agriculture to purchase conservation easements on forest areas that are threatened with development for other uses.

The 1969 U.S. Internal Revenue Code began permitting landowners to take a charitable tax deduction for qualified conservation easements, thus creating an incentive for private landowners to protect their lands. Protection of natural habitat for plants and animals is one conservation purpose that qualifies an easement for tax deduction. This legislation has led to a dramatic increase in donation of conservation easements on private land to nonprofit land conservation organizations.

The 1985 and 1990 Farm Bills contained a strong incentive, in the “swampbuster” provisions, for farmers to maintain wetlands and not convert them to agricultural production. Anyone who drained, dredged, filled, or altered a wetland in order to produce an agricultural product would lose their eligibility for federal price supports, crop insurance, the Conservation Reserve Program, and all other federal benefits for all of their land and agricultural products. The 1996 Farm Bill weakened this incentive, but swampbuster still serves to discourage conversion of wetlands to agricultural production.

D. Other National Laws

1. National Environmental Policy Act of 1969

The most significant provision of the National Environmental Policy Act (NEPA) for biodiversity is the requirement that federal agencies prepare an Environmental Impact Statement (EIS) for every federal action or major legislative proposal that can have a significant impact on the environment. The EIS is expected to address potential impacts of proposed programmatic or site-specific actions on the environment, including biodiversity. An important feature of NEPA is the requirement that there be a public comment period on the draft EIS, and that public concerns be addressed in the final EIS. Most federal agencies evaluate the effects of their programs and projects on threatened and endangered species and sensitive habitats. However, many EISs have given inadequate attention to impacts on less threatened species, to overall habitat impacts, and to cumulative impacts on biodiversity.

2. Foreign Assistance Act of 1962

The 1962 Foreign Assistance Act (FAA) authorized the creation of the U.S. Agency for International Development (USAID) and had little to say about environmental issues. Many of the areas of highest species diversity are located in developing tropical nations with insufficient economic resources to adequately protect biodiversity. As scientists and U.S. citizens became aware of the global loss of biodiversity in the late 1970s, they began to lobby the U.S. Congress to amend the FAA to ensure that U.S. foreign aid to other countries helped to conserve biological diversity and tropical forests. Beginning in 1982, amendments to the FAA have required that the United States assist developing nations with protection of habitats in protected areas, antipoaching measures for species, and research on biodiversity. Many of the most important and threatened national parks in developing countries now receive a significant portion of their operating budgets from bilateral assistance provided by developed nations.

Beginning in 1979, Section 216 of the FAA required that all U.S.-funded development projects conduct environmental impact assessments to ensure that tropical rain forests and other biologically important habitats were not adversely impacted. USAID is also required to monitor the impacts of other multilateral donor projects, particularly those funded by the World Bank, and to provide input to the U.S. representative to the World Bank about projects of environmental concern.

IV. STATE AND LOCAL LEGISLATION

The brevity of the following description of state and local laws does not imply that state and local legislation and regulations are unimportant to conservation of biodiversity. In fact, the opposite is true. The following brief sections merely skim the surface of the laws for 50 states and thousands of municipalities in the United States.

A. State Laws

All of the international and national laws and regulations outlined previously have their impacts at local levels. Primary responsibility for stewardship of biological resources is reserved to states by the 10th Amendment to the U.S. Constitution, which states that all powers not given to the federal government by the Constitution are retained by the states. Federal laws, such as those described previously, define minimum
standards for the states to follow. States can, and sometimes do, pass laws that are more restrictive than federal laws in protecting plant and animal species, wetlands, forests, and other ecosystems.

States own their wildlife resources and generally attempt to manage them to benefit all citizens. Historically, most states have focused their wildlife conservation activities on a relatively small number of hunted or fished species. States retain authority to regulate and legislate many issues related directly to conservation of biodiversity; for example, managing the natural resources on state-owned lands, setting hunting limits on deer and other resident, nonmigratory species, and approving permits for development and other projects that affect many habitats and species. Perhaps the largest impact of state laws and local ordinances on biodiversity relates to how well the laws regulate the approval of economic development plans so as to protect natural habitats and mitigate the impacts of development. Many states also require that municipalities develop comprehensive plans, which might address the conservation of biological resources along with land use and development issues.

Only a small number of states have passed laws and regulations specifically addressing conservation of genes, species, and ecosystems, although the number is increasing. In 1992, the state of Michigan passed a Biological Diversity Conservation Act that required creation of a state biodiversity conservation strategy. A 1993 statute in New York State established a Biodiversity Research Institute with a mandate to coordinate state and private efforts to collect information about the state's biodiversity. The law also directed the New York Department of Environmental Conservation to implement programs to conserve rare and endangered species on state lands. In 1996, the state of Wisconsin passed a bill that revised the state's forest statute to ensure that management of state forests is consistent with maintaining native biodiversity and sustainable ecosystems. Many states have statutes that authorize creation of conservation easements on private lands. Hawaii created a statutory tax incentive that provides owners of forest or water reserve lands with relief from all property taxes in exchange for allowing the state to manage their lands (Defenders of Wildlife, 1996). Hawaii also has strong statutes to control the introduction of exotic species.

Many states, including California and Tennessee, have created executive orders and memoranda of understanding regarding conserving biodiversity, but these do not have the force of law. Many other states have policies aimed at conserving biodiversity by creating land acquisition programs for areas with important biodiversity, prohibiting introduction of exotic plant and animal species, and requiring environmental impact assessments in order to minimize and mitigate the impacts of development projects on biodiversity.

Many states are involved in legislatively mandated, regional initiatives to conserve ecosystems that cross state boundaries. The 1987 Chesapeake Bay Agreement was signed by Virginia, Pennsylvania, Delaware, Maryland, the District of Columbia, the Chesapeake Bay Commission, and the U.S. Environmental Protection Agency. The signatories agreed to work on a program to conserve the biological resources and water quality of the Chesapeake Bay system.

### B. Local Ordinances

Local laws can be enacted by municipalities, towns, cities, counties, townships, or boroughs, depending on their regulatory authority and on what local government units are called in different parts of the United States. Although ultimately all biodiversity is conserved or lost at a local level, rarely are local laws primarily concerned with biodiversity per se; rather, they tend to focus on biological resources that indirectly influence public health, safety, or welfare (e.g., protecting wetlands or wooded slopes to control floods or regulating hunting for safety or to maintain economic benefits). Increasingly, however, concerned citizens have been working to ensure that local laws do consider conservation at some scale. Local governments can address biodiversity through broad municipal comprehensive planning and through municipal ordinances that address specific human activities that affect biodiversity on a more site-specific level.

Whether required by state law or not, many municipalities create comprehensive plans that address future land uses in the context of economic growth and quality of life, including conservation of biological resources. Many comprehensive plans include sections on natural or environmental resources. Some municipalities are beginning to prepare separate plans for "open space"—undeveloped land that often contains natural habitats. Most open space plans identify specific environmentally important habitats and recommend specific conservation strategies for those habitats.

Municipalities have adopted a variety of ordinances intended to minimize impacts on species and natural habitats; often, these ordinances support the resource conservation goals established through comprehensive and/or open space plans. An increasing number of municipalities are enacting ordinances that offer stronger
protection to stream and wetland ecosystems than the requirements of state and federal law. Some municipal ordinances focus on integrating protection and restoration of habitats with land development project approvals. For example, some local ordinances have required that development projects maintain or create natural vegetated buffers of more than 25 ft along wetland boundaries and more than 75 ft along each side of a stream.

An increasing number of municipalities have passed forest conservation ordinances that limit development in woodlands and penalize developers who harm those habitats. Some municipalities will not approve logging permits for sensitive woodlands without implementation of a forest management plan that demonstrates sustainable practices such as leaving buffers along streams, restricting logging to seasons when birds are not nesting, or leaving dead trees standing for wildlife habitat. Forest conservation ordinances have required developers to demonstrate that their plans will minimize unnecessary forest fragmentation and will not exceed the percentage of any woodland allowed to be cut. Plans that exceed the legal cutting percentage might only be approved if the developer mitigates the impact through reforestation or off-site forest conservation. Some ordinances seek to minimize tree removal by setting a limit on clearing around buildings and roads (e.g., 25 ft).

Many local municipalities in agricultural areas have ordinances that prohibit growing of introduced "weeds" that might decrease the productivity of agricultural crops. In some residential municipalities, weed laws have been replaced with natural landscaping ordinances that promote and govern natural alternatives to lawn grasses, such as wildflower meadows and reforestation.

V. CONCLUSIONS

The evolution of government legislation and regulation of biodiversity during the past 100 years reflects both the dramatic increases in human impacts on biodiversity in every part of the earth and the increased understanding of the structure and function of the natural world. Legislation has evolved from individual nations focusing on single species of economic value to multinational agreements aimed at protecting entire habitats and ecosystems. The three greatest current threats to biodiversity are habitat alteration and destruction, overexploitation of species, and introduction of exotic species; the laws summarized previously address these threats to varying degrees. Despite continuous significant improvements in legislation and regulations aimed at conserving biodiversity in the face of economic development, genes, species, and habitats continue to be lost at an increasing rate.

There are many challenges to the effectiveness of the laws and regulations described in this article. Perhaps the greatest challenge is ensuring that existing laws are implemented as intended and then enforced. Many laws require scientific information on the status of species and ecosystems; such information is often unavailable or extremely costly to obtain. In the United States, the federal agencies implementing laws such as the ESA are frequently so underfunded by Congress that they cannot achieve their conservation goals. It is also a challenge to ensure that state and local agencies receive the funding required to implement conservation laws. Federal, state, and local implementing agencies are often pressured by special-interest groups to compromise conservation mandates in favor of short-term economic gains. The laws in many countries go unenforced due to lack of oversight from private citizens and advocacy groups. Finally, most laws are based on prohibiting actions that harm biodiversity; there is still a need for more legislation that creates positive incentives for conserving and enhancing biodiversity. The need for positive incentives is particularly acute for private landowners, who control 60% of the land in the United States.

Although the laws and regulations regarding conservation of biodiversity have generally improved significantly during the past 30 years, it is important to acknowledge that they can be amended and weakened whenever the various levels of government, from international to local, choose to do so. In addition, many laws and regulations in other sectors create incentives to destroy biodiversity, such as logging subsidies, underpriced grazing rights, subsidies for economic development, and international trade policies.

Acknowledgments

Thanks to Jan Goldman-Carter, M. Lynne Cohn, Michael Hill, and David Harper for reviewing drafts of this manuscript prior to publication.

See Also the Following Articles

CONSERVATION EFFORTS, CONTEMPORARY • CONSERVATION MOVEMENT, HISTORICAL • ENDANGERED ECOSYSTEMS • PROPERTY RIGHTS AND BIODIVERSITY

Bibliography

GRASSHOPPERS AND THEIR RELATIVES

Piotr Naskrecki
University of Connecticut

I. Overview of Orthoptera
II. Tettigonioidae (Haglids and Katydids)
III. Grylloidae (Crickets and Mole Crickets)
IV. Gryllacridoidae (Camel Crickets, Cave Crickets, Jerusalem Crickets, Raspy Crickets)
V. Eumastacoidea (Monkey Grasshoppers or Eumastacids)
VI. Pneumoroidea (Bladder Hoppers or Pneumoridae)
VII. Acridoidea (Grasshoppers and Locusts)
VIII. Tetrigoidae (Pygmy Grasshoppers or Grouse Locusts)
IX. Tridactyloidea (False Mole Crickets and Sandgropers)

GLOSSARY

brachypterous Having tegmina and wings shorter than the abdomen but overlapping or touching each other on the dorsum.
cerri Paired, usually not segmented structures at the end of the abdomen, sometimes used by males during mating to grasp the female’s abdomen.
hemimetabolous Having incomplete metamorphosis, that is, showing gradual change from molt to molt, with externally developing wing pads, and lacking any larval and pupal stages.
hypognathous Position of the head when the mouth is directed toward the ventral side of the body.
kHz A unit of frequency equal to 1000 hertz; the number of hertz (abbreviated Hz) equals the number of cycles per second (e.g., sound waves).
macropterous Having wings that are fully developed, reaching or exceeding the end of the abdomen.
micropterous Having wings that are greatly shortened, not overlapping or touching on the dorsum.
prognathous Position of the head when the mouth opening is directed forward.
spermatophore A membranous package containing sperm that is transferred from the male to the female during copulation.
stridulatory apparatus An organ of sound production based on the mechanism of rubbing one part of the body against another.
tegmina (singular: tegmen) Thickened forewings.

ORTHOPTERA IS THE ORDER OF INSECTS that includes grasshoppers, locusts, crickets, katydids, and their relatives. The name of the order is derived from the Latin word orthos (straight or rigid) and the Greek word pteron (wing), referring to the simple venation and position of their wings, which are always held straight and parallel to the main axis of the body. Most orthopterans are large or medium-sized insects. Body
lengths of less than 10 mm are uncommon, while many exceed 50 mm in length, with some having bodies over 100 mm long and a wingspan of 200 mm or more.

I. OVERVIEW OF ORTHOPTERA

A. Introduction

Orthoptera are hemimetabolous insects, with nymphs resembling adult forms in their general appearance but lacking fully developed wings and reproductive organs. Mouthparts of orthopterans are of the chewing/biting type (= mandibulate). The head is hypognathous, rarely prognathous; the antennae are usually long and threadlike, consisting of fewer than ten to several hundred segments. The part of the body immediately behind the head, or the pronotum, is usually large, often shieldlike, and in extreme cases covers a large part (many katydids) or the entire body of the insect (pygmy grasshoppers).

The front and middle legs are cursorial (i.e. adapted for walking), yet in some cases the front pair of legs may be modified for digging (mole crickets, pygmy mole crickets, false mole crickets), or both the front and middle pairs may be modified for grasping (predatory katydids). In some orthopterans (most katydids and crickets) the front legs have tibial auditory organs (the ear). The hind legs of most orthopterans are saltatorial, (i.e. modified for leaping), with large, muscular femora and long, slender tibiae. Some grasshoppers can perform repeated leaps of 2.6 m without any obvious signs of fatigue. This is possible primarily because of the presence in their back legs a protein called resilin. Resilin has superb elastic properties, with 97% efficiency in returning stored energy. This allows for explosive release of energy that catapults the insect, a task impossible with muscle power alone. Certain groups of orthopterans, especially those leading a subterranean life, have lost their ability to jump and their hind legs resemble typical cursorial legs. In some grasshoppers and certain Ensiferans the inner surface of the hind femur is modified for sound production (stridulation).

The wings of orthopterans are either fully developed or reduced to various degrees. Wing polymorphism, or the occurrence of individuals with well-developed and reduced wings within the same species, is not uncommon. The forewings are somewhat thickened, forming leathery tegmina. In most katydids and crickets parts of the tegmina are modified for stridulation. The hind-wings, when present, are fanlike, hidden under the first pair in the resting position. Often the hindwings are longer than the tegmina and protrude behind their apices. Wing buds of nymphal stages are always positioned in such a way that the second pair of wings overlaps the first one, whereas in adult individuals of micro- and brachypterous species, the first pair of wings always overlaps the second pair, despite their nymphal appearance. The base of the abdomen in grasshoppers has lateral auditory organs known as abdominal tympana. Females of most orthopterans have a prominent ovipositor at the end of the abdomen, derived from the eight and ninth abdominal segments. Katydids and crickets usually have a well-developed ovipositor — sword-, sickle-, or needle-shaped — whereas females of grasshoppers and their relatives usually lack a long, external ovipositor.

The number of described species of Orthoptera is approximately 25,000. This number, however, probably represents only a half or less of the actual number of species of Orthoptera present on earth today. Tropical regions of South America, Africa, and Asia still remain virtually unexplored in terms of their grasshopper and katydid fauna, and many thousands of species are expected to be described in the future. The Australian orthopteran fauna is the best studied of all tropical regions of the world, yet more than 1500 Australian species still remain to be formally described despite being already recognized as new to science.

Members of the order Orthoptera inhabit virtually all terrestrial habitats of the world, from rock crevices of the littoral zone of the oceans, subterranean burrows, and caves; to treetops and peaks of the alpine zones of mountain ranges. Both deserts and grasslands as well as dense forests have rich and unique orthopteran faunas. There are few purely aquatic forms, but many are associated with marshes and other semiaquatic habitats. Orthopterans are important members of nearly all terrestrial ecosystems, both in the role of consumers and prey. Massive outbreaks of some species of grasshoppers (and less frequently katydids and crickets) can cause enormous losses for the food industry and forestry. Locusts (a vernacular name for certain grasshopper species that tend to produce large seasonal outbreaks, and not a taxonomic entity) have been a part of human history from the very beginning of our agricultural tradition. They still pose a great risk for agriculture in many parts of the world, although they pose less of a problem now than a few hundreds years ago, thanks mostly to better understanding of their population dy-
namics and application of various chemical and biological control measures.

**B. Sound Production**

The single characteristic most frequently associated with grasshoppers and their relatives is their ability to produce sounds. Although less widespread than generally believed, this ability is nonetheless quite common in some groups of orthopterans. The role of sound production is three-fold and similar in some respects to that of birdcalls: (a) attraction of mates, (b) territoriality, and (c) disbursement of release calls (alarm calls produced when seized by a predator). The calls of orthopterans are usually species-specific and play a very important role in species recognition. The information in the call may be coded in the form of frequency modulation (the pitch of the call changes through time, a mechanism best known in birds), time modulation (the pitch of the call remains the same throughout its duration but its temporal pattern is unique to the species), or both modes combined.

The dominant mechanism of sound production in Orthoptera is stridulation, which involves rubbing one modified area of the body against another. Contrary to popular belief, no orthopterans (nor any insect for that matter) produce sound by rubbing their hind legs against each other. Katydid (Tettigoniidae) and crickets (Gryllidae) produce sound by rubbing a modified vein (the stridulatory vein) of one tegmen against a hardened edge of the second tegmen (the scraper). The stridulatory vein is equipped with a filelike row of teeth, the number of which varies from a few to a few hundred. In most katydids, the stridulatory area is situated at the base of the tegmina, except in brachypterous species where it covers their entire surface. In crickets virtually the entire surface of the tegmina is modified for stridulation. As a rule, in katydids the stridulatory file is situated on the left tegmen and the scraper on the right one, while in crickets the situation is reversed. A membranous area at the base of the tegmen, the mirror, amplifies the sound. In addition, some katydids (e.g. *Thoracistus*) use their enlarged, shieldlike pronotum as an additional sound amplifier. Crickets, lacking the enlarged pronotum, use other methods of sound amplification, such as singing from burrows, the shape and size of which is attuned to boost certain frequencies (*Gryllotalpa*), or using the surface of a leaf for the same purpose (*Oecanthus*). The ability to stridulate is restricted almost exclusively to males, although in a few groups of katydids (Phaneropterinae, Ephippigerinae), females respond to the male’s calls by stridulating as well. Their sound apparatus is not homologous to that of the males and is usually quite simple, lacking the sophisticated mechanism for sound amplification. In addition to terminal sound apparatus, a few groups of katydids have developed other mechanisms of stridulation. For example, all members of the Australasian subfamily Phyllophorinae lack the typical wing stridulation and produce sound, instead, by rubbing their hind coxae against modified thoracic sternae. Mandibular sound production occurs in some members of Mecopodinae.

Grasshoppers use the same principle of stridulation, but instead of rubbing their tegmina against each other, these insects produce sound by rubbing the inner surface of the hind femur against one of the veins of the tegmen. In the slant-faced grasshoppers (Gomphocerinae), the inner surface of the femur possesses a file of small knobs while the vein on the tegmen acts as the scraper. In band-winged grasshoppers (Oedipodinae), the vein has a row of pegs and the femur plays the role of the scraper. In addition to these two principal mechanisms, some grasshoppers stridulate by rubbing their hind legs against the sides of the abdomen (Pamphagidae) or by kicking their legs feet against a modified area at the apex of the tegmen (*Stethophyma*). Australian false mole crickets (Cylindrachetidae) have a stridulatory file at the base of their maxillary palps, and some species of pygmy mole crickets (Tridactylidae) produce sound by rubbing a modified vein on the dorsal side of the tegmen against another vein at the base of the hindwing.

The sound frequencies produced by orthopterans during stridulation vary from a few kHz (most crickets and grasshoppers) to well above 100 kHz (some katydids). Crickets’ calls are characterized by their tonal purity, with most energy of the call allocated within a narrow range of frequencies. Katydid calls vary from tonally pure (although often well above the human hearing range) to broad, noiselike signals. Grasshoppers produce mostly broad spectrum, noiselike calls. Unlike many vertebrate calls, many orthopterans produce time modulated rather than frequency modulated signals. Crickets are a notable exception and most species produce melodious, birdlike, frequency-modulated chirps.

In addition to stridulation, some grasshoppers crepitate in flight. In this case the sound is produced by hitting certain veins of the wings against other veins. This behavior is especially common among band-winged grasshoppers (Oedipodinae) and plays an important role in courtship and territorial displays.
A few members of normally acoustic orthopterans have lost their ability to produce airborne signals and instead have developed a number of substitute mechanisms of substrate-borne communication. Males of the oak katydids (Mecoperna) lack the typical tegmental sound apparatus and instead produce sound by drumming with their hind legs against the bark of trees. Similar drumming behavior, although still accompanied by typical stridulation, is a component of the courtship behavior of the pitbull katydid (Lirometopum). Males of the cricket Phasophila are spectrum have lost their ability to stridulate and instead signal by rapidly flicking their tegmina back and forth while holding them in vertical position. The near-field motion is detected by female’s cerci, rather than her ears. Despite having a fully developed stridulatory apparatus, many neotropical members of the katydid subfamilies Suedophyllinae and Conocephalinae spend little or no time stridulating, relying instead on substrate-borne tremulations. In this case, a male stands rigidly on a leaf or stem of a plant and violently shakes his entire body. The low frequency waves are transmitted along the branches of the plant. The reason for this behavior is unclear, although a few hypotheses trying to explain it have been proposed. The most widely accepted one is the avoidance of predation by foliage-gleaning bats that are known to use insect sounds to locate their prey. Others include eluding satellite males (nonsinging males of the same species trying to intercept a female), avoidance of parasitoid flies, and helping females locate males on multi-branched plants. Thanks to the rapid development of recording techniques in recent years, it is now known that many groups of orthopterans previously believed to be silent appear to employ a number of techniques of substrate communication.

C. Reproduction and Growth

The courtship and mating behavior of orthopterans provide some of the most complex and fascinating spectacles of the insect world. In addition to sound production, as discussed earlier, many species employ visual, tactile, and olfactory signals in their mating strategies. Visual communication is especially well developed in grasshoppers, where males often carry bright, species-specific markings on different parts of their bodies, displayed in carefully choreographed sequences during copulation. Grasshoppers of the genus Syrphus are definite champions in this respect, and males of some species, in addition to calling, perform a dance consisting of 18 distinct movements. Visual signals employed by many diurnal grasshoppers include flight displays, where males flash their colorful hindwings (this is sometimes accompanied by crepitation), flagging with distinctly colored hind legs, and displays involving brightly colored, and often enlarged antennae. Courtship of katydids and crickets relies less on visual signals and more on sound and chemical cues, which are more appropriate for these mostly nocturnal animals. In both groups, males sometimes produce two different types of calls, a long-range advertisement call and a quieter courtship song, performed only in the presence of a female. Female in some species may reply using either airborne signals or tremulation.

Chemical communication in Orthoptera is little studied, but there is evidence that at least some species employ it during courtship. Females of the New Zealand giant weta Deinacrida rugosa produce a musky substance that males use to locate females, while males of camel crickets of the genus Centophyllus possess thoracic glands that may also play a role in courtship. Field crickets Teleogryllus commodus use pheromones covering the female antennae to initiate courtship. Some other crickets use airborne pheromones in locating members of the opposite sex.

Copulation in orthopterans involves transfer of a sperm sac (spermatophore), which in some groups is accompanied by a large packet of nutritious proteins, the spermatothax. Males of some orthopterans also allow females to feed on parts of their own bodies during copulation. Males of bagids (Cyphoderris) have their hindwings modified into thick, fleshy lobes, the sole purpose of which is to be eaten by the female during copulation. Females of tree crickets (Oecanthus) feed on males’ thoracic glands during copulation, and in some crickets of the subfamily Neomobinae the females feed on enlarged spines on males’ hind tibia. Males of other orthopterans, lacking such tasty incentives, must rely on their strong grasp or modified cerci at the end of the abdomen to hold the female during copulation.

Oviposition takes place in a variety of substrates, such as soil, plant tissues, or rock crevices. In some cases, eggs are protected from desiccation by a foamy mass produced by the female. Nymphs usually hatch within a few weeks or months, but sometimes the eggs undergo a year long, or longer, diapause. Few orthopterans display any kind of parental behavior, although some crickets (Anurogryllus) lay eggs in burrows guarded by the female. Female mole crickets (Grillo talpa) not only lay eggs in special egg chambers underground, but also actively care for the eggs by licking and removing fungal spores from their surfaces. The hatchlings stay with their mother for a few weeks before dispersing.
D. Food and Feeding

Orthopterans are extremely diverse in their food preferences and feeding techniques. Virtually all Ensifera (grasshoppers and their allies) are strictly herbivorous, very rarely engaging in cannibalistic behavior, and doing so only under crowded conditions. Most grasshopper species seem to be polyphagous (feeding on a wide variety of plant species), but some are oligophagous (feeding on a narrow spectrum of plant species) or monophagous (feeding on only one species of plants). The last are often associated with toxic, alkaloid-rich plant species, and these substances make the insect themselves inedible to many potential predators. Pygmy grasshoppers (Tettigidae) are some of the few insects that feed on mosses and lichens.

Ensifera (katydids, crickets, and their relatives) range from herbivorous to omnivorous to strictly predaceous. Some katydids specialize on rather unusual food sources. The members of the Australian genus Zaprochilus feed exclusively on the pollen and nectar of flowers. Most katydids, however, feed on a wide range of organic material. For example, the Central American Rhinoceros katydid (Copiphora rhinoceros) is known to eat on flowers, fruits, hard seeds, caterpillars, other katydids, snails, frog eggs, and even small lizards. Strictly predaceous katydids employ both the "sit-and-wait" strategy (Saginae) or actively forage and hunt living insects (Listroscelidinae). Some raspy crickets (Gryllacrididae) also actively search for insect prey by rapidly running along branches and grasping any sitting insect their encounter. Crickets and cave crickets tend to be generalists in their dietary preferences but rarely exhibit tendencies to feed on live prey. Some mole crickets have a behavior unique among orthopterans, and insects in general, of gathering and storing germinating seeds in circular chambers below ground for later consumption.

E. Phylogeny and Taxonomy

Despite a long history of taxonomic research on orthopteroid insects and despite their economic importance, there is little agreement as to the taxonomic position of many groups of the Orthoptera. Unfortunately, the current classification system is still based mostly on 19th-century works. Older classifications also included in this order such groups as cockroaches (Blattodea), praying mantises (Mantodea), walking sticks (Phasmatodea), and earwigs (Dermaptera), each of which is now considered a separate, albeit closely related order. For the purpose of this review, the classification system proposed by Rentz (1996) will be adopted. This system divides the order Orthoptera into two suborders: Ensifera, with superfamilies Tettigonioidae (katydids), Gryllidea (crickets), and Gryllacridoidae (cave crickets and relatives); and Caelifera, with superfamilies Eumastacoidea (monkey grasshoppers), Acridoidae (grasshoppers and locusts), Tettigoidae (pygmy grasshoppers), and Tridactyloidea (pygmy mole crickets and relatives). The following differences generally allow Ensifera and Caelifera to be distinguished (note that there are numerous exceptions):

<table>
<thead>
<tr>
<th>Ensifera (katydids, crickets and relatives)</th>
<th>Caelifera (grasshoppers and relatives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antennae long and thin, with over 30 segments</td>
<td>Antennae short, with less than 30 segments</td>
</tr>
<tr>
<td>Anterior tibiae usually with tympanal organs</td>
<td>Anterior tibiae without tympanal organs</td>
</tr>
<tr>
<td>Ovipositor long, sword- or needlelike</td>
<td>Ovipositor very short</td>
</tr>
<tr>
<td>Sound produced by tegmental stridulation</td>
<td>Sound produced by rubbing hind legs against tegmina</td>
</tr>
</tbody>
</table>

The monumental task of describing all living species of Orthoptera is far from complete. As the exploration of tropical regions of the world progresses, more and more species awaiting to be formally described accumulate in museum collections worldwide. Unfortunately, at the same time the number of specialists capable of such work is dwindling rapidly.

F. Paleontology

The Orthoptera is an old group, dating back to the Carboniferous period. Along with Titanoptera and Phasmodea (= Phasmatoptera), they represent the superorder Orthopteroidea. Both Titanoptera and Phasmodea branched off of the main stock of Orthopteroidea in the early Permian. Titanoptera disappeared at the end of the Triassic while Phasmodea (walking sticks and their relatives) and Orthoptera still flourish today. The two main lineages of Orthoptera, Ensifera and Caelifera, probably separated in the late Carboniferous. Most modern families of Ensifera appeared between the early Jurassic and the early Triassic periods. The oldest, still extant family of Ensifera, the Prophalangopsidae, appeared in the early Jurassic. The oldest family of Caelifera, the Eumastacidae, appeared in the middle Jurassic, followed by the Tettigidae and the Tridactyliidae at the beginning of the Cretaceous.
II. TETTIGONIOIDEA
(HAGLIDS AND KATYDIDS)

A. Introduction

The superfamily Tettigonioidae includes two families: the Prophalangopsidae (haglids or hump-winged crickets) and the Tettigoniidae (katydids or bush-crickets) (Fig. 1). It is the second largest group of Orthoptera (grasshoppers being more speciose), with nearly 6300 described species assigned to nearly 1100 genera and 20 subfamilies. Tettigonioidae are characterized by the presence in the males of well-developed tegmental stridulatory apparatus; long, threadlike antennae, usually longer than the body; four-segmented tarsi; short, non-articulated cerci in the males; and a large external ovipositor in the females. The wings, when present, are usually held in a roof-like fashion over the abdomen. Virtually all members of the superfamily have bilateral tympanal organs on the front tibia as well as an additional acoustic trachea on the thorax behind the pronotum.

The Tettigonioidae are distributed throughout the world, but like most other orthopterans, they show their greatest diversity in the tropical and subtropical regions. Although many groups of katydids are quite diverse in temperate regions of the world, they are less common in areas with long, cold winters. A large proportion of the Tettigonioidae are arboreal but many also inhabit grasslands, savannas, deserts or rocky mountain tops. Most species are nocturnal or crepuscular, though several major groups consist almost exclusively of diurnal species.

FIGURE 1 Tettigonioidae, Pychopulpa bicephala from Costa Rica (Tettigoniidae, Phaneropterae). See also color insert, Volume 1.

B. Major Lineages

1. Prophalangopsidae (Haglids or Hump-Winged Crickets)

This ancient family, dating back to the early Jurassic period, is represented by only six extant species. They are characterized by a rather primitive stridulatory apparatus lacking a proper mirror, head with antennae inserted near the lower margin of the eyes, and the metatarsus with a well-developed pulvillus. The rarest of these living fossils, Prophalangopsis obscura, is known from a single specimen collected over a hundred years ago at an unknown location in India, and is the sole member of the subfamily Prophalangopsinae. Naturally, little is known about this insect other than its strong resemblance to forms extinct millions of years ago. The remaining five species belong to two genera of the subfamily Cyphoderrinae and are distributed in western North America (Cyphoderris) and northeastern Asia (Paracyphoderris). The biology of Cyphoderris is quite well known and some species of the genus are often used in studies on mating behavior and parental investment in offspring. These insects are unique among the Tettigonioidae in the fact that the male's stridulatory apparatus is almost perfectly symmetrical—that is, males have fully functional files and scrapers on both tegmina and can produce sound with either the left or right tegmen (hence the name "ambidextrous crickets," sometimes applied to these insects). The hindwings are reduced to fleshy lobes, which are devoured by the female during copulation. Males who had already mated once and are missing these courtship "snacks" must resort to other methods of holding the female's attention and instead use the "gin trap," a complex system of cuticular modifications whose role is to hold the female's abdomen firmly in place during copulation. These insects are also unique in their ability of being active, and even call during winter, from snow-covered bushes, when the temperature oscillates around 0°C.

2. Tettigoniidae (Katydids or Bush-Crickets)

Katydids are a large and diverse group of orthopterans, elevated by some authors to the rank of a superfamily with numerous families. Nearly 6300 species have been described, placed in nearly 1100 genera and 18 subfamilies. Most katydids are medium sized or large insects. The largest katydid, Sialaeocerella grandis, from New Guinea, has a wingspan of over 250 mm. On the other hand, members of the Australian subfamily Microtettigoniinae are less than 10 mm long.
Males of almost all species have a well-developed stridulatory apparatus at the base of the tegmina. The exceptions are the subfamilies Phyllophoridae, Phasmodinae, and some members of Meconematinae and Mecopodinae. Males of the Australasian subfamily Phyllophoridae, despite having fully developed wings, lack the stridulatory apparatus and instead produce sound by rubbing their hind coxa against specially modified thoracic sterna. Australian Phasmodinae lack wings altogether and are presumably totally silent. This is also the only subfamily of katydids having greatly reduced tympanal tibial structures ("the ear") and lacking the thoracic auditory spiracle. All katydids have long, thin antennae, often longer than the body, in extreme cases exceeding its length three to four times. The head of katydids is hypognathus, rarely prognathous (Phasmodinae, Zaprochilinae), and sometimes bears distinct, cuticular, hornlike processes between the antennae. Their function is uncertain but is likely related to the fact that many predators (especially bats and birds) tend to seize their katydid victims by the head. The pronotum is enlarged, saddle shaped, and in some forms greatly expanded, covering part of the wings and the abdomen. In the male, the enlarged pronotum often acts as a call amplifier. Curiously, the group with a particularly enormous, boxlike pronotum, the Phyllophoridae, lacks the ability to produce sounds with their wings. The pronotum is also sometimes armed with sharp spines and processes. The wings of katydids can be fully developed or reduced to a varying extent, in some lineages (Tettigonominae) showing significant intraspecific polymorphism in the degree of their development. The tegmina are sometimes reduced to scalelike organs, the only function of which is sound production. Hindwings in such taxa are either totally missing or greatly reduced. Many katydids use their wings to enhance their remarkable mimicry of leaves, twigs, or bark to a degree that can fool professional biologists. Members of the neotropical tribe Pterochrozini are particularly excellent mimics of both live and dead leaves, perfectly imitating the look and feel of leaves, including simulated traces of herbivory, with simulated lichens and mosses "growing" on them. A few katydids are Batesian mimics of wasps.

The legs of katydids are usually long and slender. Predaceous species frequently have rows of long spines along the ventral edges of the front and middle legs, using them for grasping and holding prey. In some Listroscelidinae the spines are particularly long, forming a kind of a net, used for scooping small flies and other insect prey. The hind legs are saltatorial, except for the members of some groups living in burrows or under rocks (Heterodinae, Bradyporinae). The abdomen of males has a pair of cerci, in some groups developed to grasp the end of female's abdomen during mating. The ovipositor of females is usually long and sickle or sable shaped. The reduction of the ovipositor to short processes appears independently in ground-dwelling forms (Heterodinae) and some arboreal ones (Phaneropterae), apparently related to the fact that the eggs are laid on the surface of the soil or leaves, rather than inserted into them, as most katydids do.

The significant male investment in offspring is one of the characteristics of katydids that attracts many researchers to study their behavior and evolution. During copulation, males of many katydids produce a large, protein-rich spermatophylax, which is eaten by the female after copulation. The size of the spermatophylax can approach 60% of the male's body mass, making it an extremely costly and significant contribution to egg production. This causes the males of many species to be quite choosy in selecting their mating partners, and under certain circumstances, the females may compete for males, a role reversal remarkably rare in the animal world.

The geographic distribution of katydids reveals several large centers of endemism, the largest being Australia, where 5 out of the 13 subfamilies known from the continent are unique to it. The neotropics have the highest number of described species (approximately 4000) plus the endemic subfamilies, followed by Eurasia (approximately 1350 species), the Indo-Malaysian region (approximately 1150), and Africa (approximately 850 species, 2 endemic subfamilies). These numbers, however, are very likely to be multiplied in the future if taxonomic work in these regions is intensified.

a. Major Subfamilies

Bradyporinae. Most likely the most basal subfamily of katydids, it is characterized by having a large, stout body, greatly reduced wings, and the tarsus with a metasternal pulvillus. Members of this family are restricted in their distribution to southwestern Palaeartic. Often grouped with several other subfamilies (Heterodinae, Ephippigerinae), they nonetheless appear to be a separate clade of their own. Only about 50 species have been described. All species of the subfamily are dark colored, often black, and resemble giant crickets rather than typical, graceful katydids. Most live on the ground or low vegetation. Some species are known to produce defensive autohemorrhage and squirt their hemolymph.
from orifices on their body if the insect is seized by a predator.

Phaneropterinae. This is the largest subfamily of Tettigoniidae, with more than 2100 species described from all continents of the world (other than the polar regions). Members of this subfamily are medium to large, usually green and leaf-like katydids, characterized by a lack of lateral grooves on the tarsi, and rather primitive venation of the wings. Wing reduction is widespread in several lineages of the subfamily but the stridulatory apparatus is always present. Female stridulation is known in a few genera. Virtually all species are phytophagous, although opportunistic cannibalism has been observed in a few species. Many species are exclusively arboreal. Eggs are laid in the soil, plant tissues, or on the surfaces of leaves and bark. Immature stages of many species mimic ants, tiger beetles, and even spiders, while adults of the same species usually mimic leaves or blades of grass. Several South American genera are superb mimics of pompilid wasps.

Acridoxeninae. This is a small, aberrant subfamily of large Central African katydids, with only one species, Acridoxena hewaniama. Its relationship to other katydids is uncertain as it displays a mixture of very advanced and very primitive characters. The most interesting feature of Acridoxena is its astonishing mimicry of a dried, spiny plant. All parts of the body look like shriveled, twisted, dry leaves or twigs. Nothing is known of its biology or behavior.

Ephippigerinae. This is a small Palaearctic subfamily, with about 130 described species. All species have greatly reduced, scale-like wings, usually hidden under an enlarged, saddle-shaped pronotum. Females of many species stridulate and certain species have long been model organisms for studies on acoustic communication and courtship in insects. Some species occasionally cause minor agricultural damage.

Pseudophyllinae. This is the second largest subfamily of Tettigoniidae, with more than 1100 described species, distributed in tropical and subtropical regions of the globe. A notable exception is the true katydid (Pterophylla camellifolia) of the eastern United States, which is both the northernmost member of the subfamily and the source of the name "katydid." Its characteristic, loud call resembles (although some listeners disagree) the syllables "ka-ty-did." Most pseudophylline katydids are found in the tropical areas of South America and Southeast Asia. Many are spectacular mimics of leaves (tribes Pterochrozinini and Pseudophyllini) and bark (tribes Plaminini and Cymatomerini). Secondary loss of stridulation and widespread presence of tremulation is characteristic of most neotropical members of the subfamily. On the other hand, Old World members of the subfamily belong to the loudest night singers of many forests. This disparity in their acoustic behavior has been explained by different hunting strategies of insectivorous bats in the New and the Old World, which in turn, shaped different defensive strategies among acoustic insects. Nearly all species of the Pseudophyllinae are arboreal or at least associated with tall vegetation. A notable exception is the genus Callimenellus, which is known to inhabit marine littoral rock crevices in Hong Kong. All species seem to be phytophagous although opportunistic insectivory has been observed in a few Central American species. Virtually all species are nocturnal.

Tettigoniinae. A large subfamily, with nearly 1000 described species. Most of the species of this subfamily occur in the temperate regions of the world, and only a handful of species are present in the tropics. Tettigoniinae achieved the greatest diversity in the regions of Europe, western North America, southern Africa, and Australia characterized by the Mediterranean type of vegetation. Many species have a large, shield-like pronotum, hence the common name "shield-backed katydids." Females always have a long, sword-like ovipositor and lay eggs in soil or the stems of herbaceous plants. Males stridulate loudly and many species are active during the day. A few species of shield-backed katydids are agricultural pests, the best known being the Mormon cricket (Anabrus simplex) of the western United States. However, some species in China and Japan have been kept as pets for hundreds of years, and their pleasing calls have made a remarkable impact on the poetry and other arts of these countries.

Conocephalinae. The conehead katydids are a large, poorly defined, and possibly polyphyletic subfamily of Tettigoniidae, with more than 1100 described species worldwide. Many are characterized by a prominent fasiculum of the vertex, forming a characteristic "horn" on the head. The diet of many of these katydids is restricted to grasses and their seeds, although quite a few species are predaceous. The conehead katydids are quite common in temperate regions of North America and Europe, although they reach their greatest diversity in the tropical areas of South America. Some species of Conocephalinae may become agricultural pests and a few have been known to form large, locust-like swarms (Ruspis). Male investment in offspring can be significant in some neotropical conehead katydids, and an elaborate courtship behavior is common in such species.

Phasmodinae and Zaprochilinae. These two, closely related subfamilies are restricted to Australia. Phasmod-
iniae contains only one genus, *Phasmodes*, with three species restricted in their distribution to Western Australia. These insects are a good example of convergent evolution, resembling walking sticks (*Phasmodes*) to an extraordinary degree. Both sexes are completely wingless, and females lack tibial tympana as well as thoracic auditory spiracles. The entire body is extremely elongate and thin, and the head is prognathous. A simple way of telling these insects apart from walking sticks is looking at the proportions of thoracic segments. In real walking sticks the prothorax is extremely elongated, whereas in *Phasmodes* it is the mesothorax that underwent such modification. These interesting katydids feed on both leaves and flowers of the heath habitats.

Zaprochilinae have similarly elongated bodies and the prognathous head, but all species have wings. They are strongly reduced in the genus *Kawanaphila* but fully developed in the remaining three genera of the subfamily. Males of all 28 described species produce short, ultrasonic calls. Several species of the subfamily have been extensively studied with regards to the parental investment of males, which is significant and may lead to coevolution of reproductive strategies. All members of the subfamily feed primarily on pollen and nectar, and they play some role in pollination of the flowers on which they feed.

In addition to the above-described subfamilies, the family Tetrigonidae also includes the following subfamilies: Austrosaginiae, Hetrodinae, Lipotactinae, Listroscelidinae, Meconematinae, Mecopodinae, Microtettigoninae, Phyllophorinae, Saginae, and Tympanophorinae.

III. GRYLLOIDEA (CRICKETS AND MOLE CRICKETS)

A. Introduction

The superfamily Grylloidea includes more than 3700 described species placed in more than 520 genera and three families (Fig. 2). Crickets are small to medium size insects, the smallest species being only about 1.5 mm long (*Myrmecophilidae*) and the largest about 60 mm (*Gryllotalpidae*). The family is cosmopolitan in its distribution and some species occur even within the polar regions, albeit usually only in association with human dwellings. Most members of the subfamily are brown or black, very rarely green. The antennae are usually thin, threadlike and longer than the body (with the exception of certain burrowing forms). The head is usually large and almost globular, although it may be somewhat elongate and prognathous in some lineages (*Oecanthinae, Gryllotalpinae*). The pronotum is usually quadrate, rarely produced backward as to cover the wings (a condition common in katydids). The tarsi of the legs are three-segmented (superficially four-segmented in *Oecanthinae*), and the front legs are sometimes strongly modified for digging. The front tibia typically have well-developed bilateral tympanic organs. Hind legs are usually saltatorial but are occasionally short and not adapted for leaping. Wings are usually present but may be reduced to various degrees or totally absent (*Myrmecophilidae*). The tegmina are held flat on the dorsum and their anterior (lateral, in the resting position) margins are bent downward at an angle, forming a boxlike structure. The stridulatory apparatus of males is well-developed in most groups although it may be absent even in some fully winged taxa (some Pentacentrinae, Trigonidinae, Eneopterinae). The left and right tegmina of males have similar venation, but stridulation is performed virtually always with the right tegmen over the left. The hindwings, when folded, often project far beyond the tegmina. Many taxa lack the hindwings and wing polymorphism is common. Male cerci are usually long, flexible, often superficially annulated, and never grasping. The ovipositor is usually long and needle-like, although it may be reduced or even absent in some crickets (*Gryllotalpidae*).

Most species of crickets are nocturnal or crepuscular. They occur in almost all terrestrial habitats, from tree-stems to underground burrows, and many species are associated with aquatic and semiaquatic environments (such forms are usually very good swimmers). The great majority of crickets are opportunistic feeders and few seem to be exclusively predaceous.
B. Major Lineages

1. Gryllotalpidae (Mole Crickets)

This interesting family includes five genera and 77 described species of usually large, robust crickets, distributed worldwide. All species of the family spend their lives in underground tunnels, excavated with their extremely modified, shovel-like front legs. The head is prognathous and somewhat elongated, and so is the pronotum. The antennae are short and somewhat thickened, an apparent adaptation for their subterranean lifestyle. The wings are usually fully developed, rarely absent. The stridulatory apparatus of males is simple, lacking the mirror. The females’ tegminal venation resembles that of the males but lacks proper stridulatory organs. Cerci in both sexes are long and flexible, acting as an extra pair of antennae while the insects are moving backward in their underground corridors. Females lack an external ovipositor.

Mole crickets emerge from their tunnels only during courtship, although calling takes place underground. In order to broadcast their signals as far as possible, males build Y-shaped tunnels, the length and diameter of which are perfectly suited to amplify the dominant frequencies of their call. Females stridulate as well, but their call seems to have a territorial rather than courtship role. Eggs are often laid in special brooding chambers and actively cared for by the female. Newly hatched nymphs stay in the chamber for several weeks, feeding on humus and tender rootlets protruding into the burrow.

Several species of mole crickets are serious crop pests, not only because they feed on roots of plants but also because they damage entire root systems while digging tunnels. A species of Gryllotalpa recently introduced accidentally to Australia has quickly become one of the major threats to the Australian golfing industry by damaging carefully cared for putting greens.

2. Gryllidae (True Crickets and Tree Crickets)

The family Gryllidae is the largest lineage of the superfamily, with nearly 3600 species, assigned to more than 500 genera in 17 subfamilies (there are several alternative systems of classification of crickets, and the number and status of higher categories varies among different authors). The body size of true crickets ranges from small (under 5 mm) to large (about 50 mm). Most members of the family are rather stout insects, with short, thick legs. Phanogopinae are a notable exception, having extremely long and slender legs and other appendages, making them resemble large spiders rather than typical crickets. Delicate bodies with slender legs also characterize most tree crickets (Oecanthinae). The head is generally globular, except in tree crickets, which have an elongated, prognathous head. The pronotum is generally quadrate. The wings and stridulatory apparatus may be fully developed, reduced, or totally absent. Often, fully winged forms lack the stridulatory apparatus, while forms with greatly reduced tegmina can have a fully developed one. Species with well-developed stridulatory apparatus usually produce melodious, frequency and time-modulated calls, and many species have been kept as singing pets in Asia and Europe for many centuries. Rarely, the call is short and clicklike (Eneoptera). Courtship behavior is often complex and involves acoustic, tactile, and olfactory signals. The courtship song, produced only in the close proximity of the female, is quieter and of a different structure than the advertisement call. It is usually produced by rubbing only a part of the stridulatory file against the scraper. Some spider crickets (Phanogopinae) additionally drum with their legs. During copulation, males produce spermatophores, which in some species contain a large spermatophylax. Males of some species allow the female to feed on different parts of their bodies. Males of tree crickets have thoracic glands, which the female licks during copulation, and females of some pygmy field crickets (Nemobinae) feed on males’ abdominal spines during copulation.

The ovipositor in true crickets is usually long, needle-like, sometimes laterally flattened and swordlike (Trigonidinae) or reduced (Brachytrupes). Oviposition takes place in soil or plant tissues. Maternal care is rare but sometimes quite well developed. Females of Anurogryllus muticus excavate extensive subterranean burrows as a nursery for their eggs, which are aggressively protected against all intruders. Newly hatched nymphs stay with their mother, who feeds them with unfertilized eggs, produced by her for the sole purpose of feeding her nymphs.

a. Major Subfamilies

Gryllinae. This subfamily includes the true crickets, placed in more than 850 species and more than 100 genera. They are distributed worldwide, and a few species closely associated with human habitats are virtually cosmopolitan. Morphologically, Gryllinae are rather uniform, having stocky bodies, with a globular head and thick legs. Most have well-developed wings and a stridulatory apparatus. True crickets are some of the first singing insects to be heard in the spring in the temperate areas of Europe and North America, as they often overwinter as late nymphal instars and mature as soon as the temperature allows them to become active. North America has a particularly interesting fauna of
true crickets, with several complexes of morphologically nearly identical, so called "cryptic species." Some of the species appear to be chronospecies, that is, very closely related species that avoid interbreeding by being sexually mature during different times of the year.

Most true crickets are ground dwellers and many are quite territorial, vigorously defending their burrows. In some species a certain percentage of the males (satellite males) consistently exhibit a kind of sexual parasitism by intercepting females attracted to the call of another male. Several species form large swarms, and some (e.g., Gryllus bimaculatus in Africa) can become serious agricultural pests. The house cricket (Acheta domestica) is a cosmopolitan insect associated with human dwellings and when occurring in large numbers can cause damage to stored food and other material.

Brachytrupinae. This subfamily, often treated as a subset of the true crickets, includes about 20 genera and more than 220 species distributed mostly in the tropical and subtropical regions of the Old World, with relatively few species in the New World. The wings in some species are greatly reduced or absent and the female's ovipositor is reduced. Some of these crickets are very large and produce calls of an intensity directly proportional to their size. Some species are gregarious and live in small underground colonies consisting of adults and nymphs of different ages. Some species of the subfamily exhibit extraordinary maternal care of eggs and nymphs.

Phalangopsinae. The spider crickets are characterized by having a rather robust body supported, in most cases, by very long, thin legs. The antennae and palps in such forms are also extremely long and slender. Wings are frequently reduced or absent, and some winged species lack the stridulatory apparatus in the male. On the other hand, some species are loud singers, with a pleasing call, and males of several Asian species are greatly prized for their songs and are kept in special little cages as pets. The subfamily includes about 380 species in nearly 100 genera, achieving the greatest diversity in the tropics of Central and South America. Many species inhabit caves and rock crevices as well as spaces between buttress roots of trees in tropical forests. Others prefer tree trunks, leaf litter, and decaying wood.

Mogoplistinae. The scaly crickets are small, rarely exceeding 15 mm in length. Their name is derived from the fact that their entire body is covered with minute scales, reminiscent of the scales on the wings of butterflies and moths. Similar to these insects, the scales on the body of scaly crickets often form beautiful color patterns. Unlike butterfly wings, however, the wings of scaly crickets are not covered with scales, and are strongly reduced, often entirely concealed under a somewhat elongated pronotum. About 270 species have been described, placed in 26 genera. They are distributed worldwide but Australia and the Indo-Malaysian region seem to have the largest share of known species.

Oecanthinae. The tree crickets are rather unusual members of the family Gryllidae, having a somewhat elongated, prognathous head and a rather long pronotum. The wings can be either fully developed or completely absent. Winged forms have a very well developed stridulatory apparatus, capable of producing loud, pure sounds. Males of the genus Oecanthus enhance the range of their call by positioning themselves in a hole chewed out in a leaf, with their tegmina aligned during singing with the surface of the leaf. By doing this, they significantly increase the area from which the sound radiates, using the same principle on which speakers in radios and other musical appliances are built. Thanks to this technique, these small insects are capable of becoming the dominant singers in many environments.

Females of tree crickets lay eggs in the stems of plants, and by doing so may damage young trees in nurseries and orchards, thus becoming pests. On the other hand, many species feed on aphids, thus balancing the negative effect of their reproductive behavior.

Other subfamilies included in the family Gryllidae are Cachoplistinae, Eonephtera, Euscyrtinae, Gryllomminae, Itarinae, Malgasminae, Nemobinae, Pentacentrinae, Podoscirtinae, Pteroplistinae, Scleroptyrinae, and Trigonidiinae.

3. Myrmecophilidae (Ant-Loving Crickets)

This small family seems to be closely related to Mogoplistinae and probably does not deserve the status of a separate family. Only about 10 genera and less than 70 species are known, and all of these are closely associated with colonies of social insects, mostly ants, and less frequently termites. All species are minute, sometimes less than 1.5 mm, with an oval, wingless body and a small head with greatly reduced eyes. The hind femora are short and enormously enlarged.

Ant-loving crickets seem unable to live without a close association with ants or termites. Some species have a wide range of host species while others are restricted to only one host species. The nature of this association is unclear. At least one myrmecophilous species, Myrmecophila manni from North America, appears to be able to mimic ant behavior to an extent that allows it to be fed by workers of its host, Formica obscuripes. Another species, M. oregonensis, feeds on residues of the substance that lubricates ants' bodies. left on the walls of passageways. A few species of these crickets appear to be parthenogenetic. Their eggs are
relatively large and their embryonic development is quite long, sometimes lasting over a year.

IV. GRYLLACRIDIOIDEA (CAMEL CRICKETS, CAVE CRICKETS, JERUSALEM CRICKETS, RASPY CRICKETS)

A. Introduction

The taxonomy and phylogeny of this group of the Orthoptera seems to be in greater disarray than any other lineage of the order. Many different names have been applied to this group of insects, and frequently various subgroups have been elevated to the levels of superfamilies or even infraorders. As no modern phylogenetic study has yet been done on the Grylloidea, for the sake of taxonomic stability this review follows classification proposed by Rentz (1996).

The superfamily Grylloidea includes about 1500 species and more than 200 genera assigned to four families: the Caeliforidae (Cooooloona monsters), the Stenopelmatidae (Jerusalem crickets and wetas), the Grylloidea (raspy crickets), and the Rhaphidophoridae (the cave and camel crickets) (Fig. 3). Members of this superfamily are characterized by the lack of the tegminal stridulatory apparatus even in fully winged species. Tegmina, if present, are soft and characteristically curved around the abdomen. The head is usually large and bullet-like, and some species have greatly enlarged mandibles. The antennae tend to be very long and threadlike but in some subterranean forms (Cooooloona) may be extremely shortened. Tibial auditory structures are usually absent, and the thoracic auditory spiracle is small and simple. Acoustic communication in this group is not as common as in other orthopterans and is restricted to femoral stridulation (Grylloidea) and abdominal drumming (Stenopelmatidae).

Body size varies from about 10 mm to about 100 mm, and some species of New Zealand wetas are considered the heaviest insects in the world, reaching the weight of 70 g—nearly three times as much as an average house mouse. Coloration of species of Gryllacridoidea is rather drab, usually brown, black, or yellow, but never green.

The distribution of the superfamily is worldwide, with the exception of the Cooooloona monsters, which are known only from northeastern Australia. At least one species, Tachytes australis, has become virtually cosmopolitan thanks to its association with greenhouses and other human environments. Nearly all species of Gryllacridoidea are nocturnal or crepuscular. Their habitats include treetops in tropical forests, savannas, and deserts, but a great proportion of species dwell in caves or underground burrows. Many species have a tendency to form large aggregations.

B. Major Lineages

1. Cooooloona (Cooooloona Monsters)

Some of the most peculiar insects belong to this small family, consisting of only one genus and four species. They are restricted in their distribution to central and northern Queensland in Australia and were discovered only about 20 years ago. Their appearance is so unusual that the first specimens ever collected were suspected of being fabricated as an entomological joke. The body of Cooooloona monsters is so bulky that these insects, and especially the females, are almost incapable of walking on the surface of the ground. They spend their entire lives underground, digging through sand with their extremely thick and powerful legs. They are virtually blind and their very short antennae consist of only nine segments. Their mandibles are elongate and sharp, and the maxillae (the second part of mouthparts) bear two dagger-like processes, which probably act as piercing organs, a situation absolutely unique within Orthoptera. The wings are vestigial in males and usually completely absent in females. The internal anatomy of these insects is also unusual in the fact that the foregut extends backward into the posterior part of the abdomen.

Little is known about biology of Cooooloona monsters. They appear to "sand swim" through sand and do not construct burrows. They feed probably on body fluids of other subterranean animals.
2. Stenopelmatusidae (Jerusalem Crickets and Wetas)

This family includes about 230 described species, placed in more than 50 genera and six subfamilies. They are known primarily from North and South America, southern Africa, Madagascar, India, Australia, and New Zealand. The body of these insects is usually fairly large and bulky, and New Zealand wetas are among the largest living insects. The head is large and globular, with relatively short antennae. Legs are thick, powerful, and frequently well adapted for digging. In some species, the powerful hind legs can serve as a weapon, and these species are capable of delivering strong and painful kicks. Wings are usually absent and are rarely developed well enough for the insects to be able to fly. They do not have stridulatory organs on the tegmina, but many species communicate with elaborate percussive signals produced by drumming their abdomen against the substratum. The courtship behavior of Jerusalem crickets is a violent spectacle, involving a battle between the male and the female. If the female does not recognize the male as an acceptable partner, she devours him, thus preventing copulation. If the male is found to be acceptable, she allows him to mate and deliver to her a large spermatophore, and she attempts to eat him only afterward. This behavior may account for the rarity of male specimens of Jerusalem crickets in entomological collections. On the opposite end of the spectrum, males of some New Zealand wetas maintain harems of females in their refuges in holes and cracks in trees. Other males frequently try to take over the harems, and fights involving head-on clashing and gapping of their jaws are quite common among males.

Most species of Stenopelmatusidae spend their lives on or close to the ground, with the exception of some Central and South American species (Anabropsis), which are often found in the canopy of the rain forest. Some North American species (Chmatettix) produce silk with their mouthparts, which is used to line their burrows in coastal sand dunes. Many species are predaceous, but most seem to be opportunistic generalists. Jerusalem crickets have been reported to cause minor injuries to crops by gnawing on roots and parts of the plants close to the ground.

3. Gryllacrididae (Raspy Crickets and Leaf-Rolling Crickets)

This large family includes nearly 800 described species placed in almost 90 genera. It is quite likely, however, that these numbers represent only a fraction of the actual number of species, as the faunas of Gryllacrididae of Central and South America are virtually unknown. Most species of the family have been described from southeast Asia and Australia, but they are also known from Africa and the neotropics. North American fauna of Gryllacrididae is limited to only one species, Campito notus carolinensis, and they are absent from Europe.

Species of Gryllacrididae are characterized by the globular head with often extremely long antennae, which in some species may exceed the body length four to five times. The wings are usually present but may be reduced or absent in some species. The ovipositor of females is long and thin, and in nymphs held curved backward over the body. This morphological modification may allow them to fit more easily into their burrows or rolled leaves. Leaf-rolling behavior is quite common in this family. All stages of some species, from the tiniest nymphs to adults, are capable of producing silk, which they use to bind shelters made of leaves, twigs, sand, and other material. Burrowing species living in hot, arid zones are known to insulate themselves from the heat by sewing shut the entrances to their burrows. The unusual subfamily Schizodactylinae (by some authors considered a separate family or even an infraorder) is unique in the peculiar modifications of their hindwings, which are coiled into tight, vertical spirals. Their feet possess long, lobelike processes, which help them run on loose sand.

Tegmental stridulation is absent in species of Gryllacrididae but many species produce raspy sound (hence the common name) by rubbing the inner side of the femora against the sides of the abdomen. This behavior does not seem to play any role in courtship but is rather a defensive strategy, since all life stages of both sexes are known to display it. All species of Gryllacrididae are nocturnal. They are probably unique in the insect world in possessing individually recognizable pheromones, which allows a cricket to return to its own burrow, distinct from one of its conspecifics, after a night of foraging. Most members of the family are predators and feed on small insects. Some are extremely agile and actively hunt for prey by rapidly running along branches, searching for insects.

4. Rhaphidophoridae (Cave Crickets and Camel Crickets)

Cave and camel crickets are represented by about 470 described species assigned to over 70 genera in seven subfamilies. They are distributed worldwide and at least one species is cosmopolitan. Species of this family are rather uniform in their appearance. All are completely wingless and have smooth, dorsally convex bodies. Most have long, spidery legs and similarly long and slender antennae and palps. The ovipositor is laterally flattened and often has distinctly serrated edges.
V. EUASTACOIDEA (MONKEY GRASSHOPPERS OR EUASTACIDS)

A. Introduction

The majority of members of this large superfamily occur in the tropics, although a few are found in temperate regions (Fig. 4). Nearly 1300 species have been described, and placed in about 300 genera in two families. As in all Caelifera, the antennae of euastacids are very short, but unlike other members of the suborder, one of the last few segments of the antennae always bears a small tubercle of unknown function. The thorax can be short (Euastacidae) to extremely long (Proscoptidae), and in some groups it can have a distinct cuticular crest. The hind femora typically are long and slender, although in some forms may be greatly shortened and not adapted for leaping. In forms with long, leaping hind legs, they are characteristically splayed outward from the body when the insect is at rest. Wings are frequently reduced or absent, but when present they are very long, and characteristically widened toward the end of the body. The tegmina are frequently transparent or semitransparent. The abdominal auditory organ is absent in most species of the superfamily.

Little is known about the biology of members of this superfamily. None of the species appears to stridulate. Many species are arboreal, but some live on grasses in savanna-like environments and frequently have extremely elongated bodies, perfectly mimicking blades of grass (Australian Morabinae). Eggs are laid in batches in soil but, unlike those of other grasshoppers, they are never bound together by a foamy mass.

B. Major Lineages

1. Euastacidae (Monkey Grasshoppers)

The larger of the two families in this superfamily, the Euastacidae includes more than 1000 species in nearly 270 genera, distributed worldwide. Body size in this family is small (10 mm) to medium (45 mm). The pronotum is always short, and the front and middle legs are relatively short. Hind legs are usually saltatorial but in some grass-mimicking forms may be shortened and lacking the ability to leap. The male genitalia structures are very complex and of great value in taxonomic research. The ovipositor of females is short, with strongly serrated edges.

2. Proscoptidae (False Walking Sticks or Proscoptids)

Proscoptids are restricted in their distribution to South America, barely reaching the southernmost part of the isthmus of Panama. About 300 species are known, placed in 29 genera. This family is readily distinguishable from other grasshoppers by their greatly elongate, twiglike appearance. Some species may reach a length of 165 mm, while the smallest are about 25 mm long. The head of these insects is unusually long and pear-shaped, with large eyes situated near its narrow tip. The thorax is extremely long and slender, and the front and middle legs are well separated. Femora of the hind legs are characteristically thickened at their bases and are not adapted for jumping. Virtually all species are completely wingless. Most species live on bushes and trees, and a few in savanna-type
VI. PNEUMOROIDEA (BLADDER HOPPERS OR PNEUMORIDS)

This small group of grasshoppers includes some of the most unusual orthopterans living today (Fig. 5). The superfamily includes only 25 described species, placed in 13 genera of three families. The family Pneumoridae is restricted in its distribution to southern and eastern Africa, while families Tanaoceridae and Xyronotidae occur only in Mexico and southeastern United States.

The principal family, the Pneumoridae, includes large (50–100 mm) grasshoppers, with the wings either fully developed or somewhat reduced. The pronotum is greatly enlarged, forming a triangular helmet covering a large part of the body. The most unusual feature of the members of the family is the shape and size of the male’s abdomen, which in most species is grossly inflated, creating a balloon-like, semitransparent resonator. The stridulatory apparatus consists of a dozen or so cuticular ridges on the third abdominal segment. Thanks to the abdominal resonating chamber, the sound made by the males is among the loudest sounds produced by any insect. Females are capable of producing sound as well but lack the abdominal stridulatory apparatus. Long winged species are sometimes called “flying gooseberries” as they are frequently attracted to lights. Males who have the misfortune of being attracted to campfires explode like balloons if they fall into the fire.

The remaining two families of Pneumoroidae include only eight species of small grasshoppers, living on shrubby vegetation in deserts of North America.

VII. ACRIDOIDEA (GRASSHOPPERS AND LOCUSTS)

A. Introduction

Grasshoppers and locusts form the largest lineage of Orthoptera, with more than 8500 species in nearly 1800 genera assigned to 10 families (Fig. 6). They are distributed worldwide, and although they are most diverse in tropical and subtropical regions of the world, they also occur in cold, subpolar regions and high in the alpine zones of the mountains. The body size of these insects ranges from less than 10 mm to over 120 mm, with a wingspan of 250 mm. The body form and the head shape are very diverse, ranging from short and bulky to extremely elongate and slender. The antennae are always short and relatively thick, but often modified to form comb- or clublike structures. The hind legs are always distinctly thickened toward the base and always adapted for leaping. The wings vary from fully developed to completely lacking. Fully winged forms have relatively very narrow tegmina that are not widened toward their apices. Stridulatory organs are present only in some families and usually consist of a modified vein on the tegmen, against which the inner side of the hind femur is rubbed. The abdomen usually has lateral auditory structures, even in silent wingless species. The female’s ovipositor is always very short, with the upper and lower portions (valves) distinctly divergent.
at the apex. Eggs are laid primarily in the soil, and are always protected with a foamy mass produced with special glands in female's abdomen.

Most grasshoppers are diurnal, although nocturnal species are also known. They inhabit nearly all terrestrial habitats and a few species are aquatic (Paulinidae). Open steppes and savannas of warmer regions of the world have particularly rich faunas of grasshoppers, but tropical rain forests have their share of unique forms too. Grasshoppers are also relatively diverse in temperate zones, with some species crossing the polar circles. A number of alpine species are adapted to activity in very low temperatures. Some of these forms exhibit the interesting behavior of changing their body color depending on air temperature. In cold weather, they become much darker, and thus more likely to absorb sun rays (Rosciuscolia).

Virtually all species of grasshoppers are exclusively phytophagous. Opportunistic cannibalism has been observed in few rare instances, but only under crowded conditions. Most grasshoppers are polyphagous, feeding on a number of different plant species, although most tend to have preferences toward certain species of plants. Some grasshopper species are monophagous, feeding on only one species of plant throughout their life. Such species tend to feed on plants rich in toxic secondary compounds, and in doing so become unpalatable to potential predators themselves. They often advertise their toxicity with bright, aposematic coloration. Some grasshoppers (African Pygromorphidae) can even be harmful to humans, if ingested.

A few grasshopper species, mostly of the subfamilies Cyrtaucheniinae, Oedipodinae, Calliptaminae, and Melanoplinae, are very serious agricultural pests and cause enormous losses to crops every year. They form unbelievably huge swarms, which upon descending on fields can devour the entire annual crop within minutes. Single swarms of the African plague locust Schistocerca gregaria may fan out over hundreds of square miles and consist of 50 billion (50,000,000,000) individuals, weighing about 70,000 tons. Densities may reach 200 million locusts per square mile (= 500 million/km²). In most species of locusts, this kind of behavior is seasonal, and some have two distinct morphological forms, the so-called solitary and gregarious phases. Sometimes the differences between the solitary and gregarious phases are so dramatic that they have been described as different species. The solitary forms occur in low densities, and usually have shorter wings and cryptic coloration. Under certain conditions, especially following rains, the densities of these grasshoppers increase dramatically, prompting great physiological, morphological, and behavioral transformations. Their coloration changes, usually to black and yellow, and they exhibit a strong tendency to form tightly packed groups that start marching across the land. Upon reaching the adulthood and the ability to fly, they take off in huge swarms, usually following the prevailing winds of a cold front that could lead them to areas of fresh vegetation.

Controlling massive outbreaks of grasshoppers is difficult and many different chemical and biological agents have been used. As one can imagine, controlling a cloud of insects stretching across hundreds of miles is a nearly impossible task. Therefore, the most effort is directed toward controlling the young, freshly emerged bands of hoppers or destroying eggs laid in areas that are likely to originate massive outbreaks. Fungal pathogens have been quite successful in combating locusts, but the survival of fungal spores in arid conditions, where many locust species live, is limited.

B. Major Lineages

1. Pamphagidae (Earth Hoppers or Pamphagids)

Members of this family are large, robust grasshoppers, seldom smaller than 30 mm and often up to 90 mm long. More than 560 species in more than 90 genera have been described, mostly from arid regions of Asia and Africa, with a few representatives in southern Europe. Sexual dimorphism is often marked, with males in general being significantly smaller and often fully winged, as opposed to completely wingless females. The surface of the body is often very rough, and many species blend superbly among rocks and pebbles. Some southern African species (Trachypteryllla) have been sometimes dubbed "living rocks."

Defensive stridulation is common in both nymphs and adults of many species, and it is achieved by various mechanisms (often more than one stridulatory mechanism is found in a single species). The typical advertisement stridulation is rare but, if present, the sound produced by males by rubbing their hind legs against the tegmina is exceptionally loud (loboscellana).

2. Pygromorphidae (Bush Hoppers)

More than 550 species are known in this family, placed in more than 150 genera. They are distributed mostly in the tropics of the Old World, with only a few species known from Mexico and Central American countries.

The body is usually large, often reaching 90 mm. The pronotum is extremely variable in form, sometimes adorned with granules or spine-like processes. Pygromorphidae have no stridulatory organs but some make
rustling noises with their wings when alarmed. Many members of this family have warning (aposematic) coloration, a characteristic frequently associated with their toxic properties. Some species eject protective froth or irritating fluids as defense mechanisms. Quite a few species are poisonous. Some species are gregarious, especially during their early nymphal stages. A few have been known to cause serious damage to crops in Africa (Zonocerus).

3. Romaleidae (Lubber Grasshoppers)

This family includes more than 470 species in more than 100 genera, distributed mostly in North and South America, with very few species in Asia and eastern Africa. In the United States the best-known representative of this group is the Florida grasshopper (Romalea microptera), a large and colorful grasshopper frequently used to study insect internal anatomy and external morphology.

Romaleidae are large grasshoppers, usually strikingly colored. Species of Tropidacris and Titanacris are some of the largest orthopterans, with wingspan reaching 250 mm. The pronotum in the members of this family is often strongly keeled or crested. Wings, if present, have dense venation, and the hindwings often have a unique stridulatory mechanism formed by numerous fine teeth on one of the veins. Sound is produced by rubbing this vein against the underside of the tegmen. In addition to stridulation, some species produce a hissing sound by expelling air through their thoracic spiracles. Species of this family have little economic significance, although a few species can defoliate trees on tropical plantations.

4. Paulolinidae (Aquatic Grasshoppers)

Members of this small family are unique among other grasshoppers in their nearly fully aquatic lifestyle. They are capable of skating on the surface of the water, diving, and swimming, and their eggs are laid on stems of submerged aquatic plants. The hind tibiae are flattened and widened at their apices, forming effective swimming paddles. Only two genera with five species belong to this family, and their original distribution was restricted to South America. Recently, one species (Paulinia acuminata) has been deliberately introduced to Africa in an effort to control the noxious aquatic plant Salvinia. This is probably the only example thus far of a truly beneficial species of grasshopper.

5. Acrididae (True Grasshoppers and Locusts)

This is the largest family of Acridoidea, with nearly 6760 species in more than 1350 genera and 27 subfamil-
ies. They occur in all regions of the world, and in nearly all terrestrial or semiaquatic habitats. The body size ranges from less than 5 mm to about 100 mm. Body form and shape are extremely variable, from robust and stocky to extremely slender and elongate. Wings can be fully developed, reduced, or entirely absent. Wing polymorphism is common in some groups (Gomphocerinae). Abdominal tympanal organs are generally present, absent only in some wingless forms. Many species stridulate by rubbing their hind femora against modified veins on the tegmina. Sometimes an expanded area on the tegmen forms a “speculum,” which acts as a resonator.

All species of locusts and most other economically important grasshoppers belong in this family. Also, most grasshoppers common in temperate zones of the world belong here, and nearly the entire grasshopper fauna of Australia is the result of a massive adaptive radiation within one of the lineages of this family (Cantantopinae).

VIII. TETRIGOIDEA (PYGMY GRASSHOPPERS OR GROUSE LOCUSTS)

This interesting lineage of the Orthoptera includes nearly 1200 species in about 230 genera, placed in one family, Tettigidae (some authors divide this superorder into two families: Tettigidae and Barachideidae) (Fig. 7). Most species of pygmy grasshoppers are tropical but quite a few occur in temperate zones, often at very high elevations. Their body size is generally small, seldom exceeding 20 mm. The pronotum is always very large.

FIGURE 7 Tettigoidae. An unidentified species from Costa Rica (Tettigidae).
covering the entire body and often extending far beyond the end of the abdomen. Sometimes the pronotum is vertically expanded, forming a large crest or a leaflike lobe. A few species display a polymorphism in the development of the pronotum. Wings are usually present but may be reduced or absent in some forms. The tegmina are always reduced to small, oval lobes, with greatly reduced venation. The hindwings are well-developed, fanlike, usually completely concealed under the pronotum. The hind femora are very stout and all tarsi lack the arolium between claws. Members of Tetrigoidae have neither stridulatory organs nor the abdominal auditory tympana.

Most species are associated with moist or semi-aquatic habitats. Many can swim very well and dive when threatened. Some species are arboreal and many live on bark of trees in tropical forests. Pygmy grasshoppers feed on a variety of small plants, such as mosses, as well as on lichens and algae. They have little economic importance although some may cause some damage in rice plantations.

IX. Tridactyloidea (False Mole Crickets and Sandgropers)

This superfamily includes about 230 species in 18 genera divided among three families (Fig. 8). Most species are small, 4 to 15 mm in length, although some sandgropers (Cylindrachetidae) can reach the length of about 40 mm. Families Tridactylidae and Rhipipteridae have characteristically modified hind tibiae, which bear a pair of long, apical flaps, while the tarsi are strongly reduced. Their tegmina, if present, are short and leathery and sometimes have a row of stridulatory pegs. The Tridactylidae are distributed worldwide whereas the Rhipipteridae are restricted to Central and South America. They inhabit wet and semi-aquatic habitats, and some of them make dense networks of shallow burrows in the mud.

The Cylindrachetidae are restricted to Australia and New Guinea, and one species is known from Patagonia. The body is strongly elongated and completely wingless. The pronotum is greatly elongated and cylindrical, with its lateral edges strongly curved downward, nearly meeting on the underside of the thorax. The front legs are modified for digging, and the second and third pairs are shortened. Sandgropers spend their entire lives underground in sandy soils. They are sometimes confused with the larvae of click beetles (Elateridae) as they occasionally tunnel into stems of plants.

See Also the Following Articles

INSECTS, OVERVIEW

Bibliography

GRAZING, EFFECTS OF

Mark Hay and Cynthia Kicklighter
Georgia Institute of Technology

I. Introduction
II. Grazers as Disturbance Regimes
III. Direct versus Indirect Effects
IV. The Roles of Spatial and Temporal Scales of Grazing
V. Host Chemistry as a Promoter of Consumer Diversity
VI. The Potential Importance of Paleo-patterns
VII. Using Grazers to Manage and Restore Ecosystems
VIII. Summary

GLOSSARY

associational refuge When a potential prey escapes or deters a consumer by associating with another organism that interferes with the ability of the consumer to locate or attack the prey.
bottom-up effects When physical parameters (such as nutrients) allow increased primary productivity and the effects of this cascade up through higher trophic levels. A simple example involves nutrient additions to a lake resulting in increased phytoplankton growth, this increasing resource of food allowing increased numbers of herbivorous zooplankton, and the abundance of zooplankton leading to increased densities of zooplankton-eating fishes.
co-evolution Joint evolution of two (or more) interacting species, each of which evolves in response to selection imposed by the other.

competitive release The expansion of a species-realized niche that is associated with the absence or removal of competition from other species.
grazer A consumer that removes only a part of each prey it attacks and thus rarely kills a prey in the short term. The term is commonly applied to animals that eat plants (herbivores) but can also be argued to apply to nonlethal microbial pathogens and to animals that cause tissue loss, but not death, when they feed from colonial animals such as sponges.
resource partitioning When organisms differentially utilize resources such as food, space, or nutrients.
secondary metabolite An unusual compound that does not play a role in primary metabolism (e.g., production and storage of energy). These metabolites were initially argued to be waste products but have often been shown to defend the producer from consumers, pathogens, or competitors.
top-down effects When feeding by higher trophic-level organisms has cascading effects on lower trophic levels. A simple example is when fishes (predators) consume zooplankton (herbivores) and, by lowering the numbers of these herbivores, allow increased densities of photosynthetic phytoplankton (primary producers).

GRAZING in the traditional sense of the term describes the actions of animals that consume parts of plants without causing the death of the plant; e.g., the con-
sumption of grass plants by domestic cattle. By extension, the term applies more generally to any feeding process that involves the partial consumption of the prey organism so as to bring about tissue loss but not death; e.g., the consumption of coral reef organisms by sea urchins and certain fishes. In either sense of the term, the process of grazing has profound implications for the structure and dynamics of the given community.

I. INTRODUCTION

Herbivores play a major role in determining plant biodiversity by strongly affecting distribution, abundance, and community organization. Because plants are the major primary producers and often constitute the major biogenic structural complexity around which other community members are organized, effects of herbivores on plants often have strong indirect effects on other trophic levels and on the biodiversity of the ecosystem as a whole.

Because almost all energy on Earth is fixed via plant photosynthesis, it is an evolutionary mandate that higher trophic levels will need to consume plants and that successful plants will need to resist, tolerate, or escape herbivores, pathogens, and other natural enemies. Humans are dependent on plants for food, clothing, shelter, fuels, pharmaceuticals, and other necessities; thus, we often compete with other consumers (e.g., insect herbivores and pathogens of crop plants) for desirable plant resources. Due to this continuing struggle to grow and harvest plant resources, plant-herbivore interactions are one of the more intensively studied areas in ecology, evolution, and agriculture.

Research on this topic provides a rich resource for evaluating the effects of grazing on biodiversity. In numerous instances, examples from both marine and terrestrial studies show similar patterns. However, because of the more rapid growth of marine plants, the more intense grazing by marine herbivores, and the greater ease of manipulating seaweeds versus trees or shrubs, some of the clearest examples investigating the mechanisms of how herbivores affect plants are from marine communities.

In subtidal marine communities such as coral reefs or kelp beds, it is not uncommon for herbivores to be one of the primary forces determining the distribution and abundance of plants and often the species composition and diversity of the entire community. For example, herbivory by fishes or sea urchins on coral reefs keeps reefs largely devoid of macroalgae and allows corals to flourish by reducing competition from the more rapidly growing seaweeds. When these herbivores severely declined on the Caribbean island of Jamaica due to a combination of overfishing and urchin disease, coral reefs suffered severe overgrowth by seaweeds, and coral cover declined from more than 50% to less than 9% throughout this entire island nation (Fig. 1). Thus, removing herbivores from these habitats converted species-rich coral reefs into a completely different community dominated by a limited number of seaweeds.

Equally dramatic changes occur in temperate systems. Numerous ecological and paleontological studies have shown that subtidal communities in the temperate eastern Pacific shift from sea urchin-grazed barren to lush kelp beds depending on the presence or absence of sea otters, which selectively forage on herbivorous sea urchins, reducing grazer biomass and allowing kelps to flourish. If otters are removed from kelp communities due to hunting or predation, urchin densities increase, urchins drive kelps to local extinction, and fishes, kelp-associated invertebrates, and seals or sea lions that feed on kelp bed fishes also decline dramatically. Thus, in the tropical example, grazers increase diversity by removing seaweeds and allowing corals to produce the biogenic structure that enhances reef biodiversity, whereas in the temperate system grazers directly remove the major biogenic structure (kelps) and cause a decline in overall biodiversity.

Herbivores can also cause major alterations in terrestrial communities. When myxomatosis eliminated rabbits in Britain, many grasslands reverted to scrub woodlands. Similarly, it was argued that grazing elephants

![Figure 1](image-url)
damage trees and change closed woodlands or thickets into grassy savannas.

The previously mentioned instances are among the most dramatic examples of herbivore effects, but they serve to show the effects that grazing can have on biodiversity and community structure and to point out that there should be strong selection for plant traits that deter feeding by herbivores.

II. GRAZERS AS DISTURBANCE REGIMES

Charles Darwin noted that one grass species tended to dominate and exclude other species in unknown areas of his lawn, but that numerous species coexisted in areas that were periodically mown. He recognized that grazers might play a similar role in diversifying plant communities by preventing a few rapidly growing species from competitively excluding others. In the late 1970s, Joe Connell formalized this relationship when he demonstrated for both tropical rain forests and coral reefs that intermediate frequencies or intensities of disturbance could lead to coexistence of a larger number of species. When disturbances were severe and frequent, only a limited number of hardy species could occupy a habitat. When disturbances were too infrequent or weak to remove competitive dominants, superior competitors tended to dominate. However, when disturbances occurred at intermediate levels, they removed enough of the competitively superior species to allow invasions by others but were not so severe as to prevent the occurrence of less hardy species. Thus, these intermediate levels of disturbance promoted the maintenance of greater biodiversity within a given community type. One of the most omnipresent forms of disturbance to plant communities is grazing, and it can play the diversifying role suggested by Connell. However, the effects that grazers have on plant biodiversity may also depend on the productivity potential of the environment. A recent review of how grazing affected plant diversity used data from a diverse range of marine, terrestrial, and fresh-water communities and found that grazing almost always increased plant diversity in environments characterized as having high nutrients (and presumably high potential growth rates), but that grazing often decreased diversity or had no effect on diversity when the environment was nutrient poor. These findings can be argued to mean that there are important grazer–nutrient interactions that need to be factored into the intermediate disturbance model. Alternatively, these findings simply mean that a given level of disturbance is “high” if it occurs in a habitat in which growth

rates (and thus the ability to recover from the disturbance) are low, but it is “intermediate” if it occurs in a habitat in which rapid growth allows quick recovery.

As an example of intermediate grazing diversifying plant communities, work by Hixon and Brostoff (Fig. 2) demonstrated that fish grazing rates on open areas of coral reefs were so intense that only the most resistant (coralline algae, which are red algae that resemble a thin calcified paint applied to a rock) or most tolerant (small filamentous algae that have basal sections embedded in the coral and that grow very rapidly) seaweeds could persist. In contrast, when fishes were excluded by cages, substrates were dominated by a few species of competitively superior macroalgae. However, in defended territories in which damselfishes drove most other herbivores away, bite rates on the substrate were intermediate between those experienced on open versus caged substrates, and at these intermediate rates of herbivory there were many more species of seaweeds. Similar patterns have been documented in Mediterranean grasslands in which plant species diversity is maximized if livestock densities are maintained at intermediate levels.

III. DIRECT VERSUS INDIRECT EFFECTS

Grazers may indirectly affect biodiversity via trampling, tunneling, seed dispersal, nutrient regeneration, and selection for plant traits that have cascading effects on other processes. As herbivorous prairie dogs dig their burrows, they create patches of newly disturbed ground that can be colonized by plant species that are rare in undisturbed habitats. Their activities also create mosaics of vegetation structure and nutrient status that result in both buffalo and antelope differentially using centers or edges of prairie dog colonies versus areas remote from colonies. Prairie dogs also clip inedible bushes to low heights so that the bushes do not provide ambush sites for predators or obscure the prairie dog’s field of view. Lowering bush height decreases competitive effects and allows increased coexistence of other plants. As animals graze, they release nutrient-rich urine and feces back into the environment, which affects nutrient dynamics, the spatial patchiness of plant production, and thus plant diversity. Such effects occur in habitats ranging from open-ocean gyres to African savannas and livestock rangelands.

Initial feeding on plants can also produce a cascade of indirect effects on habitat parameters that then affect use by other species. For example, when infestations of beetles feed on goldenrod, they not only affect their
target plant but also their activities alter light penetration to the ground, soil water content, and soil nitrate content. Following these changes in the physical habitat, areas near attacked goldenrod are invaded by other plant species, increasing local biodiversity.

IV. THE ROLES OF SPATIAL AND TEMPORAL SCALES OF GRAZING

Many investigations of herbivore effects on diversity have concentrated on the maintenance of within-habitat diversity due to selective feeding on dominant plants or on the effects of small-scale disturbances that create patches and produce numerous successional states within a single community. In contrast to these studies, several investigations from marine systems show that species richness of seaweeds is often maintained by mosaics of herbivore impact that are relatively predictable in space and persistent in time. Many seaweeds are predictably found only within certain physical habitats. This restricted distribution was initially assumed to be due to physiological limitations or to fine-scale resource partitioning along physical gradients; however, experimental studies demonstrated that many of these restrictions were herbivore induced and did not result from habitat partitioning based on competitive interactions or differing physiological requirements. When spatial variations in herbivory are experimentally reduced, habitats that differ dramatically in seaweed composition often become more uniform, usually resulting in a lowering of total species richness due to the loss of between-habitat diversity. In most cases, species that dominate following herbivore removal are palatable species that exclude, or significantly reduce the abundance of, herbivore-resistant plants. This is consistent with the hypothesis that the evolution of anti-herbivore defenses can be achieved only by diverting energy and nutrients from other needs. Thus, defenses appear to be costly, and in the absence of herbivory less defended individuals or species will have higher fitness than more heavily defended individuals or species. This relationship is depicted graphically in Fig. 3. When herbivores are rare, plant communities are dominated by competitively superior, but palatable, plants. When herbivores are common, defense is favored over rapid growth and competitive ability. Under these conditions, communities become dominated by unpalatable plants that depend on the herbivores to remove better competitors. If herbivores selectively graze competitive dominants, they tend to free resources for other species and increase biodiversity. However, if herbivores selectively graze...
FIGURE 3  The hypothesized relationship between plant abundance and herbivore effectiveness for palatable and unpalatable plants. a, inedible or defended plants; b, palatable plants with a refuge; c, palatable plants without a refuge. Spatial patchiness of herbivory within a habitat allows multiple points on the x-axis to exist within the same habitat, thus increasing local biodiversity (reproduced with permission from Annual Review of Ecology and Systematics, Volume 12, © 1981, by Annual Reviews www.annualreviews.org).

competitively inferior species, they may hasten local extinction of poorer competitors and lower biodiversity. This latter scenario appears to be relatively uncommon.

In the following sections, we discuss how herbivores affect plant diversity over a range of spatial scales. Patterns occurring over a scale of hundreds of kilometers are called geographic patterns. Patterns that occur over a scale of several meters to many kilometers and involve habitats that clearly differ in physical traits are called between-habitat patterns. Patterns occurring over only a few meters and within areas that appear relatively homogeneous are called within-habitat patterns. Patterns on a scale of centimeters or millimeters are called microhabitat patterns.

A. Geographic Patterns

Large-scale patterns in plant distribution are usually assumed to be due to physical parameters such as temperature. However, numerous correlative studies suggest that some portion of geographic constraints on both specific species (or phenotypes within species) and morphological types may be due to geographic variance in herbivory. Because rates of herbivory are higher on tropical than on temperate reefs, one might predict that tropical plants would be better defended than temperate plants. There are few direct tests, but a recent study demonstrated that seaweeds from tropical reefs were significantly less palatable and significantly better defended chemically than were similar seaweeds from temperate reefs. Similarly, studies of many species of salt-marsh plants along the coast of the eastern United States show that individuals from southern regions are less palatable to a range of common herbivores than are individuals of the same species from northern regions. These patterns suggest that herbivore-generated processes shown in Fig. 3 could function on geographic and local scales and could enhance global biodiversity by creating differing selective regimes across those latitudinal scales.

Because tropical herbivores encounter more chemically defended prey than do temperate herbivores, one might expect that tropical herbivores would be more tolerant of plant chemical defenses than would temperate herbivores. This contention has not been adequately tested, but initial tests of a few tropical versus temperate fishes and sea urchins are supportive of this hypothesis.

B. Between-Habitat Patterns

Near-shore reefs often support more seaweed biomass and a different community of seaweeds than offshore reefs. Although both physical parameters and herbivores differ between these habitats, field experiments demonstrate that these differences in seaweed community structure are determined to a significant degree by differential herbivore activity. Early studies in the Caribbean demonstrated that seagrass beds, unstructured sand plains, and shallow unstructured reef flats supported different plant communities than did immediately adjacent reefs. Although the different physical regimes of these habitats appeared to be an adequate explanation for the differing seaweed communities, caging and transplant studies demonstrated that removal of herbivores allowed a more homogeneous distribution of seaweed species and that unstructured reef flats supported different plant communities than did immediately adjacent reefs. Although the different physical regimes of these habitats appeared to be an adequate explanation for the differing seaweed communities, caging and transplant studies demonstrated that removal of herbivores allowed a more homogeneous distribution of seaweed species and that unstructured reef flats supported different plant communities than did immediately adjacent reefs. Although the different physical regimes of these habitats appeared to be an adequate explanation for the differing seaweed communities, caging and transplant studies demonstrated that removal of herbivores allowed a more homogeneous distribution of seaweed species and that unstructured reef flats supported different plant communities than did immediately adjacent reefs.
ability of insect herbivores to find or effectively utilize host plants with non-host species interferes with the plant–insect literature in which interspersion of capsules. There are many instances in agriculture and in vented their exclusion from the system. provided spatial refuges for palatable species and pre-

C. Within-Habitat Patterns

By aggressively excluding other herbivores from their algal mats, territorial damselfishes generate patches of intermediate grazing intensity in which algal species richness, evenness, and diversity are increased relative to either areas that are available to all grazers or areas from which grazers are excluded by cages. Several species that are normally excluded from unprotected areas due to grazing are found only within these mats. Thus, these territorial herbivores create distinct biotic patches on an otherwise more homogeneous background, raising species richness of the community at a whole.

In a similar manner, predators can strongly impact the spatial patterns of habitat use by herbivores and indirectly create mosaics of grazing intensity. In streams, algal-eating minnows avoid both shallow areas where they are susceptible to terrestrial predators and areas near larger predatory fishes. In both cases, these behaviors create patches of increased algal mass separated by heavily grazed habitats with minimal algal mass. Similar mosaics are created in terrestrial habitats when desert rodents forage more heavily under bushes than they do in more open areas where they are more exposed to predators. Within grasslands, excavating mammals, such as badgers, prairie dogs, and pocket gophers, provide newly bared soils for fugitive plants that are unable to successfully invade undisturbed plots. These microhabitat mosaics produce a series of successional states within the community and enhance plant diversity.

D. Microhabitat Patterns

In marine hard-substrate communities, small cracks and crevices serve as microhabitat escapes from some herbivores. Under these conditions, herbivory on exposed surfaces may select for defended seaweeds, whereas selection favors competitive ability in cracks and indentations. Work in experimental microcosms has shown that algal diversity decreases with increasing numbers of herbivorous fishes when substrate is smooth, but that species number changes little with grazer densities if the substrate is more topographically complex. Thus, the presence of microhabitat escapes provided spatial refuges for palatable species and prevented their exclusion from the system.

Some types of plants can generate microhabitat escapes. There are many instances in agriculture and in the plant–insect literature in which interspersion of host plants with non-host species interferes with the ability of insect herbivores to find or effectively utilize their host. Spiny desert plants sometimes function as ‘nurse plants’ where less well-defended juvenile plants escape grazing vertebrates until they become large enough to escape because of their size. Plankton communities also exhibit these associational refuges; herbivorous copepods may quit feeding on palatable species of phytoplankton when they are mixed with adequate densities of chemically noxious species.

The effects of associational refuges on species richness have been addressed experimentally in several marine communities. Numerous seaweeds that are commonly driven to local extinction by grazers persist in herbivore-rich communities by growing on or beneath their herbivore-resistant competitors. The brown seaweed Stypopodium zonale produces cytotoxic compounds that deter feeding by Caribbean reef fishes and urchins. Numerous species of seaweeds are significantly more common near the base of Stypopodium plants than several centimeters away. When the chemically deterrent plant is removed, these less defended species are rapidly removed by herbivores, decreasing local species diversity. When plastic mimics of Stypopodium are placed in the field, they provide a partial refuge for palatable species but they are less effective than the real plants, suggesting that associational refuges are generated in part by the physical presence of a non-food plant but that the plant’s chemical repugnance makes the associational refuge more effective. In temperate communities, palatable seaweeds can reduce losses to herbivores by growing on or near unpalatable seaweeds. Growing in close association with these unpalatable competitors drastically depresses the growth of palatable species, but the associational benefits, in terms of reduced herbivory, can more than offset this competitive cost. Thus, palatable species can be dependent on their unpalatable competitors to produce spatial refuges from herbivory and prevent their exclusion from the community. Both field and mesocosm studies indicate that removal of common unpalatable competitors can cause extinction, rather than competitive release, of associated competitors that are more palatable. These associational refuges were initially interpreted as arising from simple visual crypsis, but more detailed investigations suggest that chemistry plays a significant role.

As a final example, sulfuric acid can constitute up to 18% of the dry mass of the brown alga Desmarestia; this concentration is sufficient to dissolve barnacles from coastal rocks when this alga is deposited in the intertidal by waves. In Chilean kelp beds heavily grazed by sea urchins, the palatable kelp Macrocystis cannot successfully colonize unless it invades an area encircled
by Desmarestia plants, which appear to act as acid brooms that prevent urchins from entering the area. The associational escapes discussed previously are opportunistic rather than coevolved. As such, they may be used by many organisms in a wide variety of situations. There are, however, more intimate associational refuges or defenses that may be coevolved. Many microbes are predictably associated with specific species of macro-organisms. Because of the broad ability of microbes to produce bioactive secondary metabolites, many host organisms could be coevolved with certain microbes because the microbes produce compounds that defend the host from natural enemies. For example, certain fungi infect host grasses, produce toxins, and by doing so make the grasses much more resistant to herbivores. Similarly, marine cyanobacteria grow in host sponges and produce bioactive secondary metabolites that can protect the host. There are also instances in which marine bacteria that are specialized to certain host surfaces produce metabolites that chemically defend their host from microbial pathogens. In such cases, pathogens or consumers may be selecting for specialized microbial associates that defend their host. Such associations could increase microbial diversity and the diversity of macro-organisms by allowing hosts to persist in new habitats, thus facilitating both the evolution and the retention of increased species diversity.

V. HOST CHEMISTRY AS A PROMOTER OF CONSUMER DIVERSITY

On some tropical reefs, fishes have been reported to bite the bottom in excess of 150,000 times/m2/day. In these areas, small herbivores such as amphipods, polychaetes, and crabs (collectively called mesograzers because of their size) would live short lives if they occupied plants that were preferred by fishes. Selection should therefore favor sedentary mesograzers that live on and eat seaweeds that are chemically repellent to fishes. Patterns supporting this notion have been documented for divergent types of mesograzers in several of the world’s oceans. In the temperate Atlantic, some herbivorous amphipods and polychaetes live in tubes which they attach to the seaweeds they consume. These mesograzers selectively live on and feed from brown algae that are chemically defended from fishes. By living in association with these seaweeds, which are seldom visited by fishes, they lower their susceptibility to predation.

The hypothesis that sedentary mesograzers minimize predation by specializing on toxic hosts has been tested more broadly using (i) a specialist Caribbean amphipod that eats, and builds a mobile domicile from, a chemically defended alga; (ii) several species of crabs and sea slugs from both the Caribbean and tropical Pacific that each live on and feed from only one species of chemically noxious seaweed; (iv) a decorator crab that minimizes predation by selectively decorating with a seaweed that is chemically repellent to fishes; and (v) contrasts between the palatability and susceptibility to predation or parasitism of specialist versus generalist herbivorous insects. In all these cases, predation on, or palatability of, the mesograzers was reduced as a consequence of their association with chemically noxious hosts. Additionally, mesograzers were generally stimulated or unaffected by plant compounds that deterred feeding by larger herbivores or predators. There are numerous examples indicating that seaweeds or sessile invertebrates that evolve effective defenses against reef fishes may become evolutionary targets for specialized mesograzers that can escape or deter their own consumers by evolving a resistance to these compounds and living on, feeding from, and in some cases morphologically mimicking or sequestering defensive compounds from their toxic hosts. Thus, once hosts effectively deter common fish predators, these hosts can serve as valuable spatial escapes from predation for small consumers that specialize on these noxious hosts. As in many of the earlier examples, this creates spatial mosaics of selection within otherwise more uniform habitats and allows for the evolution and retention of more species within the system.

VI. THE POTENTIAL IMPORTANCE OF PALEO-PATTERNS

Plant–herbivore interactions occurring today are the result of both modern ecological forces and an evolutionary history that has been operating for many millions of years. Major changes in plants or in herbivores can select for fundamental and often cascading changes in plant–herbivore interactions. For example, in modern African communities, grazing by elephants causes repeated deforestation of savannas, prevents larger plants from replacing grasses, and facilitates retention of other herbivore species that require open-grazing habitats. It has been hypothesized that loss of elephants and other megaherbivores from Europe and the Amer-
VII. USING GRAZERS TO MANAGE AND RESTORE ECOSYSTEMS

There is a long history of using grazers to maintain or restore ecosystem structure, function, or diversity. Both productivity and plant species composition can be affected by manipulating the duration, intensity, and timing of grazing. Herbivore manipulations have been used...
to increase species diversity, to increase the aesthetic value of habitats, and to preserve endangered species. Most grazers being used as management tools come from terrestrial grasslands, but there are indications that this practice might also be useful in a wide variety of habitats ranging from forests and heathlands to coral reefs and lakes. In the following sections, we discuss some examples by habitat type.

A. Grasslands

Grazers are crucial to the maintenance of grasslands because without them there is an accumulation of plant litter that sequesters nutrients, physically limits vegetative growth, and interferes with seedling establishment. In a 10-year experiment focused on chalk grasslands in The Netherlands, sheep grazing helped rebuild species-rich grasslands that had been degraded due to agricultural use and heavy fertilization. When grazed and ungrazed plots were compared, the grazed plots had about 50% more species and the relative mass of forbs was increased threefold. Studies of other grassland systems and grazers have produced similar results. For example, sheep grazing can be used to reclaim heathland from woodland by opening up vegetation, repressing growth of scrub, promoting low-growing plants, and encouraging growth of dwarf shrubs. Experiments with rabbits and bison also yield similar results. In both cases, when the grazers were removed from grasslands, plant diversity declined.

Although grazing increased biodiversity in the previously mentioned studies, several of these management efforts were designed to have this result, thus potentially providing a biased impression of how grazing will affect plant species richness. Under some conditions, grazers decrease biodiversity. Effects are likely to vary as a function of the evolutionary history of the plants and herbivores, the types of habitats investigated, the levels of grazing employed, and when and for how long these grazing regimes are maintained. As an example of the variable effects that grazing can have on producer diversity, a recent study reviewed 44 comparisons of plant species richness under low versus high grazing pressure in nutrient-rich versus nutrient-poor ecosystems. High rates of grazing reduced plant species richness in 100% of the nutrient-poor situations. In contrast, high rates of grazing increased plant diversity in 50% of the high-nutrient contrasts; diversity was unaffected by grazing in 36% of these contrasts and decreased by grazing in 8% of the contrasts.

Manipulation of herbivores can also be used to alter habitat traits, such as productivity, rates of nutrient cycling, and nutrient, water, and organic levels in the soil. In general, herbivores often increase organic breakdown and the mineralization of potentially limiting nutrients such as nitrogen and phosphorous. In the Serengeti, the aboveground productivity of moderately grazed plots is stimulated to about twofold greater than the productivity of ungrazed plots. In addition, ungrazed grass stands are senescent, whereas those grazed by large herbivores produce younger and more palatable shoots.

Although mowing can crudely substitute for some grazer effects, grazing will generally produce greater diversity due to spatial mosaics produced by patchy grazing (versus even defoliation by mowing) and localized trampling. Selectivity and patchiness of grazing will also be affected by herbivore morphology and behavior. For example, teeth size and mouthpart morphology are important in determining the degree of selectivity exhibited by grazers. For example, cattle are less able to graze individual plants or plant parts than are sheep, goats, or horses.

Fire has also been used as a surrogate for grazing or, in conjunction with grazing, to manage plant community structure and species composition. This is especially true in attempts to prevent exotics and woody species from altering grasslands. Although fire may be a necessary disturbance in some systems, such as tallgrass prairie, fire alone is often not sufficient for restoration of biodiversity. After 9 years, in annually burned tallgrass prairie with nitrogen-addition, species diversity was 49% lower than it was 3 years before burning started and 66% lower than in unburned plots. In contrast, mowed, annually burned, nitrogen-addition plots had more than twice the species diversity than similar un-mowed plots. Therefore, mowing prevented a decrease in species diversity that would have occurred with only fire and nutrient addition. When bison grazing replaced mowing, plots that were grazed and burned had the highest species diversity of all plots. Similarly, in a Mediterranean grassland study, cattle grazing led to an increase in species diversity, whereas burning resulted in no significant difference.

B. Forests

Many tropical forest trees produce large fruits that either fall near the parent tree or are dispersed by birds, primates, and other large mammals to sites remote from the adult tree. Dispersal is so critical for some of these species that their seeds require gut passage through
dispersal agents before they can germinate. Because of the reliance on dispersers and the heavy predation on seeds near the parent canopy, it has been suggested that seed predators and dispersers are critical for maintaining tree species diversity in tropical forests. By selectively preying on seeds that fall near parent trees, specialized seed predators may decrease the survivorship of seeds near parents, preventing the occurrence of monospecific patches of forest trees and facilitating a more even and diverse assemblage of species. Because larger animals disperse the seeds of more than half of woody plant species, they play a critical role in removing seeds from sites of high predation and in introducing new species into different patches within a forest. The hunting, poaching, and habitat modifications that have left many forests depauperate in these animal dispersers have been suggested to be a cause of lowered plant diversity in tropical forests. If this is correct, then effective management of tropical forests may require the replenishment of large-bodied seed dispersers.

Livestock introduced to serve as seed dispersers can partially replace native dispersers that disappeared during the Pleistocene megafaunal extinction. The feeding activities of these introduced species can increase the range of some plants that produce large, fleshy fruits. In a lowland deciduous forest in Costa Rica, introduced horses and cattle feed on the fruits of jicaro (Crescentia alata) and disperse their seeds. In areas with livestock, jicaro are common. In areas in which livestock are absent, jicaro is relatively rare and occurs primarily in small, spatially restricted patches.

Many seed dispersers subsist exclusively on fruit for at least part of each year. This suggests that removal of seed dispersers by anthropogenic activities could have several repercussions. First, decreased seed dispersal could eventually lead to a decrease in the number of mature seed-producing parent plants. This would in turn lead to a decrease of important wildlife food resources, which could then lead to even lower numbers of seed dispersers. This situation can be especially critical if a particular tree species is pivotal in maintaining the health of dispersers during food-limiting seasons. For example, Casearia in tropical rain forests produce rich fruit during seasons of relatively low fruit abundance and likely supply much of the diet of local frugivorous dispersers. In addition, this species supports frugivorous birds that are very important dispersers to many other tree species during seasons when fruit is more abundant. The disappearance of animals that feed on Casearia fruits, leading to decreased recruitment of Casearia, could have far-reaching effects on the forest community. This scenario is thought to have occurred with jicaro. Frugivorous bats also feed on jicaro, and it is thought that a decrease in jicaro following the loss of Pleistocene seed dispersers led to a decrease in bats and then to a corresponding decrease in other fruit-producing plant species on which bats feed.

C. Lakes

Attempts to regulate the community composition of lakes have produced a mature science in which the separate and interactive effects of both physiochemical and biotic interactions are reasonably well understood and used in novel ways to manage lake ecosystems. Success in these systems has been achieved by fusing an understanding of biotic processes such as competition and predation (traditionally studied by population and community ecologists) with an understanding of the role of physiochemical processes (traditionally studied by limnologists and ecosystem ecologists). Investigations by Steven Carpenter and colleagues have been especially important in stimulating this field. Problems of harmful algal blooms, fish kills, and general eutrophication of lakes have become increasingly apparent. Initial attempts to explain these problems, and lake productivity in general, as a function of nutrient levels revealed that nutrients could vary considerably in lakes with similar biotic communities and that nutrient levels (also known as bottom-up effects) explained only approximately 50% of the variability in lake productivity. This finding prompted lake ecologists to investigate the effects of trophic cascades (or top-down effects) in structuring these ecosystems. By understanding and manipulating both the bottom-up effects of nutrients and the top-down effects of consumers, lake ecologists have been remarkably successful at altering biotic communities and fundamental processes (e.g., productivity and nutrient cycling) occurring in lakes.

Because lake communities are structured by interactions between physiochemical conditions and biotic processes, initial attempts to manage lakes by managing nutrient input met with only variable success. When the traditional focus on nutrients was merged with top-down manipulation of trophic structure, greater success was achieved in the control of lake phytoplankton populations and productivity. Mechanisms for this are as follows: As piscivore biomass increases, their feeding causes a decline in the biomass of small fishes that feed on herbivorous zooplankton; with the decline of these planktivorous fishes, the biomass of large herbivorous zooplankton increases and causes a decline in phytoplankton biomass and an increase in water clarity.
4). In all these instances, productivity of each group is maximized at intermediate levels of piscivore biomass. This system of trophic cascades allows managers to affect planktivore, herbivore, or phytoplankton biomass and productivity via manipulation of larger fishes (which can be done using fish additions, removals, or common fisheries management practices).

Whole lake manipulations have shown that piscivorous fish populations can regulate planktivorous fish populations in both North American and European lakes. For example, in an experiment by Carpenter and Kitchell, one lake had a high bass (piscivorous fish) population and a small minnow (zooplanktivorous fish) population, whereas another had a low bass population and a high minnow population. Ninety percent of the large bass population was placed in the bass-deficient lake, and 90% of the large minnow population was placed in the minnow-deficient lake. Where bass were added and minnows removed, the size of herbivorous zooplankton increased due to decreased predation. This led to decreases in chlorophyll concentration and primary production. In the opposite experiment, trophic interactions ultimately led to increases in chlorophyll concentration and primary productivity.

The interaction between trophic cascades and nutrient addition has also been addressed. In whole lake experiments in which phosphorous was added to lakes with a high or low density of piscivorous fishes, only a high density of piscivorous fishes was effective in maintaining low levels of palatable phytoplankton. The added nutrients increased the biomass of herbivores and colonial blue-green (unpalatable) algae but not the biomass of palatable algae. Grazers can control chlorophyll levels at phosphate loading significantly higher than loads that cause eutrophication. However, grazer control of undesirable blue-green algae appears to fail at phosphate loading rates lower than those used in this manipulation.

D. Coral Reefs

Feeding by large generalist herbivores, such as fishes and sea urchins, is critical for maintaining the coral-generated topographic complexity that helps maintain biodiversity on tropical coral reefs. If these herbivores are removed, reef corals are competitively excluded by seaweeds. For example, overharvesting led to severely depleted populations of herbivorous fishes on coral reef in Jamaica. However, the effects of overfishing were not fully realized for decades because the lowered numbers of predatory and competing herbivorous fishes allowed a build-up of herbivorous sea urchins that largely compensated for the declining fish grazing. When the urchin population was severely reduced by disease, macroalgal cover increased from 4 to 92%, severely overgrowing corals and inhibiting coral recruitment. The few grazers left on the reef (small parrotfishes and surgeonfishes) were not enough to control the algal growth, which has persisted for more than a decade. From scenarios such as this, it is evident that a higher density and diversity of grazers are required to decrease algal cover and allow coral recovery.

Manipulating these herbivores would therefore be a logical way to manage critical biotic processes on coral reefs. Because of the open nature of marine populations (i.e., the long-distance larval dispersal typical of most reef herbivores means that populations must be managed regionally rather than locally), this is a more challenging proposition than in the lake or grassland systems discussed previously. However, there is considerable evidence that degrading reefs recover, or degrade more slowly, when they are protected from fishing than when fishing is allowed to continue. The increased health of the protected reefs is correlated with an increase in herbivorous fishes. Managing reefs to facilitate targeted herbivores that are especially important in removing reef macrophytes is a promising tool for aiding coral reef recovery. This more directed approach is currently being initiated but is not yet a proven tool for coral reef management.

VIII. SUMMARY

Coexistence of potential competitors within diverse plant communities has often been explained as a result of fine-scale resource partitioning. This is assumed to have resulted from an evolutionary history of competitive encounters and is usually inferred from shifts in abundance that occur between habitats with different physical characteristics. Both marine and terrestrial...
plants show predictable patterns of distribution that can be interpreted as this type of habitat partitioning. However, manipulative experiments in marine systems, and to a lesser extent some terrestrial communities, show that many plants are habitat generalists and that both between- and within-habitat patterns of distribution are often controlled by herbivores. Spatial patterns in herbivore effectiveness create a mosaic of habitats that differ in the degree to which they favor poorly defended but competitively superior species versus well-defended but competitively inferior ones. If herbivory is decreased, the mosaic nature of the habitat is reduced and many well-defended species are excluded by competition. This pattern occurs across a large range of spatial scales and can explain between-habitat differences in plant communities that occur hundreds of meters apart, within-habitat differences that occur only a few meters apart, and microhabitat differences that occur on a scale of centimeters or millimeters. A portion of the microhabitat pattern is created by the plants. For example, consumption of palatable plants can be significantly reduced when they grow on, or are intermixed with, unpalatable plants. These associational refuges increase species richness by allowing palatable forms to invade grazed areas after unpalatable forms establish and create microsites of reduced herbivory. At all of the scales examined, herbivore activities that promote spatial mosaics of herbivore effectiveness, or that differentially affect dominant plants, help to maintain higher biodiversity.

Foraging by herbivores is often constrained by factors (e.g., predation) that do not directly affect the physiological performance of primary producers. For plants, this differential use of space by herbivores often produces a spatial mosaic of selective regimes in an area that would otherwise be treated as one uniform habitat; this interaction commonly results in elevated species richness. The strong effects of herbivores on plants, and the strong effects of plants on community structure in general, mean that manipulations of plant−herbivore interactions can often be used as an effective management tool for promoting the maintenance of biodiversity or ecosystem function.

See Also the Following Articles

- CATTLE, SHEEP, AND GOATS, ECOLOGICAL ROLE OF
- COEVOLUTION
- DISTURBANCE, MECHANISM OF
- PLANT-ANIMAL INTERACTIONS
- REEF ECOSYSTEMS
- RESOURCE PARTITIONING
- SEAGRASSES
- SPECIES COEXISTENCE

Bibliography

GREENHOUSE EFFECT

Jennifer A. Dunne and John Harte
University of California, Berkeley

GLOSSARY

aerosols Microscopic airborne particles.
albedo The fraction of light hitting a surface that is reflected.
anthropogenic Resulting from human activities.
climate sensitivity Long-term change in global mean surface temperature following a doubling of equivalent carbon dioxide (CO₂) concentration in the atmosphere.
Earth energy balance Average balancing of incoming solar energy by outgoing terrestrial radiation for Earth as a whole.
equivalent CO₂ concentration Concentration of carbon dioxide that would cause the same amount of radiative forcing as a mixture of carbon dioxide and other greenhouse gases.
feedback Change in a system component that triggers effects that eventually change the original component again. Feedbacks can be positive (self-reinforcing) or negative (self-dampening).
greenhouse gases Atmospheric gases that can absorb and reradiate infrared radiation.
radiative forcing Measure used to express and compare the potential of climate change factors to perturb the Earth energy balance, reported in watts per square meter (W m⁻²). A positive radiative forcing tends to warm the Earth’s surface and a negative radiative forcing tends to cool the surface.

I. INTRODUCTION

Earth’s climate, from daily weather events to glacial and interglacial cycles, is driven by the amount of radiation received from the sun and how that radiation is distributed throughout the global Earth–atmosphere system. The atmospheric greenhouse effect acts as an important factor in establishing a temperature that is hospitable for life. The basic mechanism is simple and was first
detailed by the Swedish physicist Svante August Arrhenius in 1896. Light from the sun largely penetrates the atmosphere and is absorbed at the planetary surface. There, it is converted from energy in the form of light to energy in the form of heat (longwave infrared radiation). As the surface temperature rises because of this heat, Earth radiates more and more heat back out to space, thereby maintaining an energy balance. Several gases in the atmosphere, referred to as “greenhouse gases,” absorb some of the heat emitted from Earth’s surface and reradiate it back toward the surface, causing the temperature to rise. Without this naturally occurring greenhouse effect, Earth’s average surface temperature would be ~10°C, about 33°C colder than it is today. The term ‘greenhouse effect,’ though popular, is a misnomer because the warming effect of glass greenhouses is due primarily to suppression of convection, not trapping of infrared radiation.

II. ENHANCED GREENHOUSE EFFECT

The most important naturally occurring greenhouse gases are water vapor, carbon dioxide (CO₂), methane (CH₄), nitrous oxide (N₂O), and ozone (O₃). Although water vapor plays the biggest role in generating the natural greenhouse effect, anthropogenic emission of these gases, along with artificially produced chloro-fluorocarbons (CFCs), are most important in generating an enhanced greenhouse effect. Detailed instrument data show that concentrations of these gases have been increasing since preindustrial times (~1750) (Table I), particularly in recent decades, largely due to human industrial, agricultural, and urbanization activities. As the concentrations of greenhouse gases increase in the atmosphere, they continue to trap and reradiate more and more heat, resulting in rising surface temperature and other climatic changes.

Increases in CO₂ account for about 65% of the current direct positive radiative forcing due to anthropogenic loading of greenhouse gases (Fig. 1). The atmospheric CO₂ concentration has increased 30% since preindustrial times, as a result of increasing emissions from fossil fuel combustion, land conversion, and cement production, and is continuing to increase by 0.4% per year (Fig. 2). If future emissions of CO₂ are maintained at 1994 levels, its atmospheric concentration will be close to double the preindustrial level by the end of the twenty-first century. In the absence of strong emissions controls, given increasing global energy and resource consumption, CO₂ concentrations may double by 2040 and will continue to increase dramatically (Fig. 3). Concentrations of other greenhouse gases, particularly methane and nitrous oxide, are also expected to rise, resulting in an earlier doubling of the equivalent CO₂ concentration. Greenhouse gases tend to remain in the atmosphere for many years (see Table I) and consequently are well mixed. They continue to affect the climate long after initial emissions and later stabilization of atmospheric concentrations.

Aerosols can alter the climate by changing atmospheric albedo. These fine particles absorb and reflect solar radiation and alter cloud properties. Sulfate aerosols from fossil fuel emissions and smelting tend to have a negative effect on radiative forcing and thus cool the climate. Current estimates of direct radiative forcing are ~0.5 W m⁻² due to aerosols compared to 2.45 W m⁻² due to greenhouse gases (see Fig. 1). Unlike greenhouse gases, aerosols are very short-lived in the atmosphere and therefore are not well mixed and respond rapidly to changes in emissions.

III. CLIMATIC CONSEQUENCES: GLOBAL WARMING

A. Past Climate Change

Scientists use deep ice cores drilled from glaciers in Antarctica, Greenland, and South America to examine ancient climate and atmospheric trends over the last several hundred thousand years. The data from these ice cores reveal a strong correlation between temperature and CO₂ concentrations (see Fig. 3) and suggest that a doubling of CO₂ has been historically associated with a 3 to 4°C temperature increase (Lorius et al., 1990). However, it is uncertain whether (1) the observed increases in CO₂ drove the warming or whether (2) warming due to planetary orbital changes drove increases in CO₂. If the first hypothesis is correct, the historic temperature sensitivity to natural CO₂ increases is comparable to that expected from anthropogenic loading of greenhouse gases (see Section III,C). If the second hypothesis is correct, current climate models may underestimate how much temperature will increase in response to anthropogenic additions of greenhouse gases to the atmosphere. If an initial warming (whether from orbital changes, greenhouse gases, or some other factor) results in more atmospheric CO₂ and other greenhouse gases, the increased gases would tend to drive additional warming. This would represent a positive feedback effect that is currently not included in climate models (see Section IV).
B. Recent Climate Change

Recent observational data not only document increased atmospheric concentrations of greenhouse gases, but also reveal the first “fingerprint” of human-induced global warming. Global average surface temperature trends reveal a 0.3 to 0.6°C temperature increase during the twentieth century, with the ten hottest years on instrument record occurring after 1980. Models that take into account both the positive effects of greenhouse gases and the negative effects of aerosols predict the observed upward trend of temperature quite closely (Fig. 4), including perturbations like the temporary decrease in global average temperature due to stratospheric aerosol loading from the 1991 Mt. Pinatubo volcano eruption. The current consensus among scientists is that “the balance of evidence suggests a discernible human influence on global climate” (Houghton et al., 1996) (see Box 1).

How do the recent temperature increases compare to natural climate variability? The long-term climatic record is quite variable, with changes of up to 1°C per decade during the most volatile periods of transition from the last glacial to the current interglacial period. However, in the last 10,000 years of the current interglacial period, fluctuations have generally not exceeded 1°C per century, making it unlikely that the recent observed trends and expected increases in temperature over the next 100 years are merely due to natural climate variability.

C. Future Climate Change

What will the future bring? General circulation models (GCMs) integrate the physics of radiative forcing, ocean dynamics, and other complex Earth–atmosphere processes in order to simulate and predict climate trends. The current consensus, based on multiple GCM forecasts, is that a doubling of equivalent CO₂ concentration (referred to as “2×CO₂”) in the atmosphere over preindustrial concentrations will result in a global mean average temperature increase of 1 to 3.5°C (Houghton et al., 1996). A 2×CO₂ atmosphere is used as a benchmark by climate change scientists to establish climate sensitivity and to compare scenarios and predictions, but greenhouse gas concentrations will continue to increase beyond a doubling, resulting in even greater temperature increases.

GCM and associated simulation studies point to many climatic impacts of increases in greenhouse gases and aerosols that will be of particular importance to ecosystems and biodiversity. First, surface warming will
not be uniform. Under a $2 \times CO_2$ atmosphere, average temperature increases are predicted to be greatest at the poles ($\approx 10^\circ C$) and least at the equator ($\leq 1^\circ C$). Winter temperatures will increase more than summer temperatures, and nighttime more than daytime temperatures. The incidence of record-breaking hot days will tend to increase in the summer, and fewer frost days are likely to occur in the winter. Second, warming is predicted to increase evaporation and global mean precipitation, and may increase cloud cover. High and middle latitudes and elevations are predicted to experience increases in winter precipitation, more winter precipitation falling as rain, earlier snowmelt, and reductions of summer soil moisture in noncoastal areas. Tropical precipitation is likely to change, but how it will change is uncertain. Third, the frequency and intensity of extreme weather and disturbance events (e.g., drought, deluge, summer heat waves, hurricanes, fires) are expected to increase. Fourth, glaciers are predicted to retreat and melt, sea levels may rise up to 93 cm as a result of a $2 \times CO_2$ atmosphere, and surface waters are likely to warm. Some of these trends have already been observed in recent climate data. For example, more warming has occurred toward the poles, nighttime temperatures have increased more than daytime temperatures, warming has been greatest over the midlatitude continents in winter and spring, snowpack is decreasing and snow-melt is occurring earlier at high latitudes and elevations, glaciers are retreating, and sea levels have risen 10–25 cm during the last 100 years.

IV. FEEDBACKS

A. Geophysical Feedbacks

Recent general circulation models include not only mechanisms underlying direct greenhouse gas and aerosol radiative forcing, but also mechanisms underlying three large feedback processes: water vapor, snow/ice albedo, and cloud cover. Because the capacity of the atmosphere to hold water vapor increases as it warms, and because water vapor acts as a greenhouse gas to...
further increase temperature, a positive feedback to the climate is created that amplifies warming. The snow/ice albedo effect is also a positive feedback—as warmer temperatures melt highly reflective snow and ice at the poles and high elevations, thus lowering surface albedo, Earth will absorb and radiate more infrared radiation, which will augment warming. Cloud formation adds much of the uncertainty to GCM estimates because the impacts of warming and other changes to the atmosphere on cloud formation are myriad and difficult to predict. In addition, because clouds can form at many heights, over land or water surfaces with differing albedos, and with many shapes, clouds can have multiple negative and positive feedback effects.

B. Biogeochemical Feedbacks

As complex as GCMs are, they fail to incorporate many potentially important aspects of chemical and ecological processes that are likely to produce additional feedback effects. Though it is likely that none of these feedbacks is individually as strong as the geophysical feedbacks already incorporated into GCMs, together they represent a potentially significant perturbation to the climate system. Biogeochemical feedbacks could greatly increase climate sensitivity to a $2 \times CO_2$ atmosphere (up to 8–10°C; Lashof, 1989) compared to current model predictions (1–3.5°C). Many biogeochemical feedbacks due to global warming will also interact with other anthropogenic stresses, such as deforestation and pollution, to further exacerbate or reduce climate change effects at local, regional, and global scales.

1. Marine Feedbacks

Global warming is likely to drive many complex marine-based feedback processes, of which only a few are mentioned here. With regard to geochemistry, the warming of ocean surface waters expected under climate change will reduce the capacity of those waters to hold dissolved CO$_2$, resulting in a positive feedback since less CO$_2$ will be removed from the atmosphere. If the additional surface water warmth travels down the water column, methane could be released from temperature- and pressure-sensitive hydrates that are present in some ocean floor sediments. This would also result in a positive feedback since the released methane, a greenhouse gas, would cause more warming.

With regard to biology, more than a third of annual global primary production occurs in ocean surface waters, primarily by microscopic single-celled organisms.
GREENHOUSE EFFECT

Critical Viewpoints and Responses

Although scientists currently have reached a strong consensus regarding the relationship between atmospheric loading of greenhouse gases and recent and projected global warming trends, a few critics express alternative viewpoints about climate change. Responses to some skeptical views follow:

1. The global warming trend over the past 100 years is the result of an increasingly brighter sun. No mechanism is known that could convert the observed very slight changes in solar output into a warming trend consistent with observations.

2. Satellite data disprove the hypothesis that recent warming is due to greenhouse gases. Satellite data cover too short of a period to disprove climate models, but the available satellite data (mostly on upper atmosphere temperatures) are reasonably consistent with model projections.

3. Most of the recent warming occurred earlier in the century but the atmosphere has only dramatically changed more recently. The warming trend since 1970 has been even more dramatic than that early in the century, consistent with the increasing rate of atmospheric greenhouse gas buildup.

4. A cooling trend during the period 1940–1970 contradicts our climate models. The observed cooling trend resulted from measured increasing levels of aerosols and dust in the atmosphere during that period. Although climate modelers initially failed to include atmospheric particulates in their analyses, when particles are included, the model predictions are consistent with observation.

5. We will welcome global warming because an ice age is coming. The timescale at which Earth will cool is over several thousand years and will not be counteracted by anthropogenic climate change. Over millennia, the current anthropogenic warming episode will slowly dissipate as humans stop loading greenhouse gases into the atmosphere and as most of the excess CO₂ is naturally sequestered in the deep oceans.
called phytoplankton, which form the base of the oceanic food web. Through the process of photosynthesis, CO₂ is fixed by the phytoplankton and thus transferred from the atmosphere to ocean surface waters. Some of that fixed carbon, in the form of dead bodies and fecal matter of phytoplankton and other organisms, sinks into deep ocean layers and sediments and is sequestered there, where it can no longer be exchanged with the atmosphere on short timescales. This process is referred to as the "biological carbon pump" and has been important in maintaining a level of CO₂ in the atmosphere. This is currently about 40% lower than it would be in the absence of marine organisms. Because the amount of marine primary production is dependent on the supply of nutrients and sunlight to ocean surface layers, any way in which global warming alters those inputs can create feedbacks to warming. One of many hypothesized feedbacks, in this case positive, is that global warming will tend to diminish the intensity of the oceanic circulation of nutrients, leading to more homogeneous, diffuse ocean productivity and hence a decrease in the amount of carbon "pumped" into deep ocean layers (Rowe and Baldauf, in Woodwell and Mackenzie, 1993).

2. Terrestrial Feedbacks

About three times as much carbon is stored in terrestrial vegetation and soils than is stored in the ocean. Through photosynthesis and respiration, more than one-eighth of atmospheric CO₂ is exchanged each year with terrestrial ecosystems. Changes to terrestrial-atmospheric carbon cycling thus have the potential to produce significant feedbacks to climate change. The feedback pathways for carbon in terrestrial ecosystems are complex, representing both positive and negative feedbacks. Perhaps the most well-known potential carbon cycle feedback is the "CO₂ fertilization effect," which refers to the stimulation of photosynthesis by increased levels of CO₂, which in turn can result in increased plant growth and greater storage of carbon in vegetation. This represents a negative feedback to global warming. However, ecosystem-level experiments appear to indicate that fertilization effects may tend to disappear after a few years, and the magnitude of effects may be strongly impacted by water and nutrient availability (Lashof et al., 1997).

Climatic changes that will accompany higher concentrations of CO₂ make it even more complicated to predict the net effect of global warming on the storage of carbon in vegetation and soil versus the atmosphere. For example, under global warming, changes in water availability and temperature will reduce the CO₂ uptake and growth of some plants while favoring others. Resulting changes in plant community composition can alter the quantity and quality of litter that enters the soil, which can lead to changes in soil carbon storage. Soil microorganisms will not only respond to changes in litter inputs, but will also be directly affected by changes in climate. Microbes and fungi tend to respire more CO₂ to the atmosphere as temperatures increase. However, rates of respiration depend on levels of soil moisture, and different extremes of water availability (both too much and too little) will tend to decrease respiration. The effects of climate change on microorganisms will also alter fluxes of other greenhouse gases such as methane in wetlands (e.g., northern peatlands, which store large amounts of carbon and may be a source of strong positive feedback to warming) and nitrous oxide in moist tropical soils.

In addition to the more direct effects of changes in temperature and moisture on the terrestrial carbon cycle, indirect effects such as the alteration of fire regimes due to climate change may produce significant feedbacks. In general, predicted increases in fire frequency for many ecosystems as a result of global warming may alter the structure of plant communities and result in reductions of terrestrial carbon storage, a positive feedback.

Another set of global warming feedbacks related to changes in albedo may result from climate-induced shifts in land cover and vegetation. The drying of soils and increased desertification expected from global warming will add dust to the atmosphere that, like aerosols, can reduce warming through increases in atmospheric albedo. Surface albedo is also expected to change as the boundaries of biomes shift, since different vegetation types can have different reflectivity. For example, the predicted northward expansion of boreal forest into tundra could decrease surface albedo, resulting in increased surface warming. This mechanism may have acted as a strong positive feedback 6000 years ago when an initial warming at high latitudes as a result of orbital variations appears to have doubled in magnitude owing to changes in surface albedo from boreal forest expansion (Foley et al., 1994).

V. CLIMATE CHANGE AND BIODIVERSITY

A. Introduction

Currently, the largest reductions in biodiversity result from massive deforestation in the tropics, in conjunc-
tion with other sources of worldwide habitat destruction. Even as the raining of vast tracts of tropical forests leads to immediate direct losses of hundreds to many thousands of species per year, the carbon that is released to the atmosphere through deforestation is amplifying the anthropogenic greenhouse effect. Global warming could lead to losses in biodiversity over the next several hundred years that are similar to or greater in magnitude than losses from direct habitat destruction. Dramatic changes in global climate have the potential to disrupt every ecosystem on Earth, leading to a pervasive trend of biodiversity loss due to climate-related habitat alteration, reorganization, and destruction.

Why does anthropogenic climate change present such a threat to biodiversity? There are two main reasons. First, the rate and magnitude of climate change expected over the next several decades to centuries are greater than any changes that current organisms have experienced. Over the last 18,000 years, starting during the last full glacial period and continuing through the current interglacial period that began about 10,000 years ago, average global surface temperature has gradually increased by about 5 ± 1°C (T. Webb, in Peters and Lovejoy, 1992). If we assume conservatively that global warming will increase mean temperature by 5°C over the next 200 years, this represents a 90-fold increase over the recent natural rate of change. In terms of magnitude, a 3°C increase would result in the warmest world in 100,000 years, and a 4°C increase would result in the warmest world in 40 million or more years. Second, global warming will interact synergistically with other anthropogenic stresses such as habitat destruction, pollution, ozone depletion, and alien species introduction to reduce biodiversity by more than just the sum of losses that would occur if each factor occurred independently.

Scientists’ ability to specifically predict how biodiversity will be affected by climate change is constrained by large uncertainties associated with local and regional climate change predictions. Much of the following discussion is based not so much on how scientists specifically think biodiversity will change under global warming, but on generic ways in which biota will be affected by global warming that can lead to changes in biodiversity. The term “biota” is used as shorthand to refer collectively to individual organisms, groups of the same types of organisms (populations, species, functional types), and ecological complexes of multiple populations and species (communities, ecosystems). This section summarizes how scientists study the effects of climate change on biota, the types of responses biota can have to climate change, the kinds of biota likely to be harmed by and to benefit from climate change, and evidence for biotic responses to current anthropogenic climate change.

B. How Scientists Study the Effects of Climate Change on Biota

1. Models

Climate–vegetation classification systems are types of static models that are based on the hypothesis that climate changes are the primary determinant of the broad-scale distribution of vegetation types. In 1947, L. Holdridge developed a “life-zone” concept that used three variables based on temperature and precipitation to predict under what climates 20 vegetation types should occur. Later researchers refined and added detail to this basic concept by using a wider variety of vegetation types and bioclimatic variables that explicitly incorporate drought stress and seasonality (Table II). These types of static models can be compared to current vegetation and climate maps to determine their accuracy and then used in conjunction with maps of simulated future climate to predict shifts in distribution of vegetation types due to global warming. The Holdridge life-zone classification has been used to predict the conversion of much of today’s boreal forest into temperate deciduous forest (Emanuel et al., 1985). While classification systems provide ways to look at potential global-scale impacts of climate change on vegetation, they are limited by the fact that climate will change continuously and with considerable interannual variation, rather than shifting abruptly to a new plateau. In addition, many other factors besides climate can influence the movement and distribution of biota, and this greatly reduces the potential accuracy of such models.

<table>
<thead>
<tr>
<th>TABLE II Bioclimatic Variables Used by E. O. Box to Predict Distribution Limits of Plant Types*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td>$T_{max}$</td>
</tr>
<tr>
<td>$T_{min}$</td>
</tr>
<tr>
<td>$D$</td>
</tr>
<tr>
<td>$P$</td>
</tr>
<tr>
<td>$P_{wet}$</td>
</tr>
<tr>
<td>$P_{dry}$</td>
</tr>
<tr>
<td>$MI$</td>
</tr>
</tbody>
</table>

Simulation models incorporate and link multiple factors such as life-history traits, physiological constraints, biotic interactions, resource availability, and climate in order to predict dynamic changes in biota over the course of time. By using characteristics of individuals, populations, species, or functional types in conjunction with predicted changes in climate, modelers can simulate continuous vegetation responses to global warming at scales that vary from changes within a stand of plants to changes at landscape, regional, and global levels. Vegetation models can be used in turn to look at changes in habitat and dynamics of other types of organisms. Simulation models not only attempt to predict patterns of change, but also provide a means to explore how and why such change might occur. However, these analyses are often constrained by data requirements. Detailed and accurate information may not be available about the biota and environment of interest, and as simulations are scaled up to look at global impacts, specificity is lost. No matter how well models appear to fit observed dynamics, some factor not included in the model may be important for future dynamics, or the model may fit observed dynamics for the wrong reasons, rendering model predictions inaccurate.

2. Paleobiology

Paleobiologists combine historical climate and biological data to reconstruct past relationships between changes in climate and species distributions. Paleobiological data sources range from ice and soil cores to plant and animal fossils to tree rings and slow-growing corals. Such studies reveal the long-term and integrated direct and indirect effects of past climatic and atmospheric change on past biota and thus facilitate projections about how present-day biota might respond in the long term to anthropogenic climate change. Many crucial insights about previous and potential effects of climate change have emerged from this type of work (see Section V.C). However, the generalizability of these types of studies is limited by the lack of good fossil records for many organisms, by the lack of strict climatic parallels in the past to both the rate and magnitude of anthropogenic climate change, and by differences in the biology and geology of ancient times compared to the present. In addition, the interactions of global warming with other anthropogenic stresses such as pollution, development, agriculture, and deforestation are novel and not represented in the paleorecord.

3. Natural Climate Variability

Whereas paleobiologists look to the ancient past for insight into the future, many field biologists interested in global warming look to the present to examine how biota are regulated by and respond to natural climate variability of a magnitude similar to that expected from global warming. Two approaches are used in this type of research. First, scientists can conduct “space-for-time” analyses along elevational or latitudinal gradients. This approach suggests that the effects of climate change over time on particular biota may be represented by current differences between the biota of interest and the same type of biota found at warmer, lower elevations or latitudes. Second, researchers who conduct multiyear studies at particular sites can monitor the response of biota to the natural interannual variability of climate. In particular, biotic responses to climate in more “normal” years can be compared to biotic responses to very warm years, droughts, early snowmelts, and other climatic events that fit predicted changes due to global warming. In some cases, researchers have access to records of climate and biota from several decades ago and can compare them to current records for the same sites.

This type of research has the advantage of actually studying current biota in the field in relation to climate, but it also has several disadvantages. For example, from one year to the next and from one place to another, sites vary not only in their climate but also in many other factors (e.g., land use history, species composition, topography), making it difficult to establish whether particular climate factors and/or other nonclimatic factors underlie observed patterns. Even if sites appear to differ primarily in climate, they may differ by predicted amounts of temperature but not by predicted changes in precipitation and soil moisture. Additionally, biota have had longer time periods to adjust and adapt to current climate variation at particular sites than they will have to adjust to climate change from global warming. Finally, this approach is unable to explore the effects of increased atmospheric CO₂ with the exception of some research on natural CO₂ gradients near hot springs.

4. Manipulations

Manipulations are one of scientists' most potent research tools. By conducting controlled manipulations of various climate, atmosphere, and other resource factors expected to change as a result of global warming, climate change researchers can work toward understanding the role of single or multiple factors and their interactions in changing ecosystem structure and function. This type of research is used to predict both how specific anthropogenic climate change scenarios might impact biota and how resulting ecosystem changes may pro-
duce feedbacks to the climate system. In climate change ecology, manipulative research falls into two types of approaches: microcosm experiments and field experiments. Microcosms, which generally take the form of laboratory growth chambers of various sizes, can be used to carefully manipulate particular global warming factors (e.g., temperature, moisture, light, nutrients, atmospheric composition) and to monitor the response of soils, single or multiple organisms, assembled simple ecosystems, or intact ecosystem cores taken from the field. Field experiments also manipulate global warming factors in order to look at interactions between climate change and biota, but do so in intact natural ecosystems (Fig. 5).

The strengths and weaknesses of these two approaches are interrelated. Although it is relatively easy to manipulate, control, and replicate microcosms, field experiments may be confounded by ecosystem variability and complexity. In field experiments, typically only a very few experimental variables can be manipulated, controls can be difficult to establish, and adequate replication is often expensive and time-consuming. However, field experiments have the advantage of being conducted in a natural setting at broader scales that may be more useful for drawing conclusions about complex, ‘real-world’ ecosystem dynamics, compared to highly simplified and small-scale microcosm experiments. In both types of experiments, the relatively abrupt, short-term manipulation of climate may not be a good analog of anthropogenic climate change, which is occurring more gradually over decades and centuries. In addition, changes to disturbances such as fires and hurricanes may prove to be more important in determining the abundance and distribution of biota in many ecosystems than the usual experimental focus on ‘average’ changes in variables such as temperature, moisture, and CO2.

5. Integrated Research
Given the limitations of each type of research, the most productive strategies for exploring interactions between ecosystems and anthropogenic climate change integrate multiple research approaches. For example, responses of biota to natural climate variation can be compared to responses of the same biota to manipulated climate change to see how responses differ or stay the same over multiple spatial and temporal scales. Results from field experiments and gradient studies often suggest mechanisms that can be more thoroughly tested in microcosm experiments. Gradient, field experiment, and paleobiological data sets can be used to parameterize, calibrate, and validate models of biotic response. In general, the thoughtful integration of approaches can build on the strengths and avoid some of the limitations of each type of research, thus helping scientists to develop more rigorous hypotheses about global warming impacts and ecosystem feedbacks.

C. Impacts of Climate Change on Biota
1. Types of Responses of Organisms, Populations, and Species
a. Adjustment
The first level, short-term response of any organism to changes in their environment is adjustment, also referred to as acclimatization. All organisms have some...
degree of physiological, life-history, or behavioral plasticity that enables them to live in a variable environment. The degree of plasticity with regard to climatic and atmospheric conditions varies widely among different kinds of organisms. Therefore, some types of organisms will be able to adjust to relatively large changes in climate, whereas others will be unable to adjust to even apparently minor increases in temperature or slight variations in precipitation.

An example of climatic adjustment in animals involves thermoregulation in vertebrates. Endotherms such as mammals have built-in physiological mechanisms to cope with body temperature changes. Endotherms such as reptiles have behavioral traits that help regulate body temperature. Because of traits such as these, initial increases in environmental temperature should be well within the tolerances of many vertebrates. In plants, the concurrent increase of atmospheric CO2 with surface temperature may augment the ability of some individuals, populations, and species to adjust to and flourish under anthropogenic climate change. Increases in CO2, especially for plants with the C4 photosynthesis pathway (e.g., most trees and shrubs), can result, at least initially, in the CO2 fertilization effect mentioned before. Enhanced CO2 concentrations can increase the ability of these types of plants to tolerate water stress, higher temperatures, and lower light levels. Other kinds of plants, particularly those with the C3 photosynthesis pathway (e.g., many low-latitude and low-elevation grasses), have physiological mechanisms that enable them to withstand warm temperatures and low availability of water. Such mechanisms provide a means of adjustment to drought stress that may be associated with increased temperatures and evaporation.

While most biota will have at least some capacity to withstand, and in some cases benefit from, initial changes in climate, the rapid rate and large magnitude of climate change are likely to quickly surpass their capacity to adjust to new climate conditions within their pre-warming habitats. Biota that cannot continue to adjust will have to respond through evolution, migration, or extinction.

b. Evolution

Theoretically, populations and species could develop new adaptive traits as a result of evolution in response to anthropogenic climate change, thus enhancing the long-term survival of current taxa under new climate conditions. However, scientists generally agree that evolutionary responses to climate change are unlikely for most taxa since climate is changing rapidly compared to usual rates of evolutionary change. This view is supported by fossil data that reveal the morphological stasis of many taxa during previous periods of rapid climate change.

In the face of strong selectional pressure there is evidence that some species, especially those with fast generation times, can evolve very rapidly. For example, grass populations grown on soils polluted by heavy metals have shown signs of significant, genetic-based, heavy-metal tolerance within one or two decades. Some insects can evolve increased resistance to pesticides over the course of a few years. These types of responses depend on the presence of appropriate genetic variability in populations and species relative to a strong selective factor. As a result of global warming, populations and species will be exposed to novel environments resulting from climate change and associated shifts in ecosystem structure and function. Since many populations and species have climate-related genetic variability (e.g., differences in high temperature tolerance, drought tolerance), rapid adaptation is possible.

Will anthropogenic climate change actually result in directed selectional pressures that are strong enough to drive microevolutionary responses? For at least the next several hundred years, climate and species distributions are likely to be in fairly constant flux, which will tend to disrupt any potentially adaptive trends. Also, rapid evolutionary responses to anthropogenic climate change are unlikely in populations and species that have relatively long generation times, such as trees and many vertebrates. An added constraint on potential microevolutionary responses to global warming is the ongoing reduction in population size and thus genetic diversity of many species as a result of habitat destruction and other stresses. For most biota, other types of responses are far more likely to occur than evolution.

c. Migration

As current habitat becomes inhospitable owing to direct and indirect effects of climate change, biota will tend to track shifting climate and suitable habitat through dispersal and migration. Consequently, as a result of global warming, organisms are predicted to move generally poleward in latitude and upward in elevation. A rule of thumb is that a 3°C change in temperature is approximately equivalent to a move of 250 km of latitude or 300 m of elevation. However, migration will be restricted or made impossible to the extent that there are inherent (e.g., low mobility, slow reproductive rates) or external barriers (e.g., mountain ranges, large lakes) to movement (see Boxes 2 and 3).
Given our knowledge of climate change and biology, the kinds of biota most likely to be at risk from global warming over the next several decades and centuries can be characterized:

1. **Those at higher latitudes:** Scientists know with a high degree of certainty that temperature increases due to global warming will be greatest in polar regions. Therefore, higher-latitude temperate and Arctic/Antarctic ecosystems such as boreal forest, tundra, and peat bogs will experience both rapid and severe temperature increases, resulting in profound biotic change and disruption.

2. **Those on mountain-tops:** Temperature increases will also be greater at higher elevations. Montane biota will tend to move up in elevation as cooler, higher elevations warm. Biota already limited to mountaintops will be at serious risk of local extinction due to alteration of summit climate, the lack of potentially suitable habitat to migrate to, and the encroachment of lower-elevation species.

3. **Those in low-lying coastal areas and on islands:** Even small increases in sea level (i.e., several centimeters) can result in altered coastal marine dynamics and flooding of low-lying areas. Rising sea levels will destroy or cause severe damage to ecosystems at the terrestrial/marine interface, such as salt marshes, estuaries, mangroves, and sand dunes, and are likely to disrupt coastal marine food webs.

4. **Those sensitive to extreme disturbance events:** Even though disturbances such as fires and hurricanes are a natural part of ecosystem dynamics, any increases in frequency and intensity of such disturbances due to global warming are likely to disrupt biota. For example, ecosystems such as tropical montane forests may have less time to recover between hurricanes, limiting the development of slow-growing, late-successional species.

5. **Those with migration problems:** Because a major potential response of biota to climate change is to migrate to new areas, biota that lack the ability to readily disperse or move will be at a serious disadvantage. These include:

   (a) plants whose seed or clone dispersal rates and animals whose movement rates lag behind rates of climate change;
   (b) slow-growing populations and species that will not have time to adjust to new conditions;
   (c) biota that cannot or are slow to cross geographic barriers, for example, fish in isolated lakes, low-elevation plants bounded to the north by mountain ranges, and tropical forest birds and insects that do not cross unforest areas;
   (d) organisms that depend on other biota for habitat or food, but that have very different degrees of mobility;
   (e) species such as monarch butterflies and migratory shorebirds that have multiple habitat requirements; and
   (f) relic biota that have been left in small, unusual habitats by chance and have no nearby potential habitat.

6. **Those that are rare:** Populations and species with few numbers, low genetic variability, or limited or unusual ranges will be vulnerable to climate fluctuations and will be at increased risk of extinction. Also, unusual, unique ecosystems may break apart as populations and species respond in largely individualistic ways to climate change.

7. **Those dependent on particular hydrological regimes:** Though it is often uncertain at local and regional scales how and to what degree precipitation and moisture availability will change from global warming, it is quite certain that change will occur. Such changes could be critical in ecosystems such as tropical forests where the availability of food resources for animals is dependent on the timing of rainfall. In montane areas, many organisms will be very sensitive to changes to the snowpack and snowmelt.

8. **Those close to critical physiological thresholds:** Many organisms are adapted to living within a narrow range of limits of temperature, moisture, nutrients, light, and atmospheric composition. Others have wider tolerances but are already operating close to a threshold, beyond which their ability to live, grow, and reproduce is severely limited. Climate change may force the environment past such limits for some organisms, resulting in severe im-
pacts, especially if dispersal or growth is slow. For example, slight increases (1–2°C) in surface water temperature can induce bleaching and mortality of coral reefs, which have very slow rates of growth and provide habitat for many marine species.

9. Those that have highly specialized relationships with other organisms: Some species depend entirely on just one or a very few other species for nourishment or reproduction. If species respond very differently to changes in climate than do the species they depend on, and they cannot substitute other organisms to fulfill those roles, they will go extinct.

10. Those negatively affected by other anthropogenic stresses: Humans engage in many activities that result in deleterious impacts such as acid deposition, pollution, ozone depletion, and alien species introduction. When organisms are weakened by one of these stresses, they tend to become even more vulnerable to other stresses such as global warming. For example, insect pests can damage vegetation more when pollutants reduce plants’ resistance to herbivory and warmer temperatures encourage pest population growth. If those insect pests happen to be alien, they may cause even more damage owing to lack of local predators and because local plants may lack resistance to alien pests. Also, changes in land use such as deforestation can reduce and isolate populations as well as create barriers to migration through habitat fragmentation and destruction.

Fossil records show that for many types of organisms, warming during the last deglaciation induced significant changes in latitude and elevation of species’ ranges (Fig. 6). Those distributional changes sometimes occurred at very rapid rates. For example, peak migration rates for some tree species in North America during the last deglaciation reached 100–300 m per year, probably as a result of haphazard, long-distance transport of seeds by animals, storms, or water (Clark et al., 1998). However, even these very fast historic migration rates only translate into 10–50 km per century, whereas anthropogenic climate change will likely require latitudinal shifts of at least 200–300 km over the next century. In some cases, changes in potential range bound-

<table>
<thead>
<tr>
<th>Box 3 Biota That May Benefit from Climate Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Because of the slow rate of most evolutionary responses, including speciation, there are few ways that global warming could augment global biodiversity over the next several hundred years. However, climate change will benefit some biota through increases in abundance and range expansions, often at the expense of more at-risk biota. Types of biota likely to benefit from climate change include:</td>
</tr>
<tr>
<td>1. Those that migrate easily: Biota that are highly mobile and have rapid dispersal rates, such as some kinds of winged insects, will be equipped to track changing climate.</td>
</tr>
<tr>
<td>2. Those that are opportunistic: Opportunistic organisms that can colonize disturbed areas will be at an advantage because they will be able to migrate through marginal habitat and establish in climatically disrupted ecosystems. For example, global warming is expected to promote the spread of already weedy introduced plant species, and may facilitate the escape of more garden cultivars into natural ecosystems.</td>
</tr>
<tr>
<td>3. Those that are ecological generalists: Organisms that flourish in a wide variety of environments and have either wide tolerances for variable resource availability and climate or many possible prey items will fare better than highly specialized organisms.</td>
</tr>
<tr>
<td>4. Those that have high variability and rapid reproduction: Populations and species with lots of phenotypic or genetic variation and rapid reproductive rates have the best chances of adjusting and adapting to rapidly changing climate.</td>
</tr>
<tr>
<td>5. Those favored by new optima: Although climate and atmospheric conditions will shift away from optima for many organisms, more optimal conditions will be created for other organisms. For example, temperature increases are expected to increase parasite and insect development time, allowing parasites to spread with migrating insect hosts and promoting pest infestation and parasite infection of new hosts. These types of responses may lead to range expansions of agricultural pests and disease transmissions, as well as more frequent outbreaks (Dobson and Carper, in Peters and Lovejoy, 1992).</td>
</tr>
</tbody>
</table>
FIGURE 6 Maps of observed (upper row) and simulated (lower row) percentages for spruce tree pollen in eastern North America over the last 18,000 years. The simulated spruce pollen maps are based on the modern response of spruce pollen percentages to July temperature, January temperature, and precipitation as applied to simulated historical climates. Dark shading indicates the highest abundance of spruce. (Reprinted with permission from T. Webb, III. Past changes in vegetation and climate: Lessons for the future, in Peters and Lovejoy, 1992. As adapted with permission from COHMAP Members. (1988). Climatic changes of the last 18,000 years: Observations and model simulations. Science 241, 1043. Copyright © 1988 by AAAS.)

acies resulting from climate change due to doubled CO₂ may exceed 1000 km. For example, suitable habitat for beech trees, which currently grow over most of the eastern third of the United States, could shift almost completely out of the country and into a much smaller area of the northeastern United States and Southeastern Canada as a result of a 2 × CO₂ atmosphere (Fig. 7). The rapid pace of anthropogenic climate change will easily outstrip the capacity of some organisms to move or disperse to suitable new habitat. In addition, human destruction of habitat will create insurmountable barriers to migration for biota that have difficulty crossing large areas of urban development or agricultural use. In the short run, any migration that does occur may tend to increase local levels of biodiversity in some ecosystems as new species move in before old species completely migrate or die out. This lag effect will tend to disappear with time.

d. Extinction

Biota that are unable to adjust, evolve, or migrate are unlikely to survive long in their pre-warming habitats. As a result of climate change, organisms may be exposed to increasing physiological stress, be abandoned by their mutualists and prey, be outcompeted by more flexible neighbors or incoming species, and be attacked by new predators and pests. Although microevolutionary responses are possible for some populations, rapid genetic adaptation is likely to occur in very few. Many species will be faced with migration problems. Even in species that successfully migrate, some populations are likely to go extinct, particularly at southern and lower edges of species’ ranges, reducing genetic variability. Extinctions attributable primarily to global warming will be few over the next several decades, but will undoubtedly increase dramatically as time passes, climate change intensifies, and biotic response options narrow.

2. Impacts on Biotic Assemblages

A key insight from the fossil record is that species tend to respond to climate change individualistically rather than as a group. Thus, while the general trend is for species to move poleward and to higher elevations as climate warms, particular species can vary quite dramatically in how fast and how much their ranges contract, expand, or move, what directions they move in, and at what rate they move around or over barriers such as mountain ranges. Consequently, past communities repeatedly disassociated and re-sorted into novel combinations. In some cases, although the same species still exist, there are no modern examples of historic species associations. For example, for several thousand years at the end of the last glacial period, spruce trees grew in open parklands in association with sedges. Today spruce is found in a completely different ecosystem type, the closed-canopy
boreal forest, in association with birch, alder, and fir (Webb, in Peters and Lovejoy, 1992). Anthropogenic climate change is likely to result in the reconstitution of communities and ecosystems in unexpected ways. Biotas tend to move individualistically in response to long-term climate change because tolerances to climatic and atmospheric conditions are often specific to the organism, population, or species. Individualistic responses can lead to apparently counterintuitive shifts in range, especially if temperature is not the primary determinant of distribution. For example, the distribution of the gopher tortoise during the most recent deglaciation shifted south, rather than north. One hypothesis for this pattern is that seasonal climate extremes increased with warming, and that these extremes were more important for determining tortoise distribution than warming (Graham, in Peters and Lovejoy, 1992). In addition to climate and atmosphere, other abiotic factors (e.g., soil type, topography, disturbance regime, site history) and biotic interactions (e.g., mutualism, competition, predation, pollination) can play significant roles in determining the abundance and distribution of biota in both the short and long term. The interplay of all of these factors over time can result in complex and often unpredictable changes in the distribution, abundance, and diversity of biota.

Global warming will also precipitate many asynchronies that reduce the ability of biota to respond effectively to climate change. "Asynchrony" refers to a mismatch in timing or rate of change. One type of asynchrony already discussed is the mismatch between very rapid rates of anthropogenic climate change and slower rates of dispersal and migration for many species. Another type of asynchrony due to global warming is the potential mismatch between required resources and the availability of resources. Organisms are embedded within a network of relationships with other organisms, which may depend on for sustenance and reproduction. However, different types of organisms may be affected very differently by climate change, which can disrupt biotic relationships that are important for community and ecosystem dynamics. For example, many species of flowering plants depend on specific animal pollinators such as butterflies, hummingbirds, and bees for successful reproduction, and the pollinators depend on those plants for food. In a hypothetical example, if the timing of flowering in a plant is primarily determined by early season temperature, while the emergence and activity of its pollinator is primarily determined by the amount of daylight, the shifting of temperature but not light levels due to global warming could lead to the pollinators being active after the plants flower. This type of
asynchrony can reduce the reproduction and abundance of the plants, and reduce the availability of food for the pollinators. If the species involved are highly specialized on each other, they are likely to go extinct.

D. Evidence for Current Global Warming Impacts on Biota

During the 1990s, researchers started reporting the first evidence that global warming during the twentieth century has begun to influence populations, species, and ecosystems (see Box 4 for a bibliography of the sources cited in this section). Some evidence comes from coastal marine systems. Roemmich and McGowan (1995) reported that since 1951, zooplankton biomass in coastal southern California waters had decreased by 80% over four decades, at the same time that surface water layers warmed more than 1.5°C in some areas. They suggested that the surface warming resulted in changes to stratification and the thermocline that led to a reduction in upwelling of nutrients and thus primary production by phytoplankton, the ultimate food source of zooplankton. Barry and colleagues (1995) compared changes in abundances of 45 invertebrate species in a central California intertidal community from the 1930s to the 1990s. During that period of time, annual mean shoreline temperature at the study area increased by 0.75°C, and average summer temperature increased by 2.2°C. Eight of nine species with a southern geographic range showed significant increases in abundance, whereas five of eight species with a northern distribution showed significant decreases. Species with wide distributions showed no strong patterns of change.

At high elevations and latitudes, terrestrial vegetation appears to be responding to the warmer climate. Grabherr and colleagues (1994) compared 40- to 90-year-old records of plant distribution on high-altitude mountaintops (~2000 to 3500 m) in the Alps to 1992 data, given a 0.7°C increase in mean annual temperature during that time period. They found upward migration rates by nine “typical” species of <1 to 4 m per decade. In addition, species richness increased over time, with the increase most pronounced at lower-altitude summits. Myneni and colleagues (1997) presented evidence from satellite data from 1981 to 1991 that suggests that the photosynthetic activity of terrestrial vegetation in northern high latitudes increased by 7–14% over that time period. This increase may indicate an intensification of plant growth associated with a 12-day increase in the growing season.

Insect species are also displaying sensitivity to recent climate change. Parmesan (1996) examined historical records and recorded the current status of 151 populations of Edith’s checkerspot butterfly throughout its entire range. She found that net extinctions of populations were significantly greater at southern latitudes and lower altitudes, resulting in an observed northward and upward shift in the species’ range during the last several decades. In a later study that examined range shifts in 35 nonmigratory European butterfly species during the last century, Parmesan and colleagues (1999) reported that 63% of the species displayed northward range shifts of 35 to 240 km, while only 3% shifted south. Vertebrates also appear to be sensitive to recent climate change. Current evidence suggests that birds are both breeding earlier and shifting their ranges northward in response to warming. Brown and colleagues (1990) found that from 1971 to 1998, the average timing of first clutch in the Mexican jay in Arizona occurred 10 days earlier. This change was associated with significant increases in monthly minimum temperatures. Crick and Sparks (1999) studied 20 United Kingdom

**Box 4**

Selected Bibliography for Current Global Warming Impacts on Biota


bird species over 23 years and found long-term trends toward earlier egg laying in most species. A data set of 36 bird species over 37 years suggests that 86% of the species display significant relationships between timing of egg laying and temperature or rainfall. Furthermore, Thomas and Lennon (1999) examined the breeding distributions of British birds over a recent 20-year period and found that the northern margins of the species’ ranges moved north by an average of 19 km.

In general, the observations of all of these studies are consistent with predictions of impacts of climate change on biota, and demonstrate how just a small amount of climate change over brief time periods can lead to significant ecological changes.

VI. CONCLUSIONS

The enhanced greenhouse effect, in conjunction with other anthropogenic stresses, is likely to precipitate unprecedented changes to Earth’s climate and ecosystems. Though the details of how climate change will affect biodiversity are often hard to predict, there is little doubt that biological impacts will be pervasive and often dramatic. Studying the effects of climate change on biota can help in the formulation of strategies for conserving biodiversity and ecosystem structure and function in the face of potentially massive change and loss. Such knowledge is also crucial for refining predictions of the future rate and magnitude of global warming, since biological responses are likely to produce significant feedbacks that can augment or dampen climate change at local, regional, and global scales. Understanding and addressing the interactions between climate change and biodiversity represents one of the greatest challenges that scientists and policymakers will face in the twenty-first century.

See Also the Following Articles

ATMOSPHERIC GASES • BIOGEOCHEMICAL CYCLES • CARBON CYCLE • CLIMATE CHANGE AND ECOLOGY • SYNERGISM OF • CLIMATE, EFFECTS OF • DEFORESTATION • MIGRATION • SOIL BIOTA, SYSTEMS AND PROCESSES

Bibliography


I. The Definition and Properties of Guilds
II. The Utility of Guilds in the Search for Organization and Patterns
III. The Individualistic Nature of Species and the Vague Boundaries of Ecological Categories
IV. Attitudes Concerning Vaguely Bounded Categories

GLOSSARY

**adaptive syndrome** A coordinated set of adaptations.

**community** An assemblage of populations that coexist in an area.

**Etolian niche** The role, or occupation, of a species in a community.

**exploitative competition** An adverse interaction which results from organisms depleting their shared resources.

**family** A category, in the classification of evolutionary lineages, of related organisms that ranks above a genus and below an order. A family usually contains many genera.

**Grinnellian niche** The requirements and behaviors expressed by a species wherever it normally occurs.

**Hutchinsonian niche** The set of environmental conditions, or opportunities, that will permit a species to exist indefinitely. The set of opportunities that are available to a guild can be referred to as a "nook."

**individualism of species** Refers to the observation that the distributions and abilities of most species do not exactly coincide because each species is the product of a unique evolutionary history. Coevolution and convergence may occur, but rarely are these processes so complete that groups of species are distributed as a unit or superorganism.

**taxon** A general term used to indicate groups of related organisms at any level in a taxonomic hierarchy. In the classification of evolutionary lineages, the term can refer to species, genera, families, orders, etc.

**INVESTIGATING ENTIRE** natural communities is a formidable task because one must take a high diversity of species into account. To focus on a more manageable unit, ecologists usually restrict their attention to some portion of the larger system, such as the "plant" or "bird" community. Such "taxon"-defined communities, however, are rather unwieldy, and they may contain a heterogeneous mix of interactions. Thus, most beetle communities contain predators, herbivores, and scavengers but ignore the bugs, caterpillars, flies, and gastropods that share many of the beetles' resources. The guild concept, which groups species according to the manner in which they exploit a common resource, provides a manageable, functional unit for studying patterns of adaptation and the organization of natural communities. The existence of species with mixed requirements and different evolutionary histories, how-
ever, makes it difficult to determine standard procedures for defining guilds. The indefinite boundaries that surround most guilds are a fundamental outcome of evolution, which is also at play in all of our attempts to group species into communities, trophic levels, and other ecological categories. Discovering how to draw valid conclusions from entities with vague boundaries is a special challenge to ecologists and provides us with a quest of broad significance.

I. THE DEFINITION AND PROPERTIES OF GUILDS

Guilds are groups of species that exploit the same class of resources in a similar way (Root, 1967). This simple definition has several implications that are best explained by way of an example. Consider the birds that probe for insect prey on the bark of tree trunks (Figs. 1 and 2). In many parts of North America this guild is composed of small woodpeckers (family Picidae), nuthatches (family Sittidae), and treecreepers (family Certhiidae). The guild also contains the black-and-white warbler (family Parulidae), whose relatives are primarily adapted for glean ing insects from foliage. Guilds thus contain species with very different phylogenetic histories and, as a consequence, with different propensities for dealing with the special requirements associated with exploiting a particular resource. Thus, the woodpeckers and treecreepers have evolved stiffened tails to act as props in climbing tree trunks, whereas the nuthatches and warbler have evolved modifications of the foot to assist in moving on vertical surfaces.

In addition to the core members of the guild which take the bulk of their food by foraging on trunks and limbs, there are many species, such as titmice (family Paridae), that occasionally feed by probing bark. The presence of these infrequent bark probes illustrates that guilds can have a hierarchical structure with members that range from strict specialists to occasional opportunists. As a consequence, the size and makeup of the guild will depend on where we draw the line with respect to the proportion of a species' diet that is obtained while probing bark.

The hierarchical structure of guilds is also reflected in how broadly we interpret the “in a similar way” portion of the definition. For instance, there are ants that forage for insect prey on tree trunks. These insects have the potential of influencing the food supply available to the bark probes. Their foraging style, however, is fundamentally different: Ants rely on their small size to enter bark crevices, whereas the birds employ their relatively large mass to lever out embedded prey and to flake bark. In this example, it can be seen that the scope of a guild's delimitation needs to be matched to the question being asked. The broader definition, which includes ants, might be most useful if our primary concern is competition, and the narrower concept, which excludes ants, might be more appropriate if our primary interest is in the evolution of adaptations. Furthermore, the birds eat the ants, which introduces further complications if the ants are lumped into the same guild with the birds.

During the northern winter, the black-and-white warbler migrates to the American tropics, where it joins another representation of the bark-probing guild which includes the tropical woodcreepers (family Dendrocolaptidae) and barbtails (family Furnariidae) shown in

![FIGURE 1 A few members of the bark-probing guild in North America.](image_url)
GUILDS

Fig. 2. The warbler's experience illustrates that guild membership can vary in space and season.

The opportunities provided by the existence of bark-dwelling arthropods, many of which are in a cryptic resting stage, has been exploited by bark probers that evolved from several phylogenetic lineages throughout the world. Thus, the Australian manifestation of the bark-probing guild contains Australasian treecreepers (family Climacteridae) and sittelas (recently placed in their own family, the Neosittidae). On Madagascar, the bark-probing role is filled by the nuthatch-vanga (family Vangidae), and on the Galapagos Islands the woodpecker finch (family Fringillidae) uses cactus spines to probe for prey under bark.

In classifying the ways that organisms exploit resources, guilds hold a rank that is similar to genera in phylogenetic schemes. One may also think of a guild as a group of species that occupy similar niches. Indeed, one of the original motivations for introducing the guild concept was to clarify confusion regarding the niche concept, which was used to describe three quite different entities: (i) the Hutchinsonian niche, which refers to the set of conditions that are sufficient for a species to exist in a particular habitat; (ii) the Grinnellian niche, which refers to the requirements and behaviors that are expressed wherever a species normally occurs; and (iii) the Eltonian niche, which refers to the role or occupation of a species in a community. The Eltonian niche was usually viewed as a relatively broad category (e.g., sap-feeding insects and predators of small mammals) which could contain several species and occur in a variety of habitats. In creating the guild concept, we provide an alternative category which clarifies some of the ambiguities associated with the Eltonian niche. For instance, by substituting guild for the Eltonian niche, we can avoid the contradictions that occur when several similar species are said to occupy the same "niche," a category that is supposed to be a property of individual species according to the Hutchinsonian and Grinnellian concepts. Accordingly, species exploit "nooks"—the adaptive space that is presented by resources with similar characteristics.

There are many types of guilds in addition to those that are based on food resources. For instance, animals can be grouped according to their use of tree cavities for nest sites and plants can be grouped according to their shared use of agents for pollination (e.g., moths or hummingbirds) and seed dispersal (e.g., fruit-eating birds or wind). In the case of plant-eating insects, there has been a tendency to define "mixed" guilds that group species according to a combination of overlapping functions. Frequently, the "feeding guild," defined on the basis of the insects' mode of feeding and the plant tissues they ingest, is combined with the "sheltering guild," defined on the basis of the insects' use of different plant structures as lairs for protection from enemies and adverse conditions. As a result, we often find that plant-chewing insects are divided into categories such as "leaf-miners," "leaf-rollers," "shoot-borers," and "stem galls" that have been shaped by the interaction of nutritional and protective functions. In searching for patterns, it is important to take all these mixed functions into full account.
It should also be kept in mind that species can belong to more than one guild. For instance, the woodpeckers mentioned previously are members of both the feeding guild that probes bark and the nesting guild that utilizes tree cavities. Many insects shift from one guild to another during the course of development because larvae and adults are adapted to perform such different functions. Thus, most butterfly species have caterpillars that are in the leaf-chewing guild and adults that are in the nectar-drinking guild.

II. THE UTILITY OF GUILDS IN THE SEARCH FOR ORGANIZATION AND PATTERNS

A. Guilds Provide a Natural Unit of Manageable Size for Comparative and Evolutionary Studies of Species Interactions

Even fairly simple communities, such as old fields in the temperate zone, can harbor more than 1500 species of insects belonging to more than 175 families. Such diversity obscures our view of fundamental processes, such as competition, mutualism, and evolution, because most species interact only rarely. This requires that we sort out the insignificant interactions to appreciate the true intensity of the critical interactions. For instance, most species of plant-eating insects in an old field rarely come into direct contact with one another because they are adapted for feeding on different plant species. As a consequence, the use of standard sampling methods, such as sweeping vegetation with a net or collecting the moths that are attracted to lights, which combine the insects on different plant species, would blur our understanding of competitive interactions between plant feeders. Furthermore, the existence of many types of species presents technical difficulties. Thus, to measure the density of all the plant feeders in a community might involve censusing deer, rabbits, mice, sparrows, snails, grasshoppers, caterpillars, aphids, galls, and tiny mites. Clearly, each of these categories of plant feeders require somewhat different methods based on different assumptions, relating to different spatial scales, and having different biases. To avoid these difficulties, ecologists often restrict their attention to some portion of the community. The guild concept is ideally suited to this purpose.

In defining a guild, we create a community detach-
traits are associated with each other in different regions, and (iii) the degree to which convergence is limited by deep ancestral traits in the various phylogenetic lines that make up the guild.

B. Guilds Partition Biotas into Entities Appropriate for Detecting Community Organization

Ecologists have long debated the significance of exploitative competition in organizing communities. An important line of inquiry in this debate concerns the constancy of a community's functional relationships. The reasoning is as follows: If communities are competitively organized, we expect to find that important functional relationships are predictable in space and time because the waxing and waning of populations of one species would be countered by compensatory changes in the populations of other species that use the same resource. Furthermore, these compensatory responses are expected to be most intense within groups of species that have similar methods of using the shared resources. Finally, we might expect to find that similar types of communities have similar functional profiles because the imperatives associated with using similar resources have sorted and shaped species into convergent groups. Guilds provide a means of partitioning a community into functional entities in which these compensations might be detected.

Much of the search for community organization has focused on plant-insect associations. One reason for doing this is that the basic resources (the biomass of foliage, stems, roots, flowers, and seeds produced by a particular plant taxon) are well circumscribed and relatively easy to measure. The constancy of the association's functional profile can be judged by changes in the guild spectrum—the proportion of total herbivore biomass that is engaged in exploiting different plant parts in various ways. A matrix for sorting insect herbivores into guilds is presented in Table 1. In all likelihood, such a scheme will need to be revised to suit the needs of a particular investigation.

Evidence for community organization based on guilds has been elusive. For instance, Cornell and Kahn found extensive variation in the guild spectra associated with 28 tree species in Great Britain and Root and Cappuccino found no evidence for compensatory changes in the densities of species within guilds that feed on goldenrods. These failures draw us to question our assumption that the intensity of competition is likely to be most severe within guilds. Different types of resources are linked in nature. Thus, in plant-insect associations, the consumption of leaves will reduce the future availability of seeds and the sapping of juices from stems will reduce the nutritional quality of foliage eaten by leaf chewers. Similarly, the caterpillar eaten by a foliage-gleaning bird is unavailable as a moth to bats that forage in the air. Linked resources link guilds and open the possibility for exploitative competition to play out its effects across a diffuse network of species. In other words, the use of a common resource seems to be of primary importance in competitive interactions; the manner in which the resource is exploited is of little consequence unless differences in a guild's behavior permit members to use an exclusive subset of resources. As a consequence, there are many cases of competition between guilds. For example, Schluter

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Manner of feeding} & \text{Buds and leaves} & \text{Plant stems} & \text{Flowers} & \text{Pollen} & \text{Fruits} & \text{Seeds} & \text{Wood and bark} & \text{Crusts of algae, molds, etc.} \\
\hline
\text{Chewers} & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\
\text{Exposed} & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\
\text{Concealed} & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\
\text{Sap-tappers} & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\
\text{Gall-makers} & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\
\text{Grazers} & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\
\hline
\end{array}
\]
found that Darwin's finches in the Galapagos Archipelago consumed more nectar on islands on which nectar-feeding carpenter bees were absent.

Of course, competition could be operating in a diffuse fashion that would be difficult to detect by only studying variations in guild spectra. In plant-insect associations, however, there is increasing evidence from long-term experiments and extensive surveys that plants in natural communities are rarely devastated by herbivores and that competition between plant-eating insects is a sporadic occurrence. These results, considered in conjunction with the highly variable guild spectra that have been reported by Cornell, Root, and Strong, suggest that most plants could support more herbivore species (i.e., plants provide Hutchinsonian niches that have not been filled) and that the structure of these associations is idiosyncratic, largely determined by the characteristics of a few dominant species that happen, for a variety of historical reasons, to have evolved to become members of the community. Many ecologists have the impression that guilds and communities that are dominated by predators or vertebrates are more likely to be organized by competition and to exhibit more predictable functional profiles. This impression, however, has not been adequately evaluated.

C. Guilds Provide a Framework for Describing and Comparing the Trophic Structure of Ecosystems

The trophic structure of ecosystems is often described in terms of broadly defined levels consisting of primary producers, primary and secondary consumers, decomposers, and so on. Food webs, on the other hand, are described in terms of individual species. Various attempts have been made to develop a scheme based on guilds that can be used to describe trophic structure with an intermediate degree of refinement; thus, herbivores might be subdivided into browsers of woody plants, grazers of grasses, sap-tappers on succulent foliage, borers in stems, and so on. To accomplish this requires that the entire community be partitioned so that all its members can be placed into a standard set of feeding guilds that have a widespread occurrence in nature. When ordered in this way, it is hoped that insights about energy flows can be gained by drawing comparisons between systems. In practice, however, such schemes have been little used, probably because it is almost impossible to develop a "key" to the guilds of the world that is flexible enough to accommodate a wide array of taxa and yet sufficiently circumscribed that different ecologists will assign species to guilds in the same manner.

III. THE INDIVIDUALISTIC NATURE OF SPECIES AND THE VAGUE BOUNDARIES OF ECOLOGICAL CATEGORIES

The process of defining guilds raises some fundamental questions. Are guilds distinct entities that reflect the outcome of natural processes or are they merely artificial groupings that we invent to divide diversity into more comprehensible units? In other words, are the opportunities provided by the environment discontinuous so that species are sorted into clearly defined groups on the basis of their requirements and lifestyles? Furthermore, are similar openings (or nooks) available in different regions? Are organisms constrained from filling these opportunities by their evolutionary histories? We can begin to address these difficult questions by comparing the characteristics of the core and marginal members of various guilds.

If we census the habits of all the species that utilize a particular resource, we usually find that there is a wide variation in the level of their dependency on the resource. The most dependent species—the core users—usually possess obvious specializations. Thus, in the bark-probing guild (Fig. 1), the treecreepers, woodpeckers, and nuthatches, which take the bulk of their diet from bark, have specialized feet and tails for moving on vertical surfaces (Richardson, 1942). In our census, however, we usually encounter a host of species that use the resource only rarely. These marginal users are either generalists or they are specialized along other lines that compromise their ability to use the resource in question. For instance, bark-probing titmice often forage on horizontal surfaces, such as the tops of limbs and branches, where they take much of their prey by hammering apart acorns, galls, and similar objects that they hold against the perch with their foot. These habits are reflected in the distinctive foraging maneuvers that titmice display during their infrequent forays onto tree trunks. Other species, such as gnatchasers, vireos, and wood warblers, which are primarily adapted for glean- ing insects from tree foliage, make clumsy efforts to take insects from bark on rare occasions.

Here we confront the individualistic nature of spe-
cies—a fundamental issue, originally raised by Gleason, that complicates all of our efforts to define functional groups of species. Thus, the various taxa that can utilize a particular resource have evolved “individualistically” along independent paths. As a result of their separate histories, these taxa have different constraints and proclivities, which result in different levels and styles of specialization. As a consequence, guilds often have ambiguous boundaries consisting of several species of generalists and inept opportunists that are specialists on other resources. Since the degrees of dependency grade into one another, the line one draws to define the membership in a guild can be somewhat arbitrary.

The indefinite boundaries of guilds are merely one expression of a more general issue confronting ecologists. Community ecology and biogeography are, by their very nature, concerned with levels of organization that consist of multiple species. In their efforts to discover the processes operating at these levels, workers have developed several systems for classifying species into groups on the basis of similarities in their (i) response to physical or site conditions to define communities; (ii) possession of particular adaptive traits to define life-forms and adaptive syndromes; (iii) response to seasonal cues to define phenological aspects; (iv) geographic distributions to define biogeographic provinces, biomes, life zones, and plant formations; and (v) diet to define trophic levels. Ecologists recognize and discuss these entities because they find that several species fall into clusters on the basis of these various types of classification. As with guilds, however, we also encounter species with intermediate characteristics that tend to blur the boundaries of these various categories. This is because all the traits that underlie these classifications are subject to the same individualistic evolution that can produce the marginal members of a guild. Thus, vague boundaries are an inherent property of any species assemblage—a fact that requires ecologists to cultivate certain habits of mind in forming arguments.

IV. ATTITUDES CONCERNING VAGUELY BOUNDED CATEGORIES

People are generally uneasy about categories, such as guilds, that are subjectively defined because the criteria one chooses to define the operational boundaries will have an impact on our ability to observe patterns. Thus, if a guild is defined too broadly, the characteristics of a miscellaneous collection of marginal species could obscure interesting similarities between the core members. On the other hand, if a guild is defined too narrowly, we could overlook the full range of influences that stem from extracting a particular resource in similar ways. There have been a variety of attempts to address this problem by using quantitative procedures, such as cluster analysis and principal components analysis, to delimit the membership in guilds. The raw material for these analyses comes from “activity censuses” which measure the relative frequency that species utilize different resources in different ways. The groupings that are sorted out by these procedures, however, often seem abstract and artificial. Furthermore, they are probably no less arbitrary than those that are more subjectively defined because the person who is running the analysis must decide how broadly to cluster the species’ activities.

Despite these difficulties, the continued and widespread use of subjectively defined categories attests to the need of such concepts in ecology. The best results are obtained when we follow some simple guidelines. In the case of guilds, the classifications that provide the most insights are based on a thorough knowledge of the species’ natural history. Open-minded observations and journal keeping are good starting points to form an intuition for distinguishing resources with properties that require special means for their efficient exploitation. For instance, insects on bark constitute an appropriate resource for a guild of bird-sized predators because the core members must move on vertical surfaces and extract prey from narrow crevices or under bark—maneuvers that are difficult for birds that lack the necessary specializations.

After a guild has been tentatively defined, it helps to conduct activity censuses at different sites and in different seasons to discipline intuition and provide a quantitative basis for determining guild membership. (For sedentary organisms such as plants, one would count the incidence of the traits that define the guild along transects laid out in appropriate habitats.) When the investigators are satisfied that a guild “makes sense” with respect to the questions that are being asked, the criteria that will be used to delimit the boundaries should be clearly described and fully justified. At this point, it may be useful to assign classes of membership; for instance, core species might be those that engage in the activities that define the guild on at least 50% of occasions, and marginal or accidental species might be those that engage in such activities on less than 10% of occasions. Many of these steps seem obvious, but
they are often left out of guild classifications, especially in cases in which the investigators are attempting to partition entire communities in such a way that each species can be assigned to a single guild. Such classifications are often used to compare the functional organization of communities in which only the most general of categories can accommodate the great diversity of species that must be placed.

Part of the art of becoming an ecologist involves developing a set of attitudes for coping with the complications that stem from the individualistic nature of species. One of the most important of these is an ability to match the question one is asking with the most appropriate grouping of species—the set that will reveal valid patterns that act in nature. Thus, the guild definition that is most useful for exploring convergent evolution may be quite different from the one that is best for observing compensatory shifts in the densities of interspecific competitors. In addition, we need to be on steady guard against the natural tendency to drift into thinking that the entities we have invented are “real.” Guilds are not a fixed feature of nature. They are a convenience that we can use to cope with diversity, reveal patterns, and facilitate discussion. As a consequence, guilds can be modified as long as they provide a valid base for addressing a question and their limits are explicitly defined and justified.

See Also the Following Articles

COMPETITION, INTERSPECIFIC • FUNCTIONAL GROUPS • HABITAT AND NICHE, CONCEPT OF • TROPHIC LEVELS

Bibliography

HABITAT AND NICHE, CONCEPT OF

Kenneth Petren
University of Cincinnati

I. Contrasting the Habitat with the Niche
II. History of the Habitat and the Niche
III. Comparing Habitats and Niches
IV. Evolution of the Niche
V. The Habitat and Biodiversity
VI. Conclusions and New Frontiers

GLOSSARY

community An ecological term referring to a set of species that occur in the same location that have the potential to affect each other either directly or indirectly.

community structure The web of potential biological interactions among members of a community that may be characterized in terms of diversity, complexity, hierarchy, and stability.

ecotone A zone of transition between two different habitats that may contain a community of organisms distinct from either habitat.

habitat structure Analogous to community structure, but limited to the physical structural aspects of a habitat. The structure of habitats may be characterized by such measures as complexity, heterogeneity, regularity, stratification, and fractal dimensionality.

microhabitat Locations within a habitat where organisms may carry out important aspects of their lives, such as places for harvesting food, nesting, or taking shelter.

niche overlap The proportion of available resources that are shared by two species. Usually used in the context of a single resource that limits population growth.

WHY ARE SPECIES COMMON in some places but rare or absent in other places? Early attempts to answer this fundamental question led to the concept of the habitat. A habitat is simply where an organism can be found in nature. Habitats are described in terms of geography, geology, climate, as well as by other species commonly found within the same habitat. The habitat concept is very closely related to another concept used to characterize species, the niche. The niche of an organism can best be described as its role in the community of organisms around it. Among the many traits that may comprise a niche are the physical and climatic characteristics that an organism can tolerate, and the resources required for the species to persist, such as food and shelter. The habitat of a species is often treated as a subset of its niche, and in practice it has grown difficult to draw a distinct line between the two concepts, yet the study of habitats has a history distinct from that of the niche. Today, habitat descriptions are commonly used as a practical guide for locating and maintaining species, whereas the niche is a more abstract concept that forms the conceptual foundation for much ecological and evolutionary theory.
I. CONTRASTING THE HABITAT WITH THE NICHE

As is the case with many ecological concepts, defining habitats and niches concisely and unambiguously is difficult. Habitat definitions are particularly prone to problems of scale. At the larger end of the scale is the biogeographical term **biome**. Some biomes may be considered the habitat of larger, well-traveled species such as large birds of prey, but usually the habitat is defined more narrowly. On the smaller end of the scale, the term microhabitat is used to describe the places where an organism spends part of its time. For instance, a fish may forage in the microhabitat that occurs near the banks of a river. Thus the concept of the habitat lies somewhere in between the biome and the microhabitat.

An ecological community is a suite of species that occur in the same location that at least have the potential to interact and affect each another either directly or indirectly. These effects are usually measured in terms of population growth or changes in the density of individuals. Different communities of organisms will generally reside in different habitats. If most species in one community have little chance of affecting the populations of other species in a different community, then these communities probably occur in different habitats. In this way, microhabitats refer to places within a community, and biomes encompass a number of different habitat types. Definition of habitat and community boundaries will depend on the exact species under consideration. For instance, it may be reasonable to distinguish the small mammal communities of grassland from nearby riparian forest, though an individual jaguar may prey on species in both habitats.

The primary habitat preferences of a species can be displayed graphically. Ordination is the process of characterizing a species with regard to habitat gradients. Gradients can be displayed as axes on a graph. Usually, some measure of density or population growth is plotted, and the portion of the gradient that is occupied by the species is referred to as the habitat breadth of the species. Figure 1 shows how two species of plankton can be compared according to a single habitat dimension, depth in the water column. There are potentially limitless gradients or factors that may be involved in characterizing the habitat of a species. Graphical representation of more than three variables in a single graph is difficult, but the concept easily extends into multiple dimensions.

The niche is a more abstract term than the habitat. It encompasses all possible interactions that a species has with the environment and other species in the community. Conceptually, the niche is richer than the habitat and forms the foundation for much ecological theory. Because of the abstract nature of the concept, the definition is perhaps even more difficult to formulate for all species. There are striking similarities in the way niches and habitats are displayed. Like a habitat, a niche is often defined by axes, and there are multiple axes that can be compared.

The most common type of axes used to characterize a niche include those that define environmental tolerances and resource requirements. Humidity, sunlight, temperature, wind exposure, and pH are examples of environmental axes that help to define a niche. Resource axes may include food resources (insects, seeds, bacteria, nutrients), space requirements (breeding sites, ref-

![Figure 1](https://via.placeholder.com/150)
Plants and pathogens may even be limited to a single species successional stage, while specialized insects, parasites, trees, birds, and lichens are limited to only a certain forest, each of which may be considered a different community. Habitat descriptions include geographical information and often refer to the dominant species present (e.g., grassland, coral reef). Niche descriptions for a species usually begin with the main food items consumed. These differences make broad comparisons of niches and habitats difficult, but at any single scale, these concepts can be applied to facilitate meaningful comparisons among similar species.

II. HISTORY OF THE HABITAT AND THE NICHE

A. Niche Origins

It is tempting to think of a niche as a physical place. This is the common usage of the word, and there are examples in the early ecological literature that use the term "niche" in the purely physical sense. However, the origins of the ecological niche reside in the more general observation that no two species are exactly alike.

Many naturalists in the latter part of the 19th century turned their attention toward documenting the traits that distinguish one species from another. When two species appeared very similar, it was thought that eventually differences could be found that would distinguish the unique role of each in the community. This idea was evident even in the writings of Darwin, and over time it has evolved into a very general principle: the principle of competitive exclusion. Exploring the implications of this relatively simple notion has dominated the study of ecology for much of the 20th century.

The principle of competitive exclusion and the concept of the niche developed in parallel through the early 1900s. Theoretical work by Volterra (1926) showed that if certain assumptions are met, only one species should be able to survive on a single resource. Gause (1934) demonstrated this principle experimentally with two species of Paramecium feeding on a common resource. Two species that consume the same resource in the same way simply cannot coexist.

The term niche was first used in the context of competitive exclusion by Grinnell, a superb naturalist of the North American west, as early as 1914. Grinnell's study of the California thrasher represents one of the first that specifically set out to characterize the ecological niche of a species, though he laid out the concepts of the niche and competitive exclusion 10 years earlier in a study on chickadees.

Grinnell's study of the California thrasher illustrates some of the most basic components of a niche (a) type of food consumed (mostly insects, berries at some times of the year), (b) microhabitat preference (beneath shrubby vegetation), (c) physical traits and behaviors used in gathering food (a long beak thrashed through the top layers of soil and leaves), and (d) resources available in gather food.
HABITAT AND NICHE, CONCEPT OF

<table>
<thead>
<tr>
<th>Species</th>
<th>Habitat/Biome</th>
<th>Niche</th>
</tr>
</thead>
<tbody>
<tr>
<td>California thrasher</td>
<td>Coastal chaparral, western North America</td>
<td>Generalist insectivore, territorial; sometimes a frugivore</td>
</tr>
<tr>
<td>Toxostoma redivivium</td>
<td>Coastal chaparral, western North America</td>
<td>Generalist insectivore, short pursuit predator, territorial</td>
</tr>
<tr>
<td>Cuban crown giant lizard</td>
<td>Caribbean (Cuba)</td>
<td>Generalist insectivore, short pursuit predator, territorial</td>
</tr>
<tr>
<td>Spotted salamander</td>
<td>Ephemeral ponds in deciduous forest; temperate North America</td>
<td>Generalist carnivore, breeds in ponds with few large fish</td>
</tr>
<tr>
<td>Jaguar</td>
<td>Tropical forest of all types, savanna, Central and South America</td>
<td>Small animal carnivore, solitary stalking predator, keystone predator</td>
</tr>
<tr>
<td>Plains zebra</td>
<td>Tropical savanna, Africa</td>
<td>Greater of short grass; gregarious; migratory</td>
</tr>
<tr>
<td>Harvester ant</td>
<td>Sandy regions of desert, southwest North America</td>
<td>Scavenger of small seeds, makes terrestrial, colonial burrows</td>
</tr>
<tr>
<td>Pogonomyscus ragesas</td>
<td>America</td>
<td>Scavenger of small seeds, makes terrestrial, colonial burrows</td>
</tr>
<tr>
<td>Tawny mining bee</td>
<td>Temperate fields, gardens; Europe</td>
<td>Nectar/pollen harvester; solitary soil burrower, pollinator</td>
</tr>
<tr>
<td>Andrena fulva</td>
<td>Caves, moist environments; North America</td>
<td>Scavenger, detritivore</td>
</tr>
<tr>
<td>Camel cricket</td>
<td>Sandy marine beaches; temperate and tropical Neartic</td>
<td>Scavenger, territorial; nocturnal</td>
</tr>
<tr>
<td>Octopus quadrata</td>
<td>Mediterranean, rocky marine shores; western temperate North America</td>
<td>Filter feeds on Plankton, sesile, pelagic larvae</td>
</tr>
<tr>
<td>Mytilus californianus</td>
<td>Russia, rocky shores, western temperate North America</td>
<td>Plant and moss cell predator, meiofaunal (0.4–1.0 mm)</td>
</tr>
<tr>
<td>Echiniconus spp.</td>
<td>Intertidal, rocky shores; coastal temperate North America</td>
<td>Plant and moss cell predator, meiofaunal (0.4–1.0 mm)</td>
</tr>
<tr>
<td>Guanacaste tree</td>
<td>Tropical savanna, dry forest; Central America</td>
<td>Large canopy tree; pioneer, keystone</td>
</tr>
<tr>
<td>Eriogonum fasciculare</td>
<td>Temperate deciduous forest, North America</td>
<td>Shade-tolerant, understorey, pollen and nectar producer</td>
</tr>
<tr>
<td>Phacelia tanacetifolia</td>
<td>Mediterranean, rocky shores, western temperate North America</td>
<td>Plant and moss cell predator, meiofaunal (0.4–1.0 mm)</td>
</tr>
<tr>
<td>Drasematomya fungi</td>
<td>Moist soil, pools, cosmopolitan</td>
<td>Seed-and-wait predator of roundworms, hoop-mate structures</td>
</tr>
<tr>
<td>Archiblastys spp.</td>
<td>Moist soil, pools, cosmopolitan</td>
<td>Seed-and-wait predator of roundworms, hoop-mate structures</td>
</tr>
<tr>
<td>Slime mold</td>
<td>Moist soil, various habitats, cosmopolitan</td>
<td>Bacteria consumer, single and multinucleate stages</td>
</tr>
<tr>
<td>Acrasitae spp.</td>
<td>Moist soil, pools, cosmopolitan</td>
<td>Bacteria consumer, single and multinucleate stages</td>
</tr>
<tr>
<td>Schistosoma mansoni</td>
<td>Humans and freshwater</td>
<td>Blood parasite in humans, gut</td>
</tr>
<tr>
<td>Schistosoma mansoni</td>
<td>Snails, tropical, Africa, South America</td>
<td>Parasite in snails</td>
</tr>
</tbody>
</table>

required for shelter and breeding (dense shrubs for night roosting and nesting). These four basic factors allow one to characterize the basic niche of most animals, and most animals differ with respect to one or more of these factors.

An analogous set of core niche dimensions may be constructed for plants, which tend to partition niche space along resource axes such as available light, soil moisture, and various soil nutrient gradients. Important niche dimensions for marine organisms may include temperature, substrate type, salinity, pH, and exposure to waves. Thirteen years after Grinnell, the publication of Elton’s Animal Ecology text (1927) established the term niche in the lexicon of ecology. However it was not until the work of Hutchinson (1957) that the concept of the niche was fully developed as a cornerstone of ecological theory. Hutchinson’s niche is an n-dimensional hypervolume, implying that there are usually many factors, or dimensions, that define the role of a species in a given habitat. These included yet extended well beyond the basic niche dimensions described by Grinnell.

Other ecological concepts emerged from this founda-
The resources that a species would use if it were isolated from all potential competitors is part of its fundamental niche. The realized niche is a subset of the fundamental niche that includes the resources actually consumed by the species in nature. The difference between the realized and fundamental niche can be attributed to competition from other species in the community.

B. Habitat Origins

Originally, the habitat of a species was not viewed as a part of its niche. In the early history of these terms they were treated separately. The habitat was seen more as a guide to the kind of community in which the organism played its role. For instance, the California thrasher occupies coastal chaparral of western North America, but this habitat description can be viewed as distinct from its niche within that habitat. Also, species were viewed as possessing suites of traits and physical tolerances that enabled them to be suitable only for specific habitats.

Over time, habitat axes were compared right alongside common niche axes, and the distinction became obscured. This change may be due in part to parallel development in the field of evolution and adaptation and increased knowledge of the processes involved. Today it is not difficult to imagine that two species may have evolved to occupy different habitats, whereas at a previous time only one species may have occupied both habitats.

The history of the habitat concept can also be traced back to the work of Grinnell and his contemporaries. In one of the earlier studies, Grinnell and Storer characterized the animals of Yosemite according to the elevations at which they occurred. In this mountainous region, species tend to be found in narrow elevational zones or bands.

At about the same time as the Yosemite study, Ramsay wrote that each species was unique and possessed adaptations that enabled it to tolerate a unique set of environmental conditions. This "principle of species individuality" bears strong resemblance to the early influences on the niche concept, except that in the habitat context, competitive exclusion was not at the core. Instead, species were seen to be adapted to certain conditions and habitats, and they independently sorted themselves out accordingly.

Subsequent work has confirmed that habitat boundaries are often determined by physical tolerances. However, there are numerous examples where competition and other species interactions play an important role in determining the habitat breadth of a species in nature. This further illustrates how the habitat and the niche have converged over time. Therefore it is appropriate to treat the habitat and the associated gradients and axes as a subset of the niche.

III. COMPARING HABITATS AND NICHES

A. Multispecies Comparisons

The habitat and niche of any species can be characterized independently, but the true value of these concepts lies in comparing multiple species. The ability to display habitats and niches graphically greatly facilitates comparisons. Figure 2 shows how species can be ordered along a habitat gradient. Copepod species show different patterns of abundance with respect to the gradient from ocean to land in the intertidal region. While in this instance there is a large amount of overlap, each species has a characteristic distribution in this region. Habitat boundaries are rarely sharply defined. Ecotone is the term used to refer to this zone of transition between habitats. Because this zone is somewhat different from either neighboring habitat, an ecotone may contain several unique species. This concept was proposed to help explain distributions of species that did not appear to coincide with habitat boundaries. In a unique test of this idea, Terborgh and Weske (1975) found little evidence that subsets of bird species along an elevational gradient were responding to ecotones. Instead, it appeared that the elevational ranges of many species were independently defined by the presence of other closely related species. In this case, competition appeared to be more important than habitat gradients for setting the actual limits of species distributions.

The niche framework is typically used to compare groups of closely related species that consumed similar resources but differed in one or a few important ways. Suites of species in a community that utilize the same general resource were termed guilds by Root (1967). Just as a guild of blacksmiths all work with iron yet have different specialties, assemblages of animals that specialize on seeds, fruits, or insects all have different ways of harvesting these resources.

Early studies focused on guilds that differed mainly with respect to body size. Hutchinson (1959) observed that species within a guild often differed by a minimum size ratio. Known as Hutchinsonian ratios, these regular arrangements of species body sizes occur in a number...
of vastly different communities consuming very different resources.

One example of the kind of pattern Hutchinson was attempting to explain is the assemblage of rodents that consume seeds in the deserts of North America. Figure 3 shows how similar assemblages of rodents in different communities contain a similar set of rodents. Each species differs from others in the same community mainly in terms of body size. Convergence is a term applied to organisms or communities that are not closely related historically that have members with similar physical features because they occupy similar niches. The implication is that parallel evolutionary trajectories were followed independently by different species because of similar selection pressures.

The underlying mechanisms that create such regular patterns can be explained because each species consumes resources of a size class that is proportional to its body size. Smaller rodents consume smaller seeds, while larger rodents eat larger seeds. In a different community, smaller fruit pigeons consume smaller fruits, while larger species primarily consume larger fruit. For a medium sized bird, small fruits take too long to gather and handle, and the bird can not eat them efficiently enough to sustain itself. Larger fruits may be too big to handle or crack open.

Many types of food resources such as seeds, fruit, or arthropods have a wide, flat distribution along a size gradient. These relationships can be displayed graphically with a single niche axis, food resource size (Figure 4). One curve represents the available resources, in this case seeds, and the narrower curves underneath represent the sizes and amounts of seeds consumed by each species. This division of available resources among members of a community is termed resource partitioning.

Utilization curves are often bell shaped (normal, Gaussian) to reflect the tapering ability of each species to consume seeds away from the 'optimum' size. This type of niche representation allows direct comparison of the amount of overlap in resource consumption. If food is indeed the primary limiting resource, the overlap of species resource utilization curves directly represents the amount of competition that each species is experiencing. It is the regular spacing of curves to minimize overlap, or competition, that creates the pattern of regular size distributions of animals seen in natural communities. This simple framework forms the basis of niche theory.

The concepts of fundamental and realized niches can be clearly visualized within this framework of graphical representation of both habitat and niche dimensions. Mussels are capable of living in a wide range of elevations above the low tide mark, and they will move into these spaces in the absence of other species. However other animals, such as barnacles, are better adapted to occupy lower positions and can outcompete mussels in some zones. That is why in natural communities, the realized niche of mussels in terms of this habitat gradient is usually narrower than the fundamental niche.
A similar process may be operating in the granivorous rodent example presented earlier. Each rodent may be capable of consuming a wider range of seeds, albeit inefficiently. In nature, either there are fewer of these suboptimal resources available because of competitors, or each species deliberately chooses a narrower range of seeds to consume to avoid competition.

The effect of competition on realized niches can be seen by comparing island and mainland communities. Islands usually support fewer species than equivalent mainland habitats. It has been shown for many kinds of species that in these species-poor island communities, species tend to eat a wider array of foods and therefore have wider niches. This agrees with the expectation based on niche theory outlined earlier: there are fewer potential competitors on islands, and therefore there is less resource overlap with other species acting to constrain their niches.
Species that have very high overlap in many niche dimensions often have little or no overlap in one key dimension. This principle, called niche complementarity, lies at the heart of the guild concept. In previous examples, similar species consumed the same general resource and had high niche overlap, but they showed little overlap with respect to a single dimension: body size. Typical niche axes where complementary species can be found include size, light (nocturnal/diurnal), benthic/pelagic, and canopy/ground or coniferous/deciduous microhabitat.

In many instances niche differences among species are clear, but sometimes niche differences are extremely subtle and require careful study to be revealed. For instance, five species of *Dendroica* warblers can all be found in the very same tree in temperate zone coniferous forest habitat. All glean similar types of prey, primarily lepidopteran larvae, from the surfaces of the tree. Only careful study by MacArthur (1958) revealed that each species favors different parts of the tree, and each tends to move about the tree in different ways (Figure 5). Thus it is not only the type of food resource that is limiting the population growth of each species, but where and how that food resource was harvested. There are enough subtle differences in microhabitat use and behavior to allow coexistence of species that at first glance appear to share equally the same limiting resource.

Niches can change through time. For instance, the larval forms of many organisms often have completely different niches from adult forms. Thus, the term ontogenetic niche has been applied to amphibians, arthropods, and a wide variety of marine organisms. The great versatility of the niche concept allows incorporation of these temporal changes by simply adding a temporal axis to the species niche. The realized niches of species in a community can also change through time if the community undergoes change. Extinctions and invasions cause rapid community changes that in turn result in niche shifts among other community members.

Niche theory is the branch of ecology that has taken these informal graphical models and extended them mathematically to model the process of competition and community organization. The origins of niche theory can be traced back to the early work of Volterra, and has grown and undergone dramatic changes. Complex interactions among members of a community are very difficult to measure, and the approach of studying theoretical models has greatly enhanced our knowledge of the factors that influence community dynamics.

**B. Other Niche Dimensions**

The classical context of niche comparisons is mainly restricted to niche axes characterizing habitat and food
resources. Some ecologists have even considered the definition of a niche to be limited to the food resources consumed. However developments in the field of ecology through the 1900s have made it clear that an expanded definition of the niche that encompasses all the interactions of a species within a community is most appropriate. A broad definition of the niche includes all potential interactions that ultimately have the effect of changing the population density of another species in the community.

Mutualistic interactions are somewhat different from food resources in that they are often not literally consumed, nor are they occupied as in a nesting or sheltering place. However in niche descriptions, mutualistic interactions are treated much the same as other resources in that they may be limiting and therefore may even be competed for. For instance, pollinators interact mutually with angiosperms, yet from the point of view of a plant, they may be regarded as any other resource.

C. Indirect Interactions

Competition for resources has historically played a central role in niche descriptions. More recently, studies have uncovered processes where two or more species coexist because of factors that have little to do with resource niche axes. For example, predation can act to promote coexistence of potentially competing species in the following way: The presence of a predator may reduce the numbers of one prey species and thereby allow other species that have similar resource requirements to coexist. If two or more species are all harvested by a predator that does not show preferences for any specific prey type, then the most common prey will experience a disproportionate amount of predation. In this way, the presence of a predator can actually promote coexistence of species and an increase in biodiversity. Predators that have this effect have been referred to as keystone species because of their disproportional affect on community diversity.

An excellent example of this process can be found in the rocky intertidal communities of the eastern Pacific. Paine (1980) and colleagues conducted a unique set of experiments demonstrating that the presence of a starfish predator (Pisaster) enabled the coexistence of a number of species, whereas in the absence of the starfish, the community was dominated by the California mussel. The niche description for a barnacle in this community would therefore be incomplete without incorporating information about not only the presence of the California mussel, but also the relative vulnerability of each of these species to starfish predation. This implies that the niche characterization may change depending on the presence of potential predators and prey and their specific traits.

Predation certainly can be seen as a direct interaction, but the example above falls into the class of indirect interactions: A keystone predator can enhance the abundance of a species by interacting with a third species. Many interactions may cause ripple effects through a community in indirect ways. Elephants change the physical structure of their habitats, and many species excavate holes and nests that are used by other species. Parasites and pathogens can also have indirect effects on entire communities.

Viruses and other pathogens may initially appear to occupy a very simple niche, especially those that are confined to live entirely inside the body of a host species. How can such an organism play an important role in the community through indirect interactions? Pathogens can significantly impact the population size of the host species. This change in abundance may have consequences for other species, such as predators and prey of the infected species.

Some pathogens have complex life cycles that depend on more than one host species, and sometimes these hosts are very different. For instance, in using humans and snails at different stages of its life cycle, Schistosomiasis provides a clear link between humans and some snail species that are used as hosts for different stages of the life cycle.

Another interaction involving pathogens provides an example of apparent competition. One host species may be relatively unaffected by a pathogen, whereas other species may be severely impacted. In this case, the unaffected species can act as a carrier, spreading the pathogen to other species in the community that are more susceptible, causing their decline. On the surface, this pattern may resemble competition.

Community ecologists have used various methods to incorporate these indirect interactions into model representations of communities. It is certainly not as straightforward as simple resource utilization graphs. Mathematical matrices have been applied to study the dynamics of communities. In this format, all pairwise interactions can be explicitly considered, and broad concepts such as the relationship between community stability and complexity have been studied within this framework. Recently, web theory has been applied to this task because complex interactions within communities can be more explicitly constructed. All of these theoretical constructs of communities are based on the niche concept, thus the utility of the niche concept
has grown even as the field of ecology has undergone dramatic changes.

IV. EVOLUTION OF THE NICHE

Ecological studies can help us to understand why species occupy unique niches, but how did the great diversity of species arise in the first place? The processes underlying the evolution, adaptation, and coexistence of species that have unique niches is a fundamental aspect of understanding biodiversity.

For many organisms the primary mode of species formation, and therefore the primary generating force for biodiversity, is thought to involve differentiation of isolated populations. A terrestrial species may have populations separated by water, or coastal marine populations may be isolated by unsuitable open ocean. Over time, these populations undergo changes and begin to accumulate different characteristics. Some of these differences are due to chance events such as mutation and genetic drift. Traits change over time as the organisms adapt to the unique features of their respective habitats. Subtle environmental differences may result in the evolution of thicker shells in a population of crabs, or a preference for fish over mammals in a population of killer whales.

This differentiation by adaptation to different habitats with different food resources is in itself one engine for generating biodiversity, however it cannot explain the vast diversity of species that occur within a single habitat. Eventually, populations that have been isolated for a while may reestablish contact, and individuals will attempt to coexist within the same community. It is at this stage of speciation, referred to as secondary contact, that one can see how the concept of the niche is fundamentally related to species diversity.

When secondary contact occurs, one of three outcomes is possible. First, populations may interbreed, and speciation is not completed. Second, if populations do not interbreed, they may either coexist, or, third, one may outcompete the other. Which of these two latter outcomes is realized will depend on the amount of niche overlap between these species. If overlap is high, then one species is likely to outcompete the other. If overlap is low enough, the species may coexist.

In situations where coexistence is possible (or at least extinction is not very rapid), and the species involved still possess similar traits, we would expect evolutionary changes to occur over time. Individuals that consume limiting resources that are shared with another species will experience more competition and may survive or reproduce less than individuals that do not overlap as much. If there is natural variation in resource consumption among individuals in each population, and if the traits that cause this variation are genetically passed to offspring, then the species are expected to diverge over time to minimize niche overlap. This process is called ecological character displacement, and it is a logical extension of the competitive exclusion principle into evolutionary time. Figure 6 illustrates the process of character displacement. It is the change in a trait of a species, such as body size, that leads to a shift along a niche axis and reduction of overlap.

The empirical evidence for character displacement lies mostly in descriptive, snapshot-like comparisons of species. Typically, closely related communities are compared and character displacement is inferred if the presence or absence of one species is correlated with a change in the traits of another species. These studies infer that it is niche overlap in the use of a single limiting resource that has led to the observed species differences. Lack (1947) was one of the first to infer character displacement in this way in his study of Darwin’s finches Figure 7.

Static character differences constitute indirect evidence for competitive processes and have therefore received criticism. To document evolution directly in natural populations is a difficult task because even rapid

![Figure 6: A graphical representation of ecological character displacement. Initially upon secondary contact, species have highly overlapping resource utilization. Over time (b) resource overlap and competition is reduced through evolutionary changes that result in a shifting of the niche.](image-url)
HABITAT AND NICHE, CONCEPT OF

V. THE HABITAT AND BIODIVERSITY

We have traced the habitat concept from its independent origin through its incorporation into the concept of the niche. With growing awareness of the need to conserve biodiversity, the habitat concept has emerged again as an important tool for management. Essentially, habitats contain functioning communities of organisms. If the goal is to protect a focal species, then setting aside appropriate habitat will likely result in preserving the required niche elements of the species as well.

Assessing appropriate habitat has two great advantages over attempting to define niches. First, because of the physical nature of habitats, they are easier to recognize and quantify. Second, ecological communities tend to fall along similar boundaries as habitats, thus by preserving a habitat one is more likely to preserve communities with all niche interconnections intact. This relationship among species, communities, and habitats has become the center of the more recent conservation efforts termed habitat conservation plans.

For example, in San Diego county, a number of endangered or threatened species rely on a unique type of habitat: chaparral, or coastal sage scrub. A major goal of habitat conservation plans in this region is to create contiguous regions of appropriate habitat. In this way, the needs of many species can be met at once.

Habitat fragmentation is a major concern because small refuges of habitat are essentially islands. Islands are prone to a number of processes that degrade biological diversity. Small populations on islands are prone to extinction through random fluctuations in numbers. The geometry of islands gives them a high ratio of edge to interior. Because animals may wander across this boundary, those that rely on the interior island habitat experience more competition and predation from those outside the island. More isolated islands are also less likely to be recolonized in the event that a population has gone extinct. Therefore, natural reserves and habitat conservation plans pay particularly close attention to fragmentation and try to consolidate larger parcels of unbroken natural habitat.

Once a habitat has been severely degraded, there may be ways to return it to its former natural state. Habitat restoration is a challenging field, but one that has made great strides recently. Wetlands have been restored by controlling the water table and planting...
appropriate vegetation. Foresters have been restoring habitat for decades by replanting harvested areas. Highway construction projects aim to restore natural vegetation upon completion of a project by spraying a mix of grass, flower, and shrub seeds in a fertilizing matrix onto the bare ground. Another interesting example of habitat restoration can be seen with the construction of artificial reefs. By placing physical structures on the seafloor, a reef community can take hold and flourish. This is an example of how subtle environmental features such as physical structure can have a profound effect on habitat quality and species diversity.

Two general mechanisms act to promote species diversity within structurally diverse habitats. First, biological species diversity leads to more species diversity. A monoculture such as an orchard allows for relatively little niche differentiation of the organisms within. There is little opportunity for species to adapt to specific host species, and this applies to insects, birds, mammals, and parasites. Reduced numbers of species limits the number of possible niches in an orchard when compared to a tropical rain forest.

Artificial reefs belong to a second class of mechanisms that generate biodiversity by acting over and above the biological mechanisms. These are the purely physical features of the habitat that allow for increased niche differentiation. Dead logs, snags, distinct ground, and shrub layers will provide habitat heterogeneity that allow higher trophic levels to differentiate. Just as warblers partition the physical space within a single tree species, other competitors that use different tactics to forage on similar prey can partition a structurally diverse habitat to a finer scale. Behavior plays a key role as foraging tactics and escape strategies of prey are honed to specific microhabitats within the larger forest.

MacArthur and colleagues (1962) were among the first to correlate purely physical aspects of a habitat with species abundances. Since then, physical habitat structure has been shown to augment species diversity in systems as diverse as birds, lizards, stream invertebrates, fish in marine reefs, and microorganisms in heakers.

Keystone species have a disproportional effect on many other species in the environment, and as such are often the focus of conservation efforts. Starfish foraging in the intertidal zone have a positive effect on biodiversity. Jaguars prey on a number of small rodents and animals, and they tend to capture prey in proportion to their abundance. The net effect is to prevent any single species from becoming very common and outcompeting other prey species. The niche of the jaguar and the starfish includes this community-wide role, and they are sometimes referred to as keystone predators.

Management strategies have turned more attention toward keystone species in order to preserve communities. A similar keystone role is played by bird species such as the Guanacaste tree of Central America. In this instance the link between the niche, the habitat and biodiversity is clear. The Guanacaste tree is vital for establishing forest in grassland habitats, and it therefore plays a key role in restoring tropical dry forest. The niche of this species would include its role in developing habitat structure and therefore intersects with the niches of many other species in the community.

The habitat and the niche can interact synergistically with human activity to cause extinctions and loss of biodiversity. Niche complementarity has led to a very general pattern whereby species with similar niches are spatially segregated. Increased human movements on a global scale has had a dramatic effect in bringing these species into contact by introducing exotic species into new geographic regions. If these alien species have high niche overlap with resident species, the resident species can be driven to extinction by a successful invader. Even though there are many cases where introduced species can not establish or outcompete residents, many threatened and endangered species are at risk because of invaders. This threat also extends to introduced predators. Residents species that have adapted to a niche in one community can be driven to extinction by a predator that has evolved in a different community.

VI. CONCLUSIONS AND NEW FRONTIERS

The concepts of the habitat and the niche have grown mainly through study of the organisms most familiar to us as humans. Most conceptual advances were derived in some way from the study of larger animals and plants. Even within these groups, there has been a disproportionate amount of study of birds. Yet there have been significant contributions from other sectors of the biological world. For instance, Gause's study of Paramecium lies at the heart of the niche concept. It is fitting then that more attention is returning toward understanding the niches of microorganisms.

Experimental evolution using bacteria is a growing field that holds great promise. The short generation times and relatively simple ecologies of bacteria enable evolutionary experiments on the scale of the commu-
HABITAT AND NICHE, CONCEPT OF  315

diversity. This is one of the few instances where the interaction of ecological and evolutionary factors can be studied during the process of community formation.

There is an enormous number of microscopic, planktonic and meiofaunal (0.4–1 mm) organisms about which we know very little. These organisms are ubiquitous, species-rich, and show an amazing amount of diversity in form. However there are significant obstacles encountered in studying these smaller organisms that even make it difficult to quantify the habitats in which they can be found. Practical problems arise while trying to observe them in nature, while on the other hand, many fail to survive the transfer to laboratory environments. Yet there are compelling reasons to learn more about their ecology.

The niche of a terrestrial vertebrate or marine invertebrate is often defined largely by how it feeds or defends itself. In turn, anatomical structures have evolved that reflect these niche differences. Similar examples can be found in the nearly invisible world of small creatures around us. For instance, ciliates and other protozoans possess anatomical structures that suggest a diversity of roles even greater than those observed in larger organisms, yet there has been comparatively little study of their niches in nature.

Some organisms stretch the concepts of the niche and habitat to their limit. Some plants display an extreme amount of what is referred to as phenotypic plasticity: different individuals develop extremely different body forms depending on the environmental circumstances. Some ciliates have the ability to radically change the form of feeding or defensive appendages even within the lifetime of an individual. Many bacteria, and even some larger meiofauna like the bearlike tardigrades, can enter a more or less state to withstand extreme environmental conditions. They may persist for long periods of time in this state and may be very difficult to detect. Much like the seed bank of desert plant communities, these alternative states pose difficult problems in understanding the dynamics of their communities.

The habitat and the niche have grown, developed, and been applied with great success. Observation and theory have played important roles throughout, and experimental investigation has undergone a resurgence. Yet it is intriguing to ponder that most of this development has included only a subset of the biological world. Only time will tell whether the concepts of the habitat and niche will persist as the complex interactions of these diverse communities of small creatures are revealed.

See Also the Following Articles

CONSERVATION BIOLOGY, DISCIPLINE OF • DIFFERENTIATION • DIVERSITY, COMMUNITY/REGIONAL LEVEL • ECOLOGY, CONCEPTS AND THEORIES IN • GUILD • SPECIES DIVERSITY, OVERVIEW

Bibliography


HEMIPARASITISM

David Smith,* Todd J. Barkman,† and Claude W. dePamphilis‡
*University of Maryland and †Penn State University

I. The Natural History of Parasitic Plants
II. The Parasitic Plant Clades and Their Relationships to Other Plants
III. Evolution of the Degree of Parasitism
IV. Parasitism and Other Phenomena in Plants
V. Seed Dispersal, Germination, and Prehaustorial Events
VI. The Haustorium
VII. Physiology of the Parasite—Host Interaction
VIII. The Ecological Advantages of Parasitism
IX. Host Specificity
X. Parasitic Plants as Pests
XI. Conclusions

GLOSSARY

dodder Viney parasites of the genus Cuscuta that form haustoria on the stems of their hosts. The unrelated parasitic vine, Cassytha, is often referred to as Laurel Dodder.
edaphyte The portion of a parasitic plant that is embedded inside host tissue.
edaphytic parasite Parasites with vegetative bodies that are entirely endophytic and unobservable unless flowering.
haustorium (pl. haustoria) The organ found in all parasitic plants that penetrates vascular tissue of the host plant and forms a functional bridge for uptake of water and nutrients from the host.
hemiparasite A parasite that photosynthesizes.
holoparasite A parasitic plant that lacks chlorophyll and cannot photosynthesize.
mistletoe Parasites of the closely related families Loranthaceae or Viscaceae with haustoria that originate as root tissue and penetrate stems or roots (in Loranthaceae) of the host plants.
mycotrophic (=mycoparasitic) plant A plant that obtains energy through mycorrhizal associations with saprophytic or other biotrophic fungi.
root parasite Parasites that form haustoria on roots of the host plant.
stem parasite Parasitic plants that form haustoria on stems of the host.

AROUND 3000 SPECIES OF plants are considered parasites because they obtain water, mineral nutrients, sugars, and sometimes other materials from another plant. Parasitism in plants has much in common with other plant life-history strategies, but it is distinguished by having a direct, physiological connection called a haustorium. Modern techniques for phylogenetic reconstruction have demonstrated that parasitism has originated a dozen times within the flowering plants, including mistletoes, dodders, and many agriculturally important species. The ecology of the parasitic plants and the interaction with the host are briefly described.
I. THE NATURAL HISTORY OF PARASITIC PLANTS

Parasitic plants are strictly defined by the production of specialized feeding structures, known as haustoria, that form a functional bridge into their hosts. Through the haustorium, parasitic plants obtain water, mineral nutrients, sugars, and other host metabolic products. Approximately 3000 angiosperm species (1% of named species) are haustorial parasites; however, this total does not include any plants that feed heterotrophically on fungi (mycotrophic plants). Most parasitic plant species are hemiparasites; they photosynthesize and are only partially heterotrophic. The degree of heterotrophy in parasitic plants varies between autotrophy and the complete loss of photosynthesis, a condition known as holoparasitism.

Other than the presence of haustoria, there are few generalizations that can be made about parasitic plants because they display enormous variability in growth form and host dependence. In most species, haustoria are formed from root tissue of the parasite, but in a few species they are initiated from stem tissue. The haustorium, in turn, can penetrate the root or stem of the host plant. The major life-histories among parasitic plants are distinguished by these differences and they have important ecological implications.

Root parasites are extremely variable in growth form, occurring as trees, shrubs, and herbaceous perennials or annuals. Because their haustoria are often below ground, many root parasites do not demonstrate any obvious indications of parasitism. Experimental studies have compared hemiparasitic plants grown in pots with and without hosts. They found enormous variability among species in their ability to set seed in the absence of a host. The degree of parasitism is extremely variable in root parasites and the increase in seed set for any one species is different depending on the host.

Dodders and some mistletoes are both called stem parasites because they attach to the stems of the host; however, they have very different life histories. Mistletoes are closely related to root parasites and, in fact, their haustoria are derived from root tissue. Upon seed germination, their haustoria immediately penetrate the host's stem. Mistletoes most often grow on one individual, but in some cases they may sprawl along or hang down from the stems of the host plant. In contrast, dodders are closely related to nonparasitic vines. After germination, dodder seedlings begin to sprawl and twine along the stem of the host, forming haustoria from time to time where the new stems of the dodder contact the host. The rudimentary hypocotyl of the dodder eventually dies and it grows extensively beyond the original point of germination, often using multiple hosts.

All parasitic plants are obligate, requiring a host for survival, growth, or seed production, except for some root hemiparasites. The degree of parasitism is a useful concept for comparing the ecology of related species that vary in their dependence on the host. At one extreme, the ecology of facultative parasites is similar to their fully autotrophic relatives. At the other extreme are a few highly derived lineages of nonphotosynthetic endophytic parasites that have all of their plant body within the host tissue emerging only to flower and fruit. Among root holoparasites, there are enormous differences in the amount of biomass that is below ground. Some species are almost exclusively subterranean, growing as long-lived underground tubers, producing only flowers and fruits above ground.

Extremely derived holoparasitic plants are some of the great oddities of nature. In the 10th century, Arabic scholars wrote that parasitic plants performed the actions of animals by feeding from the juices of the host plant; they have the “souls of worms.” In 1828, a naturalist named Trattanick proposed that a separate category should be created to encompass them, the taxonomic equivalent of a mental asylum for the manic and delusional. Certainly the morphology of many holoparasites is nonplant like, and this led some early researchers to call them fungi!

The dissonance experienced on first learning about parasitism in plants betrays a common prejudice in some botanists: plants and parasites are naively regarded as opposites. As a result, parasitic plants have been misidentified and misunderstood. The “bearded-grape” was regarded as a monstrosity and assigned to its own genus; however, it was simply a grape infected by Cuscuta! Parasitism in most root hemiparasites was unnoticed by naturalists until very recently making the modern view of parasitic plants only about a hundred years old. In 1969, Job Kuijt's book, The Biology of Parasitic Flowering Plants, abolished many of the mysteries surrounding parasites. In this seminal work, Kuijt provided a critical analysis of all previous research and proposed a conceptual framework for understanding the evolution, morphology, physiology, and ecology of parasitic plants. While Kuijt's book remains the best authoritative source on the subject, new molecular evidence is being used to resolve the detailed phylogenetic history of parasitism, and a wide range of biochemical, genetic, and molecular approaches are now being applied to better understand parasite biology.
II. THE PARASITIC PLANT CLADES
AND THEIR RELATIONSHIPS TO
OTHER PLANTS

Phylogenetic relationships between parasitic plant families and their affinities to nonparasitic plants are now well known for most species. A few families of root holoparasites have been difficult to place in a phylogeny because their floral morphologies are unique and they virtually have no vegetative morphology. Molecular phylogenetic study of many parasitic plant groups has been impaired by the fact that photosynthetic and other genes commonly used in plant molecular systematics are missing or evolve at greatly accelerated rate. However, recent studies of plant mitochondrial DNA sequences have resolved 11 or 12 independent origins of haustorial parasitism (Fig. 1) distributed throughout angiosperm history.

The Sandalwood order (Santalales) are the largest, most diverse clade of parasitic plants in terms of growth habit, but all retain at least some chlorophyll and photosynthetic ability; in fact, some members may not even be parasitic. This group includes almost 2000 species of trees, shrubs, mistletoes, and at least one species of mainly endophytic mistletoe. Oleaceae and Opiliaceae are composed entirely of trees and shrubs that include both free-living as well as root parasitic members. Santalaceae, are also trees and shrubs with both root and stem parasites. The mistletoes belong to one of two large families, Loranthaceae and Viscaceae. The stem and root parasites of Loranthaceae are known as showy mistletoes because of their long, colorful, tubular flowers. Viscaceae, including the "dwarf mistletoe," Arceuthobium, are all stem parasites. In addition to the two mistletoe groups, two other families of Sandalwoods, Misodendraceae and Erechmelepidaceae, are stem parasites.

About 1700 root parasitic annuals and perennials are found in what has traditionally been classified as two closely related families, Scrophulariaceae and Orobancheae. As a whole, they represent the most diverse clade of parasitic plants in terms of variation in degree of heterotrophy. Traditional Orobancheae includes only holoparasites, while the large family Scrophulariaceae (ca. 4500 spp.) has nonparasites, hemiparasites and even some holoparasites. Molecular phylogenies based on chloroplast gene sequences have found the distinction between the two families to be artificial—a strongly supported clade contains all known parasites from both groups, but most of the rest of traditional Scrophulariaceae are not closely related to this group.

In other words, Scrophulariaceae appear to be polyphyletic, while the parasitic lineages are quite clearly monophyletic, suggesting a single origin for parasitism in this group. For this reason, dePamphilis and colleagues have proposed to include all of the parasites (plus a few closely related nonparasitic genera) in an expanded Orobancheae. Within this large group, holoparasitism appears to have evolved not less than five times. The agriculturally important genera Striga and Aleteria (both mostly hemiparasitic) and Orobanche (holoparasitic) are members of this clade, as are the large hemiparasitic genera Pedicularis, Euphrasia, and Castilleja.

The remaining parasitic plant families are not nearly as large or diverse as Orobancheae or Santalales. Plants commonly referred to as dodders are found in two families Cuscutaceae and Lauraceae. Cuscutaceae is a family of about 150 species of one genus, Cuscuta, that are distributed nearly worldwide. Lauraceae is a large family comprised of mostly nonparasitic trees that also includes one hemiparasitic genus, the vine Casytha. Casytha has about 20 species, most of which are endemic to Australia or Africa, with one species found throughout the tropics. Although strikingly similar in gross morphology, these two parasitic groups originated independently. Cuscuta is related to morning glories (Convolvulaceae), in the order Solanales, while Casytha belongs in Laurales, a basal angiosperm order. Cuscuta are usually considered to be holoparasites, but very low photosynthetic activity exists in some species; whether or not the chlorophyll-containing species can use their photosynthetic products is unknown.

Approximately 50 species of endophytic holoparasites are found in the traditionally recognized family, Rafflesiaceae. Although considered to be closely related, molecular phylogenies now indicate up to four independent origins (polyphyletic) for the genera making up this collection of extreme parasites. Molecular evidence is in agreement with previous studies of pollen morphology that suggested Rafflesiaceae be divided into four families: Rafflesiaceae, Mitrastemmacaeae, Apodanatha, and Cytnaceae. While floral morphology is diverse for these families, they were previously thought to be related because they all are endoparasites.

Rafflesiaceae includes the largest flower in the world, Rafflesia arnoldii, which measures 1 m across. Rafflesia species only parasitize species of Tetragastrina vines (Viaceae) and have an unusual reproductive biology. Emerging buds develop for approximately 1 year before they open and become receptive to pollinating flies that are attracted to the often fetid scent. Flowers die within 5 days leaving a narrow window of time for pollination to occur. Because of the short flower life span, and
because female and male flowers are on separate plants, sexual reproduction is rare. Rafflesiaceae species are all exceptionally rare and endangered throughout their range, partly due to their reproductive strategy and partly due to habitat loss. As shown in Figure 1, recent mitochondrial DNA sequence data suggest that the closest relatives of Rafflesiaceae include the rose order (Rosales).

In contrast to Rafflesia, members of Apodanthaceae, Pilostyles and Apodanthes, produce tiny flowers only a few millimeters wide. Approximately 20 species are included in Apodanthaceae, which are distributed from the southwestern United States throughout the northwestern tropics with restricted localities in Africa, Australia, and the Middle East. Curiously, Apodanthaceae appear to be most closely related to the order Fabales,
which includes their primary hosts; various species of legumes.

Cytinaceae include two genera, Cytinus and Balclayphyton. Cytinus, with 6 to 7 species, is distributed in the Mediterranean region, South Africa, and Madagascar and most commonly parasitizes species of Cistus (Cistaceae). Balclayphyton has only 1 to 4 species that are found mainly in Mexico and parts of Central America usually parasitizing members of the Burseraceae. The closest relatives of this family are classified within the cotton or hibiscus order, Malvales. Like the previously mentioned families, Raflesiaceae and Apodanthaceae, Cytinaceae are not known to cause economic loss because they do not parasitize crop plants.

Mitrastemonaceae are specialists on trees in the family Fagaceae found in southeast Asia north to Taiwan and parts of central and South America. Mitrastemon is the only genus included in this family that is made up of two named species that are totally white. Plants are protandrous and may produce large quantities of nectar.

Another small family, Hydnoraceae, is composed of about 20 species of root holoparasites native to tropical and southern Africa, Madagascar, and South America. Hydnoraceae have traditionally been considered close relatives of Raflesiaceae, based largely on the shared holoparasitic habit with a reduced vegetative body and highly modified floral structures. However, molecular phylogenies now suggest that Hydnoraceae represents an independent origin of parasitism within the basal angiosperms, and that similarities with other holoparasitic groups have evolved convergently.

Balancophasoraceae has about a hundred species in 17 genera, all of which are extremely reduced root holoparasites that have some of the smallest flowers in the world. Members of Balancophasoraceae are found throughout the tropics and have broad host ranges but do not cause extensive crop damage. Balancophasoraceae are sometimes included with the sandalwoods; however, molecular evidence does not support that relationship.

Cynomorium (Cynomoriumaceae) is a holoparasite in the Mediterranean region comprised of two species that has sometimes been included within Balancophasoraceae. As shown in Figure 1, Cynomorium is not closely related to Balancophasoraceae; instead it appears to be related to the order Sapindales, which includes the citrus family.

Krameriacae includes 17 species of root hemiparasitic shrubs and herbs in the single genus Krameria that are often found in dry, open habitats in New World deserts and subtropics. Krameria species are unusual in that they produce floral oils that are used by oil-collecting bees. Traditionally, Krameria has been placed near the legumes (Fabaceae) or milkworts (Polygalaceae) due to a superficial resemblance of their flowers. Molecular evidence, however, unambiguously suggests a close relationship of Krameriacae with Zygoophyllyaceae in spite of little morphological similarity between the two families.

Lemnaceae has two genera with five species of root holoparasites restricted to New World deserts in the southwestern United States, northwestern Mexico, and Columbia. All are herbaceous holoparasites of woody plants, including Boraginaceae (their closest known relatives) and Asteraceae. These strange parasites are spread by rhizomes often far beneath the sand or soil surface and are difficult to find unless flowering. Phorisma sonora, also known as sand food, was traditionally eaten by indigenous people.

Finally, one gymnosperm genus, Parasitaxas (Podocarpaceae), is parasitic on other free-living Podocarpaece in New Caledonia.

III. EVOLUTION OF THE DEGREE OF PARASITISM

The evolution of parasitism has been difficult to study because intermediate evolutionary stages between non-parasitic ancestor and extremely derived holoparasites are often missing. The variability in the degree of parasitism and the multiple origins of parasitism make these plants ideal for studying the evolutionary origins of parasitism and the often rapid canalization of parasitic traits. The degree of parasitism refers to increases in the dependence that parasitic plants have on their hosts. Aspects of the degree of parasitism include the amount of carbon, water, or sugar obtained from the host, as well as major changes in the way that parasitic plants relate to a host. These major shifts include the evolution of stem parasitism or the evolution of an endophytic life history. The evolutionary factors that have influenced these events remain one of the most important questions about parasitic plants.

Changes in the degree of hetrotrophy represents one of the most interesting and best-documented cases of quantitative evolution of parasitism. Some hemiparasitic root parasites and mistletoes are almost entirely autotrophic—that is, they obtain a small fraction of their carbohydrates from their host whereas holoparasitism (100% heterotrophic) occupies a presumably irreversible endpoint of this quantitative scale. Most clades of parasitic plants are entirely hemiparasitic or entirely holoparasitic. Krameriacae are all hemiparasites, while the
Hydnoraceae, Rafflesiaeaceae, Apodantheaceae, Mitragasternonaceae, Cyrtiaceae, Lenoaceae, Cynomoriaceae, and Balanoporaaceae are entirely holoparasitic. Cassytha and Cuscuta are often called holoparasites, although these species have very low amounts of chlorophyll. Cassytha species are yellow-green in color, while Cuscuta species are more orange-yellow. In Cuscuta, chloroplasts are less than 10% as numerous as in autotrophic, plants but the green color is masked by yellow pigments.

Members of the order Santalales are mostly hemiparasitic, but at least one species represents a very advanced stage of parasitism. In Arceuthobium, the endophyte is extremely well developed and the leaves are reduced to scales. The most derived parasite in the Sandalwoods is found in the endophytic parasite, Tristerix aphyllus. Tristerix aphyllus is a parasite of columnar desert cacti, whose endophytic life history may allow the parasite to escape the hot and desiccating desert conditions. Curiously, T. aphyllus and all Arceuthobium species retain some chlorophyll, and in spite of these extreme advances toward parasitism, true holoparasites are absent in the order.

In contrast to the conditions in Santalales, within Orobanchaceae, holoparasitism has evolved on several occasions. In fact, some species include hemiparasitic and holoparasitic populations and the degree of heterotrophy varies substantially. Peter Atsatt (1970) was able to produce holoparasites in a few generations in artificial selection experiments. This suggests that the complete evolutionary loss of autotrophy, as seen repeatedly in Orobanchaceaace, may be relatively common in nature. One obvious conclusion is that holoparasitism is more likely to evolve in root parasites than in stem parasites. Since root parasites are often found in the dark understorey, the selective advantage of retaining functional chloroplasts in root parasites may be much weaker than for sunbathed stem parasites. In fact, holoparasitism in beech drops (Epiplexus virginiana, Orobanchaceae) is clearly irreversible because entire sections of the genome that contain the genetic instructions for photosynthesis have been deleted.

The evolution of the mistletoe life history represents a different kind of innovation in the degree of parasitism that is different from the evolutionary loss of chlorophyll. Stem parasitism may have evolved more than once within the Sandalwoods; two families, Santalaceae and Loranthaceae, have both stem and root parasites. It is possible that these represent paraphyletic taxa, but it is also possible that study of intermediates between root and stem parasites in Santalales will lead to an understanding of the evolutionary forces that led to the evolution of stem parasitism from root parasitism.

IV. PARASITISM AND OTHER PHENOMENA IN PLANTS

The utility of the strict botanical definition of parasitism is most apparent when parasitic-host interactions are compared with other plant-plant interactions. Almost all plant-plant interactions are characterized by conflict, and many are ecologically parasitic because one species exploits another. In short, each aspect of parasitism is present in another kind of plant-plant interaction.

Parasitic plants acquire resources directly from another plant. Similar phenomena exist in root-grafted forest trees and in mychorrizal associations with fungi that indirectly connect plants. Root grafting is defined by a morphological continuum of the vascular tissue between two individuals that occurs in several species of forest trees. Interestingly, enough materials can be transferred through natural root grafts to sustain girdled trees. While root grafting is most common between closely related species, parasitic plants often form haustoria on a variety of species that are not closely related.

Mycorrhizal associations are very common in nature. Some mycorrhizal plants, such as Indian pipe (Monotropodendron, Monotropaceae) have evolved complete heterotrophy and lost functional chloroplasts, much like holoparasites; actually these myco-heterotrophs are often confused for parasitic plants. In general, mycorrhizal associations are not specific species specific, and networks of mycelia and roots from a number of individuals can be interconnected thereby allowing the exchange of resources. In root grafts as well as mycorrhizal associations, source-sink relationships may exist between individuals; however, the precise relationship depends on ecological differences that may change over time. In haustorial parasitism, the parasitic plant is generally the sink while the host is the source.

Sometimes plants exploit the habitat created by another plant at the expense of the "host." For example, epiphytes use hosts primarily as anchors. Epiphytes intercept light, minerals, and water that might otherwise be available to the host plant. The reduction of available resources, plus the mechanical stresses on the supporting structure, may impose fitness costs on the host plant. A phenomenon related to epiphytism, but potentially more harmful to the "host," involves nurse-plant relationships, in which an established plant creates a microhabitat that facilitates the establishment of another species. For example, vines use the structural wood of trees to reach the forest canopy without investing in support themselves. In an extreme case, strangler figs twine around the trunk of a host, sending down
adventitious roots to eventually support itself. Often the fig grows into a large tree that densely shades and envelopes the host and may kill it. In contrast, parasitic plants impose fitness costs by removing water, mineral, and carbon resources from the host plant. They may also physically damage conducting tissues and negatively affect the host’s metabolism through hormonal interactions.

V. SEED DISPERSAL, GERMINATION, AND PREHAUSTORIAL EVENTS

Before forming haustoria, the seeds of a parasitic plant must germinate close to a host plant and make contact. Seeds have little control over where they are dispersed and because seedlings are fragile, encountering a host is an important but difficult challenge.

New contacts between hosts and root parasites are initiated by root growth; either the host grows near the parasite or the parasite grows toward the host. Some species use chemicals exuded by host roots as a cue for finding the host such that seedlings grow chemotropically up gradients. In facultative root parasites, the plants can invest in root systems that “search” for hosts. In the meantime, these roots can absorb resources and the plants can usually persist until they contact a host. In contrast, obligate parasites do not invest in elaborate structures; they must make contact immediately or die.

Seed size and number and the degree of parasitism are correlated in parasitic plants. Holoparasites generally produce large numbers of small seeds. This correlation is probably driven by different strategies for managing the risk of finding a host. Producing larger, more provisioned seeds allows the facultative parasite to persist until it finds a host, but the production of many small seeds increases the probability that at least some seeds will land near a susceptible host root. In general there is a negative correlation in all plants between seed size and number, which is thought to be related to resource limitations. Theoretically, resources are unlimited for parasites unless the host is killed. Therefore resource limitations may not be the reason for the seed size-number tradeoff in parasites.

In some species of obligate root parasites, seed dormancy may be prolonged as a way to spread the risks associated with finding a host over time as well as space. In other words, if a seed is dispersed to a suboptimal site, dormancy allows germination to occur at later times if the site becomes appropriate. Seed dormancy is broken when chemical cues indicate the proximity of a host. The first root exudate discovered to induce parasite seed germination was Strigol. This chemical is found in cotton and it stimulates the germination of Striga seeds. It is not known how well seeds discriminate between chemical cues of hosts from nonhosts in their local environment, but seeds often make mistakes. For example, cotton has been used as a way of managing Striga in agriculture by causing seed germination in the absence of a suitable host (see Section X).

Mistletoe stem parasites must germinate on the stems of their hosts, requiring that seeds be dispersed directly onto a stem or leaf. Seeds are dispersed in two ways. Ballistic dispersal projects seeds up to 50 ft from the original host at velocities up to 90 ft/s as reported in Arceuthobium. Birds also disperse seeds of mistletoes over long distances because fruits are the main source of energy and protein for some birds. After consumption, seeds may be defecated onto hosts or the sticky seeds may attach to the bird's beak and be carried directly to another host. The seed may be wiped onto the host plant by the bird as it tries to remove the seeds from its beak.

VI. THE HAUSTORIUM

Haustoria are extremely varied structures among parasitic plants. In root parasites, haustoria are easily identified on exposed roots. They appear as swollen tissue at a contact point between parasite and host. In dodders, haustoria superficially resemble pegs or suction cups that connect the host and parasite. In general, haustorial cells occupy intercellular spaces and displace the host tissue, but enzymes also digest the host cell walls. Once a parasitic plant has encountered a host, it must penetrate the cambium and establish an interface. The interface is extremely varied. In most cases, the parasite forms a continuum with the xylem of the host plant, but in others (e.g., Cuscuta), the parasitic plant taps into the phloem. Host-derived materials may be transferred through strawlike intrusions into the host vascular tissue, or they are simply absorbed across cell walls. Haustorial cells near host tissue are usually rich in mitochondria and rough endoplasmic reticulum suggesting they are probably actively producing proteins that are likely used to produce digestive enzymes. After forming the initial haustorium, parasites may enhance local root growth to increase the number of haustoria and strength of the connection with the host. In stem parasitic mistletoes, seeds germinate and send out a radicle that grows into the host through a stomata.
VII. THE PHYSIOLOGY OF PARASITE—HOST INTERACTION

Once a haustorium has established, the question remains: Why do materials move from host to parasite? Part of the answer appears to be quite simple for most xylem-tapping species: parasitic plants usually transpire at higher rates than their hosts to maintain a gradient in water potential across the haustorial boundary. The stomata of most plants close at night to conserve water, and they may also constrict during the day to reduce rates of transpiration under drought conditions. However, in many parasitic plants, stomata remain open at night, and they tend to maintain a high transpiration rate throughout the day, even under drought conditions. Parasitic plants are among the least efficient plants at water use. In extreme cases, some species have glands to exude extra water. The net effect is that water and other dissolved substances move down the water potential gradient via mass flow from the host to the drier parasite.

If all the transfer of resources was facilitated in this manner, the concentration of nutrients should be higher in the parasite, but in the same proportion as the host plant. These general patterns are not found—instead, calcium generally has the same concentration as the host, while nitrogen, phosphorous, and potassium are enriched. Some evidence suggests that there is a component of active transport in parasitic plants. The haustorial cells at the interface have enhanced concentrations of mitochondria and show signs of being able to mobilize energy. It is possible that the main function of these cells is to pump unwanted materials back to the host. The potential benefits of reducing calcium concentrations and the mechanism(s) that make it possible are not fully understood.

VIII. THE ECOLOGICAL ADVANTAGES OF PARASITISM

In facultative parasites, the advantages of parasitism can be demonstrated experimentally. When grown with a host, parasitic plants look healthier, grow bigger, have more flowering branches, and produce more seeds. Compared with a patch of soil, the resources available in the vascular tissue of a host are extremely enriched. Parasitism allows plants access to a rich, hydroponic nutrient source. The benefit of parasitism can be mimicked experimentally by growing parasites in nutrient-rich solution culture.

A genus of annual hemiparasites, Cordylanthus, blooms during the hottest time of the year in an arid environment. Few other annual plants bloom at that time of year, and it is possible because the parasites use roots of perennial plants to gain access to groundwater. This unusual flowering time may also permit the parasites to have more exclusive access to pollinators.

In addition to nutrients and water, parasitic plants absorb secondary chemicals from hosts. Recent experiments by Michelle Marvier showed that herbivory is slowed and herbivores have lower fitness on parasitic plants that are attached to hosts that deter herbivory. Parasitic plants use a wide range of hosts, and, as a consequence, may accumulate a diverse and changing complement of protective secondary compounds. This strategy may effectively make parasites “moving targets” that are difficult for herbivores to “hit” in ecological or evolutionary time.

IX. HOST SPECIFICITY

In general, parasitic plants are considered to be host generalists. The advantage of this is quite simple: seeds can not choose where they are dispersed. Typically, a higher fraction of mistletoe seeds successfully establish on the species that is most abundant locally. In addition, dodders preferentially coil toward good hosts. Measuring host selectivity is much more difficult in root parasites than stem parasites. However, careful measurements by Gibson and Watkinson have demonstrated that root parasites are generally selective.

Some species are only found on one species or genus. The entire family, Rafflesiaceae, is composed of host specialists, as discussed earlier. Two closely related species of Orobancheae, Epifagus virginiana and Conopholis americana, are specialized on beech and oak trees, respectively. Striga (Orobanchaceae) tends to utilize grasses, and some species are highly host specific, even to the extent of requiring a particular subspecies as a host.

A tempting generalization is that species that are more dependent on their hosts tend to be more host specific. Kuijt rejected this hypothesis because most species of root holoparasites are host generalists. However, the most host specific species, including Epifagus and Conopholis, Rafflesiaceae, and Tristerix aphyllus, all represent extremely derived parasites. It may be that the endophytic lifestyle or the subterranean habit releases these plants from other selective pressures, and not that they become host specific because they are extremely
derived. New phylogenetic evidence may provide more rigorous tests of this hypothesis.

X. PARASITIC PLANTS AS PESTS

Many species of parasitic plants are economically important because they reduce crop yields. Parasitic plants remove nitrogen, phosphorus, and other mineral nutrients that may reduce the host’s ability to grow. Parasitic plants have nutritional requirements that are very similar to their hosts, therefore the damage caused by a parasitic plant is often directly proportional to its biomass. In addition, parasitic plants are less efficient at water use than their hosts, so parasitized hosts will tend to be more water stressed than unparasitized hosts. Dodders do not have the same dramatic effects on water balance as mistletoes or root parasites, but their haustoria remove sugars and stunt developing fruits. In general, crops that are infested with parasitic plants grow slower, have lower yields, and are more susceptible to disease. In addition to crop plants, parasites are important pests of trees. The invasion of mistletoe haustoria into host trees can distort wood or alter growth patterns, thereby decreasing the value of the timber. Some mistletoes and dodders damage fruit trees or ornamental plants. The damage caused by these parasites is usually measured in lower yields, the loss of aesthetic value to ornamentals, or the increased effort required to control them. An enormous amount of research has focused on the root parasitic weeds, mistletoes, and dodders.

A. Root Parasitic Weeds

The most economically important root parasites are the witchweeds (Striga, Orobanchaceae). Striga includes approximately 17 species, of which, 11 are known to be pests. Most witchweeds have a tendency to parasitize grasses including sorghum, pearl millet, finger millet, rice, maize, and sugarcane. The most important witchweed species is S. hermonthica, which ranges from Ethiopia and Sudan through Sahel, north into Arabia, and south to the Ivory Coast, Nigeria, Angola, Namibia, and the Lake Victoria basin. Striga asiatica has a much wider range that includes much of Africa, parts of India, China, and Australia. In addition, Striga asiatica was accidentally introduced into North America in the 1950s. One species, S. gesneroides, has very different host range from the other witchweeds; it uses a diverse set of broad-leaved hosts. It is best known for the damage it causes to cowpeas in West Africa, but it also damages tobacco in East Africa.

Among root parasites, the holoparasitic broomrapes (genus Orobanche, Orobanchaceae) are second to the witchweeds in terms of their economic impact. The geographical ranges of the economically important species are centered in the Middle East, but they are found further west and south into Africa, west and north into Europe, east into India, Pakistan, and Nepal, and north into Afghanistan, and several countries that were once in the Soviet Union. In general, broomrapes have wide host ranges. Orobanche ramosa attacks crops in Solanaceae, Brassicaceae, and Fabaceae. It sometimes attacks onions, but in general it is never found on grains (it was once reported on maize). The host range of O. cernua is more restricted; it parasitizes only Asteraceae (mainly sunflower) and Solanaceae (tomato, potato, and eggplant). Orobanche crenata Forsk. has a wide host range, but it is most important as a pest of Fabaceae, especially faba beans and some Apiaceae (carrots).

The genus Alectra (Orobanchaceae) is closely related to the witchweeds. Of the 30 species, 4 are notable pests. Alectra vogelii and A. pica are pests of cowpeas and other pulse crops in semiarid Africa. Alectra orobancheoides occasionally attacks sunflowers or tobacco in South Africa and A. fluminensis parasitizes sugarcane in Central and South America. Several other genera in Orobanchaceae have agriculturally important species, but they are not nearly as important as the witchweeds, broomrapes, or Alectra. Other genera with notable or potentially important pest species include Buchnera, Ramphicarpa, Odontites, Rhinanthus, Aeginetia, Melanpyrum, and Christoria.

Most control methods for root parasites focus on reducing host damage as well as seed set. Witchweeds and broomrapes produce hundreds of thousands of small, dust-like seeds per plant. Furthermore, the seeds can lie dormant in the soil for several years, and therefore reducing the size of the seed bank involves enormous effort. The most effective available method of control involves hand pulling the plants before they set seed. Although most of the damage to the host has been done by the time the parasitic plants emerge, hand pulling can reduce some of the damage. More important, early hand pulling prevents reseeding. If reproductively mature plants are pulled, care should be taken to prevent the seeds from being dispersed, and the adult plants should be burned.

More direct chemical methods can be used to reduce the seed bank. Ethylene gas, or artificial chemicals that mimic germination stimulants, can be applied to break dormancy and induce germination at times when no appropriate hosts are available. Alternatively, fumigants are applied to fields that kill seeds as well as other
soil organisms. Raising the temperature of the soil by covering it with black plastic can also kill seeds.

Trap cropping, catch cropping, and crop rotation are other effective ways of managing the land. One option is to let the land lie fallow for 10 to 20 years to reduce the seed bank. This may not be an option where the demands on land use are heavy. Planting an unsuitable host is a more effective alternative that may allow the land to be used. Some alternative crops, called trap crops, induce seed germination but are not suitable hosts. A third option involves catch cropping. The land is seeded with susceptible hosts to induce germination of parasitic plants, but the crops are destroyed before the parasite can reproduce. Mixed cropping involves planting suitable and unsuitable hosts, and it has been demonstrated to reduce the impact of some witchweeds. The mechanisms, however, are not well understood.

Herbicides can also be used to control witchweeds. Herbicides applied directly to the soil after planting but before crop seedlings emerge can reduce or delay witchweed establishment. Systemic herbicides that are applied to the crop immediately after emergence may become translocated and concentrated in the parasites, killing them. The timing of application of systemic herbicides is critical because it must degrade before the crop begins to fruit. Some herbicides can be applied directly to the emerged witchweed to control them.

Damage to crops by witchweeds is minimized when nitrogen fertilizer is applied to host crops. This method has the added advantage of helping increase yields, but the benefits of using this method alone over a long period of time have not been demonstrated. The mechanisms that allow nitrogen fertilization to decrease the impact of the parasitic plants are not fully understood. It does seem clear, however, that on poor soils, the witchweed infestations increase in intensity as the quality of the soil decreases; the result is a feedback that continues until the land is almost useless and must be abandoned.

One of the most promising control options is the development of resistant varieties of crop plants through plant breeding. As new resistant varieties are found other methods of control become unnecessary. Resistant varieties of many crops have not been discovered, however, therefore this is a very active area of research.

Parasitic plants may pose the largest problem to rural, subsistence farmers. As pressures on land use increase, the ability to use crop rotation methods decreases. Furthermore, subsistence farmers often do not have the means to buy fertilizers, herbicides, or expensive resistant seeds. For these farmers, the feedbacks between poor soils and parasitic plant infestation may eventually cause the farmers to abandon plots of land and move elsewhere.

**B. Mistletoes**

Several species of mistletoes are economically important pests of fruit trees, ornamental trees, or timber. Most mistletoes have a similar ecology, so the impact and control measures are easy to generalize. The exceptions are dwarf mistletoes (genus *Arceuthobium*, Visceae) that have a much more highly developed endophytic component and inconspicuous, scaly leaves. Because of the highly developed endophyte, physical removal of dwarf mistletoes is more difficult than other mistletoes.

Most species of dwarf mistletoes are extremely host specific. For example, *A. douglasii* is a parasite of Douglas fir, and *A. tsugense* is a parasite of the western hemlock. The dwarf mistletoes are parasites of pine trees in North America, the Himalayas of India, Pakistan, and Bhutan, and in southwestern China. Dwarf mistletoes are common throughout the Rocky Mountains where large regions of forests are infested and the timber volume and quality are affected.

Two other genera in Visceae are notable for their economic impact. *Phoradendron* is a genus of about 190 species distributed throughout North and South America. Of these, the most important are *P. serotinum*, found on an extremely wide variety of hosts (but never conifers), and *P. piperoides*, found on cocoa in Costa Rica. *Viscum* is a genus of about 60 species that include *V. album* that has an extremely wide host range including fruit trees, pines, and poplars in Europe and persimmon in China. Another agricultural pest is *V. cruciatum* that parasitizes olives in the Middle East.

Many species of Loranthaceae cause economically significant loss. *Dendrophthoe falculata* is a pest of fruit trees in India and teak in Kerala. *Dendrophthoe pentandra* parasitizes rubber and kapok in Indonesia. *Helianthera mannii* is a pest of citrus and coffee in West Africa, and *H. parasitica* is a pest of citrus in the Himalayas. *Tapinanthus bangwensis* is an important pest of cocoa and cola in Ghana, while *T. globiferus* grows on coffee, citrus, and other fruit trees in Ethiopia. Some other genera that have economically important species are *Amyema*, *Englerina*, *Loranthus*, *Macrosolen*, *Oryctanthus*, *Phragmanthera*, *Scruella*, and *Struthanthus*.

Some resistant varieties of poplars, pines, and oaks do exist, but very little variability for resistance to mistletoe infections exists and, in general, breeding for resistance has not been a viable option. Biological control by insects, pathogens, or other mistletoes remains...
a possibility, but none of these methods have been successful. Instead, control of the leafy mistletoes involves pruning off infected limbs, or chemical control. Pruning infected limbs removes the mistletoe, but the limb must be cut below the point of infection to remove the entire endophyte. This practice is extremely labor intensive and impractical on a large scale. With *Arceuthobium*, the endophyte is often extensive, and the limb must be severed 20 to 30 cm below the lowest parasite shoots. In extreme cases, it may be more cost-effective to burn the forest. In less extreme cases, early harvest followed by controlled burns may reduce the mistletoes. Some successful chemical control methods have been found. The leaves of mistletoes may be sprayed with ethephon once the host drops its leaves. Alternatively, herbicides can be injected into infected limbs below the point of infection, accumulate in mistletoes, and eventually kill them without severely affecting the host.

C. Dodder

The economic impact of *Cuscuta* species is greatest in forage crops such as alfalfa and clover; however, crops like citrus, coffee, peach, litchi, flax, linseed, and other crops as well as ornamental plants all suffer from dodder attack. The most economically important dodder is *Cuscuta campestris*. This species attacks alfalfa and has been shown to reduce forage yield by as much as 57% over a 2-year period. It also affects Niger seed in India and many vegetable and flower crops. Of the 20 species of *Cassutha*, the most damaging is *C. filiformis*. It is distributed throughout the tropics and parasitizes a wide range of hosts. It is a problem on citrus in India and Tanzania and on *Pinus massoniana* in China.

Mechanical control methods for *Cuscuta* include hand pulling, crop rotation, burning, delaying planting until after *Cuscuta* have germinated, or deep ploughing to reduce the seed bank. Few resistant varieties of crops are known. Chemical control methods include fumigants to eliminate the seed bank, herbicides applied to the soil to prevent seedling growth, and herbicide application after seeds have germinated to prevent establishment. Some methods of biological control have been established with insects and pathogens, but the scope of these has been limited.

XI. CONCLUSIONS

Parasitic plants have been understudied, but recent studies have improved our understanding of their evolution, ecology, molecular biology, and physiology. Haustorial parasitism has evolved independently at least 12 times within angiosperms and, surprisingly, complete loss of photosynthesis has occurred multiple times within some lineages. Most parasitic plants exhibit near-normal levels of photosynthesis like their autotrophic ancestors, whereas others are incapable of photosynthesis such as the derived endophytic holoparasites. Study of parasites that photosynthesize at intermediate levels will be important to understand the evolution of holoparasitism. Development of sophisticated molecular methods will be important for understanding the bases of haustorial development and should ultimately enhance efforts to control economically important crop pests. Further ecological and physiological studies aimed at understanding host-parasite interactions will also enhance efforts to control parasitic weeds.

Acknowledgments

We thank the NSF (grants DEB-9102925, DBI-9604814, and DEB-9811362 to C. W. D.) for funding our research on parasitic plants, and Joel McNeal for critical reading of the manuscript. Other important contributions of data, samples, or discussion, were also made by other members and friends of the dePamphilis lab, past and present, and by numerous botanical gardens and botanists with an interest in parasitic plants.

See Also the Following Articles

COEVOLUTION • PARASITISM • PARASITOIDS

Bibliography


HERBACEOUS VEGETATION, SPECIES RICHNESS IN

J. P. Grime
The University of Sheffield

I. Introduction
II. Plant Traits or Species Richness?: The Current Debate
III. Components of Species Richness: Dominants, Subordinates, and Transients
IV. Declining Species Richness and Ecosystem Reassembly
V. Conclusions

GLOSSARY

**dominant species** Species that have the greatest influence on ecosystem structure and function by virtue of their abundance, biomass, or coverage.

**mass ratio hypothesis** Hypothesis stating that ecosystem processes are largely determined by the dominant contributors of the overall plant biomass, that is, dominant species will exert greater influence on processes than will subordinate species.

**species richness** Number of species present in a given habitat or ecosystem.

**subordinate species** Species that have a minor influence on ecosystem structure and function, presumably because of their lesser abundance and biomass compared to dominant species.

**transient species** Species that are present as scattered seedlings or small immature individuals; many of these species occur as dominant or subordinate species in neighboring vegetation associated with different environmental conditions or management regimes.

**IN THE SECOND HALF** of the twentieth century, reductions in plant diversity have been observed in many types of ecosystems. One of the best-known and most lamented of these losses has been associated with the widespread disappearance and degradation of ancient species-rich meadows and pastures in western Europe. However, from census data collected in several countries (Ratcliffe, 1984; Thompson, 1994; Thompson and Jones, 1999), we have learned not to regard this phenomenon in isolation from processes taking place in landscapes as a whole; the fate of such highly prized communities is merely one conspicuous element in a widespread attrition affecting both species-rich and species-poor ecosystems.

I. INTRODUCTION

Ecological research into losses in plant diversity has a long history, but since 1990 the agenda has shifted from a primary concern with its mechanisms to an assessment of the extent to which losses impair the functioning of ecosystems and reduce their utility to humankind. It is relatively easy to manipulate the spe-
cies composition and diversity of herbaceous communities in both natural and model ecosystems, and experimenters have applied themselves with vigor to this new field of enquiry. The objectives in this research are twofold. First, we need to identify where and how losses in diversity affect ecosystem functions such as primary productivity, carbon storage, mineral nutrient cycling, and ecosystem resistance and resilience. Second, we require information on the plant constituents necessary for successful ecosystem restorations; we need to establish which plant species are irreplaceable and what population sizes and pools of genetic variation are required for their persistence.

As a participant in both the current and the preceding phases of research on plant diversity, I have become convinced of the value of connecting theory and data from the earlier phases of research into the causes of plant diversity to the current (post-1990) effort to examine its consequences for ecosystems. The main objective of this article is to identify these connections. An additional purpose is to place in perspective a current debate in which divergent conclusions have been drawn concerning the importance of high plant species richness as an immediate controller of ecosystem properties. This debate provides a useful introduction to the issues that will be considered later.

II. PLANT TRAITS OR SPECIES RICHNESS?: THE CURRENT DEBATE

A. Effects of Plant Traits on Ecosystem Properties

Although the majority of investigations in plant eco-physiology have sought to explain how vegetation is determined by environment, there has also been widespread recognition of feedbacks in which plants affect their environments and influence ecosystem properties. This idea is implicit in the writings of the founders of ecosystem theory (Odum, 1963, 1969) and more recently has found expression in the search for plant functional types (Grime, 1979; Chapin, 1980; Smith et al., 1996). In which specific plant community and ecosystem properties have been attributed directly to particular plant traits such as potential growth rate, palatability to herbivores, resistance to fire, litter quality, and seed persistence. This approach has been extended further by studies in which the predictive value of plant traits has been tested by examining ecosystem responses to perturbation. An early example is the investigation by Leps et al. (1982) in which differences between two neighboring grasslands in their resistance and resilience following the severe drought of 1975 in Czechoslovakia were predicted from the life histories and potential growth rates of the main component plant species. More comprehensive tests were conducted by MacGillivray et al. (1995), in which responses to frost, drought, and fire in five contrasted grasslands were successfully predicted from a set of plant traits measured in a laboratory screening program (Grime et al., 1997).

B. The Mass Ratio Hypothesis

It is important to note that in tests such as those conducted by MacGillivray et al. (1995), predictions are weighted according to the abundance of each plant species in the vegetation; this is founded on the assumption that the extent to which the traits of a species affect ecosystem properties is likely to be strongly related to the contributions of the species to processes such as photosynthesis, mineral nutrient capture, transpiration, and provision of substrates exploited by herbivores and decomposers. There is a clear implication here that ecosystem processes are determined to a very large extent by the characteristics of the dominant contributors to the plant biomass. As a corollary of this “mass ratio hypothesis” (Grime, 1998), we would not expect minor contributors to the vegetation to exert strong effects on ecosystem properties.

Is it reasonable to conclude that ecosystem functions can be sustained by inputs originating from only a few dominant plant species? Some of the most extensive ecosystems contain very few plant species; this is particularly obvious in grasslands and heathlands on acidic soils, and here the conclusion is inescapable that controls on ecosystem processes by vascular plants are mediated through very few species. In herbaceous vegetation on soils of higher base status, analysis is complicated by the presence of more species, and in old calcareous grasslands the densities of herbs can rise to 30–40 species per m². Even here, however, it is interesting to note that when quantitative studies are made by harvest methods (Al-Mufti et al., 1977) or point analysis (Mitchley and Grubb, 1986; see also Fig. 1), it is evident that in comparison with the canopy dominants of the subordinate species of such communities account for a very small proportion of the biomass. There must be considerable doubt as to whether such minor components can, even collectively, exercise immediate effects on properties such as productivity, carbon storage, and water relations.
FIGURE 1. Vertical distribution in the leaf canopy of four dominant and four subordinate component species in an ancient limestone pasture at Buxton, North Derbyshire, England. Canopy distribution in June was estimated by measuring contacts with 375 randomly distributed, vertical pins. (S. H. Hillier, unpublished data.)
C. Effects Associated with High Species Richness

Against a theoretical background in which the traits of dominant plants were widely suspected to be acting as the overriding controllers of ecosystem properties, considerable interest and controversy were generated when, in 1994, two papers appeared purporting to demonstrate the immediate benefits to ecosystem properties arising from high species richness in experimental plant assemblages (Naem et al., 1994; Tilman and Downing, 1994). In each case it was suggested that benefits arose in the species-rich mixtures from the presence of a wider range of morphologies and physiologies, generating either complementary and more complete exploitation of resources (Naem et al., 1994) or conferring resistance and resilience in the face of an extreme event (Tilman and Downing, 1994). Interest in these publications extending beyond the realm of ecology was stimulated by commentaries (Kareiva, 1994, 1996) suggesting that studies of this kind provided a justification for the conservation of species-rich ecosystems.

Subsequently, doubts have been cast on the validity of the conclusions drawn by the authors of the two papers and these have been reviewed in detail elsewhere (Givnish, 1994; Aarssen, 1997; Garnier et al., 1997; Huston, 1997; Grime, 1998; Hodgson et al., 1998; Holmes, 1998). It appears that, in both cases, ecosystem properties attributed to high species richness were in reality due to the presence in the more diverse communities of dominant species with traits assumed, respectively, to high productivity (Naem et al., 1994) and drought (Tilman and Downing, 1994). More recent experiments (Wardle et al., 1997; Hooper and Vitousek, 1997; Tilman et al., 1997) have failed to provide convincing support for the effects of high species richness on ecosystem functions; the most parsimonious explanation for the data presented in all of these papers is that the ecosystem properties examined were controlled by the functional traits of a relatively small number of species accounting for a high proportion of the total plant biomass.

The slowness with which some commentators and conservation enthusiasts seized upon supposed evidence of the direct benefits of species richness to ecosystems emphasizes the need for caution in a complex research field. It also brings the risk that exposure of false conclusions may lead to a sceptical response if and when more substantial evidence of the beneficial effects of species richness is presented. Perhaps the greatest danger arising from the current state of research on this subject is the almost exclusive emphasis on immediate benefits of richness on ecosystem functions. Elsewhere (Grime, 1998) it has been proposed that the most potent effects of declining biodiversity are likely to become evident only in the longer term. To explore this argument it is necessary to refer to much older research on plant diversity that has been largely ignored in the recent burst of new experimental work.

III. COMPONENTS OF SPECIES RICHNESS: DOMINANTS, SUBORDINATES, AND TRANSIENTS

A. Dominants and Subordinates

Early efforts to describe and interpret herbaceous vegetation (e.g., Clements, 1905; Tansley, 1939; Ramenskii, 1938) involved listing of the plants present in selected stands of vegetation and estimating the abundance of each species. Even before the widespread adoption of experimental approaches there was a keen awareness of the potential of certain species to occupy a high proportion of the plant biomass and to control the abundance and fitness of other minor contributors. A classic example is the rapid expansion of tall coarse grasses and the coincident suppression of small herbs noted by Tansley and Adamson (1925) in their study of the consequences of excluding rabbits from chalk grassland.

A more quantitative approach followed as plant ecologists adopted stricter sampling methods and measured the relative abundance of plant species by harvesting, sorting, and weighing vegetation samples. This allowed more formalized attempts to examine the functional relationships between dominant and subordinate members of plant communities (Whittaker, 1965; McNaughton, 1978). Subsequently, it was recognized that the potential to dominate vegetation could be associated with specific plant traits such as canopy height, lateral spread, and the capacity to project shoots forcefully through litter and herbaceous cover (Al-Mufti et al., 1977; Sydes and Grime, 1981a,b; Grubb et al., 1982; Campbell et al., 1992). A further step in defining the functional differences between potentially dominant and subordinate members of plant communities became possible through the application of foraging theory to plants. From a study of the development of the roots and shoots of isolated plants exposed to standardized patchy environments (Campbell et al., 1991), it became possible to predict the dominance hierarchy that devel-
opposed when eight herbaceous species were grown together in an experimental assemblage (Fig. 2). In this investigation it was found that dominance was associated with relatively imprecise foraging both above and below ground. Such coarse-grained foraging allowed the exploitation of a large volume of habitat but was associated with an imprecise concentration in resource-rich sectors. In marked contrast, subordinate species were characterized by precise but local placements of roots and shoots in resource-rich patches, a foraging pattern permitting temporary coexistence with dominants but apparently committing them to a minor status in the community and leaving them vulnerable to extinction in circumstances where the dominants continued to enjoy unrestricted growth.

These findings raise important questions concerning the ecology and evolution of the very large numbers of plant species that consistently occupy subordinate positions within the hierarchies of herbaceous communities. Are these plants merely "also-rans" in the struggle for existence or are there fitness benefits associated with playing a minor part in plant communities? What is the precise role of subordinates in the long-term functioning and dynamics of ecosystems? These are not trivial questions in view of the fact that it is the subordinates rather than dominants that account numerically for the existence of high species richness in plant communities. Detailed attempts have been made to address these issues (Grime, 1987, 1998); here argument will be reduced to the following propositions:

1. Many subordinates achieve widespread occurrence because they are capable of exploiting similar niches associated with different dominant species. For example, Poa trivialis, the commonest vascular plant species of Lowland Britain, occupies shaded microhabitats beneath tall grasses, broad-leaved herbs, and a variety of shrubs and trees.

2. In many species-rich communities the risks to subordinates of competitive exclusion remain low because potential dominants are restricted in vigor by environmental and biotic factors (e.g., mineral nutrient stress, drought, grazing, fire).

3. Because they do not experience the costs in support materials and construction time incurred during the building of robust root and shoot architectures, subordinates enjoy a temporary advantage when communities are subject to episodes of biomass removal. Where damage is catastrophic, as in tree felling and coppicing or the ploughing and burning of grasslands (Skutch, 1929; Marks, 1974; Platt, 1975; Bormann and Likens, 1979; Pickett and White, 1985; Pons, 1989), this can lead to a temporary but massive expansion of subordinates until such time as the dominants recover.
We may conclude, therefore, that many subordinate members of plant communities achieve high levels of fitness, particularly where natural or human-inspired interventions restrict the vigor of potential dominants either continuously or intermittently. Later in this article, it will be suggested that, in the long term, this rather opportunistic ecology displayed by subordinates may be sufficient to allow them an indirect but crucial involvement in the determination of ecosystem properties.

B. Transients

Many published records of the species composition of vegetation are incomplete. This can arise from cursory sampling and recording or from a deliberate policy of discounting minor constituents. Where the objective is to recognize recurring plant communities, data analysis often involves procedures that exclude species of low frequency or inconsistent occurrence in vegetation samples. The result of these various decisions can be a divergence between published data and field reality, and it has been suggested (Grime, 1998) that important information relevant to the long-term dynamics of ecosystems may be lost. To examine this hypothesis it is necessary to refer to surveys in which a strenuous attempt has been made to include all the plant species present in the vegetation.

During the period 1965–1975, a series of vegetation surveys was conducted to produce an inventory of the herbaceous plant communities of the Sheffield region in north-central England (see Grime et al., 1988, for details). In each of approximately 10,000-m² quadrats, a complete list of species was made and the frequency of each species was recorded. When the resulting data were examined it became apparent that, in addition to dominants and subordinates, the majority of the samples contained species that were represented only as scattered seedlings and small immature individuals. Most of these species were present as dominants or subordinates in neighboring vegetation associated with different environmental conditions or management regimes. Examples of this phenomenon are provided in Table I for three contrasted sets of vegetation samples; in each case it is evident that the plant communities examined were harboring juveniles of species that occur as persistent, reproductive populations elsewhere in the landscape in vegetation of a different type.

Therefore, it seems reasonable to conclude that in addition to the dominant and subordinate members of a community, a third contributor to species richness can be identified in the form of transients originating either from the seed rain from the surrounding landscape or from seed banks occurring as a legacy of previous vegetation types that occupied the site.

IV. DECLINING SPECIES RICHNESS AND ECOSYSTEM REASSEMBLY

If immediate control of the functional properties of an ecosystem rests with dominant plants and if species richness depends on the numbers of subordinates and transients, it is pertinent to ask "Does a decline in species richness matter?" To address this question, it is

<table>
<thead>
<tr>
<th>Sampled habitat</th>
<th>Number of m² samples</th>
<th>Woodland species</th>
<th>Grassland species</th>
<th>Arable species</th>
<th>Others²</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Woodland on limestone</td>
<td>51</td>
<td>69</td>
<td>23</td>
<td>1</td>
<td>5</td>
<td>94</td>
</tr>
<tr>
<td>Meadows</td>
<td>40</td>
<td>7</td>
<td>64</td>
<td>8</td>
<td>0</td>
<td>79</td>
</tr>
<tr>
<td>Cereal arable</td>
<td>55</td>
<td>5</td>
<td>38</td>
<td>69</td>
<td>2</td>
<td>114</td>
</tr>
</tbody>
</table>

* All the species encountered within a particular habitat were classified in terms of their primary habitat (columns 3–6). Details of the sampling, recording, and habitat classification procedures are provided in Grime et al. (1988).

² Includes species primarily associated with wetland or skeletal habitats (cliffs, walls, and rock outcrops).
necessary to consider the significance of losses of subordinates and transients in the long-term dynamics of plant communities and ecosystems.

A. Filter Effects of Subordinates?

One mechanism whereby losses in subordinates could affect ecosystems is through alteration of the filter controlling the recruitment, identity, and relative abundance of dominants. To review the opportunities for subordinates to control the admission of dominants into communities it is necessary to consider the long-term dynamics of vegetation and the regenerative phases in the life cycles of dominants. Studies of vegetation succession conducted earlier this century (e.g., Watt, 1925, 1947) established that continued dominance by particular species is frequently determined by the success of seedling or vegetative reestablishment following disturbance events. As already described here, often the early course of events following a disturbance is a temporary expansion in the cover and vigor of subordinates. This is most obvious in forest clearings where a dense, low cover of shrubs, herbs, and bryophytes characterizes the environment of regenerating trees (Watt, 1925; Skutch, 1929; Marks, 1974; Bormann and Likens, 1979), but similar phenomena have been described for grasslands and heathlands (Oosting, 1942; Keever, 1950; Hillier, 1990). Establishment following disturbances involves complex interactions of seedlings and vegetative shoots with substratum conditions, and contributions to the ground cover by subordinate plants may be expected to have both positive and negative effects (Pickett and White, 1985). Benefits to establishment have been described in circumstances where seedlings survive in the shelter afforded by low-growing shrubs, herbs, and bryophytes. Detrimental effects of shrub, herbaceous, and bryophyte cover on the establishment of grassland and forest dominants have been observed (Niering and Goodwin, 1962; Webb et al., 1972; Pons, 1989), and it is widely recognized that many small-seeded herbs, trees, and shrubs are incapable of establishment in a closed cover of vegetation.

We may deduce, therefore, that there is a potential for subordinate members of a plant community to act as a filter in selecting between different potential dominants during the early phases of recolonization following a disturbance event. Selection could operate on the basis of variation in the seed reserves of dominants and on the capacity of their seedlings to penetrate a low canopy. The filter might also discriminate between dominants that rely on rapid emergence and those that regenerate by persistent juveniles. Subordinates could also control regenerating dominants through more indirect mechanisms, such as provision of sites in which seed predation is reduced, or through more complex phenomena such as the maintenance of critical pests, pathogens, herbivores, or mutualists.

The filter roles that subordinates may play during ecosystem reassembly require confirmation by carefully designed, long-term experiments.

B. Founder Effects of Transients?

If, as suggested earlier in this article, the sources of the transients are seed banks in the soil and the seed rain from the surrounding landscape, it would appear that they are an index of the pool of potential colonizing species at each site. On this basis, a diversity of transients signifies a high probability that, in the event of habitat disturbance or changes in management, there will be a rapid ingress of different plant functional types, some of which may be capable of exploiting the new conditions. An obvious example is the benefit to woodland development where an abandoned grassland already contains a diverse assortment of tree seedlings.

Current losses in biodiversity in Europe and in many other parts of the world are taking place in a complex landscape mosaic that is continuously disturbed by natural events and by urbanization, arable cultivation, forestry, and various forms of grassland management. Ecosystem sustainability depends in part on the continuous movement of populations and the reassembly of vegetation types and ecosystems. The extent to which communities and ecosystems are rapidly reconstituted is likely to be related to the reservoir of colonizers, many of which should be detectable prior to disturbance as transient constituents of the existing vegetation. As Egler (1954) recognized, we may suspect that the speed and completeness with which ecosystem reassembly occurs will depend on early colonization by appropriate dominants and subordinates; late arrivals will be delayed in their establishment and some may be excluded completely (Keever, 1950; Niering and Goodwin, 1952; Platt, 1975). It is not difficult to envisage how circumstances could then arise whereby efficiently dispersed plant species with "poor fit" to habitat and management conditions could assume dominance with damaging consequences for ecosystem function. There is an urgent need to discover the extent to which failure in the processes of plant dispersal and ecosystem reassembly can be predicted from the decline through time in the density and species richness of transients in plant communities.
V. CONCLUSIONS

Currently the balance of evidence (Huston, 1997; Grime, 1998) is shifting toward the mass ratio hypothesis and against the proposition that species richness itself controls the immediate functioning of ecosystems. However, this does not mean that losses of plant diversity should be ignored. The priority in the next phase of research on declining plant diversity should be to consider its long-term consequences for ecosystem structure and function. Losses in species richness may be associated with lesser obvious impacts that operate through failures in filter and founder effects. A progressive loss of ecosystem functions may be predicted in circumstances where vegetation patch dynamics and ecosystem reassembly continue against the background of a declining pool of colonizing propagules. The effects on the recruitment of dominants, rather than the immediate consequences of declining richness per se, deserve our curiosity and attention.

See Also the Following Articles

• Ecosystem Function, Principles of • Herbicides • Plant Biodiversity, Overview

Bibliography


I. The Concept of a Weed
II. Weed Management
III. Effects of Herbicides on Biodiversity
IV. Conclusions

GLOSSARY

herbicide Chemical used to suppress or kill plants, or to severely interrupt their normal growth processes.
herbicide resistance Inherited ability of a plant to survive and reproduce following exposure to a dose of herbicide normally lethal to the wild type. In a plant, resistance may be naturally occurring or induced by techniques such as genetic engineering or selection of variants produced by tissue culture or mutagenesis.
herbicide tolerance Inherent ability of a species to survive and reproduce after herbicide treatment. This implies that there was no selection or genetic manipulation to make the plant tolerant; it is naturally tolerant.
integrated weed management Approach for suppressing weeds that combines information on the biology and ecology of the weed with all available control technologies so that no one method is used exclusively.
postemergence herbicide Herbicide applied after the emergence of the specified weed or crop.
pree emergence herbicide Herbicide applied to the soil prior to the emergence of the specified weed or crop.
selectivity Phenomenon in which some plants are killed with doses of herbicides that have little or no effect on other plants.
weed Plant that interferes with the growth of desirable plants and is unusually persistent and pernicious. Weeds negatively affect human activities and as a result are undesirable.
weed control Reducing or suppressing weeds in a defined area to an economically acceptable level without necessarily eliminating them.
exotic species is far greater than that caused by all forms of environmental pollution, including herbicides. This article reviews concepts of weeds; principles of weed management; and categories, action, and fate of herbicides. Impacts of herbicides on biodiversity of both target (weeds) and nontarget species and the role of weed control in preserving biodiversity are also discussed.

I. THE CONCEPT OF A WEED

Weeds are most often defined in human or anthropomorphic terms, that is, as plants growing where they are not wanted or plants out of place. More useful definitions of weeds are those that describe biological traits or characteristics. A list of "ideal characteristics of weeds," developed by Baker (1974), is widely known and cited for its description of traits that confer weedi-ness on plants that possess them (Table I). These traits include germination under a broad range of conditions and over an extended period; rapid growth and prolific reproduction by sexual and asexual means; flexible breeding systems, including self-pollination, cross-pollination, and nonspecialized pollinators; effective seed dispersal; plasticity and tolerance of a breadth of environmental conditions; and adaptations for competitive-ness. Though no plant species could possess all of these traits, plants considered to be major weeds are likely to possess many of them. Traits alone do not determine whether a plant will be a weed, but Baker's list is a useful tool for categorizing and studying plants that interfere with human activities and, thus, are called weeds.

A. Impacts of Weeds

By their very definition, weeds affect many activities in which humans are engaged. In agroecosystems their primary effect is to reduce crop yield and quality through competition for limited resources. Weeds also increase the time and costs required for crop production and interfere with harvesting. Indirectly, weeds affect crops through both positive and negative interactions with insect herbivores and their natural enemies. In rangelands, weeds possessing thorns or barbs pose physical hazards to livestock, and those containing tox-ins may cause allergies or poisonings of animals or humans. Other negative impacts of weeds include ob-structing visibility around roadways, serving as a fire hazard, impeding use of recreation areas, and blocking the free flow of water in waterways, irrigation canals, and drainage ditches. In nonagricultural ecosystems, weeds often comprise the first stage of plant succession on land where the native vegetation has been disturbed. With increasing movement of humans across continen-tal boundaries, exotic (nonnative) weed species have invaded many wildland ecosystems where human activ-ities have disrupted the growth of indigenous (native) species. As a result, biological diversity has been re-duced in many wildland areas that interface with urban areas. Other impacts of weed invasions in wildlands include alteration of ecosystem processes, support of nonnative animals, fungi, or microbes, and hybridiza-tion with native species to alter gene pools.

B. Weed Science

Since the beginning of agriculture, humans have em-ployed various tactics to remove weeds from land where other uses are desired. In the United States today, bil-lions of dollars are spent annually for weed removal using chemical and mechanical means. With the discov-ery of synthetic organic herbicides in 1941, weed sci-ence developed as a formal scientific discipline. Weed scientists have been extremely successful during this century in developing techniques to remove weeds from agricultural and other ecosystems. Over the past 50 years, weed science has grown into a multidisciplinary
field of study encompassing fundamental research along with applied aspects of weed suppression. Today it encompasses researchers from numerous scientific disciplines, including chemistry, ecology, genetics, morphology, and physiology, as well as applied scientists and practitioners who focus on weed suppression. In recent years, increased awareness of environmental concerns has shifted the emphasis in weed science from a primary focus on herbicides to more integrated, ecological approaches for dealing with weeds.

II. WEED MANAGEMENT

In crop, forest, and rangeland production systems, as well as in wildlands, weed presence must usually be minimized to achieve a desired land use goal. Various tools and methods are used to suppress or remove weeds while not injuring the crop or desirable species. Weed management is a strategy that includes growing or fostering desired or beneficial vegetation while suppressing unwanted plants. For such management to be successful, knowledge of the biology, ecology, life history, and taxonomy of the weed species is required, as well as selection of the proper tools to use for their suppression.

A. Tools and Methods of Weed Management

The strategy of weed management includes three key components: prevention, eradication, and control. Prevention is keeping a weed from being introduced into an area where it does not already occur. Common preventive measures include using sanitary practices, eliminating weed spread through seeds and vegetative propagules, using quarantines, and following federal and state weed laws and regulations. Eradication is the total elimination from a particular area of a weed species and any plant parts capable of reproducing. Although eradication is often a stated goal of weed management programs, it is seldom achieved owing to the presence of seed banks and vegetative bud reserves in the soil of weed-infested areas. In contrast to eradication, weed control is the suppression or reduction of a weed species to an economically acceptable level. Under weed control programs, complete elimination of the weed is not the goal; instead, weeds are reduced to a level at which the cost of continued suppression does not exceed the value of the land or crop growing on it plus the benefit afforded by weed control. Once a weed is established in an area, weed control is the approach most commonly used to manage vegetation in that area over the long term. Methods used for weed control include biological, chemical, cultural, and mechanical techniques.

B. Integrated Approaches to Weed Management

With the growing recognition and concerns about the environmental impacts of agricultural practices, particularly the use of herbicides, integrated approaches for weed management have become commonplace. Integrated weed management (IWM) refers to a strategy for weed suppression that combines information on the biology and ecology of the weed with all available control technologies. Using this strategy, a variety of different methods are used in weed control, including nonchemical ones, as well as preventive measures, such that emphasis on herbicides is minimized. An alternative approach to crop production that is currently receiving widespread attention is sustainable agriculture, which refers to production systems in which external inputs, including synthetic fertilizers and pesticides, are minimized or avoided. However, current knowledge about the biology and ecology of weeds, their interactions with crops and wild plants, and nonchemical methods for their control is still limited. Thus, in most agricultural systems today, achieving weed control without herbicides requires extensive mechanical and hand labor, which is very costly. Until viable and economical alternatives to herbicides are available, most crop production and land management systems in the United States will continue to depend on some level of herbicide use for weed control.

C. Chemical Weed Control

Herbicides are chemicals used to suppress or kill plants or to interrupt their normal growth processes. Of all groups of pesticides (including insecticides, fungicides, and rodenticides), herbicides are the leading group in terms of tons produced, dollar value from sales, and total acreage treated. Extensive and widespread use of herbicides in agriculture continues because of their high level of effectiveness and low cost relative to other methods of weed control. Use of herbicides has resulted in improved control of weeds that grow within crop rows where cultivation is not possible. Herbicides have replaced frequent tillage operations in some systems, which conserves energy, reduces crop damage, and minimizes damage to soil structure. With herbicides, crop production is less dependent on weather and hu-
man labor, such that greater flexibility in choice of crops and management methods is possible. In the United States today, the abundance of relatively inexpensive food and fiber is due in large part to the benefits afforded by herbicides for weed control in the last half century. However, the use of herbicides also carries risks, including injury to crops and nontarget plants, herbicide residues in soil or water, toxicity to nontarget organisms, and concerns for human health and safety. For this reason, the benefits and risks of each method must be weighed carefully when developing a weed control program, particularly in wildland ecosystems. In the United States, all pesticide development and use is subject to strict regulation by the federal government.

1. Herbicides

There are approximately 1300 herbicide active ingredients, which are formulated into hundreds of commercial products. Most are organic compounds, containing carbon, hydrogen, oxygen, and various other chemical elements. Each herbicide has a chemical name that describes its structure and a common name, which is often a simplified version of the chemical name. Formulated herbicides also have a trade name assigned by the manufacturer for marketing purposes. Manufacturers formulate herbicides to enhance their handling and weed control properties. Formulated compounds include the herbicide active ingredient plus inert ingredients such as solvents, diluents, and various additives. When the same herbicide active ingredient is formulated in more than one way, each is assigned a different trade name. Herbicides can be classified in several different ways, which provide users with a convenient means of selecting herbicides for various purposes.

Herbicides are often classified according to similarities in chemical structure, which often, but not always, result in similar effects on plants. A more useful classification scheme is based on where they may be used. In agriculture, herbicides are registered for use in agronomic and horticultural crops, turfgrass, and landscape and ornamental plantings. In noncrop areas, herbicides are used in pastures and rangelands, aquatic habitats, rights-of-way, utility sites, recreation areas, forests, and wildlands. Herbicides are also classified according to the site of application, are called contact herbicides. Herbicides classified by mechanism of action include growth regulators, inhibitors of photosynthesis, pigments, lipid synthesis, cell wall synthesis, amino acid synthesis, and cell division, and cell membrane destroyers.

Other classification schemes are based on plant responses to herbicides. Selective herbicides are more toxic to some plant species than others (e.g., monocotyledonous versus dicotyledonous plants), whereas non-selective herbicides are toxic to all plant species. Selectivity is one of the most important and useful characteristics of herbicides, for it allows applications to be made to weeds without risk of injury to crops or desirable vegetation. Factors related to herbicide chemistry (structure, formulation), the plant (age, size, surface characteristics, morphology, physiology), and the environment (humidity, temperature, soil moisture) determine the selectivity of a particular herbicide. Other categories of plant response used to classify herbicides are the pathway of herbicide movement in plants and the mechanisms by which herbicides kill plants (see Table II). Systemic herbicides are those that move in plants, movement occurs in the phloem (symplast), in the xylem (apoplast), or both. Herbicides that do not move in plants, but rather exert their effect at the site of application, are called contact herbicides. Herbicides classified by mechanism of action include growth regulators, inhibitors of photosynthesis, pigments, lipid synthesis, cell wall synthesis, amino acid synthesis, and cell division, and cell membrane destroyers.

2. Fate of Herbicides in the Environment

Herbicide fate in the environment is an issue of public concern and an important consideration when herbicides are registered for legal use. For herbicides to be effective, they must persist long enough to kill the weeds for which they were intended. Persistence beyond that time, however, may result in injury to nontarget plants and other organisms, residues in crops, and environmental contamination. Herbicides that enter plants generally move to a site of action and cause a toxic reaction. Over time, most herbicides in plants are transformed into relatively less toxic forms by biochemical processes. If a herbicide is not degraded, it may remain in the plant or end up in the soil as a contaminant. Eventually, however, all herbicides that enter a plant, the soil, water, or atmosphere will be degraded by the same chemical and physical reactions that act on biologically derived compounds.

Once a herbicide reaches the soil, several processes, including adsorption to soil particles, movement to another location, and decomposition, will determine its persistence. Soil and herbicide characteristics regulate adsorption of herbicide molecules onto particles of clay and organic matter. Herbicides that are tightly adsorbed are not available for plant uptake, movement to other

Table II. Herbicide Classification Scheme

<table>
<thead>
<tr>
<th>Mechanism of Action</th>
<th>Typical Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact</td>
<td>Glyphosate</td>
</tr>
<tr>
<td>Systemic</td>
<td>2,4-D</td>
</tr>
</tbody>
</table>

*Key: Glyphosate is a common contact herbicide, while 2,4-D is a typical systemic herbicide.*
<table>
<thead>
<tr>
<th>Method of application</th>
<th>Movement in plants</th>
<th>Mode of action</th>
<th>Chemical class or individual herbicide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foliar</td>
<td>Translocated in phloem</td>
<td>Auxin-type growth regulators</td>
<td>Phenoxyl acid herbicides, benzoate acid herbicides, picolinc acid herbicides, napralam</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aromatic amino acid (EPSPS) inhibitors</td>
<td>Glyphosate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Branch-chain amino acid (ALS/AHAS) inhibitors</td>
<td>Sulfonyleurea herbicides, imidazolinone herbicides, triazolopyrimidine sulfonamide herbicides, pyrimidyl oxybenzonic herbicides</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Carotenoid pigment inhibitors</td>
<td>Amitrole, aminomethanesulfonate, isoxaflutole, norflurazon</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lipid biosynthesis (ACCase) inhibitors</td>
<td>Aryloxyphosphorylurea propionate herbicides, cyclohexanediones herbicides</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Organic arsenicals</td>
<td>Disma, MSMA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unclassified herbicides</td>
<td>Amid, difemquat, fosamine, propamochlor</td>
</tr>
<tr>
<td></td>
<td>Foliar</td>
<td>Photosystem II (PSII) inhibitors</td>
<td>Triazine herbicides, metribuzin, phenylurea herbicides, uracil herbicides</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Photosynthetic inhibitors</td>
<td>Benazon, bromoxynil, phenylcarbamate herbicides, pyrazon, pyridate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other photosynthetic inhibitors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Foliar</td>
<td>Photosystem I (PSI) cell membrane destroyers</td>
<td>Bipyridyl herbicides</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Photosynthetic inhibitors</td>
<td>Diphenylether herbicides, oxadiazole herbicides, flumifural, triazolone herbicides</td>
</tr>
<tr>
<td></td>
<td>Soil</td>
<td>Photosynthetic inhibitors</td>
<td>Glyfosanate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Microtubule/spindle apparatus inhibitors</td>
<td>Dinitroaniline herbicides, DCPA, dihydroxypropanamide, pronamid</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shoot inhibitors</td>
<td>Chloroacetamide herbicides, thioanilactone herbicides</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Miscellaneous cell division inhibitors</td>
<td>Benzinid, napropanamide, suluron</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cell wall formation inhibitors</td>
<td>Dihydrobenzaldehyde, isoxaben, quinipyrac</td>
</tr>
</tbody>
</table>

*Abbreviations: EPSPS, 3-enolpyruvylshikimate-3-phosphate synthase; ALS, acetolactate synthase; AHAS, acetohydroxy acid synthase; ACCase, acetyl-CoA carboxylase; DSMA, disodium methanearsonate; MSMA, monosodium methanearsonate; DCPA, dimethyl 2,3,5,6-tetrachloro-1,4-benzenedicarboxylate.

Source: Adapted from Ross and Lembi (1999).

Herbicide decomposition occurs in soil, air, water, plants, animals, and microorganisms and results in breakdown of the original herbicide molecule and loss of herbicide activity. Decomposition of herbicides occurs by photochemical (breakdown in sunlight), chemical, or microbiological means. Herbicide decomposition also depends on temperature and other environmental factors, plus the concentration of herbicide applied. Products of herbicide decomposition may eventually degrade into simple organic molecules. Because most herbicides are degradable, they do not build up in the soil over time, even after repeated
use. Additionally, the soil microbial populations are adaptable such that the application of a pesticide to soil is often followed by an increase in the number of microbes that can degrade it. When processes that regulate herbicide fate are understood, and the legal requirements for herbicide application and use are followed, minimal contamination of the environment should result.

3. Regulation of Herbicide Use

All pesticides, including herbicides, must be registered with the U.S. Environmental Protection Agency (EPA) before they can be distributed or sold in the United States. Two laws, the Federal Insecticide, Fungicide and Rodenticide Act (FIFRA), and parts of the Food, Drug, and Cosmetic Act (FDCA) regulate pesticide development and use. FIFRA provides for registration and cancellations of pesticides, maintains a classification system for pesticides based on toxicity, and allows states to regulate pesticide use in a manner consistent with federal regulations. The FDCA mandates establishment of tolerances for pesticides in food, feed, fiber, and water. These laws were written to ensure that benefits from the use of pesticides are in balance with concerns about health and environmental impacts. Each state also has laws regulating pesticide use, including worker safety regulations and requirements for use of the most toxic (restricted use) pesticides.

For a pesticide to be registered by the EPA, it must be subjected to over 100 safety and environmental tests. Information required before registration includes chemical and physical properties, environmental fate, amounts of the pesticide in feed and food crops, toxicological properties, and effects on nontarget plants and animals. Such data are usually required both for the pesticide and for its metabolites, or breakdown products. With this information, the relative benefits and risks of each pesticide can be determined.

Pesticide effects on human health are expressed as toxicity, the amount of the chemical that is harmful or lethal, and exposure, the probability of encountering a harmful dose of the chemical. The combination of acute toxicity plus exposure to a pesticide during its expected use determines the hazard it poses to humans. The EPA uses these data to set a tolerance level for each agricultural pesticide, which is the maximum amount of the chemical allowed on a particular crop. These and many other data, including chronic toxicity, reproductive effects, teratogenicity, and carcinogenicity, are used by the EPA in setting limits for pesticide use to ensure that hazard to humans from use of pesticides will be at acceptably low levels. Toxicology and exposure studies are also required on certain species of wildlife, including birds, fish, and invertebrates, before pesticides can be registered. These tests utilize studies of pesticide residues in foods that these species may consume, as well as potential concentrations of pesticide in water or air to determine hazard to nontarget species from pesticide use.

Pesticides vary widely in their toxicological properties. In relation to all pesticides, which includes insecticides, fungicides, and rodenticides, most herbicides are relatively nontoxic to mammals, since many of the processes or pathways they inhibit in plants are not present in mammalian systems. For the purposes of regulation, pesticides are classified according to their toxicity, which is a relative term used to describe the amount of a chemical that causes harm to a particular species. The most common unit of measurement for toxicity is the lethal dose, \(LD_{50}\) or lethal concentration, \(LC_{50}\), which is the dose or concentration that kills 50% of the test population, respectively. High values for \(LD_{50}\) or \(LC_{50}\) indicate lower toxicity, as higher doses are required to produce lethal effects. The categories of acute toxicity that must be shown on herbicide labels are listed in Table III. Because all herbicides are toxic to some degree, container labels are required to give specific directions for use as well as ingredients, properties, hazards, exposure limits, first aid procedures, and other information. When handled according to the label directions, hazard to humans and wildlife from exposure to herbicides can be avoided.

### III. EFFECTS OF HERBICIDES ON BIODIVERSITY

#### A. Weeds

All weed control practices exert selection pressure on weeds and thus can have short- and long-term effects.

<table>
<thead>
<tr>
<th>Toxicity category</th>
<th>Toxicity</th>
<th>Signal word</th>
<th>Oral (LD_{50}) (mg/kg)</th>
<th>Dermal (LD_{50}) (mg/kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Very high</td>
<td>Danger</td>
<td>0.00–30</td>
<td>0.00–20</td>
</tr>
<tr>
<td>II</td>
<td>High</td>
<td>Warning</td>
<td>31–300</td>
<td>201–2000</td>
</tr>
<tr>
<td>III</td>
<td>Moderate</td>
<td>Caution</td>
<td>501–5000</td>
<td>2001–20,000</td>
</tr>
<tr>
<td>IV</td>
<td>Low</td>
<td>Caution</td>
<td>&gt;5000</td>
<td>&gt;20,000</td>
</tr>
</tbody>
</table>
on the composition, structure, and dynamics of weed communities. Agricultural weed communities tend to have lower species diversity than natural plant communities and are often dominated by a few key species. The primary short-term effect of weed control is a reduction in weed density, particularly of the dominant species, which is desirable in order to improve crop yields or facilitate land use. Over a longer time frame, weed control practices rarely eliminate weeds altogether, rather they generally result in changes in species composition and structure of weed communities. In the case of herbicides that act on specific plant processes, selection pressure over time can eliminate susceptible genotypes and thus cause evolutionary changes in weed populations. Most research in weed science to date has emphasized reducing weed density and improving crop yields. Only recently has attention been focused on changes in weed community dynamics as a result of weed control practices. Thus, only a few generalities can be made about specific effects or directions of change in weed diversity caused by biological, cultural, and mechanical methods of control. The most information available on the role of weed control in shaping weed communities comes from documented cases of herbicide-resistant weeds, which have been selected by repeated use of the same herbicide or herbicide class. It is clear that herbicides are a powerful evolutionary force acting on weed communities. Nevertheless, the data that are available indicate that effects of nonchemical forms of weed control on weed community dynamics are also significant and warrant further study.

1. Effects of Weed Control Practices

Changes in weed community composition and structure due to various agricultural practices have been documented; however, assessments of the effects of these practices on weed species diversity have been made only rarely. In a review of integrated weed management, Clements et al. (1994) calculated diversity indices from an array of published data to compare the impacts of conventional and alternative weed management practices on weed species diversity. When compared to weed control by mechanical cultivation only, broadcast applications of herbicides resulted in lower weed species diversity over time. Where herbicide use was reduced by placement of applications only in bands over the crop row, higher diversity of weeds resulted than when broadcast herbicide applications were made. Although these findings suggest that use of herbicides reduces weed species diversity, actual case studies show that this generalization is too simplistic and that impacts on diversity depend on the persistence of the specific herbicide used. When different herbicides were evaluated, applications of preemergence herbicides (those having residual soil activity) reduced weed species diversity more than applications of postemergence herbicides (those with no residual activity). By exerting continuous selection pressure on susceptible species from the time of emergence, preemergence herbicides often reduce the richness and diversity of weed communities. In contrast, postemergence herbicides are present in the environment only after emergence and for a shorter time period, which may permit a more diverse weed flora to establish both before and after the disturbance of herbicide application is imposed. Even in cases where weed species diversity is relatively unaffected by herbicide applications, interspecific selectivity of many herbicides causes a shift within a weed community from species that are susceptible to species that are naturally more tolerant to the particular herbicide. A common example of this phenomenon is the shift in relative abundance from dicotyledonous (dicot) weed species to monocotyledonous (monocot, usually grass) weed species following repeated use of the herbicide 2,4-D for control of dicot weeds in cereal (grass) crops. Similarly, in fields where herbicides have been used over many years for control of annual weed species, shifts in the weed flora to predominantly perennial weed species are commonly observed. Weed species compositional shifts also occur when the weeds in a field are taxonomically related to the crop species grown there, since plants often respond similarly to herbicides when they are in the same taxonomic family. Thus, many cases have been documented where repeated use of a particular herbicide in a crop has selected for weeds that are in the same plant family as the crop. Despite the increasing documentation of changes in composition of weed communities as a result of herbicide use, only recently has attention been focused specifically on weed species diversity.

Integrated and alternative weed management methods, which employ a combination of tools to control weeds below a specified threshold, theoretically should not impose strong directional selection on weed populations. Thus, weed species diversity might be expected to increase, or at least not decrease to the same extent as under a regime of chemical weed control, when a variety of tools are used. Clements et al. (1994) point out the many questions that remain to be answered about the potential effects of integrated weed management techniques on diversity of weeds. In addition, the role and potential importance of weed biodiversity in agroecosystems remain to be defined.
2. Selection of Herbicide Resistance

Herbicide resistance represents an extreme shift in weed species composition caused by the selection of plants possessing a gene or genes for resistance to a particular herbicide within a species that was formerly susceptible. In cases where a particular herbicide has been used repeatedly over several years, resistant weeds may be selected and come to dominate a weed community such that species diversity declines. Since the 1970s, many cases of the evolution of resistance have been documented as a result of repeated herbicide applications for weed control. Table IV summarizes the worldwide occurrence of resistant weed biotypes to different herbicide groups.

Several characteristics of herbicides and their use contribute to a high probability for selection of resistance in weeds. These include having a single target site and specific mechanism of action, being extremely active and effective in killing a wide range of weed species, and having long soil residual activity and season-long control of germinating weeds. In addition, frequent application of a particular herbicide over several growing seasons without rotating, alternating, or combining with other types of herbicides contributes to a high risk for evolution of resistance. Some herbicides are thought to pose a low risk for selection of resistance owing to their nonspecific mechanism of action and short or no soil residual activity. Even in these cases, however, repeated use of the same herbicide will exert selection pressure on weeds. Recommendations for preventing or managing herbicide-resistant weeds include practices such as rotating herbicides from different chemical classes to avoid imposing the same selection pressure over time and integrating a combination of weed control methods. The higher level of weed species diversity that presumably would result from these approaches should reduce the potential for propagation of herbicide-resistant genes in weed populations (Clements et al., 1994).

B. Other Organisms

By suppressing, removing, or destroying vegetation, weed control modifies the environment and habitat of other organisms. In agroecosystems, crop pests as well as beneficial organisms can be affected by weed removal since weeds can serve as host plants or food sources for many types of organisms, including insects, fungi, and nematodes. In some cases, a high diversity of weed species in an agricultural field has been shown to reduce the magnitude of insect attacks on crop plants because the weeds serve as alternate food for the insect pests or harbor beneficial organisms that feed on the pests. In other cases, however, increased weed species diversity results in increased insect problems in a crop field because the weeds provide a food source or habitat so that the insect pest can remain in the field even during periods when the crop is absent. In those cases, weed control to reduce the diversity of weeds in a particular field will decrease the incidence of insect pest damage to the crop. To the extent that herbicides affect weed biodiversity, therefore, their use will also indirectly affect insect organisms.

The response of fungal and nematode populations to weed species diversity has not been well studied. In general, weed control practices result in a cleaner crop field, which usually leads to fewer disease and nematode problems. However, it is also possible that large areas of crop monocultures with few weeds may be susceptible to widespread disease epidemics because of the lack of genetic diversity in response to the disease. To date, little information is available to indicate how weed biodiversity or use of herbicides in weed control influences

### Table IV

<table>
<thead>
<tr>
<th>Herbicide class</th>
<th>Dicot species</th>
<th>Monocot species</th>
<th>Total</th>
<th>Number of countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triazines</td>
<td>42</td>
<td>19</td>
<td>61</td>
<td>22</td>
</tr>
<tr>
<td>ALS inhibitors</td>
<td>43</td>
<td>20</td>
<td>63</td>
<td>18</td>
</tr>
<tr>
<td>Bipyridiliums</td>
<td>18</td>
<td>7</td>
<td>25</td>
<td>13</td>
</tr>
<tr>
<td>Ureas/amides</td>
<td>6</td>
<td>11</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>Synthetic auxins</td>
<td>15</td>
<td>4</td>
<td>19</td>
<td>12</td>
</tr>
<tr>
<td>ACCase inhibitors</td>
<td>0</td>
<td>21</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>Dinitroanilines</td>
<td>2</td>
<td>7</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Triazoles</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Chloroacetanilides</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Thioanilides</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Nitriles</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Glycines</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Benzilamides</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Organocarbamates</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Carboxylic acids</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pyrazolines</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Totals</td>
<td>129</td>
<td>105</td>
<td>234</td>
<td>232</td>
</tr>
</tbody>
</table>

* ALS, acetolactate synthase; ACCase, acetyl-CoA carboxylase.
populations of fungal pathogens or nematode pests of crop plants.

The question of whether weed biodiversity is an asset or a detriment to overall pest management in agriculture deserves serious attention by researchers. To answer this question, more information is needed on the effects of various weed control practices both on weed biodiversity and on nonweed organisms. With this information, the costs and benefits of weed control can be weighed against potential costs and benefits of alternative strategies to control other pests that are affected by weed control. The question is complicated by the fact that in large-scale, mechanized agricultural production, increases in weed species diversity complicate weed control efforts. However, this problem could be offset if there were benefits to be gained from maintaining increased genetic diversity of weeds in a field, such as greater buffering against selection for herbicide resistance.

In wildland situations where exotic weeds have replaced native vegetation, weed control is increasingly practiced to reduce weed invasions and restore the abundance and diversity of native plants. The indirect result of these weed control activities is often a restoration of habitat for nonplant organisms, such as birds and small mammals. In these situations, therefore, weed control, including use of herbicides, can result in increased diversity of other organisms that depend on native plant communities for habitat.

C. Wild Plants

Weeds are managed in many situations to restore or preserve the biodiversity of native plant species. As described in Randall (1996), the most common methods used for weed management in wildlands in the United States are manual and mechanical weed removal, prescribed fire, release of biological control agents, use of grazing animals, encouragement of native competitors, and judicious use of herbicides. Weed management strategies in wildlands differ from those in agroecosystems because wildland managers must promote or protect large numbers of plant and nonplant species rather than one or a few crop species. Thus, wildland weed management is generally approached with a desire to manage an entire plant community rather than to control a single weed species. Although the same methods can be used in agricultural and nonagricultural habitats, wildland managers must minimize negative impacts to a wider range of nontarget species than must agricultural land managers. Therefore, the methods used for wildland weed control are often labor-intensive and more environmentally conservative than in agroecosystems.

All weed control methods incur some risk to the environment in which they are used, which must be weighed against the risk of taking no action and allowing weeds to continue to spread. Mechanical weed control disturbs the soil, destroys vegetation, and leaves gaps that may be reinfested with weeds. Biological control agents may attack nontarget species or become adapted and spread in undesirable ways. Herbicides may unintentionally kill nontarget plant species, indirectly impacting habitat for other organisms, and may become environmental contaminants if not used properly. In most cases, the herbicides used in wildland situations are postemergence compounds with very short or no soil residual activity to minimize their effect on nontarget plant species. However, few empirical data exist on either intentional or unintentional effects of herbicides on native plant communities.

Rice et al. (1997) conducted a long-term field study on the effects of herbicides used for the control of an exotic species on the structure and species diversity of native plant communities. Herbicide treatments were highly effective in controlling the exotic weed (Centau-

Rice et al. (1997) concluded that periodic application of appropriate herbicides can be used to restore native plant communities from dominance by exotic species and to maintain their diversity as well. Transient reductions in diversity that may result from herbicide use are likely to be negligible when compared to the serious impacts on native communities created by exotic plant species.

As in agroecosystems, questions remain in natural systems about the role and management of weed abundance and diversity in wildland ecosystems. Where exotic weed invasions are widespread, weed removal alone may not result in establishment of native species without additional inputs such as revegetation with desir-
able species. In some areas, invasive exotic weeds have replaced native species and become an important source of food and cover for native birds and mammals, which must be considered before the weeds are removed. The use of herbicides in wildlands is often controversial because of the risks of environmental contamination, as well as the general public perception that herbicides are dangerous. Just as in agriculture, therefore, the costs and benefits of weed control in wildlands must be weighed before any method is chosen. In cases where herbicides are used, the determination is made that the risks of herbicide use are more than offset by the benefits gained in increased or restored diversity of native species due to control of exotic weeds.

IV. CONCLUSIONS

Herbicides have been employed for decades to control unwanted vegetation in agroecosystems, yet their impacts on weed biodiversity and the implications of those impacts for weed management have not been well studied except in the case of the evolution of herbicide resistance. Increasing emphasis over the past decade on the importance of maintaining diversity of all genetic resources has rarely been extended to agroecosystems, despite the likelihood that assessments of the nature and importance of weed biodiversity might reveal an important role for weeds in agroecosystem stability and sustainability. Based on experiences with herbicide use in agriculture, deliberate use of herbicides to control invasive weed species in wildlands has become increasingly common, yet the impacts of herbicides on the composition, structure, and dynamics of nontarget native plant communities are poorly understood and usually assumed to be negative. When placed in a comparative context, however, habitat degradation and destruction, particularly by competition with exotic species, is a much more pervasive threat to biodiversity of endangered plant species than is pollution of all forms, including agricultural pesticides (Wilcove et al., 1998). Thus, it is imperative that research be continued into the effects of all forms of weed control on biodiversity of both targets and nontarget species.

See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • ECOLGY OF AGRICULTURE • HERBACEOUS VEGETATION, SPECIES • RESISTANCE TO INSECTICIDE RESISTANCE • PESTICIDES, USE AND EFFECTS OF

**Bibliography**


HIGH-TEMPERATURE ECOSYSTEMS

Richard G. Wiegert
University of Georgia

I. Extreme Environments
II. Thermal Systems—General Characteristics and Definitions
III. Geothermally Heated Systems—Classification and Description
IV. Temperature Limits to Life
V. The Evolution of Thermophily
VI. The Biodiversity of Thermal Ecosystems
VII. Summary

GLOSSARY

acidophilic Organisms preferring or requiring a low pH environment.

chemoautotrophs Organisms able to synthesize organic compounds by the oxidation of energy-rich inorganic sources. Light is not used.

cyanobacteria A group of bacteria containing chlorophyll and capable of photosynthesis.

eucaryotes Organisms possessing a defined cell nucleus and nuclear membrane. Includes all organisms above the level of the procarycotic bacteria.

frustule The hard, silica-containing skeleton of diatoms (green algae).

geothermal heating Water heated at depth in the earth and released to the surface as thermal outflows.

heterotrophs Organisms dependent for their energy on organic compounds. If oxidized, the process is aerobic. When oxygen is absent, the process is anaerobic (fermentation).

metazoa Organisms whose individuals have more than one cell (multicellular).

photoautotrophs Organisms able to synthesize organic compounds from water and inorganic nutrients, using the energy in photons of light.

procaryotes Organisms without a defined cell nucleus or nuclear membrane.

protozoa A group of eucaryotic organisms usually classified as animals and regarded as primitive. But some major problems exist with some protozoa such as slime molds and euglenoid forms, which share many of the properties of plants.

solar heated Water raised to temperatures significantly exceeding the regional temperatures of lakes and streams. Many of the thermally tolerant and thermal-opportunist groups are thought to have evolved in shallow solar-heated ponds and water margins.

thermal opportunist Organisms that have developed life history characteristics that permit the exploitation of temporary cooler spots in thermal systems.

thermal systems Outflows of geothermally heated water, usually thought of in terms of those at the surface of the earth (see thermal vent) where light is present.

thermal vent Geothermally heated water issuing from cracks in the ocean floor. Very high temperatures
and pressures and the complete absence of light are characteristic.

thermally tolerant Organisms able to live at temperatures significantly higher than the regional norm. They can also exist at cooler temperatures, but may not compete well in the latter situation.

thermophily Often restricted to mean those organisms adapted to living at high temperatures and unable to survive at lower temperatures. In this chapter the term is also used in a more general way to characterize any type of adaptation to thermal systems.

THIS CHAPTER ON THERMAL BIODIVERSITY considers the number and kinds of organisms that are found in the thermal systems and thermal vent communities of the world.

I. EXTREME ENVIRONMENTS

The diversity of organisms in thermal environments is determined, in common with all biodiversity, by the interactions of the physical/biotic environment with the adaptive abilities of individual species populations. But in some respects thermal environments (ecosystems) are quite different, even unique, compared to ecosystems dominated by weather, geological origins of soils, and patterns of precipitation and drainage. First of all, thermal ecosystems form a class of extreme environments characterized by a mean temperature markedly higher than "normal." (Other extreme environments possess one more environmental variable far outside the "norm"—for example, arctic, desert, brine lakes, low pH systems, etc.) Some of these, such as the arctic/antarctic frozen wastes and extreme deserts, are caused by widespread climatic factors and are typically very large, whereas thermal ecosystems and some other extreme environments are small and result from local changes in the physical environment.

What is "normal" and what effect does the extreme factor or factors have on biodiversity? Here I define an extreme environment as "a condition under which some organisms can grow, whereas others cannot." This is too general for the purpose of this chapter, because each different ecosystem has its adapted flora and fauna. A more narrow definition identifies the extreme as excluding entire higher taxonomic groups, not simply species. Thus the extreme environment is not only low in the diversity of species, but also low in the diversity of higher orders of classification as well.

II. THERMAL SYSTEMS—GENERAL CHARACTERISTICS AND DEFINITIONS

Among extreme environments, thermal ecosystems are defined as those with an elevated temperature, compared with the range of other ecosystems found at the same latitude and elevation. The source of the heat may be (a) solar energy, (b) anthropogenic activities, or (c) geothermal. The organisms inhabiting these heated ecosystems can be classified as thermophilic, thermal tolerant, or thermally opportunistic. The first of these, thermophiles (the general term used in this chapter), require the high temperature for survival to which they are adapted optimally. Thermally tolerant organisms have evolved mechanisms to enable them to exist at extreme temperatures, but they can also survive at lower temperatures. Indeed, in culture they might do better physiologically at these lower temperatures. Competition with other nonthermal species may, however, restrict their distribution in nature. Thermal opportunists, on the other hand, have evolved mechanisms to avoid the effects of direct immersion in hot water. They seek out and take advantage of temporary refugia by using rapid maturation times, rapid mobility, and short life histories. I will later discuss the evolution of these traits and the groups to which they apply.

Solar heating of vascular plant leaves and of the soil/litter surface is enhanced by dark coloration, abundance of sunlight, and the slope of the land surface. Temperatures of plant leaves in desert environments may reach 60°C; the same temperature applies to soil/litter systems, but it is commonly much lower, and the soil litter surface cools rapidly at night and during rains. Furthermore, the soil/litter surface (as well as plant leaves) heated by the sun is very thin because of its insulating value, and the system is very dry when heated. Thus motile organisms can move out of these thermal systems during periods of inhospitable temperature regimes. Vascular plants can avoid the effects of heat by being rooted in the cooler subsurface soil, and nonmotile microorganisms resist the thermal effects by forming spores or other heat-resistant quiescent stages. The net result is that there is little effect of solar heating on the biodiversity of these ecosystems compared to adjacent systems with more shading of the soil surface. There may, of course, be a loss of species in the driest of these systems,
but as a consequence of lack of water, not of temperature per se. Solar heating may also act on very shallow aquatic systems; the margins of lakes and ponds in temperate and tropical latitudes are examples. This can provide a habitat for certain multicelled animals adapted to warm waters, but it does not exclude animals that simply avoid the shallows during the day and forage there at night. Temperature increases are seldom very great, and biodiversity in these systems is affected very little. A case involving a thermally adapted dragonfly will be discussed later in this chapter.

Anthropogenic activities include steam vents, hot water plumbing and heating systems, and fermentation. Where liquid water condenses around steam vents, certain thermal microorganisms can be found and some of them can also be found in hot water systems. But these are too small and transient to be of interest to the student of thermal biodiversity. Fermentation, however, is a large natural phenomenon, as well as the result of certain human activities, such as composting and piling up tailings from coal mines. As a natural process, the accumulation of fermenting organic matter is seldom large enough or accumulating fast enough for the heat of fermentation to significantly increase the temperature. Plant debris added to the soil litter system during the year is simply decomposed at a relatively constant rate and the heat produced is dissipated to the environment. Humans, however, often pile up organic wastes to a degree that the insulating value of the pile causes a rise in the internal temperature, killing, for example, the seeds of vascular plants in compost heaps and excluding all but microorganisms from the water draining from coal tailing piles. However, the latter effect is often caused more by the extremely low pH of such systems than from elevated temperatures. Although these environments are of little interest from the standpoint of biodiversity, they are of great interest to the student of microbial taxonomy. For example, an entire genus, *Thermoplasma*, is restricted to coal refuse piles.

III. GEOTHERMALLY HEATED SYSTEMS-CLASSIFICATION AND DESCRIPTION

The extreme ecosystems characterized by some form of geothermal heating of water are small, both in absolute size (some are only a few square meters or less in extent) and in relation to the surrounding regional climate and soil-dominated areas. Other extreme environments caused by aridity, extreme salinity, and low temperatures stress organisms by posing problems of getting or retaining water. In contrast, geothermal systems have water, indeed they are defined on the basis of hot water. The problems they pose for organisms attempting to survive and grow involve adaptations for tolerating or avoiding the hot water. Biotic diversity is typically low, both in absolute numbers of species and diversity of higher taxonomic groups. This diversity is also usually low relative to the environments surrounding the thermal systems, unless the latter are situated within a larger extreme environment. The deep-sea thermal vents are an example of the latter.

Most thermal ecosystems of the world are caused by the surface emergence of geothermally heated water. The mechanisms of heating involve the percolation of surface water to some depth, where it is heated and pushed again to the surface. The differences in thermal and nutrient characteristics of the emerging water depend on the substrate rocks it saturates after being heated and the length of time the water travels just under the surface before emerging in heated lakes, pools, streams, and spring seeps. Thermal systems are usefully catalogued into four types. These take the form of (a) heated pools or small lakes, (b) small streams, and (c) outflows from thermal springs. These first three types all exist under one atmosphere of pressure or less and the highest temperature found is 100°C, the boiling temperature at sea level. A fourth type (d) the thermal vent communities form when heated water is ejected at depth on the ocean floor, and temperatures in the hundreds of degrees centigrade are possible. These were first discovered in 1977; they are small, usually less than a few hundred square meters, and may vent water (black smokers) that reaches 450°C.

The remainder of this chapter focuses on the various manifestations of these heated ecosystems. All of them share the characteristic of being aquatic, but they show differences related to the volume and movement of water flowing out of the ground as well as the mechanisms available to organisms for adapting to the thermal environment. If temperature is regarded as analogous to a toxin, an organism can exist by (a) adapting to exclude the factor (in the case of temperature, this amounts to avoiding lethal levels for a particular stage in the life history), (b) adapting its structure/life history/physiology to nullify the effects of the factor, or (c) simply adapt in more minor ways to live with the factor. The second and third strategies differ quantitatively but not qualitatively. For example, enzymes can be changed to different temperature optima by relatively small evo-
tionary steps, but to cope with temperatures that will destroy (denature) the enzyme, a major new structure must be developed.

Geothermally heated lakes and pools characteristically are well mixed thermally, with little or no directional current. This produces, typically, a body of water with a relatively homogeneous temperature regime. The consequences for organisms attempting to colonize such ecosystems is that they cannot escape the thermal load, thus any organisms with limited motility must adapt to survival at the temperature regime of the system. Algal/bacterial mats in such thermal ecosystems, at temperatures where photosynthesis is possible (discussed later), form on the bottom of the lake or pool. There is little or no thermal gradient in these systems, so any eucaryotic consumers must adapt to the prevailing temperature regime or fail to colonize. Thus in Yellowstone National Park in the United States and in other geothermally active areas of the world, the thermal lakes and pools can be ranked from low to higher diversity as an inverse function of the mean maximum annual temperature of the system. Because of the lack of strong currents and replacement of surface water, cooling of these systems is relatively slow and the annual variation in temperature is low, even in the cold temperate latitude and high elevation of Yellowstone Park.

Thermal streams share with thermal lakes and pools the characteristic that the layer of productive organic photo- or chemoautotrophs is on the bottom and organisms attempting to colonize must adapt to tolerate the temperature regime. But the stream has a significant unidirectional flow and hot water emerging from the substrate onto the surface immediately begins to cool. The result is the establishment of a strong thermal gradient from the source downstream, within which different specific conditions for community development can be found. In theory, the cooling curve will be exponential, but other factors intervene, among which the most important are the current turbulence and the temperature, humidity, and the wind speed immediately above the surface of the stream. For example, in a small Yellowstone thermal stream (Gentian Stream, Firehole Lake Drive), I have found that the temperature in summer can fluctuate rapidly as the sun goes behind clouds and wind velocity changes. This effect takes place in minutes, whereas longer lasting weather changes (cloudy days, for example) cause changes in hours to days. In Yellowstone, 250 m from the source, the annual fluctuation in temperature is approximately 20°C, significant, but also a testimony to the thermal capacity of water since the flow of this small stream is measured in liters per minute. Thus in contrast to the organisms in thermal lakes and pools, the inhabitants of geothermally heated streams are far from living in a constant temperature natural chemostat, unless the stream habitat is very close to the source.

Outflows from thermal springs differ from lakes, pools, and streams in the volume of water discharged, its depth and pattern of flow after discharge, and in the types of communities and organisms that colonize the outflows. In contrast to lakes and pools, spring discharges have strong flows and turbulent mixing, with a relatively steep thermal gradient established from the source to the point where temperature of the water approximates that of ambient air. Outflows do not generally follow a well-defined channel (except near some of the sources), and in any case the depth is shallow. Once the temperature reaches the point that is tolerated by the filamentous cyanobacteria (blue green algae), a mat of microorganisms forms that can thicken and directly affect the flow of the water. This creates a community with distinct cool patches; some are cool on top with hot water flowing underneath, others are devoid of flow. This temperature heterogeneity provides an additional manner in which motile organisms can colonize thermal communities, by avoiding lethal temperatures as a consequence of adapting physiology, life history, or behavior to take advantage of the temporary cooling of the algal mat. Spring outflows exhibit considerable variety. The sources are of different temperatures, primarily a consequence of the water traveling for variable distances in the soil after emergence from the underlying rock and before emergence onto the surface. The chemical characteristics of the emerging waters may differ (although I am eliminating from this discussion those waters with abnormally low pH, caused by excessive sulfur content). Finally, spring outflows are often intermittent, creating yet another form of temperature heterogeneity. These intermittent flow communities vary greatly in their biotic composition, depending not only on the temperature gradient but also on the period and volume of flow, the chemical composition of the source water, and so on. The biodiversity found at temperatures less than 40°C is, however, low compared to the biodiversity of similar temperature zones in a relatively constant thermal gradient.

Thermal vent communities are formed by the emission of superheated water from fissures in the ocean floor that form along the line marking the meeting of tectonic plates. Because of the pressure of deep water, temperatures are possible that greatly exceed the 100°C temperatures of water boiling at sea level. Unfortunately, the great depths at which these vent communities form makes their study both difficult and extremely
expensive. In general outline, the water is colonized by strains of high temperature heterotrophic bacteria utilizing the organic compounds dissolved in the superheated water. A strong thermal gradient is set up where the heated water meets the cold seawater and several groups of filter-feeding marine invertebrates have evolved to utilize these resources. Whether this evolution involves tolerance to high temperatures, however, is problematic, since samples and temperatures are hard to match up under such difficult sampling conditions. Furthermore, the thermal gradient is so steep that small distances may see a radical change in mean temperature. Numbers of individuals and biomass of the invertebrate consumers are large, because of the richness of the production by heterotrophic microorganisms, but number of species is low, although perhaps similar or even higher than that of the cold, dark, relatively sterile community away from the thermal vents at these depths. Although the microorganisms are clearly living at temperatures far higher than any found at the surface of the earth, the question of whether multicelled eucaryotes have evolved to survive at temperatures above 40 to 50°C is still an open question.

IV. TEMPERATURE LIMITS TO LIFE

Extreme environments were characterized earlier by the absence of species belonging to large systematic groups. In thermal ecosystems, as noted earlier, not only are these large systematic groups not found, but within groups, species numbers are low, compared with cool water communities. In general, in terms of temperature tolerance, there is a progression from more "primitive" to higher taxonomic groups, and from small to large individual size. Later I consider the arguments for the evolution of the various temperature tolerance ranges. What are the observed limits, based on sampling and observation of thermal ecosystems?

Four different groups of organisms need to be considered: (a) the heterotrophic procaryotic bacteria, (b) the procaryotic photoautotrophs, (c) the eucaryotic microorganisms, both heterotrophic and photoautotrophic, and (d) the eucaryotic metazoans. Much of the literature prior to 1978 on these limits has been summarized. The study of photosynthetically active thermally tolerant organisms really began with the pioneering work of W. A. Setchell. The main body of his work was never published, but it exists as a 215-page manuscript in the archives of the University of California at Berkeley. A brief summary of the temperature limits portion of this work was published in Science. Setchell reported the upper temperature for "algae" as 75 to 77°C, and for bacteria to be 89°C. A later investigator claimed, on the basis of superficial samples in Yellowstone, that organisms could not grow above 73°C, but this work relied on the uptake of radioactive phosphorus as the indicator of life; what was very likely being measured was the upper temperature for photosynthesis rather than the upper temperature for life. In other older records of maximum temperatures, unfortunately, some observers confused blue-green algae, procaryotic cyanobacteria, with green algae, eucaryotes. Within the pro¬caryotes, there has also been confusion about cyanobacteria versus filamentous bacteria. Furthermore, it was not realized how steep the thermal gradient could be, thus putting a premium on temperature measurements at precisely the point where organisms are growing in the field. More recently, the discovery of ocean thermal vents (discussed earlier) has reopened some of the controversy regarding these limits.

Heterotrophic procaryotes have colonized habitats at all temperatures up to the boiling point of water (100°C) at sea level. The most thermophilic of these organisms have not been cultured, but they are easily sampled in the parent thermal communities using glass slides and simple photomicroscopy. On the basis of these findings, the prediction is that life is possible at any temperature at which there is liquid water. This prediction has now been verified by the preliminary explorations of the thermal vent communities on the ocean floor (discussed earlier) with microorganisms living at temperatures far in excess of 100°C. However, reports of organisms living in water at 300°C are still provoking argument. As noted earlier, the most extremely thermophilic procaryotes have not even been cultured. This may be in part due to insufficient knowledge about the nutritional requirements of these heterotrophs, but it might also be due to the difficulty of maintaining cultures near the boiling point of water. The problem is exacerbated in the case of thermal vent microorganisms growing at both extreme temperatures and pressures. Many of these species are also obligate anaerobes, for which even extremely low concentrations of oxygen are poisonous. Others are endosymbionts, which are also notoriously difficult to culture. Just getting the samples to the surface in a viable condition is a problem, and special culture techniques are required for the methanogens. In summary, thermophilic heterotrophic procaryotes are apparently not limited by temperature but by the presence of liquid water. At the surface this is approximately 100°C (depending on altitude), but at depth in the ocean life is found at substan-
tially higher temperatures, limited by suitable nutrients, pH, and energy sources.

The temperature limit for procaryotic photosynthetic autotrophs (both cyanobacteria and photosynthetic bacteria) is substantially below 100°C on the surface. All of the observations and experimental evidence to date suggests a maximum limit of 73 to 74°C. Of this group the most thermophilic are the photosynthetic procaryotes, of which the most temperature tolerant is the single-celled species of cyanobacterium (*Synechococcus lividus*) and the filamentous photosynthetic bacterium (*Chloroflexus aurantius*). These species are found in nature at temperatures up to 74°C, although the optimum temperature is 63 to 67°C for the cyanobacteria and even lower, about 55°C, for the filamentous bacterium. In general, the filamentous forms of the photosynthetic procaryotes seem to have adapted to much lower optimal temperatures than the single-celled forms such as *Synechococcus* sp.

Eucaryotic microorganisms are found in nature at substantially lower temperatures than are procaryotes. Different reactions and responses to thermal environments are found in (a) fungi, (b) eucaryotic algae, (c) protozoa, and (d) metazoans, including invertebrates, vascular plants, and, vertebrates.

**Fungi:** In general, thermophilic and thermotolerant fungi are found at temperatures lower than 60°C and occur at this temperature only in acid thermal waters. In the very common and widespread “alkaline” hot spring communities, where the pH is initially somewhat acid because of dissolved carbon dioxide but rapidly rises when exposed to the air, free-living filamentous fungi are absent, even at temperatures below 40°C. There are reports of fungi parasitic on cyanobacteria that would probably be found at somewhat higher temperatures, but these have not been found in the “alkaline” thermal effluents of Yellowstone Park and are, in any case, poorly known (although there is a remarkable thermal range of the disease-causing fungus, *Dactyliaria gallopava*).

Single-celled and filamentous green algae as well as diatoms become abundant at temperatures around 40°C and below. Earlier reports in the literature of diatoms growing at higher temperatures relied only on the recovery of the resistant siliceous frustules, without demonstrating that the cells were alive or growing. The maximum temperatures for which diatoms can be proved to be surviving and growing is 43 to 44°C. This particular diatom, *Achnanthes exigua*, has an optimum temperature of 40°C. The single-celled and filamentous green algae have very similar temperature limits to those of the diatoms. Below 40°C in the “alkaline” thermal outflows of Yellowstone, the diversity of algal species rises (discussed in a later section). Although some mats of filamentous green algae may be found in these communities on top of mats of filamentous bacteria and cyanobacteria, where the underlying water is much hotter than 40°C, this is because the green algae is growing on top of the mat and is not exposed to the higher temperature water. This underscores the care needed when measuring the temperature of the environment of suspected thermally tolerant or thermophilic microorganisms. Strong gradients can occur in millimeters between where temperatures are measured and where the organisms are found. An interesting and important exception to the restriction of the eucaryotic algae to temperatures in the range of 40°C is that of the acidophilic single-celled asexual green alga, *Cyanidium caldarium*. This organism has been studied intensively in Yellowstone Park acid hot spring effluents. The upper temperature limit is 55 to 57°C, and the organism can be found down to 35°C. The high temperature adaptation of this acidophilic eucaryote has been explained as a simple matter of no other competition, but the story may be more complicated than that (see next section on the evolution of thermophily).

Protozoa have not been given the attention of other eucaryotes in terms of their occurrence in thermal environments. Determining the temperature limits of this group is difficult because of erroneous reports in the older literature reporting the growth and survival of protozoa at high temperatures. Protozoa will grow readily at temperatures up to 45°C. Other studies have reported somewhat higher temperatures (57 to 58°C seems to be the maximum), but the majority of protozoa in the natural thermal systems are found at lower temperatures. I have not found living protozoa in the “alkaline” thermal outflows of Yellowstone at temperatures greater than 40 to 43°C.

Invertebrates are ubiquitous in and around thermal outflows throughout the world. In surface thermal features many are winged and thus widely distributed. The same genera and species are found in widely separated thermal ecosystems. Some species, such as the water mites, have a juvenile stage parasitic on winged adult insects and are also widely distributed. Those invertebrates that are neither winged nor parasitic are more local in the distribution of species. Most invertebrate inhabitants of thermal ecosystems are motile, thus enabling them to choose, within limits, the space and temperature optimal to growth and survival. Those that are not motile, or only minimally so, such as many of the deep-sea thermal vent invertebrates and the eggs and juveniles of many winged insects are subject to
mass mortality when major shifts in current and in the thermal gradient occur. To date, no invertebrates have been found surviving and growing at temperatures greater than the 50°C claimed for the ostracod, Potamocypris. The thermal limits of many of the most abundant and important invertebrate inhabitants of thermal ecosystems are in fact often quite low (well below 40°C). They exist by quickly colonizing temporary cool spots in the algal mats and growing into motile adults before the thermal environment becomes hot again. These are typical examples of thermal opportunists.

Vascular plants, being rooted in place, can colonize thermal waters at their maximum temperature only if the temperature is extremely stable. Thus there has been little study of their maximum tolerances. Species of Juncus grow in thermal streams in Yellowstone Park up to about 40°C and are often so abundant that they block stream flow and cause minor flooding of adjacent soil. Many species of vascular plant grow along the sides of the streams, one of the more conspicuous is the yellow monkey flower (Mimulus guttatus). It is difficult to say to what degree this plant is thermal tolerant because it does not grow directly in the hot water; its restriction in thermal basins to the sides of streams may simply reflect the availability of water, since the soil (sinter) beside thermal outflows is very dry. The species is found in wet swampy areas throughout Yellowstone Park and shows a wide range in size, being smallest in the hottest and least fertile environments.

Vertebrates, even more than vascular plants, are absent from thermal environments. One fish, the desert pupfish (Cyprinodon) is found in nonthermal (solar-heated) streams of Death Valley, California. Here it tolerates temperatures up to 43.5°C. But geothermal outflows are too hot, too small, too ephemeral, and vary too much in temperature and flow to support populations of even the small vertebrates. Reports of fish living in very hot water suffer from the same problem of most of the early observations of thermal maxima in organisms—namely, the means of temperature measurement were not sufficient to measure small gradients. Often the natural gradient of the water was such that the fish swam quickly from cool water into warm water to feed and then retreated to their cool refuges. For example, in Yellowstone Park, the many nutrients, particularly phosphorus, added to large, cold water rivers such as the Firehole, create exceptionally fertile and productive rivers. But these inflows also raise the temperature to the point that in summer they create a naturally thermally polluted river. In summer the lower Firehole river often increases in temperature far above the optimum for trout. For example, above the geyser basins in the summer of 1952, the average temperature was 10.2°C, whereas below the geyser outflows the temperature was 18.7°C, despite the influx of cold water from nonthermal springs and streams. In September of 1962 the difference was even more marked, 10.6°C above and 20.0°C below. That fish are able to not only survive but prosper in this situation is due to the many cold spring and streamside inflows entering the river, where the trout spend the majority of their time, venturing into the warmer currents of the river only to feed for short periods.

V. THE EVOLUTION OF THERMOPHILY

Thermophily in general must be thought of as the adaptation of organisms to survive and grow at temperatures (water as used in this chapter) significantly higher than the range of temperature in nonheated waters of the geographic region or climate. I have defined the range of thermal heating and the different types of ecosystem developing from the heated outflows. The types of organisms and their maximal and, where available, their optimum temperatures were discussed earlier. In general, the number of species and the diversity of the higher taxonomic classifications is lowest at the highest temperatures and increases with temperature decrease. The absolute upper temperature for life seems to be set only by the temperature at which liquid water is present. But if heterotrophic procaryotes can evolve to live and grow at these highest temperatures, why do not the other microorganisms and the metazoans also evolve to exploit higher temperatures? In one of the first (and the most detailed) discussions of the evolution of thermophily, the early arguments about this issue are reviewed and new hypotheses presented. Prior to the mid-1970s, the question of thermophily was regarded primarily as a physiological question, not an ecological one. This despite a wealth of evidence from other types of extreme environments that showed that adaptation to the extreme characteristic did not only have to be physiological, but that competition with other organisms was an important factor.

Several earlier authors regarded the increasing temperature tolerance of metazoa to eucaryotic algae to cyanobacteria (plus fungi) to heterotrophic bacteria as being due primarily to the increasing complexity of the cellular structure. In the fungal eucaryotes and for the eucaryotic green algae, Cyanidium, both of which reach 60°C as the upper maximum, the situation is complicated by the fact that they grow only in low pH hot systems. For the most part the arguments are simple
statements that the cell structure has some fundamental property that cannot be overcome by further adaptation. This postulates not only that the fundamental limit exists, but that there are positive selective pressures for the further adaptation to high temperatures. To date there is little evidence for the idea that temperature tolerance is a simple function of cell complexity, nor has the locus of the effect been identified. At first it was thought that the deep sea invertebrates were extraordinarily tolerant of high temperature, and this might have stimulated research into the cellular question. However, recent evidence points to rather low temperature tolerances in these groups. Clearly, no animals have been found living in the “black smokers” where temperatures of 400°C or more are common. White smokers at about 50°C may have some fauna, but the majority of the vent fauna is distributed around the vents, in much cooler water at 40°C or less. Since these temperatures are far above the ambient temperatures of the deep sea, the fauna is, in the sense defined earlier, thermophilic, but will do little to settle the question of why eucaryotes have not evolved to tolerate temperatures endured by thermophilic procaryotes. The essential test of the cell complexity-membrane postulate was to separate the question into: (a) the evolutionary advantage to be achieved above a maximum limit of 62°C and (b) the evidence that the various eucaryotic genomes have the capacity to respond to the selective pressures if they exist.

Following the preceding argument, the reasoning is that adaptation of eucaryotes to high temperatures could be prevented by some inherent physicochemical limitation or by the lack of benefit in additional resources gained. Two tactics are possible: (a) to shift the range of temperature tolerance at the upper level (the maximum) independently of the lower range and (b) to shift the entire range of temperature tolerance upward. Thermal outflows vary considerably in width, depth, and flow volumes. But the nonlinear cooling process (more rapid at higher temperature differences between water and air/substrate) ensures that the total area of colonization space, and thus of resources (whether solar energy or dissolved nutrients), will become steadily smaller as one proceeds, for example, in 5°C increments, down the thermal gradient. Thus organisms that can adapt to higher temperatures by moving only the upper tolerance limit, keeping efficiencies at lower temperatures constant, will always have a positive selective pressure for doing so. If, however, the range must remain the same, shifting it upward will result in an overall loss of resource space. In the latter case, there will never be a positive advantage to be accrued in shifting the range unless other factors are involved, such as predation, competition, or the thermal range of a required food species. Therefore, any general explanation of the evolution of temperature tolerance by thermophiles must consider not only a possible physicochemical limitation, but also whether resources are present and the relative amounts of resources at different temperatures. The first of these parameters, physicochemical limits due to cellular differences, has been the focus of most of the early literature. The presence of abundant resources for the various forms of eucaryotic organisms found in and around thermal ecosystems has been well established, to the point where it cannot be invoked as a sole or even a major block to the adaptation to higher temperatures. The invertebrate grazers of the algal mat are restricted to lower temperatures than their food organisms. In the case of photosynthetic eucaryotes, there is the same amount of solar energy per square meter at high temperatures and nutrients are at least as abundant as at low temperatures, if not more so. Thus the question devolves to physicochemical limits or amount of space and therefore of resources. To examine the effect of resource availability (read area) the two tactics discussed above are used. Only populations that are genetically homogeneous or whose genotypes are distributed by random mating fit the criteria. Asexual procaryotes do not fit because changes in genotypic abundance are the result of rates of growth and death. New genotypes arise by mutation and some exchange of genetic material, but the extent of the latter was not known in the 1970s and to my knowledge is not known today in thermophiles.

Because water begins to cool once it emerges from depth into the open air, surface thermal systems typically have the highest temperature water in spring-fed pools with relatively little current and little day-to-day or even annual variation in temperature. At temperatures above the (approximate) 75°C limit for photosynthetic autotrophs, these pools are dominated by monotypic or very low diversity heterotrophic procaryotes and the pools, although small in absolute area, are virtually infinite in size compared to the size of the bacteria colonizing them. These bacteria are often attached rods or sometimes filaments. Hyperthermophilic bacteria also occur in those thermal outflows where temperatures are 80°C or above. Because of the current, these forms must be attached rods or filaments. Furthermore, the colonies in the hot outflow channels decrease rapidly downstream. Whether this is because of the decrease in downstream temperature or the development of nutrient shadows is still an open question. Experiments in the field made by enriching these channels
with nutrients suggest that the latter may be a more reasonable explanation than the former.

The dominant organism in the high temperature outflows, Thermus aquaticus, is an obligate aerobe that forms attached filamentous colonies at temperatures up to 80°C. Discovered and named in 1969, it now appears to be a complex of several different bacterial types. If the attached bacteria in very hot outflows do show changes in growth and viability with temperature, it might appear as a broadened range of temperature tolerance. However, this could also be explained as the successive downstream colonization of different genotypes, each adapted to a different optimal temperature. Alternatively, if competition for a scarce nutrient presents a major restriction on distribution, low temperature tolerance will be sacrificed to gain higher temperature tolerance. In this case successive adaptations to higher temperature result in gain in resources providing the new genotype can maintain a sufficient level of efficiency to gain an advantage over its lower temperature-adapted competitor. Since I have suggested that the apparent temperature response could, in fact, be a response to nutrient depletion, the whole question is still an open one. These colonial forms can be cultured in the laboratory and also can be easily used in field experiments. Some efforts to solve the question of nutrient versus temperature would be worthwhile. In contrast, the extreme thermophilic bacteria that exist in the source pools at temperatures up to the boiling point are very refractory to culture. Thus studying their genetics and measuring their response to temperature clines has been impossible.

In the procaryotic photosynthetic microorganism, the evolution of temperature tolerance was operating in the same way, that is, the adaptation of successive genotypes to higher and higher temperatures, where there is a gain in resources, should be observed. Instead, we observe the complete failure of these organisms to colonize water at temperatures greater than 73 to 75°C (see discussion on limits presented earlier). The two major resources of the photosynthetic cyanobacteria and the flexibacteria are sunlight and a carbon source (phosphate is abundant in these systems and the organisms fix nitrogen). Cyanobacteria in particular respond positively to increases in free carbon dioxide. Sunlight is intense at the Yellowstone altitude in all shallow-water systems, obviously independent of temperature. Sunlight only becomes limited deep in thick algal mats, thus expanding the thermal range upward by the development of higher-temperature clones will neither increase or decrease the availability of this factor to thin new mats of cyanobacteria. In the “alkaline” thermal flows however, the water typically emerges at a pH in the 6 to 6.5 range due entirely to large concentrations of dissolved carbon dioxide. Therefore, any organism capable of forming genetically homogeneous clones should be able to adapt to higher and higher temperatures by gaining resources (carbon dioxide) at hotter temperatures. Such clones are known for both of the most common genera of cyanobacteria in thermal springs, Mastigocladus and Synechococcus.

The best evidence for an inherent physicochemical limit on the temperature of photosynthesis would be if the previous conditions were met with respect to resources. But adaptation proved unable to extend the range of temperature tolerance without a significant decrease in the ability or efficiency to use the increased abundance of the resource at higher and higher temperatures. This is exactly the case for Synechococcus. When growth (doubling time) is measured as a function of temperature, it is five times longer at 70°C than at 40°C. Thus adaptation to higher temperatures results in a rapidly decreasing ability to process a limiting nutrient, even though it (presumably carbon dioxide) is more concentrated at the higher temperatures. Thus the failure to evolve tolerances above 73 to 75°C must be due to a temperature dependent reduction in growth rate or a physicochemical failure in the algal cell.

Eucaryotic microorganisms, including single-celled and filamentous forms plus diatoms and free-living fungi, are seldom found at temperatures above the range of 40 to 45°C, with the exception of some fungi and the acidophilic genus Cyanidium. Collections made from six different hot springs in Montana showed a clear and significant separation between the eucaryotic algae and the cyanobacteria; the former were never collected at temperatures greater than 50°C and the latter showed more species per sample (2 to 5 at 50 to 58°C, 10 at 35°C) than the former. But the diversity of the eucaryotic algae was increasing much faster than the cyanobacteria with decreasing temperature. This separation was present in the thermal springs of both Yellowstone and Mount Rainier National Parks. The eucaryotic alga, Cyanidium, which exists at much higher temperatures than the other eucaryotic algae, may be subject to quite different selection pressures since it also occupies a variety of adjacent habitats (wet soil) where the temperatures fluctuate greatly, as do the temperatures in the outflows, which are very cool during snowmelt in the spring and the algae disappears, only to reappear in the water when the temperature rises. The argument is made that the evolution of thermophily in this group must take into account these differences plus the fact that hot acid flows have no procaryotic
photoautotrophs that could offer competition to *Cyanidium*. The possibility of genetic mixing due to sexual selection in the eucaryotic algae may, together with the increasing area for colonization as the thermal gradient is descended, render the selective optimum temperature lower in eucaryotes than in sets of competing procaryotic clones. This idea needs testing along with the notion of a major physicochemical limit on the eucaryotic cell. The growth of *Cyanidium* at higher temperatures certainly casts some doubt on the latter as a general explanation for the failure of most eucaryotic algae to adapt to temperatures higher than 50°C.

*Metazoans* typically have the lowest maximum thermal tolerances of any group of organisms. Here again, the arguments prevailing before the mid-1970s detailing the failure to adapt to higher temperatures invoked some vaguely defined physicochemical limit. In the case of the metazoan, complexity of organization is added to the supposedly inherent limits of the eucaryotic cell. But, as pointed out for other groups, this is only one of the three lines of evidence that must be considered when pondering the origins of thermophily. The other two are the availability of resources, both space and nutrients, and the interactions of resource abundance and evolutionary mechanisms that act to determine fitness. In other words, whatever the physicochemical limit to the evolution of thermal tolerance, simply measuring maximum tolerance rates in the field does not provide evidence, necessarily, that the physicochemical limit has been reached. Clearly, the larger aquatic metazoans could not sustain viable population densities in the small thermal systems, so there would be no selective pressure for them to colonize even moderate to low temperature thermal systems. Thus no fish are found in geothermally heated waters. The few cases of the evolution of tolerance to elevated temperatures in fish involve small individual size and solar-heated habitats. Here the maximum temperatures tolerated somewhat exceed 43°C.

The *metazoans* characteristic of hot springs are the arachnids (water mites), insects (mostly flies), small mollusks (small pulmonate snails), odonates (dragon flies), and, occasionally, ostracods. These groups are neither characteristic of the cold water streams into which the thermal effluents drain, nor of running water systems in general. Rather, they are related to (sometimes the same species as) the invertebrates found in the shallow muddy margins of lakes and ponds, where ambient daily maximum temperatures, due to solar heating, are of the same order as those in which the thermal inhabitants are found. If these low to moderate temperatures are not set by a physicochemical threshold, then the problem is to explain why, in thermal habitats, the species have not evolved more thermal tolerance.

The second factor affecting distribution (mentioned earlier), competition for scarce resources, seems not to be a factor. In every case studied, the grazers on the algae of thermal ecosystems are less tolerant of high temperature than the cyanobacteria on which they feed. Examples are the ostracods in Hunter Hot Springs, Oregon, and the ephyrid flies of the “alkaline” thermal flows in Yellowstone Park. This leaves the interaction of resources and selection determining fitness. Under random sexual mating (with respect to temperature), the direction of the evolution of temperature tolerance will be determined by the relative numbers of surviving offspring from the upper half of the temperature range relative to those from the lower half. But because of the nonlinear cooling curve of water, the size of the upper temperature part of the range will always be smaller than that of the lower half. Thus under a tactic of shifting range, thermophilily cannot evolve unless there is nonrandom mating or fecundity or survival is lower at low temperatures. The randomness of mating with respect to temperature differs in the various groups of hot springs invertebrates, depending on life history and behavior. In the flies and dragonflies, mating occurs after emergence and the animals are highly nuptial, so randomness with respect to temperature of development is assured. The water mites are also highly dispersed as parasitic larvae, but they do spend their adult lives within a relative narrow temperature range and mate there. The smaller mite, *Partuaniella*, tends to spend most of its adult life in cooler patches of the cyanobacteria/flexibacteria mat where it feeds on the eggs of the ephyrid flies and, as expected, has a rather low temperature tolerance. The large predaceous mite, *Thermacarus*, lives directly in the hot water and has an upper temperature tolerance 5 to 10°C higher than the smaller species of mite and just lower than the most temperature tolerant metazoan, the ostracod. This group has no widespread dispersal mechanism, thus nonrandom mating is predicted. They also seem to be limited by food. Not surprisingly, these organisms have evolved the highest tolerance to high temperature of any metazoan.

In summary, current evidence supports the following statements concerning the evolution of thermophily: (a) the heterotrophic procaryotic bacteria seem not to have any physicochemical limit, having evolved to inhabit geothermal systems of any temperature as long as liquid water is present; (b) the photosynthetic cyanobacteria seem to have evolved to the point
VI. THE BIODIVERSITY OF THERMAL ECOSYSTEMS

Species in, using, or around (living from) geothermal ecosystems throughout the world have been the subject of many studies. The earliest intensive work was on the fauna of the hot springs of Iceland. The algae of several hot springs in western India were cataloged in the 1960s. The algae and fauna of warm springs in New Zealand have also been studied, as has the fauna of thermal springs in the Dutch East Indies. The thermal effluents of Yellowstone National Park in the United States have been the subject of both the largest number and most intensive studies of thermophiles of any area in the world. The original studies concentrated on the fauna but also included some observations of the algae (cyanobacteria). Beginning in 1964, detailed studies of the microorganisms living in the hot spring waters of Yellowstone Park were begun. These became both the most extensive survey and intensive experimental protocol yet done on thermal microbiology. At first the focus was on high temperature systems where only one or a few species were found in any given thermal ecosystem. In 1967, I and my colleagues and students began studying the lower temperature thermal ecosystems with higher diversities of microorganisms and a richer food chain involving species of animal acting as scavengers, predators, and parasites. Additional unpublished information has been obtained from 1983 to the present on the thermal ecosystems developing at temperatures below 45°C. Thus the ecology of the Yellowstone ecosystems can begin to be understood from the standpoints of species diversity and food-chain relationships. In this final section I review what is known about the diversity of these systems, in their various manifestations of temperature and pH, and compare the results with the thermal biodiversity of the deep sea thermal vents, insofar as the latter are currently understood.

High temperature is the most difficult of extreme environmental conditions to which organisms can attempt to adapt. There are many ways to adapt to exist in extreme cold or aridity and resume growth and activity during short seasonal changes in solar warming and precipitation. High salinity can be avoided or salt excreted so extremes can still support a reasonably high diversity. Indeed, many saline ecosystems may appear simple from the standpoint of vascular plant species and the food chains they support, but be relatively high in the diversity of algae and bacteria in the surface sediments. Thus, in every thermal system investigated to date, the biodiversity decreases rapidly with increases in temperature. At their most diverse, thermal system species are measured in the hundreds; their least diverse manifestations are monospecific.

Temperatures near the boiling point of water in Yellowstone thermal systems and elsewhere (65°C–100°C) are most common in source pools fed by deep subterranean springs of hot water. These pools vary widely in chemical characteristics, depending on the mineral content of the rocks through which the superheated water passes on its way to the surface. All such pools have a large volume of heated water with a relatively small inflow/outflow. Currents are small; almost all mixing is by convection or mild bubbling of various gases, often carbon dioxide. Such systems are natural analogues of chemostats, and they are eventually colonized by adapted heterotrophic bacteria (photosynthesis being impossible at these high temperatures). Not surprisingly, competition generally ensures that they are monospecific, or, if chemoautotrophy is possible, the diversity might rise by one or two additional species. Although thermal source pools may be thought small relative to surrounding mountains, forests, and so on, from the viewpoint of microorganisms, these are virtually constant environmental systems capable of supporting immense numbers of individuals. Current ecological theory supports the prediction of dominance by one or a few species in these circumstances. In contrast, the emergence of superheated water from the thermal vents in the deep sea cannot reproduce these constant conditions. Pools of hot water do not exist, but instead a turbulent mixture of water is emitted at temperatures apparently as high as 350°C or more, directly into cold ocean water near 2°C. Thus environmental variation in temperature is undoubtedly high and the higher diversity of microorganisms would be predicted. It is impossible, under such variability in temperature, for one species to monopolize the resources of the system. Thus the students of thermal vent microbiology are finding a number of thermophilic species new to science (although no novel genera have been found). The latter may be due to the culture techniques used, which have
not been designed for organisms adapted to high pressures as well as high temperatures.

Food chains in the deep sea vent communities are simple and short, because in the absence of light no photosynthesis is possible. Thus the vents are surrounded by a diverse array of marine invertebrates that ingest the cells and products from the thermal vents. Almost all of these animals are novel species. As many as 236 species of animals have been collected around thermal vents, of which 223 are new to science. This diversity of animal species is far higher than that of any single category of surface thermal ecosystems, although the data represent many different thermal vent locations. If the worldwide diversity of animal species for all known surface thermal features is used, the animal diversity per system is similar in the two types of system. Thus the thermal vent communities of the deep sea, although of great interest from the standpoint of physiological adaptations, mass mortality, and colonization strategy, appear to offer little that is new to ecological theory.

At temperatures below 75°C and above about 40°C, thermal outflows form small streams of hot water. These are colonized throughout the world by unicellular photoautotrophic cyanobacteria and filamentous bacteria. These biotic communities are characteristic of all surface thermal features throughout the world, but most early studies were simply endeavoring to determine maximum temperatures. The first rigorous studies of these systems were initiated in Yellowstone Park. In each ecosystem, there are generally two dominant species or species groups at the upper end of this temperature range (a base mat of the filamentous photosynthetic bacterium Chloroflexus, and an overlying film of the unicellular cyanobacterium, Synchococcus). As the temperature decreases, the overlying mat becomes dominated by one or more genera of filamentous cyanobacteria. In relatively constant flows, temperature-adapted strains of Mastigocladus are the common dominants; in springs with variable outflows, diversity is often higher due to the presence of additional genera of filamentous cyanobacteria, commonly Calothrix or Phormidium. When these filamentous forms are able to grow, the outflow begins to go through a definite cycle, because the filamentous mat is not limited to a thin film as are the unicellular forms. As the mat grows, it thickens unevenly; water flows around these thickened areas, producing cool patches. These patches are populated, almost immediately, by one or more species of brine fly (Ephydridae). Upon hatching, the fly larvae begin to eat and destroy the filamentous integrity of the mat. The pupae metamorphose to adults within a few days. Eventually hot water reenters the cool patch, remaining larvae and pupae are killed, and the cycle starts over in another part of the mat. These flies and their predators are not thermophilic, and in fact are rather thermally intolerant, relative to organisms living in other thermal ecosystems—compost piles, for example. They are an example of what I have termed thermal opportunists, organisms that have adapted their life cycle to take advantage of opportunities provided within the thermal ecosystem. The diversity of the algal filamentous mat-fly association is not high for most expressions of this community of 20 to 30 species. Yet a diverse set of food chains and ecological processes is found, photosynthesis, grazing, decomposition, and predation, for example.

When a strongly channeled outflow cools down to 40°C or below, the diversity goes up rapidly. Metazoan animals can adapt to live directly in the warm water, and eucaryotic algae such as green algae and diatoms invade. Although cyanobacteria are still abundant, they are heavily grazed, and do not form a thick mat. In different expressions of this community are found crustacea, midges (Diptera), and water mites (Hydrachnellae), dominated by a top predator, the larva of the dragonfly, Erythmecis coifanata. The net effect on diversity is to increase the species numbers to 50 to 60, exclusive of the nonphotosynthetic microorganisms. The latter have not been studied in detail at these temperatures, thus a total diversity of 100 or more would not be unexpected.

VII. SUMMARY

Thermal waters, as extreme ecosystems, exhibit the low diversity expected under such conditions. But high temperature is a particularly difficult factor for adaptive evolution to overcome. Thus, thermal ecosystems appear to have the lowest diversities of any of the extreme environments. Nevertheless, there is considerable variation in diversity, depending on both the temperature of emerging water, the rate of cooling, and the stability of the emerging plume. Generally, there is an increase in diversity as the temperature declines. This can be modified by short-term fluctuations in temperature caused by currents set up in the emerging plume of hot water (in deep sea thermal vents) or by intermittent flows from the source creating cooler periods downstream. Both of these effects increase diversity. When animals are able to enter the system, a rich ecological food chain (in terms of numbers of ecological processes) often develops.
Acknowledgments

I am indebted for the information in this article to dozens of previous studies of the ecology of thermal organisms. My own work in Yellowstone National Park and elsewhere has benefited from permission from the park administration to do research. I have received financial support from the National Science Foundation through a series of grants, the latest of which, DEB-8121-0142, was used to investigate the community forming at temperatures below 40 to 45°C. I am also indebted to Drs. A. Chalmers and B. Hodson, who provided useful comments on the manuscript. Support for preparation of the manuscript was received as discretionary research funds from the University of Georgia.

See Also the Following Articles

PSYCHROPHILES • THERMOPHILES, ORIGIN OF • VENTS

Bibliography


I. Biologists and Biodiversity before 1986
II. Creation of the Term “Biodiversity”
III. Problems with Other Foci of Conservation Efforts
IV. What Is “Biodiversity”? 
V. Biologists and the Promotion of Biodiversity
VI. The Values of Biodiversity
VII. Historical Awareness of Biodiversity Redux

This article discusses how we have come to know and understand the subject of this encyclopedia. The term “biodiversity” was coined in 1986 by biologists who wished to express a complicated, scientific understanding of the natural world, and who wished to inspire a rapid, widespread effort to conserve the natural world. This article traces how biologists understand “biodiversity” and how they have attempted to raise awareness on its behalf.

I. Biologists and Biodiversity Before 1986

Since ancient times, scholars have simultaneously revered the natural world, attempted to discover in that world a natural order or impose rational order on that world, and sought to understand the place of humans in the cosmos based on what they read in the natural world.

Worster (1987), Bowler (1993), and other historians offer extensive treatments of pre-20th century attempts by biologists and their intellectual predecessors to understand, categorize, and philosophize the natural world. In the mid 18th century, for example, in order to fathom God's wisdom, Carolus Linnaeus classified the riot of life into a functional taxonomy, the Systema Natuur. In “The Oeconomy of Nature,” (1749) Linnaeus offered a protoecological treatise on how these life forms fit together in mutual, stable interdependence, with humans at a central nexus in the web. For Linnaeus and his contemporaries, God meant humans to use those species that we found valuable; the complexity and richness of the natural world meant that our use of nature could not upset what God had balanced so exquisitely.

A century later, Charles Darwin built on Linnaeus's and others' ecological and taxonomic groundwork and painstakingly laid out a theory that provided an historical, deterministic explanation for the relatedness of all God's creatures—even if his work made God seem a bit less relevant and humans a bit less central to the machinery of nature. Darwin saw all forms of life as human kin and believed a key to civilization's maturity was the ability to empathize with and respect our extended family. For both Linnaeus and Darwin, the diversity of earth's life forms was a phenomenon to be revered, contemplated, and explained. Neither envisioned the object of their study as a commodity that needed protection from human advances, although a contemporary of Darwin's, George Perkins Marsh
of Invasions by Animals and Plants (1958), he foretold that pesticide use threatened this diversity. Biologists came to understand and document that the grand panoply of life forms was threatened as the fabric of ecological connections was rent. This awareness led biologists to make leaders and laypersons aware, and they searched for ways to use their carefully tended scientific authority to make nonscientists care about and conserve the objects of their reverence. Such work carries perils: advocacy threatens to undermine the perception of value neutrality and objectivity that leads laypersons to listen to scientists in the first place.

In A Sand County Almanac II (1970) wildlife ecologist Aldo Leopold proposed a “land ethic” that we should follow if we are to revere nature’s diverse organisms and protect them to play the precise roles in nature’s scheme that ecological scientists were just beginning to elucidate: “If the biota, in the course of aeons, has built something we like but do not understand, then who but a fool would discard seemingly useless parts? To keep every cog and wheel is the first precaution of intelligent tinkering” (p. 190). Leopold asserted that all species are intertwined in complex interrelationships, and the diversity of organisms and their interrelationships are crucial for a stable, functioning planet. From the “is” of ecology, Leopold derived the “ought” that forms the basis of his land ethic: “A thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends otherwise” (p. 262).

British ecologist Charles Elton also viewed “ecological variety” as a threatened commodity. In The Ecology of Invasions by Animals and Plants (1938), he foretold the dangers humans face as anthropogenic introductions of nonendemic species sweep the globe and threaten the flora and fauna that had previously been left to evolve in splendid isolation. “It is not just nuclear bombs and wars that threaten us, though these rank very high on the list at the moment: there are other sorts of explosions, and this book is about ecological explosions” (p. 15). Like Elton, Rachel Carson (1962) warned that people could no longer take this variety for granted as benign backdrop to human affairs. Rather, ecological science taught that species diversity is essential for ecological and human health and that uncontrolled pesticide use threatened this diversity. Biologists like Leopold, Elton, and Carson still revered “natural variety,” still sought to reveal the secrets of nature’s order. But unlike pre-20th century natural philosophers, they desired that their readers come to view species diversity as a quantifiable, measurable entity that was inextricably bound up with their own well-being. Species diversity was now a threatened commodity, and humans who caused this threat were simultaneously threatened as that commodity diminished.

## II. Creation of the Term “Biodiversity”

Throughout the 1960s and 1970s, biologists fomented public alarm over the deteriorating environment. Prominent biologists (e.g., Ehrenfeld, 1981; Ehrlich and Ehrlich, 1983, Myers, 1979) helped raise awareness that diversity was a threatened commodity, and U.S. legislatures responded with laws designed to assuage the threat. Most notably, the 1973 Endangered Species Act recognized and sought to protect the paramount value of species diversity (Kohm, 1991).

Still, biologists were frustrated that efforts to protect diversity were not keeping pace with the furious rate of destruction. Walter G. Rosen, a biologist and senior program officer at the National Research Council (which advises the National Academy of Sciences [NAS]), brought together prominent scientists from the NAS with the clout of the Smithsonian Institution to host the National Forum on BioDiversity in 1986. At the National Forum, biologists and others concerned about imperiled diversity staged a consciousness-raising event that sought, and received, widespread attention from the public. Rosen coined the neologism “biodiversity” for the event as a convenient shorthand, a buzzword that would at once encapsulate biologists’ understanding of a chaotic, diminishing natural world, and would raise public awareness about threats to the natural world (Takacs, 1996). In BioDiversity (Wilson, 1988), the collection of essays that chronicled the National Forum, Paul Ehrlich, Daniel Janzen, Tom Cade, Lester Brown, Michael Soule, and other scientists declared the need to rouse public attention on behalf of biodiversity and exhorted their colleagues to adopt that mission.

As biologists promote the term and the complex worldview it represents, “biodiversity” has become a widespread conservation buzzword. Biologists write about it in scientific and popular presses, both exploring its complexity and advocating its protection. Environmental groups focus on it in fundraising efforts; confer-
ences convened in its name occur regularly. Laypersons have joined biologists in attempting to shape the planet's physical, political, and normative environments to make more room for biodiversity.

III. PROBLEMS WITH OTHER FOCI OF CONSERVATION EFFORTS

Why has "biodiversity" gained prominence as a conservation buzzword, and why have biologists speaking on its behalf had some success in shaping public opinion about threats to the natural world? Various scholars (Cronon, 1995; Evernden, 1992; Williams, 1980) note that nature is so all-encompassing that what one attributes to it may say more about the speaker than it does about the natural world. Previous attempts to preserve "nature" or "wilderness" are too vaguely defined or deemed elitist in some quarters. For example, Guba (1989) sees traditional advocacy for nature or wilderness preservation as setting the aesthetic desires of the rich against the needs of the poor who need land to survive. Guba's critique has, to some extent, been incorporated into the biodiversity preservation discourse. For example, as people came to appreciate the value of cultural diversity, they might also see diversity in all its forms as a normative good, particularly when efforts to preserve both may be mutually reinforcing (Nabhan, 1997). A document prepared for the 1992 United Nations Conference on Environment and Development notes that "cultural diversity is closely linked to biodiversity. Humanity's collective knowledge of biodiversity and its use and management rests in cultural diversity; conversely, conserving biodiversity often helps strengthen cultural integrity and values" (Reid et al, 1992, p. 23).

Prior to the advent of biodiversity, the most effective conservation efforts in the U.S. focused on endangered species. Biologists' foci changed because biodiversity represents a more sophisticated ecological worldview, and a more sophisticated view of what biologists want preserved and how they want it preserved. Biologists promoting biodiversity conservation also seek to circumvent certain problems arising from efforts to preserve endangered species. Even though human activities accelerate the rate of species extinction, some opponents of conservation argue that species extinction is a natural process that we should let proceed. Biologists have difficulties defining with precision what constitutes a species (or a subspecies, like the Northern Spotted Owl) in the first place.

We are not aware of the economic or ecological benefits some individual species confer, so it is difficult to argue for their conservation in some circles. Many of the insects, bacteria, plants, and other members of what E. O. Wilson (1987) calls "the little things that run the world" remain unidentified and unloved; inve rtebrates comprise only a tiny percentage of organisms protected by the U.S. Endangered Species Act. Land set aside to protect individual endangered species may prove insufficient if global warming induces species migration (Peters and Lovejoy, 1992). When we focus on species diversity, we sometimes ignore generic, population, community, or ecosystem diversity. Endangered species conservation proves nearly impossible in poorly explored areas, in oceans, and in nations without species checklists. Focus on species that are endangered can be a last minute emergency effort to save species that may no longer be playing functional ecological roles, and this strategy may be inferior to proactive efforts to preserve healthy populations in healthy ecosystems. Finally, some view the U.S. Endangered Species Act as a mixed blessing: its unyielding allegiance to species on the verge of extinction is also a political lighting rod that leaves little room for compromise and has set many citizens against conservation efforts (Mann and Plummer, 1993).

IV. WHAT IS "BIODIVERSITY"?

Elsewhere in this encyclopedia, you may read about definitions of biodiversity. In research I conducted (Talacs, 1996), I asked prominent biologists to define the term. Little, if anything, in the natural world is excluded from these definitions of biodiversity. Responses included that biodiversity is "the complete array of organisms, biologically mediated processes, and organically derived structures out there on the globe" (Jerry Franklin); "the whole package of genes, populations, species, and the cluster of interactions that they manifest" (Daniel Janzen); "the total number of genetic lineages on earth" (Thomas Eisner); "shorthand for all the richness of life" (Reed Noss); or "the sum total of plants, animals, fungi, and microorganisms in the world including their genetic diversity and the way in which they fit together into communities and ecosystems" (Peter Raven). According to Terry Erwin (1991, p. 3), it is "the sum of earth species including all their interactions and variations within their biotic and abiotic environment in both space and time."

So biodiversity represents a rich, complicated vision of life and a corresponding rich, complicated vision of
what biologists want to see preserved. Although few biologists wish to see any species slip into extinction, what they really want—and what biodiversity really represents—is preservation of ecosystems, of vast stretches of land where the evolutionary process may continue relatively unfettered.

But ecosystems are hard to delineate, and even mentioning “evolution” lights a powder keg in some quarters. Kellert (1986, 1996) notes that the public is most likely to rally around “cognitively meaningful” organisms. Habitat has no fur, and ecosystems lack big, expressive eyes that arouse public affection and therefore attention to conservation causes. With biodiversity, biologists can focus attention on the charismatic megafauna and the layperson finds endearing. These organisms are emotionally and ecologically appealing to the conservation minded: they often are at trophic levels that require large territories to survive. When we preserve the large areas of land they require, we also fortuitously protect a myriad of species and enable continued functioning of ecological and evolutionary processes that biologists value.

Conservation efforts on behalf of biodiversity enable multiple images and multiple strategies to protect and conserve the natural world; many of those strategies place more power in the hands of biologists to realize their conservation values. If definitions of biodiversity seem complex and all encompassing, that is part of why it has been successful as a conservation tool. At once it represents the complexity of biologists’ worldview, the whole span of what biologists wish to conserve, and in it each of us can see that part of nature we cherish.

V. BIologists AND THE PROMOTION OF BIODIVERSITY

When biologists speak about biodiversity, they simultaneously refer to small parts of the ecological world, the interrelatedness between those parts, the ecological processes that sustain those parts, and the evolutionary processes that gave rise to those parts. While appearing as a purely scientific, objective entity, biodiversity also encompasses political arguments on behalf of conservation and symbolizes much that biologists do not know about the natural world.

Biodiversity represents ecological complexity that, the more they study, the more biologists realize they do not understand. Biologists do not know, within an order of magnitude, how many species live on earth, so they cannot specify how many we are losing and cannot say what the loss of these species represents to ecosystems or to humanity’s future prospects (Stork, 1997; Tilman, 1999; quotes in Takacs, 1996). Biologists believe that if people are alarmed by such ignorance, they will simultaneously take precautions about dismantling an ecological world that supports us in myriad unknown ways, will support biologists’ research efforts to understand that world, and will listen to biologists’ policy prescriptions on how and why we ought to save that world. For example, the question that E. O. Wilson (1992) is most frequently asked about the diversity of life is: “if enough species are extinguished, will the ecosystem collapse, and will the extinction of most other species follow soon afterward? The only answer anyone can give is: possibly. By the time we find out, however, it might be too late. One planet, one experiment” (p. 182). In light of this uncertainty, Wilson urges “prudence. We should judge every scrap of biodiversity as priceless while we learn to use it and come to understand what it means to humanity” (p. 351). In Wilson’s view, biologists hold the key to that understanding.

Many who have come to call themselves conservation biologists cite this uncertainty as they attempt to have the public become aware of biodiversity and become concerned about its diminution. They warn that what we don’t know about biodiversity will hurt us, and urge us that if we were to come to know it, our lives might have greater meaning. They seek a new ethic where biodiversity would be cherished. Aldo Leopold (1949) sought this, and his “land ethic” laid the groundwork for a new way of valuing diversity. Leopold knew that “no important change in ethics was ever accomplished without an internal change in our intellectual emphasis, loyalties, affections, and convictions. The proof that conservation has not yet touched these foundations of conduct lies in the fact that philosophy and religion have not yet heard of it.” Today, a number of prominent biologists seek to have our intellectual emphasis, loyalties, affections, and convictions match theirs. Some biologists see this as a crucial moment in history when they must help raise awareness of biodiversity; they see it as their responsibility to speak out about what they study and love so that subsequent human and natural history will reflect their values. For example, Janzen (1986) exhorts his fellow biologists: “Set aside your random research and devote your life to activities that will bring the world to understand that tropical nature is an integral part of human life.” Janzen is part of the “mission-oriented” discipline of conservation biology, which was founded not only to study biodiversity, but also to embrace its many values and to promote findings that will help conserve the
objects of study. Built into the discipline's foundations are normative principles or, as discipline founder Michael Soulé (1983, p. 730) says, “value statements that make up the basis of appropriate attitudes towards other forms of life—an ecosophy.” The pages of the discipline’s flagship journal, Conservation Biology, report not just the latest research findings; they are also filled with thoughtful articles on conservation education, policy implications of research, and spirited debates on what types of activism are required or appropriate for practitioners. While some biologists feel they must advocate on behalf of the natural world they study and love, others fear that such advocacy threatens the scientific enterprise: if scientists are no longer perceived as objective and value neutral, why would societies fund their work and listen to their results?

Many scientists ignore the roles of advocacy and proselytize on behalf of biodiversity. Some well-known biologists (e.g., E. O. Wilson, Peter Raven, Thomas Eisner, Thomas Lovejoy) have spoken before Congress on the value of biodiversity. Others testify in courts as expert witnesses on behalf of biodiversity. Paul Ehrlich and others appear on radio and television to raise awareness about biodiversity. Many biologists write for the popular press or speak to general audiences on the values of biodiversity. They talk to garden clubs, and preach in church pulpits. Thomas Lovejoy takes senators, movie stars, and other influential people to the Brazilian Amazon so that they may experience biodiversity “in a way words can’t touch. . . . And it has never failed to be a truly touching experience for them” (Takacs, 1996, p. 153). By helping laypersons become aware of biodiversity’s beauty and fragility—by tireless efforts to, in the words of the “Father of Biodiversity” E. O. Wilson, “educate, educate, educate” (quoted in Anonymous, 1994)—biologists hope for widespread transformation of values. That transformation of values might lead to what Paul Ehrlich (1985) calls a “quasi-religious transformation” of feelings toward the wonders of the natural world, which would presumably be translated into action to conserve those wonders.

Many prominent North American biologists work in farflung corners of the earth to raise awareness of biodiversity, particularly in tropical nations that hold the greatest concentrations of diversity, and which face the greatest imminent threats to that bounty. In Costa Rica, a small country with perhaps 5% of the world’s species diversity, biologists have spurred a national effort not only to preserve biodiversity in protected areas, but also to catalog biodiversity and to search for the economic wealth contained within. At Costa Rica’s Instituto Nacional de Biodiversidad, specimens of plants, insects, and other taxa are sorted and sent to multinational pharmaceutical companies for research. Profits from successful drugs deriving from Costa Rica’s biodiversity would be funneled back to the areas where the biodiversity flourishes. In effect, this is an experiment in which biodiversity would pay its own way, with biologists driving the process.

Specimens are collected by a large group of rural-dwelling parataxonomists. As they tromp around the nation’s forests, the idea goes, they will lead a movement in “biocultural restoration” by becoming reenchanted with biodiversity, they will be compelled to spread their new knowledge and appreciation to their neighbors. Through renewed awareness—and renewed economic dividends—rural dwellers would be empowered and inclined to protect their biodiversity patrimony (Evans, 1999; Gámez et al., 1993; Janzen, 1988).

VI. THE VALUES OF BIODIVERSITY

In Costa Rica, and in many other corners of the earth, biologists are working to raise awareness of the many values of biodiversity. Biologists hope that biodiversity’s appeal to many different audiences will spark diverse efforts to preserve it. Some of the values biologists attribute to biodiversity find their loci in biodiversity itself; in most, though, humans are the value holders. As we might expect, biologists promote the value biodiversity holds for science, as the raw material for scientists’ investigations: it is “Earth’s living library” (Lovejoy, 1992). Also unsurprisingly, biologists speak extensively about biodiversity’s ecological value. Biodiversity provides numerous “ecosystem services” (or, depending on your definition, “ecosystem services” is part of what constitutes “biodiversity”). That is to say, biodiversity keeps global ecosystems functioning. Biodiversity purifies water, controls agricultural pests, decomposes waste, pollinates plants, stabilizes global climate, creates soil, transports nutrients, controls erosion, and maintains the ecological matrix human society requires to exist at all (Daily, 1997; Tilman, 1999). And as humans attempt to repair damaged ecosystems, we will need reservoirs of intact biodiversity to restore what we have destroyed (Jordan, 1997).

For those who measure value in terms of money, biologists assert that biodiversity has vast economic value. Biologists note that biodiversity’s value to functioning ecosystems is priceless—although Costanza et al. (1997) have tried to calculate the incalculable. Costa-
dollars per year to human economies, and they believe this is a conservative estimate. Thus, they note, is higher than the global gross national product of 18 trillion dollars per year. Janzen (1986) offers an exhaustive list of the treasures biodiversity already has directly provided human society. Others (e.g., Wilson, 1992) discuss the lure that can be gained from the hidden food, medicine, chemicals, fibers, and other goods lurking in the world’s natural places. As noted earlier, Costa Rica has invested heavily in its biodiversity; through ecotourism and bioprospecting, the nation is profiting from its biodiversity resources. Biologists (e.g., Janzen, 1990) also tout biodiversity’s social amenity value: it can contribute to sustainable development efforts (through direct harvesting or ecotourism), foster national pride, or help those living nearby to lead more intellectually and aesthetically fulfilled lives.

E. O. Wilson (e.g., 1992, 1998) has been the leading proponent of biodiversity’s value to fulfill our biophilic impulses. Wilson and others believe that love of nature has been hardwired into our genes, and we can only be truly fulfilled by satisfying that love. Those whose biophilic passions have been rekindled will also be those who work hardest for biodiversity preservation: others refer to this as biodiversity’s “transformative value.” Philosopher Bryan Norton (1987) explains how interactions with biodiversity can help us reconsider our consumerist impulses and make us convert to lifeways that preserve the biodiversity that provided the impetus for transformation. Biodiversity’s aesthetic value—the beauty of an individual organism and its adaptations, of landscapes, of intricate ecological processes, of the sheer riot of different life forms—may effect this kind of transformation.

When biologists suggest that biodiversity may transform us, may reawaken our biophilic impulses, they still place the locus of its value in the human valuer. Some biologists take a different tack and assert that biodiversity has intrinsic value, independent of a human valuer. This concept is difficult to prove empirically and therefore difficult for a scientist to assert. Yet given the preceding definitions of “biodiversity,” if biodiversity is the totality of life forms on earth, their interactions, and the processes that gave rise to them and to us, it makes it a more complicated proposition to reject the notion of biodiversity’s intrinsic value out of hand. This notion takes a turn for the spiritual, and, in fact, some biologists also discuss biodiversity’s spiritual value. David Ehrenfeld (1981), who would become the founding editor of Conservation Biology, wrote of “The Noah Principle”: species “should be conserved because they exist and because this existence is itself but the present expression of a continuing historical process of immense antiquity and majesty. Long-standing existence in Nature is deemed to carry with it the unimpeachable right to continued existence.” Some biologists (see quotes in Takacs, 1996) agree that the value inherent in biodiversity makes it sacred and others put the locus of value in humans, and they discuss the spiritual nourishment that contact with biodiversity brings.

The multiplicity of meanings of “biodiversity” is reflected in the multiplicity of values biologists find in it. For those biologists who would raise awareness of biodiversity’s plight, it makes sense to speak for as many different values—and therefore to as many different audiences—as possible.

VII. HISTORICAL AWARENESS OF BIODIVERSITY REDUX

We can read the title of this encyclopedia entry in many ways. Smithsonian tropical biologist and biodiversity advocate Terry Erwin says (in Takacs, 1996, pp. 100–101): “I’ve always kind of been ahead of the game or ahead of the thinking in my own little field of entomology . . . I think there will be individuals who can influence the direction of change. And I’d like to be among them, for whatever little part I might be able to play.” Ideas can act as forces of nature. They can change ecologies. They can reshape how we value and therefore how we treat nature. On behalf of biodiversity, biologists are making history: they are changing the course of events so that human history and natural history will unfurl to their liking. This encyclopedia attempts to inform, to raise awareness of both the science of biodiversity and the conscience of those who study it, revere it, and wish to see it preserved.

History suggests that two phenomena will continue to obtain: the biological complexity of the earth will continue to diminish, and biologists will continue to look for strategies that will compel us to care about what they care about and support their authority to speak for those entities. Biologists have attempted to raise our awareness of biodiversity: the complexity of real world organisms, species, and processes commingled with biologists’ factual, political, emotional, ethical, aesthetic, and spiritual values of the natural world, all combined to shape public perceptions, actions, and feelings. To become aware of it is to understand not only how biologists understand the web of life and interconnections, but to become aware of our own obligations to future evolutionary history.
See Also the Following Articles

**Biodiversity, Definition of** - Biodiversity-rich countries • Conservation biology, discipline of • conservation efforts, contemporary • Darwin, Charles • Economic value of biodiversity, overview • Forest canopies, animal diversity • stewardship, concept of

**Bibliography**


I. Introduction

We are witnessing the opening phase of a mass extinction episode that, if allowed to persist, could well eliminate a large proportion of all species among other forms of biodiversity within the foreseeable future (Myers, 1993; Wilson, 1992). We do not have nearly enough conservation resources (funds, scientific skills, and the like) to assist all species under threat, and as the biotic crisis gathers momentum the shortfall will become ever more severe. This predicament places a premium on priority planning. Which conservation strategies offer the biggest payoff? Or, to be more precise: How can we save the most species at the least cost? This key question is likely to remain at the forefront of conservation endeavors as the Earth's biotic crisis grows worse. By concentrating on a few critical areas where needs are

HOTSPOTS are specific areas of the Earth's land surface that have a disproportionately large number of extant species. Identification of the world's hotspots—roughly 18–25 in number, depending on the criteria employed—provides a means of focusing on those areas where threats to biodiversity are most extreme and conservation efforts can be most effective. Underlying this approach is the thesis that present conservation resources are not sufficient to maintain all threatened species and thus global priorities need to be established.

HOTSPOTS are specific areas of the Earth's land surface that have a disproportionately large number of extant species. Identification of the world's hotspots—roughly 18–25 in number, depending on the criteria employed—provides a means of focusing on those areas where threats to biodiversity are most extreme and conservation efforts can be most effective. Underlying this approach is the thesis that present conservation resources are not sufficient to maintain all threatened species and thus global priorities need to be established.

II. Biodiversity Hotspots as Originally Identified

III. Revised and Expanded Hotspots Analysis

IV. Main Findings of Revised Hotspots Analysis

V. Wilderness Areas

VI. High-Value Ecosystems

VII. Megadiversity Countries

VIII. The Conservation Impact of Hotspots

GLOSSARY

biodiversity Popularity supposed to refer to the spectrum of all species on Earth, the concept should also include species' subunits (genetic diversity) and the diversity of ecosystems and ecological processes.

endemics Those species that are limited to relatively small areas, being found nowhere else on Earth.

hotspots Those areas that (a) feature exceptional concentrations of endemic species and (b) face imminent threat of habitat destruction.

megadiversity Phenomenon of at least 70% of all species being confined to 17 "megadiversity" countries.
greatest and where the payoff from safeguard measures would also be greatest, conservationists can engage in a more systematized response to the challenge of large-scale extinctions that lie ahead. This represents a focused silver bullet response, in contrast to the scattergun approach that has characterized much conservation activity to date.

This is not to say—and the point is emphasized—that biodiversity outside of species-rich areas should be ignored. The biodiversity of any country is vitally important to that country’s environmental well-being. A number of responses have been proposed. The “hotspots” thesis is one such mode of setting priorities at a time when we need to determine priorities with more scientific acumen than ever. This approach identifies areas that feature exceptional concentrations of endemic species and that face exceptional threat of imminent habitat destruction (Myers, 1988, 1990; Myers et al., 2000). A hotspots strategy can be complemented with measures that highlight “megadiversity” countries, that is, those few countries that harbor most of the world’s species, whether threatened or not (Mittermeier and Mittermeier, 1997). A further backup strategy is the protection of wilderness areas, being extensive tracts of little-disturbed wildlands with rich biodiversity stocks where conservationists can “get it right” from the start. Still a fourth response would focus on “high-value” ecosystems that, while not harboring unusual concentrations of species, encompass other remarkable manifestations of biodiversity. We shall consider each of these strategies, while giving most attention to the most promising option, the conservation of hotspots.

II. BIODIVERSITY HOTSPOTS AS ORIGINALLY IDENTIFIED

The hotspots strategy was first raised in the late 1980s, when an exploratory listing of hotspots identified 18 localities, 14 in tropical moist forests and 4 in Mediterranean-type zones (Myers, 1988, 1990). The analysis centered on higher plant species alone, with the assumption that these would serve as acceptable indicator taxa for other categories of species (mammals, birds, and other vertebrates, plus invertebrates). Analysis revealed that at least 20% of all plant species were confined to areas comprising 0.5% of Earth’s land surface—areas that for the most part have already lost the bulk of their biodiversity habitats.

III. REVISED AND EXPANDED HOTSPOTS ANALYSIS

A more recent effort focusing on 25 biodiversity hotspots (Table I) has sought to refine and expand the original hotspots analysis by including other taxa and biomes (Myers et al., 2000). It has centered not only on higher plants but also on birds, mammals, reptiles, and amphibians as indicators of vertebrate taxa. In addition, it has extended ecozone coverage to include tropical dry forests, woodlands, savanna and open savanna, temperate moist forests, grasslands, and arid lands.

This expanded approach still omits invertebrates from consideration, which is regrettable in that invertebrates comprise the great majority of all species, at least 95% and possibly 99%. Future analysis could be extended to include butterflies as indicators of invertebrate hotspots. Butterflies are more closely tied to plant communities than vertebrates, and yet are popular enough with amateur naturalists that for some areas we have quite accurate records of their populations through time. (As a bonus factor, butterflies are sometimes the best single group of animals as indicators of ecosystem health.) Additional support for invertebrate hotspots could be gained from focusing on dragonflies, damselflies, and tiger beetles, all of which are widespread and fairly well known.

This latest hotspots analysis, like the earlier one, is limited to terrestrial biota. It is also based on vascular or higher plants (which comprise around 90% of all plants, and are hereafter referred to as “plants”), since they are essential to virtually all forms of animal life. These plants are well known, with their conservation status adequately documented for the most part. The endemism data tend to be minimal for two reasons: One is the sheer lack of recent documentation in the form of, for example, modern floras. For instance, there is no up-to-date account of Brazil’s plant species even though the country is believed to harbor 50,000 species or one-sixth of the world’s total. More importantly, endemism statistics almost always relate only to individual countries or parts of countries, whereas 12 of the hotspots extend across two or more countries and 6 hotspots across four or more. In these cases, it has been singularly difficult to compute totals for hotspot-wide endemics with cross-boundary complications, and the analysis (Myers et al., 2000) has usually had to depend on best-judgment estimates by scientists with extensive on-ground experience. These estimates tend to be cautious and conservative, meaning that the
<table>
<thead>
<tr>
<th>Hotspot</th>
<th>Original extent (km²)</th>
<th>Remaining primary vegetation (km²)</th>
<th>Area protected (% of original extent)</th>
<th>Endemic plants (% of global extent)</th>
<th>Endemic vertebrates (% of global vertebrates)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical Andes</td>
<td>1,238,000</td>
<td>314,900 (25.0)</td>
<td>79,687 (25.3)</td>
<td>49,000</td>
<td>20,000 (6.7%)</td>
</tr>
<tr>
<td>Mesoamerica</td>
<td>1,155,000</td>
<td>231,000 (20.0)</td>
<td>136,477 (39.0)</td>
<td>24,000</td>
<td>5000 (1.7%)</td>
</tr>
<tr>
<td>Caribbean</td>
<td>2,613,300</td>
<td>29,840 (11.3)</td>
<td>29,840 (100.0)</td>
<td>12,000</td>
<td>7000 (2.3%)</td>
</tr>
<tr>
<td>Brazil’s Atlantic Forest</td>
<td>1,227,600</td>
<td>93,930 (7.5)</td>
<td>33,264 (33.9)</td>
<td>20,000</td>
<td>8500 (2.7%)</td>
</tr>
<tr>
<td>Choco/Darien/Western Ecuador</td>
<td>260,600</td>
<td>65,000 (24.2)</td>
<td>16,471 (26.1)</td>
<td>10,000</td>
<td>2250 (0.8%)</td>
</tr>
<tr>
<td>Brazil’s Cerrado</td>
<td>1,783,200</td>
<td>396,630 (21.0)</td>
<td>22,000 (6.2)</td>
<td>10,000</td>
<td>4400 (1.5%)</td>
</tr>
<tr>
<td>Central Chile</td>
<td>300,000</td>
<td>90,000 (30.0)</td>
<td>9167 (10.2)</td>
<td>3429</td>
<td>1609 (0.5%)</td>
</tr>
<tr>
<td>California Floristic Province</td>
<td>324,000</td>
<td>80,000 (24.7)</td>
<td>31,443 (39.3)</td>
<td>4426</td>
<td>2125 (0.7%)</td>
</tr>
<tr>
<td>Madagascar¹</td>
<td>394,150</td>
<td>99,038 (25.0)</td>
<td>11,348 (11.6)</td>
<td>12,000</td>
<td>9724 (3.2%)</td>
</tr>
<tr>
<td>Eastern Arc and Coastal Forests of Kenya/Tanzania</td>
<td>50,000</td>
<td>2000 (4.0)</td>
<td>2000 (100.0)</td>
<td>4000</td>
<td>1500 (0.5%)</td>
</tr>
<tr>
<td>Guinean Forests of West Africa</td>
<td>1,265,000</td>
<td>128,500 (10.0)</td>
<td>20,324 (16.1)</td>
<td>9000</td>
<td>2250 (0.8%)</td>
</tr>
<tr>
<td>Cape Floristic Province</td>
<td>74,000</td>
<td>18,000 (24.3)</td>
<td>14,060 (78.1)</td>
<td>8200</td>
<td>9682 (1.9%)</td>
</tr>
<tr>
<td>Mediterranean Basin</td>
<td>112,000</td>
<td>30,000 (26.8)</td>
<td>2352 (7.8)</td>
<td>4849</td>
<td>1940 (0.6%)</td>
</tr>
<tr>
<td>Iberian Peninsula</td>
<td>2,362,000</td>
<td>110,000 (4.7)</td>
<td>42,123 (38.3)</td>
<td>25,000</td>
<td>13,000 (4.3%)</td>
</tr>
<tr>
<td>Canaries</td>
<td>500,000</td>
<td>50,000 (10.0)</td>
<td>14,090 (28.1)</td>
<td>6300</td>
<td>1600 (0.5%)</td>
</tr>
<tr>
<td>Sundaland</td>
<td>1,605,000</td>
<td>123,000 (7.8)</td>
<td>90,000 (72.0)</td>
<td>25,000</td>
<td>13,000 (5.0%)</td>
</tr>
<tr>
<td>Wallacia</td>
<td>347,000</td>
<td>32,020 (15.0)</td>
<td>20,413 (30.2)</td>
<td>10,000</td>
<td>1500 (0.5%)</td>
</tr>
<tr>
<td>Philippines</td>
<td>300,000</td>
<td>700 (3.0)</td>
<td>3910 (43.3)</td>
<td>7620</td>
<td>5032 (1.9%)</td>
</tr>
<tr>
<td>Indo-Malaya</td>
<td>2,060,000</td>
<td>100,000 (4.9)</td>
<td>13,000 (100.0)</td>
<td>7000</td>
<td>5032 (1.9%)</td>
</tr>
<tr>
<td>South-Central China</td>
<td>800,000</td>
<td>64,000 (8.0)</td>
<td>16,562 (25.0)</td>
<td>12,000</td>
<td>1500 (0.5%)</td>
</tr>
<tr>
<td>Western Ghats/Sri Lanka</td>
<td>182,500</td>
<td>12,450 (6.8)</td>
<td>12,450 (100.0)</td>
<td>4780</td>
<td>2180 (0.7%)</td>
</tr>
<tr>
<td>SW Australia</td>
<td>309,850</td>
<td>33,336 (10.8)</td>
<td>33,336 (100.0)</td>
<td>5469</td>
<td>4331 (1.4%)</td>
</tr>
<tr>
<td>New Caledonia</td>
<td>18,600</td>
<td>5200 (28.0)</td>
<td>320 (10.1)</td>
<td>3332</td>
<td>2951 (0.9%)</td>
</tr>
<tr>
<td>New Zealand</td>
<td>270,000</td>
<td>59,000 (22.0)</td>
<td>52,068 (87.7)</td>
<td>2300</td>
<td>1865 (0.6%)</td>
</tr>
<tr>
<td>Polynesia/Micronesia</td>
<td>46,000</td>
<td>10,024 (21.8)</td>
<td>4913 (49.0)</td>
<td>6357</td>
<td>3334 (1.5%)</td>
</tr>
<tr>
<td>Total</td>
<td>17,444,300</td>
<td>2,122,891 (12.2)</td>
<td>800,767 (37.7)</td>
<td>133,149</td>
<td>9648 (3.3%)</td>
</tr>
</tbody>
</table>

¹ Choco/Darien/Western Ecuador stretches from the Darien of Panama to Western Ecuador; Cape Floristic Province lies in the southern sector of South Africa; the Succulent Karos lies in western South Africa; Sundaland encompasses the islands of western Indonesia together with the Malay Peninsula; and Wallacea includes islands of eastern Indonesia.

² Madagascar includes the nearby islands of Mauritius, Reunion, Seychelles, and Comores.

³ These totals cannot be summed due to overlapping between hotspots.


To qualify as a hotspot, the main determining criterion is species endemism. A second criterion is degree of threat; to qualify, an area must retain only 30% or less of its original primary vegetation. The cutoff adopted is 1500 endemic plant species, or 0.5% of the 300,000 plant species on land (Prance et al., 2000).

Unlike the earlier hotspots analysis, the expanded version includes birds, mammals, reptiles, and amphibians. It excludes the only other vertebrate group, fishes (totaling roughly half of all vertebrates), because data

true totals will likely be higher than those presented here.
about their numbers, habitats, and conservation status are generally poor. Hereafter, the term "vertebrates" refers to all vertebrates except fishes. Not that vertebrates serve as a second determinant of hotspot status; if an area qualifies by the 0.5% plants criterion, it makes the list. Vertebrates are strictly used for backup support and to determine measures of congruence. Sixteen hotspots contain at least 1.30 vertebrate endemics, or 0.5% of the 27,208 species of the four vertebrate groups worldwide, while 12 of these contain at least twice as many.

As noted, the analysis omits insects and other invertebrates. While scientists have documented the great majority of plants and vertebrates, they have documented only a tiny proportion of invertebrates. For present purposes and purely as a working proposition, it is reasonable to assume that the five categories of endemic species documented here are roughly matched by similar concentrations of endemic invertebrate species. Although this assumption is preliminary and approximate, it is supported by more evidence in its favor than against.

Many of the hotspots have already lost 90% of their original primary vegetation, and a few of them, for example, the Mediterranean Basin, Indo-Burma, and the Philippines, have lost at least 93%. It is true that disrupted and secondary vegetation can sometimes support moderate numbers of original species, but in the main this is not significant for conservation purposes.

Constraints of socioeconomic status, political commitment, or conservation feasibility in the countries concerned have not been considered. This is partly because these factors are difficult to quantify, and partly because they can be better incorporated when designing conservation projects. All the same, it is worth noting that Nepal, for instance, hardly possesses the administrative structures, the managerial know-how, or the planning capacities that can usefully absorb additional external assistance. Related questions formerly arose with respect to Zaire (now the Democratic Republic of Congo): Should that country have been granted conservation support when its military spending was twice as large a proportion of its gross domestic product as the average for sub-Saharan Africa, and when the personal wealth of President Mobutu was greater than the entire country’s economy?

### IV. MAIN FINDINGS OF REVISED HOTSPOTS ANALYSIS

A total of 25 hotspots contain the remaining habitats of 133,140 plant species (44% of all plant species) and of 96,453 vertebrate species (33% of all such species). These endemics are confined to an aggregate area equivalent to 1.4% of Earth’s land surface (Table I). The hotspots are so threatened that, having already lost at least 70% of their original primary vegetation, they all seem likely, in the absence of greatly increased conservation efforts, to lose much if not most of their remaining vegetation within the foreseeable future.

The 25 hotspots feature a broad range of ecosystem types. Predominantly are tropical rain forests (in 15) and Mediterranean-type zones (in 3). Nine are mainly or completely made up of islands; almost all tropical islands fall into one or another hotspot. Sixteen hotspots are in the tropics, which largely places them in developing countries where threats are greatest and conservation resources are in shortest supply.

Now consider the relationship of endemic species to total species. In the 17 “megadiversity” countries with some 70% of Earth’s species, the ratio of endemic non-fish vertebrates to all vertebrates ranges from a high of 1:1.3 for Madagascar to a low of 1:1.4 for the Democratic Republic of Congo (formerly Zaire). When all 25 hotspots are considered, the average ratio is 1:2.8. This high ratio of endemism demonstrates the significance and rarity of the biodiversity found in these hotspots.

Still more significant, the extent of habitat loss in the hotspots means that we can reasonably assume they harbor an even greater share of threatened species, defined here as Red Data Book species (these species are assessed by the World Conservation Union and include only species known to science and known to be threatened; the true total is far higher). So far as we can calculate, albeit in a preliminary and exploratory manner, the number of threatened species occurring in hotspots probably amounts to roughly two-thirds of all threatened species.

### A. The “Hotter” Hotspots

Some hotspots are “hotter” than others. In nine hotspots, 30% of all plants are endemics (in the Tropical Andes, an exceptional 6.7% are endemic) and 25% of vertebrate species are endemics; these hotspots account for 0.7% of Earth’s land surface (Table II). At the same time, they feature some of the most depleted habitats anywhere: Madagascar retains less than 10% of its original primary vegetation; Sundaland and Brazil’s Atlantic Forest less than 8%; the Mediterranean Basin, Indo-Burma, and the Philippines less than 3%. Five hotspots hold more than 2% of the world’s biodiversity in both plants and vertebrates, hence they are super hotspots: Tropical Andes, 6.7% and 5.7% respectively; Sunda-
TABLE II

Leading Hotspots in Terms of Endemic Species

<table>
<thead>
<tr>
<th>Hotspot</th>
<th>Endemic plants (% of global plants total)</th>
<th>Endemic vertebrates (% of global vertebrates total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical Andes</td>
<td>20,000 (6.7)</td>
<td>1367 (5.7)</td>
</tr>
<tr>
<td>Sundaland</td>
<td>15,000 (5.0)</td>
<td>701 (2.6)</td>
</tr>
<tr>
<td>Madagascar</td>
<td>9704 (3.2)</td>
<td>771 (2.8)</td>
</tr>
<tr>
<td>Brazil’s Atlantic Forest</td>
<td>8000 (2.7)</td>
<td>567 (2.1)</td>
</tr>
<tr>
<td>Caribbean</td>
<td>7000 (2.3)</td>
<td>779 (2.9)</td>
</tr>
<tr>
<td>Total</td>
<td>59,704 (19.9)</td>
<td>4385 (16.1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hotspot</th>
<th>Endemic plants (% of global plants total)</th>
<th>Endemic vertebrates (% of global vertebrates total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesoamerica</td>
<td>5000 (1.7)</td>
<td>1159 (4.2)</td>
</tr>
<tr>
<td>Mediterranean Basin</td>
<td>13,000 (4.3)</td>
<td>213 (0.9)</td>
</tr>
<tr>
<td>Indo-Burma</td>
<td>7000 (2.3)</td>
<td>528 (1.9)</td>
</tr>
<tr>
<td>Philippines</td>
<td>3612 (1.2)</td>
<td>516 (1.9)</td>
</tr>
<tr>
<td>Total</td>
<td>90,536 (30.1)</td>
<td>6826 (25.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hotspot</th>
<th>Species/Area Ratios per 100 km² of Hotspots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical Andes</td>
<td>6.4 : 0.5</td>
</tr>
<tr>
<td>Mesoamerica</td>
<td>2.2 : 0.3</td>
</tr>
<tr>
<td>Caribbean</td>
<td>23.5 : 2.6</td>
</tr>
<tr>
<td>Brazil’s Atlantic Forest</td>
<td>8.7 : 0.6</td>
</tr>
<tr>
<td>Choco/Darien/Western Ecuador</td>
<td>3.6 : 0.7</td>
</tr>
<tr>
<td>Brazil’s Cerrado</td>
<td>1.2 : 0.03</td>
</tr>
<tr>
<td>Central Chile</td>
<td>1.8 : 0.06</td>
</tr>
<tr>
<td>California Floristic Province</td>
<td>2.7 : 0.09</td>
</tr>
<tr>
<td>Madagascar</td>
<td>16.4 : 1.3</td>
</tr>
<tr>
<td>Eastern Arc and Coastal Forests of Kenya/Tanzania</td>
<td>75.0 : 6.1</td>
</tr>
<tr>
<td>Guinean Forests of West Africa</td>
<td>1.8 : 0.2</td>
</tr>
<tr>
<td>Cape Floristic Province</td>
<td>31.6 : 0.5</td>
</tr>
<tr>
<td>Succulent Karoo</td>
<td>6.5 : 0.15</td>
</tr>
<tr>
<td>Mediterranean Basin</td>
<td>11.8 : 0.2</td>
</tr>
<tr>
<td>Costarica</td>
<td>3.2 : 0.1</td>
</tr>
<tr>
<td>Sundaland</td>
<td>12.0 : 0.6</td>
</tr>
<tr>
<td>Wallacea</td>
<td>2.9 : 1.0</td>
</tr>
<tr>
<td>Philippines</td>
<td>64.7 : 5.7</td>
</tr>
<tr>
<td>Indo-Burma</td>
<td>7.0 : 0.3</td>
</tr>
<tr>
<td>South-Central China</td>
<td>9.5 : 0.3</td>
</tr>
<tr>
<td>Western Ghats/Sri Lanka</td>
<td>17.3 : 2.9</td>
</tr>
<tr>
<td>SW Australia</td>
<td>13.0 : 0.5</td>
</tr>
<tr>
<td>New Caledonia</td>
<td>49.1 : 1.6</td>
</tr>
<tr>
<td>New Zealand</td>
<td>3.1 : 0.2</td>
</tr>
<tr>
<td>Polynesia/Micronesia</td>
<td>35.3 : 2.2</td>
</tr>
</tbody>
</table>

* Hotspots with at least 2% of both endemic plants and vertebrates, and together comprising only 0.4% of Earth’s land surface (all nine hotspots amount to 0.7% of Earth’s land surface).

** This would total 30.2% but for rounding of numbers in the individual hotspots.

B. Species/Area Relationships

Some hotspots are also significant because their endemic species are concentrated in exceptionally small areas (Table III). The Eastern Arc and Coastal Forests of Tanzania/Kenya (hereafter referred to as “Eastern Arc”) contain 1500 endemic plants in 2000 km² for a ratio of 75 species to 100 km², or 75 : 1, and 121 endemic vertebrates for a ratio of 0.6 : 1—both ratios top the lists for all hotspots. New Caledonia (5200 km²) has ratios of 49 : 1 and 1.6 : 1 for endemic plants and vertebrates, the Philippines (3023 km²) has 64 : 1 and 3.7 : 1, Polynesia/Micronesia (10,024 km²) has 33 : 1 and 2.2 : 1, and the Western Ghats in India (12,950 km²) has 17.5 : 1 and 2.9 : 1. Ratios for the other areas range from 18 : 1 to 1 : 2 : 1 for plants and from 2.0 : 1 to 0.03 : 1 for vertebrates.

C. Congruence among Species Categories

In several hotspots, there is a measure of congruence between plants and vertebrates insofar as high counts for endemic plants are matched by moderately high counts for endemic vertebrates (Table IV). This factor can reinforce the conservation priority thesis, especially in those hotspots with the most endemic species. There can also be high congruence in areas with lower species counts, for example, there is 100% congruence in the Philippines with 1.9% of both endemic plants and vertebrate species worldwide, and 80% congruence in the Eastern Arc with 0.3% of plant species and 0.4% of vertebrate species. But the species percentages of these areas are low relative to those of several other hotspots.

The Tropical Andes holds 6.7% of all endemic plant species worldwide and 5.7% of endemic vertebrates, for 85% congruence. Madagascar’s endemic species repre-
TABLE IV
Congruence between Endemic Plants and Vertebrates

<table>
<thead>
<tr>
<th>Hotspot</th>
<th>Endemic plants as % of global plants total, 300,000</th>
<th>Endemic vertebrates as % of global vertebrates total, 27,208</th>
<th>% Congruence (rounded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical Andes</td>
<td>6.7%</td>
<td>3.7%</td>
<td>89%</td>
</tr>
<tr>
<td>Mesoamerica</td>
<td>1.7%</td>
<td>4.2%</td>
<td>41%</td>
</tr>
<tr>
<td>Caribbean</td>
<td>2.3%</td>
<td>2.9%</td>
<td>79%</td>
</tr>
<tr>
<td>Brazil’s Atlantic Forest</td>
<td>2.7%</td>
<td>2.1%</td>
<td>78%</td>
</tr>
<tr>
<td>Choco/Darien/Western Ecuador</td>
<td>0.8%</td>
<td>1.5%</td>
<td>53%</td>
</tr>
<tr>
<td>Brazil’s Cerrado</td>
<td>1.3%</td>
<td>0.4%</td>
<td>27%</td>
</tr>
<tr>
<td>Central Chile</td>
<td>0.5%</td>
<td>0.2%</td>
<td>40%</td>
</tr>
<tr>
<td>Madagascar</td>
<td>0.7%</td>
<td>0.5%</td>
<td>43%</td>
</tr>
<tr>
<td>Madagascar</td>
<td>3.2%</td>
<td>2.8%</td>
<td>88%</td>
</tr>
<tr>
<td>Eastern Arc and Coastal Forest of Kenyana/Tanzania</td>
<td>0.3%</td>
<td>0.4%</td>
<td>80%</td>
</tr>
<tr>
<td>Guinean Forests of West Africa</td>
<td>0.8%</td>
<td>1.0%</td>
<td>80%</td>
</tr>
<tr>
<td>Cape Floristic Province</td>
<td>1.9%</td>
<td>0.2%</td>
<td>11%</td>
</tr>
<tr>
<td>Succulent Karoo</td>
<td>0.6%</td>
<td>0.2%</td>
<td>33%</td>
</tr>
<tr>
<td>Mediterranean Basin</td>
<td>1.3%</td>
<td>0.8%</td>
<td>21%</td>
</tr>
<tr>
<td>Madagascar</td>
<td>0.9%</td>
<td>0.2%</td>
<td>40%</td>
</tr>
<tr>
<td>Sundaland</td>
<td>1.0%</td>
<td>2.6%</td>
<td>32%</td>
</tr>
<tr>
<td>Wallacea</td>
<td>0.5%</td>
<td>1.0%</td>
<td>26%</td>
</tr>
<tr>
<td>Philippine</td>
<td>1.5%</td>
<td>1.9%</td>
<td>80%</td>
</tr>
<tr>
<td>Andes-Burma</td>
<td>2.3%</td>
<td>1.0%</td>
<td>83%</td>
</tr>
<tr>
<td>South-Central China</td>
<td>1.2%</td>
<td>0.7%</td>
<td>58%</td>
</tr>
<tr>
<td>Western Ghats/Sri Lanka</td>
<td>0.5%</td>
<td>1.5%</td>
<td>54%</td>
</tr>
<tr>
<td>SW Australia</td>
<td>1.4%</td>
<td>0.4%</td>
<td>29%</td>
</tr>
<tr>
<td>New Caledonia</td>
<td>0.9%</td>
<td>0.5%</td>
<td>33%</td>
</tr>
<tr>
<td>New Zealand</td>
<td>0.6%</td>
<td>0.9%</td>
<td>83%</td>
</tr>
<tr>
<td>Polynesia/Micronesia</td>
<td>1.1%</td>
<td>0.6%</td>
<td>73%</td>
</tr>
</tbody>
</table>

sent 3.2% and 2.8%, for 88% congruence; the Caribbean has 2.3% and 2.9%, to give 79% congruence. (The Tropical Andes is a large area where one could expect high congruence; the other two are only one-fifth and one-tenth as big respectively.) By contrast, the Cape Floristic Province possesses 1.9% of all endemic plants but only 0.2% of all endemic vertebrates. Congruence tends to be high in tropical forest hotspots, and generally low in Mediterranean-type hotspots (the congruence for Cape Floristic Province is 11%) and other drier areas with their meager counts for endemic vertebrates.

The four vertebrate groups reveal varying degrees of congruence among themselves. Birds and amphibians (like plants) generally show an increase in species numbers in the tropics and still more nearer the equator, with particularly high totals in tropical forests. Their numbers also increase with altitude up to 2500 m in localities with good rainfall. Similarly, reptile abundance increases nearer the equator, though in drier zones it is comparable to that in tropical forests, as witness the situation in the two countries with the most reptiles, Australia and Mexico. Mammal numbers also increase closer to the equator, with drier areas again having a species richness comparable to that of tropical forests. Certain groups within the mammals, notably primates and bats, show similar trends to the birds and amphibians, though primates reveal their greatest species numbers in lowland rain forests, declining rapidly with altitude.

D. The “Hottest” Hotspots
It is not practicable to devise a hotspots ranking index that combines five criteria, namely, numbers of endem-
ics and endemic species/area ratios for both plants and vertebrates, and habitat loss. These criteria cannot carry equal weight (a case of comparing apples and oranges), so one cannot simply sum the rankings for each case. For comparative purposes, Table V lists the eight “hottest hotspots” that appear at least three times in the top ten rankings for each criterion. The leaders are Madagascar, the Philippines, and Sundaland, which appear in all five criteria, followed by the Caribbean and Brazil’s Atlantic Forest, which appear in four. Three of these hotspots (Madagascar, the Philippines, and the Caribbean) have small land areas, which further highlights their importance.

Two additional hotspots, the Tropical Andes and Mediterranean Basin, should be considered as candidates for conservation support in light of their exceptional totals for endemic plants (Tropical Andes ranks highest for both endemic plants and vertebrates, and the Mediterranean ranks third for endemic plants). Yet they do not appear in Table V because they ranked in the top ten in only two criteria listings. Similarly, Mesoamerica ranks second for endemic vertebrates (49% higher than the third-ranking Caribbean), but scores only tenth for endemic plants.

E. Higher Taxa Assessment

Quantitative analysis can be complemented by a qualitative evaluation of endemism among higher taxa such as families and genera. Yet sufficient sampling and exhaustive surveys are not available to measure the distribution of biodiversity overall. A top-down taxonomic approach, however, could compare the biodiversity of different areas using measures based on the number of higher taxa in each. For instance, family richness can often serve as a predictor of species richness for certain animal groups, such as mammals, amphibians, and reptiles.

F. Action Responses

To review, in just 1.4% of Earth’s land surface, there are 25 hotspots containing 44% of plant species and 35% of vertebrate species facing high risk of extinction. It is often estimated that, were the present mass extinction of species to proceed virtually unchecked, some-
where between one-third and two-thirds of all species could be eliminated within the foreseeable future. The hotspots analysis indicates that much of this threat could be countered through protection of the 25 identified hotspots.

An aggregate area of 800,767 km², 38% of the hotspots total, is now protected in parks and reserves. Some of these are little better than “paper parks” in offering a modicum of legal status. All are in urgent need of stronger safeguards, including those five hotspots that fall entirely within protected areas. The areas without any protection at all amount to 1.3 million km² or 62% of the hotspots total. In a few areas, new safeguards will not provide outright protection of a traditional sort because human settlements and other activities are well established. These areas could receive a measure of protection as “conservation units” that allow some degree of multiple use provided that species safeguards are always paramount. In short, the prospect of a mass extinction can be diminished and conservation efforts can be more effective by applying a hotspots strategy. The hotspots findings reported here complement several other priority-setting analyses. There is a 68% overlap with BirdLife International’s Endemic Bird Areas (Stattersfield et al., 1998), 82% with the World Conservation Union (IUCN)/World Wide Fund for Nature International Centres of Plant Diversity and Endemism (Davis et al., 1994–1997), and 92% with the most critical and endangered ecoregions of the World Wildlife Fund–US Global 200 List (Dinerstein et al., 1996). The hotspots approach is more comprehensive than the first two by combining five categories of species, and is more tightly focused than the third.

There are surely other hotspots that feature exceptional plant endemism and face serious threats but that are not sufficiently documented to meet the hotspots criteria. They include the Ethiopian Highlands, the Angola Escarpment, southeastern China, Taiwan, and the forests of the Albertine Rift in eastern Democratic Republic of Congo (formerly Zaire), southwestern Uganda, and northern Rwanda. Were these five areas to be added to the hotspots list, they would increase the plants endemics total by only a few percent.

In addition, there are a good many mini-hotspots.

One such is Queensland’s Wet Tropics and adjacent tropical forest tracts along the Queensland coast, which contains a host of endemic species, with an exceptionally high species/area ratio (around 1,200 endemic plants in less than 11,000 km²).

To reiterate a key point: the biodiversity hotspots are not the only areas that deserve priority treatment from conservation planners. Indeed, every country has its own biodiversity stocks, even if they are not as diverse or as concentrated as those of the major hotspots. We shall now look at three other criteria that warrant attention for conservation priority rankings.

V. WILDERNESS AREAS

There are a few tropical forest expanses known as “major tropical wilderness areas” (Mittermeier et al., 1998) or “good news” areas (Myers, 1988, 1990) that total some 6–7 million km² and feature concentrations of endemic species while retaining at least 75% of their primary vegetation. These areas also have fewer than five persons per square kilometer.

One is the island of New Guinea, which has around 15,000 endemic plants. Others include the Guyana Shield of northeastern Amazonia, the lowlands of western Amazonia, and the Congolian Forest, with a total of perhaps another 30,000 endemic plants. Were these regions part of a supplementary conservation strategy, they could increase the endemic plants total to almost 60% of all plant species in roughly 5% of Earth’s land surface.

Wilderness areas of all kinds, that is, not just biodiversity-rich areas, comprise nearly 90 million km² of little-disturbed land, or well over half of Earth’s land expanse. But when areas of rock, ice, and otherwise barren land are excluded, nearly three-quarters of all habitable land has been disturbed to a significant extent. It is unlikely that most wilderness areas in question will be settled by large human communities within the foreseeable future because of unfavorable climate, difficult terrain, remoteness from markets, and other factors that mean they can be readily conserved in a wilderness state. Thus these areas merit priority treatment from conservation planners. There is still opportunity in these wilderness areas for conservationists to get things right—notably in terms of land use planning—from the very start.

VI. HIGH-VALUE ECOSYSTEMS

Still another conservation strategy to be pursued in association with hotspots is to protect ecosystems of high value by reason of their exceptional abundance and concentrations of wildlife. These ecosystems contain large numbers of individual animals and large stocks of remarkable plants striking in appearance even though they comprise few species and few if any endemic species. A notable example is the Serengeti/Mara
ecosystem in northern Tanzania and southern Kenya, which has 4 million wildebeest, zebras, gazelles, and other large herbivores in an area of only 25,000 km² (all of these species are widespread elsewhere, though not in such extraordinary numbers). For comparison, the United States has roughly 20 million deer, moose, elk, pronghorn, caribou, and other large herbivores on 9 million km².

Other high-value ecosystems contain exceptional concentrations of endemic species, but only in a few categories. For instance, Lake Baikal has numerous endemic fish species but few species of other sorts, whether endemic or not, and Lakes Nakuru and Naivasha in Kenya possess exceptional numbers of bird species but few endemics, and has few other species of other types. Leading candidates among these high-value ecosystems include:

- the Serengeti/Mara ecosystem;
- the monarch butterfly overwintering sites in Mexico;
- the coastal Sundarbans area of India and Bangladesh, which has the largest tiger population left in the wild;
- Lake Baikal, with 2000-plus fish species (15% of all freshwater fishes), 1300 of them endemic;
- the East African Rift Valley lakes with 1200 fish species, 930 of which are endemic;
- Lakes Nakuru and Naivasha, which have 400-plus bird species each, and a joint total of 600 (compare to the United States total of 770); Lake Nakuru usually has 250,000 flamingoes and occasionally 2 million;
- the green turtle nesting grounds on Ascension Island in the Atlantic;
- the caribou migration lands in Alaska;
- the Galapagos Islands with their “museum of evolution”;
- the Great Barrier Reef with its outstanding coral reefs; and
- the California, ecosystems that harbor giant redwoods and sequoias.

VII. MEGADIVERSITY COUNTRIES

To effectively conserve a hotspot area, it is usually critical that the country’s government be committed to the conservation effort. As has been well said, hotspots do not have governments, only countries do. So a complementary approach to hotspot protection should focus on “megadiversity” countries. Such a country is defined as one that either (a) contains 20,000 higher plant species or, in the case of a country with fewer than 20,000 but more than 10,000 such species, at least 5000 endemics; or (b) contains at least 2000 species of higher vertebrates (mammals and birds), or 200 such species as endemics.

These 17 megadiversity countries encompass 60–70% of all global biodiversity (Mittermeier and Mittermeier, 1997) (Table VI). When these countries are assessed for their rankings in terms of plants, vertebrates (including freshwater fish), butterflies, and tiger beetles, the top three countries are in a class of their own, namely, Brazil, Indonesia, and Colombia. They are followed by a second group that includes Mexico, Australia, Madagascar, and Peru, and then a third group of China, the Philippines, India, Ecuador, and Venezuela. Clearly there is need to give priority attention to these 17 megadiversity countries as well as to the hotspots—though in many instances, the two lists overlap.

VIII. THE CONSERVATION IMPACT OF HOTSPOTS

The original hotspots strategy was first implemented in 1989 by the MacArthur Foundation with substantial funding for hotspots. Since then, over $400 million has been invested by international agencies and conservation groups. Yet this is only 0.8% of the total amount spent by governments during the same period on biodiversity conservation, (roughly $40 billion) and by international groups ($10 billion). These monies have been spent on across-the-board activities rather than on the tightly targeted efforts advocated here. This $400 million is almost twice as much as the cost of the Pathfinder mission to Mars, which along with many other space probes has been justified largely on biodiversity grounds, namely, the search for extraterrestrial life.

The hotspots could be adequately protected, and thus a large proportion of all species at risk, for just $20 million per hotspot per year (12.5 times the annual average over the past ten years). The traditional scattergun approach of much conservation activity, that is, seeking to be many things to many threatened species, needs to be complemented by a well-directed hotspots strategy that emphasizes the most cost-effective measures.

Such a tightly targeted strategy could generate a handsome payoff in stemming the biotic holocaust that is now under way. In the 25 identified hotspots, 35% of Earth’s nonfish vertebrate species and 44%
Table VI

<table>
<thead>
<tr>
<th>Country</th>
<th>Area (km²)</th>
<th>Total species</th>
<th>Endemics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>8,511,965</td>
<td>50,000–56,000</td>
<td>16,500–18,500</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1,916,600</td>
<td>37,000</td>
<td>14,800–18,500</td>
</tr>
<tr>
<td>Colombia</td>
<td>1,141,748</td>
<td>45,000–51,000</td>
<td>15,000–17,000</td>
</tr>
<tr>
<td>Mexico</td>
<td>1,072,544</td>
<td>18,000–30,000</td>
<td>10,000–15,000</td>
</tr>
<tr>
<td>Australia</td>
<td>7,686,810</td>
<td>15,638</td>
<td>14,458</td>
</tr>
<tr>
<td>Madagascar</td>
<td>587,045</td>
<td>11,000–12,000</td>
<td>8,600–9,600</td>
</tr>
<tr>
<td>China</td>
<td>9,561,000</td>
<td>27,100–30,000</td>
<td>~10,000</td>
</tr>
<tr>
<td>Philippines</td>
<td>302,780</td>
<td>8,000–11,000</td>
<td>3,800–6,000</td>
</tr>
<tr>
<td>India</td>
<td>3,287,782</td>
<td>&gt;17,000</td>
<td>7,023–7,873</td>
</tr>
<tr>
<td>Peru</td>
<td>1,289,210</td>
<td>18,000–20,000</td>
<td>9,336</td>
</tr>
<tr>
<td>Papua New Guinea</td>
<td>472,960</td>
<td>15,000–21,000</td>
<td>10,900–16,000</td>
</tr>
<tr>
<td>Ecuador</td>
<td>283,961</td>
<td>17,600–21,100</td>
<td>4,000–3,000</td>
</tr>
<tr>
<td>USA</td>
<td>9,372,433</td>
<td>18,916</td>
<td>4,036</td>
</tr>
<tr>
<td>Venezuela</td>
<td>912,090</td>
<td>15,000–21,070</td>
<td>3,000–8,000</td>
</tr>
<tr>
<td>Malaysia</td>
<td>320,749</td>
<td>15,000</td>
<td>6,500–8,000</td>
</tr>
<tr>
<td>South Africa</td>
<td>1,221,037</td>
<td>23,420</td>
<td>16,500</td>
</tr>
<tr>
<td>Dem. Rep. Congo/Zaire</td>
<td>2,344,000</td>
<td>11,000</td>
<td>3,200</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>51,189,393</td>
<td>155,475–183,025</td>
<td><strong>195,475–183,025</strong></td>
</tr>
</tbody>
</table>

Of plant species currently face unusually high risk of extinction. The hotspots analysis indicates that nearly half of the overall problem could be countered through protection of the 25 hotspots, covering an aggregate area the size of three Texases or one Mexico. In short, the likelihood of a mass extinction could be greatly reduced and made much more manageable. Leading players among the “global community” are international funding organizations (e.g., the World Bank and other multilateral banks, United Nations agencies, and bilateral aid agencies), international nongovernmental organizations (e.g., conservation bodies and private foundations), and business enterprises interested in biodiversity protection (e.g., pharmaceutical corporations). All of humanity has a stake in Earth’s biodiversity, and through a coordinated effort and directed actions we could make all contributions have maximum conservation impact.

Recall that the mass extinction of species, if allowed to persist, would constitute a problem far graver than all other environmental problems. We could clean up acid rain, turn back deserts, and repair the ozone layer within a matter of decades, regrow forests and restore topsoil within a century or so, and even stabilize the global climate within a few centuries. But according to evidence from mass extinctions in the prehistoric past, evolutionary processes are not likely to generate a replacement stock of species in less than 5 million years and possibly several times longer. Within the next few decades, we shall determine the future of a key feature of our biosphere, its abundance and diversity of species. The hotspots strategy offers a way to largely avoid an impoverishment of the Earth that could last at least 20 times longer than Homo sapiens has been a species.

Obviously, a mass extinction will also affect a large number of people. Suppose the average global population during the 5-million-year recovery period is 2.5 billion rather than the 6 billion we have today, on the grounds that the world’s future ecosystems will be unable to support the current global population. This means that the total number of people who will be affected will be on the order of 300 trillion individuals. This figure dwarfs the 30 billion people who have existed so far. Even one trillion is a large number. To put it in perspective, consider the length of time made up of 1 trillion seconds.

This, then, is the ultimate significance of the biotic holocaust that is overtaking the planet. Fortunately, we still have time to stem and slow the biodiversity crisis—and there are few ways to do it so successfully as by safeguarding the 25 hotspots that now harbor...
what nature has produced in its most exuberant expressions of life’s abundance and variety.

See Also the Following Articles

BIODIVERSITY, GLOBAL LEVEL • ENDERIM • INVERTEBRATES, TERRITRIAL, OVERVIEW • SPECIES AREA RELATIONSHIPS • HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW • VERTEBRATES, OVERVIEW

Bibliography


HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW

Paul R. Ehrlich and Claire Kremen
Stanford University

I. Introduction
II. Current Impacts
III. Signs of Ecosystem Deterioration
IV. Driving Forces
V. Solutions

GLOSSARY

ecosystem Community of organisms in a defined area combined with the physical factors of the environment with which the community interacts.
ecosystem goods Commodities, such as timber and seafood, supplied free to humanity by natural ecosystems.
ecosystem services Functions, such as flood control and pollination, supplied free to humanity by natural ecosystems.

TWO CENTRAL ISSUES CONCERNING BIODIVERSITY TODAY are the roles that plants, animals, fungi, and microorganisms play in the functioning of ecosystems, and the primarily negative impacts of human beings on those ecosystems. Here we focus on the latter, starting with the rise of humanity as a geophysical force and then examining in some detail current anthropogenic alteration of ecosystems and the prospects of further damage by human beings to the delivery of ecosystem goods and services. Finally, we look at ways in which human damage to ecosystems could be limited in the future.

I. INTRODUCTION

Ecosystems are the functional units composed of the elements of biodiversity (plants, animals, fungi, and microorganisms) in an area interacting with each other and their nonliving environments. Ecosystems are the “natural capital” that generates a flow of ecosystem goods and services that are essential to civilization. Natural ecosystems have been the source of all foods consumed by human beings, a large proportion of the medicines, and many important industrial products ranging from natural dyes to cotton and timber. In addition to these ecosystem goods, ecosystems also supply a wide array of ecosystem services. Among other services, ecosystems tend to buffer against sudden changes in the climate, control the hydrological cycle that brings us fresh water, control floods, generate and maintain fertile soils, dispose of wastes and recycle nutrients, pollinate crops, and control the vast majority of pests that would attack crops or carry disease to human beings.

Human impacts on ecosystems precede the dawn of history. Discernible impacts probably began when people (quite possibly Homo erectus) first began to use fire as a tool in hunting, but there are almost insurmountable difficulties in determining when that was—
perhaps more than a million years ago, perhaps as recently as 50,000 years ago. The skill of modern Homo sapiens as hunters is attested to by their probable role in the extermination of the so-called megafauna of the Pleistocene epoch (ice ages starting about a million years ago)—mammoths, cave bears, giant sloths, and the like. In dramatically changing the structure of communities of large animals, many of which were herbivores (plant-eaters), they thereby substantially altered plant communities and thus modified entire ecosystems. However, it was the agricultural revolution, some 10,000 years ago, that paved the way for the large-scale impacts that threaten the integrity of ecosystems (and thus of the critical services they supply to humanity) today. In the past 350 years, since the Industrial Revolution hit its stride, the scale of the human enterprise (as measured by overall energy use) has multiplied roughly 20 times, as the result of a 4-fold increase in population size and a 5-fold increase in per capita consumption. This, of course, has resulted in a roughly 20-fold impact of Homo sapiens on ecosystems.

II. CURRENT IMPACTS

Humanity has become a global geophysical force, altering Earth on a scale as large or larger than forces such as volcanism, erosion, natural oil seepage into oceans, and, especially, the natural extinction of populations and species of nonhuman organisms. Humans have now altered every cubic centimeter of the biosphere, at the very least by changing the climate and distributing synthetic organic chemicals, novel radionuclides, and particulate matter across the globe. This modification of ecosystems has been primarily mediated through its effects on their living elements.

A. Habitat Alteration

The most significant way in which people influence ecosystems is by degrading the habitats required by nonhuman organisms. Often this occurs simultaneously through a variety of channels, so that the types of degradation discussed here often overlap with one another. On rare occasions, however, habitat alteration may have inadvertent direct beneficial effects for people.

1. Deforestation

Forest destruction is probably the most widely recognized form of ecosystem degradation, since it so dramatically transforms the structure of the habitat. In the past few centuries about one-half of all tropical moist forests have disappeared, and in southeast Asia much of the lowland moist forest is already gone. Large areas of temperate forests have been cleared, but in some areas (such as the United States) there has been substantial regrowth. Many ecosystem services are degraded by deforestation, including flood control, maintaining the appropriate gaseous content of the atmosphere (especially through carbon sequestration), regulating the climate, control of vectors of disease, and provision of timber. The reduction of populations of dangerous predators such as tigers is a rare and debatable benefit to Homo sapiens.

2. Forest Fragmentation

The increasing patchiness of forests is receiving growing attention, since brute rates of deforestation underwrite the ecosystem impact of deforestation for a number of reasons. First, fragmentation leads to a loss of species diversity due to species-area effects. The number of species that a parcel of habitat can maintain is generally a function of its area. In general, 10 fragments averaging 1 km² each will retain less diversity than a single fragment of 10 km². Second, fragmentation increases the amount of forest edge relative to forest interior, making forest interior organisms more vulnerable to edge-specializing predators. For example, there is some evidence that predation on the eggs of ground-nesting temperate forest birds increases with closeness to edge. Third, fragmentation, especially of tropical moist forests, may lead to further forest destruction because of the penetration of winds and drying into much of the interior of small patches. Fourth, forest fragmentation can lead to secondary extinctions resulting from the loss of biotic interactions. For instance, the extermination of populations of required pollinators can lead to the loss of plants species still present in the fragments. And fifth, fragmentation often results in the loss of species from the fragments due to inbreeding depression (loss of genetic variability leading to lowered fertility or other deleterious effects) or chance (stochastic) extinctions stemming from the small size of residual populations.

3. Selective Harvesting of Forests

Highgrading is the removal of commercially valuable species while bypassing others, and this practice can dramatically alter the taxonomic composition and structure of forests. Removal of mahogany trees is a classic example, so is removal of key understory ele-
ments, such as the rattans now being extracted from the forests of Southeast Asia.

4. Fire Suppression

Intervention to prevent fires in forest and chaparral ecosystems that normally burn periodically leads to a buildup of fuels that eventually can result in uncontrollable and extremely hot fires that change the entire system by killing organisms that would ordinarily survive the smaller, cooler fires characteristic of the original fire regime.

5. Conversion to Farms

Clearing of natural ecosystems to introduce crop agriculture normally results in the near total destruction of the ecosystems. Virtually all of the larger native plant and terrestrial animal species are directly extirpated, and many of the smaller plants and animals (e.g., insects, reptiles, amphibians, birds, and bats) go extinct because of loss of habitat and changes in the microclimate.

6. Conversion of Grazing

Conversion of natural ecosystems to grazing can be as or more destructive as conversion to crop agriculture if, for example, large areas of tropical moist forest are cleared to make pastures. On the other hand, if a savanna ecosystem with abundant natural grazers simply has the composition of the grazing community changed (as when small numbers of cattle are added to the ungulate population of an East African savanna), the damage can be much less. Nonetheless, at a minimum the composition and structure of the flora are altered by the introduced herbivores, and often attempts are made to control potential predators on the grazing animals, which may also have cascading effects on the original flora and fauna.

7. Conversion to Infrastructure

The building of structures, highways, malls, and so forth competes with farming for having the greatest impacts per unit area on natural ecosystems. At best, only disconnected fragments of the original system remain, and the delivery of ecosystem services is almost totally disrupted. Conversion to infrastructure, with its destruction or covering of the soils that are normally essential to maintaining a terrestrial ecosystem, is one of the most irreversible forms of human impact on ecosystems. Division of habitat by roads is also a form of conversion to infrastructure. Roads cause impacts on ecosystems far in excess of those created by the road building itself. First, they may form barriers to the normal movement of animals. Second, they serve as conduits for the introduction of invasive organisms (including human beings). Third, dirt roads may influence the flora and fauna of adjacent areas by deposition of dust (which, for example, may function as an insecticide). And fourth, especially in tropical moist forest areas, roads increase penetration, settlement, and recreational use by human beings, all of which generate additional impacts.

8. Mining

Extraction of minerals varies in its ecosystem impacts, depending on the scale of the mines, whether they are underground, open pit, or placer (hydraulic/washing/dredging) operations, and the material mined. Large-scale open-pit mining tends to be totally destructive and expensive and very difficult to ameliorate. Underground mining usually has less impact, but can destroy aquatic systems over large areas through the runoff of toxic drainage from the mines, and it converts terrestrial systems into "slime pits" for mine wastes.

9. Damming and Other Water Development

Human attempts to intervene in the hydrological cycle have massive impacts on the ecosystems of rivers and streams by changing water flow patterns and temperatures and by blocking the movements of anadromous fishes (ones that breed in streams but spend much of their lives at sea). Conditions may be created that encourage the spread of disease, as in the increase of *Bilharzia* in Egypt after the building of the Aswan High Dam. This dam made possible the construction of irrigation canals that became ideal breeding grounds for the parasites that cause the disease. Terrestrial ecosystems, of course, are directly destroyed by inundation.

10. Wetland Drainage

The drying of wetlands eliminates the organisms and services supplied by estuaries, marshes, and bogs (flood control, water purification, and nursery services for commercially important fish species). Loss of wetland habitats also disrupts the migratory patterns of birds and reduces their food supply, including wildfowl populations valued by hunters.

11. Siltation of Onshore Ocean Waters

A flow of silt into the oceans results from timbering operations, construction, and other human activities that occur in coastal zones. In the tropics, siltation leads to the destruction of coral reef ecosystems by the
smothering of coral organisms. This in turn leads to the loss of local fisheries and ecotourism opportunities.

12. Toxins in Aquatic Systems
The toxification of onshore ocean waters, lakes, and streams may disrupt marine ecosystems and negatively impact fish populations (by direct poisoning, or alteration of food chains). Toxins may also render fish unfit for human consumption, thus disrupting the food supply service of aquatic ecosystems. Sewage dumping has converted several substantial areas of the oceans into "dead zones" in which oxygen levels are so low that most animals cannot live in them. Toxification, which has influenced the entire planet from pole to pole, is one of the most subtle forms of habitat alteration. Often toxic substances are active in concentrations that are difficult to detect, especially when they function by mimicking hormones that influence the early development of animals. The long-term ecosystemic effects of adding tens of thousands of synthetic organic chemicals to the global environment are simply unknown.

13. Ozone Depletion
Thinning of the ozone layer adds another toxic substance to Earth's ecosystems—an increased flux of dangerous ultraviolet B radiation (UV-B). The impact of such radiation can be judged from the fact that until some 400 million years ago life could not exist on the land surface because there was no ozone layer to screen out the UV-B. Until then, organisms had sheltered in aquatic ecosystems, because water rapidly reduces the UV-B flux. Ozone depletion has been most serious in the region of the "ozone hole" in the Antarctic, where impacts on oceanic ecosystems are already reported. As with other toxic substances, the long-term effects of moderate ozone depletion are difficult to predict. It has been postulated that increased UV-B or other toxic materials are involved in a global decline of amphibian populations, but the data are still too fragmentary to permit definitive judgments to be made.

14. Recreation
Recreational activities have diverse impacts on ecosystems. Off-road vehicles compact and erode soils and kill organisms that burrow in them, and are especially destructive in arid climates. Trails in parks can cause local erosion and disrupt the normal activities of wildlife. Boat anchors and scuba-diving activities contribute substantially to the destruction of coral reefs in some localities. Ecotourism, which can provide economic benefits for conservation, can also be destructive if not properly controlled. The outflow of sewage from coastal hotels pollutes the waters of offshore reefs. The disturbance of animal communities and native flora by tourist vehicles in some African game parks is another example of recreational impact.

15. Climate Change
Alteration of the climate, which may be induced or accelerated by human activities, is a major mechanism for altering ecosystems. Global warming appears to be already changing both terrestrial and aquatic ecosystems in the boreal regions. Perhaps the most serious threat is an apparent thinning by one-third of the sea ice in the Arctic. Shrinkage of the ice pack will have large, if difficult to predict, consequences for polar bears and other aquatic mammals and the food chains in which they participate. But the greatest threat of broadscale ecosystem impacts doubtless resides in the potential for changing the patterns of oceanic currents. For example, should the Gulf Stream off the U.S. Atlantic Coast fail or be deflected, all of the ecosystems of western Europe would be severely stressed, to say nothing of the human populations there.

Anthropogenic climate change could be many times faster than the relatively rapid natural changes that occurred following the last glaciation. It would certainly cause large-scale population extinctions in some areas, as already appears to have been detected in butterfly species in western North America and Europe. It might also lead to a substantial loss of species diversity. Much will depend on the speed and nature of the changes, and on the capability of organisms to migrate past extensive barriers created by near ubiquitous human development.

B. Alien Invasions
Alien invasions are little appreciated as engines of ecosystem modification, and yet they are one of the most important.

1. Islands
Many oceanic islands have lost most of their native floras and faunas to species transported by Homo sapiens. The destructive invaders include goats (which long ago played a major role in transforming the ecosystems of the Mediterranean basin), rats, mongooses, Philippine brown snakes (which devoured the avifauna of Guam), and mosquitoes (which carried bird malaria to Hawaii, resulting in the extinction of almost all species of native birds).
2. Continents

Continental ecosystems have also been highly modified by invasions. Most of the grasslands and woodland savannas of California are now nearly devoid of native herbaceous plants, having been substantially taken over by introduced annual grasses and herbaceous weeds from Europe.

3. Onshore Waters

Not only has the flora of California been greatly modified by exotics, but the coastal waters, especially around San Francisco Bay, are host to a myriad of invaders. Many of these arrived in the ballast water of ships arriving from all over the world. One, the eastern shipworm, was tolerant of the relatively dilute seawater of the Bay, as the native western shipworm was not. After it was introduced early in the twentieth century, it proceeded to destroy the wooden piers and bridges that laced the Bay, doing as much financial damage as the 1906 earthquake.

4. Disease Organisms

Pathogens moved about by human beings also greatly influence ecosystems. A classic case was the introduction of a virus disease of ruminant animals into northeast Africa around 1884. The disease raced south, reaching South Africa in 1896. It not only decimated the cattle herds of peoples like the Masai (causing severe famines), but also infected giraffes, buffalo, and wildbeest, causing a die-off that led to starvation of their natural predators. One consequence was that lions switched to eating people, so that farmers deserted their land. In the absence of farmers and native browsing animals, brush and woodlands invaded grasslands, altering the entire ecosystem. Some resistance developed in the animals, and the situation has fluctuated ever since, with rinderpest remaining a factor in East African ecosystems.

C. Overexploitation of Ecosystem Components

Habitat alteration and invasions are not the only ways in which human activities produce impacts on ecosystems. Another important pathway is the overexploitation of economically desirable organisms, which in turn alters habitats.

1. Food

Often overexploitation occurs when natural populations are harvested for food. A classic example was the persecution of the passenger pigeon, which once was the most abundant bird in North America. One flock was estimated to contain two billion birds. Flocks moved around, breeding when large crops of beechnuts, acorns, and other forest nuts were found. They were commercially hunted to extinction to supply food markets in eastern cities; the last individual died in a zoo in 1914. It has been suggested that their absence altered the eastern forest ecosystems and greatly increased the food supply of rodents, which are now implicated in the cycle of lyme disease. If true, the ecosystemic effect of their extermination had an impact on humanity beyond the loss of a once abundant and easily harvested source of animal protein.

Fishing often alters ecosystems by disrupting normal oceanic food chains, especially when there is overfishing and when there is heavy “bycatch” of nontargeted animals (some 27 million metric tons of marine life are killed and discarded by oceanic fishing fleets, equivalent to about a quarter of the total catch retained, and this does not account for the heavy local mortality caused by dynamite and cyanide fishing on coral reefs). Increasingly, trawls dragged along the bottom are destroying the biotic structure there, with unknown future consequences.

2. Natural Products

Demand for natural products can also lead to overexploitation with accompanying ecosystem alteration, as discussed earlier under forest clearing and fragmentation. But it is often difficult to measure the impact, for example, of the harvesting of specialty woods by “high-grading.” Again, far more damage to the forest is sustained by the building of roads, the associated felling of nontarget trees, and the transport of harvested trunks by “skidding” compared to the small amount of damage from the selective removal of valuable hardwoods. Similarly, although the demand for sea turtle shell and soup has had serious consequences for turtle populations, we know little about the ways oceanic ecosystems may be altered by the reduction of turtle populations.

3. Esthetic Resources

The demand for pets frequently leads to the overexploitation of natural populations. Some of the cyanide fishing on coral reefs is aimed at aquarium fishes rather than fishes sold to grace the plates of rich Hong Kong businessmen. Freshwater tropical fishes are also often overexploited for the aquarium trade. The ecosystemic changes being caused by the latter have not been seriously investigated. The pet bird trade has large-scale impacts on some species and in
III. SIGNS OF ECOSYSTEM DETERIORATION

If humanity is causing substantial impacts on ecosystems, can we detect what they are? In many cases, the answer is yes, although precise measurements are difficult to obtain, mainly because baseline measurements were never taken before impacts occurred. Most, if not all, of the observed and measurable impacts are negative; positive changes have been observed largely in cases where the pressure has been lightened (e.g., the recovery of forests in parts of the eastern United States after failing farms were abandoned).

Perhaps the most direct way to measure ecosystem deterioration is to measure an area's biological productivity (net primary productivity, NPP), if possible over time. If an area is converted from a natural to a human-directed system (forest to pasture; meadow to farm fields), NPP usually falls. A further decline in NPP (for instance, crop yields) over time indicates deterioration. Such declines, even in the face of increased fertilizer applications, have been widely observed, and it is well known that fertilizer use can mask losses of micronutrients from soil and declines in soil texture for considerable periods before declining productivity becomes apparent. In such situations, not only is crop or forest productivity lost, but the ecosystem services of soil replenishment and nutrient/waste recycling (which are also a vital part of the planet's biogeochemical cycles that control the distribution of nitrogen, oxygen, carbon, and water, among others) are usually damaged.

A. Faltering Food Production

The decline of food production may be the most significant trend in which human impacts on ecosystems play a role. Following unprecedented rates of growth in food production after the middle of the century, the global grain harvest has failed to increase on a per capita basis since 1984. Since cereal grains represent the human feeding base and are equal in weight to roughly half of all foodstuffs produced by agriculture, trends in grain harvests are the best indicator of food supplies and availability, although economic factors and changes in eating habits are also significant. About a third of the world grain harvest is used for feeding livestock, and shortages in grain supplies for human food are often compensated by reducing its use as feed. Grain shortages often cause rises in price on the global grain market, which in turn cause the prices of meat and other animal products to rise and consumer demand to fall. Similarly, reduced demand for animal products, as has occurred in some industrial nations in recent years, leads to reduced demand for feedgrains. In the United States, demand for animal products has shifted away from beef to poultry, which can be grown with about a third as much grain per pound of meat produced. In the industrial countries, more than half the grain consumed is used for feed. In developing nations, by contrast, most grain is directly consumed by people. But in some countries (especially China) this is changing, and demand for animal foods is rising rapidly.

Despite these shifting economic and consumption patterns, the overall trend of food production has been increasingly problematic in recent years, with aggregate production falling to keep abreast of population growth, which itself has slowed markedly in the past decade. Production in sub-Saharan Africa has fallen especially far behind. The reasons are many and vary from area to area. They include the Green Revolution running out of steam, diminishing returns from fertilizer applications, decline in crop genetic diversity, loss of prime farmland to urbanization, and land degradation from poorly managed irrigation. Among those of rising importance are underlying trends of land degradation and the related loss of
ecosystem services. It may well be that potential food production has actually kept up with population growth. But key issues are: How much higher would that potential be if important ecosystem services had been carefully conserved? And can humanity depend on linear extrapolations of recent production trends over either a 50-year or decadal timescale?

Of the ecosystem services that are being impaired or lost, some are basically local—soil erosion and depletion, loss of natural pest and crop disease controls, and loss of pollination services are obvious examples. Others are regional, such as the consequences of damage to nearby watersheds. Thus deforestation of a watershed area leaves adjacent or downstream areas more vulnerable to floods and droughts. The destruction by Hurricane Mitch of a substantial portion of the agricultural capacity of Central America in the fall of 1998 is a case in point. There also is evidence that maintaining natural areas interspersed among intensively farmed areas helps to preserve soil fertility as well as provide sources of natural pest control, pollinators, and impediments to the spread of diseases.

The Green Revolution, along with agricultural expansion, a dramatic increase in irrigated lands, and other factors, allowed grain production to rise by more than 73% from 1960 to 1980, while the population expanded by 46%. But from 1980 to 1997, grain production rose only 31%, just equal to the population's growth of 31%. Up to 1984, grain production stayed ahead of population growth, reaching a per capita peak then of 342 kg; since then it has fluctuated between 301 and 335 kg per capita.

The degree to which human impacts on ecosystem services are involved in the difficulties of agriculture is difficult to evaluate. Some crop damage by pests is clearly due to the diminution of natural pest control services, but pesticide use and cultural practices make up for some of the loss. Similarly, declines in natural pollinator services are occurring, but data are lacking on the degree to which exotic pollinators (honeybees in many areas) are able to take up the slack. Loss of flood control services, due primarily to deforestation of watersheds, often results in serious damage to crops in places as diverse as the Indian subcontinent and Central America. Soil loss to erosion—perhaps 25 billion tons annually—in excess of soil regeneration—also reduces agricultural production. And the anthropogenic modification of ecosystems is depleting the genetic library of crop relatives, crop pollinators, and crop biocontrol agents, which are all critical to maintaining crops in their coevolutionary races with diseases and pests.

It is important to remember that agriculture is largely dependent on the weather, and that climate stabilization is a vital service of ecosystems. Thus, for example, severe damage to agriculture in Central America from Hurricane Mitch may have been exacerbated by a decline in this service if anthropogenic global warming played a role in the unusual behavior of the storm. Floods and droughts, and changes in the timing of the onset of seasonal warming trends or rains, will obviously have major impacts on agriculture around the world.

B. Fisheries Decline

Fisheries yields tell an even more dismal story—and at sea there is no doubt that yields have not been keeping pace with demographic expansion. Some two-thirds of the world's major fisheries are being maximally harvested today or are in decline. A per capita peak in fisheries yield was reached in 1988 at 17.2 kg per capita, but yields have failed to meet that level for the past decade. Most of the blame for declining yields can be placed on overharvesting, yet more systemic environmental damage has also played a role through the pollution, modification, and destruction of estuaries, coral reefs, mangrove fringes, and coastal wetlands. Anadromous fish (such as salmon) have suffered from the damming of rivers and the siltation from bank erosion. Major oil spills have taken a toll on seabirds, marine mammals, and shellfish, causing lingering damage to coastal fauna in many areas. Aquaculture (fish farming) harvests, while offsetting declines in traditional fishery yields, have caused new pollution and habitat destruction problems that affect natural fish populations, and they rely extensively on fish products for feed. Thus aquaculture tends to produce a net loss in fish biomass. It also creates a rising demand for feedgrains and other agricultural products to support their yields, thus competing, along with livestock operations, with human beings for the food they raise.

C. Change in the Epidemiological Environment

Disruptions in the epidemiological status quo may be another sign of ecosystem modification. Anthropogenic ecosystem changes are clearly threatening the deterioration of the climate stabilization service. Those changes may have played a role in the warming that has allowed the mosquitoes that transmit dengue fever to move northward in North America. This warming is expected to have similar effects on the wider occurrence of malaria and other tropical diseases.
D. Decline in Water Quality

Increasing water pollution is often a direct reflection of the modification of ecosystems. A classic example is that of the New York City water supply. Around 1900, New York City's water was of such high quality that its was bottled and sold throughout New England. Recently the U.S. Environmental Protection Agency has notified the city that its water had dropped below acceptable quality. The reason for this decline was the degradation of the Catskill Mountains watershed that served the city—fertilizer runoff and inadequate local sewage treatment had impaired the natural water-supply service. It was estimated that constructing a plant to treat the Catskill water would have a capital cost of $6–8 billion, with roughly a $300 million annual maintenance cost. In contrast, restoring the functioning of the natural Catskill ecosystem would cost only $1–1.5 billion. This is a case of ecosystem modification and the loss of essential ecosystem services.

E. Other Symptoms

There are, of course, many other readily observable symptoms of ecosystem deterioration caused by human action, some of which have already been mentioned. On land they include flooding, landslides and heavy soil erosion, changes in microclimate, the local disappearance of sensitive species of plants and animals, silting of streams, and lowered water tables. A general decline of amphibians appears to be occurring over much of the world, and may be caused by multiple disruptions from many causes. In oceanic systems, changes in the mix and abundance of animal, plant (e.g., kelp and seagrass), and planktonic species are also observable and often, as in cases of algal blooms or precipitously declining fisheries yields, are signals of serious disruption.

IV. DRIVING FORCES

The impacts of human beings on ecosystems are a product of three multiplicative factors: population size, affluence (or per capita consumption), and the technologies and social-political-economic relations that supply the consumption. These relationships are often condensed for convenience into the "IPAT equation"

$$\text{Impact} = \text{Population} \times \text{Affluence} \times \text{Technology} = P \times A \times T$$

If one uses per capita energy consumption as a surrogate for $A \times T$, the result is that quoted in the beginning of this article—a 20-fold increase in the human impact on ecosystems over the last 150 years.

A. Overpopulation and Continued Population Growth

Population growth is probably the single most important factor leading to impacts on natural ecosystems. Early in the last century there were only 1 billion people. That number had doubled to 2 billion around 1930, doubled again to 4 billion in 1975, and reached 6 billion in 1999. In other words, it took on the order of 300,000 years for the population of Homo sapiens to reach a billion, a single century to add the second billion, 30 years to add the third billion, 15 years for the fourth, 12 years for the fifth, and another 12 years for the sixth. Simply adding so many people to the population required gigantic interventions in natural ecosystems in order to bring sufficient land under cultivation to feed them. This effect was exacerbated because each individual added to the population tended to have a disproportionate ecosystemic impact. Naturally people farmed the most fertile soils first, got their water from the nearest sources, mined the most concentrated ores, and so on. More people required the farming of more marginal land, and thus more land per person. The transportation of water over greater distances involved using more water per person to make up for losses in storage and transport, and the disruption of more area per person for dams and other infrastructure. The mining of poorer ores means that more ecosystemic destruction by mining was caused per person, since more ore must be dug up to provide the same level of metal use for each additional individual.

Since at least 1950 the entire world clearly has been overpopulated by the simple standard that the human population could not be sustained by the flow of resources generated by incoming sunlight mediated through farms, forests, the hydrological cycle, and so forth—the agricultural and natural ecosystems of Earth. Instead of preserving humanity's "natural capital," it has been necessary to expend it. Three elements of that capital are especially important. The first is deep, rich agricultural soils, which are generated on a timescale of centimeters per millennium and are now in many areas being destroyed at a rate of centimeters per decade. The second is aquifers, many of which were last filled during the ice ages and which are now, in many if not most areas of the world, being pumped at many times the natural recharge rate. The third element is biodiver-
sity, the species of microorganisms, plants, and animals
that are working parts of ecosystems, which is now
being exterminated at a rate that is unprecedented in
the past 65,000 years—at perhaps 1000 times its natural
regeneration from speciation.

Furthermore, the high rate of addition to the human
population in the second half of the twentieth century
meant that little concern could be shown for the possible
long-term effects of the interventions designed to
keep people eating. Programs such as the “Green Revo-
lation” (the exportation of high-yield agriculture from
rich temperate zone nations to poor tropical ones) were
driven in part by the fear of massive starvation, and
their consequences for natural ecosystems and the vital
services they provide were largely ignored. So were their
social impacts.

B. Overconsumption

Overconsumption (consumption fat in excess of basic
needs) is almost as important as overpopulation as a
driver of ecosystem modification. For example, a sig-
nificant proportion of the ecosystem modification
cau sed by the agricultural enterprise is created by grow-
ing grain and feeding it to animals destined for human
consumption. This is the case even though in rich soci-
ties the consumption of products containing animal fat (and, perhaps, overconsumption of animal protein)
causes serious public health problems. The directly
toxic and hormone-mimicking effluents from the pro-
duction of myriad products, many of them plastic or
packaged in plastic, also are doubtless having eco-
systemic effects, although data beyond the effects on single
species are lacking. But perhaps the most serious over-
consumption problem derives from converting societies
to dependence on the automobile, with the marketing and purchase of cars that are much too large,
heavy, and inefficient for the purposes to which they are put.

C. Use of Environmentally
Malign Technologies

The use of technologies that are unnecessarily environ-
mentally malign is related to overconsumption. The
use of inefficient automobiles is mirrored in the use of
unnecessarily inefficient lighting, heating, appliances,
and so on. But the basic technological problem in soci-
ey at the end of the twentieth century is the persistent
overdependence on fossil fuels, and especially coal, as
an energy source.

D. Faulty Economic Arrangements

Market failures are partly to blame for the overdepen-
dence on fossil fuels, because the market prices of those
fuels do not come close to reflecting their social costs.
For example, the costs incurred by society in the form of
lung disease from the effluents created by fossil fuel
burning are not factored into the market price. More
importantly, neither are the present and potential costs
to society of climate change induced by carbon dioxide
emissions from fossil fuel use. If these and other social
costs (such as the pollution due to oil spills and the
discarding of lubricants used in automobile main-
tenance, other emissions, and medical costs deriving from
automobile accidents) were captured in market prices,
the price of gasoline in the United States might be
quadrupled. This could have the beneficial result that
fuel-efficient cars would be produced and then more
people could use them, in the process creating less
pollution and reducing the demand for environmentally
destructive oil extraction.

E. Faulty Social Arrangements

Social arrangements often work against the protection
of ecosystems. For instance, society has not designed
educational systems to apprise its citizens of the bio-
physical and social realities that it faces. Few people,
even “well-educated” people in rich societies, can pro-
vide a coherent explanation of ecosystem services or
the threats to their delivery. Indeed there is a general
failure of the media and standard educational system
to make explicit the connections among the many fac-
tors and effects of human alteration of ecosystems.

F. Faulty Political Arrangements

Political biases and errors also play a substantial role
in endangering ecosystems. Societies generally lack
“foresight” institutions that have the capacity and inde-
pendence to analyze complex problems and provide
competent and independent advice to government and
the general public. In addition, there is the age-old
problem of disparities in power that now have increas-
ingly widespread—often global—implications for the
sustainability of society. Powerful supporters of politi-
cians in Rome, before its collapse, were hard-pressed
to weaken the Empire. Modern business interests who
press the politicians on their payrolls to take no actions
to ameliorate global warming could theoretically bring
down our current civilization.
V. SOLUTIONS

In light of the myriad uncertainties about both human and ecosystem behavior, it would seem prudent to try to reduce the three factors in the IPAT equation simultaneously, rather than risk a possible catastrophic collapse of services as *Homo sapiens* progressively alters natural ecosystems. What sorts of steps might be taken to avoid such a collapse? Some are proximate—steps that would directly address impacts on ecosystems. Others are ultimate—steps that would deal with the basic drivers of ecosystem deterioration.

A. Proximate Solutions

A nonexhaustive list of proximate steps might include:

1. Limiting development so that a minimum of the remaining relatively undisturbed ecosystems would be destroyed. In virtually every country where new infrastructure is needed, it should be provided as far as possible by redeveloping intensively areas that are already highly disturbed.
2. Agreements to slow climate change need to be solidified rapidly, despite the grave political and economic difficulties of doing so. The international protocol to maintain the ozone layer should be reexamined in light of recent data to see that its performance is adequate.
3. Establishment and protection of reserves designed to protect relatively undisturbed ecosystems and the elements of biodiversity they contain. This should only be done when it is possible to integrate the interests of local people into the reserve design—otherwise in the long term the effort is likely to be wasted as population growth pushes people increasingly into reserves that they do not value as such.
4. Limit toxic releases by shifting the burden of proof to those who would claim that the introduction of a novel compound into use (and thus into the environment) carries benefits that clearly exceed the social costs.
5. Sustainable-yield harvesting should be enforced for all living resources. Harvesting systems should be designed with substantial attention to the precautionary principle (err on the side of conservatism) in order to buffer society against the consequences of overoptimistic estimates of what is sustainable.
6. Apply countryside biogeography principles. Countryside biogeography is the nascent field that is attempting to develop principles for making already disturbed areas most hospitable for biodiversity and thus more likely to provide necessary ecosystem services. One step in doing this would be to pass (or strengthen) laws to protect endangered populations, species, and landscapes, and to restore moderately degraded areas, including restoration with useful, native species.

B. Ultimate Solutions

Steps to deal with the ultimate drivers of ecosystem deterioration must address the factors of the IPAT equation.

1. Population Reduction

Population growth should be halted as soon as humanly possible, and then a slow decline can be initiated toward a level that can be sustained indefinitely at whatever average level of consumption is selected. Selection of that level, which is closely related to the issue of optimum population size, need not be debated in the near future. Population size will change with available technologies and preferences. In the foreseeable future the human population will be living above the world’s long-term carrying capacity, so that a goal of simply halting growth is likely to be valid for at least a half century. After this period, there will be abundant time for scientific study and public debate over where a decline should be halted. Human beings, as far as can be told, have always intervened to adjust population sizes to needs perceived primarily through the situations of reproducing couples. In recent decades there has been a shift to substantial consideration of the needs of society as well, and that trend should be accelerated in the future.

2. Consumption Control

Consumption control will be a more difficult task than controlling population growth, if current trends and attitudes are any guide. This is an area fraught with difficult issues of equity and feasibility. For instance, if the gasoline tax in the United States were raised to the point where gasoline was as expensive as in Europe, a substantial decline in fuel consumption would be likely. But there would be a deleterious impact on poor people who must commute to work by car. Compensatory tax changes would be needed to lighten that burden.

Not only do many of the things that human beings now consume seem unnecessarily harmful, but the quantities consumed per person’s often seem outrageous. But since one person’s outrage is triggered by the use of something that another values greatly, it is
3. Substitution of Technologies

Furthermore, it is possible to substitute environmentally benign technologies for currently used environmentally malign technologies. For example, with proper education, natural pest control services, enhanced by integrated pest management, can be substituted in agriculture for the broadcast spraying of toxic pesticides. Buildings can be equipped with "more expensive" energy-saving devices, such as video conferencing and telecommunications, that can be increasingly used to eliminate energy-expensive travel to meetings and workplaces and would pay for themselves through energy savings within one to five years.

4. Revising Socioeconomic and Political Systems

All of the foregoing steps would benefit from public education and broad social discussion of the need to, as the economists say, "get the prices right." That is to say, the prices of goods and services need to be adjusted so they more closely reflect the social costs of each. In other words, external costs should be internalized. The heavy burden placed on society in environmental deterioration and public health from, for example, the overuse of large automobiles should be reflected in the prices of both the vehicles and the fuel they consume.

Success at evaluating and implementing the steps that can and should be taken in various nations and in diverse economic strata within nations may depend on substantial changes in governmental institutions and the generation of broad social consensus. Some revision of socioeconomic and political systems is likely to be required if humanity is to preserve the ecosystem services upon which it ultimately depends, and the biodiversity upon which the systems themselves depend.

See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • DEFORESTATION AND LAND CLEARING • DISEASES, CONSERVATION AND • ENERGY USE (HUMAN) • FISHING, COMMERCIAL • ISLAND BIOGEOGRAPHY • LAND USE ISSUES • PLANT INVASIONS • POLLUTION, OVERVIEW • POPULATION STABILIZATION (HUMAN)

Bibliography

HUMAN IMPACT ON BIODIVERSITY, OVERVIEW

Leslie E. Sponsel
University of Hawaii

I. Principles
II. Biodiversity Modification
III. Biodiversity Conversion
IV. Biodiversity Commodification
V. Biodiversity Toxification
VI. Biodiversity Futures
VII. Conclusions

GLOSSARY

cultural ecology Analysis of how culture influences the interactions between a human population and the ecosystems in which they reside; also called ecological anthropology.
culture System of socially learned, shared, and patterned ideas, institutions, behaviors, and their material products that distinguishes a particular society.
diversity principle General geographical coincidence between high concentrations of both biological and cultural diversity, usually in the tropics.
ecological transition Tendency for societies to be in growing disequilibrium with their biophysical environment as they increasingly deplete natural resources and degrade their habitat, thereby reaching new thresholds of environmental impact.
historical ecology Transdisciplinary and diachronic analysis of how human societies and ecosystems change and in turn transform one another through time in local and regional landscapes. Data are drawn from geology, archaeology, history, and other sources.
swiddening Umbrella term including diverse types of horticulture in which a small section of forest is cut and burned to plant crops in a temporary garden; shifting cultivation is used as a synonym, but slash-and-burn cultivation is now considered to be a pejorative term.

THROUGHOUT HUMAN PREHISTORY AND HISTORY, HUMAN IMPACTS ON BIODIVERSITY have reached progressively higher thresholds. Most likely the net impact of humanity has been to reduce biodiversity. However, at the population level, the types and magnitudes of human impacts on biodiversity vary tremendously through time and space, depending on the specifics of the particular context. Many societies have decreased local biodiversity, whereas many have sustained or even increased it. Nevertheless, because of...
the widespread direct and indirect impacts of humans on biodiversity, any consideration of biodiversity must also assess the possibilities of human influence.

I. PRINCIPLES

A. Human and Environmental Diversity

Not all humans are equal in their impacts on biodiversity. While it is widely recognized that there is tremendous diversity in nature, there is also considerable diversity in humanity. Nearly 7000 distinctive cultures exist today and there were many times more in the past, probably on the order of at least tens of thousands over some four million years of human existence. Accordingly, there is tremendous diversity in human relationships with and impacts on biodiversity. In examining human impacts, it is imperative to consider the diversity of humans through time (prehistory through history) and space (cultures and regions).

Environmental diversity is another variable in determining the human impact on biodiversity. Some environments are simply much more fragile and vulnerable than others, especially those that are relatively simpler, such as the arctic and deserts, and those that are isolated with a high proportion of endemic species, such as oceanic islands. Also, some regions are simply much less accessible or hospitable to humans than others, including the poles, mountain heights, deserts, and deep oceans. In these cases, the label “wilderness” may still apply—pristine nature that is largely unaffected by humans.

Furthermore, given the dynamism of ecosystems and ecological processes, together with the widespread impact of humans on them, it is misleading to consider most environments as pristine, virgin, primeval, or wilderness. The condition of any ecosystem is the cumulative product of previous conditions, usually including human impacts. The more important question is not whether or not a human society had or has any environmental impact, but its particular forms and magnitude, and if negative then the extent to which it is reversible and allows natural regeneration within a normal period of time.

In studying human impact it cannot be assumed that the previous environment was in a state of equilibrium or climax stage in ecological succession. Furthermore, even when humans are a factor in environmental change, it cannot be assumed that other agencies, such as natural alterations in climate or weather and natural fires, are not also contributing to change.

In any case, it is unrealistic to consider biodiversity anywhere without also considering the possibilities of human influence. Human impact is inevitable because, ultimately, biodiversity is the primary or raw natural resource that all societies rely on for their subsistence and economy. Also, biodiversity is the biotic component of ecosystems and ecological processes on which human survival, adaptation, and welfare ultimately depend.

B. Types of Impact

Human impact on biodiversity, direct or indirect, involves four basic factors: (1) over-exploitation of natural resources; (2) habitat modification, conversion, and fragmentation; (3) the introduction of exotic (nonnative) species; and (4) pollution. Any one of these four factors may influence ecosystem composition, structure, and function; ecological processes; and biodiversity, especially through species extinction. Documented extinctions during the last 400 years include 484 animal species and 654 plant species, but this must be only a fraction of the actual total (Heywood, 1995: 198).

However, in some circumstances the second and third factors may increase biodiversity locally, because some components of biodiversity are anthropogenic (human-caused). In Britain, for example, there has been a net increase in mammalian diversity as 21 of its 49 species are introductions (Heywood, 1995: 757).

Human populations may influence biodiversity at the genetic, population, species, ecosystem, biome, and/or biosphere levels of the ecological spectrum. Because species as well as ecosystems are interrelated and interdependent, human impact on any one species or ecosystem may influence others as well, much like a chain reaction.

C. Species and Population Levels

Globally humankind has increasingly tended toward greater levels of ecological disequilibrium, that is, with—many (but not all) populations exceeding the carrying capacity of their habitat, depleting resources, degrading environments, and thereby threatening and reducing biodiversity. There have been at least seven successive thresholds of environmental impact in cultural evolution, their inception approximated in years before present (b.p.) and varying regionally: the use of fire by foragers (hunter-gatherers) (0.5 million), farming (5000–10,000), cities (5000–6000), exponential growth in the human population (1000), European colonialism (1500), industrialization (150–200), and globalization (50). These thresholds are part of a continuum
or gradual process, although an accelerating one. (These thresholds are somewhat analogous to the states of water as ice, liquid, and gas. The substance—H₂O—as the same, but a gradual quantitative increase in temperature leads to major qualitative changes in its physical properties.) In the case of socio-cultural evolution, scale of complexity is a pivotal factor (Bodley, 2000) (see Table 1).

At the population level, the situation is complicated. Long ago some societies developed relatively sustainable economies that did not markedly endanger or reduce biodiversity, as evidenced by their longevity in prehistoric and/or historic records. Some societies even increased biodiversity. Furthermore, the same society may increase biodiversity in certain ways, yet decrease it in others. Thus, human impact varies in time and space as well as with different species, ecosystems, and biomes. Consequently, any generalizations about human impact must be carefully scrutinized with attention to temporal and geographic scales as well as the specific details of the particular cultural and ecological contexts. It follows that a holistic anthropological approach, one emphasizing cultural and historical ecology, is indispensable for assessing the possibilities of human impact on biodiversity in any context. It could even be argued that cultural and historical ecology are indispensable for understanding biodiversity itself, because it is so rarely free from some degree of human influence, at least indirect. Indeed, biodiversity as a concept is a Western cultural construct developed at a particular stage in history, even if the associated physical phenomena are undeniably real (e.g., Soulé and Lease, 1994).

**TABLE I**

<table>
<thead>
<tr>
<th>General Trends in Sociocultural Evolution Correlated with Increasing Thresholds of Human Impact on Biodiversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Population—nomadic to sedentary settlement pattern with increasing population density, nucleation (settlements to cities), and pollution.</td>
</tr>
<tr>
<td>2. Food—wild to domesticated foods with shift from foraging to farming</td>
</tr>
<tr>
<td>3. Energy—somatic (human and animal) to extrasomatic (water, wind, wood, fossil fuel, nuclear) sources of energy for work</td>
</tr>
<tr>
<td>4. Land—intensive (agricultural) to extensive (horticultural) land use; land tenure—community/public to private/corporate ownership</td>
</tr>
<tr>
<td>5. Economy—subsistence (satisfying basic physiological needs) to market surplus production to materialist consumerism; local self-sufficient to regionally and then globally interdependent economy (globalization)</td>
</tr>
<tr>
<td>6. Waste—organic products that are readily biodegradable to more recently those like metals and plastics that disintegrate very slowly</td>
</tr>
<tr>
<td>7. Scale—small and decentralized to large and centralized societies (states), the latter with increasing import of natural resources from ecosystems in distant regions</td>
</tr>
<tr>
<td>8. Differentiation—egalitarian to hierarchical (stratified) societies, the latter with increasing inequality in access to resources, goods, and services and institutionized warfare</td>
</tr>
<tr>
<td>9. Alienation—daily to occasional contact with and feedback (monitoring human impact) from the natural environment, eventually with alienation from nature and other humans, and increasingly decisions made by agents far removed from the locations they affect</td>
</tr>
<tr>
<td>10. Worldviews—ecocentric to anthropocentric and egocentric worldviews, attitudes, and values; also sacred/moral to secular/amoral and utilitarian orientation to nature; may include shift from biophilia (love of nature) to biophobia (fear of nature)</td>
</tr>
<tr>
<td>11. Balance—some degree of dynamic ecological equilibrium with recognition of limits to increasing disequilibrium with assumption that there are no limits (i.e., ecological transition)</td>
</tr>
<tr>
<td>12. Impact—environmental modification to conversion (natural to cultural landscapes) and fragmentation (remnant patches of nature); also toxification with industrialization, local to global impact on biodiversity and environments</td>
</tr>
</tbody>
</table>
D. Ambivalence

Biologists and others have been ambivalent about the human species. Most would readily consider Homo sapiens to be a part of nature as a product of biological evolution, but apart from nature in terms of ecology. However, in much of the past and in many cases to this day, human populations are simply one component in the dynamics of ecosystems and ecological processes. The idea that humans are unique and apart from nature may derive from the anthropocentrism (human centeredness) of Western civilization and of Christianity, Judaism, and Islam. (Anthropocentrism is also a contributing cause of the environmental crisis.) Actually, every species is unique as a closed genetic system and in other respects. However, it may be valid and useful to view H. sapiens as the dominant or most important keystone species in many if not most ecosystems. (A keystone species is one that has an unusually important influence on other species and the ecosystem. See the special issue on human-dominated ecosystems of Science, July 25, 1997).

An important example of such anthropocentrism is the common view that human impact on the environment is necessarily an unnatural disturbance. Yet all organisms affect their environment in various degrees and ways; consider the various impacts of trees, fungi, earthworms, leaf-cutter ants, bees, woodpeckers, kangaroos, bats, elephants, plankton, starfish, and sharks (e.g., Westbroek, 1991). However, non-human organisms that alter their environment are rarely considered a disturbance, except for exotic species introduced by humans. Usually only when an organism depletes resources and degrades the habitat beyond its natural capacities or normal period for regeneration might the term disturbance be appropriately applied. Yet it could be argued that some human societies are unnatural or even anti-nature, especially industrial ones. Other societies appear to be an integral part of nature, such as many more traditional indigenous cultures in the Amazon forests like the Yanomami. The impact of such societies on their environment is usually no less natural than that of other species. This was likely the case for the majority of societies through most of human existence.

II. BIODIVERSITY MODIFICATION

A. Extinction Hypothesis

Human antiquity markedly varies among regions, and consequently so does cumulative impact. Hominids (the human line) evolved in Africa some four to six million years before the present (B.P.). Human dispersal into other regions is approximately dated (B.P.) as follows: Europe and Asia (1 million), Australia and New Guinea (50,000–60,000), the Americas (12,000–20,000), and Pacific and other islands (30,000–to recent, depending on the island).

As humans colonized new regions, supposedly they would have a great advantage over prey that lacked previous experience with them. Relatively suddenly the hunters became a new top carnivore in the food web, rather than through the usual gradual process of predator-prey co-evolution. Furthermore, their technology provided advantages, particularly projectile weapons like the spear, that allow some safe distance from and surprise for prey when making a kill. Indeed, Paul S. Martin and others hypothesize that the dispersal of humans into new areas caused massive faunal extinctions around 50,000 B.P. or less, depending on the region, especially in Australia, the Americas, New Zealand, Oceania, and Madagascar. The main basis for the so-called “blitzkrieg hypothesis” is the apparent coincidence in timing of human arrival and massive megafaunal extinctions. However, with little to support it but circumstantial evidence, this hypothesis remains far from conclusively proven, and it is much more problematic and controversial than advocates usually admit. For example, convincing evidence is lacking in Australia for a temporal coincidence between human dispersal and megafaunal extinctions. Yet even if temporal coincidence were demonstrated, that does not automatically prove a cause-effect relationship nor exclude other potential causal factors such as climatic change. In New Zealand, however, there is no doubt that over-hunting and habitat destruction by indigenous people (Maori) caused the extinction of many animal species, including the large moa birds (see Flannery, 1995).

Whether examining faunal extinctions or any other aspect of human impact on biodiversity, human diversity must be taken into account. Even in especially vulnerable ecosystems, different cultures may have very different impacts, such as indigenous peoples and subsequent colonists in Hawaii. Hawaiians introduced 34 exotic species and supposedly caused the extinction of some 50 endemic species during their 1500 years of exclusive occupation. In contrast, Europeans, Asian, and other colonizers introduced 4653 exotic species, caused the extinction of 211 endemic species, and endangered more than 800 others, all in a little over two centuries (e.g., Culliney, 1998). Consequently, compared to native Hawaiians, subsequent colonizers caused 137 times as many introductions and...
four times as many extinctions, all in a mere one-seventh of the time.

In general, traditional hunter-gatherers and others may seem to have a very limited impact on their environment, given their low population density, high mobility, limited technology, subsistence economy, minimal needs and wants, intimate environmental knowledge and monitoring, and animistic worldview, attitudes, and values that are nature oriented. Nevertheless, even small bands of nomadic foragers can have some impact, for example, on seed dispersal; one way archaeologists identify prehistoric occupation sites such as hunting camps is by distinctive combinations and concentrations of useful plant species that would not likely occur naturally.

B. Fire

Natural fires have influenced ecological systems and processes for many millions of years. In contrast, so far evidence indicates that the human use of fire extends back only about a half million years in the Old World to the later portion of the period of Homo erectus. Usually the main reason hunters burn an area is to create fresh plant growth to attract game. In addition, low-intensity fires set by indigenes prevent the accumulation of fuels that might otherwise lead to devastating wildfires, especially during a dry season, drought, or from lightning strikes. (This is a lesson only recently learned by national park managers in the United States, Australia, and elsewhere.) Controlled burning of different patches of the landscape at different times by indigenes or others may help increase biodiversity, whereas wildfires may decrease it.

The long-term cumulative effect of burning by foragers can be substantial, as in the case of the so-called fire-stick farming by the Aborigines of Australia over some 50,000 years (Flannery, 1995). Repeated burning of large areas of vegetation over long periods would eventually significantly modify the plant community and consequently also the associated animal community. Sometimes a pyrosere is created, a biotic community resistant or adapted to fire that is prevented from further ecological succession or development by repeated burning. With the regular and widespread use of fire by foragers, surely a new threshold in impact was reached. Some grasslands in temperate and tropical zones, such as the Scottish highlands and Venezuelan llanos, respectively, may be the result of repeated burning by humans at least in part, although in many areas this conclusion is still controversial and climatic change may be a factor as well (see S. J. Pyne, in Baleé, 1998).

C. Horticulture

Fire is also an important component of farming technology. In tropical forests, during the dry season a plot of about a hectare or so of vegetation is cut and then, after drying out in the sun, it is burned. The organic ash from the burn provides fertilizer for the growth of crops that are planted just before the rainy season. This ash can be critical because tropical soils are usually poor. The crops are often pioneer or weed species that thrive in disturbed areas. Swiddening in some form is ubiquitous throughout tropical forests, but was also practiced by pioneer farmers in Europe and America.

Swiddening may enhance biodiversity. This practice creates a gap in the forest that provides sunshine for crops. They are harvested for a few years and then less intensively as productivity declines with decreasing soil fertility and increasing problems with weed competition and pests. Over years or even decades, the swidden gradually converts into fallow, with successional plants growing from seeds left in the soil and entering from surrounding forest through agencies like wind and birds. Consequently, swiddening leads to a mosaic of plant and associated animal communities at different stages of succession. Swiddening also forms ecotones (transition zones between two environments) that may increase biodiversity (edge effect) by harboring some species from both environments and others specializing on the ecotone. Such environmental heterogeneity increases the potential for higher biodiversity. Gardens and fallows also attract game and other animals, given the concentration of vegetation on the ground that provides cover and a concentration of edible plants, unlike a primary forest in the tropics.

Swiddening enhances biodiversity at the genetic level as well, since most swiddens are polycrops (numerous species and varieties) of domesticated plants. Crop diversity reduces the risk of total failure as a result of weather, disease, or pests, because some species and varieties are more resistant or resilient.

As a result of the variety of human uses and manipulations of forests, it can not simply be assumed that any are entirely pristine. In the Amazon it is estimated that nearly 12% of the forests beyond the floodplains are to some degree anthropogenic (Baleé, 1994). For some two decades researchers at La Selva, the famous tropical biology research station in Costa Rica, thought that they were studying natural forest. They were unaware that it was to some extent anthropogenic until the recent discovery of charcoal and pottery fragments in the soils. In Africa, conservationists assumed that forest patches in savanna grasslands were relics of formerly more
poor mainly from the northeast into the Amazon Basin. The Trans-Amazon highway as a way to relocate rural situations occurred when the Brazilian government built Java to those with low density like Irian Jaya. A similar from islands with high human population density like Indonesian government in its transmigration program subsistence. This is the case with people moved by the lack of familiarity with farming in a tropical forest, until the encroachment of Western civilization and particularly in recent decades. However, in most areas traditional conditions no longer hold.

On the other hand, there are cases where swiddening has a negative impact. As an example, changes in plant species diversity over the last 14,000 years in Panama are preserved in phytoliths (plant fossils) from lake sediments. A sharp decrease in this diversity coincides with the onset of swiddening around 7000 B.P. and may be causally related (Piperno, 1994). Thus, human impact on biodiversity from swiddening and other activities needs to be assessed on a case-by-case basis.

Still, traditional shifting cultivators contrast markedly with shifted cultivators of recent decades, the latter being immigrants who colonize tropical forests, often as economic refugees. Pioneer swiddeners are much more likely to contribute to deforestation because of their lack of familiarity with farming in a tropical forest and orientation to a market economy rather than mainly subsistence. This is the case with people moved by the Indonesian government in its transmigration program from islands with high human population density like Java to those with low density like Irian Jaya. A similar situation occurred when the Brazilian government built the Trans-Amazon highway as a way to relocate rural poor mainly from the northeast into the Amazon Basin.

III. BIODIVERSITY CONVERSION

A. Agriculture

The shift from human modification to the conversion of environments and biodiversity pivots on the decline of rotational systems of land and resource use. Traditional foragers, herders, and swiddeners usually tend to maintain sustainable economies by rotating one or more of these factors: types of resources used as well as places, times, techniques, and personnel in resource exploitation. If rotation decreases, then human impact increases markedly and eventually conversion occurs instead of modification, reflecting an emphasis on intensive rather than extensive use of land. By various estimates 30–50% of the land surface of the planet has been converted by humans.

Agriculture usually involves some kind of plow pulled by a domesticated animal, in contrast to horticulture, which usually involves only hand tools like an ax and digging stick powered solely by human muscle. Agriculture is a more intensive and permanent form of land use focused on conversion of the environment, whereas horticulture is a more extensive but temporary or rotational form of land use focused on modification of the environment. Accordingly, agriculture usually simplifies ecosystems and reduces biodiversity, whereas horticulture may sustain or even increase environmental heterogeneity and biodiversity. Agriculture also creates more environmental fragmentation than horticulture, that is, the dividing up of landscapes into isolated or semi-isolated patches of "natural" environments such as forest remnants in farm fields. However, whether horticulture or agriculture, it is necessary to assess the impact on biodiversity on a case-by-case basis.

B. Rice Paddies

Throughout Asia, large portions of the landscape, especially in the lowlands near rivers and streams, were converted centuries or millennia ago to wet rice paddies. Paddies can also be found in highlands where terracing has been developed. Conversion continues to this day, driven by increasing population and economic pressures on land and resources; it is accelerating in many areas. Conversion includes numerous types of forests, grasslands, and wetlands, and accordingly sacrifices an enormous amount of biodiversity. For instance, various types of forest covered about 70% of Thailand until World War II, whereas today only about 13% of the country is forested, largely because of agricultural expansion and logging.

Rice paddies are not simple monocrops, however. There may be hundreds or even thousands of varieties of rice in a region. It is estimated that some 50,000 local varieties of rice existed in India until recently. In addition, there may be several hundred species of wild plants and animals associated with paddies, many of them considered edible. Beyond the paddies there are home gardens, fruit orchards, pastures, various types of forest, and agroforestry (tree crops), all of which contribute considerable biodiversity to the regional
HUMAN IMPACT ON BIODIVERSITY, OVERVIEW

landscape. For instance, in parts of Bali, home gardens may contain as many as a hundred or more species of wild and domesticated plants.

C. Raised Fields

Extensive areas of raised or ridged fields are found in many parts of the world, usually along river floodplains or in other wetlands. They elevate land above water for farming. Soil is dug and piled on top of either side of a ditch to form a long mound. The ditch becomes a canal for water drainage and irrigation. On floodplains the ridges are arranged in parallel fashion perpendicular or at an angle to the river. Organic debris and sediments from the canals are periodically dredged and placed on top of the mound as a rich organic fertilizer. Fish, waterfowl, and other aquatic organisms thrive in the canals and are harvested for protein in the diet. Raised fields may increase productivity and biodiversity beyond the level that would otherwise exist in the region.

D. Vavilov Centers

Vavilov centers are concentrations of genetic and morphological diversity remaining at the several foci of plant and animal domestication from the Neolithic some 5000–10,000 B.P., the exact period depending on the region. These centers, as concentrations of landraces and their wild ancestors, are in effect in situ gene banks. (Landraces are species and varieties of domesticated plants and animals that have been genetically improved by traditional farmers and herders, but remain uninfluenced by modern breeding technology.) Prehistoric and historic farmers and herders engineered significant biodiversity in their crops and livestock, in part to reduce risk. For instance, some 8 species and 3000 varieties of potatoes were cultivated in the Andes. As another example, today a single species of sheep (Ovis aries) includes more than 800 different breeds, many quite ancient (Heywood, 1995: 111, 775). Vavilov centers are increasingly endangered and degraded by growing population and economic pressures. Thus, attempts at ex situ conservation of their biodiversity may be the main hope for preserving samples. The loss of diversity in the varieties of domesticated species and their ancestors severely limits the potential for adaptive responses to future environmental changes and perturbations, whether natural or anthropogenic.

E. Population

It has been estimated that through agriculture and other activities, humans preempt some 40% of the earth's total primary biological production annually (Vitousek et al., 1986). This magnitude of resource use surely diminishes the possibilities for many other species and the planet's biodiversity as a whole. Given projections for continued human population growth in this century before any global stabilization, surely this co-option of energy and nutrient sources from other species will get much worse.

The human population explosion is a relatively recent phenomenon. After some four million years of human evolution, the world population had only reached 300 million by A.D. 1000. Then world population grew to 500 million by 1500, 1 billion by 1900, and recently topped 6 billion. It is projected to be 10 billion by 2050. Correlated with this growth is increasing population density. For example, generally in the tropics human population densities (individuals/km²) are: foragers, <1; swiddeners, dozens; and wet rice farmers, hundreds to thousands. However, though increases in the size and density of the human population certainly increase impacts on biodiversity, not all humans create the same impact, and some are grossly disproportionate. For example, upper- and middle-income consumers use resources and produce pollution at levels many times higher than lower-income consumers and the poor, whether comparing economic classes within a single society or so-called developed and less-developed countries—the core and periphery in the world economic system, respectively. The wealthy nations, as well as petroleum and pharmaceutical companies, have a special responsibility to reinvest some of their profits to help fund biodiversity studies and conservation.

IV. BIODIVERSITY COMMODIFICATION

A. Cities

Market economies and cities are largely responsible for the commodification of nature—the commercial or monetary evaluation of biota and landscapes. Because most urbanites and suburbanites do not produce their own food, they must trade or purchase it and other natural resources as imports from the farmers and peasants working in the vast rural hinterland. Thus, commodification has been developing with urbanization for at least 5000–6000 years. Today about half of humanity lives in an urban environment, although this concentration, the megalopolis, and suburban sprawl are recent phenomena. However, it is not always easy to detect a clear boundary between urban, suburban, rural, and...
During the last 500 years, Europeans introduced plants and animals from the Old World into their overseas colonies, especially areas that were climatically similar to parts of Europe, like portions of Burma, California, Chile, Kenya, and South Africa. This Europeanization of local ecosystems reduced native biota and created neo-Europes. The biological exchange also flowed in the opposite direction, however, from the New World (Americas) to the Old (Europe and beyond), with the spread of crops originally domesticated by prehistoric native Americans, such as potatoes, corn, squash, tomatoes, tobacco, cotton, and manioc. Indeed, today much of the world’s food supply comes from introduced species (Crosby, 1986).

Biodiversity at the microbial level is another aspect of the so-called “Columbian exchange,” the biological and cultural consequences of the Old World and New World on one another. Indeed, throughout the Americas epidemics of introduced Old World diseases caused catastrophic depopulation of indigenous communities. Because these populations had been isolated from the Old World for thousands of years, they had no previous experience with the Old World diseases to develop immunological resistance. The devastating crash of their populations relieved pressure on their land, resource base, and local biodiversity. In many areas, forests and other environments regenerated rapidly, indicating that indigenes had not irreversibly degraded them. However, the apparent demographic void and supposed wilderness encouraged European colonization of what appeared to them to be an underpopulated and underused frontier. This process continues to this day in parts of the Amazon and elsewhere (Denevan, 1992). The Columbian exchange was certainly an unprecedented event in the history of human impact on biodiversity, however, it did not operate on any prior pristine nature. Long before European contact, the peoples of the Americas modified and in some areas converted many ecosystems (Denevan, 1992). For instance, in some areas of the ancient Mayan civilization there was extensive deforestation. As another example, anthropogenic prehistoric shell middens (mounds) in riverine and coastal zones may have higher diversity and density of useful plant species than adjacent “natural” sites. Europeans are not the only colonials, just those most familiar. All empires are built in part on geographic expansion through some combination of long-distance trade, military conquest, and domination and oppression (economic, political, cultural, religious) for the exploitation of the land, resources, and labor of other societies. Some examples are the Aztec in Mexico; the Inca in the Andes; Arabs in many parts of Africa, Europe, and Asia; and Chinese in much of East and South-
east Asia. All empires affect local biodiversity through biotic exchanges and environmental modification and conversion. For instance, in Asia the geographical distribution of some plant species is inexplicable without considering the role of sacred species such as the bodhi or pipal tree (Ficus religiosa) in Buddhism and temple yards.

C. Modern Transportation and International Commerce

Long-distance ocean travel and trade became possible with the development of seaworthy ships some 6000 B.C. and eventually facilitated European colonialism and widespread biotic introductions. However, intentional and unintentional introductions of species between formerly independent or isolated regions of the world have greatly accelerated with the advent of modern transportation and international commerce. Now more than 600 species of animals and plants are directly threatened by illegal international trade that involves several billion dollars annually.

The introduction of exotic species can endanger native species and ecosystems. For example, the introduction of the South American otter (Lontra scapulata) into southern Louisiana in the 1940s had profound consequences on the landscape and ecology of the Mississippi delta. The inadvertent introduction of the brown tree snake (Boiga irregularis) from Papua New Guinea to the island of Guam in Micronesia caused the extirpation (local extinction) of a dozen species of native birds and still threatens other fauna. The zebra mussel (Dreissena polymorpha) accidentally spread from Europe into the Mississippi River basin of North America when a tanker dumped ballast waters in the Great Lakes. Among other problems, this exotic mussel now threatens numerous native mussel species. Many other examples could be given, for more than 4500 exotic species have become established in the United States during this century alone. The effect of species introductions worldwide probably results in a net loss of global biodiversity. However, in some areas, such as New Zealand and Hawaii, net biodiversity has actually increased as the result of numerous introductions of exotic species, even though they variously threaten, endanger, or extirpate many native and endemic species.

Through European colonialism, Enlightenment ideals of rationalism and individualism, industrialization, capitalism, economic development, modernization, and globalization, the world’s environment, land, and natural resources, including biodiversity, have progressively become objectified and commodified. All of this has greatly facilitated the overexploitation, degradation, and destruction of biodiversity throughout the world and in its last frontiers such as tropical forests. In turn, this trend has been magnified by the insatiable greed that is inherent in the competition and growth mania of capitalist economies and the associated culture of materialism and consumerism.

This situation is compounded by the fact that the rich are getting richer at the expense of the poor who are becoming poorer. Wealthy entrepreneurs, multinational corporations, and so-called developed countries can afford to buy greater access and rights to biodiversity, just as they also exert a disproportionate impact on it. The poor are able and willing to pay less attention and money for biodiversity conservation than the rich, not necessarily because they are ignorant or unconcerned, but because they have fewer economic assets and alternatives. During the 1997 economic crisis in Southeast Asia, the harvesting and export of wildlife species and their products accelerated markedly as local people turned to their forests to generate desperately needed cash, as their regular farm crops and jobs no longer provided adequate household income. Material wealth and technology are concentrated in the developed countries in the Northern Hemisphere, whereas population growth, poverty, and biological wealth are concentrated in the so-called less-developed and developing countries of the Southern Hemisphere. Such inequalities are a serious obstacle to biodiversity conservation as well as human well-being.

D. Commercial Farming Industry

The beginnings of various agribusiness industries can be traced back to European colonialism. Most of these industries cause the massive conversion of forests and other ecosystems to monocrop plantations, such as in the tropics with bananas, cocoa, coconut, coffee, cotton, eucalyptus, oil palms, pineapple, rubber trees, sugar cane, tea, and tobacco. As a result, higher thresholds of environmental impact have been reached with extensive biodiversity loss at the genetic, species, and ecosystem levels.

Since the 1980s, the so-called Green Revolution and other forces of globalization have been threatening and diminishing the genetic and species diversity of domesticated plants and animals that have been developing since the Neolithic period. Monocrops from the Green Revolution are genetically engineered for rapid growth and high productivity in response to massive subsidies of chemicals as fertilizers, herbicides, and pesticides.
However, because of the high costs of inputs, mechanization, and transportation, these agribusiness industries are grossly inefficient even if highly productive. Such monocrops are also much more vulnerable than the diverse agroecosystems of traditional farms to climatic changes and weather perturbations, such as global warming and El Niño. Nevertheless, modern agribusiness, biotechnology and genetic engineering, and bio-prospecting are currently revolutionizing food production among other things, but with largely unknown yet probably far-reaching consequences for the maintenance of biodiversity.

The forests and woodlands of the world have declined significantly in area since the origin and spread of agriculture some 10,000 years ago. However, deforestation rates have surged dramatically in recent centuries, and especially in recent decades in the tropics. Although tropical forests cover only about 6% of the earth’s surface, they contain about half of all its biodiversity; thus they are the recent and future hot spots for biodiversity research and conservation. These forests are also increasingly becoming the hot spots of sociopolitical conflict and violence, such as in the Democratic Republic of Congo, Rwanda, Burundi, and Sierra Leone. Many temperate forests are deteriorating from loss of old growth, acid rain damage, and other forces. However, the forest cover in New England has expanded as the farm economy declined over the last 100 years, a hopeful indication of nature’s capacity for regeneration.

The combination of causes and the magnitude of deforestation and consequent biodiversity erosion have varied tremendously through time and space depending on the specific details of the particular context. The impact of a stone ax on tropical forest compared to a metal one is quite different, even allowing for cumulative impact with substantial human antiquity in an area. The introduction of more efficient metal tools and a market economy to trade local natural resources for Western manufactured goods have accelerated deforestation throughout the tropics. However, in recent decades newer technologies, including chain saws and bulldozers, together with new economic enterprises and incentives, created unprecedented rates of deforestation. In many areas of the Brazilian Amazon, for instance, cattle ranching made possible by government tax support and subsequent land speculation has been the major cause of deforestation since the 1960s. In much of Central America, forests were converted to pasture for cattle to feed beef for the fast food industry in North America and Europe, the so-called hamburger connection. Banana plantations have been another major cause of deforestation in Central America. An important factor in deforestation in many areas of Southeast Asia and elsewhere is the alienation of local communities from their land and resources by state government. Forests have been converted to monocrop plantations of eucalyptus and other fast-growing trees for export to Japan and elsewhere to supply the paper pulp industry.

Aquaculture is yet another kind of commercial and industrial farming. It has been endangering mangrove forests and other coastal ecosystems and their biodiversity throughout the tropics in recent decades. Shrimp pond farms along the southern peninsula of Thailand have grown exponentially with the investment of outside capital and technology, with the benefits going mostly to outsiders. Although aquaculture is a very ancient, productive, and efficient method for producing quality protein in many parts of Asia and elsewhere, where it has traditionally been integrated with other aspects of the economy and ecology, this new economic development is very different. The shrimp are usually much too expensive to be consumed by local populations. Instead, they are exported to distant markets in Japan, Taiwan, North America, and Europe. Yet shrimp ponds are short-lived because waste products accumulate in the sediments until they become toxic to the shrimp or for any other subsequent uses. Consequently, repeatedly aging ponds are abandoned and new ones are constructed. At the same time, this cancerous growth of shrimp farms is degrading and destroying the local economies and biodiversity of many coastal regions.

V. BIODIVERSITY TOXIFICATION

A. Industry

At least since 1962, with the publication of Rachel Carson’s classic book Silent Spring, there has been growing concern about the environmental impact of chemicals and pollution. Because in both ecosystems and the biosphere, ultimately everything is connected to everything else in some way, chemicals from agriculture, factories, cars, and other sources eventually circulate worldwide. Residues of DDT and other pollutants are even found deposited in layers of polar ice. The accumulation of nitrates and other nutrients from agricultural and household chemicals can generate the explosive growth of algae in lakes and rivers to the detriment of biodiversity, a process called eutrophication. Massive kills of birds, fish, and other species have been reported from this and other contaminants such as oil spills.
One recent possible symptom of widespread pollution, whether direct through water and soil contamination or indirect through increased ultraviolet radiation from the ozone hole, may be the gross deformities observed in frogs, a most alarming phenomenon and perhaps an early warning of impending ecological catastrophe. Another result of atmospheric pollutants such as automobile exhaust is the greenhouse effect, which in the twenty-first century may trigger global warming of a few degrees and the consequent rise of sea level by up to two meters or more. The consequences of sea level rise on the biodiversity of coastal and marine ecosystems like wetlands and coral reefs are uncertain, but are likely to be negative and potentially catastrophic. In terms of high biodiversity, coral reefs are the marine analog of tropical rain forests.

B. Oil

Industrial society, whose lifeblood is oil and other fossil fuels that are actually the remains of ancient biodiversity, has produced unprecedented types and levels of chemical pollution that undoubtedly endanger and erode living biodiversity. To illustrate, in recent decades many remaining frontier zones like the Amazon of Ecuador and Peru have been the target for oil exploration and extraction. Because of the isolation of these frontiers, the usual environmental safeguards and measures for clean-up are ignored. During the exploration phase, hundreds of miles of roads and large grids of extensive trails for seismic testing are cut into the forest. The seismic explosions not only scatter wildlife, but the resulting shock waves can kill hundreds of fish in rivers, lakes, and wetlands. During the production phase, a single oil well platform consumes about six acres of forest and about 2000 trees. (In the rain forest a single giant tree may be inhabited by thousands of insect and other species, most unknown to science.) Adjacent production wastes and treatment chemicals amount to millions of gallons every day for decades. As well sites there is no proper disposal of toxic waste, only open-air pits that eventually overflow into the soils, groundwaters, and surface waters, thereby contaminating and killing fish and wildlife. One gallon of oil can kill the fish living in a million gallons of water and adversely affect aquatic life at concentrations as low as one part per hundred billion. Over decades, many hundreds of oil wells and extensive pipelines allow for numerous leaks and spills of black crude when breaks occur through metal aging or earthquakes. The magnitude of pollution that has been occurring in frontiers like the Amazon makes the oil spills associated with the Exxon Valdez and Gulf War look like minor irritants in comparison! The costs to biodiversity are incalculable. Furthermore, beyond the corporate irresponsibility and massive environmental destruction, in the Amazon oil is even being pursued in wildlife reserves and other areas supposedly set aside for conservation by national governments (Kimberling, 1991).

C. Militarization

A grossly neglected type of human impact on biodiversity and the environment are military activities and warfare, even though armament production is the top industry in the world with more than $800 billion in sales annually. One aspect of warfare that is especially detrimental to biodiversity is the use of scorched earth tactics, which unfortunately are nothing new. One of the largest applications of this tactic was the U.S. military’s dumping of some 13 million gallons of Agent Orange to defoliate forests in the Vietnam War during 1962–1971. In recent decades, the U.S. war on drug production in Amazon forests has also involved the use of defoliants. During the Gulf War, the government of Iraq set fire to oil wells in the desert of Kuwait and created oil spills along the coasts. The tolls of such tactics on local and regional biodiversity have yet to be fully revealed. The rise of nationalism and ethnic conflicts with the end of the Cold War has spread militarization into frontier, border, and other zones. Conflicts and wars are likely to proliferate in the future, according to Mal-thusian pessimists and others who consider growing resource scarcity and competition as major contributing causes. If so, then national and international security would be well served by redirecting a significant portion of funds that now go to military and defense into biodiversity and environmental studies and conservation programs.

Whereas war may threaten and erode biodiversity, peace may promote it. For instance, the Demilitarized Zone (DMZ) between North Korea and South Korea is a corridor 4 kilometers wide and 250 kilometers long that extends across the peninsula. For nearly five decades this corridor has been rigidly enforced as a no-man’s-land. As a consequence, farmlands thousands of years old and degraded forests have both reverted to a natural condition, thus protecting threatened and endangered species of plants and animals, as well as a cross section of the ecosystems of the Korean peninsula.
Environmental agencies have been making some significant progress on resolving environmental problems and managing the human impact on biodiversity. Nevertheless, there are many serious limitations on the efficacy of government-protected areas for biodiversity conservation. First, they comprise only about 5% of the earth's surface, which is not a very large or representative sample of the tremendous biodiversity of the planet. Second, they are often little more than “paper parks” because of inadequate funding and administration. Third, they will come under increasing attack in many ways with accelerating population and economic pressure in the future. Fourth, even for the some 10,000 sites that are supposedly protected by governments, only 5% of these have been thoroughly inventoried for biodiversity, and it would take the equivalent of the current number of experts several centuries to inventory the remainder. Clearly there is an enormous challenge for inventoring, managing, and conserving biodiversity. Indigenous and other local communities can make a significant contribution to biodiversity studies and conservation, and thereby also have a positive impact on biodiversity. This is just beginning to be appreciated; an example is INBio (Instituto Nacional de Biodiversidad), the biodiversity inventory and conservation program in Costa Rica that employs local people as parataxonomists. Also, though today there are some 1600 botanic gardens in the world that help conserve plant diversity ex situ, there are nearly 7000 distinct cultures in the world and each may somehow contribute to ex situ and in situ conservation. Comanagement is usually the ideal, that is, the cooperative sharing between community and government in the design, authority, responsibility, and benefits of natural resource management and biodiversity conservation projects. Among the successes in comanagement are Manu National Park in the Peruvian Amazon and Kakadu National Park in northern Australia. However, when government administrators of protected areas ignore the needs of local people, then conservation efforts usually falter or fail.

Numerous people have grown increasingly skeptical that government, science, technology, and education are sufficient for resolving the spiraling environmental crises facing the world. They view such attempts as treating superficial symptoms of the crises, not the broader underlying causes. Some of these people are turning to their own religion as a source of worldviews, inspiration, motivation, attitudes, and values for developing a more sustainable and meaningful relationship with nature. As the pioneers in spiritual ecology, indigenes can also provide profound insights for such endeavors.

VI. BIODIVERSITY FUTURES

A. The Certain Future

As those from the “green” sociopolitical movement and many others recognize, industrial society, capitalism, and economic development are based on the false assumption that infinite growth is possible on a finite base. Here growth refers to both population and economy. Base refers essentially to carrying capacity. The latter includes not only the ability of the land and natural resources to support a certain level of population without resource depletion and environmental degradation, but also the capacity of ecosystems and the planet as a whole to absorb pollution and other anthropogenic stresses. Increasingly human impacts exceed the resilience of nature to regenerate and recover within a normal time period from any perturbations and stresses, natural or anthropogenic.

Until the demonstrably eccidal ideas and practices of modern industrial society and related factors are corrected and ecosanity with some modicum of ecological balance is restored, the net impact of humans on biodiversity will be negative. Accordingly, all life, including that of humanity, will remain endangered. There isn't much room for optimism, given the great momentum of population and economic growth combined with political pressures for so-called economic development and the elevation of the standard of living throughout the world, all at the expense of the environment and its natural capital of resources, biodiversity, and ecosystem services. Indeed, the ozone hole, greenhouse warming, acid rain, collapse of oceanic and other fisheries, soil erosion, desertification, and other global environmental problems may be symptoms of the failure of the experiment of industrial society after just two centuries. No human society is infallible and eternal—the archaeological and historical records provide many examples of those that became maladaptive, collapsed, and disintegrated like Rapa Nui (Easter Island), Harappa in the Indus Valley, or ancient Greece and Rome (Ponting, 1991; Redman, 1990).

B. Limits of Government

National and international governmental and non-governmental agencies have been making some significant
C. Indigenous Potential

Indigenous cultures contrast sharply with industrial and other “modern” cultures. Many indigenous societies, especially those that are more traditional, may provide heuristic models for biodiversity conservation through their intimate environmental knowledge (ethnoecology), sustainable economy, natural resource management and conservation practices, spiritual ecology, and protection of sacred places. (To a large extent this is because most indigenes lie to the left on the continua described in Table 1.) In these and other respects, it could be argued that many indigenous societies are actually more developed than industrial ones!

Regarding traditional environmental knowledge (TEK), for example, the Ka’apor people in the Brazilian Amazon recognize at least 768 species of plants from seed to reproductive adult stages. This reflects a tremendous amount and depth of knowledge about the biodiversity in their habitat, and that example is just in their domain of useful plants (Balee, 1994). Indeed, such knowledge of indigenous and other peoples may provide biologists and conservationists with one desperately needed short-cut for the inventory and conservation of local biodiversity.

Indigenes often consider their environment, including the biotic and abiotic components, to be sacred. Such a worldview can encompass a deep respect and reverence for nature that tempers their cultural ecology and resource use, and that may lead to environmental conservation, inadvertently if not intentionally. The role of spiritual ecology and sacred places in biodiversity conservation is just beginning to be recognized and explored by those who have an open mind to such phenomena.

Most indigenous and numerous other religions are nature centered, considering certain areas of “nature” in their habitat to be the foci of spiritual power. These sacred places are treated with extraordinary care and respect, and frequently taboos restricting resource use are associated with them. However, throughout the world aggressive Christian missionization and other dominating monotheistic religions have often destroyed sacred places and indigenous religions because they were perceived as pagan; these religion-inspired actions threatened and eroded biodiversity as well. The objectification and commodification of nature also degrade and destroy sacred places, as Australian Aborigines have found with Western mining activities. Nevertheless, numerous and diverse sacred places in nature remain and may have contributed to biodiversity conservation in the past and/or could do so in the future.

D. Conservationists and/or Exterminators

The preceding discussion does not support the romantic idea that all indigenes are always in perfect harmony with their environment, the myth of the so-called “ecologically noble savage,” which amounts to little more than a “straw man” argument. Instead, it recognizes the fact that many indigenous societies were to some degree environmentally friendly and that some still are, even though others are not or were not. These variants of the human impact on biodiversity have been documented in numerous cases by cultural and historical ecologists. Revisionist advocates who attack indigenous societies as environmentally destructive have yet to realize, let alone adequately resolve, the basic contradiction in their argument—how indigenes can be so knowledgeable about their habitats and interact with and monitor it on a daily basis, yet be so ignorant of, or amenable to, such destructive practices.

A related fallacy is that all humans are environmentally destructive, the so-called “Homo devastans” or “humans as the exterminator species” view (Balee, 1998). Such simplistic either–or, all-or-nothing, always-or-never thinking is misleading at best, but it is a surprisingly common deficiency of advocates of this view. Blaming all of humankind for a negative impact on biodiversity is simply scientifically inaccurate, sloppy scholarship, and professionally irresponsible. It needs to be emphasized again that humanity is diverse and so is its impact on biodiversity—some societies decrease it, others sustain it, some enhance it, and others affect it in some combination of these directions.

E. Diversity Principle

In general, several authorities have observed independently that the greatest concentrations of biological diversity tend to coincide with those of cultural diversity, especially in tropical forest areas, and most of all in the so-called megadiversity countries of Brazil, Colombia, Mexico, Congo, Madagascar, Indonesia, and Papua New Guinea. The present author calls this general tendency toward a geographical coincidence of high cultural diversity and high biodiversity the diversity principle. It has rarely begun to be recognized and systematically described, let alone explained. However, wherever indigenous societies thrive, biodiversity is likely to do so as well.

One thing is certain: in such megadiversity regions, threats to either cultural or biological diversity also threaten the other. Ironically, many indigenous societies that have proven sustainable and adaptive for centu-
ris or even millennia, as well as their environments, are being degraded and destroyed by industrial and other societies that have yet to stand the test of time and increasingly show clear symptoms of maladaptation. The concerns for biodiversity conservation and human rights are interdependent. Furthermore, the degradation and destruction of cultural diversity, like that of biodiversity, seriously endanger the future adaptability of Homo sapiens as well as biological evolution in general.

VII. CONCLUSIONS

Globally the net impact of the human species has most likely decreased biodiversity. However, not all humans are equal in their impact on biodiversity because of the tremendous diversity in mankind throughout its temporal and spatial distribution including cultural diversity. At the population level, clearly some societies may sustain or even enhance biodiversity. In particular, many indigenous societies, especially those who retain some core traditions despite superficial changes, have special potential in their environmental knowledge, worldviews, and other attributes to contribute to developing systems for the sustainable use, management, and conservation of biodiversity. The cumulative and collective impact of humans on biodiversity across the world is sufficient to make it imperative that anyone concerned with the biodiversity of any area must consider the possibilities of human influence. Accordingly, research on cultural ecology and historical ecology is indispensable.

Current anthropogenic extinction rates are estimated at 100 to 10,000 times higher than normal background rates. Furthermore, these recent anthropogenic extinctions also involve plants, whereas prehistoric extinctions mainly affected animals. This is an alarming fact, among other reasons, considering how fundamental plants are to other life as primary producers in capturing solar energy through photosynthesis. Another distinction of the present extinction spasm is that increasingly humans are becoming aware of what they are doing and could change their behavior to reduce their negative impacts. After all, destroying biodiversity and ecosystems is ultimately ecocidal for humanity since they are our life-support systems.

Biodiversity is unlikely to be adequately conserved only by preservationism—by isolating nature from human ‘‘disturbance’’ in a few areas of supposed wilderness. Much more effort needs to be directed to adequately recognizing and better managing human impacts from the local to the global levels and over the long-term. One problem is that decisions and actions that seem reasonable in the short-term may have negative consequences in the long-term. Thus, adequate environmental and resource management requires, among other things, much better informed politicians and policy-makers at all levels (local to international) who have the political will and morality to consider no less than the integral relationship between humanity and biodiversity for many generations to come. Of course, so far such leadership is grossly inadequate, but not unprecedented. For instance, the Iroquois in North America acted with the seventh generation into the future in mind. Perhaps the Convention on Biological Diversity from the 1992 Rio Summit is a hopeful change.

Biodiversity conservation also depends on a much more informed, concerned, and involved public that understands the nature and consequences of human impact on biodiversity. Considering the gravity and urgency of this subject, environmental and biodiversity education must be advanced at all levels and include mass media. In the process, environmental ethics must be first and foremost. One of the best places to begin is by exploring Aldo Leopold’s (1949; 262) land ethic: ‘‘A thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends otherwise.’’ (However, today the term ‘‘resilience’’ is more appropriate than ‘‘stability’’.) Every individual decision potentially has some impact on biodiversity, however small or indirect; and those of humanity collectively can be synergistic and life threatening. (For teaching, especially useful videos are: ‘‘World Population,’’ ‘‘Biodiversity,’’ ‘‘Web of Life,’’ the series ‘‘Living Planet,’’ ‘‘Can the Tropical Rainforests Be Saved?’’, ‘‘An Ecology of Mind’’ from the Millennium series, and ‘‘Spirit and Nature.’’)

Finally, because the human impact on biodiversity both locally and regionally is mixed with positive as well as negative influences, there is reason for hope as well as despair for the future. However, with the current magnitude of biodiversity loss and other environmental problems, there is no doubt that the very viability, resilience, and habitability of too many of the world’s ecosystems, and thereby the biosphere as a whole, are at risk, as are the very futures of organic and human evolution on this planet. Globally the net impact of Homo sapiens in reducing biodiversity is a dangerous reversal of the megatrends that held throughout the last 3.75 billion years of organic evolution on earth—namely, increasing diversity, complexity, and adaptability. Furthermore, recovery from this extinction spasm may require
many millions of years. Because of the gravity and urgency of the negative impact of humans on the earth’s life, this impact is perhaps the single greatest challenge facing humankind and biodiversity studies for the twenty-first century.

See Also the Following Articles

COMMONS, THEORY AND CONCEPT OF • ECOLOGICAL FOOTPRINT, CONCEPT OF • ECOSYSTEM SERVICES, CONCEPT OF • EXTINCTIONS, MODERN INSTANCES OF • HISTORICAL AWARENESS OF BIODIVERSITY • INTRODUCED SPECIES, EFFECT AND DISTRIBUTION • KEYSTONE SPECIES • POLLUTION, OVERVIEW

Bibliography

HUNTER-GATHERER SOCIETIES, ECOLOGICAL IMPACT OF

Kathleen A. Galvin
Colorado State University

I. Introduction
II. Hunter-Gatherer Societies and Natural Resource Exploitation
III. Conservation among Hunter-Gatherers
IV. Modernization Processes and Hunter-Gatherers
V. Land Tenure, Institutions, and Biodiversity
VI. Economic Development and Biodiversity Conservation

GLOSSARY

bands The basic economic, social, and political unit of hunter-gatherer societies.
exogamy The practice of a person seeking a mate outside of his or her group.
patrilocality The practice of married couple's living in the husband's community.

A HUNTER-GATHERER OR FORAGING SOCIETY is a group of people whose subsistence is based on the hunting (or fishing) of animals and gathering of plants. Whether or not foragers have an impact on their environment depends on several factors, some of which emanate from foragers themselves and others which are external to their society.

I. INTRODUCTION

Many people have impressions of hunter-gatherers as people who live in harmony with nature, who are organized into simple societies and are associated with our “pristine” paleolithic hunter-gatherer past. Many of these stereotypic impressions are false (cf. Moran, 1991; Wilmsen, 1989). Today all foragers live in nation-states, have some dependence on either crop cultivation or farmers, and are not isolated. Hunter-gatherer societies have social systems that are extremely complex and whose interactions with the biodiversity surrounding them are as complicated and variable as was probably the case 10,000 years ago when all humans were foragers. It is no accident that today, areas with the greatest remaining biodiversity are also the areas inhabited by hunter-gatherers. Many hunter-gatherers retreating from land appropriation, settler immigration, and European diseases have occupied the most remote parts of their region. Today, these homelands are often part of or adjacent to conservation areas, parks or other protected areas.

This chapter describes traditional hunter-gatherer societies and the adaptations these societies have made to the environment. However, since hunter-gatherer societies and their environments have undergone continuing changes, issues of biodiversity conservation and hunter-gatherer welfare are discussed with the context of their changing world.
II. HUNTER-GATHERER SOCIETIES AND NATURAL RESOURCE EXPLOITATION

Because hunter-gatherers live in diverse environments they manifest an incredible diversity of cultures and natural resource management adaptations. Nevertheless, there are several general characteristics of hunter-gatherer societies; these traits have a direct impact on the use of natural resources. Traditional hunter-gatherer societies are comprised of bands, social groups made up of close biological kin and friends. The composition and sizes of bands change seasonally, depending on the abundance and location of food resources. Bands are led by individual hunters who are respected for particular talents such as singing or dancing well, good storytelling, or hunting prowess. Other features of band organization are small group size, flexible but primarily patrilocal residence, and strong pair bonds between individual men and women. Marriage is exogamous, that is, females are recruited from other groups. These features of hunter-gatherer society are a reflection of ecological, economic, and social necessity. For example, Efe Pygmy hunter-gatherer men of the Ituri Forest in the former Zaire have very strong relationships with close kin, which facilitates defense of their territories against other cooperative kin groups. Moreover, related men can assure women access to valuable resources in neighboring Lese agricultural villages. And women are attracted to men who can guarantee long-standing reciprocal economic relationships with Lese villages. Competition for women is high so close relations with kin may also help to obtain marriageable women and provide protection, as some women leave Efe society to live in Lese agricultural villages. Hunter-gatherers are sedentary or nomadic depending on the distribution and dynamics of their resource base. Typically, men hunt and fish while women gather and collect foods. Sometimes women’s work contributes more to the diet and sometimes male hunting and fishing products are most important. Gathering of wild foods tends to contribute more to the diet among people inhabiting tropical and semitropical areas (e.g., San Bushmen of the Kalahari) than in northern temperate climates (e.g., the Inuit of Canada) where hunting contributes the bulk of the diet. Foragers learn about their environment and resource use through acculturation. Parents teach their children different kinds of ecological knowledge and resource exploitation strategies.

Ecological knowledge is a source of landscape manipulation. For example, the Kayapo Indians of Brazil create forest islands of planted semidomesticated crops of medicinal species, wild yams, and bush bean, as well as domesticated plants such as taro, papaya, and banana. A fully grown island has sites that vary in shade and moisture thereby creating the opportunity for cultivation of different crops. They become, through time, forest patches of varying successional stages within the savanna. Cree Indians of North America rotate their hunting and fishing lands yearly to reduce wildlife disturbance and increase harvests. Biodiversity conservation is, in this case, an indirect effect of resource management. There is evidence that until recently Indians of Canada used fire to maintain trails and to open up meadows. This provided improved habitat for ungulates and increased hunting success. Australian aborigines used fire to clear trails (of poisonous snakes) and keep game habitat open.

Appropriate use of natural resources are maintained through moral and belief systems of forager societies, which includes a strong respect for nature. Through religious belief and social conventions, people respect and exert some control over their natural resources. These beliefs, however, do not always prevent hunter-gatherers from overusing their resource base. Not all hunter-gatherers live always harmoniously with the environment. Indeed, evidence of escalating overuse is accumulating (e.g., Redford and Mansour, 1996).

III. CONSERVATION AMONG HUNTER-GATHERERS

It has been suggested that the hunter-gatherer adaptation occurred in environments where resources were freely available to all and were abundant. Thus, the environment was one where subsistence strategies emphasized short-term returns over long-term conservation. But during the Neolithic rise of agriculture, natural ecosystems were compressed and the value of resources increased as relative abundance declined. Some scholars have suggested that self-regulatory mechanisms evolved under resource limitation in some hunter-gatherer societies (Berkes and Folke, 1998).

There has been much written about how hunter-gatherers are actively engaged in conserving resources, especially animal resources. However, the limited actual data gathered on the subject suggests that subsistence hunters do not conserve prey resources. Most work...
shows that hunters are concerned about short-term gains and not about resource conservation. Small, mobile groups may use resources in a sustainable manner, for example, by maintaining small groups and ranging over a large territory, but this does not necessarily imply they are consciously conserving resources. Evidence suggests that some resources may be used intensively or even depleted in local areas while other resources are sparingly used. For example, Alvard (1998) has shown that the Piro hunters of Peru depleted the large primates in the area around their village yet have not done so to peccaries. Likewise he shows that the Indonesian Wana have nearly depleted their area of macaques (large primates) but hunt pigs in a sustainable manner. These and other studies (e.g., among the Inuit of Canada, the Ache of Paraguay, the Cree of Canada) show that both overexploitation and conservation may be practiced by hunting groups. But the point remains that hunters sometimes reduce prey species to the point of local extinction.

One plausible explanation for resource depletion is that the resources exploited by subsistence hunters are considered to be open-access resources. Open access implies that there are no controls over resource use, which is said to result in the “tragedy of the commons” (Hardin, 1968). This concept proposed that deterioration of open-access grazing land is inevitable when individuals see no benefits from resource conservation. Another reason for resource depletion is lack of concern for very abundant resources. Some level of scarcity adds value to a resource relative to when resources are quite abundant. Resource users are motivated to conserve only when they see benefits to nonuse of resources. Thus, it is only when long-term benefits outweigh the short-term benefits that conservation is expected. When tied to a specific resource base and well-defined territories, hunter-gatherers have long-term strategies for natural resource conservation (Alvard 1998). For example, traditional Maine lobstermen have strong norms of territorial ownership, which are enforced through threats of violence and damage to property.

Although foragers may or may not overuse resources, their perception of the land and its value is based on use rights. Local biological diversity is an important element of local survival strategies. This view contrasts with the western view of biodiversity conservation, which is based in Western epistemology. In the western view, nature exists apart from humankind and has value independent of human use. Biodiversity conservation implies no resource use or restraints in resource use.

IV. MODERNIZATION PROCESSES AND HUNTER-GATHERERS

Major changes in hunter-gatherer society are occurring even in the most remote regions of the world. These changes are associated with agricultural development, infrastructure advancement, resettlement schemes, tree harvesting, mining and oil exploration, and other types of development. The building of roads makes it easier for outsiders to gain access to remote areas and the resources therein. In addition, hunter-gatherer populations are growing, altering their relationship to the land. The result is that indigenous systems of resource use are changing due to both internal and external pressures. The traditional systems of resource use are less appropriate or are sometimes ineffectual under current conditions. For example, traditional sanctions to protect or at least not exhaust resources are becoming ineffectual as cash income has become increasingly important to individuals interested in commodities from the modern world. Hunter-gatherers now have, under these conditions, a growing demand for cash and market goods. Under these conditions, it is less likely that people will give priority to conservation.

V. LAND TENURE, INSTITUTIONS, AND BIODIVERSITY

One political factor that is almost universally common among hunter-gatherers today is that they do not control the land they live on. Until recently, their remoteness meant that they and the resources on which they depended were somewhat protected from outside influences. Thus, resources were locally controlled by informal norms through individual behavior. Now, however, national governments, among others, have put native lands to “productive” use. This means that if the market for some product is strong it will be exploited or cultivated regardless of environmental impact. For example, the strong local demand for aguajá (a local plant) in the Peruvian Amazon has led to destructive harvesting. In theory, most hunter-gatherer communities have use rights to their territories but old laws and treaties are continually violated. Legalizing communal resource-use rights is a way of giving hunter-gatherers a long-term stake in conserving the resources on which they depend. Securing rights to resources can occur through various management and development institutions. This means that hunter-gatherers, who for-
mally did not have institutions for collective action in the formal sense, find the need to deal with western institutions to acquire control over their lands.

The future of biodiversity, conservation, and hunter-gatherer sustainability depends on understanding that there are fundamental differences in the concept of conservation for westerners and for indigenous hunter-gatherers. Understanding that there are different worldviews toward nature is fundamental to forming a relationship between outside conservation groups and hunter-gatherer peoples. The reality is that even if hunter-gatherers are using resources, selling wild animals and cutting down trees, they perhaps remain the most effective conservationists for their region. Therefore, acceptance that there are different ways of viewing the world is a first prerequisite to working with indigenous hunter-gatherer populations. Second, it is necessary to recognize that there are no longer any “pristine” hunter-gatherers and they have needs just like the rest of us. Third, securing land tenure for hunter-gatherers and biodiversity conservation is required for a basis of a “sustainable” interaction.

VI. ECONOMIC DEVELOPMENT AND BIODIVERSITY CONSERVATION

Community-based conservation is a concept aimed at involving local people in the conservation of wildlife or protection of biodiversity. The concept developed from the realization that much of the planet’s wildlife and biodiversity exist outside protected areas and in regions occupied by rural people in developing countries. Models of community-based conservation adhere to the notion that if local communities can derive some value, nominally income, through conserving biodiversity, they will do so. This promising concept has been widely promoted as “the answer” to conservation in developing countries. Thus, several models of community-based conservation have developed. The biosphere reserve is one kind of conservation area that theoretically allows for local population involvement in management of the protected areas. Integrated Conservation-Development projects are another type of community-based development. However, results from community-based conservation projects in Africa and elsewhere suggest that there are more failures than successes. Many community-based conservation efforts involve local communities in name only. Locals are neither involved in project identification and planning nor are they beneficiaries, thus these projects are not really community-based conservation projects. Another pattern of failure includes involvement of the local people only in a cursory way. Other scenarios for failure also have in common insufficient involvement of the local people at all levels in the project. In order for community-based conservation to work, people need to be considered a component of the ecosystem being conserved and brought into the project process from the beginning.

A. It Is Useful to View Humans as Part of Ecosystems

One of the fundamental problems with community-based conservation is that hunter-gatherers as well as other indigenous populations are often viewed as an external disturbance to the natural system rather than as integral components of the ecosystem. But hunter-gatherer societies see their relationship with the environment as one; they are part of that environment. Though not a study of foragers, but rather herders who do some hunting and gathering, the South Turkana Ecosystem Project (Ellis and Swift, 1988; Little and Leslie, 1999) is one of the only truly interdisciplinary and long-term projects to study the social behavior, knowledge systems, demography, human biology, and ecology of a group of people. An important goal of this study was to understand how the environment affected human management and how people affected the environment. In this case people and livestock (camels, cattle, sheep, goats, and donkeys) lived in a harsh, dry, and highly seasonal environment. This assemblage of people, livestock, plants, and other organisms within a semiarid ecosystem produced a remarkably interactive system.

Vegetation structure in this tropical savanna and dry woodlands was shown to be hierarchically constrained by physical factors: by climate at regional scales, by topography and geomorphology at landscape scales, and by water redistribution and disturbance at local and patch scales; livestock and humans played a small role. The pastoralists did influence vegetation composition and cover by burning, woodcutting, and through seed distribution by livestock. These influences were small. Livestock ecology and production followed those of the seasonal dynamics of plants. The different patterns of forage utilization by different herbivores, plus differential habitat use, lead to almost complete niche separation among this suite of domestic herbivores; among all five species, they managed to utilize a wide variety of the available plant types in the ecosystem. Thus, physical heterogeneity on the Turkana land-
HUNTER-GATHERER SOCIETIES, ECOLOGICAL IMPACT OF

Scapes ultimately resulted in spatial and temporal variation in plant production, plant life form diversity, and refuge areas for pastoralists. These, in turn, contributed to social and ecological persistence by reducing variability of ecosystem energy flow and long-term variations in species diversity. Thus, biodiversity was important to ecosystem (which included people) maintenance. This systems approach to understanding human-environment interactions is a useful way to discern the ecological impact of hunter-gatherers and, more important, to derive appropriate management of lands where hunter-gatherers live.

The description presented here shows that indigenous concepts of conservation, ecological knowledge, and moral and religious beliefs are fundamental to understanding how hunter-gatherers use resources. Not all hunter-gatherers conserve their resources, thus whether or not and to what extent hunter-gatherers effect their environment is an empirical question that needs to be investigated, not a notion to be assumed one way or another. It is, however, the case that when hunter-gatherers have short-term strategies for resource use they may overuse some resources; when long-term goals are in place, they do not. Informal institutions control use of some resources in hunter-gatherer societies, but collective action or formal institutions are generally not well developed. With major changes in and around the lands inhabited by hunter-gatherers, it is becoming increasingly necessary for hunter-gatherers to develop institutions to gain control over their resource base. Alliances between hunter-gatherers and others interested in conservation may facilitate resource-management strategies that reduce the impact of negative changes. Hunter-gatherer natural resource-management strategies that include their social system are important attributes of these ecological systems and need to be fundamental components of any plan to conserve biodiversity.

See Also the Following Articles

Biodiversity-Rich Countries • Ethnobiology and Ethnoecology • Indigenous Peoples, Biodiversity and Land-Use Patterns, Historic • Religious Traditions and Biodiversity • Traditional Conservation Practices

Bibliography


HYMENOPTERA

Norman F. Johnson
The Ohio State University

I. Classification
II. Phylogeny and Fossil Record
III. Biology
IV. Practical import

GLOSSARY

arthronotoky Reproductive mode in which unfertilized eggs develop into haploid males and fertilized eggs develop into diploid females.
eusociality Cooperative behavior among individuals of the same species characterized by reproductive division of labor, overlap of generations, and cooperative nesting.
holometabolous Insects characterized by complete metamorphosis, a wingless larval stage, and an intermediate pupal stage.
idiohost A parasitoid that develops on a paralyzed, incapacitated host.
koinobiont A parasitoid that develops on a mobile, active host.
monophyletic A group in which all species are descended from a single common ancestor and all descendants of the ancestor are classified in the group, characterized by shared derived characters.
ovidpositor Modified appendages of the seventh and eighth abdominal segments used for egg-laying paraphyletic: A group in which only some of the species descended from an ancestor are classified together; characterized by shared ancestral characters.
parasitoid An organism in which the immature stage feeds and develops on a single host arthropod, resulting in the death of the host.
parthenogenesis Reproduction in which eggs are not fertilized by males.
phylogeny Branching pattern of evolutionary relationships among organisms.
phytophagy Plant feeders, herbivores.
thelynotoky Reproductive mode in which unfertilized eggs develop into diploid females.

THE INSECT ORDER HYMENOPTERA comprises a vast array of species that are familiar to even the most casual observer. The group includes the ants, bees, and wasps as well as less well-known groups such as the chalcids, ichneumons, sawflies, and wood wasps. Hymenoptera are extremely common on all continents of the world, except Antarctica. Hymenoptera range in size from microscopic species less than 1 mm in length as full-grown adults to species in which the females are 10 cm or more in total body size. Hymenoptera are holometabolous insects, generally characterized by having mandibulate mouthparts, complete metamorphosis, and two pairs of membranous wings. One striking feature is that the vast majority of species are characterized by arthropodous parthenogenesis: Fertilized eggs develop into diploid females and unfertilized eggs ultimately develop into haploid males. Adult Hymenoptera generally feed on nectar, honeydew, or other sugar.
sources, but some are actively predacious. The larvae have a wider variety of feeding habits: Some feed on plants or fungi, whereas others feed on animals, as predators or parasitoids. The larvae of bees and a few other groups feed on provisions of pollen collected and stored by the parent female. The egg-laying appendages, the ovipositor, often are specialized to place eggs on or into the plants or arthropods that will serve as the food source for the developing larvae. In one large group, the Aculeata, the ovipositor is modified into the sting, a structure that primarily serves to deliver toxic venoms to paralyze prey or to use as a defense. The Hymenoptera are also noteworthy for the evolution of social behavior in several groups. The most highly developed form, eusociality, has arisen many times within the order among the bees, wasps, and ants.

I. CLASSIFICATION

Hymenoptera has been traditionally divided into two suborders, the Symphyta and the Apocrita. The Symphyta comprise the sawflies and wood wasps. The larvae of the vast majority of species in this suborder consume plant or fungal material, feeding on leaves or stems, feeding within galls produced in the host plant, or boring within stems or trunks of woody plants. Symbiotic fungi comprise a significant part of the diet of many of the wood-boring species. Sawfly larvae that feed externally on leaves are generally eruciform, i.e., caterpillar-like in form, with well-developed thoracic legs, abdominal prolegs, and sclerotized heads.

The suborder Apocrita is descended from a symphytan ancestral species. As adults, the first abdominal segment of the Apocrita is intimately associated with the thorax and separated from the following segments by a mobile constriction. This segment is called the propodeum. Thus, the locomotory tagma of the body, bearing the legs and wings, is four segmented. The larvae are generally immobile and highly simplified in structure. Apocrita are generally characterized by larvae that feed on other arthropods, including predatory species in which one larva feeds on several prey items and parasitoids in which the larva feeds on a single host individual. Parasitoids are distinguished from parasites in that the host of the former is usually killed as a result of feeding. Thus, ecologically, parasitoids function as predators, feeding on and ultimately killing their host. The parasitoid life history is also found in the closest living relative of Apocrita within the Symphyta—the family Orussidae. However, numerous apocritan groups are secondarily phytophagous, consuming seeds or pollen or forming galls in plant tissue within which the larvae feed and develop.

The division of Hymenoptera into two suborders is still generally accepted, even though recent evidence indicates that the Symphyta are paraphyletic. In the formal taxonomic classification, approximately 19 superfamilies and 95 families are recognized. The Tenthredinoidea, Megalodontoidae, Cephoidae, and Xyeloidea comprise the phytophagous sawflies. Wood wasps comprise four families within the Siricoidae. In the Apocrita there are many small, relatively primitive superfamilies of parasitoids: Trigonaloidae, Ceraphronoidea, Eulophidae, Megaluryoidae, and Stephanoidea. The superfamilies Chalcidoidea, Cynipoidea, Ichneumonoidea, Proctotrupoidae, and Platygastroidea encompass the vast majority of parasitoid species. Many of these are common and extremely abundant. The family Ichneumonidae is estimated to contain more than 22,000 known species, with many others yet to be recognized. The remaining four superfamilies—the Chrysidoidae, Vespoidae, Apoidea, and Sphecoidae—make up the aculeate, or stinging Hymenoptera. The Chrysidoidae include many relatively small families of parasitoids. The Vespoidae contains the ants, velvet ants, hornets, spider wasps, paper wasps, and potter wasps. The Apoidea are the bees, and the Sphecoidae are a large diverse group of predatory wasps.

The order Hymenoptera is estimated to contain approximately 115,000 described species. This is only a rough indicator of the true total because hundreds of new species are described each year in the taxonomic literature. The largest families are the Ichneumonidae (more than 20,000 species), Braconidae (approximately 12,000 species), Formicidae (the ants, with nearly 10,000 species), and the Sphecidae (more than 8000 species). Austral disjunct distributions are characteristic of a relatively few families of Hymenoptera, such as Megaluryidae and Plumeriidae. The major groups are found in all biogeographic realms or perhaps limited to the tropics, such as the fig wasps or Agaonidae. Ichneumonoids and sawflies appear to be most diverse in the temperate realms.

II. PHYLOGENY AND FOSSIL RECORD

The oldest fossil Hymenoptera are found in deposits from the Triassic period of the Mesozoic. These include some species that are closely related to living sawflies. Parasitoid wasps, as inferred from their structure (especially the elongate ovipositor in some species) and their relationships to living species, first appear in the Jurass-
sic. The oldest fossils of social insects, ants and bees, have been found in Cretaceous impressions and amber nodules. Minute parasitic Hymenoptera and ants and bees are commonly found in Cenozoic amber from the Baltic, the Dominican Republic, and Mexico.

Although Hymenoptera are clearly members of the monophyletic group Holometabola, comprising insects with complete metamorphosis, their position within that huge complex is unclear. Most workers place them as a basal group, sometimes as the sister group to all other holometabolous orders. Within Hymenoptera, the basal position of Symphyta is supported by evidence from morphology, behavior, molecular sequences, and the fossil record. The historical grouping of parasitoid species into a taxon variously called "Parasitica" or "Terebrantes," although of some practical use, does not reflect the pattern of phylogenetic relationships among them. The relationships among the basal Apocrita are still unclear. Most of these species are parasitoids of wood-boring Coleoptera, a life history shared with their closest living relatives in the Orussidae. Numerous hypotheses of relationships and classifications have been offered through the years that have been based on varying amounts of evidence from several sources and analyzed both more and less rigorously, but none has achieved anything approaching general acceptance. The Aculenta are a well-supported monophyletic group, and a fair amount of progress has been made in elucidating the relationships among its constituents. It is also fairly clear that the bees and sphecoideid wasps are very closely related; in fact, it seems evident that bees evolved from some subgroup of sphecids.

III. BIOLOGY

A. General

1. Reproduction

Arrenotokous parthenogenesis, or haplodiploid reproduction, is characteristic of the vast majority of species of Hymenoptera and is generally thought to have been very important in its evolution. In this mode of reproduction, offspring are produced from both fertilized and unfertilized eggs. Eggs that are not fertilized develop into haploid males. Fertilized eggs develop into diploid embryos, and these are usually females. Because of the mechanism of sex determination in Hymenoptera, which is not completely understood, diploid males are rarely produced. Mated adult females store spermatozoa received from the males during mating in a diverticulum off of the reproductive tract, the spermatheca.

The female is capable of controlling the release of spermatozoa to fertilize eggs passing through the oviduct, thus controlling the sex of her offspring. The control of the sex ratio of offspring may be advantageously applied, in theory, to maximize the reproductive success of the female. In some parasitoid species, the size of the host can influence the female to lay a male or a female egg. In other circumstances, the presence of other ovipositing females on a host has been shown to influence the relative proportions of male and female eggs laid. Haplodiploidy also affects the proportions of alleles that are shared among relatives in comparison to the expectations of relatedness that are commonly found among strictly diploid species. In the simplest case of a female mating with one male, a daughter may be generally expected to share a greater proportion of its alleles with her sister (75%) than with her brother (25%). The two female offspring share one set of chromosomes contributed by the haploid male spermatozoa and, on average, would expect to share 50% of the alleles inherited from the mother. In contrast, the male offspring has only a single set of chromosomes derived from his mother. This asymmetry in relatedness has been thought to be a contributing factor to the multiple evolution of eusociality in the Hymenoptera, but its importance is strongly questioned. Although single matings may be the rule for many species under normal conditions, one female mating with many males is not uncommon: The mating flight of the honeybee, Apis mellifera Linnaeus, is an extreme example. Also, many species of social insects, especially among the ants, have numerous unrelated egg-laying queens within a colony.

Another form of parthenogenesis is relatively common within Hymenoptera: thelytoky. In this case, the female does not mate with a male, and her unfertilized eggs are capable of embryonic development and produce only female offspring. Thelytoky is not uncommon among parasitoid groups, and the funneling of all reproductive output into the production of new egg-laying daughters maximizes short-term reproductive success. It has recently been discovered that in several cases thelytokous parthenogenesis is caused by bacterial infections, principally by the genus Wolbachia. Treatment of the wasps with heat or antibiotics can cause the individuals to revert to normal arrenotokous modes of reproduction. Many strains of Wolbachia do not result in thelytokous reproduction but rather in cytoplasmic incompatibility between eggs and spermatozoa. Infections by these bacteria have been discovered in a wide range of insects and the evolutionary and ecological significance of the phenomenon is not well understood.
2. Oviposition Behavior

Hymenoptera are probably most familiar as stinging insects. The sting apparatus is a modification of the egg-laying appendages fundamental to all Hymenoptera. Because egg-laying structures are naturally associated with females, only female wasps can sting. However, the males of some species are extraordinarily adept at mimicking both the structure and behavior of females and can sometimes fool even experienced workers.

In its ancestral condition, the ovipositor is composed of two pairs of elongate appendages, the gonapophyses or ovipositor valves, associated with the seventh and eighth abdominal segments. The gonapophyses are physically linked together and function as a single articulating unit. When not in use, the ovipositor is usually enclosed within a protective pair of sheaths, the gonoplaec. In sawflies, the gonapophyses are laterally flattened, serrate, and used to cut an incision into plant material into which one or more eggs are laid. The similarity in structure and function to a saw gives rise to the common name for most of the Symphyta.

Within the Hymenoptera, the ovipositor and its function have been extended beyond its basic function in egg-laying. Among parasitoids, it is used to paralyze hosts; in the Aculeata or stinging Hymenoptera, the ovipositor and venom are often used as a defensive weapon.

B. Parasitoids

A parasitoid is defined as an animal in which the immature stage feeds on a single host individual, and this feeding activity normally results in the death of the host. Historically, such insects have been called parasites, but this was misleading. True parasites, such as lice, feed on a large host but the feeding activity has only a minor deleterious effect. Despite this distinction in names for the two type of life history, the verb "parasitize" is still normally used to describe the action of both groups. A parasitoid is essentially a predator, but one that feeds on and kills only a single prey item. They also differ from predators in that it is the parent that finds the "prey" and the offspring that feeds. The Hymenoptera are the primary group of insect parasitoids, but other orders, particularly the Diptera and Coleoptera, have species that follow this same pattern.

Parasitoid Hymenoptera attack a very broad range of arthropods. Practically all other insect groups have their suite of parasitoids as well as arachnids and myriapods. Most parasitoids attack the immature stages of their host: the egg, nympha, larva, or pupa. A few groups parasitize adult insects. In some cases, the female parasitoid oviposits on one stage and her adult offspring emerge from another. For example, some species of Braconidae oviposit within the egg of their host, but the wasp larva actually feeds on and emerges from the larva. After the completion of feeding, the parasitoid larva pupates on, within, or sometimes nearby the spent carcass of its host.

The ovipositor is used to first "sting" the host arthropod, and venom is then injected. The venom may incapacitate the host for only a short period of time or, in more extreme cases, may result in the complete suspension of most normal activities, putting the host in a state of "suspending animation." The female wasp then lays one or more eggs on or often within the body of the host. In some species, the female parasitoid attacks exposed hosts—for example, a caterpillar feeding on a leaf. In other cases, though, the hosts are hidden in the soil, enclosed within shelters made of leaves, twigs, debris, or silk, or they even bore deep within the trunks of trees. The structural characteristics, sensory capabilities, and behavioral repertoire of the female wasp combine to enable her to locate and successfully oviposit upon such protected hosts. For example, many species of parasitoid Hymenoptera that attack wood-boring beetles are equipped with elongate ovipositors that may be two or more times the length of the rest of the body. The females are capable of drilling through tens of centimeters of wood with their ovipositor in order to find and parasitize the hidden beetle larva.

The proximate mechanisms by which a female parasitoid locates, identifies, and determines the suitability of hosts within a complex environment are only beginning to be understood. Chemical cues are certainly important in all of these stages of host finding. These chemicals may be produced by the hosts as a by-product of normal activity such as feeding. Often, parasitoid females are attracted to volatile chemicals produced by the plant on which a host is feeding. Other sensory modalities used in finding hosts and assessing their quality include vision, tactile examination, detection of substrate vibrations, and possibly heat detection. The external surface of the body of an adult wasp is studded with an array of sensory structures. For host location and acceptance, the most important of these are found on the antennae, tarsi, mouthparts, and ovipositor.

Among the array of details of life history strategies, two general patterns are sometimes distinguished and parasitoids may be classified as idiobionts or koinobionts. Idiobionts are generally characterized by the fact that the development of the host arthropod is arrested through the action of the venom injected by the parent. The larval parasitoid thus develops on an incapacitated
host. In koinobionts, the host resumes feeding and development after the parasitoid female has oviposited. The larval koinobiont parasitoids typically develop within the body of the host, but this may be delayed for some time as the host grows and even pupates. Thus, the parasitoid must have the physiological mechanisms to evade the immune system of its host. Idiobionts typically attack hosts that are found in concealed locations and a broad taxonomic range of hosts may be attacked. Koinobionts, in contrast, may parasitize fully exposed hosts and, because they must be attuned much more closely to the host immune and endocrine systems, their host range is typically much narrower.

There is a tremendous range of variation found within the general life history patterns described. Parasitoid larvae may feed on the host from outside the host’s body as ectoparasitoids, or they may live and feed within the body of the host as endoparasitoids. Among solitary parasitoids, a single individual wasp develops on each host individual. The adult wasp that emerges from the pupa then must locate mates, food, water, and new hosts on which to oviposit. Many individual parasitoids develop on a single host in gregarious parasitism; the number of wasps may range from two to thousands. In these cases, males generally emerge first and mate with the later-emerging females on or near their pupation site. In some cases, this leads to a high level of sib-mating (i.e., brothers mating sisters), and in such situations it is not uncommon for the sex ratio to be significantly skewed toward the production of females. In other words, the parent female lays just enough male eggs, sometimes only one, to ensure the insemination of all her daughters. In this manner, she is thought to be able to increase her overall reproductive output. Gregarious parasitism may be effected by numerous wasps attacking a single host, by a single wasp placing more than one egg on or in the host, or by a peculiar phenomenon called polyembryony. In this case, the parent female places one egg within the body of the host, but this single egg produces from two to thousands of identical embryos, each of which eventually develops into a new adult parasitoid. Polyembryony has apparently evolved at least four times within the Hymenoptera because it is found in the families Braconidae, Pteromalidae, Dryinidae, and Encyrtidae.

Hyperparasitoids, or secondary parasitoids, are species that are actually parasitoids of other parasitoids. For example, in some species of the family Trigonalidae the adult female lays large numbers of eggs on foliage near feeding caterpillars. The eggs are inadvertently ingested by the caterpillar, the thick shell of the egg is disrupted by the caterpillar’s mandibles, and the trigo- nalid larva quickly hatches and bores through the gut of the host to enter the hemocoel. It then lies dormant until the caterpillar is subsequently parasitized by another parasitoid species, typically an ichneumonoid or a tachinid (Diptera). The new parasitoid develops on the caterpillar host but is then attacked and ultimately killed by the trichomalid. Success in such a complicated life history requires a sequence of individually unlikely events: ingestion of the egg; survival of the host caterpillar from disease, predation, starvation, etc.; and subsequent parasitization of the caterpillar by a suitable species. The low probability of such a sequence is compensated by the fact that the female trichomalid lays huge numbers of eggs.

One of the most bizarre types of life history is the heteronomous parasitism found among some species of the chalcidoid family Aphelinidae. Female offspring develop as primary endoparasitoids of Homoptera. The males, in contrast, may develop as ectoparasitoids of the same species of host or as secondary parasitoids, sometimes attacking females of their own species.

Although the parent female of most parasitic Hymenoptera finds and oviposits on the host for her offspring, in a few cases it is the larvae that locate the hosts. In the Perlampidae and Eucharitidae, the adult female wasp lays her eggs on or in plant material, often in large numbers. The first instar larva that hatch from these eggs are strongly sclerotized and mobile, in contrast to the typical larva of Apocrita. These mobile first instars, called planidia, somehow find their way to their host, attach, and feed. In the eucharitids, it appears that the planidia may attach to thrips, which are then picked up and carried by ants into their nest. Once there, the planidia disengage from their temporary transport and attach and feed on ant larvae.

C. Predators

The distinction between hymenopteran larvae that act as parasitoids and those that act as predators is fairly arbitrary, being dependent only on the number of host individuals killed and eaten in the process of larval development. The typical predatory life history is extremely similar to the idiobiont strategy. The parent female wasp finds, stings, and immobilizes the prey item. The prey may already be in a place of concealment, such as beneath a rock or within a burrow, or the female wasp may actually construct a hiding place, often a burrow in the soil or in tunnels in wood. More than one prey item may be cached in a chamber in this burrow, and then the female wasp lays an egg on the prey. The larva that hatches feeds on and develops on
this store of food that has been provided by its mother. In the case in which only one prey item is available, this is essentially identical to the life history of an external parasitoid. In many species, the parent female may use the same burrow within which to store the prey for several of her larvae, constructing side chambers for each of the offspring or building chambers in a linear series. Thus, from the starting point of ovipositioning on a paralyzed prey item within its own hiding place, there is the development of a nest—that is, a structure built by the parent female within which her young are fed and develop. Some wasps gather together all of the prey that an offspring will need for development before ovipositing. Others progressively provision their nest, providing new prey items to the developing larva as it consumes the food available.

D. Phytophagy

It is generally accepted that the Hymenoptera evolved from a phytophagous ancestor similar in many respects to the extant sawflies. The Apocrita, in turn, are derived from a parasitoid ancestor. However, within the Apocrita, there are numerous cases in which phytophagy, in its broadest sense, has reoccurred. In some cases, such as the seed chalcids of the family Eurytomidae, the host relationship is fairly straightforward: The female wasp oviposits into the developing seed and the larva feeds on the endosperm. In other cases, the phytophagy is more elaborate or even bizarre.

1. Gall Makers

Galls are abnormal growths of plant tissue caused by some sort of stimulus from another plant or animal. A wide range of insects are known to cause gall development; within the Hymenoptera it is found among sawflies, the primarily parasitic Braconidae (Ichneumonoidea) and Chalcidoidea, and especially in the gall wasps, the family Cynipidae (Cynipoidea). Gall development is induced either by chemicals injected into the plant by the ovipositing female or by secretions produced by the newly hatched larva. The plant tissue that is elaborated into the gall may take a variety of shapes and sizes, from mere swellings of the stem to large, ornate structures on the leaves or roots. The gall maker feeds on the plant tissue within the gall, pupates there, and eventually chews an emergence hole through which it escapes. The nutritive tissue, both that of the gall and sometimes that of the developing gall maker, is also made use of by a wide variety of inquiline species which oviposit into the gall and kill the original gall maker.

The life histories of cynipid species may vary greatly. Some are fairly typical of other Hymenoptera, i.e., males and females emerge from their respective galls and then mate, and the females then seek out new host plants in which to oviposit. Their eggs develop in either females or males, depending on whether the eggs are fertilized or not. Such species are typically univoltine and attack a wide variety of host plants. Other life histories may be significantly more complicated. Some species have simply abandoned the production of males and are strictly thalytous. Others alternate between sexual and asexual generations, reproducing via arrhenotokous and thelytokous parthenogenesis, respectively. The sexual generation consists of both males and females. These mate, and the inseminated females seek hosts in which to oviposit. All of the offspring of these wasps develop into females. The adults of this new generation in turn oviposit, and some of the eggs develop into males and some into females, thus returning to the sexual generation. The adults of the sexual and asexual generation of gall wasps often strikingly differ in their structure and produce very different types of galls in different positions on the host plant. In some cases, they even attack different species of plants.

2. Fig Wasps

The fig wasps, of the chalcidoid subfamily Agaoninae, are very specialized forms of gall makers. Female fig wasps burrow their way into the syconium, the inflorescence of the fig plant. Pollen from the body of the female is transferred to the flowers concealed within the syconium and then the female oviposits within some of the flowers. These flowers subsequently swell, and the emerging larva feeds within the tissues, consuming both the embryo and endosperm. The male wasps then emerge first and mate with the females while they are still within the galled flowers of the fig. The females later emerge from the flowers, pick up pollen, sometimes actively storing it in special cavities on the body, and escape through holes chewed by the males. Male fig wasps are extraordinarily aberrant creatures: They lack wings, the eyes and antennae are underdeveloped, but in contrast the legs and mandibles are sometimes enormous. A rich complement of other parasitic Hymenoptera are associated with figs, some developing on fig tissues and others acting as parasitoids of the fig wasps.

3. Leaf-Cutting Ants

The leaf cutters belong to the tribe Attini (Formicidae). These ants are found nearly throughout tropical
America, north into the southern portions of the United States. The long lines of thousands of ants, pruning pieces of vegetation and carrying them back to the nest are a familiar sight throughout the New World. Strictly speaking, the leaf cutters are not phytophagous because they do not feed directly on the vegetation that they collect. Rather, they use the leaves and stems as the substrate on which they maintain a fungal colony. The ants feed on specialized structures produced by their fungus; the fungal organism, in return, is nurtured and maintained in an environment favorable for growth. When leaving the nest on her mating flight, the new virgin queen takes a bit of the fungal mass to use as an inoculant when she begins her new colony. Colonies of leaf cutters can be extremely large, numbering in the millions. As a result, they can be extremely destructive to agriculture, forestry, and horticulture.

4. Pollen Feeders
The life history described previously for predatory stinging Hymenoptera has been modified in at least two cases in which the parent female provisions her nest with pollen rather than arthropod prey. The largest group in which this has occurred, and one of the groups most familiar to the casual observer, comprises the bees. Female bees visit flowers both to gather nectar for their own energetic needs and to gather the pollen on which their offspring will feed. Some bees are oligolectic, i.e., they are fairly restricted in the range of plant species from which they gather pollen. Others, such as the ubiquitous honeybee, are polylectic and gather pollen from a broad range of species. There are many specializations of the body of bees that facilitate the transport of pollen. Some bees will ingest the pollen grains and transport them within their crop; others have dense beds of hairs particularly on the legs or on the underside of the abdomen where pollen is packed for transport. Pollination of flowers by bees—that is, the transport of pollen from one flower to another—is often accidental and is irrelevant to the insect but of critical reproductive importance to the plant. Many anatomical features of plants are believed to be specializations for attracting bees or other insects to the nectar and ensuring that pollen grains adhere to their body.

E. Social Behavior
Social behavior is most simply defined as that of groups of individuals of the same species that cooperate with one another. Simple aggregations of individuals may occur, for example, where some limiting resource is found, such as water or nesting sites. Evolutionarily important social behavior, however, involves some sort of cooperation among individuals leading to reproductive success of some or all of the participants. Among the insects, the extremes in development of social behavior primarily are found in the termites (order Isoptera) and within the Hymenoptera. Its most highly developed level, called eusociality, is characterized by cooperation among females in nesting, overlap in generations, and reproductive division of labor. This means that some individuals sacrifice production of their own offspring in order to facilitate reproduction by other individuals of the same colony. Eusociality has clearly evolved several times within the order Hymenoptera: once in the ants, which are primitively eusocial; once in the large family Sphecidae; and several times among the bees and the social wasps (i.e., the paper wasps, yellowjackets, hornets, etc.).

Colonies of social insects can be extremely large, both physically and in terms of total numbers of individuals. The life cycle of a colony typically begins with the mating flight of a virgin queen. After mating with one to several males, the queen begins construction of a nest in the soil, a natural cavity, or in some cases in the open. Colony founding is sometimes cooperative (e.g., in some paper wasps), and the determination of which individual will become the primary reproductive ones is established through behavioral interactions among the founders. Males do not participate in colony founding and only serve to inseminate the new queens, after which they soon perish. The eggs produced by the queen develop into the first worker generation; these are all daughters of the queen and typically do not reproduce. Workers forage for food and nest materials, care for the developing brood, defend and care for the nest, and care for the queen. This range of behaviors is sometimes divided among the workers. In some species, individuals are morphologically specialized into recognizable castes for certain functions such as defense in the case of soldiers. Morphological differentiation into castes results from allometric growth of the individual workers. Individuals of overall larger body size have disproportionately large mandibles, spines, and head capsules. In other cases, individuals may not be morphologically distinguishable but specialize in particular sets of behaviors at different periods of their adult life, usually ending as foragers outside the nest. Once well established, the collective efforts of the individuals of the colony result in the production of a new generation of males and reproductive females. Some social species reproduce by swarming: in the case of the honeybee, the old queen leaves the existing nest, taking with her
a portion of the worker force, and reestablishes in a new site. Then one of the newly emerging queens takes over the remainder of the colony.

Communication among individuals is critical to social behavior. Chemical communication is probably of primary importance. Many exocrine glands have been identified on social Hymenoptera from all major parts of the body. These chemicals include alarm and defense pheromones, mating pheromones, trail pheromones, and signals from the queen that suppress the reproduction of workers. Trophallaxis, the exchange of food between individuals, may serve as the medium of exchange of some chemical signals. Tactile behavior is also important, principally involving antennal contact between individuals. The classic example of tactile communication is the waggle dance of honeybees. The location of food sources, including direction and to some extent distance, is transmitted to nestmates by a stylized reenactment of the foraging flight within the darkness of the hive. The direction of the resource in relation to the sun is indicated by the angle formed by the waggle portion of the dance in relation to the vertical axis within the hive.

Eusocial behavior is a fairly rare phenomenon in the animal world. The repeated evolution of such complex behavior within the single order Hymenoptera has led to a great deal of work in search of an explanation. One suggestion is that the asymmetric relationship among siblings may be an important underlying feature that predisposes this group to the evolution of the characteristics of sociality. However, within the group at least two separate paths leading to eusociality have been identified. The parasocial route first envisons communal nesting among related females, followed by cooperative brood care and eventually reproductive division of labor. The subsocial route to eusociality posts first the development of overlap of generations followed by continued reproduction of the parent female in the company of her own daughters.

These characterizations of social behavior are often more simplified than the real situation found in nature. Colonies may be founded by more than one queen; sometimes these are closely related. Queens are probably usually inseminated by several or even many males, thus reducing the relevance of the asymmetry of relationships associated with arrhaphenogonous parthenogenesis. It is also clear that eusociality, although a "highly evolved" trait, is not the end point of evolution in Hymenoptera. Local environmental factors are undoubtedly important in the development of eusocial traits and in their subsequent loss. Several cases have been identified in which these behavioral repertoires and both the mor-

phological and physiological features associated with them have been lost. Some of the most extreme cases are social parasitism in which reproductive females of one species live within the nest of another social species. The workers of the host species feed and care for the parasite, and the latter contributes nothing to the welfare of the colony. Social parasites are widely found among the ants, bees, and wasps.

IV. PRACTICAL IMPORT

A. Pests

In comparison to other large orders of insects, the Hymenoptera have relatively few species that qualify as pest species. A few of the phytophagous forms do cause some damage to agricultural crops or to trees. Leaf-cutting ants and leaf-cutter bees can cause damage by removing significant amounts of foliage from plants. In contrast to the ants, leaf-cutter bees do not feed on the leaves, either directly or indirectly, but use the pieces in order to line the brood chambers in their nests. Seed chalcids may destroy seeds of crops such as clover. Many of the phytophagous sawflies are important pests, particularly to forest trees. The family Diprionidae, which includes species such as the red-headed pine sawfly (Neodiprion lecontei), are often important pests of conifers. Other families such as the Tenthredinidae (e.g., the larch sawfly, Pristiphora erichsonii, and the birch leaf miner, Fenusa pusilla) and Cephidae (e.g., the wheat stem sawfly, Cephus cinctus) are occasionally important.

The most significant deleterious role played by Hymenoptera is derived from the reaction that people have to being stung. Although a few species of parasitic Hymenoptera are capable of jabbing their ovipositor through the skin of a person, this usually requires that the person actually hold the wasp in his or her hand: This is not a health concern. Aculeate Hymenoptera, on the other hand, inject venom through the sting apparatus into the subject of their attention. In some cases, the venom can cause severe pain by virtue of the properties of the chemicals. The most significant hazard, though, lies with the reaction of the human immune system to the components of the injected venom. The typical response of swelling, pain, and itching may become life threatening for individuals that become hypersensitive to the components of the venom. In extreme cases, the resulting loss in blood pressure and shock can cause death, even from the sting of a single bee.
The so-called "killer bees" or Africanized honeybees are perceived as a serious health risk in the New World. These bees are not an aggressive exotic species that has invaded the Western Hemisphere but rather the same species as the honeybee that is used in apiculture throughout the world (Apis mellifera). Colonies of the honeybee from Africa were brought to Brazil in order to try to transfer some of their useful traits into the colonies of European honeybees being maintained in the tropics through controlled breeding. Unfortunately, African bees escaped from confinement and became established as feral colonies. This strain of bees has spread and interbred with both feral and domesticated colonies of honeybees throughout South and Central America and into the southern United States. The African subspecies of the honeybee is notorious because the colony more vigorously defends its nest against marauders than the European bees and is generally more sensitive to disturbance. In defense, the bees attack the animal or person disturbing the colony by stinging, just as do European bees. The difference between the two types of bees lies not in a more toxic venom but in the fact the African forms are quicker to attack and are more persistent in pursuit and stinging. Numerous deaths of humans and livestock have been recorded that result from the high number of stings inflicted and not from hypersensitivity of the person or animal being attacked. In areas in which the honeybees have become more aggressive, greater care is needed by persons approaching the hives.

B. Biological Control Agents

Although some species of Hymenoptera have negative impacts on humans and their commerce, by and large the order is considered to be very beneficial. One of the most important categories of such beneficial insects consists of those that either parasitize or prey on other pest arthropods. These species, by eventually killing the other insects, act as biological agents that reduce the population levels of their hosts. Such control agents have many advantages: Parasitoids are often fairly host specific, and thus their effects are focused on the pest problem; a population of biological control agents is capable of maintaining itself through time, thus providing continuous control; parasitoids are often extremely effective at locating their hosts in a complexly structured environment; and parasitoid populations respond to those of their hosts, increasing when host populations increase and decreasing when the availability decreases. The importance of parasitoids as regulators of the populations of their hosts is most clearly seen when, e.g., a phytophagous species is accidentally introduced into a new region without its normal complement of natural enemies. Populations of such species are capable of explosive growth in the new environment, rapidly becoming important pests, e.g., gypsy moths (Lymantria dispar) in North America and the cassava mealybug (Phenacoccus manihoti) in tropical America. Biological control programs involve the discovery of the natural environment and enemies of such pests and the importation of these natural enemies for control. Some examples of biological control programs have been spectacularly successful, with effective long-term control of pests. The parasitoid families Encyrtidae and Aphelinidae have been particularly effective in biological control programs of many species of Hymenoptera.

C. Pollination

The successful development of seeds and fruit of many species of flowering plants depends on pollination—the movement of pollen from one flower to another. Although some species of commercially important plants, particularly grasses, are wind pollinated, a great many others rely on animal agents for the transfer of pollen. Under natural conditions pollinators can include vertebrates, such as bats and hummingbirds, and a wide variety of insects, including flies and beetles. Hymenoptera, however, and in particular the bees, are extremely important as pollinators. Recall that bees provision their nests with plant pollen on which their developing larvae feed. In the process of collecting pollen for their nests and also in collecting nectar from within the flowers, pollen grains that adhere to the hairy body of a bee can be transferred to the style of the flower. Natural pollinators include a vast array of native bees, but for many commercial crops colonies of honeybees are used to effect pollination. The honeybee is a polylectic species, that is, it gathers and its larvae feed on pollen from a wide variety of plant species. Colonies are regularly transported to fields in order to temporarily increase population levels to maximize fruit and seed production.

D. Honey and other Bee Products

The large colonies of social bees also store honey within the nests that they construct. Honey is the modified and concentrated product of the nectar gathered principally at flowers and is used as an energy source for the adult bees. The most familiar species that serve as sources of honey for human consumption belong to the
genus *Apis*, including the domesticated honeybee, but both bumblebees and stingless bees also produce and store honey. Honey production and the sale of it often provide a significant income supplement for small farmers. Another product from honeybee colonies that is of some value is beeswax, used variously in cosmetics, pharmaceuticals, and candles.

**See Also the Following Articles**

INSECTS, OVERVIEW • PARASITOIDs • POLLINATORS, ROLE OF

**Bibliography**


INBREEDING AND OUTBREEDING

INBREEDING AND OUTBREEDING

Katherine Ralls,* Richard Frankham,† and Jonathan Ballou*

*Smithsonian Institution; †Macquarie University

I. Inbreeding
II. Inbreeding Depression
III. Inbreeding, Loss of Genetic Diversity, and Extinction
IV. Outbreeding and Outbreeding Depression

GLOSSARY

coadapted gene complexes Chromosomes, loci, and/or genes that are adapted to function well together.
effective population size ($N_e$) Size of the ideal population used in population genetics theory that would have the same rate of increase in inbreeding or decrease in genetic diversity as the actual population under study.
lethal equivalent Group of mutant alleles that would cause an average of one death if homozygous; for example, one lethal equivalent might represent two mutant alleles, each with a 50% probability of causing death, or any other combination of mutant alleles that would produce an average of one death.

INBREEDING REFERS TO MATING of related individuals. Inbreeding results in a decline in survival and reproduction (reproductive fitness), known as inbreeding depression, in most species of plants and animals and can increase the extinction risk in wild populations. Outbreeding refers to matings between individuals from different populations or subspecies. Outbreeding can result in a decline in reproductive fitness known as outbreeding depression, but this is less common than inbreeding depression. This article discusses the conservation implications of inbreeding and outbreeding depression.

I. INBREEDING

Inbreeding is the mating of individuals related by ancestry. This includes self-fertilization, brother–sister, parent–offspring, and cousin matings, as well as matings between more distant relatives. Inbred offspring are more likely to inherit recent copies of the same allele from both parents, that is, alleles that are identical by descent.

Two alleles that are identical by descent are homozygous, but not all homozygous alleles are identical by descent. A homozygous individual has two alleles at a locus that are functionally similar. However, these two alleles may or may not be identical by descent. The inbreeding coefficient (usually symbolized by $F$) of an individual is the probability that the individual has two alleles at a locus that are identical by descent. Because $F$ is a probability, it ranges from 0 for noninbred individuals to 1 for completely inbred individuals. For example, the inbreeding coefficient of an individual resulting from self-fertilization is 1/2 and that for an individual resulting from a parent–offspring or brother–sister mating is 1/4.
Inbreeding reduces the frequency of heterozygotes in proportion to the inbreeding coefficient and increases the frequency of homozygotes. Natural populations contain low frequencies of deleterious recessive mutations that are normally found as heterozygotes. Inbreeding exposes them as homozygotes, so that their deleterious effects are expressed. Consequently, in most populations of animals and plants, inbreeding results in a decline in reproduction and survival (reproductive fitness), which is called inbreeding depression.

II. INBREEDING DEPRESSION

A. Evidence for Inbreeding Depression

The deleterious effects of inbreeding were known long before the discovery of the underlying Mendelian mechanisms. In the nineteenth century, Charles Darwin clearly documented inbreeding depression based on studies in 52 species of plants, as well as the experience of livestock breeders up to his time. These early observations were amply confirmed by subsequent studies. There is now extensive evidence for inbreeding depression in laboratory and domestic animals and plants, and growing evidence for inbreeding depression in wild animals and plants. Many studies on inbreeding were massive, involving large numbers of animals over many years, and the literature is extensive. For example, Sewall Wright’s classic experiments on inbreeding in guinea pigs resulted in the production of 29,310 inbred and 5105 control young from 1906 to 1924. Only the five most vigorous inbred lines survived to the end of the experiments; 30 other inbred lines went extinct or declined so severely that Wright discontinued breeding them before the end of the experiment. Nevertheless, the surviving inbred guinea pigs were consistently inferior to the controls in number of young born, percentage of young born alive and raised to 33 days, and weight at 33 days.

Inbreeding and selection have been used to fix desirable traits in modern breeds of livestock. However, reduced fertility was a major problem during the early periods of inbreeding and the inbred lines were repeatedly outcrossed to restore vigor and fertility. The North Central regional dairy cattle breeding project (involving Iowa, Michigan, Minnesota, Missouri, Ohio, South Dakota, and Wisconsin), begun in 1947, is an example of the many extensive studies on inbreeding in livestock. This study found that inbreeding usually increased juvenile mortality and decreased milk yield, fat yield, growth, and reproductive performance.

Inbreeding depression is also well known in zoo animals. For example, in the early 1980s, Ralls and Ballou (1986) found that juvenile mortality was higher in inbred than in noninbred offspring in 41 of 44 populations of mammals in zoos, including many primates, antelopes, and deer, as well as a variety of smaller mammal species (Fig. 1). On average, the progeny of father-daughter and brother-sister matings suffered a 33% reduction in juvenile survival compared to outbred offspring. However, the severity of inbreeding depression varied widely across species. Although inbreeding in zoo animals usually results in less vigorous and fertile individuals that appear phenotypically normal, it is also responsible for some genetic diseases in captive populations, including blindness in wolves and dwarfism in California condors. Zoo populations are now routinely managed to avoid inbreeding.

Evidence for inbreeding depression in wild or semi-wild environments has been reported in several species of fish, snails, sparrows, lions, shrews, deer mice, and many species of outbreeding plants. The question of whether or not cheetahs show inbreeding depression in the wild has been extremely controversial. Efforts to resolve the controversy have been inconclusive because all populations appear inbred and there are no outbred populations for comparison. Several studies in birds have reported a lack of inbreeding depression in the wild, but these results may be due to incorrect attribution of paternity (molecular genetic studies have shown that inferences regarding paternity based on behavioral observations are often incorrect) and biases in environmental quality that favor inbred matings. Because inbreeding depression is so common, management of an unstudied outbreeding species should be based on the assumption that it will suffer reductions in reproductive fitness if it is inbred.

B. Factors Affecting the Severity of Inbreeding Depression

The degree of inbreeding depression in a population depends on the extent of inbreeding, the original frequency of deleterious recessives, and the environment. Continued inbreeding results in greater inbreeding depression. For example, average levels of various components of reproductive fitness, such as juvenile survival in mammals or grain yield in maize (Fig. 2), theoretically increase in a linear fashion as the inbreeding coefficient increases. The more deleterious alleles that were formerly masked by heterozygosity, the more severe the effects of inbreeding will be. A locus must have some form of dominance to contribute to inbreeding depression—complete dominance, partial dominance,
INBREEDING AND OUTBREEDING

FIGURE 1. Juvenile mortality in inbred and noninbred young in 45 populations of mammals in captivity. Inbreeding levels were calculated with respect to the founders of the population for which pedigree data were available. Most populations were founded with wild-caught animals but some were founded with animals from other zoos or of unknown origin. The noninbred category included all animals with an inbreeding coefficient of zero and the inbred category included all those with an inbreeding coefficient greater than zero. For the larger species, all young surviving to 6 months or more were considered to have survived. For the small species, one-half the age at sexual maturity was used as the criterion age. Numbers above the bars indicate sample sizes for each species. Reprinted from Ralls and Ballou, Captive breeding programs for populations with a small number of founders, © (1986) p. 16, with permission of Elsevier Science.

FIGURE 2. Decreasing juvenile survival with increasing inbreeding in golden lion tamarins (Leontopithecus rosalia). The proportion of offspring with a given inbreeding coefficient surviving to 7 days of age is plotted against the inbreeding coefficient. The sizes of the ovals represent the number of offspring available to estimate survival for each level of inbreeding. Large ovals represent samples of more than 50 offspring, medium ovals from 30 to 50, small ovals from 10 to 30, and tiny ovals from 1 to 10.

or overdominance (heterozygotes superior to both homozygotes). Most inbreeding depression is thought to be due to dominant rather than overdominant loci. Inbreeding depression is greater in more stressful conditions. For example, Dudash (1990) found that selfed progeny of one species of plant showed 75% inbreeding depression in the field but only 59% in garden plots and 53% in a greenhouse. Because estimates of the cost of inbreeding are based on captive populations of animals and plants maintained under relatively unstressful conditions, the deleterious effects of inbreeding on natural populations are probably higher than current data suggest.

C. Measuring Inbreeding Depression as Lethal Equivalents

The extent of inbreeding depression in survival in animals can be measured in terms of lethal equivalents. The number of lethal equivalents per gamete or individual can be calculated from the rate that juvenile survival
INBREEDING AND OUTBREEDING

decreases with increasing inbreeding. Ralls et al. (1988) found an average of 4.6 lethal equivalents per individual in 40 captive populations of mammals. Thus, each individual contained deleterious mutations that would be equivalent to slightly less than 5 lethal mutations if they were homozygous. This figure is similar to estimates for humans and birds. However, the number of lethal equivalents varied widely across captive populations.

D. Inbreeding Depression in Total Reproductive Fitness

Most studies of inbreeding depression measure only one or a few components of reproductive fitness. However, all components of reproductive fitness are subject to inbreeding depression. In animals, this includes offspring survival, number of offspring per female, male mating ability, sperm quality, and the quality and quantity of maternal care. Over a wide range of species, O. H. Frankel and M. E. Soulé noted that each 10% increase in the inbreeding coefficient caused approximately a 5–10% decline in the mean values of individual components of reproductive fitness but a 25% decline in total reproductive fitness. This indicates that inbreeding depression is approximately three times greater for total reproductive fitness than for its individual components.

E. Variation in Susceptibility to Inbreeding Depression

Inbreeding depression has a large chance (stochastic) element because of the random sampling of alleles during reproduction. Individuals with the same inbreeding coefficient, that is, the same probability of carrying alleles identical by descent, differ in actual levels of homozygosity and some are more fit than others. Families and populations within a species carry different types and numbers of deleterious mutations and differ in their susceptibility to inbreeding depression. Differences in the extent of inbreeding depression among lineages within species have been reported in mice, dairy cattle, fruit flies, and flour beetles and are to be expected in all outbreeding species. These differences contribute to the occasional success in establishing inbred lines or wild populations from the progeny of a small number of founding individuals even in species that typically show high levels of inbreeding depression. Differences among populations within species have also been reported. No differences in susceptibility to inbreeding depression among major taxonomic groups are known, but the relevant data are limited.

F. Inbreeding and Outbreeding in the Wild

Only a small proportion of species regularly self-fertilize and these tend to have life histories that strongly favor mating with relatives. For example, many are colonizing species in which the chances of successfully dispersing are much greater if only a single individual has to reach new habitat. The restricted taxonomic distribution of selfing to some plants, terrestrial slugs, and marine invertebrates suggests that it is an evolutionary dead end.

The majority of species appear to be naturally outbreeding. Many plants that are pollinated by insects have elaborate morphological mechanisms that favor cross-pollination. In other plants, cross-pollination is ensured because the male and female gametes do not mature at the same time. Another mechanism that prevents self-fertilization is self-sterility, in which pollen either fails to germinate on a stigma of its own flower or germinates but does not develop sufficiently to fertilize the egg.

Many animals are also thought to avoid close inbreeding. Although a few mammals, notably the naked mole rat, normally mate with close relatives and are highly inbred, most mammals and birds rarely mate with close relatives. Sex differences in dispersal patterns often limit opportunities for inbreeding. In most mammals, males tend to disperse more frequently or farther than females, whereas the reverse is true in birds. Furthermore, many species can recognize close kin, and a variety of species such as jays, woodpeckers, mice, voles, ground squirrels, black-tailed prairie dogs, and chimpanzees are known to actively avoid mating with them.

G. Inbreeding Depression in Small Populations

Inbreeding is unavoidable in small, closed populations because all individuals eventually become related to each other. Inbreeding in a population of size \( N \) increases at a rate of \( \frac{1}{2N} \) per generation. For example, in a population of size 10, there is a 5% increase in inbreeding per generation. Consequently, small isolated populations that have existed for many generations are expected to show inbreeding depression. Small populations of plants, fruit flies, a rock wallaby, Florida panthers, and a snake have been found to suffer from inbreeding depression. However, inbreeding depression may not cause declines in population size. Reduced
INBREEDING AND OUTBREEDING

Fecondity and survival will only cause a population decline if the reproductive rate drops below replacement level. Small populations can also suffer from reduced reproductive fitness because of declines in environmental quality.

II. Reduction of Inbreeding Depression

Inbreeding depression may be reduced by selection against deleterious alleles, which eliminates, or purges, them from the population. Purging has been documented in plants, mice, birds, and fruit flies and during the development of inbred lines in a variety of species. Species that naturally inbreed generally show less inbreeding depression than naturally outbreeding species because of the greater opportunity for selection against deleterious recessives. Most populations of self-fertilizing plants show some level of inbreeding depression, but those with higher selfing rates tend to show less inbreeding depression (Fig. 3). Slower inbreeding generally causes less inbreeding depression than an equivalent amount of rapid inbreeding because there is more time for natural selection to operate, but this effect is often small.

Purging may reduce inbreeding depression, but it is unlikely to eliminate it. Ballou (1997) found that selection led to a small reduction of inbreeding depression in neonatal survival in 15 of 17 populations of captive mammals in zoos. No trends in purging effects were observed in survival to weaning or litter size. The purging effects were not strong enough to be of practical use in captive breeding programs. An earlier study by Templeton and Read (1984) reported that purging reduced inbreeding depression in captive Speke’s gazelles. However, a reanalysis of their data has cast serious doubts on this interpretation.

The effectiveness of selection in reducing inbreeding depression varied substantially in three subspecies of deer mice, probably due to different histories of inbreeding and selection. Lacy and Ballou (1998) found that purging was effective in a subspecies collected in relatively continuous habitat that probably experienced little or no inbreeding in the wild. Continued inbreeding reduced inbreeding depression in four of seven fitness components in this subspecies, suggesting that the original inbreeding depression was caused by highly deleterious recessive alleles that were quickly removed by selection. A second subspecies collected from ephemeral patches of fragmented habitat showed no reduction in inbreeding depression in any component of fitness. Episodes of local inbreeding in the wild may have already removed strongly deleterious recessive alleles from this subspecies. The third subspecies was a coastal form known to experience periodic population bottlenecks in the wild owing to devastating hurricanes. This subspecies had low productivity even prior to experimental inbreeding, and inbreeding in the laboratory resulted in a complete collapse of reproductive fitness. Repeated bottlenecks may have purged this subspecies of deleterious recessive alleles but also reduced heterozygosity to the point that further reductions in heterozygosity from inbreeding led to greatly reduced fitness.

Inbreeding depression can be completely reversed by outcrossing an inbred population to another unrelated population, either an outbred population or an independently inbred population. Partial recovery can be achieved by introducing unrelated individuals into the population. For example, Spielman and Frankham (1992) found that introducing a single immigrant into partially inbred populations of fruit flies increased reproductive fitness about halfway back to that found in the original outbred population.

III. INBREEDING, LOSS OF GENETIC DIVERSITY, AND EXTINCTION

Many conservation biologists have been concerned that inbreeding and loss of genetic diversity in small populations will increase the risk of extinction. However, some researchers have questioned this view because of the difficulty of obtaining direct evidence that inbreeding contributes to the extinction of wild populations.
A. Inbreeding and Extinction in Captivity

There is overwhelming evidence that deliberately inbred populations of laboratory and domestic animals and plants suffer elevated extinction rates. Most attempts to develop inbred lines by intense inbreeding are unsuccessful. For example, Bowman and Falconer (1960) found that 19 of 20 colonies of laboratory mice maintained by sibling mating became extinct by generation 12, but the remaining colony showed no decline in litter size and became a successful inbred line. Results with fruit flies, guinea pigs, poultry, and Japanese quail are similar. Extinctions occur even with extremely slow rates of inbreeding due to small population size. For example, 15 of 60 captive populations of fruit flies with an effective population size of about 50 went extinct over 210 generations.

However, extinctions of inbred lines can be caused by inbreeding, random fluctuations in birth and death rates in very small populations (demographic stochasticity), or a combination of both. Frankham (1995) found four data sets with sufficient information to distinguish genetic from nongenetic effects and examined the shape of the relationship between inbreeding level and extinction. Although the relationship between inbreeding level and inbreeding depression in individual components of reproductive fitness is theoretically linear, the relationship between inbreeding level and extinction showed a threshold effect in all cases (Fig. 4). Extinctions were rare at low levels of inbreeding but sharply increased when inbreeding reached intermediate levels. Populations inbred at different rates (full sibling versus double first cousin mating) to similar inbreeding coefficients showed similar levels of extinctions. The reason for the threshold relationship between inbreeding and extinction is that it generally takes several generations and considerable inbreeding until the growth rate of the population becomes negative. The amount of inbreeding required to reach the extinction threshold is likely to vary in different populations, as is the rate with which extinction increases with inbreeding after that point. The existence of a threshold relationship suggests that there may be little warning of impending inbreeding problems in endangered species, few of which are closely monitored.

The effects of inbreeding on population viability are complex and will interact with other factors affecting population growth, population fluctuations, or $\sqrt{N}$, but they will be deleterious in the long term.

B. Inbreeding and Extinction Risk in Wild Populations

Conservation biologists initially thought that genetic problems played a role in the endangerment and extinction of wild populations. It was clear that inbreeding increased extinction risk in laboratory populations and theoretical work in the 1980s suggested that small populations in the wild also should suffer from increased extinction due to the unavoidable increase in mating between close relatives. However, in an influential paper, Lande (1988) argued that random demographic and environmental events will drive small wild populations to extinction before genetic factors cause problems. Environmental events, ranging from annual variation in climatic variables (such as rainfall) to catastrophes (such as fires and disease epidemics), do increase the probability of extinction and it is extremely difficult to isolate genetic effects from these other effects because inbreeding typically interacts with demography by reducing fecundity, juvenile survival, and life span. Consequently, other researchers, such as Caughley (1994), continued to question the significance of genetic factors because of the lack of direct evidence that inbreeding can contribute to the extinction of wild populations.

The first direct evidence was provided by Saccheri and colleagues (1998) with their work on the Glanville fritillary butterfly in Finland. They studied a metapopulation (a population of populations) of this butterfly plant species. Inbreeding and extinction were monitored over several generations, and the results showed that inbreeding did contribute to the extinction of individual populations. This finding has important implications for conservation efforts, as it suggests that strategies to maintain genetic diversity may be necessary to prevent the extinction of wild populations.
that consists of numerous small populations that breed in about 1600 dry meadows of different size at varying distances from each other. Caterpillars feed in conspicuous family groups of 50 to 250 individuals and the smallest populations consist of the offspring of a single pair. Populations in individual meadows often go “extinct,” but many meadows are eventually recolonized, with an average of 200 “extinctions” and 114 colonizations per year.

Because small population size results in both inbreeding and loss of genetic variation, the researchers were able to use the degree of genetic variation in each population as a measure of the extent to which it was inbred. They sampled 42 populations and found that populations with less genetic variation were more likely to go extinct. Furthermore, genetic diversity predicted extinction risk after they had accounted for all other known causes of extinction in this well-studied butterfly metapopulation. Inbreeding reduced egg hatching rate and larval survival, lengthened the duration of the pupal period (so that inbred pupae were more likely to be parasitized), and shortened female life span (so that inbred females tended to lay fewer eggs).

Furthermore, inbreeding depression likely played a role in the decline of greater prairie chickens in Illinois. Egg fertility and hatching success declined as the population became smaller. However, when prairie chickens from larger populations in neighboring states were released in Illinois, hatching success improved.

Several indirect lines of evidence also imply that inbreeding depression can be a problem in wild populations. First, theoretical studies suggest that genetic factors probably contribute to extinctions. Second, genetics may be one of the factors that makes island populations prone to extinction. These populations have lower genetic diversity than mainland populations and many are inbred to levels where captive populations show increased risk of extinction from inbreeding. Third, empirical estimates indicate that effective population size is a smaller fraction of actual population size than previously suspected, which means that genetic problems will arise at larger population sizes than previously believed. Fourth, endangered species tend to have lower genetic diversity than nonendangered species. This would not be expected if other factors drove populations to extinction before genetic factors became important. Finally, the extinction rate of a wild plant was higher in experimental populations with low genetic variation than in those with high genetic variation when both were planted in the field.

Mating between distantly related individuals, such as individuals from different populations or subspecies, is called outbreeding. Crossing populations may increase reproductive fitness by increasing heterozygosity and thus preventing the expression of deleterious recessive alleles or may decrease fitness from the disruption of coadapted gene complexes. If outbred offspring have lower reproductive fitness than nonoutbred offspring, it is called outbreeding depression. The question of whether crossing geographically distinct populations of a species usually has beneficial or detrimental effects is important for conservation, but unfortunately there are very few data on the effects of crossing natural populations, particularly among vertebrate populations.

A. Possible Causes of Outbreeding Depression

Outbreeding depression could result from two mechanisms. The genetic mechanism requires that different populations evolve coadapted gene complexes. Crossing individuals from populations with different coadapted gene complexes could then disrupt these complexes and reduce reproductive fitness. The ecological mechanism of outbreeding depression requires that populations develop different adaptations in response to different local environments. Crossing individuals from the populations may then produce progeny that are less well suited to either local environment.

B. Evidence for Outbreeding Depression

Evidence for outbreeding depression comes primarily from organisms with extremely limited dispersal, such as some plants, copepods, and scale insects, or from crosses between individuals from vastly different geographic sources or with significant chromosomal differences. Outbreeding depression appears to be more common in plants than in animals. There is not much evidence for outbreeding depression in animals, although it has been observed in crosses where there are chromosomal differences between populations such as in dik-diks and spider monkeys. Such cases usually indicate the existence of unrecognized species or subspecies. The most widely quoted mammalian case concerns ibex in Slovakia, but this example is questionable.
### TABLE I
Relative Fitness of Plant and Animal Hybrid Classes

<table>
<thead>
<tr>
<th>Genus</th>
<th>Hybrid classification</th>
<th>Fitness measurement</th>
<th>Natural/ manipulations</th>
<th>Fitness*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quercus</td>
<td>F₁</td>
<td>Fruit maturation</td>
<td>M</td>
<td>L (L–E)</td>
</tr>
<tr>
<td>Artemisia</td>
<td>Hybrid</td>
<td>Developmental stability</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Artemisia</td>
<td>Hybrid</td>
<td>Herbivore attack</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Iris</td>
<td>F₁ jelluroide</td>
<td>Shade tolerance</td>
<td>M</td>
<td>1 (I–H)</td>
</tr>
<tr>
<td>Iris</td>
<td>Eight genotypic classes</td>
<td>Viability of mature seeds</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Eucalyptus</td>
<td>E. radiata backcross</td>
<td>Reproductive parameters</td>
<td>N</td>
<td>1 (I–H)</td>
</tr>
<tr>
<td>Iris</td>
<td>F₁-type hybrid</td>
<td>Reproductive parameters</td>
<td>N</td>
<td>L</td>
</tr>
<tr>
<td>Hyla</td>
<td>F₁</td>
<td>Developmental stability</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Hyla</td>
<td>C. c. backcross</td>
<td>Developmental stability</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Hyla</td>
<td>C. g. backcross</td>
<td>Developmental stability</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Sceloporus</td>
<td>Heitzeyngus for chromosomes 1,3,4,6</td>
<td>Chromosome segregation in males</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Sceloporus</td>
<td>Parental chromosome 2</td>
<td>Chromosome segregation in males</td>
<td>N</td>
<td>L</td>
</tr>
<tr>
<td>Sceloporus</td>
<td>Recombinant chromosome 2</td>
<td>Chromosome segregation in males</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Coloptes</td>
<td>Hybrid</td>
<td>Clutch and breed size</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Geospiza</td>
<td>G. fortis/scandens F₁</td>
<td>Survivorship, recruitment, breeding success</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td>Geospiza</td>
<td>G. fortis/scandens F₁</td>
<td>Survivorship, recruitment, breeding success</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td>Geospiza</td>
<td>G. fortis x G. fortis/scandens F₁</td>
<td>Survivorship, recruitment, breeding success</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td>Allonemobius</td>
<td>Hybrid</td>
<td>Survivorship</td>
<td>N</td>
<td>1 (I–H)</td>
</tr>
<tr>
<td>Mercenaria</td>
<td>M. mexicana recombinant</td>
<td>Survivorship</td>
<td>N</td>
<td>L</td>
</tr>
<tr>
<td>Notrops</td>
<td>Hybrid</td>
<td>Survivorship</td>
<td>N</td>
<td>1 (I–E)</td>
</tr>
<tr>
<td>Bombina</td>
<td>F₁</td>
<td>Viability</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>Bombina</td>
<td>Hybrid</td>
<td>Viability</td>
<td>M</td>
<td>E</td>
</tr>
<tr>
<td>Apsis</td>
<td>F₁</td>
<td>Metabolic capacities</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>Gastorhopis</td>
<td>Hybrid</td>
<td>Foraging efficiency</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>Gambusia</td>
<td>G. hallmichi x G. affinis för</td>
<td>Development</td>
<td>M</td>
<td>H</td>
</tr>
<tr>
<td>Gambusia</td>
<td>G. affinis x G. hallmichi för</td>
<td>Development</td>
<td>M</td>
<td>1</td>
</tr>
</tbody>
</table>

*Each of these examples involves taxa that are known to hybridize under natural conditions.


*Natural (N) refers to those measurements taken from naturally occurring hybrids and manipulations (M) from experimental manipulations.

*Fitness estimates are relative to both parental species (L = lowest fitness, I = intermediate to both species, E = equivalent to both species, H = highest fitness). The most common fitness for any particular class is given, with the range of fitness values for particular classes given in brackets.
as no quantitative data were presented in the original account. Several studies have failed to find outbreeding depression in mammals, including in rhesus macaques and saddle-back tamarins. Ballou (1995) found no outbreeding depression in survival in several captive populations, including Borneo and Sumatran orangutans. The most extensive study in mammals is that of Lacy and colleagues (unpublished), who conducted numerous crosses among five subspecies of deer mice. These subspecies included closely related populations from similar, contiguous habitats as well as very divergent subspecies that had long been isolated in dissimilar habitats. In general, the benefits of increased heterozygosity outweighed the costs of disrupted gene complexes. The reproductive success of crosses was always superior to the mean for both parental stocks, although not always above that of both parental stocks. Hybrid vigor (heterosis) was not confined to the F1 generation but continued into later generations. Experiments designed to partition the effects of heterozygosity from the effects of disrupting coadapted gene complexes indicated negative effects of disrupting coadapted gene complexes on some components of reproductive success in one long-isolated subspecies, but these effects were smaller than the beneficial effect of increased heterozygosity. A cross between two other subspecies indicated that disruption of gene complexes in the F2 and backcross generations actually increased litter viability.

C. Variation in Susceptibility to Outbreeding Depression

In general, the probability of outbreeding depression increases as populations become more distantly related. Small, isolated populations of species with poor dispersal abilities and naturally high rates of inbreeding that are subject to different selective pressures are more likely to suffer from outbreeding depression when crossed than are large populations of wide-ranging species with good dispersal abilities and low rates of inbreeding.

Small populations of naturally inbreeding species must be distinguished from populations of normally outbreeding species that have recently been reduced in size, isolated, and inbred as a result of human activities such as habitat destruction. In recently isolated populations, genetic differentiation may be the result of random drift rather than local adaptations. For example, genetic differences among extant gray wolf populations in North America are thought to reflect recent population declines and habitat fragmentation rather than a long history of genetic isolation. In such cases, outbreeding is likely to benefit, rather than decrease, fitness by restoring heterozygosity and masking deleterious alleles. For example, the small, isolated population of endangered Florida panthers shows signs of inbreeding depression, and individuals from its nearest subspecies in Texas have been added to the population in attempt to increase its reproductive fitness.

Some authors, such as W. M. Shields, have suggested that outbreeding depression is widespread and might be as important as inbreeding depression as a conservation concern. However, a recent review by Arnold and Hodges (1995) found that even hybrids between species are not uniformly unfit but may have lower, equivalent, or higher levels of fitness than their parents. Hybrid genotypes showed a wide range of fitness values, but the general pattern was that hybrids demonstrated either equivalent fitness to the two parental taxa or higher levels of fitness than at least one of the parents (Table I). The results of mixing populations or subspecies always depend on the specific animal or plant groups involved. However, outbreeding depression is rarer than inbreeding depression and is of less conservation concern.

See Also the Following Articles

ECOLOGICAL GENETICS • GENETIC DIVERSITY • POPULATION GENETICS

Bibliography

INDICATOR SPECIES

John H. Lawton* and Kevin J. Gaston†

*Imperial College, United Kingdom and †University of Sheffield, United Kingdom

I. Species as Indicators of the State of the Environment
II. Species as Indicators of Environmental Change
III. Species as Indicators of Biodiversity
IV. Conclusions

GLOSSARY

acid deposition Anthropogenic acidification of terrestrial and freshwater ecosystems by (primarily) sulfurous acid, derived from sulfur dioxide produced by burning oil and coal and deposited in rain and snow (acid rain), directly as particles (dry deposition) and as cloud droplets.

all taxa biodiversity inventory (ATBI) The idea, first suggested by D. H. Janzen, that it might be feasible to produce a complete species list for all the organisms living in one place, a hectare of tropical forest, for example. The goal has so far proved elusive.

bioassay The use of living cells or organisms to make quantitative or qualitative measurements of the amounts or activity of substances.

community An assemblage of species populations that occur together in space and time.

ecotoxicology The use of test organisms (e.g., the water flea, Daphnia) to study the toxicity, pathways of accumulation, and breakdown of chemicals, particularly those manufactured by humans (e.g., pesticides).

endemic species Species confined in their distribution to a particular geographic region. The size of the region is arbitrary (a species can be endemic to North America or to a tiny island).

hot spots A word with several distinct meanings. Here it is used to denote sites unusually rich in a particular group of species (e.g., birds), compared with average sites in the same geographic region. (The converse is a "cold spot."). Has also been used to denote centers of endemism (see endemic species), which need not be unusually species rich, it is not used in this context here.

paleoclimatology The study of past climates from fossils and other traces left in the geological record.

reserve selection algorithms Mathematical techniques used to maximize efficiency in the selection of protected areas for conservation. The efficiency criteria vary with circumstances but may, for example, be the minimum number of reserves with every species represented, or minimum cost.

THIS ARTICLE REVIEWS THE USE of species as indicators of the state of the environment and of human-induced changes to the environment. We focus on five interrelated topics, namely various types of pollution, rising concentrations of atmospheric carbon dioxide, global climate change, patterns in regional and global biodiversity, and the designation of protected areas for nature conservation. Using organisms to indicate the state of, and changes to, the environment has numerous tried and tested applications. Attempts to identify indi-
I. SPECIES AS INDICATORS OF THE STATE OF THE ENVIRONMENT

There are three distinct uses of the term “indicator species” in research in ecology and biodiversity. They are a species, or group of species, that do the following:

1. Reflect the biotic or abiotic state of an environment
2. Reveal evidence for, or the impacts of, environmental change
3. Indicate the diversity of other species, taxa, or entire communities within an area

This article explains, provides examples of, and evaluates each of these uses of the term, focusing primarily on terrestrial and freshwater ecosystems; broadly similar conclusions apply to marine ecosystems, but marine examples lie beyond the scope of the article. We pay most attention to the third use of the term “indicator species,” because this seems most appropriate for an encyclopedia devoted to biodiversity. The most up-to-date evaluation and review of indicator species in the scientific literature is by McGeoch (1998). She concentrates on terrestrial insects as “bioindicators” (in all three senses of the word) but the general principles that she discusses extend to all ecosystems and organisms, not just to terrestrial insects.

Everybody knows that living organisms are sensitive to the state of their environment. Pollution from human activities kills many species and reduces the abundance of others. These changes in abundance can be used to assay the state of the environment.

A. An Example: Acid Deposition

Sulfur dioxide, produced by burning fossil fuel, particularly coal, enters the atmosphere and is eventually deposited on terrestrial and freshwater ecosystems via three routes: (a) as tiny solid particles, (b) washed from the air in rain or snow, or (c) as droplets formed in clouds. Deposition often occurs hundreds of kilometers from the source. Dissolved in water, sulfur dioxide forms sulfuric acid, resulting in what is frequently referred to as “acid rain,” but because there are three principal routes involved in its transfer to terrestrial and freshwater ecosystems, it is more correctly called “acid deposition” (Erisman and Draaijers, 1995). Sulfur dioxide is not the only source of acidification; oxides of nitrogen, again produced by burning fossil fuel, are also involved, but sulfur dioxide is the main agent of acidification in most ecosystems.

In terrestrial ecosystems, this deposition kills lichens and acidifies the soil, leading to changes in the vegetation. Lakes become progressively more acidic as deposition loads increase, until eventually they may become virtually lifeless. A trained biologist, visiting for the first time an area subject to acid deposition, will often be able to deduce that the habitat is being polluted simply by looking at the species that are present and those that ought to be there but are not. Beautifully clear Scandinavian lakes, lacking any fish or amphibians, supporting few birds and a species-poor and taxonomically unusual invertebrate fauna, have been reduced to this impoverished state by the transnational export of sulfur dioxide from coal-burning power-stations in the United Kingdom and elsewhere in Europe. Here, living organisms act as powerful indicators of the state of the environment and the damage being done to it by human activities, often performed many hundreds of kilometers away.

B. Management of European Rivers

Because the species composition and richness of biological communities change as the environment changes, we can use species as indicators of the state of the environment for practical management purposes. The techniques have been particularly well developed to assess organic and inorganic pollution in European rivers, managed for recreation, fisheries, and drinking water. The advantages of using living organisms as indicators of water quality are that they avoid the need for expensive chemical analyses, and, probably more important, organisms integrate the impacts of pollutants over space and time. All chemical traces of a major pollution incident may disappear from a river in a matter of hours as the pollution is flushed from the system. Nonetheless, the biotic community may show evidence of the damage for many months. It is extremely difficult, and prohibitively expensive, for chemists to measure all the organic and inorganic chemical pollutants entering a river, and it is certainly impossible for them to work out what all the combined impacts of such a cocktail might be. But living communities reflect the integrated effects of all the compounds that find their way deliberately and accidentally into watercourses, and hence they act as sensitive indicators of the state...
of the environment. A valuable source of further information on the use of living organisms to monitor the environmental health of rivers and lakes is provided by Rosenberg and Resh (1993).

Two widely used European examples of this approach are the German Saprobic Index, and RIVPACS in the United Kingdom. RIVPACS is now used by the Environment Agency to manage UK rivers. Both the German and UK approaches require accurately identified samples to be taken of the organisms found along sections of the river. RIVPACS focuses on invertebrates, the German index on invertebrates, microbes, and higher plants. Both rely on the fact that some species are extremely tolerant of pollution (the aquatic larvae of some chironomid midges, for instance), while others are extremely sensitive, particularly to the low oxygen levels produced by organic pollution (for instance, the larvae of many mayflies). The species present in the samples are given scores, depending on their known tolerances, and the data from all species are combined to produce a composite and very sensitive index of pollution levels for any particular section of a river.

C. Widespread Application

Use of organisms to indicate the state of the environment is widespread, taxonomically and geographically, for a wide range of environmental issues. Use of species as indicators of the state of the environment is not confined to freshwater, or to Europe and North America. A wide variety of organisms has been suggested, or used, as indicators of human impacts. In Europe, suites of plant, fungal, and insect species are only found in, and hence are good indicators of, ancient woodland; they are entirely absent from plantations, even though these may be several hundred years old. The use of lichens as sensitive indicators of air pollution is well known, but organisms as different as mites and geckos (agile, climbing lizards) have been used, or suggested, for similar purposes. Lichens have also been used as indicators of fire history in Brazilian cerrado (a type of dry, scrubby forest), tiger beetles as indicators of tropical forest degradation in Venezuela, and day-flying Lepidoptera (butterflies and moths) as indicators of the state of seminatural grasslands for conservation in Europe. Many other similar examples exist.

D. Interpretation Requires Care

In these and similar cases, considerable care is needed before a species or group of species can be used as reliable indicators of damaging (or beneficial) human impacts on ecosystems. All populations of living organisms fluctuate over time and vary in abundance spatially, because of natural variations in the weather, normal changes in the physical environment, and fluctuations in the abundances of natural enemies, competitors, and essential resources (food and shelter). Just because one or more species is declining does not mean that human impacts are to blame. In the case of lichens and atmospheric pollution or freshwater invertebrates and river quality, the links between anthropogenic pollutants and changes in the distributions and abundances of organisms are thoroughly researched and well understood. But even quite major declines in some species have proved exceptionally difficult to link to damage to the environment caused by people.

E. Amphibian Decline

The so-called amphibian decline is a particularly dramatic example (Blaustein and Wake, 1995). In many parts of the world, population biologists interested in amphibians (frogs, toads, news, salamanders, etc.) have recently become alarmed by apparent major declines in the abundance, and the complete disappearance, of many species from areas where formerly they were common, often in regions apparently remote from human impacts. The declines are not happening everywhere, and the magnitude of many of those that have been claimed is difficult to assess because of the lack of long-term data prior to the supposed population collapses; some of them may be perfectly natural. The worrying aspects of the phenomenon are that while it is apparently global in scope, the causal mechanism (or mechanisms) remains obscure. It has been suggested, for example, that the amphibian decline is indicative of rising global levels of damaging ultraviolet light (UV-B) caused by loss of the earth’s protective stratospheric ozone layer. Amphibian eggs, exposed in shallow water, and the adults with their thin wet skins may be particularly sensitive to UV-B, as are human sunbathers without sunblock. Others doubt the explanation. More recently a global pandemic has been implicated. But what should suddenly trigger lethal outbreaks of disease in amphibians is unclear.

F. Environmental Toxicology

In all the examples so far, the organisms being used as actual or possible indicators of environmental health have been in their natural environment. There is another related but quite separate way in which biologists use the sensitivity of organisms to set environmental
standards, namely in the science of environmental toxicology, or ecotoxicology for short. In many areas of human endeavor, the aim is to apply some beneficial technology with minimum environmental damage. Crop spraying with pesticides is a good example, and so is the discharge of treated effluent from a factory. Some environmentalists claim that these types of operations should not lead to any environmental contamination; factories should have zero discharges, and if we must use pesticides, they should be targeted to reach only the crop and the pest and not, for example, the soil, nontarget organisms, or adjacent watercourses. However, zero discharges or precision pesticides, if they can be achieved at all, can often only be obtained at great economic cost. The more pragmatic solution is to ask whether there are minimal levels of discharge, spray drift into watercourses, and so forth that cause no detectable environmental damage. To provide answers to this admittedly difficult question, environmental toxicologists use a wide variety of laboratory bioassays with standard organisms. Examples from freshwater include the alga, *Chlorella vulgaris*, the water flea, *Daphnia magna*, the amphipod shrimp, *Gammarus pulex*, and the rainbow trout, *Salmo gairdneri*. The fundamental problem is to try and establish acceptable levels of contamination. Defining “acceptable” obviously requires political as well as biological judgment. However, traces of a compound in water, air, or soil that cause no detectable changes in the performance (growth, survival, or reproduction) of the test organisms are clearly more acceptable than doses that kill 50% of the population (so called LD50 levels). Basically, the bioassays seek to set environmental standards for levels of potential pollutants in soil, air, and freshwater, using a range of standard laboratory organisms as indicators (Shaw and Chadwick, 1998), but there can be no absolute standards about what is safe or acceptable. The general trend in modern societies is for standards to gradually tighten.

II. SPECIES AS INDICATORS OF ENVIRONMENTAL CHANGE

If the amphibian decline (discussed in the previous section) is real, it is an example of a group of organisms acting as indicators not only of the state of the environment, but also as indicators of ongoing changes to the global environment, albeit of an unknown nature. In other words, given that species are sensitive to the condition of their environment, monitoring organisms not only tells you about the current state of an environment, but repeated monitoring can tell you about changes in that environment. To act as indicators of change rather than current environmental health, it is necessary to have at least two sets of data on the particular indicator species in question, taken in the same way, at the same place(s), on two separate occasions. More frequent sampling allows greater confidence in the direction of apparent trends and the detection of more subtle environmental changes.

A. Not All Monitoring Is about Environmental Degradation

Not all monitoring of species seeks to record environmental degradation. Increasingly after mining operations, for example, mine operators are required to restore spoil heaps and mine pits by sowing or planting native vegetation. Monitoring selected groups of common animals on nearby undisturbed control sites and on the restored land can give a good indication of the recovery of the entire ecosystem and of the success of the restoration project. For instance, when biologists monitored ant assemblages on abandoned, replanted bauxite mines in Australia, they found that the ants provided a good indication of the recovery of these ecosystems. Even after 14 years there were still differences between the ant communities found in the natural *Eucalyptus* forest and the restored land.

B. Historical Records of Change

1. Lake Acidification

It may not always be necessary to sample in real time. When anthropogenic acidification of lakes was first discovered, many people doubted that the phenomenon was real. In particular, there was considerable opposition to the notion from the power-generating industry, because solving the problem (by burning low-sulfur coal, adding “scrubbers” to power station chimneys to remove sulfur dioxide, or switching to natural gas) was inevitably going to be expensive. After all, there were few historic data on the state of the acidified lakes. Perhaps they had always been that way?

Resolving the problem required knowledge of the fact that lake phytoplankton (the tiny, unicellular plants that float in the upper layers of lakes) are extremely sensitive environmental indicators, because different species grow best in very different conditions determined by nutrient status and pH (acidity). When algae die, they sink to the bottom where their bodies and
characteristic pigments are buried and some are preserved (incipient fossils), particularly the resistant, siliceous outer cases of a group called diatoms. An undisturbed core through the sediments records the history of a lake’s phytoplankton, with the oldest flora at the bottom. Cores showed unequivocally that many Scandinavian lakes that are acid now were not acid before the Industrial Revolution; the oldest diatoms—species not found in acid lakes—are gradually replaced in the sample column by acid-tolerant species. Diatoms are wonderfully sensitive indicators of environmental change (Fig. 1).

2. Plants and Carbon Dioxide

Herbarium specimens (pressed plants collected for taxonomic purposes) and fossil leaves can also be used as indicators of past environmental change. Another consequence of the rapid rise in the burning of fossil fuel since the Industrial Revolution has been an accelerating rise in the concentration of atmospheric carbon dioxide, one of the main agents of "global warming." We will deal with species as indicators of anthropogenic global climate change (as it is more accurately known) later. Here we want to focus on physiological and developmental responses within single species to rising carbon dioxide.

If plants are grown in a greenhouse under different atmospheric carbon dioxide concentrations, from below the pre–Industrial Revolution levels of about 280 parts per million by volume (ppm), through what are roughly present levels of 350 ppm, to levels that may be reached by the end of the 21st century (700 ppm), several interesting things happen. In particular, in the present context, stomatal densities on the undersides of the leaves decline. Stomata are the tiny pores in the leaf surface through which plants take up carbon dioxide (needed for photosynthesis), and through which they lose water vapor. It has been known for a long time that plants control the opening and closing of stomata to optimize carbon dioxide uptake and reduce water loss. More surprising, we now also know that plants grown in high carbon dioxide have lower densities of stomata; something happens during leaf development to reduce the number of stomata. How and what is currently unclear. Why is simple enough. In a high carbon dioxide world, the plant needs fewer stomata to take up the carbon dioxide it requires and hence can satisfy the needs of photosynthesis and reduce water loss by developing fewer pores in the leaves.

Now back to those herbarium specimens and fossil leaves. If you look at 200-year-old (and very precious) herbarium and modern specimens of the same species,
sure enough, stomatal densities decline as global atmospheric carbon dioxide levels increase (Fig. 2A). The same approach has recently been used to try and trace atmospheric carbon dioxide levels throughout most of the Phanerzoic, from the time when plants first colonized the land. Here the method is more contentious, because different species of truly fossil plants with presumed similar growth forms have to be used in different geological periods. Nevertheless, the pattern of apparent changes in global atmospheric carbon dioxide concentrations over hundreds of millions of years, revealed by this method (Fig. 2B and C), are in reasonable agreement with alternative, independent, and also contentious geochemical methods. Here is a really unusual use of species as indicators of environmental change.

C. Species as Indicators of Climate Change

1. The Sensitivity of Species to Climate: Fossils Again

Current, rapidly rising concentrations of atmospheric carbon dioxide are the primary cause of anthropogenic
global climate change. However, the earth's climate has always changed, naturally, with no intervention from human beings. One of the ways we know this is through the careful documentation of the types and distributions of organisms in the fossil and subfossil record. The science of paleoclimatology, which seeks to reconstruct the history of earth's climate, relies heavily on changes in fossil and subfossil species assemblages to deduce what the earth's climate was like thousands or even millions of years ago. To take one example, in the modern world many types of corals occur exclusively in tropical marine environments; it is a winning bet that fossil corals of the same type indicate an ancient tropical sea, even though the rocks bearing the fossils may now lie in much colder parts of the world.

In more recent geological time, we can use changes in the distributions and abundances of plants and animals to trace major changes in the earth's climate during the Holocene (the most recent geological past) and Pleistocene glacial and interglacial. Plant remains preserved in peat and lake sediments record in exquisite detail the march northward of European and North American forests from the end of the last glaciation 12,000 years ago (Huntley and Birks, 1983). The forests spread with remarkable speed (an average of about 200 m per year, but sometimes as fast as 2 km a year) to achieve present distributions in the northern parts of both continents from glacial refugia thousands of kilometres to the south (Fig 3A and B). The information is not won easily. It requires huge patience and great skill to identify thousands upon thousands of pollen grains extracted onto microscope slides. But once done, the record reads like a speeded-up movie, as spruce, oaks, white pine, hemlock, beech, and chestnut swept north in successive waves through what is now the United States and Canada; in the more species-poor forests of Europe, pines were followed by birch, then oak. These invasions are as dramatic as any in human history, but they were silent and recorded only by pollen grains.

2. Contemporary Changes in Species Distributions

Historical changes aside, there is now no doubt that the world is currently warming quite rapidly. An upward trend in global annual mean surface temperatures is apparent from about 1920, particularly over the last two decades (from c. 1980), global mean surface temperatures in July 1998 were the highest ever recorded. Do organisms act as indicators of these changes, perhaps, as with the freshwater species discussed earlier, acting subtly to integrate several of the changes human's find difficult to comprehend in the bald statistics? Climate change does not simply involve warming; it involves changes in rainfall, extreme weather events (droughts and storms), and even locally cooler conditions. All these complex changes should show up in changes in the distributions and abundances of organisms.

They do. Species are proving to be extremely sensitive indicators of contemporary climate change, where historical records allow decent reconstruction of former and current distributions. Populations of Edith's checkerspot butterfly Euphydryas editha are disappearing from southern California and northern Mexico, at the current southern end of its distribution, and from more lowland sites; sites where previously recorded populations still exist are on average 2° further north than sites where populations went extinct (Fig. 4). These are exactly the changes we would expect in a warming world. Twenty years ago in northwest Europe, little egrets Egretta garzetta (small white herons) used to be rare visitors from the Mediterranean. Now they are breeding in northern France and southern England in astonishing expansion of range. Populations of many other European birds, butterflies, and other organisms are spreading north at the present time, as the climate warms.

Of course, none of this tells us whether the climate change that is certainly happening is "natural"—it could have happened anyway and may have nothing to do with anthropogenically produced greenhouse gases—or whether it is indeed due to human activities. Using species as indicators of climate change tells us unequivocally that the earth's climate is changing, but so does the mercury in the thermometer. What neither tells us is why, and no end of work on species as indicators will solve that dilemma. As we have already seen, this situation is not unique to climate change. It generally holds whenever we use species as indicators of the state of the environment. Indicator species can tell us whether an environment is, or is not, changing. They do not tell us why the changes are taking place. That almost always requires additional detective work, although knowledge of an organism's biology will frequently provide valuable clues. Three examples, using birds as indicators, illustrate the problem in more detail.
D. Birds as Indicators of Large-Scale Environmental Changes

Birds are widely used indicators, because in Europe, North America, and other parts of the world where there are large armies of amateur bird watchers their populations and distributions have been recorded well enough, for long enough, to reveal major environmental trends.

1. Peregrine Falcons and DDT

The catastrophic collapse of peregrine falcon *Falco peregrinus* populations throughout the northern hemisphere in the 1950s signaled widespread contamination of the environment by chlorinated hydrocarbon insecticides, first DDT, then other compounds such as aldrin and dieldrin. The total, and rapid, disappearance of these dramatic birds signaled to ornithologists that something was seriously wrong with the environment, but what? It took a great deal of clever biological detective work (see Ratcliffe, 1980) to link the decline of peregrine populations to the accumulation of these persistent pesticides up the food chain, resulting in eggshell thinning, reproductive failure, and (in extreme cases) direct poisoning of adult birds. Although some populations have now recovered, signaling a recovery in environmental quality, the species is still missing from many parts of its former range—some coastal populations in England, for instance. Nobody knows why.

2. Migratory Songbird Declines in North America

In North America, considerable concern is currently being expressed over widespread declines in summer migrant birds, particularly warblers. Unlike the so-called amphibian decline, nobody questions the phenomenon; just like the amphibian decline, nobody re-
FIGURE 4  The fate of 151 previously recorded populations of Edith's checkerspot butterfly, Euphydryas editha, in western North America. The populations ranged from northern Mexico to southern Canada and were visited by Camille Parmesan and other biologists between 1992 and 1996. Populations that had disappeared because of habitat degradation (e.g., loss of usable host plants) were omitted from the analysis. Dividing the populations into five, evenly spaced latitudinal bands between 30° N and 53° N (A) reveals that significantly more southern populations have gone extinct than northern populations; sites where previously recorded populations still exist were, on average, 2° further north than sites where populations were extinct. Extinctions were also higher at lower altitudes (B) (n is the number of populations in each latitudinal or altitudinal band). Both results are consistent with the effects of global climate warming on the butterfly, leading to a northward and upward shift in its geographical range. Reproduced, with the permission of McMillan Journals Ltd, from Nature 282 (1996), page 766.

3. Declines in Formerly Common Farmland Birds in Northwest Europe

In the intensively agricultural areas of northwest Europe—over the whole of lowland England, for example—a whole raft of formerly "common farmland birds" are also in steep decline (Tucker and Heath, 1994). They include skylarks (Alauda arvensis), European tree sparrows (Passer montanus), corn buntings (Miliaria calandra), gray partridges (Perdix perdix), and singing thrushes (Turdus philomelos). Here the problem is now reasonably well understood, though many details remain unresolved. Modern farming is so efficient and clean that there is little for the birds to eat. Weeds are killed with herbicides, which remove both seeds and rich sources of insects that feed on the weeds. The crop itself is sprayed to remove insects and is harvested so efficiently that few seeds are spilled on the way. Modern farms are biodiversity deserts, an indication of the power of people to squeeze nature to the margins while apparently maintaining a green and pleasant land. If present trends continue, skylarks will be rare birds in Britain in 20 years.

III. SPECIES AS INDICATORS OF BIODIVERSITY

A. The Nature of the Problem

Common sense suggests that the known losses of plants and birds from European farmland will go hand-in-hand with much more poorly documented declines in many other, less familiar and cryptic taxa, from land snails to glowworms, and hoverflies to harvest spiders. In other words, changes in the distribution and abun-
dance of well-known groups should serve as broad indicators of the status of, and changes in, a much wider sample of a region's flora and fauna. The assumption here is that birds (or other conspicuous species) might serve as biodiversity indicators—that is, as surrogates of overall biodiversity. But although it seems intuitively reasonable to use familiar, well-studied, and easily censused groups as indicators of what is happening to many other taxa, despite a great deal of research, the idea is actually contentious.

Following (but slightly modifying) the work of McGeoch (1998), we can define a biodiversity indicator as a group of taxa (e.g., species, genus, tribe, family, or order, or a selected group of species from a range of higher taxa) whose diversity (e.g., overall species richness, number of rare species, levels of endemism) reflects that of other higher taxa in a habitat, group of habitats, or geographic region. The idea is simple enough, and if it can be shown to work, it is important because biologists then have a relatively simple means of assessing overall biodiversity for purely scientific reasons, for setting conservation priorities, or for monitoring the effectiveness of conservation management.

B. Taxa That Have Been Suggested as Indicators of Biodiversity

The groups of organisms whose richness has been evaluated most thoroughly in the greatest number of places on earth are also the most familiar. The natural history sections of bookshops are dominated (sometimes exclusively) by volumes on plants and birds. If insects figure at all, butterflies will be on the top of the list, although there are fascinating differences between nations. Japan loves dragonflies. Birds, higher plants, butterflies, and dragonflies are all groups that occur in most places in the world but whose individual species are seldom so widespread. In much of the world they are also groups whose species are, relatively speaking, taxonomically well known and stable, readily identifiable, and have histories that are well understood. They are easy to find, inventory, and count, and they are reasonably, but not overwhelmingly, diverse in any one place. These are all desirable attributes of groups that might be used as indicators of the diversity of many other, much less well known taxa—that is, as indicators of the overall biodiversity of a region.

Other groups have many of the same attributes but have not gained the same popularity, perhaps because they are not also large bodied or perceived as being quite so attractive. The list of those that have been advocated as useful biodiversity indicators at one time or another is very long. It includes soil nematodes, moths, beetles galore (tiger, carabid, dung, and buprestid, to name but four), termites, fish, frogs, and snakes. Whatever the group, they must also have one further attribute—namely, that they genuinely indicate levels of biodiversity or at least some of the components of primary interest. The fact that the scientific literature contains suggestions for so many different possible indicators shows that there is little consensus on the matter. Many have been called, but few are chosen. Why? There are two, related, reasons. First, scientific knowledge on the degree of coincidence in patterns of biodiversity between different taxa is surprisingly poor. Second, as knowledge improves, coincidence between many taxa turns out to be much worse than people had imagined, or indeed hoped, would be the case.

C. Knowledge Is Poor Because of the Effort Required

Gathering information on the diversity of different groups of organisms, even in one place, is enormously time-consuming. Two examples illustrate the problem. To map the presence and absence of breeding birds (conspicuous, “easy” to find and to identify) in every 100 × 10 km grid-square in Britain and Ireland (there are 3672 squares) took more than half a million individual record cards, filled in by an army of amateur bird-watchers coordinated by professional ornithologists in the British Trust for Ornithology (BTO). The task took 4 years and about 100,000 hours of fieldwork (Gibbons et al., 1993). Now imagine the effort required to do the same thing for all the other hundreds of different groups of organisms found in this one small corner of Europe. It has been done for a sample of taxa (we will return to what these data show in a moment), but many groups remain unmapped.

At a much smaller spatial scale in a tropical forest in Cameroon, a group of biologists attempted to measure the impacts of forest disturbance on just eight groups (birds, butterflies, flying beetles, canopy beetles, canopy ants, leaf-litter ants, termites, and soil nematodes). The birds and the butterflies took 50 and 150 scientist-hours, respectively, to survey. But the effort required climbed rapidly for smaller-bodied, more cryptic, less well-known groups—1600 hours for the beetles, 2000 for the termites, and 6000 for the nematodes (Lawton et al., 1998). Despite the fact that this work as a whole took about five scientist-years, inventories for most groups that were surveyed were still only partial, and most taxa remained unexamined (fungi, higher plants, spiders, soil mites, collembola, earthworms, Liz-
ard, frogs, and mammals, to name some of the most conspicuous gaps.

Given this background, it is hardly surprising that biologists do not have a complete inventory of all the species that occur even in a single, moderately sized area (a field, small wood, or lake)—a so-called ATBI (All Taxa Biodiversity Inventory) (Oliver and Beattie, 1996). A moments thought will also show that to use one or two groups (for the sake of the argument, say birds and butterflies) as indicators of the richness of other taxa in fact requires several such areas to be investigated to properly test the hypothesis that high bird diversity (or any other single group) reflects a high diversity of many other groups.

Although progress has been made in this area over the past decade, considerable work remains to be done. Even in otherwise well-studied situations, many groups remain to be examined. Hence, at the present time, and effectively by default, some groups are being used as indicators of biodiversity, even though we cannot show categorically that the richness of one or more groups of organisms truly reflects the overall, or even a major portion of the overall, biodiversity of an area. As a result there is little consensus about what a “good” indicator group, or groups, might be, because there are too few hard data, from a range of habitats and geographic regions round the world, on which to draw firm conclusions. But as data slowly emerge, they are not encouraging for those who wish to use simple, single-taxon indicators of biodiversity.

D. Indicator Reliability

Where knowledge exists, it suggests that single or small numbers of taxa will usually be poor indicators of the biodiversity of other groups.

1. Tropical versus Temperate and Other Major Diversity Gradients

It would be wrong to think that there is no coincidence between patterns of diversity in different groups of organisms. Of course there is. In the broadest terms, it is axiomatic that most major terrestrial and freshwater groups are more species rich in the tropics than in temperate regions, at low elevations than at high ones, in forests than in deserts, and on large land masses than tiny islands. Whether you are a botanist, a bird-watcher, or a bug hunter, to find the most species it is generally advisable to head to hot and humid mainland tropics with lots of trees. It is easy to assume that there must therefore be reasonably good correlations between major diversity gradients for different groups. There can be, but even at this scale often there are not. Penguin diversity peaks in Antarctica, not the tropics, and there are many other examples of similar “reverse diversity gradients” that buck the average trend. On the eastern side of North America, the diversity of breeding warblers increases from south to north—suggesting that this conspicuous taxon, which is easy to identify (at least the breeding males!) and to survey, is probably highly unsuitable as an indicator of patterns of biodiversity in most other taxa (in which diversity typically decreases from south to north).

In the case of breeding North American warblers, we can spot the problem because we have enough information about the organisms involved. But the whole point about indicator taxa for biodiversity is that typically we will not be armed with, and indeed should not need, information about “other” groups; knowledge of the indicator taxon should suffice and be reliable. The evidence suggests otherwise.

2. Hot Spots

Major gradients in diversity aside, at similarly large scales an indicator group might be used to identify local geographic hot spots in the species richness of one or more other groups (peaks in the landscape of species richness) or to determine relative levels of richness in those other groups (hot spots versus all spots) (Gaston 1996b; Reid 1998). At the continental scale, the procedure has frequently been found to fail on both counts (Gaston 1996a), with mismatches between the occurrence of peaks in the richness of different groups being commonplace.

Across the United States and southern Canada, hot spots (local areas with unusually high diversity) overlap partially between some pairs of taxa (trees, tiger beetles, amphibians, reptiles, birds, and mammals), but the pattern is not a general one. Numbers of species in different large grid cells for two groups are often significantly positively correlated, for example, birds and tiger beetles or mammals and swallowtail butterflies. But these correlations are frequently weak, of rather limited predictive value, and in some cases explained by latitudinal gradients in diversity. In other words, although such correlations may sometimes enable a very general impression of the patterns in richness of one group to be obtained from the patterns in richness of another, their predictive powers are low.

These conclusions seem to hold at finer resolutions over more constrained areas. Thus, species-rich areas for different taxa in Britain (birds with butterflies, dragonflies, etc.) frequently do not coincide at a scale of 10 × 10 km squares (Pendergast et al., 1993) (Fig. 5).
FIGURE 5 Coincidence between hot spots for butterflies and up to seven other taxa in Britain. Hot spots are unusually species-rich sites (here defined as the top 5 percentile in Britain). All of the most species-rich localities in Britain for butterflies lie in southern England. Increasingly dark shading indicates that butterfly hot spots coincide with hot spots for an increasing number of other taxa. Note that many butterfly hot spots are not unusually rich in any other species (open circles) and that only one locality (in southeast England, just in from the coast) is a hot spot for all eight taxa in this particular survey. The other taxa are breeding birds, dragonflies, moths, mollusks, aquatic higher plants, and liverworts (simple plants). Information from Prendergast et al. (1993), with additional data and figure kindly provided by John Prendergast.

Hot spots in this study are not distributed randomly, overlapping more often than expected by chance, but still at a low level. Likewise, different taxa are species poor or species rich in different areas of the Transvaal region of South Africa. At even finer scales, within the Cameroon forest mentioned earlier, disturbance impacted on the diversity of eight taxa in very different ways. All declined drastically in completely cleared areas, but intermediate levels of forest disturbance had very different effects on the diversity of different groups. As a result, changes in the diversity of one taxon could not be used to predict changes in the diversity of any other (Lawton et al., 1998). A summary of these and related studies showing similar results is provided by Gaston (1996a, 1996b) and by Pimm and Lawton (1998).

3. A Commonsense Explanation
This lack of, or relatively feeble correlation between, species rich-areas for different groups of organisms makes the search for simple, robust, single-taxon indicators of overall biodiversity look increasingly like a lost cause. With hindsight, perhaps this emerging result is obvious (Reid, 1998). Major geographic gradients in biodiversity aside, within particular geographic regions or at smaller habitat scales, the conditions favoring one group of species may be hostile to another. Mollusks like it cool and wet, butterflies like it warm and sunny, and high bird diversity is more likely in tall vegetation than short vegetation, irrespective of weather. Commonsense natural history suggests that there is unlikely to be a single indicator taxon able to predict the diversity of all, or even a majority of others.

4. Rare Species and Endemic Species
Biologists and conservationists are often interested not only in patterns of species richness but also in the distribution of unusually rare species, or of endemic species. Do sites with unusual numbers of rare species frequently coincide across different taxa? Again, the answer seems to be no, or only weakly (Pimm and Lawton, 1998; Prendergast et al., 1993), for the reasons just outlined. Endemic species may be different (Bibby et al.,
There is some evidence that areas rich in endemic birds (e.g., some tropical mountaintops or isolated islands) may also contain unusually large numbers of endemic species in other groups. However, rigorous data and analyses are few, and exceptions are easy to find. Lake Baikal has no endemic birds but supports an exceptionally rich, endemic invertebrate fauna and a unique, endemic freshwater seal.

E. Selecting Areas to Conserve Biodiversity: Conservation Planning and Reserve Selection Algorithms

1. Biodiversity Indicators and Conservation

The general lack of reliable indicator groups for biodiversity is undoubtedly unfortunate for scientists wishing to understand how life is distributed across the earth; the road to an atlas of biodiversity seems set to be a long one. In practice, it may actually prove somewhat less of a worry for one of the primary motivations in the search for indicators for biodiversity—namely, conservation planning. Networks of national parks and reserves are central planks in conservation, albeit alone they are insufficient to protect all species. Their establishment is one of the obligations placed on Parties to the Convention on Biological Diversity. A primary argument for using indicators of biodiversity is to determine the effectiveness of these protected area networks in capturing biodiversity, and the best ways in which they might be extended, in the face of stiff competition with alternative forms of land use.

Although hot spots of species richness coincide weakly for different taxa (see the previous section), if we turn the problem around and look at it from a different angle, an interesting picture emerges. Imagine that conservation priorities in Britain have been set by concentrating just on areas rich in birds (the bird hot spots). (Although the general view of many nations is that the British are mad about birds, the country’s protected area network is not based solely on birds. We use the example simply to illustrate a point.) What we discover is that the hot spots for this one group tend to embrace a high proportion of the total species in other groups. Thus, the hot spots for breeding birds contain 87% of the breeding bird species in Britain, 100% of the butterflies, 92% of the dragonflies, 92% of the liverworts, and 94% of the aquatic plants. A reserve network established around hot spots for one group does a rather good job of ensuring that most species in other groups find a place in the extended ark of protected sites.

2. Reserve Selection Algorithms

Although this message is encouraging, it turns out that designing conservation networks simply on the basis of levels of species richness is extremely inefficient, at least if the goal is to capture representative samples of all taxa. The same species may occur repeatedly in different richness hot spots for a group. Conservationists do not have unlimited resources, and this duplication wastes money on purchasing, or managing, unnecessary land. On the other hand, some species, particularly the rare ones of primary conservation interest, may not occur in any richness hot spots at all. What is required is to identify those areas that constitute the greatest complementary species richness; the complementary part of an area's biota consists of those species unrepresented in another biota with which it is being compared. To do this, mathematically minded conservation biologists have developed powerful reserve selection algorithms that help to select sites with maximum efficiency, according to some predetermined criteria (Pressey et al., 1993). The criteria may be to maximize the number of species, rare species, or endemic species in a proposed reserve network, at minimum cost, on a minimum area, closest to existing reserves, or what have you.

Echoing the conclusions of the previous section, the question then arises as to whether the patterns of complementarity of one group of organisms are congruent with those of another. The question is a new one, with few studies available to answer it. Across 50 forests of Uganda, which boasts more species for its size than almost any other country in Africa, and consistent with our earlier conclusions, there was little spatial congruence in the species richness of woody plants, large moths, butterflies, birds, and small mammals once differences in sampling effort were accounted for. However, sets of forests selected using complementarity determined for single taxa were generally similar to those for all other taxa and hence served to capture well the species richness in all these other groups (Howard et al., 1998).

If these results generalize to other parts of the world, they send an encouraging message to conservation managers struggling to identify the best areas to set aside as reserves and parks. It says that a complementary and therefore efficiently selected chain of reserves based on a single indicator taxon (or perhaps two or three indicator taxa) may efficiently capture complementary
sets of many other groups as well. Unfortunately, the Ugandan results are not supported by similar studies in the Transvaal, elsewhere in Africa (van Jaarsveld et al., 1998). It may therefore be too soon to assume that we can find simple indicators for complementary reserve sets embracing many taxa as a means of conserving biodiversity.

IV. CONCLUSIONS

The term “indicator species” has three distinct meanings. They are a species, or group of species, that reflect the biotic or abiotic state of an environment; reveal evidence for, or the impacts of, environmental change; or indicate the diversity of other species, taxa, or entire communities within an area. The uses of indicator species in the first two senses of the word are very similar, differing largely in the fact that to indicate change, organisms need to be sampled more than once in the same place and in the same way. Using organisms to indicate the state of, and changes in, the environment has numerous tried and tested applications, from detecting pollution to monitoring recovery of formerly degraded habitats, at many scales, from local to global. The use of indicator species to predict the diversity of other, unstudied taxa for scientific or conservation reasons is much more contentious and may prove to be impossible with any degree of rigor.

See Also the Following Articles

ACID RAIN AND DEPOSITIONS • BERNS, BIODIVERSITY OF • ECOTOXICOLOGY • ENDEMISM • ENVIRONMENTAL IMPACT, CONCEPT AND MEASUREMENT OF • GREENHOUSE EFFECT • HOTSPOTS • KEYSTONE SPECIES • PALEOECOLOGY

Bibliography


INDIGENOUS PEOPLES, BIODIVERSITY AND

Victor M. Toledo
Institute of Ecology, National University of Mexico (UNAM)

I. Introduction
II. Indigenous Peoples
III. Biological Diversity and Diversity of Cultures
IV. Biodiversity and Biomass Appropriation: The Role of Indigenous Peoples
V. Biodiversity and Indigenous People’s Lands and Waters
VI. Biodiversity and Ethnoecology: Indigenous Views, Knowledge, and Practices
VII. Conserving Biodiversity by Empowering Indigenous Peoples
VIII. Concluding Remarks: A Biocultural Axiom

GLOSSARY

bio-cultural axiom Recognition that biological and cultural diversity are mutually dependent and geographically coterminous.
cultural diversity Variety of human groups distinguished through beliefs, lifeways, dress, food, languages, sexual behavior, forms of productive organization, art, and conceptions of nature.
endemic languages Languages that are restricted to a single country and, like their species counterparts, hold a high percentage of the unique traits in human language.
ethnoecology Interdisciplinary study that explores how nature is perceived by human groups through a screen of beliefs and knowledge and how humans, in terms of their images and symbols, use and/or manage natural resources.
indigenous peoples Those who are the “original” or oldest inhabitants of an area or region, or who have lived in a traditional homeland for many generations, usually many centuries.

INDIGENOUS PEOPLES NUMBER BETWEEN 300 AND 700 MILLION. They are inhabitants of practically every major biome of the earth and especially of the least disturbed terrestrial and aquatic ecosystems. This article is based on an exhaustive review of recently published data, and stresses the strategic importance of indigenous peoples in the maintenance and conservation of the world’s biodiversity. Four main links between biodiversity and indigenous peoples are examined: the correlation between biological richness and cultural diversity on both geopolitical and biogeographic terms; the strategic importance of indigenous peoples in the biomass appropriation; the remarkable overlap between indigenous territories and the world’s remaining areas of high biodiversity; and the importance of indigenous views, knowledge, and practices in biodiversity conservation. The article concludes by emphasizing the urgent need for recognizing a new biocultural axiom: that global biodiversity can only be effectively preserved by preserving the diversity of human cultures, and vice versa.
INDIGENOUS PEOPLES, BIODIVERSITY AND

I. INTRODUCTION

Biodiversity as a word and concept originated in the field of conservation biology. However, as Alcorn (1994:11) states, "while proof of conservation success is ultimately biological, conservation itself is a social and political process, not a biological process. An assessment of conservation requires therefore an assessment of social and political institutions that contribute to, or threaten, conservation." One of the main social aspects related to biodiversity is, undoubtedly, the world's indigenous peoples.

Scientific evidence shows that virtually every part of the planet has been inhabited, modified, and manipulated throughout human history. Although they appear to be untouched, many of the last tracts of wilderness are inhabited and have been so for millennia. Indigenous peoples live in and have special claims to territories that, in many cases, harbor exceptionally high levels of biodiversity. On a global basis, human cultural diversity is associated with the remaining concentrations of biodiversity. Both cultural diversity and biological diversity are endangered.

Given the foregoing, this article offers a review of the multiple importance of indigenous peoples and makes the point that valuable, local-specific views, knowledge, and practices are used by indigenous peoples who have relied for centuries on the maintenance of biodiversity. Indigenous peoples may have all or some of the following criteria: (a) are the descendants of the original inhabitants of a territory that has been overcome by conquest; (b) are "ecosystem peoples," such as shifting or permanent cultivators, herders, hunters and gatherers, fishers, and/or handicraft makers, who adopt a multisite strategy of appropriation of nature; (c) practice a small-scale, labor-intensive form of rural production that produces little surplus and has low energy needs; (d) do not have centralized political institutions, organize their life at the level of the community, and make decisions on a consensus basis; (e) share a common language, religion, moral values, beliefs, clothing style, and other identifying characteristics, as well as a relationship to a particular territory; (f) have a different worldview, consisting of a custodial and nonmaterialist attitude to land and natural resources based on a symbolic interchange with the natural universe; (g) are subjugated by a dominant culture and society; and (h) consist of individuals who subjectively consider themselves to be indigenous.

It is possible to find indigenous peoples carrying out many different activities of use and management of the planet's ecosystems: as forest-dwellers in the tropical lowlands or in the mountains, as pastoralists in savannas and other grasslands, or as nomadic or seminomadic hunters and gatherers in forests, prairies, and deserts. In addition, fishing is the principal economic activity and source of food for several million coastal and island dwellers, as well as for many indigenous peoples inhabiting margins of rivers.

Large numbers of indigenous peoples are, however, peasant producers and therefore can be indistinguishable from the nonindigenous peoples living nearby. In the Andean and Mesoamerican countries of Latin America, for instance, indigenous peoples farm like mestizo peasants. Similarly, in India distinctions between scheduled (government recognized) tribes and

II. INDIGENOUS PEOPLES

Indigenous peoples number between a minimum of 300 million (Table I) and a maximum of 700 million (Harmon, 1999; the reason for this wide range is discussed in the following). They live in about 75 of the world's 194 countries and are inhabitants of practically each major biome of the earth. Indigenous peoples—also called tribal, aboriginal, or autochthonous peoples, national minorities, or first peoples—are best defined by using several criteria. Indigenous peoples may have all or some of the following criteria: (a) are the descendants of the original inhabitants of a territory that has been overcome by conquest; (b) are "ecosystem peoples," such as shifting or permanent cultivators, herders, hunters and gatherers, fishers, and/or handicraft makers, who adopt a multisite strategy of appropriation of nature; (c) practice a small-scale, labor-intensive form of rural production that produces little surplus and has low energy needs; (d) do not have centralized political institutions, organize their life at the level of the community, and make decisions on a consensus basis; (e) share a common language, religion, moral values, beliefs, clothing style, and other identifying characteristics, as well as a relationship to a particular territory; (f) have a different worldview, consisting of a custodial and nonmaterialist attitude to land and natural resources based on a symbolic interchange with the natural universe; (g) are subjugated by a dominant culture and society; and (h) consist of individuals who subjectively consider themselves to be indigenous.

It is possible to find indigenous peoples carrying out many different activities of use and management of the planet's ecosystems: as forest-dwellers in the tropical lowlands or in the mountains, as pastoralists in savannas and other grasslands, or as nomadic or seminomadic hunters and gatherers in forests, prairies, and deserts. In addition, fishing is the principal economic activity and source of food for several million coastal and island dwellers, as well as for many indigenous peoples inhabiting margins of rivers.

Large numbers of indigenous peoples are, however, peasant producers and therefore can be indistinguishable from the nonindigenous peoples living nearby. In the Andean and Mesoamerican countries of Latin America, for instance, indigenous peoples farm like mestizo peasants. Similarly, in India distinctions between scheduled (government recognized) tribes and
nontribal peoples cannot be made solely on the basis of productive activities. In these and other many cases, nonindigenous peasants and indigenous peoples produce the same crops with the same farming methods. Because in numerous countries many mestizo peasants are direct descendants of the indigenous peoples and retain most of their cultural traits, it has been pointed out that a broader definition of indigenous peoples might increase the real numbers. Thus, by considering characteristics other than language, it is possible to enlarge the number of people classified as indigenous in the contemporary world. Some authors, such as J. Burger (1987), think that the number of indigenous people may actually be double that previously estimated. Thus, in the contemporary world there may be as many as 600 million indigenous peoples. Burger's estimation echoes the figure of over 700 million given by Harmon (1995) as the total number of people speaking some of the world's 5635 "endemic languages," which are languages restricted to only one nation. Endemicity in language can be linked to small-scale societies and, therefore, to indigenous peoples.

Based on the percentage of the total population identified as being indigenous, it is possible to recognize a group of selected nations with a strong presence of these peoples: Papua New Guinea (77%), Bolivia (70%), Guatemala (47%), Peru (40%), Ecuador (38%), Myanmar (33%), Laos (30%), Mexico (12%), and New Zealand (12%). On the other hand, the absolute number of people recognized as indigenous allows one to identify nations with high indigenous populations, such as India and China.

III. BIOLOGICAL DIVERSITY AND DIVERSITY OF CULTURES

On a global basis, human cultural diversity is associated with the remaining concentrations of biodiversity. In fact, evidence exists of remarkable overlaps between global mappings of the world's areas of high biological richness and areas of high diversity of languages, which is the single best indicator of a distinct culture. According to Harmon (1996), "Species and languages are not just comparable on an abstract, conceptual level. There is also a striking pattern of congruity in the geographical distribution of the two. For instance, many countries with high numbers of endemic species also have many endemic languages." Measured by spoken language, all the world's people belong to between 5000 and 7000 cultures. It is estimated that 4000 to 5000 of these are indigenous cultures. Thus, indigenous peoples account for as much as 80 to 90% of the world's cultural diversity. On the basis of the inventories done by linguists, we can draw up a list of the regions and countries with the greatest degree of cultural diversity in the world. According to Ethnologue, the best existing catalog of the world's languages, there is a total of 6703 languages (mostly oral), 32% of which are found in Asia, 19% in the Pacific, 13% in the Americas, and 3% in Europe (Grimes, 1996). Only twelve countries account for 54% of all human languages. These countries are Papua New Guinea, Indonesia, Nigeria, India, Australia, Mexico, Cameroon, Brazil, Zaire, Philippines, United States, and Vanuatu (Table II).

On the other hand, according to the most recent and detailed analysis of biodiversity on a country-hy-
country basis (Mittermeier and Goettsch-Mittermeier, 1997), there are, similarly, 12 countries that house the highest numbers of species and endemic species (Table III). This assessment was based on the comparative analysis of eight main biological groups: mammals, birds, reptiles, amphibians, freshwater fishes, butterflies, tiger beetles, and flowering plants. The nations considered to be “megadiversity” countries are: Brazil, Indonesia, Colombia, Australia, Mexico, Madagascar, Peru, China, Philippines, India, Ecuador, and Venezuela.

Thus, the relationship between cultural diversity and biological diversity stands out in global statistics: 9 of the 12 main centers of cultural diversity (in terms of number of languages) are also in the roster of biological megadiversity nations and, reciprocally, 9 of the countries with the highest species richness and endemism are also in the list of the 25 nations with the highest number of endemic languages (Harmon, 1996, see Table II and Table III).

The links between biological and cultural diversity can also be illustrated by using the data of Global 200, a program of the World Wide Fund for Nature (WWF) that was developed as a new strategy to identify conservation priorities based on an ecoregional approach. As part of this program, WWF identified a list of 233 terrestrial, freshwater, and marine biological ecoregions representative of Earth’s richest diversity of species and habitats. A preliminary analysis conducted by the People and Conservation Unit of WWF regarding the presence of indigenous peoples in the 136 terrestrial ecoregions of Global 200 revealed interesting patterns. As shown in Table IV, nearly 80% of the terrestrial ecoregions are inhabited by one or more indigenous peoples, and half of the world’s 3000 indigenous groups, as

### Table III: Top 12 Countries by Number of Species (Richness) and Endemics (Endemism)

<table>
<thead>
<tr>
<th>Country</th>
<th>Richness</th>
<th>Endemism</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>*Indonesia</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>*Colombia</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>*Australia</td>
<td>7</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>*Mexico</td>
<td>3</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Madagascar</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>*Peru</td>
<td>4</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>*China</td>
<td>6</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>*Philippines</td>
<td>14</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>*India</td>
<td>9</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>Ecuador</td>
<td>8</td>
<td>14</td>
<td>22</td>
</tr>
<tr>
<td>Venezuela</td>
<td>10</td>
<td>15</td>
<td>25</td>
</tr>
</tbody>
</table>

*Calculated for the following biological groups: mammals, birds, reptiles, amphibians, freshwater fishes, butterflies, tiger beetles, and flowering plants. (Source: Mittermeier and Goettsch-Mittermeier (1997). Asterisks indicate countries included in the list of the 25 nations with the highest number of endemic languages (See Table II, Harmon, 1996).
estimated by WWF, are inhabitants of these ecoregions. All of the regions, except the Palearctic region, have indigenous peoples living in 78% or more of their identified ecoregions.

IV. BIODIVERSITY AND BIOMASS APPROPRIATION: THE ROLE OF INDIGENOUS PEOPLES

Biodiversity conservation cannot be separated from natural resources utilization. The human appropriation of natural materials includes minerals, water, shelter and fiber materials, solar energy, and, principally, living organisms (biomass) from ecosystems. World statistics indicate that almost half the humans on the planet are still engaged in the direct appropriation of natural resources. This appropriation is carried out by a myriad of rural or primary producers through the management of terrestrial, marine, and freshwater ecosystems.

Forty-five percent of the total human population has been recorded by the United Nations Food and Agriculture Organization (FAO) as agricultural population (FAO, 1991). It can be estimated that between 60 and 80% of this agricultural population is represented by small-scale, solar-energized productive units based on a multiuse management of nature (Toledo, 1990). In fact, the statistical record shows that by 1990 around 1.2 billion rural people were practicing agricultural activities on areas of 5 hectares or less. This figure coincides with the last available world census of agriculture by the FAO in 1970, when more than 80% of all reported holdings were smaller than 5 hectares (ha). A similar pattern is found in the world's fisheries, where more than 90% are small-scale, artisanal operators acting in a great variety of coastal habitats.

Most of these small-scale farmers and fishers develop their production activities not as socially isolated households but as familial nuclei belonging to specific village communities, many of which, in turn, correspond to cultures that can be considered as indigenous. Moreover, within the core of these community-based producers, those identified as indigenous people also carry out the biomass extraction at the lowest level in their local ecosystems. Called “ecosystem people” by some authors, such as R. F. Dasmann and M. Gadgil, these producers subsist by appropriating a diversity of biological resources from their immediate vicinity. Their quality of life is therefore intimately linked to the maintenance of certain levels of local biodiversity (Gadgil, 1993, see the following). As a consequence, they are productive actors in little transformed habitats of the world, and include forest and sea dwellers, slash-and-burn agriculturists, some 25–30 million nomadic herders or pastoralists (in East Africa, the Sahel, and the Arabian peninsula), most of the world’s 15–21 million fishers, and all of the half a million hunters and gatherers still recognized as citizens of the contemporary world.

In conclusion, indigenous peoples represent the fraction of human appropriators of biomass that cause the lowest ecological impacts. They generally live in what may be termed “frontier lands” or “refuge regions,” in other words, remote areas of great “wilderness” where the structure, not the components, of original ecosystems remains more or less untouched. In many cases, these lands and waters are untamed, unknown, owned, and as yet unclaimed.

V. BIODIVERSITY AND INDIGENOUS PEOPLE’S LANDS AND WATERS

Indigenous peoples occupy a substantial share of the world’s little disturbed tropical and boreal forests, mountains, grasslands, tundra, and desert, along with large stretches of coastline and nearshore waters (including mangroves and coral reefs) (Durning, 1993). The importance of indigenous territories to biodiversity conservation is therefore evident.

In fact, indigenous peoples control, legally or not, immense areas of natural resources. Among the most remarkable examples are the Inuit people (formerly known as Eskimo), who govern a region covering one-fifth of the territory of Canada (222 million ha), the indigenous communities of Papua New Guinea, whose lands represent 97% of the national territory, and the tribes of Australia with nearly 90 million ha (Fig. 1). Although numbering only above 250,000, the Amerindians of Brazil possess an area of over 103 million ha, mainly in the Amazon Basin, distributed in 565 territories (Fig. 2 and Table V). Nearly 60% of the priority areas in central and southern Mexico recommended for protection are also inhabited by indigenous peoples (Fig. 3), and half of the 30,000 rural communities are located in the 10 most biologically rich states of the Mexican territory. In summary, on a global scale it is estimated that the total area under indigenous control probably reaches between 12 and 20% of the earth’s land surface (Stevens, 1997). The best example of notable overlaps between indigenous peoples and biologically rich areas is the case of tropical humid forests. In fact, there is a clear correspondence between areas of remaining tropical forests and...
the presence of indigenous peoples in Latin America, the Congo Basin in Africa, and several countries of tropical Asia, such as the Philippines, Indonesia, and Papua New Guinea. The strong presence of indigenous peoples in Brazil, Indonesia, and Zaire alone is remarkable, as these countries account for 60% of all the tropical forest in the world.

In Latin America, this geographical relationship has been strikingly verified for the Central American countries by a National Geographic Society map produced by a project headed by Mac Chapin in 1992. The same pattern can be found in the tropical humid areas of Mexico inhabited by 1.6 million indigenous people, and for many regions of the Amazonia Basin (see the case of Brazil in Fig. 2). It has been estimated that in Amazonia more than 1 million indigenous people of eight countries possess over 135 million ha of tropical forests (Davis and Wali, 1994).

Many temperate forests of the world also overlap with indigenous territories, for example, in India (Fig. 4), Myanmar, Nepal, Guatemala, the Andean countries (Ecuador, Peru, and Bolivia), and Canada. Furthermore, over 2 million islanders of the South Pacific, most of whom are indigenous peoples, continue fishing and harvesting marine resources in high-biodiversity areas (such as coral reefs).

VI. BIODIVERSITY AND ETHNOECOLOGY: INDIGENOUS VIEWS, KNOWLEDGE, AND PRACTICES

Biodiversity is a very broad concept that refers to the variety of landscapes, ecosystems, species, and genes, including the associated functional processes. Therefore, the maintenance and conservation of biodiversity demand efforts on these four levels. The first level is oriented to preservation of assemblies of “ecosystems,” whereas the second level focuses on protection of habitats in which the populations of species live. At the species level, most biodiversity knowledge is of large
plants and animals such as flowering plants and vertebrates. Much of the extent of diversity of smaller plants and animals remains to be inventoried and protected. Although most biological diversity is constituted by wild plants and animals, an important subset involves the diversity among domesticated organisms. In this fourth level, interest focuses on the conservation of genetic variation in crops and domesticated animals.

This section examines the potential role of indigenous peoples in biodiversity conservation from an ethnological perspective. Ethnology can be defined as the interdisciplinary study of how nature is perceived by human groups through a screen of beliefs and knowledge, and how humans, through their symbols, use and/or manage natural resources. Thus, by focusing on the cosmos (the belief system or cosmovision), the corpus (the whole repertory of knowledge or cognitive systems), and the praxis (the set of practices), ethnology offers an integrative approach to the study of the process of human appropriation of nature (Toledo, 1992). This approach allows one to recognize the value of the belief–knowledge–practice complex of indigenous peoples in relation to the conservation of biodiversity.

A. The Cosmos

For indigenous peoples, land and in general nature have a sacred quality that is almost absent from Western thinking. Land is revered and respected and its inalienability is reflected in virtually every indigenous cosmovision. Indigenous people do not consider the land as merely an economic resource. Under indigenous cosmologies, nature is the primary source of life that nourishes, supports, and teaches. Nature is, therefore, not only a productive source but also the center of the universe, the core of culture, and the origin of ethnic identity. At the heart of this deep bond is the perception that all living and nonliving things and natural and social worlds are intrinsically linked (the reciprocity principle). Of particular interest is the research done by several authors (G. Reichel-Dolmatoff, E. Boege, P. Descola, C. van der Hammen, and K. Arhem) on the role played by the cosmology of several indigenous groups as a mechanism regulating the use and management of natural resources. In the indigenous cosmology, each act of appropriation of nature must be negotiated with all existing things (living and nonliving) through different mechanisms such as agrarian rituals and shamanic acts (symbolic exchange). Humans are thus seen as a particular form of life participating in a wider community of living beings regulated by a single and totalizing set of rules of conduct.

B. The Corpus

Indigenous societies house a repertory of ecological knowledge that generally is local, collective, diachronic,
and holistic. In fact, because indigenous peoples possess a very long history of resource-use practice, they have generated cognitive systems on their own circumscribed natural resources, which are transmitted from generation to generation. The transmission of this knowledge is done through language, hence the corpus is generally an unwritten knowledge. Therefore, memory is the most important intellectual resource among indigenous cultures.

This body of knowledge is the expression of a certain personal wisdom and, at the same time, of a collective creation, that is to say, a historical and cultural synthesis turned into reality in the mind of an individual producer. For this reason, the corpus contained in a single producer’s mind expresses a repertoire that is a synthesis of information from at least four sources: (a) the experience accumulated over historical time and transmitted from generation to generation by a certain cultural group; (b) the experiences socially shared by the members of a generation or cohort; (c) the experience shared in the household or the domestic group to which the individual belongs; and (d) the personal experience, particular to each individual, achieved through the repetition of the annual cycles (natural and productive), enriched by the perceived variations and unpredictable conditions associated with them.
Thus, indigenous ecological knowledge is normally limited to the immediate environments and is an intellectual construction resulting from a process of accumulation of experiences over both historical time and social space. These three main features of indigenous ecological knowledge—being local, diachronic, and collective—are complemented by a fourth characteristic, namely, the holistic.

Indigenous knowledge is holistic because it is intricately linked to the practical needs of use and management of local ecosystems. Although indigenous knowledge is based on observations on a rather restricted geographic scale, it must provide detailed information on the variety of scales represented by the concrete landscapes in which natural resources are used and managed. As a consequence, indigenous minds not only possess detailed information about species of plants, animals, fungi, and some microorganisms, but they also recognize many types of minerals, soils, waters, snows, landforms, vegetation, and landscapes.

Similarly, indigenous knowledge is not restricted to the structural aspects of nature, which are related to the recognition and classification (ethnotaxonomies) of elements or components of nature, but also encompasses dynamic (which refers to patterns and processes), relational (linked to relationships between or among natural elements or events), and utilitarian dimensions of natural resources. As a result, it is possible to integrate a cognitive matrix (Fig. 5) that certifies the holistic character of indigenous knowledge and serves as a methodological framework for ethnoecological research (Toledo, 1992).

C. The Praxis

Indigenous societies generally subsist by appropriating a diversity of biological resources from their immediate vicinity. Thus, subsistence of indigenous peoples is based more on ecological exchanges (with nature) than on economic exchanges (with markets). They are therefore forced to adopt survival mechanisms that guarantee an uninterrupted flow of goods, materials, and energy.
from ecosystems. In this context, a predominant use-value economic rationality is adopted, which in practical terms is represented by a multiuse strategy that maximizes the variety of goods produced in order to provide basic household requirements throughout the year (see Toledo, 1990). This main feature accounts for the relatively high self-sufficiency of indigenous households and communities.

Indigenous households tend to carry out a nonspecialized production based on the principle of diversity of resources and practices. This mode of subsistence results in the maximum utilization of all the available landscapes of the surrounding environments, the recycling of materials, energy, and wastes, the diversification of the products obtained from ecosystems, and, especially, the integration of different practices: agriculture, gathering, forest extraction, agroforestry, fishing, hunting, small-scale cattle-raising, and handicrafts. As a result, indigenous subsistence implies the generation of a myriad of products, including food, domestic and work instruments, housing materials, medicines, fuel-woods, fibers, animal forage, and others.

Under the multiuse strategy, indigenous producers manipulate the natural landscape in such a way that two main characteristics are maintained and favored: habitat patchiness and heterogeneity, and biological as well as genetic variation. In the spatial dimension, the landscapes in which indigenous peoples live become a complex landscape mosaic in which agricultural fields, fallow areas, primary and secondary vegetation, household gardens, cattle-raising areas, and water bodies are all segments of the entire production system. This mosaic represents the field upon which indigenous producers, as multiuse strategists, play the game of subsistence through the manipulation of ecological components and processes (including forest succession, life cycles, and movement of materials).

It has been demonstrated that some natural disturbances can increase biodiversity if they increase habitat heterogeneity, reduce the influence of competitively dominant species, or create opportunities for new species to invade the area. On the other hand, the number of species is commonly relatively small in highly disturbed biotic communities, because few populations are able to reestablish themselves before they are reduced by later disturbances. In contrast, a low rate of disturbance provides few opportunities for pioneer species and might allow competitively dominant species to usurp limiting resources. Therefore, biodiversity is often greater at intermediate levels of disturbances than at either lower or higher rates.
The creation of landscape mosaics under the indigenous multiscate strategy in areas originally covered by only one natural community represents a human-originated mechanism that theoretically tends to maintain (and even increase) biodiversity. Several authors have already stressed the importance of the models of low-intensity mosaic usage of the landscape by indigenous peoples and other small-landowner populations for biodiversity conservation. The same diversified arrangement found in indigenous landscapes tends to be reproduced at the local scale, with multispecies, multistory crops or agroforests favored over monocultures. As a consequence, animal and especially plant genetic resources tend to be maintained in indigenous agricultural fields, aquaculture systems, home-gardens, and agroforests (Gadgil et al., 1993). Polycultural systems managed by indigenous agriculturalists and agroforesters are relatively well known and the recent specialized literature is full of case studies illustrating such designs. Especially notable are the home-gardens and agroforestry systems of the tropical and humid regions of the world, which operate as human-made refuge areas for many species of plants and animals, notably in areas strongly affected by deforestation (Moguel and Toledo, 1999).

At the farm level, it is broadly recognized that crop populations are more diverse in indigenous farming systems than in agricultural areas dominated by industrial agriculture. Therefore, indigenous peoples are regarded as key agents of on-farm preservation of plant genetic resources that are threatened by agricultural modernization (or genetic erosion). The loss of biodiversity is also experienced in farming systems when indigenous cropping polycultural practices are replaced by fossil-fueled monocrops. Indigenous agricultural systems and landscapes are widely acknowledged as designs that preserve not only landraces of crop species, but also semidomesticated and wild crop relatives and even nondomesticated species.

VII. CONSERVING BIODIVERSITY BY EMPOWERING INDIGENOUS PEOPLES

During the past three decades, as the loss of landscapes, habitats, species, and genetic diversity has become an issue of international concern, the protected areas of the world have increased notably in both size and number. However, as protected areas expanded, it became evident that the model of uninhabited national parks originated by the so-called developed nations could not be applied worldwide. Today, there are nearly 10,000 nationally protected areas (parks and other reserves) in more than 160 countries, covering some 650 million ha, which represents over 5% of the earth's land surface. Many of the areas that have been established as protected areas and many of those that are suitable for future addition to the protected area network are the homelands of indigenous peoples. In Latin America alone, over 80% of protected areas are estimated to have indigenous people living within them. On the other hand, large tracts of the territories under indigenous control, estimated to encompass between 12 and 20% of the earth's surface, are considered to be priority candidates as future reserves. Moreover, some authors, such as J. Alcorn (1994), believe that the bulk of the world's biodiversity is now held within the limits of the indigenous territories in tropical countries.

Given this situation, as well as the evidence offered and discussed in the previous sections, the idea that successful and long-term biodiversity conservation will be impossible without the participation of indigenous communities is gaining recognition in national and international conservation circles. For example, in its latest guidelines, the IUCN (World Conservation Union) Commission on National Parks and Protected Areas (1994) considers that indigenously established "protected territories" can now be recognized as national parks, wilderness areas, protected landscapes, and managed resource protected areas. Furthermore, the international conservation community is beginning to realize that sacred forests, mountains, lakes, rivers, and deserts can also be considered protected areas, as well as managed reefs, lagoons, rivers, and grasslands.

Protected areas based on consultation, comanagement, and even direct management by indigenous peoples are likely to be increasingly important in coming years as the key role of indigenous cultures is gradually recognized and accepted. However, it is important not to idealize indigenous peoples and their resource management strategies and stewardship skills. Some conservationists have been criticized for over-romanticizing indigenous peoples, and in so doing creating a late-twentieth-century version of "the noble savage" (Redford, 1991). Acknowledgment of the positive links between indigenous peoples and biodiversity has been increasingly tempered by the recognition that under certain circumstances (high population densities, market pressures, unsuitable technologies, local disorganization) indigenous peoples can act as disruptive, not as conservationist, actors. In fact, many of the conservation strategies normally applied by indigenous peoples at the local level can be profoundly affected by exoge-
nous phenomena such as economic exploitation, cultural domination, and technological change.

Today, biological diversity and sustainable development are two of the most powerful and central concepts in environmental protection. In recent years, special attention has been paid to supporting the sustainable development of community-based peoples as a key mechanism for the reinforcement of responsible participation of local communities in biodiversity conservation. Sustainable community development can be defined as an endogenous mechanism that allows a local society to take (or retake) control of the processes that affect it. In other words, self-determination and local empowerment, conceived as a “talking of control,” have to be the central objectives in all community development.

Given the demonstrated importance of indigenous peoples for biodiversity conservation, it is essential to recognize the necessity of empowering local communities. This will require that indigenous communities be allowed to maintain, reinforce, or assume control of their own territories and natural resources, as well as have sufficient access to relevant information and technology that will assist their resource management. It is important that they hold legally recognized and enforceable rights to lands and waters, which will give the communities both an economic incentive and a legal basis for stewardship. In many countries, national recognition and policy support for existing, community-based property rights systems are crucial. In many Asian and African countries, the return of a measure of control over public lands and resources to local communities is fundamental to slowing biodiversity loss in threatened regions.

Similarly, it is crucial to establish new resource-management partnerships among local communities and provincial and national agencies. Local stewardship, in conjunction with external governmental and nongovernmental institutions and organizations, is perhaps the best way to guarantee the effective protection of landscapes, habitats, species, and genes worldwide, and especially in the biodiversity-rich tropical countries.

VIII. CONCLUDING REMARKS: A BIOCULTURAL AXIOM

The research accumulated in the three last decades by investigators in the fields of conservation biology, linguistics and anthropology of contemporary cultures, ethnobiology, and ethnoecology has converged toward a shared principle: the world’s biodiversity will be effectively preserved only by protecting the diversity of human cultures, and vice versa. This principle, which represents a new biocultural axiom, is supported by four main sets of evidence: the geographical overlap between biological richness and linguistic diversity and between indigenous territories and biologically high-value regions (actual and projected protected areas), the recognized importance of indigenous peoples as managers and inhabitants of well-preserved habitats, and the certification of ecologically sustainable behavior among indigenous peoples derived from their premodern belief–knowledge–practices complex.

This bio-cultural axiom, referred to by the late Bernard Nietschmann as the “concept of symbiotic conservation [in which] biological and cultural diversity are mutually dependent and geographically coterminous,” constitutes a key principle for conservation theory and application, and epistemologically it is an expression of the new, integrative, interdisciplinary research that is gaining recognition in contemporary science.

See Also the Following Articles

BIODIVERSITY-RICH COUNTRIES • ETHNOBIOLOGY AND ETHNOECOLOGY • HUNTER-GATHERER SOCIETIES • POVERTY AND BIODIVERSITY • SOCIAL AND CULTURAL FACTORS • TRADITIONAL CONSERVATION PRACTICES • TROPICAL ECOSYSTEMS

Bibliography


INSECTICIDE RESISTANCE

Ian Denholm and Greg Devine
IACR—Rothamsted

I. Introduction
II. Extent of Resistance
III. Origins of Resistance
IV. Mechanisms of Resistance
V. Homology of Resistance Genes
VI. How Often Do Resistance Genes Arise?
VII. Cross-Resistance and Multiple Resistance
VIII. Diagnosis of Resistance
IX. Selection of Resistance Genes
X. Combating Insecticide Resistance
XI. Resistance in Nonpest Species
XII. The Special Case of Transgenic Plants
XIII. Concluding Remarks

GLOSSARY

acetylcholinesterase (AChE) The enzyme responsible for breaking down the neurotransmitter acetylcholine (ACh) at nerve synapses, thereby preventing hyperexcitation of cholinergic pathways in the nervous system.

bioassay (biological assay) A laboratory test for evaluating the response of organisms to a toxin and for diagnosing the presence or absence of resistance.

cross-resistance The ability of a single gene or mechanism to confer resistance to more than one toxin.

cytochrome P450 monoxygenases A ubiquitous group of enzymes involved in the NADPH-mediated oxidation and metabolism of a broad range of endogenous and exogenous substrates.

GABA receptor Part of the inhibitory ion channel complex gated by GABA (γ-aminobutyric acid) in postsynaptic nerve membranes.

glutathione S-transferases (GSTs) Enzymes that catalyze the metabolism of a range of substrates following their conjugation with the endogenous tripeptide glutathione.

multiple resistance The occurrence of more than one resistance mechanism in the same individual or pest population.

synergist A chemical used, at sublethal concentrations, to inhibit particular groups of detoxifying enzymes and therefore to implicate the involvement of these enzymes in resistance.

voltage-gated sodium channel A large transmembrane protein that regulates the flow of sodium ions across axonal membranes and mediates the rising phase of action potentials.

THE PHENOMENON OF INSECTICIDE RESISTANCE reflects a genetic adaptation enabling arthropods to survive exposure to otherwise lethal amounts of insecticide. This article reviews the genetic, biochemical, and ecological basis of resistance as well as options for combating its detrimental impact on crop protection and disease management.
I. INTRODUCTION

The diversity of organisms and their genetic variation have been forged by evolution. In evolution, undirected mutations generate unstructured diversity, which is then structured by selection. The time scales over which selection and adaptation occur in eukaryotes are usually too large to observe in situ (except in a few cases, such as industrial melanism in some moths and ladybirds), but there is one microevolutionary process for which many of the factors driving selection and adaptation are well understood, a process that is being increasingly studied and manipulated by a large number of biologists. This is the phenomenon of pest resistance to insecticides, and its study can reveal much about how biodiversity originates, at the infraspecific level at least. Similar processes underpin the evolution of herbicide resistance by weeds, fungicide resistance by plant pathogens, and drug resistance by disease-causing microbes.

II. EXTENT OF RESISTANCE

Although a relatively recent phenomenon (resistance to the first synthetic insecticide, DDT, was initially reported in the 1940s), insecticide resistance is now very widespread. Available statistics (Fig. 1) show that reports of resistant arthropod species increased almost exponentially between 1950 and 1980, following the successive introduction of different classes of synthetic insecticides. By 1990, over 500 species were reported to resist chemicals of at least one insecticide class, and many of these resisted several classes simultaneously. Of the resistant species reported in 1990, 88% were insects (class Insecta) and 12% were mites and ticks (class Arachnida, order Acarina). Four insect orders—Coleoptera (beetles), Diptera (true flies), Heteroptera (aphids, bugs, hoppers, and whiteflies), and Lepidoptera (moths)—accounted for 92% of the resistant insect species; the remainder mostly comprised cockroaches, thrips, lice, and fleas.

Although almost all insecticide classes are now affected by resistance, its extent varies greatly between species. In some insects, resistance only extends to a few closely related compounds in a single class; it may be very weak or restricted to a small part of their geographical range. At the other extreme, some widespread pests such as the diamondback moth (Plutella xylostella), the Colorado beetle (Leptinotarsa decemlineata), the peach–potato aphid (Myzus persicae; Fig. 3c), and the cotton whitefly (Bemisia tabaci; Fig. 3b) now resist most or all of the insecticides available for their control. The most extensively used insecticide classes—organochlorines, organophosphates, carbamates, and pyrethroids—have generally been the most seriously compromised by resistance, and many principles relat-

![Figure 1](image) Increase in the number of arthropod species reported to resist insecticides over time, in total, and in response to the four most widely used classes of insecticide (courtesy GP Georgiou).
is introduced as they are during its use in the field. Resistance is therefore a preadaptive phenomenon reflecting the selection of individuals possessing heritable genetic traits that promote their survival or reproduction in environments treated with insecticides. Estimates of mutation rates are as imprecise as they are for other adaptive traits, ranging from $10^{-5}$ to $10^{-16}$ depending on the mutational event involved (see below). The role of enzyme induction in resistance has not been demonstrated satisfactorily but is likely to be only slight. Increased tolerance caused by environmental or biological factors such as diet, age, or climate can sometimes be significant but is outside the scope of this article.

IV. MECHANISMS OF RESISTANCE

Some of the most significant recent progress in understanding resistance as an adaptive phenomenon has resulted from the application of molecular biology to resistance research. Depending on the mechanism involved, resistance has been shown to arise through structural alterations of genes encoding target-site proteins or detoxifying enzymes or through processes affecting gene expression (e.g., amplification or altered transcription). Other mechanisms that have been demonstrated or postulated include reduced penetration of insecticides through the insect cuticle, enhanced excretion of insecticides, and behavioral traits enabling pests to reduce or avoid exposure to a toxin. However, the latter are generally considered to be relatively minor in effect or to arise only under very specialized circumstances. Figure 2 is a schematic representation of some of the resistance mechanisms discussed in more detail below.

A. Increased Detoxification of Insecticides

The three major routes of detoxification implicated in resistance are as follows:

Enhanced metabolism of insecticides by cytochrome P₄₅₀ monoxygenases can potentially confer resistance to most chemical classes. However, much of the evidence for this mechanism is indirect, based on the ability of compounds such as piperonyl butoxide (PBO), which are known inhibitors of monoxygenases, to reduce the magnitude of resistance when used as synergists in bioassays. Cases in which enhanced oxidative metabolism has been demonstrated directly through binding or metabolism studies are
still uncommon. Since it is now apparent that PBO and related molecules can affect other processes in insects, including cuticular penetration and esterase hydrolysis, claims for this type of mechanism based purely on synergism studies should be interpreted with caution.

Enhanced activity of glutathione S-transferases (GSTs) is considered potentially important in resistance to some classes of insecticide, including organophosphates. However, information on the role of GSTs in resistance is still sketchy since this mechanism cannot be diagnosed reliably from bioassays. GSTs, like monooxygenases, exist in numerous molecular forms with distinct properties, making correlations of enzyme activity with resistance very challenging and often ambiguous.

Enhanced hydrolysis or sequestration by esterases (e.g., carboxylesterases) capable of binding to and cleaving carboxyester and phosphotriester bonds undoubtedly plays a significant role in resistance to organophosphates and pyrethroids. Of the three main types of detoxification mechanisms, this is by far the best characterized biochemically, and in some cases (e.g., for mosquitoes, blowflies, and the aphid Myzus persicae) the esterases involved have been identified, characterized, and sequenced at the molecular level. Resistance due to increased esterase activity can arise through either a qualitative change in an enzyme, improving its hydrolytic capacity, or (as in mosquitoes and aphids) a quantitative change in the titer of a particular enzyme that already exists in susceptible insects.

B. Alterations to Insecticide Target Sites

Resolving target-site modifications that lead to resistance clearly requires some knowledge of the mode of action of insecticides themselves. At present, this information is most advanced for molecules binding to enzymes or receptors in the nervous system of arthropods. Three examples of target site resistance are now well understood:

Organophosphates and carbamates exert their toxicity by inhibiting the enzyme acetylcholinesterase (AChE), thereby impairing the transmission of nerve impulses across cholinergic synapses. Mutant forms of AChE showing reduced inhibition by these insecticides have been demonstrated in several insect and mite species. Biochemical and molecular analyses of insecticide-insensitive AChE have shown that pests may possess several different mutant forms of this enzyme with contrasting insensitivity profiles,
thereby conferring distinct patterns of resistance to these two insecticide classes.

Pyrethroids act primarily by binding to and blocking the voltage-gated sodium channel of nerve membranes. A mechanism postulated to reflect insensitivity of this target site was first identified in houseflies (Musca domestica) and termed knockdown resistance or kdr. Recently, there has been rapid progress in attributing kdr resistance unequivocally to structural modifications of a sodium channel protein and in locating and identifying the genetic mutations responsible. As a consequence, kdr resistance has now been confirmed in many insect pests. As with insensitive AChE, there can be different forms of kdr resistance (e.g., a more potent “super-kdr” form in houseflies), and this can have important implications for cross-resistance within the pyrethroid class.

GABA receptors are targets for several insecticide classes, including cyclodiennes (a subclass of the organochlorines), avermectins, and fipronils. The primary mechanism of resistance to cyclodiennes and fipronils in several species involves modification of a particular GABA receptor subunit, resulting in substantial target-site insensitivity to these insecticides.

V. HOMOLOGY OF RESISTANCE GENES

Although there are several possible mechanisms, genetic options for resisting insecticides can also be very limited, especially for mechanisms based on decreased sensitivity of insecticide target sites. The target-site mechanism of cyclodiene resistance has been attributed to the same amino acid substitution (alanine-302 to serine) in GABA receptors of several species of diverse taxonomic origin, including Drosophila, several beetles, a mosquito (Aedes aegypti), a whitefly (Bemisia tabaci), and a cockroach (Blatella germanica). Work on the two other principal target-site mechanisms—altered AChE and kdr—has proved more complex due to the occurrence of multiple resistance alleles at the same loci. In the case of kdr, however, there is also evidence for the same amino acid substitution (leucine-1014 to phenylalanine) in a sodium channel protein conferring a “basal” kdr phenotype in a range of species including houseflies, cockroaches, the peach–potato aphid, the diamondback moth, and a mosquito (Anopheles gambiae). This phenotype may subsequently be enhanced (to “super-kdr” resistance) by further mutations that also recur between species. Despite the structural complexity of the receptors involved, these parallel mutations imply that the opportunities for insects to modify them to avoid or reduce binding of insecticides, while retaining normal functioning of the nervous system, are very limited indeed.

When susceptible individuals of the sheep blowfly (Lucilia cuprina) were exposed to the mutagen ethyl methanesulfonate (EMS) and their progeny screened for resistance to dieldrin (a cyclodiene), surviving insects not only exhibited a GABA receptor based mechanism analogous to that found in nature but also exhibited an identical alanine to serine amino acid substitution in the channel gene. Similarly, mutagenesis followed by screening with diazinon (an organophosphate) led to the recovery of a resistance mechanism showing identical toxicological, biochemical, genetic, and molecular properties to one that had previously evolved to diazinon under field conditions. These findings again demonstrate tight evolutionary constraints on the number of viable resistance mutations, even in the laboratory, where mutations conferring deleterious effects on overall fitness might be expected to survive better than in the open field. Interestingly, they also highlight the potential of using mutagenesis to predict likely resistance mechanisms to novel insecticides well in advance of them appearing in field populations and to tailor resistance management recommendations accordingly.

In other cases, different types of resistance to the same toxin exist and can account for differences in the toxicological and genetic basis of resistance between species or between different geographical populations of the same species. Insecticidal proteins from the soil bacterium Bacillus thuringiensis (Bt) are becoming increasingly important in pest management, especially in relation to insect-tolerant transgenic crops (see below).

To date, the only species to have evolved Bt resistance in the field is the diamondback moth, Plutella xylostella. The majority of Bt-resistant populations examined have exhibited very similar characteristics, including a very consistent pattern of cross-resistance to different Bt toxins and recessive inheritance. However, there are also strains of P. xylostella in which the breadth and inheritance of Bt resistance differ markedly from this “mode,” implying the existence of distinct resistance genes and/or mechanisms.

VI. HOW OFTEN DO RESISTANCE GENES ARISE?

The recurrence of specific resistance mutations within and between taxa begs another question of fundamental
significance to the origins of biodiversity: Have such mutations arisen repeatedly within the same species, or appeared on only a limited number of occasions and subsequently spread through migration and/or human agency? This question is proving amenable to investigation by sequencing not only the resistance genes themselves but also flanking regions and introns, which would be expected to vary between alleles that have arisen independently. In the mosquito, *Culex pipiens*, organophosphate resistance is primarily conferred by allozymes at two closely linked loci (esterases A and B) coding for insecticide-detoxifying carboxylesterases. Overproduced allozymes (resulting from amplification of A or B genes) tend to recur in geographically disjunct areas. This situation could be explained by recurrent mutation generating each amplification event *de novo* or by a nonrecurrent mutation that has subsequently spread within and between populations. Restriction mapping of DNA around the esterase genes points to the latter explanation, with large-scale gene flow (even between continents) most likely attributable to passive migration of mosquitoes on ships and/or airplanes. It is notable that the recent appearance of a new resistance allele in southern France is known to have originated in the vicinity of the international airport and seaport at Marseilles.

Organophosphate resistance in the aphid *Myzus persicae* is also attributable to the amplification of a gene encoding an insecticide-detoxifying carboxylesterase. Despite the often widespread dispersion of these amplified genes in the aphid genome, restriction analyses have indicated that all copies are in the same immediate genetic background. This suggests that amplification occurred only once, with the amplified DNA subsequently being moved intact around the genome through chromosomal rearrangements, or perhaps mediated by transposable elements. Similarities in the position and structure of these genes in aphids of diverse geographic origin reinforce the likelihood of a single amplification event that has subsequently become widely dispersed around the world.

There is also molecular evidence for some resistance genes having several independent origins in the same species (e.g., for target-site resistance to cyclodiene in the red flour beetle, *Tribolium castaneum*). However, results for mosquitoes and aphids highlight the potential for large-scale inadvertent movement of resistant insects between countries or even continents. For crop pests, these risks are particularly acute due to the increasing international trade in edible and ornamental plants, many of which have been treated with insecticides at their point of origin. In such cases, growers at the receiving end of the trade network face a dual threat: (i) the establishment of new pest species or more aggressive biotypes of existing ones and (ii) the possibility that such pests are already strongly resistant to compounds that might otherwise be used to suppress them or to eradicate them entirely.

**VII. CROSS-RESISTANCE AND MULTIPLE RESISTANCE**

Arthropods seldom if ever resist just one toxin. Most commonly, they exhibit differing levels of resistance to a range of related and unrelated insecticides. In its strictest sense, the term cross-resistance refers to the ability of a single mechanism to confer resistance to several insecticides simultaneously. A more complex situation is that of multiple resistance, reflecting the coexistence of two or more resistance mechanisms, each with their own specific cross-resistance characteristics. Disentangling cross-resistance from multiple resistance, even at the phenotypic level, is one of the most challenging aspects of resistance research. However, a knowledge of the mechanisms involved is often essential in order to develop resistance management recommendations based, for example, on the alternation of insecticides to avoid continuous selection for the same resistance gene or mechanism.

Unfortunately, cross-resistance patterns are inherently difficult to predict in advance, since mechanisms based on both increased detoxification and altered target sites can differ substantially in their specificity. The most commonly encountered patterns of cross-resistance tend to be limited to compounds within the same chemical class (equivalent to the term *side resistance* as used by parasitologists). However, even these patterns can be very idiosyncratic. For example, organophosphate resistance based on increased detoxification or target-site alteration can be broad-ranging across this group or highly specific to a few chemicals with particular structural similarities. The breadth of target-site resistance to pyrethroids in houseflies is also dependent on the resistance allele present. The kdr allele itself affects almost all compounds in this class to a similar extent (ca. 10-fold resistance), whereas resistance due to the more potent super-kdr allele is highly dependent on the alcohol moiety of pyrethroid molecules, ranging from ca. 10-fold to virtual immunity. Cross-resistance between insecticide classes is even harder to anticipate, especially for broad-spectrum detoxification systems whose specificity depends not on insecticides having the
same mode of action but on the occurrence of common structural features that bind with detoxifying enzymes.

Empirical approaches for distinguishing between cross-resistance and multiple resistance include (i) repeated back-crossing of resistant populations to fully susceptible ones to establish whether resistance to one chemical cosegregates consistently with resistance to another and (ii) reciprocal selection experiments whereby populations selected for resistance to one chemical are examined for a correlated change in response to another. If available, biochemical or molecular diagnostics for specific resistance genes can assist considerably with tracking the outcome of genetic crosses or with assigning cross-resistance patterns to particular mechanisms.

VIII. DIAGNOSIS OF RESISTANCE

Although a large number of laboratory bioassay methods have been developed for detecting and characterizing resistance, most of these are limited to defining phenotypes and provide little or no information on the underlying genes or mechanisms. Thus, although bioassays remain the mainstay of most large-scale resistance monitoring programs, much attention is being paid to developing more incisive techniques that not only offer greater precision and throughput but also diagnose the type of mechanism(s) present and, whenever possible, the genotypes of resistant insects. A variety of approaches are being adopted for this purpose, including electrophoretic or immunological detection of resistance-causing enzymes, kinetic and end-point assays for quantifying the activity of enzymes or their inhibition by insecticides, and DNA-based diagnostics for mutant resistance alleles. The sensitivity of these techniques can provide is exemplified well by work on the aphid Myzus persicae, which in northern Europe possesses at least three coexisting resistance mechanisms: (i) an overproduced carboxylesterase conferring resistance to organophosphates, (ii) an altered AChE conferring resistance to certain carbamates, and (iii) target-site (kdr) resistance to pyrethroids. These mechanisms collectively provide strong resistance to virtually all available aphicides. Fortunately, it is now possible to diagnose all three in individual aphids using an immunoassay for the overproduced esterase, a kinetic microplate assay for the mutant AChE, and a PCR-based diagnostic for the kdr allele (Fig. 4). The combined use of these techniques against field populations provides up-to-date information on the incidence of the mechanisms and is used to alert growers to potential control problems.

IX. SELECTION OF RESISTANCE GENES

The rate at which resistance genes are selected reflects the combined influence of numerous biotic and abiotic factors. Resistance offers several advantages for research to resolve these factors and their interactions. First, the selecting agent (exposure to insecticides) is well understood, it can usually be carefully documented using treatment histories or manipulated to investigate its effect on selection rates. Second, the selective advantages conferred by resistance genes are often very large, leading to substantial changes in genetic composition over a measurable time frame. Third, most of the major mechanisms of insecticide resistance, unlike those for many stress-related adaptations, are controlled by single genes of major effect (monogenic) rather than many genes, each of small effect (polygenic). This renders resistance readily amenable to analysis within the conventional theoretical framework of ecological genetics. Finally, the frequent availability of bioassays for quantifying the frequency of resistance phenotypes, or even in vitro assays for specific genotypes, enables accurate documentation of responses to selection applied in population cages in the laboratory or under open field conditions.

Factors determining the selection (and hence risk) of resistance to insecticides can, for convenience, be
classified into genetic or ecological ones relating to the intrinsic properties of pests and resistance mechanisms and operational ones relating to the chemical itself and how it is applied. Some of the most important factors apparent from the large body of experimental and theoretical research on resistance selection are summarized below.

A. Genetic Influences

In order for resistance to evolve, resistance genes must confer a selective advantage over their susceptible counterparts. One of the primary challenges for describing resistance is therefore to estimate the relative fitness of different genotypes under exposure to insecticides. There are different ways of achieving this, the most direct being to release individuals of known susceptible and resistance genotypes into insecticide-treated environments and to monitor their survival. This has been attempted for a variety of pest species and has identified many, often subtle, influences on how resistance genes are expressed in the field. The dominance of resistance genes, which exerts a major influence on selection rates, is a case in point. In laboratory bioassays evaluating the relative survival of susceptible homozygotes (SS), heterozygotes (RS), and resistance homozygotes (RR) over several insecticide concentrations, dominance can be measured precisely, with RS individuals usually responding in an intermediate manner. In the field, dominance is a changeable phenomenon, depending on the concentration of insecticide applied and its uniformity over space and time. Even when the initial concentration is sufficient to kill RS individuals (rendering resistance effectively recessive), the weathering or decay of residues may result in this genotype showing increased survival and resistance becoming functionally dominant in expression. When resistance genes are still rare, and hence mainly present in heterozygous condition, this can have a profound effect in accelerating the selection of resistance genes to economically damaging frequencies.

The diverse mating systems of insects also influence the rate at which resistance evolves. Although most research has focused on outcrossing diploid species (typified by members of the Lepidoptera, Coleoptera, and Diptera), systems based on haplodiploidy and parthenogenesis also occur among key agricultural pests. In haplodiploid systems, males are produced unparentally from unfertilized, haploid eggs, and females are produced biparentally from fertilized, diploid eggs. The primary consequence of this (exemplified by whiteflies, spider mites (Fig. 3d), and phytophagous thrips) is that resistance genes are exposed to selection from the outset in haploid, hemizygous males, irrespective of intrinsic dominance or recessiveness. Whether a resistance gene is dominant, semidominant, or recessive, resistance can develop at a similar rate under haplodiploidy, whereas recessiveness can cause significant delays (initially at least) in diploid populations.

Most species of aphid undergo periods of parthenogenesis, promoting the selection of clones with the highest levels of resistance and/or the most damaging combination of resistance mechanisms. However, in holocyclic populations (ones that alternate between sexual and asexual reproduction), this effect is at least partially countered by genetic recombination and the subsequent reassortment of mechanisms during sexual reproduction. In fully anholocyclic (asexual) populations, such as those of Myzus persicae in northern Europe, the influence of parthenogenesis is much more severe and has led to strong and persistent associations between resistance mechanisms within clonal lineages exposed to a succession of different selecting agents.

B. Ecological Influences

Several aspects of pest ecology, including the dynamics, phenology, and dispersal capabilities of pest organisms, act as primary determinants of resistance development. However, their influence on selection rates can be unpredictable without a sound knowledge of how they interact with patterns of insecticide use. As an example, movement of pests between untreated and treated parts of their range may delay the evolution of resistance, due to the diluting effect of susceptible immigrants. Conversely, large-scale movement can, as described above, also accelerate the spread of resistance by transferring resistance alleles between localities.

For highly polyphagous crop pests, interactions between pest ecology and insecticide treatments play a particularly critical role in determining selection pressures. Key factors to be considered are the seasonality and relative abundance of treated and untreated plant hosts and patterns of migration between hosts at different times of the year. A good example relates to the two major bollworm species (Lepidoptera: Noctuidae) attacking cotton in Australia. Only Helicoverpa armigera (Fig. 3a) has developed strong resistance; H. punctigera, despite being an equally important cotton pest, has remained susceptible to all insecticide classes. The most likely explanation is that H. punctigera occurs in greater abundance on a larger range of unsprayed hosts than H. armigera, thereby precluding a significant increase in resistance on treated crops.

However, polyphagy can sometimes be deceptive. In the cotton/vegetable/melon production systems of the
C. Operational Influences

Although closely linked to aspects of pest genetics and ecology, operational factors are best distinguished as ones which, in principle at least, are at man's discretion and can be manipulated to influence selection rates. Factors exerting a major influence in this respect include the rate, method, and frequency of applications, their biological persistence, and whether insecticides are used singly or as mixtures of active ingredients.

Equating operational factors with selection is often difficult, since without a detailed knowledge of the resistance mechanisms present it is impossible to test many of the assumptions on which genetic models of resistance are based. Anticipating the selection pressure imposed by a particular application dose of insecticide is a case in point. If resistance alleles are present, the only entirely nonselecting doses will be ones sufficiently high to overpower all individuals, regardless of their genetic composition, or ones sufficiently low to kill no insects at all. The latter is obviously a trivial option. Prospects of achieving the former depend critically on the potency and dominance of resistance genes present. A pragmatic solution to this dilemma is to set application doses as far above the tolerance range of SS individuals as economic and environmental constraints permit, in the hope that at least RS genotypes will be effectively controlled. Even this approach can backfire badly if resistance turns out to be more common than suspected (resulting in the presence of RR homozygotes) or resistance alleles exhibit an unexpectedly high degree of dominance. Unless a high proportion of insects escape exposure altogether, the consequence could then be to select very rapidly and effectively for homozygous resistant populations.

In practice, concerns over optimizing dose rates to avoid resistance are secondary to ones regarding the application process itself. Delivery systems and/or habitats promoting uneven or inadequate coverage will generally be more prone to selecting for resistance, since pests are more likely to encounter exposure conditions under which selection is most intense. This was elegantly demonstrated through experiments assessing the relative survival of cyclodiene-susceptible and -resistant phenotypes of the coffee berry borer (Hypothemus hampei) in coffee plantations treated with this chemical in New Caledonia. The practice of spraying plantations from roadsides with vehicle-mounted mistblowers generated gradients in the concentration of endosulfan that resulted in different selection pressures in different parts of each field. Similarly, underdosing with the fumigant phosphine in inadequately sealed grainstores has been implicated as a primary cause of resistance to this chemical in a range of stored product pests.

The timing of insecticide applications relative to the life cycle of a pest can also be an important determinant of resistance. A good example relates to the selection of pyrethroid resistance in the cotton bollworm, Helicoverpa armigera, in Australia. On cotton foliage freshly treated with the recommended field dose, pyrethroids killed larvae up to 3–4 days old irrespective of whether they were resistant or not by laboratory criteria. Since the sensitivity of larvae of all genotypes to pyrethroids was found to decline with increasing larval size, the greatest discrimination between susceptible and resistant phenotypes occurred only when larvae achieved a threshold age. Targeting of insecticides against newly hatched larvae, as is generally advocated for bollworm control, not only increases the likelihood of contacting larvae at the most exposed stage in their development but also offers the greatest prospect of retarding resistance by overpowering its expression.

In practice, persistent insecticides are often essential
to ensure an acceptable period of control, especially when contending with disease vectors or continued invasion of crop pests from alternative host plants. However, persistent applications can accentuate resistance development by exposing a larger number of individuals to the selecting agent. Another problem is that residues of persistent insecticides decay or become diluted through plant growth, so that resistant insects may survive more effectively than they did at the time of application. Empirical studies with a range of pests including mosquitoes, bollworms, and blowflies have demonstrated that aged deposits discriminate more readily between genotypes or phenotypes than ones freshly applied.

In theory, the coapplication of two or more unrelated chemicals as insecticide mixtures offers substantial benefits for delaying the selection of resistance. The underlying principle is one of "redundant killing," whereby any individuals already resistant to one insecticide are killed by simultaneous exposure to another, and vice versa. However, achieving this objective requires not only that each type of resistance is still rare but also that the ingredients confer mutual protection throughout the effective life of an application. Failure to ensure that they exhibit similar biological persistence may lead to one compound exerting greater selection pressure than the other, thereby accelerating the selection of doubly resistant phenotypes. Two potentially conflicting challenges of choosing ideal mixture partners are therefore (i) to ensure maximum similarity in efficacy and persistence against the target pest(s) and (ii) to ensure maximum dissimilarity in chemical structure and mode of action to minimize the likelihood of cross-resistance. Difficulties with identifying candidate molecules that meet all these criteria have greatly limited the use of mixtures for combating resistance to conventional insecticides, although they have considerable appeal for sustaining the effectiveness of insect-tolerant transgenic crops (see below).

D. Fitness of Resistant Individuals

Despite the advantages they confer under exposure to insecticides, it is often assumed that resistance genes also confer physiological costs that could lead to counterselection when insecticides are not applied. Some of the best examples of such fitness costs come from studies conducted under harsh or stressful environmental conditions, when even slight differences in relative fitness are likely to have major consequences for the survival of genotypes. For example, resistant strains of bollworms, blowflies, and aphids have all been demonstrated to overwinter less successfully than their susceptible counterparts. Possible explanations for these fitness differentials include the reduced viability of certain life stages, a slower reproductive rate rendering resistant insects more vulnerable to adverse climatic conditions or to predation, or a reduced ability to respond to environmental cues promoting survival. In the aphid Myzus persicae, resistant individuals are less inclined to move from senescing to younger leaves and are therefore more vulnerable to isolation and starvation after leaf abscission. These costs have been implicated in the decline in frequency of resistant genotypes in the absence of insecticide pressure, leading to a balancing polymorphism for resistance rather than a consistent accumulation of resistance from one season to another.

Fitness costs associated with resistance can be difficult to demonstrate experimentally, since deleterious effects may only be expressed under particular environmental conditions or conferred by other genes closely linked to the resistance locus. The most convincing examples are ones in which costs have been found consistently in resistant populations of diverse geographic origin or have persisted after several generations of back-crossing to susceptible insects in order to exclude linkage effects. The potential for fitness drawbacks to be overcome by a process of coadaptation, i.e., the integration of resistance genes with other "modifier" loci that ameliorate fitness costs, has also proved challenging to demonstrate. In a few cases, however, repeated back-crossing of resistant insects to susceptible ones has led to seemingly fit resistance phenotypes acquiring a fitness penalty, apparently due to the uncoupling of resistance genes from modifier loci.

X. COMBATING INSECTICIDE RESISTANCE

In most studies of evolution, the primary challenge is to identify selective forces and to interpret their effects on the genetic composition of individuals and populations. With insecticide resistance, it is also necessary to intervene in the evolutionary process and find ways of reducing its deleterious impact on pest management. Failure to do so in the past has had many severe consequences, including the economic failure of cropping systems, the resurgence of insect-transmitted pathogens, and damage to the environment by way of increased insecticide applications.

The concept of insecticide resistance management (IRM) aims to address these concerns through the development of control strategies for overcoming resistance to currently used compounds, or preventing its
appearance in the first place. Although drawing extensively on the theoretical and empirical framework that evolutionary biology provides, IRM strategies must also contend with several practical, economic, and political constraints on the choice of possible management tactics and the precision with which they can be applied. The most important of these are as follows:

The properties of any resistance genes present will often be unknown, and knowledge of pest ecology may still be rudimentary. It is often necessary to contend with a whole pest complex rather than just a single pest species. There will often be a very limited number of insecticides available for use in management strategies. For highly mobile pests at least, countermeasures may need to be standardized and synchronized over large areas, sometimes whole countries.

Resistance is a dynamic phenomenon; i.e., any mechanisms already known to exist may change over time. Continued monitoring is vital to determine whether management recommendations remain valid or need to be revised in light of changing circumstances or new knowledge gained.

To promote compliance with management strategies, the countermeasure adopted should be as unambiguous, rational, and simple as possible.

A strategy first implemented on Australian cotton in 1983 to contend with the bollworm, Helicoverpa armigera, illustrates many features of large-scale attempts at resistance management. It was introduced in response to unexpected, but still localized, outbreaks of pyrethroid resistance in *H. armigera* and was based primarily on the concept of insecticide rotation. The threat of pyrethroid resistance was countered by restricting these chemicals to a maximum of three sprays within a prescribed time period coincident with peak bollworm damage. Farmers were required to use alternative insecticide classes at other stages of the cropping season, in order to diversify the selection pressures being applied.

Compliance with this strategy was excellent, and initially it had the desired effect of preventing a systematic increase in the frequency of pyrethroid-resistant phenotypes. Additional recommendations resulting from work on the ecological genetics of pyrethroid resistance, including the targeting of insecticides against newly hatched larvae (when even resistant insects can be killed) and the plowing-in of cotton stubble to destroy resistant pupae overwintering in the soil, undoubtedly contributed to this success. Unfortunately, the restrictions placed on pyrethroid use were inadequate to prevent a gradual, long-term buildup of pyrethroid resistance. As a result, pyrethroids are no longer considered reliable control agents for *H. armigera*, although they remain highly effective against a coexisting species, *H. punctigera*. The strategy has therefore been revised extensively to place greater emphasis on distinguishing between the two *Helicoverpa* species and on the strategic use of nonpyrethroids against *H. armigera*. Transgenic cotton plants expressing *Bacillus thuringiensis* (Bt) toxins have since been released commercially in Australia. Tactics for deploying these without selecting rapidly for Bt resistance are therefore being investigated as a matter of urgency (see below).

Another strategy incorporating a wide range of chemical and nonchemical countermeasures was introduced on Israeli cotton in 1987. This had the primary objective of conserving the effectiveness of insecticides against the whitefly, *Bemisia tabaci*. Under recommendations coordinated by the Israeli Cotton Board, important new whitefly insecticides are restricted to a single application per season within an alternation strategy optimized to contend with the entire cotton pest complex and to exploit biological control agents to the greatest extent possible. One major achievement of this strategy has been a dramatic reduction in the number of insecticide applications against the whole range of cotton pests, but especially against *B. tabaci*. Sprays against whiteflies now average less than two per growing season compared with over 14 per season in 1986. Most importantly of all, the strategy has generated an ideal environment for releasing additional new insecticides onto cotton and for managing them effectively from the outset.

**XI. RESISTANCE IN NONPEST SPECIES**

Compared to its prevalence in arthropod pests, insecticide resistance is still relatively rare among nonpest species including beneficial organisms. However, it has been well documented in a few species of hymenopteran parasitoids and predatory mites, some of which are being exploited in integrated pest management (IPM) systems. Its rarity among beneficial organisms is probably due in part to difficulties in locating hosts and prey (and hence surviving) under exposure to insecticides. The likelihood of resistance developing in beneficial arthropods may be increased if the insects they depend on as prey are already resistant, although this requires further research. It is also likely that, in comparison with herbivorous species, the enzyme systems of predators and parasites are less well adapted to detoxify xenobiotics.
The propensity for beneficial insects to evolve resistance obviously depends on the degree of selectable variation within their populations. Although laboratory selection has been used to enhance low levels of pesticide tolerance found in field populations, there is a concern that such selection applied at artificially low doses will promote polygenic traits that could fragment and dissipate if released into natural populations. However, when substantial resistance has evolved naturally in the field, its mechanisms have tended to be similar to ones found in pest species. Organophosphate-resistant strains of the green lacewing, Chrysoperla carnea, have been shown to exhibit increased activity of acetylcholinesterase (AChE) compared to susceptible insects. A carboxylesterase enzyme, very similar in amino acid sequence to that conferring organophosphate resistance in the aphid, Myzus persicae, has been cloned and sequenced from a malathion-resistant strain of the parasitic wasp Anisopteromalus calandrae.

Since the development of resistance is dependent on the ecology of systems in which it appears, interactions between beneficial and pest species will greatly affect the epidemiology and dynamics of resistance in both. For example, some parasitoids of stored-grain beetles are resistant to insecticides, and it is thought that this adaptation has been encouraged by the fact that their parasitic larvae are sheltered from insecticides by the grain kernels inhabited by their hosts. This is thought to protect a substantial part of the insect life cycle from insecticide selection and ensures that relatively small shifts in insecticide tolerance by the parasitoid afford significant protection against the decreased insecticide doses that do penetrate their defenses.

Natural enemies may contribute to retarding resistance in pest species by exerting sufficient control to decrease the number of insecticidal treatments required. Conversely, there are ways in which natural enemies could promote the adaptation of pests to insecticides. For example, the selection pressure for resistance would be increased if weaker, sublethally affected individuals were more easily preyed upon or parasitized than their fully resistant and therefore unaffected counterparts.

XII. THE SPECIAL CASE OF TRANSGENIC PLANTS

A new development in crop protection with an important bearing on resistance is the release of crop plants genetically engineered to express genes for insecticidal toxins derived from the microbe Bacillus thuringiensis (Bt). Bt cotton and/or corn is already being grown commercially on a large scale in the United States, Canada, Australia, Mexico, China, and South Africa. In 1998, the total area worldwide planted with Bt crops was estimated to exceed 12 million ha. Existing toxin genes in Bt cotton and corn are active specifically against certain key lepidopteran pests (especially bollworms and corn borers); another engineered into potatoes provides protection against the Colorado beetle, Leptinotarsa decemlineata.

Aside from their commercial prospects, insect-tolerant transgenic crops offer numerous potential benefits to agriculture. The incorporation of Bt genes into crops offers constitutive expression of toxins in plant tissues throughout a growing season. This could reduce dramatically the use of conventional broad-spectrum insecticides against insect pests as well as remove the dependence of pest control on extrinsic factors, including climate and the efficiency of traditional application methods. However, this high and persistent level of expression also introduces a considerable risk of pests adapting rapidly to resist genetically engineered toxins. To date, there are no substantiated reports of resistance selected directly by exposure to commercial transgenic crops, but resistance to conventional Bt sprays (selected in either the laboratory or the field) has been reported in more than a dozen species of insect. Research into the causes and inheritance of such resistance is providing valuable insights into the threats facing Bt plants and the efficacy of possible countermeasures.

Tactics proposed for sustaining the effectiveness of Bt plants have many parallels with ones considered for managing resistance to conventional insecticides. However, they are more limited in scope due to the long persistence and constitutive expression of engineered toxins and the limited diversity of transgenes currently available. Indeed, for existing “single-gene” plants, the only prudent and readily implementable tactic is to ensure that substantial numbers of pest survive in nontransgenic “refuges,” composed either of the crop itself or of alternative host plants. In the longer term, stacking (pyramiding) of two or more genes in the same cultivar, or possibly rotations of cultivars expressing different single toxins, are potentially more durable options for resistance management. Whatever measures are adopted, it is essential that Bt plants (and their successors expressing other transgenes) are exploited as components of multitactic strategies rather than as a panacea for existing pest management problems, including those arising from the development of resistance to conventional insecticides.
XIII. CONCLUDING REMARKS

Over the past 20 years, few areas of entomology have advanced as rapidly or received such widespread attention as that of insecticide resistance. Research on this topic has provided invaluable insights into the origin and nature of adaptations, and these are in turn proving of much broader significance for understanding genetic responses to manmade change in the environment. In many respects the continuing battle against resistance is analogous to an evolutionary “arms race;” in this case pitting human ingenuity in discovering new toxins against the adaptive capacity of pest species. Debates as to who will eventually win this race are of secondary importance to the realization that for many species the race is probably unnecessary. A wider adoption of resistance management practices, especially through greater exploitation of nonchemical measures, would assist with reducing both the economic impact of resistance and the deleterious effects of many existing pest and disease management strategies on biodiversity in general.

See Also the Following Articles
DIFFERENTIATION • ECOLOGICAL GENETICS • GENES, DESCRIPTION OF • INSECTS, OVERVIEW • PESTICIDES, USE AND EFFECTS OF

Bibliography


I. Introduction

II. Major Divisions

GLOSSARY

hexapod  Group including insects and their primitive relatives.

instar  Stage between molts of immature insects.

pro-, meso-, and metathorax  First, second, and third segments of the thorax.

THE INSECTS are the most species-rich group of organisms known, with the most diverse natural histories of any animals. The most successful insect groups have wings for flight, can fold these wings back over the body, and have development that is divided into four discrete stages: egg, larva, pupa, and adult.

I. INTRODUCTION

The Class Insecta, or the slightly larger Superclass Hexapoda (which additionally includes the orders Collembola, Protura and Diplura), is the world’s most species-rich group of organisms, with about 1 million described species. They are found nearly everywhere on earth, including the terrestrial, aquatic, and, to a much lesser extent, marine ecosystems. They are most diverse in tropical forests, where the undescribed fauna has been estimated to comprise 5, 10, 30, or even 50 million species. Hexapods are a well-established monophyletic group, based on the presence of three major body divisions—head, thorax, abdomen—and a single pair of locomotory appendages on each thoracic segment. Some primitive insects have retained appendages on the abdominal segments, but these are much smaller and less functional than those on the thorax. Most of the more derived groups of insects also have wings, usually one pair on each of the mesothoracic and metathoracic segments, but these have been variously lost or modified in some groups, especially the Diptera (flies).

The fossil record of hexapods extends back to the earliest record of terrestrial life, with Collembola and lower insects recorded from the lower Devonian, almost 400 million years ago, and possibly even earlier traces from the Silurian. Insects have been prominent members of the fossil record ever since, with most prominent major groups having been preserved from late Paleozoic or early Mesozoic formations (200–230 million years ago). A summary of fossil insects is given by Kukalová-Peck (1991).

II. MAJOR DIVISIONS

Insects and their relatives (together referred to as the Hexapoda) are arthropods, with a chitinous exoskeleton and jointed appendages. They grow by molting,
periodically shedding their exoskeleton to allow a new larger body to expand and harden. The closest arthro-

pod relatives to the hexapods are still unknown but probably include some or all of the myriapods (centi-

pedes and millipedes).

The insects formerly included all groups currently classified as Hexapoda, but recently authors have sepa-

rated the Collembola (springtails) and Protura into a separate group called Ellipura or Parainsecta (Table I).

The Diplura are placed either within the Insecta or in a separate order of uncertain affinity. All are soft-bodied

hexapods usually found in soil or decaying organic material. Most are small, 1 to 5 mm in length, but some

Diplura are much larger. The Collembola is the most speices-rich group and the one most commonly encoun-
tered, as they are often enormously abundant in soil, compost, under rocks, in damp places, and even on the

surface of water or snow.

The “true” insects are defined by characters of the structure and musculature of the antenna, among oth-

er. There are two relatively primitive groups, the jump-

ing bristletails (Archaeognatha) and the bristletails or silverfish (Thysanura). Both are elongate, wingless crea-
tures with a long, segmented, median “tail” and two shorter, but still prominent cerci. Both groups are scav-

engers, feeding on vegetable and sometimes animal debris.

The appearance of wings marks the Pterygota. The origin of wings and flight in insects has been controver-
sial, with a number of theories put forward. Most now agree that wings are formed from a movable part of

the of the leg, with its attendant corollary that the

protowings were formed in aquatic insects, possibly

used for ventilating gills and even for locomotion in

water. Extant insects have wings on only the mesotho-

rax and metathorax (the second and third thoracic seg-

ments), but fossils exist for which there are also protho-

racic wings or wing precursors.

Also unlike the more primitive insects, pterygote insects have hemimetabolus development. The imma-

ture stages, often called nymphs, after molting several
times, undergo a final molt that produces fully winged,

sexually mature adults (except in the Ephemeroptera,

which are unique in having a second adult molt; dis-
cussed later). More primitive hexapods are ametabolus,

with gradual growth through several molts until the

sexually mature adult stage, which differs relatively lit-
tle from the immature forms, is reached.

The most primitive winged insects are sometimes
turned into a group called Paleoptera and include the

extant orders Ephemeroptera and Odonata. Unlike their

more advanced relatives, the neopterans, paleopterans
cannot fold their wings down against the body. Both

extant orders of Paleoptera have aquatic larvae, but they
differ markedly in other aspects of their life histories.

Mayflies are generally herbivores, scraping algae and
diatoms from objects in the water. When mature, they emerge from the water in an intermediate, winged stage called the subimagos, which last only a short time. They molt once more and emerge as winged adults, ready to mate, and die soon thereafter. Dragonflies and damselflies, in contrast, have highly predaceous larvae. There is no subimagos stage, as the adults emerge directly and begin relatively long lives as predators of smaller insects. Dragonflies in particular are strong fliers that can travel great distances from water when in search of prey.

The Neoptera are defined by the ability to swirl and fold the wings back at rest. The functional significance of such a modification is obvious—it allows the insect the opportunity to slip into small places to hide from predators and seek food. It has also opened the path to using the wings as protective structures when not in flight, and the forewings of many groups have become hardened or thickened. This is especially evident in the beetles, whose hardened forewings are largely useless for locomotion and are held out of the way during flight.

The relationships among the more primitive Neoptera are not well resolved. One particularly enigmatic group are the stoneflies (Plecoptera), whose aquatic lifestyle is reminiscent of the paleopterans. The larvae live in clear, cool streams or lakes, where they are herbivores or predators. The adults are weaker fliers than Odonata but longer lived than Ephemeroptera, and they feed on algae, decaying vegetation, or detritus. The Blatnoidea, Isoptera, and Mantodea are terrestrial insects sometimes unified in a group called Dictyoptera. Their lifestyles are extremely divergent, however. Cockroaches are flattened, usually cryptic and nocturnal scavengers. They vary in shape and size from small, delicate species a few millimeters in length to large, bulky forms 6 cm long. Most are cryptically colored, but some are mimics of apparently distasteful grasshoppers, crickets, and katydids.

Isoperta are termites, the only eusocial group of insects outside of the Hymenoptera. All species feed on cellulose from wood, leaves, and plant debris that is digested in their gut by a symbiotic microbial fauna of flagellate protozoans or spirochaete bacteria. Some species also culture gardens of fungi in their droppings; the fungi are then consumed and their enzymes aid in digestion. Termites have a number of castes, with most individuals in a colony being soft-bodied workers, along with a lesser number of soldiers and a single queen and king. Nests can be in the ground, in rotting wood or in carton nests built on trees or other structures. Ground-nesting species often build prominent mounds that in some species can reach a few meters in height. Although they appear defenseless, termites have a wide variety of chemical and structural modifications that allow them to resist attack by predators, especially ants.

The praying mantids, Order Mantoptera, are voracious predators. They have modified, spiny, naptorial forelegs that grasp and crush their prey, usually other insects but sometimes even small vertebrates. They have no venom or other mechanism for killing their prey other than by eating them. Most mantids are green or brown in color, camouflageing them from attack by predators and detection by their prey. Some are flattened and specifically colored to resemble leaves or petals of flowers.

Unlike the well-known cockroaches, termites, and mantids of the Dictyoptera, the grylloblattids are poorly known and only relatively recently recognized as a separate order. They are found in cool climates, usually at high elevations, in the northern hemisphere living mostly under rocks or in caves. In the spring and summer they can forage on the snow surface, feeding on dead insects and plant material. Ice crawlers lack wings, which they have secondarily lost probably as an adaptation to their cryptic lifestyle.

Earwigs, order Dermaptera, are a distinctive group of insects that have a pair of forceps-like appendages at the posterior apex of the abdomen. Wings are absent in some species, and when present are modified, such that the forewings are shortened and hardened and the hingwings are extensively folded, exposing most of the abdomen. Most species are omnivorous, predatory, or herbivorous and are nocturnal. Usually they are encountered when searching through leaf litter, under rocks, in rotten logs, or in any other hidden crevice.

A much larger group is the Order Orthoptera, the familiar grasshoppers, crickets, and katydids. Most of these have notably enlarged hind legs, used to jump great distances or launch the insects in flight to escape predators. Many katydids are masters of camouflage, with green, leaf-like wings bearing markings that resemble leaf veins, fungal infections, and even insect feeding damage. Other orthopterans are equally cryptically colored, a testimony to the relentless selective pressure exerted by sharp-eyed bird predators. Some grasshoppers and katydids are strong fliers, whereas crickets are usually much more likely to stay on the ground. Most male Orthoptera produce sound to attract mates, and the calls of katydids and crickets are an integral part of the evening chorus in warmer parts (or seasons) of the world. Plants are the food of most species, although there are some predators as well.

Like the Orthoptera, the walkingsticks (Order Phas-
leaves and becomes a major source of food for other
insects such as flies and wasps. Some are attended by
ants, who harvest honeydew and protect the homopter-
as from predators and parasitoids.

The thrips, Order Thysanoptera, are another unusual
order of insects. They are small, slender-bodied, with
or without slender, fringed wings. The last tarsal seg-
mant of the legs have an inflatable bladder used to
improve the grip on the substrate. The mouthparts are
asymmetrical, of the piercing-sucking form, used to
feed on debris, fungi, or plants. Thysanopterans also
have an unusual form of development in which there
are two or three pupa-like stages, similar to the pupal
stage of holometabolus insects, but which were inde-
pendently evolved. At least one tropical species has
developed a degree of sociability.

The major group of insects, accounting for about
80% of the species, is the Endopterygota, or Holometa-
bolan, two names that refer to different aspects of these
insects. Endopterygota refers to the development of the
wings from imaginal disks of tissue within the pupa of
the immature insect, in contrast to the gradual, external
development in less derived insects. Holometabolans re-
fers to the development of individual insects through
a complete metamorphosis, through egg, larva, pupa,
and adult stages. The larva is transformed to an often
strikingly different adult through the pupal stage, in
which the body tissues are almost completely broken
down and then reorganized into adult structures. Imagin-
al disks, sections of tissue sequenced and formed in
the larval stage, are retained in the pupa and are the
basis of many of the adult structures, such as gonads,
legs, and wings. The almost complete separation of the
structure of larval and adult insects allows the holome-
tabolans insects to be extremely specialized in each stage
and adult stages. The larva is transformed to an often
strikingly different adult through the pupal stage, in
which the body tissues are almost completely broken
down and then reorganized into adult structures. Imagin-
al disks, sections of tissue sequenced and formed in
the larval stage, are retained in the pupa and are the
basis of many of the adult structures, such as gonads,
legs, and wings. The almost complete separation of the
structure of larval and adult insects allows the holome-
tabolans insects to be extremely specialized in each stage:
the larva for feeding and the adult for mating and dis-
persal. This specialization increases the efficiency of
each stage and probably results in the incredible success
of holometabolans insects.

Within the Endopterygota, there are four orders that
are immensely large and successful: the Coleoptera,
Diptera, Lepidoptera, and Hymenoptera. Any one of
these megadiversity groups alone is among the largest
assemblages of organisms in existence, outnumbering
any other group except plants. Together, these four
insect orders constitute about 40% of all described spe-
cies of life on earth.

The most successful order of insects in terms of
species number is the Coleoptera or beetles. They are
the largest group of organisms in the world, even out-
numbering plants. The mesothoracic (first) pair of
wings of the beetles are greatly strengthened and hard-
ened, such that they are of little or no use in flight

outside of the Endopterygota (discussed later) there
is only one truly large order of insects: the Hemiptera.
This assemblage, in the broad sense, includes groups
sometimes named the Heteroptera, or true bugs, and the
Homoptera, the cicadas, leaf-, tree-, and planthoppers,
aphids, whiteflies, scales, and others. All have distinct-
tively modified mouthparts that are in the form of a
piercing-sucking beak that they use to obtain food.
Predatory species pierce their prey, usually other in-
serts, injecting digestive enzymes to kill and begin the
process of digestion. Predation is restricted to some
heteropterans; a few also feed on vertebrate blood. Most
Hemiptera are plant feeders, including the majority of
the heteropterans and all the homopterans. Most species
are terrestrial, although some heteropterans are aquatic.
Some homopterans, especially scales and aphids, pro-
duce honeydew, as sweet excretion drips down onto
leaves and becomes a major source of food for other
is in a state of flux, with molecular data giving some
ashes to Diptera (flies). The relationships among these orders
includes the Mecoptera (scorpionflies), Siphonaptera
other soft-bodied insects found on foliage. Lacewing larvae are predatory on aphids and
larva. Lacewing larvae are predatory on aphids and
immatures. The immatures of antlions dig conical pits
adults are known to be predatory, as are most of the
bearing an elaborately reticulate wing venation. Some
owlflies. All are relatively soft-bodied, with large wings
into which prey fall and are grabbed by the waiting
madrona, giving them a snakelike appearance. Neuropter-
on the ground in rotting wood or leaf litter and feeding
on smaller insects. Adults have a greatly elongated pro-
thorax, giving them a snake-like appearance. Neuropt-
ants are more diverse, with a number of elaborate or
bizarre forms intermingled with the more normal-look-
ing lacewings and the damselfly-like antlions and
larvae, which can be predators, scavengers, herbivores,
larval chironomid midges. Some species have elaborate
courts of insects involving nuptial gifts of dead in-
ssects to the female.
Fleas are highly specialized, laterally flattened, wing-
less blood feeders on birds and mammals. Usually they
are found in the nests or other areas frequented by their
hosts. The larvae are usually free living, feeding on
organic detritus and blood from droppings of the adults,
although some are obligate ectoparasites. Adults are
extremely laterally flattened, allowing them to travel
smoothly between the hairs or feathers of their hosts,
and often have backwards-pointing combs of spines
that makes them difficult to remove. They have strongly
developed jumping legs that allow them leap to and
from their hosts.
Siphonaptera are extremely unusual parasitoids of
other insects. The adults are highly sexually dimor-
phic, with the males being free-living, winged insects,
whereas the females in all but one family are endopara-
sitoids—wingless, legless, and without only vestigial
eyes and appendages of the head. The female body ex-
trudes from the body wall of the host, emitting phero-
mones to attract males that copulate with the special
openings (external genitalia being absent). The larvae
are hypermetamorphic, with active, host-seeking first
instars that make them difficult to remove. They have strongly
developed jumping legs that allow them leap to and
from their hosts.

Robber flies are found near everywhere. Their distinctive
innovation has been the reduction of the metathoracic
wings to a pair of knoblike halteres that act as gyro-
scopes in flight. This modification has increased their
maneuverability and allowed Diptera to become unpar-
alleled masters of aero locomotion. Although some
adult flies require massive protein meals to mature
eggs and power flight, most of the feeding is done by
larvae, which can be predators, scavengers, herbivores,
parasitoids and even true parasites. Free-living Diptera larvae are found in soil, rotting vegetation, feeding on and in plants, and sometimes exposed on vegetation. Aquatic forms are found in the silt or sand underlying the body of water (sometimes interstitially), on the surface of rocks, logs, or vegetation, or in the water column. Parasitoids attack mostly other arthropods, but some endoparasites attack mammals.

The orders Lepidoptera (butterflies and moths) and Trichoptera (caddisflies) are close relatives. The lesser of the two—the caddisflies—have aquatic larvae that are found in nearly every type of freshwater environment. Most construct cases or shelters from plant particles, twigs, stones, or sand grains tied together with silk. Some also construct nets to capture debris for food, whereas others are predatory, attacking other aquatic insects. The adults are slender, mothlike insects, often with long, thin antennae.

The Lepidoptera are among the best-known insects, especially the colorful, diurnal group called butterflies. Most of the diversity of the group, however, is in the nocturnal, often drably colored moths, which constitute about 80% of the species of Lepidoptera. The larvae are usually called caterpillars and are best known as voracious feeders on plants. Larval feeding takes place on the surface of the plant or within it (as in leaf miners and stem borers), and almost every plant part—leaves, stems, roots, flowers, and seeds—can be affected. Some species are also predatory and some feed on animal material, such as wool, but almost all species are phytophages. Adults of most families have mouthparts that are modified to form a long, coiled tube that is used for taking up liquids, usually nectar from flowers. Many adult and larval Lepidoptera are cryptically colored to avoid detection by predators, whereas others are brightly colored to warn of toxic chemicals sequestered in their bodies. Some adults engage in long-distance migrations.

The final large group is the Hymenoptera, the sawflies, ants, bees, and wasps. Like the Diptera, the Hymenoptera have become extremely adept fliers, but have done so by joining the fore- and hindwings with a row of tiny hooks (hamuli) to produce a single functional pair of wings. The most primitive families are phytophagous, but from within these groups have arisen a great diversity of parasitoids; predators, and plant feeders. The parasitoids include some of the smallest known insects, which attack the eggs of their much larger hosts (such as Lepidoptera). Other parasitoids attack a diversity of immature insects, especially those of other holometabolous groups, and develop as endo- or ectoparasitoids. Some are obligatory hyperparasitoids—parasitoids of parasitoids; others oviposit in plant tissue and induce the formation of plant galls, in which the larvae feed. The predatory Hymenoptera attack a wide range of hosts, especially other arthropods, which they often subdue but do not kill with a venomous sting. Larvae of these species have a supply of fresh food to consume when they hatch from an egg laid on the paralyzed prey. Most species hide their prey in some sort of burrow or nest to prevent its being taken by other insects or scavenging animals. Some of these provisioning wasps have moved on to pollen and nectar as food, as in bees. Sociality has evolved a number of times in the Hymenoptera, with the largest and most complex colonies formed by ants and bees.

III. CONCLUSION

As this brief survey shows, the largest, most diverse and arguably most successful groups of insects are those that have undergone three major innovations: (a) the development of wings, (b) the ability to fold wings over the body, and (c) the division of the life history into four major stages. Further outstanding success was found in those species that had hardened forewings (beetles), those that transferred the responsibility of flight to a single pair of wings (Diptera), those that extensively exploited flowering plants (Lepidoptera), and those that combined a single functional wing complex with a number of other traits (such as haplodiploidy) that produced the ants, bees, and wasps (Hymenoptera).

See Also the Following Articles

BIBLIOGRAPHY


INTERTIDAL ECOSYSTEMS

A. J. Underwood and M. G. Chapman
Centre for Research on Ecological Impacts of Coastal Cities, University of Sydney

I. Introduction
II. Processes Causing Patchiness on Rocky Shores
III. Boulder Fields
IV. Mangrove Forests
V. Sandy Beaches and Mudflats
VI. Measurement of Biodiversity in Coastal Habitats
VII. Can Intertidal Biodiversity Indicate Ecological Function?
VIII. What Services Does Intertidal Biodiversity Provide?
IX. Conclusions

GLOSSARY

assemblage  The collection of animals and plants found together in a patch of habitat.
diversity index  A numerical measure combining information about the number of species present in a sample or habitat and information about their relative abundances.
ecological functions  Attributes or properties of assemblages or habitats that are dependent on the biodiversity of animals and plants present. Examples are production of nitrogen, sequestering of heavy metals, sustained production of harvested resources, maintenance of a diverse food web, and so on.
ecological processes  Direct or indirect interactions between species—such as grazing, predation, competition, responses to disturbance, and so on—that cause spatial or temporal patterns in distributions and abundances of species.
patchiness  Variation from place to place (or time to time) in the abundances of animals or plants, caused by the interaction of numerous processes.
recruitment  Arrival of new juvenile animals or plants into a habitat or an older stage of the population. For many marine animals, recruitment occurs some time after the settlement of planktonic larvae in the adults' habitat.

INTERTIDAL AREAS such as rocky shores, mangrove forests, sandy beaches, and mudflats are habitats for very diverse assemblages of plants and animals. These species are usually distributed patchily in space and time because of several important ecological interactions. Local biodiversity is therefore patchy and variable. Measurement of biodiversity is complicated by variability. The functions provided by biodiversity are not well understood in most intertidal habitats. Planning for conservation, management, and restoration of biodiversity is difficult because of variability in space and time.

I. INTRODUCTION

Intertidal habitats range from rocky beaches or platforms to sandy and muddy shores. They occur under
a wide range of physical stresses, most notably due to emersion when the tide falls and disturbances due to waves. The animals and plants must endure the full range of physical variation, from being covered by water when the tide is in, to being potentially exposed to typical terrestrial conditions when the tide is out. In addition, the amplitude or range of tidal rise and fall can vary between two tides in a day (this is called "semidiurnal inequality") and, everywhere, varies in a fortnightly cycle between spring and neap tides. Spring tides rise much higher and fall much lower than do neap tides. As a consequence, during neap tides, animals and plants at high levels on a shore are not reached by an incoming tide for several days in a row and must endure long periods of inactivity or reduced activity until spring tides start again. At these high levels, moisture from spray (particularly where wave action is great) and during rain is very important for the supply of food and for preventing marine animals from drying out during their long periods of emersion.

In complete contrast, animals and plants at lower levels on the shore are submerged continuously for several days during neap tides because the tides do not fall low enough to expose such areas to aerial conditions. As a result, the animals and plants are, effectively, in a fully marine habitat and can be subject to predation by crabs, fish, starfish, and so on, which are mostly subtidal species. Such predators can forage at will over the lower parts of a shore during neap tides. This variation in periods of emersion versus submersion causes a very stark gradient of physical conditions from the top to the bottom of the shore.

A second strong physical gradient often exists along a shore, due to variation in wave action. As any shore bends away from being fully exposed to the open ocean, the strength of waves hitting the rocks will vary. As a result, there is a general decrease, with decreasing force of waves, in the amount of splash and spray over any area. The two physical forces—the direct force of waves and the ameliorating effects of splash and spray reducing the influences of desiccation during low tide—influence the abundances of animals and plants and the nature of their activities. Thus, grazing and predation are less intense where waves are strong. The diversity of species changes with wave action because of the combined influences of waves themselves, their ameliorating influence on desiccation, and the indirect influence of both physical features on the interactions among species.

These two major processes—tidal rise and fall and wave action—are so well known that their influences were long considered the major, if not the only, influences on local ecological processes and therefore diversity in intertidal habitats. The general features of intertidal regions, particularly the general patterns of distribution and the numbers of species on rocky shores, have for a long time been described in terms of these physical forces. Summaries and details are available in Lewis (1964) and Stephenson and Stephenson (1972) and other reviews of particular biogeographical regions or comparisons of various parts of the world.

More modern syntheses, in contrast, have focused on the ecological processes that have more profound effects in terms of creating and maintaining considerable patchiness in the ways animals and plants are distributed in intertidal habitats. The variation often is considerably more conspicuous than the features considered general and easily explained in terms of gradients of physical forces. This forms the central theme of this contribution—the processes that influence patterns of distribution and abundance of species and therefore the local diversity of species in intertidal habitats. The major focus is on rocky intertidal habitats, because most of the experimental work in the past 40 years has been on this type of habitat. The animals and plants are typically relatively short lived (although individuals can live for extraordinarily long periods). They are usually small and abundant, and most of them are slow moving or sessile. As a result, experimental manipulations to test specific hypotheses from existing models that might explain patterns can be done without enormous areas of habitat being disturbed and without excessive costs of logistics and infrastructure. There has been a mass of such work and it has revealed and continues to reveal a great deal about the life histories, interactions, and the general and specific ecology of the maintenance of diversity in intertidal habitats (see review by Underwood, 2000).

In addition to rocky shores, there is also some discussion of other intertidal habitats (sandy beaches, mangrove forests), even though these are discussed in detail elsewhere in this encyclopaedia. Their inclusion here is to demonstrate that the sorts of variability in time and space found on hard, rocky surfaces are widespread and general and not confined to rocky shores.

Following an overview of ecological patchiness and some consideration of its causes in intertidal habitats, three consequences of variability in biodiversity are introduced: how to measure and compare diversity from one place to another, how descriptions of biodiversity may relate to functional aspects of diversity, and the sorts of functions that biodiversity may provide in coastal habitats.
II. PROCESSES CAUSING PATCHINESS ON ROCKY SHORES

Numerous ecological processes influence or maintain patterns of local biodiversity in coastal habitats. These include direct and indirect biological processes of recruitment, predation, grazing, and competition. They also include physical disturbances and biological responses to them. These different types of processes are illustrated here.

A. Recruitment

Variability in recruitment of coastal marine organisms is well described (Connell, 1985; Thorson, 1950). Many of the animals and plants have a relatively long dispersive phase in their life history, resulting in a large proportion of propagules being killed by predators, failing to develop or not being able to find a suitable habitat to settle and metamorphose. Inevitably, the processes of production and loss of propagules lead to great variation in numbers surviving to settle in the adults' habitat and the number actually settling in any area of habitat at any particular time.

There are two major consequences to issues of biodiversity of this variation in recruitment to adult populations. First, vagaries of types and numbers of species arriving in a patch will ensure that the diversity of species continues to be patchy at small spatial scales. Second, across the ranges over which species disperse, there will be little large-scale variation in diversity. Thus, along a stretch of coastline over which organisms disperse, the assemblages that develop will, in general, receive recruits from similar mixtures of species. At a local scale of patches of coastline in any area on the coast, there will be greater variation in the composition of assemblages.

At larger scales, speciation and changes in diversity from one side of an ocean to another are very much influenced by recruitment of larvae. Thus, the vast stretches of open, oceanic water in the middle of the Pacific and the Atlantic oceans form barriers to the spread or continuity of species. Scheltema (1971) sampled planktonic stages of life history of various gastropod mollusks in samples from the middle of the Atlantic. He found a very good correlation between the number of larvae of a given species and the distribution of adults. Species found on coastlines on both sides of the Atlantic were regularly found as larvae in his samples. Larvae that were found less commonly were from adults that had closely related or subspecies on the two coasts. This implied that lack of frequent interchange of larvae was causing differentiation and speciation. Species found on only one side of the Atlantic were rare or absent as larvae in the sample. Thus, where larvae cannot arrive, diversity will differ.

The timing of recruitment by different species can also affect local diversity of species. Thorson (1950) pioneered the study of influences of when recruits arrive to their final habitats. He described the predatory brittle stars (Amphiura spp.), which did not feed for 2 months while breeding. Several species of mollusks with short larval development bred during this period. The brittle stars had a longer period of development. As a consequence, the young mollusks could be released into the plankton, complete their development, recruit, and grow to a size where at least some of them would escape from the voracious predatory recruits of the brittle stars. If the mollusks recruited later, they were unable to avoid their predators. This anecdote needs to be examined experimentally but indicates that the timing and order of recruitment can have profound influences on which species survive in an area.

B. Competition

A second major process that can influence local diversity is competition—negative interactions among individuals of the same or different species caused by absolute or relative shortage of resources needed by all of the individuals. In most coastal habitats, resources are food, space, and, for plants, light. Space is needed to settle and become established (for sessile species) and as an indirect resource for food (for example, by mobile grazers, which need relatively open space over which to feed). Alternatively, sufficient space on the substratum is needed to allow access to the water column for filter feeders.

Generally, competitive interactions are considered to lead to reduction in abundances or actual elimination of competitively inferior species. Numerous examples demonstrate this, starting with Connell's (1961) classic experimental demonstration of destruction of one species of barnacle (Chthamalus stellatus) by another, faster-growing species (Balanus balanoides, using the taxonomy as in Connell). As a result of overgrowth by B. balanoides, C. stellatus could be eliminated from some lower areas of their vertical distribution on a shore.

Such competitive overgrowth can have dramatic effects on local richness of species. For example, Paine (1974) has described the major decrease in number of species living on primary space—the underlying substratum—where mussels are able to overgrow and even-
tually eliminate most other species (limpets, chitons, algae, barnacles).

These examples do, however, paint a biased picture. For example, where mussels overgrow a primary surface, their shells provide new hard substrata on which the inferior competitors can become established. They also serve as environmental engineers (Lawton, 1994), because the packed shells of mussels trap sediments. While feeding on particles in the water, mussels selectively ingest some sizes and types of particles. The rest are wrapped in mucus and eliminated as pseudo-feces, which add substantial substance and nutrients to the surrounding sediments in the matrix of mussels. The result of all these processes is provision of habitat and food for a very substantial number of species, such that the number of species in mussel beds is very large (Lohse, 1993).

There are also complications due to indirect interactions (reviewed by Wootten, 1994) where the outcome of competition between two (or more) species (say, A and B) influences the outcome of competitive interactions between one of them and yet other species (say, B and C).

For example, Kasteniek (1982) described the interesting case of the turfing red alga, Halidrys dioica, which outcompetes the alga, Pterocladia capillacea, by growing over and denying it access to light (Fig. 1). Where there is a canopy of a third species, Eisenia arborea, however, the inferior competitor survives well under the canopy and is able to "resist" competition from H. dioica. Thus, competition for light between E. arborea and H. dioica diminishes the latter's capacity to grow well and to take over space that is occupied by inferior competitors, such as P. capillacea. As a result, competition by E. arborea over H. dioica prevents competitive overgrowth of P. capillacea by H. dioica.

C. Predation

A major process influencing diversity of species on rocky shores is predation. Predation has three distinct influences. First is simply that where predators are sufficiently numerous or active, they may eliminate some species from areas of habitat. Thus, toward the bottom of rocky shores on the northwest coast of the United States, Connell (1970) described predation by several species of the whelk, Nucella (or Thais) feeding on barnacles. Low on the shore, the whelks had sufficient opportunity (due to prolonged submersion by the tides) to eat all of the barnacles, thus eliminating them from such low areas and thereby reducing the diversity of sessile species there. At higher levels, in contrast, the whelks were less active because they had a smaller period submerged during low tide in which they could find an item of prey and consume it.

The second influence of predation is indirect. The whelk Morula marginalis in southeast Australia consumes a variety of prey, including barnacles and small limpets, Patelloida latistrigata. It takes the whelks considerably longer to drill a hole through the shell of a
barnacle than to eat a limpet; *M. marginalba* can consume *P. laevisgrata* without having to drill a hole through their shells. Fairweather (1983) removed limpets in some areas as they arrived. As a result, predation of small barnacles by whelks increased. This demonstrated an indirect effect of preferences and differences in the time taken to consume prey; the least preferred species (the limpets) took less time to eat than the more highly preferred barnacles. Where there were no limpets, the whelks turned their entire attention to small barnacles, with consequent increased mortality and decreased abundances of the barnacles. Note also that whether or not limpets arrived from the plankton also influenced the eventual outcome in terms of abundances of barnacles and, therefore, the local relative abundances and, thus, diversity of species.

The final major influence of predation is also indirect. Large generalist predators are capable of consuming most other species in any patch of habitat. As a result, they can have direct influences on the abundances of prey and can free space on the shore for species to be able to recruit. Space is often in short supply, causing the various sessile species to compete with each other for space on which to live and causing grazing species to compete for space over which to feed. If the generalist predator disproportionately consumes “winners” of competitive struggles for space, predation will have very important effects on the local diversity of species. The original example of this phenomenon, known as “keystone predation,” was investigated by Paine (1974). The large starfish, *Pisaster ochraceus*, eats most of the other sessile and mobile animals on rocky shores on the west coast of the United States. In some areas, the starfish preferentially eat mussels (*Mytilus* spp.). Mussels are, however, among the best competitors for space, being able to smother and grow over other species (see earlier). Predatory starfish continuously remove mussels, making space available for inferior competitive species that would otherwise be locally eliminated by overgrowth due to the mussels. In this case, the predators exert an indirect positive influence on a range of species of prey, resulting in there being greater numbers of species (i.e., greater species diversity) where there are predators than where there is none.

### D. Physical Disturbances

Given that physical factors such as desiccation and wave action can have large and important influences on abundances and diversity of species in intertidal habitats, it is not surprising that physical disturbances can influence diversity. Where disturbances are large or frequent, it is likely that more delicate species that are unable to withstand the physical forces will be unable to survive. They are either prevented from settling and becoming established in the first place or are unable to complete their life cycles before a large disturbance kills them. It is expected under this model that there would be reduced diversity in more physically stressed habitats simply because of the direct loss of species due to disturbances. In contrast, where disturbances are small or rare, most, if not all, species can survive the rigors of the environment. Resources of space and food will then become critical and some species will disappear because of superior competitive abilities of other species. Under these circumstances, diversity in physically benign areas will also be reduced below what is theoretically possible.

As a consequence of these two processes, it is anticipated that the greatest diversity should be found in areas of intermediate disturbance, such that competition for resources is not too intense because there is sufficient disturbance to prevent all of the possible species from building up excessive abundances. At the same time, physical stress is not so great that some species are actually eliminated. This is the model of intermediate disturbance (Connell, 1978).

From this model, it has been predicted that in areas where physical disturbances can be increased or decreased in intensity or frequency, there will be predictable changes in diversity. For example, where disturbances are small, increasing their severity will cause increases in diversity as superior competitors are prevented from building up sufficient numbers to consume all of the resources needed by other species. If disturbances are increased even more, there will be a decrease in diversity.

Such predictions have been tested in several studies on rocky shores, with mixed success. Sousa (1980) identified patterns of diversity that were fairly consistent with the model, but the mechanisms operating were not as stated. Species were not lost by superior competition for space from other species. Rather, the life histories of the different species and the time taken to be able to recolonize disturbed areas were of greater importance. McGuinness (1987) only identified the intermediate disturbance model as being an appropriate explanation for observed patterns of diversity in a few of the cases he examined in intertidal boulder fields. Clearly, more work is needed to understand how and when the interactions of competition and disturbance lead to predictable outcomes in terms of diversity of species.
III. BOULDER FIELDS

Intertidal boulder fields form a very interesting habitat that forms part of rocky intertidal environments. Boulders have top surfaces that are exposed to waves and sunlight and that may be emersed during low tide. Underneath boulders, however, there is a very different sort of habitat—usually cool, moist, and dark. On shores affected by large waves, boulders are frequently overturned and moved. This alternately exposes each surface of the boulder to the light and causes the boulders to abrade against each other, damaging and killing many animals and plants living on them. These boulders support little biodiversity, mainly encrusting species that are resistant to harsh conditions or ephemeral species that rapidly colonize new habitat when it becomes available.

Boulders on sheltered shores, in contrast, are stable and support great diversity of animal and plant life. Different species are found on the tops of or underneath these boulders. Those on top are similar to those living on nearby rocky shores—after all, they are subjected to similar environmental conditions. Therefore, boulders may be covered with leafy and encrusting seaweeds, have patches of apparently bare rock, or support large numbers of snails, limpets, and other common intertidal animals.

The species living under boulders are completely different. There are few leafy seaweeds, but there may be many kinds of encrusting seaweeds, which seem to thrive in dark conditions. There can also be patches of attached animals—plate-like bryozoans, jelly-like masses of colonial ascidians or sea-squirts, encrusting sponges, and the calcareous and sandy tubes of different types of worms. These sorts of sessile animals are not common on rocky shores, except in rock pools or crevices or under ledges (Fig. 2).

Many of the mobile animals living under boulders are rather specific to boulders—they tend to be rare or absent from other intertidal habitats. Many show very specific patterns of behavior that allow them to move quickly from one boulder to another if the boulders are overturned or moved into unsuitable conditions by large waves. Therefore, although chitons on rocky shores tend to be very slow-moving and inactive animals, those under boulders move extremely rapidly over the rock surface or curl up and drop from the boulder as soon as it is overturned. This behavior returns them to the undersurface of the same or nearby boulders.

Intertidal boulder fields are potentially important habitats for the conservation of biodiversity. They often form rather small patches of habitat—especially when compared to sandy beaches, mangrove forests, or rocky shores, which are often continuous for many kilome-

FIGURE 2  The undersurfaces of intertidal boulders support numerous species of sessile and mobile animals, many of which are not found in other intertidal habitats. These include encrusting sponges, ascidians, and bryozoans, numerous chitons and snails, and starfish, sea urchins, and brittle stars.
ters. They are also often scattered along the coast, with large stretches of other habitats between them. There are no data about how much interchange there is among populations in different boulder fields nor how consistently species would recruit from other boulder fields if locally eliminated. Because many species living in boulder fields are relatively confined to these habitats, they are likely to be vulnerable to human activities that destroy or damage their surroundings.

Finally, many species are extremely overdispersed. This means that each species tends to be crowded onto relatively few of the available boulders, while most boulders are unoccupied by that species. Because different species are often found on different boulders, patterns of biodiversity are very complex—both within and among different boulder fields.

As yet, the processes that lead to these complex patterns are not well understood but include such factors as the size and shape of the boulder, what stone it is composed of, where it is (e.g., depth of water, what substratum is beneath it), and the different wave conditions where the boulder field is located. In addition, there are complex interactions among the animals and plants themselves (Sousa, 1980), causing complex changes to biodiversity depending on which species happen, by chance, to recruit to which boulders.

IV. MANGROVE FORESTS

In contrast to intertidal rocky shores and boulder fields, mangrove forests are dominated by trees. Unlike some other habitats containing many large plants—for example, subtidal kelp beds (Foster and Schiel, 1985) and terrestrial rain forests (Terborgh, 1992)—patterns of diversity in mangrove forests (i.e., variation in the numbers and types of animals and plants found from place to place or time to time) are not well documented (Hutchings and Saenger, 1987). This is particularly true for any measures of small-scale patchiness in diversity within individual forests. This contrasts markedly with other intertidal habitats in which patchiness of animals and seaweeds is relatively well described (see the discussion on rocky shores presented earlier).

This lack of good quantitative measurements of biodiversity in mangrove forests is due to many different factors. First, the plants that make up the forests are, themselves, not very diverse. Typically, they do not develop a complex structure of many species of canopy, understory, and ground cover, which could, in turn, support many different types of animals. Second, because mangrove forests normally grow on very sheltered shores, the vertical gradients of environmental conditions associated with tidal height are very strong (Hutchings and Saenger, 1987). There are also strong latitudinal gradients, causing very large differences between temperate and tropical zones in the types and diversity of trees that make up the forests. These are well documented elsewhere and are not discussed further here. Within any single mangrove forest, however, different species of trees typically live at different levels along the intertidal gradient, so that in any one place, even fewer types of plants are found.

The main factor that has led to poor description of the distribution and patchiness of biodiversity in mangrove forests is, however, that most diversity is due to small invertebrate animals. This is, of course, true for many habitats where much of the biodiversity is “invisible.” In such habitats, appropriate sampling can clearly measure variation in diversity from place to place or time to time, but these sampling designs are inevitably complex and costly because they need to measure diversity at many different spatial and temporal scales.

The animals that make up most of the biodiversity in mangrove forests include many species of crustaceans, such as crabs and amphipods, small snails and bivalve mollusks, worms from many different phyla, insect larvae, and so on. Because these animals are generally very small and live on or under the surface of the mud, they are not readily visible without sampling small patches of mud, sieving the animals out of the sediment, and viewing them under magnification. The arduous work that this entails means that patterns in diversity are not well described for most mangrove forests.

In addition, because the types of plants found in mangrove forests often vary along the intertidal gradient, most studies of the distribution and diversity of animals have also described similar broad-scale patterns. Therefore, changes in the diversity of snails, crabs and other animals from the seaward to landward edges of the forest tend to dominate the literature. There have been few descriptions of the small-scale variability or patchiness in this diversity within a shore level, although this may be the most important source of variation in diversity, as is the case on rocky shores. Some studies that have been done to measure small-scale patchiness in diversity in mangrove forests show that much of the variability in diversity and abundances of these animals is at very small spatial scales (i.e., among patches of habitat centimeters or meters apart). It can also change unpredictably through time (Fig. 3).

This sort of patchiness of diversity is not limited to little invertebrates that live in the mud. Many of the larger animals that spend their juvenile stages in man-
grove forests, such as fish, prawns, and large crabs, show similar patterns. A recent survey of small fishes in patches of mangrove forests in Sydney Harbour showed most of the variation in abundance and diversity was found at the scale of meters (i.e., from one net to another set in a small patch of mangrove forest). Variation from site to site (hundreds of meters apart in the same bay) and from bay to bay (kilometers apart) together accounted for less variability.

In contrast to rocky shores, there is little information about processes that cause this small-scale patchiness in mangrove forests. This is due to the fact that, with few exceptions, mangrove forests have not been as well studied, particularly using well-designed, controlled field experiments. Typically, diversity of small animals living in sediments in any aquatic habitat is primarily determined by the range of grain sizes of the sediment itself. Larger species tend to be more common in coarse-grained sediments and vice versa. The range of sediment in any particular mangrove forest can vary from patch to patch because of such processes as changes to water currents around submerged objects (e.g., the bases of trees) or biological processes.

In New South Wales (Australia), the mangrove crab, Heloecus cordiformis creates mounds of coarser and drier sediments and flat areas of wetter and finer sediments when feeding and burrowing (Warren and Underwood, 1986). Although not yet measured, it is likely

FIGURE 3  This figure shows variability in the numbers and types of animals living in a mangrove forest in New South Wales, Australia. The animals were sampled at four sites along the bank of a river. Each site was sampled using four quadrats, placed 1 to 2 meters apart. Each point on this figure represents the types and numbers of animals in each quadrat. The relative distances between two points indicates how similar were animals in those two quadrats. Points close together represent quadrats containing very similar types of animals. Points far apart mean that the quadrats contained very different numbers and types of animals. Figure 3a shows that when the animals were sampled the first time, the diversity of animals was very variable from one quadrat to another. Quadrats in the same site (meters apart) showed similar amounts of variability in these measures of diversity as shown by quadrats in different sites. Figure 3b shows what the patterns looked like only three months later. Animals were more similar at the scale of meters (among quadrats in each site), but the sites appeared to represent a gradient along the river. This relatively large change in the pattern of diversity was not seasonal, predictable, nor found in other nearby sites. It was mainly due to changes in the relative numbers of numerous small crustaceans.
that this activity directly influences local patterns of diversity, as shown for the burrowing behavior of soldier crabs in Tasmania (Fig. 4) (Warwick et al., 1990). Other small-scale structures—buried shells, the roots of the mangrove plants, worm tubes—are equally likely to alter patterns of diversity, although the importance of these small structures and the behavior of other animals as influences on biodiversity are not yet well documented for mangrove forests.

V. SANDY BEACHES AND MUDFLATS

Sandy beaches and mudflats tend to grade into each other along a continuum depending on the relative grain size of the sediment. Sandy beaches have rather coarse-grained sediments and mudflats are, of course, muddy, with fine-grained sediments. Both can occur on very sheltered shores, such as in estuaries, but mudflats do not occur on wave-exposed shores, where sandy beaches may be common (Brown and McLachlan, 1990; Reise, 1985).

Sandy beaches and mudflats have very few large plants, although heaps of decaying plant material, commonly known as beach wrack, can be found on the strandline. This material—decaying seagrasses, seaweeds, and kelps that have been washed up by the waves—is rapidly decomposed by bacteria, fungi, and small animals, many of which are dependent on wrack for food. This activity ultimately releases nutrients back into the coastal waters (Griffiths et al., 1983).

Sandy beaches and mudflats appear at first glance to resemble deserts—there are often few animals to be seen, other than birds and a few crabs. The birds are generally visitors to these shores, feeding along the shoreline or over the extensive flats during low tide, but spending the rest of their time elsewhere. There are relatively few animals high on the shore, mainly semiterrestrial species, such as ghost crabs which spend the day in burrows above the level of high tide, coming out at night to feed. Nevertheless, despite first appearances, there is diverse life lower on the shore. As is the case in mangrove forests, however, most of the animals live under the sediments.

Really small animals live in the tiny spaces among the grains themselves. These include unicellular and multicellular animals from many different phyla. Many of these very small animals resemble juveniles of larger species. Many are also extremely simple. For example, although multicellular, they may only consist of a few cells and lack structures—such as limbs, kidneys, or a circulatory system—relying on exchange of substances through the body wall to acquire oxygen or get rid of waste (Brown and McLachlan, 1990).

The larger animals living under the sand or mud either construct and maintain complex tubes or burrows, or simply move through the sediment without leaving any structure behind them. On sandy beaches exposed to relatively large waves, the sediment is continually reworked by wave action. Most animals living in these habitats do not construct tubes or burrows because these are likely to be destroyed. Instead, they remain relatively deep in the sediment or tend to have mechanisms whereby they can rapidly burrow back under the sediment if exposed. Mole crabs have paddle-shaped limbs and can burrow extremely quickly during the time available between waves. Similarly, many bivalves have an extendible, mobile foot, which rapidly digs and anchors itself into the sediment and then contracts, pulling the bivalve under the sediment in only a few seconds.

Because mudflats tend to be very extensive and not steeply sloping, they usually remain quite damp during low tide. Many different species of worms, crustaceans, mollusks, and so on can be found in these habitats. Although some simply burrow through the sediment, feeding as they burrow on detritus and particles attached to the grains, others live in semipermanent tubes or burrows. These can be very complex. Some burrows of crabs have many entrances and exits and are connected together in a complex underground labyrinth. Some animals living in burrows emerge to feed on the surface of the mud itself during high tide or at night. Other species remain in the burrow, but they may have long tentacles that are spread over the surface of the mud to pick up particles of food. Animals inhabiting burrows that feed on particles in the water often
have modified limbs or other appendages, which can create currents of water through the tubes, pulling in the food with the water current. They also have very elaborate filters, which sieve the particles of food out of the water. These filters are modified legs, mouthparts, gills, or other organs.

Despite the fact that sandy beaches and mudflats are very common intertidal habitats, their patterns of biodiversity are not particularly well known. Available data suggest that, like most other intertidal habitats, they are extremely patchy and variable through time. The numbers and types of animals differ from one patch to another only a few meters away. These patterns are caused by the different species responding to subtle changes in the sediments themselves, to local disturbances such as feeding by fish which creates depressions in the mud and to other animals living in the sediment via interspecific processes such as competition or predation.

Some species have rather aggressive behavior that spreads them out, presumably to ensure adequate food for all. Other species tend to aggregate and can create large patches of tubes or burrows, altering the sediments around them. These sorts of processes inevitably change small-scale local patterns of diversity, but there is very little information on how such processes may operate. Of course, as described earlier, such processes modify the initial patterns of diversity, which are established by differential recruitment of species among different patches of habitat.

VI. MEASUREMENT OF BIODIVERSITY IN COASTAL HABITATS

Methods for and interpretations of measures of biodiversity in coastal habitats have proven complex. As reviewed by Gray (2000), the simplest measure — the number of species in any area — does not describe the structure and variation in assemblages. For this, measures are needed that can collate information on the types of species present (the composition of an assemblage) with information about their relative abundances. Many indices have been used (see Magurran, 1988). One of the most widely used is the Shannon-Wiener index:

$$H' = -\sum_{i=1}^{S} p_i \log_2 p_i$$

where $p_i$ is the proportion of species $i$ in the whole sample. Following Whittaker (1972), Magurran (1988) described measurements of diversity in a single sample as "point diversity" — that in a set of replicate samples from the same habitat as "alpha diversity." Comparisons from habitat to habitat, for example, along an environmental gradient uses measures of "beta (or between habitat)" diversity. This is the combined diversity across a set of alpha diversities. There are also larger-scale measures, such as "gamma diversity," which are measured across larger spatial areas, such as a whole coastline.

There are serious problems of scale in these considerations. For example, a rocky headland could be a relatively large area, so diversity measured across it would be gamma diversity. Within the area, there are different habitats such as rock pools, algal beds, boulder fields, and so on. In each of these, samples would give measures of alpha diversity. Beta diversity would measure the diversity in, say, algal beds along a gradient of increasing exposure to waves.

Suppose, instead, that the study was a series of headlands along the biogeographical gradient from south to north along a coastline. It is now likely that the diversity measured in a set of randomly taken replicates from one headland would be considered to be alpha diversity. That from the set of headlands would be beta diversity. So whether or not a sample allows measurement of alpha or gamma diversity is entirely dependent on the spatial scale being examined! It is clearly crucial to have defined very carefully the hypotheses being tested so that only the relevant scales of habitats and sample areas are examined.

Another major difficulty in measures of diversity is that the number of species found is almost always dependent on the size of the sample examined or the length of time spent searching for individuals throughout the habitat. This is inevitable because of two distinct aspects of the patchy distributions of species. First, rare species are not going to occur very often in any sample and will often be missing from small samples. A lot of space needs to be covered to be sure that the rare species are actually encountered. Imagine a comparison of the animals in two habitats, which both have one hundred species in them. If in one habitat (A) many more of the species are rare than in the other (B) and similarly sized small samples are examined from each habitat, there will be apparently fewer species in habitat A than in B. This is, however, an artifact of sampling.

A second reason for there being inevitable differences in the numbers of species seen in two samples where one is much larger than the other is due to potential heterogeneity of habitats in the area studied. Suppose that the number of species is recorded in samples totaling 4 square meters of habitat in one area and 8 square
meters in the other habitat. The latter sample would normally be expected to contain more species if there are many subhabitats in the area sampled. So on rocky shores, clumps of seaweeds, small rock pools, patches of mussels, and so on may be scattered at scales of tens of centimeters. Each of these subhabitats probably contains several different species, in addition to "cosmopolitan" species found throughout the entire area. As a result, the larger area examined in samples from one habitat will encompass more different patches and therefore contain more species than would be found if a smaller area were examined.

There exist techniques to compensate for differences in the size of sample from one habitat to another. One of the most commonly used is called rarefaction, but this is known to have serious problems of overestimation of numbers of species in small samples (Fager, 1972). A superior alternative is to take random smaller samples from the largest, to match the size of the smaller samples from the other habitat. Thus random samples can be used to give reliable estimates and estimates of error associated with the number of species expected in smaller samples. Comparisons between the two habitats then consist of comparisons of the entire sample from the habitat that was sampled with smaller samples and the randomly determined estimates from the other habitat (see the review by Gray, 2000).

Some multivariate methods are widely used to compare the diversity of species from one area or habitat to another. These operate on the principle that the abundances or biomasses of every species can be recorded from a series of replicate sample units (cores, grabs, quadrats, etc.). This produces a matrix of species by replicates for a sample from each habitat or area. The differences between any one sample unit to another can then be calculated across all of the species. One popular measure of such differences is the Bray-Curtis dissimilarity index (Clarke, 1993). This is calculated as follows:

$$D_{jk} = \frac{\sum_{i=1}^{s} |X_{ij} - X_{ik}|}{\left( \sum_{i=1}^{s} X_{ij} + \sum_{i=1}^{s} X_{ik} \right) \times 100}$$

where $D_{jk}$ is the dissimilarity between sample units $j$ and $k$, $X_{ij}$ is the abundance of species $i$ in sample $j$, $X_{ik}$ is the abundance of species $i$ in sample $k$, and $s$ is the total number of species found in the two sample units. If $j$ and $k$ come from the same sample, this is a measure of variation among units within a sample. If $j$ and $k$ come from two different samples, $D$ is a measure of dissimilarity between the two samples. Because there are replicate measures, there is a sample of measures to estimate variation within each of the two samples and between the two samples. Statistical tests can then be done to determine the likelihood of there being more difference between two samples than expected by chance given the variability within the two samples (Clarke, 1993).

Such procedures have been used widely in examinations of the effects of environmental impacts and other ecological studies on diversity of animals in benthic habitats, where diversity is made up of the composition of species in any habitat and their relative abundances. They are limited in the complexity of sampling and experimental designs for which they can be used, so that natural variability at a variety of spatial scales and from time to time are difficult to examine using these procedures. There are, however, developments using much more intensive sampling to obtain large enough samples that independent measures of dissimilarity can be used in complex designs to test hypotheses about spatial and temporal scales of variability in entire diverse assemblages (Underwood and Chapman, 1998).

VII. CAN INTERTIDAL BIODIVERSITY INDICATE ECOLOGICAL FUNCTION?

Although it is clear that "biodiversity" covers many levels of variability, from genetic to population diversity within species, from diversity among species to diversity among habitats or ecosystems, the most common perception of biodiversity is still simply the numbers and types of animals and plants in different areas. Therefore, the importance of biodiversity has often focused on terrestrial habitats, the consensus of opinion being that there are more species on land than in the sea. Nevertheless, at different levels of organization, marine systems are more diverse—for example, there are more phyla in the sea than found on land (Gray, 1993). In addition, marine species appear to have more genetic diversity than related terrestrial and freshwater organisms. This suggests that they may be less vulnerable to processes that reduce genetic variation, although there is increasing concern that many nearshore aquacultural practices, including culture of fish and bivalve mollusks, are decreasing genetic variability of natural populations.

Although many small marine invertebrates are still to be identified and described, it is generally held that most diversity in the sea is of benthic organisms. There are many more species of animals and plants living on
or in close association with the seafloor, rocky reefs, and others, than in the water column itself. This is probably related to the diversity of different benthic habitats. There is still controversy about the relative importance of deep-sea habitats and coastal habitats to benthic diversity—controversy arising from the fact that these habitats are so poorly described that general patterns of biodiversity across large areas of marine habitat, even nearshore and intertidal habitats, are largely guesswork.

At whichever level biodiversity is examined, however, understanding patterns of biodiversity is essential to understanding the role(s) of diversity in the maintenance of different ecological functions. By functions, we mean such processes as maintaining diversity itself, recycling nutrients and chemicals among organisms and their surrounding habitat and maintaining appropriate habitat for adequate food, shelter, and so on. Because patterns of biodiversity are complex—biodiversity varies at many spatial scales from centimeters to hundreds of kilometers and changes predictably and unpredictably among days, weeks, seasons, years, and eons—the role of diversity in maintaining the functioning of individuals, populations, assemblages of species, and habitats will also be necessarily complex.

Relationships between biodiversity and the persistence of successful ecological functioning may be described by quite different theories. First, there may be considerable redundancy of species, so that many species may be lost without impairment of function because other species simply take on their roles. There obviously must be a limit to this—not all species can disappear and ecology persist. A second view is that all species may contribute to any ecological function, but the system may withstand loss of species until a crucial limit is reached, after which there will be rapid and inevitable degradation of function. The third view, called the idiosyncratic response, is most likely. This states that there will be loss of function with loss of species, but this is likely to be variable and unpredictable because of the complex patterns of abundance and interactions among the various component species.

Most of the experimental tests of relationships between biodiversity and ecological function have been done in terrestrial systems. The results of the different experiments are not clear-cut and there is considerable controversy about their designs and the ways in which they have been interpreted (Huston, 1997). Therefore, even in those systems that have been examined experimentally, there is no unambiguous relationship between biodiversity and any ecological functions. Because of fundamental differences in the ways that different processes operate between marine and terrestrial habitats, it is extremely unlikely that the results of such experiments, even if not controversial, could be applied to marine systems. Therefore, it is essential that the importance of biodiversity to the well-being and persistence of coastal habitats be tested experimentally in those habitats. Without such tests, it is not possible to predict what might be the results of any changes to diversity of these flora and fauna.

To date, the functional role of marine biodiversity has largely focused on off-shore processes. For example, people are concerned about diversity of plankton and its role in large-scale processes, such as commercial fisheries or atmospheric levels of CO₂ and O₂. In coastal habitats, most emphasis has been on the role of diversity in maintaining ecological function and has concentrated on single species or suites of similar species. Therefore, there is concern about the loss of large filter-feeding bivalves in estuaries and bays because the large numbers of these animals were considered to filter all of the water in the estuary in only a few days. In areas where populations have crashed, the water has become more turbid and is now occupied by different types of animals than was the case in the past. Similarly, certain large predators or grazers on intertidal shores or shallow reefs appear to have a pivotal role in maintaining patterns of biodiversity, by selectively removing competitively dominant sessile animals or plants with their associated species, thereby proving space for a host of other organisms that thrive in the altered conditions (see earlier discussion on keystone predators).

Ecological engineers (Lawton, 1994) have also received considerable attention because of their roles in altering the physical environment, creating habitat, and changing the availability of resources to other species. Many coastal animals are ecological engineers—the most obvious of which are corals which may develop into large reefs—a unique and very diverse habitat totally dependent on the corals themselves. Other lesser-known examples include burrowing bivalves and crabs in salt marshes, which can alter drainage, sedimentation, and erosion in addition to having direct impacts on other animals or plants which they may eat or for which they may be an important source of food. Some intertidal animals form habitat for a range of other species. Therefore, mussels of the genus, Mytilus, provide habitats for more than 300 other species of animals and plants on intertidal shores in Washington in the United States (Suchanek, 1992).
Mangrove trees stabilize sediments, modify shorelines, prevent erosion, and provide habitat for the juveniles of commercially exploited fish and shellfish. Many studies examining the relationship between structure and function in mangrove forests have concentrated on changes in the number of trees, especially in areas where they are cut down for firewood.

There has, however, been little consideration about the role of the real biodiversity of such habitats—the mud-dwelling bacteria and small animals—on the well-being of the trees themselves or, indeed, on the maintenance of the habitat itself. Unfortunately, as described earlier, little is known of this biota, other than the fact that it is variable, unpredictable, and complex. Because the species are poorly described and even more poorly understood, it is generally accepted that there is a lot of redundancy in the system. Therefore it has been assumed that, if the larger plants are looked after, the small invertebrate animals and the functioning of the habitat will look after itself.

The same is true of all intertidal habitats. Most ecological functions are thought of in terms of general, widespread, and large-scale processes (e.g., the cycling of nutrients throughout a mangrove forest or the maintenance of food webs across a rocky shore). Yet the fauna and flora that make up most of the diversity and that maintain these ecological functions are variable and patchy and change unpredictably through time, often at very small and localized scales. The scales at which ecological structure (the species) and ecological functions (many of the processes) are measured are often quite different. It is therefore difficult to relate the stochastic variability that we see in ecological structure with what is often perceived as the predictability in measures of ecological function. It is increasingly important that we measure structure and function at a similar range of scales before we will be able to understand the role of natural patchiness in biodiversity in ecology. This is not yet widely done, but it is crucial if we are to conserve functioning habitats in an increasingly altered world.

VIII. WHAT SERVICES DOES INTERTIDAL BIODIVERSITY PROVIDE?

Consideration of relationships between diversity of species and functioning of ecological systems leads rapidly to consideration of any "values" of biodiversity. Because biodiversity is not well described (i.e., the threats to it are poorly understood and probably variable from place to place and habitat to habitat), there is still considerable controversy about perceived value(s) of biodiversity. This is specifically true for most coastal habitats, especially those on temperate coasts, which have not received as much attention from the media and celebrities as have habitats, such as coral reefs or tropical rain forests.

There are many so-called values to biodiversity, some of which are discussed in Box 1. Whatever terms are used, however, they tend to fall into two main catego-

**Box 1**

Despite the fact that there is no unambiguously accepted list of what is meant by the values of biodiversity, many values are cited as being important.

*Evolutionary value.* This includes genetic diversity that may allow organisms to persist or change in response to localized threats to diversity. This may be very important in intertidal and nearshore coastal habitats where humans tend to cause most damage to marine habitats.

*Ecological value.* A naturally functioning habitat with its full complement of biodiversity is necessary for studying and understanding ecological patterns and processes. This understanding is, in turn, considered essential for the persistence of life on earth in the face of humanity's alteration to and degradation of the natural world.

*Economic value.* There are or may be species of direct economic value. On intertidal shores, this includes plants and animals collected or cultured in fisheries, mangrove trees exploited for firewood, and so on. Habitats that are used for recreation may also have economic value. For example, a wetland with diverse species of birds, can have economic value if bird-watchers are prepared to pay to access it.

*Aesthetic value.* Natural habitats are sometimes perceived to have value simply because they exist. This is also sometimes called "existence value."

*Ethical value.* There is argument by many sections of society, that all biodiversity has ethical value—that is, we, as human beings, are ethically bound to protect biodiversity. This argument comes to the fore when decisions are being made about the rights of humanity to eliminate diseases and other pests from the earth.
ries. Ecocentric values are those associated with the well-being of the animals and plants themselves (i.e., they are focused on maintaining the full range of functioning ecological processes). They include such terms as ecological value and evolutionary value (see Box 1). Anthropocentric values, in contrast, are centered around humanity and the role of diversity in maintaining or improving human lifestyles. These obviously include a range of values, based on our perceived moral obligations, our ideas of the sort of world that we want to live in, and obvious economic returns from our exploitation of coastal resources.

In contrast to many terrestrial habitats, nearshore coastal habitats, including nearly all intertidal areas, are generally conserved for anthropocentric values rather than ecocentric values. Following traditions lasting hundreds of years, the sea is still largely regarded as a larder to be plundered rather than a unique set of habitats to be conserved. Apart from certain coastal wetlands, which can support large populations of wading birds, most intertidal habitats do not abound with large charismatic megafauna. Therefore, although they may contain threatened species, these are unlikely to attract media attention.

Although large areas of mangroves can be extensive, diverse, and very interesting forests, they are still generally valued for their economic returns—provision of firewood, habitats for juvenile commercially exploited fishes, sites for aquaculture, and, unfortunately in many countries, prime real estate for reclamation. Similarly, rocky reefs, mudflats, estuaries, and so on are generally considered important for what direct economic value they can provide humanity.

The loss or degradation of nearshore habitats goes largely unseen because the biota that live in them are small, cryptic, and unknown. Yet they are incredibly diverse, and in many parts of the world they make up a large amount of the endemic fauna and flora. Most of these organisms do not have commercial value. As long as these coastal habitats continue to be considered valuable mainly with respect to the direct economic services that they provide to humankind, there will be little initiative to do the research and impose the necessary management to conserve this diversity.

IX. CONCLUSIONS

The management of biodiversity in intertidal coastal habitats needs a different approach from that used in terrestrial habitats. Except for species that are considered to have a key role in local ecological processes, it is unlikely to be useful to attempt conservation on a species by species level. Exceptions are the so-called keystone species or ecological engineers, which were discussed in some detail earlier. Either these species provide habitat directly or their activities serve to enhance local diversity because of their elimination of competitively dominant species. For the most, however, emphasis on conservation of individual species is unlikely to be profitable.

First, for most species, we know too little about their requirements for habitat, food, and so on to be able to impose sensible managerial options. Second, we know too little about their natural patterns of variability in time or space to be able to evaluate whether any managed populations are persisting with anything like their natural patterns of abundance. Third, we certainly know too little about their interactions with other species to know which are essential to their continued well-being and which are not (i.e., which suite of species to try to conserve). In many cases, it has so far proven impossible to determine the geographical range of habitat occupied by a breeding population of intertidal animals, the extent to which there is interchange from one population to another or the sources of recruits to any area that must be conserved.

Management of coastal biodiversity will be better focused on habitat. Removing or reducing disturbances to patches of mangrove forest, rocky shores, mudflats, and beaches is one of the more feasible and cheaper options—it is usually easier to prevent access to areas than try to control people’s behavior once they are in them. The fauna and flora will then be left to do the best they can under the circumstances. Fortunately, many species living in shallow-water, marine habitats are able to deal with many disturbances to their habitat, as long as these are neither too frequent nor too extreme. Evidence suggests that many can relatively rapidly recolonize areas once any disturbances are removed: such areas can then develop what appear to be functional ecological systems. Whether all species return to such habitats is not known because it is generally not known what species were there before the disturbance.

Nevertheless, there is obviously need to do more research on interactions among species wherever possible, so that keystone or engineering species can be identified. Where possible, management that will conserve these directly is likely to conserve those species that are dependent on them. Therefore, legislation that protects rocky shores from foragers removing mussels and other large shellfish will not only protect the tar-
See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • MANGROVE ECOSYSTEMS • MARINE ECOSYSTEMS • MEASUREMENT AND ANALYSIS OF BIODIVERSITY

Bibliography


INTRODUCED PLANTS,
NEGATIVE EFFECTS OF
William G. Lee
Landcare Research, New Zealand

I. Introduction
II. Notable Examples
III. Predicting and Evaluating Impacts

GLOSSARY

**economic impact**  Capacity of weeds to limit productive use of terrestrial environments and the costs associated with their control in both managed and natural areas.

**environmental impact**  Ability of weeds to displace native fauna and flora, alter key ecosystem functions, and change disturbance regimes.

**introduced plants**  Species transported by humans to regions outside their natural geographic range.

**invasive weed**  Introduced plant species that establish self-maintaining populations and spread, with and without human assistance, into new areas where they frustrate human intentions in production and natural landscapes.

**PLANT SPECIES THAT ARE**  either deliberately or inadvertently transferred by humans to habitats outside their native geographic range and which subsequently naturalize and have significant negative economic consequences and environmental impacts.

I. INTRODUCTION

Human migration, settlement, and trade over the past two centuries have produced a rapid globalization of floras with independent evolutionary histories that were previously kept apart by natural barriers. The scale of movement and mixing of floras in recent years is unprecedented in the earth’s history, and very few parts of the world currently retain an exclusively native assemblage of plant species. Hundreds of naturalized plants, no longer dependent on humans for their persistence, occur in most regions and countries, where they form a significant and growing proportion of the wild flora. Over large continental areas the contribution of established introduced plants can reach 23% of the total flora (e.g., Canada), but this is greatly exceeded on many islands where the naturalized flora may outnumber the pool of native species (e.g., Bermuda, Ascension, New Zealand). Despite the many benefits provided by introduced plant species, there is growing concern at the environmental damage and economic costs of naturalized invasive plants in both productive and natural ecosystems. Important negative impacts are created by a small subset, probably less than 0.01% of all introduced species growing in a region, but their control, management, and impact consume significant economic resources, especially of developed countries.

The extent, patterns, and consequences of introduced plant invasions are described in another chapter presented in this volume, “Plant Invasions.” This chapter highlights the negative impacts of introduced plant species using case studies of notable invasions, particularly those that have spread into natural areas with serious deleterious consequences for native species and communities.
therefore relatively recent, spread and impact in natural environments. They are not necessarily the most widespread plant species, or the most aggressive, and some are limited to a few regions or countries. However, they are representative of the pattern and scale of intercontinental transfer of introduced plant species and the types of dramatic impacts these species can have when released into new environments. The examples chosen include a range of plant growth forms, altitudinal and bioclimatic zones, regions of origin, and modes of negative impact.

For ten introduced invasive plant species, representing algae (Undaria pinnatifida), ferns (Salvinia molesta), grasses (Bromus tectorum, Spartina alterniflora), small herbs (Hieracium pilosella), cacti (Opuntia), shrubs and small trees (Lantana camara, Mimosa pigra), and tall trees (Pinus pinaster), an account is given of the invasion process; local ecological, cultural, and economic impacts; attempts at control; and likely long-term consequences. Not all invasions are irreversible, and several examples illustrate successful levels of population control, if not eradication.

II. NOTABLE EXAMPLES

A. Undaria pinnatifida (Laminariales)

Some of the most pronounced ecological impacts and intractable problems of introduced species invasion involve marine organisms. Both deliberately and inadvertently introduced macroalgae have become invasive in recent decades with the increase in marine farming, transoceanic and coastal shipping, and the discharge of ballast water in and around foreign ports. The Asian laminarian kelp (Undaria pinnatifida), native to Japan, Korea, and China, has been introduced and spread since 1970 in other northern hemisphere waters (e.g., England, French Atlantic, and in parts of the Mediterranean near France, Spain, and Italy). More recently (1980s), populations have been discovered in the southern hemisphere (Argentina, Australia, and New Zealand), making it the first large kelp to cross tropical waters successfully. In its native range Undaria pinnatifida is widely cultivated as an edible seaweed, and new locations are often viewed positively for their commercial potential. However, the alga is highly invasive, has significant negative conservation impacts on the local native shallow marine fauna and flora, and may be impossible to eradicate.

Native to the north-west Pacific, Undaria pinnatifida has an annual, heteromorphic life cycle, with macroscopic sporophytic and microscopic gametophytic stages. The golden brown, pinnatifid sporophyte, up to 5 m long, grows over autumn and winter, before degenerating in the summer. When growth slows, sporophytes develop as undulating extensions at the base of the stipe and release millions of short-lived (generally less than 2 days) male and female pelagic spores which, following settlement, develop into small gametophytes that may overwinter before producing either eggs or sperm. Fertilization of the female egg cell by male gametes gives rise to the straplike juvenile sporophyte. Undaria pinnatifida grows from low water to a depth of 20 m, depending on turbidity and wave exposure, and tolerates sea surface temperatures from 0 to 27°C. In Japan, it is an important edible seaweed crop and is also cultivated extensively to feed juvenile abalone (Haliotis spp.).

Undaria pinnatifida first appeared outside its native range in 1971 in a lagoon on the south coast of France, in the Mediterranean Sea. The source is thought to have been contaminated oyster spat translocated from the Pacific. It has subsequently spread along northern parts of the Mediterranean from Spain to Italy. In 1983 Undaria pinnatifida was deliberately introduced to the French Atlantic for experimental aquaculture trials, on the basis that sea temperatures would be too cool (< 14°C) for the release of zoospores. This assumption was proven incorrect when ecological surveys during the 1990s revealed natural establishment at five of the nine cultivation sites, some now abandoned, with several new populations up to 20 km away from nearest putative sources. Undaria pinnatifida appeared on the south coast of England in 1994, the population mostly likely deriving from microscopic gametophytes or juvenile sporophytes attached to ships arriving from Brittany.

Spread of Undaria pinnatifida into the southern hemisphere has been very recent, beginning in the 1980s, and accidental, presumably from plants attached to ships or in seawater ballast. Undaria pinnatifida was first discovered in New Zealand in 1987, and has subsequently spread, mainly via sporophytes and gametophytes attached to hulls of boats, to more than ten localities, up to 1000 km apart (Fig. 1). Soon after (1988) the kelp was discovered in Tasmania, in and around several major ports, and was recently (1996) reported from coastal waters on the Australian mainland. The first South American population was discovered in Argentina in 1992.

While fouling boats and underwater structures is of nuisance value, the full ecological impact of Undaria pinnatifida in these new habitats is hard to predict, as...
the species is a very recent invader in foreign waters, is mostly confined to highly modified habitats in harbors, and initially preferentially colonizes clean, artificial surfaces. Investigations in the Atlantic suggest that Undaria pinnatifida is preempted on denuded rocky substrates by several local native annual and perennial kelps. Because of this competition, it may eventually be restricted along the coast in northern France to the low tide zone. In Tasmania, following disturbance and dieback of the native kelp due to ocean warming, Undaria pinnatifida displaces Macrocystis kelp forests and dominates subtidal urchin barrens due to its high fecundity and fast growth rate. However, it is less successful against native seaweeds on exposed coasts with slow currents. In New Zealand, Undaria pinnatifida reduces habitat for native marine grazers by smothering coralline algal surfaces favored for larval settlement of the animals. The region has a depauperate flora of annual and early successional seaweeds, and Undaria pinnatifida has few effective competitors on fresh surfaces created by storms and echinoid grazing. Frequently wave-induced substrate disturbances could consolidate the dominance of Undaria pinnatifida in many shallow marine habitats.

While the transoceanic and local spread of Undaria pinnatifida has been dependent on human activity, the alga has several attributes that make it an aggressive competitor in shallow marine habitats. Sporophytes densely colonize any unvegetated surfaces, and huge populations can establish within a year on boats, wharf piles, mooring buoys, and on breakwater or reclamation debris. These can provide multiple sources for short- and long-distance dispersal. Importantly, Undaria pinnatifida occurring outside its native range can, in warmer waters, produce several generations of sporophytes each year, while local kelp species are reproductively dormant for part of the year.

The negative effects of Undaria pinnatifida on marine aquaculture and numerous algal communities are potentially enormous, and most countries where the species has established are pursuing some form of control. However, the large reproductive potential and cryptic dispersal phases of Asian kelp, together with the partial obscurity provided by the marine environment, combine to make detection and control of Undaria pinnatifida almost impossible. Reducing the rate of spread of Undaria pinnatifida may be practicable with improved ship hygiene practices in ports and the strict control of aquaculture ventures. Eradication of small populations, using direct removal of sporophytes by divers, or sterilization of mariculture equipment, may also be possible where pioneer plants are restricted to artificial structures (e.g., mooring lines), but this is unlikely to be achieved where Undaria pinnatifida has established in harbors and along exposed coasts.

B. Salvinia molesta (Salviniaceae)

Ferns are rarely successful weeds in natural ecosystems, but a notable exception is Salvinia molesta (kariba weed, African pyle, Australian azolla, water fern, giant azolla), a free-floating perennial aquatic fern native to tropical and subtropical areas of South America. In the later part of the 20th century it has been spread worldwide, becoming invasive in fresh waters in warm-temperate to tropical areas, and causing major disruptions to the utilization of important water resources.

Only formally described in 1972, Salvinia molesta is a sterile pentaploid, probably of natural hybrid origin. Native to parts of Brazil and Argentina, it grows there in lagoons, swamps, and river margins with still or slow-moving freshwater, in a diverse community of floating and emergent plant species. In these situations Salvinia molesta occurs at low densities, does not form extensive mats, and is rarely invasive.

Morphologically the species is quite plastic, exhibiting several distinctive forms. At each node Salvinia molesta supports submerged (which function as roots) and floating leaves. The size, shape, and density of the foliage changes depending on environmental conditions (temperature and nutrient concentrations) and growth phase. When Salvinia molesta first establishes, usually in calm, sheltered waters, the plants are small (20 mm), with oval, flat floating leaves. With time, leaves become rounded and larger, and in the final mat-forming stage...
are folded, crowded, and between 50 and 60 mm across. *Salvinia molesta* is efficient at accumulating and reutilizing nutrients such as nitrogen. *Salvinia molesta* has become widely distributed through aquarium and horticultural industries and is now naturalized in Australia (Fig. 2), Africa, Madagascar, India, Sri Lanka, Southeast Asia, Philippines, Indonesia, Papua New Guinea, Fiji, and New Zealand. Dispersal between water systems is also invariably by human activities (e.g., boats, aquaria), but spread within and between interconnected water bodies is assisted by wind and water movement of vegetative fragments. Outside its native environment, *Salvinia molesta* shows phenomenal growth rates, via the expansion of auxiliary buds at stem nodes. Under optimum conditions (30°C and high nitrogen levels), plant area can double every 8 days and biomass every 2.2 days. In Australia total floating biomass values of 1600 g m⁻² dry weight, or 400 t ha⁻¹ fresh weight, have been recorded. Frequently *Salvinia molesta* forms mats 1 m deep across extensive waterways.

During the 1970s and 1980s, *Salvinia molesta* was seen as potentially one of the worst invasive weeds globally, with major ecological and socioeconomic impacts. The basis for this concern is well illustrated by accounts of the species invasion in Papua New Guinea. In the early 1970s *Salvinia molesta* became established in the Sepik River flood plain, and within a decade covered 250 km² of water surface. The weed impeded water transport, preventing fishing, trade between villages, and access to medical and educational facilities. Fresh water for humans, stock, and wildlife became restricted, and *Salvinia molesta* mats proved suitable habitats for insect vectors of serious human diseases such as malaria. Similarly, in Zimbabwe, *Salvinia* threatened utilization of Lake Kariba, a massive impoundment on the Zambezi River created in the 1950s for hydroelectricity generation, commercial fishing, recreation, and transport. Within 13 years of the scheme being completed, mats of *Salvinia molesta* were limiting human access, decreasing fish populations, and modifying nutrient regimes. Elsewhere, *Salvinia molesta* has also invaded rice fields, blocked irrigation ditches, eutrophied small water bodies, and smothered emergent, floating, and submerged indigenous aquatic plants. A severe reduction in light and dissolved oxygen, together with an increase in carbon dioxide and hydrogen sulfide, have had a drastic impact on most benthic biota.

The management and eradication of *Salvinia molesta* focused initially on mechanical removal and the use of chemical herbicides, and these were occasionally effective, especially where original populations were small or isolated. However, repeated treatment was required to prevent the rapid recovery of the plants, and the techniques used were expensive and occasionally environmentally damaging. Since 1980, biological control of *Salvinia molesta*, using a foliage- and stem-feeding weevil (*Cyrtobagus salviniae*—Curculionidae) from its native habitat, has been outstandingly successful in suppressing populations and achieving control in many countries (e.g., parts of Australia, Papua New Guinea, Fiji, South Africa). For example, the weevil had reduced the cover of *Salvinia molesta* on two 16 ha dams in Zimbabwe from over 90% to under 2% in less than two years. Similarly impressive results were achieved at a much larger scale in the Sepik River. Following the weevils’ introduction, large mats of *Salvinia molesta* are typically reduced to sparse plants around the margins of water ways, interspersed with other aquatic plant species of the region. After the collapse of the *Salvinia* plants, weevil numbers also decline but appear to remain at densities sufficient to keep the weed in check. The global benefit of this successful example of biological control has been estimated to surpass $A200 million.

C. Bromus tectorum (Poaceae)

The development of rangelands for pastoral use, through stock grazing, fire, and the deliberate and incidental introduction of new plant pasture species, has involved some of the most extensive modifications to natural ecosystems ever undertaken by humans. In the semi-arid intermontane west of North America more than 400,000 km² of native sagebrush-steppe communities have been penetrated or displaced this century by...
**Bromus tectorum** (cheatgrass, downy brome, downy chess, drooping brome), a winter annual grass native to arid Europe and Central Asia. The dominance of this species has resulted in changes in land use, increases in erosion potential, reductions in native biodiversity, and alterations to the major disturbance regimes over large areas. *Bromus tectorum* is considered to be the most significant plant invasion in North America.

The native habitat of *Bromus tectorum* is centered on dry continental climates. Commonly associated with human activities, it occurs in grazed and ungrazed grasslands, among crops, and along roadsides. Flowers are hermaphroditic and self-fertile, and populations survive the dry summer period as seed. It characteristically germinates in autumn, estivates during the coldest parts of winter, and can grow rapidly, to around 30 cm tall in spring, when it flowers. The plants die back during the hot, largely rainless, summer. Seeds can be widely dispersed via animals, as they have long awns that are frequently embedded in fur or wool. Transfer by humans is usually accidental as a seed contaminant and among farm machinery.

*Bromus tectorum* is now found in temperate East Asia, North and South America, and Australasia. Prior to European settlement, the natural vegetation of the Great Basin area contained perennial cespitose grass-dominated steppes, the grasses (*Agropyron*, *Festuca*, *Stipa*, *Poa*) and forbs on mesic sites, giving way to shrubs (*Artemisia*, *Chrysothamnus*) in drier areas.

The accidental introduction of *Bromus tectorum* into the intermontane region of the Pacific Northwest late in the 19th century, and several deliberate attempts to establish it as an alternative pasture species, coincided with large-scale domestic grazing by cattle, sheep, and horses. Because of the low abundance of large ungulates (bison and deer) during the Holocene, the original grasses and shrubs were generally poorly adapted to grazing ruminants. The flora also lacked aggressive colonizing annuals that could reclaim sites following disturbance. Burning by early settlers to remove shrubs and improve stock access further depleted the vegetation. Initially a weed of cultivated fields and roadides, *Bromus tectorum* rapidly spread via wind and animal dispersal, and in seed lots, into overgrazed and vulnerable ungrazed native communities, becoming a widespread and dominant weed in the 1930s and reaching its current distribution in North America by the 1950s. Spread since then continues but at a slower rate.

The expansion of *Bromus tectorum* has caused multiple changes in natural ecosystems, largely through its impact on the grass/fire cycle. Accumulation of its highly flammable dry litter following spring growth, has resulted in an increase in the frequency of natural lightning-induced summer fires from 2 to more than 20 per century. Frequent burning has killed native shrubs and depleted the recruitment of perennial grasses. In contrast, *Bromus tectorum* benefits from fire, due to its large seed bank, prodigious and prolonged germination, rapid seedling growth, and heavy annual flowering. It is therefore able to preempt the recolonization of burned areas by native species already weakened by regular fires. Greater intensity and frequency of burning following *Bromus* invasion reduce plant cover, giving rise to erosion-prone landscapes. *Bromus tectorum* provides poor-quality herbage for stock, and its dominance has resulted in a decline in rangeland quality (Fig. 3).

The outcome for conservation and pastoralism of the transformation, in less than half a century, of natural semiarid steppe dominated by shrubs and perennial grasslands, to communities predominantly of introduced annual grasses, remains uncertain. *Bromus tectorum* appears uncontrollable under extensive management systems. It is still spreading into more isolated areas of the Great Basin and into the Great Plains region of the United States.

**D. Hieracium pilosella** (Asteraceae)

Numerous flat-weeds (rosette species) have accompanied European colonization of temperate regions around the world. Most of these plants are small, short-lived (2 years), dependent on human disturbance for local persistence, and generally inconsequential as weeds outside intensively managed systems. *Hieracium*
pilosella (mouse-ear hawkweed), a perennial, stoloniferous, mat-forming daisy native to Europe, has a worldwide distribution in temperate regions and was generally considered to be relatively benign until it started to invade grasslands in New Zealand in the 1950s. Currently it occupies about 1 million ha of montane-subalpine short tussock (Festuca novae-zelandiae) grassland in subhumid and humid zones and is seen as the greatest threat to the conservation and long-term pastoral use of native grasslands. Key ecological factors responsible for the spread of *Hieracium pilosella*, and management options for its control, are controversial, and highlight the problems of large-scale weed management in non-forest ecosystems.

Native to Europe, *Hieracium pilosella* is a prostrate rosette-forming herb that reproduces vegetatively via stolons, which are initiated following flowering. Seed can be produced by both sexual and apomictic processes. In addition, hybridization and polyploidy are common and give rise to complicated population structures. Rosettes produce an erect short scape, which supports a solitary lemon-yellow flower that produces abundant small wind-dispersed achenes. In its native range the species is characteristic of disturbed dry habitats, most often grazed pastures, wastelands, and outcrops, particularly on calcareous soils below 1000 m above sea level.

Introduced into New Zealand in the middle of the 19th century, most likely as a seed contaminant, *Hieracium pilosella* remained locally restricted and largely insignificant for more than a century. The invasive phase of *Hieracium pilosella* became obvious during the 1960s, initially in the semiarid intermontane basins in the South Island, where the species became dominant in grazed Festuca-Poa grassland induced by pastoralism. Following burning of the original native forest and woodland during early Polynesian settlement, approximately 600 to 1000 years ago, tall tussock Chionochloa grassland and shrubland extended onto these areas. Early European pastoralists, in the first few decades of the 19th century, increased fire frequency and brought in sheep and cattle which, assisted by periodic eruptions of introduced rabbit populations, destroyed dominance by the tall cespitose grass species. The decline in plant cover and increasing bare ground provided numerous sites for invasive grazing-tolerant plant species. Foremost amongst these has been *Hieracium pilosella*, which slowly occupied the most depleted areas (those with 500 to 1200 mm rainfall) over about 50 years, becoming widespread and dominant in the 1970s when it began to extend into subalpine tall tussock grassland. In the 1990s *Hieracium pilosella* became the most widespread and significant introduced weed in native grasslands (Fig. 4), accompanied in parts of its range by the introduced congeners *Hieracium praealtum* and *Hieracium lepidulum*.

During its colonizing phase *Hieracium pilosella* forms distinctive circular patches that gradually coalesce to cover large areas. In eastern parts of the South Island, instantaneous rates of increase of approximately 8% per year have been recorded, with local cover values of *Hieracium pilosella* exceeding 50%. Much of this local expansion is attributable to vegetative reproduction, with seedlings being rare. *Hieracium pilosella* smothers and displaces small native herbs and grasses and restricts the survival of productive pasture grasses and legumes. Soil changes beneath *Hieracium pilosella*, especially increasing rates of acidification, also limit the growth of introduced legumes.

There is little doubt that human activities, especially repeated fire, and the grazing of domestic and feral animals, have increased the vulnerability of indigenous grasslands to invasion by introduced plant species. However, some relatively unmodified native habitats, especially in subalpine areas, on braided river beds, in semiarid environments, and on unstable hillslopes, may also be vulnerable due to tectonically induced natural erosion maintaining bare ground. *Hieracium pilosella* possesses several attributes that enable it to outcompete many low-growing native grass and herb plant species, especially in drier regions. These include rapid growth, vigorous vegetative reproduction, drought-tolerance, and grazing resistance. Furthermore, *Hieracium pilosella* is more efficient than native plant species in utilizing seasonal pulses of nutrients and water. The wide-
spread aerial application of superphosphate fertilizer in upland areas may also have contributed to the spread of *Hieracium pilosella* in recent decades. New Zealand plants of *Hieracium pilosella* are entirely polyploid with high levels of genetic variation enabling the rapid evolution of innovative genotypes for colonizing new habitats. The future of *Hieracium pilosella* in indigenous grasslands probably depends on land-use goals. In tall-tussock grassland, in the absence of mammalian grazing, natural vegetation cover and successional processes eventually leave *Hieracium pilosella* restricted to marginal habitats. However, the outlook for induced short-tussock grassland, with or without stock or feral animal grazers, is less clear, and *Hieracium pilosella* may well monopolize these communities in the longterm, or at least until succeeded by shrubs. Active management in areas under extensive pastoralism involves grazing to reduce flowering and applications of fertilizer to improve the competitive ability of associated grasses and legumes. A biological control program in New Zealand is currently underway, and two agents collected from Europe have been utilized: a pathogenic rust fungus (*Puccinia piloselloidarum var. hieraci*); and a gall wasp (*Aulacidea subterminalis*), which forms galls on stolons.

**E. Spartina alterniflora** (Poaceae)

Estuarine habitats, characterized by waterlogged soils and regular saltwater incursions, provide a severe environment for plants and usually support a sparse halophytic flora. One kind of plant that is able to form dense monospecific stands in these environments is the grass *Spartina* (cordgrass), especially *Spartina alterniflora* and its derivatives, which have spread in estuaries in many subtropical and temperate regions of the world. The rapid evolutionary process involved and the growing recognition of the biodiversity values of estuarine areas. *Spartina* species are perennial, deep-rooted (30 cm), rhizomatous, sward-forming erect grasses that mostly occupy low- to midtidal mudflats. They are well adapted physiologically to tolerate saline conditions with special salt excretion, dilution, and restriction mechanisms. A high water-use efficiency, due to their C₄ photosynthetic system, and the unusual ability to maintain higher rates of photosynthesis than other C₃ and many C₄ species under cool temperate conditions (5–10°C), increases their environmental range. The native geographic distribution of the genus centers on the east coast of North and South America, with fewer species on the west coast of North America, Europe, and north Africa. *Spartina alterniflora*, native to the Atlantic coast of North America, was accidentally introduced to the United Kingdom early in the 19th century, and produced rare hybrids with the local congener *Spartina marinum*, which is restricted to Western Europe and Africa. The male sterile F₁ hybrid *Spartina x townsendii* subsequently produced, via doubling of chromosomes, a new fertile species *Spartina anglica*. It quickly became apparent that *Spartina townsendii*, and particularly *Spartina anglica*, were markedly more invasive than either of the parent species. Estuarine habitats in southern England were transformed from exposed mudflats into tall-grass meadows. In less than a century *Spartina anglica*, via seed dispersal, clonal spread, and human plantings, covered approximately 10,000 ha of intertidal salt marsh along the coast of Britain.

Early this century, estuarine stabilization to maintain commercial waterways and reclamation for industrial and agricultural development were seen as important goals of coastal management. Experience in the United Kingdom resulted in *Spartina alterniflora* and its derivative species (*Spartina x townsendii*, *Spartina anglica*) being widely planted in parts of North America, in regions occupied by native Spartina species, and in areas well outside its native range (e.g., Australia, and New Zealand) where they were similarly successful, especially *Spartina anglica*. The heightened competitive vigor and environmental range of *Spartina anglica* appears to derive from a greater genetic heterozygosity, due to hybridization and polyploidy, and expanded phenotypic plasticity. Spreading *Spartina* species appear to have fewer insect and avian herbivores in their new ranges.

Invasive *Spartina* species induce major geomorphological changes in tidal habitats by enhancing rates of sediment accumulation. Sediment accretion rates of up 17 cm yr⁻¹ (average 5–10 cm) have been recorded within the tall grass sward, and in southern England *Spartina* caused elevation of mudflat surfaces by 1.8 m in 37 years. Declining tidal influence, and the accumulation of biomass (dry weight up to 7500 kg ha⁻¹), transforms salt marsh areas into non-estuarine ecosystems. Impacts on biodiversity are no less dramatic. At the pioneer phase, establishment via seed or planting produces circular clumps that expand at rates of between 3 and 7 m yr⁻¹, eventually coalescing to form extensive meadows on sand and mud substrates in the low- to midtidal range (Fig. 5). *Spartina anglica* physically displaces many native indigenous low-growing halophytes (e.g., *Selliera, Salicornia, Schoenus, Puccinellia*) and modifies habitats formerly occupied by a range of bird,
INTRODUCED PLANTS, NEGATIVE EFFECTS OF

F. Opuntia (Cactaceae)

Cacti are prized as ornamentals for their novel form, spectacular flowers, edible parts, and medicinal uses. *Opuntia* (prickly pear) a genus of about 90 species native to North and South America, has been distributed globally over the past two centuries. Sixty species have become naturalized, and at least 35 are considered major weeds. *Opuntia* species provide some of the earliest and most spectacular examples of invasive plants in dry environments and the effectiveness of insects as biological control agents. In the major centers of Opuntia invasion (Australia, South Africa, and India), several species have displaced horticultural crops, reduced the grazing potential of rangelands, and threatened the conservation values of native grasslands, scrublands, and woodlands. Eradication may be unattainable, but some invasive Opuntia species are now at low population densities, while others are continuing to spread. Opuntia species are fast-growing perennial succulents with thickened, often flattened, segmented cladodes, usually supporting spines. Most Opuntia will regenerate from seed, cladode fragments, and underground tubers. Ranging in size from low-growing shrubs to small trees, the species characteristically occupy dry habitats with seasonal water deficits. As with all cacti, Opuntia species have a distinctive photosynthetic system (crassulacean acid metabolism or CAM), which enables them to fix carbon at night when evaporative stress is reduced, as well as during the day, if adequate moisture is available. A high water-use efficiency, coupled with a large internal water-holding capacity and the ability to restrict water loss by tightly shutting off stomata, enhance drought tolerance. In North and South America Opuntia species are usually sparse and restricted regionally, and several are considered threatened in the wild. The major invasive weed species come from a range of native regions: *Opuntia stricta* (common or erect prickly pear) eastern North America, West Indies and adjacent South America; *Opuntia aurantiaca* (tiger pear, jointed cactus) temperate South America; *Opuntia ficus-indica* (sweet prickly pear, Indian fig) Central America; and *Opuntia vulgaris* (drooping prickly pear, Barbary fig) eastern South America.

Initially *Opuntia* species were distributed outside the New World as ornamental plants, but a range of uses rapidly emerged, especially as stock fodder, hedges, and as fruits or vegetables. A novel use was the production of carminic acid, a commercially important red dye, derived from the crushing of dried cochineal beetles fed a diet of cacti. *Opuntia* can provide an important source of food and income for local communities, and attempted extirpation is not universally accepted.

Introduced into Australia in the early 19th century, *Opuntia stricta* had by 1925 infested more than 25 million ha of eastern Australia and was spreading at the rate of 100 ha per hour, aided by the dissemination of vegetative and seed material by floods, humans, and feral and wild animals. Areas most susceptible were pastoral and arable land cleared amongst Acacia and...
Casuarina woodland, although the cacti also penetrated the woodland understorey. The Australian flora has few native succulents, and none with the environmental tolerance of *Opuntia*. Approximately 10 years after initial establishment at a site, ground cover became dominated by impenetrable thickets of *Opuntia* reaching densities of 10,000 plants and a biomass of 250,000 kg, per hectare. The weed frustrated the farming ambitions of European settlers and caused the ruin of early rural economies.

*Opuntia ficus-indica*, a tall shrub, was established in southern Africa at least 250 years ago. By the early 20th century, infestations occupied 900,000 ha, mainly in the Eastern Cape region, occupying grassland, succulent-karoo, and the savanna biome (Fig. 6). *Opuntia aurantiaca*, a low-growing species, which may be of hybrid origin, was introduced much later (early 19th century), spreading originally from garden plantings around Cape Town. A total infested area of about 400 ha in the 1890s has grown this century to around 830,000 ha, mainly in eastern parts of South Africa. It is replacing important pastoral plants in grasslands and savanna, injuring domestic and feral animals, and degrading natural rangelands. A single cladode can produce up to 145 new cladodes over a 200-day growing period, and the potential for vegetative spread is enormous. Currently *Opuntia aurantiaca* is considered to be South Africa’s most expensive weed.

*Opuntia vulgare* has also been an important weed in the latter part of the twentieth century, mainly in western coastal areas.

In Australia and South Africa, the control of *Opuntia* assumed top priority for land management agencies early in the 20th century, when various herbicidal and mechanical methods were attempted. Poisoning, using a combination of arsenic pentoxide and sulfuric acid, and mechanical cutting were successfully used to control small infestations of *Opuntia* in open lands but were very expensive and achieved little at a regional scale. However, spectacular success has been achieved using plant-sucking cochineal insects *Dactylopius* species and the cladode-eating larva of the moth *Cactoblastis cactorum*. Indications of the potential of insects as biocontrol agents first appeared in 1795, when the accidental introduction of *Dactylopius ceylonicus* resulted in widespread death of *Opuntia* vulgaris in India. Subsequently, deliberate introductions of several *Dactylopius* species have drastically reduced Opuntia weed infestations in southern Africa, and *Cactoblastis* has been similarly used to diminish population densities of *Opuntia stricta* in Australia. Overall, biological control agents have lowered population densities of *Opuntia* by 90%, especially in drier climates. *Opuntia* remains a widespread and invasive weed in these countries, but the economic impact on agricultural land uses has diminished hugely, especially in Australia.

G. *Fallopia japonica* (Polygonaceae)

The human passion for gardens increases biodiversity in urban areas, often creating large source populations, which initiate the effective dispersal of introduced plants into adjoining landscapes. One garden ornamental that has become an invasive weed is *Fallopia japonica* (Asiatic knotweed), a large perennial herb native to the Far East. In the 20th century the species has infiltrated much of central and western Europe, North America, and several southern temperate countries, where it excludes native plant species in artificial and highly disturbed habitats, riparian areas, and in open woodlands. In the United Kingdom, *Fallopia japonica* is presently the tallest and most aggressive common herbaceous species.

*Fallopia japonica* is a rhizomatous, clump-forming, perennial native to China, Japan, Korea, and Taiwan. In these countries it is an early successional species, growing to less than 2 m tall, colonizing volcanic debris and disturbed riparian habitats, under a range of edaphic conditions, from lowland to subalpine. *Fallopia japonica* establishes on new sites by seed in its native range and persists to develop a large biomass (12 tonnes ha$^{-1}$), deep litter layer, and a rhizome system that extends for up to 20 m beyond the shoots.

Taken from Japan to the United Kingdom in 1825, for planting in gardens, *Fallopia japonica* was subsequently established in other northern (e.g., North America) and southern hemisphere (e.g., New Zealand)
countris late in the 19th century. Although mostly found on heavily modified and disturbed sites within urban environments, it has also spread in natural habitats, especially along waterways. In the United Kingdom it is presently found in over half the 10-km grid squares, being most abundant in southern and western regions (Fig. 7). In Wales, for example, Fallopia japonica occupies 84% of river systems with average flow rates greater than 2.3 m$^3$/s. In North America it has become naturalized in most mesic western coastal regions, north to Alaska, and in all of the northeastern United States, and parts of central and southern United States. In western Pennsylvania, stands extending for hundreds of hectares cover wetlands, stream sides, and adjoining hill slopes.

Around waterways, Fallopia japonica is capable of forming dense thickets up to 500 m across, outcompeting and supplanting native herbs, shrubs, and small trees; impeding water runoff and increasing flooding, and modifying nutrient availability and cycling patterns through sequestration in a large standing biomass. Early emergence and fast growth of new shoots and flowering stems, usually completed before midsummer, create a dense overstory, and this, together with prodigious annual litter production, restricts the presence of other species, except where regular flooding removes surface litter, and the shade from tall trees reduces its vigor.

In its native range, Fallopia japonica bears female and hermaphrodite flowers on separate individuals. Elsewhere the species is male-inferior and, although seeds are produced from hybridization with other introduced congeners, spread via successful sexual reproduction is rare. However, spread of rhizome fragments is extremely potent, and viable plants can establish from pieces as small as 7 g, and from burial at depths of up to 1 m. Stem fragments are also able to produce new shoots. In most instances vegetative dispersal is achieved by water currents and the movement by humans of soil and gravel contaminated with rhizome pieces.

The current geographic range in Europe extends north to 63° N and is broadly correlated with degree-days, absolute minimum temperatures (>-30.2°C), and annual rainfall (>-300 mm). The shoots are vulnerable to late frost and summer droughts, and these factors may constrain the large-scale distribution of Fallopia japonica, especially in continental climates. Tolerant of a broad range of soil conditions, Fallopia japonica is strongly light-demanding and rarely enters forests.

Once established, Fallopia japonica is difficult to remove. In Europe and North America young shoots of Fallopia japonica are occasionally eaten by stock, and this may reduce the development of clumps and slow the rate of spread. Where accessible, regular cutting or spraying with herbicides can achieve control, but follow-up treatments over several years are needed to weaken and eventually kill the rhizomes.

H. Lantana camara (Verbenaceae)

International trading in plants and the intensive search for novel cultivars for horticulture has produced new hybrids with high invasive capacity and given them near-global distributions and access to a great range of habitats. Lantana camara (lantana, white sage, wild sage, tick berry), a perennial shrub with brightly colored flowers, was taken from Brazil to Europe in the middle of the 17th century as a hothouse plant. During the next several hundred years, extensive hybridization and propagation of different varieties, augmented by new collections from the native region, produced a proliferation of forms. Commonly grown in gardens, Lantana camara has now become a major weed worldwide in tropical, subtropical, and warm temperate regions. It invades pastures, plantation crops, and a range of disturbed-open natural and modified scrubland, woodland, and forest communities, displacing the indigenous biota and limiting public access and use. Because of its broad distribution, invasive ability in both agricultural...
INTRODUCED PLANTS, NEGATIVE EFFECTS OF

and natural ecosystems, and local persistence, Lantana camara is considered one of the world’s top ten weeds. Lantana camara is an aromatic shrub with distinctive four-angled stems, often armed with recurved prickles. It has multicolored, insect pollinated, bisexual, potentially self-fertile flowers that produce purplish-black drupes by assorted sexual, semisexual, and apomictic mechanisms. Native to parts of Central and South America, Lantana camara is actually a complex of species and varietal forms of mixed hybrid origin. It is a genetically diverse group, with ploidy levels ranging from diploid to hexaploid. The most weedy variety (Lantana camara var. aculeata) is tetraploid. The growth form (e.g., ground creeper, prostrate shrub, tall shrub, liane) of Lantana camara is equally plastic, varying with light and soil conditions, but with support plants can reach 15 m into the canopy.

The species currently grows in many countries between 45°N and 45°S, but appears to be limited to minimum mean monthly temperatures above 3°C, rainfall in excess of 650 mm per annum, and nonsaline soils. It is a serious weed in the Caribbean, in eastern Africa, South Africa, India, Madagascar, Australia, and the Pacific Islands. Lantana camara is an important weed of plantation crops. In India, it has taken over land used for tea and sugarcane production, causing the displacement of whole villages, and in other countries it is a major nuisance in cotton, coffee, coconut, oil palm, bananas, pineapple, rubber, and rice crops. In continental countries such as Australia, India, and South Africa, it is natural grasslands that have been most extensively modified by Lantana camara. In eastern Australia alone approximately 4 million ha of grassland have become covered, eliminating both indigenous species and pastoral agriculture. Remnants of semideciduous forest are also invaded via edges and natural and human disturbances of the canopy. In South Africa, 44% of 1-km grid squares were infected with Lantana camara in a survey of 14 major forest reserves.

The spread of Lantana camara on remote volcanic islands with distinctive biotas has been of international concern and has usually been associated with European settlement and vegetation clearance via fire, roads, and grazing mammals (Fig. 8). On the Galapagos Islands, famous for their biodiversity and role in the origin of evolutionary theory, Lantana camara endangers both rare plants and animals. It is considered to have caused the extinction of one plant species and threatens the demise of at least eight others. The weed is also closing in on the last remaining colony of the darkrumped petrel (Pterodroma phaeopyia), and will, if not managed, seal off bird access to burrows. In the Indian Ocean, Lantana camara is one of several shrub-creepers invading and depleting native forests on La Réunion Island. Initially colonizing natural gaps, it subsequently ascends to smother and weaken canopy trees, making them more susceptible to damage during cyclones. It is predicted that Lantana camara will increase dominance with successive cyclones and eventually displace the forest cover. Thirteen years after its introduction to the Hawaiian Islands in 1858, Lantana camara had become naturalized on all the islands and is considered to be the number one problem plant. It currently occupies more than 160,000 ha, mainly in coastal and lowland areas, covering cropland, shrubland, savanna, and destroying opened up forest. Apart from a broad environmental tolerance, the species has numerous traits that facilitate its dispersal, penetration, dominance, and persistence in different ecosystems. In tropical climates flowers and seeds may be produced all year, and fruits are eaten and widely dispersed by a range of birds and mammals, including some similarly invasive introduced species (e.g., Indian mynah, Acridotheres tristis). Most disturbed vegetation types are vulnerable to invasion by Lantana camara. Grassland, shrubland, and woodlands, from coastal to mountain areas, are transformed into dense stands of Lantana camara as it smothers herbs and other shrubs. Continuous closed forest is less susceptible as Lantana has greatly reduced vigor in deep shade. However, plants can penetrate forest from around the margins, readily establish in gaps created by natural treefall, and prevent the regeneration of canopy trees. Plants are
Introduced plants, negative effects of flammable, even when green, and facilitate the spread of forest fires. Foliage and seeds are toxic for many grazing mammals and the thick, prickly stands of Lantana camara are virtually impenetrable. Thickets also provide habitats for vermin and disease vectors (e.g., tsetse fly).

Techniques for effectively controlling Lantana camara are slowly emerging, although most depend on the rapid establishment of vigorous crops or pastures following initial suppression of the weed. Mechanical methods are generally less successful, because Lantana camara generally resprouts from plant fragments, basal shoots, and roots after fire, cutting, or digging. Herbicides, especially those based on phenoxy or benzoic acid or pyridine, have variable success depending on climate, season, and growth form. Biological control appears to be sometimes effective, especially in drier areas, utilizing a mixture of seed- and foliage-feeding insects.

I. Mimosa pigra (Fabaceae)

Nitrogen-fixing shrub legumes provide some of the most aggressive and disruptive invasive plant species in many parts of the world. One of the most spectacular plant invasions this century has been the spread of Mimosa pigra (giant sensitive plant, zaraz, dormilona) in tropical wetlands in parts of Asia, Africa, and Australia. In less than a century, 450 sq km of natural habitat associated with flood plain areas and rivers around Darwin, northern Australia, have been transformed into dense stands of Mimosa (Fig. 9). The species endangers the conservation and use of natural wetland ecosystems in tropical regions worldwide.

Mimosa pigra is a prickly shrub native to Central and South America where it forms shrub lands up to 5 m tall in areas with seasonally high humidity. Its current pantropical distribution reflects human movement of plants since the 16th century, most likely because of fascination with the touch-sensitive rapid folding of the foliage. Mimosa pigra has naturalized beyond its native range in Asia (e.g., Thailand, Malaysia, America (e.g., Costa Rica, Brazil), Africa (e.g., Namibia, South Africa), and Oceania (e.g., Australia, New Guinea). Climatically, Mimosa pigra favors the seasonally dry tropical zone, with an annual rainfall of between 750 and 2250 mm, mean annual temperature above 17°C, and mild winters.

Characteristics that make Mimosa such an aggressive invasive plant species include rapid growth rate (10 mm per day), rapid maturation (germination to first flowering within 6 months), potentially autogamous, abundant seed production (9000 seeds per m² annually), a large, long-lived (>10 years) seed bank in soil, and an effective dispersal system (floation of clusters of capsules and via attachment to animals). Spines on the stem and leaf rachis deter most mammalian grazers and, outside its native range, there appear to be few invertebrate or fungal attackers. In many habitats, Mimosa pigra has the ability to completely dominate a site, forming impenetrable, monospecific shrub stands up to 6 m tall.

Rates and pattern of invasion have been best documented for northern Australia. Movement away from the entry point in Darwin has taken more than 50 years, but spread since the 1970s has been rapid. Within a river system stands expand on average at the rate of 7.6 m yr⁻¹, resulting in a doubling time of 1.4 years. Peripheral expansion is fastest following higher than average rainfall during the wet season, suggesting the importance of water for both the dispersal of seeds (in the wet season) and the survival of seedlings (through the dry season). An example of this spectacular rate of local spread has been given for a site of known history where a stand grew from a few individuals to one covering an area of 60,000 ha in approximately 10 years. At a larger scale, the doubling time for new infestations is slower (6.7 years), due to habitat heterogeneity.

The role of introduced grazing animals in the invasion of Mimosa pigra is debated. In northern Australia, the feral asinatic water buffalo (Bubalus bubalis) was thought by pastoralists to limit the spread of Mimosa, because of its conspicuous expansion since the lowering of buffalo densities as part of a disease (brucellosis and tuberculous) eradication program in the mid-1980s. However, in the two decades prior to this, both buffalo numbers and Mimosa pigra infestations increased in

FIGURE 9 Dense stands of Mimosa pigra during the wet season, northern Australia. Photo: CSIRO Entomology.
INTRODUCED PLANTS, NEGATIVE EFFECTS OF

parallel, and recent evidence shows that the most readily invaded habitats (wetland margins) are those most disturbed by large grazers, lacking tall trees, and with long periods of inundation.

*Mimosa pigra* drastically alters the composition and use of natural ecosystems. In Australia, it has supplanted native sedge land and grassland communities on flood plains and has invaded and displaced adjoining Melaleuca, Eucalyptus, and *Pandanus* woodland. The tall cover of *Mimosa* decreases native biodiversity and threatens the rich wildlife associated with open habitats. Ecosystem modifications caused by *Mimosa pigra* have reduced native resources accessible for traditional aboriginal use, pastoral grazing, and ecotourism ventures. The enormous scale and potential level of impact of *Mimosa pigra* is highlighted by estimates provided for Kakadu National Park (13,000 km$^2$), a World Heritage Area containing a full range of natural habitats within 150 km of the coast. Over 80% of the park is vulnerable to invasion (29% seasonally flooded areas susceptible to complete displacement, 54% largely open forest and woodland, which could have *Mimosa pigra* as a common element). Only 17% (extreme edaphic sites) of the area appears beyond the ecological tolerance of *Mimosa*.

The management and control of *Mimosa pigra* in natural ecosystems present a major challenge. It is susceptible to mechanical removal, herbicides, fire, and competition from grasses, and all are being used, often in combination, to eradicate new infestations and control the expansion in Australia. However, to be effective, treatments need to be repeated at regular intervals for at least a decade due to resprouting from damaged stumps and recolonization via the seed bank (10$^4$ seeds m$^2$). Currently an extensive biological control program is being undertaken testing several insects and pathogens.

**J. Pinus pinaster (Pinaceae)**

Members of the Pinaceae, notably *Abies*, *Larix*, *Picea*, and *Pinus*, are the most widely planted exotic tree species in the world and form the basis of production forestry in many countries. The evolutionary history of the family is centered on the northern hemisphere, but in the 20th century the geographic range of northern conifer species has been greatly extended to many southern hemisphere regions where they have been evaluated and used for forestry, amenity planting, reforestation, and erosion control. Translocation of *Pinus* species in particular has occurred on a grand scale, initiated by European settlers. In South Africa, for example, more than 80 of the 111 species worldwide have been planted since the late 17th century. Outside their native range, at least 20 species have established away from plantations into adjoining grassland, shrubland, and woodland communities. Because of their size, growth rate, and fecundity, pines have been responsible for generating widespread concern internationally about the potential impacts of introduced plant species. *Pinus* pinaster (cluster pine, maritime pine), native to the Mediterranean region, typifies the invasion history and types of impacts that pines can have, particularly in nonforest ecosystems.

Naturally occurring in the Mediterranean Basin from Portugal to Algeria, *Pinus pinaster* is a twoneedled, tall (40 m$^2$), open-canopied pine species frequently growing on sandy soils and dunes among oaks and heathland shrubs at low to moderate altitudes, but extending up to 1300 m. Valued for resin potential and for erosion control, *Pinus pinaster* has been planted widely in South America, Australasia, and South Africa, where it has invaded a range of natural and modified communities. Plantations have also been established in Europe, within its native range, and these are similarly spreading.

The ecology and spread of *Pinus pinaster* has been best documented in South Africa, where it was introduced in the late 17th century and subsequently widely planted for timber and drift sand stabilization. Two hundred years later *Pinus pinaster* has invaded 3256 km$^2$ of natural vegetation (Fig. 10) and is one of several pine species (*Pinus halepensis*, *Pinus radiata*) that threatens the conservation of one of world’s biodiversity gems, the fynbos biome. Extending for 71,337 km$^2$ in the Cape Floristic region, the fynbos biome comprises complex, species-rich (7,300 vascular plants), fire-adapted.

**FIGURE 10.** Thousands of hectares of *Pinus pinaster* after 100 years of spread in the Riviersonderend Mountains near Genadendal, Western Cape Province, South Africa. Photo: David Richardson.
mainly shrub-dominated ecosystems with a high proportion (80%) of local endemics. The pines have spread peripherally around source plantations and also via long-distance (up to 5 km) wind dispersed seed. These outliers gradually coalesce to form extensive stands in which Pinus pinaster is more than 23% of the cover. On the Cape Peninsula near Cape Town, pines are the most common introduced plant species in 10 of the 15 vegetation types represented at the Cape, with Pinus pinaster dominant over 560 ha or 1.1% of the total area.

The success of Pinus pinaster and other invading pines is due to a combination of factors including particular species traits, natural features of the fynbos, and various human activities. Weedy pines, in this ecological context, are distinguished from nonspreading species by their resilience to fire (e.g., short period to reproductive maturity, above-average levels of serotiny) and dispersibility (small seeds with low seed-wing loading). The fynbos vegetation is fire prone, and natural fires at intervals of a decade or more maintain a mosaic of shrub land and gully forest, depending on the timing and intensity of the burn. Fire frequencies have increased with greater human activity in the region, creating ideal invasion opportunities for fire-adapted, fast-growing, and early-maturing introduced species. Plants derived from abundant seed sources blown in from established plantations have penetrated the fynbos after fire, increasing (by up to 300%) the available flammable biomass, and gradually consolidating dominance with successive fire cycles.

In the absence of management, Pinus pinaster and the other pines would eventually transform much of the diverse fynbos vegetation into an introduced conifer forest. The shift in life-form dominance threatens some rare native plant species and will suppress many other light-demanding plants, although they may provide new habitats for arboreal native birds. The large increase in stand biomass caused by pine invasion has increased interception and transpirational losses resulting in decreased stream flow (by between 30 and 70%) for human consumption from pine-dominated catchments. The pines have induced a new vegetation structure and system, mainly as a result of increasing the intensity of burning, and few native species can cope with the new disturbance regime.

Management for the conservation of the fynbos biome must include controlling the invasive pines. Where plantations adjoin watersheds, massive clearing operations are financially justifiable in view of the major impact of invading pines on water delivery from catchments.

Options currently being undertaken involve prescribed burning at intervals of 12 to 15 years, in conjunction with mechanical clearing of pine stands and outliers. Pines are felled 12 to 18 months before burning to allow time for seeds to be released from serotinous cones, and seedlings to establish. These are subsequently killed by the fire. Biological control approaches using seed-attacking insects are being explored for Pinus pinaster and Pinus halepensis as long-term solutions for effectively managing pines in fynbos landscapes.

III. PREDICTING AND EVALUATING IMPACTS

Most countries want to limit the influx of potentially invasive plant species and to identify and contain those that already exist within the large reservoir of introduced plants confined to cultivated areas. The commonest approach to improving biosecurity at the border and enabling recognition of new weeds at a very early stage of invasion when control is both cost-effective and practicable is to apply some form of weed risk assessment analyses. These invariably include evaluations of potential impact on human activities (e.g., agricultural production, animal health) and natural environments. Information is commonly derived from weed databases on the behavior of the plant species elsewhere in the world or from assumptions about the relation between plant traits and species impact. In general, features that enable a species to either capture or produce an important resource will have the greatest effect on plant community composition and structure and will initiate major changes in ecosystem processes and functions. Key traits vary depending on the local environment and the type of land management, but those that either assist species to acquire an unequal share of key resources (e.g., relative growth rate, height), foster new disturbance regimes (e.g., fire), enhance local persistence (e.g., long-lived soil seed bank), or create a major increase in resource availability (e.g., nitrogen-fixing rhizobia) are likely to be the most important in any risk analysis. However, stochastic processes (e.g., change in land use), complexity of multiple interactions in natural ecosystems (e.g., presence of seed dispersers), and the increasing prevalence of introduced plant species lacking a weed history elsewhere collectively contribute to making weed impact prediction a major challenge. Methods for evaluating comparative benefits (e.g., food source for humans and native animals, erosion control) and costs (e.g., loss of crop production, reduced native biodiversity) of introduced invasive
weed species are only slowly emerging due to difficulties in assessing qualitative and quantititative elements of these categories in simple economic terms.

Acknowledgments
This article benefited from comments from Stephen Brown, Garry Cook, Curt Darlaker, Penny Edwards, Barrie Forrest, Peter Grubb, Simon Fowler, Mic Jeltema, Richard Mack, Wendy Nelson, Trevor Partridge, Ian Payton, Dave Richardson, Alan Rose, Leslie Seiger, and Peter Williams.

See Also the Following Articles
FIRES, ECOLOGICAL EFFECTS OF • INTRODUCED SPECIES, EFFECT AND DISTRIBUTION • MIGRATION • PLANT BIODIVERSITY, OVERVIEW • PLANT INVASIONS

Bibliography


INTRODUCED SPECIES,
EFFECTS AND
DISTRIBUTION OF

Daniel Simberloff
University of Tennessee

I. Introduction
II. The Geography and Magnitude of Invasion by Introduced Species
III. Direct Effects
IV. Indirect Effects
V. Invasional Meltdown
VI. Time Lags and Evolution
VII. Quantifying Effects

GLOSSARY

biological control Introduction of a natural enemy (parasite, predator, herbivore, or pathogen) of an undesirable species, usually itself introduced.
coevolution Evolved mutual adaptations between species, in which each species influences the evolution of the other.
hybridization Mating between individuals of two different populations, usually classified as separate species. Mixing of their genomes is not required.
introgression Gene flow between populations whose individuals hybridize, achieved when hybrids backcross to one or both parental populations.
island biogeography theory of A theory that the number of species in the biota of each island (or island-like habitat) is in dynamic equilibrium, with species frequently going extinct on the island and new species frequently arriving.

I. INTRODUCTION

What constitutes an effect of an introduced species (often called "impact") has never been formally defined. One common use of the term is for interactions with native species (e.g., predation), as well as consequences of these interactions for population, community, and ecosystem structure and function of the system to which the species was introduced. Another frequent referent for this term is the economic cost generated by an introduced species (including loss of goods or services and costs of control). A range of less frequent meanings of "effect" will be clear as effects are described. Nor is the meaning of "introduced species" clear-cut. It is widely used for any species introduced to a new region through human agency, whether deliberately or accidentally. Some authorities wish to restrict "introduced" to species deliberately imported by humans, and describe species that arrive with human assistance but that are not deliberately imported as "immigrants." In this restrictive definition, introduced species plus immigrants compose the category "nonindigenous species." A few authors also include as introduced species those that arrive on their own in a location far from their original homes—for example, the Old World cattle egret (Bubulcus ibis) in Florida. Finally, an introduced species is often defined legally as one not native to a particular nation; a species carried by humans from one location to another within a nation would not be introduced. Many other common terms—alien, exotic...
species, bioinvaders—are less precisely used. In this article, introduced species will be those that arrive in a new region with human assistance, deliberately or inadvertently, whether or not the new region is within the national borders of the original range. The effects of introduced species are numerous, often subtle, and idiosyncratic. Because intensive study of invasions by introduced species is quite recent, many statements about their effects are based on intuition rather than empirical data. Experimental data on effects of introduced species are rare. Intuition is sometimes incorrect. Thus, the decline of the native otter (Lutra lutra) in the United Kingdom was long attributed to the introduction of the American mink (Mustela vison), whereas more recent research exonerated the mink and implicated organochlorine pesticides. Conversely, the destruction of the forest birds of Guam was at first attributed to pesticide poisoning before the introduced brown tree snake (Boiga irregularis) was proven to be the cause.

Another complication in predicting and determining effects of introduced species is that, after introduction, there is frequently a time lag before they spread and begin to exert an impact. For example, Brazilian pepper (Schinus terebinthifolius), which now infests about 300,000 hectares (ha) in south Florida, was an innocuous, restricted species for over 50 years. Sometimes the causes of these lags seem evident (e.g., delay in the arrival of an obligatory pollinator for a plant); other times they are mysterious. The opposite phenomenon—dramatic increase in numbers and impact by invaders, followed by a decline independent of directed human actions—is also occasionally observed. For instance, on several Pacific islands that were massively invaded by the terrestrial giant African snail (Achatina fulica), a rather abrupt decline in density was observed. Again, causes sometimes suggest themselves (e.g., a pathogenic disease), other times they do not.

Despite uncertainties surrounding the effects of introduced species, a recent burst of research shows that they are often dramatic and, in sum, constitute a major, ongoing global change. They are widely recognized as the second greatest cause of species endangerment and extinction (after habitat change, with which they often interact). A recent estimate by D. Pimentel and others suggests that the annual cost in the United States alone is greater than $130 billion. This is not to say that most invaders have dramatic impacts. They do not; most do not even survive. M. Williamson's widely cited tens rule (about 10% of introduced species establish without human assistance, and about 10% of those have effects large enough to be viewed as pests) is certainly too simplistic, but it accurately depicts the lack of substantial impact by most invaders.

II. THE GEOGRAPHY AND MAGNITUDE OF INVASION BY INTRODUCED SPECIES

A. How Many Are There, Where Do They Come From, and Where Do They Go?

No comprehensive list of introduced species exists for most taxa in most locations; indeed, there is often no list of native species. However, for well-studied groups, some figures are impressive. In Florida, for example, 27% of established plant species, 8% of insects, 29% of land snails, 24% of freshwater fishes, 22% of reptiles, 9% of birds, and 24% of land mammals are introduced. For the Hawaiian Islands, as for many islands, some of the analogous figures are even greater: almost half the plants, 23% of insects, most freshwater fishes, and 40% of birds. In some areas (e.g., Alaska) there are far fewer introduced species, but almost no regions are immune. Although the absence of adequate quantitative data makes it difficult to describe this pattern fully, it is widely believed that Eurasian species are more likely to invade other regions than vice versa and are more likely to have large impacts. For instance, most major human pathogens and most plant pathogens that have had global impacts originated in Eurasia. Similarly, more Eurasian insects, vertebrates, and plants have invaded other regions than vice versa. The reasons for this imbalance are obscure. Some authors have argued that Eurasian species have an innate superiority, generated either by the larger numbers of species evolving greater competitive ability on the larger landmass, or by the happenstance that Eurasian species were highly coevolved—plants, pathogens, and animals (including especially grazers and humans)—and overwhelmed native species by their joint action. For example, the structure and behavior of Eurasian hoofed livestock were devastating to native tussock grasses in the North American prairie but were favorable for Eurasian turfgrasses, which now dominate vast regions. However, even if Eurasian species, singly or in groups, were not innately superior, one might have expected a preponderance of them among introductions, because the opportunities for such species to reach other regions were greater throughout most of recent history. For instance, the majority of introduced insects in the United States through the eighteenth cen-
turkey came from Europe in soil ballast, which was loaded in Europe and exchanged in North America for various raw materials. Similarly, as Europeans colonized other regions, they formed acclimatization societies to introduce the birds of their homelands. Except for game birds, there was little analogous movement of species in the opposite direction. Nowadays, as travel and trade are burgeoning worldwide, opportunities for introduction from any region to any other one are greatly enhanced, and one might expect habitat and climatic matching to become more important as limiting factors.

B. When Did They Get There?
The timing of introductions has depended heavily on available means of transport and patterns of travel and trade, and it has tended to increase strongly from the late eighteenth century through the present. For example, introductions of aquatic plants and animals into the Great Lakes rose steadily from one species between 1810 and 1839 to over 80 between 1960 and 1990. The advent of rapid steamship transport across oceans coincided with a dramatic increase in introductions of many taxa, as hitchhikers that could not have stayed alive over a month or more in transit were able to survive a voyage of two weeks across the Atlantic. Air travel decreased the need to survive a long transit period still further. Overlaying this dominant pattern of increasing rates of introduction with increased transport volume and decreased transit time are idiosyncrasies associated with particular taxa and regions. For example, beginning around 1920, there was a decrease in the rate at which foreign insect species, especially herbivorous species, were introduced to the United States; this downturn coincided with the enforcement of plant quarantine laws. At the same time, a dramatic increase of introduction of wasp species reflected greatly increased biological control efforts, especially the use of parasitic wasps.

C. Distribution of Introduced Species among Habitats
Habitats modified and/or routinely disturbed by humans generally house more introduced species and larger populations of introduced species than do pristine habitats, though even the latter are occasionally invaded. The reason for this pattern has been hotly debated. In general, the very fact of human disturbance renders a habitat less suitable to the native species that had evolved in the original habitat. On the other hand, some species, often associated with humans, are superbly adapted to habitats that humans create. Thus, certain introduced plants are routinely found on lists of serious weeds from many parts of the world. A common claim that disturbed habitats are invasible because they are species-poor is incorrect. Many habitats that are relatively undisturbed by humans, such as salt marshes and mountaintops, have few native species, but they also have few introduced species, and for the same reason—they are biologically difficult environments and few species of any kind have the physiological adaptations that permit them to thrive there. Conversely, enormously diverse tropical communities often have as many invaders (with as large impacts) as less species-rich temperate analogs. Nor does periodic disturbance per se automatically lead to invasiveness. For example, the intact upland pine forests of north Florida are far less stricken by invaders than many other habitats in the state, yet they frequently burn naturally because of lightning strikes. If anything, this disturbance favors the many native species that are adapted to a fire regime over a plethora of potential invaders that are not.

Most well-known examples of introduced species with enormous impacts come from either terrestrial or freshwater habitats. Marine habitats are poorly represented in the invasion literature. Freshwater habitats—lakes and rivers—can often be seen as habitat islands surrounded by land, and thus their communities are believed to be inherently invasible for the same reasons that islands appear to be particularly invasible, as will be discussed in the following section. Certainly, many lake and river communities have evolved without certain types of organisms, such as large, predatory or folivorous fishes, and thus their species are particularly prone to damage from introduced species of these particular sorts. Whether marine habitats have suffered less impact from introduced species is highly questionable. Terrestrial habitats are immediately visible, so the effects of introduced species are surely more likely to be detected there than in the sea, just as the general ecology of most marine habitats is poorly studied and understood relative to that of most terrestrial habitats. Additionally, it is now widely recognized that many marine species that had been viewed as native over wide regions are in fact of unknown provenance—they are “crypto- genetic.” In most places they may well have been accidentally introduced by humans. Finally, a number of marine invasions have been observed lately that vie in apparent impact with the most heralded terrestrial and freshwater invasions. The tropical alga Caulerpa taxifolia, purged from the public aquarium in Monaco in the early 1980s, has spread to cover 5000 ha of nearshore
habitats off the coasts of Spain, France, Italy, Monaco, and Croatia. It has particularly overgrown and replaced meadows of marine grasses such as Posidonia oceanica, the nursery and home of many fishes and invertebrates. In the Black Sea, the western Atlantic ctenophore Mnemiopsis leidyi was first recorded in 1982. By 1988, its biomass exceeded that of all other zooplankton combined, and, because this comb jelly is a predator, its impact on other animals, including young fishes, must be enormous. It is widely credited with the dramatic crash of several commercial fisheries.

In the Black Sea, the western Atlantic ctenophore Mnemiopsis leidyi was first recorded in 1982. By 1988, its biomass exceeded that of all other zooplankton combined, and, because this comb jelly is a predator, its impact on other animals, including young fishes, must be enormous. It is widely credited with the dramatic crash of several commercial fisheries.

D. Island Vulnerability

Islands are notoriously prone to ecological damage by introduced species. Many island bird species and sub-species have been eliminated by introduced rats, feral cats, and other introduced carnivores. In the West Indies, the small Indian mongoose (Herpestes javanicus) has extinguished several species of endemic snakes and lizards. Introduced grazers have caused many plant extinctions on islands. The introduced ctenophore Mnemiopsis leidyi, first recorded in 1982. By 1988, its biomass exceeded that of all other zooplankton combined, and, because this comb jelly is a predator, its impact on other animals, including young fishes, must be enormous. It is widely credited with the dramatic crash of several commercial fisheries.

...
smothers beds of native submerged vegetation. Decaying water hyacinth can deposit over 1000 metric tons (wet weight) of detritus per hectare per year, which in turn can heavily modify water chemistry, such as nutrient concentrations. When dissolved oxygen drops, it affects many animals and the entire biotic community changes. Terrestrial plants such as Asian kudzu (Pueraria montana) in the Southeast can similarly cover existing vegetation and eliminate it by shading, as can marine plants such as the tropical alga Caulerpa taxifolia in the Mediterranean.

3. Modified Fire Regime
In much of the American West and Hawaii, Old World grasses such as cheatgrass (Bromus tectorum) increase the frequency and intensity of fires, greatly harming native plants and the animals that use them. Similarly, the Australian melaleuca (Melaleuca quinquenervia) has a spongy outer bark and highly flammable foliage, and it also produces a more abundant litter than native herbaceous communities in south Florida. These features have produced an altered fire regime that has facilitated the invasion of 200,000 ha. Both tree and ground fires are now more intense and frequent, to the detriment of the native plants.

4. Modified Hydrology
Mediterranean salt cedars (Tamarix spp.) have invaded the U.S. Southwest. They are deeply rooted and transpire rapidly; once established, they can survive on water deep in the soil, and their transpiration is a significant pathway of water loss in arid areas. For example, at Eagle Borax Spring in Death Valley, California, within 25 years of invasion by salt cedar, the surface water of what had been a large marsh had disappeared completely, along with the majority of its associated biota. In Israel, Australian Eucalyptus trees have been deliberately used to drain swamps and bogs, thus eliminating the original vegetation and animals dependent on it. In addition to changed evapotranspiration rates, plants can also affect hydrological regimes by changing soil elevation. The heavy litter of melaleuca has had this impact in the Florida Everglades.

5. Modified Nutrient Regime
Many plant invaders fix nitrogen, and in nitrogen-poor areas this added increment can be detrimental to native species, which have evolved to thrive in a low-nitrogen environment, and favor other invaders. The Atlantic nitrogen-fixing shrub Myrica faya has invaded young, nitrogen-deficient volcanic regions of Hawaii. There are no native nitrogen-fixers, and the invader thus alters productivity, nutrient cycling, and ecological succession. Because many nonindigenous plant species in Hawaii are more successful on more fertile sites, M. faya may enhance the likelihood of other invasions.

B. Competition
Competition for resources is often difficult to demonstrate, but there is strong reason to believe that introduced species often use some resource so effectively that they deprive native species of it. In addition, an introduced species can also depress a native one not by eliminating some resource but by direct interference.

1. Resource Competition
Shading and water depletion, already discussed, are means by which some introduced plants outcompete native species for a crucial resource. Introduced animals have similar effects. For instance, the house gecko (Hemidactylus frenatus) has invaded many Pacific islands, and it depresses the insect food base locally so that some native lizard populations decline. In Great Britain, the greater foraging efficiency of the introduced American gray squirrel (Sciurus carolinensis) has led to the decline of the native red squirrel (S. vulgaris).

2. Interference Competition
The South American fire ant (Solenopsis invicta), which has spread throughout much of the southeastern United States, attacks individuals of many native ant species and is replacing some species, such as the "native" fire ants (perhaps themselves pre-Columbian introductions) S. geminata and S. xyloni in several habitats. In a plant analog of aggression, the African crystalline ice plant (Mesembryanthemum crystallinum) accumulates salt, and the salt remains in the soil when the plant decomposes. In California, this ice plant excludes native plants that cannot tolerate the salt. In both of these examples, the invader does not actually render a resource in short supply for native species; rather, it directly inhibits the native.

C. Predation
Some of the most striking effects that invaders have had on particular native species, or groups of them, entail their preying on them. The Nile perch (Lates nilotica), introduced to Lake Victoria, eliminated many species of endemic cichlid fishes. Introduced rats (Rattus spp.) on many islands have destroyed at least 37 species and subspecies of island birds worldwide. The brown tree snake and the rosy wolf snail, mentioned...
Herbivory by introduced European rabbits (Oryctolagus cuniculus) has caused enormous damage on islands worldwide. Their impact in Australia is legendary. The most serious damage there is by the stripping and killing of seedling trees and perennial shrubs. In addition, rabbitbits often cause extensive erosion.

Innumerable crop plants, most of them introduced, have been devastated by introduced insects. Damage in the United States from the Russian wheat aphid (Diuraphis noxia) alone exceeded $600 million between 1987 and 1989. The South American cassava mealybug (Pseudococcus manihoti) has invaded most cassava-growing regions of Africa and causes yield losses of up to 84%. Similarly, introduced insects attack both native and introduced trees, often with staggering impact. Losses in eastern U.S. forests to the European gypsy moth (Lymantria dispar) were estimated at $764 million in one year alone. The Asian balsam woolly adelgid (Adelges piceae) has eliminated the Fraser fir (Abies fraseri) in many areas of the southern Appalachian Mountains.

Herbivorous insects have been employed in biological control projects against many introduced aquatic and terrestrial weeds. Among striking successes are the use of the South American cactus moth (Cactoblastis cactorum) against pest prickly pear (Opuntia spp.) in Australia and the South American alligatorweed flea beetle (Agasicles hygrophila) on aquatic alligatorweed (Alternanthera philoxeroides) in Florida. In each instance the weed had covered vast areas before the insect introduction, which has subsequently limited it to small, ephemeral infestations. Such introductions of insects for control of introduced weeds have occasionally led to threats to the very existence of nontarget native species. For example, the cactus moth, introduced to control pest prickly pear on the island of Nevis in the West Indies, has reached the Florida Keys, where it attacked every individual of the native semaphore cactus (Opuntia spinosissima). The Eurasian weevil Rhinocyllus conicus, introduced to North America to control Eurasian pest thistles, especially musk thistle (Cirsium arvense), attacks several native thistles, including the endangered Suisun thistle (Cirsium hydrophilum var. hydrophilum). The key to these unplanned threats to native species, in each instance, is that the introduced insect could maintain high numbers on alternative hosts (generally the target species), so that decline of the native species did not induce a decline in populations of the herbivore.

D. Herbivory

The extinction of half of the native flora of St. Helena by goats shows that herbivory by introduced species on native ones can have major effects. There are numerous other examples in a variety of habitats. Chinese grass carp (Ctenopharyngodon idella) have been introduced to Europe, Africa, and North America for the control of aquatic macrophytes. Although the targeted plants are generally introduced species, the grass carp is not highly selective and often prefers native plants. It is so voracious that it can completely eradicate nontarget species from particular water bodies before foraging on the target species. Herbivory by introduced European rabbits (Oryctolagus cuniculus) has caused enormous damage on islands worldwide. Their impact in Australia is legendary. The most serious damage there is by the stripping and killing of seedling trees and perennial shrubs. In addition, rabbitbits often cause extensive erosion.

Innumerable crop plants, most of them introduced, have been devastated by introduced insects. Damage in the United States from the Russian wheat aphid (Diuraphis noxia) alone exceeded $600 million between 1987 and 1989. The South American cassava mealybug (Pseudococcus manihoti) has invaded most cassava-growing regions of Africa and causes yield losses of up to 84%. Similarly, introduced insects attack both native and introduced trees, often with staggering impact. Losses in eastern U.S. forests to the European gypsy moth (Lymantria dispar) were estimated at $764 million in one year alone. The Asian balsam woolly adelgid (Adelges piceae) has eliminated the Fraser fir (Abies fraseri) in many areas of the southern Appalachian Mountains.

Herbivorous insects have been employed in biological control projects against many introduced aquatic and terrestrial weeds. Among striking successes are the use of the South American cactus moth (Cactoblastis cactorum) against pest prickly pear (Opuntia spp.) in Australia and the South American alligatorweed flea beetle (Agasicles hygrophila) on aquatic alligatorweed (Alternanthera philoxeroides) in Florida. In each instance the weed had covered vast areas before the insect introduction, which has subsequently limited it to small, ephemeral infestations. Such introductions of insects for control of introduced weeds have occasionally led to threats to the very existence of nontarget native species. For example, the cactus moth, introduced to control pest prickly pear on the island of Nevis in the West Indies, has reached the Florida Keys, where it attacked every individual of the native semaphore cactus (Opuntia spinosissima). The Eurasian weevil Rhinocyllus conicus, introduced to North America to control Eurasian pest thistles, especially musk thistle (Cirsium arvense), attacks several native thistles, including the endangered Suisun thistle (Cirsium hydrophilum var. hydrophilum). The key to these unplanned threats to native species, in each instance, is that the introduced insect could maintain high numbers on alternative hosts (generally the target species), so that decline of the native species did not induce a decline in populations of the herbivore.

E. Parasitism and Disease

Introduced pathogens of various sorts have had enormous impacts. The most damaging to entire ecosystems
are plant pathogens that affect dominant native plants; in so doing, they affect entire communities that interact with these plants. The Asian chestnut blight fungus (Cryphonectria parasitica) reached New York City in nursery stock in the late nineteenth century. In less than 50 years, it spread over about 100 million ha of the eastern United States, destroying the aerial parts of almost all mature chestnut (Castanea dentata) trees. Because chestnuts had been the most common tree in many forests, the ecosystem impacts of this invasion were almost certainly enormous, although few crucial data were gathered. Several insect species that are host-specific to chestnut went extinct, and nutrient cycling rates were probably greatly modified.

A pathogen of animals that affected many entire ecosystems was the virus rinderpest, native to India. Introduced to Africa in cattle in the 1890s, it infected many native ungulate species, with mortality in certain species reaching 90%; the distribution of some species remains affected a century later. Because of the crucial role played by ungulates in aspects of vegetation structure and dynamics, the impact of rinderpest far surpassed the death of individuals it infected.

Many introduced diseases have heavily affected individual species or small groups of them without apparent major impact on entire ecosystems. Ranges of Hawaiian native birds have been drastically circumscribed by habitat destruction, but avian malaria, caused by Plasmodium relictum capristranoae and vectored by introduced mosquitoes, afflicts the remaining populations and helps restrict them to upper elevations. The plasmodium was introduced with Asian songbirds. The European fish parasite Myxosoma cerebralis causes whirling disease in salmonid fishes. Following World War II, live North American rainbow trout (Salmo gairdneri) were transferred freely among European sites. There they acquired the parasite from the native brown trout, S. trutta, which has significant resistance. Eventually, frozen European rainbow trout were widely exported. The parasite probably reached North America in imported frozen fishes. Either these were fed accidentally to fish in a trout hatchery in Pennsylvania, or the viscera may have been discarded in streams near the hatchery. In any event, hatchery fishes became infected, and these were shipped to many other states. In large parts of Montana and Colorado, the great majority of rainbow trout contract the disease, and sport fisheries have collapsed in both states.

Introduced macroparasites also often have major impacts. The parasitic plant witchweed (Striga asiatica) probably invaded North Carolina from Africa with military equipment after World War II. It attacks primarily grasses, including corn, and is a sufficient agricultural pest that it has been the target of a long eradication campaign and associated quarantine. The trematode Cyathocotyle bushiensis, which causes heavy mortality in ducks, has advanced along the St. Lawrence River recently, concurrently with its invasive introduced intermediate host, the Eurasian water snail Bithynia tentaculata.

Many parasitic wasps and flies, as well as some disease pathogens, have been introduced as biological control agents for introduced pests. For example, the yellow clover aphid (Thysanips trifolii) is controlled in California by three parasitic wasps, Pastra palhens, Trioxys utilis, and Aphelinus semijulatus. Populations of the cas-sava mealybug in Africa, discussed earlier, have been greatly reduced by the introduced South American parasitic wasp Epidinocarsis lopezi. Perhaps the best-known biological control pathogen is the New World myxoma virus, introduced to mainland Europe (where the European rabbit is native), Great Britain, and Australia (where it is introduced). Initial epizootics caused over 90% mortality in Great Britain and Australia, and over 90% in France. The initial virulent strains in all three countries largely evolved to more benign strains and, at least in Great Britain and Australia, rabbits evolved a degree of resistance. Thus, successive epizootics caused decreasing mortality, until an equilibrium appeared to have been reached. A number of pathogenic fungi have been introduced to control weeds, with varying degrees of success.

F. Hybridization and Introgression

Introduced species can eliminate native species by mating with them, a particularly strong threat when the native species is not as numerous as the introduced one. Both the New Zealand grey duck (Anas superciliosa superciliosa) and the Hawaiian duck (A. wyvilliana) are thus threatened by extensive hybridization and introgression with the North American mallard (A. platyrhynchos), introduced as a game bird. Similarly, the white-headed duck (Oxyura leucophaea), now restricted in Europe to Spain, is threatened there by hybridization and introgression with North American ruddy ducks (O. jamaicensis), which were introduced to Great Britain as an amenity, escaped, and eventually reached Spain.

Both plants and animals are threatened by such introgression, and its extent is just becoming known with the extensive use of molecular techniques that can detect it. This problem is much more common in regions that exchange closely related species, such as Europe and
North America, than in those with species so distantly related that they are unlikely to be able to mate and exchange genes, such as Australia and either Europe or North America. Exchange of genes is not even necessary for hybridization with an introduced species to affect a native species inimically. Many females of the endangered European mink (Mustela lutreola) hybridize with male introduced American mink (M. vison), which become sexually mature earlier than the male European mink. The embryos are all aborted, but the loss of reproduction by the European mink exacerbates their population decline.

IV. INDIRECT EFFECTS

The foregoing effects of introduced species are direct effects of various sorts. The actions of individuals of the introduced species are directly on individuals (and often, ultimately, on populations) of one or more native species; they may attack them, eat them, poison them, infect them, and so on. However, an introduced species can also affect other species indirectly in many ways.

A. Classic Indirect Effects

A strictly defined indirect effect occurs when one species alters the interaction between two others. For example, the chestnut blight led to the replacement of chestnut in much of the eastern United States by oak species. Red oak (Quercus rubra) increased greatly, and this species is particularly susceptible to oak wilt disease (Ceratocystis fagacearum). The increase in oak wilt disease in turn, raised the frequency of oak wilt disease on many less susceptible native oak species. Thus, the chestnut blight fungus indirectly affected the interaction of oaks and oak wilt disease, in addition to its direct effect on chestnut.

Classic indirect effects can be highly complex. For example, the mite Pemotex ventricosus, accidentally introduced to Fiji, attacked the larvae and pupae, but not the eggs and adults, of the coconut leaf-mining beetle (Promocothea corallopennis). Adult beetles then oviposited and died, so the beetle population came to have synchronous, nonoverlapping generations. The absence of larvae and pupae for extended periods caused the mite population to plummet, as did those of two native parasitoids that had previously controlled the beetle; they did not live long enough to survive the intervals between occurrences of the host stages needed for oviposition. So the mite indirectly affected the interaction between the beetle and the parasitoids.

B. Chain Reactions

A number of the introductions discussed earlier affected one species directly, but this direct effect generated subsequent impacts in chain-like fashion. Thus, the direct effect of chestnut blight on American chestnut led indirectly to the extinction of several insect species that were host-specific on chestnut. The modification of nutrient cycles as a chestnut-dominated litter was replaced by litter dominated by leaves of oaks and other species almost certainly affected populations of litter inhabitants as well as plants, but these effects were never investigated. Rabbits in Australia directly affected certain plant species, as noted earlier, and also probably directly contributed to the elimination of two small burrowing marsupials, the boodie rat (Bettongia lesueurii) and the bilby (Macrotis lagotis) by competition for burrows. However, rabbits indirectly led to the elimination of the common wombat (Vombatus ursinus) from part of its range by modifying succession and locally eliminating native perennial plants. On coastal islands, the erosion and vegetation changes caused by rabbits have been highly detrimental to seabirds that use these islands for resting and breeding.

Some chain reactions induced by introduced species are so complicated that predicting them would have been difficult. Caterpillars of the large blue butterfly (Maculina arion) in Great Britain required development in underground nests of the ant Myrmica sabuleti. The ant cannot nest in overgrown areas. Changing land use patterns and reduced livestock grazing left introduced rabbits as the main grazer maintaining the habitat. The biological control introduction of myxoma virus reduced rabbit populations sufficiently so that ant populations declined, and the butterfly went extinct.

Chain reactions can generate effects far from the site of the initial introduction. For example, landlocked kokanee salmon (Oncorhyncus nerka) were introduced to Flathead Lake, in western Montana, in 1916 and largely replaced native cutthroat trout (O. clarki) as the dominant sport fish. By 1931, the kokanee were spawning in McDonald Creek (Glacier National Park), some 100 km upstream from Flathead Lake, and the spawning population was so large that it soon attracted large populations of bald eagles, grizzly bears, and numerous other predators. The catch of kokanee rose to over 100,000 fishes per year through 1985. Between 1968 and 1975, the opossum shrimp (Mysis relicta), native to several large, deep lakes in North America, became sexually mature earlier than the male European mink. The embryos are all aborted, but the loss of reproduction by the European mink exacerbates their population decline.

INTRODUCED SPECIES, EFFECTS AND DISTRIBUTION OF
INTRODUCED SPECIES, EFFECTS AND DISTRIBUTION OF

and Sweden, was introduced to three lakes in the upper portion of the Flathead catchment in a misguided attempt to enhance kokanee productivity. The shrimp drifted downstream and reached Flathead Lake by 1983. There was an immediate, dramatic decline in the densities of copepods and especially cladocerans preyed on by the shrimp. This competition with the kokanee for prey caused the kokanee population to decline rapidly; there was no catch at all in 1988 and 1989. Bald eagle numbers fell precipitously, as did those of grizzly bears and several other predators.

C. Vectoring of Pathogens

In a number of instances discussed earlier, an introduced species carried a pathogen that greatly affected one or more native species—rinderpest, whirling disease, avian malaria. In some cases, these could be classified as classic indirect effects. For instance, in Hawaii the avian malaria plasmodium changed the interaction between introduced mosquitoes and native forest birds to the detriment of the latter. The impact of an introduced pathogen can be more complex. For example, the introduced species that carries it to a new region need never be sympatric with a native species ultimately affected. Consider the introduction of the grass carp to the United States. The carp was introduced to Arkansas in 1968; it spread to the Mississippi River, carrying with it a parasitic tapeworm from Asia, Bothriocephalus acheilognathi. The tapeworm quickly infested other fishes, including the red shiner (Notropis lutrensis), a popular bait fish. Fishermen or bait dealers introduced infected red shiners to the Colorado River, and by 1984 they had reached the Virgin River, a Utah tributary. There they infected the woundfin (Plagopterus argentissimus), a native minnow already threatened by dams and water diversions. Tapeworm infections caused woundfin numbers to decline precipitously, possibly because they are less able to compete with the red shiner, which have some resistance to the tapeworm. Another introduced species, the Asian myna bird (Acridotheres tristis), introduced to control pasture insects, has dispersed the New World weed Lantana camara widely in the lowlands of the Hawaiian Islands, including into native forest areas.

V. INVASIONAL MELTDOWN

Some ecological theories suggest that introduced species should interfere with one another and thereby lessen one another’s impact. The theory of island biogeography, for example, implies that each successive species in a series introduced to an island (or a habitat island) has a lower probability of surviving. Biological control practitioners have argued about whether releasing several species of potential natural enemies of a single pest might lessen the overall impact on that pest because of competition among them. However, positive interactions between introduced species are detected as often as negative ones. Sometimes one introduced species facilitates the existence and enhances the effect of another; other times the combined effects of two or more introduced species exceed the sum of what the same species might have accomplished individually. Such situations, in which different introduced species enhance one another’s effects, are collectively termed “invasional meltdown.” The existence of many such cases suggests that an accelerating wave of invasions with ever-increasing effects may characterize the future in many regions.

A. Introduced Animals Pollinating and Dispersing Introduced Plants

Many introduced fig (Ficus) species are frequently planted as ornamentals in South Florida. If the wasp that pollinates a fig species in its native range is absent, the fig cannot reproduce. Thus, until recently, figs in Florida remained where planted. However, within the past 15 years, breeding populations of three host-specific pollinating wasps have been introduced to Florida, and the species they pollinate now regularly produce seeds. Ficus microcarpa is spreading most quickly and has become an invasive weed, the small fruits of which are dispersed by birds and ants. The spread of the same complex of F. microcarpa trees and its wasp pollinator Parapristina verticillata is also occurring in Bermuda, Mexico, and Central America.

Some introduced animals disperse introduced plants that disrupt native plant communities. The red-whiskered bulbul (Pycnonotus jocosus) has dispersed a number of alien species, such as Rubus alceifolius, Cordia interrupta, and Ligustrum robusum on the island of La Réunion. Cordia was not viewed as a problem until it was widely distributed by the bulbul; the bird is the primary dispersal agent for Ligustrum. In the Hawaiian Islands, introduced pigs selectively eat and thus disperse several invasive introduced plant species, and their rooting and defecation favor several introduced invertebrates. Further, the pigs grow larger because of the presence of introduced, protein-rich European earthworms in the soil. The Asian myna bird (Acridotheres tristis), introduced to control pasture insects, has dispersed the New World weed Lantana camara widely in the lowlands of the Hawaiian Islands, including into native forest areas.

INTRODUCED SPECIES, EFFECTS AND DISTRIBUTION OF

and Sweden, was introduced to three lakes in the upper portion of the Flathead catchment in a misguided attempt to enhance kokanee productivity. The shrimp drifted downstream and reached Flathead Lake by 1983. There was an immediate, dramatic decline in the densities of copepods and especially cladocerans preyed on by the shrimp. This competition with the kokanee for prey caused the kokanee population to decline rapidly; there was no catch at all in 1988 and 1989. Bald eagle numbers fell precipitously, as did those of grizzly bears and several other predators.

C. Vectoring of Pathogens

In a number of instances discussed earlier, an introduced species carried a pathogen that greatly affected one or more native species—rinderpest, whirling disease, avian malaria. In some cases, these could be classified as classic indirect effects. For instance, in Hawaii the avian malaria plasmodium changed the interaction between introduced mosquitoes and native forest birds to the detriment of the latter. The impact of an introduced pathogen can be more complex. For example, the introduced species that carries it to a new region need never be sympatric with a native species ultimately affected. Consider the introduction of the grass carp to the United States. The carp was introduced to Arkansas in 1968; it spread to the Mississippi River, carrying with it a parasitic tapeworm from Asia, Bothriocephalus acheilognathi. The tapeworm quickly infested other fishes, including the red shiner (Notropis lutrensis), a popular bait fish. Fishermen or bait dealers introduced infected red shiners to the Colorado River, and by 1984 they had reached the Virgin River, a Utah tributary. There they infected the woundfin (Plagopterus argentissimus), a native minnow already threatened by dams and water diversions. Tapeworm infections caused woundfin numbers to decline precipitously, possibly because they are less able to compete with the red shiner, which have some resistance to the tapeworm. Another introduced species, the Asian myna bird (Acridotheres tristis), introduced to control pasture insects, has dispersed the New World weed Lantana camara widely in the lowlands of the Hawaiian Islands, including into native forest areas.

V. INVASIONAL MELTDOWN

Some ecological theories suggest that introduced species should interfere with one another and thereby lessen one another’s impact. The theory of island biogeography, for example, implies that each successive species in a series introduced to an island (or a habitat island) has a lower probability of surviving. Biological control practitioners have argued about whether releasing several species of potential natural enemies of a single pest might lessen the overall impact on that pest because of competition among them. However, positive interactions between introduced species are detected as often as negative ones. Sometimes one introduced species facilitates the existence and enhances the effect of another; other times the combined effects of two or more introduced species exceed the sum of what the same species might have accomplished individually. Such situations, in which different introduced species enhance one another’s effects, are collectively termed “invasional meltdown.” The existence of many such cases suggests that an accelerating wave of invasions with ever-increasing effects may characterize the future in many regions.

A. Introduced Animals Pollinating and Dispersing Introduced Plants

Many introduced fig (Ficus) species are frequently planted as ornamentals in South Florida. If the wasp that pollinates a fig species in its native range is absent, the fig cannot reproduce. Thus, until recently, figs in Florida remained where planted. However, within the past 15 years, breeding populations of three host-specific pollinating wasps have been introduced to Florida, and the species they pollinate now regularly produce seeds. Ficus microcarpa is spreading most quickly and has become an invasive weed, the small fruits of which are dispersed by birds and ants. The spread of the same complex of F. microcarpa trees and its wasp pollinator Parapristina verticillata is also occurring in Bermuda, Mexico, and Central America.

Some introduced animals disperse introduced plants that disrupt native plant communities. The red-whiskered bulbul (Pycnonotus jocosus) has dispersed a number of alien species, such as Rubus alceifolius, Cordia interrupta, and Ligustrum robusum on the island of La Réunion. Cordia was not viewed as a problem until it was widely distributed by the bulbul; the bird is the primary dispersal agent for Ligustrum. In the Hawaiian Islands, introduced pigs selectively eat and thus disperse several invasive introduced plant species, and their rooting and defecation favor several introduced invertebrates. Further, the pigs grow larger because of the presence of introduced, protein-rich European earthworms in the soil. The Asian myna bird (Acridotheres tristis), introduced to control pasture insects, has dispersed the New World weed Lantana camara widely in the lowlands of the Hawaiian Islands, including into native forest areas.
B. Introduced Plants That Modify Habitat

As noted in a previous section, the Atlantic nitrogen-fixing shrub Myrica faya may facilitate invasion by other introduced plants that are currently limited by the nitrogen-poor volcanic soil and absence of native nitrogen-fixers. This plant also enhances populations of introduced earthworms, which in turn increase the rate of nitrogen burial and thus exacerbate the impact of the plant on nitrogen cycling. The soil drying by introduced salt cedars already described favors the introduction of nonnative grasses in both the U.S. Southwest and Australia. The increased soil salinity produced by the African crystalline ice plant, described previously, does more than kill native plants. When wind or other disturbances create holes in the carpet of ice plant, they are colonized not by native plants but by the ice plant itself or by weedy introduced plants such as Malva parviflora or Erodium cicutarium.

Many invasive introduced plants increase fire frequency and/or intensity to the detriment of native species but to their own advantage and that of other invaders. Old World grasses have come to dominate many New World grasslands through this facilitation. For instance, in Hawaii the introduced perennial grass Schizachyrium condensatum invaded seasonal submontane shrub-dominated woodland. It fostered much more frequent fires over larger areas, killing most native trees and shrubs. But S. condensatum recovered quickly and the even more flammable introduced perennial grass Melinis minutiflora also invaded.

C. Introduced Animals That Modify Habitat

The disturbance wrought by large, congregating, introduced herbivores in North America favored the establishment of Eurasian grasses adapted to such animals, and helped them to replace native tussock grasses. Similarly, the Asian water buffalo (Bubalus bubalis), introduced to northeastern Australia as a beast of burden and for meat, spread throughout the floodplain of the Adelaide River by the late nineteenth century. It devastated native plant communities, compacted the soil, eroded creek banks, and altered hydrology. A Central American shrub, Mimosa pigra, had been a minor, non-invasive weed in the vicinity of Darwin for a century. This legume produces large numbers of small seeds that are readily dispersed by water. The sputtering of the flood-plains by the water buffalo created ideal germination habitat for M. pigra seedlings. In many areas, native sedgelands have been converted to a monoculture of M. pigra.

The Caspian zebra mussel (Dreissena polymorpha) arrived in the Great Lakes in the 1980s. Its huge numbers and great filtering ability have greatly affected native freshwater communities over much of the eastern and midwestern United States. In particular, they convert large amounts of seston into excreted feces, creating a soft substrate of rich organic material. Among species whose populations have increased as a result of the zebra mussel is the invasive Eurasian faucet snail discussed previously. The filtering also increases water clarity, and this change has favored certain invasive macrophytes, such as Eurasian watermilfoil (Myriophyllum spicatum).

D. Mutualisms

A number of instances in which one introduced species facilitates the existence, spread, and population growth of another introduced species constitute mutualisms, in that the latter species also aids the former. For example, introduced macrophytes like Eurasian watermilfoil provide additional settling substrate for the mussel, and they can also help the mussel disperse between water bodies. Mimosa pigra aided water buffalo by protecting them from aerial hunters trying to eradicate them. Fig wasps and their associated fig species obviously form coevolved mutualistic introduced species pairs. In the Hawaiian Islands, introduced African big-headed ants (Pheidole megacephala) tend an introduced scale insect, Coccus viridis; this is a classic mutualism. The scale is on the introduced plant Pluchea indica; among other things, the ant hinders introduced predatory coccinellid beetles and parasitic wasps. In this one case, a range of interactions are taking place among an entire assemblage of introduced species. Sometimes, as with M. pigra and water buffalo, the mutualists cannot have co-evolved, as they originate on different continents. The big-headed ant in Hawaii also tends the South American gray pineapple mealybug (Dysmicoccus neobrevipes), a particularly unwelcome pest because it helps spread a wilt disease of pineapple.

VI. TIME LAGS AND EVOLUTION

Effects of introduced species may change dramatically, for example, when an invader is quiescent during a time lag, after which it rather abruptly spreads and increases in number. As observed in Section I, such time lags complicate efforts to predict the effects of an
INTRODUCED SPECIES, EFFECTS AND DISTRIBUTION OF

Introduced species and are sometimes mysterious. It is frequently suggested that a mutation in an invader could account for a particular time lag, and that evolution in general could greatly change the effects of an introduced species. In no instance can a sudden increase in population size and impact of an introduced species be clearly attributed to a mutation, but there is ample evidence that such species evolve in their new homes. How frequently and to what extent does such evolution change their effects?

Morphological evolution of introduced species is often apparent. The small Indian mongoose, whose depredations of native island species were discussed earlier, has become larger and more sexually dimorphic on all islands to which it has been introduced. This change is probably selected for in the absence on the islands of the slightly larger gray mongoose (Herpestes edwardsi), with which the small Indian mongoose is sympatric in the region of origin for the island populations. The North American muskrat (Ondatra zibethicus), introduced as a fur bearer to Europe in 1905, has spread throughout much of Europe in less than a century. The muskrat is well known in its native range as a species with substantial variation in morphology (especially size), to the extent that various subspecies have been named. In Europe, it has already evolved approximately the same degree of variation. Similarly, the European house sparrow (Passer domesticus) was introduced in New York City in 1853 and quickly spread to become one of the most common birds in North America, covering much of the continent. It has evolved so much that distinct “races” are now easily identified in different parts of its introduced range. Many plant species are known to differ substantially between their native and introduced ranges, though there is generally far less evidence than in the vertebrate cases just cited that the differences are genetic. What is rarely if ever known in these demonstrated instances of evolution after introduction is how these changes affect the impacts of these species on native communities and ecosystems.

The hybridization of an introduced species with a native one can even lead to the formation of a new species, which can be invasive. For example, North American cordgrass (Spartina alterniflora), introduced in shipping ballast to southern England, occasionally hybridized with a noninvasive native congener, S. maritima. These hybrids were sterile, but eventually one hybrid individual underwent a doubling of chromosome number to produce a fertile new species, S. anglica, which turned out to be highly invasive.

The evolution of the introduced rabbit and myxoma virus in Australia and Great Britain certainly affected the impact of the rabbit on native systems. The fact that the rabbit became somewhat more resistant to the virus, and that the virus evolved to be somewhat more benign, lowered the rabbit mortality in each successive epizootic, thus the degree of various effects outlined earlier. As pathogens and their hosts often coevolve, changes of impact might be expected. They are not always seen, however. Chestnut blight has not become perceptibly less devastating to American chestnut, nor avian malaria to native Hawaiian birds.

Many introduced pest insects have evolved resistance to chemical pesticides and thus generate much greater impacts, generally on agricultural or silvicultural plants. Just as strains of human pathogens have evolved resistance to whole sequences of antibiotics, some pest insects have evolved resistance to several insecticides. This resistance can arise in three ways: insects can evolve to tolerate greater amounts of the chemical, they can evolve physiological means of detoxifying the chemical, or they can evolve behavioral traits (such as going to the bottom of a leaf instead of the top) that help to avoid contact with the chemical. Introduced weedy plants also evolve resistance to herbicides, though this phenomenon has not been as widely studied as insect resistance.

The evolution of resistance to chemicals by introduced pests gives impetus to the desire to control them biologically, through the introduction of natural enemies. However, this approach has fostered concerns about subsequent evolution. For pathogens, one concern is that the target organisms will evolve resistance to them. Over 90% of the market for insect pathogens today consists of products involving two organisms, the bacterium Bacillus thuringiensis and heterorhabditid nematodes. Both types have no natural association with insect pests on plants and thus have to be routinely reapplied. It is possible that target pest insects could become resistant to them simply through repeated exposure plus selection of resistant strains. This possibility is heightened for the bacterium by the fact that B. thuringiensis toxins are now being genetically engineered into crop plants, thus increasing the rate of natural selection. On the other hand, introduced biocontrol insects can evolve greater adaptation to a potential host and thus become more rather than less lethal. In the United States, the Egyptian alfalfa weevil (Hypera brunneigennis) was originally quite immune to the European ichneumonid parasitic wasp Bathyplectes cucullatus, encapsulating 35–40% of the eggs and larva in an immune response. Fifteen years later, only 5% of the eggs were encapsulated.
A second concern is that introduced biological control agents could evolve to switch hosts. For introduced insects, the fact that host range can be controlled by a single gene only heightens this concern, as does the fact that certain features of most biological control introductions favor fast evolution. These features include initially small population size, rapid population growth in new environments, and novel, abundant resources. In light of the possibility, it might seem surprising that few examples of host switching are known. Because many of the pathogens used or contemplated as biological control agents have broader host ranges than parasitic insects, the probability of host-switching accompanied or followed by evolutionary adaptation to the new host is enhanced.

VII. QUANTIFYING EFFECTS

"Effect" and "impact" have never been formally defined in invasion biology, so that rankings of introduced species in terms of greatest impact are impressionistic. Area of occupancy is a frequently used index of effect, although the number of individual invaders per unit area and what each individual actually does would be important (though more difficult to measure). Impacts on ecosystem processes, such as nutrient cycling and fire frequency, are seen by some biologists as the best measures of effect of introduced species, on the grounds that such alteration of processes could affect a large number of species simultaneously. Both areal coverage and impact on processes often envision outcome for native species, communities, and ecosystems as the crucial measure of effect. The real effect, in these terms, would require accurate knowledge of the range size of the species, the distribution of abundance of species within this range, and the impact on the native species, both on-site and elsewhere, per individual invader (or per unit biomass of the invader). The first two factors can be measured or at least estimated in straightforward fashion, although the effort for some species would be enormous. The per capita impact, on the other hand, would probably entail not only detailed observation but also experiments, perhaps over the long term, at individual, population, community, and ecosystem levels. As one example of the subtleties that might be involved, the fact that an introduced predator is observed to eat individuals of a native species does not necessarily mean that this predation affects the population size, behavior, or ecology of the prey species. Ecological methods are well enough developed that these sorts of questions can be answered, but they generally require detailed research, not quick observation.

Economic costs of an invasion are often proposed as estimates of the effect of introduced species, but these costs are more easily tallied from a human standpoint than from that of native species, communities, and ecosystems. It is a fairly straightforward matter to tabulate the costs expended in controlling a particular weed or insect pest, in terms of materiel and personnel used. And one could relatively easily add the monetary cost of many services or goods lost because of an introduced species—say, a percentage of some crop. Costs of other lost services are not so easily tabulated, even in human terms. For example, how expensive is the loss in enjoyment suffered by citizens who can no longer walk through Fraser fir or American chestnut forests in eastern forests because these were eliminated by introduced species? Economists have attempted to measure these costs. For example, travel costs of citizens to a particular natural site, as evidence of how much they are willing to pay for such an experience, are sometimes used to estimate the "value" of a site, at least to humans. But all of these methods are controversial. Furthermore, costs to nonhuman species are even harder to weigh in monetary terms. What was the cost, in dollars, to the insect species extinguished because of the chestnut blight?

In sum, even though a bewildering variety of effects of introduced species can be demonstrated, measuring and comparing these effects for different species remains as much a conceptual and practical challenge as predicting the effects.

See Also the Following Articles

CORRELATION • FIRE • ECOLOGICAL EFFECTS OF • GRAZING • EFFECTS OF • HUMAN IMPACT ON BIODIVERSITY • OVERVIEW • INTRODUCED PLANTS, NEGATIVE EFFECTS OF • ISLAND BIOGEOGRAPHY • SPECIES COEXISTENCE

Bibliography

Biological Invasions (journal, Volume 1, 1999).


INVERTEBRATES, FRESHWATER, OVERVIEW

Margaret A. Palmer* and P. Sam Lake†
*University of Maryland and †Monash University, Australia

I. Introduction

II. Factors Influencing Biodiversity in Fresh Waters

III. The Role of Invertebrates in Ecological Processes

IV. Threats to Freshwater Fauna

GLOSSARY

benthic At or on the bottom of streams, lakes, or riverbeds.
decomposition The breakdown of organic matter such as dead plants and algae to release carbon and nutrients.
endemic species A species that is unique to a particular locality.
exotic species A species that has been introduced to an area outside of its native range.
groundwater A "reservoir" of water residing below ground in saturated soils and beneath geologic formations.
larvae Early stages in the development of an organism; for invertebrates, often morphologically quite distinct from the adult.
local species richness Number of species found at a local site; distinguished from regional or global species richness that "sums" the number of species across a number of individual sites.
planktonic Organisms that reside in the water column.
watershed A geographical region in which water drains into common water bodies.

WATER IS THE MOST abundant substance on earth and is essential for all of life. It is critical to the world’s climate, to the cycling of nutrients, and is a habitat for much of the earth’s biodiversity. Fresh water links the land and oceans via groundwater and riverine flow. While invisible to the human eye, groundwaters represent a substantial portion of the fresh waters (approx. 30%) acting not only as the most important reservoir of fresh water on earth, but as a home to many unique fauna. Lakes and reservoirs are also important stores of fresh water and harbor several groups of extremely abundant invertebrates. Running waters and wetlands are known to be havens for diverse assemblages of invertebrates many of whom are important in ecological processes that ensure clean water. The amount and timing of water inputs to many freshwater systems worldwide is changing rapidly. Since many invertebrates require specific habitats and flood regimes, their abundance and diversity in many parts of the world are at risk because habitats are being lost and natural flow regimes altered at astounding rates. Also contributing to the recent declines in freshwater invertebrates are poor water quality and the introduction of exotic species that have led to the extinction of native species, especially bivalves and crayfish.

I. INTRODUCTION

A. Major Freshwater Habitats

The major freshwater habitats include running waters (streams and rivers), standing or semistanding waters
INVERTEBRATES, FRESHWATER, OVERVIEW

FIGURE 1 Schematic illustration of major freshwater habitats. Note all water bodies, including lakes, streams and rivers, wetlands, and the oceans, are connected via the groundwaters. See also color insert, Volume 1.

(reservoirs, lakes, ponds, wetlands), and groundwater (Fig. 1). Most of the 33 million cubic kilometers of fresh water on the earth (≈ 69%) exists in the form of polar ice. Of the remaining fresh water, a huge volume exists out of sight as groundwater (Table I). Lakes, streams, and wetlands represent less total water volume but are biologically rich hot spots for abundant and diverse assemblages of invertebrates. Together, the freshwater bodies form a continuous “pipeline” that links the atmosphere, land, and the oceans. Rainwater is intercepted by vegetation on land, percolates through the soils into groundwaters, and then moves into wetlands, lakes, and streams before it eventually reaches the oceans (Fig. 1). Since groundwater directly or indirectly “feeds” all other water bodies on earth, maintenance of adequate quantities of high-quality water in this underground reservoir is a high priority worldwide.

### B. Major Groups and Their Lifestyles

A disproportionate number of species live in fresh waters compared to marine systems: oceans comprise 70% of the earth’s surface while inland waters make up about 1%. Despite this, about 15% of all animal species alive today live in fresh water. Most of the freshwater animal species are invertebrates. More than 70,000 species have been described but many more remain undiscovered—new species and genera are found every year. Freshwater habitats have representatives from most taxonomic groups with the insects being particularly speciose (Table II). The freshwater insects are also quite notable for

<table>
<thead>
<tr>
<th>TABLE I: Freshwater Distribution by Continent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continent</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Africa</td>
</tr>
<tr>
<td>Europe</td>
</tr>
<tr>
<td>Asia</td>
</tr>
<tr>
<td>Australia</td>
</tr>
<tr>
<td>North America</td>
</tr>
<tr>
<td>South America</td>
</tr>
</tbody>
</table>

Modified from the World Conservation Monitoring Center’s Freshwater Biodiversity: A preliminary global assessment (1998). Data are volume of water in km³; wetlands includes marshes, swamps, lagoons, and flood-plains.
### TABLE II

A List of Some of the Most Common Freshwater Invertebrates with Estimates of the Number of Described Species Globally

<table>
<thead>
<tr>
<th>Common name</th>
<th>Aquatic stage</th>
<th>Existence</th>
<th>Global diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylum Porifera</td>
<td>Sponges</td>
<td>SW, RW</td>
<td>B</td>
</tr>
<tr>
<td>Phylum Cnidaria</td>
<td>Hydra, jellyfish</td>
<td>SW, (RW)</td>
<td>B, P</td>
</tr>
<tr>
<td>Phylum Platyhelminthes</td>
<td>Flatworms, Planaria</td>
<td>RW, SW, GW</td>
<td>B</td>
</tr>
<tr>
<td>Phylum Annelida</td>
<td>Leeches</td>
<td>RW, SW</td>
<td>B</td>
</tr>
<tr>
<td>Class Turbellaria</td>
<td>Earthworms</td>
<td>RW, SW, GW</td>
<td>B</td>
</tr>
<tr>
<td>Phylum Nematomorpha</td>
<td>Roundworms</td>
<td>RW, SW, GW</td>
<td>B</td>
</tr>
<tr>
<td>Phylum Gastrotricha</td>
<td>Wheel-animalcules</td>
<td>RW, SW, GW</td>
<td>B, P</td>
</tr>
<tr>
<td>Phylum Tardigrada</td>
<td>Water bears</td>
<td>SW, (RW)</td>
<td>B</td>
</tr>
<tr>
<td>Phylum Arthropoda</td>
<td>Class Crustacea</td>
<td>O. Cladocera</td>
<td>Water fleas</td>
</tr>
<tr>
<td></td>
<td>O. Ostracoda</td>
<td>Seed shrimp</td>
<td>SW, GW, (RW)</td>
</tr>
<tr>
<td></td>
<td>O. Copepoda</td>
<td>SW, GW, RW</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>O. Isopoda</td>
<td>Sow bugs</td>
<td>SW, RW, GW</td>
</tr>
<tr>
<td></td>
<td>O. Amphipoda</td>
<td>Scuds</td>
<td>SW, GW, RW</td>
</tr>
<tr>
<td></td>
<td>O. Decapoda</td>
<td>Crabs, shrimp, crayfish</td>
<td>SW, RW, GW</td>
</tr>
<tr>
<td>Class Arachnida</td>
<td>Subclass Acari</td>
<td>Mites</td>
<td>SW, GW, RW</td>
</tr>
<tr>
<td>Class Insecta</td>
<td>O. Ephemeroptera</td>
<td>Mayfly larvae</td>
<td>RW (SW),</td>
</tr>
<tr>
<td></td>
<td>O. Odonata</td>
<td>Dragonfly and damselfly larvae</td>
<td>RW, SW</td>
</tr>
<tr>
<td></td>
<td>O. Plecoptera</td>
<td>Stonefly larvae</td>
<td>RW, (SW)</td>
</tr>
<tr>
<td></td>
<td>O. Hemiptera</td>
<td>True bugs</td>
<td>RW, SW, (GW)</td>
</tr>
<tr>
<td></td>
<td>O. Trichoptera</td>
<td>Caddisfly larvae</td>
<td>RW, SW</td>
</tr>
<tr>
<td></td>
<td>O. Lepidoptera</td>
<td>Butterfly and moth larvae (caterpillars)</td>
<td>SW, RW</td>
</tr>
<tr>
<td></td>
<td>O. Coleoptera</td>
<td>Beetles</td>
<td>SW, RW</td>
</tr>
<tr>
<td></td>
<td>O. Megaloptera</td>
<td>Dobsonfly larvae</td>
<td>SW, RW</td>
</tr>
<tr>
<td></td>
<td>O. Neuroptera</td>
<td>Spongillaflies on sponges</td>
<td>SW, RW, GW</td>
</tr>
<tr>
<td></td>
<td>O. Diptera</td>
<td>Fly larvae (e.g., mosquitoes, midges)</td>
<td>SW, RW, GW</td>
</tr>
<tr>
<td>Phylum Mollusca</td>
<td>Class Gastropoda</td>
<td>Snails</td>
<td>SW, RW, GW</td>
</tr>
<tr>
<td></td>
<td>Class Bivalvia</td>
<td>Clams, mussels</td>
<td>SW, RW</td>
</tr>
</tbody>
</table>

Some groups are more common in running waters (RW), such as streams and rivers while others are more common in standing or semistanding water (SW) bodies (lakes, swamps, wetlands) or groundwater (GW). Some are primarily benthic (B), existing in or on the bottom of lake, stream, river sediments, or on aquatic vegetation, while others maintain a water column existence as members of the plankton or pelagic fauna (P). Parentheses are used when an organism may be found in a habitat type but is less common there than in the other habitats.
INVERTEBRATES, FRESHWATER, OVERVIEW

FIGURE 2 Example of a complex life cycle for a freshwater invertebrate. Many freshwater bivalves are benthic yet produce larvae that are released into the water column. These develop into small glochidium larvae that attach to a fish host. After some time, the larvae metamorphose into small bivalves, detach, and take on a benthic existence.

the large number whose lifestyle changes with age: many have aquatic larval stages but flying adults. This is in contrast to freshwater mollusks and crustaceans who complete their life cycles in water either as members of the plankton or benthos.

Life cycles of freshwater invertebrates can be quite complex with multiple, morphologically distinct stages (Fig. 2). For example, among the freshwater bivalves, some are hermaphroditic while others have separate sexes; some brood their eggs and then release small clams; however, many release larvae into the water. Zebra mussels produce a planktonic larval stage (a veliger larva), while some of the most common freshwater native bivalves have unique parasitic larvae (glochidia larvae) that attach to the skin, scales, or gills of fish. The host fish disperses the attached glochidium larvae, prior to the detachment of the larvae and their metamorphosis into adult bivalves. This dispersal phase may be very important to the long-term survival of bivalves particularly in freshwater bodies that are being stressed.

Damselflies and dragonflies provide other examples of complex life cycles typical of many freshwater invertebrates. These insects have flying adults that live weeks to months and lay eggs on aquatic vegetation or other substrates. The eggs hatch within 12 to 30 days, or they can remain viable in an unhatched state (diapause) if there is a drought. Small larval stages (sometimes called nymphs) hatch from the eggs and then molt multiple times (10–16) before crawling up vegetation and “emerging” as flying adults. The presence of an intact riparian zone with a diversity of native plants promotes population persistence of insects.
C. Number of Freshwater Invertebrate Species

The number of freshwater invertebrates varies considerably between lakes, running waters, groundwaters, and wetlands. Since the number of species varies vastly from locale to locale as a function of many complex factors, estimates of species richness should be taken as only approximations. Additionally, freshwater invertebrate diversity at any one time may be moderate while accumulated diversity over time may be very high. For example, there may be dramatic declines in invertebrate diversity in many streams during the flood season but high levels of diversity across the entire year. Similarly, temporary wetlands may have much higher total diversity over time than permanent wetlands even though the snapshot diversity in a temporary wetland may be considerably lower. Many freshwater wetlands are extremely speciose habitats (e.g., up to 2000 invertebrate species at a locale), followed by lakes and streams (80–1400 species of invertebrates typically found at a locale), and then groundwaters (0–130 at a locale). In all these habitats, insects are by far the most speciose group with local species richness levels of 50 to 500 and global richness levels of probably at least 45,000 species. Crustaceans may attain local species richness values of up to 150 and global richness values of 8000 species; water mites up to 75 species locally and 5000 species globally; annelids up to 50 species locally and 1000 species globally; mollusks and rotifers up to 500 species locally and 6000 species globally; and mollusks up to 100 species locally and 4000 globally.

Some freshwater habitats are hot spots of endemism harboring many unique fauna. For example, ancient lakes such as Lake Baikal in Siberia has a rich abundance of endemic fauna, especially for those species that live all of their lives in fresh water and have a poor capacity for dispersal. For one group of crustacean amphipods, the gammarids, this lake has the highest level of endemism in the world (41 genera, of which 38 are endemic). Other lakes that harbor highly endemic invertebrate faunas include the African rift lakes of Lake Tanganyika and Lake Nyasa, Lake Titicaca in Peru, and Lake Lanao in the Philippines.

D. Dominant Taxonomic Groups across Freshwater Habitats

In running waters, the dominant invertebrate groups vary considerably depending on the type of bottom substrate (e.g., mud, cobble/boulder) and the flow (Fig. 3). In rough-bottom creeks and swift-flowing streams, invertebrates live primarily on or beneath pebbles and boulders on the bottom and include many species of insects, crustaceans, and mollusks. Many of these benthic species ("scrapers") are adapted to feed on decomposing plant material that has fallen into the stream. Further downstream in the watershed, the channel widens and more sunlight reaches the bottom so that invertebrate species composition changes. Here we find animals that feed on decomposing plants and animals that feed on algae ("scrapers") growing on the rocks and sides of channels. Even further downstream in the watershed, in deep riverine areas, fine sediments predominate, the neabed waters flow more slowly, and planktonic invertebrates become common in the water column. In the soft sediments, burrowing insect larvae, bivalves, and worms are common as long as oxygen is present.

In lakes and ponds, the most common invertebrates include crustaceans, rotifers, insects, and oligochaetes; however, taxonomic composition varies dramatically depending on the size of the lake and the position within the lake (Fig. 4). Benthic invertebrates are particularly abundant and speciose along the lake margins but generally are less numerous in the deeper parts of the lakes where oxygen may be limiting and habitat diversity is much lower. However, a few species that are tolerant of silt and low oxygen such as oligochaetes, midge larvae, and nematodes may be extremely abundant in the deeper part of lakes. In the water column, abundant planktonic invertebrate communities thrive by feeding on phytoplankton, but these invertebrate assemblages typically are dominated by just a few taxonomic groups such as rotifers, copepods, and cladocerans.

In the groundwater, crustaceans, rotifers and nematodes are very common, and many of these are adapted for adept movement among rocks and particles of sand. In some parts of the world, insect larvae are common in the groundwater and may be found great distances away from the nearest surface water. Important crustacean groups in groundwater include copepods and less commonly amphipods, ostracods, isopods, and decapods. Many of these live in caves and sinkholes. Groundwater invertebrates are particularly interesting because of the unique adaptations many of them possess for life in dark, isolated habitats where food may be scarce. For example, the Edwards Aquifer in Texas harbors a unique fauna of 22 species including 10 amphipod species. Many of these species have reduced (or no) eyes, reduced body pigmentation, reduced body size, low metabolic rates, and an enhanced sensitivity to touch.
In freshwater wetlands, the number of habitats and types of food sources for invertebrates are vast. The soft sediments are filled with burrowing nematodes, oligochaetes, midge larvae (a dipteran insect), and mites. Some invertebrates feed on or around the roots of vegetation, while others burrow through the sediments consuming bacteria, fungi, or other invertebrates. In the overlying water, crustaceans, rotifers, and insect larvae (e.g., mosquito larvae) are common and feed on phytoplankton or invertebrate prey. The stems of emergent and submerged vegetation in wetlands are typically crawling with diverse assemblages of crustaceans, mollusks, and insects. Small shrimp-like crustaceans (amphipods) may reach very high levels of abundance on aquatic vegetation. Water beetles, Odonates (dragonflies and damselflies), and waterbugs (water boatmen, waterstriders, and backswimmers) are commonly the top predators.

**E. Global Patterns in Freshwater Biodiversity**

Knowledge of lineage histories is very important because regional patterns of distribution of species are often a reflection of distinctly different evolutionary histories. Adaptations to specific climates and habitat characteristics influence speciation and the ability of an organism that evolved under one set of circumstances to move broadly among different geographic areas. Terrestrial and marine biologists generally have found the highest levels of animal and plant diversity in the tropics with lower levels in temperate and polar regions. This
Species richness for freshwater invertebrates is related in very complex ways to latitude and elevation and global patterns of diversity are not straightforward for this group. There have been far fewer surveys of freshwater invertebrates in the tropics than in temperate regions; however, those to date suggest that biodiversity in fresh water is not typically higher in low latitude, tropical habitats. For example, in North America, bivalves and crayfish appear to be most diverse in temperate zones (about 260 native bivalves and 320 species of crayfish). Temperate zones have a wealth of crayfish species, but tropical fresh waters have a richer shrimp and freshwater crab fauna. Insects, mayflies, and stoneflies are more diverse in temperate latitudes than in the tropics, whereas dragonflies, water bugs, and water beetles are much more diverse in the tropics.

II. FACTORS INFLUENCING BIODIVERSITY IN FRESH WATERS

Scientists have devoted lifetimes to the study of what determines the number of species in a given habitat type or geographic region. These studies have involved thoughtful systematic observations of patterns of distribution dating back many years (e.g., Robert Pennak’s work in many freshwater systems), and more recent experimental studies that have cleverly manipulated some factor believed to influence biodiversity. From such studies, we know that biodiversity in fresh waters is determined by both historical (e.g., age of a lineage, evolutionary constraints) and contemporary factors (e.g., habitat heterogeneity, predator-prey interactions). The relative importance of the many factors varies geographically as well as between habitats in the same geographic region.

A. The Origin of Freshwater Invertebrates

Most freshwater invertebrates are believed to be derived from terrestrial or marine ancestors. Representatives from many different groups of marine crustacea (shrimp, copepods, cladocera) may have made the transition from salt water to fresh water through many years of evolution in increasingly brackish (low salinity) water. Success in fresh water required that the animals be able to maintain ionic balance in their body fluids (not gain too much body water), reproduce without...
the aid of complex tidal currents to carry eggs and larvae, and, in some cases, develop life stages that could withstand freezing or drying. Only certain species could make this transition, so at the phylum level, freshwater biodiversity is more limited than in the oceans. For example, echinoderms (e.g., starfish) did not develop the ability to osmoregulate in fresh water.

Many freshwater invertebrates are secondary aquatic forms that moved, in an evolutionary sense, into fresh waters from the land. The most notable group of secondary invaders is the insects, many of which still have terrestrial adults. Some insect orders, such as the Hemiptera, Coleoptera, and Diptera, have both many freshwater species and many terrestrial species. Hemiptera and Diptera also have a few highly specialized marine species. Water mites (Acari) coevolved with insects and many species are parasites on adult insects. With their coevolutionary links, they are also secondary invaders of fresh waters. Freshwater snails consist of two major groups, the prosobranchs and the pulmonates. Pulmonates have invaded freshwater systems from the land; freshwater forms include the two common and ubiquitous families of pond snails, the Lymnaeidae and Physidae.

B. Habitat Age and Isolation

Some freshwater habitats are extremely old and have been isolated for very long periods of time. These habitats have a high degree of endemism. For example, Lake Baikal in Siberia is extremely old and has many species that are found no where else in the world. Such “hot spots” of unique species are considered valuable for conservation purposes and for use as a “natural laboratory” for the study of invertebrate evolution.

For those ancient freshwater bodies that are highly isolated, resident fauna may be at high risk of extinction. If a freshwater species is highly specialized for conditions at a particular size and conditions change, the species may be unable to survive. Geographic or habitat remoteness is not the only factor promoting isolation of freshwater invertebrates. Some fauna are isolated because they have very limited dispersal abilities. For example, many groundwater and deep-streambed dwelling species are very poor swimmers and lack dispersal stages. Thus local biodiversity may be low even though regional biodiversity is much higher—species from the large-scale regional “pool” of invertebrates (e.g., all rivers in a geographic region) may not be able to move between all watersheds. In general, ecologists assume that local species diversity represents a subset of the regional pool of species and that the size of this subset depends greatly on both dispersal abilities of fauna and geographic barriers to dispersal.

C. Habitat Heterogeneity and Disturbance

A central tenet of ecological theory is that the greater the number and types of habitats, the greater the species diversity. Put simply, different types of organisms with different niche requirements can exist in areas that are heterogeneous over space and time. In freshwater ecosystems, spatial heterogeneity (and species diversity) is enhanced by anything that leads to a great variety of particle sizes in which benthic organisms can burrow and feed. Similarly, spatial heterogeneity and species diversity of planktonic invertebrates is enhanced by anything that promotes the growth of many types of vegetation because this provides the invertebrates with diverse food types and hiding places from predators.

Environmental variation over time (temporal heterogeneity) is also extremely important as a determinant of species diversity. Variable climate and flow regimes will promote seasonal (or longer) changes in species abundances and may determine whether a species can coexist in a location. Some freshwater invertebrates have evolved life cycles to take advantage of or avoid predictable changes in water level, flood flows, or ice scours. However, when these environmental changes become extreme or highly unpredictable (a “disturbance”), species richness may decline as fauna are unable to evolve coping strategies that keep up with such changes. Many ecologists believe that invertebrate species diversity is greatest when natural disturbances like floods are at intermediate levels with respect to severity or predictability. High levels of disturbance will wipe out most species while low levels of disturbance may favor one or two species to the exclusion of other species that are poorer competitors.

D. Species Interactions

Predation and competition are well known to influence local species diversity in freshwater ecosystems. For example, many fish predators are highly selective in their feeding and may forage on a single species or suite of species. Thus, predation pressure may reduce species diversity locally, or it actually may promote species diversity if the predator is removing a species that is a competitive dominant. Competitive dominants are species that are so good at garnering resources (space, food, etc.) that other species are unable to coexist with them. The process of competitive exclusion also may
be prevented by disturbances if the disturbance (for example, a flood or drought) has a disproportionate effect on the competitive dominant.

In freshwater ecosystems, some of the most dramatic examples of the delicate balance that species interactions play in determining biodiversity come from the introduction of exotic species. Many exotic fish have been introduced into lakes throughout the world and have had devastating effects on their invertebrate prey or on other fish who in turn prey on invertebrates. Introduced carp have reduced invertebrate populations in floodplain wetlands. Non-native trout introduced to streams have depleted populations of native fish and increased trout predation pressure on native invertebrates.

III. THE ROLE OF INVERTEBRATES IN ECOLOGICAL PROCESSES

Healthy freshwater ecosystems are those in which ecological processes continue unimpeded to ensure that water is clean and that organic matter in a lake, stream, or wetland is not lost or accumulated in excess. Freshwater invertebrates play key roles in these processes. The most important ecological processes in freshwater systems include the decomposition of organic matter; the uptake and transfer of materials such as sediments, nutrients, and contaminants; and production by plants.

A. Decomposition of Organic Matter and Recycling of Nutrients

Decomposition releases elemental carbon and nutrients (nitrates, phosphates) to the environment with the result that plants and animals that rely on these nutrients can thrive. Additionally, decomposition ensures that wastes and dead organic material do not build up in bodies of water. Terrestrially produced leaf and woody litter and dead aquatic vegetation and plankton sink to the bottom of lakes, streams, and wetlands. Once deposited within fresh waters, this organic matter can be broken down directly by bacteria and fungi. Often however, the material first must be altered by invertebrates, who fragment or shred it before ingesting it. The shredding activities of invertebrates are important because smaller fragments of dead leaves and other organic matter are decomposed by microbes much faster than larger pieces. Additionally, through their feeding activities, invertebrates can enhance the abundance and reproduction of the microbes, which further acts to stimulate the rates of decomposition.

B. The Uptake and Transfer of Materials

Movement of water, sediment, and organic material within and from freshwater systems has a profound effect on global biological, geological, and chemical processes. Freshwater invertebrates may alter the rate at which water flows through sediments. For example, burrowing oligochaetes, nematodes, midge larvae, and crustaceans can increase the rate at which water moves through lake and stream sediments and to nearby groundwaters. Instead of increasing water flow and percolation through sediments, some freshwater invertebrates bind sediments by producing mucar "biofilms" or dense mats of tubes or filaments that reduce water flow and the movement of sediments. This may act to improve water clarity so that sunlight penetrates deeper and plant production is enhanced. Similarly, water clarity may be enhanced by dense populations of bivalves that filter water as they feed and remove suspended matter. Freshwater invertebrates also may influence the concentration of contaminants in the water or bottom sediments by accumulating the contaminant in their bodies or via direct degradation of toxic materials.

C. Production by Plants

In freshwater ecosystems, the synthesis of organic matter by plants depends on the presence of sunlight and nutrients. Rates of primary production are extremely variable across freshwater ecosystems, ranging from low in poorly lit groundwaters or turbid lakes to quite high in well-lit shallow wetlands. The effect that invertebrates have on plant production is indirect but significant. The availability of nutrients for aquatic vegetation and phytoplankton depends to a great extent on the decomposition of organic matter in aquatic sediments and the movement of nutrients. Grazing and movement of invertebrates stir up the bottom sediments and mix water, increasing the availability of nutrients to plants and phytoplankton. Furthermore, planktonic invertebrates that graze on phytoplankton excrete forms of phosphorus and nitrogen that are immediately available to further enhance algal growth. Grazing may be particularly important in the removal of senescent algae, thereby increasing light as well as nutrients.

IV. THREATS TO FRESHWATER FAUNA

The loss of freshwater invertebrate biodiversity because of damage generated by human activities is an immense global problem. A major difficulty in assessing the di-
mensions of the problem is the poor state of knowledge of freshwater invertebrate taxonomy and distribution and the lack of reliable monitoring to identify loss of diversity. In spite of these problems, it is clear that invertebrate species in some places already have high levels of population decline and extinction.

The major threats to freshwater biota in general, and freshwater invertebrates in particular, are habitat degradation and loss, pollution and reduced water quality, altered hydrologic regimes, and invasion by exotic species. In many situations, the major threats do not act alone but can act synergistically. Thus, for example, streams in agricultural areas may lose their invertebrate fauna due to habitat loss by sedimentation, pollution by pesticides, altered flow regimes due to irrigation, and the introduction of exotic plants and fish. A fifth threat is overexploitation. For example, harvesting of the mussel Margaritifera margaritifera for pearls has reduced populations by 90%.

A. Loss and Depletion of Habitat
Humans have damaged water bodies by river channelization and dredging, sand and gravel extraction, wetland drainage, lake and river shore development, dam and barrier construction, water diversion, and levee bank construction. Human activities on land can have indirect but significant effects on fresh waters. For example, sedimentation from poor land use, destruction of riparian vegetation, loss of surface runoff, and loss of water by groundwater extraction for irrigation all can damage fresh waters. Temporary wetlands are common in many regions. They may have surface water only occasionally, but when they do they are very productive and harbor a distinctive invertebrate fauna adapted to survive long dry periods. Unfortunately, in many areas the importance of temporary wetlands as distinctive habitats of high biodiversity value has not been recognized and many of these wetlands have been plowed, drained, and filled in.

Sedimentation of streams and wetlands from erosion can be caused by activities in watersheds such as land clearance, plowing and tillage of soil, road building, and logging. The mass delivery of sand and silt into water bodies can reduce invertebrate diversity and change species composition by filling in pools in streams, wetlands, ponds, and even lakes by burying porous coarse sediments with layers of fine sediments with very low permeability and by covering aquatic plants. The deep streambed of running waters has a rich and distinctive fauna dependent on the high permeability of the gravels and sands. Sedimentation can clog the pores in the streambed reducing water movement and oxygen availability, which leads to the loss of fauna. Many streams that originally had stable, habitat-rich channels and a rich invertebrate fauna have been damaged when channels were filled by fine shifting sediment that consequently has led to a greatly depleted fauna.

A property vital to the nature of rivers, and essential for maintenance of invertebrate biodiversity, is connectivity—the unimpeded movement of water, longitudinally from source to mouth, and laterally between the channel and its floodplain. Loss of connectivity greatly alters invertebrate species composition and may reduce diversity. Dams on rivers are barriers to longitudinal connectivity because they disrupt the downstream movements of nutrients and organic matter and prevent movement of invertebrates and fish. Leves are barriers to lateral connectivity because they are designed to stop lowland rivers from flooding their floodplains; however, periodic floodplain inundation is essential for maintenance of the biodiversity of the river system.

B. Pollution and Reduced Water Quality
Pollutants can enter fresh waters through point sources such as sewage outfalls, or via diffuse nonpoint sources such as runoff from agricultural fields. Pollutants in sewage and agricultural runoff include organic matter and nutrients. Organic pollution of rivers usually is released from point sources, and downstream from the point source, dissolved oxygen concentrations may drop greatly. The normally diverse fauna of the river is eliminated and replaced by an abundance of a few species that can tolerate low oxygen levels. As the organic matter is processed, the river may recover further downstream. In lakes contaminated by organic pollution, decomposition of the organic matter and elevation of the nutrient levels also may cause low oxygen conditions in the benthic regions so that the normally diverse benthic invertebrate fauna is replaced by species similar, if not identical, to those found in the anoxic sections of organically polluted rivers.

Nutrients such as phosphorus and nitrogen enter fresh waters either from sewage or from diffuse sources, such as agricultural fertilizers. Nutrients may encourage growth of undesirable algae and may cause algal “blooms” and periods of low oxygen availability, especially in the deeper benthic regions. This process is called cultural eutrophication and is a major problem for lakes in both urbanized and rural areas. In such lakes, the invertebrate fauna, both planktonic and ben-
In the past century with the growth in human population and the rapidly rising demands for water for human consumption, industry, hydropower, and irrigation, there has been an enormous expansion in the number and scale of dams, barrages, and diversion works. Dams and their large upstream impoundments break the connectivity of rivers. This disruption has been recognized in situations where valuable fish stocks have been threatened. Fishways, portage schemes, and stocking have been used to reduce threats to fish, but nothing has been done to alleviate the effects on invertebrates. Thus dams may disrupt essential migrations, such as those of shrimps and freshwater crabs in the tropics, and prevent the normal movements of invertebrates down the river.

The waters impounded behind dams may flood valuable habitats—floodplain wetlands, river canyons, even lakes. In Tasmania, Australia, a large dam-created impoundment flooded an entire, isolated lake with a remarkably large beach. This lake, Lake Pedder, contained five endemic invertebrate species, four of which appear to have become extinct. Impoundments behind dams may have a greatly depleted littoral (shoreline) fauna due to rapid changes in water level incurred as a result of dam operations. Drawdown of lake levels to meet hydroelectricity or irrigation can leave many invertebrates stranded. In unimpounded rivers, organic matter and nutrients are transported downstream to be used by downstream invertebrates. This longitudinal transport ceases with impoundment. Water released from dams may come from the deep hypolimnion and consequently may be cold, deoxygenated, and low in organic material—conditions quite unfavorable for riverine invertebrates.

The purposes for which a dam is operated can have dire consequences for downstream invertebrates. Most drastic of all is the simple loss of water. Many dams are water diversion structures that leave the river channel below the dam waterless or with a small minimum flow. In many places, large rivers are left with so little water that they no longer reach the sea. Needless to say the loss of invertebrates is considerable. Dams may be built and operated to generate hydroelectricity, in which case water levels in the channel below the dam fluctuate greatly and quickly as water is released or suddenly cut off to meet electricity generation demands. Alternation between flood and drought creates disturbance levels that permit only the hardiest of invertebrates to exist in the channel. Dams operated to store water for irrigation, hold water in the wet season (winter), and release it for irrigation in the dry season. Consequently, the seasonal patterns of flow are reversed, and wet season floods are greatly diminished. The lack of water in the wet season and high water levels in the dry season disrupt the life cycles of many invertebrates. For many lowland rivers, the effects of greatly diminished floods have been devastating because inundation of the floodplain has been...
greatly reduced, resulting in a loss of invertebrates that would normally thrive on the floodplain.

D. Invasion of Exotic Species

Exotic species of plants and animals may reach new localities through deliberate introduction by humans or through accidental introduction during transport. The rate of introductions both between and within continents is rising rapidly. The most obvious introductions into freshwater systems have been plants, both aquatic (e.g., water hyacinth) and riparian (e.g., salt cedar), and fish (e.g., carp). Few invertebrates have been introduced deliberately (e.g., opossum shrimp, crayfish), but many, especially crustaceans and mollusks, have been introduced accidentally.

Introduced invertebrates may exert very strong effects on native species. Exotic invertebrates may be vectors for diseases lethal to native species. In Europe, native crayfish stocks have been decimated by the crayfish plague fungus, *Aphanomyces astaci*, which came with the deliberate introduction of North American crayfish. Exotic species may take up habitat space and outcompete local species. The zebra mussel from the Ponto-Caspian region invaded the Great Lakes and Hudson River and now is common throughout eastern North America from Oklahoma and Minnesota east. It monopolizes habitat space and has greatly reduced native mussels. Zebra mussel populations increase rapidly once introduced and through their immense filtration capacity may greatly reduce planktonic organisms increasing the clarity of the water but reducing invertebrate diversity.

Introduced aquatic plants, such as water hyacinth and alligator weed, may blanket the surfaces of tropical lakes and rivers. Such blanketing may reduce water quality by causing deoxygenation or by reducing photosynthesis of native plankton and aquatic plants. This, in turn, diminishes food resources for native invertebrates. Introduced riparian plants, such as tamarisk or salt cedar, may cause the loss of habitat for native invertebrates by dominating the stream banks and often by reducing water levels in the channel. Other introduced plants, such as eucalyptus in Spain and Portugal, may produce dead plant material that is not readily consumed by stream invertebrates, thereby causing them to starve. Many fish are predators of invertebrates. The introduction of exotic fish may increase predation pressure and deplete native invertebrate populations. Introduced carp have reduced invertebrate populations in floodplain wetlands. In New Zealand, introduced trout in invading streams have depleted or eliminated populations of native fish and native invertebrates. Released from the intense grazing pressure of native invertebrates, algae in such invaded streams proliferate and build up to high levels.

See Also the Following Articles

ENDANGERED FRESHWATER INVERTEBRATES • HOTSPOTS INVERTEBRATES, MARINE, OVERVIEW • INVERTEBRATES, TERRESTRIAL, OVERVIEW • LAKE AND POND ECOSYSTEMS • RIVER ECOSYSTEMS • WETLANDS ECOSYSTEMS

Bibliography


INVERTEBRATES,
MARINE, OVERVIEW

P. J. D. Lambshead* and P. H. Schalk†
*The Natural History Museum, London, and †ETI, Universiteit van Amsterdam

I. Physical Characteristics of the Oceans
II. The Fundamental Processes Controlling Marine Biodiversity
III. Sampling and Assessing Marine Invertebrate Biodiversity
IV. Taxonomy
V. Global Diversity
VI. Pelagic Diversity Patterns
VII. Benthic Large-Scale Diversity Patterns
VIII. Benthic Small-Scale Diversity Patterns
IX. Diversity over Evolutionary Time
X. Hydrothermal Vents and Cold Seeps

GLOSSARY

benthic That connected with the bed of the sea or a freshwater lake, river, or stream.
deep sea The seabed and immediately overlying water covered by seas at least 200 m deep.
disturbance A biological or physical factor that impacts a population or community by causing death, reduced reproduction, or increased emigration.
infauna Animals living within sediments.
invertebrate All animals that lack a backbone (vertebrae)—that is, most of the animal kingdom.
pelagic That connected with the water column.
planktonic Organisms of many different phyla that float in and are carried by water masses in the pelagic of the sea or freshwater.
production The increase in biomass of an individual, population or community as it grows by converting energy-food into biomass.

AN OVERVIEW OF MARINE invertebrate biodiversity is complicated because there are two distinct marine domains, the open ocean and the seafloor, which are affected by different ecological processes. The physical characteristics of the world ocean are described here both from the viewpoint of a benthic scientist who studies the animals that live on the seabed and a pelagic scientist who studies the organisms that float or swim freely in the water. The pelagic sections include some consideration of single-celled organisms, including examples capable of photosynthesis, because excluding them would be artificial. The processes that produce modern marine diversity patterns are described and the unique features that make marine diversity different from terrestrial diversity are listed. By convention, both pelagic and benthic fauna are not categorized by a functional ecological approach but by the type of equipment used to sample the organisms, which largely depends on their size. The different size classes are listed and examples given of the organisms included in each group. A major limitation in understanding marine invertebrate biodiversity is the inadequate state of taxonomic knowledge of marine organisms, particularly the small and deep-sea taxa. The implications of our igno-
rface for estimating regional or global diversity are discussed. Open ocean pelagic biogeography patterns are described and examples given of species with commonly occurring patterns. Benthic diversity patterns are divided into large and small scale. The large-scale patterns include latitudinal and bathymetric gradients. The change in both diversity and the dominant faunal groups over evolutionary time from the Cambrian to the present is followed. Finally, the unusual hydrothermal and cold seep fauna that have been recently discovered are described.

I. PHYSICAL CHARACTERISTICS OF THE OCEANS

The world ocean covers more than two-thirds of the surface of the earth and as much as 90% of the habitable volume. It is the largest habitat on earth and about one billion people depend on the oceans for their primary protein needs. Yearly, well over 90 million tons of fish and shellfish are harvested.

Only 10% of the oceans' area could be classed as shallow water and this is mostly located around the continental margins. Seas deeper than 2000 m cover half the earth's surface. The seashore zone, the area covered and exposed by tides, corresponds to the ecological littoral region. Shallow seas make up the sublittoral zone, also called the continental shelf, which ends at the shelf break, usually at around 200 m. The continental slope marks the boundary between the continental and ocean floor crusts. It extends from the shelf break down to about 1300 m, so it roughly corresponds to the ecological bathyal region, 200 m to 2000 m. The continental rise and the abyssal plain roughly correspond to the ecological abyssal region, 2000 to 6000 m. Other features of the seafloor are rocky seamounts and ridges, which may be of a considerable size, and deep trenches down to 11,000 m or more. Trenches are usually a feature of subduction zones where the seafloor buckles and deepens beneath a continental crust. Deep trenches correspond to the ecological hadal region. The continental rise and abyssal plain tend to be relatively flat. Geological features such as terraces and submarine canyons, however, sculpture the continental slope. Pressure increases down the water column at the rate of one atmosphere per 10 m.

Much of the seafloor is covered in sediments; bare rock is rather rare. Near the coast these sediments are of terrigenous origin and may be coarse (e.g. clay, sands, or pebbles). Terrigenous material may be found down to the continental rise and vast areas of the seafloor consist of clay sediments or, in productive areas, biogenic oozes mainly from diatoms, radiolarians, and Foraminifera.

The temperature of the water declines rapidly with depth until it stabilizes at about 4°C, which takes place at approximately 1000 m, depending on latitude. This phenomenon is known as the permanent thermocline. At greater depths, the temperature declines slowly. Water currents are too complex to be discussed here in detail but, broadly speaking, deep water is formed by the sinking of dense, cold saline water in the Antarctic and Arctic. This dense water spreads out across the world ocean, eventually returning to the surface.

In general, marine water is well oxygenated, often at values near saturation. There are areas where rapid seasonal blooms in the upper water column deoxygenate the water leading to oxygen minimum zones. Oxygen concentration falls rapidly with sediment depth, especially in litoral muds and sediments under shallow enclosed seas.

The open ocean or pelagic realm has no obvious boundaries such as the mountain ranges or large water masses (rivers, seas, or oceans) that are so important in driving speciation on land, although to a certain extent the three oceanic basins—Atlantic, Indian, and Pacific—are partially separated by landmasses. The world ocean is essentially a continuum and the most important environmental factors are temperature and salinity, which translate primarily into latitude and depth. However, despite the lack of obvious barriers, speciation has and does occur in the pelagic environment. The geographic species distribution of modern plankton is the product of the geological history of the oceans and continental barriers, limitations of species in adapting to biotic and abiotic environmental factors, and the degree of organism mobility.

The pelagic realm is divided by depth into zones. Light, temperature, and salinity are responsible for the major vertical biodiversity gradients. In the upper 200 m, the epipelagic, sufficient light is available to sustain photosynthetic processes. This is the primary production zone where light and carbon dioxide is converted into carbohydrates, the principle energy (food) source for all open ocean inhabitants. The mesopelagic zone is the realm of the diurnally (daily) migrating organisms and extends from 200 to 2000 m depth. This is the deepest into the water column that light penetrates. Diurnal migration is triggered by light and typically involves moving into shallower waters at night and deeper water layers during the day. Possible reasons for such
vertical migration may include hiding in the dark depths from predators during daylight and energy conservation achieved by spending part of the time in deeper, colder water layers, thus allowing organisms to reduce their metabolic rate. The bathypelagic (2000 to 6000 m) and abyssopelagic are the aphotic zones, where there is perpetual darkness (Table I).

II. THE FUNDAMENTAL PROCESSES CONTROLLING MARINE BIODIVERSITY

There are three major differences between marine and terrestrial systems that have an enormous impact on biodiversity. The first is that the oceans are three-dimensional; organisms live, feed, and reproduce at all levels in the water column. The second is that light has low penetration through sea water, so photosynthesis only occurs in the upper water column, the photic zone, so most of the oceanic system has no primary production. The third is that most photosynthesis is carried out by tiny single-celled organisms. Macrophotosynthetic organisms (macroalgae or seaweeds) are rare and tend to be concentrated in a thin zone around the edge of the continents in the littoral and sublittoral regions.

Pelagic organisms in the three-dimensional water column have no structure to "hide" in, either for ambush attacks or defense. Essentially, there are only three possible life strategies. An organism can evolve for speed and agility to capture prey and escape predators, or an organism can evolve as a jelly, offering a poor food return for potential predators. A third strategy is rapid growth and reproduction resulting in huge numbers: schools and swarms offer protection to those within. Many pelagic organisms show a patchy distribution for this reason. The organisms within the sediment, the infauna, exist in a two-dimensional world that has a physical structure within which organisms may shelter. The marine system, with regard to biodiversity, may therefore be divided into two separate domains, the pelagic and the benthic. Some infaunal taxa have a pelagic phase and so have a presence in both domains. We may predict that marine pelagic diversity should be lower than terrestrial diversity given the limited choice of life strategy of pelagic organisms coupled with the active or passive mobility of organisms through the water column. The benthic domain is structurally similar to the terrestrial environment and may be predicted to have a similar diversity with the caveat that the absence of large photosynthetic organisms removes an entire biotope.

The lack of primary production in most of the oceanic realm results in a dependence on an organic flux that originates on land or in the photic zone. The organic flux becomes weaker further away from the continents or deeper into the water column. The weakening flux gradient governs benthic biomass, which declines with depth and distance from the continental shorelines. The Russian grab-sampling programs of Belyaev and colleagues from the 1950s onward established the global distribution of benthic biomass (Fig. 1). Productivity has a major influence on biodiversity so the organic flux gradient also will have an impact on benthic biodiversity patterns.

Physical disturbance is also considered to be a key process controlling biodiversity. Physical disturbance was considered to be highest in the littoral region and to decline consistently with water depth down to the abyssal plain. Recent research has caused science to modify this view, as there is now good evidence for hydrodynamic disturbance and seasonal perturbations right down onto the abyssal plain. But in general, the trend is for physical disturbance to decrease with depth. The main physical disturbance processes that show a depth trend are mechanical effects (current energy and wave action), temperature variation (including exposure in the littoral region), and salinity change. Deep trenches are possibly highly disturbed due to slumping of the trench sides.

III. SAMPLING AND ASSESSING MARINE INVERTEBRATE BIODIVERSITY

A. Benthos

Marine invertebrates are divided into three size classes on the pragmatic basis of the equipment used to collect them (Table II): Megafauna are large, visible animals...

<table>
<thead>
<tr>
<th>Depth Zone</th>
<th>Measured in Meters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epipelagic</td>
<td>0 - 200</td>
</tr>
<tr>
<td>Mesopelagic</td>
<td>200 - 2000</td>
</tr>
<tr>
<td>Bathypelagic</td>
<td>2000 - 6000</td>
</tr>
<tr>
<td>Abyssopelagic</td>
<td>&gt; 6000</td>
</tr>
</tbody>
</table>
that may be seen by eye (or on photographs). Macrofauna are infaunal (sediment dwelling) organisms that are not normally visible but are retained on a 1 mm sieve (0.5 mm or less in the case of deep-sea samples). Microfauna are the infauna too small to be retained on such a sieve. It is not clear whether these arbitrary classifications also have biological significance. Some taxa are almost entirely in one size class for their entire life history; for example, nematodes are always considered to be meiofauna. Others—for example, polychaetes, although predominantly in one size class, in this case macrofauna—may also have meiofaunal and megalafaunal representatives for at least part of their life history. Warwick and colleagues have some evidence, most convincingly from coastal sandy sediments, that the division of organisms into macro and meiofauna is ecologically meaningful. It appears that organisms must be either macrofaunal-sized, to actively burrow through the sediments, or meiofaunal-sized, to slip between the sediment particles, but mechanical limitations prevent them being of intermediate size. 

Megalafauna are numerically dominated by the phylum Echinodermata, although the Crustacea are also important. The most abundant and diverse macrofauna group are the polychaete worms (Annelida), while meiofaunal nematodes (Nematoda) dominate the benthos, at least in terms of abundance and diversity of meiofauna. Arthropods are also important in both the macro and meiofaunal size classes. The average marine benthic infaunal invertebrate is a soft-bodied worm (Fig. 2), but vermiform animals are less important components of the megalafauna. Littoral benthos from sediments is sampled by inserting a corer into the deposits and washing the removed sediment over an appropriately sized sieve. Sampling offshore is more problematical. Megafauna are collected by equipment such as Agassiz trawls, anchor dredges, or epibenthic sledges towed along the bottom. Such equipment has problems collecting consistent samples and is suitable only for producing qualitative or semiquantitative data. Initially, the smaller infaunal animals were collected with grabs, and these are still used where circumstances dictate but they are inefficient. The development of the box corer by Hessler and Jumars was a critical improvement in the accurate quantitative sampling of macrofauna for biodiversity studies. Development of quantitative meiofauna samplers, notably the Scottish Marine Biological Association’s multiple corer, similarly revolutionized investigation into offshore meiofauna biodiversity. Marine samples taken with modern corers provide data that are more quantitatively accurate than most terrestrial studies, and this has influenced the type of biodiversity questions asked by marine scientists. Ma-
TABLE II
Examples of Marine Benthic Invertebrate Taxa Classified into the Three Size Groups

<table>
<thead>
<tr>
<th>Size class</th>
<th>Megafauna</th>
<th>Macrodrama</th>
<th>Meiofauna</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collecting equipment</td>
<td>Agassiz trawl, epibenthic sled, anchor dredge</td>
<td>Grab, core (notably the USNEL box core)</td>
<td>Core (notably the SMBA multiple core)</td>
</tr>
<tr>
<td>Examples of taxa</td>
<td>Ascidia (sea squirts)</td>
<td>Anthropoda (amphipods, crustaceans, cumaceans, isopods, mites, tanaids)</td>
<td>Nematoda (thread worms)</td>
</tr>
<tr>
<td></td>
<td>Asteroidea (sea stars)</td>
<td>Bivalvia (clam shells)</td>
<td>Cephalopoda (mainly harpacticoids)</td>
</tr>
<tr>
<td></td>
<td>Brachiopoda (sea pens, sea lilies, sea anemones, conchs)</td>
<td>Brachiopoda (lamp shells)</td>
<td>Gastrotricha</td>
</tr>
<tr>
<td></td>
<td>Bryozoa (moss animals)</td>
<td>Bryozoa (sea lilies, sea anemones, corals)</td>
<td>Kinorhyncha</td>
</tr>
<tr>
<td></td>
<td>Cerebrata (sea irises, jellyfish, sea pens, sea lilies, sea anemones, conchs)</td>
<td>Cerebrata (sea irises, jellyfish, sea pens, sea lilies, sea anemones, conchs)</td>
<td>Loricifera</td>
</tr>
<tr>
<td></td>
<td>Ctenophora (sea pens, sea fans)</td>
<td>Ctenophora (sea fans)</td>
<td>Ostracoda</td>
</tr>
<tr>
<td></td>
<td>Echinodermata (sea urchins)</td>
<td>Echinodermata (sea urchins)</td>
<td>Phorinda</td>
</tr>
<tr>
<td></td>
<td>Ectoprocta (sea anemones)</td>
<td>Ectoprocta (sea anemones)</td>
<td>Phorinda (tube worms)</td>
</tr>
<tr>
<td></td>
<td>Hemicentrotidae (sea urchins)</td>
<td>Hemicentrotidae (sea urchins)</td>
<td>Polychaeta (segmented sea worms)</td>
</tr>
<tr>
<td></td>
<td>Hypothoeida (sea cucumbers)</td>
<td>Hypothoeida (sea cucumbers)</td>
<td>Prorupella (worms)</td>
</tr>
<tr>
<td></td>
<td>Ophiuroidea (brittle stars and basket stars)</td>
<td>Ophiuroidea (brittle stars and basket stars)</td>
<td>Sipuncula (peanut worms)</td>
</tr>
<tr>
<td></td>
<td>Pycnogonida (sea spiders)</td>
<td>Pycnogonida (sea spiders)</td>
<td>Turbellaria (flat worms)</td>
</tr>
</tbody>
</table>

Terrestrial biodiversity analysis tends to focus on larger scale measures of diversity based on species richness per area so it can be difficult to compare published marine diversity patterns with those obtained for terrestrial environments.

B. Pelagos

Functional classification of pelagic organisms is based on locomotion, size, and trophic level or ecosystem function. Plankton are generally passively drifting organisms carried by water movements; examples include bacteria, algae, and small animals. Nekton are actively swimming organisms and so can to a certain extent migrate in the horizontal plane; examples include squid, fish, and some crustaceans. Phytoplankton, marine plants, are primarily unicellular and responsible for primary production. They are the main food source for all life in the open ocean. Zooplankton are animals with a planktonic life style.

Division of pelagic organisms into size classes is related to retention by the different mesh sizes of nets and filters used for sampling (Table III). Pico and nanoplankton are usually collected with remotely operated opening and closing bottle devices, allowing control of the depth zone sampled. To collect organisms sized from microplankton upward, various types of net sampling gear are used ranging from simple ring nets, which
FIGURE 2 The invertebrate animals found in a 0.25 m$^2$ layer of sea sediment from the central North Pacific by Hessler and Jumars. A worm shape is the most common body form, nematodes being particularly abundant. Note that the size of the animals is exaggerated so that they may be seen. (From Gage and Tyler, 1991).

are hauled vertically, to complicated remotely operated opening and closing net systems that can sample horizontal depth layers (e.g., RMT net, Omori net). Nekton are sampled with large, pelagic fish trawls. Often pelagic samplers are used in combination with other equipment that measures physical factors such as depth, temperature, salinity, and light. Due to the vastness of the pelagic realm relatively little has been studied, although there has been reasonable geographic coverage. It is estimated that less than 5% of its huge volume has been sampled once! Deep-sea and very small planktonic species are particularly undersampled; undoubtedly many new species wait to be discovered.

One of the main problems with sampling the benthic and pelagic realm is that scientists are literally grabbing in the dark. In contrary to terrestrial research, where the scientist has a visual concept of the environment and can generally “see” what he or she is doing, marine biologists are completely dependent on remotely operated equipment. This has hampered progress in forming concepts of the marine environment as an ecosystem and estimating its biological diversity. Remotely operated vehicles (ROV’s) with cameras and sampling devices have improved our understanding of the marine environment significantly in the past decade. The choice of sampling gear has a large impact on the representativeness of the samples for the community as a whole, hence the frequent discoveries of new species or new information about known species. Giant squids present a nice example of a species that have eluded biologists.

IV. TAXONOMY

Estimating, let alone listing, to the number of described species on the earth is a difficult task. For some groups of organisms, usually terrestrial, the actual number of

<table>
<thead>
<tr>
<th>Planktonic</th>
<th>Micronektonic</th>
<th>Nekton</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planktonic</td>
<td>Micronektonic</td>
<td>Nekton</td>
</tr>
<tr>
<td>Nanoplankton</td>
<td>0.2–20 mm</td>
<td>2 cm</td>
</tr>
<tr>
<td>Microplankton</td>
<td>2–20 mm</td>
<td>2 cm</td>
</tr>
<tr>
<td>Mesoplankton</td>
<td>&gt; 2 mm</td>
<td>&gt; 2 cm</td>
</tr>
<tr>
<td>Macropelagic</td>
<td>&gt; 200 mm</td>
<td>&gt; 9 cm</td>
</tr>
</tbody>
</table>

TABLE III

The Size Classes of Planktonic Organisms with Typical Examples

- Bacteria, algae, protozoa
- Algae, heterotrophs
- Small fish, squid, crustaceans
- Shrimp, fish larvae, jellyfish
INVERTEBRATES, MARINE, OVERVIEW

species is reasonably well known with most species named and described; examples include birds, mammals, and higher plants. Other groups have not been comprehensively researched and accurate estimates for species numbers are not available. This category includes beetles, mollusks, and nematode worms. The current best estimate given by the Global Biodiversity Assessment is that there are 1.75 million species known to science. The estimate for the total number of existing species on earth is approximately 12 to 13 million although some scientists think that it may be as high as 20 million.

Classifying and naming the 1.75 million known organisms is an enormous task, but it is vital for comparing and exchanging biodiversity information between different organizations and countries. A profession of approximately 5000 experienced taxonomists worldwide is performing this documentation task. The Kingdom of Plants is the best known and that of Protists least. For the Kingdom Animalia only 20,000 new names are described yearly, including 17,000 species descriptions. The natural history collections stored and maintained in various countries are an important archiving tool for research, and they provide the community at large with a systematic insight into species diversity.

Inventorying marine invertebrates poses exceptional problems. In an area of survey twice as large as the terrestrial environment, marine biology has to deal with the vagueness of the three-dimensional biogeographical borders of the pelagic realm. This makes stock assessments and the definition of distribution ranges extremely problematical.

Species richness is unevenly distributed over the various kingdoms and phyla. The Kingdom Animalia represents the greatest numerical diversity, with well over 1.3 million known species. The terrestrial order of Insecta, in the phylum Arthropoda, takes up about 70% of the known animal diversity.

In both terrestrial and marine domains, the highest diversity is found in the equatorial region. The productivity (in grams carbon per m²) is more or less equal: 2.2 g C/m²/yr in tropical rain forests and 2.5 g C/m²/yr in coral reefs. In the aquatic environment, current data suggest that coral reefs are the most species-diverse regions and that they are comparable with tropical rain forests. Although reefs only form 0.2% of the seafloor by area, they harbor over 25% of all known marine species. The deep-sea is also believed to be highly diverse. Due to the difficulties in accessing this environment relatively little is known and the discovery of many new species is expected in the coming years (Table IV).

It is often surprising for nontaxonomists to discover that most species in the world have never been described and most probably will never be described. This is especially true for the smaller size classes and marine infauna in general. Marine species have received less attention from systematists than terrestrial organisms, but the vagueness of the three-dimensional biogeographical borders of the pelagic realm makes stock assessments and the definition of distribution ranges extremely problematical.

### Table IV

Known and Estimated Species Diversity of Various Taxa, including Global Biodiversity Assessment (GBA)

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Described species (×1000)</th>
<th>Estimates (highest) (×1000)</th>
<th>Estimates (lowest) (×1000)</th>
<th>GBA estimate (×1000)</th>
<th>Accuracy of estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viruses</td>
<td>4</td>
<td>1.000</td>
<td>50</td>
<td>400</td>
<td>Bad</td>
</tr>
<tr>
<td>Eubacteria &amp; Archae</td>
<td>4</td>
<td>3.000</td>
<td>50</td>
<td>1.000</td>
<td>Bad</td>
</tr>
<tr>
<td>Fungi</td>
<td>72</td>
<td>2.700</td>
<td>200</td>
<td>1.300</td>
<td>Reasonable</td>
</tr>
<tr>
<td>Protista</td>
<td>40</td>
<td>200</td>
<td>60</td>
<td>200</td>
<td>Bad</td>
</tr>
<tr>
<td>Algae</td>
<td>40</td>
<td>1.000</td>
<td>150</td>
<td>400</td>
<td>Bad</td>
</tr>
<tr>
<td>Plants</td>
<td>210</td>
<td>500</td>
<td>300</td>
<td>320</td>
<td>Good</td>
</tr>
<tr>
<td>Nematodes</td>
<td>25</td>
<td>100/000</td>
<td>100</td>
<td>400</td>
<td>Bad</td>
</tr>
<tr>
<td>Arthropods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crustaceans</td>
<td>40</td>
<td>200</td>
<td>75</td>
<td>150</td>
<td>Average</td>
</tr>
<tr>
<td>Arachnids</td>
<td>73</td>
<td>1.000</td>
<td>50</td>
<td>750</td>
<td>Average</td>
</tr>
<tr>
<td>Insects</td>
<td>950</td>
<td>100/000</td>
<td>2.000</td>
<td>8.000</td>
<td>Average</td>
</tr>
<tr>
<td>Mollusks</td>
<td>70</td>
<td>200</td>
<td>100</td>
<td>200</td>
<td>Average</td>
</tr>
<tr>
<td>Chordata</td>
<td>45</td>
<td>55</td>
<td>50</td>
<td>50</td>
<td>Good</td>
</tr>
<tr>
<td>Others/diverse</td>
<td>113</td>
<td>800</td>
<td>200</td>
<td>250</td>
<td>Average</td>
</tr>
<tr>
<td>Total</td>
<td>1,750</td>
<td>111,659</td>
<td>3,635</td>
<td>13,620</td>
<td>Bad</td>
</tr>
</tbody>
</table>
less than 1% of described species are marine. It is estimated that there are at least 200,000 known animal species and about 50,000 known plant (algae) species in the oceans. Fish (26,000 species) and marine mammals (250 species) belong to the best known taxa.

Most marine animals live in the deep sea or are tiny and were out of sight and out of mind. Only in the second half of the 20th century has the equipment become available to comprehensively collect deep-sea and small planktonic organisms, and it is still an expensive business limited to a handful of professionals. When this is compared with the long and vigorous tradition of both professional and amateur taxonomic study in the terrestrial domain, the backwardness of marine invertebrate systematics is understandable.

The taxonomically best known marine area in the world is the coastal strip along northwest Europe and to a lesser extent the coasts of the rest of Europe and North America. Outside these zones, less than half the marine invertebrate species are known to science. For example, a recent study of 16 box cores taken from bathyal depths off the British coastline revealed 304 polychaete species, of which only 17% could be associated with a known species. A study of 6 liters of coral reef sediment from Hawaii found 158 polychaete species, of which only 30% had been previously described. The largest study of macrofauna ever undertaken in the deep sea in the bathyal, northwest Atlantic counted 798 infaunal species from 233 box cores, but less than half could be named (Table V). This study illustrates the importance of polychaetes. In general, at least half the macrofauna species of any deep-sea sample will be new to science.

The systematics of meiofauna is even less well known. For example, a nematode study in the Arctic Ocean reported that only 4% of the 92 species found were known to science, while a study in the Venezuela Basin could only name 1.5% of the 136 species found. The taxonomic breakdown of different meiofauna groups in a single region has not been properly explored due to the difficulty of gathering the disparate taxonomic expertise. Some taxa, such as turbellarians, may be important but are commonly ignored in samples because they are so difficult to identify (Table VI).

The identification of marine benthic invertebrates is problematic not just because of lack of taxonomic coverage but because the taxonomic state of the art is inadequate. For many of the soft bodied groups, the species concept (i.e., the criteria used to define a species) is poorly defined. For example, the apparently well-known opportunistic polychaete species Capitella capitata and Streblospio benedicti have been found to be a complex of genetic sibling species (15 in the case of C. capitata).

The lack of species descriptions of small marine animals makes it difficult to assess regional diversity, and this is another factor that encourages marine biologists to consider diversity only over small scales. Only marine megafauna have an adequately censused taxonomy.

### V. GLOBAL DIVERSITY

It is impossible accurately to assess the total number of species in the oceans because of limited sampling, the nonrandomness of sampling locations, and inadequate taxonomy. The best estimate has come from Grassle...
and Maciolek, who sampled benthic macrofauna for 176 km along a bathymetric gradient 1.2 km to 2.0 km deep in the bathyal zone of the northeast Atlantic. They discovered a rate of species turnover of one new species detected every km. Assuming one new species per km², given that there is 10⁸ km² of deep sea, this implies that there may be up to 10⁸ (100,000,000) species in the seas. Such extrapolation techniques are notoriously inaccurate so this figure should be treated with caution but Grassle and Maciolek’s calculation is surprisingly similar to the estimate obtained for insect diversity of rain forests by Erwin who used substantially the same technique, suggesting that marine benthic and terrestrial diversity may be similar. A much better knowledge of the geographic ranges of the smaller benthic organisms is essential for an accurate estimation of global marine diversity. This will only be possible when some of the taxonomic problems surrounding these organisms are resolved.

The marine environment certainly demonstrates a greater diversity at higher taxonomic levels than the terrestrial domain. Life evolved in the oceans and many phyla never evolved to survive in the harsher terrestrial environment. It is highly likely that new metazoan phyla remain to be discovered, especially among the small organisms in marine sediments. The free-living phylum Loricifera was only described in 1983 and the commensal phylum, Cycliophora, was first found in 1995 (Table VII).

The distribution of benthic invertebrate species richness around the global ocean basins has not been comprehensively assessed. Schopf has shown that both ektroprocts and bivalve mollusks are twice as diverse in the Pacific than the Atlantic. This may reflect the relative sizes of the two oceans, or it may be the product of the relative age of the basins; the Atlantic is younger than the Pacific.

There is still some controversy over whether the highest marine benthic invertebrate diversity is found in coastal regions or the deep sea. Certainly the deep sea is a much larger environment than shallow water, so it might be expected to be home to more species; but it is unclear whether it is more diverse per area sampled. Originally, the low abundance of animals found in the deep sea caused scientists to assume that the highest diversities must occur on the coastal shelf. However, in the 1960s Saunders found the deep sea unexpectedly diverse, indeed with a higher species richness than shallow water. A series of studies appeared to confirm this view but recently an analysis has reopened the debate. Gray carried out a similar study to that of Grassle and Maciolek (noted earlier) but on the Norwegian coastal shelf. He found a similar, or even higher, rate of accumulation of species with sampling on the shelf compared to the deep sea.

VI. PELAGIC DIVERSITY PATTERNS

Van der Spoel and Heijman summarized the general distribution pattern types and biogeographic regions in the open ocean for phyto- and zooplankton species as Arctic, Subarctic, Cool-Temperate, Warm-Temperate, Tropical, Temperate and Subtropical, Subantarctic, and Antarctic. These regions principally reflect seawater temperature and latitude. Contrary to earlier beliefs, relatively few plankton and nekton species have a true cosmopolitan circumglobal distribution. Despite the lack of obvious barriers in the oceans, there is a surprising variety of the types of species distribution. Species specific characteristics and ecosystem relations play an important role in combination with abiotic factors in determining distribution patterns, but many are further modified by regional influences such as ocean basins, currents, and divergence and convergence zones. Typical and easily recognizable biogeographic patterns are belt-shaped patterns related to latitudes (temperature regimes) and nectonic distributions (basically around the continental coasts and shallow water areas).
bution around the Antarctic continent of Euphausia superba (the shrimp commonly known as krill) is an example of a neritic distribution (Table VIII).

Many widely distributed pelagic plankton species show North-South or East-West variation. Without active migration, genetic contacts between populations from one end of the range to the other are limited or even nonexistent, promoting genetic drift. In addition, the differing environmental circumstances (abiotic and biotic) at the ends of each range can give rise to phenotypic differences. This is reflected by both morphological and ecological variation within a species, a first step toward the development of new species. An example of North-South variation is found in the shell shape and size of many species of Pteropoda (pelagic Molusca). East-West variation can be found for example in Eucalanus subtenuis (copepod, Crustacea).

It is expected that over coming decade, knowledge of pelagic biogeography and open ocean distributions will develop exponentially (Table IX).

### VII. BENTHIC LARGE-SCALE DIVERSITY PATTERNS

#### A. Latitudinal Gradients

Lattitudinal gradients of terrestrial taxa commonly display a decline in species richness from the equator to

---

**Table VIII**

<table>
<thead>
<tr>
<th>Type</th>
<th>Region</th>
<th>Example species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circum-global</td>
<td>Cosmopolitan</td>
<td>Stylocheiron maximum</td>
</tr>
<tr>
<td>Belt-shaped patterns</td>
<td>Cosmopolitan</td>
<td>Rhizosolenia alata</td>
</tr>
<tr>
<td></td>
<td>Subtropical</td>
<td>Clausocalanus paupulus</td>
</tr>
<tr>
<td></td>
<td>Central Waters</td>
<td>Stylocheiron subtilis</td>
</tr>
<tr>
<td></td>
<td>Temperate N and S</td>
<td>Sagitta plantonis / zetesios</td>
</tr>
<tr>
<td></td>
<td>High latitude N</td>
<td>Calanus glacialis</td>
</tr>
<tr>
<td></td>
<td>High latitude S</td>
<td>Physodesmac marinae</td>
</tr>
<tr>
<td>Central water patterns</td>
<td>Atlantic, Indian, Pacific</td>
<td>Euphausia brevis</td>
</tr>
<tr>
<td>Endemic patterns</td>
<td>Atlantic</td>
<td>Euphausia americanum</td>
</tr>
<tr>
<td></td>
<td>Indian</td>
<td>Desmopteris gardineri</td>
</tr>
<tr>
<td></td>
<td>Pacific</td>
<td>Sagitta pseudovirens</td>
</tr>
<tr>
<td>Nerritic patterns</td>
<td>Cold water north</td>
<td>Gammarellis lonares</td>
</tr>
<tr>
<td></td>
<td>Cold water south</td>
<td>Euphausia crystallorophias</td>
</tr>
<tr>
<td></td>
<td>Warm water</td>
<td>Charybdis smithi</td>
</tr>
<tr>
<td>Bathypelagic patterns</td>
<td>Deep water</td>
<td>Cyclothone pseudopallida</td>
</tr>
</tbody>
</table>

**Table IX**

<table>
<thead>
<tr>
<th>Species Diversity in the Atlantic (A), Pacific (P), and Indian (I) Oceans for three example groups: Chaetognatha, Pteropoda, and Euphausiidae (largest diversity is found in the equatorial region)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Arctic</td>
</tr>
<tr>
<td>Subarctic</td>
</tr>
<tr>
<td>60°N–60°S</td>
</tr>
<tr>
<td>Subantarctic</td>
</tr>
<tr>
<td>Antarctic</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
the poles. Many different processes have been suggested to explain such gradients including competition, predation, mutualism, parasitism, and host diversity, but the most convincing explanation is that the diversity gradient follows the gradient of solar energy. There are two potential mechanisms that might cause this relationship. One suggestion is that latitudinal gradients in solar radiation cause concomitant gradients in productivity and hence diversity. The other suggestion is that the relationship between solar energy and diversity is the result of increased evolutionary speed in warmer conditions. The explanations are not mutually exclusive.

The relationship between latitude and diversity for marine benthic coastal fauna could best be described as confusing, with different studies and different taxa yielding conflicting results. For example, if the diversity of core samples is compared using diversity indices, then no significant difference is found between tropical and temperate samples. This experiment has been repeated for both macrofauna by Warwick and colleagues and meiofauna by Boucher and Lambshead.

However, Roy and colleagues were able to test for a latitudinal diversity gradient for prosobranch gastropod mollusks using data based on number of species per degree of latitude. Unusually for marine studies, this data is similar in scale and format to the terrestrial diversity data used for analyzing large-scale patterns and it is noteworthy that it produces convincing latitudinal diversity gradients (Fig. 3). Furthermore, the prosobranch diversity gradient follows the solar energy gradient in a similar way to terrestrial latitudinal gradients.

The deep sea is an intriguing location to search for latitudinal gradients because solar energy can have no direct impact on deep-sea diversity. The deep sea is uniformly cold and relatively stable so it can be argued that temperature cannot be a factor in controlling diversity though influencing evolutionary rates. Furthermore, productivity gradients in the deep sea are the result of gradients in the food flux to the seafloor, which vary according to a number of processes.

Rex and colleagues found that deep-sea gastropod mollusks, bivalve mollusks, and isopods showed a decline in diversity from the equator to the Norwegian Sea in the North Atlantic (Fig. 4). The explanation is not obvious, as productivity tends to increase northward.
Similar gradients have been searched for in the South Atlantic by Brey and colleagues but have not been found, so it is possible that the North Atlantic gradient is due to some individual feature of this ocean, possibly the Quaternary glaciation.

Analysis of the smaller invertebrates by Lambshead and colleagues has shown different patterns. Nematode diversity appears to be correlated with the food flux to the seafloor (i.e., a productivity gradient). In the North Atlantic, there is a weak increase in diversity northward. In the Pacific, nematode diversity declines northward away from the equator, as does the food flux to the seafloor (Fig. 5).

B. Bathymetric Gradients

Bathymetric gradients are probably better understood than any other large-scale diversity pattern for marine benthic invertebrates and for once a consistent pattern is found in all size classes. Rex showed that a number of macrofaunal taxa followed a parabolic diversity curve with depth, diversity peaking at about 2 to 3000 m (Fig. 6). A number of studies have since confirmed this finding. A similar pattern was found for megabenthos by Haedrich and colleagues and for nematodes by Bouchez and Lambshead (Fig. 7).

The exact peak in each bathymetric diversity gradient varies with different taxa and with different loca-
INVERTEBRATES, MARINE, OVERVIEW

FIGURE 7 The species richness of marine nematodes in sediments from six biotopes, temperate estuaries, tropical shallow water, temperate shallow water, bathyal depths, abyssal depths, and hadal depths. Species richness is measured using the Sanders rarefaction index. Diversity is highest in bathyal sediments.

TIONS AND, FOR ALL WE KNOW, THERE MAY BE TEMPORAL VARIATIONS. HOWEVER, THE BASIC PATTERN OF DIVERSITY PEAKING IN THE BATHYAL OR UPPER ABYSSAL REGION APPEARS CONSISTENT. THIS CONSISTENCY SUGGESTS THAT SIMILAR PROCESSES ARE RESPONSIBLE FOR ALL TAXA AND LOCATIONS. TWO OBVIOUS EXPLANATIONS ARE THAT BOTH PRODUCTIVITY AND DISTURBANCE DECLINE WITH DEPTH. COMPUTER MODELS (E.G. THOSE PRODUCED BY HUSTON) HAVE DEMONSTRATED HOW THE INTERACTION OF PRODUCTIVITY AND DISTURBANCE GRADIENTS CAN THEORETICALLY GIVE PARABOLIC DIVERSITY PATTERNS.

THE EXACT SPECIES COMPOSITION OF THE BENTHIC COMMUNITY CHANGES RAPIDLY WITH DEPTH FOR SOME TAXA WITH SPECIES HAVING A PREFERRED DEPTH RANGE. OTHER TAXA HAVE WIDE DEPTH RANGES. THIS HAS BEEN BETTER STUDIED FOR THE LARGER ORGANISMS THAN FOR SMALL ORGANISMS DUE TO TAXONOMIC REASONS. A WIDE VARIETY OF REASONS PROBABLY EXPLAIN THE DIFFERENT STRATEGIES DISPLAYED BY DIFFERENT TAXA BUT, IN GENERAL, THE SHARPEST DISCONTINUITY IS FOUND AT THE SHELF BREAK WITH A SECOND BOUNDARY AT ABOUT 1000 AND 2000 M. SO THE BASIC PATTERN IS OF ZONES OF SLOW CHANGE IN SPECIES COMPOSITION SEPARATED BY AREAS OF RAPID CHANGE.

VIII. BENTHIC SMALL-SCALE DIVERSITY PATTERNS

ONE OF THE MOST INTRIGUING OBSERVATIONS ABOUT MARINE BENTHIC INVERTEBRATES IS THE HIGH SPECIES RICHNESS FOUND IN INDIVIDUAL CORE SAMPLES (TABLE X).

THE OBSERVATION THAT A LARGE NUMBER OF SIMILAR SPECIES CAN BE FOUND TOGETHER IN SAMPLES OF APPARENTLY FEATURELESS SEDIMENT HAS ALWAYS REPRESENTED A CHALLENGE FOR “EQUILIBRIUM” EXPLANATIONS FOR THE ECOLOGICAL PROCESSES THAT CONTROL DIVERSITY. SUCH EXPLANATIONS DEPEND ON THE HYPOTHESIS THAT COEXISTING SPECIES MUST AVOID COMPETITION BY EACH POSSESSING SOME UNIQUE ADAPTATIONAL SPECIALIZATION. THIS HYPOTHESIS IS DIFFICULT TO ACCEPT FOR SUCH A SIMPLE ORGANISM AS A FREE-LIVING NEMATODE WORM, MANY OF WHICH SEEM TO HAVE SIMILAR FUNCTIONAL REQUIREMENTS.

AN ALTERNATIVE EXPLANATION IS THAT LARGE NUMBERS OF SPECIES CAN COEXIST BECAUSE THE SEDIMENT IS DIVIDED INTO TRANSIENT PATCHES BY BIOLOGICAL AND PHYSICAL SMALL-SCALE DISTURBANCE AND UNEVEN DISTRIBUTION OF THE VERTICAL FOOD FLUX. GRASSLE AND COLLEAGUES HAVE COINiked THE PHRASE SPATIAL-TEMPORAL MOSAIC TO DESCRIBE THIS PROCESS (FIG. 8). A NEW PATCH IS INITIATED BY A DISTURBANCE EVENT, OR A FOOD FALL, AND COLONIZATION TO EXPLOIT THE NEW SITUATION INCREASES DIVERSITY. AS THE PATCH AGES AND IS EXPAITED, COMPETITION INCREASES AND DIVERSITY DROPS. BUT AT SOME POINT, A NEW EVENT WILL RECREATE A NEW PATCH. IN THIS THEORY, THE SEDIMENT CONSISTS OF A HETEROGENEOUS MOSAIC OF UNIQUE PATCHES OF DIFFERENT AGES AND HISTORIES. FOR ANY GIVEN SPECIES, THERE IS A SUITABLE HABITAT WITHIN DISPERSION DISTANCE IN THE MOSAIC, AND WHEN THAT PATCH BECOMES UNSUITABLE, ANOTHER OPENS UP NEARBY. THE THEORY DEPENDS ON LIMITED LARGE-SCALE DISTURBANCE AND A LACK OF BARRIERS FOR DISPERSION SO IT IS CONSIDERED MOST APPLICABLE TO THE DEEP SEA (AND POSIBLY RAINFORESTS1).

This spatio-temporal mosaic theory predicts that areas of the deep-sea subject to a flux of phyto-detritus would have a more patchy distribution of animals and higher species diversity than other locations. This has
FIGURE 8 The spatio-temporal mosaic theory of Grassle and colleagues. The concept is that the sediments consist of a series of patches, each of which evolves through an unpredictable cycle. The example shows three patches, A, B, and C through times T1, T2, T3, and T4. Each vertical bar is a species and the height of the bar represents the abundance of that species. At T1, patch A is physically disturbed removing all the animals by T2. The patch is recolonized from the water column (dark arrow) and by migration from adjacent patches (light arrow). By T3 and T4, the patch is diverse with six species, five of which are equally abundant. Patch B is undisturbed through time. As the patch matures and is exhausted, the diversity drops to four species through competition. One of these species becomes completely dominant. Initially, patch C is similar to patch A, but it is biologically disturbed by the feeding behavior of a large organism at T2. This empties the patch allowing recolonization, forming a diverse community. This theory explains how apparently featureless marine sediments can support such a high diversity of life. For any species there is always a suitable new patch available nearby when it exhausts or is outcompeted in its current location. (From Paterson).

been tested for nematode worms by comparing the small-scale spatial distribution and species diversity of nematodes from a station in the phytodetritus-enriched Porcupine Abyssal Plain with a station from the more oligotrophic Madeira Abyssal Plain. Nematodes from the enriched site were more diverse and more species showed aggregation. Interestingly, the species were not aggregated in concordance together around the food or evenly randomly aggregated with respect to each other but showed discordance (i.e., species tended to aggregate in different places). This is consistent with Grassle’s theory.

IX. DIVERSITY OVER EVOLUTIONARY TIME

The marine fossil record is more complete than the terrestrial because the ocean floor tends to accumulate
sediment. Even so, only certain fauna, notably those with "hard parts," will form an adequate fossil record. The dominant macrofauna group, the polychaetes, presents a limited fossil record, and the dominant meiofauna group, the nematodes, has no record at all. The available fossil record suggests that marine species have an existence of about 4 million years. This indicates on average a 29% species turnover per million years. Assuming that the oceans contain $10^7$ (10,000,000) species, we might predict that 2.5 species would become extinct due to natural processes every year. In practice, the fossil record suggests that there are periods of relative quiescence followed by mass extinction and subsequent speciation. The fossil record also implies that 95% of all the marine species that have ever existed are extinct.

The history of marine biodiversity is best tracked at the family rather than the species level to remove some of the noise associated with an incomplete fossil record. This does tend to make the mass extinction events appear less dramatic. For example, only 54% of marine families were lost in the Permian extinction event, but it is estimated that 77 to 96% of all species became extinct. The early Cambrian shows a rapid increase in diversity that tends to flatten out in the middle and late Cambrian (Fig. 9). Most animal phyla appear in the record in this phase. The Cambrian diversity includes a number of "archaic" forms such as trilobites, hyoliths, and inarticulate brachiopods that decline after the Cambrian period. Diversity is not high in the Cambrian and rather unspecialized detritus and low suspension feeding organisms functionally dominate communities, suggesting a simple ecology.

**FIGURE 9** The family diversity of marine animals as measured by their fossil record from 600 M years ago to the present day. Graph C shows the rise and fall of the diversity of the Cambrian fauna of unspecialized detritus and low suspension feeding organisms. P shows the rise and fall of the diversity of the Paleozoic fauna of sessile benthic organisms, and M shows the rise of the diversity of the modern fauna of sediment movers and shell breaking predators. Note that sampling error may have exaggerated modern diversity. The arrows indicate periods of mass extinction. 1 is at the end of the Ordovician, 2 in the late Devonian, 3 at the end of the Triassic, and 4 at the end of the Cretaceous (the dinosaur extinction). P indicates the massive Permian extinction. (Graph by Sepkoski, 1992, after Heywood, 1995).
The Ordovician sees a steep rise in diversity but the curve then flattens for some 200 million years. Periods of mass extinction are detectable, notably at the end of the Ordovician and in the late Devonian, but in general diversity stabilized. This new Palaeozoic diversity was associated with an evolutionary radiation of sessile benthic organisms such as crinoids, articulate brachiopods, stenolomate bryozoans, and tabulate and rugose corals.

The Palaeozoic period ended with the catastrophic Permian-Triassic mass extinction. Diversity then apparently rose steadily, with a small extinction at the end of the Cretaceous, past the levels achieved in the Palaeozoic until the unique peak of the present day. Current family diversity is apparently twice as high as the Palaeozoic stable level. The Palaeozoic community never seems to have recovered and the modern high diversity is associated with a new fauna of shell-breaking predators and sediment movers. This fauna includes the familiar sea urchins, bivalve and gastropod mollusks, and crustaceans such as shrimps and crabs.

A variety of explanations have been given for this pattern. A number of authors have pointed to the association between the rise in taxonomic diversity and a rise in functional diversity. This almost certainly explains the Ordovician increase in diversity, which is essentially because of increased exploitation of marine sediments. Explanations for the Mesozoic-Cenozoic explosion in diversity are more controversial. One argument is that the breakup of the Pangaea supercontinent caused increased diversity through increased climatic variation leading to more endemicity. It is not clear how this argument would apply to the deep sea, where the bulk of marine diversity is to be found. Another suggestion is that evolution in the exploitation of the terrestrial domain has increased nutrient runoff into the oceans concomitantly increasing marine diversity. A similar argument has combined these two explanations suggesting that volcanism associated with continental breakup increased nutrient flux into the oceans.

Finally, Raup has suggested that the high modern marine diversity is an artifact of the fossil record. He observed that across the Mesozoic-Cenozoic time periods there are more younger rock available for study than older rock. If Fig. 9 is adjusted to allow for this, then there is a slower recovery from the Permian mass extinction and modern diversity peaks at much the same level as in the Palaeozoic. If true, it implies that there is a maximum global carrying capacity for marine biodiversity and that the Permian-Triassic extinction was even more devastating than Figure 9 suggests since the new graph would show that the Palaeozoic benthic fauna has declined consistently since the extinction event. This is possible as the new Mesozoic fauna included many organisms that would tend to disrupt the stable sediment environments on which the sessile Palaeozoic fauna depended.

X. HYDROTHERMAL VENTS AND COLD SEEPS

The special biological communities around hydrothermal vents were first discovered in 1977 on the Galapagos Rift at a depth of 2300 m. They are now known to be widely distributed in the oceans at tectonically active sites such as subduction and fracture zones, ocean-floor spreading centers, and back-arc basins. Vents are formed where sea water penetrates through fissures in the ocean floor deep into the earth's crust. The water becomes heated and escapes back to the surface through hydrothermal vents. The water temperature at a vent varies from mildly warm, 23°C, to superheated, 350°C and can be rich in sulphide and metallic ions. The hottest vents are called "smokers" from the precipitation of minerals in the water.

Cold seeps that release sulphide and methane-enriched water have been recorded from a number of areas in the Atlantic and Pacific Oceans. Similar chemical conditions have been found around whale carcasses that are oil rich. These sulphide-rich habitats are small and ephemeral. The evidence is that many have a life span that measures in decades, although cold seeps may last longer. However, they do appear to be commonplace and may be thought of as chains of islands across the seafloor, separated by distances of 1 to 100 km.

Vent communities have a noticeably different biodiversity from the surrounding ocean floor. The fascinating feature of vent communities is that they derive their energy from chemosynthetic primary production, by reducing compounds such as hydrogen sulphide, rather than photosynthetic sources. It is not clear whether these communities are entirely chemically independent of the photosynthetic world but certainly they are energetically independent. Biomass is high at hydrothermal vents and primary productivity may be double or triple that of the overlying water. Dense mats of bacteria are found, notably *Thiomicrospira*. A number of invertebrate species feed in dense colonies either directly on these bacteria or by means of symbiotic chemosynthetic bacteria. Evolutionary selection pressure has been for large, fast-growing species at these productive but ephemeral sites.
Conversely, species diversity is much lower at vents than at other benthic habitats, illustrating once again Rosenzweig’s “paradox of enrichment.” The dominant species vary from place to place and from depth to depth. For example, the shallow vents off the Palos Verdes Peninsula in California are dominated by the black abalone, *Haliockis cracherodii*; the Mid-Atlantic Ridge vents are characterized by two species of caridean shrimp; and the Eastern Pacific Rise and Galapagos Spreading Centre vents are dominated by the red-plumed, tube dwelling vestimentiferan worms (notably *Riftia pachyptila*) and the large bivalves *Calyptogena magnifica* and *Bathymodiolus thermophilus*. The fauna around the vents are often found to be new to science. Other fauna include a number of polychaetes, crustacea (notably including a number of decapods, and a new primitive genus of barnacle, which is a relic of the Mesozoic).

More than 150 new species, 50 new genera, and 20 new families or subfamilies have been identified with sulphide-rich vents and seeps. Endemism is considered to be high and it is noteworthy that many of the fauna do not appear to be particularly well adapted for high dispersal. Genetic studies on *Bathymodiolus* have shown strong genetic similarity between organisms at sites only 8 km apart but high dissimilarity at sites separated by 2200 km. Vents that are distinctly separated from the chain of sulphide-rich habitats such as the Mariana Trough or Hawaiian volcanic seamounts do show considerable faunistic differences from the norm and the Atlantic sites are less diverse than the Pacific.

See Also the Following Articles

ENDANGERED MARINE INVERTEBRATES • INVERTEBRATES, FRESHWATER, OVERVIEW • INVERTEBRATES, TERRESTRIAL, OVERVIEW • MARINE ECOSYSTEMS • PELAGIC ECOSYSTEMS • PLANKTON, STATUS AND ROLE OF • VENTS

Bibliography


INVERTEBRATES, TERRESTRIAL, OVERVIEW

Olof Andrén
Swedish University of Agricultural Sciences

GLOSSARY

Annelida Segmented worms, a relatively species-poor phylum.
Arthropoda Joint-legged animals, the most species-rich phylum existing today.
Enchytraeids Segmented worms that are related to but smaller than earthworms.
Invertebrates Animals without a backbone (vertebrae), ranging from unicellular Protozoa to multicellular, complex organisms such as insects.
Phylum Highest level of taxonomic division in the animal kingdom, followed in descending order by class, order, family, genus, and species.

THE TERRESTRIAL INVERTEBRATES CONTAIN THE MAJORITY OF THE WORLD’S ANIMAL SPECIES, although marine environments harbor more phyla representing more widely separate animal groups. Only comparatively few of the phyla have managed to adapt to terrestrial conditions, that is drought, ultraviolet radiation, no support of water against gravity, as well as rapid temperature fluctuations. However, some of the phyla that have adapted to the terrestrial environment have been very successful from a biodiversity standpoint, particularly the insects (phylum Arthropoda, class Insecta). When the basic demands for terrestrial life are fulfilled (water management, etc.), the terrestrial environment is well suited for the evolution of new species. It contains many areally defined habitats, for example, deserts, mountains, and forests, as well as vertical strata, for example, the air volume, treetops, hilltops, valleys, bushes, grasses, the litter layer, topsoil, and subsoil—each with a different evolutionary pressure.

1. The Terrestrial Invertebrates
2. The Terrestrial Environment

A. Taxonomic Groups, Fully Adapted to Terrestrial Conditions

The truly successful and well-adapted terrestrial invertebrates belong to only two phyla, Arthropoda (joint-legged animals) and Mollusca (snails, etc.). Other taxa, such as Protozoa, Aschelminthes (nematodes, etc.), and Annelida (earthworms, etc.), are limited by their need of an aquatic or at least water-saturated environment, such as in the litter layer and in soil, and a lack of protection against ultraviolet rays. These groups can only move in the true terrestrial environment during nights with favorable weather conditions.

The Insecta are the largest class of animals. About
one million species are described, but there are possibly as many as five million species present in the world today. The bulk of the insect species are in the orders Coleoptera (beetles), Lepidoptera (butterflies and moths), Hymenoptera (wasps, ants, bees), and Diptera (flies) (Table I and Figs. 1–3). Other Arthropoda that successfully cope with the terrestrial environment include the classes Diplopoda (millipedes) and Chilopoda (centipedes) and the similarly built but smaller and less known Symphyla and Pauropoda.

In the class Arachnoidea, scorpions (order Scorpi- onidea) and spiders (order Araneae) may be the most obvious. However, the smaller mites and ticks (order Acaeri) are a highly diverse group, containing parasites, predators, fungivores, and detritivores, and are common in almost all litter layers and soils (Fig. 4).

### TABLE I

<table>
<thead>
<tr>
<th>Order</th>
<th>Typical members</th>
<th>Number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protura</td>
<td>Telsentails</td>
<td>100</td>
</tr>
<tr>
<td>Thysanura</td>
<td>Bristletails, silverfish</td>
<td>700</td>
</tr>
<tr>
<td>Collembola</td>
<td>Springtails</td>
<td>2000</td>
</tr>
<tr>
<td>Ephemeroptera</td>
<td>Mayflies</td>
<td>1500</td>
</tr>
<tr>
<td>Odonata</td>
<td>Dragonflies, damselflies</td>
<td>3000</td>
</tr>
<tr>
<td>Orthoptera</td>
<td>Grasshoppers, crickets, cockroaches, mantids</td>
<td>23,000</td>
</tr>
<tr>
<td>Isopora</td>
<td>Termites</td>
<td>2000</td>
</tr>
<tr>
<td>Plecoptera</td>
<td>Stoneflies</td>
<td>1100</td>
</tr>
<tr>
<td>Dermaptera</td>
<td>Earwigs</td>
<td>1500</td>
</tr>
<tr>
<td>Embioptera</td>
<td>Webspinners</td>
<td>150</td>
</tr>
<tr>
<td>Pscoptera</td>
<td>Psoids</td>
<td>1100</td>
</tr>
<tr>
<td>Zoraptera</td>
<td>Zorapterans</td>
<td>20</td>
</tr>
<tr>
<td>Mallophaga</td>
<td>Chewing lice</td>
<td>3000</td>
</tr>
<tr>
<td>Anoplura</td>
<td>Sucking lice</td>
<td>250</td>
</tr>
<tr>
<td>Thysanoptera</td>
<td>Thrips</td>
<td>3200</td>
</tr>
<tr>
<td>Hemiptera</td>
<td>Bugs</td>
<td>23,000</td>
</tr>
<tr>
<td>Homoptera</td>
<td>Cicadas, aphids, hoppers</td>
<td>32,000</td>
</tr>
<tr>
<td>Neuroptera</td>
<td>Dohseflies, lollies, snowflakes, antlions</td>
<td>4700</td>
</tr>
<tr>
<td>Coleoptera</td>
<td>Beetles</td>
<td>277,000</td>
</tr>
<tr>
<td>Strepsiptera</td>
<td>Twisted-wing parasites</td>
<td>300</td>
</tr>
<tr>
<td>Mystoptera</td>
<td>Scorpionflies</td>
<td>350</td>
</tr>
<tr>
<td>Tsamoptera</td>
<td>Cardiophiles</td>
<td>4500</td>
</tr>
<tr>
<td>Lepidoptera</td>
<td>Butterflies and moths</td>
<td>112,000</td>
</tr>
<tr>
<td>Diptera</td>
<td>True flies</td>
<td>69,000</td>
</tr>
<tr>
<td>Stephoptera</td>
<td>Fleas</td>
<td>1100</td>
</tr>
<tr>
<td>Hymenoptera</td>
<td>Sawflies, wasps, ants, bees</td>
<td>103,000</td>
</tr>
</tbody>
</table>

The crustaceans (phylum Arthropoda, class Crusta- cea) are mainly aquatic, but the terrestrial isopods (or- der Isopoda, the "pill-bugs") are quite widespread, even in dry climate zones. However, they lack the water regulation of, for example, insects and spiders and are restricted to moist micro-environments and nocturnal habits.

Among the molluscs (phylum Mollusca) only a few classes have terrestrial capability, and the snails (class Gastropoda) have successfully coped with the dry atmosphere by developing a shell into which they can escape.
INVERTEBRATES, TERRESTRIAL, OVERVIEW

B. Taxonomic Groups with Limited Adaptations

The subkingdom Protozoa contains microscopic animals such as flagellates, ciliates, and amoebae. These are found almost everywhere but are restricted to water films or water bodies. However, several groups have resting stages that can withstand severe desiccation and have the capacity to rapidly reproduce when conditions improve. Thus Protozoa can regulate bacterial biomass and numbers in the soil.

The phylum Platyhelminthes contains primitive flatworms that lack an anus and sometimes even guts. They are mainly parasitic (e.g., tapeworms), but there are free-living species in the class Turbellaria.

Nematodes (phylum Aschelminthes, class Nematoda) are unsegmented, mainly microscopic worms that often have the capacity to form resting stages that can withstand drought and low temperatures. Thus the nematodes have successfully invaded most environments, including arable land (several are pests, but the majority are not) and the inside of other animals (several are internal parasites of mammals, insects, etc.) (Fig. 5).

Segmented worms are in the phylum Annelida, class Oligochaeta, and contain the well-known “ecological engineers”—the earthworms. Well over 1200 species of earthworms have been described. Lesser known but very abundant in boreal forests and wetlands are the smaller enchytraeids, which are also segmented (Fig. 6).

II. THE TERRESTRIAL ENVIRONMENT

The terrestrial environment is harsher than marine or freshwater environments: shortage of water, ultraviolet radiation, rapidly fluctuating temperatures, and a num-
ber of obstacles against movement and/or the dissemination of offspring create survival problems, but also opportunities for evolution and speciation. The fact that most of the known animal species are terrestrial is due to the diversity of habitats and a fairly high probability for isolation of populations—a necessary condition for speciation.

Invertebrate biodiversity ranges from the very low in the polar deserts (a few insect and nematode species) to the extremely high diversity found in some rain forest areas, particularly of insects. Note that the high biodiversity in rain forest may be due partly to the high humidity, which reduces the environmental limitations of animal groups that have limited terrestrial adaptations (e.g., slugs). The various ecosystems and habitats that harbor terrestrial invertebrate diversity are described elsewhere in the Encyclopedia.

See Also the Following Articles

INVERTEBRATES, FRESHWATER, OVERVIEW
INVERTEBRATES, MARINE, OVERVIEW

Bibliography


I. Clarification of Concepts

A. Island Biogeography

Biogeography is a scientific approach to understanding the distribution and abundance of living things, the biota, on our planet. Island biogeographers are primarily interested in isolated areas and the study of fragmented life zones and their relation to the biota. But what living things or biota are included? Nearly all groups are studied: plants, birds, insects, other animals, humans, fungi, fishes, disease organisms, and so on. From this list it is clear that biogeography is not a single discipline. Instead, it is a unifying principle for scientists of different disciplines. The unifying principle is their interest in the distribution and abundance of the organisms with which they have a greater familiarity. Thus, botanists, ornithologists, entomologists, mammalogists, mycologists, and anthropologists can all come together and be unified by their interest of biogeography.
The study of distribution and abundance of biota can be applied at any level of scale in space and time.

1. Spatial Scales

Traditionally, biogeographers were, and still are, interested in biotic patterns occurring at global and intercontinental scales. Geographically speaking, these are small scales that provide for broad overviews on maps or satellite images. Scientists were concerned with distinguishing and mapping large-area distribution patterns, based on landscape physiognomy, such as tropical rain forests, deserts, savannahs, and temperate grasslands, commonly called “biomes.” A parallel concern was to distinguish broad patterns of species distribution, which could be mapped as “biotic provinces,” areas distinguished as different centers of biodiversity with their “own” floras and faunas. Lately, biogeographers have become concerned with distribution patterns occurring at more detailed spatial scales, such as within-archipelago migration patterns, or the distribution of biota along individual mountain slopes. Moreover, the aspect of abundance versus rareness of species and other taxa has become of concern to biogeographers, an area treated traditionally by ecologists.

2. Timescales

Present-day distribution and abundance patterns of biota are usually the result of past events and historical processes. Historical processes that led to present-day distribution patterns of biota are often measured at geological timescales in millions of years. Similarly long timescales are considered for biological evolution. But organic evolution by mutations and hybridization can occur at any timescale. Other timescales of biogeographic interest relate to the concepts of succession following disturbances and to phenological change. Successional change in terms of primary succession, which refers to vegetation development on new geological surfaces, can be considered at long timescales involving hundreds and thousands of years. Primary succession can also be related to soil development (= pedogenesis). Secondary succession, defined as following disturbances on already developed soils or previously vegetated substrates, may be considered at plant-demographic timescales, involving a few years, decades, or hundreds of years, if based on the life cycles of certain long-lived tree species. Phenology relates to seasonal changes in biota. The latter two biotic changes—time concepts—succession and phenology—were developed from ecological research in biogeography. Thus, biogeography as a research approach to discover chronological changes overlaps with those aspects of ecology that deal with distribution and abundance of biota at successional and phenological timescales and also with evolutionary research that seeks to unveil migration patterns and phylogenetic relationships over long and short timescales.

3. Geobotany

The term geobotany is derived from geographical botany, the biogeographical study of plants. In an effort to clarify Central European and Anglo-American terminology in this broad study area, Mueller-Dombois and Ellenberg (1974) synthesized the specializations within the area of geobotany in tabular form, which is reproduced in Table 1.

Biogeography was originally understood as consisting primarily of phytogeography and zoogeography. Today, following a review of recent textbooks (e.g., Cox and Moore, 1993; Hengeveld, 1990; Huggett, 1998; and others), one can consider all the ecological disciplines listed under Anglo-American equivalents (Table 1) as aspects of plant biogeography.

II. THREE INSPIRING THEORIES

A. The Theory of Island Biogeography

This theory, originally proposed by MacArthur and Wilson (1963, 1967), is practically synonymous with the concept of “island biogeography.” It proposes that species equilibria are formed on islands in relation to the size of land area and its distance from biotic source areas. An equilibrium is suggested where the rate of invasion equals the rate of extinction of island biota. These are the intersection points on Fig. 1. Curves sloping from the left to the abscissa represent decreasing rates of invasion of biota, from near to far source areas; curves sloping from the zero point on the abscissa up to the right ordinate represent increasing rates of extinction, from large to small islands. The curves are based on a complicated mathematical model. However, the model is easily understood from Figure 1. For example, a large archipelago, such as the Fijian Islands, situated nearer to continental source areas for the dispersion of biota, would have richer species equilibria than a large archipelago, such as the Hawaiian Islands, which is much farther removed from any biotic source area. The reality of species equilibria is highly questionable, yet the theory has been and continues to be very inspirational. If applied not simply to species per se, but instead to different life forms, such as indigenous trees, ferns, or shrubs, the results may show differing...
### TABLE I

<table>
<thead>
<tr>
<th>Area of specialization (and synonyms, European terms)</th>
<th>Subject matter</th>
<th>Anglo-American equivalents (and synonyms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floristic geobotany</td>
<td>Study of geographic distribution of plant taxa and their evolutionary relationships</td>
<td>Plant geography (phytogeography)</td>
</tr>
<tr>
<td>Sociological geobotany (vegetation science, plant-ecology, phytosociology)</td>
<td>Study of composition, development, geographic distribution, and environmental relationships of plant communities</td>
<td>Synecology (community ecology, plant ecology in part)</td>
</tr>
<tr>
<td>Ecological geobotany (plant ecology)</td>
<td>Study of physiological functions of individual organisms in the field environments and communities; life-history studies of species or ecosystems</td>
<td>Autecology (physiological ecology, population ecology in part)</td>
</tr>
<tr>
<td>Demography (population ecology)</td>
<td>Study of structure and function of populations</td>
<td>Population ecology</td>
</tr>
<tr>
<td>Synecology (habitat science; ecosystem research)</td>
<td>Study of habitat factors and the physiological response of species and species groups to these factors; study of community functioning, and niche functions of plant populations in an ecosystem context</td>
<td>Ecosystem ecology (community process ecology; functional ecology; system ecology)</td>
</tr>
<tr>
<td>Historical geobotany</td>
<td>Study of historical origins and development of populations and communities</td>
<td>Paleobotany (paleoecology)</td>
</tr>
</tbody>
</table>

*From Mueller-Dombois and Ellenberg (1974).*

---

**Values of species richness and endemism in relation to size of island areas and degrees of isolation.** Moreover, these in turn may lead, with additional ecological studies, to a better understanding of the function of biodiversity in different island ecosystems.

At this point, another limitation of this theory should be mentioned. This relates to the fact that size of island area is only a most general predictor of species richness. At least elevational range and substrate type should be added to make the theory more predictable. This brings us to the next theory.

**B. The Biome Theory**

This theory predicts that there are broad life zones that are indicated by groups of biota of key plant life forms, which are controlled by certain broad-area climatic and edaphic (soil) parameters. For example, the biome theory predicts that in mountainous environments there are altitudinal life zones (Holdridge, 1967) or vegetation zones (Mueller-Dombois and Ellenberg, 1974) that can be distinguished by tree species and other life forms into lowland, upland, and high-altitude zones. Also,
such familiar terms as desert, grassland, deciduous forest, coniferous forest, and tundra depict different latitudinal biome types, which in turn can be defined by climatic parameters as "zonobiomes" (Walter et al., 1975). It is assumed that conditions for life within a biome are more homogenous than life conditions across different biomes (such as grassland versus desert).

Applied to island biogeography, the biomes theory lends itself to a more appropriate refinement in the analysis of biodiversity than is offered by the two ecosystem parameters—size and isolation—in the above described theory of island biogeography. For example, comparative biodiversity research within a Pacific-wide biome type, such as the montane rain forest on volcanic high islands of basaltic origin, is scientifically more satisfying than biodiversity research based simply on size of island area. The size approach groups different types of islands into the same category, which thus can be a very heterogeneous category that includes atolls, raised limestone, and volcanic high islands. Using the biome theory, environmental gradients can be studied among physically similar islands. Examples of environmental gradients for island research of the same biome type are given in Fig. 2. Atoll and reef islands can be considered as belonging to the same biome, in this case, the same "pedobiome," meaning they have similar, marine-derived, substrates and similar low elevations.

C. The Theory of Succession

Both the island biogeography and biome theories thus outlined contain elements of succession or community and ecosystem development. MacArthur and Wilson (1967) speak of five fundamental processes as the most difficult to study in biogeography. These are listed as (a) dispersal, (b) invasion, (c) competition, (d) adaptation, and (e) extinction of species. In his concept of succession, the early, influential American ecologist Clements (1916) recognized six processes: (a) nudation, (b) migration, (c) ecesis (establishment by reproduction), (d) competition, (e) reaction (habitat change through organisms), and (f) final stabilization, the climax community. All of these processes are of concern in biogeographical research. For example, the process of "dispersal" and "migration" of species among islands and their "adaptation" in terms of speciation were chosen as the main topics in a recent treatment of Hawaiian biogeography (Wagner and Funk, 1995). Other processes, such as the establishment and regeneration of populations (ecesis) after major disturbances (mutations), their "invasion" relative to "extinction" or their "final" assemblages in communities or ecosystems, also fall into the realm of biogeographical research. Clements proposed a final stabilization, called climax, while MacArthur and Wilson proposed a dynamic (final) species equilibrium as explained earlier. Both, the climax and species equilibrium concepts have been severely criticized. Yet dynamic equilibria remain an area of ecological and biogeographical interest because an understanding of dynamic processes is essential for an improved theory of island biogeography.

In connection with our long-term research on the native Metrosideros forest dieback in Hawaii, I introduced a model of ecosystem development based on the theory of succession (Mueller-Dombois, 1986). This is diagrammatically portrayed in Fig. 3. The model addresses the concepts of climax as well as those of primary and secondary succession and habitat change with time in a single island biome, the montane tropical rain forest of Hawaii. Here, the process of nudation can be a new pahoehoe lava flow or a volcanic ash blanket. Both, the volcanic ash and pahoehoe substrates, may achieve a "climax" in vegetation development in about 1000 to 5000 years in the rain forest climate. After that, a regression phase sets in very slowly, characterized by cation leaching, increasing occlusion of phosphorus (Crews et al., 1993), formation of secondary aluminum (gibbsite) and iron (goethite) minerals, and advanced desilication (Fox et al., 1991). With time, here estimated as 1 million years, the forest undergoes a number of generation turnovers in the form of canopy breakdown or gap formation and recovery. These may be synchronized over larger or smaller areas, depending on the disturbance regime, the cohort structure, and the aging pattern in the forest mosaic. The different kinds of dieback or canopy failure, depicted on the diagram as demographic events, reflect the habitat changes in terms of soil water and nutrient relations that occur over the long timescale. The model also implies that after the plant biomass or biophilic nutrient climax, forest recovery yields successively less tall Metrosideros forests. Associated with the progressive and regressive phases of ecosystem development, different species assemblages also occur. Species equilibria have not been detected among native species. The question of invasive nonnative species is a separate major problem of island biogeography that is discussed in another chapter of this book.

III. VEGETATION ANALYSIS METHODS

A. Floristic Checklist Methods

"Floras" are lists of plant species. When more elaborate, they are books with species descriptions, keys, and illus-
FIGURE 2. Latitudinal gradients of mean annual rainfall in eastern Micronesia, from the northern Marshalls to the southern Gilberts (Kiribati) and Ellice Islands (Tuvalu), summarized by Fosberg zones. Adapted with modifications from Stoddart (1992).
tations for plant identification. Floras usually give little information on vegetation, which is the most visible component in terrestrial ecosystems, there performing the function of primary producer. "Vegetation" is the plant cover of an area in which species play different roles in terms of their abundance or rarity and in terms of their life forms, life histories, and physiology. In nature, plants often interact with one another in communities. "Plant communities" can be distinguished by differences in life-form structure—such as forests versus shrub- or grassland—or by differences in species assemblages. In the study of island biogeography, floristic checklist methods can be applied at the level of biogeographic provinces, which may serve for floristic comparisons of archipelagoes. They can also be applied among individual islands of the same archipelago. At both these broad levels of scale, tests of similarity versus dissimilarity can be applied from floristic records or subregional floras where these exist. However, field research is needed if one wants to compare the floras within the same biome on different archipelagoes, because this information is usually not available. A simple “walk-through method” can provide this information without formalized sampling designs, such as plots or transects. Such broad floristic surveys also help in familiarizing with the territory and, when used together with aerial photographs and topographic maps, they can result in vegetation maps outlining tentative communities. However, floristic comparisons among communities of the same biome in an archipelago, such as the rain forest across the high Hawaiian islands, require a more formalized checklist method in representative sample plots, such as described by the releve method (Mueller-Dombois and Ellenberg, 1974, discussed later).

### B. Vegetation Sampling Methods

A number of field methods are available. Their proper use depends on the purpose of sampling and the type of vegetation. Here, I will consider plant biodiversity as the general objective of sampling.

1. The Releve Method

This is a widely applied and proven floristic checklist method, ideal for many, if not most, plant biodiversity surveys. It is based on plots, called relevés (meaning abstracts), whose sizes are based on the “minimal area concept.”
a. Determining Relevé Size
A minimal area for vegetation sampling can be established by the "nested plot technique." This technique consists of lining out a small quadrat, 0.5 x 0.5 m, and then enumerating all species encountered in this small quadrat. Next, the quadrat size is doubled into a rectangle of 0.5 x 1 m, and additional species are noted. The next doubling is a 1 x 1 m quadrat, followed by 2 m², then 4, 8, 16, 32, 64 m², and so on. At each enlargement, the new species encountered are added. When plotting the cumulative number of species over the area sample, a species/area curve is obtained. In a homogenous vegetation segment, this usually takes the form of a steeply rising curve that levels off with each enlarged quadrat. The result is a curve of "diminishing returns" in terms of number of species encountered with still greater enlargement of area. The area over which the curve levels can be used as the "minimal area." The usual practice is to use a slightly larger area for the size of a relevé. In temperate grasslands, a 10 x 10 m quadrat is often sufficient to satisfy the minimal area requirement. In temperate forests, 20 x 20 m quadrats may give a satisfactory relevé size. In multispecies tropical rain forests, 1 ha (10,000 m²) may not be large enough (see Fig. 4). But 1 ha is now often considered a practical standard for biodiversity research of tropical rain forests.

b. Recording Species in Layers
A relevé record should contain all plant species found in its boundaries. In forest communities, recording is best done in horizontal layers by vertically defined height strata. For example, one may distinguish two tree layers T1 as trees over 10 m tall, T2 as trees from 5 to 10 m tall (subcanopy trees), then a shrub layer from 2 to 5 m (including smaller trees), then a lower shrub layer from 1 to 2 m high, and further a herb layer from 0 to 1 m high, including smaller woody shrubs and trees (usually seedlings). The vertical stratification is an aid in recording. It also can document if canopy trees are regenerating in the same sample stand (or relevé).

c. Distributing Relevés in Predefined Strata
Relevés are best laid out in predefined strata. On mountain slopes, these may be predefined elevational strata, such as 200 m contour intervals in the Alps. In other areas, these are the tentative vegetation segments or communities identified on air photos in conjunction with field reconnaissance and check listing. We have called this process "entitation" (Mueller-Dombois and Ellenberg, 1974), which means defining of entities or map units from air photographs. Depending on the scale and homogeneity of the initial strata or entities, from 2 to 5 relevés are often used to characterize a stratum or entity. Also, depending on the size and complexity of plant biodiversity in an area, from 20 to 140 relevés can amount to a satisfactory biodiversity survey in terms of plant species and plant communities.

d. Estimating Species Quantities
The relevé method as described here is merely a formalized floristic checklist method, which reports the presence (and absence) of species. For classifying plant communities it is desirable to record how abundant or rare a species is in an area. This can be done simply by adding an abundance symbol to each species on the relevé record sheet. The most widely used rating system in European vegetation surveys is the Braun-Blanquet cover/abundance scale consisting of seven symbols as follows:

5 Any number of individuals, with cover more than 3/4 of the reference area (>75%)
4 Any number, with 1/2 to 3/4 cover (50–75%)
3 Any number, with 1/4 to 1/2 cover (25–50%)
2 Any number, with 1/20 to 1/4 cover (5–25%)
1 Numerous, but less than 1/20 cover, or scattered, with cover up to 1/20 (5%)
Pronounced cross) few, with small cover
r Solitary, with small cover

This estimation scale is applied by walking diagonally through the relevé, several times if necessary. It is a crude scale and thus is often criticized, but it takes little time and tells a lot more about the plant composition of a relevé than a mere presence/absence record. However, estimation methods are often not satisfactory in more complex vegetation. In forest vegetation it may be usefully applied only to the undergrowth species and only in relevés that are small enough to allow such estimation with confidence.

2. Quantitative Methods
In forest vegetation it is not possible to get a satisfactory estimate of the quantity of tree species by the cover/abundance scaling method as described earlier. In such more complex vegetation, at least the tree component needs to be measured to obtain a satisfactory abundance or cover estimate. The reason for this is primarily that relevé sizes are too small for obtaining an adequate density or cover estimate of trees.
a. The Count-Plot Method

To get an adequate estimate of tree density per unit area, such as a hectare, one should count trees in many small plots, such as \( \frac{3}{100} \times 5 \) m. The small size of such count plots is merely a convenience for an accurate (100%) enumeration per plot. Larger plots may lead to accidental double counts or omissions, thereby introducing serious errors. Count plots are conveniently laid out along transects. They must be repeated until the cumulative mean number of trees does not fluctuate anymore radically with each additional count plot. This may be accomplished by counting at least 30 to 50 trees/species. It is obvious that rare tree species can hardly be adequately counted by this method. Rare tree species can only be reasonably assessed by searching an area of known size for such species, thereby attempting a 100% survey of their presence. In counting trees, a measure of tree density per unit area, it is necessary for plant biodiversity surveys to count trees by species and by measuring some tree-size parameters. Among tree-size parameters, usually the diameter at breast height (Dbh) is measured. From this, tree basal area is obtained by the area formula \( r^2 \pi \). However, the true basal area is the stem cover at ground level. Where this parameter is desired, one needs to measure also tree diameters at ground level. The Dbh measure is usually preferred in timber volume surveys. Tree volume is established by combining the basal area with a height measure and a tree form factor. More elaborate formulas can be developed for total tree biomass, often sought in productivity studies.

b. Distance Methods

An alternative to the count-plot method are the distance methods. They often lead more rapidly to an adequate count of trees, since they do not require the layout of plot boundaries. Instead, the average distance between trees is used to determine the mean area per tree and thereby the density per unit area. The mean area per tree is obtained by squaring the mean distance. This method works well under certain restrictions. Two methods are briefly outlined.

i. The Point-Centered Quarter Method

A random starting point is established in the forest to be measured. Four distances are measured from this point to the
nearest tree in four quarters, forming each a 90° exclusion angle over the point. The sampling point therefore is used like a rectangular cross, and the distance from the point to the nearest tree is measured in each quarter. When repeated at 20 sampling points, established at random or along a compass line, one obtains 80 distances. The mean distance squared then gives an estimate of the mean area per tree. The number of trees per acre (≈ 4,000 m²) or hectare (10,000 m²) is obtained by dividing such reference area by the mean area/tree. The point-centered quarter method is relatively simple to apply, and it compares well to the count-plot method. However, two restrictions apply: the trees should be randomly distributed, and no tree should be measured twice. The second requirement is usually easy to follow as it requires that the sampling points are far enough apart so that no tree is measured twice.

ii. The Wandering Quarter Method. This second recommended distance method is often less affected by the distribution pattern of trees (random versus clumped) as it traverses for the same number of distance measures through a larger area. The method begins with establishing a random sampling point in the forest segment to be sampled. A 90° exclusion angle is established in a certain compass direction, and the nearest tree to the sampling point is measured. That nearest tree then becomes the next sampling point over which a 90° exclusion angle is established in the same compass direction. From this tree the distance is measured to the next tree in that exclusion angle. The procedure is continued for 20 distances in that same direction, moving from tree to tree and always using the same 90° exclusion angle. This results in a wandering movement or zigzag line depending on where the next tree occurs within the 90° exclusion angle. After 20 distances, the direction may be changed following a right angle for another 20 distances, then once more at a right angle in a parallel reverse direction to the first zigzag line, and finally, 20 distances in direction of the starting point. In this case, the sampling also includes 80 distances, but the area traversed is larger than in the point-centered quarter method.

c. Measures Obtained in Count-Plot and Distance Methods.

The two types of methods provide for the same three parameters desired often in quantitative vegetation surveys. These are frequency, density, and cover.

“Frequency” is simply the occurrence of a species in any number of plots in relation to the total number of plots in the sample. In the distance methods, frequency is the occurrence of a species at a sample point out of the total number of sample points. Frequency is a mixed measure of abundance and distribution. It is a relative measure as it depends on the plot size used in sampling.

“Density” is the actual count of individuals per species in the total number of count plots. The sample plot area is then adjusted to a standard reference area, such as a hectare. In the distance methods, the number of individuals is obtained from squaring the mean distance of all trees and dividing this measure into the reference area. The number of individuals per species is then obtained from the proportion of species among the total number of individuals.

“Cover” as stem cover per species is obtained from totaling the individual basal area measurements in the sample of count plots or points and by relating this measure to the reference area. For examples, see Mueller-Dombois and Ellenberg (1974).

Another aspect relates to the tree-size stratification in count-plot and distance sampling. In addition to sampling mature trees/species, it is ecologically desirable in biodiversity inventories to also establish the density of seedlings and saplings. First, mature trees have to be defined. This is often done by setting height and/or diameter limits. “Mature trees,” for example, may be all individuals 5 m and taller, thereby forming a broad size class. “Saplings” may be defined as individuals from 0.5 to 5 m tall and “seedlings” as those from 0.1 to 0.5 m tall. Smaller individuals may be considered as “germinants” (i.e. not yet well established or ephemeral individuals). Smaller subplots are often required for seedling counts because of their smaller size and frequently greater density. In the distance methods, usually only mature trees are measured because of the random pattern requirement for accuracy. In such cases, seedlings and saplings may be counted in subplots of appropriate sizes at predetermined sampling points.

Application of appropriate sampling methods requires some basic understanding of sample theory and field experience. When the sampling objective is clarified and the nature of the vegetation experienced, it is usually possible to prescribe an appropriate approach from a combination of established field analysis methods.

d. Size of Sample Plots.

Figure 5 displays six sample plot sizes that have been used in proven studies. They are drawn to the same

ISLAND BIOGEOGRAPHY

573
scale. Like the small, $20 \times 10$ m grassland relevé, all plots were placed in predetermined entities or tentative map units to obtain representative data. Daubenmire's (1968) preferred plot size of $15 \times 25$ m was applied in the mountainous interior forests of the Pacific Northwest. He preferred to assess the undergrowth vegetation within such plots along 25 m transects by determining frequency out of 50 systematically placed 0.1 m$^2$ frames. Percent cover in terms of shoot cover was also estimated similarly as with the Braun-Blanquet cover/abundance scale. But 50 placements only gave an area sample of $5 \text{m}^2$, which likely did not fulfill the minimal area requirement as used in the relevé method. The larger plots in Fig. 5 refer to quantitative studies as discussed earlier. Their sizes are based on the principle that tree species need to be counted in numbers that give reliable density and basal area estimates. For obtaining a minimum tree count of 30 to 50 trees/species, larger plot areas are needed, since individual mature trees can occupy mean areas of 25 to 100 m$^2$ or larger.

The larger sample plots with quantification of tree counts obviously require a much greater time effort than a relevé with a species record supplied with a cover/abundance estimate. Thus, one has to be very clear about the purpose of the field effort. The relevé method is particularly useful in plant biodiversity studies aiming at classifying vegetation by species groups, while quantitative methods are essential if the objective is repeated monitoring of plant biodiversity in permanent plots.

C. Data Processing and Display Methods

1. Data Analysis: An Island Example

During the 1970s, we did a multidisciplinary study in Hawaii under the auspices of the International Biological Program (IBP). Our main objective was to study the biological organization in relatively undisturbed natural communities. We focused on two areas, an upper montane rain forest (the Kilauea Rain Forest) and the east-slope of Mauna Loa in Hawaii Volcanoes National Park (Fig. 6). In the Kilauea Rain Forest, which was dominated by three native keystone species, Acacia koa, Metrosideros polymorpha, and tree ferns (Cibotium glaucum), we established a large sample plot, $1000 \times 800 \text{ m}^2$ greater by 20 ha than the largest plot on Fig. 5. Our data sets were processed by species in life form and guild categories to get quickly from forest community structure to function by niche differentiation. This aspect is dealt with in part III of our synthesis volume (Mueller-Dombois, Bridges, and Carson, 1981), and will not be elaborated here.

On the east slope of Mauna Loa, we tried to answer a basic question asked by MacArthur (1972, p. 161), “Do different plant species change synchronously, or does each have independent distribution?” This question arose from a dispute over the nature of species distribution within and across neighboring vegetation zones or biomes. To clarify this question, we established 14 transect sampling sites (1–14 on Fig. 6) in predefined altitudinal intervals from 1100 to 3080 m. These sites were studied for 12 other organism groups besides plants. For the plant distribution study, we established 48 relevés throughout this mountain transect, approximately three relevés clustered around each of the 14 transect sites.

2. Two-Way Table and Dendograph Techniques

Following the field study, the 48 relevés were processed by the “two-way table technique.” That required entering all relevés into a single “raw table,” whereby the relevé number appears at the head of the table and all species names appear at the left side of the table. The
FIGURE 6. Map of Hawaii Volcanoes National Park with the 14 Mauna Loa transect sites and the 80 ha Kilauea forest site used for integrated biodiversity sampling during the Hawaii International Biological Program (IBP). From Mueller-Dombois, Bridges, and Carson, 1981.

TABLE II
Extract of a Final Two-Way Table

<table>
<thead>
<tr>
<th>Species group 2</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>PELTER</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>ASP ADI</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>SPECIES group 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAC SAIH</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>R1</td>
</tr>
<tr>
<td>COP IRN</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>122</td>
<td>122</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>LUG PAW</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>RAC CL</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>COP MCMN</td>
<td>112</td>
<td>112</td>
<td>112</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>PUE THU</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>POL SEL</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Note: Rows used: 50% inside, 10% outside.
species magnitude, or score values, are entered for each species in the respective "relevé columns." When all data are entered in this way, one can see in the "species rows" which species are present (and with what score value) or absent in the "relevé columns." The process then begins of sorting species of similar distribution together by repositioning the "species rows," then also the "relevé columns" are resorted to bring those relevés together in the table that contain species of similar distribution. The "row and column sorting process" is reiterated until the table displays an optimal structure in the sense that one can interpret the clustering of species and relevé group in an ecologically meaningful way. This data processing, formerly done by hand, can now be done mathematically by using one of many multivariate analysis techniques and similarity indices.

An example of a two-way table extract of our 48 Mauna Loa relevés is given in Table II. For corroboration of species distribution and relevé clustering trends one can use various mathematical ordination methods or data classification methods. Ordination methods (not illustrated here) present the same data geometrically, while classification methods present the same data in the form of dendrographs (see Fig. 7). The three data processing and display techniques are corroborative analytical tools. Particularly the ordination and dendrograph techniques are rather abstract and difficult to interpret by the reader.

3. Improved Display Methods

It is more "reader friendly" to go a step further in displaying the results of a multivariate data analysis as shown in Fig. 8. This data display over the Mauna Loa transect with its 14 IBP transect sites clarifies that both multivariate analyses, the two-way table and the dendrograph techniques, gave closely similar results. Both data treatments resulted in the definition of seven altitudinal vegetation zones as named in the legend of Fig. 8.

The answer to MacArthur’s (1972) question is given
in Fig. 9. It displays the distribution of 17 key species, which are primarily responsible for the pattern of the seven vegetation zones identified by multivariate techniques. Several species clearly change synchronously, while others have independent distributions along this island mountain slope. The answer to MacArthur’s question is this: There is more than one pattern of plant species distribution along this geologically young tropical mountain slope. Some distributions coincide closely with one another; others do not. However, the
species distribution data give clear evidence of the existence of altitudinal vegetation zones as established from mathematically independent multivariate analysis techniques.

IV. AN EXPERIMENTAL SAMPLING DESIGN FOR ISLAND BIOGEOGRAPHY RESEARCH

Transects are generally the most efficient sampling designs in ecological field research. They can often be arranged to cut through a maximum of variation in representative habitats and biodiversity assemblages of an area to be studied, using the shortest distance. A still more important aspect is the use of transects to control environmental variation in its effect on biodiversity. In this way, transects can be used as “experimental sampling designs” by aligning them along changing environmental control factors to study their effects on biodiversity in environmental settings that are relatively uniform or constant. Sampling transects along a mountain slope, on which mean annual temperatures and rainfall change in predictable ways, are good examples. Another example is given in Fig. 2, in which the western Pacific atoll islands are aligned along mean annual rainfall gradients in both the northern and southern hemispheres.

A. Horizontal within Ecosystem Gradients

In addition to the atoll islands there are several other Pacific-wide biomes. A particularly important one is the upland rain forest on the volcanic islands of the tropical Pacific. This forest is important for two primary reasons: it is still extant on many volcanic high islands as indigenous forest harboring most of the endemic species, and it performs the function as watershed cover in the mountainous interiors of these islands. The volcanic high islands typically are small, isolated land fragments that protrude as mountaneous terrain above the vast Pacific ocean surface. Their interior upland forests likewise are fragments that belong to the same larger ecosystem or biome. However, from archipelago to archipelago, these biome fragments are occupied by different sets of species due to their past biogeographic isolation. Connecting these biome fragments horizontally across the ocean by a system of transects provides for a within-biome research design. Here, the broad habitat features are kept uniform, while the indigenous biodiversity sets change from island to island. This truly is a research approach to biodiversity, which Pielou (1979) defined as “geoecology,” the study of recurrence of similar communities in similar habitats, which are occupied by different sets of species. The term geoecology is an abbreviation of geographical ecology in the conceptual sense used by MacArthur (1972), who contributed substantially to the foundation of island biogeography.

B. Vertical between Ecosystem Gradients

A functioning watershed cover is an essential resource component in all Pacific high islands. From here, the fresh water flow begins to be regulated and then influences almost all lower lying island ecosystems, the
freshwater wetlands, mangroves, estuaries, fish ponds, fringing reefs, the entire coastal zone, and agriculturally used lowland areas. The ecosystem services of the upland forest watersheds have not been studied in the Pacific islands, in spite of the fact that they have been part of the traditional land-use system in the Pacific high islands, in Hawai‘i known as the ahupua‘a system. This vertically arranged land-use system was recognized in the Hawaiian Islands as comprising four integrated management zones, “wao la‘au,” upland forest or wilderness area, to be left alone, “kula,” the more open foot hill region, “wao kanaka,” the agricultural zone in the lowlands, and “kahakai,” the coastal zone. This vertically arranged multiecosystem human support system proved to be an optimal land-management scheme for the indigenous islanders in the past. Vertically arranged transect sites, using the fresh water flow and hydrology as unifying parameter, may also prove to be a good organizing principle for interdisciplinary research focused on the function of biodiversity at the landscape level.

V. THE PABITRA INITIATIVE

The acronym PABITRA refers to the Pacific-Asia Biodiversity Transect outlined on Fig. 10. The transect system connects the high island archipelagos with indigenous tropical rain forests still extant in their interior uplands. They range from east to west, from the paleotropical outlier islands, the Hawaiian archipelago and the Marquesas, to the biotically rich continental islands of southeast Asia. The PABITRA concept arose from the concluding chapter of a recent book (Mueller-Dombois and Fosberg, 1998) with the chapter heading “The Future of Island Vegetation.” Simply put, the indigenous island vegetation only has a future if the scientific community and conservation managers redouble their efforts in learning to understand the function of island biodiversity and thereby take appropriate conservation measures. Another impetus came from a Japanese research initiative, called DIWPA, Diversitas in the Western Pacific and Asia. This program is the Pacific-Asia representative of the new worldwide program “DIVERSITAS,” promoted by the International Union of Biological Sciences (IUBS) with support from UNESCO and ICSU (The International Council of Scientific Unions in Paris, France). The promotion of DIVERSITAS started effectively with a week-long forum titled “Biodiversity, Science and Development: Towards a New Partnership” in September 1994 (Youne’s, 1996). Since then, DIWPA has focused on a “Green Belt,” a north-south transect in western Asia for monitoring
biodiversity of forests and freshwater lake systems at 42 transect sites. A parallel running “Blue Belt” is also planned with focus on coastal zones and nearshore marine habitats. The PABITRA Initiative is considered a separate, but attached, “Pacific Island Branch of DIWPA,” working cooperatively under the Ecology, Conservation, and Environmental Protection (ECEP) Division of the Task Force on Biodiversity in the Pacific Science Association, which is chaired by the writer.

See Also the Following Articles

- BIOGEOGRAPHY, OVERVIEW • DISPERSAL BIOGEOGRAPHY • SPECIES AREA RELATIONSHIPS • SUCCESSION, PHENOMENON OF

Bibliography


ISOPTERA

Takuya Abe† and Masahiko Higashi†
Center for Ecological Research, Kyoto University, Japan

I. Introduction
II. Profile of Termites
III. Symbiosis
IV. Social Life
V. The Global Diversification of Termites
VI. Termites in Ecosystems

GLOSSARY

alate An imago still possessing its wings.
caste A group of individuals in a colony that are both morphologically distinct and specialized in behavior.
imago The adult insects; the final developmental stage when insects possess wings.
larva An immature individual without an external sign of wing buds or soldier morphology.
neotenic A secondary reproductive with juvenile morphological characters. Neotenes derive from larvae, nymphs, pseudergates, or workers through at least one special moul.
nymph An immature individual on the developmental pathway to the imago and which possesses external wing buds.
presoldier A transitional morph that always precedes the soldiers; an unsclerotized individual whose head shows signs of soldier differentiation.
primary reproductives Dealkate reproductives that founded a new colony after nuptial flight.
pseudergate (false worker) A temporarily nonreproductive individual serving the colony in nutrition, construction, or brood care, which results from a late, reversible deviation from the pathways to the imago and is characterized by reduced wing buds compared to nymphs of the same stage.
secondary reproductives Reproductives that differentiated in an established colony, whatever their origin and morphology. They may be supplementary reproductives if older reproductives are still present or replacement reproductives if not.
soldier An individual with a strongly sclerotized head showing defensive adaptations, such as enlarged mandibles, a stopplerlike shape, or a frontal gland able to produce a defensive secretion.
queens and kings Females and males actively reproduce in a colony.
worker An individual resulting from an early, irreversible deviation from the pathway to the imago, and performing helper tasks. Workers are primarily characterized by the loss of the ability to proceed to the winged imago, but they need not be permanently sterile.

ISOPTERA (TERMITES) is a small order of insects containing about 2650 species. The oldest fossil is found in limestone deposits of 130 million years ago (Cretaceous). Termites consume cellulose, the most abundant organic matter on the earth, and are superabundant in tropical regions. Then termites are one of the most

† Deceased.
prosperous animals on the earth from the viewpoint of history length and abundance. This chapter presents the global diversification process of termites with special reference to their social evolution and digestive symbiosis with microorganisms.

I. INTRODUCTION

The Isoptera is a relatively small order of insects that consist of 31 orders, including about 2650 described species. The order name Isoptera originates from the fact that the imago (adult) has wings of equal size (Isos in Greek meaning the same and ptera meaning wings). The insects of Isoptera are commonly called termites in English. Due to their poor abdominal sclerotization and white appearance, they are often called “white ants” in many temperate regions. However, they are not phylogenetically closely related to ants, but to cockroaches and mantises. Termites are sometimes black, reddish brown, or yellow in body coloration in tropical regions.

Termites are widely distributed in tropical and subtropical regions, spreading from humid forests to savannas and even arid areas. Their biomass approaches 10 g wet weight/m². Only a few animal groups such as human beings, earthworms, herbivorous mammals in African savanna, and ants claim this order of biomass. The basis of this extraordinary prosperity of termites lies in their feeding habits. They consume cellulose, which is the most abundant organic matter on the earth. The ecological basis of their prosperity is their symbiosis with microorganisms and their highly developed social organization.

The symbiosis produces novel abilities that cannot be attained by either of the partners alone. In particular, the symbiosis between higher organisms and microorganisms often creates capabilities for exploiting new food resources such as dead plant material containing large amount of cellulose, which is a potential energy source but difficult for most animals including human beings to utilize. On the other hand, sociality enhances the efficiency of resource exploitation. Either of these or both together may cause a rapid growth and spatial expansion of the population, leading to diversification.

Animals in the tropical rain forests where the species diversity is the highest in the world are characterized by the dominance of social insects (termites, ants, social wasps, and social bees), and the dominance of symbiotic associations. Termites possess both of those attributes. Termites and ants occupy about one-third of total animal biomass in Amazonian tropical rain forests. The symbiosis and sociality in termites, together with their super abundance, brings them to play a keystone role of “super decomposer” of dead plant material in tropical terrestrial ecosystems. They consume and transform a large amount of nitrogen-poor dead plant material into nitrogen-rich termite body, which is in turn consumed by a great variety of animals ranging from ants and spiders to chimpanzees and human beings. Thus they form the basis for a large food web in the ecosystems, although they are well-known pests of agriculture and wooden buildings.

Furthermore, termites are prominent ecological engineers, modifying the soil properties by constructing huge mounds and long subterranean galleries and providing many animals and plants with heterogeneous habitats. It is also possible that they are an important player in global change scenarios because they have high biomass and emit a large amount of greenhouse gases such as methane as well as CO₂.

There are several landmarks in the process of global diversification of termites: (a) acquisition and loss of symbiotic flagellates, (b) social evolution, (c) change of life type from one-piece type to separate type, and (d) change of dominant feeding habits from wood feeding to soil feeding. This chapter explains an outline of the diversification of termites in relation to those events as well as their ecosystem functions.

A textbook of termites, Biology of Termites, was edited by Krishna and Weesner (1969, 1970), and monographs of termites were published by Grasse (1982, 1984, 1986). A new textbook, Termites: Their Symbiosis, Sociality and Global Diversification, was recently edited by Abe et al. (2000).

II. PROFILE OF TERMITES

A. Phylogeny and Fossil Records

The Isoptera, which includes 280 genera and more than 2650 described species, is a small order, considering that more than 0.9 millions species of insects have been described. Termites include seven families: Mastotermitidae, Hodotermitidae, Termopsidae, Kalotermitidae, Rhinotermitidae, Serritermitidae, and Termitidae. Principal genera of the world termite fauna and their distribution are given in Table 1.

The phylogeny of termites is shown in Fig. 1. Three families, Mastotermitidae, Hodotermitidae, and Termopsidae, are the oldest living lineages. These three families together with Kalotermitidae, Rhinotermitidae, and Serritermitidae are called lower termites, which are
<table>
<thead>
<tr>
<th>Division</th>
<th>Distribution</th>
<th>Division</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Family Mastotermitidae</strong></td>
<td></td>
<td><strong>Microtermitidae</strong></td>
<td>Worldwide except for</td>
</tr>
<tr>
<td>Mastotermitidae</td>
<td>Australia</td>
<td></td>
<td>North America</td>
</tr>
<tr>
<td>Family Kalotermitidae</td>
<td></td>
<td><strong>Amitermes</strong></td>
<td>Worldwide</td>
</tr>
<tr>
<td>Kalotermitidae</td>
<td>Worldwide</td>
<td><strong>Drepanotermitidae</strong></td>
<td>Australia</td>
</tr>
<tr>
<td>Cryptotermitidae</td>
<td>Worldwide, except for temperate</td>
<td><strong>Gnathostetermitidae</strong></td>
<td>North America, temperate</td>
</tr>
<tr>
<td>Neotermitidae</td>
<td>Eurasia</td>
<td><strong>Subfamily Termitinae</strong></td>
<td>Old World tropics</td>
</tr>
<tr>
<td>Rugotermitidae</td>
<td>Polynesia, New World tropics</td>
<td><strong>Basilenotermitidae</strong>,</td>
<td></td>
</tr>
<tr>
<td>Neotermitidae</td>
<td>Worldwide</td>
<td><strong>Ctenotermitidae</strong>,</td>
<td></td>
</tr>
<tr>
<td>Neotermitidae</td>
<td>Worldwide, except for temperate</td>
<td><strong>Formicotermitidae</strong></td>
<td></td>
</tr>
<tr>
<td>Neotermitidae</td>
<td>North Temperate Zone</td>
<td><strong>Megagynoleptomtermitida</strong></td>
<td></td>
</tr>
<tr>
<td>Coptotermitidae</td>
<td>New World</td>
<td><strong>Necantermitidae</strong></td>
<td></td>
</tr>
<tr>
<td>Coptotermitidae</td>
<td></td>
<td><strong>Paranotermitidae</strong></td>
<td></td>
</tr>
<tr>
<td>Psamotermitidae</td>
<td>Eurasia, Africa</td>
<td><strong>Procerotermitidae</strong></td>
<td></td>
</tr>
<tr>
<td>Coptotermitidae</td>
<td>Worldwide, mainly tropical</td>
<td><strong>Capritermitidae</strong></td>
<td>Madagascar</td>
</tr>
<tr>
<td>Psamotermitidae</td>
<td>Eurasia, Africa</td>
<td><strong>Subfamily macrotermitinae</strong></td>
<td></td>
</tr>
<tr>
<td>Heterotermitidae</td>
<td>Worldwide, mainly tropical</td>
<td><strong>Acanthotermitidae</strong></td>
<td>Africa</td>
</tr>
<tr>
<td>Ratotermitidae</td>
<td>Eurasia, North America, temperate</td>
<td><strong>Ancistrotermitidae</strong></td>
<td></td>
</tr>
<tr>
<td>Ratotermitidae</td>
<td>Only</td>
<td><strong>Psocotermitidae</strong></td>
<td></td>
</tr>
<tr>
<td>Prothorinaeidae</td>
<td>Worldwide, tropical islands and</td>
<td><strong>Sphaerotermitidae</strong></td>
<td></td>
</tr>
<tr>
<td>Prothorinaeidae</td>
<td>shores</td>
<td><strong>Synacanthotermitidae</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td>Asia</td>
<td><strong>Macrotermes</strong></td>
<td>Asia</td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Microtermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Colobotermitidae</strong></td>
<td>Asia</td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Subfamily Nasutitermitidae</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Eutermes</strong></td>
<td>Africa</td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Vertermites</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Bulgaritermes</strong></td>
<td>Asia, New Guinea</td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Bulgaritermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Sanguitermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Gnotitermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Trinervitermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Armitidae</strong></td>
<td>Asia</td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Constrictotermitidae</strong></td>
<td>New World tropics</td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Convexitermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Cornitermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Curvitermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Labiatermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Paraceritermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Proceritermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Rhynchoritermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Subulitermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Synitermes</strong></td>
<td></td>
</tr>
<tr>
<td>Termesotermitidae</td>
<td></td>
<td><strong>Velocitermes</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Family Serritermitidae</strong></td>
<td>New World tropics</td>
<td><strong>Temnoptermitidae</strong></td>
<td>New World</td>
</tr>
<tr>
<td>Serritermitidae</td>
<td></td>
<td><strong>Nasutitermes</strong></td>
<td>Worldwide, in tropics</td>
</tr>
</tbody>
</table>

*Based on Wilson, 1971.*
characterized by the presence of symbiotic flagellates (Protozoa) in the gut. A remaining family, the Termitidae, is made up of the so-called higher termites, which are characterized by the absence of symbiotic flagellates, although some higher termites harbor symbiotic amoeba in the gut. The species diversity is the highest in Termitidae (the higher termites, ca. 1900 species), occupying about 70% of all species, while the Kalotermitidae have the highest species diversity (ca. 410 species) among the lower termites.

The fossil record of termites is fragmentary and has not been known from Africa. The number of fossil and living species of each family is shown in Table II. The oldest fossil is the worker of Melatermes bertrandi (Hodotermitidae) found in limestone deposits of Spain, dating to 130 million years ago (Cretaceous). A related species, M. arrapena, was discovered in limestone deposits of Brazil, dating to 110 million years ago. The next oldest fossil species is Valditermes brenanac (Mastotermitidae) from England dating to 120 million years ago and V. acutiennis from Mongolia. Other Cretaceous species include Cretatermes carpenteri (Termopsidae), Luteitermes priscus (Hodotermitidae), and Mastotermes sarthensis (Mastotermitidae). Those Mesozoic termite fossils are found from Europe, Asia, and North and South America, indicating a broad Pangean distribution in both temperate and tropical habitats. A living termite, Mastotermes darwinensis, whose distribution is confined to Australia, may well be called a living fossil.

Many fossil species of the Kalotermitidae and Rhinotermitidae, as well as Mastotermitidae, Hodotermitidae, and Termopsidae, have been found in Tertiary period, while only a few fossil species of the Termitidae (higher termites), which have the highest living species diversity, have been found in the Tertiary period.

Thus, the adaptive radiation of the higher termites seems to have occurred in the late Tertiary or Quaternary (Pleistocene) after the breakdown of Gondwana, yet they show a worldwide distribution.

### TABLE II

<table>
<thead>
<tr>
<th>Groups</th>
<th>Cretaceous species</th>
<th>Tertiary species</th>
<th>Pleistocene species</th>
<th>Living species</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Lower&quot; termites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mastotermitidae</td>
<td>3</td>
<td>19</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Kalotermitidae</td>
<td>33</td>
<td>0</td>
<td>411</td>
<td></td>
</tr>
<tr>
<td>Termopsidae</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Hodotermitidae</td>
<td>3</td>
<td>14</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Rhinotermitidae</td>
<td>16</td>
<td>1</td>
<td>305</td>
<td></td>
</tr>
<tr>
<td>Serotermitidae</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Higher&quot; termites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Termitidae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apicotermitinae</td>
<td>4</td>
<td>186</td>
<td></td>
<td>186</td>
</tr>
<tr>
<td>Termitinae</td>
<td>4</td>
<td>702</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macrotermitinae</td>
<td>3</td>
<td>332</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasutitermitinae</td>
<td>659</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>7</td>
<td>96</td>
<td>3</td>
<td>2632</td>
</tr>
</tbody>
</table>

After Kambhampati and Eggleton, 2000; Thorne et al., 2000.
is an interesting problem remaining to be explained because the flight distance of termite alates is usually limited to some kilometers.

Based on the Cretaceous fossil record and living species diversity, two lower termite groups can be distinguished, primitive and modern. Primitive lower termites characterized by Cretaceous species and low-living species diversity include Mastotermitidae, Heterotermitidae, and Termopsidae. The three families are the oldest living lineage as shown in Fig. 1. Modern lower termites characterized by moderate Tertiary and living species diversity include Kalotermitidae, Rhinotermitidae, and probably Serritermitidae.

**B. Distribution**

Global generic distribution of termites is shown in Fig. 2. Termites occur between the latitudes of 30–45N–40–45S and are mainly distributed in tropical regions, especially in western and central Africa, southeast Asia, northeastern Australia, and South America. The species diversity drops from 10 degrees north and south of the equator, but this gradient is not uniform on both sides of the equator (Fig. 3).

The three groups of termites—primitive lower, modern lower, and higher termites—show different distributions. Primitive lower termites show a fragmental or disjunctive distribution. The Mastotermitidae have only one species, *Mastotermes darwiniensis*, which is distributed in the savanna of northern Australia, although

**FIGURE 2** Global generic distribution of termites. The darker the square the higher the generic richness. Maximum generic richness is the southern Cameroon square. From Eggleton, 2000.

**FIGURE 3** Latitudinal gradients of termite generic richness north and south of the equator. From Eggleton, 2000.
more than 20 species of Tertiary fossils are known from Europe, Australia, and North and South America. The Hodotermitidae including three genera (Hodotermes, Microhodotermes, and Anacanthotermes) and 13 species of grass feeders are distributed in the savanna and arid grassland of Africa and Asia although the family was widely distributed (Spain, Brazil, Europe, Japan, United States) in the Cretaceous and Tertiary periods. Hodotermes and Microhodotermes occur in Africa, while Anacanthotermes occurs in northern African deserts, the Middle East, and southern India. The Termopsidae containing 20 living species show an amphitropical distribution; Three species of Porotermes are distributed in temperate regions of three southern continents (one species in each of Chile, Australia, and South Africa). Several species of Stolotermes are found in Australia, New Zealand, Tasmania, and South Africa, whereas the other three genera are distributed in the temperate and subtropical north hemisphere (Archotermopsis in Kashmir of India, Zootermopsis in North America, and Hodotermitidae in Japan, Vietnam, and China).

The modern lower termites except the Serritermitidae show a worldwide distribution. The Serritermitidae includes only one species, Serritermes serrifer in Brazil, and its phylogenetic position is under debate. The Kalotermitidae are distributed worldwide but tend to be distributed in coastal forests or islands. Among them, Glyptotermes, Cryptotermes, and Neotermes are pantropical, while Kalotermes shows a bipolar distribution. The Rhinotermitidae are widely distributed across tropical, subtropical, and temperate regions, and from wet to dry areas. For example, Psammotermes is one of the few insects persisting in deserts of Africa, the Middle East, and northwest India. Coptotermes, some species of which are serious pests of wooden buildings, are pantropical, partly because they are carried by humans. The distribution of Heterotermes and that of Reticulitermes are complimentary: Heterotermes is distributed in all tropical regions except for Africa, while Reticulitermes shows a Holarctic distribution. Proxinitermes is mainly distributed on tropical oceanic islands.

The higher termites, including the four subfamilies of the Termitidae—Macrotermiteinae, Apicotermiteinae, Termitinae, and Nasutitermitinae—are characterized by central (tropical) distribution. The Macrotermiteinae (fungus-growing termites) are distributed in the central areas of tropical Africa and southeast Asia. Three genera, (Macrotermes, Odontotermes, and Microtermes) are common. The Apicotermiteinae include the Apicotermes and Anoplotermes groups. The Apicotermes group, most species of which are soil feeders, is distributed in tropical Africa, while the Anoplotermes group, which is a soldierless soil feeder, is distributed in tropical Africa, southeast Asia, and South America. The Termitinae are pantropical, including dominant genera of Cubitermes, Termes, Amitermes, and Microtermites. The Nasutitermiteinae are also pantropical, including dominant genera of Nasutitermes, Trinervitermes, and Cornitermes.

It is interesting to note that the higher termites and some lower termites (the Termopsidae, the Kalotermitidae, and Proxinitermes of the Rhinotermitidae) are complementary in geographic distribution: the former show a central distribution (occupying tropical mainland), while the latter show a marginal distribution (localized in temperate regions and tropical coastal areas and islands).

C. Termites as Eusocial Insects

An excellent introduction to eusocial insects was given by Wilson (1971). Eusociality is generally characterized by two generations of conspecific adults living together (that is, forming a group called a colony) and cooperation with each other to the extent that a strong reproductive skew is observed—that is, only one or a few adults in the group lay viable eggs (i.e., division of reproduction). The interior of the nest of a higher termite, Amitermes hastatus, is shown in Fig. 4. The primary queen (the largest individual) and king sit in the middle cell. Workers attend to the queen, to which they offer regurgitated food. Developmental pathways of the fungus-growing termites are schematically shown in Fig. 5.

The eusociality of termites has evolved at least four times as sterile castes (once for the soldier caste and three times for the worker caste). Some wasps, some bees, all ants (Hymenoptera), and all termites (Isoptera) are typical eusocial insects, and sterile members have been found also in gall-forming aphids (Hemiptera) (soldiers), gall-forming thrips (Thysanoptera)(soldiers), an ambrosia beetle (Coleoptera) (workers), some mole-rats (workers), and a coral reef shrimp (soldiers), which are included in eusocial animals.

Termites have similar social organization to the Hymenoptera but differ from them in several important aspects (Table III). Termites are characterized by the diversity of soldiers, while the Hymenoptera are characterized by the diversity of workers. Furthermore, in termites, (a) females and males are both diploid (whereas females are diploid and males are haploid in the Hymenoptera), (b) females (queens) and males (kings) both take care of brood and male remains with female through life (whereas males are produced just before the mating season and die soon after mating in
the Hymenoptera), (c) workers and soldiers are either male or female or both (whereas they are all females in the Hymenoptera), and (d) termites are hemimetabolous (whereas they are holometabolous in the Hymenoptera).

D. Food and Predators

Animals are divided into two large groups in relation to their feeding habits: cytoplasm consumers and cell wall consumers. Most termites are cell wall consumers, whereas the Hymenoptera, including social insects such as bees, wasps, and ants, are cytoplasm consumers. The cell wall of plants consisting of cellulose, hemicellulose, and lignin is very poor in nitrogen but tends to be abundant and clumped, while cytoplasm is rich in nitrogen but tends to be sparsely distributed. Cellulose, a major component of plant cell wall, is decomposed to glucose, an important source of energy and body construction for animals. However, cellulose is difficult for most animals to decompose, especially when it is combined with lignin, which is extremely resistant to chemical degradation.

Termites as a whole feed on live and dead vegetation, lichens, algae, fungi, dung, soil rich in organic matter (so-called humus), other individuals of the same colony (cannibalism), organic-rich portion of termite nests, and skins of vertebrate corpses (Wood, 1978). Most termites in natural ecosystems do not attack living plants, although Microcerotermes sometimes attacks native living trees in Malaysia and some termites consume living native grasses in savannas and grasslands.
Lower termites eat dead wood except Hodontitermitidae, which are grass feeders, while higher termites are rich in feeding habits. Some higher termites take specialized food. Some genera of the Nasutitermitinae feed on lichens on tree trunks: *Hospitalitermes* in southeast Asia, *Constrictotermes* in South America, and *Grallatoritermes* in east Africa, India, the Philippines, and New Guinea. *Ahamitermes and Incolitermes* in Australia and *Ophiotermes and Tuberculitermes* in Africa are obligate inquilines in the nest of other termite species and feed exclusively on the carton of nest material of their hosts.

Soil feeders, which are recognized in the Termitidae, Apicotermitinae, and Nasutitermitinae of higher termites, include 1100 species, which is approximately 45% of the described termite species. The development of soil feeding is a landmark in the evolution of termites, considering that it is rare among animals. The real food source for soil feeders is unknown.

Termites are included in the diets of a wide range of animals. Ants are the most dominant predators. They are either obligatory or opportunistic predators. Some ponerine ants, such as *Termitopone commutata* in South America, *Leptogenys chinensis* in Sri Lanka, and *Megapponera foetens* and some subterranean Dorylus spp. in Africa, are typical obligate predators. The ant with lethal effects on a termite colony changes with the colony's development. For a colony of the fungus growing termite, *Macrotermes michaelseni* in Kenya, it is opportunistic predatory ants (individual foraging ants such as *Pachycondyla* and *Ophitaxanopone* and ants with recruitment system such as *Camponotus* and *Pheidole*) at their early stage, ants with marked recruitment by trail pheromone (e.g., *Megapponera foetens*) at the middle stage, and subterranean group foraging doryline ants (e.g., *Dorylus juvenculus*) at the late stage.

Mammals that specialize in digging out termite nests are pangolins in Africa and India and aardvarks in tropical Africa. Other vertebrates specializing in predation on termite brood centers include armadillos in Texas of the United States, anteaters in South America, sloth bear (*Melursus ursinus*) in India, and the numbat and echidna in Australia. Wild chimpanzees in Africa make and use tools for "fishing" termites (Fig. 6).

### E. Nests and Life Types

Termites are ecosystem engineers. They construct nests of various shapes and sizes, sometimes reaching to 5 m in height (Fig. 7). The diversity of nests is related to the diversity of social evolution, colony size, and feeding habits as well as the establishment of a microclimate suitable for termites to defend their colony against predators. The inner parts of carton nests of some species contain a large amount of wood fibers, which are eaten by the builders or utilized by iniline termites. The most remarkable example is fungus gardens of Macrotermitinae.

Nest material has two origins: exogenous material such as particles of soil and wood transported by colony members and fecal. The cohesion of the nest construction is achieved by saliva moistening of exogenous material. Termites usually use both of fecal and exogenous materials for nest construction except for the Macrotermitinae, which do not use fecal material. Soil feeders and arboreal nest builders mainly use fecal material.

Soil material is important in most epigeous nests (Abe, 1987).

Three life types are distinguishable in relation to nest and feeding sites: (a) a one-piece type, which nests in a piece of wood and consumes only that piece, (b) an intermediate type, which nests in a piece of wood and consumes it as food but also constructs galleries to consume other wood pieces, and (c) a separate type, which nests in various sites such as on living tree trunks, in soil, on the ground surface, and so on, constructing...
### TABLE III
Basic Similarities and Differences in Social Biology between Termites and Higher Social Hymenoptera (wasps, ants, bees)

<table>
<thead>
<tr>
<th>Similarities</th>
<th>Termites</th>
<th>Differences</th>
<th>Eusocial Hymenoptera</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The cases are similar in number and kind, especially between termites and ants</td>
<td>1. Caste determination in the lower termites is based primarily on pheromones; in some of the higher termites it involves sex, but the other factors remain unidentified</td>
<td>1. Caste determination is based primarily on nutrition, although pheromones play a role in some cases</td>
<td></td>
</tr>
<tr>
<td>2. Trophallaxis (exchange of liquid food) occurs and is an important mechanism in social regulation</td>
<td>2. Most species possess soldiers</td>
<td>2. A few species possess soldiers</td>
<td></td>
</tr>
<tr>
<td>3. Chemical trails are used in recruitment as in the ants, and the behavior of trail laying and following is closely similar</td>
<td>3. The worker castes consist of both females and males</td>
<td>3. The worker castes consist of females only</td>
<td></td>
</tr>
<tr>
<td>4. Inhibitory case pheromones exist, similar in action to those found in honeybees and ants</td>
<td>4. Larvae and nymphs contribute to colony labor, at least in later instars</td>
<td>4. The immature stages (larvae and pupae) are helpless and almost never contribute to colony labor</td>
<td></td>
</tr>
<tr>
<td>5. Grooming between individuals occurs frequently and functions at least partially in the transmission of pheromones</td>
<td>5. There are no dominance hierarchies among individuals in the same colonies</td>
<td>5. Dominance hierarchies are commonplace, but not universal</td>
<td></td>
</tr>
<tr>
<td>6. Nest odor and territoriality are of general occurrence</td>
<td>6. Social parasitism between species is almost wholly absent</td>
<td>6. Social parasitism between species is common and widespread</td>
<td></td>
</tr>
<tr>
<td>7. Nest structure is of comparable complexity and, in a few members of the termite family (e.g., <em>Apicoterms</em>, <em>Macrotermes</em>), of considerably greater complexity; regulation of temperature and humidity within the nest operates at about the same level of precision</td>
<td>7. Exchange of liquid anal food occurs universally in the lower termites, and trophic eggs are unknown</td>
<td>7. Anal trophallaxis is rare, but trophic eggs are exchanged in many species of bees and ants</td>
<td></td>
</tr>
<tr>
<td>8. Cannibalism is widespread in both groups (but not universal, at least not in the Hymenoptera)</td>
<td>8. The primary reproductive male (the “king”) stays with the queen after the nuptial flight, helps her construct the first nest, and fertilizes her intermittently as the colony develops; fertilization does not occur during the nuptial flight</td>
<td>8. The male fertilizes the queen during the nuptial flight and dies soon afterward without helping the queen in nest construction</td>
<td></td>
</tr>
</tbody>
</table>

Modified from Wilson, 1971.

galleries to consume various dead plant material outside of the nest.

The one-piece type is supposed to be the most primitive. An evolutionary trend of life type is a progressive separation between food and nest. The first step is the development of a subterranean gallery system, which allows the colonization of new wood pieces (intermediate type), and the second step is the distinction of the nest from food (separate type). This change of life types seems to have occurred independently at least three times (Mastotermitididae, Hodotermididae, and Rhinotermitidae) in the evolution of termites. The one-piece type is observed in primitive lower termites (Termopsidae) and modern lower termites (Kalotermitidae and *Prokrinotermes* of Rhinotermitidae). The Termopsidae usually nest in large damp wood and are called damp wood termites, while the Kalotermitidae usually nest in dry wood of living and fallen trees and are called dry wood termites. *Prokrinotermes* of the Rhinotermitidae nest in ordinary wood. In the one-piece type, the size, growth, and longevity of the colony are constrained by the size of a wood piece.

The intermediate type is observed in primitive lower termites (Mastotermitididae) and the modern lower termites (most Rhinotermitidae). The separate type is observed in primitive lower termites (Hodotermitidae),
modern lower termites (a few Rhinotermitidae), and the higher termites (Termi- tidae). Separate type nests are classified into three categories: (a) subterranean nests constructed below ground, (b) epigeous nests protruding above soil surface, and (c) arboreal nests built on a tree trunk or a tree branch but always linked with soil by covered galleries.

III. SYMBIOSIS

A. Digestive Tube

Termite symbiosis is an obligate nutritional mutualism between the termite and microorganisms in the gut and nest of termites, although symbiosis, in the general sense, means the living together of two or more organisms that are not closely related in phylogeny, without the implication of beneficial exchanges.

The digestive organ of lower termites is schematically shown in Fig. 8. The digestive tube consists of foregut, midgut tubular (site for secretion of digestive enzymes and for absorption of soluble nutrients), and hindgut (voluminous site for digestion and for absorption of nutrients). The swollen portion of the hindgut is called the paunch. Malpighian tubules, which transport urine and urinary metabolites for excretion, empty at or near the junction of mid- and hindgut. Midgut and salivary glands produce the enzyme for decomposing cellulose into glucose. In the hindgut, cellulose is decomposed into acetic acid by symbiotic flagellates.

The microenvironment becomes more anoxic from foregut to hindgut. The hindgut used to be considered an anaerobic fermentation chamber, but recently this has been found to be false. The hindgut consists of an anoxic lumen surrounded by a microoxic periphery (Fig. 9). This is consistent with the occurrence of both anaerobic and O₂-dependent microbial metabolism in

FIGURE 6 A chimpanzee fishing termites with a stick in a fungus-growing termite mound. From Behnke, 1977.

FIGURE 7 Nests of Nasutitermes triodiae in Australia. See also color insert, Volume 1.
the hindgut. Bacteria on or near the gut wall constitute an oxygen sink, consuming the inwardly diffusing O₂ and thereby creating anoxic conditions favorable for fermentative production of acetate, a major energy source for termites. The hindgut of termites typically has a low redox potential (from −50 to −270 mV) and a pH around neutrality (6.2 to 7.6), although some portions of the hindgut of soil-feeding termites have a pH as high as 11.

B. Two Types of Symbioses

Woody tissue, a major food of termites, contains only 0.03 to 0.1% nitrogen and their C/N (carbon/nitrogen) ratio is 350 to 500, whereas termite tissues contain 5 to 14% nitrogen and their C/N is 4 to 12. Therefore termites have to solve the "carbon-nitrogen balance" problem in addition to "cellulose decomposition problem" (Higashi et al., 1992). Termites use free-living and symbiotic fungi, gut bacteria, and protozoa for assistance in decomposing cellulose and metabolizing and conserving nitrogenous compounds.

Thus, two symbioses are distinguished in termites: "cellulose decomposition symbiosis" and "C-N balance symbiosis" (Fig. 10). All the species of termites are, though to different extents, able to produce their own cellulase. This seems reasonable because they have no way to gain glucose, which is used not only for energy production but also for biological synthesis, other than
digesting cellulose for themselves. The lower termites, which harbor symbiotic flagellates, need cellulose decomposition symbiosis to increase the efficiency of cellulose decomposition.

On the other hand, C-N balance symbiosis is necessary for all termites, because they are not able to solve the problem alone. Two mechanisms are possible for C-N balance: adding N to their food, referred to in Fig. 10 as route N-1, and selectively outputting C, referred to as routes C-1 and C-2. Termites obtain N either through their food or by fixing atmospheric N. Termites tend to feed selectively on food with above average N context. Some species feed on plant tissue only when it is partially decomposed, and by this means, take advantage of the reduced C/N ratio that results from fungal and microbial decomposition. As selectively eliminating carbon is a "wood-consuming" manner, one-piece type termites, which stay in their wood nest and consume it, are expected to develop mechanisms to add N, such as symbiosis with nitrogen-fixing bacteria. On the other hand, separate type termites that are free from the limit in resource utilization are expected to make a full choice among possible mechanisms for C-N balance, including association with methanogenic bacteria. Among the separate type termites, the fungus-growing habit brings additional mechanisms for C-N balance; termites both add N through local N-recycling (route N-2b) and selectively eliminate C through fungal respiration (route C-2).

C. Cellulose Decomposition Symbiosis

The phylogenetic trend of symbiotic microorganisms in termites and cockroaches is shown in Fig. 11. Lower termites harbor flagellates in their hindgut. More than 400 species of flagellates of three orders, Oxymonadida, Trichomonadida, and Hypermastigida, have been described from 205 species of lower termites. Each species of lower termites usually harbors more than one species of flagellates. The class Parabasealea, to which Trichomonadida and Hypermastigida belong, is one of the oldest groups among euarkaryotes, lacking mitochondria and peroxisomes. Members of this class use hydrogenosomes for energy production and, like the prokaryotes, have a 70S ribosome.

All colony members of lower termites except eggs, newly hatched larvae, and individuals just after molting harbor flagellates. They are transmitted through proctodeal feeding (i.e., transmission of gut contents from anus of a donor to mouth of a receptor). The flagellates of the mother colony are transmitted to offspring colonies by alates. The punch of the worker is filled with flagellates and bacteria; they occupy 61% by weight of the hindgut contents of Reticulitermes flavipes (Rhinotermitidae). The population of flagellates in a gut of Reticulitermes speratus often reaches more than 105 individuals.

The association between lower termites and flagel-
lates is well known, since Cleveland (1924) showed that the lower termite died in 3 to 4 if the flagellates were removed, but reinfection by them enabled their host to survive. However, more than 70% of termite species belong to the Termidae, which lack symbiotic flagellates. Lower termites in addition to higher termites are more or less able to produce cellulases by themselves. They secrete endo-β-1,4-glucanases from the salivary gland, while higher termites secrete it from the midgut. The first cellulase gene for the endo-β-1,4-glucanase was sequenced from a Japanese lower termite, *Reticulitermes speratus*, in 1998.

A biochemical overview of the major processes involved in the symbiosis between termites and their gut microorganisms is shown in Fig. 12. The flagellates ingest wood taken by the lower termites and decompose cellulose as follows:

\[ C_{6}H_{12}O_{6} + 2H_{2}O \rightarrow 2\text{CH}_{3}\text{COOH} + 2\text{CO}_{2} + 4\text{H}_{2} \]

The flagellates provide acetate for termite and bacterial metabolism as well as H_{2} and CO_{2} for bacterial metabolism.

Cellulose in the plant cell wall is combined with lignin to form lignocellulose. Except for fungus-growing termites, the lignin degradation in termite guts remains ambiguous. Some termites such as *Mastotermes darwiniensis* and *Coptotermes acinaciformis* are incapable of digesting lignin, whereas *Nasutitermes exitiosus* seems to be able to degrade lignin at least in part.

D. C/N Balance Symbiosis

Lower and higher termites harbor a high density and diversity of bacteria in the gut: 109 to 1011 cells per ml of gut fluid. Most bacteria are distributed in the paunch of the hindgut, but no clear evidence to support a significant role for bacteria in cellulose hydrolysis has been given. Intestinal bacteria are important in nitrogen fixation, nitrogen enrichment of the diet, pyruvate metabolism, or the metabolism or acetogenic reduction of CO_{2}.

Carbon elimination through the activities of microorganisms is the most general mechanism of C/N balance. It entails the release of carbon-rich, nitrogen-poor products of digestion leaving an assimilable fraction of digesta enriched in nitrogen. Gaseous forms of carbon released are principally carbon dioxide and methane, which is formed through respiration, fermentation, and methanogenesis. Almost all termites emit methane and soil-feeding termites tend to emit more methane than wood-feeding termites. Diverse prokaryotes are capable of participating in such transformations.

As long ago as 1925, Cleaveland suggested that ter-
mites use nitrogen fixation to explain their survival on a diet low in nitrogen. Nitrogen fixation, which is energy demanding, is only known from bacteria. Nitrogen fixation in termites was first reported by Benemann (1973) and Breznak et al. (1973). Tayasu et al. (1994) estimated the extent of nitrogen fixation in termites by comparing the natural abundance of $^{15}$N in termite tissues with that of their food and air, showing that 30 to 60% of the nitrogen in the body of Neotermes koshunensis (Kalotermitidae) was derived from $N_2$ fixation. However, the variation in the importance of nitrogen fixation in the termite colony is not clear.

Nitrogen is conserved by internal recycling of nitrogenous waste from termites, by recycling of feces, or by cannibalism and nectophagy. Nitrogenous wastes excreted into the gut through the Malpighian tubules are broken down by gut microorganisms and recycled. Termites have symbiotic, uricolytic bacteria, which use uric acid as an energy source, producing ammonia as the major end product. Ammonia provides a readily available source of nitrogen for protein synthesis in many bacteria.

E. Symbiosis with Fungi

Higher plants and animals have developed various symbioses with fungi. The Macrotermiteae of higher termites cultivate the basidiomycete fungus, *Termitomyces*, on fungus gardens (fungus combs) in the nest and are called fungus-growing termites. *Termitomyces* is known only from macrotermite nests. The fruiting body of *Termitomyces* growing on the comb of *Odontotermes formosanus* at the beginning of rainy season in Okinawa, Japan, is shown in Fig. 13. Ascomycetes of the genus *Xylaria* grow rapidly on the comb, when the comb is removed from the nest.

Fungus-growing habits are known from ants as well (e.g. *Atta*). It is a curious fact, possibly coincidental, that fungus-growing termites, which consume both fungi and fungus combs, are distributed in Africa and southeast Asia, whereas fungus-growing ants, which consume only fungi growing on fungus gardens, are distributed in America.

In *Macrotermes michaelseni* in the grassland of Kenya, old workers (mainly major workers) go out of the nest and collect dead grass pieces. Young workers in the nest consume the pieces and build fungus comb in the mound chambers by depositing faecal pellets of partially digested grass pieces. The comb, which is shaped into round forms resembling brain or sponge, is inoculated with *Termitomyces*, which develop small round white nodules called conidia. Young workers continually deposit the faecal pellets on the top of the comb and consume the conidia, which contain cellulase as well as glucose. Old workers and soldiers consume the old part of the comb from underneath.

The role of *Termitomyces* has been controversial since König and Smeathman discovered the fungus in the 18th century. Suggested roles of the fungus are (a)
degradation of lignin in the fungus comb, (b) ingesting and utilization of fungal cellulytic enzymes in conidia to complement its own digestive enzymes, and (c) improvement of nitrogen economy in termite nutrition.

*Termiomyces* is able to break down lignin in addition to cellulose in the comb to simpler compounds, which are reinjected by old termites together with the fungi. In *Macrotermes muelleri*, a synergistic interaction between the cellulytic enzyme of the termite and fungal origin in the gut is known, while in *M. subhyalinus* and *M. micaelensi*, most (90%) of the glucose requirements of workers can be met by the activity of endogenous intestinal cellulases. The extent to which fungal enzymes contribute to cellulose digestion after ingestion by termites is under debate.

The symbiotic fungi seem to be an important nitrogen source for termites, because nitrogen concentration in the fungi is much higher than that in the fungus comb. Thus, all of the three suggested roles of fungus are partly true. The size and morphology of the combs are different among macrotermite species, and the role of fungus may vary among them. The comb plays a role as reserved food for termites, especially at times when food is scarce out of the nest.

Alates of *Macrotermes bellicosus* and *Microtermes* spp. carry the spores of *Termiomyces* in their rectal chamber, while *Anisotermes cavithorax*, *A. guineensis*, *Macrotermes subhyalinus*, *Odontotermes pauperus*, and *O. sneathmani* do not carry the fungal spores. They may depend on the source of spores from fungi growing on the soil surface, which the first workers collect.

IV. SOCIAL LIFE

A. Caste Differentiation

A termite colony usually consists of a pair of reproductives and their offspring including workers or pseudergates (false workers), soldiers, nymphs, larvae, and eggs. Termites are hemimetabolous, and their developmental pathway consists of several instars with little morphological change until the final molt to alates. The differentiation into castes such as workers, soldiers, and neotenics occurs when immatures deviate from the pathway leading from egg to winged imago (alate). The more evolved the species, the earlier the caste differentiation.

The most important difference in the caste differentiation exists between the one-piece nest type and separate and intermediate nest types. The one-piece type produces pseudergates (false workers), while separate and intermediate types produce workers. The pseudergates, which functionally correspond to workers in separate and intermediate types, retain the capability to be reproductive (alate and neotenic), presoldier, and another pseudergate instar. In this sense, pseudergates are helpers.

The developmental pathways of *Kalotermes flavicollis* of Kalotermitidae (one-piece type) and *Macrotermes michaelseni* of Termitidae (separate type) are shown in Fig. 14. *K. flavicollis* is characterized by flexible or late caste differentiation. On the way to the imago, which develops from the egg through seven molts, two irreversible deviations to sterile soldiers and neotenics and one reversible deviation to pseudergates occur. There is no sexual dimorphism: the sex ratio among pseudergates and soldiers usually depart little from 1.

Caste differentiation is controlled by pheromones. Reproductives (king and queen) secrete sex-specific pheromones that inhibit the metamorphosis of immatures into reproductives, although the inhibitory pheromones have not been chemically identified (Fig. 15). Two inhibitory pheromones are produced by the queen and the king and are passed out of their anuses (proc-todeal feeding). When the reproductives die, pseudergates or nymphs molt to neotenics and inherit their parents' colony. The presence of soldiers inhibits the metamorphosis of pseudergates into soldiers, but its physiological basis is not yet known.

The flexible caste differentiation in *K. flavicollis* is associated with its nesting habits. The Kalotermitidae to which *K. flavicollis* belongs nest in the dead branches of living trees, standing dead trees, and rarely fallen trees, consuming the wood. Those nests are unstable, because wood consumption by nesting termites leads to nest destruction.

In *Macrotermes michaelseni* of the Termitidae, the second instar is an irreversible decision point separating two developmental pathways to imago and sterile workers and soldiers. The number of larval instars is three in the Macrotermitinae and two in the other Termitidae. A sexual dimorphism frequently occurs in workers, while soldiers are of one sex. In *Macrotermes*, major workers are males and minor workers are females, while major and minor soldiers are females. In *Trinervitermes* of the Nasutitermitidae, soldiers are males and workers are females. Replacement reproductives of the Termitidae arise either from alates and nymphs or workers.

B. Evolution of Workers and Soldiers

All species of termites, with some exceptional groups in higher termites, have soldiers. They are adapted to living in the darkness and lack compound eyes. The soldiers of most species are sterile, but fertile soldiers
that are able to be neotenes are present in Zootermopsis, Archotermopsis and Stoloterms of Termopsidae, a family of primitive lower termites.

Soldiers, which are specialized in the defense of nest and foraging workers, are rich in the variety of morphology. Two types can be distinguished; mandibulate and nasute types. Mandibulate soldiers occur in all families and rely on large, prominent mandibles to defend the nest and colony members by biting or snapping with mandibles. In nasute soldiers occurring only in Nasutitermitinae of the Termitidae, the front of the head is modified to chemical warfare and the mandibles are in many cases reduced. Defense by nasute soldiers are very effective to ants, major predators of termites (Fig. 16). Soldiers are of one size in most species, but are dimorphic in Schodarhinoterms of Rhinotermitidae and Termitidae (e.g., Macrotermes and Acanthoterms of the Macrotermiteinae).

The proportion of soldiers to workers (or false workers) in mature colonies varies considerably. Soldiers occupy up to 30% of colony members in the Nasutitermitinae with nasute soldiers, which are usually smaller than workers, while they occupy about 5% in most termites with mandibulate soldiers, which are usually larger than workers. The proportion of soldiers decreases in soil feeders and some of them (mainly Apicotermiteinae of Termitidae) lack the soldier caste.

Workers are numerically the largest caste in the colony and consist of sterile males or females which are wingless, in most cases unpigmented, and lack special external modifications. Compound eyes are usually absent. Workers provide labor for gathering food, feeding dependant castes (larvae, soldiers, nymphs, and reproductives), tending eggs, and repairing and enlarging the nest and gallery system. Worker polyethism is based on age and sex; old workers tend to predominate in
evolved once in one-piece type prototermites with false workers, and disappeared secondarily in some soil-feeding termites. On the other hand, the sterile worker caste has evolved polyphylogenetically. Separate and intermediate type termites (Mastotermitidae, Hodotermitidae, most Rhinotermitidae, and Termiidae) possess a worker caste, whereas one-piece type termites (Termopsidae, Kalotermitidae, and Proxhinotermites of Rhinotermitidae) possess a false worker caste. The sterile worker may have evolved from the false worker at least three times independently in accordance with the change from one-piece nest to intermediate and separate types.

C. Diversity in Social Life

The social organization of termites is similar to that of social Hymenoptera in that the colony is organized based on the division of labor due to caste and age. However, there is an important difference between them. In contrast to Hymenoptera, whose active members are imago (adult, which are suppressed to be reproductive), active members such as workers and soldiers of termites are individuals which are suppressed to be imago.

The degree of eusociality, fecundity of reproductives, and the other life history characteristics are diverse among termite species, and they are summarized in relation to life type in Table IV. The reproductive skew, defined as the probability that workers and soldiers forego direct reproduction, is a good indicator of eusociality (0: all colony members reproduce; 1: colony members forego direct reproduction except a single colony breeder). The skew is low in the one-piece type, where pseudergates have the potential to be reproductive, while it is high in the intermediate or separate type, where workers have little chance to become reproductives.
The reproductive skew is negatively correlated with the extent of neotenic production. The percentage of genera producing neotenics is high in the one-piece type (100% in Termopsidae and Proxhinotermes of the Rhinotermidae) except the Kalotermitidae (48%), medium (100% in Mastotermitidae and 71% in Rhinotermidae) in the intermediate type, and low (67% in Hodotermitidae and 0–20% in Termitidae) in separate types.

A remarkable difference among life types is the fecundity of primary reproductives. It is low in the one-piece type, whereas it is sometimes extremely high in the separate type (e.g., a queen ofMacrotermes subhyalinus lays 36,000 eggs per day).

In comparison with most insects, primary reproductives survive for a long time—10 to 12 years for Incisitermes minor (one-piece type) and 20 to 50 years in Mastotermitinae (separate type). It is noteworthy that the longevity of reproductives is similar to that of pseudergates in the one-piece type, whereas it is much longer than that of workers in the separate type.

The age at which termites begin to reproduce is early (beyond the second year in Kalotermes flavicollis, beyond the fourth year in Incisitermes minor, after the fourth year in Zootermopsis, after the sixth year in Neotermes tectonae) in the one-piece type, and late (after 5 to 10 years in many species of Rhinotermidae and Termitidae) in intermediate and separate types. The size of mature colonies producing alates ranges from a few hundred individuals for Kalotermitidae (one-piece type), many thousands or much more in Mastotermitidae, Rhinotermitidae (intermediate type), and Termitidae (separate type).

The colony size is shown in Table V, although the accuracy of estimation is different. The one-piece type such as the Kalotermitidae and Termopsidae is small, usually up to 10,000, while the separate type such as mound builders of the Termitidae, in particular Macrotermes and Nasutitermes, is very large up to some millions. The intermediate type, including the Rhinotermitidae and Mastotermitidae, is sometimes very large up to some millions.
### Table IV

**Distinguishing Social Characteristics Associated with Each Nest Type**

<table>
<thead>
<tr>
<th>Life type:</th>
<th>One piece type</th>
<th>Intermediate type</th>
<th>Separate type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Termite group</td>
<td>Termitidae, most Kalotermitidae and Prothorntermitidae of Rhinotermitidae</td>
<td>Mastotermitidae, few Kalotermitidae and most Rhinotermitidae and some Termitidae</td>
<td>Hodotermitidae, some Rhinotermitidae and most Termitidae</td>
</tr>
<tr>
<td>Reproductive skew</td>
<td>0–0.5</td>
<td>0.4–0.75</td>
<td>0.7–1.0</td>
</tr>
<tr>
<td>Task specializations</td>
<td>Sterile soldiers, false workers and age polyethism</td>
<td>Sterile soldiers, sterile workers and age polyethism</td>
<td>Sterile soldiers, sterile workers and age polyethism (except in some soil feeders), sterile workers and age polyethism</td>
</tr>
<tr>
<td>Fecundity of primary reproductive (eggs/day)</td>
<td>Small Cryptotermes havilandii: 8</td>
<td>Massotermes darwiniensis lab-reared reproductives up to 17 years, C. formosanus at least 9 years, R. lubricus primaries and neotenics up to 7 years in the lab</td>
<td>Large Odontotermes obesus: 26,000–86,000 Macrotermes subhyalinus: 36,000 Nasutitermes surinamensis: 3,900 Cubitermes severus: 50–600</td>
</tr>
<tr>
<td>Age of primary reproductives</td>
<td>Field-observed reproductive 4–5 years in Zootermopsis spp. and 10–12 years in Incisitermes minor</td>
<td>Massotermes darwiniensis lab-reared reproductives up to 17 years, C. formosanus at least 9 years, R. lubricus primaries and neotenics up to 7 years in the lab</td>
<td>Primary reproductives 20 years; queens 20–25 years in Macrotermitidae spp. (estimates based on size of physonogastric queen)</td>
</tr>
<tr>
<td>Age of workers and soldiers</td>
<td>Zootermopsis false workers up to 4 years, soldiers 4–5 years in the lab; Neotermes indicus soldiers and false workers more than 6 years in the lab</td>
<td>Reticulitermes lucifugus soldiers up to 5 years in the lab; R. hesperus workers 3–5 years in the field; Cryptotermes acinaciformis workers and soldiers 4 years in the lab</td>
<td>Macrotermes spp. workers and soldiers less than a year in the field and lab; Macrotermes natalensis up to 1.5 year in the lab; Cubitermes ugaformis worker 196–339 days</td>
</tr>
<tr>
<td>Colony foundation</td>
<td>A male and female devalue pair</td>
<td>A male and female devalue pair by a cohort of neotenics and workers</td>
<td>A male and female pair (sometimes multiple reproductives) or sociotommy or secondary devalues reproductive, brachypterous neotenics with workers, rarely ergate reproductive</td>
</tr>
<tr>
<td>Dispersal distance of alates</td>
<td>Short Cryptotermes: 1–45m Kalotermes: 20–50m Incisitermes: 120m Zootermopsis: 300m</td>
<td>Short Reticulitermes: 10–200m</td>
<td>Long Odontotermes: over 0.8km Cubitermes: a few km</td>
</tr>
</tbody>
</table>

*After Shellman-Reeve, 1997.*

## D. Organization of Foraging

Foraging of termites has evolved in concert with the change of life type. The one-piece type of termites do not forage out of nest, and the nest site selection by alate reproductives plays a role of food research. Deicate reproductives of *Zootermopsis* (Termopsidae) look for a colony founding site in the wood cambium layer containing nitrogen of relatively high concentration.

In intermediate and separate type nesters, a group of workers and soldiers go out of the nest and search for new food sources. Many species show aggressive behavior toward the other colony members of the same or different species, and the foraging territory, which is intra- and interspecifically defended, is known from many termite species.

The division of labor in foraging activities is based on caste and age. Soldiers usually defend foraging workers, and in some species they explore new areas for food and recruit workers to new food sites. Workers mainly collect and process food, and they also defend foraging territory.

Foraging activities are mediated by trail pheromones emitted from the sternal gland. Several sub-
<table>
<thead>
<tr>
<th>Behavioral-physiological</th>
<th>Primitive condition</th>
<th>Derivative condition</th>
<th>Characters showing regression</th>
<th>Images (primary reproductive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small numbers in the nature colony</td>
<td>Large numbers in the nature colony</td>
<td>Y-suture of head present</td>
<td>Y-suture of head reduced or absent</td>
<td></td>
</tr>
<tr>
<td>Excavated nests with little construction</td>
<td>Elaborately constructed carton nests</td>
<td>Two well-developed ocelli present</td>
<td>Ocelli reduced or absent</td>
<td></td>
</tr>
<tr>
<td>Relatively little care of eggs, nymphs, and adults</td>
<td>Relatively great care of eggs, nymphs, and adults</td>
<td>Antennae with numerous segments</td>
<td>Antennae with fewer segments</td>
<td></td>
</tr>
<tr>
<td>Food not stored</td>
<td>Food stored</td>
<td>Mandibles with 2 or 3 prominent marginal teeth with sharp basal notches</td>
<td>Mandibles with reduced teeth and notches</td>
<td></td>
</tr>
<tr>
<td>Nutritive dependence upon symbiotic intestinal flagellates</td>
<td>Nutritive independence from symbiotic intestinal flagellates</td>
<td>Five tarsal joints present</td>
<td>Second tarsal joint reduced or absent</td>
<td></td>
</tr>
<tr>
<td>Damp-wood eaters</td>
<td>Dry-wood eaters</td>
<td>Arolium present between tarsal claws</td>
<td>Arolium absent</td>
<td></td>
</tr>
<tr>
<td>Wood eaters</td>
<td>Deal leaf and grass eaters</td>
<td>Genitalia with numerous segments (up to 8)</td>
<td>Cerci reduced, usually with 2 segments</td>
<td></td>
</tr>
<tr>
<td>Fungus gardens</td>
<td>Fungus gardens</td>
<td>Styli present</td>
<td>Styli absent</td>
<td></td>
</tr>
<tr>
<td>Food foragers only within excavated tunnels in wood</td>
<td>Termitephiles, often species-specific and highly modified</td>
<td>Egg mass in cluster</td>
<td>Eggs laid separately</td>
<td></td>
</tr>
<tr>
<td>Reproductive cases</td>
<td>Workers forage outside nest, sometimes in covered tunnels and sometimes on exposed odor trails</td>
<td>Genitalia similar and clearly homologous to those of cockroaches</td>
<td>Genitalia reduced or absent</td>
<td></td>
</tr>
<tr>
<td>Small abdomen of queen with small ovaries and glandular tissue</td>
<td>Large abdomen of queen with large ovaries and glandular tissue</td>
<td>Hind wing with anal lobe</td>
<td>Hind wing without anal lobe</td>
<td></td>
</tr>
<tr>
<td>Capacity to produce substitute (neotenic) kings and queens</td>
<td>Reproduction confined to primary reproductive caste (image)</td>
<td>Pronotum wide and flat</td>
<td>Pronotum narrow and saddle-shaped</td>
<td></td>
</tr>
<tr>
<td>Frontal gland absent</td>
<td>Frontal gland present</td>
<td>Compound eyes faceted and pigmented</td>
<td>Compound eyes reduced or absent</td>
<td></td>
</tr>
<tr>
<td>Front coxae smooth</td>
<td>Front coxae with ridge or projection</td>
<td>Mandibles as in primitive images</td>
<td>Mandibles as in derivative images</td>
<td></td>
</tr>
<tr>
<td>False worker caste present</td>
<td>Workers</td>
<td>Soldiers with many segments (up to 29)</td>
<td>Soldiers with fewer segments (As few as 10)</td>
<td></td>
</tr>
<tr>
<td>Head elongate, large, smooth, somewhat flattened, with large, curved, and toothed biting mandibles</td>
<td>Head round, or extremely flat, or phragmotic; surface rough or with ridges and projections, mandibles elongate and thin, or twisted for snapping, or with projection for frontal gland</td>
<td>Compound eyes pigmented and faceted</td>
<td>Compound eyes unpigmented, or nonfaceted or even absent</td>
<td></td>
</tr>
<tr>
<td>Mandibles with smooth cutting edges or with reduced teeth</td>
<td>Mandibles with serrated cutting edges</td>
<td>Ocellus spot present</td>
<td>Ocellus spot absent</td>
<td></td>
</tr>
<tr>
<td>Pronotum wide, flatly convex, and with smooth edges</td>
<td>Pronotum narrow, saddle-shaped and sometimes with serrated edges</td>
<td>Mandibles with 2 or 3 large marginal teeth</td>
<td>Mandibles with reduced marginal teeth</td>
<td></td>
</tr>
<tr>
<td>Front coxa smooth</td>
<td>Front coxa ridge or projection</td>
<td>Soldiers present</td>
<td>Soldiers absent</td>
<td></td>
</tr>
</tbody>
</table>

These changes have occurred at various places in termite evolution. Modified from Wilson, 1971.
stances are isolated from the gland: n-hexanoic acid in Termopsidae, dodecanol in Rhinotermitidae, and neocembrene in Nasutitermitinae of the higher termites. Many chemicals that are not endogenous to termites cause an orientation response. 2-phenoxethanol from the ink of certain ball-point pens have a trail-following effect on Coptotermes formosanus.

Territoriality and elaborate division of labor in foraging were studied well in the Macrotermesinae and the Nasutitermitinae of higher termites, because some species of them forage on the ground. A fungus grower, Macrotermes michaelensi in Kenyan savanna, constructs a large mound and has a very large foraging territory based on subterranean galleries that extend up to 50 m from the mound, run horizontally, and lead to foraging holes on the ground surface. The species possesses four sterile castes, major and minor soldiers, and major and minor workers. Major soldiers mainly defend the mound, while minor soldiers mainly defend foraging workers. Major workers are foraging specialists, while minor workers mainly perform intranidal works. The division of labor between major and minor workers is more or less different among species of Macrotermes.

During the night, major workers emerge from the foraging hole, cut off dead grasses, fallen twigs and cow dung on the ground, and bring them into the foraging hole. They form foraging columns (up to 1 m) between the foraging hole and the foraging site, both sides of which are guarded by minor soldiers and rarely major soldiers. The column is attacked by various species of ants. The foraging areas of Macrotermes carbonarius, a mound builder in southeast Asia, change irregularly but seem to be arranged to avoid redundant foraging in the long term (Fig. 18).

Black marching termites of Hospitalitermes of the Nasutitermitinae in southeast Asia form soldier-led foraging columns reaching up to 100 m in length from the nest at the base of trees to feeding sites on standing tree trunks where workers collect lichens. Soldiers explore the feeding site and recruit workers to the feeding site, defending the workers. Three worker castes show division of labor during food collection; minor workers gnaw at lichens and form a food ball that is passed on medium and major workers, which carry the food ball to the nest. Medium workers also form food balls.

Two arboreal species of Nasutitermitinae, Nasutitermes nigriceps, and N. nigriceps form a clear foraging territory in a Panamanian mangrove forest. They build arboreal nests and forage along carton-converted tunnels on branches, which radiate from the nest, and have a nonoverlapping foraging area that is intra- and interspecifically defended.

![Figure 18](https://example.com/figure18)

**FIGURE 18** Daily change of foraging areas and pavement trackways of a fungus-growing termites, M. carbonarius in Thailand. From Sugio, 1995.

### E. Life History of Fungus-Growing Termites

A new colony is founded by dealates or budding of the parent colony. In most cases, a pair of dealates found a new colony to be primary reproductives. Multiple alate-derived queens are restricted to the Termitidae (ca. 40 species are known). Colony budding is achieved by active migration of swarms containing all castes or by division of a diffusely organized colony. The former, called a “sociotomy,” is known only in the Termitidae (e.g., Trinervitermes bettonianus, T. gratiosus, Syntermes territus, and Anoploterme spp.). The latter is known in Mastotermitidae, Kalotermitidae, Termopsidae, Rhinotermitidae, and Termidae, where the extension of nests is coupled of production of supplementary reproductives.

The life cycle of Macrotermes michaelensi, a socially
advanced fungus-growing species of the Termitidae in the grassland of Kenya, is shown in Fig. 19. After the nuptial flight at the beginning of rainy season, a male (king) and female (queen) pair of dealates form a tandem, dig a small cell in the soil, copulate, and initiate a subterranean incipient colony. Occasionally three or more dealates found a new colony together. The primary queen begins to lay eggs within a few days after mating. They hatch after some weeks.

The first brood is tended by the king and queen and is differentiated into soldiers as well as workers. The first workers begin foraging at about 10th week after colony founding in the fungus grower Microtermes. The king and queen are similar in body size at colony founding, but the abdomen of the queen becomes much larger than that of the king with time.

Within a few years, M. michaelensi constructs a mound that reaches 3 m in height at the mature stage. A large mound builder has a territory based on subterranean galleries, which run horizontally in a shallow layer (10–15 cm). Young subterranean colonies can survive within the territories of large mound builders, because they are located at a deeper layer (20–40 cm) than the subterranean galleries of large mound builders and are able to avoid their attack. However, with colony growth they construct foraging galleries in the shallow layer and are killed by large mound builders. Thus in Macrotermes michaelensi, large mounds tend to be uniformly distributed due to territoriality, and small (young) mounds are rare in Kenya.

On the other hand, both large and small mounds of M. bellicosus are distributed randomly in Nigerian savanna. The mortality rate of large mound builders of M. bellicosus is very high due to predation by doryline ants and aardvarks. Thus, young subterranean colonies are not seriously affected by the proximity of mound builders and survive to construct small mounds, which are randomly distributed.

A marked increase in the number of small mounds following the death of large mound-building colonies due to the severe drought has been reported M. bellicosus in Senegal and M. subhyalinus in Kenya. Thus, when the mortality rate of large mounds is low, they show uniform distribution due to their territoriality, whereas when their mortality rate is high, young mounds are continuously added and mounds as a whole tend to show random distribution. The duration of mound occupation has been poorly understood: 80 years for Macrotermes in Africa, more than 100 years for a large mound of Nasutitermes triodiae in Australia, and 20 to 40 years for Amitermes virosus in northern Australia.
V. THE GLOBAL DIVERSIFICATION OF TERMITES

A. Major Events in Termite History

Three major events of global diversification that Isoptera has experienced are identified (Higashi and Abe, 1997): (a) the evolutionary radiation of original termites, (b) the expansion and diversification of separate type termites and the diversification of one-piece type termites driven by the expanding separate type termites into fragmented habitats, and (c) the radiation of higher termites.

As stated earlier, Cretaceous fossil species belonging to Hodotermitidae, Termopsidae, and Mastotermitidae are known from Europe, Asia, and North and South America. This shows that primitive lower termites had been globally distributed and diversified in the Cretaceous. Therefore, the first event, the evolutionary radiation of original termites, must have occurred at latest in the Jurassic and Cretaceous periods of the Mesozoic. All living species of Hodotermitidae are specialized in grass feeding, while living species of Termopsidae and Mastotermitidae are wood feeders. Grass feeding habits of the Hodotermitidae must have evolved in the Tertiary period, and Cretaceous Hodotermitidae must have consumed wood, because grasses evolved in the Tertiary.

The fact that the oldest fossil record is the worker of Hodotermitidae is interesting, because this suggests that the evolution of the worker caste occurred in the Mesozoic. We do not know if Cretaceous termites of Hodotermitidae nested in wood or soil.

Many fossil species of the Kalotermitidae whose living species are mostly of the one-piece type have been found in the Tertiary period. It is probable that the second major event, the expansion and diversification of separate type termites and the diversification of the one-piece type termites driven by the expanding separate type termites into fragmented habitats, occurred in the Tertiary.

Only a few fossil species of the higher termites are known from the Tertiary. The third event, the radiation of higher termites, may have occurred in the late Tertiary or the Quarternary. On the other hand, the worldwide distribution of higher termites suggests their early evolution and dispersal before the breakdown of Gondwanaland in the Cretaceous. As shown later, it is quite difficult for termites to cross the oceans. It is probable that the radiation of higher termites may have occurred in the late Tertiary or Quarternary, although the origin of higher termites may be before the breakdown of Gondwanaland in the Cretaceous.

B. Dispersal Ability of Termites

Islands sterilized by volcanic eruption are good for studying the process of colonization. In 1883, Krakatau, located about 60 km from Java and 40 km from Sumatra, erupted violently and more than half of the island disappeared, leaving the remaining Krakatau and two neighboring islands. The entire flora and fauna of the islands were almost destroyed, but they were rapidly colonized by organisms of Java and Sumatra.

In the Krakatau Islands, three, three, and seven species of termite were collected in 1908, 1919–1933, and 1982, respectively. They were characterized by the dominance of the lower termites (Kalotermitidae and Rhinotermitidae), the one-piece type, and the absence of soil feeders, although a wood-feeding arboreal species of Nasutitermes of the higher termites was found in the Krakatau Islands. On the other hand, the supposed source area (the tip of West Java) was characterized by higher species diversity (13 species), the dominance of higher termites and separate type, and the presence of soil feeders. In the tip of West Java, the one-piece type termites of the Kalotermitidae are confined to the coastal forest, while soil feeders are confined to inner forests.

Alates of termites are specialized for dispersal within a zone of calm air near the ground. The dispersal distances are short in lower termites (up to 300 m) and a little long in higher termites (a few kilometers). It is not impossible but difficult for termites to disperse widely or to reach remote islands by flight. Therefore, termites usually cross the sea by rafting. The Kalotermitidae, the first colonizers, have some preadaptive attributes for oversea dispersal: (a) they tend to be confined to the coastal area and may be easily swept into the sea, (b) they are more tolerant to sea water than other types, and (c) their caste differentiation is flexible.

On the other hand, it is difficult for soil-feeding termites to cross the sea because they have little opportunity to be in rafting wood. Thus the dominance of the Kalotermitidae (one-piece type) and the Rhinotermitidae (intermediate type) in many islands is explained in terms of the differential dispersal ability of termites.

C. The Evolutionary Radiation of Original Termites

The first event, the evolutionary radiation of the original termites into a new habitat of wood throughout the world, may have been driven by an efficient utilization of wood as an abundant and stably supplied food resource and nest substrate, which became possible by
the evolution of two symbioses for termites (i.e., the cellulose digestion symbiosis with cellulolytic flagellates and the C-N balance symbiosis with nitrogen fixing bacteria). All descendants of the Mesozoic families of Hodotermitidae, Termitopsidae, and Mastotermitidae harbor symbiotic flagellates and nitrogen-fixing bacteria in the gut.

Once a termite obtains a means for digestion and C-N balance through association with microorganisms, then wood, which is superabundant but extremely hard to digest and C-N unbalanced, becomes a "well-protected" food resource that can be monopolized by the termite. The diversification of termites into the Hodotermitidae, Termitopsidae, and Mastotermitidae together with their worldwide distribution in the Cretaceous may be viewed as an example of niche opening, which often leads to an adaptive radiation.

Conquering the wood habitat should promote the evolution of a false worker caste, because life in a piece of wood causes (a) greater need for concentrating nitrogen, (b) higher possibility of nest success due to the longer duration time of the nest relative to the longevity of reproductives, and (c) lower possibility of finding an alternative nesting place due to the heterogeneity of resource (wood) distribution. The evolution of this subsociality should further have enhanced the opportunity for eusociality—that is, the production of sterile soldier caste.

D. Separate Type Expansion and One-Piece Type Diversification

The second major event came with the development of separate type termites, which separate their feeding sites from their nest. They may have expanded, replacing the existing one-piece type termites through a competitive exclusion process, the detail of which was examined by Higashi and Abe (1997). This is a consequence of asymmetrical competition between one-piece and separate types due to the fact that a piece of wood is food and nest for the one-piece type but only food for separate type termites. On the other hand, some groups of one-piece type termites, which were driven out by the expanding separate type termites into fragmented habitats, should have gone through a diversification process.

This process could be reconstructed based on the present peculiar distribution patterns of one-piece type termites and separate type termites. The three groups of one-piece termites show quite different marginal distribution patterns. Damp wood termites (all species of Termitopsidae) show a clear amphitropical distribution (Fig. 20). Ordinary wood termites of Probiotermes are widely distributed in tropical and subtropical regions but tend to be confined to islands. Dry wood termites (Kalotermitidae) are also widely distributed in the tropical and subtropical regions but tend to be confined to dry dead parts of standing trees.

On the other hand, separate type termites, represented by the Termitidae, are widely distributed in tropical mainlands, the central areas of termite distribution. Intermediate type termites, represented by Rhinotermitidae, are widely distributed from tropical to temperate regions. Thus the distribution of the one-piece type and separate type is complimentary, while that of intermediate type covers the ranges of both types.

Although a full explanation of such distribution patterns requires one to take into consideration other factors such as plate tectonics and global change of climate, ecological mechanisms, specifically interspecific interactions and differential dispersal ability, have been a major force in determining the ranges of distribution.

The asymmetrical competition theory predicts that as the latitude (ocean) gets higher (closer), the resource supply rate of the one-piece type increases relative to that of the separate type, because more wood is available to the one-piece type due to the slower decay rate of damp (dry) wood at higher latitude (seacoast) areas under a given predation pressure. Thus, the one-piece type may compensate its disadvantage from its asymmetrical competition with the separate type and becomes dominant at higher altitude (seacoast) areas. This explains the questioned marginal distribution patterns of damp and dry wood termites. The distribution of ordinary wood termites confined to islands may simply be explained as a result of escape from asymmetric competition due to their dispersal ability.

Separation of feeding site from nest (reproductive site) should enhance the C-N balance symbiosis by adding the option of selective carbon elimination and promoting the evolution of a true worker caste. These should further have enhanced the superiority of the separate type in resource utilization, thus driving the expansion and diversification of separate type termites.

Among the one-piece type termites driven out by the expanding separate type termites, those groups driven into fragmented habitats along seacoasts and islands are expected to have gone through a diversification process. In fact, dry wood termites show a high species diversity as a family.

E. The Radiation of Higher Termites

The third major event, the radiation of higher termites that are free from flagellates, may have been driven by the advantage that a termite would have if it could digest cellulose efficiently enough and eliminate celu-
lolic flagellates, which became possible by eliminating the lignin cover, the major obstacle in cellulose utilization, perhaps through the shift of their diet from wood to a food with less lignin cover.

Those termites that became higher termites must have obtained their own cellulase activity. The present-day termites have their own cellulase activity. What prevents the elimination of cellulolitic flagellates from lower termites may be the inefficiency of the termites in cellulose digestion, which should cause the termites a shortage of (not glucose but) energy supply. Thus, cellulolitic flagellates are still symbionts to lower termites, though they are not for cellulose digestion but for energy supply. Higher termites are those that succeeded in increasing the efficiency of cellulose digestion to the extent that they no longer need help from cellulolitic flagellates for energy supply either. This may be the driving force for their radiation.

For increasing the efficiency of cellulose digestion, the elimination of the lignin cover is expected to be the most effective. An obvious means to achieve the elimination of the lignin cover is to shift the diet from wood to a food with less lignin cover. The nest-food separation makes possible the selection of food. Thus, it is expected that higher termites may appear only from separate type termites, which is indeed the case.

Three major groups of higher termites are recognized: nasute termites (Nasutitermitinae), soil (humus)-feeding termites, and fungus growing termites, although there is some overlap among them. They seem to have succeeded in the elimination of the lignin cover.

The Nasutitermitinae show the broadest distribution and the highest species diversity among the higher termites. Their radiation was driven by the evolution of a new type of soldier with a chemical weapon against ants, their major predator, which made it possible to forage out more freely and select food with less lignin cover. The radiation of soil feeders was driven by the shift of diet from wood to humus, which contains less lignin. The radiation of fungus-growing termites was driven by broadening the diet and a new symbiosis with fungi for C-N balance and for lignin reduction.

VI. TERMITES IN ECOSYSTEMS

A. Terrestrial Ecosystems

Terrestrial ecosystems have some characteristics different from aquatic ecosystems—for example, the dominance of the detritus chain among grazing, detritus,
and microbial chains; the rare cascading effect of predators on vegetation; the abundance of fungi; and so on. This is partly because vascular plants in terrestrial ecosystems produce lignin, an organic matter extremely resistant to chemical degradation. Plants cover cellulose with lignin to form lignocellulose as a major component of the cell wall. Cellulose is an energy source for many heterotrophs, but its availability decreases markedly when covered with lignin.

Lignin makes it possible for terrestrial plants to support leaves horizontally and transport water from root to leaves and branches, providing various heterotrophs with heterogeneous habitats. Furthermore, lignin regulates the decomposition process of dead plant material, and its remnants play an important role on the formation of humus associated with soil fertility.

Therefore, lignin is a key substance for understanding the architecture and the material flow in terrestrial ecosystems. Considering that microorganisms, especially fungi, are major decomposers of lignocellulose, fungi are reasonably abundant in terrestrial ecosystems, whereas they are poor in aquatic ecosystems that lack lignocellulose.

Brown rot among fungi mainly decomposes cellulose, whereas white rot decomposes both lignin and cellulose. Most soil animals are not able to decompose lignin chemically but decompose it physically and promote the chance of contact between lignocellulose and microorganisms. Termites are distinct among soil animals in that some of them decompose lignocellulose efficiently with the aid of microorganisms in the gut or in the nests.

Many termites live in lowland tropical forests, where 50 to 80 species are found in 1 ha area in Fig. 21. Their food is rich in variety including dead branches of living trees, fallen logs, fallen leaves, lichens, and soil. Those termites are functionally largely divided into two groups: soil feeders and litter feeders. Litter feeders are furthermore divided into lignocellulose feeders and cellulose feeders.

Soil feeders, confined to three subfamilies of higher termites, are concerned with the formation of humus.

associated with soil fertility. Lignocellulose feeders (fungus growers of Macrotermitinae) decompose both lignin and cellulose rather completely with the aid of the white rot, Termitomyces. As they do not leave many lignin remnants, they may functionally well be called giant moving white rot. Cellulose feeders, including most of the other termites, decompose cellulose in dead plant material, leaving lignin remnants. They may functionally well be called giant moving brown rot. Roughly saying, soil feeders, wood feeders, and lignocellulose feeders occupy about 45%, 45%, and 10% of living termite species, respectively.

We do not know what substrates soil feeders really eat. Soil feeders can generally be distinguished from litter feeders by intestinal morphology, the stable isotope ratios of carbon and nitrogen (Fig. 22), and by the higher activity of certain gut bacteria, notably methanogens and organisms able to ferment reduced and recalcitrant substrates, including aromatics. An interesting idea that soil feeders eat microorganisms such as bacteria and fungi in ingested soil, and therefore are important members of microbial chain in the tropical forests, is partly supported by the stable isotope analysis and the detection of lysozyme decomposing cell wall of bacteria, but this theory has not been proved.

B. Functional Role of Termites

Taxonomic composition and feeding habits of termite assemblages in some African habitats are shown in Table VI. The species diversity is the highest in the tropical forest and decreases from forests to arid ecosystems. The dominant feeding habits gradually change from soil-feeding to wood- and litter feeding (cellulose and lignocellulose feeding). The species diversity of fungus growers (Macrotermiteae) is the highest in slightly dry areas such as savannas and dry forests where soil fungi are poorer than in moist forests. A similar tendency is recognized in tropical forests of southeast Asia.

The change may be explained in relation to the activities of fungi. Soil feeders are abundant and diversified in tropical rain forests where fungi, in particular white rot, may be active and decompose lignocellulose efficiently, decreasing the litter accumulation on the ground. On the other hand, lignocellulose feeders cultivating white rot of Termitomyces in the nest chambers with high humidity and temperature are dominant in slightly dry tropical areas where the low humidity may suppress the decomposition of lignocellulose by white rot. Cellulose feeders are widely distributed in areas where fungi may be inactive due to low temperature and inadequate humidity.

Abundance of termites sometimes exceeds 10000/m² and 10 g wet weight/m², although it is usually up to 4500/m² and 10 g wet weight/m². The biomass of 10 g wet height/m² is equivalent to 250 persons/km² and 40 kg/person in human beings. In the tropical rain forest of Malaysia where fungus growers are dominant, as much as 30% of the leaf litter supply is consumed by termites. In a Guinea savanna in Nigeria where fungus growers are dominant, termites consume up to 55% of surface litter. Termites are responsible for up to 20% of total carbon mineralization in the savannas of Africa. Consumption of dead plant material by termites increases when fungus growers are dominant, because they consume five to six times more food per unit of biomass than other termites.

Termites are consumed by a great variety of animals ranging from ants to human beings, forming a large detritus chain. Furthermore, a large number of alates produced by mature termite colonies are eaten by birds.
<table>
<thead>
<tr>
<th>Locality</th>
<th>Type of vegetation</th>
<th>Latitude</th>
<th>Annual rainfall (mm)</th>
<th>Grass and litter</th>
<th>Fresh wood and leaves litter</th>
<th>Decomposing litter</th>
<th>Humus and soil</th>
<th>Kalotermiteidae</th>
<th>Rhoenotermiteidae</th>
<th>Termitinae</th>
<th>Apicotermitidae</th>
<th>Macrotermiteidae</th>
<th>Nhamibermitidae</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fout Ouh, Senegal</td>
<td>Wooded steppe (small sand dunes)</td>
<td>16°N</td>
<td>375</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>Tsavo east, Kenya</td>
<td>Bush and wooded steppe</td>
<td>3.5°S</td>
<td>900</td>
<td>8</td>
<td>17</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>Cap Vert, Senegal</td>
<td>Woody sub-Guinean savanna</td>
<td>15°N</td>
<td>550</td>
<td>6</td>
<td>17</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>12</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>Minas, Nigeria</td>
<td>Primary and secondary woodland</td>
<td>9°N</td>
<td>1175</td>
<td>10</td>
<td>17</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>3</td>
<td>11</td>
<td>7</td>
<td>31</td>
</tr>
<tr>
<td>Lomtso, Ivory Coast</td>
<td>Grassy to woody derived savanna</td>
<td>6°N</td>
<td>1200</td>
<td>10</td>
<td>18</td>
<td>3</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>4</td>
<td>9</td>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>Youthoussi, Ivory Coast</td>
<td>Grass savanna</td>
<td>9°N</td>
<td>2000</td>
<td>10</td>
<td>(0)</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>Edna, Cameroon</td>
<td>Equatorial rain forest</td>
<td>3.5°N</td>
<td>&gt;3000</td>
<td>0</td>
<td>8</td>
<td>4</td>
<td>31</td>
<td>0</td>
<td>1</td>
<td>25</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>43</td>
</tr>
</tbody>
</table>

From Josens, 1985.
and spiders in the canopy, joining the grazing chain of terrestrial ecosystems.

C. Termites and Soil

While the distribution and abundance of termites are influenced by climatic conditions (air temperature and humidity), soil properties, and vegetation, termites modify local climate conditions, soil properties, and vegetation through the activities of building mounds and excavating subterranean galleries. The effect of termites on soils has been well examined in the textbook *Termites and Soils* by Lee and Wood (1971).

Termites can modify the soil profile by removing soil from various depths (up to 70 m) and transporting it to the ground surface in the form of runways, sheetings, or mounds, from where it is redistributed by water and wind erosion. In central Africa, termites of *Macrotermes* form huge mounds up to 10 m in height and 30 m in diameter. Excavation of subterranean galleries has positive effects on the hydraulic conductivity and infiltration rates.

Large mounds built by species of Macrotermitinae in Africa and Asia often support a vegetation comprising trees and shrubs that is quite distinct from that of the surrounding soils. This difference is considered due to protection from fire, improved drainage, greater soil depth, higher soil moisture, and improved soil fertility. Cation change capacities (CEC), exchangeable cations (Ca, Mg, Na, K), and base saturation decrease with increasing distance from the mound of *Macrotermes michaelseni* in the grassland of Kenya and available phosphorous and NO₃ nitrogen are high near the mound. This causes the change of vegetation; maximum standing stock of plants occurs near the mound base and the number of plant species increases with increasing distance from a mound. Figure 23 shows a schematic diagram depicting the influence of termite mounds on plant growth.

The termite mound supports a higher level of bacterial activity than adjacent top soil, especially of cellulose decomposers, and the activities of these microorganisms result in the release of nutrients into the mound soil. Many species of higher termites have developed to store the fragments of grasses and leaves in the mounds mainly in somewhat dry regions. The stored fragments are heavily colonized by microorganisms. The storage chambers may have the specific function of decomposition chambers and the process may be similar to the fungus gardens of the Macrotermite in that it provides the colony with organic material enriched in nitrogen.
D. Termites and Human Disturbances

The abundance and species diversity of termites tend to decrease when tropical forests and savannas are cleared and modified (Table VII). At first soil feeders are replaced by fungus growers and then by wood feeders in Africa and Asia. However, the destruction of natural ecosystems and their replacement by other ecosystems in some cases provides suitable habitats for some groups of termites, leading to an increase in their population. Zimmerman et al. (1982) were the first to draw attention to the topic in relation to the production of greenhouse gas by termites.

The global concentration of atmospheric methane, a greenhouse gas, is increasing in recent years at an annual rate of about 1%. This has been mainly attributed to paddy rice cultivation, enteric fermentation (mainly by ruminants), biomass burning, and fossil fuel consumption. Termites have high biomass and emit methane in many tropical ecosystems.

Much research has been done on the global estimates of termites, methane production by various groups of...
termites, oxidation of methane produced by termites in the soil, and the importance of methane production by termites on global climatic change. The rate of emission of methane by termites differs between species and feeding habits. Soil feeders tend to produce more methane than wood feeders on a biomass-specific basis. An important fact is that a large amount of methane produced by subterranean termites is oxidized by microorganisms in the soil. Thus the annual contribution by termites is estimated to be less than 20 Tg and probably less than 10 Tg (ca. 4% and 2% of global total from all sources, respectively).

Termites usually consume dead plant material and do not attack living plants in natural forests. However, some termites that are able to survive in agricultural ecosystems become harmful to crops (Harris, 1971). This tendency is stronger when the crops are exotic. A few hundred species of termites cause damage and only about 50 species are serious pests (Pearce, 1997).

See Also the Following Articles
AMAZON ECOSYSTEMS • INSECTS, OVERVIEW

Bibliography


KEYSTONE SPECIES

Bruce A. Menge and Tess L. Freidenburg
Oregon State University

I. Historical Perspective
II. Critique and Reevaluation
III. Identification of Keystone Species
IV. Conservation Implications

GLOSSARY

**dominant species** Species that owe their influence to their high abundance. Such organisms account for most of the biomass in a community, and thus are the primary components of community structure. Trees in forests, mussels in rocky intertidal habitats, grasses in grasslands, and kelps or corals in near-shore subtidal habitats are all dominant species.

**interaction webs (functional webs)** Subset of species that through their interactions and responses to abiotic factors make up the dynamic core of food webs or communities. These webs include keystone species, dominants, and other strong interactors.

**key-industry species** Prey of intermediate trophic status that support a large group of consumers.

**keystone species** Consumers that have a large effect, and one that is disproportionately large relative to their abundance, on communities and ecosystems. Uniquely, the strong effects of keystone species on their interacting species exert extensive influence, often indirectly, on the structure and dynamics of communities and ecosystems. They are a distinct subset of a more broadly defined set of “strong interactors” that also include species having strong effects on interacting populations but not necessarily on communities or ecosystems. Keystone species can include predators, parasites, pathogens, herbivores, pollinators, and mutualists of higher trophic status, but generally are not plants, sessile animals, or “resources.”

**strong interactors (foundation species)** Species that have a large effect on the species (one or a few) with which they interact. Communities and ecosystems may have many strong interactors, and such species may occur at all trophic levels. “Strong interactors” is a more general term that can include keystone species, but not all strong interactors are keystone species. A similar idea is “foundation species;” defined as the group of critical species whose effects and interactions define much of the structure of a community.

**weak interactors** Species that have little effect on other species, at least under average conditions. Under some circumstances, weak interactors may occupy important roles in ecological communities as a result of changes that lead to temporary increases in their abundance, size, or biomass.

KEYSTONE SPECIES WERE originally defined as species high in the food web that greatly modify the composition and physical appearance of an ecological community. The original keystone species was a carnivorous predator that greatly modified the composition and physical appearance of a marine community.

Encyclopedia of Biodiversity, Volume 3
Copyright © 2001 by Academic Press. All rights of reproduction in any form reserved.

613
A sea star occurring in a marine rocky intertidal habitat dominated by macroalgae, mussels, barnacles, and other invertebrates (Paine, 1966). Implicit in the initial definition was the idea that keystone species have effects that are disproportional to their abundance. This concept was formalized by Robert Paine in the 1960s: “the patterns of species occurrence, distribution and density are disproportionately affected by the activities of a single species of high trophic status” (see Power et al., 1996). With time, the number of examples of apparent keystone species grew, leading to heightened awareness that the phenomenon was more general, and therefore more important, than originally thought. This increasing familiarity with, and embracing of, the concept of keystone species led workers to broaden its definition and, perhaps inevitably, to extend it to questionable applications (Mills et al., 1993). Thus, in addition to keystone predators, workers defined “keystone herbivores,” “keystone prey,” “keystone mutualists,” “keystone hosts,” “keystone resources,” “keystone guilds,” and “keystone modifiers.” One idea, the “extended keystone hypothesis,” considered that every community or ecosystem has several keystone species, which are those that dominate community structure and dynamics. Simultaneously, many researchers simplified the definition of keystone species to refer to those members of a community that had some large effect, regardless of the keystone’s relative abundance or trophic status. These alterations led Mills and coworkers to question the usefulness of the concept. Their critique stimulated a comprehensive review of these and related issues (Power et al., 1996; see Section II). One outcome of the recent controversy has been to refocus attention on keystone species and the overall importance of this concept in understanding the dynamics of ecological communities. This article surveys the roots of the issue, considers the precise meaning of the term, and reviews several case studies. It then addresses problems in methods of documenting keystone effects, in the identification of keystone species, and in determining the context under which keystone effects are likely. Finally, it evaluates the concept in a broader context of the strength of interactions among species, and concludes with comments on future directions of research.

I. HISTORICAL PERSPECTIVE

A. Identification of Pisaster ochraceus as a Keystone Species

In 1966, Paine (1966) reported the results of a study of the role of predatory sea stars in structuring marine intertidal communities on a rocky shore on the outer coast of Washington State. In temperate regions, such communities typically display a striking spatial pattern called “zonation.” The hallmark of this pattern is that dominant space-occupying organisms are arranged in vertically stacked horizontal bands. For example, temperate shores commonly have a band of barnacles and fucoid algae on the upper shore, a band of mussels on the middle shore, and a band of macroalgae on the lower shore. On the Washington coast, the macroalgadominated “low zone” also harbors various mobile and sessile invertebrates, including limpets, chitons, sea urchins, whelks, anemones, and, significantly, the sea star Pisaster ochraceus. A particularly intriguing feature of the zonation pattern was the sharp demarcation between the mussel bed and the low algal zone. Observing that low-zone-dwelling sea stars fed regularly on mussels, and that the only abundant source of mussels was higher on the shore in the middle zone, Paine postulated that predation by sea stars determined the sharp lower limit to mussels. Specifically, he speculated that at high tide, sea stars moved up from their low tide resting places in the low zone to remove prey from the lower edge of the mussel bed, then retreated back to the low zone where the prey were consumed.

To test the hypothesis that sea star predation determined the lower limit of the mussels, he periodically removed Pisaster from a section of a rocky outcropping and compared this section to an area in which sea stars had been left at natural densities. Within 3 years the mussel Mytilus californianus had increased in abundance, causing a reduction in local species richness of macroorganisms from 13 to 8 species (Paine, 1986).

Ultimately, after 10 years of excluding sea stars, a single-species monoculture of M. californianus dominated the shore (Fig. 1). Ancillary studies suggested that this change depended on two important mechanisms. First, Pisaster preferentially fed on mussels. This sea star actually will feed on many invertebrate species, but given a choice, it selects mussels. Second, mussels were dominant competitors for space. Field observations, and later experiments, indicated that mussels could displace all other occupants of space, including macrophytes, sessile invertebrates, and mobile invertebrates, by crowding them out. In so doing, mussels also created habitat for a sharply different set of invertebrate cryptofauna that occupied the byssal “forest” (the fibers that mussels use to attach to the rock) beneath the mussels.

Paine later repeated this experiment on a nearby coastal island (Tatoosh Island) with essentially identical results, further demonstrating that under natural conditions Pisaster indirectly enhances the persistence of other space occupiers by preventing mussels from in-
vading the low intertidal zone. The striking difference in species composition and physical appearance of the low zone community wrought by the star, a community member that was nowhere close to being "domi-
nant" either numerically or in biomass, was the basis for Paine's identification of Pisaster as a keystone predator. Note that contrary to some uses of the concept, the original definitions did not comment on keystone species as determinants of species diversity. Although the changes in species diversity that occurred in Paine's (1966) field experiments revealed an important role of such species, Paine and other early workers on the topic (Estes et al., 1978) evidently did not consider these changes to be fundamental to the concept of key-
stone species.

The formulation of the keystone species concept contributed to a gradual shift in ecological thought. Throughout the 1960s and 1970s, most ecologists believed that competition was the most important process structuring communities. By the late 1960s, advances in ecological theory and modeling, in conjunction with comparative and experimental field studies, demonstrated that processes other than competition could be important. Experiments such as those leading to the keystone species concept were among the first to de-
scribe how predation can dramatically affect commu-
nity structure. Following Paine's pioneering work, ecol-
ogists began to identify keystones in other systems.

B. Are There Others? Keystone Predators Elsewhere

One of the first documented examples of a keystone predator in another system was the sea otter, Enhydra

---

**FIGURE 1** Results of Pisaster removal experiments. Upper diagrams: "Reference" shows the natural interaction web in the low zone, and "Removal" shows the interaction web (i.e., Mytilus) 10 years after removal began. Arrows point to the species or group affected. Solid arrows show direct effects; dotted arrows show indirect effects. "Minus" by an arrow indicates that the effect was negative; "plus" by an arrow indicates that the effect was positive. Lower diagrams: left and center show covers of Mytilus, other sessile invertebrates (barnacles, anemones), and algae in 1963 (beginning of experiment) and 1973 in the presence and absence of Pisaster. Right diagram shows difference between covers in 1973 and 1963 for each group of sessile organisms. "Direct" indicates that increase in mussel cover was due to direct predation by sea stars. "Indirect" indicates that decreases in sessile invertebrates and algae were an indirect consequence of sea star predation, through competitive elimination of sessile invertebrates and algae by mussels. (Data from Paine, 1974, cited in Power et al., 1996.)
FIGURE 2  Total cover of macroalgae and sea urchin density by depth at three locations on Amchitka Island—Kirilof Point, Bat Island, and Kirilof Rocks (all with sea otters)—and one at Shemya Island (without sea otters). — vegetation cover; •, sea urchin density. (From Estes et al., 1978.)

lutris. Estes and coworkers (Estes et al., 1978) observed that subtidal communities in the Aleutian Islands, Alaska, where sea otters were abundant, differed from those in nearby areas where sea otters had been locally extirpated by hunting. In waters surrounding islands with sea otters, kelp forests at shallow depths dominated nearshore subtidal communities, and within these kelp forests sea urchins, an important consumer of kelp, were relatively small and urchin biomass was low (Fig. 2). On neighboring islands without sea otters, however, sea urchins were large; urchin biomass was high, and kelps were absent. On the basis of these data and the observation that sea otters preferentially feed on sea urchins, Estes and coworkers postulated that sea otters are a keystone species in this community. They argued that predation by sea otters reduced sea urchin grazing, thereby maintaining kelp forests and associated species.

By documenting the “keystone predator” phenomenon in a different ecosystem, this research bolstered the view that such dynamics were potentially widespread.

Subsequent work both supported this hypothesis and expanded our understanding of the conditions facilitating a keystone role for sea otters. In a large-scale study published in 1995, Estes and his colleague David Duggins determined the spatial generality of sea otter effects and tested the prediction that after sea otters colonize new areas, sea-urchin-dominated subtidal communities would become kelp-dominated. Surveys in the Aleutians and in southeast Alaska showed that the differences associated with sea otter presence or absence in the earlier studies were general in space—kelps were abundant in the presence of sea otters and scarce in the absence of sea otters. Further, considering all surveyed sites, sea urchin and kelp abundance were strongly inversely correlated (Fig. 3). Finally, sea urchin abundance declined sharply at sites invaded by sea otters but did not change at sites where sea otter abundance remained constant. An interesting regional difference was that rates of increase in kelp abundance in the presence of invading sea otters were higher in southeast Alaska. Estes and Duggins suggested that higher rates of kelp increase resulted from differences in the recruitment of sea urchins, which was much greater in the Aleutians than in southeast Alaska. Coupled with sea otter preference for larger, adult urchins, higher recruitment rates of juvenile urchins presumably led to moderate rates of grazing on kelp sporelings and juveniles in the Aleutians, slowing rates of increase in kelp abundance relative to that in southeast Alaska.

A recent complication with troubling conservation implications suggests that killer whale (Orcinus orca) predation has begun to eliminate sea otters from their native habitats with predictable consequences for abundances of sea urchins and kelp (Fig. 4). During the 1990s, sea otter abundance declined sharply at several Aleutian Islands. Once freed from sea otter predation, sea urchin populations increased, leading to severe declines in kelp and associated species. Estes and coworkers suggested that killer whales may have shifted their diet from large marine mammals to sea otters ultimately as a consequence of human activity. Overfishing and increased ocean temperatures (due to global climate change) are associated with declines in forage fish for sea lions and seals, and consequently these marine mammals have declined sharply in number. Historically these large marine mammals have been the primary prey of killer whales; in their absence, some orcas have shifted their diet to otters. Because sea otters are small.
FIGURE 3  Density of kelp (individuals/0.25 m²) versus estimated sea urchin biomass (g/0.25 m²) for the Aleutian Islands and southeast Alaska. Points show averages for sites within locations. Sea urchin biomass was estimated from samples of population density, size–frequency distribution, and the functional relation between test diameter and wet mass. [From J. A. Estes and D. O. Duggins, (1995). Sea otters and kelp forests in Alaska: Generality and variation in a community ecological paradigm. Ecol. Monographs 65, 75–100.]

relative to pinnipeds, rates of predation loss have evidently been high, with substantial direct and indirect consequences for subtidal communities.

C. Generalization and Liberalization of the Definition: Other Types of Keystone Species

These earliest examples of keystone species were of consumers that modified the community through predation. However, some investigators suggested that species other than top predators could play keystone roles in communities. Subsequently, a variety of different types of organisms were termed “keystone” species, including herbivores, plants, pollinators, pathogens, habitat modifiers, and mutualists. The remainder of this article presents examples of such alternative keystones, summarizes a critique of these ideas, and offers proposals for further clarification of keystone and other functionally important species.

1. Herbivores

By consuming primary producers, keystone herbivores can have dramatic impacts on community structure. Among species identified as keystone herbivores are kangaroo rats and bison. In the Chihuahuan Desert, Brown and colleagues (see summary in Brown, 1998) suggested that a guild of kangaroo rats played a keystone role. In plots where kangaroo rats were excluded, grass cover increased threefold. This was accompanied by a change in species composition; species shifted from those typical of desert shrubland to those characteristic of arid grasslands (Fig. 5). The mechanisms responsible for this transition were seed predation and soil disturbance by kangaroo rats. By preferentially eating the seeds of competitively dominant grasses, kangaroo rats indirectly released subordinate plant species from competition. Furthermore, kangaroo rat burrowing favored disturbance-tolerant plant species. Hence, the presence or absence of kangaroo rats determined whether the community was a desert shrubland or an arid grassland.

Note that this example both extends the concept to herbivores and attributes the keystone effect to a multispecies group instead of a single species. Recently Brown (1998) suggested that the kangaroo rat effect was attributable primarily to two species, Dipodomys merriami and D. spectabilis, and one of these (D. spectabilis) went locally extinct in 1994. If the system persists unchanged well beyond this extinction, these fortuitous changes may suggest that D. merriami is a keystone species. In our view, however, the original interpretation of these herbivores as a keystone “guild” may contribute to confusion in terminology. This study is an excellent example of the strong effects of consumers on community structure, but as explained later, it was not designed to discern between keystone versus “dif- fuse” (multispecies) predation.

Bison have also been identified as keystone herbivores (Knapp et al., 1999). Studies at Konza Prairie have documented strong effects of bison on species composition, diversity, and several physical and chemical aspects of ecosystem function of grassland commu-
KEYSTONE SPECIES

FIGURE 4 Changes over time in (A) sea otter abundance at several islands, and (B) sea urchin abundance, (C) grazing intensity, and (D) kelp density at Adak Island in the Aleutian archipelago. Error bars are 1 SE. Diagrams to the left and right of the data panels suggest the mechanisms leading to the changes. The left diagram shows kelp forest dynamics without orcas and the right diagram shows changes induced by the addition of orcas as the top predator. Thick arrows show strong effects, and thin arrows show weak effects. [From J. A. Estes et al. (1998). Killer whale predation on sea otters linking oceanic and nearshore ecosystems. Science 282, 473–476. Copyright 1998 American Association for the Advancement of Science.]
KEYSTONE SPECIES

2. Plants (Resources)

Terborgh (1986; cited in Power et al., 1996) suggested that certain plants in Peruvian rain forests can provide keystone "resources." Most tree species in these tropical rain forests fruit in synchrony. However, when these species are out of season, palm nuts, fig trees, and nectar-bearing plants can provide a crucial resource to frugivores. Less than 1% of the plant biomass of tropical forests is made up by these species and yet virtually all frugivorous animals rely on them during the 3 months when other food sources are rare. Consequently, Terborgh argued that these plants maintained high animal diversity in these communities and therefore occupied a keystone role. Although Terborgh's work focused on Peru, he suggested that keystone plant resources may be widespread in tropical rain forests.

In another example, anadromous fishes swimming upstream to spawn, or their carcasses, were suggested to be keystone resources. It was argued that a large number of vertebrate predators and scavengers, especially terrestrial species, rely on the energy input provided by anadromous fishes returning to their natal streams. Such energy sources can be strong interecosystem links and are suggested to have an influence on the diversity of terrestrial mammal assemblages.

We do not dispute the notion that such species or resources can be important to community dynamics, but for two reasons we question calling them keystones. First, in most such cases, the evidence for keystone effects is weak and conjectural; the investigators have not demonstrated that removal of the resources would lead to wholesale community changes. Second, with many others we believe that top-down effects (such as keystone predation, keystone herbivory) are qualitatively distinct from bottom-up effects (e.g., primary production, food supply). We therefore suggest that examples in which plant-level processes are thought to determine consumer effects, and which thus suggest that the plant is the keystone species, serve to confuse rather than clarify the concept by combining under a single idea a heterogeneous collection of ecological processes. Although there is increasing evidence in favor of the role of bottom-up effects as important determinants of community dynamics, these effects commonly lead to variation in top-down effects, often with single species (i.e., keystone) having large effects that are disproportionate to their abundance. We therefore argue that retaining the original distinction of keystone species as consumers is simpler and clearer, and hence better serves both understanding and insight into community dynamics and policy decisions regarding ecosystem management and conservation. Studying the

Table 1: Densities of perennial and annual grasses in the presence and absence of kangaroo rats (Dipodomys).

<table>
<thead>
<tr>
<th>Vegetation Type</th>
<th>Inside Enclosures</th>
<th>Outside Enclosures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eragrostis lehmanniana (ERIN)</td>
<td>Mean percent cover: 10% ± 2 SE</td>
<td>Mean percent cover: 20% ± 3 SE</td>
</tr>
<tr>
<td>Other tall perennial grasses (PERG)</td>
<td>Mean percent cover: 15% ± 3 SE</td>
<td>Mean percent cover: 30% ± 4 SE</td>
</tr>
<tr>
<td>Aristida adscensionis (ARAD)</td>
<td>Mean percent cover: 5% ± 1 SE</td>
<td>Mean percent cover: 10% ± 2 SE</td>
</tr>
<tr>
<td>Other tall annual grasses (ANNG)</td>
<td>Mean percent cover: 15% ± 3 SE</td>
<td>Mean percent cover: 25% ± 4 SE</td>
</tr>
<tr>
<td>Bouteloua aristidoides (BOAR)</td>
<td>Mean percent cover: 5% ± 2 SE</td>
<td>Mean percent cover: 10% ± 3 SE</td>
</tr>
<tr>
<td>Bouteloua barbata (BOBA)</td>
<td>Mean percent cover: 5% ± 2 SE</td>
<td>Mean percent cover: 10% ± 3 SE</td>
</tr>
</tbody>
</table>

Figures show significant differences between kangaroo rat and kangaroo rat enclosures. Asterisks above bars indicate significant differences between kangaroo rat enclosures. Bottom: Data from transects outside the respective enclosures. Asterisks between panels indicate significant differences between transects inside and outside plots where kangaroo rats were present or absent.
role of bottom-up processes as determinants of community structure is an exceedingly important area of research, but we believe progress both in this area and in the role of keystone species will be enhanced if they are treated as distinct but dynamically linked concepts.

3. Mutualists (Pollinators, Seed Dispersers)
Flying foxes (Pteropus spp.) have been proposed to be keystone “mutualists” in tropical rain forests. Flying foxes are large tropical bats that are important pollinators and seed dispersers in Old World forests and on tropical islands. Especially on isolated islands where other vertebrate pollinators are scarce, these bats can be responsible for dispersing 80–100% of the seeds. Cox and coworkers (cited in Power et al., 1996) argued that, through these pollinatation and dispersal “services,” flying foxes may be responsible for maintaining high plant diversity in the forest communities in which they occur. Comparisons of the proportion of fruits and seeds dispersed by flying foxes on Guam, where flying foxes have been driven nearly to extinction, and on Samoa, with abundant flying foxes, were consistent with this idea (see Ramey et al., 1995, cited in Power et al., 1996). Between 0 and 100% of fruits were dispersed by flying foxes on Western Samoa, whereas 0–1% of seeds were dispersed by flying foxes on Guam. Such data are only suggestive, however; further study is needed to demonstrate that forest community structure reflects these differences in fruit and seed dispersal, and that flying foxes play a keystone role in determining community structure.

4. Habitat Modifiers
Organisms that influence the availability of resources for other species by modifying the physical environment also have been considered keystone species (more recently the term “ecosystem engineers” has been coined to identify such organisms; see Lawton and Jones, 1995, cited in Power et al., 1996). Beavers, red-naped sapsuckers, kangaroo rats, and prairie dogs are all examples of species that, through their nonforaging activities, modify and/or create habitat for other species. By altering the hydrology of rivers, beaver dams can have profound effects on sediment retention, nutrient cycling, and the condition of the riparian zone (Naiman et al., 1986). Sapsuckers create habitat for two species of swallowbirds by drilling holes in aspens. In addition, their feeding holes in willows create sap flows, providing a resource for several species of birds, mammals, and insects (see Daily et al., 1993, cited in Power et al., 1996). The activities of kangaroo rats and prairie dogs cause soil disturbance, which affects community structure. These habitat modifiers were all thought to have impacts disproportionate to their abundance and were thus dubbed keystone species.

II. CRITIQUE AND REEVALUATION

Paine’s original definition of keystone species referred to top predators that greatly modified the species composition and physical appearance of the ecosystem by preferentially feeding on dominant competitors for space. Although not quantitative, the term was clear in concept and interpretation. As documented in the previous section, however, the term was subsequently applied to a variety of organisms other than top predators and to actions that affected communities in ways other than feeding on competitive dominants. Application to groups of species (e.g., keystone guilds) rather than to a single species blurred the concept even further. These definitional liberalizations, combined with a lack of rigor in determining if candidate species meet the necessary criteria to be termed “keystone,” eventually undermined the usefulness of a potentially powerful concept (Mills et al., 1993). Here we evaluate these issues and suggest a scheme of classification and definition (slightly modified from that offered by Power et al., 1996) that we believe further clarifies and advances the usefulness and application of the concept of key- stone species.

A. Reevaluating the Keystone Species Concept

The seeming lack of a working definition of keystone species prompted Mills et al. (1993) to argue that careless usage had sufficiently degraded the value of the concept to justify removing it from usage. They further argued that because the term was poorly defined it was functionally useless and conservation strategies therefore should not be based on protecting keystone species. They advocated focusing on interaction strength rather than on a species’ keystone or nonkeystone status as a more useful management strategy.

In response to the criticism by Mills et al. (1993), a group of ecologists with expertise in the study of keystone species and strongly interacting species met to evaluate the keystone species concept. The publication that resulted from this meeting was a signal achievement, and much of the present article is patterned after this synthesis (Power et al., 1996). The group agreed
with Mills et al. (1993) that through misapplication and questionable redefinition, ecologists and conservation biologists had obscured the meaning of the term keystone species. Rather than abandoning the concept, however, the group proposed clarification and adherence to a set of more sharply defined concepts for community dynamics. They defined a keystone species as "one whose impact on its community or ecosystem is large, and disproportionately large relative to its abundance" (Power et al., 1996). This definition retained the essence of Paine's original usage, but expanded it more broadly to include species other than predators.

Power et al. (1996) stressed the importance of having a rigorous, quantitative method of assessing the community- or ecosystem-level effects of a species when determining if it is a keystone. To assess a species' impact on a community they proposed a community importance (CI) index, which in practical terms is quantified experimentally as

$$CI_i = \frac{(t_i - t_{N})/t_{D}}{(1/p)}$$

where $p_i$ is the proportional abundance of the species $i$ before it was deleted, $t_i$ (for "normal") is a quantitative measure of a community or ecosystem trait under usual conditions (e.g., productivity, nutrient cycling, species richness, relative abundances of species), and $t_{D}$ (for "deleted") is the trait in the absence of species $i$. Keystone species are those whose community importance (CI) is large relative to that of other species (Fig. 6). On the basis of these considerations, Power et al. (1996) suggested that if the total impact of a species is plotted against its proportional abundance, keystone species would cluster toward the upper left of the graph in Fig. 7. Power et al. (1996) defined another group of species, "dominants," as those that had large total impacts but in proportion to their abundance—these would lie to the upper right of the abundance–impact diagram.

Importantly, keystone species are not necessarily simply "strong" interactors, at least as we define the terms. Keystone species not only have disproportionately large effects, but also have a community-wide or ecosystem-wide impact through direct and indirect effects cascading through the system (Power et al., 1996; see Fig. 1). Strong interactors are species that have a large impact on the species with which they interact, but this large impact could affect only a single interacting species. Only when strong interactors have multispecies effects that alter the structure of the community, or the functioning of the ecosystem, are they also keystone species.

### B. Related Concepts

The idea that not all species have equal significance in community dynamics is an old one, and many efforts have been made to assign names or terms to distinguish species or groups of species with important roles. Elton (1927) advanced the concept of "key-industry species" as single species of animals supporting a large number of consumers (e.g., copepods, herring, anchovies, sea pens). "Foundation species" were defined as the "group of critical species which define much of the structure of a community" (Dayton, 1972). Related ideas include "interaction webs" and "functional webs," which are defined as the subset of strongly interacting species that regulates community structure. A similar idea, the "extended keystone hypothesis," states that ecosystems are controlled by a small set of key plants, animals,
KEYSTONE SPECIES

FIGURE 7 Total impact of a species (absolute value of community impact divided by proportional abundance of a species) versus its proportional abundance, $p_i$. Species whose total impact is proportional to their abundance would fall along the diagonal line $X = Y$. Keystone species have effects that greatly exceed their proportional abundance, both on a per-capita (or per-biomass) basis and a per-population basis, and would cluster toward the upper left region of the diagram. To illustrate, some organisms (e.g., rhinovirus, $V_R$, which causes colds in wildlife) may have effects that are greater than expected from their biomass, but because the impact on the community is relatively minor, they are not keystone species. Others (e.g., distemper virus, $V_D$, a killer of lions or wild dogs) may have collective effects on the community that are disproportionately large, and would be keystone species. Examples of keystone species are $Pisaster (P)$, sea otters ($O$), predatory whelks ($Concholepas$, $C$), and freshwater bass ($B$). Dominants are high proportional biomass species whose large effects are not disproportionate to their biomass, such as trees ($T$), giant kelp ($K$), prairie grasses ($G$), and reef-building corals ($Cr$). Positions of each species or group represent educated guesses. (From Power et al., 1996.)

C. Recommended Terminology

Although the various generalizations of the keystone concept and the lumping of several related concepts into a broad keystone species definition were well intentioned, we agree with Mills et al. (1993) and Power et al. (1996) that such usage sharply reduces the utility of the keystone species concept. On the basis of current usage and the documented need and desire for clarity in standardized terminology, we propose the adoption of the following terminology. Species in communities can be defined as:

- **Keystone species**: Consumers having a disproportionately large effect on communities and ecosystems. By this definition, keystone species can include predators, parasites, pathogens, herbivores, pollinators, and mutualists of higher trophic status but not plants, sessile animals or "resources" (e.g., salmon carcasses, salt licks, deep pools). To date we know of no convincing examples of communities with more than a single keystone species. We therefore suggest the hypothesis that most communities will have at most a single keystone species. At the ecosystem level, which might include several distinct communities, there may be several keystone species. Evaluation of these predictions awaits future research.

- **Strong interactors (critical species)**: Species having a large effect on the species (one or more) with which they interact. Communities and ecosystems may have many strong interactors, and such species may occur at all trophic levels. Strong interactors would lie toward the upper side of the abundance-impact diagram (Fig. 7), and therefore include both keystone species and dominants.

- **Weak interactors**: Species having little effect on other species, at least under average conditions. Under some circumstances, weak interactors may occupy important roles in ecological communities as a result of changes that lead to temporary increases in abundance, size, or biomass (Berlow, 1999). Weakly interacting species would all lie toward the lower portion of Fig. 7.

- **Dominant species**: As noted above, dominant species are those strongly interacting species that owe their influence to their high abundance (Fig. 7). Such organisms are the species that comprise a large proportion of the biomass in a community, and thus are the dominant components of community structure. Trees in forests, mussels in rocky intertidal habitats, grasses in grasslands, and kelps or corals in nearshore subtidal habitats are all dominants.

- **Key-industry species**: As defined earlier, key-industry species are prey that support a large group of consumers. Following Elton's (1927) usage, key-industry species are therefore animals of intermediate trophic status.
Groups of species in communities or ecosystems can include:

- **Interaction webs**: Interaction webs (= functional webs) are the subset of species that through their interactions and responses to abiotic factors make up the dynamic core of food webs or communities. Interaction webs include keystone species, dominants, and other strong interactors.

### III. IDENTIFICATION OF KEYSTONE SPECIES

Lack of experimentation and other rigorous approaches to identify keystones continues to be a pervasive problem. In many cases, species are named keystones based on superficial evidence such as natural history observations. A lack of rigor in identification could result in mislabeling a species as a keystone and inferring that it is the primary determinant of the structure of its community when it is in fact not. As noted by Mills et al. (1993), this can yield a serious loss of credibility with respect to policy and management decisions.

Despite limitations of spatial and temporal scales and other shortcomings, experimentation supplemented by comparison remains the most powerful tool available for revealing the dynamics of communities and ecosystems (Paine, 1994). In many cases, however, practical, legal, and ethical concerns preclude manipulations of suspected keystones (e.g., sea otters, killer whales, or prairie dogs). It is clear, therefore, that identification of keystone species must necessarily use a variety of approaches. In addition to the comparative-experimental approach, alternative approaches include path analysis, sensitivity analysis, the study of natural or accidental invasions, the study of the consequences of overexploitation, and exhaustive and detailed comparison coupled with natural history and ideally small-scale ancillary experimentation (Power et al., 1996). Inferences based solely on descriptive natural history knowledge (e.g., diet composition and frequencies, behavior, abundance) are likely to be misleading. For example, under natural conditions a keystone predator may rarely consume the competitive dominant in a system (because it has sharply reduced the prey's availability) and thus be overlooked as a possible regulator of the dominant or the community. In the sea star system studied by Paine (1966), for example, barnacles, not mussels, were the most frequently consumed prey. Thus, although natural history knowledge is fundamental to the understanding of the dynamics of any ecological system, much additional evidence is necessary before the ecological role of a species can be determined.

### A. Experimental Approaches

Keystone species are often identified by removal or exclusion experiments; that is, the presumed keystone is deleted from a portion of the habitat, and the effect of the removal on the community is compared to a separate, control portion of the habitat. One problem with this approach is that if a suspected keystone species is removed but there is no detectable response by the community, it is not possible to conclude that the community lacks a keystone species. In such a case, three alternative interpretations are possible. (1) Predation is weak overall such that no predator deletion will produce an effect. (2) Predation is strong and there is a keystone species in the community, but it was not the species removed. (3) Predation is strong but diffuse, and multiple predators would need to be removed to produce an effect.

To tease apart these alternatives, an appropriate experimental protocol in testing for keystone predation (Menge et al., 1994; Navarrete and Menge, 1996) should include:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Explanation</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. +Predators</td>
<td>Intact community, all predators present</td>
<td>Control or reference (&quot;natural&quot; community state)</td>
</tr>
<tr>
<td>2. –Predators</td>
<td>Quantifies total predation effect, all predators removed or excluded</td>
<td>Strong versus weak predation</td>
</tr>
<tr>
<td>3. Single Predator</td>
<td>Quantifies single-species effects, deletion of each predator species singly, leaving the others present</td>
<td>Keystone versus diffuse predation</td>
</tr>
</tbody>
</table>

For example, if consumer species are removed both collectively and singly, but predation is weak, the total effects of predation and the effects of each species with respect to their impact on community structure will be small (Fig. 8a). With strong predation regimes, keystone predation would be indicated if total removals and removal of just one of the predator species were similarly large (Fig. 8d). Finally, diffuse predation would be indicated if removal of each predator species...
FIGURE 8  Total effects (solid bars) of a consumer group and effects of single species (open bars) under three regimes of predation effect, with possible responses to sequential loss of single species. Panels a, d, and g represent the intact assemblage; b, e, and h represent the changes in the absence of species 1; and c, f, and i represent the changes if both species 1 and 3 are removed or lost. Panels a, b, and c show change with weak predation; d, e, and f show changes with strong keystone predation; and g, h, and i show changes with strong diffuse predation. In the absence of the keystone species, the remaining species have a small impact on the community (f). In contrast, with diffuse predation, the remaining species compensate strongly for the loss of a consumer so that predation remains strong. (From Allison et al., 1996. Copyright John Wiley & Sons. Reproduced with permission.)

singly led to significant but substantially smaller effects than removal of all predator species (Fig. 8g).

This protocol also has important implications for understanding how communities or ecosystems might respond to losses of trophically high species (Allison et al., 1996). In a system where predation is weak, of course, loss of species should have little effect (Fig. 8a–c). With strong diffuse predation, where each consumer has an impact on community structure, species loss may lead to little change in the system as a result of compensation by the remaining consumers, at least until most consumers have been lost (Figs. 8g–i and 9). In both cases, the community response to species depletion is relatively predictable (Allison et al., 1996). With strong keystone predation, however, the consequence of species loss is relatively uncertain (Figs. 8d–f and 9). By definition, compensation by the weakly interacting predators, while possible, is not likely to fully account for the loss of the keystone species, and if the identity of this species is not known a priori, the system response is highly uncertain (see Fig. 9).

This protocol has rarely been used, but two examples illustrate its efficacy. Menge et al. (1986) evaluated the separate and combined impacts of most combinations of four groups of consumers in rocky intertidal communities in Panama. The extremely high diversity of this tropical community imposed an immediate compromise in their design: simultaneous single species removals were essentially impossible because >40 consumer species were relatively abundant. The compromise was to remove major consumer groups (i.e., omnivorous crabs, omnivorous fishes, predatory whelks, and grazing molluscs), each consisting of several common species. Their partial factorial design (difficulties in separating crab and fish effects meant only 12 of a possible
KEYSTONE SPECIES

16 treatments could be conducted) included treatments assessing the effects of removing consumers in single groups and in total. After three years, none of the single-group removals produced an effect similar in magnitude to excluding all the consumers (Fig. 10). Removals of two groups and three groups produced changes in prey that were intermediate between single-group and total exclusion treatments. Thus, total predation pressure in this system was high but the effect of individual groups was small. The lack of prey responses from single-group removals compared to the increasingly strong responses from removals of two, three, and all four groups suggests three things: that, up to a point, groups can compensate for reductions in other groups; that there was no keystone species in this system; and that predation was diffuse.

In another study, Navarrete and Menge (1996) tested the relative intensities of predation on the small mussel *Mytilus trossulus* by the original keystone species, the sea star *Pisaster ochraceus*, and coexisting predatory whelks, *Nucella emarginata* and *N. canaliculata*, on the Oregon coast. In a fully factorial design they quantified rates of predation on mussels in all combinations of presence or absence of each. Their results indicated that, as predicted, *Pisaster* had the strongest effect, by far, on mussel survival (Fig. 11). Although whelk effects were almost undetectable in the presence of sea stars, they were relatively strong in the absence of sea stars. Thus, *Pisaster* was the keystone predator in these experiments. Whelks were weak interactors in the presence of sea stars, but, via compensatory increases in density, had moderately strong effects on mussel survival when sea stars were absent.
FIGURE 11 Results of a mussel transplant experiment. Data are the mean proportions of mussels (±1 SE) surviving in each of four treatments at wave-exposed and wave-protected sites at two study sites, Boiler Bay (BB) and Strawberry Hill (SH). Treatments were all combinations of presence and absence of whelks (Nucella spp.) and sea stars (Pisaster).

The experiment was begun 3 July 1993 and completed 30 August 1993. (From Navarrete and Menge, 1996.)

B. Nonexperimental Approaches

To be convincing, comparative evidence needs to be extensive in space and time and needs to incorporate a wide range of approaches. One of the most convincing examples of the determination of keystone species on the basis of description and comparison is the sea otter example described earlier. Another convincing example is that of beavers as keystone species (Naiman et al., 1986). By building dams across streams and through feeding activities, beavers can have profound effects on the structure and dynamics of aquatic ecosystems. Beavers alter hydrology, channel geomorphology, productivity, and biogeochemical cycling at a magnitude far exceeding their proportional biomass in these systems.

Bison are another good example of keystone species designation based on largely nonexperimental evidence (Knapp et al., 1999). Extensive and detailed studies at the levels of individual leaves, individual plants, plant populations, and landscapes showed that selective grazing at the level of species and patch greatly altered patterns of community structure. Bison preferred grasses in burned areas, and revisited these sites repeatedly, establishing a distinct, high-diversity mosaic pattern of vegetation. Their grazing on grasses and avoidance of forbs resulted in high levels of species diversity, high spatial heterogeneity, and high nitrogen availability in tallgrass prairie. Moreover, bison grazing interacted with fire, another natural element of tallgrass prairies, in ways that enhanced the patchiness, diversity, and nitrogen availability, and generated intermediate levels of net primary productivity. As mentioned earlier, estimates of the CI of bison indicate that these large grazers clearly have a large effect, one disproportional to their biomass, on plant community structure and ecosystem functioning.

Space does not permit further detailed citation, but other convincing examples of keystone species are available from marine, freshwater, and terrestrial habitats, including those cited in Table 1 in Power et al. (1996). Many more examples than these are available in the literature that are based on limited evidence and thus remain conjectural. It seems clear, however, that the phenomenon of keystone species is real and widespread, and that a variety of approaches are available to document their existence and importance. What remains unclear from the foregoing is (1) whether or not there are keystone traits that allow a priori identification of keystone species, and (2) whether or not there are predictable conditions under which keystone species will evolve and persist. We consider these issues next.

C. Identification of Keystone Species a Priori?

Field experimentation is not always feasible. Moreover, even where and when feasible, such studies are often time-consuming and expensive. It seems obvious that we cannot hope to study each and every system on Earth to determine the nature of interactions among species and how these influence community dynamics. For these reasons an ultimate goal in ecology is to gain the ability to predict the roles of species and the processes underlying the dynamics of communities and ecosystems. Ideally, predictions would be based on observation and measurement of organismal, population, and community traits and patterns. Evaluation of the reliability of such predictive traits would depend on cycles of study that identified important traits in particular systems and tested their predictive capacity in novel systems.

Traits originally suggested to characterize keystone species were differential impacts on prey species, high consumption rates relative to prey production, and a
focus of the keystone's impact on dominant competitors in the system. A recent survey of well-studied keystone predator examples in marine and freshwater habitats, however, suggested that these, and several other postulated 'keystone-identifying' traits, were not predictably associated with keystone species (Menge et al., 1994; Table I). In all, 17 studies of strong predation were considered. Eleven were dominated by keystone predation and six were dominated by diffuse predation. As indicated in Table I, differential predation and dominance of the community by a single prey species in the absence of the predator were characteristic of both keystone and diffuse predation systems, and thus do not serve as a distinguishing trait of keystone predation. Although data were limited, prey production rate was the only trait that consistently distinguished keystone (high prey production) and diffuse (low prey production) predation systems. No other potential trait was consistently associated with either type of system. Clearly further research is needed, but based on present knowledge, a priori identification of keystone species using traits alone does not seem possible.

D. Context Dependency

The role of a species as keystone can be context dependent (Menge et al., 1994; Power et al., 1996), that is, a species that serves a keystone role under some set of biotic and/or abiotic conditions may not under other conditions. For example, at exposed headlands along the Oregon coast, the sea star *Pisaster ochraceus* was found to occupy a similar role to that documented for this species in comparable habitats on the Washington coast (Menge et al., 1994; Paine, 1966). On more wave-sheltered rocky shores, however, *Pisaster's* role was weak to absent, despite the presence of populations of this sea star at all locations studied (Menge et al., 1994). The basis for this change was that environmental conditions changed with diminished wave impact. At one sheltered site, prey production rates (a combination of recruitment rates and prey growth) were so low that the sea star population consisted of a few scattered individuals. This and other studies suggest that sea star abundance is highest where food concentrations are high. At the other sheltered site, periodic and unpredictable burial by sand, not sea star predation, eliminated the competitively dominant mussels (*Mytilus californianus*) from the lower shore (Menge et al., 1994). Keystone predation in this system, then, depended on the spatial and environmental context in which the community occurred.

Keystone effects may also be altered by temporal environmental changes. Recent studies on the Oregon coast showed that *Pisaster* foraging activity is strongly
suppressed by cold water temperatures such as those occurring during periodic upwelling (Sanford, 1999; Fig. 12). Although further study is needed to determine the broader significance of this result, sea temperature variation in time or space could also modify the community role of this species. In particular, alteration of upwelling patterns by global climate change could lead to community- and ecosystem-level modifications or shifts in community pattern over a broad geographic region. Thus, investigations conducted over a wide range of environmental conditions indicated that whether or not a species can fill a keystone role is dependent on physical and biological conditions (i.e., context).

In a second example of context-dependent keystone predation, juvenile steelhead were demonstrated to control summer food web structure in California rivers in years experiencing winter flooding. Steelhead were the top predator in a four-level food chain, and when present they controlled invertebrate predators, freeing herbivorous insects from consumer control and allowing them to control the abundance of algae (Power, 1990; cited in Power et al., 1996). When steelhead were excluded, invertebrate predators increased and controlled herbivorous insects, resulting in a high biomass of algae. Winter floods swept these communities away, leading to annual cycles of community redevelopment, with shifts in algal biomass reflecting the development of successive trophic levels through the summer into autumn.

In years without winter flooding, predation-vulnerable herbivorous midge larvae (Pseudochironomus) were replaced by predator-inulnerable herbivores (e.g., aquatic moth larvae Petropbilia, caddisflies Tinodes and Glossosoma) (Power, 1992, 1995; cited in Power et al., 1996). The invulnerable herbivores sharply suppressed algal abundance and, due to their effective defenses against predators, effectively shortened the food chain from four to two levels. Thus, fishes (juvenile steelhead) were keystone predators in the context of years with winter flooding and were not in the context of years without winter flooding.

A variety of similar examples are available from all major habitats (see Table 2 in Power et al., 1996), suggesting that context dependency is common and perhaps near-universal in ecological communities. Such variation may be a useful tool in investigating the conditions under which keystone species will evolve and persist. Understanding the conditions under which a species plays a keystone role is critical if ecologists are to build a predictive framework to evaluate the effects of global changes in biodiversity.

E. Interaction Strength

As noted earlier, keystone species are distinguished by both the strength of their interactions with other species and the large indirect consequences of these effects. Since determination of whether or not a species is a keystone depends on rigorous application of methods that can quantify these effects, it is important to briefly consider the meaning of the concept of interaction strength and the methods used to estimate it.

The concept of interaction strength has different meanings depending on the context. Theoretical models, such as the Lotka–Volterra models of interspecific competition, include a coefficient, $a_{ij}$, which “measures how strong the interactions are” (MacArthur, 1972). This coefficient reflects the per-individual or “per-capita” effect of one species $j$ on another, species $i$. Similar coefficients are included in predator–prey models as well. MacArthur (1972) also suggested a second meaning of interaction strength: interactions can be considered as strong if their “removal would produce a dramatic effect.” In this case, interaction strength refers to
population-level, not per-capita, impacts of one species on others. Despite its long history in ecology, the concept of interaction strength has only recently been explored by experimental field ecologists. Paine (1992), for example, quantified per-capita interaction strengths in field enclosure/exclosure experiments involving herbivores (sea urchins, limpets, and chitons) grazing on kelp sporelings. Specific impacts of individual herbivore species were determined by comparing the densities of kelp sporelings that occurred in the complete absence of all herbivores (a measure of total herbivore pressure) to those occurring in the presence of single herbivore species (a measure of a single species' effect). Per-capita effects were estimated by dividing by the number of herbivores used in each single-species treatment.

This analysis showed that per-capita species interaction strengths varied among the species of this herbivore assemblage. Three interactors were described as “strong.” Sea urchins and a large chiton had large negative effects on algae. A smaller chiton had relatively large positive effects on kelp sporeling densities in one experiment and weak effects in another, showing that interaction strength can vary within species. The other four species had no distinguishable per-capita effects and were therefore weak interactors. Similar efforts have been made with preying mantis interactions with its insect prey, with predation effects on muddy invertebrate species (see Berlow et al., 1999, and citations therein).

One important outcome of these studies, consistent with expectation, was that these communities or assemblages consisted of a few strong and many weak interactors (Paine, 1992). More broadly, this approach is a potentially powerful tool in distinguishing systems with weak predation from those with strong predation. Further, in those systems with strong predation, this method should separate keystone predation regimes from “diffuse” predation regimes.

Quantification of interaction strength is sensitive to the particular metric used in its estimation and to the particular theoretical concept of interaction strength (Berlow et al., 1999). Estimates of per-capita effects can vary with density of the prey and predator, with the proximity of the system to equilibrium, and with the particular index used. For example, both the CI index given earlier and the index used by Paine (1992) yield measures that are asymmetric about zero, inflating either negative (CI) or positive (Paine's index; PI) values in one direction but bound at a maximum of 1 (or –1) in the other direction (Berlow et al., 1999). Another index, the “dynamic index” (DI, an index that is based on the rate of change immediately after a perturbation), provides symmetrical estimates of interaction strength bounded at –1 and +1. The CI and PI indices provide consistent estimates when abundances are at or near equilibrium but DI does not; when abundance is changing, DI provides consistent estimates but CI and PI do not. It is thus important that investigators pay close attention to the specific experimental protocol, the underlying model for the interactions studied, and the type of index used.

Using these approaches, keystone species will be those with large per-capita effects. Identifying such interactors in this way might shed light on the traits, both individual and community, that generate species with high per-capita effects. As implied by the alternative meanings for interaction strength offered by MacArthur (1972), however, species impacts will also vary with density. Regardless of how large its per-capita effect, a single individual will never have the same impact as that of many individuals (Berlow et al., 1999; Navarrete and Menge, 1996). Thus, fuller understanding of interaction web dynamics, and whether a species is a keystone, a strong interactor, a weak interactor, or a dominant, will depend on knowledge of both per-capita interaction strength and population interaction strength or “species impacts.” Note also that only a keystone species will have high values of both per-capita and per-population indices of interaction strength, at least at the community or ecosystem level.

IV. CONSERVATION IMPLICATIONS

Biologists are being increasingly asked to inform management decisions concerning the preservation of biological diversity and ecosystem “integrity.” Because keystone species can be critical to the maintenance of species diversity and ecosystem functioning, some have argued that focusing conservation on them should be a priority. A current topic of debate is how to best design reserves, both marine and terrestrial, that will protect species diversity.

Many species have been presumed to be keystones based on observation and conjecture. In many cases no experimental removals were done to determine the actual effect of the species presumed to be a keystone on the rest of the community. As outlined earlier, identification of keystones should involve manipulations to quantify the effect of the suspected keystone on the community. However, from a management standpoint this is not always feasible. It sometimes takes years for
field manipulations to yield results, while conservation decisions often need to be made in a much shorter period of time. Also, as noted earlier, in addition to potential time constraints, manipulations of some suspected keystones may not be practically feasible or ethical. Removing killer whales, sea otters, lions, polar bears, or other large, charismatic, and often endangered species is not a realistic option. The logistic difficulties in designing controlled enclosure or removal experiments for animals of large size and/or with large ranges are daunting.

Despite the difficulties, and accepting the view that conservation should focus on ecosystems rather than species, the keystone species concept offers several important insights that are relevant to management (Power et al., 1996):

- Seemingly scarce and unimportant species may have unexpectedly large, dramatic effects on communities and ecosystems;
- Conserving species may depend strongly on other species in the community with which the target species has little seeming association, whether as prey, competitor, mutualist, or predator; and
- Loss of a species, particularly one high in the food chain, may have surprising and extensive consequences for the remainder of the community or ecosystem.

All of these points suggest that great caution is necessary before decisions are made that may result in the loss of a native species or the introduction of exotic species. We recognize, and agree with, the need to make management and policy recommendations on the basis of present knowledge. We also argue that, although much has been learned and the pace of new knowledge is increasing, the current state of knowledge is still insufficient to allow making such recommendations with a high level of confidence. In particular, we need additional study on those putative keystones for which evidence is scant, and continued study on ways to detect keystone species and systems with keystone species. Other urgent issues are to determine the commonness of keystone species and keystone-dominated systems, the patterns of interaction strengths in representative communities, and the environmental and organismal conditions and traits that foster keystone species.

We conclude that the concept of keystone species is a powerful one, having broad importance and application to ecological theory, ecosystem dynamics, and conservation. Many problems remain to be answered, but we have made great progress and are in the midst of a period of dramatic advances on these issues.

Acknowledgments

We thank Kraloha Freidenburg, Heather Leslie, Bob Pane, Mary Power, Eric Sanford, David Skelly, and Michael Webster for constructive criticism. This article was written while the authors were supported by grants from the Andrew Mellon Foundation (B.A.M.), the Wayne and Gladys Valley Foundation (B.A.M.), the David and Lucile Packard Foundation (B.A.M.), and an NSF Graduate Fellowship (T.L.F.). Contribution number 9 from PISCO, the Partnership for Interdisciplinary Studies of Coastal Oceans: A Long-Term Ecological Consortium, funded by the David and Lucile Packard Foundation.

See Also the Following Articles

Competition, Interspecific • Conservation Biology, Discipline of • Ecology, Concepts and Theories in • Ecosystem Function, Principles of • Food Webs • Predators, Ecological Role of • Species Coexistence

Bibliography


LAKE AND POND ECOSYSTEMS

Christian Lévêque
Institut de Recherche pour le Développement (IRD) Paris

1. Broad Characteristics of the Biodiversity in Lakes and Ponds
2. Evaluating Biological Diversity
3. Biological Diversity and Ecosystem Functioning
4. The Case of Ancient Lakes Species Flocks
5. Major Threats to Biodiversity in Lakes

GLOSSARY

ancient lakes Lakes with a persistence of more than 100,000 years are called long-lived or ancient lakes.
eutrophication The process of enrichment of a water body due to an increase in nutrient loading.
species flocks An aggregate of closely related species that share a common ancestor and are endemic to a geographically circumscribed area.

COMPARED TO THE SEA, freshwater habitats are deficient in major taxa and there are no uniquely freshwater metazoan phyla (May, 1994). In river lakes, aquatic biota is similar to the biota of the river basin. The great majority of existing isolated lakes (around 10,000 exceeding 1 km² in extent) are geologically very young and their flora and fauna are usually depauperate compared to ancient lakes that exhibit a rich endemic fauna for several major groups of animals. Species flocks for fish and invertebrates are known from ancient lakes and represent a unique biological heritage to be preserved. Major threats to the lakes biota include habitat alteration, fisheries practice, pollution, and the introduction of exotic species.

1. BROAD CHARACTERISTICS OF THE BIODIVERSITY IN LAKES AND PONDS

Freshwater habitats are widely considered to be transient in time and space in comparison with both terrestrial and marine habitats. This is true for many lakes and ponds. However, depending on the origin of lakes there are great differences in the nature and the diversity of their biota. Three broad categories of lakes may be recognized:

1. Lakes and ponds that are permanently or frequently connected to large river systems. This category includes river-lakes (i.e., Lake of Geneva) or lakes that are part of a large floodplain system such as ‘várzea’ lakes. In these lakes, exchanges of flora and fauna occur with the main river system so that their biota is usually greatly similar to the biota of the river system itself with the exception of a few species adapted to still waters. Many endorheic lakes (Lake Chad, Aral Sea) also belong to this group.
2. Isolated lakes with a limited drainage system. The biota of the lakes in this category has evolved in isolation from others for a more or less long period of time leading to speciation and endemism when the period...
is long enough. The associated ice ages at higher latitudes and altitudes were the phenomena that created most of the lakes in existence today. Therefore the great majority of existing lakes (around 10,000 exceeding 1 km² in extent) are geologically very young and occupy basins formed by ice masses or glacial erosion after the retreat of continental ice sheets some 10,000 years ago. All such lakes are expected to fill slowly with sediment and to disappear in the future, along with any isolated biota. Compared to ancient lakes, they acquired their fauna and flora via the rivers that supply them with water as a result of runoff in their basin and from aerial transport by wind or animals. Only a few existing lakes are known to be much older, and most of them occupy basins formed by large-scale subsidence. They may date back at most 20 million (Lake Tanganyika) or 30 million (Lake Baikal) years. These so-called ancient lakes are of particular interest for biodiversity because they exhibit a rich endemic fauna for several major groups of animals. There is also good evidence that some extinct lakes were also very large and long-lived under different climatic and tectonic conditions.

3. Temporary lakes and ponds whose water budget is controlled by the climate regime. The fauna and flora in these lakes exhibit special biological adaptations to seasonal drying.

A. Origin and Peculiarities of Freshwater Biota

It is thought that the early evolution of all the major animal phyla took place in the sea. Most phyla are predominantly marine and benthic: 32 phyla are found in the sea with 11 exclusively marine, whereas 14 are represented in freshwater and only 12 are found on land (May, 1994). Compared to the sea, freshwaters are deficient in major taxa and there are no uniquely freshwater metazoan phyla. The osmotic challenges of life in freshwaters probably discouraged invasion of the habitat by many marine invertebrates. It explains probably the tendency in freshwater invertebrates for larger but fewer eggs than in marine relatives: they must eclose with fully developed osmoregulatory capacities to be at a more advanced stage to cope with the highly dilute surrounding.

Another difference between the species richness of marine and freshwater zooplankton derives from the necessity of diapause or other resting mechanisms as a condition for persistent successful radiation in freshwaters (Lehman, 1998). Freshwater invertebrates developed anaerobic devices: special resistant eggs, cysts, and other resting stages that are produced to tide the animal over periods of desiccation, extreme cold, heat, anaerobic situations, lack of food, and other adverse conditions. In addition to withstanding unfavorable conditions, resistant stages have the further function of making overland transport and geographical dissemination possible. Without such a function, colonization of freshwater areas would be slow and difficult in the discontinuum of isolated lakes and ponds.

B. The Latitudinal Gradient

It is usually assumed that species diversity increases from high to low latitudes for most of the major groups of plants and animals and that highest values occur at low latitudes. Indeed, the diversity of marine plankton decreases from low latitudes to high ones, so that tropical and subtropical ocean waters exhibit rich diversity of zooplankton whereas Arctic and Antarctic waters tend to be dominated by copepods and euphausiids. In freshwaters however, the latitudinal trend in species richness is the opposite. Tropical lakes have abbreviated zooplankton faunas compared with temperate locales (Fernando, 1980); they are depauperate in large-bodied species of copepods and Cladocera, and limnetic rotifers are likewise poorly represented. It could be assumed that the associated ice ages at higher latitudes and altitudes were the phenomena that created most of the lakes in existence today. They are therefore very young compared to ancient lakes, and they acquired their fauna and flora via the rivers that supply them with water via runoff in their basin and via aerial transport by wind or animals.

For fish, the species richness is actually smaller in north temperate lakes of glacial origin than in long lived lakes from tropical areas. At least the endemicity is much lower in temperate lakes than in tropical lakes. Dumont (1994), in a review of the species richness of the pelagial zooplankton in ancient lakes, provided also evidence that these water bodies have simple pelagial communities. Among 14 pre-Pleistocene lakes across the world, at least one Cyclopoid copepod species is present, often in the genus Cyclops or Meso-cyclops, a group of microplanktonic species feeding on rotifers, small crustaceans, and immature stage of other copepods.

The number of species regularly found in the pelagic plankton of ancient lakes (pre-Pleistocene) varies from 3 (Lake Tanganyika) to approximately 15 to 20 (up to 5 copepods, 3 cladocerans, 10 rotifers) in Lakes Victoria, Biwa, and Titicaca. In contrast, “young” lakes may have up to 10 species of copepods, 10 of cladocera, and 10 to 15 species of rotifers occurring together. In the oldest
lakes (Baikal, Tanganyika), which also happen to be the deepest, this simplification has gone extreme and the food web reduces to a linear chain. The question has been raised as to why Cladocera have been almost completely eliminated from some ancient lakes such as Tanganyika, Baikal, and even Malawi (Dumont, 1984). They are able to eat both large items and microplankton and seem all but competitively inferior to other species for food acquisition. Predation has been advocated as a possible cause. In clear-water lakes such as Tanganyika and Baikal, visual predation by fish is more effective than in turbid lakes, and large clumsy swimmers like big Daphnia are likely to be preyed to extinction before the relative small transparent, agile swimmers like the Calanoïds. An experimental demonstration of this hypothesis has been the disappearance of all Cladocera from the pelagial of Lake Kivu, within a decade, after the introduction of the zooplanktivore clupeid Limnothrissa miodon, native from Lake Tanganyika (Dumont, 1986).

C. Vertical Distribution in Lakes
Aquatic organisms are not evenly distributed along depth. Water characteristics are relatively uniform in shallow lakes, which are mixed by winds. However, deeper lakes exhibit patterns of vertical gradients for temperature and light. Briefly speaking, the lake is divided by a thermocline into an upper layer, the epilimnion, and a lower layer, the hypolimnion. Life occurs in the oxygenated upper layer while the lower layer is deoxygenated. In most stratified lakes therefore, biota is very depauperated, except for bacteria, below a few 10 meters in depth. There is an exception, Lake Baikal, which is the deepest lake in the world with a maximum depth of 1620 m. The mechanism of mixing of the deep water zone is still not completely understood, but the entire water column is well oxygenated. Life for fish and invertebrates is therefore possible from surface to maximum depth, which is exceptional for freshwater systems. Lake Baikal is therefore unique among inland systems to the study of bathymetric segregation and includes some of the deepest occurring freshwater animals. Among fishes, the family Abyssocottidae contains 20 species distributed throughout the depths of the lake. Species of the genera Abyssocottus, Cotinella, and Neocottus are adapted to the deep water way of life in that they do not occur above 400 m, the size of the eyes is reduced, and they are physiologically adapted to resisting high pressures (Sidleva, 1994). The discovery of a deep sea hydrothermal fauna in the 1980s was a surprise for marine biologists. Similarly, hydrothermal vents have been discovered in Lake Baikal, at a depth of 440 m on the sediment floor of Frolikha Bay (Crane et al., 1991), at the foot of an east-west trending fault. Photographs reveal that the center of the vent field is covered by a near-continuous bacterial mat. A white sponge encroits small cobbles at the periphery of the vent field. Coiled gastropods and whitish translucent amphipods are found among the sponges and on the sediment at the edge of the bacterial mat.

D. Relationship between Species Richness and Area in Lakes
Community ecologists used to compare isolated freshwater systems to biogeographic islands. The relationship of species number to area containing those species is a well-known empirical observation, and a power function is widely used to describe this pattern mathematically: \( S = aZ^c \), where \( S \) is the number of species, \( A \) the area, \( Z \) the slope of the regression line, and \( c \) a constant. It can also be expressed as \( \log S = \log a + Z \log A \).

The effect of lake size on species richness of invertebrates has been demonstrated. For crustacean zooplankton, species richness is also significantly correlated with lake surface area (Dodson, 1992). The species area curve for North American lakes is statistically different from and steeper than the corresponding European curve (slopes, respectively, 0.094 and 0.054). The log species richness is also correlated to log of the average photosynthetic flux per cubic meter and log number of lakes within 20 km of the target lake. For 66 North American lakes, the three variables can be combined in a multiple linear regression model, which explains 73% of the variation in log species richness (Dodson, 1992).

Species richness of aquatic birds also increases with lake size. In Swiss lakes, the species number increase steeply with lake size up to 50 km² and species richness depends more closely on lake area in fish eaters and diving ducks than in dabbling ducks (Suter, 1994). Actually lake size explains 70 to 85% of the variation of abundance and species richness in fish eaters and diving ducks but only 64% of species richness in dabbling ducks. In Florida lakes, bird species richness was also positively correlated to lake area and to total water column phosphorus concentration value (WCP) for each lake. The multilinear Log (species richness) = 1.12 + 0.56 Log (Lake area) + 0.12 Log (WCP) and accounts for 77% of the variance in species richness (Hoyer and Canfield, 1994).
The species diversity in a lake is a function of the diversity of habitats: the more ecological niches in the lakes, the more species may be expected. The lake's morphometry is basic to its structure: deep, steep-sided lakes do not offer as many biotopes than shallow, flat lakes. For the latter, most of the lake bottom may be colonized by plants and animals (the benthic flora and fauna), while in deep lakes, only a small part of the lake bottom is colonized. Generally speaking, deep lakes are dominated by planktonic organisms, which are floating or weakly swimming organisms, usually associated with suspended particles. In shallow lakes, benthic organisms are dominant and the heterogeneity of lake bottom, as well as the development of macrophytes, may increase the diversity of benthic species.

II. EVALUATING BIOLOGICAL DIVERSITY

Despite the efforts of taxonomists, a good estimation of the total number of species occurring in freshwater lakes and ponds does not exist. We shall provide here some recent findings about aquatic biodiversity.

A. Diversity of Plankton and Microbial Loop

Three major size classes are usually recognized in pelagic plankton: microplankton (20–200 μm), nanoplankton (2–20 μm), and picoplankton (0.2–2.0 μm). In the late 1970s, phototrophic picoplankton was discovered in great abundance in both marine and freshwater ecosystems. However, identifying picoplankton causes considerable taxonomic problems due to the very small sizes of these organisms. We do not know how many bacterial species exist in the world, because bacteria cannot be differentiated under the microscope; we do not even know the right order of magnitude. A new way of classification has been proposed, based on the sequences of ribosomal RNA that led to a phylogenetic classification of bacteria. It is becoming apparent that the genetic diversity among bacteria is much wider than that among the animals and plants. Most heterotrophic nanoplanckton are small (2–5 μm), colorless flagellated protists. They grow at about the same rate as bacteria and are capable of consuming the entire bacterial production. Meanwhile, they regenerate significant amounts of nutrients and serve as prey for micro- and mesoplankton.

The importance of bacteria and protozoa activities in the trophic structure of lacustrine food chain has been largely underestimated in the past. The major role played by microorganisms in controlling energy and nutrient fluxes is now better understood following the discovery of the microbial loop and its role as a source or a sink for carbon and energy flow to higher trophic levels in pelagic systems. We know now that these microorganisms can control major fluxes of energy and nutrients. In some cases, 30% of the photosynthetic production does not pass directly to higher trophic level but is diverted into a microbial loop where nutrients are rapidly remineralized and fed back to the dissolved inorganic pools.

B. Diversity in Freshwater Sediments

About 175,000 species of organisms associated with freshwater sediments have been described, but the true number is much higher than this (Palmer et al., 1997). The number of species in most taxa can scarcely be estimated and global estimate of microbial diversity remains controversial. For example, some specialists estimate that there are hundreds of thousands of aquatic nematodes and only a small percent of these have been described. Rotifer species diversity is also poorly known for freshwater sediments, but it is estimated that there are thousands of undescribed species.

Most freshwater sediment species are small and concentrated in the upper sediment layers. Availability of light limits the development of plants and photosynthetic bacteria, which are therefore scarce or absent in most sediments. Moreover, oxygen level may influence species richness and the number of species is low in anoxic waters (see Table I).

C. Diversity in Fish

Presently, 25,000 fish species have been described. Some 10,000 species are found only in freshwaters, a large proportion of which occurs in lakes and ponds. The freshwaters are therefore disproportionately rich in species of fishes on the basis of area when compared to oceans. That could be viewed as the result of the patchy nature of inland waters and the resulting high endemicty of the biota. Fish live in almost every conceivable type of aquatic habitat. They exhibit enormous diversity in size, shape, and biology. Other vertebrates species occur in freshwaters: a few mammals, several reptiles, and many birds and amphibians. There is no quantitative evaluation of the number of vertebrates whose life cycles include lakes or ponds, but it is far from negligible (see Table II).
### TABLE I

**Species Richness of the Freshwater Sediment Biota for Many Habitat Types**

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Number of species described</th>
<th>Probable number of species</th>
<th>Range of local species richness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td>&gt;10,000</td>
<td>Unknown</td>
<td>&gt;1,000</td>
</tr>
<tr>
<td>Algae</td>
<td>14,000</td>
<td>20,000</td>
<td>0–100</td>
</tr>
<tr>
<td>Fungi</td>
<td>600</td>
<td>1,000–10,000</td>
<td>0–300</td>
</tr>
<tr>
<td>Protista</td>
<td>&lt;10,000</td>
<td>10–100,000</td>
<td>20–900</td>
</tr>
<tr>
<td>Plants</td>
<td>1,000</td>
<td>Unknown</td>
<td>0–100</td>
</tr>
<tr>
<td>Invertebrates</td>
<td>70,000</td>
<td>&gt;100,000</td>
<td>10–1,000</td>
</tr>
<tr>
<td>Aschelminthes</td>
<td>4,000</td>
<td>&gt;10,000</td>
<td>5–900</td>
</tr>
<tr>
<td>Annelida</td>
<td>1,000</td>
<td>&gt;1,500</td>
<td>2–80</td>
</tr>
<tr>
<td>Mollusca</td>
<td>4,000</td>
<td>5,000</td>
<td>0–50</td>
</tr>
<tr>
<td>Acari</td>
<td>5,000</td>
<td>&gt;7,500</td>
<td>0–100</td>
</tr>
<tr>
<td>Copepoda</td>
<td>8,000</td>
<td>&gt;10,000</td>
<td>5–300</td>
</tr>
<tr>
<td>Insecta</td>
<td>45,000</td>
<td>&gt;50,000</td>
<td>0–500</td>
</tr>
<tr>
<td>Others</td>
<td>1,400</td>
<td>&gt;2,000</td>
<td>0–100</td>
</tr>
</tbody>
</table>

Numbers are rough estimates and derived from many sources. Collected by Palmer et al. (1997)

### TABLE II

**Number of Fish Species Recorded from Several Lakes Connected to Rivers Systems**

<table>
<thead>
<tr>
<th>Lake</th>
<th>Latitude</th>
<th>Area</th>
<th>Number of fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chad</td>
<td>13°N</td>
<td>10–20,000</td>
<td>137</td>
</tr>
<tr>
<td>Turkana</td>
<td>3°N</td>
<td>6,750</td>
<td>51</td>
</tr>
<tr>
<td>Chilwa</td>
<td>15°S</td>
<td>875</td>
<td>31</td>
</tr>
<tr>
<td>Ngami</td>
<td>20°S</td>
<td>130</td>
<td>48</td>
</tr>
<tr>
<td>George</td>
<td>0°</td>
<td>270</td>
<td>30</td>
</tr>
<tr>
<td>Huron</td>
<td>44°N</td>
<td>39,600</td>
<td>99</td>
</tr>
<tr>
<td>Erie</td>
<td>42°N</td>
<td>25,700</td>
<td>113</td>
</tr>
<tr>
<td>Michigan</td>
<td>44°N</td>
<td>38,000</td>
<td>114</td>
</tr>
<tr>
<td>Superior</td>
<td>47°N</td>
<td>82,400</td>
<td>67</td>
</tr>
<tr>
<td>Great Bear</td>
<td>66°N</td>
<td>31,150</td>
<td>12</td>
</tr>
<tr>
<td>Great Slave</td>
<td>61°N</td>
<td>27,200</td>
<td>26</td>
</tr>
<tr>
<td>Big Trout</td>
<td>34°N</td>
<td>616</td>
<td>24</td>
</tr>
<tr>
<td>Chapala</td>
<td>20°N</td>
<td>1,080</td>
<td>14</td>
</tr>
<tr>
<td>Nicaragua</td>
<td>11°N</td>
<td>8,200</td>
<td>40</td>
</tr>
<tr>
<td>Maggare</td>
<td>Europe</td>
<td>46°N</td>
<td>676</td>
</tr>
<tr>
<td>Windermere</td>
<td>Europe</td>
<td>34°N</td>
<td>13</td>
</tr>
<tr>
<td>Ladoga</td>
<td>Europe</td>
<td>61</td>
<td>48</td>
</tr>
<tr>
<td>Aral sea</td>
<td>Europe</td>
<td>45°N</td>
<td>64,300</td>
</tr>
</tbody>
</table>

See also Table III for ancient lakes.
III. BIOLOGICAL DIVERSITY AND ECOSYSTEM FUNCTIONING

Energy and nutrients in an ecosystem are transferred through successive trophic levels. Photosynthesis provides the basic food for herbivorous animals, which are eaten by the carnivores. Therefore, knowledge of the role of individual species and their relationships in aquatic systems is critical to understanding the functioning of the system as a whole. Limnologists pointed out several key issues to the study of the relationships between species diversity and lake functioning.

A. Food Webs

Food webs are diagrams depicting which species in a community interact in feeding and describing which kinds of organisms in a community eat which other kind. Food webs are thus caricatures of nature, but they give a picture of the processes at work in ecosystems. Connectivity food webs are describing pathways along which feeding interactions occur. These interactions change at least seasonally and not all interactions are equally strong. The interaction web emphasizes connections that appear to have a large effect on the dynamics of the food web structure and function. Food webs occupy a central position in community ecology. Many important interactions (e.g., competition, predation) cannot be isolated from a food web context because the outcome of these interactions can be modified directly and indirectly by other members of the web.

For a long time food webs served principally as heuristic devices, useful in depicting complex ecosystems as diagrams composed of many interactive parts and enhancing our understanding of pathways of energy and material transfer in aquatic ecosystems. However, the recent surge of interest in food webs seems related to the question of the functional role of biodiversity (discussed later).

Few if any of the aquatic food webs are unimpacted by humans both at a local and a global scale. For example, fisheries food webs are complex, involving multiple trophic levels at several spatial and temporal scales. Fish species offer a wide range of body sizes and feeding habits, and thus have a variety of food web roles and interactions with other species. Exploitation of fishes may result in major changes in food webs. However, the consequences of species removal through fisheries are an almost unexplored field of research in most freshwater systems.

B. The Top-Down Control

In the classical limnological approach, it was usual to consider freshwater ecosystems as operating in a physical-chemical milieu that, largely through nutrient availability, conditions the food chain from primary producers to top predators. In this "bottom-up" control, competition between primary producers for limiting nutrients determines the state of higher trophic levels. A reverse viewpoint, the "top-down control" slowly became prominent. It argues that the effects of top predators cascade down through the trophic chain and are responsible for controlling the state of the entire ecosystem. The predators, near or at the top of the trophic pyramid, may be fishes but also may be birds, mammals, and so on, as well as invertebrates. Through grazing, for instance, fish have direct effects on the composition and abundance of phytoplankton, periphyton, and macrophytes, as well as on the dynamics of plankton and benthi communities. Size-selective predation by fish may not only play a major role in the population dynamics of prey species, but also result in shifts in the relative abundance of species.

Predation is now considered to be a major driving force in shaping zooplankton communities. A great number of papers emphasized the size-related alterations in zooplankton communities as a consequence of planktivorous fishes, which select the largest available prey and may rapidly reduce the density of large zooplankters, resulting in a shift of the prey community to small species, predominantly rotifers and small cladocerans. Extinction of large zooplankton has been documented in several habitats, usually following the introduction of new species of planktivorous fish.

Trophic cascades from fishes to water quality in lakes are among the clearest examples of feedbacks from population to ecosystem processes (Carpenter and Kitchell, 1993). A shift in the species composition and size distribution of the fish assemblages alters the community composition and size distribution of the herbivorous zooplankton. The impact of herbivory on phytoplankton depends on the relative abundance of certain herbivores with wide diets, high grazing rates, and rapid population growth rates. Population of these keystone herbivores is sensitive to fish predation. In addition, the size distribution of fishes and zooplankton and their migratory behavior largely determine the rate and spatial pattern of nutrient recycling in pelagic ecosystems. In whole-lake experiments, manipulations of fish community structure have caused significant changes in primary production, algal biomass, nutrient recycling, and sedimentation rates.
C. Relationships between Biodiversity and Ecosystem Stability
A major concern for limnologists is to predict response to stress. For a long time, the so-called conventional wisdom in ecology was that increased complexity within a community leads to increased stability. Complexity is used here to mean more species, more interactions between them, and more pathways. The basic assumption is that if the number of pathways increases, any blockage at one point of the network would be compensated for by the opening of another pathway. However, until now this conventional wisdom has not received much support from field or experimental work. Therefore some basic questions remain open and are of particular concern for freshwater lakes:

• How is system stability and resistance affected by species diversity, and to what extent could the integrity and sustainability of ecosystems be maintained in spite of species deletions resulting from degradation of environmental conditions?
• Are rare species an insurance of ecosystem stability? Do these rare species play a role in ecosystem functioning? One hypothesis is that the most stable ecosystems in terms of key functions are those richest in species. However, well documented studies of rare species substituting for declining common species in the maintenance of key freshwater ecosystem functions following disturbance are scant.

D. Role of Intra- and Interspecies Communication Systems in Ecosystem Dynamics
The structure of the biota is determined by complex interactions between individual organisms acting at different trophic levels. It is now becoming clear that besides energy transfer from one trophic level to another, there is an exchange of information between trophic levels through infochemicals. Moreover, sexual pheromones, but also sounds, electric signals, and visual communication play a role in shaping the structure of the biotic community. This diverse communication network has biological consequences and may modify the behavior of aquatic animals, such as migrations. This is a fairly new field of research.

E. Biological Productivity and Biodiversity: The Case of Eutrophication
The biological structure and internal biological control mechanisms of freshwater lakes are highly affected by lake water nutrient level and by the extent of nutrual loading. Limnologists distinguish oligotrophic lakes, which are generally deep with steep slope and are characterized by low nutrient levels and clear water. The biomass at all trophic levels is small. On the opposite, eutrophic lakes are often shallow with gradually sloping edges. The most characteristic features are high nutrients levels, the abundance of plankton, and low water clarity.

One concept of lake succession considers that lakes pass through different trophic states, beginning with low fertility or oligotrophy, gradually moving to a moderately productive or mesotrophic state to reach finally the eutrophic stage. This succession may happen in undisturbed lakes. However, eutrophication (sometimes called cultural eutrophication) is now widespread as a result of human activities. Eutrophication may be defined as the process of enrichment of a water body due to an increase in nutrient loading. The most important nutrients causing eutrophication are phosphates, nitrates, and ammonia. All these chemicals are abundant in waters released from sewage treatment works and from surface and groundwater runoffs in intensive agriculture areas.

The most obvious consequence of eutrophication is the increased aquatic plants and phytoplankton growth, an overall increase in biomass, and a shift in species composition of the lake. For example, at low P-concentrations, north European shallow freshwater lakes are usually in a clearwater stage; submerged macrophytes are abundant, potential piscivores are present in large numbers, and predation pressure on zooplankton is consequently low. At some higher P-concentrations, there is a shift to a turbid stage: submerged macrophytes disappear and the fish stock changes. The fish biomass rises and there is a shift from a system dominated by pike (Esox lucius) and perch (Perca fluviatilis) to one exclusively dominated by planktivorous-benthivorous fish, mainly bream (Abramis brama) and roach (Rutilus rutilus).

IV. THE CASE OF ANCIENT LAKES SPECIES FLOCKS
About a dozen lakes in the world are up to three orders of magnitude older than most others (Table III) (Mar-
tens, 1997). Such lakes have exceptionally high faunal diversity and levels of endemicity.

An important characteristic of ancient lakes biodiversity is the existence of "species-flocks." An aggregate of several species should be identified as a flock only if its members are endemic to the geographically circumscribed area under consideration and are each others' closest living relatives. Briefly speaking, a species flock has to be monophyletic. At present, different rich species flocks for fish and invertebrates have been identified in various ancient lakes, which are therefore exceptional natural sites for the study of speciation patterns.

The processes accounting for these radiations are a matter of debate, but there is more and more evidence that sympatric speciation may occur in isolated water bodies. These species flocks are sometimes considered to be a world heritage that is endangered and has to be preserved from destruction by human activities (Coulter et al., 1986; Nagelkerke et al., 1995) (see Table III).

### A. Fish Species Flocks

The most striking feature of the Great East African Lakes (Victoria, Tanganyika, Malawi) is that each has its own highly endemic lacustrine Cichlid fauna. In Lake Victoria, according to our present knowledge, there is a cichlid species flock of more than 500 endemic haplochromine species. The true species number is almost certainly even higher (Seehausen, 1996). The age of this flock was estimated at 200,000 years, but it is most likely that Lake Victoria had entirely dried up as recently as 12,400 years ago, so that most of the endemic cichlid flock would have evolved during the past 12,400 years (Johnson et al., 1996). In Lake Malawi, the diverse cichlid fauna of this lake could also total much more than 500 species. Species flocks are also reported for the clarid catfish Dinopterus (10 species).

The Lake Tanganyika cichlids are slightly less diverse. However, morphological and electrophoretic data both suggest that several lineages of cichlids from Lake Tanganyika are much older than the Lakes Victoria and Malawi lineages and can be traced back to at least seven distinct ancestral lineages. Species flocks also occur in noncichlid families: seven Mastacembelid species, six species of the bagrid Chrysichthys, seven species of Synodontis, and four species of the Centropomid Lates (De Vos and Snoeks, 1994).

Rates of speciation in cichlids can be astonishingly fast. That has been known since the discovery of five endemic species of cichlids in Lake Nabugabo, a small lake less than 4000 years old and separated by a sandbar from Lake Victoria. Still faster speciation rates were suggested by the finding that the southern end of Lake Malawi was dry only two centuries ago, while it is now inhabited by numerous endemic species and "color morphs" that are only found there and are believed to have originated during the past 200 years.

In South America, the native fish fauna of Lake Titicaca includes the genera Trichomycterus and Oreotus, both endemic to the Andean Altiplano. Twenty-four Oreotus species are presently recognized in Lake Titicaca (Lauzanne, 1992). However, it is probably not a monophyletic group but rather an assemblage that includes, in part, several species flocks.

### TABLE III

<table>
<thead>
<tr>
<th>Lake</th>
<th>Age (My)</th>
<th>Max. depth (m)</th>
<th>Area (km²)</th>
<th>Number of animal species</th>
<th>Number of endemic fish species</th>
<th>Number of fish endemic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baikal</td>
<td>35–30</td>
<td>1,700</td>
<td>31,500</td>
<td>1,825</td>
<td>982</td>
<td>56</td>
</tr>
<tr>
<td>Tanganyka</td>
<td>6–12</td>
<td>1,470</td>
<td>32,600</td>
<td>1,470</td>
<td>632</td>
<td>330</td>
</tr>
<tr>
<td>Malawi</td>
<td>4.5–8.6</td>
<td>785</td>
<td>30,800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Victoria</td>
<td>0.6?</td>
<td>70</td>
<td>70,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Titicaca</td>
<td>3</td>
<td>284</td>
<td>8,448</td>
<td>533</td>
<td>61</td>
<td>29</td>
</tr>
<tr>
<td>Ohrid</td>
<td>2–3</td>
<td>295</td>
<td>348</td>
<td>29</td>
<td>94</td>
<td>17</td>
</tr>
<tr>
<td>Buja</td>
<td>4</td>
<td>104</td>
<td>674</td>
<td>600</td>
<td>54</td>
<td>57</td>
</tr>
</tbody>
</table>

In part from Martens, 1997.
LAKE AND POND ECOSYSTEMS

Lake Baikal (East Siberia) hosts a very diverse fauna, with some 2500 described species (most of which are endemic), which might constitute 50% of the total amount. At present, Lake Baikal comprises 56 species and subspecies of fish, which belong to 14 families (Sudeleva, 1994), a group of six species and subspecies, belonging to three families (Thymallidae, Coregonidae, and Acipenseridae), which are relatively endemic; another group of nonendemic fauna includes 21 species and subspecies, which belong to Cyprinidae, Percidae, Cobitidae, Esocidae, Gadidae, Salmonidae, Siluridae, and Eleotridae.

In Asia, Lake Lanao was formed by a lava flow that dammed the streams flowing southwest. The cyprinid fauna presents a widely acknowledged example of adaptive radiation. Of the 23 cyprinid species presently known from Mindanao Island, 15 are reported from Lake Lanao (Kottelat and Whitten, 1996). Unfortunately overexploitation and exotic introductions have decimated the fauna, so that now only few endemic cyprinids are still present.

Lake Biwa is the largest and oldest lake in Japan. The deep basin as seen presently is supposed to have been formed 300,000 years ago. Most endemic fishes exist also since that time (Kawanabe, 1996). Some 500 plant and 600 animal species have been recorded. At present there are 71 species and subspecies if freshwater fishes found in Lake Biwa and its tributaries. There are 13 endemic species and subspecies of fish in Lake Biwa.

The only more or less pristine species flocks left in Asia are to be found in the lakes the Malili River drainage of the Sulawesi Island (Celebes) in Indonesia (Kottelat & Whitten, 1996). Malili lakes (lakes Towuti, Matano, Mahalona, Wawontoa, and Masapi) constitute a system of lakes partially isolated from each other and completely isolated from other freshwaters. As a result, most of the animal species known from the lakes are endemic.

B. Invertebrates

Mollusks focus our attention on parts of the world that seem to be hot spots of endemicity, where the resident clades are remarkably more diverse than in other similar environments. Long-lived lakes are prime examples of these evolutionary theaters. Particular clades, such as the hydrobioid and ceratioidean prosobranchs and the planorbid pulmonates, show repeated patterns of diversification in both extant and fossil long-lived lakes, revealing the common patterns that make them prone to speciate (Michel, 1994). The process of diversification is tied to intrinsic characters shared by many of these clades: reproductive and dispersal strategies (brooders and poor dispersers), genetic structure (highly constrained genetic systems), morphology (often relatively thick and ornamented shells), substrate specificity (hard bottom stenotopy), and physiology (depth tolerance). The most notable examples of these evolutionary theaters are extant lakes Tanganyika, Baikal, Ohrid, Titicaca, and fossil lakes Idaho, Biwa, and Turkana. However, examples of gastropod radiations are also found in river systems (see Table IV).

Among Crustacea, gammarids have also undergone an enormous evolutionary radiation in Lake Baikal, with a total of 239 species, 98% of which are endemic. There are also several species flocks of Ostracods reported from ancient lakes (Marten et al., 1994).

V. MAJOR THREATS TO BIODIVERSITY IN LAKES

The concentration of people around freshwater systems has resulted in a much greater degree of degradation to these systems than most open marine or even terrestrial systems.

A. Competition for Water

Competition for water may result in the total or partial desiccation of lakes and ponds through various diversion and impoundment of tributaries. Water is withdrawn most often from aquatic systems for irrigation, flood control, and urban and industrial consumption. A spectacular example is provided with the Aral Sea, a large saline lake in the terminus of an extensive inland drainage basin in south-central Asia. Water diversion for irrigation purposes of most of the waters in inflowing rivers of the Aral sea, as well as poor agricultural practices, resulted in a marked fall of water level (c. 15 m) and an increase in salinity (from c. 10 to 30 g/l) since the 1960s. This changes have resulted in the degradation of the natural environment. Fish have virtually disappeared from the lake and the diversity of associated bird and wildlife communities has decreased. Many invertebrates also disappeared (Williams and Aladin, 1991).

B. Habitat Alteration

Siltation from erosion of the lake basin has direct adverse effects on fish by covering spawning sites, destroying benthic food sources, and reducing water clarity to
visual feeding animals. However, the increase in turbidity may have also indirect effects on biodiversity in lakes. Seehausen et al. (1997) provided evidence that increasing turbidity (as the consequence of deforestation and agricultural practices) by curbing the impact of sexual selection on sexual isolation is responsible for the decline in cichlid diversity in Lake Victoria. Actually, mate choice in these cichlids is determined on the basis of coloration, and strong assortative mating can quickly lead to sexual isolation of color morphs, which is increasing and probably started in the 1920s. By constraining color vision, turbidity interferes with mate choice (Seehausen et al., 1997). The reduced effectiveness of signals causes relaxation of sexual selection for color, with consequent loss of male nuptial coloration and erosion of species diversity due to a breakdown of reproductive barriers. Dull fish coloration, few color morphs, and low species diversity are found in areas that have become turbid as a result of recent eutrophication. This is proof that human activities that increase turbidity destroy the mechanism of diversification and the maintenance of diversity.

C. Species Introductions
The introduction of alien fish into inland waters has occurred all around the world. The main goals of deliberate introductions by fishery officers were initially to improve sport fisheries and aquaculture, or to develop biological control of aquatic diseases, insects, and plants, or else to fill supposed "vacant niches" and improve wild stocks in old or newly created impoundments.

The introduction of alien species has been considered as the main causes of extinction of endemic species flocks in several ancient lakes. In Lake Lanao, the introduction of the white goby (Glossogobius giurus) in the early 1960s resulted in the elimination of numerous species of endemic cyprinid fish. In Lake Titicaca the rainbow trout Salmo gairdneri was accused of seriously threatening the endemic Orestias fauna and for having been responsible for the disappearance of species such as Orestias cuvieri. In Lake Biwa, the recent increases in numbers of the exotic bluegill Lepomis macrochirus, black-bass Micropterus salmoides, and Channa maculata, have been mirrored by serious declines in the native species Onchorhynchus rhodurus (an endemic), Hemigrammocypris rasborella, and Hymenophysa curta. Much has been said about the impact of the introduction of the Nile perch on the hundreds endemic haplochromines of Lake Victoria (Levéque, 1997). In the early 1980s this impact was considered an ecological and conservation disaster (Coulter et al., 1986). However, it was later recognized that predation by Lates may not be solely responsible for the depletion of haplochromine stocks, and that the haplochromine stock was already affected by fisheries before the establishment of Lates, particularly by unregulated fishing or by trawling techniques introduced in the Tanzanian part of the lake. Lake Victoria is now invaded by water hyacinth, and the remaining fish fauna is therefore more and more threatened.

Transport through ballast water is probably one of the most important pathways for alien species invasions in several places, including the North American Great Lakes (Mills et al., 1993). That is the case for the zebra mussel introduced into the Great Lakes, apparently in 1983 or 1986, which spread dramatically throughout the waterways of both Canada and United States expansion with serious economical and ecological consequences. The recent finding of individual mitten crabs (Eriocheir sinensis), a European flounder (Platichthys

<table>
<thead>
<tr>
<th>Lakes</th>
<th>Baikal</th>
<th>Ohrid</th>
<th>Tangan</th>
<th>Titicaca</th>
<th>Biwa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mussels</td>
<td>2</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>% end.</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sphaeridae</td>
<td>10</td>
<td>9</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>% end.</td>
<td>30</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prosobranchs</td>
<td>72</td>
<td>47</td>
<td>32</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>% end.</td>
<td>93</td>
<td>91</td>
<td>84</td>
<td>95</td>
<td>72</td>
</tr>
<tr>
<td>Pulmonates</td>
<td>61</td>
<td>23</td>
<td>16</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>% end.</td>
<td>77</td>
<td>48</td>
<td>6</td>
<td>33</td>
<td>62</td>
</tr>
<tr>
<td>Total species</td>
<td>145</td>
<td>82</td>
<td>81</td>
<td>19</td>
<td>27</td>
</tr>
</tbody>
</table>
flesus), and the establishment of the alien gastropod Potamopyrgus antipodarum in Lake Ontario in 1995 demonstrate that the process of invasion is still going on at a fast rate.

One of the major problems in freshwater species introductions is their irreversibility, at least on scale of a human’s lifetime. Once introduced and established, it is impossible, given current technology, to eradicate a fish, a mollusk, or a plant species from a large natural water body. As a consequence, we are likely to see a continued reduction in native aquatic biodiversity and an increased homogenization of the world’s freshwater biotas.

D. Fisheries Practices

One of the major threats to the unique species flocks of ancient lakes are the fishing practices and particularly overfishing and introduction of new fishing practices. According to Coulter et al. (1986), the collective experience in recent years on the African Great Lakes seems to show that large-scale mechanized fishing is incompatible with the continued existence of the highly diverse cichlid communities. Cichlids appear especially vulnerable to unselective fishing because of their particular reproductive characteristics. The structure of endemic cichlid fish communities in the African Great Lakes can change dramatically within a few years when trawlers and other such fishing gear are used. Actually, a number of authors have recorded the effects of overfishing in Lake Victoria, from the decline of some species to the virtual disappearance of others, and the history of the fishery was briefly reviewed by Witte et al. (1992).

It has also been suggested that parks should be developed (Coulter et al., 1986) and that fishing should be rendered impossible in certain areas by placing obstructions on the bottom that would snarl trawls. Lake Malawi National Park will very probably afford protection to widespread species, but no data are at present available to confirm this hypothesis. It is unknown yet whether these reserves can adequately preserve the integrity of populations, but that is probably only possible for stenotopic populations whose distribution coincides with the park area. The size of the reserves, the intensity of fishing in nearby areas, and the possible influence of pollution or introduced alien species should also be taken into account.

E. Pollution

Pollution can affect aquatic biota through direct mortality at any life stage or by sublethal effects influencing predation, foraging, and reproduction. The eutrophication of Lake Victoria during the past 25 years is quite well documented. Enhanced quantities of nutrients appear to have been entering this lake for many years, both through rivers and from aerosols as a result of human activities in its watershed. The eutrophication could lead to increased oxygen demand in the lake’s deep water and thus decrease the hypolimnetic volume habitable by fish during seasonal stratification. This phenomenon is partly responsible for the threatening or disappearance of cichlid species belonging to the haplochromine flock.

The release of sulfur and nitrous oxides from the burning of fossil fuels may be transported great distances before being transformed chemically into sulfuric and nitric acids and deposited as rain, snow, or dust. When acid rains occur over areas where waters are poorly buffered, the chemistry and biology of freshwaters can be changed dramatically. Many softwater lakes have been acidified both in North America and Europe, but evidence has accumulated for its occurrence in China, the former Soviet Union, and South America. Monitoring studies indicate a general impoverishment of species numbers in lakes as they become more acidic. Many lakes in the northeastern United States have lost 30% or more of the species in some taxonomic groups. In many northern European countries, acidification strongly modified the fish composition and abundance in lakes.

See Also the Following Articles

ENDANGERED FRESHWATER INVERTEBRATES • EUTROPHICATION AND OLFOTROPHICATION • FISH BIODIVERSITY • FRESHWATER ECOSYSTEMS • FRESHWATER ECOSYSTEMS, HUMAN IMPACT ON • INVERTEBRATES, FRESHWATER, OVERVIEW • RESOURCE EXPLOITATION, FISHERIES

Bibliography


644___LAKE AND POND ECOSYSTEMS___
I. Introduction

While biodiversity is usually considered at the species level, maintenance of biodiversity requires management at higher levels of organization, particularly at the landscape scale. It is difficult to manage for each threatened species individually. Alternatively, management can focus on the ecosystems that contain these species, and on the landscapes in which ecosystems are found. The relatively new discipline of landscape ecology provides insight into both landscape diversity and species diversity, and suggests a theoretical and practical basis for conservation planning.

There are three basic characteristics of landscapes that affect their diversity: structure, function, and dynamics. Structure is the most well-understood element of landscapes. It is also the most obvious—nearly any
aerial view will show a mixture of different landforms, habitats, or vegetation types. The patch is the basic unit of landscape structure. The characteristics of patches and the spatial relationships among patches are important components of landscapes. The distributions of energy, materials, and species among patches differing in size, shape, abundance, and configuration are particularly important to patterns in diversity at the landscape scale. The other two elements of landscapes go beyond a description of spatial heterogeneity. Function is concerned with interactions among the spatial elements of a landscape, including flows of energy, materials, and species among patches. Landscape dynamics includes characteristics of both structure and function in order to examine changes in pattern and process over time. The conservation and management of biodiversity require understanding of all three elements, including the effects of human activities on the system. This article discusses each element in turn, and also considers the underlying determinants of landscape structure, including environmental heterogeneity and disturbance patterns. We then discuss classical and current issues in biodiversity management, and conclude with a case study of landscape diversity at the Sevilleta National Wildlife Refuge Long-Term Ecological Research site in central New Mexico, United States.

It is essential to keep the concept of “scale” in mind when considering landscape diversity. Spatial scale has two elements: grain and extent. Grain is the minimum resolution sampled, usually the cell size or quadrat size for ecological studies. The extent is the domain of the study, which is typically the size of the study area. Ecological processes often have characteristic spatial and temporal scales. This means that the grain and extent of sampling in both space and time may strongly affect the results of a study. For example, as quadrat size (grain) increases, species richness may increase, yet diversity of patch types within a landscape may decrease since fewer large quadrats can be found within a given area (Fig. 1a). As landscape size (extent) increases, more species and more patches of a constant size may be found that would increase both species and landscape diversity (Fig. 1b).

II. DESCRIPTION OF LANDSCAPE STRUCTURE

Landscape structure can be most easily described at two hierarchical spatial levels, both of which are relevant to landscape diversity as well as to species diversity. At the lower level, the focus is on the attributes of individual patches, particularly size and shape. Description at the higher spatial level is concerned with the composition and pattern of the entire landscape and its mosaic of patches. The ability to quantify landscape structure at both levels allows the comparison of different landscapes. More importantly, interactions between landscape structure and function have implications for both species and landscape diversity.

A. Patch Description

A patch is a relatively homogeneous nonlinear area that differs from its surroundings. The definition and identification of individual patches and their boundaries are important steps in characterizing the structure of a landscape. In some systems, boundaries may be easily identified, such as between patches of agricultural field and adjacent woodland in human-dominated systems. In many cases, however, the boundary is not so
clear, and patches are more difficult to delineate. Most methods of patch identification combine qualitative and quantitative approaches. A subjective determination of how different two areas must be in order for them to be considered separate patches is often needed. A number of quantitative techniques have been developed to group similar cells into homogeneous patches or to identify repeating patterns across a landscape (Turner and Gardner, 1990). Approaches such as blocking techniques, spectral analysis, and nearest-neighbor analysis are commonly used. Other techniques rely on the detection of edges or boundaries rather than identifying patches directly. These methods include moving window analysis and image analysis to characterize landscapes with sharp transitions.

Patch identification provides an excellent example of the importance of the spatial scale of the observer. From inside a forest, clumps of trees and grass-dominated openings appear to be separate patches with different vegetation and resource availability. From an aerial view, the entire forest appears to be a single patch. This illustrates the importance of the selection of spatial scale based on study objectives prior to the determination of patches and their edges.

Once the patches in a landscape have been identified, there are many ways to describe and quantify them (Riitters et al., 1995). Only patch size and shape will be discussed here, since the relevance of these two attributes for species diversity is the most well understood. The relationship between patch size and species richness goes beyond the familiar species–area curve (Fig. 2). Although the number of species present in a patch tends to increase with patch size up to a certain limit, the kinds of species found also tend to vary with size. Two general types of species can be distinguished. Interior species are found primarily in the interior of large patches. These species often have very specific habitat requirements and are relatively rare. Migratory songbirds that are particularly sensitive to patch size and adversely affected by habitat fragmentation are interior species. By contrast, edge species are found near the edges of large patches and throughout small patches that consist mostly of edge habitats. Edge species are commonly occurring generalists that can use various habitat types, and are often introduced species. Because small patches consist mostly of edge with little interior area, they often have the highest species densities, but contain few or no rare species. Large patches, on the other hand, are mostly interior area with lower species densities per unit area, but they contain more rare species and a higher total number of species.

In an important study of tropical deforestation in the Amazon rain forest, species in patches of various sizes were compared to evaluate the importance of patch size to species number (Lovejoy et al., 1984, 1986). Large patches were richest in species and small patches were found to contain only edge conditions. Patch size had important effects on different species, including trees, insects, birds, and mammals, that were noticeable in a short time. This study is one of the few where patch size was experimentally manipulated to allow comparison with pretreatment conditions as well as control patches.

A simple measure of patch shape is the perimeter : area ratio. This measure is often standardized so that the most compact possible form, either a square or circle, is equal to 1. More complex shapes have increasingly higher numbers. Another common index of shape complexity is the fractal dimension, which is also derived from the perimeter and area of a patch. The fractal dimension of a patch is between 1 and 2; a simple shape will have a lower fractal dimension than a more complex shape. Figure 3 illustrates the amount of interior area available in patches of different shapes. Both patches have an area of 25, but the perimeter of shape a is 20 while the perimeter of shape b is 32. Using a scaled perimeter : area ratio, a has a value of 1 and b has a value of 1.6. Assuming that interior area is at least 1 unit from any patch edge, a has an interior area of 9, but b has an interior area of only 2. Thus, a is more compact and less convoluted than b, where more of the area is closer to its edge and can interact with the area surrounding the patch. This suggests that the overall flow of species and resources between b and its surroundings is higher than that between a and its sur-

---

**FIGURE 2** Species–area curve showing the increase in species number with increasing patch size tending toward a regional limit.
roundings. It is also expected that a would have higher richness of interior species than b, which would have higher richness of edge species.

B. Landscape Description

At the landscape level, there are two basic components of structure: composition and pattern. Composition refers to the parts (i.e., patch types) that make up the landscape, and pattern refers to how these patches are arranged. Though these two components are conceptually different, in practice they are often related. For example, the pattern of agricultural fields on a landscape is likely to be different from the pattern of undisturbed woodland.

Landscape composition can be measured in ways analogous to measurements of species composition (Romme, 1982). The most straightforward approach is landscape richness or the number of different patch types in a landscape. Another approach includes the relative abundance or dominance of different patch types along with richness. These landscape indices were derived from information theory and are closely related to species diversity measures, such as the Shannon–Wiener and Simpson indices, which are used to describe alpha species diversity (Turner, 1989; Huston, 1994). Using one of these indices, a landscape containing many small patches of different types would have a higher diversity value than a landscape consisting of one large patch and several smaller patches, even if the total number of patches is the same for both landscapes. Landscape measures of richness and evenness were used in a study conducted in different patch types in Yellowstone National Park (Romme, 1982). Changes in landscape diversity through time were related to fire frequency, and were hypothesized to have important effects on species diversity as well as wildlife habitat (Romme and Knight, 1982).

Measurements of landscape diversity are analogous to common measurements of species diversity (Whittaker, 1960, 1972). Alpha species diversity is a measurement of species richness (number) and evenness (dominance or distribution) within a patch. Similarly, alpha landscape diversity is a measure of the number of patch types in a region (O’Neill et al., 1988). Large-scale species diversity is called gamma diversity. Gamma landscape diversity of ecosystems is sometimes called ecdiversity (Rowe, 1992; Lapin and Barnes, 1995). The third form of species diversity, beta diversity, describes species turnover along a gradient. Beta diversity has no analog at the landscape level, but is sometimes estimated as gamma diversity/alpha diversity, which gives an average regional beta diversity.

Because different patch types provide different habitats and species compositions, one might expect that the total number of species in a landscape would increase as landscape richness increases. This idea was supported by a study that compared plant species richness in Rhode Island Audubon refuges varying in terrain and soil properties (geomorphological measures) (Nichols et al., 1998). In a related study of one landscape, the diversity of trees and shrubs was higher on plots with the greatest geomorphological heterogeneity, indicating an important connection between landscape diversity and species diversity (Burnett et al., 1998). Although this simple relationship between landscape and species diversity is generally true, the interactions between landscape composition and species diversity are more complex, in part because of species preferences to edge or interior types of habitats. The species found in a diverse landscape with many small patches are mostly edge species. Interior species are found only in landscapes with large patches, even though these landscapes have a lower diversity. The total number of interior species increases with the number of large patches on a landscape, similar to the species–area relationship in Fig. 2. Thus, the type of species that increases with increasing landscape diversity depends on the change in the size and configuration of patches within the landscape.

Landscape pattern, or the spatial arrangement of patches, can be measured in a number of ways, some of which are extensions of the patch-level metrics already discussed. These measures focus on patch abundance without regard to location in the landscape. The distribution of patch sizes can be determined within a landscape and used as information in the management of

![Figure 3](https://example.com/figure3.png)  
**Figure 3** A simple shape and a more complex shape of equal area. The interior area is shaded.
habitat patches for species that are sensitive to patch size or spatial arrangement, such as the spotted owl in the Pacific Northwest. The effects of forest clear-cutting on changes in patch structure and implications for interior and edge species provide another example of the importance of these measures. Shape complexities and boundaries can also be scaled up from the patch to the landscape level using the fractal dimension and perimeter : area ratio (Milne, 1988).

The second type of measurement for landscape pattern explicitly considers the location of patches relative to each other and includes patch abundance as well. Dispersion indicates the tendency of patches of one type to be distributed either uniformly, randomly, or aggregated. Contagion describes the tendency of patches of two different types to be near each other. Connectedness can be quantified using nearest-neighbor probabilities that reflect the degree of fragmentation in the landscape. All three of these indices have implications for the flow of species and resources between patches of the same and different types, and thus have important effects on species diversity.

Two additional structural elements other than patches may be recognized in many landscapes. The second element is the matrix or the background landscape. The matrix is characterized by extensive cover, high connectivity, and major control over landscape dynamics. Forest patches contained within a matrix of subdivisions are functionally very different from forest patches surrounded by agricultural land. Corridors, strips that differ from the adjacent landscape on both sides, are the third element of landscapes. Corridors are usually linear and always highly connected; stream networks and roadways are common examples. Corridors may also connect larger patches of a similar type, such as a stream flowing between two lakes.

The patch–matrix–corridor model of landscape structure is illustrated in Fig. 4. Corridors may be particularly important for preserving species diversity by allowing movements of species across diverse landscapes. Corridors can also adversely affect species diversity by allowing nonnative or exotic species to invade and reduce the number of native species in an area. An example is the extensive spread of cheatgrass, an annual introduced to North America in shipments of grain from Asia and Europe in the 1880s (Mack, 1981). Movement of cheatgrass seed along railroad and cattle trail corridors in the early 1900s spread this grass throughout much of the northwestern United States, resulting in changes in species composition and dominance as well as losses of diversity.

III. CONTROLS ON LANDSCAPE DIVERSITY

Heterogeneity or diversity of landscape structure arises from a number of factors. Patches can be produced
through biotic or abiotic causes, including natural- or human-caused disturbance, fragmentation, regeneration, and persistent differences in environmental resources. Once a patch is formed, environmental conditions or interactions among organisms may change through time, leading to successional dynamics on the patch. A landscape consisting of patches in various successional stages is called a “shifting mosaic” (Bormann and Likens, 1979). The spatial pattern of patch formation and the changes within patches are collectively called “patch dynamics” (Pickett and White, 1985). The patch dynamic mosaic is part of the broader landscape transformation that includes changes in corridors and the matrix, as well as in the dynamics of species and ecosystem processes. These dynamics are discussed in Section V.

Biotic causes of patch generation include the local dispersal of seeds into a landscape, such as by an invasive weed, and the spatial segregation of populations or communities as a result of competition. Spatial structure can also be generated by differences between species in their dispersal abilities and rates of mortality. Naturally occurring and human-created disturbances are common causes of patch formation. A wide variety of natural disturbances are possible, including mud slides, avalanches, windstorms, ice storms, herbivore outbreaks, animal grazing, trampling, and digging, as well as fire. Mounds produced by badger digging activities in tallgrass prairie are an example of patch-producing disturbances that have important influences on patch structure as well as species diversity (Platt, 1975). Human activities, such as forest cutting, altered fire regimes, cultivation, urban development, introduction of pests, and strip mining for surface coal and minerals, also produce disturbance patches. Many landscapes are influenced by both natural- and human-caused disturbances, and distinguishing the separate effects on landscape diversity can be difficult. In a recent study, pollen and charcoal were collected from small lakes in Massachusetts in order to reconstruct long-term vegetation dynamics as related to disturbance history (Fuller et al., 1998). This reconstruction over the past 1000 years included the period of time prior to European settlement when the primary disturbances were fire and wind. Landscape patterns in forest composition following settlement by Europeans were largely influenced by clearing of forests for agricultural purposes and timber. These researchers found that the past history of disturbance as a result of settlement has persistent effects on current landscape patterns. Landscape fragmentation is closely related to disturbance. Many forms of disturbance effectively break up large patches into smaller pieces. Decreases in patch size, connections between patches, and total interior area as a result of fragmentation have important implications for species and landscape diversity. As landscapes become more fragmented, patch diversity increases with subsequent increases in edge species, exotic species, and generalists. Richness in interior species tends to decrease. Fragmentation of landscapes by human activities is considered a major threat to biodiversity worldwide (Saunders et al., 1991; Bierregaard et al., 1992). A major focus of the field of conservation biology is the design of nature reserves to maximize the likelihood of species existence and to minimize the loss of species to extinction. These processes are discussed in Section VI.

Another cause of patch formation is environmental heterogeneity, which refers to variation in soils, topography, and other landscape features. This variation in the physical environmental leads to heterogeneous or patchy spatial distribution of resources, including water, nutrients, and light. Plant species found in a resource patch can differ from species in other patches containing different levels of resources. The importance of spatial heterogeneity to species diversity has been well documented, and is most closely related to beta species diversity. The extension of these ideas to landscape diversity has occurred more recently, and several studies have linked measures of alpha and beta species diversity with landscape diversity (Romme, 1982; Lapin and Barnes, 1993). Large-scale gradients in landscape diversity can also be related to broad-scale patterns in the environment. For example, spatial variation in climate, topography, and soils was found to be strongly related to latitudinal gradients in richness of land cover types across the continental United States (Wickham et al., 1999).

IV. LANDSCAPE FUNCTION

Interactions among the spatial elements of a landscape are the major components of landscape function. These flows of energy, materials, and species among patches, or among patches, corridors, and the surrounding matrix, are at least as important to the maintenance of diversity as patch size and configuration. However, these flows have not been as well studied as landscape structure. An example of flows among different patch types is the dispersal of seeds from forest patches into clear-cuts, which has important effects on vegetation dynamics in these open areas. Boundaries or edges between patches or between patches and the mosaic often
V. LANDSCAPE DYNAMICS

Landscape structure and function can change for many reasons and in many ways. Changes can happen over very small or very large areas, and over short or long time spans. The gap caused by a single tree falling in the forest during a storm is small and temporary, while an entire forest may be leveled by a hurricane and take decades to centuries to recover.

Vulnerability or sensitivity to change varies from landscape to landscape. This vulnerability (or, conversely, stability) is traditionally divided into two components: resistance and resilience. Resistance is the ability of a patch or landscape to remain unaffected by a disturbance. A grassland is much more resistant to wind damage than a forest, since grasses can bend with the wind without breaking. Resilience is the ability of a patch or landscape to recover after a disturbance. Temperate forests recover after clearing much more quickly than tropical forests (which may never recover) owing to differences in soil depth and fertility.

Change in a patch or landscape can be caused by many factors. Some of these are intrinsic to the population being studied, including recruitment, growth, mortality, and spread or migration, which can lead to invasions or extinctions as well as changes in patch boundaries. Other causes are extrinsic to the ecosystem and are imposed by outside forces, such as climate change and disturbance events. Human transformations of the landscape include deforestation and reforestation, urbanization, corridor construction, and agricultural conversion. The effects of consumer, pathogens, and especially people can be considered either intrinsic or extrinsic depending on the particular point of view. The potential causes of change may be interrelated in complex ways. A drought may make a forest more vulnerable to pathogens, or a new clearing may increase the vulnerability of adjacent trees to windthrow.

Changes in landscape structure can have several spatial and temporal forms. Patches can shrink or expand, or be lost entirely. Successional dynamics on patches can lead to a shifting mosaic of patch types through time. Species interactions with other species and with their environment, as well as dispersal of new species into patches, are primary determinants of the regrowth of plants on these successional patches. Changes in patch size and shape can occur along edges, such as the clearing of forest to increase the size of a cultivated field (Fig. 5a). A new patch type may spread outward from a corridor (Fig. 5b). For example, housing developments often spread from the course of new roads. Alternatively, a patch type may spread out from a nucleus that could be a remnant of a previous vegetation type, or an introduction site for a new patch type (Fig. 5c). Some changes are nearly instantaneous and occur over very short periods of time, such as the effect of fire. Other changes occur slowly and take a longer period of time to develop, such as suburbanization and desertification.

Patch configuration on a landscape can also change. Patches can become perforated by other patch types,
and large patches can be fragmented into several smaller patches. Landscape fragmentation, particularly in the tropics, is having severe effects on species biodiversity. Some of the potential consequences of fragmentation include the loss of patch types and their characteristic species, decreased connectivity with its repercussions for species movements, and decreased interior area. The biggest consequence for species diversity is the associated loss of interior species and the increase of generalist or edge species.

Landscape-level dynamics are often studied with ecological models since the temporal scales of interest are often greater than the human life span, and experiments are difficult to perform at large spatial scales. There are four general classes of models that are used to predict landscape dynamics: transition probability models, individual-based models, ecosystem process models, and biogeographic models. Transition process models are useful when the factors causing landscape change are not represented mechanistically. For example, assume that a landscape with three patch types was sampled twice, before and after an event. A table can be constructed showing the percentage of each patch type that stayed the same or that was transformed into a different patch type over a single time step. Over a single time step, 60% of the forest land remained forested, 25% was converted to agricultural uses, and 15% was developed. These transition probabilities can be used to extrapolate into the future by individual time steps. Figure 6 shows the projected change over 25 time steps.

<table>
<thead>
<tr>
<th>Forest</th>
<th>Agriculture</th>
<th>Developed</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>75</td>
<td>15</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>98</td>
</tr>
</tbody>
</table>

* Each cell shows the percentage of the landscape area that changed from the patch type in that row to the patch type in that column over a single time step. For example, 25% of the original forest land was cleared for agriculture and 15% was developed. Figure 6 shows this projected change over 25 time steps.

Individual-based simulation models are useful when information is known about the mechanisms underlying changes in landscape structure. These models incorporate life-history traits of individuals and the mechanisms by which they interact with their environment in order to predict landscape-level dynamics. landscapes are simulated by linking plots together in a grid or transect. Plots are spatially interactive through processes such as seed dispersal. Spatially interactive individual-based models can represent a variety of environmental conditions, including differences in soil properties, climate, and disturbance regime (Coffin and Lauenroth, 1994). These models are most commonly used for evaluating changes in the diversity of groups of similar species (i.e., functional types) rather than species diversity.

A third class of models simulates ecological processes, including rates of nutrient cycling, water balance, and primary production. These models have been linked with geographic information systems (GIS) to simulate large regions. Effects of climate, soil texture, and management on soil organic carbon dynamics were simulated for the central grasslands of the United States (Burke et al., 1991). Across this large region, soil organic carbon increased with precipitation and decreased with temperature and percentage sand content. Biogeographic models are a fourth class of models that can be used to investigate vegetation responses to environmental heterogeneity. These models incorporate large-scale variations in climate and soils, as well as water
and energy constraints on plant growth, to simulate continental and global patterns in vegetation. Biogeographic models are most useful for simulating responses of plant functional types at large spatial scales, to either equilibrium or transient environmental conditions (Prentice et al., 1992; Nelson and Drapek, 1998). These models do not explicitly incorporate landscape-scale processes or flows among patches.

Although each of these types of models has traditionally been used independently, the linking of different models together is a recent development that has considerable potential for addressing issues related to landscape diversity. Because of important feedbacks between species and rates of ecosystem processes (Schulze and Mooney, 1994), linking a spatially interactive individual-based model with an ecosystem model has a large potential for simulating the dynamics of landscape structure and function as well as changes in species diversity. An important first step was illustrated by connecting a nonspatial individual-based model with a nutrient cycling model to examine the importance of soil heterogeneity to forest responses to global climate change (Pastor and Post, 1988). The incorporation of landscape-scale flows of water, carbon, and nutrients into a spatially interactive individual-based model is an important research area for predicting landscape diversity dynamics. Recent linkages between biogeochemical and biogeographic models are another area of potential application to landscape diversity, especially if the plant functional types become more resolved and landscape-scale processes, such as disturbance regime, are included.

VI. BIODIVERSITY PLANNING AT THE LANDSCAPE LEVEL

To preserve species diversity most effectively, management plans must preserve the habitats and landscape structures needed by the target species, rather than simply preserving the species in isolation from the larger, potentially changing environment. Management practices aimed directly at a particular species run the risk of losing ecosystem functions that might actually be crucial for the target species, but that were unknown
when the management plan was created. Furthermore, maximizing benefits for one species may threaten others. The ideal is to preserve overall ecosystem health, including species diversity. Unfortunately, this is easier said than done. Much of the effort of conservation biologists has been directed toward learning how to manage ecosystems, at both theoretical and practical levels.

One of the classic debates in conservation biology centers around the “best” reserve design. If limited resources are available to purchase land, is it better to establish one big reserve or a few smaller ones? This has become such a famous and controversial issue that it has its own acronym, SLOSS (single large or several small). A large reserve provides the most potential habitat for interior species, which are usually the ones most in need of protection. However, a single reserve is vulnerable to all sorts of disasters. If a major hurricane or a pathogen hits that reserve, there are no other reserves to take its place. The establishment of several smaller reserves minimizes the risk of losing everything at the same time. However, a minimum size is needed to sustain populations of interior species as well as to preserve the characteristic species diversity and species composition of the ecosystem. Furthermore, reserves do not operate like isolated islands; thus connections between reserves and the surrounding habitats are also important.

A related concept involved in determining the optimum size of a nature reserve is the minimum dynamic area (Pickett and Thompson, 1978). Assuming the disturbance regime of an area is known, the frequency, areal extent, and recovery time can be used to determine the smallest reserve area in which there will always be some mature patch types to provide a species source for the rest of the area as it recovers from disturbance. If a patch is smaller than this minimum dynamic area, it will likely be eliminated through time simply from natural disturbances.

Given the large number of species on the planet, it is impossible, or at best impractical, to manage for every one of them. Instead, conservation biologists are now trying to identify ways to simplify the task of landscape-level management. The most promising methods identify one or a few important species, and concentrate on their management. One tactic is to manage keystone species, those on which important ecosystem functions or other species depend. Another approach is to target umbrella species, those with large ranges or broad habitat requirements. Managing for these species will automatically save many other species with smaller or less inclusive requirements. A similar method identifies a set of focal species, each of which is sensitive to a particular aspect of landscape structure or function. One of the focal species might be especially vulnerable to habitat fragmentation, whereas another might require a high level of connectivity. The protection of this set of sensitive species provides the management goals. When the requirements of the sensitive species are met, other species will be provided for as well.

VII. CASE STUDY: SEVILLETA NATIONAL WILDLIFE REFUGE

The Sevilleta National Wildlife Refuge (SNWR; 34.3°N, 106.9°W), located approximately 75 km south of Albuquerque, New Mexico, provides an excellent example of landscape diversity and its relationship to species diversity. This 100,000-ha wildlife refuge was established in 1973 and is currently managed by the United States Fish and Wildlife Service. The refuge is also a Long-Term Ecological Research site funded by the U.S. National Science Foundation. The climate at the SNWR is semiarid to arid, with low amounts of precipitation and high temperatures during the April to October growing season. Mean annual precipitation over the past 65 years was 23.4 cm. (sd = 70.4 cm) and average annual temperature was 14.1°C (sd = 0.7°C).

The SNWR is uniquely located at the ecotonal boundary between four major grassland–shrubland biomes found within the continental United States. Two of these biomes, shortgrass steppe ecosystems and Chihuahuan desert grasslands, form transition zones in the eastern part of the refuge (Figs. 7a and 7b). Patches of variable size (<10 m² to >1000 m²) and shape occur and result in high landscape diversity (Fig. 7b). These patches can be differentiated into one of two patch types based on the cover of the dominant plant species (Gost, 1995; Kroel-Dulay et al., 1997). The vegetation of some patches consists mostly of blue grama (Bouteloua gracilis), the dominant species in shortgrass steppe ecosystems. A second patch type occurs where the majority of cover is black grama (Bouteloua eriopoda), a dominant grass in Chihuahuan desert ecosystems. Species richness is similar for both patch types, although the cover of plants is higher in black grama compared to blue grama patches (Coffin, 1997). A transect across the conceptual landscape shown in Fig. 7b goes through each patch type as well as the matrix vegetation where similar cover of both species occurs (Fig. 7c).

Within each patch, a smaller scale of heterogeneity also exists due to disturbances associated with the burrowing activities of bannertail kangaroo rats (see Fig.
Figure 7  Patch types at the Sevilleta National Wildlife Refuge LTER site (SNWR). (a) Location of the SNWR in central New Mexico, United States, and location of patch types in the eastern part of the refuge. (b) Location of two patch types on the landscape. The transect across the landscape goes through patches dominated by either blue grama, black grama, or the matrix containing both species. (c) Cover of blue grama and black grama along the transect, showing one approach to identifying and delineating patches.
Small-scale heterogeneity within one black grama patch from Fig. 7b. Three types of mounds can be defined based on plant species composition and burrowing activities of bannertail kangaroo rats. (a) Mounds change from one type to another through time as kangaroo rats abandon some mounds and invade others. (b) Number of plant species and average plant size for each mound type. (c) Because the number of mounds within each of the three types is constant in this hypothetical landscape, the total number of species found on the landscape is also constant.
Mounds can be distinguished into one of three types based on plant species diversity as well as composi-
tion (Fig. 8a). Active mounds are the site of frequent burrowing, thus only plant species well adapted to dis-
turbance can survive there. Typically this vegetation consists of small plants representing few species (Fig.
8b). After mounds are abandoned and burrowing activi-
ties cease, more plant species can survive to larger sizes
on these early-successional mounds. Through time, compe-
petition among plants typically reduces the number of
species, although plant sizes can be quite large as
one or a few plants can come to dominate late-successi-
ional mounds. This invasion–abandonment cycle results in
a shifting mosaic of mound types through time across
the landscape (Fig. 8a). Although the species diversity
on mounds changes through time, and the location of
mound types varies spatially across the landscape, the
total numbers of species and patch types remain con-
stant on the scale of the landscape (Fig. 8c). Therefore,
landscape diversity both reflects and determines pat-
terns in diversity at smaller levels of organization, and
in particular species diversity.

VIII. CONCLUSIONS
Although much of the current emphasis on biodiversity
has been at the level of species, landscape diversity is
also important. The preservation and maintenance of
multiple levels of organization, including species, popu-
lations, communities, and ecosystems, require an un-
derstanding of how these various levels interact with
their environment across a range of spatial scales. Main-
tenance of landscape diversity provides a spatial tem-
plate for the preservation of these smaller levels of orga-
nization, and in particular for species biodiversity.
Changes in landscape structure and function through
time have important effects on the distribution of re-
sources, with resulting influences on the survival of
species. Because of the overwhelming numbers of spe-
cies, it may be impractical to attempt to conserve species
diversity per se. By focusing on landscape diversity and
the perpetuation of dynamic processes across multiple
scales, an attempt can be made to preserve entire ecosys-
tems with their full complement of genetic diversity.

See Also the Following Articles
DEFORESTATION AND LAND CLEARING • DESERTIFICATION • LAND USE ISSUES • KEYSTONE
SPECIES • RANGE ECOLOGY • TEMPERATE GRASSLAND AND SHRUBLAND ECOSYSTEMS

Bibliography
Bierregaard, R. O., J. T. E. Lovejoy, V. Kapros, A. A. dos Santos, and
Forested Ecosystem: Disturbance, Development, and the Steady State
Based on the Hubbard Brook Ecosystem. Springer-Verlag, New York.
Burke, I. C., T. G. F. Krol, W. K. Laur한다면, P. Smokol, C. M. Venker,
(1998). The influence of geomorphological heterogeneity on bio-
diversity. I. Patch-scale perspective. Conservation Biol. 12(2),
362–372.
Collin, D. P. (1987). Spatially-explicit propagule sources and vege-
tation dynamics at a grassland–shrubland ecotone. US-
International Association for Landscape Ecology conference,
of a semiarid grassland: Effects of soil texture and disturbance
size. Vegetatio 100, 67–82.
interactions in pastoral, ranching, and native ungulate ecosystems.
J. Range Management 44(6), 330–342.
and Regions. Cambridge University Press, Cambridge, United
Kingdom.
Wiley & Sons, New York.
Impact of human activity on regional forest composition and
on Changing Landscapes. Cambridge University Press, Cambridge,
United Kingdom.
Kremen-Dunly, G., T. Hochmarter, and D. P. Collin. (1997). Composi-
tional comparison of grass-dominated patches at a semiarid–
approach to assess species and ecosystem diversity. Conservation
Biol. 7(3), 1148–1158.
Lovejoy, T. E., J. M. Ranken, B. O. Bierregaard, Jr., K. S. Brown, Jr.,
L. H. Emmons, and M. Van der Vooit. (1984). Ecosystem decay of
Lovejoy, T. E., R. O. Bierregaard, Jr., A. R. Byrd, J. R. Malvick,
C. E. Queevara, L. H. Harper, K. S. Brown, Jr., A. H. Powell,
G. V. H. Powell, H. O. B. Schubert, and M. B. Hayes. (1986). Edge
and other effects of isolation on Amazon forest fragments. In
Conservation Biology: The Science of Scarcity and Diversity (M.
Soule, ed.), pp. 251–263. Sierra Associates, Sunderland, Massa-
chusetts.
Mack, R. N. (1983). Invasion of Bromus tectorum L into western North

7b).


LAND USE ISSUES

John M. Marzluff and Nathalie Hamel
University of Washington

I. Land Use Issues
II. Effects of Land Use on Biodiversity
III. Methodological Issues in the Estimation of Land Cover and Land Use

GLOSSARY

afforestation Process of planting trees in areas currently devoid of them. It is done in desert regions and is increasingly common in Europe, where grasslands and pastures are being converted into woodlands and forests.

agricultural intensification Management of agricultural land to increase yield principally through the use of high-yielding crop varieties, chemical fertilizers and pesticides, irrigation, shorter rotations, larger fields, and mechanization.

desertification Reduction of the biological potential of land that ultimately leads to the creation of desertlike conditions. It occurs primarily through overuse of marginal lands by humans for agriculture.

grassland Land that has a vegetation cover dominated by grasses.

habitat fragmentation Splintering of once contiguous land cover into isolated pieces. Fragmentation happens when habitat is lost from the interior, rather than the edge, of a large block of cover. The resulting habitat patches are sometimes referred to as “habitat islands” and the intervening, converted land is called the “matrix.”

land cover Physical and biotic character of the earth’s terrestrial surface. Land cover is typically the vegetation (e.g., tropical forest, marsh, desert, cornfield, shrubland) or human construct (e.g., road, dwelling, industrial area) that covers the surface. It may be grossly defined as in the preceding examples, or defined at finer scales (e.g., moist deciduous tropical forest, moist evergreen tropical forest, dry tropical forest).

land use The use that humans make of the earth’s terrestrial surface, usually to obtain goods or benefits. Examples include agriculture, mineral exploration, settlement, natural reserve, and timber production. Land use refers to the dominant human activity that occurs on the earth’s surface, and it often modifies natural characters to change land cover (e.g., replacing native grassland with a cornfield or a road).

settlement Land used for human habitation, including for the construction of cities, towns, villages, rural settlements, and roads. In a gross sense, settlement is also a land cover when used to refer to the land occupied by the variety of trappings accompanying human living space, including homes, gardens, roads, transportation centers, and industrial areas.

LAND USE ISSUES CONCERN THE PROCESSES by which human activities determine land cover. Important issues are agricultural development and intensification, settlement, and extraction of natural resources.
In response to human land use, the earth’s land cover has changed from a mosaic of native woodlands, forests, and grasslands to an increasingly impacted mixture of degraded and fragmented native habitats, exotic croplands, and impervious urban surfaces. In the last three centuries, models suggest that forests have declined 19%, grasslands have declined 8%, and cropland has increased over 400%. This article discusses how land use processes have changed through time and how they have caused the natural pattern of land cover to change. This transformation of the planet’s landscape is widely recognized as the primary driver in the current global loss of biodiversity. Several examples of how land use can influence biodiversity are also considered.

I. LAND USE ISSUES

Humans obtain food, energy, and shelter from the earth by harvesting natural resources and creating new resources. Both of these activities, especially the creation of new resources by agricultural and industrial actions, modify our environment at a scale unknown for other organisms. Our unparalleled recent success as a species (increasing in population size by nearly 2% per year from 1950 to 1985) and our tremendous ability to modify the environment have allowed us to dramatically transform the earth (Turner et al., 1990). It is estimated that we have transformed or degraded from 39 to 50% of the earth’s surface and now command 25% of the world’s total net primary productivity (Vitousek et al., 1986, 1997).

We can conceptualize our effects on the earth’s surface by relating our land and resource use to changes in land cover (Fig. 1). As our population increases, we garner more of the earth to supply our needs. The aspects of the human population that drive changes in land use and resulting land cover are known only in a general way. Population size is obviously an important determinant of our land use. But its importance relative to global affluence, technological capacity, political organization, and social structuring is hotly contested (Meyer and Turner, 1992). Recent insights suggest that population size may directly determine land use at a global scale, but land use locally is determined by the interaction of regional population size, technology, socioeconomic development, industrialization, and policy (see Fig. 1). Total biological demand for resources and resulting conversion of land cover are determined by the total world population. Where on the planet changes in land use occur is affected by regional population size, technology, policies, conventions, cultural attitudes, and economic forces that discourage or reward land conversion at the local scale (Meyer and Turner, 1992).

Three main land uses are responsible for changes in the earth’s land cover. These are (1) agriculture, which includes cultivation, grazing, irrigation, and drainage; (2) extraction of natural resources, primarily for forestry and energy production; and (3) settlement, which we define loosely to include use of land for human settlements ranging from rural villages to cities. Our domination of the earth is clearly reflected by the earth’s current land cover, which is primarily a mixture of fragmented and degraded native habitats, exotic croplands and pastures, and settlements (Fig. 2). This transformation of land cover per se is recognized as the driving force in the loss of biological diversity worldwide. Additionally, land use affects the biosphere indirectly by modifying ecosystem processes and introducing exotic species (Vitousek et al., 1997).

A. Agriculture

Our agricultural practices have had the greatest influence on the earth’s land cover. Before the development of agricultural society, several lines of evidence suggest
FIGURE 2  Modern land cover (gross scale) of the world. Human domination is evident by the preponderance of agricultural land cover (arable, irrigated, grazing, and range lands). (From The National Geographic Atlas of the World, National Geographic Society, Washington, D.C., 1992.)
that the earth was covered with a complex mosaic of forest, woodland, grassland, wetland, tundra, and desert (Fig. 3). During the last 300 years we have cultivated approximately 12% of the earth’s surface. This represents an approximate increase of 390 to 466% since 1700 (Fig. 4; Richards, 1990). Increases in cultivated land came mostly at the expense of native forests (other than tropical rain forest) and native grasslands (Fig. 5). We transform the earth by burning and clearing forests, tilling grasslands, draining wetlands, and irrigating arid lands. Agricultural development is responsible for the overwhelming majority of wetland conversion (the
global loss of wetlands is not accurately estimated; Meyer and Turner, 1992). The pattern of increasing human population, increasing cropland, and declining forests and grasslands is consistent across the continents except in Europe, where current afforestation is increasing forest and woodland cover (Fig. 6).

Although conversion of native land to exotic croplands has slowed, our use of agricultural land has intensified since “the Green Revolution” began in the 1960s (Matson et al., 1997). Agricultural intensification alters land cover by (1) creating highly productive monocultures, (2) diverting water resources, and (3) fueling technology to produce better seeds and more fertilizers, pesticides, and insecticides.

Increasingly intensive use of grasslands for agriculture has resulted in worldwide overgrazing (Fig. 7). A direct effect of this is “desertification” or the degradation of land to desert-like conditions. Current estimates indicate that 6% of the earth’s land has been converted to deserts by agricultural activity; nearly 25% of the earth is threatened with desertification (Meyer and Turner, 1992).

B. Natural Resource Extraction

We extract a variety of resources from the earth for shelter and energy, but at a global scale timber harvest and management have the largest effect among extractive activities on landscape composition and land cover. Roughly one-third of the earth is forested and 18% (650 million ha) of forests are currently maintained as plantations or managed for goods and services (Noble and Dirzo, 1997). Humans affect all forests by hunting, gathering, recreating, and suppressing fire. However, only fire suppression has the potential to change forest cover type (if successful, it initially slows forest growth by reducing natural thinning but may allow for greater development of late-successional forests by reducing disturbance).

Natural resource extraction has a less direct effect on forest cover than does agriculture. Resource extraction modifies the forest by changing seral stage, degrading forest integrity, and fragmenting continuous forest, but is associated with reduced forest cover in only a few locations (southeastern Asia and parts of Africa, Latin America, and India). Agriculture, in contrast, is the primary agent of forest loss, although it also modifies forest function. Some of the forest loss to agriculture is compensated for by maintenance and planting of forest for timber. From 1980 to 1995, one-tenth of the 13 million ha of forest lost per year to agriculture was compensated for by the creation of forest plantations.

Much of this trade-off is occurring in Europe, where afforestation campaigns are increasing forest land at the expense of cropland (see Fig. 6c).

C. Settlement

The clearing of native land cover and drainage of wetlands for human settlement are the most permanent forms of land conversion. Overall, 2.5 to 6% of the earth is now occupied by human dwellings and their supporting infrastructure (roads, industrial areas, reservoirs, etc.). About 10% of this area is actually built area with impervious surface (Meyer and Turner, 1992). Human settlements typically occupy less than 20% of even the most developed country’s area, but this varies widely (Fig. 8). Each night we advertise our settlements by turning on lights. Viewing this spectacle from the air provides a striking representation of the degree to which settlements occupy the earth (Fig. 9). Human settlement affects nearly all land in Europe, eastern North America, India, and parts of Asia.

Settlement has increased dramatically over the last 300 years as evidenced by the increasing number of large cities (Fig. 10). As human populations have grown, we have settled with increasing frequency in urban areas. Today nearly half of all humans live in urban areas (primarily in the developed countries), compared to only 14% at the start of the twentieth century (Meyer and Turner, 1992). A result of our increasingly urban life is an increased occupation of area by cities and surrounding lower-density housing and industrial areas (urban and suburban sprawl). This expansion of urban areas in developed countries and the increasing density of large urban areas in the developing tropics are responsible for the dramatic and likely irreversible transformation of the earth’s surface.

D. History Repeats Itself: Past and Future

Effects of Land Use on Land Cover

Our transformation of the earth during the last five centuries has proceeded by pioneers expanding frontier settlements in a fairly regular way. When humans settle an area, they open a new frontier where natural resources are abundant, cheap, and readily obtained. As resources are consumed and population increases, we expand our “footprint” on the surrounding land in a series of radiating waves. We transform the land at successively greater distances from a settlement’s center by clearing and draining land, extracting resources, and converting land cover to suit our needs (Berry, 1990).

The conversion of land cover was fueled on a global
FIGURE 6  Land use trends and population estimates of major regions of the earth. Land use trends between 1700 and 1980 (closed symbols) are compiled from Richards (1990). After 1950 (open symbols), land use estimates are derived from FAOSTAT (www.fao.org, October 1998). Population estimates in (a) are for Africa only. Panel (f) includes Japan, North and South Korea, and Oceania.
scale by Western European nations and their seaborne pioneers from the 1700s to the early 1900s (Richards, 1990). Once the European frontiers were settled, the appetite for amenity and subsistence resources propelled the opening of new frontiers around the world. These frontiers then had to satisfy local and European needs, so resources were harvested with frenzied zeal as opportunities seemed boundless on the frontier. Landscapes were quickly degraded and transformed, and as resources dwindled, the frontier closed and pioneers moved on to new frontiers. However, the old frontier was rarely left to return to a natural state. Rather, a more orderly settlement usually persisted, leading to slower, but more permanent, land transformation. Many of these settlements became new centers in a changing world economy, thereby broadening the base from which remaining frontiers were exploited.

The global pattern of urbanization over the last 300 years (see Fig. 10) illustrates the consequences of our opening and closing the earth’s frontiers. New frontiers open in areas increasingly far from Europe, and develop large urban areas as they close. The frontiers of Europe, eastern North America, India, and Asia have closed, leaving dense human settlement to permanently change land cover. As the remote frontiers of the tropics, mountainous regions, and northern forests and tundras are exploited, settlement will likely increase and land will be transformed to a lasting urban state. Europe can be viewed as a crystal ball for much of the earth’s future land use. The forecast is for drastic simplification of a once diverse native land cover to one dominated by settlements, intensive agriculture, and fragmented vestiges of native habitat.

II. EFFECTS OF LAND USE ON BIODIVERSITY

The current extinction crisis is fueled, at a proximate level, by (1) habitat conversion, (2) habitat fragmentation, and (3) habitat degradation. These are the most common reasons given to explain why modern species are at risk of extinction (IUCN, 1996). These three processes are the direct result of our land use and they function as the mechanisms that connect changing land cover to biodiversity (see Fig. 1). Conversion of native land cover to agriculture and settlement is unquestionably the main cause of habitat loss on a global scale. This conversion typically occurs in a piecemeal fashion and is associated with extensive road building that fragments habitat as it is being lost. Native and converted habitats that are already utilized by humans are further degraded by agricultural intensification, resource extraction, and recreation.

The effects of land use on biodiversity can be illustrated by considering habitat conversion, fragmentation, and degradation in several regional studies, and we have selected well-documented studies to illustrate the general process. It is important to consider regional changes in land cover because the factors that drive land use and link it to land cover are best understood...
FIGURE 9  The world at night. Viewed from space, the distribution of lights (white blurs) in human settlements indicates the omnipresence of urban space. This image taken from a satellite is a composite of cloud-free observations made over a 6-month period. The red is the result of fires, and the yellow indicates natural gas flares. (From National Geographic Maps, October 1998, National Geographic Society, Washington, D.C.)

FIGURE 10  Distribution of the sizes of world cities, between 1700 and 1985. The growth rate of cities in the Americas is more rapid relative to that of cities on the European and East Asian continents. The black dots represent cities of varying sizes (●, 200,000 to 500,000; ○, 500,000 to 1,000,000; ◆, 1,000,000 to 10,000,000). From Berry (1990).
at “subglobal” levels (Kummer and Turner, 1994). The following sections examine (1) the complex ways that biophysical and socioeconomic factors affect land use, (2) the linkage between land use and habitat conversion, fragmentation, and degradation, and (3) the effects of land use on biodiversity.

A. Habitat Conversion

1. Deforestation in Southeast Asia

The tropical rain forests of Southeast Asia have declined dramatically since the end of World War II (in the late 1940s). Before humans arrived, the Philippines, Indonesia, Malaysia, Vietnam, and Thailand were completely forested (see Fig. 3). Their rich dipterocarp forests were heavily logged (primarily for export) after the war ended. The Philippines currently retains less than 20% of its original forest cover; Malaysia retains more forest, but it is being extensively harvested (Fig. 11). The rates of deforestation in the Philippines are well documented and staggering (Collins et al., 1991; Kummer, 1992). From 1948 to 1987, 55% of the forest was lost. From 1969 to 1988, the rate of deforestation was over 2% per year—or 2 ha every 5 minutes!

The process of habitat loss (deforestation) in the Philippines serves as a model for all of Southeast Asia (Kummer and Turner, 1994). Two linked land uses, logging and subsistence agriculture, are responsible for converting land cover from tropical rain forest to arable land. Logging dipterocarps converted the primary rain forest to a simpler secondary forest. The resulting partial clearing and road construction allowed poor, landless people to remove the secondary forest and expand agriculture.

![Diagram showing current land cover in Southeast Asia](image)
This two-step process illustrates how human density, global markets, and local socioeconomic conditions work together to cause changes in land cover. Despite the Philippines’ extremely dense (217 people/km² in 1993) and rapidly growing population (2.3% in 1993), deforestation is not directly driven by the local population. Rather, a global demand for wood, coupled with corrupt and unrestricted logging, cleared land and made it accessible to people in need of agricultural space. A surplus of landless people was needed to drive the forest conversion to agricultural land, but without logging to open a new frontier, urbanization might have increased and agriculture might have remained more centralized and intense rather than dispersed as it is now.

Southeast Asia in general, and the Philippines in particular, has tremendous biodiversity. Diversity in the Philippines is poorly known, but there are >12,000 plants and fungi and >950 terrestrial vertebrates. This incredible diversity is primarily found in a large number (>59) of scattered and poorly protected national parks. Less than 1.3% of the Philippines is managed for conservation, and much of this area is still logged and settled by landless agriculturalists.

The combination of high biodiversity, high endemism, and extreme habitat conversion results in extreme risks of extinction. The magnitude of threat to birds and mammals in Southeast Asia is particularly well known (IUCN, 1996). Worldwide, Indonesia has the largest number of birds (104) and mammals (128) threatened with extinction. The Philippines leads the world in the percentage of its avifauna that is threatened with extinction (15%). These countries are second only to Madagascar in the percentage of mammals threatened with extinction (32%). Vietnam, Thailand, and Malaysia are all in the world’s top 20 countries for species endangerment.

To summarize, deforestation in Southeast Asia is an example of how our land use for amenity (dipterocarp logging) and subsistence (agriculture) resources changes land cover and affects biodiversity. Habitat conversion threatens biodiversity, but the greed (exotic wood) and need (agriculture) of a growing global human population fuel the conversion.

2. Urbanization in Western Washington State

An expanding population and the world’s appetite for resources have transformed much of North America from extensive forest and grassland to agricultural and urban land (compare Figs. 2 and 3). The general story is similar to the case in Southeast Asia. However, in the Pacific Northwest a more dramatic and permanent conversion of land is occurring at a rapid rate. Logging during the last 50 years has converted more than two-thirds of the primary forest in Washington to secondary forest. Rather than continuing to manage these forests for timber production, much of the land is now being converted to human settlements. In western Washington, for example, human populations have doubled in the last 50 years and are expected to double again in the next 50 years. From 1998 to 2000, Washington is expecting a net gain of one person every 5 minutes! This places a premium value on land for settlement. The predictable result is a rapid conversion of forest to urban and suburban settlements. Indeed, from 1970 to 1997, Washington lost 2.3 million acres of managed forestland. Urban expansion was responsible for about half of this loss. Rights-of-way and agriculture accounted for the rest.

The increasing urbanization in western Washington threatens to reduce biodiversity. Water flows altered by settlement have reduced the spawning and rearing habitat for the region’s spectacular salmon diversity. As a result, several salmon runs have been extinguished and many are now listed as endangered. Loss of forests and intensification of resource extraction and recreation on remaining forests have contributed to the endangerment of several birds (e.g., Spotted Owls, Strix occidentalis, and Marbled Murrelets, Brachyramphus marmoratus), mammals (e.g., Grizzly Bear, Ursus arctos, and Gray Wolf, Canis lupus), and amphibians (Larch Mountain Salamander, Plethodon larselli). Drainage of wetlands and settlement of native woodlands and grasslands have endangered amphibians (Oregon Spotted Frog, Rana luteiventris), insects (Oregon Silverspot butterfly, Speyeria zerene), mammals (Western Gray Squirrel, Sciurus griseus), and reptiles (Western Pond Turtle, Clemmys marmorata).

Some species benefit whenever land cover changes. A good example of such a species is the American Crow (Corvus brachyrhynchos) in western Washington. Crows are found only in close association with human settlement. As a result, their numbers have increased 10-fold from 1960 to 1995. Increases in human commensals like crows may have a ripple effect through the native biota. Crows prey on the eggs and young of other birds, which may limit their reproduction and reduce overall biodiversity close to human settlements even in remaining forest reserves.

B. Habitat Fragmentation

The loss of native habitat as lands are converted for human use splinters the remaining habitat into small
fragments afloat in a human-dominated matrix. Such “habitat fragmentation” is most evident on a global scale when we compare the original and current distributions of forests (Fig. 12). contiguous primary forests have been removed to varying degrees, resulting in a patchwork of forest “islands” in a “sea” of settlement, agriculture, and second growth forests.

Forest fragmentation affects biodiversity by (1) outright loss of habitat (recall Southeast Asia), (2) creation of forest edges that differ from interior forest in many physical and biological properties (e.g., wind speed, humidity, temperature, and predator populations), and (3) disruption of movement and dispersal patterns of forest species. The effects of habitat fragmentation depend on the type of land cover that surrounds the fragments (the “matrix”). In general, when the matrix is similar to the native habitat (e.g., secondary forest surrounding native forest), the effects of fragmentation are less than when the matrix is dominated by human settlements or agriculture. The size of the remaining fragment is also important; large fragments conserve more biodiversity than small ones.

A variety of recent studies of fragmentation in the tropics suggest that habitat fragmentation reduces
global biodiversity but not necessarily local biodiversity (Schelhas and Greenberg, 1996; Laurance and Bierregaard, 1997). This seeming paradox occurs because some species benefit while others suffer from fragmentation. The result is that local communities change composition after fragmentation, but the overall number of species in a local area does not necessarily decline. Forest interior specialists and wide-ranging species, notably predators, may go extinct with extensive fragmentation, thus lowering global biodiversity. Generalist and exotic species typically increase in fragmented habitats. Research suggests that many forest species tolerate considerable fragmentation because they can utilize the surrounding matrix. The extent of their ability to sustain viable populations depends on resources being available in the matrix and similarity of the matrix to the remaining native habitat.

Research in Singapore and Hong Kong is beginning to reveal the long-term effects of fragmentation. Forest fragments on these islands have been isolated for 150 to 350 years. Extinction of birds and mammals has been severe, but many plants are able to persist. Forest interior plants that are sensitive to changes in temperature and humidity (such as orchids and epiphytes) are also extinguished by fragmentation.

C. Habitat Degradation

Changes in land use can affect biodiversity without changing land cover. Land cover can remain essentially the same while the intensity of use increases and degrades habitat. Three forms of land use that degrade land cover rather than drastically changing it are (1) agricultural intensification, (2) recreation, and (3) forest management for sustainable timber production.

1. Agricultural Intensification in Europe

Agriculture is the dominant form of land use in Europe. Conversion of forests to agricultural lands has resulted in dramatic changes in land cover. Forests once covered more than 80% of Europe, but today they cover only one-third of the land. In contrast, slightly more than 40% of the European landmass is now agricultural. Over one-third of this land is permanent grassland maintained for grazing; the remainder is arable land. Conversion of forest to agriculture is no longer a major conservation concern in portions of Europe like Britain that have few remaining primary forests. Instead, the intensification of agricultural practices is degrading the conservation value of lands with grave consequences for biodiversity.

Agricultural intensification in Europe comes in many guises (Newton, 1998). (1) Increased use of chemical pesticides and fertilizers. (2) Removal of hedges to increase field size. (3) Plowing in late summer after harvest rather than waiting until the following spring. (4) Draining wetlands to increase arable and pasture lands. (5) Conversion of mixed farms that produced a variety of animal and plant products to monoculture cereal farms. (6) Earlier harvest dates. (7) Intensive grassland management to increase grass growth. (8) Increased stocking density of sheep in hill country.

These changes in agricultural practices cause declines in bird distribution and density for three main reasons (Newton, 1998): (1) chemicals and early harvests reduce breeding success and survivorship of birds; (2) breeding and foraging habitat is reduced by hedge removal, land drainage, late summer plowing, intensive grassland management, and increased grazing; and (3) habitat diversity and resulting bird community diversity are reduced when mixed farms are converted to monocultures. The results of these changes have been dramatic. It is suspected that intensification has affected 42% of bird species of conservation concern in Europe. Nearly 90% of the 26 farmland birds in Britain decreased their range of occurrence from 1970 to 1990, and those species using pastures were negatively affected by increasing grazing intensity. Specific aspects of agricultural intensification affect particular species, but on the whole this form of habitat degradation threatens a significant portion of the common, as well as the rare, European avifauna.

2. Influence of Recreation on Wildlife Communities

As our landscape becomes more dominated by urban and agricultural land, our psychological need to visit and explore the last remaining wild places increases. We often fulfill this desire by camping, hiking, nature watching, hunting, and fishing in wildlands. These seemingly benign activities can affect biodiversity by harassing animals, trampling plants, and altering competitive, facilitative, and predator–prey relationships (Knight and Gutzwiller, 1995). Although recreation in wildlands rarely changes land cover, its effects are often greatest near settlements, thereby extending the actual area modified by settlement.

How campers and fishermen affect salmon-scavenging birds in the Pacific Northwest of North America is a good example of the subtle ways that our land use affects
biodiversity without changing land cover. Three species of birds [American Crow, Common Raven (Corvus corax), and Bald Eagle (Haliaetus leucocephalus)] vie for salmon that die after spawning each fall and winter. Bald Eagles are the largest and dominant scavenger, but they are also the least tolerant of human activity. Sport fishermen float and wade rivers in search of live salmon and often flush eagles from their meals. This disrupts eagle foraging and may lead to lower overwinter survival and reduced population viability. The rarity of eagles and their status as a national icon increase the desire of nature watchers to observe them during the winter, thereby disturbing them even further. Recreation effects do not stop here. Camping areas along rivers are common and they provide consistent sources of food that fuel and increase in the wildland crow abundance and distribution. Crows are tolerant of humans and quickly return to carcasses after disturbance. They attempt to eat carcasses dominated by eagles and ravens but rarely succeed until humans displace the eagles (and often ravens), thereby incidentally allowing crows to eat without the check of competitors.

3. Effects of Intensive Timber Management

Timber harvest directly affects biodiversity by (1) facilitating land cover change through clearing land for agriculture and settlement and (2) fragmenting contiguous primary reserves as discussed earlier. Intensive forest management to produce timber may also affect biodiversity in more subtle ways by (1) reducing the average age of the forest, (2) simplifying the age structure and diversity within forest stands, and (3) increasing the frequency of disturbance in forests.

An intensively managed forest is like an intensively managed agricultural field, with the focus on producing the most marketable timber in the least amount of time. However, the effects of timber management on biodiversity appear to be much less than the effects of agricultural intensification. Animals with large space requirements, especially requirements for the interiors of mature forests, do not fare well in managed forests. But animals and plants that utilize clearings, thickets, and young forests thrive in managed forests, which can actually increase local biodiversity by providing habitat for such species. Nonetheless, the loss of sensitive interior forest species usually leads to lower global biodiversity.

The long-term effects of intensive timber management on biodiversity are poorly known. We do not know if managed forests provide adequate resources for many birds during the nonbreeding season, nor do we know how the shifting nature of forest patches in time and space, so typical of managed forests, affects the ability of plants and animals to maintain viable populations connected by dispersal. It is safe to conclude that maintaining forested land cover is an important contribution to global conservation. The maintenance of managed forest, that perpetuate native forest trees and some of the biodiversity dependent on them certainly helps stem the global loss of native habitat to agriculture and settlement.

III. METHODOLOGICAL ISSUES IN THE ESTIMATION OF LAND COVER AND LAND USE

Much of the information presented in this article depends on the estimation of current land use and land cover on a global scale. Our conclusions depend directly on the quality of these estimates. Therefore it is important for the reader to have some insights into how these data are obtained and the degree of uncertainty they contain.

A. How Are Land Cover and Land Use Estimated?

Land cover is estimated directly by aerial photography at small scales and by satellite imagery at large scales. Aerial photographs resolve fine details in land cover at local or regional scales. Satellite imagery detects features over hundreds of kilometers, but small details are missed because the resolution is coarser.

An important satellite used for imaging the earth is Landsat. Its imaging abilities have improved over the years: current Landsat TM (thematic mapper) images have better resolution (30 m × 30 m) that permits the fine-scaled discrimination of land covers. Another satellite that is widely used is the AVHRR (Advanced Very High Resolution Radiometer). Its wider swath and coarser resolution (1.1 km) are good for describing large areas of vegetation.

While modern land cover can be estimated directly using photography and satellite imagery, land use cannot always be derived from the images. Hence statistics of land use supplement or substitute for imagery. The Food and Agriculture Organization of the United Nations (FAO) has reported tabulated data on land use and cover on an annual basis since the 1950s (these
data are used in Fig. 6). The FAO does not gather data independently but rather compiles national data from returned questionnaires. Globally, the FAO is an important source for time-series estimates of land use and cover over the last half of this century.

B. Limitations in Estimating Land Cover and Land Use

Virtually every determination of land use and land cover is an estimate that includes some level of smoothing, averaging, and guessing. The general conclusions we make about how land use affects land cover and biodiversity are not affected by these limitations. However, the application of the maps and processes to specific areas on the planet may fail because of estimation.

An important limitation occurs in large-scale mapping of land use and cover. Some minimum mapping unit (the resolution) is used in all compilations of land cover and use. Only one use or cover can be assigned to this minimum unit, and it is the dominant use or cover that is assigned and assumed to occur uniformly over the entire unit. For example, a unit that is simply recorded as “forest” might actually be covered by 40% forest, 30% lake, 20% grassland, and 10% building. Recording only the dominant cover reduces the occurrence of dispersed (rarely dominant) covers that typify many human-dominated land covers. Therefore, large-scale maps likely underestimate the actual effect that humans have on land cover.

Projected early estimates of land use and cover are by necessity approximations. The map of preagricultural land cover (see Fig. 3) was derived entirely by modeling the plant growth suspected under varying climatological and soil conditions. Cropland estimates in the 1700s were estimated by assuming that 0.2 ha of land was cleared for each human (Richards, 1990). Even the population size at that time was an estimate. Regional estimates of land cover and use are also of inconsistent quality. FAO data vary greatly in quality by country because uniform scales and cover categories are not used. Yet even these large sources of estimation and inaccuracy do not affect the general conclusion that at a global scale agriculture has increased tremendously, replaced native forests and grasslands, and caused declines in biodiversity.

Another important bias is an underestimation of habitat degradation. Mapping and imagery capture the basic pattern of habitat conversion and fragmentation, but rarely detect changes in the quality of habitats that retain their basic structure. Selective logging, understory trampling and removal, and disturbance by human presence are not recorded on global scales. Regional, local, and more subjective measures are needed to assess these changes in land use.

See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • AGRICULTURE, SUSTAINABLE • DEFORESTATION AND LAND CLEARING • DESERTIFICATION • LANDSCAPE DIVERSITY • LOGGED FORESTS • RANGE ECOLOGY • TEMPERATURE GRASSLAND AND SHRUBLAND ECOSYSTEMS • TIMBER INDUSTRY

Bibliography


LAND-USE PATTERNS, HISTORIC

Oliver Rackham
Corpus Christi College, University of Cambridge

I. Introduction
II. Human Activities and Climatic Change
III. Properties of Crops and Domestic Animals
IV. Fields, Hedges, and Terraces
V. Forest and Woodland
VI. Wood-Pasture and Savanna
VII. Grassland
VIII. Heathland
IX. Moorland
X. Buildings and Built-Up Areas
XI. The Rise and Fall of Historic Land-Use Patterns
XII. Implications for Biological Conservation

GLOSSARY

climax The stable ecosystem that is supposed to result from a sufficiently long period of unchanging environment.
clone A group of individuals formed by vegetative reproduction from a parent. This term applies particularly to those trees and plants that reproduce by creeping stems or by suckers from roots, forming circular patches of genetically identical individuals.
coppice Regrowth of a felled tree from the stump.
cultural ecosystem Ecosystem produced by the long-term interaction of wild plants and animals with human activities.
endemic Plant or animal species limited to a small area of the world, for example, one island or mountain range.

heath Open vegetation dominated by dwarf undershrubs usually of the family Ericaceae.
meadow Grassland maintained by mowing.
moor Open vegetation, dominated especially by dwarf Ericaceae and Sphagnum mosses, on peaty soil.
pasture Grassland maintained by grazing domestic animals.
pollard A tree repeatedly cut at about 6–10 feet above ground to produce successive crops of wood.
savanna Grassland (or other nonforest vegetation) with scattered trees.
seminatural Vegetation that owes its character to human activity, but in which the plants are wild, not sown or planted. Examples: coppice woods; most heathland; some savannas.
wildwood Wholly natural forest not affected by sedentary human activities such as agriculture and pastureage, widespread in prehistoric times (including earlier interglacial periods) and still surviving in remote places.
woodland Forest forming part of a cultural landscape.
wood-pasture Cultural ecosystems, often savanna-like, combining trees and domestic animals.

SOME OF THE WORLD’S BIODIVERSITY is associated with "virgin forest" and other ecosystems that are supposed to have escaped human interference until now. This category shrinks as archaeological research reveals more of the extent and pervasiveness of past human activities. The landscapes of most of the world’s land surface result from long and complex interactions be-

56

Deyclopedia of Biodiversity, Volume 3
Copyright © 2001 by Academic Press. All rights reserved.
I. INTRODUCTION

Human intervention often does not efface all signs of previous ecosystems. The traveler in Texas, despite 160 years of often intense activity by settlers, has no difficulty in recognizing the presettlement vegetation regions (Pineywoods, Blackland Prairie, Post Oak Savannah, etc.).

Even in Britain, after 6500 years of settlement, the landscape is still not wholly artificial. By careful search, one still finds remains of the mid-Holocene vegetation categories (limewoods in lowland England, oakwoods in the north and west, pinewoods in east Scotland). Superimposed on these are the largely artificial ecosystems of arable farmland, sown grassland, and forestry plantations; the seminatural ecosystems produced by the responses of wild plants and animals to continuing land use (abandoned mines, derelict forestry plantations); and the wealth of biodiversity contained in cities, old pits, and industrially derelict land.

Cultural ecosystems are often contrasted with wildwood or “virgin forest,” the aboriginal vegetation of much of the world’s land surface before the impact of settled human cultures. Wildwood itself, however, may already have been altered by the activities of nonsettled peoples. There is abundant evidence to show that hunting and gathering were not simple, unorganized activities, but could have profound effects on landscape and ecology, even in places remote from the activities themselves. An example is the extermination of elephants and other great mammals by Paleolithic peoples. The loss of the only animals capable of breaking down big trees could hardly fail to have profound, if as yet little known, consequences. From the early Holocene, if not earlier, land management by burning affected the ecology where vegetation was combustible, especially in Australia and North America. These activities, as far as we know, created only simple land-use patterns: they would alter, for example, the boundaries between forest and prairie or between combustible and noncombustible vegetation. Some of the Texas ecosystems, such as the Edwards Plateau savannas, were probably produced by Native American land use and taken over by settlers.

The introduction of settlements, domestic animals, and cultivated plants—in European terms, the coming of Neolithic cultures within the past 8000 years—began to create the diversity and complexity of land-use patterns that is familiar today.

II. HUMAN ACTIVITIES AND CLIMATIC CHANGE

It used to be thought that, given enough time, the plant and animal communities in an area would settle down to a stable climax state, determined only by the climate and (maybe) geology and changing only slowly in response to evolution. This may have been so in the Tertiary geological period, but in the Quaternary the dominant factor has been climatic change. Glacial cycles in high latitudes destroyed all vegetation over large areas, including much of Britain. In middle latitudes and even in the tropics, although there has been little ice cover, climatic changes such as drought have been sufficient to displace or destroy most of the interglacial plant communities.

There have been many glacial cycles over the past 2 million years. In much of the world, forest has been the “normal” vegetation only in interglacial periods like the present, of at most a few tens of thousands of years each. Changes of climate have been too rapid for plants and animals, apart from annuals and other fast-reproducing species, to keep up with them by evolutionary change.

Human activity has recently become the predominant force shaping the world’s ecosystems. However, it did not burst into a world of stable environment in which evolution had previously been the predominant force. It came on top of a period of unusual instability: climatic change had been operating on a time scale not much longer than that of human activity itself. In Britain humanized ecosystems have existed for about 6000 years; the “virgin forest” preceding them had existed for at most another 6000 years and was itself affected by the activities of Mesolithic and Paleolithic people.

Climatic change and human activity overlap in time, and their effects can be difficult to separate. In the Mediterranean, the present, rather arid, ecosystems are variously attributed to the strongly seasonal climate with its hot dry summers and to the activities of people “degrading” what would otherwise be a more forested landscape. In reality, the present climate there is only about 6000 years old and set in at the same time that Neolithic people were beginning to have a profound effect on the landscape.
III. PROPERTIES OF CROPS AND DOMESTIC ANIMALS

Land-use patterns depend on the ecological behavior of a few species of major crop plants and domestic animals, which are usually not indigenous to the region.

Most of the crop plants of European-style agriculture originated in the steppes of southwest and central Asia and have a particular set of ecological requirements. They have modest moisture needs, rather high needs for mineral nutrients, and are very intolerant of water-logging and shade. As they spread, first to Europe and later to European colonies, farmers strove to convert more and more of the world's ecosystems to an imitation of a Turkish steppe in which wheat and barley are at home. Of the multitude of indigenous American crop plants, only a few—corn, potato, and tobacco—have been widely grown in Europe; these are the few whose ecological requirements are not too different from those of Southwest Asian crops, although they need more moisture. Rice is a Southeast Asian crop with very different requirements, growing in wetlands and making possible the great wetland civilizations of Asia. Hence it comes about that generations of farmers have sought to destroy Europe's fens and convert them into artificial steppe, whereas Japanese farmers, with the utmost ingenuity, have contrived to turn hillslides into artificial fen (Fig. 1).

Many tropical crops, on the other hand, will grow in moderate shade, and some require it. The distinction between forest and farmland, which is so sharp in most European-style agriculture, is not always clear in the tropics.

Keeping domestic animals also involves destroying or modifying natural ecosystems. Cattle, sheep, and goats eat tree leaves but are not forest animals. If kept in the shade of trees, they soon eat all the leafage within reach, creating a "browse-line" beneath which nothing is allowed to grow. To keep them in manageable numbers requires at least some gaps between the trees, in which grasses and low herbage will grow. Pigs, although often thought of as acorn feeders, cannot be kept in forest alone. Acorns are available for only a few months of the year and contain insufficient protein to sustain pigs by themselves. The world's cultural ecosystems would be very different had monkeys been domesticated for human food.

IV. FIELDS, HEDGES, AND TERRACES

In some parts of the world, arable fields are impermanent, being created out of forest or grassland, cultivated for a few years, and then allowed to revert to natural vegetation. The landscape consists of a mosaic of areas in different stages of reversion. This shifting cultivation is typical of regions too infertile to support even primitive agriculture permanently, especially in the tropics. In more fertile regions, such as most of Europe, people avoid having to repeat the immense labor of creating new farmland. Fields are permanent and tend to be private property with fixed boundaries.

In regions not yet reached by weed killers, arable fields are seminatural ecosystems. Many weeds came from the original homelands of the crops they accompanied, and some annuals have evolved in parallel with particular crops: for example, cow-wheat (Melampyrum arvense), whose seeds resemble a grain of wheat, so that they are sown along with the wheat crop. In Japan, rice fields are complex aquatic ecosystems involving cyanophytes, water ferns, amphibians, and insects; some of the plants fix nitrogen and so benefit the rice crop.

Fields tend to be discontinuous. As well as boundaries where they adjoin forest, pasture, or roads, they have boundaries separating ownerships or subdivisions within an ownership (Fig. 2). These are determined both by practicalities (e.g., erosion control) and by anthropological factors, such as land tenure, inheritance, or collectivization. The boundaries (hedges, banks, walls, etc.) provide a network of relatively natural ecosystems permeating the farmland. The size of fields varies from country to country. In parts of Canada or Australia, 100 acres is a small field. In Texas, the nineteenth-century allotments of 1 square mile (640
acres) are by now divided into at least 30 compartments in several ownerships. In the hedged parts of medieval England, a typical field would be of about 4 acres. In Crete, fields tend to be of about 1 acre. In Japan, with rice as an immensely high-yielding crop, even this would be a big field.

A hedge (in North America, a fence-row) is a row or strip of trees or shrubs forming the boundary of a field. In Europe, these are often maintained by cutting in particular styles to form a barrier to keep cattle or sheep from straying into the next field; often trees are left at intervals either uncut or pollarded (Fig. 3). In America, they are less managed.

In America and to some extent Europe, most hedges have arisen spontaneously as trees have taken root from wind- or bird-dispersed seed at the bases of fences or on balks or walls between fields. In England, hedges are thought of as deliberately planted, but can arise naturally, as many of the older hedges probably did. In Australia, hedges are rare, evidently because most eucalyptuses and other native trees have poor powers of seed dispersal.

Hedges presumably go back in some countries to the beginnings of agriculture. In Europe, the earliest evidence for them is about 2000 years old; they probably go back much further, but leave little archaeological record. In England, they are well documented for about 1200 years. Many individual hedges are at least 500 years old; these tend to be mixed in their composition, unlike the monotonous hedges of more recent origin.

Hedges are often thought of as linear wood-lots. This is partly true: in England, many hedges were periodically felled as if they were woodland, yielding crops of fuel wood and timber. However, in England and (from what I have seen of them in America), hedges are distinct vegetation types in themselves, seldom corresponding in detail to the local forest. They usually lack the more exacting woodland plants, apart from those few hedges that are the surviving edges of grubbed-out wood-lots.

Hedges are important as diversifying what might otherwise be featureless farmland. They introduce tree-living (though not forest birds) to open country that would lack nesting sites. They are often regarded as corridors for movement of mammals across farmland.

In many mountainous and densely populated parts of the world, slopes are shaped into terraces to create flattish surfaces on which to grow crops (Fig. 4). Banks or walls between terraces, like hedges between fields, can be significant sites for seminatural vegetation.

V. FOREST AND WOODLAND

It is often supposed that wildwood consisted of trees, trees, and nothing but trees, and only shade-bearing ground vegetation; gaps arising through the death of trees would promptly be filled by the growth of new trees. Examples can be seen where the dominant trees are of shade-bearing kinds such as Abies and Fagus species. The argument that wildwood was like this is often based on the general doctrine of climax vegetation, rather than on evidence of the history of specific forests. Pollen analysis can exaggerate the dominance of trees, many of which produce more pollen than her-
land-use patterns, historic

Figure 4: Cretan cultural landscapes in a semiarid climate. In the bottom of the basin are houses grouped into hamlets, with gardens, orchards, and small fields demarcated by dry-stone walls. Extending up the mountains are remains of terraces, on which grain used to be grown. Asphendou, Sphakia, Crete

Figure 5: A coppice wood. The foreground has just been felled, leaving the stumps from which new growth will arise. In the middle is an area of 2 years’ growth since last felling. An older part of the wood forms the background. Bradfield Woods, Suffolk, England

Baceous plants. If this is allowed for, some wildwoods were much more diverse.

In some tropical forests, there is vast biodiversity, owing to the many species of tree and the abundance of vines, epiphytes, and termites, which enable the trees themselves to support complex ecosystems. In other parts of the world, continuous forests, especially of densely shading trees, are relatively poor habitats: many of their plants and animals are concentrated in gaps, cliffs, watercourses, burnt areas, and other breaks in the continuous shade.

Ancient peoples destroyed some areas of forest by various forms of land management. Continuous forest is not very productive for most human purposes. Edible animals and plants occur sparsely, if at all, and the animals are difficult to catch. Tree fruits are out of reach in the high canopy.

Forests were not generally destroyed by people cutting down the trees in order to use them. I know of no instance in European history of a forest being destroyed—converted to nonforest—solely by people using up the trees. Normally they would cut the trees for the purpose in hand and leave the other trees. The result would be a depleted forest, unsuitable for that purpose until a new generation of trees had grown. This is not to be confused with a destroyed forest.

Great trees do not easily furnish wood for fuel and timber for construction. A big tree, when cut down, is a very intractable object. Until the coming of sawmills, vehicles, and railroads capable of reducing giant trees, people preferred to use the smallest log that would serve the purpose and to manage forests to produce a succession of trees small enough to handle.

Coppicing is an important factor. Most European and North American trees, other than conifers, survive being cut down and sprout from the stump. Clonal trees, like European elms and American beech, sucker from the roots. Coppicing is one of the world’s most important practices in historic forest management. By cutting down woodland every 5–30 years and allowing it to grow again from the stumps, a permanent succession of small stems can be assured, of sizes that are easily handled and suitable for light construction and fuel. It was often the practice to leave a scatter of trees of selected species to grow on for three or four cycles to yield constructional timber (Fig. 5).

Coppicing has been the nearly universal woodland management in Britain, well documented for the past thousand years and known on archaeological evidence for some 6500 years. It is the basis of historic forest management in many parts of the broad-leaved and Mediterranean zones of Europe and also in Japan. It was apparently not much practiced by Native Americans but was widely introduced by European settlers in America, where there are large areas of ex-coppiced wood-lots.

Coppicing affects biodiversity by drastically reducing the shade at the start of each cycle of felling and regrowth. Two or three years of relatively open conditions follow, ending as the new growth closes in. This favors various woodland plants and animals. Low-growing herbs such as species of Viola and Primula flower in abundance in the years of extra light (Fig. 6). Others
such as *Euphorbia* species appear from buried seed produced by their parents at the last felling. Many insects feed on the leaves or nectar of these plants. The middle stages of regrowth, when there is a thicket of young stems, favor warblers and similar small birds—a famous English example is the nightingale (*Luscinia megarhynchos*). The dormouse (*Muscardinus avellanarius*), an English woodland mammal, favors the later stages.

Coppicing is often thought to be artificial, but the ability to coppice is widespread among the world’s trees and presumably was an adaptation to some process in wildwood. Sometimes it is a response to fire, but it is not correlated with flammability: few pines (among the world’s most flammable trees) will coppice, but fireproof trees such as elms and poplars coppice or sucker. American, European, and Japanese species of *Tilia* (lime, basswood) are self-coppicing and grow naturally in a multistemmed form; so do American and Japanese species of *Magnolia*. Possibly coppicing behavior is an adaptation to tree-breaking mammals: the axes of woodcutters are a replacement of the missing elephants etc.

I have observed something like a coppicing ecosystem developed in eucalyptus forests in New South Wales after only the third or fourth successive logging (Fig. 7).

Another historic woodland practice is the creation of permanent edges and open areas. In England, woodlots have permanent edges (many of them are over a thousand years old) defined by banks and ditches, constructed as a conservation measure (Fig. 8). They may also have permanent tracks and other open areas in the interior. In a typical woodlot, well over half the plant species are associated with the boundary, with recently felled areas, or with permanent openings. The species of permanent openings tend to constitute plant communities of their own, distinct from those of temporary clearings and of grassland away from woodland. On pollen evidence, something like these permanent open areas already existed in wildwood.

**VI. WOOD-PASTURE AND SAVANNA**

Where the environment becomes too dry or cold for forest or where there is too much grazing, forest changes into nonforest in various ways. The trees may suddenly stop, or there may be a zone of trees reduced to the stature of shrubs, or a mosaic of patches of forest and...
LAND-USE PATTERNS, HISTORIC

FIGURE 9  Savanna in middle Texas. The prairie (rich in spring and summer flowers) is dotted with motts—clonal patches—of Texas live oak and Texas elm. This is a cultural landscape of Native American origin, which was already in place when the area was settled in 1846. It needs to be maintained by grazing or burning, without which it is invaded by juniper (*Juniperus ashei*) and turns into forest. Valley Mills, Texas

patches of nonforest. Alternatively, there may be a zone, often of great extent, of scattered trees among grassland. This constitutes tropical and subtropical savanna ecosystems and their extensions into northern latitudes. The trees may be single, as in the wood-pasture ecosystems of England, the *dehesa* landscapes of southwestern Spain, and the eucalyptus savannas of Australia. Or they may be in groups, as with the oak and elm motts of middle Texas (Fig. 9). This depends on the properties of the trees: Texan *Quercus fusiformis* and *Ulmus crassifolia* are clonal trees that sucker.

It used to be thought that the “natural” savannas of Africa and Australia were very distinct from the seminatural treed grasslands and treed heaths of Europe, which were regarded as “degraded forest,” meaning forest from which some trees had been removed. Recent research makes the distinction less clear: many tropical savannas are at least partly the result of ancient land management, especially by burning. The pollen record of European wildwood, moreover, often contains non-shade-bearing undershrubs and herbs, showing that it must have had open areas as well as trees.

Cultural treed grasslands are usually differentiated from forest by there being enough browsing animals to prevent trees from occupying all the space. The amount of browsing needed to do this varies. Where the climate is so dry or cold as to be marginal for tree growth, it takes less browsing than in regions, such as England, well within the climatic limit of forest. (In some hot-climate savannas, conversely, the grasses are the dominant partners, and too much browsing encourages trees.) The limits of “natural” treed grassland are vague and controversial. The amount of browsing—the numbers and behavior of the animals—is the most difficult aspect of prehistoric ecosystems to ascertain. Trees tend to be exaggerated, because many non-tree plants produce sparse or nondiagnostic pollen grains and also because pollen analysts are trained to interpret tree pollen in terms of forest rather than of single trees.

In glacial times, parts of Europe not covered by ice were evidently dry as well as cold. The natural vegetation consisted of either patches or scatters of trees, which pollen analysts call “forest steppe.” This was the situation for at least four-fifths of the past 2 million years, punctuated by interglacials where large areas of forest became possible. In the present interglacial, when the climate became favorable for trees, non-shade-bearing plants became very local—except in the drier parts of southern Europe, where forest was still patchy and savanna-like.

When people took to keeping cattle, sheep, and goats, they replaced an unknown, but probably sparse, scatter of wild herbivores by local concentrations of domestic animals. To feed them, they turned forest into grassland; but they also kept domestic animals in the remaining forest. This involved managing the forest in a different way from where wood was the product. The trees would have to be sparse enough for pasture to grow between them on which the livestock might feed. Trees were also cropped, not by coppicing—for the animals would eat the young shoots and kill the stools—but by the practices of pollarding and shedding, cutting the tree high enough for the shoots to arise out of reach. Sometimes wood was the objective, but in climates where grass did not grow all year, it was the practice to harvest leafy tree branches to store instead of hay for feeding the animals in the cold of winter or drought of summer.

Such seminatural wood-pasture ecosystems came to occupy large areas of Europe and Asia, often associated with communal land management. In England, there was a further development with deer as semidomestic livestock from the eleventh century A.D. onward. Landowners would keep deer (native and introduced) for meat and as a status symbol in enclosed parks. The king, moreover, had the right to keep deer on certain commons, such as Epping Forest and the New Forest, and to kill and eat them. Parks and Forests often contained treed grassland on which deer and other livestock fed. There could also be elaborate systems of temporary enclosure to combine livestock with coppice woods.
FIGURE 10  A complex medieval wood-pasture, combining old grassland, ancient pollard trees, and coppice wood-lots: a particularly important juxtaposition of habitats, reproducing some characteristics of wildwood. Hatfield Forest, Essex, England

Wood-pasture is a most important ecosystem, especially for insects and birds that require both trees (for example as nest sites) and grassland (Fig. 10). The inland bird fauna of England as a whole falls into this category. Wood-pasture also can preserve ancient trees. A large proportion of the fauna and flora of trees, especially beetles and lichens, is associated with the specific habitats of old trees: old dry bark, heartwood rotted by particular fungi, and cavities of various sizes. These are specialized, poorly dispersed animals and plants and presumably existed in wildwood. Forest trees are not an ideal habitat, for the lives of old trees are often shortened by the competition of neighboring trees. Coppice woods and managed forests are inhospitable to them, unless the ancient bases of coppice stools are a suitable habitat. However, wood-pasture is often very good for old trees and the organisms that live on them. The trees are free from competition and can live to a great age, especially as pollarding prolongs the lives of many species. Most of the information so far comes from England, where the New Forest, Windsor Forest, and Sherwood Forest are major sites for the invertebrates of ancient trees.

Wood-pasture is now much reduced in northern latitudes, but I have found traces of it as far north as Sweden. The area surviving in England, though small, contains some supremely complex and important examples. It still flourishes in southern Europe, covering about one-sixth of Portugal and one-eighth of Spain, although few of these contain ancient trees. It extends eastward into the Himalayas. At one time, it was evidently extensive in Japan, where old pollard trees can still be found embedded in forests. In North America, Native Americans and settlers both operated savanna-like landscapes, but pollarding was rare. In Australia, savannas are widespread and sometimes contain trees dating from presettlement times, but often the grassland component has been displaced by introduced grasses and herbs.

Wood-pasture is an important ecosystem for biodiversity: it perpetuates a different set of plants and animals from coppice woods. The most significant examples remaining are those few in which ancient trees and seminatural grassland are both preserved.

VII. GRASSLAND

Many of the world’s grasslands are regarded as natural prairies, in regions too dry, or too cold, or too much grazed for trees. Others are part of a cultural landscape in regions that would naturally be forested if not grazed, they revert to forest. There are large intermediate areas where natural grassland has been extended by burning and other forms of land management.

Much grassland in Europe, America, and Australia is sown pasture, little different from other arable crops. This practice goes back only three centuries. Previously, grassland was permanent, managed as either pasture (grazed) or meadow (mown for hay). Pasturage often goes with communal land use, as in the American Wild West or the chalk downlands of England.

Seminatural pasture grassland in England goes back to the Neolithic period, when the forests that covered the chalklands were converted into open land; prehistoric monuments such as Stonehenge were meant to be visible from a distance. Often there was a phase of farmland before the historic grasslands were established.

Meadow is a more productive kind of grassland, yielding hay which is dried and stored to feed animals in winter when the grass is not growing. This requires iron tools with which to cut the grass. As a large-scale land use, meadow is mainly of the past 1500 years. It was located especially in floodplains, which in medieval England were the most valuable land. In the sixteenth century, elaborate irrigation systems were developed to advance the growth of meadow in spring. Grasslands were often part of a highly integrated pattern of land uses that transferred nutrients, via the dung of cattle and sheep, to plowland and meadow.

A contrasting type of hay meadow is that on mountains, for example, in northern England or in the Alps. In the northern Apennines, the grasslands, originating in the Iron Age, were maintained for centuries to feed
local livestock and those of the city of Genoa; recently, they have largely turned into forest and their distinctive plant communities have been lost.

Old grasslands are the classic "seminatural" vegetation. They are artificial in that they were created by human activity, and disappear if that activity is withdrawn. They are natural in that nobody sowed the grasses and other plants of which they are composed. They are of many kinds, depending on whether they were pasture or meadow, on the geology, and on which animals fed on them. Chalk pasture is a dry, relatively infertile form of grassland that is one of the richest plant communities (in terms of plant species per square meter) in Britain. When grazed by sheep or rabbits, it is particularly rich in dwarf herbs and small undershrubs. At the other extreme, wet meadow is rich in tall grasses, sedges, and orchids.

VIII. HEATHLAND

Heathland in northwestern Europe is open vegetation other than grassland, dominated by dwarf undershrubs especially of the family Ericaceae. It can be a very rich plant community, with many herbs and also characteristic butterflies and birds. It is a seminatural ecosystem, maintained by grazing and cutting; without these, it quickly turns into woodland (Fig. 11).

Heathland began in a small way in the Mesolithic period, about 10,000 years ago; people may have encouraged it as part of land management in favor of deer. It expanded in later prehistory, especially as browsing by domestic livestock encroached on forest. Throughout the historic period, it was extensive in England, especially on acid soils in low-rainfall areas.

Heathland should perhaps be seen as a northward extension of the undershrubby plant communities of the Mediterranean, known as garrigue and phrygana. These, too, are often thought of as the result of destruction of forest; on palynological evidence, they were relatively local and specialized ecosystems until the Neolithic period. However, they are more stable than heath, occurring in places too dry for tree growth. Often they fail to turn into forest if grazing is removed. Phrygana is one of the richest ecosystems in endemic plants.

The heaths of northwestern Europe have fared very badly in the past 200 years. Little remains outside Britain, and even there, heaths, where not made into arable fields or forestry plantations, have been allowed to turn into woodland.

IX. MOORLAND

Moorland, although often confused with heath, is really a southward extension of the peaty tundra of the Arctic. It is characteristic of northwestern Europe, especially mountains (Fig. 12); it still covers half of Scotland.

Moorland is dominated usually by Ericaceae undershrubs, especially heather, Calluna vulgaris. Unlike heath, it occurs in high-rainfall areas and has a peaty...
substratum. It grades into blanket bog, with pools and abundant Sphagnum.

Moorland is more stable than heath and less obviously a cultural ecosystem. It gradually extended during the Holocene, covering areas that once had forest. Sometimes there was an intermediate stage of pasture or even arable land: on the plateaus of Dartmoor, southwest England, the boundaries of Bronze Age fields can still be traced. However, it is not clear how far the spread of moorland was due to the lapse of time and accumulation of peat, to a change to a wetter climate encouraging peat to grow, or to human activities tipping the balance against forest.

In historic times, moorland has been much used for pasture and for digging peat as fuel. Large supplies of peat have made up for the scarcity of wood fuel in Scotland and Ireland. A skilled tradition of periodic burning has grown up to maintain the pasture. In the past 200 years, moorland regions have lost human populations, but without forest returning. The main uses of moorland now are for grouse (Lagopus scoticus) and deer as game animals.

X. BUILDINGS AND BUILT-UP AREAS

Many plants and animals live on buildings, including some quite rare species: for example, the stork in Europe, which nests almost exclusively on buildings and structures, sometimes on platforms provided for it. In England, buildings and abandoned mines are the chief strongholds of bats.

Cities are often the chief concentrations of wildlife in their regions: well-known examples are Cambridge and Glasgow, which contain more wild plant species in a given area than anywhere else in their surroundings. This is because of the diverse habitats provided by gardens (despite the prevalent use of garden chemicals), buildings, and especially ruins and derelict land.

Derelict industrial land can be an important habitat. In England, the quarry that provided the excellent Barnack building stone is now one of the best examples of limestone grassland. Long-standing chemical contamination can produce its own ecosystems, especially where heavy-metal-tolerant plants have evolved.

Sacred sites develop in peculiar ways. In Japan, temples and shrines often own forests, which may preserve ancient trees and evidence of past management practices. In England, many churchyards are excellent habitats (Fig. 13). They often contain their own peculiar types of grassland. Tombstones and the stonework of the church itself are important sites for rare lichens, especially in areas that lack natural rock outcrops.

XI. THE RISE AND FALL OF HISTORIC LAND-USE PATTERNS

Human intervention creates new ecosystems as well as destroying old ones. Cultural ecosystems are often complex, diverse, and stable. Sometimes they are modifications and extensions of ancient natural ecosystems. chalk grassland, for example, is in some respects a prolongation in time of the grasslands of the glacial periods. Moorland is in some respects an extension of tundra into southern latitudes. Some, however, seem
to be new creations: I hesitate to guess what the habitat of the plant *Fritillaria meleagris* might have been before hay meadows had been invented.

Some conservationists regard ancient human intervention as destructive, like the modern logging of the Oregon redwoods or the early-modern ravaging of oceanic islands. Sometimes this was so: we would undoubtedly now disapprove of the destruction of almost all the native mammals of Crete by its early human inhabitants. Some interventions were experimental and unsustainable, like the prehistoric attempt to farm Dartmoor. Others may have been almost too gradual to notice, such as the erosion of tree cover on a common pasture.

The ecosystems of cultural landscapes could be preserved for centuries, both by ownership patterns and by deliberate conservation. By A.D. 1350, some 94% of the forest of England had been converted to nonforest uses. Woodland management had been invented long before, probably for other reasons, but as forest became scarce, conservation practices gradually spread to the remaining forest. It may be no coincidence that the idea of sustained yield is first set down in writing in 1356. Many medieval wood-lots still exist, and are often now nature reserves: "Ancient Woodland" is a scheduled category in modern conservation. Historic land-use patterns usually worked on a small scale to result in useful juxtapositions of habitats. They often perpetuated habitats for non-shade-tolerant plants and animals. Because of the glacial history of the past 2 million years, many species are adapted to nonforest conditions and are threatened by the advance of forests in interglacial periods. They would have fared particularly badly in the present interglacial, because of the widespread extinction of giant trees-breaking mammals. For example, the greatest concentration of endemic plants in Europe is in the Mediterranean, where nearly all of them are nonforest species.

Historic land-use patterns have been in retreat for some 230 years. Usually this was because of outside ideas and interventions, not because they were unsustainable. Commonlands were not destroyed because of the operation of the "Tragedy of the Commons, the supposed tendency of communal systems to break down because each participant pursues his own short-term advantage regardless of the rights of other participants. That had already been foreseen, and the participants in most commons had drawn up rules of use to prevent it.

In the eighteenth century, multiple and especially communal land uses became unfashionable. The only proper use of cultivable land (or land thought to be cultivable) was conventional agriculture, carried on by private owners in rectangular plots of not less than 10 acres. For uncultivable land the proper use was timber production organized by the state on the model of either French or German "scientific" forestry (depending on which state). Agricultural writers invented these ideas and bent the ear of governments to get them put into practice; they were not deterred if their schemes failed to work. In this way, the open fields of England were destroyed, and most of the heathland was converted into poor-quality arable. In countries where modern forestry was influential, such as Sweden, Germany, and Italy, most of the wood pasture ecosystems were destroyed.

In the twentieth century there have been four contrasting processes: extension of cultivation (including plantation forestry) into areas not previously cultivated; intensification of cultivation in areas already cultivated; urban development; and land abandonment. Often, as in Crete, all four occur almost side by side. They destroy existing seminatural ecosystems without creating an equivalent. A common feature is that the grain of the landscape is coarsened: the small-scale pattern of juxtaposed land uses is replaced by monotonous expanses of the same land use. One consequence is loss of habitat for creatures that require more than one habitat: birds that feed in the open but nest in tree holes; insects whose larvae feed in rotten wood but whose adults need a nectar source; plants that are weakly competitive when growing in the open but do not flower in shade—all these were favored by the mosaic of historical land uses, but not by what has replaced them.

Land abandonment does not re-create wildwood (wildwood either as it was before human intervention or as it would be by now had that intervention never happened). It tends to produce uniform expanses of even-aged, densely shaking trees: the herbaceous plants of grassland, field edges, etc. disappear without being replaced by woodland herbs. In Mediterranean Europe, where secondary forest often consists of fire-adapted trees, the effect is a recurrent cycle of fires, which destroy whatever escaped the increasing shade. Many conservationists disapprove of goats, but the consequence of removing goats is usually to convert a browsing-dominated landscape into a fire-dominated one (Fig. 14).

Land use has become polarized: for example, land is now either forest or pasture but not both. Most of rural Japan, for example, is forest, and the remainder is intensively cultivated rice fields. The forests are now little used, and within them are the remains of many
land-use patterns, historic ecosystems: coppice woods, wood-pastures (Fig. 15), terraced rice fields, pine savannas, Japanese early-modern forestry, and the growing of the giant grass with which house roofs were thatched.

Forests, on the whole, have become denser as savannas infill, coppice woods grow up, and foresters encourage the growth of timber trees. The older trees in a forest are commonly more spreading than the younger ones, having grown up when they had more room. I have seen examples in many countries from Canada to New South Wales, often in forests with little obvious human intervention.

An important innovation is the cult of tidiness: the urge among professional planners to destroy old quarries, spoil heaps, ruined buildings, patches of roughland, and even old trees and neglected tombstones, on the grounds that the public might not like them—even though, in practice, derelict land is often a much-loved public amenity.

XII. IMPLICATIONS FOR BIOLOGICAL CONSERVATION

Some conservationists disdain cultural ecosystems on the grounds that they result from the ‘degradation’ of wilderness ecosystems, especially forest. The only proper business of biological conservation (they say) is to preserve pristine natural ecosystems where they still exist or to restore them where they do not.

However, this cannot be the whole purpose of conservation. In many parts of the world, such as England, ecosystems have been modified by human activity for much of the Holocene; not enough is known about pristine natural ecosystems to make restoring them a practical objective. Conservation depends on public interest, which it cannot retain if it is limited to remote wildernesses that few of the public ever see. It is unrealistic and impolitic to cut off biological conservation from other kinds of conservation: to separate the archaeological interest in an earthwork from the biological interest in the vegetation growing on it. It is essential for students to gain a knowledge of how local ecosystems work before pronouncing on those of distant countries.

Wilderness philosophy often ignores the realities of archaeology or vegetation history. In many countries, when a national park is scheduled, the authorities automatically expel the human inhabitants (however long and respectable their history), play down their part in the development of the landscape and ecosystems, and pretend that the park is wilderness. In Yellowstone National Park they even tried to efface all evidence that settlement had ever happened. In reality, few national parks lack human influence altogether; many of them are the last strongholds of precisely those old-fashioned land uses that biological conservationists ought to support. (The National Trust, biggest landowner in the Lake District National Park in England, goes to much trouble to sustain farming and prevent land abandonment.)

Conservationists are tempted to start from theories of what natural ecosystems ought to be and to re-create them in the image of those theories. They regard large, continuous, mixed-aged forests of tall trees as natural and try to preserve those forests that most nearly ap-
LAND-USE PATTERNS, HISTORIC

approach that ideal. The notion owes more to climax theory and to foresters’ notions of an ideal forest than to weighing the evidence of what particular natural forests were like; yet its influence leads to neglect of discontinuous forests, savanna, and forests of short trees, however remarkable.

Biological conservation has many aspects. Priority is rightly given to preventing logging of remote forests that have never been logged, whether or not later research shows that they are really wilderness. However, the real merits of cultural ecosystems should not be overlooked. Wildwood included a great diversity of ecosystems. Those that involved open ground often predominated during glaciations. When game herbivores were exterminated, this probably made the forests more shady, and the survival of non-shade-bearing plants through interglacials more difficult. Historic land management, to some extent, supplied the place of the missing elephants and restored the habitats of savanna and open ground. Modern changes tend to destroy historic ecosystems altogether or to replace them by uniform, very shady forest.

Historic land-use patterns need to be respected and historic cultural ecosystems maintained. Conservationists rightly campaign against destroying trees in tropical rain forests. They often need to campaign against allowing trees to grow in unsuitable places near home.

Acknowledgment

Some of the plates are the result of travels with Mr. A. T. Grove, Dr. D. Lunney, Professor J. A. Moody, Professor Jun-ichi Ogura, Dr. J. Sleath, Professor Kathuyoshi Tsuchida, and the Sphakia Survey.

See Also the Following Articles

AGRICULTURE, TRADITIONAL • CATTLE, SHEEP, AND GOATS, ECOLOGICAL ROLE OF • ECOLOGY OF AGRICULTURE • FOREST ECOLOGY • HUNTER-GATHERER SOCIETIES, ECOLOGICAL IMPACT OF • INDIGENOUS PEOPLES, BIODIVERSITY AND • TRADITIONAL CONSERVATION PRACTICES

Bibliography


I. Introduction
II. Deforestation and the Living Dead
III. When Is a Tree Not Living Dead?
IV. What of Small Plants?
V. What of Animals, Those Things That Move?
VI. And What of Those Things That Eat the Living Dead?
VII. Are There Living Dead Habitats and Ecosystems?
VIII. Restoration Biology

GLOSSARY

agroscape The agricultural, ranching, and plantation countryside, with its roads, irrigation ditches, buildings, and so on. The agroscape stands in contrast to the wildland countryside that is not directly managed by humanity (though it is strongly impacted by it). The agroscape intergrades with wildlands in the form of woodlots, abandoned fields, poor soil sites, hedge-rows, and edges of wildlands.
living dead An individual stripped of the ecological circumstances that allow it to be a reproductive member of its population, but which is living out its physiological life. Living dead are most easily observed as large trees remaining on the agroscape, but they are also present in natural ecosystems.
megafauna Large mammals that are wolf-sized, deer-sized, and larger. Commonly used in reference to the many species of extinct “Pleistocene megafauna” that 9000 years ago populated the New World. The elimination of this megafauna by hunting (of the herbivores) and starvation (of the herbivore-deprived carnivores) was probably the first, and certainly the most dramatically irreversible, of the anthropogenic macroalterations of New World ecosystems. Today, of the extinct Pleistocene megafauna, only the horse remains—evolutionarily invented in the New World but surviving in the Old World until brought back as a gift from the Pleistocene by Spanish soldiers.

TREES AND MANY OTHER organisms that dot the tropical agroscape are often living dead. These are those individuals that have been stripped of the ecological circumstances that allowed them to be a reproductive member of their populations but are living out a physiological life. Living dead are most easily observed as large trees remaining on the agroscape, but they are also present in natural ecosystems.
psychology of tropical conservation because their presence obfuscates pending extinction. But living dead are also primary elements of natural processes of local extinction, immigration, and population-community structural dynamics in response to short- and long-term environmental change, be it natural or anthropogenic. Extra-tropical habitats and extreme tropical ecosystems may have fewer living dead than do complex tropical ecosystems, but they are nevertheless present. As magnificent as the living dead may be on the tropical countryside, I suggest that we not be distracted by attempting to save them, but rather that we focus our conservation efforts on saving large blocks of wildland ecosystems that are relatively complete and (it is hoped) relatively poor in living dead.

I. INTRODUCTION

The idea of the living dead has gradually emerged in my ecological understanding as I have lived past and around the majestic forest giants left standing as the agroscape creeps into Costa Rica’s forest ecosystems over the past 4 decades (Fig. 1 and Janzen, 1986a, 1986b). This creep gradually converts the forest to an agroscape of pastures, fields, and roadsides dotted with the occasional adult tree but few or no juveniles. This is an agroscape where a magnificent flower crop now stands bee-less, an agroscape where fruit crops lie rotting below the pasture tree, an agroscape where tree seedlings wither in the dry-season sun or are turned to smoke in the dry-season anthropogenic fires.

I begin this article with a focus on adult large trees and use familiar examples from the Costa Rican countryside. To create breadth, I suggest that you join these verbs with the nouns from the ecosystems you know. This is a conservation biology question, but it applies to more than that, and it applies across the once-forested tropics as well as elsewhere.

Looking across the tropical landscape, the eye is greeted by stately single trees (Fig. 2), by patches of forest, by the blaze of a colorful flowering episode. Put an inventory to the plant species in a field, in a valley, in an ecosystem. All these species appear in the list. All is more or less well, we conclude, as 96.4% of the species that were here 50 years ago are still present. But are they? How many of them are living dead, part and parcel of latent extinctions?

We live a perceptual lie as we bustle about our agrosapes. That single stately green Dipteryx or Hymenaea or Swietenia or Enterolobium, standing in a field, pasture, or roadside, is often just as dead as if it were a log in the litter or the back of a logging truck. That tree was birthed in some favorable circumstance, a circumstance for pollination, seed dispersal, seed germination, and sapling survival.

FIGURE 1 Living-dead trees isolated in pasture at the edge of the agroscape (background) as it creeps into old growth forest (foreground): Los Naranjos, Sector Cacao, Área de Conservación Guanacaste, July 29, 1987.
LATENT EXTINCTION—THE LIVING DEAD

FIGURE 2 A living dead *Terminalia* tree stands in silhouette, left behind as the rain forest was cleared around it, the natural tree falls in which its seedlings might have survived long since removed. Rincon Rainforest, Area de Conservación Guanacaste, January 6, 2000.

But one or more of these circumstances is now gone. It was carried away with the forest, put on the hunter’s table, pesticided out of existence, or global warmed into oblivion. The long-lived tough adult lives out its physiological life, in the absence of the carpenter with a chain saw, but it is evolutionarily dead. Its pollen no longer flows to other members of the population, its seeds are no longer carried away from seed predators, or its seeds are no longer carried to a favorable site for seedling growth and sapling survival to adulthood.

But because the adult lives on, we are lulled into thinking that the environmental damage really is not all that bad, that extinction has not already occurred. If we can still show the tree to our children, it seems not to be extinct. It is so big and green and strong. Every year we see its flowers, and maybe we even see its fruits on the ground below. And after all, it has clearly weathered all that we have thrown at it. What ever can the matter be?

Humanity’s interaction with the world’s ecosystems has an enormous perceptual element. We act on what we perceive, be it threat or opportunity. Much of our conservation pragmatics and understanding is based on our knowledge that we really are losing species, losing ecosystems, losing the capacity of the environment to absorb our footprints. But that knowledge comes from what we see and measure. If all members of a tree species were to have the trait that each abruptly falls over dead the moment that it ceases to be a reproductive member of its population in its ecosystem, there would be far stronger alarm cries across the tropics about extinction rates and realities. If trees, the largest organisms on most of our landscapes, were very short lived as compared with humans, there would be less of a perceptual problem—though just as large a conservation problem.

When the terrestrial world was covered with forest ecosystems, the single tree left standing in an aboriginal cornfield may well have been living dead, but the population from which it was derived was not usually at risk of anthropogenic extinction, unless perhaps dependent on a seed disperser targeted by that aboriginal population (Janzen and Martin, 1982). But when the agroscape is dotted with living dead in the wake of contemporary omnipresent ecosystem alteration, latent extinction is very real. A tree species may be ranked as “common”—meaning visible from a car window along many roads—yet be effectively extinct in a county, state, or region. And since the agroscape now stretches from horizon to horizon, the plant may well be absolutely extinct, since all of its former range may be populated by living dead.

II. DEFORESTATION AND THE LIVING DEAD

The forest need not be removed to convert trees to living dead. It is just that when the forest is partly removed, there is a very high chance that this alone will ecologically deprive many individuals of the remaining tree species sufficiently to convert them to living dead status. And, it certainly leaves the living dead very visible.

But even when the forest is left in place, that is no guarantee of a healthy tree population. When the Pleistocene hunters and their carnivorous helpers hunted out the neotropical mastodons and gomphotheres, the glyptodonts and camels, the ground sloths (Janzen, 1983b; Janzen and Martin, 1982), they did not do it by forest clearing. For decades to millennia after this 9000-year-old event, many of the remnant individuals of the tree populations that these big mammals
dispersed (Fig. 3), and for which they created safe sites for seedlings by their browsing and trampling, would have been living dead scattered in the forest.

If some particular species—a pollinator or dispersal agent, for example—in the forest is extinguished, by whatever cause, there will often be surrogates and alternates that will assume, in some form, some portion of the "role" of the extinguished mutualist. The tree species will live on, albeit in some other ecological morph, and therefore in some technical sense will not be extinct. The tree that was "dependent" on the extinguished species will not, then, be living dead. But the devil is in the details. We need to go case by case. The suite of interactants with a tree species generates a given seed shadow, pollen rain, sapling demography, and microgeographic distribution. Remove one species of interactant. The entire n-dimensional hyperspace shifts in this or that direction. In some places this is toward eventual extinction, in other places it is just a change in demography and microgeographic distribution. The history of any surviving species is that it must have survived thousands of such handoffs from one mutualist to another, from one moment to the next (e.g., Hallwachs, 1986). What bumps individuals into the category of living dead is the serendipitous event of losing irreplaceable partners. Humanity has a way of removing not only partners, but whole suites of them, as well as altering the physical environment. Our thoroughness and omnipresence creates ecological irreplaceability. Yes, when we lose one ground sloth, a glyptodont picks up some of the slack, though the tree is now a different beast. And at some time, likely as not, some new slothoid arrives by evolution or immigration over the millennia. But lose all these big mammals at once, and the result is guaranteed to be large arrays of living dead.

We have all been nourished by the marvels of evolutionary understanding, leading to the temptation to wonder if rapid evolution will not resuscitate a living dead population, if not many of its individuals. Novel pollinators, dispersal agents, fruit morphology, flowering phenology—all could save the day. In theory yes, but in reality not on the timescales ordained by humanity’s charge across the landscape. How long will it take to evolutionarily renvent a neotropical herbivorous/frugivorous megafauna? Fracture the remaining forest, with its living dead, into small ecological islands (also known as national parks and reserves). Thereby create ideal circumstances for rapid and novel evolution. We still cannot expect natural selection to create a mastodon from a white-tailed deer in anything like the speed required to be an antidote for neotropical rain forest anthropogenic alteration, beginning with the megafauna extinctions.

Certain kinds of habitat destruction are compatible with some tree natural histories. Two common trees, the guanacaste (Enterolobium cyclocarpum, Fabaceae) and jicaro (Crescentia alata, Bignoniaceae), owe their
FIGURE 4 An earthbound extinct megafauna returned from the Costa Rican Pleistocene by Spanish immigrants, breaking a *Crescentia alata* fruit (Fig. 3) to eat the molasses and seeds inside. Sector Santa Rosa, Area de Conservación Guanacaste, 1980.

contemporary prominence on the Mesoamerican Pacific coastal landscape to a particular kind of habitat destruction. For both, large mammals—such as free-ranging horses—swallow the seeds while eating the content of indehiscent fruits fallen below the parent tree (Figs. 3 and 4), and later defecate them in open sunny habitats (Janzen, 1981, 1982a, 1982b). Forest clearing unto brushy pastures and scraggly roadides, populated by widely circulating working horses, maintains a healthy population of reproducing guanacaste and jicaro trees in a precarious balance with humanity.

What did these trees do before the Spaniards brought the horse back from its Old World refuge after its neotropical extinction by Pleistocene hunters (Janzen and Martin, 1983)? They probably survived in a peculiar habitat characterized by ample insolated ground yet sufficient rain for there to be large trees and sloppy seed predator rodents (or human fruit and seed harvesters), which offered sufficient seed dispersal. River edges, marsh edges, and the interface between tropical dry forest and desert are such habitats, and the aboriginal village/field edge adds a serendipitous fourth. The Spanish working horse (Fig. 4) found the fruits abandoned by their extinguished ancestors and spread these two trees so thoroughly that today they are viewed by Mesoamerican societies as native and natural. And, in the case of *Enterolobium cyclocarpum*, cattle are surrogate horses (Janzen, 1982a).

However, as the motorbike and car replace the horse today, and as the cattle industry fades, these two trees are left as very visible living dead scattered across the former ranch lands, their abundant fruits rotting below the parent tree, the newly germinated seedlings killed by fungal pathogens nourished by the annually replenished seed crop, and the rare escaped seedling killed by herbicides, grass fires, and cosmetic cleansing.

III. WHEN IS A TREE NOT LIVING DEAD?

Earlier I noted that if each member of a tree species were to abruptly fall over dead the moment that it ceases to be a reproductive member of its population in its ecosystem, there would be far stronger alarm cries across the tropics about extinction rates and realities. However, the isolated tree, left an adult in the open as the forest is mined away from around it (Fig. 2), is not necessarily or automatically a member of the living dead, or at least not necessarily at that moment. At least two circumstances may help to avoid this label. First, the pollinator community and the seed dispersal community for that tree may still be of a structure such that they confer sufficient amounts and patterns of their services and do so with the new reproductive phenology that will be expressed by the tree in its “new” habitat. And males do have fitness. A plant may never set a fruit or never have a surviving seedling from its seed crop, yet it still may be very much a member of the reproducing population (e.g., Aldrich and Hamrick, 1998). Plants contribute pollen “outward” as well as receive it from unseen members of the population. There may be some circumstances where this or that member of the pollinator guild will in fact carry pollen from that isolated tree back into the forest. At least potentially this may remove the living dead label.

Second, the new pattern of seed/seedling/sapling safe sites for that species may be sufficient for population survival, even if different. A novel demography, reproductive phenology, and microgeographic structure will ecologically emerge, reflecting the serendipitous matching of the tree’s traits to these new conditions.
For the survivor, such ecological fitting (Janzen, 1985) of an individual (or a population) into the environment newly thrust upon it is the same process as occurs when a tree species is anthropogenically introduced to a new place. Whether introduced by humans or by natural processes, its survival there demonstrates that it has ecologically fit in. Such introduction may occur into a natural ecosystem or one variously anthropogenically perturbed. Sloppy deforestation may create many living dead; only mildly impact some other species, and favor yet new introductions into the region by having removed competitors or consumers.

A population of plants in a newly altered landscape is not necessarily at a given moment either “dead” or “alive.” Just as the relationships of an individual to its ecological circumstances may decay slowly, it is also easy to visualize a population being sufficiently anthropogenically impacted that it gradually decays over several decades-to-centuries-long generations. This state of decay is an intermediate between living dead and “normal surviving.” The portion of a population of trees at some geographic point may be in a constant state of swinging between being “okay” and living dead, as its associated climate and community of interactors goes through their own changes.

A species’ population in its totality may also be waxing or waning in geographic coverage, density, “living deadness,” or all three. Living dead are found at the geographic or demographic margins of all populations. It is just that human activity in ecosystem modification (elimination, simplification) simultaneously impacts so many species, and is so omnipresent, that it creates large numbers of living dead in the same place at the same time. These then carry the tragic perceptual load of tricking us into thinking that all is much more well than it actually is.

But ecological neutering, expressed as here in the terms “living dead” or “latent extinctions,” is not restricted to the circumstance of the single tree in the field or a single portion of a population. The living dead are an integral part of natural age-structured mortality. Any field biologist can identify a large number of young individuals—seeds, seedlings, saplings—that have a vanishingly small chance of survival as individuals. The forest understory is densely populated with them, as is each squirrel’s winter seed cache, as is the patch of seedlings below the healthy parent tree, as is the ground covered with epiphyte seeds that fell past the branches of the trees above, as is the floor of the cave littered with bat-dispersed seeds. A very large part of the world’s herbivore machine is run with this fuel and actually should be labeled “detritivore” rather than herbivore.

The implications for evolutionary biology are huge, given that no matter how much herbivory occurs on these living dead, there can be no natural selection inflicted on the food populations.

Living dead adult individuals are also a prominent part of many undisturbed habitats and ecosystems. These are the wafers, the strays. Each of these is a plant whose seed arrived, grew to an adult, but found itself in a place lacking whatever is needed to maintain a viable population (Janzen, 1986c). In complex interwoven tropical habitats and ecosystems, the species list in a given place may contain as many as 10 to 20% of these kinds of living dead. For example, if a valley-bottom forest is eliminated, over time a significant number of tree species may disappear from the adjacent ridge, not because of any direct impact on the ridge forest but because the portions of the populations that were there are no longer maintained by seed flow into them from the valley bottom. This phenomenon is particularly visible where a particular soil or slope is thoroughly cleared for a crop, and the natural vegetation is left relatively intact in a neighboring habitat, ostensibly to protect it. Some species disappear because the conserved habitat did not really have its mutualist animals and physical climate conserved, or because it is too small, but others disappear simply because they were naturally occurring living dead.

Not to belabor the obvious, a tree standing dormant in the tropical dry season is not reproducing in the narrow sense, but it is also not necessarily living dead. But this is tricky for the observing human. We are very accustomed to being around trees that are not, at that moment, undergoing anything that appears to be reproduction, yet are members in good standing of quite surviving populations. The living dead tree does not display anything much different at first glance. Recognition of living dead status requires in-depth knowledge of its activities over decades, requires knowing if and where its pollen is going, and requires knowing where its seeds are moving to and what happens to them when they get there. This understanding is not acquired with the casual glance (e.g., Aldrich and Hamrick, 1998; Curran et al., 1999; Hallwachs, 1986).

IV. WHAT OF SMALL PLANTS?

The isolated tree in the pasture has been a convenient illustrative example, but the world to which these ideas apply is far greater than that of large tropical trees. A small herbaceous plant may be a perennial with longevity like that of a tree. When the euglossine bees are
extinguished through forest partial clearing, a Cat-
senam orchid they pollinated is left high on the main
trunk of a shade tree left behind, a living dead in its
own right. The orchid may flower for a century, waiting
in vain for its long-distance pollinators (Janzen, 1974).
They are long gone, their year-round nectar and pollen
sources turned to charcoal. A living dead clump of
perennial grass on a landslide scar may for many de-
cades produce its small hard seeds, designed millions
of years ago for a trip through a seed-dispersing, now-
extinguished, large herbivore to a new disturbed site
(Janzen, 1984). It finally succumbs to its individual
sterile fate as the landslide scar revegetates to forest. A
living dead herbaceous morning glory (Convolvula-
ceae), sprouting and flowering year after year into the
insolated roadside ditch from its underground tuber,
may never again see the bees that once moved among
its flowers and the flowers of the many other forest-
edge species that once sustained them (e.g., Frankie et
al., 1998).

But as mentioned earlier for a population of trees,
even a population of annuals may also be a living dead
distribution map not of what is today on the Costa
 Rica, Nicaragua, Guatemala, Veracruz, and San Louis
Potosi lull one into thinking “surely over that huge

V. WHAT OF ANIMALS, THOSE THINGS
THAT MOVE?

Reproduction—that is, membership in the popula-
tion—has two components. On the one hand, it is self-
evident that the individual needs to be physiologically
able to reproduce. On the other hand, if it is ecologically
neutralized, it is as dead as if sliced off with a chain saw.
Selection has not generally favored the ability of a tree
to “know” that it has been ecologically neuterized by the
removal of its pollinators, its dispersal agents, or the
safe sites for its juveniles, and then take remedial action.
What would the mutant tree have to be able to do?
Walk back to the forest? Animals, with their chance
to move to a new ecological circumstance, get horny. They
search for nesting sites and mates, they may fight harder
for their surviving fewer children, or they may migrate
or emigrate to other places. But, in the face of the
sweeping and omnipresent hand of humanity, busily
extending its extended genome to cover the globe with
both people and their domesticates (Janzen, 1998),
where is the potentially living dead animal to go, and
how long does it have to get there? One can search
only so long before dying of old age, becoming a road
kill, or running out of stored food reserves.

The tropical agroscape, and most wildlands as well,
are awash with living dead animals, animal populations,
and animal arrays (also known as “communities,” what-
ever those are). Latent extinction is everywhere, but it
operates more rapidly on animals with their high turn-
over rate and their lower capacity for extended lives as
ever those are). Latent extinction is everywhere, but it
operates more rapidly on animals with their high turn-
over rate and their lower capacity for extended lives as
humans contributing in a curious perceptual manner
to us being less aware of the animal living dead. At
the level of the large animals, “everybody knows” that
jaguars and tapirs are still “here” because everyone
knows someone who knows someone who saw one
once. One sighting of one 10-year-old jaguar crossing
the road at noon 12 years ago will sustain the living
dead jaguar in that area for decades, long past its con-
signment to the litter. It has taken more than three
decades for the myth of Costa Rican giant anteaters,
which once ranged these forests, to die a natural death.
Collectors and collections do their part as well. There
is a snapshot of history present in our museum drawers,
each specimen with its neat locality label. These collec-
tions continue the illusion of survival long past the
reality. Retroactive data capture from museums gives a
distribution map not of what is today on the Costa
Rican countryside, but rather what once roamed where
today sweeps unbroken waves of sugarcane, pasture,
plantations, and horticulture. Intellectually every tax-
onomist knows this, but the orderly march of specimens
across the museum drawers that read Panama, Costa
Rica, Nicaragua, Guatemala, Veracruz, and San Louis
Potosi lull one into thinking “surely over that huge
geographic range there are still viable populations. Plants are not immune to these processes. It is just that with the more illusive, the shorter lived, the more mobile, the animal living dead may be more easily manifest in historical collections than on looking out the car window at 70 kmph.

And, when one descends from a field vehicle somewhere, a rare butterfly flutters from the museum drawer and down the roadside ditch, the cruel illusion is reinforced. Highly mobile animals are particularly effective at hiding the living dead from perception. The last living dead Costa Rican green macaws will fly across the countryside for decades. One small viable population of butterflies can create hundreds of living dead individuals searching across the food-plant-free agroscapes until dying on windshields, of pesticides, or in the collector's net.

Some animals, like some plants, thrive in the agro-scape. Are they living dead as well? The agro-scape changes its biotic and its physical traits at the whim of some combination of the market and our technical ability to (re)engineer our domesticates (and produce new ones). Overnight the agro-scape can flip from heaven to hell for a particular species. When cotton was the crop of choice on the Costa Rican countryside, the world was an ocean of food for native Dysdercus cotton-stainer bugs (as well as for a number of other native cotton herbivores). The local extinction of the bugs' original wild food plants (Malvaceae, Sterculiaceae, Bombacaceae) that accompanied the forest clearing for crops that are destined to fall and rot below the parent tree, may be living dead as well. But the extinction of animal dispersal agents and safe sites for juvenile plants goes unheralded, not living dead. But the extinction of animal dispersal agents and safe sites for juvenile plants goes unheralded, with not even a potential feedback loop.

VI. AND WHAT OF THE THINGS THAT EAT THE LIVING DEAD?

All have their predators, their parasites, their mutualists, their scavengers. Many of these are quite dependent on the traits of their hosts. Food is not food if it is not food. Narrowly host-specific specialists abound.

For every living dead individual, population, or species, there is a large suite of consumers—individuals, and even species—living at the margin of their existence. A seed predator weevil—Rhinochermes stigma—passes its larval stages in the pods of guapinol (Hyptis meara courbaril) on the Costa Rican countryside (Janzen, 1974). It maintains what appears to be a healthy population in the annual supra-anual fruit crops that are destined to fall and rot below the parent tree in the absence of both the Pleistocene megafauna and the agouti (Dasyprocta punctata), contemporary inheritor of the guapinol (Hallwachs, 1986). But as each of those old guapinol trees dies at the end of its 200 to 500 year life span, the weevil population takes another hit. One day the last living dead guapinol trees will die, and along with them will go what appears today to be a perfectly healthy community of weevils.

The guapinol is also fed on by leaf-eating caterpillars. One, a large saturniid, Schausia santarosensis, eats only guapinol leaves and will go the way of the Rhinochermus weevil. Another, Dirphia avia, also a large saturniid, feeds also on the foliage of Spanish cedar (Cedrella odorata), mahogany (Swietenia macrophylla), oak (Quercus oleoides), and guarea (Guarea excelsa) (Janzen and Hallwachs, 2000). As the adult guapinol trees dwindle in number, how the Dirphia avia population will respond and change will depend in part on how many individuals of the other living dead remain. (You guess: How many Spanish cedar, mahogany and oak trees will be left standing by the Costa Rican roadside?) Perhaps Guarea excelsa, its wood of no commercial value, will be the only host plant left. Enough to sustain Dirphia avia? Who knows, but it certainly won't be the same moth population that it was before.

The flowers of the living dead Andira trees were once a primary food source for tens of thousands of individuals of hundreds of species of bees; today they are visited by only a pale shadow of this bee community (Frankie et al., 1998). But those old adult Andira con-
FIGURE 5 A living dead patch (left center) of natural vegetation, composed primarily of living dead individuals, among rice fields. There is essentially no gene flow between the patch and the secondary successional wildland in the foreground despite the thin connecting strip of riparian vegetation. Southwest of Liberia, Guanacaste Province, Costa Rica, December 14, 1999.

continue to produce their massive flower crops and will do so for many decades to come. Its copious fruits, now largely from pollination by domestic honey bees, lie rotting below their parents in the absence of the masses of frugivorous bats that once dispersed them (Janzen et al., 1976).

As noted earlier, the living dead are a "natural" part of any plant population. They are those individuals that have fallen where they have no chance of survival to reproduction. There are even living dead that have lived past their reproductive age. However, these living dead differ from the tree in the field in a very critical way for those who consume them. These living dead are being continually replenished by the natural dispersal process. They do not herald an invisible walk to extinction for the consumer.

VII. ARE THERE LIVING DEAD HABITATS AND ECOSYSTEMS?

Even when heavily agroindustrialized, the tropical agroscape often has patches of wildlands (Fig. 5)—forests along rivers and ravines, broken topography, swamps and marshes, vegetation on bad soil, no-man's land between rival owners, woodlots, hunting preserves, industrial accidents, parks, and parklets. This remaining natural vegetation is a patchwork and a dot map, and it appears to be 1 to 20% of the original vegetation. And it gives one hope.

One says, "aha, there are remnants. There is wild biodiversity on the countryside, in the agroscape. There is hope outside of the reserves" (which are so hard to maintain and seem so expensive in national park status). This is a cruel illusion. Descend to one of these patchlets of forest, so green, so tree-filled. It is a biodiversity desert, lacking 90 to 99% of its original biodiversity that it had when it was once part of a forested landscape. As a package it is a vegetational living dead. Its species list is a mix of actual living dead and a few species that can maintain viable populations under these circumstances. Our major problem is that we visit these patches as tourists. We were not there in 1965 to see their earlier biodiversity, to compare it with its pale shadow in 1999 (but see Frankie et al., 1998).

Why are the survivors living dead, and what happened to those that have gone locally extinct? Part of them went when the area got so small that there were no longer circumstances for a viable population size. Part of them were explicitly mined or hunted. Part of them went when their mutualists, prey, and hosts went. Part of them went when the neighboring habitat, a habitat that spit seeds into the remaining forest and thereby maintained a population there, went to crop-lands. Part of them went when the seasons got drier, or wetter, or windier, or more fire-rich, or longer, or shorter, or, or, or.

Even those national parks that seem so secure are at major risk from this phenomenon. When the Southeast Asian dipterocarp trees fruit, the wild pigs come from everywhere and the collective seed crop of the preserved forest patch has no chance of satiating these seed predators (e.g., Curran et al., 1999). It may be better to surround a conserved wildland with wild animal-free rice fields than oceans of secondary succession subsidizing waves of animals that then turn the small old-growth forest into yet more secondary succession by defecating seeds all over it (e.g., Janzen, 1983a).

The bottom line is that the complex fabric woven from thousands of interacting species has been ripped to bits. Many of those that seem to have survived are living dead, or the serendipitous few that find this new impoverished habitat to their competitive liking. In short, these patches are only pseudo-remnants, not really smaller pieces of what once was. Even those ecosystems and habitats that have always existed as small units—a marsh, a landslide scar, a volcano top, a patch of serpentine soil—did not live in isolation. Rather, each was maintained by a complex ebb and flow of immigrants, waifs, and influences from the neighbors. When the neighboring natural system is turned to crop-land, the integrity of the small natural patch (e.g., Fig.
living an even yet more impoverished biodiversity exists. For those to whom a forest is just a bunch of large woody plants, for those who cannot or will not read the differences between an advertising ditty and a complex poem, the agroscape with its living dead and pseudo-remnant natural vegetation appears to be not much different from a glade and forest mix in a national park. All seems to be well. But when humanity expects something from that wildland patch, it discovers that almost all of its tropical biodiversity is gone.

These patches have also played a mean trick on the conservation community. A huge portion of the world’s conservation policy is based on the understandings of nature held largely intuitively by those who have grown up extra-tropical and learned their lessons from extra-tropical ecosystems. They easily adopt the mantra of trying to save the biodiversity remnants scattered across the agroscape. They are especially prone to do so in the face of the frustration of trying to save very large (and commercially juicy) blocks of intact vegetation.

The forest-patchlet-dotted agroscape of Minnesota or Sweden still collectively contains easily more than 80% of the species that were there when the European colonists arrived. However, the same snapshot of a Costa Rican agroscape contains at best 5 to 20% of what once was. And the percent is still falling rapidly because a huge fraction of what remains today is living dead. The more biodiverse and the more complex an ecosystem, the more likely that human perturbation will create anthropogenic living dead among the species with longer-lived individuals. This is because perturbations strip away mutualists and other biointeractors, leaving behind the physiologically functional individuals to live out their neutered life spans. The more biodiverse and the more complex, the more likely any given individual is to be dependent on one or more of these interactants to remain a member of the population.

These impoverished patches are especially deceptive for the bioliterate. For those to whom a forest is just a bunch of large woody plants, for those who cannot or will not read the differences between an advertising ditty and a complex poem, the agroscape with its living dead and pseudo-remnant natural vegetation appears to be not much different from a glade and forest mix in a national park. All seems to be well. But when humanity expects something from that wildland patch, it discovers that almost all of its tropical biodiversity is gone.

The future of real conservation in the tropics lies in by-and-large focusing our efforts on the survival of a relatively small number of very large and diverse biophysical units, each complicatedly integrated with local, national, and international societies (Janzen, 1998, 1999). Painful as it may be, resources spent on trying to save individual species and small habitat fragments scattered across the agroscape, often living dead, is bad conservation economics and creates an angry antagonistic Homsaptions.

We have no option in the tropics but to recognize that conserved wildlands are and always will be islands in an ocean of agroscape. Our task is to get on with rendering them into the highest quality islands possible, and not be distracted by, nor lulled by, the living dead individuals and islandlets. Yes, if there remains but just one Rembrandt painting, we of course save it even if it is bullet-holed and faded. However, we must recognize it for what it is and not convince ourselves that by doing so we have preserved our knowledge of European history.

VIII. RESTORATION BIOLOGY

The living dead are largely a negative force in the algebra of conservation biology and conservation reality. However, in those few cases where ecosystem restoration is desired or serendipitous, their life span delimits a window of opportunity for the reintegation of their species into the restoring ecosystem. Reintegration is not an unqualified given, however. A single large tree in a pasture being restored to forest may be dropping its seeds and fruists into an early successional old-field community that for decades is still way too unattractive to contain the seed dispersal coterie that will begin to restore the demography of that tree species. Equally, the pollinators of its flowers may already be extinct, or abhor the young secondary succession coming up below the large old parent. And finally, the physical climate of the highly deciduous and dry-season blasted secondary succession may well be a dismal place for a seedling or sapling of that old-growth giant. As every plantation initiator knows, the act of stuffing seeds into the ground does not a plantation make.

Until a very short time ago, the California condor was made up of living dead individuals. They were brought into captivity (e.g., transplanted to a safe field), reproduced (e.g., seeds collected and grown in pots), and have been put back out, hopefully in an agroecosystem with a friendly sociology. This habitat is, however, very seriously impoverished through reduction of ma-
rine mammal populations that so kindly generated the cadavers for lunch, and the California condor may always be dependent on human subsidy.

Many species of living dead may be rescued in this manner, if we care enough to spend the resources on them and gather information about them. But before racing out to apply the same technique to the living dead guapinol trees in the centers of Costa Rican pastures, a question very much needs to be addressed. Would not the same money spent on saving large blocks of guapinol-occupied wildlands, complete with their pollinators and dispersal agents, not generate vastly more conservation of guapinol and its hundreds of thousands of compatriot species? Yes, even these large blocks of wildland will contain some living dead. The wildland’s biodiversity will attain an equilibrium density at whatever number of species survive the reduction from a continent of wildland to a large island of wildland. Those who are extinguished during this process will suggest the list of who were the living dead.

See Also the Following Articles

CENTRAL AMERICA, ECOSYSTEMS OF • CONSERVATION BIOLOGY, DISCIPLINE OF • DEFORESTATION AND LAND CLEARING • EXTINCTIONS, MODERN EXAMPLES OF • POLLINATORS, ROLE OF • RANGE ECOLOGY • RESTORATION OF BIODIVERSITY • TROPICAL ECOSYSTEMS

Bibliography


LATITUDE, COMMON TRENDS WITHIN

Michael R. Willig
Texas Tech University

I. Context
II. Patterns
III. Mechanisms
IV. Amplification and Synthesis
V. Assessment

GLOSSARY

alpha diversity The diversity of species, often estimated as species richness, within a local community or site.
beta diversity The degree of turnover in species (and changes in their abundances) among communities or sites along a gradient or within a larger area.
gamma diversity The diversity of species, often estimated as species richness, in a larger area as a consequence of both alpha and beta diversity.
latitudinal gradient A gradual change in a characteristic of interest (e.g., species richness) with a gradual change in latitude; a gradient is well defined if it adheres to a particular mathematical relationship.
middomain effect A gradient wherein species richness increases symmetrically from latitudinal extremes to the middle of a region as a consequence of the random placement of species ranges within a geographic domain (also known as Pernet effect).
rapoport effect A latitudinal gradient wherein the sizes of the distributional ranges of species decrease with decreasing latitude.
scale dependence A condition in which either the form or the parameters of a relationship between two variables (e.g., richness and latitude) is contingent on spatial or temporal attributes.
species density The number of species within a sampling unit of fixed size.
species diversity A feature of biological communities or assemblages that reflects the variety of organisms in an area and that includes two components, species richness and species evenness (the degree to which all species have the same proportional abundance).
species richness The total number of species in an area.

LATITUDINAL gradients of diversity are biogeographic patterns that define the way in which the number of species changes with latitudinal position on the surface of the earth. The general pattern is for species richness to increase from polar to tropical regions (Brown, 1995; Gaston, 1996; Rosenzweig, 1995), regardless of the taxonomic affiliation of the organisms (e.g., mammals, fishes, insects, and plants) or geographic setting in which they occur (e.g., Africa, South America, and the Atlantic Ocean). This is true for extant organisms (Fig. 1) as well as for those organisms alive during the past 70 million years (Fig. 2). An increase in species richness with decreasing latitude is the pattern generally observed at three spatial scales, including the level of
FIGURE 1 Species richness gradients in Canada and the United States for trees (A), mammals (B), amphibians (C), and reptiles (D). Contour lines connect localities with approximately equal species richness (reproduced from Currie, 1991, with permission from The University of Chicago Press).
LATITUDE, COMMON TRENDS WITHIN

I. CONTEXT

Since the voyages of Darwin and Wallace, biologists have been fascinated with the high species diversity of tropical regions compared to those in temperate or boreal zones. Indeed, this fascination with tropical diversity catalyzed in many ways the conceptual development of the theory that currently constitutes modern ecology. Moreover, increasing concern about the loss of diversity, especially in tropical regions, has led to the rapid development of the science of conservation biology. Documenting the way in which diversity differs across the globe and understanding the mechanisms that produce such variation are critical steps in the design of global conservation strategies and the implementation of regional management plans.

The mid-1950s to early 1960s saw the emergence of rigorous quantification of broad-scale relationships between species richness and latitude (Fischer, 1960). Within the next 25 years, scientists convincingly had documented the ubiquity of gradients in which species richness increased toward tropical areas. Similar gradients also were documented for diversity of higher taxonomic groups (e.g., genera, families, and orders) (Figs. 3 and 4). Indeed, the increase in species richness for terrestrial and marine environments was quantified successfully for a wide variety of taxonomic groups, such as mammals, birds, reptiles, amphibians, fish, tunicates, crustaceans, mollusks, bryozoans, corals, foraminifera, and vascular plants. Nonetheless, some taxa representing lower levels in the systematic hierarchy (i.e., orders or families) were notable exceptions in having maximal diversity in polar (e.g., seals, penguins, and sandpipers) or temperate zones (e.g., voles, salamanders, ichneumonid wasps, and coniferous trees). Caution must be employed in considering such exceptions because other groups of equivalent rank within the same higher taxon often are restricted to lower latitudes, and the higher taxon exhibits a tropical maximum in species richness.
II. PATTERNS

A gradient implies a gradual change in species richness with a gradual change in latitude. In an unambiguous fashion, the form of that pattern is the precise mathematical or statistical relation that describes how species richness changes with latitude. As a consequence, three considerations are important in assessing patterns: the general shape of the curve (e.g., symmetry, kurtosis, or linearity), the parameters that characterize the relation, and the degree to which the fit of empirical data to the predicted curve is equivalent to the north and south of the equator. Knowledge of these three aspects of gradients suggests the kinds of causal mechanisms that are in operation. In addition, it facilitates comparison of gradients among taxa within the same geographic domain (Fig. 1; birds versus mammals versus reptiles versus amphibian in North America) as well as comparisons among different geographic domains for the same taxon (Fig. 4; North American versus South American for mammalian orders).

Patterns are often scale dependent, with particular mechanisms more likely operating at some areal scales than at others. Consequently, patterns will be elucidated for each of two foci: biotic assemblages occupying broad areas and ecological communities occupying local sites. These scales are intimately associated with each other. In part, the diversity of regions, biomes, or climatic zones is a consequence of the species richness that is accumulated within local communities. Similarly, the species richness and composition of local communities are affected by the set of taxa that constitute regional species pools (Putman, 1994).

A. Assemblages

Most of the empirical research concerning the relationship between species richness and latitude that has been done using arbitrary sampling units has been based on (i) latitudinal bands, (ii) quadrats of fixed area, or (iii) quadrats of unequal area defined by lines of longitude (meridians) and latitude (parallels). Alternatively, research has focused on the species richness of biomes or broad latitudinally defined climatic zones. Because the area of any sampling unit may have as large or larger an effect on variation in species richness than does its latitude, it is critical to understand how area may affect latitudinal patterns in different ways, depending on the method or approach.

Generally, analyses of quadrats defined by meridians and parallels are inferior to those based on other sampling units because such quadrats differ in area in a systematic fashion and bias quantitative conclusions. As meridians converge toward the poles, the size of the
Analyses based on latitudinal bands also must control for the effect of area because the width of a continent is not constant at all latitudes. A variety of methods have been used to compensate for this problem. Importantly, care must be employed when the area of bands varies in a systematic fashion with latitude because of the shape of the continent (e.g., progressive decreases in area with decreasing latitude in North America versus progressive increases in area with decreasing latitude in South America). In such cases, analyses based on bands may be plagued with the same confounding effects as those for analyses based on quadrats defined by meridians and parallels. Indeed, if classical statistical techniques are used to control for the effect of area in bands defined by 5-degree meridians in North America, the areal relationship is contrary to both common sense and ecological theory in that species richness increases as area decreases. Subsequent assessment of latitudinal effects may be severely compromised because the width of North America decreases as latitude decreases. Hence, adjusting the latitudinal gradient to account for area also removes an appreciable latitudinal effect. Nonetheless, if continental shape does not confound the effects of latitude and area, then regression techniques hold great promise, especially if appropriate nonlinear approaches are taken to adjust species richness in accord with species area theory (i.e., adjust richness of bands to a common area based on nonlinear regression of nested quadrats within each band).

Quadrats of fixed size also have been used to analyze broad-scale patterns of diversity. Nonetheless, variation in species richness among sampling units could still be a consequence of area, at least partially, when quadrats occupy coastal positions along continental borders. Adjusting for area in these cases may obscure the effects of rapid transition zones in terrestrial communities as they approach land–sea margins. Hence, the consensus is to not consider quadrats unless they are full of land. Subsequent variation among quadrats that is due to latitude can be assessed through a variety of statistical models. However, here too it is important to note that patterns are scale dependent. That is, the pattern detected for quadrats encompassing 1,500 km$^2$ could be quite different from those at 10,000 km$^2$. The importance of scale dependence in ecology and biogeography has been emphasized increasingly during the past decade. Nonetheless, broad-scale biogeographic patterns for sampling units between 1000 and 25,000 km$^2$ have been shown recently to be scale independent and little affected by area (Lyons and Willig, 1999).

Much of the early literature on latitudinal gradients in diversity was based on the species richness of broadly defined climatic zones (e.g., north polar, north temperate, tropical, south temperate, and south polar) or geopolitical units (e.g., countries, states, and provinces). Taxonomic richness was documented to increase from polar to tropical regions (Fig. 3). Even when values for richness were not adjusted for the areal extent of geopolitical regions (e.g., snake species in Argentina,
TABLE I
Mechanisms Potentially Affecting the Latitudinal Gradient in Species Richness*

<table>
<thead>
<tr>
<th>Circular</th>
<th>Empirically unsubstantiated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competition</td>
<td>Environmental stability</td>
</tr>
<tr>
<td>Mutualism</td>
<td>Environmental predictability (contingency)</td>
</tr>
<tr>
<td>Predation</td>
<td>Productivity</td>
</tr>
<tr>
<td>Epidemics</td>
<td>Abiotic rarefaction</td>
</tr>
<tr>
<td>Biotic spatial heterogeneity</td>
<td>Physical heterogeneity</td>
</tr>
<tr>
<td>Population size</td>
<td>Angle of the sun above the horizon</td>
</tr>
<tr>
<td>Niche width</td>
<td>Area</td>
</tr>
<tr>
<td>Population growth rate</td>
<td>Aridity</td>
</tr>
<tr>
<td>Patchiness</td>
<td>Seasonality</td>
</tr>
<tr>
<td>Epiphyte load</td>
<td>Number of habitats</td>
</tr>
<tr>
<td>Host diversity</td>
<td>Rapoport's rule (range size gradient)</td>
</tr>
<tr>
<td>Harshness</td>
<td>Ecological time</td>
</tr>
<tr>
<td></td>
<td>Evolutionary time</td>
</tr>
<tr>
<td></td>
<td>Temperature dependence of chemical reactions</td>
</tr>
<tr>
<td></td>
<td>Solar energy</td>
</tr>
<tr>
<td></td>
<td>Evolutionary speed</td>
</tr>
<tr>
<td></td>
<td>Stochastic placement of species ranges</td>
</tr>
</tbody>
</table>

* Modified from Rohde (1992).

ant species the Americas, and breeding bird species the Americas, the polar to tropical gradient was obvious. Nonetheless, controversy currently surrounds the interpretation of such data when the focus is on broad climatic zones associated with latitude. Some suggest that the increase in diversity with decreasing latitude primarily is a result of more tropical regions having larger areas than their extratropical counterparts. In contrast, others suggest that extratropical areas are often larger but have fewer species than their more tropical counterparts, suggesting that latitudinal correlates other than area are the driving forces behind the gradient.

B. Ecological Communities

The way in which latitudinal variation in diversity at broad spatial scales (gamma diversity) is related to patterns of species richness at the level of local communities (alpha diversity) is unclear. In part, this is because the geographic boundaries of a community are difficult to designate and are ultimately arbitrary decisions. Moreover, until recently, little was known about the composition of local communities in tropical regions, making assessment of broad-scale latitudinal patterns of community richness a premature endeavor. Finally, it is unlikely that a single research scientist could gather sufficient data across many sites to assess latitudinal gradients in community richness with sufficient power to distinguish pattern from noise. Hence, compositional data must be compiled from the work of many different individuals, who often use different methods, designs, and sampling intensities, to quantify the gradient in a meaningful way. The concern expressed by Voss and Emmons (1996) regarding adequate sampling within a community must be tempered by the realization that community composition has a temporal dynamic. Communities represent suites of populations with the potential to interact, and thus they must be constrained to some extent by both time and space. Sampling regimes which extend over protracted periods of time (e.g., decades or longer) may have inflated estimates of richness and inaccurate assessments of species composition. Nonetheless, recent progress in this regard allows quantitative evaluation of patterns in a rigorous way, at least for some taxonomic groups. When care is taken so that a local community is delimited as a geographic area in which constituent species have a high likelihood of interaction, it becomes clear that species richness increases from polar
through temperate to tropical regions. The gradient can arise from an increase in the number of functional groups (e.g., guilds or ensembles) within a community as well as from an increase in the number of species per functional group. For example, data for volant mammals (bats) from 17 local communities (Stevens and Willig, 1999) that met rigorous standards for inclusion in analyses (i.e., well-delimited local areas must have been sampled intensively for at least 1 year to include all seasons in which species are active) clearly exhibit a latitudinal gradient of increasing richness with decreasing latitude (Fig. 5A). A similar increase in the number of functional groups or feeding guilds within a community (Fig. 5B) contributes to this trend. Specifically, north temperate bat communities are depauperate with 10 or fewer species, all of which are members of carnivorous feeding guilds (e.g., aerial insectivore, molluscid insectivores, or gleaning animalivores). In contrast, tropical and subtropical communities on average contained 33.7 and 20.5 species, respectively. Moreover, functional diversity of communities in these two regions was high (equaling 5.9 and 4.5 feeding guilds, respectively) and included sanguinivorous, piscivorous, frugivorous, and nectarivorous taxa in addition to the guilds represented in the temperate communities.

Species richness of nonvolant taxa (tetrapods) in mammalian communities from North America also shows a significant increase from polar to tropical regions (Fig. 6; Kaufman, 1998). Tropical communities contained approximately 43.5 species on average, whereas extratropical communities contained only approximately 25.8 species. Within these two latitudinally defined climatic regions, the relationship between richness and latitude was nonsignificant. Most important, an examination of species turnover for these communities revealed that beta diversity increased toward the tropics as well. Hence, increasing diversity of broad geographic or climatically defined regions (gamma diversity) with decreasing latitude may be a consequence of the increase in local diversity (alpha diversity) as well as the increase in differentiation among local communities within latitudinal regions. This appears to be the case for nonvolant mammals in the New World, at least some spatial scales.

Many factors (e.g., productivity, competition, predation, and disturbance) have been suggested as the dominant forces affecting the composition and structure of local communities. Early theoretical and empirical work stressed the role of deterministic factors such as competition in molding community attributes. Subsequent focus on the distinction between equilibrium and nonequilibrium communities cast doubt on the universality of deterministic mechanisms in general and competition in particular and raised serious questions about the degree to which local communities were saturated by species. Indeed, variation in the degree to which local communities attain equilibrial richness may contribute strongly to the latitudinal gradient in diversity (Putman, 1994). The latitudinal progression from polar to tropical regions may represent a gradient in the degree to which stochastic density-independent mechanisms or biotic interactions dominate the forces affecting the abundance of local populations and the composition of local communities. Specifically, the species richness of a community may be a consequence of the severity, variability, and predictability of local environmental conditions. Low predictability and harsh conditions predispose communities to be regulated by abiotic parameters and to have low diversity. High predictability and conditions that are clement favor high diversity. To the extent that high solar insolation and warm temperatures represent favorable conditions, and low intra-annual variation in temperature and rainfall represent predictable conditions, tropical communities should be more species rich than their extratropical counterparts. In essence, the factors that affect elevated richness at the local scale likely contribute to enhanced gamma diversity of regions as well (see Section IV).
III. MECHANISMS

A grand proliferation of hypotheses (Table I), along with subsequent a posteriori modifications, is a characteristic feature of the literature concerning the relationship between diversity and latitude. A variety of factors contribute to this. Each of these hypotheses represents a conceptual model with only qualitative predictions. As with much of macroecological research, broad-scale data concerning the distribution of species is not available for many taxa. Manipulative experiments designed to disentangle the effects of competing hypotheses are not feasible or ethical. In addition, the inherent geographic factors that might affect richness are often correlated so that efforts to remove the effect of one to assess the other can lead to spurious results due to the confounded nature of the mechanisms.

A. Geographic Area Hypothesis

The latitudinal gradient in which richness peaks in the tropics may be a consequence of the larger landmass of the tropics compared to other geographic zones. This simple idea had its genesis in the work of Terborgh (1973), with considerable development and refinement by Rosenzweig (1995) in subsequent years, during which the effects of productivity and zonal bleeding have been incorporated into a more comprehensive conceptual model. Nonetheless, the geographic area hypothesis has generated considerable controversy and contention (Rohde, 1997, 1998; Rosenzweig and Sandlin, 1997). The controversy does not surround whether an areal mechanism operates; rather, it focuses on the degree to which variation in area is the dominant factor molding latitudinal gradients in richness.

Two features of the earth's geometry predispose the sizes of tropical regions to be greater than those of their higher latitude counterparts. First, the earth is essentially a sphere. The distance between longitudinal meridians at the equator is greater than that elsewhere on the globe, and intermeridian distance decreases in a regular fashion toward the poles. Second, northern and southern tropical zones are adjacent, whereas the northern and southern variants of other latitudinally defined zones are isolated from each other. Nonetheless, the positions, sizes, and configurations of the earth's continents will affect the proportion of land or water in tropical versus extratropical regions, and this has varied over geological time as a consequence of plate tectonics. In addition, the number and breadth of zones used to subdivide latitude will affect the perception of areal dominance associated with the tropics. For example, a tripartite division (tropical, temperate, and polar) reveals that the tropics ranks second in area to north temperate regions at the global scale, with considerable variation in the proportional area represented by the tropics among continents (approximately 38, 12, 80, 41, and 0% of America, Eurasia, Africa, Australia, and Antarctica, respectively). In contrast, finer resolution of zones to tropical, subtropical, temperate, boreal, and tundra indicates the areal predominance of tropical lands globally.

Most important, the degree of environmental variation within the tropics is less than that in other geographic zones, at least with respect to incident solar radiation and temperature. Specifically, a band of 50° centered on the equator evinces no or little change in mean annual temperature with latitude (constant at approximately 27°C), whereas mean annual temperature decreases thereafter by approximately 0.75°C per degree latitude. Hence, regardless of the size of zonal subdivisions chosen to define tropical or extratropical regions, tropical landmasses are larger than any other landmasses with similar variation in temperature.

As a consequence of the area extent and homogeneity of temperature and solar insolation in the tropics, speciation rates there should be higher and extinction rates lower than in extratropical regions. Specifically, the larger area of the tropics allows its species to have larger ranges than do their extratropical counterparts. Larger ranges allow species to be represented by more populations or populations of larger size, both diminishing the likelihood of extinction resulting from
accidental cause or from environmental perturbation. Simultaneously, larger areas are more likely to contain or experience geological events that produce geographic barriers that enhance the rate of allopathic speciation. The dynamic balance between the rates of speciation and extinction therefore yield higher equilibrial richness in the tropics than in extratropical areas (Fig. 7).

Rosenzweig and colleagues marshal many lines of evidence in support of the geographic area hypothesis by citing two kinds of observations. First, larger biotic provinces, regardless of latitude, have more taxa than do their smaller counterparts (e.g., generic, familial, and ordinal richness of mammals increases with provincial area). Second, diversities from the same biome but from different continents or provinces differ as a function of their areal extent (e.g., rain forest vertebrates and plants as well as tropical freshwater fish increase in richness as their areal extent increases). Situations in which the general pattern does not occur usually include large but unproductive climatic zones with few species—effectively the richness-diminishing effects of low productivity may countermand the dominant role of area in these systems. In contrast, Rohde (1997) considers area not to be the dominant factor that affects high species richness in the tropics. He illustrates the point with Eurasian freshwater fishes, and shows that much smaller tropical regions have much greater species richness than do larger cold-temperate regions. Similarly, the expansive deep-sea biome with more or less constant temperature has far fewer species than its smaller tropical counterparts. Clearly, consensus is elusive concerning the relative importance of area in affecting species richness compared to other mechanisms.

B. Evolutionary Speed

After a broad and incisive review of the various mechanisms purported to cause latitudinal gradients in species richness, Rohde (1992) found them all to be lacking. Instead, he suggested that the gradient was a consequence of differential rates of speciation associated with an important latitudinal correlate, temperature, rather than being a product of equilibrium-based ecological processes that presupposed that local communities are saturated with species. His conceptual model is erected on the foundation of three premises. First, tropical environments support shorter generation times for many homiotherms and poikiliotherms. Second, mutation rates increase as temperature increases and are highest in the tropics. Third, faster physiological processes occur at higher temperatures; this, coupled with the first two relationships, suggests an accelerated rate of fixation of favorable alleles in tropical populations. This effectively results in greater evolutionary time in the tropics for mechanisms of diversification to attain fruition.

Much of the subsequent dialog concerning the evolutionary speed hypothesis has been embedded in the debate concerning the efficacy of area, productivity, and zonal bleeding in producing empirical gradients of diversity with respect to latitude. Although this debate has been fruitful in crystallizing assumptions of equilibrial conditions and species saturation as they relate to the area hypothesis, it has not furthered our understanding of the possible role of temperature in affecting speciation rates. Rhode has challenged the research community with this supposition; unfortunately, the response is deafening in its silence.

C. Rapoport–Rescue Hypothesis

As its name implies, this hypothesis is a hybrid of two mechanisms operating in tandem: the Rapoport effect and the rescue effect. A geographic pattern in which species range size decreases from high to low latitudes recently has come to the forefront of the macroecological literature (Stevens, 1989, 1992) and been termed Rapoport’s rule after the Argentine scientist who first discussed the pattern in the context of many other areographic principles. Stevens hypothesized that the latitudinal propensity for range size to decrease toward the tropics, when combined with differential movement
of individuals from source to sink habitats (rescue or mass effect), can generate the latitudinal gradient of diversity. Specifically, at any one locale in the temperate zone, an individual must be able to tolerate considerable intraannual variation in climatic conditions; thus, species that occur in the temperate zone can attain a wide latitudinal distribution because of the broad tolerance of its constituent individuals to varying local conditions. In contrast, an individual in the tropics experiences little seasonal variation in climatic conditions; consequently, species comprising individuals that occur in tropical zones are predisposed to have narrower latitudinal distributions. This creates the Rapoport effect. The rescue effect is a phenomenon whereby local extinction of a population, often in marginal or sink habitats, is prevented because of immigration of individuals from source or high-quality habitats. Because smaller ranges, which are differentially situated in the tropics as a consequence of the Rapoport effect, have greater perimeter to area ratios, they are predisposed to having greater rescue effect areas relative to range areas. This differentially inflates species richness in tropical areas, generating the latitudinal gradient of diversity.

The generality of Rapoport’s rule, as well as the degree to which empirical patterns are generated by the hybrid mechanisms embodied in rescue and Rapoport effects, is controversial. The Rapoport effect has been documented for a diversity of taxa (mammals, reptiles, amphibians, fish, crayfish, amphipods, mollusks, and trees) in aquatic and terrestrial environments and quickly has become engrained as the explanation for species diversity gradients in a variety of ecology textbooks. Additional circumstantial evidence was derived from the observation that taxa, which do not show the general latitudinal gradient in richness, do not adhere to Rapoport’s rule, suggesting that both patterns had a shared mechanistic basis (Stevens, 1989). Nonetheless, a growing body of evidence suggests that the pattern is far from universal (Rohde et al., 1993; Lyons and Willig, 1997). Moreover, reanalyses of data on marine mollusks that was used to corroborate the Rapoport effect (Stevens, 1989) failed to produce the same patterns in a subsequent study, even though the methods were the same in both studies (Roy et al., 1994). In addition, New World bats and marsupials (Lyons and Willig, 1997; Willig and Lyons, 1998), as well as nonmigratory marine teleosts from surface waters (Rohde et al., 1993), each exhibit strong latitudinal gradients in diversity but do not adhere to Rapoport’s rule. Hence, occurrences of latitudinal gradients in diversity do not have one-to-one correspondence with the existence of a Rapoport effect.

Recent simulation models provide added insight into the phenomenon of Rapoport’s rule. The three commonly used methods (i.e., Stevens’, midpoint, and most-distal point) for assessing a Rapoport effect suffer from serious limitations. Stevens’ method is problematic because of a lack of independence associated with counting the same species multiple times in the same analysis (Rohde et al., 1993). Midpoint and most-distal point methods suffer from severe mathematical biases—the bounded nature of continents or oceans predisposes correlations between range size and latitude even when ranges are distributed stochastically with respect to latitude (Colwell and Hurtt, 1994; Lyons and Willig, 1997). Finally, a comprehensive set of simulation models (Taylor and Gaines, 1999) suggests that the Rapoport effect causes a latitudinal pattern in species richness, but the gradient is opposite of the pervasive pattern found in nature in that species richness increases with increasing latitude. Moreover, incorporation of a rescue effect into the model so that it reflects the Rapoport–rescue mechanism still fails to rescue the hypothesis; the predicted pattern remains a decrease in richness toward the tropics. Only the incorporation of competitive effects to either the simulation model based on the Rapoport effect alone or to the combined Rapoport–rescue mechanism produces latitudinal gradients that are consistent with real-world patterns, and in both scenarios, communities must be saturated at equilibrial compositions.

D. Geometric Constraint Hypotheses

The ubiquity of the latitudinal increase in diversity with decreasing latitude stimulated the search for a predominant mechanism effecting variation in richness. The rationale was essentially that because almost all taxa on all continents as well as in the oceans followed the same pattern, and the pattern has persisted through geological time, a single mechanism must be causing it. Interestingly, the search for a predominant mechanism during the past 25 years has generated increasingly more hypotheses rather than leading to consensus or a synthetic understanding of mechanisms producing the universal pattern.

Insights during the past 5 years concerning the nature of geographic constraints in affecting patterns of species distribution document clearly that modal patterns of diversity peaking in the tropics can be a consequence of the bounded nature of terrestrial and aquatic habitats. Indeed, both simulation (Colwell and Hurtt,
LATITUDE, COMMON TRENDS WITHIN

FIGURE 8  Graphical representation (parabola) of the gradient of species richness that arises from the random placement of species ranges within the latitudinal bounds of the New World, scaled from 0 in the south to 1 in the north (modified from Willig and Lyons, 1998). The number of species at any latitude is determined by its proportional distance ($p$) from the southern boundary and is given by $\frac{2p^2}{(1 + p^2)}S$. The vertical lines labeled A and B represent the latitudinal extents of bats and marsupials, respectively.

1994; Pineda andCaswell, 1998) and analytical (Lees et al., 1999; Willig and Lyons, 1998) null models suggest that species richness of a biota should increase toward the center of a shared geographic domain in a quasi-parabolic or parabolic fashion as a consequence of the random placement of species ranges [the mid-domain effect of Colwell and Lees (1999)]. Three null models have been developed that differ in the manner in which species ranges are randomized. Unlike the other mechanisms proposed to account for latitudinal gradients, which only suggest qualitative increases in richness with decreasing latitude, geometric constraint models make quantitative predictions concerning the form of the latitudinal gradient so that expected values for richness occur for each latitude and can be compared to empirical data.

In the fully neutral model (Colwell andHurtt, 1994; Willig and Lyons, 1998), the placement of termini for each species distribution is random, with the geometric constraint that they occur within a shared boundary or domain. As a consequence, the number of species at any point within the domain ($S_p$) is only related to the proportional distance of that point from the boundary ($p$) and the number of species in the species pool ($S$), and is given by $S_p = 2p(1 - p)S$. The model (Fig. 8) is an incarnation of both the two-hit broken stick model of MacArthur and the binomial distribution. In essence, if the latitudinal domain of a biota is rescaled to range from 0 in the south to 1 in the north, then the likelihood of a species range overlapping any point $P$ that is exactly $p$ from the southern terminus (and hence $1 - p$ from the northern terminus) of the domain is

$$Pr(P) = 1 - p^2 - (1 - p)^2 = 2p - 2p^2,$$

where $p^2$ is the proportion of species whose northern and southern boundaries lies to the south of $P$, and $1 - p^2$ is the proportion of species whose northern and southern boundaries lie to the north of $P$. The functional form of the distribution of $Pr(P) = 2p - 2p^2$ is a parabola that peaks at 0.5, and as a consequence the average size range of species in the biota is half the extent of the domain ($0.5$).

The other two models have constraints concerning either the range size distributions of the biota or the distributions of midpoints within the domain. In the second model, the placement of ranges within the domain is constrained such that the simulated ranges have a size distribution exactly equal to that of the biota of interest. Essentially, the size of a species range limits the options for the feasible placement of midpoints: Species with broad distributions must have midpoints located near the center of the domain, whereas species with increasingly narrow ranges can have midpoints
located ever more distant from the center of the domain. Hence, randomization of location for the midpoint of each species range is constrained to only occur within a subset of the domain, and within this subdomain its position is determined from a uniform random distribution. This produces a quasi-parabolic curve that becomes increasingly flattened as mean range size diverges from 0.5 (the mean value in the totally null model). The third model constrains the distribution of midpoints to be exactly the same as that in the empirical data. It essentially allows the extent of each species range to vary in a uniform random manner with the only constraint being that its distributional midpoint not change. Although it exists as a simulation null model, its analytical analog has not been developed (Colwell and Lees, 1999).

Empirical support for geometric constraints is limited but increasing. In a comparison of empirical latitudinal gradients of richness with predictions of a geometrically constrained null model in the New World, Willig and Lyons (1998) were able to account for 69–95% of the variation in species richness for bats and marsupials (Fig. 9). Nonetheless, systematic deviations from the null distribution were observed for both taxa. The null model overestimated bat species richness near the edges of the domain and underestimated richness near the center of the domain. In contrast, the null model consistently underestimated species richness for marsupials at all latitudes. Although both taxa gradually increase in species richness toward the tropics, the manner in which they deviated from the predictions of the null model is in sharp contrast, and the residuals were not related to the area or the width of the continent at each latitude.

The geometric model constrained to conform to empirical range size distributions accounted for the spatial distribution of endemic rain forest taxa (e.g., butterflies, frogs, rodents, tenrecs, chameleons, and birds) in Madagascar far better than did mechanisms related to area, elevation, temperature, precipitation, habitat diversity, or productivity (Lees et al., 1999). Richness peaked at or near the midpoint of domains even though values for numerous environmental characteristics did not do so. This was true for the one-dimensional domain of latitude (the model accounts for 85% of variation in richness) as well as for the two-dimensional domain defined by latitude and longitude (the model accounts for 73% of the variation in richness).

In an analogous fashion, applications of geometric constraints models to elevational distributions of birds (Rahbek, 1997) or bathymetric distributions of gastropods and polychaetes (Pineda and Caswell, 1998) have led to considerable insight concerning diversity gradients. This is particularly relevant because elevational gradients are thought to recapitulate latitudinal patterns and mechanisms. For New World tropical birds, a peak in species richness occurred at an intermediate elevation, after controlling for the effects of area. Indeed,
Based on first principles, Kaufman et al. (2000) provided a synthetic model to explain the latitudinal gradient in diversity based on variation in abiotic stress, productivity, and biotic interactions. The model is comprehensive because it simultaneously considers and makes predictions about other macroecological patterns, such as latitudinal gradients in range size, habitat specificity, and species dominance. Central to their model is the premise that abiotic stress increases with increasing latitude. The stress is a result of three interacting phenomena that vary with latitude as a consequence of the spherical nature of the earth and the fact that it rotates on a tilted axis with respect to the sun. Solar radiation is the ultimate source of energy for all food webs, and the daily input of energy per unit area decreases from the tropics to the poles. Similarly, the two essential metabolic pathways that dictate energy transformation in the earth’s living systems—photosynthesis and respiration—are temperature dependent, and average daily temperature decreases from tropical to polar latitudes, increasing the costs of executing all life-sustaining physiological processes. Finally, intra-annual variation in both temperature and solar insolation increases from tropical to polar regions.

Together, these phenomena result in increasingly energy-poor and stressful environments toward high latitudes which reduces the richness of local communities and assemblages. Increased costs of maintaining metabolic rates combined with reduced inputs of energy in extratropical regions result in population densities of a species diminishing at high latitudes, whereas increased biotic pressures resulting from interspecific interactions reduce densities at the tropical edge of a species boundary. Consequently, species are expected to have modal geographic distributions of population size. Second-order biotic feedbacks (e.g., parasitism, disease, and diffuse competition) are hypothesized to further increase the costs of survival where richness is high, thereby truncating the extent of species distributions and the abundance of local populations in tropical regions. This would generate a Rapoport effect and increase the rate of species turnover (beta diversity) and diminish dominance toward the tropics. Although data are insufficient to test all of the patterns predicted by the synthetic model, and there is controversy surrounding the ubiquity of the Rapoport effect, for heuristic reasons alone the conceptual model provides a springboard from which enhanced understanding of the causes and consequences of the latitudinal gradient may be forthcoming.

The ontogeny of theory can be viewed from a variety of perspectives that deal with the detection of patterns, the linkage of patterns to particular mechanisms, and ultimately the integration of those constructs to other theories in the discipline (Pickett et al., 1994). The theory of latitudinal gradients of diversity has matured considerably in the past 5 years. The general patterns of latitudinal increase are well documented from an empirical perspective. In addition, the manner in which particular mechanisms could affect patterns of diversity has become clearer. Elements of the theory have been used to understand other gradients of diversity, such as those related to elevation or depth. Finally, latitudinal patterns of diversity are being integrated with other broad-scale patterns concerning the dominance and turnover of species as well as to latitudinal patterns of range size and abundance. This represents a significant advancement in understanding and integration.
Although few of the hypotheses postulated to affect the latitudinal gradient in diversity have been eliminated in a conclusive manner, research concerning many of the mechanisms appears more likely to advance theory in the near future. Indeed, recent synthetic works have focused on them to a great extent. Understanding the contexts and degrees to which area, climatic variability or stress, geographic constraints, productivity, temperature, and their interactions mold the latitudinal gradient in diversity remains a challenge for the next decade of scientists to address in a synthetic manner.

See Also the Following Articles

Biodiversity-Rich Countries • Biogeography, overview • hotspots • species-area relationships • tropical ecosystems

Bibliography


LIFE HISTORY, EVOLUTION OF

Derek Roff
McGill University

I. Introduction
II. Two Frameworks for Analysis
III. Tradeoffs: A Central Feature of Life History Analysis
IV. The Evolution of Life Histories

GLOSSARY

**genotype** The genetic constitution of an organism.

**heterozygosity** The presence of alternate alleles at a given locus in a diploid organism (e.g., \(A_1A_2\)).

**homozygosity** In a diploid organism, the presence of the same alleles at a given locus (e.g., \(A_1A_1\)).

**iteroparity** Repeat breeding (see semelparity).

**semelparity** Breeding once and dying; sometimes called ‘big bang’ reproduction.

THE ANALYSIS OF LIFE HISTORY evolution includes any trait that impinges on the reproductive success of an organism. More specifically, life history evolution is typically concerned with the evolution of the age and size at first reproduction, reproductive effort, clutch size, and propagule size. While many analyses focus only on a single trait, it should be remembered that selection acts on fitness (as defined later) and not solely on single traits. Therefore, the appropriate framework for the analysis of life history evolution is the whole suite of traits that interact to determine the fitness of an organism. The analysis of components of fitness is appropriate in many circumstances but the limitations of such an analysis must always be remembered.

I. INTRODUCTION

Plants and animals show profound variation in all aspects of their life histories, which include age at maturity, age-specific fecundity, survival rate, size at birth, and so on. This variation is evident at both the interspecific and intraspecific levels. For example, at the interspecific level, species of flatfish range in size from 2 cm-long tropical species that reproduce within their first year of life to behemoths such as the Pacific Halibut (Hippoglossus hippoglossus), which exceed 200 cm and take over 10 years to mature. Though the range in variation within a species is not as dramatic as between species, it is still impressive, as illustrated by variation in the flatfish, Hippoglossoides hippoglossoides. In this species maturation occurs at age 3 years at a length of 20 cm in populations off the coast of Scotland while the same species requires 15 years to reach maturity at a length of 60 cm on the Grand Banks of Newfoundland. Longevity and maximum size are equally different in the two populations, Scottish fish reaching a maximum length of 25 cm and an age of 6 years, compared to 60 cm...
and 20+ years on the Grand Banks. Similar observations on variation in life history characteristics could be made in most taxa. But though the diversity of life histories is readily apparent, attempts to understand its origin and maintenance are still in their infancy.

Because fitness varies quantitatively among different life histories, a necessary tool for analysis is mathematical modeling. An influential factor encouraging the use of mathematical investigation into life history variation was Lamont Cole’s 1954 paper, “The Population Consequences of Life History Phenomena,” which set out the basic mathematical framework by which the consequences of variation in life history traits can be analyzed. Cole’s paper ushered in an era of research predicated on the integration of mathematics and biology in the study of the evolution of life history patterns.

In his review, Cole analyzed how changes in demographic attributes, such as the age at first reproduction, influenced the rate of increase of a population. Cole’s paper gained widespread notice because of an apparent paradox with respect to the value of semelparity versus iteroparity: “For an annual species, the absolute gain in intrinsic population growth which could be achieved by changing to the perennial reproductive habit would be exactly equivalent to adding one individual to the average litter size” (Cole, 1954, p.118, Cole’s italics). Thus, an annual species with a clutch size of 101 would increase in numbers as fast as a perennial that produces 100 young every year forever. There is obviously something amiss with this result, for perennials are common. While there is good evidence that survival and reproduction are negatively correlated it seems highly unlikely that perennial species are committing so much energy into reproduction that they cannot produce one more offspring. Cole’s paradox derives from a failure to consider the consequences of juvenile and adult survival rates; specifically Cole implicitly assumed that the survival rate of juveniles is the same as that of the adults. In general, juvenile survival is lower (often by a great amount) than that of the adults and this means that the reproductive output of the annual must be greater than one for the annual to have the same gain in rate of increase as a perennial, the precise amount depending on the mathematical details of the model. Cole’s paradox illustrates the need to formulate an analysis of life history variation within a mathematical framework where the biological assumptions are explicitly stated. Since Lamont Cole, the theoretical and experimental analysis of life history evolution has made enormous strides using two different analytical perspectives.

II. TWO FRAMEWORKS FOR ANALYSIS

The evolution of life history variation has been approached via two different modes of analysis, here termed phenotypic models and genetic models.

A. Phenotypic Models

In this approach no attempt is made to model the genetic basis of traits: it is simply assumed that there exists sufficient genetic variation that evolution is not constrained by genetic architecture. Is this a reasonable assumption? Most life history traits—such as age at first reproduction, fecundity, and survival—are not determined by simple Mendelian mechanisms such as single locus, two allele systems. More generally they show continuous variation, or if dichotomous can be best interpreted under a threshold model in which there is an underlying continuously varying trait. Nevertheless, polygenic determination is not essentially different from the simpler case of a single locus with two alleles. In the one locus case the phenotype is determined by the additive action of the alleles, or one allele may show dominance. When one allele is dominant, the effect on the phenotype can be statistically decomposed into additive and dominance components. With two or more loci controlling a trait there may be interaction between loci. These interactions, termed epistasis, can produce a wide range of responses but are generally assumed to be absent (because most variation can be subsumed under the additive or dominance portion of genetic variance, this assumption is not as unreasonable as it might appear).

The average similarity between parent and offspring can be measured by the linear regression of mean offspring value (Y) on midparent value (X),

\[ Y = \left(1 - h^2\right)m + hX \]  

where m is the population mean of the parents and the slope of the regression, \( h^2 \), is termed heritability in the narrow sense. More generally, heritability is defined as follows:

\[ h^2 = \frac{V_a}{V_t} \]

where \( V_a \) is the additive genetic variance and \( V_t \) the total phenotypic variance. Heritability in the narrow sense, or generally simply heritability, must not be confused with heritability in the broad sense, \( V_a/V_t \), which is a measure of the overall variance in the trait attributable to all genetic influences.
Obviously $h^2$ varies between 0 and 1: at 0 there is no resemblance between parent and offspring due to the additive effects of genes, while at 1 the mean offspring value equals the midparent value. The importance of heritability for evolutionary theory is obvious: the higher the value of $h^2$, the faster genetic changes in the population can occur. Most important from the point of view of life history theory, if heritability is zero for a particular trait or traits under consideration, then the optimal combination cannot be attained because the effects of selection on parents are not manifested in the offspring. Heritability estimates for different types of traits are shown in Table I. While heritabilities for life history traits are typically lower than those less directly related (in general) to fitness, there is nevertheless sufficient genetic variation to permit rapid evolutionary change even under modest selection pressure.

While the assumption of sufficient additive genetic variation is reasonable with respect to the equilibrium conditions, the genetic architecture can considerably influence evolutionary trajectories and hence the phenotypic approach is contingent on the state having attained an equilibrium. The method seeks to construct the fitness surface and hence the optimal combination of trait values.

1. Optimality Modeling

The concept of tradeoffs is central to present theories of the evolution of life history traits, for tradeoffs limit the scope of variation. Within the set of possible combinations there will be at least one combination that exceeds all others in fitness. Optimality analysis assumes that natural selection will drive the organism to that particular set. To initiate an analysis using the principle of optimality, we must designate some parameter to be optimized. In the present case we assume that there is some measure of fitness that is maximized by natural selection. The second step is to construct a set of rules that define the life history pattern of the organism, hypothetical or real, under study. Within these rules there will exist a variety of possible life histories; the optimal life history is that which maximizes fitness.

What do we do if the predicted life history does not correspond to that which is observed? The first point to note is that the principle of optimality is not under test. Failure to get a correct prediction is not taken as evidence that fitness is not being maximized, but it is taken to imply that the model is deficient. Having found that the initial model does not work we inquire into the assumptions of the model, namely the rules that define the range and scope of life history variation. These rules are changed either arbitrarily or based on further observation until agreement is gained between prediction and observation. Having found congruence, we are not able to say that therefore the component relationships are correct. The underlying components must be independently verified; a model is only the logical outcome of a set of interactions and the onus is on the person specifying those rules to demonstrate that they are indeed valid. Note again, that the assumption that fitness is being maximized is not under test, except to the extent that the particular measure chosen may be inappropriate. This does not mean that we assume that all traits and trait combinations are the result of adaptive evolution. But we do choose those that we have a priori reason to suppose are under selection.

A primary purpose of optimality modeling is to organize a program of experimentation and data collection. An adequate fit of a model to data gives us reassurance that a sufficient number of factors have been taken into account. Nevertheless, the validity of a model is continually under question and is challenged by the addition of more information. The more tests the model survives the greater the assurance that it is realistically capturing the important elements that determine the set of traits being studied.

TABLE I

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Life history</th>
<th>Behavior</th>
<th>Physiology</th>
<th>Morphology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila</td>
<td>0.12</td>
<td>0.18</td>
<td>n</td>
<td>0.32</td>
</tr>
<tr>
<td>All animals</td>
<td>0.26</td>
<td>0.30</td>
<td>0.31</td>
<td>0.40</td>
</tr>
<tr>
<td>Mammals</td>
<td>0.26</td>
<td>0.31</td>
<td>0.31</td>
<td>0.34</td>
</tr>
</tbody>
</table>

* Data are from Roff and Mousseau (1987) and Mousseau and Roff (1987).

Obviously $h^2$ varies between 0 and 1: at 0 there is no resemblance between parent and offspring due to the additive effects of genes, while at 1 the mean offspring value equals the midparent value. The importance of heritability for evolutionary theory is obvious: the higher the value of $h^2$, the faster genetic changes in the population can occur. Most important from the point of view of life history theory, if heritability is zero for a particular trait or traits under consideration, then the optimal combination cannot be attained because the effects of selection on parents are not manifested in the offspring. Heritability estimates for different types of traits are shown in Table I. While heritabilities for life history traits are typically lower than those less directly related (in general) to fitness, there is nevertheless sufficient genetic variation to permit rapid evolutionary change even under modest selection pressure.

While the assumption of sufficient additive genetic variation is reasonable with respect to the equilibrium conditions, the genetic architecture can considerably influence evolutionary trajectories and hence the phenotypic approach is contingent on the state having attained an equilibrium. The method seeks to construct the fitness surface and hence the optimal combination of trait values.

1. Optimality Modeling

The concept of tradeoffs is central to present theories of the evolution of life history traits, for tradeoffs limit the scope of variation. Within the set of possible combinations there will be at least one combination that exceeds all others in fitness. Optimality analysis assumes that natural selection will drive the organism to that particular set. To initiate an analysis using the principle of optimality, we must designate some parameter to be optimized. In the present case we assume that there is some measure of fitness that is maximized by natural selection. The second step is to construct a set of rules that define the life history pattern of the organism, hypothetical or real, under study. Within these rules there will exist a variety of possible life histories; the optimal life history is that which maximizes fitness.

What do we do if the predicted life history does not correspond to that which is observed? The first point to note is that the principle of optimality is not under test. Failure to get a correct prediction is not taken as evidence that fitness is not being maximized, but it is taken to imply that the model is deficient. Having found that the initial model does not work we inquire into the assumptions of the model, namely the rules that define the range and scope of life history variation. These rules are changed either arbitrarily or based on further observation until agreement is gained between prediction and observation. Having found congruence, we are not able to say that therefore the component relationships are correct. The underlying components must be independently verified; a model is only the logical outcome of a set of interactions and the onus is on the person specifying those rules to demonstrate that they are indeed valid. Note again, that the assumption that fitness is being maximized is not under test, except to the extent that the particular measure chosen may be inappropriate. This does not mean that we assume that all traits and trait combinations are the result of adaptive evolution. But we do choose those that we have a priori reason to suppose are under selection.

A primary purpose of optimality modeling is to organize a program of experimentation and data collection. An adequate fit of a model to data gives us reassurance that a sufficient number of factors have been taken into account. Nevertheless, the validity of a model is continually under question and is challenged by the addition of more information. The more tests the model survives the greater the assurance that it is realistically capturing the important elements that determine the set of traits being studied.

2. Game Theory

Game theory is really a subset of optimality modeling: it is appropriate when interactions are frequency dependent. The approach comprises two essential elements. First, it is assumed that particular patterns of behavior will persist in a population provided no mutant adopting an alternate behavior can invade. Such stable combinations are termed evolutionarily stable strategies (ESS). The concept of the ESS is not unique to game theory: the maximization of fitness measures in optimality models are all ESSs within the context in which they are appropriate. Second, for each type there must be an assigned gain or loss in fitness when this type interacts with another individual. From this payoff matrix we compute the expected payoff for each behavior.
For two behaviors to be evolutionarily stable, their fitnesses must be equal.

**B. Genetic Models**

Genetic modeling proceeds by first defining the genetic mechanism determining the phenotypic trait and then cranks the model through the appropriate mathematical machinery to obtain the equilibrium trait values. The critical problems are the correct definition of the genetic architecture and the estimation of the selection gradients.

1. Simple Mendelian Models

Consider a trait whose expression is governed by a single locus with two alleles, A₁, A₂. A selection coefficient (fitness) representing their relative contribution to the next generation can be assigned for each of the three genotypes (Table II). The change in the frequency of A₁, p, is then easily obtained as

\[ \Delta p = p [w_1 (1 - p) + w_2 p] \]

At equilibrium \( \Delta p = 0 \), and hence it is a trivial matter to predict the frequencies of the three genotypes at equilibrium and hence the mean trait value. Complexity can be added by making the fitness a function of density or frequency, but this does not change the basic mathematical approach.

Most traits of ecological interest, fecundity, age at maturity, clutch size, egg size, and so on are continuous in character. Even traits that appear dichotomous, such as liability to disease, wing dimorphism, diapause, and sex ratio, are best understood as being the result of some underlying continuously varying factor exceeding or not attaining a threshold for the expression of the trait. The expression of traits that show continuous variation are not, in general, the result of a single gene, nor two genes, but a large number of genes that acting additively produce a continuous spectrum of phenotypes. The analysis of such traits is the domain of quantitative genetics. This is largely a statistical approach to genetic variation and is founded on a mathematical analysis of variation rather than an understanding of how groups of genes interact to determine a particular trait.

2. Quantitative Genetic Models

The concept of heritability has already been introduced. For nonzero heritabilities, the rate at which the appropriate combination of trait values can be realized depends in part on the value of the heritability and the intensity of selection, measured as the selection differential, S, the difference between the mean of the population and the mean of the selected parents. For a single trait it can be shown from Equation 1 that the response to selection (the difference between the mean population values in parental and offspring generations), R, is equal to h²S. Since selection changes the gene frequencies, heritability must change at each generation. This problem is avoided in the application of quantitative genetic theory to natural populations by assuming that selection is weak and population size large, thereby permitting mutation to replace variation eroded by selection. Whether this assumption is reasonable can only be answered empirically and at present there are insufficient data to draw any meaningful conclusion.

Tracts are typically not inherited as separate independent units, rather, phenotypic values of several traits are typically controlled in part by a common set of genes. Thus selection on one trait will produce a change not only in the selected trait but also traits that are genetically correlated by virtue of shared genes (genetic correlations can also arise through linkage disequilibrium, but these are transitory and are ignored here). Selection on two traits, X and Y, will produce standardized (in phenotypic standard deviation units) responses \( R_x \) and \( R_y \),

\[ R_x = \beta_x h_x + \beta_y h_y r_{xy} \]

\[ R_y = \beta_x h_x + \beta_y h_y r_{xy} \]

where \( r_{xy} \) is the genetic correlation between the two traits and \( \beta_x, \beta_y \) are the standardized selection differentials on X and Y, respectively. The preceding equation can be more conveniently written in matrix notation, for which the typical notation is \( \Delta z = G\beta \), where \( \xi \) is a vector of
trait means. \( G \) is the genetic variance-covariance matrix, and \( \beta \) is a vector of (unstandardized) selection gradients for each character.

The evolutionary trajectories for two traits can be easily visualized by constructing a plot of isoclines of equal fitness (defined by the selection gradients) and plotting the changes in trait means each generation as illustrated in Fig. 1. If there is no genetic correlation between the two traits, the population will move rapidly to its optimal combination. For genetic correlations lying between \(-1\) and \(1\), the trajectory is "warped" and if the genetic correlation is exactly \(\pm 1\), the optimal combination may be unattainable (Fig. 1).

The theory described here assumes that the genetic covariance (i.e., the genetic correlation) remains constant. As with heritability, selection changes allelic frequencies and hence must in principle change the genetic correlations, or more generally the genetic variance-covariance matrix. The \( G \) matrix will remain constant provided selection is weak and population size large so that mutation can replace variance lost due to selection and drift.

### III. TRADEOFFS: A CENTRAL FEATURE OF LIFE HISTORY ANALYSIS

Tradeoffs are essential elements of all the above approaches. In the "phenotypic approach" it is implicitly assumed that the tradeoffs are genetically determined, otherwise there could be no evolutionary response. In
quantitative genetic terms, this is interpreted as a negative genetic correlation between two traits. Thus it has been supposed that to demonstrate an evolutionarily important tradeoff it is sufficient to demonstrate that the tradeoff is expressed not only as a phenotypic correlation but also as a negative genetic correlation. However, as discussed earlier, a negative genetic correlation is not, in principle, a barrier to movement anywhere in parameter space. Thus while it is necessary for a tradeoff to be expressed as a genetic correlation for it to be evolutionarily significant, this is not a sufficient demonstration that the optimal combination will be governed by the tradeoff.

Empirical investigations of tradeoffs can be placed into three categories: (a) phenotypic relationships based on field or laboratory observations of unmanipulated situations, (b) experiments in which organisms were manipulated to vary the value of one trait (e.g., manipulation of clutch size to investigate subsequent survival of adults or young), and (c) demonstration of a negative genetic correlation between two traits, obtained by sib analysis or selection. Only the last measure provides definitive proof that the tradeoff is evolutionarily significant but there is, however, merit in the second approach because it can establish the mechanism generating the tradeoff. The first approach is suspect because of the covariation of traits that could produce false conclusions if such covariation is not taken into account. For example, reproduction may be determined by condition, those in poor condition not breeding; consequently it would not be surprising to find the survival rate of nonbreeders to be less than breeders.

There is no reason to suppose that all tradeoffs between life history traits will be negative, but we should expect a significant proportion to be so. In fact, the distribution of genetic correlations between life history traits is very broad (Fig. 2), with 39% being negative but few estimated to be $-1$. In contrast, only 2.3% of genetic correlations between morphological traits are negative. Thus there do appear to be tradeoffs that will at least impede the evolution of life history traits.

IV. THE EVOLUTION OF LIFE HISTORIES

The starting point of many analyses of evolutionary change, most particularly that of optimality modeling, is the assumption that there exists some variable maximized by selection. The issue of what is being maximized has been the subject of much discussion, partly because the appropriate measure of fitness changes with circumstance and method of analysis. Broadly, measures of fitness can be separated into two groups: global measures and local measures. A global measure of fitness is one that involves the interaction of all life history components, the best example being Fisher's Malthusian parameter $r$. Local measures assume that maximization of a fitness component will also maximize the overall fitness of the organism: for example, a common local measure used in foraging theory is the net rate of energy intake. This is an appropriate measure if it can be shown that maximizing this rate does not detrimentally affect other components of fitness, in which case maximizing net rate of energy intake will also increase global fitness. Local measures are generally tailored for the particular analysis under consideration, but there now exists a general consensus, and more important sound theoretical rationale, of what global measures are likely maximized by natural selection.

A. Static Environments

A population growing in an unlimited, homogeneous, and constant environment follows the simple exponential growth function

$$ \frac{dN(t)}{dt} = rN(t) $$

$$ N(t) = N(0)e^{rt} $$

where $N(t)$ is population size at time $t$ and $r$ is the intrinsic rate of increase, comprising the difference between instantaneous rates of birth and death. Equation
The symbol $r$ is called the finite rate of increase and is sometimes used instead of $\lambda$. Suppose there are two clones with population sizes, $N_1(t)$ and $N_2(t)$, respectively, the first with an intrinsic rate of increase of $r_1$ and the second with $r_2$, with $r_1 > r_2$. The ratio of population sizes after some time $t$, given that both clones start with the same population size is,

$$\frac{N_1(t)}{N_2(t)} = e^{(r_1-r_2)t}$$

It is clear that as time progresses this ratio will increase, clone 1 becoming numerically more and more dominant in the combined population. This conclusion does not depend on the two clones beginning with the same population size: differences in starting condition simply accelerate or retard the rate at which clone 1 increases in frequency relative to clone 2.

For these two clones, an appropriate measure of fitness is $r$, since the frequency of the clone with the highest value of $r$ will increase toward unity. Thus any mutation in a set of clones that increases $r$ by changing rates of birth or death will increase in frequency in the population. There are no difficulties in assigning $r$ as a measure of fitness in the preceding circumstances. Difficulties arise, however, when sexual reproduction and age structure are introduced. Suppose we have a random mating population in which a mutation arises that increases birthrate or decreases death rate. Since the mutant will initially be rare in the population, its fate can be ascertained by considering the birthrates and death rates of the heterozygote alone. If the heterozygote's rate of increase is enhanced, the mutation will increase in frequency in the population, but its ultimate fate depends on the relative birthrates and death rates of the homozygotes and heterozygotes bearing the mutant allele. If the homozygote carrying both mutant alleles has a higher birthrate or a lower death rate than the heterozygote, the mutant allele will eventually be fixed in the population; otherwise the population will reach a stable polymorphism.

The general assumption, stemming from the work of Fisher, has been that $r$ can be associated with genotypes that follow particular life histories and that selection will favor that genotype with the highest value of $r$. In an age-structured population, the rate of increase is obtained by solving the characteristic equation

$$\int_0^\infty e^{-\xi}l(x)m(x)dx = 1$$

where $l(x)$ is the probability of surviving to age $x$ and $m(x)$ is the number of female births at age $x$. The discrete time equivalent of this is

$$\sum_{x=0}^{\infty} e^{-\xi}l(x)m(x) = 1$$

Note that in the discrete version the initial age is subscripted as $x=0$. The important issue to be considered is the fate of a mutant that increases $r$. Charlesworth demonstrated that to a rough approximation the rate of progress of a rare gene eventually becomes directly proportional to its heterozygous effect on $r$. In this case, the probability of survival of a mutant gene of small effect in a near-stationary population is largely determined by its effect on $r$. Following a more detailed analysis, Charlesworth (1994) concluded that for the case of weak selection and random mating with respect to age, the intrinsic rate of increase of a genotype or, more generally, the mean of the male and female intrinsic rates, provides an adequate measure of fitness in a density-independent and constant environment.

Lande tackled the problem of applying quantitative genetic theory to the evolution of $r$ in a population. Assuming weak selection, large population size, and a constant genetic variance-covariance matrix, Lande showed that in a constant selection regime, life history evolution continually increases the intrinsic rate of increase of a population until an equilibrium is reached. The intuitive appeal of $r$ as a suitable measure of fitness that receives qualified support from both population genetic and quantitative genetic theory.

If the population is stationary, $r$ is zero and the characteristic equation reduces to

$$R_e = \int_0^\infty l(x)m(x)dx = 1$$

$R_e$ is termed the net reproductive rate and is the expected number of female offspring produced by a female over her life span. The use of $R_e$ makes analysis easier and can be justified if $r$ is very close to zero or if the density-dependent or other factors that maintain the population at some relatively stable value do not impinge on the traits under consideration. For example, population size might be controlled by density-dependent mortality in the larval stage, while the object of study is the evolution of female age at maturity. In this case we can examine the relationship between the age
at maturity and fitness under the working assumption that genotypes do not differ in the characteristics of their larvae. Since population size is stable, the expected lifetime fecundity, $R_0$, is then the appropriate measure of fitness.

In some analyses, fitness is determined from the reproductive value at age $x$, $V(x)$

$$V(x) = \frac{1}{R_0} \int_0^x e^{-\lambda y} m(y)dy$$

The reproductive value of an individual of age $x$ is a measure of the extent to which it contributes to the ancestry of future generations. Williams (1966) postulated that natural selection maximizes $r$ by maximizing reproductive value at every age. A mathematically correct statement of the principle is "reproductive value at each age is maximized relative to reproductive effort at that age, although not necessarily with respect to effort at other ages."

The preceding approach has been very successful in accounting for variation in life history traits. An illustration of this is the analysis of the optimal size at first reproduction in *Drosophila melanogaster*. In *D. melanogaster* body size increases with development time and fecundity increases allometrically with adult body size, which is fixed on eclosion from the pupa. The age schedule of female births is triangular in shape (Fig. 3) and is described by the equation

$$m(x) = \frac{1}{2} c_1 L^2 (1 - e^{-d(L)}e^{-c_1 x})$$

where the $c$'s are constants, $L$ is thorax length, and $x$ is age. Thorax length, $L$, scales the age schedule of reproduction, larger females producing more eggs, but does not change its position. The constant $c_1$ is the product of two constants: the coefficient of proportionality within the allometric relationship between length and fecundity, and the proportion of eggs that fail to hatch.

Development time, $d(L)$, scales to body size according to the relationship (Fig. 3)

$$d(L) = c_2 L^2 + c_4$$

Thus development time scales allometrically with size except for the constant $c_4$, which represents time required for the eggs to hatch and the development within the pupa, both of these components being independent of size.

Because information on mortality rates are so poorly known, it was assumed that instantaneous rates remain constant in the adult and larval phases at $M_a$ and $M_l$, respectively. Probability of surviving the larval period, $f(L)$, is thus (Fig. 3)

$$f(L) = e^{-M_a L}$$

*Drosophila melanogaster* is a colonizing species and hence the appropriate measure of fitness is $r$. Combining the previous relationships, we obtain the characteristic equation

$$\sum_{i=1}^{n} \frac{1}{R_0} e^{-d(L)} e^{-c_i x} e^{-c_4 x} L^i = 0$$

where, for convenience age has been rescaled to begin at the first day of egg laying. Though this equation is tediously long, its solution presents no great difficulty. Briefly, the method is, first, to evaluate the series making

\[ \frac{1}{R_0} e^{-d(L)} e^{-c_4 x} L^i \]
use of the fact that it is a geometric progression and, second, to differentiate implicitly to obtain the optimal length at maturity. The optimum thorax length depends on adult mortality, larval mortality, and the constant initial “mortality” component (\( p \) or \( c_0 \)). Figure 4 shows the relationship between \( r \) and thorax length using values obtained from laboratory stocks and estimates from wild populations. The predicted maximum, 0.95 mm, falls very nicely within the observed range in thorax length of 0.90 to 1.15 mm.

### B. Stochastic Environments

Environments are typically variable in both time and space. While the assumption of a constant environment may be a reasonable first approximation in many cases, there will be many others in which variation cannot be ignored. The simplest case to consider is where the environment is stochastic and there is no cue as to future conditions.

In a stochastically, temporally fluctuating environment, the correct measure of fitness is the geometric mean of the finite rate of increase. The rationale for this measure is as follows: the size of a population after \( t \) time intervals is given by

\[
N(t) = N(0)\lambda_1 \cdots \lambda_t = N(0) \prod_{i=1}^{t} \lambda_i
\]  

from which it is readily apparent that the correct measure of fitness is the arithmetic average of \( r \).

Consider two genotypes living in an environment that comprises two types of year, “good” and “bad,” each occurring with equal frequency. In “good” years, genotype A has a finite rate of increase of 2 and in a “bad” year a finite rate of increase of 0.5, while genotype B has finite rates of increase of 1 and 1.1, respectively. The arithmetic averages of A and B are 1.25 and 1.05, respectively, but the geometric averages are 1 and 1.1. Thus genotype B has the highest long-term fitness although it has a smaller arithmetic finite rate of increase. Genotype A increases more than genotype B in “good” years but suffers a greater reduction in “bad” years. The relatively high fitness of genotype B resides in the fact that although it has a smaller arithmetic finite rate of increase, it also has a smaller variance in its finite rate of increase. In a stochastic environment the highest fitness may be gained not by the production of a single phenotype but by a genotype producing a range of phenotypes. Such bet-hedging strategies are common, an example being the production of both diapausing and direct-developing offspring in both invertebrates and plants.

### C. Predictable Environments

Most environments show predictable changes either as a result of biotic factors (e.g., overgrazing, community succession) or abiotic factors (e.g., winter, monsoons). Suppose factors that affect one or more life history characters are fully predictable—that is, given some
cue $E$ the life history characters will be changed from their present values to values that can be related to $E$ by some function. For each case there is some optimal combination of trait values: given the predictable nature of the cue means that the organism’s life history can be altered to accommodate the new conditions. Suppose, for example, that when the environmental conditions are $E$, the optimal clutch size is $C_i$. Upon receipt of the environmental cue, the organism produces the optimal clutch size, $C_i$. This relationship between trait value and environment is called a norm of reaction, whether or not it is optimal. Reaction norms are ubiquitous. Two examples of reaction norms are shown in Fig. 5. A reaction norm may be adaptive or a nonadaptive (even maladaptive) physiological response to the environment. To demonstrate that a reaction norm is adaptive requires that we use the set of tradeoff functions related to the environment to predict the optimal reaction norm. Consider, for example, the question asked by Stearns and Koella (1986, p. 894): “How should an organism encountering an unavoidable stress that results in slower growth alter its age at maturity to keep fitness as high as possible despite the constraints imposed by slower growth?”

To address this question, Stearns and Koella used the following hypothetical model

1. Fecundity increases with body size, which increases with age:

$$W(x) = c_1(1 - e^{-kx})$$

2. Adult mortality rate, $M_a$, decreases with growth rate according to the function:

$$M_a = \frac{c_2}{k}$$

3. Mortality rate of juveniles, $M_j$, is a function of the component of adult mortality, $\alpha$, plus an amount that decreases with age at maturity, $\alpha_1$, and growth rate, $k$:

$$M_j = c_3 + \frac{c_4}{\alpha + \alpha_1 k}$$

Stearns and Koella assumed that fitness is maximized by maximizing $r$. The predicted norms of reaction between age and size at maturity for four possible scenarios are illustrated in Fig. 6:

1. Mortality rate and growth rate independent ($c_6 = 0, c_9 = 0$). Decreases in growth rate, $k$, favors an
increase in the age at maturity and a decrease in the size at maturity.

2. Juvenile mortality rate increases as growth rate decreases ($c_1 > 0$, $c_4 = 0$). As with the previous case, decreases in $h$ favor an increase in the age at maturity and a decrease in the size at maturity.

The actual norm of reaction may have the same shape as case 1 or be sigmoidal.

3. Adult mortality rate increases as growth rate decreases ($c_6 > 0$, $c_7 = 0$). The optimum norm of reaction has the same shape as case 1 but is reversed, size at maturity and age at maturity being positively related.

4. The adult and juvenile rates are variable ($c_i > 0$, $c_i > 0$). The relationship between age and size at maturity is "klee-shaped." The portion of the curve in which size at maturity is an increasing function of age at maturity only occurs when adult mortality rates exceed juvenile mortality rates by a factor of 100 or more. Thus, generally the trajectory is determined by mortality rates of juveniles.

The message from the preceding analysis is that reaction norms can be quite complex and dependent not only on the functional relationships determining the tradeoffs but also the parameter values. The situation is made even more complex if the environment is spatially tradeoffs but also the parameter values. The situation is made even more complex if the environment is spatially

variable: in this case the appropriate fitness measure is the overall $r$, not the $r$ that is characteristic of a particular environment. This requires modifying the characteristic equation to

$$
\int p(h) \int f(x, h) m(x, h) e^{-x \mu} dh = 1
$$

where $p(h)$ is the probability of habitat $h$ occurring. The optimal combination of trait values can be markedly different from the average set determined by considering each habitat separately.

For reaction norms to evolve to produce the optimal response, there must be genetic variation in how individuals respond to the environment (i.e., there must be genetic variation for phenotypic plasticity). The evolution of phenotypic plasticity can be addressed using two apparently different mathematical perspectives: the character state approach and the reaction norm approach. Both approaches are actually interchangeable and each has advantages and disadvantages.

The character state approach is based on the assumption that the same trait measured in two environments can be considered as two traits that are genetically corre-
FIGURE 7 Phenotypic plasticity in two environments as viewed from the character state perspective. The panels on the left show the character states in the two environments $E_1$ and $E_2$, each line joining the trait values of a single genotype (the reaction norms of the genotype). The panels on the right show the regression of the trait value from the second environment, $X(E_2)$, on the trait value from the first environment, $X(E_1)$. A: The reaction norms meet at a single point beyond the range of the two environments. The correlation between trait values is $r = 1$. Note that parallel reaction norms also give a genetic correlation of $r = 1$ (mathematically this is because the lines meet at infinity). B and C: The reaction norms cross at several points within the range of the two environments. Depending on the distribution of intersections the genetic correlation will be positive but less than 1 (B), zero (C), or negative but greater than –1 (not shown). D: The norms of reaction intersect at a single point between the two environments. In this case the genetic correlation is $r = –1$. The triangles show a hypothetical optimal combination of trait values. Because in cases A and D all the points lie on a single line ($r = 1$), this combination may not be achievable. In all other cases, because in principle the distribution about the regression line is normal (i.e., no value is excluded), selection can move the population to the joint optimum.

Gorman determined that survival was a complex function of gall diameter, with an optimum at approximately 24 mm (Fig. 9). On the basis of analysis of 16 full-sib families, they obtained a linear relationship between final gall diameter and the time lag between oviposition and gall initiation (Fig. 9). Because the time lag showed variation associated with plant sibship but not among insect sibship, Weis and Gorman argued that time lag is a trait of the plants and not the insects. Thus time lag is, from the insect’s perspective, an environmental gradient associated with the plant. The heritability of the gall-size/lag time intercept was determined as 0.21 (SE = 0.18) and the slope as 0.54 (SE = 0.25). Although the former estimate is not significantly different from zero, the ANOVA indicated significant family effects.

Maximum survival occurs when gall diameter is approximately 24.3 mm, and hence we would expect selection to favor reaction norms that produced a gall of this size, regardless of lag time. In this regard it is significant that the greatest intersection of the reaction norms occurs at the modal lag time but gives a gall diameter of approximately 19 mm (Fig. 9). Thus the fitnesses of the different families are approximately equal in the most frequent environment but are lower than the maximum fitness. Selection should act to shift this area of intersection upward, which could be done by increasing the intercepts or changing the slopes. In the former case an overall increase of approximately 4 mm would achieve the required maximization of fitness within the modal environment, but in the latter case it appears that a much greater range in slopes must be achieved. Therefore, we might expect that selection will act most strongly in a
directional sense on the intercept, since the same change in all families is likely to have a greater impact on fitness than a similar change in the slope. This prediction can be tested by estimating the relative strengths of selection on the two parameters. As predicted, directional selection was stronger on the slope than the intercept (Intercept/Slope = 4.4), but stabilizing selection acted approximately equally on the two components (Intercept/Slope = 1.4).

This study reinforces the point that the presence of genetic variation is necessary for reaction norms to evolve to their optima, but at the same time the continued presence of genetic variation remains a phenomenon that itself must be explained.

D. Environments with Both Stochastic and Predictable Components

The most realistic scenario of the environment is that it has both stochastic and predictable elements. Consider, for example, the problem faced by an organism in which its propagules (eggs or seeds) can survive frost, and so can overwinter, but all other stages are killed by a frost. Now suppose the season available for growth and reproduction is long enough to accommodate on average two generations. Individuals that mature early in the season will lay direct developing eggs that give rise to the second generation. Individuals of the second generation will all lay diapausing eggs that will overwinter. However, individuals of the first generation that emerge late in the season face the problem that there may be insufficient time for a second generation in which case they should lay diapausing eggs. If the season is long enough to accommodate a second generation, diapause eggs have a lower fitness than direct developing eggs, because the former give rise to only a single descendant for that year whereas the latter give rise to many descendants. In an environment in which the day of the first frost always occurred on the same day or could be predicted without error by the first generation females, there would be a sharp switch between the production of diapause and direct-developing eggs. On the other hand, in general, the day of the first frost could only be assigned as a probability function: under this circumstance there will exist a period during the season during which the most fit strategy will be to lay a mixture of diapausing and direct-developing eggs, the proportion depending on the reliability of the cue.

The phenomenon in which a mixture of types is produced is termed bet hedging or risk spreading. It has been explored theoretically fairly extensively, but there have been few empirical tests. In particular, risk spreading could be achieved by a strategy of avoiding risks (e.g., switching to production of diapausing eggs whenever the probability of not getting in the second generation is less than 1), which has been termed conservative risk spreading or by a single
genotype producing a range of phenotypes, which has been termed diversified risk spreading. In an extensive review of insect data Hopper (1999) concluded that there is little evidence that risk spreading has been a major factor in the evolution of insect life histories. The importance of risk spreading in the evolution of life histories remains to be satisfactorily investigated.

See Also the Following Articles
Biodiversity, Evolution and • Biodiversity Generation, Overview • Phenotype, a Historical Perspective

Bibliography


LIMITS TO BIODIVERSITY (SPECIES PACKING)

L. B. Slobodkin
State University of New York-Stony Brook

I. Introduction: The Biological Packing Metaphor
II. Two Species in One Container
III. Expanding the Theory of Packing
IV. Species Diversity and Species Packing
V. Conclusions

GLOSSARY

chemostat Apparatus for growing microorganisms in a continually replenished medium.

community A multispecific aggregation of organisms in a particular location that may interact with each other.

competition More than one species utilizing one or more common resources.

ecological niche The set of requirements that must be met if a particular species is to survive. It is sometimes used to mean the place in which those requirements are met.

ecosystem A region with more or less clear boundaries, which contains a particular set of species and may be characterized by some set of meteorological, climatological, and geochemical properties.

invasive or alien species Species that have recently colonized some geographic regions different from the one in which they were initially described.

isocline A line along which some property remains constant.

multidimensional niche A tempo-spatial region defined by meeting a set of different requirements for viability of a particular kind of organism.

niche dimensions Ranges of values of environmental measurements in an ecological niche. A range of temperatures, salinities, or oxygen concentrations may be niche dimensions for a population of fish.

packing The placement of objects in a container.

reification The assignment of empirical reality to the referenda of a word or theory, regardless of the existence of any such referenda.

species diversity The number of different species in some area of interest. Global species diversity refers to all species. Local species diversity refers to some geographic region such as Hawaii or New York City.

species packing The study of how species on the same trophic level coexist in a limited region or container.

SPECIES PACKING is the study of the co-occurrence of different species of organisms in a particular sampling area. Certain regions are known to contain many species of particular groups. Samples of coral reefs, the middle depth sea bottoms, and some tropical forests contain many species while samples from arctic lakes, northern forests, and estuarine bottoms have relatively few. Since it is known that several species simultaneously introduced into a container cannot persist, there is a question about how so many species can be packed into certain situations and not into others. The theory of species packing was a response to this problem.
I. INTRODUCTION: THE BIOLOGICAL PACKING METAPHOR

Different locations vary conspicuously in the number of species that are present. Some, like coral reefs and tropical rain forests, are famous for their high species diversity, while others, like some Alaskan forests, are notably monotonous in the number of kinds of trees that are present. Differences in diversity have attracted many researchers, although as yet there is no one all-encompassing theory of species diversity. In fact, the overall term "species diversity" seems more useful at the level of scientific policy and public relations than it is at the level of scientific discourse.

There are many theories and models of species diversity, which differ in their foci and assumptions. The theory of species packing is concerned with one aspect of diversity, specifically the coexistence of populations of different species located in a particular space. It carries elements of the concept of "species richness" in that the greater the species richness of a region, the more species have presumably been "packed" into it. It also carries implications for "realized species niches" because it might be inferred that the more species are packed into a region, the narrower the realized niches.

Like so many of the terms of theoretical ecology, "species packing" often is used in a metaphorical sense, which can be misleading. It is best to begin by analyzing the concept in a literal and elementary way.

The word "packing" has at least two meanings, which can cause confusion. It may mean placing objects inside a wrapper or container without altering them, or it may mean squeezing them in to a fixed-size container. Contrast the packing of glass tumblers with that of pillows. Tumblers are packed into a container but if they press hard against each other they shatter. Pillows can be squeezed into a container and fluffed up again when they are taken out.

The possibility of either type of species packing presupposes some sort of container into which things are packed. Is the inferred container equivalent to an environment? Does an environment have fixed geometric dimensions like a wooden box? A pond, a log, a rotting fruit—all can, in fact, be thought of as containers. However, environments can be defined in ways that may have more meaning to investigators than to inhabitants. Are the boundaries of a study area like La Selva, a political unit like Costa Rica, or a geographic area like Central America in some sense meaningful packing containers?

We will introduce the mathematical theory of species packing, examine how the theory has been exemplified in laboratory and field experiments, and, finally, consider the implications of species packing for field observations.

II. TWO SPECIES IN ONE CONTAINER

A. Theory

The problem of species packing was crystallized in the apparently simple question: "Why are there so many kinds of animals?" Thereafter there has been a continual stream of often controversial, difficult, and sometimes acrimonious papers on this subject.

1. Lotka-Volterra Equations

Species packing theories are concerned with species in an implicitly assumed container. The boundaries of the container are usually considered to be permeable to the flow of resources and energy but, unless specified, impervious to the passage of organisms. These are important assumptions.

The formal theory of species packing originated from the equation systems of Lotka, Volterra, and D'Ancona. Lotka considered that in the entire biosphere all individual organisms have similar resource requirements, and no resource is in infinite supply. Therefore, at some level there must ultimately exist a limit to the abundance of organisms and an inverse relation between the number of species and the number of individuals per species. Further analysis is required before this assertion becomes interesting.

Volterra and D'Ancona considered several species competing with each other for resources in a space through which resources flowed at a constant rate. Competition between species A and B can be considered in terms of pairs of isoclines on a two-dimensional phase space with axes $N_A$ and $N_B$.

$$\frac{dN_A}{dt} = 0 = K_A - N_A - \alpha N_A$$

$$\frac{dN_B}{dt} = 0 = K_B - N_B - \beta N_A$$

in which the subscripts refer to the two species. The $K$ values refer to the number of organisms present at what was called "population equilibrium" or "population saturation."

It was assumed that any single species population would remain numerically constant after $K$ had been achieved, if the flow of resources and the physical condi-
tions did not change. These simple equations are the starting point for many species packing theories.

The K and N values were initially considered as counts of organisms, but as the theory developed the K values were taken to designate some resource divided up among the organisms. It often helps to think of species abundance as being directly translatable to rate of consumption of a temporarily renewable resource.

It was usually assumed that any individual organism will negatively affect the well-being of all the others in the designated space, as measured by their birth and death rates.

Organisms can affect each other by reducing the availability of resources. They can also alter the chemical or physical properties of the local environment. These alterations are referred to as “conditioning” or “crowding.” Yeast cells secrete alcohol, which affects other yeasts and bacteria. Beetles and mice change the odor of the environment. Mussels and barnacles physically crowd each other.

In elementary packing theory, a constant, α, represents the relative amount of the resources of species A consumed by each individual of species B and, conversely, β represents the crowding of the individuals of species B by one individual of species A. In the multispecies case, α, is taken as the effect of species j on members of species i.

2. The Ecological Niche

The possible outcomes for two species were often represented graphically in elementary texts by using a phase diagram with abundance or concentration of each of the two species as axes. The persistence of both species populations is only possible if the two isoclines cross.

For species Ns, the isocline connects the point (Ns = Ks, Ns = 0) to the point (Ns = 0, Ns = Ks/α). At each point on this isocline, individuals of the relevant species are crowded enough that births and deaths are equal. The crowding is exercised by either con-specifics, the members of the other species, or some mixture of the two.

If two species are inoculated into a space in which each species can survive if alone, there are basically three possible outcomes:

• If an individual of one species impacts the con-specifics more than hetero-specifics, both species may persist. The two isocline cross at a “knot.”
• Another possibility is that individuals of both species are more sensitive to the presence of the other species than to the presence of con-specifics. In this case, the outcome depended on the initial concentrations of the two species. The isoclines cross at a “saddle.”
• The final case is one in which one species has a stronger effect on the other, but the second species does not have a very strong effect on the first. The isoclines do not cross.

These three cases are related to simple definitions of the ecological niche. In the first case, the container is considered to contain regions from two ecological niches. In the second case, the container is in the intersection between two niches, and in the third case the container is within the niche of the victorious species.

B. Laboratory Experiments

While the elementary theory is primarily of pedagogical rather than practical value, it did inspire several laboratory population studies in which populations of two species were actually placed in a container, provided with renewable resources, and permitted to interact.

The early investigations on macroscopic organisms required long periods of observation and repeated tedious counting. Some later laboratory population experiments were done. Current studies of laboratory population dynamics mainly use microorganisms and are primarily of genetic and evolutionary significance. Even in these studies there are significant ecological differences between results derived from liquid culture in test tubes, in chemostats, and those from colonies on semisolid gels.

Without attempting to review all laboratory studies it is possible to summarize insights that have proved relevant in designing field studies and that have suggested theoretical analyses or may be expected to do so in the future.

• Laboratory studies have demonstrated that a single container can be divided in ways that can be more interesting than simply differences in quantitative utilization of resources.

For example, when populations of a moth and of a beetle were placed in a container with intact wheat grains or with flour, the moths were eliminated. Both species survived in cracked wheat. Broken bits of glass capillary tubing added to flour also permitted two species persistence. The moth larvae were being eaten by the beetles unless they were sheltered inside the grains or in the glass tubes. This illustrated that spatial complexity can modify the relation between competing populations. In the absence of physical refuges, the relation-
ship between the moth and the beetle becomes a predator-prey interaction combined with some competition for food. Examples in which the interaction between several species can vary between competition and predation are not uncommon.

No populations, except for those of some microorganisms, grow according to the Verhulst Pearl logistic equation of sigmoid growth. This is interesting because there still are those that will defend the use of this counter-to-fact equation in development of theory.

- As populations become more crowded there are changes in birth and death rates. These changes are always accompanied by other physiological, anatomical, and behavioral changes.

There is abundant evidence of psychological and physiological change caused by individual social history in mammals. While the effects of past history may be most dramatic in mammals, they have been found in essentially all animals and plants. Crowded trees show different shapes than those grown in uncrowded situations.

- The levels of population density that can be reached during population growth experiments are very much in excess of what the organisms would tolerate in nature, as indicated by the ubiquitous tendency of animals to escape from these containers in any way possible.

In one study it was possible to compare responses of house mice in closed and open spaces. When the confined mice became sufficiently numerous to empty their food trays they immediately developed reproductive and behavioral pathologies. Unconfined mice remained in a local area around their food trays until they could empty them after each feeding. They then immediately dispersed.

FLOUR beetles when crowded move to the surface of their medium and show a strong tendency to fly. This cannot be demonstrated in population vials but becomes apparent if the flour containing a crowded population of beetles is removed from the vials or the vials are unstopped (personal observation). Dispersal when crowded is also found in other grain pests.

In Daphnia, winter eggs (ephippia) are produced at an early stage of food stress. This permits descendants of the population to survive despite whatever occurs in the experimental container. Any pond or lake that contains Daphnia or other Cladocerans will contain abundant ephippia in its sediments. These may hatch after a long dormant period, introducing genetically distinct animals from an animal seed bank for many years, even if the emergent animals of a particular genotype cannot survive in most years.

Even Hydra, which have no locomotory organs, respond to crowding and food shortage by floating free of their substrate, which removes them from a locally crowded situation.

- Laboratory populations can sometimes be used to make some inferences about what may be occurring in nature, but the inferences are not necessarily direct.

Early students of Daphnia assumed that Daphnia condition the water in which they lived. It was later demonstrated that the sizes of Daphnia populations in the laboratory, over a crowding range 10 times that ever observed in nature, were linearly dependent on food supply. If conditioning was of importance, this linearity would be impossible and there would have been a curvilinear relation similar to that found in the relation between nutrient flow and population size in chemostat cultures of microorganisms.

This does not imply that Daphnia populations in nature are regulated by their food supply. It does imply that they are not regulated in any significant way by causing chemical deterioration of their environment, because over a range of crowding that greatly exceeds any that has been found in nature, no chemical effect has been found.

Clearly, laboratory experiments and theoretical models cannot adequately imitate nature. However, an experiment can be designed to consider a wider range of conditions than those that occur in nature. In that sense, nature becomes a subset of the experimental world.

In some cases there is a clear logical transition between theory, laboratory experiment, and field experimentation. Gause mathematically considered predation on a two-species competition system. He concluded that the coexistence of two species could, in principle, be stabilized by predation.

In laboratory populations, the coexistence of brown and green species of Hydra was facilitated by either removal of some of the animals or by maintaining them in low illumination, demonstrating that coexistence could be stabilized by either a biological or a physical change in the container.
In a field experiment, Paine found that heavy predation of starfish on mussels prevented the mussels from eliminating their competitors. In this case, an hypothesis proposed a possibility, laboratory experiments demonstrated its reality, and the field experiment demonstrated its importance in nature. But even in this case there were dangers of misinterpretation. The theory and the laboratory and field experiments demonstrated that under some circumstances a predator might enhance the number of surviving species on the next lower trophic level.

Any loss of organisms imposed on a multispecies system relieves, to some extent, competitive pressure on the survivors, so that weaker competitors may survive. There is, however, an abundance of cases of predators locally eliminating prey species, as is now occurring due to feral cats in Australia and Hawaii.

III. EXPANDING THE THEORY OF PACKING

Many formal theories of species packing begin by complicating this two-species situation. Among the obvious complications are those in which the isoclines are permitted to be curved lines, which may be interpreted as social interactions of various sorts. Also, changing the sign of the interactions could crudely mimic predation or symbiosis. Loss of organisms across the "container" boundary could be thought of as mimicking predation.

A. Expansion of the Theory to Multiple Species

The more general question is: How many species can be introduced into a space? Multiple species packing theory is mathematically more difficult because a multispecies theory cannot be conveniently represented as isoclines on a two-dimensional phase space. If a third species were to be added, the phase space would have to be three-dimensional and isoplanes would be needed instead of isoclines. For coexistence of three competing species, the three isoplanes would have to intersect somewhere in the region below the intersection of any two of the species isoclines.

In the case of multiple species packing, there can be a region in the multidimensional phase space in which resources are divided in such a way that all species can coexist, but only if each species is relatively strongly self-limiting and relatively mild in its effects on the other species present. The multiple species theory requires clarification of the notion of ecological niche. The ecological niche of a species is often considered as the full set of measurements of environmental properties that are relevant to the survival and persistence of that population. Hutchinson and his students developed this into an image of a multidimensional hyperspace.

Any loss of organisms imposed on a multispecies system relieves, to some extent, competitive pressure on the survivors, so that weaker competitors may survive. There is, however, an abundance of cases of predators locally eliminating prey species, as is now occurring due to feral cats in Australia and Hawaii.

B. How Similar Can Coexisting Species Be?

MacArthur and Levins asked: What might be expected if a third species invaded a space in which two species are already resident? They simplified the situation by considering that all species concerned interact through competition on a single niche dimension, or at least a linear representation of a multidimensional niche. They also assumed simple distribution of resources along that dimension. If, for example, the niche dimension is some particular prey, one of the species might eat large specimens while the other eats only small ones.

They considered that the niche of any species would have a shape on any single niche dimension. If the niche can be represented by a straight line, the probability of an individual finding a bit of resource at some point on the line can be represented as an ordinate, generating a two-dimensional shape. If the shapes for several species overlap, then the intensity of competition is measurable in terms of the probability of the several species striving for the same resource—the product of the individual species-specific probabilities.

They could then consider species packing along a niche dimension. Given their assumptions, packing can be closer on any one dimension as the number of niche dimensions increases. If species are arranged in a one-dimensional array along a line, the closest stable packing is admissible when the niche spaces are rectangular and all K values are equal. These conclusions were strongly dependent on their assumptions.

More general conclusions were that closer packing is possible if niche dimensionality is high, niche breadth...
is small, and the environment is predictable and has high productivity. Centers of adjacent niches are in a multiplicative series of numerical value approximately 1.1.

If the single linear dimension of competition is considered more realistically, so that there is a multiplicity of niche dimensions, MacArthur and Levins concluded that interspecific competition could be less important. In fact, it is sometimes considered that multispecies packing involves as many niche dimensions as there are species. Tilman considers that this is only valid if the organisms are not fixed spatially (i.e., they are free to move about in the container).

Various authors, by slightly modifying the mathematical assumptions, could reach a variety of conclusions about the packing of niches. For example, it was possible to plausibly argue that niches could sometimes be packed infinitely closely.

MacArthur and Levins suggested, as a matter of convenience, that the shapes of niches projected onto a single niche dimension might be thought of as approximating normal distributions. Other shapes are, of course, imaginable.

Species packing became a rich field for construction of theories, which were not necessarily connected to specific biological observations and more. These theories each make slightly modified assumptions or focus on special cases of previous theories. For example, Roughgarden and Feldman considered the significance of shape differences as if the shapes could be empirically demonstrated. Distributions with "thick" tails permitted more species to coexist than those with thinner tails.

There were also studies of multispecies biological situations under natural or artificial circumstances, which used the term "species packing." In these the connection with species packing theory was sometimes explicit but very often the general conclusions of theory were presented in the context of field data without explicit theoretical derivation. This appears to be due to the fact that explicit formal theory production becomes very difficult once more realistic elements of natural complexity are involved. This is typical of mathematical models, at least prior to the recent development of powerful computing machinery.

Sometimes the models were sufficiently complex, combining ecological and evolutionary predictions, that self-contradictory assertions were made within the same theoretical argument and were not noticed for years. For example, one study published in 1983 assumed absence of any genetic heterogeneity and then proceeded to discuss the genetic effects of selection, as noted by Taper and Case.

C. Niche Separation and Anatomical Difference

The exact ratio of minimum possible distinction between adjacent species can be derived from special, simplified, and somewhat arbitrary assumptions. Depending on assumptions, it can be asserted that an infinite number of species can coexist, that an infinite number of species could coexist were it not for statistical variance in the parameters, that coexistence is contingent on number of dimensions in the hyperspace, or that species must differ anatomically by a specific ratio in order to coexist.

Hutchinson suggested the possibility that one might find in nature that niche hyper-volumes of the most closely similar, often congeneric, coexistent species would actually differ from each other by a factor whose value might be approximately 1.3. This seemed reasonable based on the common observation that genera are often represented in particular locations by large, small, and medium-sized species. Attention was focused on this by Lack for Darwin's finches, but its occurrence can be easily confirmed by examining intertidal snails of the American northeast coast or the eastern shores of the Mediterranean Sea (personal observation).

"Assembly rules" for competing species on islands were developed in part on the basis of size. These results were criticized on statistical grounds. Were the species distributions really departures from random?

One major step was missing in attempting to demonstrate that a theoretical niche overlap in a simplified mathematical model predicted a morphological ratio between anatomical parts. While Hutchinson expected that trophic structures like jaws might show the appropriate tightly packed ratio, the theoretical connection between niche shape or size and body shape or size seems absent except in very special cases.

It was never really clear why there should be a relation between the overlap of niche hyperspaces and any obvious anatomical differences. There is no rigorous and general way of theoretically deciding which of the infinite possible number of anatomical measurements that might be made is appropriate in any given case. By choosing the measurements correctly or by manipulating the data, almost any ratio might be achieved. Simberloff provided a firm critique of measuring the niche separation by anatomical differences. There is a clear connection between body size and the utilization of preexisting nest holes by snakes, owls, chipmunks, weasels, and others, but any pair of nest hole utilizing
species is almost certain to differ in many niche dimensions.

IV. SPECIES DIVERSITY AND SPECIES PACKING

Does species packing set limits to species diversity in nature? This would require that natural aggregations of organisms are organized into communities and their component guilds, which would act as "containers." Are natural aggregations of organisms in nature strongly mutually interdependent so that "packing in" more species or removing species already present should have discernible effects on those already present? Do cohesive communities of organisms or ecosystems actually exist in nature? Early ecosystem theorists believed in a "holistic" or emergent concept of ecosystems and communities. If these were real, they would constitute containers in which packing theory might be applicable.

There is an enormous literature related to how many species occur in particular locations. Recently there has been a weakening of the basic notion of a natural community as a "container" in any serious sense. There is massive concern about the dangers of invasive species, but recent studies have returned emphasis to individual species and almost discarded the idea of integrated communities. There is a general condemnation of invasive species but they are particularly interesting from the standpoint of species packing. If invasions do not occur, it may be due to lack of opportunity. Large mammals cannot float across an ocean or logs or hide away in ballast tanks. However, if a species can invade a region, then in some sense the region was not packed full, or perhaps successful invasions necessarily result in displacements of resident species.

Ruiz estimated that more than 90% of alien species in estuaries have made no discernible impact on the species diversity or species abundance distribution of the estuary. Levine and D'Antonio report that a "consistent positive relation between exotic species abundance and resident species diversity ... [suggests] that invaders and resident species are more similar than often believed." Apparently some invasive species fit into niches that were in some sense empty.

To what degree is competition, and therefore species packing theory, a serious factor in nature and how could one tell? Attempts to answer this question sometimes hinged on naturalists' insights, sometimes on various statistical analyses, and sometimes on detailed natural history and experiments conducted in more or less natural circumstances. Each of these has its strengths, advocates, and opponents.

In what sense do communities depend for their continuity on particular species? American chestnut trees constituted almost a third of the large trees of the southeastern American forests in the early 19th century. While small specimens still exist, the chestnut trees were essentially wiped out by the Chestnut blight. The forest did not disappear. As noted by Hairston et al., the southeastern forest is as dense with trees as it has ever been. The general appearance of the forests has not changed, but the oaks, which have to a large degree replaced the chestnuts, produce volatile organic compounds that are of biogeochemical significance (M. Lerda, personal communication, 2000).

Davis, studying the northward expansion of forests after glacial retreat, noted that each individual species seemed to migrate at its own rate—there was no movement of the forest as a community marching together. Whittaker noted that distribution of trees on various environmental gradients was as individual trees, not as communities.

V. CONCLUSIONS

Theory, experiments, and natural history all suggest that communities are not tightly organized, so that species packing may not be strongly relevant in nature. In short, the popular image of an ecological community as an airplane in which each part has a vital role for the integrity of the whole is dubious. There may be groups of species, in which each one is closely connected to a few others but only loosely connected to other groups.

The term "community" was once extremely useful and is still of pedagogic value if carefully used. As reified objects for research, the concept of communities is now threatening to become what Simberloff and Dayan have referred to as a "pancreston," an idea as likely to generate confusion as enlightenment.

The prospect of a general species packing theory has melted away. Despite the unreality of the models, they did direct attention to what seemed to be a real phenomenon and encouraged experimentalists and field biologists to ask important questions.

Differences in species richness can be partially explained by a multiplicity of factors that do not necessar-
AILY relate to each other so as to permit formation of a single coherent theory. Each of the multiplicity of theories of diversity focuses on one or a few of the empirical factors that are known to enhance or diminish the possibility of species coexistence.

Many of these can be seen in laboratory experiments, which have the advantage of clarity but may be of questionable applicability. For example, local geometry complexity influences species diversity in nature as well as in the laboratory.

Certainly in some cases the term “species packing” is used in the sense of species being squeezed into a space. In several cases individuals of particular species in species-rich regions are believed to have a narrower range of activities than individuals of the same species in species-poor situations. Diet or nest sites may be more restricted. In these cases the individual organisms can be imagined to have been constricted by some packing process, like individual pillows in a crate but even this has two possible meanings depending on whether we are concerned with the population level or with individuals.

In comparing different locales, the range of variation among organisms within a unispecific natural population may be reduced when more other species are present. In the case of comparisons of different islands or lakes, this might be tentatively attributed to species packing.

Returning to the initial analogy of packing actual objects, an island is clearly a container. But if packing means filling the ecological space, either by pillows or tumblers, we would not expect islands or speciose lakes to easily admit invasive species. In lakes there may be enormous species richness of fishes, as in the ancient African lake cichlids, but I don’t know of any comparisons showing that species-rich lakes show less within species, among individual variation, than species-poor lakes. Is the attempt to crowd multispecific collections of fish together in the same lake equivalent to packing glass tumblers rather than pillows? There is a general impression that species rich systems seem at least as likely to be invaded by exotic species as species-poor systems, violating our sense of what packing might mean.

If the container walls are not apparent but among individuals variation is reduced in one or more populations of a species, is this a sign of species packing? Does the mere fact that among individuals variation is reduced when more species are present imply that there must exist a container wall which may not be obvious?

The overall conclusion is that the theory of species packing does not conveniently predict very much about natural systems but that the images of packing that it generates do informally suggest interesting phenomena to look for.

See Also the Following Articles

**COMPETITION, INTERSPECIFIC • DIVERSITY, COMMUNITY/REGIONAL LEVEL • HABITAT AND NICHE, CONCEPT OF • PLANT INVASIONS • SPECIES-AREA RELATIONSHIPS • SPECIES COEXISTENCE • STRESS, ENVIRONMENTAL**

**Bibliography**


Treberg, M., and Husband, B. (1999). Relationship between the abundance of Lythrum salicaria (purple loosestrife) and plant species richness along the Bar River, Canada. Wetlands 19, 118–123.
LITERARY PERSPECTIVES ON BIODIVERSITY

William Howarth
Princeton University

I. Literature and Science Relations
II. Language in Science and Literature
III. Etiological Tales of Nature
IV. Pastoralism in Poetry and Scripture
V. Colonial Expansion and Taxonomy
VI. Darwinism and Literature
VII. Ecology and Industrialism
VIII. Land Ethics and Environmentalism
IX. Natural Rights and Animal Fables
X. Ecocriticism and Biodiversity

GLOSSARY

contingency Possibility or uncertainty; an event that may occur but is not likely.
ecocriticism Interdisciplinary study of literature, history, religion, and philosophy with an emphasis on places, evolutionary biology, and environmental problems.
etiology The study of causes or origins; stories that explain the origins of phenomena.
land ethic Valuing land as part of a biotic community, not merely as property.
literature Imaginative and crafted writings, in the form of poetry, prose, fiction, or drama.
nature The material world and the physical forces or processes that control it.
pastoralism Agrarian life and work; also literary accounts of rural life, often simplified or idealized.

“DIVERSITY” REFERS TO DIFFERENCES or variations; “biodiversity” refers to the variety among biological species, sometimes called species richness (Nudds, 1999). Biodiversity is a recent concept, used since the 1980s to publicize losses in wildlife habitat, genetic variety, and number of species (Wilson, 1992). Field and laboratory studies confirm that natural diversity sustained a healthy planet until human beings began to acquire global dominance through technology and urbanization. Recent policies of sustainable development, especially in first-world nations, seek to restore biodiversity with land-use planning, protection of endangered species, and preservation of wild habitats. Although most definitions of biodiversity focus on its economic value, something beyond cash receipts must account for the continuing interest of societies in nature, wildlife, and open space (Werner, 1990).

The idea of biodiversity traces back to early concepts of variety and unity, which have their sources in history, philosophy, and literature. These cultural fields have recognized the need for biological variety since the first days of recorded time. Literature comprises many acts of language, from ancient folk chants to poetry on the Internet, often for entertainment and instruction, the dulce et utile praised by Cicero in his writings on rhetoric. The main goal of literature is to expand and extend...
the human imagination. By simulating experiences and performing ideas, literature helps affirm the value of diverse places and species, through a wide range of styles and themes. Not all writings are considered literature, just as all experiments are not science. Like professional science, the ideal of literature implies standards of custom and cultural decree (Turner, 1985).

I. LITERATURE AND SCIENCE RELATIONS

Over the centuries, literature has come to mean imaginative writing, the product of disciplined craft that creates drama, fiction, poetry, or essay. Critical theorists disagree as to whether the prime agent in literature is content or form, conveying a message or shaping an attractive verbal structure, like a story or a poem. In the balanced formula of British poet and critic Samuel Coleridge, poetry is "the best words in the best order," which affirms that an ideal of quality guides literary writers. But "quality" is an intrinsic value, one that clashes with the extrinsic aims of science. Science seeks an unbiased view of what nature is and how it operates, while literature observes those conditions through the filters of human belief and emotion.

Although novelist C. P. Snow (1959) believed that science and literature are opposing cultures, recent observers find more congruity in their aims. Both disciplines use observational methods to detect patterns and to record them for the benefit of posterity. Biologist Stephen Jay Gould (1989) sees science and literature as forms of historical narrative that describe contingency, the recording of uncertain, unexpected events just beyond current knowledge. Literature uses contingency in plot sequences—as events connect and readers turn pages “to find out what happens next.” Scientists unravel nature’s “plots” by reconstructing them as predictive narratives.

Literary forms represent natural diversity with varying degrees of accuracy. Drama is imprecise because it relies on social situations and voiced sentiments, enacted on a physical stage. Painting and photography may document a place or species, but pictures without words are not sufficiently interpretive. Poetry (whether epic or couplet) and the various forms of prose (novel, story, essay, memoir) are considered more comprehensive because they both describe and evaluate, mingling sensory experience with calculated thought. The meaningful connection of image and word: literature pursues that goal in order to represent biodiversity for the benefit of human understanding.

II. LANGUAGE IN SCIENCE AND LITERATURE

Human beings share a common attribute in language, the instrument they use to construct meanings. The path of language moves from variety to unity: many words exist; when combined they form phrases that express thoughts. A thought may be social or personal, as it creates a send-and-receive cycle that communicates. To meet the human need to understand both surface and depth, literature often asserts that natural objects are not alien but connected to human emotions. As Henry David Thoreau wrote in Walden (1854), “Shall I not have intelligence with the world? Am I not vegetable mould myself?” Seeking knowledge of nature's variety is thus a mutual goal for literature and science, but sharp differences exist in their methods.

Scientists prefer to use words precisely, to convey strictly limited meanings. Literary writers recognize that words have long histories, in which meanings change across time and space. As philologists learned in the 19th century, words behave like organisms (growing from root to stem to branch; cross-breeding and migrating to other habitats). Some words are viruses, infecting a host, while others lie in dead heaps, like sediment. Through these morphological changes, words shape human values to form and reinforce perceptions, beliefs, and opinions.

Words are also representational, using speech sounds to create signs for rain, snow, leaf, and rock. Combinations with other words produce symbolic and metaphorical overtones, suggesting layers of possible meaning. Ezra Pound's haiku-like poem, “In a Station of the Metro” (1916), compresses a startling comparison of culture and nature into two lines: “The apparition of these faces in the crowd/Petals on a wet, black bough.” Such powers of analogy give words a suggestive and speculative function, which lends itself to stories with a didactic purpose.

III. ETIOLOGICAL TALES OF NATURE

The earliest surviving fables and myths are etiological tales, told to explain how the world began, why rain falls, and when crops will bloom or die. Among prehistoric people, nature stories about diverse and complex phenomena became a form of cosmology, providing the basis for ritual ceremony, whether for medical or spiritual healing. Early ideas of disease, as a condition of imbalance or disharmony, encouraged the symbolic
belief that the earth was also a vast "body" that may be cleansed and healed. Some observers (Buell, 1998) note that medicinal rhetoric continues to shape current discourse on toxicity, pollution, plague, and other forms of ecodisaster.

Early tales are efforts to define the myriad natural forms that shape everyday life, from land and species to climate and seasons. At the same time, literature may portray human efforts to resist or control nature, commonly through acts of labor and art. Farmers drain wetlands in order to plant crops on dry land. They weave nets to lift fish from the sea, then bury the fish with seeds to help new life emerge. Stories about these actions form early contributions to understanding biodiversity by recognizing the complexity of nature and the intricacy of its related elements. In a Navajo creation tale, "Changing Woman and Her Hero Twins," the storyteller pours tinted sands on the earth, forming a series of objects and events that are as multiple and shape shifting as the identity of Changing Woman, an earth goddess who has the power to create.

Yet by emphasizing human impact, many early stories also teach that humans and nature are divisible and separate. This idea is distinctly Judeo-Christian and supports the very concept of "nature," meaning all that is born, rising from a source. No equivalent term appears in Asian culture, where the physical world connects humanity to the wider universe (Torrance, 1998). The Old Testament tells two different stories of human creation. Man is made separately, in the image of God (Genesis 1:26), but is also formed from dust (Genesis 2:7). The first story appears to justify human dominion over other creatures, while the second suggests a humbler bond with the earth. Belief in a separate, special creation of human beings continues to guide modern fundamentalists, who fiercely resist Charles Darwin's unitary theory of evolution. By including two narratives, the Bible leaves open the problem of literal versus metaphorical interpretation.

IV. PASTORALISM IN POETRY AND SCRIPTURE

In the early civilized era, 4000 to 500 B.C., ideas and tales of nature's variety emerged from the extractive activities of hunting, fishing, farming, herding, logging, and mining. By learning to harvest natural materials and process them into goods, humans became experienced observers of seasonal events, of plant and animal lives, and of locations favorable to cultivation. Land use and property rights became the basis for literacy, trade, law, and tribal identity (Crumley, 1994). Nations root themselves in physical places, and when people are forcibly displaced, as in the case of the early Jews, they endure their nomadic years with consoling legends of a Messiah, a warrior-shepherd who will return them to the lost homeland. Such early stories reveal a basic function of literature, to reflect the value of places and how human presence changes them.

In the Old Testament, land is an element made by God, then granted to humans to clear and own. They receive this gift in innocence, but later sin and must accept the burdens of mortality: labor, pain, disease, and death (Genesis 1:3). The status of wilderness is even less benign: uncultivated places, like mountains and deserts, display God's power and offer a test to the faithful. In the story of Noah (Genesis 6-9), God directs Noah to build a great Ark and bring to it at least two of every creature that lives, male and female. Then God floods the earth with rain, until only the Ark and its inhabitants survive. The story illustrates not just divine power, but also how human stewardship may preserve the earth's vast diversity of species. The later vision of Isaiah predicts a savior who will restore the earth to its unevent state, when all species live together in harmony: "The wolf also shall dwell with the lamb, and the leopard shall lie down with the kid; and the calf and the young lion and the fatling together; and a little child shall lead them" (Isaiah 11:6). This scene, oft painted by the 19th-century Quaker artist Edward Hicks, represents the differences between a creation marred by original sin and one healed by the world's innate bounty.

In times of constant social change, literature offered comforting accounts of bygone days and places. The very concepts of city and country, or urban and rural land, arose out of literature that invested both realms with a complex relation, possessing zones and boundaries of separation but also connection (Williams, 1974). As urban centers grew into the great city-states of Alexandria, Athens, and Rome, the poets Hesiod, Theocritus, and Virgil wrote pastoral poetry, a literary form known variously as bucolic, georgic, and eclogue; verses that portray rural life as gentle, principled, and close to the soil. Jesus of Nazareth also adapted pastoralism in his parables about fish and seeds to preach transcendence of earthly limitations, a literary device equally effective with audiences in rural Judea and in the fast-growing cities of the Mediterranean basin.

After the Roman empire shifted from paganism to Christianity in the third century A.D., early Christians saw their pastoral mission as spreading the Gospels
throughout Europe. Monastic enclaves preserved both classical science and literacy in the Dark Ages (400–800), between Rome's fall and the revival of learning in courts and cathedral towns. During the Middle Ages (800–1400), infectious plagues reduced populations and relieved pressure on Mediterranean ecosystems, still recovering from Roman-era deforestation, erosion, and pollution (Hughes, 1994). Medieval literature and science became intertwined endeavors, as clerics and scholars sought to describe the fixed Ptolemaic universe, an effort that culminates in Dante's epic poem, *The Divine Comedy* (1321). Medieval culture's fascination with alchemy likewise fed early studies of natural history and chemistry, described in several of Chaucer's *Canterbury Tales* (1400).

**V. COLONIAL EXPANSION AND TAXONOMY**

Chaucer's use of folk tales reflects the popular literacy that spread rapidly after Johannes Gutenberg invented movable type (1455), as printed tracts, manuals, and libraries of books fed a new hunger for learning. Literacy spawned the Renaissance (1500–1700), a period of European expansion into the continents of Africa, Asia, and the Americas. Global exploration refined mapping and navigation, while imperial conquest decimated native populations through warfare and disease. Discoveries of new lands and species also spurred growth in the descriptive sciences, especially geography, geology, and biology. A buoyant verbal fluency stirred the language arts, epitomized in the writings of Shakespeare, Cervantes, and Molière. Colonial empires used their literacy to record events and claim land as property. Native cultures often regarded words as totemic: their oral chants and stories were prayers, repeated to bless earthly cycles. Such radically different uses of language also expressed conflicting views of nature. Europeans sought to own the earth, even as their sciences and arts revealed the immense biodiversity of wild, untouched land.

With the rise of natural science after 1700, literature began to express growing respect for nature, especially those aspects of it that were rare and varied. Optical devices peered into the heavens and the human body, while theorists and engineers created new branches of knowledge, publicized and marketed by print, commerce, and coffeehouse (Jardine, 1999). New forms of technology changed the economic value of natural resources. Land remained a source of food and energy, but wealth increasingly came from products and services. Cities expanded by drawing rural tenants from the countryside; new social classes formed around labor, management, and capital. An expanding, literate middle class called for political reforms, replacing monarchy with representative democracy. Revolutions in the Americas pulled masses of immigrants to the new Western nations, where open territory and cheap land began to transform agrarian life into republican independence. (As the world has grown tamer in recent decades, wilderness has come to have a higher value, as land existing for itself, rather than for human enterprise.)

The study of natural diversity advanced through what Thomas Kuhn (1967) has called changing paradigms, theories or conventions that promoted experimentation. Early naturalists described organisms with confusing and inconsistent local vernacular until Carolus Linnaeus compiled *Systema Naturae* (1734), using Latin nomenclature to assign names by genus and species. Linnaeus was dead and thus fixed, and Linnaeus used social distinctions (kingdom, phylum, class, order, family) to create taxonomy, a systematic language that described the bounty and variety of nature. While literary authors from Jonathan Swift to Fenimore Cooper mocked scientists for obsessive categorizing, taxonomy became a consistent model for geology, archaeology, and philology. Those sciences declared that prehistoric changes in the earth and in language follow traceable lines of descent across long stretches of time. Such discoveries replaced mythic accounts of creation with rational studies of the unity and diversity of organic events.

**VI. DARWINISM AND LITERATURE**

The new sciences spawned a second age of exploration, 1750 to 1850, as voyagers from James Cook to Alexander von Humboldt surveyed the continents of Australia and the Americas. In their wake followed a generation of literary naturalists like Gilbert White, William Bartram, and John James Audubon, who drew field sketches and wrote evocative accounts of seasonal journeys through rural or remote lands. The expeditionary narratives of Peter Kalm or Meriwether Lewis and George Rogers Clark are read today as literature and history, for they verify the immense variety of species that led to theories of biodiversity. This tradition of gentleman amateur molded Charles Darwin, who spent a 5-year expedition on the *HMS Beagle*, examining patterns in geology and zoology along the coasts of South America and on Pa-
cific islands. Darwin was one of the last major scientists to be an independent field naturalist, living on his own income. After 1870 science moved to the laboratory bench and university departments, a shift in professionalism that also affected writers, who became teachers and seekers of literary prizes (Jardine, 1996).

In his major works, *Voyage of the HMS Beagle* (1839), *The Origin of Species* (1859), and *The Descent of Man* (1871), Darwin argued that species evolve through the effects of heredity, variation, and natural selection. He had several predecessors in science, from Georges-Louis Leclerc Buffon, who recognized adaptation, to Alfred Russell Wallace, who anticipated and clarified several of Darwin’s ideas about selection. The term “evolution” is quite ancient, going back to Latin evolvere, to unroll, and may describe inscriptive scrolls, the writings that physically expressed a gradual process of change. The Biblical account of creation (Genesis 1) frames a sequence of originating events, each a division (night from day, land from water, plant from animal) that measures the binary logic of doubling, branching, and splitting, all recurrent patterns in nature. Poetry, tragedy, and biography all depict similar sequences of choice and accident, selection and reproduction. Those lines of cause and effect link the Bible to Homer and Sophocles, and then to Shakespeare or Milton. A constant theme in Western literature is the relation of creator, man, and nature, as chains of interactive events lead to human survival or extinction.

Literary traditions emerged from worship and theology, while Darwin stated that evolution proceeds according to physical laws, without divine intervention, and that human beings did not arise separately from other species. Early confirmation of these views appeared in the research of the Austrian monk Gregor Mendel, whose experiments in hybridization found that reproduction is a process of encoding and transmitting genetic material. Despite the strong controversy these views aroused among people of faith, by the 1860s evolution and genetic science began to alter popular ideas of nature and produce the first authors who consciously recognized biodiversity.

Two leaders in this movement were Henry David Thoreau (1817–1862) and George Perkins Marsh (1801–1882). Their books (*Walden, Cape Cod, The Maine Woods, Man and Nature*) often examine the physical environment of New England, a region of low economic value but rich biodiversity. A self-trained land surveyor, Thoreau spent several years studying landforms and seasons. After reading *The Origin of Species* in 1860, he wrote the first modern accounts of forest succession and the dispersal of seeds (Howarth, 1982).

Marsh, a lawyer and diplomat, published early studies of erosion, blaming land clearance and overgrazing, and he argued vigorously that human beings must learn to control their destructive impact on nature.

Among later writers known as social Darwinists (chiefly Herbert Spencer, Walter Bagehot, and William Graham Sumner), application of evolutionary theory to human behavior produced racist theories of cultural history, favoring white and Nordic peoples over Mediterranean and African. This bias, based on fears of immigration by darker races toward the wealthier nations, slants the novels of Jack London, Frank Norris, and Theodore Dreiser. Later these voices were replaced by modernist writers like T. S. Eliot, Willa Cather, and Ernest Hemingway, who saw racial diversity as healthy and progressive, a human check on industrial monopolies that had lain waste to natural resources and open spaces. Cather explored these issues on the plains of Nebraska, while Hemingway examined waning biodiversity in several lands, from upper Michigan to the savanna of east Africa, where great forests and wild herds are steadily destroyed by human incursion.

**VII. ECOLOGY AND INDUSTRIALISM**

The concepts of ecology were slow to gain acceptance among biologists and thus entered literary culture at a later date. Ernst Haeckel, a German biologist and advocate of Darwin, coined the term “ecology” in 1869 to describe how organisms form alliances that shape their number and distribution. Based in part on Haeckel’s political socialism, his views also echoed rising concern that industrial growth in Germany, France, and England had begun to damage both physical environments and public health (Bowler, 1994).

A vivid picture of decline lay in the western United States, where rapid agrarian settlement after the Civil War destroyed native tribes, bison, and grasslands in a few decades. The conservation movement of 1870 to 1920 sought to stem those losses by creating public lands and parks for protection of natural resources. Yet in the popular arts, writers of dime novels and outlaw ballads glorified “winning the West,” a form of historic triumphalism sanctioned by the theories of Frederick Jackson Turner (1893), who held that the frontier shaped fundamental American values of self-reliance and innovation. Later scholars note that the settlement process was destructive, as pioneers destroyed the very Eden they had sought (Mitchell, 1981).

With the experimental writings of James Joyce, Gertrude Stein, and William Faulkner, a modern critique
of mass-industrial civilization began to emerge. Drawing on their perspectives as cultural outcasts (Irish Catholic, Jewish Lesbian, Southern Gentry), they mounted strong attacks on the spiritual vacuity of urban life and, in Faulkner’s case, on the economic abuses that had sapped agrarian America. In his 14-novel saga about a fictional county in northern Mississippi, Faulkner assailed profiteering, an ideology that created slavery, miscegenation, one-crop farming, and soil erosion. He was the first modern novelist to comprehend ecological concepts, which gained attention in the 1930s from New Deal efforts to counter the wide-scale failure of capitalism.

The word “ecology” connotes a desire to coexist, if not in symbiosis, at least in a mix of competition and collaboration that marks the dynamics of a healthy, vigorous ecosystem. Darwin’s image of nature was “an entangled bank,” and throughout the 20th century, ecologists created an interdisciplinary science that depicted nature as a web of interacting, interdependent forces that sustains itself—not as an unchanging constant, but as a vigorous dynamism sustained by frequent, patchy disturbances (Hagen, 1992).

Three writers who affirmed these principles, in sharply varying styles, were all Californians: poet Robinson Jeffers, “hard-boiled” novelist Raymond Chandler, and John Steinbeck, a social novelist who was trained in marine biology and in the investigative methods of documentary journalism. In his literary works, especially The Grapes of Wrath (1939) and The Sea of Cortez (1941), Steinbeck built a vision of interlacing natural and cultural forces, locked in struggles that were shaped by biological conditions of production, consumption, and adaptation (Beegle, 1997).

VIII. LAND ETHICS AND ENVIRONMENTALISM

After World War II, two writers emerged who were trained ecologists and willing to apply their influence to public policy. In their writings, Aldo Leopold and Rachel Carson issued firm warnings of environmental destruction by human means. Leopold’s The Sand County Almanac (1949) celebrates a glaciated wetlands in Wisconsin, poor in economic value but rich in its diversity of habitats. His farm lies at the intersection of prairie, forest, and marsh, forming a mosaic of ecosystems that measure the “intricate tangle” of nature. Out of this vision, Leopold sketches the story of how he abandoned early conservationist ideas (a wildlife ranger, he was paid to kill wolves) for a preservationist vision he calls “the land ethic,” the sense that land deserves respectful protection as a vital element in the biotic community.

In Silent Spring (1962), biologist Rachel Carson intensified that sense of respect by analyzing the destructive force of chemical pesticides. In an incisive passage, she demonstrates how toxicity spreads through groundwater, a system of transport so invisible that it is easy to ignore. For her efforts, Carson was attacked by the petrochemical industry as hysterical and unscientific. But her logic and eloquence impressed the Kennedy administration, brought a ban on DDT and other pesticides, and eventually helped create the Environmental Protection Agency (Howarth, 1999).

Both Leopold and Carson saw that human values are formed by ideas of land, seeing that property is also a shared earth. Their books brought attention to emerging “green” political movements of the late 20th century, from which two different land ethics developed: landscape, shaped by people and reflected by aesthetic beauty, and land use, focused on food and energy extraction and marked by utilitarian security. Although often in conflict, the two ethics also reflect agreement that human identity rests on a sense of place, while alienation is feeling displaced, homeless, or unlanded (Spinn, 1999).

IX. NATURAL RIGHTS AND ANIMAL FABLES

The perception that land deserves ethical regard soon led to extending rights to nonhuman species, both plant and animal, granting them value and respect while seeking to conserve their populations. Herman Melville’s epic novel Moby-Dick (1851) anticipates this generosity, expressed in an iconoclastic rhetoric that denounces whaling. The cause of animal rights began with 19th-century reformers, who sought protection for religious dissenters, children, women, Indians, slaves, and other powerless groups. These figures came to be seen not as “lower” orders, but creatures having what Darwin called “the mutual affinities of organic beings.” That phrase, the title of his penultimate chapter in Origin of Species, delivers his belief that humans are descended from all the creatures that have existed and still share the earth.

This theme enters literature in many animal tales written for young readers, from Anna Sewell’s Black Beauty (1877) to E. B. White’s Stuart Little (1945) and
Charlotte's Web (1952). They descend from ancient beast and fairy tales, in which animals have the power to create language, before passing it on to humans. Such stories describe how things came into being, and they still have great attraction for writers drawn to environmental themes (Flynn, 1999). A contemporary example is David Quinns's novel Ishmael (1991), a series of dialogues between man and ape, in which the ape explains history as a conflict between Takers and Leavers, identified by their divergent approaches to the natural world. Takers dominate resources and evolve into modern urbanized man; Leavers are indigenous tribes who lead subsistence lives and eventually vanish.

While such accounts are fantastic, they also urge readers to think of animals, water, soil, rocks, and food as more than commodities. The cultural information provided in literary texts clarifies that natural objects result from processes, including the moral and ethical choices made by humans. As such choices recur in history, the value of a creature or a place changes: once the sea was thought to contain monsters, later it became a therapeutic place to bathe and sun; now it is emptying of fish and clogged with sewage (Hendrickson, 1984). Formerly the forest world was seen as profane and lawless, the home of pagans; now it is sacred, a haven for rest and retreat (Harrison, 1992).

X. ECOCRITICISM AND BIODIVERSITY

Scientists are now warning that a global environmental crisis will be the future outcome of several current trends: human overpopulation, the rise of carbon dioxide emissions, atmospheric warming, and the loss of biodiversity. These dangers transpire across vast frames of time and space that are difficult for nonscientists to imagine. A crisis that occurs in a locality, such as flood or earthquake, is understandable, but one that works over a longer period and through evolutionary selection is invisible to most eyes.

Because its domain is virtual or imaginative, literature helps its readers to envision what they will never directly experience. Yet many variables also affect the ability of literature to depict the threat to biodiversity. A writer's personal involvement in a story will shape its tone, bringing nostalgia (Wallace Stegner) or urgency (Bill McKibben) to the subject matter. The work's form may illustrate its content, as when Peter Matthiessen shapes The Snow Leopard (1973) into a mountaineering ascent, using elevation to provoke meditations on higher metaphysical themes. An object described will take on the character of its subject, as when Annie Dillard expresses her notions of wildness through a weasel's alleged indifference to pain.

The work may also have a distinct cultural audience (John McPhee) or veer between standards of elite and popular reception (Edward Abbey). Some writers are sensitive to social and economic conditions (Edward Hoagland), others contrast regional and national concerns (Barry Lopez), still others attend to differences of race and gender (Sue Hubbell), or to the impact of other arts on literature (Anne Matthews). Environmental writing is often a mixed collection of genres; part scholarly research, part sermon-editorial, part whimsy and adventure. Similar manifestations affect the contemporary novels of Thomas Pynchon (Gravity's Rainbow, Vineland, Mason and Dixon) and Don DeLillo (White Noise).

Because of the great variation in environmental writing, a new mode of reading called "ecocriticism" has emerged to establish interpretive standards. While not agreeing in all respects, ecocritics often seek to translate the specialized language of science, with its jargon and acronyms, into a common vernacular. Ecocritics are interpreters and bridge builders, looking for ways to connect the disciplines that university departments have long separated (Glotfelty and Fromm, 1995). A few scientists, such as Lewis Thomas (1974) and E. O. Wilson (1998), have written popular accounts of their environmental concerns, but mostly the public has turned to literary writers to gain insight into the causes and effects of waning biodiversity. At times these voices cannot agree about solutions, for the global biosphere is too intricate, too complex, and too unpredictable to describe readily, and some scientists believe that it will not be managed until it returns to pre-industrial levels of stable composition (Peters, 1994).

Other ecologists have grave concerns about genetic erosion, through both species loss and also genetic engineering, which in the long term may produce fewer varieties of plants and animals. (Karaim, 1999). The role of species biodiversity suggests that its purpose may be to provide redundancy, which has a stabilizing effect on whole communities and ecosystems. In the face of drastic environmental change, diversity may buffer ecosystems against the collapse of ecological function. (Nudds, 1999).

If biodiversity works because variety protects a community against stress and disaster, then literary accounts of biodiversity must remain equally diverse and inventive for the sake of textual survival. A review of literary history suggests that the continuance of human life on earth will depend in part on what storytellers make of a changing world. The gift of language brings
insight, but also the wisdom to heed the silence that Annie Dillard calls "nature's one remark." The central question for human beings is not whether they will survive as a species, but whether the survivors will inherit a world worth sharing (Wilson, 1998).

See Also the Following Articles

AESTHETIC FACTORS • DARWIN, CHARLES • HISTORICAL AWARENESS OF BIODIVERSITY • RELIGIOUS TRADITIONS AND BIODIVERSITY • SOCIAL AND CULTURAL FACTORS • STEWARDSHIP, CONCEPT OF

Bibliography

LOGGED FORESTS

Reinmar Seidler* and Kamaljit S. Bawa*†
*University of Massachusetts, Boston and †Ashoka Trust for Research in Ecology and the Environment, India

I. Introduction
II. Management Systems
III. Structural Alterations
IV. Abiotic Changes
V. Biotic Changes
VI. Ecological Interactions
VII. Genetic Effects and Evolutionary Processes
VIII. Synergetic Anthropogenic Effects
IX. Concluding Remarks

GLOSSARY

alpha diversity The variety of organisms occurring in a particular place or habitat; often called local diversity.

beta diversity The variety of organisms within a region arising from turnover of species between habitats.

canopy Usually the highest tree layer in the vertical stratification of forests. The herbaceous foods for canopy-dwelling animals are found among the branches and leaves of the canopy.

dipterocarp A member of the Dipterocarpaceae family (comprising 22 genera) of South Asian and African timber trees.

ecosystem A community of interacting living organisms and its physical surroundings.

forest management Broadly, a pattern of human activities to derive economic or other utility from a forest. In its narrower sense—as a concept or guiding principle for the multiple and sustainable use of a forest—it is a complex ecological and sociological concept in which exceptional skill has to be exercised on the part of forest manager. Forests could be managed for the extraction of timber or non-timber forest products, forage for animals, watershed protection, or recreational use.

gap An opening in the vegetation created from disturbances such as clearing, logging, fires, diseases, storms or the natural death of a tree.

inbreeding Mating between closely related individuals more often than would be expected by chance.

logging The operation of harvesting trees, sawing them into appropriate lengths (bucking), and transporting them (skidding) to a sawmill. In modern day, this method is mostly mechanized.

mast fruiting A period or year in which a heavy crop of fruits/seeds is synchronously produced by trees and shrubs. This phenomenon, uniquely characterized by synchronicity, high variability, and periodicity of heavy fruit production, distinguishes it from non-masting plants.

natural regeneration The process of replacement by natural seedlings in gaps created by the selective cutting of marketable timber.

pioneer guild The first groups of species to colonize a newly formed or denuded habitat.

primary forest Forests that appear to be undisturbed by human influence.

secondary forest Successional forests growing in areas where in the past the forest cover had been completely removed.
shifting agriculture The practice of clearing a plot of land for cultivation for a short period of time, then abandoning it and allowing it to revert to its natural vegetation when the cultivation moves to another plot.

silviculture An applied science and branch of forestry concerned with the theory and practice of controlling forest establishment, composition, and growth.

succession Replacement of one kind of population/community by another in a habitat through a regular progression over time.

taungya The Burmese term for an agricultural system in which crops are interplanted with plantations of trees. As the trees grow and shade the areas, cultivation of crops is abandoned.

understory The broad spectrum of plants at the ground level of forests that often provide forage for grazing animals. These transitory plants are often relatively sparse in dense forests with closed canopies.

OVER MILLENNIA, HUMAN SOCIETIES have managed forests for the production of a range of goods and services. As long as population densities remained low, the impact of humans on forest ecosystems was minimal, except near population centers of ancient civilizations. With the start of the industrial revolution, the demand for forest products increased considerably. At the same time, rapidly expanding human populations started to exert tremendous pressure on forests. The demand for timber (and nontimber forest products) grew at an accelerated pace. Timber extraction on an industrial scale became the principal goal of forest resource managers in the latter half of the 19th century and remained so throughout the 20th century. Today, there is virtually no forest on the planet without at least some signs of human influence. With mounting losses of forest cover and increasing interest in the conservation of biodiversity, attention has now again shifted to managing forests for a wider array of goods and services. In this new paradigm shift, management plans may include logging as an element among others, or they may not include logging at all.

I. INTRODUCTION

Clearly, forests can be managed for a variety of goals. The impact of different management systems on biological diversity is still largely unclear, although there have been a number of studies of forest ecosystems managed for the production of timber. Here we provide a perspective on changes in biodiversity in forests managed largely for timber. The concepts and principles of large scale forest management for a diverse range of goods and services are still being developed, even though traditional societies have managed forests for such purposes for centuries. We first outline various systems of management and then describe the extent of our knowledge about their effects on forest biodiversity. The effects of these systems vary according to a number of factors, including the type of forest, the landscape matrix in which the forest is embedded, the nature and intensity of the management system, and the scale and method of analysis chosen by the observer. Forests are complex ecosystems with subtle and complex interrelationships, many of them unknown in detail and changing over time. The potential for variation in each of these factors is great, making generalization across specific cases tenuous.

Forest diversity may be examined at various levels, including ecosystem, landscape, population, and genetic levels. It is advantageous to keep these multiple levels in mind as one considers forest management systems. Since the changes effected by management on forest biodiversity do not happen all at once, but are mediated through a chain or sequence of influences both spatially and over time, our discussion is organized sequentially. We begin by describing the structural changes imposed on forests by management operations or activities. These structural changes determine the ensuing abiotic or environmental conditions within the forest at the small and medium (local and regional) scales. Structural changes and altered abiotic conditions combine to produce a variety of biotic changes, including changes in species composition and abundance, productivity, population density and distribution, and so forth. It is thought that these changes may also affect long-term ecological interactions and processes. Thus, the original structural alterations to forests result in changed environmental conditions, which in turn trigger changes at the species, population, and community levels. Large reductions in effective population sizes may result in genetic alterations and impoverishment. In addition, it is important to keep in mind that virtually every managed forest experiences a variety of indirect synergistic anthropogenic effects, namely human activities not part of the management itself but facilitated by it. These may include increased incidence of wildfire, hunting, shifting agriculture, and others. Taken together, these activities may multiply the overall impact of the management system.
II. MANAGEMENT SYSTEMS

Systems of forest management vary widely throughout the world in response to ecological, social-institutional, and political factors. Use of the phrase management system need not imply that management is in each case applied in a planned, thoughtful, or consistent manner, but only that it is systematic in the sense of being a pattern of human manipulative activities governed by a specific common goal. These activities may be carefully planned and controlled interventions or they may be haphazard patterns of one-time individual extractive activities, without much overall guidance. Thus, we use the word management rather broadly.

A. Timber Management

1. Clearing Systems

At one extreme of a continuum of management systems for the extraction of timber are systems of forest clearing, with or without replanting to stimulate regeneration. Clearing systems have been used most often in temperate forests. These forests have relatively low numbers of tree species and therefore a relatively high density of commercially useful species per hectare. However, some tropical forests, such as some of the dipterocarp-dominated forests of Southeast Asia, may also have sufficiently dense stands of commercial trees to make clearing an economic option. If there are residual noncommercial trees, they may be girdled or poisoned to produce even-aged stands in which commercial species predominate. There have been experiments with various geometries of clear-cuts, intended to aid in natural regeneration and the preservation of wildlife habitat. Examples are the checkerboard patterns in the Pacific Northwest of the United States and Canada and the strip shelter-belt system of Palcazu in Peru. Patterns of this type have as their goal the reduction in area of the individual gaps, and thus the reduction in distance between the edge of the remaining forest and the center of clear-cut areas. The underlying idea is to increase the likelihood of propagules from the forest finding their way into the clear-cut area and regenerating the stand. There have also been attempts to develop timber management systems that use natural gaps as models, but these are considered selective logging systems (discussed below).

Other forest management systems involving clearing are replacement systems, in which plantations of indigenous or exotic species are established after clearing. These have been widely used in temperate forests and less so in the tropics, partly because they often involve a larger initial investment in silviculture. The even-aged stands which result are seen as advantageous in terms of accounting and the efficiency of silvicultural operations, but they have been vulnerable to outbreaks of pests and disease and, as highly simplified ecosystems, have been criticized as deleterious to biodiversity.

2. Selective Logging Systems

In many species-diverse tropical forests, clear-cutting is neither an economically nor an ecologically interesting option, since only a few tree species are commercially accepted. Moreover conspecifics are widely scattered, and infrastructure including roads and rural mills may be sparse. In such managed forests, trees are selectively logged. Selective logging is a general term encompassing a wide array of management systems that vary widely with respect to spatial and temporal scale, harvesting intensity, planning, and oversight. Selectivity of species to be removed does not necessarily translate into selectivity of overall impact, so the residual forest may be affected indiscriminately even by the selective extraction of only a few trees per hectare.

Tropical forest management systems were adapted from the German forestry tradition, which was exported to India and Burma during the 19th century. Natural regeneration systems, or polycyclic systems, minimize silvicultural interventions by relying on natural regeneration after the harvest of a relatively low number of trees per hectare. These systems result in uneven-aged and multispecies stands, which are thought to provide the best opportunities for biodiversity conservation as part of the management plan. However, a variety of assumptions about regeneration patterns must be made, many of which have been questioned by some foresters, and conclusive evidence of long-term sustainability is lacking.

Natural regeneration systems with adaptations to regional conditions have been developed in Malaysia (selection management system), Ghana (modified selection system), Suriname (Celos silvicultural system), Trinidad (periodic block system), and Queensland, Australia (Queensland selective logging system). Each of these has made important strides toward solving technical silvicultural problems, but in many cases sociopolitical obstacles have been more severe.

B. NTFPs

Timber is not the only product to be taken out of forests and marketed, nor is it the only product for which forests are managed. Non-timber forest products
(NTFPs) include fruits, nuts, fungi, fibers, medicinal and ornamental plants, mosses, dyes, resins, gums, fuel-wood, charcoal, leaves as fodder, poles for local construction, honey, syrup, fish, and game, as well as other animal products. In some forests, these may constitute important and large-scale commercial resources; in others, they may have great local importance but fail to appear in commercial markets. Management systems for NTFPs, too, run the gamut from traditional, socially sanctioned systems to those that are legally organized and monitored. In some cases, extraction of NTFPs may affect forest biodiversity even more than timber extraction, since NTFP extraction is often done over a long period of time—over many human generations in parts of southern Asia, Southeast Asia, and Africa—and may constitute an intensive use of some species or areas. Most of the economically significant NTFP extraction systems are today found in tropical forests and include extraction of Brazil nuts, rattans, ornamental plants, and animals, fruits, and medicinal plants. Maple syrup is an economically significant and commercially developed NTFP of temperate forests.

C. Shifting Agriculture and Traditional Long-Term Intensive Forest Management Practices

As primarily agricultural systems, the effects of shifting agriculture on biodiversity are treated extensively in another chapter. Shifting agriculture is by no means the only traditional forest management practice, however. Some traditional practices have been very intensive, especially in high population density areas of Asia, Europe, and Central America, and some have been sustained over the very long term. In Japan and Europe, coppicing and pollarding systems, in which trees are repeatedly pruned back for their leaves or poles, have been practiced for centuries. These traditional systems have preserved habitat for a diversity of plants, birds, arthropods, fungi, reptiles, and amphibians that are absent from younger successional forests in the same areas. Traditions of pasturing domestic animals in commons woods, and the retaining of hedgerows, copses, and windbreaks within intensively managed agricultural landscapes, have all contributed to the survival of woodland diversity in Europe (Kirby and Watkins, 1998). In the tropics, traditional systems of taungya and other intensive agroforestry systems have for centuries combined high forest productivity with high species diversity. Taungya is a Burmese word, but similar traditions of agroforestry and the combination of tree crops with ground crops and kitchen gardens may be found in Brazil, Indonesia, Malaysia, and the Indian subcontinent. There are currently attempts to reintroduce and adapt some of these traditional techniques to modern conditions, since they are thought to hold promise both ecologically and economically.

Sacred groves found throughout the world that have been "managed" for centuries for spiritual, religious, and utilitarian reasons are beyond the scope of the present discussion.

D. Managed Forests as Distinct from Secondary Forests

The phrase secondary forest has been used in a variety of ways in the scientific literature of the recent decades. We reserve the phrase secondary forest for successional forests growing in areas where the forest cover has at some time in the past been completely removed, causing a break in the continuity of the vegetative cover over time, and where this break in continuity can be detected structurally or floristically. Forest historians, geographers, and archeologists have been pushing back the date for earliest detectable human impacts on forests throughout the world, and it is becoming clear that few if any forests are primeval in the literal sense of the word. Nevertheless, we call forests that show to obvious structural or floristic traces of human influence undisturbed primary forests.

From the perspective of forest biodiversity, the distinction between secondary and managed primary forests is important. Secondary forests consist of earlier successional stages; they are dependent on seed dispersal from outside for their regeneration and the continued process of succession. The implications of this for biodiversity will be discussed later.

III. STRUCTURAL ALTERATIONS

Forest structure is the three-dimensional arrangement of trees and other plants, in combination with nonliving spatial elements such as soils, slopes, and hydrology. In short, structure is the physical geography of the forest, considered at a range of spatial scales. Structure includes such characteristics as canopy and understory geometry, continuity or fragmentation of canopy cover, homogeneity or patchiness of species distribution through the landscape, soil structure, and the species composition and age structure of stands. Most of these elements may be considered at micro, local, or land-
scape scales. Forest structure influences forest biodiversity directly through the formation of microhabitats as well as the determination of larger-scale habitat characteristics, but the relationship between structure and species diversity is complex and not well understood.

The most immediate effects of selective timber extraction systems on forest structure, of course, are the removal of the individual target trees, together with the associated incidental damage. Incidental damage generally includes removal of surrounding vegetation to construct access roads and skid trails, damage to vegetation during felling and skidding, compaction, and scraping of soils. Harvest of as little as 3% of the trees in an area, as is not atypical in the diverse neotropical forests, may reduce canopy cover by half through incidental damage. Canopy reduction by 75% is not uncommon in Asian dipterocarp forests, where a total of nearly half the basal area of the forest may be removed. On the other hand, damage to basal area of as little as 4 or 5% has been documented in Amazonia, when less than a single mahogany tree per hectare was removed. And noncommercial harvesting projects have obtained even better harvest-to-damage ratios in Costa Rica, Ghana, Madagascar, and Queensland, Australia.

In conventional operations, the impact on undergrowth can also be substantial. In a logged area in neotropical French Guyana, nearly half of the undergrowth was destroyed during the removal of an average of only three trees per ha. Removal and damage of vegetation opens the canopy, creating gaps and artificial edges, and often lowers the average height of the canopy, thus altering the internal vertical habitat structure of the remaining forest and, in some cases, removing most or all individuals of the larger and older size classes. Little precise information exists on the sorts of structural damage ensuing from NTFP extraction. It is clear that damage from intensive harvesting of leaves for fodder, poles for construction, fruits, and firewood can all be substantial.

In the temperate zone forests, the structural changes in managed forests often result from changes in species composition and age structure of overstory trees, as the multispecies stands are replaced by even aged populations of commercially valuable native or exotic species. Such changes set in motion a number of changes in ecosystem structure (biodiversity) and function.

A. Vertical Structure

The vertical stratification of forest vegetation has been the subject of some debate among plant ecologists. It is clear that many forest organisms (birds, insects, and other arthropods, certain herps) partition habitat in the vertical dimension. It is less clear what effect timber management may have on this dimension of forest structure. Selective timber extraction may simplify structure, for instance, by eliminating or preferentially damaging one or several strata. On the other hand, it has been suggested that one reason for the increase in local abundance of certain species groups in secondary forests is the greater complexity of the understory structure in younger, less well developed forests. The enhanced presence of short-lived treellets, shrubs, and herbaceous plants in disturbed forests is a matter of record (Richards, 1996).

B. Canopy

The forest canopy, the highest layer of the vertical structure, includes most of the interface between leaf and light. As such, it is the area of greatest energy input into the forest ecosystem. In tropical or temperate evergreen rain forest, this interface may be virtually unbroken over large areas, and there is often a very distinct division, in terms of biodiversity, between the canopy layer and the understory. In seasonal, dry, or gallery-type forests, the canopy may be much more permeable to incoming light and correspondingly less distinct in its flora and fauna.

C. Gaps and Edges

The creation of forest gaps through disturbances such as selective logging both enhances and diminishes species diversity and richness in different ways. Gaps increase the local heterogeneity of habitats and create microhabitats that are rare in mature forest, but they simultaneously limit the extent of undisturbed forest, thereby reducing regional habitat heterogeneity. The question of which tendency will predominate is at the heart of discussions about management for biodiversity. The answer depends partly on the characteristics of the focal species or group, partly on the extractive methods employed, and partly on the spatial scale of analysis. Alpha diversity, which is a measurement of species diversity at the community level, does not necessarily covary with beta diversity, namely the overall turnover of species associations in contiguous habitat patches across a landscape. In other words, local diversity may not mirror or represent regional diversity. There have been a number of suggestions that the creation of structural gaps through management systems can or should mimic the dynamics of natural gap creation through treefall, with the idea that community dynamics would thus be
minimally affected. Other authors, however, have been less sanguine about the practicality of this.

Gaps and forest edges may often constitute habitats not unlike the forest canopy in respect of abiotic factors such as insolation, wind, temperature extremes, and rainfall, and may therefore exhibit an analogous flora and fauna.

D. Forest Fragmentation

The term *fragmentation* is often used to summarize the landscape-level structural changes to the forest exerted by a range of human activities. Forest fragmentation is a process that may take place over centuries, decades, or years. Often, this process starts with selective logging within a matrix of natural forest and may extend over time through the development of progressively more intensive agricultural landscapes (*agroecosystems*). These agroecosystems may eventually themselves become the matrix, finally leaving only isolated "islands" of residual forest. In developed countries, urbanization, suburbanization, "exurbanization," and the establishment of extensive transport infrastructure have all contributed to forest fragmentation, especially over the past 50 years. The process of forest fragmentation, and its effects on various aspects of biodiversity, has proven rich mine for ecological study. Several studies show that fragmented forests have less biodiversity than contiguous forests and that ecological and evolutionary processes that maintain biodiversity are compromised in forest fragments.

E. Spatial Mosaic of Forest Types

Both short- and long-term forest management actions have been shown to make radical changes in the spatial distribution of forest types and communities at the landscape scale. Natural forests are mosaics of species associations responsive in their distribution to aspect and degree of slope, altitude, exposure to winds and storms, fire history, and soil characteristics. A forest’s spatial pattern and landscape context are thought to influence the regeneration of logged areas through seed and pollen dispersal dynamics and recruitment rates. Management can alter the natural forest mosaic by superimposing on it separate patterns reflecting land-use history, regeneration dynamics and history, or simply the pattern of access for logging (Cannon et al., 1998; Foster, 1992). Altered distribution of forest types, even within a continuous forest cover, can have potentially harmful (or beneficial) effects on populations of some wildlife species.

F. Soil Effects

Effects of management systems on soils include alteration of the microstructure of large areas of forest soil through compaction and scraping, alteration of the forest floor profile through the creation of artificial pits and mounds (or the elimination of natural ones) by bulldozers, changes to the hydrology of the affected region, and losses of organic matter and nutrients that potentially threaten fertility of the site. The latter problem is perhaps most serious in dryland areas and some areas of the humid tropics, where soil structure is often less robust, lacking the highly developed humus layer and deep mineral soil profile of many forests in temperate climates. In some tropical and temperate rain forests, heavy rainfall tends to exacerbate soil problems through increased runoff from the exposed surfaces and compacted soils of roads and trails, the associated erosion of soil surfaces, and increased clogging and silting of lakes, streams, and wetlands. Old tractor tracks and landings may remain sources of direct runoff for decades, due to their low infiltration capacities. Amounts of sediment washing into streams have regularly been shown to be higher and more variable after logging disturbances. Increases in sediment off-flow by a factor of between 2 and 20 have been measured. These sediments often contain substantial amounts of organic material. Leaching of nutrients, as large volumes of rainwater fall directly onto exposed earth and percolate through the soil, is another problem, the severity of which depends both on the characteristics of the soil substrate and on the logging techniques employed. These soil changes may over time alter species composition and structure of forest vegetation at all levels.

The absence of downed logs and decomposing woody debris in temperate zone production forests has been blamed for reductions in soil fertility. A number of plant, animal, and fungi groups depend on complex forests floor structure for habitat and nutrient availability. All of these problems with soils have been said to be amenable to improvement through careful logging procedures. The U.S. state of New Hampshire and the Canadian province of British Columbia have each recently published suggested guidelines for maintenance of coarse woody debris in forestry operations.

IV. ABIOTIC CHANGES

A. Insolation, Temperature, and Wind Patterns

Direct structural changes may cause changes in a series of abiotic factors, again on several scales. Opening of the
V. BIOTIC CHANGES

Up till now we have been describing those measurable structural, abiotic effects of logging that have been documented in forests of various types worldwide. What effects do these changes have on the biodiversity, especially the species, which make up the forest?

Plants form the first, or autotrophic, layer of the forest ecosystem. The autotrophs are able to build organic compounds from inorganic building blocks, using energy from the sun. The other (heterotrophic) organisms in a forest are directly or indirectly dependent on the transformative power of photosynthetic plants for their livelihood. Therefore we will consider plants first. Fungi have been implicated in crucial symbiotic relationships with many forest plants and will be considered next. The most numerous primary consumers of plants in forests are the invertebrates. Primary consumers transform and recycle the organic compounds they derive from ingesting plants. This group will be discussed third. Some vertebrates, the group we review fourth, are also primary consumers (herbivores), others are secondary or tertiary consumers (carnivores and omnivores). Detritivores, which make up the final stage in the recycling of organic materials, are a very important component in forest ecosystems; we will discuss them in the context of effects on soils.

A. Plants

Since plants are stationary, they are most immediately influenced by changes in their abiotic surroundings. Such changes may influence plant species composition, density, evenness, community structure and associations, and plant interactions including mutualisms and parasitisms.

1. Early Successional Phases

A pioneer guild of plants is characterized by lack of shade tolerance and therefore tends to be favored in the light-rich microhabitats created by gaps. Pioneer species tend to be widespread, with airborne pollen and small airborne seeds, fast growth rates, and tolerance of a wide variety of environmental conditions, allowing them to disperse quickly into disturbed habitats. Hence, they are characteristic of managed forests, and the pioneer guild is relatively homogeneous in species composition over large regions (Finegan, 1996). This can be confusing to researchers, since localized increases of plant species diversity are common following logging. Logging expands the area of forest occupied by early successional phases, boosting representation of pioneer
species in small samples. These local increases in diversity are not, however, matched on the regional level, since the local increases themselves come from the regional pool. Thus, logged forests tend to become increasingly homogeneous on a larger scale.

2. Exotic Invasive Species
Loss of tree cover often permits the recruitment of naturalized exotics (species that are not natural to the area), since the special conditions under which native species are competitively superior no longer exist. The presence of exotic plants tends to homogenize habitats and hence biodiversity. Examples of the many widespread and aggressive exotic plant problems are Impatiens (an Asian ornamental) in the neotropics, and the reed Phragmites australis in the United States.

B. Fungi and Mycorrhizae
A number of studies have shown that the presence in the soil of inoculate of certain species of symbiotic fungi is necessary for the regeneration and successful growth of a number of tree species, and the same may be true for other rooted plants. Soil mycorrhizae invade growing root tips and make available to them, in the course of their own metabolic processes, nitrogen and other nutrients that are otherwise locked up in the soil. Mycorrhizal fungi require certain soil conditions to flourish. If soil environmental conditions change radically, as can occur as a result of clear-cutting, for instance, these fungi may go locally extinct. Without healthy populations of the appropriate fungal symbionts, many tree species may be unable to regenerate successfully.

C. Invertebrates
We still know the least about the effects of logging on the first tier of forest-inhabiting animals, the invertebrates. One of the central unknowns is the number of invertebrate species inhabiting tropical forests, which is not known even to the nearest order of magnitude. Nevertheless, data on some taxa are available, and it is possible to distinguish some trends. Canopy denizens, especially flying insects such as butterflies and moths as well as flies, wasps, and bees, may be more easily seen or trapped in lower, disturbed forest, and hence may show up more easily in surveys. Like nectar-feeding and frugivorous bird species, some invertebrates may also benefit from the increased flowering and fruiting among early successional trees and vines in logged tropical forests. And many groups are adapted to the microclimates of forest edges and gaps, and so may benefit on the population level from structural changes and forest fragmentation. Nevertheless, lepidopterans and some dipterans and bee species are clearly reduced on a regional scale in disturbed tropical forests, because there are groups with very specific adaptations to the environmental conditions of closed canopy forest. For example, the hives of the large Asian honeybee (Apis dorsata) are usually found on very large old trees. Certain groups of tropical flying insects depend on vertebrate dung for part or all of their nutritional requirements. These may be strongly affected by reductions in mammal and bird populations that may accompany disturbance, although some may equally benefit from the presence of horses, cattle, or other domestic animals in the landscape. Other tropical winged insects are well adapted to the conditions of closed forest and have coevolved with particular resources that may be absent or rare in disturbed forest. Examples are orchid bees and orchid flowers, parasitic wasps and fig fruits, or lepidopteran larvae. These groups may become rare when their hosts become rare.

In the temperate zone, lepidopterans, as well as other flying invertebrates, such as dragonflies, mayflies, and damselflies, are rarely found under closed forest canopy due both to their temperature requirements and to their dependence on bodies of water during part of their life cycle. Nevertheless, these too may be affected by uncontrolled logging activities that disturb their habitats within forests, such as bogs, swampy areas, ponds, and streams. Since boreal forests are relatively species poor, approaching the conditions of a monoculture, it is unclear to what extent cyclic events—such as insect infestations or fire—may be a part of their natural disturbance regime.

The invertebrate fauna of leaf litter and soil surface shows divergent responses to logging activities. Some groups (beetles, cockroaches, and millipedes) increase significantly in both species richness and abundance in logged tropical forests. Others (spiders, mites, scorpions, springtails, and termites) decrease. Still others, such as ants, seem to be reduced in species richness while remaining abundant in absolute numbers. Many detritivores (wood-boring beetles, certain termites, and soil mites as well as fungi) are habitat-dependent on standing or downed rotting wood. Such groups appear to have been negatively affected by long-term timber management in Europe and the American Southeast, where old growth forest structure including snags and large rotting logs has been effectively eliminated over large areas.

The effects of forest management on another important group of invertebrates, the parasites, are little un-
nderstood. It is known that parasites are some of the strongest regulators of agricultural pests, and there are data showing that tropical agriculture, chronically vulnerable to pests, may actually be more successful where patches of natural forest are left between fields. Parasites and predators on insect pests may breed in forests and disperse into nearby fields, holding down pest populations.

It has often been observed that a dramatic increase in mosquito populations is a common phenomenon in logged areas, where drainage patterns are often affected. In the tropics this may constitute an important human health concern, and malaria is increasing in some areas.

In general, the invertebrates are still the least well known of the macroscopic fauna, yet their aggregate biomass is very large and their ecological influence is difficult to overestimate. E. O. Wilson has written about the essential role played by the “little things” in maintaining a variety of ecological functions (Wilson, 1987). The reduction or absence of groups of these forest denizens may have as yet unforeseen long-term consequences for forest ecology and ecosystem functioning.

D. Vertebrates

John Terborgh has observed that, in addition to the critical role played by Wilson’s “little things” as the foundation of an ecological pyramid, there is an important top-down regulatory role played by the “big things,” namely vertebrates and especially the large predators (Terborgh, 1988). In sites where top predators are absent, their prey populations, most of which are herbivores, tend to proliferate. This dynamic has been documented in tropical sites such as at the Smithsonian Institution Tropical Research Institute station on Barro Colorado Island, Panama, which lacks jaguar, in temperate parks devoid of wolf and puma populations, and in expanding suburbia all across North America, where white-tailed deer and Canada geese are increasingly being viewed as pests by the human population. Many of the proliferating tropical herbivores, such as the large rodents agouti and capybara, are seed predators and may be affecting the regeneration of large-seeded tree species. In temperate parks, large deer and elk populations are browsing deciduous seedlings to the ground, effectively preventing their regeneration. The lack of top predators in managed forests may significantly alter forest structure and composition, both in the short and the long term.

Vertebrate populations in forests coming under management may be affected by losses or gains in (a) habitat area or (b) food resources. The direction of change is determined both by the particular needs of each species and by the nature of the management system. The many permutations of these variables explain why we see divergent responses to management among vertebrates.

1. Habitat Loss or Increase

The loss of important habitat elements tends to affect the more specialized and the less mobile species most strongly, since they may not be able to disperse to new habitats. Loss of the appropriate microclimate, sufficient kinds and amounts of cover, nesting sites, or even perch sites may force species to local extinction. We know that some vertebrates depend on habitat elements found only in old, undisturbed forests. A well-known example is the northern spotted owl in the Pacific Northwest of the United States, which nests in the large hollow trees that are usually eliminated by timber management. A range of large temperate and tropical birds, including large woodpeckers, hornbills, and quetzals, likewise nest exclusively in hollows and may use and reuse the same holes year after year. Some mammals, too, are critically dependent on old-growth habitat elements for survival. The American marten, for instance, which has very little body fat yet remains active throughout the winter in cold and snowy environments, needs resting places below the surface of the snow, often under rotting stumps or large logs. Thus, although animals like the omnivorous marten may benefit from the increased seasonal availability of berries or other resources in logged forest gaps, their territories must also include areas of older forests with woody debris.

It has been shown that a number of smaller forest birds, both temperate and tropical residents, avoid gaps in forest cover such as those created by logging. Some are reluctant to cross gaps of as little as a few tens of meters. This may result from fear of predation or from physiological adaptation to a narrow range of understory temperatures and humidity levels. Under the increasingly predominant conditions of forest fragmentation, these species may find it difficult to locate appropriate habitat, even when it is available nearby.

Terrestrial amphibians, including frogs and salamanders, may often find the relatively desiccating conditions of open-canopy disturbed forest too warm and dry. Because they "breathe" partly by the exchange of gases directly across the skin, amphibians are particularly vulnerable to alterations in their microclimates. One element in the present worldwide crisis among amphibian populations may be reductions in suitable cool, moist, and shaded old-growth habitat.

On the other hand, some vertebrate populations may do well in disturbed habitats in secondary succession.
Examples of this group that have been studied are mice in temperate forest fragments, tenrecs in Madagascar, and other rodents in tropical and temperate forests, all of which may benefit from the increase in undergrowth and herbaceous ground cover accompanying management.

2. Loss or Increase of Food Resources
Opening the forest canopy may often stimulate a flush of new leaves or fruits as the trees and understory respond to the more generous light allowance. Many browsing and fruit-eating vertebrates benefit from this change and may seek out areas of secondary succession such as logged areas. Populations of large folivorous mammals, including elephant, tapir, duiker, gaur, bearded pig and peccary in the tropics, or deer and elk in the temperate zone, often do well in secondary forest. In the absence of predation, including human hunting, they may themselves become agents of serious ecological change by grazing and trampling large areas of forest understory. On the other hand, many of these large browsing mammals are important game animals, especially in tropical forests where hunting is particularly difficult to control. Many have evolved shyness responses to humans, and most have greatly reduced populations in areas where hunters have access.

3. Birds
The effects of logging on the species richness of birds are highly variable. In logged tropical forest, certain feeding guilds experience a reduction in population ranging from 29% up to 100%. Hard-hit guilds include terrestrial foragers (insectivores and frugivores) and the small insectivorous species associated with understory habitats, particularly those of the forest interior. These decreases are likely the result of reductions in the insect resource base or microhabitat changes. These findings are not completely consistent among studies, however. Some studies, for instance, have found small increases in abundance for terrestrial, understory, and foliage-gleaning species after logging in Malaysia. Substantial local increases in abundance have also been noted for generalist feeders that can supplement nectar and fruit resources with insects. The opening of the canopy often, though not always, increases the availability of nectar and fruits in early successional patches, in turn supporting more abundant populations of generalist feeders.

4. Primates
Primates have particularly complex responses to logging and to the presence of humans in managed forests. In the Neotropics, generalist feeders (Pithecia spp., Saginus spp.) have been found to do well in disturbed areas, whereas frugivores (e.g., Chiroptera spp., Ateles spp.) tend to leave logged-over areas, especially when their food trees have been extracted. However, there are exceptions, such as Callithrix torquatus, a generalist feeder that was found to be absent from logged forest. Old World frugivores such as Pongo pygmaeus and Pan troglodytes decline sharply in population density after logging, whereas folivores (e.g., Colobus spp., Gorilla gorilla) may seek out disturbed forest for the leaf flush.

F. Behavioral Changes
Behavioral changes among vertebrate populations have received little attention until recently. There are substantial difficulties inherent in studying and quantifying them. Nevertheless, they should be considered a significant group of changes because their effects may be delayed until one or several generations after the disturbance itself. Examples, which have been noted, are changes in social organization and breeding behavior among primates in Malaysia, and in nesting behavior and dispersal among birds in fragmented habitats. Among animals with complex response patterns, there are doubtless many other instances yet to be studied of behavioral changes stemming from human interventions. These may affect population dynamics years or decades after an event such as logging.

F. Domestic Animals and Nonforest Species
Domestic animals and introduced species have been important influences on forests throughout history. In European oak and beech forests, for instance, pigs and other grazing animals were grazed in commons forests on a seasonal basis to take advantage of fruit masting. Sheep and goats have virtually eliminated forest regeneration on the once lushly tree-clad slopes of the Mediterranean and throughout the Middle East (Thirgood, 1981). Today, cattle browsing is a major influence on remaining forest fragments in many areas of the tropics. The movement of nonforest species into logged and disturbed forests is the origin of some of the apparent increases in species diversity in logged forest plots. More species of coleoptera, for instance, have been counted in logged tropical forests than in neighboring primary forest. The species lists in logged areas, how-
ever, included the names of field species that had entered the forest after the canopy had been degraded.

**G. Aquatic Fauna**

The loss of aquatic biodiversity is understudied in the tropics. Concerns regarding changes in hydrology resulting from forest disturbance overlap with local community concerns, since the need to protect sources of clean water is becoming acute in many places as human populations grow. There is more information about the effects of logging on streams and aquatic life in the temperate zone. The removal or degradation of tree cover on slopes, including the construction of logging roads, often results in topsoil loss and erosion. Soil is washed downhill and ends up in streams and rivers. Silt deposition causes declines among anadromous fishes, bottom-feeding river fishes, bivalves, and other invertebrates. Heavy silt loads can have major impacts far downstream—even in some cases degrading coastal habitats such as tropical coral reefs, mangroves, and offshore fisheries. Changes in the temperature of stream water as a result of streamside vegetation loss can also have far-reaching effects on faunal diversity.

**VI. ECOCLOGICAL INTERACTIONS**

Structural, abiotic, and biotic changes are likely to alter a range of ecological interactions and evolutionary processes. Changes in hydrology, soils, and soil fauna, for example, should influence nutrient dynamics, which may in turn have further impacts on biodiversity. Biotic changes alone can alter species interactions. Again little is known about the impact of extractive regimes on ecosystem function and ecological interactions. The impacts are likely to be most dramatic where extracted species constitute dominant elements of forest ecosystems as, for example, the dipterocarp tree species in Southeast Asian dipterocarp forests. Indeed, much of the recent evidence that logging may influence vital interactions and ecological processes comes from such forests.

Pollination, seed dispersal, and seed predation are basic processes that profoundly influence reproductive output and regeneration of plants. By lowering the density of harvested species, logging can decrease the resources available to pollinators, animal seed dispersers, and seed predators. The vast majority of tree species are outcrossed, and in tropical forests, where most tree species are pollinated by animals, substantial declines in the abundance of food plants can result in changes in pollinators' foraging behavior. One study in Southeast Asian dipterocarp forests confirmed that logging in forests containing Shorea siamensis increased the average distances among the remaining trees of that species. The large distances made it difficult for the bees that pollinate S. siamensis to move from tree to tree, concomitantly limiting outcrossing potential. Opening up the canopy also led to colonization of the forest floor by flowering plants that attracted bees to the understory, further depriving the canopy trees of their pollinators (Ghazoul, 1999).

Reduction in the density of reproductive individuals can also result in a substantial decline in the number of seedlings. A recent study compared seedling production in logged and unlogged dipterocarp forests of Indonesia (Curran et al., 1999). Logging reduced the number of reproductive individuals per hectare to 3% of the original. Seedling production in the logged forest, following a mast fruiting year, was a mere 15% of that in the nearby unlogged forests.

The same study also documented indirect effects of logging. Dipterocarps undergo mast fruiting every few years associated with local El Nino-Southern Oscillations (ENSO). Logging around the Gunung Palung National Park, the site of the study, has fragmented the once contiguous forest. Changes in land use have apparently affected local ENSO climate conditions. As a result, spatial synchrony in mast fruiting and the level of seed production have been reduced. One result of the natural mast fruiting pattern is that copious amounts of seed are produced simultaneously. It is assumed that this "satiates" seed predators, leaving a substantial surplus of intact seeds for regeneration. However, with the reduction in seed production, seed predators are not getting "satiated," and there is little seed surplus. Even in the national park, where there is no logging, there has been inadequate regeneration because of the decrease in seed production and the movement of seed predators into the park from surrounding areas lacking mature reproductive trees.

These two studies from Southeast Asian dipterocarp forests illustrate several of the ways in which logging can have far-reaching and perhaps unexpected effects on forest dynamics. (a) Removal of mature dipterocarps and the opening of the canopy (both structural changes) produce new microclimatic conditions and an altered array of understory species (a biotic change). (b) These changes combine to produce altered foraging patterns among bees, compromising their efficiency as pollinators—an essential ecological interaction. (c) Mean-
while, fragmentation and degradation of the forest cover alters local climatic patterns (an abiotic change), which again alter an ecological process, the mast-fruiting pattern. (d) This alters seed predation dynamics and affects regeneration, potentially influencing forest structure and composition into the long term.

Considerably more work is needed to fully understand the consequences of the altered ecological relationships brought about by logging and other extractive activities. Studies on the effects of logging have largely been restricted to the impacts on structural and biotic components. Removal of nontimber forest products may also be expected to have an impact on a diverse range of ecological interactions and processes but has received even less systematic study.

VII. GENETIC EFFECTS AND EVOLUTIONARY PROCESSES

Alterations in density of tree species may also have a profound effect on evolutionary processes. Most trees are strongly outcrossed; outcrossing rates are highly dependent on population density. Low population density can restrict the movement of pollen among trees and result in a high level of inbreeding. Although it has been difficult to document the deleterious consequences of a decrease in outcrossing rate, it has been confirmed in several species that inbreeding does increase with a decrease in density.

In the tropics where many tree species typically occur in low densities, reduction in density due to extraction may further decrease population sizes. Small population sizes may also contribute to inbreeding. Thus, inbreeding may be increased both by the effects of decreases in density on the mating system and by decrease in overall population size.

Extraction can also remove the better-adapted genotypes from the population, leading to dysgenic selection, or selection for less well-adapted genotypes. Populations subject to harvest have lower levels of heterozygosity and lower overall levels of genetic variation. Inbreeding in logged populations resulting from a decrease in density may further decrease overall levels of heterozygosity.

Genetic and evolutionary effects can be cumulative. Fragmentation, decrease in density, and reduction in effective population size all contribute to inbreeding. In trees that are highly outcrossed, increasing levels of inbreeding may be particularly deleterious in the face of a changing environment. The genetic consequences of management interventions in forest ecosystems remain poorly explored.

VIII. SYNERGISTIC ANTHROPOGENIC EFFECTS

Probably only a minority of extractive forest operations today are performed under a management plan, and the vast majority of these are in the developed, less biodiversity-rich countries. In the tropics, in contrast, only about 2% of forests are currently under any active management plan, according to one authoritative study (Poore et al., 1989), although most are experiencing extractive pressure. The absence of planning and long-term control most often results in a series of secondary effects, including increased hunting, fire, invasion by exotic species, mining activities, shifting agriculture, and illegal logging, which may far outweigh the effects on biodiversity of the original management operations.

It has been noted that large amounts of secondary vegetation and brush tend to alter the understory structure of cut-over areas. The role of such vegetation and brush in fueling forest fires, both in temperate forests and even in the relatively fire-resistant moist tropical forests, has received increasing attention in recent years (e.g., Nepstad et al., 1999). It is clear that large-scale fires are altering microclimatic conditions over wide areas of tropical Asia and South America. In some of the drier temperate forests, for instance parts of the U.S. West and the pine barrens of the Atlantic seaboard, vegetation communities have evolved under the influence of regular low-intensity fires. The suppression of such fires during this century has altered the makeup of these communities and in some cases lowered their resistance to large destructive conflagrations. Reintroduction of controlled fire regimes is beginning to allow the native species to return and compete successfully with exotic species. A somewhat contrasting case is found in certain areas of the dry tropics, which are, as a group, some of the most highly altered and degraded forest ecosystem types. Here, degradation of the native plant assemblages and hydrology have virtually eliminated the forest's ability to regenerate itself in the face of continual incursions of fire. In such areas, putting a stop to forest fires over at least the next several decades is seen as a key step to restoring the natural plant communities.

Sharp increases in hunting after logging, too, have become the object of attention recently (Oates 1999; Robinson et al., 1999; Struhsaker, 1997). Logging re-
quires roads. Particularly in the tropics, where hunting is often unregulated, logging roads provide easy access for hunters and their vehicles into previously remote areas. Logging crews themselves are often the first hunters in a newly accessed area, both for their own consumption and for commercial markets. Others quickly follow them, and pressure on the game animal species can become intense.

In many parts of the tropics, one of the most intractable effects of logging is settlement by colonists eager for new agricultural land. Colonists follow the logging crews, entering the forest along logging roads, carving out homesteads, and planting crops. Since such movements tend to fragment and finally convert forest to agricultural land, they are clearly some of the most deleterious influences on biodiversity. Further extractive activities associated with wresting a living from the forest, such as mining, fishing, and trapping animals and birds, tend to accompany colonization and increase pressure on forest resources in a multitude of ways.

IX. CONCLUDING REMARKS

Clearly the extraction of forest products in managed ecosystems has an impact on biodiversity. However, natural ecosystems are dynamic and subject to all sorts of perturbations. The critical questions in the case of managed ecosystems are the extent to which biodiversity is lost and the degree to which the losses are irreversible. A landscape perspective is important for addressing these questions, because biodiversity may increase locally while decreasing at the regional level. Moreover, some local changes may be a manifestation of regional level changes and vice versa. The landscape perspective is also important to unraveling the confounding but synergistic effects of habitat fragmentation and climate change on biodiversity, as is evident from the recent work of Curran and colleagues in Southeast Asia (see also Bawa and Dayanandan, 1998). A wider outlook is also critical to understanding the impacts on the ecological and evolutionary processes and the ecosystem services that have been so far ignored in studies concerned with biodiversity in managed ecosystems. Finally, we do not imply that biodiversity in managed forests will always substantially decline. The low-impact harvesting regimes that are being explored all over the world may allow significant and critical goods and services to be conserved and sustainably used, particularly if such management systems are well integrated with their social and culture milieus and are responsive to the social needs of specific regions.

Acknowledgments

This represents publication number 75 of a program in biodiversity conservation supported in part by the MacArthur Foundation, involving the Ashoka Trust for Research in Ecology and the Environment (ATREE), Bangalore; the University of Agricultural Science, Bangalore; Pondicherry University, Pondicherry; the Institute of Social and Economic Change, Bangalore; and the University Massachusetts at Boston.

See Also the Following Articles

AGRICULTURE, SUSTAINABLE • DEFORESTATION AND LAND CLEARING • FIRES, ECOLOGICAL EFFECTS OF • FOREST CANOPIES • ANIMAL DIVERSITY • FOREST ECODY • PREDATORS, ECOLOGICAL ROLE OF • TIMBER INDUSTRY

Bibliography


LOSS OF BIODIVERSITY, OVERVIEW

Robert Barbault
Institut Féderatif d’Ecologie Fondamentale et Appliquée
Université Pierre et Marie Curie

GLOSSARY

biodiversity The variety of organisms considered at all levels, from genetic variants within the same species to the whole range of species and ecosystems.

extinction The disappearance of any lineage of organisms, from populations to species and higher taxonomic categories (genera, families, phyla). It can be local or global (total).

genetic drift The process of random sampling of genes that leads to changes in the genetic composition of a population. The effect of this process is particularly important at small population sizes.

homozygous Possessing the same gene form (allele) on both chromosomes.

locus The position on a chromosome of a gene. It is determined by any number of allelic forms.

metapopulation The set of populations (or subpopulations) of the same species, linked by migration. It is characterized by the processes of local extinction and recolonization.

Pleistocene The geological time that ends with the last glacial period and the appearance of humans. It started 2 million years ago and finished 10,000 years ago with the end of the last Ice Age.

stochastic The result of chance and random processes.

THE EXTINCTION OF A SPECIES, like the death of an individual, is a natural phenomenon—its inevitable destiny. In fact, during the long history of life, earth has experienced several periods of mass extinction. But the crisis of current extinction differs from the preceding ones in that it is the direct result of human activities. By our ecological success—haven’t we invaded the entire planet?—amplified by our industrial and technological revolutions, our species exerts such an impact on the biosphere that one witnesses today an acceleration of extinction phenomena without precedent. This erosion of biodiversity is expressed on three interdependent levels; it affects (a) the diversity of our planet’s ecosystems and landscapes, (b) the richness of species in the faunas and floras in most parts of the world, and (c) the genetic diversity of many natural and domesticated species. The objective of this chapter is to address the patterns, causes, and extent of such losses of biodiversity.

I. LESSONS FROM THE PAST

The longest phase in the evolution of life on our planet extends for the 2 billion years from the appearance of...
the simplest molecules capable of autoreplication to the appearance of the first prokaryotes. Unfortunately, nothing is known about the evolution of biodiversity throughout this very long period, dominated by microorganisms. Our attention concerning extinction must therefore be restricted to two kingdoms, the plants and the animals or, in other words, to the past 600 million years.

A. Extinction Rates Deduced from the Fossil Record

Table I shows the estimates for the average life span of species in various groups of fossils from their origin to their disappearance. Overall, the average life span of the species is about 5 to 10 million years. Based on this estimate, and remembering that the total interval of time considered is equal to 600 million years, one would estimate that the present stock of animals and plants accounts for only approximately 1 to 2% of all species that have ever lived. However, this evaluation obviously depends on the quality of the fossil record. One should also note that the estimated life spans vary considerably according to the groups considered (Table I). In particular, according to Raup, the average duration of mammal species is on the order of 1 or 2 million years, much lower than the 10 million years estimated for (mainly marine) invertebrates. Taking account of this variability, Sepkoski (1992) estimated that the current species of plants and animals account for approximately 2 to 4% of those that have ever existed.

<table>
<thead>
<tr>
<th>Group</th>
<th>Species' average life span in millions of years</th>
</tr>
</thead>
<tbody>
<tr>
<td>All invertebrates</td>
<td>11</td>
</tr>
<tr>
<td>Marine invertebrates</td>
<td>5–10</td>
</tr>
<tr>
<td>Marine animals</td>
<td>4–5</td>
</tr>
<tr>
<td>Mammals</td>
<td>1–2</td>
</tr>
<tr>
<td>Dinoflagellates</td>
<td>12</td>
</tr>
<tr>
<td>Foraminifera</td>
<td>10</td>
</tr>
<tr>
<td>Brachiopods</td>
<td>6</td>
</tr>
</tbody>
</table>

* Based on data from various sources (from R. May et al., in Lawton and May, 1995, simplified).

Rates of extinction tend to vary considerably over time (Fig. 1). Thus, four of the five major extinction periods in the fossil record (Fig. 1) each eliminated between 65 and 85% of the marine animal species (Or dovician, Devonian, Trias, and Cretaceous), and one (Permain) eliminated 95% or more.

B. Mass Extinctions Since the End of the Pleistocene

During the last 50,000 years, several extinction episodes have concerned almost exclusively the terrestrial mammals, some birds, and, with a lesser degree, some reptiles. A comparison of the data gathered on the various continents makes it possible to refine the explanatory assumptions (Fig. 2). Africa gained and lost several genera of mammals throughout the Pleistocene up to the Holocene, 10,000 years ago, but was left with a primarily intact fauna. In Europe, the extinctions touched only the largest mammals (>44 kg). Australia was subjected to a massive peak of extinctions, with 86% of the large-bodied animal genera, including the large reptiles and giant birds, going extinct at the end of Pleistocene. The majority of extinctions of the mega-fauna occurred about 30,000 years ago. Finally, in North America mass extinction occurred between 10,000 and 12,000 years ago. Of the 37 kinds of mammals that died out, 33 weighed more than 44 kg. Similarly, extinctions in South America affected only the genera of animals exceeding 44 kg.

Two classes of interpretations have been advanced to explain these extinctions, one focusing on human activities as an extinction force, the other focusing on climate changes. Thus, the episode of extinction of North America occurred at the end of Pleistocene, when temperatures rose, leading to changes in the composition and the structure of plant communities, thus disturbing the interactions between plant and their browsers, and consequently leading to the extinction of many herbivores (particularly the large ones) as well as the predators that depended on them. Similar arguments may be used in Australia, where the climatic change at the end of the Pleistocene was associated with a higher frequency of drought periods. However, the asynchrony of the extinction crises in America and Australia may lead one to favor the responsibility of humans, since the colonization of these continents by humans was also asynchronous.

The well-documented case of Madagascar is particularly interesting. Humans arrived there 1300 to 2000 years ago. At the beginning of the Christian era, 17 species of lemurs inhabited the island. Seven of them disappeared about 1000 years ago, all but one were
large species, probably diurnal and mostly terrestrial. The surviving species were, in contrast, either nocturnal or small, arboreal species. Three species of mammals from other families disappeared at about the same time: a pigmy hippopotamus, Hippopotamus lemeriei, and two large carnivores, Cryptoprocta spelea and Pleisiorcterus madagascariensis; furthermore, two species of the giant tortoises, Geochelone and the elephant bird Aepyornis (which is, with a weight of 500 kg, the largest known bird), went extinct. While it is not clear whether these extinctions are caused by the direct effects of hunting or by a degradation of the environment, it is certain, however, that they are linked to human activities.

II. CURRENT EXTINCTION RATES

With the increasing concerns about the erosion of the biodiversity and the developing polemic about the Convention on Biological Diversity since the Earth Summit of Rio in June 1992, it is becoming more necessary to have scientific information about the current rates of extinction. From the development carried out on the initiative of John Lawton and Robert May (1995), it is possible to draw some general conclusions.

A. Recent and Current Extinction Rates

The available data, summarized in Table II, show a number of points of general interest. Half of the extinctions listed since 1600 occurred during the 20th century; the majority affect species inhabiting islands. Extinctions among the insects are much rarer, relative to the number of known species, than among vertebrates (0.006 versus 0.3%). Although we have noted that the average life span of insect species, estimated from the fossil records, could be 10 times higher than that of...
The extinction of large vertebrates coincided with the arrival of humans in Australia, North America, Madagascar, and New Zealand. In Africa, where humans and animals evolved together for millions of years, the damage was less severe. From Wilson (1992).

tetrapods, the difference in extinction rate of two orders of magnitude is difficult to explain.

Of course, the differences in extinction rates among groups could result partly from the difference in interest granted to one or the other group. Thus, the number of taxonomists working on vertebrates is 100 times the number working on invertebrates, and 10 times higher for vertebrates than for the vascular plants. This differential attention appears in the rates to which new species are discovered: three to five new species of birds are discovered each year (i.e., 0.03 to 0.05% of the known total), while the tropical botanists can hope to find a new species of plant among every 100 specimens; collections of insects, mushrooms, or marine microrganisms carried out in unstudied areas reveal proportions of new species ranging from 20% to 80%.

A final point worth emphasizing is that, even at groups known comparatively well such as the birds and the mammals, the listed extinctions are certainly underestimated.

B. Estimate of the Future Rates of Extinction from the Area—Species Richness Relationship

Future rates of extinction can be estimated by merging the estimates on the losses of the environment (deforestation, for example) with the expected relationship between area and species richness. This relationship is an empirical rule, based on a number of studies, that connects the number of species $S$ for a given taxonomic group (coleopterans, birds, vascular plants, etc.) and the area $A$ of the “islands” (true islands, or ecologically isolated areas such as lakes, mountain peaks, clearings in a forest) inhabited by the species (Fig. 3). Generally, the relationship between species richness and the area has the form:

$$\log(S) = c \log(A)^z$$

where $c$ is a constant and the parameter $z$ is usually between 0.15 and 0.35. Assuming that this equation can also describe the reduction of species richness driven by deforestation or other processes decreasing the area of suitable habitat, and taking into account the current rate of deforestation in the tropics (which varies between 0.8 and 2% per year), one can deduce an annual rate of extinction of between 0.2 and 0.5%. In other words, with a total richness of currently 5 million species inhabiting the planet (our minimal estimate), between 10,000 and 25,000 species will disappear every year.

However, this kind of extrapolation is associated with several problems. It is not known, for example,
up to what point the effects of the fragmentation of the environment, particularly in the tropics, can be deduced from the area-species relationship that has been described mainly for insular biogeography. Thus, Simberloff stressed that, during the past 2 centuries, only three species of birds are known to have gone extinct in the forests in the eastern United States, even though the forest has been reduced to small fragments with a total of only 1 to 2% of their original area.

In spite of such uncertainties about our estimates, it is certain that the reduction and the fragmentation of the environment, even if not always resulting in a reduced density of populations, will increase the risk of their becoming extinct (Table III in Box 1).

### III. MECHANISMS OF EXTINCTION

What aspects of the environmental conditions make a species or population go extinct? It is clear that, if the mortality rate continuously exceeds the birth rate, a species will eventually become extinct. However, any

---

**TABLE II**

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of extinctions</th>
<th>Number of species threatened(^a)</th>
<th>Approximate number of recorded species (in thousands)</th>
<th>Percentage extinct</th>
<th>Percentage threatened</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molluscs</td>
<td>191</td>
<td>354</td>
<td>100</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>Crustaceans</td>
<td>4</td>
<td>126</td>
<td>40</td>
<td>0.01</td>
<td>3</td>
</tr>
<tr>
<td>Insects</td>
<td>61</td>
<td>853</td>
<td>1000</td>
<td>0.006</td>
<td>0.07</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>220</td>
<td>2212</td>
<td>47</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td>Fishes</td>
<td>29</td>
<td>432</td>
<td>24</td>
<td>0.1</td>
<td>2</td>
</tr>
<tr>
<td>Amphibians</td>
<td>2</td>
<td>59</td>
<td>3</td>
<td>0.1</td>
<td>2</td>
</tr>
<tr>
<td>Reptiles</td>
<td>23</td>
<td>167</td>
<td>6</td>
<td>0.4</td>
<td>3</td>
</tr>
<tr>
<td>Birds</td>
<td>116</td>
<td>1020</td>
<td>9.3</td>
<td>1.2</td>
<td>11</td>
</tr>
<tr>
<td>Mammals</td>
<td>59</td>
<td>305</td>
<td>4.3</td>
<td>1.3</td>
<td>11</td>
</tr>
<tr>
<td>Plants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gymnosperms</td>
<td>2</td>
<td>242</td>
<td>0.8</td>
<td>0.3</td>
<td>32</td>
</tr>
<tr>
<td>Dicotyledons</td>
<td>462</td>
<td>17,474</td>
<td>100</td>
<td>0.06</td>
<td>9</td>
</tr>
<tr>
<td>Monocotyledons</td>
<td>120</td>
<td>442</td>
<td>32</td>
<td>0.9</td>
<td>9</td>
</tr>
</tbody>
</table>

\(^a\) Reprinted with permission from Nature (Smith et al., 1993). Copyright 1999 Macmillan Magazines Ltd.

\(^b\) "Threatened" includes the IUCN (World Conservation Union) categories of "vulnerable," "endangered," and "probably extinct."

---

**TABLE III**

Number of Species Considered to be "Threatened" by the World Conservation Monitoring Centre (the WCMC considers all species classified as "endangered," "vulnerable," "rare," or "indeterminate" as threatened):  

<table>
<thead>
<tr>
<th>Group</th>
<th>Endangered</th>
<th>Vulnerable</th>
<th>Rare</th>
<th>Indeterminate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammals</td>
<td>177</td>
<td>199</td>
<td>89</td>
<td>68</td>
<td>433</td>
</tr>
<tr>
<td>Birds</td>
<td>108</td>
<td>241</td>
<td>237</td>
<td>176</td>
<td>882</td>
</tr>
<tr>
<td>Reptiles</td>
<td>47</td>
<td>86</td>
<td>79</td>
<td>43</td>
<td>237</td>
</tr>
<tr>
<td>Amphibians</td>
<td>32</td>
<td>32</td>
<td>35</td>
<td>14</td>
<td>133</td>
</tr>
<tr>
<td>Fishes</td>
<td>158</td>
<td>226</td>
<td>246</td>
<td>304</td>
<td>934</td>
</tr>
<tr>
<td>Invertebrates</td>
<td>382</td>
<td>702</td>
<td>422</td>
<td>941</td>
<td>2047</td>
</tr>
<tr>
<td>Plants</td>
<td>3032</td>
<td>3087</td>
<td>14,403</td>
<td>3302</td>
<td>26,307</td>
</tr>
</tbody>
</table>

---
Box 1
Future Extinction Rates Deduced from IUCN Red Lists

The International Union for the Conservation of Nature (IUCN) proposed a number of criteria to classify the level of threats imposed on plant and animal species. Seven categories are defined:

Extinct: A taxon that has not been observed in nature for more than 50 years.

Endangered: A taxon close to extinction and whose survival is at risk if no action is taken to eliminate the causes of its disappearance. Included in this category are species whose populations are reduced to a critical level or whose habitats are threatened.

Vulnerable: A taxon at risk of entering the ‘‘endangered’’ class if detrimental factors continue to exert their effects. This category includes those taxa whose populations decrease as a consequence of overexploitation, large-scale habitat destruction, or any other perturbation of their environment, and those taxa whose populations remain abundant but that are nonetheless threatened by a variety of detrimental factors.

Rare: Taxa whose populations are globally rare even though they are not endangered nor vulnerable. Such taxa are generally found in very specific areas or habitats or have widely dispersed small populations.

Indetermined: A taxon that has to be included in one of the above categories, but about which insufficient information is available to assess the level of risk.

Insufficiently known: A taxon thought to belong to one of the preceding categories, but without data to substantiate this.

Threatened: A taxon included in any of the preceding categories.

On this basis it appears that the number of animal species cited as threatened has grown by more than 30% between 1986 and 1990. During this interval, 15 vertebrate species (33 animal species in all) have been added to the list of recent extinctions; if it continues unabated, this would result in the extinction of half of the 47,000 vertebrate species in 700 years. Similarly, between 1990 and 1992, 163 plant species have been added to the list of extinct species; this corresponds with the extinction of half of the earth’s 250,000 plant species in approximately 3000 years.

Though all of these estimates remain fragile, the table of threatened species (Table III) constitutes a useful starting point for conservation strategies.

Prediction (which is based on perhaps a decade of observation) about extinction during the next, say, 3000 or 7000 years assumes that nothing will change. Since we are assuming that the major cause of extinction is humankind, this assumption appears unlikely.

Therefore, we will discuss more short-term risks of extinction, risks associated with an environment where, in principle, the population should persist (where births exceed deaths, at least at low population densities; see Box 2). In doing so, we will distinguish isolated populations constituting so-called metapopulations. The main difference between the two is that (unless due to human intervention) only interconnected populations can benefit from immigration.

A. Isolated Populations

1. Low Population Densities: The Allee Effect

At low densities, individuals may be distributed over distances that are much larger than the distances they usually move in, so that the probability of meeting one another for mating is small. Therefore, the growth rate of the population may drop below unity when the density is below a certain threshold. A well-established example of such an Allee effect leading to the extinction of the mottled woodpecker Dendrocopos medius in Sweden in 1982.

2. Demographic Stochasticity

Changes in population density due to the births and the deaths necessarily imply the role of chance, as birth and death, but also the sex of offspring and other parameters are random processes. It follows that there is a positive probability that all of the individuals in a popu-
**Box 2**

*Centaurea Corymbosa* from the Massif de La Clape: A Species on the Brink of Extinction

*Centaurea corymbosa* is a cliff-dwelling plant species, endemic to the Massif de la Clape in Southern France. It apparently cannot stand competition since it is only observed on cliffs or on rocks with very few other plant species; it cannot be found either in the pinewoods or vineyards downhill, or in the garrigue on the plateau. Although the massif is about 50 km, and that cliffs suitable for the species seem to occur all over it, only six natural populations are known, within a 3 km area, and they are separated by distances from 0.3 to 2.3 km. *C. corymbosa* is an outcrossing monocarpic perennial. Seeds germinate mostly during the autumn following summer dispersal (almost no seed bank); plants then remain at the vegetative rosette stage up to at least the fourth year following germination. Flowering takes place from April to August and plants are pollinated by small hymenoptera and diptera. About 500 individuals flower every year (from 5 to 250 per population), corresponding to about 6000 to 10,000 rosettes for the whole species (including seedlings).

Using both genetical and ecological approaches, French and Spanish researchers (Colas et al., 1997) carried out collaborative work to understand the factors limiting the colonization ability of *Centaurea corymbosa*.

Although the populations are at most separated by 2.3 km, population genetic structure suggests that gene flow among populations is highly restricted, as shown by the high level of differentiation: it is much larger than usually observed for other species with similar biology.

An ecological study has shown that seed dispersal distances are very short: a few tens of centimeters from the mother plant. Contrary to other species of the same family (Asteraceae), long-distance seed dispersal has never been observed, even when the wind reached 100 km/h. At maturation, seeds fall near the mother plant and the wind just pushes them on rocks before landing into a rock or a cleft. Pollen dispersal also seems to limit seed production since the fertilization rate of plants isolated by more than 4 m is about half that of plants with a close flowering neighbor. Possible colonizers of a new cliff, a few hundreds meters from the source populations, would thus have little chance to set seeds.

Successful experimental introductions on new cliffs nearby natural populations confirmed that they are suitable and were unoccupied because of the lack of colonization ability of *C. corymbosa*.

One might wonder why dispersal ability has not evolved in this species. It appears that landscape structure selects against larger dispersal and colonizing ability. Hence, if by chance some seeds are dispersed a long way to a suitable site, germinate, and grow into adult plants, they are likely to die without producing any offspring if they do not flower simultaneously.

The species has thus become trapped in its own deleterious system. One could argue that the species looks perfectly well adapted to the rocky habitat of the Massif de la Clape. However, it is clear that its survival entirely depends on that of the six extant populations that are very small. Moreover, there is some demographic evidence that some populations might go extinct in the near future because of demographic stochasticity, Allee effect, or inbreeding depression. To prevent this species from becoming extinct, more introductions into new sites will have to be performed.

---

3. Environmental Stochasticity

In natural populations, the probability of having offspring or of dying are influenced by the environment, and these environmental effects have some degree of correlation among individuals. If the correlation is low, a failure, say, of one individual in producing offspring, may be canceled out, at the level of the population, by a large number of offspring produced by another individual. If, however, the correlation is high, the stochasticity will reduce the variance of the demographic parameters among individuals and therefore increase the risk of extinction. In the extreme, during a total catastrophe the correlation is complete, and all of the individuals die at the same time. Thus, high population densities are by no means a guarantee of population survival.

More formally, environmental stochasticity is evi-
dent in the temporal fluctuations of the growth rates of the populations. The increase in the variance of population growth increases the risk of extinction, in particular when mean growth rate is low, population density is low, or the carrying capacity of the environment is low.

4. Genetic Factors

Two widely discussed genetic factors associated with extinction, both of which are important in small populations, are inbreeding depression and the loss of evolutionary adaptability to new environments.

If a population goes through the bottleneck of a low population size, individuals are constrained to mate with close relatives. This is often associated with reduced viability or fertility, a phenomenon called inbreeding depression. The mechanisms leading to inbreeding depression are unknown for the majority of the species. In Drosophila about half of the effect is due to the increased number of homozygous loci carrying recessive detrimental alleles (up to 5000), the other half is due to the accumulation of slightly deleterious and recessive mutations. Inbreeding depression has been established in many domestic species and zoo populations, but unfortunately, information is rare for natural populations.

5. The Critical Role of Population Size

The central message that emerges from most empirical and theoretical studies on extinction is that extinction risk increases, and thus that the life expectancy of the population decreases, as population size decreases (Fig. 4). Demographic stochasticity greatly increases the extinction risk if population size decreases below about 50 individuals; inbreeding depression is likely to become significant only in circumstances when a previously abundant population goes through a population bottleneck.

B. Metapopulations: Sets of Several Local Populations

Considering a metapopulation introduces an additional dimension to the risk of extinction, due to the possibility of dispersal from one local population to another.
Seeds, spores, or animals may disperse passively or actively from their natal population over the landscape. If they arrive at a favorable site that is not already occupied by members of their species, they may be able to found a new local population. In the long run, populations of a species may thus be established and disappear repeatedly at the local scale, while at the level of the landscape, one observes the dynamic equilibrium of a species moving among local habitats.

The persistence of a metapopulation depends on processes that affect rates of extinction of local populations, but also on processes that affect the establishment of new local populations. Two conditions must be fulfilled to get a new population established: favorable sites must exist, and a sufficient number of individuals must migrate from existing sites to the unoccupied habitats. That makes immigration and dispersal processes important mechanisms for the survival of metapopulations.

The survival of a metapopulation is threatened by two types of stochasticity, which are analogous to the demographic and environmental stochasticities involved in the dynamics of local populations: (a) the stochasticity of local colonization and extinction rates and (b) the regional stochasticity. The first involves the stochastic processes discussed earlier; the second, analogous to environmental stochasticity in isolated populations, applies if stochasticity has effects that are correlated among several local populations within the metapopulation. Since a significant source of environmental stochasticity is related to weather conditions, which are typically strongly correlated over a large spatial scale, one may generally expect a high degree of correlated regional stochasticity in metapopulations and thus high risks of extinction.

IV. CAUSES OF THE LOSS OF SPECIES AND GENES

We have argued as if there were no change in the environment experienced by the populations, but there is, of course, hardly any doubt that the current increase in the rates of extinction is primarily caused by drastic changes in the environment. This affects all of the key parameters associated with the risk of extinction: population sizes, the carrying capacity of the environment, the mean and variance of growth rates, the genetic structure of a population, the number and area of suitable habitats, and the number of local habitats forming a metapopulation. With this perspective, we will now consider three types of environmental changes that may initiate the extinction process: the destruction of natural habitats, overexploitation by humans, and changes concerning other species.

A. Destruction and Changes of Natural Habitats

All animal and plant populations are adapted to the local conditions of their environment; their persistence depends on the maintenance of these conditions. Large climatic changes of the past, the periods of glaciation, for example, led to local extinctions and shift of species' geographic ranges. These kinds of changes, however, were very slow, being spread out over periods of millions of years, so that most species were able to gradually adapt or alter their geographic range according to the novel conditions. Humankind has introduced a novel dimension to the changes of natural habitats: human activities are more similar to volcanic eruptions than to gradual climatic change and thus could amplify its effects. At the time of the last glaciation, for example, the extinctions of the large mammals during the Pleistocene are partly manmade.

The main effects of humans on the environment are the following:
1. Destruction of habitats (clearing forests, drying lakes and humid areas, etc.)
2. Degradation of natural habitats (pollution, dams, etc.)
3. Fragmentation of habitats

1. Destruction of Habitats

The most severe threat imposed on biological diversity is the loss of suitable habitat—and reducing habitat loss is the most important way of preserving the environment. The loss of habitat has been the main source of extinction, at least in the recent extinctions of vertebrates (Table IV). In many areas of the world, particularly on the islands and everywhere where the human density is high, a large portion of the natural habitats has already been destroyed (Box 3). Thus, in tropical countries, the destruction of the habitat ranges from 29% (Zambia) to 89% (Gambia) in Africa and from 41% (Malaysia) to 97% (Hong Kong) in Asia, and in most countries the number is above 70%.

Currently, the global rate of deforestation is between 1 and 2% per year, depending on the region. The main reason for such a rapid destruction of humid tropical forests is agriculture and other sources of human pressure, though the exploitation of forests for the trade of wood is also a significant cause.
2. Fragmentation of Habitats

The fragmentation of habitats is the process by which a large area consisting of a given type of environment is reduced in size and divided into two or more fragments. These fragments of the original environment are often isolated from one another by a highly modified and degraded landscape. For example, the large stands of vegetation that used to cover vast areas of the world have been more and more divided into separated fragments. Europe as well as densely populated areas of Asia have experienced this process for centuries.

Such isolated fragments differ from the original environment in two critical ways:

1. The ratio of edge to area is increased.
2. The center of each fragment is nearer the edge.

Fragmentation of the habitat threatens the persistence of the species associated with it in various ways:

First, fragmentation may limit the capability of a species to migrate to, and thus to colonize, other habitats, because the landscape between the fragments creates an effective barrier to dispersal. In an undisturbed habitat, individuals can disperse over the whole landscape, which can lead to a stable metapopulation despite local extinctions (discussed earlier). If, however, the habitat is fragmented, the potentials for dispersal and for colonization are often reduced. Thus, many birds or mammals, and many insects, are reluctant to leave the cover of their forest habitat and to cross small open spaces, as, if they do, they may become the victim of predators prowling forest edges. Cultivated fields, measuring only 100 m in width, can constitute an insurmountable barrier to numerous invertebrate species. If the dispersal ability of, say, some mammals is reduced, plants might also suffer, as they may rely on these species for the dispersal of their seeds. Therefore, isolated fragments of habitat may develop that, though in principle suitable for colonization, cannot be reached by dispersing species; they will therefore remain uninhabited.

As some species disappear from a fragmented landscape as a result of natural processes in a metapopulation, while others do not succeed in recolonizing the habitat, the number of species will decline over time. It must be noted here that, worldwide, most natural parks and reserves are too small and too isolated to harbor a large number of species; almost half of the protected sites have an area of less than 100 km², and 98% have an area less than 10,000 km².

By splitting a large and contiguous population into subpopulations restrained to limited areas, the fragmentation of the environment can precipitate the decline of species and their extinction. As we have seen, small populations are indeed more vulnerable to inbreeding depression, to genetic drift, and to several other problems. While a large contiguous area may support a single large population, it may well be that none of the fragments can harbor a population large enough to ensure its long-time persistence.

### TABLE IV
Factors Responsible for Some Extinctions and Threatened Extinctions

<table>
<thead>
<tr>
<th>Group</th>
<th>Extinctions</th>
<th>Threatened extinctions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentage due to each cause</td>
<td>Percentage due to each cause</td>
</tr>
<tr>
<td></td>
<td>Habitat loss</td>
<td>Over-exploitation</td>
</tr>
<tr>
<td>Mammals</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>Birds</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>Reptiles</td>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>Fishes</td>
<td>33</td>
<td>4</td>
</tr>
<tr>
<td>Mammals</td>
<td>68</td>
<td>54</td>
</tr>
<tr>
<td>Birds</td>
<td>58</td>
<td>30</td>
</tr>
<tr>
<td>Reptiles</td>
<td>53</td>
<td>83</td>
</tr>
<tr>
<td>Amphibians</td>
<td>77</td>
<td>29</td>
</tr>
<tr>
<td>Fishes</td>
<td>78</td>
<td>12</td>
</tr>
</tbody>
</table>

*a* From Reid and Miller (1989).

*b* The values represent the percentage of species that may be influenced by more than one factor; thus some rows may exceed 100%.
Wildlife habitat loss is one of the foremost threats to biodiversity and the destruction of tropical rain forests has come to be synonymous with the loss of biodiversity. In fact, tropical moist forests occupy 7% of the earth’s land surface but are estimated to contain over 50% of its species (Myers, 1986).

The original extent of the tropical rain forests has been estimated at 16 million km², based on current patterns of rainfall and temperature. From a combination of ground surveys and remote sensing data from satellites it is possible to assess the present area of tropical forests as well as the percent deforested per year: it ranges from 0.5% per year (Brazil) up to 3% (Ivory Coast), with an average around 2% year.

However, estimates of deforestation rates may vary tremendously, depending on the definitions and assumptions used in quantifying both the forest’s original extent and the amount of forest that has been converted to another land use. On a global scale, the primary cause of rain forest destruction is small-scale cultivation of crops by farmers (45,000 km² per year). Another 45,000 km² per year is destroyed through commercial logging. A further 25,000 km² is degraded for fuelwood production, mostly to supply local villagers with wood for cooking fires. The remaining 20,000 km² per year is cleared for animal breeding and pasture for cattle (Table V). The relative importance of these activities varies by geographical region, with logging being a more significant activity in tropical Asia, cattle ranching being more prominent in tropical America, and farming and fuelwood gathering more important in tropical Africa.

The national figures obscure the probable impact of humans in the forest, because no allowance is made for the effects of forest fragmentation. This phenomenon has not yet been assessed at the global level.

Second, the increased proportion of habitat edges in fragmented environments brings with it changes in microclimate—light, temperature, humidity, wind—but also increased risk of fire and higher vulnerability to novel species of predators and competitors (by, for example, invasion of weedy species from disturbed environments).

B. Persecution and Exploitation of Populations

The persecution and exploitation of natural populations generally threatens a certain number of species with extinction, in particular the large-bodied ones. These, and in particular the vertebrates, are also those that attract most attention from the general public, from protection agencies, and from the scientific community. This is reflected in the considerable amount of scientific literature specializing on a rational and controlled exploitation of natural populations, be it by hunting or fishing.

One generally acknowledges that large animals require larger reserves than small ones, due to their low population densities. However, though it is generally true that large animals have small population densities, this may be due in many cases to the past persecution or exploitation of these species. It must be emphasized that large-bodied species may survive in disturbed habitats, if they are not persecuted. Thus, in Finland, large birds and mammals, including predatory species, have become more abundant recently in a range of manmade landscapes, most likely because they are now protected.

C. Changes in the Biotic Environment

A common cause of extinction is the interaction with exotic species that had either been introduced by humans or arrived naturally (Box 4). The introduction of exotic species threatens in particular small isolated islands with specialized endemic species. Large islands, however, are not protected from this effect: after being introduced to Australia, the red fox became the major cause of extinction of several small marsupials.

The isolation of insular habitats favours the evolution of endemic species, but it also makes these particularly vulnerable to invading exotic species. Only a limited number of species can reach islands; plants, birds, and invertebrates are the most common colonizers. On the other hand, insular communities generally have few or no predators and browsers (due to the difficulties of colonizing the island or of establishing a population on an insufficient area), and indeed, the species representing the highest trophic levels (e.g., the carnivorous mammals) may be missing completely. Since many endemic species on islands have thus evolved in the absence of the selective pressures of predators and mam-
Examples of Extinction Cascades due to Introduced Species

What happened to Guam's avifauna is a good example of the destructive effects of introduced species. Guam is an island belonging to the Mariana Archipelago, situated between Japan and New Guinea. Eighteen species of indigenous bird species were known to inhabit the island, as well as seven introduced ones. Then, during the past 20 years, populations declined spectacularly: today seven species are considered to be extinct and four others have become so rare that their survival is endangered. Nevertheless, on other islands of the same archipelago, no comparable decline has been observed. Strikingly, the 10 forest species were all affected and in a similar manner: the birds disappeared first from the southerly forests during the 1960s, and then their decline spread progressively toward the north of the island. This decline coincided exactly with the introduction and subsequent spatial expansion of a tree snake, Boiga irregularis. Since this snake is absent on neighboring islands, there is good evidence that it was the cause of the observed declines and extinctions. Because of its arboreal and nocturnal habits, this snake is a voracious predator of perching birds and birds sitting on their nests, as well as of eggs and nestlings. Moreover, because it predates on small mammals and lizards as well (the latter particularly abundant), it can attain high densities even while exterminating its most vulnerable prey species.

On Santa Catalina, an island close to California, 48 native plant species have been eliminated mainly as a consequence of overgrazing by goats and other introduced mammals.

At Madagascar, where the ichthyofauna is highly endemic, with 14 out of 25 genera unknown elsewhere, a recent inventory of freshwater environments was not able to retrace more than 5. Introduced fishes dominate all aquatic environments. The combination of habitat degradation and the introduction of exotic fishes seem to lead the original ichthyofauna toward extinction.

The same has been found in continental aquatic environments, which for aquatic species are a kind of island surrounded by inhospitable terrestrial space. Originally, Lake Victoria had more than 350 endemic fish species. Today, many are rare or have gone extinct after the introduction of the Nile perch, Lates nilotica, in 1960. In fact, the phenomenon is more complicated than that: in 1978, the perch did not yet represent more than 2% of the annual catch; in 1986 it represented 80%. Other factors than predation have played a role (algal blooms creating anaerobic conditions as a consequence of pollution by fertilizers and other pollutants).

Mammalian browsers, they have not evolved (or have lost) any means of defense: birds have lost the ability to fly and lay their nests on open ground; plants do not produce any chemical substances and do not have any protective tissues (spines) that could deter browsers.

Therefore, endemic species that can spread in the absence of these selective pressures may go extinct rapidly when the pressures appear: animals introduced into the islands eliminate them by predation or overgrazing. The introduced plants, however, equipped with a protected or toxic foliage, have previously evolved to withstand browsing, so that they outcompete the endemic plants, enhancing the selective pressure by the introduced browsers and thus accelerating the extinction of the endemic species.

Moreover, insular species usually do not have an immune defense against foreign diseases; once introduced with the invading animals, these can spread epidemically and can devastate the native populations. It is thought, for example, that the almost complete destruction of the Hawaiian avifauna—since 1830, 89% of the endemic species have gone extinct or have been reduced to very small populations—is due to the spread of variola and of bird malaria, introduced together with the mosquito Culex quinquefasciatus in 1826. Manmade extinctions in insular communities are much more frequent than generally recognized. Thus, the rate of extinction on the order of 1% given earlier is a gross underestimate of the true value. Current estimates are that 25% of endemic bird species have disappeared from islands as a result of human activities. There are, for example, only three species of winged rails today, while thousands have existed during the past 2000 years. We are rightly worried by the current wave of mass extinctions threatening the tropical forests, but should also recognize that the oceanic islands have already experienced such mass extinctions.
D. Ecological Reflection on the Primary Causes of the Current Crisis of Extinction

All organisms modify their environment, and humans are no exception. Thus, in a strategic document, jointly published by the World Resources Institute (WRI), the World Conservation Union (IUCN), and the United Nations Environmental Programme (UNEP), six fundamental causes for the impoverishment of biodiversity are recognized:

1. The rapid and unsustainable growth of human populations and the consumption of natural resources
2. The continued reduction of the range of products in agriculture, forestry, and fishing
3. The economic and political systems that do not take into account the environment and its resources
4. The inequalities in the possession, the management, and the sharing of the advantages related to the use and the conservation of biological resources
5. The legislative and institutional systems that favor unsustainable exploitation of resources
6. The lack of knowledge and its applications.

In fact, diversity decline results from specific choices of path during the course of human development (Swanson, 1995). It is clear that, with population growth, demographic changes, and the associated technological development, humankind is using an increasing share of the planetary resources. Indeed, it consumes, diverts, or holds approximately 39% of plant productivity, the fundamental source of energy for most living systems (Vitousek et al., 1986). The rate of conversion from natural environments into agricultural systems, high in the developing countries (Table V), remains a major threat for biodiversity. This is alarming not only for the future of many animal and plant species, but also for our own.

In addition to the dangerous reduction of the renewable resources from the environment, population growth leads to increases in pollution (including the gases responsible for the greenhouse effect), which threaten the balance of ecosystems and of our planet’s climate. Novel models of development are necessary so that population growth does not exceed the carrying capacity of the planet.

The five other primary causes blame either our economic behaviors or the legal uses of the human societies. For thousands of years, the human world formed a mosaic of relatively autonomous areas. The state of knowledge, the strategies of subsistence, and the social structures evolved more or less independently in each area. What the populations required of the environment seldom exceeded the capacity of nature.

With the globalization of the economy, starting at the end of past century, the uniformity and the interdependence have increased. In agriculture, for example, a small number of cultivated plants have come to dominate the global economy (Box 5). In fact, the human societies today mainly depend on four species of plants (maize, wheat, potato, and rice) for their basic needs. Furthermore, these species are used in the form of a decreasing number of high-output varieties. Thus, on Sri Lanka, the number of 2000 varieties of rice used in 1959 has decreased to 5 main varieties today; in India 10 varieties out of originally 30,000 represent 75% of production; about 62% of the varieties in Bangladesh and 74% in Indonesia are estimated to be derived from a single maternal plant (Box 5).

The process of “conversion,” which consists in investing capital in a type of resource that is economically advantageous, is the principal economical force threatening biodiversity: ‘Development and conversion go hand in hand, and conversion is the process by which habitat and its resident species are lost’ (Swanson, 1995).

The reduction of the number of cultivated species is accompanied by the disappearance of nitrogen-fixing bacteria, mycorrhiza, predators, pollinators, and many other species that have coevolved during centuries with the traditional systems of agriculture. The use of ma-
It is rather surprising to note that human civilization uses a severely reduced number of species for agriculture and livestock: a couple of hundred plant species and a couple of tens of animal species. Most of the agricultural production is represented by at most 20-odd of these species. Almost all of present-day agronomical research, whose aim is to increase the yield of agriculture, focuses on the improvement of cultivars of the 20 main plant species that provide about 80% of the annual world harvest. Four of those species—wheat, maize, rice, and potato—already account for almost half of it.

Furthermore, the Green Revolution has led to the adoption of modern varieties and at the same time a considerable genetic erosion. Thus, in Indonesia, 1500 local varieties of rice have disappeared during the past 15 years. In Kenya, of 13 wild coffee varieties, only two are not endangered and two have already disappeared. In 1991, the genetic homogeneity of orange trees in Brazil has favored the worst epidemic of root cancers seen in the country. In Asia, following the Green Revolution, a considerable proportion of rice varieties originate from a single mother-plant (62% in Bangladesh, 74% in Indonesia, 75% at Sri Lanka): similar epidemics can emerge at any moment.

Such an economic concentration on such a low number of species is even more paradoxical when it is taken into account that over 3000 plant species are edible! Worse even, during the past couple of centuries, species known and cultivated by indigenous populations have seen their use reduced, a consequence of cultural homogenization due to colonization. One of the most well-known cases is that of the amaranth (three species of the genus *Amaranthus*), of which the Aztecs consumed the seeds and whose culture was forbidden by the Spanish when they conquered Mexico. The same is true for a plant of the high Andean plateaux, the passerage (*Lepidium meyenii*), whose roots resemble those of the black radish and are rich in saccharose and starch. These plants, which used to be grown by the Incas of Peru and Bolivia, are on their way to extinction as they no longer cover more than about 10 ha.

Similarly, with regard to marine ecosystems, the vast world markets favored the development of "blind" fishing, for example, gigantic drifting nets that trap not only enormous quantities of target species, but also a considerable number of other species of fish, mammals, and birds.

In addition, several reasons led to an underestimate of the use of environmental resources. First, many resources are directly consumed without appearing on the markets. Second, the benefits of biodiversity are largely public goods. For example, while the protection of wetlands may benefit a population, the benefit is so diffuse that traditional economic theory cannot incorporate mechanisms that would preserve them. Correctly evaluated, the natural systems and their biodiversity are major economic assets. But as these systems are often underestimated, economic theory considers the conservation of the biodiversity an expenditure rather than an investment.

In response to this challenge, a new discipline is being developed that integrates economics, ecology, and public policy: ecological economics. As Timothy Swanion emphasized, "diversity decline is a specific form of institutional failure: the failure to create institutions that internalise the values of biodiversity within the decision-making of states and individuals making conversion decisions."

The reduction in the number of the species and the destruction of the habitats are the standard in many countries where a minority of the population has or controls most of the territory. The rapid profits obtained from excessive logging or fishing go to a minority, while the local population, which depends on the sustainable production of the resources, pays the price. In fact, the ownership is often more likely to be granted to those that cut down and colonize forests and other areas covered with natural vegetation than to the inhabitants of the forests living on the sustainable harvest of the natural products. Any uncertainty on the ownership dissuades good management practices and encourages overexploitation.

A second problem stems from the concentration of the control of the resources and to the responsibilities for decisions about environmental policy in the hands of townsmen. However, in many societies, it is the women that manage the environment and understand the value of the biodiversity for agriculture and health.

A third problem is that of international trade, the debt and the policies of technology transfer that support the inequalities and often reinforce the imbalances ob-
served within the nations. To preserve the biodiversity, the industrialized countries must reverse this flow. If the developing countries continue to be excluded from the markets, to be deprived of accesses to technologies, and to be blocked by their debts, they will have neither the means nor the incentives necessary to preserve their biological resources.

See Also the Following Articles

- Extinction, Causes of
- Extinction, Rates of
- Human Impact on Biodiversity, Overview
- Mass Extinctions, Notable Examples of
- Metapopulations

Bibliography


World Resources Institute (1992). Global Biodiversity Strategy. World Resources Institute (WRI), the World Conservation Union (IUCN), and the United Nations Environmental Programme (UNEP).
I. Introduction to Mammals
II. Mammalian Phylogeny
III. Phylogeny and Biodiversity: Patterns of Ordinal Diversification
IV. Evolutionary Trends
V. Geography, Biogeography, and Biodiversity
VI. New Discoveries
VII. Current Extinction Crisis

GLOSSARY

brachyodont Low-crowned cheek tooth commonly found in omnivorous or burrowing animals.
cloaca The chamber into which the gut and urinary and reproductive tubes empty before exiting the body.
craniomandibular Where the head, or cranium, and the jaw, or mandible, meet.
cryptic species Species that are virtually indistinguishable by normal morphological analysis and are, instead, defined by a combination of genetic, behavioral, and other characters.
cursorial Adaptable for running.
diastema A large gap between incisors and premolars that is thought to permit an animal space to manipulate food with the tongue.
digitigrade Literally, walking on the digits; a posture where the majority of the weight is borne by the metacarpal and metatarsal bones and the heel/palm is raised off the ground.

endothermy The metabolic condition of being hot-blooded and internally regulating body temperature.
eutherian A placental mammal of the subclass Eutheria.
fossorial A burrowing animal.
hypsodont High-crowned cheek tooth commonly found in grazing mammals.
keystone species A species that exerts an effect on the structure and function of a community out of proportion to its numerical abundance.
macropod A marsupial in the family Macropodidae, the kangaroos and wallabies.
marsupium Folds of skin that envelope the mammary glands and provide protection to infants; found in metatherian and protherian mammals.
megatherian A member of the infraclass of mammals that includes the marsupials.
montemone A mammal in the family Monotremata, which includes spiny anteaters and the platypus.
morphological Descriptive of a clade, or branch, of an evolutionary tree that has a single root, indicating that all members of that branch are descended from a common ancestor.
patagium The skin that stretches across the arms and legs of flying and gliding mammals.
phylogeography The study of the relationship between and among genetics, morphology, paleontology, and ecology to better understand the spatial aspects of evolution.
placental mammal See eutherian.
polyphyletic Descriptive of a group of animals that,
although grouped together in a phylogeny or evolutionary tree, do not share a common ancestor.

Ricochetal Bouncing movement that involves the release of stored energy in stretched tendons, movement that involves quick changes of direction, as in to ricochet.

Rumen The first, and usually largest, chamber of the four-chambered stomach found in ruminant ungulates, the place where cellulose digestion occurs.

Semelparous Species that breed once and die.

Sympatric Populations or species that co-occur in a single location.

Ranging in size from a shrew, which may weigh only a few grams, to the blue whale, which weighs over 150 metric tons, mammals are found in the air, on the ground, under the earth, and in the oceans, rivers, and lakes of the world. Modern mammals are represented by 135 families divided into 1135 genera, not an impressively diverse group of animals. Throughout geological history, there have been just over 400 families of mammals and perhaps as few as 5000 genera. Paleontologically, the structure of the mammalian middle ear is unique, with three small bones that were originally found in the jaw of more primitive vertebrates now forming the structure of the middle ear. Mammals are endothermic; i.e., they regulate their internal temperature. To support endothermy, mammals have greater cellular production of energy, lungs with a large capacity, and numerous red blood cells to transport oxygen; an efficient four-chambered heart; and increased digestive efficiency. Soft-tissue characters that are diagnostic of mammals include sweat glands, hair, a folded cerebellum, an epiglottis, and a complex lung structure. Of course, mammary glands are the single most common diagnostic feature for mammals: the production of milk with which to feed and nurture young is a shared feature of all mammals.

I. INTRODUCTION TO MAMMALS

Modern mammals are not an impressively diverse group of animals, at least not numerically. Mammals are an extremely well described class of animals, with, perhaps, only Aves being better examined. Well described and relatively well studied, modern mammalian diversity is limited to 26 orders, divided into 135 families with 1135 genera and classified, at present, into approximately 4650 species (Fig. 1). Mammals have about half

![Graph showing species diversity of recent mammalian orders.](image-url)
the diversity of modern birds and pale in significance when compared to the nearly million described species of insects and the perhaps 10 million species that remain undescribed. Historical consideration does little to mediate these discrepancies in diversity: throughout all geological history, there have been just over 400 families of mammals and perhaps 5000 genera.

What mammals lack in numeric diversity is made up in their physical and ecological variation. Ranging in size from just a few grams (e.g., shrews in the family Soricidae) to over 150 metric tons (the blue whale, Balaenoptera musculus), mammals are found in the air, on the ground, under the earth, and in the oceans, rivers, and lakes of the world. Their diversity of body types, diets, and habits has captivated both the general public and scientists, taking an entirely disproportionate share of scientific study and popular examination.

A. What Is a Mammal?

There are literally dozens of diagnostic characteristics of mammals, most of which are related to soft-tissue structure and metabolic function. For paleontological purposes, soft-tissue differentiation of mammals is of little or no value. Soft tissue is rarely preserved and its structure can only occasionally be deduced from the structure of the bones that it covers and from which it is suspended. Paleontologists, therefore, look for skeletal and cranial features that distinguish mammals from other classes. One defining character is found in the structure of the mammalian middle ear. The middle ear contains three small bones, the stapes, incus, and malleus, which were originally found in the jaw in the cranio-mandibular joint of more primitive vertebrates (the joint where the jaw meets the cranium). The simplification of the cranio-mandibular is also reflected in further simplification and solidification of cranial bones overall. This simplification and increased ossification are also reflected in the postcranial skeletal structure of mammals. Skeletal growth occurs early in the life history of mammals, with further growth of bones limited to a cartilaginous area at the end of the bone, under the articular surface, called the epiphysis. Eventual closure and ossification of the epiphyses occur after growth is complete.

Physiologically, a critical difference between mammals and reptiles is the development of endothermy, or internal regulation of temperature. While it is unclear whether the mammal-like Therapsids were endothermic, all modern mammals exhibit endothermy, while extant reptiles are ectothermic, using behavioral adaptations to regulate temperature in harsh environments or allowing their temperature to fluctuate widely. The cellular and structural differences that endothermy require are all hallmarks of mammalian life: greater cellular production of energy, lungs with a large capacity and numerous red blood cells to transport oxygen, efficient circulatory systems, and increased digestive efficiency. Soft-tissue characters that are diagnostic of mammals (many of which are related to endothermy) include sweat glands, hair, a four-chambered heart, a folded cerebellum, an epiglottis, and a complex lung structure. Of course, mammary glands are the single most common diagnostic feature for mammals: the production of milk with which to feed and nurture young is a shared feature of all mammals.

B. Overview of the Class Mammalia

Mammals are usually divided into three subclasses: the Allotheria, which was represented by the extinct multituberculates; the Prototheria, which reached the peak of its diversity in the Mesozoic and which today is represented by a single order, Monotremata, containing the two species of echidna (Tachyglossidae) and the duck-billed platypus (Ornithorhynchus anatinus); and the Theria, which contains two infraclasses, the Metatheria (marsupials) and the Eutheria (placental mammals).

1. Monotremata

Viviparity, or live birth of young, is so common among mammals that it is usually, wrongly, considered a defining character of the class. The three species of the order Monotremata all lay eggs. The platypus lays its eggs into a nest, similar to a bird's nest, while both species of the family Tachyglossidae, the echidna, or spiny anteaters, lay their eggs directly into a marsupial-like pouch. While this curiosity is the root of the common name for the monotremes—egg-laying mammals—the egg is actually a rather insignificant aspect of the monotreme's life history. Incubation is brief, under 2 weeks, following which monotreme development does not differ significantly from that of other mammals. Clearly mammals, they nurture their young with milk that is expressed from mammary glands that lack nipples. Like all mammals, they are endothermic, have hair, possess a single jaw bone, and have the diagnostic three-bone middle ear structure.

Divergence of the monotremes from other mammals occurred approximately 175 million years ago in mammalian history. Fossil monotremes have only been found from Australasia, and all extant species share this distribution. Monotremes appear to be extremely primitive in their reproductive habits, with not only an
egg-laying habit but also a single opening, or cloaca, into which both the excretory and reproductive tracts exit. The cloaca (or single exit) gives the order its name.

2. Theria

The therian mammals all share the derived character of live birth of their young. The subclass Theria is divided into two infraclasses: metatherians and eutherians. Metatherians are usually called marsupials because of the pouch, or "marsupium," that many members of the subclass possess. A far more diverse order than Monotremata, there are 273 described species of marsupials, which are categorized into 19 families. Metatherians and eutherian mammals probably diverged at about the same time as the monotremes, approximately 100 million years b.p. Since that time, the two major groups have had a divergent, and sometimes convergent, evolutionary history.

3. Metatheria: Marsupials

In Recent times, metatherian mammals have only been found in the Australian subcontinent, the adjacent large islands of New Guinea (the island consisting of Papua New Guinea and the Indonesian state of Irian Jaya) and Tasmania, and the Neotropics. This distribution is more restricted than that given by fossil evidence. Arising in the Mesozoic, about 100 million years ago, marsupials are first found in North and South America. The metatherians make their first appearance in the fossil beds of the Eocene of Europe, 50 million years ago, and it is thought they arrived in Asia, and Australasia, some time thereafter. The fossil evidence for this period from these regions is poor, and some have suggested that this, combined with the present distribution of marsupials, strongly implies an Antarctic origin.

In the New World, the Recent marsupial fauna has not developed significant diversity in competition with placentals. The New World family Didelphidae is represented by 63 Recent species in 15 genera. While found across North, Central, and South America, there is relatively little variation in size, with the smallest member of the family the size of a mouse and the largest only the size of a large cat. The most common species in the Americas is the Virginia, or common, opossum (Didelphis virginiana), which is widespread and extremely successful in adapting to human presence. The earliest members of this genus are found in South America and it appears that they migrated to North America just over a million years ago after the establishment of the Central American land bridge. While marsupials, for the most part, did poorly when they invaded North America, Didelphis is a clear exception to this rule.

In South America during the late Miocene and Pliocene epochs (approximately 5–10 million years b.p.), a diverse radiation of marsupials occurred. Convergent evolution produced a suite of marsupials that resembled contemporaneous placental mammals and included forms similar to large felids (in the family Thylacooleonidae). The most diverse family of carnivorous, South American marsupials was the Borhyaenidae, often called the family of dog-like marsupials because of the general resemblance of some species to canids (e.g., Lycopsis). Borhyaenids persisted only until the late Pliocene. Other families of South American fossil marsupials included rodent-like, mole-like, and rhino-like animals and perhaps the best known, the Thylacosmilidae, or the "false" saber-tooth tigers (Thylacosmilus).

In the Australian region, until Recent times, the metatherians have been free from competition with placental mammals, with the exception of bats and some murid rodents (rats and mice). Here, the metatheria have undergone a remarkable and fascinating radiation. Many species resemble, in basic form, the eutherian moles, squirrels, and small carnivores. These species tend to be relatively undifferentiated and follow a relatively similar body form within groups. In some cases, however, marsupials in Australia have evolved into forms that little resemble in structure their ecological replacements in areas where placental mammals dominate. While most grazing placental mammals are, to a greater or lesser extent, relatively similar in their body plan, the kangaroos and wallabies (Macropodidae) represent a distinctly different solution to the design of a grazer. The origin of the large rictocmathrm hind feet and vestigial, but highly flexible, forearms suggests an arboreal origin for the family.

4. Eutheria: Placental Mammals

The greatest diversity of mammals is seen in the eutherians. Both through geological time and in the present, eutherians are the most diverse of the three main branches of mammalian evolution. Representing nearly 95% of the extant species of mammals, eutherians have shown a remarkable variety of forms and adaptations. Fully aquatic forms are represented by the 10 families and 78 species of cetaceans. The 17 families of Chiroptera, representing over 900 species, have a near monopoly on the nocturnal skies and are keystone pollinators in tropical forests. But by far the most diverse order is Rodentia, with 29 mammalian families (approximately 25% of the total) and over 2000 species, nearly half the known mammals alive today. A more thorough discussion of eutherian diversity is given below.

One would expect that the defining character for eutherian mammals is the well-developed placenta,
from which the infraclass derives its name. The use of the term is unfortunate as marsupials possess one of four types of more primitive, but nonetheless well-developed, placenta, and placenta-like structures are seen in some fish and reptiles. In all mammals, the placenta serves to provide exchange of nutrients and waste products and acts as a respiratory organ for the developing fetus. Increased complexity of the placenta found in eutherian mammals allows for more efficient transfer and for added functions, including endocrine production, including progesterone, which helps sustain pregnancy.

C. Differentiating between Placental and Marsupial Mammals

1. Cranial and Skeletal Differences

Differentiation of marsupial and placental mammals is easiest in the examination of soft tissues; nonetheless there are a set of diagnostic characters that distinguish the metatherian and eutherian mammals. While the palates of placental mammals are smooth, those of marsupials have a series of diagnostic holes or vacuities. Marsupials also differ in their dental morphology: in contrast to the serial replacement of deciduous teeth with permanent teeth found in placental, only a single tooth is replaced in each jaw. With the exception of the third premolar, all premolar, canine, and incisor deciduous teeth are resorbed before they erupt in the jaw. Only a single set of molars forms and molars erupt sequentially. Recent fossil evidence from the Mesozoic marsupial Alphadon shows that this pattern of tooth replacement and development is a derived character and suggests that the trait is ancestral to marsupials.

A further primitive character common to the Monotremata and Metatheria, but thought to be lost in the eutherian mammals, is the epipubic bones, which extend forward from the pubic bone in both sexes. Hypotheses about the function of the epipubic bones have been various. Initial speculation was that the structure provided support for the attached young found in marsupials or support for the marsupium, in which they are often contained. Others have suggested it was involved with locomotion, while more recently it has been proposed that epipubic bones provided a place for the attachment of abdominal muscles, thus allowing for greater expansion of the abdominal cavity, increasing the function of the diaphragm and improving efficiency of respiration. Recent paleontological finds show that epipubic bones were present in Cretaceous eutherian mammals. Because the loss of epipubic bones is associated with prolonged gestation, these finds have been interpreted as evidence that the complex suite of adaptations that typify placental mammals were derived later in the evolution of eutherians.

2. Reproduction

While placental mammals tend to have long gestation periods and relatively short periods of lactation, gestation in the marsupials is always relatively brief and is followed by a much longer period of lactation. Typically, neonatal marsupials are extremely small, only a few grams, and show very incomplete development. The exception to this generality is seen in the jaw and forearm structure of the neonate: the nearly embryonic young need strong arms to navigate up to the mother's midline to the teat and well-developed jaws with which to form an immovable attachment to the teat once arriving. A greater proportion of development occurs outside the mother's uterus after the neonate has attached to a teat. In approximately half the marsupial species, the teats are found inside a marsupium. By the time the neonate leaves the pouch, or detaches from the teat, it weighs more or less the same as a placental mammal of an equivalent adult size.

Diagnostic differences in the structure of the reproductive tract of female marsupials and in the development of the embryo may explain many of the life history differences between eutherian and metatherian mammals. While eutherian mammals form a complex placenta that nourishes the embryo, marsupials retain an eggshell membrane with a simple yolk sac to nourish the young. This arrangement limits the nutrients that a mother can transfer to her embryos.

The size of a neonatal marsupial is also limited by the structure of the female reproductive tract. In eutherian mammals, the reproductive tract is arranged linearly (Fig. 2): the vagina leads directly into the uterus, with

![FIGURE 2 Comparison of eutherian and metatherian reproductive tracts. Reprinted with permission from Sharman, Reproductive physiology of marsupials © (1970) American Association for the Advancement of Science.](image-url)
two ureters transporting urine from the bladder to the urethra. Growth of the fetus is essentially constrained only by the size and elasticity of the vaginal canal. Given that the spotted hyena gives birth through a vagina that has evolved to look like a pseudopenis, this constraint does not appear to be too great. In contrast, in marsupials, the structure of the vagina and uterus is like a double jug handle (Fig. 2). Just above the urethra, two lateral vaginas form loops on either side of a pseudovaginal canal. The two vaginal loops rejoin and feed into a bipartite uterus. The ureters pass inside these vaginal loops. Because birth takes place through the lateral vagina, placement of the ureter inside this loop may limit the size of the neonate.

II. MAMMALIAN PHYLOGENY

A. Early History

Mammals have their origins deep in geological history. Approximately 200 million years ago, in the late Triassic period, primitive cynodont mammals evolved from their mammal-like reptile progenitors, the Therapsids. The Therapsids, whose members dominated the terrestrial landscape during the Triassic, faded into insignificance by the end of the Triassic/Jurassic boundary, leaving the stage of the late Mesozoic era open to the radiation of great sauropods, the dinosaurs. For 140 million years, dinosaurs filled most terrestrial (and many aerial) ecological roles, with mammals for the most part relegated to small, terrestrial rodentlike forms.

The earliest evidence of true mammals occurs in the late Triassic, although by this time, mammals were found worldwide, suggesting a somewhat earlier divergence from the mammal-like cynodonts. Poor fossil records from many areas leave gaps, which slowly are being filled. While the Mesozoic mammals were once thought to be lacking diversity, recent finds suggest this was not the case. Fossil evidence suggests they were never abundant and rarely showed any great size—the first mammal weighing more than 1 kg does not appear in the fossil record until the early Cretaceous. Yet a variety of unusual forms evolved in the Mesozoic, including Symetrodonia, which are characterized by well-developed, triangular molars, the predatory Triconodonta, and the omnivorous and herbivorous members of the Multituberculata.

The great radiation of modern mammals began in the mid-Cenozoic era, in the late Cretaceous period, approximately 100 million years ago. More rapid evolution of larger body sized mammals did not begin until the mid-Cenozoic, beginning about 65 million years ago during the Paleocene epoch of the Tertiary. Following this, mammalian radiation accelerated.

B. Ordinal Diversification

1. Structure of the Mammalian Evolutionary Tree

The ordinal diversification of the mammals was more or less completed in the late Eocene, 50 million years ago, and has been relatively stable since that time. Of the 32 major orders that were present 50 million years ago, 27 are extant (Fig. 3). Reconstruction of evolutionary paths of ordinal diversification can be made using morphological analysis of modern or fossil taxa, molecular analysis (limited to species from which DNA can be extracted), or some combination thereof. Extant species have often lost defining skeletal and cranial characters present in earlier fossil forms but have the advantage of soft tissue being available for study. Fossils offer a long time frame, but significant gaps in the fossil record and the absence of soft tissue for comparison may be problematic. Molecular genetic studies have become increasingly sophisticated but suffer from their own inaccuracies and, of course, are useless for independently reconstructing phylogenies from taxa that have gone extinct, leaving no modern (or DNA-bearing subfossil) forms.

Morphological and molecular data agree remarkably well on the general structure of the phylogeny of the major mammalian orders, although significant questions remain about the exact relationship of several orders. While there was hope that the addition of molecular data to the discussion of mammal phylogeny would clarify some of the more confusing aspects of ordinal diversification, present molecular methods have not met that expectation.

A good example of the problem is the placement of the Hyaenidae, the hyraxes. Both protein sequence data and morphological studies suggest a close affiliation of the hyraxes with the Proboscidea (elephants) and the Sirenia (dugongs and manatees). Yet other morphological analysts suggest that hyraxes and perissodactyls (rhinos, tapirs, and equids) share a common lineage. Gene sequence data, while not without problems of its own, may offer some resolution of such problems.

2. Timing of Ordinal Diversification

While scientists agree that there was a period of rapid diversification of higher orders of mammals, the exact

timing of this event is still hotly debated. At the root of this debate are the differences in timing derived from molecular and fossil evidence. The data supporting the timing of events presented above, with the major ordinal diversification completed in the early Tertiary, are derived from the near lack of any fossil evidence for the existence of modern orders of mammals before the Cretaceous/Tertiary boundary 65 million years ago.

Advances in molecular biology have permitted more sophisticated molecular analyses of ordinal diversifica-
tion of both birds (which are purported to have a similar timing of ordinal diversification) and mammals. These analyses, which employ a number of different genetic "molecular clocks," all point to an early ordinal diversification of birds and mammals some time in the middle to late Cretaceous, approximately 90–100 million years B.P. That the estimated timing of diversification is similar for birds and mammals supports this interpretation. However, of greater interest is the coincidence of the timing of ordinal divergence and the fragmentation of emergent land areas at about the same time.

By the early Cretaceous, the two main land masses formed by the breakup of Pangea, Gondwanaland, and Laurasia had further subdivided into seven continents. By the late Cretaceous, 70 million years ago, 11 continents had appeared, the peak of land mass isolation and diversity. Earlier explanations had suggested that the rapid ordinal diversity resulted as mammals (and birds) radiated into a diversity of empty niches on the new continents. The molecular data suggest, however, that the breakup of land masses itself provided the opportunity for diversification.

If the molecular data are correct, fossil evidence for two dozen orders is "missing" for upward of 30 million years, and by some estimates 60 million years, of stratigraphic history. The real question asked by many paleontologists, therefore, is, "How likely is it that an increasingly good fossil record could be devoid of evidence for such diversification?" The answer appears to be, not likely. Using standard models to examine the probability of not finding a species in the fossil record, scientists estimate that, even under the most severe conditions, there is a 2% chance of not finding earlier fossils if, indeed, ordinal diversification occurred before the Cretaceous/Tertiary boundary. One possible reconciliation of these two viewpoints would be the Cretaceous splitting of a few lineages, thus giving rise to earlier ordinal diversification, followed by a rapid radiation during the Paleocene.

Whatever the timing of the ordinal diversification of mammals, by the mid-Tertiary, continental drift resulted in a relatively large number of isolated land masses, evolutionary caldrons in which mammalian diversification occurred. Peak continental isolation occurred at the Cretaceous/Tertiary boundary when landmasses, which in modern times are joined, were isolated. This included present-day continents (e.g., South America and Africa) and subcontinents (e.g., India). Even landmasses that were joined, such as Asia and Europe, exhibited greater isolation as a result of landscape features.

Island continents, in and of themselves, would have likely produced greater mammalian diversity over time. However, in addition to the isolation of continents, the Tertiary was a time of greater differentiation and seasonality of climates. Much evidence exists for this, from diversification of Tertiary flora (particularly in the late Tertiary) to developments such as the Antarctic ice cap, which formed in the Miocene. Continental drift, with landmasses moving toward the poles, combined with concomitant changes in ocean currents resulted in increasingly temperate environments. For the homothermic mammals, this diversification of ecosystems provided a unique opportunity for radiation. In Recent times, geologically isolated continents and large islands, such as Madagascar and Australia, continue to show unique assemblages of mammals that are typified by high levels of endemism and diversification of a small number of families.

III. PHYLOGENY AND BIODIVERSITY: PATTERNS OF ORDINAL DIVERSIFICATION

Mammalian generic diversity appears to be rising exponentially throughout the Cenozoic (Fig. 3). While there is some fluctuation in the levels of diversity, the progressive increase in diversity is notable. This pattern may be driven, in part, by an increasingly heterogeneous environment. However, for the most part, the exponential rise is probably an artifact of an increasingly well represented fossil record. Mammal diversity probably peaked during the mid-Tertiary, although patterns of diversification (and simplification through extinction) vary from continent to continent. The following sections will briefly describe patterns of diversity in the extant orders of mammals.

A. Monotremes

The three extant species of the Monotremata, found only in Australasia, all represent highly specialized representatives of an order that had an early branching from all other mammalian orders. Living monotremes are toothless, with the exception of young platypuses, a trait that has not made reconstruction of fossil history a simple task. Early fossil remains are known from Australia. A single fossil found in southern Argentina, dated to about 60 million years B.P., represents the only occurrence outside of Australasia.

Monotremes possess what are surely primitive mammalian characters such as egg laying and a cloacal
reproductive/excretory structure. While these primitive characters reflect an early divergence, monotremes are not primitive in all their characters. For instance, all three extant species of monotremes share complex and well-developed nervous systems. The most unusual sensory adaptation is the development of electroreceptors across the platypus’s bill and in the tip of the echidnas, producing snout or beak. Unique among mammals, electroreception in monotremes is complemented by a complex and large brain.

Monotremes are also the only mammals that have the ability to deliver venom as a defensive measure. While venom glands are vestigial in the echidnas, in the male platypus a well-developed venom gland is found in the hind legs. More active during the mating season, the venom is assumed to be a defense against intruding males. The venom is toxic to humans and lethal to small mammals.

The sole surviving representative of the Ornithorhynchidae is the platypus, Ornithorhynchus anatinus, found only in eastern Australia and Tasmania. The morphology of the species is so distinctive that after early examination of skins sent to London, they were pronounced a fraud. Weighing around a kilogram and with an overall length of 45 cm, the platypus has a broad, duck-shaped bill. The animal is covered in short, dense fur and has webbed feet with claws for digging. The animal is highly aquatic and solitary outside of the breeding season. Females will expand their burrows before breeding and, after mating, will “gestate” one or two eggs for 10–12 days. After laying, the egg is incubated for a further 12 days. Lactation is protracted, approximately 4 months, but the young becomes independent soon after weaning.

The Tachyglossididae is represented by two extant monotypic genera: Zaglossus bruijnii, the long-beaked echidna, and Tachyglossus aculeatus, the common echidna, or spiny anteater. Zaglossus is endemic to the mountains of New Guinea and feeds primarily on earthworms and arthropods, while Tachyglossus has a distribution across much of Australia, found wherever its primary prey, ants and termites, are abundant. Introduction of the dingo 10,000 years ago probably has reduced the range of Tachyglossus, much as it has been thought to have a similar effect on the platypus. The echidnas are larger than the platypus, with Zaglossus weighing up to 10 kg and Tachyglossus reaching about half that weight. The two species have thin protruding snouts and thick coats of dark brown fur interspersed with sharp quill-like spines. Patterns of reproduction differ from those of the platypus. The egg is deposited into a pouchlike structure, where it hatches 10 days later. The young stays within the pouch for 2 months and then nurses outside the pouch for a further 3 months.

### B. Marsupials

Extant marsupials are grouped into 7 orders, with 21 families, 83 genera, and 276 species (Table 1). In a pattern commonly seen among mammals, some lineages contain disproportionately more families, genera, and species than others. For instance, more than half of all extant marsupial families are in the order Diprotodontia, while three orders are represented by only a single family. The distributions of genera within families, and species within genera, are also highly skewed. Three orders of marsupials are found only in the New

### TABLE 1

<table>
<thead>
<tr>
<th>Order</th>
<th>Common name(s)</th>
<th>Number of families</th>
<th>Number of genera</th>
<th>Number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Didelphimorphia</td>
<td>American opossum</td>
<td>1</td>
<td>15</td>
<td>63</td>
</tr>
<tr>
<td>Pseudoherpestidae</td>
<td>Rat opossum</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Microbiotheria</td>
<td>Mimito del monte</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dasyuromorphia</td>
<td>Thylacine, numbat, quoll, amecinus,</td>
<td>3</td>
<td>17</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Tasmanian devil</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peramelomorphia</td>
<td>Bandicoots</td>
<td>2</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>Notoryctomorphia</td>
<td>Marsupial mole</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Diprotodontia</td>
<td>Koala, kangaroo, wombat, wallaby,</td>
<td>10</td>
<td>35</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>glider, feathertail, cuscus, phalanger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>21</td>
<td>83</td>
<td>276</td>
</tr>
</tbody>
</table>
World, while the remaining four orders are found in Australia. The Old World continents of Africa, Asia, and Europe have no extant marsupial representation.

1. Didelphimorphia

The American possums, or didelphid marsupials, represented by a single modern family, are found throughout South America, with the Virginia opossum (Didelphis virginiana) found widely over eastern, central, and southern North America. A generalist capable of living on garbage, the Virginia possum has shown a remarkable range expansion in the past two centuries, expanding its range from the southeastern United States north into the central states and the southern fringes of eastern Canada and along the west coast up to and just over the Canadian border. In the past 50 years alone, the species has extended its range over an area of two million square kilometers. This expansion has been facilitated by the opossum’s extreme potential fecundity. A female may give birth to up to 60 very altricial young, up to a quarter of which may successfully attach to one of her 15 teats.

Despite ranging in size from a small mouse to a large cat (25 g to 5 kg), most didelphids show a clear set of similarities. Arboreal habits are common in the Didelphidae as is reflected in the general morphology of family members. Hands and feet are well developed, with an opposable big toe on a hind foot that has five toes. Many species also have prehensile tails.

The dentition of most didelphids is relatively unspecialized, with 50 teeth found in most species. The family is dominated by generalists, many of which have wide geographic ranges. However, some of the more remarkable specialists are also found in the didelphids. The yapok, or water possum (Chironectes minimus), is the only truly aquatic species, with a fully webbed hind foot and a carnivorous diet consisting of a variety of crustaceans, fish, and amphibians.

2. Paucituberculata

Paucituberculata is also represented by a single Recent family, the Caenolestidae. All six extant species, placed in three genera of shrewlike marsupials, are found only at higher altitudes in the Andean cordillera, from Ecuador south to Chile. Capable of killing small vertebrates with their sharp teeth, the caenolestids usually subsist on invertebrates, particularly earthworms. Weighing under 20 g, the shrew opossums are nocturnal omnivores, using their long snouts and tactile whiskers, or vibrissae, to locate and kill prey. The extant paucituberculats are represented by a much more diverse fossil fauna. In the Miocene, seven genera of caenolestids were represented by a large number of species with widespread geographic distribution and a relatively large diversity of life forms, including specialized predators.

3. Microbiotheria

A single Recent species of Microbiotheria represents an order whose fossils have been found on three continents: South America, Australia, and Antarctica. With origins in the Paleocene, the microbiotherids are thought to have invaded South America while it was connected to Antarctica and Australia via a southern land bridge. The historical diversity of the group in South America is poorly known. The extant arboreal marsupial mouse inhabits the forests of southern Chile and is nocturnal.

4. Dasyuromorphia

Dasyuromorphia is known from fossil evidence in Australia in the mid-Miocene (Dasyuridae), but by this time the order already exhibited a relatively large diversity of morphotypes. Extant members of the order are represented by 3 families and 63 species, approximately 20% of all Recent marsupials. All but two of these species can be found in the family Dasyuridae.

In historical times the Thylacinidae were represented by a single species, the thylacine Thylacinus cynocephalus, also called the Tasmanian wolf or Tasmanian tiger. This species has been considered extinct since the last specimen died in captivity in 1936. Thought to have had a wide distribution across Australia, introduction of the feral domestic dog, or dingo, by Polynesians approximately 10,000 years ago probably led to its extinction in Australia. The cause of extinction of the relictual population in Tasmania is debated, although it was most likely the result of a number of interacting factors common to human predator extinctions: habitat loss, disease, and declining prey supplanted by sheep farmers, who hunted the thylacine because their livestock was perceived to have been threatened by predation.

The Myrmecobiiidae consists of a single, monotypic genus, represented by the numbat, or banded anteater. Weighing approximately 400–500 g, the numbat resembles a small grayish-red squirrel, with a long bushy tail, a half dozen lateral white stripes circling its abdomen, and a prominent protruding snout. As its family name would suggest, however, the numbat is a termite specialist. Like other termite specialists (e.g., the bat-eared fox of Africa), the historical distribution of the species was dictated by the occurrence of its primary prey. While termites are found across southern and central
Australia, today the numbat persists in small fragments of *Eucalyptus* forest and woodlands (hardly core habitat), its last refuges from agriculture and predation by the introduced red fox (*Vulpes vulpes*). The 61 species of the family Dasyuridae are often called the marsupial carnivores despite the fact that most members of the family are insectivores. Their dentition, while similar to that of most marsupials, is similar to the dentition of many carnivores, characterized by bladelike incisors, large, sharp canines, and upper molars modified with large, sharp cusps. Found across Australia and New Guinea, dasyurids are usually small, with most species weighing under 50 g. The family occupies a wide diversity of terrestrial niches, with the smallest species convergent on shrews (family Soricidae) found in Asia, Europe, and North America.

Some of the dasyurids, particularly species in the genus *Antechinus*, have remarkable life histories. For instance, Stuart’s marsupial mouse, *Antechinus stuartii*, has a semelparous life cycle more similar to that of many insects than to that of a mammal. The species is usually arboreal but forages terrestrially for invertebrate prey. Mating is highly seasonal, with males competing aggressively for mates. Females produce an enormous litter for their size (10–12 offspring). The young forage independently at 3 months and are sexually mature by 10 months. What is most remarkable is that after mating, the males die, while females rarely survive to breed in a second breeding season. Similar patterns of reproduction are seen in other *Antechinus* species.

The two largest species of dasyurids rightfully deserve the name marsupial carnivore. The tiger quoll, *Dasyurus maculatus*, may weigh up to 7 kg (males) while the largest extant species of Australian carnivore, the Tasmanian devil (*Sarcophilus harrisii*), can reach the size of a medium-sized dog, weighing up to 9 kg. Like the thylacine, the Tasmanian devil once had a widespread distribution across Australia but is now extinct on the continent. It persists in Tasmania, perhaps because it is primarily a scavenger, not a predator, and thus is more catholic in its dietary requirements and less of a perceived threat to farmers. Both quolls and Tasmanian devils are usually solitary; however, aggregations of up to 20 Tasmanian devils have been observed around carcasses.

5. Peramelomorpha

Found from the deserts of Australia to the rain forests of New Guinea, the 21 species of bandicoots and bilbies are placed in 8 genera and 2 families, the Peramelidae and the Peroryctidae. The bandicoots are insectivorous or omnivorous and for the most part are ecological counterparts of insectivores. Most species forage by digging insects out of the ground. They have small, sharp teeth, and ecological studies show that, despite feeding specialization for insects, they are omnivorous and opportunistic in their diet choice.

Bandicoots vary in size from the rodentlike mouse bandicoot (*Miconycteris murina*) at 15–18 cm up to the giant bandicoot, *Peroryctes broadbentii*, at 45–50 cm and up to 5.5 kg. In recent times, the evolution of bandicoots has taken separate paths in Australia and New Guinea. Of the 11 species of peroryctids, or spiny bandicoots, only two are found outside New Guinea. One species is found on the Indonesian island of Seram, in the Molucca or Spice Island archipelago, due west of New Guinea, while another has a relictual distribution in the rain forests of Cape York in northern Australia. Little is known about the fossil history of this family as New Guinea is all but lacking in fossil record. On New Guinea, species occur at different altitudes, with the most species (six) occurring in midaltitude regions of 1000–1500 m above sea level.

Like perrisodactyls and artiodactyls, peramelid bandicoots show relatively extreme reduction in their limb morphology. Many peramelids have digitigrade elongated hind limbs, with expanded growth of the fourth digit, and lateral digits showing varying ranges of reduction. Reduction in limb digits suggests strong selection for running in the bandicoots. While the bandicoots are remarkably cursorial, some of the elaboration of the hind limb may also have evolved for improved digging and burrowing.

6. Notoryctemorpha

The only Australian marsupial adapted for a fossorial existence is the sole species of the order Notoryctemorpha, *Notoryctes typhlops*. The marsupial mole has a pale, golden yellow coat and a long, tubular snout about 15 cm long, weighing 35–40 g. There are rudimentary, almost vestigial eyes, compressed neck vertebrae, and stubby, almost vestigial forelimbs with two well-developed claws on the third and fourth digits. Other digits have essentially disappeared. The hind limbs are flattened, with only three small claws, and are used to expel dirt from a burrow. Little is known about the behavior or ecology of the species, although it is believed to be solitary and has been observed to eat burrowing insects and their larvae.

7. Diprodontia

This order includes both the largest and smallest marsupials. The smallest member of the family Burramyidae, or pygmy possums, weighs 6–8 g (*Cercartetus leptus*).
Known only from fossil remains until 1996, the family is now known to have five extant species placed in two genera. At the other extreme of the diprotodont scale are the members of the family Macropodidae, the kangaroos and wallabies. While the smallest macropod, the rock wallaby (Petrogale nuridei), weights in at 1 kg, the great gray kangaroo may reach 90 kg and stand flat-footed at just over 2 m. Diprotodontia also shows the greatest morphological variation on any extant marsupial order. While it is beyond the scope of this article to review all 117 species in the 10 families that make up the diprotodonts, a review of some of the more divergent taxa is informative.

The single species in the family Phascolarctidae, the koala (Phascolarctos cinereus), like the kangaroo, is emblematic of the Australian continent. Highly selective in its diet, the koala subsists entirely on the leaves of a few species of smooth-barked eucalyptus trees. Unlike many marsupials, the koala is slow to reproduce, with a single young born to a female each year and the time of dependency between mother and young stretching to up 1 year. Sexual maturity is not reached until age 3 or 4.

The feathertail glider (Acrobates pygmaeus) and the feathertail possum (Distoechurus pennatus) are the only two species in the aptly named family Acrobatidae. Found only in New Guinea and with no known fossil record, both species have long stiff hairs on either side of their tails, from which they derive their name. The gliders, which weigh a mere 10–15 g, have membranes between the elbows and knees, further aiding gliding flight. The membrane is lacking in the possum, which weighs up to 50 g.

Gliding as a strategy appears to have evolved twice in the diprotodonts. Five species of lesser glider, along with the striped and Leadbeater's possum, make up the family Petauridae. Found from Tasmania to New Guinea, the petaurids resemble small squirrels. Weighing from 100 to 700 g, they have long bushy tails and stripes down their backs. The striped possum has black and white dorsoventral stripes along the back, resembling a North American skunk in coloration.

The five gliders bear a remarkable resemblance to flying squirrels of the genus Glaucomys. A membrane that stretches between their wrists and ankles provides an almost rectangular, kitelike gliding surface that these animals use to move from tree to tree. While the striped possum is an insectivore, all other members of the family specialize on eating plant exudate (sap and gum) as well as insects. The sugar possum (Petaurus breviceps) is particularly specialized, concentrating its efforts on the exudate from a single species of Eucalyptus.

Resembling fossorial rodents, the three extant species of wombat have short, muscular limbs, stocky bodies, and broad long claws. They are large animals, weighing up to 40 kg. Prodigious diggers, they live in complex burrow systems, which caused them to be called marsupial badgers on their first discovery. Unique among marsupials, their teeth are ever-growing like those of rodents and lagomorphs. Wombats are herbivores, concentrating their diet on tussock grasses. Wombats are found only in southern Australia and Tasmania, with the northern hairy-nosed wombat (Lasiorhinus krefftii) found in one isolated population of 70 individuals living on 3 square kilometers in central Queensland. Wombats have a long fossil history in Australia, with two much larger species occurring throughout the Pleistocene and going extinct about 10,000 years ago, coincident with the arrival of humans and dogs.

The best known family of diprotodonts, and probably the most widely recognized family of marsupials, are the 11 genera and 54 Recent species of macropods, the kangaroos and wallabies. Ranging in size from 1 to 90 kg, all species share the same basic body plan. Forelimbs are reduced and have five toes, all bearing small claws. The hind limbs are elongated, and, as in the bandicoots, on the hind leg all but the fourth digit has been reduced or lost. The large hind foot allows macropods to take a bipedal stance.

The limbs are highly adapted for ricochetal or hopping locomotion that, because of the limb structure, is highly energy efficient. Each time the kangaroo bounces down, the large elastic tendons in the foot are stretched taut like a rubber band, storing energy from the previous bounce. As the foot is released, the tendons snap back and push the foot off the ground, throwing the kangaroo into the air. On landing, the tendons are stretched taut again, completing the cycle.

In addition to the remarkable limb adaptations, macropods have a digestive system remarkably similar to that of the eutherian ruminant herbivores they ecologically replace. Both groups have evolved segmented stomachs, with a true chambered stomach in the ruminants and a divided, three-part stomach in macropods. This segmentation allows for the acquisition of foregut fermentation, in which symbiotic bacteria digest the tough outer cell walls of the plant materials that the animal eats, thus releasing nutrients from within the cell, providing a digestible form of cellulose (which is broken down by the bacteria), and nutrients from the bacteria themselves as they are absorbed when they pass through the gut.

Macropods have a wide geographic range and habits. While macropods are usually thought of as long-dis-
tance, open-plains grazers, in New Guinea tree kangaroos in the genus *Dendrolagus* have taken to an arboreal habit. While these animals still spend much of their time on the ground foraging for fallen fruits, they are agile climbers and use their large hind feet to propel them from tree to tree. Many of the smaller wallabies are also habitat specialists, such as the 10 species of rock wallabies (*Petrogale* sp.).

C. Placental

Placental dominate both the fossil and recent history of the Mammalia. With fewer than twice the number of orders of marsupials (18 vs 10), placental have shown remarkable diversification at all lower levels. Recent species are grouped into 18 orders with 115 families, 1049 genera, and 4375 species (Table II). As in the marsupials, some lineages contain disproportionately more families, genera, and species than others. With a fossil record stretching back into the Cretaceous, eutherian mammals have a long evolutionary history derived from marsupials and monotremes. Found on all continents, the placentals show remarkable niche diversification, occupying a wide variety of terrestrial, aerial, and aquatic habitats and subsisting on diets ranging from the true omnivorous habits required to live on the varied detritus of human populations to a single group of grasses (pandas and bamboo) or termites.

1. Xenarthra

A group found in Recent times only in the New World, xenarthrans, are thought to have occurred in Europe, although the fossil record is, at best, fragmentary. To appreciate the diversity of the Xenarthra, the sloths, anteaters, and armadillos, take a journey back into Miocene South America. Giant sloths, commonly called ground sloths or sloth bears, were as huge as their name would indicate. Numerous families of armadillos, including the Glyptodontidae, lumbering creatures up to 3 m long, occupied the landscape. While Tertiary radiations included a variety of species exploiting numerous habitats and niches, the members of the extant families of Xenarthra are relatively or highly specialized: all four species of anteaters (*Myrmecophagidae*) are obligate ant and termite eaters, using their long, sticky tongues to collect their food. The sloths (five species

---

**TABLE II**

Diversity of Recent Eutherian Mammals

<table>
<thead>
<tr>
<th>Order</th>
<th>Common name(s)</th>
<th>Number of families</th>
<th>Number of genera</th>
<th>Number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xenarthra</td>
<td>Sloths, armadillos, and anteaters</td>
<td>3</td>
<td>13</td>
<td>29</td>
</tr>
<tr>
<td>Insectivora</td>
<td>Tenreces, moles, hedgehog, shrew, mole</td>
<td>7</td>
<td>66</td>
<td>429</td>
</tr>
<tr>
<td>Scandentia</td>
<td>Tree shrews</td>
<td>1</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>Dermoptera</td>
<td>Colugo</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Chiroptera</td>
<td>Bats</td>
<td>18</td>
<td>178</td>
<td>928</td>
</tr>
<tr>
<td>Primates</td>
<td>Apes, gibbons, marmosets, lemurs, galagos, Old World and New World monkeys</td>
<td>13</td>
<td>60</td>
<td>236</td>
</tr>
<tr>
<td>Carnivora</td>
<td>Canids, felines, mongoose, hyena, seals, weasels, raccoons, otters</td>
<td>11</td>
<td>129</td>
<td>271</td>
</tr>
<tr>
<td>Cetacea</td>
<td>Whales, dolphins, and porpoises</td>
<td>10</td>
<td>41</td>
<td>78</td>
</tr>
<tr>
<td>Sirenia</td>
<td>Sea cows, manatees, dugongs</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Proboscidea</td>
<td>Elephants</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Perissodactyla</td>
<td>Equids, tapir, rhino</td>
<td>3</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>Hyracoida</td>
<td>Hyrax</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Tubulidentata</td>
<td>Aardvark</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Artiodactyla</td>
<td>Pigs, peccaries, hypos, deer, antelopes, bovids</td>
<td>10</td>
<td>81</td>
<td>220</td>
</tr>
<tr>
<td>Pholidota</td>
<td>Pangolins</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Rodentia</td>
<td>Rodents</td>
<td>29</td>
<td>443</td>
<td>2024</td>
</tr>
<tr>
<td>Lagomorpha</td>
<td>Rabbits and pikas</td>
<td>2</td>
<td>13</td>
<td>81</td>
</tr>
<tr>
<td>Macroscelida</td>
<td>Elephant shrew</td>
<td>1</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>115</td>
<td>1049</td>
<td>4370</td>
</tr>
</tbody>
</table>
in two families) have acquired an odd but apparently effective strategy, shifting along tree branches upside down, their spines hinging in a catenary curve. Sloths are the only green mammals; however, their color is not intrinsic to their fur but is derived from algae and cyanobacteria that grow in grooves in their hair. Sloths come down to the ground only to defecate. Armadillos (20 species in the family Dasypodidae) are distinguished by their jointed armor.

2. Insectivora
Modern insectivores show a wide range of adaptations and are found on every continent except Australia. Taxonomy of the insectivores has been problematic, not only because the order contains some of the most primitive eutherian mammals, as well those with highly derived characters, but also because the order was used as the equivalent of a taxonomic wastebasket in which problematic families were thrown. The 6 families and 66 genera now classified as insectivores form a monophyletic (single-root clade) with ancient origins in the late Cretaceous or Paleocene.

Two families of insectivores, Solenodontidae and Nesophontidae, were found in modern times only in the Caribbean. All of the Nesophontidae (also called West Indian shrews) are extinct. The two remaining Solenodon species are found in Haiti and, tenuously, in Cuba. Expansion of agriculture and introduction of rats, mongoose, and companion animals (dogs and cats) have led to decline of these species. Solenodon are shrewlike, have a distinctive, highly flexible snout and large hind feet, but are distinguished by an unusually large body size, weighing up to 1 kilogram. Fossil solenodons are known from North America from Oligocene deposits, approximately 30 million years B.P.

Tenrecs (family Tenrecidae) show a disjoint distribution, found on islands of the western Indian Ocean (Madagascar, Comoros) and central Africa, and have probably been linked to Africa for all of their evolution. The family can best be described as diverse. While only ranging in body size from a shrew to a small cat (maximum weight of a kilogram), tenrecs have undergone a remarkable radiation on Madagascar, expanding into vacant niches usually occupied by animals as diverse as otters, shrews, hedgehogs, and moles. Many species are omnivorous.

The family Chrysochloridae, the golden moles, is restricted to southern Africa. The 18 species in 7 genera are all fossorial and closely resemble both true moles (also insectivores, family Talpidae) and the marsupial mole, Notoryctidae. While the entire evolutionary history of the chrysochlorids is in eastern and southern Africa, true moles, or talpids, are found throughout the Northern Hemisphere and have long evolutionary histories in both Europe and North America. The talpids are more diverse than the chrysochlorids, with a range of morphotypes including the star-nosed mole (Condylura cristata), which has nearly two dozen small, fleshy fingers arranged in a starburst pattern on its nose. This represents the most bizarre elaboration of a general phenomenon found in the true moles, which have thousands of receptors arranged in a structure called the Eimer's organ on the snout. Cortical development of the brain is linked with the Eimer's organ. The Eimer's organ assists moles in navigating and foraging in their tunnel networks.

3. Scandentia
The 5 genera and 19 species of tree shrews (family Tupaiidae) are the only extant members of the Scandentia. The group is dispersed across the rain forests of southern Asia and Southeast Asia and has a relatively long fossil history (first found in the Eocene of Asia). Ressembling a small squirrel, weighing 45–50 g, tree shrews have an elongated snout and a long, bushy tail. Like true shrews, many species are known to consume insects; however, the few scientific studies that have been conducted on the family have found fruit is often the dominant component of the diet. Parental care is unusual, with the female giving birth to a small number of young (1–3) and hiding them in a nest. In a pattern more common in antelopes, the female visits the young to nurse and then leaves them lying in. Maturation is rapid, with independence from the mother at about a month and full sexual maturity at 4 months. Tree shrews are enigmatic in their phylogenetic affiliations. They have variously been assigned to the primates, flying lemurs (Dermoptera), and rabbits (Lagomorpha), with both molecular and morphological analysis producing contradictory results.

4. Dermoptera
Gliding, as opposed to true flying, is a mode of locomotion that has evolved independently in a number of mammalian orders. The two living species of Dermoptera (family Cynocephalidae), also called colugos or flying lemurs, are a relictual group of gliders found today only in the forests of Southeast Asia. Not closely related to lemurs, the dermopterans first appear in the fossil record in the Eocene of Thailand. Throughout their evolutionary history, only two families of colugos have been described, one North American and the other Asian. Affiliations of the order are unclear, with some paleontologists placing them as relatively close relatives.
to bats: they are not related to lemurs despite their common name. The gliding membrane, or patagium, stretching over the animal’s entire body, connecting along both sides of the tail to the legs, from legs to arms and from the arms to the neck, makes the animal look like a kite. Despite their relatively large size, reaching a maximum size just under 2 kg, the colugos are spectacular gliders, moving over 100 m from tree to tree in the forest to reach the trees on which they feed. The colugo diet is herbivorous but may include leaves, fruits, and flowers.

5. Chiroptera

Bats represent the second most diverse order of mammals. Two major suborders of bat have been distinguished, the Megachiroptera, also called fruit bats or flying foxes (166 species in 42 genera), and the Microchiroptera (762 species in 136 genera). Bats come as small as 1.5 g with a wingspan of 15 cm (Craseonycteridae, Kitu’s hog-nosed bat) and as large as the flying foxes in the genus Tropus (Tadaridae), which may weigh up to 1.5 kg and have wingspans of 2 m.

The two suborders have long, independent evolutionary histories, although the very nature of their morphology, and in particular their fine, light bone structure, makes preservation unusual. In contrast to modern diversity, sufficient fossil material has been found to describe just over 30 genera. The lack of fossil evidence has provoked repeated controversy and called into question whether bats are monophyletic or whether the fruit bats are actually more closely related to another group (usually primates are postulated because of shared evolution with megachiropterans of details of their neural pathways for vision). If this were the case, the remarkable flight structures would have to have evolved twice independently. Recent molecular data suggest, relatively unambiguously, that the two bat groups are monophyletic.

Bats are the only mammals capable of true flight (as opposed to gliding). The wing is formed by a thin membrane that stretches across the arms, elongated fingers, and along the body, forming a diaphanous umbrella-like structure, or patagium. Like all flying animals, bats need wings that are at once light but strong. The wings are greatly reduced in weight with muscles pulled in close to the body and bones reduced in size and volume. Torsional stress is reduced by simplification of joints. The remarkable flying abilities of bats ensure a global distribution, with bats often the only mammal naturally occurring on remote islands.

While birds monopolize diurnal aerial niches, bats are nocturnal specialists. In some areas where birds and bats do not co-occur, bats may become more diurnal, providing thin evidence of competitive exclusion. Flying at night, visual acuity is of relatively little value; nonetheless all but one species of Megachiroptera rely on vision to navigate. In contrast, microchiropteran bats have evolved navigational tools that are independent of sight. Using their larynxes, microchiropterans produce extremely high frequency sound that they emit through their nose or mouth. The sound produced by bats is referred to as ultrasonic because it is above the range that humans can hear. The ultrasound bounces off both potential prey and obstacles and is received back at the large, elaborated ears of the bat. Because the bat navigates using the sound that bounces back, this form of navigation is called echolocation.

Despite their high specific diversity, patterns of reproduction (in the few species that have been studied) all are remarkably similar. Females carry their young with them while in flight, and this constraint limits bats to producing one, or occasionally two, offspring. Paternal care of offspring is rare; hence social systems tend to be promiscuous, with few lasting bonds between males and females. Bats tend to be highly gregarious, hanging upside down in roosts containing hundreds, thousands, and in rare cases hundreds of thousands of individuals. Caves, cliffs, and the eves of large houses all provide appropriate roosting areas.

Bats are usually insectivorous, gleaning their prey while airborne, but the diversity of diets is enormous. Diet specialists cover a range of vertebrates and include species that are adapted to fish, much like a small eagle, frog eaters, blood-eating specialists (the vampire bats), nectar feeders, and two groups of specialized fruit eaters. Two families of bat dominate fruit eating: in the Old World, Pteropodidae (flying foxes), and in the New World Phyllostomidae (spear-nosed bats). In the forests in which they live, fruit-eating bats are keystone species, dispersing many of the largest forest seeds and fruits. Of course, a frugivorous diet and the large aggregations in which fruit-eating bats can occur make some bats extremely unpopular with fruit farmers around the tropics.

6. Primates

In Recent times, primates have been found on every continent and in every habitat on earth. If one excludes the most abundant and successful primate, humans (Homo sapiens), the range of the order is much reduced, with representatives found in the tropics and subtropics of Africa, Asia, and Latin America. One macaque species (Barbary macaque, Macaca sylvanus) is found in Europe on the island of Gibraltar across narrow straits
from Africa. Another macaque species, the Japanese macaque, *M. fuscata*, survives in a temperate environment by relying on hot springs during the winter.

With the first fossil representation in the late Cretaceous of North America, primates are an ancient order. The modern order is represented by 13 families, including the apes, gibbons, marmosets, lemurs, and galagos. Just over half of the familial diversity is found in the suborder Strepsirhini, or prosimians, which includes the five families of lemurs, all of which are endemic to the island of Madagascar, the Loridae (potos and lorises), and the Galagonidae (galagos), small nocturnal monkeys of Africa often called bushbabies. Some include the Tarsiridae as well.

The 30 species of lemurs show a remarkable range in size from 50 g (lesser mouse lemur, *Microcebus murinus*) to over 10 kg (*Indri, Indri indri*) and a diversity of diet from generalized herbivores, insectivores, a larve specialist (the aye-aye, the only extant species in the family Daubentoniidae), and three species of bamboo specialist (*Hapalemur* species). Many members of this suborder are under threat, victims of the rapid and seemingly irreversible deforestation of Madagascar.

While 11 species in 4 genera of galago are recognized, the specific status of the family Galagonidae is in question. Recent, and as yet unpublished, genetic data indicate that within a single species of galago there may be several cryptic species, virtually indistinguishable morphologically from one another.

The remaining primates are placed in the suborder Haplorhini, which includes the five families of New and Old World anthropoid primates and the tarsiers. Found across the forests of Southeast Asia, the tarsiers bear a strong physical resemblance to galagos and for many years were grouped with them in the prosimians. Recent genetic evidence, and a reconsideration of morphological characters, place the tarsiers firmly as distant, but distinct, relatives of the anthropoid primates.

The extant South American primates fall into two families, the Callitrichidae (tamars and marmosets, 26 species) and the Cebidae (58 species, including capuchin, squirrel, howler, spider, and night monkeys and the sakis). Marmosets and tamarins are unusual in that all but one species regularly produces twins. They have a monogamous or polygynous mating system (one female, several males), and a diet that is highly omnivorous and may include fruit, seeds, insects, and small vertebrates. Some species, including the diminutive pygmy marmoset (150 g), are exude specialists, eating the gum of trees.

The Cebidae have been described as monkeys with five legs due to the extensive use they make of their prehensile tails, an adaptation not found in Old World monkeys. Body size ranges from just over 500 g for squirrel monkeys (*Saimiri sciureus*) up to 12 kg for the woolly spider monkey (*Brachyglottis arachnoides*). The diet is predominantly fruits and other vegetation and competition for resources is common among cebids, with larger species dominating smaller ones. As a result, many smaller species have evolved diet or behavioral specialization, including nocturnal foraging (night monkey, *Aotus trivirgatus*), the ability to eat green fruit (tit monkey, *Callithrix sp.*), or habitat specialization (swamp living in *akaris*, *Cacajao sp.*).

Old World monkeys in the family Cercopithecidae dominate the primates numerically, with 81 species in 18 genera. Found across Asia and Africa, cercopithecids include animals with divergent life history strategies, sizes, and diets. The guenons, macaques, and baboons (subfamily Cercopithecinae) range in size from 1 to 50 kg and are usually found in large, extended matrilineral groups. Many species are tree-living, but the baboons and some macaques have adopted a terrestrial or semiterrestrial existence. Primarily fruit eaters, most species are omnivorous and will supplement their diet with practically anything that they can capture and eat.

The columbine monkeys, langurs, and leaf monkeys are placed in the subfamily Colobinae. A diverse group of animals, this group tends to be more arboreal and lighter than the cercopithecines. While the evolutionary history of this group is mainly African, modern representation is dominated by Asian forms (4 genera, 28 species). Diet is reflected in an unusual stomach structure, with a large, saclike upper stomach that, like the rumen of ungulates, allows for fermentation of the coarse leaves that constitute much of this group's diet.

The 11 species of gibbons (family Hylobatidae, genus *Hylobates*) are commonly called the lesser apes and are often grouped together with the true apes, or Hominidae. Found throughout tropical Asia, gibbons are renowned for their brachiating (or swinging) locomotion and their echoic loud vocalizations. The family is primarily monogamous, and its diet is primarily fruit, supplemented with young leaves and flowers.

The best studied group of primates is the family Hominidae. Formerly broken into two families, Pongidae (orangutans, gorillas, chimps, and bonobos) and Hominidae, recent data, particularly molecular genetic analyses, have made the isolation of human beings and their ancestors in their own family impossible. While largely vegetarian (gorillas are almost exclusively folivores), some chimps have been known to actively hunt other monkeys and to kill conspecifics. While most species are extremely social, orangutans tend to be solitary except when mothers are in association with young or during the mating season. Slow breeders and habitat
specialists, great apes are declining in number and distribution as a result of the rampant expansion of their con-familial Homo sapiens.

7. Carnivora

A diverse group of animals, the Carnivora share a key diagnostic feature. The carnivore dentition is typified by the evolution of a single pair of slicing teeth, the carnassials. Formed from the upper fourth premolar and the lower first molar, the carnassials are the key to understanding the diversity and plasticity of Carnivora. By isolating the shearing function to a single set of teeth, true carnivores have retained flexibility in their diet while allowing for specialization of shearing of meat. Molars retain their grinding function, while premolars in front of the carnassials are retained for gripping, crushing, or puncturing food. The canines are often elaborated into spikes of one size or another, allowing a sharp stabbing of prey. While carnassials were critical to carnivore evolution, many extant members of the order have reverted to more generalized dentition and lost the carnassials (e.g., pandas grinding bamboo).

Modern carnivores are divided into two major groups, the Feliformia, which includes hyenas, felids, mongooses, and viverrids, and the Caniformia, a diverse group that includes seals, sea lions, walruses, canids, bears, procyonids, and the mustelids.

Pinnipeds are closely related to other carnivores, but there has been a running debate as to the relationship of the three families: the eared seals (Otariidae, 14 species, 7 genera of eared seals and sea lions), the earless seals (Phocidae, 19 species and 10 genera of seals), and the walruses (Odobenidae, 1 species). For many years, these taxa were considered monophyletic, but a recent suggestion was made they are diphylectic, with walruses and sea lions branching off from a common ancestor with bears and phocids (seals) being more closely related to the mustelids. More recent molecular and morphological work confirms the monophyly of the pinnipeds.

The Canidae include 34 species of foxes, wolves, jackals, and dogs. Ranging in size from the fennec fox (Vulpes zerda, just under a kilogram) to the wolf (Canis lupus, up to 80 kg), canids show a wide range of diet and social organization. The family is best known for its pack-living, social animals (wolves, the Asian wild dog or Canis alpinus, the African wild dog Lycaon pictus, and the South American bush dog Speothos venaticus), species that raise a single litter of pups communally and that are capable of hunting animals up to 10 times their body weight through coordinated group hunting. Most species, however, live in small family groups and subsist on rodents and small vertebrates. One species, the African social bat-eared fox (Otocyon megalotis), specializes on termites. Another species, the stilt-legged South American maned wolf (Chrysocyon brachyurus), is a solitary omnivore.

Mustelids, with 65 Recent species in 25 genera, are among the most diverse of the modern carnivores. A family that includes weasels, badgers, otters, and more often than not the skunks, Mustelidae are found from the Arctic Circle to the tropical rain forests and on every continent except Australia. The group includes efficient terrestrial hunters of small mammals (mustelids, badgers) and species that are oceanic and specialize on bivalves for their diet (the sea otter, Enhydra lutris). Otters have become the dominant aquatic carnivore in many lakes and rivers across much of the globe. The giant otter, Pteronura brasiliensis, of the Amazon basin is the largest otter, weighing 30 kg, and is highly social, “herding” fish to improve success of prey capture.

Absent from South America, the nine species of badger are mostly nocturnal omnivores of stocky build and with strong, sharp claws. The Eurasian badger (Meles meles), a highly social animal that lives in extended burrows or setts, is an earthworm specialist. Weasels and polecats are found from northern Greenland to the tip of Africa. While the European weasel (Mustela nivalis) can be as small as 50 g, most weasels follow a remarkably common body form. Predominantly terrestrial and weighing 1–2 kg, with long tubular bodies and short legs, they get their speed from the compression—extension of the body; sharp teeth deliver a killing blow to a number of small vertebrates, such as frogs, rodents, rabbits, or birds. Some species once trapped for their fur are now raised in large commercial farms—e.g., mink (Mustela sp.).

The best known felid is the largest, the lion (Panthera leo). In many ways, the lion is an anomalous cat. While most of the 36 species of cat are solitary hunters, the lion is a highly social, cooperative hunter. Even congeneric large cats such as the leopard (P. pardus, Asia and Africa), the tiger (P. tigris, Asia), the snow leopard (P. uncia, central Asia), and the jaguar (P. onca, South and Central America) are more similar in their social structure and hunting patterns (solitary, territorial, with male territories overlapping those of females) to the smaller cats in the genera Felis. All cats are highly carnivorous, but diet ranges from the largest bovid species to frogs, fish, and mollusks. While most are nocturnal stealth predators, the cheetah (Acinonyx jubatus) is a true diurnal pursuit predator, capturing prey with bursts of speed of up to 90 km/h.

Viverrids (civets and genets) and mongooses (Herpestidae) are often classified as a single family, although
the most recent phylogenies separate them. Both families are exclusively Old World. Viverrids show a wide range of habits, many are truly arboreal, but there is a semi-aquatic form (otter civet, Cryptoprocta ferox). Mongooses are either solitary, nocturnal predators that resemble polecats or highly social, diurnal predators.

8. Cetacea

The recent discovery of Eocene fossil whales confirms an early branching of the Cetacea from the primitive ungulates, closely related to modern artiodactyls. The 10 families of whales, dolphins, and porpoises show a wide diversity of body size from the finless porpoise, Neophocaena phocoenoides, at 30 kg, to the world’s largest mammal, the blue whale, Balaenoptera musculus, at 150 metric tons. There is an equally large diversity of diet from tiny shrimp or krill to seals, fish, and squid—practically anything that swims. Social structure is variable and ranges from solitary individuals to the communal pods of the Orca, Orcinus Orca, or killer whale, whose social system more closely resembles that of wolves than other whales. Cetaceans share many adaptations critical to true aquatic living. The skulls are “telescoped,” with the premaxillary and maxillary bones forming the roof of the skulls and the occipital bones forming the back of the skull. In all whales the tail fin is horizontal.

Whales are divided into two suborders, the Mysticeti, or baleen whales, with 11 Recent species in 4 families, and the Odontoceti, or toothed whales, which includes 67 species in 6 families. Baleen whales, which include the aforementioned blue whale, are filter feeders. Incongruously, these leviathans live on tiny prey, from krill and small planktonic copepods, often less than 1 cm in length, to small, schooling fish. Migratory patterns across entire oceans from summer to winter feeding grounds are not uncommon, with many baleen whales being truly cosmopolitan species. Baleen whale populations, severely decreased by aggressive hunting during the nineteenth and twentieth centuries, are still in the process of recovering.

The toothed whales in the family Odontoceti are found in all oceans and all seas. Some of the smaller members of the family are also found in rivers and lakes on all continents except Australia. Like bats, the toothed whales all use echolocation, allowing them to hunt and live in turbid waters where visibility is low. While echolocation “pings” or clicks are used for locating prey, longer, more continuous tones are also important for communication among conspecifics. The toothed whales include three species of sperm whale, including the giant sperm whale (Physeter catodon, family Physeteridae) made famous by Herman Melville in Moby Dick; six species of Phocoenidae, or porpoises, found throughout the Northern Hemisphere and in near-shore environments of South America; nineteen species of the poorly known beaked whales (Ziphiidae), some of whose species have never been seen alive; and two species of highly gregarious Monodontidae, limited to the Arctic oceans, the beluga whale (Delphinapterus leucas), or white whale, and the narwhal (Monodon monoceros). The narwhal has a long, spirally grooved single upper tooth, used by males in intrasexual competition, which may provide the origin of the unicorn myth.

9. Sirenia

The four species of dugong and manatee (one species Dugong dugon and three species of manatees in the genus Trichechus) are the sole modern representatives of the Sirenia. An anomalous, nearly hairless group of mammals weighing up to 1600 kg, the common name for these animals, “sea cows,” is remarkably accurate. The only large, fully aquatic mammal that lives on grasses, the sirens have extremely long intestines. Like perissodactyls, cellulose fermentation takes place in the hindgut, enabling the sea cows to process large volumes of relatively coarse seagrass vegetation. Reproduction is both delayed and slow, with long interbirth intervals and the production of a single offspring that remains with its mother for up to 2 years.

Sirens have a near-global distribution in tropical and subtropical waters. Manatees are found up the Amazon and in major river systems of West Africa. They suffer little competition for their main resource, aquatic grasses. Yet, despite this, all species are threatened with extinction. These large mammals are docile, represent a large package of meat to human hunters, and are supposed to taste good. This, combined with slow reproduction, has led to overharvesting throughout their range. Pollution, dam building, and injury from the propellers of motorboats all amplify the problems of overhunting.

10. Proboscidea

Elephants were widely spread across North America, Asia, and Africa until recent times. Throughout the Tertiary, proboscideans showed a relatively wide range of adaptations and a diversity of species. Today, only two species remain, the African elephant (Loxodonta africana) and the Asian elephant (Elephas maximus). These hulking creatures are the largest terrestrial mammals, weighing up to 6 metric tons. While the Asian elephant is now limited to the forest of Asia, this probably reflects a historical shift from open grassland and
woodland habitat which has now almost universally been converted to human agricultural use. In Africa, elephants are found throughout the forests, woodlands, and savannas, but at highest densities in savanna woodlands. Two subspecies have been defined, the forest and savanna elephant, but many ecologists believe this reflects a more recent fragmentation of populations, with a former, clinal structure of genetic variation now lost.

In both Africa and Asia, elephants are ecological keystone species, one of the few species, other than humans, having a direct impact on the form and structure of the environment in which it lives. Elephants keep savanna ecosystems from reverting to woodland when they knock down trees. "Catastrophic" destruction of trees may occur in droughts but is clearly part of a long-term natural cycle. In Asia and Africa, the fruits and seeds eaten by elephants are transported and then defecated, seeding the plains and the forests and depositing the seeds in their own fertilizer.

Elephants, fossil and modern, have a highly specialized dentition with six cheek teeth on either side of the jaw. At any one time, one to two teeth are exposed in the jaw. As these teeth wear down, they are replaced by a tooth erupting from behind the current tooth. This pattern of sequential replacement of grinding teeth is unique in mammals. Tusks are formed from elongated upper incisors.

Highly social, female elephants live in extended, matriarchal groups. Adolescence is prolonged, with males remaining in the herd until they are 10–15 years old and females not breeding for the first time until they are 15 years old. Males are solitary, and breeding priority is determined by an interaction of sheer size (which increases indeterminately with age) and an endocrinological sexual state called musth. Found in both elephant species, musth is associated with heightened testosterone levels and makes males extremely aggressive. The length of musth increases with age, amplifying the reproductive success of older, larger males.

Elephants have been hunted for their ivory, and in several populations in Uganda and southern Africa, tusklessness in females and reduced tusk size in males have occurred as a result of repeated selection through hunting. While poaching for ivory is an immediate threat to elephants, the real threat is loss of habitat and conflict between man and elephants.

11. Perissodactyla

If you are interested in horses, tapirs, and rhinos, the mid-Tertiary was the time to be alive. While modern perissodactyls are represented by only eight species of equid in one genus (*Equus*), four species of tapir in one genus (*Tayassu*), and five species of rhinoceros in four genera, this represents but the faintest hint of the diversity of this order in the past. Of 14 Cenozoic families, only three families and five genera are represented in the modern mammalian fauna.

Modern equids all look pretty much the same: highly cursorial with a single hoof and consistent compact body, long neck, and highly hypsodont dentition in a large jaw with a distinct gap or diastema. Modern equids are only found in Africa and Asia, although there is evidence of the historical extinction of the tapan, or European wild horse. Equids went extinct in North America in the Pleistocene, but asses and horses have been reintroduced in many parts of North America, the center of evolution of the Equidae. In contrast, the Pliocene of North America was populated with a few low-crowned, three-toed equids and three-toed horses, hypsodont species of various sizes, from the small, gracile *Nam Hippus* species to the larger *Hipparion* species. The first one-toed horse, *Plihippus*, lived side by side with the hippo-like rhino *Teleoceras*, while in Asia the elephant-sized rhinoceros, *Sinotherium*, coexisted with the hippo-like rhino *Chilotherium*.

Only two forms of social organization are seen in modern equids. Those species that live in relatively arid environments, the asses and the Grevy's zebra (*Equus grevyi*), exhibit flexible associations among females and male territoriality. Females move widely in search of food and water, while males tend to aggregate around water holes, the limiting resource. Mating is promiscuous. In contrast, horses and the two remaining zebra species, the mountain zebra, *E. zebra*, and the plains zebra, *E. burchelli*, form harems of unrelated females and a single male, a social structure also seen in gorillas.

All five species of rhinoceros are threatened with extinction, the victim of the value of their horn. Two uses of rhino horn have been identified: shavings of the horn are used in traditional Asian medicine to bring down fever, while in Yemen, the highest value has been placed on dagger handles made from rhino horn. Income from a booming oil economy in the 1970s led to increased demand from Yemen, while economic growth in Asia in the 1980s and 1990s made rhino horn always expensive, accessible to many more people, thus increasing demand. The black rhino (*Diceros bicornis*) was found widely across eastern and southern Africa until the late 1970s, now all but extinct except in reserves and conservancies. The white rhino (*Ceratotherium simum*), almost extinct in South Africa at the turn of the century, has recovered to where populations number 10,000. Outside of South Africa, however, the northern subspecies is all but extinct, and introduced populations in Zimbabwe were decimated in the early
1990s. The Sumatran rhino (Diceros sumatrensis) and the Javan rhino (Rhinoceros sondaicus), formerly found across Southeast Asia, are also nearly extinct. Of Asian species, only the Indian rhinoceros (Rhinoceros unicornis) exists in any numbers, and only as the result of major conservation efforts.

12. Hyaenidae

Modern hyraxes are classified in a single family with three genera and six species. Limited to Africa, fossil hyraxes are also found in the Middle East and Europe. While modern forms are all more or less rabbit-sized (2–5 kg) and resemble a very large guinea pig, some extinct forms weighed up to 50–70 kg. Tree hyraxes (Dendrohyrax sp.) are entirely arboreal, while both bush hyraxes (Heterohyrax) and rock hyraxes (Procavia) live on talus slopes, rock outcrops, and cliffs. Affiliations with other animals are uncertain, although it appears that hyraxes represent an early branching from ungulates, perhaps sharing an early evolutionary history with elephants.

13. Tubulidentata

Found across most of sub-Saharan Africa, the nocturnal and solitary aardvark (Orycteropus afer) is the last survivor of the Tubulidentata. Never a very diverse order, the majority of the aardvark’s evolutionary history is in Africa and Europe, although aardvarks are known from Asia. Aardvarks weigh about 50 kg and resemble pigs with extremely long snouts, large ears, and long fleshy tails. Like most animals specialized for eating ants or termites, aardvarks have long, sticky tongues that are used to collect their prey. Powerful arms, with large claws, are used to excavate burrows and search out food.

14. Artiodactyla

The decline of the perissodactyls coincides with the rise of the artiodactyls. Starting in the Miocene, artiodactyls showed a remarkable diversification of families, many of which are still extant. Of 36 families found throughout the Cenozoic, 10 families and 81 genera are still found, with a near-global distribution in North and South America, Europe, Asia, and Africa. Found in all habitats, from the far north of Greenland to the southern tip of Tierra del Fuego, artiodactyls are clearly the most successful of the extant ungulates.

Diagnostic characters of artiodactyls are found in both hard and soft tissues. All members of the order have two- or four-toed hooves, with a plane of symmetry passing through the third and fourth toes. In four families (Suidae, Hippopotamidae, Tragulidae, and Tayassuidae), four fully functional digits are present, although peccaries have reduced digits on the hind limbs. All limbs have springing ligaments that capture energy when the leg flexes and return the energy to the foot as the limb pushes off, thus increasing efficiency of locomotion.

The suborder Ruminantia is represented by six families: Tragulidae, mouse deer; Giraffidae, giraffes and okapis; Moschidae, musk deer; Cervidae, deer; Antilocapridae, pronghorns; and Bovidae, cattle, sheep, goats, and antelopes. The suborder is typified by ruminant digestion, in which the rumen, an expanded first segment of a four-chambered stomach, provides an environment for the bacterial digestion of cellulose. This allows the animal to make use of highly siliceous grasses and vegetation. The cell contents, liberated by digestion in the rumen, and dead bacteria are digested farther along in the intestines. Ruminants also typically lack upper incisors, with lower incisors used to scrape grasses against the maxilla.

The tragulids, or chevrotains, are small, primitive ungulates, represented by four species in two genera, found in the forests of Africa, southern Asia, and Southeast Asia. Weighing 3–12 kg, the family lacks horns although the upper canines are unusually large and grow continually. Canines are used in intrasexual competition by males and may be used in antipredator defense by both sexes.

Cervids, or deer, are found across the Northern Hemisphere, in most of South America, and in the northern fringes of Africa. The family is absent from sub-Saharan Africa, perhaps excluded by the remarkable radiation of the antelope fauna in that region. Antlers are a defining character and are found in all but 2 of the 43 extant species. With the exception of caribou, only males carry antlers. Unlike horns, antlers are grown and shed annually. Antlers are found in a variety of forms, from spike or stubs to the elaborate branched and palmate forms seen in the moose (Alces alces). Deer are found in nearly every biome, from the snowfields of the Arctic, where they subsist on lichen, to the tropical forests of South America and Asia, where fruit is a common dietary staple.

The four species of musk deer in the Moschidae are found in central Asia, the Himalayan plateau, and east through parts of China and Vietnam. A primitive deer, the family was originally subsusted under the Cervidae. Lacking antlers, they have sharp, swordlike upper canines, similar to those found in the chevrotains. Musk is produced by a gland in its abdomen. This waxy substance is used as the base for many expensive perfumes.
and demand has led to near extirpation of all members of the family across much of its range. While captive breeding of musk deer by the Chinese might reduce demand on wild populations, prospects are not good for these animals.

Antilocaprids are only found, and have only been found, in North America. Although the family has a long and relatively diverse fossil history, only one species (Antilocapra americana) survives. The horns of the modern species are relatively simple when compared to the much more elaborated fossil forms. Unlike giraffids and all bovids, Antilocapra sheds its horn sheath annually, much like a deer losing its antlers.

The Giraffidae were more speciose during the mid-Tertiary and are represented in modern times by only two species, the giraffe, Giraffa camelopardalis, and the okapi, Okapia johnstoni. The family, exclusively Old World, has a long fossil history in Europe, where the modern genus Giraffa is first found in the Pliocene. Having evolved in Europe, giraffes dispersed to Africa. While modern forms have two, short stubby horns, fossil giraffes such as the genus Giraffokeryx had four long horns. Camel-like forms, resembling more closely the modern okapi, were also found in Europe.

By far, the most diverse of the ruminants is the family Bovidae. With 45 genera and 137 species, the family includes the greatest radiation of modern herbivores. Bovids are more widely dispersed than any other ungulate family, with a near-global distribution, absent only from Australia and the oceanic islands of Asia. But it is in Africa that the family has its greatest diversity. Members of all subfamilies are found in Africa, including 17 species of duiker (Cephalophinae, exclusively found in Africa); 35 of 40 species of antelope (Antelopinae), all 7 species of hartebeest and wildebeest (Alcelaphinae), and the 1 species of impala (Aepycerotinae). The only subfamilies of the Bovidae that are relatively poorly represented in Africa are the Bovinae (11 of 24 species: African buffalo, eland, and Tragelaphus sp.) and the Caprinae (4 species in North Africa, none south of the Sahara).

The suids, peccaries, and hippos make up the last suborder of Artiodactyla, the Suiformes. Despite their global distribution, wild pigs and boars (family Suidae) are an Old World family, with 9 genera and 23 species introduced into North and South America, Australia, Tasmania, and New Guinea. They range in size from the pygmy hog (Sus salvania) of the Himalayan foothills (9 kg) to the giant forest hog (Hylochoerus meinertzhageni) of central and East Africa (275 kg). With 16 species in 5 genera, there are relatively few modern pig species, all of which closely resemble the domestic hog.

Omnivores, pigs usually are forest dwellers, although the warthog (Phacochoerus aethiopicus) is found throughout the savanna woodlands and grasslands of Africa. Lacking horns, the upper canines of most pigs are ever-growing and form large, slashing tusks. This adaptation has gone to its extreme in the island endemic, the babirousa (Babyrousa babyrussa) of Sulawesi, Indonesia, whose tusks curl up over the head.

Closely related to the pigs are the three monotypic genera that represent the extant peccaries (family Tayassuidae). Thought to be derived from Old World pigs and with a fossil record in the Old World and New World, the peccaries are South America's pig equivalent. Smaller than pigs (20-40 kg), peccaries are critical frugivores and seed eaters thought to be important for seed dispersal in South America's tropical forests.

The third family in the suborder Suiformes includes the two species of extant hippos, the hippopotamus (Hippopotamus amphibius), found in Africa from the Nile basin south throughout West, East, central, and southern Africa, and the pygmy hippo (Choeropsis liberiensis), found only in a small section of West Africa. Recent fossil hippos are known from Madagascar, and in the Miocene, the family was spread throughout Africa and Asia. The hippo is a nocturnal grazer, leaving rivers and lakes at night to forage along their banks.

The first fossil remains of the Camelidae, the family that includes camels and llamas, are in the Old World, although by the Eocene there are camels in the fossil record of the Americas. Former diversity was much greater than that of the present, where three species in the family Llama and one Vicugna species are found in South America, and two Camelus are found in the Old World: the dromedary (C. dromedarius, southwest Asia and North Africa) and the Bactrian, or two-humped, camel (C. bactrianus, Mongolian steppe). Camels are highly adapted for desert life. The feet are splayed to prevent the large (650 kg) animals from sinking into the sand, a reversal of the usual ungulate simplification of foot structure that is evident in the South American camels and most fossil forms. Water conservation is achieved through well-insulated bodies, a high tolerance for dehydration, dry feces and concentrated urine, and metabolic conversion of fat (stored in the hump).

15. Pholidota

Pangolins, also called scaly anteaters, are not very diverse—seven species in a single genus, Manis—yet they have a wide geographical range, occurring in a variety of habitats across southern Africa and in much of Southeast Asia. The order is also found in the Tertiary fossil
record of all continents except Australia and Antarctica. Resembling a cross between an armadillo and an ant-eater, pangolins are large (5–35 kg) and both ecologically and culturally important. The scales, which cover the body completely, may constitute up to half the animal’s body weight and are prized for their medicinal value across Asia, although no evidence exists that the scales, which are composed, like rhinoceros, of agglutinated hair, have any therapeutic value. Pangolins have the distinction of being the only extant mammals with no teeth.

16. Rodentia

The diversity of the order Rodentia deserves a volume of this encyclopedia of its own. With nearly half of all mammalian species (2024), over 400 genera, and 29 families, rodents are the most successful group of mammals living today. One family alone, the Muridae, includes two-thirds of the living species (hence, one-third of all mammals) and is subdivided into 17 subfamilies. The order includes rats, mice, squirrels, guinea pigs, beavers, kangaroo rats, dormice, jerboas or jumping mice, hamsters, mole rats, porcupines, chinchillas, agouit, and nutria. Rodents are used by humans for food (e.g., guinea pigs), fur (nutria, beaver, chinchilla), and pets. Most rodents weigh about 100 g and have relatively similar body plans: long nose, large eyes, and long tails. Size varies four orders of magnitude, however, with the smallest member of the order weighing about 5 g (pygmy mouse, Balomys species of Central America) while the largest, the capybara (Hydrochoerus hydrochaeris), may weigh up to 60 kg and is found near water across the northeast of South America.

Rodents are an ancient order of mammals, found in the Paleocene of both North America and Asia, indicating an earlier first occurrence. Rodents have always been successful, and myriad fossil forms are found in the Eocene of Asia and North America. Modern rodents are usually divided into three major groups: the cavy-like rodents, or Caviomorpha; the mouse-like rodents, the Myomorpha; and the squirrel-like rodents, the Sciuroterida. While a convenient division, both paleontological and molecular data suggest that this tripartite classification may be more useful than it is real.

While supporting a phenomenal diversity in both Recent species and fossil forms, all rodents have adopted a similar jaw structure, diagnostic of the order. The incisors are used for gnawing and clipping and are sharp and ever-growing. Molars are adapted for grinding. The jaw musculature is also modified, with the muscles moving some of their attachments off the zygomatic arch and cranium forward onto the rostrum, with chewing thus providing a forward movement of the jaw.

Rodents have a global distribution and are found in every habitat type from the high Arctic to the driest deserts and the wettest tropical forests. Species like the Norway rat, Rattus norvegicus, have hitched rides on ocean-going vessels since people started sailing the seas 10,000 years ago. A crop pest, the Norway rat is responsible for billions of dollars of damage each year. Of greater historical significance, rats have acted as secondary hosts for a number of diseases that plague humans, including the bubonic plague itself, which killed 30 million people in Europe from the fourteenth to seventeenth centuries. Rats have also been one of the main agents of island extinctions.

17. Lagomorpha

Rabbits (Leporidae) and pikas (Ochotonidae), while not diverse when compared to the rodents (80 species in 13 Recent genera), are nearly worldwide in their distribution. Found in both the New World and Old World, rabbits naturally occur on all continents except Australia, where they have been introduced, as they have been to many larger islands around the world. Fossil remains are first found in Paleocene China, and they occur in a wide variety of habitats, from the Arctic snowshoe hare (Lepus americanus) to the tropical species of the genus Nesologus (the Sumatran and Annamite rabbits). Rabbits are herbivores, with hypsodont molars, and, like rodents, ever-growing incisors are used to clip vegetation. Body size varies over a relatively narrow range (300 g to 5 kg), and body form is consistent: round heads, large eyes, big ears, and extended, rictaloteal hind legs which converge on those of macropod marsupials. Pikas, which are smaller than rabbits (approximately 200 g), closely resemble small caviomorph rodents and are most commonly found on rocky outcrops and talus slopes. Pikas are found in North America along the northwest coast and across the central Asian steppe into Russia.

18. Macroscelidea

The taxonomic position of the elephant shrews is unclear. Similar to kangaroo rats with long noses, with only 15 living species and a poor representation in the fossil record, the order is usually grouped with the lagomorphs and rodents, although earlier taxonomies have suggested an affiliation with insectivores. Found only in Africa, the order has a narrow range of body size, from the tiny 45-g short-eared elephant shrew (Macroscelides proboscideus) to the rather larger 500-g golden-rumped elephant shrew (Rhynchoecyon
chrysopteryx). Absent in West Africa, elephant shrews occupy a diversity of habitats, from the deserts of Namibia to the lowland forests of central Africa. Despite the diversity of habitats occupied, all species are monogamous, terrestrial, and omnivorous, although insects form a large part of their diets where studied.

IV. EVOLUTIONARY TRENDS

A. Evolution of Brain Size

The defining characters of hominid evolution are an upright gait and an increasing large and complex brain. Humanoids have the largest brain for their body size of any mammal, extant or extinct. Brain size has, not surprisingly, been correlated with intelligence, with other relatively large-brained forms (e.g., dolphins and great apes) being imputed to be more intelligent than the smaller brained mammals. If larger brains do, indeed, confer greater intelligence and survival of an individual is in some way correlated with intelligence, then one would predict strong selection for increasing brain size in mammalian lineages.

At some point in each of their fossil histories, progressive increase in brain size is observed in primates, cetaceans, carnivores, and ungulates. However, increasing brain size has not been a linear effect, but has been punctuated with periods of rapid increase followed by stasis or relatively slow rates of change. Rapid evolution of brain size occurred early in primates, with modern prosimian brain size occurring by the late Eocene. In anthropoids, long held as the best example of progressive evolution of brain size, a rapid increase in brain size in the Oligocene was followed by relative stasis in most lineages, with the exception of the hominid line. Because carnivores, ungulates, primates, and whales have some of the largest brains and are also among the more charismatic species, generalization about the occurrence of progressive brain size evolution in mammals through time may result from a certain large mammal myopathy.

In many lineages of mammal there has been no such progressive increase in brain size—marsupials, edentates, and some lineages of rodents have shown little change in relative brain size since the orders are first seen in the fossil record in the early Tertiary. This stasis is often attributed to differences in predator pressure, with Miocene South American and Australian marsupials not requiring “higher” brain function and hence there having been no evolutionary pressure on brain size. This explanation is unsatisfactory given the high variation in when brain size evolution occurs in different mammalian lineages.

Because brains tend not to shrink in evolutionary time, the variance in brain size will increase with time, and hence any increase in variance, with size bounded at the lower end of a distribution, will result in a larger average brain size. If selection in evolutionary time for increased brain size in different lineages reflects modern patterns of brain size variation, then examining the ecological and social correlates of brain size variation may give us an insight into the selective forces that may have shaped brain size. For instance, growth of the cerebellum has been correlated with locomotion in three dimensions (flight, swimming) as compared to terrestrial motion. Elaboration of the neocortex has been associated with various aspects of learning. Taxa that have prolonged periods of maternal dependence, and presumably long periods of information transfer, have relatively greater neocortex development than taxa with minimal parental association. Within lineages, in monogamous species males and females have similar brain structure, while in those species where males are promiscuous, there is elaboration of the hippocampus, the part of the brain correlated with spatial memory. The hypothesis is that in searching for receptive females, promiscuous males search over large areas, thus requiring greater spatial skills.

B. Cope’s Rule

One of the earliest rules applied to mammalogy is Cope’s Rule: species within a lineage will show increasingly large body size through evolutionary time. Cope made his findings in the late nineteenth century working on North American fossil assemblages (some of the best preserved of the Tertiary fossil record), yet most studies since Cope have failed to find support for this generalization. Some have argued that because most lineages originate at small body sizes, it is axiomatic that the only place they have to go is up. Others have stated that the pattern seen in the fossil record is a statistical artifact, the result of passive diversification of species within clades rather than the result of any kind of directed evolution. Recently, it was even suggested that while Cope was an advocate of directed evolution, he did not even do any significant analysis of body size trends and that the attribution of the first observation of this phenomenon is misplaced.

As with all paleontological examination, the detection of such trends relies on the quality of the data set used. Most studies of Cope’s law have suffered from either a telescoped time frame or data sets that cover
a long time run, but focus on a small number of taxa. But a recent study of Cope’s law, which used a data set of 1534 species of North American mammals ranging in age from the late Cretaceous to the Pleistocene, suffers from neither of these faults. Species within a genus were followed through their evolutionary histories, and, on average, new species were 9% larger than the older species within the same genus. Diversification in size was not gradual, but changed rapidly at the Cretaceous/Tertiary (K/T) boundary, coincident with the rapid ordinal diversification of mammals. Average body weight of 29 species in the late Cretaceous was 150 g; by the early Tertiary, the average weight of 33 matched species was just over 1000 g, an order of magnitude higher. There is an unambiguous directional trend to larger size in North American mammals, but what could possibly explain this trend?

While data convincingly show that Cope’s law is valid, at least for North American taxa, no good explanation has been offered as to why such persistent increases in body size are observed. Not only does body size increase within lineages, but at some point the middle of the size distribution drops out, leaving relatively large and relatively small species. This has led some to suggest that there are, perhaps, optimal body sizes for homeothermic mammals. Scientists have noted in modern assemblages of mammals that there are a disproportionately large number of species that weigh about 100 g (coincident with the average weight of Cretaceous mammals) and that this might be a lower end optimal body size for mammals. This hypothesis is supported by data on the relationship between minimum home range and body size. While home range size scales roughly with body size in mammals, the relationship is U-shaped, with smaller mammals requiring relatively larger areas than expected. An “optimal” body size is one in which a minimum area is required to support the activities of an individual (the bottom of the U-shaped curve): this minimum occurs between 80 and 200 g.

Upper end optima are not well defined, and gaps in the body size distributions of extant North American mammals are more illusory than real and are easily explained by random statistical variation. If Mesozoic and Cretaceous mammals were competitively excluded by dominant terrestrial vertebrates, the extinction of the dinosaur fauna at the K/T boundary may have opened up the larger end of the body size distribution, allowing mammals to evolve to larger sizes. If larger mammals show higher rates of extinction or lower rates of origination, gaps will develop in the upper end of the mammalian size distribution, perhaps mimicking the random body size gap distribution observed in an ecological time scale and allowing new taxa to continue to evolve to larger sizes.

C. The Island Rule

When mammals colonize islands, a strange thing happens: small mammals, such as rodents, tend to increase in size, while larger mammals, such as carnivores, lago- morphs, and artiodactyls, tend to become smaller. Numerous examples of dwarfism and giantism have been found in the fossil record and can be documented in extant species, the most notable being the discovery of fossil evidence of pygmy elephants on Mediterranean islands and fossil remains of dwarf mammoth on the Channel Islands of California.

Explanations for this Island Rule have been numerous, and none is completely convincing, but for the most part they focus on a combination of resource limitation, reduced competition as a result of a depauperate fauna, and predator release. For larger species, a limited food supply would favor a smaller body size as individuals of smaller body size require fewer resources, can reproduce more efficiently, and are more likely to leave surviving offspring. Niche partitioning through character displacement will result in selection for the smaller competitor in a feeding guild to become yet smaller. Therefore, it is not surprising that the absence of the larger competitor results in an observed larger body size in the smaller species of a feeding guild. Predation has been hypothesized to have two different effects: in larger mammals, where size is a form of predator defense, release from predators may select for reduced body size. This argument, of course, is confounded with the explanations provided by resource limitation arguments. For smaller mammals, where stealth and crypsis are defense strategies, absence of predators may reduce the adaptive value of small size, thus allowing small mammals to evolve to larger sizes.

A combination of bioenergetics and hypotheses put forth for Cope’s Rule may provide a somewhat more synthetic argument for the Island Rule. Studies of mammalian feeding guilds suggest that within a guild of large-bodied mammals, the individuals of smaller species frequently monopolize a greater proportion of the available resources. In guilds dominated by smaller mammals the opposite appears to be true: individuals of larger species within the guild appear to control a greater proportion of the available energy. If there is an optimum size for mammals, arrival on an island in which there was release from predators and guild
competitors would allow rapid evolution toward that optimum. Animals that were significantly larger, or smaller, than the optimum would show the greatest divergence between insular and continental body size.

D. These Legs Are for Walking: Predators and Their Prey

The evolution of ungulate locomotion, as typified by the evolution of the equid leg, is a topic covered by every high-school textbook. Cursorial specialization has evolved independently in a number of mammalian lineages, but there is a general pattern of morphological changes that results from a simple calculation: for an animal to move faster, it must increase either the length of its stride or the number of strides it takes. Hence, shifts in morphology must make the leg longer, faster, or both. The lateral reduction and fusing of bones (especially the hand and foot bones, the metacarpals, and metatarsals), the elongation of limbs, the reduction or loss of the calcaneus, and the shifting of muscle mass toward the torso with a concomitant increase in the use of tendons to move the limbs are all part of the suite of changes which improve running speed.

Like their herbivore prey, carnivores have also evolved increasingly long-legged, faster forms through time. Carnivores have gone from being short-legged, small creatures to the long-limbed, slender-bodied forms typified by modern pursuit predators such as the cheetah or wolf. These forms appear to have relatively short evolutionary histories, with repeated evolution of similar body types in different carnivore lineages.

The co-evolution of faster herbivores and faster predators has been described as a predator–prey arms race: slight increases in prey speed lead to higher rates of survival of those swift individuals who pass on their genes at disproportionate frequencies. Similarly, faster predators have greater hunting success, produce more offspring, and are better represented in successive generations. There is only one problem with this scenario: the fossil data do not support the hypothesis. Ungulates have tended to evolve ever more cursorial forms throughout the Cenozoic, beginning 55 million years ago in the Eocene. But is was not until the Pliocene, only 5 million years ago, that carnivores “caught up.” By this time, most of the “novel” ungulate adaptations for speed were tens of millions of years old.

Long legs confer the ability to move quickly, but what if this ability were a secondary adaptation, the result of selection for another phenomenon? An alternative explanation for cursorial adaptation suggests that the suite of adaptations that has evolved in ungulates has to do with the energetics of movement rather than the energetics of being eaten. All the observed adaptations do one of two things—increase stride length or reduce the costs of moving the leg. Hence, long legs lead to increasing efficiency of movement, whether the animal is walking 10 km to get to a water hole or dashing to get away from a predator. If the opening of the savanna led to increased distance movement in ranging patterns, long, simplified, lightweight legs would confer a great advantage in terms of energy efficiency of movement, whether daily or migratory.

E. Convergent Evolution

Mammalian evolution has been rich with novelty, whether in the evolution of flight in bats, the suite of adaptations evolved by whales that have brought mammals into the sea to live, or more specifically in one lineage the well-developed venom gland found in the hind legs of the platypus. Throughout mammalian evolutionary history, however, some of the more remarkable evolutionary patterns have involved the repeated evolution of derived characters in widely divergent mammal lineages. Convergent evolution spans a wide array of adaptations, from the suite of carnivorous Miocene South American marsupials that resemble extinct and contemporaneous placental carnivores to the striking similarity of burrowing forms in the marsupial mole (family Notoryctidae), the golden moles (family Chrysochloridae), and true moles (family Talpidae).

Throughout their evolutionary history, mammals have shown trends to increasing specialization in many morphological and correlated ecological functions. For instance, as grasslands became widespread in the Miocene, ungulate tooth morphology shifted from a dominance of low-crowned, or brachyodont, teeth, which are easily ground down through the animal’s lifetime, to a dominance of hypsodonty, or high-crowned teeth, which last longer when an animal eats siliceous grasses covered in dust and dirt. This evolution has been complemented by an increased complexity of molar teeth to enable herbivores to more thoroughly grind their food. This pattern is observed in both artiodactyls and perissodactyls. Rodents have solved the problem differently, evolving ever-growing teeth, while proboscideans roll their teeth out of their jaw, each tooth having a limited useful life.

Convergent evolution may address the same problems but find different solutions. The expansion of grasslands and a high-cellulose diet have led to two different solutions to digesting an essentially indigest-
ible substance. Ruminant digestion, widespread in the
tartiodactyls, uses the fore-gut, or rumen, as a fermenta-
tion chamber in which bacteria break down cellulose,
making the grass cell contents available for absorption
and converting the cellulose into digestible material.
Because the passage of materials through the rumen
limits intake, tartiodactyls tend to be relatively selective,
choosing grasses of high quality. In perissodactyls and
some kangaroos, a similar solution has evolved—
bacterial breakdown of cellulose—but the site of bacte-
rial digestion is in the hindgut and the process is less
efficient. While this means that hindgut fermenters can-
don’t draw out as many nutrients from a given pulse of
food, by rapid processing of food, hindgut fermenters
can gain sustenance from larger volumes of lower qual-
ity forage.

Table III lists a few of the more widespread patterns
of convergent evolution in mammals.

V. GEOGRAPHY, BIOGEOGRAPHY,
AND BIODIVERSITY

A. Gradients of Species Richness:
Diversity, Density, Range, and
Rapoport’s Rule

The tendency for species richness to increase with de-
creasing latitude was first observed by Alfred Russel
Wallace and has been the subject of significant study.
Mammals, as well as nearly every taxa studied, appear
to show this pattern. While nearly a dozen hypotheses
have been suggested to explain the phenomenon, no
single explanation appears dominant.

In a study of nearly 200 North and Central American
mammals which matched subspecies living at lower
and higher latitudes, it was found that those living at
lower latitudes had significantly smaller overall latitudi-
nal geographical distributions. The same pattern is ob-
served in a review of 679 North American mammals in
which latitude is compared to the range of the species.
In this study, in addition to the increase in range size
with increasing latitude, a similar pattern is observed
in a west-to-east gradient of longitude, with range sizes
increasing to the east. Between the tropics and the
northern Arctic regions, the average range of a species
increases by a factor of 30. Interestingly, this same
pattern is not observed in Australian mammals. The
continent has a divergent geographic structure, with an
arid center fringed by moist mountains, creating habitat
diversity (and stability) very different from that ob-
served in northern continents.

The vast majority of North American mammals have
a narrow geographical species range that covers but a
few habitat types and appear to be habitat specialists.
Not surprisingly, patterns of variation are not consistent
across orders, with the ranges of carnivores and tart-
ioidactyls being larger, on average, and bats and rodents
much smaller. Nonetheless, the observed pattern of in-
creasing range with increasing latitude is not an artifact
of taxonomy: these patterns are seen both across all
mammals and independently within mammalian orders.

Overall, range size shows a log-normal distribution,
with very large and very small ranges being exceptional.
Species with small ranges would probably result in a
greater risk of extinction through severe catastrophic
events such as abnormally cold, hot, or dry weather,
through disease, or because of predation or competi-
tion. Species with unusually large areas may be prone
to extinction because of the low densities at which
they occur.

It has long been noted that species diversity (or more
accurately for mammals, species density, the number of
species per unit area) decreases with increasing latitude.
But species density also has a strong trend, with species
density increasing from east to west. The increase in
density is greatest in the more southerly latitudes (30–
40°) and correlates with increasingly complex topogra-
phy, and hence greater habitat diversity.

Is there a connection between these two patterns,
with narrowly distributed species packed in more
closely at southern latitudes and in more complex habi-
tats? Selection for wide tolerances of temperature and
moisture extremes is more likely to occur in the tempera-
ture zones, where temperatures can range from freezing
or subfreezing in the winter to 35°C in the summer.
Patterns of climatic variation are similar over large
areas; hence moving from place to place does not usu-
ally expose an individual to temperature or moisture
extremes that have not previously been experienced.
At more northern latitudes, the range of extremes of
moisture, temperature, and light become more extreme,
and unless a species is migratory, an individual will be
exposed to these extremes year after year throughout
its lifetime.

While wide climatic tolerance might have few costs,
it would confer few benefits in the tropics. In the tropics
the range of climatic variation at any given location
is low. Microhabitat variation operates over a smaller
spatial scale; hence staying in one place does not expose
an individual to climatic variation, but moving a few
tens of kilometers, or a few hundred meters in elevation,
can expose an individual to conditions to which the
individual, or its ancestors, have not been exposed.
<table>
<thead>
<tr>
<th>Adaptation</th>
<th>Holartic</th>
<th>Madagascar</th>
<th>South America</th>
<th>Australia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carnivory</td>
<td>Carnivora</td>
<td>Carnivora</td>
<td>Berhysanidae*</td>
<td>Dasyuroid</td>
</tr>
<tr>
<td></td>
<td>Many families</td>
<td>Many families</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pursuit predators</td>
<td>Carnivora</td>
<td>Viverrids</td>
<td>Didelphoid</td>
<td>Dasyuroid</td>
</tr>
<tr>
<td></td>
<td>Dogs</td>
<td>Hyenas</td>
<td>Berhysanid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dog bears*</td>
<td>Cheetah</td>
<td>Thylacinus*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bear dogs*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saber-tooth carnivores</td>
<td>Smilodon*</td>
<td>Felids</td>
<td>Berhysenid*</td>
<td>Phalangeroid</td>
</tr>
<tr>
<td></td>
<td>Felids</td>
<td>Viverrids</td>
<td>Fossa</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nimravids*</td>
<td></td>
<td>Thylacosmilus*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thalacoleo*</td>
<td></td>
</tr>
<tr>
<td>Semiaquatic</td>
<td>Insectivora</td>
<td>Insectivora</td>
<td>Didelphoid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Desmans</td>
<td>Otter shrew</td>
<td>Water opossum</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Water voles</td>
<td>Otter teurec</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pantolessids*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rodentia</td>
</tr>
<tr>
<td></td>
<td>Rodentia</td>
<td></td>
<td></td>
<td>Water rat</td>
</tr>
<tr>
<td>Aquatic</td>
<td>Carnivora</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Otters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herbivores with cellulose fermentation</td>
<td>Ungulates</td>
<td>Ungulates</td>
<td>Ungulates</td>
<td>Phalangeroids</td>
</tr>
<tr>
<td>Artiodactyla</td>
<td>Proboscidians</td>
<td></td>
<td>Notoungulates*</td>
<td></td>
</tr>
<tr>
<td>Perissodactyla</td>
<td>Hyrictoids</td>
<td></td>
<td>Lipotermes*</td>
<td></td>
</tr>
<tr>
<td>Some rodents</td>
<td>Arsinotheres*</td>
<td></td>
<td>Pyrotheres*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Some rodents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hornlike structures</td>
<td>Perissodactyls</td>
<td>Arsinotheres*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhinos</td>
<td></td>
<td>Notoungulates*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broniotheres*</td>
<td></td>
<td>Toxodonida*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artiodactyls</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procervividae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oreodonta*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pigs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ruminants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artic ungulates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unthaitheres*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rodents</td>
<td>Mylagaulida*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anteaters</td>
<td>Pholidota</td>
<td>Pholidota</td>
<td>Edentates</td>
<td>Dayroids</td>
</tr>
<tr>
<td>Pangolins</td>
<td>Pangolins</td>
<td></td>
<td>Anteaters</td>
<td>Numbat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Armadillos</td>
<td>Monotrems</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Echidna</td>
</tr>
<tr>
<td>Gliding</td>
<td>Rodents</td>
<td>Rodents</td>
<td>Phalangers</td>
<td>Flying possums</td>
</tr>
<tr>
<td>Flying squirrels</td>
<td>Rodents</td>
<td>Sealy-tailed squirrels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dermopteryans</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flying lemur</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paromomyidae*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This has prompted one scientist to note that "mountain passes are higher in the tropics." The physical structure of such passes is, of course, no different in the tropics. But the perception of a dispersing animal trying to cross that pass may vary widely, depending in large part on the way adaptation for extremes in temporal and spatial on the way adaptation for extremes in temporal variation of ecological variables has equipped the individual to deal with new microclimates encountered on its journey.

**B. Hot Spots of Species Richness and Endemicity**

Broader patterns of species diversity and species density may explain large-scale geographic variation in observed diversity of mammals, but within these patterns there are finer grained anomalies in distribution that result in areas with extremely high diversity, so-called hot spots. Across all species of terrestrial vertebrates, 35% of all species are confined to a total of 1.4% of the Earth's land surface. Almost by definition, this includes a large proportion of geographic endemics, and not surprisingly, such hot spots tend to be much more common in the tropics than in temperate areas for reasons obvious from the above discussion of geographical patterns of diversity. This pattern has strong implications for priority setting in conservation of mammals in the tropics, with a greater bang for the buck (as measured by species protected for dollar expended) if conservation investments are focused in these hot spot areas.

The geographical and geological precursors that lead to high levels of species endemicity are not well understood. Despite the radical difference in life histories and evolutionary histories, patterns of endemicity frequently are highly congruent across taxa, indicating necessary and sufficient conditions are probably common to the evolution of endemicity in mammals and other taxa. Indeed, in Africa and its associated islands, endemicity of plants and animals converges, with the hottest of the hot spots for all taxa being found in places like Madagascar, the Karoo, the Cape Province of South Africa, and the Ethiopian Highlands.

The diversification of a small number of families on islands is probably as much due to vicariant or random arrival of animals on the island as it is to any set of geological or climatic conditions. The diverse lemur fauna of Madagascar and the rapid radiation of marsupials on the Australian continent owe as much to isolation as they do to anything else. The rapid, and astonishing, diversification of cichlid fishes in the rift lakes of Africa or the convergent pattern seen in the fish genus Semio-

_**VI. NEW DISCOVERIES**_

**A. Patterns of Discovery**

The discovery of a new mammal, even a large mammal, is not such an unusual event. In the past 60 years,
17 species of large mammals have been described and several others are known but await formal description. Consideration of the pattern of description of new species of mammals shows that the curve is more or less bell shaped (Fig. 4), with the peak of activity in describing mammals coming early in the twentieth century. While new mammals continue to be discovered each year, the rate at which these discoveries have occurred has not changed significantly in recent decades.

If finding new species is not a particularly rare event, can we predict how many species will eventually be discovered? Recent analysis of the discovery and description of large (>2 m) open-water marine mammals suggests perhaps that 47 species await formal scientific description, that we should continue finding species at a rate of one new species every 5 years, and that we are most likely to find new cetaceans. Given the long and intensive history of commercial whaling, such a finding is surprising.

New species can be named in one of several ways. In some cases, specimens collected many decades earlier were neither properly examined nor adequately described. Hence, when a systematist revises a taxon, "new" species may be found sitting in a museum drawer. Similarly, new forms of evidence from field, laboratory, or genetic studies may force a revision of a species group. Molecular genetics has played a large role in this field, forcing scientists to reassess previously assigned taxonomic names. Finally, previously unknown forms may be found in nature.

B. The Annamites: The Last Frontier for Large Mammals

The Annamite mountain range, which forms the border between the Lao People’s Democratic Republic and Vietnam, has yielded a large number of discoveries of new species and rediscoveries of previously described species of large mammals in the past decade. The Vietnamese warty pig, Sus bucculentus, long thought to be extinct, was found in a food market in Laos in the early 1990s. The field has been particularly rich for discovery of barking deer, primitive deer found throughout the tropics of southern Asia and Southeast Asia. In 1994, the giant muntjac was discovered in the forests of the Annamites. Large in size and with unusual antlers, the species was thought sufficiently morphologically distinct to deserve its own genus, Megamuntiacus. DNA analysis suggested that the species belongs with other muntjacs in the genus Muntiacus (M. vaquangensis) and not in a genus of its own. Two further Annamite Muntiacus species have been found, one a new species from Quang Nam Province, Vietnam (Muntiacus truongsonensis), and another from the Lao side of the border, a rediscovery of Roosevelt’s muntjac, M. rooseveltorum. In neighboring Myanmar (Burma), a 1997 expedition
to the far north of the country, an area not visited by scientists since the 1930s, yielded discovery of the world’s smallest deer, another species of muntjac, the leaf deer (M. putaoensis).

The Annamites have also been the source of two discoveries that have both puzzled and captivated scientists. The first is the saola, or Pseudoryx nghetinhensis. The species, which weighs approximately 100 kg, derives both its scientific name and Lao common name from the shape of its horns: long and arced, the species’ horns resemble both the Africa oryx and the arc of a lyre-like Lao musical instrument. First found in Vu Quang forest reserve, Vietnam, the species is now known to range widely at higher altitudes in the Annamites. Most closely related to bovids, the saola is clearly unusual and represents a deep branch in the phylogeny of the bovids.

Another, more curious, Annamite discovery was that of a new rabbit species (Nesolagus timminsi). The rabbit, which has distinctive, dark brown stripes running down both its face and back, a reddish rump, and short ears, was first identified in a food market in Ben Lak, Laos, but since has been photographed in Vietnam. The rabbit’s closest relative is a critically endangered species found in Sumatra, about a thousand miles away (Nesolagus netscheri). Despite extreme morphological similarity, genetic data suggest that the two species may have diverged about 8 million years ago. The two are so divergent genetically that it was debated as to whether the new (as yet unnamed) species should be placed in its own genus.

C. Cryptic Species, Phylogeography, and Evolutionary Significant Units

New discoveries do not require going to the ends of the earth. The pipistrelle (Pipistrellus pipistrellus) is one of the most common bats in Europe, and one of the best studied. In the British Isles it was found that the bat used two frequencies to echolocate, which led scientists to classify these populations as either the 45- or 55-kHz phonic type. Because roosts consisted of bats of a single phonic type, to these scientists it was suggested that the phonic types represented sympatric, but distinct, species. Because of their morphological similarity, they are called cryptic species. Two recent studies have provided further evidence that these two phonic types are, indeed, good species. The two show relatively good separation of diet: while both eat mostly the dip- teran suborder Nematocera, the dominant prey groups for the 45-kHz phonic type were in the families Psychodidae, Anisopodidae, and Muscidae, while the families Chironomidae and Ceratopogonidae were the main prey groups of bats of the 55-kHz phonic type. Studies of the cytochrome b gene also showed unambiguously that the two phonic types are distinct species. That such cryptic species can be found in such a well-studied group suggests that further genetic studies of mammalian phylogenies will yield some surprises.

Molecular genetics can also be a powerful tool for disentangling the relationship between the geographical and morphological history of subspecies within a species group, the study of phylogeography. Because morphological, geographic, and genetic data frequently tell different stories (as in the striped rabbit example above), an understanding and reconciliation of these data are critical to understanding both the evolution and conservation of mammals. The recent study of Australia’s largest extant carnivore, the marsupial tiger quoll (Dasyurus maculatus), is instructive. The quoll occurs on Tasmania and on the mainland in two distinct populations, one in the north and one in the south. Previous to genetic analyses, strong body size and morphological convergence led to the southern and Tasmanian populations being grouped as one subspecies (D. m. maculatus), while the northern population was classified as a separate subspecies (D. m. gracilis).

Genetic evidence suggests, however, that despite their size differences, which may be as much as 50%, the two mainland populations are more closely related to one another, while the Tasmanian population is genetically divergent. Why does this matter? If the two mainland populations constitute one evolutionary significant unit (ESU) while the Tasmanian population constitutes a different ESU, then one management solution would be to manage the mainland and island populations separately. This would ease management concerns as the northern population is classified as endangered, while the southern population is relatively more common. Yet, it is likely that if a female northern quoll were placed in a breeding center with a southern male, because of their great size difference, the male would view the female as a dinner, not as a potential mate. Similarly, translocation of southern quolls to the north to augment that population could be disastrous. Hence, while phylogeography suggests one management regime, a more complex solution, with each population managed separately, is recommended.

VII. CURRENT EXTINCTION CRISIS

A. How Many Species Have Gone Extinct?

The current rate of extinction of mammals is, by any measure, frightening. In the past 500 years, approxi-
Mammals, biodiversity of

Approximately 88 species of mammals are thought to have gone extinct. This represents approximately 1.9% of the extant species of mammals. Using the “background” or natural rate of extinction derived from an examination of the fossil record, one would have expected at most one species of mammal to have gone extinct in the same time period. This accelerated rate of extinction can be ascribed, directly or indirectly, to a single cause: humans. While determining what factor is ultimately (as opposed to proximately) involved in a species extinction, habitat loss, introduction of exotic species (including zoonotic diseases), and overharvesting all have contributed to the high historic rates of mammalian extinction.

One would think that assessing the rate of extinction in such a well-studied group as mammals would be relatively easy. Yet, despite relatively good taxonomy and detailed study of the question, there remains considerable debate among mammalogists and conservationists about the number, the precise identity, and the timing of disappearance of those species that may have gone extinct in historical times. Two recent efforts to categorize historical extinctions came up with approximately the same number of species going extinct: 85 by one estimate and 88 by another. What is disconcerting, however, is that the two lists contain only 27 species in common, resulting in, overall, 116 different species being listed by one or another of these studies. This suggests that, even among experts, there is some debate about what actually constitutes an extinction.

How much does this matter when assessing patterns of modern loss of mammalian species? On the one hand, documented extinction rates are far over background rates that definitive listing will not change the implications for conservation of mammals. On the other hand, such lists are playing an increasingly important role both in setting priorities for conservation (the process of keeping species off the list) and in public debate, and an accurate, defensible list is essential.

A clear ability to agree on what is meant by extinction is critical both to conservation planning and to conservation science. While all differences in lists of extinct species need to be reconciled, much can be learned by analyzing why one set of authors excludes species that other authors list as extinct (Table IV). The reason(s) why authors disagree can tell us much about the source of such errors, or differences of opinion, but also help us focus our efforts more clearly on those species where changes in status affect conservation action.

Most lists examine the extinction of species over a particular time frame. For instance, one might examine mammalian extinctions since the beginning of the age of exploration starting in 1500. The first question one must ask is, Did the species in question really exist at any time during the past 500 years, or was extinction during this time inferred incorrectly either through data in the literature or from misplaced stratigraphy of subfossil material? In Table IV, clearly a plurality of the 76 species excluded were in this category. There is no dispute whether or not these 31 species went extinct, rather just a question of whether their extinction predated the period under study. While inclusion or exclusion of these species will change the calculation of rates of extinction, such changes have little practical applications to conservation.

Similarly, removing or adding a species from a list for reasons of taxonomic uncertainty or taxonomic revision does not manifestly change conservation status of a species. Take for instance the quagga. Distinct from other zebras, with only vestigial stripes, the quagga, Equus quagga, was initially thought to be a distinct species of zebra extirpated from the southern tip of Africa at the turn of the twentieth century. Recent molecular analysis of museum specimens showed unambiguously that the quagga was a subspecies of the common zebra, E. burchelli—one less extinct species, to be sure—but such revision does not change the way we manage the existing populations of E. burchelli. Similarly, if two extinct species are found to synonymous, as in the case of Johnson’s huita (Plagiodontia ipanuem) and P. veloci (no common name), management plans remain unaffected.

When a species thought to be extinct is determined to still be extinct, more likely than not the size and status of the extant population will be either totally unknown or known to be critically endangered. Hence, transition from extinct to extant, unlike other categories, has important conservation implications. In Table IV, such cases represent about one-quarter of those species listed. Such a transition should be a red flag indicating that further study, and conservation action, are likely needed.
B. Patterns of Modern Mammalian Extinctions

While the rate of extinction is extremely high in comparison to the expected background rate of extinction, the time frame over which historical extinctions have occurred is so short that it is often difficult to discern patterns in data collected on mammalian extinction. Nonetheless, both taxonomic and geographical patterns do emerge, both of which may be informative if we try to project future patterns of extinction in mammals.

Perhaps the most striking pattern seen when one examines data on historical extinctions of mammals relates to the geographic distribution of extinct species. In the past 300 years, the great majority of extinctions have been on islands, and within island groups, the Caribbean islands have suffered the most extreme loss, accounting for nearly 40% of all recorded extinctions (Table V). Mammalian extinctions are, in this case, no different from those of other well-studied taxa, with over 90% of modern avian extinctions and 89% of modern molluscan extinctions also occurring on islands. In recent times, continental extinctions have, for the most part, been rather rare, and nearly all continental extinctions have been on the continent of Australia.

No sophisticated statistical analysis of the effect of body size and phylogeny has been made of recent extinctions, nor have analyses adjusted extinction rates for extant patterns of diversity. Yet the observed patterns suggest that smaller mammals have been particularly susceptible to extinction. This pattern contrasts sharply with that observed in the fossil record of the Americas during the Pleistocene–Holocene transition, in which the majority of known extinctions were in the megafauna of the region.

Perhaps the most striking set of extinctions is in the Insectivora, where an entire family (Nesophontidae, the Antillean island shrews) has gone extinct. Eleven percent of all recorded mammalian extinctions in the Modern era have occurred in this family. Insectivora, as an order, appears to have a greater propensity to extinction, with 11 species having gone extinct, representing 2.5% of the ±450 described species of Insectivora. This is about 30% above the overall rate of 1.9% for mammals.

Other orders that have overrepresentation in recent extinctions include two other orders, the most diverse being Rodentia (52% of all extinctions, 44% of described mammals). Similarly, among therian mammals, the most diverse order, Diprotodontia, with 117 species, which includes possums, cuscuses, wombats, and the koala, shows overrepresentation in the extinction table (6.8% of extinctions, 2.5% of all mammalian species) as does one of the least diverse, the Peramelida, or bandicoots (3.4%/0.45%).

In contrast, while nine species of the order Chiroptera have gone extinct, representing 10% of the known modern extinctions, bats represent nearly 20% of the extant modern mammalian fauna; hence, one could argue bat extinctions have been underrepresented in recent times. Similar arguments could be made for both carnivores (2.2% of extinctions, 5.8% of species) and primates (3.4% of extinctions, 5.1% of species).

C. Projecting Future Extinctions

Compiling lists of recently extinct species and studying current extinctions tells us something of the recent history of changes in mammalian biodiversity. The strength of such lists is that they are an assessment of the global patterns of extinction and tell us something about the persistence of a species, family, order, or size class. To better understand the immediate future of mammalian extinctions (perhaps the next 200 years) rather than reflect on what has gone extinct, there is greater value in examining which species, families, and orders are under threat and discussing how threats will lead to decline, and ultimately to extinction. These processes are more fully discussed elsewhere in this encyclopedia, as are patterns for particular orders of mammals.

For mammals overall, however, patterns of threat do not correlate well with patterns of recent extinction. A
recent assessment of threatened species by the World Conservation Union (IUCN) suggests that 25% of extant mammal species are threatened with extinction. Some of the groups that are overrepresented in analyses of recent extinctions are prominent in the list of threatened species: rodents are somewhat underrepresented, with "only" 17% of the species in the order threatened with extinction; insectivores, however, look likely to dominate the lists of extinct species in centuries to come, with 36% of the species in this taxon threatened. What is more interesting is that some orders that have shown relatively low rates of extinction such as carnivores and primates have high proportions of their taxa threatened. In carnivores nearly 26% of all species are under threat, while nearly half the primates, 46%, may go extinct in the next century.

Analysis of the extinction of subspecies or populations of animals gives some insight into what is generating this pattern. Among both carnivores and primates, rare species have been subject to increasing threats from habitat loss and fragmentation, isolating endemics in smaller and smaller patches. But wide-ranging species are also showing promise for future extinction through similar processes acting at larger scales. The spotted hyena, while still abundant and occurring at relatively high densities for a large carnivore, requires a matrix of resources that are widely dispersed across a landscape. Combined with a relatively low level of dispersal, habitat fragmentation and isolation suggest the species will come under increasing threat. The African wild dog still has a nearly continental distribution, but most populations are small (fewer than 100 individuals) and increasingly isolated. Threats to the species amplify one another: habitat loss and fragmentation, loss of prey, disease introduced by domestic companion animals, active hunting, and road kills all contribute to mortality. The decline of the species in West Africa, where populations became isolated and then went extinct one by one, like candles being snuffed, is probably indicative of patterns that are evolving in central and East Africa. Increasingly, while the species is still extant, and will remain so for decades in southern Africa, it is being lost as a component of the great majority of ecosystems in which it once lived.

While extremely valuable both for planning and for developing an understanding of which species are most likely to go extinct, lists of extinct or threatened mammals only address the concern of evolutionary extinction, the ultimate loss of a species. Increasingly, scientists who study mammals, and lobby for the preservation of mammalian diversity in particular and biodiversity more widely, will have to address issues related to the local, regional, and global ecological consequences of a species march toward extinction.

See Also the Following Articles

BIRDS, BIODIVERSITY OF • ENDEMSM • ECTINCTION, MODERN EXAMPLES OF • HOTSPOTS • LATITUDE, COMMON TRENDS WITHIN • MAMMALS, CONSERVATION EFFORTS FOR • MAMMALS, LATE QUATERNARY, EXTINCTIONS OF • MAMMALS, PRE-QUATERNARY, EXTINCTIONS OF • VERTEBRATES, OVERVIEW

Bibliography


MAMMALS, CONSERVATION EFFORTS FOR

E. J. Milner-Gulland* and R. Woodroffe†
*Imperial College, London and †University of Warwick, United Kingdom

I. Introduction
II. Threats to Terrestrial Mammals
III. Types of Conservation Effort
IV. Case Studies
V. Conclusions

GLOSSARY

conservation efforts Any action that aims to reduce the probability of extinction of a taxon over a specified time period.

ex situ conservation Conservation activities that involve individuals held outside their native habitats (e.g., in zoos or seed banks).

in situ conservation Activities which aim to conserve wild populations in their native habitats.

mammal Member of the order Mammalia. A species that provides milk for its young and has fur. They typically (but not exclusively) bear live young.

persecution Deliberate killing of animals perceived to be a nuisance.

sustainable use Exploitation of wildlife in a manner that avoids depletion of the resource (e.g., limited hunting, ecotourism).

terrestrial A species that lives on land (as opposed to marine or fresh water) for the majority of its life cycle.

CONSERVATION EFFORTS for terrestrial mammals must start with a consideration of the particular threats that face these species. Terrestrial mammals are a disparate group, but similarities in their conservation needs can be discerned, particularly when considering large mammals. Large-bodied species tend to be vulnerable to extinction both because of their biology and their interactions with humans. They tend to need large areas for survival and to be relatively slow growing, making their populations less resilient. They also tend to be more likely to be killed by humans for their meat, trophies, or because they are a danger or a nuisance. Because of these varied threats, conservation efforts for terrestrial mammals are broader than for some other taxa; they encompass both habitat protection and direct protection. Conservation efforts for terrestrial mammals are also noteworthy because they are often treated as flagship species for the rest of biodiversity conservation, sometimes at the expense of other species. This makes them important more broadly for conservation policy.

I. INTRODUCTION

In discussing conservation efforts for terrestrial mammals, we first need to identify what is particular about their conservation needs. It is hard to generalize about such a broad range of species, from bats and mice through antelopes and lynx to elephants, but there are some common characteristics that predispose terrestrial
mammals to threat, at least among larger-bodied species. These characteristics relate both to natural history and to human attitudes.

Large mammals have high energetic demands. They grow slowly, range widely, and occur at low population densities. Together these linked characteristics mean that large mammals are vulnerable. Their population dynamics are less resilient than those of other species, since they cannot bounce back quickly from population declines. Furthermore, they are more vulnerable to habitat loss than other species, because they rely on larger areas of habitat.

The natural history of mammals is not the only feature that predisposes them to vulnerability—large mammals are often under threat because they are worth killing. Being large bodied, most are valuable for meat; those that have trophies such as horns, tusks, or furs can be extremely valuable. Large mammals can also be prone to human persecution because they represent a threat, both to humans themselves and to their livelihoods, as predators on livestock or competitors for land and resources.

Large mammals can have a significant impact on their environments. This creates an intimate association between the conservation of terrestrial mammals and the conservation of other species in their ecosystems; conservation efforts targeted toward mammals might, therefore, have added conservation value over and above the individual species being conserved. Some such species have a profound impact on the ecosystems they inhabit. For example, the destructive effect of elephants (*Loxodonta africana*) on vegetation means that a trend in plant species richness with elephant density is visible; in the 1980s elephant densities inside Amboseli National Park, Kenya, were very high as they sheltered from poachers in the relative safety of the park. Numbers were very low outside the park, producing a gradient of elephant density that could be measured from damage to Acacia trees, a favorite food plant. Species richness of all plants (whether or not eaten by elephants) was highest at the boundary of the Park, at intermediate elephant densities (Fig. 1). Likewise, predators may have a marked effect on community structure. For example, local extinction of coyotes (*Canis latrans*), in parts of Southern California was associated with disappearance of a variety of smaller species (Crooks and Soulé, 1999), yet the distribution and densities of coyotes themselves have been shaped by the recent decline of a still-larger predator, the wolf (*Canis lupus*).

In this chapter, we shall first discuss in general terms some of the threats facing terrestrial mammals and

![Graph A](image1)

**FIGURE 1** (a) Plant species richness along a 12 km transect from the center of Amboseli National Park, Kenya. The Park boundary is at 6 km. (b) Changes in elephant density along the transect from the center of the park, measured by elephant damage to acacia trees. From Western, D. (1989). The ecological value of elephants: A keystone role in Africa’s ecosystems. In The ivory trade and the future of the African Elephant. Ivory Trade Review Group, Queen Elizabeth House, Oxford, UK.
which species are particularly vulnerable to these threats. We shall then discuss conservation measures that have been used to address some of the threats. Finally, we shall use in-depth case studies to examine how conservation efforts have been targeted at particular species, concluding by placing terrestrial mammal conservation into the broader context of international conservation efforts.

II. THREATS TO TERRESTRIAL MAMMALS

Like other taxa, many terrestrial mammals are at risk from habitat loss. As people become more numerous, they modify their surroundings in a variety of ways; this process creates new environments hostile to many wild species. Direct conversion of natural habitat threatens the high proportion of mammal species that cannot survive in farmed or urban environments. Many terrestrial mammals' large body size and high energy requirements make them particularly sensitive to human activities.

Direct killing by people is a serious threat, which perhaps affects terrestrial mammals more than other groups (with the possible exception of marine mammals and commercial fish species). Some forms of killing involve exploitation. Subsistence hunting of deer, antelopes, and many primate species can cause population declines, as can commercial hunting to fuel local and international trade. The bushmeat trade (trade in meat from wildlife for local consumption) is predominately based on mammal species; recent estimates have put the annual output for this trade at 1000 tonnes of meat a year in Central Africa alone (Robinson et al., 1999). There is a strong interaction between hunting and habitat loss; as an area becomes more accessible through the creation of logging roads, for example, hunting rates increase. Hunting for the international trade has caused some of the most serious and well-publicized species declines and includes hunting of elephants to make ivory carvings and spotted cats to make fur coats. Particularly insidious is the demand for animal body parts used in traditional Asian medicines; this has caused declines of tigers (Panthera tigris, in demand for their bones) and bears (Ursidae, in demand for their gallbladders).

Not all killing of mammals by people constitutes exploitation; several species, especially the large carnivores, have suffered serious declines through persecution. Large-bodied species such as elephants and bears are a risk to human life and are rarely tolerated in areas with high human densities. Much of the problem with human-elephant relationships in recent years is concerned less with ivory than with the fact that elephants are dangerous neighbors. Elephants can flatten a crop in minutes and are responsible for large numbers of human injuries and fatalities each year. Likewise, large carnivores are feared and perceived as a threat to livestock and have been the victims of organized eradication campaigns that saw, for example, the extirpation of wolves from most of the coterminous United States, and African wild dogs (*Lycaon pictus*) across much of Africa, in the first half of the 20th century.

Humans may also cause inadvertent declines through exploitation or control of relatively common species. Traps set to catch rabbits (*Oryctolagus cuniculus*) are a serious threat to endangered Iberian lynx (*Lynx pardinus*), which are accidentally captured and killed. Likewise giant panda (*Ailuropoda melanoleuca*) populations are seriously affected by accidental capture in snares set for other species. More seriously still, extermination campaigns for prairie dogs (*Cynomys* spp.) inadvertently caused the collapse of black-footed ferret (*Mustela nigripes*) populations.

Human activities also influence wild mammals' exposure to "natural" threats such as disease and predation. Frequent local extinctions of bighorn sheep (*Ovis canadensis*) in the Western United States, leading to serious species decline, have been largely attributable to infectious diseases contracted through contact with domestic sheep. Small, isolated populations are especially vulnerable to the threat of disease; thus disease may deliver the final coup de grace to species already crippled by other factors. Likewise, these populations may be unable to cope with predation pressures readily sustained by larger populations. When exotic predators or pathogens are introduced to native species, the impact can be more severe still. The introduction of red foxes (*Vulpes vulpes*) to Australia decimated populations of the smaller marsupial species, and the introduction of rinderpest—a viral disease of Asian cattle—to Africa caused widespread and often irreversible declines of many African ungulates.

The recent updating of the World Conservation Union's (IUCN) Red Lists of threatened species provides an opportunity to analyze the relative importance of different threatening processes. The IUCN categories of threat have limitations, reflecting experts' subjective assessments of which threatening processes are the most important for a species. They are also biased in
that poorly known or unknown taxa are not well represented. However, mammals are relatively well studied in comparison with other taxa, so that for them, the IUCN categories of threat can be used as a basis for assessing the relative importance of various threatening processes. This has been done by Mace and Balmford (in press), who show that about 25% of mammal species are threatened. This is higher than the threat levels others have estimated for birds, but not as high as for some other taxa. Mace and Balmford identify five major threat types: habitat loss makes up 47% of the threats to mammals, overexploitation makes up 34% of threats (66% of which is local exploitation, 21% commercial exploitation, 3% by catch and incidental catch, and 8% persecution and hunting as a pest). Other threats, such as introductions (13%) and rarity (6%), are less common among mammals. Thus, any analysis of the threats to mammals and the conservation efforts needed to counteract them is likely to emphasize habitat loss and overexploitation.

Another useful technique for looking at the effect of threatening factors on extinction risk compares the characteristics of populations that have become extinct with those that survive. These have shown the following links:

**Vulnerability correlates with species’ natural history.** Among carnivores, vulnerability is clearly associated with the extent of a population’s contact with people. For example, carnivores with large home ranges frequently travel beyond the borders of protected areas, where they come into contact with people. Such contact is often fatal, with the result that wide-ranging species require larger parks and reserves for effective protection (Woodroffe and Ginsberg, 1998).

**Vulnerability also correlates with degree of human use of the mammal’s habitat.** The direct and indirect effects of human activity on mammals lead to general associations between high human population density and local extinction of mammal populations. Elephants, primates, large and small carnivores all demonstrate such associations. It is not clear to what extent these associations reflect direct killing of mammals by people, destruction of mammals’ habitat by people, or, on the small scale, simple avoidance of people by mammals.

Individual species may be threatened by a number of different factors, so that the type of conservation effort needed can vary widely between species. However, it is possible to discern some broad patterns that can help to identify species that may be at risk of extinction. These patterns relate both to the biology of the species and to the level and type of human pressure on the species and its habitat.

### III. TYPES OF CONSERVATION EFFORT

The threats described here demand a variety of conservation measures, many of which have been applied to terrestrial mammals. We discuss these measures briefly, giving more detailed evaluations of their costs and benefits in the case studies.

**Legal protection of areas** is the most traditional form of conservation; this approach has the advantage of protecting habitats, as well as particular species, against a multitude of threats associated with human activities. While reserves are often targeted at a particular species, and even named as such (e.g., Addo Elephant Park, Gemsbok National Park), well-planned reserves can protect multiple species, habitats, and landscapes. Large mammals are often used as “flagships” to promote such reserves (e.g., India’s Project Tiger). Despite their effectiveness at protecting some species, limitations on local people’s rights to inhabit or exploit such areas have attracted controversy and may undermine the longer-term sustainability of reserves.

**Legal protection of species or populations** may also be used as a conservation measure. For example, stringent legislation in the United Kingdom protects the European badger (*Meles meles*) from killing and disturbance. Difficulties of enforcement mean that such legislation may have limited value as a conservation tool, however; for example, babirusa wild pigs (*Babyrussa babirussa*) are still openly traded for meat in Indonesia despite having full legal protection. Species threatened by exploitation may be protected by legislation governing trade rather than preventing killing. For example, in the Gambia it is not illegal to kill leopards, but international trade in their skins would be prohibited through the Convention on International Trade in Endangered Species (CITES).

Problems surrounding the enforcement—and natural justice—of legislation prohibiting the exploitation of wild populations or use of protected areas have led to initiatives aimed at conservation through sustainable use. The rationale behind such initiatives is that sustainable exploitation, with the benefits accruing to local communities, should ensure that local people value and protect natural resources. Like the establishment of reserves, this approach has attracted controversy; the array of political, legislative, and logistical concerns that it entails is different from, but no less complex than, that surrounding more traditional legal protection.
Threatened populations may also be protected through direct management. Habitat management involves measures such as provision of waterholes or den sites, or control of predators or prey. Population management describes a variety of interventions on the population being conserved, such as vaccination against infectious diseases and supplementation from other populations, wild or captive. Species conservation may also involve reestablishment of extinct populations through reintroduction of captive-bred or wild-caught mammals. Such measures have been important in the recovery of some endangered species (e.g., golden lion tamarins, *Leontopithecus rosalia*, red wolves, *Canis rufus*), as well as the reestablishment of populations of species that remain widespread but have suffered local extinctions (e.g., gray wolves).

Finally, the conservation of some endangered species may benefit from captive breeding, maintaining a population in captivity that may or may not be intended for release into the wild. Such efforts have proven valuable for critically endangered species reduced to their last few individuals (e.g., Arabian oryx, *Oryx leucopus*, black-footed ferret), in which managed breeding can minimize the loss of genetic variation.

IV. CASE STUDIES

A. African Wild Dogs

African wild dogs provide a case study of the interactions between persecution and habitat fragmentation in leading to a threat of extinction, and this example illustrates how, once depleted, a species can become more susceptible to other factors such as disease and competition with other species. Current conservation efforts for this species include protection in large reserves and reintroductions. We critically examine these efforts and suggest others that may also be useful.

1. Background

The African wild dog, *Lycaon pictus* (Fig. 2), is one of the world’s most endangered carnivores. Formerly widespread throughout sub-Saharan Africa, wild dogs have disappeared from 25 of the 39 countries in which they were formerly found, leaving a wild population of around 5000 animals (Fig. 3; Woodroffe et al., 1997). The species’ extreme rarity is illustrated by the fact that Africa’s elephants currently outnumber its wild dogs by about 100 to 1.

FIGURE 2  Picture of a wild dog. African wild dogs live in close-knit social groups but their overall population density is low and their home ranges are enormous.

2. Threats

Why have wild dogs populations collapsed so dramatically, when other large carnivores have experienced less serious declines? The answer probably reflects both the

FIGURE 3  Distribution of wild dogs circa 1995. Wild dogs’ historical distribution covered the majority of sub-Saharan Africa. The population marked with a question mark (?) is known only from unconfirmed reports. Data are from Woodroffe et al. (1997).
particularly intense persecution that wild dogs have suffered from people and wild dogs' unusual natural history, which seems to have made them especially sensitive to human activities.

The decline of wild dogs can be linked to the growth and expansion of human populations—they persist only in countries and areas with low human densities (Fig. 2). Because they occasionally kill livestock, wild dogs have long been shot and poisoned by farmers anxious to protect their stock. In the earlier parts of the 20th century, however, wild dogs were the victims of organized eradication campaigns in many parts of Africa. In Zimbabwe, for example, vermin control units shot 2430 wild dogs between 1956 and 1960, and the species was classified as vermin until 1975 (Childes, 1988). Control programs were not intended only to protect livestock. Wild dogs were believed to "kill wantonly far more than they need for food, and by methods of the utmost cruelty," so that "When the Uganda national parks were established it was considered necessary, as it had often elsewhere, to shoot wild dogs in order to give the antelope opportunity to develop their optimum numbers" (Bere, 1955). Wild dogs were shot in Serengeti National Park, Tanzania, as recently as 1973, and in Niger until 1979 (Woodroffe et al., 1997). Thus wild dogs experienced human-caused declines in areas where other species were protected.

Two lines of evidence suggest that such organized persecution is not the sole explanation for wild dogs' ongoing decline. First, wild dogs continue to disappear from protected areas, even though persecution inside reserves halted more than 20 years ago. Second, wild dogs were not the only large carnivore subject to organized control. Other species (e.g., spotted hyenas, Crocuta crocuta) were equally reviled and persecuted with similar enthusiasm, but persisted relatively well both inside and outside protected areas (Woodroffe, in press).

Wild dog ecology and behavior appears to have made them particularly sensitive to the persecution and habitat fragmentation associated with the growth of human populations. Even where they are protected, wild dogs live at extremely low population densities (average circa 2 adults per 100 km²) and range very widely (average circa 600 km² per pack). This appears to be a behavioral response to predation by lions (Panthera leo), which are a major cause of mortality for both adults and pups. Wild dogs avoid their larger competitor, favoring areas of low prey density where lions are scarce; thus wild dogs effectively experience low prey density even where prey are abundant, and wild dogs are commonest where lions are scarce. Wild dogs' low density means that very large areas (≥10,000 km²) are needed to support viable populations (Woodroffe et al., 1997). Worse still, their wide-ranging behavior means that few packs remain entirely within the borders of protected areas. Shooting, poisoning, accidental snaring, and road accidents—most of which occur on or outside reserve borders—account for more than half of the adult mortality recorded by studies of protected populations (Woodroffe et al., 1997). Ranging beyond reserve borders also increases contact with domestic dogs, a reservoir of infectious diseases threatening wild dogs (Woodroffe et al., 1997). Such edge-related mortality is most severe where substantial human populations abut reserve borders, and it is powerful enough to cause extinction in small reserves with high perimeter-to-area ratios (Woodroffe and Ginsberg, 1998). Thus, wild dog populations inhabiting small reserves (<5,000 km²) face a double jeopardy: small reserve size and high mortality due to strong edge effects. Wild dogs' behavioral response to competition with lions seems to predispose them to local extinction in fragmented habitats.

3. Conservation Efforts

Thus far, few in situ conservation measures have been implemented specifically to protect wild dogs. The species has persisted primarily in very large protected areas (Woodroffe and Ginsberg, 1998), and the maintenance of such areas has the highest priority for continent-wide wild dog conservation (Woodroffe et al., 1997). As Africa's growing human population makes increasing demands on the landscape, and reserves are de-gazetted, this is certainly not a trivial recommendation. Any measures that expand the area of contiguous land available to wild dogs will be beneficial, establishing cross-border parks and buffer zones and encouraging land use such as game ranching on lands bordering reserves will all contribute to wild dog conservation. Such landscape-level management would conserve wild dogs by

1. maximizing the size of wild dog populations, making them more resilient to perturbations such as disease outbreaks,
2. ensuring that wild dogs' prey base is maintained, and
3. minimizing the proportion of packs exposed to human activity by lowering perimeter-to-area ratios of wildlife areas.

Few of Africa's reserves exceed the 10,000 km² minimum needed to provide adequate protection for wild dog populations large enough to remain viable. Popula-
tions remaining in smaller reserves, or outside protected areas, are likely to require active population management to minimize mortality due to persecution, accidental snaring, road accidents and disease (Woodroffe et al., 1997). Such measures will demand

1. working with local farmers to devise husbandry methods that effectively protect livestock from wild dogs, using the findings as the basis for local education campaigns,
2. reducing snaring in protected areas through intensified antipoaching activities,
3. strict limitation on the construction or improvement of roads in and around protected areas, and
4. controlling the growth of domestic dog populations around wildlife areas to limit the spread of infectious disease.

Reintroduction has been advocated as a means of conserving wild dogs. However, although wild dogs have been extirpated across large tracts of Africa, especially in West and Central Africa, it is unlikely that reintroduction has any role to play in restocking these areas. Trials in southern Africa have shown that reintroduction is technically possible, but there are no suitable reintroduction sites in west or central Africa, and few in eastern and southern Africa; candidate sites are too small, or too poorly protected, or entail uncontrolled disease problems (Woodroffe et al., 1997). Furthermore, no wild dogs with the appropriate west, central, or east African genotypes are available for release in the areas where reintroduction is most needed (Woodroffe et al., 1997). Plans are underway to establish a "managed metapopulation" of wild dogs in small, fenced reserves scattered across South Africa's highly fragmented landscape, to be maintained by frequent translocation of animals between sites. This approach is locally valuable but will not establish a population likely to remain viable without intensive management in perpetuity. For these reasons, protection of remaining wild dog populations—and the vast landscapes required to support them—currently represents a better investment than any attempt at reintroduction.

B. Rhinoceroses

Rhinos are threatened predominately by the lucrative international trade in their horns. Conservation efforts have focused on strong protection in small areas, leading to the species existing in increasingly unnatural conditions. Strong laws have also been implemented, banning the commercial international trade in rhino products, although there is now a move toward conservation through trade, facilitated by dehorning. Captive breeding is also an important component of conservation efforts for some species.

1. Background

There are five species of rhino, spread across Asia and Africa. None of the species is entirely secure. The Sumatran (Diceros sumatrensis) and Javan (Rhinoceros sondaicus) rhinos are on the brink of extinction in the wild, with population estimates of 300 and <70, respectively (Foose and van Strien, 1998). The Indian rhino (Rhinoceros unicornis) population is around 1600 (Martin, 1996), with strongholds in the Royal Chitwan National Park, Nepal, and Kaziranga National Park, India. In Africa, there was a reversal of fortunes in the 1980s and 1990s; previously the black rhino (Diceros bicornis) was relatively widespread, with a 1970 population estimated in the high tens of thousands. By 1987 the population estimate had dropped to 3800, and the 1995 estimate was 2400 (Brooks, 1996). These estimates are disputed, but the rapid decline is clear. White rhinos (Ceratotherium simum), however, were already at very low numbers by 1900, and have only recently stabilized and started to increase under strong conservation protection. A recent population estimate was around 7500 individuals (Brooks, 1996). There are two subspecies of white rhino; the southern white rhino is currently secure, but the northern white rhino is confined to a population of a few tens of individuals in the Garamba National Park, Democratic Republic of Congo; this population is entirely reliant on continued protection.

2. Threats

Rhino horn has a long history of use and value. Ctesias of Knidos, the Persian court physician from 416 to 398 B.C., wrote of the value of Indian rhino horn drinking cups as poison detectors. Asian rhinos had already been heavily hunted by the beginning of the 19th century, to satisfy demands for horn in the Asian traditional medicine markets (Leader-Williams, 1992). In the 19th century, European game hunters targeted African rhinos along with other big game animals. In the 1830s, Cornwallis Harris saw white rhinos "in almost incredible numbers" in the Cape of South Africa (Martin and Martin, 1982), while Selous (1908) commented on the thousands of white rhinos being killed by hunters in the 1870s, and by 1899 was writing that "the two white rhinos which I shot in 1882 are the last of their species that I have ever seen alive . . . , and when I left Africa towards the end of 1892, I fully expected that these animals would become extinct within a short time."
Wars and the European withdrawal from Africa reduced hunting pressure in the first half of the 20th century, but the rhino horn trade increased to high levels in the 1970s. Oil revenues in Yemen led to large-scale imports of horn in the mid-1970s for dagger handles, while increasing prosperity in the Far East led to an increase in demand for rhino horn for medicinal use (Leader-Williams, 1992). Contrary to popular myth, rhino horn is not primarily used as an aphrodisiac, but as an antifever drug. CITES came into force in 1976, prohibiting the international commercial trade in rhino horn; however, key horn consumers such as Japan and Yemen did not ban horn imports until the 1980s.

Although total volumes of rhino horn traded might be significantly lower since the CITES ban, the populations from which it originates are now so small that any level of poaching is a threat to their survival. Thus the 1980s and early 1990s saw a battle between conservationists attempting to protect rhino populations through aggressive law enforcement and poachers killing rhinos for the lucrative illegal horn trade. Several authors have demonstrated a link between the amount of investment in law enforcement and a country’s success in protecting their rhinos (e.g., Martin, 1996). Hwange National Park, Zimbabwe, was extremely successful at protecting its rhino population until a 4-month hiatus in funding led to the entire rhino population being wiped out by poachers.

It is clear that the driving force behind rhino population declines has been large-scale commercial hunting over a long period, rather than other factors such as habitat loss. Rhino biology makes them particularly vulnerable to overexploitation. They are large, slowly growing species, with low rates of population increase. Correspondingly, they are able to sustain only a low level of hunting mortality (less than 5% of the population a year for a black rhino) before declining to extinction.

3. Conservation Efforts

Controlling rhino poaching can be approached from the supply end (preventing people from killing rhinos), the demand end (eliminating consumer demand for rhino horn products), or at some point along the supply chain (blocking off trade routes). The CITES international trade ban of more than 20 years’ standing has not been successful in blocking off trade routes, even if it has reduced overall volumes; as one country has clamped down, trade has shifted to others (Leader-Williams, 1992). A medicine as valued as rhino horn, used to treat serious illness, will continue to have a market even at very high prices. Thus demand is not easily reduced, despite efforts to find substitutes such as buffalo or saiga antelope horns. Whether or not demand for horn can be reduced in the long term, rhinos are so threatened currently that even a low level of exploitation in the short term could be disastrous. Thus conservation efforts have tended to concentrate on preventing the illegal killing of rhinos.

Experience of the disastrous decline in black rhino numbers over the past 20 to 30 years has shown that the key requirement for rhino conservation is small, heavily guarded reserves—an approach known as “fortress rhino.” However, there has been a recent shift in perspective, particularly in the southern African countries, caused by recovery of the southern white rhino populations and the general interest in sustainable use. The costs of heavy law enforcement to keep people out of reserves are very high, financially and socially; the shoot-to-kill policies instituted by several countries for anyone found poaching in a reserve are neither popular with local people nor particularly effective (Leader-Williams and Milner-Gulland, 1993). Since rhino poaching is done mostly by commercial poachers for the international trade, not by local people hunting for meat, the usual community initiatives to provide a sustainable income for would-be poachers are not possible; and although local attitudes can help in reducing poaching (for example, if local people do not shelter poachers), communities should not be expected to carry the burden of the expensive and dangerous law enforcement needed to prevent commercial poaching.

Rhino horns could be significant sources of income to support law enforcement; rhinos do not have to be killed for their horn to be used. Pioneering dehorning programmes (e.g., in Namibia) are seen by some as the answer to the problem of rhino conservation—not only does dehorning protect rhinos from poaching by making them unattractive targets, but selling the horn could provide revenue for conservation and development programs. Thus southern African countries are pushing for a limited legal horn trade, similar to that agreed by the 1998 CITES Conference of the Parties for ivory. However, although dehorning is a potentially promising conservation tool, the issues involved are complex and not fully understood. There are worries about the effects on rhino behavior and possible increased mortality caused directly or indirectly by dehorning. Dehorning is expensive and needs to be repeated often enough to make the rhinos unattractive targets for poachers; this may need to be every 1 to 2 years. The Hwange rhinos that were dispatched so rapidly by poachers in 1991 had been dehorned 2 or 3 years previously; it may be that the horns were large enough to be attractive to poachers once again. Although the legalization of the
international rhino horn trade may be extremely positive for Southern African countries, it might have negative effects on the already precarious Asian species by sending confused messages to traders and consumers. There has been debate about the relative merits of in situ and ex situ conservation of rhinos, particularly the highly endangered Javan and Sumatran species. Rabinowitz (1995) suggests that captive breeding is not a suitable conservation option for the Sumatran rhino and Leader-Williams (1993) suggests that in situ conservation (protection from poachers) is a much more cost-effective method of reducing rhino mortality than captive breeding. The problem is that these species have small population sizes, causing vulnerability to stochastic extinctions, but a dearth of the high-quality population estimates needed to calculate a suitable number for removal to captivity. There are also uncertainties concerning the sustainability of captive herds, given that some species have not bred successfully in captivity, and about the ability of in situ conservation to protect scattered individuals from a poaching threat.

Rhinos are severely threatened mammals, for which any continuation of exploitation could lead to the extinction of one or more species. They require intensive and expensive protection to guard against this threat, on an ongoing basis. It seems highly unlikely that any truly wild rhino populations will remain extant for long into the 21st century, with zoos and heavily guarded reserves being the only places where they are likely to survive.

C. Saiga Antelopes

Like rhinos, saigas are chiefly threatened by hunting for international trade, as well as for meat. However, there are major biological differences between them, and this means that their reaction to heavy hunting and thus the emphasis of conservation efforts are also very different. Conservation efforts for saigas focused in Soviet times on promoting managed commercial harvesting; recently these efforts have collapsed, and the situation is now being reexamined.

1. Background

The saiga antelope (Saiga tatarica, Fig. 4) is a nomadic herding species found in the semiarid rangelands of Kazakhstan, Russia, and Mongolia (Fig. 5). It has two subspecies, S.t. tatarica and S.t. mongolica; the latter is found only in Mongolia. The Kazakhstan populations currently make up more than 80% of the species; their ecology and management are reviewed in Bekenov et al. (1998).

The saiga has been hunted for its meat, horns, and hide since prehistoric times. In the 18th and 19th centuries, it was hunted in large numbers by the St. Petersburg Imperial court. By the early 20th century, hunting had reduced it to near extinction. Horn prices were very high, with horns exported for use in Chinese medicine. During the Soviet period, up to 1990, the population was well managed, with legal protection and regulated commercial hunting, and grew to relative stability. However, the situation changed dramatically with the breakup of the Soviet Union. Reports on the horn trade over the past few years are now mirroring those of the late 19th century and early 20th century in noting high prices, large quantities exported to China, and worries about the effects on populations. International concern about the species led to it being listed on Appendix II (monitored trade) of CITES in 1994.

The key features of the saiga's life history that make it resilient to hunting pressure are its nomadic behavior, its ability to give birth in the first year of life and to twin consistently, and its harem breeding system. The nomadic habit of the species is a response to the harsh climatic conditions of the continental ecosystem in which it lives, allowing herds to move rapidly away from areas where there is bad weather or where other threats exist. Saigas live in desert areas in winter, where there is lower snow cover, migrating up to 600 km to the steppe areas where they spend the summer; here higher rainfall levels provide abundant grass. The females' ability to twin consistently is unusual among ungulates and means that the population can increase.
very rapidly in good years, allowing populations to recover quickly from harsh winters or disease outbreaks. The harem breeding system, in which each adult male controls and mates with a group of 12 to 30 adult females, is important because there is an incentive for hunters to target adult males. Adult males bear horns, which are highly prized in Chinese medicine. Although heavy hunting for adult males has caused the population sex ratio to drop to less than 10% adult males (compared to 25–30% in an unhunted population), this has not led to any discernible decrease in population fecundity, possibly because the change in sex ratio is buffered by the harem breeding system (Milner-Gulland et al., 1995). However, if hunting of adult males continued to depress the sex ratio, sudden collapses in fecundity might result, with serious consequences for the population. Overall, though, the saiga antelope is a resilient species that is capable of withstanding relatively heavy hunting pressure and of recovering quickly from episodes of high mortality.

2. Threats

The saiga population is still numerous, at least in Kazakhstan. However, it is under increasingly heavy pressure and concerns are growing that it may rapidly become threatened. Threats to the saiga stem from the large-scale political and social upheaval that has happened in the former Soviet Union in the 1990s, leading to a collapse in the rural economy, with dramatic reductions in livestock numbers and crumbling infrastructure. This is good news for the saiga in that it is likely to have led to a recovery of overgrazed vegetation and a reduction in contact with livestock, reducing parasite and disease transmission rates. However, this benefit is more than offset by an increase in hunting of saigas for meat, as people who have lost their livelihoods turn to saigas for food and income. As the economic crisis has also hit scientific research, population counts are no longer as frequent and reliable as they once were, so that although there are anecdotal reports of large-scale population declines in recent years, these are unconfirmed. There is also an increased risk that mass mortality from disease may strike; collapses in veterinary services in Kazakhstan have led to outbreaks of foot and mouth disease in livestock, which could cause a major epidemic in saigas. In 1967, before livestock vaccination started, the disease spread over an area of more than 100,000 km² in central Kazakhstan, resulting in the deaths of 50,000 saiga calves (Bekenov et al., 1998).

Thus although saigas face threats from habitat loss and degradation, particularly in Kalmynia, the major current threat is illegal hunting for horns and meat, combined with the threat of mass mortality from disease. Threat levels vary; one population in Kazakhstan is under intense pressure, another is relatively unaffected because of its remoteness. In Mongolia, the threats are less severe, but are magnified because of the very small population size, making it vulnerable even to low levels...
of human pressure. The threat to the Mongolian population should also be taken more seriously because of its genetic distinctiveness. Thus although overall the saiga is not in imminent danger of extinction, it is under threat in most of its range. It is an important species, being the only wild ungulate present in significant numbers throughout most of its range, as well as being economically significant; this suggests that sustainable management could have significant benefits for local people.

3. Conservation Efforts

The traditional approach to mammal conservation of establishing protected areas is unfeasible for saigas. Because the saiga is nomadic, it is reliant on huge areas of relatively natural vegetation continuing to exist in the semi-arid rangelands of Central Asia. Another approach that is often used is to establish protected areas in key locations, for example, in calving or mating areas. Again, this is problematic for the saiga; although they do tend to visit particular areas, they are not predictable in their use of any one area from year to year. One of the best aspects of Soviet conservation measures was the ability to establish temporary reserves (zarazniki) at particularly sensitive times such as the birth season, placed wherever the saiga happened to be, in which there were specific prohibitions on activities such as grazing, hunting, and driving motor vehicles. These flexible restrictions are much more likely to be of practical benefit to the saiga than the necessarily small permanent reserves that might suit other species.

However, any system of protection is only as effective as its enforcement. This relies not only on high levels of investment in law enforcement agencies, but also on the engagement of local people in conservation efforts. The importance of local people having a stake in conservation, usually through revenues from the use of these species, is well accepted nowadays. The saiga appears to be an ideal species for sustainable use, being resilient to harvesting and of high value. However, the problem in giving local people a stake in its conservation is its nomadic behavior; the range area of each saiga population in Kazakhstan is hundreds of thousands of km². Without a direct and predictable link between a community's conservation actions and the status of the population they are conserving, community-based conservation will not work.

This discussion suggests that saiga conservation needs to be carried out on a larger scale than the local community. Again, the previous practice of Soviet saiga managers is a good model to follow. Commercial hunting organisations were responsible for saiga management from the protection of birth areas to the processing and distribution of meat. This chain of responsibility within a single organization meant that the correct incentives existed for conservation of the resource on which future profits depended. However, problems have increased recently; large-scale selective poaching has led to the commercial organizations being unable to harvest males and thus they cannot realize the potential profits from horn sales. Underfunding of law enforcement efforts, no investment of poaching fines into saiga conservation, and a lack of support for hunting inspectors’ powers of arrest have all contributed to the problem (Bekkenov et al., 1998). Other problems include corruption, particularly at customs, the lack of international control on saiga horn sales, and the huge areas that must be covered to protect saigas. A system that could have continued to work well for saiga conservation has not been able to cope under the stresses of sociopolitical change, leaving the saiga effectively unprotected.

The key issue identified here is the overwhelming effect that sociopolitical changes can have on efforts to conserve mammals. General economic conditions, particularly in the rural economy, or other external events such as war (as seen recently for the wildlife of Rwanda), can render even the best conservation system powerless to resist the overexploitation of a species.

D. Themes of the Case Studies

In the case studies presented here, we discussed the contrasting cases of three mammal species. In all three cases, threatening factors are modulated by the species' biology—two mammal species facing similar threats will often react differently. These differences are reflected in the conservation efforts that have been directed at each of the species.

African wild dogs were perceived as vermin and severely persecuted in the past, both inside and outside protected areas. However, although this persecution has caused the population to decline to very low levels, it is not the whole story, because other persecuted species like hyenas have not suffered the same fate. Because the wild dog lives at very low densities, it requires very large areas to maintain a viable population. Ranging widely, the species is also more vulnerable to mortality from road traffic, disease, and persecution at the edge of protected areas, so that the smaller the protected area, the higher the mortality rate from these factors.

Rhinos are also highly endangered species, but in this case, their threatened status can be firmly linked to high levels of hunting for their horns throughout
recorded history. Again their biology is a contributory factor in their susceptibility to hunting, because they have very low population growth rates so cannot sustain high levels of hunting. As their populations decline, even low levels of exploitation become disastrous. Because of the nature of the threat they face, protection measures center around very intensive (and expensive) protection of small reserves where they can be guarded from poachers. This is in stark contrast to the wild dogs, which require very large protected areas.

The theme of the interaction between a species' biology and its threatened status continues with the saiga antelope. This species was hunted to very low numbers at the turn of the century to supply the market for Chinese traditional medicine, just like rhinos. However, unlike rhinos the saiga is a biologically resilient species, and saiga numbers bounced back to peak at about 1 million animals in the 1970s. Since then, heavy hunting and habitat alteration have reduced numbers, but it is only in the past few years, since the fall of the Soviet Union, that hunting threatens the species again. The saiga, like the wild dog, is wide ranging and reliant on extremely large areas of relatively undisturbed habitat for its survival; this is available for saigas at the moment, but is always vulnerable to conversion for human needs.

Critical examination of the conservation efforts directed at these species throws some light on current debates in conservation. For example, conservation efforts are currently often focused on promoting sustainable use by local communities, under the philosophy that only if a species is given a monetary value will people have the incentive to conserve it. However, none of the species described here appears a particularly suitable candidate for community-based conservation based on sustainable use. The wild dog has little obvious commercial value, the rhino has too much and thus needs expensive protection, while the migratory saiga is not easy to assign to the ownership or control of any particular community.

Another suite of conservation efforts involves intensive intervention, using techniques such as captive breeding and reintroductions. Rhino conservation has included captive breeding and ranching, and reintroductions are seen as one solution to the problems faced by wild dogs. Even saigas, which are still very numerous in the wild and do not yet have serious problems with habitat loss, are being proposed as candidates for ranching, captive breeding, and reintroduction. Our analysis suggests that these technological solutions of ex situ conservation are generally not useful; wild dogs need such large areas to receive adequate protection that hardly any suitable reintroduction sites are available, saigas are too flighty to be kept in captivity, and several of the rhino species have not bred successfully in captivity. The southern white rhino, which was brought back from near extinction through careful ranching and re-introduction programs, is the exception rather than the rule. Thus traditional conservation in situ is the best way forward for most of the populations of the species discussed in these particular case studies.

Conservation efforts for threatened mammals will always involve a suite of responses; which is appropriate in a given case depends on the interaction between the threats that it faces and the biology of the species. In the case of the species discussed, the efforts that may work best include for wild dogs, a landscape-level approach to habitat management, complemented by community-level conflict-minimization programs; for rhinos, intensive guarding of small protected areas; and for saigas, control of poaching followed by the potential reinstatement of commercial hunting. For other species, different approaches will be suitable. However, the key issue is money—protection from anthropogenic threats is expensive, and someone needs to pay. In the case of the saiga antelope, and possibly the rhino, the trophies from the species themselves could provide substantial revenues to offset conservation costs, but the wild dog is not able to fund itself in this way. In the end, conservation comes down to the willingness of people to pay for it, either directly or through the opportunity costs of not converting a species’ habitat, rather than to the specifics of its biology.

V. CONCLUSIONS

Terrestrial mammals are often used as flagship species for conservation. The classic example of this is the panda used by the World Wide Fund for Nature, but tigers, elephants, and other charismatic species are also featured in publicity material about conservation because they are appealing to the public. Many people, particularly in wealthier countries, have very high intrinsic valuations of mammals and are extremely keen to preserve them, give them rights, and ensure the welfare of individual representatives of the species. The recent controversial proposal in New Zealand for legislation to extend some human rights to ape species provides an example of people regarding mammals as somehow different to other taxa. Other people, particularly those who live alongside wild mammals in poorer countries, are very keen to use them. Terrestrial mammals are often valuable species because of their meat, furs, or trophies, or destructive neighbors that invite
persecution. At the same time, because of their biology, these species can be very susceptible to overexploitation.

Thus terrestrial mammals are high-profile species among both the general public and conservationists. This combination of attributes means that mammals often provide a battleground on which conservation battles of more general significance are fought. An example of this is the debate about the appropriate place for the listing of the African elephant on the CITES Appendices. This began in 1989 with a successful proposal to transfer the species from Appendix II (controlled international trade) to Appendix I (no international commercial trade), on the basis that the elephant was threatened with extinction from an uncontrolled, and largely illegal, international ivory trade. Although the international ivory trade could in itself explain the observed 12% per annum decline in the African elephant population over the period 1979–1986, there were other issues involved, including escalating conflicts between elephants and local people as elephant habitat declined. The debate about the ivory trade continues to rumble on, with the latest issue being CITES’ agreement to a limited resumption of ivory sales from March 1999. However, the ivory trade debate was also the catalyst for a public airing and popularization of the issue of sustainable use, and the sovereign rights of nations to decide on the management of their own biodiversity. Popular articles explained the needs of local people to make a living from their wildlife if they are to have an incentive to conserve it in the face of growing pressure on natural resources. Others suggested that the sustainable use of elephants was either morally wrong, scientifically unwise given the current state of the African elephant population, or impractical because of the endemic corruption among officials and the very high value of the product.

Thus the conservation of terrestrial mammals is high profile and an area in which general conservation policy is debated and eventually made. The danger is that this policy may not be at all relevant to other components of biodiversity, with their own specific conservation requirements and biological peculiarities. Policies made with respect to one high-profile mammal species may not even be positive for other mammals. Flagship species have to be chosen wisely, since experience shows that species’ conservation needs cannot be predicted from their body size or trophic position. For example, in the 1980s, India designated a network of reserves designed to protect tigers, under the banner of “Project Tiger.” The project was expected to have the added benefit of protecting other species, notable the dhole (Cuon alpinus). This expectation was not met; dhole require much larger areas to persist than do tigers and have disappeared from several Project Tiger reserves (Woodroffe and Ginsberg, 1998).

Conservation efforts for terrestrial mammals are particularly noteworthy because of their impacts on biodiversity conservation in general. This is partly because mammals can be powerful forces structuring ecosystems. But it is also because they are often flagship species for conservation policy; this can be a mixed blessing for the conservation of biodiversity as a whole, as policy made with one particular group of species in mind will generally be inappropriate for others. Though this calls for caution in extrapolating the needs of terrestrial mammals to conservation policy on the broader scale, it is also important to remember that terrestrial mammals are particularly in need of conservation efforts in their own right.

See Also the Following Articles
CAPTIVE BREEDING AND REINTRODUCTION • CONSERVATION EFFORTS, CONTEMPORARY • EX SITU, IN SITU CONSERVATION • MAMMALS, BIODIVERSITY OF • MAMMALS, LATE QUATERNARY, EXTINCTIONS OF • MARINE MAMMALS, EXTINCTIONS OF • PREDATORS. ECOLOGICAL RULE OF

Bibliography
Mace, G. M., and Balmford, A. (in press). Patterns and processes in contemporary mammalian extinction. In Future Priorities for the


MAMMALS (LATE QUATERNARY), EXTINCTIONS OF

Paul S. Martin
University of Arizona

I. Introduction: Prehistoric Extinctions
II. The Comparative Approach
III. Pattern and Causes

GLOSSARY

extinction spasm or pulse A catastrophic burst of extinctions, peaking in less than a millennium.
first contact Initial human arrival on a landmass followed by human colonization.
late Quaternary extinction event Selective prehistoric extinctions, typically catastrophic, eliminating within the past 40,000 years two-thirds or more of large land mammals of America, Australia, and Madagascar and at least half the species of land birds on remote islands of the Pacific. Humans are present or suspected to be present in virtually all cases.
mega fauna Large terrestrial vertebrates variously defined as >1, >10, and >44 kg adult body weight. Equivalent to 100 pounds and similar to the average weight of adult humans, a body size exceeding 44 kg is easily visualized and is adopted here.
quaternary The ice age of at least the past 1.81 million years, including the Pleistocene and the Holocene, the latter representing the past 10,000 years.
radiocarbon dating An isotopic or nuclear decay method for inferring age of organic materials. Carbon 14 is produced in the upper atmosphere by cosmic ray bombardment and oxidized to form C14O2. Distributed through the earth's atmosphere and oceans, a small percentage is incorporated into a variety of organic materials to decay with a half-life of 5700 years. By dating tree rings of known age, and other methods, radiocarbon determinations can be calibrated. Although routine application of radiocarbon dating is usually limited to dates of less than 40,000 years, ages up to 75,000 years have been measured.

I. INTRODUCTION: PREHISTORIC EXTINCTIONS

The late Quaternary extinction event (LQEE) is best known by the loss of large mammals (megafauna) from certain continents. Familiar faunal examples of the LQEE include woolly mammoths of the Northern Hemisphere, woolly rhinos in northern Eurasia, ground sloths and glyptodonts in the Americas, and diprotodonts and giant kangaroos in Australia. Comparisons between landmasses are revealing. The continents differ significantly in both the magnitude and in the timing of their extinctions. First Australia, then North and South America, and lastly Madagascar (the Island Continent) rapidly lost two-thirds or more of their large mammals (Fig. 1). In contrast, losses were comparatively gradual and much less severe in Eurasia and Africa. Although small animals of the continents seldom
suffered extinctions, thousands of endemic mammals, birds, and reptiles vanished from oceanic islands.

At least 85 genera of large (>44 kg) mammals, half of the terrestrial megafauna of the planet, disappeared from the continents in the past 100,000 years. Now "out of sight, out of mind," the extinct mammoths and other large animals of the Quaternary coevolved with modern plant species, thereby shaping certain features of vascular plant anatomy and biochemistry. Undoubtedly, the extinct species were just as influential as living megaherbivores in determining the structure of natural communities (Owen-Smith, 1988). Before 13,000 years ago the diversity of large mammals in the New World would have resembled that currently found in African or Asian game parks. Megafaunal extinctions have made it difficult for ecologists to envision the multiple stable states prevailing in evolutionary time. In the absence of their megafauna, the savannas, grasslands, riparian habitats, and other biotic communities of the New World, Australia, and Madagascar are not functionally "natural."

Prehistoric settlement on deep-water islands in the remote Pacific began 4000 years ago. Extinctions followed, eliminating many endemic birds, reptiles, and land snails along with depletion or size reduction brought by overharvesting of near-shore fish, shellfish, and marine turtles. The relatively small number of oceanic islands that escaped prehistoric discovery suffered losses during historic time of such animals as the dodo, Steller's sea cow, giant rats, and parrots (Fig. 2). No late Quaternary extinction pulse is seen in the fossil record of the whales, other marine fauna, or in Antarctica.

An intense and often controversial search during the past 40 years for an explanation for the cause or causes of the LQEE has failed to yield any widespread consensus among paleontologists, archeologists, paleoecologists, geographers, and other interested parties, many of them convinced of the efficacy of climatic change and of climatic extinction models. In recent years, as the global pattern of the LQEE has become better known, much more attention has been focused on anthropogenic models (compare MacPhee (1999) with Martin and Klein (1984)). Always on the scene, humans can no longer be ignored in the search for the cause or causes of "extinctions in near time."

II. THE COMPARATIVE APPROACH

In the view of most geologists the Quaternary or Pleistocene ice age of multiple glaciations embraces at least the past 1.81 million years (Ma), with the late Quaternary representing the past 200,000 years. In North America the late Quaternary includes the Rancholabrean mammalian age stage, marked by the appearance of the genus Bison.

With the possible exception of Australasia, most late Quaternary extinctions fall within the reach of radiocarbon assay, essentially the past 40,000 years. Fossils uncontaminated by groundwater and yielding bone collagen or gelatin are suitable for dating. Environments favorable for the preservation of bone collagen include dry caves of arid regions, the frozen ground of high latitudes, lacustrine deposits, and saline or petrolic sediments such as those at Rancho la Brea, California.
Burned bone and associated charcoal are ideal for dating, although charcoal may prove intrusive. A variety of geochemical dating methods, especially radiocarbon dating, allow biogeographers to compare both the rate of extinction within a landmass and the timing of extinctions between landmasses. A robust chronology allows intercontinental and insular comparisons of considerable value in inferring cause of the extinctions.

Although details regarding just how the extinctions were triggered remain speculative, it is increasingly clear that the two major contending explanations of what forced LQEE—that is, climatic changes and cultural impacts—can now be evaluated globally. When approached on a comparative basis, human expansion and climatic change track a punctuated extinction episode in North America and perhaps in South America, whereas climatic change is less clearly involved in Australia or Madagascar and appears to be of no significance in New Zealand and other remote Pacific islands.

A. North America

Focusing on the past 4 million years only (the Pliocene and the Quaternary), it is possible to evaluate the LQEE using age-stage divisions of biostratiographers (Table 1). The overwhelming importance of extinctions of North American large mammals at the end of the Quaternary, shown in Figs. 3 and 4 and in Table 1, has been evident for at least the past 50 years.

The widespread adoption of screen washing of fossil deposits has vastly enriched the fossil record of small

FIGURE 2. Map silhouettes of continents and islands showing time-transgressive sequence of late Quaternary extinctions and F.C. Africa and Eurasia (1–3): Sequential extinctions of large mammals during the past 100,000 years including Nauman’s elephant and giant deer in Japan (1–2] ca. 30,000 years ago. Megacerom (2): Humans arrive and major extinctions occur 30,000–40,000 yr bp; neither event is well dated or constrained. Americas (3): Well-dated evidence for human arrival and for megafaunal extinctions center on 11,000 yr bp (13,000 calendar years). Mediterranean Islands: Epipaleolithic arrival and dwarf hippo and dwarf elephant extinction on Cyprus (3) 10,500 yr bp; mid-Holocene colonization and extinction of goat antelope (Myotragus) on the Balearic Islands (4); F.C. and LQEE chronology elsewhere are uncertain. Antilles: Humans arrive in Cuba and Hispaniola in early or mid-Holocene (4) and in Jamaica and the Lesser Antilles in the late Holocene (5); few radiocarbon dates on extinctions. Madagascar (5): Humans arrive 2000 yr bp; major episode of extinction terminates 1500 AD. Mascarenes, east of Madagascar (6): Humans arrive 1600 AD, followed by extinctions of dodo, solitaire, and giant birds, and giant tortoises. New Zealand (5): Humans colonize by 1300 AD; the Polynesian rat perhaps colonized much earlier. Giant flightless birds (moas) are extinct by 1500 AD. Wrangel Island (5): Last woolly mammoths dated 4000 yr bp. 1000 years older than oldest cultural material. Commander Islands (6): Humans arrive 1741 AD; Steller’s sea cow extinct within 30 years. Galapagos Islands (6): Bishop of Panama arrives 1535 AD; extinction rates increase over background by two orders of magnitude (after Martin and Steadman, 1999).
### Table I
Large (>44 kg) Plio-Pleistocene Terrestrial Mammals of North America North of Mexico*

<table>
<thead>
<tr>
<th></th>
<th>BLANCA1</th>
<th>IRV</th>
<th>RLB</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Stage duration (Ma)</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

| Xenarthra            |        |     |     |     |     |     |     |     |
| + Glyptotherium, glyptodont | X | X | X | X | X | X | X | X |
| + Helmetina, northern pampatheres | X | X | X | X | X | X | X | X |
| + Pachyermatherium, ground sloth | X | X | X | X | X | X | X | X |
| + Ereotherium, giant ground sloth | X | X | X | X | X | X | X | X |
| + Nothrotheriops, Shasta ground sloth | X | X | X | X | X | X | X | X |
| + Megalonyx, Jefferson's ground sloth | X | X | X | X | X | X | X | X |
| + Paramylodon, big-tongued ground sloth | X | X | X | X | X | X | X | X |

| Carnivora            |        |     |     |     |     |     |     |     |
| + Borophagus, plundering dog | X | X | X | X | X | X | X | X |
| * Canis, dire wolf, gray wolf | X | X | X | X | X | X | X | X |
| + Protoxylocyon's dog | X | | | | | | | |
| Ursus, bears         | X | X | X | X | X | X | X | X |
| + Tremarctos, Florida cave bear | X | X | X | X | X | X | X | X |
| + Arctodus, short-faced bear | X | X | X | X | X | X | X | X |
| + Chasmaporthetes, hunting hyena | X | X | X | X | X | X | X | X |
| + Megantereon, western direwolf | X | X | X | X | X | X | X | X |
| + Smilodon, sabertooth | X | X | X | X | X | X | X | X |
| + Ischyrotherium, Idaho sabertooth | X | | | | | | | |
| + Homotherium, scimitar cat | X | X | X | X | X | X | X | X |
| + Dinofelis, false sabertooth | X | X | X | X | X | X | X | X |
| [*] Panthera, American lion, jaguar | X | X | X | X | X | X | X | X |
| + Miracinonyx, American cheetah | X | X | X | X | X | X | X | X |
| Felis, cougar, puma   | X | X | X | X | X | X | X | X |

| Rodentia             |        |     |     |     |     |     |     |     |
| + Procateroides, large beaver | X | X | X | X | X | X | X | X |
| + Castoridae, giant beaver | X | X | X | X | X | X | X | X |
| + Neocapromys, giant capybara | X | X | X | X | X | X | X | X |
| [*] Hydrochoerus, Holmes's capybara | X | X | X | X | X | X | X | X |

| Proboscidea          |        |     |     |     |     |     |     |     |
| + Mammut, American mastodon | X | X | X | X | X | X | X | X |
| + Stegomastodon, stegomastodon | X | X | X | X | X | X | X | X |
| + Rhynchotherium, rhynchothere | X | X | X | X | X | X | X | X |
| + Cuvieriornis, gomphotheres | X | X | X | X | X | X | X | X |
| + Mammutus, extinct mammoths | X | X | X | X | X | X | X | X |

| Sirenia              |        |     |     |     |     |     |     |     |
| + Hydrodamalis, Stellar's sea cow | X | X | X | X | X | X | X | X |
| Trichechus, manatee  | X | X | X | X | X | X | X | X |

| Pterosodactyla       |        |     |     |     |     |     |     |     |
| + Cimolotherium, extinct eucyn | X | X | X | X | X | X | X | X |
| + Macrohipparion, gazelle-horse | X | X | X | X | X | X | X | X |
| + Plesihipparion, extinct eucyn | X | X | X | X | X | X | X | X |
| [*] Equus, horse species | X | X | X | X | X | X | X | X |
| [*] Tapirus, tapir species | X | X | X | X | X | X | X | X |

*Continues*
<table>
<thead>
<tr>
<th>Artiodactyla</th>
<th>BLANCAN</th>
<th>IRV</th>
<th>RLB</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>+Megalalopex, large camelid</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>+Blanocamelus, giraffe—camels</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+Tiananlopus, giant camelid</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Camelops, camel species</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Hemiancieros, llama species</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Palaeolama, stout-legged llama</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Mylohyus, long-nosed peccari</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Platygonus, flat-headed peccari</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Brevicia, false elk</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odontocerus, dear species</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Torincoceros, extinct large cervid</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+Navahoceros, mountain deer</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rangifer, caribou</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Alces, moose, broad-fronted moose</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+Cervid, stag-moose</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cervus, wapiti (elk)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Tetraceros, four-horned pronghorns</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Hoceros, Hay's pronghorn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+Slocomos, four-horned pronghorns</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antilocapra, pronghorn</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+*Saiga, saiga</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*Oreamnos, mountain goat</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Ovis, bighorn or mountain sheep</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+Eucranium, shrub ox</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Sorex, squirrel's ox</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>+Bootherium, horned-headed musk ox</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+Praevald, extinct musk ox</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ovibos, musk ox</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>[</em>] Bison, bison species</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+Placental, flat-horned ox</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Primates                     |         |     |     |   |
| +Mammals, modern H. sapiens  | X       |     |     |   |

**Originations**: 18
**Extinctions**: 24
**Total genera**: 26

---

*Abbreviations used: IRV, Irvingtonian; RLB, Rancholabrean; H, Holocene; E, early; M, middle; L, late; *, extinct genus; +, extinct species; brackets indicate generic survival on other continents. Living genera including extinct taxa of large size that vanished with late RLB extinction include Dasypus bellii (giant armadillo), Canis dirus (dire wolf), Panthera tigris (American lion), Alces latifrons (broad-fronted moose), and Oreamnos harringtoni (Harrington's mountain goat). Age estimates in millions of years (Ma): Blancan 1.4—0.35 Ma; Blancan 2, 3.5—3 Ma; Blancan 3, 3.0—2.5 Ma; Blancan 4, 2.5—2 Ma; Blancan 5, 2.0—1.8 Ma (Blancan—Irvingtonian Boundary); early Irvingtonian, 1.8—0.9 Ma; middle Irvingtonian, 0.9—0.5 Ma; late Irvingtonian, 0.5—0.3 Ma; early Rancholabrean 0.3—0.1 Ma; late Rancholabrean, 0.1—0.01 Ma; Holocene, last 10,000 years. There may not be specimen records for all cells within the temporal range of a genus as plotted here. Checkpoint and age estimates are courtesy of Elaine Anderson (1996), Denver Museum of Natural History, Denver, Colorado, revised from Kurten and Anderson (1980).*
vertebrates. The harvest of identified small mammals is vital in supporting the conclusion that no appreciable extinctions of small mammals occurred (Fig. 3)—certainly not matching the extinction spasm of large mammals approximately 11,000 radiocarbon years ago (Fig. 4).

The fossil record is based on a much larger sample of late Pleistocene than early Pleistocene and Pliocene

![Graph showing mammal extinctions and origins over time](image1)

**FIGURE 3** North American small mammal originations (dotted line), extinctions (solid line), and total fauna (dashed line) since the Miocene from data shown in Table 1. Note the change of scale in the past 0.5 Ma (reproduced with permission from Martin and Steadman, 1999).

![Graph showing mammal extinctions and origins over time](image2)

**FIGURE 4** North American large mammal originations (dotted line), extinctions (solid line), and total fauna (dashed line) since the Miocene from data shown in Table 1. Note the change of scale in the past 0.5 Ma (reproduced with permission from Martin and Steadman, 1999).
faunas (Table 1 and Figs. 3 and 4). Presumably this accounts at least in part for the increase in number of genera of both large and small mammals in the later part of the record. Although *Mammuthus, Equus,* and extinct *Bison* included multiple species, most of the large genera listed in Table 1 are monotypic (represented by one species only). Although a few survived on other continents, most large genera totally vanished (Table 1).

In North America north of Mexico, 33 genera of large mammals disappeared in the late Quaternary, which constitutes two-thirds of the total late Rancholabrean fauna of 48 genera and more than all the generic extinctions recorded during the previous 4 million years prior to LQEE (Table 1). Events that might trigger regional or global extinctions in the early part of the past 4 million years include intercontinental surging of the Americas with faunal intermingling, extraterrestrial accidents such as the impact of the Eltanin bolide, and the initiation of continental glaciation. However, only at the end of the Quaternary is there a major extinction spasm—one that impoverished America's large mammalian faunas (Fig. 4).

Unlike the case of the large mammals, a much larger number of small genera (48) vanished in the past 4 million years. Within the minifauna (rat, mouse, or shrew-size mammals), which make up the majority of the 75 living genera of small mammals, no generic extinctions are known in the late Quaternary. Among the medium-sized mammals, three genera were lost: the antilocaprid *Stockoceros,* the skunk *Brachypotoma,* and the rabbit *Azilianolagus.* The difference deserves close inspection. Although small mammals escaped virtually unscathed, the LQEE (late Rancholabrean losses) blighted all terrestrial orders of large land mammals and eliminated one, the Proboscidea (Table 1).

Since the vast majority of Cenozoic (last 65 Ma) extinctions occurred long before radiocarbon time or F. C., it is essential to know something of the Cenozoic pattern. Do earlier bursts of extinction match the LQEE pattern at the very end of the Cenozoic? As a result of a relatively rich fossil record of the Cenozoic, the question is readily approached in North America.

Paleontologist John Alroy (cited in MacPhee, 1999) recently analyzed 65 Ma of change in mammalian faunas through 1.0 Ma sampling bins. His data consist of extinction time rate series that are computed from a multivariate ordination of 4015 faunal lists from 2415 publications that span the Late Cretaceous through Sangamonian (last interglacial). The lists can be viewed on the World Wide Web at the North American Mammalian Paleofaunal Database: http://homebrew.si.edu/nampfd.html.

According to Alroy (as quoted in MacPhee, 1999, p. 120), standing diversity is defined by

\[
\text{number of species that cross each boundary between sampling bins, e.g., the time planes at 65.0, 64.0 \ldots 1.0 Ma. Extinctions are computed by counting the number of species in each cohort that fail to survive until the next time plane. Note that this excludes single-interval species (those that appear and disappear in the same bin), which makes the analysis less vulnerable to sampling artifacts.}
\]

Alroy's results indicate that within the past 55 Ma there were more extinctions of large species in the LQEE than there were during any earlier time. The difference in the mean body masses of the victims and the survivors peaks at the end of the Pleistocene (Fig. 5).

Although there is a general increase in mean mass of mammals through the Cenozoic, the mean mass of victims (extinct species) is highest in the past 1 million year bin within the Cenozoic (Fig. 6), which includes the LQEE. In other words, when compared with 64 turnovers earlier in the Cenozoic, the 65th has a unique property—excessive extinction of large mammals.

With the exception of bison (*Bison*) and deer (*Odocoileus,* the common fossils of large herbivores found in Quaternary deposits before the LQEE are not the

---

**FIGURE 5** Diversity and extinction in the Cenozoic. Data are prepared by multivariate ordination and randomized subsampling. The figure plots the correlation among 1.0-Ma intervals between extinction intensity and the difference in the mean body masses of the victims and the survivors. The end Pleistocene value (●) is among the highest during the past 55 Ma (reproduced with permission from MacPhee, 1999).
productivity, and evolutionary potential of the continent.

B. South America, Australia, and Madagascar

The continents of South America, Australia, and Madagascar share a common property with North America—heavy extinction of large mammals late in the Quaternary. South America lost all large mammals more massive than a tapir (300 kg); Australia lost all mammals larger than a gray kangaroo (60 kg); and, excepting the bush pig which may have been introduced, Madagascar lost all terrestrial vertebrates larger than the living Indri, a lemur which weighs 7–10 kg. Timing of the extinctions varies; it was earliest in Australia, later in South (and North) America, and latest in Madagascar. Although Australian small mammals (0.5–2.0 kg) suffered historic losses, there is no indication that severe late Quaternary extinction of small mammals accompanied the loss of large ones on the continents.

1. South America

The Quaternary of South America includes the Ensenadan and Lujanian land mammal ages dated at 1.5 to 0.5 and 0.5 to 0.01 Ma, respectively. The preceding Uquian land mammal age from 2.5 to 1.5 Ma is of great interest because it immediately follows the intercontinental suturing of North and South America. The Uquian shows the effects of the famous Great American Faunal Interchange, a natural experiment in which two continents exchanged long-isolated terrestrial faunas across a new land bridge. The land bridge connecting the continents at the end of the Pliocene brought Northern Hemisphere carnivores, gomphotheres, artiodactyls, and perissodactyls into contact with South American endemic orders, the xenarthrans, notoungulates, and litopterns. Interestingly, millions of years before the intercontinental conjunction, megalonychid ground sloths managed to colonize both the Greater Antilles and North America.

Occasionally, biogeographers have sought to invoke intercontinental exchange as a cause of the LQEE. A detailed summary of the South American fossil faunas by Marshall and Cifelli (1990) shows that although small-mammal extinctions occurred in the Uquian, heavy megafaunal extinction did not occur until the Lujanian, approximately 2 million years after the exchange had begun and apparently coeval with the LQEE in North America.

South American extinctions of the late Quaternary...
were even heavier than those in North America north of Mexico, involving 50 genera of which 35 belong to extinct families. The losses were differential and involved all megaherbivores such as giant ground sloths, glyptodonts, gomphotheres, and two endemic orders, Litopterna and Notoungulata (Table II). It is not obvious that the interchange had anything to do with the LQEE.

Defensible radiocarbon dates on most of the South American fauna remain to be assembled and may be difficult to obtain given the poor preservation of bone collagen in the open sites where Lujanian faunas are often found. One source of ideal material for radiocarbon dating—dung, hair, and perishable tissue—has been obtained on mylodontid ground sloths from Cueva del Mylodon in southern Chile. Fifteen dates from 13,000 radiocarbon years before present (yr bp) to 10,600 yr bp support the extinction chronology for ground sloths in North America.

2. Australia

Flannery and Roberts (cited in MacPhee, 1999) report that near the time of first contact (FC) Meganesia (Australia) lost approximately 28 genera and 55 species of large vertebrates (body weights exceeding 10 kg). The more important extinct genera of mammals in the Australasian (Australia and New Guinea) extinct megafauna are all marsupials. They include the extinct “lion” Thylacoleo; the wombats Phascolomys, Phascolonus, and Ramsay; a Palorchestid, Palorchestes; four diprotodontids—Diprotodon, Eucada, Nototherium, and Zygomaturus; a potoroid, Propleopus; and five giant kangaroos—Fissistrix, Procoptodon, Protemnodon, Sthenurus, and Troposodon. In addition, there was an extinct giant flightless bird, the mihirung (Genyornis); a giant lizard, Megalania; an extinct horned turtle, Melocharina; and a giant python, Woumba. Although older, the pattern of extinction in Australia resembles that in the Americas and Madagascar. In all three landmasses kill sites or processing sites are few, in dispute, or absent. As elsewhere, the lack of kill or butchering sites has deterred many Australian archeologists and paleontologists from invoking an overkill. Although no extinctions of Australian mammals less than 10 kg are known at FC, a sizable number of medium-size mammals are threatened or have vanished in historic time, presumably a side effect of European settlement in Australia (Flannery, 1994).

Despite the problem in direct radiocarbon dating on extinct fauna, the absence of extinct species from well-dated Tasmanian archeological sites up to 25,000 years in age supports the view that the Australian LQEE was over by then. The Franklin River region in Tasmania serves as an example. As in other parts of eastern Australia, Tasmania once harbored a variety of extinct species of giant marsupials. Although 21 dated late Pleistocene

<table>
<thead>
<tr>
<th>Endemics</th>
<th>Order Artiodactyla</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family Dasydopidae</td>
<td>Order Carnivora</td>
</tr>
<tr>
<td>Eutatus</td>
<td>Family Canidae</td>
</tr>
<tr>
<td>Holmesina</td>
<td>Thelotrolidae</td>
</tr>
<tr>
<td>Panopatherium</td>
<td>Family Felidae</td>
</tr>
<tr>
<td>Proprocaeus</td>
<td>Smilodon</td>
</tr>
<tr>
<td>Family Glyptodontidae</td>
<td>Family Ursidae</td>
</tr>
<tr>
<td>Chlamysotherium</td>
<td>Arctodus</td>
</tr>
<tr>
<td>Doedicurus</td>
<td>Order Proboscidea</td>
</tr>
<tr>
<td>Glyptodon</td>
<td>Family Gomphotheriidae</td>
</tr>
<tr>
<td>Hoplophorus</td>
<td>Cuvieronius</td>
</tr>
<tr>
<td>Lamaphorus</td>
<td>Hapalemuridae</td>
</tr>
<tr>
<td>Neotheracosaurus</td>
<td>Notiomastodon</td>
</tr>
<tr>
<td>Nesoeucalyptus</td>
<td>Stegomastodon</td>
</tr>
<tr>
<td>Panochthus</td>
<td></td>
</tr>
<tr>
<td>Paraprotocetus</td>
<td></td>
</tr>
<tr>
<td>Platybolus</td>
<td></td>
</tr>
<tr>
<td>Sclerochelys</td>
<td></td>
</tr>
<tr>
<td>Family Megalonychidae</td>
<td></td>
</tr>
<tr>
<td>Nothrocti</td>
<td></td>
</tr>
<tr>
<td>Neothereuctis</td>
<td></td>
</tr>
<tr>
<td>Octopus</td>
<td></td>
</tr>
<tr>
<td>Valgipes</td>
<td></td>
</tr>
<tr>
<td>Family Megatheriidae</td>
<td></td>
</tr>
<tr>
<td>Eremotherium</td>
<td></td>
</tr>
<tr>
<td>Megatherium</td>
<td></td>
</tr>
<tr>
<td>Paramegatherium</td>
<td></td>
</tr>
<tr>
<td>Family Mylodontidae</td>
<td></td>
</tr>
<tr>
<td>Glossotherium</td>
<td></td>
</tr>
<tr>
<td>Lestodon</td>
<td></td>
</tr>
<tr>
<td>Mylodon</td>
<td></td>
</tr>
<tr>
<td>Scelidodon</td>
<td></td>
</tr>
<tr>
<td>Sclidotherium</td>
<td></td>
</tr>
<tr>
<td>Order Perissodactyla</td>
<td></td>
</tr>
<tr>
<td>Family Equidae</td>
<td></td>
</tr>
<tr>
<td>Equus</td>
<td></td>
</tr>
<tr>
<td>Hippidion</td>
<td></td>
</tr>
<tr>
<td>Onohippidium</td>
<td></td>
</tr>
</tbody>
</table>

sites in caves or rock shelters in southwestern Tasmania ranging in age from 10,000 to at least 25,000 yr bp have yielded hundreds of thousands of animal bones. All are of living species and mainly of Bennett’s wallaby, Macropus rufogriseus (J. Allen as cited in Kirsch and Hunt, 1997). If the extinct mihirung, giant macropods, diprotodonts, etc. were still alive 25,000 years ago or later, it is difficult to believe that they would not be represented in such rich zooarchaeological material. This negative evidence supports the view that extinctions occurred earlier.

Until more fossil faunas are dated, there will be unavoidable uncertainty about when extinction occurred. In this regard, new environmental information from stable isotopes from samples of Genyornis eggshell is especially promising. The record to date indicates extinction of Genyornis at by 45,000 yr bp in several parts of its wide range and in more than one climatic province; these findings argue against extinctions driven by some climatic bottleneck (Miller et al., 1999). In Australia extinctions long predate any hypothetical late glacial climatic forcing.

Finally, in terms of human origins, a comparison between Australia and America is instructive. In the New World archeologists have seen a variety of sites claimed to be 13,000–22,000 years or older fail to be verified or locally replicated, despite regular assertions of their proponents. In contrast, in the past two decades dozens of Australian sites have repeatedly yielded geochemical dates indicating dozens of sites older than Clovis in North America.

Given the smaller area of Australia compared with that of the New World, its much less productive soils and much more variable precipitation (Flannery, 1994), and the much smaller number of archeologists, geologists, paleontologists, and amateurs searching for artifacts and fossils, the abundance of sites 10,000–40,000 years old and older in Australia, compared with the half a dozen claims of pre-Clovis sites in the New World and their debatable status, is a red flag to environmentalists. Unless proposed early sites in the Americas are critically replicated by the geoarcheological community at large, as in Australia, the claims for a pre-Clovis occupation and a pre-Clovis culture, no matter how detailed or how often repeated, remain in limbo (Martin and Steadman, 1999).

3. Madagascar

Within the past two millennia, at a time when no other continents suffered appreciable extinctions of large mammals (Martin and Klein, 1984), Madagascar lost many large mammals, including 20 species of lemurs up to the size of a gorilla. All are larger than the largest living lemur. In addition, Madagascar lost two hippos; the highly endemic antelope known as bibymalagasy (Plesiorycteropus), recently assigned to its own order; a large (bobcat-size) species of carnivore (Cryptoprocta spelea); a rabbit-size rodent (Hypogeomys australis); two genera of giant flightless birds; and two giant tortoises, one exceeding 100 kg in mass (MacPhee and Marx as cited in Goodman and Patterson, 1997). Although most if not all of these extinctions in Madagascar coincide with human colonization, none have been found in well-defined kill sites, leading MacPhee and Marx to model hyperdisease. They generalize their model to include all lands of prehistoric human colonization. Despite its implausibility, the model invites testing since the fossil record has yet to reveal how LQEE occurred. In the Americas the lack of many Clovis kill sites may be overcome by modeling “blitzkrieg,” a rapid elimination of preferred prey, leaving minimal field evidence, by a potent human predator (Martin, 1990).

If aridification contributed to extinctions in Madagascar, it was evidently a potential problem only in the southwest. Sudden overkill by a “blitz” sweep of the vulnerable megafauna, seems unlikely if the start-up of colonization took more than 1000 years. Robert Dewar (as cited in Goodman and Patterson, 1997) suggested that the impact of wild cattle on the native fauna rather than human hunting forced the extinctions.

All this may be resolved eventually by enrichment of the stratigraphic and geochronological record, including a major effort at extending the radiocarbon chronology of the extinct Malagasy megafauna into glacial times to determine if late glacial climatic change forced extinctions long in advance of human presence. Radiocarbon dates of 13,000 and 26,000 yr bp on a large (~70-kg) extinct lemur, Megaaladapis (Simons as cited in Goodman and Patterson, 1997), indicate that glacial-age faunas exist. Apart from Megaaladapis, virtually nothing is known of the glacial-age faunas of Madagascar. Like other extinct genera, Megaaladapis survived until the end of the Holocene and may have been known historically.

Unlike Madagascar, rich glacial-age faunas are known in New Zealand, Tonga, and Hawaii. To date, no glacial-age extinctions have been detected in these islands, strengthening the case for an anthropogenic agency as the unique cause of LQEE (Martin and Steadman, 1999).

Perhaps the most interesting case of comparative extinctions is the global decline of giant tortoises. Giant tortoises can be regarded as extremely vulnerable to human foragers. They survived historically only on re-
mote warm-water oceanic islands that escaped prehistoric colonization, such as the Mascarenes, Seychelles, and Galapagos. In the late Quaternary in Madagascar, giant tortoises vanished with the giant lemur. In Africa they disappeared much earlier, perhaps more than 2 million years ago, an event that may represent the earliest example of an anthropogenic forcing of megafaunal extinctions.

C. Eurasia and Africa

Compared with the sweeping, catastrophic loss of late Quaternary megafauna in the New World, Madagascar, and arguably Australia, the losses in Eurasia were relatively minor and they occurred gradually (Fig. 7; Martin and Steadman, 1999; Stuart, 1991). Outside the tropics, the extinct Eurasian fauna can be divided into two parts. One was a “warm” or interstidal fauna with straight-tusked elephants (Elephas [Palaeeleaxodon] antiquus), temperate rhinoceroses (Stenorhinus hemitoecus and S. kirchbergenis), hippo (Hippopotamus amphibius), and a cave bear (Ursus spelaeus). The other was a “cold” or stadal fauna with woolly rhinoceroses (Coelodonta antiquitatis), woolly mammoth (Mammuthus primigenius), and musk ox (Ovibos moschatus). This stepwise sequence of change in temperate and boreal faunas of Eurasia is gradual compared with the sudden extinction of both cold- (woolly mammoths and woodland musk oxen) and warm-adapted species (tapir and jaguar) in North America.

Intriguing differences in the pattern of Old and New World extinction are best illustrated in the case of proboscideans. Extinction of straight-tusked elephants did not begin until early in the last cold stage of the late Quaternary. Extinct throughout their European range by 60,000 years ago (Stuart, 1991), Elephas persisted in Japan until 30,000 years ago or less. The largest Holocene survivors in the megafauna of Japan were Sika deer (Cervus), wild boar (Sus), and black bear (Selenarctos). Although dwarf relatives of Elephas evidently did not survive on Mediterranean islands as late as the Holocene, as was once thought (Martin and Klein, 1984), mid-Holocene woolly mammoths have been found on Wrangel Island off Siberia, 6000 of years younger than any other known mammoths—contemporary with the Pharaohs. The Akrotiri Aetokremnos bone cave, an epipaleolithic site on Cyprus, yielded dwarf straight-tusked elephants along with abundant bones of dwarf hippo. The deposit is radiocarbon dated at 10,500 years old. Just as Wrangel Island served as a refugium for the last of the boreal woolly mammoths, thousands of years after they were extinct throughout the rest of their range, Japan and Cyprus apparently sheltered the last populations of temperate Eurasian Elephas long after its extinction on the continent (Martin and Steadman, 1999).

Unlike the slow pace of extinction of the Proboscidea in northern parts of Eurasia, in North America prior to an extinction spasm 10,000–11,000 yr bp the animals suffered no loss of taxa, no clear-cut reduction in range, and no apparent decline in numbers (see maps in Graham and Lundelius, 1994). Had American Proboscidea followed the Old World extinction pattern for elephants, the temperate elephant of midlatitudes, the Columbian mammoth, would have predated high-latitude extinction of woolly mammoths by tens of thousands of years while the order would endure in tropical America, represented by a living species.

Both ivory carvings and cave paintings depicted by Stone Age artists reveal their intimate knowledge of the anatomy and behavior of Old World woolly mammoths and many other animals. Although field evidence of mammoth kills is scarce, low rates of increase of modern megaherbivores means that modern elephants, rhinoceroses, and hippo would have been highly vulnerable to human predation (Owen-Smith, 1988). Even with governmental protection in this century, the future of surviving megaherbivores is by no means ensured. The gradual decline of Eurasian mammoths could readily have resulted from a minimal amount of human hunting or interference, especially if younger age classes were targeted.

Although the nature of the association and its meaning in modeling insular faunal extinctions in the Mediterranean are uncertain, the record at Akrotiri Aetokremnos suggests a Paleolithic “commando raid” in which Stone Age hunters or foragers overran an island, slaughtered preferred prey (dwarf hippo, dwarf elephant, and endemic deer), did not find sufficient resources for a sustainable economy, and soon left with no other evidence of their passage besides the remarkably rich contents of one cave. Possibly Crete, Ireland, Wrangel Island, and the Channel Islands of California also lost megafauna to commando raids that left scant evidence or at least none discovered to date. Given the scarcity of convincing kill or processing sites on continents, the rarity of such features on offshore islands is not surprising. Only in the remote Pacific such as on the low limestone islands in Tonga are extinct animals (bones of megapodes) abundantly associated with artifacts of the first colonists. In other cases (Hawaii and Madagascar), such associations are rare or unknown (MacPhee, 1999).

Africa has been held up to modelers of overkill as a
FIGURE 7  Extinction or extirpation chronology of large mammals in northern Eurasia. Note how the inferred extinction dates are very staggered compared with the strong pulse of extinctions in North America (right). During at least 50,000 yr BP the loss of Eurasian large mammals was sequential; the loss in North America was sudden. NE, northern and central Europe; It, Italy; Ja, Japan; SF, southern France; Sp, Spain; Lv, Levant; E, England; Sc, Scandinavia; Ir, Ireland; NS, north-central Siberia; Elephas (Palaeoloxodon) antiquus, straight-tusked elephant; Hippopotamus amphibius, hippopotamus; Stenotherus hemitoechus and/or S. kitchingensis, temperate rhinoceroses; Ursus spelaeus, cave bear; Crocuta crocuta, spotted hyena; Coelodonta antilope, woolly rhinoceros; Megaceros giganteus, giant deer; Mammuthus primigenius, woolly mammoth (after MacFhee 1999) and Stuart 1991).

conundrum. If people are such effective exterminators, how did the African megafauna survive? However, the argument cuts two ways. If late Pleistocene climatic change played a major role in triggering megafaunal extinctions elsewhere, how did so many large animals manage to survive in Africa and Asia, continents no more immune to Pleistocene climatic change than other corners of the globe?

One reply to the conundrum is that in Africa and Asia there was sufficient time to evolve a balanced predator–prey relationship, perhaps with human populations locally suppressed by sleeping sickness and other
D. Oceanic Islands

Late Quaternary losses were limited neither to continents nor to mammals. They swept through oceanic islands. For example, on the larger oceanic islands the extinction of large terrestrial vertebrates included flightless moas in New Zealand, gorilla-size extinct lemurs, extinct hippo, giant tortoises, and giant flightless birds such as <i>Aepyornis</i> in Madagascar; house cat to bear-size ground sloths in the West Indies; dwarf elephants and dwarf hippo on islands of the Mediterranean; and dwarf elephants and/or giant tortoises on oceanic islands beyond the continental shelf in southeastern Asia such as Timor, Flores, and Sulawesi. In one unusual case, prehistoric extinction only 4000 years ago eliminated elephants (woolly mammoths) not from a deep-water island but from a shelf island, Wrangel, in the Arctic Ocean (Fig. 2).

With their discovery 25 years ago on Hawaii of a flightless goose (<i>Thambetochen</i>) and a flightless ibis (<i>Aptenibis</i>), ornithologists Alexander Wetmore and Storrs Olson triggered the search for unknown extinctions on Pacific archipelagos. In the following years, archeologists and paleoecologists began to uncover rich fossil faunas reflecting prehistoric (Holocene) extinctions. During the past two decades their efforts on the Cooks, the Marquesas, the Kingdom of Tonga, and the Line Islands, to name a few examples from the South Pacific, have yielded many extinct species of small birds, especially flightless rails, pigeons, and parrots (Steadman, 1995; Steadman as cited in Kirch and Hunt, 1997).

On the basis of fossil deposits from a sample of the 800 islands in the Pacific that are more than 1 km² in area, Steadman (1995) estimates the loss of 10 species or populations from each—a total loss of 8000 species or indigenous populations of land birds (including an estimated 2000 endemic species of taxonomically distinctive flightless rails). In addition to human harvesting and possible surplus killing, the introduction of Pacific rats may have eliminated many flightless rails and other birds. In some cases, the rats may even have colonized in advance of their human vectors (Holdaway as cited in MacPhee, 1999).

Extensive colonies of petrels, shearwaters, and other seabirds vanished from many Pacific islands, dimming the avian beacon which helped guide the prehistoric explorers to remote uninhabited islands (Steadman as cited in Kirch and Hunt, 1997). The fossil records of the relatively small number of islands that experienced FC and heavy extinction within historic time, such as the Galapagos, the Mascarenes, and remote Atlantic Islands such as St. Helena, have yet to reveal any pulse of prehistoric (Holocene) extinctions to match those found on islands that were colonized prehistorically.

New Zealand provides the best fossil record of island extinctions, starting with the loss of flightless wrens and small petrels, which were eliminated by accidental Polynesian introduction of Pacific rats (<i>R. exulans</i>), and beginning perhaps 1000 years in advance of Polynesian settlement. After Captain Cook's arrival in 1770, more species of mammalian predators reached New Zealand, including the dog, Norway rat, feral pig, feral cat, house mouse, black rat, ferret, and stoat (Holdaway as cited in MacPhee, 1999).

To bring prehistoric extinctions into focus, it is necessary to scan the globe and to probe the magnitude, timing, pattern, and natural history of all prehistoric losses on all landmasses, continents, and islands. This simple comparative approach in paleontology opens new vistas—ones that the popular MacArthur–Wilson model of island biogeography, based on modern distributions, will not reveal. High rates of vertebrate extinction on oceanic islands in the past five centuries, the focus of attention by conservation biologists (MacPhee and Flemming as cited in MacPhee, 1999), pale in magnitude to the extinction spasm of the previous four millennia.

Can extinctions of the LQEE be reversed? They certainly can if we do not insist on replacing lost species with taxonomically identical populations. On oceanic islands restoration ecologists need to consider restarting evolution of lost lineages of rails, using Guam rails on islands once occupied by flightless rails. In North America, a Pleistocene park should minimally include proboscideans, camelids, and equids as well as bison and other historic megafauna that survived the LQEE.

III. PATTERN AND CAUSE

Through an understanding of the pattern and timing of late Quaternary extinctions it may be possible to establish their cause. As the chronology, stratigraphy,
paleoecology, and global pattern of the extinctions become better known, the ultimate questions of rate are clarified. For example, Darwin's view in Origin of Species that prehistoric extinctions of large mammals occurred gradually and sequentially is supported by the LQEE in Eurasia (Fig. 7). Although an extraterrestrial accident such as an asteroid or cometary impact is often linked to some earlier mass extinctions, especially at the end of the Cretaceous, the space rock scenario can be ruled out in the late Quaternary. Over radiocarbon time the global pattern of loss was sequential or time transgressive, from more than 30,000 years ago in Australia to 1000 years ago in New Zealand (Figs. 1 and 2), virtually eliminating any possibility of a one-shot global climatic catastrophe, such as a unique lethal cold snap, hyperdrought, or a great flood.

The prehistoric extinctions reviewed here are increasingly suspected of being the preamble of a vastly larger number to follow, a true mass extinction event in the making. Its cause would be anthropogenic.

A. Anthropogenic Models

The two major competing hypotheses of models to LQEE involve either anthropogenic or climatic forcing or both in combination (Binney, 1993; MacPhee, 1999; Martin and Klein, 1984). The anthropogenic model includes a variety of possible human impacts: direct overkill, surplus killing, a predator pit (doomed but still functioning predators, such as saber cats, reinforcing human predation), introduction of pandemic disease (“hyperdisease”), and human-initiated changes in habitat. The anthropogenic model is based on the close timing of extinctions to the global spread of anatomically and behaviorally modern people. In addition, some propose that Neanderthals or earlier hominids also had a role both in shaping the evolution of large mammal communities in the Old World and in the extinctions of the most vulnerable prey species providing rich resource packages such as giant tortoises, animals whose antipredator strategies were ineffective against humans. The kinds of animals lost in the LQEE appear to represent species known to be or likely to have been preferred prey.

A new player among the possible contenders for anthropogenically driven LQEE pulses is infectious disease, brought in at FC by human colonists themselves or, more likely, everything and anything that came along with them, such as domesticated, commensal, and synanthropic species. Especially on oceanic islands, the impact of human colonists would be amplified by the introduction of aliens, such as the Pacific rat R. exulans, chickens, domestic pigs, and dogs, along with their exotic diseases. The logic of hyperdisease is similar to that of classic overkill, but it contemplates humans as passive rather than active agents in causing faunal crashes. The basic idea is that “emerging” diseases, introduced into species without any natural immunity to them, would have induced incredible levels of mortality in susceptible populations, conceivably leading to their extinction. Once a “new” infectious disease process got started in a susceptible population, it would run its course quite independently of the rate or direction of human expansion.

Critics of anthropogenic models note that kill sites indicating human predation on and processing of allegedly preferred prey are rare, with few in the Americas and none in Australia or Madagascar and few in the Americas. The case of Madagascar is especially curious because its losses occurred only within the past 2000 years. Skeptics challenge the chronology of extinction, denying that within North America or Australia all extant large mammals vanished abruptly, on the heels of human colonization, as implied by overkill. Some contend that humans arrived in both continents significantly before extinctions occurred. If human arrivals triggered megafaunal extinctions in the New World and Australia, critics ask, how do we explain the coexistence of Homo sapiens and large mammals in Africa and tropical parts of Eurasia? Vertebrate paleontologists have traditionally appealed to climatic or environmental changes as the main forcing function for many mammalian extinctions. During the past 65 million years the vast majority of extinctions occurred before any possibility of human involvement.

B. Climatic Models

Climatic models are based on the highly variable nature of late Quaternary environments with rapid switches from cold or cool and dry to warm and wet accompanied by changes in CO₂ pressure. In some cases, such as the Allerød–Younger Dryas shift within the late glacial, the switch appears to coincide with megafaunal extinctions, especially in North America. At least one habitat, the steppe–tundra of polar regions in the Northern Hemisphere, has been identified as an extinct biome (Guthrie as cited in Martin and Klein, 1984) whose end doomed woolly mammoths and other subarctic megafauna. Some LQEE models invoke a switch from less extreme to more extreme seasonality with out-of-step breeding cycles eliminating ruminants whose life cycles could not accommodate the climatic changes (Klit and Graham and Lundelius as cited in Martin and Klein, 1984).
Critics discount late Quaternary climatic change as a forcing function since the paleoclimatic record of the Quaternary is rich in rapid and severe changes long before as well as during episodes of extinction. Fewer mega faunal extinctions occurred in the 3.5 million years combined, before human arrival, than in the late Quaternary, suggesting that species in Quaternary biotas were buffered against environmental switches typical of the last ice age. In any case, a late Quaternary extinction spasm is not evident in the oceans or in freshwater habitats and involves small animals only as parasites or on oceanic islands on which the reduced area would have amplified human impacts. Finally, the character of LQEE universally points to species that would either be preferred prey for human foragers and predators or be vulnerable (as in the case of minute endemic island snails) to aliens introduced from the continent.

Although details of how prehistoric humans might have triggered prehistoric extinctions are not easy to interpret from the fossil record, on a global scale the LQEE strongly reflects a deadly syncopation between extinction and human arrival. Viewed from the Cenozoic, Alroy (as quoted in MacPhee, 1999, p. 105) observed,

The event's timing, rapidity, selectivity, and geographic pattern all make good sense according to the anthropogenic model and no sense at all otherwise. I believe that the overkill hypothesis, at least in general terms, already has been "proven" as thoroughly as any historical hypothesis can be.

See Also the Following Articles

EXTINCTION, CAUSES OF • MAMMALS, BIODIVERSITY OF • MAMMALS, PRE-QUATERNARY, EXTINCTIONS OF • MASS, CONCEPT OF • MASS EXTINCTIONS, NOTABLE EXAMPLES OF •

Bibliography

MAMMALS
(PRE-QUATERNARY),
EXTINCTIONS OF

W. A. Clemens
University of California, Berkeley

I. Mammals and Other Synapsids
II. Patterns and Causes of Extinction
III. Origin of Synapsids and the Permian–Triassic Mass Extinction
IV. The Second "Therapsid" Radiation and Late Triassic Faunal Turnover
V. Mammalian Evolution during the Age of Dinosaurs
VI. Cretaceous–Tertiary Mass Extinction and Its Consequences
VII. Recovery of Mammalian Faunas during the Early Paleocene
VIII. Mammalian Evolution during the Tertiary
IX. Conclusions

GLOSSARY

background extinction Rates of extinction that characterized the major part of the evolution of life.
clade A group of organisms including its common ancestor and all its descendants.
crown group A clade including a group of modern species, the common ancestor, and all its descendants, including extinct lineages.
extinction The termination of a lineage of organisms.
mass extinction A short period of Earth history in which the rate of extinction reached an exceptionally high level.

paraphyletic A group of organisms that includes the common ancestor but not all of its descendants.

EXTINCTION HAS PLAYED A major role in shaping the course of mammalian evolution from the differentiation of their early synapsid ancestors, during the interval between 354–290 million years ago, up to the Quaternary, approximately 1.8 million years ago. Throughout their history, relatively low rates of extinction contributed to a continuing background of evolutionary change. Major decreases in diversity occurred at times of mass extinctions, when many lineages were terminated during relatively short periods of time, or during periods of longer duration characterized by high levels of faunal turnover. Such major events divide mammalian evolution into three phases, which will be analyzed after a brief introduction to the history of the synapsids and the general patterns and causal factors of extinction.

I. MAMMALS AND OTHER SYNAPSIDS

The evolutionary history of mammals and their ancestors, which constitute the Synapsida, began in the Carboniferous Period approximately 354–290 million years ago (Ma) (Table I). During this interval terrestrial vertebrates underwent their first evolutionary radiation. One of these lineages, characterized by a specialized type of egg (the amniote egg), split into two clades—
TABLE I
Geological Timescale for the Interval of Earth History
Considered in the Text

<table>
<thead>
<tr>
<th>Era</th>
<th>Period</th>
<th>Epoch</th>
<th>Age of boundary (millions of years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quaternary</td>
<td>Pleistocene</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pliocene</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>Cenozoic</td>
<td>Tertiary</td>
<td>23.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oligocene</td>
<td>33.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eocene</td>
<td>54.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paleocene</td>
<td>65.0</td>
<td></td>
</tr>
<tr>
<td>Cretaceous</td>
<td></td>
<td>144</td>
<td></td>
</tr>
<tr>
<td>Mesozoic</td>
<td>Jurassic</td>
<td>206</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Triassic</td>
<td>248</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Permian</td>
<td>290</td>
<td></td>
</tr>
<tr>
<td>Paleozoic</td>
<td>Carboniferous</td>
<td>334</td>
<td></td>
</tr>
</tbody>
</table>

* Data taken from the 1999 Geological Time Scale, Geological Society of America.

reptiles (including birds) and synapsids (including "nonmammalian synapsids" and crown group mammals; Fig. 1). Early members of these clades are distinguished by different structural patterns of the posterior parts of their skulls. Although clearly part of the ancestry of mammals, early nonmammalian synapsids lacked specialized features of skull structure that characterize members of the crown group Mammalia—for example, only one pair of bones (the dentaries) making up the lower jaw or the presence of three bones (malleus, incus, and stapes) in the middle ear. Direct evidence is not preserved in the fossil record, but it is probable that these early synapsids lacked hair, mammary glands in the females, and endothermy—traits commonly used to distinguish modern mammals from reptiles.

Unfortunately, in many older and some recent texts, early synapsids are referred to as "mammal-like reptiles." These early synapsids definitely were not members of the ancestry of reptiles. Morphological similarities to reptiles are the product of inheritance and retention of traits found in their common ancestor. In the following discussion, these early synapsids will be dubbed "nonmammalian synapsids," which accurately expresses the fact that they were members of the early part, the stem, of the synapsid clade but not members of the crown group Mammalia (Fig. 1).

In a broad-brush summary, the evolutionary history of the synapsids can be divided into three phases. Beginning in the Carboniferous, the first evolutionary radiation of "nonmammalian synapsids" was interrupted by the mass extinction event used to mark the Permian–Triassic boundary. The surviving lineages of "nonmammalian synapsids" radiated again and regained their position as a dominant group of terrestrial vertebrates. Prior to the next mass extinction marking the Triassic–Jurassic boundary (approximately 206 Ma), almost all lineages of "nonmammalian synapsids" became extinct. The following Jurassic and Cretaceous Periods, which lasted from approximately 206 to 65 Ma, constitute the "Age of Dinosaurs." Within this interval all surviving lineages were limited to animals of small body size, but it was a time of major evolutionary radiation that included the origin of the crown group Mammalia. The Cretaceous ended with another mass extinction event; this was the time of extinction of dinosaurs other than birds. During the following Tertiary Period (65–1.8 Ma), mammals flourished and their diversity rapidly increased. Biogeographic differentiation, extinction events, and periods of faunal turnover complicated Tertiary evolutionary radiations of mammals. Changes in global climate appear to have been among the causal factors of these extinctions and faunal turnovers, which led to the origin Quaternary mammalian faunas.

II. PATTERNS AND CAUSES OF EXTINCTIONS

Although discussed in much more detail later, a few issues relating to the nature and causal factors of extinctions need to be reviewed here as background for analy-
sis of the role of pre-Quaternary extinctions in shaping the course of mammalian evolution.

A. Extinctions and Pseudoextinctions

In any attempt to assess changes in biodiversity resulting from extinctions, care must be taken to differentiate between true, biological extinctions and what have been termed pseudoextinctions. In a biological sense, extinction is the termination of the group or groups of organisms comprising a species. No longer is there successful reproduction and survival of the young to sexual maturity, population size dwindles to zero, and the species, as well as the lineage it represents, is lost from the biota.

Pseudoextinction refers to an artifact produced by zoological nomenclature; a named unit becomes extinct but the lineage it represents survives. The taxonomy (the system of recognition and naming of groups of organisms) of the Equidae, the family that includes modern horses, asses, and zebras, provides many examples of pseudoextinction. The oldest known species of the family are a group of lineages included in the genus "Hyracotherium" (= "Eohippus"). These were diminutive, multitoeled inhabitants of tropical forests. The fossil record documents increase in body size, loss of toes, and other significant morphological changes in descendants of "Hyracotherium"—for example, "Orohippus," "Miocippus," "Merychippus," and "Dinohippus." These extinct genera are members of the ancestry of modern horses, zebras, asses, and other closely related species, all of which are included in the genus Equus.

The genus "Hyracotherium" is composed of the common ancestor of all the species referred to "Hyracotherium," including the lineage leading to "Orohippus," the next member of the equid family tree. In the jargon of modern taxonomy, "Hyracotherium" is paraphyletic, a unit that includes its common ancestor and some but not all of the common ancestor's descendants; in this case, "Orohippus" is excluded. All genera of the family tree ancestral to Equus are now extinct. The extinctions of these paraphyletic groups, however, did not mark the termination of the equid lineage; they are products of the system of recognizing and naming organisms. (To highlight their status, the names of paraphyletic groups are enclosed in quotation marks.)

Current studies of patterns of extinction usually do not differentiate between biological (lineage) and pseudoextinctions. Therefore, the relative proportions of these two types of extinction in the fossil record are uncertain. Distinguishing between them might significantly modify some current interpretations of the patterns and processes of evolution.

It also must be stressed that in most studies of pre-Quaternary extinctions, species are not the basic units considered. Usually, these studies analyze the differences in numbers of genera or families, a procedure that opens the possibility of significantly muting patterns of change in biodiversity. For example, if a genus consists of only one species, loss of the species results in the loss of the genus. On the other hand, if the genus contains many species, the loss of one species is not reflected at the generic level. Thus, a major extinction event that resulted in the loss of 60% of the species extant in a region might be reflected in the loss of only 30% of the genera or 10% of the families represented.

B. Background and Mass Extinctions

Since the origin of life approximately 3.5 billion years ago, extinction has been a major process influencing the course of its evolution. Current estimates from the fossil record indicate that more than 90% of the species that once existed are now extinct; some suggest that the number is more than 99%. Many workers, particularly J. J. Sepkoski Jr., D. M. Raup, and their associates (Raup, 1991), have contributed greatly to our understanding of the contribution of extinction to patterns of change in diversity. The majority of these studies have been based on extensive databases of records of the times of origin and extinction of families and genera of marine invertebrates and vertebrates compiled by Sepkoski. Temporally, these records of extinctions (including both lineage extinctions and pseudoextinctions) are ordered on a scale of geological stages. These stages vary in length but have an average duration of about 6 million years, an interval about three times as long as the Quaternary. This difference in degree of resolution must be emphasized in attempts to apply the results of studies of pre-Quaternary extinctions to analyses of Quaternary, particularly modern, patterns of extinction.

Sepkoski's initial studies indicated that rates of extinction fell into two categories, which were termed background extinctions and mass extinctions. Background extinction rates appeared to characterize the majority of the history of life. Several times during the past 600 million years rates of extinction increased greatly over short periods of geological time; these were recognized as mass extinction events. Subsequent research led Raup to develop what he termed the "kill curve," which highlighted the fact that rates of extinction varied over a considerable spectrum with background extinctions and mass extinctions merely repre-
senting the extremes. His analysis, however, did not negate the observation that during the past 600 million years the pattern of evolution of marine organisms was greatly modified by mass extinctions.

The most severe of the mass extinction events frequently are identified as the "big five" mass extinctions. Although the available database for evolution of terrestrial organisms is not as detailed as that compiled by Sepkoski for the marine record, enough is known to show that some of these mass extinctions modifying the marine biota had significant impacts on the evolution of terrestrial life. In particular, the mass extinctions used to mark the Permain–Triassic and Cretaceous–Tertiary boundaries significantly influenced the patterns of evolution of "nonmammalian synapsids" and early mammals.

C. Temporal Scale of Mass Extinctions

Mass extinctions were intentionally defined subjectively as short periods of Earth history during which rates of extinction reached exceptionally high levels. In geological terms, how short is short? In studies of recent extinctions in which events are followed on an ecological timescale, short periods of time are measured in terms of years, decades, centuries, or possibly a millennium or two. Current radiometric methods for determining ages and durations of pre-Quaternary events lack this level of resolution. For example, using the $^{40}Ar/^{39}Ar$ method of age determination, which is the most precise method of radiometric age determination available for pre-Quaternary deposits, the age of the Cretaceous–Tertiary boundary is placed at 65.16 ± 0.04 Ma. As data on extinctions of Cretaceous lineages of marine and nonmarine organisms are collected and correlated, all extinction events that occurred within an interval of 80,000 years would necessarily be treated as having occurred simultaneously. The farther one goes back in Earth history, the longer the error bars become; for example, some paleontologists currently argue that the extinctions of lineages that comprise the Permain–Triassic mass extinction might have occurred over an interval of 1 or 2 million years.

D. Distal and Proximal Causal Factors

In studies of extinction it has proven useful to distinguish between distal and proximal causal factors. Proximal causal factors are those that impinge directly on individuals and are the immediate causes of their death. Distal causal factors are the regional or global events that result in changes in local (proximal) physical and biological environments affecting individual organisms. As a current example, the various sources of atmospheric pollution thought to contribute to global warming would be ranked as distal factors. The impacts of changes in annual temperature regime on individuals would be proximal factors.

In the much longer period of mammalian evolution prior to the Quaternary, several distal causal factors emerge as major contributory factors to mass extinctions and other significant changes in the course of evolution of the group. Recognition of some of these can be traced to the development of our understanding of plate tectonics, the mechanisms involved in movements and changes in configurations of continents and oceanic basins. Immediate consequences of these events have been the modification of patterns of circulation of oceanic currents and continental topography, which contributed to climatic changes. Hallam and Wignall (1997) and other workers have noted that many of the major mass extinctions were associated with times of withdrawal of shallow continental seas (marine regressions) and reduction of coastal habitats.

Another product of changes in configurations of continents during periods of marine regression is the establishment of terrestrial connections between long-isolated continents. Two recent examples of this process are the formation of dry land connections between Siberia and Alaska and between North and South America via the Panamanian Isthmus. Both facilitated mammalian interchange and extinction of some lineages, probably through some combination of competition, predation, and introduction of new diseases.

Volcanic activity has played and continues to play a significant role in modifying the earth's environments. Major eruptions produce large quantities of particulate matter, aerosols, and acids. These clouds can reach magnitudes sufficient to limit or block the sun's radiation from reaching the surface of the earth, causing a decrease in temperature. Later, they would trap heat radiated from the surface to produce a greenhouse warming of the climate. Increased amounts of acid rain also are predictable consequences of volcanic activity.

Although the hypothesis had precursors in earlier decades, the assertion by Alvarez et al. (1980) that the impact of an asteroid was the causal factor of the Cretaceous–Tertiary mass extinction focused attention on the potential role of extraterrestrial bombardment in controlling the course of evolution of the earth's biota. Their initial hypothesis argued that the impact of an asteroid was the distal cause of the extinction of nonavian dinosaurs and numerous other kinds of terrestrial and marine organisms. The hypothesized
proximal cause was the formation of dust clouds that first produced global refrigeration by blocking solar energy and then a greenhouse effect by trapping heat in the atmosphere. Acid rain and extensive wildfires also were suggested as contributory factors. Qualitatively, these are the same kinds of perturbations of the environment that could be generated by intense volcanic activity. As discussed later, the pattern of survival and extinction of mammals across the Cretaceous–Tertiary boundary has played a significant part in testing the impact hypothesis and other hypotheses concerning the causal factors of this mass extinction.

It has been argued that a mass extinction is an exceptional event and, therefore, requires an exceptional causal factor. Two aspects of this assertion require comment. Analyses of the geological and fossil records indicate that many of the potential distal causal factors, for example, regressions of the seas, outbursts of volcanic activity, and impacts of large extraterrestrial bodies, were contemporaneous with mass extinctions and lesser changes in the earth's biota. None of them has been demonstrated to be contemporaneous with all the major mass extinctions. Examples of each of these potential causal factors have been recorded at times when mass extinctions in the marine and/or terrestrial biota have not occurred. Now, some scientists suggest that the causes of mass extinction are to be found in combinations of these and possibly other factors that compounded to produce major, deleterious changes in the environment.

Second, the level of ecological interdependence of members of the terrestrial biota has, no doubt, varied through time. Terrestrial ecosystems characterized by relatively low levels of interdependence of their members might well be able to withstand the loss of one or a few species. In contrast, ecosystems with highly interdependent members might not be able to withstand the loss of one or a few species, resulting in the collapse of the whole system and consequent extinction of many other species. Thus, a period of global cooling, for example, could have a catastrophic effect on highly interdependent ecosystems but cause little or no change in those that were characterized by a lower level of interdependence.

Finally, the focus of this article is extinctions of lineages and their effects on patterns of evolution and diversity. Extinction is but one-half of the equation for calculating changes in biodiversity. Variation in the rate of origination of new species can have a marked effect on biodiversity. For example, analysis of the fossil record of nonavian dinosaurs during the Jurassic and Cretaceous shows that they evolved rapidly; a genus of nonavian dinosaur rarely lasted more than a geological stage (on average 6 million years in duration) or two. Throughout most of their history, rates of extinction and origination of nonavian dinosaurs were high. In the last two stages prior to their extinction at the end of the Cretaceous, a high rate of extinction of genera appears to have been maintained, but the rate of origination of new genera was depressed.

III. ORIGIN OF SYNAPSIDS AND THE PERMIAN–TRIASSIC MASS EXTINCTION

During the Carboniferous Period (Table 1), approximately 354–290 Ma, the reptilian and synapsid clades differentiated. From this beginning, the first phase of synapsid evolution, characterized by their prominent position in terrestrial faunas, would last until the end of the Triassic. It was interrupted, and the course of synapsid evolution altered, by the massive extinctions that mark the Permian–Triassic boundary.

In the Early Permian, basal "nonmammalian synapsids," the "pelycosaurs," were the most diverse terrestrial vertebrates. Notable within the evolutionary radiation of "pelycosaurs" were two lineages that evolved long extensions of the neural spines of their vertebrae, which supported high dorsal fins. These fins, which greatly increased the surface area of the body but not its volume, have been interpreted as early experiments in thermoregulation. Exposed to the sun, the fin would have sped up heating of the body; in the shade it would have served as an efficient radiator of heat.

The "pelycosaur"-dominated terrestrial faunas of the Early Permian were composed of many more carnivores than herbivores; at first glance, this appears to be an ecologically anomalous situation. It is now thought that the dominance of carnivores is not a product of a bias in preservation or collecting but reflects an initial stage in the evolution of the terrestrial ecosystem. Many of the species of these early "nonmammalian synapsids" may well have been amphibious, returning to streams and lakes to feed on fish and freshwater invertebrates. In about the middle of the Permian, the composition of terrestrial faunas changed with the appearance of a diversity of "therapsids" (Fig. 1). Herbivorous forms, including the dicynodonts, were much more abundant, which suggests that fully terrestrial ecosystems had evolved.

The mass extinction used to mark the end of the Permian was the most severe yet recorded. It has been
estimated that more than 90% of the species and 50% of the families of marine invertebrates were lost during this event. Terrestrial ecosystems also were greatly modified by extinctions. Of the approximately 50 families of terrestrial vertebrates present in the past 5 million years of the Permian, approximately 75% died out. This dramatic loss included large and small herbivores and carnivores. Notable survivors among the “nonmammalian synapsids” were a few lineages of carnivorous “therapsids” and the herbivorous dicynodonts.

Debate continues over the causal factors of the mass extinction marking the end of the Permian (Hallam and Wignall, 1997). A variety of lines of evidence suggest that this was a time of global warming, probably related to a marine regression followed by development of anoxic oceans. Extensive volcanic activity, the eruption of the massive flood basalts of western Siberia, was probably another distal causal factor. Currently, students of the Permian–Triassic mass extinction appear to favor some combination of changes in the oceans and volcanic activity as its distal causal factors. No evidence of the impact of an asteroid or other extraterrestrial body has been discovered.

IV. THE SECOND “THERAPSID” RADIATION AND LATE TRIASSIC FAUNAL TURNOVER

As a result of the decimation caused by the Permian–Triassic mass extinction, the taxonomic diversity of earliest Triassic terrestrial faunas was greatly reduced. These faunas were dominated by a dicynodont, Lystrosaurus, but included representatives of a few other lineages of “nonmammalian synapsids,” which were the basal stocks of another evolutionary radiation. Later, in the Early Triassic, numerous large and small herbivorous forms, primarily dicynodonts, became prominent elements of terrestrial faunas. In parallel, the surviving carnivorous “therapsids” radiated, producing a variety of new lineages of large and small carnivores as well as some herbivorous species.

Analysis of the evolutionary radiation of the carnivorous “therapsids” during the Triassic reveals trends in modification of skull structure and morphology of the postcranial skeleton that increasingly resemble the characteristics of modern mammals. For example, the structure of the lower jaw was modified through expansion of the dentary bone and reduction or losses of other bones characteristic of early synapsid or reptilian jaw structure. These were preliminary steps toward acquisition of an articulation of the dentary with the squamosal bone of the skull, a key osteological character for distinguishing modern mammals from reptiles. Limb posture began to be modified from a primitive, sprawling stance to a more typically mammalian stance with the elbow and knee tucked close to the body and the limbs brought into an upright posture.

A few million years before the end of the Triassic, some synapsids had evolved a functional articulation between the dentary bone of the lower jaw and the squamosal. Many textbooks and popular articles present the acquisition of this structure as the hallmark of the Mammalia and identify the end of the Triassic (approximately 206 Ma) as the time of origin of the group. In current phylogenetic classifications, however, the clade Mammalia is restricted to a crown group including the common ancestor of living monotremes (the echidna and platypus), marsupials, and placentalists and all its descendants. Phylogenetically, some of the very advanced “nonmammalian synapsids” with a dentary–squamosal jaw articulation and other mammal-like specializations lie outside the crown group; they are sister lineages. In many recent research papers, these very advanced “nonmammalian synapsids” are referred to as “mammaliaforms.”

During the Late Triassic, the composition of terrestrial faunas underwent a major change that played a significant role in reshaping the course of synapsid evolution. All the lineages of Triassic synapsids, except for a few lineages of “mammaliaforms” and one group of herbivorous “nonmammalian synapsids,” the tritylodonts, became extinct. All the surviving lineages were represented by species of relatively small body size. Two groups of reptiles, the rhynchosauras, and several lineages of dinosaurs became the dominant terrestrial vertebrates.

Recent studies indicate that the Triassic–Jurassic mass extinction’s primary impact was on the marine biota. Some evidence suggests that terrestrial floras were modified at this time. The admittedly limited fossil record of Late Triassic and Early Jurassic “nonmammalian synapsids” and “mammaliaforms” suggests that these groups were not greatly affected. Clearly, the radiation of “mammaliaforms,” which began in the Late Triassic, continued in the Jurassic.

V. MAMMALIAN EVOLUTION DURING THE AGE OF DINOSAURS

The extensive changes in terrestrial faunas during the Late Triassic brought the first phase of synapsid evolution to a close. During the Jurassic and Cretaceous,
popularly dubbed the Age of Dinosaurs, "mammaliaforms" and early mammals were not dominant members of terrestrial faunas. For the most part, they were very small animals, in the size range of the smallest modern mammals. A few evolved larger body size, rivaling modern opossums or raccoons, but these species were exceptions to the rule.

During the Jurassic, "mammaliaforms" differentiated. Functional interpretations of their dentitions suggest that most were probably carnivorous, feasting on small prey including small terrestrial vertebrates and invertebrates. One group, the multituberculates, evolved a remarkably rodent-like style of dental specialization and is interpreted as having been omnivorous or herbivorous in dietary preferences. This interpretation is strengthened by the persistence of the group into the Paleocene. Then, as rodents began their evolutionary radiation in the late Paleocene, the taxonomic diversity of multituberculates dwindled by the end of the Oligocene, when almost all families of rodents had differentiated, multituberculates were extinct.

Tracing the course of mammalian evolution during this second phase of synapsid evolution is complicated by distinct biogeographic differentiation. Unlike the Permian and Triassic, when most of the terrestrial regions of the world were parts of one supercontinent, Pangaea, through the Jurassic and Cretaceous terrestrial areas and their faunas began to be fragmented. In part, this was the product of plate tectonic processes that broke up the Pangaea supercontinent and shifted the positions of the resulting continental blocks. Additionally, the Age of Dinosaurs was a time characterized by extensive marine transgressions when shallow seas covered many areas and added to fragmentation of terrestrial areas. By the beginning of the Cretaceous, when the currently sparse fossil record of these faunas gives us our first real picture of biogeographic differentiation, there is evidence of distinct northern (North American–Eurasian) and southern (Australian–Antarctic–South American) terrestrial faunas. Unfortunately, very little is known of the history of "mammaliaform" and mammalian evolution on the African continent during this interval.

The Cretaceous was a time of major change in the terrestrial biota. Angiosperms made their appearance and began to diversify. First records of most modern families of insects document their diversification. Particularly in the Northern Hemisphere, dinosaurian faunas underwent a major reorganization. Global climates of the Jurassic and Cretaceous were, in general, distinctly warmer and more equable than modern climates. The latitudinal climatic gradient, reflecting the degree of difference between equatorial and polar conditions, was much lower. Throughout most of the interval there is no evidence of glaciation. Finally, to round out this description of the setting of mammalian evolution during the Age of Dinosaurs, it must be noted that evolution of the terrestrial biota was not interrupted by a mass extinction event.

By the middle of the Cretaceous, the mammalian fauna had undergone a major turnover in composition. In the Northern Hemisphere, with the exception of the hardly multituberculates, most of the Jurassic lineages of "mammaliaforms" had dwindled in taxonomic diversity or become extinct. Concurrently, two of the three major lineages of the crown group Mammalia, marsupials and placentals, appeared and began to radiate.

Mammalian evolution on a southern continent, composed of the modern Australian, Antarctic, and South America continents, followed a different course. The terrestrial fauna of this area includes the first records of the third major lineage of the crown group Mammalia, the monotremes. Additionally, at least two groups of "mammaliaforms" diversified there.

Currently, the first records of the crown group Mammalia are of Early Cretaceous age. The biogeographic dichotomy, the first appearances of marsupials and placentals in the Northern Hemisphere and the first appearance of monotremes in the southern continent, strongly suggests that the first records of the crown group Mammalia will be found in the Jurassic.

VI. CRETAUCEOUS–TERTIARY MASS EXTINCTION AND ITS CONSEQUENCES

The effects of the Permian–Triassic mass extinction only overshadow the extent of devastation of marine and terrestrial biotas during the Cretaceous–Tertiary mass extinction. On land, the Cretaceous–Tertiary mass extinction is marked by the demise of nonavian dinosaurs and some other vertebrate lineages. Similarly, many dominant groups of marine invertebrates and aquatic reptiles died out at this time.

The primary source of information documenting the effects of the Cretaceous–Tertiary mass extinction on a terrestrial biota is a series of studies of faunal and floral evolution in an area of the Western Interior of North America that extends from Colorado northward into Alberta, Canada. For decades, a wide range of geological and paleontological research projects have been under way in this area, and the resulting database is extensive. This is both a blessing and a curse. There has been a tendency to regard the results of studies on
biotic change in the North American Western Interior as typical of global patterns. Recent discoveries in other areas show that this was not the case. On the other hand, although geographically limited, the extensively documented patterns of survival and extinction in the North American Western Interior can be used to test hypotheses concerning the kinds and severity of environmental changes that caused the Cretaceous–Tertiary mass extinction.

Archibald (1996) and others have analyzed the patterns of survival and extinction of terrestrial vertebrates based on extensive collections from northeastern Montana. In the Cretaceous–Tertiary mass extinction all the nonavian dinosaurs, approximately 20 genera, became extinct. A small group of freshwater sharks and their relatives shared the fate of the nonavian dinosaurs. These losses, plus extensive extinctions among the Cretaceous lineages of lizards and marsupials, account for almost 75% of the extinctions of terrestrial vertebrates in the North American Western Interior. In contrast, at the species level, no extinctions of lineages of frogs, salamanders, and placental mammals have been documented. Only a few species of turtles, crocodilians, and crocodilian-like reptiles died out at this time.

Since Archibald published his analysis, several paleontologists have begun to analyze the very fragmentary fossil record of the evolution of birds in the Late Cretaceous. Their research shows that the evolutionary diversification of the crown group of birds had begun in the Cretaceous. Although many lineages of more primitive birds died out before or at the Cretaceous–Tertiary boundary, extinction appears to have taken little toll among the lineages of the crown group.

Among the mammalian lineages present in the North American Western Interior, approximately 50% of the species of “mammaliaform” multituberculates, which are thought to be the ecological equivalents of rodents, became extinct during the Cretaceous–Tertiary mass extinction. All lineages of placental (also dubbed eutherian) mammals survived this event. In stark contrast, all but one of the Cretaceous lineages of marsupials appear to have become extinct at this time.

Recent discoveries in Australia and, particularly, South America are beginning to outline the course of mammalian evolution through the Cretaceous–Tertiary mass extinction in the Southern Hemisphere. At this time, Australia, Antarctica, and South America were closely approximated and global climates were much warmer. Direct evidence documents interchange of terrestrial vertebrates between Australia and South America. In South America, deposits laid down several million years prior to the end of the Cretaceous contain records of a mammalian fauna dominated by members of several lineages of “mammaliaforms” accompanied by a lineage of monotremes. A gap in the fossil record separates records of this fauna from deposits documenting the earliest Tertiary fauna known from the area. The change in mammalian faunal composition is striking. Most of the “mammaliaform” lineages had gone extinct. The lineage of monotremes survived into the Tertiary but soon became extinct. The dominant group of mammals in the earliest Tertiary of South America is a small but diverse group of marsupials that were accompanied by a few lineages of placental mammals. These marsupials and placentals appear to be derived from stocks that dispersed southward from North America about the time of the Cretaceous–Tertiary boundary.

Other continents have yet to yield extensive records of mammalian evolution across the Cretaceous–Tertiary boundary. The available data strongly suggest that the mass extinction at this time involved the termination of mammalian lineages in these areas. Globally, mammalian diversity appears to have been greatly decreased.

The distal causal factors of the Cretaceous–Tertiary mass extinction are the subject of continuing debate. The pattern of survival and extinction of species of terrestrial vertebrates at the Cretaceous–Tertiary boundary in the North American Western Interior argues strongly against, if not falsifies, hypotheses calling for massively catastrophic changes in the environment as distal causal factors of this mass extinction. Discovery of a large crater in Yucatan adds evidence supporting the hypothesis that an asteroid, or other extraterrestrial body, impacted the earth at the close of the Cretaceous. Additionally, massive volcanic deposits in peninsular India document a period of extensive eruptions that began before and continued across the Cretaceous–Tertiary boundary. Finally, the end of the Cretaceous is a time of a major marine regression. Certainly, all three of these events contributed to changes in global environments that triggered the mass extinction; their relative importance remains an open question (Archibald, 1996; Hallam and Wignall, 1997).

**VII. RECOVERY OF MAMMALIAN FAUNAS DURING THE EARLY PALEOCENE**

The bias favoring the fossil record from the Northern Hemisphere continues to limit analyses of the course
of mammalian evolution immediately after the Cretaceous-Tertiary extinctions, which marked the end of the second phase of synapsid evolution. Earliest Paleocene faunas are well-known only in the North American Western Interior. In northeastern Montana and immediately adjacent areas, a refined fossil record provides the basis for study of the recovery of the mammalian fauna during the first million years of the Paleocene (approximately 65-64 Ma). Radiometric age determinations make it possible to subdivide this period of faunal recovery into a first interval of approximately 400,000 years and a second of approximately 600,000 years.

In the first 400,000 years of the Paleocene, diversity of the mammalian fauna of the North American Western Interior was depressed in comparison to that of the latest Cretaceous. The most numerous mammals were placentals, primarily represented by the parapelyptic “condylarths.” This group was allied to lineages of modern ungulates, hoofed mammals, that would make their appearance later in the Paleocene or in the Eocene. The second most diverse and abundant group was the “mammaliaform” multituberculates. Only one lineage of marsupials remained. All these earliest Paleocene mammals were characterized by small body size. As far as can be determined from their fragmentary remains, none was larger than the largest latest Cretaceous mammals known from the area, which were about the size of modern opossums or raccoons.

Less than half of the mammalian species in this earliest Paleocene fauna represent lineages present in the preceding latest Cretaceous fauna of the Western Interior. The majority is composed of immigrants that appear to have differentiated in other areas and dispersed into the Western Interior after the Cretaceous-Tertiary mass extinction. This is one line of evidence that suggests the diversification of major placental lineages had begun in other, as yet unsampled, areas prior to the end of the Cretaceous.

Support for this view also comes from recent comparative molecular studies of modern placental mammals. Through application of the molecular clock technique, these have produced hypotheses suggesting very ancient times of differentiation of many placental lineages. The molecular clock technique is based on the assumption that the rate of accumulation of molecular differences has been relatively regular. If this is correct, the time of differentiation of two lineages can be estimated on the basis of their current degree of molecular difference. In large part, these molecular studies suggest that the differentiation of many major lineages of placentals took place in the middle or early part of the Cretaceous, if not earlier. This conclusion directly contradicts the current fossil record, which suggests the early differentiation of these lineages occurred in the Late Cretaceous or was a product of the rapid evolutionary radiation of placentals in the Paleocene. Probably the true history of placental evolution is to be found somewhere between these extremes (Novacek, 1999).

Regarding charting the course of mammalian recovery after the Cretaceous-Tertiary boundary, during the next 600,000 years of the Early Paleocene there was a major increase in mammalian, particularly placental, diversity in the Western Interior. In part, this increase was the result of diversification of lineages present in the area at the beginning of the Paleocene, but immigration of additional groups of placentals and multituberculates from other areas played a significant role. During this interval, body size increased in several lineages. For the first time since the Late Triassic, placentals and multituberculates achieved sizes that approximated those of modern middle- and large-sized mammals. These trends of an increase in taxonomic diversity and the appearance of species of large body size continued throughout the later part of the Paleocene.

VIII. MAMMALIAN EVOLUTION DURING THE TERTIARY

From the late Paleocene onward, mammalian faunas of the Northern Hemisphere are increasingly well documented by assemblages of fossils from both Eurasia and North America. After the initial recovery of the mammalian fauna of the Northern Hemisphere from the effects of the Cretaceous-Tertiary mass extinction, diversity continued to increase, and many lineages were characterized by evolution of increasingly larger individual body size. Before the end of the Paleocene some species had attained the bulk of modern rhinos or hippos.

The late Paleocene and Eocene was a period of marked faunal turnover in the Northern Hemisphere, with the appearance of an increasing number of lineages of placentals that are represented in modern faunas; for example, this was the time of appearance of the equids and other kinds of modern ungulates. Additional contributions to mammalian diversity came with the first records of flying mammals (bats) and the appearance of one group of marine mammals (whales). Evolution of the marine seals and sea lions occurred later, in the Oligocene. Also, there were instances of decreases in diversity or extinction of major lineages. For example,
during the late Paleocene and Eocene the diversity of the “mammaliform” multituberculates decreased with the appearance and diversification of rodents. The correlation suggests that competition with rodents, and possibly other small, placental herbivores, played a significant role in the demise of the multituberculates. In general, however, the diversity of terrestrial mammals increased.

Near the Eocene–Oligocene boundary, approximately 33.7 Ma, many mammalian lineages became extinct. In comparison to other extinctions, this was not a mass extinction event, but it significantly modified the composition of mammalian faunas of the Northern Hemisphere. The extinctions occurred at about the time of a marked cooling in global climates and were most frequent in groups of placentals thought to have been particularly adapted to tropical forests, which were widespread in the Eocene. In addition to having been cooler, Oligocene climates were characterized by a general increase in aridity and decrease in equalibility, which was reflected in the evolution of greater areas of savanna and steppe. Analysis of patterns of evolution of mammalian body size during the Eocene and Oligocene suggest a polarization with increases in diversity of species of large and small body size but a decrease in diversity of species of intermediate size.

Mammalian diversity in the Northern Hemisphere, measured at the family levels, reached a maximum in the early Miocene, particularly reflecting a radiation of several lineages of ungulates. Through the remainder of the Miocene and Pliocene, the diversity of large ungulates decreased as the climate became cooler and dryer. Although differing in detail, broadly similar patterns in the incidence of extinctions and faunal turnover characterize the evolutionary history of mammalian faunas of the continents of the Southern Hemisphere.

Until recently, climatic changes were thought to be the major, if not the only, driving force in mammalian evolution during the Tertiary. During the Paleocene global climates became distinctly warmer and reached maximum levels in the Eocene, when tropical forests extended into high latitudes. During the transition from the Eocene into the Oligocene, global climates appear to have changed rapidly, on a geological timescale. This was a time of extinction of many lineages of terrestrial and marine mammals. Although at a more gradual rate, cooling of the environment continued up to the beginning of the Quaternary and inception of major continental glaciation. The circumstantial evidence of a correlation of periods of mammalian diversification or extinction with patterns of climatic change has been challenged by recent studies (Alroy, 1998; Prothero, 1999). A variety of high-resolution paleontological and geological methods were used to describe and test these correlations. The results indicated that correlations of times of climatic change and periods of increased diversification or extinction were not precise. Other environmental or biological factors apparently contributed to controlling the course of Tertiary mammalian evolution.

IX. CONCLUSIONS

The approximately 300-million-year history of synapsid evolution can be subdivided into three major phases. During the first phase, “nonmammalian synapsids” were among the dominant terrestrial vertebrates. The second phase, which began in the Late Triassic and lasted until the end of the Cretaceous, was characterized by the dominance of the dinosaurs in terrestrial faunas. “Mammaliform” synapsids and early mammals were a lesser component of these faunas. The third phase began after the Cretaceous–Tertiary mass extinction and was characterized by a major diversification of mammals, returning them to a dominant position.

In addition to the ongoing effects of background extinction, mass extinctions played a significant role in directing the course of synapsid evolution. The Permian–Triassic and, particularly, the Cretaceous–Tertiary mass extinction events greatly reduced synapsid diversity and reset the course of its evolution. Other major turning points in the evolution of the group, which involved extensive extinctions of lineages, were not coincident with mass extinctions—for example, the periods of faunal turnover late in the Triassic and during the transition from the Eocene into the Oligocene.

Attempts to compare patterns and causal factors of pre-Quaternary and Quaternary mammalian extinctions must consider the significant differences in resolution of the timescales at which these events can be studied. In some instances, the fossil record of Quaternary, particularly later Quaternary, extinctions permits analysis of biological change on ecological timescales of hundreds or thousands of years. Resolution of the pre-Quaternary record is much less. The severity of these extinction events may be artificially compounded by our inability to distinguish events that occurred at different times over intervals of hundreds of thousands or a few million years.

Finally, a survey of current research on the distal causal factors of mass extinctions and periods of rapid faunal turnover in synapsid evolution indicates that there probably was not a single grim reaper. Varying combinations of the effects of modification of the con-
figuration of oceans and their patterns of circulation, increases in the intensity of volcanic activity, and the sporadic impacts of extraterrestrial bodies all contributed to sometimes deleterious changes in the physical environment. Similarly, biotic factors played a part. The degree of interdependence of the members of an ecosystem, for example, contributes to the system's resistance to environmental change. Competition and predation also have contributed to triggering extinctions. At different times in the pre-Quaternary history of mammalian evolution, all these and probably other factors caused extinctions of lineages.

See Also the Following Articles

DINOSAURS, EXTINCTION THEORIES FOR • EXTINCTION, CAUSES OF • MAMMALS, BIODIVERSITY OF • MAMMALS, LATE QUATERNARY, EXTINCTION OF • MASS EXTINCTIONS, CONCEPT OF • MASS EXTINCTIONS, NOTABLE EXAMPLES OF

Bibliography


I. Mangrove Trees
II. Mangrove Animals: Fauna of Terrestrial Origin
III. Fauna of Marine Origin
IV. Connections
V. Mangrove Diversity
VI. Uses and Abuses of Mangroves

GLOSSARY

aerenchyma A spongy plant tissue composed largely of air spaces enabling gas exchange to take place by diffusion in underground mangrove roots.
aerial roots In mangrove species such as Rhizophora, roots branch out from the stem some distance above the soil surface. Lenticels (pores) in the aerial portion of these roots enable gas exchange to take place, through aerenchyma tissue, with the respiring underground portions of the root.
mangal A term sometimes used to specify the mangrove habitat as a whole as opposed to "mangrove" applying specifically to the trees themselves. For the most part, however, mangrove is considered to apply to both trees and habitat.
pneumatophores In some species of mangrove, such as Avicennia and Sonneratia, underground roots spread laterally from the main stem. Pneumatophores grow vertically from these, typically standing 10–20 cm above the soil surface, enabling gas exchange to take place with the underground roots.

pseudofecal pellet Fiddler crabs and their relatives collect soil with their mouthparts, separate organic particles from mineral components by a complex flotation process, ingest the former, and discard the latter in the form of compact pellets. These are known as pseudofecal because, although extraction has taken place, the waste material has not passed through the gut.

MANGROVES ARE a group of trees and shrubs, mostly evergreen, which have convergently evolved physiological and morphological adaptations to shallow intertidal environments. These are mostly composed of soft sediment, in which other vascular plants are rare. Mangroves are almost exclusively tropical in distribution and often dominate large areas of coastline or estuary.

I. MANGROVE TREES

Currently, the total mangrove area in the world is estimated at 170,000 km². They are the principal source of primary productivity in such areas. By their presence, they also provide shelter for other organisms. Mangroves are therefore the energy base, and physical substrate, of an often complex and diverse ecosystem. Mangrove faunas, to a unique extent, comprise organisms of both marine and terrestrial origin.
A. The Mangrove Habitat

The mangrove environment is a demanding one. Typically, mangroves are regularly inundated by tides and are therefore usually in a permanently waterlogged state. The tidal water is saline, so mangrove trees have the problem of coping with salt and acquiring sufficient water against an osmotic gradient. In hot climates, evaporation may make the salinity even greater than that of seawater. In the Indus Delta (Pakistan), for example, the prevailing salinity may be as much as twice that of seawater. Among the vascular plants, only mangroves flourish in such an inhospitable environment (Fig. 1).

Mangroves are defined physiologically as trees that can survive in the mangrove habitat, or mangal. The term is not a taxonomic one, nor does it indicate phylogenetic divergence from a common mangrove ancestor. The approximately 50 species generally recognized as mangroves belong to 20 genera in 16 families, although 2 families, Avicenniaceae and Rhizophoraceae, dominate in number of species (as they do also in abundance) (Table I). In most cases these genera and families also contain nonmangrove members. Mangrove species have evolved their specialist features as the result of convergent evolution, and mangrove attributes have probably evolved independently at least 15 times.

In addition to true mangrove species, there is also a loosely defined category of mangrove associates. These are species often occurring in mangrove habitats but which also occur elsewhere. Some are found only at the landward margins of the mangal, whereas others, such as creepers and lianas, have their roots above the intertidal zone but invade the mangal by using the mangrove trees purely for support. Other plants associated with mangrove trees are epiphytes, which include ferns and the "ant-house" plants (see Section II.A), and parasitic mangrove mistletoes.

B. Adaptations to the Mangrove Environment

1. Salinity

Three principal mechanisms enable mangrove trees to survive saline environments. Some species exclude salt at the root surface while continuing to take in water. In Aegiceras and Avicennia, up to 97% of the salt is excluded, apparently by a physical rather than a metabolic mechanism. This has the effect of locally increasing the salinity of the soil around the roots, with implications for other organisms: mangrove trees modify their environment as well as respond to it. In other instances, trees take in salt but sequester it within cells in such a way that sensitive metabolic processes are protected from contact with excessive salt concentrations. Finally, several mangrove species secrete excess salt, at considerable metabolic cost, from specialized salt glands on their leaves. Many mangrove species use a combination of these mechanisms, as shown in Table II.
2. Waterlogging

The major problem of waterlogged soils is lack of oxygen. Underground roots, like all tissues, require oxygen for respiration. In a normal soil, gas exchange takes place readily through air-filled spaces between soil particles. In water, the rate of diffusion of oxygen is very low, and in consequence waterlogged soils are generally virtually lacking in free oxygen. One of the most widespread mangrove trees, Rhizophora, adapts to anoxic soils by keeping much of the root mass above the mud surface, surrounded by air. The stretches of these aerial roots (Fig. 2) close to the soil carry numerous gas-exchange pores, or lenticels, whereas the underground portions are honeycombed with air-filled spaces.

This air-filled tissue, or aerenchyma, is also a feature of Avicennia and Sonneratia, whose roots are horizontal and close to the surface. These species respire by growing numerous pencil-like pneumatophores which protrude above the mud surface and allow gas exchange with the underground tissues (Fig. 3). Pneumatophore growth is facultative: The less waterlogged the soil, the lower the pneumatophore density. In the extreme and atypical case of Avicennia growing in sand between the Egyptian Sinai desert and the sea, the soil is so well oxygenated that no pneumatophores develop.

The aerial roots of Rhizophora and the intertwined underground horizontal roots of Avicennia physically support the trees in what is often a relatively unstable and shifting soil. Aerial roots and pneumatophores provide attachment sites for epibionts and facilitate the accretion of sediment by impeding water movement.

3. Reproduction

Many mangrove species show some form of vivipary. Rhizophora is an example. The ovum is fertilized while still on the parent tree and grows by a combination of photosynthesis and acquisition of nutrients from the parent until it may reach a length of 50 cm (Fig. 4). This structure—neither a seed nor a fruit, and hence usually termed a propagule—then falls to the ground. The propogules of some species root almost immediately, but others appear to have an obligatory floating period before they sink and establish themselves. The majority of floating propogules probably settle close to the parent, but long-distance dispersal is also possible. Floating mangrove propogules may remain viable for a month or longer: Depending on current speed and

---

### TABLE I

<table>
<thead>
<tr>
<th>Family</th>
<th>Genus</th>
<th>Number of mangrove species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avicenniaceae</td>
<td>Avicennia</td>
<td>8</td>
</tr>
<tr>
<td>Combretaceae</td>
<td>Laguncularia</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Lumnitzera</td>
<td>2</td>
</tr>
<tr>
<td>Palmae</td>
<td>Nypa</td>
<td>1</td>
</tr>
<tr>
<td>Rhizophoraceae</td>
<td>Bruguiera</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Ceriops</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Kandelia</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Rhizophora</td>
<td>5</td>
</tr>
<tr>
<td>Sonneratiaceae</td>
<td>Sonneratia</td>
<td>5</td>
</tr>
<tr>
<td>Bombacaceae</td>
<td>Camptospermon</td>
<td>2</td>
</tr>
<tr>
<td>Euphorbiaceae</td>
<td>Exocoecaria</td>
<td>1</td>
</tr>
<tr>
<td>Lythraceae</td>
<td>Pemphis</td>
<td>1</td>
</tr>
<tr>
<td>Meliaceae</td>
<td>Xylocarpus</td>
<td>2</td>
</tr>
<tr>
<td>Myristacae</td>
<td>Aegiceras</td>
<td>2</td>
</tr>
<tr>
<td>Myrtaceae</td>
<td>Osbornia</td>
<td>1</td>
</tr>
<tr>
<td>Pectiniaceae</td>
<td>Peltiiera</td>
<td>1</td>
</tr>
<tr>
<td>Plumbaginaceae</td>
<td>Aegiceras</td>
<td>2</td>
</tr>
<tr>
<td>Persicaceae</td>
<td>Acrostichum</td>
<td>3</td>
</tr>
<tr>
<td>Rubiaceae</td>
<td>Scyphyphora</td>
<td>1</td>
</tr>
<tr>
<td>Sterculaceae</td>
<td>Hirticera</td>
<td>3</td>
</tr>
</tbody>
</table>

*This follows the classification of Tomlinson (1986); there are alternative views on the status of certain species as true mangroves or mangrove associate species.*

### TABLE II

<table>
<thead>
<tr>
<th>Species</th>
<th>Exclude</th>
<th>Secret</th>
<th>Accumulate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acanthus</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aegiceras</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Avicennia</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Bruguiera</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ceriops</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exocoecaria</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laguncularia</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osbornia</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Rhizophora</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Sonneratia</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Xylocarpus</td>
<td>+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
direction, they could travel a considerable distance. It is not uncommon for mangrove seedlings from Mexico, for instance, to be stranded and take root in Texas virtually across the length of the Gulf of Mexico. An even greater dispersal may explain the mangrove species Rhizophora samoensis, which is found only in Samoa and adjacent islands, at the opposite extremity of the Pacific from its presumed ancestor, the species R. mangle of Central America. The significance of dispersal ability for the geographical distribution of mangrove species is discussed in Section V.B.

II. MANGROVE ANIMALS: FAUNA OF TERRESTRIAL ORIGIN

Although mangrove roots are periodically immersed, the branches and leaves provide an environment little
difference from that in adjacent terrestrial forests, with which they consequently share much of their fauna. Mangrove animals of terrestrial, rather than marine, origin include arthropods (particularly insects, but also spiders and myriapods), amphibians, reptiles, birds, and mammals. Virtually none are found exclusively in mangroves.

A. Insects

Anyone who has worked in mangroves can testify to the abundance of biting insects, particularly mosquitoes and "sand flies" or biting midges (Ceratopogonidae). Mosquito larvae develop in rot holes in mangrove trees, in semipermanent brackish pools, or in the water retained in crab burrows. In the latter case, one East African species, Aedes pempbaensis, ensures a suitable burrow environment for its larvae by laying its eggs directly onto the claws of the crab Sesarma meinerti. Prey of adult mosquitoes includes, apart from humans, a variety of mangrove mammals and birds, and in some cases it extends to fish.

Ants are often abundant in mangroves, including the aggressive nest-building weaver ants (Oecophylla) of the Indo-Pacific and leaf-cutter ants (Atta) of South America. Particularly complex relationships have evolved between ants, epiphytic "ant-house" plants, and mangrove trees (Fig. 5). Ant-house plants have bulbous stems (which may weigh several kilograms) honeycombed with passages inhabited by ants. One such plant, Hydnophytum formicarium, has specialized chambers in which ants deposit the remains of their prey, and from which the plant can absorb nutrients released by fungal action. The situation is further complicated by the presence of butterfly larvae (Hypochrysops) which feed on the ant-house plant and which are tended by the ants. The relationship therefore involves interac-

FIGURE 4 Mangrove propagules on a Rhizophora tree, Indus Delta, Pakistan. See also color insert, Volume 1.

tions between two plant species, two animal species, and one or more fungus.

Most mangrove ants are arboreal and essentially terrestrial animals. In many cases they nest outside the intertidal zone and forage in the mangal only at low tide. One Australian species, *Polyrachis sokolova*, is truly intertidal, retreating at high tide to nests within the mangrove mud. Nothing is actually known of its physiology; like other intertidal insects, it may retain a surface film of air and therefore avoid the need for any special adaptations to immersion or varying salinity.

Probably of greater ecological significance are the various plant-eating insects. Termites play a major role in disposing of dead wood. Some species construct nests of mud on tree trunks several meters above high-tide level, with access galleries snaking down the trunk to the aerial roots and upwards to the canopy.

The most important herbivores are those that eat mangrove leaves and seedlings, particularly the larvae of moths and beetles. Typically, only a small proportion of leaf production falls to herbivory. Sometime, however, it reached epidemic proportions. Individual trees in an otherwise healthy forest may be completely defoliated, and occasionally areas of many hectares are stripped of leaves. Canopy loss may result in the defoliated trees dying and being replaced by other species that are more tolerant of unshaded conditions. Insect herbivory therefore may alter mangrove community structure.

Other mangrove insects include the spectacular synchronously flashing fireflies of Malaysia (*Pteroptyx*), which occupy the mangrove *Sommeratia* for their displays, and numerous species of butterfly and moth. Hawkmoths, bees, and drosophilid flies are among the species which are probably of importance in pollinating mangrove flowers.

**B. Amphibia and Reptiles**

Amphibia are rare in brackish or salt water, but one species, the crab-eating frog (*Rana cancridora*), is common in mangrove habitats of Southeast Asia. Tadpoles survive well in salinity up to 50% that of seawater.

Reptiles are more abundant. Numerous species of snake forage within the mangal at low tide, including terrestrial or arboreal species but also some for which the mangal is their primary habitat. Mangrove snakes eat crabs (sometimes reciprocated), insects, and fish. In Southeast Asia, one of the most formidable mangrove predators is the monitor lizard (*Varanus indicus*), which may reach 1 m in length. Crocodiles, caimans, and alligators also occur in mangroves, although these are now rare in many areas due to human activities.

**C. Birds**

Birds are highly mobile. Many spend only part of their time in mangroves, migrating seasonally, daily, or tidally. Mangroves provide a feeding area, a nesting site, a refuge from the rising tide, or some combination of these. Waders probe for invertebrates in the mud of the mangal or adjacent mudflats. Kingfishers, egrets, and herons catch fish or invertebrates in the shallow water of mangrove creeks. Larger fish eaters, such as pelicans, ospreys, and cormorants, range further afield and may return to the mangal to roost or breed. In the Caribbean, roosting and nesting colonies of cattle egrets (*Bubulcus ibis*) and scarlet ibis (*Eudocimus ruber*) are so densely packed that the consequent enrichment of the soil with guano leads to significantly enhanced local growth of the mangrove trees.

Mangrove forests typically include numerous passerine species. Nectar feeders such as sunbirds in Malaysia, boneyeaters in Australia, and hummingbirds in South America move seasonally into mangroves and may be important pollinators. Insectivorous passerines specialize in hawking for insects in the canopy or, among low-lying vegetation, in picking insects off leaves or from bark crevices or from different species of tree. Broadly similar guilds of insectivorous birds, comprising different constituent species, seem to occur in different geographical regions.

Few of the species found within the mangal are mangrove specialists, and those which are restricted to mangroves in one part of the world may occupy different habitats elsewhere. One example is the cosmopolitan Great tit (*Parus major*), distributed from western Europe to China; only in Malaysia is it a mangrove species. The lack of mangrove specialists is probably due to the relative simplicity of the mangrove forest structure compared with typical tropical forest, allowing less scope for niche specialization. Another reason is probably the proximity of a pool of competing species in adjacent tropical rain forest. There are proportionally fewer mangrove specialists in New Guinea, where rain forest usually abuts mangrove habitats, than in Australia, where this juxtaposition is less common. Within Australia, there are few specialists in the mangroves of Queensland, which are extensive and contiguous with rain forest, than in northwestern Australia, where this is not the case. Most mangrove birds are probably using the habitat opportunistically.
D. Mammals

As with birds, many mammal species use the mangal opportunistically. These include small rodents, agoutis, wild pigs, antelopes, deer, and rhinoceroses; the Sundarbans of Bengal are the last major redoubt of the Bengal tiger (Panthera tigris). Domestic animals, such as camels and buffalo, are often a major element in the mangrove fauna. Otters may also be abundant, feeding on fish and crabs from the mangrove creeks.

Monkeys are common in mangroves. In Southeast Asia these include macaques (Macaca) which forage on the mud for crabs and mollusks. They also uproot large numbers of mangrove seedlings: Because these are seldom eaten or even greatly damaged, the purpose is not clear. Herbivorous monkeys are found in the forest canopy, including leaf monkeys (Presbytis) and, in the mangrove forests of Sarawak, the striking proboscis monkey (Nasalis larvatus). This is found only in mangroves and riverine forests, and it specializes in eating foliage, which is digested in an elaborate multichambered stomach with the aid of resident bacteria.

Bats are often abundant in mangroves. Resource partitioning in insectivorous bats parallels that of insectivorous birds, with species specializing in different zones of the mangrove vegetation and catching their prey with different flight techniques. A single bat may eat up to one-third of its body weight of insects each night: A 30-g bat might therefore consume 5000 insects nightly. The impact on the insect population of foraging bats must be considerable.

The exclusively Old World fruit bats often occur in mangrove forests in vast numbers: Roosts of an estimated 220,000 individuals have been recorded. Most fruit bats feed on nectar and fruit, and it is this which attracts many species into the mangal. In Malaysia, the long-tongued fruit bat Macroglossus minimus is an important pollinator of the mangrove Sonneratia: the long tongue is specialized for insertion into the Sonneratia flower, which carries large projecting stamens to deposit pollen onto the fur of the feeding bat. Sonneratia flowers last for only a single night, possibly because of the wear and tear resulting from visits by such a large pollinator. This species of bat is a true mangrove specialist, and in western Malaysia at least, it has not been recorded from other habitats. Mangrove specialization is possible only because the three species of Sonneratia in the area have different flowering patterns so that nectar is available throughout the year. Other fruit bats switch seasonally between mangrove and non-mangrove species.

III. FAUNA OF MARINE ORIGIN

One of the principal reasons for the high faunal diversity of mangrove ecosystems is their accessibility to occupation by organisms from both terrestrial and marine habitats. Of these, the marine invaders are the more numerous in terms of numbers and diversity of species. These include more or less sessile organisms settling on aerial roots and pneumatophores as well as more mobile species living on and under the mud. Many animal groups are represented in the mangal, the most conspicuous and ecologically most significant being teleost fish, crustacea, and mollusks. As with the land-derived mangrove fauna, the majority of species occur elsewhere and accumulate in mangroves because of the availability of food, shelter, or suitable substrate.

Considering mangrove communities at a scale of, for example, hectares, the diversity of such animals is often high. At smaller scale, however, the anoxic conditions caused by waterlogging, exacerbated by microbial decomposition of detritus, may greatly reduce both species diversity and abundance.

A. Root Communities

Mangrove roots and pneumatophores provide a hard substrate often covered with a rich and diverse growth of sponges, sea anemones, bryozoans, tunicates, barnacles, tubeworms, and mollusks as well as epiphytic algae. These in turn may attract a more mobile population of browsers or predators. The epibionts are mostly filter feeders, extracting organic particles suspended in the water, or predators of zooplankton, with no direct interaction with their mangrove host. A particularly thick growth, however, can adversely affect the host tree by occluding lenticels and restricting gas exchange with the underground roots. The relationship is sometimes mutually beneficial, as encrusting sponges may transfer nitrogenous nutrients to their host, and encrusting fauna can protect the root from attack by wood borers.

The labyrinthine aerenchyma tissue of the roots is easily penetrated by wood-burrowing organisms. The isopod crustacean Sphaeroma is a common root borer and may cause severe damage and even death. Sphaeroma-induced damage near the growing tip of a root may induce forking, with a resulting increase in the number of roots entering the soil: This may benefit the tree. The "shipworm" Teredo (which is in fact a mollusk) also bores dead roots and trunks extensively and plays a similar role to that of termites in disposing of woody...
debris. Like termites, *Teredo* relies on symbiotic microorganisms to digest the more intransigent components of wood.

B. Fish

Mangrove creeks and inlets are frequently occupied by abundant and diverse fish populations. In Southeast Asia, for instance, records of more than 100 species are by no means unusual. Many of these species spend only part of their time within the mangal, often moving to other habitats seasonally or at different stages of their life cycle. Mullets (*Liza*) eat significant amounts of mangrove detritus, such as shed leaves; most hunt small crustacea or other invertebrates. Some fish are permanent creek residents, commuting into the forest when it is submerged at high tide and foraging among the mangrove roots.

At low tide, Asian mangroves are occupied by mudskippers, which are relatives of the gobies (Fig. 6). As their name suggests, they skip across the exposed mud surface using their tails and leg-like pectoral fins, sometimes even climbing up aerial roots or pneumatophores. This amphibious life requires appropriate physiological adaptations, particularly in relation to respiration. Mudskippers are largely air-breathing, with gas exchange taking place not just across the gills but also at highly vascularized areas of the skin. Some store air within their burrows to enable aerial respiration even at high tide.

All mudskippers are probably to some extent omnivorous, but some are predominantly deposit feeders and others carnivores. Prey of the latter include crabs, insects, spiders, shrimps, and snails.

C. Crustacea

Mangrove habitats, particularly in the Indo-West Pacific, are dominated by crabs belonging to two families, Grapsidae and Ocypodidae. The former are predominantly herbivores or detritus feeders and the latter deposit feeders, extracting fine organic particles from mangrove mud. Predatory crabs, such as the formidable *Sylla*, may also be important components of the mangrove fauna. Shrimps (*Penaeoidea*) and mud lobsters (*Thalassina anomala*), and smaller crustacea such as amphipods and isopods, may also be significant as scavengers, in breaking down leaf litter, or as predators of smaller organisms.

1. Grapsid Crabs

Grapsid crabs of the subfamily Sesarminae, particularly of the genus *Sesarma*, are characteristic of mangroves, although a few species of this genus occur in other habitats (Fig. 7). More than 40 species of sesarmine have been reported from the mangroves of Malaysia alone, and many species, here and in other regions, undoubtedly remain to be described.

*Sesarma* are small (usually less than 3 cm in breadth) and inconspicuously colored. They are amphipod, re-
processing of leaf material by crabs increases the rate of breakdown of leaf litter 75-fold compared with the rate of decomposition under microbial action alone. Therefore, sesarmine crabs collectively play a very important role in facilitating energy flow through the mangrove ecosystem. By eating propagules, they also affect species distribution and community structure of mangrove trees (see Section V.C.1). However, there are geographical differences: in Southeast Asia and Australia, sesarmines are crucial in litter breakdown and selective removal of propagules, whereas in Florida and the Caribbean they are of lesser significance.

2. Ocypodid Crabs
Some crabs of the family Ocypodidae, such as the Central American hairy land crab Ucides, consume mangrove detritus. The majority are deposit feeders. Among these, the most conspicuous are the gaudily colored fiddler crabs (Uca spp.), widespread throughout the mangroves of the Old and New World (Fig. 8).

The common name derives from the one greatly enlarged claw of male fiddlers, which is used in courtship and in deterring rival males. The smaller claw of males and both claws of females are devoted to feeding. Mud is scraped into the buccal cavity in which, by a complicated process of flotation and manipulation by the mouthparts, fine organic particles are separated from the mineral components. The former is ingested and the latter deposited as a ball of sand, or “pseudofecal pellet.” The process of separation may be quite selective. In some species, what is extracted consists almost entirely of microbial cells rather than, for example, fragmented leaf material. Others have subtly different extraction techniques and may specialize in the smaller meiofaunal animals. There may be as many as 60 fiddler crabs per square meter, resulting in 500 g of soil being processed daily. The toll on meiofauna is probably considerable, and the effects on soil texture and composition are profound.

3. Other Mangrove Crustacea
Other crabs found in mangroves are important predators. The most conspicuous is the mud crab Scylla serrata of the family of swimming crabs (Portunidae). Scylla reaches a carapace width of up to 20 cm, making it the largest invertebrate predator found in mangroves. Equally formidable predators are the mantis shrimps (Stomatopoda), which live in burrows in the mud and lacerate prey by rapidly shooting out their viciously spiked raptorial appendages. Other rarely seen burrowing crustaceans include pistol or snapping shrimps.
(Alpheus spp.) and the mud lobster *Thalassina* (see Section III.C.4).

More general mangrove scavengers include hermit crabs, particularly *Clibanarius*, which forage on the mud surface at high tide. Shrimps may also be abundant in mangroves and mangrove creeks. Penaeid shrimps, which in at least some parts of the world depend heavily on mangroves for feeding and breeding, are an important commercial crop. The shrimp *Merguiia* apparently lives only in mangroves and has the distinction of being the only semiterrestrial shrimp; it actually climbs trees. Only two species are known. One occurs in the Indo-West Pacific region, from Kenya to Indonesia, and the other occurs in Panama, Brazil, and Nigeria. Indo-West Pacific and Atlantic regions differ in the composition of their mangrove floras, and the separation of the two species of mangrove-associated shrimps may have occurred in parallel with the divergence of the mangroves themselves.

4. Crustacea as Ecosystem Engineers

All species have an impact on their environment, at the very least exchanging materials in the form of food, waste materials, and respiratory gases. Some species have effects beyond these simple transactions and alter the nature of their environment in ways that affect species other than their direct competitors, predators, or prey. Such species are often termed “ecosystem engineers.”

In a mangrove ecosystem, the trees are the greatest engineers, influencing sedimentation rates and creating a physical environment. Crustacea also, in important ways, transform their surroundings. The topography of mangrove swamps in Southeast Asia is often visibly modified by mud lobsters (see section III.C.3). While processing mud, *Thalassina* throws up waste material from beneath the surface, forming mounds which may reach 2 m in height. These create patches of dry mud which provide habitats for other species, including the mangrove fern *Acrostichum*, fiddler crabs, and a variety of other burrowing crustacea and mollusks. Between the mounds the mud surface is lower, and more waterlogged, than it would be otherwise. Burrowing crabs also contour their environment, although less dramatically.

Much of the microbial activity of mangrove mud occurs in the surface layer, to a depth limited by the diffusion and exchange of gases with the atmosphere. As fiddler crabs process surface mud, they continually expose fresh material, facilitating microbial activity, while the active surface of the mud is increased in area by crab burrows. Burrowing activity also oxygenates the deeper soil and creates an underground labyrinth of interconnecting passages, through which significant underground water flow occurs. Experimental evidence suggests that crab activities significantly affect nutrient recycling and enhance growth of mangrove trees. Crustacea therefore alter the state of their environment in ways that significantly affect other species.
D. Mollusks

1. Bivalves

The most visible bivalve mollusks of mangroves are the oysters and mussels found attached to roots. Within the mud, however, there is often an abundant population of burrowing species. These, like the oysters and mussels, are largely filter feeders, extracting fine organic particles from suspension. A less typical group of bivalves are the shipworms of the family Teredinidae (see Section III.A), including the giant mangrove shipworm Dictyathiera, which may reach 2 m in length.

2. Snails

Gastropod snails are also generally abundant in mangroves. As with the crustacean fauna, these include herbivores, detritus and deposit feeders, and predators. Although a few species are uniquely found in mangroves, the majority of surface-living species also occur on open mudflats.

The principal predatory snails are species of Thais, found in mangroves worldwide. These nurse over mud and mangrove roots, feeding on barnacles or smaller gastropods. In the mangroves of Costa Rica, for example, T.里斯quiformis densities may reach more than 200 per square meter, and the species plays a major role in maintaining the function of mangroves by removing encrusting fauna from their roots.

Many gastropod species are deposit feeders, ranging in size from tiny and almost invisible species to the massive Terebralia and Telescopium of the Indo-Pacific region, which may reach a length of 10 cm (Fig. 9). One species, Terebralia palustris, feeds on small detritus particles when young, but on reaching a length of approximately 3 cm it switches to a diet of fallen leaves. The teeth on the radula (the ribbon-like tongue) of gastropods metamorphose appropriately to a form suitable for the altered diet. In Florida, snails are important consumers of mangrove seedlings, at some locations destroying nearly three-fourths of the seedling population. This is an interesting geographical contrast with other regions, such as Malaysia and Australia, where crabs fulfill this role (see Section III.C.1).

The most abundant snails on the mangrove trees are often species of Littoraria, close relatives of the periwinkles of temperate rocky shores. In Central America, on both sides of the Isthmus of Panama, the common species is L. angulifera. In the Indo-Pacific, this species is replaced by many others, which partition between them the slightly different habitats afforded by a tree. In Papua New Guinea, L. scabra prefers the back of trees on the seaward side of a forest, L. intermedia prefers trees next to freshwater creeks, whereas the polymorphic species L. pallescens is found solely on leaves.

E. Meiofauna

Within the mangrove mud lies a rich fauna virtually invisible to the naked eye—the meiofauna. Beneath an area of 10 cm² of mud there may be many thousands of individuals. Orders of magnitude smaller than the more conspicuous macrofaunal crabs and snails are meiofaunal herbivores, detritivores, and formidable predators, with food chains probably dependent on
photosynthetic cyanobacteria ("blue-green algae") and heterotrophic bacteria. Meiofauna colonize fallen leaves, and the stages of leaf breakdown are accompanied by complex interactions and successional shifts in species composition and community structure which parallel, on a microscopic scale, the processes of macroecology.

The numbers of meiofaunal individuals are immense, and their diversity is astonishingly high. Not only are there many species but also the species show a higher level of taxonomic diversity. Among the macrofauna virtually all species belong to just three phyla: arthropods, mollusks, and chordates. The meiofauna from just one mangrove area in Australia, for example, yields turbellarian flatworms, nematodes, copepods, Ciliophora, Foraminifera, bivalve mollusks, oligochaete and polychaete annelids, hydrozoa, archiannelids, kinorhynchs, tardigrades, and gastrotrichs.

Very little is understood about the meiofauna of mangroves, their interactions, their functional significance in the ecosystem as a whole, and the relationship between the meiofaunal and macrofaunal worlds. Their small size belies their great importance.

IV. CONNECTIONS

The salient features of typical mangrove ecosystems are relatively high rates of primary productivity, much of the results of which enter decomposition pathways, either directly or after initial breakdown by leaf-eating crabs or mollusks. This is true of leaves and reproductive structures and, on a more protracted timescale, of the woody components of the trees. Particulate organic matter, either small leaf fragments or bacterial cells, is ingested by mollusks and crustacean deposit and filter feeders, enters meiofaunal food chains, or accumulates in the mud.

The ecosystem can be viewed physically as well as in terms of the flow of energy or matter. Mangrove trees supply hard surfaces on which other organisms settle, and they modify (as well as respond to) the physical environment by stabilizing the soil, facilitating accretion of mud, and retarding erosion. The environment is further modified by the physical activities of burrowing crustaceans and other animals.

Mangrove ecosystems cannot be considered in isolation. They interact with adjacent habitats through the trapping of exogenous sediment or export of particulate or soluble organic matter or inorganic nutrients. Animals, by moving between mangroves and other habitats, also contribute to import and export of matter. Commercially important penaeid shrimps use mangroves as nursery areas so that shrimp catches many miles away may depend critically on mangrove productivity. Hard evidence for such connections between mangroves and other ecosystems, however, is sometimes elusive, and the strength of such linkages is almost impossible to quantify.

V. MANGROVE DIVERSITY

Mangrove diversity must be considered at a range of spatial scales, from global patterns of species richness to the pattern of distribution, at a particular location, at a scale of a few meters. In considering mangrove fauna, even smaller spatial scales become relevant. At all scales, diversity is affected by the past history of the area, by physical factors, and by biotic interactions, but the importance of each of these and the timescales over which they operate vary with scale.

A. Global Patterns

1. Latitudinal Range and Species Diversity

Mangroves are almost exclusively tropical or subtropical. This distribution is a reflection of a temperature limitation: The global distribution of mangroves correlates very closely with, for example, the winter position of the 20°C isotherm (Fig. 10). The number of mangrove species declines with increasing latitude, with the most northerly and southerly mangroves being species of Avicennia. In temperate regions, mangroves are replaced by salt marsh vegetation: plants which, like mangroves, are adapted to conditions of salinity and waterlogging but which do not carry the additional burden of being a tree or of producing large propagules.

2. Longitudinal Differences

Within their temperature and latitudinal constraints, mangroves show interesting patterns of species distribution. The principal biogeographic division is between the Indo-West Pacific (IWP) and Atlantic–Caribbean–east Pacific (ACEP) regions. These two regions have broadly similar areas of mangrove habitat, but the IWP has four times more genera and six times as many species of mangrove: 17 genera compared to 4, and 40 species compared to 7. It is apparent that none of the mangrove genera are very diverse, possibly because of a general limitation on species diversification in harsh intertidal conditions. Genera occurring in the IWP, however, are slightly more species than those of the ACEP: 2.35 compared with 1.75 species per genus.
The differences between the IWP and ACEP regions are maintained by major barriers. The most obvious of these is the African continent (Fig. 10). Less obvious is the barrier represented by the central Pacific. This results principally from dispersal limitations rather than from the absence of suitable habitat. Suitable environments are present on many Pacific islands without natural mangrove populations, as shown by the success of the artificial introduction of mangrove species to Hawaii.

Further dispersal barriers, including the Isthmus of Panama, open ocean, and arid coasts unsuitable for mangrove occupation, divide the major regions into smaller subregions, each with a more or less distinctive mangrove flora (Fig. 11). Only one species occurs in all six subregions: the mangrove fern Acrostichum aureum. Two genera, Avicennia and Rhizophora, are common to both IWP and ACEP regions. All other genera are found exclusively in either the IWP or the ACEP, although the close similarity between Laguncularia (ACEP) and Lumnitzera (IWP) suggests a recent separation of these two genera.

The traditional explanation of mangrove species distribution is of a center of origin and of diversification in Southeast Asia, followed by dispersal restricted by physical barriers. This clearly makes little sense in relation to the current dispersal barriers. Fossil evidence of mangroves is widespread and reveals a much wider distribution during the Eocene and earlier epochs: Fossil Nypa, Avicennia, and Rhizophora pollen and other remains, for instance, have been identified in Eocene and Miocene deposits that now form part of North and South America, Europe, and North Africa as well as


![FIGURE 11 Comparison of the mangrove flora in six geographical subregions. Because of the recent closure of the Isthmus of Panama, the eastern Pacific and western Atlantic (including Caribbean) are most similar in species composition. Note also the separation between Atlantic–Caribbean–eastern Pacific (ACEP) and Indo-West Pacific (IWP) regions.](https://example.com)
south and east Asia. At the time, these locations were connected by the Tethys Sea, continuous through what is now the Mediterranean and Indian Ocean.

Subsequently, this pantropical distribution was partitioned as a consequence of continental movements. Cosmopolitan genera such as Avicennia and Rhizophora were separated into regional populations by the approach of Africa to Asia 30–35 million years ago which closed the Tethys Sea, and separation of the sister genera Laguncularia and Lumnitzera followed the widening of the Atlantic barrier. The emergence of modern species ensued within the isolated subregions. Closure of the Isthmus of Panama was geologically very recent (a mere 2 or 3 million years ago) so that differences between eastern Pacific and Caribbean species are slight. One species (Pelliciera rhizophorae) is found on both sides of the Isthmus, presumably reflecting a separation into two populations too recently for allopatric speciation to have occurred.

An originally pantropical mangrove distribution was therefore partitioned into regions and subregions, with subsequent evolutionary divergence. Climatic conditions then eliminated mangrove species from areas such as southern Europe and the Mediterranean fringes. The current distribution pattern results from a combination of large-scale geographical factors and more regional climatic ones.

3. Diversity of Mangrove Fauna

It might be expected that faunal species diversity would follow a similar pattern to that of mangrove tree diversity both because the mangrove fauna has presumably been exposed to the same influences and because of a presumption that faunal diversity should respond to tree diversity.

The IWP region, richer in plant diversity than the ACEP, is also richer in species of mangrove-associated crustacea and mollusks (Table III). The reverse is true of other taxonomic groups, particularly those that form constituents of the root communities, such as sponges, coelenterates, and echinoderms. This may reflect regional differences in tidal range and availability of roots for settlement. For many groups, unfortunately, little comparable data are available and recorded species numbers reflect taxonomic interest and effort rather than the composition of actual species assemblages.

B. Regional Patterns of Diversity

Species diversity varies within regions in response to many different factors. The ACEP region, in addition to having fewer mangrove species in total, shows less differentiation between localities within the region, and all the species available in the geographical vicinity are likely to be represented at most locations.

Various factors may result in local variation in species diversity. Mangroves do not grow on rocky shores or in areas where fresh water is completely lacking (which is in part why all tropical shores are not dominated by mangroves). Stretches of inhospitable coastline therefore act as barriers which affect mangrove dispersal and geographical distribution. The arid shores of Soma-

<table>
<thead>
<tr>
<th>Taxonomic group</th>
<th>Atlantic–Caribbean–east Pacific</th>
<th>Indo-West Pacific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Caribbean/W. Atlantic</td>
<td>East Africa</td>
</tr>
<tr>
<td>Sponges/bryozoans</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>Coelenterata/cnidophora</td>
<td>42</td>
<td>12</td>
</tr>
<tr>
<td>Nonpolychaete worms</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Polychaetes</td>
<td>33</td>
<td>72</td>
</tr>
<tr>
<td>Crustacea</td>
<td>87</td>
<td>163</td>
</tr>
<tr>
<td>Mollusks</td>
<td>124</td>
<td>117</td>
</tr>
<tr>
<td>Echinoderms</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>Ascidians</td>
<td>30</td>
<td>13</td>
</tr>
<tr>
<td>Fish</td>
<td>212</td>
<td>114</td>
</tr>
<tr>
<td>Reptiles</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>Birds</td>
<td>138</td>
<td>177</td>
</tr>
<tr>
<td>Mammals</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

*From Saenger et al., 1983.
lia, for example, result in the reduction in species number northwards so that Avicennia marina is virtually the only mangrove species found in the Red Sea. Separation of mangrove estuaries from each other by arid coastline, and regional-scale variation in physical variables, also affects the species distribution of mangroves around the Australian coasts.

Dispersal ability also affects species distributions within regions. The number of mangrove species on islands of the western Pacific shows clear attenuation with increasing distance from the species-rich areas of Australia and Papua New Guinea. Similarly, among islands off the West African coast there is a clear relationship between the number of mangrove species present and the distance from the nearest landward neighbor (Fig. 12). Species number also correlates with island size, with larger islands containing more species.

C. Local Variation in Species Distribution and Diversity

1. Tree Distribution

At a specific location, the distribution of mangrove species responds to physical variables in the environment. These often vary as gradients: in an estuarine mangal, for instance, salinity and the influence of tidal fluctuations tend to diminish with distance up the river. Sediment composition and nutrient dynamics also alter with distance from the open sea. Mangrove species respond differentially to such upriver/downriver gradients, resulting in zonation of species.

Similarly, in areas dominated by tide rather than river flow, tidal fluctuations establish gradients of physical variables, particularly in salinity and the extent of waterlogging of the soil. Again, mangrove species respond differentially to these physical variables and tend to form distinct zones. Where both river and tidal influences interact, the pattern of species distribution can be extremely complex.

In relation to salinity, species generally grow better at low salinity and differ more in the tolerance range than in their salinity optima. Low salinity, in consequence, tends to be associated with higher species diversity. At higher salinities, tolerance differences result in differing competitive success and translate into zonation of mangrove species along a salinity gradient, with species dominating zones at which they compete best, rather than those corresponding to salinity growth optima.

Although response to physical gradients suggests a gradual transition from one species to another as the determining physical variable gradually alters, this is often not the case. Mangrove species are frequently found in virtually monospecific stands or zones, with a more or less abrupt transition from one dominant species to another. This suggests that interactions between tree species, and mutual exclusion, may play a part in defining zone boundaries. Other physical variables, such as the degree of waterlogging and soil an-
oxia, nutrient availability, and biotic interactions between species, similarly affect species distribution within the mangal.

Superimposed on the sorting of species under the influence of physical variables are variations resulting from interactions with the mangrove fauna. Of the faunal influences, the most significant is the selective destruction of mangrove propagules by sesarmine crabs (see Section III.C.1). At least in Southeast Asia and Australia, this is a major factor determining mangrove species distribution. Mangrove animals respond to physical gradients of salinity and inundation regime. Sesarmine abundance is often greatest at midshore, and it is therefore here that mangrove propagules are most vulnerable. For reasons related to nutritional value and the levels of aversive tannins, *Avicennia* is generally the preferred food of sesarmines: hence at some locations the distribution of *Avicennia* in the upper and lower shore and their virtual absence from intermediate shore levels.

Random factors can also affect mangrove species distributions. If a gap is created in a mangrove forest because of the death of a tree, it is most rapidly filled by the species that are the best colonizers and best able to flourish in unshaded conditions. In Southeast Asia, the result is often an initial invasion of the mangrove fern *Acrostichum*. This may be succeeded by seedlings of *Bruguiera parviflora*. This species has relatively small and easily dispersed seedlings, whose growth is suppressed by the shade of an intact canopy. These in turn are replaced by slower growing shade-tolerant species such as *B. gymnorrhiza*. *Avicennia marina* is less tolerant of shade but is less likely to occupy a small gap because of propagule destruction by crabs. If, however, the gap is large one, *Avicennia* is more likely to establish itself, probably because foraging crabs are vulnerable to predation in large open spaces. The distribution of species within a mangrove forest may therefore be patchy and reflect the stochastic nature of tree death and the subsequent successional history. On a larger scale, extensive death of trees by typhoons, by widespread defoliation by insect attack, or even by oil spills can have profound and long-lasting effects on species composition.

The structure of a mangrove forest is therefore in part explainable in terms of “patch dynamics”—of gaps appearing by chance and being filled by a changing assemblage of species differing in composition (at least for a time) from the surrounding forest. Eventually, something similar to the surrounding forest emerges. With a high incidence of gaps, a mangrove forest could be seen as a mosaic of patches of different successional age: if patches appear relatively rarely, the effect would be transient aberrations in an otherwise homogeneous, or consistently zoned, environment.

2. Distribution of Mangrove Animals

The species distribution of the mangrove fauna is less well understood since small, cryptic, and often mobile animal species are less easy to describe and analyze than large and immobile trees. A high level of taxonomic confusion compounds the problem. Nevertheless, it seems likely that the same general considerations apply. The distribution of mangrove crabs, for instance, forms zones related to shore level, salinity, and soil texture, whereas mollusks show zonation patterns in relation to shore level and to vertical position on the roots and trunks of mangrove trees.

The distribution of species of mangrove animals may also be related to patch size and the distance between neighboring patches, on a smaller spatial scale than applies to the distribution of mangrove species themselves, corresponding to the more limited dispersal ability of the species in question. This was demonstrated in the classical experiments of Simberloff on the terrestrial arthropod fauna (principally insects) of mangrove islets in the Caribbean. The species richness on a range of mangrove islets increased with the area of the islets and decreased with increasing distance from potential sources of fresh colonists. When the fauna of islets was completely eliminated with pesticides, recolonization soon established an equilibrium species richness similar to that before the elimination. In terms of the representation of different functional groups the previous situation was largely replicated, but the actual species comprising the new assemblages differed. Finally, artificially reducing the area of mangrove in experimental islets reduced species richness, showing that it was causally related to habitat area rather than to habitat diversity.

At an even smaller scale, individual mangrove roots can be regarded as “islands” of habitat suitable for epibiont settlement, surrounded by areas of unsuitable habitat. Here, too, the composition of root epibiont communities appears relatively stable in terms of functional groups. The actual species present are much more unpredictable and particularly affected by physical variables and by the supply of colonizing larvae. These factors are of different significance at different time and spatial scales.

Meiofaunal diversity has scarcely been investigated, although the same considerations apply as in the macrofaunal world. Variation in physical variables, species interactions, patchiness, dispersal, and the other factors relevant to larger organisms must also affect the meio-
fauna. To date, limited research interest (and the intrinsic difficulty of studying species interactions or measuring, e.g., nutrient gradients at a scale of millimeters) has restricted our knowledge of mangrove metazoan diversity and the factors which determine it.

D. Genetic Diversity of Mangroves

The advent of molecular genetic techniques has made it possible to study diversity at levels lower than the species. To date, few species have been studied, and clear general conclusions cannot be drawn. In some cases, such as the self-pollinating Rhizophora mangle of Florida and the Caribbean, populations appear to be genetically homogeneous, with slightly more genetic variation toward the northern extremes of the species' range. The extent of intraspecific genetic variation varies with the breeding structure of the population, with dioecious species showing much greater polymorphism. Genetic variation between populations is naturally greater than that within a population at a particular location, although West African mangroves have greater levels of genetic diversity than the same species in the Florida and Caribbean. This confirms the belief that western Atlantic mangroves derive from African populations rather than the reverse. As research proceeds, no doubt many such insights into the causes and consequences of intraspecific diversity will emerge.

VI. USES AND ABUSES OF MANGROVES

Mangroves are of interest not just to biologists. Their diversity and productivity make them the source, directly and indirectly, of many products of use (and commercial importance) to humans.

Mangrove trees are exploited for timber for construction and firewood. This ranges from the casual collection of fallen wood to major charcoal industries based on the intensively managed mangroves of, for example, western peninsular Malaysia. Foliage may also be grazed directly or harvested for fodder for domestic animals. On a smaller scale, mangrove products are collected for a host of other purposes, including thatching houses, the manufacture of fish traps, for use in medicine, for tanning leather, and for use in various foods and drinks. Indirectly, mangrove productivity supports fisheries, both within the mangal and offshore. Less tangibly, mangroves can be of considerable importance in consolidating shorelines and limiting coastal erosion.

The significance of mangroves to humans varies greatly from place to place, but attempts have been made to achieve an overall economic valuation of the goods and services supplied. One recent estimate indicates that, on average, the annual value of a hectare of mangroves is approximately $10,000, resulting in a worldwide total contribution of $1,648,000,000.

An asset of this magnitude is worth conserving. Unfortunately, sustainable management of mangrove resources is the exception rather than the rule. In almost all parts of the world, mangroves are under pressure from irrigation schemes which divert rivers and prevent fresh water from reaching mangroves and from pollution, overexploitation, or deliberate clearance for construction or for the planting of alternative crops. One of the most destructive processes in many countries of Southeast Asia and Central America has been the clearance of mangroves for the construction of shrimp ponds—an attempt to increase the production of species dependent on mangroves while simultaneously reducing the primary production on which they depend. Not surprisingly, this has not been a success.

During the past few decades, loss of mangrove area in many countries has been dramatic. In the Philippines, for example, 60% of the mangrove area has disappeared, whereas in other countries such as Malaysia, Thailand, and Pakistan, annual losses are on the order of 1–3%. It may be, however, that the tide is turning. The virtual collapse of the shrimp industry in several countries and a greater awareness of the value of mangroves as a natural resource have focused attention on rational management strategies and on the possibility of reversing some of the damage. Much effort is now being put into replanting mangroves in abandoned shrimp ponds and the rehabilitation of denuded areas for coastal protection or in support of local fisheries as well as into developing suitable mangrove areas for ecotourism. The destruction of mangroves has largely been due to human activities: In the future their survival may also depend on mankind.

See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • INTERTIDAL ECOSYSTEMS

Bibliography


I. Introduction
II. Extrapolating from Lakes to the Sea
III. Nearshore Ocean Nutrient Response
IV. Stoichiometry
V. Contaminants and Stress

GLOSSARY

eutrophication  Overenrichment of aquatic systems with nutrients, often leading to harmful algal blooms and subsurface oxygen depletion.

harmful algal blooms  Development of sufficient numbers of cells of algal species to cause "harmful" effect to ecosystem.

microbial response  Expected responses of microscopic algae to nutrient enrichment is excess production beyond what can be consumed by grazers and species shift to noxious species; this expected response is not necessarily what happens.

nutrient enrichment  Usually excess of nitrogen and phosphorus nutrients to aquatic systems.

redfield ratio  From large-area and time averaging, ratio of carbon, nitrogen, phosphorus, and oxygen for normal ocean plankton and deep-ocean nitrate and phosphate pools.

subsurface oxygen depletion  Due to isolation of waters below the surface, metabolic breakdown of organic matter from the surface waters can cause depletion of dissolved oxygen concentration, leading to very low oxygen (hypoxia) or no oxygen (anoxia).

MOST LAKES, ESTUARIES, AND COASTAL OCEAN REGIONS in the proximity of large human populations have experienced significant nutrient enrichment when contrasted to more "pristine" conditions. This stress of nutrient enrichment is viewed as the cause of eutrophication, with classical symptoms of subsurface water oxygen depletion from excess algal production and proliferation of noxious algal species. The typical eutrophication phenomenon has been defined in lakes as a direct cause-and-effect relationship to a single macronutrient, usually phosphate. For both ecological research and resource management, this lake concept has been extended to estuarine and coastal waters, with some incorrect conclusions. Estuaries and coastal waters, as well as lakes, have complex microbial communities of phytoplankton algae plus heterotrophic bacteria and protozoa that together act as the beginning of the food web. The response of different members of these communities to different stimuli as well as differences of grazer pressure on different ecosystems must be taken into consideration. In estuaries and coastal waters, more complex biogeochemical reactions and tides and currents further complicate understanding the impact of nutrients. Ratios of the macronutrients nitrogen and phosphorus to each other and to silicon vary in both time and space,
exerting varying influences on the microbial response. In addition, trace nutrients, which are often considered unimportant in lakes, probably have a major influence on estuarine and coastal primary production. Trace metals and organic compounds also have potential contaminant influences on the overall microbial response. Eutrophication as a stress in estuarine and coastal marine waters is not a simple cause-and-effect phenomenon.

I. INTRODUCTION

Perhaps the single anthropogenic stress on aquatic and marine environments that is considered to be most ubiquitous is described with the broad term eutrophication. Human activities have mobilized nitrogen and phosphorus through agriculture, urban and suburban sewage, and atmospheric emissions sufficiently to greatly increase fluxes to aquatic environments, especially to lakes, estuaries, and coastal ocean waters. The overly simplistic view of eutrophication is that loading of nitrogen and phosphorus causes increased production and biomass of planktonic algae, with decreased species diversity. Thus, nutrient loading creates a stress to the aquatic community causing adverse community impact. Characteristic symptoms of eutrophication are depletion of oxygen in subsurface waters from excess algal biomass and development of blooms of noxious algal species (Richardson, 1997). Limnological studies of eutrophication have given concepts and observations about the phenomenon which, when applied to estuarine and coastal waters, often lead to incorrect conclusions. Contrary to simple models, nutrient loading in nearshore marine waters often does not support the level of primary production and phytoplankton biomass that would otherwise be expected from the high nutrient concentrations. Depletion of oxygen in subsurface waters is probably more a function of the physics of the specific aquatic system and lack of grazer consumption than it is of nutrient-stimulated excess algal growth. A bloom of a noxious algal species is probably more a function of the response by the entire aquatic community to the overall chemical milieu than it is nutrient-stimulated growth of the algal species.

This lack of a simple cause-and-effect response to nutrient loading is especially important to evaluate in light of two essentially opposite resource management actions. In nearshore waters, the traditional approach to eutrophication is removal of nutrient inputs to prevent algal production and growth. This approach has been successful in some cases. However, its success may be overestimated and probably, nutrient removal will have little of the intended impact in many estuarine and coastal waters. The second and opposite action with nutrient enrichment comes from recent proposals for large-scale engineering projects to fertilize waters of the open ocean to increase primary production (Cullen and Chisholm, 1999). In the one case, there is the simple goal of decreasing algal production (which is considered bad) by reducing nutrient inputs and in the other case a goal of increasing algal production (which is considered good) by adding nutrients; it’s not that simple. A principal reason that we often misinterpret marine eutrophication is that insufficient consideration is given to the hydrodynamic and biogeochemical complexity of the environment as well as the biodiversity of the microbial community (Paerl, 1998).

In discussing biological diversity, Norse (1993) considers hierarchical levels that range from genetic to species to ecosystem. Stress appears to decrease community diversity, often also decreasing the number of species within an individual function, such as primary production. It is critical to consider that increased diversity at one trophic level may decrease diversity at another. An interesting example comes from a recent examination of thermal stress on a planktonic community. Microcosms were studied with examination of several trophic levels (Petchy et al., 1990). It was found that environmental warming caused losses of top predators and herbivores, with increasing dominance by autotrophs and bacterivores. The warming increased extinction of predators with little effect on primary producers and bacterivores. Primary producer and bacterivore biomass increased, bacterial biomass did not change, and there were idiosyncratic impacts on total biomass. Warming directly increased primary production through temperature-dependent physiology and indirectly through food web structure. The results of this quantitative study are consistent with the general descriptive view of eutrophication where primary production and bacterial decomposition appear to increase while much of the rest of the trophic structure decreases (Vollenweider et al., 1992).

Many anthropogenic inputs to the aquatic environment that may be considered potential stressors have a dual impact, both stimulating growth at low concentrations and decreasing growth at higher concentrations. At a functional level, like primary producers, it is well known that temperature has a stimulating influence on individual species until a threshold is reached, above which it can inhibit. There has been
considerable interest recently about the influence of
trace metals on primary producers showing this dual
stimulating–inhibiting influence (Bruland et al., 1993).
The impact of nutrients is clearly similar if one looks at
the full ecosystem. Overcoming limiting levels increases
primary production and can increase entire ecosystem
production, but only to an extent. Then too much nutri-
ent loading may become a negative stress on the ecosys-
tem. It is necessary to discriminate whether the effect is
from nutrients or other inputs.

In the next section, the extension to the marine
environment of the limnological example of eutrophica-
tion is examined. Following that is a section examining
relatively low growth in high-nutrient environments
and a section on ecosystem stoichiometry. The last sec-
tion addresses the need for more thorough understand-
ing of multiple stressor effects on the entire community,
including grazer and trophic transfer.

II. EXTRAPOLATING FROM LAKES TO
THE SEA

A. Lake Eutrophication

The classic example of reducing input of one nutrient
to ameliorate eutrophication can be seen with the rela-
tively successful cleanup effort that has taken place with
lakes (e.g., Edmondson, 1991). Vollenweider's models
of the 1960s have led to the single-element, phosphorus,
approach to lake eutrophication. These were success-
fully applied to the experimental lakes program, where
focus on P proved successful (Schlinder, 1981). In these
studies, if carbon or nitrogen were in shortage, it could
be brought in from atmospheric or sedimentary sources.
Schlinder also concluded that micronutrients were not
important except in rare cases. The accepted overview
for lakes is that phytoplankton production is controlled
by annual P-loading and that P-loading is directly re-
lated to P concentration and P concentration and chlo-
rophyll are well correlated. Generalizations of this na-
ture may be more valid for lakes than for marine and
estuarine waters due to more thorough experimental
information and the hierarchical approach used to study
lake eutrophication (Hecky and Kilham, 1988). In addi-
tion, physical properties in estuarine and marine waters
may make these systems more complex. Indeed, Schlinder
(1981) has stated, "The control of eutrophica-
tion in estuaries has appeared to be much more complex
that it is in lakes. Estuarine and coastal ocean waters have considerable
transport circulation driven by both tidal energy and
riverine discharge and this circulation is variable on
predictable and unpredictable periodicities. These envi-
ronments also have complicating factors from variations
in suspended sediments through specific input sites,
circulatory resuspension of bottom sediments, and co-
agulation along estuarine salinity gradients. Addition-
ally, the salinity gradients of estuaries cause significant
physical stress such that populations of organisms will
change and moreover the change in ionic strength will
influence chemical speciation and solute–solvent inter-
actions. The more complicated and varying physics
gives rise to less predictability of populations on a sea-
sonal basis. This pertains not only to primary producers
but also to food web structure. There is a large literature
on the subject of coastal ocean eutrophication, with
recognition of difficulties in applying eutrophication
concepts to estuarine and coastal waters, where physics
and biogeochemistry complicate the picture (Vollen-
weider et al., 1992).

B. Food Web Complexity

In addition to the complicated nature of coastal waters,
it is well recognized that food web structure can bring
significant variability to all aquatic ecosystems, lakes
as well as coastal marine waters. Schlinder et al. (1997)
demonstrated that the primary production enhance-
ment from nutrient enrichment was less in a piscivore-
dominated lake than in a planktivore-dominated lake.
The reason for this difference was the suppression of
phytoplankton by large zooplankton in the piscivore
lake. The study focus included evaluation of drawdown
of atmospheric CO₂ from nutrient-enhanced produc-
tion and confirmed that changes in aquatic CO₂ fugacity
could be successfully manipulated in lakes and open-
ocean ecosystems.

The recognition of complexity in oceanic microbial
ecosystems led to the concept of the "microbial loop,"
where organic matter from phytoplankton is rapidly
consumed by bacteria, which are in turn consumed by
protozoans that are consumed by small zooplankton
that otherwise would be herbivores exclusively (Azam
et al., 1989). This shunt from the phytoplankton–
herbivore direct trophic transfer means lower efficiency
in the path to higher metaizoans and led to debates of
bacteria as "source or sink" in oceanic as well as estua-
rine environments (Sherr et al., 1987). Subsequent work
with the microbial loop has led to considering the com-
bined phytoplankton and microbial heterotrophs (bac-
teria and protozoans) as the primary producer commu-
nity supporting the metaizoans (Sherr and Sherr, 1991).
With the primary producer function from a multicom-
parment ecosystem, it is not surprising to find phyto-
plankton and heterotrophic bacteria being influenced
by different stimuli. For example, Pace (1993) showed
bacteria being controlled by phosphorus while phyto-
plankton were controlled by nitrogen and phosphorus
in lake nutrient enrichment experiments. In most cases
with any aquatic environment, nutrient addition will
bring variable influences on phytoplankton and hetero-
trophic bacteria and may also influence heterotrophic
protozoans and metazoan grazers.

It has been well recognized that phytoplankton pro-
duction, or the entire primary producer community, is
influenced by removal (top-down control) as well as
by resource limitation (bottom-up control). In a recent
review of eutrophication in planktonic ecosystems,
Gilbert (1998) pointed out that grazing and nitrogen
recycling are intricately connected in controlling plank-
tonic nitrogen availability. Another important recogni-
tion is that top-down control has a major impact on
export from the pelagic system (Wassman, 1998).
Wassman warned that to view only bottom-up controls
(nutrient influence) will not successfully guide biogeo-
chemical studies of marine systems. Thus, it seems obvi-
ous that a nutrient influence on phytoplankton should
not be considered in the absence of the rest of the
beginnings of the ecosystem.

C. Response of Nearshore Waters to
Nutrient Enrichment

The idea that a single nutrient controls primary produc-
tion comes from classical ecological theory. With a sin-
gle nutrient, a single phytoplankton species should have
an advantage over others and dominate by outcompet-
ing. The fact that multiple species can coexist within an
apparent niche was considered a paradox (Hutchinson,
1961) and was the subject of a massive amount of
excellent aquatic research. Recognizing the necessity
to consider guilds rather than a single species makes
understanding of primary production more compli-
cated. Add to this the more recent recognition that
multiple compartments of phytoplankton plus bacterial
and protozoan heterotrophs may be considered as a
primary producer community, and the need for whole-
ecosystem experiments is obvious.

There is an oversimplified view that nutrient concen-
trations (or loading) above those of some "pristine"
condition directly cause phytoplankton response, with
negative impact in estuarine and coastal waters. The
overall impression is that increased nutrients cause in-
creased algal growth with the consequence of excess
algal production causing oxygen depletion or the conse-
quence of a bloom of noxious algae. Oxygen depletion
from nutrient-enriched phytoplankton growth occurs
in environments where summer stratification isolates
bottom waters, for example, Chesapeake Bay and mid-
Atlantic coastal waters. Oxygen depletion is often quite
variable on an interannual basis and moderated by
meteorological forcing. Thus, the occurrence and extent
of oxygen depletion are complicated and not simply
predictable as a function of nutrient loading.

There is concern that unusual and noxious algal
blooms are increasing globally in both geographic ex-
tent and intensity, although there is debate on the quan-
titative significance of such claims (Andersson, 1997).
It is important to be careful in defining harmful algal
blooms (HABs) as Smayda (1997) has indicated. In
most cases, the sign of the "bloom" is the appearance
of numbers of cells of a species of a harmful alga suffi-
cient to have a negative environmental impact. Analysis
of individual HABs shows that generally the HAB taxa
have no unique ecophysiology, including higher affinity
for nutrients, and often that the HAB taxa have growth
rates lower than those of phytoplankton in general
(Smayda, 1997). Many HAB taxa have allelochemically
enhanced competition with other algal species and have
allelopathic defense against predators as well as against
a broad group of other microbial taxa. It appears that
noxious algae bloom from their ability to dominate
rather than their ability to outcompete other species
for nutrients or to grow fast. What actually stimulates
these taxa to express their domination is an area in need
of more research. Full-ecosystem studies are needed to
better understand noxious algal proliferations.

The experimental lakes program mentioned above
has provided great empirical evidence to combine with
theory in limnology. It is more difficult to manipulate
whole parts of estuaries and coastal oceans than it is
to do so with small lakes. Controlled mesocosms are a
good compromise. The Marine Ecosystems Research
Laboratory (MERL) at the University of Rhode Island
has been one of the largest and most successful versions
of realistic estuarine ecosystems (Oviatt et al., 1986).
In these, sufficient volumes of water have been used to
overcome many problems of confinement and attempts
have been made to simulate estuarine physical and bio-
geochmical influences. Some excellent research has
been done and much learned about the complexity of
estuarine responses to nutrients and other stressors.

A much used picture that was developed with infor-
mation from the MERL experiments and from compari-
sion of phytoplankton biomass and production in vari-
ous estuaries and coastal waters has been shown by
Nixon. The picture indicates phytoplankton increasing
proportionately with increasing nutrient concentrations or loading. Figure 1A shows a generic version of this with phytoplankton production versus nitrogen loading. Note that the nitrogen loading is portrayed on a logarithmic scale (as was done in Nixon and Pilson, 1983), giving the appearance of regularly increasing production as a function of increasing nitrogen over a broad range of nutrient loadings. With transformation to a linear axis for N-loading, it is obvious that production becomes asymptotic after an initial linear increase. There have been several articles in which this conceptual picture has been shown and expanded upon; most of them are extensive evaluations of data from many published works, with the authors indicating that the relationship is complex (e.g., Nixon et al., 1986). As has been cautioned by Nixon and others using it, the relationship is intended to cover a large range of nutrient conditions and to compare a number of different environments. However, a simplistic extension has been made suggesting that there is a simple linear relationship between nutrient loading and adverse phytoplankton production.

The behavior shown in Fig. 1B is probably more correct to indicate that phytoplankton response to increased nutrients is not linear along a very long loading or concentration scale. Relatively small increases in nutrient concentrations and loadings will cause a large increase in primary production, but continued increases do not. In fact, it is likely that with very high nutrient concentrations a decreased phytoplankton response will be encountered as is shown with the theoretical curve in Fig. 1C. Thus, it is not necessarily the case that nutrient enrichment leads to excess algal production. Perhaps, we should be addressing the question of why there is not greater phytoplankton production in estuarine and coastal waters from nutrient enrichment.

III. NEARSHORE OCEAN NUTRIENT RESPONSE

A. The Oceanic HNLC Concept

Although it is well known that enhancement of primary production is not simply and directly proportional to enrichment by a single nutrient, there is a tendency to oversimplify this relationship in both nearshore and oceanic waters. I would like to use the oceanic HNLC (high nutrient, low chlorophyll) concept to examine enrichment of nearshore waters. The HNLC concept came from interest in iron limitation in the ocean that has a history going back to at least the 1920s (Martin et al., 1990) but is best recognized in relation to Martin’s proposed Antarctic and equatorial Pacific experiments (Coale et al., 1996). Underlying the HNLC is the observation that some areas of the open ocean have relatively high concentrations of nitrogen and phosphorous nutrients but do not support proportionately large phytoplankton biomass; see Table 1. In open-ocean experiments, iron as a trace element added to overcome limitation has been shown to increase phytoplankton production with concomitant drawdown of atmospheric CO2 in both the equatorial Pacific and the Antarctic oceans. The very nature of considering a trace
constituent as critical to oceanic production indicates that the ecosystem is more complex than a simple cause-and-effect relationship with a major nutrient. In addition, the details of responses and trophic complexities in these oceanic experiments need more study before simple conclusions should be reached. However, this does not stop engineering plans for commercial fertilization of the ocean that are based on very simple cause-and-effect relationships. The same oversimplification in these oceanic experiments need more study before simple conclusions should be reached. However, this does not stop engineering plans for commercial fertilization of the ocean that are based on very simple cause-and-effect relationships. The same oversimplification in nearshore waters leads to proposals to solve eutrophication based upon very simple cause-and-effect relationships. The same oversimplification in enrichments over time in the San Francisco Bay estuary. 

Table I suggests that estuarine waters have alternating controls by nitrogen, phosphorus, and light. It is important to consider that nutrient enrichment is not necessarily an extension of nutrient limitation and that nutrient enrichment will not necessarily cause a direct proportional increase in phytoplankton. In some cases, fairly direct increases can be shown such as the indication that chlorophyll levels in the Chesapeake Bay have increased appreciably, with a doubling in dissolved inorganic nitrogen (DIN) over several decades. In contrast to this are examples in estuaries that do not show predicted increases. For example, Balls et al. (1996) show no change in chlorophyll in the Ythan River estuary between 1960 and 1990 when there was a fourfold increase in DIN. Alpine and Cloern (1992) showed a decline in primary production with increasing nutrient enrichment over time in the San Francisco Bay estuary. Reviewing conditions in a number of shallow coastal environments, Cloern (1999) showed a nonlinear response between N-loading and phytoplankton production and suggested that the simple eutrophication model in lakes does not have a current analog in coastal eutrophication. Examining data on DIN concentration versus measured primary production for summer estuarine tran-
sects along the Delaware Estuary, we can show a negative relationship (Fig. 2). While I will not suggest that the high nitrogen concentrations directly cause a decrease in phytoplankton production, it is clear that there is not a simple continual increase in production proportional to high nitrogen content. Note the similarity of Fig. 2 to the theoretical Fig. 1C, where a small increase in DIN causes a large increase in production at concentrations near limiting, reaching an asymptote followed by a decline at very high DIN concentrations. To normalize measured primary production, the ratio of productivity to chlorophyll biomass (P/B) is often used. With data for the 1980s and 1990s from the Delaware Estuary, average summer P/B from the high-nutrient upper estuary is 61 (grams of C fixed per gram of chlorophyll) contrasted to the lower estuary production maximum region P/B value of 225 (J. H. Sharp et al., unpublished data).

In most estuaries, nutrient increases have not been uniform. Often increases in DIN are accompanied by smaller increases or decreases in dissolved inorganic phosphorus (PInf) such that the N/P ratio has increased. For example, this is the case with the nitrogen increases in the Chesapeake Bay and the Delaware Estuary (Sharp, 1988).

IV. STOICHIOMETRY

A. Redfield Ratios

From extensive averages of planktonic CNP composition and deep-ocean NO3 and PO4, nutrient concentrations, it was noted that a regular and predictable N/P ratio was found (Redfield et al., 1963). This Redfield ratio concept was extended to carbon and oxygen for ecosystem utilization of elements. The concept has been extensively applied to aquatic systems for over a half century and is a valuable guide in understanding biogeochemical fluxes. Analysis of particulate CNP in a number of fresh-water lakes indicates that a variety of conditions exist, ranging from N to P deficiency to sufficiency (Hecky et al., 1993). They concluded that ocean plankton is not as N and P deficient as is lake plankton. From a different viewpoint, Flynn (1990) concluded that only in the presence of excess NH4 is the cellular response to N-stress fully suppressed. Thus, plankton throughout the ocean show some symptoms of N-stress. He suggested that there are three forms of N status: N-replete (no stress), N-sufficiency (enough stress to depress NO3 transport and assimilation), and N-deplete (maximum stress, no growth). With this classification, most estuarine, coastal, and open-ocean waters are N-sufficient. To evaluate whether or not there is nutrient stress from too little or too much, it is probably necessary to look over sufficient time to be in essentially steady-state conditions. For instance, it has been suggested that in Southern Ocean waters, Redfield CNP ratios for plankton use were only obtained when averaging over the full vegetative season of the Austral summer (Hoppe and Goeyens, 1999). They suggested that the Redfield ratio was reached only because of nutrient-replete conditions.

B. Adding Silicon

The concept of the Redfield ratio has been extended to silicon for many studies of aquatic ecology. From laboratory culture studies, the average Si/N ratios for small and large diatom species is close to 1/1 (Brezinski, 1985). This would give an overall Redfield ratio for CNPSiO of 106/16/1/16/1/16/1/16/1/11002/276 for balanced diatom growth. For a long time, it has been assumed that with deficiency of Si, phytoplankton populations will shift from diatom domination to that of other groups of algae. However, recent research would suggest a more absolute limitation of “healthy” ecosystem production by Si. In oceanic HNLC environments, it has been suggested that “new” production (that which is supported by upwelled NO3) is reduced by Si limitation and thus export from pelagic primary production is controlled by Si availability. In coastal upwelling regions, it has been demonstrated that iron limitation will cause diatoms to increase the Si/N uptake ratio, depleting the water of Si, leading to secondary Si limitation. Clearly,
Si is very important and can influence N response of the primary producers. Comparative Si availability may be a major feature in the apparent eutrophication response seen in nearshore waters. In a recent 20-year comparison in the Bay of Brest, there was a large decrease in the Si/N ratio, "but, contrary to what has been observed in other coastal ecosystems, phytoplankton stocks have not increased" (LePape et al., 1996). In light of the discussion in Section III.B, maybe this is less of an exception than LePape et al. interpreted. In some cases, an increase in phytoplankton biomass is seen, but not always with a shift from diatoms, and rarely is there an increase in higher trophic level consumption of the primary production. In an extensive study of a very long term record for the Mississippi River outflow into the Gulf of Mexico, Rabalais et al. (1996) have shown a large decrease in the Si/N ratio accompanied with an increase in primary production but also an increase in the deposition of biogenic silica in the sediments underlying an increasingly large hypoxia region. The explanation is that with relative Si scarcity, diatoms that are in the plankton are not grazed efficiently, and they fall to subsurface waters and contribute to hypoxia.

C. Changing Nutrient Ratios

Probably most estuarine waters with impact from human activities show greatly changed N/Si as well as N/P ratios. In Table II, average values for total dissolved inorganic nitrogen (NO₃ plus NH₄), PO₄, and Si are shown for several nutrient-enriched estuaries. All of the examples show large increases in DIN and most have smaller proportional increases in P so that the N/P is usually considerably higher than would be the case without the anthropogenic influence. Since Si is not usually a byproduct of human activity, the Si concentration has not changed much; there is probably a large natural variation depending upon the nature of the land drained for the estuary. A few systems probably have had significant decreases in Si due to decreased natural land erosion (dams, diked river banks); this is definitely the case with the Mississippi (Rabalais et al., 1996). As a result, the N/Si ratio is much different from that prior to human impacts. Inverting this as Si/N, the pristine condition is about 10/1 and most of our nutrient-enriched systems show values of 1/1 or lower. This very likely has a serious negative impact on the primary production community. The importance of Si in relation to eutrophication has been recognized in the past, but usually only in relation to shift from diatom to flagellate flora (e.g., Officer and Ryther, 1980). With more recent information on interactive influences of Si, N, and Fe and on the fate of Si-limited diatom production, it is timely to reinvestigate the role of Si on eutrophication. While species responsible for HABs do not necessarily show greater affinity for nutrients in general, giving them ability to outcompete more "normal" phytoplankton like diatoms, it is probable that changing ratios of N and P to Si do favor some of the HAB flagellates (Smayda, 1990).

The large changes in N/P ratios are often not documented because of lack of complete nutrient records from long-range monitoring. In a data set from the Delaware Estuary, dissolved inorganic nitrogen has been measured regularly along the full axis of the estuary for over 35 years, but parameters for phosphorus measurements have varied over that period. Total P, a composite that includes dissolved organic and particulate phosphorus as well as PO₄, has been measured.

### Table II

<table>
<thead>
<tr>
<th>Estuary</th>
<th>DIN</th>
<th>PO₄</th>
<th>Si</th>
<th>N/Si/P</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scheldt</td>
<td>550</td>
<td>19</td>
<td>250</td>
<td>37/17/1</td>
<td>Zwolsman, 1994, 1999</td>
</tr>
<tr>
<td>Delaware</td>
<td>230</td>
<td>5</td>
<td>125</td>
<td>40/20/1</td>
<td>Sharp, unpublished data</td>
</tr>
<tr>
<td>Mississippi</td>
<td>114</td>
<td>7</td>
<td>108</td>
<td>15/14/1</td>
<td>Rabalais et al., 1996</td>
</tr>
<tr>
<td>Chesapeake</td>
<td>73</td>
<td>1</td>
<td>30</td>
<td>57/30/1</td>
<td>Maloney et al., 1996</td>
</tr>
<tr>
<td>Northern San Francisco Bay</td>
<td>40</td>
<td>2</td>
<td>200</td>
<td>20/100/1</td>
<td>Peterson et al., 1983</td>
</tr>
<tr>
<td>Princeton</td>
<td>10</td>
<td>0.9</td>
<td>100</td>
<td>20/100/1</td>
<td>Fanning and Meynard, 1978, Meybeck, 1982</td>
</tr>
</tbody>
</table>

*Average concentrations of nutrients (in μmol/L) approximated from publications listed. Averages for total dissolved inorganic nitrogen (DIN), dissolved phosphorous (PO₄), and silicate (Si) and ratios normalized to P are listed. Values for pristine estuaries approximated from data for the Zaire and Magdeleni River outflow systems.
A comparison of the N/P ratio change, based on the total P, over a 30-year period is shown in Fig. 3. Recognizing that the majority of the P in the estuary today is PO₄, and that in the past PO₄ was probably a larger portion of the total, it is possible to view the N/P ratio as indicative of available P. This dramatic N/P ratio change is probably largely due to reduced input of detergent phosphorus and the same change has occurred in many U.S. estuaries (N. A. Jaworski, unpublished data, 1998). In the earlier situation, almost the entire estuary would appear to be replete in relation to P since the N/P was considerably below the Redfield ratio; in the more modern situation, N/P ratios are in the 30–60 range. However, it must be recognized that transport and availability of phosphorus in estuaries is a complex function that also involves geochemical influences. In the past 20 years, the P geochemical reactivity in the Delaware Estuary has changed due to increased pH and dissolved oxygen. As a result, the N/P ratio based on PO₄ only for average concentrations of the entire salinity gradient of the estuary has decreased in the past 20 years from about 90/1 to 40/1. More thorough analysis of many estuaries may show this dual trend of long-term decrease of N/P loading in an upper estuary but of more available dissolved PO₄ being delivered to the lower estuary. It is important to understand the full biogeochemical picture of estuarine phosphorus before accurate conclusions of nutrient impacts can be made.

In addition to long-term changes in nutrient ratios, there are large spatial changes in estuaries at any single time. Figure 4 shows nutrient ratios along the full length of the Delaware Estuary from sampling in the spring. In the spring bloom condition in the estuary, NH₄, PO₄, and Si are exhausted from the mouth of the estuary moving up toward a strong-light-limiting turbidity maximum (Pennock and Sharp, 1994). The very high ratios of P and Si to N are due to the large excess NO₃ concentration of the river water as it is advected downstream. It is interesting to see that different regions of the estuary appear to have large differences in the nutrient that could be most limiting. Also, from this picture, it appears that Fe could be more limiting in the upper estuary than near the mouth of the bay. This greater proportional availability of Fe is a year-round occurrence. A noxious algal group that has caused considerable international concern recently is responsible for brown tides. A recent analysis of brown tide occurrence suggests the macronutrient levels are not implicated but that Fe is.

In the Delaware Estuary, the area of the greatest phytoplankton production throughout the year is the lower bay (Pennock and Sharp, 1994). Figure 5 shows
macronutrient ratios for this region averaged on a seasonal basis. The maximum biomass is achieved in the spring bloom, where N/P and Si/N ratios indicate exhaustion of P and Si with residual NO₃; at this time, nutrient regeneration is minimal and there is also little herbivore grazing. Usually, the highest seasonal primary production is found in the summer. At this season, it would appear that grazing controls chlorophyll (Pennock and Sharp, 1994) and grazing is in balance with nutrient regeneration, which is sufficient to allow measurable NH₄, PO₄, and Si levels. At this season, the N/P ratio is close to Redfield and the Si/N ratio is sufficient to support healthy diatom growth although the flora is dominated by small flagellates. Acknowledging the caution of Hecky and Kilham (1988) that nutrient concentration does not equal nutrient utilization, we have demonstrated limitation in several ways (Pennock and Sharp, 1994). Looking at the entire estuary with highly nutrient enriched tidal freshwater region, a light-limited turbidity maximum in the oligohaline region, and clearer and nutrient-diluted lower bay, it would appear that the only time and place that near to Redfield ratio of nutrients is found is in the lower bay in the summer. We have performed preliminary simple mesocosm experiments and find that the CNP ratios approach Redfield stoichiometry only in the lower estuary in the summer (Sharp et al., unpublished data). In the lower estuary in the spring bloom, close to Redfield ratios of NPSiO occur but there appears to be an accumulation of C with the biomass accumulation. In the nutrient-rich upper estuary, nothing close to expected stoichiometry is seen. Further research on this anomalous stoichiometry is currently underway in my laboratory.

V. CONTAMINANTS AND STRESS

The lack of high growth rates of estuarine and coastal phytoplankton in the presence of high nutrient concentrations leads to the suggestion that anthropogenic contaminants, other than nutrients, may have an influence on ecosystem response. The MERL research facility mentioned earlier has extensive and expensive controls of the large tanks for the studies. A number of excellent experiments have been carried out also in less sophisticated smaller mesocosms in other estuarine areas. One interesting study done in Chesapeake Bay mesocosms over several seasons showed that nutrient enrichment by N and P caused growth of 'beneficial' diatom species over flagellates (Sanders et al., 1987). A conclusion of that study was that many other factors probably control community dynamics. As discussed earlier, the lack of predation is often a factor in noxious algal species proliferation. Also discussed earlier is the suggestion that ratios of nutrients may be more important than quantities in causing eutrophication responses. It is possible also that contaminants cause stress to the "desirable" algal species and to grazers that would otherwise consume the primary production.

Most estuaries with large anthropogenic influences have had chronic exposure to many chemicals in addition to nutrients. Chronic exposure to arsenic appeared to cause reductions in phytoplankton cell size with less trophic transfer while chronic exposure to silver resulted in essentially the opposite. Thus, arsenic would appear to be partially responsible for eutrophication response. Many estuaries also have frequent or continuous inputs of chlorination byproducts which must have a selected influence. Sanders (1984) showed that one diatom and a chrysophyte would not grow in aged chlorinated water (with levels below detection for total residual chlorine) where a more resistant flagellate would grow. This result would also favor species other than "normal" with potential to decrease trophic transfer.

Paerl (1998) has illustrated positive and negative interactions and feedback from nutrient loading, emphasizing that negative influences on grazing can increase the impact of primary production. In lakes, it has been shown that combinations of nutrient additions and zooplankton size can have major influence on phytoplankton sizes and thus on trophic transfer. The variability in nutrient and trace metal impacts on phytoplankton, bacterioplankton, heterotrophic protozoa, copepods, fish, and benthos caused variable trophodynamic responses in estuarine mesocosm experiments (Breitburg et al., 1999). These authors concluded that trace elements may mask the response of high nutrient loadings in eutrophic systems. In lake studies, N, P, and C had different controlling effects on zooplankton, phytoplankton, and bacterioplankton, with variable responses in different seasons.

An overall conclusion is that eutrophication as a stress in estuarine and coastal marine environments is not a simple cause-and-effect phenomenon. Nutrient enrichment elicits complex and variable responses from the phytoplankton, bacterioplankton, and protozoa that make up the primary producer community. Increased nutrient concentrations and loadings may cause an overall increase in phytoplankton biomass, but not invariably from higher growth rates. Shifts in dominant species and shunts through the microbial loop may decrease the trophic transfer to higher metazoan levels.
The biodiversity of the aquatic community may allow some resilience to the system (Patrick, 1988), but ultimately changes in the diversity are probably more important than a direct response to increased levels of nutrients.

See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • ESTUARINE ECOSYSTEMS • LAKE AND POND ECOSYSTEMS • PLANKTON, STATUS AND ROLE OF • RIVER ECOSYSTEMS

Bibliography


MARINE ECOSYSTEMS

J. Frederick Grassle
Institute of Marine and Coastal Sciences, Rutgers University

GLOSSARY

benthic Pertaining to the bottom of the sea or other aquatic environment.
benthos Organisms living on, in, or near the seabed or at the bottom of some other aquatic environment.
coastal Estuaries, semi-enclosed seas, and shallower regions of the ocean, including areas influenced by rivers and runoff from land.
community A group of species co-occurring in an area and interacting through trophic and spatial relationships.
coral reef Benthic environments characterized by reef-building corals with symbiotic dinoflagellates.
deep sea Volumes of water or areas of ocean bottom at depths greater than 200 m.
ecosystem A community of organisms and their physical environment interacting as an ecological unit.
habitat The locality or three-dimensional space occupied by an organism.
mangrove Environments characterized by mangrove trees.
nekton Actively swimming pelagic organisms.
pelagic Pertaining to the water column in aquatic environments.
plankton Organisms that float freely in the water column and do not maintain their position independent of water movements. Phytoplankton (literally plant plankton) is plankton with photosynthetic pigments and zooplankton is animals of the plankton.

MARINE ECOSYSTEMS may be defined as major units of ecological function in the marine environment. Ecosystems are communities of organisms and their physical, chemical, and geological environment—distinct assemblages of species coevolved with a particular environment over long periods of evolutionary history. As units of function, ecosystems have measurable imports and exports of material and energy. In comparison to ecosystems on land, ocean ecosystems have less clearly defined boundaries, a greater variety of major taxonomic divisions of organisms, and a long evolutionary history that preceded colonization of land. As the diversity of life in the oceans is explored, the importance of previously unrecognized aspects of ocean circulation, flux of energy and materials, and bottom characteristics to marine ecosystems are becoming better understood.
I. MARINE ECOSYSTEMS

A. Ecosystem Units

On land, ecosystems are separated into two-dimensional biomes, land areas defined by characteristic primary producing plants such as trees, grasses, and shrubs. Most shallow lakes and streams are similarly two-dimensional; however, a few freshwater deep, ancient lakes, such as Lake Baikal in Siberia, and large rivers such as the Amazon have spatial complexity comparable to many coastal marine ecosystems. The ocean biosphere has an average depth of 4 km and comprises 99.3% of the biosphere. The dense seawater medium allows at least part of the life cycle of almost all marine organisms to be transported and dispersed by ocean currents. One ocean phylum is entirely pelagic, and about a third of the ocean phylla have representatives that spend their entire life cycle in near-surface waters as plankton. The boundaries that define ocean habitats and communities may involve a variety of overlapping criteria such as depth, distance from land, separation by landmasses, ocean currents, water masses of characteristic salinity and temperature, depth, and sea bottom topography. In addition, interactions with and between land and rivers and patterns of ocean circulation, light, nutrients, hydrology, and physical energy of water movements can strongly influence the distribution of species.

Descriptions of species boundaries are few and biogeographical classification depends heavily on the groups of organisms considered and how well they have been sampled. The ocean generally lacks the obvious barriers to dispersal characteristic of terrestrial environments. There may be multiple criteria for defining biogeographical provinces or marine ecosystems.

Major estuaries, where fresh water from rivers mixes with ocean water, are among the smallest individual ecosystem units in area. The largest units are regions defined by major boundary currents such as the Gulf Stream, Kuroshio, and Brazil currents, and the north and south subtropical ocean gyres (the Sargasso Sea and South Atlantic Gyre in the Atlantic and the North Pacific Subtropical and South Pacific Subtropical Gyres in the Pacific). In the far north, the Arctic Ocean ecosystem is a distinct ocean basin covered by ice and the southern ocean around Antarctica is separated from the circulation of the Atlantic, Indian, and Pacific Oceans by the cyclonic circulation of the Antarctic Circumpolar Current.

As with terrestrial environments, marine ecosystems may be classified by their characteristic primary producers (i.e., single-celled phytoplankton that float in the near-surface layers of the ocean, marsh grasses, sea grasses, mangrove trees, seaweeds such as those forming kelp beds, the single-celled plants called zooxanthellae that live symbiotically with corals, and the chemosynthetic bacteria living in water, sediments, or symbiotically with other organisms at hydrothermal vents or other seep environments rich in chemically reduced compounds such as sulfide or methane).

Using combinations of coastline, coastal bathymetry, ocean current systems, surface winds, and biota, the near-surface pelagic layer of the ocean where primary productivity occurs has been classified into 31 provinces (Fig. 1) by Longhurst (1998). Similar criteria have been used to classify coastal areas (Briggs, 1974). Marine sediments cover almost the entire surface of the ocean floor, yet a consistent global biogeographic classification of these benthic ecosystems has yet to be developed (Snelgrove et al., 1997).

B. Comparison of Marine Environments with Land

The ocean occupies 71% of the surface area of the globe and the deep sea at depths below 200 m occupies 63.3% of the earth’s surface. Seawater is 830 times more dense than air and supports most of the biomass in the ocean. The volume of seawater in the ocean provides 99.5% of the livable volume of the earth (Cohen, 1994). Concentrations of near-surface chlorophyll in the ocean are measured according to wavelengths of light reflected from the surface of the ocean, which are sensed by earth-orbiting satellites. Extensive studies of the relationship between near-surface chlorophyll and primary production allow satellite-derived information on chlorophyll to be converted to maps of primary productivity. Until very recently, overall primary production was thought to be approximately half that on land. Using distribution of chlorophyll in satellite photographs and models, primary productivity of the oceans has been shown to be about the same as that on land (~45–50 Pg C per annum in the ocean and ~53 Pg C per annum on land; Falkowski et al., 1998). For regions without ice cover, average net primary productivity (NPP) per area in the ocean is a third of that on land (ocean: 140 g C m⁻² year⁻¹, and land: 426 g C m⁻² year⁻¹). Only about 1.7% of the ocean surface area has NPP greater than 500 g C m⁻² year⁻¹ compared to 29% for land. Most productivity in the marine environment is from phytoplankton. Attached, multicellular algae contribute only about 2%. The highest productivity occurs in estu-
aries and upwelling areas—these highly productive areas contribute approximately 18% to net ocean primary productivity. In the open ocean, the greatest primary productivity is near the equator and at midtemperate latitudes in the Northern Hemisphere where there are regional maxima in terrestrial productivity. A smaller peak in productivity occurs in the Southern Subtropical Convergence where physical processes supply high concentrations of nutrients to surface waters (Falkowski et al., 1998; Field et al., 1998).

Marine primary producers are small and mobile whereas terrestrial primary producers are mostly large and rooted in the ground—trees account for approximately 80% of the primary production in terrestrial systems. By contrast, in central ocean gyres, phototrophic bacteria less than 2 μm in diameter and short generation times account for most of the primary production. Oceanic biomass is extremely dilute and filtering of organic particles is an important mode of feeding in marine environments.

Oceanic food webs have an average food chain length of nearly six trophic links as opposed to four trophic links in terrestrial systems (Cohen, 1994). The number of species of smallest marine organisms, such as the various groups of one-celled marine organisms, are extremely poorly known. The relationship between the spectrum of individual body size and the spectrum of rates of population growth differs in marine and terrestrial systems (Fig. 2). In open ocean food webs, the hierarchy of size is not apparent at the lower trophic levels because of the broad overlap in size of consumers and primary producers (Fig. 3, Karl, 1999).

The pattern of temporal variability of the physical environment differs between oceans and land. Marine ecosystems are characterized by about the same environmental variation over weeks and years as over days—variability is constant at frequencies ranging from days to decades. In terrestrial environments the variance of environmental parameters (e.g., temperature) increases steadily from frequencies of hours to millennia. Beyond 50 years the variance increases with increasing frequency as it does over the entire time spectrum on land (Steele, 1983).

On land, individual organisms have a high probability of surviving the relatively predictable patterns of environmental variation that occur over time periods up to decades. For example, individual trees and many vertebrate animals resist adverse effects of variation at all frequencies up to several decades because of their large size and long generation time. In the open ocean, time series measurements at a single station show that primary production varies significantly on periods from days to decades (Karl, 1999). Both seasonal and daily differences in cloud cover may result in three-fold variation in light at the surface. Vertical displacements of phytoplankton by internal waves further increase the amount of light absorbed by seawater before it reaches the photosynthetic organisms, creating a further source of variability. Small bacterial and flagellate primary producers have reduced the adverse effect of this variation.
in light by supplementing their diet from the pool of dissolved organic matter excreted by other organisms.

Other distinctive features of marine populations are outlined by Cohen (1994) and in a U.S. National Academy of Sciences book on marine biological diversity (National Academy of Sciences, 1995). Plant and animal populations in marine ecosystems generally spend part of their life cycle as floating or swimming stages in the plankton. Unlike most terrestrial systems, the connections between benthic and planktonic life-history stages assume great significance and there is an unusually broad range of dispersal abilities, reproductive rates, and generation times. Almost all species have the ability to disperse in the water column as larval stages produced by some form of sexual reproduction. As a consequence, marine ecosystems are largely open and distant marine habitats can be linked by dispersing larvae. Terrestrial systems are more localized functionally and localized extinction of species occurs more frequently. Invertebrate predators and grazers generally have very high reproductive output, which makes population fluctuations more likely. Fluctuations at the highest trophic levels affect interactions among species at successively lower trophic levels. This cascading effect often has unpredictable consequences, and even the lowest trophic level of primary producers may be controlled from the top down. Bottom-up control of food webs is exerted through the effects of nutrients and physical processes on primary productivity.

II. BIODIVERSITY OF MARINE ECOSYSTEMS

A. Higher Taxa

The three main biological lineages are the Bacteria, Archaea, and Eukarya (includes plants, fungi, protists, and animals). Recent advances in molecular-biological techniques permit the first measurements of highly diverse oceanic assemblages of bacteria and archaea that cannot presently be cultured in the laboratory. Bacteria are more abundant in the photic zone and archaea are more abundant in deeper water. The Eukarya (all taxa except the Bacteria and Archaea) are divided into 71 well-defined monophyletic groups with no apparent taxonomic affinity with one another on the basis of cell organization (Patterson, 1999). Each of these groups includes taxa formerly assigned to the protists. By this classification animals and their relatives the choanoflagellates, and fungi and their relatives the chytrids, are defined as a single group. Plants are in another group altogether with 11 categories (~7000 species) of green algae.

Important groups of primary producers have affinities with several other monophyletic groups. The red algae are a distinct group with about 4000 known species; the ~1000 species of dinoflagellates are related to the ciliates. The ~10,000 species of diatoms are in a highly diverse lineage that includes kelps and other brown algae. The conspicuous red, green, and brown seaweeds of rocky shores are divided among three separate lineages. The two most important primary producers in the open ocean were formerly called blue-green
algae. They are actually prokaryotic bacteria in two groups: the Synechococcus with three lineages, and the Prochlorococcus group with two lineages. These organisms account for most of the phototrophic standing stock and primary production in the open ocean (Andersen et al., 1996).

Among the many nonphotosynthetic unicellular marine organisms, the ubiquitous Foraminifera are common both on the bottom at all depths and as pelagic organisms. Two abundant, poorly described benthic groups, the Komokiacea and the Xenophyophora (~40,000 known species), are big enough to be seen on the surface of deep-sea sediments. A leaflike form of Xenophyophora may be as large as 25 cm in diameter. These groups are separate lineages with no obvious relatives.

In the classification of marine, free-living, multicellular animals there are 29 phyla. Figure 4 (modified from May, 1994) compares the described diversity and abundance among marine benthic, marine pelagic, freshwater, and terrestrial environments. Of the 29 known Phyla, all are known to have lived in the ocean and 14, or about half, are known only from the ocean. Living representatives of the Phylum Onychophora are presently found only on land in the Southern Hemisphere, but are also known from fossil organisms that...
lived in the ocean more than 300 million years ago. Species diversity on land is dominated by insects and trees, groups that play a significant role only at the margins of the marine environment. Only about 13% of described species are found in the marine environment, but this may reflect the much greater cumulative effort devoted to species descriptions on land, rather than an actual difference in the number of species (May, 1994).

B. Species

Species are the basic units of evolution and represent the biological variability for future generations of life. For whole collections, species diversity is measured as the number of species and their relative abundances within and between habitats, regions, or other ecological or geographical units. Species richness is measured by collecting enough samples to represent very large numbers of individuals over very large areas. Ideally, communities should be sampled until the rate at which new species are found declines, and a plot of species versus area approaches a constant number of species. This level of sampling effort is achieved for groups with few rare species (e.g., larger animals including most vertebrates, planktonic organisms, and macrophytic plants). For species-rich taxa of bottom-dwelling invertebrates from coral-reef or deep-sea habitats, this level of sampling has not been attained. Where habitats are patchy and the vast majority of species are rare, it is seldom possible to collect and process enough samples to estimate species richness accurately.

For individual samples, indices based on the absolute number of species and the relative abundance of species are used to study species diversity. The most commonly used index is the Shannon-Wiener information function, $H'$, which equals the frequency of each species $p_i = \frac{n_i}{N}$, multiplied by $\log(p_i)$, summed over the number of species $n$ collected (e.g., $\Sigma n \log(p_i)$). Another measure, Hurlburt rarefaction, calculates a species versus
individuals curve for each sample based on the expected number of species in successively smaller subsamples drawn from an actual sample. These species diversity curves are especially useful in comparing samples of unequal size.

There are approximately 200,000 described species of animals in the marine environment (Table I). The most species-rich and least well known areas are coral reefs and the sediments of the deep-sea floor. There are no precise estimates for these environments but estimates for coral reefs alone exceed 600,000 species (Reaka-Kudla, 1997). Based on quantitative analysis of 233 box core samples from the Atlantic Ocean continental slope and rise off the east coast of North America, Grassle and Maciolek (1992) estimated 1 to 10 million macrofaunal species in the deep sea (Gage and Tyler, 1991). May (1994) estimated 0.5 million based on the portion of species previously undescribed in the Grassle and Maciolek study. Poore and Wilson (1993) analyzed samples from the Southern Pacific Ocean off Australia and, on the same basis, estimated that there are 5 million species in deep-sea sediments. Multicellular animals small enough to pass through a 1 mm sieve (meiofauna), such as nematode worms, are even less well known and Lambshead has argued that there may be 100 million species if nematodes are included (Lambshead 1993). Reasons for high diversity of species in the ocean include the long evolutionary history of the ocean, the vast area of deep-sea floor (3 \times 10^8 km^2) with relatively few barriers to dispersal, and the episodic nature of patch formation within and between habitats on a variety of spatial and temporal scales.

### Genetic Diversity

Genetic diversity is the heritable variation among individuals measured as allelic diversity at a broad sampling of genetic loci or as genetic sequence information at the molecular level within populations. Genetic variation occurs among subpopulations as well as within populations. Differentiation among subpopulations results from natural selection for genetic variants adapted to local patterns of environmental variation or random loss of genetic variants in small isolated subpopulations. Species with relatively high rates of dispersal are less likely to form subpopulations and species with very poor dispersal ability are more likely to diverge from parent populations as a result of random processes. In coastal areas, genetic divergence is related to the length of life of dispersal stages and barriers to current flow from one place to another along a coastline. For some shallow-water species, genetic isolation of island populations is related to distances among islands. The archipelagos in the central Indopacific in the vicinity of Indonesia and Papua New Guinea have a high richness of species, which then declines eastward to relatively isolated peripheral island archipelagos (Planes and Galzin, 1997; Stehli, 1965). In the same region, in a study of population differentiation in four species of sea urchins,
Palumbi (1997) found high genetic diversity (mitochondrial DNA sequence diversity) in the central area (1.6% variation among individuals) and much lower genetic diversity (0.5% variation among individuals) in peripheral island localities to the east. For these species, genetic diversity and species diversity covary across gradients suggesting a similarity in the processes maintaining gradients in diversity despite different mechanisms for the origin of the variation. Fluctuations in population size in relatively isolated populations could result in both loss of genetic variants and reductions in number of species (Palumbi, 1997).

In the deep-ocean, hydrothermal vents are analogous to islands in the sense that these fluid flows support widely separated biological communities, linearly aligned along the Mid-Ocean Ridge. The patterns of deep-sea ocean currents that transport dispersal stages of species restricted to hydrothermal vents are poorly understood, but it is possible to make estimates of gene flow from the extent of genetic differentiation among populations of individual species. The flow of hydrothermal fluids, containing energy-rich reduced compounds such as hydrogen sulfide, supports chemosynthetic primary productivity. At East Pacific Rise vents, the flow of hydrothermal fluid may last only a decade or two at any one site and all populations are maintained by dispersal over considerable distances. Species can be divided into three categories: those that show no geographic pattern of genetic differentiation, those that are isolated by distance, and species without a free-living larval dispersal stage, which apparently have good dispersal to sites along a single ridge segment but poor dispersal between separated ridge segments (Vrijenhoek, 1997). The latest methods for measuring genetic diversity have been applied to very few marine species and rapid advances in this area of research can be expected.

III. ECOSYSTEM FUNCTION

It is useful to classify members of species assemblages according to their feeding relationships with other species in the ecosystem. A trophic unit includes all species that eat the same kinds of foods or are consumed by the same kinds of consumers. Within a food chain, there is a hierarchy of consumers from primary producers to primary consumers followed by a further sequence of consumers. Each step in a food chain results in a reduction in biomass, and simple food chains are often described as a pyramid with plants at the base and apex predators at the top. In the water column, unicellular phytoplankton form the first trophic level of marine food chains. The second level is formed by herbivores and detritivores and subsequent levels are formed by successive levels of predators. Species at the highest trophic levels can affect the food web relationships among species at lower levels. For example, removal of a top predator can have cascading effects on herbivores and ultimately on primary producers.

Because of the dilute seawater medium, a great many marine species have developed both active and passive means for filtering or trapping food particles from the dilute seawater medium. Copepods, the most common animals in the water column, have filtering appendages and gelatinous zooplankton cast mucous nets to feed on phytoplankton. Baleen whales filter zooplankton (krill) from the water column. On the sea bottom, clams and sea cucumbers pump water past internal filters and many animals in sediments pump water through burrows in order to feed. Other bottom animals have appendages protruding above the sediments that trap or filter food particles. In many marine organisms, the distinction between producers and consumers is blurred. Reef-building corals use their tentacles to trap zooplankton yet may take most of their sustenance from photosynthetic dinoflagellates living symbiotically in their tissues. Other animal-plant relationships of this sort are found in tropical clams and one-celled radiolarians and foraminifera.

Some marine species play another important functional role by providing habitat for other species, either on a large spatial scale—as with coral or coralline algae reefs, polychaete worm reefs, seagrasses, kelps, marsh grasses, and mangrove trees. On a smaller scale, bionomic sediment structures (tubes, burrows, mounds, fecal aggregations) and more persistent structures made by tube builders, sponges, or shell-bearing animals may serve as habitat for other species.

Some species significantly affect the ecosystem by regenerating nutrients that limit primary production. Burrowing animals release nitrogen into the water column and stimulate phytoplankton growth. In chemically reduced sediments, animals pump water into sediments for respiration or feeding and supply oxygen to chemosynthetic primary producers living in the burrow. The role of single species is often not obvious, and several different criteria may be used to assign species to functional groups within an ecosystem. In general, redundancy of ecosystem function within a functional group has the potential to stabilize ecosystem processes despite fluctuations in the environment. Loss of functional groups implies drastic changes in ecosystem function.
IV. ECOSYSTEM DIVERSITY

A. The Edge of the Ocean

1. Intertidal Beaches

Beaches can be classified according to topography, organic content of sediments, and wave action. Reflective beaches are dominated by low wave energy, low organic content, and coarse sand. Reflective beaches have waves 1 m high or less and are generally found on open coasts with deep embayments, tropical coasts, and coasts of polar seas. Surging wave action filters and drains large volumes of water through the interstices of the sediments, resulting in well-flushed and highly oxygenated coarse sand deposits (Alongi, 1998). Dissipative beaches, at the other extreme of a continuum, are produced by a combination of high waves (>2.5 m) and fine sand deposits with higher amounts of organic matter. These are common on the west coasts of Australia and Southern Africa and seasonally on the west coast of North America where high wave swells and fine sands are abundant. Intertidal sand and mudflats are common on dissipative beaches.

Many beaches have adjacent seagrass beds, kelp beds, or other sources of macrodetritus, which are deposited as thick layers of wrack on the beach. These accumulations support communities that include both marine and terrestrial invertebrates (e.g., beach hoppers, beetles, and kelp-flies larvae). Other beaches are more dependent on growth of diatoms in the sediments and input of small, filterable organic particles. Many animals live in the sediments, and in some high energy situations animals such as mole crabs and small bivalves move up and down the beach with the tides filtering particles from the waves. Large areas of sand flats, such as the Wadden Sea in the Netherlands, may be especially productive and support high standing stocks of grazing invertebrates.

2. Kelp Beds

Kelps attach to the bottom and form a surface canopy at depths up to ~20 m. Under the most favorable conditions these large marine plants form subtidal forests and attain rates of primary production in excess of 1000 g C m⁻² d⁻¹. These forests provide protection and food for a rich community of fish and invertebrates. The biomass and abundance of kelps may be regulated by wave action resulting in communities of attached seaweeds and filter-feeding bivalve mollusks, such as mussels that provide physical structure for other species. Wave energy enhances the productivity of these ecosystems by continually renewing nutrients and food. The shore face and the organisms that reside on the shore can be divided into zones according to tidal height and length of exposure to air and the interactions of the dominant species with herbivores such as snails (gastropod mollusks) and predators (particularly snails, starfish, and birds). The large-scale pattern of rocky-shore communities depends on the distribution of rocky outcrops and sporadic changes in climate resulting in unusually heavy waves, ice cover, or sedimentation from rivers. The interaction of physical change and biological relationships among species at a variety of spatial scales (from local to regional) and temporal scales (from annual storm events to interdecadal climatic change) are most clearly worked out for rocky intertidal ecosystems.

3. Rocky Shores

Rocky shores are exposed to the open ocean and characterized by wave action resulting in communities of attached seaweeds and filter-feeding bivalve mollusks, such as mussels that provide physical structure for other species. Wave energy enhances the productivity of these ecosystems.

4. Coral Reefs

Coral reef ecosystems occur where conditions are favorable for growth of reef-forming corals with dinoflagellate primary producers living symbiotically in their tissues. Growth of corals over many generations in geologic time results in major limestone structures such as coral atolls or the Great Barrier Reef off Australia. Dense growths of coral can sometimes occur in the deep sea, but these species lack photosynthetic symbionts, grow relatively slowly, and do not form major reef structures. Reefs grow in strong light and clear water at temperatures from 18°C to 30°C at latitudes between 30°N and 30°S. Coral reefs are adversely affected by high nutrient concentrations, runoff of sediments from land, direct removal, and overfishing. The midrange of primary production of corals in combination with their symbiotic dinoflagellates is about 25 g C m⁻² d⁻¹ and...
varies greatly from species to species. Over large areas, net primary productivity of the most actively growing reef crests and slopes ranges from 1 to 5 g C m$^{-2}$ d$^{-1}$. Reefs support an enormous species richness and complexity of interactions among species. Conspicuous large animals include enormous coral heads and large fish such as groupers, stingrays, and manna rays. Many of the colorful reef fish do not move far and develop complex behavioral relationships both within and between species. Some live symbiotically with other species, for example, individual anemone fish live in close association with patches of anemones. Cleaner fish set up cleaning stations where they feed on the ectoparasites attached to the gills of other fish. Some species mimic the cleaner fish and take bites out of the fish expecting to be cleaned of parasites.

B. Continental Shelves

Continental shelf coastal areas, on the order of 10,000 km$^2$ or more, have been called “large marine ecosystems” (Sherman, 1993). These are separated from other areas of the ocean by continental shelf depth and ocean currents, and the shapes of coastlines form major seas, bays, or gulfs. Examples include the Baltic, North, Mediterranean, Black, Caspian, Red, Arabian, Barents, Bering, Okhotsk, Japan, Yellow, East China, Sulu, Celebes, and Caribbean Seas; Bay of Bengal and Walvis Bay; and Gulfs of Alaska, California, and Mexico. Primary productivity in these systems ranges from below 35 g C m$^{-2}$ yr$^{-1}$ in the low latitude, warm waters of the Red Sea and high latitude, cold waters of the Beaufort Sea (10–20 g C m$^{-2}$ yr$^{-1}$) to the very high primary productivity of Eastern Boundary Current upwelling areas in the Southern Hemisphere (1000–2000 g C m$^{-2}$ yr$^{-1}$) of the Peru Current and Walvis Bay (Walsh, 1988).

Most of the world’s major fisheries are on continental shelves in midlatitudes.

C. The Open Ocean and Deep Sea

1. Pelagic

The largest ecosystems in the ocean are the central gyres of the Atlantic, Pacific, and Indian Oceans. Ecosystem processes in the North Pacific Subtropical Gyre (NPSG) have been summarized by Karl (1999). This ecosystem is the largest circulation feature on the planet (2 $\times$ 10$^7$ km$^2$) and one of the most persistent, its boundaries having remained approximately the same for the past 10$^4$ years. The NPSG has a clockwise circulation of less than 4 cm s$^{-1}$ and forms a circumscribed, stable, and relatively homogenous habitat. The surface mixed layer varies from 40 m to 120 m depth and is characterized by surface temperatures are 24°C or higher low nitrate concentrations but relatively high dissolved organic nitrogen, and low standing stocks of organisms. The zone of primary productivity can be divided into two layers: an upper layer where chlorophyll increases in the winter and decreases in the summer and lower layer (100–175 m) where chlorophyll increases in the spring and declines in the fall. Recharge of nutrients is from deeper water below as a result of vertical eddy diffusion and episodic mixing events leading to considerable spatial variability in mixing processes and nutrient concentrations varying by as much as three orders of magnitude. Phytoplankton primary production was once thought to be mostly by Eukaryotes (diatoms and flagellates), but is now known to be more than 90% from the small bacterial taxa Prochlorococcus and Synechococcus. The standing stock of these autotrophic bacteria groups comprise 80% of chlorophyll a and feed a microbial loop that internally regenerates nutrients and maintains a pool of dissolved organic matter, which supports them (Fig. 4). The abundance of these auto-heterotrophs is controlled by light, nutrients, and predation by bacteria and a mixed assemblage of protists. Viral infection may also be an important source of mortality for these organisms. Archaea are abundant but it is not clear whether these are significant chemosynthetic primary producers because little is presently known about these organisms.

Very little organic matter escapes remineralization and the microbial loop provides negligible subsidy to the rest of the food web. The classic food chain pathway of eucaryote phytoplankton to copepod herbivores and on to higher trophic-level fish is ephemeral and occurs more frequently in surface waters during the summer. Organic matter produced by the eucaryotic phytoplankton food chain produces most of the exportable carbon during aperiodic, pulsed events. Falkowski et al. (1998) provide a summary of biogeochemical processes controlling primary production in the open ocean. The central ocean gyres in the Atlantic, Pacific, and Indian Oceans have been considered analogous to deserts on land with low primary productivity and contain only 0.2 mg m$^{-3}$ of chlorophyll. Coastal upwelling regions, seasonally mixed regions of temperate and boreal seas, divergent subpolar gyres, and mesoscale features with eddy-induced pumping have sufficient vertical flux of nutrients to support 3 mg m$^{-3}$ of chlorophyll. Throughout most of the coastal and open ocean, primary production is limited by the availability of inorganic fixed nitrogen. In some instances, the cyanobacteria that fix nitrogen in the open ocean are limited
by iron and an important source of iron to the ocean is dust carried from land by winds. Limitation of primary production by lack of iron is especially notable in the South Pacific (Falkowski et al., 1998).

2. Benthic

The deep-sea floor is divided into major ocean basins by continents and the Mid-Ocean Ridge. Communities within ocean basins may be further divided according to depth, sediment type, and level of energy of deep-sea currents. The deep ocean floor is the least-known part of the planet but, through use of manned and unmanned submersibles, distinct ecosystem processes at hydrothermal vents, continental margin seeps, seamounts, ocean trenches, and areas of strong bottom currents are being explored and described.

The largest ocean basins and deep ocean trenches each have some species that live only in that basin and nowhere else. Hydrothermal processes along the Mid-Ocean Ridge mix seawater through porous rock at high temperatures yielding an energy-rich fluid containing reduced compounds. These compounds support chemosynthetic microorganisms that provide primary production for a discrete ecosystem clustered around each hydrothermal vent. Flow of subsurface fluid seeps out of sediments deposited along some ocean margins providing similarly energy-rich fluid to chemosynthetic organisms.

The food supply for the deep sea comes from the productivity of surface waters. When diatoms bloom, or gelatinous animals such as salps multiply rapidly, they die and sink, so that organic matter accumulates in low areas of the uneven surface of the sea floor and in burrows and depressions left by the larger inhabitants. Even in the central ocean gyres where export production is low, the dead remains of fish, marine mammals, or terrestrial plant material carried seaward sink and form widely separated organic patches on the sea floor. Species responding to these patches at different rates and the probability that two species reach the same patch at the same time is low. This reduces the likelihood of species competing and of one species eliminating another. Most deep-sea species are small and many species, including most fish species, are relatively slow growing, long lived, and late in maturation. Attempts to sustain deep-water fisheries have proven unsuccessful because low rates of population growth cannot keep up with rates of removal.

Species that grow relatively fast (characteristically respond to patchy but concentrated sources of food from the ocean surface, such as wood from rivers, or the bodies of pelagic animals that settle to the bottom. For example, wood-boring bivalves rapidly colonize pieces of wood, grow to maturity in a few months feeding on their wood habitat, and produce thousands of eggs and larvae to colonize the next piece of wood that settles to the sea floor. Other species of bivalves grow very slowly in relatively homogeneous sediments, take several decades to reach maturity, and may produce only one egg at a time—in contrast to the rapid maturation and production of millions of eggs produced by most shallow-water bivalves.

Submarine canyons form conduits for sediment from continental shelves into the deep ocean. Unpredictable events of sediment erosion or scouring by intense currents result in relatively few species in the soft sediments at the bottom and sides of canyons. Seamounts are undersea mountains formed by the same processes at the hot spots on the ocean floor that form volcanic islands. Seamounts often support large populations of fish, and more than 70 species of commercially important fish have been reported. Interactions of currents with the steep topography of seamounts results in areas of enhanced primary productivity and concentrations of zooplankton that provide food for fish and dense concentrations of bottom animals (Rogers, 1994).

D. Mid-Ocean Ridges and Hydrothermal Vents

The 40,000 nautical mile Mid-Ocean Ridge system is the largest feature on the deep-sea floor. In 1977 a unique ecosystem was discovered at sites where a plume of high-temperature fluid rich in reduced compounds pours out into the water column. It is now known that sulfur oxidizers are among the most numerous bacteria and form a major base of the food chain. Other energy sources include reduced iron, manganese, and hydrogen. In the Pacific, large, red-plumed worms up to 2 m long and large clams and mussels dominate the vents. These animals feed on organic compounds produced by symbiotic sulfur bacteria living in their tissues. Vents in the Atlantic have some of the same kinds of animals, but the most conspicuous are shrimp, which swarm over the surface of vent chimneys. Vents usually have a restricted distribution on any given ridge segment and persist for about 10 to 20 years, until there is local extinction of the vent community. Animals colonize new vents rapidly, grow fast, and produce enough offspring to colonize the next vent. In comparison with the rest of the deep sea, few species have adapted to the extreme thermal (4°C up to temperatures in excess of 150°C), chemical (high concentrations of cadmium, lead, cobalt, and arsenic) conditions at hydrothermal
vents (Grassle, 1986). Most species found at hydrothermal vents live exclusively in this environment. Of the 443 species found at hydrothermal vents, 15 have been found in other sulfide-rich environments and only 30 species are known from elsewhere in the deep sea (Tunnick et al., 1998).

V. POTENTIAL CONSEQUENCES OF ANTHROPOGENIC CHANGE

A. Eutrophication

Eutrophication is the increase in the rate of supply of organic matter to an ecosystem. Increases in global inputs of nitrogenous fertilizers and the mining of phosphate rock have generated increased concern about the effects of eutrophication on enclosed marine ecosystems (Nixon, 1995). Eutrophic ecosystems have algal production in excess of 300 g C m$^{-2}$ y$^{-1}$, which results in areas of anoxia and loss of habitat for fish and other organisms. Relatively high rates of denitrification on continental shelves remove excess nitrogen originating from land sources and, in concert with dilution, help prevent adverse eutrophication effects in open coastal areas.

B. Overfishing

Globally, about 30% of commercial fish stocks are overfished and another 44% are being fished at or near the maximum potential long-term catch rate. Atlantic halibut, cod, orange roughy, and many species of salmon are now severely depleted. Significant changes in community structure as a result of overfishing have occurred in ecosystem structure in the Bering, Barents, and Baltic Seas (National Academy of Sciences, Committee on Ecosystem Management for Sustainable Marine Fisheries, 1999). Bottom-fishing has been shown to result in physical destruction of some bottom habitats. Fishing gear, when dragged over the bottom, levels structures such as worm tubes, burrows, and shell hash necessary for the survival of many species. Overfishing has resulted in major changes in coral reef ecosystems. Normally, herbivorous fish heavily graze the attached algae, ensuring enough open reef surface for corals to settle and grow. This is especially true following major storms when wave action reduces coral coverage and circumstances are favorable for rapid algal growth. In the Caribbean, under normal circumstances, sea urchin grazing may compensate for reductions in fish grazing. A combination of overfishing and the decimation of sea urchin grazers by disease favored algal growth following a hurricane, which has resulted in reefs dominated by algae (National Academy of Sciences, 1995).

C. Invasive Species

Unwanted, exotic species are sometimes introduced to new geographic regions both deliberately to start new fisheries and accidentally through release from aquaria or ballast water carried by ships, sometimes with disastrous consequences. The Asian clam became established in the San Francisco Bay in 1986 and quickly displaced other species from large areas of the seabed and altered the water chemistry of the bay (National Academy of Sciences, 1995). The introduction of predatory green crabs to coastal environments on the east coast resulted in major reductions in shellfish beds. In short, invasive species have become a significant problem in many marine coastal environments and considerable effort is needed to curb this severe problem. In summary, the oceans encompass a broad array of habitats that differ in their diversity, function, and vulnerability. Much of the vast area of the oceans is poorly described, but we have some understanding of a variety of globally essential ecosystem processes, and species loss may threaten not only the organisms themselves but also the many ecological processes that serve the rest of the planet and its human populations.

See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • ENDANGERED MARINE INVERTEBRATES • ESTUARINE ECOSYSTEMS • INTERTIDAL ECOSYSTEMS • INVERTEBRATES, MARINE, OVERVIEW • MANGROVE ECOSYSTEMS • MARINE ECOSYSTEMS, HUMAN IMPACT ON • PELAGIC ECOSYSTEMS • REEF ECOSYSTEMS • VENTS

Bibliography


MARINE ECOSYSTEMS, HUMAN IMPACTS ON
Juan Carlos Castilla
Pontificia Universidad Católica de Chile

I. Introduction
II. Human Impacts on Marine Communities and the Effects on Species Diversity and Functioning
III. Nonanthropogenic Environmental Changes and Variability
IV. Conclusions

GLOSSARY

alien: introduced, exotic, nonindigenous, nonnative, invasive species A species that has been transported by human activity (i.e., mariculture), intentionally or accidentally, to a site at which it does not naturally occur.
ballast water Water carried by a vessel to improve stability.
benthic organism An organism pertaining to the seafloor; bottom-dwelling.
biodiversity The variability among living organisms from all sources and the ecological systems of which they are a part.
disturbance Any relatively discrete event in time that disrupts ecosystem, community, or population structure and changes resources, substrate availability, or the physical environment.
ecosystem A complex nonlinear community of organisms in their physical environment.
ecosystem engineer species Species that directly or indirectly modulate the availability of resources (other than themselves to other species) by causing physical state changes in biotic or abiotic material and in so doing modify, maintain, and/or create habitats.
eutrophication Enrichment of a body of water with nutrients causing excessive growth of phytoplankton, seaweed, or vascular plants and often accompanied by a depletion of oxygen.
food web, trophic web A network of interconnected trophic chains in a community. A network of consumer–resource interactions among a group of organisms, populations, or aggregate trophic units.
guild A group of species having similar functional roles in the community (i.e., herbivores).
keystone species A group of species whose effects on the structure, dynamics, and functioning of the community is disproportionately large relative to its abundance.
pelagic organism A free-swimming (nekton) or floating (plankton) organism that lives exclusively in the water column.
resilience The resistance to a disturbance of a system and the speed of return to an equilibrium point, or the disturbance that can be absorbed before the system changes in structure by the change of variables and processes that control system behavior.
species diversity The number of species in a given community (= species richness) and the way the species' abundances (i.e., number, biomass, and cover) are distributed among species (= species evenness).
trophic level Feeding level in a food chain or pyramid (e.g., carnivores).
MARINE ECOSYSTEMS represent the greater part of the earth’s total biological system. At the present time these marine communities are threatened by human effects, both direct and indirect, such as resource extraction (e.g., fishing), introduction of alien species, pollution, and water temperature modification. These effects demonstrate the unique ability of humans to profoundly influence the status of ecosystems.

I. INTRODUCTION

The main threats to marine ecosystems are the human alteration of habitats, the excessive extraction of resources, pollution (Castilla, 1996), invasive species (i.e., introduction through mariculture and ballast water; Cohen and Carlton, 1998), eutrophication, and nonanthropogenic environmental changes (National Research Council (NRC), 1999; Castilla and Camus, 1992). Furthermore, multiple and compounded perturbations related to physically and biological based disturbances are resulting in communities entering new domains or “ecological surprises” (Paine et al., 1998), with important modifications in their structure (i.e., species composition) and dynamics (i.e., alternative states).

Single, multiple, or compounded impacts on ecosystems may directly or indirectly affect their structure, including species diversity and functioning. Ecosystems are complexly linked nonlinear systems and their dynamics may be sensitive to past conditions and subjected to shifts when exposed to anthropogenic and nonanthropogenic environmental stress (NRC, 1999).

The concept of biological diversity (biodiversity; Heywood, 1995) is defined as: the variability among living organisms from all sources and the ecological system to which they are part. The analysis of biodiversity considers four levels: genetic, species, community, and ecosystems. This article focuses on the species diversity (richness, the number of species in a given community; evenness, species abundance), community resilience, and ecosystem functioning. One of the best avenues to integrate species diversity functioning and community resilience (Holling, 1973) is to study their dynamics through long-term manipulations. The article reviews long-term experiments and impacts on marine communities and ecosystems in which humans are one of the key ecological factors (Castilla, 1999).

II. HUMAN IMPACTS ON MARINE COMMUNITIES AND THE EFFECTS ON SPECIES DIVERSITY AND FUNCTIONING

A. Rocky Intertidal Communities

Castilla (1999), based on a 16-year intertidal human exclusion experiment in central Chile (Las Cruces fenced Marine Coastal Preserve; ECIM), summarized the ecological roles played by humans as top predators on rocky mid-intertidal marine communities. The functional intertidal food web, without humans (inside the ECIM preserve) and with humans (outside ECIM), differed substantially. On these rocky shores the impact of intertidal food gatherers is significant (Durán et al., 1986). The collectors target mainly the keystone murexid snail *Concholepas concholepas*, locally known as ‘‘loco’’ (Castilla et al., 1998). The high density of locos inside ECIM, following its closure to collectors in 1982, resulted in strong loco predation on the competitive dominant mussel *Perumytilus purpuratus*, which cannot “escape in size” from its predator. Therefore, a few years after the fencing of ECIM, the original dense mid-intertidal mussel beds inside ECIM were almost completely eliminated by the locos (Castilla, 1999). The primary space, so liberated, was readily invaded by two species of barnacles, *Jehlius cirratus* and *Notochthamalus scabrosus*, and several species of algae. Despite the fact that the loco also consumes barnacles, they have persisted for several years since they have a “weed recruitment strategy” (Castilla, 1988): After removal they keep reinvading the shore. This is not the case for *P. purpuratus*, which requires special substratum conditions to reinvade the shore (Navarrete and Castilla, 1990). Following the closure of the rocky shore at ECIM, species richness and evenness of sessile organisms using primary substrata decreased inside ECIM. Outside ECIM (control), under reduced loco density due to food gathering, primary space is still dominated almost exclusively by the competitive dominant mussel *P. purpuratus*, and the biological diversity of the sessile primary substrata users is reduced since the mussels are long-term winners and appropriate the rock resource (Fig. 1). Castilla (1999) provided a detailed account of direct and indirect human impacts on these communities and discussed differences in their functioning. For instance, it was noted that the settlement of keyhole limpets, *Fissurella spp.*, was indirectly negatively impacted inside ECIM since their recruitment substratum, the beds
of the mussels *P. purpuratus*, were absent due to *loco*'s direct predatory impacts (Fig. 1).

Nevertheless, in the papers previously noted, no mention was made that rocky intertidal species diversity should be viewed in a more comprehensive way so as to include the secondary substrata generated by *P. purpuratus*, an ecosystem engineer species (Jones et al., 1994). Mussel matrices allow for the establishment of a rich macroinvertebrate and algal community composed of dozens of species (Paredes and Farazona, 1980; Lohse, 1993) which live inside the matrices and on mussel shells. Although in central Chile this effect has not been evaluated, the *P. purpuratus* matrices enhance species richness (for southern Chile, see López and Osorio, 1977) in sites impacted by humans (outside ECIM) compared to those not impacted (inside ECIM, J. Castilla, unpublished results).

Similar ecological direct and indirect human impacts and drastic modification in rocky intertidal species evenness and intertidal community functioning (Fig. 2) have been reported at Mehuín’s southern Chile coastal preserve (Moreno et al., 1984). Lindberg et al.
(1980), through manipulative and "natural" experiments, described a three-trophic-level interaction among the American black oystercatcher (*Haematopus bachmani*), limpets (*Lottia spp*.), and erect fleshy algae in rocky intertidal bench communities of central and southern California. Human disturbances, such as the selective collection of large-size limpets and the reduction of shorebirds (in shores frequented by humans), drive the communities to a state dominated by small limpets and high cover of fleshy algae. Intertidal benches in relatively isolated islands (e.g., San Nicolas in central California) with large densities of oystercatchers and an absence of limpet human collection present communities in a different alternative state, which is characterized by large-size limpet populations and comparatively reduced fleshy algal cover.

**B. Rocky Subtidal Communities**

The Cape rock lobster *Jasus lalandii*, commercially the most important lobster species in South Africa, causes profound direct and indirect effects on subtidal competitive dominant mussel species, such as *Choromytilus meridionalis* and *Aulacomya ater* (Griffiths and Seiderer, 1980), severely modifying species diversity and community functioning. Barkai and Branch (1988a, b) compared the nearshore benthic communities of two adjacent islands on the west coast of South Africa: Malgas and Marcus Islands (33°S, 18°E), which are approximately 4 km apart. The biotas of both islands have been protected from human exploitation since 1929. In the 1960s both islands supported populations of rock lobsters, but later, due to overfishing, a management plan...
was established which included a catch quota. Currently, Malgas still supports an unusually dense population of J. lalandii (probably partly due to the management plan) with densities of up to 10 individuals per square meter, whereas Marcus has a very reduced adult population of lobster. The benthic communities of both islands have only 34% of species in common. The benthos of Malgas is dominated by numerous species of algae, whereas that of Marcus consists of thick beds of the black mussel C. meridionalis, an autogenic ecosystem engineer species that has a rich and diverse associated fauna (Barkai and Branch, 1988a). At Malgas, the predatory lobsters have eliminated a large proportion of spatial competitors, including mussels and barnacles, and sea urchins are absent. As a consequence, macroalgae proliferated. At Marcus, due to the absence of lobsters, the competitive dominant C. meridionalis formed dense beds, outcompeting other species of mussels, such as A. ater and algae; sea urchins are common (Castilla et al., 1994). Barkai and Branch (1988a, b) discussed this ecological situation and argued for the existence of alternative stable states on the contrasting islands. Figure 3 provides a summary of the main species involved, relative biomass, and direct, indirect, positive, and negative interactions between organisms on both islands.

The ecological impact of the Cape rock lobster at Malgas was experimentally demonstrated by Barkai and McQuaid (1988). The experiments showed that the drastic community differences between the islands were due to the dense population of lobster at Malgas and its absence at Marcus. In fact, the introduction of 1000 lobsters at Marcus ended amazingly: The lobsters were attacked by thousand of snails, Burunepia sp., which exist at Marcus in densities of up to 250 per square meter, and the lobsters perished within 30 min. This may explain their absence at Marcus, supporting the existence of an alternative ecological state.

In South African waters, it is unknown to what extent the commercial exploitation of rock lobsters or conservation measures (i.e., coastal closures) have impacted the nearshore rocky subtidal communities or in how
many cases (other than Marcus and Malgas Islands) alternative stable states have been reached. This is a classical example in which both extreme attitudes—overexploitation and total conservation (no-take areas)—can result in drastically different species diversity and community functioning, mediated by the role of a high-trophic-level predator.

C. Humans and Linkages between Coastal and Oceanic Waters

*Enhydra lutris*, the northern sea otter, is found in near-shore environments ranging across the Pacific rim from Hokkaido (Japan) to Baja California (Mexico). The exploitation of their pelts led to the near extinction of otter populations in approximately 1911, when unregulated hunting was ended. Since then, the recovery of otter populations has occurred, particularly in the Aleutian Island chain, where by the 1970s the populations reached near maximum densities in some areas, were growing rapidly in others, and remained absent from others. Otters as keystone species (Power et al., 1996) control the local biomass and the abundance of sea urchins, which regulate benthic algae biomass and productivity. Aleutian interisland comparisons (Estes et al., 1998) have shown that kelp deforestation occurred in islands with low sea otter densities due to the increased density of sea urchins, whereas islands with high sea otter densities showed high kelp biomass. Estes et al. reported the complete transformation of a subtidal kelp forest in islands of the Aleutian Archipelago from three to four trophic-level systems and the release of sea urchin populations from the limiting influence of their predator, *E. lutris*. In the original circumstances, in the absence of sea otters, sea urchin populations increased rapidly and overgrazed the kelp forest, setting in motion a suite of different ecological impacts which drastically transformed the coastal ecosystems. These transformations had implications in the functioning of the communities and affected species diversity. Humans are highly involved in Estes et al.’s findings. In recent years in western Alaska, declines of *E. lutris* populations have been observed. The authors have argued that this is probably due to the recent increased predation on sea otters by killer whales, *Orcinus Orca*. Orcas may have initiated predatory influences that cascaded down successively lower trophic levels, first through the reduction of densities of sea otters, which triggered the increase of sea urchin populations, and ultimately the depletion of kelp biomass due to overgrazing. Estes et al.’s paper includes documented information on declines of sea otter populations and increases in the density and intensity of grazing of sea urchins on the kelp beds. Sea otters and killer whales have cohabituated the Aleutian Archipelago for millennia and Estes et al. attributed the sudden change of behavior of killer whales to a shift in their prey resource base. This has probably resulted from the collapse of pinniped populations, such as the Stellar sea lion and harbor seals, which were among the killer whale’s main prey items. It has been suggested that the pinniped populations may have collapsed due to the Northwest Pacific midwater-trawl overfishing of walleye pollock (*Theragra chalcogramma*) (Alverson et al., 1994) and/or increases in the ocean temperature. The authors recognized that some of their arguments contained speculations and that the critical one refers to the direct/indirect impacts of humans on marine ecosystems. In fact, sea otters, pinnipeds, and whales are under national and international protection in the Aleutians through different treaties and agreements signed dozens of years ago, but it also has to be recognized that their food resources have been depleted independently through overfishing. For instance, there is evidence that in the case of the pinnipeds a reduction (population collapses in some cases) has occurred mostly due to overfishing of pinnipeds, or of their fish resources, and also to climate changes. Overfishing is directly linked to human activities, and in Estes et al.’s scenario, humans and not killer whales may be considered as the apex predator. Humans have redirected the functioning of oceanic and coastal marine ecosystems in these localities and modified trophic linkages. These examples indicate that there are at least two aspects of human ecological influences on marine communities that are difficult to evaluate and hence demonstrate an indisputable cause-effect situation. First, in many cases, the functioning of the marine communities is affected indirectly by anthropogenic activities—for example, human overfishing of pinnipeds’ fish resources, collapse of pinniped populations, a shift in the prey item of killer whales, predation on the sea otter, population explosion of sea urchins, and overgrazing of kelp beds. The cascading down to successively lower trophic levels is complex and requires long-term observation and experiments to be understood. Furthermore, nonanthropogenic impacts also need to be considered. Second, limited knowledge exists on the resilience properties of marine communities and ecological conclusions on linkages between marine ecosystems are based on preliminary data.
Ecosystem engineer species are species that directly or indirectly modulate the availability of resources (other than themselves) to other species by causing physical state changes in biotic or abiotic materials, and in so doing they modify, maintain, and/or create habitats (Jones et al., 1994). Jones et al. distinguished (i) autogenic engineers, when the changes in the environment occurred via their own physical structure, living or dead tissues (e.g., coral reefs), and (ii) allogenic engineers, when they produced changes in the environment through the transformation of living or nonliving materials from one physical state to another via mechanical means (e.g., rabbits and burrows). In marine coastal communities, there are numerous autogenic engineer species playing roles in the functioning of the community and ecosystem and creating the physical conditions for other species to exist (e.g., mussels; Lobel, 1995). In the Southern Hemisphere, rocky littoral zone tunicates of the genus Pyura play such a role (see Fielding et al., 1994, for P. stolonifera in S. Africa). These tunicates are also important as species extracted for food and/or bait by recreational fishers, divers, and intertidal food gatherers (for Pyura praepatulis in Australia, see Fairweather, 1991; for P. praepatulis in Antofagasta, northern Chile, see Castilla, 1998). The tunicates form dense intertidal and subtidal belt monocultures and attain collective cemented beds, creating microhabitats used by several dozen macroinvertebrates and algae. Fielding et al. identified 83 taxa of macroinvertebrates and algae in intertidal and subtidal Pyura stolonifera beds around Durban, South Africa, whereas more than 100 taxa of macroinvertebrates and algae have been found in intertidal P. praepatulis beds in Antofagasta. The P. praepatulis beds in Chile present a very restricted geographical distribution of only 60–70 km around Antofagasta Bay (Clarke et al., 1999). According to a working hypothesis (J. Castilla, work in progress), the species might have been introduced recently to Antofagasta by ships or arrived on floating objects from Australia. In Antofagasta, a contrasting situation concerning species richness is found in sites with P. praepatulis, with more than 100 taxa in the Pyura beds, as opposed to sites without the tunicate, which have about one-third to one-fourth of the species. It is unknown how much ecological damage, if any, human extraction causes on the dynamics of Pyura populations or on species diversity at a local scale. However, preliminary information at Antofagasta indicates that following Pyura removals by waves, predators, or humans, the species reinvades intertidal sites (the center of its distribution) within 1 year (J. Castilla, work in progress). A higher rate of anthropogenic or nonanthropogenic removal of engineer species than the rate of recovery may be key to local species diversity. Invasive species are displacing native species throughout the world. They are altering the physical nature of habitats (e.g., the effects of the Asian clam Potamocorbula amurensis in the San Francisco Bay) and causing changes in food webs of economically important species (NRC, 1999). The best reported case is that of the Bay of San Francisco, in which ship activities (i.e., the elimination of ballast waters) have increased drastically the number of exotic species in the bay's benthic communities (Carlton, 1996). At the pelagic level the introduction in the bay of the zooplanktonic mysid Acanthomysis sp., which displaced another species of mysid, Neomysis mercedis, a major food item of the striped bass Morone saxatilis, is partly responsible for a severe decline in the bay's bass population (NRC, 1999). Furthermore, there are recent reports showing that the predator green crab Carcinus maenas has invaded the San Francisco Bay and is spreading throughout the coastal waters of California (Cohen and Carlton, 1998).
St. Lawrence that caused serious oyster stock depletions were ascribed to the transplant of oysters in 1914 from New England to Canada (Edwards, 1990). No comprehensive ecological reports on the ecological effects of these species introductions and diseases on local species diversity or community functioning have been published.

The intensive farm-raising of high-value species, such as shrimp and salmon, is far from trouble-free. There are concerns about the increase in the deposition of particulates and accumulation of organic matter under salmon cages in intensive mariculture installations due to unwanted effects, such as anoxic conditions and the production of toxic gases (Beveridge, 1996). Coastal ecosystem destruction, nutrient loading, antibiotics wastes, accidental release of alien or genetically altered organisms, and disease spreading to native species are some of the threats to community and ecosystem functioning.

F. Human Overfishing, Diseases, and Trophic Cascades

Hughes (1994) and Jackson (1997) reported major ecological effects on coral reef communities as a consequence of the overexploitation of herbivorous fishes and a disease killing sea urchins. In Caribbean coral reefs, a chain of effects, starting with the overfishing of herbivorous fishes, appeared following category 5 hurricane Allen in 1980. Allen severely damaged coral reefs in Jamaica, but by 1983 there was evidence of their recuperation. Nevertheless, at that time a disease devastated the herbivorous populations of the sea urchin Diadema antillarum. The elimination of the herbivore guild caused dramatic food cascading effects, resulting in reefs overgrown by algae and the detention of their recuperation. Species diversity and community functioning severely changed: The coral cover was reduced from approximately 52% in 1977 to 3% in the early 1990s, and cover of macroalgae increased from approximately 3 to 92% (Hughes, 1994).

G. Pollution and Artificial Reefs

The cases exemplified are among the best known ecological situations in which human impacts and the function of communities or ecosystems, combined with changes in species diversity, have been observed or studied. However, there are additional examples showing anthropogenic negative, as well as positive, impacts on marine communities and ecosystems. Among negative impacts on marine species diversity and community functioning, the most conspicuous (not discussed here) is pollution (Castilla, 1996). Among positive impacts is the building of marine reefs for fishing enhancement and recreational purposes. Artificial habitats may locally enhance species diversity and resources and drive community structure toward alternative states (Buckley, 1982).

III. NONANTHROPOGENIC ENVIRONMENTAL CHANGES AND VARIABILITY

Nonanthropogenic environmental changes and impacts on marine populations and communities have been well documented. For instance, Souar and Isaacs (1974) reported large fluctuations in the density of scales of hake, anchovy, and sardines in sediment cores during the past 2000 years, well before fishing was a factor. Large-scale ocean climate changes, such as El Niño Southern Oscillation (ENSO) events, have dramatic negative (Arntz and Fahrbach, 1996) or positive (Castilla and Camus, 1992) impacts on fish, shellfish, and algae populations in the Southeastern Pacific. ENSO also causes multiple positive and negative oceanic, freshwater, and terrestrial impacts throughout the world.

Barry et al. (1995) reported changes between 1931 and 1933 and between 1994 and 1995 in species richness and evenness of intertidal invertebrates at a rocky intertidal transect in the Hopkins Marine Station, Monterey, California. They reported species’ latitudinal range shifting northward, suggesting a consistency with predictions associated with anthropogenic-linked climate warming (but see alternative explanation by Denny and Paine, 1998). Nevertheless, it is debatable whether the current global warming trend, due partly to the build-up of several greenhouse gases, is part of a long-term climatic trend. In any case, marine species with different geographical origins would have different responses to water temperature alterations (Castilla and Camus, 1992). Moreover, in the case of the oceans, water temperature modifications would be just one of the potential factors affecting the distribution of species. For instance, temperature effects on the turbulence of the ocean waters, and their association with wind stress, may
have major implications for plankton dispersal. Also, the predicted north–south interhemispheric asymmetry, due to the thermal inertia in the south, must be considered before drawing firm conclusions on marine species latitudinal shifts (Bernal, 1994). Furthermore, since the ocean is affected simultaneously by several climate forces (including anthropogenic greenhouse effects), it is difficult to determine the real cause of any observed change, such as that in surface seawater temperature. Shifts in marine populations, community structure, and their functioning represent the integrated response of species assemblages to nonanthropogenic long-term climate changes superimposed on the effects of numerous short-term factors, including anthropogenic forcing.

IV. CONCLUSIONS
This article discussed several marine examples in which direct anthropogenic and nonanthropogenic impacts (or combinations), such as species extraction and oceanic water temperature modifications, caused drastic ecological shifts on marine benthic intertidal, subtidal, and coastal–oceanic communities, and thereby modified species diversity and the functioning of associated communities. Interestingly, extreme conservation measures (e.g., the establishment of no-take areas) to protect species, habitat, community, or ecosystem may also cause drastic modifications in the functioning of marine communities and drive communities into alternative ecological states (Castilla et al., 1994; Estes et al., 1998; Castilla, 1990). This article highlighted that anthropogenic activities (e.g., mariculture) and impacts (e.g., overfishing) on different ecological categories of species (predator, keystone, invasive, and competitive dominant) translate into differential responses and functioning at the species diversity and community level. The unique ecological role played by humans and their apex keystone position in trophic webs were discussed.

Acknowledgments
I acknowledge financial support received through the Consejo Presidencial de Ciencia, Chile (awarded in 1997) and the Pew Charitable Trusts supported fieldwork in Antofagasta, Chile. Comments and suggestions on a preliminary manuscript made by a referee are acknowledged.

See Also the Following Articles
AQUACULTURE • INTERTIDAL ECOSYSTEMS • MARINE AND AQUATIC COMMUNITIES • STRESS FROM EUTROPHICATION • MARINE ECOSYSTEMS • RESOURCE EXPLOITATION, FISHERIES

Bibliography
Castilla, J. C., Branch, G. M., and Barkas, A. (1994): Exploitation of


Echeverría, E. (1990). Be careful with introductions. Fish Farming Int. LT


I. INTRODUCTION

A. Taxonomic Definition of “Marine Mammals”

The marine mammals include one extinct order and three major extant taxa that were or are fully aquatic, in most cases occurring entirely in the marine habitats of the major ocean basins and associated coastal seas and estuaries. In addition, a few species of largely terrestrial taxa are currently regarded as marine mammals. We consider 127 recent mammal species in total to be marine mammals for purposes of this review. We acknowledge that species numbers within any taxon are subject to revision as new systematic methods and philosophies emerge. Our primary bases for defining our list of marine mammal species are the protocols of the U.S. federal government, determined largely by the U.S. Marine Mammal Protection Act (MMPA) of 1972 [16 U.S.C. §§1361-62, 1371-84, and 1401-07 (Supp. IV 1974)] as amended (MMPA) and managed by two U.S. federal agencies, the National Marine Fisheries Service (NMFS) and the Fish and Wildlife Service (FWS). Our choice of defining criteria is arbitrary. Our principal source for taxonomic nomenclature, including common names, is the recent review of Rice (1998).

The order Cetacea includes whales, dolphins, and porpoises (Table I). The “pinnipedia” is a group of species in three families in the mammalian order Carnivora (Table I). The pinnipeds include the seals, fur seals, sea lions, and walrus. The term pinnipedia is no longer recognized formally by marine mammal taxonomists, but it continues to appear in the systematic vernacular as a matter of tradition and convenience. The order Sirenia includes the extant manatees and dugong and the extinct Steller’s sea cow (Table I). The order Desmostylia is the only recognized order of marine mammals to become entirely extinct.

Two largely terrestrial families of the order Carnivora also include species recognized as marine mammals (Table I). Sea otters and chungungos (family Mustelidae) live entirely or primarily in marine habitats. Polar bears (family Ursidae) also spend a significant proportion of time at sea.

Many other species of mammal utilize aquatic or marine habitats, including monotremes, ursids, mustelids, canids, primates, rodents, bats, and ungulates. Ultimately, the distinction among aquatic, marine, and terrestrial taxa is arbitrary. Thus, our reliance on definitions and protocols of MMPA, NMFS, and FWS is subjective, although it is consistent with common practice at least in the United States.

We use the term ‘marine’ to refer to large, contiguous aquatic habitats with significant dissolved salt content in ambient waters. Thus, we apply the term marine
Marine mammals are characterized by a number of striking modifications in anatomy, physiology, and ecology (Table II). In some cases, the modifications are sufficiently extreme that phylogenetic linkages to terrestrial ancestry are obscured and difficult to resolve. The degree of modification is correlated approximately with the duration of the evolutionary history of the major marine mammal taxa.

1. General Features and Habitat Boundaries

Compared to terrestrial mammals, marine mammals are characterized by many striking modifications in anatomy, physiology, and ecology (Table II). In some cases, the modifications are sufficiently extreme that phylogenetic linkages to terrestrial ancestry are obscured and difficult to resolve. The degree of modification is correlated approximately with the duration of the evolutionary history of the major marine mammal taxa.

Although marine mammals are largely defined by marked departures from the terrestrial mammalian model, it is instructive to consider major features of terrestrial mammals retained in marine mammals. In the context of extinction processes in general, and anthropogenic extinctions in particular, two retained features are of particular importance. First, although most marine mammals spend most of their lives immersed at sea, they retain largely terrestrial respiratory architecture and must surface and breathe in order to exchange respiratory gases. Second, marine mammals are homeothermic, with core body temperatures typically near 38°C, like their terrestrial relatives. The need to breathe at the surface and the need for major anatomical adjustment to minimize rates of heat loss are constraints that foster vulnerability to unsustainable rates of exploitation and to certain types of pollution. The significance of these constraints is developed in the case studies we present later.

The diving capabilities of marine mammals define the three-dimensional nature of their habitats at sea. Nearly all extant marine mammals dive to forage, although the ranges of diving capability and pattern are broad. Most marine mammals also spend signifi-
can time submerged while traveling, socializing, or breeding.

Among cetaceans, sperm whales and beaked whales likely dive the deepest and longest compared to other species. Sperm whales can dive to 1500 m, remaining submerged for 20 min or more. The diving behavior of beaked whales is poorly known, but there is emerging evidence that beaked whales may also routinely make repetitive dives of long duration to great depth. Baleen whales may make long deep dives during breeding season. Foraging dives of baleen whales normally are relatively shallow and brief. Many of the smaller cetaceans commonly dive for less than 10 min at a time to depths no greater than a few hundred meters.

Among pinnipeds, elephant seals (Phocidae) have maximum diving capabilities comparable to the sperm whales, and they are known to make remarkably long sequences of repetitive deep (to 1300 m), long (20 min or more) dives with surface intervals of only 2 or 3 min. These sequences may be maintained day and night for tens of days at a time. Many other phocid seals are thought to have similar capabilities. The sea lions and fur seals (Otaridae), in contrast, usually dive for only a few minutes at a time, and usually to maximum depths of a few hundred meters, although many otariids are known to be capable of continuous sequences of repetitive shallow dives of 10–12 hr or more. Walruses are known to dive as deep as 80 m, with maximum durations of 10 min.

In contrast to cetaceans and pinnipeds, sirenians are weak divers, normally remaining in shallow water (<20 m) and diving for only 2 or 3 min when active. Deeper dives (to 70 m) may occur on occasion, and dive duration can be quite long (up to 24 min) when animals are resting at the bottom. Sea otters are capable of diving to 1000 m depth and remaining submerged for a maximum of approximately 5 min, although most dives are to 30 m or less and last only for 1 or 2 min. To our knowledge, there are no data available on the diving capabilities of the chungungo.

Few field observations of Steller’s sea cow were made prior to extinction, but morphological analysis suggests that sea cows were unable to dive below the sea surface, surviving instead by foraging on macroalgae floating on the surface. Polar bears are able to make shallow dives but do not typically engage in the extended repetitive dive sequences typical of many marine mammals. They apparently do not forage while diving. Polar bears instead use stealth, quickness, and great strength to capture seals, their primary prey, at seal breathing holes on the ice surface. When confined in ice-bound seas with small breathing holes, beluga whales and narwhals are also taken as food by polar bears. Thus, the extent to which the at-sea habitat of marine mammals is truly three-dimensional varies widely among the major taxa and the individual species. Within species, there is also marked ontogenetic variation in diving capability and pattern.

The marine mammals are geographically ubiquitous in the world’s oceans, seas, and estuaries. Cetaceans occur in marine environments at all latitudes. For example, killer whales and minke whales may have the largest natural geographic ranges of the earth’s mammals. Most of the mysticetes and some of the larger odontocetes have global ranges or are distributed antitropically. Smaller cetaceans are widely dispersed as well, although individual populations typically concentrate in regions of predictably high local biological productivity. Several species of small cetacean, including two delphinids, a phocoenid, and the four monotypic families of river dolphins, are found in major river systems in South America and Asia. Beluga whales also spend significant time in river habitats. Pinnipeds occur in all the world’s major marine habitats, but most species are concentrated in middle or high latitudes, in close association with regions of high productivity. In addition, there are several pinniped species or populations confined to isolated large lakes in Europe, Asia, and North America. Most sirenians are limited to tropical or subtropical latitudes, in shallow seas that provide adequate macrophytic food and refuge from predation, and are thermally tolerable. The sea otter is confined to the coastal North Pacific Rim and the chungungo to the temperate coastal southeastern Pacific. Polar bears occur only in the Arctic and subarctic, rarely traveling south of 60°N latitude except in the relatively frigid northwestern Atlantic and Hudson Bay.

2. Cetacea

The distinguishing anatomical and functional features of the cetaceans are summarized in Table II. The two major taxonomic subdivisions of cetaceans are the suborders Odontoceti, or “toothed whales,” and Mysticeti, or “baleen whales.” The odontocetes are the most diverse of the major marine mammal taxa, with 10 families and 71 species (Table I). The best known families are the Delphinidae and Phocoenidae. In ecological terms, the family Ziphiidae is among the most sparsely studied groups of mammals on Earth. The mysticetes (Table I) include species that, by a wide margin, are the largest animals in the earth’s history. The blue
Most cetaceans dwell in the open sea or in the seas and estuaries of the continental margins. The exceptions are two delphinids, a phocoenid, and four odontocete species known as river dolphins (Table 1). Of the river dolphins, one species, the Franciscana or La Plata dolphin, is found mainly in coastal marine waters of Brazil, Uruguay, and Argentina. The other three river dolphins are exclusively aquatic. One of these, the boto, is known to leave river channels and travel within the adjacent flooded forests of the Amazon basin during the wet season. The Irrawaddy dolphin, a delphinid, occupies coastal marine habitats and major river systems in southeastern Asia, some Indo-Pacific islands, and the northeastern coast of Australia. The tucuxi, another delphinid, occupies the Amazon watershed and coastal marine habitats of tropical Atlantic South America and southern Central America. The finless porpoise, a phocoenid, occurs in the Yangtze River watershed and other large southern Asian rivers and coastal marine habitats from the Persian Gulf to Japan.

Mysticetes often segregate feeding and breeding activities, both in time and space, connecting the two categories of activity with extensive seasonal migration. Feeding is done primarily at high latitude during summer, and breeding and parturition are done primarily at low latitude during winter. Thus, a significant poleward migration is required during spring, and an equatorward return trip is necessary in autumn. The segregation of feeding and breeding, and the associated migratory behavior, is best developed and understood in the largest cetaceans. However, some large mysticetes, such as the bowhead whale of the Arctic region and the Bryde’s whale of tropical and subtropical latitudes, undertake only modest seasonal migrations in contrast to species such as the humpback whale or gray whale. The largest mysticetes are adult males of the highly dimorphic southern elephant seal, exceeding 5 m in length and reaching 3700 kg in mass. Adult male northern elephant seals are only slightly smaller. Walruses are also quite large, reaching 3.5 m and 1300 kg. Despite many anatomical, physiological, and ecological features obviously associated with life at sea, pinnipeds are best regarded as amphibiaous. All species utilize solid substrata for breeding or for postbreeding maternal care. Solid substrata are also used as short-term resting sites and for protracted periods in some species during molting of the skin and pelage. Although the proportion of time spent on land (“hauled out”) over the long term varies significantly among species and age and sex categories, generally the pinnipeds spend a major portion of their lives on land near shore or on pack or shorefast ice at sea.

Otarids use terrestrial habitats near shore for breeding, postpartum maternal care, molting, and resting. Preferred hauling sites are those near areas of high oceanic productivity and those free of large terrestrial predators. Thus, hauling grounds for otarids typically are islands or mainland locations protected by cliffs or rough terrain from land predators. Often, the hauling grounds are localized at high latitude or the upwelling zones of mid-latitude eastern boundary currents, such as the Humboldt, California, and Benguela currents, where production of preferred food species is high and temporally predictable. Optimal hauling grounds for breeding otarids are few in number and often limited in size. Breeding activities of otarids are highly synchronous, occurring during a narrow time window when food availability for lactating females and newly produced juveniles is seasonally optimal. As a consequence of the various spatial and temporal constraints, pinniped breeding typically involves conditions of extreme crowding on haul-out sites. Phocids and walruses haul out for the same purposes as otarids. Unlike otarids, phocids and walruses use two very different kinds of substrata. About half of the phocid species use coastal land for hauling grounds, selecting sites for largely the same reasons described for otarids. Thus, timing and location of breeding for
land-breeding phocids and otariids are generally similar. As a consequence, reproductive activities for many land-breeding phocids also occur under conditions of extreme crowding. Known exceptions include some populations of harbor seals with spatially dispersed, largely aquatic breeding systems and the three recent species of monk seal (one now extinct) with temporally asynchronous breeding systems at low latitude. The walrus and the remaining phocids breed or care for young, rest, and molt on ice at high latitudes rather than on land. Ice as a substratum varies widely over time and among locations in stability, vulnerability to predators, and provision of access to the surrounding sea. Thus, among ice-breeding pinnipeds there are significant resultant variations in social and breeding strategies and in the degree of crowding at hauling sites. A major predator of ice-hauling phocids, the polar bear, is present in ice-covered marine habitats of the Arctic region but not the Antarctic. This pattern has many interesting consequences for interhemispheric differences in the ecology of ice-hauling pinnipeds.

Some phocid species or subspecies occur only in aquatic habitats. Two subspecies of ringed seal occur only in Lake Saimaa, Finland, and Lake Ladoga, Russia, respectively. The Caspian seal is found only in the Caspian Sea and the Baikal seal only in Lake Baikal, both in Russia. A population of harbor seals occurs year-round in Lake Iliamna, Alaska, but the degree of exchange, via river connection to populations of harbor seals in nearby Bristol Bay, is unknown.

4. Sirenia

Several of the defining anatomical and functional features of the sirenians are convergent with those of the cetaceans (Table II). However, in contrast to the cetaceans there are few Holocene sirenians (Table I). The sirenians are the only extant herbivorous marine mammals, sharing common ancestry with desmostylians and terrestrial subungulates (e.g., aardvarks, elephants, and hyraxes).

Sirenians are large in body mass and linear dimension compared to most terrestrial mammal species. The Steller’s sea cow was the largest of the modern species, reaching 7 m in length. To our knowledge, body mass of the Steller’s sea cow was never directly measured, but the estimated maximum is 10,000 kg. Maximum adult lengths and masses of the three manatee species range from 3 to 4 m and from 450 to 1600 kg, respectively. Adult dugongs reach maxima of 3.3 m in length and 400 kg in mass.

All modern sirenians are fully aquatic or marine and are incapable of leaving the water. Sirenians feed on large plants growing on the bottom, in midwater, at the surface, or closely overhanging the surface. They forage exclusively in shallow habitats. Manatees utilize freshwater, estuarine, and fully marine habitats, often interchangeably. Use of rivers by manatees is influenced by rainfall patterns and river discharge rates. Manatees generally concentrate in habitats that are relatively warm and physically protected from extremes of weather and sea. Dependence on relatively warm-water temperatures may result from the combination of obligate homeothermy and a relatively low basal metabolic rate compared to that of other marine mammals. The limited tolerance of low water temperature likely contributes to seasonal migration and a tendency to concentrate at high density in warm-water refugia during the winter. Manatees may also congregate near sources of fresh water, although fresh water is not a physiological requirement. The smallest of the modern sirenians, the Amazonian manatee, occurs only in the freshwater habitats of the Amazon River watershed of South America. Dugongs are fully marine and forage primarily on benthic seagrasses in shallow coastal tropical marine habitats. Steller’s sea cow, known only from isolated parts of the subarctic North Pacific, was the most aberrant of the holocene sirenians. Sea cows likely foraged exclusively on kelps and other large algae along exposed shores.

5. Desmostylia

Desmostylians are the only known extinct order of marine mammals. The small number (<10) of recognizable species in the fossil record are of Oligocene and Miocene age and are confined geographically to the North Pacific region. Desmostylians were quadrupedal amphibians sharing common evolutionary ancestry with the sirenians. Habitats of desmostylians probably were shallow waters supporting productive populations of algae and aquatic vascular plants, their primary food, in latitudes ranging from subtropical to cool temperate. In habits and superficial morphology, desmostylians often are characterized as similar to the modern hippopotamus. Some have argued that at least some of the desmostylians fed on clams and other benthic invertebrate prey, but the consensus is that they were primarily herbivores.

6. Marine Otters

There are 13 recognized extant species of otters worldwide, comprising the mustelid subfamily Lutrinae. Here, we consider 2 species, the sea otter of the North Pacific Rim and the chungongo of Peru, Chile, and southernmost Argentina. Both are marine species with...
amphibious characteristics. Other otter species may utilize marine environments, but they also have obligatory associations with aquatic and terrestrial habitats. The sea otter and chungongo do not appear to utilize freshwater habitats, except occasionally and facultatively.

The sea otter is arguably the most derived of the lutrines. It is the largest of the mustelids, with some adult males reaching 45 kg in mass and 1.6 m in total length, but among the smallest of the marine mammals. Sea otters are relatively weak divers compared to most marine mammals, and they feed almost entirely on large-bodied, sessile or slow-moving benthic invertebrates. Sea otters often haul out on coastal beaches and reefs to rest and conserve heat, especially in the northern portions of their geographic range during periods of harsh weather or reduced sea surface temperature. Sea otters are not known to utilize freshwater habitats for any purpose.

The smallest of the marine mammals, the chungongo, reaches maxima of 6 kg in mass and 1.1 m in length. Chungongos are morphologically similar to the seven congeneric species of otters. The ecological characteristics of chungongos are not well known. They feed primarily on small crustaceans, mollusks, and fish taken during dives in nearshore marine habitats along open coasts. They may also forage in fresh water, taking small crustaceans. They haul out between foraging periods on exposed rocky shores and appear to maintain shoreline dens that are focal areas for social and reproductive behavior.

7. Polar Bears
Polar bears are one of seven recognized bear species. Several of the other bear species utilize marine and aquatic habitats for foraging, but polar bears are more dependent on marine habitats for food than are other bears. Although generally similar to other bears morphologically, polar bears have several distinguishing features that reflect their associations with frigid terrestrial and sea ice habitats and with Arctic marine ecosystems. Polar bears are small compared to many marine mammals but large compared to most other bears. Adult males reach 2.5 m in length and 800 kg in mass. Along with chungongos, polar bears are perhaps the least modified morphologically, compared to terrestrial mammals, of the world’s recognized marine mammal species. Although polar bears do not dive repetitively in the manner typical of many marine mammals, they are efficient swimmers, able to traverse large expanses of open water. Polar bears also cover large distances at sea by walking or running across sea ice. Although primary prey are pinnipeds taken from the surface on sea ice, polar bears may have a diverse diet. They are the only recognized species of marine mammal that travels extensively on land away from the shoreline and the only species that consumes both plant and animal species as a regular part of the diet. In addition, polar bears are unique among marine mammals in producing altricial young.

B. Synopsis of Evolutionary Histories of Major Marine Mammal Taxa

1. Cetacea
The oldest recognized cetaceans are Eocene fossils of the cetacean suborder Archaeoceti. Archaeocete fossils are found primarily in rocks of present-day Egypt, Pakistan, and India in strata thought to be associated with the Tethyan Sea of ancient times. Thus, it is presumed that cetaceans originated in the Old World Tethyan environment. Cetaceans share common ancestry with an extinct terrestrial ungulate taxon known as the mesonychid. The earliest recognizable cetacean fossils date to approximately 55 million years ago (Ma). The archaeocetes included many “missing link” fossils, displaying intermediate forms with regard to progressive reduction and loss of the hindlimbs, elongation of the anterior skull, and modification of dentition.

Archaeocetes were largely extinct at the beginning of the Oligocene, approximately 38 Ma. The first precursors to the modern suborders Odontoceti and Mysticeti appear in the fossil record during the Oligocene, but the first fossils linked unambiguously to modern cetacean families appear primarily during the Miocene. For example, the earliest sperm whales appear in early Miocene strata. Beaked whales appeared first in the middle Miocene, and the earliest dolphins and porpoises appeared in the late Miocene approximately 11 Ma. In the mysticetes, the earliest rorquals and right whales also are in Miocene strata. The oldest gray whale fossils are from the Pleistocene. Thus, there are no known fossils providing insight to the early evolution of the modern gray whales. There is general agreement that the cetaceans are monophyletic.

It is apparent from the fossil record that the modern taxa of cetaceans have been preceded by many extinct species, likely outnumbering extant species by a considerable number. For example, an early mysticete group, the family Cetotheriidae, contains about 60 known species dating from the Oligocene. The existence of extinct large taxa implies significant episodes of diversification and subsequent extinction well before the modern families of cetaceans appeared. The record also suggests a
dynamic pattern of biogeographic variation, such as the occurrence of ancestral monodont fossils in Miocene strata of Mexico. The Monodontidae now include only the belugas and narwhals of high northern latitudes. The dynamic evolutionary record almost entirely predates hominid evolution and the emergence of anthropogenic influences on pattern of extinction.

2. Pinnipeds
The earliest known pinnipeds are represented by fossils of the late Oligocene or early Miocene, approximately 25 Ma. Two major recognized lineages culminate in the modern Phocidae and the modern superfamily Otarioidea, the latter including the Otariidae and the Odobenidae. The earliest fossils of both lineages are of similar late Oligocene or early Miocene age. In contrast to the cetaceans, there are recognized linkages of good quality between the earliest pinniped fossils and the major modern pinniped taxa.

The traditional view of pinniped evolution is that the phocids and otarioids evolved independently. Phocids are said to have emerged in the North Atlantic region from ancestral forms linked to modern mustelids. Appearance of phocid species in the Pacific likely occurred much more recently, possibly during events related to the extinction of early Pacific otariids and odobenids. Otarioids are thought to have evolved from ancestral ursids in the North Pacific. The diphyletic model is supported by traditional analyses of cranial morphology and by the absence of early fossil otarioids from strata of the North Atlantic region. Recent analyses of postcranial material and of molecular data support an alternative model in which the pinnipeds are a monophyletic lineage sharing common ancestry with the modern mustelids. Current consensus favors the monophyletic model.

Recently published evidence indicates controversy regarding the affinities of the Odobenidae. Some analyses indicate that odobenids are in fact more closely related to phocids than otariids, whereas others favor the more traditional view, with odobenids closely allied to otariids in the Otarioidea. We follow the traditional view here but acknowledge the evidence in support of the alternative scenario.

Ancestral taxa of otariids and odobenids show a high level of diversity compared to modern forms. The recognized modern genera of odobenids and otariids appeared primarily during the late Oligocene or the Pleistocene. The fossil record for phocids is not well developed, especially in some North Atlantic regions thought to be important in understanding early evolution in the group. Miocene fossils from the Southern Hemisphere include significant numbers of monachine (phocid subfamily Monachinae) seals (ancestral to the modern monk seals), elephant seals, and Antarctic ice seals. Ancestry and relationships of the more derived forms, including the modern phocine (phocid subfamily Phocinae) seals of northern temperate and polar latitudes, have not been resolved definitively. As with the cetaceans, the pinniped fossil record indicates significant episodes of diversification and extinction before the emergence of the modern forms, clearly predating anthropogenic influences.

3. Sirenians
The earliest fossil sirenians are from early Eocene strata, approximately 55 Ma. Significant radiation into at least three families of sirenians had occurred by the time the earliest cetacean fossils appeared in the Eocene. Thus, sirenians appear to be the oldest order of living marine mammals on Earth. The fossil record does not provide clear evidence of the ancestral groups that gave rise to the sirenians, although studies of modern forms indicate common ancestry with other subungulates. The oldest sirenian fossils are from Jamaica, but the sirenians are thought to have emerged first in the Old World Tethyan environment. The early radiations of the sirenians appear to share common ancestry with the extinct family Protosirenidae. Protosirenids appear to have given rise to the modern families Trichechidae (manatees) and Dugongidae (dugongs and sea cows). Dugongs probably first appeared in Mediterranean waters during the Eocene, whereas trichechids apparently first evolved along the South American coast during the Miocene. Sea cows first appeared in the southeastern Pacific during the early Miocene, later radiating throughout the Pacific basin. Some sea cows were unusual among sirenians in their great body size and use of relatively exposed cold-water habitats. They occurred along the coasts of California, Japan, and the subarctic North Pacific Rim during the Miocene and Pleistocene. Although the predominant sirenian family in the Holocene, trichechids have a relatively poor fossil record. The allopatric distribution of the three modern species apparently resulted from temporary geological isolation of the Amazonian watershed and from a chance colonization of West African coastal waters from an ancestral Caribbean population.

4. Desmostylians
As noted previously, desmostylians apparently first appeared in the Oligocene, approximately 35 Ma. The oldest known fossils are from the coasts of Washington and Oregon in the northeastern Pacific, and all desmo-
5. Marine Otters and the Polar Bear
The sea otter, chungungo, and polar bear represent three separate, relatively recent entries into marine environments by largely terrestrial or aquatic taxa. Chungungos still resemble other otters so closely that a meaningful fossil record of their evolution as marine mammals does not exist to our knowledge. There are some Pleistocene fossils of polar bears, but neither fossil nor modern forms represent significant departures from the ancestral ursid morphology. Thus, for both chungungos and polar bears, the history of adaptation to marine life is best inferred from modern biological data.

The sea otter appears to have a somewhat more extensive fossil ancestry in the marine environment than the polar bear or chungungo. There are two extinct genera of sea otters in the fossil record. Species of *Enhydridodon* have been found in Africa and Europe in late Miocene and Pliocene strata. *Enhydritherium luciae* is known from the late Miocene of Europe, and *Enhydritherium terraenovae* is known from the late Miocene through the middle Pliocene in Florida and California. *Enhydritherium* is thought to be ancestral to *Enhydra*. *Enhydra* is confined to the North Pacific region. The extinct *Enhydra macrodonta* is known only from the late Pleistocene. The single surviving species of sea otter (*Enhydra lutris*), dates from the early Pleistocene.

C. General Factors Contributing to the Vulnerability of Marine Mammals to Overexploitation and Extinction

1. Obligatory Dependence on the Sea Surface for Respiration
Marine mammals must exchange respiratory gases directly with the atmosphere, in the same manner as their terrestrial relatives. Thus, unlike marine fishes and invertebrates, marine mammals at sea are never entirely free of their association with the sea surface. Marine mammals must return periodically to the surface to breathe. The process of breathing at the surface is often associated with conspicuous activities, such as splashing, exhalations audible over long distances, and the production of clearly visible clouds of condensed water vapor associated with exhalations. Marine mammals are often physiologically obligated to remain at the surface for several minutes, allowing multiple exchanges of gas volumes contained in their lungs, in order to set the biochemical stage for successful subsequent dives.

Human travel in watercraft at the sea surface is relatively efficient and advanced. By understanding the respiratory behavior of marine mammals, people can position themselves to facilitate close contact with surfacing animals. The result is high vulnerability of marine mammals at sea to human hunters.

2. Large Body Mass and Linear Dimensions
The marine mammals, on average, are large compared to most other animals. Large body size probably evolved in response to certain constraints associated with life at sea, most notably those associated with thermoregulatory and hydrodynamic efficiency, foraging ecology, reproductive ecology, and physiology, and habitat preference.

The energetic return to the consumer per unit of hunting effort will increase with the mean size of the prey, all other factors being equal. Among the marine mammals, all the mysticetes and many of the odontocetes, pinnipeds, and sirenians are large enough in body mass to be highly desirable as targets by human hunters. With twentieth-century refinements to the technology of marine mammal hunting at sea, pursuit of even the most mobile and dangerous of the marine mammals produced highly desired rates of economic return as long as stocks were not depleted. Thus, large body size, per se, increases the vulnerability of marine mammal populations to extinction simply by improving the economic return on investment in hunting activity.

3. Relatively High Predictability of Spatial and Temporal Distributions in Association with Regions of High Biological Productivity at Sea
With certain exceptions, marine mammals are rarely far removed from locations in which they can forage efficiently. The metabolic generation of heat, fueled by food consumption at high rates, is the only option available to marine mammals at sea for replacing heat lost continuously in a heat-consumptive environment. In species with significant annual migrations to food-poor areas for breeding, high rates of intake during the feeding season are vital for survival of extended travel and for maximizing rates of reproductive fitness. In species whose breeding systems include extensive seasonal lasts, seasonal hyperphagy, requiring proximity to abundant food, may be crucial to reproductive fitness and to long-term survival of both sexes.

Most marine mammals feed on planktonic inverte-
bmates and small schooling fishes. Over the long term, seasonal and spatial patterns of production of zooplankton and forage fish are relatively predictable. Successful tracking of resources that vary predictability in space and time is vital to survival and reproductive success. Accumulated ecological data for marine mammals indicate that most populations successfully track food source productivity most of the time. The result is an array of stereotypical seasonal and spatial movements by marine mammals that, in many cases, are readily identified. Clearly, an understanding of movements of marine mammal stocks over time reduces the investment risk in developing strategies for efficient hunting of marine mammals. Thus, spatial and temporal predictability in marine mammal foraging facilitates efficient hunting by humans and adds to the risk of anthropogenic extinction as a result of direct exploitation of marine mammal populations.

4. Impaired Mobility, Contagious Dispersion, and Temporal and Spatial Predictability When Hauled Out on Land

Pinnipeds, the two marine otters, and polar bears are amphibious marine mammals. Hauling behavior of pinnipeds is particularly synchronous and predictable, producing seasonally dense hauled-out aggregations that can be anticipated readily in space and time. These patterns are most extreme for some of the land-breeding pinnipeds that dwell at middle or subpolar latitudes, but they are prevalent in many other pinniped species as well. Two primary factors contribute to the pattern. First, good hauling sites that are near seasonally predictable foraging locations, and that are free of the disruptive effects of terrestrial predators such as bears or wolves, are typically few in number and small in size. Pinnipeds depending on such hauling sites have evolved high site fidelity and strong navigational capabilities in order to minimize the risk that good hauling sites will be overlooked when needed for breeding, molting, and resting. Second, in the case of pinnipeds at middle and high latitudes, highly productive foraging locations near good hauling sites tend to be strongly seasonal in food availability.

Pinnipeds are awkward on land and can be captured easily by human hunters on hauling sites, even if methods are primitive. Because seasonal hauling patterns are easily recognized in space and time, human hunters can plan highly efficient hunting of hauled pinnipeds with a very low risk of poor return on invested effort. In addition to directed harvest, human activity on preferred hauling grounds can cause significant un-
pelts may be used for clothing and for construction of boats or dwellings. The pelts of sea otters and fur seals have been sought for centuries as commodities of high commercial value for clothing or as adornments for ceremonial robes and artifacts. Teeth and other body parts from sea otters are also known to have had ceremonial significance to indigenous peoples. Specialized tusks in narwhals and walruses are extremely valuable and have put these species at continued risk during much of the past two centuries. Polar bears and walruses have been hunted for meat, oil, hides, and pelts by subsistence cultures for millennia and were targets of trophy hunters through much of the nineteenth and twentieth centuries.

Much of the demand for marine mammal products centers on tissues involved in thermoregulation. The blubber of cetaceans, pinnipeds, and sirenians and the pelts of fur seals, sea otters, and polar bears are the respective primary organs of heat retention, allowing the preservation of homeothermy in a chilling environment. Thus, the homeothermic physiology of marine mammals underlies their desirability as commodities and is a significant contributor to the vulnerability of marine mammals to anthropogenic extinctions.

6. Low Demographic Potential for Rapid Recovery from Disturbance or Overexploitation

The marine mammals are significantly convergent in many aspects of life history. Mean litter size is one for all species except polar bears and chungongos, and multiple simultaneous births are rare in all species of cetaceans, pinnipeds, sirenians, and in sea otters. The age of first reproduction is often relatively high, especially in sirenians and the larger odontocetes. None of the extant marine mammals has a birth interval of less than 1 year, and for many species the birth interval is at least several years. In all species, parental care is entirely maternal, and the energetic costs of lactation and other forms of care are extensive for the adult female. Reproductive success of newly mature females is often very low in marine mammals, increasing only with experience. Survival rates of weaned offspring may be low during the first few years of independence. The combined effects of the previously discussed characteristics are low potential rates of growth in marine mammal populations, even when free of the constraints imposed by food limitation, competition, predation, or natural disturbance. Maximum potential annual rates of population growth typically are 2–8% for the cetaceans and sirenians and 10–15% for the pinnipeds and sea otters. Exceptional cases, both higher and lower than mean rates, are known for both groups. Realized rates of growth, affected by variations in food supply and the effects of disturbance, predation, competition, and possibly other factors, often are much closer to zero and may be constrained to negative values. Given these patterns, recovery from depletion associated with excessive exploitation or disturbance may require many years, and intervening additional harvest or disturbance can increase the risk of extinction to high levels.

7. Morphological, Physiological, and Ecological Predisposition for Bioaccumulation of Lipophilic Anthropogenic Contaminants and Toxins

Most marine mammals utilize a well-developed subcutaneous blubber layer as the primary means of thermoregulation. The blubber layer is also a primary organ for energy storage, as are fat deposits in polar bears. The blubber layer consists of lipids, fatty acids, and connective tissues. Relative proportions vary by species, age, and reproductive status of individuals and by location on the body.

Many species of cetaceans and pinnipeds, the Amazonian manatee, and some polar bears have an extensive seasonal fast each year, relating to migration away from primary feeding areas, an extended haul-out or denning period in association with reproductive activity, or shifts in habitat structure. During such fasts a significant proportion of the blubber layer or fat deposits is metabolically mobilized to meet energy and water demands during the fasting period. Following the fast, animals return to the feeding grounds and forage intensively to reconstitute the reduced blubber layer or fat deposits. In the case of the odontocete cetaceans, the acoustic portion of the blubber is a second concentration of lipid-based tissue. Odontocetes, pinnipeds, the two marine otters, and polar bears occupy high trophic levels in marine food webs. Many stable lipophilic contaminants are transmitted through food webs, such that top-level consumers may be exposed to high levels of contaminants. Such patterns are particularly well-known for environmental contaminants such as organochlorines, a group including the polychlorinated biphenyls (PCBs) and the various derivatives of dichlorodiphenyltrichloroethane (DDT). In this context, many marine mammals face "double jeopardy." First, a high position in their respective food webs confers the risk of high exposure to stable lipophilic contaminants. Second, extensive, met-
abolically active lipid-based tissues are vital to their survival but vulnerable as sites for accumulation of contaminants. In several cases, lipophilic contaminants have been correlated with reduced immune competence, disease outbreaks, and significant mass mortalities in marine mammal populations. There is also evidence that contaminants may cause endocrine disruption and reproductive malfunctions such as premature deliveries of pups. Thus, the combination of lipophilic contaminants in the marine environment, the pattern of foraging at high trophic levels by marine mammals, and lipid-based tissues that are functionally significant and metabolically active in many marine mammals results in increased vulnerability of marine mammals to anthropogenic extinction.

8. Overlap of Diet or Habitat with Commercial or Recreational Fisheries

Marine mammals often feed preferentially on prey species also sought by commercial, recreational, or subsistence fisheries, or they forage in habitats in which significant commercial fishing activity occurs. These patterns create two types of problems that may enhance the vulnerability of marine mammals to anthropogenic extinction. In the first case, marine mammal populations are viewed by commercial, recreational, or subsistence fishers as competitors for a common resource. As a consequence, legal recourse may be sought to actively reduce the range or numbers of marine mammals by killing, translocation, or harassment in order to reduce the intensity of competition in favor of harvesting interests. Illegal activities may also result, including unauthorized killing or harassment also intended to reduce the intensity of competition between marine mammal populations and fisheries. Such circumstances can lead to conflicting management goals by interested parties, particularly if the involved marine mammal populations are small. From the perspective of the involved marine mammal species, competition from harvesting interests may alter the quantity or distribution of food availability and may produce significant consequences at the population level.

The second type of problem is inadvertent or incidental take of marine mammals by entanglement in fishing gear. Such taking may include injury or death of individual animals, again producing possibly significant effects at the population level. Potential responses to such problems include alteration of management authority, changes to fishing techniques or effort to reduce rates of taking, or displacement of fishing effort to other locations. Such interactions may still result in population-level effects if reduction in rates of taking is inadequate, or if parties affected by displacement of fishing effort resort to illegal taking of involved marine mammals as a form of retribution. Thus, both types of problems may contribute to increased risks of extinction, especially in cases in which the involved marine mammal population is small at the time that conflicts are recognized.

D. General Factors Hindering Effective Identification and Monitoring of Marine Mammal Populations Vulnerable to Extinction

1. Availability Bias

Marine mammals at sea spend most of their time below the sea surface. Depending on water clarity, typical depth of dive, angle of observation, and platform of observation (i.e., surface vessel or aircraft), a varying proportion of individual marine mammals in a field of view cannot be seen and thus cannot be enumerated in a population survey. The proportion of animals not visible because of submergence is the availability bias of the survey. Availability bias reduces both accuracy and precision of population estimates, and it reduces the statistical power of a survey effort to detect population trends correctly. Availability bias can be estimated with detailed information on water clarity in the survey area, diving characteristics of target species, and detection characteristics of survey observers. Estimates of bias allow the effects of the bias to be incorporated into calculations of correction factors and coefficients of variation (CVs) for population estimates. Elimination of availability bias generally is not possible for surveys at sea.

Availability bias may also be a problem for surveys directed to hauled animals on shore. For example, topographic irregularities such as overhangs may obscure animals that are present in the defined survey area, and animals at high density may obscure one another. Judicious timing and modified survey angles may sometimes reduce availability bias in surveys of hauled animals to nearly zero.

2. Observer Bias

Observer bias, also known as detection bias, results from the inability of survey observers to correctly enumerate the number of animals visible in a field of view. Unlike availability bias, observer bias can be either a high or low bias. Observer bias has the same implica-
tions for population estimation as summarized previously for availability bias. Observer bias can be a significant source of error in both surveys at sea and surveys of hauled animals on shore. Observer bias can be reduced with increased experience of individual observers, and it can be estimated using double-counting techniques with paired independent observers or by comparing observer counts in the field with counts from aerial photographs taken at the same time and place.

An important form of observer bias involves errors in estimates of group size in marine mammal surveys. Typically, marine mammal surveys involve counting of groups in the survey area. The group count is then multiplied by the mean group size, often based on a separate survey effort, as the first step in population estimation for the survey area. Estimates of group size are subject to observer bias, contributing to an increase in the CV for population estimates. Observer bias in group size estimates can be assessed under good conditions by comparing group size counts by observers with group size counts based on aerial photographs of the same group of animals taken during the surveys.


In many cases, the single most important type of information for assessing the status of a marine mammal population is the trend in population size over time. A trend is simply a time series of counts in which the slope of a fitted line is significantly different from zero. In a statistical context, the ability to detect a trend correctly is influenced by four factors. The first is the strength of the trend. Strong trends are those in which the absolute value of the slope of the fitted line is large. When other factors are constant, strong trends are more likely to be detected correctly than weak trends. The second factor is estimation error. Other factors being equal, trends in survey-based estimates of population size over time are more likely to be detected correctly if the associated CV is small than if it is large. The third factor is the number of replicate surveys available for a given time period, to be used to calculate a single estimate of population size. Increased replication reduces the CV associated with a given estimate. Thus, the probability of correctly detecting a trend increases with the number of replicate surveys used to calculate each point in the population time series, other factors being equal. The fourth factor is the number of years in which surveys are done. The probability of correctly detecting a trend, with other factors being equal, is increased with an increasing number of survey years.

Many marine mammal surveys have characteristics that reduce the probability of correctly detecting trends. Weak trends may portend significant conservation concerns for marine mammals but are inherently difficult to detect if CVs are large and replication is minimal. Large CVs are common in all types of marine mammal population surveys, although CVs are gradually being reduced by the extraordinary efforts of involved researchers and managers. The level of replication generally is directly dependent on levels of funding. Implementation of field surveys in marine mammal science is often compromised by the challenge of limited funding and competing priorities. Well-executed marine mammal population survey programs generally detect strong population trends successfully. However, many surveys lack the statistical power to detect weak trends that may nevertheless be important in the context of avoiding eventual extinction of marine mammal populations. The only solution is to extend survey effort over many years, thus improving the odds of recognizing a trend. Such an approach carries the obvious risk of potentially delaying the recognition of a significant conservation problem for the target population and inevitably increases the overall monetary cost of the monitoring effort. In many cases, there is no alternative to extension of the timescale of the monitoring project. Policies that require decisions about trends over a short time frame are likely to fail in providing appropriate management if trends are weak.

4. Inadequate Understanding of Vital Demographic Parameters and Population Structure

Determination of effective measures to eliminate a negative trend in a marine population often depends on a reasonable understanding of the demographic characteristics of the subject population. Such an understanding improves the odds that limited resources for conservation work will be applied where the greatest benefits will accrue. The demographic parameters of marine mammal populations are often known only with poor levels of accuracy or precision. In such cases, conservation effort can be readily misdirected. For example, measures to reduce preweaning mortality in a marine mammal population will be ineffective if the larger problem is a high rate of adult female mortality obscured by imprecise or inaccurate estimates of adult female survival rate. Misdirection of limited resources for conservation has obvious effects on extinction probabilities for populations in jeopardy.

Accurate, precise measurements of population parameters in marine mammals are difficult to obtain.
Because marine mammals are long-lived species, the time line necessary to obtain good parameter estimates is always lengthy, and the best data come from studies that extend beyond a decade in duration. One of the important results of long-term demographic research on marine mammals is evidence, in some cases, of marked interannual variability in demographic parameters. Such patterns further emphasize the critical need for a long timescale in such studies. For many species, good estimates of demographic parameters require tagging of individuals, an invasive process that can impose risks of reduced survival for the tagged individual and risks of disturbance to groups of animals such as pinniped breeding colonies on haul-outs. In some cases, tagging has the potential to bias the demographic data recorded from the tagged individual. Finally, because of the lengthy duration and labor intensity involved, demographic studies may be quite costly and therefore difficult to implement.

Population structure is poorly understood in many marine mammal species. Species with large geographic ranges and high mobility may appear to be genetically panmictic. However, recent data suggest significant within-species genetic structure in several cases. Examples include several delphinids and baleen whales. Uncertainties about population structure impose a risk of inappropriate scale in the application of management policies. For example, bycatch in fishing nets may occur at high rates only in one portion of the geographic range of a delphinid. If the population is panmictic across the entire range, management authorities may deem the bycatch rate acceptable. Such a policy would, however, increase extinction risk if the area of high bycatch rate supports a population genetically distinct from the remainder of the species’ range.

5. Inadequate Understanding of Effects of Environmental Uncertainty on Dynamics of Populations

There is limited evidence that apparently stochastic environmental fluctuations may have important effects on the dynamics of marine mammal populations. The best known cases involve the pinnipeds, primarily because pinnipeds breed or care for young on solid substrata and therefore have more readily observed and better known population dynamics than most other marine mammal taxa. For example, during the early 1980s and mid-1990s, food supplies for many temperate pinniped populations in the North Pacific were disrupted by oceanographic disturbances generally known as “El Niño Southern Oscillation” (ENSO) events. ENSO events involve a suite of changes in global wind and ocean current patterns, with major large-scale effects on patterns of biological productivity. Although often drastic, local changes in productivity typically are temporary, returning to normal levels over time periods from a few months to a few years. ENSO events have stochastic characteristics regarding both the frequency and the intensity of occurrence and may be variable in their effects on survival rate and population trajectory of marine mammals. Both of the referenced ENSO events caused significant reduction in some pinniped population sizes and in mortality rates in pups of the year. There have been very few opportunities to observe the effects of stochastic disturbance during other time periods or on other marine mammal taxa.

In the context of extinction, the major problem in incorporating naturally occurring stochastic events into conservation planning is the lack of good quality data from an adequate number of events. Thus, it is not possible to generalize about effects of stochastic events, nor is it possible to reasonably consider mitigation measures to minimize increased extinction risks associated with stochastic events.

6. High Cost of Survey Efforts

Despite their obvious conservation value, survey efforts for marine mammal populations often are compromised or eliminated by funding constraints. The central problem is the high cost per unit effort of a good-quality survey for marine mammals at sea. Most surveys at sea use either aircraft or large surface vessels as observation platforms. Both types of platforms are extremely costly to operate at the level of rigor and safety necessary to obtain statistically defendable estimates of target population size. Indeed, the monetary costs of good at-sea surveys may be a significant proportion of the research and management budgets of resource-oriented government agencies, even in wealthy countries such as the United States. In most cases, low-cost alternatives simply do not exist. Thus, only the most serious issues of population trend in marine mammal conservation can reasonably attract a quantitatively rigorous level of survey effort. These funding realities increase the risks of extinction by improving the odds that trends of concern will be overlooked for lack of the necessary survey effort.

7. Lack of Consensus and Consistency in Definition of Objective Criteria for the Determination That Particular Populations Are Vulnerable

There are many forms of institutional or organizational protocols designed to provide protection for depleted
populations of wildlife, including marine mammals. Many such protocols originate from and have the backing of government agencies. However, few protocols for the protection of species in peril include explicit, objective criteria for determining the level of jeopardy or for specifying recovery from jeopardy status. Moreover, those cases in which objective criteria are specified rarely accommodate the quantitative uncertainties in population estimation, demographic parameterization, characteristics of ecological disturbance, and related issues that are universal problems in population data for depleted wildlife. Surveys of criteria for status determination of wildlife in peril reveal little consistency, and often only minimal consideration of fundamental biological patterns, even within particular political jurisdictions.

Resolution of the problem of objective criteria will be difficult. Different taxa of organisms have different population characteristics and may warrant different approaches to the development of objective criteria for jeopardy and recovery. Differences in culture, values, and political traditions among jurisdictions also complicate any effort to establish common objective criteria. However, the separation of jeopardy criteria from arbitrary judgments, often intertwined with political considerations, is crucial to ensure that extinction risk is measured by biological criteria. A failure to achieve such separation may increase the odds of extinction for some populations in peril.

E. Problems in Distinguishing “Natural” from “Unnatural” Extinction

As indicated previously, marine mammals have been present on Earth since the early Eocene. The majority of the marine mammals that have evolved on Earth are now extinct, and virtually all extinctions occurred before the evolution of the homonid primates, particularly Homo sapiens. In the context of conserving the earth’s biodiversity, however, we are primarily concerned about anthropogenic loss of species. Thus, our time window is narrowed to the Pleistocene and Holocene and our conceptual focus to extinctions that result from conscious human action rather than from other, natural processes such as changing habitat, disturbance and catastrophe, or displacement of one species by another. We suggest that not all anthropogenic extinctions are necessarily unnatural. However, separating the natural from the more philosophically objectional unnatural is a great challenge, much burdened by the complicating considerations of values, culture, politics, and economics. Thus, the distinction, however important, is well beyond the scope of this article.

II. PATTERNS AND CASE STUDIES OF EXTINCTION IN MARINE MAMMALS

A. Extinction or Near Extinction of Major Taxa over Evolutionary Time

1. Odobenidae

Odobenids have a fossil record of striking diversity, including some species resembling modern walruses and others superficially similar to the modern sea lions and fur seals. The odobenids seem to have diverged from the otariids in the early Miocene. Peak fossil diversity is in strata from the late Miocene and early Pliocene. Diversity declined abruptly during the late Pliocene and Pleistocene. The single surviving species, the modern walrus, appeared in the fossil record during the Pleistocene. Extinct odobenids were also broadly distributed across latitude. The modern walrus is limited to high northern latitudes and is a distributionally aberrant relative to the extinct odobenids.

The fossil record indicates that modern walruses are the single relict of a largely lost taxon. Most extinctions of odobenids seem to be associated with global-scale cooling and related large-scale habitat change during the Pliocene. However, ecological characteristics of the extinct odobenids are in many cases difficult to understand because the surviving contemporary model seems atypical. Thus, it is difficult to develop meaningful functional models of extinction in the odobenids.

2. Sirenia

Sirenian diversity clearly peaked during the Miocene, a period of warm global climate coincident with extensive warm shallow coastal marine habitats. Dugongids, now represented by a single surviving species, were the most diverse family of sirenians through the fossil record. Sirenian diversity declined sharply during the relatively cool Pleistocene and Pleistocene, and modern forms are relicts of a largely extinct order. Two of four recognized sirenian families are now fully extinct.

The surviving sirenians seem to provide good ecological models for the extinct forms. Modern sirenians typically consume macrophytes in shallow waters and, except for Steller’s sea cow, clearly prefer warm protected waters. Thus, it is likely that Pleistocene cooling and the coincident widespread loss of warm shallow seas were major factors in the decline of the sirenia.
Surviving species were those able to retreat to low-latitude refugia or, in the case of Steller's sea cow, a subpolar refuge with abundant food and no significant predators.

3. Desmostylia
Desmostylians did not diversify to nearly the extent of the other major marine mammal taxa, and they did not survive beyond the end of the Miocene. Lacking modern ecological models, we prefer not to speculate on specific ecological mechanisms of extinction in the desmostylians. However, loss of the taxon coincided in time with the decline of sirenians, with which desmostylians share common ancestry. Thus, habitat needs and associations may have been similar between late Miocene sirenians and desmostylians, and loss of optimal habitats during Pliocene cooling could have had similar effects on both groups.

B. Modern Anthropogenic Extinctions of Species, Subspecies, and Major Populations

1. Species Level
Following the taxonomic format of Rice (1998), three species of modern marine mammal are known to be extinct. In all cases, the probable causes are anthropogenic. In addition, there is evidence that a recent species of pinniped, undescribed and entirely unknown to science, once occurred in the Chagos Archipelago and Seychelles Islands of the tropical southwestern Indian Ocean (Rice, 1998). If such a species occurred, it is now extinct as a result of unknown factors. There is substantial skepticism that such a species ever existed.

a. Caribbean Monk Seal: Monachus tropicalis (Gray, 1850)
The Caribbean monk seal is one of three recent species of Monachus. All occur in tropical or subtropical latitudes. The preexploitation range of M. tropicalis included the islands of the Caribbean Sea, the coastal regions of Venezuela and Caribbean Colombia, southern Florida, the east coast of Mexico south of the Bay of Campeche, and the Caribbean coasts of Belize, Guatemala, Honduras, Nicaragua, Costa Rica, and Panama. Estimates of preexploitation population sizes are not available. Intensive hunting of seals for meat and oil began with the arrival of Columbus in 1492. Seal populations were reported as depleted as early as the seventeenth century, but a few animals survived into the middle of the twentieth century. The last confirmed sighting of a Caribbean monk seal was at Seranilla Bank, west of Jamaica, in 1952. Directed surveys for seals in the later 1950s and 1960s found no living animals.

b. Steller’s Sea Cow: Hydrodamalis gigas (Zimmerman, 1780)
Steller’s sea cow was first observed by a scientist in 1741, at which time it occurred only along the shorelines of the Commander Islands, east of Kamchatka in the Russian Far East. At the time of discovery, sea cows probably numbered no more than a few thousand individuals in total, but population surveys were never done. There are reports that the sea cow had previously occurred along the eastern shore of mainland Kamchatka and in the Near Islands of the Aleutian Archipelago, but the primary evidence is stranded material that could have resulted from the drift of carcasses. Sea cows were subject to intensive hunting, soon after discovery, by crews of sea otter hunters working in the Commander and Aleutian Islands. Harvest of sea cows provided high-quality meat and blubber in great quantity and facilitated extended harvesting expeditions by otter hunters. The last sea cow observation was recorded in the Commander Islands in 1767. The only scientific observations of living Steller’s sea cows were made by G. W. Steller while shipwrecked at Bering Island in the Commander Islands during the winter of 1741–1742.

Three factors probably contributed to the extinction of Steller’s sea cow. First, directed hunting was intensive and unmanaged. Second, at the time living sea cows were observed by Steller, they occurred only on islands lacking aboriginal human populations. The pattern suggests that aboriginal hunters may have previously reduced the range and numbers of sea cows, predisposing them to extinction when hunting intensified. Third, sea cows foraged on nearshore benthic kelps. The rapid depletion of sea otter populations by otter hunters in the Commander Islands may have allowed sea urchins, the primary prey of otters in the region, to overgraze their preferred food, the same kelps utilized by sea cows. Thus, catastrophic loss of food supply could have facilitated the rapid extinction of sea cows after reduction of population size(s) and range.

c. Japanese Sea Lion: Zalophus japonicus (Peters, 1866)
The Japanese sea lion is sometimes considered a subspecies (Z. californianus japonicus) of the California sea lion. Rice (1998) regards the Japanese sea lion as a separate species. The Japanese sea lion originally ranged
along the shores of Japan and Korea and the southern Pacific shores of Russia. The species was subject to a long history of hunting for meat and oil. No population surveys were done. The sea lion was thought to be extinct at the end of the nineteenth century, but a group of animals was reported from a Japanese island in 1952. There have been no subsequent sightings and the species is now considered extinct. All available data indicate that directed harvest was the primary factor contributing to extinction, but other unreported factors could have facilitated loss of the species.

2. Subspecies or Population Level

a. North Atlantic Gray Whale: Eschrichtius robustus (Lilljeborg, 1861)

Whaling records and subfossil remains indicate that a population of gray whales (Eschrichtius robustus), apparently not taxonomically distinct from North Pacific populations, once occurred along both coasts of the North Atlantic. Available evidence suggests that gray whales occurred along the Atlantic coast of North America into the seventeenth century but probably not beyond. We are not aware of any available data on population size, foraging or breeding habitats, or migratory corridors. Subfossil specimens have been found in Europe from the central Baltic coast of Sweden to Cornwall in the United Kingdom. In North America, subfossil finds range from New York to South Carolina. In our opinion, the most likely explanation for extinction is prolonged excessive harvest by the whaling industry, but a definitive explanation for extinction does not exist.

C. Species, Subspecies, and Populations in Imminent Peril of Extinction

1. Synopsis

Here, we summarize the status of 11 marine mammal taxa or populations that in our opinion face a substantial probability of extinction during the twenty-first century (Table III). We provide more detailed summaries for five examples of the category, chosen arbitrarily based on our relative familiarity with the cases. In all but one of the eleven cases, available data suggest total abundances of less than 1000 individuals. Most of the taxa or populations considered in this category have several significant past or current anthropogenic sources of mortality that facilitate extinction risk. In nearly all cases, the scope of operational resource investment and human societal adjustment necessary to avoid extinction seems to us very high.

2. North Pacific and North Atlantic Populations of the Right Whale, Balaena glacialis Müller, 1776

Right whales occur in three major regions: the North Pacific, the North Atlantic, and the Antarctic Ocean, together with adjoining regions of the South Pacific, South Atlantic, and South Indian Oceans. Because of intervening landmasses and the antitropical distribution of the species, rates of migration and genetic exchange among the three regions are very low. The primary conservation problem for the northern populations is very low population sizes, imposed largely by centuries of unregulated commercial whaling.

Right whales were identified by the earliest whalers as targets of choice because of their abundance, large size, high yields of meat and blubber, relatively docile behavior, and relative buoyancy postmortem. Historical records suggest that significant exploitation of the North Atlantic populations began along the European coast early in the second millennium A.D. Changes in hunting effort over time and space followed the stereotypical pattern of overexploitation. As coastal European stocks of whales were depleted, whalers expanded efforts westward to Greenland, Newfoundland, and Labrador beginning in the fifteenth century. Subsequent stock depletion led to further expansion southward in the seventeenth century to the waters of Nova Scotia and the United States. Right whale populations off the northeastern United States were depleted to low numbers after the middle eighteenth century. The historical record of right whale harvest is less lengthy for the North Pacific, but it almost certainly followed the same general pattern. Commercial whaling was most intensive in the nineteenth century. What ever recovery of North Pacific right whales occurred during the twentieth century was damaged by substantial illegal harvest by whalers from the USSR during the 1960s. Approximately 300 individual right whales are thought to be currently present in two populations in the North Atlantic. The eastern population, off the coast of Europe, probably contains only a few individuals and is in extreme jeopardy of extinction. The western population is found mainly along the coast of Canada and the United States. During the twentieth century, estimated annual growth rates of the western population have never exceeded 2.5%, and the population is now declining. The two populations recognized in the North Pacific are also quite small, although estimates of size and trend have not been made. The western population, ranging from the Sea of Okhotsk to mainland China, is thought to be somewhat larger than the eastern population, which ranges from the eastern Be-
### TABLE III

<table>
<thead>
<tr>
<th>Taxon or population</th>
<th>Range/habitat</th>
<th>Population identity structures</th>
<th>Estimated population size</th>
<th>Primary risk factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eastern population of the North Pacific right whale: Balaena glacialis Müller, 1778</td>
<td>Pelagic, eastern margins of North Pacific at middle and high latitudes</td>
<td>One population</td>
<td>No survey data, perhaps a few hundred animals</td>
<td>Commercial whaling, including illegal Soviet whaling, incidental take in fishing nets</td>
</tr>
<tr>
<td>North Atlantic right whale: Balaena glacialis Müller, 1778</td>
<td>Pelagic, eastern and western margins of North Atlantic at middle and high latitudes</td>
<td>Two populations—eastern and western</td>
<td>300</td>
<td>Commercial whaling, ship strikes, incidental take in fishing nets</td>
</tr>
<tr>
<td>Davis Strait, Hudson Bay, Spitsbergen Barrens Sea, and sea of Okhotsk populations of the bowhead whale: Balaena mysticetus Linnaeus, 1758</td>
<td>Coastal and pelagic, in regions as indicated</td>
<td>Four separate populations</td>
<td>Davis Strait and Hudson Bay, about 450 combined; Spitsbergen, less than 30; Okhotsk, a few hundred</td>
<td>Commercial whaling, disturbance from offshore petroleum exploration and development</td>
</tr>
<tr>
<td>Western North Pacific gray whale: Eschrichtius robustus (Lilljeborg, 1861) Russia, during summer; winter range thought to be off Korea or China</td>
<td>Coastal, western Pacific, northern Sakhalin Island</td>
<td>One population</td>
<td>&lt;100</td>
<td>Commercial whaling, offshore petroleum exploration and development in summer range</td>
</tr>
<tr>
<td>Vaquita: Phocoena sinus Norris and McFarland, 1958</td>
<td>Coastal, Northern Gulf of Mexico, California</td>
<td>One population</td>
<td>367 (95% CI: 177–1073) in 1997 survey</td>
<td>Incidental take in fishing nets</td>
</tr>
<tr>
<td>Baiji: Lipotes vexillifer Miller, 1918</td>
<td>Aquatic in lower and middle Yangtze River, China</td>
<td>Population structure unknown</td>
<td>&lt;100</td>
<td>Incidental take in siltine fisheries, directed harvest for oil, contaminants, watercourse impoundments and diversions, habitat loss for prey</td>
</tr>
<tr>
<td>Indus River population of the Indian river dolphin Platanista gangetica (Bosc, 1801)</td>
<td>Aquatic in main channels and tributaries of the Ganges, Indus, Brahmaputra, and Karnaphuli Rivers of India, Bangladesh, Nepal, and Bhutan</td>
<td>Population structure unknown</td>
<td>No data; population is thought to be very small</td>
<td>Incidental take in fishing gear, directed harvest for meat and oil, contaminants, watercourse impoundments and diversions, habitat loss for prey</td>
</tr>
<tr>
<td>Gulf of Alaska population of beluga whale: Delphinapterus leucas (Pallas, 1776)</td>
<td>Coastal marine habitats and river mouths, primarily in Cook Inlet, Alaska, also along the coast of the central Gulf of Alaska</td>
<td>One population</td>
<td>Approximately 350</td>
<td>Hunting by indigenous peoples, commercial whaling, contaminants, habitat disturbance</td>
</tr>
<tr>
<td>Gulf of St. Lawrence population of beluga whale: Delphinapterus leucas (Pallas, 1776)</td>
<td>Coastal marine and coastal, St. Lawrence River estuary and Gulf of St. Lawrence, Canada</td>
<td>One population</td>
<td>1238 (standard error = 119) in 1997</td>
<td>Commercial whaling, culling to protect fisheries, contaminants</td>
</tr>
<tr>
<td>Mediterranean monk seal Monachus monachus (Hermin, 1779)</td>
<td>Coastal in Mediterranean Sea and eastern warm-temperate North Atlantic; currently limited primarily to coastal areas of Turkey, Greece, western Sahara, and Mauritania</td>
<td>At least two populations likely: Cape Blanco in Mauritania, western Sahara, and eastern Mediterranean</td>
<td>275–490</td>
<td>Directed harvest for meat and oil, illegal hunting, habitat destruction, overfishing of prey, contaminants</td>
</tr>
<tr>
<td>Lake Saimaa ringed seal Pusa hispida saimensis (Nordquist, 1890)</td>
<td>Aquatic in Lake Saimaa, Finland; river connection to the Gulf of Finland is thought to be too swift to allow passage of seals</td>
<td>One population</td>
<td>Approximately 200</td>
<td>Incidental take in fishing nets</td>
</tr>
</tbody>
</table>
Failure of northern right whale populations to recover from the cessation of commercial whaling has been difficult to understand, excepting the obvious damage from recent illegal whaling in the eastern North Pacific. Southern right whale populations protected from exploitation have grown at rates estimated as high as 6 or 7% per year, and there is no evidence of major differences in fundamental vital rates between northern and southern populations. Thus, other factors must be retarding population growth in the north. The current consensus view is that two factors, incidental entanglement in fishing gear and inadvertent collisions with large commercial ships, are the primary causes for failure of North Atlantic populations to grow in recent decades. These right whales tend to be concentrated in regions that support productive and highly capitalized fisheries, facilitating damaging rates of incidental entanglement. Right whales are also concentrated in areas frequently transited by large ships. Vulnerability to ship strikes is enhanced by the apparent tendency of right whales to rest quietly at the surface for long periods. Given the long period of depletion, factors in addition to bycatch and ship strikes may have contributed to failed recovery over the long term. For example, social dysfunction resulting from inability to find potential mates (the Allee effect) may diminish the growth potential of a small population.

Survival of northern right whale populations beyond the twenty-first century requires a major reduction of all forms of directed harvest and the largest possible reductions in rates of incidental taking and ship strikes. Eastern populations in the North Atlantic and North Pacific may be destined for extinction within the century regardless of actions taken.

3. Western North Pacific Gray Whale, *Eschrichtius robustus* (Lilljeborg, 1861)

The western North Pacific population of gray whales summers in the Sea of Okhotsk and probably migrates south to winter habitats in southern China. Very little is known about the ecology of the population, and focused studies have been done only within the past 5 years. The western gray whale population is among the smallest of large whale populations in the world and is one of the most vulnerable to extinction in the twenty-first century. The principal conservation concerns for western gray whales are the very small population size, resulting from both commercial whaling and subsistence harvests in previous centuries and risks posed by offshore oil development in the current summer feeding range of the population. Most of the available information on the western gray whale population comes from recent studies summarized by Weller et al. (1999). Western gray whales forage on benthic invertebrates during summer months in shallow coastal waters off northeastern Sakhalin Island, Russia. Foraging activity is particularly intensive in a small nearshore area off the entrance channel to Piltun Lagoon (52°50’N, 143°20’E). Data from photo-identification studies indicate high fidelity of foraging gray whales to the area north and south of the lagoon, within and among years, during the period 1995 to 1999.

Analyses of photographic data indicate that as few as 90 individual whales regularly use the summer foraging area of northeastern Sakhalin Island. Based on current data, the population numbers near 100 individuals. Data from the recent studies indicate that the crude birth rate in the Sakhalin summering aggregation was 4.3% in 1997 and 13.2% in 1998. Adult females with calves show particularly high fidelity to the feeding area near Piltun Lagoon, suggesting that the area is important to calf rearing and weaning. The summer feeding area near Piltun Lagoon has been subject to intensive exploration in recent years and production of oil began in 1999 at offshore petroleum production facilities. Offshore petroleum exploration involves frequent use of intensive low-frequency sounds, and development of located petroleum resources involves increased shipping activity, placement of structures, modification of adjacent sediments and benthic communities, and the risk of unintended spills of drilling fluids, vessel and machinery fuels, and extracted crude oil.

The western North Pacific population of gray whales is in imminent peril of extinction. Recovery planning for the population will benefit from acquisition of significant additional data. The winter range, calving grounds, and migratory corridors of the population must be determined, and the associated risk factors must be identified and analyzed for significance with regard to extinction. The affinity of summering animals, and particularly females with dependent calves, for the feeding grounds off Piltun Lagoon must be better understood. Activities associated with development of petroleum resources in the coastal zone of northeastern Sakhalin Island must be carefully evaluated in the context of risks to gray whales using the area for feeding. All forms of harvest of western gray whales must be prevented indefinitely if extinction is to be avoided.
Marine Mammals, Extinctions Of — 35

4. Gulf of Alaska Beluga Whale, Delphinapterus leucas (Pallas, 1776)

The Gulf of Alaska population of beluga whales has always concentrated primarily in Cook Inlet on the southern mainland coast of Alaska. Animals have been seen on occasion in coastal waters of the Kodiak Archipelago, in Prince William Sound and Yakutat Bay, and along the outer coastal waters of the central Gulf of Alaska. However, it is not known if these animals are part of the Cook Inlet population. Historically, belugas have been widely distributed in Cook Inlet, a well-mixed, turbid, highly productive marine inlet that receives runoff from several large river systems. The whales concentrate seasonally near and in the river mouths to forage on migrating anadromous fish.

Belugas in Cook Inlet were hunted at low levels in previous decades by commercial whalers, although commercial harvest no longer occurs. Passage and implementation of MMPA in 1972 ensured the rights of native peoples in Alaska to pursue the traditional practice of hunting beluga whales, including those in Cook Inlet. Hunting of belugas in Cook Inlet by natives increased in the 1980s and 1990s compared to earlier decades. During the 1980s, it was recognized that the geographic range of beluga whales, based on opportunistic observations from research vessels, was much smaller than it had been in the 1970s. Few sightings of belugas in other locations of the Gulf of Alaska have been made in recent years. Intensive surveys in the 1990s confirmed that the population was declining. The population estimate for 1998, determined by NMFS, was 347 (CV = 0.29). Based on known whale kills by native hunters and evaluation of survey data, it was determined that native harvest of belugas may be excessive, possibly placing the Cook Inlet beluga population at risk of extinction. As a result, a temporary moratorium on native harvest of belugas in Cook Inlet was implemented in 1999. Native tribal leaders and U.S. federal agencies are now planning for recovery actions and a more carefully managed harvest.

The population of traditional control over the harvest.

There is concern about risks of contaminants and marine oil activity for belugas in Cook Inlet. However, despite substantial offshore oil production activity, there is no firm evidence of negative effects of oil production on belugas in Cook Inlet, nor is there substantive evidence of negative effects of chemical contaminants. We suggest that the primary conservation concern for Cook Inlet belugas at the present time is the risk of excessive harvest by native hunters disconnected from the cultural regulation of their respective home villages. In our opinion, harvests of belugas in Cook Inlet must remain suspended until population recovery is apparent and a pragmatic harvest comanagement plan supported by tribal and agency authorities is in place.

5. Vaquita, Phocoena sinus Norris and McFarland, 1938

The vaquita is a small porpoise limited to the northern part of the Gulf of California, Mexico. The vaquita has the smallest natural geographic range of any marine cetacean, and the single population probably has never contained a large number of individuals. Numbers were estimated at about 600 individuals based on a comprehensive survey in 1997 (Table III). The population biology of the vaquita is poorly known, and the very best surveys of the vaquita population have low precision. Thus, trends in the population are difficult to discern with confidence, but available data indicate that the population may be declining. Studies of DNA in sampled vaquitas indicate the complete absence of polymorphisms in the hypervariable region of the mitochondrial genome.

During the last two decades, the primary concerns for vaquita conservation have been high rates of incidental take in gillnet and shrimp trawl fisheries, effects of contaminants, effects of reduced genetic diversity,
and effects of diversion of the Colorado River away from the northern Gulf of California on regional rates of biological productivity. The significance of the listed risk factors has been evaluated by Rojas-Bracho and Taylor (1999). Losses due to incidental take were identified as the primary conservation problem for the vaquita. Concentrations of organochlorines, including PCBs and DDT congeners, were found to be low in vaquita tissues, low in other consumer species within the vaquita range, and low in the vaquita habitat. Organochlorine levels in the region are generally below minimum levels considered harmful to human health and are presumed to be innocuous for vaquitas we well. The levels of reduced genetic variability in vaquitas do not necessarily result in genetically based reduction of reproductive rates, particularly in the context of populations such as vaquitas that probably have always been small. Biological productivity in the upper Gulf of California is high compared to other coastal marine ecosystems despite diversion of the Colorado River. Although other risk factors may be detrimental to vaquita conservation over the long term, mortalities due to incidental take are the primary current problem.

Continued incidental take of vaquitas will cause extinction of the species during the twenty-first century, given even the most conservative estimates of the current rate of take. Elimination of the risk of anthropogenic extinction requires significant reduction in the level of fishing effort or changes in gear design or deployment strategy to reduce take rates in fisheries responsible for incidental take.


The Mediterranean monk seal was found originally in the western Black Sea, throughout the Mediterranean Sea, and along the coast of northwestern Africa from the Strait of Gibraltar to about 21°N latitude. Currently, there are thought to be no more than 275–460 individuals, occurring primarily in two populations. Prior to 1997, the largest group of approximately 300 seals occurred in a small population at Cabo Blanco, at the border of the western Sahara and Mauritania, on the outer coast of northwestern Africa. A second population of unknown size occurs in the eastern Mediterranean, primarily in the coastal waters of Turkey and Greece. Mediterranean monk seals probably have been subject to directed subsistence harvest for meat, oil, and hides for several millennia. The precarious status of modern populations seems to result from many factors associated with the large, multicultural human populations of southern and eastern Europe and northern Africa. For years, monk seals have been perceived as direct competitors of fisheries and have been harassed and killed in substantial numbers, often illegally, as a result. Harassment has included directed destruction of caves and other shoreline locations favored by seals for breeding and resting. Monk seals likely have also been affected by loss of prey due to overfishing and to various forms of contamination of the habitat and food webs. In 1997, a mass mortality event was observed in the colony at Cabo Blanco, reducing the local seal population to about 100 individuals. The cause and magnitude of the event have not been determined to our knowledge. Although comprehensive demographic and population survey data are lacking, the consensus opinion is that the total number of Mediterranean monk seal is probably declining over time.

In addition to the small size of the two known populations, two factors add great difficulty to the prospects for implementation of a successful recovery strategy for Mediterranean monk seals. First, the habitat of the monk seal is bounded by many culturally disparate political jurisdictions. Historically, the political and cultural diversity has interfered with cooperation across jurisdictions. Thus, the attainment of consistent, broadly supported conservation priorities for monk seals may be an unrealistic political objective. Second, ongoing damage to the monk seal populations appears to result from many factors acting in concert rather than one clearly predominant problem. Thus, agreement on conservation priorities and actions may be difficult even within jurisdictions. Mediterranean monk seals appear to be destined for extinction, possibly within the twenty-first century, unless marine conservation authorities in countries bordering seal habitat can agree on two issues. First, risk factors for the seals must be evaluated dispassionately and placed in order of significance. Second, involved authorities must agree on a plan for recovery of seal populations based on the assessment of risk factors and convince the human populations of their respective jurisdictions that seal conservation is a worthwhile objective.

D. Species, Subspecies, and Populations of Significant Concern with Regard to Extinction

1. Synopsis

Here, we consider a group of taxa and populations that, we believe, are of significant concern with regard to extinction (Table IV). In contrast to the discussion in...
### TABLE IV

<table>
<thead>
<tr>
<th>Taxon or population</th>
<th>Range/Habitat</th>
<th>Population identity/structure</th>
<th>Estimated population size</th>
<th>Primary risk factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue whale: <em>Balaenoptera musculus</em> (Linnaeus, 1758)</td>
<td>Pelagic</td>
<td>Three subspecies—two in Southern Hemisphere, one in Northern Hemisphere; five populations in North Pacific, two in North Atlantic; one in north-west Indian; number in Southern Hemisphere unclear</td>
<td>2000 in eastern North Pacific, 400–2000 in eastern North Atlantic; others unknown</td>
<td>Commercial whaling, including illegal Soviet whaling, ship strikes</td>
</tr>
<tr>
<td>Hawaiian monk seal: <em>Monachus schauinslandi</em> (Matschie, 1905)</td>
<td>Coastal leeward Hawaiian Archipelago sites on different islands; some exchange of individuals among islands</td>
<td>Approximately 1300</td>
<td></td>
<td>Directed harvest for meat, oil, and skins; disturbance by military conflict and peace-time human activity; entanglement in marine debris; declining ecosystem productivity; aberrant breeding behavior associated with anomalous sex ratio; natural toxins in prey</td>
</tr>
<tr>
<td>Baltic Sea ringed seal: <em>Pusa hispida botnica</em> (Gmelin, 1788)</td>
<td>Coastal, in areas with pack or shorefast ice at least part of the year</td>
<td>Probably one population</td>
<td>Unknown</td>
<td>Pollution and contaminants</td>
</tr>
<tr>
<td>Lake Ladoga ringed seal: <em>Pusa hispida ladogensis</em> (Nordquist, 1899)</td>
<td>Coastal, in areas with pack or shorefast ice at least part of the year</td>
<td>Probably one population</td>
<td>Unknown</td>
<td>Pollution and contaminants</td>
</tr>
<tr>
<td>Western North Pacific harbor seal: <em>Phoca vitulina stejnegeri</em> Allen, 1902</td>
<td>Coastal along the shores of the western Aleutian and Commander Islands, southeastern Kamchatka Peninsula, Kuril Islands, Sea of Okhotsk, and Hokkaido</td>
<td>Population structure unknown; island breeding colonies probably somewhat isolated; although exchange among colonies may occur</td>
<td>Unknown</td>
<td>Incidental take in fishing nets</td>
</tr>
<tr>
<td>Western North Pacific Steller's sea lion: <em>Eumetopias jubatus</em> (Schreber, 1776)</td>
<td>Coastal and pelagic, islands and isolated rocky shores of the western Gulf of Alaska, western North Pacific, and Bering Sea</td>
<td>Thought to be a single population; some isolation of breeding colonies probably occurs; exchange among breeding colonies is thought to be minimal</td>
<td>Approximately 40,000 individuals; decline in numbers of approximately 6% since the mid-1970s</td>
<td>Unknown; primary possibilities are large-scale declines in ocean productivity, changes in the species composition and diversity of prey species, and competition for prey with commercial fisheries</td>
</tr>
<tr>
<td>Australian sea lion: <em>Neophoca cinerea</em> (Pieron, 1816)</td>
<td>Coastal and continental shelf waters of southern and southwestern Australia</td>
<td>Population structure unknown; long-distance movements are uncommon, so separate breeding colonies may be isolated</td>
<td>Approximately 3000 individuals</td>
<td>Commercial sealing, incidental take in fishing nets, illegal directed killing, entanglement in anthropogenic debris</td>
</tr>
<tr>
<td>Taxon or population</td>
<td>Range/habitat</td>
<td>Population identity/structure</td>
<td>Estimated population size</td>
<td>Primary risk factors</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------</td>
<td>-------------------------------</td>
<td>---------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Hooker’s sea lion: <em>Phocaena hookeri</em> (Gray, 1844)</td>
<td>Coastal waters of New Zealand</td>
<td>Probably one population, breeding primarily at the Auckland Islands</td>
<td>10,000–12,000 individuals</td>
<td>Commercial sealing, incidental take in fishing nets</td>
</tr>
<tr>
<td>Guadalupe fur seal: <em>Arctocephalus townsendi</em> (Merriam, 1897)</td>
<td>Coastal and continental margin habitats off Baja California, Mexico, and southern California</td>
<td>Probably one population, breeding only at Guadalupe Island, Mexico</td>
<td>3000–4000 individuals</td>
<td>Commercial sealing</td>
</tr>
<tr>
<td>Juan Fernandez fur seal: <em>Arctocephalus philippii</em> (Peters, 1866)</td>
<td>Coastal and pelagic habitats of the Juan Fernandez Archipelago, Chile</td>
<td>Population structure unknown</td>
<td>7000–10,000 individuals</td>
<td>Commercial sealing</td>
</tr>
<tr>
<td>Atlantic walrus: <em>Odobenus rosmarus rosmarus</em> (Linnaeus, 1758)</td>
<td>Coast with extensive pack ice in the Arctic North Atlantic, from eastern Canada to the Kara Sea</td>
<td>Four separate populations are recognized</td>
<td>Unknown</td>
<td>Commercial hunting, illegal hunting</td>
</tr>
<tr>
<td>Laptev Sea walrus: <em>Odobenus rosmarus laptio</em> (Chapman, 1940)</td>
<td>Coastal habitats of the eastern Kara, Laptev, and western Eastern Siberian Seas</td>
<td>Probably one population</td>
<td>Unknown</td>
<td>Commercial hunting</td>
</tr>
<tr>
<td>Amazonian manatee: <em>Trichechus inunguis</em> (Natterer, 1883)</td>
<td>Aquatic habitats of the Amazon watershed, South America</td>
<td>Population structure unknown</td>
<td>Unknown; thought to be declining</td>
<td>Commercial hunting, hunting by indigenous peoples, watercourse impoundment and diversion, sedimentation associated with forest modification, contaminants associated with mining, incidental take in fishing nets</td>
</tr>
<tr>
<td>West African manatee: <em>Trichechus senegalensis</em> (Link, 1795)</td>
<td>Warm shallow coastal marine habitats and rivers of West Africa from Senegal to Angola</td>
<td>Population structure unknown</td>
<td>Unknown</td>
<td>Commercial hunting, hunting by indigenous peoples, killing to reduce damage to fishing gear and rice crops</td>
</tr>
<tr>
<td>West Indian manatee: <em>Trichechus manatus</em> (Linnaeus, 1758)</td>
<td>Warm shallow coastal marine habitats and rivers of the Georgia Bight, Florida, the Caribbean, and northeastern South America; southward to central Brazil</td>
<td>Two subspecies—one in the southeastern United States, especially in Florida, and the other along the mainland east coast of central America from northern Mexico to central Brazil, and in the islands of the Caribbean Sea</td>
<td>Approximately 2000–3000 in Florida; trend in Florida populations unclear; numbers elsewhere unknown</td>
<td>Habitat destruction, modification, and disturbance associated with growing human populations in Florida; commercial hunting; hunting by indigenous peoples, contamination; incidental take in fishing nets; ingestion of debris</td>
</tr>
<tr>
<td>California sea otter: <em>Enhydra lutris nereis</em> (Mer- riam, 1904)</td>
<td>Shallow marine habitats along exposed outer coast of California</td>
<td>One primary mainland population, plus small separate colony at San Nicolas Island</td>
<td>2000–2300 individuals; numbers declining at 1 or 2% per year</td>
<td>Commercial hunting, incidental take in fishing nets, contamination, disease, parasites</td>
</tr>
</tbody>
</table>
Section II.C, we do not regard entries in this group to be in imminent peril of extinction during the twenty-first century. In most cases, population sizes are large enough and conservation issues tractable enough that less dire predictions seem reasonable. However, we suggest that vigilance and positive action will be required to prevent taxa and populations in this group from falling to a more precarious status. We list 17 taxa or populations (Table IV), providing more detailed summaries for four arbitrarily selected examples.

2. Blue Whale, *Balaenoptera musculus* (Linnaeus, 1758)

Blue whales occur in all the world’s coastal and pelagic marine habitats. Currently, three subspecies are known. The pygmy blue whale (*Balaenoptera musculus brevicauda*) occurs in southern cool-temperate and subpolar latitudes. The “true” southern blue whale (*B. m. intermedia*) summers in the Antarctic Ocean, and the northern blue whale (*B. m. musculus*) is found in the North Pacific and North Atlantic. Eastern and western populations are known in the North Atlantic, and at least five populations have been described in the North Pacific. The population structure in the Southern Hemisphere is unclear. However, “true” blue whales must have different populations in the southern Indian, Atlantic, and Pacific oceans, respectively. Available data suggest that blue whales migrate seasonally, utilizing higher latitude habitats in summer for feeding, primarily on euphausiid crustaceans, and lower latitude habitats during winter for courtship, breeding, and parrunirritation.

Blue whales are an obvious target of choice for commercial hunting because of their great body size. However, blue whales are swift swimmers and are negatively buoyant postmortem. Prior to exploitation, blue whales were most abundant in the Antarctic Ocean and other marine habitats distant from human population centers. Thus, they were beyond the technological capabilities of commercial whalers prior to the twentieth century. Blue whales became priority targets of whalers only after the development of the steam engine, factory ships, explosive harpoons, and air compressors to inflate carcasses after killing. Thus, the harvest and depletion of blue whales occurred primarily during the twentieth century. At least 360,000 blue whales were killed in the Antarctic region before commercial whaling declined in the 1960s, and other populations were exploited as well. Illegal harvests by Soviet whalers occurred after a moratorium was imposed, taking at least 8000 additional pygmy blue whales.

The current consensus opinion is that blue whale populations are now but a fraction of pre-exploitation size, but there are few data available to defend the perception. Recent data indicate that the eastern North Pacific population numbers approximately 2000 individuals and may be increasing. The eastern North Atlantic population has been estimated at 400–2000 individuals. A circumpolar survey of the Antarctic Ocean estimated 710 (CV = 0.64) individuals. There are published arguments that blue whale populations off southern Japan and in the eastern Gulf of Alaska are locally extinct or very small. Because of small population sizes and large CVs associated with surveys, it is not possible to identify trends in most blue whale populations. Thus, the status of the world’s blue whale populations is generally unknown, and prospects for confidently understanding the size of populations in the foreseeable future are virtually nil.

The current status of blue whale populations seems to be entirely the result of excessive past commercial harvests. There is evidence that ship strikes may cause some mortality of blue whales off the California coast.
but current anthropogenic mortalities are probably minimal. Recovery of small blue whale populations will require indefinite suspension of all forms of harvest and prompt detection and elimination of emerging sources of anthropogenic mortality.

3. Western North Pacific Steller’s Sea Lion, *Eumetopias jubatus* (Schreber, 1776)

Steller’s sea lions occur in coastal waters of the North Pacific Rim from southern California to northern Japan and in the Bering Sea and Sea of Okhotsk. Recent genetic data are the basis for dividing the species into two populations, western and eastern, with the boundary at Cape Suckling, Alaska (144°W longitude). The eastern population is dispersed along the west coast of North America, numbers more than 20,000 individuals, and is increasing gradually, particularly in southeastern Alaska. The western population occupies the Bering Sea, Aleutian and Commander Islands, and remote locations of the Russian Far East. The western population numbered approximately 130,000 animals in the 1930s but has since declined precipitously, with current numbers estimated at approximately 39,500. The rate of decline has varied over time, with highest rates (approximately 15% per year) from 1985 until 1990. The decline currently continues. The cause or causes of decline in the western population of Steller’s sea lions are not understood. Possible risk factors include incidental take in fishing gear, competition with fisheries for prey in common, hunting by indigenous peoples, illegal hunting or harassment, inadvertent rookery disturbance, disease or parasitism, predation by killer whales, contaminants, and changes in the structure and productivity in the marine ecosystems of which Steller’s sea lions are a part. Based on extensive research since the decline was first recognized, the current consensus opinion is that ecosystem change or competition with fisheries are the most likely factors driving the decline. Resolution of the question of cause has become the focus of intensive political debate. Several viability analyses have been applied to Steller’s sea lion population data. Model results lead to extinction of the western population in all cases. Median estimates of time to extinction range from 62 to 160 years. The combination of ongoing decline, projected extinction risks, and uncertainty about primary risk factors leads to significant concern about the persistence of the western population. Because natural oceanographic changes can neither be predicted nor controlled, management authorities have no choice but to focus on understanding and minimizing anthropogenic risk factors, despite the political consequences, in order to reduce the probability of eventual extinction.


West Indian manatees occur in coastal habitats and the lower reaches of rivers in the southeastern United States, the islands of the Caribbean Sea, and the mainland shores of the Gulf of Mexico, Central America, and northeastern South America. Two subspecies are recognized: The Florida manatee (*T. m. latirostris* Harlan, 1824) is found in U.S. coastal waters, especially in Florida. The Antillean manatee (*T. m. manatus* Linnaeus, 1758) occupies the remainder of the range of the species. The current population in Florida is the largest of the species, numbering 2000–3000. Despite a broad perception of increasing numbers in Florida, population data lack the statistical power to document a positive trend with confidence. Population structure, numbers, and trends in other locations are not well-known, but most populations probably number no more than a few hundred animals. Manatees have been hunted by indigenous peoples for meat and other products for centuries. Commercial hunting probably contributed to a reduction of population sizes, but apparently there are few data available to assess the rates or significance of commercial harvest.
Manatees have been protected from all forms of directed harvest in Florida since the 1960s, but subsistence harvest for meat and oil and for ceremonial purposes continues in other populations. During the twentieth century many risk factors for manatees emerged, all in association with an expanding human population in immediate proximity to manatee habitat. The primary problems are a variety of modifications and ongoing disturbances of habitat. The latter include watercourse diversions and impoundments in aquatic manatee habitats and disturbance and collision risk associated with recreational boating in aquatic and marine manatee habitats. There is also continuing concern about increased levels of contaminants.

Long-term studies of stranded carcasses indicate three major sources of mortality in Florida manatees. Perinatal mortalities are newborn animals with the proximate cause of death uncertain but possibly linked to contaminants. Significant mortality rates are also associated with entrapment in dam floodgates and collision with recreational powerboats. Manatees clearly prefer areas without significant powerboat traffic. The ongoing expansion of human populations and associated demand for recreational opportunities inevitably leads to continuing reduction in the size of available manatee habitat.

Manatees are sensitive to water temperature and typically congregate in warm-water refugia during winter, especially when ambient sea surface temperatures drop below approximately 18°C. Refugia include natural warm aquatic springs and lagoons associated with thermal effluent from electrical generating plants. The limited size and number of natural refugia may be threatened by the encroaching effects of human development and activity. Use of power plant lagoons may be risky if plant operations dictate precipitous shutdown, resulting in cutoff of thermal effluent and rapid chilling of lagoons.

Manatees in Florida have experienced several significant mass mortalities in recent years. The mortality events apparently result from an interaction of compromised immune systems, disease, and natural toxins associated with phytoplankton blooms.

Habitat loss and disturbance is the primary conservation problem for all populations of West Indian manatee, and subsistence or illegal harvest remain significant risk factors for populations outside of Florida. Over the long term, avoidance of extinction will require cessation of all forms of human harvest and an effective, broadly supported strategy for balancing the habitat needs of manatees with the consequences of human population growth, especially in Florida.

5. California Sea Otter, *Enhydra lutris nereis* (Merriam, 1904)

Sea otters originally ranged throughout the coastal waters of California, including San Francisco Bay and the southern California islands. The population was originally contiguous with other otter populations ranging from the central outer coast of Baja California, Mexico, through the North Pacific Rim to northern Japan. Sea otters in California comprise one of three subspecies currently recognized.

Sea otters have probably been hunted by indigenous peoples of the North Pacific Rim for several millennia for meat and pelts and for ceremonial purposes. Observations of the Bering Expedition of 1741 and 1742 and other voyages of exploration in the North Pacific found abundant sea otters throughout their range. Commercial harvest of sea otters for pelts began with the Bering Expedition. In California the hunt was pursued by Russians, often utilizing enslaved Aleut hunters, and by hunters from Spain, Mexico, and the United States. All commercial hunts for sea otters were terminated in 1911 with approval of the Treaty for the Preservation and Protection of Fur Seals (37 Stat. 1592, T.S. No. 542) by Japan, Russia, Great Britain, and the United States. The treaty included a passage affording protection to sea otters. However, in California and Mexico the sea otter populations were depleted commercially by the 1860s. By 1900, only two small populations survived. One was along the Big Sur coast south of the Monterey Peninsula, numbering approximately 50–100 animals. The second, of unknown size off the San Benito Islands, Mexico, was extinct by the 1920s. The Big Sur population grew at a rate of approximately 3% per year through much of the twentieth century and now numbers approximately 2000–2500 animals, ranging from San Francisco southward to Point Conception. The observed rate of growth in California has been much lower than rates commonly observed in more northerly populations with protection from harvest and available adjacent vacant habitat. In the late 1980s a new colony was established by translocation at San Nicolas Island off southern California. The colony now numbers approximately 20 animals and persistence remains uncertain.

Several risk factors are known for sea otters in California. Nearshore net fisheries are responsible for significant rates of incidental take. Changes in fishery regulations have reduced but not eliminated incidental take. Sea otters are known to compete with nearshore marine shellfish fisheries in California, particularly for abalones, sea urchins, clams, and crabs. Concern about
interactions of sea otters and shellfisheries is sufficient that it was until recently the policy of the U.S. federal government to actively remove and relocate individual otters found along the California coast south of Point Conception in order to minimize damage to shellfisheries. The management effort was done despite the small size of the sea otter population, its precarious status, and various state and federal legal protections including a listing as "Threatened" under the auspices of the U.S. Endangered Species Act of 1973 (ESA) [16 U.S. Code §§1531-43 (Supp. IV 1974)] as amended. Relocation efforts were inefficient and costly, and alternative management protocols are now in active development. There is also concern about illegal killing of sea otters to protect shellfisheries, but little direct evidence of a significant problem. The California sea otter population is also regarded to be at high risk of potential damage from oil spills, although oil spills have yet to cause sea otter mortality in California. Oiling mats the pelage and eliminates thermoregulatory function, rapidly leading to chilling and death. Oiling is also known to cause other respiratory and physiological pathologies.

During the 1990s, the growth of the population slowed, and currently the population is declining at 1 or 2% per year. A clear consensus on the cause of the decline has not emerged, but primary risk factors are thought to be ongoing incidental taking in fishing nets, effects of contaminants, and the emergence of mortalities from diseases and exotic parasites not previously associated with sea otters. The long-term survival of the sea otter population in California will require ongoing research to more effectively characterize current risk factors and the development of strategies to minimize associated mortalities. The large and growing human population in California is a major underlying cause of the jeopardy status of sea otters, and broad popular support, including economic compromise by shellfishery and marine oil interests, likely will be required if recovery is to be successful.

E. Species, Subspecies, and Populations Once Thought to Be Near Extinction But Now Showing Evidence of Recovery

1. Synopsis

Here, we summarize data for eight species, subspecies, or populations of marine mammals that have been near extinction in the recent past but are now either recovered or en route to recovery (Table V). We regard the taxa or populations in this group as unlikely to become extinct in the foreseeable future as long as risk factors currently unknown do not appear. We present case summaries for five taxa, selected arbitrarily, to illustrate patterns typical of the group.

2. Western Arctic Population of the Bowhead Whale, Balaena mysticetus Linnaeus, 1758

Bowhead whales occur in at least five populations in northern habitats characterized by frequent sea ice. Four of the five populations are small and at risk of extinction (Table III). The western Arctic population is by far the largest of the bowhead populations, numbering approximately 8000 individuals with annual growth rates averaging 2%. The population ranges in the Bering, Chukchi, and Beaufort Seas, migrating northeasterly to the Beaufort Sea in spring and returning to the Bering and Chukchi Seas in the fall.

The western Arctic population of bowhead whales probably numbered approximately 23,000 individuals when first exploited by U.S. whalers during the late 1840s. By the demise of the commercial harvest in 1919, approximately 20,000 whales had been taken, and the population was thought to include approximately 1000 individuals.

Bowhead whales have been hunted for subsistence and ceremonial purposes by the indigenous peoples of Alaska, western Canada, and Russia for many centuries. Russian and Alaskan native villages continue to harvest whales from the western Arctic population based on a quota approved by the International Whaling Commission (IWC) and, in Alaska, managed jointly by representatives of native villages and U.S. federal agencies. Management policies have been successful, allowing continued growth of the population despite annual harvests for whales by several villages. In recent years, the annual native harvest has been 20–50 whales in Alaska. For many years, there has been concern about effects of marine petroleum exploration and development activities in the habitat of bowhead whales, especially in the Beaufort Sea. Offshore oil activity is known to influence movement patterns of whales during migration, but demographic effects have not been demonstrated to our knowledge.

The western Arctic bowhead whale population appears to be recovering while other bowhead whale populations are not. The most likely reason is that other populations were pushed much closer to extinction by commercial whaling, limiting the capacity for recovery. It appears that the western Arctic population will continue to recover as long as harvests by native villages are regulated conservatively and other significant risk factors do not emerge.
TABLE V
Species, Subspecies, or Populations Once Thought to Be Near Extinction But Now Showing Evidence of Increasing Numbers

<table>
<thead>
<tr>
<th>Taxon or population</th>
<th>Range/habitat</th>
<th>Population identity/structure</th>
<th>Estimated population size</th>
<th>Primary risk factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western Arctic population of the bowhead whale: <em>Balaena mysticetus</em> Linnaeus, 1758</td>
<td>Coastal and pelagic, Arctic and subarctic</td>
<td>One population</td>
<td>Approximately 8000</td>
<td>Commercial whaling, disturbance from offshore petroleum exploration and development</td>
</tr>
<tr>
<td>Humpback whale: <em>Megaptera novaeangliae</em> (Bennett, 1846) Arctic and subarctic</td>
<td>Prox populations in the Northern Hemispheres; three in the North Pacific; one in the North Atlantic; and one in the Arabian Sea</td>
<td>Recent surveys in the North Pacific indicate 6000–8000 individuals, with aggregate population growth rates of approximately 7% per year; numbers are not well-known for other populations</td>
<td>Commercial whaling, incidental take in fishing nets, habitat disturbance from recreational vessels</td>
<td></td>
</tr>
<tr>
<td>Eastern North Pacific gray whale: <em>Eschrichtius robustus</em> (Liljeborg, 1861)</td>
<td>Coastal from the Bering and Chukchi Seas southward to the southern Gulf of California, Mexico, along the west coast of North America</td>
<td>One population</td>
<td>6000–8000 individuals with aggregate population growth rates of approximately 7% per year; numbers are not well-known for other populations</td>
<td>Commercial whaling, incidental take in fishing nets, habitat disturbance from industrialization and recreational boating in and near breeding lagoons in Mexico and along migration route</td>
</tr>
<tr>
<td>Northern elephant seal: <em>Mirounga angustirostris</em> (Gill, 1866)</td>
<td>Breeding and molting seasons: coastal, on island and mainland haul-outs along the Pacific coast of northern Mexico and California. Other seasons: pelagic in cool-temperate and subpolar latitudes of the North Pacific</td>
<td>Separate breeding colonies at each of the major haulouts, exchange among colonies is known to occur</td>
<td>Approximately 200,000 individuals, increasing at approximately 9% per year</td>
<td>Commercial whaling, disturbance of breeding colonies</td>
</tr>
<tr>
<td>Galapagos fur seal: <em>Arctocephalus galapagoensis</em> (Fischer, 1844)</td>
<td>Coastal and pelagic habitats of the Galapagos Archipelago</td>
<td>Separate breeding colonies at each of the major haulouts, exchange among colonies is known to occur</td>
<td>Approximately 30,000–35,000 individuals, thought to be increasing in numbers</td>
<td>Commercial whaling, habitat disturbance</td>
</tr>
<tr>
<td>Subantarctic fur seal: <em>Arctocephalus tropicalis</em> (Gray, 1822)</td>
<td>Subantarctic islands north of the Antarctic convergence zone in the Antarctic Ocean</td>
<td>Separate breeding colonies at each of the major haulouts; population structure is otherwise unknown</td>
<td>Approximately 100,000–460,000 individuals; thought to be increasing in numbers</td>
<td>Commercial whaling, habitat disturbance</td>
</tr>
<tr>
<td>Antarctic fur seal: <em>Arctocephalus gazella</em> (Peters, 1875)</td>
<td>Subantarctic islands north and, in a few cases, slightly north of the Antarctic convergence zone in the Antarctic Ocean</td>
<td>Separate breeding colonies at each of the major haulouts; population structure is otherwise unknown</td>
<td>Approximately 1,200,000 individuals, thought to be increasing in numbers</td>
<td>Commercial whaling, habitat disturbance, entanglement in marine debris</td>
</tr>
<tr>
<td>Northern and Russian sea otter: <em>Enhydra lutris</em> (Wilson, 1803, and E. l. lutris* (Linnaeus, 1758)</td>
<td>Shallow marine habitats along the exposed outer coast</td>
<td>Northern subspecies ranges from western Alaskan Islands to the Columbia River mouth; Russian population ranges through the Commander and Kuril Islands, southern Kamchatka, and northern Hokkaido</td>
<td>Approximately 100,000 individuals in Alaska and approximately 20,000 individuals in Russia; data for many Aleutian locations are obsolete; populations in the Aleutian Archipelago may be declining rapidly</td>
<td>Commercial hunting, incidental take in fishing nets, contamination, increased predation rates by killer whales</td>
</tr>
</tbody>
</table>
3. Humpback Whale, Megaptera novaeangliae (Borowski, 1781)

Humpback whales are distributed globally in marine habitats. At least five populations are recognized in the Southern Hemisphere, three in the North Pacific, and two in the North Atlantic. An unusual nonmigratory population occurs in the Arabian Sea. Humpback whales often are found near shore. Lengthy seasonal migrations between high-latitude summer feeding areas and low-latitude winter breeding areas are well documented for all populations except as noted.

Humpback whales were among the earliest target species of commercial whalers because of their abundance, large size, and coastal distribution. During twentieth-century commercial hunts in the southern ocean, several hundred thousand individuals were taken. Illegal harvest by whalers from the USSR was extensive and was extended to the populations following the decline of large-scale legal commercial whaling. Subsistence harvest by indigenous peoples is common in previous centuries. Additional modern risk factors include incidental take in fishing nets and habitat disturbance by large recreational cruise ships.

 Although trends in most populations have not been confidently documented, there is a broad perception that humpback whale numbers are increasing on a global scale. Recent estimates are 9000–12,000 individuals in the North Atlantic and a combined total of 6000–8000 individuals in the North Pacific. Data from the Southern Hemisphere and the Arabian Gulf are not sufficient to develop confident population estimates. There is concern that the population wintering near Tonga, in the western South Pacific, may have experienced excessive subsistence harvests in recent decades, although the hunt has been stopped. The eastern and western North Pacific population remain small enough to be of ongoing concern. The relatively small Arabian Gulf population came to be known because of extensive Soviet whaling in the 1960s. Recent concerns have been expressed about the sustainability of current levels of incidental taking by local fisheries. We suggest that prospects for survival of humpback whales, as species, are good as long as commercial harvests are not resumed, incidental taking is reduced, and habitat disturbance by human activities continues to be restricted and carefully monitored.

4. Eastern North Pacific Gray Whale, Eschrichtius robustus (Liljeborg, 1861)

The eastern North Pacific population of gray whales ranges from the Bering and Chukchi Seas of the Arctic region during summer to the shallow protected coastal lagoons of Baja California, Mexico, during winter months. The annual migration of this population is among the longestest of the world’s mammals and is the best documented and most familiar of the cetaceans.

Eastern gray whales were hunted extensively by commercial whalers from the eighteenth through the early twentieth centuries. Preexploitation estimates of population size are inconsistent and difficult to interpret. The population may have reached a minimum of approximately 4000 individuals in the late 1890s. The population numbered approximately 11,000 individuals when regular quantitative surveys were begun in 1967. Based on 19 surveys over a span of 30 years, the current population is thought to be about 27,000 individuals. Patterns in population trend, observed adult mortality rates, and the number of whales in unusual locations during summer all support the perception that the population is approaching carrying capacity. In 1994, the eastern gray whale was the first marine mammal to be removed from the ESA “List of Endangered Species.”

Several issues are of current concern for eastern gray whales. Modest rates of incidental take in fishing nets are reported. Disturbances by boats supporting “whale-watching” activities are of concern, both in migratory corridors along the heavily populated west coast of North America and in the breeding and calving lagoons in Mexico. Recently suspended plans to expand salt extraction industries at Mexican lagoons drew substantial opposition from those interested in gray whale conservation. Possible demographic damage to gray whales by existing saltworks remains a matter of speculation.

Eastern gray whales are subject to ongoing annual subsistence harvests by indigenous peoples in Russia, Alaska, and Washington state. The subsistence harvests are authorized by IWC quota and, in the United States, are comanaged by village elders and government agencies. Subsistence hunts have taken over 100 whales per year for several decades, with 95% of the harvest occurring along the Russian coast of the Bering and Chukchi Seas during summer months. In 1998, the Makah tribe of northwestern Washington state was allocated a harvest quota of 3 whales per year for 4 years in exchange for a comparable reduction of the quota for Russian and Alaskan hunters. The allocation of harvest quota to the Makah people was consistent with the Treaty of Neah Bay of 1855 (12 Stat. 930) between the Makah tribe and the U.S. government. One whale was taken in 1999 by Makah hunters.

In our opinion, the eastern gray whale population is no longer at risk of extinction in the foreseeable future.
future. The status of the population should remain good as long as subsistence harvests are conservative and carefully managed in the context of emerging knowledge of population structure. Increasing problems with incidental taking, disturbance by boats, and industrialization at the calving/breeding lagoons in Mexico must be monitored and regulated to ensure that these issues are specifically managed.

5. **Northern Elephant Seal**, *Mirounga angustirostris* (Gill, 1866)

Northern elephant seals breed and molt on coastal islands of Baja California, Mexico, and southern and central California. Since 1973, at least four mainland breeding and molting colonies have developed in central California. When not hauled out, northern elephant seals forage in pelagic habitats of the temperate and subpolar North Pacific. Two complete migrations are made each year between hauling sites and foraging habitats. Seals pup, nurse, and wean young and breed from December through mid-March at the hauling sites, then they swim north to forage. They return to haulouts to molt during spring and early summer, and then return again to foraging areas. Schedules for migration vary among age and sex categories. When hauled out, elephant seals are concentrated at high density. Hauled seals are easily approached by humans, even more so than other land-breeding pinnipeds of North America.

Northern elephant seals have been utilized for food and oil over the centuries by native peoples of North America. Intensive commercial harvests for oil began early in the nineteenth century. Most colonies of seals were severely depleted by 1850, but commercial harvests continued until 1884, at which point the species was considered extinct. Subsequent discoveries of small numbers led to continued harvests by scientific collectors working for museums of natural history, including the Smithsonian Institution. Scientific collecting continued until at least 1911. The total number of surviving seals is thought to have been as low as 20–100 in 1880. At the time harvests finally ended, elephant seals survived only at Guadalupe Island off the west coast of Baja California.

Northern elephant seals numbered approximately 127,000 individuals in 1991. The total is probably now approaching 200,000. They occupy at least 16 major hauling sites in Mexico and California. Population growth has been approximately 6% per year through the latter half of the twentieth century. Recent studies indicate a marked lack of genetic diversity at examined loci, almost certainly a result of the severe population "bottleneck" associated with commercial and museum harvests. In contrast, several of the southern elephant seal populations currently in decline have much higher levels of measured genetic diversity than those of the northern species.

The elimination of commercial harvest has allowed northern elephant seals to recover fully from near extinction despite a loss of genetic diversity. Most current management concerns for the species involve perceived problems of overabundance rather than rarity. Indefinite survival should be ensured as long as harvests and disturbances to habitat can be monitored and controlled.

6. **Northern and Russian Subspecies of the Sea Otter**, *Enhydra lutris* kenyoni

Wilson, 1991, and *E. lutris lutris* (Linnaeus, 1758)

The northern subspecies of sea otter ranges from the Aleutian Archipelago eastward and southward along the coasts of Alaska, British Columbia, and Washington to the mouth of the Columbia River. The Russian subspecies ranges from northern Hokkaido, Japan, through the Kuril Archipelago to northern Kamchatka and the Commander Islands. Precommercial estimates of population size are not available. As noted in Section II,D,6, large-scale commercial hunting of sea otters for pelts began in 1741. By the time legal hunting ended in 1911, the combined number of individuals in both subspecies was probably less than 2000 individuals, scattered among 10 isolated remnant populations from the Queen Charlotte Islands, Canada, to the Kuril Islands. The Queen Charlotte population was extinct soon after. The two subspecies of sea otter have been subject to ongoing subsistence harvest by native peoples for meat and pelts, probably for many centuries.

Populations from Prince William Sound westward have largely recovered without assistance, other than prohibition of harvest, during the twentieth century. Observed annual population growth rates have been as high as 10–15% in some cases. From the 1950s through the early 1970s, several translocation projects were attempted, moving groups of animals from Prince William Sound and Amchitka Island eastward to the Pribilof Islands, southeastern Alaska, British Columbia, Washington, and Oregon. The projects in the Pribilofs and Oregon failed, but all others succeeded, producing large populations in southeastern Alaska and small but rapidly growing populations off British Columbia and Washington.

Illegal commercial harvest of sea otters has been an occasional problem throughout the range. Poaching increased in Kamchatka with the development of eco-
nomic crises following the fall of the Soviet Union in 1991. The current scope of the poaching problem is unknown. Legal subsistence harvest by native villages in Alaska has averaged approximately 500 animals per year since the mid-1980s. The harvest is concentrated primarily in Prince William Sound and southeastern Alaska. The Exxon Valdez oil spill of 1989 killed several thousand sea otters in Prince William Sound and nearby coastal areas. Despite the intensive public interest and media coverage that facilitated apocalyptic scenarios, it appears that sea otter numbers in the sound did not experience long-term reduction, with the possible exception of a few local areas most heavily affected by the oil spill. Some effects on sea otter prey populations and habitats may persist to the present, but sea otter numbers are large and increasing in most areas affected by the spill.

Currently, Russian sea otter populations include approximately 20,000 individuals and are viewed as stable and near carrying capacity. Alaskan populations number approximately 100,000 individuals, although many of the survey data are obsolete. Recently, some populations in the Aleutian Islands have declined rapidly for unknown causes. There has been speculation about possible effects of contaminants and of increased rates of predation on sea otters by killer whales. A clear understanding of the scope and causes of the apparent declines will require new survey data and more definitive studies of population risk factors.

A. Synopsis of Factors and Processes Known to Be Facilitating Modern Anthropogenic Extinctions of Marine Mammals

In our review of taxa and populations of marine mammals currently in jeopardy of anthropogenic extinction, there are two major, recurring categories of vulnerability. The first is prolonged excessive harvest, primarily in the past and usually for commercial purposes, reducing numbers to a small fraction of preexploitation status. In this category we find taxa or populations from all marine locations on Earth, including many that are distant from human population centers. The second is a combination of risk factors strongly linked to a proximate and encroaching human population. The factors include habitat loss or chronic habitat disturbance, incidental taking in fishing gear, and contaminants as well as directed harvests. The second category primarily includes taxa or populations restricted ecologically to a limited geographic range in nearshore marine or aquatic habitats. Here, we consider the vulnerabilities of taxa or populations in the two categories.

The first category primarily involves excessive directed harvest. All species of mysticetes, the larger odontocetes, nearly all pinnipeds, the marine otters, and the polar bear have been hunted extensively at least during the past few centuries, in most cases in order to obtain articles of commerce. Most exploited taxa lack the necessary demographic features to sustain viable populations at the level of harvest experienced (see Section 1.C.6). There are two important results. First, such exploitation has often reduced populations to small sizes. Second, populations so affected require decades or even centuries to recover to levels free of the risks of extinction. Directed, commercial-scale harvests of marine mammals ended for most species during the twentieth century, and many species are now subject to rigorous protection. However, the risk of extinction persists for reduced populations despite relaxation of harvest activity.

Small populations are vulnerable to any factor that reduces survival or collective reproductive success. Survival and reproduction can be impaired by anthropogenic factors, such as contaminants or disturbance to critically important breeding locations, or by natural fluctuations in the biological habitat. Anthropogenic factors should be controllable in principle by the appropriate management actions, but in reality effective risk management is difficult and often fails. Natural fluctuations are effectively stochastic in timing, duration, and intensity and cannot be anticipated or controlled by any form of management authority. Precautionary management can, however, reduce risks associated with stochastic events in some cases.

Once marine mammal populations are reduced, they recover slowly and are therefore at risk of the damaging effects of anthropogenic or natural disturbances over an extended period even under the most rigorous protection. For example, North Atlantic right whales have not been harvested for decades, but the harvest reduced them to low numbers. Now, even modest rates of ship-
strike mortality or incidental take in fishing gear are adequate to hold the population at a dangerously small size. A major change in ocean productivity, however brief, could easily push the populations to virtual extinction without any possibility of effective human intervention.

The second category involves a group of factors associated with increasing human populations in coastal regions, interacting with taxa or populations constrained to life in the coastal zone. Species in this category include the sirenians, river dolphins, and several coastal odontocetes. The essential problem here is that increasing human populations produce a suite of effects, each damaging to proximate marine mammal populations. Reduction of the effects often requires a deliberate curtailment of economic enterprise such as fishing or of institutional infrastructure such as waste disposal, flood control, or the provision of drinking water. The emergent dilemma is the perception by political institutions that there must be a choice between human welfare and the welfare of nearby marine mammals. Our case studies suggest that, given the choice, human cultures of major population centers act in favor of human needs. Thus, for example, recreational boating activities continue to crowd needed habitat for manatees in Florida despite 20 years of documentation that manatees do not tolerate powerboat activities. Fishing interests continue to set nets in vaquita habitat despite general recognition that incidental take is driving the population to extinction. The latter case is complicated by the lack of economic alternatives for the artisanal fishers of the region.

Species in our first category have reasonable probability of survival, given some luck. Sea otters and northern elephant seals have escaped the window of vulnerability associated with small population size, and other taxa seem well on their way. Some taxa or populations probably will not persist. Northern right whales and western gray whales will survive the new century only with good fortune and the most rigorous imaginable protection. We are less optimistic about species in our second category. Their ultimate survival depends on conscious economic restraint by human cultures and a possible reevaluation of values regarding the survival of marine mammals and other species in habitats also used or coveted by people.

Excessive subsistence harvests, anthropogenic noise, contaminants, oil spills, and depletion of genetic diversity are issues that have at least occasionally been invoked as risk factors for extinction of marine mammals. We find that there are relatively few taxa or populations clearly falling toward extinction as the direct result of any one of these factors. Subsistence harvest by native peoples is without question a serious risk factor for Cook Inlet beluga whales, and it may have been a crucial precursor to the extinction of Steller’s sea cow. However, the western Arctic bowhead whale population has been increasing steadily for years despite regular annual subsistence harvest. Thus, subsistence harvests are manageable risk factors and need not be regarded universally as unacceptable practice. None of the other listed factors are alone causing widespread extinction risk, although there are isolated examples for each. Of greater concern here is the problem of significant effects from the interaction of multiple factors. The best known cases involve mass mortalities that result from disease outbreaks. Such outbreaks often result from immunosuppression, which in turn may result from contaminants or from natural disturbances such as the toxic by-products of certain phytoplankton. Interacting factors are often a problem near human population centers and are difficult, if not impossible, to manage. Thus, we believe that the danger of many risk factors discussed here is not the direct effect of a single factor but rather the synergistic effects of multiple factors that may be less damaging when separated from one another.

The detection of extinction risks for both categories of species will continue to be confounded by the inherent uncertainties of population data for marine mammals. Several kinds of errors are possible in the future, although predictions of error rate are beyond the scope of this review. Populations that are numerically stable may be categorized incorrectly as declining and will receive unwarranted research effort. Declining populations may be categorized incorrectly as stable and may be denied attention from survey efforts necessary to detect problems and plan recovery actions. Recovery efforts may be directed to one component of life history, such as juvenile survival, when similar efforts toward other components, such as adult female survival, could produce far greater return for the same effort. All the error types will be more likely to influence perceptions of small populations in remote locations, where the incremental cost of survey effort is prohibitively large.

B. Geographic Regions of Greatest Concern with Regard to Anthropogenic Extinctions of Marine Mammals

As previously suggested, marine mammal taxa and populations constrained to life near human population centers are in general most vulnerable to anthropogenic extinction. Extinction risks will be greatest where human cultures have the fewest economic options when
confronted with the need for restraint in order to solve conservation problems. Such circumstances are most likely in "developing" countries at lower and middle latitudes in which large concentrations of people face ongoing economic shortfalls. During the next century, we anticipate the greatest extinction risks for coastal marine or aquatic marine mammals off southern and southeastern Asia, eastern Europe, Central America, and central Africa. High rates of per capita resource consumption in the "developed" countries are also linked to risk factors for marine mammal populations. Thus, extinction risks are a significant concern for coastal marine mammals off North America and western Europe. Example cases include manatees in Florida, sea otters off California, and ringed seals in Lake Saimaa, Finland.

Marine mammals in the Southern Hemisphere generally should be at lower risk of extinction during the next century than those in the north. Most major human population centers and most cases of coastal habitat degradation occur in the Northern Hemisphere. However, some southern populations remain small because of past excessive harvests. Thus, distantly located southern taxa and populations will be removed from the risk of extinction only to the extent that all forms of harvest are regulated with extreme caution and conservatism. Moreover, small distant populations are at risk of incorrect conclusions about number of individuals, trend in numbers, or demographic characteristics because of the statistical limitations associated with survey and demographic data.

C. General Approaches toward Minimizing the Rate of Anthropogenic Extinctions of Marine Mammals

Excessive directed harvests have caused more cases of jeopardized marine mammal taxa and populations than any other single factor. Thus, the most direct and straightforward approach to the control of extinction risk for marine mammals is a precautionary approach to the concept of marine mammal harvest on a global scale. Fortunately, this is the approach currently adopted by many governments, and by international regulatory cooperatives such as IWC, for some species. In this context, we offer three points of caution. First, the protections provided by international treaties and conventions such as the IWC, and by individual governments, do not necessarily extend to all marine mammals. There are high-profile protective protocols, with active ongoing oversight, for larger cetaceans, some pinnipeds, sea otters, and polar bears. Many small cetaceans, some pinnipeds, and some sirenians are not actively and explicitly protected at the national or international level. Second, many of the populations subject to active protection are very small and as a consequence will be subject to risks associated with stochastic events, both natural and anthropogenic, for many decades. Thus, some extinctions are possible because not all species are equally protected, and because the most aggressive protection cannot eliminate all risk factors. Third, the crucial process of detecting trends in small distant populations is so costly that errors are likely in determining which populations are most seriously jeopardized. Thus, despite the best of human intentions, protective effort may be directed inadvertently to the wrong taxa or populations.

The group of risk factors associated with human population growth will almost certainly cause some extinctions of coastal marine mammals in the next century. Here, we believe that the outcome is more certain, and the methods of prevention more intractable, than in the simpler cases of small populations distant from major concentrations of people. Effective protection of imperiled species requires that human cultures forego economic benefits for the good of jeopardized marine mammals. Such sacrifices must extend indefinitely to be effective, given the demographic limitations of most marine mammals. The acceptance of foregone benefits is most needed in cultures least able to accommodate the loss. Cultural acceptance of economic loss motivated by a conservation ethic will require education, reorientation, and provision of meaningful economic alternatives at a level of effectiveness that we find difficult to imagine. We conclude with the reminder that both evolution and extinction of species have been characteristic features of the history of marine mammal taxa on Earth since the early Eocene. Although the anthropogenic loss of species is both regrettable and inevitable, there is no reason to deny that new species will evolve as well. For example, killer whales off Washington and British Columbia occur in three distinctive social configurations that correlate with subtle but reliable morphological differences. Individuals in the different categories have entirely different diets, different acoustic repertoires, rarely co-occur in space, and do not interbreed. Some have characterized these patterns as a step in the process of speciation, although there are alternative viewpoints. The human mind can easily perceive extinction as a finite event, but not speciation because of differing timescales and imprecise definitions of the
latter. Thus, we are biased toward greater conscious awareness of extinction than of speciation. Although anthropogenic extinctions should not be accepted without exhaustive attempts at recovery, it is perhaps more reasonable to view such events as parts of a dynamic process of evolution rather than as catastrophic failures of human behavior.

See Also the Following Articles

EXTINCTIONS, CAUSES OF • EXTINCTIONS, MODERN EXAMPLES OF • MAMMALS, BIODIVERSITY OF • MARINE ECOSYSTEMS, HUMAN IMPACT ON

Acknowledgments

We thank Carlos Alvarez-Flores, Amanda Bradford, Allan Fukuyama, Laura Lively, Josh London, Robert Suydam, and Alexandre Zerbini for technical guidance and constructive reviews of the draft manuscript. We received financial support from the Biological Resources Division of the U.S. Geological Survey, the School of Fisheries of the University of Washington, the National Center for Ecological Analysis and Synthesis at the University of California, Santa Barbara, and the Southwest Fisheries Science Center of the National Marine Fisheries Service (U.S.).

Bibliography


MARINE SEDIMENTS

Paul V. R. Snelgrove
Memorial University of Newfoundland

I. Benthos and the Benthic Environment
II. Marine Sedimentary Habitats
III. Global Estimates of Species Numbers
IV. Global Patterns of Biodiversity
V. Regulation of Diversity
VI. Ecosystem Services and Sedimentary Diversity
VII. Threats to Sedimentary Diversity
VIII. Summary

GLOSSARY

abyssal plains Relatively flat areas of the ocean bottom below ~4000 m depth.
benthic Aquatic bottom habitat, encompassing areas on or below the interface between the water and the bottom.
benthos Bottom-living organisms, including those that reside on hard and soft bottom surfaces and others that reside between sediment grains.
continental rise An area at the base of the continental slope between 3000 and 4000 m where the bottom slope is slight and sediments often accumulate.
continental shelf A region of ocean bottom extending from the low water mark at the edge of continents to a depth (~200 m) at which the incline increases markedly and the continental slope begins.
continental slope Ocean bottom extending from the edge of the continental shelf at an ~4° incline to a depth (3000–10,000 m) at which the slope decreases and the continental rise begins.
deposit feeders Organisms that feed primarily by ingesting organic material occurring on or between sediment grains.
emergent vegetation Plants that are attached to the bottom but extend up through the water column above the ocean surface.
infrauna Organisms living below the sediment–water interface and between sediment grains.
microfauna Animals large enough to be retained on a 300- or 500-μm sieve.
mega fauna Animals that are sufficiently large to be identified from bottom photographs.
meso fauna Animals small enough to pass through a 300-μm sieve but large enough to be retained on a 63-μm sieve.
phytoplankton Unicellular and chain-forming algal plants that are suspended and transported in the water above the bottom.
suspension feeders Organisms that feed primarily on organic material suspended in the water above the bottom.
trench A steep-sided depression in the ocean floor ranging from 6000 to 10,000 m depth and associated with areas of tectonic plate subduction.

MARINE SEDIMENTS COVER ALL but a few percent of the ~70% of the earth’s surface that is ocean and as such comprise the largest single ecosystem on Earth in area. Given the vast global coverage, it is not surprising
that marine sediments are species-rich systems, particularly concerning the range of sedimentary habitats that occur. Sediments range from rubble and coarse gravel to fine clays. They occur from tidally exposed areas, where temperature, salinity, and sediment water content fluctuate greatly, to the greatest depths of the oceans, where no light penetrates, temperatures approach freezing, environmental variables are largely invariant, and pressure exceeds 100 times that experienced at sea level. The species that inhabit these environments, how they are regulated, the ecological roles that they play, and activities that threaten them vary considerably from one habitat to the next and are the basis for this article.

I. BENTHOS AND THE BENTHIC ENVIRONMENT

Depending on the specific habitat and species, different benthic (bottom-dwelling) organisms may reside just above the bottom but closely associated with it (hyperbenthos), on the sediment surface (epifauna), or among the sediment grains (infauna). Because sediments have limited permeability, very little oxygenation (penetration of millimeters) is achieved through diffusion. Greater oxygen penetration is achieved only when bottom currents mix sediments or, more commonly, when bioturbation (biological mixing of pore water and sediments by benthos) facilitates transfers of oxygenated water from the water column into the sediment pore water. Bioturbated sediments are usually oxygenated within the top few centimeters, although active bioturbation can lead to deeper pockets of penetration. Light does not penetrate well through sediments and most productivity is provided by phytoplankton detritus sinking from surface waters above, macrodetritus transported from coastal areas (e.g., kelps and seagrasses), or from primary producers living attached to the sediment surface in shallow areas where light penetrates all the way to the bottom. Thus, most of the primary production or detritus in benthic systems is concentrated near the sediment surface rather than within the sediments. This combination of limited food and oxygen penetration results in the vast majority of organisms being confined to the upper few centimeters of sediment near the sediment-water interface. Smaller organisms that can tolerate anoxia and larger organisms that maintain a burrow or appendage to the surface can live deeper, but even then distributions are generally limited to 10–20 cm or less.

As in other environments, feeding modes include herbivores, predators, scavengers, parasites, and omnivores. However, most organisms that reside in marine sediments are either deposit feeders or suspension feeders, and some species are capable of either feeding mode. Deposit feeders ingest particles associated with sediments or, in many cases, they ingest the sediment particles themselves. Nutrition is provided by detritus associated with the sediment grains and also by microorganisms decomposing the detritus. In shallow areas, benthic algae (diatoms) may provide a food source for some deposit-feeding species. Deposit feeders are important to sediment geochemistry because, as they move through the sediments, they facilitate pore water movement in sediments and also move around sediment grains (bioturbate). Suspension feeders are organisms that remove particles such as phytoplankton and detritus from the water column. Because they rely on suspended particles, suspension feeders tend to be most abundant in energetic environments and absent from areas in which currents are weak and horizontal flux of food particles is reduced. Differences in the physical energetics of an environment will also influence the stability of the benthic habitats and the organisms that are able to live there. Quiescent areas usually have fine-grained sediments that are relatively stable except where storms pass through. High-energy environments are more dynamic in terms of sediment movements; fine-grained sediments are usually absent and the coarse-grained sands that remain are often moved around by ambient flow conditions. Not surprisingly, the most dynamic environments are generally in shallow areas, whereas much of the deep sea is relatively quiescent because wave energy rarely penetrates beyond 50–60 m with sufficient strength to move sediment grains.

For reasons of sampling and taxonomic practicality, benthic researchers routinely focus on just one of several size groupings. Megafauna are defined as organisms that are sufficiently large to be identified in bottom photographs and include organisms such as flatfish, crabs, and scallops. Macrofauna are defined as those that are retained on a 300- or 500-μm sieve (standards vary). This size grouping includes polychaete annelids, crustaceans, bivalves, and many other phyla (Fig. 1). Meiofauna refers to organisms that pass through a 300- or 500-μm sieve but are retained on a 44- or 63-μm sieve and include nematodes, foraminifers, tiny crustaceans, and many others (Fig. 2). Microorganisms include living organisms that pass through a 44- or 63-μm sieve and include bacteria and protists. These size groupings are not absolute in that the larval or juvenile stages of one group may be comparable in size
to individuals from a smaller group, but this division of organisms is necessary because of sampling logistics. The sample size appropriate for enumerating megafauna, for example, is inappropriate for sampling meiofauna that are orders of magnitude more abundant. Moreover, the taxonomic challenges even within each of these size groupings are considerable, and synthesis across groups is therefore rare in a single study.

These groups of organisms do not operate independently of one another. Megafauna prey on infauna, and macrofauna prey on meiofauna. There is also evidence that meiofauna prey on juvenile macrofauna. Microbes form an important dietary component for many meiofaunal and macrofaunal taxa, and macrofaunal bioturbation impacts microbial composition because it oxygenates sediments and thereby regulates the relative importance of aerobic versus anaerobic bacterial activity. Microbes also initiate breakdown of food material that would otherwise be undigestable by metazoans and play a major role in cycling of organic matter. Thus, there is considerable interaction between these groups of organisms.

Many factors are thought to influence the spatial distribution of benthic sedimentary organisms, includ-
ing sediment type, productivity, temperature, salinity, pressure, depth, oxygen, sediment stability, air exposure, current speeds and wave action, and biological interaction with other species.

II. MARINE SEDIMENTARY HABITATS

Oceans range in depth from intertidal habitats to ocean trenches at 10,000-m depth (Fig. 3). Light can penetrate to 1000 m in the clearest oceanic waters, but in coastal waters penetration may be limited to from 200 m to only a few meters. Like most ecosystems, marine sedimentary communities are fueled by photosynthesis, and because many marine sediments occur below photosynthetic depths they rely on phytoplankton sinking from surface waters above and macrophyte detritus transported from nearshore environments. Not surprisingly, water depth is a major variable in categorizing marine sedimentary habitats.

A. Intertidal Areas

The shallowest sedimentary habitats are those that occur at the land–sea interface. They include habitat with emergent vegetation such as mangroves and salt marshes as well as sandflats and mudflats without emergent vegetation that may or may not include diatom mats (Table 1). Mangroves and salt marshes occur intertidally, and as such they experience wide ranges of salinity and temperature. They are also extremely productive, and organic matter in the form of decaying vegetation is abundant. Because of the high productivity of these systems, bacterial respiration is extremely high and sediments are often hypoxic or even anoxic unless they are right at the sediment–water interface. The organic matter produced by vascular plants is also relatively refractory and difficult for most species to digest. In combination with oxygen limitation and tidal exposure, this refractory material contributes to the low diversity of organisms residing in the sediments. Unvegetated sandflats and mudflats tend to be much lower in productivity unless they are immediately adjacent to vegetated areas, and they have slightly higher diversities. Nonetheless, intertidal areas experience high variation in environmental variables and are characterized by a low-diversity fauna consisting of organisms able to cope with this variability.

B. Subtidal Areas

At depths below low tide, species diversity increases relative to that of intertidal systems. In the shallowest
TABLE I
Summary of Sedimentary Habitats, Local Diversity, Ecosystem Services Provided, and Current Threats

<table>
<thead>
<tr>
<th>Sedimentary habitat</th>
<th>Diversity</th>
<th>Ecosystem services</th>
<th>Current threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salt marshes</td>
<td>Low</td>
<td>Nutrient cycling, critical habitat, shoreline stability, filtration, productivity, pollutant cycling</td>
<td>Habitat destruction, pollution, global climate change</td>
</tr>
<tr>
<td>Mangroves</td>
<td>Low</td>
<td>Nutrient cycling, critical habitat, shoreline stability, filtration, productivity, pollutant cycling</td>
<td>Habitat destruction, pollution, global climate change, exotic species</td>
</tr>
<tr>
<td>Sandflats</td>
<td>Low to modest</td>
<td>Nutrient cycling, filtration, productivity</td>
<td>Habitat destruction, pollution, exotic species</td>
</tr>
<tr>
<td>Mudflats</td>
<td>Low to modest</td>
<td>Nutrient cycling, productivity, pollutant cycling</td>
<td>Habitat destruction, pollution, exotic species</td>
</tr>
<tr>
<td>Estuaries</td>
<td>Low</td>
<td>Nutrient cycling, critical habitat, filtration, productivity, pollutant cycling</td>
<td>Habitat destruction, pollution, global climate change, exotic species</td>
</tr>
<tr>
<td>Coral reef sediments</td>
<td>Very high</td>
<td>Nutrient cycling, productivity</td>
<td>Habitat destruction, pollution, fishing</td>
</tr>
<tr>
<td>Continental shelf</td>
<td>Moderate to very high</td>
<td>Nutrient cycling, critical habitat, filtration, productivity, pollutant cycling</td>
<td>Habitat destruction, pollution, fishing, exotic species</td>
</tr>
<tr>
<td>Continental slope</td>
<td>Very high</td>
<td>Productivity, nutrient cycling (global)</td>
<td>Habitat destruction, fishing</td>
</tr>
<tr>
<td>Abyssal plains</td>
<td>High to very high</td>
<td>Nutrient cycling (global)</td>
<td>—</td>
</tr>
<tr>
<td>Hydrothermal vent sediments</td>
<td>Low</td>
<td>Nutrient cycling (global)</td>
<td>—</td>
</tr>
<tr>
<td>Deep ocean trenches</td>
<td>Low</td>
<td>?</td>
<td>—</td>
</tr>
</tbody>
</table>

*The table summarizes generalities for which there are exceptions. In the case of global climate change, the large-scale circulation changes predicted in some models would threaten all habitats, and only more immediate impacts are included here.*

C. Deep-Sea Areas
The deep sea, defined here as habitats beyond the edge of the continental shelf at ~200 m, encompasses continental slopes, continental rises, abyssal plains, midocean ridges, and trenches. Although the sediments that cover the bottom can vary in terms of grain size composition, environmental variability is considerably reduced. Temperatures are largely invariant, salinity does not change, and sediments are rarely moved around other than by bioturbation. Depending on the region, productivity is generally less seasonal than in shallow water, but for the most part deep-sea benthic environments have much less productivity entering the system than their shallow-water counterparts. Most deep-sea areas are below light penetration and productivity consists of whatever material sinks through the thousands of meters of water column to the bottom. As a result of this reduced productivity, densities of organisms are very low and individuals are relatively small; surprisingly, species diversity is quite high. A very notable exception to low productivity in the deep sea is hydrothermal vents and seeps, for which chemoautotrophic bacteria

Subtidal areas, vegetation in the form of seagrass and kelp may occur, and some shallow sediments may be covered in diatom mats. However, all these plant forms are confined to depths of tens of meters or less, and most subtidal sedimentary habitats rely on sinking plant material as the base of their food chain. Sedimentary environments may range from gravel to fine clays. Fine-grained sediments are characteristic of low-energy environments and coarse-grained sediments are characteristic of high-energy areas. The latter are often dynamic environments, with sediments often moving around in response to bottom flow. Continental shelves are largely covered in sediments and are inhabited by a wide variety of organisms. The shallower portions of these environments (up to tens of meters) can experience strong seasonality in productivity and temperature, and where offshore runoff is significant (e.g., estuaries) salinity may vary considerably. The greater the environmental variability in these systems, the lower the diversity. With increased depth, the variability in salinity and temperature is generally reduced, although productivity can still be seasonal. Continental shelf areas are large, productive areas of the oceans and support the majority of marine commercial fisheries.
provide the basis for a very productive localized community.

Not all deep-sea communities are diverse. Because hydrothermal vent environments are characterized by fluids rich in compounds such as hydrogen sulfide and heavy metals, most species are unable to tolerate these toxic compounds and diversity is quite low. Deep-sea trenches are subject to mud slumping and poor circulation. Not surprisingly, species diversity is very low. Upwelling regions are also generally low in diversity, likely because large amounts of organic matter accumulate on the ocean floor and decompose, leading to hypoxia. A few deep-sea areas are subject to intense “storms,” in which currents become intense and sediment resuspension occurs. Evidence suggests that macrofaunal diversity is depressed in such areas, but surprisingly meiofaunal diversity is not. Presumably, the meiofauna are able to cope with the disturbance more effectively than are the macrofauna.

III. GLOBAL ESTIMATES OF SPECIES NUMBERS

Several attempts have been made to estimate the total number of species that live in the ocean. The Danish marine scientist Thorson estimated in the early 1970s that 100,000 species would eventually be described from the marine environment, and until recently this was the ballpark figure most thought appropriate. In the late 1960s, deep-sea biologists Sanders and Hessler documented that the deep sea was not the low-diversity habitat that many had previously thought and could contain a significant portion of global marine species diversity. The debate on species estimates heated up considerably when Grassle and Maciolek published data from a collection of samples along a 176-m-long depth contour along the east coast of the United States, constituting the most extensive quantitative data set ever assembled for a single area of the deep sea. Based on the rate that species were added with increased area sampled, and the total area of the deep sea, they extrapolated to 10 million macrofaunal species. May pointed out that approximately half of the species in their study were new to science, and one could extrapolate from known to unknown species using a similar ratio. Given that approximately 230,000 marine species have already been described, May’s extrapolation suggests ~500,000 total species living in the oceans. Recent work from the Pacific suggests that only 1 in 20 species has been described for that area of the deep sea, in which case May’s approach would yield ~5 million species.

Although most of the discussions on species numbers have centered on macrofauna, Lambshead, a nematode ecologist, extrapolated upward from Grassle and Maciolek’s estimate based on the fact that nematodes are more abundant and species rich than macrofauna, at least on local scales. Scaling up from their number, he estimates that there may be 150 million species of nematode alone. Currently, we know little about how widely distributed individual nematode species may be, making extrapolation even more tenuous than that for macrofauna. For microbial groups we know very little. There is evidence that protists may not be very diverse on a global scale, but studies on water column bacteria indicate vast amounts of undocumented diversity. Clearly, more data are needed to improve estimates of global species numbers. Current estimates for metazoans range from 500,000 to 10 million species, but vast areas of the ocean and some taxonomic groups remain unsampled or poorly sampled. The South Pacific and tropical latitudes are among the least known areas. Even many shallow environments remain poorly known, and recent evidence (described later) suggests that some of these areas may be very species rich. A United States National Research Council committee summarized recent surveys and showed that more than two-thirds of the polychaetes in Hawaiian coral reef sediments were previously undescribed; even Georges Bank, a relatively well-studied area, was found to have only two-thirds of its polychaete species previously described. The situation is even more severe in deep-sea areas and for meiofaunal taxa such as nematodes, for which proportions of known species can sometimes be less than 10–20% of total species present. Even for taxa thought to be well-known, there is increasing evidence based on genetic data that “cosmopolitan taxa” are often species complexes that are difficult to distinguish based on morphological characters. Regardless of which projection of total species number is appropriate, there remains a huge amount of undescribed species diversity in marine sediments.

IV. GLOBAL PATTERNS OF BIODIVERSITY

The historical perception of the deep sea as an azoic or species-poor environment changed markedly with a series of papers published by Sanders and Hessler. From data collected along a transect running from Martha’s Vineyard, Massachusetts, to Bermuda, they demon-
strated that diversity in deep-sea sediments exceeded that in most shallow areas and rivaled that observed in shallow tropical areas. These findings contradicted a previous notion that diversity would generally decrease with depth. Rex followed up this work with extensive analysis of large-scale patterns with depth in the North Atlantic and found that diversity is highest at intermediate depths. Diversity in shallow water is low in estuaries and seasonal shallow areas, is somewhat higher on the continental shelf, increases down the continental slope, and peaks on the lower slope. At greater depths it declines again. Other workers have observed a similar pattern for other taxa, although the exact depth of the diversity peak is not always consistent. Moreover, diversity in the abyssal plains of the Pacific appears to exceed that observed at shallower depths, suggesting that a parabolic diversity–depth pattern is not universal. The generality of lower diversity in shallow water has been questioned by Gray and colleagues, who presented convincing data that the local diversity of sedimentary fauna of Bass Strait, Australia (11–51 m deep) may rival or exceed that of the deep sea. The pattern observed by Rex and others has been based largely on data from the North Atlantic, and Gray’s findings emphasize the fact that patterns described from one area may not be easily extrapolated globally.

Many studies have documented latitudinal gradients in shallow-water benthic diversity, with decreasing diversity toward the poles. In the deep sea, the existence of such a trend is the subject of debate. The pattern has been proposed for North Atlantic deep-sea macrofauna, but isopod data from the South Pacific do not indicate such a pattern.

The inconsistency of conclusions regarding global patterns in diversity indicates a need for more complete sampling coverage with latitude and depth. Improved coverage would allow more effective comparison of different oceans and reduce the impact of “noise” added by regional differences in variables such as productivity that are bound to obscure any global patterns. Within this coverage, areas that have been poorly sampled historically, such as tropical sediments and the Southern Hemisphere, must be emphasized to reduce the North Atlantic bias. An additional problem with these large-scale comparisons is that most studies focus on just one of the size groupings of organisms and generally use different sorts of sampling techniques. For example, we must ask whether described differences in large-scale patterns in different taxa and ocean basins reflect differences in biology or sampling coverage. Studies that are more comprehensive in taxonomic and geographic coverage are necessary to resolve these debates.

V. REGULATION OF DIVERSITY

A. Shallow Water

Our understanding of diversity regulation in marine systems is limited. In environments with extremes of temperature (e.g., intertidal sediments), salinity (e.g., estuaries), geochemistry (e.g., sediments near hydrothermal venting), or oxygen (e.g., eutrophied coastline, mangrove sediments, and fjords with poor circulation), there are obvious physiological limitations for many species that allow only the most robust species to exist. The same may be said for physically disturbed environments such as sandflats, which tend to be relatively low in diversity. Similarly, highly seasonal environments are reduced in diversity, presumably because relatively few organisms are able to cope with the variability in the physical environment. However, beyond these extreme habitats, there are few obvious conclusions on how diversity is regulated.

Historical events and modern conditions both play major roles in determining global pattern in diversity. Certainly, the evolutionary history of a region must be considered. Jablonski’s analysis of fossil invertebrates suggests that the tropics provide evolutionary novelty and as such contribute to the higher diversity noted in tropical latitudes. Glaciation effects are well-known to have played a role in establishing current marine patterns. For example, the reduced diversity observed in Arctic relative to Antarctic fauna likely relates to extinctions during Arctic glaciation and the slow reestablishment of the Arctic community. Thus, historical events set the stage for current ecological processes, which may subsequently play a role. In terms of modern conditions, the high level of total energy input in the tropics has been proposed to explain shallow-water mollusk diversity gradients, and the high degree of seasonality with increasing latitude may be a factor that depresses diversity.

The role that ecological interactions play in sedimentary diversity is less well understood. In shallow-water experiments, it has been observed that the exclusion of predators seems to enhance diversity. This finding is the direct opposite of Paine’s classic findings in rocky intertidal communities and suggests that it is inappropriate to generalize from the relatively well-studied rocky intertidal system to the less understood sedimentary environment. Whether generalizations regarding predators depressing diversity can be extended to the deep sea remains to be tested. There is a dearth of experiments linking diversity in the water column to sedimentary faunal diversity and little data on how
predator diversity might impact benthic diversity. One might predict, for example, that increased diversity in food sources (e.g., phytoplankton) might lead to enhanced diversity in the infauna that feed on those food sources. Increased diversity of predators might also allow tighter species packing in an environment.

### B. Deep Sea

Regarding the deep sea, there has been considerable debate on how an environment that appears so physically homogeneous is able to support such a broad array of species. Sanders, who played a pivotal role in recognizing the high diversity of marine systems, suggested that the high level of stability over evolutionary time in the deep sea has allowed greater specialization and niche diversification. If this were the full explanation, then diversity should be highest on abyssal plains; the depth patterns described above do not consistently support this notion. There is also little evidence of niche specialization in deep-sea organisms. Most species appear to be relatively nonselective deposit feeders. In the early 1970s, several contrasting theories were put forward. First, cropping by predators could prevent competitive equilibria from being attained by infauna. If this were the case, however, then fast growth, early reproductive maturity, and an age structure dominated by young individuals might be expected, but this pattern is not typical for the deep sea. Moreover, evidence from shallow water suggests that predators decrease sedimentary diversity. In any case, the role that predators may play in maintaining deep-sea diversity remains largely unknown. A contrasting theory held that small patches may form disequilibria habitats that promote certain species. Tests of this patch mosaic model have included sampling of natural patches and creation of experimental patches. Both approaches have yielded a similar result: The species that occur in many patch types are usually rare or absent from background sediments, although diversity in patches is usually reduced. Such a pattern is consistent with the patch mosaic model, but experiments so far have demonstrated that patches promote only a modest number of species. Whether it is necessary to sample more patch types, or whether other explanation needs to be invoked, remains unclear.

The large area of the deep sea almost certainly impacts its species richness in that species richness of most environments increases with area. However, the vast rolling plains of the deep sea do not seem to add habitat heterogeneity with area as seen in other species-rich habitats, such as tropical rain forests or coral reefs.

Given that abyssal plains are the largest deep-sea habitat in area, the parabolic diversity–depth relationship described earlier is also inconsistent with a simple species–area relationship.

Huston’s intermediate disturbance hypothesis is another explanation for the high diversity of deep-sea systems. The idea is that at one end of the spectrum, high levels of disturbance depress diversity by eliminating sensitive species. At the other end of the spectrum, very benign habitats allow superior competitors to eliminate weaker species and thus depress diversity. Highest diversity would then be expected in habitats with levels of disturbance that prevent competitive dominants from taking over but not so frequent and harsh that few species can cope. The midslope peak in diversity described earlier is consistent with this hypothesis but does not test the hypothesis explicitly. Most disturbed deep-sea environments, such as hydrothermal vents, low-oxygen areas, environments with benthic storms, and slumping areas, have reduced diversity. However, such extreme examples are far from conclusive in drawing comprehensive generalities.

Efforts to understand regulation of biodiversity in the deep sea have been hampered by the considerable logistical challenges of deploying and recovering experiments in deep, oceanic environments. In shallow-water environments, diversity regulation has received considerably less attention, and there is a pressing need to move some of the deep-sea framework and experimental effort into shallow water where logistics are simplified and outcomes are bound to produce interesting comparisons with the deep sea. Our understanding of diversity maintenance in sedimentary systems is generally very poor, and we are unable to predict, for example, whether loss of pelagic diversity will impact benthic diversity. The problem is that humans are impacting diversity of marine benthic organisms in many areas through fishing, habitat destruction and modification, pollution, exotic species introductions, and global climate change, but we have little idea how different components of the ecosystem will be affected.

### VI. ECOSYSTEM SERVICES AND SEDIMENTARY DIVERSITY

Although the huge fraction of the earth’s surface covered by marine sediments and the large numbers of species residing within sediments provide strong motivation to understand the pattern and regulation of benthic diversity, there are also good ecological motivators.
The oceans provide a variety of ecosystem services (Fig. 4), and although we understand little about the role that biodiversity plays in maintaining these services there is very clear evidence that benthic species play critical roles. There is a very real chance that ongoing losses of sedimentary biodiversity will result in loss of ecosystem services.

A. Nutrient Cycling

Benthic marine organisms play a critical role in global cycles of nitrogen, sulfur, and carbon. As organic material sinks to the sediment surface, it carries with it organic carbon and nitrogen. This material may be directly ingested by benthic organisms or microbes may colonize it. In areas with high sedimentation rates or low densities of organisms, there may be substantial burial loss of this material, but most is consumed or decomposed. Even the organic material that passes through organisms undigested will be colonized and broken down by microbes. Organisms that ingest the material respire, are preyed on, or die and decompose in the sediment. Depending on how decomposition and digestion occur, nitrogen may be released as ammonia, nitrite, nitrate, or nitrogen gas that diffuses or is physically mixed into the water column above. In coastal sediments, this cycle is a critical part of regenerating the dissolved nitrogenous compounds that are critical for primary producers; without this regeneration the cycle would end and primary production would cease. Carbon is cycled in a similar way. Material sinks to the bottom and may eventually be buried (e.g., fossil fuels) or it may be decomposed into carbon dioxide or other carbon compounds. The relative rates at which carbon is buried, tied up in living organisms, or respired to carbon dioxide are closely tied to microbial activity and thus to the animals that influence microbial activity through feeding and oxidation of sediments. As a result, benthic organisms play a key role in global carbon budgets. Sulfur cycling through marine sediments is also dependent on sediment oxygenation and therefore bacterial activity. Whether sulfur is buried and stored in sediments or recycled is determined in part by metazoan and microbial activity.

B. Pollutant Cycling

Because benthic organisms move around and ingest particles, and may themselves be ingested by predators, they can greatly impact the burial fate and mobility of pollutants. As sediment particles are bioturbated, so are any pollutants linked to them. As a result, benthic organisms can dilute pollutants at the sediment–water interface by mixing them downward. By the same process, the continual remixing of sediments by infauna may increase burial time of pollutants that might otherwise be buried more quickly by sedimentation. Bioturbation also tends to destabilize sediments and increase the likelihood that they will be resuspended. Some sedimentary organisms can also stabilize sediments by excreting mucous (e.g., bacterial mats) or creating tubes that they inhabit. Benthos may also increase effective
grain size by repackaging grains as fecal pellets. All these processes impact whether sediments are likely to resuspend or not. Infauna that ingest pollutants may also provide a conduit up the food chain if they concentrate the pollutant in their tissues. For example, fish that feed on polychaetes with high levels of a heavy metal may also have high concentrations of that metal. Finally, benthic organisms have some capacity to break down pollutants. Microbes in particular may have the capacity to break down some toxic compounds and render them harmless.

C. Sediment and Shoreline Stability
Because benthos can destabilize (bioturbation) and stabilize (microbial mats, mucus excretion, and tube building) sediments, they can have a major impact on sediment transport and coastal geology. Mucus excretion, for example, binds sediment grains together and therefore increases the amount of energy required to move sediments. However, as some organisms produce products that bind sediment grains together, other organisms move among the sediment grains, breaking apart binding structures, increasing the water content, and increasing the likelihood of resuspension.

The vegetation in seagrass beds, mangroves, and salt marshes acts as flow baffles and therefore traps fine sediments, a process that is further aided by root structures that add further stability. The presence of these plants protects the shoreline and results in accretion rather than erosion. Indeed, rapid shoreline growth has been documented in areas with mangroves and marshes, where land is extended seaward on scales of kilometers over timescales of hundreds to thousands of years.

D. Filtration
The benthos can also impact water clarity in coastal areas. The sediment-trapping capacity of emergent vegetation reduces sedimentation to the coastal waters beyond the marsh, mangrove, or seagrass bed. In tropical areas, for example, mangroves and seagrasses can trap sediments that might otherwise be transported onto coral reefs, where the suspended sediment would reduce water clarity and productivity and potentially smother corals. In temperate latitudes, marshes are well-known not only for their ability to trap sediments but also for their capacity to take up excess nutrients. Agricultural runoff to the coastal environment can result in eutrophication (discussed later), but the presence of salt marshes can significantly offset this problem and provide a natural water filtration system. Mangroves have a similar capacity.

Suspension-feeding benthos can also have a major impact on water clarity by filtering out suspended particles. The reduction of oysters in estuaries along the east coast of the United States through overfishing, destruction of oyster reef habitat, and disease has resulted in a marked decrease in water clarity. Impacts of this magnitude are generally confined to shallow nearshore habitats with very high densities of suspension feeders, but the effect is nonetheless important.

E. Secondary Production
Benthic megafauna and macrofauna provide an important source of secondary production. Some of this production, such as clams, scallops, shrimp, and crab, is directly consumed by humans and forms the basis of important commercial fisheries. However, even groups such as the polychaetes, which have no obvious economic importance, form an important food source for benthic-feeding fishes and crustaceans. Because sinking organic material collects on the bottom, the secondary production in the benthos can often be considerably higher than that in the water column above.

F. Linking Biodiversity and Ecosystem Services
Although it is inarguable that benthic organisms provide key ecosystem services, there is no evidence that species diversity is important in maintaining these services. This deficiency does not necessarily mean that species diversity is unimportant but rather that we have not studied it. In virtually all the categories listed previously, there is a particular functional group that is critical to the service, but in most cases there is more than one species of a particular functional group in a given habitat. For example, deposit feeders are the major bioturbators in marine sediments, and bioturbation has major impacts on nutrient cycling, pollutant burial and mobilization, and sediment stability. However, most sediments that contain deposit-feeding organisms will contain multiple species. At least superficially, it appears that different deposit feeders often do the same sorts of things, although there are species that feed on particular grain sizes and at particular depth horizons in the sediment. What we do not know is whether those species that remain after others are eliminated can provide the same ecosystem services as the more diverse community. Will the removal of a given species
result in fewer total numbers of deposit feeders or will other species increase in number and/or activity to compensate? Elmgren and Hill provide evidence from the Baltic Sea that complete removal of a functional group (in this case, suspension feeders) will indeed fundamentally impact trophic linkages. Experiments to test these ideas are sorely needed. It is clear that changes in functional groups will alter the way in which ecosystem services are carried out, but any linkage between regional or local biodiversity and ecosystem services in marine systems remains to be demonstrated.

In some instances, there are likely to be keystone species that provide services disproportionate to their abundance. Clearly, the loss of these species will have a major impact. In other instances, there may be only a single species providing a service so that loss of that species will have an easily predicted outcome. Seagrasses and mangroves, for example, are often composed of only one or two species of emergent plant in a given area, and the loss of that species will have a clear effect. The loss of seagrasses to disease in Europe earlier this century had very clear effects on coastal sediments and shoreline erosion. Similarly, the loss of oysters described previously had an obvious impact. Unfortunately, we are not always able to predict key species and we are even less able to predict the cumulative impacts of multiple species that belong to a given functional group.

VII. THREATS TO SEDIMENTARY DIVERSITY

Documented extinctions in the marine environment are relatively few, and many of these are birds or mammals. However, there are certainly biodiversity losses occurring in the sedimentary fauna. Particularly in coastal areas, where human impacts have been most severe, large areas of habitat are being damaged or lost. At the very least, genetic diversity is disappearing with these habitats. Species that live in coastal habitats are often adapted to local conditions and may therefore have a specific tolerance to variables such as salinity and temperature. Once this gene pool has been lost, there is no guarantee that conspecifics from other areas will be able to invade that specialized environment. There is also increasing evidence, summarized earlier, that a significant portion of sedimentary diversity is undescribed and may subsequently be lost without our realizing it even existed. With these losses, changes in ecosystem services could also be missed or misinterpreted. It is possible to generalize that with increasing proximity to the shoreline, and therefore to human populations, the number and magnitude of threats to sedimentary diversity increase (Fig. 5). Thus, salt marsh and mangrove habitats are being reduced at a rapid rate, continental shelf habitat is being damaged in many although not all areas, and deep-sea habitats are currently being impacted in mostly isolated instances. The types of impacts are far from uniform between systems.

A. Fishing Impacts

Most of the major marine fisheries in the world occur on continental shelf areas, and they can impact sedimentary habitats in several different ways. Among the most destructive forms of fishing are dredges and trawls that are dragged across the bottom to collect benthic fishes and invertebrates. These types of fishing gear are often heavy and specifically designed to dig into sediments, and in doing so they alter the sediment composition and geochemistry, damage infaunal organisms, and destroy any benthic epifauna and plants and the habitat they create for other species, including some commercial fish. Many fisheries are also notoriously nonselective, and subsequently remove (as by-catch) large numbers of species other than those intended. By-catch, which is often 50–80% of the total catch, is often thrown overboard; most organisms discarded in this manner will be unable to recover, and pockets of decaying organisms then create localized areas that are similar to eutrophied habitat.

Fisheries also tend to remove large and abundant predators, and because these are large and abundant taxa they are species that have a major impact on their ecosystem. There are well-documented cases of shifts in species composition and trophic mode that occur as dominant predators are fished out.

In general, fisheries tend to reduce benthic diversity and homogenize habitat, although the specific means by which diversity loss occurs is only well documented in the case of physical destruction of habitat. Indirect effects, such as predator removal, require additional study.

B. Habitat Destruction, Degradation, and Shoreline Modification

Fishing activities are certainly the greatest source of degradation of bottom habitat beyond the immediate coastline, but coastal development has resulted in considerable “reclamation” of intertidal habitat ranging
from salt marshes to mangroves and mudflats. A significant portion of The Netherlands and Boston’s Logan Airport are just two examples in which shoreline has been engineered by filling or erecting dykes to allow development at the land–sea interface. Globally, large areas of marsh habitat have been modified to allow building, and large areas of tropical mangroves have been eliminated to allow development of shrimp aquaculture farms. Clearly, the loss of large areas of habitat changes local species composition and likely reduces genetic diversity of individual species. Extinctions are likely occurring in some areas.

C. Pollution and Eutrophication
The large human populations and industries that occur along coastlines throughout much of the world create problems of eutrophication, pollution, and habitat degradation. Eutrophication occurs when excess nutrients are supplied to coastal water from agricultural runoff and sewage. Phytoplankton and shallow vegetation respond to increased nutrients by growing, and as they die and sink to the bottom microbial decomposition occurs. Where large amounts of organic matter are available, microbial respiration exceeds photosynthetic production of oxygen and sediments, and bottom waters may become hypoxic or anoxic.

Pollutants are also problematic in coastal environments. Industrial and storm drain runoff deliver heavy metals, hydrocarbons, synthetic compounds (PCBs, DDT, dioxin, etc.) and countless other toxic compounds to sedimentary communities. Some of these compounds result in increased respiration, reduced reproductive output, and countless other physiological responses that may ultimately lead to death. Alternatively, the physiological stress caused by contaminants may weaken an organism and increase its susceptibility to disease. Transfer of contaminants may occur by diffusion, ingestion of contaminated particles, or feeding on contaminated organisms.

Problems of eutrophication and pollution are most intense in industrialized and densely populated coastal areas where circulation with the open ocean is limited. Thus, partially enclosed estuaries and harbors provide no mechanism for rapid dilution and removal of contaminants. Not surprisingly, the deep ocean is largely buffered from most of these impacts by the size of the continental shelf, given that most contaminants arrive as runoff of some sort from land or rivers. Nitrogenous compounds may also be delivered over broader distances by atmospheric deposition. For the coastal areas that are impacted by eutrophication and pollution, many benthic species are unable to survive under these conditions, and bottom habitats become low in diversity and dominated by weedy species that thrive under such conditions.

D. Exotic Species Introductions
Within recent years, there has been increasing concern regarding the impact that introduction of nonnative
(exotic) species can have on benthic communities. Although many people assume that the oceans are open systems and lack barriers to dispersal, there are actually many discontinuities in temperature, land barriers, depth, and ocean currents that limit the dispersal of reproductive propagules from different areas of the world. Historically, ships have provided a mechanism by which exotic species are transported to nonnative habitats. In the case of hard-substrate fauna, fouling of ship hulls allows adults to cross deep ocean basins that they could not otherwise cross because of a lack of suitable shallow habitat. However, for sedentary invaders, ship ballast water is the main culprit. A tanker may take on large volumes of ballast water in one port, sail to a new port, and dump the ballast water before loading goods. The dumped ballast water may contain adults and/or reproductive propagules that have been carried through unfavorable open-ocean habitats.

In recent years, the increasing interest in developing marine aquaculture has also helped spread nonnative species. Nonnative species are transported into new environments in which adults or reproductive propagules can escape into the wild, invade nonnative habitats, and displace native forms. An additional problem is that parasites and pathogens can be transported along with species imported for aquaculture, creating additional problems. Another transfer mechanism that is less common, but embarrassing, is the transport of organisms throughout the world for scientific study; although scientists are usually conscientious about preventing individuals or reproductive propagules from escaping into the natural environment, there are instances of accidental release.

In many instances, nonnative individuals will not successfully colonize the new environment, but particularly when conditions are stressful for native organisms (e.g., pollution and unusually low salinity or high temperatures), the exotic species may not only invade but also displace native organisms. When displacement occurs, even after conditions return to normal, the invasive species may remain the dominant taxon. San Francisco Bay and Long Island Sound are two good examples in which it is well documented that many of the dominant species today are nonnative. In some instances, the exotic and native species may be able to coexist, resulting in slightly higher biodiversity, but the complexity of species interactions can sometimes result in elimination of native species. In other words, the introduction of nonnative species is a very dangerous and unpredictable disturbance whose potential impacts are only beginning to be understood.

E. Global Climate Change

Perhaps the least easily predicted impact of human activity on sedimentary biodiversity is the long-term effects of global climate change. In a simplistic sense, species distributions might be expected to shift toward the poles to compensate for warming of the oceans. However, community response is likely to be far more complicated. Because the habitat of a given species is more complex than just temperature, it will often not be possible to simply shift distributions because suitable sediment and geomorphic features (e.g., productive banks) may be unavailable. As sea level rises, intertidal communities will not simply encroach landward because in populated areas we will prevent such encroachment to preserve expensive coastal development. Moreover, models of global warming predict potential circulation changes in the oceans, and if this indeed happens then a whole suite of ecological processes ranging from primary productivity to dispersal of reproductive propagules will be impacted. It is very likely that coastal habitats and species will be lost if significant ocean warming occurs, and it is possible that deep areas will also be impacted if these changes alter large-scale circulation patterns.

VIII. SUMMARY

Although it is likely that ocean sediments contain in excess of 500,000 species, we know little of how this diversity is maintained and how loss of biodiversity may impact the ecosystem services that benthos provide. It is critical that benthic ecologists working in sedimentary habitats ride the current wave of interest in biodiversity to improve on these deficiencies in our knowledge. Marine sedimentary systems are widespread, rich in biodiversity, and ecologically important in terms of the processes that occur within them. It is therefore reasonable to expect exciting discoveries regarding biodiversity in these habitats in the new millennium.

See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • ESTUARINE ECOSYSTEMS • EUTROPHICATION AND OLIGOTROPHICATION • INTERTIDAL ECOSYSTEMS • MARINE ECOSYSTEMS • RESOURCE EXPLOITATION, FISHERIES

Bibliography


MARKET ECONOMY AND BIODIVERSITY

R. David Simpson and Pamela Jagger
Resources for the Future

I. Biodiversity and Market Economics
II. Biodiversity in Existing Markets
III. Making Markets
IV. The Market as a Mechanism for Conserving Biodiversity

GLOSSARY

certification Procedure of determining whether a particular process is conducted in an ecologically benign fashion (also referred to as eco-labeling).

economic efficiency State of affairs in which all goods in an economy are allocated in such a way that no one can be made better off without making someone else worse off, also known as Pareto optimality.
hedonic price Price of a good expressed as a function of the particular attributes it embodies.
marginal benefits and costs Additional benefit or cost incurred in response to a small incremental change in the quantity of some variable (in the limit, the first derivative of the benefit or cost function).
market-based incentives Policy designed to change behavior by altering the price paid, or cost borne, by a person engaged in an activity that affects biodiversity.
market economy Organization of economic activity in which private individuals buy and sell goods at prices that balance supply and demand.
open access resource Resource that can be exploited by any person, regardless of the consequences of that exploitation for other would-be users; a resource for which property rights (see below) are not well defined.
perfectly competitive economy Ideal organization in which all goods are owned by perfectly rational, perfectly informed individuals, all goods can be bought and sold in markets, and no one individual can influence the prices at which transactions can take place. A perfectly competitive economy is efficient.
property right Legal and social construct under which the owner of any good cannot be deprived of it without the payment of compensation that is acceptable to her.
public good A good which, if one person consumes it, everyone else does as well, and for which one person's consumption does not diminish another's.

A MARKET ECONOMY IS A SYSTEM in which private individuals make decisions regarding how much to produce and how much to consume so as to best meet their own wants and needs. Market economies typically do not result in the most socially desirable production and consumption of all goods, as there are some instances in which the individual self-interest that drives the market economy should be tempered by a concern for the welfare of others. This is the case with biodiversity. Often one person can make herself better off by diminishing the biodiversity around her, but in doing
so she ignores the effects of this reduction on others. Ideally, people would be required to incorporate the effects of their actions on others in making their production and consumption decisions. There are, however, some activities that are consistent with both improving the welfare of those conducting them and preserving biodiversity. These market-based approaches to preserving biodiversity may be promising in some contexts, but they are not without a number of difficulties. It seems unlikely that private markets will ever be able wholly to reflect the many values to which biodiversity gives rise.

1. BIODIVERSITY AND MARKET ECONOMICS

Biodiversity confers a great many benefits on society. These benefits are most easily recognizable when embodied in a commodity of market value. In such instances, social benefits can be proxied by the prices people are willing to pay. Many other benefits arising from biodiversity are not bought and sold in markets. The market values of biodiversity, in combination with less tangible values that cannot easily be measured by market prices, represent the total value of biodiversity to society. The challenge to the natural or social scientist, policymaker, or resource manager is to recognize both the market and the nonmarket values associated with biodiversity and to direct social policy accordingly.

The full range of biodiversity values is discussed elsewhere in this Encyclopedia. The task here is to perform a more detailed analysis of the relationship between biodiversity and the market economy. To begin with, the phrase “market economy” can be used in two very different ways. The first is the textbook abstraction more often referred to in the economics literature as “perfectly competitive general equilibrium,” or a “perfectly competitive economy.” In this sense a free market economy is an idealized state in which all goods and services are owned and can change hands in voluntary transactions. Because only some goods and services are subject to private ownership, this “free market economy” is distinctly not efficient. Those who suffer as a result of the economic choices of others are not afforded an opportunity either to offer payment to dissuade others from harming them or to demand compensation for such harm.

We will use the phrase “market economy” in the second sense in this article. In particular, we will concentrate on the relationships between biodiversity and a real-world economy that is clearly imperfect in the ways in which it produces and allocates some goods and services. In particular, we will ask to what extent “the market” can be relied on to provide incentives for the preservation of biodiversity, and to what extent a more active public policy may be required.

It is important to note that “active public policy” may be a more effective and efficient thing in theory than it is in practice. The “benevolent social planner” who appears in economic analyses to make such policy is every bit as much a deus ex machina as is the idealized perfectly competitive market itself. Choices must be made between relying on an imperfect market economy on one hand and, on the other, government officials who can only be expected to advance the general welfare imperfectly. Which option is better is not clear in general, and is likely to depend on particular circumstances. Despite our focus on the existing market economy and its imperfections, it can still be useful to compare it with the benchmark provided by the textbook abstraction of the perfectly competitive economy. It is revealing to see how and why actual economies fall short of this ideal. In the following section we discuss the conditions that characterize a perfectly competitive economy, and why these conditions are unlikely to be met with respect to biodiversity. We turn next to a discussion of why and how biodiversity can be undervalued in an imperfect economy. Following this discussion, we consider some of the ways in which some biodiversity values are reflected even in imperfectly organized real-world economies. Next, we consider steps that could be taken to rectify some of the “market failures” that result in the
undervaluation of biodiversity, and review some evidence on attempts that have been made to correct these failures. We conclude by noting that while markets should be relied on to preserve biodiversity to the extent to which they are able, private markets will probably not soon, if ever, provide completely adequate incentives for that purpose.

A. Biodiversity in an Ideal Economy

A good place to begin is to ask what conditions would characterize the amount of biodiversity preserved in an economy that recognizes the total value of all aspects of biodiversity. An important principle in economics is that the maximization of total value is achieved by meeting conditions on marginal values. The marginal value of something is its incremental contribution to total value (in mathematical models, it is the first partial derivative of total value with respect to the good in question). In a perfectly competitive economy, all goods would be provided in such a way as to afford equal marginal value. The intuition behind this notion is straightforward. Finite resources place constraints on what we as a society can have and use. An optimal allocation of goods would be one in which we cannot have a little more of one thing without losing a quantity of something else that is at least as valuable. If one of the “goods” in question is the amount of biodiversity we maintain, the value an ideal market places on it would be equal to its marginal contribution to welfare. This would, in turn, be equal to the marginal contribution afforded by any of the other things that we could have instead of that incremental amount of biodiversity. In short, the value of biodiversity—is its price in an ideal economy—would, as with every other good, be set in accordance with its marginal contribution to the general welfare. The marginal contribution made by biodiversity to welfare might be realized directly, in enjoyment of the preservation of biodiversity for its own sake, or indirectly, via its contribution to the generation of other goods and services.

B. Practical Problems in Preserving Biodiversity in Markets

Biodiversity differs in an important respect from bread, automobiles, or other goods that might be better candidates for allocation in an unregulated economy. Biodiversity is (at least in some aspects) an example of what is known in economics as a public good; it is something that, if provided to one person, is necessarily available to others. Inasmuch as many people can benefit from the preservation of the same biodiversity, its marginal contribution to welfare should properly be calculated as the sum of the incremental contributions to each affected person’s welfare. Closely related to this consideration, however, is the fact that no one can be excluded from the enjoyment of biodiversity. If a rare species is saved from extinction, all those people who value its continued existence benefit, regardless of whether they have made any sacrifice themselves to maintain it. In short, if one person pays for the conservation of biodiversity, many benefit. This is what is known as the “free rider” problem in economics.

The existence of public goods typically means that market economies do not efficiently allocate goods. This means that the real-world economy, in contrast to the perfectly competitive ideal, provides insufficient incentives for preservation. In theory, one could create appropriate incentives for the provision of public goods through elaborate rules for the punishment of free riders and for revealing individuals’ true attitudes toward resources. Given the formidable practical impediments to applying such fixes to a problem as complicated as the preservation of biodiversity, however, it seems unlikely that private markets could ever fully handle the problem in practice.

It should be noted, if only for completeness, that the preceding passage assumes that there is a problem. If the nonmarket aspects of biodiversity prove to be negligible, then by definition there would be little problem with allocating what we might label “biodiversity-related” goods in markets. One should also bear in mind what was said earlier concerning the difficulties inherent in correcting “market failures” with public policy options that are themselves inadequate or poorly designed. While there is considerable debate about the extent of the problem, a conclusion that the values to which biodiversity gives rise are negligible is surely premature. Current evidence suggests that the issue is not so much whether biodiversity conservation is a challenge facing society but rather, given that it is a significant challenge, how it compares in relation to other social and economic issues. We will continue to assume that this is, in fact, an important problem.

The observation that markets do not allocate public goods efficiently is not tantamount to saying that markets cannot, and for that matter do not, affect the preservation of biodiversity. In the following section we consider some of the ways in which biodiversity enters into market transactions. The focus will be on determining the extent to which market forces may be counted on to alleviate the biodiversity conservation problem.
II. BIODIVERSITY IN EXISTING MARKETS

It is important to avoid a logical fallacy in thinking about the relationship between biodiversity and the market economy. Private markets tend to undersupply biodiversity because of the free rider problem. This does not necessarily imply that data from market transactions are irrelevant in estimating the benefits that biodiversity provides society. To give an example, a farmer may not be required to pay for the services that bees from an adjoining forest provide by pollinating his crops, but the value of those services is reflected in the earnings the farmer receives from his crops and, by extension, in the market value of the farmer's land. Thus, the social value of the bees is not incorporated in the forest landowner's decision to fell her trees, but is reflected in the change in earnings the farmer experiences as a result.

This section considers the importance of biodiversity in a number of sectors of the economy. Some values of biodiversity are reflected in a number of market transactions. The contributions of biodiversity to many economic activities are the subject of dispute among experts, however, and in the opinion of some of those experts they are negligible. From this observation the reader should draw two conclusions. The first is simply a restatement of what we have just said: Experts disagree as to how effective market incentives can be for preserving biodiversity.

The second conclusion to be drawn from an overview of the contributions of biodiversity in the market economy is that, whether the accurately and fully measured contributions of biodiversity to the observable and quantifiable output of the economy prove to be large or small, they may still provide something less than complete guidance to policymakers. It would still be important to consider the incentives to preserve biodiversity for ecological, ethical, aesthetic, or even spiritual reasons. Few economists would go so far as to say that everything of value is recorded in the measured economy.

Let us now consider some of the ways in which biodiversity enters into the measured economy.

A. Agriculture

The market for agricultural produce provides us with an example of biodiversity integrated into the market economy. All agricultural products are derived from what were, at one time, wild organisms. Today, though, agricultural productions relies on only a relatively small handful of species. Wilson (1992), for example, noted that the vast majority of human nutritional needs are met by only about 20 species. This is despite the fact that many thousands of species are edible and potentially cultivable.

The observation that this base is narrow is not necessarily a compelling argument that diversity is undervalued; however. Specialization in particular crops would not occur if there were not some savings in production or marketing costs. In addition, consumer preferences with respect to variety, and how these preferences vary geographically and culturally, also factor into the issue of the degree of specialization in agricultural production. The optimal degree of product diversity is an open question in economics more generally. About the only thing that can be said definitively is that the answer depends on the specifics of the situation, and therefore that different results will obtain under different circumstances.

There are, however, some reasons to suppose that existing markets for agricultural products do not generate sufficient variety. One such reason has to do with the inability of producers of differentiated products to realize the full benefits from their provision. Consumers receive what is called in economics “consumer surplus” from the purchase of products. If I only have to pay $1 for a product that I would have been willing to pay $2 for, I receive $1 = $2 − $1 in consumer surplus from my purchase. Because this surplus cannot, by definition, be appropriated by the seller of the product, the seller does not consider it in making the decision of whether or not to introduce a new product.

Another way in which unregulated markets may offer too limited an agricultural product line concerns the spread of pests and disease. Every farmer, by making his decision of what to plant, affects not only what is going to be available to consumers, but also what may befall other farmers. Mile upon mile of uninterrupted fields of nearly identical crops may comprise a low-resistance passageway for the spread of pests or disease. A farmer who plants a different crop might expect to achieve greater profits in the event that others’ crops are wiped out, but he or she would probably have inadequate incentives to provide such a pest-break for several reasons. First, as noted earlier, the farmer would not receive the full benefits of his contribution, as consumers would also receive some additional surplus. Second, by providing the barrier to the spread of pests, the farmer is not only benefiting himself and consumers, but also neighboring farmers who would be less vulnerable to the spread of pests.
Finally, there are circumstances in which political interference with what might otherwise be free markets prevents desirable outcomes. Farmers would have greater incentives to plant different, disease-resistant varieties to the extent that they believe they would be able to achieve high profits by changing high prices should their crops be successful while their neighbors crops failed. Yet ample historical evidence suggests that governments generally step in to prevent such “price gouging.” It is a situation in which de facto restrictions on profiting after the event of a catastrophic disturbance create perverse incentives for protecting against such occurrences in the first place (we are grateful to Professor Brian Wright of the University of California for making this point).

We might also remark in passing that while agricultural markets represent an instance in which one component of biodiversity is integrated into the economy, agricultural practices themselves affect the status of biodiversity more generally. Each farmer may be better off over the short term by applying more pesticides and fertilizers to her fields, but these actions affect adjoining, and sometimes even distant, ecosystems and the diversity they sustain. Perhaps most importantly, the sheer scale of agriculture affects patterns of land use and conversion. More land in agriculture typically means less land in forests, wetlands, grasslands, or other habitats more conducive to the support of indigenous biodiversity. As we will see, this reduction in surrounding biodiversity may also have an effect on the productivity of agriculture.

B. The Harvest of Wild Organisms

Settled agriculture might be regarded as the culmination of a process that began with hunting and gathering. While the vast majority of food products and fiber of organic origin produced today comes from farms, hunting, gathering, and particularly fishing remain important in many areas. Unlike the case of settled agriculture, where many management practices (e.g., soil conservation) are intended to maintain production into the indefinite future, the harvest of organisms from their indigenous habitats can present a threat to the continued existence of the targeted population or species.

This can be true for two reasons. The first is that there is often a market failure that prevents private agents working in their own self-interest from behaving efficiently. Many wild populations are subject to open access. By this we mean that any or all of a number of individuals are able to harvest organisms without having to compensate the others for the privilege of doing so. This leads to a situation that has come to be known as the “tragedy of the commons” (Hardin, 1968). If a resource base is subject to open access, benefits from its exploitation will be dissipated by excessive entry. So long as it is profitable to engage in harvesting wild organisms, more and more people will do so until there is no further advantage to be gained from additional exploitation.

This outcome is perhaps most clearly seen in the case of fishing in international waters. Since no one owns the open ocean—and hence fishers from any nation cannot be excluded from fishing on the high seas by those of any other nation—international exploitation of fish stocks has often proved excessive. There are also terrestrial examples. Animals have been hunted to the point of extinction in various places and times. Similarly, in situations where property rights are nonexistent, poorly enforced, or subject to frequent abrogation, forests may be felled at excessive rates.

The solution to the “tragedy of the commons” is either to define property rights in the resources—make it illegal for me to catch a fish or shoot an elephant that “belongs” to you—or to achieve the same end by regulating timing, location, or other attributes of harvest. Regrettably, this is often more easily said than done. It can be difficult to define ownership in fugitive resources. If animals are not fenced in, property rights in real estate confer rights to the animals resident on it only during such periods as they are actually on the property. Extending rights to the animals themselves is difficult when individuals cannot be readily identified. The alternatives of selling rights to entire populations or of designing schemes for defining the ownership of particular animals may not be easily practicable. Despite these practical difficulties, some promising experiments are under way with assigning such property rights (see Section III.A). Regulatory approaches are also problematic. Outright prohibitions can be difficult to enforce. More flexible approaches, such as tradable fishing quotas, can result in unintended consequences, such as discarding “bycatch” of species that the fisher does not have the right to catch.

The “tragedy of the commons” gives rise to both static and dynamic problems. From a contemporaneous viewpoint, the more I harvest today, the less there is for you to harvest today. It is also true, however, that the more I harvest today, the less there will be for either you or me to harvest tomorrow. Now if I do not trust you to leave anything for me to harvest tomorrow, I will take as much as I possibly can today. If everyone behaves this way, the target population may be extirpated.
While imposing property rights may mitigate the "tragedy of the commons," the assignment of property rights does not assure that whatever harvesting practices are then adopted will be sustainable. It can be shown (Clark, 1991) that the optimal path for exploiting a biological resource depends on the rate of interest. In this analysis, the target population is regarded as an asset, which must compete with other assets in the economy. The hypothetical owner of, say, a fishery, would manage it so as to equate the rate of return on owning fish with that from holding an alternative asset, such as a bank account. It can be shown that under some circumstances the profit-maximizing strategy of a private owner would be to completely extinguish the fishery and invest the proceeds of this effort in an interest-bearing bank account.

This finding is, understandably, regarded as morally repugnant by many. The fault lies not with the logic driving the conclusions, but with the assumptions underlying the analysis. To suppose that a species should be extinguished if its rate of intrinsic growth does not compare favorably with the rate paid on Treasury Bills presumes that only the profits for those harvesting the species matter. Defining the problem more broadly to include the benefits to society of maintaining a species may yield a different answer. Thus we see that even a relatively well-functioning "free market economy," one in which the rights to harvest organisms are completely defined and allocated, will not achieve truly optimal outcomes unless all those who care about the existence of biodiversity per se are afforded opportunities to purchase and retire such rights of harvest.

Before leaving the subject of hunting and gathering, we might briefly consider the special topic of its conduct in less-developed countries. Some research has suggested that biodiversity-rich natural habitats can be considerably more valuable as sources for the sustainable collection of non timber forest products (NTFPs) than they would be if clear-cut and converted to pasture. Despite these findings that land would be more valuable in a more natural state, land is, in fact, being clear-cut and converted to pasture in many of the areas studied. There has been a great deal of criticism, on both conceptual and empirical grounds, of such studies. Moreover, some commentators have expressed doubts that the collection of forest products in commercial quantities can be truly "sustainable," in the sense of preserving a natural, diverse forest ecosystem.

This is not to say that profitable, sustainable NTFP collection is impossible everywhere or that it does not occur in some areas. In most parts of the world, however, technological progress and increasing scarcity of land motivate more concentrated production via ex situ cultivation. As with the other strategies for the conservation of biodiversity in developing countries, the important considerations in evaluating the efficacy of NTFP collection as a conservation incentive concern measuring its true profitability and assuring that it is, in fact, conducted in such a way as to preserve biodiversity-rich habitats.

C. Biodiversity Prospecting

Perhaps the area in which the market economy has come closest to pricing biodiversity per se, as opposed to elements of it, is in "biodiversity prospecting." Biodiversity prospecting is the search among wild organisms, both plants and animals, for new products of industrial, agricultural, or, especially, pharmaceutical value. Researchers from a number of industries have undertaken searches in many nations and ecosystems for valuable products. In some instances cash payments have been made for rights of access to indigenous biodiversity, while in other cases search is undertaken under contracts that specify royalties to be paid in the event that commercial products are developed from natural sources.

Despite the fact that such compensation has been paid or promised, the potential of these market forces to motivate conservation remains unclear. This is so for several reasons. First, it is not clear that the property rights necessary to motivate truly remunerative payments have yet been established. Some commentators have, in fact, suggested that fuller delineation of such property rights will lead to greater payments for access to biodiversity and, by extension, for its conservation. On the other hand, most existing arrangements offer compensation to individuals or organizations that do more than simply provide access to samples for testing. In many instances, the initial steps of collection, identification, preparation, and sometimes even preliminary testing are conducted by the sample seller. It is unclear, then, what share of observed payments and prospective royalties are allocated for the natural samples themselves as opposed to the labor and other inputs involved in their processing. Finally, many existing agreements specify royalties in the event of discovery. The probability with which discoveries will be made is unknown, however, and in many cases the rates at which royalties would be paid are not made public. Thus, it is impossible to infer the values attributed to the resources themselves.

For these reasons, valuation of genetic diversity remains highly controversial. Absent clear evidence from
market transactions, attempts to value biodiversity for its use in the development of new products have been based on indirect inferences. One thing is obvious, however. The aggregate value of biodiversity for its use in new product development is astronomical. We would not have food, nor many of the other necessities of life, were it not for the natural organisms from which they are derived. Even the value of extant biodiversity in the development of new products and the improvement of existing ones can be quite substantial.

Economic value, however, is determined by “value on the margin.” Whatever philosophical reservations one may have concerning this principle, it is the way in which businesses make decisions. Hence the value that private companies assign to biodiversity will be determined by the marginal contribution that additional biodiversity makes to their profitability. This marginal contribution is composed of two elements. The first concerns the expected contribution of an incremental component of biodiversity if it proves to be the best source of a new or improved product. The second concerns the probability with which other potential sources will prove to provide better leads for new product development. Simpson, Sedjo, and Reid (1996) argue that this marginal value is negligible. Thus the “free market” value of biodiversity with respect to the search for new products is neither a compelling reason for preservation nor a plentiful source of funds for habitat preservation (they also emphasize that this conclusion says nothing about the value of biodiversity in any of its myriad other uses and aspects). This analysis has been revisited by Rausser and Small (in press), who argue that earnings could be higher in exceptional circumstances. There seems to be little disagreement, however, that appreciable earnings could only be realized in such exceptional circumstances. The reasons for this are related to the difficulties in defining and measuring biodiversity, and in separating its value from the values of undeveloped habitat more generally. Empirical economists must rely on natural experiments for the generation of their data. Controlled experiments, at least of a phenomenon as complicated as the valuation of biodiversity, cannot be conducted in the laboratory. Only the extreme points of the observed distribution provide unambiguous data. At one end of the spectrum, market economies typically place the highest value on land in densely populated, largely “unnatural” areas. At the other extreme, lands containing the greatest diversity of (relatively large, at least) organisms are often virtually valueless, as reflected in market prices. In many instances biodiversity continues to thrive in habitats far enough distant from concentrations of population and industry so as not to have been brought within the modern market economy.

Yet between these extremes it is undeniable that proximity to nature does contribute to the values of some properties. The city lot adjoining the park is the canonical example. However, one must be clear what it is one is considering. The park in the city, often laid out on land cleared decades or centuries earlier, cannot compete in diversity to truly pristine habitat. The evidence that markets value open space is considerably stronger than that they value diversity per se.

D. Biodiversity Values Embedded in Property Prices

Economists often suppose that the price of some property reflects the advantages of its location. Statistical techniques exist for inferring the “hedonic prices” implicit in such properties. For example, if a hectare of land in a city sells for $10,000, while another, similar in all ways save that it adjoins an area of parkland, sells for $15,000, the value of proximity to natural amenities may be inferred to be $5000. This technique of hedonic pricing has been used in a variety of contexts. In addition to its common use in the valuation of differentiated products, hedonic pricing studies have been employed to measure the market value of environmental amenities, such as clean air. There have been few hedonic pricing studies that attempted to measure the market value of proximity to biodiversity, however.

E. Tourism and Conservation Incentives

One way in which biodiversity may be reflected in the market economy is through its impact on patterns of, and expenditures on, tourism. Earnings arising from rain forest excursions, river rafting, photo safaris, hunting, and other such nature-based tourism options have proved more lucrative in many areas than are alternative, less ecologically benign activities. Travel to destinations that derive their appeal from their proximity to living assets but that does not degrade such assets (one might define this as “ecotourism,” although the precise definition of the term is a matter of contention) is a promising option in many parts of the world. Such nature-based tourism is not without its pitfalls, however. Encouraging tourism in ecologically sensitive areas can have unintended consequences: hotel, road, and trail construction, the physical passage of tourists.
through sensitive habitats, and the depletion of forests for firewood are examples. Making a comfortable and attractive destination for tourists can be a different thing than preserving an area’s full biodiversity.

The relationship of even innocuous ecotourism to the market economy may be complicated as well. If there were sufficient money to be made from the operation of tourism facilities or related activities, one would expect that more agents in the market economy would identify and exploit the opportunities for doing so. It is problematic, then, that some of the impetus behind nature-based tourism comes from government, international, or private donors. If there is not money to be made in tourism, one has to ask if conservation-related funds might not be better spent on other activities. Moreover, a number of analyses suggest that a disproportionate share of the expenditures related to tourism are often received by the providers of transportation and other services rather than by those who “maintain” the natural habitats of the destinations, commonly assumed to be the local populations who forgo other use of the resources. Incentives for conservation at the local level may, then, be attenuated.

A fundamental principle of economics is that value is related to scarcity. As was the case with biodiversity prospecting, money might be made at truly spectacular and unique locations. Responsibly conducted tourism in such places is clearly desirable. The truly spectacular and unique is inherently scarce, however. Thus the potential of tourism to conserve biodiversity must also be limited. Tourists can choose from a variety of spectacular destinations: there are rain forests, coral reefs, high mountains, and arid deserts on most of the continents. Inasmuch as all of these destinations are in competition with one another, this competition will reduce willingness to pay for travel to a particular destination. Moreover, there are any number of potential tourist destinations that could be developed. We return again to our earlier points. Private incentives may be adequate for the establishment of tourism locations where they are economically justified, but more effective conservation strategies might be pursued when ecotourism is not justified.

F. The Diffuse Benefits of Biodiversity

The benefits of biodiversity may be widespread and diffuse, and for that reason not easily identified in the market prices of particular properties. This begs the question as to whether such values can be identified at all within the prices of the existing market economy. The answer to the question depends on the nature of the values.

On one hand, a host of services are provided by diverse natural ecosystems that, while they themselves are not bought and sold in market transactions, are essential inputs into things that are. Ecosystems purify water, cycle nutrients, capture and break down pollutants, harbor pollinators, moderate local and global climate, reduce the frequency and intensity of flooding, retard erosion, and provide many other services. In the case of most of these services, however, natural ecosystems provide these services to some particular set of properties. If, for example, the supply of ecosystem services makes particular properties more agriculturally productive, then the measurable product of the agricultural properties should reflect the contributions of the ecosystems. Although it may be difficult to determine which properties benefit by how much, total effects would appear in aggregate statistics. Thus national income accounting would reflect, if imprecisely and indirectly, the services of natural ecosystems. On this score, one can read the evidence in several ways. The advanced industrial economies that have achieved the greatest measured economic performance are typically not the best endowed with biodiversity. Conversely, the less-developed countries are often rich in biodiversity and other natural resources. To suppose that the historical “success” of the industrialized nations indicates the insignificance of biodiversity to measurable economic performance may be premature, however. First, such a statement may not be true in a broader historical—or even prehistorical—and geographical context. Jared Diamond (1997) has argued that the achievements of the wealthiest nations are attributable in part to their inhabitation of environments with greater exposure to broader arrays of organisms. Second, some would argue that the apparently wealthier nations are incurring an unrecognized debt by straining their natural asset base to a point at which decline may be inevitable or recovery long and expensive. Third, countries differ in any number of respects. The advanced industrial economies differ from the less-developed countries in the education of their populations, the vintage of their technologies, the nature of their institutions, and in innumerable other respects over and above the state of their biological resources. To say that the nations of the “first world” have achieved their status despite the initial paucity or subsequent degradation of their biological resources is not so say that their performance might not have been better still had their economies grown less profligately.

Even if one were to claim that the leading industrial
MARKET ECONOMY AND BIODIVERSITY

MARKET ECONOMY AND BIODIVERSITY


economies have gotten where they are today either because of or despite their relative lack of biodiversity, and even if one were to project that this leadership would continue, it would not provide a satisfying answer concerning the importance of biodiversity to society. There may be very important things that are not, and may never be, traded in markets.

Foremost among these important things are those that go under the rubric of nonuse, or existence, values. These are benefits that are wholly unrelated to any actual or potential current or future consumption. It is postulated that people derive satisfaction merely from the continued existence of some elements of the natural world, regardless of whether they, their descendants, or anyone else will ever benefit directly from such elements. Philosophical debate continues on the possibility and magnitude of values that are, by construction, totally divorced from the price of any good bought or sold in markets. It seems reasonable to suppose, however, that the relationship between some of the things people care about and market transaction is so tenuous as to make inferring values from market price data impossible. At best, then, the market economy can provide only a partial guide to the social benefits of biodiversity.

III. MAKING MARKETS

The conclusion to the previous section notwithstanding, the incentives provided by the market economy should be aligned as closely as possible with the social and economic values afforded by biodiversity. This can be done in two ways. First, while the establishment of property rights in wild organisms and their habitats cannot be expected to solve all problems relating to the preservation of biodiversity, it can at least amplify the incentives now provided by private markets. Second, “market-based incentives” in biodiversity preservation can be established by government action, and may achieve the socially desirable level of biodiversity preservation at less cost to society than would be incurred under more direct regulation. This section treats these two themes in turn.

A. The Extension of Property Rights and the Preservation of Biodiversity

We noted earlier that the extension of property rights is not a panacea for the preservation of biodiversity. Vesting an owner with some property rights encourages her to maximize the value she can obtain from the resource, but unless ownership is assigned to absolutely all the things people care about, inefficient choices may be made in resource uses. For example, providing legal title to a forest may remove the incentive to cut it down before somebody else does, but does not mean that the forest will remain standing if its owner would make more money by converting the land to residential or agricultural use.

Traditional societies have often been able to avoid the “tragedy of the commons”—the tendency of all users of a resource to overexploit it—by developing societal rules for sharing a resource. A community in which a fishery, game population, or forest has managed to survive for centuries is probably one in which rules have evolved over time for allocating rights to exploit the resource within the community. This may take the form of de jure ownership of certain areas, or of restrictions on the amount of the given resource that any particular individual is allowed to harvest. As traditional social structures are supplanted, however, establishment of de jure property rights may again result in sustainable management being undertaken.

Some recent developments in property rights and conservation policy can be interpreted as efforts to replicate the procedures that traditional societies commonly establish. The Community Areas Management Programme for Indigenous Resources (CAMPFIRE) in Zimbabwe gives local people the right to manage the wildlife that crosses their land. CAMPFIRE illustrates an interesting paradox that arises in many biodiversity management situations. If local people are not given the right to manage herds—including the right to benefit from hunting—they may have no incentive to preserve endangered species at all. In fact, to the extent that large animals such as elephants can be extremely destructive of farmer’s fields, local people would otherwise have a clear incentive to eradicate them. By providing an incentive to cull animals in a sustainable fashion, CAMPFIRE removes the incentive to eradicate the entire population. This is not to say, of course, that a program like CAMPFIRE will always be successful—CAMPFIRE itself has received some criticism. However, this case may demonstrate that “exceptions prove the rule.” Areas in which such programs are not successful are often those in which community involvement and cohesion are low.

In the United States, Defenders of Wildlife, a non-profit organization devoted to conservation, has initiated programs to encourage the reintroduction of wolf populations. The “Wolf Reward Program” makes payments to landowners who can demonstrate that wolves...
have reproduced on their properties. A related program compensates livestock owners for losses from predation by wolves. In each instance, Defenders of Wildlife encourages protection of the wolf population by making payments for occurrences that had previously not been the subject of market transactions: the reproduction of wolves in the first case, the loss of livestock in the second.

Efforts such as the Defenders of Wildlife programs and CAMPFIRE may be the prototypes for more widespread future imitation. One might expect to see increasing extension and enforcement of property rights as time passes. The economic theory of property rights holds that property rights come into existence when the benefits of their definition exceed the costs of their enforcement. When populations are large it makes little sense to sell the right to cull individuals; because there is little scarcity, such rights would sell for less than the cost of instigating and monitoring the transaction. As populations grow smaller, however, the benefits from allowing hunting may grow very large.

B. Market-Based Incentives

Traditional regulatory approaches to environmental matters have often taken the form of so-called “command and control” instruments. Under these approaches a regulator “commands” people to do certain things (e.g., reduce pollution or preserve wetlands) and “controls” them by imposing civil or criminal penalties. National and international regulation pertaining to biodiversity has generally taken this form. The Convention on International Trade in Endangered Species (CITES) outlaws trade in certain listed species. The United States Endangered Species Act prohibits harming of listed species.

Command and control regulations are generally not favored by economists. Their argument consists of two parts. The first is that regulations typically should not totally forbid activities. The depletion of biodiversity is clearly not desirable per se, but our lives would indeed be “nasty, brutish, and short” if our ancestors had not reconfigured the natural world to at least some extent. Given this fact, the second component of the argument against command and control regulation is that the regulatory burden should fall most heavily on those individuals or firms who can most easily bear it. This is exactly what “market-based incentives” are intended to do. If some biodiversity loss is believed to be the necessary consequence of economic development, we should at least design programs to get the most development in exchange for the least loss.

Some such programs are now in place for biodiversity-related matters. In the United States, for example, private parties can buy and sell wetlands and the obligations to establish wetlands. A number of suggestions have been made for modifying the Endangered Species Act by introducing market-based incentives. Perhaps the most prominent of these suggestions concern “tradable development rights” (TDRs). Under such plans a certain number of permits for the conversion of endangered species habitat would be established. Any conversion of habitat would require the purchase of a permit, and with it the obligation on the permit seller’s part to maintain a specified area of land as habitat for the endangered species. The advantage of such a plan would be that it would allocate the preservation obligation in the most cost-effective fashion. Owners of endangered species habitat with high commercial value would buy permits from those whose lands have low commercial value, and the species would be preserved at the lowest overall cost.

Of course, real-world complications can generate more controversy than is suggested by this thumbnail sketch. The problem with any program that involves a trade-off of one unit of habitat for another (or, perhaps in the future, one unit of biodiversity for another) lies in defining the “unit.” A hectare of land in one location may be “worth” more or less than that in another with respect to its capacity to support particular species, the issue becomes one of which species to protect. Until an operational consensus emerges as to how to measure and trade off biodiversity, all but the most rudimentary tradable development right programs will be impossible to implement.

Another way in which market-based incentives for biodiversity preservation can be established is through “certification” or “eco-labeling.” In these programs certain products often associated with biodiversity loss (most prominently timber) are certified to have been produced in such a fashion as to minimize biodiversity loss (ideally, on sustainable plantations so that no further clearing of natural habitat would be required). Because the process of growing, harvesting, and processing products is typically more expensive if conducted sustainably, certification programs presume that consumers are willing to pay higher prices for certified products. Moreover, an effective certification program must necessarily be one in which consumers are able to distinguish between certified and uncertified products, as well as between truthfully certified and untruth-
fully certified products. The more effort that must be expended in order to credibly certify sustainable production, of course, the more expensive will be the final product. Thus, there is some concern that certification programs may not yet be financially feasible.

Nevertheless, certification programs have begun under the auspices of conservation organizations such as the Rainforest Alliance and industry groups such as the Forest Stewardship Council, and by the beginning of 1999 it was estimated that some fifteen million hectares of forestlands were covered under independent certification programs. Agreements by major purchasers, such as home improvement chains, as well as smaller “alternative” outlets, to purchase certified wood may also indicate an increasing willingness on the part of distributors, as well as producers and consumers, to participate in certification programs.

The ultimate success of market-based incentive programs will require both additional information on the biodiversity trade-offs that society faces and the emergence of a social consensus that these trade-offs should be resolved in a fashion more conducive to biodiversity preservation. This is not to say that these problems constitute an insurmountable impediment to market-based incentive programs. A generation ago there was little support for market-based incentive programs for the reduction of industrial emissions, but in recent years programs such as sulfur dioxide trading have reduced the costs of environmental compliance by millions of dollars. We are likely to see increased experimentation with, and eventually reliance upon, market-based incentives for biodiversity preservation in the future.

IV. THE MARKET AS A MECHANISM FOR CONSERVING BIODIVERSITY

We have seen that biodiversity enters into a number of aspects of the free market economy, even when the phrase “free market economy” is used to describe our existing, imperfect mechanism for deciding what to produce, what to consume, and what to preserve. The services that biodiversity provides are partially reflected in agricultural production, land prices, and aggregate measured economic production. Biodiversity can also generate values via biodiversity prospecting, nontimber forest product collection, and tourism, although expectations for these activities must be measured.

The real-world “free market economy” still falls short of how an ideal free-market economy, the perfectly competitive economy of textbook abstraction, would value biodiversity. We can approximate that ideal somewhat more closely by making broader use of market-based incentives. We must also realize that an ideal economy cannot be achieved in a real world of imperfect institutions and behaviors. Given these constraints, the pivotal question of how much biodiversity we can and should save will likely remain unresolved for many years to come.

See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • BIOPROSPECTING • COMMONS, THEORY AND CONCEPT OF • ECONOMIC VALUE OF BIODIVERSITY, MEASUREMENTS OF • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • PROPERTY RIGHTS AND BIODIVERSITY • RESOURCE EXPLOITATION, FISHERIES • TOURISM, ROLE OF

Bibliography


MASS EXTINCTIONS,
CONCEPT OF

J. John Sepkoski, Jr.†
University of Chicago

I. History of the Concept
II. Models of Mass Extinction
III. Interpreting Data From the Fossil Record
IV. Magnitudes of Mass Extinctions
V. Hypothesis of Periodicity
VI. The Kill Curve and Self-Organized Criticality
VII. Selectivity of Mass Extinction: Victims and Survivors
VIII. Recoveries from Mass Extinction
IX. The Modern Biodiversity Crisis
X. Summary

GLOSSARY

benthos Organisms living on or in sediment, below the water.

extinction The disappearance of a species upon death of its last surviving individual. In the fossil record, this is treated as the last fossil occurrence of individuals of a species.

foraminifera An order of animal-like protists, many of which secrete calcareous skeletons ("tests").

mass extinction The simultaneous extinction of a disproportionate number of species over timescales of 10^5 to 10^6 years resulting in loss of biodiversity.

phanerozoic The geological interval of abundant animal fossils, beginning approximately 545 years ago.

stratigraphic section An outcrop of rock (with fossils in this case) or a drilled core. The term can also refer to a composite for a region in which fossil ranges and stratigraphic events have been summarized into a synthesized rock column.

tetrapod vertebrates Vertebrate animals with four limbs (or vertebrates that have evolved from such animals, such as snakes).

MASS EXTINCTION refers to the disappearance of large numbers of organisms over relatively short geologic spans of time. The result is diminished biodiversity, which can take millions of years to recover, depending on the magnitude of the extinction event. This chapter presents topics related to this concept, including its history, current measurements of the magnitude and timing of mass extinctions, and consequences for the recovering biota.

I. HISTORY OF THE CONCEPT

The concept of extinction of species goes back at least several centuries. The extirpation of aurochs (wild relatives of cattle) and disappearance of lions from Europe were well known in the era of enlightenment and ascribed to human interference. The fact that species could become extinct from nonhuman causes was promoted by Cuvier at the end of the 18th century through his exquisitely detailed studies of mammalian fossils of the Paris Basin. His arguments were not accepted by...
all intellectuals at the time, and, in fact, Thomas Jefferson, the third president of the United States, doubted species could disappear before humans; he assigned Lewis and Clark a secondary mission in their explorations of the American northwest to search for living mammoths and mastodons.

The division of sequences of sedimentary rock into geologic systems by British geologists and paleontologists in the first half of the 19th century reflected a concept of major changes in marine faunas between these still-used time periods. But the first quantitative depiction of mass extinctions—major declines in biodiversity followed by recovery—appears to be Phillips’s (1860) count of known numbers of fossil species and interpretive graphing of massive drops in diversity between the Paleozoic and Mesozoic eras and the Mesozoic and Cenozoic eras (terms he coined; see Fig. 1).

The study of mass extinctions rested largely in limbo from Phillips’s pioneering work into the mid-20th century. This was perhaps because of emphasis on documenting evolutionary continuity in the fossil record and an assumption of substantive uniformitarianism, inherited from Lyell. However, with accumulation of paleontological data, the greatest of all Phanerozoic mass extinctions—the end-Permian, or “Permo-Triassic,” event—could not remain unnoticed. Schindewolf (1963) wrote a seminal paper discussing this event and invoking lethal radiation from an extraterrestrial catastrophe of a nearby supernova explosion. In response, Newell (1967) carefully counted fossil taxa (mostly described families) and argued that there were at least five events of mass extinction in addition to the Permo-Triassic. These papers set the stage for modern studies: examining detailed biostratigraphic data on local species disappearance and global compilations of taxonomic ranges.

Despite the contributions of Schindewolf and Newell, work on mass extinctions remained largely a “cottage industry” among paleontologists until 1980. Workers would examine one of Newell’s events (usually in isolation of others) and posit some associated physical event as the cause, such as fall of sea level, or invent ad hoc hypotheses, such as heavy metal poisoning in the oceans as a result of mountain building.

Maturation of the study of mass extinction came with the bold hypothesis of Alvarez et al. (1980) that the Mesozoic-Cenozoic event, recognized 120 years before by Phillips, was caused by impact of a 10-km meteorite. The initial evidence of Alvarez and coworkers was concentration of the rare terrestrial element, iridium, at solar-system abundances in a clay layer at the Cretaceous-Tertiary boundary. But the hypothesis implied other testable questions: (a) Is there additional physical evidence of an impact at the Cretaceous-Tertiary boundary and (b) Is the abruptness of biological extinctions at the event consistent with a catastrophe induced by meteorite impact? Impressive evidence affirming the first question has been assembled, including global identification of the iridium-rich clay layer (in both
MASS EXTINCTIONS, CONCEPT OF

marine and terrestrial stratigraphic sections); presence of shocked mineral grains (e.g., quartz), macrospHERules of shock-melted minerals, and abundant soot in the global clay layer; and the presence at 65 Ma of the largest impact crater known from the Phanerozoic in the Yucatan, Mexico. Answers to the second question haveproved more difficult and will be considered later.

The recognition of extinction events in the geologic past demands comparison to modern extinctions, from the auroch to the myriad other species documented to have disappeared from human interference. Questions include not only how modern rates of extinction compare to those in the fossil record but also the consequences of extinction: What kinds of organisms are most susceptible to extinction and what are the patterns of biotic recovery after the pressure of extinction has been removed? These questions will be considered further later in this discussion.

II. MODELS OF MASS EXTINCTION

A variety of patterns of extinction have been observed in the fossil record around events, ranging from abrupt termination of species at an extinction level to gradual disappearance up to it, and perhaps beyond, the extinction level. Three scenarios for mass extinction have been proposed based on empirical observations: abrupt extinction, gradual extinction, and stepwise extinction.

A. Abrupt Extinction

This is the pattern hypothesized by Alvarez et al. (1980) of species disappearing in a geologic instant (which could in fact be 10$^3$ to 10$^4$ years). Observed declines in diversity before the event (such as seen in detailed records of foraminifera or broader records of dinosaurs before the Cretaceous-Tertiary) are a result of the Signor-Lipps sampling effect (discussed later).

B. Gradual Extinction

The fossil record is taken on face value, especially if there has been extensive sampling around the horizon of extinction. Slow attrition of species up to the end of a mass extinction has been claimed for extensively sampled foraminifers around the Cretaceous-Tertiary event, where several large, but rare foraminifers seem to disappear below the stratigraphic boundary and some small, generalized foraminifers occur above the boundary.

C. Stepwise Extinction

The fossil record is again taken at face value but exhibits a series of pulses of species terminations. This pattern has been hypothesized for situations such as the Late Devonian mass extinction ("upper Kellwasser Event") where intensively sampled taxa of different groups appear to disappear in small pulses separated by 10$^3$ to 10$^4$ years. The model can be expanded to intervals such as the end Ordovician (the second largest marine mass extinction of the Phanerozoic) where many trilobites and other marine animals of tropical areas appear to become extinct at the onset of major glaciation, and then, perhaps 10$^5$ years later, surviving deep-shelf benthos disappear as normal conditions of low oxygen return with the end of glaciation. (In this case, each pulse of extinction could be dissected to determine if it had been abrupt, gradual, or stepwise at finer timescales.)

These scenarios need to be distinguished from the low levels of extinction that are observed in all geologic intervals. These levels are normally termed "background extinction" to distinguish them from events of mass extinction. Background extinction for marine animals appears to decline through Phanerozoic time. Thus, smaller extinction events are more obvious in Newell-type data over the Mesozoic and Cenozoic eras than during the early Paleozoic when background rates were high.

III. INTERPRETING DATA FROM THE FOSSIL RECORD

The fossil record provides direct evidence of previous mass extinction but only an incomplete accounting because of differences in preservability of organisms (e.g., bivalve mollusks versus polychaete worms) and in scientific sampling (e.g., Europe and North America versus Antarctica). With an incomplete record, observed last occurrences of fossil species are only a minimum estimate of actual times of extinction. This consideration was formalized for mass extinction by Signor and Lipps (1982) who modeled how observed terminations of species would appear around an abrupt extinction (Fig. 2). With less intensive sampling or less complete preservation, the expectation is a pattern of gradual disappearance of fossil species up to a boundary of abrupt mass extinction. This sampling pattern holds true whether one is examining detailed stratigraphic sections of fossils or analyzing compilations of fossil taxa, like those of Newell.

Raup presented a very intuitive example of the Sig-
mass extinctions, Concept of

Figure 2. Hypothetical and calculated diversity curves reflecting the “Signor-Lipps effect”—that is, imperfect sampling around a catastrophic mass extinction. The dashed curve represents true diversity, which is treated as constant until time unit 4 at which there is a catastrophic mass extinction reducing diversity by three-quarters. If fossil taxa are not sampled up to the times of their true extinctions, declines in diversity will appear more or less gradual, depending on the intensity of sampling. Curve a represents the poorest sampling, and curve c the best (three times better than a). From Signor and Lipps (1982 with permission from the Geological Society of America).

nor-Lipps effect (Fig. 3). He used data collected for ammonites below the Cretaceous-Tertiary boundary to investigate how the fossil record of extinction would have appeared if the mass extinction had occurred at an interval 100 m lower in the stratigraphic section. (A meteorite impact or other unpredictable catastrophe potentially could occur anywhere in a stratigraphic section.) The decline in observed diversity up to the artificial event does not appear substantially different from the gradual decline of diversity actually observed below the true Cretaceous-Tertiary boundary. This suggests that problems of variable preservation and incomplete sampling of fossils can indeed influence empirical patterns of species disappearance around extinction events, hindering discrimination among the three models of mass extinction.

More powerful statistical methods have been developed to investigate how apparent last occurrences of fossil species may relate to actual times of extinction. These methods involve calculating “confidence intervals” on the time of last sampled occurrence of a species (Fig. 4). The basic concept is that the last fossil occurrence of a species that is rarely sampled is a poorer predictor of actual time of extinction than the last occurrence of a species that is densely sampled. Using various models of the density distribution of sampling of a fossil species, probability statements can be made about how far true extinction lies above the last observation of a species.

IV. Magnitudes of Mass Extinction

Questions of abrupt, gradual, or stepwise extinction involve patterns in the fossil record resolved over 10^3 to 10^5 years (encompassing the range from the late Pleistocene extinctions of large mammals to the historical extinction of species). On larger timescales, general magnitudes of mass extinction can be measured from global fossil data. These data are best for the marine record of animals from continental shelves and seas and fall into roughly three classes of magnitude (Fig. 5).

A. The End-Permian Mass Extinction

This class of magnitude stands alone in its effects on the biota (Erwin, 1993). Compilations of taxa lost indicate that more than 50% of animal families and 80% of genera in the oceans became extinct. Extrapositions of species loss have been attempted, using ecological rarefaction (how many species would be lost given measured declines in genera or families, assuming some distribution of species within higher taxa); results range from 90 to 96% loss of marine species. This loss of marine biodiversity at the end-Permian is unprecedented. Recent work suggests that on land important groups, including insects, tetrapod vertebrates, and plants, also experienced substantial declines in diversity.

B. Four Other Events of Marine Mass Extinction

This class of magnitude eliminated substantial proportions of marine animals and seem to have had nearly equal magnitudes: the end-Ordovician, Late Devonian, end-Triassic, and end-Cretaceous events. (The occurrence of these events at or near the end of geologic periods reflects the use of faunal change to define intervals of geologic time.) The four events have measured family extinction in the oceans of 15 to 29% and extrapolated species extinctions of 64 to 83%.
FIGURE 3  Apparent gradual extinction at an imaginary sudden extinction event. (a) Actual observed geologic ranges of 20 fossil species. The y-axis is stratigraphic position, in meters, below the Cretaceous/Tertiary boundary. Vertical lines show the total observed ranges of the species and horizontal ticks indicate horizons at which they were actually sampled. The curve at the right, labeled "Diversity," is the sum of the ranges of the species. (b) Resultant ranges if an imaginary catastrophic extinction were imposed at 100 m. Apparent last occurrences of species again are graded below the mass extinction and diversity appears to decline, both as a result of species being irregularly sampled through the geologic interval. (Note that the species in b have been reordered based on their highest geologic occurrence.) From Sepkoski and Koch (1996), based on Raup.
MASS EXTINCTIONS, CONCEPT OF

C. Other Intervals of Unusual Amounts of Extinction

This class of magnitude is now termed “extinction events” (Fig. 6). These appear in more detailed compilations of the kind Newell made as well as in precise biostratigraphic and paleogeographic analyses. Paleogeographic analyses suggest that many of these third-order events were not global in extent or taxonomic effects, unlike the two previous categories. Examples include the Cambrian biomere events, marked by nearly complete extinction of North American trilobites but with uncertain effects in other parts of the world or on other taxa (although some events are recognized in Australia, China, Siberia, and Europe); the lower Toarcian event affecting mollusks in Europe but not around the Atlantic Ocean and Mediterranean Sea but not evident in the tropics or in the Pacific.

There are extinction events that could be classified at lower ranks. For example, it has been demonstrated that marine animal communities persist on order of $10^6$ years during the middle Paleozoic, punctuated by rapid changes in faunal composition (Miller, 1997). This raises the question of Raup (1991a) as to whether extinction events represent a continuum between rare major events and frequent small events (discussed later).

V. HYPOTHESIS OF PERIODICITY

Given the varying magnitudes and geographic extents of extinction events, one would expect differing forcing agents. With a plethora of forcings acting independently over geologic time, it would be expected that extinction events would be distributed at random through the fossil record. Thus, it came as a surprise when Fischer and Arthur (1977) and Raup and Sepkoski (1984) observed that extinction events of first through third rank appeared regularly distributed through time. Raup and Sepkoski performed extensive statistical analyses of Newell-type data and concluded there was a strong periodicity of 26 myr for events during the Mesozoic and Cenozoic eras (Fig. 7). This suggested some sort of clocklike mechanism behind mass extinction with a periodicity unknown in terrestrial processes. Because one of the periodic events was the end-Cretaceous mass extinction, Raup and Sepkoski speculated that the clocklike mechanism might be extraterrestrial.

This speculation engendered both intriguing hypotheses from astronomers and geologists and critical scrutiny of data and statistical methods from paleontologists, geologists, and statisticians (Raup, 1991b; Sepkoski, 1989). The best-known hypothesis is Nemesis, sometimes called “Shiva”; this is an hypothesized small binary companion to the sun with a large and eccentric orbit that brings it through the Oort Cloud of comets beyond the planets every 26 myr. Nemesis’ gravitational perturbation scatters up to $10^9$ comets of which $10^1$ to $10^2$ might impact the earth, disrupting climate and causing mass extinction.

Nemesis has not been observed, and astronomical models cast doubt on the possibility that a small star could maintain a stable orbit at large distance from the sun through the 4.6 billion-year history of the solar system. Also, there is direct evidence of extraterrestrial impact for only a few of the periodic extinctions, despite intensive investigation: Shocked quartz has been found...
Mass extinctions, concept of

FIGURE 5  Time distribution of notable extinction events in the marine fossil record, classified as major ("first order"), intermediate ("second order"), and lesser ("third order") events. Only major and intermediate events are normally referred to as "mass extinctions." Abbreviations are mostly as in Figure 1; l = lower, u = upper. Names refer to stratigraphic intervals (mostly form stages). (?) indicates an event that has not been well documented. From Sepkoski (1992).

at several end-Triassic localities, and concentrated iridium, microtektites (annealed globules of impact melt), and a crater have been found in the Late Eocene, although their temporal correspondence to extinctions remains unclear. Searches of this nature are difficult and time consuming, and one should not be surprised if more discoveries are made.

Questions about data and analyses with respect to the hypothesis of periodicity have been largely technical (Sepkoski, 1989). Questions of data concern the veracity of fossil families and genera as indicators of extinctions in the geological past and the accuracy of geochronological dating of intervals of extinction. Statistical questions involve problems of analyzing data that are not sampled evenly through time (geologic intervals vary in duration; Fig. 7) and problems of significance when performing multiple tests when searching for periodicities of best fit. Some of these questions entered new analytic ground, leading to new kinds of analyses, but not all analytical issues have yet been resolved. Thus, a nonrandom, periodic distribution of extinction events remains a hypothesis on the table, with neither a clear forcing mechanism nor a definitive analytical test.

VI. THE KILL CURVE AND SELF-ORGANIZED CRITICALITY

Raup (1991a) treated extinction in the fossil past not as a hierarchy of events but as a continuum, in which the obvious events of large magnitude are rare relative to less obvious events of small magnitude. An analogy is to the historical record of river floods, in which the 1000-year event stands out but is really part of a spec-
Mass extinctions, concept of

Figure 6 extinction intensities for marine animal genera through the Phanerozoic. Intensity is measured as percent extinction (= 100 × number of extinctions/observed diversity) in geologic intervals averaging about 5.5 myr in duration. Estimated errors of estimate are indicated by error bars (± 1 standard deviation). Significant events are labeled: Tom = Tommotian; uhBo = upper Botomian; Llvi = Llanvirnian; Ashg = Ashgillian (end-Ordovician mass extinction); Lud = Ludlovian; Fra = Franian (Late Devonian mass extinction); Serp = Serpukhovian; Tar = Tarazian (end-Permian mass extinction); uNor = upper Norian (= end-Triassic mass extinction); Pli = Pliensbachian; Cal = Callovian; Tith = Tithonian; Apt = Aptian; Ceno = Cenomanian; Maas = Maastrichtian (end-Cretaceous mass extinction); uEoc = upper Eocene; mMio = middle Miocene. From Sepkoski (1996) with permission.

Raup used a standard function for these kinds of events to analyze probabilities of species extinction of a given magnitude during some duration of geologic time (Fig. 8): the longer the duration, or “waiting time,” the larger a pulse of extinction that can be expected. The limitation of Raup’s analysis is that it was based on generalized geological intervals, the same as illustrated in Figs. 6 and 7, and cannot distinguish between sudden events, such as observed at the end of the Cretaceous, and cumulative small events, such as may also have occurred during the last stratigraphic stage of the Cretaceous during the 9 million years prior to the impact-induced mass extinction.

Raup’s analysis was postulated on an assumption that external perturbations of varying magnitude caused most extinction through geologic time, and he in fact demonstrated a linear relationship between his empirically based kill curve and the similar curve established for the flux of meteorites of varying magnitude (Raup, 1992). Other approaches have postulated extinction to be a result of internal dynamics of the biota over time. Sole et al. (1997), for example, analyzed Phanerozoic time series of extinction and diversity fluctuation to determine if there was evidence of “self-organized criticality.” The prediction was that a power series from fourier transforms of the data would exhibit a linear \(1/f^x\) relationship when power was plotted logarithmically against the logarithm of frequency, \(f\). The exponent, \(x\), is expected to be between 1 and 2, which is indeed what was found.

Self-organized criticality refers to systems of interactive components that grow to a degree of complexity that leads to cascades into chaos or collapse. The standard analogy is to a pile of sand onto which one grain is added at a time. As the pile approaches critical size, one more grain can cause a small shuffle of sand while, far less frequently, another grain can cause a major
FIGURE 7. Time series for extinction intensity of marine animal genera through the Mesozoic and Cenozoic showing the putative 26-myr periodicity of mass extinction (vertical bars). The abbreviations are as in Fig. 6. Extinction intensity here is measured as an average of percent extinction in major taxa that have been standardized to zero means and unit standard deviations. From Sepkoski (1990).

There are three criticisms of this argument. First, some mass extinctions are clearly associated with physical disturbances, such as the meteorite impact at the end of the Cretaceous. Second, many other processes can produce 1/f power laws when analyzed in the way organized self-criticality is; for example, the secular decline in background extinction over time, perhaps induced by changing taxa with different characteristic rates of extinction (Sepkoski, 1984), can cause power series to decline relative to increasing frequency. Finally, there is no independent evidence that species within marine communities, for which data have normally been analyzed, are as interactive as posited.

There still may be some interesting avenues of inquiry to pursue. The state of the biota is not constant, given variations in the earth’s climates, changing positions of continents, and varying tectonic events and their effects on environments and the time over which the biota evolves to these varying conditions. A perturbation at one stage of the earth’s conditions and the biota’s development could have very different effects than an identical perturbation at a different time (Sepkoski, 1989). Thus, an expectation of some chaotic element in the record of extinction cannot be ruled out.

VII. SELECTIVITY OF MASS EXTINCTION: VICTIMS AND SURVIVORS

An obvious effect of the end-Cretaceous mass extinction is that all (nonavian) dinosaurs disappeared whereas some mammals survived. At other mass extinctions there are also cases of major taxonomic groups disappearing and others surviving. This observation has led to a search for rules as to what makes some kinds of organisms more vulnerable to mass extinction than others. Patterns that have been found are largely probabilistic, and all seem to have unexplained exceptions:

1. Taxa that have high rates of extinction during times of background extinction suffer disproportion-
MASS EXTINCTIONS, CONCEPT OF

FIGURE 8 Raup’s “kill curve” showing the expected extinction of marine species, as a percentage of standing diversity, over intervals of time (“mean waiting time”). During intervals of less than 10^6 years, percent of diversity expected to become extinct is low, but over longer intervals, especially greater than 10^8 years, extinction events of large magnitude become quite probable. The solid line is the best fit curve to fossil data; the dotted curves fall at bounds of uncertainty for empirical data. From Raup (1991a).

2. Animals, particularly terrestrial vertebrates, with large body size seem to be particularly prone to extinction. Rates of extinction measured for higher taxa (e.g., classes) during background times tend to be good predictors of magnitudes of extinction at major events. Thus, trilobites, with higher rates of background extinction than articulate brachiopods, suffered more loss at the end-Ordovician, and brachiopods, with higher characteristic rates of extinction than bivalves, suffered greater proportional declines at the end-Permian and even the end-Cretaceous, when brachiopods were relatively marginalized (Sepkoski, 1984). It remains a major question as to why major taxonomic groups have different characteristic rates of extinction during background times and why these rates are conserved over vast stretches of macroevolutionary history and species turnover.

3. Environment seems to have played only a minor role in selectivity of extinction among marine animals at mass extinctions. Rates of extinction among Paleozoic marine animals tended to increase toward shallower waters during background times but to be equal across continental shelves at mass extinctions. There is some suggestion, however, that deep-water animals fared better: Marine shelves seem to have been repopulated by descendants of deep-water trilobites after Cambrian extinction events; shelf communities contained dispro-
portionate numbers of deep-water sponges and corals after the Late Devonian mass extinction (McGhee, 1996), and descendants of mollusks typical of low-oxygen deep-water communities of the late Paleozoic are common after the end-Permian mass extinction (Erwin, 1993).

These reports of taxonomic selectivity at mass extinctions are largely confined to analyses around a few events of extinction, and much larger comparative studies need to be conducted across all mass extinctions for both marine and terrestrial organisms. However, if there are consistent probabilistic biases in terms of the properties of species and higher taxa that survive mass extinctions, then extinction events on frequencies of $2.6 \times 10^7$ years, eliminating 30 to 90% of the biota, could be a major factor in the history of life.

VIII. RECOVERIES FROM MASS EXTINCTIONS

Global data on mass extinctions exhibit not only the geologically rapid declines of the biota at mass extinctions but also subsequent recoveries of biodiversity in the aftermaths. With modern data, these recoveries are seen to encompass longer intervals of time than the extinction events, ranging up to about 5 myr for third-order extinction events to around 10 to 15 myr for second-order mass extinctions; recovery from the great end-Permian mass extinction took even longer but was interrupted by the end-Triassic event, some 30 myr later.

On long geologic timescales, recoveries from mass extinctions follow simple patterns expected for a diversifying biota. At more detailed scales, there is great complexity, with some patterns repeated after every mass extinction.

A. Large-Scale Rebound of Diversity

To a first approximation, diversification in the oceans can be described as a system with equilibrium constraints imposed by limitations of resources and their utilization (Sepkoski, 1984, 1992). This is most obvious in the long plateau of diversity during the Paleozoic era, spanning approximately one-quarter billion years. This interval witnessed two second-order mass extinctions, the end-Ordovician and Late Devonian. Significantly, after each, animal diversity rebounded to previous, unperturbed levels and then continued the Paleozoic plateau (Fig. 1b). Also, the rate of rediversification after these mass extinctions was approximately the same as during the Ordovician radiations that established the Paleozoic plateau of global marine diversity. Such a system can be modeled if it is assumed that Paleozoic animals were constrained by an equilibrium diversity, conditioned on the environment and the way these animals subdivided resources (Fig. 9). The simplest mathematical description of diversification is a logistic model, assuming decreasing rates of origination as diversity increases. A mass extinction is a perturbation to the environment that increases extinction rates (Fig. 9b). Following the perturbation, when extinction rates return to background levels, taxa rediversify to the normative level (Fig. 9c).

Patterns are more complex through the quarter-billion years after the Paleozoic era, as diversity in the oceans generally increased. The increase reflects expansion of the modern marine fauna, which may have subdivided resources in ways different from the fauna of the Paleozoic era. Nonetheless, rebounds after mass extinctions during the Mesozoic and Cenozoic eras were much more rapid than the long-term increase in marine diversity when measured over millions of years and were congruent with initial rates of diversification of the taxonomic groups involved (cf. Miller and Sepkoski, 1989).

B. Delayed Diversification

When rebounds from mass extinction are analyzed at finer timescales, many complications become apparent (Erwin, 1998). One is delayed recovery: rediversification does not commence immediately after perturbation. For example, there are only slow rates of rediversification for 10$^5$ years after the end-Cretaceous event among planktonic foraminifers (actually, fast at first and then slow for some four myr; D’Hondt et al., 1996), benthic mollusks (Jablonski, 1998), and terrestrial mammals (Maas and Krause, 1994). Recovery seems to be delayed for nearly 5 myr after the much larger end-Perman event, with only depauperate faunas in the oceans and on land (Erwin, 1998).

Explanations for this pattern have varied. Some workers have suggested that the extended post-extinction intervals of low diversity reflect lingering effects of the external perturbation—that is, continued environmental stress or instability. Others have hypothesized that substantial amounts of time are required for the evolution of new species that can reestablish normal ecosystem function which, in turn, can support high diversity. An example is the evolution of planktonic...
MASS EXTINCTIONS, CONCEPT OF

C. "Disaster Species" and "Lazarus Taxa"

A feature that appears consistent during early phases of recovery from mass extinction is the appearance of "disaster species." These are remarkably abundant and geographically widespread species that appear in the waning stages or aftermaths of mass extinctions. Examples include the small foraminifer, *Guembelitria*, that spread out from marginal environments to form nearly monospecific assemblages immediately after the end-Cretaceous extinction and the terrestrial "fern spike" observed at the same time on land. Another example is the terrestrial synapsid (''mammal-like reptile''), *Lystrosaurus*, that produced monospecific assemblages in Gondwanaland following the end-Permian mass extinction.

Disaster species are characterized by not only great abundance in fossil deposits but also short geologic durations. Marine disaster species seem to flourish for around $10^5$ years and then disappear, often to be replaced by another "disaster species" with similarly short duration. This appears like an evolutionary "boom and bust" cycle. Thus, the earliest Tertiary foraminifer, *Guembelitria*, declined and was replaced in dominance by a succession of bursts of *Paravularugoglobigerina*, "Eoglobigerina," and finally *Woodringina* over timescales of $10^5$ years as the planktonic ecosystem seemed to regain stability.

Perhaps the antithesis of disaster species are "Lazarus taxa." These are lineages that disappear around mass extinctions—seemingly to have died—only to reappear in the fossil record some $10^6$ to $10^7$ years later (Erwin, 1998; Jablonski, 1986). These can be conspicuous but usually never dominant members of the biota prior to a mass extinction. Presumably, these taxa survived in some environmental or geographical refugium until the ecosystem regained sufficient function so that their dispersal—and recovery of sufficient abundance to be encountered by paleontologists—was possible.

IX. THE MODERN BIODIVERSITY CRISIS

The collapses of ecosystems in the fossil past can inform thinking and actions with respect to the contemporary
loss of biodiversity. The fossil record is incompletely known, but the data seem little worse than estimates of living biodiversity and current rates of species extinction. Estimates of modern global biodiversity range over more than an order of magnitude, from 5 to 10^9 species. Estimates of present-day loss of species range over nearly two orders of magnitude, from 5 to 1.5 x 10^7 species extinction per day. Most of these extinctions are terrestrial animals (largely insects), whereas the best data from the fossil record are for marine species. Still, in the geologic past, mass extinctions appear to have occurred largely contemporaneously both on land and in the seas (except for the end-Permian event), so some simple calculations can be made.

If a median contemporary extinction rate of 41 species per day is assumed and species attrition is treated as a negative exponential, then it would take only about 16,000 years for 90% of the modern biota to become extinct (Sepkoski, 1997). This is a long time by human standards, but it is beyond the limits of geologic resolution at 250 Ma (the end-Permian), the only other time when 90% of Earth’s biota disappeared.

This discussion has been largely in terms of fossil diversity. Another variable that can be measured in the geologic record is ratios of stable carbon isotopes. Because organisms tend to utilize 12C in slight bias over the heavier 13C, the ratio of the two isotopes preserved in the rock record can indicate, among other things, the activity of primary producers. For example, after the end of the Cretaceous, there was no difference in isotopic ratios of carbon in the skeletons of surface planktonic foraminifers and deep-sea benthic foraminifers. This indicates that organic carbon stopped sinking to the deep ocean. This change must have resulted from either (a) a major decline in oceanic productivity or (b) a major collapse of community structure in marine producers or consumers (D’Hondt et al., 1998).

X. SUMMARY

Exploration of the fossil record has demonstrated that the earth’s biota is fragile at timescales of 10^7 years, suffering numerous declines in diversity. The magnitudes of these declines have been variable, but at least five times in the last 500 myr animal diversity was reduced by more than 50%, with the most severe event, at the end of the Permian, eliminating more than 90% of animal diversity. Recoveries from these events have been slower than the mass extinctions, often taking 5 to 10 myr or more. Although the long-term rebouts of biodiversity are predictable, the detailed patterns of recovery are complex, involving outbreaks of disaster species and considerable ecological instability over timescales of 10^5 to 10^6 years. Current rates of species extinction could eliminate as many species as seen in past mass extinctions in a geologically short interval, and biotic recovery could be long and unpredictable.

See Also the Following Articles

Bibliography


MASS EXTINCTIONS,
NOTABLE EXAMPLES OF

Douglas H. Erwin
National Museum of Natural History

I. Great Mass Extinctions
II. Lesser Mass Extinctions

GLOSSARY

carbon cycle Photosynthetic organisms preferentially use the carbon-12 isotope instead of carbon-13, enriching living organisms in C-12. This differentiates the organic from inorganic carbon reservoirs. Quantifying shifts between the two isotopes of carbon reveals shifts between the two reservoirs.

mass extinction A rapid loss of a large fraction of biodiversity on timescales of 10^6–10^7 years, generally involving a variety of unrelated groups.

supercontinent The amalgamation of many continental masses into a single mass through continental drift. The supercontinent of Pangea (from 300 to approximately 190 million years ago) included most continental areas other than those that comprise east Asia.

THE 3.6-BILLION-YEAR FOSSIL record is interrupted by numerous mass extinctions, but only the 600 million years since the appearance of animals is sufficiently dense to provide an adequate record of these great biotic crises. Five great mass extinctions punctuate the record, several of which had major impacts on the course of evolution. Additionally, there are many smaller biotic crises, but not all of these are well studied.

I. GREAT MASS EXTINCTIONS

The five most severe biotic crises occurred during the Late Ordovician Period, the Late Devonian, the Late Permian (two closely spaced episodes), at the end of the Triassic Period, and at the end of the Cretaceous (the great Cretaceous–Tertiary (K/T) mass extinction that eliminated the dinosaurs) (Figs. 1 and 2). Despite numerous attempts to find common mechanisms for these events, a variety of different causes appear to have been involved, from global cooling during the end-Ordovician event to the impact of an extraterrestrial object with the earth during the K/T mass extinction.

A. Late Ordovician

The second largest of the great mass extinctions occurred during the Late Ordovician [439 million years ago (Ma)] in two pulses separated by approximately 1 million years. With few, if any, organisms on land, this marine extinction of approximately 25% of all families and 60% of genera affected most marine groups. Despite the magnitude of the extinction, it had few lasting ecological effects. With some minor exceptions, Silurian faunas look much like Ordovician ones, in contrast to the profound differences between Triassic and Permian faunas.
MASS EXTINCTIONS, NOTABLE EXAMPLES OF

Graptolites, a group of colonial, floating marine organisms, were commonly found on the outer shelves and were almost completely eliminated by this extinction, with only a few species surviving. Among other swimming and floating groups, conodonts, nautiloids, and planktonic groups all appear to have suffered considerable extinction. The articulate brachiopods were one of the major components of Ordovician ecosystems, and approximately 83% of brachiopod genera became extinct, primarily during the first of the two crises. A brachiopod fauna associated with the brachiopod *Hirnantia* developed in many areas between the two extinction pulses. At a single locality, fewer than 10 species are present and all appear to be adapted to cold water. This *Hirnantia* fauna appears to have reached equatorial latitudes during this interval, providing one index of the extent of cold-water conditions. The second phase of the extinction wiped out this fauna. Trilobites display a similar pattern of extinction followed by the spread of a low-diversity, broadly distributed group between the extinctions, however, the second extinction appears to be more significant than the first. Trilobites living in deeper water and floating forms suffered greater extinctions than those occupying shallow-water habitats. Bivalves and echinoderms also experienced considerable extinction, whereas bryozoans, a colonial group of filter feeders, were not affected as severely. Approximately 70% of rugose corals disappeared, but they recovered quickly so the extinction did not have a major impact. In fact, this is the only major mass extinction that did not trigger an extensive turnover in reef communities. The decrease in sea level associated with glaciation eliminated deposition of marine rocks in many areas, of course, thus complicating analysis of changes in biodiversity during this interval.

The causes of the Late Ordovician mass extinction are relatively clear, although there is disagreement about the relative significance of different aspects. Global climates during the Ordovician were warm; geologists term this a greenhouse climate. One consequence of this was a decline in vertical circulation in the oceans, leading to a depletion of oxygen in the deep oceans. As a large amount of continental area moved near the South Pole, the climate cooled and glacial conditions set in; as ice developed, sea level decreased, eventually exposing the continental shelves and leading to the first phase of the extinction. Deep-water groups were particularly affected. A low-diversity, broadly distributed community adapted to relatively cool conditions developed in the wake of this extinction. The end of the glaciation 500,000–1 million years later led to an increase in sea level and ocean temperatures. The rising

---

**FIGURE 1** Diversity of marine families through the Phanerozoic (the past 600 million years). The thick black line shows the total number of marine families in each of 83 geologic stages. The circles correspond to the five major mass extinctions: 1, Late Ordovician; 2, Late Devonian; 3, Late Permian; 4, end-Triassic; and 5, end-Cretaceous. The lower solid line shows the proportion of family diversity assignable to the Cambrian evolutionary fauna (Cm); the dotted line shows the proportion of family diversity assignable to the Paleozoic evolutionary fauna (Pz), and the area between the dotted and thick solid line is the proportion of family diversity assignable to the modern evolutionary fauna (Md). V, Vendian; C, Cambrian; O, Ordovician; S, Silurian; D, Devonian; C, Carboniferous; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous; T, Tertiary; Cz, Cenozoic; Ma, million years ago [after Sepkoski, J. J., Jr. (1984), A kinetic model of Phanerozoic taxonomic diversity. III Post-Paleozoic families and mass extinction. *Paleobiology* 10, 246–267; used with permission).

**FIGURE 2** Major and lesser mass extinctions during the Phanerozoic. Not all the lesser mass extinctions are discussed in the text. Some only appear in compilation studies, as in Fig. 1, and may be statistical artifacts. See the legend to Fig. 1 for abbreviations.
seas are associated with the spread of low oxygen, or anoxygenic conditions, leading to pervasive extinctions among shallow-water species. Changes in the carbon cycle, chronicled as a shift in the ratios of the isotopes of carbon-12 to carbon-13, coincide with the major events of this scenario.

Although rocks recording this extinction are common in Europe and China, few deposits in North America span the extinction interval other than an important sequence in Anticosti Island, Canada, and a newly described sequence in central Nevada. The rocks in Nevada preserve an excellent record of the decline in sea level, the biological extinction, and consequent perturbation of the carbon cycle (Finney et al., 1999). As the onset of glaciation lowered sea level, graptolites suffered the first extinctions, and the extinctions can be tracked from near-shore regions into deeper water habitats. Finally, as the decrease in sea level exposed most of the shelves, most marine communities were affected. The Nevada sections cover a diversity of environments within a small region, allowing the extinctions to be tracked between different environments. This is rarely the case in other areas; therefore, the difference in timing of extinctions has been difficult to resolve. The Nevada sections also reveal that a sharp shift in the carbon cycle was related to the onset of glaciation and the resulting decrease in sea level but was not directly related to the extinction.

B. Late Devonian

Although extraterrestrial impact is commonly associated with the Cretaceous/Tertiary mass extinction today, it was first proposed in 1970 for the Late Devonian mass extinctions. These biotic crises have been recognized since the 1870s, but the rapidity of these events remains unclear. For the Givetian and Frasnian Stages of the Late Devonian, Sepkoski (1996) records a 57% extinction of marine genera (47% for filtered data) and approximately 22% extinction of marine families. Particularly in European rocks, a sequence of events has been distinguished: a Taghanic event at the end of the Givetian Stage, the Kellwasser event at the end of the Frasnian, and a subsequent extinction episode near the Devonian-Carboniferous boundary (the Hangenberg event), although it is unclear whether the last event is part of the same series or has a different cause. Terrestrial plants had evolved and evidently experienced their first biotic crisis during this interval as well, although the available data are too spotty to reveal much information (McGhee, 1996; Hallam and Wignall, 1997).

Perhaps the most geologically obvious effect of the mass extinction is the elimination of the widespread Devonian reefs. These reefs were dominated by rugose and tabulate corals (the corals of the Paleozoic) and stromatoporoids, a group of calcareous sponges, in addition to some brachiopods and large foraminifera. Major Devonian reef complexes are known from Australia, western North America, Russia, and Europe. The extinctions suffered by these groups essentially eliminated true reefs until the Middle Permian. Brachiopods were one of the most diverse members of Devonian marine communities and illustrate the magnitude of the extinction. At least six families disappeared during the Taghanic crisis. McGhee (1996) indicates that 73% of brachiopod genera disappeared during Frasnian, including at least two major orders; many of the extinct genera were from tropical regions, suggesting, as for other groups, that higher latitude, cold-water groups were less susceptible to extinction. Brachiopods recovered quickly during the Famennian and were relatively unaffected by the Hangenberg event. Bryozoans have perhaps the strangest pattern of extinction. They suffered one of their largest biotic crises at the end of the Givetian, but no other groups are known to be affected at this time; it is suspected that the data may be faulty in this case. Trilobites, although much reduced in diversity from the Cambrian and Ordovician, were still important during the Devonian. The Givetian assemblage included eight orders and 13 families, but only a single family and eight genera survived to the Famennian. Although there is heightened extinction from the Givetian through the Frasnian, the only discrete crisis occurred during the Kellwasser event when numerous important groups disappeared. Deep-water and some shallow-water trilobites diversified immediately after the extinction during the Famennian, but trilobites suffered considerable extinction again at the Devonian-Carboniferous boundary and never regained their previous diversity. Two other arthropod groups, the eurypterids and the ostracods, also belonged to Devonian marine communities. Eurypterids were relatively rare, and their diversity appears to be declining throughout the interval, although one study suggests the extinction of several families during the last two stages of the Devonian. The ostracod record is far better. The pelagic ostracods suffered considerable extinction during the Kellwasser event but little extinction during the Hangenberg crisis. Among the benthiic ostracods significant extinctions occurred during the Kellwasser, with groups tolerant of low oxygen showing greater survival. Ammonoids, the coiled cephalopods that were distant cousins of the modern chambered Nautilus, almost disappeared, with an estimated 88% species extinction.
Ammonoids display a boom-and-bust pattern throughout their history, and this includes massive extinctions during the Taghanic event, in which only two genera survived, and the Kellwasser event, with only few surviving genera, mostly from deep waters. Famennian ammonoids diversified considerably, producing some bizarre morphologies, but only a single family, and perhaps a single genus, survived the Hangenberg event at the end of the Devonian. Ammonoids once again diversified strongly in the early Carboniferous and again became important members of marine ecosystems. Nautiloids, the other major group of fossilizable cephalopods, encountered only a minor perturbation during the Kellwasser event but suffered perhaps their greatest crisis during the Hangenberg event when 11 of 19 families disappeared. In contrast to the ammonoids, early Carboniferous nautiloids did not expand greatly in diversity. The other major molluscan groups are the bivalves and gastropods. Neither was particularly important during the Devonian and each suffered only minor losses in diversity.

The only prominent vertebrate group was the fish, which were amazingly diverse during the Devonian (hence the nickname ‘the age of fishes’). Fish have proven to be very resistant to mass extinction, but all fish groups other than sharks did poorly during this crisis. Among the jawless, armored fish all nine families disappeared. More advanced fish suffered as well during both the Kellwasser and Hangenberg crises. Diversifications occurred after each extinction, however, and that following the Hangenberg eventually led to the teleosts, the modern bony fish. Conodonts are tiny phosphatic elements that formed part of the jaw of small eel-like chordates. Because of their diversity, conodonts are important tools in correlating rocks from different parts of the world. The Kellwasser represents a major biodiversity crisis for the group, at least at the generic and species level. Less is known of their history during the Hangenberg crisis.

What patterns of ecological selectivity are apparent from the variations in extinction intensity among different groups? Often, patterns of differential extinction are among the best clues to the causes of mass extinctions and other biotic crises. The patterns of change are complex, with long-term decline in many groups as well as during the discrete Taghanic, Kellwasser, and Hangenberg events. The first of the three biotic crises primarily affected shallow-water benthic groups. The Kellwasser is the best known of the three events (McGhee, 1996). Warm-water groups, particularly brachiopods and reef forms, suffered greater extinction, whereas groups adapted to high latitudes and colder waters seemed to have done well. Most swimming groups either became extinct or declined greatly in diversity. Whether the Hangenberg crisis qualifies as a mass extinction is unclear. Many swimming groups, including placoderm fish, nautiloids, and ammonoids, as well as trilobites and stromatoporoids experienced considerable extinction. The overall pattern is the reverse of that of the Taghanic event (Hallam and Wignall, 1997). The causes of the first and third events, the Taghanic and Hangenberg, are poorly understood, so emphasis will be placed on the better studied Kellwasser episode.

Extraterrestrial impact, sea-level fluctuations with attendant spread of anoxic waters, and climatic changes and global cooling are among the more prominent suggested causes of the biotic crisis (McGhee, 1996). The evidence for impact can charitably be described as spotty. Such evidence would include demonstration that extinctions are rapid and virtually simultaneous; signs of enrichment of elements found in extraterrestrial objects, particularly iridium; indications of impact including shocked quartz; fused glass from the impact site; and geochemical anomalies, including a shift in the carbon cycle. Although evidence of impact is present in Late Devonian rocks, including the 52-km Siljan crater in Sweden and heightened levels of iridium, these are not closely associated with extinction horizons. The most recent advocates of impact have placed considerable emphasis on shifts in the carbon cycle, but such shifts can arise through a variety of causes and are not diagnostic of impact. This does not imply that no impact was associated with the extinctions, only that no compelling evidence for such impact has been advanced (Copper, 1998).

Global cooling due to the closure of an equatorial seaway with an associated decrease in sea level has been advanced based on the heightened survival of cold-water groups, the presence of glacial sediments in Brazil, and the rapid sea level fluctuations, reminiscent of Pleistocene glacially induced fluctuations (Copper, 1998). The correlation between the physical events and the extinction is doubtful, however, and the enhanced survival of cold-water taxa can also be interpreted as survival of deep-water forms. Given the time span over which these changes operate, this model also seems to require a fairly prolonged extinction. In McGhee’s (1996) excellent analysis of the extinction he concludes that global cooling best explains the pattern of extinction and survival, but he argues that the extinction was relatively sudden and thus glaciation is not an adequate explanation. Although acknowledging the difficulties with the im-
The two pulses of the end-Permian mass extinction collectively extinguished approximately 95% of all marine species and perhaps 70% of species on land. The destruction caused the most pervasive reorganization of marine communities since the beginning of the Paleozoic, and modern marine communities continue to reflect the heritage of this event (Erwin, 1993). The only mass extinction ever suffered by insects occurred during the end-Permian extinction. Echinoderms suffered pervasive extinction as well, with the demise of several major groups of stalked echinoderms, including an entire class, the Blastoidea. Although echinoids (sea urchins) are ubiquitous members of marine communities today, only a single genus (Mio-
cidaris) survived the end-Permian extinction. Echinoderms are present in the Early Triassic, but most fossils are fragmentary and appear to represent very few species that were highly abundant. They were essentially weeds, and are found in many parts of the world.

A group of microfossils that produced a calcareous test, or shell, were a major part of Permian marine communities. Known as the Foraminifera, they have been well studied, in part because they never recovered from the extensive extinctions they suffered during the Late Permian. As with several other groups discussed later, they suffered considerable tropical extinction at the end of the Guadalupian, during which 53 of 82 genera died out. A small increase in diversity occurs after this extinction, followed by the second phase of extinction, with the loss of 90% of families and 95% of the genera. The colonial bryozoans suffered catastrophic extinctions, although only the lacy fenestrates became completely extinct. Echinoderms suffered pervasive extinction as well, with the demise of several major groups of stalked echinoderms, including an entire class, the Blastoidea. Although echinoids (sea urchins) are ubiquitous members of marine communities today, only a single genus (*Mio-
cidaris*) survived the end-Permian extinction. Echinoderms are present in the Early Triassic, but most fossils are fragmentary and appear to represent very few species that were highly abundant. They were essentially weeds, and are found in many parts of the world.

A group of microfossils that produced a calcareous test, or shell, were a major part of Permian marine communities. Known as the Foraminifera, they have been well studied, in part because they evolved rapidly and are useful in determining the age of rocks. They were also ecologically diverse and provide some clues to the nature of the extinction. One important group of shallow-water forams almost completely disappeared.
in the first extinction pulse and the few remaining spe-
cies became extinct during the second phase of the
extinction. Many new species appeared after the first
wave of extinction. The second phase of extinction
eliminated many architecturally complex forms found
in the tropics.

Among molluscs, the extinction was more muted.
 Gastropods suffered considerable extinction at the ge-
neric level, but this eventually led to new groups of
gastropods becoming important in the Triassic. Broad
geographic range and occupation of a wide range of
environments evidently aided survival at the end of the
Permian. Bivalves experienced relatively minor extinc-
tion, but took magnificent advantage of the postextinc-
tion possibilities, becoming dominant members of mod-
ern communities. The ammonoids, as usual, suffered
near-catastrophic extinction.

Extensive biotic crises occurred on land as well, pro-
viding a critical clue to the cause. Clearly, any proposed
cause which affects the oceans to the exclusion of land
is ruled out. Twenty-two orders of insects had appeared
by the end of the Permian, and 9 became extinct and
another 10 suffered serious diversity declines. This is
the only mass extinction ever suffered by insects. Many
of the groups that disappeared could not fold their
wings back over their body (as do flies and butterflies)
but rather held them straight out to the side; dragonflies
are a relic of these Permian groups. In contrast to studies
in the 1980s which suggested that plants experienced
relatively little extinction, recent work has revealed
massive perturbations, including a relatively sudden
extinction of plants in Australia. In many parts of the
world a sharp spike of pollen from fungi has been
documented at the Permo-Triassic boundary. What caused
this remains unclear, but it suggests a profound and
rapid disruption of terrestrial ecosystems.

Earliest Triassic sediments in the Karoo Basin of
South Africa contain abundant specimens of the pug-
nosed therapsid Lystrosaurus and occasional specimens
of a few other species. In contrast, the Permian rocks
below have a diverse fauna of several different groups of
carnivores and herbivores. Although the exact posi-
tion of the Permno-Triassic boundary in these terrestrial
sequences is unclear, as is the correlation to the marine
Permno-Triassic boundary, abundant evidence demon-
strates a considerable extinction of vertebrates. Approx-
imately 78% of tetrapod families apparently became
extinct at about this time.

The end-Permian mass extinctions occurred during
an interval of considerable change in tectonics, climate,
and other aspects of the physical environment. Virtually
all these changes have been invoked by one or another
extinction model at some time. Perhaps the most com-
mon link has been between the extinction and the for-
mation of the supercontinent of Pangea. During the
Permian virtually all the continents, with the exception
of some pieces that today comprise China and east Asia,
collided to form a single supercontinent. The climatic
and biological effects of the formation of Pangea have
figured prominently in many extinction scenarios, but
geologic data have shown that it formed approximately
20–25 million years before the extinction. broke apart
slightly, and then reformed in the Late Triassic. Neither
event is associated with any biodiversity crisis. Addi-
tionally, the rapidity of the extinction is much too fast
to be explained by the slow dance of the continents.
Glaciation has also been invoked and may have been
involved in the first extinction pulse, but there is scant
evidence for glaciation in the latest Permian.

Deposits of rock spanning the Permno-Triassic
boundary extinction interval are relatively sparse, and
this led several generations of workers to suggest a
profound decrease in sea level at this point, and many
extinction scenarios have focused on this issue. Recent
studies of many regions have shown that the regression
ended well before the onset of the extinction, and sea
level was actually rising during the extinction (Hallam
and Wignall, 1997). As with many of the extinction
events, a sharp change in the carbon cycle has been well
documented at the extinction horizon. Unfortunately,
several different extinction scenarios can produce the
observed shift.

More significant, however, is the close connection
in time between the eruption of extensive flood basalts in
Siberia and the extinction. These comprise one of the
two largest known continental flood basalts and rep-
resent at least 13 large lava flows complexes and more
than 45 individual flows. The flood basalt covers an
area of 2–3 million cubic kilometers, with a depth of
more than 4 km in some places. The entire complex of
flows appears to have erupted in approximately 1 mil-
lion years—a phenomenal rate of volcanic production.
The link to the extinction is more obscure, however.

The cause of the end-Permian extinction remains
ambiguous, despite the tremendous research effort
since 1990. The following facts are clear, however
(Bowring et al., 1998): Low-oxygen levels occurred in
deep and shallow waters during the extinction interval;
sea level was rising; the extinction coincides with an
abrupt shift in the carbon cycle, although the magnitude
of the shift is less than previously thought; the extinc-
tion pattern is consistent with poisoning from carbon
dioxide; and there is increasing evidence for a sudden
climatic warming immediately after the extinction, per-
happens associated with the extinction. Several extinction scenarios are consistent with this scenario, including the impact of an extraterrestrial object (despite many searches, no evidence for impact has been discovered), release of CO$_2$ and sulfur dioxide from the Siberian flood basalts triggering greenhouse warming and acid rain, and a shift in ocean circulation.

**D. End-Triassic**

The end-Triassic extinction is the second smallest event in most analyses and probably the most poorly known of any of the five great mass extinctions. Sepkoski's (1996) data show a 53% extinction for all marine genera (40% when the data are filtered) and 22% for marine families, similar to the K/T mass extinction. Among marine groups, Sepkoski's database shows significant extinctions among ammonoids, bivalves, gastropods, and brachiopods, but the most reliable studies have focused on ammonoids and bivalves in northwestern Europe; little data are available from elsewhere. Whether this extinction is entirely at the end-Triassic or includes a substantial event earlier, during the Carnian Stage, has been the subject of active debate.

Ammonoids always experienced a boom-and-bust pattern, and this event marks one of their largest crises. The group was quite diverse during the latest Triassic, but at least six superfamilies became extinct near the boundary, followed by a rapid rediversification in the Early Jurassic. Some estimates suggest that only a single ammonite genus survived the crisis. Although no family-level extinctions occurred among bivalves, approximately half the genera disappeared and turnover at the species level was particularly high. Endemic genera and those in deep offshore facies appear to have experienced greatest extinction. Conodonts, a group of phosphatic microfossils formed by simple chordates, survived the crisis. Whether or not any vertebrate extinctions occurred remains highly contentious. Analysis of the Newark group vertebrates indicated extensive extinctions, but this conclusion has been challenged by many vertebrate paleontologists who have argued that the Newark Super-pgroup record is sparse and there are no other data supporting an extinction of terrestrial vertebrates.

The causes of this extinction remain enigmatic, with most attention focused on flood basalt volcanism, possible bolide impact, and marine anoxia. Recent age dates reveal that the Central Atlantic Magmatic Province, an extensive region of flood basalts extending across eastern North America, Europe, northwest Africa, and northeastern South America, erupted coincident with the end-Triassic mass extinction (Marzoli et al., 1999).

As with the end-Permian and K/T extinctions, the causal link between the volcanism and extinction remains unclear. The assumption is that some combination of rapid climatic shifts, introduction of volcanic gases (carbon dioxide and sulfur dioxide), and acid rain would be sufficient to cause the extinctions. The available terrestrial data are consistent with a sharp increase in carbon dioxide and a disruption of the carbon cycle. The extinction had previously been associated with the Karoo flood basalts in South Africa, but these are now dated as Early Jurassic.

Changes in sea level, perhaps induced by uplift of the supercontinent of Pangea during the initial breakup that produced the Atlantic Ocean, have been invoked as well. Hallam and Wignall (1997) invoke a sharp decrease in sea level and then a transgression of low-oxygen waters, with the extinction coincident with the anoxia. Such a mechanism fails to explain the widespread extinctions of terrestrial plants, however.

The final possibility is the impact of an extraterrestrial object. The evidence is less secure than for the K/T extinction, but it includes the fern spike associated with the K/T extinction. Whether or not any vertebrate extinctions occurred remains highly contentious. Analysis of the Newark group vertebrates indicated extensive extinctions, but this conclusion has been challenged by many vertebrate paleontologists who have argued that the Newark Super-pgroup record is sparse and there are no other data supporting an extinction of terrestrial vertebrates.
and coincides with the boundary. The diverse record noted previously, the ammonite extinction is abrupt of large (up to 1 m), flat clams, the inoceramids. As evidently did so well before impact event, as did a group of another group of cephalopods, the belemnites. The bi-
tination pattern. The ammonites finally succumbed, as did the foraminifera suffered much less extinction, as did phytoplank-
ous microfossils also suffered species extinction of per-
haps 85%.

Considering the marine organisms first, planktonic foraminifera, a group of floating single-celled organ-
isms, experienced a fairly catastrophic extinction. There has been considerable debate regarding the abruptness of these disappearances (MacLeod et al., 1997), which can only be resolved by appropriate statistical analysis. Preservation of fossils is so uneven that even a cata-
strophic mass extinction will produce an apparently gradual pattern of disappearance leading up to the ex-
tinction horizon. Statistical analysis of fossil occur-
rences can correct this problem, however. Such an anal-
ysis of ammonites from the coast of France and Spain demonstrated at most three phases of extinction (Fig. 3; Marshall and Ward, 1996). Of the 28 species of ammonite, 23% probably became extinct during the several million years leading up to the mass extinction; a decrease in sea level just before the extinction may have been responsible for the extinction of up to 33% of species, but 40–75% of the species went extinct at the K/T boundary. The disappearance of planktic fora-
minifera is likewise abrupt, as emphasized by the recent recovery of a core from the ocean bottom in the western Atlantic (Norris et al., 1999). Another group of calcare-
ous microfossils also suffered species extinction of per-
haps 85%.

Among other marine groups, the benthic foraminifera suffered much less extinction, as did phytoplank-
ton. The various mollusc groups vary greatly in extinc-
tion patterns. The ammonites finally succumbed, as did another group of cephalopods, the belemnites. The bi-
zarre, reef-forming rudistid bivalves disappeared, but evidently did so well before impact event, as did a group of large (up to 1 m), flat clams, the inoceramids. As noted previously, the ammonite extinction is abrupt and coincides with the boundary. The diverse record

of bivalves and gastropods has been exploited for some of the most detailed studies of the evolutionary implica-
tions of mass extinctions. Different groups display very different patterns of extinction selectivity. Despite many claims for increased extinction in the tropics, a detailed study of bivalve extinction patterns found no such pattern (Jablonski and Raup, 1993). Earlier, however, Jablonski showed that survival of bivalves and gastropods, two groups with relatively low rates of extinction, was dependent on the broad geographic range of an entire lineage. During the interval preceding the extinction, the broad geographic range of species (rather than multispecies lineages) and the type of larvae best predicted long-
term survival. This pattern suggests that the factors controlling survival during mass extinction events may be very different than those controlling evolution be-
tween mass extinctions.

Despite the attractiveness of dinosaurs, the scarcity of bones makes them largely unsatisfactory for studies of the pace of the extinction. Northeastern Montana contains one of the best records of vertebrate diversity. Different groups experienced very different extinction patterns. Sharks, marsupial mammals, lizards, and the dinosaurs suffered very high extinction, whereas frogs, salamanders, turtles, placental mammals, and croco-
diles suffered virtually none. Both large and small verte-
brates are part of the list of extinct forms, countering earlier claims that only large vertebrates became ex-
tinct. Clearly, animals in freshwater settings encoun-
tered much less extinction than did fully terrestrial forms. Approximately 90% of freshwater forms sur-

vived but only 12% of the land species survived. Dinosaurs had high extinction throughout their his-
tory, although this is not generally appreciated. The last dinosaurs disappear about 1 m below the extinc-
tion horizon. Although claims for dinosaurs persisting after the end of the Cretaceous have generally been dismissed, there is still considerable uncertainty re-
garding whether dinosaurs were in decline prior to the impact. Recent work indicates that the fossil record of vertebrates, particularly dinosaurs, is simply too poor to resolve the question in a statistically rigorous fashion. A remarkable increase in the abundance of fern spores coincides with the boundary at sites throughout North America and coincides with the iridium spike and the occurrence of shocked quartz. An extinction of almost 80% of plant species has been documented in western North America, but this declines away from this area so that in Antarctica and New Zealand the extinction is almost unrecognized.
Scientists have proposed extraterrestrial causes of mass extinctions for at least 250 years, but only recently have methods of testing such suggestions been available. In 1980, Luis Alvarez and colleagues provided the first definitive test when they announced the discovery of an increased amount of iridium associated with the Cretaceous/Tertiary boundary at Gubbio, Italy. Iridium is more common in material from space than on the surface of the earth so is a good index of the rate of delivery of extraterrestrial material. This iridium anomaly has since been discovered at many other K/T boundary sections, both marine and terrestrial. Other indicators of an extraterrestrial influx have also been recovered, including shocked quartz and other chemical anomalies. The discovery of a crater approximately 200 km in diameter at Chicxulub in the Yucatan gave a significant boost to the impact scenario. The object is believed to have come from the southeast and vaporized.
about 3 km of rock, creating a massive cloud of water vapor and sulfur dioxide. This cloud would cool the earth for a long period and produce extensive acid rain. North America probably felt the greatest effect of the blast, accounting for the higher terrestrial extinctions there. Although smoke and dust would have been produced as well, recent studies indicate that the effects would have ameliorated within a few weeks to months and probably played a minor role in the extinction. Although most attention has focused on the impact at Chicxulub in the Yucatan, the latest Cretaceous was a time of considerable environmental change and many other causal mechanisms have been proposed. Like the end-Permian and Late Triassic mass extinctions, the K/T coincides with a very large flood basalt eruption—in this case, the Deccan traps in India. As with the earlier events, the eruptions could generate acid rain and climatic change. Proponents of this mechanism have been unable to explain many of the features of the K/T boundary that are explained by the impact scenario. The extinction also coincides with a major marine regression, which has also been invoked as a cause (Hallam and Wignall, 1997). Low oxygen levels in the oceans have been proposed, although Hallam and Wignall, generally advocates of such anoxia hypotheses, acknowledge that this is unlikely for the K/T. Currently, the existence of the impact seems indisputable, as does the disappearance of most species at that point. Of continuing interest, however, is the influence of other environmental changes on extinction leading up to the impact. In other words, is the impact solely responsible for the Cretaceous/Tertiary mass extinction, or are other processes also responsible?

II. LESSER MASS EXTINCTIONS

During the past decade, paleontologists have identified many smaller mass extinctions (Hallam and Wignall, 1997; for postextinction recoveries, see Erwin, 1998). Many of these events are not well studied, and a few only appear on large-scale, synoptic analyses and may be statistical artifacts. Several of these events are significant however, and they provide a useful perspective on the larger mass extinctions. They also emphasize the continuum in biotic disturbances between the great end-Permian mass extinction and the common biotic crises which do not qualify as mass extinctions. Only reasonably well-characterized events are noted here.

A. Neoproterozoic and Paleozoic

What happened to the enigmatic, soft-bodied Ediacaran biota of the late Neoproterozoic (575–543 Ma)? This diverse assemblage of probable metazoans is found worldwide and includes a diverse assortment of cnidarian-grade organisms and a few probable ancestors to the clear metazoans of the Cambrian. Virtually all the Ediacaran forms disappear from the record approximately 543 Ma at the base of the Cambrian. Whether this represents a mass extinction, a change in preservation style so that Ediacaran fossil were no longer preserved, or a gradual replacement of these forms by diversifying animals is unclear. A very sharp shift in the carbon cycle occurs at this point, similar to but larger than that which marked the end-Permian mass extinction. This geochemical anomaly is consistent with a mass extinction, although whether one actually occurred remains unclear.

The Early Cambrian is characterized by a diverse assemblage of small shells, spicules and tubes, trilobites, and the sponge-like archaeocyathids. Two pulses of extinction near the end of the Early Cambrian each eliminated approximately 40–50% of marine genera. Reefs, small shelly fossils, and some brachiopods were particularly affected. These extinctions mark the end of this characteristic fauna and paved the way for the trilobite-dominated marine communities which dominate the remainder of the Cambrian. The causes of these events are unclear, but marine anoxia, a decrease in sea level, and possibly changes in atmospheric composition have been suggested. These crises mark the end of reefs until the Ordovician, when the origination of the Paleozoic corals (tabulates and rugosans), bryozoans, and some sponges produced a new suite of reef architectures.

Equally puzzling are a series of crises affecting near-shore trilobites and other marine benthos during the Middle and Late Cambrian in North America. Each event begins with a radiation of new trilobite families across marine shelves, evidently from deeper water ancestors. These new forms rapidly produce a diverse, endemic trilobite assemblage, in which individual taxa have relatively narrow environmental tolerances. The new forms rapidly produce a diverse, endemic trilobite assemblage, in which individual taxa have relatively narrow environmental tolerances, that is then wiped out by a sharp extinction event that cuts across different marine environments. Brachiopods, conodonts, and other marine taxa are also affected. Although the extinctions were initially linked to decreases in sea level, with the diversifications occurring as sea level increased, other paleontologists have suggested temperature decreases or incursions of low-oxygen waters. At least four of these events occurred from the Middle Cambrian into the earliest Ordovician, with the interval between extinctions about 4 Ma, which is approximately half the time thought only a few years ago.
B. Mesozoic

Many Mesozoic mass extinctions have been proposed, but most are small and not particularly well characterized. Hallam and Wignall (1997) note a late Early Triassic event which occurred 5 Ma after the end-Perman mass extinction and particularly affected ammonites, conodonts, and some bivalves; some vertebrates may also have disappeared. Interestingly, the recovery after the end-Perman mass extinction does not really begin until after this episode. Other crises occurred in the early Late Triassic (Carnian) in both marine and terrestrial realms, but these are not well characterized.

The Cenomanian/Turonian (C/T) mass extinction during the Late Cretaceous (93.4 Ma) eliminated approximately 8% of marine families and 26% of marine genera. Statistical analysis of extinction patterns from the western United States suggests the extinction occurred in many discrete steps. Benthic organisms were particularly affected, whereas cephalopods and other groups living high in the water column suffered relatively little extinction. The crisis occurred during warm, equable climates, relatively high sea level, and is associated with marine anoxia and the formation of black shales. In England, for example, the C/T boundary forms a well-known “black band” in quarries. The greenhouse climate evidently led to sluggish oceanic circulation and warm, saline bottom waters, which stratified the oceans and aided in the formation of anoxic conditions (Harries and Little, 1999).

Harries and Little (1999) noted some remarkable similarities between the C/T event and an earlier mass extinction during the early Toarcian (eTo; Early Jurassic). The extinction intensities were essentially identical, as were the environmental settings. Northwestern Europe was particularly affected by this event, although this may reflect the extent of sampling in this area. As with the C/T event, the eTo primarily eliminated benthic taxa, particularly bivalves, gastropods, and brachiopods, as well as many ammonites. In both extinctions diverse faunas were replaced by a much less diverse, stressed fauna dominated by epifaunal bivalves and ammonites.

C. Cenozoic

The Cenozoic experienced many smaller biotic crises, each associated with climatic shifts. The causes of the most recent extinction, that of large mammals during the late Pleistocene, are also associated with human hunting, and whether climate or hunting was the primary cause is the subject of continuing debate.

The late Eocene to early Oligocene was an interval of heightened extinction both on land and in the oceans, although correlation between the two realms remains difficult. Approximately 80% of the foraminifera disappeared, with most of the extinctions coinciding with a sharp shift in the carbon cycle. Both molluscs and echinoids (sea urchins) display a prolonged period of increased extinction, evidently due to a period of global cooling, and this appears to have been the primary factor in the marine extinctions. The same is true on land, where the latest Eocene or earliest Oligocene marks the disappearance of many primitive mammal groups and the rise of modern mammals. Primates and rodents suffered considerable extinction, as did aquatic crocodiles and turtles; the terrestrial tortoises did well, however (Hallam and Wignall, 1997). Tekites, melted and then cooled terrestrial rock ejected from an impact, are well-known in the upper Eocene, leading to suggestions of possibly several extraterrestrial impacts. Re-study of this data has revealed that there is probably only a single layer of tekite material and that no increase in extinction is associated with it.

The most recent mass extinction occurred within the past 100,000 years, and some claim it continues today. Many large mammal species disappeared from Europe, North and South America, and Australia; only in Africa were they unaffected. In North America, mammoths, mastodons, camels, saber-tooth cats, ground sloths, and a host of other animals disappeared approximately 10,000–12,000 years ago. A similar pattern is found in northern Europe. The leading explanation for this extinction is the overkill hypothesis, with human hunting the most important factor. Proponents of this view note that humans became important ecological factors in each area about the time that the pace of extinction increased. Opponents have favored climatically driven extinction, although the nature of the climatic change has been continuously modified to accommodate new information. The overkill hypothesis has also been extensively criticized.

See Also the Following Articles
CARNIC CYCLE • DINOSAURS, EXTINCTION THEORIES FOR • EXTINCTION, RATES OF • EXTINCTION, CAUSES OF • MAMMALS, LATE-QUATERNARY, EXTINCTIONS OF • MASS EXTINCTIONS, CONCEPT OF

Bibliography
MASS EXTINCTIONS, NOTABLE EXAMPLES OF...


MEASUREMENT AND ANALYSIS OF BIODIVERSITY

Wade Leitner and Will R. Turner
The University of Arizona

I. Introduction
II. Definition of Symbols
III. Theoretical Properties of Richness Estimators
IV. Practical Sampling Considerations
V. Estimators Based on Sampling Theory
VI. Estimators Based on Extrapolation
VII. Which Estimation Method to Use?

GLOSSARY

abundance The number of individuals of a given species in a region.

estimator A statistic calculated from data to estimate the value of a parameter.

incidence The number of samples containing at least one of a given species in a census.

sampling scope The area, interval of time, and taxonomic grouping over which sampling takes place.

spatial heterogeneity In community sampling, variation in species capture probabilities across space; may be due to habitat variation, intraspecific clumping, interspecific association, or other factors.

species abundance distribution The number of species found in each interval of abundance in a community.

species accumulation curve A plot of the total number of species observed in a census against some measure of cumulative sampling effort.

species diversity Any of many measures concerning the number of species in an area and/or the distribution of their abundances.

species richness The number of species present in a region.

temporal heterogeneity In community sampling, variation in species capture probabilities over time; may arise from temporal environmental variation, migration, speciation, or other factors.

THE MEASUREMENT AND ANALYSIS OF BIODIVERSITY encompasses sampling strategies and statistical estimation methods for assessing the diversity of species in biological communities and provides the basis for comparison of these measurements across time and space.

I. INTRODUCTION

A. Why Measure Diversity?
Biological diversity pervades every aspect of community ecology. As the key distinguishing feature of communities, it forms the basis of many ecological studies. Many studies have used a wide range of techniques to quantify diversity. These have included studies focused on patterns of species number in both time and space. Projects with more mechanistic aims have employed diversity measures to draw inferences about processes influenc-
ing diversity (e.g., migration, habitat selection, and competition at ecological timescales; and speciation, colonization, and extinction over evolutionary scales). The measurement of diversity applies to studies of the potential consequences of diversity (e.g., stability and resistance to invasion). In addition, diversity measurement plays a critical role in the study of human impacts on biological systems. Its uses in conservation include estimation of extinction rates due to habitat loss, development of conservation strategies, indication of effect due to disturbance, and use as a barometer of ecosystem status. These and other fields of inquiry require adequate methods for quantifying species diversity.

B. Measures of Diversity

Diversity measures are classified into two broad categories: species richness (SR) and shape of the species abundance distribution. The first simply reflects the number of species present in an area. The second effectively measures the probability distribution of population sizes of the species in an area. Evenness and equitability measures belong to the second category. Evenness is highest when all abundances have equal probability in a certain range, whereas equitability is highest when all species have the same population size. Depending on the question under investigation, both categories of diversity measures may be used to document patterns. Many hypotheses concerning patterns in diversity use the relatively simple concept of SR to make concise statements connecting pattern and process. Although richness simply refers to the number of species, it encompasses many sophisticated issues in sampling design and estimation. For these reasons, most studies use species richness. We restrict our focus to species richness as well.

C. The Problem

Species richness, the number of species present, is conceptually the most straightforward of diversity measures. Measuring richness, however, is not so simple. The number of species observed in a sample, \( S_{obs} \), will always tend to underestimate the true species richness because we lack the resources for exhaustive sampling of communities. Even if we had such resources, properties of community structure may change during the course of sampling. This problem is exacerbated by the fact that many biological communities have many rare species, which are unlikely to be detected by sampling efforts. These issues are not unique to ecology.

The problem of determining the number of classes of objects in a collection has a long history. Many different disciplines have sought ways to estimate the number of undiscovered classes from the classes already observed. For example, how many dies were used to mint a collection of ancient coins? How many undiscovered bugs remain in a large computer program? No one method has been successful with all such problems because we often find that we have the most information where it is the least useful. That is, a few classes account for most of the observations, whereas a few observations are scattered over most of the classes. Therefore, we wind up with small sample sizes for the rare observations where we need to characterize variation the most. In the most extreme case, imagine a community in which every species is represented by just one individual. Then, for any sample size below the true number of species we shall always observe fewer than the true number of species. Of course, in this case we could observe that the number of species was a linear function of the number of individuals. This might suggest to us a way to estimate the number of undetected species if we knew how many individuals were present in the study area. This example illustrates the strategy followed by richness estimators: Model the regularity in the behavior of the number of species detected as a function of sample size. Then, use this model to predict the number of species when the sample size becomes large. The regularity could involve a clear relationship between the average sample size and the average number of species observed; this is the basis of extrapolation-based estimators. On the other hand, the regularity could mean that we can replace a complicated sampling process with a more tractable model; estimators based on sampling theory follow this approach.

D. Article Overview

An exhaustive review of all methods used for estimation of species richness is beyond the scope of this article. Therefore, this article reviews the most widely used methods for the estimation of species richness. Section II defines the symbols used in this article. Section III lists some theoretical properties of richness estimators. Section IV discusses practical considerations of using community sampling to estimate species richness. The next two sections present the estimators: richness estimators based on fine-scale, theoretical models of sampling are detailed in Section V, whereas those based on coarse-scale, global modeling of species accumulation (extrapolation techniques) are discussed in Section VI. Section VII addresses the practical problem of evaluating and selecting estimators for use on new data sets.
II. DEFINITION OF SYMBOLS

We adopt the convention that random variables are written as uppercase symbols, whereas lowercase is reserved for deterministic variables or fixed constants. The definitions we shall need are given in Table I.

III. THEORETICAL PROPERTIES OF RICHNESS ESTIMATORS

Salient features of richness estimators include the following:

**Type of data:** Estimators differ in whether they use incidence or abundance data. The incidence of species \(i\) is the number of samples in which \(i\) occurs. Formally, \(Y_i = \sum_{j=1}^{s} x_{ij}\). The abundance of species \(i\) is the number of individuals of \(i\) contained in the census. Formally, \(Y_i = \sum_{j=1}^{s} x_{ij}\).

**Sampling assumptions:** Many estimators make assumptions about the sampling process (e.g., invariance of capture probabilities across all samples in a census). **Parametric/nonparametric:** The estimator \(\hat{s}_\text{parametric} = \phi(X)\) is nonparametric if the function \(\phi\) does not depend on a given distribution of \(X\).

**Bias:** Bias measures average deviation from the true value. We define \(\text{bias} = E[\hat{\phi}(X)] - \phi\). An estimator of \(s\) is unbiased in the strict statistical sense if \(E[\hat{\phi}(X)] = s\) for every \(s\).

**Variance:** An estimator is a random variable. Therefore, we can use variance as a measure of the uncertainty in an estimate.

**Sufficiency:** Let \(w = \phi(X)\) be an estimator of \(s\). The statistic \(W = \phi(X)\) is sufficient for \(s\) if \(p(X \in A | W) = p(X \in A | W, s)\). Suppose \(\hat{s}_\text{parametric} = \phi(X)\) is sufficient for \(s\). Then there are no other estimators (other than functions of \(\phi\)) that we could compute using \(X\) that could increase or decrease our confidence about our estimate of \(s\).

IV. PRACTICAL SAMPLING CONSIDERATIONS

A. Limiting the Influence of Sampling Biases

We lack the resources for complete censusing of communities. Thus, sampling is the window through which we view the ecological world. The species we observe when we sample are determined both by underlying ecological processes and by biases associated with sampling. The simplest of these is bias due to sample size. If we plot the number of species observed in a census against some measure of sampling effort, such as individuals counted, we get a species accumulation curve as shown in Fig. 1. Sample size bias accounts for much of the increase in a species accumulation curve with increasing sampling effort. This bias may be reduced by increasing sample size or by estimating the number of species actually present from incomplete census data. An increase in the accumulation curve also arises from sampling over different habitats and at different times. Spatial and temporal heterogeneity are difficult to quantify, as are their effects on the performance of richness estimators. It may be possible, however, to reduce such effects through careful sampling design. As a general rule, one should minimize sources of variation in capture probabilities over sampling. Where possible, the spatial and temporal scale of a census (see Section IV,
B) should be small relative to the scale of the processes generating diversity.

B. Sampling Scope

Any sampling effort can be described as taking place within some area, over an interval of time, and involving a particular taxon or group of taxa. Even within the area sampled, some habitats may not be sampled. Likewise, within an interval of time, some periods are often not considered (e.g., sampling that takes place over a week but never at night). We define sampling scope as the area (and habitats) sampled, the interval of time (and periods within that interval) over which sampling occurred, and the group of taxa sampled. The consideration of sampling scope is important because it means that the sampling design must be tailored to the question of interest.

The number of species predicted by richness estimators reflects the scope of sampling. In general, these methods estimate only the number of species in the taxa, area, and time interval sampled.\(^1\) For example, estimates from ground-level plant surveys do not predict epiphyte diversity. Similarly, daytime sampling does not predict nocturnal animal species, estimates based on beetle abundance do not predict ant diversity, summer bird diversity does not inform us of winter migrants, and so on.

C. Abundance and Incidence

With quadrat-based sampling, the area to be sampled is partitioned into many subsamples (quadrats) of equal size. A subset of \(i\) of these quadrats is then sampled using capture or observation techniques appropriate for the taxon of interest. The data recorded may take the form of either species incidences \(Y_{ij}\) or abundances \(X_{ij}\).

Abundance data may not be possible to collect in all cases. For some taxa (e.g., some fungi), individuals are not readily recognizable as such. Additionally, some sampling methods may not allow abundances to be recorded. For example, point counts based on bird vocalizations are a practical method of recording species in bird communities but do not allow accurate counting of individuals. Even when abundances can be observed, obtaining abundances may require more effort than obtaining incidences alone. Nonetheless, a wider range of estimators may be used if abundance information is available. Thus, despite the greater effort involved in

\(^1\) There are exceptions. Methods based on extrapolation of accumulation curves, in particular, might be used to estimate species richness at larger spatial (and temporal) scales than that sampled. There are many scaling considerations to be made that are unique to this type of extrapolation.
collecting abundance data, it may be useful to record species abundances so that abundance-based estimators may be used as well.

D. Other Sampling Issues: Surrogates for Quadrats and Species

Two other sampling issues bear mention here. First, it is important to note that the use of areal units (quadrats) for sampling units is not required. Indeed, although quadrats are often appropriate for plants and other sessile organisms, they are less so for mobile organisms. For mobile taxa, it is useful to substitute a more practical measure—such as observer-hours, trap-days, or volume of substrate—as a sampling unit. Additionally, all is not lost if sampling units are of unequal sizes. Equal-sized samples will meet the assumptions for a larger number of estimators. However, some estimators exist which may be applied when it is not possible to keep sample sizes approximately constant over a census.

Second, techniques for estimating species richness are potentially applicable to taxonomic levels other than species. For any of a variety of reasons, it is often impractical or impossible to key all organisms censused to the species level. For example, in studies of paleontological data, it may only be possible to place an individual to the generic level. In other cases, the presence of undescribed species or other lack of knowledge of species status may necessitate the use of a morphospecies concept. In any case, it is important to note that richness estimators will estimate the number of classes (genera and morphospecies in the previous examples) at whatever taxonomic level was used in sampling.

V. ESTIMATORS BASED ON SAMPLING THEORY

Species richness estimators are all alike in the obvious way: They all estimate species richness. However, SR estimators take different approaches to estimation. This section presents approaches involving fine-scale, detailed models of the sampling process. Section VI discusses extrapolative methods based on coarse-scale, global models of species accumulation.

A. Models of the Sampling Process: General Concepts

For the fixed collection from which we sample, \( r_0, \ldots, r_s \) specifies the distribution of abundances among the species. Sampling introduces randomness into our data so that when we census the area, we record \( X_0, \ldots, X_S \) from which we compute \( R_0, \ldots, R_S \). When we sample a set of \( t \) quadrats, we may keep track of the abundance of the \( i \)th species within the \( j \)th quadrat, \( X_{ij} \). Then, we compute the abundance of species \( i \) by \( Y_j = \sum_i X_{ij} \). The distribution of \( X_{ij} \) depends on how we sample individuals.

Suppose we sample from a region (an urn) containing \( n \) individuals (marbles) distributed among \( s \) species (colors). Write \( x_i \) for the number of individuals contained in species \( i \). We collect \( N \) individuals out of \( n \) and \( S_{obs} \) species out of \( s \). If we sample with replacement, we find that the number of observations of species \( i \), written \( X_i \), has probability

\[
P(X_i = a) = \frac{\binom{N}{a} (\frac{x_i}{n})^a (1 - \frac{x_i}{n})^{N-a}}{\binom{n}{N}}
\]

(1)

If we sample without replacement

\[
P(X_i = a) = \frac{\binom{x_i}{a} \binom{n-x_i}{N-a}}{\binom{n}{N}}
\]

(2)

Under certain conditions, both of these distributions can be approximated by the Poisson distribution according to

\[
P(X_i = a) = \frac{(N a_i)^a}{a_i!} e^{-N a_i}
\]

(3)

where \( a_i = x_i/n \). Strictly speaking, the approximation of Eq. (1) is exact only for communities containing infinite species richness and infinite abundance. Practically speaking, the Poisson approximation should work best on large, species-rich communities. A similar argument holds for the approximation of Eq. (2). The requirement that \( n \) be considered infinite is sufficient to get to Eq. (3), but it is not necessary. For example, if the successive times between discoveries of an individual of species \( i \) are exponentially distributed and the number of individuals found in disjoint samples are independent, then we will get Eq. (3). There is a natural way in which this can happen. First, detections of individuals must be independent events. Second, once we know the current value of \( X_i \), the distribution of future values
The number of species observed in a census, $S$, can be modeled with a Poisson process.

### B. Species Observed ($S_{ob}$)

As a starting point, we consider the number of observed species, $S_{ob}$, as our first estimate of $s$. Our presentation of $S_{ob}$ serves as the template for the more complicated estimators to follow. First, we consider the type of data we record. Then, we state the assumptions made in deriving the estimator and its properties. We next turn to theoretical properties of the estimator, for example, its bias and variance. Finally, we discuss issues related to the application of the estimator to real data.

Estimators may be applied to two types of data: species abundance and species incidence. If we record abundance, denoted $X_i$, our data set will be a matrix of the number of individuals of species $i$ captured on sampling occasion $j$.

We define the abundance statistics, $R_i$, to be the number of species containing exactly $i$ individuals.

$$R_i = \sum_{j=1}^t I(X_i = i)$$  \hfill (4)

We define the incidence statistics, $F_i$, to be the number of species occurring in exactly $i$ samples in a census.

$$F_i = \sum_{j=1}^t I(Y_i = i)$$  \hfill (5)

When we wish to refer to the vector of $R$, or $F$, values for a given data set we will write $R$ and $F$, respectively. The number of species observed in a census $S_{ob}$ may be written in terms of either the abundance or the incidence statistics: $S_{ob} = \sum_i R_i = \sum_i F_i$. In some cases, we use only the lower order statistics to correct estimator bias. We can visualize the difficulties inherent in estimator design by considering how information accumulates over the $R$. Let the number of species having fewer than $k$ individuals be $S_k$. Similarly, $N_k$ is the number of individuals representing these $S_k$ species. Therefore,

$$S_k = \sum_{i=1}^k R_i$$  \hfill (6)

$$N_k = \sum_{i=1}^k iR_i$$  \hfill (7)

Figure 2 shows a plot of $N_k$ and $S_k$ vs $k$ for real data collected on mist-netted birds. In Fig. 3, we plot $N_k$ as a function of $S_k$. This plot reveals an apparently exponential relationship between $N_k$ and $S_k$. These figures show that information may be distributed very unevenly in the data set. This means that we shall have few data points (e.g., $N_1$ and $N_2$) where we wish to concentrate our estimation effort ($S_1$ and $S_2$). This is the worst possible distribution of information if we want to estimate the number of rare species from a simple transformation of the data. For example, an estimator using only $E[N]$ and $E[S]$ will probably do poorly for many real data sets.

Suppose we model sampling in terms of the probability of detection of individuals of each species, $q = (q_1, \ldots, q_s)$, and the number of individuals observed, $N$. We assume that each individual of a species has the same capture probability throughout the entire census. Thus,

$$E[S_{ob}(q_1, \ldots, q_s)] = s - \sum_{i=1}^s (1 - q_i)^s$$  \hfill (8)

We record the incidence, $Y_i$, of species $i$ during sample $j$ giving a data matrix whose entries are either $0$, for no detection, or $1$ when any individuals of $i$ get detected on occasion $j$. For an entire census, the number of detections of species $i$ is $Y_i = \sum_j Y_{i}$. Then we may define $p = (p_1, \ldots, p_s)$ to be the vector of species detection probabilities. After $t$ samples, we compute

$$E[S_{ob}(p_1, \ldots, p_s)] = s - \sum_{i=1}^s (1 - p_i)^s$$  \hfill (9)

Equations (8) and (9) show that $S_{ob}$ is biased over finite sampling. In the following, consideration of incidence and abundance versions of $S_{ob}$ parallel on another. Therefore, we focus on the incidence data.

As we increase sample size we have for any $i$, $\lim_{n \to \infty} P[Y_i > 0] = 1$. However, in practical terms, we would like to know just how fast $S_{ob}$ gets close to $s$. Suppose that the $p_i$ are independent and identically distributed (iid) with distribution $h(p)$ and probability density $h(p)$ and consider $E[s - S_{ob}]/s$. The relative rate of convergence of the mean error is given by this expectation taken over all $p$:

$$\frac{s - E[S_{ob}]}{s} = \int [1 - \exp(-p)]h(p)dp \approx (1 - E[p])^s$$  \hfill (10)

When we have few rare species, $E[p]$ will be close to $1$ and $S_{ob}$ will converge quickly to $1$. Note that in absolute terms if $s$ is large we may still have many rare.
FIGURE 2 The accumulation of species and individuals with increasing abundance class size for a bird community in southeastern Arizona (W. Leitner, unpublished data). Note that $N_k$ changes relatively slowly where $S_k$ changes quickly and vice versa.

FIGURE 3 The accumulation of individuals as a function of accumulated incidence for a bird community in southeastern Arizona (W. Leitner, unpublished data). Note that $N_k$ is nearly exponential with $S_k$. We shall have few data points (e.g., $N_1$ and $N_2$) at which we wish to concentrate our estimation effort ($S_1$ and $S_2$).
undetected species. When we have a large proportion of rare species, the convergence will of course be much slower. Now, \( S_{\text{obs}} \) is a random variable and the variance in \( S_{\text{obs}} \) is

\[
\text{VAR}(S_{\text{obs}}) = \sum_{i} \left[ (1 - p)^i - (1 - p)^{i-1} \right] (p^i(1-p)) \quad (11)
\]

The root mean square error will then decrease approximately like the square root of the error in the mean. Perhaps one of the indices of diversity that measure evenness could be used to sharpen confidence intervals in some way. Before undertaking the design of such a method, it would be beneficial to know if any function of \( F \) could be used to increase our confidence in our estimate of \( E[S_{\text{obs}}] \). Burnham and Overton (1978) show that \( \{F_1, \ldots, F_t\} \) are sufficient for \( p \) under the assumption of iid \( p \). Now, the distribution of \( S_{\text{obs}} \) is determined by the distribution of \( p \) because \( S_{\text{obs}} = \sum_{i} F_i \). Therefore, to increase our confidence in our estimator we should focus directly on reducing its variance. Thus, our goals in estimating SR should include getting estimators that decrease the bias faster than \((1 - E[p])^t\) and whose standard errors decrease faster than \(\sqrt{1 - E[p]}\). Often, these goals will conflict with one another. However, knowing that the bias decreases like \((1 - E[p])^t\) tells us that for large enough \( t \), the bias decreases faster than \( t^{-k} \), \( k = 1, 2, \ldots \). Resampling methods can use this fact to accelerate the rate of bias reduction.

C. Resampling Methods (\( S_b, S_d \))

We can very rarely expect

\[
E[f(X)] = f(E[X]) \quad (12)
\]

When \( f(x) \) is linear in \( x \) then Eq. (12) holds. Suppose that \( f(x) \) is nonlinear. Then as \( x \) gets spread away from \( x = E[X] \) the function \( f \) can over- or underweight the contribution of \( X \) to the mean. Thus, both the nature of \( f \) and the distribution of \( X \) contribute to bias. For this reason, outliers (points far from \( E[X] \)) can severely alter the approximation \( E[f(X)] \approx f(E[X]) \). Most estimators derive from just such an approximation. Therefore, when we have outliers, we can sometimes improve our estimate by deleting them from the data set. A version of this approach applied to data not so clearly identified as extreme is the intuitive basis for the jackknife and bootstrap estimators. If the order in which data are recorded does not matter, then we should expect that the relationship between the distribution of \( t \) and \( t - 1 \) data points should inform us about the relationship between the distribution of \( t \) and \( t + 1 \) data points. Thus, the common property that underlies both techniques is the exchangeability of the data. There are two resampling approaches we consider: jackknifing and bootstrapping. In each case, we describe the resampling strategy that gives rise to the estimator. However, in each case it turns out that actual resampling need not be done. The \( F \) statistics contain the same information we would obtain from resampling.

1. Jackknife Estimator (\( S_b \))

Jackknifing involves computing the average value of a statistic on a reduced data set. One removes each combination of \( t \) data points in a data set and computes the statistic of interest to get a set of new pseudostatistics. Then, by taking a suitably weighted average of these statistics we get a reduced-bias version of the original statistic. The number of data points removed at a time gives the order of the jackknife. The most obvious statistic to which we can apply the jackknife is the number of observed species, \( S_{\text{obs}} \). As discussed previously, \( S_{\text{obs}} \) is a biased estimator of \( s \). To apply the jackknife technique to \( S_{\text{obs}} \) we need three assumptions:

1. The capture probabilities may vary across species,
2. The capture probabilities do not change during the census,
3. The bias in \( S_{\text{obs}} \) decreases at least as fast as \( 1/t \).

Under these assumptions, Burnham and Overton (1978) propose a generalized jackknife estimator for species richness, simplified by Smith and Van Belle (1984):

\[
S_b = S_{\text{obs}} - \frac{1}{t} \sum_{i=1}^{t} \left[ \sum_{j=1}^{n} (-1)^{j+1} \binom{j}{i} \left( \frac{t-j}{j} \right)^i \right] \quad (13)
\]

where \( k \) gives the order of the estimator. \( S_b \) does not require that we actually do the resampling. To better understand the properties of \( S_b \), we consider how the first-order jackknife estimate of \( S_{\text{obs}} \), written \( S_{\text{obs}} \), would be produced.

Suppose we compute a statistic, \( D(X_1, \ldots, X_n) = D_k \) from \( t \) observations of \( X \). Now, we want to consider the effect of each observation in turn by removing it from the data set. However, we do not want the order in which we do the removal to matter. Within a given sampling occasion, there may be time of day or other
local differences. For example, during bird banding operations we may consistently catch flycatchers later in the day than warblers because it may take time for flying prey to become active. Then, resampling within a trapping occasion may introduce heterogeneity. Thus, we shall conduct jackknife resampling by adding or removing entire quadrats. Let us remove the $i$th quadrat and compute $D$ without this quadrat. Call this value $D_{i-1}$. Do this for each $i$. The mean of these $D_{i-1}$, $$D_h = \frac{\sum_{i=1}^{t} D_{i-1}}{t}$$ (14)
gives rise to the first-order jackknife estimate
$$D_h = D_t - (t - 1) D_{i-1}$$ (15)
Now, $D_h$ will be less biased than $D_t$ if
$$E[D_h] = s + \sum_{i=1}^{t} \frac{b_i}{t}$$ (16)
where the $b_i$ do not depend on $t$. If the bias of $D_t$ decreases like $1/t$—that is, $b_i = 0$—then the bias in $D_h$ decreases like $1/t^2$.
Suppose $D = \lambda_{\text{obs}}$. If assumption 3 holds, then we can easily get
$$S_h = S_{\text{obs}} + \frac{t - 1}{t} F$$ (17)
The higher order estimates are considerably more complicated but depend on the assumptions in the same way. As we increase the order of the jackknife, we get greater bias reduction. Unfortunately, we can expect that removing data to reduce bias might increase estimator variance. We see from Eq. (13) that $S_{\text{obs}}$ is a linear combination of the $F_i$, with constant coefficients, $a_{ij}$, given $t$ and $h$. Burnham and Overton (1978) give the unconditional variance as
$$\text{VAR}(S_h) = \sum_{i=1}^{t} a_{ij}^2 E[F_j^2] - \frac{E[F_j]^{1-1}}{s}$$ (18)
Note that as we increase the order from $k$ to $k + 1$ the coefficient of $F_{k+1}$ becomes nonzero and therefore increases the variance. Further more, a quick check of the lower order coefficients reveals that they also increase with $h$. Therefore, as we expected, the variance increases with $h$. Thus, we tradeoff bias reduction against increased noise. Two methods have been proposed for choosing the order of the jackknife. The first method tests the hypothesis that $E[S_{\text{obs}} - S_{\text{obs}}] = 0$ using the test statistic
$$T_h = \frac{S_{\text{obs}} - S_{\text{obs}}}{\sqrt{\text{VAR}(S_{\text{obs}}) - S_{\text{obs}}}}$$ (19)
If we can reject this hypothesis for $h = 1$ then we proceed to test successive values of $h$ until we fail to reject. The difficulty with this test is that quadrat information comes in discontinuous jumps so that the order used in estimation may jump around during the course of a census. Burnham and Overton (1978) suggest an interpolation technique that smooths this behavior.

The power of the jackknife technique is that it requires very little knowledge about the distribution of the data. It is thus nonparametric. An argument like the one presented previously on species observed tells us that assumption 3 is reasonable for finite collections of species. The first two assumptions, however, require biological information. For example, suppose we sample a site containing migratory animals. Then, when we sample determines, to a large extent the distribution of number of individuals observed. If we sample across a productivity gradient then we would not expect the distribution of a given species’ abundance, $(X_{11}, \ldots, X_{1t})$, to be symmetric. That is, the quadrats would not be exchangeable. This changes the importance of each quadrat on removal, making some have a deterministically larger effect. Consequently, the effective rate of bias reduction may become much smaller than $1/t$. Nonetheless, with attention to census design, such issues can be minimized. If in addition we assume that the $p_i$ are iid random variables, then Burnham and Overton (1978) show that $F$ is sufficient for the distribution of the capture histories. This result does not depend on the resampling protocol. This assertion will apply to some of the other estimators we examine.

2. Bootstrap Estimator ($S_b$)
Bootstrap estimation begins by constructing a surrogate data set by sampling the original data set with replacement. We shall bootstrap to remove bias in $S_{\text{obs}}$, so we shall once again focus on incidence-based data. Smith and van Belle (1984) analyzed the bootstrapping process to derive
$$S_h = S_{\text{obs}} + \frac{S_{\text{obs}}}{t} \left(1 - \frac{t - 1}{t} F\right)$$ (20)
Just as in the jackknife estimator, we analyze the resampling process to get the estimate that would result if we actually did the resampling. We would like to think of resampling as a surrogate for doing multiple replicate censuses. Let \( s = \text{bias} + E[S_{\text{obs}}] \). If we knew the bias and \( E[S_{\text{obs}}] \) we would be done. To estimate the bias over all possible sampling outcomes, we would need to know the distribution of the \( X_t \). Let \( S_{\text{obs}} \) play the role of \( s \), and let \((X_1, \ldots, X_{t})\) approximate \((s_1, \ldots, s_t)\). Then, if we resample the \( X_t \) multiple times, we can estimate \( S_{\text{obs}} \), the mean number of species found in the replicates. Then bias, \( s = S_{\text{obs}} - S_{\text{obs}} \) is an estimate of \( s = E[S_{\text{obs}}] \), the true bias. Now, we have a single point estimate of \( E[S_{\text{obs}}] \), namely, \( S_{\text{obs}} \). Therefore, we estimate \( s \) by \( S_{\text{obs}} = S_{\text{obs}} + \text{bias} \). We just need to compute bias. We assume that

1. Detection probabilities, \( p_i \), may vary between species,
2. The \( p_i \) remain fixed during sampling.

Then, on resampling we find that

\[
\text{bias} = \sum_{i} \left( 1 - \frac{Y_i}{t} \right) \tag{23}
\]

which leads us to Eq. (20).

The variance in \( S_{\text{obs}} \) given by Smith and Van Belle (1984) is

\[
\text{VAR}(S_{\text{obs}}) = \sum_{i} \left( 1 - \frac{Y_i}{t} \right) \left( 1 - \left( 1 - \frac{Y_i}{t} \right) \right) \tag{22}
\]

\[
+ 2 \sum_{i\neq j} \left( \frac{Y_i Y_j}{t^2} \right) \left( 1 - \frac{Y_i}{t} \right) \left( 1 - \frac{Y_j}{t} \right)
\]

where \( Z_{ij} \) are the number of quadrats that jointly lack species \( i \) and \( j \). For large \( t \) we can replace \( (1 - Y_i)/t \) by \( e^{-Y_i} \). Then, as \( t \) gets large the variance in the estimator goes to zero as long as the tendency of species to co-occur decreases sufficiently fast as \( t \) increases. Therefore, species that interact strongly, for example, core members of a mixed foraging flock, will tend to inflate estimator variance.

### D. Moment Estimators (\( \alpha \), \( S_{\text{M1}}, S_{\text{M2}} \))

These methods all derive relationships between the moments of various distributions. The moments of a distribution, defined by \( E[X^r] \), where \( r \) is a positive integer, are used to simplify the analysis. At a minimum, the relationship between various moments reflects assumptions about the sampling process. In other cases, we require special assumptions about the abundance distribution.

1. Fisher's \( \alpha \)

Fisher's \( \alpha \) is intermediate between species richness and species diversity. It measures species richness but does so in terms of the parameters of the abundance distribution and the sampling process. It is a parametric estimator of richness when applied as described later. Unfortunately, \( \alpha \) involves some rather technical arguments. This has led to two different ways to define \( \alpha \). The first follows from Fisher's use of the gamma distribution to model the abundance distribution. The second results from a limit taken on the functional relationship between the mean sample size and the mean number of observed species in a sample. As we shall see, properties that are exact for one definition of \( \alpha \) are approximations for the other. We shall focus on two key properties of \( \alpha \): (i) sample size dependence of \( \alpha \) and (ii) estimating species richness from \( \alpha \). We begin by defining \( \alpha \), as Fisher did, in terms of the sampling process and the abundance distribution under study. Then, we derive the second version of \( \alpha \) from the first by considering what happens when the distribution under study represents an arbitrarily large number of species. Finally, we consider the practical problem of calculating and interpreting \( \alpha \) when we know very little about the true nature of the abundance distribution.

#### a. Derivation of \( \alpha \)

Fisher's \( \alpha \) was first introduced (Fisher et al., 1943) as a means for estimating the parameters of empirical abundance distributions that had a negative binomial shape. Fisher made three key technical assumptions in his derivation of \( \alpha \) that show up in its properties. First, Fisher assumed that the sample was drawn from a community of infinitely many individuals. As we shall see, this allowed Fisher to use the Poisson process to model the sampling of individuals. Second, he assumed that the community was composed of species whose mean abundances were distributed according to a gamma distribution. Finally, Fisher considered the limiting case of a community containing an infinite number of species.

If we assume that the order in which we discover species is unimportant (i.e., individuals are exchangeable) then \( P(N = x) \) specifies the probability that we observe a species represented by \( x \) individuals. With \( P(N = x) \) we can compute \( E[S_{\text{M1}}] \) and \( E[N] \) because
\[ E[S_{ai}] = s \sum_{x=0}^{s} P[N_i = x] \]

\[ E[N] = s \sum_{x=0}^{s} xP[N_i = x] \]

Fisher begins with the assumption that \( P[N_i = x] = (m/n)^x e^{-m/n} \), where \( m \) is the number of individuals of species \( i \) expected in the sample. The number of individuals in a sample, \( N \), or the sample size, is usually a random variable because trapping and other forms of detection are random processes. Suppose that detection is a Poisson process, or very nearly so, and that our sampling region has extent \( A \). Then, we expect that \( m_i/A = \lambda_i \) where \( \lambda_i \) is the density of species \( i \).

Fisher’s second assumption was that the \( \lambda_i \) possess the gamma distribution with parameters \( 1/p \) and \( k \). The density of a gamma random variable, \( f(x, p, k) \), is given by

\[ f(x, p, k) = \frac{1}{\Gamma(k) p^k} x^{k-1} e^{-x/k} \]

where \( \Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx \). The gamma distribution has three features which make it an obvious first choice. First, gamma random variables must be nonnegative just like species’ densities. Second, the gamma distribution can take on a wide range of shapes. Lastly, computations using the gamma distribution are often tractable.

To spread out densities evenly over a community, \( A \), the gamma distribution can be used in many cases at very small areas. Thus, for smaller sample sizes in smaller \( A \), the number of species from \( s \) to \( s+1 \) is expected to be Poisson-distributed.

Now, if the relative abundance of species is the same at all sampling scales it must be that the distribution’s shape and thus \( k \) are fixed. Although we expect smaller sample sizes in smaller \( A \), the number of species from which we sample remains fixed at \( s \). This means that \( \mu_i \) can only change as a function because of \( A \) or \( p \). Therefore, we collect all the constants from Eq. (28) into \( \alpha \) and write \( \mu_i = \alpha p A \) and set \( \alpha = s \). This is Fisher’s definition of \( \alpha \) (with the effect of sampling extent made explicit). When defined this way, \( \alpha \) depends on only the shape of the distribution and not on the expected sample size. Using this definition for \( \alpha \), we can rearrange Eq. (27) and substitute Eq. (28) to get

\[ 1 - \frac{p_i}{s} = 1 - \frac{p_i}{\alpha s \mu_i} \]

Equation (29) defines \( \mu_i \) as a function of \( p_i \) with parameters \( \alpha \) and \( s \). In fact, Eq. (27) is a species–area relationship. At \( A = 0 \) we expect no species and for \( A \to \infty \) we expect \( s \) species. A little calculus shows that

\[ \frac{\partial p_i}{\partial \ln(A)} = \alpha s \frac{p_i}{\alpha s p_i} \]

Consequently, if we plot \( \ln(s - s_{obs}) \) vs. \( \ln(A) \) we get a line of slope \((\alpha s)/(1/(1 + (\alpha s)/\mu_i))\). Now, \( \mu_i \) quickly exceeds \( s \) in many cases at very small areas. Thus, for smaller most sample sizes we have constant slope \( \alpha s \). Therefore, we may estimate \( s \) if we know \( \alpha \). Unfortunately, this will give only a lower bound on \( s \) because the sensitivity of the slope to changes in \( s \) decreases as \( s \) increases. Nonetheless, we need a method for estimating \( \alpha \) that does not involve \( s \). To do this, we will want to consider communities that contain many species. Again, calculus tells us that if we let \( s \to \infty \) in Eq. (29) we get

\[ e^{-\alpha s} = 1, \quad \mu_i = \alpha s \]

The solution to Eq. (29) gives us the second way of defining \( \alpha \). In order to avoid confusion, we write \( \alpha^* \) for the solution of Eq. (31). We do not need to assume anything about the sampling procedure or the abundance distribution, except that \( \mu_i \) and \( p_i \) exist, in order to compute \( \alpha^* \). Note that \( \alpha \to \alpha^* \) as \( s \to \infty \). Unlike \( \alpha \), \( \alpha^* \) may depend on sample size. This can be seen from

\[ \frac{\partial a}{\partial A} = -\frac{\alpha^*}{\mu_i + \alpha^* p_i - \alpha^* p_i} \frac{\partial \mu_i}{\partial A} \]

(32)
because, in general,

\[
\frac{\partial \mu_s}{\partial a} = \mu_s + a \cdot \frac{\partial \mu_s}{\partial a} \neq 0 \quad (33)
\]

If we apply the assumptions Fisher made about sampling from a gamma abundance distribution, then

\[
\frac{\partial \mu_s}{\partial a} = \frac{\mu_s + a \cdot \frac{\partial \mu_s}{\partial a}}{a} = \mu_s + a \cdot \frac{\partial \mu_s}{\partial a}
\]

\[
\left(1 - \frac{\alpha(\mu_s + a)}{a(\mu_s + \alpha)} \right) \rightarrow 0 \quad \text{as} \quad s \rightarrow \infty
\]

Therefore, the independence of \( a \) from \( \mu_s \) carries over to \( \mu_s \), once again, in the limit of large \( s \). Most often, however, computations of \( \mu_s \) do not in fact use \( \mu_s \) or \( \mu_s \). Rather than average values, the raw numbers of species and individuals observed serve as estimates of the respective expectations. This leads to \( \hat{a} \), an estimate of \( a^* \):

\[
N = \hat{a}(1 - e^{-\frac{\hat{a}}{a^*}}) \quad (35)
\]

Unfortunately, Eq. (33) cannot be solved directly for \( \hat{a} \) in terms of \( S_{\text{obs}} \) and \( N \).

We can get an approximate solution to Eq. (33) by expanding the function \( a(N, S_{\text{obs}}) \) about the point \( (\mu_s, \mu_s) \) in a second-order Taylor series using the fact that \( \hat{a}(\mu_s, \mu_s) = a \). Then, we find the bias is approximately

\[
E[\hat{a}(N, S_{\text{obs}}) - a] = \frac{a^2(\mu_s + a)}{(\mu_s \mu_s + a \mu_s - \alpha \mu_s)^2}
\]

\[
(\mu_s) \operatorname{VAR}(N) + \mu_s^2 \operatorname{VAR}(S_{\text{obs}}) - \mu_s \mu_s \operatorname{COV}(N, S_{\text{obs}}))
\]

and the sample variance of \( \hat{a} \) is

\[
\operatorname{VAR}(\hat{a}) = \left(\frac{a^2(\mu_s + a)}{(\mu_s \mu_s + a \mu_s - \alpha \mu_s)^2}\right) \left(\operatorname{VAR}(N) + \frac{\mu_s + a}{\alpha} \operatorname{VAR}(S_{\text{obs}})ight)
\]

\[
+ 2 \operatorname{COV}(N, S_{\text{obs}})
\]

To this point, we have needed no information regarding abundance distributions or sampling methods. Therefore, any conclusions we can draw from Eqs. (36) and (37) apply to any study system. Let \( \gamma(X) \) be the coefficient of variation of \( X \). Then, even for moderate sample sizes,

\[
E[\hat{a}(N, S_{\text{obs}}) - a] = \frac{a^2(\gamma(N)^2 + \gamma(S)^2)}{\mu_s \mu_s} - \gamma(N)\gamma(S)\operatorname{CORR}(N, S_{\text{obs}})
\]

For moderate \( a \) the sampling process will be nearly Poisson. Therefore, \( N \) will be Poisson because it is a sum of Poissons. Thus, \( \gamma(N) \) will decline with sample size. As we increase sampling effort, \( a \) it must be that \( \gamma_N \) is nondecreasing because we cannot undiscover species. For most collections, the probability of finding new species decreases as we increase sampling effort because we generally find common species first and rare species after prolonged sampling. Thus, \( \gamma(S_{\text{obs}}) \) must also decline with increases in \( a \). From \( \gamma(N) \leq \gamma(S_{\text{obs}}) \leq 1 \) we can see that bias decreases as fast as \( a^2/\mu_s \). This means that \( \hat{a} \) should depend weakly on sample size. Once a sample large enough to get good estimates of the various coefficients of variation has been gathered, \( \hat{a} \) will show little further change. This holds for any abundance distribution for which the variances and means of \( N \) and \( S_{\text{obs}} \) exist. The fluctuations of \( \hat{a} \) can be seen from Eq. (37) to decrease as \( a^2/\mu_s \) as well so that \( \hat{a} \) closely approximates \( a^* \) for large \( a \). However, we should remember that \( \alpha \) expresses a relationship between the means of \( N \) and \( S_{\text{obs}} \). Therefore, we will need more information on the abundance distribution if we are to interpret the number of undiscovered species remaining in terms of \( \hat{a} \).

2. Chao's Abundance-Based Estimator ($S_{\text{AM}}$)

Abundance-based methods apply to data sets in which individuals can be readily identified and counted. From our data we compute the statistics \( (R_1, \ldots, R_n) \) and seek an estimate of \( E[R_1] \), the expected number of unobserved species. Chao (1984) proposed

\[
S_{\text{AM}} = S_{\text{obs}} + \frac{R_1}{2R_2}
\]

as a lower bound for \( s \) in the limit of infinite sample size. \( S_{\text{AM}} \) makes the following two assumptions:

1. Capture probabilities, \( q_j \), may vary among species.
2. Capture probabilities remain fixed during sampling.
where \( q_j \) gives the probability of capture of an individual of species \( j \). The starting point for the derivation of \( S_{AM} \) begins with a model of the sampling process.

Using the Poisson approximation as described previously, we get

\[
E[R] \approx \sum_{i=1}^{N} \frac{(Nq_i)^i}{i!} e^{-Nq_i} \tag{40}
\]

From Eq. (40) we see that this approximation works best when \( i \) is small relative to \( N \) and \( q_i \) is small. Thus, this method should perform best when applied to collections that have relatively few common or abundant species but many rare species. The exponential form of the Poisson approximation permitted Chao (1984) to do some clever analysis using a judiciously chosen distribution function, \( F_{AM}(x) \). The key consequence of the Poisson approximation is that the \( i \)th moment of \( F_{AM}(x) \), written \( \mu_i \), can be used to estimate \( E[R] \) because

\[
\mu_i \approx (i+1)! E[R_{[i]}] \tag{41}
\]

Next, Chao (1984) connects the seen with the unseen species through

\[
E[R_i] \approx E[R] \frac{1}{\int_{x} F_{AM}(dx)} \tag{42}
\]

Once again, we have used the Poisson approximation to estimate the expected number of species. This integral representation reveals two pathways to further analysis: approximate the integrator or approximate the integrand. Chao (1984) approximated the integrator rather than the integrand. By constructing a suitable distribution function that possesses the same first and second moments as \( F_{AM}(x) \), Chao shows that the smallest value taken by Eq. (42) must be \( E[R]/\mu_i \). Using Eq. (41) and taking the \( R_i \) as estimates of \( E[R] \), we find

\[
S_{AM} = S_{un} + \frac{R_1^2}{2R} \tag{43}
\]

The derivation incorporates the following assumptions into \( S_{AM} \):

1. Sampling is a Poisson process.
2. Third-order and higher moments of \( F_{AM} \) may be neglected.
3. The \( R_i \) are good estimates of \( E[R_i] \).

These assumptions determine the statistical properties of \( S_{AM} \). The first assumption could be used to get the sufficiency of \((R_0, \ldots, R_N)\) if the \( q_i \) were iid random variables with distribution \( H(q) \). Then,

\[
P(t_0, \ldots, t_N) = \left( \prod_{i} t_i \right)^{\theta(H)} \tag{44}
\]

where \( \theta(H) = \int H(q) (1 - q)^{-1} h(q) dq \). Paralleling Burnham and Overton (1978), Eq. (44) can be rewritten to permit use of a factorization theorem to assert the sufficiency. If we relax the assumption that the \( q_i \) be identically distributed, then we need a model for selecting \( H(q) \), the distribution function for \( q_i \). For example, if the \( H_i \) corresponded to habitat heterogeneity, then the sufficient statistics would presumably incorporate information on variation in habitat type during sampling. The first and the last assumptions allow us to avoid assuming a form for the abundance distribution. \( S_{AM} \) is therefore nonparametric. Assumption 2 makes \( S_{AM} \) a lower bound for \( s \) that increases as sample size increases, indicating that \( S_{AM} \) is a biased estimator. Assuming that we have a large enough sample size, we expand the function \( S_{AM} \) as a first-order Taylor polynomial about \( E[S_{un}], E[R] \), and \( E[R_i] \). Then we can use

\[
\text{VAR}(ax + by) = a^2 \text{VAR}(x) + b^2 \text{VAR}(y) + 2ab \text{COV}(x, y) \tag{45}
\]

to estimate the variance. Working through the analysis and assuming the covariance terms are small gives

\[
\text{VAR}(S_{un}) \approx \left( 1 + \frac{R_1}{R} \right) \text{VAR}(R) + \frac{(1 - R_1)^2}{2R} \text{VAR}(R_i) + \sum_{i} \text{VAR}(R_i) \tag{46}
\]

Even when we have the abundance distribution, the indicated variances can be challenging to compute. Chao (1984) used bootstrapping techniques coupled with simulations and tests on well-characterized data sets to evaluate the performance of the \( S_{AM} \). When applied to data sets with known \( s \), \( S_{AM} \) generally had smaller bias and narrower confidence intervals than \( S_1 \) or \( S_2 \).

3. Chao’s Incidence-Based Estimator \( S_{IM} \)

Incidence-based methods apply to data sets in which individuals are not readily identified or counted. We compute \((F_1, \ldots, F_s)\) from the data matrix \( X_1 \) and seek an estimate of \( E[F_1] \), the expected number of unob-
served species. Chao (1987) used the same approach as used for $S_{AM}$ to derive

$$S_{AM} = S_{AM} + \frac{F_{1}^{\frac{1}{2}}}{2F_{2}} \quad (47)$$

for data sets of species incidence. Like the jackknife estimator, the original application of this method was to estimate population sizes. It produces a lower bound for $s$. The starting point for the derivation of $S_{AM}$ begins as $S_{AM}$ did with a model of the sampling process. Let $p_i$ be the probability of detection of species $i$. In this case, sampling is as in the jackknife estimator. Therefore, we assume

1. Detection probabilities, $p_i$, may vary between species.
2. The $p_i$ remain fixed during sampling.
3. The $p_i$ are iid random variables with distribution $H(p)$.

As we saw for $S_{AM}$, if these assumptions are met then $(F_1, \ldots, F_n)$ will be sufficient for $H$. The last assumption allows us to write

$$E[F_i] = \int_0^1 \left( \frac{1}{i} \right) p(1-p)^i H(p) dp \quad (48)$$

From here we can once again use the Poisson approximation to simplify Eq. (48). In this case, the focus on data sets that concentrate probability over $F_1$ and $F_2$ is less critical. The Poisson approximation will still work well if the probability of large $p_i$ is small. From here, Chao (1987) parallels the derivation of $S_{AM}$ using an absolutely continuous version of $F_{AM}(x)$. This distribution function, $F_{AM}(x)$, has moments that satisfy

$$\mu_i = \left( i + 1 \right) \frac{E[F_i]}{E[F_1]} \quad (49)$$

An integral relationship between $F_2$ and $F_1$ analogous to Eq. (42) once again leads us to approximate either integrand or integrator. While $S_{AM}$ arises from approximating the integrator as before, we gain some insight into the relationship between $S_{AM}$ and $S_{IM}$ by considering the other alternative first. If we approximate the integrand, $1/x$, with a $k$th order polynomial then we find

$$E[F_i] = \sum_{i=1}^n \left( -1 \right)^{i+1} \binom{k}{i} E[F_i] \quad (50)$$

Now, the bias should decrease as the expected number of undiscovered species decreases. This expectation, given by

$$E[F_i] = \int_0^1 \left( 1 - p \right) H(p) dp \quad (51)$$

must decrease at least as fast as $1/t$ because $(1 - p)^t \leq e^{-t}$ for $0 \leq p \leq 1$. Equation (50) has, as Chao (1984) points out, the same form of the $k$th-order jackknife estimator. The connection between $S_{AM}$ and $S_{IM}$ reflects the assumption made in any jackknife estimator that the bias can be expressed as a power series in $1/t$. Chao (1987) obtains $S_{AM}$ by approximating the integrator rather than the integrand. Then, using an analog to Eq. (41) and taking the $F_i$ as estimates of $E[F_i]$, Chao proposes

$$S_{AM} = S_{AM} + \frac{F_{1}^{\frac{1}{2}}}{2F_{2}} \quad (52)$$

As with $S_{AM}$, we have several assumptions built into the derivation of $S_{IM}$:

1. The $p_i$ are small enough to justify the Poisson approximation to the binomial probabilities of Eq. (48).
2. Third-order and higher moments of $F$ may be neglected.
3. The $F_i$ are good estimates of $E[F_i]$.

First, we note that $S_{IM}$ is nonparametric. Assumption 2 makes $S_{IM}$ a lower bound for $s$ that increases as sample size increases, indicating that $S_{IM}$ is a biased estimator. Chao (1987) estimates the variance of $S_{IM}$ as

$$\text{VAR}(S_{AM}) = \frac{F_1}{4} \left( \frac{F_1}{F_2} \right) + 4 \left( \frac{F_1}{F_2} \right)^2 + 5 \left( \frac{F_1}{F_2} \right)^3 \quad (53)$$

Equation (53) can be used to get approximate confidence intervals for $S_{AM}$ provided that $p_i$ is not too large.

As we increase $t$, we expect the lower order $F_i$ to go to 0 if we are sampling a finite collection of individuals. This would seem to imply that the variance goes to zero with increasing $t$. However, large $t$ means large $p_i$ and so makes Eq. (53) a poor approximation. Chao (1987) used real data sets with known $s$ to evaluate $S_{AM}$, finding that it performed well relative to $S_{AM}$. For low $t$, $S_{IM}$
displayed smaller standard error and less negative bias than $S_J$. Notably, the data sets tended to have more recaptures with increasing $t$. Chazdon et al. (1998) found similar results.

E. Estimates Derived from Sample Coverage ($S_{AC}$, $S_{IC}$)

Sampling of a finite community must eventually capture each individual multiple times. Thus, we can expect that exhaustive sampling would lead to $F_i = 0$ and $R_i = 0$ for the lower order $i$. An alternative that measures representation of species in the data set while retaining analytic tractability is sample coverage. The sample coverage, $C$, is defined by

$$C = \sum_{j=1}^{s} p_j I(Y_j > 0) \text{ for incidence data}$$

Chao and Lee (1992) propose a nonparametric estimate of $s$ using estimates of $E[S_{obs}]$, $E[C]$, and $E[q]$. They define the abundance coverage-based estimator $S_{AC}$ as follows:

$$S_{AC} = \frac{E[S_{obs}]}{E[C]} = \frac{E[R]}{E[C]} \cdot (q^{\gamma})$$  \hspace{1cm} (54)

Intuitively, we see that $S_{obs}$ is inflated by $C$ to adjust for the fraction of the distribution left uncovered. Additional bias correction reflecting the shape of the distribution comes through $R$ and the $\gamma$ of the distribution. Thus, we are using information from more than just the lower order abundances and may expect $S_{AC}$ to perform better than $S_{AM}$ and $S_{IM}$ when higher order abundances support significant probability. We start with the same two basic assumptions that apply to $S_{obs}$:

1. Individual detection probabilities, $q_i$, may vary between species.
2. The $q_i$ remain fixed during sampling.

Thus, the $q_i$ can be regarded as a sequence of random variables with mean $q$, variance $\sigma^2_q$, and coefficient of variation $\gamma_q$.

Then, we find that

$$E[S_{obs}] = s - \sum_{i=1}^{n} (1 - q_i)^{y_i}$$

$$E[C] = 1 - \sum_{i=1}^{n} (1 - q_i)^{y_i}$$

Of course, if we knew the true values of the $q$ we could compute $s$ directly. Instead, we get estimates of $E[S_{obs}]$ and $E[C]$ from our data. Now, $E[S_{obs}]$ and $E[C]$ are functions of $s$ and the $q$. Therefore, we would like to eliminate the dependence on the $q$ in order to estimate $s$. If our estimates of $E[S_{obs}]$ and $E[C]$ are close to the truth, we can use Taylor’s theorem to expand about the point $(q_1, \ldots, q_n)$ to get

$$s \approx \frac{E[S_{obs}]}{E[C]} = \frac{E[R]}{E[C]} \cdot \gamma_q$$  \hspace{1cm} (55)

Now, all we need are estimates of the expectations and the coefficient of variation for the abundance distribution. Chao and Lee (1992) suggest

$$E[S_{obs}] \sim S_{AC}$$  \hspace{1cm} (56)

$$E[R] \sim R_i$$  \hspace{1cm} (57)

$$E[C] \approx C = 1 - \frac{R}{N}$$  \hspace{1cm} (58)

for the various expectations. The coefficient of variation is somewhat more complicated. First, it must always be positive. Second, when the true $\gamma$ is large, we can expect to need more data to estimate it. It can be easily shown that

$$\gamma^2 = \frac{1}{s} \sum_{i=1}^{n} \frac{q_i^2}{q_i} - 1$$  \hspace{1cm} (59)

and

$$E \left[ \frac{\sum_{i=1}^{n} i(i-1)R^2}{\sum_{i=1}^{n} R} \right] = n(n-1) \sum_{i=1}^{n} q_i$$  \hspace{1cm} (60)

Chao and Lee (1992) then suggest that

$$\gamma = \max \left\{ \frac{N}{N} \frac{\sum_{i=1}^{n} i(i-1)R^2}{(\sum_{i=1}^{n} R)^2} - 1, 0 \right\}$$  \hspace{1cm} (61)
where $N_0 = S_{\infty}/C$ when $y^2 \leq 0.5$. Then we have

$$S_{\infty} = \frac{S_0}{C} - \frac{R_s}{C} y^2$$  \hspace{1cm} (62)$$

In Eq. (61) the term $N_0$ is an initial estimate of $s$. If $y^2 > 0.5$ the bias can be reduced by using $S_{\infty}$ in place of $S_{\infty}/C$ to yield

$$y^2 = \max \left\{ \frac{\sum_{i=1}^{N} (i-1)R_i}{\left(\sum_{i=1}^{N} R_i\right)^2} - 1, 0 \right\}$$  \hspace{1cm} (63)$$

Then, for larger $y$ we have

$$S_{\infty} = \frac{S_0}{C} - \frac{R_s}{C} y^2$$  \hspace{1cm} (64)$$

It turns out that $S_{\infty}/C$ is biased low when the $q_i$ differ from one another [when sample sizes are large enough to permit estimation of $y(q_i)$]. Thus, $S_{\infty}$ should be biased low. The variance of $S_{\infty}$ is approximately a linear combination of $\text{COV}(R_i, R_j)$. It can be shown that

$$\text{VAR}(R) = E[R]\left(1 - \frac{E[R]}{s}\right)$$  \hspace{1cm} (65)$$

$$\text{COV}(R_i, R_j) = \frac{E[R_i E[R_j]]}{s}$$  \hspace{1cm} (66)$$

Now $R \leq s$ so the variance is bounded and decreases with increasing richness. However, for a fixed $s$, variance depends on how the $R_i$ spread out as we increase $i$. This in turn depends on the sampling model we use. Here, we have assumed that the $q_i$ remain fixed so that we sample with replacement. Therefore,

$$E[R_i] = \sum_{j=1}^{N} \left(\frac{N}{N-j}\right) p_i (1-p_i)^{N-j}$$  \hspace{1cm} (67)$$

Evidently, as $N$ increases, $R_i \rightarrow 0$ with probability 1. Now, the $R_i$ represent $N$ different abundance classes. However, we have only $s$ species. Therefore, the probability that $R_i > 0$ will be spread out over an increasing set of the $R_i$ as $N$ increases. Thus, $E[R]$ should become small relative to $s$ as $N$ increases. The simulations performed by Chao and Lee (1992) indicate that at high sample coverage $S_{\infty}$ is nearly unbiased. However, Colwell and Coddington (1994) found that $S_{\infty}$ tends to overestimate $s$ when sample sizes are small. The difficulty appears to occur for large values of $y$. In this case, many of the individuals in a sample represent common species, so there are a few dominant $q_i$ in the collection. In this case, $C$ is a poor estimate of sample coverage. However, we can easily ascertain the number of common species in a collection. Chao et al. (1993) suggest truncation of the data set to species with $k$ or fewer representatives. The new estimates for $E[C]$ and $E[S_{\infty}]$ that result are

$$S_{\text{obs}} = \frac{S_{\infty}}{C} + \frac{R_s}{C} y^2$$  \hspace{1cm} (68)$$

$$\hat{C}_s = 1 - \frac{R_s}{\sum_{i=1}^{N} R_i}$$  \hspace{1cm} (69)$$

Similarly, we have

$$y^2 = \max \left\{ \frac{\sum_{i=1}^{N} (i-1)R_i}{\left(\sum_{i=1}^{N} R_i\right)^2} - 1, 0 \right\}$$  \hspace{1cm} (70)$$

so that

$$S_{\text{obs}} = S_{\infty} - S_{\text{obs}} + \frac{S_{\text{obs}}}{C} + \frac{R_s}{C} y^2$$  \hspace{1cm} (71)$$

2. Chao’s Incidence-Based Coverage Estimator ($S_{\text{IC}}$)

The incidence-based coverage estimator, $S_{\text{IC}}$, applies to data sets that have the quadrat structure outlined for $S_{\infty}$. Briefly, when we have the presenceabsence of data then incidence-based estimators apply. Lee and Chao (1994) found ways to relax many of the assumptions present in $S_{\text{AM}}$, $S_{\text{IM}}$, $S_{\text{JK}}$, and $S_{\text{AC}}$. They follow Pollock (1976) by partitioning variability into between-species, within-species, and temporal components. The results involve more complex notation and more complicated proofs but use the same conceptual framework as $S_{\text{AC}}$. For brevity, we present only the case of unequal detection probabilities among species. The estimator for incidence-based coverage is

$$S_{\text{IC}} = \frac{E[S_{\text{obs}}]}{E[C]} + \frac{E[S_{\text{obs}}]}{E[C]} \left(1 - p\right)^3$$  \hspace{1cm} (72)$$

We define $p_i$ as the probability of detection of species $i$. For $S_{\text{IC}}$ we assume that

1. Detection probabilities, $p_i$, may vary between species.

2. The $p_i$ remain fixed during sampling.
Thus, the $\pi_i$ can be regarded as a sequence of random variables with mean $\bar{p}$, variance $\sigma^2_p$, and coefficient of variation $\gamma(p)$. Then, we find that

$$E[S_{\alpha}] = s - \sum_{i=1}^{s} (1 - p_i)$$

$$E[C] = 1 - \frac{\sum_{i=1}^{s} (1 - p_i)}{\sum_{i=1}^{s} p_i}$$

$E[D]$ and $E[C]$ are functions of $s$ and the $p_i$. Therefore, we would like to eliminate the dependence on the $p_i$ in order to estimate $s$. We can use Taylor's theorem to expand about the point $(\bar{p}, \ldots, \bar{p})$ to get

$$s \approx E[S_{\alpha}] + E[F] = E[C] + \gamma(p)$$

Using the approximations

$$E[S_{\alpha}] = S_{\text{obs}}$$

$$E[F] = F_1$$

$$\hat{C} = 1 - \frac{F_1}{\sum_{i=1}^{t} F_i}$$

$$\hat{C} = 1 - \frac{(t-1)F_1 - 2F}{(t-1)\sum_{i=1}^{t} F_i}$$

and

$$\gamma(p) = \max \left\{ \frac{\sum_{i=1}^{t} (1 - F_i)}{(\sum_{i=1}^{t} F_i)^2} - 1, 0 \right\}$$

where $\hat{N}_i = S_{\alpha}/\hat{C}$ when $\gamma(p) \leq 0.5$, we get

$$S_{\text{IC}} = \frac{S_{\text{obs}} + F_1}{\hat{C}} \gamma(p)$$

If $\gamma(p)$ is large, the bias can be reduced by using $S_{\text{IC}}$ in place of $S_{\alpha}/\hat{C}$ to compute $\gamma$. Then, using $\hat{C}$ for $\hat{C}$, we have

$$S_{\text{IC}} = \frac{S_{\text{obs}} + F_1}{\hat{C}} \gamma(p)$$

As $\hat{C}$ increases, the bias must decrease. Therefore, we can expect that the assumptions of fixed $p_i$ will lead to $S_{\alpha}$ underestimating $s$ in most cases. The variance of $S_{\alpha}$ is approximately a linear combination of $\text{COV}(F_i, F_j)$. It can be shown that

$$\text{VAR}(F_i) = E[F_i] \left( 1 - \frac{E[F_i]}{s} \right)$$

$$\text{COV}(F_i, F_j) = \frac{E[F_i]E[F_j]}{s}$$

The variance is thus bounded and decreases with increasing richness because $F_i \leq s$. Now, as we increase $t$ each of the species will be detected a different number of times with probability 1 (even if they have the same $p_i$). Therefore, there will be at most $s$ nonzero $F_i$ among $(F_1, \ldots, F_t)$. Thus, $E[F_i] \Rightarrow 0$ as $t \Rightarrow \infty$. As with $S_{\alpha}$, the effects of a few dominant $p_i$ can be adjusted for using the methods of Chao et al. (1993). If we truncate the data set to species with $k$ or fewer representatives, the new estimates for $E[C]$ and $E[D]$ that result are

$$S_{\alpha,k} = \frac{\sum F_i}{k}$$

$$\hat{C}_k = 1 - \frac{F_1}{\sum_{i=1}^{k} F_i}$$

Similarly, we have

$$\gamma_k = \max \left\{ \frac{S_{\alpha,k} \sum_{i=1}^{k} (1 - F_i)}{(\sum_{i=1}^{k} F_i)^2} - 1, 0 \right\}$$

so that

$$S_{\text{IC},k} = \frac{S_{\text{obs}} + F_1}{\hat{C}_k} \gamma_k$$

F. Estimation on Heterogeneous Samples

Increasing sample size often improves estimator performance through the operation of the laws of large numbers. We get better estimates because the means of many summary statistics become closer to the truth and the variation in these estimates decreases with increasing sample size. Of course, increasing sample sizes increases the cost of performing the census. However, a potentially more serious problem accompanies increases in sample size. In most cases, larger samples take more time or cover larger areas. Inevitably, this means increasing heterogeneity in the sampling process. For example, seasonal variation or habitat heterogeneity will cause detection probabilities to change within a species.
as sampling progresses. A similar problem occurs in estimating the population size of a single species. Pollock et al. (1984) and Lee and Chao (1994) modeled the contribution of heterogeneity in individual capture probabilities to population size estimates. Just as most of the estimators presented so far have been adapted from population size estimators, Nichols et al. (1998) applied Pollock's robust estimation method to species richness estimation. These methods incorporate many parameters to characterize the effect of heterogeneity on capture probabilities. Consequently, the estimators are substantially more complicated. Although we may better adjust an estimator to the needs of a given census, the added complexity tends to obscure the overall relationship between sample size and species richness.

VI. ESTIMATORS BASED ON EXTRAPOLATION

In order to approximate exhaustive sampling or to elucidate large-scale patterns, many studies extrapolate beyond the extent of a given census. These studies focus less on the sampling process and more on biological processes that operate over scales that are too large to survey in detail. As noted previously, studies that increase the scope of sampling usually encounter a wider range of heterogeneity. Heterogeneity increases due to the responses of individual species to the range of environmental and biological conditions as we increase the study area or the time span of the sample. As we increase sample size, we expect that the correlation in the abundance of arbitrarily selected species should decrease. This decreasing correlation permits us to average over the fine-scale, detailed distribution of the components of \( S_{\alpha} \), and focus on the coarse-scale, global behavior of \( S_{\alpha} \). When most species in large communities interact weakly with most other species, our experience with the laws of large numbers suggests that there should be a trend relating species richness to area or time span. If the interactions are weak enough relative to sample size, the distribution of fluctuations about the trend may display regular, perhaps even Gaussian, behavior. Our models of species accumulation can then be built with two pieces of information. First, we shall need the average rate of increase in richness as a function of sample size. Second, we need the variability about the mean. The first piece gives us a trend we can extrapolate. The second allows us to fit the model to the data. Most extrapolation methods focus on the trend model and assume a form for the noise that makes curve fitting convenient.

Once we have a suitable model, we can then extrapolate beyond the areas in which we actually sample. We must keep in mind that this extension confounds the contribution of ecological process and sampling effort to diversity. At smaller scales, sampling effort dominates the accumulation of diversity. However, at larger scales, we encounter heterogeneity effects that may scale differently than the effects of sampling effort. Often, we have no general guidelines for selecting one of the several functional forms for species accumulation. Most of the difficulty in applying these models stems from fitting the nonlinear models to data. For brevity, we omit the technical details of curve fitting. We briefly consider a few representative forms here. These are species–area curves, birth chains, and the Michaelis–Menten equation.

To evaluate extrapolation models, we need at least three pieces of information. First, is the trend nondecreasing? Even though it may seem counterintuitive, there are models whose rate of species accumulation increases as small sample sizes (Leitner and Rosenzweig, 1997; Soberón and Llorente, 1993). Second, does the trend have an asymptote? Third, what curvature does the functional form of the model predict? The fitting of the model to data depends on the overall shape of the function. Less obviously, the model transforms the noise that gives rise to the residuals used to evaluate the goodness of fit.

A. Species–Area Curves

Species–area curves may use samples from disjoint areas or nested areas. This approach has a long history due primarily to its flexibility. There are several functions that may be fit to the data in order to extrapolate to a large area. Chief among these are \( \log(S_{\alpha}) \) vs \( \log(N) \) and \( S_{\alpha} \) vs \( \log(N) \). Note that these curves have no asymptote. Furthermore, their curvature is often too shallow to give the best fit to the data. The key parameters of interest are the slope (the \( z \)-value) and intercept of the plot. The slope in particular has drawn much attention in both theoretical and applied circles. Given the abundance distribution, one may use methods similar to Fisher’s approach in deriving a to make theoretical predictions about the magnitude of the slope. Preston (1962) and May (1975) both claim theoretical predictions for \( z \)-values built solely on the assumption of the lognormal distribution abundances. This work implied that the species–area curve, at least for lognormal distributions, reflected nothing more than sampling. However, this conclusion rests on the tacit assumption that the abundance distribution be lognormal.
at all scales. This assumption cannot be achieved without invoking ecological process. Leitner and Rosenzweig (1997) demonstrate that realistic species–area curves must include information about spatial heterogeneity. The curves obtained in this work did have finite asymptotes and variable curvatures. This work does not, however, provide insights into the selection of log–log or log–linear models.

B. Birth Chains

Many different functions can exhibit similar asymptotic behavior and curvature. Suppose we select the candidate functions that fit the data best. Functions that give equally good fits to the data can make very different asymptotic predictions. One way to select among these candidate models would be to choose the model that best captures the mechanisms of species accumulation. Soberón and Llorente (1993) use the framework of birth chains for this purpose. Let \( S(t) \) be the number of species observed by time \( t \). Then \( S(t) \) is a birth chain if

\[
\begin{align*}
P[S(t) + dt] - S(t) = 1[S(t)] = \lambda(S(t), dt) \quad (87) \\
P[S(t) + dt] - S(t) > 1 = 0 \quad (88) \\
P[S(t)] = 0 = 1 \quad (89)
\end{align*}
\]

This set of conditions means that the time between discoveries of new species is exponentially distributed. However, the rate at which new species occur depends on the time spent sampling and the number of species already found. The rate function, \( \lambda(S, t) \), captures the details of the mechanism of species discovery. For example, suppose we assume that the rate of discovery declines linearly with species discovered. Then \( \lambda(S, t) = a - bS \). Soberón and Llorente (1993) show that

\[
E[S(t)] = \frac{a}{b}(1 - e^{-bt}) \quad (90)
\]

This method facilitates curve fitting because it allows us to derive expressions for the variance of \( S(t) \). Existing methods that lack mechanistic foundations may be justified using this approach. For example, Soberón and Llorente (1993) provide insights into the popular Michaelis–Menten equation using this birth chain model.

C. Michaelis–Menten

The Michaelis–Menten equation has often been fit to species accumulation data. It is a widely known function that has a finite asymptote, is nondecreasing, and accommodates a wide range of curvatures. The equation states that

\[
S_m(t) = \frac{at}{1 + bt} \quad (91)
\]

Soberón and Llorente (1993) use a birth chain model to suggest a mechanistic basis for the Michaelis–Menten equation. The birth chain model assumes that the number of new species found is an inhomogeneous Poisson process with rate

\[
\lambda(S, t) = a + \frac{bt}{a + bt} - \frac{b}{a} \quad (92)
\]

where \( j \) is the number of species found by time \( t \) and \( a \) and \( b \) are constants to be fit to the data. Soberón and Llorente (1993) argue that if the chances of adding a new species improve with effort, to some upper limit, then a form such as Eq. (92) obtains. The key drawback with this and other curve-fitting models is that we do not have a distribution for the changes in \( S_m \). This means that we cannot rationally assign different weights to the data during fitting.

VII. WHICH ESTIMATION METHOD TO USE?

Many estimators have been put forth, leaving the investigator with the practical question of which estimators are appropriate for use with a given new data set. Two facts may give us confidence a priori in an estimator’s applicability to a new data set. As mentioned previously, each estimator makes assumptions about properties of the community being sampled (e.g., variation in capture probabilities across species or samples) and the sampling strategy used. If the community sampled and the sampling procedures used satisfy the assumptions of the estimator, then we can have some confidence in applying that estimator to a new data set.

However, meeting a method’s explicit assumptions is not always sufficient. Another difficulty with relying solely on the theoretical basis of estimators is the fact that one or more assumptions are often violated in the sampling of real communities. Thus, theoretical support alone may not justify the use of a particular estimation method.

A second way to have confidence a priori in an estimator is to know its performance on similar communities sampled using similar methods over similar scales.
Evaluating estimator performance on data allows testing implicit assumptions. Also, many estimation methods (e.g., extrapolation) have little biology built in. They exist because they seem to work in a few cases, and our confidence in them may come only from verification with relatively complete data sets. One should take care, however, not to conclude too much about an estimator based on its performance on single data sets. Rather, only through comprehensive testing on real data sets from a variety of taxa, locations, and spatial/temporal scales, as well as testing on simulated data, will we gain insight into estimator performance.

A. Estimator Evaluation

1. Evaluation on Real Data and Simulated Data

The testing of estimators has used data from real communities and from computer simulation. Each type of data has advantages and drawbacks. Real data take considerable effort to obtain. As a result, replication of real data sets is difficult. Simulated data sets, on the other hand, can be turned out in comparatively large numbers. This allows replication of data and testing of various sampling strategies on the same data.

To test estimators effectively with data, we must have some idea of the true parameters underlying our data (otherwise, how do we know what estimators should predict?). With simulated data, each parameter's true value is known because it is either specified by the investigator or is directly measurable from the simulated community. With real data, the truth may be approximated by testing estimators on smaller sets of data from communities in which the parameters are reasonably well-known from separate, intensive sampling.

Finally, there are many properties of ecological communities which are likely to influence estimator performance. These include factors such as species richness, various properties of the species abundance distribution, intraspecific clumping, and interspecific associations. Each of these can vary widely. Simulated data allow us to explore large parts of this parameter space. More important, with simulated data we can cover the range corresponding to anticipated field conditions. This allows us to test and select estimators before taking them into the field. Nevertheless, only real data can tell us which parts of the vast parameter space are relevant. Thus, we cannot ignore either approach. The biological reality of actual data is complemented by the versatility of simulated data.

Some progress has been made in the evaluation of estimation methods with data. Every such study cannot be detailed here, and not enough work has been done to allow generalizations to be made. Chazdon et al. (1998) provide one useful example of an approach combining both real and simulated data. They examined the performance of eight estimators using young woody regeneration in six tropical forest sites. Noticing that patchiness (intraspecific clumping) varied across sites, they then resampled the data sets to create simulated data sets with a range of patchiness levels. Testing estimators on these simulated data sets revealed the effect of patchiness on the performance of various estimators.

2. Estimator Evaluation Criteria

Clearly, one comprehensive study of the performance of all estimators over all possible data is impossible. Evaluation of estimators, then, will necessarily be done by many investigators with different communities and over different ranges of simulated data. To facilitate this cooperative approach, we must identify a common set of evaluation criteria that can be used to measure the performance of an estimator on a data set. These criteria are slightly different from theoretical properties of estimators. Numerous such criteria have been put forth, although they differ in calculation methods, they are often variants of a few common themes, which are listed as follows:

- **Bias**: As defined previously, bias measures deviation from the true value. Common measures of bias include \( E[\hat{S}_n] - S \), or deviation from \( S \), and \( E[\hat{S}_n] - s \). Relative deviation from \( S \) in the evaluation of estimators with data, bias can be calculated as a mean across replicate data sets or can be assessed with increasing sample size within a data set.

- **Variance**: As discussed previously, variance measures uncertainty in an estimate. If we have high variance, we can have little confidence that a single observation is a good indication of the mean. Variance may be calculated numerically over replicates, or an analytical estimator of variance may be used. These measures can also be expressed as confidence intervals and may be used in hypothesis testing when comparing estimated values. Note that this measure does not depend on \( S \), which is considered by other criteria.

- **Sample size independence**: This measures the rate of convergence of an estimator’s mean to its asymptotic value. An estimator that rapidly approaches its asymptotic value requires less sampling effort to obtain an equally good estimate. Such an estimator is relatively...
sample size independent. This measure, too, does not depend on \( s \).

Note that the previous three criteria may be calculated over multiple realizations of the same underlying parameters (e.g., the capture probabilities \( q \)). The following two may only be considered among multiple data sets in which these parameters vary.

**Correlation with \( S \):** Since \( s \) is a fixed parameter, we introduce \( S \) as a random variable representing the true richness of any of a number of data sets, between which \( S \) and capture probabilities \( q \) may vary. If the correlation between an estimator \( S_{\text{est}} \) and true richness \( S \) across these data sets is high, \( S_{\text{est}} \) may be useful in comparing SR between sites or census occasions (Palmer, 1990). Note that correlation with \( S \) does not require that an estimator have low bias. Rather, correlation reflects a relative deviation of \( S_{\text{est}} \) from \( S \) that is somewhat constant across data sets.

**Robustness:** A robust estimator is one for which performance (as measured by any of the previous criteria) changes little across a range of data sets.

It is important to note that the relative importance of each of these criteria depends on the problem of interest. In measuring absolute species richness, minimizing bias may be a priority. In comparisons across time or space, variance and correlation with \( S \) may be more important, whereas bias may become more tolerable.

A final criterion represents a minimum requirement of sorts for estimator performance:

**Beating \( S_{\text{obs}} \):** An estimator should perform better than \( S_{\text{obs}} \), the number of species observed. Most estimators should have lower bias than \( S_{\text{obs}} \), most of the time. However, there are trade-offs. For example, the variance of some estimators may be higher than that of \( S_{\text{obs}} \).

**B. Selecting an Estimator**

Although it is tempting to think that an estimator may exist which is robust to all census conditions, such an estimator is unlikely given the difficulty of the problem. Therefore, estimators should be chosen based on the context of the problem of interest. To select the estimators that are most likely to perform well on a new data set, one might use the following approach. First, consider the anticipated properties of a data set—based on knowledge of the system’s biology—in relation to the modeling assumptions behind estimators. Select those estimators whose assumptions are best met by the expected data. If possible, select or modify the sampling strategy so that assumptions will be better met and/or more estimators may be used. Then, test estimators using data (simulated or previously existing) similar to that anticipated. The previous approach should indicate (i) which estimators might perform best on new data and (ii) whether these estimators perform well enough to meet the needs of the problem at hand. Although evaluating a number of estimators on multiple replicate data sets may seem a daunting task, computer programs (e.g., EstimateS, R. K. Colwell, http://viceroy.eeb.uconn.edu/estimates; W2SM, W. R. Turner et al., unpublished) may be used to automate many of these calculations. However, use of such programs should complement, not replace, thoughtful consideration of modeling assumptions, sampling design, and the biology of the community under study.

**See Also the Following Articles**

- ECONOMIC VALUE OF BIODIVERSITY, MEASUREMENTS OF
- ECO SYSTEM FUNCTION MEASUREMENT, AQUATIC AND MARINE COMMUNITIES
- ECO SYSTEM FUNCTION MEASUREMENT, TERRESTRIAL COMMUNITIES
- FRAMEWORK FOR ASSESSMENT AND MONITORING OF BIODIVERSITY
- MICROBIAL BIODIVERSITY, MEASUREMENT OF

**Bibliography**


MEDITERRANEAN-CLIMATE ECOSYSTEMS

Philip W. Rundel
University of California, Los Angeles

I. Introduction
II. Natural Disturbance Regimes
III. Patterns of Speciation
IV. Biodiversity in California
V. Biodiversity in Chile
VI. Biodiversity in the Mediterranean Basin
VII. Biodiversity in the Cape Region of South Africa
VIII. Biodiversity in southwestern Australia
IX. Conclusions

GLOSSARY

chaparral Evergreen sclerophyllous-leaved shrublands that cover large areas of California.
coastal sage scrub Sem woody shrublands dominated by drought deciduous species and occurring in semi-arid areas along the coast of Southern California and the interior transition from chaparral to desert.
drought-deciduous Descriptive of plant species that lose their leaves during the dry season as soil moisture becomes limited.
fynbos Evergreen sclerophyll vegetation dominating the Cape Region of South Africa.
garrigue Low-growing secondary evergreen shrublands that dominate extensive areas of the Mediterranean Basin.
Gondwanaland Former supercontinent of the Southern Hemisphere from which South America, Africa, Australia, and India are derived.
kwongan Evergreen heathlands that cover extensive areas of southwestern Australia.

maquis Tall vegetation cover of the Mediterranean Basin dominated by evergreen sclerophyllous shrubs and trees.
matorral Evergreen sclerophyllous-leaved shrublands that dominate large areas of central Chile.
Mediterranean-type ecosystem (MTE) Habitat characterized by mild wet winters and warm dry summers. MTEs occur in California, central Chile, the Mediterranean Basin, the Cape Region of South Africa, and southwestern and South Australia, all at N and S latitudes of 30 to 35°.
phrygana Dwarf shrub land of the eastern Mediterranean Basin characterized by a dominance of species with seasonal leaf dimorphism.
renosterfeld Evergreen needle-leaved shrubland on richer soils of the Cape Floristic Region of South Africa and dominated by the resinous shrub Elytro pappus rhinocerotis (Asteraceae); the name literally means rhinoceros bush.
sclerophyllous Descriptive of leaves with a leathery texture due to the presence of sclerenchyma with large amounts of lignin and cellulose in their tissues.
seasonal leaf dimorphism A phenological trait in which plant species change the morphology of leaves through the year to better adapt them to prevailing temperature and water stress.

MEDITERRANEAN-TYPE ECOSYSTEMS (MTEs) are those habitats located in five regions of the world characterized by mild wet winters and warm dry summers.
TABLE I
Comparative Area, Topographic Heterogeneity, Climatic Heterogeneity, and Estimated Natural Fire Frequency for the Five Mediterranean-type Climate Regions of the World

<table>
<thead>
<tr>
<th>Region</th>
<th>Area (10^6 km^2)</th>
<th>Topographic heterogeneity</th>
<th>Climatic heterogeneity</th>
<th>Natural fire frequency (yrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>0.32</td>
<td>High</td>
<td>Very high</td>
<td>40–50</td>
</tr>
<tr>
<td>Central Chile</td>
<td>0.14</td>
<td>Very high</td>
<td>Very high</td>
<td>Fire-free</td>
</tr>
<tr>
<td>Mediterranean Basin</td>
<td>2.30</td>
<td>High</td>
<td>High</td>
<td>25–30</td>
</tr>
<tr>
<td>Cape Region</td>
<td>0.07</td>
<td>Moderate</td>
<td>High</td>
<td>10–20</td>
</tr>
<tr>
<td>Southwestern Australia</td>
<td>0.31</td>
<td>Low</td>
<td>Moderate</td>
<td>10–15</td>
</tr>
</tbody>
</table>

due to the influence of subtropical high-pressure centers. These five regions are located on the western margins of all of the major continental landmasses at latitudes of about 30 to 35° N and S. A variety of individual climate regimes occur within the MTEs under the influence of topography and marine influences. The characteristic form of vegetation in MTEs is a shrubland dominated by evergreen sclerophyllous shrubs, but woodlands are widespread as well in most MTEs.

I. INTRODUCTION
Mediterranean-climate ecosystems (MTEs) with their characteristic and unique climatic regimes of mild wet winters and warm and dry summers occur in just five regions of the world. These regions are California, central Chile, the Mediterranean Basin, the Cape Region of South Africa, and southwestern and South Australia (Table I). Biodiversity is particularly notable in the MTE regions for vascular plant species. Although the combined area of these five regions is little more than 2% of the land area of the earth, MTEs are home to approximately 50,000 species of vascular plants (Table II), 20% of the world's total. Nowhere outside of lowland tropical rain forests are there ecosystems with higher regional diversities of species, providing a strong justification for each of these regions being designated as a global "hot spot" of evolution.

The five Mediterranean-type ecosystems of the world share the characteristic climate regime of dry summers and cool wet winters. Typically 90% or more of the annual precipitation falls in the 6 months centered on winter. Mean annual precipitation is as low as 230 mm in coastal areas of the MTEs and reaches to as high as 900 mm at the upper margins of the classic evergreen

TABLE II
Comparative Species Richness of Vascular Plants and Major Vertebrate Groups for the Five Mediterranean-type Climate Regions of the World

<table>
<thead>
<tr>
<th>Vascular plant species</th>
<th>Resident bird species</th>
<th>Reptile species</th>
<th>Amphibian species</th>
</tr>
</thead>
<tbody>
<tr>
<td>California*</td>
<td>4,300</td>
<td>130</td>
<td>337</td>
</tr>
<tr>
<td>Chile**</td>
<td>4,000</td>
<td>99</td>
<td>285</td>
</tr>
<tr>
<td>Mediterranean Basin</td>
<td>23,000</td>
<td>197</td>
<td>366</td>
</tr>
<tr>
<td>Cape Region</td>
<td>8,530</td>
<td>127</td>
<td>288*</td>
</tr>
<tr>
<td>Western Australia</td>
<td>8,000</td>
<td>62</td>
<td>190</td>
</tr>
</tbody>
</table>

*Excluding desert areas.
**All of Chile.
#Excluding seabirds.
##Frogs only.
Some caution must be used in interpreting these numbers because of varying contexts of what constitutes core Mediterranean-climate habitats.
The evergreen, sclerophyll shrublands of California and Chile commonly grade off at their arid interior margin and along drier coastal margins to a plant community dominated by drought deciduous shrubs. This community, which may also dominate early successional disturbance sites or arid microsites in evergreen shrublands, is termed coastal sage scrub in California and coastal matorral in Chile. Structurally similar communities with mixed dominance of low evergreen and deciduous shrubs are called phrygana in Greece and batha in the eastern Mediterranean Basin. Deciduous shrubs are largely lacking, however, from similar habitats in the Cape Region of South Africa and southwestern Australia.

There has been a long history of comparative ecological studies between Mediterranean-climate regions. Particularly prominent in this respect have been comparative studies of California chaparral versus Chilean matorral and of South African fynbos versus Australian kwongan. Regular international meetings of scientists representing all five Mediterranean-climate regions have led to a series of edited volumes investigating individual environmental themes across these regions. Such syntheses have included themes of fire, nutrient relationships, ecosystem resilience, water, plant-animal interactions, biodiversity, and landscape disturbance (see the bibliography).

II. NATURAL DISTURBANCE REGIMES

Although the five Mediterranean-climate regions share many aspects of natural climatic and environmental disturbance regimes, there are significant differences as well. While all five regions experience characteristic summer drought, for example, the magnitude of this drought is particularly severe in California, Chile, and the much of the subard portion of the Mediterranean Basin where 6 to 8 months or more may pass without measurable rainfall. Extreme drought such as this is rare in South Africa and southwestern Australia where summer months frequently have light showers.

The dense canopies of evergreen sclerophyllous shrubs, which form the major component of vegetation cover over large areas of MTEs, are highly flammable. These structures, combined with summer drought, make fire an important component of natural disturbance regimes. Natural fires, however, are significant ecological events in consuming above-ground vegetation in only four of the five MTEs. Chile, where natural fires are a rare event, is the exception among MTEs. The Mediterranean-climate region of central Chile is

The Mediterranean-climate region of central Chile is

shrub zone. Mediterranean-type climates in the broad sense, however, also include adjacent arid regions that maintain winter rainfall regimes (e.g., the Mojave Desert in California, Atacama Desert in Chile, and Succulent Karoo in South Africa), as well as montane areas with winter rainfall where the majority of precipitation falls as winter snow. Many areas in the Sierra Nevada of California, for example, experience 10 m or more of annual snowfall. Although frosts may occur throughout much of the MTE regions, these are infrequent and relatively mild in lowland areas. Generally there is an upper limit of 3% of the annual hours with temperature below freezing in these regions.

A characteristic ecological feature of MTEs is the widespread prevalence of evergreen shrublands dominated by species with sclerophyllous leaves. These shrublands are called chaparral in California, matorral in Chile, garrigue and maquis in the Mediterranean Basin, fynbos in South Africa, and kwongan or heathlands in southwestern Australia. The dense cover of these shrublands burns readily under dry summer conditions with low humidity, although with differing frequencies characterizing individual MTE regions as described later. Thus, morphological, ecophysiological, and phenological adaptations to post-fire regeneration of these stand through resprouting and fire-stimulated reseeding is characteristic. Adaptations in tolerance of low nutrient availability and low summer water potentials are common in MTE sclerophyll shrubs.

Despite the general characterization of MTE regions as having dominance by evergreen, sclerophyll shrublands, other vegetation forms are also present. Woodlands are widespread in most MTEs, particularly in areas with deeper or richer soils, or as riparian woodlands or gallery forests in wetter sites. Oak woodlands dominated by species of Quercus are widespread in California and the Mediterranean Basin, with both evergreen and deciduous species as dominants. These communities can take the form of closed canopy evergreen woodlands grading into shrublands as in live oak woodlands of Southern California and the maquis of Europe, or open savannas of deciduous oaks that are widespread in both regions. Central Chile once had widespread dry sclerophyll and wet sclerophyll woodlands, but the dominance of these has been dramatically reduced by human activities. Evergreen woodlands dominated by Eucalyptus are widespread in western Australia. Only the Cape Region of South Africa of all of the MTEs is largely lacking in woodlands, with such communities restricted to scattered stands of relict Afro-montane forest along the southern coast in areas lacking in strong seasonal drought.
protected from summer storms and lightning moving westward across Argentina by the high Andean Cordillera. The native flora of Chile shows little evidence that fire has been an important ecological disturbance regime in the evolution of life history characteristics.

Natural fire frequencies are quite different among the other four Mediterranean-climate regions (Table I). In South Africa, for example, fynbos vegetation in the Cape Region commonly burns at intervals of 10 to 15 years, while in California natural frequencies are thought to be 40 to 60 years or more.

There are other strong environmental differences in addition to potential drought stress that make the Cape Region of South Africa and southwestern Australia distinct from the other three MTEs. These two areas lie in geologically ancient and stable landscapes, resulting in highly leached and nutrient-poor soils. In contrast, earthquakes, volcanic activity, orogenic uplift, and other dynamic processes create natural disturbance regimes in California, Chile, and the Mediterranean Basin that are absent in South Africa and Australia. Unlike South Africa and southwestern Australia, the younger landscapes of these three regions have experienced tremendous changes in climate regime and landscape structure in Quaternary and even Holocene times, and these changes have had profound impacts on community structure and speciation. The remarkable patterns of speciation in fire-sensitive shrub lineages in the Cape Region of South Africa and southwestern Australia likely has resulted from a combination of relatively mild and stable Quaternary climatic conditions coupled with high fire frequencies in these nutrient-poor habitats.

III. PATTERNS OF SPECIATION

Although there has been widespread speculation for many years on the causal factors responsible for the high levels of vascular plant biodiversity in MTEs, few clear generalizations have emerged. What is known is that there are conditions promoting the coexistence of seemingly ecologically equivalent shrubs, graminoids, and geophytes in frequently burned shrublands on nutrient-poor soils of fynbos and kwongan communities. Slow growth rates and diverse strategies of post-fire regeneration and reestablishment appear to promote this coexistence. MTEs on less infertile soils and with longer intervals between fires (chaparral, matorral, garigue) have lower diversity as short-lived species are excluded by rapidly growing shrub dominants. Open woodlands in California and the Mediterranean Basin, however, may have very high local diversity of annuals and short-lived perennials where grazing maintains open habitat for species establishment.

Regional topographic and climatic heterogeneity per se, which might be thought to be logical correlates of comparative diversity between the five MTEs, is a poor predictor of diversity (Table I). Instead, natural selection has operated in a predictable manner to allow for a fine-scale discrimination of habitats and niches under the selective pressures of stable climates, predictably frequent fires, and periodic drought that promote community turnover and diversification. Thus, fynbos in the southwestern Cape Region and kwongan communities in southwestern Australia have evolved species-rich landscapes in topographically homogeneous areas through rapid speciation coupled with low extinction rates.

IV. BIODIVERSITY IN CALIFORNIA

The political boundaries of the state of California cover an area of $4.1 \times 10^6$ km$^2$, but the area includes more than the core area of Mediterranean-type climate. These political boundaries include winter rainfall portions of the Sonoran Desert and Mojave Desert, as well as areas of cold desert habitats east of the Sierra Nevada. The California floristic province, as generally defined, excludes these desert areas and adds northwestern Baja California and southern Oregon to the floristic province. Under their definition, the California floristic province covers $3.24 \times 10^6$ km$^2$. Because of the differences between the political and floristic province boundaries of California, some caution must be used in assessing figures on California biodiversity in the literature.

The geomorphic structure of California is complex and the topographic diversity within the floristic region is very high. Thus, this region covers the Coast Ranges extending north and south along the state, the broad Central Valley, the Sierra Nevada range, and the Transverse Ranges of Southern California. The Coast Ranges reach elevations as high as 2700 m, while Mount Whitney in the Sierra Nevada is the highest point in the continental United States at 4420 m elevation. The Transverse Ranges in Southern California have a number of peaks reaching above 3000 m. The dynamic geologic history of uplift, faulting, and tectonics has produced complex mosaics of soil structure and parent material and sharp climate shifts over the Quaternary with associated glaciation in the high mountains.

The foothill regions throughout most of California are typically dominated by mosaics of evergreen chape-
ral shrublands and both evergreen and deciduous woodlands with oak species as the typical dominants. These areas commonly receive 400 to 800 mm annual rainfall. Rainfall is strongly centered on the winter months, and 6 months without rain is common. Drier areas along the coast and inland at the transition to desert environments support coastal sage scrub dominated by drought deciduous shrubs and a few species of deeply rooted evergreen sclerophyllous. Mountain areas above 1500 m in Northern California and 1800 m in Southern California show a transition to montane conifer forests, subalpine forests, and alpine communities with increasing elevation. Higher rainfall areas along the central and northern coast of California support mixes of conifer and hardwood forests, extending into massive coast redwood forests along the northwest coast. Mean annual rainfall reaches its highest levels above 2300 mm in this region.

A. Vascular Plant Species Diversity

California as a political unit contains 4839 vascular plants species. This total includes 99 ferns and fern relatives, 60 gymnosperms (33 conifers), 823 monocots, and 3862 dicots. The largest family in this flora is the Asteraeaceae with 627 native species, followed by the Fabaceae with 297 species, and the Poaceae with 251 species. The largest five genera make up more than 10% of this total and include Carex (131 species, Cyperaceae), Eriogonum (112 species, Polygonaceae), Astragalus (94 species, Fabaceae), Phacelia (93 species, Hydrophyllaceae), and Lupinus (71 species, Fabaceae). All of these genera are composed largely of herbaceous perennial and annual species. Notable speciation has also occurred in two shrub lineages, Arctostaphylos (Eriaceae) and Ceanothus (Rhamnaceae), in response to adaptations to post-fire regeneration. Older published data give a figure of 4452 species for the California flora.
its extreme of conditions in the Tacna-Arica region along the Peruvian border. Here a virtual absence of rainfall separates the Peruvian floristic elements of the desert from the Chilean elements. To the east, the Mediterranean-climate region of central Chile is strongly delineated by the high Cordillera de los Andes. Many peaks in this range reach well above 6000 m and effectively shield Chile from weather fronts moving westward across Argentina. Although major uplift of the Andes began in the mid-Tertiary, at least 14 million years ago, the range continues to be tectonically active today. The elevation of mountain passes along the Cordillera de los Andes in northern and central Chile is too high to allow easy migration of either plants or animals and thus has helped to isolate the flora and fauna of Chile. Only in southern Chile where the Andes are lower has there been easy migration across this range. However, the severe climatic conditions of cold that characterize Patagonia in southern Chile strongly reduce the biological diversity in this area.

Comparisons of species diversity between Chilean organisms and those of other Mediterranean-climate regions deserve some caution in terms of the area included. As with California, the political boundaries of Chile include desert and wet forest ecosystems that are not comparable to core Mediterranean-climate habitats. Most figures for Chilean diversity in the literature are based on political boundaries. An additional issue of political boundaries comes in assessing levels of endemism for the Chilean flora or fauna. Levels of endemism for all groups are much lower if strictly adhering to the political boundaries of Chile rather than to the more natural boundaries of the Chilean/Patagonian biogeographic province.

Much of central Chile, which covers an area of about $140 \times 10^3$ km$^2$, shares a physiographic structure parallel to that of California. Moving inland from the Pacific Ocean, there is a coastal range of mountains, a broad central valley, and a high mountain range to the east. The geologically recent Cordillera de la Costa in Chile is relatively tall. West of Santiago at about 33° latitude, the major peaks are Cerro Campana (1910 m), Campanita (1510 m), and El Roble (2220 m). These peaks are high enough to intercept moisture from humid southwestern winds, producing woodland areas with significant fog interception and thus improved water relations. Further north in the winter-rainfall desert regions, the coast ranges reach up to 3000 m in the Cordillera Vicuña Mackenna.

The dominant vegetation in central Chile is matorral, an evergreen shrubland similar in general form to chaparral. Along the coast and to the north, this community grades into a coastal matorral with a greater dominance by drought deciduous shrubs. At higher elevations and on sites with greater water availability, matorral grades into a sclerophyll woodland community, and on the higher parts of the coast range into hygrophilous woodland with species characteristic of the Valdivian forest region of south-central Chile. Much of the central valley of Chile today is dominated by a savanna community termed espinal, with Acacia caven as the sole dominant. This community is almost certainly the result of human intervention on landscape processes over the past four centuries. Sclerophyll woodland and matorral would once have covered much of this area.

Unlike California and the Mediterranean Basin, the high mountains of the Andean Cordillera in central Chile do not have a forest zone. The young age of these mountains and lack of soil weathering have produced unstable geological conditions on the west-facing slope of the Andes. Matorral communities on the lower foothills of the Andes give way to a low and scrubby monane matorral community at about 2000 m elevation.

The biological diversity of Chile has an ancient origin that dates back to Gondwanaland. Southern Chile in particular exhibits many broad biogeographical linkages with New Zealand and Australia. Central Chile shows other biogeographical connections with southeastern Brazil, a linkage dating back to mid-Tertiary times before the uplift of the Cordillera de los Andes. Since the Andean uplift, however, the biota of Chile has evolved largely in isolation from other biogeographical regions.

Biosystematic and biogeographical knowledge of the flora and fauna of Chile have increased greatly in recent decades and thus the biodiversity of most groups of vascular plants and vertebrates is relatively well known. These syntheses have been applied more generally to the country as a whole, however, rather than to discrete regional areas. Thus, there is a strong need for more regional studies that evaluate patterns of alpha, beta, and gamma diversity in relation to environmental gradients.

A. Vascular Plant Diversity

The age and evolutionary isolation of the Chilean flora is clearly indicated by the great number of families that are largely endemic to the Chilean floristic region, which includes adjacent areas of Austral forest in southern Argentina. One family of ferns, the Thyrsopteridaceae, and 9 families of Angiosperms are endemic to Chile. Endemic families entirely restricted to Chile are...
the Aextoxicaceae, Gomortgaceae, and Lactoridaceae (restricted to the Juan Fernandez Islands). Other families that are largely Chilean in distribution but cross political boundaries into desert areas of Peru or Austral forests of southern Argentina are the Malesherbiaceae, Nolanaeae, Mizidendraceae, Vivianiaceae, Framoaceae (a segregate from the Saxifragaceae), Heterosylaceae, and Arachnitaceae (sometimes placed in Corsiaceae). Of these endemic families, the Aextoxicaceae and Gomortgaceae, have distributions centered in the Mediterranean-climate regions of central Chile, while the Malesherbiaceae and Nolanaeae are centered in the coastal deserts and adjacent and montane regions of northern Chile and southern Peru, although one species of the latter reaches the Galapagos Islands.

For continental Chile as a political unit, the vascular plant flora consists of approximately 4600 native species, divided into about 850 native genera and 180 families. This flora includes 124 species of ferns and fern relatives, 13 species of gymnosperms, and about 4500 species of Angiosperms. The flora of central Chile, excluding the desert areas north of La Serena and the forest and moorland areas south of Concepción, is estimated to include about half of this total.

Endemism is high at the generic level in the Chilean flora. Extrapolating from literature, 16% of Chilean genera are strict endemics restricted to the political boundaries of the country. Another 17% of the genera are endemic to Austral regions of Chile and adjacent areas of Argentina. Together, then, a third of the genera are endemic to the Chilean-Patagonian floristic province. There are 22% of the genera with broad South American patterns of distribution and 10% of the genera have Gondwanaland origins with extant species in New Zealand or Australia.

The levels of species-level endemism within groups of the Chilean flora are high, but the values reported vary somewhat depending on whether or not the author is quoting distributions within the strict political boundaries of Chile. A more natural view of endemism would include Andean species or Austral forest and moorland species, which occur within communities of the Chilean-Patagonian biogeographic province that may extend into Argentina. Using the political boundaries of Chile, one authority estimated that 62% of the vascular plant species were endemic, while others estimate about 50 to 55%.

A rich flora high in endemism occurs on the oceanic Juan Fernandez Islands, which lie about 300 km off of the central Chilean coast and contain floristic elements from both the mainland of Chile and Polynesia. The flora includes 361 species of vascular plants, 60% of which are endemic. In addition, there is an endemic family, the Lactoridaceae. The 20 largest genera of Chilean vascular plants make up about 30% of the flora. This group is led by Senecio (Asteraceae) with 218 species, Adromia (Fabaceae) with 140 species, Oulth (Onalidaceae) with 111 species, Calceolaria (Scrophulariaceae) with 86 species, and Calandrinus sensu lato (Portulacaceae) with 70 species. As might be expected, the degree of endemism within the largest families is higher than that for the flora overall. Eleven of the 20 largest families have 80% or more of their species endemic to Chile. These numbers on species diversity and endemism will no doubt change somewhat as these large and difficult genera become better studied. One of the most charismatic of endemic species is the Chilean wine palm, Jubae chilensis. This large palm was once widespread through central Chile but has a more restricted distribution today.

No manual or complete checklist of the flora of central Chile has been produced. For the latitudes of 32° to 36°S, it has been estimated that there are about 2100 species of native vascular plants. Regional floras exist for several areas of central Chile, but these are generally older and incomplete.

Vascular plant species diversity at small scales of habitat in central Chile were studied during the International Biological Program (IBP) in the 1970s. Data were prepared on species diversity in 1 m², 10 m², 100 m², and 0.1 ha plots located at a single site in each of four habitats—coastal matorral, matorral, sclerophyll forest, and espinal. The number of species in the 0.1 ha plots ranged from a high of 114 at the matorral site to only 25 at the espinal site. The sclerophyll forest site (102 species) and coastal matorral site (85 species) were also high in diversity. Excepting the espinal site, all of these values are remarkably high and almost certainly not typical of species diversity in central Chile. More studies of alpha species diversity at these scales are needed.

B. Mammal Diversity

Mammal diversity in Chile, as in other parts of temperate South America, is low. It has been suggested that severe climatic conditions associated with Pleistocene glacial movements in the Andes may have had a strong impact in reducing the diversity of temperate mammal faunas in South America. Major faunal extinctions were also present in South America at the end of the Pleistocene, as in North America, and these extinctions have been associated with the arrival of early humans. Many ecological niches in temperate South America appear
to be incompletely occupied by mammals in comparison to North America.

Chile has a mammal fauna of 99 native terrestrial species. The largest single group is the Rodentia, which comprise 60% of this total. Next in abundance are the Carnivora with 14%, and the Chiroptera (bats) with 10%. Large mammal species are particularly low in number. These include five species of felids (e.g. puma, Geoffroy’s cat, and colo colo), three species of canids (all fox species of Pseudolopex), four camelids (guanaco, vicuna, and the domesticated llama and alpaca), and three cervids (huemul, northern huemul, and pudu). Mammal faunas of Chile have been placed into six biogeographical groupings: the summer rainfall Altiplano region of northeastern Chile, the Atacama Desert and adjacent winter rainfall Andes of northern Chile, the Andes of central Chile, the Mediterranean-climate region of central Chile, the Austral forests of southern Chile, and the Patagonian region. In addition to these native mammal faunas, 15 species of terrestrial mammals have become naturalized in Chile. Five of these occur only on the Juan Fernandez Islands where they have had an extreme impact on the structure and composition of the native flora.

Considering Chile in the broad sense to include communities that extend across political boundaries, there are functionally 37 endemic species. Endemism is highest among Chilian mammals, however, in the summer rainfall Altiplano region (63 species) and in Patagonia (10 of 34 species). The Mediterranean-climate region of central Chile is the primary biogeographical range of only 12 endemic species. Observing the strict political boundaries of Chile, only 15 of the 99 species of terrestrial mammals are endemic. These include 13 species of rodents, 1 marsupial, and the Chiloe fox (Pseudalopex fulvipes).

On a site basis of comparison for Mediterranean-type shrublands, the small mammal fauna of Chile is also relatively low. IBP studies at matorral, coastal matorral, and dry coastal scrub communities in central Chile found only a total of 10 small mammals. These were nine rodents and a marsupial mouse. A similar gradient of ecological sites in California yielded 17 species.

C. Bird Diversity

The total bird diversity of Chile is only moderate. Including the oceanic islands of Pascua and Juan Fernandez and the Antarctic territory claimed by Chile, there are reports of 491 native species and 5 introduced species for Chile. These are distributed in 56 families and 222 genera. Of the total native species, 285 are residents, 90 are visitors, and 76 are accidentals. Among the administrative regions of central Chile, the diversity of native bird species is remarkably consistent with a range of 179 to 194 species per region (excluding the smaller metropolitan region of Santiago). There are 10 endemic bird species in Chile, all of them terrestrial. Seven of these occur on the mainland of central Chile, while the remaining three are restricted to the Juan Fernandez Islands.

At a regional scale, the structure and diversity of bird faunas in the Mediterranean-climate region of central Chile have been studied in considerable detail, using structurally parallel communities from the coast into montane habitats of the coast ranges in Chile and California as a comparison. Excluding such guilds as aerial feeders, most parasites, raptors, and nocturnal species, there were only 39 species in the Chiloean gradient compared to 69 species in California. Looking more specifically at individual sites, sclerophyll woodlands in Chile supported only 18 bird species compared to 31 species in the structurally similar oak woodlands of southern California. The changes in β-diversity observed along the Chiloean and Californian gradients showed a similar pattern, with much higher levels of diversity in California. Chile differed in the opposite direction, however, with much higher γ-diversity than California. When these differences were all merged into assessing the total diversity of resident bird species, Chile and California prove to be almost identical.

The biogeographical isolation of the Chiloean bird fauna shows parallels with that of the mammal fauna in the manner in which niches have been filled in unusual manners. Corvids, for example, are absent from Chile. Their role as scavengers has been filled by caracaras.

D. Reptile and Amphibian Diversity

There are 87 native species of terrestrial reptiles in Chile, divided among 7 families and 18 genera. These include 6 snakes (four genera) and 81 lizards (14 genera). Two of these lizard species (two genera) occur only on Isla Pascua. Seven turtle species (six genera), all marine, are known from the coasts of Chile and its oceanic islands. The diverse lizard fauna of Chile is strongly dominated by the family Tropiduridae, in particular the iguanid genus Liolaemus with 53 species. This diverse genus and its evolution have been the focus of a large number of ecological and evolutionary studies. The occurrence of distinctive populations within individual species has led to the designation of a large number of subspecies. Only Sceloporus and Anolis among New World lizards have a species diversity of this magnitude. Studies in central Chile during the IBP
VI. BIODIVERSITY IN THE MEDITERRANEAN BASIN

The Mediterranean Basin represents the largest area of MTEs in the world, covering a complex landscape with a large amount of topographic and climatic heterogeneity. Covering an area of approximately 2.3 × 10^6 km², it is nearly 10 times greater in size than any other MTE. This area includes more than 20 nations arrayed on both sides of the Mediterranean Sea. While coastal areas are extensive because of the large archipelagos and islands within the Mediterranean, much of this area consists of mountainous terrain with many areas above 2000 m elevation and peaks reaching above 4000 m. The geographic position of the Mediterranean Basin is also an important factor in understanding the biodiversity of this region. Lying at the juncture of three continental landmasses, it holds a geologic history with dynamic changes associated with plate tectonics, mountain uplift, and active volcanism. In contrast to other MTEs, the great majority of the Mediterranean Basin is underlain by limestone. Local areas of volcanic or siliceous parent material are present, however. Strong climatic shifts that took place during the Plio-Pleistocene period, most notably major glacial episodes, resulted in a telescoping of many communities into the Mediterranean Basin and provided opportunities for geographic isolation and speciation.

The climatic features of the Mediterranean Basin are often used to define this region, but the range of dominant and widespread woody species such as Quercus ilex and Olea europaea also are used as bioindicators of the region. To the north, the Mediterranean-climate region grades into more mesic regions with summer or year-round patterns of rainfall. To the south, the Mediterranean region integrates with the winter rainfall desert of the northern Sahara. Climates of the Mediterranean Basin are notable for their high interannual variation in both rainfall and temperature extremes. The large area of the Mediterranean Basin, coupled with its topographic and climatic heterogeneity, makes for complex assemblages of vegetation types. There are extensive woodlands dominated by both evergreen and deciduous species of oak, and evergreen sclerophyllous shrublands of many forms. These shrublands are often differentiated into types depending on the height of the vegetation. Tall sclerophyllous shrublands that may include small evergreen trees are termed maquis. Several species of Mediterranean pine may be present in this community. A middle height shrubland, generally occurring on calcareous substrates, is termed garrigue. Finally, low semi-arid evergreen shrublands in the eastern Mediterranean Basin are commonly termed phrygana in Greece and barba in Israel. A long history of human impacts on the natural landscape has strongly impacted community structure and diversity.

A. Vascular Plant Diversity

The vascular plant flora of the Mediterranean Basin is estimated to include about 25,000 species, making this region the richest among MTEs in total plant diversity. By comparison, the remaining portions of Europe with non-Mediterranean-type climate regimes cover four
times as much area but have only about 6000 vascular plant species. This large flora in the Mediterranean Basin is a broad mixture of species with disparate evolutionary histories and biogeographical origins. One group of species evolved under subtropical conditions that existed in this region prior to the Quaternary. This group includes such woody plant genera as Arbutus and Calluna (Ericaceae), Ceratonia (Fabaceae), Chamaerops (Arecaceae), and Laurus (Lauraceae). Another group of taxa represents neo-Mediterranean elements that migrated into the Mediterranean basin after the establishment of a Mediterranean-type climate. Examples of woody genera in this group include Amelanchier (Rosaceae), Clematis (Ranunculaceae), and Cistus, Halimium, and Helianthemum (Cistaceae). Three groups of nontropical woody elements that evolved after the onset of Mediterranean-type climates have been identified. These groups are Mediterranean elements evolved in situ in mountain areas and high in endemism, a desert and cold steppe group of species entering from Africa and the Middle East, and a Holarctic element of species with Eurasian temperate affinities. Endemism is high at the species level in the Mediterranean Basin, with a level of about 50%. No family of vascular plants is strictly endemic to the Mediterranean Basin. Species richness at the scale of 0.1 ha is often remarkably high in large or disturbed Mediterranean woodlands and shrub grasslands, with as many as to 119 to 179 species in such stands. In Israel, with 21 to 29 species in a single 1 m², annual species contributed from 47 to 72% of these 0.1 ha diversities. Protected areas without disturbance and highly disturbed sites support fewer species. This suggests significant coevolution between plants and herbivores.

B. Vertebrates

Vertebrate faunas of the Mediterranean Basin share with vascular plants the characteristic of multiple biogeographical origins. Dramatic climatic oscillations during the Pleistocene led to periodic turnovers of Eurasian and African faunal elements and a resulting high beta diversity in species diversity. The present fauna of land mammal species for the Mediterranean Basin numbers about 197 species, of which 23% are endemic. Because of the biogeographical barriers of the Mediterranean Sea and the Saharan Desert, mammal faunas of Mediterranean Europe, the Middle East, and North Africa are somewhat distinct. For North Africa the mammal fauna of the Mediterranean Basin shows its strongest affinities with tropical Africa. A decline in species richness occurred at the end of the Pleistocene with a combination of sharp climate shifts and human pressures through hunting.

Bird diversity of the Mediterranean Basin includes about 366 breeding species. This number compares with a total of 419 breeding bird species reported for all of Europe. In contrast to mammals, the affinities of bird faunas are more strongly linked to the Asian steppes than to tropical Africa. The evolution of these elements of bird faunas can be linked to Eurasian (135 species) and Eremian semiarid habitats (85 species), where Plio-Pleistocene conditions led to ongoing isolation and speciation. Forest birds of boreal origin are widespread and dominant throughout both middle Europe and the Mediterranean Basin. Shrubland bird species characteristic of the region represents only about 12% of the total. Overall, there are 62 endemic bird species in the Mediterranean Basin, 17% of the total.

Reptiles and amphibians of the Mediterranean Basin include 165 and 83 species, respectively, and show distinct holarctic affinities. Much of the endemism within these groups appears to represent archaic lineages that differentiated during the middle Tertiary. Reptile diversity is highest in the eastern Mediterranean Basin and drops steadily moving westward. Species diversity on Mediterranean islands is relatively low. Important reptile groups include lizards of the Lacertidae with 60 species (30% of the world total), snakes of the Viperidae with 14 species (7.4% of the world total), and tortoises of the Testudinidae with 4 species. Overall, 68% of Mediterranean Basin reptiles are endemic to this region. For amphibian diversity, the pattern is reversed compared with reptiles as the highest levels of diversity are found in the Euro-Mediterranean areas compared to the North African and Middle Eastern portions of the region. Notable groups of amphibians include the Discoglossidae with 10 species (71% of the world total) and the Salamandridae with 19 species (36% of the world total). Endemism for amphibians in the Mediterranean Basin is 94%.

VII. BIODIVERSITY IN THE CAPE REGION OF SOUTH AFRICA

The Cape Region forms a small area on the southwestern tip of the African continent; it is renowned for its showy and diverse flora that is unlike that of any other area of the world. The landscape of the Cape Region is dominated by steep and strongly folded mountain ranges of quartzitic sandstone that predate the separation of the African continent from Gondwanaland.
These once high mountains have been eroded over the past 200 million years to form low ranges capped by resistant Table Mountain sandstone. Separating the mountains are gentle valleys and undulating plains that are largely underlain by shales with greater nutrient availability. Relatively young Tertiary and Quaternary limestones and sands mantle extensive areas of the coast.

The characteristic vegetation of the Cape Region, particularly on the nutrient-poor quartzite soils, is fynbos. Fynbos is an evergreen shrubland dominated by four major plant morphological groups. These include two shrub groups (the proteoids and ericoids), a sedge-like group (restioids), and geophytes. The proteoids, formed by woody Proteaceae, make up the tallest matrix of the fynbos community and commonly reach 2 to 4 m in height. The ericoid group gains its name from the Ericaceae but includes more than 3000 species of small-leaved shrubs representing many families. The restioids are primarily members of the Restionaceae, a family with origins in Gondwanaland but its primary diversification in the fynbos. Finally, the Cape Region contains the highest diversity of bulbs and other geophytes in the world, with more than 1300 species. Many types of fynbos have been described, but a simple classification scheme includes proteoid fynbos, ericaceous fynbos, restioid fynbos, asteraceous fynbos, and grassy fynbos.

Another important shrubland vegetation type of the Cape Region is renosterveld (or rhinoceros veld, referring to black rhino that once grazed here). Renosterveld contains thousands of species, many of which have 100 species.

A. Vascular Plants

The Cape Region contains what is arguably the most unique and diverse flora of any temperate area of the world. Despite the relatively tiny area of this region, phytogeographers have uniformly separated it out from the other parts of Africa and designated it as the Cape Floral Kingdom, one of only six floral kingdoms that comprise the entire world. This status as a distinct floral kingdom is reinforced by the presence of six endemic families—Geissolomataceae, Grubbiaceae, Penaeaceae, Retziaceae, Roridulaceae, and Stilbaceae. Moreover, there are 193 endemic genera, 19.5% of the total.

Covering an area of only 71,000 km², the Cape Region contains approximately 8550 species of vascular plants. About 7000 of these species grow in fynbos communities. More recently it has been suggested that the winter-rainfall area of the arid Succulent Karoo should be added to the Cape floristic region. Such a change would almost double the number of species present. For the purposes of this review, however, only the Mediterranean-climate shrublands of the Cape Region are considered. The 10 largest genera of the Cape Region account for 20% of the flora. These are Erica (Ericaceae, 566 species), Aspalathus (Fabaceae, 245 species), Phyllica (Rhamnaceae, 133 species), Agathosma (Rutaceae, 130 species), Ornithopus (Oxalidaceae, 129 species), Pelargonium (Geraniaceae, 125 species), Senecio (Asteraceae, 113 species), Cliftonia (Rosaceae, 106 species), Muraltia (Polygonaceae, 106 species), and Ruschia (Aizoaceae-Mesembryanthemata, about 100 species).

Species richness is greatest in the southwestern Cape Region centered around Cape Town. The Cape Peninsula, for example, supports 2256 species (including 90 endemics) in an area of 471 km². Cape Hangklip on the eastern shore of False Bay near Cape Town has 1383 species in 240 km².

Levels of species endemism in the Cape floristic region are among the highest in the world. For the entire region, endemism at the species level is about 68%. The high levels of endemism present in the Cape Region are largely due to the presence of neoeconomics rather than paleoeconomics. This dominance of neoeconomics is indicated by the predominance of endemic diversity in large, species-rich genera, the widespread presence of sympatric congeners, and the edaphic specialization of many endemics on geologically young substrates. Rather than being a random ecological or phylogenetic assemblage of species, the great majority of endemics are low shrubs killed by fire and dependent on closely dispersed seeds for regeneration. Four families are notably rich in endemics: the Proteaceae, Ericaceae, Rutaceae, and Polylalaeaceae. Smaller regional centers of endemism exist within the Cape Region. Dividing the Cape Region up into five floristic zones on the basis of species distributions within seven large families, regional levels of endemism are highest in the southwestern and northwestern Cape (about 30%) and lowest in the eastern Cape and inland...
B. Mammal Diversity

The Cape Region lacks a distinctive mammal fauna. This region contains 127 species of native mammals, with 90 of these being present in the Southwest Cape area. The regional total is less than half of the mammal species occurring within all of South Africa. The largest orders present are the Rodentia and the Carnivora. The rodents are represented by 2 species of mole rats (Bathyergidae), a porcupine (Hystricidae), 2 dormice (Muscardinidae), and at least 21 species of Muridae and Cricetidae. There are 27 species of the Carnivora, ranging from mustelids and civets to larger hyenas, jackals, and cats. Large browsers and grazers play an important role in this ecosystem in comparison to other MTEs. There are 20 species of Artiodactyla and 5 species of Perissodactyla. Very few of these depend on grazing, however, because of the paucity and poor nutritive value of Cape Region grasses. The Chiroptera is a large group with 1 fruit bat and 14 species of Microchiroptera in the Southwest Cape. Orders and families present include eland, buffalo, and hartebeest. Elephants, essential absent from this region today, once grazed commonly along the south coast. Two other grazers, the bluebuck and quagga, are extinct. The absence of large carnivores from many areas today has led to increased populations of small mammals such as rock hyraxes, which may exert a significant impact on plant demography through selective grazing and granivory.

Endemism, as might be expected, is quite low among mammals in the Cape Floristic Region as most of this fauna extends northward or westward into arid or savanna ecosystems. Only 7% of the mammal fauna (nine species) are endemic. These endemic mammals include two bats, three rodents, an insectivore, and three antelope (Cape grysbok, bontebok, and the extinct bluebuck). The fossorial Cape dune mole rat and burrowing gerbil among the endemic rodents are associated with sandy soil substrates rather than with any specific vegetation type. The colonial behavior and feeding specialization of mole rats on bulbs may be linked to the remarkably high diversity of geophytes in the Cape Floristic Region.

C. Bird Diversity

Ecological analyses of avian diversity in relation to community structure and function in the Cape Region have shown that species richness in fynbos communities is comparable to that in other MTEs. Thus, the high species richness for vascular plants is a poor predictor of bird species diversity. Fynbos birds generally occupy narrow feeding niches and replace themselves rapidly across gradients of changing fynbos communities. In comparison to bird faunas in California chaparral and Chilean matorral, fynbos birds are more stereotyped to a narrow habitat range and less behaviorally plastic. The nature of low nutrient soils that have led to the evolution of highly sclerophyllous leaves low in nutrient content and high in lignin have made these communities difficult resources for insect herbivores. As a result, the fynbos is low in both the abundance and diversity of insectivorous birds. For the Cape Region avian fauna overall, however, the average density of

33 species.

and diversity of large mammals was present in renosterveld plains included the bontebok, eland, buffalo, and hartebeest. At the time of European colonization, the highest numbers of large mammals in the Cape Region were present in the teeming savanna regions of the continent. At the Cape Region lack a distinctive mammal fauna. This region contains 127 species of native mammals, with 90 of these being present in the Southwest Cape area. The regional total is less than half of the mammal species occurring within all of South Africa. The largest orders present are the Rodentia and the Carnivora. The rodents are represented by 2 species of mole rats (Bathyergidae), a porcupine (Hystricidae), 2 dormice (Muscardinidae), and at least 21 species of Muridae and Cricetidae. There are 27 species of the Carnivora, ranging from mustelids and civets to larger hyenas, jackals, and cats. Large browsers and grazers play an important role in this ecosystem in comparison to other MTEs. There are 20 species of Artiodactyla and 5 species of Perissodactyla. Very few of these depend on grazing, however, because of the paucity and poor nutritive value of Cape Region grasses. The Chiroptera is a large group with 1 fruit bat and 14 species of Microchiroptera in the Southwest Cape. Orders and families present include eland, buffalo, and hartebeest. Elephants, essential absent from this region today, once grazed commonly along the south coast. Two other grazers, the bluebuck and quagga, are extinct. The absence of large carnivores from many areas today has led to increased populations of small mammals such as rock hyraxes, which may exert a significant impact on plant demography through selective grazing and granivory.

Endemism, as might be expected, is quite low among mammals in the Cape Floristic Region as most of this fauna extends northward or westward into arid or savanna ecosystems. Only 7% of the mammal fauna (nine species) are endemic. These endemic mammals include two bats, three rodents, an insectivore, and three antelope (Cape grysbok, bontebok, and the extinct bluebuck). The fossorial Cape dune mole rat and burrowing gerbil among the endemic rodents are associated with sandy soil substrates rather than with any specific vegetation type. The colonial behavior and feeding specialization of mole rats on bulbs may be linked to the remarkably high diversity of geophytes in the Cape Floristic Region.

C. Bird Diversity

Ecological analyses of avian diversity in relation to community structure and function in the Cape Region have shown that species richness in fynbos communities is comparable to that in other MTEs. Thus, the high species richness for vascular plants is a poor predictor of bird species diversity. Fynbos birds generally occupy narrow feeding niches and replace themselves rapidly across gradients of changing fynbos communities. In comparison to bird faunas in California chaparral and Chilean matorral, fynbos birds are more stereotyped to a narrow habitat range and less behaviorally plastic. The nature of low nutrient soils that have led to the evolution of highly sclerophyllous leaves low in nutrient content and high in lignin have made these communities difficult resources for insect herbivores. As a result, the fynbos is low in both the abundance and diversity of insectivorous birds. For the Cape Region avian fauna overall, however, the average density of mountain regions with nonseasonal rainfall (18–28%). These patterns of regional endemism have been further demonstrated in studies of distribution of the Proteaceae in the Cape Region. For the entire Cape Region, 99.4% of the 330 species of Proteaceae are endemic. At a regional level, 63% of the Proteaceae in the southwesternern region are endemic to that region, compared to only 19% endemism for the coastal mountain and southeastern regions. Point endemism is also widespread involving species that are restricted to highly specific edaphic habitats.

Fynbos species diversity is also extremely high at the alpha-diversity level of small stands. Typical fynbos communities support a mean of about 65 vascular plant species in 0.1 ha, with a range of 31 to 126 species. Renosterveld shrublands have even higher diversities with a mean of 84 species per 0.1 ha and a range 28 to 143 species. Fynbos and renosterveld average about 14 to 17 species in 1 m² areas, with a range of 5 to 33 species.
birds is quite comparable to that found in California and central Chile in ecologically comparable communities. Excluding seabirds, there are 288 species of resident birds in the Cape Region, with a notable diversity of Falconiformes with 22 species. For fynbos communities specifically, there are 101 reported species. The richest habitat for birds is found to be renosterveld with about 77 species.

As with mammals, endemism is low among Cape Region birds. Only 2% (six species) are endemic. Endemic species are largely dietary specialists such as the Cape sugarbird, orange-breasted sunbird, and Protea seedeater that are tied to specific plant resources in the fynbos. The originally direct communication of fynbos habitats with semi-arid and savanna shrublands has probably been a factor in limiting the number of endemic fynbos birds. The savanna region of South Africa is far richer in endemic species. At least four of the endemic fynbos birds are characteristic of montane areas or are allied to montane species of East Africa.


d. Reptile and Amphibian Diversity

The Cape Region is moderately diverse in reptiles, with 109 species known to occur in this area. Fynbos communities may contain more than 50 species of lizards. The Gekkonidae are the most important group with 18 species. There are 32 species of snakes reported from the Western Cape area. Among snakes, the Colubridae have the highest diversity with 23 species. There are 19 endemic species among the reptiles, 17% of the total. One notable endemic to the southwestern Cape Region is Psammobates geometricus, one of the rarest tortoises in the world. The life history of this species seems to have evolved to adapt to fynbos fire cycles, with hatchlings appearing in late autumn after the danger of summer fires is past.

Amphibians are relatively low in diversity in the Cape Region with 38 native species. The largest single group of amphibians is the Ranidae with 13 species. Compared to other vertebrate groups in the Cape Region, endemism among amphibians is relatively high at 50% (19 species), and much of this endemism is centered in the Southwest Cape area. One of the most interesting endemics among amphibians is the arum lily frog, Hyperolius horstocki, which lives in the flowers of the common arum lily Zantedeschia aethiopica.

E. Invertebrate Diversity

Relatively little is known about the diversity of invertebrates in the Cape Region, but what data exists suggest that very high levels of endemism exist in many groups. Among 234 species of butterflies known from the region, 31% are endemic. A regional study of invertebrates on the Cape Peninsula reported a large number of endemic species. The high diversity and rates among insects in the Cape Region is not surprising in view of the important ecological links between insects and plants. Specific pollination mechanisms have been involved in maintaining many sympatric species populations. Ants perform an important role as seed dispersers. About 20% of fynbos species have seeds dispersed by ants, a pattern seen only in western Australia among other MTEs.

VIII. BIODIVERSITY IN SOUTHWESTERN AUSTRALIA

Core Mediterranean-climate conditions are present in southwestern Western Australia and in South Australia around Adelaide. The core southwestern floristic province of western Australia is approximately 310 x 10³ km² in area, similar in size to the California floristic province. A broader region with transitional rainfall patterns with biseasonal distribution is often included as well, almost doubling the core MTE area. While mean annual rainfall is commonly 350 to 800 mm over much of southwestern Australia, it reaches as high as 1500 mm at the extreme southwestern corner of western Australia and drops to a low of about 250 mm at the eastern edge of the Mediterranean-climate region in a transition to arid communities. Topographic heterogeneity is very low over this region, with elevations almost entirely below 1000 m. Much of the region is a low, laterized plateau dissected into broad valleys with deep in situ weathering. Soils of southwestern Australia, as in the Cape Region of South Africa, are generally very old, highly weathered, acidic, and low in nutrient availability.

For the true Mediterranean-climate regions, the highest rainfall zones with 800 to 1200 mm of annual rainfall support evergreen forests and woodlands dominated by Eucalyptus marginata (jarrah), E. calophylla (marri), E. obvallata (karri), and E. globulus (stringybark). Low Banksia woodlands and coastal heath are also present. At intermediate rainfall regimes of 300 to 800 mm, the dominant vegetation is a mosaic of low woodlands (mallees) and heathland communities termed kwongan. The vegetation and dominant species are finely tuned to small changes in edaphic conditions that influence nutrient and water availability. Human impacts on these ecosystems have been severe over the past century.
High levels of species diversity and endemism characterize the vascular plant flora, with only moderate diversity and endemism present in most vertebrate groups. The high levels of vascular plant diversity has been related to (a) the development of a complex mosaic of landforms and soils during the Tertiary and Quaternary, (b) the geologic history of oscillating moisture regimes through the Quaternary in the absence of glaciation, (c) isolation of southwestern Australia from the east by the arid Nullarbor Plain, and (d) interactions of gene pools from both paleotropical and temperate assemblages.

A. Vascular Plant Diversity

The South West Botanical Province, which includes the broadly defined Mediterranean-climate regions of western Australia, is estimated to include approximately 8000 species of vascular plants. Thus, this region overall is similar in plant species diversity to the Cape Region of South Africa. The core Mediterranean-climate area of southwestern Australia includes about 3611 species, with 2540 species in kwongan habitats. Woody perennials in four families—the Myrtaceae, Proteaceae, Fabaceae, and Epacridaceae—dominate the flora. Much of the high species diversity within these families is due to extensive adaptive radiation within a few large genera. Large genera for the region include Acacia (400+ species), Eucalyptus (100+ species), Grevillea (200+ species), Stylidium and Melaleuca (150+ species each), and Hakea and Caladenia (100+ species each).

Nodes of unusual species diversity are present along the south coast of western Australia (Stirling Range, Fitzgerald River area) and in the sandplains north of Perth (Mount Lesueur area). As in the synobs of South Africa, large numbers of endemics with highly local patterns of distribution also characterize kwongan. Local scale plant diversity reaches almost as high as areas in the western Cape Region of South Africa. Vascular plant diversity in a sample of 0.1 ha stands of heathland in southwest Australia exhibited a range from 43 to 103 species, while jarrah forests and mallee stands had a lower range of 17 to 55 species. As many as 48 species in 100 species in 1 m², and 110 species in 100 m² quadrats, have been found at Mount Lesueur, with a high turnover as species move to comparable nearby areas.

At the regional level, southwestern Australia exhibits major differences in centers of highest diversity among the most important woody genera. Some genera—as, for example, Banksia (Proteaceae), Adenanthera (Proteaceae), Leschenaultia (Epacridaceae), and Eucalyptus (Myrtaceae)—are most speciose near the south coast or in southern kwongan and mallee communities. Other large genera have their highest diversity in northern kwongan—Grevillea (Proteaceae), Conostylis (Haemodoraceae) and Lechenaultia (Goodeniaceae). Finally, a large group of genera show bimodal patterns of diversity reflecting the both nodes of high species diversity in northern and southern kwongan—Calothamnus (Myrtaceae), Melaleuca (Myrtaceae), Hakea (Proteaceae), Darwinia (Myrtaceae), and Dryandra (Proteaceae). Two other genera, Acacia (Fabaceae) and Verticordia (Myrtaceae), are most diverse in the inland transition area of rainfall.

Three families are endemic to southwestern Australia: Cephalotaceae, Phyllocladaceae and Emblingiaceae. The latter of these is restricted to kwongan habitats. Endemism is notably high at the generic level, 103 endemic genera in southwestern Australia, and 33 of these are restricted to kwongan habitats. At the species level, endemism for the southwestern botanical province has been estimated to be about 48%.

B. Vertebrate Diversity

The vertebrate fauna of Australian MTEs shows no unusual level of high diversity of endemism. The majority of the fauna in these MTEs in made up of populations of more typically arid or mesic habitat species whose ranges extend into southwestern or southern Australia. Thus, the majority of the vertebrate appears to have relatively broad ecological niches rather than specialized requirements unique to the MTEs.

For mammals, the Mediterranean-climate Australian fauna of 82 species (which represents one-third of the Australian mammal fauna) has only 16 species (20%) endemic to MTEs. Interpreting the patterns of distribution of large mammals is made difficult, however, by the strong impact that both aboriginal and European populations have had on this fauna. The extinction of a large megafauna in the late Quaternary left the Australian continent without large grazers or predators. The low level of endemism among southwestern and southern Australian birds is particularly evident. Of the 190 species in western Australia, only one are endemic to Mediterranean-climate regions.

Endemism is moderate among reptiles. For snakes, 23% of the 43 species native to Mediterranean-climate Australia are endemic. The level of endemism is slightly higher for lizards, with 29% of the 144 resident species endemic. The level of endemism for terrestrial vertebrates in southwestern Australia is highest for amphibians. Among the 36 frog species in this region, 64% are endemic.
MEDITERRANEAN-CLIMATE ECOSYSTEMS

IX. CONCLUSIONS

Mediterranean-type ecosystems provide many opportunities for comparative studies of the controlling factors in the evolution of biodiversity. These five regions share common characteristics of climatic regime, with an independent evolution of plant and animal species adapted to these conditions within each region. Thus, they provide a natural system experiment with five independent replications. The value of such comparative studies stems from the fact that these regions lie not only with their similarities but with subtle differences in climatic conditions, topographic diversity, evolutionary history, and human impacts that have led to the patterns of diversity that we see today.

There has been a long history of comparative ecosystem studies between California and Chile, and between South Africa and southwestern Australia; and these have traditionally devoted some attention to comparing and contrasting how natural processes operate. The remarkable biodiversity of MTEs, together with the large number of rare and endangered species in these regions, gives a special significance to expanding studies of these regions to better understand the evolution of diversity, particularly by vascular plants. Serious threats of habitat transformation and degradation today make it critical that there be a better understanding of the conservation biology and sustainable resource management in all five MTEs.

See Also the Following Articles

AUSTRALIA, ECOSYSTEMS OF: FIRES, ECOLOGICAL EFFECTS OF: NEAR EAST ECOSYSTEMS, ANIMAL DIVERSITY: NEAR EAST ECOSYSTEMS, PLANT DIVERSITY

Bibliography


Mediterranean-type ecosystems provide many opportunities for comparative studies of the controlling factors in the evolution of biodiversity. These five regions share common characteristics of climatic regime, with an independent evolution of plant and animal species adapted to these conditions within each region. Thus, they provide a natural system experiment with five independent replications. The value of such comparative studies stems from the fact that these regions lie not only with their similarities but with subtle differences in climatic conditions, topographic diversity, evolutionary history, and human impacts that have led to the patterns of diversity that we see today.

There has been a long history of comparative ecosystem studies between California and Chile, and between South Africa and southwestern Australia; and these have traditionally devoted some attention to comparing and contrasting how natural processes operate. The remarkable biodiversity of MTEs, together with the large number of rare and endangered species in these regions, gives a special significance to expanding studies of these regions to better understand the evolution of diversity, particularly by vascular plants. Serious threats of habitat transformation and degradation today make it critical that there be a better understanding of the conservation biology and sustainable resource management in all five MTEs.

See Also the Following Articles

AUSTRALIA, ECOSYSTEMS OF: FIRES, ECOLOGICAL EFFECTS OF: NEAR EAST ECOSYSTEMS, ANIMAL DIVERSITY: NEAR EAST ECOSYSTEMS, PLANT DIVERSITY

Bibliography


I. Introduction

II. Strict Metapopulations

III. Quantitative Effects of Spatial Variation

IV. Lessons from Metapopulation Theory

GLOSSARY

local population That part of the population of a species found in a particular habitat patch.

metapopulation The collection of local populations in a region.

patch An area of suitable habitat for a particular species or particular collection of species, ideally bounded by unsuitable habitat or habitat with different physical properties. Normally, it is one of many such areas in a region.

spatially structured population A population whose reproductive and survival rates vary over the region that it inhabits, and whose members stay long enough in a locality to experience the local reproductive and survival rates.

strict metapopulation (Also called a classical metapopulation.) A metapopulation satisfying the following conditions: (i) Local populations are partially isolated from one another and are frequently capable of sustaining themselves for several to many generations in the absence of immigration from other local populations, (ii) local population extinction occurs on a timescale of several to many generations, and (iii) migration between local populations leads to reestablishment of local populations following local extinction.

THE CONCEPT OF A METAPOPULATION has its beginnings in the suggestion of Andrewartha and Birch and others that some, if not most, local populations of organisms in nature frequently go extinct and are reestablished later by immigration from surrounding areas (Hanski, 1999). The metapopulation—the collection of all local populations—only persists if local extinction is balanced by recolonization. How this balance is achieved is an important focus of metapopulation theory.

I. INTRODUCTION

Because many human activities fragment natural habitats, humans may artificially create metapopulations or may decrease the density of local populations within existing metapopulations and may put the continued survival of natural populations at risk. Hence, metapopulations are a major topic in conservation biology. Regardless of whether or not a natural population is distributed over easily defined patches of habitat, on which distinct local populations may be recognized, essentially all natural populations are patchily distributed in space. Most ecologists believe that patchiness in space and time has functional roles in population dynamics, i.e., in the manner in which population densities change over time. Most important, population
One very obvious way that patchiness is important is through mate finding. A sparse population evenly distributed over an area may have a low reproductive rate because males and females encounter each other too infrequently for many eggs to be fertilized. Patchiness can be a solution to this problem. High local concentrations or aggregations of a species solve the problem of low encounter rates of males and females consistent with a low average density. Similar effects occur from predation. Individuals may find safety from predators in a group but be vulnerable when isolated. Such effects of sparse populations are called Allee effects and are potentially one reason why patchiness in space is important. However, disadvantages from clumped distributions in space might also occur, for example, due to competition. Clumps of individuals in space deplete resources locally and reduce population growth rates compared to what would be experienced if the population were distributed more evenly. A similar effect results if predators are attracted to clumps of prey. Rather than prey finding safety in numbers, it is possible that they may be more vulnerable in clumps if predators increase their numbers at these locations.

These explanations for the importance of patchy distributions involve interactions between organisms in the same species and therefore involve density dependence—dependence of the probability that an individual survives, or dependence of its reproductive rate, on population density. The most striking theoretical predictions for the effects of patchiness, however, are for interactions between species for which one or more interactions are involved with the dynamics of this occupancy fraction: the extinction rate of local populations ($c$) and the rate $e$ at which empty patches are recolonized per fraction occupied. In essence, the fraction of local populations going extinct in one unit of time is $e$. The probability of an extinct population being recolonized in one unit of time is $c$ because $c$ is measured proportional to the fraction of occupied patches and incorporates the idea that the probability of recolonization should depend on the fraction of other patches that are occupied. Thus, the change in the fraction occupied with time may be expressed as the following differential equation, called the Levins equation:

$$\frac{de}{dt} = cp(1 - p) - ep$$

If the rates $e$ and $c$ remain constant with time, it follows that in a metapopulation with a large number of local populations, the fraction occupied will reach an approximate equilibrium at the value $1 - e/c$, provided $c > e$, because at this value extinction is balanced with...
recolonization. If $e$ is greater than $c$, local extinction exceeds local colonization and the metapopulation as a whole goes extinct. Conversely, $c$ must exceed $e$ for the metapopulation to persist. This result is often called the threshold condition, and it is of major concern for applications of metapopulation theory in conservation.

Because actual metapopulations in nature never have infinitely many local populations, the system is never exactly at the equilibrium $1 - ec/c$ but rather fluctuates about this value. In metapopulations with only a few local populations there is a danger that all local populations will simultaneously go extinct, causing the extinction of the metapopulation even though $e > c$.

On what biological features do these colonization and extinction rates depend? Consider extinction. Local extinction can occur in many ways, some of which stem directly from the activities of humans in habitat destruction or modification and hunting; however, metapopulation models are concerned with the case in which local extinction is a natural and repeated phenomenon, whose rate may be influenced by human activities but whose presence is not. Thus, local populations may become extinct by chance because of small population size. The phenomenon of demographic stochasticity refers to independent chance events in the lives of individual organisms. It is impossible to predict how long any individual will live and how many offspring it will have, and these effects summed over individuals in a population cause it to fluctuate. In a small local population, there is always a definite probability of below replacement reproduction in any year; and in a small population a chance run of such years can lead to extinction. These probabilities decrease dramatically as population size increases, but there are other factors that can also lead to local population crashes, including unstable population dynamics and disturbance.

Two different forms of disturbance are commonly considered. An environmental disturbance is caused by an environmental event such as a storm or fire, and a biological disturbance is caused by the invasion of predators or disease. In the strict metapopulation scenario, the interest is in disturbance events that lead to extinction of the local population either coincident with the disturbance or shortly after, following weakening of the population by disturbance. At any time, disturbances must strike patchily in space to cause extinction of only a proportion of local populations, or the metapopulation as a whole would be lost. It is very common for environmental disturbance agents, such as fires and storms, to distribute mortality very patchily in space even though they may affect a large area in a short period of time.

In a strict metapopulation, to which Levins equation applies, recolonization of a patch is from immigration from occupied patches. Many organisms may have means of recovering from disturbance by regrowth from seeds, roots, resting eggs, or spores within a recently disturbed patch, but with a strict metapopulation such populations have not gone extinct. They have merely died back and have been regenerated from dormant stages in the life cycle or belowground parts. The strict metapopulation idea that colonization occurs from external sources is the reason for the formula $p$ for the recolonization probability. Indeed, the rate of regeneration from sources within a patch should not depend on $p$. Confusion of regeneration with recolonization could potentially lead to serious errors in calculations and inferences.

Recolonization can be from mobile individuals, seeds, spores, or other propagules that arrive from external sources in sufficient quantity to reestablish a local population. The value of $c$ of $c$ will be affected by a variety of factors, such as the spacing of local populations. With large distances between local populations, many propagules may perish after leaving their source population before arrival in suitable habitat. The sizes of local populations will also influence $c$ because larger local populations should be expected to send out more propagules. The size of the metapopulation, in terms of the number of local populations, is also important. Because the number of habitat patches is always finite, the fraction of patches occupied is not at the equilibrium value $1 - ec/c$ but rather fluctuates about this value; the smaller the number of habitat patches, the larger these fluctuations. If the number of habitat patches is too small, there is a danger of simultaneous extinction of all local populations eliminating the metapopulation even if $c > e$.

Habitat fragmentation by human activities may lead to a metapopulation structure of local populations even if this was not the preexisting situation for the natural population of these organisms. With continuing habitat destruction, $c$ will decline due to increased spacing and smaller size of suitable patches, and there is the danger that $c$ will decrease below $e$, especially because $e$ may increase as suitable patches decrease in size and support smaller local populations. Thus, even though there is habitat capable of sustaining local populations, the danger exists that local populations may be either small or so vulnerable because of other habitat alterations that the rate of local extinction exceeds local recolonization, which would lead to loss of the metapopulation altogether. Alternatively, if the organisms are not able to disperse between habitat patches, habitat fragmentation
is accompanied by zero colonization rates, and the metapopulation decreases to extinction at the rate $e$ the moment it is created.

The simple destruction of habitat reduces natural population sizes, putting species at risk independently of these metapopulation considerations. However, metapopulation theory emphasizes that the viability of populations in the remaining habitat may be seriously affected by its connectedness. A final feature of habitat destruction is that by reducing the areal extent of a population, the probability that extinction events occur simultaneously in different parts of the metapopulation increases. For example, fire might sweep through the whole area occupied by the metapopulation, or climate fluctuations might have severe effects in all parts of the metapopulation, endangering the metapopulation as a whole.

So far, the rates $c$ and $e$ have been treated as though they are independent of $p$, the fraction occupied. This need not be the case. It has been suggested, for example, that extinction rates may decrease as the fraction occupied increases because extant local populations may tend to be larger and less vulnerable to extinction when they are subsidized by immigration. This could be especially important in the recovery of local populations after chance population declines, thus staving off extinction. As such, this phenomenon is called the rescue effect. If the rescue effect is required for $c$ to exceed $e$, the metapopulation as a whole may be especially vulnerable to extinction due to chance fluctuations in the fraction occupied; if this fraction becomes too low, $e$ will exceed $c$, promoting further decline and total extinction of the metapopulation. There are many other characteristics and complications that one can consider in single-species metapopulations, such as variation in the sizes of local populations, their distances from each other, and variation in habitat quality—features that are discussed by Hanski and Gilpin (1997) and Hanski (1998).

## B. Multispecies Considerations: Predators and Parasitoids

Extinction and colonization rates may depend on other organisms. Indeed, as mentioned previously, as biological disturbance, local extinction may be a consequence of invasion of the local population by predators or disease. In this case, extinction rates vary with the abundance of predators within the metapopulation, which in turn depend on the abundance of prey. It is possible that only the prey has a metapopulation structure because predators may be more mobile, occupying all of the landscape (Harrison and Taylor, 1997). Their density must be taken into account in determining the extinction rate of the prey, but otherwise their inclusion does not result in major differences in the understanding of metapopulation dynamics, except possibly in terms of the effects of habitat destruction, as discussed later.

If both predator and prey have the same metapopulation structure, a metacommunity results. Simple predator–prey models of isolated local communities are often unstable, causing large fluctuations in their abundances and potentially leading to extinction of the prey or the predator (see Fig. 7). A metacommunity structure has long been suggested as one means by which predator–prey interactions may be stabilized in nature. The interaction between the two locally in space remains unstable. The prey is driven extinct by the predator, but provided the colonization rate $c$ for the prey exceeds its extinction rate due to predation and other causes, it is able to persist in a metapopulation. Similarly, although the predator depends on the prey, whose local extinction leads to local predator extinction, provided the predator can find patches sufficiently well it can also persist in the system. In this way, it is possible for predator and prey to have a stable equilibrium at the metapopulation level despite unstable dynamics on the local scale (Nee et al., 1997). This metacommunity structure may have important conservation implications. Habitat destruction may more strongly affect the predator than the prey because prey numbers in a region may be buffered at first by reduction in predator numbers, with the effects of habitat destruction on prey being stronger after the predator has been eliminated (Nee et al., 1997).

The potential for metacommunity structure for predator and prey is also of great interest in the study of biological control of insect pests and introduced weeds. Commonly, in biological control, one seeks as a control agent a predator or parasitoid (a parasite that invariably kills parasitized hosts) that is specialized on the pest organism and causes sufficient damage to the pest to reduce its numbers greatly. Such a control agent, however, may have an unstable interaction with its host, leading to large oscillations in agent and pest densities, which is an undesirable outcome. However, there are many successful examples of biological control in which the pest is maintained at relatively low and stable numbers. One possible explanation for such control is that local population dynamics are indeed unstable, but regional dynamics are stable because they are metapopulation dynamics (Briggs et al., 1999).
Communities are in different successional states and high species diversity due to the fact that different local when considered over space, the metacommunity has low species diversity because only the best competitor persists there for long. However, when disturbance occurs, interrupting the process and starting it over again.

Because both extinction and colonization occur patchily in space and time, the metacommunity consists of a mosaic of local communities in different stages of succession (Fig. 1). Each local community may have low species diversity because only the best competitor in the local community persists there for long. However, when considered over space, the metacommunity has high species diversity due to the fact that different local communities are in different successional states and have different species. It is possible for the metacommunity to be in equilibrium and maintain this high species diversity regionally. This equilibrium is dependent on the inverse ranking of competitive ability and colonizing ability, but it is not sufficient for an inferior competitor simply to be a better colonizer. It has to be better by more than a critical amount that depends on the colonizing abilities of superior competitors and the disturbance frequency (Hastings, 1980).

The disturbance frequency has a particularly important role in the maintenance of regional diversity. A higher disturbance frequency means that fewer patches have the best competitor and there are more patches available for other species. At too low a disturbance frequency, lower ranked species do not have sufficient opportunity for colonization and therefore may not persist regionally. Their \( s \) values may exceed their \( c \) values. However, at higher disturbance frequencies, lower ranked species may persist and coexist regionally with higher ranked species. If the disturbance frequency is too high, the best competitors, which are by assumption poorer colonizers, have too high an extinction rate, \( c \), relative to their colonization rate, \( s \), and disappear from the system. At extreme levels, only the very best colonizers can persist. It follows that with such competitive hierarchies, both high and low disturbance frequencies lead to low regional diversity. Diversity is maximized at some intermediate value of the disturbance frequency—an idea that is often referred to as the intermediate disturbance hypothesis (Connell, 1978). Disturbances here can be of the physical sort or of the biological sort provided the predators or other agents causing disturbance are generalists and therefore are dependent for their own persistence on all species in the system, not just particular species which they may control separately from other species. Specialist mortality agents may be important in the maintenance of regional diversity (Chesson and Huntly, 1997), but this idea is not included in the intermediate disturbance hypothesis.

The critical feature of the previously discussed mechanism of diversity maintenance is a metacommunity that exists as a successional mosaic, i.e., the system is diverse because local communities range over a variety of successional stages having different species that provide colonists of other local communities. Colonization moves succession along or reestablishes local communities following local disturbance (Chesson and Huntly, 1997). This successional mosaic model has been suggested to work with local populations consisting simply of single individuals, for example, single herbaceous plants in a meadow (Lehman and Tilman,
METAPOPULATIONS (1997). Although this is not exactly what one might think of as a local population, none of the previous discussion depends on the size of a local population. Death of an individual is then equivalent to local population extinction. Death of an individual may be brought on by the arrival of a superior competitor at the site, disturbance, or simply senescence. Disturbance or senescence open the site to invasion by any species. Otherwise, the site can only be invaded by a superior competitor. The idea that many species may coexist in an area by this mechanism operating on a small spatial scale is referred to in the literature as coexistence by competition–colonization trade-off. This idea is particularly useful in the case of plants or sessile aquatic animals that hold space (Lehman and Tilman, 1997).

These ideas on diversity maintenance in metapopulations have important implications for conservation. If habitat destruction lowers colonization rates, as discussed previously, competitively superior species may be most at risk (Tilman and Lehman, 1997) because their already poor colonizing abilities make them less tolerant of decreases in colonization rates. These models predict that habitat destruction eventually negatively affects all species, however. When the per unit colonization rate, $c$, of the best colonizer is reduced below the extinction rate of local populations, all species are doomed to regional extinction even though patches of suitable habitat remain.

III. QUANTITATIVE EFFECTS OF SPATIAL VARIATION

The discussion so far has made the assumption that the only thing we need to know about a local population is its presence or absence: Is it extinct or not? This is a rather crude accounting because surely the number of colonists or propagules sent out by a local population depends on the size of that local population. Patches may be of different sizes and qualities and therefore support local populations of different sizes and densities, issues that are active areas of research in metapopulation theory (Hanski and Gilpin, 1997). Moreover, local populations vary in size over time in any one patch and in space from patch to patch even when patches of identical size and quality are compared. Ignoring changes in local abundance over time (i.e., ignoring the dynamics of local abundance) would be justified if local population buildup after colonization occurs quickly to some local population equilibrium (a population size at which, on average, reproduction and immigration balance deaths and emigration), around which population fluctuations occur until extinction. A broader variety of behaviors of metapopulations can be examined by taking actual local population sizes into account. At the same time, we can depart from the strict metapopulation assumption that recolonization occurs from ex-
METAPOPULATIONS

ternal inputs. We shall now assume that there are two spatial scales in the system. The smaller scale is the spatial scale on which interactions between individual organisms occurs (i.e., it is the spatial scale of positive and negative density-dependent effects within species), the scale of competition between species, or the scale of predation, depending on which of these are important in the system of concern. This scale corresponds to the local population scale in the strict metapopulation sense and shall still be referred to in terms of the patch and the local population. The larger scale is the scale of the whole population, which corresponds to the metapopulation or regional scale in the previous discussion.

A. Single-Species Dynamics

To introduce the fundamental concepts, consider first density-dependent population dynamics applicable to organisms with annual life cycles. The dynamics of a local population in the absence of migration can be defined by plotting local population density (numbers per unit area), \( N_t \), at time \( t \) as a function of its density, \( N_{t-1} \), the previous year, as represented in Fig. 2. These curves are dynamical relationships, i.e., by applying them repeatedly, one can plot density as a function of time (Fig. 3). The straight line (relationship I) represents the density-independent case in which individual organisms do not interact with one another, and so an individual’s contribution to future genera-

\[ N_t = N_{t-1} \]

Thus, \( N_t \) is simply proportional to \( N_{t-1} \). Relationships II and III are two cases of density-dependence and this density dependence means that they are nonlinear, i.e., they are curved or humped relationships (certainly not straight lines). For relationship II, the resources sustaining a local population are strictly limited, fixing an upper bound on the local population density, which is achieved if the resources are efficiently utilized. The larger the population at time \( t \), the closer the population comes to using all resources and the closer the upper bound on population density is approached. Annual plant populations commonly accord at least approximately with this relationship, which is sometimes referred to as contest competition. Some insect populations may have dynamical relationships more like curve III. Above a certain density, the number of insects in year \( t + 1 \) is a decreasing function of the number of insects in year \( t \). One explanation for a relationship such as this is scramble competition: At high densities, a high proportion of the population may starve to death, which leads to loss of the resources that these individuals consumed before death. Thus, a large fraction of the resources that could have been turned into new individuals at time \( t + 1 \) is lost, and the population crashes.

The three different dynamical relationships of Fig. 2 give very different local population dynamics (Fig. 3). The linear case (relationship I) gives simply a geometric increase. Relationship II (contest competition) gives highly stable dynamics: The population quickly converges on an equilibrium population size, which is determined by the point at which the diagonal line in Fig. 2 intersects the dynamical relationship. Relationship III (scramble competition), however, gives irregular fluctuations referred to as deterministic chaos which result here from the tendency of scramble competition to cause population crashes following population buildup. What happens when local populations with these various dynamics are connected regionally? If there is no variation in population density in space, then regional population dynamics are the same as local population dynamics. However, if there is a large amount of spatial variation in population densities (a common occurrence in nature), regional population dynamics can be very different from local dynamics (Fig. 4). The density-independent case (relationship I), however, does not show different local and regional dynamics because individuals are not affected by density, and therefore population dynamics cannot be affected by spatial variation in density. On the other hand, the strong density dependence arising from scramble competition leads to corre-

\[ N_{t+1} = N_t \]

\[ N_{t+1} = N_t \]

\[ N_{t+1} = N_t \]

\[ N_{t+1} = N_t \]
FIGURE 3  Local population dynamics generated by the dynamical relationships of Fig. 2. Relationship I (density independent), ■ relationship II (contest competition), ■ relationship III (scramble competition), ▲.

FIGURE 4  Regional dynamics generated by the interaction of the dynamical relationships of Fig. 2 with spatial variation in population densities. Relationship I, ■ relationship II, ■ relationship III, ▲.
spondingly striking differences between regional and local dynamics. For the situation depicted, which is defined in detail later, chaotic fluctuations have been replaced by a stable equilibrium. For contest competition (relationship II), the effect of spatial variation on regional population dynamics is quantitative and not qualitative: The equilibrium density is decreased.

What is the explanation for these changes in dynamics at the regional level in density-dependent situations? To answer this question, we need to work out the dynamical relationship between \( N_t \) and \( H_t \), defining the dynamics of population density at the regional level. These regional densities can be defined as the averages of the population densities in local populations, weighted if necessary by patch area. Dynamical relationships such as those in Fig. 2 continue to define dynamical change within patches, but disturbance and fluctuations due to small population size lead to random deviations in these relationships. Thus, these local dynamical relationships are mean relationships converting inputs of population density into outputs, and some of the output population in any patch may then disperse to other patches.

The essence of the difference between local and regional dynamics can be understood by imagining that dispersal occurs early each year and for the rest of the year patches are isolated with mean change governed by dynamical relationships in each patch. Further imagining that after dispersal local populations exist at just two densities (high and low) in equal abundance, then the regional dynamical relationship is easy to derive. This is done in Fig. 3, in which it is assumed that low-density patches have inputs 1/3 the regional density, \( N_t \), and high-density patches have inputs 5/3 of \( N_t \). The solid curve defines the relationship of local outputs to local inputs (scramble competition) and the dashed curve defines the relationship between \( N_t \) and \( H_t \), i.e., the regional dynamical relationship. This regional relationship is found by connecting pairs of points on the local relationship corresponding to low- and high-density inputs and finding the midpoint of the lines joining these pairs of points. For example, the points A and B in Fig. 3 are one such pair of points, and M is their midpoint. The x coordinate of M is thus the regional input, \( N_t \), which is the average of the x coordinates (local inputs) of A and B. The y coordinate of M is the regional output, \( H_t \), which is the average of the y coordinates (local outputs) of the points A and B. The complete regional relationship is found by repeating this procedure for every possible value of the input density \( N_t \).

Comparison of the point M on the regional relationship with the point \( M' \) on the local relationship reveals the reason for the difference between local and regional relationships. These points both have the same input density, but \( M' \) is the output of the local dynamical relationship, whereas M is derived from averaging the outputs of the local dynamical relationship at two different input densities, which both give outputs less than \( M' \) due to nonlinearity (curvature) of the local relationship. Thus, nonlinearity in the local dynamical relationship combines with variation in local inputs reducing the hump on the regional dynamical relationship. A similar effect acting on the nonlinearity in the local dynamical relationship after the hump reduces the severity of the decline in the regional dynamical relationship, and it is the combination of these two effects that is responsible for the strong stability of the regional dynamics shown in Fig. 4. Nonlinearity and spatial variation also affect regional dynamics in the case of contest competition, but the effects are quantitative and not qualitative. The regional relationship for contest competition is shown in Fig. 6, in which it is assumed that 50% of local populations are extinct and the others have a density twice the regional density. Not surprisingly, the regional equilibrium is 50% of the local equilibrium, but this reduction, and the regional relationship in total, can be understood using the same averaging approach discussed for scramble competition.

Having local populations take on just two densities for any given average density is clearly unrealistic. However, the qualitative features of the previous results

![FIGURE 5](image-url)
occur with general sorts of variation commonly observed in natural populations. The most common situation in nature is a negative binomial distribution of possible local population densities, which means that densities are somewhat more clumped than random in space. The striking change in dynamics from local to regional found for scramble competition is enhanced in this case (Chesson, 1998). In the density-independent case (relationship I) the dynamics at the regional level are unaffected by variation from patch to patch: $M$ always equals $M'/H^{1/2}$, regardless of the pattern of variation. The nonlinearities, i.e., the deviations from a straight line relationship, present in relationships II and III, are necessary for variation in space to have an effect on the dynamics of regional density, and these nonlinearities occur from density dependence.

There are many causes for spatial variation in local densities, of which disturbance and fluctuations due to small population size were mentioned previously. Simple random dispersal processes also lead to some variation in density, but spatially varying environmental factors may have large effects on dispersal and on the ultimate distribution of organisms among patches. For example, in marine organisms such as reef fishes, local currents may transport larvae away from some places and toward others, and these currents may vary with time of year, storms, and other atmospheric conditions, potentially delivering dispersing larvae to reefs in a very patchy fashion. Moreover, organisms commonly actively seek particular places, and the relative attractiveness of different places may vary greatly. Organisms may also seek out their own species, and therefore any existing spatial variation in density may be magnified by later dispersal. Unstable population dynamics occurring locally in space can be a cause of local variation in density, at least in models. In the illustrations here, spatial variation in density was assumed to have a fixed relationship to regional density [e.g., low-density patches ($N_t/H^{1/3}$) and high-density patches ($N_t/H^{1/3}$)]. More realistically, local densities will not have fixed relationships to the mean, and therefore the regional relationship will in fact change over time as the relationship between spatial variation and regional density changes (Chesson, 1998). Nevertheless, the way in which nonlinearity and spatial variation change the regional relationship, illustrated in Figs. 5 and 6, remains the same despite all these complications. In particular, the stabilizing effect of spatial variation in densities on scramble competition continues to be seen in much more complex circumstances than illustrated here.

These quantitative considerations are very different from the colonization and extinction issues discussed previously for strict metapopulations, but the two approaches can be related: Colonization and extinction dynamics involve implicit density dependence: To count all occupied patches as equivalent, recently colonized patches must quickly increase in density to some sort of steady state, such as a local equilibrium density or fluctuation about a local equilibrium. Local dynamical relationship II would serve in this circumstance, with the mean output taking into account the probability of local disturbance. The example given previously with half the patches extinct corresponds to a $c$ to $e$ ratio of 0.5.

These quantitative considerations also apply to spatially structured populations that are not strict metapopulations, for example, when recovery after disturbance depends to a large extent on regeneration rather than recolonization and also when only part of the life cycle is spatially confined. For example, it is very common for marine organisms that live on a reef, intertidally, or on the ocean floor to have widely dispersing larvae so that only the adult stage has clear spatial structure. The adult may be physically attached to the bottom, as in corals, barnacles, and mussels, or if it remains mobile it may have a territory or at least a home range that is much restricted in extent relative to the distance traveled by dispersing larvae, as in many reef fishes. Such spatial restrictions mean that interactions between individuals are localized because the individuals are localized and are likely most strongly affected by individuals living nearby, for example, on the same coral head or part of the reef. Similarly, in terrestrial plants a spatial unit of major importance is the immediate neighborhood of other plants close enough.
METAPOPULATIONS

171

Insect populations often have spatially confined larvae and dispersive adults. Although it is the adults that disperse in this case, the fundamental principles are the same in all of these cases, and the effects of variation in density in space are similar, even though in no case does spatial structure of one part of the life cycle qualify these populations as metapopulations in the strict sense. The effects of nonlinear dynamical relationships that we have explored depend on the smaller scale, which corresponds to the local population in a strict metapopulation, being the scale on which these nonlinear relationships apply. Because these nonlinear relationships derive from density dependence, the smaller scale is also the scale of density dependence.

How the scale of density dependence is determined depends on how the individuals in a population affect each other. For example, for territorial coral reef fishes, one source of density dependence may simply be competition for space to set up feeding territories. Other fishes may compete for hiding places from predators. The scale on which fishes may move to secure territories or hiding places defines the scale of density dependence. Predators may cause density dependence in their prey populations by responding to the density of the prey, for example, aggregating in areas of high prey density.

The scale of density dependence is the spatial scale on which predators respond to variation in prey density. Predators may also increase in numbers where prey are dense simply because they have more to eat and therefore reproduce faster. The scale on which this effect occurs is bound to be much larger than the scale of aggregation of predators to variation in prey density. The changes in dynamics with a change in scale demonstrated in Figs. 3–6 should occur in all these instances when one compares the dynamics defined on the scale of density dependence with the dynamics that emerge at the regional or metapopulation scale (Chesson, 1998).

B. Predators and Parasitoids

Like the dynamics of colonization and extinction, a rich array of phenomena is uncovered when the quantitative effects of spatial structure are examined in terms of interactions between species. In particular, predator–prey and host–parasitoid dynamics (Fig. 7) are often stabilized by local interactions and spatial variation (Hassell, 1997). This phenomenon is best understood in host–parasitoid systems in which a critical issue is spatial variation in the risk of parasitism experienced by a host, which may result from variation in parasitoid
FIGURE 8 Local host survival rates and regional host survival rates as a function of parasitoid density with half the patches having no par-
asitoids. The risk of surviving is thought to be an exponentially decreasing function of parasitoid density, as shown by the solid line. Very simple two-level variation in parasitoids is as-
sumed for ease of illustration: no parasitoids in half the patches and twice the mean parasitoid density in the other half. The effect of averaging different risk levels, exemplified by Fig. 8, is to moderate the decline in host survival with increasing parasitoid density, which means that host crashes following increases in parasitoid numbers are diminished, tending to stabilize the host–parasitoid interaction regionally (Fig. 7). Although the reasons for spatial variation in parasitoid density are poorly understood in general, an important feature of predator–prey and host–parasitoid interac-
tions is the potential for the unstable nature of the interactions to generate variation in risk, which may then stabilize population dynamics on the larger spatial scale (Hassell, 1997).

C. Competitive Interactions

Competition–colonization trade-offs are one mecha-
nism allowing coexistence of competing species in a patchy environment. Quantitative approaches allow the consideration of other mechanisms and more general spatial variation. Species fail to coexist (i.e., competitive exclusion occurs) when the species best adapted to the environment depletes resources to levels that are too low for other species to persist. Because the environ-
ment varies in space, different species may be the best
adapted under different environmental conditions. Then, many species that depend on the same resources may coexist regionally because each has patches in which it is the best competitor, i.e., each has its own spatial niche. If each of these patches were a closed community (i.e., not connected to other places by mi-
gration), then only a single species would dominate in each locality. However, in the presence of migration, each patch may have many species because a species that is not the best competitor in that local community has its reproduction supplemented by immigration from places where it is the best competitor.

If environmental characteristics vary over time also, favoring some species some times and others at other times, local diversity is even more likely to be main-
tained at high levels because trends in local species abundances will be reversed with reverses in the rank-
ing of competitive superiority in a particular locality: Species tending toward low relative abundance will be partially restored to higher values. A phenomenon with the opposite effect has been discussed by Rosenzweig (1995) and may be particularly important in animals, especially those with complex behavior. Animals de-
tecting a competitively superior species may leave or keep away from a patch, an effect that tends to divide a landscape into exclusive areas containing only the best species for those areas (Rosenzweig, 1995).

The various effects of spatial environmental variation do not require strict metapopulations but simply spatial structure. This more general theory can be applied, for example, to the coexistence of insects, especially various communities of flies such as Drosophila and carrion flies, which lay eggs in ephemeral patches of food such as fruit, mushrooms, or dead animals. In general, although the larvae of several species are com-
monly found developing in the same patch of food, it is also a common finding that there is segregation between species in choice of food patches. Although the underly-
ing cause of the choice is not well understood, the effect is that it promotes stability very strongly (Ives, 1988).

A second example of the potential importance of spatial environmental variation is marine communities, espe-
cially coral reef fishes but also various sedentary inverte-
brates that compete for space, with similar effects appli-
cable to plant species. In marine communities, as discussed previously, dispersing larvae characteristi-
cally arrive at potential settling sites in a highly patchy manner. Such systems are said to be strongly affected by recruitment variation, where recruitment refers to the process of a larva settling at a particular in patch. Differences in the spatiotemporal patterns of recruit-
ment have the potential to contribute to species coexis-
there is a very simple graphical technique for understanding the effects of variation in recruitment rates in space, which serves to illustrate how spatial environmental variation and its interaction with population density may have important effects. Figure 9 assumes a community with two competing species and plots the proportion of new inputs to a local community as a function of the local environment and the regional abundance of the species. The solid curves correspond to two environment types. The x-axis is the proportion, $p_i$, of species 1 in the system as a whole. The y-axis is the proportion, $s_{i1}$, of the available space at the site taken by new settlers of species 1 during the time interval from $t$ to $t + 1$. The top curve refers to patches favoring settlement of species 1, whereas the bottom curve refers to patches favoring settlement of species 2. The difference between the different sorts of patches is in the relative arrival rates of larvae of the two species: More members of species 1 arrive at the top curve sites, and more members of species 2 arrive at the bottom curve sites. The species compete for available space at a site, which is assumed to be limiting and thus filled by the larvae arriving each year. Note that if there were only one sort of patch in the system, eventually one species would eliminate the other because it would have higher settlement success everywhere and would increase in abundance relative to the other species. For example, if all sites favored species 1 (Fig. 9, top curve), then the fraction of species 1 settling would always be higher than $p_i$. It follows that species 1 would steadily increase until it had taken over altogether, assuming that settled individuals of both species have the same mortality rates (a simple adjustment, however, extends this diagram to different mortality rates; Chesson, 1985).

In Fig. 9, it is assumed that two-thirds of patches favor species 1 and one-third favor species 2. This means that the average proportion of settlers, $s_{i1}$, that belong to species 1 regionally is $(2/3)$ the top solid curve + $(1/3)$ the bottom solid curve, which is given by the points dividing the vertical lines between the two solid curves in the ratio 1:2. The dynamics of regional settlement are then given by the wavy line between the two curves. Depending on the actual adult mortality rates of settled organisms, the regional proportion, $s_{i1}$, of species 1 at time $t + 1$ will lie somewhere between $p_i$ and the regional $s_{i1}$. Thus, the point where the wavy line crosses the diagonal is the equilibrium point for the system, i.e., the point where $p_i$ does not change over time. It is a stable equilibrium point: If the value of $p_i$ is less than the equilibrium, then the regional $s_{i1}$ is always greater than $p_i$, and therefore species 1 increases in relative abundance ($p_{i1} > p_i$). However, if $p_i$ is above the equilibrium, the regional $s_{i1}$ is less than $p_i$, and therefore species 1 decreases in the system. Hence, whenever away from equilibrium, the system moves back toward it over time.

These results apply no matter whether a patch continues to favor the same species forever or favors different species at different times, for example, if environmental conditions responsible for larval transport change over time. These results have to be modified, however, if instead all patches favor a particular species at a particular time but the favored species changes over time, for example, if environmental conditions applicable to a particular year favor one species over the other everywhere. In this case, the environment is varying purely temporally. The way temporal variation affects the dynamics of an organism is greatly influenced by life history parameters such as the adult death rate because this determines how long past effects of fluctuations in settlement are reflected by the age structure of the population. A low adult death rate means that a
species can persist over periods when it is at a competitive disadvantage without its population density declining too greatly. It follows that the ability of species to coexist as a result of temporal variation varies inversely with the adult death rates (Chesson and Huntly, 1997). With irregular temporal fluctuations, the system does not have a traditional equilibrium but nevertheless can still exhibit stable fluctuations about a mean value as depicted in Fig. 10.

Temporal variation is an important feature of many systems, and a mechanism such as that described here has been suggested for coexistence in tropical trees in which highly variable fruit production occurs, in annual plants in which seed germination may vary greatly between years but seeds survive during unfavorable times in the soil seed bank, and also in annual zooplanton in lakes in which a resting egg bank takes the place of the seed bank.

IV. LESSONS FROM METAPOPULATION THEORY

A. Conservation

Few populations in nature exist without important spatial structure. Although there is debate regarding how often this structure conforms to a strict metapopulation, metapopulations are an important tool in conservation biology because habitat fragmentation creates distinct local populations even if they could not be defined before human interference. There is no question that the inhabitants of small isolated habitat fragments are at risk of local extinction. Although a collection of such habitat fragments may genuinely be thought of as a metapopulation, recolonization rates may be so low that each local population will go extinct before any are recolonized. Metapopulation theory emphasizes the need, at a very minimum, for the recolonization rates to exceed the extinction rates. Habitat destruction not only puts populations at risk by reducing their total size and spatial extent but also, by altering the connectedness of the remaining habitat, it reduces the value of the remaining habitat.

Metapopulation theory also implies that even uniform reduction in recolonization rates with habitat destruction may have different impacts on different species in the system. Therefore, the system may change in character with habitat destruction even under conditions that may not appear to be biased in favor of one species or another. Such effects have the potential to disrupt the functioning of ecosystems. This is not specific to strict metapopulations. The metapopulation perspective also indicates that not just inhabited habitat is
suitable habitat. A species may be only temporarily absent or in low abundance in an area, and this habitat may in fact be playing an important role in the system. Empty habitat must be judged in terms of its potential role as part of a larger connected system. Short-term environmental fluctuations may play a role in changing the suitability of habitat patches. Patches that are currently unsuitable may have the potential to become suitable in the future. With the prospect of global climate change, the maintenance of a diversity of patch types and a landscape structure suitable for migration of organisms between patches are of major importance in reducing the loss of biodiversity.

B. Nonequilibrium Dynamics
The metapopulation perspective highlights the role of nonequilibrium conditions in space, i.e., situations in which local populations do not remain near an equilibrium population size. Simple predator–prey and host–parasitoid models commonly do not have both species coexisting stably in isolated patches. Nonequilibrium locally in space that occurs asynchronously in different patches has an essential role in stabilizing the spatial extensions of these models. Similarly, in successional mosaic systems, perturbation from equilibrium (domination by one species), locally in space, is critical to the maintenance of the system. Such essential roles for nonequilibrium conditions have sometimes been considered as meaning that the system is particularly robust to perturbations, including habitat degradation, by human activities. There is no reason to expect this to be true. Indeed, the message here is that equilibrium occurs on the larger spatial or temporal scales. Interfering with the system through habitat destruction, alteration in the disturbance frequency, alteration in mean environmental states, and alteration of the variance in environmental states all have major consequences and are therefore of serious concern even though on small scales the system may not be at an equilibrium.

C. Population Regulation
These metapopulation and structured population perspectives also have the potential to resolve the long-standing debate regarding the relative importance of environmental fluctuations versus density-dependent interactions within and between species. Andresen and Birch were the first to clearly articulate metapopulation structure as an important feature of nature (Hanski, 1990), but because of the prominent role of stochastic factors (factors with random elements), such as extinction and colonization, they believed that density dependence was of minor importance. The quantitative analysis here shows that, far from being of minor importance, density dependence within local populations has a major role in shaping the nature of population dynamics at the metapopulation or regional scale. Density-dependent processes lead to nonlinear dynamical relationships, and it has been shown how such relationships interact with fluctuations in time and space to give outcomes at larger temporal and spatial scales that are not predictable on the basis of the separate effects of these factors.

See Also the Following Articles
- Competition, Interspecific
- Disturbance, Mechanisms of
- Loss of Biodiversity, Overview
- Parasitoids, Population Dynamics
- Predators, Ecological Role of
- Species-Area Relationships
- Species Interactions

Bibliography
- Ives, A. R. (1988). Covariance, coexistence and the population dy-


I. Introduction

II. Methods for Measuring Microbial Diversity

III. Future Directions

GLOSSARY

clone A population of cells all descended from a single cell, a number of copies of a DNA fragment obtained by allowing the DNA fragment to be replicated by a phage or plasmid.

fingerprint A pattern produced by DNA fragments or lipids that represents a community or species.

function A process carried out by an organism or group of organisms.

genotypic Relating to the genetic composition of an organism.

oligonucleotide A short, single-stranded nucleic acid molecule either obtained from an organism or synthesized chemically.

phenotypic Relating to observable characteristics of an organism.

phylogeny The ordering of species into taxonomic groups based on evolutionary relationships.

polymerase chain reaction A method for amplifying DNA in vitro involving the use of oligonucleotide primers complementary to nucleotide sequences in a target gene and replication of the target sequences by the action of DNA polymerases.

pure culture An organism growing in the absence of all other organisms.

species A collection of closely related strains.

strain A population of cells all descended from a single cell.

MICROBIAL DIVERSITY is the measure of the number or relative abundance of microbial species in a local area or region. Measuring microbial diversity may involve counting individual species, numbers of functional groups, or units operationally defined by the particular method being used.

I. INTRODUCTION

Microorganisms encompass an extraordinary diversity, in both their taxonomy and their ecological functions. As a group, microorganisms range in size over several orders of magnitude, span the three taxonomic domains (Archaea, Bacteria, and Eukarya), and are uniquely capable of performing all known biochemical transformations. The existence of such an immense variety of organisms, combined with the fact that microorganisms are too small to see without magnification, makes the task of measuring microbial biodiversity challenging and even daunting.

Measuring the biodiversity of microbial communities
is important for immediately practical and more fundamental considerations. Understanding microbial biodiversity may translate into benefits for biotechnology; management of agricultural, forest, and natural ecosystems; biodegradation of pollutants; reclamation of damaged lands; waste treatment systems; and biological control systems. On a more basic level, microbial processes underlie many essential processes shared by all ecosystems. However, because of our current minimal understanding of these groups of organisms, microbial communities are treated as black boxes in most studies of community, ecosystem, and landscape ecology. Increased knowledge of microbial biodiversity would clarify the details of many ecosystem processes and help generate the baseline information needed to predict and perhaps ameliorate the effects of global climate change.

This article discusses major issues involved in measuring microbial biodiversity and provides an overview of methods for measuring microbial diversity in natural, managed, and engineered ecosystems. The content is intended for biologists and ecologists who desire to incorporate the often-neglected microbial taxa into surveys of biodiversity.

A. Biodiversity of Microorganisms

The unit of measurement used in biodiversity studies of macroscopic organisms is usually a taxonomic one: the species. Microbial taxonomy has undergone enormous changes in the past two decades. Traditionally in microbiology, a species was phenotypically characterized with organisms classified into taxonomic groups based on morphology, physiology, and metabolism. Recently, however, phenotypic classification has been replaced or supplemented by the use of genotypic analysis. Nucleic acid sequence information derived from the small subunit of the ribosome [16S ribosomal RNA (rRNA) for prokaryotes or 18S rRNA for eukaryotes] is used to determine the degree of similarity among groups of organisms and the evolutionary relationships of microorganisms and all other life-forms. Certain analytical methods use the DNA sequence that codes for the rRNA rather than the rRNA sequence itself. The term used to describe this DNA is ribosomal DNA (rDNA). Although genetic sequencing for individual microbes is labor-intensive, particularly when characterizing complex microbial systems, some microbial ecologists maintain that an in-depth understanding of recently described systems is not possible until this information is gathered.

Far more species of bacteria, fungi, and other microorganisms exist than have been described to date. For example, between 300,000 and 1 million species of bacteria are estimated to exist based on initial in-depth phylogenetic surveys of several specific ecosystems and on calculations based on the reannealing kinetics of DNA extracted from natural environments. In contrast, only approximately 3860 species have been isolated and described by culture-based analyses. As genetic sequence analyses progress, libraries of small subunit rRNA sequences from specific ecosystems are created and compared to sequences in the existing database of known species. Often, these new sequences exhibit a low degree of similarity compared with the sequences of known species.

B. Background on Measuring Microbial Biodiversity

Although measuring the biodiversity of animal and plant communities involves counting and identifying individual species, such an approach is neither practical nor feasible for microorganisms. Though individual microorganisms can be counted using microscopy, their morphologies do not reflect their diversity. Visualization reveals little that is defining about identity beyond the observations that an individual cell is a fungus or bacterium, is a particular shape (rod or coccus), or has certain cell wall properties (gram positive or gram negative). Furthermore, counting is often hindered by interference from the physical environment in which microorganisms live. Some of the traits used as classification criteria for microorganisms are relatively plastic and, given that horizontal gene transfer appears common among microorganisms, using phenotypic information to classify species in microbiology is problematic. Additionally, the end point of many methodologies used to characterize microorganisms is not species identification. Thus, extrapolating the concept of species used in classifying large eukaryotes (plants and animals) to microorganisms is questionable.

Another important issue in exploring microbial diversity is the question of whether taxonomic (e.g., species or family groups present) or functional (e.g., types of processes present) diversity is of interest. Some important microbial processes include the transformation of elements in biogeochemical cycles, decomposition of organic residues, biodegradation of pollutants, fluxes of gases to and from the atmosphere, symbiotic relations with plants, and parasitism of other organisms. In ecological studies, the processes carried out by microorganisms within an ecosystem may be of greater relevance than knowledge of the specific identities of organisms involved. A considerable amount of functional redundancy...
dancy among microorganisms is thought to exist in most communities based on the fact that there are far more taxa than processes. To date, specific functions have been difficult to measure and interpret, especially by nonmicrobiologists, in studies of ecosystem biodiversity. However, with the development of rapid molecular assays, measuring microbial community compositions has become more feasible, and taxonomic characterization is more widely used. Compiling functional and taxonomic diversity measurements is important given that the ecological roles played by organisms represented by newly discovered ssrDNA sequences are often unknown.

Numerous methods are available for characterizing microbial communities. Determining which approaches to use is a critical step and depends on the objectives of the study. Should the diversity of a community be measured by characterizing its individual members, a task perhaps too large to be feasible, or by directly characterizing the community as a whole? If taxonomic diversity is of concern, what level of resolution is needed? Given that microorganisms are so small, what is the appropriate physical scale on which to sample in a study of biodiversity? These points are discussed later.

C. Enrichment/Isolation versus Direct Measurement of Microbial Communities

Early studies of microbial diversity involved culturing bacterial isolates on enrichment media supporting their growth ("culture-based") followed by phenotypic characterization of these isolates based on their morphology, physiology, and specific metabolic capabilities. As the tools of molecular biology began to be applied to microbial ecology, it became evident that culture-based methods were failing to detect large portions of microbial communities (no more than 0.1–5% of bacteria in communities are detected using culture-based techniques). Rapidly growing strains of microorganisms were often favored over other members of the community. A culture-based approach therefore probably offers a skewed perspective of microbial diversity in many environments. Consequently, there has been a shift in studies of microbial diversity toward the use of molecular biology and other biochemical and molecular tools that do not require culturing of organisms. Characterization of taxonomic and functional diversity is conducted using information provided by nucleic acids, lipids, and other cellular constituents. If culture conditions of a microorganism are known, data are gathered after individual strains have been separated from a community using enrichment and isolation techniques. In other cases, especially in soils and sediments, biochemical constituents may be directly extracted from the environmental samples, bypassing the isolation step. Many of these methods are currently under development and new ones are being developed at a rapid pace. Those seriously engaged in studies of microbial diversity have to keep close tabs on emerging trends in the field.

Newer methods may introduce their own biases in the assessment of microbial biodiversity, although these biases are presumably not as severe as the culturing bias. For example, molecular methods often make use of the polymerase chain reaction (PCR), during which original copies of specific DNA sequence are amplified exponentially. Two important considerations are that amplification efficiency varies among DNA sequences and therefore PCR results are rarely considered to be quantitative, and that a DNA sequence with low amplification efficiency could conceivably be prominent in the environment but not appear in PCR-based diversity measurements (see Section II.B.1.).

These direct measurement techniques may also be used to amplify sequences of rRNA or messenger RNA (mRNA) in order to shed light on the distributions, relative activities, and gene expression (functions) of specific groups.

D. Level of Taxonomic Resolution

When measuring microbial taxonomic diversity, a decision must be made about the level of taxonomic resolution (e.g., species, genus, family, order, and domain) needed to sufficiently assess diversity. The very definition of species in microbiology is subject to intense debate. Though counting numbers of individual species can be relatively straightforward in studies of plants and animals, this approach is not feasible for microorganisms. There is an overwhelming amount of diversity among microorganisms and categorizing all types is beyond the scope and resources of most projects. In some cases, determining families, or even kingdoms, within a previously uncharacterized community or ecosystem provides novel and valuable information.

For instance, the DNA from microbial communities as a whole may be presented in the form of a "fingerprint," which provides a means to classify and describe systems. DNA fingerprints are an abstraction of the original genetic information, but they may be an efficient way to characterize complex microbial habitats in which there is little information linking community structure and community function. They also offer the means to generate hypotheses, test responses of systems to environmental changes and variables, and detect
shifts in microbial communities. Efforts may be concentrated on conducting sequence analysis on changing elements of a DNA fingerprint rather than on sequencing all the members of a microbial community.

E. Environmental Heterogeneity

Obtaining a representative sample of the microbial diversity of a particular environment may be very difficult. It is obvious that most environments are heterogeneous at the spatial and temporal scale of humans. This heterogeneity is even more pronounced at the scale of microorganisms, often in ways that are not immediately evident to humans. For example, microbial communities within the root zone of plants, in which available carbon is high and soil chemistry has been altered, differ substantially from immediately adjacent communities in unrooted soil. These different communities are often only millimeters apart from one another. Also, a community that temporarily forms in a pocket of decaying plant or animal matter may be strikingly different from a community only several centimeters away. Within sediments, communities change substantially with depth as a function of oxygen availability and carbon distribution. This concept is also applicable on the scale of individual soil aggregates. In assessing biodiversity, one approach is to identify important microenvironments within the ecosystem of interest and measure diversity within each major type of microenvironment. With this approach, the number of samples obtained may soon become overwhelming. Another approach is to combine samples from different microenvironments in an attempt to understand the heterogeneity within the entire ecosystem, in which case it may be difficult to interpret factors governing the diversity. An understanding of the underlying physical structure of an environment is a prerequisite to a good sampling design.

II. METHODS FOR MEASURING MICROBIAL DIVERSITY

Far more methods for measuring microbial diversity exist than can be presented here. The methods described here are commonly used and, in composite, represent the breadth of approaches available. Methods can be divided into those that involve the counting of cells, measurement of cellular constituents, or determination of activity. Most methods provide either taxonomic or functional information about communities, whereas only a few (e.g., nucleic acid-based methods) have the potential to integrate both types of information. The following methods were originally developed in pioneering studies of bacteria, but most are also suitable for studies of fungi and other eukaryotes. Commonly, studies of fungal diversity involve isolation and characterization of strains rather than being community-based. Increasingly, community-based molecular methods are being used for eukaryotic microorganisms.

A. Counting Microorganisms

Since the development of the microscope, microbial population numbers, and to some degree diversity, have been quantified by examining cells under high magnification. Selective growth media, provided either in petri plates or in most probable number (MPN) assays, permit enumeration of culturable organisms that can metabolize specific compounds. Plates and MPN assays thus give potential functional information about a community. Counting organisms by the latter method requires that organisms grow sufficiently to be detectable.

1. Microscopy

Major microbial groups, such as bacteria and fungi, can be distinguished by their morphologies as observed under a light or fluorescence microscope after staining the cells. The microscope provides a quick but relatively superficial survey of microbial diversity based on the sizes, shapes, and staining properties of microorganisms. The relatively low number of morphotypes among bacteria limits descriptions of diversity. To some extent, fungal diversity can be investigated by examining hyphal and fruiting body structures and spore morphology, if present. Taxonomic resolution can be improved by using dyes tagged with fluorescent monoclonal antibodies specific to groups or strains, which permit selective visualization of members of the group to which the antibody binds. Recently, development of fluorescent in situ hybridization (FISH) with specific DNA probes (described later) has made the microscope a powerful and highly quantitative tool for use in studies of diversity. The antibody and DNA tagging methods require initial characterization of the organism or group of interest in order to develop the antibody or probe.

2. Plate Counts

Plate counts exploit the fact that individual microscopic cells quickly grow into a colony visible to the unaided eye if provided with suitable growth conditions. Microbial diversity has been assessed by culturing microorganisms on solid nutrient media to which substances are added to specifically promote the growth of target
organisms and/or inhibit the growth of unwanted groups. For example, media containing complex organic substrates combined with low pH tend to select for fungi instead of bacteria. Colonies that grow on the plates are counted and may be further differentiated based on their color or other morphological properties. The diversity of cultivable microorganisms is estimated by counting the number of different colonies present on agar media inoculated with a dilution series of the microbial community. Diversity measured based on this technique is unlikely to yield comprehensive, ecologically relevant information about the microbial community as a whole. An advantage of using plating techniques is that strains that grow on plates can often be easily isolated, characterized, and identified by traditional methods of microbiology.

3. MPN Methods

MPNs make use of specifically designed culturing conditions to estimate the number of microorganisms in a community able to carry out specific functions. The MPN medium is designed to select a specific trophic group by providing carbon, energy, nutrients, and environmental conditions needed to support the growth of that group. Thus, iron reducers are selected by providing anaerobic conditions, a carbon source, ferric iron, and other nutrients. A series of dilutions of an environmental sample are inoculated into the MPN medium. Measurements of turbidity, substrate utilization, or product formation confirm cell growth or activity. The number of organisms capable of carrying out the particular function being investigated is estimated from cell counts from the series of dilutions and with standard statistical calculations. MPN estimation of numbers is obviously a culture-dependent method, biased toward those organisms able to grow and compete under laboratory conditions.

B. Analysis of Cellular Constituents

1. Nucleic Acids

Methods using information contained in nucleic acids, DNA and RNA, provide the most specific information about an organism and thus potentially the most detailed information about the biodiversity of a microbial community. Figure 1 provides an overview of nucleic acid-based methods used in studies of microbial biodiversity. Extraction of DNA or RNA from environmental samples is the first challenge in using nucleic acid-based methods. One approach is to first extract cells from environmental samples (usually from aquatic environments) and then extract cellular DNA or perform analyses on cells collected on filters (e.g., using in situ hybridization). In situ approaches are quantitative because taxonomic information can be correlated with actual cell numbers within the community. The other approach, direct extraction of DNA without first extracting cells, is more commonly employed in complex environmental media such as soil and sediment. Numerous approaches have been developed to overcome extraction problems associated with specific types of environments, such as soils with high clay or humic acid contents. Other methods have been optimized for rapid extraction to accommodate large numbers of samples. There is usually a trade-off between speed of extraction and quality of extracted DNA (e.g., with respect to purity of the DNA). The efficiency of DNA extraction associated with either the direct or indirect method is difficult to estimate and may range from 10 to 99%. Extraction efficiencies associated with different groups of microorganisms differ, which could skew conclusions about a community's diversity.

Nucleic acid-based approaches can be divided into those that employ PCR and those that do not. The PCR makes multiple copies of a specific fragment of RNA or DNA from a mixed pool of nucleic acid fragments and permits detection of sequences originally present at very low densities (Fig. 2). After PCR amplification, fragments are separated in a polymer gel on the basis of their length or nucleotide composition. Concentrations of the fragments form “bands” in the gel which are visualized by staining. The pattern made by the bands constitutes a “fingerprint” of a community. PCR is also useful for detecting specific strains, species, or phylogenetic groups in environmental samples and, if used this way, requires a priori knowledge about sequence variability/conservation in the species or group of interest. The use of primer sets that target evolutionarily conserved regions, such as universal bacterial or fungal primers, circumvents this problem. However, in highly diverse communities, universal primers may amplify so many different nucleic acid sequences that it is not possible to separate them on a gel (e.g., a “smear” may result). Primer sets that target evolutionarily variable regions are useful for differentiating closely related organisms.

PCR amplification efficiencies differ among individual DNA or RNA sequences, introducing PCR bias. This is of particular concern because it is difficult to distinguish artifacts of PCR from true contributions from members of the microbial community. For example, PCR-based diversity measurements may be biased toward populations with sequences most complemen-
Fig. 1. A schematic of commonly used approaches in nucleic acid-based microbial ecology studies (reproduced with permission from Head et al., 1998).

Finally, PCR may be used as a preliminary step in random cloning and sequencing DNA or RNA from environmental samples from which a 'library' of DNA sequences (clones) is generated to represent the diversity present within a community. PCR-independent methods provide the option of obtaining taxonomic or functional information from nucleic acid extracts or even from intact microbial cells. Hybridization methods entail combining complementary strands of nucleic acids. The complementary strands are genomic DNA in the case of solution hybridization. Other hybridization methods utilize short, single strands of DNA or RNA (oligonucleotide probes) complementary to conserved or variable portions of rRNA or genomic DNA. If the sample contains small amounts of target tRNA or DNA, some of these techniques can be used in conjunction with PCR to reach the threshold of detection. Table I provides a summary of nucleic acid methods commonly used in studies of microbial biodiversity.

a. PCR-Dependent Methods

i. Denaturing or Thermal Gradient Gel Electrophoresis. Gradient gel electrophoresis resolves double-stranded DNA fragments of the same size based on properties defined by their genetic sequence. First, the small subunit rRNA is amplified from microbial com-
The polymerase chain reaction is used to amplify portions of double-stranded DNA from environmental samples. PCR enables detection of microbial populations from very small sample sizes and from extreme environments containing uncultivable microorganisms.

Small subunit rDNA fragments of the same length are separated based on nucleotide sequence to produce fingerprint patterns. Isolated fragments may be excised and cloned for sequence analysis.

Intergenic Transcribed Spacer (ITS) Analysis

Fragments between the small and large rDNA genes are separated based on length producing fingerprint patterns. Isolated fragments may be excised to differentiate strains by sequence analysis.

Restriction Fragment Length Polymorphism (RFLP)

Small subunit rDNA PCR products are cut into smaller fragments with restriction enzymes and separated based on length to produce patterns that differentiate simple communities or strains.

Cloning and Sequencing

DNA PCR products are inserted in plasmid vectors which are taken up by bacterial cells. The cells are then cloned to generate many copies for sequence analysis and phylogenetic classification.

Hybridization techniques have all the specificity of PCR-based techniques, are without PCR bias, and have the additional advantage of a means to quantify microorganisms.

Solution Hybridization

This method takes community double stranded DNA through thermal dissociation and reassociation process to estimate community DNA complexity.

Membrane Hybridization

Group or species-specific probes hybridize with community DNA or RNA immobilized on a membrane to produce estimates of relative abundance.

Fluorescent In-Situ Hybridization (FISH)

Fluorescently labeled group or species-specific probes hybridize with RNA in intact cells. Individual cells may be counted and types of organisms quantified by microscopy.

Oligonucleotide array

Group, species-specific, or functional gene probes are immobilized on a solid support. Hybridization between probes and community DNA is detected by laser technology.
denaturing double-stranded DNA fragments. PCR products of the same size migrate down the gel at a similar speed until a portion of a particular fragment denatures. At that point the migration of the fragment virtually ceases. Individual bands often correspond to individual microbial strains; however, some strains may produce more than one band (if the strain contains more than one ssrDNA sequence).

The result of denaturing or thermal gradient gel electrophoresis (D/TGGE) analysis is a banding pattern that reflects the composition of a microbial community. Statistical analysis involves scoring the bands in the sample as present or absent for a particular migration location on the gel, performing a similarity index based on band information, and then evaluating the relationships of samples with cluster analysis algorithms that create dendrograms. Although D/TGGE patterns have been used directly to determine Shannon–Weaver diversity indexes, the potential for PCR bias necessitates that extreme care be taken in using these banding patterns as the basis for traditional diversity measurements. Often, up to 60 dominant bands can be discerned in a DGGE profile, whereas it is known that thousands of different species may be present in a community. Thus, D/TGGE does not reveal information about the entire community but rather provides a survey of the dominant members in the community (at least those sequences that amplify by PCR). D/TGGE patterns are particularly rich in information since bands in the gel can be analyzed by hybridization methods or excised for cloning and sequencing.

Figure 3 depicts a polyacrylamide gel of a DGGE fingerprint obtained from a bacterial mixed culture grown on the pollutant MTBE (lane 6) or on trypticase soy broth (a rich medium) (lane 7). Also shown on this gel are bands obtained from individual bacterial strains isolated from the mixed culture (lanes 2–5). The organism depicted by the bands in lane 5 is able to biodegrade MTBE.

**ii. Intergenic Transcribed Spacer Analysis** The final portion of the ssrRNA gene and the beginning of the large subunit ribosomal RNA (lsrRNA) gene contain highly conserved sequences, and primers can be designed to amplify the intergenic transcribed spacer (ITS) region (also called ribosomal intergenic spacer analysis). The lsrRNA gene is located downstream from the small subunit gene. The intergenic spacer region consists of both noncoding and coding regions; coding regions include transfer RNA genes in the case of bacteria (Fig. 4) and 5.8S rRNA genes in the case of eukaryotes such as fungi. The noncoding regions are poorly conserved and contain insertions, deletions, and mutations that do not have a direct effect on the fitness of the organism. Using ITS primers that are universal to bacteria or eukaryotes in a PCR reaction with microbial community DNA generates fragments of different lengths and sequences. These complex banding patterns, like D/TGGE patterns, can be interpreted as a measure of the diversity of the microbial community. As with D/TGGE patterns, a single species may be represented by more than one band. In addition, a single species may have more than one length of ITS region, and one band may represent more than one species.
Bacterial ITS analysis of whole soil DNA extracts from the crops cotton, tomato, or almond. Each bacterial DNA fingerprint is represented by a pattern that may be categorized by band location and relative intensity and then compared to other patterns by clustering algorithms or other statistical methods.

Because closely related strains may have similar 16S rDNA sequences but dissimilar ITS sequences, ITS analysis is an effective tool to detect diversity in situations in which DGGE cannot.

Figure 5 shows microbial communities from agricultural soils planted with three different crops discriminated by ITS analysis. The fingerprints were generated from PCR amplification of total soil DNA using universal bacterial ITS primers. Different lanes within each crop category represent different replicates or PCR reactions.

iii. Restriction Fragment Length Polymorphism
Small subunit rDNA may also be analyzed by restriction enzymes that recognize and cut double-stranded DNA at precise sequence locations. The fragments derived from the procedure are of different sizes and indicate restriction fragment length polymorphisms (RFLPs). In this method, microbial community ssrDNA is either amplified and then digested directly or individual strains are first cloned and RFLP analysis is performed on each clone separately. Restriction enzyme digestion of the ssrDNA for a single strain typically results in 2 to 20 DNA fragment lengths, which are separated by gel electrophoresis. The resulting banding pattern is unique and can be used as a fingerprint to distinguish strains or mixtures of organisms from each other. Whole community RFLP fingerprints are extremely complex, resulting in sometimes indecipherable patterns, making this method useful only in systems of low diversity. Cloning the gene first has the advantage of avoiding overly complex community patterns but the disadvantage of limiting investigations due to the additional cost and workload. In terminally labeled RFLP, one of the primers used in amplifying ssrDNA is fluorescently labeled for detection in an automated DNA sequencer. As a result, the fingerprint is less complicated than standard RFLP fingerprints because only one of the fragments formed by the restriction enzyme is detected in the analysis.

iv. Sequencing of ssrDNA Clone Libraries
The most direct method for analysis of ssrDNA sequences obtained from environmental samples is to clone and sequence them. In cloning, individual PCR products are integrated into a vector (vehicle for reproduction), such as a bacterial plasmid, so that individual ssrDNA sequences can be reproduced and analyzed. Although this has traditionally been a time-consuming approach, the increasing availability of automated DNA sequencers is making direct sequence analysis more rapid and widespread. The sequence data are used to construct phylogenetic trees, which represent the evolutionary relatedness of microbial species present in a given habitat. Figure 6 shows phylogenetic affiliations of 16S rDNA sequences of bacterial clones from a hot spring cyanobacterial mat community.

To date, it has been impossible to exhaustively sam-
FIGURE 6 A distance matrix phylogenetic tree constructed to include cloned and sequenced members of a cyanobacterial mat community. The sequences of previously undescribed clones are analyzed and fit with their most closely related relatives in the cluster diagram (reproduced with permission from Weller et al., 1992).

ple a microbial community using this approach. For instance, in diversity studies of soil bacterial communities, if 100 clones are picked for analysis, each sequence may be unique, suggesting that species richness is far greater than 100. The clones are chosen randomly; thus, the fraction of clones represented by a common sequence can be used to estimate the relative predominance of that population in the community. In addition to PCR amplification bias, artifacts produced by selective cloning of some fragments over others must be considered.

b. PCR-Independent Methods

i. Solution Hybridization Microbial diversity of an environmental sample may be estimated using solution hybridization (DNA–DNA reannealing kinetics) approaches similar to those used to determine the genetic similarity of two bacterial strains. In this approach, double-stranded DNA is extracted directly from an environmental sample, denatured, and then allowed to reanneal (return to double-stranded form). Reannealing of DNA can be monitored using spectroscopic techniques. The rate at which the environmental DNA strands reanneal can be compared with the rate at which perfectly matched DNA strands reanneal, and the difference gives a measure of the number of different genomes present in the sample. Results of these assays should be interpreted with two potential artifacts in mind. First, contaminants in the samples (i.e., humic acids in soil DNA extracts) can affect DNA reannealing. Second, since the size of a genome affects reannealing kinetics, necessary assumptions about the "average" size of a genome may skew absolute estimates of the number of genomes present. The genetic similarity between two communities is best estimated by labeling extracted DNA from one community and measuring the amount of label that hybridizes to DNA from the other community and then vice versa.

ii. Membrane Hybridization Specific taxonomic groups can be detected and quantified by Southern blotting with labeled oligonucleotide probes to complementary ssrDNA sequences. With this method, the sample is fixed to a membrane, and then through a series of steps the probes are allowed to hybridize to complementary portions of microbial nucleic acids. The probes are either radiolabeled or enzyme linked with fluorescent dyes for quantification by densitometry or fluorimetry. The DNA on a single membrane may be stripped of one probe and then hybridized with others to provide estimates of relative abundance of certain taxa or other phylogenetic information. This technique is more rapid than cloning and sequencing, but the appropriate quantification standards must be used. Precise quantification of the number of individuals per species, however, is not possible by this method because the naturally occurring number of copies of the ssrDNA varies among microorganisms. The membrane hybridization technique can also be used with ribosomal RNA, in which case it is called Northern blotting. Northern blots provide quantitative estimates of active taxa because the rRNA concentration is higher in those members of a community that are physiologically active.

iii. Fluorescent In Situ Hybridization FISH is a modification of oligonucleotide probing that allows spatial location of specific populations in their native habitat. With this technique, fluorescently labeled oligonucleotides hybridize to ssrDNA inside cells and are detected with an epifluorescent or laser confocal microscope. Microorganisms are first pretreated to create gaps
in their membranes that permit the labeled oligonucleotides to enter the cells. Group- or species-specific oligonucleotides may be labeled with differently colored fluorescent molecules such that multiple groups of organisms can be observed in a microscopic field. Microscope images can be analyzed automatically to reduce operator errors in identifying and counting fluorescent cells. In soil systems, in which the majority of cells are presumably dormant, FISH results may lead to an underestimation of the microbial diversity due to detection limits. Because this technique allows visualization of individual cells, results are not confounded by variation in the number of ssrDNA gene copies among different microorganisms. This method is problematic in soil systems because soil organic matter fluorescence interferes with the probes' fluorescent signals and because probes may bind to soil constituents.

iv. Oligonucleotide Arrays
A promising new variation of hybridization analysis involves attaching arrays of oligonucleotides to a solid surface rather than attaching sample DNA to traditional blotting membranes. In this technique, multiple oligonucleotide sequences are immobilized on a glass or plastic slide, and the sample DNA is radioactively or fluorescently labeled. After labeled PCR fragments or rRNA hybridize to complementary oligonucleotides on the slide, the signal is quantified using laser technology. As many as 10,000 different sequences can be deposited on a small microscope slide. Due to variation in hybridization kinetics, the results are not strictly quantitative but can show trends in population dynamics. An advantage of this technology is that, given sufficient DNA or RNA, it can be used with nucleic acids extracted directly from the environment without an intermediate PCR step. Interpretation of the wealth of data that results from this method remains a formidable challenge. However, it is currently the only technique that can detect thousands of different populations rapidly, and it likely will have a large impact on microbial diversity studies. As more environmental isolates are characterized and sequenced, oligonucleotide arrays will become more useful for studying biodiversity.

2. Fatty Acid Analysis
In addition to DNA-based identification methods, microorganisms can be classified based on their cellular lipid composition. Lipids can be used to identify individual strains or to characterize whole communities. The most commonly analyzed lipids are the phospholipid fatty acids (PLFAs). PLFAs are the primary component of cell membranes and are present in all living organisms. Fatty acids are distinguished on the basis of chain length, number and location of unsaturations, and location of substituents (e.g., methyls, hydroxyls, and cyclopropane rings).

In lipid analysis of individual strains, cells are grown according to a standard protocol and their cellular fatty acids are extracted. Chromatographic analysis of the fatty acids provides a fingerprint, which is compared to a database to determine identity species. In a modification of the method, PLFAs derived primarily from the cell membrane are analyzed instead of whole cell fatty acids. Analysis of PLFAs in environmental samples is based on an extraction procedure in which polar lipids are extracted and purified using silicic acid chromatography. The ester-linked fatty acids of bacteria, or ether-linked fatty alcohols of Archaea, are analyzed by capillary gas chromatography or gas chromatography with mass spectrometry (White et al., 1979). Figure 7 shows a correspondence analysis plot of PLFA abundance data for 100 sulfate-reducing bacterial cultures together with uncharacterized sulfate-reducing cultures enriched from sediments in Clear Lake, California. Bacterial strains (represented by points) which lie close together have similar fatty acid compositions. PLFA analysis is also used to characterize entire microbial communities in environmental samples. There

![FIGURE 7 A diagram summarizing PLFA abundance data for pure cultures of sulfate-reducing bacteria and uncharacterized strains enriched from polluted sediments in Clear Lake, California. A, Desulfovibrio; B, Desulfobacter, C, Desulfobulbus; D, Desulfotomaculum; E, Desulfobacterium, F, Desulfomicrobium, G, Desulfuromonas; H, other (see text); I, Desulfomonas, J, Desulfo bacterium; K, Desulfo cococcus, M, Desulfolobus; N, Desulfoisarina; O, Desulfomena.](image-url)
are three ways in which PLFAs can provide information about diversity: (i) The entire PLFA profile can be used as a fingerprint which reflects the composition of the total soil community; (ii) signature fatty acids can be used to detect specific subgroups within the community, e.g., sulfate reducing bacteria, methane oxidizing bacteria, fungi, and actinomycetes; and (iii) certain lipids are physiological indicators of environmental stresses, e.g., the ratio of saturated to unsaturated fatty acids, the ratio of trans- to cis-monoenoic unsaturated fatty acids, and the proportion of cyclopropyl fatty acids.

C. Measurement of Process or Function

1. Substrate Utilization Patterns

The measurement of substrate utilization patterns is commonly used to evaluate functional diversity of bacterial populations. The commercially available Biolog system, which contains 93 carbon and energy sources on one microtiter plate, is widely used because of its convenience. The plate is inoculated directly with a sample of a microbial community (e.g., a dilution of a soil or sediment suspension) and then incubated under aerobic conditions. Substrate utilization leads to the development of a blue color originating from a redox-sensitive dye. Comparison of well color development patterns among different microbial communities coupled with multivariate statistical analysis may yield information on the functional relatedness of the bacterial populations in the sample. A major drawback of this method is reliance on the activity of culturable organisms; thus, organisms unable to grow under the incubation conditions are not represented. There is also debate regarding the environmental significance of the carbon sources used in the microtiter plates. These carbon substrates were originally chosen for their ability to distinguish pathogenic bacterial isolates from one another, and therefore bear little relation to substrates encountered by microorganisms in their natural environments. In response to this criticism, plates that contain substrates more common in aquatic and terrestrial environments have been developed and used in some recent applications.

2. Enzyme Assays

A variety of enzymes have evolved to carry out the diverse metabolic processes in which microorganisms are engaged. Six major classes of enzymes have been defined that, in turn, are divided into subclasses on the basis of the type of reactions they catalyze. Table II provides an overview of common enzymes found in soil. Enzymes found in the environment may be associated with cellular cytoplasm or may be attached to the lipid membranes or surfaces of microbial cells. Other enzymes may be present outside of cells, either in solution or attached to particles in the environment. Enzymes are usually measured based on the type and amount of activity they catalyze under controlled conditions in response to added substrates. "Potential" rather than actual activity is commonly measured. Some enzymes are present in all organisms (protease and hydrolyase) and give general information about the intensity of biological activity in a sample. Other enzymes catalyze more specific reactions and provide information about the diversity of processes that potentially can be carried out by microbial populations within a given environment. These include enzymes associated with decomposition of polysaccharides (e.g., cellulase, chitinase, amidase, and xylanase), the transformation of phosphorus (e.g., specific kinds of phosphatase) and sulfur (e.g., arylsulfatase), lipid metabolism (e.g., lipase), lignin degradation, and the metabolism of specific environmental pollutants.

D. Computational Needs

Within the past decade, many large data sets describing microbial communities have been generated using the new biochemical approaches described previously. Much of the information potentially available in these data sets is overlooked because biologists often do not employ the statistical approaches necessary for handling such large data sets. Interesting information may be buried in improperly or unanalyzed data; therefore, serious attention must be given to this problem. One challenge is reducing these large data sets to smaller sets of variables or components. In this way, classification tools can be used and models can be constructed to predict the response of microbial processes to changes brought on by human activity or environmental changes.

Standard multivariate techniques are commonly used to explore the "natural" relationships within a data set (principal component analysis and correspondence analysis) and to determine the correlation of environmental variables to specific components (constrained ordination techniques such as canonical correspondence analysis). In addition, regularized discriminant analysis (RDA) and partial least squares (PLS) are used to test the significance of a priori knowledge about the data sets (e.g., the significance of environmental...
variables or sample category). RDA uses a regularized covariance matrix estimate for the conventional statistical discriminant analysis methods. PLS is a multivariate calibration tool and can also be applied as a modeling method to solve pattern classification problems. PLS finds a set of latent variables in the measurement variable space (e.g., lipids or DNA banding patterns or sequences) that have a maximum covariance with the dependent variable space. Variable selection can then be used to choose which variables are most important for modeling. Artificial neural net analysis can be used to explore nonlinear relationships within the data sets, (e.g., relating community composition to process rates). Using these approaches, it may be possible to address such questions as which traits are shared by all members of certain classes (e.g., communities on a particular land use and vegetation type) and which components of a community are linked to specific processes such as pollutant degradation and nitrogen cycling.

### III. FUTURE DIRECTIONS

Great care must be taken to not overstate the reliability of microbial diversity measurements currently available. None of the diversity measurements currently available have the ability to (i) exhaustively sample a community, (ii) quantify population densities, and (iii) resolve individual populations. Ecologists who study macroorganisms have the ability to satisfy all three of these criteria and therefore it is difficult to compare the diversity of large organisms with that of microorgan-
isms. It is likely that a large subset of microorganisms exist at low population densities and are dormant. When this fraction of the community encounters environmental conditions suitable for growth, the dormant populations expand, and only then do we have the ability to measure them. Thus, current techniques tend to be biased toward dominant members of a microbial community. Despite this limitation, ecological studies of microorganisms and this limited perspective have already revealed overwhelming biodiversity.

Once the tools are developed to characterize and analyze data describing microbial communities, numerous questions can be considered. Questions of interest may include the following:

1. What environmental gradients (e.g., vegetation, time, soil texture, and topography) are most strongly correlated with the composition of specific communities or patterns of communities in general?
2. What other (e.g., nongradient) factors also regulate community diversity?
3. Do communities undergo succession within a year? How much do communities vary year to year? Are there seasonal patterns common to all communities across biomes?
4. Are there associations between microbial communities (e.g., facilitation, competition, and interaction)?
5. Does human disturbance alter communities to a greater degree and in different ways than do seasonal and vegetation changes, and are there general patterns that are indicative of disturbance?

See Also the Following Articles

**MEASUREMENT AND ANALYSIS OF BIODIVERSITY**
- MICROBIAL BIODIVERSITY
- NUCLEIC ACID BIODIVERSITY

Bibliography

MICROBIAL DIVERSITY

Paul V. Dunlap
University of Maryland Biotechnology Institute

I. Introduction
II. The Scope of Microbial Diversity
III. The Biological Significance of Microbial Diversity
IV. A New Era in Biological Sciences
V. The “Delft School” of General Microbiology
VI. The “Woesean Reformation” of Microbiology
VII. Major Groups of Microbes
VIII. Concluding Comments

GLOSSARY

aerobe An organism that utilizes or requires the presence of oxygen for growth.
anaerobe An organism able to grow in the absence of oxygen.
aerotrophic An organism able to utilize carbon dioxide as its source of carbon.
barotolerant and barophilic Able to tolerate high pressures and growing better under high pressure.
biofluorescence Light production by living organisms.
chemotrophic Organisms that utilize chemicals as sources of energy.
cryptoendolithic Living within the surface of rocks.
elective culture The provision of appropriate physical and chemical conditions that elicit the growth of specific metabolic types of microbes.
extremophile An organism that grows better at, or requires for growth, extremes of temperature, pressure, salinity, or other environmental factors.
halophile An organism requiring high levels of salts for growth.
heterotrophic An organism that obtains its carbon from organic carbon compounds.
microbe Single-celled organisms, such as bacteria, archaea, protists, and unicellular fungi.
phototrophic An organism utilizing the energy of light, as in sunlight, for growth.
polytrophic An organism that grows better at low temperature or requires low temperature for growth.

MICROBIAL DIVERSITY can be defined as the range of different kinds of unicellular organisms, bacteria, archaea, protists, and fungi. Various different microbes thrive throughout the biosphere, defining the limits of life and creating conditions conducive for the survival and evolution of other living beings. The different kinds of microbes are distinguished by their differing characteristics of cellular metabolism, physiology, and morphology, by their various ecological distributions and activities, and by their distinct genomic structure, expression, and evolution. The diversity of microbes presently living on earth is known to be high and is thought to be enormous, but the true extent of microbial diversity is largely unknown. New molecular tools are now permitting the diversity of microbes to be explored rapidly and their evolutionary relationships and history...
to be defined. The purpose of this article is to define the scope of, and highlight major themes in, our current understanding of microbial life and to describe recent progress in expanding knowledge of the evolution and biological significance of these organisms.

“The key to taking the measure of biodiversity lies in a downward adjustment of scale. The smaller the organism, the broader the frontier and the deeper the unmapped terrain.” (Wilson, 1994)

I. INTRODUCTION

Rapidly accumulating evidence indicates that microbes most likely account for the vast majority of kinds of organisms on earth. Microbes carry out a stunningly diverse array of metabolic activities, several of which were instrumental in creating conditions for the evolution of other life forms. Through their colonization of diverse and extreme environments, their geochemical cycling of matter, and their biological interactions among themselves and with all other organisms, microbes define the limits of the biosphere and perform functions essential for ecosystem development and health. However, because microorganisms are predominantly unicellular life forms that generally are smaller than can be seen with the unaided eye, they historically have received disproportionately little scientific attention compared to that given to animals and plants. This lack of attention has begun to shift recently as awareness of the diversity of microbes and their biological importance has grown. Of the three presently recognized domains of life, two, the Bacteria and the Archaea, are entirely microbial, and the third, the Eucarya, through its vast array of protists and fungi, is primarily microbial. The essence and full scope of the diversity of microbes is revealed in the dramatic differences among these microorganisms in their phenotypic characteristics of cellular metabolism, physiology, and morphology, in their ecological distributions and activities, and in their genomic structure, expression, and evolution. Appreciation for the true extent of microbial diversity is growing rapidly through the development and use of molecular phylogenetic approaches, which are enabling rigorous analysis of the origins and evolution of microbial life. In combination with classical methods of elective culture, isolation, and phenotypic analysis, the approaches of molecular phylogeny are stimulating the discovery of multitudes of new microorganisms, opening up their biology for study, and providing a clear understanding of the importance of microorganisms as the functional and evolutionary foundation of the biosphere.

II. THE SCOPE OF MICROBIAL DIVERSITY

We live on “a microbial planet” (Woese, 1999) in the “Age of Bacteria” (Gould, 1986). Microorganisms, the first cellular life forms, were active on earth for more than 3.0 billion years before the development of multicellular, macroscopic life forms. During that time and continuing into the present, through the invention of a spectacular array of different metabolic and physiological capabilities, microbes evolved to exploit the multitude of environments and microhabitats presented by the abiotic world. They thereby obtained the cellular building materials and energy necessary for growth and reproduction. In so doing, however, they progressively altered the geochemical conditions of the planet, leading to a continual development of new conditions and habitats, abiotic and biotic. Those new conditions and habitats presented both challenges to survival and opportunities to exploit, leading to continuing evolution of distinct microbial types able to endure or take advantage of the biogeochemical changes taking place on earth. Once cellular life began, it is likely that no place on earth containing the molecules and energy conducive to life remained abiotic for long.

The evolutionary trend toward greater complexity, seen in the relatively recent appearance of multicellular life forms (e.g., plants and animals), however, did not cause microbes to be displaced. The appearance of plants and animals did not shunt the unicellular microbes to forgotten corners of the biosphere to hang on and eke out a marginal existence. Instead, multicellular organisms, which themselves can be viewed as highly evolved, complex assemblages of microorganisms, have provided unicellular microbes with a wide variety of new habitats to colonize and exploit. Consider the various microbes whose growth is favored by the different and changing habitats provided by the growth and senescence of roots, stems, leaves, flowers, and fruits during the life of plants. Consider the multitude of physico-chemically distinct habitats of the human skin, of our mucous membranes, and the changing environments of our complex intestinal system. Along with these habitats, colonized often by assemblages of several different kinds of microbes, consider the species-specific developmental and metabolic symbioses certain bacteria have established with plants, such as nitrogen-fixing Rhizo-
Microbial Diversity

FIGURE 1 Light-micrograph of a section of the light organ of the sepiolid squid Euprymna scolopes. The animal harbors a dense population of the luminous marine bacterium Vibrio fischeri extracellularly within a ventral tissue complex, the light organ, and uses the light produced by the bacteria to predator avoidance. Reprinted from Claes and Dunlap (1999). Copyright © 1999 Wiley-Liss, Inc.

Microbes of one kind or another, especially bacteria, survive and grow almost everywhere on earth. Whether widely distributed, having been spread globally by winds, water currents, and animals, or occurring only in localized areas where they have adapted to grow under specific environmental conditions, microbes dramatically extend our perception of the limits of the biosphere. Previously, that perception was shaped to a large extent by our notion of the conditions under which plants and animals can live. As aerobic organisms, we require the high levels of oxygen present in air for survival. While many bacteria utilize oxygen as we do, most, however, grow best at lower oxygen levels, and anaerobic microbes of many different types require the strict absence of oxygen to survive, as found, for example, in sediment and gut tracts. Temperatures that from a human perspective are extreme—from just below the freezing point of water in some oceanic waters and sea ice, to several degrees above its boiling point in waters near hydrothermal vents and in hot springs—are not extreme at all to the bacteria that colonize these habitats. Indeed, water temperatures considered cold for humans (i.e., 15°C) can be lethally hot to true cold-loving, psychrophilic bacteria. Barotolerant and barophilic bacteria, active at and requiring the extremely high pressures of the deep sea, exist at pressures that would crush the human body, while other microbes, the halophiles, found in salt lakes and solar evaporation ponds, require salt concentrations that would quickly pickle our tissues. Add to this the cryptoendolithic microbes living just below the surfaces of sandstone in the Antarctic, the acid-tolerant and acid-requiring bacteria of acid mine drainage and sulfur springs, and the alkalinity-requiring bacteria of desert soils and alkaline lakes that live at pH levels that would be caustic to our skin. This array of microbial attributes gives a sense of the physical and chemical extremes at which certain bacteria survive and grow (Madigan et al., 1999). These organisms have no “protection” from the environment other than their inherent metabolism and physiology.
which, instead of protecting them, exquisitely suit them to living in these extremes, on which they are dependent. Major research initiatives around the world seek to expand knowledge of ‘extremophile’ biology identifying these ‘exotic’ microorganisms and bringing into study the metabolic and physiological attributes that adapt them to life at the physical and chemical limits of the biosphere.

What about microbial diversity in less extreme environments, away from the limits of the biosphere? Many environments, such as garden soil, coastal seawater, and lake sediments, do not exhibit such dramatic extremes of temperature, acidity, pressure, or other factors. Microbial life in these environments is strikingly diverse. While many different types of microbes can be isolated and grown in laboratory culture from these environments, most so far cannot. A typical expectation is that much less than 1 percent to a few percent of the microorganisms seen in an environmental sample will grow in culture (e.g., Amann et al., 1995). That means the vast majority of microbes, even from commonly studied environments such as coastal seawater, have not been brought into study or identified. Thus, discoveries of new types of microbes are waiting to be made, and they are being made. Examples of the unexpected include the discovery of magnetotactic bacteria, which synthesize intracellular chains of magnetic granules that orient the cells to magnetic fields (e.g., Amann et al., 1995) and the occurrence of new members of presumably anaerobic Archaea in oxygenated seawater at shallow depth (e.g., De Long, 1998). These reports demonstrate that microbial life in more accessible and commonly encountered environments is still poorly understood and far more diverse than presently known or expected. Therefore, most habitats that are known to be rich in microbial life and that have been sampled, such as seawater, soil, and animal guts, have not yet yielded for study anywhere near their full complement of microbes. And what about easily accessible habitats likely to be rich in microbial life but that have not yet been examined at all or at best have been sampled only minimally? One type of habitat is the gut tracts of the several hundred thousand known insect species, only very few of which, such as the termite, have been examined microbiologically. Yet the gut tract of each species—because of the different foods the animal eats, the gut's specific morphology and physiology for digesting that food, and the environmental conditions under which the animal lives—is likely to host its own very different kinds of microbes. Our ‘microbial planet’ is largely unexplored.

III. THE BIOLOGICAL SIGNIFICANCE OF MICROBIAL DIVERSITY

Microorganisms are ‘the foundation of the biosphere’ (Staley et al., 1997), providing its ‘essential, stable underpinnings’ (Woese, 1999). Microorganisms have played and continue to play fundamental roles in the evolution of higher life forms on earth. They have done so and continue to do so through the essential ecological processes they carry out in obtaining the materials and energy needed for growth and reproduction. A primary example of the evolutionary role microbes have played is oxygenic photosynthesis, invented by phototrophic bacteria more than 2 billion years ago and which releases oxygen as a by-product of energy generation. Over time, this release of oxygen led to a gradual change in the earth’s atmosphere from reducing to oxidizing. The oxidizing atmosphere, as it developed, allowed energetically more efficient aerobic organisms to evolve and provided a protective shield of ozone against ultraviolet radiation for terrestrial and aquatic organisms. Equally striking examples are the bacterial endosymbiotic origins of chloroplasts, light-harvesting organelles in plants, and of mitochondria, energy-generating organelles, major events in the evolution of plant and animal lineages in the Eucarya. Furthermore, the fixation of atmospheric nitrogen, reducing nitrogen gas to ammonium and converting it into organic nitrogen forms, is an entirely bacterial activity, carried out by various symbiotic and free-living microbes (Madigan et al., 1999).

The ecological processes carried out by microorganisms are equally fundamental. For example, global biogeochemical cycles of major elements, carbon, nitrogen, sulfur, and iron, essential components of all living cells, operate through microbial activity. Specifically, degradation of complex carbohydrates such as chitin, forming the exoskeleton of arthropods, and cellulose, hemi-cellulose, and lignin, structural polymers in plants, is essential. Without microbial conversion, these polymers would accumulate, removing huge amounts of carbon from the biosphere and blocking a multitude of biological processes that allow micro- and macroorganisms to live. In the absence of these microbial degradation processes, life on earth would soon falter. Besides the microbial degradation of complex organic compounds, consider the range of metabolic diversity in microbes, from oxygenic and anoxicogenic photosynthesis, sulfate reduction, methanogenesis, denitrification, iron oxidation, nitrite oxidation and nitrate reduction,
hydrogen and methane oxidation, and so on, all ways by which microbes obtain the energy necessary for growth and reproduction. These considerations form in part the basis for a commonly held view that bacteria and other microbes, in carrying out these processes, serve humans and other higher organisms as environmental recyclers and bioremediators. That view, while essentially correct, overlooks an essential point—these activities and processes are the fundamental biology of this planet. Microbes are the biosphere; their activities create and provide the foundation for all other life.

To gain a perspective on the significance of microbial diversity, imagine a biological survey crew tasked with discovering and documenting life forms on a newly encountered planet. Consider that upon landing, the crew found, remarkably, no macroscopic life. However, suppose that an initial sampling of a cubic centimeter of the planet’s surface revealed the presence of millions of discrete microscopic cellular entities. Imagine that with much additional analysis these entities were found by the crew to represent thousands of different kinds of organisms, distinguishable by their morphology or their dramatically different ways of obtaining the energy and nutrients necessary for metabolism and reproduction. Would the survey crew be surprised? What if upon analysis of the genetic material from the different types present, these microbial life forms were confirmed to be different and were found in many cases to be dramatically more distinct from each other evolutionarily than are seaweeds and humans, would the crew be impressed? What if the crew continued sampling, spreading out and choosing other locations of the planet’s surface, and found similar “species richness” wherever they looked, but often with little or no overlap in the types of entities present from one environment to the next. Would that start the crew thinking about, to paraphrase E. O. Wilson (1994), “a strange and vastly complex living world virtually without end”? It would, of course. However, there is no need to invoke new planets. This imaginary scenario describes the reality of microbial diversity on earth. The only difference is the microbially driven evolution of macroscopic life forms on earth, giving rise to the plants, animals, and macroscopic fungi.

IV. A NEW ERA IN BIOLOGICAL SCIENCES

Despite the complexity and richness of microbial life and the essential evolutionary and ecological roles played by microbes, awareness of microorganisms remains limited. The diversity and scientific importance of microbes have been largely passed over in human society, in science, in biology, and even in discussions of biodiversity (Hawksworth, 1991), overshadowed by attention to macroscopic forms. The prevailing and erroneous view for many biologists is that bacteria are “primitive, simple and relatively uniform” (Pace, 1996). This view developed naturally from early technical and scientific limitations for discovering and studying bacteria and other microbes in the 18th and 19th centuries, such as the need for high-resolution microscopy and the need to understand cellular structure, biochemistry, polypeptides, and nucleic acids. Such limitations did not hinder as starkly the beginnings of microbiology. Later, as limitations to the study of microbial life were overcome in the first three-quarters of the 20th century, the bias against microbes as scientifically important biological systems was nonetheless maintained and reinforced through the lack of a comprehensive phylogeny of microbes. This situation was especially true for bacteria, which generally lack distinctive morphological characters and for which sexual reproduction like that of plants and animals is absent. That bias remains largely extant today. For example, most college and university departments of biology and biological sciences are staffed predominantly with animal and plant biologists, with relatively few if any microbial biologists. Yet “the incongruity between the scientific perception of microbiology and the preeminence of microorganisms in the real world is astounding; it is worrisome; it cannot be scientifically justified or tolerated” (Woese, 1999). Fortunately, the bias and incongruity are beginning to be eliminated.

V. THE “DELFT SCHOOL” OF GENERAL MICROBIOLOGY

The confluence of two distinct but complementary approaches in biological science, one classical and one more recent, is leading to a shift in awareness about microbial diversity and its scientific importance. The classical approach is that of elective culture, also referred to as enrichment culture, by which new microbial types are brought into culture and isolated for phenotypic analysis. The elective culture approach, through the careful design of growth media and conditions, seeks to provide an appropriate physical and chemical environment that, when inoculated with an environ-
mental sample (mud, for example), will elicit the growth of specific metabolic types of bacteria postulated to be active or present in the sample. For success, the approach requires a thorough knowledge of biochemistry, good observational skills, and sensitivity to potential novelty in microbial metabolism and physiology. A recent example of this approach, demonstrating its central importance and value in microbial research, is the isolation of acetogenic spirochaete bacteria from the hindguts of termites (Leadbetter et al., 1999). Spirochetes are major members of the diverse microbial consortium resident in the termite hindgut (Fig. 2), and they have been thought to play key roles in the insect's nutrition, which is based on microbial degradation of cellulose and conversion to acetate, a major carbon and energy source for the insect. However, no spirochetes from the termite gut previously had been isolated in pure culture. That inability limited knowledge of the contribution these morphologically distinct and numerically significant bacteria make to host animal nutrition. Leadbetter and coworkers, however, successfully established culture conditions that favored the growth of spirochetes over other bacteria and that simultaneously encouraged the growth of bacteria able to form acetate, from H₂ and CO₂, breakdown products of cellulose degradation. In this way, acetogenic spirochetes from the termite hindgut were brought into pure culture for the first time. H₂/CO₂ acetogenesis, a type of metabolism previously unknown among spirochetes, reveals an important way, formation of acetate, in which these bacteria contribute to termite nutrition. The ability to design culture conditions that elicit the growth of specific bacterial types known or suspected to be present in an environment is central to development of our understanding of microbial diversity and microbial ecology. Once a novel microbial type has been brought into culture, that organism's special or unique cellular metabolism, physiology, and genetics can then be studied in detail.

This now classical elective culture approach developed through the insights of Sergei Winogradsky, a Russian soil microbiologist, and Martinus Beijerinck and later Albert Kuyper, Beijerinck's successor, in Delft, Netherlands. Their work was instrumental in forming the foundations of general microbiology and microbial ecology. When Cornelis van Niel, a student of Kuyper, moved in 1928 from Delft to the Hopkins Marine Laboratory in Pacific Grove in California, he brought with him the "Delft School" tradition in microbiology, which he continued and further developed through his research and through a course he taught in general microbiology and comparative biochemistry (van Niel, 1949). Through his course van Niel trained a generation of microbiologists. Those individuals have gone on to use the Delft School, van Niel approach, in their research, and they in turn have taught other generations of microbiologists, further disseminating the Delft School tradition. Most notably today, the elective culture and isolation approach to the study of metabolically and ecologically diverse bacteria is fostered in young microbiologists and other scientists through the Microbial Diversity summer course of the Marine Biological Laboratory in Woods Hole.
VI. THE "WOESEAN REFORMATION" OF MICROBIOLOGY

Despite the progress arising from the Delft School tradition in exploring and defining the diversity of bacteria and their metabolic capabilities, a major problem limited the development of general microbiology in the first three-quarters of this century: the lack of a unifying phylogeny of microorganisms. Various efforts at classifying bacteria and systematizing relationships among them, based on phenotypic characters of morphology and biochemical growth attributes, were attempted and abandoned. The frequent lack of characters, the instability or the shared nature of many characters, and the awareness that phenotypic characterization was not grounded necessarily at the genetic level left these efforts with major flaws. The inability to place microbes, especially bacteria, in an evolutionary context caused general microbiology and microbial diversity largely to languish as mainstream sciences at a time when other areas of biology and microbiology (e.g., clinical microbiology and biotechnology) were developing rapidly (Woese, 1999). What was needed was a unifying phylogenetic framework, founded at the genome level, which would allow the true evolutionary relationships among microbes to be analyzed critically and defined.

The use of informational macromolecules, begun more than 30 years ago, is now fulfilling that need for a unifying phylogeny of microbes (Woese, 1987; 1990) and is reforming our view of the evolution and diversity of microbial life. Analysis primarily of ribosomal RNA (rRNA) sequences, especially for the small subunit 16S and 16S-like rRNAs, has created the "first valid microbial phylogenetic systematics" (Jannasch, 1997). The functionally constant 16S and 16S-like molecules, common to all organisms, contain evolutionarily highly conserved regions, suitable for comparing less closely related organisms, and more variable regions, suitable for assessing evolutionary relationships in more closely related organisms. The universality of ribosomal RNA extends the value of this sequence comparison approach to all life forms, but importantly for bacteria it has established a unified phylogenetically based system with which to begin defining bacterial evolutionary relationships, a "first step in microbiology's reformation" (Woese, 1999). Along with 16S rRNA, other molecules, such as elongation factor Tu, 23S rRNA, and F1F0 ATPase, also provide substantial phylogenetic information and can serve as alternative markers for inferring relationships (Ludwig and Schleifer, 1999). A variety of opportunities exist at various institutions for learning the principles and application of molecular phylogenetic analysis in microbes and higher organisms. Notable among them is the Marine Biological Laboratory's course in Molecular Evolution, an intensive 3-week course dedicated to these topics.

A second step in the Woesean reformation of microbiology is the application of rRNA-based molecular phylogeny to microbial ecology (e.g., Hugenholtz et al., 1998; Pace, 1996). Sequence analysis of 16S rRNAs, for example, extracted from natural environments provides direct access to the diversity of bacteria in that environment, bypassing the need for culturing microbes and giving a rapid and potentially comprehensive assessment of microbial community composition. rRNA-based approaches have opened up for study many microbial activities and associations. The result is a rapidly expanding awareness of the diversity and ecology of microbes, both culturable and not-yet-cultured (Amann et al., 1995; Hugenholtz et al., 1998; Pace, 1996; Pace, 1999).

The strengths of the elective culture and molecular phylogeny approaches make them naturally complementary. The cultivation of a new microbe leads to acquisition of the organism's 16S or 16S-like rRNA sequence in the context of data on its metabolism, physiology and habitat. That sequence provides a defined point of phylogenetic reference and a highly specific tool with which to examine the organism's distribution in the environment. Equally exciting is the opportunity to examine and explore the environment for the full, natural microbial diversity present, without concerns about the bias inherent in and finesse required for culturing. Those explorations identify and define more deeply within a unified phylogenetic framework the diversity of microbes present while also offering potential insights into their metabolism and physiology. That information, in turn, can motivate more refined or novel attempts at cultivation of the sequence-identified microbes. Each approach nurtures and magnifies the strengths of the other. For microbiology in the next century truly to "emerge as the primary biological discipline" (Woese, 1999) the continuing confluence of these approaches must be encouraged.

VII. MAJOR GROUPS OF MICROBES

Through analysis primarily of 16S and 16S-like rRNA genes, microorganisms can now be placed in a potentially comprehensive phylogenetic framework, one that includes all living organisms and therefore is universal. Examination of the 16S and 16S-like rRNA-based uni-
The universal phylogenetic tree (Fig. 3) shows the three currently proposed domains of life: Bacteria, Archaea, and Eucarya. Future analyses, using expanding sequence data sets and markers of phylogenetic relationships other than 16S and 16S-like rRNA (Ludwig and Schleifer, 1999) will serve to test and refine the validity of this evolutionary grouping. Regardless, the current universal tree reveals in a simple, compelling way the dominance of microbial life forms. The true diversity of life is microbial.

A. Domain Bacteria

The unit of biological diversity is the species. The classical concept of a biological species, a reproductively isolated interbreeding or potentially interbreeding pop-
ulation of individuals, however, does not work for bacteria and other microbes that do not interbreed, that have undefined or cosmopolitan distributions, and that generally lack distinguishing morphological characters. So, in discussing bacteria, members of the domains Bacteria and Archaea, how is a species defined? The phylogenetic species concept, a group of individuals for which phylogenetic analysis has demonstrated a shared genealogical relationship (Alexopoulos et al., 1996), seems more workable for bacteria and other asexually reproducing microbes. The question then becomes the extent of genealogical relationship between two individuals necessary to designate them as members of the same species. Operationally for bacteria, if the genomic DNA of two strains is 70% or more similar, as determined by DNA-DNA hybridization analysis, they are considered the same species. With respect to rRNA, 16S rRNA sequence, similarity values below 97% are a strong indication that the two bacteria are different species (Amann et al., 1995; Madigan et al., 1999). The combination of these two approaches, both based at the genomic level, is powerful. When further combined with the identification of distinguishing phenotypic characters, such as cell morphology, motility, and flagellation, response to oxygen, requirement for organic growth substrates, growth temperature range, special metabolic attributes, characteristics of the habitat, and so on, a character rich, biologically meaningful description of a bacterial species can be obtained.

Conservative estimates place the total number of species of bacteria at 30,000 to 300,000 (e.g., Staley et al., 1987). As of 1999, approximately 5000 bacterial species had been described (Pace, 1999); a very small portion of the estimated total number, though many of these descriptions do not yet include rRNA sequence information. Most easily accessible environments, even those commonly sampled for microbial life, are likely to contain a multitude of as-yet-uncultured bacteria. Another important consideration in estimates of the number of bacterial species is the extent to which microhabitats, habitats relevant from the microbe’s viewpoint, have been sampled. A reasonable assumption is that most biotic and many abiotic surfaces are colonized by bacteria. Each surface presents a different microhabitat and therefore is likely to be colonized by a different individual bacterial type or assemblages of types, with the “many different microenvironments creating an almost infinite variety of selective conditions” (Palleroni, 1994), that is, conditions selecting for the specific metabolic and physiological types. An estimated total number of plant and animal species is approximately 9 million, with insects making up the majority of those species (Staley et al., 1997). Presumably, the external surfaces of these organisms, and also the mouths, gut tracts, and internal tissues of the animals, provide myriad types of microhabitats for bacterial colonization by assemblages of different microbes (Amann et al., 1995) and in many cases by specific bacterial symbionts, as in the various nutritional endosymbioses of between bacteria and insects. Consequently, these estimates of bacterial species must be too low. When the approximately 2 million species of fungi and protists, many of which undoubtedly also have largely different assemblages of microbes as well as specific bacterial symbionts, also are factored in, then the total number of bacterial species easily could exceed tens of millions. Future generations of microbiologists will likely find even this estimate to be conservative. Regardless, however, of what the actual total number of bacterial species turns out to be, bacteria are stunningly diverse.

Previously, all bacteria were grouped together taxonomically as prokaryotes. That grouping was based primarily on the common lack of a membrane-bounded nucleus in bacteria. Prokaryotic organisms, however, are now separated phylogenetically into two domains, the Bacteria (or Eubacteria) and the Archaea (formerly the archaebacteria). The Archaea exhibit many similarities to the Eucarya, demonstrating that the prokaryotic body plan is not a phylogenetically definitive character. The bacteria exhibit a wide variety of ways of obtaining growth the necessary carbon, as in the various kinds of heterotrophic and autotrophic microbes, and energy, as in phototrophic and chemotrophic microbes.

The domain Bacteria presently contains well over two dozen divisions, or kingdoms, of organisms (Fig. 4). Madigan et al. (1990) suggest, however, that the true number of kingdoms in the Bacteria may be 30 or more. As the number of these groups indicates, physiological diversity within the Bacteria is profound. The diversity is especially striking in two of the kingdoms, the Proteobacteria and the Gram-positive bacteria (Madigan et al., 1999), members of which exhibit a wide range of different metabolisms, physiologies, morphologies, and habitats. Examples among the Proteobacteria include the human enteric bacterium Escherichia coli, pathogenic pseudomonads, light-producing marine photobacteria, the fever-causing rickettsias, gliding bacteria and fruiting-body formers, stalked and budding bacteria, nitrogen-fixers, sulfate reducers, and so on. Within the Gram-positive kingdom are staphylococcal parasites of humans, the milk-sugar fermenting lacticocilli, and spore-forming soil bacteria, among many others. This diversity, however, probably reflects more the relative ease of cultivation and long-
B. Domain Archaea

The domain Archaea includes the majority of presently known "extremophiles," organisms that live at physical or chemical extremes. Archaea increasingly are being discovered, however, in less extreme types of environments, including the marine plankton, lakes, and sediments (e.g., DeLong, 1998; Ventricelli et al., 1999). Diversity within the Archaea is presently less well understood than in the Bacteria and Eucarya because the Archaea often require particular care to culture. Knowledge of how to culture the Archaea has expanded in recent years.

FIGURE 4 Diversity and phylogenetic relationships among members of the domain Bacteria. The division level grouping are of two types: currently recognized divisions, represented by cultivated bacteria (black), and provisional divisions, represented to date by environmental sequences (outline). The scale bar represents 0.1 changes per nucleotide. From Pace (1999), with permission.
years, and additional habitats supporting their growth are being actively studied, so the disparity in the numbers of different kinds of Archaea compared to Bacteria should progressively diminish. The Archaea consists at this time of three kingdoms, the Euryarchaeota, the Crenarchaeota, and a provisional kingdom, the Korarchaeota (Fig. 5).

Two major types of bacteria, the methanogens and the extreme halophiles, are included within the Euryarchaeota. Methanogens are bacteria that produce methane as an end-product of energy conversion reactions; they occur in a variety of strictly anaerobic habitats, such as sediments, sewage sludge digestors, the rumen of cattle, and the termite hindgut. CO₂ and similar compounds, methanol and other methyl-containing compounds, and acetate are used by different members of this group as substrates for methanogenesis, often with H₂ as an electron donor. Methane, primarily from methanogenic bacteria, is an important greenhouse gas, accounting for possibly as much as a few percent of total primary production. Some methanogens, such as Methanococcus jannaschii, are hyperthermophiles. The extreme halophiles, represented by Halobacterium salinarum, are bacteria that require for survival and growth the exceptionally high salt concentrations found in salt lakes and solar evaporation ponds. Other groups of Euryarchaeota include a lineage of extremely acidophilic bacteria containing Thermoplasma and Picrophilus, and two lineages of hyperthermophiles, represented by Pyrococcus and Archaeoglobus (Fig. 4). The Crenarchaeota contains a large and physiologically diverse group of hyperthermophilic, sulfur-metab-
olizing bacteria from terrestrial and marine hot springs and hydrothermal vents. Recently, several Crenarchaeota (or Euryarchaeota) have been identified at the 16S rRNA level as members of the plankton in cold oceanic surface and deep waters, coastal sediments, lakes, and in association with animals, indicating that the Archaea are more cosmopolitan in their distribution than believed earlier (e.g., DeLong, 1998; Vetriani et al., 1999).

A third kingdom of Archaea, the Korarchaeota, was established provisionally based on rRNA sequences obtained from samples of the Obsidian Pool hot spring in Yellowstone National Park and distinct from those of other Archaea (Pace, 1996). Attempts to bring these hyperthermophiles into pure culture are underway. The discovery of this third Archaean kingdom and the recent discoveries of Archaea in cold, oxygenated habitats clearly indicate that the true diversity of Archaea is likely to far exceed that based on presently identified species and sequences obtained from environmental samples. Implicit in this newly evolving view of Archaea is the substantially broader metabolic capabilities and wider ecological roles for this group of bacteria (DeLong, 1998; Vetriani et al., 1999).

C. Domain Eucarya

Microbial groups in the Eucarya are the Protista and the Fungi, organisms that, in contrast to the Bacteria and the Archaea, have a membrane-bounded nucleus. Endosymbiotic events are likely to have been major driving forces in the evolution of eukaryotes. Bacteria are thought to have diverged early from a universal prokaryotic ancestor, followed by the Archaea, both of which retained the prokaryotic body plan. Fusion of an archaean and a bacterium may have led to the nuclear line, which through symbiotic acquisition of phototrophic and nonphototrophic bacteria resulted in chloroplasts- and mitochondria-bearing eukaryotic cells. Loss of one or both of these organelles, or failure to acquire them initially, along with secondary symbiotic events, can be seen to account for much of the diversity of modern eukaryotes (Madigan et al., 1999).

Previous groupings placed eukaryotes into four kingdoms: animals, plants, fungi, and protists, with a fifth kingdom, Monera, to contain all the bacteria. Current understanding of phylogeny indicates that this five-kingdom system greatly underemphasized the diversity of bacteria while overemphasizing animals and plants, as described earlier. The five-kingdom system also underrepresented the diversity within the protists. The current, developing view, based on a rapidly increasing database of 16S-like rRNA sequence information, is that the Eucarya consists predominantly of unicellular microorganisms and that many phylogenetically deep, kingdom-level divisions exist within the protists (Sogin et al., 1996). Diversity within the protists dominates that of the other eukaryotic lineages (Fig. 6).

1. Protista

The Protista is a large complex grouping of mostly unicellular eukaryotic organisms. They are morphologically diverse and can be found in most terrestrial, aquatic, and marine habitats as free-living forms and as parasites of other protists, of fungi, and of plants and animals. With their nutritional modes restricted primarily to osmo- and phagoheterotrophy and phototrophy, protists are metabolically much less diverse than Bacteria and Archaea. Along with various independent amoeboid groups, major groupings include the Alveolata, composed of ciliates (e.g., *Paramecium*), dinoflagellates (e.g., *Alexandrium*), and apicomplexans (e.g., *Plasmodium*), and the Stramenopiles, composed of the brown and golden-brown algae, diatoms, cryptophytes, oomycetes, and distinct groups of slime molds, among other groups. Cryptophytes, Rhodophytes, and Haptophytes are other major groupings of protists. Along with these groups are the diplomonads, trichomonads, microsporidia, amoeba-flagellates, and euglenoids (Fig. 6).

2. Fungi

The fungi, sensu strictu, are commonly filamentous, multicellular heterotrophic organisms. Though previously thought to be similar to plants but lacking chlorophyll, fungi phylogenetically are not closely related to plants: Instead, fungi are seen now to have diverged from the animal lineage (Alexopoulos et al., 1996). Many fungi are saprobic, feeding osmotrophically on dead organic matter, and many are also parasites or symbionts of animals. As such, they share the limited range of metabolic capability of animals. Approximately 69,000 species of fungi have been identified, and more than 1,500,000 species are estimated to exist (Hawksworth, 1991). Four major groups (phyla) of true fungi have been defined (Alexopoulos et al., 1996) (Fig. 7). The Chytridomycota contains a single class, Chytridiomycetes, which uniquely among fungi produces motile cells during its life cycle. The motility organelle, a typical eukaryotic flagellum, probably was retained from ancestral protists (Beebee and Taylor, 1990). Chytrids play important ecological roles in decomposing organic materials. The Zygomycota contains two classes, the Zygoymy-
Microbial Diversity

Figure 6: Diversity phylogenetic relationships among members of the domain Eucarya. This evolutionary tree reveals the dominance of the protists over the fungi, animals, and plants. From Sogin et al. (1996), with permission.

cetes, which form thick-walled resting zygospores, and the Trichomycetes, obligate symbionts of arthropods. The Ascomycota, which form ascospore-carrying asci, contains most of the lichen-forming fungi. The Basidiomycota, which produces sexual basidiospores on specialized basidia, contains many of the commonly recognized fungi, such as mushrooms, puffballs, and bracket fungi. Plant pathogens in this phylum include the rust and smut fungi. Other fungi-like microbes commonly studied by mycologists now are grouped with the protists. These include the Stramenopile groups of oomycetes, hyphochytrids, and labyrinthulids, the plasmodiophorids, and the dictyostelid, plasmodial, and acrasid slime molds. The phylogenetic diversity and evolution
FIGURE 7  Diversity of fungi and the timing of their divergences. See Berbee and Taylor (1999) for details. Species names are abbreviated with the first five letters of the genus followed by the first three letters of the specific epithet for these fungi (e.g., Dermocystidium salmonis = Dermosal). From Berbee and Taylor (1999), with permission.
VIII. CONCLUDING COMMENTS

Understanding the extent and significance of microbial diversity “is the primary goal, the necessary objective, of biology in the next century” (Woese, 1999). Where progress toward that goal will take biology, however, is difficult to even guess, since the unexplored, unrecognized diversity of microbes is immense. However, the fusion of elective culture and molecular phylogeny is creating a revolution in our understanding of that diversity and of the ecological and evolutionary significance of microbes, especially bacteria. Molecular phylogeny also is overturning the previously limited view of protist diversity and is enhancing awareness of fungal phylogeny and diversity. Add to this microbial revolution, just underway in the beginning years of the 21st century, the developing influences of genomic sequencing, differential display and proteomics, and one sees a turbulent, radical, and exciting restructuring of biology underway, one that is being spearheaded by new awareness of the extent and significance of microbial diversity.

Important beyond measure in the success of that restructuring will be attracting talented students to pro-
fessions in microbial research. While biotechnological aspects of microbiology will continue to appeal to capable young researchers, many of the best students will be attracted not by the usefulness of microbes but by their inherent biological beauty and by the potential their study offers for insights into the evolution of life on earth.

Acknowledgments

The author is indebted to M. Sogin, N. Pace, J. Taylor, M. Berbee, K. Suberkropp, J. Breznak, T. Schmidt, J. Fuhrman, C. Vetriani, and K. Suberkropp, J. Breznak, T. Schmidt, J. Fuhrman, C. Vetriani, and J. Taylor for permission to use published figures. Research in the author’s laboratory is supported by a grant from the National Science Foundation.

See Also the Following Articles

MICROORGANISMS, ROLE OF
BACTERIAL BIODIVERSITY
MICROBIAL BIODIVERSITY, MEASUREMENT OF
MICROBIAL DIVERSITY

Bibliography


MICROORGANISMS (MICROBES), ROLE OF

Tom Fenchel
University of Copenhagen

I. Definition of Microbes
II. Prokaryotes (Bacteria)
III. Unicellular Eukaryotes (Protists)
IV. Roles of Microbes over Geological Time and in the Contemporary Biosphere
V. Microbes and Man

GLOSSARY

archaebacteria (Archaea) One of two major groups of prokaryotes, including methane-producing bacteria, extreme halophiles, and extreme thermophiles.
chemoautotrophy Metabolism exclusively based on the oxidation of inorganic compounds.
cyanobacteria A large group of eubacteria with chlorophyll a and oxygenic photosynthesis; formerly known as "blue-green algae."
eukaryotes Organisms with a compartmentalized nucleus in their cells; include animals, fungi, and plants as well as a large number of microbial groups collectively referred to as protists.
phagotrophy Feeding on particulate matter.
phototrophy Energy metabolism based on light energy.
prokaryotes Bacteria; cells without a compartmentalized nucleus.
protists Protozoa + protophytes; all microbial eukaryotes.
protophytes A diverse assemblage of partly unrelated groups of eukaryotic, phototrophic microorganisms.
protozoa A diverse assemblage of partly unrelated groups of eukaryotic, mostly motile and phagotrophic microorganisms; some phototrophic flagellates and fungi-like protists are traditionally included.
sulfur bacteria Bacteria that depend on the phototrophic or chemotrophic oxidation of reduced sulfur compounds.

MICROORGANISMS (microbes) are those life forms too small to be seen by the naked eye, i.e., that require a microscope or other form of magnification in order to be observed. The term microorganism is thus a functional description rather than a taxonomic one, and the grouping includes a wide variety of organisms.

I. DEFINITION OF MICROBES

Microbes are defined mainly by their size; most are unicellular, but some form colonies which may be macroscopic. Some "microorganisms" even exceed the size of the tiniest animals (metazoa). Taxonomically, microbes include the prokaryotes or bacteria (the eubacteria and the archaeabacteria) and members of various unrelated eukaryotic groups, together referred to as protists or protozoans (including some fungi-like organisms) + protophytes. Contemporary understanding
of phylogeny (Fig. 1) reveals that biological diversity is largely microbial. In this representation, multicellular organisms (plants, animals, fungi, and macroscopic algae) appear as minor and evolutionarily recent additions to the “tree of life.” This is also reflected perhaps by the fact that except for vascular plants (which are largely responsible for primary production on land) biogeochemical cycling of elements is mainly the result of microbial activity.

Although we have defined microbes in terms of size, it should be emphasized that the size range of microbes is immense. The smallest, free-living bacteria measure less than 0.5 $\mu$m and the largest protozoan cells measure 1 mm (some foraminiferan species measure 1 cm). The size range of microbes, 1 $\mu$m to 1 mm (a factor of 10^3 or a factor of 10^9 in terms of volume), approximately equals that of all vertebrates (guppies to whales).

Even so, some general properties of microbes that relate to their relatively small size can be identified. Among these, the most important is a high “rate of living”—that is, high weight-specific metabolic rates and high growth and reproduction rates. When organisms spanning a wide size range are compared, weight-specific metabolic rates tend to decrease proportionately to the 1/4 power of body weight, and generation times tend to increase proportionately to the 3/4 power of body weight. Under optimal conditions the generation (doubling) time of bacteria may be as short as 20 min and unicellular eukaryotes typically have generation times between 4 and 24 hr. Consequently, a relatively small microbial biomass may be responsible for a relatively large part of the flow of energy and materials in ecosystems. Certain microorganisms may in some circumstances form conspicuous accumulations of biomass in nature, but usually they are macroscopically invisible constituents of, for example, soils and seawater. However, due to their high weight-specific metabolic rates they dominate as agents of chemical transformations in the biosphere.

There are other characteristics of being small. Uptake of solutes from the surroundings (e.g., dissolved oxygen or organic substrates) is by molecular diffusion, a fact that is important in shaping microbial communities and which imposes constraints on physical transport rates. Due to their small size and low swimming velocities, microorganisms live at “low Reynolds numbers”; this implies that motility depends on viscous rather than on inertial forces. Many properties relating to motility and motile behavior of microorganisms therefore appear counterintuitive.

However, functional diversity of microorganisms is (as suggested by Fig. 1) more evident than their similarities. It is therefore expedient to discuss prokaryotic and eukaryotic microorganisms separately.

II. PROKARYOTES (BACTERIA)

A. Principle Properties of Bacteria

The prokaryotes include two major groups of life: the archaeabacteria (sometimes referred to as Archaea) and the eubacteria, which are differentiated by numerous genetic and biochemical traits. Structurally and functionally, however, they show so many similarities that it is appropriate to discuss them together. In contrast to eukaryotic cells, bacteria do not have a cytoskeleton and almost all bacteria are enclosed by a rigid cell wall. These features result in certain general properties. The limitation of diffusional solute transport from the surrounding water and within the cell typically constrains bacterial size to 1 or 2 $\mu$m. Certain giant bacteria (mainly among cyanobacteria and sulfur bacteria) measure 5–10 $\mu$m or more, but they usually include a large internal vacuole. Bacteria take up only low-molecular-weight solutes from their surroundings. Bacteria that depend on high-molecular-weight polymeric compounds as a source of energy and organic carbon must first hydrolyze their substrates extracellularly, using membrane-bound enzymes, before the resulting monomers can be transported into the cell. This transport
may be "passive" (facilitated diffusion) or "active" (ATP dependent). The small size of bacteria means that they can exploit and grow in extremely dilute substrate solutions (e.g., approximately 50 nmol per liter of amino acids or glucose).

Two additional properties explain the important role of bacteria in the biosphere: these apply to bacteria as a group rather than to individual species. One is that many bacteria are "extremophiles": Extreme thermophilic bacteria live at temperatures >80°C (or >100°C under hyperbaric pressure); extreme halophiles may grow in saturated brine; acidophilic and alkalophilic bacteria occur at a pH <2 and >10, respectively; and some bacteria tolerate toxic metal ions (e.g., copper and zinc) in millimolar concentrations. A variety of habitats (hot springs, brines, subsurface, etc.) are inhabited exclusively by prokaryotes. Many natural habitats lack oxygen (e.g., aquatic sediments, stratified water bodies, water saturated soils, or even the interior of soil aggregates and detrital particles) because diffusive supply of O₂ cannot meet the demand for aerobic microbial respiration. Many bacteria have an anaerobic energy metabolism and they are practically the only inhabitants of anoxic environments. It seems that bacteria have only one absolute requirement for metabolism and growth: liquid water. Active "terrestrial" bacteria (and other microorganisms) are therefore confined to water films surrounding soil particles, leaf litter, plant surfaces, and the like.

Another collective property of bacteria is their metabolic diversity, which far exceeds that found among eukaryotes. Metabolism of bacteria includes processes such as the oxidation and reduction of inorganic sulfur, nitrogen, iron, and manganese compounds and the production and oxidation of methane. Some bacteria can grow using hydrocarbons as a substrate: Some phototrophic bacteria use H₂, H₂S, or Fe³⁺ as electron donors (rather than H₂O as in algae, plants, and cyanobacteria) and many bacteria can assimilate atmospheric nitrogen. These are all fundamental (but exclusively bacterial) processes in the biosphere. In addition, all natural and many xenobiotic polymers can be hydrolyzed by at least one type of bacterium.

B. Metabolic Diversity of Bacteria

It is convenient to distinguish between dissimilatory (energy and catabolic) metabolism and assimilatory (anabolic) metabolism, although the distinction is not always sharp. Heterotrophic bacteria typically use a given organic substrate for dissimilation and a carbon source for growth, whereas autotrophic bacteria must use energy gained in dissimilatory metabolism for the reduction of CO₂. Energy generation and growth are coupled since most energy generated is used for the synthesis of macromolecules (DNA, RNA, and proteins), for active transport of substrates into the cell, and sometimes for assimilatory reductions. In microorganisms there is therefore typically an almost linear proportionality between energy generation and reproductive rate. Here, we concentrate on energy metabolism because it provides an overview of the functional diversity and roles of bacteria in the biosphere, notwithstanding that the synthesis of bacterial biomass is also an important process in terms of providing a basis for phagotrophic food chains.

Table I presents a (simplified) overview of major types of bacterial energy metabolism. Different types of microbial processes predominate in different habitats according to the chemical environment (availability of substrates and electron acceptors), to thermodynamics (energy yields of the different processes), and to certain physiological constraints. Many of the listed processes are interdependent in nature in that one functional type of bacteria requires the presence of other types of bacteria.

Under oxic conditions, aerobic bacteria completely dominate in accordance with the fact that oxygen respiration yields more energy than any other metabolic process. Bacterial diversity in aerobic habitats is to a large extent due to differential abilities to hydrolyze different polymers, but most species are capable of complete mineralization of organic monomers to CO₂, H₂O, and mineral N or of the complete oxidation of reduced inorganic compounds.

Anaerobic bacteria, on the other hand, are specialists and a complete mineralization under anaerobic conditions requires several physiological types of bacteria. Fermenting bacteria play a key role in anaerobic habitats because they are (nearly) the only anaerobes capable of hydrolyzing polymers. The complete fermentation of, for example, carbohydrates (to acetate + CO₂ + H₂) is thermodynamically possible only if the ambient H₂ tension is low (< about 10⁻¹⁰ atm). Low H₂ tension results from the activities of H₂-consuming bacteria, notably sulfate reducers and methanogens. This syntrophic interspecies H₂ transfer is an essential feature of anaerobic habitats. In general, metabolic end products of fermenting bacteria serve as substrates for anaerobic respirers. The type of respiration that will dominate depends on the availability of external energy acceptors and on the energy yield of the different types of oxidation. After aerobic respiration, nitrate respiration is the energetically most favorable process, but its...
TABLE 1
Principal Types of Energy Metabolism in Bacteria

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fermentation</td>
<td>Anaerobic processes in which (organic) molecules are dismutated. No external electron acceptor, redox equilibrium; low energy yield. Principal end products (of carbohydrate fermentation): acetate and CO₂ ( \rightarrow H₂ ), but many fermentation types are incomplete and metabolites then include other low-molecular-weight fatty acids and alcohols.</td>
</tr>
<tr>
<td>Respiration</td>
<td>Use of an external electron acceptor, electron transport phosphorylation. Aerobic respiration uses ( O₂ ) as terminal electron acceptor. A variety of organic substrates are used by different heterotrophs. Chemoautotrophs use ( O₂ ) to oxidize various inorganic substrates (reduced S, N, Fe, or Mn compounds and ( H₂ ) and ( CH₄ )). Aerobic respiration provides the highest energy yield of any known metabolic process.</td>
</tr>
<tr>
<td>Anoxic respiration</td>
<td>Use of external electron acceptors other than ( O₂ ). Denitrifiers use ( NO₃^-/H₂O ) (reduced mainly to ( N₂ ) or ( N₂O )) to oxidize various organic compounds and reduced S compounds (but not ( CH₄ )). Sulfite reducers produce ( H₂S ) while oxidizing especially ( H₂ ), or low-molecular-weight fatty acids; front and manganese reducers use ( Fe^{III} ) and ( Mn^{IV} ) for the oxidation of substrates.</td>
</tr>
<tr>
<td>Methanogenesis</td>
<td>Methanogenesis is found only among certain anaerobic archaeabacteria. Acetoclastic methanogens dismutate acetate into ( CO₂ + CH₄ ). ( CO₂/\text{H}_₂ ) methanogens produce ( CH₄ ) from ( H₂ + CO₂ ). Some methanogens can also produce methane from reduced C-1 compounds (e.g., methanol).</td>
</tr>
<tr>
<td>Phototrophy</td>
<td>Phototrophs use light energy for generation of ATP and (with an external electron donor) for the reduction of ( CO₂ ) to organic compounds. Oxygenic phototrophs (cyanobacteria) use ( H₂O ) as electron donor and produce ( O₂ ) as metabolite. Anoxygenic phototrophs (several groups of bacteria) use ( H₂S ), ( H₂ ), or ( Fe^{II} ) as electron donors, thus producing ( S^{2-}/H₂O ) or ( Fe^{III} ).</td>
</tr>
</tbody>
</table>

**FIGURE 2** A biosphere model incorporating different microbial respiration processes. Oxygenic phototrophs use light energy to produce \( O₂ \) and organic matter \( ([CH₂O]) \); this represents the chemical potential that maintains the biosphere. Microorganisms derive energy by redox processes involving oxidants (left) and reductants (substrates) (right). The standard potential of the half-cell processes decreases from above so that the oxidant has to be situated higher than the reductant if the process is to be thermodynamically possible. Oxidations of organic matter may take place stepwise (e.g., first by sulfate reduction and the resulting sulfide can subsequently be oxidized by \( O₂ \)). A few processes (e.g., \( N₂ \) oxidation) are not realized due to kinetic constraints. Data on free energy changes are not precise under physiological conditions because the processes do not necessarily occur under standard conditions and because only a part of the potential energy is conserved as ATP; however, the energetic hierarchy of the processes holds.

Quantitative importance is often limited by the availability of \( NO₃^-/H₂O \). When nitrate becomes exhausted, \( Mn^{IV} \) followed by \( Fe^{III} \) are the electron acceptors of choice; when these are exhausted sulfate respiration prevails (Fig. 2). In marine sediments, sulfate reduction is typically the dominant process because of the high concentration of \( SO₄^{2-}/H₂O \) in seawater (about 25 mM compared to about 250 \( \mu M \) at atmospheric saturation; the capacity for oxidation of organic matter through sulfate reduction is about 200 times more than through \( O₂ \) reduction and so, in contrast to \( O₂ \) \( SO₄^{2-} \) supply is not rapidly exhausted). Only when sulfate is depleted will methanogenesis be quantitatively important; this rarely happens in marine environments, but it is important, for example, in lake sediments. In anaerobic habitats, the microbial processes are therefore spatially or temporally
structured, for example, when moving downwards from the surface of sediments, O\textsubscript{2}, Mn\textsuperscript{4+}, Fe\textsuperscript{3+}, and SO\textsubscript{4}\textsuperscript{2-} are depleted sequentially. The reduced end products of anaerobic metabolism diffuse upwards and are eventually reoxidized by O\textsubscript{2} (or NO\textsubscript{3}\textsuperscript{-}) by chemooautotrophic bacteria residing at the aerobic–anaerobic interface. The complex interactions between different types of bacteria in anaerobic habitats and at the anaerobic–aerobic interface are often referred to as a “food chain” by microbiologists. It differs fundamentally from phagotropic food chains, however, in that the organisms do not feed on each other but rather one functional group of bacteria utilizes the metabolites of another functional group.

Hydrolysis of polymers is often the rate-limiting step in mineralization. Some substrates are inherently difficult to hydrolyze, notably the ligno–cellulose complexes of woody tissue and many other natural plant compounds such as phenols, waxes, tannins, and cork substances (whereas pure carbohydrates, such as cellulose and hemicelluloses, are relatively rapidly degraded under aerobic and under anaerobic conditions). Hydrolysis may also be limited by the availability of mineral nutrients (N and P) that are assimilated (“immobilized”) by bacteria during decomposition of mineral-poor substrates (such as most plant tissue and litter). Crude oils (which contain a mixture of normal and branched paraffins, aromatic hydrocarbons, and other compounds) are degraded by a variety of aerobic bacteria and anaerobically by nitrates or sulfate reducers. Otherwise, easily degradable monomers (e.g., amino acids) may bind to clay minerals or humic substances to become temporarily or permanently inaccessible to bacterial attack. Humic substances are complexes consisting mainly of aromatic rings deriving from lignin, quinones, and phenols. Humic substances are very resistant to microbial attack and their rate of turnover in nature must be measured in centuries.

Some polymers are not, or are only very slowly, degraded under anaerobic conditions. This applies to, for example, lignin and some other plant compounds. Anoxia in conjunction with a low pH, such as found in moors and swamps, strongly limits mineralization, and peat, in which the original structure of plant material is often preserved, accumulates. Even in marine sediments, a certain fraction of the organic input is not mineralized but fossilized as kerogen in sedimentary rocks. Peat is eventually (over geological time through abiotic processes) transformed into lignite and coal. Large accumulations of organic matter in marine deposits may eventually transform into crude oils and natural gas.

C. Roles of Bacteria for Element Cycling in the Biosphere

The single most important role of bacteria in the biosphere is the degradation and mineralization of organic matter produced by (mainly eukaryotic) phototrophs. In some habitats, however, primary production by prokaryotic phototrophs is also significant.

Mineralization processes are often complex due to spatial and temporal heterogeneity, limiting physical transport rates, and kinetic and physiological constraints. This complexity is especially evident where anaerobic processes are important as in sediments. There are considerable differences between major habitat types and these will be discussed separately.

1. The Water Column of Oceans and Lakes

Prokaryotic oxygenic phototrophs play a specific role in the water column. Thus, the unicellular, 1-μm-large Synechococcus cells are ubiquitous in lakes and in seawater, typically occurring at densities of approximately 10\textsuperscript{5} ml\textsuperscript{-1}. Their role in primary production (relative to that of eukaryotic phototrophs) is especially important in oligotrophic oceanic waters where it may account for 30–70% of the total production. The tiny Prochlorococcus has recently been shown to be important in the deepest parts of the photic zone (>100 m) in oceanic waters. Blooms of large colonial filamentous cyanobacteria (Trichodesmium) occur periodically in some marine waters; the organism is capable of N\textsubscript{2} fixation and is thus favored by N-rich but N-depleted water masses. Macroscopically conspicuous blooms of various species of colonial cyanobacteria are common phenomena in eutrophic fresh waters.

Aerobic heterotrophic bacteria, however, are the most important group of prokaryotes in the water column. They typically occur in numbers ranging between 5 \times 10\textsuperscript{5} and 5 \times 10\textsuperscript{6} cells ml\textsuperscript{-1} constituting a volume fraction in seawater (~10\textsuperscript{-2}) comparable to that of other important biotic components (e.g., phytoplankton). Their relatively constant numbers imply that they are controlled by grazing (mainly by protozoa) but they are also subject to viral attack, and that their turnover is relatively rapid (varying from <1 to several days).

Bacteria depend on dissolved organic compounds which again derive from several sources. The most important one seems to be the passive excretion of photosynthate from phytoplankton cells; it is estimated that between 3 and 40% of primary production is lost in this way from algal cells to become mineralized by bacteria. Other sources include allochthonous material (e.g., runoff from land), degradation of macroalgae,
MICROORGANISMS (MICROBES), ROLE OF

and “sloppy feeding” by zooplankton. Dissolved organic matter in natural waters covers a wide range of molecular size from monomers (amino acids and monosaccharides) to colloidal matter. Small monomers have a rapid turnover (sometimes <1 hr) and occur at very low concentrations (nM range) due to efficient bacterial utilization. Larger molecules are utilized much more slowly due to their lower diffusion coefficient and because their utilization requires extracellular hydrolysis. Some macromolecular compounds (humic substances) are very recalcitrant; they have a very low turnover rate but constitute the bulk of dissolved organic matter.

The fact that a relatively large part of primary production is channeled via dissolved organic matter to bacteria, which then enter phagotrophic food chains, has been termed the microbial loop (Fig. 3). The relative amounts of the primary production which is channeled through the microbial loop compared to the “classical plankton food chain” (large phytoplankters → copepods → planktivorous fish → carnivorous fish) depends on circumstances. The classical food chain is particularly favored by large or periodic influx of mineral nutrients (such as in upwelling zones); otherwise, the microbial loop seems to dominate in terms of carbon flow.

Special microbial biota are associated with aggregates (“marine snow”) which consist of different types of particles (diatom frustules, fecal material from plankton organisms, etc.) held together by colloidal material. They form in the water column where they constitute approximately 10% of nonliving organic matter; they eventually sink to the bottom and become the basis for benthic life. They are, however, rapidly colonized by bacteria (and other microorganisms) and so a substantial part of the organic fraction is mineralized while still suspended in the water column; it has been estimated that 75–80% of the organic matter is lost during a passage of 2000 m depth.

2. Aquatic Sediments

The most important property of sediments is that vertical solute transport depends on molecular diffusion. Organic material—produced in the water column—sediments and is in part mixed into the sediment through the activity of burrowing animals. The aerobic metabolism of microbes and animals deplete O2 at a certain depth, depending primarily on the input of organic material. In shallow-water sediments, anoxia typically prevails a few millimeters beneath the surface; at greater water depths the oxic zone typically measures several centimeters. Beneath the oxic–anoxic interface the (“chemoautotrophic”) mineralization is anaerobic. As previously explained, various other electron acceptors are used by bacteria for the oxidation of their substrates (Fig. 2). In marine sediments sulfate reduction dominates and, in shallow offshore sediments, may be responsible for more than 50% of the terminal mineralization. The resulting metabolite, hydrogen sulfide, is thus produced in copious amounts. Some of it combines reversibly with iron to form black FeS, but most diffuses upwards to become reoxidized by chemoautotrophic sulfur bacteria at the chemocline and therefore a major portion of the O2 uptake of marine sediments is due to the reoxidation of microbially produced sulfide. Some of the ferrous sulfide reacts (abiotically) with sulfide to form the somewhat more stable pyrite (FeS2), which appears as a type of fossil fuel in sedimentary rocks.

In sulfate-poor freshwater systems methanogenesis plays a major role as the terminal anaerobic mineralization process. Some of the CH4 formed is reoxidized aerobically in the sediment surface or in the water column; since CH4 (in contrast to sulfide) has a low solubility, much escapes to the atmosphere via ebullition.

In productive shallow waters the anaerobic zone may reach the surface of the sediments. This is macroscopically visible as a white cover of chemoautotrophic sulfur bacteria (they owe their color to intracellular elemental sulfur). If the sediments are exposed to light, the surface may instead be dominated by a purple layer of phototrophic sulfur bacteria. In stratified water bodies (such as thermally stratified lakes or where there is a salinity difference between the surface and deep waters), vertical, turbulent transport of dissolved O2 from the surface layers to the deep layers is prevented. The chemocline may then extend above the sediment surface and into the water column. The phenomenon is
common (at least seasonally) in deep, productive lakes and in fjords with a sill, the largest permanent anoxic water body on Earth is the Black Sea, which is every-
where anoxic and sulfidic below 150- to 200-m depth.

3. Terrestrial Soils
The microbial life of soils differs in many ways from that of aquatic sediments. The chemical composition of soil organic matter differs because it mainly derives from vascular plants and therefore includes large amounts of structural polymers (especially lignin and cellulose). Vascular plant tissue is very poor in mineral nutrients; C : N ratios typically exceed 60, which affects the microbial nitrogen cycle of soils. Another important feature is that soils represent a complex matrix that includes air-filled spaces. Since diffusion coefficients are approximately 104 times higher in air than in water, soils are not anaerobic throughout, although anaerobic microniches may occur within individual waterlogged soil particles, and therefore anaerobic microbial pro-
cesses do occur. In general, the biological activity of soils exceeds that of aquatic sediments reflecting the higher input of dead organic matter. Microbial activity in soils affects soil fertility and some processes (e.g., CH4 oxidation and denitrification) affect element cycling on a global scale.

The overall important, and variable, factor control-
ling soil microbial activity—qualitatively and quantita-
tively—is water content. Fungi (ascomycetes and basid-
iomycetes) are better adapted to water stress and their hyphae can penetrate air-filled spaces; in soils and litter they are therefore important rivals to bacteria as primary decomposers. Certain bacterial groups (including the spore-forming gram-positive bacteria and the fungi-like actinomycetes and myxobacteria) are better adapted to water stress than other types of bacteria and thus play a relatively larger role in soils. For a given type of soil there is an optimum water content for microbial activity; when a large fraction of pores are water filled anaerobic conditions are more widespread, thus inhibiting aerobic processes. Water contents affect, for example, soil nitrogen cycling; low water contents favor ammonia oxidation (nitritification, an aerobic process), whereas high water content favors denitrification (an anaerobic process leading to loss of reactive N from the soil).

The highest microbial activities are found within the root zone (rhizosphere). Microbes (in part directly attached to plant roots) exploit dissolved organic mate-
rial which is excreted from the roots. They also form the basis for protozoan grazing. In addition to direct

4. Extreme Environments
The microbiology of extreme environments (here in-
cluding hyperthermal and subsurface habitats and brines) has recently drawn considerable interest. Stud-
ies have expanded our knowledge of microbial diversity (especially of thermophilic archaeabacteria) and led to the discovery of novel microbial communities; the po-
tential for isolating microbes producing industrially useful enzymes has also spurred interest in extreme en-
vvironments.

Microbial mats are communities that develop in envi-
ronments with no or little animal grazing and distur-
bance. These are complex biota that include a great variety of interacting physiological types of prokaryotes and in some cases also include eukaryotic microbes. Cyanobacterial mats develop in illuminated thermal springs and brines (salterns); when undisturbed, they develop 1-m-thick laminated deposits (stromatolitic mats) over centuries, although almost all biological ac-
tivity is concentrated in the upper few millimeters. They are a modern analog to Precambrian stromatolites, which seem to have represented the dominating shal-
low-water biota for approximately 2.8 billion years of Earth’s history; they almost disappeared from the geo-
logical record in the past 600 to 700 million years, when metazoans made their appearance. Cold or hot seeps of sulfidic water support mats of chemosynthetic sulfur bacteria. These biota—well-known from deep-sea hydrothermal vents but also occurring else-
where—are interesting in that their energy support (sulfide, methane, and reduced Fe) derives from geo-
thermal processes rather than from solar energy via photosynthesis (however, the oxygen, which is neces-
sary for exploitation of these energy sources, derives from oxygenic photosynthesis).

Subsurface microbial biota are a recent discovery. Thus, bacterial life has now been found hundreds of meters underground in shales, limestone, and other rocks, and groundwater ages and other evidence indi-
cate that there has been no contact with the surface for at least 107 and up to 109 years. The microbial activity seems mainly to be hydrogen based (methanogenesis, acetogenesis, and sulfate reduction), but organotrophs also occur. Oil reservoirs in deep sediments have also been shown to harbor biota of hydrocarbon-degrading thermophilic sulfate reducers.
D. Symbiotic Bacteria

There are numerous examples of bacteria living in or on other organisms; in many cases the adaptive significance of these associations is not understood. Some cases of symbiotic between prokaryotes and eukaryotes, however, have been studied in detail; they are of scientific interest and sometimes even attract substantial economic interests.

Symbiotic polymer degradation is one such example. With few exceptions, animals are not capable of hydrolyzing and utilizing structural plant polymers directly. Animals do not produce the necessary enzymes (cellulase, xylanase, etc.). Also, plant tissue generally contains insufficient amounts of N and P to sustain growth and various secondary plant compounds are toxic to animals. Some herbivores therefore eat copious amounts of plant tissue with a low digestion efficiency. They are typically confined to a few (taxonomically related) food plants because each grazer has only developed detoxification mechanisms against specific plant toxins; this applies to many insects. Many animals, however, have solved the problems of herbivory by maintaining consortia of anaerobic microbes in their gut system. Such systems have evolved independently in several groups of mammals and in a few groups of birds (ostriches, ptarmigans, and hoatzins), reptiles, and fish; they are also found in termites and cockroaches, in some echinids, and in shipworms.

The best known example is the rumen of cows and sheep. Ruminants exemplify pregastric fermentation. The rumen, which constitutes 10–15% of the volume of the animal, is anatomically a part of the esophagus. Its anaerobic content is a mixture of rumen fluid, ingested plant fragments, and microbes (approximately 10^11 bacteria and 10^9 protists ml^-1). Ruminant saliva does not contain hydrolytic enzymes, but it is strongly buffered. Urea (the principal nitrogenous excretion product of mammals) is also excreted into the saliva, supplying the microbial community with N; it is a mechanism for preserving this essential element for the ruminant through internal recycling. Within the rumen a consortium of different bacteria (and some eukaryotic microorganisms) ferment carbohydrates (including cellulose and xylan) principally into acetate, butyrate, propionate, and CO₂ + H₂. The fatty acids are absorbed in the intestinal tract and they constitute the energy and carbon source of the ruminant. Protein is exclusively provided by the digestion of microbial cells in the true stomach; thus, a cow does not profit from eating proteins because these are microbially hydrolyzed, deamidated and fermented in the rumen prior to the stomach.

In the rumen, methanogens convert H₂ into CH₄; the methane represents a (necessary) loss to the cow (approximately 15% of the food intake) and it rids itself of the gas through belching. Ruminants are thus totally dependent on their microbial symbionts; ruminants that have artificially been deprived of their rumen biota cannot survive on a normal diet. In addition to ruminants, pregastric fermentation is found, for example, in kangaroos, sloths, and colubrine monkeys.

Postgastric fermentation is more widespread (e.g., horses, rhinos, elephants, lemur, koalas, lagomorphs, and several others; even humans probably profit slightly from their intestinal microflora in this way). Animals with postgastric fermentation possess one or more caeca containing a consortium of fermenting bacteria. Ingested food is first subject to acid digestion in the stomach, with subsequent absorption of monomers, whereas undigestable plant polymers are eventually fermented in the caecum. In most termites and cockroaches, the principal cellulose decomposers (harbored in the hind-gut) are protozoa (various types of specialized flagellates), although bacteria also play a substantial role in these microbiota.

Nitrogen fixation (the reduction of N₂ to NH₄) is a very energy-requiring, anaerobic process which is known only from bacteria. It is widespread especially among free-living anaerobes and microaerobes, and some aerobes have special adaptations to protect the nitrogenase complex from O₂ (e.g., some cyanobacteria and Azotobacter). Nevertheless, it is estimated that in terrestrial habitats more than 90% of N₂ fixation is symbiotic.

Symbiotic nitrogen fixation is of immense economic importance and since antiquity it has been known that legumes improve soil fertility. The great majority of plants belonging to the Leguminosae form root nodules that harbor the N₂-fixing bacteria (Rhizobium). These bacteria normally live in soils; they especially thrive in the rhizosphere of legumes and often adhere to root hairs. Appropriate strains enter root hairs and induce the formation of root nodules. The bacteria transform the nondividing "bacteroids" that fix N₂ on the basis of carbohydrates provided by the plant. A hemoglobin (leghemoglobin) secures a sufficiently low O₂ tension while simultaneously allowing for aerobic metabolism, and the plant assimilates the formed ammonia.

Another type of symbiotic N₂ fixation is found in many unrelated trees (e.g., Alnus, Myrica, and Hippophae); it is due to the bacterium Frankia (an actinomycete), which also forms root nodules. The adaptive significance of different examples of symbiotic cyanobacteria in plants seems also to be that of N₂ fixation.
with H2 production; the symbionts in turn are H2 scavengers (mainly certain ciliates) have a fermentative metabolism. Some anaerobic protozoa maintain symbiotic methanogenic or sulfate-reducing bacteria. The protozoa are gutless and entirely dependent on their symbionts for food; the hosts in turn provide the bacteria with necessary oxygen and sulfide via their circulatory system. Other bivalves maintain sulfur bacteria in their gills, but they are simultaneously capable of filter feeding. Some anaerobic protozoa maintain symbiotic methanogenic or sulfate-reducing bacteria. The protozoa (mainly certain ciliates) have a fermentative metabolism with H2 production; the symbionts in turn are H2 scavengers, thus maintaining a low intracellular H2 tension which again enhances the energy efficiency of the host metabolism.

III. UNICELLULAR EUKARYOTES (PROSTIS)

Protists are known to include many unrelated eukaryotic groups. Eukaryotic cells differ from prokaryotes in possessing a cytoskeleton and membrane-covered organelles, among which mitochondria and chloroplasts are recognized as being descendants of endosymbiotic aerobic bacteria belonging to the as a group of proteobacteria and to cyanobacteria, respectively. Protists have a much greater size range than bacteria; the smallest free-living species measure approximately 3 μm.

Phagocytosis is probably a primary property of eukaryotic cells and most extant species depend on particulate food (mainly bacteria or other protists). Many groups have acquired chloroplasts; this has apparently happened independently in different taxa and examples of “chloroplasts” which represent intermediate stages between an endosymbiont and an organelle are known. Many species have secondarily lost chloroplasts and thus the ability of photosynthesis. Within many groups of phototrophs the ability of phagocytosis has been retained (mixotrophs, found for example among chrysomonads and dinoflagellates); in other groups (e.g., diatoms and green algae) phagocytosis has been irreversibly lost. A few species (some ciliates and foraminifera) are capable of retaining chloroplasts from their prey cells, the chloroplasts remain functional for some days and this is exploited by the “host.”

Many heterotrophic protists harbor endosymbiotic phototrophs. Very few protists subsist on dissolved organic matter; in most habitats they would be inferior competitors to bacteria simply due to their larger size. The majority of protists are aerobes; a few specialized protists, however, are obligate anaerobes depending on a fermentative metabolism. Tolerance to extreme conditions is limited relative to that of some bacteria, but otherwise protists are omnipresent; primarily they are the principal consumers of bacteria in all types of environments and phototroph protists are largely responsible for primary production in aquatic habitats.

A. Phototrophs

Macrophytes (macroalgae and vascular plants) are responsible for much of the primary production along shallow coastal waters and lakes. In offshore waters and in the oceans, unicellular phototrophs are solely responsible for light-driven CO2 reduction and they thus largely constitute the basis for aquatic life. Characteristic annual succession patterns (with respect to species composition) occur in temperate seas and lakes. During winter, mineral nutrients are brought to the photic zone from deeper waters. In early spring, when light conditions allow, a bloom of diatoms develops; the bloom is brought to an end by zooplankton grazing and by nutrient limitations. During summer, phytoplankton is dominated by small forms (chrysomonads and prymnesids) which perform better at low nutrient concentrations; autumn is often characterized by a bloom of large dinoflagellates. In tropical waters which are thermally stratified throughout the year, competition for mineral nutrients is intense and small species prevail. Upwelling areas, such as off the coast of South America, support production of large phytoplankton species and thus represent an exception to this generalization.

Phytoplankton biomass is primarily limited by mineral nutrients—in most marine areas by inorganic N; although P limitation is also known and may be the rule in limnic systems. Diatoms may be limited by available Si and it has been inferred that Fe limits primary production in the Southern Ocean. Increased nutrient additions, especially from fertilizers, via river runoff...
and groundwater seepage have caused increased phytoplankton production and biomass in offshore waters, bays, and estuaries in addition to lakes in many parts of the world. Such eutrophication may be detrimental to rooted vegetation (due to competition for light) and in some cases may result in widespread anoxia or hypoxia of bottom waters due to an increased input of organic material.

Large phytoplankton cells are consumed by various zooplankters (copepods, etc.) which again serve as food for fish. Small phytoplankton are a large extent consumed by protozoa which then enter planktonic food chains. Many bentonic filter feeders (e.g., mussels) also depend mainly on phytoplankton cells.

In shallow aquatic sediments phototrophic protists are important primary producers even though the photic zone extends only 2 or 3 mm beneath the sediment surface. In these biota, diatoms, euglenoids, and dinoflagellates are especially important (together with unicellular and filamentous cyanobacteria).

B. Phagotrophic Protists
Phagotrophic protists seem to be the principal consumers of bacteria in almost all habitats. In plankton, it is particularly the small (4–10 μm), heterotrophic flagellates that play a role as bacterial consumers, thus closing the microbial loop (Fig. 3). They occur at densities of approximately 10^3 cells ml^-1 in most natural waters. Other groups of protists, especially ciliates, mainly consume larger prey such as heterotrophic flagellates and small phytoplanktons, whereas heterotrophic or mixotrophic dinoflagellates feed on prey which often exceeds their own size. In oceanic plankton, the large scapharcaea, radiolaria, and the planktonic foraminiferan Globigerina are mainly phagotrophs feeding on large prey; most of them also harbor endosymbiotic phototrophs. The three (unrelated) groups all produce skeletons (made of strontium sulfate, silica, and calcium carbonate, respectively) and the sedimented skeletal remains of the two latter groups characterize many oceanic sediments.

Sediments, especially from shallow waters, harbor high densities of a wide variety of phagotrophic protozoa, including amoebae, flagellates, and ciliates, filling a variety of niches especially with respect to food requirements and oxygen tension. Foraminifera are characteristic of marine sediments; their niches are in part taken over by testate amoebae and by heliozoans in lake sediments. Deeper marine sediments harbor large foraminifera and in deep-sea sediments the peculiar xenophyophoreans grow to a size of several centimeters and possess a skeleton made of barite.

C. Symbiotic Protists in Animals
Like prokaryotes, many protists occur as symbions in animals; probably all animal species (including humans) harbor several protozoan symbions. Symbiotic polymer degradation by flagellates in termites has already been mentioned. The most important type of symbiosis involving protists is that between animals and intracellular phototrophs. Many aquatic invertebrates harbor such symbions. In fresh waters the symbions are usually the green alga Chlorella (e.g., in freshwater sponges Chloryheda and in the ciliate Paramecium bursaria). Most marine cases are based on dinoflagellates (Symbiodinium), but other groups (e.g., diatoms, chlorophytes, and prymnesids) are also represented. In most cases the host combines the nutrition derived from the symbiont (usually in the form of carbohydrates) with particulate food; in some cases, it has been shown that the hosts can subsist entirely on the basis of the symbions, and in a few cases the ability of phagotrophy has been lost. The most important marine example is that of phototrophic symbions in reef-building corals; the symbions are not only responsible for a significant share of primary production of coral reefs but also facilitate carbonate deposition of the host during active photosynthesis. A similar situation applies to giant tropical shallow-water foraminifers; like corals, they are responsible for the formation of limestone deposits (the Cheops Pyramid is built from limestone consisting of the calcareous remains of the Eocene foraminiferan Nummulites). Other marine invertebrates harboring phototrophic symbions include the giant clam Tridacna and various coelenterates.

IV. ROLES OF MICROBES OVER GEOLOGICAL TIME AND IN THE CONTEMPORARY BIOSPHERE
This section briefly explores the role of microbes in the evolution of biogeochemical element cycling on a global
MICROORGANISMS (MICROBES); ROLE OF

scale. Earth (and the solar system) arose approximately \(4.6 \times 10^9\) years ago; the earliest unambiguous sign of life dates to approximately \(3.5 \times 10^9\) years ago (mid-Archean). It is generally accepted that life arose on Earth, perhaps \(4 \times 10^9\) years ago, under anoxic (and presumably chemically reducing) conditions, and that the atmosphere eventually became oxic as a result of oxygenic (cyanobacterial) photosynthesis. There is convincing evidence that atmospheric \(O_2\) slowly increased from very low levels during the Precambrian. The earliest evidence of life (from Warrawoona in Australia and the Fig Tree Formation from southern Africa) is in the form of stromatolites, which are fossil remains of microbial mats. Some fossils are sufficiently well preserved to reveal organisms that were very similar to present-day cyanobacteria. Stromatolites (sometimes with well-preserved fossil bacteria) are known from throughout the remaining Precambrian. Evidence of an increasing \(O_2\) tension of the atmosphere derives from banded iron formations—laminated (\(\geq 3\) to \(2 \times 10^9\) years old) deposits containing partly oxidized \(Fe\). The general interpretation is that as \(O_2\) (resulting from photosynthesis) appeared in the atmosphere, dissolved reduced iron, in the otherwise anoxic oceans, was oxidized at the sea surface to become insoluble oxidized \(Fe\) that again gave rise to the deposits. Atmospheric \(O_2\) levels thus initially remained low due to oxidation of reduced minerals on the surface of Earth (and presumably due to the early origin of biological \(O_2\) respiration). Later, the \(O_2\) level increased, reflecting the accumulation of fossil organic material in sedimentary rocks. There is evidence that the \(O_2\) level had increased to \(\geq 1\%\) of the current atmospheric level approximately \(2 \times 10^9\) years ago (early Proterozoic). When metazoa arose \(600\) to \(700\) million years ago, atmospheric \(O_2\) must have reached at least \(10\%\) of the current level (according to requirements of extant invertebrates), but it could also have been higher.

The conclusion is that oxygenic phototrophs, quite similar to modern cyanobacteria, must have evolved \(3.5 \times 10^9\) years ago. It seems likely that all basic types of bacterial metabolism had also evolved by then, although this is speculative.

The time for the origin of eukaryotic cells is unknown, although molecular data (Fig. 1) suggests “deep roots”; there is (molecular) evidence suggesting that mitochondria (and thus “modern” eukaryotic microbes) arose approximately \(2 \times 10^9\) years ago (or slightly earlier than fossils that have been interpreted as remains of protists). In all circumstances, when metazoa and the first macroalgae arose \(600\) to \(700\) million years ago, almost all major biochemistry and metabolic pathways had evolved in microorganisms and with them the basic biogeochemical cycles as we know them today. Colonization of land approximately \(100\) million years later and the evolution of vascular plants and fungi, however, must have had a profound effect on the biosphere and on mineral cycling.

There is not enough space here to describe the major biogeochemical element cycles (notably the \(C\), \(N\), and \(S\) cycles) in detail. A view of the nitrogen cycle (Fig. 4), however, further illustrates the immense importance of microbes in the biosphere. Atmospheric \(N_2\) is fixed by bacteria, but reactive \(N\) is also added to the biota as \(N\) oxides formed during electric discharges in the atmosphere and by industrial \(N_2\) fixation. In cells, \(N\) occurs in a reduced form in organic matter and it is released as \(NH_3\) when organic matter becomes mineralized. Through nitrification (a bacterial process) it is oxidized to \(NO\) which can be utilized (through assimilative reduction) by plants. The only pathway back to \(N_2\), however, is denitrification, an anaerobic bacterial process that is necessary for the completion of the global \(N\) cycle and ultimately regulates the amount of reactive \(N\) available to the biota.

The important mechanisms that control particular microbial processes are largely understood. However, the complexities of the cycles and the many positive and negative feedbacks render it difficult, or perhaps impossible, to make predictions of the effect of environmental changes on an ecosystem or on a global scale. An example is atmospheric \(CH_4\), which is a greenhouse gas. The basic sinks (abiotic atmospheric oxidation and microbial oxidation in soils and other biota) and sources
ever, the effect on CH4 levels as a function of climate change is in practice impossible to predict. Both CH4 reduced Fe of pyrite (FeS2), which is omnipresent in 

lus ferrooxidans philic, chemoautotrophic bacteria (including underlining mechanism is that a consortium of acido-

recovered from the leachate by chemical methods. The 

crushed copper ore leads to acid conditions and dissolu-
ing, which was also in use long before the underlying 

mechanism was understood. Circulating water through 

crushed copper ore leads to acid conditions and dissolu-
tion of the ore; metallic Cu can then subsequently be 

recovered from the leachate by chemical methods. The 

underlying mechanism is that a consortium of acido-

phile, chemoautotrophic bacteria (including Thobacil-
lus ferrooxidans) oxidize both the reduced S and the 

reduced Fe of pyrite (FeS2), which is omnipresent in 

many ores. The resulting sulfuric acid in turn dissolves 

the ore.

Biological sewage treatment serves primarily to min-
eralize organic material. Various types of sewage treat-
ment are in use, depending, among other factors, on 
the scale of the plant; most systems involve aerobic and 
an aerobic microbial processes. An important aspect is 
the flocculation of bacteria, a process which is enhanced 
by the presence of protozoa. Removal of nitrate by 

microbial denitrification is another important function 

of biological sewage treatment, whereas phosphate is 
primarily removed by chemical precipitation. Mainly 
in smaller plants, anaerobic mineralization can be ex-
ploded to produce methane, which can be collected 
and subsequently used for heating.

In recent times, mass production of certain species of 
bacteria for the production of enzymes and antibiot-
ics has played an important industrial role. Recently, 
genetically engineered bacteria that express human pro-
tein genes (e.g., insulin and other hormones) have been 
used in the pharmaceutical industry.

Microbial diseases, of which there are many, repre-
sent the most direct encounter between humans and 
microbes. Through recorded history such diseases have 
played an important role for human populations; most 
dramatically illustrated, perhaps, by the recurrent 
plague epidemics in Europe from medieval times to 
approximately 1700; however, many other bacterial dis-

cases, such as cholera, tuberculosis, leprosy, and ty-
poid fever, and protozoal diseases such malaria were 
also important. In North America, Europe, and in some 
other parts of the world serious bacterial and protozoal 
diseases have, especially after World War II, largely 
been brought under control due to the combined effects 
of hygienic measures, vector control (mosquitoes and 

rats), immunization programs, and antibiotics and 
other forms of chemotherapy. However, globally, tuber-
culosus and malaria remain among the most frequent 
causes of death. Many bacterial and protozoal diseases 
of livestock also remain economically significant. Evolv-
ing resistance to antibiotics and other types of chemo-
therapy in agents of disease in man and animals may 
represent an increasing problem and indicate that our 
interactions with pathogenic microorganisms is not a 
closed chapter in human history.

See Also the Following Articles

ARCHAEA, ORIGIN OF • BACTERIAL BIODIVERSITY • 
ECOSYSTEM SERVICES, CONCEPT OF • EUKARYOTES, ORIGIN OF • MARINE SEDIMENTS • MICROBIAL BIODIVERSITY • 
NITROGEN AND NITROGEN CYCLE • PROTOZOA • 
PSYCHROPHILES • THERMOPHILES, ORIGIN OF
Bibliography

MIGRATION

Mace A. Hack and Daniel I. Rubenstein
Princeton University

I. Introduction
II. What Is Migration?
III. Migratory Patterns: A Taxonomic Survey
IV. Attributes of Migrants Affecting Susceptibility to Human Disturbance
V. Ecological Consequences of Human Disturbance on Migrants
VI. Evolutionary Consequences of Human Impacts on Migrants

GLOSSARY

circannual rhythms Endogenous, or internal, rhythmic cycles of one year in duration that govern the onset and cessation of migratory behaviors.
compass orientation Navigation in a particular direction without reference to landmarks or sites of origin or destination. Migrants are known to use compass information from magnetic fields, chemical gradients, and visual features such as the stars, sun, and planes of light polarization.
diadromy Migrations that take individuals between fresh and salt-water habitats, a common phenomenon for many migratory fish species.
partial migration The case where intrapopulation variation in migratory behavior leads some individuals to migrate while others within the same population may only migrate locally or remain sedentary.
zugenruhe Restlessness exhibited by some migratory species, especially birds, if not allowed to migrate during their usual migratory period. It reflects an underlying physiological transition to a migratory state.

MIGRATION DESCRIBES the movement of individuals between spatially separate ecological communities, typically on a seasonal or annual schedule. Several characteristics of migrations distinguish them from other forms of animal ranging behavior, including more persistent movements, of greater duration, that follow a more direct path with fewer turnings. Furthermore, migrants do not respond to resources along their path, but show a heightened response to the same resources near the migratory journey’s end. This latter feature distinguishes migrations from typical foraging and dispersal movements. Behavioral specializations may include specific activity patterns particular to departure and arrival, and unique patterns of energy allocation to support long-distance movements. The ecological consequences of migration are that they take a species from one community of organisms to another and they partition life histories so that specific phases or events occur in different ecological communities.

I. INTRODUCTION

Migrations capture the human imagination like few other animal behaviors. The single-minded struggle of
the salmon fighting its way upstream, the barely visible formation of geese piercing the sky overhead, and the thundering line of wildebeest snaking across the open savanna—all speak of ancient rhythms that drive life on our planet. The often great distances moved and the large numbers of individuals involved make animal migrations a conspicuous and essential aspect of many regions’ biodiversities. Animals migrate for many reasons but in general do so to avoid temporarily unfavorable conditions or to locate particularly favorable areas that can meet specific biological needs, such as reproduction. However, it is somewhat ironic that migrations might actually increase a species’ risk rather than reduce it. Protecting migratory species requires preservation not only of their final destinations but their migratory routes and stopover points as well. In effect, migration inextricably links the fates of biotas across the length and breadth of the globe. No natural phenomenon makes the point better that biodiversity on local and global scales shares the same actors and the same processes that drive them. In this chapter we explore the nature, scope, and patterning of animal migrations as a prelude to discussing how human-caused changes in the environment are likely to impact migratory species. We not only address direct effects on migratory species behavior, ecology, population dynamics, and evolutionary potential, but we also consider indirect impacts on the ecological communities and regional biodiversities that disruption of migrations can produce.

II. WHAT IS MIGRATION?

Definitions of ‘migration’ abound in the scientific literature, many of which capture the nature of the behavior for specific taxa but fail to generalize across taxa. Commonly cited elements include long distances traveled and movements from one place to another, then back again. But what is ‘long’? Dark-eyed juncos (Junco hyemalis) and Blackburnian warblers (Dendroica fusca) are both small passerine birds that migrate with the seasons; both breed at higher latitudes in northeastern Canada (among other sites) yet juncos may migrate only a few hundred kilometers to their overwintering range while Blackburnian warblers fly several thousand kilometers to overwinter in the Andean forests of Ecuador. The lengths of these respective trips differ by an order of magnitude, yet in examining the ecology and behavior of each species, we might well consider both to constitute migrations. For example, both involve moving between distinct ecological communities on a consistently timed, seasonal basis. Similarly, ‘typical’ migrants, like most birds, travel the same circuit each year. Others, however, may only complete a single circuit in their lifetime (e.g., salmon) or only part of a circuit before they die, such as insects in which successive generations continue the journey their predecessors began. Should we exclude such species from the ‘migratory’ category despite other aspects of behavior, physiology, and life history held in common with typical migrants?

From a biodiversity perspective, migration drives a species’ life history and pattern of resource use, and it ties together ecological communities in different regions. Thus, defining migration in ecological terms of a species’ use of space over time is imperative to understanding its functional impacts on biodiversity. However, one also needs to ask, does biodiversity refer to just the specific organisms found within a region, or does it encompass the specialized behaviors such organisms display? Ecological consequences aside, are sedentary populations of Canada geese interchangeable with migratory ones when summing up a region’s biodiversity? If migratory and nonmigratory races or subpopulations—whether genetically or phenotypically different—each represent unique components of biodiversity, then we must also define ‘migration’ in terms of the behavioral mechanisms distinguishing them.

The combined ecological and behavioral dimensions of migration have plagued attempts to come up with a single definition suitable from both perspectives. Ecological definitions have included “the act of moving from one spatial unit to another” by Baker (1978) and “the persisting change that is left over when all other, minor excursions are removed” by Taylor and Taylor. Both capture the essential idea of migrations extending the ecological space used by individuals, although ambiguity over what constitutes a spatial unit or minor excursion renders each too imprecise to be of much use in distinguishing migratory from nonmigratory species.

Dingle (1996) has taken Kennedy’s lead to propose a definition of migration that emphasizes the differences between migrations and other forms of animal movement. He uses clear behavioral terms to distinguish migratory and nonmigratory taxa, while rendering the differences themselves as their own reference for differentiating movements with different ecological consequences. Generally put, nonmigratory (ranging) movements are driven by the immediate need to acquire or safeguard resources, and they cease when suitable resources are encountered or are adequately defended. The length and timing of foraging and territorial defense movements will vary daily, or on even shorter time scales, depending on the frequency of encountering resources or potential threats. Commuting daily between refugia and feeding
locations, such as the diel movements of zooplankton and other aquatic animals, is similarly directed by resources and is responsive to their location and abundance. Even dispersal to establish a new home range or to leave a natal group usually ceases as soon as a suitable new home range is found. In contrast, once they set out, migrants will ignore many of the resources they encounter, or use them briefly to refuel, and only become responsive again to resource abundance and quality after all or a critical part of their journey has been completed. Migratory movements thus entail specific changes in behavior and physiology that distinguish them from nonmigratory movements, even those that are circular in nature. Defining migrations relative to a species’ other forms of movement results in a more generally applicable definition and one that encompasses the diversity of the migratory behaviors animals display. Specifically, five mechanistic attributes distinguish migratory behavior: movements are (a) more persistent and of greater duration than ranging movements and (b) follow a more direct path with fewer turnings. There is (c) an initial suppression of responses to resource-derived stimuli, but often a heightened response to such stimuli near the migratory journey’s end. Migrants may also have (d) specific activity patterns particular to departure and arrival and (e) unique patterns of energy allocation to support long-distance movements. Of course, not all migrants will show all five characteristics, but as a group these traits circumscribe the suite of distinct and specialized behaviors entailed in migration. To these behavioral attributes we add two functional hallmarks of migrations: (f) they take a species from one community of organisms to another, and (g) they partition life histories so that specific phases or events occur in different ecological communities. Two common uses of the term “migration” will not be considered in this chapter: (a) migration in the paleontological sense of species shifting their historical distributions with climate change or geological events and (b) migration as geneticists use the term to refer to gene flow between populations.

III. MIGRATORY PATTERNS: A TAXONOMIC SURVEY

The diversity of migratory behavior in animals overwhelms attempts to neatly summarize its character and function. Generally, animals migrate to escape unfavorable conditions or to exploit favorable ones, yet defining “favorable” or “unfavorable” is often specific to the taxon examined and the life-history function at hand. Polar and cold temperate habitats tend to have more migrant species than tropical ones because they vary so strongly in productivity and habitability, although this tendency differs widely among the major taxonomic groups. Migration distance and duration similarly vary to a great degree, even within groups of species that migrate for the same reason. Differences in physiology account in part for this variability, since larger body sizes can store relatively more energy to fuel longer trips and lower the weight specific cost of transport while certain forms of locomotion, such as flying or swimming, are more efficient than others. In addition to physiology, physical forces such as winds and currents act in concert with habitat topography to further shape the migratory route and schedule. Biogeographic history may even play a role in these features of migration since routes may be hard-wired genetically and slow to adapt, or learned and dependent on knowledge within lineages. Indeed, migrations present a truly fascinating mix of evolutionary puzzles in behavior, ecology, physiology, and biogeography (see Dingle, 1996, for an excellent introduction to these). The broad surveys that follow provide an overview of migration prevalence, function, and character for the major taxa with specific emphasis on the biodiversity of migrants per taxonomic group and region.

A. Birds

Birds epitomize the act of migration for many people. Whether one lives in the temperate zone or tropics, New World or Old, Southern or Northern Hemisphere, migratory species constitute a significant and conspicuous fraction of the avifauna. At the most northern and southern latitudes, close to 100% of bird species migrate out of the region for part of the year. For example, 135 species breed in the arctic zone, yet all migrate south to spend most of their year elsewhere. In more temperate regions, a majority of species, and nearly all insectivores, migrate to more tropical latitudes after breeding. Approximately 200 species migrate from North America to the West Indies, Central, and South America each year, while many more make shorter distance migrations into the southern United States and Mexico. In Europe, 177 species, or 40% of the region’s avifauna, migrate from temperate breeding grounds to overwinter in Africa; 104 species from western, central, and eastern Asia join them there. Moreau has estimated that 3 billion individuals make this migratory journey south to Africa alone; the comparable number for the entire
globe is surely several times this. Seasonal fluctuations in temperature ultimately drive food production cycles at higher latitudes, setting the schedule for migratory movements, but in tropical climates distinct wet and dry seasons may function in an identical manner. Substantial numbers of birds follow regular migratory routes that track resources such as fruiting or flowering trees, seeding grasses, or invertebrate flushes brought about by rains. Indeed, this form of short-distance movement may be the most common type of migration in birds. For example, out of 1450 breeding species in Africa, 532 species have been classified as intra-African migrants.

Bird migration is clearly a ubiquitous feature of our planet’s biodiversity, yet migrant species are not distributed equally around the globe. In particular, birds that migrate in excess of 1000 km—moderate and long-distance migrants—are more prevalent as breeders in the northern hemisphere. This is largely due to the much greater land area at higher latitudes in the Northern hemisphere than in the Southern Hemisphere. For example, only 20 species among the hundreds that breed in the temperate regions of southern Africa migrate as far north as the equator, while a much larger number and proportion of Palearctic species overwinter in the same equatorial region (noted earlier). Similarly, only 8% of the nearly 600 species breeding in Australia and Tasmania migrate in and out of the temperate zone. In the New World, 31 species of shorebird (e.g., plovers, sandpipers, curlews) breed at higher latitudes in the Northern Hemisphere and migrate thousands of kilometers to spend their nonbreeding seasons in the Southern Hemisphere, yet not a single shorebird species that breeds in the south even migrates as far north as the equator during its nonbreeding season. Although the actual bases for these fascinating patterns remain relatively unexplored, they suggest a greater sensitivity of northern avifaunas to factors that threaten moderate and long-distance migrants off their breeding grounds. Long-distance migrants accomplish truly stunning feats. Blackpoll warblers (Dendroica striata), weighing only 10 to 20 gm, embark from Cape Cod on a nonstop, 86 hr, 3500 km flight 2000 m above the waves of the Atlantic to the northeastern coast of South America. Tundra-breeding shorebirds such as the Pacific golden plover (Pluvialis fulva) make nonstop 3000 to 7000 km journeys to wintering grounds in the South Pacific. Arctic terns (Sterna paradisaea) have the longest migration of any bird, and perhaps any animal, traveling an annual circuit that can exceed 40,000 km. They breed in the boreal summer along the northern edges of the Old and New World continents, then migrate along the western edges of North America, South America, Europe, and Africa to feed in the rich waters off Antarctica’s pack ice during the austral summer. Close to 8 months of each year are spent in transit between these two endpoints.

The Arctic tern’s migratory journey seems extraordinary but it typifies why many birds migrate. In doing so, they exploit highly seasonal flushes of food resources, especially to meet the increased demands of breeding. High variability in food supply at higher latitudes, especially in insects and other invertebrates, makes it difficult for species to reside permanently in the habitat. However, the very predictable nature of these fluctuations allows mobile species to rely on them for part of their annual cycle. The absence of resident competitors may make these ephemeral resources even more abundant and accessible to migrants. Thus, the typical migratory movement for birds involves making an annual round trip between seasonally resource-rich breeding sites and nonbreeding areas where resource abundance may be lower but less variable over time. Seasonal resource variability explains why many birds migrate, but what determines how far they go and where their final destinations lie? Proximal determinants of migration distance and route include weather patterns, history, the distribution of resources along the way, and the character of the landscape. The north-south orientation of mountain ranges in North, Central, and South America tend to funnel migrants along corridors also running north-south, whereas Eurasia mountain ranges run east-west, forcing many Palearctic migrants to make large westerly movements before they are able to fly south. The importance of trade winds or winds generated by weather fronts to migratory journeys can be seen in the often very different routes followed when moving north versus south (e.g., Arctic tern study by Alerstam). The migratory path of the white-rumped sandpiper (Calidris fuscicollis) extends the length of the Western Hemisphere and appears carefully choreographed to coincide with seasonal pulses of invertebrate prey along the route. History too seems to shape migratory paths and destinations as suggested by the very indirect route (via the eastern Mediterranean) taken by red-backed shrikes (Lanius collaris) migrating from the Iberian Peninsula to their nonbreeding range in Central Africa. Habitat suitability is a more ultimate determinant of migration distance and destination. Species with very specific habitat requirements may have limited options for suitable nonbreeding areas. For example, the upland sandpiper (Bartramia longicauda) breeds in North American grasslands and must travel 110,000 km to over-
winter in the pampas of Argentina—the only similar Southern Hemisphere grassland habitat in the New World. Many tundra-breeding shorebirds migrate to the southern coastlines of South America, Africa, Southeast Asia, and Australia where extensive intertidal mudflats and rocky beaches provide abundant invertebrate prey. Most Neotropical migrants—the warblers, flycatchers, vireos, swifts, hummingbirds, swallows, tanagers, orioles, and raptors—migrate more moderate distances to forest and scrub habitats in Central America, the West Indies, and northern South America. In addition to the general structure of the habitat, competition with resident forest species and other migrants certainly influences where migrants settle during their nonbreeding seasons.

Often the question of how far migrants move depends on which population one examines. Many migrant birds conduct “partial” migrations where some individuals migrate while others remain as year-round residents on the breeding grounds or move only short distances. Different age and sex classes may pursue different migratory strategies, as in the dark-eyed junco of the eastern United States. Alternatively, all individuals may migrate but different populations travel different distances. A common pattern in this case is for higher latitude breeding populations to migrate the longest distances. A common pattern in this case is for higher latitude breeding populations to migrate the longest distances. “Leap-frogging” beyond the migratory movements of lower latitude populations (e.g., fox sparrow, Passerella iliaca). Some sanderling (Calidris alba) populations migrate only a short distance from their tundra breeding grounds to overwinter in the northwestern United States while others travel 7500 km to nonbreeding areas in Chile. Apparently, the much greater energetic and exposure costs of migrating to South America are offset by much richer food resources and a more hospitable climate, so that the payoffs for each strategy are equivalent, and both short- and long-distance migrants persist.

The presence of both migratory and nonmigratory strategies in the same species underlines the opportunistic and flexible nature of bird migration. As environmental changes occur or as a species expands its range, migratory behavior can often adapt to fit the new circumstances. The European starling (Sturnus vulgaris) is a widespread, permanent resident of Britain today, yet in the 18th century starlings regularly migrated out of Britain to overwinter in warmer regions of Europe. Long-term climate change has presumably made Britain a more hospitable place for starlings to spend their winters. Conversely, several species resident year-round in Europe have extended their breeding ranges north over the past century. These more northerly populations have developed full migrations to southern Europe for overwintering. Migrants may also establish nonmigratory populations along traditional migratory routes as resources become more abundant or habitats are altered, for example, the barn swallow (Hirundo rustica) in Argentina and the Canada goose (Branta canadensis) in eastern North America. With industrial parks in North America supporting extensive expanses of grassy lawns some Canada geese cannot only forage sufficiently well at temperate latitudes during the winter, but high levels of vegetative production during the spring and summer enable them to stay and breed.

A fascinating form of bird migration not related directly to food abundance is molt migration. It is particularly common among waterfowl that molt all their flight feathers simultaneously and thereby lose the ability to fly. To escape predation, they may migrate to coastal areas, large lakes, or far off-shore until their new feathers have grown in.

B. Mammals

Small (<5 kg), nonvolatile mammals represent a large fraction of our planet’s mammalian biodiversity—rodents alone account for 40% of all mammal species—yet very few species are known to migrate. Rather than leave highly seasonal habitats when resource abundances decline, many small mammals hibernate or reduce activity levels and wait for conditions to improve. Species known to migrate include lemmings in the arctic tundra and a variety of rats, mice, and shrews living on the Kafue River flats of Zambia studied by Sheppe. In both cases, these small mammals move several kilometers to escape flooding conditions as snow melts (lemmings) or rivers flood (Kafue mammals).

Migration occurs more commonly among a second, highly diverse mammal group, the bats (25% of all mammals). The ability to fly and travel large distances more efficiently than via terrestrial forms of locomotion must account in large part for this difference between bats and other small mammals. Fruit-eating bats in West Africa migrate 1500 km annually, following the movement of rains and the consequent pattern of fruit abundance. As in birds, different sex and age classes may migrate different distances; juveniles migrate further, perhaps trading-off reproduction for more abundant food and greater survivorship. Similar migrations tracking fruit, nectar, and insect abundances have also been found in Australia and the New World tropics. Insectivorous bats breeding in the temperate regions of North America and Europe commonly migrate,
showing two distinct patterns of movement. Like many birds in these regions, species such as the Mexican free-tailed bat (Tadarida brasiliensis) and the hoary bat (Lasiurus cinereus) travel a maximum of 800 to 1700 km south to overwinter in warmer sites where their insect food can still be found. These individuals do not hibernate, but other individuals of the same species may make only part of the southward journey to hibernate at sites along the migratory route. The second migratory pattern involves movements of shorter distances (10–200 km), and more variable orientation, from summer ranges to particular winter hibernation roosts. For example, as Griffin discovered, the little brown myotis (Myotis lucifugus) hibernates in a few select caves and mine tunnels in northern New England, yet migrates usually less than 100 km back to summer ranges both north and south of these wintering sites. Seasonality determines the timing of migration in these species but the very specific requirements of hibernation, such as temperature stability, humidity, and protection from predators, drive migration distance and direction. It is not uncommon for 75 to 93% of the population in a large geographical area, or even an entire species, to migrate and hibernate in only a few caves. Both migratory patterns are particularly common among species that roost in trees during the summer, presumably because trees make very poor, exposed winter hibernation sites.

Among terrestrial mammals, migrations are most common, and most spectacular, in the ungulates, or hoofed mammals. In the Serengeti ecosystem of Kenya common, and most spectacular, in the ungulates, or

tions; eastern North American populations of "wood-

to taiga during winter. The more wooded taiga also provides greater protection from harsh winter condi-
tions; eastern North American populations of "wood-
land" caribou do not migrate at all.

Ungulates in temperate regions often migrate altitu-
dinally with the seasons, moving to lower altitudes in
winter to avoid harsh weather, low food abundance, and low food accessibility due to deep snow. Such species include bighorn sheep (Ovis canadensis), elk (Cervus elaphus), mule deer (Odocoileus hemionus), and feral horses (Equus caballus) in the Nearctic and chamois (Rupicapra rupicapra) in the Paleartic. Migratory dis-
tances in these species are usually short. An interesting exception is the mule deer, where each individual or family has a definite summer and winter range yet the migratory route is often not the most direct link be-
tween them. Individuals have been observed to travel 150 km in straight-line distance, crossing six mountain ranges in a winding route to do so. Tradition appears to have a tremendous effect on defining seasonal ranges and migratory routes in this species.

For marine mammals, annual migrations are the rule rather than the exception. The baleen whales, or mysti-
cetes, migrate the longest distances, some moving thou-
sands of kilometers from tropical to polar waters and then back again each year. Northern and Southern Hemisphere populations both migrate in this way, but their opposite schedules prevent overlap in the tropics. Food availability and quality at higher latitudes drive the timing of these movements. During the polar sum-
mer, long day-lengths lead to phytoplankton blooms which, in turn, generate huge abundances of the zoo-
 plankton, such as krill, that baleen whales specialize on when feeding. The lipid content of krill and small fish prey, or energy per mouthful from the whale’s perspective, also rises during the summer months. In-
deed, polar waters provide such an amazingly rich food source that many species forego feeding the rest of the year, consuming enough during a 3-month summer binge to not only maintain themselves for the remaining 9 months but to complete a hemispheric round-trip and breed. Species such as the fin whale (Balaenoptera physalus) birth mate and give birth 11 months later in warm, low latitude waters. Lactation lasts for 6 to 7 months and the calf is weaned on the summer feeding grounds; fat reserves for lactation just after birth may constitute 50 to 75% of a female’s body mass. Lower thermoregulatory costs in warm water presumably allow the baleen-migrate-and-fast strategies of baleen whales to be successful, especially since warm, tropical waters may also provide a thermal environment more conducive to calf growth.

Odontocetes, or toothed whales such as dolphins, pilot whales and beaked whales, may show regular sea-
sonal movements between breeding and nonbreeding

grounds, but long-distance migrations rarely occur. Most toothed whales feed on single food items like fish
and squid and, as a consequence, must feed every day to maintain themselves. Many have large feeding ranges through which they travel 20 to 60 km per day. Short-distance migrations typically occur when food abundance in an area changes seasonally, such as when fish schools move off-shore or concentrate in river mouths. Some Arctic odontocetes, such as the narwhal (Monodon monoceros) and beluga (Delphinapterus leucas), migrate short distances seasonally as the polar pack ice retreats and opens up calving grounds in warmer-water estuaries. When the ice returns, they move back with its advancing edge to deeper, ice-free waters. Only the sperm whale—the largest odontocete—migrates distances comparable to those seen in the baleen whales. Adult males move two to three times the distances of females, reaching richer feeding areas at higher latitudes in order to achieve and maintain their much larger body sizes.

Among pinnipeds, 44% of phocid (true or hair seals) species migrate seasonally while only 14% of otariids (fur seals, sea lions) do so. This difference may arise from the phocid’s greater reliance on seasonally changing pack ice, rather than solid land, as a substrate for giving birth, raising young, and molting. For example, harp (Phoca groenlandica) and hooded seals (Cystophora cristata) whelp on pack ice in March near Newfoundland, among other sites, then migrate north with the retreating pack ice to feed. Annual round-trips may cover 400 km, although there is much variation in migratory distance and route among individuals and populations throughout the North Atlantic. Northern elephant seals hold the current record for the longest migration of any mammal: 18 to 21,200 km annually as they move twice between Californian islands used for breeding (January–February) and molting (July–August) to higher-latitude feeding areas rich in cephalopod prey. Northern fur seals, perhaps the only true migrants among the otariids, migrate in the opposite direction of most phocids, breeding at higher latitudes in the summer months and migrating south to temperate waters to overwinter.

Although clearly not mammals, sea turtles closely resemble pinnipeds in their migratory behavior. Adults migrate to specific island or continental beaches where eggs are laid. Hatchlings venture immediately to the sea where they may spend 30 or more years before maturing and returning to the same beach to complete the cycle.

Tropical marine mammals, such as dugong and manatees, are generally nonmigratory, although the West Indian manatee (Trichechus manatus) population in Florida may move several hundred kilometers north in summer to exploit new feeding areas, retreating back to warmer waters in the winter. Their need for warm water in which to overwinter has in some cases made them dependent on the warm-water discharges of coastal power plants.

### C. Fish and Other Aquatic Species

Fish show a range of migratory patterns, both among different species and different populations of the same species. Migration is relatively common in this group, particularly as a means of linking rich feeding habitats with specialized spawning grounds that provide refuge for eggs and young fish. For diadromous species—the most studied and conspicuous migratory fishes—this journey requires moving from marine to freshwater environments in order to breed (anadromous: e.g., salmonids, shad, sticklebacks, lampreys) or, more rarely, the converse (catadromous: e.g., freshwater eels, southern trout, southern smelt). Anadromous species predominate in cold-temperate and subpolar waters in both hemispheres, while catadromous species more commonly occur in warm-temperate and subtropical waters. The benefits of moving from lower to higher productivity habitats to feed and grow may account for this pattern since marine productivity is higher at high latitudes while freshwater productivity is higher at low latitudes. Anadromous species constitute a higher proportion of total coastal fish diversity in the Pacific than they do in the Atlantic, largely due to the Atlantic’s much higher coastal fish diversity overall.

As with marine mammals (e.g., baleen whales), migratory fishes with separate feeding and spawning areas must feed intensively and store sufficient energy reserves for both migration and spawning. Although many species migrate only a few kilometers between feeding and breeding sites, some fish migrations are truly remarkable in length (also discussed later). Upstream distances in diadromous species include 300 to 400 km in lampreys, 500 km in shad, and well over 1000 km in some salmon and sturgeons. Sockeye salmon migrating upstream expend 70% of their available energy in reaching the spawning grounds and the rest on spawning itself; not surprisingly, both sexes die shortly thereafter. The combined toll of fasting, migrating upstream, and spawning results in semelparity for most diadromous species, but some species, such as the Atlantic salmon (Salmo salar) and northern populations of American shad (Alosa sapidissima) may migrate and spawn more than once. The dramatic physiological transition required to move between freshwater and marine environments may also be a factor in the prevalence of semelparity among diadromous species. Salmonid migrations typify anadromous life cycles,
but they also illustrate the very flexible migratory strategies of fish. Pacific salmon (Oncorhynchus spp.) spawn in streams at cold-temperate and subpolar latitudes on both sides of the Pacific. After hatching, they either (a) travel immediately downstream to the ocean or, (b) before migrating to the ocean, move downstream to lakes where they spend the next 2 to 4 years, or (c) remain in their hatching stream for one or more years. Once in the ocean, some species migrate the breadth of the Pacific to reach feeding areas while others remain closer to their spawning streams; accordingly, the time spent feeding and growing in the marine environment varies among species but is generally 2 or more years. The same species may have several distinct upstream migration strategies for the two sexes. Larger at maturity suggesting very different life history stages the lobsters feed in during warmer months. Migration thus allows spiny lobsters to exploit a resource-rich, yet seasonally stressful, habitat.

Oceanic long-distance migrants include cod (Gadus morhua) and plaice (Pleuronectes platessa) at higher latitudes and several tuna species in temperate and tropical waters. In most cases, these migration circuits are on the order of 1000 to 3000 km in diameter. Littoral migrations also occur commonly among oceanodromous species, but they are generally much shorter. Those species moving to track food resources usually move offshore in winter, while species migrating inshore during winter may do so to spawn in relatively predator-free habitats.

Species living solely in freshwater (i.e., potamodromous) commonly migrate from deep to shallow waters in order to spawn. A large variety of potamodromous fish migrate throughout the world's lakes and rivers, including freshwater rays, sturgeons, suckers, minnows, pikes, sunfishes, darters, perches, and catfishes. Shallower, upstream habitats may exclude certain predators or have water characteristics (e.g., oxygen, temperature, or silt levels) more suitable to the development and growth of eggs and young fish. The South American characin, Prochilodus mariae, has both non-migratory, lake-breeding and migratory, stream-breeding individuals within the same population. The observation that lake breeders expend five times more energy on egg production than stream breeders suggests that stream habitats provide sufficient benefits in offspring survival to compensate for the high energetic costs of migration.

Beyond fish and marine mammals, migration in aquatic animals is less common—or perhaps just understudied. Spiny lobsters in the family Palinuridae demonstrate a very curious migratory behavior, queuing up in long lines of up to 50 individuals that snake their way along the ocean bottom towards deeper, more sheltered habitats. These movements, as much as 30 to 50 km in length over several days, usually occur seasonally in response to a greater incidence of polar storm fronts, which bring colder water temperatures and greater wave disturbance to the shallow water habitats the lobsters feed in during warmer months. Migration thus allows spiny lobsters to exploit a resource-rich, yet seasonally stressful, habitat.
Plankton show distinct migrations in the water column driven by the seasonal availability of nutrients. During months of high productivity, plankton migrate to surface waters; as productivity declines, they return to lower depths and often enter into a state of diapause. Since changes in light work in concert with endogenous rhythms to set the schedule for these movements, the maximum depth reached during the nonproductive season will depend on how far light penetrates the water. In the Northeast Atlantic, typical maximum depths are 1200 meters. Seasonal planktonic migrations are most prevalent at higher latitudes and in regions where currents cause seasonal upwellings of nutrients. (Many planktonic species demonstrate diel vertical "migrations" in the water column, but because these constitute daily ranging movements rather than migrations by the definition given earlier, we do not consider them further.)

D. Amphibians and Reptiles

Amphibian life cycles often require a return to water to reproduce, resulting in migrations from feeding areas or refugia to seasonal ponds, streams, and other water sources. However, salamanders and anurans are not known for their great mobility and consequently most species do not travel far from where they were born. Red-bellied newts (Taricha rivularis) may hold the distinction for their migrations of 1 km from feeding range to breeding site. Similar migrations in anurans may extend several kilometers. Still, migrating anurans are remarkable for their fidelity to specific breeding sites and the precision with which individuals are able to return to the same few meters of shoreline each year. Lizards very rarely migrate, but those that do also move to seek out suitable nesting sites. Green iguanas and a few related species that nest on islands may move as much as several kilometers to lay eggs at specific sites, the scarcity of appropriate soil in their island habitats presumably drives these migrations. The red-sided garter snake (Thamnophis sirtalis parietalus) migrates to winter hibernacula, such as limestone sinkholes, aggregating in the thousands to buffer the harsh winter conditions of Manitoba.

E. Insects and Other Terrestrial Invertebrates

Our current state of knowledge regarding insect migrations contains a number of spectacular examples but gives the overall impression that the phenomenon is not common in this group despite its huge contribution to our planet's biodiversity. Well-known examples include the Eurasian milkweed bug (Lygaeus equestris) and the North American ladybird beetle (Coleomegilla maculata), which perform seasonal movements from summer feeding and breeding sites to more protected areas a few kilometers away. There they diapause and last out the winter surviving on stored fat. Individuals may coalesce at specific sites and form spectacular aggregations. As in other migratory species we have considered, migration allows both species to exploit rich, but seasonally variable, habitats. Because many insects have highly specialized feeding and breeding requirements, even small seasonal variations in temperature, moisture, and light levels may be sufficient to trigger migrations to sites where individuals can wait for local conditions to improve (i.e., diapause). The general migration pattern illustrated by ladybirds and milkweed bugs may thus be widespread among both temperate and tropical species. However, small insects are likely to migrate only short distances and be overlooked and understudied, especially if migrations occur without individuals aggregating into conspicuous groups.

An alternative migration pattern takes individuals from habitats declining in quality to richer sites allowing feeding and breeding to continue. Among the noctuid, or armyworm, moths—a group found throughout the world—several species follow annual round-trip migratory circuits of several hundred kilometers that take them to successively higher latitudes or to habitats recently freshened by rain with host vegetation more suitable for breeding. In many cases, these species have become significant agricultural pests having a life history ideally suited to quickly exploiting ephemeral but rich food sources. Interestingly, single generations may complete only part of the migratory circuit, raising fascinating questions regarding the evolutionary genetics of migration timing and navigation. Perhaps the most famous migrant of this type is the monarch butterfly, Danaus plexippus, of North America. Monarch larvae feed on milkweeds and the timing of migration coincides with seasonal milkweed growth. In autumn, adult butterflies from as far north as eastern Canada migrate more than 1000 km south to overwinter in huge aggregations at only a few critical sites in the mountains of central Mexico. In spring, these same adults migrate 200 to 300 km to the northern edge of the Gulf of Mexico where they breed and die. Successive generations then move north tracking the milkweed growing season. When large-scale migrations occur in insects, they are truly astounding behaviors if one considers the size
of the travelers and the distances covered. Because insects have only limited capacity for fuel storage, they depend more greatly than other migrants on winds to propel and direct their movements. North American aphids, for example, fly for 2 to 24 hours and travel from 50 to 1100 km before setting down. Differing wind speeds at different altitudes account for the variation in flight time and distance moved, but aphids probably exert some control over these features by choosing an altitude at which to fly.

In response to deteriorating local conditions, either due to seasonal reductions in food or increased densities, some insects develop long-winged (i.e., macropterous) forms able to travel relatively large distances to find more suitable habitat. Examples include planthoppers and aphids. These irruptive movements also occur in other taxa, such as birds, but they resemble dispersal processes more than the migrations we have so far considered.

Terrestrial crustaceans, such as hermit crabs and ghost crabs, must migrate from nonbreeding, feeding habitats to high-salinity water in order to breed. These movements occur seasonally, especially in response to tidal cycles, and may require movements from 10 to 3000 m each way.

IV. ATTRIBUTES OF MIGRANTS AFFECTING SUSCEPTIBILITY TO HUMAN DISTURBANCE

As we have seen, migrations appear to serve a few common functions. In general, they enable species to temporarily avoid harsh conditions or to meet important biological needs that are separated by great distance. Not surprisingly, there are certain attributes of migrants that affect their susceptibility to human disturbance.

First, migratory species often use a variety of habitats, leaving them vulnerable to multiple points of disturbance. Often harm is felt mostly at one destination. Bachman’s warbler (Vermivora bachmani), for example, was driven to extinction by the destruction of its overwintering habitats in the tropics. For others, however, such as many Neotropical migrant birds, the impacts of human activities are felt at both endpoints of their migrations. Globe-trotting species like the white-rumped sandpiper typify the precarious dependence of migrants on habitat health across a tremendous geographic scale. But harm need not be limited to the endpoints of a migrant’s journey. Any diminution in quality of refueling sites along the way could winnow a population and limit its ability to replenish its numbers before the next cycle of migration. And given that migratory trajectories for many species are shaped by the vagaries of prevailing winds or currents, conservation strategies entailing the protection of all stopover areas becomes almost impossible. Even where resting points are protected, unintended consequences associated with the normal nonintrusive behavior of naturalists can put migrants at risk. Steady viewing by bird-watchers at refuges along coastal fly routes has been known to force birds to move too far offshore where they can no longer feed on inshore marine invertebrates exposed at low tides.

Sea turtles probably illustrate best the effects of migratory species being vulnerable at many life-history stages to the excesses of human behavior. Not only are their eggs sought after and easily harvested by indigenous peoples, the beaches themselves are often degraded by the activities of affluent humans. Entire breeding populations are eliminated when beach habitats are developed or severely compromised in their abilities to launch young when dune buggies destroy nests or excessive night lighting disrupts the water-seeking behavior of newly hatched young. In addition, for those young that mature to subadults and return to the breeding grounds, nets of fishermen, particularly shrimpers, provide the coup de grâce to the species by drastically reducing the pool of future breeders. In fact, it has been shown for that the most vulnerable period in the life cycle is not the nestling, but rather the subadult, stage. Although protecting beaches and increasing the number of functioning beaches gives the species a head start by diversifying recruitment sites, sea turtles are most vulnerable to extinction if the number of subadults is reduced. It is at this point in the species’ life history where a long life of breeding commences and reproductive value is highest. Only by insisting that shrimpers insert turtle excluder devices (TEDs) in nets so that the turtles can escape before drowning will there be any hope that those migratory species already endangered will survive.

Second, migratory species often aggregate when traveling. While this behavior might reduce each individual’s risk of falling prey to nonhuman predators, both aquatic and terrestrial species in groups are easy targets for the advanced harvesting gear employed by commercial fishers and hunters. In prehistoric times, humans stampeded herd animals off cliffs and into canyons where massive kills occurred. Such actions have been implicated in the extinction of many of the North American megafauna. Today even the smaller, more elusive...
and difficult to capture prey are at risk. Many marine fisheries involving migratory schooling fishes, such as those of cod and haddock, have been overfished. And despite moratoria, many are not recovering.

Third, since many migratory species move according to very precise schedules, any delays caused by human alteration of the habitat or disruption of normal movements could lead to a cascade of effects, jeopardizing a species’ ability to replenish its population numbers reduced by ordinary mortality. As noted earlier, migrating shorebirds if forced to linger longer at refueling refuges, could have breeding seasons shortened sufficiently to lose the ability to lay one or more clutches. And since selection favors renesting because nest predation rates are already high, any force constraining such efforts could severely limit a species’ recruitment potential.

Despite these attributes that threaten species survival there are some beneficial traits exhibited by migrants. By occupying, at least temporarily, many different habitats, migrants can spread risks and thus can escape catastrophes that befall one location. Unless all individuals in a species simultaneously occupy the same habitat, survivors from habitats not impacted by the catastrophe can serve as sources for increasing species numbers. Migratory species are also more likely to find and colonize new habitats as they are opened up by climate change or other human-induced and natural changes in the environment.

V. ECOLOGICAL CONSEQUENCES OF HUMAN DISTURBANCE ON MIGRANTS

As humans fragment landscapes, many migratory species will find themselves in peril. For species that must move from one region to another in order to meet a variety of biological needs, any barrier disrupting these movements could lead to local extinctions. In southern Africa, fences are being built either to prevent mixing of wildlife and livestock and the concomitant transmission of disease (ensuring that exported beef is disease free) or to prevent wildlife from gaining a competitive edge over livestock when both compete for critical resources such as food near water. Without access to these resources at the time of year most critical for developing juveniles, any migratory population’s recruitment will be severely curtailed. Many local populations will either go locally extinct or will be transformed from “sources” of new emigrating individuals to “sinks” where excess individuals from healthy populations find refuge. Even for populations that still boast numbers in the hundreds of thousands, such as the plains zebra or even African elephants (Loxodonta africana), fragmentation of the large populations is well underway. With human populations expanding into habitats only marginally suited for horticulture, the cry for fencing areas to prevent crops being consumed by migrating herbivores is increasing.

When the movements of these populations are disrupted, a cascade of indirect effects on both the recycling of nutrients in ecosystems and the structuring of animal communities is likely to occur. Since migratory species represent some of the largest aggregations and highest densities of individuals seen in the animal world, any disruption to migration is likely to reduce local density. Thus it is quite possible that the important impact on the structure of vegetation, or more generally the recycling of nutrients, will be altered. It has even been proposed that the overharvesting of sperm whales and the removal of their large carcasses from the ecosystem have impacted the community stability of the deep-sea communities associated with deep-sea vents. By preventing carcasses from falling to the ocean floor, an important source of renewing “island” resources has been eliminated, thus making it more difficult for species tightly associated with vents to hop from one vent to another. With respect to structuring communities, if plains zebras are unable to move in large numbers across the landscape from the tops of catenas where they forage during the rains to the wetter valleys where they move when the rains cease, then their facilitative effect on diversifying the herbivore community will be reduced. Typically, plains zebra move to the wetter valleys and graze down the tougher, fibrous forage that dominates these areas making the greener and more nutritious forage available for the ruminants that require such higher quality vegetation. Thus by disrupting the movements of such a “keystone” species, the diversity of large grazers could be reduced. Furthermore, enclosure experiments suggest that the grazing community could shift from one dominated by large ungulates to one populated by small rodents and lagomorphs. Since rodents often come in contact with people and harbor human diseases, the cascade of effects could be quite profound.

Perhaps one of the greatest impacts of disrupting movements will be seen via the effects that elephants have on the landscape. As populations become restricted to reserves if their numbers are allowed to increase, then their ability to transform a mosaic landscape of trees and grasslands into one of mostly grasslands will mean the disappearance of resident for-
VI. EVOLUTIONARY CONSEQUENCES OF HUMAN IMPACTS ON MIGRANTS

While it is clear that human behavior disrupting migrations will have ecological consequences, the effects on the evolutionary potential, or even the future characteristics, of species are also likely to be large. Nowhere is this impact more evident than in North America where the range of the bighorn sheep has been severely reduced and fragmented. Bighorn migrate seasonally from meadows to mountaintops to avoid harsh climate and its search of grass that changes seasonally in quality at each location. As human populations have expanded, populations have been isolated into refuges, numbers are shrinking, and genetic diversity is being lost. In addition, hunting for trophy males is removing just those individuals with the best and most diverse genotypes. As a result, the ability of the isolated populations to remain genetically diverse enough to forestall problems associated with inbreeding or to avoid the ravages that would occur if a novel disease appeared has been severely compromised.

Migratory Atlantic salmon (Salmo salar) provide perhaps the clearest example of what human activities are causing a species to change under our own eyes. Atlantic salmon typically are born in fresh water and develop for 1 to 2 years before smolting and heading to sea to grow and fatten by feeding on a rich supply of marine invertebrates. Once they attain a certain size, they become sexually mature and return to rivers. There they travel long distances upstream until they reach clear and cool breeding grounds. Some individuals, however, never head to the sea. Instead, they remain in their natal streams and mature sexually at young ages and at small sizes. Such individuals are called parr and they never go through the smolting process that adapts them to a marine lifestyle. Under pristine environmental conditions, the fraction of the population that becomes parr is small since the reproductive gains of such a strategy are low. When competing for mates with larger more aggressive males, parr fare poorly. What matings they obtain are derived by “sneaking” among a mating pair and releasing milt at just the right time. Such events are rare and the mixing of milt is poor so such sneak matings result in few young being sired by parr. But the survival prospects of parr are high since they do not incur the risks of going to sea and, moreover, they begin breeding at a very early age. Thus over a lifetime, reproductive success is modest. But as human fishing increases and the netting of the older larger salmon intensifies, the relative lifetime reproductive success of parr is improving. Since the costs of migrating to sea and back again have increased dramatically, the long life span of parr gives them a relative advantage. And given that competition with the larger males for mates is also declining, the chances of parr securing matings are also improving. Thus it is not...
surprising that the composition of the population is changing as parr increase markedly in abundance. The impact on the long-term stability of the population is unclear. But the long-term impact on the species appearance is clear, just as it is for yet another fishing industry that will go into decline as the large fish disappear.

See Also the Following Articles
ARCTIC ECOSYSTEMS • BIRDS, BIODIVERSITY • DISPERSAL, BIOGEOGRAPHY • MOTHS

Bibliography

I. INTRODUCTION

The molluscan body plan typically consists of four body components: (1) a head with tentacles and eyes, (2) a ventral muscular foot, (3) a dorsal visceral mass, and (4) the enveloping mantle that secretes the shell. A space between the covering mantle and the side of the foot forms a mantle (or pallial groove). In most molluscs the groove deepens in the posterior region to form a cavity that contains a pair of gills, or ctenidia, as well as openings of the rectum, paired renal organs, and gonads.

The mouth typically is equipped with jaws and a radula. Numerous glands and sacs are associated with the mouth (or buccal chamber) and esophagus, which opens into a stomach. A digestive gland also opens into the stomach. The hindgut, or intestine, is typically long and highly looped or coiled. The nervous system is concentrated in the head and consists of four ganglia—cerebral, visceral, pedal, and pleural. They are connected by commissures; the paired pedal nerve cords extend ladder-like through the foot. The circulatory system consists of a heart enclosed within the pericardium. The heart has multiple auricles and a single ventricle. The system is typically open except in cephalopods, which have capillaries. The excretory system is paired and connected to the pericardium as well as the gonads in some taxa. The gonads are also paired, but can be fused into a single structure (Polyplacophora) or reduced to a single organ (Gastropoda). Separate gonoducts are present in some taxa, and in other taxa
the gonads empty into the kidneys. These connected, mesodermal structures (pericardium, kidneys, and gonads) likely represent the reduced coelom of the Mollusca.

The characters that diagnose the Mollusca are a shell-secreting mantle with a tripartite mantle edge divisible into outer, middle, and inner folds; the secretion of an outer proteinaceous layer, or periostracum; the secretion of multiple crystal types and arrangements (microstructures) within the shell; and a pericardium. Plesiomorphic characters include a mantle groove with flow-through water current patterns, bilateral symmetry, and cephalic tentacles.

This body plan has been substantially modified both among and within groups. Diversification appears to have occurred very early in the history of the Mollusca, but there has been surprisingly little change in some groups. For example, the shells of late Cambrian monoplacophorans are almost identical to those of living taxa despite 450 million years of evolution. Other examples of little change include protobranch bivalves, nautiloids, and scaphopods. There have also been some large and notable extinctions, including those of the bellerophonoids, hyoliths, rostroconchs, and ammonites.

II. TEMPORAL AND SPATIAL DISTRIBUTIONS

The Mollusca include some of the oldest metazoans found in the fossil record. Late Precambrian rocks of southern Australia and the White Sea region in northern Russia contain bilaterally symmetrical, benthic animals with a univalved shell (Kimberella) that resembles those of molluscs in many respects. The earliest unequivocal molluscs are found in the early Cambrian, and there is even some support for a Precambrian presence (Coeloscleritophora). Most of the familiar groups, including gastropods, bivalves, monoplacophorans, and rostroconchs, all date from the early Cambrian, whereas cephalopods are first found in the middle Cambrian, polyplacophorans in the late Cambrian, and the Scaphopoda in the middle Ordovician. Most of these taxa tend to be small (<10 mm in length). After their initial appearances, taxonomic diversity tends to remain low until the Ordovician, when gastropods, bivalves, and cephalopods show strong increases in diversity. For bivalves and gastropods this diversification increases throughout the Phanerozoic, with relatively small losses at the end-Permian and end-Cretaceous extinction events.

Cephalopod diversity is much more variable through the Phanerozoic, whereas the remaining groups (monoplacophorans, rostroconchs, polyplacophorans, and scaphopods) maintain low diversity over the entire Phanerozoic or became extinct (Fig. 1). Molluscs occur in almost every habitat found on Earth. Large concentrations of gastropods and bivalves are found at hydrothermal vents in the deep sea, whereas other gastropods live above tree line in Arctic tundra and on mountain slopes. Molluscs occur in the wettest environments of tropical rain forests and in the driest deserts, where annual activity patterns may be measured in hours. They live below ground in the lightless world of aquifers and caves, and even interstitially in groundwater (stygobionts), but are also familiar organisms of the seashore, rivers, and lakes and even our gardens. The only thing molluscs do not do is fly, unless the passive aerial dispersal of minute land snails among the Pacific Islands by high winds is considered to constitute flying.

Like many other organisms, marine molluscs reach their highest diversity in the tropical western Pacific and decrease in diversity toward the poles. Marine diversity is also highest nearshore and becomes reduced as depth increases beyond the shelf slope. In terrestrial communities gastropods can achieve great diversity and abundance: as many as 60–70 species may coexist in a single habitat and abundance in leaf litter can exceed more than 300 individuals in 4 liters of litter. Abundance and diversity for some groups can also be higher in temperate communities than in tropical settings. In freshwater communities, species diversity tends to be lower but abundances of sessile molluscs such as the zebra mussel Dreissena polymorpha can exceed more than 3000 individuals per square meter.
III. MAJOR GROUPS

The major groups of living molluscs are clearly dissimilar from one another and have long been recognized as distinct taxa. However, not all were originally recognized as belonging to the Mollusca. For example, the worm-like bodies of the aplacophorans were perplexing to early biologists and required study of their internal anatomy to ultimately recognize their affinities with the other molluscan groups. This problem becomes especially acute with fossil taxa; the extinct groups (indicated with a dagger) discussed below may or may not be molluscs in our current delimitation of the taxon based on living representatives. However, there is little doubt that at some more inclusive grouping these fossil taxa share common ancestors with living molluscan groups.

The converse problem obtains for living taxa. For example, whereas it is possible to relate living taxa to one another using both morphologic and molecular characters, there exists the real possibility that the living taxa do not share a single most recent common ancestor, but may have had multiple, independent derivations from distantly related molluscs or mollusc-like taxa that are now extinct (see below). These alternatives require that both fossils and living taxa be studied and incorporated into evolutionary scenarios and hypotheses of molluscan relationships, especially when the fossil record provides such a wealth of fossils and putative relatives.

A. Coeloscleritophora†

The worldwide presence of small, hollow, calcareous sclerites in numerous Precambrian and Cambrian sediments (collectively referred to as “small shellies”) was an enigmatic component of molluscan evolutionary studies. However, in the early 1990s, an articulated fossil was found in the lower Cambrian of Greenland that was covered with small shellies. It was immediately apparent that what had been thought to be the remains of individual organisms were actually parts of a single larger animal (Fig. 2). Recent work in the Cambrian of Europe, Asia, and Australia has greatly expanded our knowledge of Coeloscleritophora, and although their relationship to the Mollusca remains uncertain, they likely share a common ancestry with the molluscs as well as with annelids and brachiopods.

B. Hyolitha†

Hyoliths (Fig. 3) had cone-shaped shells with a (presumably ventrally) flattened side. The aperture was closed by an operculum, and a single pair of anterior appendages (or Helens) is present in some taxa. The operculum was attached to the shell by pairs of symmetrical muscles. Shells were constructed of crossed lamellar calcium carbonate crystals, and the apical tips bore larval shells. Two forms of larval shells—a smooth globose shell and a point shell with growth lines—perhaps reflected direct and indirect development, respectively. A looped alimentary system connected the ventral mouth to the dorsal anus. Hyoliths were sessile, epifaunal deposit feeders, scrapping sediments from the seafloor. They first appear in the early Cambrian and became extinct at the end of the Paleozoic.

C. Rostroconchia†

Rostroconchs (Fig. 4) begin life with a single conch (or valve), but the single shell transforms ontogenetically into a nonhinged bivalved shell that gapes at its margins. Based on muscle scar patterns, the animals appear to...
D. Polyplacophora

Polyplacophoran molluscs, or chitons (Fig. 5), have a dorsal shell that is secreted as eight valves or plates. The plates of individual animals differ in size and shape based on their sequential position; this plate dimorphism is also present in the earliest chitons of the late Cambrian. Although known since the early Paleozoic, chitons do not show a marked increase in diversity until the Cretaceous. Chitons occur worldwide in intertidal habitats and at depths in excess of 7000 m. There are about 850 species and they live on a variety of firm substrates ranging from rocks to algae.

Shell plate morphology has been an important character in determining taxa and relationships. The plates are composed of two distinct shell layers. The outer surface of each plate is high in organic content layer and overlies an inner calcareous layer. Plates may be covered by a thin periostracum. The posterior and anterior edges of each plate have projecting flanges and slits for articulation with other plates, and the outer surfaces are sculptured into distinct regions. The outer plate layer is penetrated by numerous vertical canals that connect to the mantle edge via a canal system that lies between the outer and inner layers. The canals come in two sizes and connect sense organs (including eyes) on the dorsal surface of the valves with the mantle’s nervous system.

The plates are surrounded by the girdle, which is formed by a strongly differentiated mantle epidermis. The mantle secretes the plates and surrounds their articulated form. Around the outer plate boundaries the mantle epidermis becomes thicker and secretes a vast array of scales, spines, bristles, and other structures on its dorsal surface. In some species the anterior portion of the mantle epidermis is modified into a large hood that is used to trap food; most species feed by grazing on hard substrates. A poorly defined head and elongated foot that is surrounded by the ventral mantle epidermis of the girdle mark the ventral surface of the animal. The gills are located in the mantle groove that is formed between the girdle and the lateral sides of the foot. The excretory pores and gonoducts also open into this groove.

The visceral mass of the chiton is dominated by a long looped alimentary system. Above the digestive system is a bilobed gonad that appears to result from the fusion of a single pair of gonads. Two separate gonoducts lead from the gonads to opposite sides of the mantle groove. The excretory organs are paired and each connects with the pericardial cavity via a short canal. The center of the pericardial cavity contains an elongated ventricle that is surrounded on either side by enlarged auricles that collect blood from the lateral gills. The polyplacophoran nervous system is ladder-like and almost identical to that of the aplacophorans and monoplacophorans; however, it differs from both by lacking cerebral ganglia. The radular musculature is also similar to that of the monoplacophorans, and the radular tooth configuration is docoglossate like that of the monoplacophorans and patellogastropod limpets. The typical chiton radula has a weak central tooth that is flanked by strong lateral teeth and weaker marginal teeth, some of which are only plates.

Chiton development has been well studied. The larvae come from relatively large eggs and are nonfeeding. The eggs are covered by a thick hull with elaborate spines and knobs. The chiton embryo hatches as a trochophore, a globular cell mass with an apical tuft and central band of cilia, and subsequently elongates and differentiates into a small chiton. Larval eyes are...
Molluscs

E. Aplacophora

The morphological distance between an aplacophoran mollusc (Fig. 6) and, for example, a cephalopod mollusc stretches anyone’s concept of the rank of phylum. Aplacophora were the last class to be placed in the Mollusca and have until recently been one of the least studied groups. The aplacophoran molluscs include two different morphological groups: the Chaetodermomorpha or Caudofoveata, and the Neomeniomorpha or solenogastres. The Chaetodermomorpha are footless vermiform molluscs that live in sediments. The Neomeniomorpha are more elongate, have a narrow foot, and typically live in association with cnidarians such as hydroids and alcyonaceans. They are found throughout the world and live at depths between 18 and 6000 m. It is estimated that there are about 250 species of aplacophorans.

The shell of the aplacophoran mollusc is a myriad of calcareous spicules that are secreted by the mantle epidermis. Spicule morphology varies over the body of the aplacophoran, and in some taxa spicules are modified into scales.

The calcareous spicules that cover the bodies of most aplacophorans give the animals a striking sheen. In the Neomeniomorpha there is a pedal groove on the ventral surface of the animal. The pedal groove expands into an anterior pedal pit and contains the narrow foot. The Chaetodermomorpha lack a pedal groove but do have a posterior mantle cavity that contains a Gill. In some Chaetodermomorpha there is a slight constriction and distinct change in spicule morphology at the midline of the body.

It is the internal anatomy that provides evidence of the molluscan identity of the aplacophorans. In both groups the anterior end of the alimentary system includes a radula and odontophore. In the Chaetodermomorpha the radula and odontophore are strongly developed, and the alimentary system is more differentiated than in the Neomeniomorpha. Both groups have a dorsal gonad that opens into the pericardium, which contains the heart. From the posterior portion of the pericardium there extends a coelomoduct that loops or bends and ultimately opens into the mantle cavity. In the Neomeniomorpha the posterior portion of the coelomoducts are modified for reproductive functions such as sperm storage or brooding young. The nervous system is ladder-like with a well-developed cerebral ganglion. Radular configurations are quite variable and show a wide range of tooth development and modifications that include jaw-like structures, denticles with cones, and sweepers. This is one of the greatest ranges of radular variation found in a Mollusca, and it stands in marked contrast to the lack of variation in the radular configurations found in both the Monoplacophora and Polyplacophora.

Development of the aplacophoran mollusc includes a test cell larval stage in which the three tissue types (mesoderm, ectoderm, and endoderm) align and differentiate within an exterior cell layer constructed of large test cells. Aplacophoran eggs are relatively large and free-spawned in the Chaetodermomorpha and fertilized internally in the Neomeniomorpha; some Neomeniomorpha brood their young to various stages of development. After the formation of the test cell larva with an apical tuft and prototroch, the posterior development of the differentiating larva quickly outgrows the constraints of the exterior test and develops directly into the juvenile aplacophoran.

After many years without students, the Aplacophora are now the subject of extensive studies. These studies and interpretations of aplacophoran phylogeny have focused attention on this small group of molluscs. Overall, the aplacophoran body plan is very similar to that of the chiton. Aplacophorans and polyplacophorans differ from the monoplacophorans by having a dorsal gonad rather than a posterior gonad. The pericardium is similar in all three groups as are many of the other organ systems and positions. Major differences are found in the type of shell secreted by the dorsal mantle epidermis.

F. Bivalvia

Bivalves (Fig. 7) are laterally compressed molluscs with a hinged, two-part shell. A nonglobal ligament serves as a hinge and connects the two valves along their dorsal surfaces and acts to force the valves apart. The remaining shell surfaces may tightly close or gape with
contraction of the musculature. Adult valves can vary greatly in shape and size and certain shell forms appear associated with specific habitats. The shell can be internal and reduced (or even absent) and the bivalve animal worm-like such as in “shipworms” (Teredo). Although the first occurrences of taxa referred to the Bivalvia are found in lower Cambrian deposits, it is not until the lower Ordovician that bivalve diversification, both taxonomic and ecological, explodes in the fossil record. This diversification continues unabated through the Phanerzoic, with relatively small losses at the end-Permian and end-Cretaceous extinction events. Today over 10,000 species of bivalves are found in most marine, brackish, and freshwater communities. They may be infaunal or epifaunal, and epifaunal taxa may be either sessile or motile. Bivalves typically display bilateral symmetry both in shell and in anatomy, but there are significant departures from this theme in taxa such as scallops and oysters.

Bivalve shells are constructed of different shell fabrics, including crossed lamellar, nacreous, and foliated microstructures. Most of the variability in shell structure sorts along higher taxon divisions. For example, nacreous structures are present primarily in the basal members of the group (Protobranchia) whereas crown taxa have primarily crossed lamellar shells (Heterobranchia). The outer shell surface is often covered by a periostracum, and the free margins of the shell of some taxa may be uncalcified and flexible. In some taxa such as oysters, one valve is cemented to the substrate. The hinge area often contains an array of sockets and pegs, referred to as “teeth” or hinge dentition. These teeth align the valves as they close and prevent shearing of the valves. The inner surfaces of the valves have scars formed by the adductor shell muscles, typically positioned in the anterior and posterior regions and connected by the pallial line that marks the ventralk most attachment point of the mantle.

The mantle that envelops the bivalve animal forms a large ventral chamber or mantle cavity with paired ctenidia. The mantle edge along the posterior portion of the shell often forms a pair of siphons that facilitate inflow and outflow of water through the mantle cavity. Although the pleisomorphic feeding state for bivalves is likely deposit feeding, the ctenidia provide an effective filter-feeding mechanism in most taxa with numerous levels or grades of organization. Bivalves lack a well-defined head and a radula. The foot is often triangular and in some species, such as mussels, the larval byssal gland is retained into the adult and attaches the individual to the substrate with strong fibers.

The visceral mass is primarily situated above the mantle cavity and continues ventrally into the foot. The mouth is often flanked by palps and opens into a stomach. The intestine is irregularly looped and opens dorsally into the excurrent flow near the exhalant siphon. Also opening into this region are the paired kidneys and, when separate from the kidneys, the gonopores of the paired gonads. The heart typically lies below the center of the valves and consists of two auricles and a single ventricle that supplies both anterior and posterior aorta. The nervous system is made up of three pairs of ganglia. These innervate the musculature, mantle, viscera, ctenidia, and siphons. They receive sensory input from statoliths, osphradium, various siphonal sensory structures, and photoreceptors along the mantle margin.

Bivalve development has been well studied, primarily because of the economic importance of bivalves (see below). The eggs are typically small and not very yolk-rich. Fertilization is usually external, but in brooding species occurs in the mantle cavity. Cleavage patterns are spiral and both polar lobes and unequal cleavage patterns are present throughout the group. The first larval stage is typically a trochophore that transforms into a veliger. Although morphologically similar to the gastropod veliger stage, phylogenetic analyses suggest that these larval stages are homoplastic rather than homologous. The initial uncalcified shell grows laterally in two distinct lobes to envelop the body. Larval bivalves have a byssal gland that may assist with flotation while planktic but later attaches the juvenile to the substrate.

G. Monoplacophora

For more than 50 years the Monoplacophora (Fig. 8) were known only from the fossil record and were recognized by the presence of multiple, symmetrical muscle
scars in patelliform shells. The discovery of living monoplacophorans, beginning with *Neopilina galatheae* in 1957, provided anatomical characters for the group. About 20 living species of monoplacophorans have since been discovered worldwide, living at depths between 174 and 6500 m. They are found both on soft bottoms and on hard substrates on the continental shelf and seamounts. Paleozoic taxa are associated with relatively shallow water faunas (>100 m).

In Recent and fossil patelliform monoplacophoran shells, the apex is typically positioned at the anterior end of the shell, and in some species actually overhangs the anterior edge of the shell. Aperture shapes vary from almost circular to pear-shaped. Shell height is also variable and ranges from relatively flat to tall. Shell sculpture varies from rugose concentric ridges to fine growth lines. In some taxa radial sculpture produces a fine reticulate pattern.

The monoplacophoran animal has a poorly defined head with an elaborate mouth structure on the ventral surface. The mouth is typically surrounded by a V-shaped, thickened anterior lip and postoral tentacles; postoral tentacles come in a variety of morphologies and configurations. Below the head lies the semicircular foot. In the mantle groove, between the lateral sides of the foot and the ventral mantle edge, are found five or six pairs of gills (less in minute taxa).

Internally, the monoplacophoran is organized with a long, looped alimentary system, two pairs of gonads, and multiple paired excretory organs (four of which also serve as gonoducts). A bilobed ventricle lies on either side of the rectum and is connected via a long aorta to a complex plumbing of multiple paired atria that in turn are connected to the excretory organs. The nervous system is ladder-like and has weakly developed anterior ganglia, paired muscle bundles enclose the visceral mass. The most interesting anatomical structures in the original description of a living monoplacophoran were the paired "dorsal coeloms" that are connected to the anterior excretory organs and are topographically similar to the fused gonads in chitons. Later studies suggest that these structures are more likely extensions of glands associated with the pharynx. The monoplacophoran radula is docoglossate: it has a rachidian tooth, three pairs of lateral teeth, and two pairs of marginal teeth. The lateral teeth are in a stepped configuration: the first and second pairs of lateral teeth are aligned with the rachidian tooth, and the third lateral pair is slightly posterior and lateral to the first two pairs. The cusp of the innermost marginal tooth is frilled.

Developmental studies of monoplacophorans have not been done. Recent monoplacophorans form a distinct clade, and their similarities and differences with the other extant molluscan groups are easily recognized. There is little question that some Paleozoic taxa also are members of this clade. However, the characters that distinguish some Paleozoic monoplacophorans from torted molluscs (i.e., the Gastropoda) and vice versa are open to alternative interpretations, and the relationships of several major groups of early-shelled molluscs have therefore been the subject of much lively debate.

### H. Scaphopoda

The scaphopod mollusc (Fig. 9) combines gastropod-like shell morphology with bivalve-like development with cephalopod-like anatomy. Members of the class first appear in the early Paleozoic and the taxon has maintained a slow and steady rate of increase in morphological diversification since then. Our knowledge of the biology of recent species comes primarily from members of a single genus, *Dentalium*. Scaphopods are infaunal organisms and feed on foraminiferans and other interstitial prey. Approximately 350 species occur from the intertidal zone to depths in excess of 7000 m and are present in all the major oceans.

The scaphopod shell is a calcium carbonate tube with equal or unequal apertures; the tube may be either inflated or bowed. The shell microstructure includes...
prismatic and crossed-lamellar components; the latter is similar in structure to elements seen in members of the Bivalvia.

The head and foot extend through the ventral aperture of the scaphopod shell. The head bears numerous tentacles (capitacles) that are used to capture and manipulate prey. Foot morphology is variable and has been used as a taxonomic character. The pattern of water circulation through the scaphopod mantle cavity is also unique among the molluscs, because water both enters and exits through the small dorsal aperture. Scaphopods also lack gills.

Unlike the previously discussed groups, scaphopods have a U-shaped gut rather than an anterior–posterior configuration of the mouth and anus. The stomach and digestive gland are in juxtaposition, and the intestine loops before passing through the excretory organ and opening into the mantle cavity. The posterior portion of the digestive gland overlies the gonad that connects with the mantle cavity via the excretory organ. There may be a reduced pericardium, but a heart is absent.

The radula consists of a rachidian plate, a single first lateral tooth, and a lateral plate.

The ontogeny of several species has been documented. The trochophore larva has an apical tuft and prototroch and develops mantle folds that divide the larva into equal lateral halves. A velum forms from the prototroch on the anterior end and the dorsal and lateral sides of the larvae begin to secrete a shell that fuses along the ventral surface of the larva but leaves an opening at the posterior end. The trilobate foot forms ventrally under the velum, and the larva attains the tubular shell morphology of the adult scaphopod. The loss of the velum and completion of organogenesis after settling complete metamorphosis.

Scaphopods have an intriguing set of molluscan characters that have been allied to several scenarios of molluscan evolution and relationships. Their shell structure and development suggest bivalvian affinities, but scaphopods also have a radula. The gross morphology of the scaphopod gut is U-shaped, like that of the gastropods and cephalopods, rather than linear as in monoplacophorans, polyplacophorans, and aplacophorans. However, like monoplacophorans, scaphopods have a gonad that lies under the alimentary system rather than on top of it as in aplacophorans and polyplacophorans. Some workers have suggested that aplacophoran larvae and scaphopod larvae share some features; however, others are unconvinced by these arguments and compare the scaphopod velum to that of higher bivalves.

It has been suggested that scaphopods are descended from riberid rostroconchs and they have been grouped with the bivalvia. Though there is little doubt that scaphopods share some characters with the Bivalvia, the direct derivation of the scaphopod mollusc from a riberid rostroconch is contradicted by the U-shaped gut present in scaphopods. Rostroconchs are thought to have had a linear gut. Thus, the transition from the riberid condition to a scaphopod morphology would include the construction of a U-shaped alimentary system from a linear one, although the ancestral linear gut would have been an ideal exaptation for a scaphopod shell morphology open at both ends and with flow-through water circulation. Like many other characters, novelties and confusing patterns are common in molluscan morphology but so are convergent grades (especially in molluscan shell form).

I. Gastropoda

Externally, gastropods (Fig. 10) appear to be bilaterally symmetrical; however, they are one of the most successful clades of asymmetric organisms known. The ancestral state of this group is clearly bilateral symmetry (e.g., chitons, cephalopods, bivalves, see above), but gastropod molluscs twist their organ systems into figure-eights, differentially develop or lose organs on either side of their midline, and generate shells that coil

FIGURE 10 External morphology and diversity of the Gastropoda. Top, typical benthic gastropod mollusc. Bottom, a pelagic heteropod. Modified from Gray, Maria; Figures of molluscous animals, selected from various authors. 1857–1859.
MOLLUSCS

243

to the right or left. And although they have left the more common molluscan state of bilateral symmetry, gastropod molluscs are one of the most diverse groups of animals, both in form and habitat. They occupy habitats ranging from the deepest ocean basins to the highest mountains and from the tropics to high latitudes. Estimates of total extant species range from 40,000 to over 100,000, but may be as high as 150,000, with about 15,000 named genera for both Recent and fossil species. They have figured prominently in paleobiological and biological studies, and have served as study organisms in numerous evolutionary, biomechanical, ecological, physiological, and behavioral investigations. They have a long and rich fossil record that shows periodic extinctions of subclades, followed by diversification of new groups.

The best documented source of gastropod asymmetry is the developmental process known as torsion. Like other molluscs, gastropods pass through a trochophore stage, and then form a characteristic stage of development known as the veliger. During the veliger stage a 180° rotation of the mantle cavity from posterior to anterior places the anus, and renal openings over the head, and twists organ systems that pass through the snail’s ‘waist’ (the area between the foot and visceral mass) into a figure eight. This rotation is accomplished by a combination of differential growth and muscular contraction. In some taxa the contribution of each process is about 50:50, but in other taxa the entire rotation is accomplished by differential growth. Although the results of torsion are the best known asymmetries in gastropods, numerous other asymmetries appear independent of the torsion process. Anopod flexure, sometimes considered a feature of torsion, is widely distributed in the Mollusca; it is present in the extinct hyoliths as well as in the Scaphopoda and Cephalopoda (and to a lesser extent in the Bivalvia).

Trends in gastropod evolution often feature a reduction in the complexity of many characters. These include reduction of the number of radular teeth, simplification (thought to be due to shell coiling) of the renopericardial system (loss of right auricle and renal organ), reduction of ctenidia (loss of the right gill), and associated circulatory and nervous system changes. There is also a reduction of diversity of shell microstructures, simplification of the buccal cartilages and muscles, reduced coiling of hindgut, and simplification of the stomach. Other characters show an increase in complexity, such as life history characters (e.g., internal fertilization with penis and spermatophores and associated reproductive organs). This increase in complexity is correlated with the ability to produce egg capsules and the evolution of planktotrophic larvae and direct development. There is also an increase in chromosome number, and greater complexity of sensory structures (e.g., eyes, osphradium). In the pulmonates (land snails) the pallial cavity is modified into a pulmonary cavity or lung, while in the opisthobranchs (sea slugs) there are secondary gills and elaborate neurosecretory structures.

Generally, gastropods have separate sexes. One major exception is the heterobranch clade (pulmonates & opisthobranchs), which is exclusively hermaphroditic. Basal gastropods typically spawn gametes directly into the water column, where they are fertilized and develop. Brooding of developing embryos is widely distributed throughout the gastropods, as are sporadic occurrences of hermaphroditism. These basal groups also have non-feeding larvae. Eggs are relatively small in most taxa, although eggs of taxa with non-feeding larvae tend to be a little larger than those taxa that have feeding larvae. Egg size is reflected in the initial size of the juvenile shell or protoconch; when preserved in fossil taxa, this feature has been useful in distinguishing feeding and non-feeding taxa in the fossil record. The first gastropod larval stage is typically a trochophore that transforms into a veliger and then settles and undergoes metamorphosis to form a juvenile snail.

J. Cephalopoda

Cephalopods (Fig. 11) are dorsoventrally elongated marine molluscs that may or may not have a recognizable external or internal shell. Cephalopods are the most complex and mobile of the nonvertebrate metazoans and...
show numerous modifications of the general molluscan body plan. The nautiloids first appear in the late Cambrian and undergo a rapid diversification in the Ordovician. Ammonites (Coleoidea) appear later in the Devonian. Cephalopods are much more variable in their diversity through time than other molluscan groups. They are hit by numerous extinctions (e.g., terminal Permian, Triassic, and Cretaceous events) but typically showed rapid replacement by the survivors. The cephalopods include the largest living as well as the largest extinct molluscs: ammonite shells range to over 2 m across and body sizes of living squid range up to 8 m, with tentacles exceeding 20 m in length. Over 10,000 species are known with about 700 living taxa. All are active carnivores in marine benthic and pelagic habitats from nearshore to abyssal depths.

Cephalopods are thought to have evolved from mol-luscophoran-like ancestors. Septa formed at the apex as the animal grew and withdrew into a newly formed body chamber. The old chambers are gas-filled and provide buoyancy for the organism. The foot was modified into a funnel and provides jet propulsion for movement. Prehensile arms with suckers surround the mouth on the head (cephalic in origin) and capture prey. In cephalopods with external shells (ammonites and nautiloids), the shells are composed of an inner nacreous layer and an outer prismatic layer. In other cephalopods the shell is typically internal and reduced to a linear stiffener as in cuttlefish and squid or virtually nonexistent as in the octopus, whose pelagic ancestors reinvaded the benthic realm.

Much of the external anatomy of cephalopods appears associated with their highly motile, pelagic habits. These modifications are so strong that it is difficult to readily identify the typically molluscan body axes in a cephalopod. For example, as a squid jets through the water, the most posterior portion of the body, relative to the direction of movement, is the squid’s head with its prehensile arms. The most anterior portion is the dorsal surface of the muscular mantle and the posterior mantle cavity is located ventrally. In addition to jets of water expelled through the funnel, squid and some octopi use undulating movements of paired fins at the distal end of the mantle for swimming as well.

The cephalopod mouth contains strong beak-like jaws that are used to deliver a lethal bite to prey items as well as shred them; a radula is also present. Salivary glands can produce highly toxic venoms in some squids and octopi, and ink sacs are present in most taxa except Nautilus. The digestive system differs from that of most other molluscs in having a cecum in juxtaposition with the stomach and a relatively short intestinal tract. There are two pairs of ctenidia in living Nautilus, but all remaining cephalopods (cuttlefish, squid, and octopus) have a single pair of ctenidia. The circulation system also differs markedly from that of other mollusks in being a closed system capable of maintaining a high blood pressure, rather than open and diffuse as in other taxa. The heart consists of two auricles and a single ventricle that supplies multiple arteries. In all cephalopods except Nautilus, a pair of branchial hearts situated at the base of the ctenidia pump blood through the ctenidia. The nervous system is highly concentrated and developed in cephalopods. The three major ganglia are fused and organized into lobes, each with its own specific function. Nautilus has fewer lobes than other cephalopods, and lobe size may vary among groups with different lifestyles. Coleoid cephalopods also have two large stellate ganglia on the mantle that control both respiratory and locomotory functions of the mantle. Sensory structures include statocysts, olfactory organs, and eyes. Coleoid eyes are surprisingly convergent with vertebrate systems and are capable of resolving brightness, shape, size, and orientation. Cephalopods are also able to change color patterns using an elaborate chromatophore system that is under nervous control; it provides them with the ability to display incredible camouflage.

Cephalopods have separate sexes and spermato- spheres are transferred between males and females by modified tentacles. Eggs are large and yolk-rich, and embryoic development of cephalopods is different from that of all other mollusks. There is no larval form, just direct development into juveniles. Both the eggs and young may be brooded, benthic, or pelagic.

IV. ECOLOGY

Molluscs occur on a large variety of substrates, including rocky shores, coral reefs, mussel beds, and sandy beaches. Gastropods and chitons are characteristic of these hard substrates; bivalves are commonly associated with softer substrates, where they burrow into the sediment. However, these patterns are not inflexible. The largest living bivalve, Tridacna gigas, occurs on coral reefs, whereas microscopic gastropods live interstitially between sand grains. On hard substrates gastropods are often grazers that feed either selectively or indiscriminately on algal, diatom, or blue-green films and mats or animal aggregations. On soft substrates bivalves typically are suspension or deposit feeders.

The adoption of different feeding habits appears to have had a profound influence on molluscan evolution.
The change from grazing to other forms of food acquisition is one of the major features in the adaptive radiation of the group. Based on our current understanding of relationships, the earliest molluscs were likely carnivores and detritivores that indiscriminately grazed on encrusting animals and detritus. Truly herbivorous grazers are relatively rare in the molluscs and are limited to the polyplacophorans and gastropod groups, whereas living aplacophorans, monoplacophorans, scaphopods, and cephalopods are carnivorous. Some protobranch bivalves are also carnivorous; most bivalves are either suspension or deposit feeders that indiscriminately take in particles but then elaborately sort them based on size.

Cephalopods are typically active carnivores specialized on mobile prey such as fish, crustaceans, and other cephalopods. Because they are so abundant in pelagic systems, cephalopods are often important food sources for larger fishes, marine mammals, and even seabirds. In the gastropods several groups such as Janthinidae are planktic pelagic carnivores whereas the heteropods (Caenogastropoda) and the pteropods (Opisthobranchia), like the cephalopods, are active swimmers in search of prey. These taxa spend their entire lives in the water column where they primarily feed on Cnidaria, other molluscs (including small cephalopods), and even fishes.

Bivalve-like suspension feeding has evolved in some gastropods such as the turritellid Umbonium and in the pelagic gastropod group Thecosomata. Some groups with carnivorous diets have undergone what appear to be true, explosive, adaptive radiations (e.g., the Neogastropoda). Other carnivorous taxa such as the aplacophorans and scaphopods have low diversity and abundance.

In addition to these more typical trophic strategies and interactions, some molluscs (especially among the gastropods) have also become both endo- and ectoparasitic in and on other invertebrates—most often echinoderms. Galeommatoxid bivalves and eulimid gastropods occur as ectoparasites on holothurian and asteroid echinoderms, respectively. The reduced, worm-like Entomobranchia occurs in and on various substrates in the deep sea, in fast-flowing rivers, or in quiet lakes and ponds. It is often suggested that strong wave action selects for limpet morphologies in several groups, but it is obvious from their current distributions that limpets do very well in a wide range of habitats.

V. MOLLUSCS AND HUMANS

Molluscs and humans are most often associated economically. Molluscs have many important commercial benefits such as fisheries and mariculture but can also be responsible for tremendous economic loss and human suffering. Molluscs are found in some of the earliest human habitation sites in southern Africa over 100,000 years ago, and it is likely that humans have included molluscs in their diet and as material resources for millions of years. More recently, commercial fisheries have focused on bivalve, gastropod, and cephalopod taxa for food (oysters, abalone, scallops, clams, cephalopods, and mussels) as well as pearl and button fisheries.

In freshwater habitats molluscs may serve as intermediate hosts for parasites such as digenetic trematodes and are an integral part of the Schistosomiasis contagion that infects millions in the tropical regions of the world. Nonhuman, mollusc-borne parasites are also responsible for more than $2 billion worth of losses worldwide to the livestock industry each year. Most people are familiar with “garden snails” and the damage they can do to favorite plants, but this local damage is minimal compared to the role of molluscs as pests. Molluscan pest species cause millions of dollars worth of damage to horticultural and food crops throughout the world every year, and as the global economy expands, the
probability of potentially damaging introductions increases dramatically. For example, over the past 5 years, almost 5000 molluscs from 100 different countries were intercepted entering the United States. One of the most notable introductions has been that of the freshwater bivalve *Dreissena polymorpha*, or zebra mussel. Zebra mussels were first detected in Lake St. Clair (Great Lakes chain) in 1988. It is thought that zebra mussels entered the Great Lakes in ballast water of commercial ships from eastern Europe, and without natural enemies to control their numbers, they have thrived and spread to 19 states in less than 10 years. In the United States and Canada, they have caused widespread economic and environmental damage. They typically concentrate where water flows rapidly and therefore can quickly clog intake pipes and cut off water flow to hydroelectric and nuclear power plants, industrial facilities, and public water supplies. One mussel removal project in Lake Michigan removed 400 cubic yards of zebra mussels at a cost of $1.4 million. Zebra mussels can also have tremendous effects on freshwater ecosystems by displacing native species and increasing levels of toxic algae. As zebra mussels have spread across the country, they have devastated native freshwater clam populations (especially the unionid species) that were already endangered and threatened by pollution and habitat modification.

On a more positive note, molluscs are also important in biomedical and biotechnology research. Although not as well known as model organisms such as the fruit fly *Drosophila*, opisthobanch molluscs are commonly used in neurological studies of learning and memory functions, and bivalve molluscs have served as models for environmental and genetic interactions in tumor research. Extracts of numerous other molluscs have been examined for antibacterial and other properties potentially useful in the medical and health sciences. Molluscs also provide important model organisms for environmental health studies. For example, bivalve molluscs are frequently used in studies on the toxicology and environmental monitoring of heavy metals and other pollutants. Molluscan conservation issues are primarily centered on freshwater and terrestrial taxa. The largest number of documented extinctions has involved nonmarine molluscs, and nonmarine molluscs are second to the arthropods in numbers of currently threatened species. Terrestrial species are impacted primarily by the destruction of native vegetation and cover, and freshwater species by alter flow, sedimentation rates, and pollution. Although marine taxa are thought to be less impacted by human activities, populations of some taxa have been substantially reduced by overfishing. These include abalone (*Haliotis*), clams (*Tridacna*), and scallops (*Pecten*). Marine taxa, like freshwater and terrestrial taxa, are also threatened by habitat destruction, pollution, changes in sedimentation patterns, and other factors.

As discussed above regarding zebra mussels, the introduction of exotic species can also seriously impact native taxa. This is especially true on islands where the effective population size may be small. One of the best-documented cases was the apparent extinction of the endemic land snail taxon *Partula* from the island of Moorea in French Polynesia. This resulted because of the intentional introduction of the predatory snail *Euglandina rosea* to the island as a biocontrol agent for the Giant African Snail, *Achatina fulica*.

### VI. CLASSIFICATION

Given the state of flux in molluscan and metazoan phylogeny and the wealth of new data that is now appearing from molecular, morphological, and paleontological work, any classification proposed here would be rapidly outdated. Several traditional classifications are available in the references cited in the bibliography. However, few are based on hypotheses of relationships, but are instead based on overall similarity and ad-hoc scenarios of evolution.

Classifications based solely on morphology have been especially problematic, and much of this confusion has resulted from problematic taxa such as the aplaco- phorans, scaphopods, and bivalves where possible reduction and loss of organs or other secondary simplification has produced morphologies that may be argued as either primitive or highly derived. Many classifications have also focused exclusively on the morphology of living taxa and have ignored potential, fossil members of the Mollusca. If extinct fossil taxa are included in evolutionary scenarios, they are typically limited to distinctive clades such as the Rostroconchia and Bellerophonta. Other more problematic extinct taxa (e.g., hyoliths) are systematically ignored, arbitrarily excluded from the Mollusca without analysis, or shoehorned into extant groups.

The sister taxa of Mollusca have included the Platyhelminthes, Annelida, Sipuncula, and the Kamptozoa. Within the Mollusca both Polyplacophora and Aplacophora have been argued as the most primitive taxon, and thus the outgroup to all Conchifera.

Most classifications have assumed a single cladogenetic event in the origin of the Conchifera from the...
MOLLUSCS

supposedly more primitive placophoran groups. Alternative hypotheses have derived the conchiferans in an unresolved polytomy from a hypothetical ancestral mollusc, or HAM. Some workers have interpreted the Cambrian Burgess Shale taxon Wiwaxia and other less complete halkieriid-like fossils as molluscan while others have argued Wiwaxia to have annelid worm affinities. However, the discovery of an articulated halkieriid from the lower Cambrian and the existence of these and other multishelled placophorans necessitate the reexamination of long-held assumptions of molluscan ancestry and monophyly. The rapidly increasing knowledge of Coeloscleritophora diversity suggests that they may harbor independent ancestors for extant molluscan groups.

Molecular phylogenies for the Mollusca have not fared much better than the morphological studies. Nuclear and mitochondrial DNA sequences have had very limited success in resolving a monophyletic molluscan clade or even producing robust or reasonable groupings within the Mollusca (e.g., the bivalves and gastropods). These problems most likely result because of the deep, Paleozoic divergence of many of the molluscan taxa and the variable rates of change across molluscan genomes.

The following rank-free classification is conservative and only denotes the major clades within the Mollusca.

Coeloscleritophora
Sipuncula
Hyolitha
Mollusca
Chaetodermomorpha
Neomeniomorpha
Polyplacophora
Conchifera
Rostroconchia
Bivalvia
Protobranchia
Prosobranchia
Heterodonta
Anomalodesmata
Monoplacophora

Scaphopoda
Dentaliida
Gastropoda + Cephalopoda
Gastropoda
Patellogastropoda
Vetigastropoda
Neritopsina
Caenogastropoda
Heterobranchia
Cephalopoda
Nautiloidea
Coleoidea
Aminonotoidea
Decapoda
Octopoda

See Also the Following Articles

INVERTEBRATES, FRESHWATER, OVERVIEW • INVERTEBRATES, MARINE, OVERVIEW • INVERTEBRATES, TERRESTRIAL, OVERVIEW

Bibliography


MOTHS

David L. Wagner
University of Connecticut

I. Introduction
II. Higher Classification
III. Fossil History
IV. Species Diversity and Biogeography
V. Life History
VI. Feeding Biology
VII. Natural Enemies and Chemical Defense
VIII. Dispersal and Migration
IX. Key Adaptations
X. Pheromones
XI. Coloration, Diurnality, Nocturnality, and Attraction to Light
XII. Importance
XIII. Conservation

GLOSSARY

aposematic Warningly colored; boldly colored, usually involving reds, oranges, or yellows, as well as black and white.
crochets The minute hooklets on the fleshy abdominal prolegs of a caterpillar.
diapause A period of delayed development that is generally associated with weather conditions that are unfavorable to survival or reproduction.
detrivory Feeding on fallen and generally dead organic debris; although the term covers both plant or animal matter, in moths the term is especially apt to apply to leaf litter feeders.
hypermorphogenic A life cycle that includes two or more larval forms, with each often specialized for a different feeding function.
instar A larval stage; the first instar hatches from the egg and upon molting enters the second instar. Most moths undergo five or six instars prior to pupation.
maxillae The second pair of mouthparts, located between the mandibles and the labium (the third pair of mouthparts).
monophyletic A group with a single evolutionary origin, which includes a common ancestor and all of that ancestor's descendants.
Mullerian mimicry When two or more distasteful species come to resemble one another. In Batesian mimicry there is a model (unpalatable species) and one or more mimics (palatable species).
parasitoid An insect whose larval stage feeds on a second (host) species, killing the host in the process; a parasitoid attack is equivalent to delayed predation and therefore the phenomenon is distinguished from true parasitism.
pharate A "cloaked" or hidden stage, for example, the adult moth just prior to its emergence from the pupa.
pheromone A chemical released by one individual that elicits a response in a second individual of the same species.
polyphagous Eating plants from more than two unrelated plant families.

World moth diversity is believed to be in excess of 225,000 species, virtually all of which are dependent...
on the explosive Cretaceous radiation of the flowering plants. This chapter begins with a synopsis of the higher classification and diversity of the order Lepidoptera, then reviews several aspects of the order's biology. Throughout the chapter, emphasis is placed on the basal lineages on which the group's success is rooted. Special effort is made to discuss the key innovations and circumstances that may have led to the enormous global diversity and importance of moths.

I. INTRODUCTION

Every night at dusk one and a half million Mexican free-tailed bats flood out from beneath the Congress Avenue Bridge in Austin, Texas. Bats from this single colony harvest more than 10,000 pounds of insects every evening, most of which are moths. Moth caterpillars account for much of the above-ground insect biomass in temperate forests; without them one wonders whether there would be songbirds to usher in each spring. Estimates of species-richness for moths have climbed steadily over the past decade, with the more modest extrapolations falling between 200,000 and 300,000 species. Whether measured in terms of biomass, influence on terrestrial ecosystems, or species-richness, moths must be held as one of the most successful lineages of macroscopic organisms on this planet.

Body sizes span two orders of magnitude: Nepticulidae and Heliozelidae, some with wingspans under 3 mm, are little more than aerial plankton (Fig. 1). The world's largest moth, a neotropical noctuid, has a wingspan that sometimes exceeds 275 mm. All members of the order Lepidoptera can be diagnosed by the presence of scales, which cover the wings and body. A second readily observable structure that is part of the order's ground plan is the epiphysis, a cuticular flap on the inside of the foreleg, which is used by many moths to clean the antenna. A curious lepidopteran attribute is the production of apyrene (anucleate) sperm—in some species 80% or more of the sperm is apyrene and therefore incapable of fertilization. Kristensen (1984, 1999) identified nearly two dozen other synapomorphies, which, taken as a whole, incontrovertibly established the monophyly of the Lepidoptera and its sister-group relationship with the Trichoptera (caddisflies). Molecular studies of both nuclear and mitochondrial genes have corroborated the major findings of his benchmark studies.

Butterflies and moths make up the order Lepidoptera. Because butterflies derive from within one (or two) moth lineages, it is not possible to identify uniquely evolved features shared by moths that are not also found in butterflies. Recent phylogenetic reconstructions indicate that butterflies have their closest relatives in the Hedyloidae, an odd-looking group of neotropical moths that were formerly classified in the Geometroidea (Fig. 2). But regardless of their phylogenetic position within the order, since butterflies are, evolutionarily speaking, nothing more than diurnal moths (and negligible in terms of species richness), “Lepidoptera” and “moths” are used synonymously throughout this chapter.

II. HIGHER CLASSIFICATION

Phylogenetic relationships among the early lineages of Lepidoptera are among the best known of any insect
order, so well that they have even been used as a "truth table" for testing the phylogenetic signal in several gene sequencing studies. Despite this agreement, there is little consensus as to how to render the phylogeny (Fig. 3) into a classification. Kristensen (1999, p. 28), frustrated by the redundancy and instability of higher classifications for the order, adopted a phyletic sequencing scheme (i.e., he arranged taxa in a hierarchical outline without formal assignation of many higher category names).

A. The Four Suborders

The most recent higher classification for the Lepidoptera recognizes 46 superfamilies, all but two of which (the Hesperioidea and Papilionoidea) are moths (Kristensen, 1999, and authors therein; Fig. 3 and Table I). The 120 families have been arranged in four suborders: the Zeugloptera, Aglossata, Heterobathmiina, and Glossata. The first and most basal lineage within the order contains but a single family, the Micropterigidae, with circa 180 known (121 of which are described) species distributed across the major continents. Decidedly primitive, this family shares many morphological traits with the Trichoptera. Noteworthy are the functional mandibles in the adult (and absence of a proboscis). The larvae are bizarre creatures with little morphological affinity to other lepidopteran caterpillars (Fig. 4a), so odd that at one time these moths were placed in their own order. The Aglossata contains but a single family (Agathiphagidae) with two species, restricted to Australia, New Caledonia and a few Pacific islands. The third suborder, the Heterobathmiina, is represented by nine (three described) species from temperate areas of Chile and Argentina. All are classified in a single genus in the Heterobathmiidae. The remaining diversity occurs in Glossata, which possess, among other traits, a coiled proboscis.

B. The Glossatan Infraorders

The most basal lineage within Glossata is the Dacnonypha, with a single principal family, the Eriocraniidae. These are small leaf-mining moths of the Northern Hemisphere with 24 recognized species. The remaining glossatan infraorders have been grouped in the Coelolepida, all of which share "normal" wing scale ultrastructure and lack ocelli (light-sensitive but non-image-forming lenses over the dorsum of the head). (Ocelli reappear in some derived moth lineages.) The placement of the Acanthopteroctetidae, another small family, remains in question; some authors place it in the Eriocraniidae (Dacnonypha) and others in the Coelolepida. Australia's lophocoronids are superficially similar to both of these families but are accorded their own family, superfamily, and infraorder. Although the life history of all six lophocoronids remains unknown, the piercing ovipositor indicates that the larvae are internal...
# TABLE I
**Diversity of the Lepidoptera**

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Approx. no. families</th>
<th>Approx. no. described species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suborder Zeugloptera</td>
<td>1</td>
<td>121</td>
</tr>
<tr>
<td>Micropteroidea</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Agathiperoidea</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Heterobathmioidea</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Infracrateroidea</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Acathompheroidea</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Lepchoerutoidea</td>
<td>5</td>
<td>609</td>
</tr>
<tr>
<td>Infracrateroidea</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Nepticulioidea</td>
<td>2</td>
<td>902</td>
</tr>
<tr>
<td>Palaephatoidea</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>Tischeroidea</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>Infracrateroidea</td>
<td>1</td>
<td>588</td>
</tr>
<tr>
<td>Division Ditrysia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section Tineina</td>
<td>4</td>
<td>2,320</td>
</tr>
<tr>
<td>Symotheroidea</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Timoidea</td>
<td>9</td>
<td>4,530</td>
</tr>
<tr>
<td>Gascinleridoidea</td>
<td>8</td>
<td>1,428</td>
</tr>
<tr>
<td>Gelechionidea</td>
<td>13</td>
<td>15,940</td>
</tr>
<tr>
<td>Elachidoidea</td>
<td>3,210</td>
<td></td>
</tr>
<tr>
<td>Oecophtheridae</td>
<td>3,130</td>
<td></td>
</tr>
<tr>
<td>Gelechidae</td>
<td>4,530</td>
<td></td>
</tr>
<tr>
<td>Section Apolobrora</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>Gaebetidoidea</td>
<td>11</td>
<td>2,548</td>
</tr>
<tr>
<td>Sammioidea</td>
<td>3</td>
<td>1536</td>
</tr>
<tr>
<td>Cosmidoidea</td>
<td>2</td>
<td>676</td>
</tr>
<tr>
<td>Tortricidoidea</td>
<td>1</td>
<td>8000</td>
</tr>
<tr>
<td>Choristidoidea</td>
<td>1</td>
<td>405</td>
</tr>
<tr>
<td>Urodidoidea</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>Schneckeraltoidea</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Epertemoidea</td>
<td>1</td>
<td>85</td>
</tr>
</tbody>
</table>

| Infracrateroidea | |
| Nepticulioidea | 2 | 902 |
| Palaephatoidea | 1 | 60 |
| Tischeroidea | 1 | 80 |
| Infracrateroidea | 1 | 588 |
| Division Ditrysia | |
| Section Tineina | 4 | 2,320 |
| Symotheroidea | 1 | 4 |
| Timoidea | 9 | 4,530 |
| Gascinleridoidea | 8 | 1,428 |
| Gelechionidea | 13 | 15,940 |
| Elachidoidea | 3,210 |
| Oecophtheridae | 3,130 |
| Gelechidae | 4,530 |
| Section Apolobrora | 1 | 17 |
| Gaebetidoidea | 11 | 2,548 |
| Sammioidea | 3 | 1536 |
| Cosmidoidea | 2 | 676 |
| Tortricidoidea | 1 | 8000 |
| Choristidoidea | 1 | 405 |
| Urodidoidea | 1 | 60 |
| Schneckeraltoidea | 1 | 5 |
| Epertemoidea | 1 | 85 |

Continued

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Approx. no. families</th>
<th>Approx. no. described species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alucoidea</td>
<td>1</td>
<td>148</td>
</tr>
<tr>
<td>Protocreneidea</td>
<td>1–3</td>
<td>986</td>
</tr>
<tr>
<td>Whalleyoidea</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Immidoidea</td>
<td>1</td>
<td>246</td>
</tr>
<tr>
<td>Copromorphoidea</td>
<td>2</td>
<td>313</td>
</tr>
<tr>
<td>Hyllioidea</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Thordila</td>
<td>1</td>
<td>760</td>
</tr>
<tr>
<td>Pyralidae</td>
<td>2</td>
<td>17,763</td>
</tr>
<tr>
<td>Pyralidae</td>
<td>6,133</td>
<td></td>
</tr>
<tr>
<td>Crambidae</td>
<td>11,630</td>
<td></td>
</tr>
<tr>
<td>Menniulidoidae</td>
<td>1</td>
<td>240</td>
</tr>
<tr>
<td>Lasiocampoidae</td>
<td>2</td>
<td>1,373</td>
</tr>
<tr>
<td>Rombidoidea</td>
<td>9</td>
<td>3,354</td>
</tr>
<tr>
<td>Saturinidae</td>
<td>1,395</td>
<td></td>
</tr>
<tr>
<td>Sphinxidae</td>
<td>1,269</td>
<td></td>
</tr>
<tr>
<td>Asinidae</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Callidulidae</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>Hedyloidea</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>Neptunioidea</td>
<td>1</td>
<td>3,300</td>
</tr>
<tr>
<td>Papilionoidea</td>
<td>4</td>
<td>13,600</td>
</tr>
<tr>
<td>Lycaridae</td>
<td>6,000</td>
<td></td>
</tr>
<tr>
<td>Nymphalidae</td>
<td>6,000</td>
<td></td>
</tr>
<tr>
<td>Drepanoidea</td>
<td>2</td>
<td>673</td>
</tr>
<tr>
<td>Geometridae</td>
<td>3</td>
<td>21,340</td>
</tr>
<tr>
<td>Noctuoidae</td>
<td>8</td>
<td>41,300</td>
</tr>
<tr>
<td>Arctiidae</td>
<td>11,000</td>
<td></td>
</tr>
<tr>
<td>Lymantriidae</td>
<td>2,300</td>
<td></td>
</tr>
<tr>
<td>Noctidae</td>
<td>23,000</td>
<td></td>
</tr>
<tr>
<td>Notodontidae</td>
<td>2,800</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>149,677</td>
<td></td>
</tr>
</tbody>
</table>

1 With but few exceptions numbers were taken from Kristensen 1999. The figure for the Tortricoidea was suggested by John Brown; those for the Bombycoidea and Noctuoidea were suggested by Ian Kitching. The classification is adopted from Kristensen (1999) and Heppner (1998). Species-numbers are given for all 46 superfamilies. Families, other than the nominate, that contain more than 2500 described species are also given. A few additional families that are mentioned repeatedly in the text are also included.

feeders. The neopseustids are peculiar insects that in some ways resemble lacewings. Larval stages are unknown, but morphological evidence suggests that they feed internally; the 11 species are found in South America, China, and Southeast Asia. The Exoporia, with six families and more than 600 species, are the only
basal lineage with appreciable species diversity. Three of the families are Gondwanan: Mnesarchaeidae (New Zealand, 14 species; 8 described), Protothoroidae (South Africa, 12 species), and Anomosetidae (Australia, 1 species). The largest family, the Hepialidae, with 387 species is cosmopolitan, the most generic diversity within the family also is Gondwanan. The largest family, the Hepialidae, with 587 species is cosmopolitan; the most generic diversity within the family also is Gondwanan. The remaining 37 superfamilies usually are placed into a single taxon, the Heteroneura—the name referring to the venational differences between the fore- and hindwing—the monophyly of which is still in question.

C. The Heteroneurans

Phylogenetic relationships among many heteroneuran groups remain equivocal even in the face of detailed morphological and molecular studies. This uncertainty may be because these early evolutionary splits correspond to a period of rapid diversification within the order—the events occurring so closely together in time that it will be difficult to accurately reconstruct their sequence. There is weak morphological evidence suggesting that four superfamilies (Nepticuloidea, Incurvarioidae, Paleaphatoidea, and Tischertoidea) form a monophyletic group, the Monotrysia. All share an unusual configuration of the female reproductive system, whereby there is a single gonopore for copulation and oviposition (and in the Incurvarioidae this is fused with the anus to form a cloaca). But other morphological and molecular data indicate that the characters holding this group together are symplesiomorphic (i.e., that they should be regarded as part of the ground plan for all heteroneurans). In Freidlander et al.’s recent molecular analyses employing gene sequence data from dopa decarboxylase, there was little support for the monophyly of the Monotrysia.

D. The Ditrysia

The remaining 33 heteroneuran superfamilies are placed in the Ditrysia, a group whose monophyly is well supported by the presence of separate copulatory and ovipositional pores, an internal duct that connects the two, and the unique organization of the proboscis.
musculature. This morphologically homogeneous clade contains 98.3% of the described species of Lepidoptera. Published phylogenies for ditrysian superfamilies and families are replete with uncertainty. In all likelihood considerable gene sequence data will have to be collected before consensus is reached about phylogenetic relationships within this remarkably successful group of insects. Many guides and faunal works divide the order into the so-called Microlepidoptera and Macrolepidoptera. Microlepidoptera roughly corresponds to the basal groups, most of which are small to minute, possess a fused CuP vein in the hindwing, have the larval crochets arranged in a circle, and possess a jugum or the remnants of one (all of which are primitive features that were part of the ground plan for the order). It is a grouping of convenience as the “macros” are clearly derived from the “micros,” and therefore some micros are more closely related to the macros than they are to other micros. The monophyly of the Macrolepidoptera, which includes the last 11 superfamilies in Table I, is still being debated: the crochets are arranged in a linear series parallel to the body axis, the first axillary sclerite in the wing base is elongate, and, obviously, the wingspans average consistently larger than those of the micros. Scott discussed a number of other characters that argued for the monophyly of the Macrolepidoptera, but exceptions are so numerous that some authors are reluctant to employ the term in their classifications.

III. FOSSIL HISTORY

While the fossil record of the Trichoptera, the sister group to the Lepidoptera, is adequate and dates into the Permian, the record for lepidopterans is meager, especially for Mesozoic-aged fossils, when the basal lineages were diversifying. There are five putative Lepidopteran fossils from the Triassic; these cannot be assigned to the order with certainty. The oldest unequivocal moth, Archeolepis manura, dates to the Lower Jurassic. By definition, sister taxa are of equal age, and thus it is curious that the oldest caddisfly and moth fossils are of such disparate ages. Skalski has raised an interesting point—that some early stem group Lepidoptera may be unrecognized because currently they are assigned to the Trichoptera.

More than a dozen moth fossils from the Upper Jurassic and Cretaceous are unassignable to a family. While most of these clearly belong to prodiglossatan taxa, this is not true of all. Protolepis caprealata from Upper Jurassic deposits in Kazakhstan appears to have a coiled proboscis. This finding has special significance because phylogenetic reconstructions for the order have an angiosperm-feeding lineage (the Heterobathmima) branching off prior to the evolution of the Glossata (Fig. 3). Because the oldest documented angiosperms are Lower Cretaceous (and Protolepis is Upper Jurassic), this suggests that either specialized angiosperm-feeding evolved at least twice within the Lepidoptera or that the date for the origin for angiosperms is, in fact, much older.

Microptergids have been identified from several amber and sedimento deposits. Examples from the lower Cretaceous represent extinct genera, while those from the Tertiary are assigned to modern genera. There are no fossil records for either the Aglossata or Heterobathmima. Glossatans, including a putative lepcho- nirid, are well substantiated from the Upper Cretaceous. By 97 mya, modern genera of ditrysians are recognizable, suggesting that early evolution of the order occurred on nonseed plants and gymnosperms during the late Jurassic (Lahaneira et al., 1994). Not surprisingly, most of the modern superfamilial and species diversity appear after the angiosperms began radiating in the early Cretaceous. The oldest unambiguously macrolepidopteran fossils are Paleocene.

The most enigmatic macrolepidopteran fossil is the noctuid egg reported by Gall and Tiffney from the Upper Cretaceous (75 mya). Because noctuids are thought to be among the most derived ditrysian superfamilies (Fig. 3), their presence in the Cretaceous would push back the origin of virtually all lepidopteran higher taxa. Arguing against Gall and Tiffney's identification is the observation that all present-day noctuids have a thora-cic “ear” sensitive to the ultrasonic frequencies emitted by bats and that the oldest bat fossils date back only to the Eocene. Perhaps then, it is not surprising that the oldest unequivocal noctuid fossils also date to the Eocene.

Leaf-mining families from the Northern Hemisphere have an extensive fossil record that is believed to reflect their ages of origin as well as their relative abundances. Labandeira et al.’s (1994) discoveries of mines of Ectoedemia and Stigmella (both Nepticulidae) and a phyllocnistine (Gracillariidae) from Cenomanian deposits (97 mya), which agree in detail with those of modern species from closely related host taxa, are noteworthy in that they demonstrate the antiquity of many moth-plant associations. Another especially well-represented group of fossil moths is made up of the tineoids, which are richly represented in Dominican amber deposits that date from 13 to 40 mya.
IV. SPECIES DIVERSITY AND BIOGEOGRAPHY

There is still considerable doubt about the species richness within the order, an alarming observation given that Lepidoptera are among the most well-studied group of invertebrates. Two recent compilations, which are far from independent, give estimates of 145,677 (Table I) and 146,565 (Table I) for the world lepidopteran fauna. Heppner (1998) suggested that half again as many species remain to be described, and that species richness for the order would climb to 255,000 (Table II). Kristensen and Gaston have offered even higher estimates, in the range of 300,000 to 500,000 species.

The world’s least speciose biogeographic realm, the Palearctic region, is the only one that can be regarded as well known. Heppner’s tallies indicate that 92% of all the Lepidoptera that occur in the Palearctic have received names (Table II). The richest biogeographic realm, the Neotropics, with an estimated 35% of the world’s butterfly and moth fauna, has only about 50% of its fauna named. Yet these figures are probably overly optimistic. In recent revisions of neotropical tineids or tortricids, Davis and Powell, respectively, found that 75 to 90% of the species were new. Of the more than 250 species of Gracillariidae that have been collected at the La Selva Biological Station in Costa Rica, 98% is believed to be undescribed.

There can be little doubt that microlepidopterans will be a frontier for descriptive taxonomy for decades if not centuries. In general, small cryptically colored taxa have greater proportions of unnamed species. Also, species-rich taxa that are intractable taxonomically have proportionately greater numbers unstudied. Microlepidopteran superfamilies with especially high numbers of undescribed species include the Gracillarioidea, Tineoidea, Gelechioidea, and Pyraloidea. But even among the Macrolepidoptera, considerable taxonomic work remains, especially for neotropical taxa. The only two appreciably diverse moth families where the descriptive taxonomy is approaching completion are the Saturniidae and Sphingidae, veritable behemoths among insects that have long been favored among collectors.

Present tallies suggest that the Noctuoidea is without parallel in richness, with more than 41,300 described species, most of these falling within the nominate family (i.e., the Noctuidae, Table I). It is the most species-rich superfamily on all continents except Australia, where its preeminence is superseded by the Gelechioidea. Current tallies of described species place the Geometroidea second in richness with some 27,140 species worldwide, nearly all of which are in the nominate family. Probably its ranking will fall to at least fourth once world microlepidopteran faunas are studied. Presently, there are approximately 15,500 described species of Gelechioidea and 16,000 Pyraloidea. The former likely contains in excess of 50,000 species and may well prove to be the most species-rich superfamily within the order; the latter contains at least double that of the described number. The butterflies, with 17,300 species, most of which have been given names, will rank no higher than fifth in global importance. In the most well-studied faunas (i.e., those of the north temperate zone and Australia) moths make up 93 to 95% of the lepidopteran fauna, with butterflies making up the remainder. In terms of biomass or number of individuals, moths and their caterpillars probably have even greater importance, except in open and early successional habitats where butterflies thrive.

Because tropical faunas are very incompletely known, especially for microlepidopterans, a quantitative assessment of global diversity patterns for the entire order is not yet possible. Nevertheless, virtually all lepidopterists are in agreement that greatest species, generic, familial, and superfamilial diversity is found in

---

**TABLE II**

Worldwide Lepidopteran Species Diversity by Biogeographic Region (1758–1990)

<table>
<thead>
<tr>
<th>Biogeographic Region</th>
<th>Neotropical</th>
<th>Palaearctic</th>
<th>Ethiopian</th>
<th>Oriental</th>
<th>Australia</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Described species</td>
<td>11,332</td>
<td>44,791</td>
<td>22,465</td>
<td>20,491</td>
<td>27,683</td>
<td>19,603</td>
</tr>
<tr>
<td>(% of world fauna)</td>
<td>(7.9)</td>
<td>(30.6)</td>
<td>(15.3)</td>
<td>(14.0)</td>
<td>(18.8)</td>
<td>(13.4)</td>
</tr>
<tr>
<td>Estimated richness</td>
<td>14,000</td>
<td>50,000</td>
<td>25,000</td>
<td>30,000</td>
<td>30,000</td>
<td>255,000</td>
</tr>
<tr>
<td>(% of world fauna)</td>
<td>(9.5)</td>
<td>(35.3)</td>
<td>(19.9)</td>
<td>(19.9)</td>
<td>(19.9)</td>
<td>(100)</td>
</tr>
<tr>
<td>% of fauna known</td>
<td>82</td>
<td>51</td>
<td>92</td>
<td>54</td>
<td>53</td>
<td>50</td>
</tr>
</tbody>
</table>

From Heppner (1998).

---
MOTHS

tropical rain forests. Preliminary estimates of the moth fauna of Costa Rica (51,100 sq. km) run between 13,000 and 16,000 species and thus are roughly comparable to the species diversity of the entire North American continent north of Mexico or that of Australia, areas 380 and 130 times greater in size, respectively. Families with exceptional tropical richness include the Arctiidae, Cosmopterigidae, Gelechiidae, Gracillariidae, Oecophoridae, Lymantriidae, Notodontidae, and Sphingidae.

Given that moths are quintessentially phytophagous, their species diversity should parallel that of vascular plants. If this is correct, the highest global richness should occur on the low to mid-elevational slopes around the Amazon basin. The forests of Southeast Asia and Indo-Australia also would be expected to harbor hyperdiverse moth faunas. In temperate regions, the floras of South Africa and southwestern Australia are among the richest in the world. These latter areas are noteworthy because they support comparatively high percentages of nonditrysian moths and hence can be regarded as two of the world’s most important “living museums.” Obviously, not all taxa are hyperdiverse in the tropics. Familial, generic, and species diversity for the nonditrysian moth lineages is highest in extratropical latitudes. In temperate areas of the Northern Hemisphere, the Coleophoridae are among the most numerically important Lepidoptera in grassland, scrub, and desert ecosystems, yet the family appears to be absent in the rain forests of Central America. The Tortricoidea also appear to be less diverse than would be predicted based on the latitudinal richness gradients seen in other Lepidoptera.

On a local scale, Lawton has shown a correlation between herbivore richness and host plant architecture. Trees have more architectural complexity and therefore more herbivores than shrubs, shrubs more than herbs, and so on. A poignant example of the importance of trees is provided by white oak (Quercus alba)—across its range in eastern North America it hosts more than 400 species of moths (this number includes both specialist and generalist feeders). Island faunas, especially those of distant islands, are always unbalanced with respect to those of the nearest continents. Macrolepidopterans that tend to be well represented on islands include the Arctiidae, Geometridae, Noctuidae (especially the Catocalinae, Heliothinae, and Noctuinae), and Sphingidae. Despite their small size, microlepidopterans are present even on remote islands, presumably because they disperse effectively as aerial plankton. Hawaii has been colonized by dozens of different microlepidopteran lineages, and all can be characterized as weak fliers. Interestingly the richness of all leaf-mining families (minute to small moths) on Santa Cruz Island, off the coast of California, is nearly that of the mainland, while most groups of larger moths and other insects are represented by less than half of the mainland fauna.

V. LIFE HISTORY

A. Egg

The ova are variable in shape and especially surface ornamentation. Microlepidopterans often have flat, smooth eggs, many of which are nearly transparent once laid on a leaf surface; those of macrolepidopterans tend to be raised. The latter may be round, square, disk shaped, or spindle shaped, the chrotion (outer surface) varies from smooth to richly ornate. Females generally deposit the egg within or on the surface of the host; in the latter case, secretions released from the accessory glands help bond the egg to the substrate. Females may leave behind abdominal scales, which discourage natural enemies or advertise to other females that the resource already has been found. A few polyphagous species (e.g., arctiids and hepialids) broadcast the eggs as they fly. The latter family also includes the world’s most fecund lepidopteran—a single female Abantiades hyalinatus laid 30,000 eggs (and died with another 20,000 ova remaining in her abdomen).

B. Larva

The larva (Figs. 4–9) is the principal feeding stage, increasing its mass by four to five orders of magnitude.
MOTHS

FIGURE 6  Phobetron pithecium (Limacodidae). Phobetron caterpillars are thought to mimic the cast “skins” of tarantulas, which contain thousands of decussate urticating hairs—Phobetron caterpillars do not sting and are thought to be completely harmless.

FIGURE 7  Automeris sp. (Saturniidae). Automeris caterpillars are furnished with a battery of urticating setae. The effect of which is not unlike that of stinging nettle.

FIGURE 8  Euproctis chrysorrhoea (Lymantriidae). Euproctis caterpillars have red middorsal glands on the sixth and seventh abdominal segments that yield a potpourri of defensive secretions—this caterpillar is responsible for at least two human deaths.

during its development. There are from three to more than a dozen larval instars, with five being the mode. Dyar was the first to note that there is commonly a linear increase of 1.3 to 1.4 in head dimensions between successive instars. Dyar’s rule holds across the Lepidoptera and often is employed to infer the number of larval instars when only incomplete life history data are available. In bombycoids and a few other groups the females may pass through an additional instar. Development is remarkably fast in some species, especially those that feed on ephemeral tissues such as fruits and new leaves. Larger moths of high latitudes may take 2 or 3 years to mature. In Gynaephora groenlandica—a lymantrid found well north of the Arctic Circle—caterpillars mature over a 14-year period.

Hypermetamorphosis—where two or more larval
Moths

forms occur, each specialized for different functions—as rare within the order. The best-known example occurs in the minute to small leaf-mining moths of the family Gracillariidae. In basal gracillariid genera, the first two or three instars are flat, legless, possess anteriorly directed mouthparts, lack a spinneret (silk-spinning organ) (Fig. 4b), and feed in a single cell layer, usually the epidermis. The diet is liquid, the larva filtering plant fluids from damaged cells. In later instars the larva takes on a rounded form that is legged, possesses a spinneret and ventrally directed mouthparts (Fig. 4c), and consumes whole cells and tissues. The spinneret is critical in that it is employed by the larva to lay down silk within the mine drawing it into a bubble, thereby creating a space where a rounded body and legs are appropriate. Metamorphosis in some Bucclatoidae, epipyropids, cyclocternids, opostegids, and doubtless others is also hypermetamorphic. Nonfeeding instars are rare, generally occurring in the first or last instars. An example of the former is provided by some zygaenoids. In a number of gracillariids (e.g., Cameraria and Chrysaster) there may be as many as two nonfeeding prepupal instars. Whereas a general loss of color is common among pre-pupae, some caterpillars turn almost scarlet prior to pupation (e.g., some Prododoidae, Gracillariidae, Elachistidae, and Notodontidae).

C. Pupa

In the most basal lineages the wings, legs, and antennae are weakly fused to the main body of the pupa (Fig. 10). In several of these families, the pupal mandibles, which may be impressively large (Fig. 10a), are used to cut through the cocoon or work the pharate adult free from the pupal crypt. In hetereoneurans the appendages are fused to the body and mandibles are very reduced and immobile (Fig. 10c). Another phylogenetic trend involves the degree of fusion between adjacent abdominal segments. In the most primitive moths the first seven abdominal segments are movable. Intermediate degrees of fusion occur among microlepidopterans. In the Obcctomerina, a clade which includes 17 of the most derived lepidopteran superfamilies (Fig. 3), the first four abdominal segments are fused and are immobile with respect to one another. Taxa with appreciable abdominal mobility often extrude the distal end of the pupa from the cocoon or pupal crypt prior to emergence (e.g., most microlepidopterans). Those with a high degree of fusion hatch within the cocoon or pupal crypt and must crawl free before expanding their wings (e.g., all Obtectomera). Male pupae have a single set of paired genital marks on the venter of the ninth abdominal segment; females have an additional paired set of genital ‘scars’ on the eighth segment as well (e.g., 10b). The pupal stage may be as short as 10 days or last several years, the latter is not unusual among species that inhabit arid environments.

Most moths spin a cocoon on or within the larval substrate or find their way to the ground and pupate in litter or drift. The cocoon may be a highly elaborate affair that takes days to spin. The most famous example is that of the Oriental Silk moth (Bombyx mori), the cocoon is fashioned from a single strand of silk that, when unwound, stretches for more than half a mile (0.8 km). A curiously elaborate cocoon is made by Marmara and related genera in the Gracillariidae: the cocoon is ornamented with up to 100 anally extruded bubbles, each of which is nearly the diameter of the caterpillar’s body. Cocoons represent a significant energetic investment in that silk is entirely protein, and the nitrogen that goes into its production is generally in short supply in plants. In addition to silk, the cocoon...
MOTHS


may incorporate excrement, scrapings from the larval substrate, or larval setae (especially if these are stiff, poisonous, or urticating). Preupal limacodids discharge a calcium oxalate solution over the inner surface of the cocoon that hardens the cocoon as it dries. Open, netlike cocoons are often spun by species with aposematic pupae. Taxa that pupate in concealed sites such as below ground or in wood often forego spinning a cocoon. A few moths (e.g., elachistids, pterophorids, sterrhine geometrides, and hedyliids) pupate naked, exposed on surfaces, like butterflies.

D. Diapause and Quiescence
Diapause may occur in any of the four life stages. The largest fraction of temperate moths pass through the winter months in diapause as prepupal larvae or pupae; the second largest fraction overwinter as eggs. Moths that overwinter as adults invariably need to feed over the five to seven months of winter—some of these are among the unwelcome guests in the buckets used to collect maple syrup. Reproductive diapause is the rule among moths that overwinter as adults, ovarian matura-
tion mating and being delayed until the arrival of spring. Summer diapause and aestivation occurs in many groups that live in habitats with a pronounced dry sea-
son. Enormous aggregations involving tens of thou-
sands of moths (usually noctuids) are known to assem-
ble in caves or on (hill) mountain tops. Moths in these aggregations become important sources of protein for vertebrates, even for grizzly bears. The diapause longevity record belongs to the desert-inhabiting prodoxids—single larval cohorts may yield adults over a period of 3 decades.

VI. FEEDING BIOLOGY
A. Larvae
Given the enormity of the order it is remarkable how little diversity there is in feeding habits. More than 99% of all moths are fundamentally herbivorous. Within this
realm, however, they have radiated to exploit virtually every corner of niche space. Terrestrial algae (including blue-greens), lichens, fungi, bryophytes, pteridophytes, and seed plants are eaten. All plant tissues and organs are consumed, but especially leaves, meristems, and reproductive structures. The majority require living tissues, sometimes of a very specific age, while others subsist on dead and fallen leaves, fruits, seeds, and wood. Nymphalid pyralids, which number more than 700 species, are aquatic as larvae, with some even adapted for life in rapidly flowing water. Other lineages have members that are subaquatic or bore into aquatic plants.

The most basal lepidopteran lineages have diets that include ancient plant lineages. Among the Micropterigidae, there are many species that include bryophytes in their diet. Larvae of agathiphagids—representing the order's second most ancient lineage—feed within the seeds of Agathis (Araucariaceae), a family that traces back to the Jurassic. Heterobathmiid larvae are leaf miners in new foliage of Nothofagus (Fagaceae), an archaic Gondwanan angiosperm. Basal families of the Glossata are strictly associated with angiosperms, hence the great species, generic, and even familial radiations within the Lepidoptera were reliant on the diversification of the angiosperms.

Few of the basal lineages account for appreciable species diversity. The first radiation of note, as measured by extant richness, was that of the Hepialidae, with 387 species worldwide. Their success may be due to their diet, which is arguably the most diverse of any moth family. Larvae eat bryophytes, pteridophytes, conifers, and hardwoods; leaf and especially woody root and stem tissues; leaf litter and fungi, especially in early instars. Many are cannibalistic, and some even attack and consume soft-bodied arthropods. Another early lineage with remarkably catholic diets are the teneid families, a high percentage of which are fungivores and detritivores. Members of the Tineidae (e.g., the clothes moth family) are noteworthy for their ability to metabolize keratin, the normally undigestible protein that makes up fur and feathers. Others feed in bird and bat guano, dung, and even the dead insects that accumulate below spider webs.

As the Lepidoptera radiated so did their host associations and feeding habits. The majority of microlepidopterans that followed are concealed feeders, tunneling into plant tissues, forming leaf mines or shelters, or feeding from within silken cases. Galls are made by some members of more than a dozen families of Lepidoptera, but none of the lineages can be regarded as particularly successful. Diets tend to be specialized in internal feeders because the host is not only the food but also the environment for some or all of the immature stages. The most intimate host associations occur in taxa where the adults and larvae utilize the same host plant, the most famous example being that of the yucca moth in the genera Tegritioida and Panagystisidae (Prodioideae). The female yucca moth uses special maxillary “tentacles” to collect yucca pollen and then flies to a second flower where she scrapes the pollen over the stigmatic surface. She then deposits one to a few eggs in the developing ovary, and in so doing, leaves behind a marking pheromone, which can be detected by other females; she then collects more pollen and flies to another flower. The larvae are seed predators but do not consume all the seeds in a fruit. The mutualism is obligate, as neither moth nor yucca can survive without the other.

Leaf miners, minute moths whose larvae feed between the two surfaces of a leaf, account for the most significant early radiations. The Nepticulidae, whose fossil history traces back more than 97 million years to the Cenomanian, contains more than 800 species with at least this number remaining to be described. The Gracillariidae, with 2000 recognized species and perhaps three times as many unrecognized, also traces its fossil history back to the Cenomanian. The latter family contains members that are among the first shelter-forming Lepidoptera. Early instar gracillarines mine in leaves as do other gracillarids, but middle instars may exit the mine to form a leaf shelter in which they complete their development.

The bulk of the microlepidopterans, including the three largest superfamilies (the Gelechioidea, Tortricoidea, and Pyraloidea) are concealed feeders, mostly leaf rollers or shelter forming, but others are borers, leaf miners, gall formers, and detritivores. A handful are predators, inquilines in galls or nests of social insects, and so on. Leaf rollers and shelter forming exhibit intermediate levels of host plant specialization, but polyphagy is not common. An obvious exception to this generalization is provided by many tortricids, especially members of the nominate subfamily. Several microlepidopteran families have representatives that are external feeders for one or more instars (e.g., the Bucculatricidae, Schreckensteinidae, Periphoridae, and others), but none accounts for appreciable present-day diversity. The largest group of externally feeding microlepidopterans are the Zygaenidae, which as a group appear to be well protected—their arsenal includes urticating spines, abundant, long, or dense setae and hairs, and cyanoglucosides (Figs. 5 and 6).

Approximately 60% of the described Lepidoptera are...
mots

macrolepidopterans, which are principally external feeders. Host plant associations are diverse—some macrolepidopterans are exceptionally specialized while others are widely polyphagous. Polyphagy is far more common among external feeders and is the mode for many macrolepidopteran families (e.g., Lasiocampidae, Saturniidae, Ennomine Geometridae, Lymantriidae, and several subfamilies of the Noctuidae). It is worth noting that many species that are host plant generalists across their geographic range specialize on just one or a few plants at any one locality. Regardless of the mode of feeding, internal or external, caterpillars that feed on highly poisonous plants tend to be host plant specialists. Fungivory appears throughout the order but is especially common in the Hepialidae and Tineidae. The line between fungivory and feeding on fallen plant material is often an arbitrary one. Not surprisingly, both of these groups are also detritivores. Other sizable lepidopteran clades, with members specializing on fallen leaves, decaying wood, and other nonliving plant tissues, include the herminiine Noctuidae, Blastobasidae, Oecophoridae, and a few pyralid subfamilies. In dry oak forests of the eastern United States, herminiines may account for more than half of all macrolepidopterans in light trap collections during the month of July. Australia has an exceptionally rich fauna of litter-feeding Oecophoridae and Tortricidae. There, numbers are so great in eucalyptus forests that the larvae are believed to play a significant role in litter processing. The sloth moths of the neotropics are an oddity. Adults of three chrysaugine pyralid genera are found on or about sloths, awaiting the once-weekly release of dung, which serves as the substrate for the developing larvae.

The most common type of carnivory in the order is cannibalism, often showing up in taxa where the host substrate is limited (e.g., annual plants and architecturally reduced perennials) or where movement between hosts is either impossible or exposes larvae to considerable risks (e.g., internal borers). Essentially phytophagous species that occasionally eat other caterpillars occur sporadically across the order. The Euphytopriadeae and Cychlotosomatidae are obligate ectoparasites of homopterans. The latter is only parasitic in the first instar, as the second to final instars are brood predators in ant nests. Only about 200 lepidopterans are regarded as predacious, many of which are butterflies (see De Vries this volume). Nearly all predatory caterpillars feed on sessile or nearly sessile prey, such as the brood of hymenopterans or aphids, scale insects, and related homopterans. Hawaii has spawned the most exceptional predaceous caterpillars—members of the genus Eupithecia (Geometridae) are sit-and-wait predators that snatch small insects that are so unwise as to crawl over the caterpillar’s body.

B. Adults

Micropterigids are mandibulate and use their jaws to grind up pollen and fern spores. A crude proboscis, formed from the maxillae in adult Dacnonypha, is used to collect fluids from bark and other surfaces where the adults are perching. These fluids may include dissolved honeydew, but adults do not visit flowers for nectar. Nectar feeding is first seen in the Incurvariaidae and is retained by virtually all the larger lineages of moths that follow. In addition to nectar and other sources of dissolved sugars, moths use the proboscis to imbibe water, liquids from injured plants and tree flows, dissolved salts (including human perspiration and lacrimal secretions), pus, animal excreta, the crushed bodies of other insects, rotting fruit, dung, carrion, many other substrates. A few anomalous noctuids use the proboscis to pierce fruits, related species have even taken to blood feeding on large mammals. In both cases the tongue is short, stout, and the two sides of the proboscis are able to slide back and forth against each other as the tongue is worked into the fruit or flesh.

A favorite collecting technique for moths is “sugaring.” A standard bait is made by mixing fruit, beer, and sugar. While overripe bananas seem to be the most universal ingredient, collectors sometimes add watermelon, apricot, grapes, rum, jellies, molasses, and a rather long list of other substances, some of which might best be left unnamed. Once the mixture has been allowed to ferment it is then applied to tree trunks or offered on sponges. Virtually all of the winter-active moths (e.g., Oecophoridae, Tortricidae, and Noctuidae) feed at bait. During summer months, many moths in dry woodlands also come to bait. Presumably, moths that are attracted to bait normally would be feeding at tree wounds, sap flows, rotting fruit, or the honey dew secreted by homopterans. Plant tissues tend to be low in sodium relative to the metabolic needs of animals. Many lepidopterans reach adulthood with inadequate supplies of this essential element, but they are able to acquire it by imbibing fluids from mud puddles and water courses, urine, sweat, lacrimal secretions, and so on. Although “puddling” is well known among butterflies, it also occurs in a diverse array of diurnal and nocturnal moths. In most moth and butterfly species, 90 to 100% of the individuals that puddle (or feed at other sodium-rich substrates) are males. A remarkable case is provided by Gephispa septentrionis (Notodontidae). Puddling males...
European gypsy moth (*Lymantria dispar*). The introduction of a fungal pathogen was considered unsuccessful until 1989 when *Entomophaga maimaiga* (†Entomophthorales). This fungal pathogen was introduced into North America from Japan in 1910 and 1911 as a biological control agent to stem the ever-increasing populations of the gypsy moth. Adults of late fall and early spring geometrids tend to be non-feeding, whereas those that fly during the summer commonly consume nectar.

VII. NATURAL ENEMIES AND CHEMICAL DEFENSE

A. Natural Enemies

Moths are fecund invertebrates that produce 100 to 10,000 times more ova than will ultimately survive. A panoply of biotic and abiotic mortality factors come into play during every life stage to limit population growth. Biotic factors include an enormous range of pathogens, parasites, and predators. Population numbers of many pest species, especially in forests and other stable communities, are often controlled by viruses, bacteria, and fungi. Viruses are often highly specific, and one wonders if there are not nearly as many lepidopteran viruses as there are moths and butterflies. There are comparatively few recorded species of bacteria and fungi that attack insects, but the taxonomy of insect pathogens, especially those that cannot be easily cultured, is grossly understudied; while some are highly specific, others are thought to have broad host ranges. Bacillus thuringiensis, the most widely used pathogen for the control of Lepidoptera in gardens, agriculture, and forests, has a host range that includes hundreds of species across many superfamilies. Fungal pathogens, mostly in the Entomophthorales and Fungi Imperfecti, often require special conditions such as a period of high humidity before appreciable infection will occur. A species of special note is *Entomophaga maimaiga* which is used to control the gypsy moth. Adults of late fall and early spring geometrids tend to be non-feeding, whereas those that fly during the summer commonly consume nectar.

Another 20 or so families of insect parasitoids attack Lepidoptera, but two of which (Diptera: Bombyliidae and Sarcophagidae) are parasitic Hymenoptera, mostly in the Chalcidoidea and Proctotrupoidea. All four life stages are exploited by parasitoids, although most are specialized to attack the larva. Some of the smallest insects, trichogrammatid and mymarid hymenopterans, with body lengths of only 0.2 mm, are egg parasites. The great majority of insect parasitoids are specialists that utilize but one or a few related species. Large, long-lived caterpillars support rich guilds of parasitoids: across eastern North America, 29 fly and wasp parasitoids are recorded from *Hyalophora cecropia*, at least 3 of which are believed to be specialists on this saturniid and its congeners. Such caterpillars may be at the base of an entire food web, as evidently there are six hymenopteran hyperparasitoids that parasitize the parasitoids of *Hyalophora*, and one wasp hyperparasitoid that principally attacks the hyperparasitoids. Askew and others have noted that leaf miners and gall-forming caterpillars have an exceptionally rich parasitoid fauna. *Phyllonorycter apparella* (Gracillariidae) has a wingspan of only 7 mm, yet larvae and pupae of this leaf miner are known to host 20 species of hymenopteran parasitoids and hyperparasitoids.

Invertebrate predators of major importance include mites, spiders, predaceous stink bugs, beetles, robber flies, ants, and wasps. Many of the physical, chemical, and behavioral defenses of caterpillars are known to be effective in thwarting the pillages of ants. No doubt, these defenses are of special importance in tropical ecosystems where ants make up much of the insect biomass. Mammals, birds, reptiles, and amphibians are the principal vertebrate predators. Birds are such an important selection pressure for externally feeding caterpillars that there is a growing consensus that they have shaped not only what many caterpillars look like, but also what, when, and how they feed. Their influence on adult color patterns and behavior also seems to be preeminent. Bats, too, have been an exceptionally important evolutionary force in the diversification of the order. Hearing organs, sensitive to the frequencies emit-
MOTHS

263

ted by bats, have evolved independently at least five times (discussed later); within one of these groups, the Arctiidae, some chemically protected taxa have evolved the ability to signal back to bats. Upon detecting ultrasound frequencies, these arctiids begin clicking, presumably advertising their whereabouts (and toxicity).

B. Defensive Chemistry

The defense strategies of caterpillars and adults are legion and make for fascinating reading and study (see Stamp and Casey, 1993, and Scoble, 1992, for larvae and adults, respectively). Chemical defense against parasitoids and predators likely has played a prominent role in the diversification of the order. The ecological and evolutionary constraints facing unpalatable species differ widely from those faced by palatable taxa. For example, chemically protected caterpillars are often brightly colored, feed day or night, and rarely take measures to conceal their feeding damage; many are gregarious. Conversely, palatable species are apt to be cryptic in color, to be nocturnal, to move slowly, and to only rarely leave conspicuous signs of feeding. Across the order there is a spectrum of palatability, ranging from those species that can be eaten in large numbers (by birds or indigenous peoples) to those where contact or ingestion can be fatal. There is an enormous literature on this subject—this discussion and examples presented here are only meant to be illustrative.

Stinging setae, hollow and filled with poison, occur widely in the Limacodidae, Megalopygidae, and Hemileucinae (Saturniidae). In rare instances, human fatality has been reported after encounters with the latter two. The body setae of Euproctis chrysorrhoea (Lymantriidae) (Fig. 8) contain histamine, protease, lipase, and phospholipase, and are thought to be responsible for at least two human deaths. Gypsy moth caterpillars and other lymantriids drag the body setae through the dorsal abdominal glands, ‘arming’ them with a histamine and a diverse mixture of noxious substances that cause itching and other allergic responses, which can be severe in some individuals. Other taxa with large numbers of species with urticating setae include the Lasiocampidae, Thaumetopoeinae (Notodontidae), and Arctiidae.

In some taxa, caterpillars and adults synthesize noxious compounds from simple dietary precursors. Zygaenid larvae produce cyanoglucosides from two basic amino acids. Chemical protection in many Notodontidae also appears to be based on simple manufactured compounds (e.g., the late instars of many Heterocampi- nae eject formic acid and ketones from a cervical gland if molested). Other moths sequester toxins or precursors from their (poisonous) host plants; toxin concentrations in the hemolymph may greatly exceed those that occur in any of the plant tissues on which the caterpillar has fed. Examples include Euphorbia-feeding caterpillars, which retain dietary terpenes, and the milkweed caterpillar (Euchaetes egle) (Arctiidae), which, like the monarch butterfly caterpillar, sequesters high concentrations of a cardenolide. Other arctiids actively seek out and consume host plants high in pyrrolizidine alkaloids (PAs). Experimental studies have demonstrated the effectiveness of PAs in repelling both vertebrate and invertebrate predators as well as their importance in courtship, mating, and paternal investment (discussed later). If bright coloration can be used as a proxy for chemical protection, there is still a universe of protected moth species whose natural chemistry remains unstudied and unappreciated.

In many species the chemicals sequestered or manufactured by the caterpillars are retained by the adult. PAs collected by male Utetheisa bella (Arctiidae) caterpillars are used to synthesize the male courtship pheromone. Eisner and his students found that female Utetheisa (Fig. 11) raised on diets devoid of PAs are readily consumed by spiders, but females that have mated with males laden with PAs are cut out and dropped from webs. The eggs and even young caterpillars, resulting from the latter type of pairing are also chemically protected by the alkaloids collected by the father. Arctiids are not completely dependent on the larval diet because adults of some species are able to locate and ingest

FIGURE 11 Tiger moths (Arctiidae) (also see Figs. 12 and 13). Utetheisa bella; in addition to sperm, the male transfers alkaloids during mating that not only protects his mate subsequently, but also her offspring.
dissolved PAs. They will imbibe fluids from the surfaces or wounds of PA-containing plants (e.g., Asteraceae, Boraginaceae, and Fabaceae) or even the crushed bodies of alkaloid-rich caterpillars or moths. Many arctiids bubble out a defensive secretion from the prothorax upon disturbance. In *Grammia*, the fluid is mostly clear and green and may represent nothing more than hemolymph (Fig. 12); in others it is more frothy, sometimes milky, and loaded with acetylcholine, histamine, and possibly pyrazines. Adult female lymantriids collect larval hairs from the cocoon and incorporate these into their egg mass. Not all adults are able to retain the protection enjoyed by their larvae (e.g., adult saturniids and limacodids are generally palatable). Both caterpillars and adults of chemically protected species tend to be aposematic (Fig. 13). Adults are often diurnal and may be appreciably slower fliers than their unprotected sister lineage. Not a small number are members of Mullerian mimicry complexes.

VIII. DISPERSAL AND MIGRATION

Long-distance movements occur in many groups, not an insignificant number of which are pest species. One of the most common patterns is for populations to cycle through the winter in tropical and subtropical latitudes, then to move north (or south) during the spring and summer. Some migrants move on the leading edge of storm fronts, where both high wind and high humidity facilitate long-distance dispersal. Occasionally their numbers may be so great that they show up on radar screens. Swallows, swifts, and other insectivorous birds follow these fronts as well, gleaning insects as they go. A second common pattern is for moths to move between wet and dry forests, with adults trickling out of the latter over the course of the dry season then reinvading with the return of the rainy season. The dazzlingly beautiful sunset moths (*Uraniidae*) are famous for their mass diurnal migrations in Africa and Latin America.

IX. KEY ADAPTATIONS

One can only speculate as to why there are so many moths. One morphological feature that stands out as universal is the presence of scales over the wings and body. Those covering the wings are easily abraded, so much so that the wings feel slippery. Regardless of the context in which scales evolved, there can be little doubt that they provide considerable protection from natural enemies: adults can fly into a spider web and, if they are lucky, they will fly out, leaving behind a hundred or a thousand scales still attached to the web. Scales allow the wings to slide past one another and often out of the bill or jaws of a would-be predator.

The collapsible siphon or proboscis has allowed Lepidoptera to exploit liquid diets. More than any other feature, this structure has been tied intricately to the coevolutionary history of moths with flowers. Nearly all insect-pollinated flowers with deep corolla tubes (or nectar spurs) are pollinated by Lepidoptera. Charles Darwin knew this, and upon being told of a Madagascar flower with a 23 cm corolla, predicted that a sphingid hawkmoth would one day be found on the island with a tongue of at least this length—and it was. The tongue also allows the collection of dietary constituents that are scarce in plants, such as nitrogen and sodium (noted earlier). No nonfeeding moths are known to migrate,
suggesting that the presence of a tongue is a critical attribute for long-distance movement. Similarly all temperate moths that overwinter as an adult have a tongue.

A tympanum or analogous hearing structure, sensitive to the ultrasonic frequencies used by hunting bats, has evolved independently in no less than five moth lineages (Scoble, 1992, and Fig. 12). It may be no coincidence that three of these (the Geometroidea, Noctuoidea, and Pyraloidea) are among the most species-rich moth clades. Fenton and Fullard estimated that 85% of all macrolepidopterans possess some type of hearing organ. Noctuids that detect ultrasound first attempt to fly away from the source, but if the hunting frequencies are sensed at close range, these moths either drop to the ground or initiate a highly erratic flight while spiraling downward. Further testimony as to the importance of the ear is provided by the mites that infest the tympanal cavities of certain noctuids (especially Leucania) (Fig. 14). Treat found that mites placed onto Leucania always move into one ear, leaving the other fully operational.

Silk has played a major role in the evolutionary success of the order. Its most universal use is in the formation of the cocoon or in the attachment of the chrysalis to the pupation substrate. Most microlepidoptera use silk to form a shelter or to line a chamber within the larval feeding substrate. Many caterpillars lay down silk whenever they walk, which presumably allows them to quickly retrace their step, for example, in times of danger. Case making, where the caterpillar constructs a portable silken case, has evolved in no less than a dozen moth groups; two especially successful lineages include the Psychidae and Coleophoridae (both families are appreciably more species rich than their non-case-making sister taxon). Some gregarious caterpillars spin conspicuous nests. Not only do these nests offer considerable protection from natural enemies, but they can be important in thermoregulation, heating up well above ambient temperatures on cool days. Many species, especially among those taxa with flightless females (e.g., some Psychidae, Geometridae, and Lymniidae), disperse by "ballooning." Early instars drop from silken lines and wait for winds to blow them about. Among microlepidopterans, it is common to see prepupal larvae dropping from trees on silken lines on their way into leaf litter where pupation will occur. At night many geometrid caterpillars drop from their perches, suspended on a silken thread, and dangle for hours before climbing back onto foliage, always before daylight. In a similar fashion many caterpillars, if disturbed, drop from their perch on a silk cable and crawl back up after the danger has evidently passed. Processionary caterpillars (Notodontidae: Thaumantopinae) and tent worms (Lasiocampidae) lay down a silk trail, impregnated with pheromone; that siblings follow. Some microlepidoptera (e.g., Bucculatricidae) form a molting cocoon within which the vulnerable transition between instars takes place. Prior to molting, many species lay down a light sheet of silk into which the crochets are engaged; in so doing the new caterpillar can crawl more easily out of its old integument.

One behavioral trait that stands out is the ability of Lepidoptera to exploit the hours of darkness. In many ecosystems moths are the most diverse and numerically important group of nocturnal flying insects. Perhaps 90% of all Lepidoptera are night active. Associated with their nocturnal habits is the order's remarkable elaboration of morphological, neurophysiological, and behavioral traits that facilitate pheromonal communication (discussed later).

Their success may also just be a matter of being in the right place at the right time. Because the order was essentially phytophagous and beginning to diversify on nonseed plants and gymnosperms prior to the origin of angiosperms, it was well poised to ride along with the explosive radiation of flowering plants that was to follow. Indeed, the moths are the largest monophyletic lineage to have evolved with plants.

X. PHEROMONES

While visual communication is thought to be of modest importance in nocturnal moths, chemical communication is highly developed and perhaps universally important. The first animal pheromone was identified from the Chinese Silk Moth (Bombyx mori) and was appropri-
ately dubbed bombykol. In nearly all nocturnal moths, virgin females release a pheromone that attracts males from distances of up to several hundred meters (reports of up to a mile are likely exaggerated) (Fig. 15). Males move upwind, in and out of the pheromone plume, until they arrive at its source. The male antenna and associated neurophysiology make up one of the most sensitive chemical detection systems known in nature. Not surprisingly, males often have more elaborate antennae than females, sometimes with secondary or tertiary branches that increase the surface area several fold (e.g., many bombycoids, Fig. 16). One interesting exception, and one that remains unexplained, are Artenophanes females (Tineoidea)—the female’s antenna is considerably more complex than that of the male.

Female sex pheromones usually are released from glands on the abdomen, especially the intersegmental regions between abdominal segments seven and eight or eight and nine. The complete pheromone contains an isomeric “cocktail” of two to several components, usually with one or two major components that are responsible for most of the male attraction. The individual components of the pheromone are often odorless (at the concentrations produced by moths), straight chain hydrocarbon backbone, with one or more double bonds, and an active moiety (e.g., an acetate, aldehyde, epoxide, ketone, or alcohol). Closely related species tend to have different (enantiomeric or isomeric) blends, minor components, or “calling” times.

Once the two sexes are in proximity, male courtship pheromones may come into play. The structures by which males deploy their pheromones are as varied as any structures in the order. In the simplest case they are individual scales peppered over the wings or other parts of the body. Male sex scales (or androconia) that are grouped into tufts, brushes, or “pencils” occur over almost any part of the body including the wings, legs, mouthparts, and antennae. They are especially common about the genitalia, where they provide final signals prior to coupling (Fig. 17). Remarkably complex scent organs occur in a number of arctiids, hepialids, and noctuids (Fig. 18). A hallmark of androconia is that their abundance and development may differ markedly among closely related species, which suggests that they are under strong sexual selection.

Male pheromones are thought to serve in species recognition and isolation and to promote female acceptance. In general they are used at close range in the vicinity of the female. Role reversal, with males calling females from long ranges, is very rare but does occur (e.g., see Fig. 18). Chemically, male pheromones are exceptionally heterogeneous, ranging from simple molecules to long-chain hydrocarbons and ring compounds. Many have a detectable and often pleasant odor. The male pheromone blend of the oriental fruit moth contains both (2)-methyl epijasmonate and (E)-
FIGURE 17. Male scent-releasing structures (also see Fig. 18). *En-chephalia* (Noctuidae); end view of abdominal brush that encircles the male genitalia. Courtesy of Ken Haines.

ethyl cinnamate; the former is the active constituent in the odor of jasmine, a common ingredient in women’s perfumes. Many male scents are based on secondary plant compounds that are sequestered by the larva or collected by the adult. In some arctiids, the female is reluctant to copulate with males that lack appreciable quantities of pyrrolizidine alkaloids (PAs) on their androconia.

Beyond mate location, courtship, and mating, pheromones are also known to serve in a number of lesser roles. Some female moths (e.g., yucca moths) leave behind a “marking” pheromone at the time of oviposition that alerts other females that a resource already has been located. The aggregation behavior displayed by some arctiids and noctuids is likely mediated by a pheromonal signal. During mating, some male Lepidoptera transfer an “anti-aphrodisiac” pheromone to the female that, at least temporarily, repels other would-be suitors.

XI. COLORATION, DIURNALITY, NOCTURNALITY, AND ATTRACTION TO LIGHT

As a general rule, moths that are active principally at night are subtly colored in earth tones, especially browns, grays, black, and white. The patterning in nocturnal taxa is often variable, presumably because details of the coloration are under relaxed selection pressure. But exceptions to both of these statements abound. Most notably, there are dozens of lineages with brightly colored hindwings that are covered by the forewings—the bright coloration being exposed as a flash or warning signal after a moth has been disturbed. Bright coloration is associated with moths that are diurnal or both diurnal and nocturnal. Aposmotic and highly reflective (e.g., blue, silver, and gold) coloration are associated with diurnal taxa.

Assuming nocturnality as the ground plan for the order, or at least the Glossata, it is evident that diurnality has evolved repeatedly, perhaps hundreds of times within the order. The most conspicuous example is that of the butterflies, which are, from an evolutionary sense, nothing more than a group (or two) of diurnal moths. Taking the Geometridae as an example, there are dozens of species and genera that are diurnal, yet there are essentially no species-rich clades (genera, tribes, subfamilies) within the family that are active principally during the day. Diurnality is common among moths that live in alpine or arctic habitats or that fly during the coldest months of the year.

From a phylogenetic perspective, diurnality is well distributed across the order. Members of the most basal family, the Micropterigidae, are both diurnal and nocturnal in habit, with some genera exhibiting splendidly.
Moths with metallic golds and reflective purples. These same bright colors are found in a host of other primitive families (e.g., Heterobathmiidae, Eriocraniidae, and Incurvariidae). Like in the Micropterigidae, adults may be active both day and night. Across the Microlepidoptera diurnality is widespread; essentially diurnal lineages include the Adelidae, Choreutidae, Heliodinidae, Heliozelidae, Scythrididae, and Sesiidae. Many moths are crepuscular, some flying at both dawn and dusk and others limiting their activity to one of these two periods of rapidly changing light intensities.

Nearly all nocturnal moths can be collected at light. Indeed, light traps are routinely employed to conduct moth inventories, to monitor populations of agricultural pests, and to collect gravid females for life history studies. The most commonly asked question about moths is “Why are they attracted to light?” Although it is difficult to answer this question with certainty, the moon-compass explanation has great appeal. If a moth had a need to fly in a straight line at night it could do so by maintaining a constant angle to any celestial light source. When this same behavior (constant orientation to a distant light source) is applied when a light is proximate, a moth will spiral into a nearby light source (Fig. 19). Perhaps the greater mystery is why only a fraction of the moths present in a given area are attracted. Anyone who runs flight interception traps, collects at bait, or knows what caterpillars are feeding in the garden appreciates the fact that the individuals attracted to light represent but a small fraction of the whole. Sex ratios in light-collected samples are often heavily biased toward males. Hepialids in the genus Sthenopis provide a counterexample as females predominate in museum collections. There are many moths known from but one sex—the sex that is attracted to light.

Moths are not entirely dependent on celestial navigation. There is ample evidence that moths also possess a geomagnetic compass that allows them to move in a directed fashion, even on moonless nights and under overcast skies.

XII. IMPORTANCE

Lepidoptera, in particular moths, are among the most important forest and agricultural pests. There is almost no plant tissue that is not eaten by one or more species of moths. Most of the forest pests are of cyclical occurrence with natural enemies eventually controlling population numbers. Introduced species offer an obvious exception to this generalization—their populations may go unchecked for many years before natural enemies can be introduced or recruited from the native fauna. Agricultural pests tend to be more chronic and, as noted earlier, many are migrants preadapted to be crop pests (because they are highly dispersive insects that locate resources quickly). Moreover, crop pests often have short life cycles that allow them to complete one or more generations before natural enemies can build to appreciable numbers. A few, such as the Indian Meal Moth (Plodia interpunctella) (Pyralidae), are stored product pests that attack stored grains, cereals, bird seed, dry dog food, and even candy that has sat too long in the pantry. Clothes moths (Tinea species) (Tineidae) are well known for their spoils of woolens. Most moths should be regarded as beneficial principally for their role in terrestrial food webs, especially those that include birds, bats, and rodents. Many songbird nestlings are dependent almost entirely on caterpi-
lars. With the exception of large, bat-pollinated flowers, most white-flowered plants are pollinated by moths under cover of night. Scents from moth pollinated flowers may be exceptionally fragrant (e.g., honeysuckle, jasmine, and some orchids). Sphingidae are the primary pollinators for about 10% of the tree species in Costa Rica’s seasonal dry forests. These large, strong fliers are especially important in tropical forests where individuals of the same species can be widely spaced. Many moth caterpillars feed on leaf litter, decaying wood, and fungi and thus play at least a minor role in litter decomposition and nutrient cycling. The most famous product derived from a moth is silk. This fiber, with a tensile strength greater than that of steel, is still made the same way it has been for more than 4600 years by boiling the cocoons of Bombyx mori and unraveling them a single thread at a time. Moths have been used successfully as biological control agents of invasive weeds. Australia had lost 50 million acres of pastureland to introduced prickly pear cactus (Opuntia) before the pyralid, Cactoblastis cactorum, was introduced from southern South America. Within a decade of its release, most large stands of the cactus had been utterly destroyed. The moth also has been effective in controlling invasive Opuntia in Hawaii, Mauritius, South Africa and the West Indies. This story recently has taken an odd and unfortunate twist, as this same moth has found its way into the Florida Keys, where it now threatens stands of three rare species of Opuntia.

XIII. CONSERVATION

Across much of Europe and the United States lepidopteran faunas are being increasingly employed as bioindicators by state, provincial, and federal agencies and nongovernmental conservation organizations to evaluate natural areas and management regimes. Moths offer many advantages as bioindicators: many macrolepidopterans are easily identified and their habitat requirements are reasonably well known. By any standard they are hyperdiverse and therefore provide considerable data relative to the amount of effort that goes into their sampling and sorting. Part of their value as bioindicators is that they represent another trophic level—two grasslands that are botanically equivalent would have different moth faunas if one were being managed too intensively (e.g., by fire). One major disadvantage of moths as bioindicators is that samples often include many transients. For example, more than one-third of the species of moths that occur in Denmark have been collected in a light trap atop the Zoological Museum on the University of Copenhagen campus at the periphery of the city.

The principal threat to moths is the same as for other terrestrial invertebrates—habitat destruction and fragmentation—and nowhere is this more true than for tropical forests where moth diversity is thought to be highest. In other ecosystems the obverse, reforestation and succession, pose threats to moths. Fire suppression in grasslands, chaparral, barrens, scrub forests, and other fire-maintained ecosystems has been detrimental to many local invertebrate populations. And the reverse, ill-timed and too aggressive prescribed burning, can be detrimental. There has been much discussion about the impacts of artificial lighting on moth populations, much of it focusing on the decline of saturniids and other large moths. Up to this point the published commentary has been anecdotal, but there can be little doubt that street lamps and other high-intensity lamps have had some impact on moth populations—within days, bats learn to forage at them by night, and they are among the first places that birds visit at dawn.

Introduced or alien species can be a threat at many levels. Invasive plants can displace native hosts upon which either the larvae or adults of native moths depend. In exceptional cases, introduced plants can provide strong ovipositional cues for species whose larvae have either fail entirely or experience high levels of mortality. Because introduced moths often arrive without natural enemies, their numbers can cause widespread overexploitation of host plants. It is hard to imagine that the widespread defoliation by the gypsy moth (Lymantria dispar) has not been responsible for at least local impacts. Top-down threats from the release of introduced predators, pathogens, and parasites for biological control efforts merit special attention. The tachnid, Compsilura cocinata, was introduced from Europe into eastern North America to control the gypsy moth. The parasitoid’s host range includes more than 200 native species. The moth has one generation a year, the fly three to four. It has been estimated that a single acre of a gypsy moth infested forest will yield up to 10,000 adult Compsilura in June, which then must seek out alternate (native) hosts.

Although DDT and other pesticides have been implicated in the local extirpation of Lepidoptera, the evidence is invariably circumstantial. While suppression activities (single applications) would rarely be expected to result in local extinctions, eradication efforts (multiple applications, especially in a single season) certainly impact many moths and the insectivorous animals that are dependent on them. The most widely used biological pesticide to control Lepidoptera is Bacillus thuringiensis...
var. kurstaki (Btk)—a bacterium that either kills a caterpillar outright or leads to a systematic infection that results in death. Btk is used by home owners to control garden pests, by farmers in agricultural systems, and by government and industry to combat forest defoliators. Product labels indicate that Btk is effective against a taxonomic array of Lepidoptera (and Coleoptera). Susceptibility to Btk ranges from taxa that are not affected to those that are killed weeks after a single application; saturniids and papilionids (swallowtail butterflies) are among the groups reported to be especially susceptible.

From one perspective Btk is a great control agent because it affects few nonlepidopterans and has no mammalian or avian toxicity. Yet from a lepidopterists’ perspective it is hard to view Btk in such a positive light. The use of Btk and other pesticides in natural ecosystems is most likely to be a problem over critical habitats that are isolated, restricted, or in some other way imperiled, especially when these occur within a matrix that contains a primary pest. For example, the glades, balds, or barrens communities that occur in an otherwise forested landscape are apt to suffer from the aerial pesticide applications targeted for forest pests.

With the exception of the Chinese silk moth, which no longer exists in the wild, there is no evidence that collectors have ever represented a significant risk to moths. Collecting is likely to represent a serious threat only to highly imperiled and geographically restricted taxa. Where population sizes are very small, the use of light traps that employ a killing agent should be avoided. A far greater threat is undercollecting, as it seems probable that less than half of the world’s moth fauna has been described. And even in those regions where most of the basic taxonomic work has been completed, it is not until we have reliable occurrence data on a broad geographic scale that conservation biologists will know with certainty what species are slipping toward extinction.

See Also the Following Articles

- BUTTERFLIES
- INSECTS, OVERVIEW
- INVERTEBRATES, TERRESTRIAL, OVERVIEW
- PARASITOIDS
- PESTICIDES, USE AND EFFECTS OF

Bibliography


MUSEUMS AND INSTITUTIONS, ROLE OF

Paul Henderson and Neil Chalmers

The Natural History Museum

I. Collections
II. Research and Biodiversity
III. Exhibitions
IV. Education and Training
V. Publications and Outreach
VI. Advisory

GLOSSARY

bioprospecting The search for commercially valuable biochemical and genetic resources in plants, animals, and microorganisms.
descriptive taxonomy The description of new species usually involving written accounts, with illustrations, of the characteristics of a specimen.
parataxonomist A layperson who has received training in practical basic biology, ecology, and taxonomy, as well as in the collection and preparation of biological specimens, so that he or she can undertake a specific part of a biodiversity inventory.
revisionary taxonomy The reevaluation of entire groups of organisms based on new and old evidence.

THE COLLECTIONS HOUSED in natural history museums and similar institutions are essential to research, education, and a better appreciation of the natural world and its diversity. Since it is estimated that only approximately 13% of the species alive today have been described and named, taxonomic and related research is still essential if we are to understand, conserve, and manage biodiversity properly. Furthermore, increasing awareness of the value of biodiversity has led some countries to establish institutions with the express aim of making inventories of their biodiversity and to become the agents for the information that arises from those surveys. The research and collections of museums are applied to a wide range of issues, including human and animal health, bioprospecting, environmental quality, resource management, and the implementation of legislation. Museums are also active contributors to a wide range of educational objectives, including those incorporating biodiversity. This is achieved through exhibitions, training, field studies, and outreach programs. New technology, such as the Internet, is enabling museums to order and make readily available the vast amounts of information on biodiversity they already hold.

The role of museums and related institutions is to conserve and develop the collections in their charge; to improve public understanding through exhibitions and other means; to undertake research involving the collections; and to make their information, expertise, and objects as widely accessible as possible.

I. COLLECTIONS

Collections are the raison d'être of natural history museums and similar institutions. They form the prime basis
for the museums’ research, education, exhibition, and advisory work. For the larger national and equivalent museums, the collections of animals, plants, fossils, and sometimes rocks and minerals have been assembled over more than two centuries. They are therefore a major source of information on biodiversity in both space and time. They also reflect our knowledge of the natural world and something of the processes that we have used—past and present—to describe and understand it.

A. Content

It is estimated that the world’s 6500 natural history museums contain approximately 3 billion specimens covering the animal, plant, and mineral kingdoms, excluding microorganisms. Add to these the untold numbers of specimens (living and dead) in the numerous botanical gardens, arboreta, zoos, and institutions with culture collections and it is clear that the resource is huge. To this must also be added the collections of books, monographs, research papers, and artwork amassed over many years by the libraries of these institutions. They form an invaluable part of the collections and represent, in part, the developing knowledge and perceptions of biodiversity. The particular nature of the collections in each museum or institution will depend on its function and a host of other factors. Two European examples (Paris and London) showing the broad makeup of their collections are given in Fig. 1.

In most museums the collections are maintained by curators or collection managers whose task is to ensure that the specimens are well conserved and available for users. Collection management and conservation is a skilled job often requiring considerable patience and a sound taxonomic knowledge of the organisms. As interest in biodiversity increases, so does the use of related collections, and this brings associated threats principally to the specimens themselves. Conservation practice must therefore be proactive so as to allow increased access to the material while developing improvement in its care. Conservation is thus seen as the employment of best practice to prevent or arrest the long-term physical deterioration of natural history specimens and associated artifacts and documents to preserve their scientific and cultural worth.

The methods used in conservation are the subject of global interest and conferences on the topic are now being held; the first major one was the International Symposium and First World Congress on the Preservation and Conservation of Natural History Collections, which was held in Madrid in 1993. A significant literature now exists. One topic of particular interest is the preservation of an organism’s DNA in museum specimens. The evidence suggests that most current practices are not particularly deleterious but additional research is needed.

The collections of any institution are a manifestation of its research, collecting, and acquisition policies possibly over a considerable period of time. The British Museum, established in 1753 (and as such is one of the earliest museums to house natural history and other objects), developed from the initial collections of Sir Hans Sloane who had acquired specimens from many parts of the world. The natural history part of this collection grew by donation, purchase, and active collection by its staff over the succeeding decades to the extent that accommodating it became a major problem. As a result, the Natural History Museum was opened in 1853 in South Kensington. The collections continued to grow and have become one of the world’s most renowned, along with those in Washington, DC, Paris, New York, and Berlin. This long history of collection development involving major periods of exploration in different parts of the world—on land and at sea—has led to this museum having specimens from numerous countries and waters, collected over a wide time span. It is perhaps no surprise to find numerous specimens of extinct species as well as those that have been crucial to the development of ideas on evolution (such as the finches collected by Darwin from the Galapagos Islands) and other concepts. Other museums have acquired their own identity through complementary histories leading to distinct collections. No two museums have very similar holdings of specimens.

Culture collections are of increasing importance and require special storage and conservation approaches. Such collections are usually housed in institutions specifically established for this purpose. An example is the American Type Culture Collection (ATCC) comprising mainly cultures of bacteria and fungi. Others may be very specific, such as the Provasoli–Guillard National Center for the Culture of Marine Phytoplankton, also in the United States. A principal function of these centers is to distribute cultures to the research community. The ATCC is also working to set up local culture collections in developing countries.

B. The Uses and Role of Collections

1. Taxonomic Studies

For most specimens in the collections, the question “what exactly is it?” usually has to be answered. The
initial information usually sought is that which is intrin-
sic to the specimen itself, i.e., properties such as shape
(or morphology), color, molecular characteristics (e.g.,
DNA), and sometimes its behavior. With such informa-
tion the specimen can be named and placed within a
classification. This process is the key to virtually all the
other uses of the collections: Without a proper basic
knowledge of the material, other studies and applica-
tions would probably be worthless.

For accurate taxonomic nomenclature "type" speci-
mens are essential. When formally naming a species,
it is normal practice to select a particular preserved
specimen (the type) to act as a "name bearer." It is
given a unique and permanent name. If, subsequently,
any doubt arises about how the given name should be
used, then identification of the type specimen will show
which name belongs to which recognized species. As
the taxonomic system changes because of new discover-
ies, reexamination of type specimens is often essential
for accurate use of names, and this need may arise
repeatedly over decades and centuries. Since type speci-
mens are very important, they are usually deposited in
museums or other public institutions. It follows that
there is a major obligation for such institutions to con-
serve them and make them available for study by the
international scientific community.

Systematists, by examining the intrinsic properties
of different species, work to establish the relationships
between organisms. Multiple specimen samples can
provide information on variation within species, on
geographical distribution, and on changes over time. If
other relevant information is available, such as the
habitat from which the specimen was collected, then
other findings are clearly possible. The collections are
therefore curated and developed with these needs in
mind.

2. Biological and Related Research
Collections are an essential resource for a range of bio-
logical research activities, including evolution, genetics,
ecology, and epidemiology. Some of this research is
carried out in museums and related institutions (see
Section II), but much is also done elsewhere. The muse-
ums, while sometimes providing the material, usually
act primarily as a resource for verifying the status of organisms on which research has been done. Without this backing, the findings can lack credibility or repeatability.

Bioprospecting can only be undertaken on a sound basis if the species are accurately identified. Previously collected museum specimens are perhaps rarely used directly in such work, especially when it must be destructive, but the collections and the taxonomists’ expertise are often vital for verification and for planning field-collecting programs. Moreover, cladistic and other predictive classifications, usually based in whole or part on comparison of museum specimens, can provide helpful guidance for selecting taxa for bioprospecting. They help in searching for taxa that are likely to have similar properties to those of valuable species already discovered.

3. Environmental Conditions

Specimens collected at different times and in different places can be of significant value in documenting changes in environmental conditions, including those resulting from human activity. Examples include documenting the presence of DDT in different areas and times, documenting the presence of radioactivity from weapons testing or environmental pollution from radioactive waste, and assessing the spread of invasive species.

4. Inventories and Conservation

Management of biodiversity, in almost any sense of the term, requires knowledge of the species (or at least the genera) in the region of interest. For example, conservation work requires an inventory of those organisms that are of relevance. It is sometimes argued with good reason that a requirement to conserving biodiversity is to discover, identify (and possibly describe), and inventory the species in a region or locality. The museums can and do provide the resources for such work, but these are far too little for them to have a sufficient impact on their own. It is thus necessary for others (often local people) to be trained and given the tools for the task. For this, a local or regional collection of what are often called “voucher specimens” may be established. These are correctly identified specimens that serve as a reference for the fieldworkers in identifying the specimens that they have collected. National and international publications (such as the International Union for the Conservation of Nature’s Red Lists of Threatened Animals and Plants) giving information on endangered species are also based on information stemming initially from collections.

Since collections have often been developed over extended periods of time (some for two or more centuries), it follows that some of the information associated with a specimen, such as date and place of collecting, is very useful for establishing past distributions and their changes. Unfortunately, although many collections have these data, a significant amount of work may be necessary to extract the data from labels or registers.

5. Implementation of Legislation

Many countries have legislation in place to help protect against pests and diseases, whether they originate from or via organisms that are domestic or imported. Legislation prevents the growing or rearing of certain species. Some countries do not allow the export of certain endemic or other species for a variety of reasons, including the concern that others might benefit commercially from them in some way but without any benefit accruing to the country of origin. Implementing this kind of legislation requires people, such as customs officers, to either have or be able to call on the appropriate expertise or identification facility. The tools for such jobs (e.g., guides and charts) are based on the work and collections of museums and other institutions. Some cases are relatively straightforward. One well-known case involved efforts to prevent the introduction of the Colorado potato beetle into the United Kingdom. Posters showing this bright and distinctive beetle and warning of its potential impact were displayed at points of entry into the United Kingdom for many years. Most other cases, however, are much more complex and require regular contact between customs and other authorities and museums scientists or similar specialists.

Some international treaties and conventions require the resources of collections and associated expertise for them to be enforced or implemented. The Convention on International Trade in Endangered Species (CITES) is one example which involves the collaboration and cooperative working of museums, botanic gardens, etc.

6. Resource Management

The numerous related commercial activities of farming, cultivation, breeding, and fishing involve the use of collections through the need for pest identification or even occasionally for the identification of the specimens collected (e.g., fishing). The exploitation of resources through activities such as mining or oil exploration and extraction entail impact studies in which the affected land or water is monitored before and after for its biodiversity. Correct identification of organisms is essential for understanding these impacts and any programs for possible remedial work and restitution.
II. RESEARCH AND BIODIVERSITY

A core research activity for natural history museums, in addition to other institutions such as botanical gardens, zoos, and aquaria where these have research staff, is the pursuit of systematics and taxonomy. This research is effectively aimed at the fundamental units of biodiversity, namely, species, by describing and naming them and understanding their relationships. Much of this research refines what we know and also charts the unknown. Phylogenetic studies, in addition to giving important insight into evolution and its processes, help to provide a sound and strongly predictive framework for inferring unknown properties of taxonomic groups.

The systematic research carried out in museums is done at different levels. The important base level is descriptive taxonomy (sometimes called “alpha-taxonomy”), which involves the description of new species and is heavily reliant on their morphological characteristics. Revisory taxonomy is a higher level involving studies on the relationships of species and it can involve the establishment of phylogenies. Other levels of systematic study can be concerned with life cycles, ecology, and genetic aspects.

It is variously estimated that we have named and described only approximately 10–15% of the species alive today. The huge task of describing the rest falls for the most part to taxonomists in museums and related institutions. Such institutions, however, are expected to increase the breadth of their research to embrace ecology, conservation, and theoretical studies. These pressures arise because of the demand for information on a wide range of environmental and resource issues at a time when there is a much greater extinction rate of species (recently estimated to be approximately 10,000 times greater) relative to the normal rate over geological time. The research is quite labor-intensive and requires relevant skills, many of which are acquired through years of experience. Modern methodologies and technologies, through approaches such as molecular biology and electronic pattern recognition, are helpful, but the task and its importance remain enormous. It is thus becoming increasingly important to establish priorities for applying the research resource.

Different institutions have responded in different ways to this imperative. Several museums are organizing their research around broad themes or programs that address significant current and often global issues. By focusing different types of expertise onto specific and timely problems, greater impact can be achieved while still providing some of the basic information for more generic uses. Some authors advocate that priority should be given to groups of organisms that are ecologically and economically important (Raven et al., 1998), including those that we know relatively little about such as bacteria, fungi, and nematodes. However, despite international recognition of the need for good priority setting, there has been little discussion or agreement on how this should be done, primarily because of the current shortage of taxonomists and financial and other constraints placed on the relevant institutions.

Some institutions, however, do have a very specific research agenda, especially those concerned with charting the biodiversity of their own country. Indeed, a few institutions have been set up with this express purpose, as private or government agencies charged with the management of biodiversity information. Principal examples include the Environmental Resources Information Network in Australia (http://haas/ervs.gov.au/pein/erv/index.html), the National Commission for the Knowledge and Use of Biodiversity (La Comisión Nacional Para El Conocimiento Y Uso de la Biodiversidad) in Mexico (http://www.conabio.gob.mx/), and the National Biodiversity Institute (Instituto Nacional de Biodiversidad, INBio) in Costa Rica.

INBio is different from the other examples in that it holds collections and has wider functions. Its mission is to promote a new awareness of the value of biodiversity and thereby achieve its conservation and use to improve the quality of life. Its functions involve making inventories of biodiversity, including those concerned with bioprospecting, and administering and disseminating biodiversity information. It works in collaboration with other research centers in Costa Rica to obtain corroboration of inventory information and to help conduct chemical processing in prospecting. It was established in 1989 and has since developed a significant staff ranging from researchers to trained parataxonomists. Its four principal objectives are

• To assume responsibility for developing and executing a national biodiversity inventory
• Where feasible, to locate national collections within one physical space and under one administration
• To centralize biodiversity information now and in the future
• To put this information into easily accessible and user-friendly formats available to a wide variety of users and to promote its use by Costa Rican society

Further information is available from its web site (http://www.inbio.ac.cr/). Research in the form of inventories can provide sig-
MUSEUMS AND INSTITUTIONS, ROLE OF

Significant benefits in addition to those accruing from bioprospecting and the development of new drugs. These include a better knowledge of pests, invasive species, and disease vectors and the control of all of these. These in turn lead to a more sustainable agriculture, better resource management, and improvements in human health.

The type of inventory can vary significantly. Commonly, this will involve collecting examples of all members of a particular set of organisms, such as plants, in a particular area using the expertise of both trained taxonomists and parataxonomists. Perhaps more rarely they will be an All Taxa Biological Inventory in which most, if not all, organisms in an area are identified, together with information on ecological relationships. However, inventories in any form usually require considerable resources, especially manpower, if they are to achieve a significant degree of usefulness. In this regard, this activity involves long-term research goals.

Research on systematic biology carried out in museums also has other uses:

1. In medicine through studies involving disease vectors and pathogens
2. In the pharmaceutical industry, especially through bioprospecting (see Section I.B.2)
3. In fisheries and agriculture, including studies of pests and pathogens
4. In environmental quality assessments involving water and air pollution and the effects of food production and human and animal health
5. In appreciation of our environment, including raising awareness and knowledge of biodiversity through good-quality guides and keys

The research agenda sometimes includes work directed toward conservation of biodiversity by modeling and manipulating data on the distribution of species. The aim is to give assessments of species richness (i.e., the number of species in a given area), species turnover, habitat occurrences, and other parameters that are useful in planning programs with strategic conservation objectives.

III. EXHIBITIONS

For modern museums of natural history and equivalent institutions, biodiversity provides a new and integrating focus to their exhibition program. In many institutions, there is a move away from displays on particular taxonomic groups toward exhibitions encompassing ecology, environmental change, resource exploitation, conservation, and related topics. Concepts are being emphasized, which is an approach that helps to aid education and understanding. Issues are also being addressed, such as the strongly deleterious effects of certain human activities on the environment and on biodiversity. Increasingly, members of the visiting public are being given the opportunity to find answers to their specific questions and to see how they might help in conserving biodiversity and to understand some of the implications of its loss.

The approach to some exhibitions on biodiversity is to demonstrate the nature and variety of the animal and plant kingdoms. These tend to be based on what we know. Few museums deal with the problems of our limited knowledge of living species and habitats, and how we are endeavoring to overcome them, despite the recognition by the systematics profession and others that both our understanding and our fundamental knowledge base are relatively poor.

The scale and remit of a museum well, of course, determine the broad nature of the exhibits. Local and regional museums tend to show the biodiversity of their area and address particular educational needs for their community. They may be able to stage activities, including field days or courses, on particular aspects. These can provide an important educational and social contribution and help raise awareness of the scope and processes of local natural history and some of the associated threats and opportunities. Larger museums are less likely to focus on particular areas, although some describe the biodiversity, in selected ways, of their country or large area.

Even allowing for the fact that museums are less likely to emphasize taxonomic groupings, the approaches to exhibitions can be very different from one museum to another. Most attempt to engage visitors and will strive to meet their needs. The Grande Galerie de l’Evolution (part of the Museum National d’Histoire Naturelle) in Paris achieves this by the scale and overall visual design of the displays. Here, for example, on the topic of the African savanna is a memorable display of mammals parading as if in some type of Serengeti migration. (For such displays, the skill of the taxidermist is still paramount.) The displays, covering approximately 6000 m², emphasize the diversity of life in different marine and terrestrial environments as well as cover the principles of evolution and the role of mankind.

In contrast, the Life Galleries at The Natural History Museum in London have several themes with a focus on particular processes or concepts, whereas others are
devoted to particular animals and their habitats. A gallery on the subject of ecology impresses because of its architectural novelty, its use of modern methods of communication, and its integration of different topics, including the role of mankind. Another gallery deals with the principles of evolution. Others cover particular animal groups, such as the arthropods (in a gallery titled "Creepy Crawlies"), marine invertebrates, and mammals. In these galleries, attention is paid not only to highlighting descriptions of some of the representative organisms but also to habitats and behavior.

A related approach is adopted in the Zoological Museum of Copenhagen, Denmark (founded more than 350 years ago), at which a major display titled "From Pole to Pole" deals with animals in specific habitats, including the tundra, the Antarctic, the "bird cliff," the temperate forests, the desert, and the rain forest. Another display on animal life in the oceans covers subjects such as whales, the photic zone, deep-sea animals, and animal behavior. Much of the rest of the permanent exhibitions is devoted to animal life in Denmark. The overall aim is to show animals in their natural habitats and to educate about biodiversity in an integrating and aesthetically pleasing way, especially through the use of large modern dioramas. School parties comprise approximately 30–40% of the visits.

The American Museum of Natural History in New York was probably one of the first, if not the first, to mount a major exhibition specifically on biodiversity (titled "Hall of Biodiversity"). This permanent exhibition of approximately 1200 m² was opened in 1998 with the aim of alerting the public to the biodiversity crisis and its implications and also to show the beauty and abundance of life on Earth. There is significant use of new technology to aid communication, such as multiscreen video installations and use of high-resolution imagery. An electronic "BioBulletin" presents a regularly updated account of events that affect biodiversity and the threats it is currently experiencing. A large diorama (occupying approximately 23% of the total space) of a portion of rain forest of the Central African Republic emphasizes the interaction of animals and humans with the environment. By showing the rain forest in three different states—pristine, altered by natural forces, and degraded by humans—it is possible for the visitor to grasp readily the broad impact of current forces on forest ecology. The diversity of life is covered by a display of approximately 1500 specimens and models mounted along one side of the Hall. The visitor is able to learn more about the organisms, their distribution, their habitats, and ecological processes by using interactive computer stations. Finally, a so-called "Solutions Wall" enables visitors to learn how mankind can help to manage and protect biodiversity.

These four examples (necessarily highly selective) of the approaches adopted by some modern museums exemplify the importance attached to biodiversity and the issues surrounding it. The visiting public is treated to a visually interesting and informative experience of direct relevance to the modern world. The experience is essentially unique to museums and closely related institutions. As such, they are providing an important role in educating and influencing a wide range of people.

"Traveling exhibitions" form part of the program of some larger museums. The aims are to enable a greater audience to experience a particular exhibition and to recoup some of the production costs. They may be shown in several countries. The topics can also be diverse, although biodiversity is often featured in one form or another. Two recent examples are the "Ocean Planet" organized by the Smithsonian Institute (Washington, DC), which includes coverage of the plant and animal life of the oceans, and "Voyages of Discovery" produced by The Natural History Museum (London), which tells of several major voyages, from Hans Sloane's to Jamaica (1687–1689) to that of the oceangoing Challenger (1872–1876). These voyages were responsible for many major discoveries about the earth's biodiversity.

Rental, transport, and installation costs, however, can be a significant deterrent to an extensive use of traveling exhibitions. The costs may have to be met by an admission charge or through sponsorship. Exhibitions about biodiversity and related exhibitions therefore compete with a mass of other attractions for people's attention.

Finally, electronic media allow for the development and transmission of the "virtual museum" in which images of specimens with associated information can be remotely accessed, sometimes in an interactive way. Museums are experiencing a revolution in the ways that they display and disperse their information, representations of their holdings, and their skills.

IV. EDUCATION AND TRAINING

The educational role described previously extends in different ways once the greater functions of museums are considered. Many museums provide a special service for visiting school groups, such as activity programs,
MUSEUMS AND INSTITUTIONS, ROLE OF

Schemes to help teachers guide children around the exhibitions so as to learn more from them, and special talks and demonstrations, sometimes involving live organisms. Local schools often find these services invaluable for teaching more about biodiversity and helping the students gain a genuine and long-term interest.

A community approach can be particularly beneficial by bringing together the aims of educating people about biodiversity while simultaneously achieving a useful product or new research on it. The Chicago Wilderness Project involves individuals from the community together with the Field Museum of Natural History and other U.S. organizations and NGOs to document and restore the wild areas in and around the city of Chicago. All are learning about, and contributing to, the environment through an interactive process.

Since biodiversity is a complex concept, it is recognized that careful thought should be given to the educational role. What should be the aims? In what way is the museum trying to give the visitor a greater awareness or new skills relating to biodiversity and other topics? One recent example of analysis of this issue is provided by the work of the Council for Environmental Education (1997), which is part of the University of Reading, England. The council researched, with government backing, site-based biodiversity education provision. Museums were included in its site reviews. The council’s findings suggested that biodiversity education should enable people to:

- Understand what biodiversity means
- Understand that biodiversity is a dynamic concept; that species, habitats, and ecosystems are part of a balanced system that changes naturally over time
- Become more aware of biodiversity as part of their cultural and spiritual as well as economic heritage
- Be more aware of, informed about, and understand the significance of biodiversity around them and define their own level of interaction with it
- Recognize the relationship between biodiversity and the maintenance and quality of life
- Know that factors influence biodiversity and understand that human activity can both damage and enhance biodiversity
- Be aware of the impact of their own and others’ actions on biodiversity, including lifestyle and consumer choices
- Improve their own skills in relation to biodiversity, including those skills that enhance understanding and promote appropriate action
- Be aware of what actions they can take to preserve and enhance biodiversity and act on that awareness

Museums do not, as a matter of course, address all these issues in their educational programs, but in practice several programs are being implemented in some of the larger institutions. The research helped to formulate and bring together much of what had been recognized in other ways.

The teaching of systematics and the training of systematists have been in decline during the past 20–30 years in many parts of the world. The decline has occurred in both the schools and the universities at a time when the need for advances in the field is increasing so as to meet the challenges of biodiversity conservation, of environmental protection, and of sustainable development. The issue is compounded by the fact that there is an estimated decrease in the number of systematists throughout the world. Furthermore, the age distribution of those remaining is skewed toward older age groups.

Museums are attempting to counter these trends by providing courses and training in systematics and in related fields of direct relevance to biodiversity. They work with universities in postgraduate research training and also in 1-year “masters” courses to teach taxonomy. The scale of these operations is still limited and is insufficient to meet the current and predicted future needs.

Many larger museums and gardens provide training and workshops to meet specific needs, often including programs related to biodiversity. These inevitably change with time as demand changes, but these institutions are providing an important service, whether it be for training in biodiversity surveys and inventories or training in handling biological informatics.

V. PUBLICATIONS AND OUTREACH

Museums and related institutions are increasingly the providers of biodiversity information and interpreters of that information. The forms of the information are numerous and range from monographs on particular genera or families of organisms to popular books for a wide readership. Interactive keys and guides through the medium of the CD-ROM have also gained prominence in recent years. They can be a fast and user-friendly way of identifying a specimen within a particular group. The CD-ROM has also been used for educational packages on insects, birds, and other wildlife. Development costs, however, can be high, and with the growth of the Internet as an effective competitor, the
CD-ROM is seen as having only a limited applicability in the marketplace. The Internet provides a major challenge for collections-based institutions to distribute their relevant information to a wide range of potential users. As discussed in Section I, the information contained in the collections and in other related databases is of important use in many areas. The challenge in making that information available is primarily through the substantial resource demands (in staff, software, and hardware) not only to establish the large-scale databases but also to maintain them. It is also still very early to be able to assess realistically what the user requirements are likely to be and how these will change with time.

Because of the scale of the operation and the need for many institutions to play a role in these kinds of information dissemination so as to gain much greater benefits, some planned approaches are international in scope. Data sets held by different institutions and maintained by them will form part of a global network. Developing and other countries will be able to access information, held in developed countries, about their own biodiversity. In this sense, information is being "repatriated" to those who need it most.

Currently, the World Wide Web gives any user a considerable degree of access to information that was difficult to access only a few years ago. Much remains to be done, but many museums and institutions have their own web pages with associated databases and other packages. Inevitably, the quality of these pages varies markedly from one institution to another, primarily because we are at a relatively early stage of development and not all institutions have the necessary resources to take on these additional tasks. In 1999, a link list of "natural history museums" gave the web addresses of several institutions (including universities) in the Americas, Africa, Asia, Europe, Australasia, and Oceania: http://www.lib.washington.edu/sla/natmus.html.

VI. ADVISORY

Museums have been contributing to broader issues in addition to the pursuit of systematics, the curation of collections, and the provision of exhibitions. They contribute through the provision of advice and consulting services on biodiversity to a wide range of users. This is achieved primarily in three broad ways: consulting work, aiding governments in policy formulation, and helping to implement regional and global initiatives or policies.

A. Consulting

In an age when there is increasing awareness of and activity regarding conservation, environmental quality, and sustainable development, the role of natural history museums in servicing the associated demands is becoming more important. Museums, with sufficient resources, can provide the most fundamental of data in almost all environmental studies or developments, including information on the species present and their behavior as well as on the changes that an activity or process can bring to the local biodiversity system. Assessments of the impact of any activity, such as mining, deforestation, or urban development, can be undertaken. The information will also be used to provide predictions for work in other areas. Planning authorities will be one of the prime clients of such a service, as will other government agencies and industrial companies. In some countries (e.g., Australia), it is a requirement that proper assessment of the biodiversity of an area is made, together with assessment of the likely impact, before a development can be given approval to proceed. Museums are often the only institutions that can provide much of the necessary information and therefore are increasingly being called on to undertake an advisory role with related research. Other consulting work done by museums includes waste management, chemical analysis, resource exploration, health and veterinary studies, and collections management. Some are also helping others to develop their public exhibitions, including those on biodiversity.

B. Aiding Governments on Policy

Governments call on museums because of their special skills, to advise them on many policy issues and their implementation. Implementation of the CITES is a good example because museums are key to sound identifications that can carry legal weight.

In the realm of biodiversity, museums have been used to help governments create their country's action plans under the Convention on Biological Diversity (CBD). Several museums have staff that serve on their official national delegation to the Conferences of the Parties of the CBD. Similarly, they may serve on the delegation to the CBD's Subsidiary Body on Scientific, Technical and Technological Advice. This body advises on the implementation of the convention. Governments recognize that museums have a special role to play which complements those of the other institutional and government members. This is particularly important
since taxonomy and systematics are currently topics on the agendas of the CBD.

The expertise of museum staff is being used in many ways by countries to implement their obligations under the CBD. Several are undertaking biodiversity inventories or assessments. In aiding this process, guides on procedures and organisms have been developed. One particularly relevant example is a guide to good practice in biodiversity assessment (Jermy et al., 1995), the purpose of which is to help those who have the responsibility to survey and assess the biodiversity of their own country."

C. Regional and International Initiatives

The CBD operates at both national and international levels. Biodiversity information, through its sheer scale, also operates in this way. Natural history institutions are partners or contributors to several regional and international initiatives, especially those related to biodiversity. The European Union is currently supporting some European museums and botanic gardens in planning and subsequently developing an integrated collections databasing facility at the specimen level. Governments are also discussing how a major international program bringing together many aspects of biodiversity information might be established.

See Also the Following Articles

BIOPROSPECTING • CONSERVATION BIOLOGY, DISCIPLINE OF • CONSERVATION MOVEMENT, HISTORICAL • EDUCATION AND BIODIVERSITY • GOVERNMENT LEGISLATION AND REGULATION • TAXONOMY, METHODS OF • ZOOS AND ZOOLOGICAL PARKS

Bibliography


MUTUALISM, EVOLUTION OF

Egbert Giles Leigh, Jr.
Smithsonian Tropical Research Institute

I. Why Is Mutualism So Important and So Puzzling?

Natural selection is a competitive process. What keeps competition from annihilating the possibility of mutualistic cooperation? In this article, I outline conditions that allow mutualisms to evolve, describe various devices and circumstances that prevent one partner from becoming parasitic or overly exploitative of the others, and discuss why mutualism is so important in evolution.

GLOSSARY

ethology The study of the behavior and social relations of animals.

genetic terms A locus is a particular place on a chromosome (strictly speaking, an “address” on a chromosomal map) where a gene resides. This gene may be any one of several alternative types or alleles. A diploid organism contains two genes at each autosomal locus—one from its mother and one from its father. A haploid organism contains one gene at each locus. A genome is the set of all an organism’s genes.

mutualism A mutually beneficial association or interaction, temporary or permanent, among organisms of the same or different species.

parasitism An association among members of different species in which members of one species (the parasite) live inside or on the resources of a body of the other.

symbiosis A mutualism among members of different species in which members of one species (the symbiont) live inside or on a body of the other.

1. WHY IS MUTUALISM SO IMPORTANT AND SO PUZZLING?

Natural selection is a competitive process. The winners are those genes—and the individuals carrying them—which reproduce most successfully (Fisher, 1999). However, many animals live in groups because they depend in some way on each other’s activities. The partners in an animal’s group are likely to be its closest competitors—for food, mates, or living space. Life in groups thus poses two questions: What are the advantages of living in a group? and What factors prevent competition among group members from undermining their cooperation, thereby destroying the benefits of living together? Indeed, how the mutualism of social behavior could possibly evolve is a central problem of ethology. Solving this problem means understanding (i) the potential benefits of group life (Maynard Smith
and Szathmáry, 1995), (ii) the social mechanisms and ecological circumstances which minimize or prevent destructive competition among the partners in a group (Moynihan, 1998), and (iii) how such social mechanisms could evolve (Leigh and Rowell, 1995).

The central problem of ethology has a political analog. Human beings benefit from, and indeed depend on, social life. Nevertheless, some individuals can siphon off for their private advantage part of the common wealth created by social activity. A standard problem of political philosophy from the time of Aristotle onward is how to organize a society so that individual advantage coincides most nearly with the common good.

The “ethological” problem recurs at other levels of biology. Especially in the tropics, natural ecosystems are more luxuriant, productive, and diverse than the sterile grasslands which so often replace them when the land is abused (Leigh, 1999). Indeed, the human influence on natural ecosystems is called “disturbance” and considered damaging, as if ecosystems were organized for functions which include the maintenance of productivity and diversity—organization which a random alteration of the system would disrupt. This (rarely stated) assumption parallels Aristotle’s argument that organisms are adapted because visibly mutant individuals are usually less functional than their normal counterparts. Indeed, despite the abundance of natural enemies (predators, parasites, and pathogens), organisms depend on their ecosystems for the necessities of life, even the air they breathe. Therefore, organisms share a common interest in their ecosystem’s integrity, as if ecosystems were true “commonwealths.” Does the common interest of organisms in the integrity of their eco-system influence the evolution of individual species? If so, how? These are among the greatest mysteries of biology (Leigh, 1999).

Genes are the “ultimate” units whose self-interest drives evolution in the sense that no characteristic evolves nonrandomly unless it serves the self-interest of some gene (Dawkins, 1982). However, an individual gene, divorced from the rest of its genome and an organism appropriate for the genome’s expression, is useless as a piece of computer program without the rest of the program, the right computer for running the program, and an operator who can run the program on the computer. Thus, these ultimate units of self-interest are utterly dependent on each other. Chromosomes express the mutual advantage that different kinds of genes derive from each other’s presence—their mutual dependence on each other’s functions. In particular, chromosomes maintain their genes’ continued association by ensuring their simultaneous replication (Maynard Smith and Szathmáry, 1995). However, some types of genes, which do nothing to benefit the organisms that carry them, multiply independently of the chromosomes, threatening to fill their genome with selfish DNA. How is this sort of destructive competition prevented from draining away the common wealth of the genome?

Conflicts can also occur between chromosomal genes and their genome. Meiosis, the process whereby a gamete is assigned one gene at each locus, is usually as fair a lottery as nature can devise, in the sense that at each locus a gamete has equal chance of inheriting its grandmother’s or grandfather’s gene. When meiosis is fair, natural selection favors an allele only if it enhances the survival or reproduction of the organisms carrying it. On the other hand, a few alleles which injure their bearers spread by biasing meiosis of heterozygotes in their own favor. Such alleles are called segregation distorters. Some experimental populations have been wiped out by the spread of such alleles. Why is meiosis normally so fair when segregation distortion can spread alleles so effectively?

Eukaryotic cells are cells whose chromosomes are separated from the cytoplasm by a nuclear membrane. Nearly all eukaryotic cells contain organelles, such as mitochondria and chloroplasts. Organelles are essential to their host cells. Cells require mitochondria to obtain energy by oxidizing carbohydrates to carbon dioxide and water. Leaves cannot photosynthesize unless their cells contain chloroplasts. Organelles contain DNA and can reproduce with some help from nuclear genes. Their DNA reveals that organelles descend from free-living bacteria which invaded, or were ingested by, ancestors of their host cells more than a billion years ago (Margulis, 1993). Although organelles appear to be integral structures of their host cells, conflicts can arise between organelles and their hosts. Organelles are usually passed on by the mother. Natural selection thus favors organelles which can survive inside the female, or all-female, sex ratios among the young of their host organisms, although selection usually favors hosts with equal numbers of young of each sex. Moreover, zygotes (fertilized eggs) which receive organelles from both parents may be impaired by the struggle for dominance between organelles from different parents. How did the symbiosis between cells and their organelles evolve? Why does harmony usually prevail between cells and their organelles?

Metazoans are complex multicellular animals. A metazoan’s cells usually serve their organism well. However, cancer suggests that conflict between an individual and one of its cell lines is a real, often devastating,
possibility (Buss, 1987). How did the harmony among a metazoan's cells evolve? How are conflicts between metazoans and their cells reduced or suppressed?

II. THE EVOLUTION AND MAINTENANCE OF MUTUALISM

A mutualism evolves only if the potential partners all benefit by cooperating. In other words, there must be advantages which are most easily or effectively obtained by cooperating, and all partners must benefit from cooperating, even if unequally. Cooperating must truly serve a common good.

A. What Can Be Gained by Cooperating?

Among members of the same species, the most basic form of cooperation is sexual reproduction, by which two individuals jointly produce offspring more varied than those that either could produce alone (Maynard Smith and Szathmary, 1995). This cooperation is extended in many species, including insectivorous birds and some fish, in which the parents share the labor or divide the tasks involved in feeding, sheltering, and defending the young.

Members of the same species initially join in larger groups, as a rule, if grouping enhances safety or effectiveness (Moynihan, 1998). They may combine for greater safety against predators because, with more eyes, predators are detected sooner or because members of a group can coordinate activities to confuse or repel predators. A group may also be able to overcome a competitor, or a large prey animal, which could defeat any one of its members. Among social insects, however, advantages of a suitable division of labor soon influenced the evolution of social life.

Mutualisms among different species usually involve complementation of different functions. Thus, a coral provides its symbiotic algae with nutrients from small animals that the coral polyps catch. In return, the coral receives carbohydrates from the photosynthesis of these algae. A plant provides its pollinators with nectar or pollen in hopes that these mobile animals will convey some of this pollen to another plant of the same species. Some plants provide food and shelter for specific kinds of ants, which repel most of their host plant's herbivores and destroy the growing tips of vines that would otherwise overgrow their host plant.

B. What Keeps Mutualisms from Becoming Parasitisms?

Cooperation involves the pooling of labor to create a common good. What keeps a group member from exploiting the good without helping to create it? Why don’t some group members cheat their partners in mutualism?

1. Kin Selection

If members of a group are related to each other, and much less closely related to members of competing groups, an individual propagates copies of its own genes by helping fellow members of its group reproduce (Holldobler and Wilson, 1990). A child inherits half its mother's genes. Thus, either mother or child can propagate copies of its own genes by helping the other reproduce.

Kin selection plays a crucial role in the evolution and maintenance of insect societies. A wasp lays an egg where the hatching larva has immediate access to food. Often, this is a provisioned nest cell. Sometimes, solitary nesting is futile because unguarded nests are robbed of their eggs or provisions. Communal nesting makes for better defended nests but intensifies competition for nest cells and provisions, creating winners and losers. If winners and losers are related, it may be more profitable (as measured by the numbers of copies of the loser's genes propagated to offspring) for the loser to help the winner reproduce, perhaps by provisioning the winner's nest while the winner guards it. Relatedness to the winner enables the loser to benefit from this mutualism, but their common interest in a defended nest is equally essential to this mutualism.

In more complex insect societies, a single queen produces all the young (Fisher, 1999), and most of her daughters are workers who help the queen reproduce. Workers share a common interest in helping their queen if it is impossible, or at least much less effective, for them to reproduce on their own. Among ants, workers are either sterile or lay eggs so slowly that they do better to help their queen (Holldobler and Wilson, 1990). A queen honeybee attains this end differently. She mates with many males and mixes their sperm thoroughly. Thus, most of a worker's colleagues are half-sisters. Because a worker is more closely related to her mother's egg than to a half-sister's egg, workers eat eggs laid by half-sisters, rendering worker reproduction futile (Seeley, 1995). The queen thereby creates a circumstance in which mutual policing among workers enforces their common interest in helping their queen. This community of interest among the workers makes
among groups on a characteristic is the proportionate
increase in the number of groups founded per preex-
isting group per group lifetime conferred by its mem-
bers’ possession of this characteristic, just as the inten-
sity of selection among individuals on this characteristic
is the proportionate increase that this characteristic
confers on an individual’s lifetime reproductive success,
relative to fellow group members. Selection among indi-
viduals balances an equally intense selection among
groups on the same characteristic if (i) each group
descends from a single parent group, and one migrant
is exchanged per two groups per group lifetime, or (ii)
no migrants are exchanged but one of every $N$ groups
is founded by the joining of colonies from two parent
groups, where $N$ is the number of mature individuals
per group (Leigh, 1999). Moreover, selection among
groups is unlikely to override strong within-group se-
lection unless there are many more groups than indi-
viduals per group. In sum, selection among groups
is decisive only if groups are definite individuals in
their own right.

Selection among groups plays a crucial role in main-
taining mutualism between organelles and their host
cells. Cells do not exchange organelles. Each cell re-
cieves its organelles from a single preexisting cell. This
is true even of zygotes. Nearly all kinds of eukaryotic
organisms have evolved means to ensure uniparental
transmission of organelles. Organelles are therefore
subject to a stringent selection among groups (each
group being the organelles of given type in one cell)
in their host’s interest. In other words, organelar repro-
duction depends utterly on the reproductive success
of their hosts.

How did organelles become susceptible to selection
among their hosts? Consider mitochondria. The mito-
ochondrial ancestors which first invaded host cells were
probably parasites (Margulis, 1993). In some parasites,
selection favors increased dispersal from one host to
another, although this injures their current host’s wel-
fare. For others, selection favors “caring for” their cur-
rent host even if this reduces the effectiveness of dis-
persal. It depends on how greatly increasing dispersal
impairs the current host’s usefulness, which differs from
case to case. The complementation of functions between
a host cell which procures food and an ancestral mito-
ochondrion which metabolizes food far more effectively
than the host can do so presumably favored symbionts
which “cared” for their current hosts. As a house cat
defends its new owner’s house against conspecifics, so
ancestral mitochondria defended their valuable hosts
against conspecific invaders. Thus, migration of mito-
ochondria between cells was prevented, thereby sub-
jecting the mitochondria to selection among host cells.
Selection on hosts favored mechanisms ensuring uniparental transmission of organelles when their hosts reproduced sexually, thereby perfecting the mechanism maintaining mutualism between host cells and their organelles (Leigh, 1999). Selection among groups also played an integral role in the evolution of metazoans. All metazoans descend from sexually reproducing ancestors, whose young began life as a fertilized egg or zygote. A zygote grows by dividing mitotically so that an animal's cells all have the same genotype. Each individual is therefore genetically unique, but (mutations excepted) an individual's cells are genetically identical. Moreover, selection favors individuals which can distinguish "self" from "nonself" and prevent invasion by unrelated cells, thereby preventing the exchange of migrant cells among individuals. These circumstances identify the reproductive interests of an animal even more closely with those of each of its cells (Maynard Smith and Szathmáry, 1993) than strict endogamy does for a group and its individuals.

Mutants do occur, however, and cancerous cell lineages can run riot at their organism's expense. Buss (1987) reviewed features of different species which enhance the harmony between individuals and their cells. In some phyla, the "germline" is sequestered, preventing a cancerous lineage from spreading to an individual's offspring. This circumstance is analogous to a honeybee or ant queen sterilizing her workers to create a common interest among them in helping her reproduce. In some species, maternal genes control the early stages of her offspring's development, just as, by judicious distribution of food or hormones, the queen of a complex ant society apportions her workers among different "castes" or morphotypes, thereby programming a division of labor suitable for her colony (Hölldobler and Wilson, 1990).

In summary, selection among groups played a crucial role in the evolution of both eukaryotic cells and metazoans.

3 Mutual Enforcement

Mutual enforcement is vital for the maintenance of many mutualisms. I mentioned previously the mutual enforcement by honeybee workers of their common interest in helping their queen reproduce. By eating each other's eggs, they make it pointless for workers to lay eggs on their own rather than helping their queen. An analogous mechanism enforces the fairness of meiosis. Fair meiosis ensures that selection favors only those alleles that benefit their carriers. Some "distorter alleles," however, bias meiosis in their own favor (Leigh and Rowell, 1995). Up to 95% of the sperm of an individual heterozygous for a segregation distorter carry the distorter allele. If individuals homozygous for the distorter die childless, the distorter allele spreads until death of its homozygotes balances its spread from biased meiosis. Imagine a mutant at a locus on another chromosome whose only effect is to restore fair meiosis in all its bearers. Because alleles on different chromosomes assort independently at meiosis, no allele at this mutant's locus can benefit from the distorter's bias. This mutant spreads because it prevents the distorter's spread among its bearers and thus spares some of its bearers from death caused by being homozygous for the distorter. Because mutants which restore fair meiosis are favored at loci on every chromosome but the distorter's, fair meiosis expresses the common interest of the genome as a whole. This selection has been demonstrated empirically. Moreover, the rarity among birds and mammals of alleles affecting sex ratio suggests that this selection has closed off most of the "easy" ways to bias meiosis, ensuring that selection favors alleles only if, on balance, they benefit the individuals that carry them. Therefore, even though genes are the ultimate units of self-interest which drive evolution, we analyze the adaptations of whole animals.

The most familiar form of mutual enforcement is reciprocal altruism—help those who help you, and make it stick by retaliating against those who cheat you (de Waal, 1996). Reciprocal altruism is illustrated by the sexual behavior of hamlets, simultaneously hermaphroditic coral reef fish, studied by Eric Fischer (Leigh and Rowell, 1995). Even though a sperm is much smaller and "cheaper" than an egg, a successful sperm contributes as many genes to future generations as a successful egg. Thus, most hermaphrodites devote as much energy to male functions—making great quantities of sperm, fighting or otherwise competing with each other for mates, and so on—as to the female functions of bearing and raising young. The proportion of a population's reproductive effort devoted to male functions represents the "50% cost of sexual reproduction." Hamlets avoid this cost by pairing off at mating time and alternating sex roles in successive spawns (matings). The fish which releases eggs during one spawn fertilizes the other's eggs during the next. This "egg trading" short-circuits competition for mates and permits much lower sperm production, greatly reducing the cost of male functions. Reciprocation is favored even though it is cheaper to play the male role. At successive rounds of reciprocal spawns, each partner releases more eggs for the other to fertilize, as if mutual trust were increasing, whereas a fish that tries to play the male role twice in a row is dismissed by its partner.
Mutualism among members of a chimpanzee group is also based in large part on reciprocal altruism (de Waal, 1996). The community of interest among a group's chimpanzees is based on two fundamental features. First, a chimpanzee is in danger outside a group; it needs to belong to one. Second, a group's effectiveness depends on all its members. Thus, all must accommodate, even the dominant, because without his fellow members the dominant would be alone (W. J. Smith as cited in Leigh and Rowell, 1995). Chimpanzees help those who have helped them, and they retaliate against those who have attacked them or failed them in time of need. Chimpanzees who are most generous to others when they obtain food are most likely to be given food when they beg for it. Chimpanzees' strategies of deception show that they are sufficiently self-aware that they can predict another's behavior by "putting themselves in the other's shoes." A chimpanzee who has supported another in a fight expects support from that other if it is attacked in consequence, and it will show anger if that support is not forthcoming.

Chimpanzees, however, seem to have a sense of justice (defined as what promotes the group's common good) that transcends reciprocal altruism (de Waal, 1996). Females may gang up on an alpha male who is taking excessive revenge on a subordinate. The group supports alpha males who mediate fairly in disputes, protecting lower ranking animals. If an alpha male always favors his allies when he interferes in fights, females will keep him from interfering. Indeed, alpha males retain their dominance by consent of the dominated. Finally, chimpanzees try to reconcile with animals whom they have just attacked, while the group usually celebrates such reconciliations. Because fights endanger the welfare of the whole group, reconciliations serve the "common good," including that of the reconciler, who needs its fellow group members.

4. Mutualism and the Common Good

Mutualism evolves only if there is genuine community of interest among the potential partners. In his Politics, Aristotle expressed this point very clearly: States whose constitution and social organization serve the common good more nearly are less susceptible to overthrow by popular revolution or factional putsch. Often, a common interest permits mutualisms to develop that lack any means of enforcement. On the other hand, even coreplication of host and symbiont (whereby symbionts can propagate only to the offspring of the host), such as that maintaining the mutualism between eukaryotic cells and their organelles, will not create a stable mutualism if host and symbiont do not share a stable common interest.

Mixed-species bird flocks are examples of mutualisms without enforcement (Leigh and Rowell, 1995). A typical Neotropical mixed flock contains one pair each of several nuclear species, some with attendant young. Adults of these nuclear species jointly defend a common territory. As the flock progresses regularly over its territory, certain birds with smaller territories join the flock as it passes through them, whereas other birds with larger territories move from flock to flock. Each species of bird eats different food, or feeds in different types of places, as if to minimize competition among flock members. The advantage of a flock is more eyes to watch for predators and more birds to "mob" a predator when need be (Moyghan as cited in Leigh and Rowell, 1995). Even an alarm signal benefits its giver—not only by helping to keep safe the group on which it depends but also by informing a potential predator that, since it has been seen in time to be avoided, it might as well go elsewhere. This mutualism of seemingly unenforced common interest has caused the evolution of certain colors and behaviors among members of certain mixed flocks (Moyghan, 1998).

One of the most striking mutualisms of the tropical forest is that between fig trees and their pollinating wasps (Herre, 1996). Each fig fruit is a flowerhead turned outside in to form a ball, lined with flowers on the inside, with a hole at one end. Each fruit is pollinated by one or more female wasps. These enter the fruit, pollinate its flowers, and lay eggs in up to half these flowers. Each wasp's larva grows within a single fig seed. When a fruit's adult wasps "hatch," they mate among themselves and the fertilized females fly out in search of new trees to pollinate. These pollinator wasps are parasitized by nematodes. In some fig species, each fruit is pollinated by several wasps, each carrying nematodes into the fig. The young of these different nematodes compete to enter the fertilized females leaving the fig, reducing the wasps' ability to reach other figs and their reproductive success should they reach one. In other, "single-foundress" fig species, almost every fruit is pollinated by a single wasp. In these species, nematodes can only infect the young of their current host. Because of this coreplication, their reproductive success depends on that of their host. Selection thus favors nematodes which minimize the damage they inflict on their host. Such nematodes sometimes even enhance their host's reproduction.

Coreplication shaped a stable mutualism between eukaryotic cells and their organelles, which persists in
those few species whose members receive organelles from both parents. Nonetheless, coreplication of wasps and nematodes in single-foundress fig species has had no such effect. Some fig species with several pollinators per fruit are descended from single-foundress species. Nonetheless, their nematodes have reverted to the status of damaging parasites: They must lack a stable community of interest with their host wasps.

Another striking mutualism involves reef-building corals and their zooxanthellae, the symbiotic algae that supply them with carbohydrates in return for nutrients and a place in the sun. Many of these corals call their zooxanthellae from the surrounding water rather than inheriting them from their mothers by mutualism-enforcing coreplication. Wasps which inherit their nematodes from several hosts apiece suffer grievously from these parasites. Multiple origins of a coral’s zooxanthellae do not disrupt the coral–algal mutualism, presumably because the complementation of functions between corals and their zooxanthellae establishes an effective community of interest between these organisms. Community of interest makes the mutualism, not the breeding system.

Similarly, a mutualism does not outlive the community of interest among its partners. When the queen dies in a colony of the social wasp *Metapolybia aztecoides*, no one member of the colony can lay enough eggs to maintain the colony’s pool of workers. Several would-be queens thus join forces to undertake this task. Practice, however, increases their reproductive rate. Mary Jane West Eberhard found that when each reproductive no longer needs the others, they fight it out for the profits of their cooperation. In summary, a community of interest among the potential partners is the one thing needful for the evolution of mutualism. Coreplication, social system, reciprocal altruism, etc. can only shape a mutualism if there is a genuine community of interest among the potential partners.

III. WHY IS MUTUALISM SO CRUCIAL TO PROGRESSIVE EVOLUTION?

Mutualism has allowed the successive formation of more complex and effective wholes from parts already tested by natural selection (Maynard Smith and Szathmáry, 1995). Moreover, wholes composed of modules already adapted to respond “constructively” to challenging circumstances are more capable of further adaptive evolution (Gerhart and Kirschner, 1997). Finally, the conflicts between selection at different levels, and the mechanisms by which these conflicts are controlled or suppressed, represent a series of distinctive footprints of the decisive role played by natural selection in macro-evolution. Indeed, study of the role of mutualism in evolution allows evolutionary history to testify decisively to the mechanisms of evolution.

A. Mutualism and the Major Transitions of Evolution

Mutualism played a crucial role in all the major transitions of evolution. Perhaps the first of these transitions, almost effaced by the mists of time, was the transformation of nucleic acid molecules which replicated themselves by parasitizing protein-based metabolism into genes programming the metabolism and development of discrete organisms (Dyson, 1985). Whether Dyson is correct or not, the origin of life as we know it did involve the transformation of mutually dependent genes into coherent genomes (Maynard Smith and Szathmáry, 1995). Later transitions include the evolution of bacterial cells containing a variety of smaller parasitic and commensal microbes into genuine eukaryotes (Margulis, 1993); the evolution of sexual reproduction, whereby two individuals cooperate to produce offspring more varied than either could alone (Maynard Smith and Szathmáry, 1995); the evolution of meiosis, which ensures an appropriate apportionment of the genes of a diploid cell to its haploid descendants (Maynard Smith and Szathmáry, 1995); the evolution of complex, multicellular organisms (Buss, 1987); and the evolution of animal societies (Leigh and Rowell, 1995).

All these transitions involve entities cooperating toward achievements which no one entity could accomplish unaided. Many of these transitions, such as the evolution of eukaryotes, metazoans, and complex insect societies, transformed groups of formerly independent entities into coherent, integral wholes. Insofar as evolutionary progress has occurred, the evolution of mutualisms has made it possible.

B. Mutualism, Modularity, and Evolvability

Herbert Simon compared two watchmakers, Tempus and Horus. Horus put together a watch by making sub-assemblies of 10 parts apiece and then combining them into 2d-level assemblies of 10 subassemblies apiece, and so forth, until his watch was finished. Tempus would try to put together a watch without benefit of subassemblies. Both were subject to interruptions. An interrup-
tion forced Tempus to begin his watch again from scratch, whereas it forced Horus to begin his current subassembly again. As a result of his subassemblies, only Horus could finish watches.

The relevance of this parable to evolution is that mutualism is a mechanism that can combine preexisting "subassemblies," each already tested by natural selection, into larger, more effective wholes (Dyson, 1985). Mutualism thereby allowed the evolution of complex organisms by manageable stages. Moreover, when cooperating entities share a common interest strong enough for selection to transform groups of cooperators into better integrated, more effective wholes, the resulting combines benefit from two properties. First, the self-interested adaptability of the component entities makes for a more effective and adaptable combine. Second, the adaptability of its components enhances the combine’s capacity for adaptive evolution.

The first of these propositions is illustrated by the behavior of the oldest of intracellular organelles. The microtubules and centrioles (centrosomes) of eukaryotic cells, responsible for the “asters” and “spindle” that form during mitosis and meiosis, are not programmed by nuclear DNA. Instead, these organelles appear to be remnants of spirochaetes which invaded the ancestors of their hosts countless ages ago (Margulis, 1993). When a cell divides in mitosis, the centriole, just outside the nucleus, divides first, and its two descendants move to opposite sides of the nucleus. Microtubules grow out from the centrioles, forming asters. A microtubule quickly dissolves unless it encounters a part of a chromosome called a "kinetochore," which stabilizes a microtubule touching it. This "random exploration" by microtubules allows chromosomes to be attached to their respective asters in a few minutes, accomplishing the essential first step in apportioning a complete set of chromosomes to each daughter cell (Gerhart and Kirschner, 1997). At the level of cells within organisms, the generation and multiplication of a great variety of ‘‘T cells” allows selection for and mass production of those antibodies required to protect their organism from disease (Gerhart and Kirschner, 1997). At the level of organisms within societies, the search by foraging honeybees for new sources of food and the communication of successful searches to the hive, thereby recruiting other foragers to the newly found food, allow honeybee colonies to allocate foraging effort effectively (Seeley, 1995). At all three levels, forms of random exploratory search, conjoined with mechanisms rewarding success, play a crucial role in the development of well-adapted morphology and behavior. Search by quasi-autonomous organisms often serves these purposes more effectively than more centrally controlled processes (Gerhart and Kirschner, 1997).

Adaptability of cells and tissues greatly increases the probability that a major change, genetic or environmental, will not prove lethal. E. J. Shriver described a goat born without forelegs. This goat learned to hop about like a kangaroo. To suit this goat’s peculiar two-legged hop, its hind legs became enlarged, its spine became curved and the sizes and shapes of its vertebrae changed, and the sizes and points of attachment of many muscles and ligaments also changed. These adjustments testify to the adaptability of this goat’s cells and tissues. Similarly, N. Smythe (as cited in Leigh, 1999) was able to create a social strain of pacas, otherwise fiercely territorial animals, by rearing one generation of young under special conditions. These young passed on their new social “traditions” to their descendants. The adaptability of animals, developmental and social, allows them to adapt to some very unusual circumstances.

C. Mutualism and Unmistakable Footprints of Natural Selection

I have shown how several evolutionary transitions involved smaller parts combining into larger, more integral wholes. Thus, certain parasites and commensals of bacteria were transformed into organelles of eukaryotic cells, certain multicellular aggregates were transformed into coherent multiscellular individuals, and certain insects which lived together in groups were transformed into units of cohesive insect societies with a single reproductive and complex division of labor.

A trace of each of these transitions is left in the potential conflicts between the erstwhile parts and the whole they formed. Competition between organelles inherited from different parents can injure their hosts. The spread of cancerous cells can kill the individual to which they belong. Worker bees can lay eggs of their own rather than help their queen.

How a transition occurred is often revealed by the mechanisms which suppress conflicts between the resulting whole and its parts. Most organisms are designed to ensure uniparental transmission of organelles, thereby preventing competition between organelles from different parents and ensuring that the reproduction of organelles depends utterly on the reproduction of their hosts. In most metazoan species, sexual reproduction ensures that each individual is genetically identi-
cal save for the accidents of somatic mutation, a circumstance that ensures that selection discriminates among individuals and not the cells within a single individual. The other means by which metazoans reduce the threat of "rebel" cell lineages—sequestration of the germline (so that cancers of the body cannot be passed on to offspring), maternal control of the early stages of development of her young, and so on—are paralleled by the means whereby queens of complex insect societies maintain harmony within their colonies.

Since these conflicts are suppressed by the very mechanisms that joined the parts into larger wholes, the mechanisms involved represent unmistakable footprints of the critical role natural selection played in these transitions—the most important events of macroevolution. The conflicts involved in each such transition, and the means by which they are suppressed, reveal the historical path of evolution and the selective mechanisms that directed it. Thus, study of the role of mutualism in evolution enables evolutionary history to testify to the mechanisms of this evolution.

See Also the Following Articles

BIODIVERSITY, EVOLUTION AND • COEVOLUTION • COMPETITION, INTERSPECIFIC • DIVERSITY, COMMUNITY/REGIONAL LEVEL • GENETIC DIVERSITY • PARASITISM • SOCIAL BEHAVIOR • SPECIES COEXISTENCE

Bibliography


I. Introduction

II. Number of Species Known and Expected

III. Geographical Patterns

IV. Habitats and Adaptations

GLOSSARY

anthropochoric Geographic distribution due to human agency.
cerotegument Hydrophobic secretion layer covering the body of several arthropods adapted to temporary submersion.
disharmonic, fauna A fauna in which a strongly reduced number of major lineages is represented, as in oceanic islands, due their improbable colonization from distant sources.
laurisilva Warm-temperate woodlands characterized by evergreen broadleaves such as laurel (Laurus) trees.
pupoid Earliest postembryonic stage with incompletely developed appendages, as in insect pupae, hence motionless.
stenotopic A species (or higher taxon) with very restricted geographic range.

MYRIAPODS have long been treated as a natural class (Myriapoda) of the phylum Arthropoda, but many doubts have been raised regarding the close affinity between the four main groups of terrestrial, tracheate, and multilegged arthropods traditionally classified as myriapods. The term, however, is still universally used as vernacular to contrast these arthropods to the insects, also terrestrial and tracheate, with these latter being easily identified by their smaller number of legs (three pairs), the strong differentiation of the trunk into thorax and abdomen, and the generalized presence of wings.

I. INTRODUCTION

Myriapods are wingless terrestrial arthropods with at least nine pairs of walking legs in the adult and a trunk not distinctly subdivided into thorax and abdomen. Unlike many hexapods (insects), myriapods never undergo complete metamorphosis in their life cycle. Some of these and other features of myriapods have long been regarded as primitive within a large arthropod lineage called the Atelocerata (or Tracheata, Uniramia, or Laurata), traditionally recognized as including myriapods and hexapods, as opposed to the other main arthropod lineages Chelicerata and Crustacea. However, recent molecular investigations (Friedrich and Tautz, 1995; Boore et al., 1998) suggest that insects may be closer indeed to crustaceans than to myriapods, a scenario that invites fresh reinvestigation of the real affinities of the myriapod groups among themselves and to the remaining arthropods.
The centipede body is divided into head and trunk. Compound eyes are only present in Scutigeromorpha, whereas groups of simple eyes (ocelli) are present in most Lithobiomorpha and many Scolopendromorpha; all Geophilomorpha and many Scolopendromorpha are blind, as are several cavernicolous representatives of the Lithobiomorpha. The antennae may be longer than the body, as in the Scutigeromorpha and in some Lithobiomorpha. In these two groups, the number of antennal articles is mostly high and variable, but it is fixed and generally lower in the other groups (14 in all Geophilomorpha and 17 in many Scolopendromorpha). The first trunk segment bears a pair of specialized appendages—the poison claws (or forcipules), each of which contains a voluminous poison gland—which are used in the capture of prey and occasionally in defense. The mouthparts are generally adapted for cutting and chewing hard matter, such as wood or dead leaves, but some millipedes have evolved adaptations to sucking. The trunk is normally elongate, more or less flattened dorsoventrally, but subcylindrical in several orders. The body wall is rarely soft and flexible (subclass Pselaphognatha with the only order Polydesmida and the exoskeleton is usually rigid (subclass Chilognatha, with all remaining orders) due to the presence of calcium salts in the internal layers (endocuticle). When the millipede is about to undergo a molt, these salts are dissolved. Accordingly, these arthropods need to obtain from the food sizable amounts of calcium. There are no waxes in their epicuticle, but some protection from evaporation is obtained through the lipid compounds in their exocuticle.

The first trunk segment (the collum) is legless; it is followed by three (four in the Spriothrobida) “thoracic” segments with one pair of legs each and a further number, sometimes very high, of “abdominal” segments with two pairs of legs each. There are 11–17 pairs of legs in the Polydesmida and at least 17 (but usually many more) in the remaining groups, with the highest number being 375 pairs of legs recorded in Illacme plenipes (Engel, 1990). Most millipedes are provided with chemical defenses in the form of glands (ozadens) producing noxious substances which pour out from lateral series of repug-
FIGURE 1 Habitus of representatives of the four main centipede orders: A, Scutigera coleoptrata (Scutigeromorpha); B, Geophilus carpophagus (Geophilomorpha); C, Scolopendra cingulata (Scolopendromorpha); D, Lithobius rutilus (Lithobiomorpha).
natorial openings (ozopores). A diversity of defense substances have been found in these animals, with distinct classes of chemicals being characteristic of the different millipede groups. The juloids rely on benzoquinones, hydroquinones, phenol, and the acetates of some long-chain carboxylic acids; the polydesmoids rely on several carboxylic acids but especially on benzoic acid, benzaldehyde, hydrogen cyanide, and mandelonitrile; and the pill millipedes (Glomeridae) rely on very peculiar and complex heterocyclic compounds (glomerin and homoglomerin) and the Polyzoniiidae on compounds known as polyzonimine and nitropolyzonimine (Eisner et al., 1978). Several millipedes are adapted to coiling onto themselves (volvation), thus becoming a smooth ball difficult to seize; many others can roll themselves into a spiral flat coil. However, despite such precautions, many millipedes fall victim to insectivorous birds and predatory arthropods.

Special modifications of one or two leg pairs in the male may involve their transformation into clasping and/or genital organs. Posterior claspers are characteristic of the Pentazonia (the pill millipedes and their closer relatives). The male uses these claspers to fix the female
3. Symphyla

Symphylans are tiny myriapods, mostly 2–9 mm long (but Hanseniella magna is 25–30 mm long alive), generally with a whitish body (Fig. 3A). All species are blind and bear one pair of elongate multisegmented antennae, 12 (rarely 11) pairs of legs, and a pair of posterior unarticulated appendages called spinnerets. The mouthparts are of masticatory type. The genital opening is found on the fourth trunk segment. Symphylans are the only arthropods with a pair of respiratory openings on the head, but the whole of their very soft and permeable cuticle is important for gas exchange. In these tiny myriapods there is no cuticular protection against evaporation.

4. Pauropoda

Pauropods are the smallest among myriapods; most species are only 0.3–0.7 mm long, and the largest is just 1.9 mm long (Fig. 3B). There are no eyes. The antennae are of unique structure, with four (order Tetramerocerata) or six (order Hexamerocerata) basal articles and two apical or subapical branches, each bearing in turn one or two flagella. The mouthparts are comparable to those of millipedes but are generally adapted to the suction of fluid aliment. The trunk bears 9–11 pairs of legs. Most species lack a respiratory system; one pair of tracheae, with spiracles on the coxae of the first pair of legs, is present only in the Hexamerocerata.

B. Basic Biology

Nearly all myriapods are strictly terrestrial and most species are found in forest leaf litter or in rotten wood, under the bark of dead trees, in the soil, or in caves. Several millipedes also occur in dung, compost, and almost any other kind of plant debris; in nests of ants, termites, and birds; in worm and mammal burrows; near water; on the seashore; on open terrain; under stones; in buildings; etc. That is, they are found in virtually any terrestrial environment.

A few species of diplopods (e.g., Blaniulus guttulatus) and symphylans (Scutigerella immaculata) are agricultural pests. Due to their swarming habits, a few millipedes can be of danger for transport when smashed in myriads on roads or railways or for housing, especially due to their smell, but such extremes are exceptional. In temperate regions, migrating swarms of many thousands of millipedes have been recorded, especially swarms of polydesmids (e.g., Fontaria virginiana, Parafoantarila laminata, and Polydesmus collaris). Causes for and mechanisms of this behavior are little known, but it has been suggested that heavy rainfalls may cause the swarming behavior of a large spirostreptid (Plusioporus setiger) in southern Brazil. Larger millipedes have quite long life spans lasting between 2 and 11 years, whereas their life histories usually take 2 or 3 years even in tropical environments. In high-montane habitats, the development is normally retarded by a year or two. However, some tiny tropical species live for less than...
A few large scolopendromorphs include feeding on barnacles by swarms of specimens hunters; gregarious habits have rarely been observed, fangs, are basically predatory. They are mainly solitary centipedes, as revealed by their poisonous forcipular structures, which can habits differing from those prevailing in forest litter. Tree dwellers are not common among myriapods, being mainly represented by small or slender, if not flat-bodied, millipedes and by some representatives of lithobiomorph and geophilomorph centipedes. Species confined to tree crowns or living in suspended soil are even fewer. Some scolopendromorphs, however, are fairly common in suspended soils in the tropical forests of America. Semicolotropic, littoral, or nidicolous myriapods are exceptional. Myriapod species capable of tolerating seawater are particularly few, although these are often quite common and widespread. Interestingly, no conspicuous morphological types can be discerned among myriapods as obvious adaptations to any of such definitely marginal environments. Again, no common morphological adaptation is shared by the relatively few anthro- poid millipedes that have attained more or less strongly reduced, are known to live on fungi and/or semiliquid end products of decaying plant mate- rial. Most pauropods feed on fungal hyphae and spores, but Millitauropus (Hexamerocerata) feeds on small arthropods (springtails and their eggs). Symphylans are largely vegetarian, rarely saprophagous. Hence, the roles played by various myriapod groups in the terrestrial food chains differ, with millipedes being of particular importance as primary destructors and soil-forma- tion factors and centipedes as active and restless predators. Representatives of only a few millipede or- ders are capable of manufacturing silk for cocoon pro- duction during molting and/or egg-laying.

C. Reproduction

A few dozen cases of parthenogenesis are known (Eng- hoff, 1994). One is the lithobiomorph centipede La- mycetes corculus, a synanthropic species known from several continents. Two small European millipedes (Nemataoma vanicornis and Polyxenus lagurus), both mainly living under tree bark, exhibit geographical par- thenogenesis, with bisexual populations in less dis- turbed areas and unisexual ones in areas where the species appear to be recent colonizers. Similarly, Muy- xoecus orlatus is bisexual in its native South Ameri- can habitats, but a stable population recently estab- lished in a hothouse in Germany is parthenogenetic. There is also evidence of parthenogenesis in pauropods. All myriapods are oviparous. Two groups of centi- pedes (Scolopendromorpha and Geophilomorpha) have evolved brood care. In these animals, the female retains the eggs (for as long as 3 months in some species) around her brood until hatching. Parental cares are virtually unknown in the other myriapod classes, with two of the rare exceptions being the millipedes Polyzonium germanicum (order Colobognatha) and Bericostenus hu-
MYRIAPODS

Newborn millipedes have an incomplete number of segments and legs (Fig. 4). In most cases, only three pairs of legs are present in the first active or inactive postembryonic stadium. New segments and leg pairs are added at each molt, sometimes inside of constructed molting chambers, according to a precise schedule which is different in different orders, families, or genera. Three main developmental modi can be distinguished (Enghoff et al., 1994).

In hemianamorphosis (as in the pill millipede Glomeris) the final number of segments and leg pairs is obtained after some molts, but the millipede must undergo further molts without segmental increase before achieving adult size and sexual maturity. In teloanamorphosis (as in polydesmids) the achievement of the final number of segments and leg pairs coincides with obtaining the adult condition and no further molt occurs. In euanamorphosis (as in the cylindrical julids), new segments and leg pairs are added at every molt and the animals continue molting even after reaching sexual maturity. The males of some western European representatives of the order Julida exhibit the peculiar developmental behavior known as periodomorphosis. Under specific adverse environmental conditions, sexually mature males of these species molt to intercalary males with regressed gonopods and suspended reproductive activity. After one or more additional molts, these animals reach a further mature stage, with newly differentiated gonopods.

Hemianamorphosis is also known in centipedes: Scutigeromorphs, lithobiomorphs, and craterostigmomorphs hatch with 4, 7, and 12 pairs of legs, respectively, and obtain the final number of 15 leg pairs following one (craterostigmomorphs) or more molts but continue molting until they reach maturity. The re-mating centipedes (scolopendromorphs and geophilomorphs), however, develop by epiomorphosis. That is, they reach the final number of body segments during the embryonic development; no segment or leg pair is thus added during the postembryonic life.

In symphilans, 6 or 7 of the final 11 or 12 pairs of legs are present in the first postembryonic stadium. Quite peculiar is the following developmental schedule in that first-order and second-order molts regularly alternate until maturity is reached. After each first-order molt a new two-segment unit is added, but only 1 more pair of legs is differentiated, whereas the second-order molts are accompanied by the differentiation of another pair of legs without addition of new segments. Mature symphilans continue to molt, up to 40 times in Scutigrella immaculata.

In pauropods, the first postembryonic stadium is a motionless pupoid with inarticulated traces of the antennae and of the first two pairs of legs. After a molt it changes into an active stadium with three pairs of legs. The development then proceeds by anamorphosis.

II. NUMBER OF SPECIES KNOWN AND EXPECTED

Our knowledge of myriapods is highly fragmentary and incomplete (Table I). Not only does the bulk of existing species diversity remain to be described, but also the number of higher taxa appears to vary considerably as estimated by different taxonomists. For example, according to E. H. Eason (personal communication), of the approximately 150 genera and 1730 species/subspecies of Lithobiomorpha named to date, only approximately 17 and 950, respectively, are probably valid. In the Diplopoda, by far the largest of the myriapod classes, the situation concerning the number of orders and families, let alone genera and species, is badly equivocal to the point that Hoffman (1980) could equate the state of the art in millipede systematics to that in entomology in the middle of the nineteenth
century. Thousands of taxa awaiting description are available in the collections. For example, according to P. Johns (personal communication), the collections of Australian millipedes already kept in Australia's natural history museums contain at least 2000 undescribed species. Indeed, it is mostly tropical myriapod faunas that appear to be particularly poorly explored.

Judging from the species described thus far, one millipede order outnumbers all others in terms of species counts—Polydesmida (up to one-half of the genus and species richness of the Diplopoda), followed by Spirostreptida, Chordeumatida, and Julida. The remaining orders are minor to marginal. Thus, the pentazonian order Glomeridesmida contains only a single genus with a handful of species in the Indo-Australian and Neotropical realms. The order Siphoniulida is extreme and obviously relict, being known by a single species each in Malaya and Central America. Because no male siphoniulidan has ever been described, and the structure of the male gonopods (if any) is absolutely unknown, the affinities of this order remain dubious. Probably less than 20% of the world fauna has been described to date; this is particularly true of the largest class Diplopoda (Table I). Indeed, updated counts/checklists are only available for the millipede faunas of some European countries (Kime and Golovatch, 2000) and those of North and Central America and the Caribbean, whereas the fauna of Australia is assessed very crudely at the familial and ordinal levels only and that of Eurasia remains unpublished. The fact that the numbers of genera and species of Diplopoda currently known in Europe and the Mediterranean are similar to those known from the whole of Asia shows how poorly investigated these are is the Asian, largely tropical fauna. A world list of Pauropoda is available, but the distributional data concerning Chilopoda and Symphyla are still scattered in numerous local publications and no recent overview is available at the world or at least the continental level.

### III. GEOGRAPHICAL PATTERNS

The bulk of species richness is definitely confined to tropical countries, mainly woodlands (Black, 1997), with certain notable exceptions. Thus, some orders or families appear to be largely temperate and more or less restricted to the Northern Hemisphere, e.g., Lithobiomorpha among Chilopoda or Glomerida, Julida, Callipodida, and Polydesmidae among Diplopoda. For these groups, the hot spots of diversity are mountainous lands at the periphery of the Holarctic Region within the temperate to subtropical areas such as the Atlas, Pyrenees, Alps, Balkans, Caucasus, Tien-Shang, Himalaya, and Appalachians. Figure 5 shows the patterns of millipede generic diversity in Eurasia. Harsh environments such as tundra or deserts are populated by myriapods patchily to marginally if at all, especially by Diplopoda and Symphyla. However, Chilopoda, Pauropoda, and even some Symphyla occur in permafrost areas, and a few scolopendromorph centipedes are quite characteristic desert dwellers. In contrast, high-montane habitats, however adverse, are regularly populated by myriapods, inclusive of millipedes, and sometimes even by endemic species. The highest altitudes recorded are approximately 4500 m above sea level for both millipedes and centipedes, the latter being encountered well within the nival belts of the Himalaya and the Andes.

Given the hundreds to thousands of species in each of the major myriapod groups, one may get the wrong impression that these basically cryptic animals should be highly diverse in most of the local environments, in particular in productive temperate or tropical forests. However, virtually no local fauna is known to include more than two or three dozen species, depending on habitat. Given that myriapods are basically composed of very ancient, widespread, but poorly vagile and largely stenotopic forms, most of the faunules, even among the...
Cold areas host a very reduced fauna of millipedes. In Europe, only two species (*Polyxenus lagurus* and *Proteroiulus fuscus*) live north of the polar circle; no millipede is known from such areas of Siberia or from Greenland. Only relatively few nonanthropochoric species display vast distributions; these tend to be particularly vast among predators (= Chilopoda), especially soil dwellers. Several geophilomorphs are examples of this pattern of distribution, e.g., *Arctogeophilus glacialis* and *Pachymerium ferrugineum*. Forms populating special but widespread habitats likewise appear widely distributed, e.g., the littoral geophilomorph *Strigamia maritima* or the huge, largely arboricolous, Neotropical scolopendromorph *Scolopendra gigantea*. Active colonizers of relatively young areas/habitats, such as the trans-Siberian *Angarozonium amurense*, the pan-European *Onnia toliula* sabulosus, or the subcorticolous pan-European *Proteroiulus fuscus* among millipedes or the subcorticolous pan-European *Proteroiulus fuscus* among millipedes or the Euro-Siberian lithobiomorph centipedes and some other congeners among lithobiomorph centipedes, reach up to the taiga and, in part, tundra belts in the north, the youngest boreal biomes. The Mediterranean chilopod *Scolopendra cingulata* is a notable colonizer of semi-arid badlands, whereas the European symphyllan *Symphyplealla vulgaris* is a colonizer of fresh mines.

Worldwide distribution due to human agency is known for several myriapod species, mainly of European origin such as the millipedes *Cylindroiulus lateritias* and *Ophylus pilosus* and the centipede *Lithobius forficatus*. The origin of the synanthropic centipede *Lamycosis marginatus*, also widely distributed, is unknown but it is possibly from the Southern Hemisphere.

### A. Endemism and Speciation

High numbers of endemic species and high percentages of endemics within the total millipede faunas are peculiar to some environments such as caves and some areas, such as islands and some mountain chains in the temperate regions. The best investigated cases of intense speciation in millipedes are by far those of the genera *Cylindroiulus*, *Dolichotubulus*, and *Acipes* in the Macaronesian area (Enghoff, 1992; Enghoff and Baez, 1993). On Madeira, of 60 millipede species known from the island, at least 29 belong to the endemic *Cylindroiulus* madeirae group (Julidae) and another 6 to an endemic lineage of the genus *Acipes* (Blaniulidae). Still more conspicuous is the *Dolichotubulus* (Julidae) species swarm in the Canary Islands, which has 46 species of a total of 79 millipede species thus far recorded from the archipelago. These *Dolichotubulus* species have colonized the most diverse habitats, from the laurisilva to the xeric habitats and the caves, whereas the Madeiran *Cylindroiulus* are all confined to laurisilva forest habitats.
TABLE II

Distribution of Centipede and Millipede Diversity among the Major Continents/Regions*

<table>
<thead>
<tr>
<th>Continent/region</th>
<th>North and Central America</th>
<th>Afrotropical region</th>
<th>Australia and South Pacific</th>
<th>South America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chilopoda</td>
<td>1.3</td>
<td>0.3</td>
<td>0.3</td>
<td>1.1</td>
</tr>
<tr>
<td>Diplopoda</td>
<td>7.5</td>
<td>3.6</td>
<td>0.4</td>
<td>2.8</td>
</tr>
</tbody>
</table>

* Approximate number of genera in hundreds

In the Canarian Dolichoiulus species swarm, conventional allopatric (intersland) speciation is only marginally responsible for the genesis of the huge diversity of the group. According to Enghoff, speciation occurred mainly within individual islands, with distribution patterns governed mainly by habitat differences between closely related and nearly sympatric species. In several cases, those species which coexist in the same habitat differ conspicuously in size. The same is often seen in Madeiran Cylindroiulus and Acipes species; this rule, however, is far from universal.

B. Distribution by Continents

In general, the overall distribution of myriapods is quite consistent with the world's biogeographic regionalization known since the times of J. Hooker and A. Wallace. In other words, the traditional division of the globe's territory into the Holarctic (= Palaeartic + Nearctic), Afrotropical (including Madagascar and the Cape), Neotropical, Oriental (Indian + Malesian), and Australian (including New Guinea and the islands of the southern Pacific) realms is supported by the patterns demonstrated by myriapod higher taxa. Antarctica is the only continent harboring no Myriapoda, whereas, due to the particularly poorly explored tropical faunas, the distribution of genera and species between the remaining continents or biogeographic regions is highly provisional and appears extremely uneven and patchy (Table II).

C. Hot Spots of Biodiversity and Endemism

The bulk of species richness is confined to tropical countries, mainly woodlands (Black, 1997), although there are some notable exceptions. The opposite extreme is represented by numerous species endemic, e.g., to a single cave or mountain slope; this is especially characteristic of Diplopoda. Local cavernicolous endemics, notably troglobites, are particularly numerous in southern Europe, especially in the Balkan area, which is rich in karst. Similarly, nearly every patch of Amazonian primary rain forest near Manaus, Brazil, contains (much of) its own millipede fauna, with differences being particularly drastic between various kinds of inundation vs nonflooded (= terra firme) woodlands. Approximately the same pattern has been observed among Pauropoda.

D. The Fauna of Islands

Analysis of insular biotas clearly shows the differences in colonizing capability and radiation patterns of centipedes vs millipedes. In most faunas, the ratio of millipede to centipede species is approximately 4:1, but this ratio is much lower in Hawaii, Tahiti, Samoa, and the Fiji Islands. This reveals that centipedes are apparently better dispersers than millipedes. The distribution patterns of millipedes on continental vs oceanic islands are those of poor dispersers. For instance, there are 18 millipede families represented by indigenous species in Sumatra, 17 in Japan, and 14 in Java, but there are only 10 in Australia, 9 in Madagascar, 5 in New Zealand, 3 in New Caledonia, and 1 in the Hawaii Islands. The number of endemic genera is very high in Indonesia and the Antilles, but these island groups do not have endemic families and suprageneric indigenous taxa are only known from Borneo. The myriapod fauna of most oceanic islands is not simply disharmonic but also poor in species. Native myriapods are virtually absent from Kerguelen, Tristan da Cunha, and St. Helena.
IV. HABITATS AND ADAPTATIONS

At least five morphological body plans can be distinguished among millipedes. The pin-cushion millipedes (i.e., the representatives of the worldwide distributed order Polyxenida) are soft bodied, very small (usually <3 mm long), extremely brittle, swift, and capable of withstanding much drier conditions than most other Diplopoda. Many polyxenids live under loose tree bark or are characteristic inhabitants of microcaverns and small crevices under stones, in the uppermost soil, in litter, and in similar substrates.

The pill millipedes (the largely Holartic order Glomerida and the Afrotopical + Indo-Australian order Sphaerotherida) are collectively termed “rollers” because of their ability to roll themselves into mostly glossy balls. Pill millipedes are generally litter dwellers, but some smaller Glomerida are either troglobionts or geobionts.

The colobognathans (i.e., the “sucking millipedes” with reduced mouthparts) and some Chordeumatida, which mostly possess flexible, worm-like, strongly tapering bodies and shorter legs, are termed “borers.” Many Polydesmida and numerous Chordeumatida, often conspicuously ornamented and relatively shorter, long-legged, and displaying more or less strong parapodia (wing-like dorsolateral projections of the diplosegments), are referred to as “wedge types” and are characteristic of forest litter.

The lifestyle, or ecomorphological type, most common and widespread among Diplopoda is that of “bulldozers” or “rammers.” Their long, cylindrical, hard body with numerous diplosegments (hence, numerous pushing legs) penetrates the substrate like a bulldozer, using the broad head as a ram. Most juliform millipedes (Julida, Spirobolida, Spirostreptida, Callipodida, and some others) belong to this ecological type. This burrowing habit must have been critical in ensuring diplopods much or even most of their present-day high ecological and geographical performance. In particular, unlike the remaining ecomorphotypes, only juliforms, among Diplopoda, appear to have colonized virtually any suitable habitat. In fact, in Juliformia belong the few littoral dwellers and deserticoles, but this body plan also provides most of the millipede diversity in troglobionts and geobionts, in addition to anthropochores. Indeed, the northernmost record of a millipede belongs to the subterranean European species Protostreptus fuscus (Julida) in the forest–tundra belt of Yamal Peninsula, Russia’s north, whereas perhaps the most characteristic deserticole among millipedes is Orthopus ornatus (Spirostreptida) in the southern United States and adjacent parts of Mexico. Both these species are rare except in their native environments.

There are sound reasons to believe that, due to their capability to escape adverse conditions by burrowing in the soil, rotten logs, and similar shelters, juliforms (namely, species of the order Julida) dominate in Europe the youngest, fully migratory nucleus of the millipede fauna, showing clear-cut inclinations to dwelling on open terrain. In contrast, all remaining millipede orders display very evident geographical trends in diversification. In Europe, even the relatively uniformly distributed Polydesmida have a remarkable center of secondary diversification in Slovenia, whereas the Chordeumatida are particularly species rich within the Atlantic climatic zone of western Europe (Kime and Golovatch, 2000). Furthermore, only a few species of Julida, apparently in response to the strongly adverse conditions that existed in Europe during the Ice Age, appear to have developed a population strategy unique among other terrestrial animals—the periodomorphosis strategy mentioned previously, i.e., the extension of male life by means of intercalary stadia (Hopkin and Read, 1992).

In summary, most millipede lineages are currently in a phase of rapid evolution and speciation; however, there are a few apparently relict groups (Hoffman, 1980). Polydesmida seem remarkably diverse largely due to their wedge type of burrowing allowing ecological niche partitioning mainly in the litter and at the soil–litter interface. This is particularly obvious in tropical and subtropical faunas. However, examples of troglobionts, geobionts, and myrmecophilous as well as arboricolous species are about as numerous among Polydesmida as among juliforms, whose type of burrowing and lifestyle seem the most characteristic, widespread, and ecologically progressive among all recent Diplopoda.

A. Burrowing Myriapods

Several major myriapod groups tend to be soil dwellers. Usually, adaptations to geophily involve body elongation due to an increased number of segments. This is quite evident, for instance, in several lineages of geophilomorph centipedes, especially the Himantariidae and the Oryidae, which have a flexible worm-like body and peculiarly short appendages. This is also the case for certain scolopendromorph centipedes and some juliform and colobognathan millipedes, which are active burrowers. Body miniaturization also appears advantageous to dwelling in the crevices and/or burrows in the
soil. This adaptation is observed in most Symphyla, numerous Pauropoda, and some Diplopoda (e.g., some representatives of Glomerida, Julida, and Polydesmida). However, only more or less loose soils appear suitable for such myriapods because harder grounds are apparently too difficult to penetrate, especially by the actively burrowing species.

B. Open-Country Myriapods

Short-living Polydesmida are dominant in the savanna. The adults are active during the rain season, whereas the juveniles spend the dry season in resting conditions. The resistance to desiccation varies much between species, with the females being mostly less sensitive than the males, for unknown reasons. Cuticular waxes seem to be present in very few millipede species. One of these species is the Orthoporus ornatus of the American semideserts; when the temperature rises over 40°C, it rests in a coiled condition to reduce water loss. Lowest lethal temperatures for millipedes are in the range of −3 to −7°C, and the highest temperatures are between 36 and 43°C. Preferred temperatures are clearly lower for hygrophilous forest species (4–18°C) than for xerophilous ones (e.g., 26–32°C for Ommatothorax subsulatus, a species common in open environments in southern Europe: the Pyrenees, Sardinia, Italy, the Balkans, Greece, and Turkey, Algeria, and Morocco. A few species have semiaquatic habits. One of them is Serradium semiaquaticum, which inhabits caves near Verona (northern Italy), and its sister species S. hirsutipes, whose juveniles spend up to 11 months under water feeding on algae.

C. Special Habitat Myriapods

A few species have semiaquatic habits. One of them is the littoral julid Thalassobates litoralis, which can be found among stranded decaying matters on the seashore. However, this species does not seem to spend much time under water; at some cavernicolous species from southern Europe (several julids and the polydesmid Serradium semiaquaticum; Enghoff et al., 1997) and some Amazonian species, including Myrmecestodesmus adisi, whose juveniles spend up to 11 months under water feeding on algae.

Both morphological and behavioral adaptations have been observed in myriapods that spend prolonged periods of time under water. Conspicuous are the microtrichia in their spiracles and the cerotegument covering their body which enable flood resistance for weeks and even months compared to the simple flood tolerance (up to a few days) exhibited by nonspecialized forms. These morphological adaptations are present in a few millipedes of the seasonally inundated forests of Amazonia; they are also present in the polydesmid Serradium semiaquaticum, which inhabits caves near Verona (northern Italy), and its sister species S. hirsutipes, which does not have spiracular microtrichia and, as expected, does not have the flood resistance of its congeners.

Regarding centipedes, resistance of seashore geophilomorphs to submersion in seawater is not very conspicuous, being 12–24 hr in Strigamia maritima and up to 36 hr in Hydroschendyla submarina. Longer resistance to submersion in fresh water has been shown for the lithobiont Lamyctes emarginatus from European floodplains and from populations living along the Amazon River (J. Adis, personal communication).

Behavioral adaptations to the seasonally flooded forest of Amazonia have been also recorded in the other myriapod groups. Regarding Symphyla, members of the Scoleopendrellidae have a dormant stage of 5–7 months duration, which in the case of Ribautiella amazonica is spent among roots. Due to the anoxic conditions of the flooded soil, these arthropods are likely to recur to anaerobic metabolism. A similar situation is presumed for Pauropoda living in the same environment because Scleroporus taramantoni has been found to spend 5–7 months under water (J. Adis and U. Scheller, personal communication). Myriapods lacking similar morphological and/or physiological adaptations must rely on vertical migrations along the tree trunks to survive the flood period in the upper trunk or canopy region [e.g., Cuterebresmus adisi (Diplopoda); Hensentiella arbores (Symphyla); and some Geophilomorpha]. The migratory behavior of these species, however, is the most conspicuous aspect of their complex ethological and physiological adaptations. No reproduction occurs during the time they spend on the trunks, and trunk ascent is associated with changes in local climatic conditions (‘‘dry’’ versus rainy season) and with macroclimatic influences such as the El Niño Southern Oscillation.

A large fraction of myriapods are represented by cave dwellers. Adaptations to life in caves are often opposite those to geophily, often including larger size compared to their epigeic counterparts (Causey, 1960), frequently with strong paranota and/or hairs, elongated antennae and appendages, and reduced or missing ocelli. Among myriapods, troglobites are known mostly among millipedes and, to a lesser degree, centipedes. Many myriapods have developed adaptations to cave life and can be defined as troglobionts.

Regarding centipedes, approximately 50–60 troglobitic species of lithobiont myriapods are known from caves of southern Europe: the Pyrenees, Sardinia, Italy, the Balkans, Greece, and Turkey, Algeria, and Morocco. A troglobitic geophilomorph (Geophilus persephones) has recently been described from the cave Gouffre de la Pierre Saint-Martin in the French Pyrenees; it is blind,
as expected, but in this case the blindness is not a specific adaptation to cave life because all geophilomorph lack eyes (Forddai and Minelli, 1999). Modified mouthparts have evolved convergently in several unrelated genera of hygrophilous cave-dwelling millipedes from southern Europe and the Caucasus (Julidae: Leuconeurina, Typhloniulus, and Tringatulus; Illanthiidae: Vasconobrillius; Polysdesmidae: Serradium). In these forms, the biting or masticatory part of the mandible is reduced, whereas its pectinate lamellae are hypertrophied, suggesting a function like a filter, used in collecting organic material suspended in water. Recent investigations on the gut content of several unrelated genera of hygrophilous cave-dwelling millipedes (Polydesmidae: Blaniulidae: Vascoblaniulus; Julidae: Serradium; Polydesmidae: Serradium) lack eyes (Foddai and Minelli, 1999).

Acknowledgments

We gratefully acknowledge the help provided by Joachim Adis with his precious comments on a draft of the manuscript and by Edward Holt Eason (deceased), Peter Johns, and Ulf Scheller, who shared his precious comments on a draft of the manuscript and by Edward Holt Eason (deceased), Peter Johns, and Ulf Scheller, who shared his precious comments on a draft of the manuscript and by Edward Holt Eason (deceased), Peter Johns, and Ulf Scheller, who shared his precious comments on a draft of the manuscript and by Edward Holt Eason (deceased), Peter Johns, and Ulf Scheller, who shared his precious comments on a draft of the manuscript and by Edward Holt Eason (deceased), Peter Johns, and Ulf Scheller, who shared his precious comments on a draft of the manuscript.

See Also the Following Articles

ARTHROPODS, AMAZONIAN • INVERTEBRATES, TERRESTRIAL, OVERVIEW

Bibliography


I. Introduction

Most species have durations that are very short relative to the history of life, and it is likely that 99% of all species that have ever lived are now extinct. Extinction, therefore, is very much a natural process that produces continuous turnover in the membership of biological assemblages as species are steadily lost by extinction and replaced by speciation. Natural extinctions may be divided into two kinds: those that happened during geologically brief periods of crisis when many taxa disappeared (the mass extinction events) and those during the long intervals of "background" time between mass extinctions. Mass extinctions have attracted a great deal of interest because they hit so deeply into standing biodiversity and because the biotas that reformed after mass extinctions were often quite different from those...
that existed before; they dramatically reshaped world patterns of biodiversity. However, 95% or more of the total number of extinctions have taken place outside mass extinctions (May et al., 1995). These background extinctions have been most important in producing turnover in species assemblages.

Differences in the probability of extinction for different types of organisms interact with rates of origination of new lineages to shape patterns of biodiversity. To understand patterns of biodiversity it is therefore essential that the process of extinction be understood. In particular, we need to know whether extinction rates vary among different types of organisms and to identify the characteristics of taxa that make them more or less vulnerable to extinction. Extinction can be studied at the level of higher taxa—genera, families, and so on—but this article is concerned primarily with extinctions of species.

II. NATURAL CAUSES OF EXTINCTION

The causes of extinction can be divided into two types: changes in the environment and interactions with other species.

A. Changes in the Environment

Populations of all species fluctuate in abundance. A large part of this variation is due to chance variation in environmental conditions, such as variations in weather or events such as cyclones, which can be regarded as environmental 'accidents.' Species vary in their ability to resist accidents, but for any species there will eventually come an environmental event of sufficient magnitude to wipe it out, or an unlucky succession of smaller blows that drive its abundance down to zero, even when there is no trend in average conditions.

Additionally, species may be driven extinct by directional changes in the environment, such as changes in temperature to levels outside their tolerance or the disappearance of key habitats. Species can respond to such changes in the environment by (i) acquiring adaptations to the changed conditions ("evolving out of trouble"), (ii) shifting their geographic ranges to remain within a set of suitable conditions ("moving out of trouble"), or (iii) going extinct. If the pace of environmental change is fast relative to the rate of evolution, the first of these responses is much less likely than the other two. The fossil record provides many examples of extinctions apparently driven by environmental changes of many different kinds, especially during mass extinctions events. Environmental changes have been much more extreme and rapid during some periods of Earth history than during others, but environmental conditions have never been truly static. Global temperatures have trended downwards for much of the past 30 million years, and at a finer temporal scale variations in the earth's orbit produce fluctuations in climate with periodicities in the range of 10,000 to 400,000 years. These are called Milankovitch cycles, and during the past 2 million years they have produced the succession of ice ages, but they must have forced rapid swings in climate throughout Earth history (Bennett, 1997).

B. Interactions with Other Species

The population growth rates of species are constrained by the species predators, parasites, and competitors. The abundance of a species will be reduced if its natural enemies become more abundant or evolve greater efficiency, or if its geographic range is invaded by a new enemy to which it has not evolved defenses.

It is generally difficult to reconstruct such interactions in the fossil record, but some patterns may reflect the impact of predators on prey species. For example, the number of species of endobyssate bivalves gradually declined through the Mesozoic. These species were abundant, immobile, and lay partly exposed on the open seafloor. Their decline coincided with radiations of marine crabs, teleost fish, and carnivorous snails, groups that account for most predation on modern bivalves and to which the endobyssates must have been vulnerable (Stanley, 1979). The few present-day survivors of the group live in conditions with low predator pressure. Recent experience shows that prey species can be rapidly driven extinct by unfamiliar predators that invade their habitat. The carnivorous snail Euglandina rosea, for example, has caused the extinction of hundreds of species of endemic snails, including 600 of more than 1000 original species from the Hawaiian Islands, since its introduction to many Pacific islands in the 1970s. Such a rapid course of events would be unsolvable in the fossil record.

Probably, the impacts of environmental changes and of other species interact in subtle ways to cause extinction. For example, a species might experience a slight environmental change that reduces its population without driving it extinct and to which it could readily adapt given sufficient time. However, the same change might favor a competitor or trigger a range expansion by an unfamiliar predator suited to the new conditions. Interactions of this kind could mean that quite small envi-
Environmental changes could have dramatic consequences leading ultimately to extinction.

III. LIFETIMES OF SPECIES

A. Variation among Taxonomic Groups

The oldest extant species known is the tadpole shrimp, *Triops cancriformis*, a small freshwater crustacean found in temporary pools in arid regions of Eurasia and north Africa that is indistinguishable from 180-million-year-old fossils bearing the same name. Most species do not live to this age, as shown in Table I.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Estimated mean species duration (millions of years)</th>
<th>Extinction rate (extinctions per million species years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-celled organisms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diatoms</td>
<td>8</td>
<td>0.12</td>
</tr>
<tr>
<td>Dinoflagellates</td>
<td>13</td>
<td>0.08</td>
</tr>
<tr>
<td>Planktic foraminifera</td>
<td>7–20</td>
<td>0.14–0.09</td>
</tr>
<tr>
<td>Benthic foraminifera</td>
<td>25</td>
<td>0.04</td>
</tr>
<tr>
<td>Plants</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early vascular plants</td>
<td>10–14</td>
<td>0.1–0.07</td>
</tr>
<tr>
<td>Pteridophytes</td>
<td>7–15</td>
<td>0.14–0.07</td>
</tr>
<tr>
<td>Gymnosperms</td>
<td>2–15</td>
<td>0.1–0.07</td>
</tr>
<tr>
<td>Monocots</td>
<td>4</td>
<td>0.25</td>
</tr>
<tr>
<td>Diatoms</td>
<td>3</td>
<td>0.35</td>
</tr>
<tr>
<td>All invertebrates</td>
<td>11</td>
<td>0.09</td>
</tr>
<tr>
<td>Reef corals</td>
<td>20</td>
<td>0.05</td>
</tr>
<tr>
<td>Mollusks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marine gastropods</td>
<td>10</td>
<td>0.1</td>
</tr>
<tr>
<td>Marine brachiopods</td>
<td>15</td>
<td>0.07</td>
</tr>
<tr>
<td>Mesozoic ammonites</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Upper Cambrian invertebrates</td>
<td>1.3</td>
<td>0.77</td>
</tr>
<tr>
<td>Marine ostracods</td>
<td>8</td>
<td>0.12</td>
</tr>
<tr>
<td>Silurian graptolites</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>Echinoderms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Echinoids</td>
<td>6</td>
<td>0.17</td>
</tr>
<tr>
<td>Cenozoans</td>
<td>6.7</td>
<td>0.15</td>
</tr>
<tr>
<td>Bryozoa</td>
<td>12</td>
<td>0.08</td>
</tr>
<tr>
<td>Freshwater fish</td>
<td>3</td>
<td>0.33</td>
</tr>
<tr>
<td>Birds</td>
<td>2.5</td>
<td>0.42</td>
</tr>
<tr>
<td>Mammals</td>
<td>1.4</td>
<td>0.71</td>
</tr>
<tr>
<td>Cenozoic mammals</td>
<td>1–2</td>
<td>1–0.3</td>
</tr>
<tr>
<td>Horses</td>
<td>4</td>
<td>0.25</td>
</tr>
<tr>
<td>Primates</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

* Compiled from Stanley (1979), Niklas et al. (1993), May et al. (1995), and McKinney (1997).
tion can be made, but most estimates include an unknown combination of real and pseudoextinctions.

There is considerable variation in species life spans among different groups of organisms. In general, it seems that species of unicellular organisms last longer than species of multicellular organisms, and invertebrates appear to last longer than vertebrates. The most long-lived animal groups tend to be marine rather than terrestrial, although too little is known about species durations of marine vertebrates or terrestrial invertebrates to judge differences between marine and terrestrial species independent of the difference between vertebrates and invertebrates. Some marine invertebrate groups that are now entirely extinct (the trilobites, ammonites, and graptolites) had short species durations, even though they were successful, abundant, and species rich for long periods of geological time. Among plants, species durations were longer in groups that appeared early in the evolutionary history of plants but have tended to shorten in recently evolved groups.

The inverse of the typical duration of species in a given taxonomic group can be used as a measure of the probability of extinction per unit time for species belonging to that group. Long durations equate to low probabilities of extinction. In Table 1, probabilities of extinction have been expressed as the rates of extinction per million species years. Thus, marine gastropods typically last 10 million years, so a sample of 1 million marine gastropod species would be expected to experience 0.1 extinctions per year.

B. Variation within Taxonomic Groups

The distributions of species lifetimes within higher taxa are typically right-skewed. Most species have short durations, but a small minority are long-lasting (Fig. 1). There could be two causes for this pattern. First, only a few species might have characteristics that make them intrinsically resistant to extinction, whereas most are sensitive to extinction risk. Second, species might not vary in their susceptibility to extinction, but if many independent factors can cause extinction only a small minority of species will be lucky enough to avoid extinction for long. Probably both factors play a part in shaping distributions of species durations.

C. Risk of Extinction in Relation to Species Age

Does the likelihood that a species will go extinct depend on its age? One might think that it should because the longer a species stays in existence the more adaptations to its environment it can accumulate; older species should therefore be better at avoiding extinction. This idea can be tested by examining species survivorship curves. A species survivorship curve shows the relationship between age and the proportion of species in a taxon surviving to each age. Such curves are typically log-linear; that is, when the proportion surviving is plotted logarithmically its decline with age approximates a straight line. This shows that the proportion

![Figure 1: Species durations of Silurian graptolites of the British Isles (From Macroevolution: Pattern and Process, Stanley, S. M. © 1979 by W. H. Freeman and Company. Used with permission.)](image)
of species going extinct is approximately constant with respect to age. The constancy of extinction probability with species age can be explained by assuming that the environment of any species is continually changing so that even if the species continually adapts, its fitness lags behind the current condition of the environment. Thus, no matter how long a species lasts it can never reach a condition of optimal adaptation. There are two forms of this explanation. The first assumes that the probability of extinction for a species is determined primarily by its interactions with other species and that each species is under constant selection pressure to increase its fitness relative to these other species. However, adaptations in one species that increase abundance at the expense of other species will be met by counteradaptations from them so that, on average, no species improves its fitness with time. This is the Red Queen hypothesis, named after the character in Lewis Carroll’s Alice Through the Looking Glass in whose world “it takes all the running you can do, just to stay in one place.” The Red Queen hypothesis predicts that extinction rates will be approximately constant with respect to absolute time. The second form of the explanation assumes that it is the abiotic environment that is continually changing. Changes in the abiotic environment will tend to be episodic and will therefore produce fluctuations in extinction rates with time, but their effects will still be independent of species age. It is not clear which form of the hypothesis is closer to the truth.

IV. THE SELECTIVITY OF EXTINCTION

Extinction may be largely a matter of the bad luck of environmental change, exacerbated perhaps by pressure from other species; however, the fact that species durations vary so widely among higher taxa suggests that biological characteristics of species influence their susceptibility to extinction. Three types of studies have helped to identify these characteristics.

First, the attributes of species that have short versus long durations in the fossil record can be compared. A common variant of this technique compares species that persisted through mass extinction events with those that did not. Second, the living faunas of “land-bridge islands” can be compared with their presumed original faunas. A land-bridge island is an island formed as a result of isolation from the mainland by rising sea levels during the Pleistocene. Such islands typically have fewer species than do similar areas of mainland habitat to which they were once connected, reflecting extinctions caused by the pressure of reduced habitat area.

The following traits consistently emerge from many studies of the selectivity of extinction: rarity, dispersal ability, body size, and specialization.

A. Rarity

The commonness or rarity of a species is a function of two factors: its geographic range and its population density where it occurs. Both components of rarity influence extinction risk.

1. Geographic Range

There is strong evidence from the fossil record (mainly for marine invertebrates) that species with large ranges have lower extinction rates (Jablonski, 1995; McKinney, 1997). There are probably two causes for this effect. First, at a given scale, a disturbance will affect a smaller proportion of the range of a widespread rather than of a geographically restricted species. In the extreme, a single catastrophic event may wipe out a localized species but have little impact on a widespread species. Second, even if an environmental change affects a very large area, it is likely that some populations of widespread species will persist in isolated refuges that provide some local protection from the change. Species that hang on in such refuges may reinvent their original ranges once conditions improve. This is probably part of the explanation for the existence in the fossil record of “Lazarus species” that vanish during periods of environmental crisis and then reappear much later. The effect of range size on extinction risk tends to be strongest for background extinctions, and it may weaken or disappear in mass extinctions. This seems to have been the case for marine mollusks, for example (Jablonski, 1991), and is probably because mass extinctions were caused by events that affected such large areas and had such profound impacts that even widespread species were susceptible to them.

2. Local Abundance

Population density is generally not well represented in the fossil record, but patterns of extinction of populations on land-bridge islands and in the present day show that local extinction is more likely for species
with low population densities. For example, Foufopoulos and Ives (1999) found that reptile species with low population densities were more likely to go extinct from land-bridge islands in the Mediterranean Sea. The causes of extinction of small populations have been widely discussed in the literature on conservation. Briefly, small populations are more vulnerable than large populations because (i) they may go extinct more quickly when chance environmental variation causes fluctuations in abundance; (ii) they are affected by chance demographic fluctuations, such as occasional production of biased sex ratios of offspring, that would be averaged out in large populations; (iii) they may lose genetic variation and experience high levels of inbreeding and inbreeding depression; and (iv) they may be subject to Allee effects—that is, social or reproductive dysfunction as a direct result of low numbers.

3. The Relationship between Range and Abundance

The components of rarity—geographic range size and population density—are generally positively related among species: Species with small geographic ranges also tend to have low population densities, and wide-ranging species have high population densities. This pattern has been found in many terrestrial plant and animal taxa, although it remains almost unstudied in marine organisms. Why there should be a positive relationship between range and abundance is not clear. Several ecological mechanisms have been proposed as its cause, but two ideas have been especially influential. First, Brown (1995) argued that niche breadth is positively correlated with both range size and population density because species with broad niches can exist under a wider range of conditions and use more types of resources than can species with narrow niches; variation among species in niche breadth therefore produces a positive relationship between range and density. Second, species that reach high local densities should be both more resistant to extinction on habitat patches because their populations are larger and produce more migrants that are able to recolonize habitat patches after local extinctions (Hanuki, 1999). High local abundance therefore results in wide geographic distribution.

To the extent that geographic range and local abundance affect extinction risk independently of one another, the correlation between them should exaggerate differences among species in extinction risk. Rare species face double jeopardy: The vulnerability that comes from having a small range is compounded with the vulnerability due to low abundance. Common species, on the other hand, are likely to be highly resistant to extinction because they combine high abundance with large ranges. For example, Johnson (1998) showed that among Australian marsupials, ancient species (those that diverged from their closest relative more than 4 million years ago) are very unlikely to have both small ranges and low population densities, although this combination occurs frequently among young species (Fig. 2). This suggests that species with low abundance and small ranges tend to be short lived. There are ancient marsupials with small ranges but they have unusually high densities, and conversely there are species with low densities but they have unusually large ranges,
implying that a high density can compensate for the vulnerability that comes from having a small range and vice versa.

4. Spatial Structure of Populations
Another aspect of rarity is ubiquity or patchiness of distribution within the geographic range. The distributions of practically all species are discontinuous at some scale, but the degree of patchiness varies, reflecting specialization for habitats or resources that are themselves patchily distributed. In a species that has a patchy distribution, metapopulation dynamics are likely (Hanski, 1999). That is, the risk of extinction of populations on discrete patches of habitat may be high, but local populations are eventually reestablished by migrants from other populations. Recolonization will involve a time lag that depends on the distance between patches and the dispersal ability of the species.

A general deterioration in the quality of habitat for a species is likely to cause contraction of larger habitat patches and the disappearance of smaller ones, causing local extinctions and simultaneously increasing the distance between surviving patches and thus reducing the probability of migration. Patchily distributed species should therefore be more likely to go extinct as a result of environmental change than should continuously distributed species. The sensitivity of patchy species to environmental change has been demonstrated for insects in Great Britain subject to human-caused changes (Webb and Thomas, 1994), but it is likely to be general to any form of shift in environmental conditions. Patchiness of distribution is loosely correlated with the other components of rarity so that widespread and locally abundant species tend to be continuously distributed within their ranges, whereas rare species are likely to be patchy.

B. Dispersal Ability
Dispersal ability is positively associated with species longevity for two reasons. First, species that disperse widely tend to have large geographic ranges, and their resistance to extinction can be a direct result of large range. Marine gastropod snails, for example, have two different forms of development. In some, the egg is released into surface waters and develops into a larval form that feeds in the plankton and drifts widely before settling and developing into an adult snail. This form of development promotes wide dispersal. Others have direct development in which eggs and young grow up close to the parent. Species with planktonic development have larger geographic ranges and longer durations in the fossil record than do species with direct development (Jablonski, 1995). Second, species that disperse widely are better able to recolonize after going extinct from part of the range, and they can more rapidly shift their ranges to track changes in the distribution of their preferred climates and habitats. During the Quaternary in the Northern Hemisphere, the distribution of habitats changed very rapidly as ice sheets repeatedly advanced and receded. These changes produced remarkably few extinctions among Northern Hemisphere snails. Instead, species shifted their geographic ranges to keep pace with the changing distribution of habitats, an effect that was more dramatic in flighted than in flightless species (Coope, 1995). Probably, dispersal ability and local abundance interact to confer high resistance to extinction because high local abundance means that large numbers of dispersers are produced, resulting in the potential for rapid shifting of range boundaries as conditions change. This combination of characteristics might protect species that have specialized habitat requirements and small geographic ranges and would otherwise be vulnerable to extinction.

C. Body Size
It often seems to be the case that large-bodied species are at higher risk of extinction than small-bodied species. Although there are exceptions, this pattern is reasonably consistent in the fossil record both for background extinctions and for mass extinctions (McKinney, 1997), and it also emerges in extinctions on land-bridge islands (Brown, 1995). There are several reasons why large-bodied species might be particularly vulnerable to extinction:

1. Potential rate of population increase declines with body size because large-bodied species generally have longer generation times and lower fecundity than do small-bodied species. This means that large-bodied species will generally recover more slowly from population declines, and because they produce fewer offspring they may be slower to track habitats than will small-bodied species.
2. Individuals of large-bodied species generally need larger areas, and large-bodied species are therefore strongly affected by declines in habitat area.
3. Population density tends to decline with body size so that large species are often naturally rare. This is partly compensated by a tendency for size of geographic range to increase with body size, but this relationship is often weak or absent so that total population size is
usually much less for large than small species. The decline in population density with body size is clearest when species of very different size (e.g., mice and elephants) are compared, but in guilds of ecologically similar species it is often the case that density increases with body size, possibly because larger species are better competitors for resources (Cotgreave, 1993). At this finer scale of comparison, therefore, large-bodied species might be more resistant to extinction than small-bodied species.

D. Specialization

Species can be classed as "specialists" or "generalists" on three criteria: the range of conditions that they are adapted to tolerate, the range of resources that they are able to use, and the degree of their evolved dependence on a small number of other species.

1. Conditions

Species that are narrowly adapted to environmental conditions are likely to be the first to go extinct when the environment changes. This is difficult to demonstrate in the fossil record, however, because it is not possible to measure directly the environmental tolerances of fossil species. Instead, environmental tolerances are inferred from geographic distributions. Species with small geographic ranges will necessarily occupy a narrow range of climate zones and habitats, but it does not follow from this that their environmental tolerances are narrow. Some species that could potentially be widely distributed have small ranges because of population history, geographic barriers to movement, or poor dispersal ability or because range expansion is prevented by interactions with other species (competitors, predators, and so on). Therefore, although there is abundant evidence from the fossil record that species with small geographic ranges are extinction-prone, there is much less evidence for the commonsense view that the breadth of tolerance of environmental conditions directly affects extinction risk.

2. Resources

Species that depend on a narrow range of types of resources are likely to be more sensitive to changes in resource abundance than generalists that can easily switch resources. This vulnerability may be partly compensated for by the fact that specialization on a particular resource may be more likely to evolve if that resource is abundant and widespread. For example, feeding on grasses by mammals requires extensive specialization of the teeth and digestive system, and increases in body size, to overcome the abrasiveness and poor nutritional quality of grasses. However, because of the abundance of grasses many mammals have evolved these adaptations, and grazing mammals typically have high local abundance and large geographic ranges. Nonetheless, grazing mammals in Africa have suffered higher extinction rates since the Miocene than mixed grazers/browsers, as the extent of grasslands has fluctuated during the Pliocene and Pleistocene (Vrba, 1987).

3. Interactions

Specialization is taken a step further in species that have evolved a close dependence on one or a small number of other species. For example, many herbivorous insects feed on only one species of plant, and many predators attack only one or a small number among many possible species of prey. Of particular interest are mutualistic interactions, in which species provide benefits to one another. Figs, for example, rely on fig wasps for pollination, and fig wasps in turn lay their eggs only in the flowers of figs. This interaction tends to be highly species specific, with each species of fig visited by only one species of fig wasp and each partner in the interaction completely dependent on the other for reproduction. Such tight species specificity results from coevolution, in which each species in the interaction evolves special characteristics in response to evolutionary changes in the other to increase its benefit from the interaction.

Specialization of this kind is classically regarded as an extinction trap because the specialist will inevitably go extinct if the species that it depends on goes extinct or becomes very rare. This view is probably an oversimplification. Careful study of some species-specific interactions has revealed more flexibility and greater potential for rapid evolutionary response to changes in the abundance of interacting species, including the ability to switch to new partners, than was previously assumed (Thompson, 1994). Also, mutualistic interactions have the general effect of increasing the geographic range and abundance, and stabilizing the population dynamics, of both partners in the interaction. These ecological benefits may at least partially compensate for the vulnerability caused by dependence on another species. Because interactions between species are not revealed in detail in the fossil record, there is little direct information on rates of secondary extinction. It is sometimes possible, however, to evaluate the risks of secondary extinction from study of living communities. Many plants cannot set seed without cross-pollination and rely on animals to transfer pollen. Such plants should be vulnerable to reproductive failure, and possibly ex-
NATURAL EXTINCTIONS (NOT HUMAN INFLUENCED)

313

FIGURE 3 The relationship between an index of dependence on animal pollinators for seed set and dependence on seeds for regeneration for species of spring flowering herbs in a temperate deciduous forest. Any species that was both dependent on pollinators for seed set and dependent on seeds to regenerate would be considered at high risk of extinction due to pollinator failure (redrawn from Bond, 1995).

E. Reproductive and Life History Traits

A wide variety of reproductive and life history traits may influence the likelihood of extinction in some circumstances. Perhaps the most important is sexual reproduction. Sexual lineages appear able to evolve more quickly than parthenogenetic ones because genetic recombination generates more variability among individuals on which natural selection can act. Sexual species may therefore be better able to adapt to changed environmental conditions or to meet new challenges from other species. The taxonomic distribution of parthenogenetic lineages suggests that they are short lived: With few exceptions, no higher taxa (genera, families, etc.) consists solely of parthenogens; almost all parthenogens have close sexual relatives and are likely to be recently derived from sexual species (Maynard Smith, 1978).

Species with resting or resistant stages in their life cycle may be better able to pass through environmental catastrophes than species in which all life stages are vulnerable to environmental changes. In the islands of the Lesser Antilles, for example, butterflies are better able to persist on small islands than are birds or mammals, perhaps because they are able to pass through unfavorable seasons as diapausing eggs and pupae. Similarly, reptiles and amphibians are resistant to extinction, possibly by virtue of their ability to survive for long periods in protected microenvironments without having to feed (Ricklefs and Lovette, 1999).

V. GEOGRAPHIC PATTERNS IN EXTINCTION

Natural extinction rates vary along two major geographic axes: Extinction rates are lower for marine than terrestrial organisms, and extinction rates may be higher near the equator than at high latitudes. The low extinction rates of marine organisms are presumably due to their large geographic ranges—the oceans are bigger and less subdivided than the continents—and the high dispersal ability of many species that have larval life stages that drift in the plankton. The effects of latitude on extinction rates are less distinct, but where extinction rates have been shown to vary with latitude they tend to be higher in the tropics. Tropical species often have smaller geographic ranges and lower population densities than species at higher latitudes and may thus be more vulnerable to the effects of environmental change (Lawton, 1995). Also, some particularly species-rich tropical environments, such as the low-sediment, low-nutrient, shallow-water platforms that support reef and related communities, are easily disrupted by changes in climate and sea level (Jablonski, 1995).

VI. NATURAL EXTINCTION AND TRENDS IN BIODIVERSITY

The diversity of any lineage of organisms through time reflects the balance of the rates of loss of species by extinction and gains by speciation. Because extinction rates vary among lineages, extinction is an important factor shaping patterns of biodiversity. Much of the explanation for the very high diversity of insects is that,
owing presumably to traits such as small size, high abundance, and high mobility, they have low extinction rates. Essentially no insect families have gone extinct during the past 50 million years, and some living genera and species with good fossil records are of Miocene age or older (Labandeira and Sepkoski, 1993). Species diversity of insects has therefore accumulated over long periods of time. Mammals, on the other hand, appear to evolve and speciate rapidly, but they are a minority taxon because their extinction rates are high. Insects and mammals provide an extreme example of contrasts in the balance of extinction and speciation rates, but in general extinction and speciation rates are positively correlated among different lineages (Fig. 4). Such a relationship could arise if total species number in a lineage is held at equilibrium by competition between species so that a new species can only establish by occupying a niche left vacant by a species that has gone extinct. However, the relationship was first shown for groups of animals undergoing rapid increase in diversity (Stanley, 1979). There are three possible causes of the correlation between rates of speciation and extinction:

1. The effects of dispersal and gene flow: Dispersal rates should be correlated with resistance to extinction, but the gene flow resulting from high dispersal should also prevent genetic divergence of populations, thus impeding allopatric speciation. This explanation may apply to marine gastropods during the Late Cretaceous, when lineages with planktonic larvae and large ranges not only were less likely to go extinct but also were less likely to produce new species than those with direct development and restricted dispersal (Hansen, 1983).

2. The effects of niche breadth: Ecologically specialized species are likely to be vulnerable to environmental change, but they also tend to have patchy distributions, and this spatial subdivision should promote divergent evolution of local populations and the generation of many species. Generalists may be resistant to extinction, but their lack of spatial population structure also impedes divergent evolution among local populations.

3. The effects of rarity: Species with small populations are prone to extinction. Somewhat controversially, it is believed that speciation may be enhanced by small or fluctuating population sizes (Futuyma, 1998).

Because extinction is selective for particular characteristics of species, it can act as a filter through which certain types of species are more likely to pass. Selective extinction therefore shapes the composition of ecological assemblages of species. This filtering effect can be seen most clearly in mass extinctions, which typically leave in their wake “survival faunas” dominated by small numbers of wide-ranging and abundant generalist species of a few simple ecological types. The examples shown in Figs. 2 and 3 illustrate some more subtle effects of the continuous operation of extinction filters, shaping respectively the patterns of distribution and abundance in living marsupials and the combinations of demographic characteristics of living plants.

In some cases, directional trends in evolution and the development of communities can be opposed or curtailed by selective extinction. This effect can produce taxon cycles, which occur if derived species tend to evolve characteristics that make them vulnerable to extinction. For example, species that colonize islands are often generalists that are abundant and competitively dominant and quickly become widespread. Their descendants, however, tend to become more specialized and less mobile, occupy fewer habitats, lose their competitive ability, and sustain smaller and more subdivided populations. These trends eventually lead to extinction and the repetition of the cycle as sets of derived species are replaced by newly invading generalists (Ricklefs and Miller, 1999). Taxon cycles are most visi-
ble on island chains, where ecological communities are relatively simple and the patterns of distribution of related species on different islands provide clues to the direction of evolution. However, similar processes probably operate over larger areas on continents and in the oceans.

What can the study of natural extinctions teach us about the coming wave of human-caused extinctions? There are at least two messages. First, species durations among different taxa in the fossil record are positively related to rates of endemism of living species in the same taxa (McKinney, 1997). This suggests that the characteristics that have made species vulnerable to extinction under natural conditions also make them sensitive to human impacts on the environment, and in the absence of better information these characteristics could be used to identify species likely to be at risk in the near future. Second, the selectivity of extinction means that extinction rates will be higher in certain ecological types and taxonomic groups than in others, and if environmental pressures continue for a sufficient amount of time some groups may be removed entirely. The result will be a much greater impoverishment of biodiversity than would be the case if extinctions were randomly distributed among species.

See Also the Following Articles

EXTINCTION, CAUSES OF • EXTINCTION, RATES OF • MAMMALS, LATE QUATERNARY, EXTINCTIONS OF • MAMMALS, PRE-QUATERNARY, EXTINCTIONS OF • MASS EXTINCTIONS, CONCEPT OF • POPULATION DENSITY • SPECIES INTERACTIONS

Bibliography

NATURAL RESERVES AND PRESERVES

Alexander N. Glazer
University of California at Berkeley

I. Origins of Protected Areas
II. Classification
III. Nature Alongside Humanity
IV. Research in the National Parks
V. Biological Field Stations
VI. The Long-Term Ecological Research Network
VII. The International Long-Term Ecological Research Network
VIII. The International Long-Term Ecological Research Network
IX. Sustained Ecological Research
X. GAP Analysis

GLOSSARY

endemic Confined to a particular region or locality; native (not introduced).
hectare A metric measure of surface equal to 10,000 square meters or 2.471 acres (U.S.).

NATURAL BIOLOGICAL COMMUNITIES, undisturbed by humans, are very valuable. Our understanding of ecological processes is still rudimentary and areas in which natural communities are protected are irreplaceable sites for the scientific study. In total, such areas also make a very important contribution to slowing the loss of the earth’s biological diversity. Protected sites serve as ecological and baseline control areas where long-term studies provide early warnings of climatic change and of undesirable outcomes of human activities such as acid rain. There are many different types of protected areas: national parks, refuges, and sanctuaries, both terrestrial and marine. There are also numerous biological field stations whose primary mission is to support university-level research and instruction. This article describes protected areas as well as such research sites. Examples are also given of the types of scientific research carried out in the various protected areas and in field stations and of the contribution of research to the understanding and management of the biosphere.

I. ORIGINS OF PROTECTED AREAS

Preserves, such as sacred groves or royal hunting parks, were established in ancient times, but protection of natural habitats in those areas was incidental. Preservation of natural biological communities for their own value originated in the United States relatively recently. The world’s first national park, Yellowstone, was established in 1872. It occupies more than 8000 km² mainly in the northwestern corner of Wyoming, and it extends into Idaho and Montana. The Yellowstone Act of 1872 reflected the contemporary faith in the inexhaustibility of natural resources. The act prohibited the wanton destruction of fish and game for profit, but it permitted hunting, trapping, and fishing for recreation or to provide food for park residents or visitors. Soon after, several other nations established national parks: Australia (Royal National Park, 1879), Canada (Banff National...
Park, 1885), and New Zealand (Tongariro National Park, 1897). There are currently approximately 1,500 national parks in the world, which are among the most popular tourist destinations. In 1990, more than 275 million people visited the national parks in the United States alone.

In Africa, and to a lesser extent in the Indian subcontinent, steps toward conservation during the colonial era emphasized large mammals, particularly those which had been intensively hunted. For example, Jim Corbett National Park, a haven for tigers in the Himalayan foothills of Uttar Pradesh, was established in 1936. It is named in honor of Jim Corbett, a well-known hunter-naturalist, who became an author and a photographer and helped define the park’s boundaries. In contrast, in long-settled parts of Europe, protection was extended to humanized landscapes.

Protection has also been given to mangroves and coral reefs. Mangroves are found in the tropics and subtropics on river banks and along coastlines. Mangrove plants include trees, shrubs, ferns, and palms, which have adapted to the largely anaerobic conditions of muddy, aquatic environments by producing stilt roots which project above the water to absorb oxygen. Approximately 900 protected areas include mangroves. Coral reefs, many formed over periods of thousands of years, cover between 300,000 and 600,000 km² mainly between the Tropics of Cancer and Capricorn. The Great Barrier Reef Marine Park off the coast of Queensland, Australia, covers 349,000 km² of reefs and surrounding waters.

The diverse national parks, reserves, and preserves throughout the world, now referred to as “protected areas,” comprise more than 3% of the earth’s surface.

II. CLASSIFICATION

The International Union for the Conservation of Nature (IUCN) has defined a protected area as “an area of land and/or sea especially dedicated to the protection and maintenance of biological diversity, and of natural and associated cultural resources, and managed through legal or other effective means.” IUCN noted that the purposes for which particular protected areas are managed vary widely and noted specifically (i) scientific research, (ii) wilderness protection, (iii) preservation of species and genetic diversity, (iv) maintenance of environmental services, (v) protection of specific natural and cultural features, (vi) tourism and recreation, (vii) education, (viii) sustainable use of resources from natural ecosystems, and (ix) maintenance of cultural and traditional attributes. IUCN classified protected areas into the six categories listed in Table I reflecting the management purpose. Note that IUCN categories IV–VI encompass areas where limited or extensive use of biological resources is allowed. The following discussion of the characteristics of representative protected areas belonging to each of the categories in Table I clarifies the distinctions.

The subantarctic Snares Islands are located 209 km southwest of Bluff on South Island of New Zealand. The Snares are classified as la (Strict Nature Reserve) because the vegetation is nearly pristine and the islands are free of introduced mammals. The total area of the Snares Islands Nature Reserve is 328 ha, covering the islands and their foreshores. The islands are home to millions of breeding seabirds, including the endemic Snares crested penguin, and are strictly protected. Research and monitoring of natural plant and animal communities that does not cause long-term disturbance or damage is permitted. There is no access for tourists.

NorthEast Svalbard Nature Reserve, owned by the Norwegian government, exemplifies category 1b (Wilderness Area). The reserve includes the islands of Nordaustlandet, Kvitoya, and Kong Karls Land and the surrounding territorial waters, a total of 1.9 million ha, in the northeast of the Svalbard archipelago. The archipelago extends approximately from 74° to 81°N and 10° to 35°E, an Arctic area about one and a half times the size of Switzerland. This wilderness is covered by high arctic tundra vegetation, with many species near their northernmost distribution in Europe. The large mammals include polar bear, reindeer, and walrus. Access is limited to nonintrusive scientific research and recreation.

Canaima National Park (category II, National Park), owned by the Venezuelan government, covers 1 million ha centered on the Guayana Shield in Bolivar State, south of the Orinoco River. Canaima includes the uplands of the Grand Sabana and the eastern tepuis of the Roraima Range. The tepuis are flat-topped mountains with almost vertical slopes. The world’s highest waterfall, Angel Falls, with a fall of 102 m, descends from one of these mountains. The summits of the tepuis have many endemic species. A small population of traditional occupants of the area, the Pemon, continues swidden (slash-and-burn) agriculture, hunting, and gathering. The park preserves representatives of the geological, biological, and cultural features specific to the Guayana Shield. Hunting and collecting of wildlife is forbidden. Recreational activities, research, and education are encouraged.

Ngerukewid Islands Wildlife Preserve (category III,
TABLE I

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Approximate number</th>
<th>Approximate area (ha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ia</td>
<td>Strict Nature Reserve: managed mainly for science</td>
<td>1460</td>
<td>86,474,000</td>
</tr>
<tr>
<td>Ib</td>
<td>Wilderness Area: managed mainly for wilderness</td>
<td>2040</td>
<td>376,784,000</td>
</tr>
<tr>
<td>II</td>
<td>National Park: managed mainly for ecosystem protection and recreation</td>
<td>250</td>
<td>13,686,000</td>
</tr>
<tr>
<td>III</td>
<td>Natural Monument: managed mainly for conservation of specific natural features</td>
<td>3810</td>
<td>308,314,000</td>
</tr>
<tr>
<td>IV</td>
<td>Habitat/Species Management Area: managed mainly for conservation through management intervention</td>
<td>2275</td>
<td>141,092,000</td>
</tr>
<tr>
<td>V</td>
<td>Protected Landscape/Seascape: managed mainly for conservation and recreation</td>
<td>Not available</td>
<td>Not available</td>
</tr>
<tr>
<td>VI</td>
<td>Managed Resource Protected Area: managed mainly for the sustainable use of natural ecosystems</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* See http://www.wcmc.org.uk for updated information and information by country.

National Monument) is an archipelago of high limestone islands sharing a common reef platform. The islands belong to Koror State, Palau. The reserve covers 1200 ha, of which 90 ha is terrestrial. These islands have never been inhabited and are home to many endemic and threatened species. Management is focused on protection of the virtually pristine habitat and on promoting research and public education. Tourist visitation is frequent, but there are no facilities.

Situated in southern Tanzania, the 5 million-ha Selous Game Reserve (category IV, Habitat/Species Management Area) is one of the largest protected areas in Africa, with altitudes ranging from 100 to 1200 m. The reserve protects the world’s largest area of miombo woodland. This woodland requires a particular burning regime for its preservation. Selous is famous for its elephant, hippopotamus, and black rhinoceros populations, although only a few of the rhinoceros remain. Sizable populations of other large animals include buffalo, Nyasaland gnu, brindled gnu, hartebeest, Greater Kudu, sable antelope, eland, reedbuck, bushbuck, waterbuck, wart hog, zebra, giraffe, and wildebeest. Predators include lion, leopard, spotted hyena, hunting dog, and a small population of cheetah. There are more than 350 species of birds and reptiles, including crocodiles. Ecological monitoring and research provide information for the quotas for regulated sport hunting, the main use of the reserve.

Dartmoor National Park (category V; Protected Landscape/Seascape) was designated one of the national parks of England and Wales in 1951. It is a moorland landscape with tors of exposed granite and wooded valleys. Approximately 31,000 people live within the 91,300-ha area of the park. There are approximately 10 million visits for recreation each year. Dartmoor is a rich habitat for wildlife and includes lowland farming areas of meadow, pasture, and woodland, deep valleys, and upland moors. There is a wealth of archaeological remains. Management of the area aims to ensure preservation of the natural habitat, the cultural heritage of the area, and continuation of traditional agricultural uses.

Tamshiyacu-Tahuayo Communal Reserve (category VI, Managed Resource Protected Area), established in 1990, is an area of 332,500 ha of Amazonian rain forest in the state of Loreto in northeast Peru and encompasses a rich flora and fauna. The reserve is divided into three land zones: a totally protected core, a subsistence zone, and an area of permanent settlement. More than 6000 people exploit the resources of the subsistence zone and the bordering areas. These people, known as ribereños, depend on hunting, fishing, swidden agriculture, and gathering of nontimber plant products.

III. NATURE ALONGSIDE HUMANITY

Do the protected areas represent adequate “Noah’s Arks” for Earth’s currently known biological diversity? Regrettably, this question must be answered in the negative. Natural communities vary in time and space and are vulnerable to introduced animals and plants. Moreover, existing protected areas cover only a small fraction of biological diversity. Many of these areas are small and isolated and lose species over time. Many
places designated as "protected areas," even those in IUCN categories I–III, are either inadequately protected or not protected at all and are referred to as "paper parks." Such parks are subject to rapid exploitation for subsistence or commercial gain. Conflicts between conservation and exploitation of natural resources impact the establishment of new protected areas and affect existing ones. Finally, even adequately protected areas suffer impacts from visitation and from changes that take place in adjoining lands. Glimpses of these global problems are offered here.

There are approximately 7000 nationally protected areas in more than 125 nations encompassing ~5% of the earth’s surface. Table I provides an estimate of the approximate number of sites and of the area worldwide protected within each of the six IUCN categories, whereas Table II indicates the percentage of land area protected for countries in which that area exceeds 10%. Many other nations have stated their intention to protect 10% or more of their land by the Year 2000. Even the realization of such goals is unlikely to stem the rapid loss of critical habitat. Protected areas are increasingly becoming islands encircled by intensively exploited landscapes.

The well-founded concern about lack of appropriate representation of biological diversity in reserves is illustrated by a few examples. Consider the Chilean National Park system. This system, first established in the early 1900s, was one of the earliest in Latin America. In the temperate forest region of south-central Chile, parks and reserves protect 13 million ha, 29% of the land in this region. However, more than 90% of the protected land is at high latitudes (>43°), largely outside the areas of the highest tree species richness, endemic woody genera, and maximum species richness of native mammals, amphibians, and freshwater fishes, which lie in the region of 33.6° to 41.3°S. Areas of high endemism and species richness in the latter region correspond with areas of high human population density and intense land use. Moreover, forest protected in the reserves in the 35.6–41.3°S region lies at high Andean locations (above 600 m), often on large active volcanoes. This example highlights a worldwide problem. Reserves larger than 100,000 km² protect species-poor high mountain, tundra, and the driest desert areas.

Utilization of natural resources, particularly in the poorer regions of the world, takes precedence over protection of habitat. In the Philippines, for example, the market demand for hardwoods and plywood led to a high rate of deforestation, less than one-fourth of the rain forest remains and <1% of the original rain forest area is officially preserved. Even for countries with a

<table>
<thead>
<tr>
<th>Country</th>
<th>Area protected (ha)</th>
<th>% land area protected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antigua and Barbuda</td>
<td>6,128</td>
<td>13.88</td>
</tr>
<tr>
<td>Australia</td>
<td>93,943,357</td>
<td>12.18</td>
</tr>
<tr>
<td>Austria</td>
<td>2,059,473</td>
<td>23.92</td>
</tr>
<tr>
<td>Belize</td>
<td>323,121</td>
<td>14.07</td>
</tr>
<tr>
<td>Bhutan</td>
<td>3,666,320</td>
<td>20.72</td>
</tr>
<tr>
<td>Botswana</td>
<td>10,642,200</td>
<td>18.54</td>
</tr>
<tr>
<td>Brunei Darussalam</td>
<td>113,133</td>
<td>19.97</td>
</tr>
<tr>
<td>Cameroon</td>
<td>2,997,730</td>
<td>16.36</td>
</tr>
<tr>
<td>Chile</td>
<td>13,723,323</td>
<td>18.26</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>638,364</td>
<td>12.53</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>1,096,808</td>
<td>13.53</td>
</tr>
<tr>
<td>Denmark</td>
<td>1,388,750</td>
<td>36.24</td>
</tr>
<tr>
<td>Dominican Republic</td>
<td>1,048,284</td>
<td>21.64</td>
</tr>
<tr>
<td>Ecuador</td>
<td>11,113,893</td>
<td>24.08</td>
</tr>
<tr>
<td>France</td>
<td>9,601,486</td>
<td>10.30</td>
</tr>
<tr>
<td>Germany</td>
<td>9,149,702</td>
<td>25.77</td>
</tr>
<tr>
<td>Greenland</td>
<td>98,230,000</td>
<td>44.93</td>
</tr>
<tr>
<td>Israel</td>
<td>307,833</td>
<td>14.82</td>
</tr>
<tr>
<td>Lao P.D.R.</td>
<td>2,440,000</td>
<td>10.50</td>
</tr>
<tr>
<td>Latvia</td>
<td>774,724</td>
<td>12.16</td>
</tr>
<tr>
<td>Liechtenstein</td>
<td>6,000</td>
<td>37.50</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>36,000</td>
<td>13.93</td>
</tr>
<tr>
<td>Malawi</td>
<td>1,058,800</td>
<td>11.23</td>
</tr>
<tr>
<td>Namibia</td>
<td>10,217,777</td>
<td>12.40</td>
</tr>
<tr>
<td>New Zealand</td>
<td>6,147,794</td>
<td>23.19</td>
</tr>
<tr>
<td>Norway</td>
<td>9,356,312</td>
<td>17.09</td>
</tr>
<tr>
<td>Oman</td>
<td>3,736,250</td>
<td>13.74</td>
</tr>
<tr>
<td>Panama</td>
<td>1,326,352</td>
<td>16.69</td>
</tr>
<tr>
<td>Rwanda</td>
<td>327,000</td>
<td>12.42</td>
</tr>
<tr>
<td>S. Korea and N. Korea</td>
<td>2,612</td>
<td>10.00</td>
</tr>
<tr>
<td>S. Yemen and the Seychelles</td>
<td>11,803,979</td>
<td>72.36</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>793,953</td>
<td>12.13</td>
</tr>
<tr>
<td>Switzerland</td>
<td>730,727</td>
<td>17.70</td>
</tr>
<tr>
<td>Taiwan</td>
<td>426,997</td>
<td>11.54</td>
</tr>
<tr>
<td>Tanzania</td>
<td>13,089,473</td>
<td>14.78</td>
</tr>
<tr>
<td>Thailand</td>
<td>7,020,276</td>
<td>13.66</td>
</tr>
<tr>
<td>Togo</td>
<td>646,906</td>
<td>11.39</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>9,127,966</td>
<td>20.94</td>
</tr>
<tr>
<td>United States of America</td>
<td>104,238,016</td>
<td>11.12</td>
</tr>
<tr>
<td>Venezuela</td>
<td>26,322,306</td>
<td>28.86</td>
</tr>
</tbody>
</table>

*IUCN 1996 global protected areas summary statistics.

4 Values may be inflated in some instances due to the inclusion of marine protected areas in the calculation (e.g., Ecuador, which includes the Galapagos marine reserve).
relatively high per capita income, detailed biota surveys and resulting conservation plans do not guarantee appropriate representation of biological diversity in protected areas. Such detailed planning was carried out during the 1990s for the diverse forest ecosystem of northwestern New South Wales in Australia. However, extensive logging is to take place in this area and land to be included in reserves is mostly in unloggable escarpment forest well represented elsewhere. Struggles between development and preservation are commonplace. In South Africa, Greater St. Lucia Wetlands Park may become the site of a titanium mine; construction of a uranium mine is under way in Australia's Kakadu National Park.

An increasing number of parks appears on the IUCN’s list of threatened protected areas of the world. Many of the threats arise from internal use and external development. The density of automobile traffic in Yosemite valley approaches that on urban freeways. In the Everglades, decades of water diversion have greatly reduced the wetlands, with a decrease in the number of wading birds by approximately 90% since the 1900s. Some protected areas have sustained long-term, or even irreparable, damage from massive accidental contamination, as illustrated by a recent example from Spain. The 200,000-ha Guadalquivir ecosystem, the largest wetlands in Europe, encompasses the World Heritage site of Doñana National and Natural Parks. This is one of the most important bird breeding and overwintering sites in Western Europe. A dam on a massive tailings pond used by a zinc mine on one of the tributaries of the Guadalquivir River collapsed in April 1998 releasing approximately 5 million cubic feet of acid sludge from the processing of pyrite ore. The waste entered ecologically sensitive areas of the park, resulting in massive contamination of the wetlands with high concentrations of lead, zinc, arsenic, and other heavy metals. The accident caused considerable fish and invertebrate kills and an impact of unknown extent on the protected bird species. High-level heavy metal contamination is likely to persist for many years.

**IV. RESEARCH IN THE NATIONAL PARKS**

Each year, 1000 scientific papers or more report results of research projects which examine questions concerning natural communities or particular organisms in national parks throughout the world. Surveys of flora and fauna, as well as of microbial communities, continue to reveal the existence of new species. Many studies address narrow questions specific to a particular park or species. For example, the Bwindi-Impenetrable Great Ape Project studies the ecological relationship between the mountain gorillas (*Gorilla gorilla beringei*) and chimpanzees (*Pan troglodytes schweinfurthii*) in Bwindi-Impenetrable National Park in southwestern Uganda, the only forest in the world in which these two apes occur together. Kanha National Park, a wildlife conservation area in Central India, is home to the highly endangered swamp deer (*Cervus duvauceli branderi*). Analysis of the evolutionarily conserved repeat sequence motifs in the genome of the swamp deer at Kanha National Park indicates a high level of genetic homogeneity that may underlie the gradual extinction of this species.

In the United States, the National Park Service has focused on six issues that threaten park resources: loss of ecosystem integrity and aesthetic degradation, polluted air, altered water quality or quantity, resource consumption, invasion by nonnative species, and visitor impacts. As demonstrated in the examples that follow, much of the research effort in the U.S. national parks addresses such concerns.

Until the latter half of the twentieth century, protection of forests from naturally occurring fires was considered essential to their preservation. A dramatic illustration of the impact of this policy is provided by the data for four Canadian national parks in which suppression of fires led to 10- to 100-fold decrease in burned area per decade relative to periods preceding fire suppression (Table III). Years of fire suppression have allowed unnatural, dangerously high accumulations of fuels as well as shifts in the proportion of fire-tolerant to fire-intolerant tree species. Research in Yosemite and in Sequoia/Kings Canyon National Parks strongly supported the need to reintroduce fire into these ecosystems.
past 30 years have seen the introduction of a closely monitored prescribed burn regime. This is a clear example of a major change in management driven by the outcome of scientific research.

Removal of predators to protect ‘desirable’ prey species was among the first resource management actions in the U.S. national parks. This practice was driven by an a priori belief-based consensus and was ultimately abandoned because of objections based on scientific research. The 30-year study in Isle Royale National Park of the interaction of wolves (Lupus canis), moose (Alces alces), and the vegetation on which moose feed was a particularly influential multifaceted study which demonstrated the long-term stability of this predator–prey system.

Mammoth Cave National Park lies in a classic karst terrane, an irregular limestone region, with sinks, underground streams, and caverns. Concerns about pollution of water at Mammoth Cave led to an extensive hydrological research program. An essential element of this program was the development of methods to measure and monitor conduit flow characteristics within the park and the surrounding area. This research documented the existence of an extensive underground conduit network through which water could flow several miles in a few hours. The source of pollution was shown to be drainage of untreated or inadequately treated sewage from communities well outside the park into the underground water system that enters the cave system. The hydrologic research influenced sewage treatment both within the park and in communities outside the park. This example provides another illustration of the critical importance to management of understanding the relationships between protected areas and adjacent lands.

V. BIOLOGICAL FIELD STATIONS

The several hundred field stations throughout the world are located in a wide variety of natural settings and explore a huge array of topics in the field sciences. Most are supported by or affiliated with universities, whereas others are operated by governmental organizations or by private foundations and institutes. Field stations are outdoor laboratories, the bases of research and teaching in environmental sciences. Field stations also fulfill important long-term monitoring functions. In a particularly apt description, these sites have been termed “sense organs” in nature. Descriptions of many individual stations and of their activities are available on the World Wide Web (for links, see http://www.capital.net/com/)

The research at a particular field station is strongly dependent on the biogeographic features of the area around the station, but it is also strongly influenced by the size of the station, the past history of the research at the site, and the interests of the researchers at the institution(s) with which it is affiliated. As the following examples show, research at some field stations covers a very broad spectrum of fields. Other stations support highly specialized research. The majority lie in-between.

Kristineberg Marine Research Station, the largest field station for marine research in Sweden and one of the oldest such stations in the world, is located on the west coast of Sweden, at Fiskebäckskil, 120 km north of Gothenburg. The station, founded in 1877 by the Royal Swedish Academy of Sciences, is operated by the academy and by Göteborg University. Located on the shore of a protected bay, a nature reserve within the Gullmarsfjord, the station has easy access to coastal and offshore marine habitats—sand and mud flats, steep rocky cliffs, and deep basin sediments. The brackish surface waters in the bay originate from the Baltic and the deep oceanic waters from the North Sea. These features endow the area with very high biodiversity.

Research at Kristineberg encompasses a broad spectrum of marine ecology and environmental research. Specific research areas include behavioral ecology, ben-thic ecology and monitoring, biological oceanography, ecophysiology, ecotoxicology, functional morphology, larval ecology, pelagic monitoring, physiology of macroalgae, plankton research, and trophic relationships in shallow coastal communities.

The Konrad Lorenz Research Station in the northern Alps of Austria, approximately 250 km west of Vienna, offers an example of a much narrower research program. The station was established by Konrad Lorenz in 1973 to continue his behavioral studies on the ecotology of social life in greylag geese (Anser anser) and other animals. Such research continues. Projects examine status dependence of within-flock competition, risk-sensitive foraging, and noninvasive behavioral endocrinology. Other research deals with the behavioral ecology and aspects of cognition in ravens (Corvus corax) and the functions and ecological roles of chemosensory organs in fish.

An intermediate situation is exemplified by the Alpine Research Center—Finse owned and operated by the universities of Bergen and Oslo. The center is 1222
m above sea level, approximately 250 m above the current tree line, on the northwestern end of Hardanger-vedda in south-central Norway (60°36′N, 7°30′E). Annual mean temperature is −2°C and the mean precipitation −1030 mm, much of which falls as snow. The focus of biological research at the center has been on the population dynamics of birds and small mammals and on alpine plant ecology. The center is also a base for geological and glaciological studies. Ongoing research includes a 30-year series of rodent trapping data, 13 years of glacier monitoring, as well as meteorological monitoring.

Field stations play an indispensable role in the training of undergraduate and graduate university students in a wide variety of disciplines. For example, the universities of Bergen and Oslo both conduct undergraduate and graduate-level courses at the Alpine Research Center–Fisne. Two of the longest running courses are alpine ecology and snow and winter ecology.

In the United States, the University of California Natural Reserve System manages 33 reserves that encompass more than 30,000 ha across 12 ecological regions in one of the most physiographically diverse regions in the United States. It is the largest university-operated reserve system in the world. Approximately 3500 University of California students utilize reserve system sites each year in courses in disciplines such as botany, entomology, zoology, geology, geography, meteorology, archaeology, paleontology, ecology, environmental planning, and wildlife management. When students from other educational institutions are included, the annual number increases to approximately 10,000. Hundreds of doctoral thesis projects worldwide depend on the use of the resources of a field station or on the use of a field station as a base of operations in a particular region.

VI. THE LONG TERM ECOLOGICAL RESEARCH NETWORK

The U.S. National Science Foundation initiated the establishment of the Long Term Ecological Research Network (LTER; http://www.lternet.edu) in 1980. The 21-site network (Table IV) carries out long-term studies of phenomena with broad spatial and temporal scales on the following core topics: (i) the pattern and control of primary production, (ii) the spatial and temporal distribution of populations selected to represent trophic structure, (iii) the pattern and control of organic matter accumulation in surface layers and sediments, (iv) the patterns and movements of inorganic inputs through soils and ground- and surface waters, and (v) the patterns and frequency of disturbances. A few examples provide a glimpse of the diversity of the LTER sites and of the broad range of research they make possible. Of the LTER sites described here, H. J. Andrews Experimental Forest, Coweeta Hydrologic Laboratory, and Konza Prairie Research Natural Area were among the initial group of six sites established in 1980.

The H. J. Andrews Experimental Forest LTER site in the Cascade Mountains of Oregon comprises a temperate coniferous forest biome. The main communities are Douglas fir, western hemlock, western red cedar, true fir, and mountain hemlock, interlaced with streams. The research centers on succession changes in ecosystems, forest–stream interactions, population dynamics of forest stands, patterns and rates of decomposition, carbon sequestration, and disturbance impacts on hydrologic response. The Arctic Tundra LTER site at Toolik Lake, Alaska, includes tussock and heath tundra, riverine willows, oligotrophic lakes, and headwater streams. The research includes studies of the movement of nutrients from stream to lake, changes due to anthropogenic influences, and control of ecological processes by nutrients and predation.

The Coweeta Hydrologic Laboratory LTER site in the southern Appalachian Mountains of North Carolina is an eastern deciduous forest biome. The research includes long-term hydrology, nutrient cycling, and productivity responses to management practices and natural disturbances (such as drought, flood, wind, and insects), impacts of atmospheric deposition on forest ecosystems, and physiological studies of carbon balance and competition.

The Konza Prairie Research Natural Area LTER site in Flint Hills, Kansas, is a tallgrass prairie biome. Research at Konza focuses on the effect of fire, grazing, and climatic variability. Data obtained by remote sensing and geographic information systems (GIS) are used to evaluate grassland structure and dynamics. The Sevilleta National Wildlife Refuge LTER site in central New Mexico lies at the intersection of diverse communities: montane mixed-conifer forest and meadows, riparian Rio Grande cottonwood forest, interior chaparral, Great Plains grasslands, Colorado Plateau shrub–steppe, Chihuahua Desert, juniper savanna, and pinyon–juniper woodlands. The primary emphasis in the research program is to examine long-term changes in ecosystem properties, such as population dynamics of plants and animals, nutrient cycling, hydrology, and productivity and species diversity, re-
TABLE IV

<table>
<thead>
<tr>
<th>LTER site</th>
<th>Location</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. J. Andrews Experimental Forest</td>
<td>Oregon</td>
<td>Temperate coniferous forest</td>
</tr>
<tr>
<td>Arctic Tundra</td>
<td>Alaska</td>
<td>Tundra, lakes, streams</td>
</tr>
<tr>
<td>Baltimore Ecosystem Study</td>
<td>Maryland</td>
<td>Urban and agricultural watershed</td>
</tr>
<tr>
<td>Botswana Creek Experimental Forest</td>
<td>Alaska</td>
<td>Taiga</td>
</tr>
<tr>
<td>Central Arizona–Phoenix Urban</td>
<td>Arizona</td>
<td>Sonoran desert scrub, urban environments, regulated river and floodplain</td>
</tr>
<tr>
<td>Cedar Creek Natural History Area</td>
<td>Minnesota</td>
<td>Eastern deciduous forest and tall grass prairie</td>
</tr>
<tr>
<td>Coweeta</td>
<td>North Carolina</td>
<td>Eastern deciduous forest</td>
</tr>
<tr>
<td>Jornada Experimental Range</td>
<td>New Mexico</td>
<td>Hot desert</td>
</tr>
<tr>
<td>Kellogg Biological Station</td>
<td>Michigan</td>
<td>Row-crop agriculture</td>
</tr>
<tr>
<td>Konza Prairie Natural Research Area</td>
<td>Kansas</td>
<td>Tallgrass prairie</td>
</tr>
<tr>
<td>Luquillo Experimental Forest</td>
<td>Puerto Rico</td>
<td>Tropical rain forest</td>
</tr>
<tr>
<td>McMurdo Dry Valleys</td>
<td>Antarctica</td>
<td>Polar desert oases</td>
</tr>
<tr>
<td>North Temperate Lakes</td>
<td>Wisconsin</td>
<td>Lakes, eastern deciduous forest</td>
</tr>
<tr>
<td>New Mexico Ridge</td>
<td>Colorado</td>
<td>Alpine tundra</td>
</tr>
<tr>
<td>Palmer Station</td>
<td>Antarctic</td>
<td>Polar marine</td>
</tr>
<tr>
<td>Plum Island Sound</td>
<td>Massachusetts</td>
<td>Estuarine</td>
</tr>
<tr>
<td>Scoville</td>
<td>New Mexico</td>
<td>Subalpine mixed-conifer forest/meadow, riparian forest, dry mountains, grassland, cold desert, hot desert</td>
</tr>
<tr>
<td>Shortgrass Steppe</td>
<td>Colorado</td>
<td>Shortgrass steppe</td>
</tr>
<tr>
<td>Virginia Coast</td>
<td>Virginia</td>
<td>Coastal barrier islands</td>
</tr>
</tbody>
</table>

VII. THE INTERNATIONAL LONG TERM ECOLOGICAL RESEARCH NETWORK

By 1998, 14 nations had initiated long-term ecological research programs (International Long Term Ecological Research Network; http://www.lternet.edu/ilter) with more than 200 research sites, modeled after the U.S. LTER: Brazil, Canada, China, China-Taipei, Costa Rica, Czech Republic, Hungary, Israel, Korea, Mexico, Poland, United Kingdom, Uruguay, and Venezuela. Other countries are working to establish similar programs.

VIII. SUSTAINED ECOLOGICAL RESEARCH

Establishment of long-term ecological research networks worldwide connotes a commitment to continuous research and monitoring to study long-term changes in ecosystems and to understand ecological processes that vary over long periods of time. One characteristic of such "sustained ecological research" is that it must last at least as long as the phenomenon under study or must be scaled to the frequency of the events being studied. The indispensability of sustained ecological research is well illustrated by considering some of the outcomes of 30 years of such research at the Hubbard Brook Experimental Forest (HBEF). HBEF, a 3160-ha reserve in the White Mountains of New Hampshire, was established by the U.S. Department of Agriculture Forest Service in 1955 and designated as a LTER site in 1988. The reserve represents a northern hardwood forest ecosystem with streams and lakes. HBEF has one of the most extensive and longest continuous databases on the hydrology, geology, chemistry, and biology of natural ecosystems in the world. Measurements at HBEF as early as 1964 were among the first to direct attention to acid rain in North America. More than 90% of the sulfur and nitrogen oxide emissions to the atmosphere in North America are anthropogenic and originate primarily from the combustion of fossil fuels by electrical power plants and from smelters. Sulfuric acid contributes approximately 65% and nitric acid approximately 33% to the acidity of the precipitation in the eastern United States. Studies
at HBEF revealed that dry deposition of acidic substances (gases and particles) is also a quantitatively important atmospheric input to aquatic and terrestrial ecosystems. Long-term monitoring at HBEF showed that both the hydrogen ion and sulfate concentrations in the precipitation gradually decreased from about 1970 on, after passage of clean air legislation by the U.S. Congress. These monitoring data provide invaluable measures of the outcomes of emission controls and reveal shortcomings. For example, the concentration of nitrate in the precipitation has remained virtually unchanged.

Dispersal of toxic metals, particularly lead, from anthropogenic activities is also an important environmental problem. In 1980, global emissions of lead from human activity were estimated at 2000 metric tons compared with approximately 6 metric tons derived from natural sources. In 1975, the concentration of lead in rain and snow at HBEF averaged 25 μg per liter, a concentration deemed unsafe in drinking water. The use of lead additives in gasoline was sharply restricted on species distribution and habitat is frequently lacking. Experience has shown that the available biological data for both existing reserves revealed that 10% of the ranges of endangered forest birds were within the reserves. This analysis led to the establishment of the 6693-ha Hakalau Forest National Wildlife Refuge in one of the areas of highest species richness. Gap analysis leads to natural resource management planning on a regional, national, or even global scale because it allows an integrated examination of species, habitats, and human ownership patterns and management practices over any desired area. The value of gap analysis depends on the quality of the data. Experience has shown that the available biological data for both public and private lands is generally fragmentary. Data on species distribution and habitat is frequently lacking. Finally, ready incorporation of biodiversity data into GIS requires that it be spatially referenced, a criterion not always met.

Such shortcomings can be addressed, and gap analysis has received widespread acceptance as the approach of choice to guide conservation planning. The desired outcome is that large-scale planning will lead to a world in which humans and the other
inhabitants of the biosphere live side by side in relative harmony.

See Also the Following Articles
CONSERVATION EFFORTS, CONTEMPORARY • EDUCATION AND BIODIVERSITY • FIRES, ECOLOGICAL EFFECTS OF • PREDATORS, ECOLOGICAL ROLE OF • ZOOS AND ZOOLOGICAL PARKS

Bibliography


NEAR EAST ECOSYSTEMS, ANIMAL DIVERSITY

Joseph Heller
The Hebrew University of Jerusalem

I. The Levant
II. The Extent of Biodiversity
III. Historic Zoogeography
IV. Endemic versus Widespread Animals
V. Factors Determining Diversity
VI. Freshwater Diversity: Historic Factors
VII. Biodiversity: The Human Impact

GLOSSARY
allopatric Describing two taxonomic entities whose geographic ranges do not intersect with each other.
endemism The fact of being found in a specific location of limited size, rather than being widely distributed.
eremic Relating to or occurring in desert regions.
parapatric Describing two taxonomic entities whose geographic ranges are distinct but adjacent to each other.
xeromorphic Describing a form that has developed in response to highly arid conditions.

THE ANIMAL FAUNA OF THE LEVANT (the eastern shoreland of the Mediterranean Sea) is exceptionally rich and diverse. To what extent is this high diversity due to geological history, climate, and substratum? What is the human impact on the fauna of the Levant? It is these and other questions that I address in this article.

I. THE LEVANT

The Levant is the eastern shoreland of the Mediterranean (Por, 1975), a stretch of land approximately 800 km long and approximately 150 km wide (Fig. 1). It is wedged between the Mediterranean Sea in the west and the Arabo-Syrian Desert in the east, stretching from the mouth of the River Orontes in the north to the Isthmus of Suez in the south. It consists of four basic north–south-oriented features: the coastal plain, the western (cis-rift) mountains, the rift valley, and the eastern (trans-rift) mountains. In the east it merges gradually into the Arabo-Syrian Desert.

The 800-km-long Levant has several climatic and floral belts. In its northern (and high-altitude) areas the overall habitat is Mediterranean: Summers are dry and warm, and winters are rainy and mild with occasional snow in the higher mountains. Natural plant associations consist of xerophytic shrubs and trees. In its southern and eastern areas (the Negev, Sinai, and fringes of the Syrian desert) the overall habitat is eremic: Temperatures are high, precipitation is low and very irregular, and the vegetation cover is poor. Between these two major habitats there is an intermediate habitat of steppe character. The combination of the four major north–south relief patterns with the three climatic belts imparts an unusual heterogeneity to this relatively small zoogeographic province and has resulted in a highly dynamic faunal history.
II. THE EXTENT OF BIODIVERSITY

It is difficult to estimate animal diversity throughout the whole Levant because there is better knowledge about the fauna of its southern than of its northern areas. Unfortunately, the term “southern Levant” is loosely applied by various authors to include the area of Israel, and sometimes also the Palestinian Autonomy, and/or Sinai, Jordan, and the Golan Heights. These limitations should be taken in account when considering the data (of representative groups) in Table I. They suggest that animal diversity in the Levant is high and perhaps that this area is one of the richest and most diverse natural regions among temperate regions of the world (relative to its size). For example, the area between the Mediterranean Sea and the Jordan River, together with the Golan Heights, is only approximately 28,000 km², but it contains approximately 170 bird and 100 mammal species (breeding). These numbers are not markedly fewer than those in California (210 birds and 110 mammals in an area of 404,000 km²), whose southern part lies at the same latitude as Israel.

Several factors contribute to this high diversity. One is the lack of a major disastrous disturbance; the general pattern of the landscape in the Levant has persisted since the early Pleistocene. Only minor areas of the Levant (the tops of Mt. Lebanon and Mt. Hermon) were ever glaciated to any considerable extent, and only small regions (e.g., the Golan) were subject to volcanic eruptions that covered large areas with basalt. Consequently, the Levant never suffered the catastrophic wipeouts that hit northern and central Eurasia many times throughout the Pleistocene.

Local, region-scale changes did occur, however, repeatedly separating and then uniting different patches of landscape, thus permitting possible speciation under allopatry. These local geological events probably contributed, through speciation and through invasion, to enrich rather than to impoverish the Levantine fauna. This long, generally evenless but locally eventful history of the Levant is one of the reasons for its rich fauna. A further reason is the heterogeneous climate. Mean annual rainfall may decrease from 1000 mm to less than 100 mm over a short distance of 150 km. The landscape therefore changes from Mediterranean oak maquis to batha and then to barren deserts, and the fauna gradually changes as one species replaces another. The net result is a fauna richer than would be expected were the landscape more uniform.

In contrast to factors contributing to biodiversity of the Levant, certain environmental factors limit it. The major limiting factor is that southern and eastern areas of the Levant are an extreme desert that, for many animals, is too hostile a habitat for survival.

III. HISTORIC ZOOGEOGRAPHY

Levantine biota constitute one of the most complex ecosystems in the world: Ethiopian and Oriental taxa coexist with Euro-Siberian, central Asiatic, and Mediterranean species, mixed with Saharan and Arabian desert faunal elements. This kaleidoscopic admixture of Paleartic, Palearctic, and Saharo-Arabian elements has been constantly changing during the Neogene and the Quaternary periods, disposing a new biogeographic configuration during each geological period. Consequently, the Recent fauna is a living monument to a
TABLE I
Animal Diversity in the Levant or in Regions of It

<table>
<thead>
<tr>
<th>Animal group</th>
<th>No. of species</th>
<th>Region</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Crustaceans</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ostracods (Ostracoda)</td>
<td>53</td>
<td>Israel</td>
<td>Martens and Ortal (1999)</td>
</tr>
<tr>
<td><strong>Arachnids</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scorpions (Scorpiones)</td>
<td>16</td>
<td>Southern Levant</td>
<td>Levy and Amitai (1980)</td>
</tr>
<tr>
<td>Camel spiders (Solifugae)</td>
<td>54</td>
<td>Israel</td>
<td>Levy and Shulov (1964)</td>
</tr>
<tr>
<td>Crab spiders (Thomisidae)</td>
<td>40</td>
<td>Southern Levant</td>
<td>Levy (1985)</td>
</tr>
<tr>
<td>Cobweb spiders (Theridiidae)</td>
<td>82</td>
<td>Southern Levant</td>
<td>Levy (1998)</td>
</tr>
<tr>
<td><strong>Insects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dragonflies (Odonata)</td>
<td>82</td>
<td>All Levant</td>
<td>Dumont (1991)</td>
</tr>
<tr>
<td>Termites ( Isoptera )</td>
<td>31</td>
<td>Southern Levant</td>
<td>Kugler (1988)</td>
</tr>
<tr>
<td>Bushcrickets ( Tettigoniidae )</td>
<td>42</td>
<td>Israel</td>
<td>Ayal et al. (1999)</td>
</tr>
<tr>
<td>Grasshoppers (Acridoidea)</td>
<td>110</td>
<td>Southern Levant</td>
<td>Freidberg (1988)</td>
</tr>
<tr>
<td>Flies (Diptera)</td>
<td>3500</td>
<td>Southern Levant</td>
<td>Freidberg (1988)</td>
</tr>
<tr>
<td>Asilidae</td>
<td>158</td>
<td>Southern Levant</td>
<td>Theodor (1980)</td>
</tr>
<tr>
<td>Tephritidae</td>
<td>89</td>
<td>Southern Levant</td>
<td>Freidberg and Kugler (1989)</td>
</tr>
<tr>
<td>Butterflies (Rhopalocera)</td>
<td>126</td>
<td>Israel</td>
<td>Benyamini (1988)</td>
</tr>
<tr>
<td>Caddisflies (Trichoptera)</td>
<td>73</td>
<td>All Levant</td>
<td>Reissmann (1992)</td>
</tr>
<tr>
<td>Wesps (Vespidae)</td>
<td>7</td>
<td>Southern Levant</td>
<td>Kugler (1988)</td>
</tr>
<tr>
<td>Ants (Formicidae)</td>
<td>140</td>
<td>Southern Levant</td>
<td>Kugler (1988)</td>
</tr>
<tr>
<td><strong>Mollusks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Land snails (Pulmonata)</td>
<td>100</td>
<td>Southern Levant</td>
<td>Heller (1988)</td>
</tr>
<tr>
<td>Vertebrates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fish, freshwater (Pisces)</td>
<td>69</td>
<td>All Levant</td>
<td>Krupp (1987)</td>
</tr>
<tr>
<td>Amphibians + Reptilia</td>
<td>96</td>
<td>Southern Levant</td>
<td>Werner (1988)</td>
</tr>
<tr>
<td>Birds (Aves), breeding</td>
<td>170</td>
<td>Southern Levant</td>
<td>Yom Tov (1988)</td>
</tr>
<tr>
<td>Mammals (Mammalia), breeding</td>
<td>100</td>
<td>Southern Levant</td>
<td>Yom Tov (1988)</td>
</tr>
</tbody>
</table>

fascinating animal history and ecology. Current zoogeographic patterns in the Levant are the result of four major formative events (Por, 1975; Tchernov, 1988):

1. The breakup of the Tethys Sea during the Miocene and the consequent establishment of the Eurasian–African land bridge: Approximately 18–24 million years ago (late Oligocene–early Miocene) northern regions of today's Levant were still under the large Tethys Sea, whereas southern regions were terrestrial and part of the vast Afro-Arabian continent, with its unique faunal realm (Fig. 2). Approximately 17 million years ago (Ma) the Tethys contracted, the emerging lands combined the southeastern areas of Asia Minor with the northwestern areas of Arabia, and the Levant was formed (Fig. 3).

Bridging the two continents, the newly formed Levant triggered an interchange of terrestrial biota between Africa and Eurasia. A wide spectrum of habitats were present in the Levant during much of the Miocene, including rivers, lakes, marshy habitats, dense woods, and Paleotropic rain forest, together with open country. Early Miocene deposits in the southern Levant contain a diverse community of animal taxa, most of which were of African origin (proboscoideans, crocodiles, soft-shelled turtles, and catfish) but with others representing Eurasian elements (viverrines and cricetines). It seems that during this period African forms found it relatively easy to emigrate into the Levant, whereas contemporaneous Eurasian forms found it more difficult.

2. The late Miocene and Pliocene drying out of the Old World subtropics and the formation of the vast belt of the Saharo-Arabo-Syrian deserts, which restrained terrestrial migration between Africa and the Levant: The desertification process was very gradual. At first, approximately 12–15 Ma, open-land biota of savanna and semisteppe began to expand, and in doing so they
created an extremely diversified mosaic of biotopes that ranged from the East African highlands through (still connected) Arabia to the Levant and enabled large-scale biotic interchange. Gradually, however, aridity along the Saharo-Arabian latitudes increased (associated with intensification of seasonality) so that by approximately 5.5–6 Ma (late Miocene) a vast arid belt of almost continental size formed which became a severe biogeographic barrier. Eventually, this desert barrier was settled by strongly eremic-adaptive species recruited from its immediate surroundings. The desert fauna of the Levant thus originates from both Eurasia and Africa and is established along adaptive rather than historical lines. It is more between than within faunal realms.

3. Quasi-isolation and biogeographical provincialism of the Levant during the Pliocene: Beginning approximately 5 Ma three geomorphological events had profound effects on the biota of the Levant. To the north, orogeny of the Taurus–Zagros mountain chains restrained many northern Palearctic elements from migrating southwards. To the south, the Red Sea opened up (Fig. 4) and connected with the Indian Ocean, thereby establishing, for the terrestrial fauna, a marine barrier between Africa and the Levant (in addition to the already existing desert). The third event was the brief flooding of the Mediterranean by oceanic waters through the Gibraltar straight. All along the Levant coastal plains were submerged, and deep marine embayments penetrated into such lowlands of the Levant as
(today’s) Dead Sea Rift Valley and Plain of Beer-Sheva (Fig. 5). These marine embayments of the Mediterranean (which gradually retreated later in the Pliocene) formed biogeographic barriers that enabled local speciation.

Pliocene mammals of the Levant (from a site near Bethlehem) are of open-country species, suggesting a lowland African savanna-like landscape, with some freshwater bodies nearby. The absence of deer, beavers, dormice, and bears (common in the Pliocene of Turkey) is noteworthy.

4. The successive creation of the Jordan–Orontes rift valley in the late Pliocene and early Pleistocene: During the Plio-Pleistocene transition (approximately 1.8 Ma) large-scale tectonism occurred in the southern Levant, which established a longitudinal geomorphological feature that opened initially in the south and gradually northward. The Jordan–Orontes valleys, with their chains of lakes and swamps, were then created, the cis- and trans-rift mountain ranges were then lifted, and basalts were extruded in the eastern regions of the Levant. It was during the Pliocene–Pleistocene that the contemporary relief structure of the Levant was molded (Por, 1975; Tchernov, 1988).

Levantine faunas during the lower Pleistocene were characterized by a bloom of Palearctic elements and a deterioration of tropical ones, suggesting that the Zagros Mountains filter barrier had by then been overcome and that the Saharo-Arabian barrier, between the Levant and Africa, had already become highly effective (Tchernov, 1988). By 1.4 Ma, two-thirds of the mammal genera (from a site in the Jordan Valley) were of Eurasian origin. The remaining one-third were mainly of tropical origin, and many African mammals entraped in the Levant underwent local speciation (Hippopotamus amphibius, Praomys minor, and Arvicanthis eczo). By the mid-Pleistocene (800,000–150,000 years ago) additional Palearctic immigrants had invaded the Levant (squirrel, hare, Syrian bear, and wild cat), various tropical elements had become extinct (giraffe, monkeys, the fossil artiodactyls Pelorovis and Kolpochoerus), and there were no further massive invasions of tropical elements; the fauna was becoming similar to that of today. From 125,000 years ago onwards (upper Pleistocene) there were no drastic changes in faunal turnover rates. In many species of snails, voles, mole-rats, wolves, bear, and gazelle, however, there were considerable changes in body size to adjust to environmental fluctuations. These size alterations suggest that Pleistocene climatic changes in the Levant affected population levels more than species extinctions (Heller, 1988; Tchernov, 1988).

Today, the majority of animal genera in the Levant (more than three-fourths) occur outside the Levant exclusively to the north or northeast. Faunal connections with Eurasia are thus obvious. The extent of the Eurasian element varies from group to group. Almost all species of land snails are of Palearctic origin, as are three-fourths of the fruit flies (Tephritidae), two-thirds of the butterflies, and one-fifth of all chironomids. Furthermore, the extent of the Eurasian element within each group may vary from one region of the Levant to another. From Mt. Hermon to Sinai frequencies of Palearctic butterflies decrease from 88 to 43%, within the Palearctic ant genus Formica, the number of species decreases from Turkey (12) through Lebanon (1) to...
Israel (absent; Kugler, 1988; Benyamini, 1988; Heller, 1988).

Tropical elements comprise approximately 5–13% of the fauna, but again this varies from group to group. Approximately 12% of the ant species are tropical or have strong tropical affinities; Diptera (in general) comprise approximately 5–10%, chironomids 40%, land snails 1%, and there are no aquatic oligochaetes. Within the Levant, tropical components in butterflies increase from approximately 8% on Mt. Hermon to 32% in Sinai.

Although most tropical elements are either Ethiopian or Paleotropical, a small Oriental element is also present (Table II). This is noteworthy because the Orient is distant from the Levant and separated from it by a vast desert barrier. The influence of the Oriental species is sporadic and does not reach sizable percentages in the whole fauna. Freshwater fishes and aquatic insects are exceptions, however.

In conclusion, the Levantine land bridge serves as an important crossroads of biotic exchange. Along the tropical wet and savanna-like landscapes of the rift valley, African animals entered Palearctic. Along the Mediterranean woodland and dry steppe landscapes of the bordering mountains, Eurasian animals invaded Africa. Among the dunes of the shoreland, eremic, Saharan biota advanced northward. The Levant thus functions as a complex biogeographic corridor in which representative species of the Ethiopian, Palearctic, and even
 Oriental biogeographic realms are encountered side by side. This mixture of faunas is outstanding and perhaps unique.

IV. ENDEMIC VERSUS WIDESPREAD ANIMALS

The Levant is characterized by the high frequency of marginal populations of species which range beyond the Levant. However, it is also rich in endemics (Table III). Endemism is a function of a group’s low mobility, combined with its tendency to speciate. It is therefore not surprising that within the vertebrates, endemism is higher among (the more sedentary) fish and reptiles of the Levant than among (the more mobile) mammals and birds; within arthropods it is higher among (the more sedentary) ants than among the (highly mobile) butterflies and wasps.

Comparisons within and between groups should not be carried too far, but the very high extent of endemism among land snails is noteworthy. This may be because they are, as a whole, very ineffective in crossing even minor ecological barriers of substratum or aridity. Consequently, many land snail species of the Levant have a very small range, sometimes only a few square kilometers. The endemic land snail Pene galilaea is confined to a very small area of approximately 1.5 $\times$ 2.5 km in northwestern Upper Galilee, where it is surrounded by another species, P. sidoniensis (Figs. 6 and 7). Also, the endemic hygromine Trochoidea picardi is confined to an area of 3 $\times$ 5 km in the coastal plain, where it is completely surrounded by T. davidiana. Similarly, the high level of endemism among bushcrickets is probably due to limited dispersal ability due to their loss of ability to fly. Approximately 45% of the Israeli bushcricket species are brachy- or micropterous, and 84% of these flightless species are endemic compared to 13% of the fully winged flying species.

FIGURE 6 Narrow-ranged endemism in the Levant: total global distribution of the land snail Pene galilaea (●) is only 8 km$^2$ in the western Galilee (©, sites of P. sidoniensis) (adapted with permission from Heller, 1993).
FIGURE 7 Wide-ranging distribution in the Levant: *P. sidoniensis* ranges over ca. 1200 km² in the southern Levant alone; it also occurs in the northern Levant and southern Turkey (adapted with permission from Heller, 1993).

It is tempting to seek “hot spot” regions, in which the extent of endemism is high and cuts across many animal groups. For the Levant, it is difficult to single out such regions. Endemics occur in many different regions.

In contrast to the endemics, several genera of the Levant enjoy an incredibly wide, intercontinental distribution. The land snail genera *Truncatellina*, *Lauria*, *Cecilioides*, and *Punctum*, minute snails of the litter or subterranean environment, are found outside the Levant not only throughout Eurasia but also in much of the African continent. *Oxyloma*, an amphibious genus that lives on the vegetation above water, is also included in this list. All these taxa probably reached the Levant by aerial dispersal that, whether by wind or by birds, puts a premium on small size and light weight.

Aerial dispersal separates these widespread animals from many genera of Palearctic origin, which migrated into the Levant by the diffused, gradual movement of populations across hospitable terrain over many generations. It enables constant colonization. In the Levant, for land snails, birds are probably a more important aerial transporter than the wind. They migrate very regularly over considerable distances and are large enough for the accidental transport of minute animals without inconvenience. Birds migrating from Eurasia to Africa via the Levant could thus be responsible for at least some of the minute fauna of the Levant.

V. FACTORS DETERMINING DIVERSITY

Diversity fluctuates considerably in the different geographic regions of the Levant. The following sections focus on rain and substratum, the major physical environmental forces that, by determining distribution, influence diversity, and on vegetation, an important biotic factor.

A. Rain

The Levant is situated on the margin of the extreme desert and the mean annual rainfall decreases from 1000 mm to less than 100 mm over the short distance of 150 km. This sharp gradient is reflected in animal ranges, with borderlines coinciding with isohyets so closely that there can be little doubt that rainfall is a major factor and of paramount importance in regulating animal distributions (Fig. 8).

In rain-regulated distributions the species are classified into two categories: allopatric and parapatric. Allopatric species do not share borders with any other species of their genus. The dependence of such a species’ range on rain, directly or indirectly, seems unequivocal. (For example, the Mediterranean-dwelling snails *Paramastus episomus* and *Pene sidoniensis* are the southernmost species of their genera. *Paramastus episomus* does not range below the 500-mm isohyet and *P. sidoniensis* not below the 400-mm isohyet.) Parapatric species pairs, on the other hand, share a common border of distribution. The location of the frontier within such pairs may well be influenced by interspecific competition, and rain could simply be the factor determining the point of equilibrium between the species *Sphinc-
In many groups, overall biodiversity is a function of the amount of rain. In this context, the 200-mm isohyet, representing the end of the Mediterranean and the beginning of the desert landscape, appears to represent a barrier that many animal species cannot cross. In the Negev and Judean Deserts (areas receiving less than 200 mm of rain annually) only 25 land snail species are found compared to 73 species in the Mediterranean landscape. Of these 25 desert species, 7 occur in both the desert and the Mediterranean habitats. A generally similar pattern occurs at the generic level: Only 10 genera are recorded from the deserts of the Levant compared to 40 genera in the Mediterranean landscape, of these 10 desert genera, 8 occur in both desert and Mediterranean landscape and only 2 occur exclusively in desert areas. These comparative data indicate that adaptations to desert conditions are generally restricted to the species level rather than to higher taxonomic categories.

A similar trend of decrease in diversity also occurs among dipteran families (Table IV). In four latitudinal sections of Israel that broadly represent a decreasing gradient in rainfall, Freidberg (1988) noted that of 77 families, 20 were progressively restricted to the more northern sections, with 2 families occurring only in the extreme north. Only 57 families were detected from all four sections.

Among arachnids, the 350-mm isohyet (rather than the 200-mm one) demarcates the end of the Mediterranean and beginning of the desert fauna.

Studies of species distributions along climatic gradients may aid in predicting potential faunal responses to global climatic change. However, although a “univariate” approach in which a separate model is constructed for each species is more accurate than a multivariate one, it is not very efficient for analyzing climatic responses of large species assemblages or high taxonomic groups. However, scaling up from the level of individual species to higher taxonomic levels is crucial for evaluating faunal responses to global climatic change. Early suggestions (Heller, 1988) that rainfall is important in determining patterns of land snail diversity in Israel were based on nonquantified generalizations deduced

### Table IV

<table>
<thead>
<tr>
<th>Latitudinal section</th>
<th>A-D</th>
<th>A-C</th>
<th>A-B</th>
<th>A</th>
<th>Total no. of families</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (N of 33)°</td>
<td>37</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>77</td>
</tr>
<tr>
<td>B (32–33°)</td>
<td>37</td>
<td>10</td>
<td></td>
<td></td>
<td>77</td>
</tr>
<tr>
<td>C (31–32°)</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td>67</td>
</tr>
<tr>
<td>D (S of 31°)</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td>57</td>
</tr>
</tbody>
</table>

*From Freidberg (1988).
from distribution maps of single species. Recent research (Kadmon and Heller, 1998) analyzes thegleosa of the entire land snail fauna of Israel to gradients in mean annual rainfall by integrating geographical information system (GIS) tools and standard multivariate techniques. Faunal variation of land snails was found to be significantly correlated with underlying variation in rainfall (as expected), but the effect of rainfall on the fauna was much greater in drier regions than in more rainy areas. Above 450 mm, no relationships could be detected between the patterns of faunal variation and rainfall.

1. Adaptations to Dry Conditions

Many desert species of the Levant differ from closely related Mediterranean species, or subspecies or populations, in their physiological adaptations to stress (Shkolnik, 1988). In the harsh environment of the desert, shortage of water and scarcity of adequate food are the major factors that pose a threat to many animals. I briefly discuss the various strategies that land snails and mammals employ to overcome these threats.

Concerning water economy, important adaptive features in determining a snail’s ability to inhabit dry deserts include not merely resistance to desiccation but also instant adjustment of the rate of water loss to oncoming desiccation. Within the snail genus Sphincterochila (Sphincterochilidae), desert-dwelling *S. conanu* is capable of immediate response to desiccating conditions compared to Mediterranean species that take 3 or 4 days to recruit their water-preserving mechanisms; the steppe-dwelling species *S. fimbrata* is intermediate in its speed of adjustment (Arad et al., 1989). A similar picture emerges among hygromiid and helicid land snails of the Levant. Desert-dwelling *Trochoidea simulata* and steppe-dwelling *Sxipops* *vestalis* respond rapidly to desiccation, whereas the Mediterranean *Theba pisana* and *Monacha haifensis* require up to 4 days to adjust their rate of water loss (Arad et al., 1992).

Barriers to water loss in land snails of Levantine deserts include the shell; consequently, land snails in which the shell is missing, internal, reduced to a small external remnant, or without considerable calcium enforcement do not enter the desert of the Levant (Limaux, Limacae, Milax, Deroceras, Dardelardia, Eopolita, Oxyloma, Vittrea, and Oxyloma). The epiphragm (a mucous-calcareous sheet spread over the shell aperture during seasons of inactivity) is also an important barrier to water loss, and *S. conanu* has the thickest epiphragm and also lower water loss than the Mediterranean species of *Sphincterochila*. Another adaptation is the extrabody water reservoir that snails carry inside their shell; species that lose significant amounts of water during desiccation do so almost equally from both body and extrabody compartments. However, in desert species the water content of the body is more closely controlled at the expense of extrabody water, thereby avoiding severe dehydration of soft body tissue.

Superiority in resistance to desiccation, however, is not always unique to desert animals. Mediterranean-dwelling *X. vestalis* enjoy a water loss of only 0.13% per day and cope better with desiccation than do desert-dwelling *T. simulata* (Arad, 1990). However, *X. vestalis* is a semelparous, annual species. It cannot survive even one single rainless year in the desert since all the populations would be wiped out in such an event. *Trachooda simulata*, with somewhat less efficient mechanisms, reaches a life span of at least 3 years and can survive a rainless year (Heller, 1988; Arad, 1990).

Furthermore, within the genus Sphincterochila, desert-dwelling *S. prophetarum* has water regulatory capacities similar to those of the Mediterranean species of this genus. *Sphincterochila prophetarum* dwells underneath stones, where humidity is high compared to other microhabitats of the Levantine desert.

Desert snails of the Levant, to conclude, are adapted to withstand desiccation stress in that they usually have lower and slower rates of water loss and are capable of closer regulation of stable water content of the body compared to Mediterranean snails.

Desert mammals of the Levant differ from Mediterranean mammals in that they have a lower energy metabolism. Among murid rodents of Mediterranean landscapes, *Apodemus sylvaticus* and *A. mystacinus* possess the basic metabolic rate as expected from their body mass (according to the “mouse to elephant curve”). However, in the common spiny mouse *Acomys calthi*- *nus*, a species of both Mediterranean and desert habitats, this rate is only 75% of the expected value, and in the golden spiny mouse *A. rusauus*, a species of the extreme desert and regularly active during the day, it is only 55%. Similar comparative patterns have been found among gerbils, hedgehogs, carnivores, and ruminants of the Levant. A low metabolic rate, implying a lower rate of heat generation, saves the water otherwise needed to dissipate heat in a hot environment (Shkolnik, 1988).

In addition to a shortage of water, food in the desert is also at a premium. The frugal food requirements of desert animals may be related not just to a low demand for metabolizable energy. An efficient digestion of the food consumed may also reduce the amount of food required for maintenance. Desert mammals require less food for their maintenance than do nondesert ones.
Bedouin goats herded in the southern Levant consume less food, retain it in the gut for a longer time, and digest it more efficiently than do European breeds of goat, thereby gaining more energy from a given mass of food. In addition, the Bedouin goat has a remarkably spacious rumen, which functions not only as a fermentation vat but also as a voluminous water reservoir, enabling the goat to graze in the water-depleted desert without depending on frequent drinking (Shkolnick, 1988).

Balancing their nitrogen metabolism is as much a challenge for desert mammals as the maintenance of a balanced energy metabolism. The camel, instead of wasting the nitrogenous end products of protein catabolism, first retains them in its kidney and later allows them to be recycled through the gut. Here, the urea and other nitrogenous wastes are used by the microbial symbiotic population for resynthesis of protein, from which the host animal will eventually benefit. The potential capacity for recycling urea is far greater in desert than in Mediterranean species. Recycling of urea, in addition to helping the animal survive on low-protein feed, attenuates the load on the water otherwise required for the elimination of that waste (Shkolnik, 1988).

In conclusion, rain is the major environmental factor determining the diversity of animals within the Levant. The (approximately) 200-mm isohyet marks the arid limit beyond which many species of the Levant cannot exist.

B. Substratum

Within the boundaries of the relevant isohyets, animal distribution is further limited by lithic factors. The checkered geomorphologic map of the Levant presents a wide range of substrata, including granite, basalt, limestone, chalk, marl, heavy alluvium, calcareous sandstone, and light sand dunes. Calcium-rich environments are very abundant in the Levant. Concerning one particular animal group, land snails, the abundance of limestone, chalk, or calcium-rich soils derived from them is a prerequisite for the flourishing of any testaceous fauna. Acid environments, which restrict the diversity in many regions of Europe and which in many evergreen forests throughout the world exist due to the gradual addition of acidic elements to the litter, are not known to exist in the Levant to any sizable extent. Granite and basalt, from which the extraction of calcium is difficult, are restricted to small areas. This lithic composition is an important reason for the high diversity of land snails in the Levant.

In some cases, the distribution of land snails varies in close association with geomorphology. The endemic Cristatarius genezarethana occurs almost exclusively on mid-Eocene formations and is absent from the lower Eocene formations (Fig. 9). However, among many groups, only two major categories of substratum-dependent groups can be distinguished: the fauna of the limestone mountains and that of the loose sands. Limestone mountains in the Levant are the substratum that is richest and most diverse in species because they produce a rich mosaic of habitats and provide many shelter niches. Complex systems of rock crevices, moderately dense litter, soil pockets, generous vegetation cover, and a variety of exposed versus shaded sites are
all to be found on limestone but not always on other substrata. Accordingly, a 100 × 100 m of limestone in the Judean Hills (in a Mediterranean landscape that receives approximately 570 mm of rain annually) yields 15 species of snails. A similar nearby area consisting of chalk, marl, or alluvium produces only 4 species, all of which also occur on limestone. Chalk, marl, or alluvium merely contain an impoverished, depauperated element of the rich limestone fauna but without any characteristic component of their own. Crevices provide shelter from the severe heat and drought prevailing throughout the summer. Whereas limestone has a well-developed system of crevices, chalkstone, marl, and alluvium do not and therefore, presumably, crevice dwellers cannot inhabit them. Also, the rich assemblage of litter habits, inhabited mainly by small invertebrates, is often well developed in limestone but nonexistent in nearby chalk or marl. Accordingly, there are no litter dwellers in such substrata. Extensive alluvial plains, such as Yizre’el, Zevulun, or the hinter parts of the Levant’s coastal plain, harbor little diversity.

Sands of the Levant contain a fauna that differs considerably from that of the nearby calcareous mountains and alluvial soils (Kadmon and Heller, 1998). Furthermore, sands of the Mediterranean climatic region may differ from those of the Negev (Fig. 10), which is arid.

1. Adaptations to Sand

Sand differs from other substrata in its “near-fluid” texture, which makes locomotion and burrowing more difficult than in other habitats. However, it holds more water than many other substrates so that more water is eventually translated into a more productive biotic community.

Many taxa of the animals restricted to sands converge in their morphological adaptations, thereby signifying the importance of sand as an evolutionary agent. Among reptiles of the Levant, a suite of morphological, physiological, and behavioral adaptations enables certain taxa to live in the sands or endows them with a competitive advantage (Werner, 1985). Locomotion over sand is facilitated in lizards by expansion of the feet. In *Acanthodactylus*, the extent of fringing along the toes varies intraspecifically and interspecifically (Fig. 11A). Also, the toes of the gecko *Stenodactylus* spp. are fringed, and those of the skink *Scincus scincus* are flatly expanded (Fig. 11B). Sand-dwelling viviparous locomote by sidewinding, as do *Cerastes* spp. and *Pseudocerastes fieldi*. *Lytorhynchus diadema* employs a unique variety of serpentine locomotion mechanisms wherein the loops push the sand down rather than sideways. Since burrows tend to collapse, on the one hand, and loose sand enables instantaneous submergence and sand swimming, on the other hand, problems arise from submergence and locomotion within the sand medium. Wedge-shaped snouts, frequent in burrowing reptiles, are prominent in the sand-swimming skinks *S. scincus* (Fig. 11B) and *Sphenops* seposides. In these and in the snake *Lytorhynchus* the mouth opening is protected from sand during submerge progression by its ventral position. *Lytorhynchus* has an expanded rostral, common but not unique...
Some adaptations of Levantine psammophilous reptiles to sand. (A) Foot of *Acanthodactylus schreiberi syriacus* showing fringe of spiny scales along third and fourth toes. (B) *Scincus scincus* showing shovel-shaped snout, countersunk jaw, angularly set-off venter, and expanded toes (photographs by Y. L. Werner).

Sand-dwelling snakes. *Cerastes* spp. can submerge stationally by shuffling the flattened body sideways. The vertical flanks of *Scincus* (and *Sphenops*) may improve the lateral thrust during sand swimming. The relatively high species density of *Gekkonidae* on the sand may reflect the preadaptation of spectacled forms to sand. Breathing below sand is problematical not only at the nasal passage level but also at the thoracic: Upon exhalation the sand caves in and stifles inhalation. *Scincus* and *Sphenops* are structured for breathing with the protected ventral surface angularly set off from the lateral ones (Fig. 11B; Werner, 1985).

Sand-dwelling scorpions of the Levant can also be distinguished by their morphological adaptations. The most well-known are bristles to aid traction, locomotion, and burrowing in sand. These bristles occur on the tarsomeres and tibiae of all sand-living scorpions. Another leg modification in sand scorpions that acts to increase the effective surface area of the tarsi is long claws (Fet *et al.*, 1998). The bristles and claws create an expanded surface area on the tip of the legs and prevent sinking into the sand. The genus *Buthacus* (represented by *B. arenicola*, *B. yotvatensis*, *B. l. leptochelys*, and *B. l. nitzani*) is a typical sand dweller in the southern Levant, with up to 25 bristles on its walking legs (Levy, 1980).

Sand-dwelling grasshoppers can be distinguished from those of other biotopes by their behavioral adaptations. Sand-dwelling *Hyalorrhipis calcarata*, *Leptopternis* spp., and *Eremogryllus hammadari* dig into the sand with their hindlegs, throwing the sand behind them. They are thus able to cover themselves, leaving only the upper part of the head and the antennae exposed above the surface (species of other substrata usually settle ‘head to sun,’ thus casting the smallest shadow; Fishelson, 1985).

In conclusion, substratum is the second major environmental factor determining animal diversity in the Levant. Limestone rocks offer calcium, shelter, and an additional microhabitat—that of litter. Sands form a separate substratum category that usually has a unique fauna.

### C. Vegetation

Both animal and plant distribution are influenced by rainfall and substratum, and consequently their distribution patterns may overlap. Whether vegetation influences animal distribution directly, in that different animal species have food preferences for different higher plants, varies among groups. Among fruit flies (Tephritidae), the larvae of practically all species in the southern Levant are phytophagous. All representatives of the Myopitinae, Oedaspidinae, Tephritinae, and Schistopterinae (65–70 species, 25 genera) develop only in plants of the composites (Compositae), whereas none of the Aciurinae and Dacinae do so. The daciine *Dacus oleae* attacks olives (Oleaceae), whereas other daciines are associated with fruits of milkweed (Asclepiadaceae). Many tephritid species are oligophagous; a few are polyphagous, using plant hosts of different families. The parts of plants attacked by tephritid larvae also vary considerably among groups. Larvae of Dacinae and Trypetinae develop in fleshy or juicy fruits. However, *Euleia heracleii* is a leaf miner, and *Capparimyia savastani* (Fig. 12) produces larvae that develop in the flower buds of the caper *Capparis*. Most species of the five subfamilies that attack Asteraceae in Israel develop in fruit heads. The larvae of different species are usually restricted to feeding on certain parts of the flower head, such as flowers, achenes, or the receptaculum. Some species induce the formation of galls (e.g., *Myopites*...
FIGURE 12. Distribution patterns that correlate with vegetation: Cap- parimyia savastani (Tephritidae), the larvae of which develop in flower buds of Capparis (reproduced from Freidberg and Kugler (1989) with permission from the Israel Academy of Sciences and Humanities, frontispiece).

spp. and Urophora spp.). Several species bore into the stems of their hosts, often forming galls (e.g., Spathulina spp. and Oedaspis spp.). Orellia falcata, unlike most members of the Terelliinae, mines along the stem of the goat’s beard Tragopogon and pupates in the upper part of the root (Freidberg and Kugler, 1989).

Some grasshopper species use plants as a food source, resting source, and hideout. Morphological adaptations to this habitat include large arolia to serve for attachment, streamlined and smooth body structures, heads with an acute or subacute angle, long and delicate colorless wings, and usually green or greenish-gray or brown coloration. Behavioral adaptations include resting along branches or stems and, if flying up, moving from plant to plant. When disturbed, many of these species crawl around the stem or branch, remaining on the opposite side to the intruder; others drop from the plant and dig among the dense branches close to the ground. There is a strong correlation between special types of vegetation and their fauna of plant-dwelling acridoids within the Levantine fauna: The Tropidopola–Ochrilidia association of grasshopper species is found on reed Phragmites, reed mace Typha, and rush Juncus; the Heteracris–Eyprepocnemis association shows a great affinity for knotweed Polygonum sp. and wormwood Artemisia monosperma, and the Dociostaurus–Morphacris association occurs mainly in areas with low, usually dry, ephemeral grass (Fishelson, 1985). It is not known if these associations are obligatory.

In contrast to fruit flies, among land snails of the Levant a direct influence of vegetation on diversity is unlikely. Snails feed on a wide spectrum of decaying vegetable matter and saprophytic fungi, occasionally on green tissues, and there is no evidence that food requirements for specific plants determine their wide-scale distribution patterns. Vegetation, however, may play an important role in microgeographic distribution, especially in preventing some snails from occupying certain habitats. Some species of the Mediterranean region (Bulimus labrosus, Sphincterochila cariosa, Euchondrus septendentatus, E. saulcyi, Paramastus eiptus, and Xeropicta vestabita) are found on the southern slopes of hills, where annual vegetation prevails, and are very uncommon on the northern slopes that are covered with dense oak maquis. In regions in which the vegetational distinction between northern and southern slopes is not obvious, the distinction of the snails between the slopes also becomes obscure, and they also occupy the northern slopes. A preference for the south slopes could perhaps be due to the annual vegetation prevailing on it. Annual plants do not develop xeromorphic characters and are therefore easily eaten and digested. The oak maquis of the north-facing slope may be cool and damp and offer more shelter, but it consists mainly of perennial vegetation. The development of xeromorphic characters (such as the thick cuticle of the oak leaves) necessary for plant survival in the dry season could perhaps be an obstacle for snail feeding and digestion. Too little information is available on the precise diet of many animal groups. Once this information is gained, I believe that a major breakthrough will occur in our understanding of biodiversity in the Levant.

VI. FRESHWATER DIVERSITY: HISTORIC FACTORS

The inland water system of the Levant is dominated by the north–south-oriented topography of the Rift Valley. The three major rivers of the Rift (the Jordan, Litani, and Orontes) run along successive segments of the valley and create a “steeple chase” waterway. Between the Rift basin and the Mesopotamian basin, watershed shifting and headwater capture may well have enabled a faunal transition of transcontinental dimensions.

During the Pliocene the Mediterranean flooded the southern Rift Valley and upon its retreat it left, in the freshwater of the Jordan Valley, several marine relicts. Crustacean species of the Jordan Valley suggested by Per (1975) as marine relicts of the Pliocene invasion include Loxocyma gallaecia, Pseudobradya barroisi, Nitocra balnearia, Monodella relicta, Typhloleptolana reichii,
Typhlocaris galilea, and Bogidiella hebraea. The blind prawn Typhlocaris galilea (Fig. 13) is endemic to a subterranean warm, sulfide-rich spring near the shores of Lake Kinneret, where it feeds on oligochaetes (Isochaeta israelis) and snails (Theodoxus jordani).

From the Pleistocene onwards there was a continuous existence of freshwater or saline lakes in the Jordan Rift Valley. In this large endorheic system, the smallest climatic change and the changing balance of evaporation and precipitation (in addition to any slight tectonic movement) resulted in considerable fluctuations of water levels, with the creation, fusion, separation, and disappearance of lakes and rivers. This resulted in richly diverse faunas that partly replaced one another throughout the Pleistocene. Some old lacustrine freshwater endemics have survived in the Jordan system since the early Pleistocene (Table V).

The fish fauna of the Levant is a mixture of Palearctic, Oriental, and Ethiopian species. Thirty-four species of freshwater fishes occur in the Orontes, the largest river of the Levant. Of these, approximately two-thirds are Palearctic, one-third are Oriental, and only 3% are African elements. The Orontes shares 18 of its species with the Tigris–Euphrates but only 7 with the Jordan River. The number of species occurring in the Jordan, the second largest river, is only slightly lower, but the species composition is totally different: Of 28 species, the Palearctic, Oriental, and African elements each comprise about one-third. The Jordan River shares only 3 species with the Tigris–Euphrates, but 6 additional species have their closest relatives in Mesopotamia. Apparently, the Jordan and various courses of the Orontes were separately colonized via branches of the Mesopotamian river system, and the Jordan drainage basin became an important center of speciation for several fish lineages with Palearctic affinities. Later, faunal elements of the Jordan reached the Orontes (through the Litani). Some freshwater fish of African origin reached the Jordan Valley by way of freshwater connections to the south. They could have done so up to Pliocene times, and in the Jordan Valley they differentiated specifically or even generically (Astatotilapia flavitopisophi and Tristramella spp.). During later periods, some African fish may have reached the southern Levant from the Nile via the Mediterranean Sea (the euryhaline cichlid Tilapia zilli survives well in brackish and even marine waters) (Fig. 14).

The fish fauna of the Rift’s lakes and rivers was thus recruited from the north (Palearctic), east (Euphrates), and south (the Nile, Krupp, 1987).

The headwaters of Jordan constitute a typical temperate Palearctic freshwater fauna, with a ‘normal’ level of diversity that includes pulmonate snails, cold-water copepods, and many stream-living insects such as stone flies (Plecoptera), Elmidae among the Coleoptera, and typically cold-water-like Rhyacophila (Trichoptera). There are no less than 12 species of caddis flies (Trichoptera) in the springs of the River Dan; there are 13 species of Ephemeroptera in the River Hatzbani.

**TABLE V**

Endemic Species of the Jordan Rift Valley

<table>
<thead>
<tr>
<th>Group</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porifera</td>
<td>Cortispongilla barroisi</td>
</tr>
<tr>
<td>Tricladida</td>
<td>Dugesia salina</td>
</tr>
<tr>
<td></td>
<td>D. biblica</td>
</tr>
<tr>
<td>Oligochaeta</td>
<td>Isochaeta israelis</td>
</tr>
<tr>
<td>Halacarida</td>
<td>Limnohalacarus capernaumi</td>
</tr>
<tr>
<td>Lohmanella heptapegoni</td>
<td></td>
</tr>
<tr>
<td>Crustacea</td>
<td>Limnocaenis galilae</td>
</tr>
<tr>
<td></td>
<td>Hylodiplocaenis barroisi</td>
</tr>
<tr>
<td>L. nitida</td>
<td></td>
</tr>
<tr>
<td>Protothoe barroisi</td>
<td></td>
</tr>
<tr>
<td>Nitocra incerta</td>
<td></td>
</tr>
<tr>
<td>Stenocheta galilea, Barlowa rubrae</td>
<td></td>
</tr>
<tr>
<td>Sclerothoe turriculata</td>
<td></td>
</tr>
<tr>
<td>Parakathylella caudata</td>
<td></td>
</tr>
<tr>
<td>Monachella rubrae</td>
<td></td>
</tr>
<tr>
<td>Typhlocaris galilae</td>
<td></td>
</tr>
<tr>
<td>T. steinitzi</td>
<td></td>
</tr>
<tr>
<td>Monodella relicta</td>
<td></td>
</tr>
<tr>
<td>Bogidiella hebraea</td>
<td></td>
</tr>
</tbody>
</table>

*From Por (1975), select groups.*
Elsewhere in the southern Levant, there is little of this rich freshwater fauna. In all the springs and brooks of Israel, usually only two species of Ephemeroptera can be found, *Cloeon dipterum* and *Calenis macrura*, which are well-known for their broad ecological valency. Only a small number of freshwater invertebrates are found in every spring and brook. This is a monotonous fauna characterized by *Melanopsis lucinoides* (Mollusca), *Prouellia caenalis* (Echinogammarus syriacus and *Eucyclops serrulatus* (Crustacea), and *Dina lineata* (Hirudinea). However, even this impoverished fauna does not advance much further than the oasis springs of En Avdat and of Qadesh Barnea, on the brink of the extreme desert. In most springs of Sinai, usually only insects or passively transported animals occur (Por, 1973).
VII. BIODIVERSITY: THE HUMAN IMPACT

Since the dawn of the Pleistocene, the Levant has been inhabited by the evolving hominids, starting with the early Homo erectus of the Jordan Valley. It was in the Levant that the 10,000-year-old Neolithic revolution of the Old World, with its agriculture and animal domestication, occurred.

The history of the Levant is characterized by tides of highly sophisticated agricultural civilization and ebbs of nomadic destruction. Agricultural apogees saw the development of soil protection, terracing, aqueduct and irrigation works, and drainage of swamps, whereas nomadism brought erosion, overgrazing, destruction of irrigation works, and swamping. Nomads semiconductorly created the deserts, of which they are often called the sons. Throughout history there was a progressive destruction of the woodland habitats by ever-increasing deforestation.

However, the most profound effects of man on biodiversity of the Levant started at the beginning of the twentieth century with the tremendous increase in human populations, use of firearms, extent of cultivated areas, and use of pesticides. These changes have had a pronounced, and in many cases fatal, effect on wildlife in the Levant. The effects of various human activities on animal life in the Levant (mainly in Israel and mainly concerning vertebrates) have been studied in depth by Yom-Tov and Mendelssohn (1988).

A. Hunting

The proliferation of firearms into the Levant by the end of the nineteenth century was followed by the overhunting and extermination of many game animals, mainly at the beginning of the twentieth century. Roe deer Capreolus capreolus and fallow deer Dama dama mesopotamica used to occur in the woods of Galilee and of Mt. Carmel; the last roe deer was shot in approximately 1912 and the last fallow deer probably at the beginning of the twentieth century; the last onagers in Jordan was shot in 1932; now this species is extinct in the Levant. The Nile crocodile Crocodylus niloticus was found in swamps near Mt. Carmel as late as 1912 (Yom-Tov and Mendelssohn, 1988).

Hunting also caused the almost total elimination of the green and loggerhead turtles (Chelonia mydas and Caretta caretta) which used to nest along the sandy Mediterranean beaches. During 1920–1930 approximately 30,000 turtles were killed along Israel's seashores; in 1985 only 14 nests were found along its whole Mediterranean coast.

Today, effects of hunting on wildlife in the Levant vary in broad correlation to the enforcement of hunting laws (and hence to political borders). In Israel, hunting today has only minor effects. There are only approximately 5000 licensed hunters and the main game species are chuckar partridges, rock pigeons, doves, various wintering waterfowl, hare, porcupine, wild boar, and mountain gazelle where they damage agricultural crops. Many of these may be hunted only during well-defined seasons, and none are seriously affected by hunting. Indeed, increasingly efficient enforcement of hunting laws has enabled several species to recover: the leopard Panthera pardus nimr, of which approximately 20 individuals now dwell in the Judean Desert and Negev; the wolf Canis lupus pallipes; and the hyena Hyaena hyaena. All wild mammals, birds, and mollusks in Israel are completely protected by law, with the exception of designated pest species (rats, house mice, voles, fruit bats, house sparrows, etc).

B. Habitat Destruction

Most habitats of the Levant have been severely or completely destroyed by urban and agricultural development. Two examples are presented in the following sections (Yom-Tov and Mendelssohn, 1988).

1. Coastal Sand Dunes

Coastal sand dunes form a narrow strip along the Mediterranean coast of the Levant, and until the beginning of the twentieth century were almost uninhabited by man. Today, however, the majority of the population of Israel lives on the coastal plain. The many urban settlements built on the coastal sands have caused loss of sand dune areas to such considerable extent that of

used to be found in Turkey, Lebanon, and Syria and was not rare in the Galilee at the beginning of the twentieth century; the last individual was killed in 1965, and today this subspecies is extinct. The other subspecies, P. p. nimr, survives in the Judean desert and Negev. An ostrich Struthio camelus syriacus was seen in Jordan in 1932, now this species is extinct in the Levant. The Nile crocodile Crocodylus niloticus was found in swamps near Mt. Carmel as late as 1912 (Yom-Tov and Mendelssohn, 1988).
sand dunes are the most eastern extension of Saharan sand-dwelling animals into the Mediterranean, temperate zone of the Levant: the Egyptian tortoise *Testudo kleinmanni*, snakes (*Cerastes vipera*, *Macroprotodon curtus*, and *Lycorhinus diadema*), skinks (*S. scincus* and *Sphenostrictus sepioides*), geckos (*Stenodactylus silenoides* and *S. petrii*), lizards (*Acanthodactylus stellatus* and *Agama savignii*), a monitor (*Varanus griseus*), a chameleon (*Chamaeleo chamaeleon musae*), rodents (*Gerbillus gerbillus*, *G. pyramidum* and *G. andersoni allenbyi*, the endemic jerboa *Jaculus jaculus schuleri*, and the endemic jird *Meriones sacramenti*), and a hedgehog, *Hemiechinus auritus*. The destruction of the coastal sand dunes (accompanied by predation by feral domestic cats and dogs) has reduced the distribution and abundance of these species.

2. Wetland Habitats

Wetland habitats that existed in the Levant at the beginning of the twentieth century included swamps, lakes, rivers, springs, and temporary (mainly winter rain) pools; man-made fish ponds were then barely existent. Today, however, water in the Levant is a scarce and much needed resource. All these habitats have been affected.

a. Swamps

At the beginning of the twentieth century, the Hula swamp ranged more than 40 km² north of the (13 km²) Hula Lake (Fig. 16) and contained a rich diversity of (mainly widespread) animals. The swamp and lake were drained in the 1950s, but a small nature reserve (3 km²) was created that represents an impoverished version of the original swamp. Even so, 22 species endemic to the Levant have not been recorded from the Hula since its drainage (Table VI). Although the Hula nature reserve is today rich in bird life, eight bird species that once bred in the Hula Valley do not breed there anymore.

b. Riverine Habitats

The rivers, streams, and wadis in the Levant drain, in general, either to the Mediterranean Sea or to the Rift Valley. Today, most Mediterranean-flowing rivers of the southern Levant are polluted, and consequently most fish have disappeared from them. Gray mullet (*Mugil cephalus* and *M. ramada*) and eel (*Anguilla anguilla*) fingerlings once used to spend their first years in these rivers (they reproduce in the sea) but now do not occur in most of them. Eight other, entirely freshwater, species are today very rare and the endemic *Acanthobrama telavivensis* now faces extinction. The soft-shelled turtle *Trionyx trianguliss*, which once lived in many of the
Mediterranean-flowing rivers and nested on their shores, has disappeared from most of them. Only in Alexander River does a breeding population persist, but many of the eggs fail to hatch and the population consists only of very large specimens. The river otter (*Lutra lutra*), a former inhabitant of all these rivers, survives only in the Bezet, Keziv, and Na’aman, in which it is only occasionally seen.

In the Rift Valley-flowing streams the major threat to biodiversity is the extreme reduction in the amount of water that flows in them since most springs are piped by man and used for irrigation. This has reduced and even eliminated the populations of several animal species. The river otter, once common in each major tributary in the Jordan River system, is now rare. The blue-cheeked bee-eater (*Merops superciliosus*), which bred along the Jordan River in the early 1930s, does not breed there any more (Yom-Tov and Mendelssohn, 1988).

### c. Temporary Winter Rain Pools

A multitude of temporary winter rain pools once existed in temperate regions of the Levant. Many existed near Arab villages, where their waters once served for watering livestock and for irrigation. Many of these village pools were man-made, built as reservoirs by erecting a dam across a wadi or hewn into rock at the foot of a water-collecting slope, and were probably ancient. These pools supported a rich and varied invertebrate fauna and several species of amphibians.

The number of winter rain pools decreased sharply during the 1960s and 1970s. Many of the formerly undrained valleys were turned into cultivated fields. Other pools were eliminated due to road construction and building. Continuous spraying of herbicides on roadsides and of insecticides on pools near human settlements destroyed this ecosystem or turned these pools into breeding sites for pesticide-resistant mosquito strains. One result of these operations is that the once abundant amphibians have become a threatened group. The spadefoot toad (*Pelobates syriacus*) (Fig. 17) breeds exclusively in rainwater pools and its tadpoles require a long (3 months) development period. Only large seasonal ponds of sufficient duration can support a viable population of spadefoot toads, and only 27 such ponds remain in Israel (including the Golan). Lack of suitable breeding localities has also caused the newt (*Triturus vittatus*) to become a threatened species, and even the formerly very common tree frog (*Hyla arborea*) and green toad (*Bufo viridis*) are becoming rare (Yom-Tov and Mendelssohn, 1988).

### C. Poisoning

More than 700 compounds are registered in Israel’s Ministry of Agriculture for agricultural use, including insecticides, acaricides, nematicides, fungicides, herbicides, bactericides, molluskicides, rodenticides, insect attractants, bird and mammal repellents, fumigants,
<table>
<thead>
<tr>
<th>Animal Group</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sponges (Porifera)</td>
<td>Ephydatia syriaca</td>
</tr>
<tr>
<td>Turbellarians (Turbellaria)</td>
<td>Dendrocoelum sp., Dugesia biblica, Phagocata? armeniaca</td>
</tr>
<tr>
<td>Crabs (Crustacea)</td>
<td>Anhyzella respiensi affinis</td>
</tr>
<tr>
<td>Turbellarians (Turbellaria)</td>
<td>Dendrocoelum sp., Dugesia biblica, Phagocata? armeniaca</td>
</tr>
<tr>
<td>Copepods (Copepoda)</td>
<td>Attheyella trispinosa affinis</td>
</tr>
<tr>
<td>Insects (Insecta)</td>
<td>Rhyothermis semihyalina syriaca, Urothermis edwardsi hulae, Calopteryx hyalina</td>
</tr>
<tr>
<td>Caddisflies (Trichoptera)</td>
<td>Orthotrichia? melitta</td>
</tr>
<tr>
<td>Beetles (Coleoptera)</td>
<td>Herophydrus cleopatre, Herophydrus galileae, Hydrometra n.sp.</td>
</tr>
<tr>
<td>Mosquitoes (Diptera)</td>
<td>Anopheles pharoensis</td>
</tr>
<tr>
<td>Bivalves (Bivalvia)</td>
<td>Potomida littoralis semirugata, Unio terminalis terminalis</td>
</tr>
<tr>
<td>Fish (Pisces)</td>
<td>Tristramella simonis intermedia, Nun qalilaeus, Mirogrex terraesanctae hulensis</td>
</tr>
<tr>
<td>Amphibians (Amphibia)</td>
<td>Discoglossus nigriventer</td>
</tr>
<tr>
<td>Mammals (Mammalia)</td>
<td>Arvicola terrestris hintoni, Amniotes phario</td>
</tr>
<tr>
<td>Birds (Aves)</td>
<td>Aquila clanga, Aquila pomarina, Falco tinnunculus, Falco naumanni</td>
</tr>
</tbody>
</table>

*Adapted from Dimentman et al. (1992).*

Plant growth regulators, and defoliants. Effects of pesticide residues on wildlife started shortly after DDT began to be used at the end of World War II. The most dramatic effect, however, was during the early 1950s, when thallium sulfate-coated grain was widely used to control rodents. These poison campaigns were not necessary. The main damage of these rodents was caused before 1950, when fields were plowed with the traditional shallow-plowing plows that did not disturb rodent burrows. When, after 1950, deep plowing was introduced, rodent burrows were destroyed, the rodents were exposed to predation, and the damage they caused decreased (pest control officers, however, attributed the decrease in rodent damage to their intense poison campaigns). In unplowed areas, as in alfalfa fields, damage caused by voles is still considerable and no rodenticides can prevent it. Since the 1960s, fluoroacetamide has replaced thallium sulfate. The residues of both substances accumulate in bodies of secondary consumers such as raptors, bats, and carnivores.

Of 39 species of birds of prey that occurred in Israel before the use of pesticides, all but 2 were seriously affected. Species that once were common breeders became very rare breeders (the black kite Milvus migrans, griffon vulture Gyps fulvus, long-legged buzzard Buto ferox, Bonelli’s eagle Hieraaetus fasciatus, Egyptian vulture Neophron percnopterus, kestrel Falco tinnunculus, lesser kestrel Falco naumanni, and lanner falcon Falco biarmicus). Rare breeders went extinct or their populations decreased drastically (the raptor-faced vulture Torgos tracheliotus, spotted eagle Aquila clanga, peregrine Falco peregrinus brookei, marsh harrier Circus aeruginosus, black eagle Aquila verreauxi, white-tailed eagle Haliaeetus albicilla, bearded vulture Lammergeier, and Gypaetus barbarus). Thallium sultate also affected the wintering raptors, whose populations decreased considerably; some species disappeared completely for many years (the sparrowhawk Accipiter nisus and merlin Falco columbarius). Insect-eating birds, such as lesser kestrel Falco naumanni, and lanner falcon Falco biarmicus, and probably due to its specialized reptilian diet and to its absence in winter when most of the poison grain was used. The latter factor also saved the hobby Falco subbuteo, the breeding population of which returns to the Levant only in May.

Today, after prohibition of the use of DDT, other chlorinated hydrocarbons, and thallium sultate, some species have made a comeback but the existing populations of breeding raptors are only a fraction of their former populations. Hundreds of pairs of griffin vulture...
once bred in Galilee and on Mt. Carmel, but only 20 breed today; only one pair of lapper-faced vultures bred in 1986 in the Negev, where approximately 25 pairs had bred before the widespread use of agricultural poisons began. Continuous breeding attempts by long-legged buzzards, Bonelli’s eagle, and golden eagle Aquila chrysaetus often fail; only the kestrel seems to have almost fully recovered.

The removal of many raptors had secondary effects on the fauna. The population increase of the blackbird Turdus merula, the hubbul Pycnonotus barbarus, the palm dove Streptopelia (Stigmatopelia) senegalensis, the Syrian woodpecker Dryobates syriacus, and the jay Garrulus glandarius may be partly attributed to the decrease in the numbers of their predators (mainly the sparrowhawk A. nisus).

Secondary poisoning by insecticides also affected some insectivorous birds, particularly species that lived near fields and human settlements. Populations of the swallow Hirundo rustica, red-rumped swallow Hirundo daurica, white-throat Sylvia communis, rufous bushchat Cercotrichas galactotes, nubian shrike Lanius nubicus, spotted flycatcher Muscicapa striata, roller Coracias garrulus, bee-eater Merops apiaster, and Egyptian nightjar Caprimulgus aegyptius all decreased considerably.

Also, insectivorous bats (Microchiroptera) have been affected by chemicals. Twenty-eight species are known in the southern Levant, and many caves used to be inhabited by thousands of roosting bats. The numbers of both bat-inhabited caves and individuals roosting in them have decreased drastically since the end of the 1950s and most species have become rare. This decline is attributed to two main factors. The first is fumigation of caves with ethylene dibromid and later with lindane (Gammexan). Fumigations were intended to control the fruit-eating bats Rousettus aegyptiacus, which were considered fruit pests and which often roost together with insectivorous bats. It was later found that the damage allegedly caused by fruits bats was marginal at best and had been grossly exaggerated by pest control officers. The population of one insectivorous bat, Pipistrellus kuhli, remained less affected than the others, probably because it roosts not in caves but in hollow trees and in wooden houses. The second factor affecting insectivorous bats is secondary poisoning. Most bat species occur in the cultivated regions and are highly vulnerable to secondary poisoning because noctuid moths are an important part of their diet. Larvae of noctuids such as Spodoptera litura are serious agricultural pests, and fields are regularly sprayed with insecticides to eliminate them. Hence, secondary poisoning may be another factor contributing to the decrease of bat populations in the Levant.

The remaining bat populations suffer now from an-
other threat. Increasing numbers of hikers visit various caves and disturb roosting and hibernating bats, thus causing wastage of fat reserves and desiccation of the hibernators.

Carnivores are another group which suffered from poisoning. In 1964, the Plant Protection Department of Israel decided that jackals (Canis aureus), one of the few mammals not protected at that time by the “Wild Animals Protection Law,” were a nuisance. They were being blamed for damaging plastic sheets used to cover certain crops. A large antijackal campaign was started and tens of thousands of chicks injected with 1081 (fluoracetamid) were spread over the Mediterranean area of Israel in an effort to eradicate the jackals. Several mammal predators were affected, including the jackal, the wolf Canis lupus, the red fox Vulpes vulpes, the Egyptian mongoose Herpestes ichneumon, the jungle cat Felis chaus, and the African wild cat F. sylverstris. Most species recovered within a few years and their numbers today are similar to those before the poisoning campaign. The jackal, however, is much slower to recover and its numbers in Israel are still low. The wolf, whose populations in the Negev have increased in recent years due to greater food availability at garbage dumps, is endangered in central and northern Israel, where individuals are larger than those in the Negev and occasionally prey on livestock. Although the wolf is legally protected, stock owners often retaliate by poisoning.

The caracal Caracal caracal, which was formerly known only in the Negev, has increased its distribution area northwards into the Galilee. This range extension coincided with the population decrease in many Mediterranean carnivores following the jackal poisoning operation, suggesting that range extension of the caracal was possible due to the absence of competitors. Apparently, the main competitor of the caracal is the jackal, which preys on hares, the main food item of the caracal. After the jackal poisoning campaign, hares increased considerably. Today, after most predator species have recovered from the effects of poisoning, caracals still occur in the Mediterranean region of Israel but seem to be rare.

D. Changes in Agricultural Practices

Current agricultural practices in the southern Levant are different from those of the past mainly in that today approximately half of the cultivated area is irrigated. As a result, areas which formerly were left fallow in summer are now cultivated in the dry season. Some species have responded to the change in agricultural practices with an increase in both abundance and distribution. Population size of the mountain gazelle Gazella g. gazella in central and northern Israel increased between 1948 and 1985 from less than 300 to 10,000. Availability of succulent, nutritious food and of water throughout the year in irrigated agriculture enables female gazelles to deliver their first fawns at the age of 1 year (2 years in nonagricultural areas) and average 1.8 young annually (only 1 in nonagricultural areas). Gazelles have become a pest in several areas because of their high numbers. The males damage fruit trees by rubbing their horns against the bark, and both sexes eat cotton, wheat, corn, and other crops. Increasing food and water availability in agriculture has also affected the dorcas gazelle G. dorcas in the southern Negev. Whereas populations near agricultural settlements increased between 1964 and 1985 by 9% annually, the rate of increase in other populations was only approximately 3%. Normally, dorcas gazelles in the southern Levant have only one fawn per year, in spring. Even if this fawn is lost, the dam does not become estrous until the normal breeding season in autumn. In recent years, increasingly more fawns are born in autumn so that either some females breed twice per year, as do mountain gazelles, or females become estrous again after losing their fawn. Autumn fawns are seen mainly near agricultural areas.

Availability of green food in summer and leftover grain in wheat and barley fields has enabled several seed-eating birds to increase their populations to such an extent that they have become agricultural pests: the collared dove Streptopelia decaocto, palm dove Stigmotis ennealepis, house sparrow Passer domesticus, and feral domestic pigeons Columba livia domestica. The planting of exotic trees and ornamental plants in human settlements has made new areas available to several bird species. Typical woodland species, such as the syrian woodpecker D. sylriacus, blackbird Turdus merula, great tit Parus major, and jay G. glandarius, once restricted mainly to woodlands of the Galilee, Mt. Carmel, and Judean hills, are now widespread and common in the gardens and parks that accompany many human settlements, sometimes even in the Negev and Jordan Valley. The Syrian woodpecker has become a pest in certain areas because it drills holes in irrigation pipes and in telephone cables.

Approximately 1500 plant species, mostly ornamental plants, were imported to Israel during the twentieth century. The orange-tufted sunbird Nectarinia osea is favorably affected by this enriched flowering vegetation of ornamental exotic plants in human settlements. When first discovered (in the nineteenth century) it occurred only in the lower Jordan Valley and near the
Dead Sea, its distribution coinciding with that of the mistletoe Loranthus acaciae. Today, however, owing to the presence of ornamental plants with nectar-bearing flowers, the sunbird is widespread and common in settlements throughout much of the southern Levant. Another species benefiting from the spread of plantations and gardens is the Arabian bulbul Pycnonotus barbatus capensis. It feeds mainly on fruits, flowers, and leaves of introduced plants such as Melia azedarach, Lantana sp., Erythrina sp., and other fruit trees and flowers. Due to the ample available food it has increased and become a pest in vineyards and orchards.

Man-constructed fish ponds and water reservoirs present an alternative habitat for some wetland species. The night heron Nycticorax nycticorax, little egret Egretta garzetta, and to a smaller extent squacco heron Ardeola ralloides are the main beneficiaries, but glossy ibis Plegadis falcinellus also breed near some reservoirs. All four species were rare breeders or nonbreeders in the Levant until the late 1950s, when they started breeding near fish ponds, which provide food and nesting sites. Fish ponds also enabled the pied kingfisher Ceryle rudis to extend its breeding distribution from the Jordan Valley to the coastal plain. Dirt roads surrounding fish ponds and reservoirs are a favorite nesting site of the spur-winged plover Recurvirostra avocetta, which once bred only in the Hula swamp area, are now common breeders along the coastal plain on banks of fish ponds and reservoirs. Also, the little-ringed plover Charadrius dubius and the Kentish plover C. alexandrinus, which formerly bred mainly on the Mediterranean seashore, now breed on reservoir banks. Another wader that breeds on banks, especially of salt ponds in the coastal plain, and become a pest in vineyards and orchards.

All four species were rare breeders or nonbreeders in the Levant until the late 1950s, when they started breeding near fish ponds, which provide food and nesting sites. Fish ponds also enabled the pied kingfisher Ceryle rudis to extend its breeding distribution from the Jordan Valley to the coastal plain. Dirt roads surrounding fish ponds and reservoirs are a favorite nesting site of the spur-winged plover Recurvirostra avocetta, which once bred only in the Hula swamp area, are now common breeders along the coastal plain on banks of fish ponds and reservoirs. Also, the little-ringed plover Charadrius dubius and the Kentish plover C. alexandrinus, which formerly bred mainly on the Mediterranean seashore, now breed on reservoir banks. Another wader that breeds on banks, especially of salt ponds in the coastal plain, and become a pest in vineyards and orchards.

Among mammals, the coypu Myocastor coypus was introduced into the southern Levant for the fur industry during the 1950s. Some individuals escaped from breeding farms, and others were released when their farming was discontinued. Currently, coypu are common near water bodies, from northern Israel to the Gaza Strip. During the 1960s, it has become rather rare. However, the poisoning of jackals also reduced the number of mongooses. This was followed by an increase in the number of snake bites in Israel by the Palestine viper Vipera palmeata (Fig. 18) from approximately 200 per year during 1960–1964 to 430 in 1967. Recovery of the mongoose populations was followed by a decrease of snake bite cases to the former level.

In the future, animal diversity in the Levant will be far from the great ecological and zoogeographical variety which makes it so unique.

See Also the Following Articles

DESERT ECOSYSTEMS • MEDITERRANEAN-CLIMATE ECOSYSTEMS • NEAR EAST ECOSYSTEMS, PLANT DIVERSITY

Bibliography


![Viperidae](image_url)
Diptera: Tephritidae. Israel Academy of Sciences and Humanities, Jerusalem.


NEAR EAST ECOSYSTEMS, PLANT DIVERSITY
Avinoam Danin
The Hebrew University of Jerusalem

I. Environmental Conditions
II. Flora
III. Vegetation

GLOSSARY
batha A biblical term for semishrub communities, mainly of seral vegetation of old field succession in the Mediterranean part of the Near East.
contracted vegetation Vegetation restricted to wadis that receive additional water supply.
diffused vegetation Vegetation occupying all slopes and most habitats.
semisteppe batha Semishrub communities developing on most soil types at the boundary of the Mediterranean zone of the Near East.
wadi Dry water course.

THE HIGH RICHNESS OF PLANT SPECIES OF ISRAEL, expressed as species to area and being 9.06 species/100 km² (with a similar number for Jordan), is related mainly to its position in a meeting zone of plant geographical regions, each with its own typical flora. This wealth is supported by the existence of many habitats needed to support these species. The wealth of habitats derives from the climatic transition between the relatively moist area in the northern part of Israel and Jordan and the extreme desert areas in their southern part. A long history of human pressure of cultivation and grazing by domestic animals led to strong stress on the existing flora and enabled the introduction of many alien species, many of which occupy habitats created by human activity. In this article, these factors are discussed and examples are provided.

I. ENVIRONMENTAL CONDITIONS
Currently, the most important factors affecting the distribution of species are environmental conditions. The evolutionary history of the area reviewed influenced the composition of the flora, but the environment influences the composition and their ability to coexist and prosper. Before discussing the vegetation and its gradients of species diversity, the physical setup of the study area will be discussed.

A. Topography
The topography of Israel and Jordan can best be described as north–south topographic belts, which are influenced by the geomorphological features of the area. The Mediterranean coastal plain is narrow in the north of Israel (Fig. 1, 1) and becomes wide in the south (Fig. 1, 5). Low hills with gentle topography and wide valleys with deep soil constitute the foothills. The mountainous area reaches elevations of approximately 1000 m at the Judean Mountains (Fig. 1, 12) and the Negev Highlands.
FIGURE 1 Geographical subdivision of the study area (based on Danin, 1999). Israel: 1, coastal Galilee; 2, Acco Plain; 3, coast of Carmel; 4, Sharon Plain; 5, Phthierson Plain; 6, Upper Galilee; 7, Lower Galilee; 8, Mt. Carmel; 9, Esdraelon Plain; 10, Samaria; 11, Shefela; 12, Judean Mountains; 13, northern Negev; 14, western Negev; 15, Negev Highlands; 16, southern Negev; 17, Hula Plain; 18, Kinneret Valley; 19, Beit Shean Valley; 20, Mt. Gilboa; 21, Samaritan Desert; 22, Judean Desert; 23, Lower Jordan Valley; 24, Dead Sea Valley; 25, Arava Valley; 26, Mt. Hermon; 27, Golan Jordan; 28, Gezeid; 29, Ammon; 30, Moab; 31, Edom; 32, Mediterranean sands and salt marshes; 33, anticlines of northern Sinai; 34, gravely plains covered with shifting sands; 35, gravely plains of central Sinai; 36, Table mountains of central and western Sinai; 37, Gebel el Sgna; 38, coastal plain of the Gulf of Suez; 39, coastal plain and foothills of the Gulf of Elat; 40, Tiran and Sinaifi Islands; 41, sandstone belt; 42, Lower Sinai Massif; 43, Upper Sinai Massif.

(Fig. 1, 15) and 1200 m at the Galilee (Fig. 1, 7). The Jordan–Dead Sea–Arava rift valley is part of the Afro-Syrian Rift Valley and constitutes the lowest topographic element in the area. This valley is 400 m below sea level in the deepest part of the Dead Sea (Fig. 1, 24). The terrain ascends to 200 m above sea level at the foot of Mount Hermon (Fig. 1, 26) in the north and at the water divide of the Arava Valley (Fig. 1, 25) to the south. A few transversal valleys, such as the Esdraelon Plains (Fig. 1, 9) running from east to west, dissect the mountainous area into large blocks. Steep escarpments, dissected by canyons, typify the areas east and west of the rift valley. There are a few internal lakes in this valley. Lake Hula between the Golan and Galilee was drained in the 1950s. The Kinneret, also known as the Sea of Galilee, is the largest freshwater natural reservoir of Israel. The Dead Sea is a highly salty water body. The large salt marshes at the Arava Valley are in fact a large underground lake. At the vicinity of these water bodies there are small or large springs. Mount Hermon, at the common border of Israel, Lebanon, and Syria, reaches an elevation of 2800 m and terminates sharply at the basalt-covered plateau of the Golan to the south. This plateau descends gently from an elevation of 1200 m toward the Yarmouch River and toward the Kinneret at 200 m below sea level. The Jordanian plateau starts south of the Yarmouch River (Fig. 1, 28). Its escarpments delimiting the rift valley are as steep as, and in many places steeper than, the escarpments in Israel. The Jordanian plateau has several peaks 1200–1600 m high. A few deep rivers dissect the plateau from east to west toward the rift valley. East of the water divide, which is near the western edge of the plateau, the landscape becomes flat and gradually descends toward Mesopotamia.

Sinai has a wide coastal area along the Mediterranean Sea (Fig. 1, 32), ascending gradually southwards. The northern Sinai sand belt reaches the anticlinal ridges of Gebel Maghara, Gebel Halal, and Gebel Yiallaq (Fig. 1, 33). The central Sinai gravel plains (Fig. 1, 35) are surrounded by crescent-like ranges of mountains or plateaus (Fig. 1, 36), dissected by a few valley passages such as those of Wadi Goldh, Wadi Mitlah, and Wadi Sdr. South of the gravelly plains there are two large plateaus ascending gradually in a north–south direction. The Gebel el Igha plateau (Fig. 1, 37) peaks at 1600 m and the Gebel et Tih plateau peaks at 1450 m. The erosion escarpments of the two plateaus are steep and that of the latter descends to the sandstone belt. The morphology of the sandstone belt (Fig. 1, 41) is highly diverse in different parts of Sinai. The Southern Sinai Massif (Fig. 1, 42 and 43), built of magmatic and metamorphic rocks, is highly dissected topographically. The mountain peaks reach 2300–2600 m. The escarpments toward the Gulf of Elat–Aqaba are not as steep as those of the rift valley in Israel and Jordan. The
coastal plain along this gulf is wide in the southern part of Sinai, whereas in many places the mountain slopes descend directly into the sea without any coastal plain. There is a wide coastal plain along the Gulf of Suez.

B. Rock Types, Geomorphology, and Edaphic Conditions

The most common rock types of northern Israel and Jordan are sedimentary limestones, dolomites, chalks, and marls of the Cretaceous and the Tertiary. Terra Rossa soils occur in the mountainous areas on hard rocks, and Rendzinas occur on the soft ones. Basalt rocks typify much of the Golan plateau and northeast Galilee. Both are covered with basaltic brown or red Mediterranean soils. Tertiary and Pleistocene rocks and derived soils fill up the small and large valleys. Calcareous sandstone, locally known as kurkar, and sandy-loamy soils, known as hamra, typify the coastal plain near the coast and deep clay soils, known as grumusols or vertisols, far from the Mediterranean coast. Cretaceous and older sandstone typify the vicinity of the rift valley of the Jordan River, the Dead Sea, and Arava in Jordan. The edaphic conditions in northern Jordan and south to Amman are similar to those of the area west of the Jordan River.

The steppe and desert areas of Israel, Jordan, and Sinai develop on a much more diverse assemblage of rocks. The main contributors are limestone, chalk, marl, chert, sandstone, magmatic, metamorphic rocks, and gravel of alluvial origin or rocks weathered in situ. Loess is important eolian sediment, composed mainly of silt and clay particles, that influences much the soils of the transition zone of the Mediterranean territory and the steppelands. Large flatlands south of this transition zone occur in the northern Negev of Israel and east of Irbid, Jarash, and Amman in Jordan. The poor moisture regime of loessial soils with low quantities of rainfall makes the desert boundary prominent.

Dan et al. (1975) provide soil maps of Israel, and Al-Eisawi (1996) provides a map of Jordan; a detailed soil map of Sinai is not available. The specific influence of each substratum type on the vegetation is discussed in detail elsewhere (Damin, 1983). The relationships between rock type, soil, and vegetation in the desert areas of Israel and Sinai are similar to those in many areas of Jordan. Geomorphological structures in Israel are small, and they are larger in Sinai and Jordan; however, these are smaller than those in countries such as Saudi Arabia in which huge areas of similar structures occur. This is the expression of relatively high geomorphological diversity. One may expect to find a high diversity of rock and soil types under a certain climatic regime. Hence, habitat diversity in such areas is expected to be high.

C. Climate

1. Rainfall

The climate of Israel, Jordan, and Sinai is Mediterranean, characterized by a cold and rainy winter and a rainless and warm summer. Rainfall quantities vary in two directions: Rainfall decreases gradually from north to south—sharply from the water divide eastwards in Israel and gradually in Jordan from the water divide eastwards (Fig. 2). The north–south gradient is influenced mainly by the intensity and frequency of rain.
contributing systems, whereas the west–east gradients are influenced by the topography. Thus, the ascent of air bodies from the Mediterranean Sea toward the mountains of Judea, Samaria, and Galilee results in cooling of the air and an increase in rainfall with increasing elevation. Air bodies descending toward the rift valley become warmer and drier; hence, the Samarian and Judean Deserts and the eastern Galilee occur at the rain shadow of the Judean, Samarian, and Galilee mountains, respectively. The ascent to the Jordanian plateau mirrors the rainfall map with maximum at the mountaintops. There is a gradual decrease in mean annual rainfall eastwards with increasing continentality far from the Mediterranean toward the desert. The increase in mean annual rainfall in Moav and Edom, located south of the west–east section of the 300-mm isohyet in Israel, is related to the orographic influence of the high elevation of this area. Snow may fall in winters of cold and wet years on the highest summits of mountains in Israel, Jordan, and Sinai and remain for a few days.

2. Dew and Fog
Dew and fog are important sources of humidity for the poikilohydric organisms growing on rock outcrops. The measurement of dew by Duvdevani's dew gage indicated a high similarity to events of efficient dew when lichens imbibed on stones and rocks at the Negev Highlands. The mean annual number of nights with more than 0.02 mm per night of dew has been 191 ± 22 for the past 15 years at Avdat; of these, there are only 124 ± 28 nights annually with dew amounts of 0.11–0.5 mm. Dew measurements in other parts of Israel, Jordan, and Sinai are not available. However, lithobiont communities near Avdat were used to extrapolate the regional distribution of dew. There are many similarities in weathering features found at the top of the northern Sinai anticlines (Fig. 1, 33) and on limestone outcrops at the area marked 11d in Jordan (Fig. 3). I assume that the dew and fog regime is similar in the latter two areas to that of the Avdat area where real measurements of dew and rainfall were carried out.

3. Temperature
Temperature regime in the study area is influenced by the altitudinal and latitudinal position of the site under review. Mean annual temperature in the desert areas varies from 9° to 23°C. The lowest temperatures prevail in the peaks of southern Sinai and Jordan near Shoubak, whereas the highest are those of the Dead Sea and Arava Valleys, where elevation reaches 400 m below sea level. Mean annual temperature maps of Israel are presented in the atlas of Israel and of Jordan in Al-Eisawi (1996).

II. FLORA
There are 2882 plant species in Israel, including approximately 200 species which occur in the Mt. Hermon area and are absent from the other districts of Israel. The number of species in Jordan is 2078 according to Al-Eisawi (1996). There are 2885 species in the combined lists of plants of Israel and Jordan (A. Danin,
TABLE I

<table>
<thead>
<tr>
<th>Country/State</th>
<th>Species</th>
<th>Area (km²)</th>
<th>Species/100 km²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Israel</td>
<td>2682</td>
<td>29,600</td>
<td>9.06</td>
</tr>
<tr>
<td>Sinai</td>
<td>889</td>
<td>61,100</td>
<td>1.43</td>
</tr>
<tr>
<td>Egypt (excluding Sinai)</td>
<td>1120</td>
<td>1,000,000</td>
<td>0.11</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>2200</td>
<td>2,200,000</td>
<td>0.1</td>
</tr>
<tr>
<td>Syria</td>
<td>3040</td>
<td>189,000</td>
<td>1.69</td>
</tr>
<tr>
<td>Spain</td>
<td>5950</td>
<td>304,000</td>
<td>1.90</td>
</tr>
<tr>
<td>Tunisia</td>
<td>2250</td>
<td>164,000</td>
<td>1.37</td>
</tr>
<tr>
<td>Greece</td>
<td>4200</td>
<td>132,362</td>
<td>3.17</td>
</tr>
<tr>
<td>Italy</td>
<td>5600</td>
<td>301,100</td>
<td>1.86</td>
</tr>
<tr>
<td>Britain</td>
<td>1666</td>
<td>229,850</td>
<td>0.72</td>
</tr>
<tr>
<td>California</td>
<td>3037</td>
<td>411,000</td>
<td>1.23</td>
</tr>
</tbody>
</table>


The high species richness of Israel, expressed as the parameter of species to area (Table I), is related mainly to the following factors:

1. Its position in a meeting zone between plant geographical regions, each with its own typical flora.
2. The existence of many habitats needed to support these species. The wealth of habitats derives from the climatic transition between the relatively moist area in the northern part of the two countries and the extreme desert areas in their southern part and also in Jordan’s eastern part. Topography is also a factor in creating the warm climates of the rift valleys and the relatively cold climate of the mountainous areas. Similarly, other highlands and lowlands have local climatic influence which increases the habitat diversity of Israel and Jordan and increases the number of habitats which support high plant species diversity. The high number of rock types influences the development of many soil types in a small area, increasing the diversity of habitats available for plants.
3. A long history of human pressure of cultivation and grazing by domestic animals led to strong stress on the existing flora and enabled the introduction of many alien species, many of which occupy habitats created by human activity.

According to various studies of the flora of Israel, it may be divided into the following groups:

1. Mediterranean species, which are distributed around the Mediterranean Sea.
2. Irano-Turanian species, which also inhabit Asian steppes of the Syrian Desert, Iran, Anatolia in Turkey, and the Gobi desert.
3. Saharo-Arabian species, which also grow in the Sahara, Sinai, and Arabian deserts.
4. Sudano-Zambesian species, typical of the sub-tropical savannas of Africa.
5. Euro-Siberian species, also known in countries with a wetter and cooler climate than that of Israel. These species grow mainly in wet habitats and along the Mediterranean coast.
6. Biregional, triregional, and multiregional species that grow in more than one of the regions mentioned previously.
7. Alien species derived from remote countries; these plants propagate without human assistance.

III. VEGETATION

Danin (1999) discusses the vegetation of the entire area in detail. The most prominent factors influencing the high species diversity of the study area are discussed in the following sections. The reader may combine the general information on the vegetation of Israel (Danin, 1999) with that of Sinai (Danin, 1983) and Jordan (Al-Eisawi, 1996).

A. Maquis and Forests

The principal spontaneous woodland areas of Israel are found in the mountains of Judea, Carmel, and Galilee and at the foot of Mt. Hermon; those of Jordan occur north of Amman (Fig. 3, 1). These forests or maquis are dominated by the sclerophyllous evergreen Quercus calliprinos and the deciduous Pistacia palaestina and grow on hard limestone with terra rossa soil. The companions of Q. calliprinos vary according to edaphic and climatic conditions. In the mesophytic aspect of the oak woodlands, found mainly in Upper Galilee, there are several trees, shrubs, vines, and geophytes that are not found in the xerophytic aspect of the southern Judean Mountains. In the driest maquis stands, Rhamnus lycioides subsp. graecus is the only arboreal companion of Q. calliprinos. Typical vines in these maquis are Rubia tenuifolia, Lonicera etrusca, Asparagus aphyllus, and Ephedra foeminea.

Marly chalk is a common rock type that has high moisture-holding capacity and is covered with a shallow light rendzina soil. The aeration of the rhizosphere, of trees and shrubs that penetrate into the soft rock, is poor and leads to nutritional stresses, which only specially adapted plants can withstand. Much of the nitrogen in these soils is in the form of ammonium ions, whereas...
in the terra rossa it is in the nitrate ion form. The vegetation cover of the light rendzinas on marly chalk is poor when compared with that on terra rossa and includes very few annual species. The principal trees of these soils are *Arbutus andrachne* and *Pinus halepensis*. In most of the area, cultivated plants have replaced the spontaneous trees. A few thousand years ago, people in the eastern Mediterranean countries started to clear the natural vegetation to create agricultural land. Trees such as olives (*Olea europaea*) and almonds (*Amygdalus communis*), which grow spontaneously in the area, have been domesticated. Timber derived from the forests and maquis was used for the construction of houses, for agricultural tools, and for fire fuel. For the past few millennia shepherds have burned large woodland areas to open paths for domestic animals, and the pasture quality has been improved through the replacement of trees and shrubs with palatable herbaceous plants.

As a result of thousands of years of deforestation and agricultural and urban development, large areas of Israel and Jordan look like mosaics of serial communities. Abandoned cultivated ground becomes populated for dozens of years by herbaceous and low lignified plants. Habitat diversity, and thus species diversity, is extremely high due to the removal of the shading trees and the opening of poor and temporary fertile soils to annual plants which could not grow in the closed woodlands. Frequent human-induced fires reopen areas, which become populated with shading trees and shrubs and ephemeral plants that increase species diversity. Some terra rossa soils that are leached and poor become fertilized during the years of crop cultivation. Once abandoned, a series of replacement annual plant communities occurs, from those using fertile ground, including tall plants, to those using poor soil and small in size. By sampling approximately five such stages, each a square of 1 m × 1 m², in the vicinity of Jerusalem, I found approximately 100 different species in 8 m². By sampling approximately five such stages, each a square of 1 m × 1 m², in the vicinity of Jerusalem, I found approximately 100 different species in 8 m². Local enrichment of the poor soils by gazelle droppings and at the vicinity of nests of harvesting ants creates an ecotone of soil fertility which enables the coexistence of many species.

Light rendzinas were often cultivated due to the high water-holding capacity of the substratum of soil plus rock. When abandoned, this soil may support the growth of a few annual species. After time passes and the aerated plowed soil becomes flattened, microbiont crust, including cyanobacteria, lichens, and mosses, develops on the surface, thus influencing soil aeration and changing edaphic conditions. Such minute edaphic changes along the precipitation gradients of the mountainous areas (Fig. 2) contribute to the high species diversity of the country.

B. *Quercus calliprinos* Woodlands on Basalt

The basalt flows of the northern Golan are younger and differ in rock types from those of the rest of the Golan. The woodlands constituting these woodlands (Fig. 3, 2) are rich in their herbaceous companions. Contrary to the vegetation of the calcareous rocks, they are almost devoid of semishrub communities at the early succession stages because of destruction and abandonment. The main environmental factor influencing this is the high phosphorous content of the rock and soil which favors the development of ephemeral plants. Thus, a dense maquis of *Q. calliprinos* covers the gentle north-facing slope of the ancient volcanic cone of Har Odem, the Golan, at an elevation of 900–1000 m. Regarding the trees, the rich ephemeral vegetation includes approximately 20 species of *Trifolium*, many other Papilionaceae, and rich annual flora of other families. The phytomass is low, thus indicating some nutrient deficiency when compared to that of the vegetation of other basalt flows. Under such conditions, without one dominant species there is high species diversity.

Other basalt rocks derived from older basalt flows at lower elevation, down to 500 meters above sea level (m a.s.l.), support denser and taller ephemeral vegetation with different composition due to differences in soil fertility and climatic conditions, which follow elevation. The oak trees and their arborescent companions are found only occasionally and they have no impact as shadow creators in most of the Golan and eastern Galilee. Ancient volcanoes provided much volcanic ash to large areas surrounding them. Because of a different pedogenic nature, this substratum supports a rich flora, many components of which are not common in other kinds of soil in this area. Many of the volcanic ash layers have a porous and well-aerated texture. In contrast to the grumusols and protohumusols of much of the area surrounding it, this substratum supports many narrow-range species.

C. Montane Forest of Mt. Hermon

The montane forest (Fig. 3, 3) stretches at Mt. Hermon from 1300 to 1700 m a.s.l., where there are only scattered remnants of the woodlands which are believed to have been there before their destruction by human activity. The dominants are deciduous trees such as...
rocks in southern Sinai and in southwestern Jordan. These occur at high elevations in crevices of smooth-faced Astragalus. Nonetheless, a few representative shrubs of the genera Crambe and Cotoneaster, common at Mt. Hermon, occur at high elevations in crevices of smooth-faced rocks in southern Sinai and in southwestern Jordan.

D. Open Forests of Quercus ithaburensis

The tabor oak is the most frequent tree in the open forests prevailing in large areas of Israel and Jordan (Fig. 3, 4). Syringa officinalis also appears with the tabor oak when developed on chalky ground of the Lower Galilee, Israel, and the Gilead in Jordan. Pistacia atlantica is the companion on basalt rocks of the Golan. Plenty of herbaceous plants cover the open space among the trees in all areas of this category. The few semishrubs found in this community when developing on chalky ground are mainly of Majorana syriaca. This is contrary to the dominance of Sarcopterium spinosum, the typical component of open communities on terra rossa. This type of forest, which once dominated the Sharon Plain, is restricted in the Sharon (Fig. 3, 4) to solitary, sporadic tabor oak trees which survived urbanization and agricultural development. There are large woodlands of the tabor oak in the Lower Galilee and in the Golan, lower than 500 m. Some of the largest indigenous trees in Israel are those of tabor oak at the Hula and Dan Valleys in the northern section of the rift valley in Israel. The tabor oak woodlands of the northwest of Jordan are well preserved. They cover large areas of rocky terrain west of Irbid, at the western escarpments of the Jordanian plateau, from sea level to 300 m a.s.l. The drier and more continental climate of the Jordanian woodlands allows for a different assemblage of companions to the oak.

E. Open Forests of Ceratonia siliqua and Pistacia lentiscus

Open woodlands dominated by carob trees and shrubs of P. lentiscus cover the lower altitudinal belt of the main mountain ranges, 0–300 m on both sides of the central mountain range on terra rossa soils. This community inhabits rendzina soils at the foothills of the Judean mountains and sandy soils of the Sharon Plain in the littoral aspect of the community (Fig. 3, 5). Generally, the community is more drought and heat resistant than the communities dominated by Q. calliprinos and has a similar position in the aridity sequence of communities as those dominated by the tabor oak. Wild olive (Olea europea var. sylvestris), which resembles the cultivated olive but has much smaller fruits, is an important companion of the carob in rocky sites at Mt. Carmel and Galilee. In Jordan, this category is almost missing, however, scattered carob trees occur in the open woodlands of Quercus ithaburensis of the Gilead, mainly in the transition zone from the belt dominated by the tabor oak to that of Q. calliprinos at elevations of 300–600 m. The main companion of the carob at the southern Judean foothills is Rhamnus lycioides subsp. graecus. The latter totally replaces all the other arboreal components in dry habitats.

The flora of this community is rich due to high habitat and microhabitat diversity. The hard limestone rocks have many crevices and soil pockets, which enable diverse microhabitats supporting different species to coexist that are isolated from each other by the rock body. Soft chalk rocks, covered by hard nari crust, also known as caliche, constitute additional diverse habitats. The nari layers have an upper hard crust layer of 5–10 mm which covers softer rock but is still consolidated chalk. Trees and shrubs, the roots of which penetrate into the rock, may establish themselves, thus avoiding the competition of the rich herbaceous flora which accompanies them in the deep soil pockets. Thus, in many places the local boundary of the land dominated by shrubs and trees coincides with that of the nari rocks. Various kinds of shade are provided by evergreen trees and have a similar position in the aridity sequence of communities as those dominated by the tabor oak. Wild olive (Olea europea var. sylvestris), which resembles the cultivated olive but has much smaller fruits, is an important companion of the carob in rocky sites at Mt. Carmel and Galilee. In Jordan, this category is almost missing, however, scattered carob trees occur in the open woodlands of Quercus ithaburensis of the Gilead, mainly in the transition zone from the belt dominated by the tabor oak to that of Q. calliprinos at elevations of 300–600 m. The main companion of the carob at the southern Judean foothills is Rhamnus lycioides subsp. graecus. The latter totally replaces all the other arboreal components in dry habitats.

The flora of this community is rich due to high habitat and microhabitat diversity. The hard limestone rocks have many crevices and soil pockets, which enable diverse microhabitats supporting different species to coexist that are isolated from each other by the rock body. Soft chalk rocks, covered by hard nari crust, also known as caliche, constitute additional diverse habitats. The nari layers have an upper hard crust layer of 5–10 mm which covers softer rock but is still consolidated chalk. Trees and shrubs, the roots of which penetrate into the rock, may establish themselves, thus avoiding the competition of the rich herbaceous flora which accompanies them in the deep soil pockets. Thus, in many places the local boundary of the land dominated by shrubs and trees coincides with that of the nari rocks. Various kinds of shade are provided by evergreen trees and have a similar position in the aridity sequence of communities as those dominated by the tabor oak. Wild olive (Olea europea var. sylvestris), which resembles the cultivated olive but has much smaller fruits, is an important companion of the carob in rocky sites at Mt. Carmel and Galilee. In Jordan, this category is almost missing, however, scattered carob trees occur in the open woodlands of Quercus ithaburensis of the Gilead, mainly in the transition zone from the belt dominated by the tabor oak to that of Q. calliprinos at elevations of 300–600 m. The main companion of the carob at the southern Judean foothills is Rhamnus lycioides subsp. graecus. The latter totally replaces all the other arboreal components in dry habitats.

The flora of this community is rich due to high habitat and microhabitat diversity. The hard limestone rocks have many crevices and soil pockets, which enable diverse microhabitats supporting different species to coexist that are isolated from each other by the rock body. Soft chalk rocks, covered by hard nari crust, also known as caliche, constitute additional diverse habitats. The nari layers have an upper hard crust layer of 5–10 mm which covers softer rock but is still consolidated chalk. Trees and shrubs, the roots of which penetrate into the rock, may establish themselves, thus avoiding the competition of the rich herbaceous flora which accompanies them in the deep soil pockets. Thus, in many places the local boundary of the land dominated by shrubs and trees coincides with that of the nari rocks. Various kinds of shade are provided by evergreen trees and have a similar position in the aridity sequence of communities as those dominated by the tabor oak. Wild olive (Olea europea var. sylvestris), which resembles the cultivated olive but has much smaller fruits, is an important companion of the carob in rocky sites at Mt. Carmel and Galilee. In Jordan, this category is almost missing, however, scattered carob trees occur in the open woodlands of Quercus ithaburensis of the Gilead, mainly in the transition zone from the belt dominated by the tabor oak to that of Q. calliprinos at elevations of 300–600 m. The main companion of the carob at the southern Judean foothills is Rhamnus lycioides subsp. graecus. The latter totally replaces all the other arboreal components in dry habitats.

The flora of this community is rich due to high habitat and microhabitat diversity. The hard limestone rocks have many crevices and soil pockets, which enable diverse microhabitats supporting different species to coexist that are isolated from each other by the rock body. Soft chalk rocks, covered by hard nari crust, also known as caliche, constitute additional diverse habitats. The nari layers have an upper hard crust layer of 5–10 mm which covers softer rock but is still consolidated chalk. Trees and shrubs, the roots of which penetrate into the rock, may establish themselves, thus avoiding the competition of the rich herbaceous flora which accompanies them in the deep soil pockets. Thus, in many places the local boundary of the land dominated by shrubs and trees coincides with that of the nari rocks. Various kinds of shade are provided by evergreen trees and have a similar position in the aridity sequence of communities as those dominated by the tabor oak. Wild olive (Olea europea var. sylvestris), which resembles the cultivated olive but has much smaller fruits, is an important companion of the carob in rocky sites at Mt. Carmel and Galilee. In Jordan, this category is almost missing, however, scattered carob trees occur in the open woodlands of Quercus ithaburensis of the Gilead, mainly in the transition zone from the belt dominated by the tabor oak to that of Q. calliprinos at elevations of 300–600 m. The main companion of the carob at the southern Judean foothills is Rhamnus lycioides subsp. graecus. The latter totally replaces all the other arboreal components in dry habitats.

The flora of this community is rich due to high habitat and microhabitat diversity. The hard limestone rocks have many crevices and soil pockets, which enable diverse microhabitats supporting different species to coexist that are isolated from each other by the rock body. Soft chalk rocks, covered by hard nari crust, also known as caliche, constitute additional diverse habitats. The nari layers have an upper hard crust layer of 5–10 mm which covers softer rock but is still consolidated chalk. Trees and shrubs, the roots of which penetrate into the rock, may establish themselves, thus avoiding the competition of the rich herbaceous flora which accompanies them in the deep soil pockets. Thus, in many places the local boundary of the land dominated by shrubs and trees coincides with that of the nari rocks. Various kinds of shade are provided by evergreen trees and have a similar position in the aridity sequence of communities as those dominated by the tabor oak. Wild olive (Olea europea var. sylvestris), which resembles the cultivated olive but has much smaller fruits, is an important companion of the carob in rocky sites at Mt. Carmel and Galilee. In Jordan, this category is almost missing, however, scattered carob trees occur in the open woodlands of Quercus ithaburensis of the Gilead, mainly in the transition zone from the belt dominated by the tabor oak to that of Q. calliprinos at elevations of 300–600 m. The main companion of the carob at the southern Judean foothills is Rhamnus lycioides subsp. graecus. The latter totally replaces all the other arboreal components in dry habitats.

The flora of this community is rich due to high habitat and microhabitat diversity. The hard limestone rocks have many crevices and soil pockets, which enable diverse microhabitats supporting different species to coexist that are isolated from each other by the rock body. Soft chalk rocks, covered by hard nari crust, also known as caliche, constitute additional diverse habitats. The nari layers have an upper hard crust layer of 5–10 mm which covers softer rock but is still consolidated chalk. Trees and shrubs, the roots of which penetrate into the rock, may establish themselves, thus avoiding the competition of the rich herbaceous flora which accompanies them in the deep soil pockets. Thus, in many places the local boundary of the land dominated by shrubs and trees coincides with that of the nari rocks. Various kinds of shade are provided by evergreen trees and have a similar position in the aridity sequence of communities as those dominated by the tabor oak. Wild olive (Olea europea var. sylvestris)
barley (Hordeum spontaneum), and wild oats (Avena sterilis). This area (Fig. 3, 6) contains the highest gene-
netic diversity of these species, which are believed to have been domesticated in the Near East thousands of years ago (Zohary and Hopf, 1993). A similar formation covers the west-facing slopes of the Gilead, Jordan, below the tabor oak belt. In drier and warmer sites or sites at which the soil is shallow, Stipa capensis is the dominant plant. The lignified and spiny shrub Ziziphus lotus, spread as green patches over the area, typifies most areas of these grasslands. Sites of relatively high soil fertility, common in this plant community, are the nests of the harvesting ant Messor messor and piles of dung of the male gazelles (Gazella gazella). This categ-
ory is presented as a mosaic with that of the 'savannoid Mediterranean vegetation' (Fig. 3, 7) in the steep topog-
raphy at the vicinity of the rift valley north of Jericho to north of the Sea of Galilee.

G. Mediterranean Savannoid Vegetation
Warm, stony/rocky slopes of the Galilee, Golan, Gilead, and Samaria descending to the rift valley (Fig. 3, 7) at sea level and below are covered with grasslands of Mediterranean annuals with large seeds, such as wheat, barley, and oats, and scattered Ziziphus spinos-
christi trees. The latter is a low, spiny tree with edible fruits that also grows in true savannas in Africa, where its companions are Sudanian perennial grasses. There-
fore, the vegetation here is named “savannoid.” Stands of this tree are also established along the transversal valleys on deep clay soils and along the Sharon Plain, where it grows on sandy–loamy soils together with the grass Desmostachya bipinnata. Local diversity of nutri-
ents and derivative high species diversity are caused by harvesting ants and gazelles as discussed previously.

At the Bet-Shean Valley south of the Sea of Galilee there are considerable areas of slightly saline springs, the salts of which remain in the soil while water evapo-
ration occurs. This results in a wide range of microhab-
itals which vary in soil salinity and the availability of spring water. The northernmost site of Balanites aegyp-
tina, an African savanna tree, is in the vicinity of one of these springs. The Sudanian mistletoe Phceoesphalos acutae (= Loranthus acutae) has its northernmost site on these Balanites trees.

H. Semisteppe Batha
Semishrub communities of the Balanites undulatae at the boundary of the Mediterranean zone (Fig. 3, 8), where mean annual rainfall is 320–400 mm, are re-
garded as semisteppe bathas. There are several commu-
nities dominated by Mediterranean plants such as Sarco-
poterium spinosum, which dominates bathas in more mesic parts of the country. Others, such as Artemisia sieberi and Noaea mucronata, dominate steppe areas in the Negev, Sinai, Jordan, and eastwards to Afghanistan. Many plants which play an important role in the xer-
al communities in fallow fields at the center of the Medi-
terranean region grow here in what is regarded as their primary habitats. No anthropogenic disturbance assists or enables their growth here, and they occupy their typical habitats which are related to certain edaphic and climatic conditions. Sarcopterium spinosum, which becomes etholcont and dies under the shade of trees and shrubs in the course of plant succession in maquis and forests, has no such competitors in the semisteppe bathas. The semisteppe bathas extend further south and east of the regional boundary of the maquis in Jordan and in Israel. Because semishrubs are the dominant tallest growth form, the semisteppe bathas are rich in their flora. Many plants of mesic and xeric origin coexist here, contributing to the high species diversity. Phyto-
geographical analysis of each association in this vegeta-
tion indicates a higher diversity in origin when com-
pared to associations of drier and moister conditions.

I. Tragacanth Vegetation of Mt. Hermon
Semishrubs which look like spiny cushions constitute the most prominent formation of vegetation developed on the windward slopes of the peaks of Mt. Hermon above 1900 m (Fig. 3, 9). Many species of section Traga-
cantha in the genus Astragalus and of the genus Acant-
tholimon are components of the cushion plants forma-
tion throughout the Middle East. The spiny cushion seems to have some biological advantages which make it adaptable to some of the components of the harsh conditions of this habitat: cold winter with high-veloc-
ity winds, precipitation mainly as snow, and dry sum-
mer. The short growth season and the harsh environ-
ment are the main factors enabling the success of very few annual species as companions of the cushion plants. Snow accumulates on the slopes in the wind shadow of small local ridges or crests because of the low wind velocity. Consequently, snow may reach a depth of 10 m for a few months. Water drainage from the slopes of this area is through karst systems and not through wadis as in many areas of the country. A common feature of the karst topography is the occurrence of small valleys which are filled with fine-grained soil known as dolinas. The soil in these valleys is waterlogged for a long time,
thus increasing habitat diversity and the growth of many additional species compared to those of the rest of Israel. Several species of tragacanth *Astragalus* occur in areas which are far from Mt. Hermon and the Anti-Lebanon mountains. *Astragalus bethlemiticus* is a typical companion of steppes and rock vegetation in the shrub–steppes and semisteppe baths of the Negev Highlands and of southwestern Jordan; *A. echinus* is a representative of this group at the high elevations of southern Sinai, where it is confined to crevices of smooth-faced granite.

**J. Vegetation Patterns in the Dry Areas of Israel, Jordan, and Sinai**

When dealing with the vegetation of drylands, which receive less than 200 mm mean annual rainfall, one has to consider general patterns of distribution of vegetation. The two principal patterns of distribution of vegetation in the dry areas of the Near East are related to dry watercourses, which are known as wadis from the Arabic language or arroyos in the southwestern United States. Wherever plants grow on slopes and depressions, the pattern is regarded as “diffused.” In extremely dry desert areas, where vegetation is restricted to wadis that receive additional water supply, the pattern is “contracted.” In an area in which the climatic conditions enable the development of diffused vegetation on most soil types, silty or clayey soils which are relatively dry soils in desert will locally support contracted vegetation. Plants demanding relatively high quantities of moisture may inhabit special habitats of wadis in zones of contracted vegetation. Such special habitats are derived from local contribution of runoff water from rocks to their crevices and soil pockets. In hard and fissured limestone, dolomite, granite, and metamorphic rocks, much of the rainfall is available to the semishrubs growing there because water infiltration is good and the soil in the rock fissures is leached. Accumulating deep in the soil and the weathered rock, this water is protected from direct evaporation by rocks and stones. Trees which grow on slopes of the Mediterranean zone that receives 300--700 mm of annual rainfall occur in desert areas with 100 mm or even less in proximity to outcrops of smooth-faced hard rocks. These rocks do not absorb much water and their crevices receive high amounts of water through runoff (Danin, 1999).

One of the ways in which plants are adapted to desert conditions is that they avoid the extremely dry season, which may last from 6 months to several years. Thus, annual plants may be seen in some habitats every few years. This depends on edaphic and climatic conditions; for example, annual halophytes may germinate and grow to produce seeds only on salty soils and only in rainy years when the soil solution is diluted to the appropriate level. On the other hand, annuals which develop in crevices of smooth-faced rocks may have sufficient resources to be sustained almost every year. Thus, species diversity in the desert may be high in areas in which habitat diversity is high, especially if there are sufficiently large outcrops of smooth-faced hard rocks which contain many different microhabitats.

**K. Shrub–Steppes**

The area of the Negev Highlands, the Judean Desert, Sinai, and southwestern Jordan, which receives 80--250 mm of rainfall annually, is covered with semishrubs at a diffused pattern, thus forming shrub–steppes (Fig. 3, 10). The most common dominants in these steppes are *Artemisia sieberi*, *Noaea mucronata*, and *Cymopteris decandra*. The phytomass produced by annuals in the plant communities developing on stony or rocky shallow soils is always small when compared with that of fine-grained and deep soils. The latter types hold much of their water close to the soil surface, thus losing much of it through direct evaporation. The minute quantities of 8 ppm of salts carried by clouds and later by rain from the Mediterranean sea by climate systems remain in the soil and accumulate there. The soil may be too dry or too saline for the growth of annuals in regular or dry years. Nearly monospecific communities of semishrubs exist, each best adapted to the specific local saline conditions. The most common dominants under these conditions are *Kraunharia hirtella*, *R. negevensis*, *Salvia vermiculata*, *Bassia (Chenopodium) arbores*, and *Atriplex glauca* on chalk- and marl-derived soils; *Anabasis syriaca* and *Haloxylon scoparium* are the shrubby dominants on loess-derived soils. However, in most years there is much development of annuals, although in patches with high salinity there are monospecific patches of salt-resistant annuals. Showy geophytes such as species of *Tulipa*, *Iris*, *Ixiolirion*, *Ranunculus*, and *Anemone* may bloom in large quantities in the shrub–steppes in moist years. Outcrops of smooth-faced hard limestone support plants that differ much from those on the other soil types. This vegetation is typically characterized by *Chilanthus* (phloxes), *C. montanus*, *Globularia arbutifolia*, *Stachys aegyptiacus*, *Polypagia negevensis*, *Tanacetum sini- cam*, and *Capparis aegyptica*. Isolated populations of dozens of Mediterranean relics and many rare desert plants are found in this habitat in the Negev, the Judean Desert, Sinai, and Jordan. The semishrub *Sarcopoterium*...
Shrubs of Retama raetam and Achillea fragrantissima are the dominants in wadis of the terrain of hard limestone. Atriplex hallusmae prevails in this habitat, in which the catchment area is built up from the salty soils on chalk, clay, or marl. At lower elevations, Acacia raddeana, A. pachyceras, Tamarix mitis, and T. aphylia also occur. Many small springs even in the limestone hills of western Sinai and the sandstone hills of southwestern Jordan. Most of these springs can be detected from afar by the date palm (Phoenix dactylifera), which is confined to sites with a high water table of fresh water. It is accompanied by Euphorbia retusa, Janus arabicus, Phragmites australis, and Cressa cretica. Canyons occur in many wadis and may have long-lasting water pools supplied by floods and support rich flora of hydrophytes such as Zannichellia palustris and Potamogeton spp. and green algae such as Chara spp.

L. Shrub–Steppes with Trees

Shrubs–steppes are the most characteristic vegetation type of the Central and Eastern Mediterranean. They represent a natural refugium which has functioned for a long time in the evolutionary history of the region. The Shrub–Steppes with Trees subunit 11c are the richest both in nondesert trees and shrubs and in companions. The large outcrops of smooth granite and the high elevation of southern Sinai Massif result in a high number of habitats available for the survival of rare species. The typical trees of the rocky environment in 11c are Pistacia khusjuk, Crataegus sinaicus, and Ficus pseudosycomorus. The west-facing escarpments of Gebel Serbal are rich in Moringa peregrina, which also grows in the vicinity of Wadi Feiran oasis. The typical shrubs of the rocky habitat of 11c are Rhamnus disperma, Rhus tripartita, Colocsinster orbicularis, Pteroloxus aphyllos, and Sageretia thea. Most of the endemic and rare species of Sinai occur in the crevices of rocks that also support trees. The Q. calliprinos and J. phoeica woodland is marked 11d on the vegetation map. The climatically controlled belt of the arborescent Mediterranean vegetation terminates in the vicinity of Amman, approximately 120 km north of the Dana–Tafila area. Large areas of shrub–steppe, semisteppe batha, and steppe–forest typify the western ridge of the Jordanian plateau between At–Tafila and Petra. These are dominated by Artemisia sieberi, Noaea mucronata, and spiny species of Astragalus. The occasional arborescent components include Pistacia atlantica, Crataegus aronia, J. phoeica, and Q. calliprinos. These formations develop on fissured limestone, basalt, and chalk rocks. The smooth-faced hard sandstone outcrops of the Dana–Petra area support the richest relic Mediterranean flora in the Near East. The trees and shrubs growing here are Q. calliprinos, P. atlantica, P. palaestina, P. khusjuk, A. areonia, J. phoeica, Amygdalus korschikiewitschi, Ceratonia siliqua, Olea europeica, Arbatus andrubus, Rhamnus punctata, R. lycioides, R. disperma, Ficus carica, F. pseudosycomorus, and Sageretia thea. Typical root parasites of the Mediterranean maquis, such as Oxyris alba and Thesium bergeri, occur in the rock crevices, as do typical vines of the maquis—Rubia tenaliula, Ephedra foeminea, Hedera helix, Bryonia cretica, and Lonicera etrusca, which are represented by a higher number of individuals than other Mediterranean vines in any other refugia in the Near East. The rich Mediterranean flora with many endemics may be regarded as existing in a successful refugium which has functioned for a long time in the evolutionary history of the region.
M. Desert Vegetation

Sparsely vegetated shrub–steppes dominated by Anabasis articulata and Zygophyllum damarum occur at the broad boundary area of the desert with the steppes zone (Fig. 3, 1, 2 and 10). These typical semishrubs of the desert grow in this transition zone in a diffused pattern. Chalk and marl outcrops are populated with the xerohalophyte communities of Salsola australis, Salsola teretandra, and Haloxylon negevensis. The nearly monospecific communities of semishrub xerohalophytes are accompanied by a diverse assemblage of herbaceous plants that grow on the leached soil only in rainy years. The extreme desert areas may have no higher plants growing out of wadis, but the entire area may be covered by a highly diverse microbiotic crust. Dor and Danin (1996) studied the microbiotic succession and changes in floristic composition of the crust in the Dead Sea area.

The sequence of plant communities along the wadis of a more arid zone of this category is typified by a stretch of annuals which occur in rainy years. The nature of the annuals community is heavily influenced by the edaphic and climatic conditions of the site. A lower section of the wadi receives higher quantities of water and supports small and short-living semishrubs such as Pulsatilla incisa, which may function locally as an annual. Further down, larger and long-living semishrubs such as A. articulata or Z. dumosum grow. A section dominated by shrubs such as R. raetam, Ochradenus baccatus, or Lycium shawii prevails further down the wadi system. In the lowest section of the wadi system, trees, mostly Acacia species or Tamarix species, may be found. The nature of plant communities and the sequence of their occurrence along the wadis are in close affinity to rock and soil types, which greatly influence the moisture, salinity, and nutrient regime in the wadi systems.

In similar gravel plains in the large desert area south of Ma’an to the Saudi Arabian border, A. articulata grows with trees of A. pachyceras at the fifth-order section of the wadi system. The rocky terrain at the escarpments of southwestern Jordan to the Arava rift valley supports diverse communities in a diffused pattern as occurs in Sinai and the Negev.

N. Sand Vegetation

The main areas in which sand dunes or sand sheets occur in Israel (Fig. 3, 13) are the Mediterranean coastal plain, the western Negev, a few valleys of the northeastern Negev, and the Arava Valley. Different climatic regime prevails in each of these areas, sand texture differs, and hence the vegetation and processes of its development differ accordingly. Much of this vegetation, patterns of species diversity, and the nature of the microbiotic crust of the sand in desert areas are discussed in other publications. Thus, some background concerning the sand vegetation of southern Jordan is presented here.

Weathering of nubian sandstones in the Jordanian plateau contributes mobile sand to the Arava Valley. The relatively high water table at the Arava enables the successful development of the tall shrub Haloxylon persicum at sites in which sand is sufficiently deep. However, most stands of this plant in Israel became intensively irrigated agricultural areas. Nonetheless, large areas of Haloxylonetum persici cover considerable areas of the Arava and the southeastern desert of Jordan (Al-Eisawi, 1996). Annuals accompany the shrubs of sand sheets in rainy years. The sandstone area of the Jordanian plateau is built of sandstone inselbergs which project from large flat valley and are filled with stable sand. Huge areas of sand sheets in the Wadi Rum area are dominated by Haloxylon salicornicum, A. articulata, and occasional patches of H. persicum. The sandstone hills support many relict Mediterranean species and there is particularly interesting vegetation at the vicinity of the contact zone between the sandstone and the Precambrian crystalline rocks.

O. Oases with Sudanian Trees

The climate of the rift valley is much warmer than that of the hilly terrain surrounding it. Springs in which a high quantity of fresh water is available throughout the year (Fig. 3, 14) support thermophilous Sudanian trees such as date palms (P. dactylifera), A. raddiana, Acacia tortilis, Calotropis procera, Moringa peregrina, Balanites aegyptiaca, Cordia sinensis, Maerua crassifolia, Dolichegia siso, Capparis decidua (in Jordan), and Z. spina-christi. Rich annual flora may accompany sites which are not under the influence of the springwater but successfully grow in rock crevices, soil patches, and abandoned cultivated land. In many freshwater springs/oases near the Dead Sea in Israel and Jordan and in southern Sinai, Adiantum capillus-veneris grows on dripping water at shady places with the orchid Epipactis venetofolia. A prominent species of the freshwater springs in the Jordanian desert, Nerium oleander, is absent from the desert springs of Israel and Sinai.

P. Desert Savannoid Vegetation

A considerable part of the Arava and Dead Sea Valleys is covered by savannoid vegetation in which Acacia

363
The upper soil layers support in wadis the typical desert vegetation. *Acacia pachyceras* is the dominant tree of areas which are at a relatively high elevation, such as the upper tributaries of Nahal Paran in Israel, Wadi Jirali in Sinai, and large areas of gravel plains in Jordan south of Ma‘an. *Acacia raddiana*, which is less resistant to low temperatures, is the dominant tree at lower elevations, and *A. tortilis*, which has the highest demands for high temperatures, grows in the southern Arava or below sea level in the northern Arava and Dead Sea Valley. In a few places between Wadi Watir and Sharm el Sheik in eastern Sinai, rare trees of *Capparis decidua* grow, which is an important component of the savanna vegetation of southern Egypt and Sudan.

**Q. Swamps and Reed Thickets**

Only small nature reserves, such as that of the Hula Lake, remain (Fig. 3, 16) from the swamps which covered large areas at the beginning of the twentieth century. Freshwater springs still flow and their vegetation is composed of a small number of species that produce high quantities of phytomass. The most common hydrophytes are *Phragmites australis*, *Arundo donax*, *A. pliniiana*, and *Typha domingensis*. Large areas of swamps in the Hula Valley supported *Cyperus papyrus*, which reached its northernmost station at this valley. Several rivers of the coastal plain that supported riparian vegetation in the past have become sewage canals and their polluted water has destroyed most of the past rivers’ vegetation.

**R. Wet Salinas**

Wet, salty soils occur in places in which springs of salty water occur or at sites in which the water table is near the surface and the evaporating water leaves salt in the upper soil layers (Fig. 3, 17). A salt crust may occur at the soil surface in arid regions and plants are restricted to small wadis in which leaching occurs. Most plants growing in the desert salt marshes are perennials that establish themselves in the rare event of leaching. The typical plants of desert salty soils are *Saussara montana*, *S. fruticosa*, *S. vermiculata*, *Nitraria retusa*, *Seidlitzia rosmarinus*, and a few species of *Tamarix*.

**S. Synanthropic Vegetation**

The category of synanthropic vegetation (Fig. 3, 18) is further divided in Israel into three subcategories according to the remnants of trees found in the intensively cultivated areas: In 18a, it is trees of *Quercus ithaburensis*, in 18b the tree is *Z. spina-christi*, and in 18c it is *A. raddiana* and *Z. spina-christi*. The synanthropic vegetation of Jordan and the current status of the synanthropic vegetation in Sinai need further investigation.

**T. Mangroves**

The mangrove of Nabq, eastern Sinai (Fig. 3, 19), typically growing in muddy soils of the tidewater, constitutes the most northerly population of *Avicennia marina* on Earth at 28°10′N. This species was recorded even farther from the equator in South Australia at 37°S. The only population found in the Gulf of Suez shares its water, and possibly warmth, with the Gulf of Elat at Ras Mohammed.

See Also the Following Articles

DESERT ECOSYSTEMS • MEDITERRANEAN-CLIMATE ECOSYSTEMS • NEAR EAST ECOSYSTEMS, ANIMAL DIVERSITY

Bibliography


NEST PARASITISM

Scott K. Robinson* and Stephen I. Rothstein†
*University of Illinois at Urbana-Champaign and †University of California at Santa Barbara

I. Natural History
II. Coevolution between Brood Parasites and Their Hosts
III. Modeling Host–Parasite Coevolution
IV. Impacts on Host Population Dynamics
V. Ecology and Social Behavior of Brood Parasites
VI. Conspecific Brood Parasitism
VII. Conclusions and Research Needs

GLOSSARY

acceptance Hosts treat brood parasitic eggs as if they were their own eggs.
coevolution Cycles of adaptation and counteradaptation that occur between interacting lineages.
coevolutionary arms race The continuing bouts of co-evolving defenses and counterdefenses that occur between hosts and parasites.
conspecific brood parasites Brood parasites that lay their eggs in the nests of other individuals of the same species.
cowbirds American blackbirds in the genus Molothrus, which contains important brood parasites.
cuckoos A family of birds of which approximately half the species (61 of 125) are obligate interspecific brood parasites.
egg rejection Host responses to parasitism that include ejection of parasitic eggs, abandonment of parasitized nests (usually with renesting), or a new cup built over a parasitized clutch.
evolutionary equilibrium hypothesis The hypothesis that frequencies of acceptance of brood parasitism reflect an equilibrium between costs and benefits of host defenses against parasitism.
evolutionary lag hypothesis The hypothesis that hosts lack defenses against brood parasites because the defenses have not yet had time to evolve.
generalist brood parasites Species that parasitize many (up to more than 200) host species.
genets (singular gen) Lineages of cuckoos in which individual females specialize on a single host and lay mimetic eggs.
host The individual that is parasitized.
interspecific brood parasites Brood parasites that lay their eggs in the nests of other species.
mafia effect Interspecific brood parasites that may destroy clutches from which parasitic eggs have been ejected by the host.
mimicry Brood parasitic eggs or nestlings that closely match those of the hosts.
specialist brood parasites Species that parasitize only one or a few host species.

BROOD PARASITISM, also called social parasitism, is the exploitation by one individual (the brood parasite) of the parental care of another (the host). Brood para-
sites can deposit eggs in the nests or broods of another individual of the same (conspecific brood parasitism) or of a different (interspecific brood parasitism) species. Hosts often raise young of the brood parasite, typically at the expense of their own young. This article presents a comprehensive overview of the natural history, evolution, and consequences of brood parasitism, with a special focus on birds, the taxon in which it has been best studied.

I. NATURAL HISTORY
Brood parasites lay their eggs in the nests of other individuals, which then raise the parasitic young at the expense of part or all of their own brood. Brood parasitism was noted by Aristotle and even earlier (approximately 2000 B.C.) in India. Brood parasitism is best known in birds, approximately 1% of which are obligate interspecific brood parasites that only lay their eggs in the nests of other species. Interspecific brood parasitism, however, has also been documented in insects and fish. Conspecific brood parasitism, in which individuals facultatively lay eggs in nests of conspecific individuals, is more widespread.

Brood parasites have generated intense interest in the public and scientific communities. Brood parasites tend to be vilified in the media because of a human tendency to moralize about such things as killing baby birds (which parasitic birds often do) and because at least some may pose a conservation threat to some of their hosts. Among scientists, brood parasites have generated intense debate about the coevolutionary processes that may be responsible for the seemingly maladaptive acceptance of parasitic eggs by hosts.

A. Adaptations of Obligate Avian Brood Parasites
Brood parasites search for host nests, synchronize their laying with that of the host, and often remove a host egg from nests they parasitize. Brood parasitic eggs often have unusually thick eggshells, are usually small relative to the size of the parasite (but large relative to the size of the host), often mimic the coloration of the hosts’ eggs, and have rapidly developing embryos. All these traits increase fitness of the parasite by reducing competition with host nestmates and making parasitic eggs more difficult to detect and remove. Egg mimicry is most pronounced in specialist brood parasites. In the well-studied common cuckoo, the species as a whole is a generalist, but each individual female is a specialist member of a “genus.” Any one region has only one to several coexisting genotypes but different arrays of genotypes occur in different regions. Brood parasitic nestlings also have a variety of mechanisms that enhance their ability to compete with host nestlings. Cuckoo nestlings have concave backs that they use to push out host eggs and nestlings, whereas several other brood parasites have specialized bill hooks that they use to kill nestmates. Some brood parasites apparently increase the amount of food hosts bring to them by mimicking the juvenile plumages or complex mouth markings (the “gapes”) of host nestlings or by having large mouths and intense begging behavior.

Because brood parasites do not have to engage in costly breeding activities such as incubation and nestling feeding, they often have more time and energy to devote to egg production. Some generalist brood parasites lay 60 or more eggs in a season; estimates for one tropical brood parasite suggest that more than 100 eggs/year are typical. Other brood parasites, however, may be only slightly more fecund than their hosts.

B. A Survey of Avian Brood Parasites
There are 90–95 species of obligate avian brood parasites in five unrelated families (Table I).

1. Cuculidae
The cuckoos are a diverse family that contains both parasitic and nonparasitic species. The Cuculinae has traditionally been recognized as an Old World subfamily, all of whose approximately 50 species are obligate

<table>
<thead>
<tr>
<th>Taxon (family/subfamily)</th>
<th>No. of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old World cuckoos (Cuculinae)</td>
<td>50</td>
</tr>
<tr>
<td>New World cuckoos (Narinemphinae)</td>
<td>3</td>
</tr>
<tr>
<td>Honeyguides (Africa and Asia) (Indicatoridae)</td>
<td>17?</td>
</tr>
<tr>
<td>Volume finches (Africa) (Ploceidae)</td>
<td>14</td>
</tr>
<tr>
<td>Cuckoo-finches (Africa) (Ploceidae)</td>
<td>1</td>
</tr>
<tr>
<td>Cowbirds (New World) (Icteridae)</td>
<td>5</td>
</tr>
<tr>
<td>Black-headed duck (South America) (Anatidae)</td>
<td>1</td>
</tr>
</tbody>
</table>

* Recent work indicates that at least one genus in this subfamily belongs in another cuckoo subfamily (see text).
brood parasites. Most species are host specialists that are relatively uncommon and only a few have been well studied (the common cuckoo and the great spotted cuckoo). Another traditionally recognized subfamily, the Neomorphinae or New World ground cuckoos, has 11 species, 3 of which are relatively rare obligate parasites. The latter tend to parasitize hosts that build domed nests and are poorly known overall. Nestlings of one species, *Tapera naevia*, have pincher-like bills that they use to kill host nestlings, a case of convergence on honeyguides (Indicatorinae). Recently published DNA sequence data indicate that at least 1 genus (*Clamator*) within the Cuculinae is more closely related to the Phaenicophaeinae, a subfamily of New World cuckoos. None of the currently recognized members of the Phaenicophaeinae are obligate parasites but some parasitize conspecifics and occasionally other species. These new DNA data place obligate parasitism within three groups of cuckoos, two of which have both parasitic and non-parasitic species, and indicate that parasitism evolved separately up to three times in cuckoos or that some parasitic cuckoo lineages reverted to parental behavior.

2. Indicatoridae
The honeyguides, a family named for the habit of one species that guides people and the African honey badger or ratel to beehives, has 17 obligately parasitic species. The species of *Indicator* and two related genera parasitize cavity nesters, whereas 3 species of *Proctodacys* parasitize open-cup nesters. Most species are poorly known and the eggs of many species have never been described. Some honeyguides have raptorial hooks on their bills, which they use to kill host nestlings.

3. Viduine Finches
The approximately 16 species of obligate brood parasitic viduines occur in Africa and provide some of the best examples of nesting mimicry, particularly of the intricate gape patterns of their hosts. Compared with other brood parasites, viduines exert relatively low costs on their hosts, which can usually raise mixed broods. Most *Vidua* species specialize on single species of grassfinches (*Estrildidae*), to which they may be related. Unlike most brood parasites, which are insectivorous, hosts of viduines are granivorous. Some species of viduines incorporate the songs of their hosts into their own songs, which are then used by females to choose males that have been raised by the same species.

4. Cuckoo-Finch (*Anomalospiza imberbis*)
This African species is closely related to Viduine finches and parasitizes approximately 11 species of warblers in grasslands. Recently reported DNA evidence suggests that parasitism evolved once in a single lineage that then gave rise to both *Anomalospiza* and the Viduines.

5. Icterinae
The five species of New World brood parasitic cowbirds include the most generalized of all brood parasites, the brown-headed and shiny cowbirds, both of which parasitize more than 200 host species. Another species, the tropical giant cowbird, parasitizes almost exclusively colonial American blackbirds. Of the two remaining species, the bronzed cowbird is a host generalist (>80 host species), whereas the screaming cowbird is one of the most specialized of all brood parasites. It mainly parasitizes a single species of blackbird, which was formerly thought to be a species of cowbird. Unlike most brood parasites, cowbirds are often more abundant than many of their hosts and can pose a significant threat to populations of rare, localized host species. Most cowbirds benefit greatly from changes in the landscape caused by humans, which has enabled several species to expand their geographic ranges and increase their population sizes. As a result, they are coming into contact with new hosts that have not had recent contacts with brood parasitism. Some of these new hosts suffer extremely high levels of parasitism. Cowbirds appear to be extremely fecund; the shiny cowbird may lay more than 100 eggs a year. The invasion of North America by this cowbird may pose an additional threat to species that are not currently being parasitized by the brown-headed cowbird. The brown-headed cowbird may be the most intensively studied North American bird.

6. Black-Headed Duck (*Heteronetta atricapilla*)
This species is the only obligately parasitic species with precocial young that feed themselves when they hatch. For this reason, this brood parasite has little effect on the nesting success of its hosts, which only have to incubate an extra egg or two. Because incubation is so similar among all bird species, *Heteronetta* successfully parasitizes a wide range of birds, such as gulls and ibises, as well as other ducks.

7. Conspecific Brood Parasitism
Recent studies have shown widespread conspecific brood parasitism, especially in waterfowl, gallinaceous birds, and in a small number of songbirds (generally species with rare nest sites or that breed colonially).
8. Other Brood Parasite Systems

a. Fish
An African catfish is the only fish species known to be an obligate brood parasite. It parasitizes mouth-breeding cichlids and apparently consumes all of its host’s young while in the mouth of its foster parent. Conspecific brood parasitism and occasional parasitism of other species occur in many nest-building fish. Some host defenses such as increased nest guarding may have developed in response to this parasitism but egg recognition has not been reported.

b. Insects
A range of insects practice brood parasitism, especially members of the order Hymenoptera (bees, wasps, and ants). Some of these species are obligate brood parasites, such as certain ants that kill the queens of nonparasitic species and use the workforce of the entire colony to rear their own young. Other species are facultative parasites that victimize conspecifics or raise their own young. To our perception, some of these parasitic insects look remarkably different from their hosts so their acceptance by hosts seems inexplicable. However, olfaction is critical to these insects and it is likely that parasitic species have evolved chemical cues that mimic those of their hosts. Because both humans and birds perceive the world largely via vision and audition, and have relatively poor olfactory abilities, it is not surprising that parasitic birds have been subjects of much more research than have parasitic insects, even though the latter undoubtedly have many equally interesting examples of coevolved adaptations.

II. COEVOLUTION BETWEEN BROOD PARASITES AND THEIR HOSTS

A. Cuckoo–Host Systems
Cuckoo–host systems are much more highly coevolved than the other well-studied systems, those of cowbirds, and DNA divergence data show that they are also much older. Most cuckoo species use only one or a few host species. Some species, such as the Eurasian common cuckoo, might be called generalists when viewed over their entire ranges because they parasitize 50–100 or more host species. However, even these cuckoos are more properly viewed as specialists because they use only one to several host species in each region and individual females, or members of a gens, specialize on a single host species or several similar host species. In regions where several cuckoo species coexist, they tend to show little or no overlap in host use, which reduces competition for hosts. Most cuckoos are much less abundant than their hosts and probably have little effect on host population dynamics because they affect only a small proportion of their host populations.

Some cuckoo–host systems may have reached an evolutionary equilibrium, whereas others show strong evidence of ongoing coevolution and provide some of the strongest examples in vertebrates of microevolution that has occurred during ongoing research.

B. Adaptations of Cuckoos
1. Behavior
Cuckoos typically approach nests stealthily and may quickly drop their thick-shelled eggs onto host eggs to increase chances of breaking host eggs. Males of some cuckoos resemble hawks. This resemblance or simply host recognition of cuckoos incites host aggression and males in some species use this aggression to draw hosts away from nests. This male distraction is coordinated with and facilitates nest entry by the stealthy and drab-colored females in several cuckoo species. Nestlings of some cuckoos mimic the begging calls or plumage of host young. There is evidence that great spotted cuckoos revisit nests they have parasitized and destroy host eggs or nestlings if the parasitic egg has been removed by the host. Such a protection racket or ‘mafia effect’ could select for acceptance of host eggs.

Cuckoos find nests by watching hosts from hidden perches and laying is synchronized with that of the hosts. Most cuckoos lay one egg per nest, except in a few species in which cuckoo nestlings do not kill those of the host. Cuckoos defend territories against other cuckoos and usually remove a host egg before laying one of their own.

2. Fecundity
Cuckoos lay 8–25 eggs per season, which is generally more than their hosts lay. Eggs are laid every other day.

3. Egg Adaptations
Cuckoo eggs usually mimic the coloration of their hosts’ eggs. Cuckoo that parasitize many host species have individually specialized females, with each group of females or gens (plural gentes) mimicking the eggs of a single host species or a group of hosts with similar eggs. Egg mimicry is usually assumed to have arisen in response to host egg recognition. An alternative hypoth-
ESIS argues that egg mimicry evolved to keep a second cuckoo from identifying and removing a cuckoo egg from an already parasitized nest and then laying its own egg in the nest. However, it is not clear that more than one cuckoo attempts to parasitize the same host nest often enough for this to be a significant selective pressure. Most cuckoos parasitize hosts that are much smaller than themselves and have correspondingly small eggs for their body sizes. However, cuckoo eggs are usually slightly larger than those of the hosts. The few species that parasitize hosts larger than themselves do not have smaller eggs. The thick shells of cuckoo eggs may be adaptive because they reduce chances of breakage of cuckoo eggs during rapid laying, facilitate breakage of host eggs (thereby reducing competition with hosts), or make cuckoo eggs resistant to pecking by hosts. A counteradaptation to egg mimicry that some hosts have developed is variable egg types. In response, some cuckoo species have evolved an equivalent range of egg types; but hosts still seem to benefit because individual female cuckoos lay only one type and may not always parasitize a host female who lays the same type. The highly variable host eggs may also function in the context of conspecific parasitism.

4. Nestling and Fledgling Adaptations

Nestlings in most cuckoo species have concave backs that they use to push host nestlings or eggs out of nests. Cuckoos that parasitize larger hosts do not show this eviction behavior but can compete successfully with host nestlings because they beg more loudly and incessantly, hatch earlier, and develop more rapidly. Mimicry of host begging calls occurs mainly, but not exclusively, in species that do not evict host nestlings.

C. Host Defenses Against Cuckoos

1. Defenses against Adults

Hosts often respond aggressively to cuckoos, but this may actually provide a cue to cuckoos about their stage of the nest cycle and the proximity of nests. Mobbing cuckoos has both genetic and learned components. One possible advantage of mobbing cuckoos is that it may trigger ejection of parasitic eggs.

2. Egg Rejection

Hosts reject cuckoo eggs by abandoning parasitized nests or ejecting eggs directly. Small hosts are more likely to abandon (and renest), perhaps because ejection results in the incidental breakage of some of their own eggs and is therefore costly or because they are simply too small to eject cuckoo eggs. Egg rejection is most prevalent in hosts that have a long coevolutionary history with cuckoos. These hosts, which often show fine abilities to discriminate among egg types, are species that have intrinsic characteristics that make them suitable for parasitism, such as animal food appropriate for the nestling parasites and nests accessible to adult cuckoos. Species unsuitable as hosts, such as ones that have specialized nestling diets (e.g., seeds in some finch species), have presumably never been parasitized intensely in the past and generally lack egg recognition altogether. Among suitable hosts, egg recognition is more prevalent in species that are currently rarely parasitized than in species that are currently common hosts. This trend suggests that cuckoos shift away from using suitable hosts with well-developed egg recognition. It is possible that cuckoos have a dynamic system of host usage in which they repeatedly switch from hosts with well-developed defenses only to switch back to these hosts after their defenses have declined and defenses in other hosts have increased. Such a system could result in never-ending cycles of host switches and coevolution, but it is also possible that hosts retain high levels of egg recognition for long periods in the absence of parasitism, which would force cuckoos to become increasingly more specialized. Long-term retention is indicated by high levels of egg recognition in New World magpies and shrikes, which are currently not parasitized by any brood parasites but are descended from Old World ancestors that are cuckoo hosts. Populations of some suitable hosts that have only recently come into contact with cuckoos have apparently undergone rapid increases in egg ejection and discrimination, apparently because they possessed some recognition even in the absence of parasitism. This possession could be due to retention of recognition from past bouts of parasitism or to gene flow from parasitized populations of the same species.

3. Nestling Discrimination

Despite there being many species that have some degree of mimicry of host nestling appearance or begging calls, there are no known cases in which cuckoo hosts show outright rejection (i.e., removal) of nonmimetic nestlings. This lack of outright rejection appears to be true in all other systems of avian parasitism that show equivalent or even better mimicry of host nestlings. Therefore, the selective value of nestling mimicry may relate to ensuring that parasitic nestlings receive high-quality care from their hosts rather than avoiding outright rejection.
D. Cowbird–Host Systems

The five brood parasitic cowbirds include one of the most specialized of all brood parasites and the two most generalized of all brood parasites. We have chosen to highlight cowbird–host interactions in this article because there is a vast literature on cowbird–host systems and because these systems are quite different from cuckoo–host systems.

1. Cowbird Adaptations

a. Egg Adaptations

As described previously, some cowbirds are extremely fecund, an apparent reflection of a trade-off between egg production and parental care. In captivity, brown-headed cowbirds can lay up to 77 eggs in the short temperate breeding season and field estimates of shiny cowbird productivity exceed 100 eggs during the much longer tropical breeding season. As in cuckoos, the eggs of the smaller cowbirds are small relative to the size of the cowbird but are larger, thicker shelled, rounder, and faster to develop than those of most of their hosts. Only the giant cowbird, which parasitizes hosts as large as itself, has normally proportioned eggs. The other smaller species usually parasitize smaller hosts. Egg mimicry has not been firmly established in any cowbirds, but cowbird egg shape and patterns have evolved, perhaps in response to diffuse selection by dominant hosts or groups of hosts. The two most generalized cowbirds, the shiny and brown-headed, have generalized spotted egg colorations that are similar to those of many passerines. Although cowbird egg coloration is not finely tuned to host egg coloration, as in cuckoos, it has undergone some shifts, presumably in response to host use, during the evolution of the cowbird lineage. Different cowbird species have generalized spotted eggs or immaculate unspotted ones; however, two species, the giant and shiny cowbirds, have both kinds of eggs.

b. Behavioral Adaptations of Adults

Similar to cuckoos, cowbirds search for nests primarily by observing the behavior of hosts. Cowbirds approach nests stealthily and lay eggs very early in the morning before most hosts appear at their nests during the egg-laying period. In contrast, giant cowbirds appear to circumvent the defenses of colonial hosts by visiting colonies as a group in which males appear to distract hosts while females stealthily enter nests. Screaming cowbirds also visit their communally breeding hosts in groups, but males do not appear to help distract the hosts. Cowbird eggs are usually synchronized properly with host laying but improperly timed eggs that appear before the host eggs or after hosts have finished laying occur more commonly than in cuckoos. Although generalist cowbird species avoid parasitizing some species that are clearly unsuitable as hosts, such as doves, there is little evidence that they select the best host species in a community. Many cowbird eggs are seemingly wasted in nests of hosts that feed their nestlings inappropriate diets of seeds or fruit or that eject cowbird eggs. However, parasitism of the latter species may sometimes be adaptive for two reasons. First, some species that exercise egg recognition and ejection have low nest predation rates. Second, these species learn the appearance of their own eggs by imprinting on the first egg or eggs that they lay during their lives. Accordingly, naive hosts parasitized about the time they begin laying may come to learn cowbird eggs as their own eggs and may provide cowbirds with nests that have a relatively low likelihood of failing due to nest predation. Multiple parasitism (two or more cowbird eggs in a nest) occurs in approximately one-third of all nests parasitized by cowbirds, especially with some large hosts. Several female cowbirds sometimes lay in the same nest even though multiple parasitism usually reduces the success rate of cowbird eggs and many host nests receive no cowbird eggs. Female cowbirds are usually highly aggressive toward one another in habitats in which hosts occur (but not in which cowbirds feed), but there is still substantial overlap among home ranges. Cowbirds sometimes, but not always, remove host eggs by puncturing them. Egg removal is rarely done during the same nest visit but may occur the afternoon before or later in the morning on which cowbird eggs are laid. There are many observations of cowbirds depredating unparasitized host nests and there is evidence that cowbirds do this regularly to stimulate renesting, which will increase the future availability of host nests to be parasitized. However, some studies have found no evidence for this depredation-renesting hypothesis and the behavior may occur only in special circumstances, if at all.

c. Nestling and Fledging Adaptations

In the specialized screaming cowbird, there is almost perfect mimicry of host plumage and vocalizations of its usual host species, the bay-winged cowbird (which is not closely related to the parasitic cowbirds despite its name). This mimicry does not seem to be essential for nestling care but instead seems to be an absolute requirement for receiving host care after the parasites fledge from host nests. Such perfect mimicry is lacking in the other, more generalized cowbirds. However, giant cowbird nestlings have white bills, like those of their...
NEST PARASITISM

oropendola hosts. Bills turn the usual cowbird black after the fledglings become independent. Nestling cowbirds grow more rapidly than the nestlings of many, but not all, of their hosts and beg more loudly. They have relatively large mouths and show intraspecific variation in their gape colors, unlike the majority of nonparasitic nestlings. This gape variation may relate to host use in some unknown way. There is a recent report of a cowbird ejecting a host nestling, although the ejection could have been inadvertent.

2. Host Defenses against Cowbirds
   a. Preventing Access to Nests
      Many hosts react aggressively to cowbirds and colonial nesting may provide protection for some hosts. Aggressive responses to cowbirds, however, may reveal nest locations to cowbirds and be counterproductive for small hosts, which may not be able to drive cowbirds away from nests. Some hosts sit on their nests when approached by a cowbird, but there are records of cowbirds pulling hosts off their nests. Some hosts reduce parasitism by nesting in cavities or in dense vegetation, but even cavity nests are often parasitized in many species.
   b. Egg Rejection
      Cowbird host species can be divided into acceptors, which accept cowbird eggs, and rejecters, which abandon parasitized nests, build a new nest on top of the old nest, or, most commonly, eject cowbird eggs. Within acceptor species, nearly 100% of the individuals accept cowbird eggs. Rejecter species show nearly 100% rejection and there is little geographic variation in a species' responses to cowbird eggs. Birds that eject cowbird eggs either puncture them or grasp them and usually drop them at least several meters from the nest. Small hosts tend to puncture–eject, whereas larger hosts grasp–eject. Nest abandonment occurs most frequently when cowbird eggs are laid too early in a host's laying cycle and when hosts are too small to eject eggs (or to eject without breaking too many of their own eggs). Nest desertion appears to be triggered by encounters at the nest between hosts and cowbirds. In nearly all host species, detection of cowbird eggs in nests does not appear to play a role in eliciting desertion, even in hosts whose eggs are highly divergent from cowbird eggs. Some hosts may begin incubating early in the morning to prevent undetected cowbird parasitism. Most species that reject cowbird eggs occur in open habitats where they may have a long coevolutionary history with cowbirds, all of which forage in open areas in association with grazing ungulates. Hosts that eject cowbird eggs recognize their own eggs and remove cowbird eggs even if these outnumber their own eggs, unless they have misimprinted on cowbird eggs laid in their very first nest.

III. MODELING HOST–PARASITE COEVOLUTION

There is little doubt that the adaptations and counteradaptations described in Section II result from coevolution between parasites and their hosts. Parasitic counterdefenses to host defenses must correspond to the characteristics of hosts, as in mimicking of host eggs. Thus, parasites cannot have effective defenses against more than a few hosts at any one time because hosts vary in key characteristics such as egg coloration. The traditional view therefore hypothesizes that parasites specialize on increasingly fewer hosts as a parasite–host system becomes older and more hosts evolve defenses. This view is supported by studies showing that (i) recently derived brood parasites tend to be more generalized than older species (e.g., cowbirds versus cuckoos); (ii) the costs of parasitism are often much higher than the costs of potential host defenses, which suggests that there is a continuing evolutionary arms race; (iii) some systems in which hosts and parasites have only recently come in contact show either an absence or low levels of host defenses; (iv) host defenses have evolved since recent contact with a brood parasite; (v) parasites have evolved egg mimicry in response to host defenses; and (vi) cuckoo–host systems are dynamic as evidenced by host shifts in parasitic cuckoos. In accord with these generalizations, most Old World passerines in Africa and Eurasia, which are exposed to numerous species of an ancient group (the cuckoos), possess egg recognition. However, most New World passerines, which are exposed to a smaller and relatively younger group of parasites (the cowbirds) lack egg recognition.

The question then arises, why do so many hosts accept easily distinguishable parasitic eggs? Two hypotheses have been proposed: “evolutionary lag” and “evolutionary equilibrium.” There is a burgeoning literature by proponents of each hypothesis. The evolutionary lag hypothesis is that these acceptors have not yet had time to evolve antiparasite adaptations. Acceptance of parasitic eggs by some hosts in heavily parasitized, unproductive populations may also result from gene flow from populations in areas where cowbirds are absent and reproduction is high. The lag hypothesis is clearly applicable to situations in which parasites have
began to use host species with no recent history of exposure to parasitism but may even apply to some host species with long histories of parasitism because egg recognition/rejection is a totally new feature that may be difficult to evolve. In the evolutionary equilibrium hypothesis, hosts accept parasitic eggs because of the high costs of rejection. In this scenario, parasitism is costly, but trying to reject parasitism would be even more costly, i.e., acceptors are making the best of a bad situation. Distinguishing between lag and equilibrium has proven to be difficult. Testing either hypothesis requires measurements of the costs of rejection in species that accept, which has not yet been done effectively. Sometimes, likely costs for acceptors can be estimated from similar species that reject. For example, the number of their own eggs that small ejecting hosts break when they eject cowbird or cuckoo eggs can be extrapolated to acceptor hosts of similar size. These rejection costs, generally less than 0.5 host eggs per ejection, do not seem to be sufficiently high to make acceptance a more adaptive option because acceptance usually results in the loss of all the host’s young for cuckoo hosts and for small cowbird hosts. There is evidence that rejecter hosts mistakenly eject their own eggs on occasion, but it is not clear if this cost is sufficient to outweigh the demonstrable costs of egg acceptance. Equilibrium may be most likely in species for which parasitism is relatively infrequent and in which parasitic eggs resemble those of the host, which greatly increases the probability of mistakes. Nevertheless, distinguishing conclusively between lag and equilibrium will require new experimental and phylogenetic studies.

IV. IMPACTS ON HOST POPULATION DYNAMICS

Brood parasitism is undoubtedly almost always costly for individual hosts, but the extent to which brood parasites threaten host populations is much less clear. Most brood parasites have little impact on host populations, presumably because coevolutionary processes reduce the frequency of parasitism and because most parasites are much less abundant than their hosts. In Europe, for example, less than 5% of most host nests are parasitized by the common cuckoo. One brood parasite, the giant cowbird, has even been hypothesized to benefit its hosts because its nestlings remove parasitic booby eggs from its host nestmates.

Nevertheless, brood parasites can pose threats to host populations, especially generalist brood parasites and those that have undergone recent range expansions and population increases. In a few cases, cuckoos have expanded their ranges as a result of human-caused increases of early successional habitat. In areas recently colonized by cuckoos in Japan, levels of parasitism often exceed 40%, although in most of these areas the incidence of host defenses appears to be increasing rapidly, probably because these host populations already had some level of defense before being parasitized. Such rapid development of host defenses, however, appears to be absent from hosts of the brown-headed and shiny cowbirds. Furthermore, range expansions are much more pervasive in these two generalist cowbirds than in any cuckoo species. Cowbirds have undergone enormous population increases as the result of human activities, especially those associated with cattle. As a result, many host populations with no recent history of parasitism are being exposed to massive levels of brood parasitism. For some species, especially forest-nesting ones, widespread parasitism may be occurring for the first time in the history of the species’ lineage. Parasitism levels of many forest species nesting in the midwestern United States, for example, exceed 80% with most nests receiving multiple cowbird eggs. Most newly exposed host lineages lack effective defenses. Several recent studies have linked large-scale population declines of songbirds with increasing levels of cowbird parasitism. For some species, especially forest-nesting ones, receiving multiple cowbird eggs. Most newly exposed host lineages lack effective defenses. Several recent studies have linked large-scale population declines of songbirds with increasing levels of cowbird parasitism. Concerns about cowbird parasitism have stimulated an enormous amount of research since the late 1980s and two major symposia have been published recently on this subject. This increased focus on cowbirds has resulted in pressure to trap and kill cowbirds to reduce their impacts on hosts. However, host species that have declined have experienced massive loss of habitat and increased rates of nest predation in addition to increased rates of cowbird parasitism. For example, most passerine species that breed in riparian habitat in the southwestern United States have declined in the past century but so too has their habitat. Dams, water diversions, overgrazing, urbanization, and exotic plants have resulted in the degradation or loss of more than 90% of the riparian habitat present a century ago in the Southwest. Even if cowbirds are not the primary cause of some of all declines, they may now be exposing some reduced host populations to additional stresses that threaten the populations with extinction. Importantly, a cowbird population can be stable or grow even as it pushes a rare host
to extinction because individual female cowbirds do not specialize on single host species. Nevertheless, it is not clear whether cowbird parasitism threatens more than a few host species. Species that are at most at risk are those with small geographic ranges that are wholly included within areas that contain abundant cowbird foraging habitat (pastures, feedlots, mowed grass, and bare soil) and that cannot be rescued by emigrants from more productive populations. This list includes several endangered species that are brown-headed cowbird hosts (Kirtland’s warbler, black-capped vireo, and southwestern willow flycatcher) and many species restricted to islands in the Caribbean that have recently been invaded by shiny cowbirds (yellow-shouldered blackbird and Puerto Rican vireo).

Most species in North America have very large geographic breeding ranges that include regions in which cowbirds are rare and restricted to areas near human habitations. These refugia from cowbird parasitism tend to occur in large, unfragmented habitats that may act as “sources” of surplus host young that can recolonize populations in more fragmented habitats in which levels of parasitization (and nest predation) are often very high. In these population “sinks” in fragmented habitats, levels of reproduction may be too low to compensate for adult mortality; that is, such populations can only be sustained by immigration from source habitats. Evidence for this source–sink scenario derives from well-documented sink populations in fragmented midwestern U.S. forests. These populations are nevertheless relatively stable, probably due to immigration from populations in large, unfragmented forests in the region in which levels of both parasitism and nest predation are very low and reproductive success is sufficiently high for populations to act as sources. Such large-scale source–sink population dynamics can slow the evolution of host defenses because most young are being produced in areas in which cowbird parasitism is rare.

A. Cowbird Management

Many cowbird control programs have been initiated as a result of concern regarding several endangered songbird species. Controlling cowbirds, which is easily done because their social nature attracts them into traps that contain cowbirds that function as decoys or “bait,” has become a multimillion dollar a year business in the American Southwest and there are also active programs elsewhere in North America and in the Caribbean. Local control efforts, in conjunction with habitat restoration, may have prevented the extinction of several endangered host species and races, and two endangered hosts have increased in population since cowbird control began. However, control programs for two other endangered hosts resulted in no increases in the sizes of host breeding populations, even though all control programs have resulted in increases in host reproductive output. Unfortunately, even when host populations increase, cowbird control must be done every year because cowbird removal has little or no year-to-year effect on the numbers of cowbirds that occur in an area due to high dispersal by cowbirds. Therefore, most workers view cowbird control as a temporary, stop-gap measure, although a U.S. Fish and Wildlife recovery plan for the least Bell’s vireo in California advocates cowbird control in “perpetuity.” Larger scale control programs, such as killing cowbirds by the millions at winter roosts, have been suggested by some workers but do not seem justified given that regional breeding season control programs are effective in eliminating nearly all cowbirds from the ranges of endangered hosts. Such large-scale killing programs may also raise important ethical issues and some workers have argued that an undue emphasis on cowbirds may detract from more productive and more long-lasting management actions such as habitat restoration.

V. ECOLOGY AND SOCIAL BEHAVIOR OF BROOD PARASITES

The lack of parental care in brood parasites potentially sets them apart from most other birds in their foraging ecology, mating systems, spacing behavior, and vocal development.

A. Foraging Ecology

Most brood parasites have unusual diets and foraging behavior. Honeyguides eat wax, many cuckoos eat hairy and toxic caterpillars, and cowbirds prefer to forage in short grass close to ungulates such as cattle, horses, or bison. In each of these cases, either the diet of the brood parasite would be difficult for nestlings to digest (wax and hairy caterpillars) or the foraging habitats may be so ephemeral (the proximity of ungulates) that it may not be available for an entire nesting cycle. It is not known, however, if these unusual foraging ecologies were precursors of brood parasitism or if they were made possible by the evolution of brood parasitism.
B. Mating Systems

Mating system theory predicts that birds freed from the needs of parental care should be promiscuous. Nevertheless, avian brood parasites show a remarkable array of mating systems, including monogamy and resource-based polygyny. This variation in mating behavior has been attributed to the diverse ways in which parasites gain access to nests (some of which require cooperation between several individuals), to the great variety of foraging ecologies of brood parasites, and to such unusual features of individual host–parasite systems as strongly male-based sex ratios in cowbirds. In general, there have been few detailed studies of the mating systems of brood parasites. The mating system of the brown-headed cowbird has received the most study among parasitic birds and has been described as promiscuous, polygynous, or monogamous. Promiscuity has been proposed because it is a common sight to see two or more males associating with a single female. However, studies of color-marked cowbirds that enabled researchers to determine which males and females mated together and associated with one another the most have indicated monogamy. A recent DNA fingerprinting study confirmed that monogamy prevails and even found that cowbirds have fewer matings outside of their pair bonds than do most nonparasitic songbirds. Field observations show that the mate faithfulness in cowbirds is due both to males guarding females from the advances of other males and to females being reluctant to mate with males other than their usual consort. Because males outnumber females, many males do not acquire a mate. A single study demonstrated promiscuity in an area in which extremely high cowbird abundances may have made it difficult for the birds to maintain pair bonds.

In one Asian honeyguide, males defend beehives and only allow access to wax if the females mate with them. In some viduine weavers and cuckoos, males display from prominent perches and are chosen by females presumably on the basis of mate quality, which may be indicated by song elements or plumage.

C. Spacing Behavior

Many brood parasites defend breeding areas that are rich in hosts. At least some cowbirds and many cuckoos defend home ranges against conspecifics. However, the home ranges are technically not true territories because several males and females often occupy the same area. In areas of home range overlap, some host nests are often parasitized by several female cowbirds. Cuckoos may occupy areas that are more mutually exclusive and thus this is closer to classic territoriality. In areas of low abundance, cowbirds may occupy mutually exclusive areas and thus may appear to be territorial. Cowbirds often breed and feed on a daily basis in different areas that can be separated by as much as 15 km. After searching for nests in the morning, cowbirds commute to feeding sites in pastures, plowed fields, and in other habitats with short mowed grass. The uncoupling of breeding and feeding areas made possible by brood parasitism allows cowbirds to select a wide variety of breeding habitats, even if there are only a few foraging sites in a region. Some cuckoos also appear to have very large home ranges. Home range size is related to a bird’s body size and trophic level, with predators needing especially large areas. Although cowbirds and cuckoos are not predators of other birds, adults breeding birds are a key resource for them, and they have home ranges similar in size to those of raptors that feed on adult birds.

D. Vocal Behavior

Brood parasites have provided some of the clearest examples of genetically hard-wired vocal behavior, but recent studies have also shown a key role for subsequent learning in the modification of vocalizations. Because cowbirds are raised by different species, it is not surprising that at least one of their songs, the perched song, is genetically programmed. Cowbirds, however, learn to modify this song in response to female preferences and interactions with other males and develop individual repertoires of approximately five perched song types. A second song type, the flight whistle, is almost totally learned. It occurs as discrete spatial dialects generally tens of kilometers in diameter, which are so variable that trained observers may fail to recognize as cowbird vocalizations whistles from dialects they have not yet experienced. These dialects are true examples of culture and show that considerable amounts of biodiversity within a species can be related to learned/cultural differences rather than to genetic differences. Unlike the majority of songbirds that complete their vocal development by the time they are 1 year old, male cowbirds in some regions do not master local versions of whistles and perched songs until they are 2 years old. Vocal learning occurs among one other parasitic group. Male viduine weavers incorporate elements of
NEST PARASITISM

their host’s song into their repertoire, which presumably enables females to choose a male raised by the same host species. As such, vidiuine weavers show a high potential for sympatric speciation by cultural learning. Such assortative mating maintains coevolutionary adaptations such as the gape mimicry that is specific to just one host. Assortative mating is essential because gape appearance is due to paternal and maternal genes. In contrast, common cuckoo and cowbird males do not incorporate host songs into their repertoires, perhaps because egg appearance, which is essential for acceptance by their hosts, is determined solely by the maternal genotype.

VI. CONSPECIFIC BROOD PARASITISM

Female birds sometimes lay their eggs in the nests of conspecifics. This behavior has been documented in more than 100 bird species and has also been documented in some insects and fish. Among birds, conspecific brood parasitism (CBP) is most prevalent among precocial species, such as ducks and gallinaceous birds, and among altricial species that breed colonially and/or use specialized nest sites, such as cavities. CBP is rare or absent in most songbirds, even some that are colonial, and it has been well studied in swallows.

In many cases, CBP occurs when individuals lose their nests during the laying period or are unable to obtain a territory or nest site. In these situations, individuals may be making the best of a bad situation. In some species, CBP may be an alternative reproductive strategy that increases reproductive success by spreading the risk among several nests, reducing competition within a female’s own nest, and exploiting the parental care of others. Recent theoretical work suggests that CBP is a necessary precursor to the evolution of interspecific brood parasitism. Conspecific brood parasitism is especially prominent in New World cuckoos of the genus Coccyzus, which tend their own nests and also parasitize conspecifics and congeners. In at least some species, such as the ostrich and some ducks, individuals may benefit from CBP because extra eggs or young dilute losses to predators.

Nest defenses against CBP parallel those used to combat interspecific brood parasitism. Hosts guard their nests, especially at high population densities in barn swallows, which may exert a significant cost of nesting colonially. Parasitic females of another swallow species, the cliff swallow, actually carry their eggs between nests. The rejection of parasitic eggs laid before the host has begun laying is a common adaptation against CBP but does not necessarily involve egg recognition. True egg recognition, in which a bird discriminates among egg types, is a rare response to CBP because it is difficult to evolve because parasitic and host eggs are similar in appearance, given that they are from the same species. Also, CBP is less deleterious to hosts than interspecific parasitism because it does not involve adaptations for killing host nestlings or asymmetries in size and incubation period that accomplish the same end. However, egg recognition in response to CBP does occur in some birds such as ploceid weaver finches of Africa and Asia, which have extensive variation among the eggs of conspecific females. Indeed, it has been hypothesized that the extreme intraspecific egg variation of these weavers is an evolutionary response to CBP. Some species with typical amounts of variation in egg appearance may lay smaller clutches than they could potentially feed to leave room for parasitic eggs. If hosts cannot recognize parasitic eggs, then laying a smaller clutch might avoid starvation of most or all nestlings when parasitism does occur.

CBP may actually affect population sizes. In species with frequent CBP, fitness may be reduced at high population densities, which may cause populations to be cyclic. Nests of wood ducks that are placed too close together are subject to extreme levels of CBP, which can reduce population fitness.

VII. CONCLUSIONS AND RESEARCH NEEDS

Brood parasitism raises fascinating questions about co-evolution and conservation. Studies of brood parasitism provide some of the strongest evidence of microevolution yet documented in vertebrates. The extreme mobility of cowbirds and their potential threats to host populations illustrate the importance of landscape-level processes in conservation biology. Cowbird song development has become a model system showing that both genetic and environmental factors are important to the development of behavior in general. Despite these and other lessons learned from studies of brood parasitism, there are many unanswered questions that are vital to our understanding of this subject.

We still have much to learn about topics such as
(i) the frequency of recognition errors in hosts with
rejection behavior, (ii) reasons for the general lack of parasite nestling recognition by hosts, (iii) the frequency of "mafia"-like behaviors and nest predation by brood parasites, (iv) how brood parasites choose from available hosts, (v) whether or not cowbirds pose a significant threat to populations of widespread host species, and (vi) the genetic mating systems of brood parasitism. The recent invasion of the United States by a new generalist brood parasite, the shiny cowbird, also offers an excellent opportunity to study its interactions with other cowbirds and its new hosts. Finally, most host–parasite systems remain very poorly known, especially those of the New World cuckoos, many Old World cuckoos, and some old honeyguides. Some members of the two latter groups are so poorly known that there is no direct evidence that they are parasitic. However, they are assumed to be so because closely related species are known to be parasitic.

See Also the Following Articles

BIBLIOGRAPHY


NITROGEN, NITROGEN CYCLE

Sandy L. Tartowski and Robert W. Howarth
Cornell University

I. An Overview of the Nitrogen Cycle and Its Ecological Importance
II. Human Impacts on the Nitrogen Cycle
III. Effects of Nitrogen on Biodiversity
IV. Effects of Biodiversity on Nitrogen Dynamics
V. Considerations for the Future

GLOSSARY

anoxia The absence of molecular oxygen.
biogeochemistry The discipline which studies biotic controls on the chemistry of the environment and geochemical controls on the structure and function of ecosystems.
denitrification The reduction of nitrate or nitrite to gaseous nitrogen products, mainly N₂ and N₂O, by bacteria.
deposition The delivery of material inputs to the earth's surface from the atmosphere.
eutrophic High in production or species typical of high-production environments.
mesotrophic Intermediate in production or species typical of intermediate-production environments.
mineralization The conversion of an element from an organically bound form to an inorganic form.
nitrification The oxidation of ammonia to nitrite and nitrate by bacteria.
nitrogen fixation The reduction of atmospheric N₂ to ammonia or other organic or inorganic compounds by bacteria, lightning, or photochemistry.
nitrogen saturation Nitrogen supply in excess of the capacity of the ecosystem to retain nitrogen.
nutrient limitation The addition of a nutrient or nutrients to an ecosystem causes an increase in net primary production.
oligotrophic Low in production or species typical of low-production environments.
opportunistic species Species typical of transient, unstable, unpredictable, frequently disturbed, or periodically extreme environments, usually having strong dispersal abilities and faster growth, smaller size, and shorter life spans than potential competitors.
primary production Synthesis of organic matter from carbon dioxide by photosynthesis.

NITROGEN IS AN ESSENTIAL ELEMENT FOR LIFE and it frequently controls productivity and community structure in both terrestrial and aquatic ecosystems. During the past few decades, human activity has changed the nitrogen cycle more than any other element cycle, resulting in many deleterious environmental changes, including decreased biotic diversity.

I. AN OVERVIEW OF THE NITROGEN CYCLE AND ITS ECOLOGICAL IMPORTANCE

Nitrogen (N) is an essential element for all life on Earth, a required component of DNA, all proteins (including
enzymes), chlorophyll, and many other critical constituents of biological structure and function. More than 99% of nitrogen on Earth is present as molecular N₂, with most of this in the atmosphere and some dissolved as a gas in the oceans. However, most organisms cannot use the relatively inert N₂ molecule. Paradoxically, molecular N₂ is not a thermodynamically stable compound in the presence of oxygen, but these two gases coexist and comprise virtually all the atmosphere. The nitrogen in N₂ becomes biologically available only when the N-N bond is broken and nitrogen is combined with other elements to form compounds such as ammonium (NH₄⁺) and nitrate (NO₃⁻) in a process called nitrogen fixation. Prior to the massive alteration of the global nitrogen cycle by human activity, small amounts of nitrogen were fixed by lightning and by volcanic activity, but the vast majority of nitrogen fixed on Earth each year (more than 95%) was fixed by bacteria. Not all bacteria are capable of fixing nitrogen, although the ability is fairly widespread among diverse groups of bacteria including cyanobacteria, non-photosynthetic bacteria (such as purple sulfur bacteria), and many classes of heterotrophic bacteria. No other organisms are capable of nitrogen fixation, although many plants, lichens, mosses, and some animals have bacterial symbionts which fix nitrogen. These plants with symbiotic nitrogen-fixing bacteria include many species of trees (such as alders in the temperate zone and many species in moist tropical forests) and many agricultural crops, including alfalfa, peas, peanuts, soybeans, and rice. Plants with symbiotic nitrogen-fixing bacteria are often loosely called "nitrogen fixers," although it is always the bacteria that are actually fixing nitrogen.

For most of the history of life on Earth, the demand by plants for biologically available forms of nitrogen such as nitrate and ammonium has been greater than the rate of supply by nitrogen fixation. Consequently, nitrogen limits the primary productivity in many ecosystems, and when more biologically available nitrogen is added to these systems plant production increases. This relative lack of availability of nitrogen globally has also been a major factor shaping the evolution of life and structuring the communities in many different types of ecosystems, both terrestrial and aquatic.

The general patterns which relate productivity and diversity in ecosystems are well established. In aquatic ecosystems, species diversity declines with increasing productivity, and diversity is greatest in systems of very low productivity, such as oligotrophic lakes and areas of low productivity in the oceans in which co-occurrence by multiple nutrients (including nitrogen) may be common. In terrestrial ecosystems, the greatest diversity of plant species usually occurs in sites of intermediate fertility and productivity. In sites of very low fertility, the resources can support only a few viable populations able to tolerate the severe conditions. In very high fertility sites, competition intensifies and may shift from competition for nutrients to the more asymmetrical competition for light or space, resulting in dominance by a few species. For both aquatic and terrestrial systems, when nitrogen is limiting production, it is also influencing diversity.

The importance of nitrogen as a factor limiting productivity and regulating community structure varies among the earth's ecosystems, with nitrogen being of paramount importance in some systems and not limiting at all in others. Although there are exceptions, among terrestrial ecosystems nitrogen availability generally controls production in temperate forests, boreal forests, arctic and alpine tundra systems, temperate and tropical grasslands, and agroecosystems. On the other hand, phosphorus or other elements—and not nitrogen—tend to limit production in most tropical forests if any element is limiting. Among aquatic ecosystems, nitrogen limits production in most estuaries and coastal seas of the temperate zone and perhaps in some tropical seas, although these are sometimes phosphorus limited. Nitrogen with phosphorus as a co-limiting element may regulate production in much of the oceans away from shore as well. However, phosphorus is generally more important in regulating production in freshwater lakes, at least in moderately or highly productive lakes in the temperate zone.

These patterns of nutrient limitation are the result of many processes which regulate the availability of nitrogen in comparison to other elements essential for plant growth. For terrestrial ecosystems, one important aspect is the change in phosphorus biogeochemistry as soils develop over geological time. In a geologically young soil, the initial amount of phosphorus is often high, but as the soil is weathered this phosphorus is exported from the system or stored in mineral forms which make the phosphorus biologically unavailable. In the temperate zone, soils are often young due to glaciation within approximately the past 20,000 years. On the other hand, many tropical soils are extremely old and highly weathered and therefore have very low availabilities of phosphorus. This is one factor which tends to make tropical forests phosphorus limited and temperate forests nitrogen limited.

Differences in nitrogen processes among systems are also important in determining which element is most limiting to primary productivity. Nitrogen fixation is one critically important process, and it is much more
prevail most at higher rates in tropical forests than in temperate forests. This contributes to the greater tendency for nitrogen limitation in temperate forests.

An interesting paradox is that nitrogen fixation does not alleviate nitrogen shortages in temperate forests and thereby alleviate nitrogen limitation. Why should any ecosystem be nitrogen limited if nitrogen fixation could provide sufficient inputs? Why is nitrogen fixation in temperate forests so much less than that in tropical forests? The reasons for the relatively low rates of nitrogen input to nitrogen-limited temperate forests remain poorly known and are an active area of research. A similar paradox occurs in aquatic ecosystems. One of the reasons that freshwater lakes tend to be limited by phosphorus is that planktonic cyanobacteria in these systems can often fix sufficient nitrogen to alleviate any potential nitrogen limitation. In contrast, nitrogen fixation by planktonic cyanobacteria rarely occurs in even strongly nitrogen-limited estuaries. Recent research suggests that the relative lack of nitrogen fixation in estuaries is due to an interaction of two factors: Trace metals that are required for the process of nitrogen fixation (molybdenum and iron) are less available in estuaries than in lakes, slowing the potential growth rate of nitrogen-fixing organisms, and this slow growth rate makes cyanobacteria highly susceptible to grazing mortality by zooplankton and benthic animals.

Nitrogen fixation is only one process affecting nitrogen availability in an ecosystem. Also important are the processes which regulate the loss of nitrogen from the system and the input of biologically available forms of nitrogen from neighboring ecosystems and from the atmosphere. A brief summary of the nitrogen cycle illustrates the complexity. The nitrogen that is fixed accumulates over time in both the living and nonliving components of ecosystems and is transported and transformed by numerous processes (Fig. 1). Ammonium and nitrate are taken up by plants as well as by microorganisms and converted to organic nitrogen compounds. This organic nitrogen is cycled through food webs but also accumulates in soils, water, and sediments as organics die, excrete matter, or drop tissues such as the fall of leaves from trees. Much of the nitrogen in this nonliving organic matter is not readily available for use by organisms, but microbial decomposition of the organic matter releases ammonium, which can then be oxidized to nitrate by bacteria. Both nitrate and ammonium are highly soluble in water, but the negatively charged nitrate ion is more easily leached from soil and transported to aquatic ecosystems. In ecosystems where oxygen concentrations are low (such as in the sediments of wetlands, lakes, and estuaries), the nitrogen in nitrate can be converted back to molecular N₂ (or the slightly oxidized gas N₂O) by bacteria in a process called denitrification. Also, in soils and waters where the pH is sufficiently high to favor the dissociation of ammonium to ammonia (NH₃), nitrogen can be lost to the atmosphere as ammonia gas. Fires can also volatilize both ammonia and oxidized gases of nitrogen (such as NO and N₂O) to the atmosphere. Most of these forms of nitrogen are redeposited from the atmosphere onto the earth's surface in precipitation or as dry deposition.

II. HUMAN IMPACTS ON THE NITROGEN CYCLE

Human activity has altered the nitrogen cycle globally more than that of any other element cycle, and much of the change has occurred during the past 50 years. The increase of carbon dioxide and other greenhouse gases in the atmosphere is, appropriately, a cause of widespread alarm, but the rate at which humans have altered nitrogen availability on land is even greater. As recently as 30-40 years ago, the major source of fixed nitrogen was still natural biological nitrogen fixation, but today more nitrogen is fixed through human activities than by all nitrogen fixation in natural terrestrial ecosystems on Earth (Fig. 2). The greatest change has been the widespread use of inorganic nitrogen fertilizer, and the production of such fertilizer today accounts for more than half of the total amount of nitrogen fixed by all human activities. The process for making nitrogen fertilizer from atmospheric N₂ was invented in the early twentieth century, but such fertilizer was not widely used until the 1950s. The rate of use has increased steadily since then, with only a modest interruption caused by the collapse of the former Soviet Union and the disruption of agricultural in Russia and eastern Europe in the early 1990s. Despite that interruption, half of the inorganic nitrogen fertilizer that has ever been used on the planet has been used during the past 15 years.

The synthesis of inorganic fertilizer accounts for approximately 60% of the total amount of nitrogen fixed globally through human activities, with the rest coming primarily from the production of nitrogen-fixing crops in agriculture and from the combustion of fossil fuels. Fossil fuel combustion not only releases nitrogen from geological storage in the fuel but also catalyzes the reaction of O₂ and N₂ in the air drawn into internal combustion engines. This nitrogen is released as ox-
dized nitrogen gases, generically referred to as NOₓ, which are subsequently deposited onto the earth's surface in precipitation or dry deposition. Overall, human fixation of nitrogen (including production of fertilizer, combustion of fossil fuel, and production of nitrogen-fixing agricultural crops) increased globally approximately two- or threefold between 1960 and 1990 and continues to increase. By the mid-1990s, human activities made fixed nitrogen available at a rate of approximately 140 Tg year⁻¹, or more than 0.8 g N m⁻² year⁻¹ when averaged over the entire land surface of the earth.

Nitrogen from human sources can travel for fairly long distances in the environment. For instance, much of the nitrogen deposited from the atmosphere in the northeastern United States comes from the combustion of fossil fuels in the Midwest, and the majority of nitrogen flowing down the Mississippi River and causing eutrophication in the Gulf of Mexico originates from agricultural sources in Illinois, Iowa, Indiana, Minne-
sota, and Wisconsin. Of the nitrogen fertilizer applied to agricultural fields, some leaches into groundwater and surface water and affects downstream ecosystems; on average in the United States, approximately 20% of nitrogen fertilizer is thus exported from agroecosystems, although this percentage ranges from only 3 to 80% depending on the soil type, climate, and agricultural practices. Nitrogen from agricultural sources also gets volatilized into the atmosphere and thereby redistributed onto nonagricultural lands, including forests. Some of this nitrogen is volatilized directly from the agricultural fields, as both ammonia and NO\textsubscript{x}, with the flux of NO\textsubscript{x} being particularly important in tropical areas. Also of importance is the volatilization of ammonia to the atmosphere from animal wastes. In developed countries, approximately half of the nitrogen used as fertilizer is exported from the field in crop harvest, on average, with most of these crops being fed to livestock and poultry. Much of the nitrogen in these feedstocks ends up in the animal wastes, with a high percentage being volatilized. The total flux of such volatilization in the United States is almost as large as the leaching of nitrogen directly from fertilized fields.

Reactive nitrogen compounds are usually transported for only a few hundred kilometers to at most approximately 1000 km through the atmosphere, and rivers and ocean currents transport biologically available forms of nitrogen on the same sort of spatial scales. As a result, the alteration of the nitrogen cycle is not uniform over the earth, and the greatest changes are concentrated in the areas of greatest population density and greatest agricultural production. Generally, the largest changes have occurred in the northern temperate zone. The tropics have seen less change, at least to date. Most terrestrial ecosystems and most coastal marine ecosystems in the temperate zone are nitrogen limited, and therefore the acceleration of nitrogen cycling in the northern temperate zone has had a great impact. Current deposition of nitrogen from the atmosphere onto the landmasses of the Northern Hemisphere temperate zone is almost seven times more than the preindustrial deposition, and in some regions, such as the northeastern United States, Western Europe, and eastern Asia, the increase has been far greater. The natural rate of biological nitrogen fixation in European watersheds of the North Sea and in the northeastern United States is approximately 0.5 g N m\textsuperscript{-2} year\textsuperscript{-1} or less and human activity has increased nitrogen input to the watersheds of the North Sea to 7.5 g N m\textsuperscript{-2} year\textsuperscript{-1} and in the northeastern United States to 4.1 g N m\textsuperscript{-2} year\textsuperscript{-1}. The export of nitrogen in major rivers to coastal oceans is closely related to the increased inputs of nitrogen to the basins from human activities (Fig. 3). Downstream transport of nitrogen to estuaries and coastal oceans has increased up to 20-fold in some areas such as the North Sea.

The rate at which human activity has accelerated nitrogen cycling varies widely among regions of the

---

**FIGURE 3**  Export of nitrogen from rivers emptying into the North Atlantic Ocean is strongly correlated with the net nitrogen inputs to the watersheds caused by human activities, especially agriculture and combustion of fossil fuels (reproduced with permission from Howarth et al., 1996).
world. Globally, the use of inorganic nitrogen fertilizer continues to increase, but in the United States and most other developed countries fertilizer use has changed little since 1980. In China, on the other hand, the use of inorganic nitrogen fertilizer has more than doubled since 1980. By 1995, China was using almost one-third of all inorganic nitrogen fertilizer used globally, and the rate of use there continues to increase. The acceleration of nitrogen use in developing countries has allowed major increases in food production and has greatly reduced starvation, but the environmental consequences are becoming increasingly apparent.

III. EFFECTS OF NITROGEN ON BIODIVERSITY

Human acceleration of the nitrogen cycle has tremendous implications for biodiversity at all scales, from the genome to the globe. Reduced genetic diversity within species exposed to high levels of nitrogen deposition has been documented in tree species, soil mycorrhizae, and *Rhizobium spp.* in beans. Changes in species composition and reductions in species diversity are obvious and widespread. Declining ecosystem and landscape diversity are becoming evident. In return, changes in biodiversity influence nitrogen dynamics and processes, creating complex nonlinear interactions and feedbacks between nitrogen cycling and biodiversity. Nitrogen enrichment can reduce biodiversity by several mechanisms including acidifying soil and water as well as increasing the growth and dominance of a few particularly responsive species in nitrogen-limited ecosystems.

A. Effects of Eutrophication on Biodiversity

The substantial increase in nitrogen inputs to nitrogen-limited ecosystems as a result of human activities alters species composition (which species are present) and decreases species diversity (the number of species and the evenness of their relative abundance). A few fast-growing opportunistic species, with greater ability to take advantage of the increased availability of nitrogen to produce additional biomass, usually become dominant in fertilized ecosystems, while the species characteristic of less fertile environments (oligotrophic species) decline and disappear. Often, a relatively diverse flora, including rare endemic plants, characteristic of an oligotrophic ecosystem is replaced by a more productive, but less diverse, flora dominated by a few common weedy species, normally found in mesotrophic or eutrophic sites. Usually, invasive pest plants respond strongly to increased availability of nitrogen, and generally eutrophication facilitates the spread of introduced and native pest plants. As plant species are lost from an ecosystem, additional animal species associated with those plants may be lost as well, further decreasing species diversity. The conversion of whole ecosystems from oligotrophic to eutrophic and the disappearance of entire types of ecosystems encompass the extinction of large groups of species, not just a few particularly vulnerable endangered species, and reduces biodiversity on the local, landscape, and regional scale.

The decrease in species diversity caused by the eutrophication of terrestrial and aquatic ecosystems has been observed consistently across a wide variety of nitrogen-limited ecosystems. However, the speed and magnitude of the effects of increased inputs of nitrogen on biodiversity depend on the characteristics and history of the affected ecosystem. The mobilization and redistribution of nitrogen by human activity can cause local areas of nitrogen depletion, even though the amount of fixed nitrogen has increased overall. Semi-arid and arid lands and unfertilized agricultural and overgrazed systems are particularly susceptible to the loss of nitrogen-containing topsoil. In some cases, nitrogen depletion can threaten biodiversity much as nitrogen addition reduces biodiversity. In very unproductive ecosystems, the addition of nitrogen may cause an initial, usually transient, increase in species diversity by allowing the invasion of native species from more fertile sites in the same region or exotic species introduced from other regions. Even when species diversity remains elevated, often the oligotrophic species are lost from the ecosystem, contributing to a regional decline in species diversity. The effects of nitrogen enrichment may be delayed in systems which are severely nitrogen limited due to previous depletion of nitrogen by fire or human activities, such as forest clearing or timber harvesting.

Enrichment of grasslands with nitrogen causes productivity increases in a few dominant grasses and severe declines in species diversity, as shown by experiments in North America, Australia, and Europe (Fig. 4). In one long-term study of fertilizer addition to a grassland in England, species diversity was reduced by more than fivefold. Oligotrophic, calcareous grasslands in Europe, with high biodiversity and many endemic species, are particularly sensitive to species loss due to eutrophication and most sites have already experienced changes in species composition. Fertilization and increased productivity decrease species diversity by allowing fast-growing, tall plants to shade competitors and sometimes
Nitrogen, Nitrogen Cycle

by increasing production of litter which prevents the establishment of seedlings. Increased availability of nitrogen suppresses nitrogen-fixing plant species and consequently many of the forbs lost from eutrophied grasslands and heaths are nitrogen-fixing species. The mechanisms by which nitrogen enrichment reduces biodiversity may be quite specific to the circumstances of each ecosystem. For example, when nitrogen was added to fertile wet and infertile dry sedge meadows in Colorado, diversity declined in the wet meadow but not in the dry meadow. In the infertile dry meadow, nitrogen additions released rare species from nitrogen limitation, increasing the evenness of species abundances, but in the more fertile wet meadow the already dominant species increased in abundance, suppressing subordinate species. Nitrogen inputs affect not only plant diversity but also the diversity of the animal community. It has recently been suggested that nitrogen deposition near San Francisco may be a contributing factor leading to the decline of a threatened, endemic butterfly.

In many ecosystems, moderate levels of grazing—within the range typical of the evolutionary history of the system and plants—tend to increase plant species diversity. Grazing can interact with nitrogen availability to affect diversity, and ungrazed or lightly grazed grasslands are more susceptible to decreases in biodiversity caused by nitrogen enrichment because heavy grazing usually suppresses the dominant grass species stimulated by nitrogen. Nitrogen enrichment usually reduces the spatial heterogeneity of nitrogen and therefore reduces the potential for different species to coexist using microsites with different availabilities of nitrogen. Grazing redistributes nitrogen and usually increases the spatial heterogeneity and temporal variability of nitrogen availability, contributing to the maintenance of biodiversity. However, when fertilization increases productivity sufficiently to support greatly increased grazing, grazing-intolerant plant species may be lost from the ecosystem.

Diversity in forests is also affected by nitrogen additions, as has been particularly well studied in parts of Europe in which nitrogen deposition from the atmosphere is high. In both deciduous (the Netherlands and France) and coniferous forest (Sweden and Finland), nitrogen enrichment has caused shifts in the species composition of the understory, favoring opportunistic species with high nitrogen requirements and reducing biodiversity of lichens, mosses, and vascular plants. Increased nitrogen deposition has caused weakened tree trunks, decreased frost hardiness, and increased pest outbreaks in some forests. After many years of high nitrogen deposition, some forests of Western Europe have become “nitrogen saturated,” unable to store additional nitrogen in soil or vegetation. In nitrogen-saturated ecosystems, continuing nitrogen inputs are balanced by comparable nitrogen outputs, accelerating both the eutrophication of downstream waters and the loss of other essential elements from the forest.

In the Netherlands, the high population density, with accompanying industry and vehicles, is combined with intensive livestock production to generate perhaps the highest rates of nitrogen deposition in the world. More than 35% of oligotrophic, diverse heathlands in the Netherlands have suffered nitrogen accumulation, invasion by grasses, and reduced diversity. In dry heathlands, this nitrogen fertilization has stimulated outbreaks of the heather beetle, accelerating the conversion to grassland. Some heaths have been completely converted to much less diverse grasslands and forest, reducing the landscape and regional diversity as well as the diversity within the heathlands. The beginnings of similar eutrophication of heathlands have been observed in Norway and in the United Kingdom.

Estuaries and other coastal waters receive a variety of insults from human activities. Perhaps the greatest pollution problem, however, is the eutrophication that results from increased nitrogen inputs. Eutrophication of coastal marine ecosystems frequently results in hypoxia (the depletion of dissolved oxygen) or anoxia (the absence of dissolved oxygen), both of which result not only in fish kills and the loss of other biotic resources but also in sharp decreases in biotic diversity. Hypoxia has severe effects on the diversity and abundance of benthic species, shifting dominance from large, long-
lived, less mobile species to smaller, mobile, opportunistic, short-lived species. Frequent hypoxia may prevent successional development beyond the early colonizing community. The loss of burrowing benthic organisms that irritate and oxygenate the sediments may strongly influence biogeochemical processes such as phosphorus adsorption and nitrification and denitrification, which can feed back to influence diversity. Increased nitrogen inputs have caused increased hypoxia and anoxia in Chesapeake Bay, the Baltic Sea, the Black Sea, Long Island Sound, Florida Bay and the Florida Keys, the northern Adriatic Sea, and many other areas globally. The hypoxic “dead zone” in the northern Gulf of Mexico increased from 9500 km² when first reported in 1991 to 20,000 km² in 1999 and is the direct result of nitrogen inputs coming down the Mississippi River.

Not all the effects of eutrophication on biodiversity are due to hypoxia and anoxia. Nutrient enrichment also leads directly to changes in the composition of the phytoplankton community, and these can cascade up the food web generally leading to lowered diversity at each trophic level. Eutrophication of coastal waters is often accompanied by a decrease in silica availability, due both to increased sedimentation of silica within estuaries and to trapping of silica in upstream, eutrophic fresh waters. Since silica is required by diatoms, but not by other types of phytoplankton, eutrophication often results in a relative loss of diatoms from the community. A 4-fold increase in the relative inputs of nitrogen compared to silica along the German coast during the past few decades was accompanied by a 10-fold decrease in diatoms and a comparable increase in flagellates, which are less conducive to supporting food webs leading to commercially valuable fisheries. Decreased silica availability and the concomitant loss of diatoms may also be responsible, at least in part, for the increased frequency of occurrence and duration of blooms of harmful algae that seem to accompany coastal eutrophication. Harmful algal blooms ("red tides," "brown tides," and "green tides") can kill animals, such as sea lions, far up the food chain. The resulting loss of important top predators such as sea otters can cascade back down the food chain, altering species composition and reducing biodiversity.

Coastal systems that are both high in biodiversity and severely affected by eutrophication are seagrass beds and coral reefs. Temperate seagrass beds are usually nitrogen limited, whereas tropical seagrass beds and coral reefs are often phosphorus limited. However, even these tropical systems probably become nitrogen limited once eutrophication begins, and therefore additional inputs of nitrogen can cause immense harm. For the seagrasses, eutrophication leads to an immediate loss in diversity and can lead to a complete loss of the grasses due to shading by the increased biomass of phytoplankton in the water column, shading by mats of opportunistic species of macroalgae, or an accumulation of toxic decomposition products in the sediments. Eutrophic shallow estuaries and lagoons with large accumulations of macroalgae may experience frequent episodic oxygen depletion, further reducing biodiversity. Seagrasses provide food and shelter for a rich and diverse fauna, and reduced seagrass depth distribution or replacement by macroalgal blooms can result in marked declines in the abundance and biodiversity of the associated fauna.

Coral reefs occur in the oligotrophic shallow waters of the tropics, support extraordinary biodiversity, and are extremely sensitive to damage from eutrophication. The biodiversity of coral reefs declines dramatically with eutrophication from nitrogen inputs, especially when species-poor turf algae or macroalgae communities overgrow and replace corals and coralline algae. Nutrient enrichment disrupts the coral–zooxanthellae symbiosis, inhibiting calcification and contributing to increased coral “bleaching” (loss of zooxanthellae). Increased phytoplankton biomass and production increase turbidity and sedimentation. The decreased quantity and quality of light penetrating to the corals slow growth and reduce the maximum depth at which the corals can survive, reducing the available habitat. Algal blooms can cause hypoxia, decreasing the biodiversity of reef organisms through mortality and reduced habitat quality. Also, eutrophication can damage reefs indirectly by increasing predation on corals, facilitating the expansion of opportunistic filter feeders, and reducing the recruitment of corals.

B. Effects of Acidification by Nitrogen on Biodiversity

A variety of oxidized nitrogen compounds (NOx) are released into the atmosphere in the exhaust of vehicles, electric power plants, and other fossil fuel combustion. Oxides of nitrogen are also released from soils and burning vegetation. In the presence of sunlight, NOx catalyzes the formation of photochemical (or brown) smog and reacts with oxygen and hydrocarbons from automobile exhausts to form ozone, an air pollutant damaging to many plant species. In the atmosphere, NOx reacts to form nitric acid, one of two major components of acid rain (the other is sulfuric acid).

Approximately 70% of global ammonia (NH3) emissions are caused by humans, mostly by volatilization
from fertilized fields, animal wastes, and forest burning. In the atmosphere, \( \text{NH}_3 \) neutralizes some of the acidity in aerosols, cloud water, and precipitation. However, once deposited on the earth's surface, ammonium \( (\text{NH}_4^+) \) is taken up by plants or converted to nitrite \( (\text{NO}_2^-) \) and then nitrate \( (\text{NO}_3^-) \) by bacteria. These processes of biological uptake and nitrification release hydrogen ions, acidifying the soil and downstream aquatic systems. Other processes, such as increased biological nitrogen fixation or increased rates of decomposition, which increase the amount of ammonium in the soil, also contribute to acidification.

As nitrogen inputs to ecosystems increase and the capacity of the ecosystems to retain nitrogen is lessened, ecosystems begin to release increasing amounts of \( \text{NO}_3^- \). When the negatively charged ion leaves the ecosystem, it carries with it positively charged base cations, such as \( \text{Ca}^{2+} \), \( \text{Mg}^{2+} \), and \( \text{K}^+ \). These elements are essential plant nutrients and their removal can deplete soil fertility and cause serious nutrient imbalances which are detrimental to plants. As the other soil cations are depleted, toxic \( \text{Al}^{3+} \) is mobilized. Soils with low capacity to neutralize, or buffer, the increased acidity from nitrogen and acid deposition are especially sensitive to acidification. Increased acidification may be the major impact of increased nitrogen deposition in ecosystems which are not nitrogen limited, such as many tropical forests and most temperate lakes, or are already nitrogen saturated, such as some temperate forests.

Until recently, most of the impacts of humans on nitrogen dynamics were focused in the temperate zone, but industrial and agricultural practices are changing rapidly in the tropics, where most of the earth's biodiversity resides and for which our ecological knowledge is rudimentary. By 2020, approximately two-thirds of the global application of industrial nitrogen fertilizer and energy-related nitrogen inputs will occur in the tropics and subtropics. The majority of tropical forests are probably not nitrogen limited and nitrogen enrichment will likely have little direct effect on plant production but instead will substantially increase the output of dissolved and gaseous nitrogen from these ecosystems. The acid soils of the tropics may be particularly susceptible to further acidification, depletion of base cations, decreased availability of plant nutrients, and release of toxic \( \text{Al}^{3+} \), all possibly reducing productivity and biodiversity.

The effects of acidification on sensitive terrestrial and aquatic ecosystems are clear and severe. Increasing acidity of thousands of lakes and streams and the loss of fish and amphibian populations have been documented in Scandinavia and eastern North America. Experimental acidification of an entire Canadian lake eliminated shrimp, minnow, and crayfish species, all important food sources for the lake trout, which subsequently ceased reproduction. Increased nitrogen deposition has been implicated in the dieback of poorly buffered high-elevation forests in Europe and damage to high-elevation red spruce and fir forests in the northeastern United States. Acid deposition causes direct damage to trees by leaching base cations from leaves. Exposure to acid fog and cloud water, usually much more acidic than rain, increases sensitivity to frost damage. Acidification of forest soils causes nutrient imbalances, cation loss, aluminum toxicity, reduced growth rates, and increased mortality in trees. The diversity of lichens is greatly reduced by acidic deposition and the species composition of vascular plants, soil animals, and microbes shifts toward a less diverse community of acid-tolerant species. Acidification of grasslands and heathlands is widespread in Europe and results in decreased plant diversity when acid-tolerant species replace the species typical of less acid soils.

Terrestrial and aquatic ecosystems subjected to increased inputs of nitrogen often experience both eutrophication and acidification. The combined effects on biodiversity may be independent, additive, or synergistic depending on the characteristics of the ecosystem, but interactions between eutrophication and acidification are common. For example, nitrogen saturation increases the loss of base cations from the soil, accelerating acidification, while acidification strips base cations from plants, increasing the severity of the nutrient imbalances caused by eutrophication.

### IV. EFFECTS OF BIODIVERSITY ON NITROGEN DYNAMICS

The global biodiversity crisis has prompted increased concern for the possible effects of extinction of native species and introduction of invasive species on ecosystem function and the provision of ecosystem services essential to humans. The effects of biodiversity on nitrogen dynamics have not been studied as thoroughly as the converse effects of nitrogen on biodiversity, but there are numerous examples of significant effects of particular species or functional groups of species on ecosystem processes, including nitrogen dynamics. There are far fewer examples of the effects of species diversity (the number and relative abundance of species), independent of the species composition (specific identity of the species present), on ecosystem processes such as primary production and biogeochemistry.
At larger scales, the effects of ecosystem and landscape diversity on nitrogen dynamics vary with the specific ecosystems and their spatial relationships. Ecosystems which capture or accumulate nitrogen, such as riparian zones bordering farm fields or wetlands upstream of estuaries, can be very important in controlling the transport and downstream impact of nitrogen. Declines in biodiversity associated with the fragmentation and homogenization of landscapes. Land clearing, and development may increase the magnitude and variability of exports of nitrogen from the disturbed or converted ecosystems.

Particular species sometimes play unique or dominant roles in ecosystems and the introduction or elimination of these key species can dramatically alter the structure and function of particular ecosystems. For example, the introduction of a nitrogen-fixing tree to a Hawaiian ecosystem which contained no other nitrogen-fixing trees rapidly increased the availability of nitrogen and led to other changes in species composition of the vegetation. Burrowing animals, such as pocket gophers or prairie dogs, increase the rate of mineralization of nitrogen and increase the heterogeneity and availability of nitrogen. In many grasslands, grazers accelerate the rate of nitrogen cycling and influence the distribution and availability of nitrogen as well as the species composition of the plant community. In the boreal forest, in contrast, selective browsing by moose on deciduous trees containing more readily mineralized nitrogen slows nitrogen cycling and speeds succession toward coniferous forest. Addition or removal of whole groups of species that are performing the same functional role ("functional group") has effects similar to removing or adding a single species which is performing a unique function (a functional group with one member). For example, functional groups including nitrogen fixers, leaf-chewing insects in streams, and early spring ephemeral plants of the forest understory all have significant effects on nitrogen dynamics. To date, most of the investigations of the effects of diversity on ecosystem processes have focused on the diversity within groups of primary producers, usually herbaceous plants, but the effects of diversity at other trophic or organizational levels (consumers, predators, microorganisms, soil invertebrates, or whole ecosystems) may differ from the effects on primary producers. The influence of species or functional group diversity on ecosystem processes may differ for each functional group.

Undoubtedly, the idiosyncratic traits of individual species can influence nitrogen dynamics, but species also share many similar characteristics, compete for some of the same limited resources, and carry out some of the same ecosystem functions. If species are very similar, then redundancy is high and the removal of a single species would have little impact on ecosystem function. Alternatively, if species are very different (idiosyncratic) and there is little niche overlap (rather, niche differentiation), then the removal of a particular species would have a greater impact on ecosystem function. As the number of species increases, the probability of including an extremely effective dominant species in the more diverse community increases. This simple "sampling effect" is a result of changing species composition which occurs simultaneously with increasing species diversity.

Spatial and temporal variability in resource availability and environmental conditions are characteristic of most ecosystems; therefore, there are many ways in which species might differ in their use of resources and the environment. One species may be capable of symbiotic nitrogen fixation, whereas another can use organic nitrogen and another prefers NO$_3^-$ or NH$_4^+$. One species may be better adapted to drier environments or another may flower earlier in the growing season. Theoretically, a greater number of species, with different traits, will make more complete use of the available resources, increasing productivity beyond the maximum of any less diverse mixture of species. Complementarity of species in the use of nitrogen as a limiting resource and facilitation of one species by another as occurs when a nitrogen fixer increases the availability of nitrogen to other species can cause productivity of diverse mixtures of species to be greater than the productivity of any of the single species.

Species diversity is often correlated with other factors, such as productivity, biomass, predation, or resource availability, that complicate the assessment of the effects of species diversity separate from other factors. Only a few studies have examined the effect of species diversity, independent of species composition, on nitrogen cycling. These initial results suggest that greater species diversity among the primary producers more completely uses the available nitrogen and increases plant uptake of nitrogen. In a Minnesota grassland, the NO$_3^-$ concentration in the soil of the rooting zone (Fig. 5) and below the rooting zone decreased with increasing diversity, indicating that less NO$_3^-$ was lost via leaching and more nitrogen was retained within the ecosystem. In these same Minnesota grasslands, the nitrogen concentration in plants and the total nitrogen content of the plant biomass increased with the number of functional groups as well, and the composition and diversity of the functional groups had approximately equal influences on ecosystem processes. In California
serpentine grasslands, temporal partitioning of resources and facilitation by nitrogen fixers allowed more complete use of the greater available nitrogen, but the identity of the functional groups was more important than the number of functional groups. Hypothetically, increased diversity and resource exploitation increase nitrogen retention, thereby creating a positive feedback loop of increased fertility and productivity. With more complete capture or immobilization of a limiting resource such as nitrogen, more diverse communities may be less susceptible to invasion.

Variability of NO$_3$ in the rooting zone decreases as the number of plant species increases, but this is partly due to a simple statistical artifact—a “portfolio effect” in which the aggregate variance of several species (like the diversified portfolio) is less than the variance of individual species. More diverse communities may show less variability in ecosystem function under the varying conditions typical in the natural world. Thus, species diversity may provide “insurance” that regardless of the varying environmental conditions, ecosystem functions will be carried out by some species suited to the prevailing conditions.

In general, it appears that the number of species or functional groups may influence nitrogen dynamics, but these are probably less influential than the particular identities of the species or functional groups, especially within the context of the many other factors which strongly influence nitrogen dynamics, including climate, geology, hydrology, type and history of disturbance, etc. Species and functional group diversity may be most important at the very low diversity characteristic of intensively managed and exploited ecosystems typical of modern agriculture or very degraded systems. It appears that the upper limit of the number of plant species for species diversity, independent of species composition, to have an impact on nitrogen cycling may be quite low (1–12) in most ecosystems. Of course, there may be some ecosystems in which nitrogen cycling cannot be maintained by a few species—perhaps extremely variable and heterogeneous ecosystems such as semiarid grasslands.

V. CONSIDERATIONS FOR THE FUTURE

Unfortunately, there is no obvious reason to expect human impact on nitrogen cycling or other human impacts on biodiversity to decrease. Rather, we can expect modification of the nitrogen cycle and declining biodiversity to continue, within the context of all the other modifications which we humans are making to our environment, including climate change, air and water pollution, and overexploitation of natural resources. Many indirect effects of alterations of the nitrogen cycle on biodiversity are possible. For example, human activities have increased the release of N$_2$O, a greenhouse gas which contributes to climate change. Climate change has potentially major effects on biodiversity. In addition, N$_2$O which reaches the stratosphere participates in the destruction of the stratospheric ozone shield, allowing more ultraviolet light to penetrate to the earth’s surface, potentially decreasing biodiversity. The variety of important ecological interactions which involve nitrogen, the complexity of controls on nitrogen dynamics, and the huge scale of human alterations to the biogeochemistry of nitrogen imply that many more interactions and feedbacks between nitrogen and biodiversity remain to be discovered as the domination of the environment by humans intensifies. The interactions between nitrogen and biodiversity are multiple and complex, but it is clear that humans cause both the changes in the biogeochemistry of nitrogen that usually reduce biodiversity and the changes in biodiversity that alter the biogeochemistry of nitrogen.

See Also the Following Articles
ACID RAIN AND DEPOSITION • ATMOSPHERIC GASES • BACTERIAL BIODIVERSITY • BIOGEOCHEMICAL CYCLES • CARBON CYCLE • ESTUARINE ECOSYSTEMS • EUTROPIFICATION AND Oligotrophication • SOIL CONSERVATION
Bibliography


I. Purpose
II. Origins
III. Hierarchical Systems
IV. Principles
V. Codes of Nomenclature
VI. The Draft BioCode
VII. Special Cases
VIII. Other Systems
IX. Why Names Change
X. Naming Newly Discovered Species
XI. Determining Current Names
XII. Indices of Names

GLOSSARY

available Of a scientific name of an organism, one that must be taken into account when deciding the correct (q.v.) name to be used for it.
binomen, binomial The name of a species, consisting of the name of the genus in which it is placed followed by a word peculiar to that species.
Code In nomenclature, one of the sets of internationally agreed rules governing the scientific names of organisms.
conserved Of a scientific name, one that the appropriate international body has decided should continue to be used for an organism in cases where a strict application of the Code (q.v.) would mean it had to be replaced.
correct Of a scientific name of an organism, one that conforms to the appropriate international Code (q.v.) for the position in which it is placed.
priority Of a scientific name, its date of valid publication; the first published name is generally the one to be used, subject to the provisions of the pertinent Code (q.v.).
synonym One of two or more scientific names applied to the same organism.
taxon Taxonomic entity or group of any rank, including all subordinately ranked taxa placed within it.
type Of an organism, the specimen, culture, or other element on which a scientific name of an organism is based and that fixes its application, the name-bearing type.
valid Of a scientific name, one that conforms to the conditions of valid publication proscribed in the relevant Code (q.v.).

SYSTEMS OF NOMENCLATURE are the procedures developed to provide the scientific names of organisms. This article concentrates on the consensual systems that operate in different groups of organisms through a series of published, internationally adopted Codes. It also discusses why scientific names change and the procedures involved in the naming of newly discovered species, and provides hints for ascertaining the current name of an organism. The article is concerned only with the nomenclature of organisms, that is, bionomenclature, and not with systems developed for naming.
NOMENCLATURE, SYSTEMS OF

parts of the genome or communities of plants or animals.

I. PURPOSE

The purpose of systems of nomenclature is to provide an unambiguous mechanism that enables biologists and all others who work with organisms to communicate by a scientific name so as to avoid misunderstandings and confusion. Scientific names are latinized and given to organisms of all kinds, whether living or fossil, and are in effect a universal currency. The same species may have different common or colloquial names in a variety of current languages, but can bear only a single correct scientific name.

Nomenclature is not the same as taxonomy, although the two are often confused. Taxonomy, a component of systematics (or biosystematics), is concerned with drawing up classifications, with deciding, for example, to what genus a species belongs, whether two species are really the same, or whether one species should be divided into two. Nomenclature is the process of determining what scientific names or labels should be applied to the units that taxonomic research considers to merit independent names. Nomenclature is subservient to taxonomy; it is not science itself, but the method by which scientific results are made available for general use.

The application of naming systems is complicated because new research may show that a particular taxonomy does not reflect the phylogenetic relationships of the organisms concerned. Further, the same species may have been described more than once by different scientists, and placed in a number of genera according to different taxonomic opinions. The net result is that there are many more scientific names than known species in the scientific literature. The full extent of the problem is unclear, but in the case of nonfossil botanical groups there are about 1.7 million species names available for 406,000 accepted species. It is the challenge of nomenclatural systems to provide clear procedures of how to handle a morass of names that may have been applied over the centuries to the same organism.

II. ORIGINS

From the earliest times, humans needed to communi- cate taxonomic information: to tell each other which animals or plants are a threat, which can be eaten, which make tools or dyes, or which can help cure particular ailments. Indigenous peoples today confront the same problems, and often have sophisticated naming systems for organisms of importance to them; inter- estingly, the species concepts they employ are often not too divergent from those a taxonomist would adopt. The basic need to name organisms was recognized even in the Bible, where one of Adam’s first tasks was to name the animals.

Problems arose when different peoples needed to communicate with each other, and even the use of Latin as a language of scholars failed to solve the problem. Descriptive phrases written in Latin came to be used, but these ‘polynomials’ were cumbersome and difficult to remember. For instance, the tea plant was referred to as Evonymus affinis arbor orientalis nuclera, flore rosco by Leonard Plukenet in 1696. By the early eighteenth century the situation was becoming impossible, and it was Carl Linnaeus (1707–1778) in Sweden who developed a solution that remains the basis of our current nomenclatural systems. Linnaeus at first followed several earlier authors in grouping his species into cate- gories that had a single name, in effect a genus. How- ever, in 1737 he started to use a single second name to denote the species in the indexes in his books. This shorthand practice entered the main texts of some of the works he was associated with in 1751, and he then went on to produce accounts of all organisms known throughout the world using this approach; the species names were printed in the margins offset from the main text.

III. HIERARCHICAL SYSTEMS

Nomenclatural systems are based on a series of ranks. Each species is referred to a genus, a genus to a family, a family to an order, and so on. The principal ranks available are, in descending order: domain, kingdom, phylum (or division), order, class, family, tribe, genus, section, species, and subspecies. Additional ranks can be intercalated using the prefixes super- and sub- (e.g., superfamily above family, subfamily below family). In practice, all the possible ranks are rarely used.

Though the number of available ranks may seem large, the current system cannot adequately satisfy those who wish to recognize each branching point in a clado-
gram, and various proposals for modified systems have been made, but are not yet accepted in any of the current Codes.

Some of the ranks above genus have standardized suffixes that enable them to be recognized (Table I), but there are exceptions for some well-known names approved by the relevant international nomenclatural bodies; for example, the grass family Gramineae (not Graminaceae). There are no rank-indicating endings used above phylum in botany, none over order in bacteriology, nor any over superfamily in zoology.

The names of families are based on an included genus, such as Nostocaceae on the genus Nostoc. Order names are also based on generic names in botanical groups, but not necessarily so in zoology, where the characters of the order tend to be employed. At ranks above order, there are no strict rules, but the names usually reflect some key distinguishing feature.

The only rank below species that is formally recognized in bacteriology and zoology is that of subspecies. In zoology the subspecies name follows that of the species without any indication of rank, for example, Mus musculus domesticus, whereas under the other main Codes the rank is made clear by the insertion of “subsp.” or “ssp.,” as in Silene vulgaris ssp. maritima. The insertion of a rank term in those cases is essential to make clear which is referred to: “var.” denotes variety and “f.” denotes a form. In botany, additional terms are used in particular groups, most notably “cultivar” (“cv.”) for stable propagated cultivated plants and “special form” (“f. sp.”) for pathogenic fungi attacking different host plants that cannot be separated morphologically.

Scientific names in the rank of genus and above always start with a capital letter. Generic, specific, and formal ranks below species (see Table I) are always printed in italic script. The practice at other ranks varies and is a matter of editorial style as this is not ruled on in any of the existing Codes. While in zoology it is not

| TABLE I Hierarchical Ranks and Suffixes Indicating Them |
|-------------|-------------|-------------|-------------|
| Rank        | Bacteriology | Botany      | Virology    | Zoology     |
| Domain      | -ales       | -ales       | -ales       | -ales       |
| Kingdom     | -aceae      | -aceae      | -aceae      | -aceae      |
| Subkingdom  | -aceae      | -aceae      | -aceae      | -aceae      |
| Phylum      | -aceae      | -aceae      | -aceae      | -aceae      |
| Subphylum   | -aceae      | -aceae      | -aceae      | -aceae      |
| Class       | -aceae      | -aceae      | -aceae      | -aceae      |
| Subclass    | -aceae      | -aceae      | -aceae      | -aceae      |
| Order       | -ales       | -ales       | -ales       | -ales       |
| Suborder    | -aceae      | -aceae      | -aceae      | -aceae      |
| Superfamily | -aceae      | -aceae      | -aceae      | -aceae      |
| Family      | -aceae      | -aceae      | -aceae      | -aceae      |
| Subfamily   | -aceae      | -aceae      | -aceae      | -aceae      |
| Tribe       | -aceae      | -aceae      | -aceae      | -aceae      |
| Subtribe    | -aceae      | -aceae      | -aceae      | -aceae      |
| Genus       | -aceae      | -aceae      | -aceae      | -aceae      |
| Subgenus    | -aceae      | -aceae      | -aceae      | -aceae      |
| Section     | -aceae      | -aceae      | -aceae      | -aceae      |
| Species     | -aceae      | -aceae      | -aceae      | -aceae      |
| Subspecies  | -aceae      | -aceae      | -aceae      | -aceae      |
| Variety     | -aceae      | -aceae      | -aceae      | -aceae      |
| Form        | -aceae      | -aceae      | -aceae      | -aceae      |
usual to italicize any other ranks, in botany scientific names at all ranks are italicized in the most recent Code (see Section V) and this practice is increasingly being adopted in botanical publications.

IV. PRINCIPLES

All Codes share a common objective: the scientific name of an organism must be unique and stable in a particular taxonomy. Nomenclature can be thought of as a tool used when it is necessary to decide on the name to apply to a particular organism or other hierarchical rank; it must not constrain taxonomy. To achieve this objective, each name needs to be formally introduced by agreed upon procedures. There must also be a means of ensuring that the application of every name is fixed so that it can always be recognized. In cases where more than one name has been proposed for the same organism, the basis on which to make a choice must be unambiguous.

Names of genera and species can be formed from any source, but are treated as if they were Latin. Ideally they should be informative, giving some clue to their features, but they are often based on the names of their discoverers or the places where they were found. This is not always the case; they can even be derived from acronyms or be anagrams. Today, the etymology of names that are introduced is generally expounded when they are first introduced.

A consequence of treating species names as if Latin is that their endings must agree with the gender of the genus in which they are placed. The ending of the name of the same species may thus change if it is transferred from a genus that has a female gender to one that is male. Sometimes, the name of the person (or persons) who first introduced a scientific name is often indicated after the name itself, sometimes with the date of the publication. This device is essentially a much abbreviated reference to a particular publication, and while often interpreted as attributing responsibility for the person that burdened us with it—especially as many names proved to be unnecessarily coined. If an author's name appears without brackets, it was introduced just like that. If the name is in brackets, the name was originally introduced in another genus or rank.

The citation *Simulium albellum* Rubtsov tells us that Rubtsov published that species name in the genus *Simulium*. *Obuchovia albellula* (Rubtsov) makes clear that Rubtsov introduced the species name of this fly but did not refer it to the genus *Obuchovia*. In bacteriology and botany, the name of the person making the change in generic placement or rank follows any name placed in brackets. For example, the citation of the Fool’s Watercress, *Apium nodiflorum* (L.) Lagasca, indicates that "L." coined the name (actually as *Sium nodiflorum* L.) but that Lagasca first placed it in the genus *Apium*. Names such as *Simulium albellum* and *Sium nodiflorum*, which give the name being used in a different rank or combined in other genera, are termed "basionyms"; names using such basionyms in a different rank or position are "combinations."

The abbreviation of the names of authors, such as "L." for Linnaeus in this case, is commonplace, especially in botany where an internationally recognized list of abbreviations to be used for particular authors was issued in 1992; initials are inserted where there is more than one author with the same surname. An "ex" will sometimes appear inserted between the names of two authors; this means that the one before it first used the name but did not meet some requirements of validity or availability (see the following), whereas the second author did.

The way in which names are published is also specified in the different Codes. These relate to the publication itself (effective publication), which must be available to the scientific community, and the matter that must be associated with a name when it is introduced (valid publication), for example, a description and details of the material on which the name was based (the name-bearing type, discussed next). Only if the specified criteria are met does a name formally enter the body of scientific literature and become available for use.

The name-bearing type irrevocably fixes the application of a name. This is usually a specimen permanently preserved in a museum or other public institution, but in some cases it can also be a microscopic preparation, drawing, photograph, or microbial culture. The original material used and specified by the author is a "holotype"; a "lectotype" is material from the original author’s collection selected to act as the name-bearing type by a later researcher; a "neotype" is a later collection chosen to serve as the name-bearing type where no original material remains. Types are not necessarily representative of the range of a species or subspecies, but are simply a nomenclatural device to fix the usage of names. They are consequently a very important reference point in biological communication and merit careful preservation and handling as they will remain of relevance for centuries.

A key feature in all Codes is the priority by date accorded to the names to be used for an organism. The
name to be adopted is generally the first published in the same rank or group of ranks, depending on the Code.

Surprisingly, with the exception of bacteriology and some groups of cultivated plants, there is no obligation on a person introducing a new scientific name to register it with some international authority. Proposals to introduce such a system for botanical groups were accepted in principle in 1993 and were to be introduced for the year 2000, but the 1999 International Botanical Congress reversed that decision.

Conscious of the fact that the application of a rigid system of rules could lead to unwelcome and confusing name changes, the Codes have developed a variety of mechanisms for overcoming this. These involve cases being published for open comment and consideration by international committees who debate them and express an opinion or recommend a particular course of action. Names can be "conserved" against ones that otherwise would have to be used, or "rejected," in which event they cannot be taken up. In some instances, whole publications can be "suppressed" from usage.

An Approved List of names is published in bacteriology. Virus Taxonomy approaches this, and there is an Official Lists and Indices of Names and Works in Zoology. There is no exact equivalent in botany, although some of the appendices to that Code approach them, at least in part; however, proposals to develop protected lists of botanical Names in Current Use are currently being debated.

Names other than the one to be used that have been proposed for the same biological unit are termed "synonyms." These can be of two types: "homotypic" ("subjective" or "nomenclatural" synonyms) and "heterotypic" ("subjective" or "taxonomic"). Homotypic synonyms are based on the same name-bearing type, whereas heterotypic ones are based on different types.

V. CODES OF NOMENCLATURE

The five current Codes differ in their detailed provisions. All are complex documents approved by particular internationally mandated bodies. Each has its own history of development and special characteristics. Some of their distinctive features, differences, and procedures are summarized here.

A. Bacteriological

This Code arose as an offshoot of the botanical Code in 1939 in large part due to the need to accept living cultures as name-bearing types. In consequence, there are many similarities between the two Codes, although they have diverged in some respects. Most strikingly, and as a consequence of the large numbers of bacterial names of uncertain application, a start date of 1980 was established and linked to an Approved List of Bacterial Names; names published prior to 1980 and not on the list do not exist for nomenclatural purposes, that is, they are not treated as validly published and so are no longer available for use. At a stroke, the number of species names that had to be considered by bacteriologists was reduced from around 32,000 to 3000.

New names of bacteria now have to be published in or be registered by the International Journal of Systematic Bacteriology, and will then be added to the next edition of the Approved List; a continuously updated on-line system is available through the Internet.

The realization that many bacteria, and especially archaebacteria, can be identified on the basis of molecular sequence data but not grown in culture is challenging the requirement for living cultures to be designated as name-bearing types. In addition, many cyanobacteria have name-bearing types that are dried specimens or microscopic preparations.

While the only rank below species to be formally recognized in this Code is that of subspecies, "pathovar" ("pv.") is employed for plant pathogens essentially distinguished only by the ability to cause diseases in particular plants; for instance, Xanthomonas campestris pv. durantae (on Duranta repens) and pv. erythrinae (on Erythrina indica). Regulations regarding pathovar names and lists of those proposed are prepared by a committee of the International Society for Plant Pathology. Strain or serological type names or numbers are commonly used in bacteria of medical importance, but these have no standing under the Code.

The Code operates under the aegis of the International Committee on Systematic Bacteriology (ICSB), which is a part of the International Union of Microbiological Societies (IUMS). The Committee has a Judicial Commission that reports back to the ICSB and has wide powers to propose changes to the Code, which have to be ratified by IUMS congresses.

B. Botanical

The botanical Code covers all groups traditionally studied by botanists even if they are no longer classified in the plant kingdom (Plantae), namely, cyanobacteria (Bacteria), fungi (Fungi), slime-moulds (Protozoa), and various algal groups (Chromista or Straminopila).

Botanical nomenclature is considered as starting from the publication of Linnaeus' Species Plantarum in
1733, but the first internationally agreed Code dates from the Lois de La Nomenclature Botanique prepared by Alphonse de Candolle in 1807 and adopted by an International Botanical Congress in Paris that year. Later starting dates, all linked to particular major publications by authors other than Linnaeus are used for some groups, notably mosses (1803), certain groups of algae (1848–1900), and fossils (1820). Fungal names formerly dated from 1801 or 1821 depending on the group, but that practice was discontinued in 1981. Names of fungi accepted in the previous starting point works remain “sanctioned” for continued use even if earlier competing names exist.

In addition to the insertion of terms to denote ranks below species and differences in the practice of how describing authors are cited, several features of this Code are unique to it. New scientific names published after 1935 must, with a few special exceptions, have a description or diagnosis (i.e., a statement of how the organism differs from others) in Latin. The desirability of retaining this practice is questionable and regularly debated now that English has occupied the role held by Latin in the eighteenth century.

The botanical Code recognizes the priority of names only within the particular rank under consideration. This means that even if a plant was recognized as a subspecies long before a species name was coined, the species name is nevertheless the one to be used. In addition, this Code rules as “illegitimate” names that have been introduced unnecessarily when another should have been adopted by the author. Also “illegitimate” are names spelled in exactly the same way; these are termed “homonyms” and only the oldest is generally available for use; for example, Erica hibernica (Hook & Arnott) Syme 1866 is illegitimate because of the existence of E. hibernica Unnert 1839, which represents a different species and is based on a separate name-bearing type. That Syme’s name was based on E. mediterranea var. hibernica Hook & Arnott 1835 does not affect the situation, as that name has priority from 1835 only in the rank of variety and not of species.

Name-bearing types have had to be cited in the Code since 1958, and from 1990 the institution where they are preserved must also be cited. Living type material is not permitted, but dating from 1993 freeze-dried (lyophilized) material or specimens preserved in liquid nitrogen are acceptable as they are in a metabolically inactive state.

The botanical Code has special provisions for hybrids, fungi with pleomorphic life cycles, and fossils, all of which are considered separately in Section VII below, and also appendices of conserved and rejected names, and suppressed publications. The provisions of the Code are now debated at each six-yearly International Botanical Congress, after which a new edition is published. Any changes proposed have to be published in the journal Taxon and are balloted first by mail and then at the Congress itself, where a 60% majority is normally required to effect any change. The Congress establishes a series of permanent Committees that are charged with considering and making recommendations on proposals to reject or conserve names in different groups; those also have to be published in Taxon and ratified by a subsequent Congress.

C. Cultivated Plant

The cultivated plant Code split from the botanical one in 1932. It is primarily concerned with the naming of cultivated varieties or “cultivars,” which are propagated horticulturally by any method that retains their characteristics in a stable way. It is complementary to the botanical Code and does not compete with it. Species names and those in other formal recognized ranks of cultivated plants remain subject to the botanical Code.

Cultivar names, with a few established exceptions, are not latinized and since 1938 have been required to be in a modern language; they are placed within single quotation marks and never printed in italic type, for example, Rubus nitidoides ‘Merton Early,’ in which ‘Merton Early’ is the cultivar name and R. nitidoides is the botanical species name. The abbreviation “cv.” can be inserted before the cultivar name but is generally omitted. It is also not uncommon for the species epithet to be dropped as well in garden catalogs (e.g., Rubus ‘Merton Early’). Similar cultivars can be grouped into cultivar groups; this is achieved by placing a group name in brackets before the cultivar name and without quotation marks; for instance, Rosa (Hybrid Tea Group) ‘Richmond’; if the particular cultivar is not being cited, the brackets are omitted (Rosa Hybrid Tea Group).

In the case of orchids, groupings of cultivars based on parentage are called “grexes” (or more correctly “greges”); their application follows separate provisions in the Handbook of Orchid Nomenclature and Registration (1993).

Graft-chimeras, in which living tissues of two species are made to grow together by grafting, can also be named under this Code. These are indicated by the use of a “×” sign in a manner equivalent to the use of “x” in the names of hybrids (see Section VII,C). The form Cruciferae monogyna × Mespilus germanica indicates a graft with tissues of both of those plants. A latinized
name can be used instead provided that the + sign is retained and the resultant name is not identical to a name of a hybrid, in this example, the name + Crotac-gomespilus dardarii has been coined. Cultivar names are generally preceded by an indication of the host and (or) the symptoms caused; for instance, the "Desmodium mosaic potyvirus". For viruses, that register newly proposed names and make lists of these available.

The cultivated plant Code operates under the auspices of the International Commission for the Nomenclature of Cultivated Plants.

D. Virological

The naming of viruses was controversial for many years. Two different systems were operating in parallel: a traditional system recognizing families, genera, and species with latinized names, and a divergent one employing the categories of group, subgroup, and type (or virus) and in a modern language. The former held sway among those working with viruses in animals, and the latter among plant pathologists. In the case of plant viruses, a scientific name indicative of the structure of the virus was generally preceded by an indication of the host and (or) the symptoms caused; for instance, the "Desmodium mosaic potyvirus" and the "Desmodium yellow mottle potyvirus".

The situation started to change rapidly in 1991 when the International Committee on the Taxonomy of Viruses decided to use the traditional approach (the Committee operates within the International Union of Microbiological Societies). New guidelines were drawn up and adopted by the International Congress of Virology in 1993. Proposals to modify the rules relating to virus names may be aired in a herbarium, or in some cases an illustration. There is then a system of 45 International Registration Authorities, each dealing with particular groups of cultivated plants, that register newly proposed names and make lists of these available.

The cultivated plant Code operates under the auspices of the International Commission for the Nomenclature of Cultivated Plants.
VI. THE DRAFT BIOCODE

The need for a more unified approach to biological nomenclature has long been recognized, and this has been of particular concern to the International Union of Biological Sciences. Concerted action to address the problem emerged from a symposium held during the International Congress of Systematic and Evolutionary Biology in 1985. Pressure for harmonization between the various Codes arises from the unification of biology as a discipline and the consequent difficulties in teaching nomenclature, the mismatch between the Codes and the kingdoms of life now recognized, the needs of users for increased stability in names, the needs of systematists to spend less time on nomenclature, and the difficulties faced by developing nations. In addition, the existing Codes were finding that they had to confront similar issues, for example, in relation to electronic publication, living name-bearing types, organisms that could be treated under different Codes depending on whether they were interpreted as animals or plants, protected lists of names, and the registration of newly proposed names.

The move toward harmonization is operating along three fronts: harmonization of terminology (see Table II), bringing similar provisions into all Codes where new problems are being confronted, and developing a unified Code that would apply to all kinds of organisms. In 1995 the IUBS and the International Union of Microbiological Societies, the two bodies that between them oversee all five Codes, jointly established an International Committee on Bionomenclature to promote these initiatives. This Committee, which consists of representatives of the bodies responsible for the five organismal Codes, issued a Draft BioCode: The Prospective Interna-
The Draft BioCode, subject to approval by the appropriately mandated bodies, is planned to operate for names introduced after a date to be agreed on and for groups where lists of protected names are in existence. It was recognized early on that a single retrospective Code would be too disruptive of names in use. The existing Codes would consequently continue to apply to names introduced before that date, but be expected to adopt a common terminology and similar approaches to new problems. Among the provisions of the Draft BioCode are the adoption of coordinate status for names, a requirement for descriptions and diagnoses to be in either English or Latin, provision for electronic publication, registration of newly proposed names, unacceptability of homonyms, and the conservation or rejection of names to promote stability. The starting point for the new Code will be lists of protected names that cannot be displaced by undisturbed names; a more open and more stable system will result, which at the same time will reduce the time inputs required from systematists.

Debates on the latest version of the Draft BioCode took place in 1998–1999, and final implementation decisions by the international bodies will not be made for some years; at the 2005 International Botanical Congress in the case of botany.

VII. SPECIAL CASES

Particular provisions or problems in the Codes relate to the special cases of ambireginal organisms, fossils, hybrids, lichens, and pleomorphic fungi.

A. Ambireginal Organisms

Ambireginal organisms are ones that can be treated under different Codes depending on whether they are considered to be plants or animals, algae or bacteria, etc. The problem is particularly acute in the case of unicellular and often flagellate organisms now placed in the kingdoms Protocista or Protozoa, and especially where related species may or may not have chloroplasts and so have traditionally been treated as plants or animals, respectively. Because different rules apply when a “plant” is found to be an “animal,” name changes in both the generic and species names may result. For example, the euglenoid name Entosiphon B Stein 1878 is acceptable under the zoological Code, but when treated as a plant it has to be replaced by Entosiphonomas Larsen & Patterson 1991, as there is an earlier flowering plant genus Entosiphon Beddome 1864.

The resultant frustrations and confusions have led to demands for a separate Code for protocistian groups, but the preferred route forward is to have parallel provisions in the existing Codes. The problem would not arise under the Draft BioCode as no decisions as to whether an organism is a “plant” or an “animal” would be needed.

In the case of slime moulds (Myxomycota), because these organisms have been traditionally treated under the botanical Code as fungi, that practice is continued to prevent the disruption in names that a switch to the zoological Code would entail. The Cyanobacteria, traditionally treated under the botanical Code as blue-green algae, are pragmatically best left there rather than transferred to the bacterial Code in order to avoid unfortunate and unnecessary name changes.

B. Fossils

Special provisions for fossils are made only under the botanical Code. Pieces of trunks, stems, roots, and leaves, and seeds, pollen, and spores can all be given latinized binomials. These are considered to be “form-” or “organ-” genera and species in the absence of intact leaves. Even when such pieces are later found connected together, as in the case of the Archaeopteris leaves and Callixylon wood, the two systems are often maintained because of the time spans involved. In particular, it is not known whether a Callixylon leaf always had an Archaeopteris leaf system.

Where names of living and fossil plants compete, the living always has precedence. The names of fossils thought to be close but not necessarily identical with a modern genus are often given the suffix “-ites,” as in Ginkgites for fossil plants and Ginkgo for species now living.

C. Hybrids

Only the botanical Code has special arrangements for the naming of hybrids. The zoological Code excludes them from consideration.

Botanical hybrids can be referred to by indicating their parentage and the use of the multiplication symbol “×” to make a “hybrid formula.” Alternatively, and particularly if the hybrid is regularly encountered in nature, it can be given a separate scientific name prefixed by the same symbol. For example, the hybrid between Potentilla anglica x P. erecta can be indicated as Potentilla anglica × P. erecta or alternatively as Potentilla
Lichens are mutualistic symbioses composed of two or sometimes more different kinds of organisms. A fungal partner combines with a green alga or cyanobacterium, or both, to form characteristic shapes for the particular association. The different partners, when isolated into pure culture, fail to produce the same morphological structures. The classification of both the fungi and algae or cyanobacteria involved is based on their own characteristics and not those of the combined association; both have independent scientific names.

Before the dual nature of lichens was recognized in 1867, and for many years afterward, names were given to the composite lichen structure. The botanical Code rules that for nomenclatural purposes the names given to lichens are to be considered as belonging to the fungal partner. Pedantically, this means that lichens do not have names, and that it is impossible to “name a lichen.”

**D. Lichens**

Although generally appearing as discrete individuals, lichens are mutualistic symbioses composed of two or sometimes more different kinds of organisms. A fungal partner combines with a green alga or cyanobacterium, or both, to form characteristic shapes for the particular association. The different partners, when isolated into pure culture, fail to produce the same morphological structures. The classification of both the fungi and algae or cyanobacteria involved is based on their own characteristics and not those of the combined association; both have independent scientific names.

Before the dual nature of lichens was recognized in 1867, and for many years afterward, names were given to the composite lichen structure. The botanical Code rules that for nomenclatural purposes the names given to lichens are to be considered as belonging to the fungal partner. Pedantically, this means that lichens do not have names, and that it is impossible to “name a lichen.”

**E. Pleomorphic Fungi**

Fungi that have a life cycle with separate sporulating stages, known as pleomorphic fungi, present particular problems in naming. In many cases the different stages were named separately in different genera and were never thought to be connected until fresh research was carried out, for example, by germinating single sexual spores. In many cases the names given to the sexual (“teleomorph”) and asexual (“anamorph”) stage are both well established and mycologists have been reluctant to simply apply the earliest name regardless of the stage it represents. This resistance is understandable, as it is often a particular sporulating form that is associated with a plant disease or the production of toxins.

In basidiomycete and ascomycete fungi, with the exception of lichen-forming species, the correct name for a fungus with a pleomorphic life cycle that covers all its stages is that which produces sexual spores, that is, the teleomorph. The anamorph (or anamorphs) can also have independent names so that just those stages can be referred to; the name-bearing types of the names of anamorphs must represent only the asexual (“mi- tosporic”) sporing stage, and the sexual stage must not be mentioned in its establishing description.

For example, when Penicillium brefeldianum B. O. Dodge 1933 was introduced, the description and type covered both the teleomorph and anamorph; that name is therefore applicable to the teleomorph, even though the generic name Penicillium is based on an asexual species. In this case, the teleomorph belongs to the sexual genus Eupenicillium and is correctly named E. brefeldianum (B. O. Dodge) Stolk & Scott 1967. The anamorphic name Penicillium dodgei Pitt 1980, with only asexual structures mentioned in its description and represented on the type, was introduced for those wishing to refer just to the asexual Penicillium stage.

Now that molecular techniques enable the position of asexual fungi to be determined with confidence in overall fungal classifications, even in the absence of a sexual stage, doubt is being expressed as to the desirability of maintaining this special provision for dual nomenclature in the botanical Code.

**VIII. OTHER SYSTEMS**

This article focuses on the nomenclatural systems used in and operating under the mandate of internationally mandated committees established by the scientific community. But, in addition to infraspecific systems such as those for special forms and pathovars that are specifically excluded from these Codes, alternative nomenclatural systems have been proposed, although none has found general approval.

Concern over the adequacy of the currently accepted systems has been voiced most strongly by those involved in the determination of phylogenetic relationships using cladistic analyses. If formal scientific hierarchies were to be given to equivalent branching points, considerable numbers of new ranks would be needed and there would be an exponential expansion in the numbers of such names. A further complication is that phylogenetic analyses may show that different families or orders generally regarded as distinct have some recent ancestors in common (i.e., are monophyletic) and arguably could therefore be united. A pragmatic approach to this problem has been suggested, namely, continuing the current taxonomic practice while recognizing that all higher category names may not be equivalent, even to the extent that some may be nested within units bearing names of the same rank.

In a system proposed by K. E. Kinman in 1994 to take
the best from cladistic analyses and include them in a practical system, markers are used to indicate where orders or higher ranks include organisms belonging to different evolutionary lines (i.e., are paraphyletic) while retaining established names.

A more radical nomenclatural system is the New Biological Nomenclature, which arose in 1971 out of the frustration felt by Belgian zoologists with the current system of Codes. Names being introduced have to be agreed on by a group of specialists and are presented in Esperanto and not latinized. Generic names under this system relate more to common perceptions than those of scientists and can embrace several traditionally accepted genera. For example, the New Biological Nomenclature Esperanto generic name Delfeno embraces 17 genera of cetaceans; Delfeno diverskolora is used for the Common Dolphin (i.e., Delphinus delphis) and Delfeno ondoropa for the Bottle-nosed Dolphin (i.e., Tursiops truncatus). The system has not gained general acceptance, but does have an Association devoted to its promotion; detailed rules have been issued, the latest edition in 1991, and 522 names had been “officialized” by the end of 1990.

A novel approach to the nomenclature of fossils of all groups, a Paleontological Data Handling Code, was proposed by N. F. Hughes in 1989. Instead of first matching a newly discovered fossil to what has already been described, each specimen is recorded separately with particular reference to its stratigraphic position; only then are comparisons made with what has already been described in that particular position in the fossil record.

Various systems of coding using letters and/or numbers rather than latinized names have been proposed, the first by P. Hartig in 1871. A particularly elaborate system was that published by G. Tornier in 1898, which took the first letters of words of the higher ranks to which a species belongs, followed by a number indicating the particular species; where more than one name started with the same first letter, subscripts were employed. Numerical codes regained popularity in the 1960s in response to the need to compress data in the computing systems in use at the time; one such code extended to 36 digits. With the advances in current technology, the interest in alphanumeric systems has waned. Humans find it easier to remember words than numbers.

IX. WHY NAMES CHANGE

Changing scientific names are a constant cause of irritation and frustration to all who use them. That this occurs at all is sometimes interpreted as a criticism of the discipline, and while that may be justified in some instances, in most cases it is in the long-term interests of users.

New research can show that a certain species is actually a member of a different genus than the one in which it was placed, and it is therefore transferred. Because classifications are predictive of the properties of an organism, accurate placement will benefit users searching for species with particular attributes or wishing to determine the risks they may pose. Detailed studies may also show that one species is a complex of several that merit independent recognition, or that two species currently treated as distinct are really the same; again, these are categories of information that are relevant to users.

Name changes that arise from new taxonomic research, or “taxonomic changes,” are therefore to be welcomed, but that is not so for “nomenclatural changes.” Nomenclatural changes are ones that arise from the provisions of the Codes and not from any new scientific data. Two categories can be distinguished. First, some changes are due to a failure to apply the relevant Code, for example, the discovery of an earlier name for the same species, that a name did not meet all the requirements for establishment, or that a name was superfluous under the botanical Code when published. Second, some changes arise from rules of the Code itself that are retroactive, for example, in relation to starting point dates or new provisions relating to issues ranging from recommendations on spelling to reworkings of particular sections.

Problems leading to name changes also arise from the application of the system of name-bearing types. Species names are sometimes found to have come to be applied to a species different from that represented by its type. This category of changes is especially unfortunate as it can mean that a name that has been used in one sense must be applied in a different one.

The responsibility for deciding when enough data are at hand to justify a change in taxonomy will always lie with the taxonomist. Provided that the taxonomist is self-critical in approach and conscious of the need to be conservative in the interests of the image of the subject, most taxonomic changes will prove to be in the users’ interests. That is not the case with nomenclatural changes where all Codes, and particularly the Draft Bio-Code, are increasingly addressing this fundamental issue. The powers now given to the botanical and zoological Code Committees mean that, provided a sound case is made, rules may be set aside to protect familiar names; this can even extend to the designation of different
name-bearing types in order to maintain the current usage of names.

X. NAMING NEWLY DISCOVERED SPECIES

As only about 1.75 million of the estimated 13.7 million species on Earth have yet been given scientific names, the task of naming the balance is one of the major challenges facing biology as we enter the twenty-first century. However, it is always easier to describe an organism as new to science than to establish if it really has never been described before, albeit in a different genus or even in another family. The repeated unnecessary description of species that are not new is a major cause of the inflation of taxonomic literature. In the case of the fungi, one analysis showed that each accepted species had been given on average 6.5 names! Once introduced into the body of scientific knowledge, names can never be expunged and will have to be taken note of by all future workers on the group concerned. Before deciding to introduce a new name, careful checking is a prerequisite.

This checking proceeds in a series of steps. First, see if the prospective species matches any in the same (or allied) genera in any checklists for the geographical area from which it comes. Then proceed to any regional or world monographs of the group if they exist. If these steps yield no matches, check names in the world indices of names (see Section XI), investigating possible names in allied genera and under names no longer in use. This will entail making lists of possible names and their places of publication, looking up original descriptions, and locating and making comparisons with name-bearing types for likely candidates. If a specialist in the group concerned can be located, that can be a most valuable shortcut as the checking process can be exorbitantly time-consuming if dealing with, for example, a genus of over 1000 described species.

When novelty has been confirmed in the checking phase, the procedures detailed in the appropriate Code for introducing (and in some cases registering) new names must be followed. These include clear statements of why the species is different, a full description (which may have to be in Latin or English), illustrations (photographs as well as drawings are recommended), and a name-bearing type deposited in a secure public institution (ideally with duplicates deposited elsewhere for security). Publication should ideally be in a journal that will be readily accessible to other specialists in that group and geographical region. Much can be learned from recent publications of new species in the same group by other authors.

In some instances, notably certain bacteria, internationally developed minimum standards for descriptions are available and should be followed. In some cases codes of practice have been developed, for example, there is one for fungi by the International Commission on the Taxonomy of Fungi.

XI. DETERMINING CURRENT NAMES

When taxonomic work is undertaken, a number of different names are often found to have been given to the same species. The steps to be followed in determining which is to be used can be considered as a series of sieves, a "nomenclatural filter." Although there are differences in the detailed requirements at each step, and some will depend on the rank, the sequence in which these need to be followed is the same.

Step 1: Publication If a name has not met the requirements for effective publication, generally being printed in a book or journal available to other researchers, it need not be considered further.

Step 2: Establishment All of the following requirements must be met: species description (and in some cases illustrations), clarity concerning the rank (or coordinate group of ranks in zoology) and the intention to introduce a new scientific name, registration where appropriate, citation of the name-bearing type, and in some cases its place of preservation. If any requirement is not fulfilled, those names also need not be considered further.

Step 3: Typification The most critical step and the one that requires scientific judgment is checking the status of the type and examining the specimen, microscopic preparation, illustration, and living culture as appropriate. In some cases, no name-bearing type may have been designated and it may be necessary to select a lectotype or neotype to fix the application of the name. Names whose types are different from the taxon to be named can be excluded at this step.

Step 4: Acceptability Does the name contravene any special provisions, such as being superfluous under the botanical Code? Is the spelling identical to another name of the same rank (or coordinate group of ranks in zoology)?
Step 5: Precedence  What is the date of establishment of each name and which is the earliest in the same rank (in botany) or family, genus, or species-group (zoology)?

Step 6: Accepted Name  The name to be used under the Code will be the one with precedence, that is, the earliest acceptable name to be established. This name may have to be used in a new combination. Under the botanical Code, a new combination cannot be made if the resultant binomial would be spelled exactly like an already existing name (i.e., a “homonym”). The situation in zoology differs in that a new homonym can be made if the basionym takes precedence by date; the already existing homonym becomes a “secondary homonym” and can no longer be used. Names applied to the same suprageneric rank or organisms other than the accepted one are “synonyms.” Synonyms are of two main types: “homotypic” when based on the same basionym or name-bearing type, and “heterotypic” when based on different basionyms or name-bearing types. Homotypic synonyms will always appear together, as their applications are irrevocably linked by a common synonym” and can no longer be used. Names applied to the same suprageneric rank or organisms other than the accepted one are “synonyms.” Synonyms are of two main types: “homotypic” when based on the same basionym or name-bearing type, and “heterotypic” when based on different basionyms or name-bearing types. Homotypic synonyms will always appear together, as their applications are irrevocably linked by a common type, whereas heterotypic synonyms may be treated differently depending on the disparate views of taxonomists on the particular types. If the steps potentially result in the disruption of a well-known name, the procedures for the conservation and rejection of names are available as a safeguard (see earlier discussion).

This daunting series of steps, once performed, leads to an accepted name. The advantage of the international agreement on Codes is that whoever follows the steps faithfully will come to the same answer regarding the name to be used. That name will remain stable unless new nomenclatural or taxonomic information comes to light (see earlier).

Faced with such complexity, users of names wishing to check the currently accepted name for an organism need shortcuts. Recent world monographs and checklists are the first port of call, and then any regional or national treatments. If these are not available, and there is no taxonomist specializing in the group that can be consulted, it will be necessary to check the main indices of names and supporting bibliographic databases or abstract publications (e.g., Biological Abstracts, Kew Record, or Bibliography of Systematic Mycology). Online checking will improve markedly when the SPECIES-2000 system is fully operational (see the next section). Checking should be viewed as a necessary evil. If research is published using an unfamiliar obsolete name, it may never be retrieved from the hundreds of thousands of scientific articles that appear each year.

XII. INDICES OF NAMES

There is a series of major reference works of the scientific names of organisms that focus on different major groups and that vary with respect to the information captured. There are too many to list here, but attention should be drawn to those that are currently active: Index Kewensis, Index of Fungi, Zoological Record, and the Approved Lists of Bacterial Names. These are increasingly available as on-line systems through the internet or as CD-ROMs. Unfortunately, there is still no one-stop-shop for information on names, though there is now an international initiative, SPECIES-2000, that is linking key databases around the world that have not just lists of names but ones that are taxonomically up-to-date for particular families, orders, and so on. International funding is currently being sought to complete the linking of existing data sets, and also to establish additional databases where no taxonomically vetted ones exist. For zoology, the BIOSIS Taxonomic Reference File is particularly valuable in the interim.

See Also the Following Articles

SPECIES, CONCEPT(S) OF • SUBSPECIES, SEMISPECIES • SYSTEMATICS (OVERVIEW) • TAXONOMY, METHODS OF

Bibliography


NOMENCLATURE, SYSTEMS OF


NORTH AMERICA, PATTERNS OF BIODIVERSITY IN

Martin J. Lechowicz
McGill University

I. Continental Diversity
II. Terrestrial Bioregions of North America
III. Diversity in Major Groups of Organisms
IV. Changes in North American Biodiversity
V. Conclusion

GLOSSARY

Beringia A region at the northwestern corner of North America separated from Asia only by the shallow waters of the Bering Strait. When ocean levels drop during glaciation, Asia and North America are connected by land in this region.

bioregion A landscape subunit at the scale of the continent that is set apart by its coherent geological and biotic history.

continental shelf The continental margin, which is submerged to depths of approximately 180 m at current sea levels and is bounded by the abrupt drop to the abyssal depths of the oceans.

coregion A landscape unit within a bioregion in which a distinct assemblage of organisms interact ecologically, usually within a spatial context defined by drainage systems, mountain ranges, or similar natural boundaries.

Holarctic A biotic realm in the temperate, boreal, and arctic regions of the Northern Hemisphere with strong affinities among the regional floras and faunas.

Neotropics A biotic realm in tropical South and Central America with strong affinities among the regional floras and faunas but different from the tropical biota of Africa and Southeast Asia (the Paleotropics).

physiography The landforms that give shape and character to the continental landscape, for example, the configuration of mountains and drainage basins.

THE DIVERSITY OF THE FLORA AND FAUNA OF NORTH AMERICA has been the focus of study by naturalists and scientists for the past few hundred years. Studies of species diversity have proceeded even as parts of North America have undergone dramatic changes driven by activities associated with settlement, mining, forestry, agriculture, and industrial development. Large areas of the midcontinent have less than 5% of their primeval landscape intact. This article attempts to identify the patterns of species diversity that existed in North America 500–1000 years ago, before the arrival of the Europeans. The focus is on the diversity and distribution of native or indigenous species, as opposed to species that people have introduced from other places. Current or historical records from relatively undisturbed habitats provide the basis for estimates of species diversity from region to region. This article is restricted to the diversity of the terrestrial and freshwater biota of mainland North America, including offshore islands but excluding the rich marine biota associated with the continental shelf and coastal estuaries. Although the continental shelf is quite properly considered part of...
North America, its biota is determined not only by continental but also oceanic influences of considerable complexity. The evaluation and summary of North American marine diversity is a separate and important task, but it is beyond the scope of this article.

I. CONTINENTAL DIVERSITY

Mainland North America, with an area of 24,298,000 km², accounts for 16% of the land surface on Earth and stretches over 65° of latitude. The ancient core of North America is well over 3 billion years old, but parts of the continent are much younger. The continent is bounded on the west by a series of mountain ranges along the coast of the Pacific Ocean. A more extensive coastal plain and less rugged mountains characterize the Atlantic coast of North America. The Arctic Ocean lies at the northern limit of the continent, which includes the extensive Arctic Archipelago. Rugged mountains mark the commonly accepted limits of the continent in southern Mexico, but in fact distinguishing the boundary in this mountainous and geologically complex region where North America grades into Central America is arbitrary. The continental interior is an extensive plain with relatively little topographic relief, although highlands, escarpments, and erosional features occur regionally. Major drainage systems flow both north and south from the interior and also through the Great Lakes basin to the east. The Mississippi River system flowing south into the Gulf of Mexico is among the largest drainage basins in the world. The Caribbean Sea lies southeast of the continent. North America has a land connection to South America through the Isthmus of Panama, a link that formed only approximately 3 million years ago (Ma). In the past few million years, as ocean levels have fluctuated during glacial cycles, there has also been an intermittent land connection to Asia through Beringia. People arrived in North America by this route, and perhaps others, only very recently—sometime in the past 15,000–35,000 years. With the exception of Antarctica, the Americas were the last continents to be colonized by people. Although recently much influenced by people, the biota of North America evolved for hundreds of millions of years without a human presence. The current diversity of the continental biota is founded on this long period of geological and evolutionary history.

Two biological processes interact over millions of years in the evolutionary history of the continental biota: speciation and extinction. Populations, isolated from one another by changing landforms or climates, diverge to form new species. Species go extinct in the face of changing environments or by chance events in small populations. The number of species in a continental biota at any point in time reflects in part an ongoing balance between speciation and extinction on the continent as a whole. The evolutionary history of a continent is intimately interwoven with its geological history, especially through the influence of continental drift. In the 400 million years since life appeared on land, interchange among different continental biota has been possible whenever continents have drifted into contact with one another. Species that have evolved separately on other continents have had opportunities to disperse and colonize North America. This occasional mixing of continental biota together with in situ evolution in North America has created the total pool of species on the continent today. Although this pool of species defines the current magnitude of diversity in North America, the regional patterns of diversity are set by the different distributions of individual species within the continent. Some regions have more species, and others have fewer. The range of a species within a continent reflects both environmental requirements and the ability to disperse to suitable environments. Species differ in their rates of dispersal and establishment in the face of changing environments and in their ability to cross barriers such as mountains, bodies of water, or inhospitable habitat. The physiography of the continent is thus an important influence on the current patterns of species diversity within North America. Species distributions also depend on climatic patterns within the continent, which arise from both the local influence of physiographic features and the global patterns of atmospheric circulation from season to season. In the end, the complement of species found in a region depends on the ability of species to disperse to the region, establish there, and survive and reproduce there. Ecological processes acting on relatively short timescales (years to millennia) are the proximate controls on species distribution, but the availability of the species and the distribution of suitable habitats on the continent ultimately have arisen in geological and evolutionary processes acting over much longer timescales (millions of years). It is this interplay of processes on ecological and evolutionary timescales that affects the natural patterns of diversity in North America.

II. TERRESTRIAL BIOREGIONS OF NORTH AMERICA

Mainland North America and its offshore islands can be divided into six bioregions: (i) the Canadian Shield...
NORTH AMERICA, PATTERNS OF BIODIVERSITY IN

Figure 1. Schematic diagram of the six terrestrial bioregions of North America.

in the north, (ii) eastern North America, (iii) western North America, (iv) southwestern North America, (v) a northern extension of the Central American biota in southern Mexico, and (vi) outliers of the Caribbean biota on the southern tip of the Florida peninsula. Figure 1 illustrates and depicts in a general way the boundaries of these six terrestrial bioregions in North America. There is, of course, substantial physiographic and biotic diversity within each bioregion, and the boundaries between adjacent bioregions are not precisely identifiable. Fairly distinct ecoregions can be recognized at smaller spatial scales within each bioregion, and these are important for understanding and conserving biodiversity (Abell et al., 1999; Ricketts et al., 1999). Larger drainage systems, such as that of the Mississippi River and its tributaries or the Great Lakes basin, subsume many terrestrial ecoregions. The geography of the terrestrial versus freshwater biota of North America differs at the scale of ecoregions and to some degree at the scale of the continental bioregions. Similarly, the recognition of bioregion boundaries at the current continental margins is arbitrary. Continental watersheds are directly linked ecologically to the marine diversity of the continental shelf, and this shelf typically shares a geological history with the adjacent continent. Despite these uncertainties in the boundaries and internal structure of the six bioregions, each does have a reasonable coherence in its geological history, current physiography, and current biota. These bioregions therefore provide a useful working framework in which to summarize the general patterns of terrestrial and freshwater species diversity within North America.

The first three of these bioregions are part of the Holarctic, a zonal assemblage of flora and fauna that has coherence across the whole of the middle and higher latitudes of the Northern Hemisphere. Species may differ among North America, Europe, and Asia, but many families and genera of plants and animals are represented throughout the Holarctic realm, including the Canadian Shield, have especially strong biotic affinities. High-latitude sites in North America, Europe, and Asia are
This bioregion takes its name from the Canadian Shield, which is the ancient core, or craton, of the North American continent. The bioregion and the craton are not geographically identical. The craton extends as basement rock well under the central and eastern parts of the continent, but this does not affect the regional development of the modern biota. The southern boundary of this bioregion is marked by the shift in surface geology from the ancient, crystalline bedrocks of the craton to younger, sedimentary bedrocks. The bioregion actually extends beyond the craton in northern Alaska and the Yukon, even though the bedrock in these western regions is not nearly as ancient. Two major biomes occur with the Canadian Shield bioregion: the boreal forest and the arctic tundra.

Tundra occupies the coasts and islands of the Arctic Ocean and higher elevations in interior Alaska and the Yukon. These barren lands with a short, cold growing season are relatively poor in plant and animal species. From one ecoregion to another within the Canadian Shield bioregion, there are only 100–750 plant species in tundra. Some tundra ecoregions have as many as 50 mammal and 160 bird species, and from 0 to 75 fish species in tundra. Tundra diversity decreases seasonally as birds migrate to more southern bioregions for the winter. The most diverse tundra biota occurs at high elevations in the Wrangell–St. Elias Range in southern Alaska and the adjacent Yukon. The least diverse tundra occurs along the coasts of the heavily glaciated Axel Heiberg and Ellesmere Islands, both of which extend north of 80°N latitude. There are only approximately 110 species of vascular plants in this northernmost part of North America.

The boreal forest, which occupies the more southern parts of the Canadian Shield bioregion, generally has greater biodiversity than the tundra. Different boreal ecoregions have between 250 and 1250 vascular plant species that include 5–30 tree species. There are never more than 10 coniferous tree species in any ecoregion, even though conifers dominate much of the boreal landscape. Ecoregions in the boreal forest usually have approximately 20–30 mammal species, 105–185 bird species, 35–95 butterfly species, and 20–60 fish species. Amphibians are less diverse, usually 1–5 species in an ecoregion and occasionally more than 10; there may be 1–3 reptile species, but often there are none in a given ecoregion. The boreal forests of central and eastern Canada are the most diverse overall, and those of Newfoundland are the least diverse.

B. Eastern North America

The Appalachian Orogeny essentially created the character of this bioregion. These mountains on the eastern margin of North America arose approximately 300 Ma in the collision of all the ancient continents to form the supercontinent Pangea. The Appalachians are in fact a part of the African continent left behind as the mid-Atlantic Rift developed beginning approximately 240 Ma, gradually breaking up Pangea, creating the Atlantic Ocean and moving Europe, Africa, and the Americas to their current geographic positions. As the young and rugged Appalachians eroded, huge deposits of sediment accumulated in the continental interior forming the bedrock of much of the interior plains that comprise the central part of this bioregion. Their sedimentary deposits also account for an extensive continental shelf, rich in marine biota today and also an important corridor for plant and animal migrations during and after glacial
peaks when ocean levels are low. From approximately 170 to 70 Ma much of the more western interior of this bioregion lay beneath shallow continental seas in which marine biota built up extensive reef systems and sediments. The Canadian Shield forms the basement rock of essentially the entire bioregion, but it is these much younger mountains and sedimentary features that influence the current patterns of biodiversity.

The eastern parts of this bioregion are generally forested with some combination of conifer and broadleaf trees. The western parts are more often grasslands, except for gallery forests along watercourses. The transition from forest to grasslands, which is more a mosaic dependent on local topography and drainage than a smooth gradient, generally occurs to one side or the other of the Mississippi River. Savannahs, grasslands with a scattering of trees, are common in the transitional ecotones. Many of these transitional ecotones, such as the Edwards Plateau in Texas, have exceptionally high biodiversity. There are approximately 6,500–33,500 species of plants in the different ecoregions of eastern North America, 3,050–7,500 mammals, 160–270 birds, 3–65 amphibians, 3–8 reptiles, 75–230 butterflies, 40–230 fish species, 2–200 land snails, 5–125 mussel species, and 1–65 crayfish. The mixed forests of the southeastern United States comprise the ecoregion most rich in plant species within eastern North America, and the Pine Barrens of the Atlantic Coastal Plain are the least rich. The western short grasslands are richest in mammal species, and the Florida Pine Scrub is the least rich. Very high numbers of both amphibian and land snail species occur in the forested Blue Hills are least rich. Very high numbers of both amphibian and land snail species occur in the forested Blue Ridge and southern Appalachian Mountains. The Tennesseee and Cumberland River systems, which drain the unglaciated Appalachian–Blue Ridge region, are the most rich in fish, mussel, and crayfish species of any region in the bioregion; the glaciated, northern parts of the bioregion are most poor in freshwater biota.

C. Western North America

Western North America is younger and more rugged than the northern and eastern parts of the continent, with some coastal mountains having been created only in the past 1 million years. All of western North America was created as the Farallon Plate, which has been squeezed between the Pacific and North American Plates, gradually subducted under the North American Plate. The subduction of the Farallon Plate caused a series of major episodes of mountain building in western North America: the Sonoran Orogeny approximately 245 Ma, the Nevadan Orogeny approximately 150 Ma, the Sevier Orogeny approximately 120–90 Ma, the Laramide Orogeny approximately 70–60 Ma, and the Basin and Range Orogeny approximately 15 Ma. The continued subduction of the Juan de Fuca Plate, a remnant of the now broken up Farallon Plate, accounts for the current geologic activity in the Pacific Northwest. All this mountain building associated with the expansion of the North American continent has created and defined the essential character of this bioregion. The landscape is physiographically diverse, with a high degree of regional isolation leading to many distinct ecoregions and a biota rich in local endemics. The mountains that dominate the landscape in the bioregion are predominantly forested, but various grasslands, shrublands, and deserts occur in basins and on plateaus amid the mountains. Between ecoregions, there is considerable variation in species diversity. There are approximately 450–2,350 species of plants in the different ecoregions of western North America, 30–110 mammals, 140–240 birds, 1–20 amphibians, 0–60 reptiles, 3–225 butterflies, 0–30 land snails, 3–60 fishes, 1–5 mussels, and 0–5 crayfish. Plant species diversity is lowest on the isolated Queen Charlotte Islands and high in the Sierra Nevada, in the Great Basin, and on the Colorado Plateau. The Colorado Plateau and the Sierra Nevadas are notably rich in bird species. The Colorado Plateau and the Colorado Rocky Mountains are notable rich in butterfly species, whereas the Queen Charlotte Islands are very poor in butterfly species. The forested coastal mountains have the most amphibians in the bioregion, and the Colorado Plateau and the Great Basin have the most reptile species. The Siskiyou Mountains, the Sierra Nevada, and the north Central Rocky Mountains are all notably rich in species of land snails. Although low in terrestrial biodiversity within the bioregion, the watersheds of the mountains along the northern Pacific Coast are the most rich in aquatic species, the isolated and arid watershed of south-central Oregon is the least rich. The Colorado Plateau shrublands are the most species-rich ecoregion overall, whereas the Queen Charlotte Islands and the adjacent northern Pacific Coast are the most species poor.

D. Southwestern North America

Northern and central Mexico and immediately adjacent parts of the American Southwest form a distinct bioregion.
region within North America, which like western North America has rugged and relatively young terrain. The Sierra Madre Occidental and Sierra Madre Oriental running down the western and eastern coasts converge on a belt of high mountains of volcanic origin in southern Mexico, which form a natural boundary at the southern limits of this bioregion. These mountainous regions surround an interior plateau that grades north into increasingly arid deserts. The mountain ranges stand in contrast to the low relief of the coastal plains along the Pacific Ocean and the Gulf of Mexico and to the Baja California Peninsula. Ecoregions in southwestern North America are especially rich in bird, mammal, reptile, and butterfly species. Among the ecoregions, bird species number 90–280, mammals 30–110, reptiles 10–105, fishes 10–50, and there are at least 130–260 butterflies. Numbers of plant species are also high—500–4000 species among the ecoregions—but there are as many as 8000 species along the Sierra Madre Oriental. Not surprisingly given the aridity of this bioregion in general, the diversity of amphibians is low—only 5–20 species in most ecoregions. There is insufficient data to designate an ecoregion that is the most rich in biodiversity overall, but the more diverse ecoregions are almost certainly in Mexico as opposed to the southwestern United States.

E. Southernmost North America

The tropical highlands of the Sierra Madre del Sur and the Sierra Madre de Chiapas and the tropical lowlands on the Pacific and Gulf Coasts of Mexico comprise this bioregion, which is located at the southern limits of North America. This part of North America has a complex and poorly understood geological history involving the convergence of five plates and various crustal fragments of uncertain continental affinity. It appears certain, however, that the rugged mountains of Central America and the land connection between North and South America have developed only in the past 2–5 Ma. For the preceding 70 Ma or more, North and South America had been separated by an ocean channel deep and wide enough to severely limit biotic exchange between the two continents. Once Central America arose as a bridge between the two continents, its biota was enriched by migrations in both directions, but especially by Neotropical flora and fauna from South America that were well-suited to its tropical climate. Most ecoregions in this part of southern Mexico have on the order of 8000 plant species, although this decreases to only approximately 2000 on the Yucatan Peninsula where water quickly drains deep underground in the karst landscape. Some ecoregions have as many as 125 amphibian species, but most have only 10–30. Many ecoregions have as many as 120–160 reptile species, 100–300 bird species, and 80–160 mammal species. The overall diversity of this bioregion is high and increases toward the south as the influence of the South American biota strengthens. Defining the southern limit of North America at the Isthmus of Panama would greatly increase the biodiversity of the continent as a whole.

F. Caribbean

The Caribbean Plate has a poorly understood geological history. Most of the plate is submerged in a shallow sea dotted with island arcs that are closest to North America at the Florida Peninsula. The southern tip of Florida, the only part of North America that is located in the Caribbean bioregion, has a biota that is a mix of continental and Caribbean elements. There are approximately 1000 species of plants, 25 species of mammals, 175–200 species of birds, 13–20 species of amphibians, 45–50 species of reptiles, 130–155 species of butterflies, and 60–65 species of land snails. Despite the addition of Caribbean species, the biota is not exceptionally rich compared to adjacent parts of the continent.

III. DIVERSITY IN MAJOR GROUPS OF ORGANISMS

An alternative perspective on the general levels and patterns of biodiversity in North America is to focus on the distribution of species in particular taxonomic or functional groups of organisms. Although our knowledge of the biodiversity of North America is better than that for many other parts of the world, it is far from complete and we can only adopt this perspective for the better studied organisms in North America. A few groups of vertebrate animals are fairly well studied, but the diversity of even these is poorly known in parts of Mexico. Some groups of insects and vascular plants are fairly well studied throughout the continent, but the sheer numbers of species in these groups undoubtedly indicate that many have yet to be discovered and properly mapped. Well-founded summaries of diversity in microbial groups, many invertebrate animals, fungi, and nonvascular plants currently do not exist for North America as a whole. Despite our incomplete knowledge of North American diversity in individual groups of organisms, the numbers of species and distributional patterns in the better known groups yield insights and provide guidance for conservation efforts.
A. Plants

There are approximately 1320 species of mosses in North America north of Mexico and approximately 1200 species known from Mexico. The liverworts and hornworts in North America north of Mexico number approximately 650 species, with about 800 species known from Mexico. All these numbers are expected to increase as the continental bryoflora is better explored. There is no comprehensive North American flora for these plant groups.

The pteridophytes (ferns, horsetails, club mosses, spike mosses, and quillworts) are among the most primitive vascular plants represented in the North American flora. There are approximately 440 species of pteridophytes in North America north of Mexico, more than 75% of which are ferns. Mexico has on the order of 1000 pteridophytes species, most of which are ferns of moist habitats and Neotropical affinity. For example, there are only approximately 129 pteridophyte species in the Chihuahuan Desert of northern Mexico compared to almost 700 species in the southern state of Oaxaca. The moist, tropical forests of southern Mexico harbor the greatest diversity of pteridophytes in North America.

The gymnosperms, an ancient group of plants that includes the pines, spruces, and other conifers, are commercially important and especially well-known in North America. There are approximately 90 species of conifer north of Mexico; numbers of conifers in Mexico are higher but uncertain. The well-studied pines of Mexico offer one specific point of comparison to the rest of North America. There are approximately 33 pine species in North America north of Mexico, and only about one-third of these have ranges extending into Mexico. There are another 60 species of pine in Mexico! On the other hand, there are 11 spruce species in North America, but only 2 occur in Mexico. Frequently isolated, arid mountain habitats of Mexico clearly have provided a center for the evolutionary diversification of pines in North America. North of Mexico the greatest concentration of conifer species is found in the Klamath–Siskiyou ecoregion on the mountainous coast at the California–Oregon border; approximately 30 species of conifer are found there. Conifer diversity is generally greater in the western mountains than in eastern or northern North America.

Trees in general, angiosperm as well as gymnosperm species, are well studied in North America. There are approximately 650 native species of trees in North America north of Mexico, approximately 89% of which are angiosperms; the arboreal flora of Mexico is not known with certainty but there are at least as many, and probably substantially more, tree species as there are in the rest of the continent. The oaks, which are well studied in North America including Mexico, provide a useful point of comparison. As many as 150 species of oaks are known from Mexico, approximately 30 of which are also found farther north. There are only about 60 additional oak species in the rest of North America. The Mexican oaks predominate in arid habitats, with the mountains of central and eastern Mexico being a center of diversity. The center of tree diversity north of Mexico is in the southeastern United States, where as many as 190 tree species can be found in a given ecoregion. With the exception of Mexico, the diversity of hardwood species declines steadily to the west and north of this concentration of species in the southeastern United States.

The angiosperms, both herbaceous and woody, are the most recently evolved and most species rich of the vascular plant groups in North America and indeed throughout the world. The angiosperms comprise the largest part by far of any regional flora. They are also the least completely identified and mapped in terms of their North American patterns of biodiversity. There is no comprehensive floristic inventory for North America, only a series of regional descriptions. There may be as many as 17,000–20,000 native plant species in North America north of Mexico. At midlatitudes, the highest diversity of angiosperm species in an ecoregion is in the forests of the southeastern United States, where there are approximately 3000 native angiosperm species; diversity decreases to the west and north on the continent but increases dramatically in Mexico. The state of California, which includes many diverse ecoregions, has more than 5000 native plant species. The whole of Canada has about only 3125 native angiosperm species, and the Canadian Arctic Archipelago has only 330. Mexico, on the other hand, has approximately 19,000 angiosperm species; the state of Oaxaca alone has about 8000. Mexico has approximately 5000 of the 5500 species in the Asteraceae (sunflower family); there are only 400 native species of Asteraceae in Canada. The Fabaceae (bean family) has approximately 1700 species in Mexico, with more than half occurring only in Mexico, compared to only about 130 in Canada. The 900 species of Cactaceae (cacti) in Mexico represent more than half the species in the world. The greatest angiosperm diversity in North America is clearly found at the southern end of the continent.
B. Butterflies

There is no reliable estimate of the total number of butterfly species in North America. Comprehensive surveys of western and southwestern North America report only approximately 930 species of butterflies, although this is believed to be from the part of the continent north of Mexico where butterfly diversity is greatest. Surveys of remnant prairie in the midcontinent yielded only approximately 90 species of butterflies, but the prairie ecosystems have been much disrupted in the past 200 years. There are estimated to be approximately 2200 butterfly species in Mexico, and it is certain that butterfly diversity within North America is greatest in Mexico. The swallowtails (Papilionidae), which are among the more conspicuous and better known butterflies, provide evidence of the relative richness of the Mexican butterfly fauna. There are almost 60 swallowtail species in Mexico, about twice as many as in the more northern parts of the continent.

C. Land Mammals

There are approximately 4500 mammal species in the world. Canada has approximately 140 species of mammals, whereas the United States has about 350 and Mexico about 440–450. There is no good composite estimate for the continent as a whole. North of Mexico, the Colorado Plateau harbors the greatest diversity of mammal species but the Chihuahuan Desert, which extends into Mexico, is equally rich in mammal species. The species richness of mammals in North America decreases steadily to the north, and at midlatitudes it is relatively low in the east compared to the west.

D. Birds

There are approximately 9000 bird species in the world. Canada has approximately 425 species of birds, the United States approximately 650 species, and Mexico approximately 440–450. There is no good composite estimate for the continent as a whole. The geographic pattern of species richness for birds in North America is generally similar to that of mammals: richest in Mexico and the adjacent southwestern United States and steadily decreasing to the north.

E. Amphibians

The amphibians in North America primarily include frogs and salamanders. There are approximately 4000 amphibian species in the world. Canada has approximately 40 species of amphibians, none of which do not also occur farther south on the continent. Mexico, in contrast, has approximately 285 amphibian species, approximately 170 of which do not occur elsewhere in North America. The United States has approximately 230 amphibian species. There is no good estimate of the number of amphibian species for the continent as a whole. With 55–70 amphibian species, the Appalachian Mountains comprise the region north of Mexico that is most rich in amphibian species, but this is only one-fourth of the number of species in the tropical highlands of central and southern Mexico.

F. Reptiles

The reptiles in North America include turtles, lizards, snakes, and a negligible number of crocodilians. There are about 6000 reptile species in the world. Canada has only about 40 species of reptiles, none of which do not also occur farther south on the continent. The United States has approximately 280 species of reptiles. Mexico has about 690–720 reptile species, approximately 370 of which do not occur elsewhere in North America. The Chihuahuan Desert is the part of the continent most rich in reptiles.

G. Freshwater Fishes

North America has approximately 1200 species of freshwater fish. There is, however, substantial variation in the diversity of fishes in glaciated and unglaciated parts of the continent. The glaciated regions are notably poor in fish species. Canada, despite its very large land area rich in lakes and river systems, has only approximately 180 native fish species. The extensive Hudson Bay Basin, which was entirely glaciated 18,000 years ago, has only approximately 100 fish species and most of these occur in the southern headwaters of the drainage system. The Canadian Arctic Archipelago has fewer than 10 fish species. In contrast, the largely unglaciated Mississippi River Basin has approximately 375 fish species. The southeastern United States has about 485 freshwater fish species, with the greatest diversity in the Appalachian and adjacent Interior Plateau. Despite the aridity of much of Mexico, about 380–500 freshwater fish species occur there, with notably high numbers of species in the Panuco and Papaloapan River systems along the southwestern coast of the Gulf of Mexico. The arid regions of southwestern North America had wetter climates and more extensive wetland systems as recently as 7500 years ago, but the fish fauna now is restricted
to remnant bodies of permanent water. Although the habitats available for fishes in southwestern North America are currently limited, there are many remnant species with a high degree of endemism.

IV. CHANGES IN NORTH AMERICAN BIODIVERSITY

There is a long history of patterns of biodiversity in North America. The native species that occupy North America today belong to groups of organisms that have developed over the long history of life on this continent. Some of the earliest life on land is represented by fossil plants more than 400 million years old that are found along the Acadian coast of North America. The continent has shared in the evolutionary diversification of life on land. Ancestors of conifers in the family Pinaceae grew in what is now North America 135 Ma. Angiosperm fossils from North America date to approximately 120 Ma, near the origins of this now dominant group of plants. Dinosaurs that dominated the animal life of North America beginning approximately 200 Ma are ancestors to the modern birds; the mammals rose to dominance beginning 65 Ma with the mass extinction of dinosaurs. Primitive insects are among the earliest fossils of terrestrial life in North America, but the rapid increases in species numbers of groups such as the butterflies and beetles that have high diversity in North America today also began only 65 Ma. Patterns of continental drift and climate change during the 400 million years that life has been on land have facilitated the exchange of biota from different continents. The strong affinity in the flora and fauna of the Holarctic regions of North America, Europe, and Asia has its origins in the period 170–80 Ma, when these continents were still in close proximity to one another. Similarly, the richness of the Central American flora and fauna in southernmost North America stems in part from biotic exchanges initiated in the past 2 or 3 Ma when the Americas were reconnected by land after 60–70 million years of relative isolation from one another. Some contemporary patterns of diversity in North America that we might seek to attribute to current environmental conditions in fact originated in the ancient history of life on the continent.

The influence of past events on contemporary patterns of biodiversity has been most marked during the past 1 million years, during which climatic changes associated with glacial cycles have repeatedly disrupted the North American biota. Species ranges in North America have shifted dramatically from glacial to interglacial, and current distributions of species reflect patterns of dispersion from the most recent glacial refugia. The relative poverty of species in the more northern parts of North America in part stems from the failure of some species to disperse back into these glaciated regions in the 5000–14,000 years since the ice melted. Major climatic changes from glacial to interglacial periods also have influenced the biodiversity of unglaciated regions of the continent. For example, during the most recent glacial maximum parts of western and southwestern North America had a wetter climate. Large pluvial lakes and extensive river systems existed that are now mostly reduced to playas and dry channels. Today, desert pupfish occur only in isolated springs and pools, whereas once they were widespread in extensive lake and river systems throughout the Southwest. Once widespread palms now exist only as remnant populations in isolated canyons. In another cycle of changing climate, such remnant populations may again become common, but now we value them as rare species that enhance the biodiversity of North America. The rich and sometimes peculiar patterns of biodiversity in North America are built on a dynamic history of changing environments and changing species distributions.

In terms of the most recent history of the North American biota, the arrival of humans coincident with the end of the most recent glacial cycle is especially noteworthy. As the interglacial was beginning, hundreds of the larger animals in North America went extinct. It is difficult to account for these extinctions by changing climate alone, and it appears that human hunting pressures were a significant factor. In a brief period, the mammoths, mastodons, camels, horses, saber-tooth tigers, giant ground sloths, and many other large mammals that undoubtedly were major elements in the functioning of North American ecosystems simply disappeared from the continent. The patterns of biodiversity at the time of European settlement not many thousands of years later must reflect this fairly recent mass extinction as well as the continued influence of the few million aboriginal people who already lived in North America at that time. Whatever patterns of continental diversity we identify, we should not assume that they occurred independent of human influence.

Human influence on patterns of diversity in North America has increased greatly in the past 500 years. The European colonists consciously or inadvertently brought many new plant and animal species to this continent. Indeed, people continue to introduce species from outside North America, some of which establish
and spread to natural ecosystems well beyond the confines of urban or agricultural lands. Approximately 25% of the current flora of Canada consists of alien species; the United States has approximately 3700 alien plant species, on the order of 20% of the flora. Approximately 6% of the fishes of Canada are not native species; the United States has about 75 fish species native outside of North America and approximately 200 that have been introduced outside their native North American ranges. The fishes currently dominating the ecosystems of the Great Lakes are entirely the outcome of introduction and management of nonnative species by people. Inadvertently introduced zebra mussels are now altering ecosystem function in the Great Lakes basin and ousting native mussel species. Introduced starlings and sparrows dominate the avifauna in many settled landscapes. The landscapes and habitats of large parts of the continent have been drastically altered from pre-Columbian times. Almost all rivers in the United States have had their flow altered by dams or their drainage basin hydrology has been manipulated. Terrestrial habitats reasonably unaltered from pre-Columbian times now occupy only 48% of North America north of Mexico, and these are largely concentrated in remote mountainous or northern regions. Remnants of natural prairies and forest habitats represent only a few percent of many ecoregions in eastern North America. In western North America natural habitats are essentially obliterated in the grasslands of central California and in the interior grasslands of the Pacific Northwest. We are faced with the challenge of identifying natural patterns in the biodiversity of our native flora and fauna while the continental biota is increasingly disrupted by the introduction of alien species and the alteration of natural landscapes.

V. CONCLUSION

From the preceding discussion, assembled from a variety of sources, one point is most clear: We have really only begun to know the patterns of biodiversity in North America. The biodiversity of no group of organisms in North America is completely characterized. Some of the groups of organisms discussed in this review are reasonably well studied, but all are incompletely known. Remote parts of arctic North America are little explored and may yield new species despite their general tendency of species diversity to decrease at higher latitudes. There are unusual habitats in the high arctic that are unlike anywhere else on Earth, such as the mineral-rich hot springs on Axel Heiberg Island that emerge through hundreds of meters of permafrost. Such habitats, which are being studied as analogs for Martian polar environments, may well harbor previously unknown microbial life. The more heavily settled and biologically explored midcontinent seems well-known, but every year new localities and range extensions, and even new species, are recorded. The remote and rugged parts of southern North America undoubtedly harbor unknown species even in groups that have been well studied on the continent as a whole. Mexico is recognized as one of the top five countries in the world in terms of overall levels of biodiversity. By far the most species-rich part of North America, Mexico is ironically the least well studied.

In addition to our incomplete knowledge of species diversity in well-studied groups of organisms in North America, we must recognize that there are many other groups of organisms on the continent whose distribution and abundance are scarcely known. There are almost 91,000 species of insects described from North America north of Mexico, but this number is certainly only a small fraction of the total insect species on the continent. Almost all the insect groups are too little studied to assemble any sort of meaningful summary of their patterns of continental biodiversity. Similarly, there may be as many as 120,000 species of macrofungi in the United States alone, only one-fourth of which have been described and far fewer have been studied sufficiently to map and inventory them for the continent. The 3000–10,000 species of macrofungi (mushrooms) estimated to occur in the United States are only marginally better known. It is too early to write the definitive treatise on patterns of biodiversity in North America. There is much exciting and worthwhile work yet to do.

See Also the Following Articles

CENTRAL AMERICA, ECOSYSTEMS OF • SOUTH AMERICA, ECOSYSTEMS OF

Bibliography


I. Gene Scrambling

II. RNA Editing

GLOSSARY

DNA
Linear polymer of nucleotides encoding information for a cell.

eukaryote
An organism whose cell or cells have a membrane-bound nucleus.

mitochondrion
A eukaryotic cellular organelle which is used for cellular respiration and energy production.

nucleus
Compartment of a cell in which DNA is stored on chromosomes.

protein
A three-dimensional macromolecule constructed of amino acids which is formed based on an RNA sequence.

protist
A member of a diverse collection of eukaryotes, defined only by their exclusion from the groups plants, animals, and fungi.

RNA
A linear polymer of nucleotides which is transcribed from DNA.

DNA IS OFTEN DESCRIBED as a “blueprint for life,” implying that knowledge of the primary sequence of nucleic acids in a genome can give biologists a complete picture of the organism built from these plans. However, neither life nor DNA is that simple. In reality, the sequence of nucleotides in a genome is often just a starting point for the construction of DNA genes and RNA transcripts, which undergo many alterations before organisms actually use the information to fashion their building materials.

Two fascinating modes of nucleic acid sequence modification are gene scrambling and RNA editing. Gene scrambling is the rearrangement of DNA segments between a transcriptionally active copy and an archived germline copy. Some genes are broken into more than 50 unordered fragments along a germline chromosome and then unshuffled during formation of an active somatic chromosome. While gene scrambling assembles existing DNA information into a new order, RNA editing of transcripts creates completely novel sequences which are not even found in the genome. Procession of edited RNAs alters the transcript length and nucleotide identity, and the transformations can render the expressed RNA form unrecognizably different from the DNA molecule of its origin. Although gene scrambling and RNA editing complicate our analysis of genomes, understanding the methods by which organisms achieve these genetic revisions gives us an appreciation for the diversity of ways by which nucleic acids can store and recombine information.

I. GENE SCRAMBLING

Gene scrambling occurs when coding segments of DNA are mixed in a randomly or nonrandomly shuffled order along a chromosome. Before the stored information can
be expressed by the organism, the pieces of the gene must be cut apart and reassembled in the proper sequential arrangement. These stunning acrobatics are performed in the genomes of spirotrichous ciliates, protists which have many unique and mysterious features. However, the fundamental lessons learned from spirotrichs about the flexibility of nucleic acid storage mechanisms are universally important.

A. Ciliates and Nuclear Dualism

Ciliates are unicellular protists closely related to the "eukaryotic crown taxa," meaning that on most phylogenetic trees they diverge as one lineage near the neighboring cluster of plants, animals, and fungi. The ciliates are a diverse monophyletic group, with certain ciliates estimated to be as evolutionarily distant from one another as corn from rats. All ciliates share two features: a coating of cilia on their cell surfaces and two types of nuclei within single cells.

The two nuclei types in each ciliate cytoplasm are different sizes; they are called the micronucleus and the macronucleus. The tiny germline micronucleus is transcriptionally inert and functions solely in sexual exchange. In contrast, the large somatic macronucleus is responsible for gene expression, but its contents are only transmitted to clonal offspring. Ciliates reproduce asexually by fission but are capable of exchanging genetic information in a sexual manner independent of reproduction. Conjugation between ciliates leads to an exchange of haploid micronuclei, which fuse to form a zygotic nucleus (Fig. 1). The biparentally created zygotic nuclei in each mating partner form new micronuclei and macronuclei as the old macronuclei are destroyed.

B. DNA Processing during Macronuclear Formation

The micronuclear and macronuclear genomes do not have the same chromosomal structure. As the new macronucleus is formed, diploid micronuclear chromosomes are polytenized and reproducibly broken at certain sites, portions of DNA are deleted, chromosomes are rejoined, and new chromosome ends are healed by telomerase action. DNA in the new, shortened chromosomes is differentially amplified so that copy numbers of individual chromosomes in Tetrahymena thermophila range from 45 to 9000. Spirotrichous ciliates tend to have higher copy numbers of chromosomes than the oligohymenophorean Thermoplasma; whereas spirotrichs such as Styloynchia and Oxytricha eliminate up to 98% of their micronuclear genomes to create macronuclear chromosomes that bear single genes and very little noncoding sequence.
FIGURE 2. Overview of gene unscrambling. Dispersed coding MDSs 1–7 reassemble during macronuclear development to form the functional gene copy (top), complete with telomere addition to mark and protect both ends of the gene.

C. Scrambled Genes and the Unscrambling Process

The drastic genome rearrangements of spirotrichous ciliates are not confined to the extreme quantity of DNA deletions. The protein-coding MDSs in *Oxytricha* and *Stylonychia* species are sometimes disordered relative to their final position in the macronuclear copy. For example, Prescott and colleagues found that in *O. nova*, the micronuclear copy of three genes (actin I, α telomere binding protein, and DNA polymerase α) must be reordered while intervening DNA sequences are removed in order to construct functional macronuclear genes. In *O. nova*'s micronuclear genome, the MDSs destined to construct the gene for α telomere binding protein (α-TP) are arranged in the cryptic order 1-3-5-7-9-11-2-4-6-8-10-12-13-14 relative to their conventional order in the macronucleus of 1-2-3-4-5-6-7-8-9-10-11-12-13-14. Most impressively, the gene encoding DNA polymerase (DNA pol α) in *O. trifallax* and *S. lemnae* is apparently scrambled in 48 or more pieces in their germline nuclei (Fig. 2).

Homologous recombination at MDS boundaries probably helps guide the unscrambling process. A segment of DNA sequence at the junction between a particular MDS and its downstream IES usually matches the sequence at the junction of the next MDS and its upstream IES, leading to the correct ligation of the two MDSs over a distance. However, the presence of shared repeat regions as short as an average of 4 base pairs (bp) for nonscrambled MDSs and 9 bp for scrambled MDSs suggests that although these recombination guides may be necessary, they are certainly not sufficient to guide accurate assembly of the genes. Hence, it is more likely that the repeats satisfy a structural requirement for MDS joining rather than perform any role in substrate recognition. Otherwise, incorrectly spliced products of promiscuous recombination would dominate genomes since 2–4 bp repeats occur many thousands of times throughout the micronucleus. This incorrect hybridization could be a driving force in the production of newly scrambled patterns in evolution. Nonetheless, if this sort of ambiguous unscrambling actually occurs during macronuclear development, then only unscrambled molecules which contain both 5′ and 3′ telomere addition sequences are selectively retained in the macronucleus, ensuring that most haphazardly ordered genes would be lost.

The pattern of MDSs in the micronuclear genome provides a strong clue to both the ciliates' orchestration of the unscrambling process and the mutational events which led to the scrambling in each gene sequence. For example, in the previously mentioned gene encoding *O. nova*'s α-TP, the micronuclear order of MDSs (1-3-5-7-9-11-2-4-6-8-10-12-13-14) predicts a spiral mechanism in the unscrambling path to link odd and even segments in order (Fig. 3).

In contrast, DNA polymerase α has at least 44 MDSs in *O. nova* and 51 in *O. trifallax*, scrambled in a nonrandom order with an inversion in the middle; some MDSs are located at least several kilobases away from the main gene in an unmapped fragment. A hairpin structure
Gene scrambling in ciliates may have evolved as a product of an increased capacity for homologous recombination. The reason why it has been detected in only a restricted group of ciliates might reflect their increased levels of recombination, which generate both the scrambled arrangements and the subsequent process of unscrambling them. Although it is difficult to propose an adaptive argument for the presence and maintenance of such a complicated gene-decoding procedure, the forces that have led to ciliate nuclear dualism might be at the root of its origin. For example, Yao and Doerder independently proposed that the different roles required of the micronucleus and macronucleus, within the micromolecular genome. Consequently, selection would favor the appearance of more scrambled MDS segments in such a nonrandomly scrambled gene since each additional MDS adds more paired junction sequences to stabilize the hairpin necessary for unscrambling the already existing MDSs. This explanation is consistent with the additional MDSs in the DNA pol α gene in *O. trifallax*. The arrangement of MDSs 2, 6, and 10 in *O. nova* could have given rise to the arrangement of eight new MDSs in *O. trifallax* (Fig. 4) by multiple crossovers in the germline micronucleus. Thus, the appearance of an inversion leads to the introduction of new MDSs in a nonrandomly scrambled pattern.

Gene scrambling in ciliates may have evolved as a product of an increased capacity for homologous recombination. The reason why it has been detected in only a restricted group of ciliates might reflect their increased levels of recombination, which generate both the scrambled arrangements and the subsequent process of unscrambling them. Although it is difficult to propose an adaptive argument for the presence and maintenance of such a complicated gene-decoding procedure, the forces that have led to ciliate nuclear dualism might be at the root of its origin. For example, Yao and Doerder independently proposed that the different roles required of the micronucleus and macronucleus, within the micromolecular genome. Consequently, selection would favor the appearance of more scrambled MDS segments in such a nonrandomly scrambled gene since each additional MDS adds more paired junction sequences to stabilize the hairpin necessary for unscrambling the already existing MDSs. This explanation is consistent with the additional MDSs in the DNA pol α gene in *O. trifallax*. The arrangement of MDSs 2, 6, and 10 in *O. nova* could have given rise to the arrangement of eight new MDSs in *O. trifallas* (Fig. 4) by multiple crossovers in the germline micronucleus. Thus, the appearance of an inversion leads to the introduction of new MDSs in a nonrandomly scrambled pattern.

Gene scrambling in ciliates may have evolved as a product of an increased capacity for homologous recombination. The reason why it has been detected in only a restricted group of ciliates might reflect their increased levels of recombination, which generate both the scrambled arrangements and the subsequent process of unscrambling them. Although it is difficult to propose an adaptive argument for the presence and maintenance of such a complicated gene-decoding procedure, the forces that have led to ciliate nuclear dualism might be at the root of its origin. For example, Yao and Doerder independently proposed that the different roles required of the micronucleus and macronucleus, within the micromolecular genome. Consequently, selection would favor the appearance of more scrambled MDS segments in such a nonrandomly scrambled gene since each additional MDS adds more paired junction sequences to stabilize the hairpin necessary for unscrambling the already existing MDSs. This explanation is consistent with the additional MDSs in the DNA pol α gene in *O. trifallax*. The arrangement of MDSs 2, 6, and 10 in *O. nova* could have given rise to the arrangement of eight new MDSs in *O. trifallax* (Fig. 4) by multiple crossovers in the germline micronucleus. Thus, the appearance of an inversion leads to the introduction of new MDSs in a nonrandomly scrambled pattern.
Figure 5. Proposed model for the origin of a scrambled gene. (Left) Birth of a scrambled gene from a nonscrambled gene by a double recombination with an IES or any noncoding DNA (new MDS order 1-3-2 with an inversion between MDSs 3 and 2). (Middle) Generation of a scrambled gene with a nonrandom MDS order from a nonscrambled gene with an inversion between two MDSs. (Right) Creation of new scrambled MDSs in a scrambled gene containing an inversion. Inversions may dramatically increase the production of scrambled MDSs by stabilizing the folded conformation that allows reciprocal recombination across the inversion.

as well as intragenomic competition, may have led to disruptive selection acting on their genetic organization and chromatin structure. The accommodation of both types of nuclei within a single cytoplasm may promote or at least permit profound differences to arise that distinguish active genes from transmitted genes. Likewise, the gene scrambling present in certain sporozoan ciliates may be a profound exaggeration of this solution to problems presented by ciliate life.

II. RNA EDITING

"RNA editing" is the alteration of RNA sequences by base modifications, substitutions, insertions, and deletions. The process of rewriting RNA transcripts by editing produces major effects, even adding more than half of the nucleotides in some mitochondrial transcripts of kinetoplastid protozoa. In contrast, the impact of editing can still be large even in cases in which the physical extent of editing is small. For instance, in human apolipoprotein B transcripts, replacement of a single cytidine (C) by uridine (U) results in the conversion of a glutamine codon to a stop codon; the early termination of apob translation shortens the resultant polypeptide by one-half and removes a functional domain. Since the initial discovery of RNA editing as extensive U insertions and deletions in trypanosome mitochondrial mRNAs, many additional and apparently unrelated examples of editing have been found in organisms ranging from Ebola virus to humans. Substitution/modification editing exists in certain nuclear and organellar RNAs among a diverse set of eukaryotes; however, insertions/deletions editing has only been found in mitochondrial RNAs of two protist groups (Table I).

Upon first inspection, the use of RNA editing in gene expression seems inefficient. Why not encode genes in their final edited form rather than require an additional revision step? However, organisms exploit the ability to edit RNA in amazingly clever ways. RNA editing allows the persistence of "deleterious" mutations in DNA genomes: New point mutations and frameshifts may persist if they are repaired at the RNA level, and DNA copies of genes and control sequences in crowded genomes may overlap but generate two or more discrete RNA sequences through editing. RNA editing also offers an array of posttranscriptional modes of genetic regulation through the formation of start and stop codons, intron splice sites, and open reading frames. Editing even permits single genes to produce multiple peptides, allowing combinatorial protein diversity.

The following sections review several models which explain our current understanding of the molecular and evolutionary basis of numerous forms of RNA editing.

A. Kinetoplastids: Cryptogenes to Proteins by U Insertion/Deletion

Kinetoplastids are a group of unicellular parasites which include the pernicious trypanosome and leishmania parasites responsible for deadly human diseases such as African sleeping sickness and Chagas disease. Trans-
TABLE I
RNA Editing in Eukaryotes and Viruses

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome</th>
<th>Class of RNA</th>
<th>Form of editing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinetoplastids</td>
<td>Mitochondria</td>
<td>mRNA</td>
<td>U insertion, U deletion</td>
</tr>
<tr>
<td>Myxomycetes</td>
<td>Mitochondria</td>
<td>mRNA, tRNA</td>
<td>C insertion, U insertion, mixed dinucleotide insertion, C → U</td>
</tr>
<tr>
<td>Plants</td>
<td>Mitochondria</td>
<td>mRNA, tRNA</td>
<td>C → U, U → C</td>
</tr>
<tr>
<td>Plants</td>
<td>Chloroplast</td>
<td>mRNA</td>
<td>C → U, U → C</td>
</tr>
<tr>
<td>Humans, rodents</td>
<td>Nucleus</td>
<td>mRNA</td>
<td>C → U</td>
</tr>
<tr>
<td>Humans, rodents, fish</td>
<td>Nucleus</td>
<td>mRNA</td>
<td>A → U</td>
</tr>
<tr>
<td>Humans</td>
<td>Nucleus</td>
<td>mRNA</td>
<td>U → C</td>
</tr>
<tr>
<td>Drosophila</td>
<td>Nucleus</td>
<td>mRNA</td>
<td>U → A</td>
</tr>
<tr>
<td>Marsupials</td>
<td>Mitochondria</td>
<td>tRNA</td>
<td>C → U</td>
</tr>
<tr>
<td>Monostomates</td>
<td>Mitochondria</td>
<td>tRNA</td>
<td>U → A, U → C, A → C</td>
</tr>
<tr>
<td>Land snails</td>
<td>Mitochondria</td>
<td>tRNA</td>
<td>C → A, U → A, G → A</td>
</tr>
<tr>
<td>Ledge (squid)</td>
<td>Mitochondria</td>
<td>tRNA</td>
<td>G → A</td>
</tr>
<tr>
<td>Chicken</td>
<td>Mitochondria</td>
<td>tRNA</td>
<td>G → A</td>
</tr>
<tr>
<td>Acanthamoeba</td>
<td>Mitochondria</td>
<td>tRNA</td>
<td>U → A, U → G, A → G</td>
</tr>
<tr>
<td>Spicellomyces (fungus)</td>
<td>Mitochondria</td>
<td>tRNA</td>
<td>A → G, U → G, U → A, C → A</td>
</tr>
<tr>
<td>Hepatitis delta virus</td>
<td>Viral</td>
<td>mRNA</td>
<td>A → I</td>
</tr>
<tr>
<td>Paramyxoviruses</td>
<td>Viral</td>
<td>mRNA</td>
<td>G insertion (by polymerase ester)</td>
</tr>
<tr>
<td>Ebola virus</td>
<td>Viral</td>
<td>mRNA</td>
<td>A insertion (by polymerase ester)</td>
</tr>
</tbody>
</table>

lation of mitochondrial mRNAs of kinetoplastids is impossible without massive RNA editing of the transcripts by insertion and deletion of uridine. This editing drastically rewrites the coding regions of the transcripts by introducing and fixing frameshifts as well as creating stop and start codons in some kinetoplastid genes, editing creates more than 90% of the amino acid codons. Since the DNA copies of many genes are barely recognizable as sources of their cognate mRNAs, they have been named "cryptogenes."

Although all kinetoplastids display specific, reproducible editing patterns in certain mitochondrial mRNAs, comparison of the editing patterns in a particular gene transcript across a variety of species reveals a trend in the extent of RNA editing within each species. The earlier diverging kinetoplastids exhibit more editing within the cytochrome c oxidase III transcript than their later diverging relatives. This implies that this type of RNA editing is ancient within this lineage, and that its use has decreased over evolutionary time. However, the lack of U insertion/deletion editing in any other type of organism suggests that it arose specifically within the kinetoplastid lineage. Cavalier-Smith proposed that glycosomes, special organelles found only in kinetoplastids, may provide a clue to the origin of editing in these protists. Glycosomes permit an efficient anaerobic lifestyle, during which RNA editing may have become fixed by drift in the seldom-used mitochondria of the early kinetoplastids.

Guide RNAs (gRNAs) provide the specificity of the kinetoplastid editing mechanism. gRNAs are tiny RNA transcripts which guide the editing machinery through base pairing with the mRNAs in edited regions. The gRNAs are complementary to small segments of the fully edited mRNAs and thus serve as minitemplates to guide the addition and deletion of uridines from the preedited mRNA as they form a locally double-stranded RNA helix. For instance, an unpaired A in a gRNA signals for an editing event to insert a U into the mRNA. Complete editing of a gene requires a set of many overlapping gRNAs. The affiliated set of overlapping gRNAs sequentially edit the mRNA from its 3’ end to its 5’ end. During editing, a complex of proteins called the “editosome” catalyzes the insertion and deletion of uridines until the length of each paired gRNA/mRNA region is maximized. The extent of G • U wobble pairs
FIGURE 6. Editing of a gene region by four overlapping gRNAs. Thick lines in the mRNA are encoded in the mitochondrial DNA. Thin shaded lines are inserted U's; the two asterisks are deleted U's (Maslov and Simpson, 1992). Thin lines in the gRNAs are guide nucleotides (A or G) that pair with inserted U's. Vertical lines indicate Watson–Crick base pairs; colons indicate G:U wobble base pairs, illustrating formation of well-paired ‘‘anchors’’ between the 5’ ends of gRNAs and the corresponding region of the mRNA.

and conventional A–U and G–C pairs of each gRNA/mRNA pair controls the cascade of editing. Each subsequent gRNA binds the mRNA more stably than the last by possessing more Watson–Crick base pairs (A–U and G–C) in the so-called ‘‘anchor region’’ on one end of the gRNA. These Watson–Crick base pairs displace the less stable G·U base pairs on the opposite end of the previous gRNA to dislodge the upstream gRNA and lead to an overall 3’ to 5’ directionality of editing (Fig. 6).

gRNAs are encoded in the kinetoplastid mitochondrial genome. Kinetoplastids are named after the unusual structure of their mitochondrial genomes, which consist of a ‘‘kinetoplast,’’ a network of DNA inside a single, large mitochondrion located at the base of the flagellum. In one kinetoplastid subdivision, the trypanosomatids, the kinetoplast consists of a few identical maxicircles of 20–40 kb intertwined with thousands of heterogeneous minicircles of 0.5–3 kb. The maxicircles encode the mitochondrial genes and cryptogenes for respiratory proteins, ribosomal proteins, and rRNA. The only known role of the minicircles is to encode gRNAs. Bodonids, the other group of kinetoplastids, share the presence of a DNA network with the trypanosomatids but have a different structure for their minicircle homologs. In Bodo caudatus, the ‘‘minicircles’’ are noncatedated 1.4-kb circles. Trypanoplasma borreli has minicircle-like structures of 170 and 200 kb, with tandem 1-kb repeats encoding most gRNAs.

How did the gRNAs and minicircles come to exist as they do today? A product of mitochondrial DNA recombination provides a plausible explanation. Lunt and Hyman (1997) identified a minicircle of DNA excised from the major mitochondrial genome circle in a nematode. Similar intramolecular recombination within a kinetoplastid maxicircle might have led to the formation of minicircles encoding tiny portions of mRNAs while other unrecombined maxicircles still contained complete functional copies of the mRNA. If the minicircle copy of a gene fragment acquired a promoter that allowed transcription of an antisense RNA, such a gene-encoding complementary RNA could have given rise to a proto gRNA. If the protein equipment responsible for catalyzing RNA insertions and deletions simultaneously arose or was recruited from another function to serve in RNA editing, mutations in the mitochondrial mRNA could have been repaired by editing. gRNA-containing minicircles would be selectively retained in the genome, ensuring their survival.

B. Myxomycetes: The Slimes, They Are A-changin’!

Physarum polycephalum is a myxomycete, or plasmodial slime mold. It takes on many shapes and sizes throughout its life, morphing from microscopic amoeba to a multinucleate syncytium which can be as large as several feet across, and then forming millimeter-scale delicate, mushroom-like fruiting bodies. Physarum’s mitochondrial transcripts for almost all messenger and structural RNAs require several types of RNA editing to create functional products. Editing positions are sprinkled regularly throughout the entire length of each edited transcript, spaced an average of 27 bases apart, but never closer than nine nucleotides. The majority of editing events are insertions of cytosine, but there are also a small number of specific and reproducible insertions of uridines and dinucleotides into sequences. Although there appears to be no consensus sequence that defines the hundreds of insertion sites, they usually follow a purine/pyrimidine dinucleotide and show significant preference for third codon position. Visomirski-Robic and Got (1997) found that under conditions of stalled RNA polymerization, sites only 14–22 nucleotides from the 3’ end of the RNA have correct insertion editing, suggesting that the editing reactions closely follow transcription and they may even be coupled. The tight association of transcription and editing yields a high editing efficiency; unedited or partially edited transcripts are rarely detected. Physarum insertional editing must occur within a narrow window of possibility:
If editing is restricted in vitro by low concentrations of "the required nucleotide" CTP, and CTP is subsequently restored, then sites that were "missed" by the editing process never get edited. Myxomycete insertional editing seems distinct from the uridine insertion/deletion editing found in the kinetoplastids. This is supported by the dissimilar pattern of edited sites, the different identity of nucleotides involved, and Physarum's apparent lack of gRNA-like template molecules. Although these differences set myxomycete editing apart, another remarkable feature of myxomycete editing makes it unique among all other editing systems. Myxomycetes are the only organisms known to combine multiple types of editing within the same transcript. For example, *P. polycephalum*'s 1.5-kb transcript encoding cytochrome c oxidase subunit I (coI) undergoes 5' C insertions, a single U insertion, three dinucleotide insertions, and, astoundingly, four sites of C/L50478U base conversion (Fig. 7). C insertion has been separated from C/L50478U conversion in isolated mitochondria, implying separate mechanisms or components. We have found that insertional and base conversion editing have distinct evolutionary histories. The col gene of *Stemonitis flavogenita*, another myxomycete, shares all three types of insertional editing with *P. polycephalum* but lacks the C/L50478U conversion editing in this transcript. Even the three types of insertional editing did not all evolve simultaneously because U insertional editing is found in all myxomycetes to date, whereas C insertion and dinucleotide insertion are found only in some slime molds.

The sites of C/L50478U editing in *P. polycephalum* are all in first or second codon position, similar to the case in plant organelle editing. Also similar is that the process of editing restores the conserved peptide sequence in this region so that the unedited *S. flavogenita* transcript codes for the same amino acids at these positions as does the edited *Physarum* transcript. In contrast, the role of the inserted nucleotides, usually in third codon positions, may be primarily to restore the correct reading frame since they rarely change crucial coding information.

**C. Plant Organelles: C → U and U → C Editing Restores Conserved Amino Acids**

Plant mitochondria employ rampant RNA editing. In nearly every mitochondrial mRNA, RNA editing converts many cytidines to uridines, and some mitochondrial mRNAs have uridine to cytidine conversions as well. Several plant chloroplast RNAs also show similar editing patterns. Although the mechanism for plant organellar editing is not fully understood, when CTP residues of in vitro transcripts were labeled on their 5'-phosphate and on their cytosine base, the labels were retained after editing. This means that the C/L50478U changes occur through a deamination of cytidine rather than a base or nucleotide substitution. Plants with different nuclear genotypes exhibit different degrees of editing of transcripts from identical mitochondrial genomes, indicating that at least part of the editing machinery is nuclearly encoded. Although no consensus sequence for editing site recognition has been determined, recent analysis (Bock et al., 1997; Williams et al., 1998) of natural mutants and creation of mutant sequences indicate that the specificity for editing sites in both organelles lies in
RNA editing is present in both the chloroplast and the mitochondrial genomes of all land plants, but it is absent in green algae and the liverwort Marchantia polymorpha. This distribution suggests that the plant editing systems share common components and may have arisen simultaneously in both organelles or have been transferred from one to the other. However, although the editing seems to appear suddenly in the land plant lineage, the degree of mitochondrial or chloroplast editing does not correlate with phylogenetic position, indicating multiple evolutionary losses and gains of editing at particular gene sites.

Plants which possess the ability to perform base conversion may exploit the conversions as a mechanism for repairing sequences disturbed by mutational drift. The observation that most of the edited sites in plant genes are in first or second codon positions supports this hypothesis because editing produces nonsynonymous changes in amino acids. Furthermore, Malek et al. (1996) found that the level of mitochondrial editing in plants correlates with G/C content. Indeed, Lu and associates (1998) determined that editing "reverses" all nonsynonymous U → C substitutions at the RNA level in col genes from eight gymnosperm species, eliminating almost all variation in the predicted protein sequences. Editing can regulate translation or enzyme activity; it creates stop and start codons and proteins of varying function. However, the pressure toward loss of editing appears to be greater than any selective advantage it confers since Shield and Wolle's (1997) comparison of edited sites over many plant species reveals a high rate of mutation of the genes toward the edited (uridine) nucleotide.

D. Mitochondrial tRNA Editing: Acceptor Stems and Anticodons

Mitochondrial genomes of some organisms seem pressed to economize space. "Junk" DNA is held to a minimum, and genes immediately about their neighbors. Some tRNA genes overlap their 5’ and 3’ extremities. Base pairing of a tRNA’s 5’ and 3’ ends is essential to form an acceptor stem for aminoacylation. In six described cases of tRNA editing, mitochondria repair incompletely paired tRNA acceptor stems by exchanging mismatched bases on one side of the acceptor stem for ones that complement those bases present on the opposing stem. Thus, editing and the presence of an internal acceptor stem template allow the genome to repair mutations that may occur in response to selection for either nucleotide compositional bias or sequence compression.

In land snails, squids, and chickens, the overlapping ends of certain tRNAs disturb the crucial base pairing of the acceptor stems. RNA editing activity restores complementarity to these tRNAs by replacing mismatched bases with adenosines. In a platypus tRNA, there are three exchanged nucleotides (U → A, U → C, and A → C) in the 3’ half of the stem to complement the 5’ half of the acceptor stem. There is also editing of mitochondrial tRNA acceptor stems in the amoeboid protist Acanthamoeba and the fungus Spizellomyces, using the second half of the stem as guides. However, the editing mechanism in these two organisms may be unrelated to the RNA editing described in animals since it occurs on the 5’ half of the stem. Editing of the Acanthamoeba and Spizellomyces tRNAs exchanges uridines, adenosines, and cytidines in the first three 5’ nucleotides for purines complementary to a corresponding pyrimidine on the 3’ template side of the stem.

In opossums and other marsupials, the RNA encoding mitochondrial tRNAAsp has the wrong anticodon. Although the rest of the tRNA has a canonical tRNAAsp sequence, the anticodon, GCC, will pair with glycine codons rather than aspartic acid codons. About half of the tRNAAsp transcripts are not edited and consequently are aminoacylated with glycine. The remainder are edited by a C → U change that restores the GUC aspartic acid anticodon. These are charged correctly with aspartic acid since this anticodon is used as a determinant of tRNA charging. Surprisingly, the standard tRNAAsp (with a UCC anticodon) is present in the genome as well, and it could theoretically recognize all four GGN glycine codons by wobble. However, a mutation just outside of the anticodon region of this tRNAAsp renders it incapable of recognizing the two glycine codons which the mutated tRNAAsp can bind.

Boerner and Paabo (1996) suggest that marsupial mitochondrial editing became fixed through two mutational steps. First, a mutation occurred in the anticodon for tRNAAsp, transforming it into a functional tRNAAsp. Genomes with this mutation were still viable since the altered tRNA could retain its necessary function as tRNAAsp through editing. A subsequent mutation in the original tRNAAsp eventually left it unable to recognize the two glycine codons that the newly mutated
tion in the original tRNA^Gly, back-mutation of the tRNA^Gly/Asp anticodon (to form simply a tRNA^Asp anticodon) would be deleterious since the mutant tRNA^Asp/Asp serves double duty by pairing with both aspartic acid and glycine codons. Thus, RNA editing permitted a deleterious change in a genome and then became fixed when a second mutation made editing a requirement for expression of mitochondrial proteins.

E. Human Apo B and NFI: Shorter Peptides by Editing

Apolipoprotein B is present in two different forms in both humans and rodents. The long form, apo B100, is part of very low-density lipoprotein particles which transport dietary lipids. The two forms of the apo B protein are actually encoded by a single gene which undergoes tissue-specific C → U RNA editing at nucleotide position 6666. The editing converts an encoded glutamine codon into a stop codon to create the short form of the peptide.

A complex of proteins edits the apo B transcript. The main catalytic peptide is apo B RNA editing cytidine deaminase subunit 1 (APOBEC-1). APOBEC-1 expression varies throughout development, and apo B editing levels vary correspondingly. Specificity of the cytidine deamination is determined by the primary sequence of the apo B RNA transcript, particularly an 11-base “mooring sequence,” located just four nucleotides downstream of the edited site. In addition to this required mooring sequence three other “efficiency” elements are found in the 140-base region surrounding the edited cytidine. These upstream and downstream elements increase the effectiveness of the editing reaction.

APOBEC-1 is a part of a family of cytidine deaminases. The family is divided into two groups of larger and smaller deaminases with various structural features in common. APOBEC-1 is categorized as a member of the group of larger deaminases, as in Escherichia coli cytidine deaminase (ECCDA). The two enzymes are approximately the same size, form homodimers, and share structural features such as the carboxy-terminal core domain, which is absent in the smaller, homotetramer-forming deaminases. ECCDA catalyzes the deamination of single cytidine nucleosides as part of bacterial biosynthetic pathways. Comparison of the APOBEC-1 primary sequence with the ECCDA crystal structure predicts the presence of an additional hollow space within the APOBEC-1 structure. Navaratnam and associates (1998) noted that this cavity is the correct size and shape to accommodate an RNA transcript, which suggests how this type of editing might occur.

Cytidine deamination may even play a role in human tumorigenesis. For example, an imperfect APOBEC-1 mooring sequence and efficiency elements are present within the coding region of the neurofibromatosis type 1 (NFI) tumor suppressor gene. Despite slight differences in position and sequence context, normal individuals exhibit very low levels of editing. In comparison, tumor tissues from NFI-affected individuals showed more than eight times the normal quantity of edited transcript. The NFI gene encodes the neurofibromin protein, which is a putative homolog of yeast proteins in the ras signal transduction pathways. The proposed GTPase activating domain of the neurofibromin lies just downstream of the NFI editing site, and indeed C → U RNA editing at the site transforms an arginine codon into a premature translation stop upstream of the domain. Thus, the editing of the NFI transcript most likely cripples the tumor-suppressing activity of neurofibromin.

F. A → I Deamination: One Gene, Many Peptides

Deamination of another type alters additional RNA sequences of humans and rodents. Removal of adenosine’s amino group yields inosine, a nucleotide that the translation machinery reads as guanosine. Several such A → I transitions cause predicted amino acid changes in glutamate and serotonin receptor subunits in the human and rodent brain. Teleost fish also share one of these A → I editing sites in their glutamate receptors. The proposed GTPase activating domain of the glutamate receptors may contain an imperfect APOBEC-1 sequence, which could translate. After this second mutation in the original tRNA^Gly, back-mutation of the tRNA^Gly/Asp anticodon (to form simply a tRNA^Asp anticodon) would be deleterious since the mutant tRNA^Asp/Asp serves double duty by pairing with both aspartic acid and glycine codons. Thus, RNA editing permitted a deleterious change in a genome and then became fixed when a second mutation made editing a requirement for expression of mitochondrial proteins.

A complex of proteins edits the apo B transcript. The main catalytic peptide is apo B RNA editing cytidine deaminase subunit 1 (APOBEC-1). APOBEC-1 expression varies throughout development, and apo B editing levels vary correspondingly. Specificity of the cytidine deamination is determined by the primary sequence of the apo B RNA transcript, particularly an 11-base “mooring sequence,” located just four nucleotides downstream of the edited site. In addition to this required mooring sequence three other “efficiency” elements are found in the 140-base region surrounding the edited cytidine. These upstream and downstream elements increase the effectiveness of the editing reaction.

APOBEC-1 is a part of a family of cytidine deaminases. The family is divided into two groups of larger and smaller deaminases with various structural features in common. APOBEC-1 is categorized as a member of the group of larger deaminases, as in Escherichia coli cytidine deaminase (ECCDA). The two enzymes are approximately the same size, form homodimers, and share structural features such as the carboxy-terminal core domain, which is absent in the smaller, homotetramer-forming deaminases. ECCDA catalyzes the deamination of single cytidine nucleosides as part of bacterial biosynthetic pathways. Comparison of the APOBEC-1 primary sequence with the ECCDA crystal structure predicts the presence of an additional hollow space within the APOBEC-1 structure. Navaratnam and associates (1998) noted that this cavity is the correct size and shape to accommodate an RNA transcript, which suggests how this type of editing might occur.

Cytidine deamination may even play a role in human tumorigenesis. For example, an imperfect APOBEC-1 mooring sequence and efficiency elements are present within the coding region of the neurofibromatosis type 1 (NFI) tumor suppressor gene. Despite slight differences in position and sequence context, normal individuals exhibit very low levels of editing. In comparison, tumor tissues from NFI-affected individuals showed more than eight times the normal quantity of edited transcript. The NFI gene encodes the neurofibromin protein, which is a putative homolog of yeast proteins in the ras signal transduction pathways. The proposed GTPase activating domain of the neurofibromin lies just downstream of the NFI editing site, and indeed C → U RNA editing at the site transforms an arginine codon into a premature translation stop upstream of the domain. Thus, the editing of the NFI transcript most likely cripples the tumor-suppressing activity of neurofibromin.

F. A → I Deamination: One Gene, Many Peptides

Deamination of another type alters additional RNA sequences of humans and rodents. Removal of adenosine’s amino group yields inosine, a nucleotide that the translation machinery reads as guanosine. Several such A → I transitions cause predicted amino acid changes in glutamate and serotonin receptor subunits in the human and rodent brain. Teleost fish also share one of these A → I editing sites in their glutamate receptors. These editing events adjust the calcium channel permeability and the speed of the desensitization response in glutamate receptors. In the serotonin receptors, protein loops encoded by the edited region of the RNA interact with G proteins to turn on signaling pathways. The editing, along with alternative splicing, allows the exquisite fine-tuning of neural responses by increasing combinatorial protein diversity available from a single transcript. Posttranscriptional regulation demands the expenditure of less time and energy than transcription of multiple gene sequences.

A family of enzymes called ‘‘ADARs’’ (adenosine deaminases acting on RNA) are the candidates for A → I editing activity based on in vitro experiments. At least three different ADAR enzymes are expressed differentially in human tissues, and their specific targets have not been definitively resolved in vivo, although they
display discrete editing abilities at various mRNA sites in vitro. The double-stranded RNA regions required for ADAR action are elegantly provided by pairing of the exonic sequences with intronic sequences in the pre-mRNAs. Neither mature mRNAs whose introns have been removed by splicing nor pre-mRNAs with mutations disturbing intron/exon complementarity act as substrates for editing in transfected cell lines.

Mice genetically disabled for A → I editing at a single site demonstrate the importance of RNA editing in gluta-

tamate receptors. Bruscia et al. (1995) removed an intron region crucial for creating a double-stranded editing substrate from the mouse GluR-B gene. Heterozygous for this editing-disabled allele displayed severe epileptic seizures and premature death within 3 weeks of birth. The decrease in edited mRNA levels led to a fivefold increase in the calcium flow into their neurons and serious damage to their nervous systems.

Vertebrates are not alone in using the A → I editing activity to generate diversity in nervous system compo-

nents. Recent work (Petschek et al., 1996; Smith et al., 1996) reveals that two Drosophila mRNAs have appar-

ent A → G substitutions between their genomic and RNA copies. The if-mp gene encoding a putative RNA-

binding protein is expressed in several alternatively spliced variants during fruit fly development. The adult form, expressed heavily in the central nervous system, has probable deamination editing of 263 adenosine sites. Another Drosophila gene, which encodes a subunit of a calcium channel protein, has seven putative A → I editing sites, five of which alter codon identity, in addition to many alternatively spliced forms.

Scott (1995) suggests that ADARs might have evolved for an antiviral function, destabilizing RNA genomes by modification and subsequent unwinding. Interestingly, a devious viral genome coopted this func-

tion and uses host adenosine deaminases to its own benefit. Hepatitis delta virus (HDV), a single-stranded RNA subvirus of hepatitis B, has only one known gene product, the delta antigen. The short form of delta anti-
gen stimulates replication of the genome, whereas the long form of the antigen suppresses replication and is necessary for packaging HDV. The short and long forms of the delta antigen differ only by a single A → I base change of the negative strand, or antigenome. The base conversion replaces a stop codon at the end of the shorter peptide’s mRNA with a tryptophan codon to allow read through of the longer peptide. In vitro, ADAR1 is capable of inducing this A → I change on the antigenome of the HDV virus, suggesting that either this or a related enzyme edits HDV in vivo. In exquisite control of the reaction, the long-form delta antigen acts as a repressor of editing in a negative feedback loop. The repression prevents accumulation of unduly high levels of long-edited antigen, which would lower the level of viral replication.

G. Paramyxoviruses and Ebola Virus: Poly
erase Stutter Unites

Reading Frames

Two other types of viruses also capitalize on RNA edit-

ing processes as a method of increasing information storage while under pressure for genome size con-

straint. The RNA genomes of the paramyxoviruses have several overlapping genes. Some of these genes are acti-

vated for expression by rhosomal choice, whereas oth-

ers become available by cotranscriptional RNA editing. The polymerase may slip as it travels through a purine run, adding from one to six extra guanosines, depending on the sequence present in the particular virus species. The stuttering polymerase thereby fuses coding regions together to make extended versions of genes.

Many viruses use an apparently similar mechanism to add poly A tails to mRNA transcripts. When the RNA polymerase reaches the U-rich regions, it adds additional nontemplated adenosines to create the tails. In paramyxoviruses, RNA editing activity may be related to a process that corrects genome length, maintaining length in multiples of six nucleotides. The substrate for the RNA polymerase is the RNA genome complexed with the capsid in hexamer-length segments. Hausmann and colleagues (1996) note that if the genome is not a multiple of six, the polymerase inserts or deletes guano-
sines or adenosines from the same region in which the RNA editing occurs.

Another RNA virus, the infamous Ebola virus, uses a comparable mechanism to insert adenines into a se-

quence to unite two reading frames. In this case, the sequence of the site where adenine addition occurs appears similar to vital poly A addition sites. In contrast to the paramyxovirus editing, the additions occur even when the RNA is produced by a nonnative polymerase and thus must be intrinsic to the template sequence or structure. T7 transcription of the Ebola mRNA in vitro results in edited product, although in a lower quantity than when transcribed by Ebola’s RNA polymerase.

Thus, once considered a molecular anomaly unique to protists, RNA editing is now recognized as a vital part of gene expression in a wide distribution of eukary-

totes and their viruses. Organisms can alter RNA se-

quences through many different mechanisms by recruit-

ing enzymes such as deaminases, nuclease, ligases,
and special polymerases to specific RNA editing tasks. Editing clearly arose multiple times in evolutionary history in the various forms of both insertional and substitutional editing, all of which profoundly affect the expression of RNAs, the products they form, and the biological processes in which they participate. RNA editing is significant in cancers, cholesterol regulation, and neural function in vertebrates as well as in the de novo creation of coding sequences from obscured mitochondrial genes.

Furthermore, the impact of editing on the genomes which employ it is astounding. Genomes that use editing may become increasingly lenient toward persistence of deleterious mutations. Large amounts of genetic information and control devices may be confined to small spaces. Lastly, RNA editing, like gene scrambling, establishes a device for generating combinatorial sequence diversity.

Acknowledgments

We thank Laura A. Katz and Andrew L. Goodman for helpful discussions. TLH is supported by a National Defense Science and Engineering Graduate Fellowship.

See Also the Following Articles

DIVERSITY, MOLECULAR LEVEL • GENES, DESCRIPTION OF • GENETIC DIVERSITY

Bibliography


I. History of the Ecosystem Concept
II. Utility of the Concept
III. Partitioning the Ocean into Natural Functioning Units
IV. Characteristics of Ocean Ecosystems
V. Ocean Ecosystems and Global Change

GLOSSARY

convective mixing Vertical mixing produced by the increasing density of a fluid in the upper layer, especially during winter in temperate and polar regions.
euphotic zone The surface layer of the ocean in which there is adequate light for net positive photosynthesis.
nutrients Dissolved mineral salts necessary for primary productivity and phytoplankton growth: Macronutrients are phosphate, nitrate, and silicate; micronutrients are iron, zinc, manganese, and other trace metals.
phytoplankton Photosynthetic, usually single-celled, plants that drift with ocean currents.
pycnocline The layer in which density changes most rapidly with depth and separates the surface mixed layer from the deep ocean waters.
Southern Ocean The circumpolar ocean in the Southern Hemisphere between the Subtropical Front and the continent of Antarctica.
stratification The formation of distinct layers with different densities; stratification inhibits mixing.
subpolar Pertaining to the regions between the polar and temperate zones, but for the oceans the boundaries are the Subtropical Front and the Polar Front.
subtropical Pertaining to the regions which, under the influence of the trade winds, are permanently stratified.
upwelling Upward vertical movement of water through the bottom of the surface mixed layer produced by a divergence at the surface.
zooplankton Animals that float or drift with ocean currents: Microzooplankton are protozoan plankton that graze on small phytoplankton; mesozooplankton are crustaceans that graze on larger phytoplankton such as diatoms.

AN ECOSYSTEM IS A NATURAL UNIT in which physical and biological interactions cause the organized flow of energy, mass, and information to have characteristic trophic structure, material cycles, and a successional pattern. Ecosystems have some degree of internal homogeneity, objectively definable boundaries, and a predictable pattern of temporal development. Ocean ecosystems are those ecosystems that exist in the open ocean more or less independent of solid substrates.

1. HISTORY OF THE ECOSYSTEM CONCEPT

Natural scientists have long recognized that beyond populations or communities there is a higher level of...
organization, but it has been difficult for physical and biological scientists to reach consensus about its nature. The origins of the ecosystem concept are indistinct and, indeed, involved polemics between some of the principals. Ernst Haeckel (1834–1919) coined the word "ecology" and he envisioned ecology as a science that studies the environment as a stage within which selection pressures and adaptations affect the evolution of species. Haeckel brought the physical environment to the forefront, but he was interested primarily in evolution, which acts on species or genetically interacting populations. Victor Hensen (1835–1924) coined the word "plankton" to describe the organisms that drift in the ocean following the paths dictated by currents and mixing. Hensen was interested in the integrated working of natural systems, the subject today called "ecosystem science." He was particularly interested in the functional behavior of ocean systems from the smallest phytoplankton up through the food web and leading finally to fish, birds, and marine mammals. Hensen and Haeckel had a long professional dispute which seems ironic today because the eventual synthesis of their ideas by others led indirectly to the ecosystem concept.

The linguistic route to the modern word "ecosystem" was tortuous. From 1877 to 1939 the words "biocenos," "microcosm," "naturkomplex," "holocen," and "biorosystem" were all proposed as names for this idea. In 1935 an English scientist, Tansley, introduced the neologism "ecosystem" and this term rapidly eclipsed the competing names, but ecosystem science was not a well-organized endeavor until 1939, when Eugene Odum and Howard Odum published a groundbreaking textbook, *Fundamentals of Ecology*. There were many treatises on ecology before the Odums' book, but their book stressed the principle that the ecosystem (not the population or community) is the fundamental unit of ecology and indeed of biology. The 1939 text, which develops this concept in each of its chapters, has remained in print for many years. Although Eugene Odum's definition of ecosystem (given at the beginning of this article) is widely accepted, the principle of the ecosystem as the fundamental unit of ecology is controversial. Populations, communities, adaptation, and evolution have all been explosively successful research areas. Ecosystem research, because of its inherent complexity and requirement for interdisciplinary work, has not enjoyed the success of the more reductionist areas of ecological research. The study of ecosystems is usually supported by encouraging words, but our institutions and agencies, which are organized along disciplinary lines, have trouble coping with a science composed of equal parts of physics, chemistry, and biology. Ecosystem studies are placed with the life sciences, in which there is little institutional or agency enthusiasm for the complex and expensive atmospheric and physical oceanographic work required by the study of ocean ecosystems.

II. UTILITY OF THE CONCEPT

The ecosystem concept was originally developed for terrestrial, intertidal, and benthic habitats, in which ecological succession on timescales of 10–100 years is an obvious and salient characteristic of the ecosystem. Succession on this timescale is difficult or impossible to detect in the fluid medium of pelagic ocean ecosystems. In the ocean, winds, mixing, and currents appear to reset the successional time clock to time zero each annual cycle or, perhaps, with each storm or passage of a front or large eddy. It has been shown that there are low-frequency (over the timescale of 10–100 years) changes in species abundance and community structure. These biological changes are sometimes so far reaching that they are called "regime shifts," the term used by John Steele in 1998, but they are not functionally analogous to succession, which is (i) orderly and directional, (ii) involving biological modification of the physical environment, and (iii) resulting in a more stable climax community (Odum, 1969). Succession in ocean ecosystems is dramatically evident on the day to month timescale after spring stratification in temperate and subpolar waters or after an episode of upwelling. This short timescale succession, which appears to be cyclic, is not unidirectional and definitely not orderly. If long-term directional succession does not appear to occur in pelagic ocean ecosystems and succession is a central tenet of ecosystem theory, isn't the very existence of ocean ecosystems in doubt? The answer is that ocean ecosystems, as Steele (1985) noted, are clearly different from terrestrial, aquatic, intertidal, or benthic ecosystems, but ocean ecosystems meet most of the requirements of the definition set forth by Eugene Odum. The ocean ecosystems described here all have a characteristic and distinct trophic structure, characteristic and distinct maternal cycles, some degree of internal homogeneity and commonality, and definable hydrographic boundaries.

There is considerable heuristic power in the ecosystem concept because understanding gained in one ocean ecosystem can be used to predict the response of another ecosystem of the same kind that is geographically distinct from it. This predictive power is perhaps the
greatest benefit of ecosystem theory and provides evidence that each distinct kind of ocean ecosystem has characteristics that can be generalized and used in prediction.

Pomeroy and Alberts (1988) emphasized that the concept involves emergence of new properties. One consequence of the hierarchic organization of ecosystems is that as components (both biotic and abiotic) are integrated into a larger functional unit, new properties emerge that cannot be detected by study of the component populations or processes, no matter how thorough the reductionist study. This aspect of ecosystem theory makes the concept useful, even necessary, for predictive understanding of ocean ecosystems.

Determining specific emergent properties of a system as large as ocean ecosystems is difficult. Ocean ecosystem spatial domains of thousands of kilometers are difficult to sample adequately with ships, however, with satellites that measure wind, ocean temperature, ocean currents (from sea surface topography), and phytoplankton biomass, a new era has begun in understanding ocean ecosystems. Are there appropriate benefits to justify the large societal investment in remote sensing and data handling required to achieve this new level of understanding? One benefit deals with fisheries management. The historic approach to management of these fisheries has involved analysis of local populations and environment. It has become clear that ecosystem properties, especially physical conditions, operating on scales much larger than the range of the exploited population, may be responsible for changes in reproductive success and adult abundance. In this context, a valuable societal payoff of understanding ocean ecosystems is improved ability to predict local variations in living resources as shown by Sherman and others in 1986.

Odum repeatedly emphasized that ecosystem science and economics are parallel disciplines and expressed regret that they are not perceived as such, particularly by the economic community. Investors and political leaders making policy decisions on resource development need an understanding of the probability, frequency, and intensity of natural variability for realistic economic decision making. In this context, economics and the ecosystem concept are closely related: By understanding the variability of ocean ecosystems, decision makers will also understand that variation in marine resources is a normal and inevitable characteristic that must be accommodated in economic plans. This kind of resource-related benefit, although important, is not the only societal benefit that will accrue.

Odum (1977) said, ‘It is the properties of the large-scale integrated systems that hold solutions to most of the long-range problems of society.” Although few in the scientific community recognized the wisdom of this comment two decades ago, the validity of Odum’s prediction has now been well demonstrated. For example, carbon dioxide modification of the planet’s heat budget, and therefore climate, is a phenomenon that can be understood only if the emergent properties of large-scale biogeochemical systems are understood.

III. PARTITIONING THE OCEAN INTO NATURAL FUNCTIONING UNITS

A. Central Problem

The central problem for the lower trophic level of ocean ecosystems is obtaining light (energy) and inorganic nutrients (mass). Odum’s definition requires that for a functioning system to be a distinct ecosystem it must possess characteristic trophic structure and material cycles. That is, how one kind of ocean ecosystem captures light and passes that energy on in the form of primary productivity, secondary productivity, and so forth is different from how another kind of ocean ecosystem processes and transfers its energy. Likewise, how mass (C, N, P, and Si), initially in the form of inorganic compounds, is taken up and transferred through the food web and eventually released back to the environment is different and very poorly understood.

What controls the supply of light and nutrients to an ocean ecosystem? Sverdrup in 1955 was the first oceanographer to note that the spatial and temporal patterns of physical processes, particularly the seasonal patterns of mixing, stratification, and upwelling as well as the seasonal pattern of irradiance, control the patterns of physical processes, particularly the seasonal patterns of mixing, stratification, and upwelling as well as the seasonal pattern of irradiance, control the patterns of biological organization. The division of ocean ecosystems into six distinct types is based fundamentally on the pioneering work of Sverdrup and that of others in the intervening years.

B. Biome Concept

Longhurst (1998) presented a scholarly discussion of the attempt to partition the oceans into natural functional provinces. He described with considerable elan a regional ecology of the oceans which is clearly distinct from partitioning the oceans into ecosystems. The difference can be illustrated using one of the best described ocean ecosystems, the coastal upwelling ecosystem. It is well accepted that there is one coastal upwelling
ecosystem and that it is replicated at five coastal regions in the world ocean: the California Current off California and Mexico, the Peru Current off Peru and Chile, the Benguela Current, the Canary Current, and the Arabian Sea off Oman and Yemen. Longhurst recognized the same regions but counted them as five distinct ocean “provinces.” The ecosystem concept emphasizes that there are geographic ocean regions that have much in common; ecological geography emphasizes the discreteness or autonomy of various ocean provinces. Both concepts are useful.

Longhurst (1998) proposed that the following information is required to define the functional nature of an ocean region:

- **Latitude**, which determines if wind stress induces mixing or wind-driven advection
- **Depth of water** because stratification may be broken down in shallow seas by tidal mixing
- **Proximity to coastline**, which determines the effects of terrestrial runoff, river discharge, and release of nutrients (especially Fe) from sediments
- **Seasonal irradiance**, which forces photosynthesis, stratification, and freezing or melting of ice
- **Winds**, which force mixing or upwelling of subsurface waters and their nutrients up to the euphotic zone
- **Precipitation**, which may induce strong stratification by making the surface layer less salty
- **Nutricline depth**, which modifies the vertical flux of nutrients by wind mixing and upwelling
- **Strength of the vertical nutrient gradient**, which determines the magnitude of the upward nutrient flux
- **External source of iron**, because insufficient iron may limit uptake of macronutrients, phytoplankton growth, or primary productivity

Using these criteria, Longhurst partitioned the ocean into four major biomes: the westerlies biome, in which convective mixing and stratification are forced largely by the strong seasonal progression of winds, irradiance, and heat flux, the trades biome, in which upwelling and mixing are forced on the ocean-basin scale by both local and remote wind forcing, the polar biome, in which there is no significant thermal stratification and mixed layer depth is constrained by a surface low-salinity layer which forms each summer as the marginal ice melts, and the coastal biome, in which coastal processes such as tides or currents break down stratification and force mixing.

Longhurst’s (1998) partitioning of the ocean into four biomes and 35 provinces is a significant accomplishment. Eventually, his ideas will be merged with the ecosystem concept to produce an internally consistent hierarchy of biomes, ecosystems, and provinces.

### C. Ecosystem Concept

Using the formal Odum definition of ecosystem, six or less well-defined ocean ecosystems can be delineated (Table I), but note that the boundaries of these ecosystems are usually oceanographic, not geographic, features. [See Tomczak and Godfrey (1994) for a description of these oceanographic features.] Although it is possible to describe approximately where these ecosystems exist, the actual domain is determined by dynamic processes such as fronts where two kinds of ocean water converge. The approach is to define the characteristic trophic structure and material cycles by applying the same analysis to each system. Each of the six systems has a different combination of stratification, nutrient supply (Table I), primary productivity (Table II), and biotic characteristics (Table III). This analysis indicates that there are six distinct ocean systems that meet Odum's criteria:

1. A low-latitude gyre ecosystem is present in each of the five great low-latitude gyres—North Atlantic, South Atlantic, North Pacific, South Pacific, and South Indian Oceans—as well as in the warm pool of the western Pacific Ocean, the equatorial Indian Ocean, and in large marginal seas such as the Mediterranean Sea and the Gulf of Mexico. This ecosystem is also present in the Western Boundary Current regions between the western edge of each of the five great gyres and the coastal waters of the adjacent continent.

2. The Southern Ocean ecosystem occupies the circumpolar area between the continent of Antarctica and the Subtropical Front at approximately 40°S latitude; the Southern Ocean ecosystem has a subantarctic region from the Subtropical Front south to the Polar Front and an Antarctic region from the Polar Front to the Antarctic continent.

3. The equatorial upwelling ecosystem occupies an equatorial band from 3°N to 3°S and from South America westward to 180° in the eastern and central Pacific Ocean. This region is often called the “cold tongue” because upwelling keeps these tropical waters surprisingly cool. In the Atlantic the 5°N to 5°S band of equatorial upwelling reaches from Africa across to South America. In the Indian Ocean equatorial region there is no manifestation of this ecosystem because the upwelling, if present, does not extend upwards into the euphotic zone.
### TABLE 1
Summary of Heat Flux, Stratification, Nutrient, and Light Characteristics of Ocean Ecosystems

<table>
<thead>
<tr>
<th>Ocean ecosystem</th>
<th>Heat flux</th>
<th>Stratification</th>
<th>Nutrient</th>
<th>Light</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strength</td>
<td>Duration</td>
<td>Level</td>
<td>Source</td>
</tr>
<tr>
<td>Low-latitude gyre</td>
<td>Neutral, negative in Western Boundary Current</td>
<td>Strong</td>
<td>Permanent</td>
<td>Low (&lt;&lt;$K_s$)$^a$</td>
</tr>
<tr>
<td>Southern Ocean</td>
<td>Negative (ocean loses heat)</td>
<td>Very weak, except strong when ice melts in summer</td>
<td>Seasonal</td>
<td>High ($&gt;K_s$), except $K_s$</td>
</tr>
<tr>
<td>Equatorial upwelling</td>
<td>Positive (ocean gains heat)</td>
<td>Strong stratification following vertical transport</td>
<td>Permanent</td>
<td>High ($&gt;K_s$), except $K_s$</td>
</tr>
<tr>
<td>Subarctic gyre</td>
<td>Seasonally positive and negative</td>
<td>Moderate stratification following winter mixing</td>
<td>Seasonal convective mixing</td>
<td>High ($&gt;K_s$, in winter; $K_s$, rest of year)</td>
</tr>
<tr>
<td>Eastern Boundary Current</td>
<td>Positive</td>
<td>Medium</td>
<td>Permanent</td>
<td>Medium ($&gt;K_s$)</td>
</tr>
<tr>
<td>Coastal upwelling</td>
<td>Positive</td>
<td>Strong stratification following vertical transport</td>
<td>Continuous</td>
<td>High (&gt;&gt;$K_s$)</td>
</tr>
</tbody>
</table>

$^a$ See text for the location and boundaries of these ecosystems.

$^b$ $K_s$ is the nutrient concentration at which nutrient uptake occurs at one-half the maximal rate; at concentrations <<$K_s$, uptake is limited.

$^c$ $E_k$ is the light level at which the photosynthetic rate is light saturated; at light levels <<$E_k$, photosynthesis is light limited.
### TABLE II
Summary of Size, Primary Productivity, Export, and Limiting Factors of Ocean Ecosystems

<table>
<thead>
<tr>
<th>Ocean ecosystem</th>
<th>Size**</th>
<th>Primary productivity amount and pattern (mmol C m$^{-2}$ day$^{-1}$)</th>
<th>Export ratio</th>
<th>Factors limiting primary productivity</th>
<th>Factors limiting fish yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-latitude gyre</td>
<td>164 $10^{12}$</td>
<td>35 continuously Low, weak seasonality</td>
<td>Macronutrients</td>
<td>Low primary productivity, small size of organisms and long food web</td>
<td></td>
</tr>
<tr>
<td>Southern Ocean</td>
<td>77 $10^{12}$</td>
<td>120 summer mean; $\approx$30 in spring and fall; $\approx$0 in winter</td>
<td>High in summer; very low in winter</td>
<td>Iron and light</td>
<td>Short duration of summer bloom</td>
</tr>
<tr>
<td>Equatorial upwelling</td>
<td>22 $10^{12}$</td>
<td>100 continuously Low</td>
<td>Episodically high; moderate rest of year</td>
<td>Iron and silicate</td>
<td>Low export productivity</td>
</tr>
<tr>
<td>Subarctic gyre</td>
<td>22 $10^{12}$</td>
<td>150 spring bloom mean; $\approx$50 rest of year</td>
<td>High in summer; moderate rest of year</td>
<td>Unclear, but iron is a factor; and light is also in winter</td>
<td>Unclear</td>
</tr>
<tr>
<td>Eastern Boundary Current</td>
<td>21 $10^{12}$</td>
<td>150 summer mean; 75 rest of year; grades into gyre $\approx$35</td>
<td>High, but spatially variable</td>
<td>Iron</td>
<td>Small size of ecosystem</td>
</tr>
<tr>
<td>Coastal upwelling</td>
<td>6 $10^{12}$</td>
<td>300 continuously close to coast; grades into Eastern Boundary Current $\approx$130</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Size was calculated for the world ocean exclusive of noncoastal upwelling, continental shelf regions; total area was $312 \times 10^{12}$ m$^2$.

**Export ratio is the relative proportion of primary productivity that is exported vertically, horizontally, or to higher trophic levels. The maximum observed export ratios relative to total primary productivity are approximately 0.50.
TABLE III
Summary of Biotic Characteristics, the Zooplankton Component of Ocean Ecosystems

<table>
<thead>
<tr>
<th>Ecosystem</th>
<th>Endemic spp</th>
<th>Species richness</th>
<th>Species evenness</th>
<th>Variability of biomass</th>
<th>Size of organisms</th>
<th>Importance of protistan micrograzers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-latitude gyre</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Small</td>
<td>Dominant</td>
</tr>
<tr>
<td>Southern Ocean</td>
<td>High</td>
<td>Low</td>
<td>Very low</td>
<td>Very high (&gt;3×)</td>
<td>Large</td>
<td>Moderate to low</td>
</tr>
<tr>
<td>Equatorial upwelling</td>
<td>Moderate</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Small</td>
<td>Dominant</td>
</tr>
<tr>
<td>Subarctic gyre</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>High (&gt;2×)</td>
<td>Moderate</td>
<td>Moderate to low</td>
</tr>
<tr>
<td>Eastern Boundary Current</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Low (&lt;2×)</td>
<td>Moderate</td>
<td>Dominant to low</td>
</tr>
<tr>
<td>Coastal upwelling</td>
<td>Low</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate (=2×)</td>
<td>Large</td>
<td>Low</td>
</tr>
</tbody>
</table>

*Adapted from McGowan (1974).

4. A subarctic gyre ecosystem is present in both the North Atlantic and the North Pacific in the region north of the Subtropical Front at approximately 40°N. In the Atlantic, the subarctic gyre ecosystem is bounded on the south by the North Atlantic Current and Subtropical Front, on the west by the Labrador Current, on the north by the East Greenland Current, and on the east by Europe. The North Atlantic subarctic gyre is a complex, but very well-delineated, ocean feature. The North Pacific subpolar gyre is bounded on the south at approximately 40°N by the North Pacific Current and the Subtropical Front, on the west by Siberia, on the north by the Aleutian Islands and Alaska, and on the east by Canada.

5. An Eastern Boundary Current (EBC) ecosystem is represented in each of the four great EBCs: the California Current off the west coast of the United States and Mexico, the Peru Current off Peru and Chile, the Canary Current off Northwest Africa, and the Benguela Current off Namibia and South Africa. These locations are characterized by a broad and weak equatorward-flowing current that usually extends offshore 400–600 km. Each of these ecosystems receives upwelled water from a coastal upwelling ecosystem and, in turn, exports water to the adjacent low-latitude ecosystem.

6. There are four great coastal upwelling ecosystems: Peru Current, Benguela Current, Canary Current, and California Current. This ecosystem is narrow but long; it extends for great distances along the west coasts of South America, North America, northwest Africa, and South Africa. This ecosystem is also present in the northeastern Arabian Sea, along the coasts of Oman and Yemen, where evidence of upwelled water is present up to 600 km off the Omani coast.

IV. CHARACTERISTICS OF OCEAN ECOSYSTEMS

A. Shared Characteristics

The following characteristics of ocean ecosystems are based on a 1974 synthesis by McGowan for the oceanic Pacific Ocean:

*They are few in number (only six). This small number of ocean ecosystems contrasts dramatically with the much larger number of well-defined terrestrial, intertidal, and coastal habitats that occur in much smaller areas. The reason, of course, is that the fluid medium of the open ocean is fundamentally partitioned by the dominant physical processes that occur in an oceanic area, and there are only a few (six) of these overriding physical patterns. In contrast, terrestrial, intertidal, or coastal ecosystems are fundamentally partitioned by spatial or geographic features and there is a much larger number of unique geographic features associated with the solid surface of the earth’s crust. The same fundamental concept pervades all aspects of the comparison of oceanic biodiversity with that of terrestrial, intertidal, or coastal biodiversity. The fluid medium of the ocean is relatively homogeneous, and the boundaries that do exist are dynamic processes such as the presence or absence of winter convective mixing. It is obvious that dynamic processes of this nature form physical boundaries that permit much more biological exchange than the physical barrier associated with the solid earth.*

*They are large relative to terrestrial, intertidal, or coastal ecosystems. The largest ocean ecosystem, the low-latitude gyre ecosystem, occupies 52% of the open*
The coastal upwelling ecosystem, however, may have disappeared entirely during periods of lower sea level. The coastal upwelling ecosystem, has received thorough, comprehensive, and multinational study during the past three decades. This ecosystem is small relative to other ocean systems; it has very strong physical, chemical, and biological processes; it has huge economic importance and considerable geological importance because much of the earth's oil is assumed to have been formed 18,000 years ago. Oceanic cores suggest that the defining oceanographic features have been in place for the past 200,000 years and perhaps for the past 1 million years. The coastal upwelling ecosystem, however, may have disappeared entirely during periods of lower sea level such as during the last Glacial Maximum approximately 180,000 years ago.

They respond to climate but not to weather. Subtle and pervasive changes in mixed layer depth, the depth of convective mixing, or the strength of stratification cause profound changes in ocean ecosystems. On the other hand, strong storms or the passage of a violent hurricane cause no change that is detectable in the ocean ecosystem 2 weeks after the event. El Niño events cause profound changes in equatorial upwelling, EBC, and coastal upwelling ecosystems, but when El Niño ends the lower trophic levels of these ecosystems (phytoplankton, bacteria, microzooplankton, and mesozooplankton) return to their pre-El Niño condition within 1 month. Higher trophic levels, of course, require several years to recover.

They have considerable internal homogeneity, i.e., they tend to be monotonous. The key word is “internal.” There are clear changes in almost all biological properties when physical or oceanographic boundaries are crossed; however, if oceanographic boundaries are not crossed, biological properties will remain surprisingly similar over distances of thousands of kilometers. They are relatively undisturbed by anthropogenic processes compared to other ecosystems. The largest human disturbance thus far has been ruthless overfishing that removes top predators. The removal of these long-lived and slow-growing carnivores appears to set off a trophic cascade that changes the ecosystem even down to the level of nutrient regeneration by bacteria and protozoa. A second and more threatening change is that as the earth's heat budget changes the processes of precipitation, winter mixing, stratification, and depth of the nutricline all change in such a way that there is less transport of nutrients into the euphotic zone. This change, already apparent in the North Pacific Ocean, will reduce productivity and, hence, the yield of food resources.

The basic organization (Fig. 1) is like that of other ecosystems. It appears that the basic assembly rules for all ecosystems are very similar. There has to be a large number of primary producers, somewhat fewer herbivores, still fewer carnivores, and many fewer top predators. There have to be efficient recyclers to return a portion of the nutrients back to the euphotic zone. In addition, the ecological “rules” regarding diversity, species composition, and adaptation are expressed similarly in terrestrial and ocean ecosystems.

B. Distinguishing Characteristics

There are qualitative differences in the basic biological processes. The most dramatic illustration of this is the observation that processes limiting productivity, biomass, export, and yield are different in almost every one of these six ecosystems. Although our understanding of limiting factors may change, currently it appears that the low-latitude gyres are limited by fixed nitrogen, especially nitrate and ammonia; the Southern Ocean appears to be limited by iron and light; the equatorial upwelling ecosystem appears to be limited by iron and silicon; coastal upwelling ecosystem rates appear not to be limited by any macronutrient, but the space where optimal coastal upwelling occurs is highly constrained and iron supply is involved; the EBC ecosystem appears to be limited by light and iron in winter and by iron...
FIGURE 1  Generalized organization of ocean ecosystems. Physical forcing and input functions are shown in dark gray, chemical and geologic input functions are white, and biological components are in light gray. Physical transfers of energy, momentum, or mass are shown by arrows with solid lines; biological transfers of mass and energy between living components are shown by arrows with dotted lines. The export process at the extreme right, labeled fishing, is a proxy for all processes that remove biomass from the euphotic zone of the represented ecosystem.
alone in the summer; and the subarctic gyre ecosystem appears to be limited by the depth of the winter mixed layer or convection.

There are quantitative differences in basic processes such as primary productivity among different kinds of ocean ecosystems. One of the important milestones in the study and understanding of these differences was the work of John Ryther, who in 1969 published a work that provided a quantitative explanation of why fish yields vary by approximately 200-fold from the richest ocean ecosystems to the poorest. Variations in productivity, of course, are well-known from terrestrial ecosystems, but on land either aridity in deserts or freezing in polar regions is responsible for the low productivity of the poorest regions. Understanding why the benign low-latitude gyre ecosystem was so poor in fish production was much more difficult than understanding why productivity was low in deserts and polar regions. Part of the explanation was proposed in 1955 by Sverdrup, who said simply that the physical supply of nutrients to the euphotic zone is the reason for low fish yields in stratified ocean ecosystems. Ryther's contribution amplified the physical explanation by considering the nature of the ecological processes that lead to fish production.

First, Ryther estimated that approximately half the fish caught in the world are caught in coastal upwelling ecosystems, the smallest of the ocean ecosystems. Why? To begin, Sverdrup was correct: The physical processes of upwelling bring abundant nutrients to the surface layer, so primary productivity is very high in upwelling ecosystems. However, much more is involved. The phytoplankton that thrive in the rich coastal upwelling ecosystems are very large—so large that some are eaten directly by fish. This means that in coastal upwelling ecosystems the food chain leading to fish is very short. Ryther estimated that half the fish diet was phytoplankton and half was zooplankton. On average, then, the length of the food chain leading to fish had 1.5 transfers: large phytoplankton to fish or large phytoplankton to zooplankton to fish. At each ecological transfer, a large portion of the energy of the food is used to support the organism and this portion cannot be passed up the food chain. The shortness of the food chain is a factor that multiplies the high nutrient effect. Ryther also noted that in the small and food-rich coastal upwelling ecosystem the fish and zooplankton did not have to work as hard to get food, so the efficiency of transfer was increased relative to that of a poor environment such as the low-latitude gyre. Ryther proposed that fish yields were high in the coastal upwelling ecosystem because of high nutrients, high primary productivity, large size of the primary producers, short food chains with few transfers, and increased efficiency of transfer. These effects multiply each other, leading to very high yields of fish. For the same reason, the abundance of seabirds and marine mammals is also very high in the coastal upwelling ecosystem.

The same arguments in reverse explain the low fish yields of the low-latitude gyre ecosystem. The other four ocean ecosystems range between these two extremes. The order of fish yield per unit area can be approximately estimated, from highest to lowest, as (1) coastal upwelling ecosystem, (2) subarctic gyres ecosystem, (3) EBC ecosystem, (4) Southern Ocean ecosystem, (5) equatorial upwelling ecosystem, and (6) low-latitude gyre ecosystem.

V. OCEAN ECOSYSTEMS AND GLOBAL CHANGE

Human intervention in the material cycles and trophic structure of ocean ecosystems may already have caused some changes. The interventions are so varied and play out at so many different scales that it is difficult to know how to describe their impact on future climate change. One approach is to focus on model studies of how anthropogenic ocean warming will affect ocean ecosystems. Considerable effort has gone into investigating how increased atmospheric carbon dioxide will affect the physical ocean-atmosphere system. Other effort has gone into determining how the ocean's carbon dioxide system will behave in an anthropogenically warmed ocean. Today's atmosphere-ocean climate models are fully coupled to an ocean carbon model, including a full carbon system into investigating how increased atmospheric carbon dioxide will affect ocean ecosystems. Other effort has gone into investigating how increased atmospheric carbon dioxide will affect ocean ecosystems. Considerable effort has gone into investigating how increased atmospheric carbon dioxide will affect ocean ecosystems. Today's atmosphere-ocean climate models are fully coupled to an ocean carbon model, including a full carbon system into investigating how increased atmospheric carbon dioxide will affect ocean ecosystems. Other effort has gone into determining how the ocean's carbon dioxide system will behave in an anthropogenically warmed ocean. Today's atmosphere-ocean climate models are fully coupled to an ocean carbon model, including a full carbon system into investigating how increased atmospheric carbon dioxide will affect ocean ecosystems.
have the major impact on ocean ecosystems. The changes predicted vary in intensity from one ocean ecosystem to another, but they are most severe in productive ecosystems such as the subantarctic gyre, equatorial upwelling, and coastal upwelling ecosystems. Very little change in nutrient supply is predicted for the world ocean’s five great low-latitude gyres. The predicted change is an expansion of the size of the world ocean’s most oligotrophic ecosystems: Approximately half of the ocean area will have decreased nutrient supply; the low-latitude gyres, the poorest half, will have no change; and a very small area (less than 9%) in the high Arctic and Antarctic will have an increase in nutrient flux to the surface.

The El Niño phenomenon has provided evidence of the biological consequences of reducing an ocean ecosystem nutrient supply. New populations that became established under the low nutrient conditions are healthy, highly diverse communities, but they are dramatically different. Another consequence of reduced nutrients is that the biological pump sequesters less carbon dioxide. This change will feed back into the global carbon system to accelerate the increase in concentration of atmospheric carbon dioxide. A reduction of new nutrients by 37% will have a strong impact on both the quantity and the kind of fish present. Because the increase will occur in what are now rich fishing regions, the coastal upwelling and subarctic gyre, the societal impact will be larger than the biological impact. A change of 37% in new nutrients is significant, but it is not enough to destroy ocean ecosystems. The current gradient in nutrient flux from the low-latitude gyre ecosystem to a coastal upwelling ecosystem region is more than a factor of 10. The predicted changes will reduce the yield from the world’s rich fishing banks, but the expanded low-latitude gyre ecosystems should maintain their ecological integrity.

See Also the Following Articles

CLIMATE CHANGE AND ECOLOGY, SYNERGISM OF
ECOSYSTEM, CONCEPT OF
MARINE ECOSYSTEMS
PELAGIC ECOSYSTEMS
PLANKTON, STATUS AND ROLE OF

Bibliography

ORIGIN OF LIFE,
THEORIES OF

Susanne Brakmann
Evotec Biosystems AG

I. Historical Overview
II. Theoretical Concepts

GLOSSARY

autocatalysis Self-acceleration of certain chemical reactions that yield catalytically active products. The reaction rate of autocatalytic processes increases exponentially.

hypercycle Cyclic sequence of self-supporting reactions between primitive prebiotic biomolecules, nucleic acids, and proteins. Hypercyclic organization of reacting systems was postulated to explain the spontaneous emergence of replicating systems. As a result of the hypercyclic reaction principle, functional characteristics of replicating molecules (i.e., phenotypes) are connected to their hereditary characteristics or their genotypes. Consequently, both phenotypes and appertaining genotypes are evaluated and selected in feedback loops.

natural selection Preferential reproduction and survival of certain species under given environmental conditions. Growth advantages of selected individuals reflect the efficiency of their reproduction pathways.

quasispecies Hierarchically ordered population of mutants that results from erroneous copying of a genotypic ancestor. Whole mutant distributions behave like single species. They are subjected to natural selection. The average sequence is defined as a consensus sequence that represents the “center of gravity” of the distribution. The consensus sequence is synonymous to a wild type.

self-organization Spontaneous formation of complex structural aggregates from biological macromolecules such as proteins, nucleic acids, or lipids. The assembly pathways are usually determined by physicochemical properties of the constituent molecules.

LIFE IS A STATE OF MATTER that results from controlled cooperation of highly complex structural assemblies. Living matter shows more than the sum of properties of the isolate entities since none of them could survive separately. Many general features distinguish living from non-living matter, predominantly self-reproduction, functional self-organization, and evolution by natural selection. The origin of life is a consequence of prebiotic evolution of organic material. The process of prebiotic evolution is the object of various and contradictory hypotheses, none of which has been proven experimentally.

I. HISTORICAL OVERVIEW

A. What Is Life?

Currently, life appears on Earth with an immense variety of structures and functions. Despite this diversity,
some invariant features exist that are common to living systems and distinguish them from non-living matter. All life takes shape with defined individuals. Cells are the smallest individual entities capable of performing all processes for living and reproducing. They represent the elementary structural units of life. Cells do not spontaneously originate from suitable mixtures of non-living matter. Louis Pasteur (1822–1895) was the first to disprove the idea of abiogenesis by showing that each cell results from cell division. Rudolf Virchow (1821–1902) confirmed this, noting that fusion of cellular precursors (germ cells) is the only alternative for generating new cells. Two general types of cellular organization are known: the prokaryotic type, which lacks a nucleus, and the more highly developed eukaryotic type, which harbors a nucleus. Basically, both cell types exhibit the same structure and functional organization suggesting a common origin. Simple organisms such as bacteria consist of single cells which are either prokaryotic or eukaryotic, whereas complex organisms, such as plants and animals, are highly organized multicellular assemblies of the eukaryotic type.

The different forms of life do not originate from nothing as Charles Darwin (1809–1882) discerned in his epoch-making work. He, and independently Alfred Russel Wallace (1823–1913), postulated that, instead, a continuous process of successive adaptation to environmental conditions led to the origin and development of species. Darwin also realized that evolutionary progress results from mutagenous reproduction and natural selection. Today, this theory can be explained with twentieth-century insights into structural and chemical characteristics of living cells.

Each single cell contains an extraordinarily efficient machinery for respiration and production of energy; for absorption and assimilation of nourishment; for synthesis, storage, utilization, and transport of new products; and for production of more or less identical copies of itself. The cellular machinery is instructed by hereditary information that is carried by large molecules which are composed of nucleic acid and known as genes. They comprise the architect’s plan of all cellular components and also include structural elements regulating the formation of the different characteristics of an organism. During the process of reproduction, genes also reproduce, thereby passing all instructions on to the next generation. Proteins represent the executive that is coded by genetic information. They cooperate closely with nucleic acids in order to fulfill the diverse functional and structural tasks that maintain the cell machinery. Many proteins are enzymes that act as catalysts for the reproduction of nucleic acids and for the synthesis of essential organic compounds, such as carbohydrates, amino acids, lipids, and hormones. Other proteins care for the cellular metabolism and the production of energy. The teamwork of these two molecular species, polynucleotides and polypeptides, predominantly characterizes the chemistry of living matter as we know it.

Generally, high-energy macromolecular compounds play a major role in living organisms. In addition to nucleic acids and proteins, biomolecules such as lipids, polysaccharides, and phosphates are found. They serve for demarcation of individual cells, for their division into different subcompartments, for their structural stability, and for energy transfer processes. However, the splendid variety of life on Earth is mainly due to the heteropolymeric nature of proteins. They consist of long chains that are built from 20 constituents, chemically defined as amino acids. Similarly, nucleic acids encoding the natural proteins are polymeric chains based on 4 different monomers. The building principle realized with both molecular species allows a variability that is far beyond the imagination—even with chain lengths of 20 monomers, more than $10^{12}$ different nucleic acids (with 4 different building blocks) or $10^4$ alternative proteins (with 20 different building blocks) can exist.

Structural variability is not the only feature that characterizes the molecules of life. Nucleic acids are also capable of serving as templates for their own replication, thereby guaranteeing the reproduction of living organisms. Self-reproduction is an inherent property of polymeric substances. The chemistry of living matter dominantly characterizes the chemistry of living matter as we know it.
ward higher fitness and often toward greater complexity. In summary, complex organisms evolved through time according to Darwin's theory because of replication, mutagenous replication, natural selection, and replication of resulting mutants. During the process of Darwinian evolution, information is created continually.

Another viewpoint of "life" is presented by the discipline of thermodynamics that distinguishes between open and closed systems. Corresponding to the second law of thermodynamics, no processes can occur in closed systems that increase the net order of the system or that decrease their entropy. In this perspective, any living organism is an open system that exchanges energy and matter with its surrounding, thereby establishing and maintaining a highly ordered structural assembly at the expense of a larger decrease of order of the universe outside. Living organisms participate in ecosystems and depend on resources found in their environment. In particular, most organisms are dependent on the flow of sunlight, which is absorbed by plants and utilized by them to synthesize high-energy molecules from simpler ones. These products serve as a source of food and energy for other links of the food chain (e.g., animals). By excretion and degradation of dead organic compounds, all material will be recycled without altering the organisms' state of life, which is characterized as being far from thermodynamic equilibrium. A state of thermodynamic equilibrium would be identical to a state of maximum entropy and thus correspond to a thermodynamic equilibrium.

B. Hypotheses and Theories on the Origin of Life

The traditional position of theology and some philosophy views the origin of life as the result of a supernatural event which is permanently beyond the descriptive powers of chemistry and physics. In its most general form, this view is not necessarily contradictory to contemporary scientific knowledge about prebiotic evolution, although the biblical descriptions of creation given in the first two chapters of Genesis, taken literally and not metaphorically, are inconsistent with modern knowledge.

Until the mid-nineteenth century, the prevailing opinion was that God created man together with higher animals and plants, but that simple forms of life such as worms and insects arise steadily from mud, wase, and putrefied matter during short periods of time. The physiologist William Harvey (1578–1657), who studied reproduction and development of deer, was the first to challenge this view by postulating that every animal comes from an egg ("omnia viva ex ovo") a long time before Karl-Ernst von Baer (1792–1876) discovered the existence of human egg cells by microscopy. An Italian scientist, Francesco Redi (1626–1698), found Harvey's idea to be true, at least for insects; he found that maggots in meat arise from fly eggs. Later, Lazzaro Spallanzani (1729–1799) discovered that spore-containing zoospores were necessary for the reproduction of mammals. Before Pasteur, Spallanzani also showed that living matter ("infusories") does not originate from boiled fluids kept in closed containers. Although Redi's and Spallanzani's findings definitely proved that insects and larger animals develop from eggs, it remained obvious to a large majority that at least microorganisms, because of their ubiquity, are generated continually from inorganic material. The debate of whether life is spontaneously generated from non-living matter or not culminated in the famous controversy between Louis Pasteur and Félix-Archimède Pouchet (1800–1872) which Pasteur won triumphantly. He showed that even microorganisms in fluids come from germs floating in the air, and he also demonstrated that nutrient solutions could be guarded against these creatures by suitable sterilization such as filtration or boiling. However, contemporary scientists were not satisfied by Pasteur's experiments because a delicate question remained: If living organisms do not arise from non-living matter, how had life come about in the first place?

In the late nineteenth century, another hypothesis was initiated by the Swedish chemist Svante Arrhenius (1859–1927). He strongly believed that the whole universe is replenished with living germs, a phenomenon that he called "pansperrma." He suggested that microorganisms and spores of cosmic origin spread from solar system to solar system, and thus they arrived on Earth. Although Arrhenius' view avoids rather than solves the problem of the origin of life, and despite the extreme unlikeliness of microorganisms surviving the interstellar effects of cold, vacuum, and radiation, a few twentieth-century members of the scientific community returned to the idea of panspermia. Among these scientists are astronomer Fred Hoyle (1915--) and mo-
l e cular biologist Francis Crick (1916–), who are convinced that the time span between the origin of Earth and the appearance of first cellular organisms on this planet was too short for life to have occurred spontaneously.

Darwin's theory of "natural selection as motive power for evolution" resulted in a new view on the phenomenon of life that is still valid. Although Darwin did not commit himself on the origin of life, contemporary scientists such as Thomas Huxley (1825–1895) extended his idea, asserting that life could be generated from inorganic chemicals. Pursuing this opinion, Alexander Oparin (1894–1980) was the most influential advocate of the successive origin of cellular organisms from non-living matter. He suspected this transition was proceeded by a series of regular and progressive chemical reactions under the physical and chemical conditions on early Earth. Together with John Scott Haldane (1860–1936), Oparin recognized that the biological production of organic molecules in the current oxidizing atmosphere of Earth is highly unlikely. Instead, both suggested that the beginning of life occurred in primordial hot waters under more reducing (i.e., hydrogen-rich) conditions. Furthermore, Oparin postulated the existence of pre-cellular coacervates—globular units with membrane-like surface structures—that may have high concentrations of certain chemical compounds. Coacervates indeed form spontaneously from colloidal aqueous solutions of two or more macro-molecular compounds.

However, many fundamental problems on the transition from non-living to living matter remained unsolved. The central question concerned the role of the second law of thermodynamics, which defines the equilibrium in an isolated system as a state of maximum entropy that appears to contradict the origin and existence of highly ordered living organisms. Erwin Schrödinger (1887–1961) gave a decisive answer to this question, stating that "living matter evades the decay to equilibrium" or death by steadily compensating for the production of entropy. In any organism, this is achieved by feeding it free energy or energy-rich matter which is used by the cellular machinery to drive essential chemical reactions. Schrödinger and others also realized that living organisms can thermodynamically be described as open systems, but they could not explain the general physical conditions for self-ordering processes. These were perceived by Ilja Prigogine (1917–) and Paul Glandorff (1904–1999), who worked on a thermodynamic theory of irreversible processes. According to Prigogine, selection and evolution cannot occur in equilibrated or nearly equilibrated reaction systems, even if the right types of substances are present. Instead, certain combinations of autocatalytic reactions with transport processes may lead to peculiar spatial distributions of reaction partners, called "dissipative structures." These ordered structures are of importance for the formation of functional order in the evolution of life, especially for early morphogenesis. However, the first steps of self-organization probably involved little organization in physical space but extensive functional ordering of a tremendously complex variety of chemical compounds. Manfred Eigen (1927–) explained the process of ordering among molecules by augmenting the Prigogine–Glandorff principle with phenomenological considerations on the behavior of self-replicating molecules: A certain quantity approaches a maximal value in any open system that is replicating autocatalytically with sufficient fidelity, and thereby continually consuming energy and matter. This quantity is called "information" and is closely related to the "negative entropy" postulated by Schrödinger. In addition to setting the stage for a molecular interpretation of biological information, Eigen developed the mathematical models for describing "selection." According to Eigen's theory, selection is the fundamental natural principle that brings order into any random arrangement of autocatalytically replicating species. With selection, information is generated successively, leading to a steady optimization of species, which can either be organisms or molecules.

The mathematical models developed by Eigen support a detailed hypothesis of the origin of life which comprises multiple, successive steps for the transition from inorganic to living matter. However, it should be mentioned that some scientists have theories on the emergence of life that differ from Eigen's theory. Among these is Stuart Kauffman (1939–), who believes that natural selection is important but not the sole ordering principle of the biological world. Instead, he considers spontaneous self-organization to represent the predominant source of natural order. Kauffman demonstrated that sets of inter-related autocatalytic reactions can undergo a transition to a newly ordered (i.e., self-organized) state as soon as their connectivity reaches a certain threshold value. Furthermore, Kauffman emphasizes that the phenomenon of autocatalysis, which plays the central role in his theory, is not limited to nucleic acids. Therefore, he concludes that even genes were not necessary for the origin of life. In contrast to Kauffman, Eigen distinguishes "random" autocatalytic or self-replicating activity which is observed for a variety of molecular species from the "inherently" self-replicating nucleic acids. Inherent capability for self-replica-
tion, in turn, represents the molecular basis for natural selection according to Eigen's theory.

Well-defined experiments were invented in order to simulate the principles that were postulated for molecular evolution. With certain experimental set-ups, replication and selection can be performed in a test tube. Similarly, the chemical conditions on primordial Earth can be mimicked in the laboratory. Several scientists attempted to verify experimentally the twentieth-century ideas on biogenesis. Their experiments are discussed in the following section.

C. Experiments

1. Production of Building Blocks and Polymers

The ideas of Oparin and Haldane inspired the first simulations of prebiotic conditions on early Earth. Under the auspices of Harold Clayton Urey (1893–1981), graduate student Stanley Miller (1930–) attempted to simulate the primordial chemical processes in a famous experiment. He continually exposed a mixture of methane, ammonia, hydrogen, and water vapor above an "ocean" of boiling water to sparks of a corona discharge in a closed apparatus (Fig. 1). The gaseous reaction products were condensed in a cooler, dissolved in boiling water, and recirculated to the gaseous atmosphere. After several days of exposure to sparking, the aqueous solution changed color. Subsequent analysis indicated that several amino and hydroxy acids had been produced by this simple procedure. Because on early Earth probably much more energy was available in ultraviolet light than in lightning discharges, astrophysicist Carl Sagan (1934–1996) and colleagues altered Miller's experimental conditions. They synthesized amino acids by applying long-wavelength ultraviolet irradiation to a gaseous mixture of methane, ammonia, water, and hydrogen sulfide. Similar to Miller's experiment, these attempts also revealed that amino acids, particularly the biologically relevant ones, form readily under simulated primitive conditions. Also, it became obvious that reducing conditions were necessary for prebiological formation of amino acids since no yield was observed with oxidizing atmosphere. Further modification of these types of experiments showed that alkaline conditions led to the spontaneous synthesis of a variety of sugar molecules, including the five-carbon sugars fundamental for the formation of nucleic acids and six-carbon sugars such as fructose and glucose, which represent common metabolites and structural building blocks in contemporary organisms. As was shown by Juan Oro (1923–) and several other investigators, nucleotide bases and porphyrins can also be synthesized from simple mixtures of ammonia, water, and hydrogen cyanide. Therefore, it seems probable that most—if not all—essential building blocks of proteins, carbohydrates, and nucleic acids can be produced under quite general primitive reducing conditions.

The formation of polymers, long-chain molecules made of repeating units of the building blocks, would have to be the next major step during chemical evolution. However, polymerization of monomers was not achieved in the simple experiments described previously. These reactions involve the loss of one water molecule during the formation of each two-unit product and therefore are facilitated under dehydrating conditions. By dry heating of amino acid mixtures, Sidney Fox (1912–) accomplished the condensation reaction in the laboratory to yield polypeptide acids. He observed that these molecules were not random polymers and they also exhibited some distinct enzyme-like catalytic activities; therefore, he called them proteinoids. In addition to this work, various experimental approaches concerned the catalysis of polymerization reactions by mineral surfaces, which may also protect growing chains from degradation by water molecules, as well as combinations of heating under dry and wet conditions mim-
The environment of submarine hydrothermal vents. Some scientists cast considerable doubt on any contribution of prebiotic syntheses of organic compounds to the origin of life. Predominant among these is Günther Wächtershäuser (1938–), who imagines an origin without reducing atmosphere and without the primordial soup. According to Wächtershäuser, prebiotic chemistry began on the surface of minerals in the lithosphere. If these minerals were positively charged, they would form ionic bonds with negatively charged chemical groups of phosphate, carbonyl, or sulfide. As a result, a layer of negatively charged molecules would form, coating the minerals in two dimensions. Once a diversity of molecules accumulated, inorganic reactions involving the minerals could supply reducing power for the synthesis of more complex organic molecules from various possible interactions within the two-dimensional layer.

The most important implication of Wächtershäuser’s theory is that the first forms of life could have been autotrophic instead of heterotrophic. Autotroph organisms get their carbon from carbon dioxide, whereas heterotrophs get it from organic molecules such as glucose. The fact that the oldest microfossils appear to be photoautotroph blue-green bacteria supports the autotroph hypothesis. Nevertheless, these ideas have not been adequately tested.

2. Evolution Experiments
The onset of self-replication—following the formation of polymeric nucleic acids—represents a major difficulty for the explanation of the origin of life. Experimental attempts by Leslie Orgel (1927–) and coworkers revealed that, under prebiotic conditions, it is possible to add monomeric units, called nucleotides, to complement a small template nucleic acid molecule and to combine the adhering nucleotides to the correct negative copy of the positive original. This is merely the first step of a self-replicating process that must be continued by a separation of both strands, positive and negative, and by repeated addition and concatenation of monomers. Without proteins supporting the polymerization of the building blocks, identical copies of the original nucleic acid molecule, the positive strand, could not be synthesized. However, because the simultaneous emergence of self-replication of nucleic acids and specific catalysis of their duplication by proteins seems very unlikely, Orgel, together with Francis Crick and Carl Woese (1928–), supposed that the first and primitive self-replicating systems involved nucleic acids only. The discovery of catalytically active nucleic acids, or ribozymes, by Thomas Cech (1947–) and Sidney Altman (1939–) decisively supports this hypothesis. Much contemporary work therefore focuses on searching for ribozymes catalyzing the replication of nucleic acids.

Experimental studies on selection among replicating polynucleotides were first performed by Solomon Spiegelman (1914–). He purified nucleic acid and the replicating enzyme of bacteriophage Qβ and implemented their replication reaction in a test tube which contained all essential reagents. Then he replicated the viral nucleic acid in serial transfer experiments by applying constraints that select exclusively for higher replication rates (Fig. 2). Beginning with infectious nucleic acid molecules, sequences were obtained after several rounds that replicated several times faster and had much lower molecular weights than the original molecules but were unable to infect bacteria. On the basis of more detailed knowledge of this replication reaction, Christof Biebricher (1941–) and coworkers, together with Manfred Eigen, repeated Spiegelman’s experiments. They applied conditions which were chosen to be optimal with respect to rate of evolution, and they analyzed the stepwise progress of the optimization procedure. Their attempts at facilitating artificial evolution established an ‘‘irrational’’ technique for the design of new drugs known as directed molecular evolution. Eigen and coworkers also performed evolution experiments with various different replicators, either molecules or viruses, and they observed functional organization between replicators and their catalysts as well as some of the molecular mechanisms of evolutionary optimization.

II. THEORETICAL CONCEPTS
A. Quantitative Treatment of Darwinian Evolution
The process of complementary reproduction of self-replicating molecular assemblies can be described in detail by applying the mathematical formalism of chemical kinetics. For any population of molecules within a certain environment, consideration from the viewpoint of kinetics accounts for three necessary conditions: metabolism, self-reproduction, and mutability.

Metabolism describes the rates of formation and decomposition of each molecular species taking part in the competition. The term metabolism expresses the continuous influx of energy-rich matter, which main-
FIGURE 2  Scheme of the serial transfer principle performed by Spiegelman. A series of test tubes contain QL replicase, RNA building blocks (nucleoside triphosphates of A, G, C, and U), and necessary growth factors. The mixture is supplied with RNA template, and the reaction is started by increasing the temperature to 30°C. When a certain product concentration is reached, a small fraction of the mixture is transferred to the next tube and incubated in the same way. The procedure is repeated many times, continually reducing the incubation period. Eventually, a single optimal RNA product is selected.
with the expectations drawn from the Prigogine–Glansdorff principle: It results in the breakdown of the formerly (meta-)stable quasispecies and its substitution by a new quasispecies.

The quasispecies behaves physically and mathematically like a single species, or more precisely like the selected wild type of a species, but does not define a single species. It includes all mutants of the dominant form and represents a widely dispersed, not necessarily symmetrical, distribution of similar and related, but by no means identical, sequences. A quasispecies is usually characterized by a defined average sequence, called a consensus sequence, and can be dominated by one or more so-called master sequences. Evolutionary progress of the quasispecies results in new mutants that appear on the periphery of the distribution. Selection, or the emergence of mutants which replicate more efficiently than their ancestors, is a manifestation that new genetic information has been generated.

B. The Informational Aspect

1. What Is Information and How Does It Originate?

In the most common sense, information is associated with the content and meaning of a ‘news’ message. From the viewpoint of a recipient, the news can be understood only if the recipient and the transmitter agree on the meaning of the symbols used to code the message. In other words, new information is understood only when the error rate is near the threshold value. Indeed, it has been established experimentally by Esteban Domingo (1942–), Charles Weissmann (1931–), Lawrence Loeb (1936–), and several other research groups that viruses (which, due to their short replication periods, are ideally suited for evolutionary studies) behave like quasispecies and replicate close to their error thresholds. In accordance with theory, the reciprocals of the error rates that have been experimentally determined for many viruses define the limiting information capacity of their genomes. Loeb was also the first to prove the hypothesis that an additional increase in the mutation rate would abolish viral replication. By applying deoxynucleoside analogs to HBV replicating in vitro, he observed a complete loss of viral replicative potential and thus a lethal mutagenesis.

Evolutionary progress is achieved at reasonable velocities only when the error rate is near the threshold value. Therefore, natural evolution processes are bounded by an error threshold. Far below this boundary, selection can be reduced to all-or-none decisions. Evolutionary progress is achieved at reasonable velocities only when the error rate is near the threshold value. Indeed, it has been established experimentally by Esteban Domingo (1942–), Charles Weissmann (1931–), Lawrence Loeb (1936–), and several other research groups that viruses (which, due to their short replication periods, are ideally suited for evolutionary studies) behave like quasispecies and replicate close to their error thresholds. In accordance with theory, the reciprocals of the error rates that have been experimentally determined for many viruses define the limiting information capacity of their genomes. Loeb was also the first to prove the hypothesis that an additional increase in the mutation rate would abolish viral replication. By applying deoxynucleoside analogs to HBV replicating in vitro, he observed a complete loss of viral replicative potential and thus a lethal mutagenesis.

Generally, the theoretical and experimental results on replication error thresholds are not just true for viruses. They also remain valid for all replicating individuals that can be described by the formalism developed by Eigen (Table 1).

The complexity that life has developed by successive evolutionary optimization does not just result from vertically transmitted point mutations. (Vertically’ expresses the hereditary transmission of errors to the direct offspring.) Additional principles for the modification and expansion of genomes exist that enable the horizontal transfer of genetic information, i.e., the transfer between individuals or species. These additional principles are homologous genetic recombi-
TABLE I
Stepwise Expansion of Information Capacity during the Evolution of Life

<table>
<thead>
<tr>
<th>Replication process</th>
<th>Subject/organism</th>
<th>Error rate</th>
<th>Maximal genome length</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzyme-free replication (base pairing)</td>
<td>AU polymer</td>
<td>$10^{-1}$</td>
<td>$10$</td>
<td>$80$ (tRNA ancestors)</td>
</tr>
<tr>
<td></td>
<td>GC polymer</td>
<td>$10^{-2}$</td>
<td>$10^2$</td>
<td>$4000$ (bacteriophage Qb)</td>
</tr>
<tr>
<td>Enzyme-catalyzed RNA replication</td>
<td>RNA viruses</td>
<td>$10^{-7}$</td>
<td>$10^7$</td>
<td>$4 \times 10^6$ (Escherichia coli)</td>
</tr>
<tr>
<td>Enzyme-catalyzed DNA replication</td>
<td>Prokaryotes</td>
<td>$10^{-10}$</td>
<td>$10^{10}$</td>
<td></td>
</tr>
<tr>
<td>Recombinative, enzyme-catalyzed DNA replication</td>
<td>Eukaryotes</td>
<td>$10^{-14}$</td>
<td>$10^{14}$</td>
<td>$3 \times 10^9$ (man)</td>
</tr>
</tbody>
</table>

...
genes allowed their mutual survival and regulated their growth. It also enabled a more refined competition–selection behavior than that among the individuals of a quasispecies. Instead of single RNA mutants, cooperative ensembles of nucleic acid and protein were evaluated according to their self-replication qualities and according to their stabilities. First, those sequences were selected as fittest that were best able to get themselves replicated as quickly and as accurately as possible by the enzyme responsible for their replication. Second, with the continual introduction of new nucleic acid mutants, new catalytic couplings were constantly tested.

The cooperation results in a double-feedback loop in which both the enzyme encoded by the nucleic acid template and the sequence information contribute to the replication of the template; a process called second-order autocatalysis. Phenomena of this particular kind were named “hypercycles” by Eigen and coworkers.

Hypercycles exhibit three major characteristics. First, they consist of coexisting and cooperating quasispecies, each of which is replicating autocatalytically and thereby maintaining competitive growth, selection, and its specific genetic information. Second, some of the coexisting quasispecies replicate and evolve independently. Finally, others form cooperative units within the hypercyclic organization and evolve together with mutual advantage. In an extreme case, some species as well as quasispecies may exist only by participating in hypercyclic organization. As a whole, the cooperating units can compete with other hypercyclically replicating systems.

The overcoming of the information crisis of self-replicating nucleic acids probably proceeded via hypercyclic coupling as shown in Fig. 3, and it supported the emergence of “translation” from nucleic acid into protein. It is assumed that hypercycles represent one major step in the organization of complex reaction networks which arose as naturally and continuously as did quasispecies. Experimentally, hypercyclic organization and its efficiency were demonstrated by Eigen and coworkers to mainly determine the infection and replication cycle of a simple bacteriophage such as Q8.

D. From Hypercycles to Cells

All life on Earth is cellular and reproduces only when this complex replicative unit, the cell, divides. Among the obvious advantages of cellular organization are protection of the cell’s content from fluctuations of resources in the external environment and the maintenance of internal concentration gradients necessary to drive chemical reactions. Nevertheless, these facts do not explain the necessity for cellular organization: Instead, the origin of spatially limited reaction systems from a homogeneous prebiotic soup appears to be a result of problems in information processing.

Although hypercycles are the essential organizational form for the transition from self-replicating molecules to reproductive, multi-molecular machineries, they do not represent the ultimate optimum of organization. Hypercycles and quasispecies share a substantial evolutionary disadvantage: Both quasispecies competi-
tion and hypercyclic cooperation evaluate only the phenotypic properties of replicating nucleic acids, i.e., their replication rates and their stabilities. Hypercyclic coupling enables the discovery of molecular replicators which are beneficial to one or more others, and it supports the selection of those coexisting replicator combinations which gain mutual advantage by amplifying themselves. However, hypercycles cannot evaluate genotypic properties of nucleic acid sequences or their genetic messages. In other words, they do not selectively amplify mutant sequences that demonstrate their advantages only before translation.

Primitive translation mechanisms probably gave rise to proteins that were more helpful to self-replication than miscellaneous proteins randomly occurring in a homogeneous distribution. With time, preferences between translation products and certain sequences became pronounced, and these more distinct interactions effected advantages due to more specific catalysis. Finally, the differences among the various template-
catalyst interactions became so influential that each enzyme had a particular catalytic role. At this stage, selection acted not only on the kinetic characteristics (i.e., the phenotype) of the replicating sequences but also on the information content (i.e., the genotype) of the sequences. Within the enlarged information system, the amount and quality of coding genes were strongly determined by efficiency, fidelity, and rate of the catalyzing enzymes, and these in turn were the translation products of their templates. By spatial separation of individual hypercycles, selective genotypical advantages in competition with other compartments were exploited. Favorably mutated compartments could develop more strongly than others and thus evolve more quickly. The compartmentalization was accomplished by individualization of all replicative units represented by a hypercycle. This event can be viewed as the birth of the cell, the smallest living entity known to us today. Figure 4 summarizes the complete view that results from these theoretical descriptions.

See Also the Following Articles

ARCHAEA, ORIGIN OF • BIODIVERSITY, ORIGIN OF • DARWIN, CHARLES • EUKARYOTES, ORIGIN OF • EVOLUTION, THEORY OF • NUCLEIC ACID BIODIVERSITY

Bibliography

I. Introduction

PALEOECOLOGY is the study of the composition and distribution of past ecosystems and their changes through time on scales of decades to hundreds of millions of years. Paleoecologists derive their inferences mainly from fossil and geological data and have assembled data sets with local to global coverage. These data sets provide the possibility of having a zoom lens view of past changes both in space and in time. I use Quaternary pollen and plant macrofossil data to illustrate how these data can be organized spatially, temporally, taxonomically, and numerically to provide a zoom lens view of a variety of ecological phenomena.

II. Scaling Factors

III. Zoom Lens Perspective

GLOSSARY

AMS dates Radiocarbon dates obtained by directly measuring the amount of carbon 14 in a sample by using an accelerator mass spectrometer.

Bryophytes Mosses and liverworts; nonvascular plants.

Cyperaceae Sedge family.

Fagus grandifolia American beech trees.

Late Quaternary The past 21,000 years since the last time of maximum glaciation.

Pollenology The study of pollen and spores.

Pliocene The period of geological time from 3.5 to 1.8 million years ago.

Poaceae Grass family.

Quaternary The past 1.8 million years of geological time.

radiocarbon dates Dates of organic matter in geological samples using the radioactive decay of carbon 14, which has a half-life of 5730 years.

Tracheophytes Vascular plants.

Past ecosystems are a continuous creation of those who study them. What we map through time or reduce to time series depends on the data available and the training of those who interpret and display those data. Studies in paleoecology, therefore, are an exercise in perception and interpretation and depend on the sensitivity of the data in time, space, and taxonomy. The ultimate goal is to understand what occurred ecologically in the past, but this understanding can only be obtained if reliable records exist of past taxon distributions (or other components of ecosystems) that can be displayed and interpreted in informative ways. Simple time series are not enough because of the many spatial variations and processes that affect them. Geographic networks of temporally varying data are required to show how past and present taxa have varied in abundance, location, and association within and among their associated ecosystems.
The focus here is on the rich data sets of Quaternary pollen and plant macrofossils that provide records of the major changes in vegetation at a variety of space and time scales. These data provide a valuable bridge from modern ecological data to the vast fossil record in geological time that Jablonski, Sepkoski, and others have compiled from the work of paleontologists. The Quaternary data sets yield a remotely sensed view of past vegetation, landscapes, taxon migrations, and invasions and illustrate various temporal changes in ecological assemblages and communities—all of which show changes and patterns in biodiversity. Direct measurement of changing species through time, however, is not possible with pollen data, even when supplemented by plant macrofossil data, although the extinction and arrival of some species are evident in records dating back to the Phocene. The focus with these paleoecological data, therefore, is on vegetation and its diverse changing patterns of composition and distribution through time and space and how environmental factors, such as climate and human activities, have influenced these changes. These data therefore show the ecological and environmental setting within which modern patterns of biodiversity have arisen. They also demonstrate, as long claimed by Richard West and Margaret Davis, that the "present plant communities have no long history.''

A. Data and Sensing System

For map views of the globe and continents, the pollen and plant macrofossil data can be viewed as if they came from a video recorder from high in space with resolution in places of up to 10 m. The images that are retrieved can be of high or low resolution temporally, spatially, taxonomically, and numerically, and they can illustrate local to global changes in plant populations, vegetation, biodiversity, human activity, fire frequency, and plant diseases over decades to millions of years. Because each of these entities or phenomena varies spatially and temporally, records of data covering a breadth of scales in space and time are needed. To obtain the highest quality images of a specific phenomenon requires an understanding of the sensing system that accumulated the data. How does the paleoecological video recording system work and what are the scaling characteristics of the images that it registers? These characteristics include breadth of coverage, sampling resolution, and sampling density in time, space, and taxonomy. Studies of pollen and plant macrofossil data covering a variety of temporal and spatial scales within the Quaternary have helped provide this understanding. Quaternary paleovegetation data pose several problems for interpretation. The main one arises because pollen samples in sediments represent death assemblages of immature microgametophytes (i.e., pollen grains) that differ manifestly from the assemblages of sporophytes (i.e., plants) that produced the pollen. Studies of modern data from surface sediments have therefore focused on identifying the features of plant assemblages that appear in pollen. This work parallels that of scientists deciphering the remotely sensed data from satellites because pollen data are remotely sensed data from plant populations and vegetation. Just as the current vegetation emits or reflects radiation that remote sensors on satellites intercept, too does (and has) the current (and past) vegetation shed pollen that accumulates "remotely" (i.e., well away from the source) in lakes and bogs. Both types of 'remote' sensors (satellite instruments and lake sediments) record data with certain sampling characteristics (e.g., spatial and temporal resolution) and their data need calibration and ground-truthing in terms of vegetation attributes such as composition (taxon abundances), structure (height and mixture of growth forms), and pattern (geographic gradients and mosaics and aspect of beta and gamma diversity). Studies of spatial arrays of modern data by Bradshaw, Prentice, and Jackson have provided this calibration and ground-truthing. Down-core studies of pollen bring time into the picture, and temporal resolution becomes one of the factors controlling what is recorded in the data. Pollen in annually laminated sediments can provide seasonally distinct samples as nearly shown by Peglar, but most samples integrate 10 or more years and can be independently dated back 40,000 years with an average precision of ±200 years for data from within the past 12,000 years. However, if vegetational and ecological phenomena are the target for study, then spatially distributed data are needed because vegetation and biodiversity are inherently spatial entities that vary on virtually all time and space scales. Recording their full dynamics requires time series of geographically distributed data that yield a zoom lens space–time perspective. I focus on how the different scaling characteristics of the data sets control what we see and on how data can be organized to yield such a zoom lens perspective. I discuss each of the sampling characteristics in taxonomy, space, and time. I then describe different ways to display and visualize the data and some findings that they have revealed.

II. SCALING FACTORS

Most Quaternary paleontologists study sediments from lakes or bogs that accumulate pollen relatively continu-
ously, for which the data from the surrounding vegetation can be averaged, and whose sediments are radiometrically dated. In these sediments, pollen grains are morphologically distinct and numerous, and plant macrofossils are found in a subset of these sediments. (Pollen and plant macrofossil data are also available from deposits that are discontinuous in time but can be organized into time series; Betancourt et al., 1990). Once analyzed, the data exist as point estimates of the abundance for each taxon in time series of samples at a site (or at nearby sites and middens for some discontinuous records). These time series can be expanded to transects of time series in latitude–, longitude–, or elevation–time diagrams, in networks of samples from an area for mapping, or to networks of time series, which are also time series of maps, to form a space–time box for displaying the data (Fig. 1). Each of these displays illustrates different views in different dimensions of the space–time variations in the data. A zoom lens perspective can then be achieved in space, time, or both simultaneously by moving from data at a single site for a short time interval to progressively longer time series for either that site or for data on maps from increasingly larger areas. In this way, records of local succession can be seen in the broader context of long-term migrations and global climate changes.

Palynologists control the ability of their data to display selected patterns in vegetation and biodiversity (both in time and in space) by making choices about the numerical aspects of their data and by choosing the temporal, spatial, and taxonomic sampling characteristics of their data. These sampling characteristics have three elements—breadth of coverage, resolution of individual samples, and sampling density—which for a photograph are comparable to its frame of reference, grain size, and number of grains (or pixels) that are exposed, respectively. The temporal and spatial breadth of the data is the total time or area covered by a data set, and the taxonomic breadth is the total set of taxonomic groups (e.g., seed plants) included. The temporal and spatial resolution of each sample is defined in terms of how much time or area is represented in each individual sample within the data set, and taxonomic resolution depends on whether the data are lumped into groups or listed at their finest level of morphological distinction. The uncertainties of radiocarbon dates and age estimates also affect temporal resolution in correlations among sites. The number of samples in time and space defines the temporal and spatial sampling density. The total number of taxa listed in a data set defines the taxonomic sampling density. Choices about these characteristics can influence what changes in vegetation and biodiversity appear in palaeoecological times series and map sequences (Figs. 1–3).

A. Taxonomic and Numerical Characteristics

In most studies of lakes and bog sediments, the potential taxonomic breadth is Tracheophytes and Bryophytes along with certain algal remains (Prasinophyceans) and fungal spores. When non-wetland vegetation is a primary focus, spores, algae, remains, and pollen and plant macrofossils from aquatic plants are excluded from detailed study. Careful study and morphological characteristics of pollen grains determine the taxonomic resolution in samples and permit identification of genera and some species, although some grains can only be determined to the family level (e.g., Poaceae and Cyperaceae). Regional restriction of single species, otherwise identifiable pollen morphologically only at the genus level (e.g., Fagus grandifolia within the northeastern United States), can allow designation of plant species from pollen, but plant macrofossils, if present, give the best information on species identification.

Palynology is a direct beneficiary of the ubiquity and inefficiency of wind pollination. Millions of grains are produced for every successful pollination, and some of the abundant residual reaches the sediments of lakes and bogs where the pollen grains are well preserved. Entomophilous grains are abundant in honey and on the legs of bees, but too few insects perish in lakes and bogs to leave a record that can be discerned among the overwhelming numbers of anemophilous grains. From the point of view of the remote-sensing metaphor, wind-pollinated plants are bright lights on the landscape and little or no signal comes from the other plants. As a result of this bias, the focus is on using pollen data to record the vegetation rather than species lists. For temperate to boreal forests in which the diversity of wind-pollinated trees is greatest globally, pollen records provide a fairly direct representation of the vegetation; however, in tropical regions and deserts, the pollen records yield a much more indirect representation of the vegetation because most species are not wind pollinated.

In arid lands of southwestern North America, where lakes are rare, packrats create middens with embedded plant macrofossils that can be radiocarbon dated and identified to the species level for those plants collected (Betancourt et al., 1990). These biased records provide valuable information about arid vegetation and environments.
In most studies of Quaternary pollen, taxon percentages are used. Three hundred to 500 grains are typically counted in a sample, and the counts for each taxon are divided by the total count. Because counts for individual taxa can be thought of as being binomially distributed (i.e., each grain is either taxon x or not taxon x), the percentages for multiple taxa are multinomially distributed, thus permitting direct calculation of confidence intervals. Pollen concentrations (grains/cc or grains/g) are seldom used because of their dependence on sedimentation rates. When radiocarbon or stratigraphic dates allow estimation of sedimentation rates,
pollen accumulation rates (grains/cm²/year) can be calculated, and procedures exist for calculating and potentially minimizing the confidence intervals for both pollen concentrations and accumulation rates.

Pollen accumulation rates, first introduced by Margaret Davis, have proven valuable for checking the ambiguity of certain changes in pollen percentages. However, pollen accumulation rates are unsuitable when large sets of comparable pollen data are required for mapping. These rates are highly sensitive to sedimentation differences within and between lakes as shown by W. Pennington and M. Davis, and percentages of pollen taxa are generally used in mapping studies (Figs. 1–3).

Many studies show that the percentages for pollen taxa represent well the relative abundance of plants on the landscape. The relationships vary with spatial scale and pollen type, and the uncertainties of these relationships add to the number of uncertainties for the pollen data when estimates of the vegetation are attempted.
Finding reliable quantitative measures for plant macrofossils is much more difficult than for pollen data. Most data are recorded in presence/absence terms or categorically, but sometimes percentages or concentrations are calculated. Local biases can make the data difficult to interpret because, for example, 100 needles may all come from the same tree.

B. Spatial Characteristics

The breadth of coverage for sets of pollen data can be broad or fine and thus match the different scales for mapping the vegetation (Fig. 2). Continental and global data sets exist (Prentice et al., 2000) along with those at regional (1000–300,000 km²) and local (10⁻⁴ km²)

FIGURE 3  Zoom-lens view of the mapped distributions of oak tree and pollen percentages at the scales of a subcontinent (10⁷ km²), state (10⁵ km²), and county (10³ km²). These maps show how well pollen percentages can reflect the distribution of tree percentages. Data used in contouring at the subcontinental scale were also used in contouring at the state scale, but a different data set was used at the county scale (modified from Solomon and Webb, 1985).
scales, and an embedded series of these data sets can provide a zoom lens view of past vegetation change if the sample resolution and sampling density are appropriate.

The potential spatial resolution of individual pollen samples varies among pollen taxa depending on the dispersability of their pollen. Pollen from taxa with large grains is dispersed less far than that with small grains. Individual pollen samples are therefore variable-area samples, with the area of vegetation contributing pollen depending on which pollen type is studied; the percentages for pine pollen, for example, represent a larger area than those for beech or maple. Such differences in sampling area among pollen types can be important for studies of local or even regional vegetation but are less important for networks of data in which the distance between samples is much larger than the average dispersal distances for the different taxa.

The potential spatial resolution also depends on the basin characteristics. Key factors include the presence or absence of a canopy over the basin, the size of the water or wetland surface accumulating airborne pollen, the presence or absence of horizontal mixing of sediments within the basin, and the watershed area that supplies additional waterborne pollen to lakes. Samples from within a forest canopy accumulate most of their pollen in radiuses from 10 to 100 m, and data from samples that preserve pollen are comparable spatially to vegetation data from large permanent plots in forests. In contrast, samples from lakes accumulate their pollen on scales from 10 to 50 km. The former data have enough spatial resolution to illustrate succession within forests, whereas pollen data from lakes integrate across the mosaic of vegetation on landscapes and are sensitive to succession only if it is occurring across much of the pollen-source area. Bogs, mires, and shallow wetlands accumulate pollen at both regional (1–10 km) and local (10 m) scales because many plants grow on these wetlands. In samples within mires, pollen from local taxa can vary systematically and abruptly, whereas the pollen from regional taxa remains reasonably uniform. (This behavior of pollen percentages for regional taxa led von Post to develop pollen analysis as a method in 1916.) The spatial resolution of pollen samples, therefore, varies with pollen type within certain basin types, but selection of samples by pollen type and basin type can help keep the spatial resolution reasonably uniform. Such choices are key to "seeing" well with the observations afforded by sets of pollen samples.

For plant macrofossils, Dunwiddie and Jackson have shown that needles and seeds in lake sediments have a spatial sampling resolution of 10–20 m, which is much finer than the distance sampled by most pollen in lakes or open wetlands. Plant macrofossils can therefore help resolve local changes at or near a site and fine-scale elevation differences. Therefore, sampling schemes are possible that mix open and canopy-covered basins, mire and lake sediments, local and regional pollen types, and plant macrofossils in order to represent different scales of spatial pattern in the vegetation. Such mixtures of data sets can show vegetation differences that reflect specific soil and elevation differences within a broader geographic data set.

The geographical sampling density for pollen data varies. To illustrate the variations in recent pollen within a basin or a forest, some studies include samples at 10-m or finer intervals along 500-m or longer transects through a basin or forest. For selected taxa, these studies show high sensitivity in the pollen data to local variations in vegetation cover and biodiversity. The variations are also evident in sets of time series from closely spaced cores in a basin (Simmons, 1993). For studies whose breadth of coverage is regions to continents (Fig. 2), sets of modern data exist with average densities of 1/14 km² in 1000 km², 1/500 km² in 10,000 km², and 1/5000 km² in 100,000 km². These densities are sufficient for the contour patterns of oak pollen percentages to match the corresponding patterns of oak tree percentages at each spatial scale (Fig. 3). Sets of fossil data provide less dense geographic coverage and vary in density from 1/50 km² for the Adirondacks to 1/6000 km² for the northern Midwest and 1/40,000 km² for eastern North America and Europe.

These data sets have been used to illustrate how well pollen data represent or remotely sense spatial vegetational features such as range boundaries, ecotones, and abundance gradients (Figs. 2 and 3). Detecting the range boundary with pollen can be difficult because of pollen transport and relatively high counting uncertainties at low pollen percentages, but Davis et al. (1991) used a data set from the Midwest with a sampling density of 1 sample/1000 km² to show that the past species limit for beech and hemlock can be identified within an area 20 km wide. For widely dispersed taxa such as oak and pine, such fine-scale resolution may not be possible, but the abundance gradients for these taxa can be used to locate ecotones at scales of 10–100 km (Fig. 3).

C. Temporal Characteristics

The time range for records from individual cores can vary from 50 years to millions of years ago (Fig. 4), with the bulk of the late Quaternary cores covering 14,000 years and fewer extending back 21,000 years,
FIGURE 4 Zoom lens view of pollen data in time from a 1.5-million-year record down to a 300-year record. Different ecological processes are evident on different timescales. Temporal coverage, resolution, and density all vary appropriately for the Colombian lake record of alder pollen to show (a) alder's arrival at the site long after North and South America joined and (b) its response to orbitally driven climate changes (Hoogheemstra and Ran, 1994). The California lake record of oak pollen shows its response to orbitally driven climate variations over 130,000 years and the transition from glacial to interglacial climate over 16,000 years. In Wisconsin, pollen data from a lake that is free from bottom-organism mixing are sufficiently high in resolution to record how oak and pine populations responded to climate and local disturbance events (e.g., fire) over 11,000 and 2500 years, respectively. Local succession is evident in the pollen data from a 300-year record from a small hollow within the forest canopy in Massachusetts (Foster and Zebryk, 1993). Gray indicates a portion of a time series that is expanded in the graph on the right.
the last glacial maximum. Temporal density generally correlates with time range because short cores are often taken to study short-term changes and may contain 10–20 samples for 500 years, which yields 1 sample per 25–30 years, whereas 200 samples in a 2-million-year record can only yield 1 sample per 20,000 years. Twenty to 40 samples for 10,000 years yield 1 sample per 250–300 years, which is typical for most cores in the data sets from eastern North American and European (Huntley and Webb, 1988). At a high level of temporal density are records from millimeter-scale sampling of peats that have yielded 1–5 year sampling for 50–200 years for selected times in the past, as reviewed by Turner and Peglar (1988) and Simmons (1993).

One of the longest continuous records for the Quaternary is from Sabana de Bogota in Colombia and covers 1.5 million years (Fig. 4). This study is remarkable in having relatively high sampling density of 1 sample/1000 km2 and shows the appearance from North America of Alnus and Quercus at 1.3 million and 250,000 years ago, respectively. Other long cores with records of 50,000 years or longer exist on all continents but Antarctica; however, only in Europe are these long cores numerous enough to allow some mapping of vegetation from previous glacial and interglacial periods.

At the other extreme in terms of temporal coverage are (i) short cores of the upper sediments of lakes or bogs and (ii) profiles from forest soils that preserve pollen. The former provide records of human impact on the vegetation and of recent changes, whereas the latter show combined fine-scale sampling in both time and space and provide the view most closely approximating scale within which succession into forest gaps occurs. One of the ironies of paleoecological research is how seldom the data reveal succession as the dominant dynamic in the vegetation even though much ecological research has focused on succession.

These short-term records are often analyzed in high temporal density, e.g., 1 sample per 50 years or even 1 per decade. The short cores (50 cm to 2 m) from lakes give coverage in high detail for 100–500 years or more in transects or in networks of relatively high spatial density (1 sample/1000 km2) from states or regions. Dating can be fairly accurate and precise because historical dates can be assigned to events in cores, and the date for the core top is usually known. High-resolution mapping in 50-year intervals is possible.

### D. Temporal Resolution

A key metaphor for thinking about temporal resolution within individual samples and cores is to consider stratigraphic sections of lake sediments as strip charts along which pollen data accumulate to form "palygraphs." The sediment accumulation rate (which varies but averages 0.7 mm per year in most Holocene lakes) measures the drum speed (as on a thermograph) that drives the strip chart past the recorder, and the mixing of sediments within 1 to a few cm of the sediment surface before burial measures the degree to which the pen recording the data wiggles up and down vertically in time while registering horizontally the changing abundances of pollen. Within this sediment mixing zone, older sediments are moved both up from below (2–6 cm) and in from other parts of the basin by reworking. Such processes can make for tricky pen work on the palygraph. If the paper in the imagined palygraph were to absorb the ink slowly such that the scatter of tracings in the upper mixed zone is reduced to a single tracing of the abundances below this zone, then our metaphorical instrument would yield what is finally observed down-core. This image of the palygraph helps by indicating the separate components controlling temporal resolution in a core. The mixing depth and sedimentation rate largely control postdepositional time averaging, which regulates the time interval (i.e., temporal resolution) represented by pollen in a sample, whereas the temporal density of samples recorded and the accuracy and precision of assigning dates to each of many cores (strip charts) control analytical time averaging, which determines the temporal resolution among a set of cores used for mapping.

### E. Dating Uncertainties

The dating uncertainty of an arbitrary pollen sample at some depth in a core depends on the uncertainties of the dating method, what is dated, the uncertainties of the age model, and variations in the sediment accumulation rate. The most accurate and precise dates are from lakes with annually laminated sediments, whose annual layers can be counted. These dates are in calendar years and accurate, with dating errors of only 1 or 2% at most. Unfortunately, lakes with such precision are too sparsely distributed for detailed mapping work. For lakes without varves, the dating near the tops of cores is more accurate and precise than that lower in the core because historical dates can be assigned to events in the core. These dates have uncertainties of 50 years or less and can be pinpointed in the core to 1 or 2 cm.

In studies of late Quaternary data, most cores are dated by conventional radiocarbon dates of bulk sediments. The bulk dates give the dates for the sediment matrix (i.e., strip chart material) and not for the pollen embedded within the sediments (i.e., what the pollen records). The bulk dates have counting uncertainties of...
20–400 years, with most approximately 100 years. These average “errors” give ±200 years as the 95% confidence interval for each date. The dates can also be derived from 2–10 cm of sediment and therefore may average sediments covering 20–1000 years in age depending on the sediment accumulation rate. When dates for “synchronous” pollen events are averaged, such as the elm decline in Europe and the hemlock decline in eastern North America, standard deviations of ca. 300 years result. This uncertainty incorporates both the depth and the counting uncertainties. As Olsson (1986) observed, “There is . . . no point in determining the radiocarbon age with an uncertainty of ±30 years if the relevant pollen-analytical level has an uncertainty of ±200 years.”

Olsson (1986) also notes the problems with contamination, old carbon, and sampling depths. Old carbon can lead to systematic errors of 2000 or more years and requires correction. Dates on wood and accelerator mass spectrometer (AMS) dates on macrofossils in cores are helping to reduce these errors, and AMS radiocarbon dates on wood and accelerator can lead to systematic errors of 2000 or more years nation, old carbon, and sampling depths. Old carbon requires correction. Dates on wood and accelerator mass spectrometer (AMS) dates on macrofossils in cores are helping to reduce these errors, and AMS radiocarbon dates on wood and accelerator mass spectrometer (AMS) dates on macrofossils in cores are allowing depth intervals of 1 cm or less to be dated. Pilcher (1993) concludes that standard errors for radiocarbon dates from sediments are potentially ±200 years but not more than ±225 years.

### F. Age Models and Mapping Intervals

Most pollen samples (30–50 per core) are not dated directly by a radiocarbon date (3–6 per core). Rather, their age is estimated by an age model based on linear interpretation, regression, or some other curve-fitting method. For cores with continuous sedimentation, the age model estimates the drum speed for the strip chart on the palygraph.

The uncertainties for all data plotted on the map for 12,000 years ago give an estimate for the amount of time before and after 12,000 years ago that the pollen data may cover on the map. The uncertainties are likely to be between 300 and 500 years. This dating uncertainty for the data mapped matches the average temporal sampling density of one sample per 300–500 years in late Quaternary pollen studies. Given such large uncertainties, maps are best restricted to 1000-year intervals. Attempts to map at finer intervals will only produce maps whose data are not independently observed from the data in the maps for the next earlier and later dates.

For sets of short cores in which historical dates yield a precision of 50 years or less, age models back to 500 years can yield dating sufficiently precise for maps in 100- to 50-year intervals. For regions over which stratigraphic markers such as volcanic ash represent synchronous layers in cores, some fine-scale mapping intervals are possible for the distant past, but even the dating of tephras has uncertainties of a century or more (Brown et al., 1989). Pollen-assemblage zones are sometimes used chronostratigraphically, especially in Europe, and can yield temporal resolution to 30 years or less among cores, but synchrony must be assumed for processes (e.g., abundances changes for individual taxa) that are ultimately time transgressive, especially at scales of 200 km or more. The geographic area for such maps is therefore relatively small.

### III. ZOOM LENS PERSPECTIVE

Time series and mapped networks of pollen and plant macrofossils record late Quaternary vegetation patterns at the scale of local vegetation, landscapes, regions, and continents (Fig. 2). A hierarchy of nested data sets can reveal vegetation patterns at a variety of scales and provide information from general to detailed about the spatial vegetation patterns (Figs. 2 and 3). Such hierarchies can link local changes to global events by recording each simultaneously (Kutzbach and Webb, 1991). A paleoecological “zoom lens” can then be used to examine the vegetation patterns that dominate different spatial scales for areas of 10° to 10° km² with spatial resolution of 10° to 10° km over timescales covering 10–10° years (with time series to 10° years) with resolution from 1 to 500 years (Fig. 4). From a global perspective over thousands of years with a relatively coarse grid of temporally and spatially averaged samples, investigators can zoom in ultimately to the locally small but statistically significant short-term changes in species abundances within one wetland or forest stand. A zoom lens perspective is possible for eastern North America and Europe and is becoming possible in western North America and other regions.

Two examples using current data illustrate this zoom lens view. The first holds time relatively constant and focuses up spatially from local to continental (Figs. 1–3), and the second zooms in from a 1-million-year timescale down to successional changes during the past 300 years (Fig. 4). The first example covers the past 21,000 years since the last glacial maximum in the northern Midwest and focuses up from a local to a continental perspective. It shows how the local changes in a wetland during the past 12,000 years in central Minnesota are part of regional climate changes that ultimately reflect the global climate changes during de-glaciation and the current interglacial period.

At the local scale of the sampling basin, plant macro-
fossils, aquatic pollen, and diatoms at Kirchner Marsh in east-central Minnesota illustrate that 3000 years after the Laurentide ice sheet retreated from the sites ca. 19,000 years ago, a buried ice block melted to form the kettle basin and its associated lake (Kutzbach and Webb, 1993). Water levels were high until approximately 11,000 years ago, when the seeds of damp ground and weedy annuals began appearing in the core. These appeared when the mixture of oak and herb pollen data show the first evidence of oak-savanna conditions within the landscape and region near the site (Fig. 1). When the herb pollen values increased sufficiently to indicate prairie vegetation near the site, large fluctuations in aquatic seeds indicate intermittent droughts that continued until at least a marsh developed at the site 2000 years ago. This history of local wetland development is closely linked to the vegetation and climate changes indicated at the landscape, regional, and continental levels (Figs. 1 and 2).

Here, I shift from the time series view provided by local pollen and macrofossils to the continental view that the maps of pollen data provide. These show how 15,000 years ago spruce trees grew initially in a parkland that became populated with deciduous trees, first ash and then birch and oak, from 14,500 to 12,000 years ago as forests began to develop near the site for the first time (Fig. 1). This early vegetation was unlike any growing today and illustrates the major compositional changes in vegetation that accompanied the major changes in climate. The sudden decrease in spruce abundance near Kirchner Marsh 12,500 years ago reflects the general northward movement and decrease in regional and continental spruce populations as climate warmed and the ice sheet retreated. Pine populations were then invading from the east to replace the birch trees that briefly grew abundantly near the site in the northern Midwest generally. Oak and elm populations then increased to replace the pines as the climate warmed. The regional addition of herb pollen after 11,000 year ago signals the development of savanna conditions as the climate dried and the ice sheet retreated. Pine populations then decreased regionally as part of a general retreat of oak in the north as conifer populations increased and spruce populations increased south from Canada into the northern Midwest and New England. The broad-scale regional and continental maps show how the local and landscape changes in vegetation fit within the broader context of changes.

The second example is a zoom lens perspective in time (Fig. 4) and shows time series for taxa reflecting the long-term temporal beat in climate at long timescales (1.5 million years to 16,000 years) but then shifting to reflect fires, disease, and other disturbances at shorter timescales, especially when the pollen data are derived from within a forested hollow. Within the latter, the pollen reflects local changes in trees next to the site. This arrangement and description of the data allow paleoecologists to zoom in or out in space–time to observe different aspects of vegetation dynamics and patterns of biodiversity from long-term changes on several spatial scales (from continental to inside the forest canopy) to competitive interactions after disturbance within communities. With such an ordering of data and images, the interconnected roles and impacts of climate change, disturbance, disease, succession, soil development, competition, and evolution can all be observed and potentially distinguished from one another in studies of past biodiversity.

My own studies in mapping and studying pollen data in space–time have convinced me of the potential for resolving many different views of the past. No single fixed perspective is possible because there is no preferred viewing scale or perspective, nor is there any one way to describe the vegetation and biodiversity. Many conventional displays and perspectives are used, but some, such as traditional pollen diagrams, may not be optimal for interpreting key variations in the vegetation. A zoom lens perspective in space and time is required to help explore all the possibilities and to allow for diverse narratives and explanations. Succession and patch dynamics are evident with data scaled to meters and decades, whereas climatically induced migrations and abundance changes appear best over thousands of kilometers and years. To zoom out from a local to a global view, the paleoecological data need to be organized so that highly resolved images of species near a site give way to more generalized averages of data values whose patterns will stand out sharply when viewed across continents or the world. Appropriate study of assembled databases should allow ecologists to resolve and perceive a variety of phenomena, such as ecotones, disturbances, horizons, species distributions, and migration patterns, that are important for biodiversity.
I. Related Phenomena
II. Types of Parasites
III. Adaptations to Parasitism
IV. Origins of Parasitism and Complex Life Cycles, the Evolution of Virulence, and Coevolution of Hosts and Parasites
V. Host-Parasite Interactions
VI. The Ecological Niches of Parasites
VII. The Structure of Parasite Communities
VIII. Parasite Population Dynamics
IX. The Diversity of Parasites (Distribution of Parasites in the Animal and Plant Kingdoms)
X. Zoogeography of Parasites
XI. Economic and Hygienic Importance of Parasites

GLOSSARY

adult parasite A parasite associated with a host during part or the whole of its mature phase.

commensalism An association of animals in which one uses food supplied in the internal or external environment of a host without affecting the host in any way.

ectoparasite A parasite living on the surface of a host.

endoparasite A parasite living inside a host.

facultative parasite A parasite that can also live without a host.

final (= definitive) host A host that harbors sexually mature stages of a parasite.

hyperparasite (of first, second, etc. degree) A parasite living on or in another parasite.

intermediate host A host that harbors sexually immature, developing stages of a parasite.

intraspecific parasitism A parasitic association of members of the same species.

larval parasite An organism that is parasitic only at a larval stage.

latent parasitism Parasitism without obvious symptoms.

mutualism An association of organisms in which both partners benefit from the association.

obligatory parasite A parasite that cannot survive without a host.

parasitism A close association of two organisms in which one, the parasite, depends on the other, the host, deriving some benefit from it without necessarily damaging it.

periodic parasite A parasite visiting a host at intervals.

permanent parasite A parasite associated with a host for long periods.

phoresis An association in which one organism uses another as a means of transport and/or protection.

symbiosis (sensu lato) Any association between organisms (parasitism, commensalism, mutualism, phoresis).

symbiosis (sensu strictu) An association of organisms in which both partners benefit from the association and cannot live without each other.

temporary parasite A parasite found in or on a host only for short periods.

transport host A host that harbors sexually immature stages of a parasite that do not develop.
PARASITISM IS DEFINED IN DIFFERENT WAYS by different authors, usually reflecting their research interest and bias. Parasitism, as used here, is defined as a close association between two organisms in which one, the parasite, depends on the other, the host, deriving some benefit (usually food) from it without necessarily damaging it. Traditionally, fungi, bacteria, and viruses, many of which are parasitic, are studied by microbiologists, whereas parasitologists study protozoan and metazoan parasites. In this contribution, only protozoan and metazoan parasites as well as higher plants (angiosperms) are included. Several types of associations resemble parasitism in various ways and cannot always be clearly distinguished from it, either because of insufficient knowledge or because genuine intermediate forms exist. Such associations are discussed in the following.

I. RELATED PHENOMENA

A. Commensalism

A commensal is an organism that uses food supplied in the internal or external environment of the host, without establishing a close association with the host, for instance by feeding on its tissues. Examples are the amoeba Entamoeba coli, an endocommensal of humans feeding on bacteria in the lumen of the intestine, and the ciliate protozoan Ephemera globularis, an ectocommensal on various marine invertebrates.

B. Phoresis (Phoresy)

In a phoretic association, one organism uses another as a means of transport and/or protection. An example is barnacles living on whales.

C. Mutualism

A mutualistic association is one in which both organisms derive a benefit, but the association is not compulsory. The cleaner fish Labroides dimidiatus feeds on parasites and diseased tissues of various marine fishes, and both partners derive a benefit, the cleaner obtaining food and the host fish getting rid of their parasites and diseased tissues.

D. Symbiosis

Symbionts live in a compulsory association in which both partners derive a benefit. An example is the symbiosis of fungi and algae in lichens. It should be noted, however, that the term symbiosis is sometimes used in a wider sense, including all types of associations between organisms (parasitism, commensalism, phoresis, mutualism).

E. Predation

A predator is an organism that attacks another, the prey, and usually kills and eats it. Most predators are larger than their prey.

An organism may be a parasite under certain conditions, but a commensal, mutualist, or predator when conditions change. For example, Entamoeba histolytica is often a harmless commensal feeding on bacteria in the intestine of man, but may become a dangerous parasite feeding on red blood cells, apparently induced by some changes in the host that are not fully understood. A number of normally pathogenic parasites even improve the health and fitness of their hosts at low infection intensities. Thus, Lincicome (1971, cit. Rohde, 1993) made some controlled experiments using rats and mice infected with two species of trypanosomes and the nematode Trichinella spiralis which showed that parasitized animals grew faster, are more, could compensate better for deficiencies in the diet, had livers richer in certain vitamins, and were more active and more responsive to the human presence. For these reasons, it is best to consider parasitism as a type of association that is not clearly delimited from the others.

II. TYPES OF PARASITES

There are many types of parasites, distinguished by the site of infection, kinds of hosts, state of development, etc. Ectoparasites are parasites that live on the external surface of hosts, for example fleas and lice of various terrestrial vertebrates, and Monogenea and Copepoda of freshwater and marine fishes. Endoparasites are parasites that live in the tissues and organs of their hosts, such as tapeworms, flukes, and protozoans of vertebrates. An obligate parasite is a parasite that cannot survive without a host, such as the malaria parasite, and a facultative parasite is a parasite that can also live without a host, such as maggots, which normally are saprophagous but can infect living hosts as well. A permanent parasite is associated with a host for long periods, whereas a temporary parasite is found in or on a host only for short periods: examples of the former are human helminths and blood protozoans, and examples of the latter are mosquitoes and leeches that visit
hosts for blood-sucking only for short periods. Larval parasites, like the praniza larva of isopods, are parasitic only at a larval stage, and adult parasites, to which most metazoan parasites belong, are associated with a host during part or the whole of their mature phase. Periodic parasites (leeches, mosquitoes) visit a host at intervals. Intraspecific parasites parasitize individuals of the same species; for example, males of certain deep-sea fish are permanently attached to females of the species and derive food from them. Hyperparasites (of the first, second, etc. degrees) are parasites living on or in other parasites.

An example of a hyperparasite of the first degree is Udonella, a monogenean parasitic on copepods, which themselves parasitize marine fishes. Microparasites, i.e., protozoans, bacteria, viruses, and some helminths, are small, with short generation times; they reproduce on or in a host usually at high rates, the duration of infection is usually much shorter than the life span of the host, and they induce immune responses in vertebrates. Macroparasites, i.e., most helminths and arthropods, do not reproduce on or in the host, they have longer generation times than microparasites, immune responses are lacking or weak and depend on infection intensities, and infections are often chronic and lead to morbidity rather than mortality. Parasitoids, many species of Hymenoptera, lay their eggs into insect hosts which may survive for some time but are invariably killed by the growing larvae of the parasitoids. Hence, parasitoids are predators rather than genuine parasites.

III. ADAPTATIONS TO PARASITISM

A. Size of Parasites

Most parasite species are much smaller than their hosts. Malaria parasites, for example, are microscopic protozoans, and human pinworms are less than 1 cm long. Nevertheless, some species reach a remarkable size. For example, a didymozoid trematode infecting the sunfish, Mola mola, reaches a length of 12 m, although its diameter is very small and the volume of the parasite is still much smaller than that of the fish, which reaches a weight of 1 ton. Likewise, the broad fish tapeworm, Diphyllobothrium latum, which lives in the intestine of various fish-eating mammals, including man, reaches a length of over 10 m, but its volume is much smaller than that of its hosts. At first glance, perhaps surprisingly, parasites often are considerably larger than their free-living relatives. This phenomenon is clearly shown in flatworms, phylum Platyhelminthes. Most free-living flatworms are very small, from less than 1 mm to a few millimeters long, whereas parasitic flatworms such as flukes and tapeworms, as a rule, are much larger, up to several centimeters in length for the flukes and many meters long for the tapeworms. There may be two reasons for this. First, parasitic flatworms in the organs and tissues of their hosts have a much richer and consistent food supply than do free-living species, and food supply therefore does not present a limit to size. Second, selection may have favored multiple and larger gonads and, therefore, a large body size because parasites have to produce many offspring in order to overcome the hazards involved in infecting other hosts. But selection may also favor smaller body size, because parasites depend on living hosts and it is important that hosts survive at least until the parasite has produced offspring that can infect other hosts. A smaller size of parasites relative to that of their hosts is therefore of advantage.

B. Reduction and Increase in Complexity

A general misconception is that all parasites have a less complex structure than free-living forms, a phenomenon that has been named sacculization after the parasitic barnacle Sacculina, which infects marine crabs and
which indeed shows a remarkable reduction in complexity: The only parts of the parasite visible on the outside of the host are the so-called externa, a saddle-like structure containing the gonads. Most of the parasite consists of an extensive system of cytoplasmic processes reaching into the various host tissues (Fig. 1). All the crustacean characteristics have been lost, and only the free-living larval stage indicates that the parasite is indeed a crustacean, related to barnacles. However, in most parasites such a reduction in complexity has not occurred, and in many species there is, in fact, an increase in complexity. A well-studied example of the latter case is the trematode *Lobatostoma manieri*, a parasite of marine fishes, about 3–5 mm long, which has not only a remarkably complex nervous system with more longitudinal nerve cords than free-living flatworms but also about 20,000–40,000 sensory receptors, belonging to about a dozen different types distinguishable under the electron microscope (Fig. 2). These numbers are much larger than those of most related free-living flatworms, and this in spite of the fact that there is not a single free-living stage in the life cycle of this parasite. The adult worm lives in the small intestine of marine fish, it produces eggs containing an infective larva, the egg is eaten by a snail in which the larva hatches and develops to a stage infective to fish, and fish become infected by eating snails.

C. Increase in Reproductive Capacity

Almost all parasites that have been studied produce a remarkable number of offspring, greater than that of related free-living forms. An example is given in Table 1. Free-living turbellarians produce the fewest offspring of any flatworms, ectoparasitic Monogenea produce more, and the highest number of offspring is produced

| TABLE 1 | Approximate Estimates of Fecundity of Free-Living and Parasitic Flatworms* |
|---------------------------------|--------------------------|---------------------|
| Free-living Turbellaria          | 10                       | None                |
| Ectoparasitic Monogenea          | 1000                     | None                |
| Endoparasitic trematodes         | 10 million               | ×1000 at least      |
| Endoparasitic tapeworms          | 10 million               | ×1–1000             |

* After Jennings, Calow, and Rohde, from Rohde (1993).
by endoparasitic trematodes and cestodes. Most trematodes and some tapeworms not only produce large numbers of eggs, but there is a secondary increase in numbers of offspring in the intermediate hosts by asexual or parthenogenetic reproduction. Thus, a single egg of a trematode can produce thousands of cercariae, i.e., larval stages infective to the final vertebrate host. A prerequisite for this increased reproductive capacity is the safe and rich food supply available to parasites, as well as the relatively large body size of many parasites compared with their free-living relatives (see above). However, it is likely that selection has favored an increased reproductive output, because the hazards encountered by parasites in their often complex life cycles are enormous. Very few larvae will survive and establish an infection.

D. Mechanisms of Dispersal

Dispersal is important for any species, whether free-living or parasitic, because a population restricted to one small area risks becoming extinct if conditions become unfavorable and because dispersal reduces in-breeding and the loss of evolutionary adaptability. For parasites, a third point is important: dispersal may reduce the chances of hosts becoming overinfected. Three aspects of dispersal are important: dispersal over short distances away from an individual host, dispersal in space and range extension over larger distances, and dispersal in time. Trematode larvae illustrate that all three aspects of dispersal can be brought about by the same stage. Larvae ( cercariae) are often forcibly ejected in the respiratory currents of the snails in which they have developed, bringing about dispersal away from the host. They actively swim and keep aloft by means of their tails and can thus be dispersed over long distances by water currents. In many species special flotation devices of the tail prolong duration of floating (Fig. 3). Adult flukes produce eggs, and larvae in the snail hosts are produced over long periods, months or even many years, leading to dispersal in time.

E. Mechanisms of Infection

It is essential for a parasite to ensure entry into a host, and this is achieved by an amazing variety of mechanisms. Table II lists examples of infection mechanisms used by human parasites. Many parasite species have evolved remarkable behavioral adaptations that facilitate transmission to a host. For example, microfilariae, i.e., larvae of various species of filariae (nematodes), circulate in the peripheral blood of vertebrates, where they are ingested by mosquitoes for further development. Depending on the activities of the mosquito species involved, the microfilariae appear in the peripheral blood either during the day or the night, and strains of the same species may be nocturnal in one area and diurnal in another. Larvae of some monogeneans have endogenous hatching rhythms (rhythms not induced by external factors) adapted to the hosts' behavior. Some species (and strains) hatch in the evening, others in the morning, whenever chances to infect a host fish are greatest. Perhaps most remarkable are the adaptations
of some trematodes. For example, the liver fluke of sheep, *Dicrocoelium lanceolatum*, uses terrestrial snails as the first and ants as the second intermediate hosts. Cercariae produced in the snails aggregate in slime balls inside the snails, are expelled by the snails, and are eaten by ants. The first cercariae that enter an ant migrate into the subesophageal ganglion, inducing spastic behavior of the ant, which makes it cling to a plant, where their chances to be eaten by sheep are enhanced. The trematode *Leucocysterium macrostomum* has a larval stage, the sporocyst, which forms colorful outgrowths that extend into the snail’s tentacles. They pulsate rhythmically, mimicking worms, the natural food of small birds, which are the final hosts. Birds bite off the tentacles and become infected. Infection also induces snails to move to more exposed sites.

**F. Aggregation**

The distribution of organisms in space may be uneven (more or less equal distances between individuals), random (random distances between individuals), or aggregated (or clustered, or overdispersed; individuals occur in clusters). Applied to populations of parasites, as a rule, parasites show an aggregated distribution in host populations; some individuals are heavily infected, but most are very lightly infected or not at all. Very often, such distributions can be described best by a negative binomial distribution. Reasons for aggregated distributions are manifold: a series of exposures, each with different chances of infection; nonrandom distribution of infective stages; increase or decrease in chances for further infections by a first infection; variations in susceptibility of host individuals; changes of infection of individual hosts over time. Selection may even have favored aggregation in order to restrict damage to a few heavily infected individuals or to facilitate mating, but experimental evidence does not exist.

**G. Hermaphroditism, Parthenogenesis, and Asexual Reproduction**

Contact between parasites and hosts is usually sporadic and, in most cases, only single or a few parasites will manage to infect a host. It is important that populations can be built up from these individuals. A single individual can produce large populations by parthenogenetic or asexual reproduction. In the first case, gametes develop without fertilization; in the second, somatic cells develop. An example of the second case is malaria parasites developing by schizogony in human red blood cells. An example for the first case is (probably) trematode larvae developing in snails. Almost all trematodes, as well as the cestodes and monogeneans, are hermaphroditic, thereby doubling the chance of meeting a mating partner. Furthermore, some species can at least sometimes self-fertilize.

**H. Host Specificity**

All parasites are restricted to certain host species, i.e., they cannot infect all species available, although some
parasite species are more restricted than others. For example, the human pinworm, Enterobius vermicularis, is found only in humans, whereas the protozoan Toxoplasma gondii has been found in a wide range of mammals and birds.

Some parasites, although found in many hosts, nevertheless infect a single or a few host species much more strongly than others. It is therefore useful to distinguish host range (the number of host species infected irrespective of how frequently and how strongly they are infected) and host specificity (taking frequency, or prevalence, and intensity of infection into account). A parasite that infects many host species, but one of them much more strongly than the others, may have a greater host specificity than a parasite species that infects fewer host species, but all of them strongly. Indices to measure host specificity are available (Rohde, 1993).

I. Site Specificity

There is no "universal parasite" that infects all organs and tissues of a host equally, but the degree of site specificity within or on a host varies greatly. Entamoeba histolytica infects not only the intestine of humans but also the liver, lungs, brain, and other organs, whereas adult schistosomes are restricted to the blood vessels. Site specificity is extreme in many Monogenea on the gills of fishes. Different species occupy different parts of the gill on the same host (Fig. 4).

J. Simple and Complex Life Cycles

Many parasites use a single host; i.e., they have a direct life cycle. Others use a final (definitive) host as well as one or several intermediate hosts; i.e., they have indirect life cycles. Parasites with direct life cycles include fleas and lice on various vertebrates as well as many intestinal nematodes and Monogenea on the skin and gills of fish. Usually, the adult stage is parasitic. It produces eggs or larvae that infect the same or other host individuals. In the case of some isopods, the only parasitic stage is the so-called praniza larva: adults live on the sea floor, larvae attach themselves to the gills of fish, suck blood, and drop off to mature in the benthic environment. Parasites with indirect life cycles include the trematodes, some with a single intermediate host, others with several. An example of the former is the aspidogastrean trematode Lobatostoma manteri: the adult infects the intestine of marine fish; eggs are produced and shed in the feces; they are eaten by marine snails, in which the larvae hatch and grow to (almost) adult body size; and fish become infected by eating snails (Fig. 5). An example of a trematode with two intermediate hosts is the

FIGURE 4 Monogenean gill parasites on the gills of mackerel, Scomber australasicus, off southeastern Australia. P, pseudobranch; 1–4, gills 1–4; ext, external gill filaments; int, internal gill filaments. A, Kuhnia sproti; B, Kuhnia scrobilis; C, Kuhnia scomberecola; D, Gruberia australis; x, Pseudokuhnia mina. Note that species A–D have identical copulatory organs, and species x has different copulatory organs. A–D are spatially segregated from each other, and species x overlaps with B, C, and D.
Chinese liver fluke, *Opisthorchis* (Clonorchis) sinensis. Adults infect the liver of fish-eating mammals, including man; snails are the first and fish the second intermediate hosts. An example of a trematode with three intermediate hosts is the bird fluke *Strigea falconispalumbi*. Adults live in predatory birds, the first intermediate hosts are snails, the second intermediate hosts are tadpoles/frogs, and the third intermediate hosts are amphibians, snakes, mammals, and birds, which are eaten by the final hosts (Fig. 6). The life cycles of some trematodes with four hosts can be extended by incorporating various transport hosts in which larvae accumulate but do not develop further. Nematodes with indirect life cycles include filariae which are transmitted by blood-sucking dipterans. Some nematodes with direct life cycles include species whose larvae undergo a peculiar migration through the body of the host. The human roundworm *Ascaris lumbricoides*, for instance, produces eggs that are swallowed by the host. They hatch in the digestive tract but do not grow there to the adult stage; instead they penetrate the wall of the intestine, invade the blood vessels, are carried into the lungs, penetrate through the alveoli, migrate up the trachea, are swallowed a second time, and now develop to the adult stage in the intestine. Reasons for this curious phenomenon are not known, but it may reflect a different kind of life cycle in the evolutionary past. Hookworms, which infect their hosts by penetrating through the skin, undergo a similar body migration involving passage through the lung alveoli and final maturation in the intestine.

**K. Some Physiological Adaptations**

Parasites belong to a wide range of taxa and infect a wide range of hosts. Hence, their physiological adaptations are of a great variety as well and cannot be discussed in any detail here. Jennings (in Rohde, 1997) has made detailed comparative studies of free-living, ecto- and endoparasitic flatworms. These studies are particularly informative because they show the transition from free-living to commensal to ecto- and endoparasitic forms. Some free-living turbellarians living on the body surface and in the gill chambers of their hosts feed on the same kind of food as their free-living relatives, but in addition, they also feed opportunistically on food scraps of the host. Food reserves and digestive physiology do not differ from those of free-living species. Turbellarians living in the interior of mollusks, arthropods, and echinoderms have become increasingly dependent on their hosts. Some feed on protozoans also found in the hosts as well as on food ingested by the host, intestinal cells, and coelomocytes. Others feed mainly on intestinal cells, using some digestive enzymes from the host, and others entirely lack digestive enzymes and depend on enzymes ingested with host tissues. Most of the endoparasitic turbellarians do not
IV. ORIGINS OF PARASITISM AND COMPLEX LIFE CYCLES, THE EVOLUTION OF VIRULENCE, AND COEVOLUTION OF HOSTS AND PARASITES

A. Origins of Parasitism and Complex Life Cycles

Few fossil parasites are known. They include schistosome eggs from ancient Egyptian mummies a few thousand years ago and galls on the arms of feather stars, probably produced by Myzostomida (parasitic annelids) from the Silurian and Devonian periods, 350–430 million years ago. Conclusions on the origins of parasitism and parasite life cycles must therefore be based on inferences from comparative studies of extant species. The Platycladida have been studied most thoroughly, using DNA studies of several genes, in particular 18-S rDNA and 28-S rDNA, and phylogenetic systematics (cladistics). Phylogenetic systematics seeks to establish branching patterns in phylogeny on the basis of shared acquired characters (synapomorphies). Ultrastructural characters are of particular use because of their complexity. Cladistic and DNA studies by many authors have consistently shown that the Neodermata, the major groups of parasitic flatworms (Trematoda, Monogenea, and Cestoda) all share a common ancestor, i.e., are monophyletic. The Neodermata, or the Neodermata plus some of the parasitic turbellarians, are the sister group of a very large taxon, including most Turbellaria, with which they share a common ancestor. This means that the parasitic groups evolved very early in evolutionary history. Among the Neodermata, the trematodes are the sister group of the other Neodermata and, among the trematodes, the Aspidogastrea are the sister group of the other trematodes, or Digenea. Three of the four families of Aspidogastrea occur in eelworms, whereas almost all digeneans parasitize teleost fishes, amphibians, reptiles, birds, and mammals; very few species of Digenea have been recorded from eelworms, to which they have secondarily adapted. Fossil records indicate that eelworms are 450 million years old and teleosts are 210 million years old. This suggests that the Aspidogastrea are the oldest extant

store lipids (as free-living and ectocommensal forms do), but they store glycogen instead. Glycogen storage is also characteristic of the Neodermata (the major groups of parasitic flatworms, including the Monogenea, Trematoda, and Cestoda). Some endoparasitic turbellarians possess physiologically active hemoglobins, which permit preferential abstraction of oxygen from host tissues.

L. Adaptations of Flowering Plants

Parasitic plants have haustoria, which form a close connection with the vascular system of host plants, either in the roots or in the shoots. They depend entirely or partly on the host for water and inorganic and organic solutes. "Hemiparasites," i.e., plants only partly dependent on the host, have chlorophyll, whereas "holoparasites," i.e., plants entirely dependent on the host, lack chlorophyll.
trematodes and possibly neodermatans. It also suggests that the simple life cycle of Aspidogastrea (Fig. 5) is the original life cycle of trematodes, including a final and an intermediate host (which in some species is not obligatory), without multiplication of larval stages in the intermediate host. The complex life cycles of digenean trematodes, including a final and at least one, and up to three, intermediate hosts and several transport hosts, with multiplication of larval stages in the first intermediate host (Fig. 6), may have evolved from this primitive kind of life cycle to make transmission to the final host more effective.

Among the crustaceans, most barnacles are free-living, attached to rocks or other hard substrata. Some barnacles live in a phoretic association, for instance with whales, attached to their skin and feeding on plankton in the environment. Other, closely related species have become parasites. Thus, *Anelasma* parasitizes the skin of sharks, processes of its stalk branching in the host's muscles and extracting food from it. All the approximately 120 species of rhi zocephalans, also related to the barnacles, are parasites and strongly modified in adaptation to their way of life (see, for example, *Sacculina*, Fig. 1). This suggests that parasitism in the barnacles may have evolved from free-living to phoretic to parasitic.

**B. Evolution of Virulence**

Intuitively, one might suspect that it is not in the parasite's interest to severely damage or even kill its host, because this would also affect the fitness of the parasite. On the other hand, parasite transmission to another host may well be facilitated by such damage to an intermediate or transport host. In other words, evolutionary pressure may have led to an increase in virulence in some cases and to a decrease in others. Anderson and May have developed an epidemiological model that considers virulence, as follows:

$$R_0 = \beta(N)/(\mu + a + v).$$

$R_0$ is the fitness of the parasite, its lifetime reproductive success; $\beta$ is the rate of transmission of the parasite by its host to other hosts, which is dependent on host density $N$; $\mu$ is the mortality rate of uninfected hosts; $a$ is the virulence or the morality rate induced by the parasite; and $v$ is the recovery rate of the host. This model (applied and further developed by various authors) permits some predictions on the evolution of virulence under different conditions. For example, a parasite that can achieve a large increase in transmission rate by a small increase in virulence should have lower optimal virulence than one that achieves less by such an increase. The transmission rate depends, among other factors, on the likelihood of transmission for instance facilitated by the behavior of the host. Low virulence will evolve when contacts between hosts and thus opportunities for transmission are frequent. High virulence may evolve in vector-transmitted infections, and optimal virulence may differ in parasites that are vertically transmitted (from parents to offspring) from those that are horizontally transmitted. Although the parameters in the model are difficult to measure, some of the predictions have been verified empirically. Thus, parasites transmitted by contact (e.g., lice) are usually less virulent than those transmitted by vectors (e.g., malaria), and sexually transmitted parasites (e.g., *Trichomonas vaginalis*) are usually less virulent and longer lasting than parasites transmitted in other ways. In vertically transmitted nematodes of fig wasps and in vertically transmitted microsporans of mosquitoes, virulence is less than in horizontally transmitted related species.

**C. Coevolution of Hosts and Parasites**

The comparison of phylogenetic trees of hosts and parasites has given contradictory results concerning possible coevolution. Reasons may be biological differences between the groups studied, or they may be due to different interpretation of results by different authors. For example, according to several authors, chewing lice infecting rodents of the family Geomyidae, as well as lice of seabirds, seem to have coevolved closely with their hosts. In contrast, lice of rock wallabies in Australia, apparently, have switched hosts frequently and there is little evidence for coevolution. Altogether, knowledge is insufficient to state how common coevolution is.

**V. HOST-PARASITE INTERACTIONS**

**A. Cleaning Symbiosis**

Hosts use a variety of behavioral methods to rid themselves of parasites. Preening of birds, bathing of birds in dust and water, and passive and active anting (letting ants passively crawl over the body or actively squeezing ants over the plumage, respectively) are thought to help in reducing parasite loads, although evidence is scarce. Some other behavioral patterns (dolphins rubbing against rocks, fish jumping out of the water, etc.) may also play a role. Most widely distributed and well studied,
particularly in the marine environment, is cleaning symbiosis, in which one animal (the cleaner) cleans another (the host), removing its parasites and diseased tissues. Birds remove ticks and other ectoparasites from cattle, hippopotamuses, and even large marine fish floating on the surface, and several species of shrimps, as well as over 100 species of fish, are cleaners in freshwater but mainly in the sea. Beside fishes, the Galapagos marine iguana, whales and dolphins, invertebrates, etc. are hosts to cleaners. Cleaner fish often have special morphological adaptations, such as a terminal mouth and fused anterior teeth and conspicuous color patterns, so-called guild signs. In addition, the Indo-Pacific cleaner wrasse, Labroides dimidiatus, performs a cleaning dance that attracts host fish. Hosts, on their part, show invitation postures signaling to the cleaner that they are ready to be cleaned, and fish of many species, usually hostile to each other, queue peacefully up at "cleaning stations" (territories where cleaning occurs). They even allow cleaners to enter their mouth cavity. In well-established cleaning symbioses, hosts rarely or never eat cleaners. Some fish were observed to spend as much time at cleaning stations as they spend on feeding, and some cleaner species feed exclusively on parasites and diseased host tissue. Certain fish mimic the color pattern and behavior of cleaners to approach hosts in order to attack them and bite off pieces of fin or skin.

B. Immune and Tissue Reactions, Resistance

Defense reactions to parasites at the humoral and tissue levels are based on the ability of the host to distinguish self (its own cells) from nonself (foreign cells and material). Vertebrates have three types of such reactions: phagocytosis, inflammation, and adaptive immunity. The first two are nonspecific tissue reactions, and the third is specific to a certain type of nonself material. They are found in all vertebrates, but (particularly immunity reactions) are best developed in birds and mammals. All three defense mechanisms usually interact and occur in most tissues and organs. Typically, the reactions occur in a certain sequence: degeneration or necrosis of cells due to the infection leads to an inflammatory response with edema (swelling of tissue). Phagocytic cells engulf small parasites, but if parasites are not eliminated, a chronic inflammation develops, leading to a connective tissue capsule around the parasite; macrophages in the capsule engulf damaged cells and often the parasite. Special defense mechanisms are active on the surface. Thus, fish continually shed mucoid material from the skin even if uninfected, but in infected fish the slough increases and leads to the removal of monogenean and other ectoparasites.

Immune reactions are induced by antigens of the parasite that lead to the formation of specific antibodies in the host. In microparasites (protozoans and some helminths, e.g., Strongyloides stercoralis, but particularly in bacteria and viruses, which are not discussed here), immune responses are more effective than in macroparasites (most helminths and arthropods), where they are either lacking or only short-lasting. Immune responses have been particularly well studied in trypanosomes. The antigen is a surface glycoprotein that completely covers the parasite. The host's antibodies eliminate most parasites, but a few trypanosomes of different antigenic types survive and build up a new population. This is repeated over and over again, leading to marked fluctuations in infection intensities. The host finally dies, because the immune reactions do not suffice to destroy the whole parasite population. The number of antigenic types is very large and apparently limited only by the host's life span: more than 100 types were shown to develop in a single clone strain. In parasites with complex life cycles, such as the malaria parasite, each stage in the life cycle has different antigenic properties, and there are many variants because of the large numbers of daughter cells produced by schizogony. These are the reasons that attempts to develop effective vaccines have failed so far.

Invertebrates, apparently, cannot acquire specific immune responses; they rely on phagocytosis and capsule formation. Pearl formation in bivalves, for example, is the result of encapsulation of foreign bodies.

Hosts show different degrees of resistance to infections that are not due to acquired immunity. For example, some sheep may be less susceptible to nematodes than others because of their genetic makeup, and individuals of the same species of different age may differ in susceptibility. Older individuals are often less infected than young ones, a phenomenon referred to as age resistance. For example, the cestode Austrantilina elongata uses turtles as final and crayfish as intermediate host, but only young crayfish can be infected experimentally: the cuticle of older individuals prevents successful penetration of the larva.

C. Effects on Host Individuals and Populations

There is a great variety of effects on host individuals, depending on the parasite and host species, site of infection, virulence of parasites, and susceptibility of hosts.
A few examples of human parasites may illustrate this. The mite *Demodex folliculorum*, intestinal nematodes at low infection intensities, *Toxoplasma*, etc. often do not cause symptoms. On the other hand, *Toxoplasma* and the nematode *Onchocerca volvulus* can lead to blindness, filariae can cause elephantiasis, a sometimes enormous swelling of scrotum, legs, and arms, etc. Malaria causes a range of symptoms from influenza-like to death. Hookworms, depending on infection intensity, can cause severe anemia and death. Cysticerci of *Taenia solium*, infecting the brain, may lead to epilepsy-like symptoms and death. Nematodes of sheep often cause death, and monogeneans may severely damage the skin, fins, and gills of fish, particularly in aquaculture. Larval trematodes often cause partial or complete castration of their snail hosts, and they sometimes induce gigan-tism in the snails, i.e., a markedly larger body size of infected than uninfected snails.

Models predict that microparasites should be highly effective in controlling host populations, and this is indeed often the case. For example, trypanosomes and malaria severely affect human populations and the former also livestock, and oyster beds were decimated by various protozoan parasites in several countries. However, almost all such reports deal with populations in abnormally high densities (humans, oysters) or with populations affected after introduction of a parasite into an area where it was not present originally or after introduction of a host species into a new area (livestock in Africa). The same applies to macroparasites. Host populations under natural conditions may sometimes be severely affected by macroparasites, but evidence is usually circumstantial and mass mortalities may not be due to parasites alone, but due to synergistic effects involving, for instance, environmental degradation, as well as parasites. Best documented are cases of mass mortalities caused by parasites in livestock and aquaculture, as well as after introduction of parasites or hosts into a new habitat. Of historical interest is an epizootic caused by the liver fluke *Fasciola hepatica*, which caused the death of 3 million sheep in Great Britain in 1879/1880 and led to the clarification of the life cycle of this parasite. A well-documented case of an epizootic due to the introduction of a parasite is that of the monogenean *Nitzchia sturnionis*, which was introduced into the Aral Sea with sturgeon from the Caspian Sea in the early 1930s. It devastated the local sturgeon population and led to the collapse of the sturgeon and caviar industry in the Aral Sea in 1937, which did not recover for 20 years. Recently, the first experimental proof was given that a gastrointestinal roundworm can indeed regulate wildlife populations of red grouse in England.

VI. THE ECOLOGICAL NICHES OF PARASITES

A. The Niche Concept: Niche Dimensions of Parasites

A niche is defined as the total of an organism's relations to its biotic and abiotic environment. These relations define the organism's place in nature, or, in ecological jargon, its place in "multidimensional niche space." Important niche dimensions of parasites are hosts, microhabitats, geographical range, sex of host, age, season, and food. Different parasite species use different host species and different microhabitats within or on the host. Thus, particular nematodes of the cat use different microhabitats (Table III), and even within each microhabitat, there is further subdivision, some species using certain parts of the small intestine, the stomach, etc. Male lambs in the United States have more nematodes of certain species than females, and many parasites prefer hosts of a certain age, occur only at certain seasons, or are restricted to host populations that use certain food.

B. Saturation of Niches with Parasites

An important question in ecology is whether habitats are saturated with species or whether "empty niches" exist. The most extensive and intensive studies were made of parasite communities on the heads and gills of marine fish. Results on 112 host species showed that maximum component species richness was 27 parasite species but that most species had less than six (Fig. 7). Similar differences were found for abundances (intensities of infection with all parasite species), which ranged from less than five for most host species to over 3000, with no apparent signs of damage in the most heavily infected fish. This strongly suggests that most fish species, at least, could accommodate more parasite species and individuals, and even in fish with the greatest parasite species richnesses and abundances, some parts of the gills that are occupied in other species were not infected.

C. Proximate and Ultimate Causes of Niche Restriction

Proximate causes are physical and/or chemical factors determining niche selection of a parasite, whereas ultimate causes refer to the biological function of niche
TABLE III

Nematodes of the Cat: Sites of Infection

<table>
<thead>
<tr>
<th>Site</th>
<th>Species of Nematode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidney</td>
<td>Diocotephysa renale</td>
</tr>
<tr>
<td>Colon</td>
<td>Strongyloides mystax, S. steneciens, Trichuris canicola</td>
</tr>
<tr>
<td>Stomach and small intestine</td>
<td>Abbreviata geminata, Ankylostoma braziliense, Gastrostoma spinigerum, Mastophorus mars, Othelanus tricuspis, Physaloptera brevispiculum, P. jellida, P. paciae, P. cantis, Rictularia cahriensis, Soboliphyma basurini, Spirura rhipheutites, Toxascaris leonina, Toxascaris canis, T. mystax, Trichinella spiralis (adults), Urinauria stenocephala, Ankylostoma tubaforme, Physaloptera praestuatae</td>
</tr>
<tr>
<td>Skin</td>
<td>Dirofilaria repens</td>
</tr>
<tr>
<td>Lungs</td>
<td>Aebrasomystlyus abstrusus, Anphyloides rostratus, Capiliaria aequincola, Triangolymphus subcrenatus, Vogeloides massinoti, V. tamanjachairi</td>
</tr>
<tr>
<td>Heart</td>
<td>Dirofilaria immitis</td>
</tr>
<tr>
<td>Middle ear</td>
<td>Mammonemogonimus auris</td>
</tr>
<tr>
<td>Spinal cord</td>
<td>Garetia paralysans</td>
</tr>
<tr>
<td>Lymph vessels</td>
<td>Brugia malay</td>
</tr>
<tr>
<td>Diaphragm and other striated muscles</td>
<td>Trichinella spiralis (juveniles)</td>
</tr>
</tbody>
</table>

selection: why has selection favored one niche over another? There are many proximate causes responsible for niche selection of different parasites, although they have been poorly studied. Parasite larvae, for example, may use chemical stimuli to locate a particular host and, even if they can infect many hosts, they may survive only on the "correct" ones because a factor or factors produced by the "wrong" hosts may kill them. Exsheatment and development of parasite larvae of different species may depend on different stimuli, redox potential, and pH in different parts of the intestine, thus determining microhabitat differences. Males and females of a host species may simply acquire different parasites because of different feeding habits etc. Ultimate causes may be extrinsic, due to interspecific effects, or intrinsic, due to intraspecific effects. Among the interspecific effects, interspecific competition is generally considered to be the most important one, as evidenced by occasional exclusion of one species by another, microhabitat shifts in the presence of other species, etc. However, altogether evidence is scarce. Many cases of microhabitat differences that have been explained by interspecific competition are likely to be the result of reinforcement of reproductive barriers. Thus, one the monogeneans that have identical copulatory organs are spatially segregated in different microhabitats, whereas species with dissimilar copulatory organs coexist in the same microhabitat (Fig. 8, also Fig. 4), suggesting that microhabitat segregation does not have the function to avoid competition but hybridization between closely related species. Only one intrinsic factor has been suggested: facilitation of mating. Many parasites live at low infection intensities and prevalences (Fig. 7). Restriction to certain hosts and microhabitats may therefore vastly increase the chances of meeting a mating partner. Experimental evidence, however, is scarce.

VII. THE STRUCTURE OF PARASITE COMMUNITIES

A. Concepts of Community Ecology

Community ecology deals with the relations between organisms in a certain habitat, in the case of parasites with relations between parasites infecting a certain host. Such relations can be studied at different levels, the level of infracomunity, component community, and compound community. Infracommunities of parasites consist of all the infrapopulations within a host individual. Component communities consist of all infrapopulations within a host population, and compound communities consist of all parasite communities within an ecosystem. An infrapopulation is the total of all the individuals of a parasite species within a host individual. Two of the major questions of community ecology are (1) whether communities show predictable patterns of
species composition, relative abundances, and resource use and (2) what processes are responsible for the patterns. Studies of endo- and ectoparasite communities of various vertebrate hosts have led authors to distinguish interactive communities, characterized by species regularly occurring in great densities with much interaction between them, and isolationist communities, characterized by species occurring at low densities with little interaction between them. Both types, supposedly, are extremes at the ends of a spectrum, with many intermediate kinds of communities between them. Authors also distinguished between core and satellite parasite species, the former dominant species with high prevalences and intensities of infection and the latter with low prevalences and intensities of infection.

B. Empirical Evidence

Even in host species with extremely species-rich parasite communities, core and satellite species usually cannot be distinguished, and evidence indicates that many more parasite species than actually present could be accommodated (i.e., that vacant niches exist; see above). This, and the frequency of positive and the scarcity of negative associations between species, scarcity of nestedness, greater intra- than interspecific aggregation, and no or
only minor effects of the number of parasite species on microhabitat size, strongly suggest that interspecific competition is not of great importance (see above). Concerning the structure of communities as indicated by predictable co-occurrence and relative abundance of species, studies of ectoparasite communities of marine fish showed that the most dominant species usually represented between 60 and 80% of all parasites in an infra-community but that different species were dominant in different infra-communities. For communities of larval trematodes in snails, Sousa (see Rohde, 1993) has shown that there is a significant difference between communities studied at different scales. Interactions are important at the infracommunity level, but at the level of component community, interactions are not important and communities are largely structured by external processes.

C. Parasite Communities as General Ecological Models

The vast majority of animal species are probably arthropods parasitizing plants, and they share many characteristics with parasites of animals: their habitats are resource-rich and seldom exhausted by the parasites. The major problem for such parasites is not to avoid competition with other species but to find the appropriate hosts and sites for feeding and mating. The conclusions based on the study of parasites of animals can therefore be applied to them.

VIII. PARASITE POPULATION DYNAMICS

A. Concepts of Population Growth

Population growth is described by the logistic growth equation $\frac{dn}{dt} = rN(K - N)/K$. $\frac{dn}{dt}$ is the rate of population growth, $N$ is the number of individuals at the time $t$, $r$ is the per capita rate of natural population increase, and $K$ is the carrying capacity of the habitat (the maximum number of individuals a habitat can support). The equation shows that population growth is exponential when population density is small and that it decreases with increasing $N$ and approaches 0 when $N$ approaches $K$. In a graphic representation, the curve becomes asymptotic, i.e., it flattens out, when population density reaches the carrying capacity. In habitats in which, for example, populations repeatedly become impoverished as the result of external disturbances, species will be favored that can reproduce and develop rapidly, without great investment in the individual offspring. In stable habitats where population size is usually close to the carrying capacity, it pays off to have few but well-adapted offspring. Selection for large numbers of fast-developing offspring is referred to as $r$-selection, whereas selection for few well-adapted offspring is referred to as $K$-selection. However, the general applicability of this distinction has been questioned because population density is only one of the factors that determine selection pressure. Nevertheless, the distinction is still useful as an approximation to the real scenario.

B. Ecological Strategies of Parasites

Most parasites rely on the production of many offspring (see above) and therefore tend to be $r$-strategists. In some cases, numbers are controlled by abiotic factors and not by population size; i.e., the factors are density-independent. This has been demonstrated for some helminth species, for example the cestode Bothriocephalus acheilognathi in the freshwater fish Gambusia affinis, whose prevalence and intensity of infection varied
strongly with water temperature over several years in a lake in North Carolina. However, few such studies have been made and generalizations are therefore premature. Density-dependent factors, i.e., factors dependent on the population size of the parasite, also play a role in some cases. For example, predation by trematode larvae in snails on other trematode larvae, immune responses of hosts that depend on infection intensities (as in many helminth infections), and competition between parasites of the same and of different species are such density-dependent factors. There is experimental evidence for all of these, but studies are too few to permit generalizations on how important such effects are. For example, high infection intensities often lead to stunted growth and reduction in the number of offspring, an example of intraspecific competitive effects. The microhabitat of a parasite species may shift or become smaller if other species are present, or one parasite species may reduce the numbers of another or even completely eliminate it, examples of interspecific competition. However, the importance of interspecific competition, in particular, generally seems to have been overestimated (see above).

Data in Table IV show that 13 large taxa (phyla, subphyla, or classes), some of them very large, consist entirely of parasites, and many other groups include a high proportion of parasitic species. Even some vertebrates are parasitic. Males of a number of deep-sea fishes parasitize females of the same species (Fig. 9). The needle fish, Carapus acus, feeds on the viscera of sea cucumbers, which represent an almost inexhaustible source of food because they regenerate, cleaner numics (fish that imitate genuine cleaners) approach fish and bite off pieces of fin or skin, and gulls were observed feeding on the flesh of whales. Cowbirds and about 50 species of cuckoos are brood parasites; i.e., they lay their eggs into the nests of other birds, which then incubate them. Skuas and frigate birds chase other birds in flight and feed on their regurgitated food ("kleptoparasitism"). Kleptoparasitism in its various varieties is widespread among humans.

Concerning plants, estimates are that 1% of flowering plants, about 3000–4000 species, are parasitic. Parasitism has evolved at least eight times in the angiosperms, and 16 plant families have parasitic species.

X. ZOOGEOGRAPHY OF PARASITES

A. Latitudinal Gradients in Species Richness and Abundance

The most dangerous parasites of man are found in warm countries: malaria (originally also found in temperate countries such as those of northern Europe), filariasis, guinea worm, sleeping sickness, Chagas disease, and several other diseases are restricted to tropical—subtropical countries. However, latitudinal gradients in species richness have been quantified best for parasites of marine fishes. Species number of marine fishes is greatest in the tropics. Species richness of metazoan ecto- and endoparasites also rises toward the tropics. However, interestingly, both the number of ectoparasitic species per host fish (their "relative species richness") and their abundance increase at lower latitudes, whereas such a trend does not exist for the endoparasites (Fig. 10). The reasons for these differences are not understood.

B. Latitudinal Gradients in Reproductive Strategies

Such a gradient has been studied only in marine Monogenea. Most monogeneans produce eggs in which ciliated larvae (oncomiracidia) develop, one family, the
<table>
<thead>
<tr>
<th>Phylum</th>
<th>Subphylum (Protozoa) or class</th>
<th>Total number of species</th>
<th>Number of parasite species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subkingdom Protozoa</td>
<td></td>
<td>70,000</td>
<td></td>
</tr>
<tr>
<td>Sarcomastigophora</td>
<td>Mastigophora</td>
<td>?</td>
<td>2,000</td>
</tr>
<tr>
<td></td>
<td>Opalinata</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>Sarcolemma</td>
<td>?</td>
<td>250</td>
</tr>
<tr>
<td>Labyrinthomorpha</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microspora (Microsporida)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ascetospora (Haplosporida)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ciliophora</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subkingdom Metazoa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mesozoa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dicyemida</td>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orthonecotida</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portiera</td>
<td>1,500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crinidaea</td>
<td>2,700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ctenophora</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myxozoa (Myxosporidia)</td>
<td>6,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platyhelminthes</td>
<td>&gt;1,500</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;1,500</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;ca. 80,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;ca. 65,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;35</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;2,500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gastrotricella</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Priapulida</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enteropodia</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nemertina</td>
<td>750</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nematelminthes</td>
<td>1,300</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>150</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;20,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;20,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;100</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;100</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5,600</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ca. 420</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;4,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>111</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>111</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>250</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annelids</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Many millions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Many millions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Onychophora</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arachnida</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4,300</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylum</td>
<td>Subphylum (Protista) or class</td>
<td>Total number of species</td>
<td>Number of parasite species</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------------------</td>
<td>-------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Pycnogonida</td>
<td>350</td>
<td>&lt;350</td>
<td></td>
</tr>
<tr>
<td>Crustacea</td>
<td>20,000</td>
<td>&gt;2,500</td>
<td></td>
</tr>
<tr>
<td>Myriapoda</td>
<td>10,500</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Insecta</td>
<td>Many millions</td>
<td>Many millions(^a) (&gt;90%)</td>
<td></td>
</tr>
<tr>
<td>Tardigrada</td>
<td>180</td>
<td>1(?)</td>
<td></td>
</tr>
<tr>
<td>Pentastomida</td>
<td>75</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>Nemaculida</td>
<td>ca. 5,000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Phoronidea</td>
<td>18</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Bryozoa (Ectoprocta)</td>
<td>&gt;4,000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Brachiopoda</td>
<td>280</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Mollusca</td>
<td>112,000</td>
<td>&gt;100</td>
<td></td>
</tr>
<tr>
<td>Solenogastres</td>
<td>140</td>
<td>few(?)</td>
<td></td>
</tr>
<tr>
<td>Placophora</td>
<td>1,000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Gastropoda</td>
<td>85,000</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Scaphopoda</td>
<td>300</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Bivalvia</td>
<td>25,000</td>
<td>&gt;10</td>
<td></td>
</tr>
<tr>
<td>Cephalopoda</td>
<td>&gt;600</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Echinodermata</td>
<td>70</td>
<td>Few (intraspecific parasites)</td>
<td></td>
</tr>
<tr>
<td>Echinodermata</td>
<td>250</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Hemichordata</td>
<td>80</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Enteropneusta</td>
<td>60</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Pterobranchia</td>
<td>20</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Echinodermata</td>
<td>ca. 6,000</td>
<td>≥2</td>
<td></td>
</tr>
<tr>
<td>Crinoidea</td>
<td>620</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Holothuroidea</td>
<td>1,100</td>
<td>1(?)</td>
<td></td>
</tr>
<tr>
<td>Echinidea</td>
<td>860</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Asteridea</td>
<td>1,500</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Ophiuroidea</td>
<td>1,900</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Pogonophora</td>
<td>47</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Chaetognatha</td>
<td>50</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Chordata</td>
<td>62,000</td>
<td>Some</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Many millions</td>
<td>Millions</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Groups that consist entirely of parasitic species are printed bold. After various authors. Only extant species are included.

\(^b\) Many authors refer to many of the plant-parasitic insects as herbivorous insects.

Gyrodactyliidae, consists exclusively of small viviparous species. Whereas gyrodactylids are very rare in warm waters, they represent the vast majority of all species in cold waters (75–89% in the Bering, White, and Barents Seas). This corresponds to a trend in marine benthic invertebrates, which produce large numbers of eggs from which pelagic larvae develop in warm waters and small numbers of larvae by viviparity etc. in cold waters (Thorson's rule).

C. Latitudinal Gradients in Host Ranges and Host Specificity

Such trends have not been studied and are not obvious in terrestrial and freshwater parasites. Studies of trematodes and monogeneans of marine fish have shown that the former have greater host ranges, i.e., they use more host species, in cold than in warm waters whereas monogeneans do not show such a trend, using very few
ties of infection in different host species into account) is the same at all latitudes for both groups.

D. Parasites as Biological Markers

Parasites are frequently used as biological markers to study host populations and migrations, particularly in the sea. A prerequisite for such study is that parasites have a long life and are acquired only in the population or area studied. Larval helminths fulfill these conditions and are therefore usually used. Examples for the successful use of parasites as biological markers are studies that demonstrated different populations of herring with different Anisakis infections in the Baltic Sea, different populations of Atlantic salmon from different tributaries of the Miramichi River system in Canada, and different populations of sockeye salmon in the North Pacific, one harboring the nematode _Dactylopis truttae_ acquired in Kamchatka.


TABLE V
Human Parasites with the Most Important Parasites Worldwide

<table>
<thead>
<tr>
<th>Parasite Infection</th>
<th>Number infected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intestinal roundworms</td>
<td>1400 million</td>
</tr>
<tr>
<td>Malaria</td>
<td>300 million</td>
</tr>
<tr>
<td>Bilharzia</td>
<td>200 million</td>
</tr>
<tr>
<td>Lymphatic filariasis</td>
<td>120 million</td>
</tr>
<tr>
<td>Amebiasis</td>
<td>90 million</td>
</tr>
<tr>
<td>Food-borne trematodes</td>
<td>40 million</td>
</tr>
<tr>
<td>Chagas disease</td>
<td>16–18 million</td>
</tr>
<tr>
<td>Leishmaniasis</td>
<td>12 million</td>
</tr>
<tr>
<td>Sleeping sickness</td>
<td>0.3 million</td>
</tr>
<tr>
<td>Guinea worm</td>
<td>0.1 million</td>
</tr>
</tbody>
</table>


and one harboring the larval cestode *Taeniaeformis crusidis* acquired in western Alaska.

E. The von Ihering Method

About 100 years ago, von Ihering first used parasites and ectocommensals to clarify places of origin and dispersal of hosts and ancient land connections between land masses. One assumption of the method is that animals have acquired the greatest diversity in the area where they have been longest. Mamaev and Margolis applied the method to Pacific salmon. On the basis of their finding that salmonids harbor many freshwater parasites, some of them with complex life cycles, but no marine endoparasites, they concluded that salmonids are of freshwater origin. Many similar studies have been made.

XI. ECONOMIC AND HYGIENIC IMPORTANCE OF PARASITES

A. Human Parasites Are among the Most Important Disease Agents of Man

The web pages of The World Health Organization, Division of Tropical Diseases (CTD), and of the Centers for Disease Control (CDC) contain information about the current status of the important parasitic diseases, which is continually updated. Information about the most important parasites of man are given in Tables V and VI. Note that some of the most widespread species, such as *Demodex folliculorum*, a mite infecting the skin, and *Toxoplasma gondii*, are not included because of lack of data and usually symptomless infections.

A number of parasite species, most of them protozoans, but also the nematode *Strongyloides stercoralis*, have recently become more important because of AIDS. All these species multiply within the human host and may become fatal in immunodepressed patients. Protozoans involved are, among others, *Cryptosporidium* and *Giardia*, sometimes acquired in polluted drinking water. *Pneumocystis carinii*, *Microspora*, and *Entamoeba histolytica*.

TABLE VI
Infections of the Most Common Parasites in Some Countries/Regions

<table>
<thead>
<tr>
<th>Region</th>
<th>Infection</th>
<th>Number infected</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>Pinworms (<em>Enterobius vermicularis</em>)</td>
<td>50 million</td>
</tr>
<tr>
<td></td>
<td><em>Strongyloides stercoralis</em></td>
<td>&lt;1 million</td>
</tr>
<tr>
<td>South America</td>
<td>Biliharzia</td>
<td>45 million</td>
</tr>
<tr>
<td></td>
<td>Chagas disease</td>
<td>16–18 million</td>
</tr>
<tr>
<td></td>
<td><em>Strongyloides stercoralis</em></td>
<td>ca. 7.4 million</td>
</tr>
<tr>
<td>Africa</td>
<td>Ascaris lumbricoides</td>
<td>&gt;200 million</td>
</tr>
<tr>
<td></td>
<td>Biliharzia</td>
<td>100 million (mortality, 0.01 million)</td>
</tr>
<tr>
<td></td>
<td>Malaria</td>
<td>23 million (mortality, 2.6 million)</td>
</tr>
<tr>
<td>China</td>
<td>Ascaris lumbricoides</td>
<td>ca. 100 million</td>
</tr>
<tr>
<td></td>
<td>Opisthorchis sinensis</td>
<td>5 million</td>
</tr>
<tr>
<td>Middle East</td>
<td>Ascaris lumbricoides</td>
<td>100 million</td>
</tr>
<tr>
<td></td>
<td>Hookworms</td>
<td>60 million</td>
</tr>
<tr>
<td></td>
<td>Biliharzia</td>
<td>50 million</td>
</tr>
</tbody>
</table>

Note: *Demodex folliculorum* and *Toxoplasma gondii* not included.
Resistance to drugs used against various parasite species has developed, and to insecticides used against various vectors. Hence, the epidemiological situation is getting worse. Malaria is becoming more prevalent in many countries, partly due to insufficient financial and human resources for control but also because of man-induced climatic and environmental changes, migration, and war. Likewise, the prevalence of intestinal parasites worldwide is increasing, partly due to increasing urbanization. In contrast, the global prevalence of guinea worm (Dracunculus medinensis) has decreased from about 10 million ten years ago to about 150,000 (most of them in the Sudan), as a result of a campaign to eradicate the infection. Almost certainly, malaria and several other important parasite diseases would spread into areas presently not affected if global warming as the result of the greenhouse effect should occur.

B. Parasites of Livestock Also Have Very Great Economic Importance

For example, nagana, a disease of many domestic animals caused by Trypanosoma brucei brucei and transmitted by tsetse flies in Africa, has made large areas of sub-Saharan Africa unsuitable for livestock production, and nematodes of sheep and cattle cause large economic losses in many countries. Ostertagia ostertagi, one of the nematodes infecting cattle, is estimated to cause an annual loss of $600 million to the cattle industry in the United States alone. The situation is not improving
because of drench resistance of nematodes. Thus, in Australia about 90% of sheep farms have worms resistant to one or more drench classes.

C. Parasites Have Repeatedly Caused Large Economic Losses in the Fisheries Industries

In particular, aquaculture is affected. On the East Coast of North America, for example, several protozoan parasites have repeatedly decimated oyster culture in several areas. The haplosporidian *Haplosporidium nelsoni* caused a decline in oyster production in the New Jersey waters of the Delaware Bay from about 5–8 million pounds between 1950 and 1955 to 167,000 pounds in 1960, which has never fully recovered (Fig. 11).

D. Many Parasites, Particularly Nematodes, Are among the Most Important Pests of Plants

For example, wheat is attacked by the nematode *Anguina tritici*, potatoes are attacked by the potato cyst nematode *Globodera rostochiensis*, and rape, rice, etc. all have their specific nematodes that lead to large economic losses. The estimated annual loss to crop production due to nematodes in the United States is $8 billion (12%), and it is $78 billion globally (http://anrwww.umn.edu/son/nm/nema.htm).

E. Parasitic Angiosperms Attack Many Crops

At least 11 species of witchweed, *Striga* spp., attack crops, including all the important tropical cereals; control is difficult because of the high production of seeds and longevity of seeds up to 20 years in the soil. Dwarf mistletoes almost always cause deformities of trees, leading to substantial reduction in yield and quality of timber. The mistletoe *Viscum album* reduced the yield of a particular apple variety in the United Kingdom by 7–54% (Press and Graves, 1993).

F. Parasites, Including Nematodes, Are Increasingly Used to Control Insect Pests

*Neoaplectana carpocapsae* is bred in the laboratory and sprayed on rape to combat insects attacking it, and *Howardula sp.* effectively kills the dried-fruit beetle *Carposphillus mutatus*. *Heterorhabditis bacteriophora* kills insects by carrying bacteria into them.

See Also the Following Articles

COEVOLUTION • HABITAT AND NICHE, CONCEPT OF • PARASITOIDS • SPECIES INTERACTIONS

Bibliography


PARASITOIDs

H. C. J. Godfray
NERC Centre for Population Biology and Imperial College at Silwood Park

I. Life History Variation
II. Host Location
III. Host Acceptance and Oviposition Strategy
IV. Resistance and Virulence
V. Population Dynamics
VI. Community Ecology
VII. The Importance of Parasitoids

PARASITOIDs are a type of animal that have a life history intermediate between that of a predator and prey. As normally defined, parasitoids are invariably insects and their larvae feed at the expense of other insects, with the exception of a few species that attack different types of arthropods or mollusks. The adult female parasitoid lays her egg on, in, or occasionally near the body of the host; the eggs hatch and the developing parasitoid larvae consume the host, eventually killing it. Like a parasite, only a single host is required for full development, but like a predator the host is invariably destroyed. The term parasitoid, originally coined by Reuter at the beginning of the twentieth century, is now used nearly universally to describe species with this life cycle, although in the older literature they are sometimes just called insect or protean parasites. Some solitary wasps have a similar life cycle except that the female parent carries paralyzed hosts to prepared nests or caches, and normally the requirement that hosts are attacked and oviposited on in situ is included in the definition of a parasitoid.

More than 50% of parasitoids belong to the insect order Hymenoptera: the sawflies, ants, bees, and wasps. The least derived suborder are the sawflies (Symphyta), which are nearly all phytophagous, although one family, the Orussidae, contains species whose larvae are parasitoids. From a lineage that was probably related to the Orussidae, the remainder of the Hymenoptera (the Aprocrida) evolved, and these are in turn divided into a huge group: the Parasitica or parasitoid wasps, which are almost exclusively parasitoids, and the smaller Aculeata in which the ovipositor has become a sting. The Aculeata, derived from the Parasitica, are primatively parasitoids but have radiated into a wide range of feeding habitats, including predators, scavengers, and pollen feeders. The Aculeata contain nearly all the Hymenoptera familiar to the general public, including all the species of ants, bees, and wasps that show advanced sociality (eusociality). The Parasitica are classified into a large number of families and superfamilies, the most important of which are the Ichneumonidae, Braconidae, and Chalcidoidea.

The second major group of parasitoids belong to the two-winged flies or Diptera. The parasitoid habit has evolved on many occasions in this order, with the most important taxon being the family Tachinidae. These are almost exclusively parasitoids, with the adults frequently resembling houseflies (Muscidae, like the Tachinidae in the Calyptrata). The parasitoid habit has also evolved on many occasions in the beetles (Coleoptera), although a relative small number of species feed in this way. The most frequently encountered beetle
parasitoids are probably those in the rove beetle family (Staphylinidae). Outside the three insect orders just mentioned, parasitoids are extremely scarce and taxonomically scattered.

As described in more detail later, parasitoids are important members of nearly all terrestrial ecosystems, but there is little agreement about how many species exist. Currently, about 10% of the approximately 1 million described insects are parasitoids, but parasitoids are almost certainly comparatively poorly known compared with other groups—parasitoid taxonomy is legendarily difficult. Most workers today would argue that there are between 4 and 8 million species of insect on Earth, of which 1.5–2 million are parasitoids. To put this into perspective, recall that there are less than 10,000 species of birds and approximately 5000 species of mammals.

1. LIFE HISTORY VARIATION

A simple way to classify parasitoids is by the host stage that they attack. Holometabolous insects have three juvenile stages (egg, larva, and pupa); whereas hemimetabolous insects have nymphal stages with no pupal metamorphosis. The most common category of parasitoids is composed of those that attack larval or nymphal stages, although egg and pupal parasitoids are also common. The least common type is adult parasitoids, probably because the adult is best able to defend itself against parasitoid attack. Egg parasitoids deposit their own eggs within that of their host and are obviously very small insects. This life history has evolved several times but is most common in two families of chalcidoid wasp, the Trichogrammatidae and the Mymaridae. Members of the latter family, which have very narrow wings with pronounced cilia (fine hairs) and are known as fairy flies, include the smallest of all known insects. Some parasitoids oviposit into the eggs of their host, but their larvae only kill the host later when it is a larva or pupa. These are called egg–larval or egg–pupal parasitoids, whereas another common life history is found in larval–pupal parasitoids.

Most parasitoid eggs are deposited in or on the host, but a substantial minority (especially among the flies and beetles) are deposited in the vicinity of the hosts, and the final stages of host location are carried out by a specialized first-instar larva (a triangulum or planidial larva). There are two main modes of parasitoid larval development: as endoparasitoids or as ectoparasitoids. As their names suggest, the former feed internally within the host and the latter externally. Occasionally, a parasitoid species may begin life as an ectoparasitoid and switch midway through development to an endoparasitoid or vice versa. Ectoparasitism is found most frequently in species that attack concealed hosts in plant tissue or other refuges, where the external feeding parasitoid has some protection from the environment and predators. Parasitoids can also be classified as idiobionts or koinobionts. Idiobionts kill or permanently paralyze their host at oviposition, whereas koinobionts either do not or only temporarily paralyze the host so that it recovers and continues feeding. The parasitoid suspends its own development, normally as a first instar, and waits for the host to reach a size at which it can support its development, whereupon it resumes growth that results in host death. The koinobiont life history, normally only found in endoparasitoids, allows the adult parasitoid to attack hosts when they are still too small to support the full development of the progeny. However, while in a state of suspended development, the parasitoid larva is at risk from the host insect’s immune system.

Part of the definition of a parasitoid is that its larva requires just a single host for development; however, it is possible for several parasitoids to share the same host. Species in which only a single parasitoid develops per host are termed solitary as opposed to gregarious parasitoids. The larvae of solitary parasitoids tend to have large mandibles which they use to attack other parasitoids in the host, whether conspecific or of a different species, and thus the difference between the two types of parasitoid is more than just one of degree. Occasionally, very large numbers of parasitoids emerge from a single host, for example, several hundred bracoid wasps from a single hornworm (hawkmoth, Sphingidae) larva, although the largest broods are known from certain specialized wasps with polyembryonic development. In these species (in particular certain genera of chalcidoid Encyrtidae), one or two eggs are laid into the host, but these divide asexually to produce many embryos—occasionally hundreds and even more than 1000.

Unlike predation, in which feeding removes a prey item, a parasitized host may be discovered by a second parasitoid individual prior to its destruction. Sometimes, parasitoids avoid previously parasitized hosts, a practice called host discrimination, and this may be aided by chemical marks deposited by the first female. However, in many circumstances the second female will be selected to add its own eggs to the host. If the second individual is of the same species as that of the first, this is termed superparasitism, whereas if it is of a different species the term multiparasitism is used. Occasionally,
a female may parasitize a host that she had previously attacked (self-superparasitism). Some, but not all, parasitoid species can distinguish between hosts previously parasitized by self and by other conspecifics. In superparasitism and multiparasitism, the parasitoid larvae compete with each other for host resources; when hyperparasitism occurs, one parasitoid larva actually feeds on the other. There are two main categories of hyperparasitism, obligate and facultative. As the name suggests, an obligate species can only develop as a parasitoid of a parasitoid, whereas a facultative species develops as a normal parasitoid when it encounters an unparasitized host and as a hyperparasitoid when it discovers a previously attacked host. Hyperparasitism is sometimes called secondary parasitism (as opposed to primary parasitism) and cases of tertiary and even facultative quaternary parasitism have been recorded. An unusual life history is found in some chalcidoid wasps of the family Aphelinidae: Females develop as primary parasitoids, but males develop as secondary parasitoids of other aphelinids (including females of the same species).

II. HOST LOCATION

Parasitoids are small to tiny animals that search for hosts that are themselves small or tiny animals in a very complex environment. They have solved the formidable problems of host location by making use of a battery of clues and stimuli that reveal information about the presence or potential presence of hosts. The majority of these cues, especially those acting over long ranges, are chemical, but visual, tactile, thermal, and auditory information is also used. Moreover, the parasitoid is not a hard-wired automaton but constantly updates its estimates of the values of different stimuli to modulate its strategies for host location.

One can envisage a parasitoid as utilizing a hierarchy of cues and stimuli to reveal information about the presence or potential presence of hosts. The majority of these cues, especially those acting over long ranges, are chemical, but visual, tactile, thermal, and auditory information is also used. Moreover, the parasitoid is not a hard-wired automaton but constantly updates its estimates of the values of different stimuli to modulate its strategies for host location.

One can envisage a parasitoid as utilizing a hierarchy of stimuli of increasing information content. Perhaps the lowest grade stimuli are to habitats or microhabitats that might contain hosts. Many parasitoids attacking plant-feeding hosts orientate toward green objects, that might contain hosts. Many parasitoids attacking the lowest grade stimuli are to habitats or microhabitats of stimuli of increasing information content. Perhaps its strategies for host location.

Estimates of the values of different stimuli to modulate not a hard-wired automaton but constantly updates its information is also used. Moreover, the parasitoid is chemical, but visual, tactile, thermal, and auditory information is also used. Moreover, the parasitoid is not a hard-wired automaton but constantly updates its estimates of the values of different stimuli to modulate its strategies for host location.

One can envisage a parasitoid as utilizing a hierarchy of stimuli of increasing information content. Perhaps the lowest grade stimuli are to habitats or microhabitats that might contain hosts. Many parasitoids attacking plant-feeding hosts orientate toward green objects, whereas the parasitoids of Drosophila feeding in different types of fermenting substrate are attracted to volatile chemicals characteristic of their hosts' microhabitat. Plants damaged by herbivores release a cocktail of volatiles which are often highly attractive to parasitoids. In some cases, parasitoids respond to artificially damaged hosts, whereas in other cases they are only attracted to the 'host–host plant' complex. It has been shown that the plant's response to particular herbivores and their salivary and other secretions can be very specific, prompting the intriguing suggestion that the plant may have been selected to recruit 'bodyguards' to protect itself against its herbivores.

Host waste products reveal valuable information about their originator's whereabouts. Many parasitoids are highly attracted to host frass (feces), whereas volatiles associated with salivary or mandibular gland secretions, or with silk, are also used in host location. Some hosts have symbiotic relationships with other organisms that may provide clues to their location. For example, wood-boring sawflies are associated with symbiotic fungi that digest the wood, and their parasitoids are strongly attracted to chemicals released by the fungi. Clearly, hosts will be under strong selection to reveal as little about their location as possible to searching parasitoids. Many parasitoids use vision in the final stages of host location and camouflage and immobility may help protect hosts against parasitoids and more traditional predators. Vibration is particularly important for species attacking hosts feeding in concealed material, and some hosts 'freeze' if they detect the presence of a searching parasitoid. There is evidence that parasitoids may also use thermal gradients to detect the presence of concealed hosts.

Hosts should maintain as much as possible the chemical equivalent of 'radio silence,' although sometimes hosts have to advertise themselves for different reasons and in doing so may provide clues to searching parasitoids. Some bark beetles (Scolytidae) can only overcome tree defenses by mass attacks of many hundreds or thousands of individuals, and these are coordinated by the beetles releasing an aggregation pheromone. However, not only other beetles but also their larval parasitoids are attracted to the pheromone. Similarly, parasitoids are often sensitive to their host's sex pheromones—both species that attack the adult hosts directly and species that attack other stages and use the presence of adults as evidence of the presence of eggs, larvae, or pupae. A particularly devilish instance of parasitoids subverting their hosts' sexual signaling is that of the dipteran parasitoids of grasshoppers that home in on stridulating males.

Parasitoids are born with an innate hierarchy of responses to different stimuli, but they will change the position of cues in this hierarchy and incorporate new stimuli as they gain experience in host location. For example, if a novel volatile chemical such as vanilla is presented to the parasitoid at oviposition, the parasitoid learns to associate this chemical with hosts and subsequently flies to baits releasing this chemical. In a more natural setting, Drosophila parasitoids preferentially fly toward certain potential host microhabitats if they have...
Hosts are often found in patches that parasitoids exploit, but with decreasing returns as the fraction of unparasitized hosts decline. There will thus come a time when the parasitoid will be selected to abandon the current patch and search for a new site. Studying parasitoid patch use behavior has many parallels with predators searching for prey that are distributed in patches. This is one of the classical problems of optimal foraging theory, a field of behavioral ecology. In the simplest case, predators should leave the patch when their instantaneous rate of gain of food equals the maximum long-term rate that can be achieved in that environment. Longer patch residence is thus predicted in poorer environment, for example, when interpatch travel times are long.

The qualitative predictions of patch use theory have been supported by studies of parasitoids, but specific aspects of their searching behavior complicate the theory's application. For example, previously parasitized hosts do not disappear (as do eaten prey) but remain as potential sites for superparasitism. The parasitoid may also have to waste time in identifying them as parasitized, although they may provide useful information about host distributions. Parasitoid behavior may also change critically in the presence of conspecifics: When a parasitoid is searching a patch alone, any superparasitism is normally damaging self-superparasitism, but with decreasing returns as the fraction of unparasitized hosts decline. There will thus come a time when the parasitoid will be selected to abandon the current patch and search for a new site. Studying parasitoid patch use behavior has many parallels with predators searching for prey that are distributed in patches. This is one of the classical problems of optimal foraging theory, a field of behavioral ecology. In the simplest case, predators should leave the patch when their instantaneous rate of gain of food equals the maximum long-term rate that can be achieved in that environment. Longer patch residence is thus predicted in poorer environment, for example, when interpatch travel times are long.

The qualitative predictions of patch use theory have been supported by studies of parasitoids, but specific aspects of their searching behavior complicate the theory's application. For example, previously parasitized hosts do not disappear (as do eaten prey) but remain as potential sites for superparasitism. The parasitoid may also have to waste time in identifying them as parasitized, although they may provide useful information about host distributions. Parasitoid behavior may also change critically in the presence of conspecifics: When a parasitoid is searching a patch alone, any superparasitism is normally damaging self-superparasitism, which is not the case if other females are active in the site. Current research in this area using statistical models based on survival analysis seeks to identify the precise factors determining when a parasitoid abandons the patch.

III. HOST ACCEPTANCE AND OVIPOSITION STRATEGY

After a parasitoid locates a host, it then has to make a series of decisions that are very important for its future reproductive success. First, it must decide whether the host is of sufficient quality to be used as an oviposition site or whether not to waste eggs and search for a better host, perhaps using the current host as food. Given that the host is of sufficient quality, gregarious parasitoids have to decide how many eggs to lay, while hymenopterous parasitoids have to decide whether to lay male or female eggs. Only Hymenoptera have to make this last decision because they have a genetic system called haplodiploidy in which females develop from fertilized eggs and males from unfertilized eggs. By choosing whether or not to fertilize the egg, the parent can control the sex of its offspring. Unlike groups in which sex is determined by the random segregation of sex chromosomes, natural selection can operate on the behavior of haplodiploid females to produce sex ratios adapted to local conditions.

Before laying an egg, most parasitoids determine whether a potential host is suitable by examining it externally, and in the case of endoparasitoids often internally as well by using sensilla on their ovipositors. At either stage, a host may be rejected if unsuitable, perhaps because it is of the wrong species. As part of biological control campaigns, parasitoids sometimes have to be mass reared, and much effort has gone into investigating whether they can be reared on artificial hosts. Critical to a successful strategy of this type is an understanding of exactly what host stimuli trigger oviposition.

Even if a host can potentially support a parasitoid larva, it may still not be worth laying an egg if the resulting offspring is of poor quality, perhaps very small in size. Understanding how natural selection molds host acceptance behaviors has many parallels with understanding how natural selection influences diet breadth in predators and other consumers, a second major strand of optimal foraging theory. Theory assumes that predators should maximize their rate of gain of energy (or other resources) and hence should not take 'time out' from searching for good-quality prey items to attack poorer quality types from which the energy returns do not justify the time it takes to subdue and process them. Applying these ideas to parasitoids, searching females should not waste time ovipositing on poor-quality hosts when that time could be better spent searching for good hosts. Exactly what constitutes a poor host thus depends on the quality of the environment. Attacking a particular host of medium quality might be economically advantageous in a habitat in which very good hosts are rare or absent, but it may not make sense in a environment in which the best hosts are abundant. There is experimental support for this prediction.

In many cases, the simple assumption that parasitoids should seek to maximize their rate of gain of...
reproductive success with time is probably overly sim-
plistic. All parasitoids have a limited number of eggs, and maximizing rate of gain of fitness may lead to
exhaustion of egg reserves prior to death. More sophisti-
cated models have been developed that maximize repro-
ductive success over the lifetime of the insect and pre-
dict that the parasitoid may become more selective as
its egg supply diminishes. Again, there is experimental
data supporting this prediction, although whether egg
limitation is a major feature of parasitoid biology is a
highly contentious topic. Another factor that may
influence oviposition decisions is host feeding. Many
adult parasitoids can feed from hosts, although this
normally results in the host’s destruction. Host feeding
provides the parasitoid with energy and nutrients and
hence allows it to produce more eggs. Thus, there is a
trade-off. Should the parasitoid use a host for reproduc-
tion and possibly risk later starvation or egg exhaustion,
or should it forgo an immediate increment in fitness in
the hope of deferred payback? In these circumstances,
it clearly makes sense to feed from suboptimal hosts.

The question of optimal clutch size in parasitoids
can be explored using similar economic arguments, in
this case ideas originally developed to understand how
natural selection operates on bird clutch size. An obvi-
ous prediction is that parasitoids should adjust their
clutch size to the amount of resources available, and
there is ample evidence that parasitoids do this, with
the numbers of eggs laid typically being closely corre-
lated to the size of the host. Careful experiments have
shown that parasitoids can often estimate host size from
very subtle cues; for example, some egg parasitoids can
assess host size by very limited information about the
local curvature of the host egg. As eggs are added to
the host, the fitness returns from each egg typically
decline because the members of the brood compete
with each other for a fixed pot of resources. The optimal
clutch size is defined as the number of eggs for which
the marginal returns from adding additional eggs ex-
actly equal the marginal opportunity costs in terms of
wasted time or wasted eggs on the current host. This
argument suggests that in good-quality environments
in which fresh hosts are easy to discover, smaller clutches
should be laid. Similarly, small clutches should be found
in circumstances in which egg reserves are low. Experi-
ments on gregarious parasitoids have supported both
predictions.

Are solitary and gregarious parasitoids just part of
a single continuum in clutch size behavior? There are
arguments that suggest this is not so. Solitary parasiti-
odis typically have large mandibles which they use to
attack any other parasitoid in the host, including sib-
lings, whereas gregarious parasitoids do not have this
aggressive behavior. For evolution to convert a solitary
parasitoid to a gregarious parasitoid, natural selection
must operate not only on parental clutch size behavior
but also on larval morphology and fighting. Although
adult parasitoids and their offspring have many evolu-
tionary interests in common, these interests are not
identical and there is the potential for evolutionary
parent–offspring conflict. Models incorporating this ge-
netic conflict predict that the solitary life history is
qualitatively distinct from gregariousness, and that
there is hysteresis in the transition between the two
states with gregariousness far more difficult to evolve
from the solitary state than vice versa.

The study of the sex ratio is one of the most success-
ful areas of evolutionary ecology. The problem is to
predict why some species of plants and animals produce
sex ratios that differ from 50 : 50. For a deviation to
occur, there must not only be selection for a biased sex
ratio but also the flexibility to control son and daughter
ratios. Haplodiploid parasitoid wasps have been critical
in developing this field because they have both the
incentive and the means to produce unusual sex ratios.
As Fisher demonstrated in the 1930s, a 50 : 50 sex
ratio is expected to evolve because were the population
to be biased toward females, sons would be a more
efficient way of getting genes into future generations
(because they would, on average, mate with more than
one female), whereas if the population were to be biased
among males, daughters would be a more efficient way
of getting genes into future generations (because sons
would mate, on average, with less than one female). How-
ever, this argument assumes that the population
is completely mixed and that all males compete for all
females. This assumption breaks down if sons compete
for mates, including their sisters. In such situations,
the ovipositing female is selected to produce fewer sons
to avoid wasteful within-family conflict and also to pro-
vide more potential mates for the male offspring. Hamil-
ton, who in the late 1960s developed these ideas, called
this process local mate competition and marshaled a
long list of examples in which mating among siblings
was associated with female-biased sex ratios: 18 of the
26 examples he quoted were parasitoids.

Recent studies have shown that some parasitoids are
able to assess the degree of local mate competition that
their progeny are likely to experience and to adjust
their sex ratio accordingly. For example, Nasiona vitri-
ennis is a gregarious parasitoid of fly pupae (e.g., of
blowflies that live in birds’ nests). Wasps emerge from
the bird’s nest and mate prior to dispersal (the male has
limited powers of flight). If a single female colonizes a
bird's nest, all her offspring will mate among themselves and the wasp should produce just enough sons to mate all her daughters; if two females colonize a nest then local mate competition still occurs, although it is less intense, and a sex ratio of approximately 25% males is predicted. With greater numbers of females, the sex ratio quickly asymptotes at 50%. In laboratory experiments in which the number of insects searching together is manipulated, females respond to the presence of conspecifics by producing less female-biased sex ratios.

It had long been noted that parasitoid wasps often lay male eggs in small hosts and female eggs in large hosts, but why this evolved was only understood in the late 1970s by Charnov as part of a wider theory of environmentally correlated sex ratios. In parasitoids, there is normally a strong correlation between host size and the size of the adult wasp that eventually emerges from it. In general, wasps suffer from being small; they tend to have reduced longevity and fecundity and are less able to withstand the insults thrown at them by the environment. However, the negative consequences of being small are probably not experienced equally by the two sexes. In particular, female reproductive success may depend more critically on size than does male reproductive success for the simple reason that eggs are expensive and sperm cheap. This assumption is difficult to test because it requires size-dependent fitness to be measured in the field, which is difficult for such small animals, but the limited amount of evidence available supports this notion. If female fitness does increase more rapidly with size than does male fitness, then Charnov's theory predicts that males should be laid in relatively small hosts and females in relatively large hosts. Moreover, there should be a sharp threshold in host size below which only males, and above which only females, are produced. Laboratory experiments tend to show a threshold in host size as predicted by theory, although typically it is not as sharp as expected, perhaps because the wasp is using subtly different cues to assess host size than those used by the experimenter. To test whether relative or absolute size is important, many workers have presented medium-sized hosts to wasps in conjunction with either larger or smaller hosts. Theory predicts that the wasp should produce male eggs in medium-sized hosts when they are relatively small and female eggs when relatively large. Some wasps do indeed do this, although others have a much more fixed behavior, responding only to absolute host size. Possibly, these wasps always encounter the same host distribution in the field and hence have never evolved the flexibility to deal with variable patterns of host size.

**IV. RESISTANCE AND VIRULENCE**

Endoparasitoids that delay their development while the host continues to feed and grow in size have to protect themselves from the host's defenses. The main host defense against parasitism is a cellular immune response called encapsulation. Cells (hemocytes) circulating in the blood cavity (hemocoel) recognize a parasitoid egg or larva as foreign and cause other cells to aggregate to the intruder, forming a capsule. The cells in the capsule fuse, and the whole structure becomes hardened and melanized, thus leading to the death of the parasitoid either through suffocation or through the release by the parasitoid of necrotizing substances. Encapsulation is not a specific antiparasitoid measure but rather operates against any foreign body.

There are two broad parasitoid strategies to counter host resistance: passive and active defense. In the first, parasitoids either camouflage themselves so they are not recognized as foreign or insert their eggs in specific organs away from circulating hemocytes. How this camouflage works can be seen in some species in which the egg is coated with a layer of protein prior to injection into the host. If the egg is dissected out of the wasp and the protein layer removed, the egg becomes susceptible to encapsulation if artificially “oviposited” into the host. Many parasitoids carefully position their eggs in the host brain or ganglia where they avoid hemocytes, whereas even eggs placed in the hemocoel may obtain some protection by sticking to fat body and other host organs.

The most common form of active defense is the injection of toxins at oviposition or their later secretion by the developing larva. It is common for the ovipositing female to temporarily paralyze the host so that it is easier to handle, but the same or other substances can also damage the host's immune system. Many parasitoids, though so far only Hymenoptera (particularly Ichneumonoidea), inject viruses into the host at oviposition. The best studied type is polydnavirus, so named (poly-DNA-virus) because the genetic material is composed of many separate DNA molecules within a single protein coat. The virus DNA is stably integrated within the wasp's genome, but in certain cells of the female reproductive tract it is copied, greatly amplified, and encapsulated. In the host, the virus does not replicate, but it does invade the hemocytes responsible for encapsulation which it can destroy by triggering apoptosis. How polydnaviruses evolved is not clear. In particular, are they part of the wasp's genome that has acquired virus-like properties or an originally autonomous virus that has been captured and tamed by the wasp? In-
stances are known of wasps injecting viruses that appear unrelated to polydnaviruses and also of wasps injecting virus-like particles that resemble the protein coats of viruses but that contain no DNA.

Parasitoid eggs are surrounded by embryonic membranes which normally disintegrate after hatching. However, in some species, the cells of the membranes dissociate and float free in the hemocoel where they grow enormously in size. These cells are called teratocytes, and their primary role is most likely secretory (they are packed full of endoplasmic reticulum), producing substances that limit the host’s immune defense. When the parasitoid is fully grown it consumes the teratocytes, which may thus also have a nutritive function.

Hosts and parasitoids almost certainly exert very strong selection pressures on each other, and observed levels of resistance and virulence will be determined by this coevolutionary interaction. Two major factors determining the outcome are the asymmetry in the interaction and the nature of the costs of resistance and virulence. The interaction is asymmetric in that every parasitoid has to overcome its host’s defenses in order to survive, whereas not every host is parasitized. Models of the dynamics of host–parasitoid coevolution predict that in certain circumstances hosts may be selected not to defend themselves but to “gamble” on not being attacked. A major determinant of whether this strategy is favored is the nature of the costs to the host of investing in resistance mechanisms and of the parasitoid in investing in countermeasures. Costs are difficult to measure, but artificial selection experiments in the fruit fly, Drosophila, have shown that increased resistance to its parasitoids is accompanied by a decrease in competitive ability under conditions of resource stress. It appears that the fly switches limiting resources from optimizing feeding efficiency to immune function.

V. POPULATION DYNAMICS

Host–parasitoid population dynamics have received considerable study for two main reasons. First, parasitoids are important biological control agents and understanding how they may regulate their host has clear applied relevance. Second, host–parasitoid interactions are a useful model system for answering broader questions in population biology applicable to all resource–consumer interactions. Host–parasitoid interactions are attractive as model systems because of the simplicity of the trophic relationship: In most cases, a parasitoid attack results in one (solitary species) or a certain number (gregarious species) of recruits to the next generation of parasitoids. This is in contrast with predation, in which it is far more difficult to predict how a single instance of prey capture influences the future number of predators.

Modern consideration of host–parasitoid population dynamics began in the 1930s, particularly through the work of the Australian ecologist Nicholson and his collaboration with the mathematician Bailey. They considered the case of a host–parasitoid interaction with discrete, synchronized generations. Hosts are exposed to parasitoids during a fixed developmental stage, and those that survive parasitism and other sources of mortality produce the next generation of hosts the following years. Parasitized hosts are the source of the subsequent parasitoid generation. In their basic formulation, known as the Nicholson–Bailey model, they assumed parasitoids searched randomly in the environment. The average number of times a host is encountered is then simply a product of parasitoid density (P) and a constant (a) known as the attack coefficient. Providing parasitoid search is random, the probability of escaping parasitism is simply the zero term of a Poisson distribution with this mean (i.e., $e^{-a}$).

The Nicholson–Bailey model is important because it encapsulates the minimum bare bones of a host–parasitoid interaction. It tells us what to expect in the absence of any biological complications, and what it tells us is perhaps surprising. If the Nicholson–Bailey model is iterated over many generations, the densities of hosts and parasitoids oscillate with ever-increasing amplitude until one or both species go extinct. In essence, the parasitoid overexploits the host, which crashes to low densities, and this is followed by a drastic decline in parasitoid number. The host then recovers and in the temporary absence of the parasitoid increases to the higher densities than before; this is followed, after a lag, by a huge growth in parasitoid numbers and even more cataclysmic overexploitation. The combination of random search and the time lags inherent in this discrete-generation framework do not allow a host–parasitoid interaction to persist.

The intrinsic instability of the Nicholson–Bailey framework has set the agenda for much of host–parasitoid population dynamics in the past 60 years. What aspects of real host–parasitoid interactions allow the persistence that is so palpably clear in the field? Introducing a realistic functional response (a limit on the number of hosts a parasitoid can attack due to egg or time limitation) does not help, and if anything it makes nonpersistence more likely. In the 1960s and 1970s, many workers thought that interference between parasitoids was the key. Interference is said to occur if female parasitoids disrupt each other’s searching, per-
happens because when two parasitoids meet they fight or in other ways interact. Such interference might reduce the efficiency of high-density parasitoid populations and break the cycle of overexploitation and recovery. However, although interference occurs and can be measured in the laboratory, a consensus soon formed that in the field it was unlikely to be strong enough to allow persistence. Another idea was to assume that in addition to parasitism, hosts were also subject to direct density dependence due to competition for food. This additional density dependence can stabilize the Nicholson–Bailey interaction and may very well be important in some host–parasitoid interactions in the field. However, it is unlikely to be the complete answer because at the stable population densities that are predicted by this model, there is still substantial competition for food. This theory can neither explain why so many hosts are regulated at levels well below their food-carrying capacity nor why parasitoids can cause such dramatic reductions in host density after their release in biological control programs.

Most workers today believe that host–parasitoid persistence involves a general phenomenon called heterogeneity of risk (this idea actually dates back to the last papers of Nicholson and Bailey in the 1960s). Overexploitation of the host population occurs because all hosts are equally susceptible to parasitoid attack, and therefore nearly everyone succumbs when parasitoid densities are high. However, suppose that some hosts are in a refuge and protected from parasitoids. These individuals can survive periods of high-parasitoid densities and prevent the crash in host density that initiates the diverging oscillations (of course, if too many hosts are in a refuge then the parasitoid will not be able to prevent the host population increasing without restraint).

What might these refuges be? There are many possibilities. First, there may be physical refuges in which hosts are protected from parasitism. For example, the host might be a stem borer and a certain fraction of hosts may burrow so deeply in the plant tissues that they are inaccessible to parasitoids. Second, a certain fraction of the hosts may be physiologically immune to parasitoid attack. Third, there may be a phenological mismatch between host and parasitoid: Some hosts may avoid parasitoids by emerging either sufficiently early or late in the season so that they miss all or most searching parasitoids.

Finally, and probably most important, the refuge may not be physical, physiological, or phenological but statistical: A certain number of hosts, more than would occur if parasitoid search was random, may escape parasitism by chance each year. For example, suppose that hosts are distributed across the environment in patches, and that certain patches are more attractive than others to parasitoids (for reasons not associated with host densities, some patches may be in areas that are more likely to be encountered by parasitoids). Provided enough hosts occur in rarely encountered patches—statistical refuges—the interaction will be stable. (A popular way to model this is to replace the Poisson distribution in the Nicholson–Bailey model with a different, more aggregated statistical distribution such as the negative binomial. The resulting interaction is stable provided the variance of the distribution is sufficiently large). All these ways of stabilizing the Nicholson–Bailey model rely on there being differences among individuals in the risk of parasitism; hence their general labeling as models of heterogeneity of risk.

Recent developments in host–parasitoid population dynamics have largely involved the study of interactions with overlapping rather than discrete, synchronized generations and interactions in which the spatial distribution of the population needs to be taken into account. Models for species with overlapping generations typically abandon the discrete time difference equations of the Nicholson–Bailey model and use instead differential equations, although they often incorporate time delays to represent developmental lags. Two new insights have come from this approach. First, an interaction that is otherwise identical to the Nicholson–Bailey model may be persistent if there is a long-lived, invulnerable host stage. The reason for this relates to the arguments regarding heterogeneity of risk discussed previously. The long-lived invulnerable stage provides a refuge that allows the host population to ride out periods of very high parasitoid densities.

The second insight to emerge concerns when hosts and parasitoids differ in generation time, in particular when the parasitoid generation time is approximately half that of the host. The population can then display persistent cycles with a period approximately equal to the host generation time. This is of more than technical interest because such cycles are common among pests of tropical plantation trees such as coconut, oil palm, and cocoa, and these pests tend to be attacked by parasitoids with shorter generation times. The cycles arise because of dynamic interference between the two time lags in the system. Suppose there is a temporary increase in the densities of host: The progeny of this "blip" will form a secondary blip one host generation later. However, the parasitoids will also be able to take advantage of the increase in host population density, and this will result in a secondary
increase in their densities one parasitoid generation later. If the two generation times are equal, the two secondary blips coincide (the Nicholson–Bailey situation), but if the parasitoid generation time is approximately half that of the host the two blips do not coincide, and the parasitoid causes a trough in host density, an "antiblip," half a host generation later.

There are two main ways to incorporate a spatial component into host–parasitoid models. The first is to assume that the system consists of discrete subsystems linked by migration, whereas the second is explicitly to incorporate space and write equations that describe how host and parasitoid densities change at all spatial coordinates. Species that have isolated populations linked by migration are normally described as having a metapopulation structure. Metapopulations are an area of very active research in population biology, although studies of interacting metapopulations, such as that of a parasitoid and its host, are still rare. The most interesting aspect about host–parasitoid metapopulations is that the interaction they describe can persist, even if every component population would be nonpersistent in isolation. For this to happen, component populations must fluctuate out of synchrony so that one population doomed to extinction can be rescued by immigration from a different population in a different phase of the host–parasitoid cycle. Random, local effects that promote asynchrony favor the persistence of a metapopulation, whereas regional fluctuations in characteristics such as climate will tend to synchronize populations and hence act against persistence.

In fully spatial models, the rescue effect can still operate, but now more elaborate spatial processes can also be observed. For example, a wave of hosts can spread through the environment, chased by a following wave of parasitoids. The hosts escape parasitoids by moving into uncolonized areas ahead of the advancing waves, whereas the parasitoid destroys the host population behind the wavefront and would itself be destroyed were it not able to chase the host population. These waves tend to form spiral patterns, a feature common to many spatial interactions in population dynamics, physiology, and even inorganic chemistry. For certain parameter combinations, the organized spiral waves break down into a seemingly random jumble of interactions and short-lived wavefronts called spatial chaos.

Experimental studies of spatially extended parasitoid populations are still somewhat behind theory and one of the greatest challenges of contemporary host–parasitoid population dynamics is to devise ways of studying spatial dynamics in the field.

VI. COMMUNITY ECOLOGY
Few if any host–parasitoid interactions exist in isolation. Most parasitoids attack more than one species of host, and the majority of hosts are attacked by several species of parasitoid. How does one begin to understand the workings of a large community of interacting hosts and parasitoids? One approach is to build up from the simple one-host, one-parasitoid interactions discussed previously. Alternatively, one can survey the properties of real communities and try to deduce patterns that may reveal how they are structured.

The simplest possible community that is more complex than those discussed previously consists of one host and two parasitoids, or two hosts attacked by a common parasitoid. Both tell us interesting things. In the first case, the simplest models predict that one of the two parasitoids will invariably go extinct. This is a corollary of a basic finding in population ecology that identical species cannot coexist on the same resource—the competitive exclusion principle. For coexistence to occur, other factors must be brought into play. In particular, coexistence can occur if the ecological niche (i.e., the host) is divided in such a way that the parasitoids are no longer identical. Thus, one parasitoid may use the host at one time of year or in one microhabitat, or it may specialize on a different developmental stage of the host than the other. Coexistence can also occur by a more statistical route if hosts differ in their susceptibility to attack by parasitoids so that some are in the statistical refuge discussed previously and if the probability of being protected from attack by one parasitoid is independent of other. Spatial processes can also promote coexistence. Consider two species of parasitoid, one of which is always the winner in a straight competition at a single locality, whereas the other has a superior dispersal capability. Coexistence is now possible in an environment in which new host populations are constantly being generated because the poorer competitor is able to find unexploited hosts. This is known as a competition–colonization trade-off.

I now discuss the two-host, one-parasitoid case. Again, the simple models predict extinction, this time of one of the host populations. This occurs because the equilibrium population of parasitoids maintained on the surviving host species is sufficient to prevent the second species from replacing itself. A recent laboratory experiment cleverly illustrated this using two flour moths attacked by the same species of wasp. This is a similar situation to competitive exclusion, but the trophic structure is reversed: instead of two species co-
Parasitoids

...mechanisms for the same resource, two species are subject to the same natural enemy. In fact, there are deep biological and mathematical symmetries between these two cases, and this has led to the phenomenon of two species interacting through a common natural enemy being referred to as apparent competition to stress the parallels with direct competition in which two species interact through a common resource. Again, for a two-host, one-parasitoid system to persist, something extra must be added to the simplest models. The two hosts might be present at different times of the year or might be spatially segregated, or the parasitoids might preferentially exploit the most abundant host (behavioral switching).

Much of traditional community ecology has stressed how biological communities may be structured by resource competition, and thus it does not apply to many insect communities in which the majority of species feed on different host plants and thus never come into contact. Apparent competition is significant because it can at least potentially structure a community in exactly the same way as does direct competition. The degree to which this actually occurs is a major theme in current insect ecology.

Progressing from models of three species to those with larger numbers of components becomes increasingly difficult. There are two problems: First, more assumptions have to be made about how different species interact and about the values of large numbers of parameters. Seldom are there field data to reduce this burden of supposition. Second, the dynamic behavior of larger communities becomes increasingly more complicated. For example, a model of a five-species community consisting of two hosts, two specialist parasitoids, and a generalist parasitoid showed the same range of population dynamic behaviors as those exhibited by simpler communities. However, it also showed more complex behaviors in which the full five-species community was unstable with one or more species going extinct, but with the resultant smaller communities, after they had reached equilibria, then being susceptible to invasion by the species that had recently gone extinct. Other multispecies models have shown complex chaotic dynamics. Currently, it is unclear the extent to which the bottom-up approach, modeling explicitly the dynamics of each member of a large community, is a feasible way to approach parasitoid community ecology.

The top-down approach to community ecology consists of searching for patterns in multispecies assemblies that provide evidence of structuring forces. For example, workers have searched for patterns in the number of species of parasitoids attacking different species of hosts. The attraction of this approach is that there are numerous studies in the literature providing information on the parasitoid complexes of different insects. The major result to emerge from these studies is that host feeding niche influences parasitoid species numbers. Leaf-mining insects are attacked by the largest number of species, with successively smaller numbers attacking more concealed hosts (gall formers, shoot borers, and root borers) and less concealed hosts (species living in leaf rolls and tis and those living externally like typical caterpillars). There are two explanations for this pattern. One suggests that the number of parasitoid species that can coexist on a single host is influenced by the fraction of that host's population that inhabits a refuge from (all) parasitoid attack. Propagules of this view argue for a correlation between feeding niche and the size of the refuge. Alternatively, many species are likely to occur on hosts that are taxonomically and ecologically similar to other hosts because there is less of an evolutionary hurdle to overcome in incorporating a new species in a parasitoid's host range. Leaf-mining insects have a far more uniform ecology than insects feeding in other host niches. They are also relatively taxonomically homogeneous, and this may have facilitated host transfer and broad parasitoid host ranges.

There are less data in the literature on the host range of parasitoids because this requires all potential hosts in a area to be surveyed. However, the data that are available support one major conclusion: Idiobionts have broader host ranges than koinobionts. Recall that idiobionts kill or permanently paralyze their host at oviposition, whereas koinobionts delay development until the host is fully grown. During this suspension of growth, the parasitoid has to cope with attack by the host's immune system, and the need to evolve to be finely attuned to the host to counteract its defenses probably limits koinobiont host range.

A different top-down approach to parasitoid community ecology has recently been taken by a few groups, although it is still too soon to assess its value. There is a long tradition of constructing food webs in ecology, and one of the aims of this research program is to search for patterns that are common across different webs. A major limitation of this research is the heterogeneity in published food webs, which are typically collected in very different ways and which differ greatly in their taxonomic resolution. Host-parasitoid interactions have many advantages for food web studies, the most prominent of which is the relative ease with which trophic links can be identified and quantified. This has led to the construction of several quantitative food webs.
in which all hosts, parasitoids, and links are expressed in the same units. The webs published to date illustrate the extent to which different hosts are linked by shared natural enemies and also the extent to which indirect effects such as apparent competition may act as forces structuring the community. Of course, a limitation of this approach is that only one guild of natural enemies is included in the web, but it will be interesting to determine whether common patterns emerge as more communities are studied.

VII. THE IMPORTANCE OF PARASITOIDS

Why are parasitoids important to students of biodiversity? First, they are abundant members of nearly all terrestrial communities. The populations of few herbivorous insects are unaffected by parasitoid attack, and understanding how they influence and possibly regulate the densities of their host is immensely important in understanding the mechanisms underlying biodiversity. Much of the modern work in parasitoid ecology has been motivated by these considerations.

The second reason why parasitoids are important is also the reason why there is such an enormous and valuable literature on parasitoids dating back nearly a century: Parasitoids are our allies in the war against agricultural and forestry pests. There are two main strategies of parasitoid biological control: classical and inundative. In classical biological control, a parasitoid is released into an area to control a pest. Typically, the pest is a nonnative species that has been accidentally introduced into a new region, and normally the parasitoid is a species that attacks the pest in its region of origin. The parasitoid is released once or on a few occasions until it is established. Although only 10–20% of biological attempts are successful, they can have dramatic effects in controlling pests, and the savings in crop yield are normally many multiples of the investment in the biological control program. There are interesting patterns in the probability of success against different hosts, for example, biological control has an excellent record against scale insects and mealy bugs but is far less successful against aphids.

The second strategy is inundative release in which large numbers of parasitoids are released onto a crop to destroy a pest. Applications are made as often as required, and long-term establishment of the parasitoid (which is normally already there at low density) is not a goal. In effect, the parasitoid is acting here as a biological insecticide. There are two main situations in which this strategy is used today. The first is when labor is relatively cheap so that the costs of mass rearing the parasitoids are not too great. In many tropical countries, huge numbers of egg parasitoids are mass reared and released to attack cutworms, armyworms, and bollworms. Advances in the mechanization of mass rearing are making inundative release more attractive in developed, high-wage countries. The second situation concerns high-value crops in enclosed growing conditions such as greenhouses. Control of tomato whitely is typically by parasitoids in northern Europe. Inundative control has been favored in these situations because of the decreased tolerance of insecticide residues by regulatory authorities and because of the general need to reduce chemical input so as not to disrupt insect pollination and the existing successful biological control of other pests.

In the post-Silent Spring era biological control has often been viewed as an unmitigated good (particularly by biocontrol workers) because it does not involve pesticides. However, in recent years, increasing attention has been paid to the dangers of biological invasions, particularly by plants, and this has led to a reassessment of biological control and to worries regarding the effect of introduced parasitoids on native insects. These are important issues and need to be carefully scrutinized, although there is a real danger of an overreaction and a return to the automatic reliance on environmentally harmful chemical insecticides. Currently, the major problem is a lack of good data on the consequences of biological control introductions. The challenge to applied ecological entomologists is to develop a sufficient understanding of parasitoid population and community ecology building on the ideas described previously, so that they can provide rules and guidelines that will ensure that biological control has a net positive rather than negative effect on biodiversity.

See Also the Following Articles

HYMENOPTERA • PARASITISM • SPECIES INTERACTIONS

Bibliography


PELAGIC ECOSYSTEMS

Andrea Belgrano,* Sonia D. Batten,† and Philip C. Reid†

*The Royal Swedish Academy of Sciences and †Sir Alister Hardy Foundation for Ocean Science

I. Phytoplankton and Primary Production
II. Zooplankton
III. Pelagic Fish
IV. Outlook

GLOSSARY

calanoid copepods These belong to the Crustacea and are very abundant in the zooplankton. They play a major ecological role in the food web of the pelagic ecosystem as grazers of phytoplankton and as a food source for, e.g., larval and adult fish.

clupeid fish Pelagic fish, including the anchovy, sardine, and herring, which feed on plankton.

El Niño Southern Oscillation (ENSO) Regarded as quasi-periodic fluctuations occurring in the equatorial region of the Pacific Ocean. The ENSO is associated with changes in the sea surface temperature, sea surface levels, and rainfall patterns in the tropics and with a possible influence on the weather at higher latitudes. ENSO events have also been associated with changes in fisheries and ecosystem processes.

gadoid fish Includes the cod Gadus morhua L.; the life cycle of cod can be summarized as four stages: spawning, larvae, juveniles, and adults. During the larval and early juvenile phases they feed largely on copepods on the nursery grounds; after this phase they become largely stationary and feed on benthic or epibenthic animals.

North Atlantic Oscillation (NAO) The difference in atmospheric pressure between Ponta Delgada, Azores, and Stykkisholmur, Iceland. The variations in the NAO are usually associated with a positive or negative phase related to changes in the direction and strength of the westerly wind as well as in sea surface temperature. During a positive phase, winters over Scandinavia are warmer and vice versa. Recently, the fluctuations observed in the NAO have been related to changes in ecosystem processes.

phytoplankton Unicellular photoautotrophic plant cells ranging in size from 1 μm to 1 mm. They are divided into 12 taxonomically defined divisions, including diatoms, dinoflagellates, and coccolithophorids.

primary production An estimate of the rate of carbon fixation by phytoplankton; this can account for up to 75% of the photosynthetic production on Earth.

Russell cycle Named after F. Russell, the Russell cycle describes a major change in the biota of the English Channel from 1923 to 1972.

zooplankton From the Greek zoon meaning animal, zooplankton comprises those animals that are found passively drifting or weakly swimming in the water column. Zooplankton can be divided into two major categories: holoplankton, which are organisms that spend their entire lives as plankton, and meroplankton, which are organisms that spend part of their life cycle as plankton and part on the seafloor as benthic invertebrate larvae or as nekton (e.g., fish larvae).
PELAGIC ECOSYSTEMS cover more than 70% of the surface of the earth. Their species diversity and richness are related to physicochemical and biological processes acting at a range of temporal and spatial scales. They are strongly influenced by atmosphere–ocean (coupling) interactions related to hydrodynamic processes. Climatic oscillations, such as the El Nin˜o Southern Oscillation and the North Atlantic Oscillation, may be associated with changes in ecosystems on a decadal scale. Pelagic biodiversity is part of the ocean’s complex, adaptive ecosystem, and the interactions between species diversity and ecological processes such as food web dynamics occur at a variety of temporal and spatial scales. The study of changes in biodiversity patterns in relation to the natural and anthropogenic (e.g., overfishing) variability in pelagic ecosystem structure and function need to be considered for a sustainable global ecosystems policy for the next millennium.

I. PHYTOPLANKTON AND PRIMARY PRODUCTION

The etymology of the word “plankton” derives from the ancient Greek meaning wandering. Plankton have very limited motility and are dependent on currents and the physical environment for their location. Phytoplankton comprises a very diverse group of single-celled photosynthetic planktonic plants that are divided into 12 taxonomic divisions (3000–4500 species of oceanic plankton) and live in the surface water of the oceans. The role played by phytoplankton is at the base of the trophic planktonic plants that are divided into 12 taxonomic divisions (3000–4500 species of oceanic plankton) and live in the surface water of the oceans. A typical phytoplankton community contains a mixture of these, although the species which are dominants vary through time. The majority of the species present in the ocean gain their energy via photosynthesis and, therefore, environmental factors such as light availability, temperature, and the supply of major nutrients in the form of ions are important factors that influence their distribution on both temporal and spatial scales. Several hypotheses have been proposed which explain the coexistence of several species; one, “temporal succession,” supposes that the nutrient regime of the environment changes rapidly so that all species are in different states of approaching to or declining from their maximum abundance. In low-turbulence environments microhabitats may develop which favor the growth of

a distinct patch of plankton. Alternatively, different species may be limited in growth by the availability of different nutrients so that coexistence is possible. In neritic temperate waters of the North Sea and in the open temperate North Atlantic, the mean seasonal pattern of phytoplankton abundance shows a distinct peak in the spring, followed by a summer decline and a second, lesser, peak in the autumn. Work by Colebrook (1986) showed a clear differentiation between those species associated with the spring bloom, which tended to be diatoms, and those associated with the autumn bloom, which tended to be dinoflagellates. There were changes in the North Sea in terms of phytoplankton species composition after 1987, for example, Ceratium spp. (Dickson et al., 1992) and Thalasiosira longitudinalis (Reid et al., 1992), which signaled a change in the ecosystem. These changes have been related to an increase in the sea surface temperature (SST) and to changes in the extension of the Baltic/Norwegian waters in the North Sea and an increased input of more oceanic water. Variations in the seasonal cycle are caused by different preferences of the groups for light intensity, water stability, temperature, and nutrient availability. A bimodal pattern of the seasonal cycle is evident across the temperate North Atlantic but other patterns exist in different geographical regions. Polar regions show a single summer peak because light intensity is only sufficient for a short period of phytoplankton growth. The North Pacific Ocean shows no spring peak, and in fact there is little change in the phytoplankton standing stock throughout the year. The principal reason is that the dominant zooplankton (the copepod genus Neocalanus) overwinter as adults which produce young in the spring without feeding. The young are then able to immediately take advantage of any increase in the phytoplankton and effectively prevent a spring bloom. The north Pacific Ocean shows no spring peak, and in fact there is little change in the phytoplankton standing stock throughout the year. The principal reason is that the dominant zooplankton (the copepod genus Neocalanus) overwinter as adults which produce young in the spring without feeding. The young are then able to immediately take advantage of any increase in the phytoplankton and effectively prevent a spring bloom. The tropics show little seasonality and both phytoplankton and zooplankton fluctuate slightly throughout the year, dependent on local processes. Large blooms of phytoplankton can be monitored via satellite remote sensing (e.g., the coccolithophorid Thalassiothrix longissima). They have the potential to modify, to some extent, the carbon exchange taking place between the ocean and the atmosphere. Figure 3 shows an example of a cyanobacterium, Synchococcus sp., which is an extremely small photosynthetic cell 1–3 μm in length. Picocyanobacteria and nanoflagellates are responsible for assimilating excretory products of zooplankton consumers in the euphotic zone in nutrient-poor waters characteristic of the central ocean gyres. As noted by Falkowski et al. (1998), the fixation of carbon by phytoplankton results in approximately 45 gigatons of organic carbon per year, of
FIGURE 1. Phytoplankton species illustrating the taxonomic diversity as well as the shape and size of these organisms, which are an important component of the pelagic food web. (A) Diatomophyceae, Skeletonema costatum, a worldwide diatom species with distinctive features such as long chains of small cells with long external tubes. Diameter varies between 2 and 21 μm. (B) Dinophyceae, Ceratium furca, a worldwide dinoflagellate with a solitary or paired life-form, with a straight body, and the epithea gradually tapers into an anterior horn. The size ranges in length between 70 and 200 μm and in width between 30 and 50 μm. (C) Dinophyceae, Dinophysis norvegica, a solitary dinoflagellate, normally found in cold waters, varies in size with length between 48 and 67 μm and width between 39 and 53 μm. Dinophysis norvegica can be regarded as a toxic species since it produces a toxin that causes diarrheic shellfish poisoning and therefore it is a potentially dangerous species when found in coastal waters in large numbers (photos courtesy of Mats Kuylenstierna and Bengt Karlson). See also color insert, Volume 1.

which approximately 16 gigatons is exported out of the surface waters. Changes in the total and export production can affect the marine food web structure, ultimately leading to changes in the fish stocks. The photosynthetic pigments such as chlorophyll a can be used as a proxy measurement of phytoplankton biomass, and with the aid of satellite-derived measurements a color index can be derived to illustrate changes in primary production. Behrenfeld and Falkowski (1997) generated several maps of primary production estimates for the world’s ocean (Figs. 4 and 5). Measurements of phytoplankton color on samples taken by the Continuous plankton recorder (CPR) survey from the central northeast Atlantic and North Sea (Fig. 6) shows a similar increasing trend and longer growing season as that of the satellite-derived vegetation index and this trend dates back to 1948. The phytoplankton increase is not ubiquitous because an inverse trend is evident
in eastern Atlantic waters between latitudes 59°N and 63°N. This index of phytoplankton change may reflect a possible long-term biological response in the North Atlantic to increasing global temperatures which are causing melting of ice, sea ice, and permafrost in the Arctic (Reid et al., 1998).

In the central North Pacific environment, Venrick (1990) observed 245 species of phytoplankton in the shallow waters and 231 in the deep zone during the period between 1972 and 1985. In both areas, 21 species accounted for 90% of the individuals and most of the variance in abundance occurred at small spatiotemporal scales and samples within a zone were similar. This study showed that epipelagic populations of the North Pacific central environment were more stable when compared with planktonic populations from other ecosystems. In relation to the pelagic food web structure, phytoplankton is an important food source for the larval stage of pelagic fish and thus plays an important role in the trophic structure of pelagic communities.

II. ZOOPLANKTON

Zooplankton are the secondary producers in pelagic ecosystems and comprise an extraordinarily wide range of organisms. The zooplankton community of continental shelf waters, for example, may contain larval stages of littoral and benthic invertebrates (meroplankton) in addition to the species that spend all their lives in the plankton (holoplankton). Phytoplankton production in these nutrient-rich waters is generally regulated by the grazing activity of zooplankton (Fig. 7). Most zooplankton belong to the crustacean, with the principal organism group being copepods. Two copepod species, Calanus finmarchicus and C. helgolandicus, constitute the major components of the northeast Atlantic and North Sea zooplankton in terms of biomass, abundance, and trophic role (Marshall and Orr, 1972). These popula-
tions follow distinct spatial and seasonal dynamics such that \( C. \) finmarchicus is a northern spring species, whereas \( C. \) helgolandicus is located in warm-temperate waters and reaches its maximum abundance in autumn.

The distribution of these species of Calanus has been studied during the period 1938–1995 in the North Atlantic (Figs. 8–10) in relation to the year-to-year variability in the North Atlantic Oscillation (NAO).
NAO is a climatic oscillation affecting the hydrology and climate in the North Atlantic Ocean and adjacent regions, and it can be compared to El Niño in the Pacific Ocean. After 1987, the zooplankton community in the North Sea changed, for example, an increase in the abundance of the copepod *Corycaeus* spp. and it may also have affected the distribution and abundance of the two *Calanus* species. The increase in temperature, wind, and alteration of the winter circulation pattern observed during years of a high positive NAO index have resulted in unfavorable conditions for the *C. finmarchicus* population leading to a significant decrease in the abundance of the species. Conversely, these hydro-climatic shifts have proved beneficial to *C. helgolandicus*, the abundance of which has increased during these years (Planque and Taylor, 1998; Stephens et al., 1998).

In the northeast Pacific the plankton community structure (Fig. 11) responded to the warming in 1958 to 1960 in the California Current in relation to an El Niño event that caused a sudden change in the SST distribution (McGowan et al., 1998). The zooplankton abundance has generally decreased in the California Current system while, in a synchronous manner, the zooplankton abundance in the Gulf of Alaska gyre increased. These interdecadal regime shifts in these two systems driven by climatic variations in the atmosphere are reflected first in changes in the physical structure of the ocean resulting in large-scale biological responses in the ocean. These regime shifts have a severe effect on the temporal and spatial distribution of planktonic species, leading to changes in the secondary production as well as in community structure (McGowan et al., 1998). In terms of species richness and diversity in the open area of the Atlantic Ocean, the maximum species richness is found at depths between approximately 1000

and 1500 m. Angel (1991) showed, for example, that the species richness of planktonic ostracods in the northeastern Atlantic at 20°W occurred between 500 and 1000 m. Moving up the water column, and at higher latitudes, there tend to be fewer species but these occur at higher abundances. The zooplankton, and especially the Calanus species, are an important part of the pelagic food web since they are the major source of food for fish species including cod. The nauplii of Calanus in particular are a key component of the diet of fish larvae and a recent model of Calanus off the Georges Bank (Miller et al., 1998) showed clearly the importance of the availability of C. finmarchicus for the cod and haddock populations along the bank (Lough, 1984).

Recent evidence shows that SST explains almost 90% of the geographic variation in planktonic foraminiferal diversity for the whole Atlantic Ocean (Rutherford et al., 1999). The peak in foraminiferal diversity was found at middle latitudes in the Atlantic, and this pattern may be extended to other taxa such as euphausiids (krill species), pteropods, and chaetognaths in the North Pacific. This is additional evidence that changes in pelagic diversity in the world’s ocean result from changes in the upper-ocean physical structure and especially that water column and ecosystem structure may depend on vertical niche separation.

In the North Sea, particularly the northern areas, there has been an increase in zooplanktonic species richness, as recorded by the CPR survey, since the 1950s. The increase, particularly of calanoid copepods, in the most northerly areas can be attributed to increased inflow from the Atlantic Ocean, but currently the species which are increasing or adding to the numbers are not permanent components of the North Sea plankton. Resident and colder water holoplanktonic species have declined in abundance, and meroplankton (adult and larval) and expatriates from warmer oceanic and mixed waters have increased. This constitutes a significant challenge to analysis of pelagic diversity because the species richness of the meroplanktonic groups


Regional variation in the seasonal cycle of C. finmarchicus showing how the distribution varies in the North Atlantic and how the peak of the log abundance differs between different locations (maps from Planque, 1997, courtesy of SAFHOS).

FIGURE 10

Regional variation in the seasonal cycle of C. finmarchicus showing how the distribution varies in the North Atlantic and how the peak of the log abundance differs between different locations (maps from Planque, 1997, courtesy of SAFHOS).

In the marine environment, approximately 80% of the known fish species are found in coastal areas, whereas approximately 2% are epipelagic and the remaining are deep-sea species. For example, at a 1000-m depth in the North Atlantic 44°N 13°W, the fish were the most taxonomically diverse group found in the micronekton (Angel, 1993). In the pelagic food web, pelagic planktivorous fish are regarded as predator species and changes in their prey items in relation to physical changes in the ocean circulation pattern are reflected in fish stock fluctuations (Fig. 12) and changes in the catches of one species of sardine, Sardinops sagax (Kawasaki, 1991; Sharp and McLain, 1993; Stromberg, 1997). The increase in the sardine populations may reflect changes in phytoplankton biomass in the Pacific Ocean during the 1970s and 1980s (Verrick et al., 1987) as well as changes in zooplankton abundances in the northeast Pacific. Zooplankton populations off the Peruvian coast declined from 1972 coincident with the severe decline in the anchoveta fishery. Shifts in the Kuroshio current southeast of Japan, and shifts in the wind stress, may be related to the fluctuations in sardine populations. The effects of El Niño Southern Oscillation events on the fluctuations observed in the fish stocks are unclear as, for example, in the case of the Chilean sardine and the Peruvian anchoveta that do not indicate a clear connection. During the El Niño events in the late 1980s, 1990s, and early 1990s, the Peruvian anchoveta stock was high, whereas the Chilean sardine stock was low. In 1971 and 1972, the El Niño event decreased dramatically and as a result the Chilean sardine stock increased while the Peruvian anchoveta stock crashed. At the same time, both the Japanese sardine and the Californian sardine increased. This trend was confirmed dur-
ing the El Niño events that took place in 1982 and 1983. In the Black Sea the introduction of a new species, the comb jellyfish Mnemiopsis leidy, caused a strong decline in the Black Sea anchovy stock. As the abundance of the jellyfish increased, it had a great impact on the recruitment of the anchovy due to changes in the food web and competition for food. These population collapses are very important since pelagic fish such as sardine and anchovy species are among the top 10 species that dominate the world catches.

If we examine the long-term studies of plankton in the North Atlantic observed by the CPR, we find that in the North Sea after 1987 an increase in phytoplankton biomass coincided with taxonomic changes in the phytoplankton and zooplankton species and a large increase in the catches of the western stock of the horse mackerel Trachurus trachurus L. in the northern North Sea. This reflected a northerly expansion of the stocks along the shelf edge from the Bay of Biscay to the North Sea after 1987. The increase in the western horse mackerel may be related to the fact that this species tracked down an increase in food resources available in the North Sea during high NAO years. The presence after 1988 of an unusual plankton species (for the North Sea), Eucheta
The covariation of three different populations of Pacific sardine compared with the temperature anomaly from the North Pacific in relation to ENSO events. The data for the California sardine are part of the California Cooperative Fisheries Investigations (CALCOFI) [data from Kawasaki (1991) and Sharp and McLain (1993); redrawn from Stromberg (1997)].

hebes, which normally has a more southerly/oceanic distribution, suggests an increase in the shelf edge current that in turn favored the advection of western horse mackerel from the western margin of Europe, allowing them to reach the northern North Sea. As noted by Stromberg (1997), the causes and events related to the fluctuations in the standing stocks of pelagic fish in relation to changes in the plankton and to production are very complex, and these changes may be related to wind-driven advection processes and to a change in the behavior of predator and prey in response to changes in the small turbulent flow. In the coastal areas, and in particular in the shelf seas, very abrupt changes in the community structure can occur and persist on a decadal scale (Steele, 1998). The observed changes in the dynamics of pelagic fish stocks may be related to low- and high-frequency climatic variability that in turn affects the interannual variation in recruitment, growth, and abundance. For example, in a time series analysis of variation in O-group cod Gadus morhua along the Norwegian Skagerrak coast, variation was found to not be directly associated with the NAO and with the abundances of C. finmarchicus, but the fluctuations in the cod stock may have been related to changes in the bottom habitats and especially in the seagrass coverage by Zostera marina (Fromentin et al., 1998). Further analysis of this data set demonstrated the importance of age-structured interactions such as asymmetric competition and cannibalism between cohorts, causing fluctuations in the abundance of the O-group and I-group juvenile cohorts of cod (Bjørnstad et al., 1999).

The northern stock of cod on the Labrador Shelf and Grand Banks of the northwest Atlantic declined and, although overfishing occurred, other factors also affected stock recruitment. The match/mismatch hypothesis was revisited by Conover et al. (1995). This hypothesis basically states that the spawning of cod occurs at the same time each year and if the spawning of copepods (the major food source) occurs at the same time there is a "match" and a successful recruitment. If the spawning of copepods is delayed through dependence on the occurrence of the phytoplankton spring bloom there is a "mismatch" and the recruitment of cod larvae will be low. Changes in the physical structure in the Labrador Sea and Grand Banks occurred with unusually low temperatures in the Labrador current and this was followed by the appearance in the zooplankton of the copepod Calanus glacialis rather than C. finmarchicus. Calanus glacialis spawns later than C. finmarchicus and therefore does not provide a good food source for the cod larvae. Changes in the Labrador current may be the result of climate change and in turn affects the structure of the pelagic food web by creating a shift in the species composition as just described, thus resulting in a change in cod recruitment. Ultimately, the increase in fishing activities may have a severe effect on fish populations (Stromberg, 1997) since indirect effects considerably alter the biogenic habitat for fish recruitment, as does the removal of top predators. These indirect interactions by the fisheries can cause a trophic change in the food web as a result of top-down control of community structure, whereas a change in the resource availability induced, for example, by the El Niño event represents a bottom-up process that influences the dynamics of pelagic ecosystems. The interplay between these two types of perturbation may be the cause of shifts in species composition which result in changes in the food web structure. In light of these processes, we should also consider the importance of the microbial loop (Azam, 1998) and be aware of bacterial control of the fluxes of organic matter and its role in ecosystem dynamics. For example, fish production in the eastern Mediterranean decreased through a dominant microbial loop, and the uncoupling of bacteria from primary productivity.
was found in a fishery off the coastal region of Newfoundland that was causally related to the intensity of the fishery.

IV. OUTLOOK

The structure of pelagic ecosystems varies at both temporal and spatial scales. At an ocean basin spatial scale, phenomena such as El Niño can be linked at a temporal scale of approximately 1 year to the stock recruitment of pelagic fish, and species and single species models can be used as predictive tools. At a larger spatial scale and on a decadal timescale, climatic oscillations such as the NAO can be linked to regime shifts in the species composition and community structure of ecosystems, and models should be at a community or ecosystem level. A retrospective analysis of historical data could reveal important biodiversity patterns.

The Russell cycle (Russell, 1973) explained the changes in the biota of the English Channel very well. In 1929, a decline in the herring stock was observed and the Plymouth herring stock (on the southwest coast of the United Kingdom) collapsed in 1936. In 1931, the decline in the macroplankton resulted in a change in the presence of the arrowworm from the species Sagitta elegans to S. setosa. In the following decade, there was a drastic decrease in the standing stocks of gadoid fish such as cod (G. morhua). The English Channel ecosystem essentially shifted from one dominated by herring and that had a good supply of dissolved nutrients to one with less nutrients and that was dominated by pilchards and smaller planktonic organisms. It is interesting that in 1965 there was an increase in the fish larvae concomitant with an increase in nutrient supply. These events were followed in the 1970s by a reverse switch in the Sagitta species and an increase in copepod abundance, whereas the number of pilchard eggs decreased. The Russell cycle is a well-documented example that shows how ecosystems can shift or oscillate from one state to another after reaching a bifurcation point (May and Oster, 1976). Pelagic ecosystems seem to show decadal or longer oscillations and this is reflected in the fossil records and paleoceanographic studies. Pelagic ecosystem diversity and the maintenance of biodiversity appear to be influenced by climatic oscillations and through pressure exerted by overfishing and removal of top predators from the food web. These can be regarded as macroscopic system properties that in turn are related to different flux that may have an effect on the structure and ecosystem functioning (Levin, 1998). An understanding of changes in biodiversity patterns in relation to the variability in ecosystem structure and function needs to be incorporated in a sustainable global ecosystem policy for the next millennium.

See Also the Following Articles

FISH, BIODIVERSITY OF • FISH STOCKS • FOOD WEBS • OCEAN ECOSYSTEMS

Bibliography


I. Classification and Uses
II. Efficiency
III. Ecotoxicology and Management

GLOSSARY
pesticide A chemical substance used for controlling, preventing, destroying, or mitigating a pest organism.

PESTICIDES HAVE BEEN IN RECORDED USE since 1000 B.C. Arsenic was in regular use as a garden insecticide in China by A.D. 900 and chemicals of one form or another have protected humans and their crops and livestock throughout the development of modern civilization. In comparison with this long time scale, we are still in the earliest phases of the use of synthetic organic pesticides, which were first used over large areas in the 1940s. There has, nonetheless, been sufficient time for several generations of pesticide chemistry to evolve, and pesticides have influenced all habitats and have affected the lives of all their inhabitants over this period.

I. CLASSIFICATION AND USES
Nowhere have the costs and benefits of modern technology been more difficult to reconcile than with pesticides. Benefits that include improved yield, crop quality, and food safety and reductions in vector-borne disease incidence have driven, and will continue to drive, their use. The earlier synthetic organic compounds were, however, flawed in their environmental behavior. They were persistent, they had a very broad spectrum of toxicological activity, and they displayed a tendency to be magnified in concentration through food chains, such that damage was inflicted to animal populations that lived beyond the treated area in habitats that were not intentionally contaminated. The discovery of some of these limiting impacts was only made possible through technical advances in, for example, analytical chemistry and conceptual advances where, for example, the ability to predict environmental behavior from chemical properties only developed after pesticides had been in use for decades. New pesticide discoveries are no longer accompanied by the marvel and optimism that characterized the first synthetic insecticides. Our ability to exploit these materials has, however, advanced considerably in recent years, and the chemicals themselves are increasingly specific in their impacts and appear to pose reduced risks. Scientists still question the value of reliance upon chemical pesticides, however, and modern pest control is characterized in general by a cautious approach to their management and use.

Set in a volume that will be consulted largely by biologists and ecologists with an interest in biodiversity, this article summarizes the chemicals that are in most widespread use and outlines the processes that contribute most to efficient delivery to the biological target. The
TABLE I
Classification of Pesticides by the Type of Pest Controlled

<table>
<thead>
<tr>
<th>Type of Pest Controlled</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mites, ticks, and spiders</td>
<td>Acaricide</td>
</tr>
<tr>
<td>Adult insects</td>
<td>Adulticide</td>
</tr>
<tr>
<td>Algae</td>
<td>Algiacide</td>
</tr>
<tr>
<td>Trees, brush, and scrub</td>
<td>Arboricide</td>
</tr>
<tr>
<td>Birds</td>
<td>Avicide</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Bactericide</td>
</tr>
<tr>
<td>Fungi</td>
<td>Fungicide</td>
</tr>
<tr>
<td>Plants</td>
<td>Herbicide</td>
</tr>
<tr>
<td>Insect and sometimes related arthropod pests, including mites</td>
<td>Insecticide</td>
</tr>
<tr>
<td>Ticks</td>
<td>Ixodicide</td>
</tr>
<tr>
<td>Insect larvae</td>
<td>Larvicide</td>
</tr>
<tr>
<td>Mites, ticks, and spiders</td>
<td>Miticide</td>
</tr>
<tr>
<td>Mollusks, such as slugs and snails</td>
<td>Molluscicide</td>
</tr>
<tr>
<td>Nematodes</td>
<td>Nematicide</td>
</tr>
<tr>
<td>Invertebrate eggs</td>
<td>Ovicide</td>
</tr>
<tr>
<td>Fish</td>
<td>Pesticide</td>
</tr>
<tr>
<td>Vertebrate Predators</td>
<td>Predacide</td>
</tr>
<tr>
<td>Rodents, including rats and mice</td>
<td>Rodenticide</td>
</tr>
<tr>
<td>Trees, brush, and scrub</td>
<td>Termiticide</td>
</tr>
</tbody>
</table>

nature and importance of toxicological and ecological effects are then reviewed, followed by a summary of procedures through which pesticides are regulated and managed. This article does not provide detailed reviews of biochemical mode of action, application, formulation, or environmental fate and behavior in any detail, and readers are recommended to pursue some of the literature in the bibliography to gain insight into these important areas of pesticide science.

A. Uses of Pesticides

Pesticides may be classified by the types of pest they control (terms often being the suffix-cide; Table I) or by the effects that they have upon the pest organism (terms that do not bear the suffix-cide; Table II).

B. Classification of Pesticides

1. Formulations

Pesticides are marketed in complex mixtures, or formulations, containing the pesticide chemical itself, the active ingredient (AI), and additives that enhance mixing and dilution in water or oil and release or deliver the toxic material once it has been applied. Common ingredients in a pesticide formulation include the AI, solvents, carriers, surface-active agents, and specialized additives. Solvent choice is determined by the solubility of the AI, potential toxicity to target plants (phytotoxicity), toxicology, flammability, volatility, and cost. Some solvents are not miscible in water and cause emulsions to be formed (e.g., xylene), whereas others are selected because of their ability to dissolve the AI and their ability to dissolve in water (e.g., isopropyl alcohol). Carriers can include inert clays that disperse the active ingredient through a powder or granular formulation. Surface-active ingredients are used to assist in the process of emulsion formation (the dispersion of the pesticide liquid within the liquid diluant), but they also include wetting agents (materials that reduce surface tension and enhance wetting and coverage of surfaces), dispersing agents (materials that maintain the emulsion as microscopic droplets within the diluted formulation), and spreading agents (materials that enhance coverage of waxy plant foliage and insect cuticles).

The pesticide formulation may be in solid (i.e., powder or granule) form, or it may be a liquid or gas concentrate. Modern pesticides may achieve their intended effects at rates of less than 10 g per hectare (10,000 m²/1 ha), and the formulation blend enables these tiny quantities to be distributed evenly over the intended target surface.

A two-letter code denotes the formulation type on all pesticide labels, and this is a fundamentally important aspect of the selection of a pesticide for a particular use. There are four groups of formulation types:

- Group 1: Concentrates for dilution in water, including DC (dispersable concentrate), EC (emulsifiable concentrate), SC (suspension concentrate), and WP (wettable powder)
- Group 2: Concentrates for dilution with organic solvents, including OL (oil-miscible liquid) and OP (oil-dispersable powder)
- Group 3: Formulations to be applied undiluted, including GR (granules) and UL (ultra-low-volume (ULV) liquids)
- Group 4: Miscellaneous formulations, including RB (bait) and AE (aerosol dispenser)

2. Pesticide Types

Pesticides are characterized by their chemical diversity also. The summary below includes many of the most important groups of chemicals used in crop protection, but it is not exhaustive. Some pesticide properties are
TABLE II
Classification of Pesticides by Effects on Pests

<table>
<thead>
<tr>
<th>Effect Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antifeedant</td>
<td>Inhibits feeding while insects remain on the treated plant</td>
</tr>
<tr>
<td>Attractant</td>
<td>Lures pest to a specific location</td>
</tr>
<tr>
<td>Chemosterilant</td>
<td>Prevents reproduction</td>
</tr>
<tr>
<td>Defoliants</td>
<td>Removes foliage without immediately killing plant</td>
</tr>
<tr>
<td>Desiccant</td>
<td>Causes plant parts to dry</td>
</tr>
<tr>
<td>Disruptants</td>
<td>Destroys or inactivates harmful organisms</td>
</tr>
<tr>
<td>Feeding stimulant</td>
<td>Causes vigorous feeding</td>
</tr>
<tr>
<td>Growth regulator</td>
<td>Stops, speeds up, or retards growth in insects or plants</td>
</tr>
<tr>
<td>Repellent</td>
<td>Drives away pests without killing them</td>
</tr>
<tr>
<td>Semiochemical</td>
<td>Pheromones and other substances emitted by plants or animals that alter animal behavior</td>
</tr>
<tr>
<td>Synergist</td>
<td>Substance that enhances the effects of a pesticide</td>
</tr>
</tbody>
</table>

sufficiently uniform within the major classes of chemicals for this broad classification to be used in selection of chemicals for a particular use. Biochemical mode of action, for example, tends to be similar within major classes, and many crop protection programs use materials from several classes to avoid excessive selection pressure for resistance to pesticides. Other properties, are, however highly variable within pesticide classes as well as between them. These include vapor pressure (volatility) and the various partitioning coefficients that determine the distribution and fate of the active ingredient in the environment. These properties determine the effectiveness of the pesticide against a specific target or in a specific climate type or habitat, and detailed knowledge of these properties is required for pesticide selection to be effective. Some aspects of properties will be dealt with in Section II.

3. Major Classes of Insecticides
a. Inorganic
These are mainly nonvolatile and water-soluble compounds that often have high mammalian toxicity and may also be cumulative poisons. Many are now discontinued or banned. They include materials such as borax acid, copper sulfate, mercuric chloride, and sodium arsenite.

b. Organic
i. Botanical These pesticides are derived from plant materials, some of which have low mammalian toxicity. Some have adverse effects on wildlife, including toxicity to fish. Examples include neem tree (Azadiracta indica) oil, which is used to protect stored products from insects; pyrethrum, from Chrysanthemum cinerariaefolium, a powerful but rapidly degraded insecticide that affects the peripheral nervous system, causing paralysis, known as "knockdown"; nicotine (from Nicotiana tabacum), an alkaloid insecticide that is also a neuromuscular poison in vertebrates, and rotenone, from the roots of Derris and Lonchocarpus spp., also used as a piscicide. Some of these materials are widely used in locally made or commercial formulations, and although they may be highly toxic, many are not persistent in the environment and degrade rapidly in sunlight or when exposed to microbial activity in soil or water. Each material has unique properties, and it is not possible to make generalizations about toxicology or environmental impact.

ii. Organochlorines Organochlorines are synthetic pesticides, among the first to be developed. Early materials, including DDT, had dramatic early successes in vector-borne disease control, but most have adverse environmental impacts, resulting from persistence in the environment, accumulation through food chains, global redistribution of residues, and ecotoxicological impacts. Many organochlorines are now banned internationally. Examples include diphenyl aliphatics (DDT, dicofol, methoxychlor), benzene derivatives (HCH, pentachlorophenol), cyclodiene (chlor dane, endosulfan, endrin), and polychloroterpenes (camphechlor).
iii. Organophosphates

Organophosphates, or OPs, are esters of phosphoric acid. Many have high mammalian toxicity and may require frequent application because they are generally not persistent. Organophosphates are nerve poisons, acting through inhibition of cholinesterase. They fall into three groups: (1) Aliphatic organophosphates, which are the oldest group, some with low mammalian toxicity (e.g., malathion, which has been in use for 40 years) but others having high mammalian toxicity, but short persistence. Short persistence makes many OPs of use in short-season crops, where the plant or the consumer has a low tolerance for pests (e.g., vegetable crops). Examples include acephate, dichlorvos, dimethoate, malathion, and phorate. (2) Phenyl organophosphates, which are more stable and persistent but which include materials with high mammalian toxicity. Examples include fenitrothion, methyl parathion, and temephos. (3) Heterocyclic organophosphates, which may be more persistent and active in soil. Examples include azinphos-methyl, chlorpyrifos, phosmet, and pyrazophos.

iv. Carbamates

Similar to organophosphates, carbamates are esters of carbamic acid, with short residual life and a wide spectrum of activity. This group also includes some herbicides and fungicides. Some of the pesticides with highest mammalian toxicity are carbamates. Again, they divide into three groups: (1) Methylcarbamates with a phenyl ring structure, including carbarlyl and methiocarb; (2) methyl and dimethyl carbamates with heterocyclic structures, including aldicarb and methomyl; and (3) methyl carbamates of oximes, having a chain structure, including bendiocarb.

v. Formamidines

Formamidines are used as insecticides and acaricides, have the characteristic nitrogen structure –N=CHN–, and are effective against the eggs and larvae of Lepidoptera (butterflies and moths). They are useful alternatives to organophosphates, where resistance has developed. Examples include amitraz.

vi. Dinitrophenols

Dinitrophenols are broad-spectrum insecticides that have two nitro groups (NO2) attached. Examples include binapicryl and dinocap.

vii. Organotins

Organotins are tin-based organic compounds that act as acaricides and fungicides, some with long residual activity. Examples include cyhexatin.

viii. Pyrethroids

Pyrethroids are synthetic compounds based upon pyrethrins, found in chrysanthemum flowers. They have similar toxicological properties but tend to be much more photostable and persistent. Natural pyrethrins occur in mixtures of six esters, and resistance is rare in the insects that they are used against. The synthetic pyrethroids are generally marketed as a single ester, and resistance may develop rapidly. Pyrethroids are generally of low mammalian toxicity but are highly toxic to fish and bees and beneficial preditory or parasitic invertebrates. Pest resurgence, or population outbreaks that result from the destruction of natural enemies, may occur in certain crops. Examples include allethrin, biorresmethrin, cyfluthrin, deltamethrin, and permethrin.

ix. Fumigants

Fumigants are insecticides that kill by vapor or gas action, have low molecular weight, and often contain halogen radicals. They can be highly toxic to vertebrates, particularly in enclosed spaces. Examples include chloropicrin, ethylene dibromide, and methyl bromide.

x. Petroleum Oils

Petroleum oils, also known as mineral oils, have insecticidal properties. Light applications of refined paraffin oils can be made to trees in leaf. Lower viscosity, semirefined oils can be applied to dormant trees in winter to kill invertebrates and their eggs.

xi. Antibiotics

One antibiotic, abamectin, derived from Streptomyces, is an effective insecticide and antiparasitic agent also in veterinary use. There is evidence that when used in cattle, it can harm dung-feeding invertebrates. It has some systemic properties (i.e., it can penetrate leaves and be carried within the plant vascular system).

xii. Semiochemicals

Semiochemicals are behavior-modifying compounds, including pheromones, particularly the sex attractants of female moths. These may attract male moths for monitoring or be dispersed by spraying or formulation into controlled-release devices to establish false trails and disrupt mating of pest species. These compounds have very low environmental hazards, and the development of resistance is rare.

xiii. Insect Growth Regulators

Insect growth regulators are compounds derived from, or inhibitory to, insect hormones and include methoprene, a juvenile hormone mimic, effective against mosquitoes, and diflubenzuron, which disrupts the synthesis of chitin and kills insects when they molt.
PESTICIDES, USES AND EFFECTS OF

xiv. Microbials Included in this group are bacterial agents, particularly Bacillus thuringiensis, which is used against Lepidoptera, Diptera, and Coleoptera. A proteinaceous inclusion or endotoxin, which constitutes 40% of cell weight at sporulation, breaks down in insects with alkaline midguts to an active protoxin. This category also includes fungal diseases of insects, including Beauvaria bassiana, Metarhizium anisopliae, and Verticillium lecanii, some of which can be formulated and sprayed like conventional insecticides. Some viral agents are also used on a large scale, including nuclear polyhedrosis viruses (NPV) of Lepidoptera. Microbial insecticides tend to have very limited effects on organisms that they do not directly parasitize, and vertebrate effects are almost unknown.

xv. New Insecticides A series of new insecticides belonging to several previously unexploited classes of chemistry have been introduced in recent years. Some of these pose greatly reduced environmental risks, although this does not apply to all of the newer materials. These include imidacloprid (in the class chloronicotinyls), a highly systemic pesticide that is now used on a very wide scale. Research on these newer insecticides is limited at present, but there is optimism that ecological impacts and risks to wildlife will be greatly reduced as they become more widely used.

4. Major Classes of Fungicides

a. Inorganic

Inorganic fungicides are derived from sulfur or simple metal salts. They are generally stable, persistent, and insoluble in water. They include sulfur, which was originally applied as flowers of sulfur, in dust form, and which is still used, but in a more highly ground colloidal suspension. It has both direct contact and fumigant activity at temperatures above 20°C, but above 32°C the vapor may cause phytotoxicity (toxic harm to the target plant). Environmental damage and toxicological impacts to nontarget organisms are otherwise limited. Copper-based fungicides include Bordeaux mixture, an early fungicide that consists of a solution of copper sulfate and hydrated lime. With 12% copper, this fungicide has low mammalian toxicity and is no longer available for any purpose. (2) Organocupper compounds, including copper acetate, which was first synthesized in 1899. These fungicides are not easily washed off leaves, being insoluble in water, and give persistent protection. They act by the nonspecific denaturing of proteins. Examples include cuprobam. (3) Organotins, triphenyltin salts, including fentin, which can be both toxic and phytotoxic.

b. Organic

Organic fungicides have the reputation of being safer and less persistent than some of their inorganic counterparts, and they are often used at very low doses.

c. Dithiocarbamates Dithiocarbamates are derivatives of sulfur-containing dithiocarbamic acid, in combination with zinc salts, ferric salts, and manganous salts. These fungicides have greater efficacy, better stability, and less phytotoxicity than elemental sulfur. Their toxicity derives from the formation of the isothiocyanate radical (–N=C=S−) in breakdown. Compounds in this group include thiram, maneb, metham sodium, and zineb.

d. Organometallics This group of fungicides includes the following: (1) Mercury compounds, formerly popular for disinfective and protective action as well as volatility. They have high mammalian toxicity and are no longer available for any purpose. (2) Organocupper compounds, including copper acetate, which was first synthesized in 1899. These fungicides are not easily washed off leaves, being insoluble in water, and give persistent protection. They act by the nonspecific denaturing of proteins. Examples include cuprobam. (3) Organotins, triphenyltin salts, including fentin, which can be both toxic and phytotoxic.

e. Substituted Aromatics These compounds include derivatives of benzene and phenol, with hydrogen atoms replaced by chlorine, nitrogen, and oxygen, and are suited for seed treatment and control of soil-borne fungi. Examples include chlorothalonil and pentachlorophenol (PCP).

f. Dicarboximides Also called sulfonylureas, these compounds are considered to be among the safest pesticides in seed treatment and protectant sprays. Examples include iprodione and vinclozein.

g. Phthalalimides Phthalalimides are used as nonsystemic, broad-spectrum, foliar fungicides on fruit, vegetables, and ornamentals. Examples include captan and captafol, which is used widely in the tropics, but increasingly subject to restrictions.

h. Dinitrophenols These nonsystemic fungicides are used against powdery mildew and include the compounds binapacryl and dinocap.
vii. **Triazines** This group consists mainly of herbicides, with anilazine as the only fungicide, used as a protectant treatment in vegetables.

c. **Systemic Compounds**

These fungicides are absorbed and translocated through plant tissues and provide longer term, protective control, with some curative and therapeutic effects for plants that are already infected. There are many groups, some of which are summarized below:

i. **Oxathiins** Oxathiins control basidiomycete fungi and include the compounds carboxin and methfuroxam.

ii. **Benzimidazoles and Thiophanes** Benzimidazoles and thiophanes are broad-spectrum fungicides and are widely used in the tropics; intensive use has often led to resistance. Some of these compounds replaced organomercury fungicides as seed dressings. They include benomyl, carbendazim, and thiabendazole.

iii. **Pyrimidines** Included in this class of compounds are bupirimate and ethirimol.

iv. **Acylalanines** This class of compounds includes metalaxyl.

v. **Ergosterol Biosynthesis Inhibitors (EBIs)** Ergosterol biosynthesis inhibitors are a heterogeneous group of compounds with a common mode of action. They can have systemic, protective, and curative properties. They include (1) imidazoles (e.g., imazalil and prochloraz), (2) piperazine, pyridine, and pyrimidine compounds (e.g., pyrifenox and triforine), (3) morpholines (e.g., dodemorph and tridemorph), and (4) triazoles (e.g., flutriafol, myclobutanil, and propiconazole).

vi. **Organophosphates** Organophosphates are a group consisting mainly of neurotoxic insecticides but also include fungicides such as pyrazophos.

vii. **Phenylamide and Other Fungicides against Oomycetes** This is another heterogeneous group, sharing the property of toxicity to oomycete fungi. High specificity and systemic activity help to confer resistance when used intensively. Included in this group are (1) phenylamides (e.g., benalaxyl and metalaxyl) and (2) the compounds cymoxanil and prothiophene.

viii. **2-Aminopyrimidines** This small group of systemic fungicides has activity against powdery mildew and includes bupirimate.

ix. **Quinones** This group includes benodanil and futonil.

x. **Other Organic Fungicides** A number of important fungicides do not belong to the groups above, but comprise a random selection of compounds with unrelated chemical structures. They include chlorothalonil, dodine, guazatine, and thioethoxam. Most are nonsystemic, protective fungicides.

5. **Major Classes of Herbicides**

Herbicides kill or interrupt the growth of plants. Some are selective, but others kill all plants that come into contact with them and are used in industrial settings and in rights-of-way maintenance.

a. **Inorganic Herbicides**

This class includes mostly salts that have been in use for a considerable time. Sodium arsenite solutions and arsenic trioxide were popular in the 1960s but are now banned. This class also includes iron and copper sulfates, which are used for foliar application. Sulfuric acid has been widely used in horticulture and cereal crops as a selective herbicide. These materials are water soluble and readily leach from soil. Further examples include sodium salts of boron acid (borates) and sodium chloride.

b. **Organic Herbicides**

i. **Organic Arsenicals** These materials are much less toxic to mammals than inorganic arsenic salts. They inhibit metabolism, competing with phosphate in essential reactions. These compounds include disodium methanearsenate (DSMA).

ii. **Phenoxyalkyl Phosphonic Acids** Herbicides in this group are a series of compounds in which the phenoxy nucleus is linked with acetic, propionic, and butyric acids. Solubility in water is high relative to that of many herbicides. They have selective, hormone-type effects on broad-leaved weeds, but grasses are tolerant. 2,4-D, 2,4,5-T, and MCPA were immensely popular and once regarded as safe. In the 1970s, dioxin contamination was suspected in 2,4,5-T and its registration was canceled.
iii. Substituted Amides  This is a large group of organic nitrogenous herbicides, including amides and anilides. Amides primarily act in soil against annual grass weeds, causing stunting. They have low mammalian toxicity and include chlorbromid and propyzamide. Anilides, which have numerous subgroups, are used in postemergence grass and broad-leaved weed control as well as in preemergence compounds to control germinating weeds and grasses. All have low mammalian toxicity (e.g., alachlor and propanil).

iv. Diphenyl Ethers  These compounds possess two benzene rings, joined through oxygen or a more complex chain of molecules. They include preemergence and selective postemergence compounds of low mammalian toxicity. They are fairly insoluble in water, do not leach, and may be persistent for several months. Examples include diclofop methyl and oxyfluorfen.

v. Dinitroanilines  Dinitroanilines are the most widely used group of herbicides in agriculture. They have a dinitroaniline nucleus in common and are effective against annual grasses and broad-leaved weeds when applied preemergence. Degradation occurs through volatilization and photodecomposition. Mammalian toxicity is low, and in soil persistence can be quite long. This class of compounds includes benthylamin and trifluralin.

vi. Substituted Ureas  Based on the simple nitrogen-containing molecule urea, the first example, monuron, was discovered in the 1950s as a total herbicide. More modern compounds are used as selective, preemergence herbicides that are strongly adsorbed to soil. They inhibit photosynthesis, causing chlorosis. These compounds have low mammalian toxicity and their efficacy is influenced by temperature, rainfall, and soil type. Examples include chloroxuron, diuron, isoproturon, and linuron.

vii. Carbamates  In addition to many insecticides and fungicides, a number of herbicides have been developed from this chemical group. They include pre- and postemergence materials that inhibit germination and cell division. They have low mammalian toxicity and short persistence. Examples include asulam and phenmedipham.

viii. Thiocarbamates  The thiocarbamates are a group of carbamates containing sulfur, including selective pre- and postemergence compounds. The allate compounds may persist for many months in soil.

Metham sodium is a soil fumigant, converted in the soil to methyl isothiocyanate. Examples include dithiophosphate, metham sodium, and thiobenzcarb.

ix. Heterocyclic Nitrogens  These compounds have a ring structure where the carbon atoms are replaced by nitrogen or sulfur. They include triazines, triazinones, triazoles, pyridines, and uracils. Triazines are a very selective group of compounds; selectivity depends upon the plant’s ability to metabolize the AI. They are soil applied and absorbed through roots, inhibiting photosynthesis once the plant has emerged. Mammalian toxicity is generally low, and intense application in some areas has led to groundwater contamination and restrictions upon use. Examples include atrazine and simazine. Triazinones have six-membered rings and include hexazinone. Triazoles, with a five-membered-ring structure, are active against broad-leaved weeds (e.g., amitrole). Some pyridines are effective against deep-rooted herbaceous weeds, and others are selective brush killers (e.g., picloram and trichlopyr). Uraclls, another substituted, six-membered-ring family of chemicals with two nitrogen atoms and a double bond, are primarily for preemergence application and uptake by roots. They are effective at controlling annual grasses and broad-leaved weeds over an extended period by inhibition of photosynthesis. Uraclls may be very persistent. Examples include bromacil and 2,4-D.

x. Bipyridiliums, Also Termed Pyridines  Containing two pyridyl rings, this group includes some extremely well known and widely used compounds (e.g., diquat and paraquat). They have contact action on the above-ground parts of plants and in the plant they are reduced to free radicals that destroy tissue under light. They are widely used as desiccants. These compounds are deactivated by sorption to the soil and slow to degrade, leading to increasing restrictions. Both AIs are hazardous, being the main cause of pesticide-related death in many countries. There is no known antidote.

xi. Aliphatic Acids  This group includes chlorinated derivatives of acetic acid, trichloroacetic acid (TCA) and dalapon. These compounds are used on non-cropland and in forestry, killing the plant by causing precipitation of proteins within cells.

xii. Phenol Derivatives  Herbicides in this group are highly toxic, selective, foliar herbicides and include dinitrophenols and one chlorinated phenol, pentachlorophenol. Dinitrophenols, primarily contact herbicides,
TABLE III

<table>
<thead>
<tr>
<th>Pesticide</th>
<th>Method of application</th>
<th>Target organism</th>
<th>Utilization efficiency (%)</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demeton-S-methyl</td>
<td>Foliar spray</td>
<td>Aphids on sugar beet</td>
<td>0.000008</td>
<td>Insects located in heart leaves, an effective refuge from direct spraying</td>
</tr>
<tr>
<td>Dieldrin</td>
<td>Seed treatment</td>
<td>Wheat bulb fly larvar</td>
<td>0.0015</td>
<td>Pests less likely to encounter seed as plant develops</td>
</tr>
<tr>
<td>Dimethoate</td>
<td>Foliar spray</td>
<td>Aphids on field beans</td>
<td>0.03</td>
<td>Relatively low efficiency of spray retention by plant</td>
</tr>
<tr>
<td>Lindane</td>
<td>Foliar spray</td>
<td>Capsids on cocoa</td>
<td>0.02</td>
<td>Pesticide losses due to drift when treating large trees</td>
</tr>
<tr>
<td>Lindane/dieldrin</td>
<td>Aerial spraying</td>
<td>Locustis</td>
<td>6.0</td>
<td>Pesticide applied by aircraft within swarm, maximizing chance of contact</td>
</tr>
</tbody>
</table>


Inhibit respiration and photosynthesis (e.g., dinoseb (DNBP), DNOC) and are now banned in the United States. Pentachlorophenol is now being banned in Europe and the United States.

xiii. Benzonitriles or Substituted Nitrites These compounds consist of a benzene ring with a cyanide (CN-R) radical and are broad spectrum, acting on various processes of growth and tissue disruption (e.g., dichlobenil and ioxynil).

xiv. Miscellaneous Herbicides Herbicides with unrelated chemical structures include some widely used materials such as endothal sodium, an aquatic weed killer, glyphosate, a nonselective residual, postemergence herbicide, and alloxynil sodium, flazifop-butyryl, and other selective, postemergence systemic herbicides, used against perennial and annual grasses.

II. EFFICIENCY

Ideally, the receiving organism receives a toxic dose of pesticide and nontarget organisms escape injury, achieving high efficiency of use. Efficiency is, however, very rarely measured. When it has been, efficiency is far from ideal, and is expressed as the proportion of the applied dose taken up by the target organism, or by the amount that reaches the organism under the conditions of application. The amount that the organism is exposed to depends upon placement (which is a function of application method), pattern of release (a function of the formulation), redistribution and transfer processes (a function of the physicochemical properties of the compound, environmental conditions, and the nature of the substrate), and the location and receiving characteristics of the organism. The latter is a function of the degree of contact or avoidance of the chemical, the activity of the organism in question, and, on a large scale, the spatial dynamics of the organism as it relates to the pattern of treatment and the persistence of the chemical.

A. Delivery and Bioavailability

The toxic effects of a pesticide against any intended target or unintended nontarget organism are a function of intrinsic toxicity (the activity when it is directly applied to the organism) and the amount that reaches the organism under the conditions of application. The amount that the organism is exposed to depends upon placement (which is a function of application method), pattern of release (a function of the formulation), redistribution and transfer processes (a function of the physicochemical properties of the compound, environmental conditions, and the nature of the substrate), and the location and receiving characteristics of the organism. The latter is a function of the degree of contact or avoidance of the chemical, the activity of the organism in question, and, on a large scale, the spatial dynamics of the organism as it relates to the pattern of treatment and the persistence of the chemical.
**B. Application, Formulation, and Delivery of the Toxic Dose**

Through application and formulation, it is possible to manipulate the placement, size, and composition of the discrete units or drops of pesticide in which it is dispensed and the rate of release from those units or particles.

1. **Soil Treatment**

Some of the most difficult challenges relate to the problem of delivering sufficient pesticide to small targets distributed in a bulk medium in which mobility is restricted and the chemical subject to rapid degradation. It is therefore important to localize the pesticide to the vicinity of the target and plant part that needs protection. For example, granular formulations and seed treatments can both be used to increase efficiency of transfer of pesticide to organisms in the soil. Once applied, the chemical must move into the soil matrix and the rate of release determines the concentration in the soil. Distribution and redistribution can be achieved by bulk flow of water moving down through the soil profile after rain or irrigation. Contact between the water network in soil pores and the coated seed is very important.

Efficiency, or the toxicity of soil-applied AI per unit of material released, is determined by the way in which the chemical partitions between the solid, liquid, and air phases of the soil. The partition coefficients between these different phases can be used to accurately predict the rate of movement of pesticides within a given soil type.

2. **Spray Application**

Spray application is the commonest method of pesticide delivery. It is convenient, flexible, and simple, but it lacks selectivity and there is a risk of contaminating nontarget areas. Hydraulic sprayers give a range of pesticide drop sizes, at the smallest extreme of which there is a tendency to drift away from the target area. Controlled droplet application (CDA) sprayers are much rarer, but they control drop size and as a consequence may reduce drift and increase efficiency.

The main function of spray machinery is to transfer energy to the liquid pesticide formulation to atomize it into droplets by means of a nozzle. The drop size distribution is typical for the nozzle in question and it is characterized by the volume median diameter and the number median diameter of the drops.

- **Volume median diameter (VMD):** If a sample of the volume contains drops with a diameter less than the VMD, and the other half contains drops with a diameter greater than the VMD.
- **Number median diameter (NMD):** If the drops are divided in half by number, independent of volume, then the diameter of half of the drops is larger than the NMD, and the diameter of the other half is smaller.

The NMD : VMD ratio indicates the spread of drop sizes: the larger the ratio, the broader the spectrum. If the ratio approaches 1, the drops are all of a similar size and behave in a similar way.

Conventional spray application through hydraulic nozzles presents a major challenge for optimization of pesticide use efficiency, and their widespread use accounts for a great deal of pesticide drift and off-target exposure and contamination. Table IV provides details of the drop spectrum from a conventional fan jet hydraulic spray nozzle. Drops of between 1 and 50 μm in diameter account for only 4.2% of the volume, but 90.8% of the total drops in the sample. In addition, 61.9% of the volume of the spray liquid is in drops of greater than 160-μm diameter. In 1 liter of spray liquid, this would yield nearly 90 million drops with a diameter of less than 50 μm, and therefore susceptible to drift.

Drop behavior has been the subject of detailed research. Drops are subjected to various forces as they leave the nozzle and before they impact on a surface.

- **Gravitational forces:** Small drops have a low terminal velocity. For example, a 5-μm-diameter drop falls at 0.075 cm/sec and often drifts from the target area, whereas a 500-μm-diameter drop falls to the ground at 213.9 cm/sec.
- **Movement induced by air movement:** Wind carries drops away from the target area. For example, in a wind of 1 m/sec, a 5-μm-diameter drop released from 1 m experiences 350 m of sideways transport, whereas a 500-μm drop moves 0.48 m. To deposit within 3 m of the target area, a drop must have a diameter of greater than 67 μm in a wind speed of 0.7 m/sec and it must have a diameter of greater than 168 μm in a wind traveling at 3 m/sec.
- **Movement induced by air movement:** The friction of the air rapidly decelerates drops from the velocity at which they left the nozzle. For example, a 5-μm drop is decelerated to its deposition velocity in 0.33 cm and a 200-μm drop is brought to the equivalent velocity in 630 cm.

Drift is worse where relative humidity is low. Drops become smaller as spray liquid evaporates and sedimen-
### TABLE IV

Drop Sizes from a Hydraulic Spray Nozzle Measured with a Laser Particle Size Analyzer

<table>
<thead>
<tr>
<th>Drop range (μm)</th>
<th>Volume (%)</th>
<th>Cumulative volume (%)</th>
<th>Number (%)</th>
<th>Cumulative number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>563–262</td>
<td>26.17</td>
<td>51.94</td>
<td>0.07</td>
<td>0.80</td>
</tr>
<tr>
<td>262–160</td>
<td>35.77</td>
<td>87.71</td>
<td>0.73</td>
<td>8.73</td>
</tr>
<tr>
<td>160–113</td>
<td>17.32</td>
<td>105.04</td>
<td>1.36</td>
<td>10.36</td>
</tr>
<tr>
<td>113–64</td>
<td>8.78</td>
<td>113.82</td>
<td>1.82</td>
<td>12.16</td>
</tr>
<tr>
<td>64–30</td>
<td>4.81</td>
<td>118.63</td>
<td>2.31</td>
<td>14.47</td>
</tr>
<tr>
<td>30–15</td>
<td>2.66</td>
<td>121.29</td>
<td>2.86</td>
<td>15.33</td>
</tr>
<tr>
<td>15–10</td>
<td>1.12</td>
<td>122.41</td>
<td>4.83</td>
<td>16.16</td>
</tr>
<tr>
<td>10–5</td>
<td>0.56</td>
<td>122.97</td>
<td>6.45</td>
<td>16.81</td>
</tr>
<tr>
<td>&lt;5</td>
<td>0.27</td>
<td>123.24</td>
<td>21.90</td>
<td>90.81</td>
</tr>
</tbody>
</table>

* Data from Micron Sprayers, UK. Nozzle, Fan Jet SS8002 operating at a pressure of 3 bar and a flow rate of 700 ml/min. VMD = 192 μm, NMD = 12 μm, VMD : NMD ratio 16.

Evaporation is therefore exacerbated by the shrinkage of drops and the increase of relative surface area.

Large drops pose an equivalent problem: not deflected by air currents, they gain momentum and either miss the plant target or hit the soil or strike the plant, shatter, and fall to the ground. The cubed relationship between diameter and volume means that the volume of a 100-μm-diameter drop is 1 million times that of a 1-μm-diameter drop. The variation in pesticide dose between a 4- and 400-μm-diameter drop (common in a hydraulic nozzle drop spectrum) is 1 million fold. Many drops of a diameter greater than 300 μm fall to the ground, and a large proportion of the pesticide applied by a hydraulic nozzle never comes into contact with the pest, disease, or weed it is intended to control.

### 3. Deposition on Obstacles

Objects are surrounded by a cushioning layer of air, and spray drops require considerable momentum to hit a target. Larger drops strike targets, and smaller drops move around them. The probability of striking depends upon the shape of the object; hairs, for example, pick up small drops. Flying insects such as mosquitoes are best hit by drops of 10–20 μm in diameter; resting flies require drops of 30–40 μm, diameters of 90–130 μm are needed where crop penetration is needed, and a diameter of >250 μm is optimal for horizontal surfaces such as weeds.

Continuing reliance upon hydraulic spray technology limits the efficiency, effectiveness, and safety of the spray application process.

### III. ECOTOXICOLOGY AND MANAGEMENT

Pesticides pose challenges through being toxic to organisms other than the specific pests, diseases, or weeds they are targeted against. Relative toxicity is expressed in a number of ways, but the most prevalent relates to oral and dermal exposure of mammals, which equates to both human and vertebrate wildlife toxicological risks for organisms that are directly exposed (Table V).

Table VI lists some representative active ingredients by toxicity and demonstrates that in all major pesticide types and range of classes, there is a considerable variety in intrinsic toxicities to mammals.

It is extremely difficult to translate these data, or toxicological data for any other species, into predictions of ecological risk. Very toxic materials may be relatively nonhazardous in the environment, because they are rapidly sorbed or dissipated, whereas less toxic materials may prove hazardous because they readily leach or runoff into water or because they are persistent.

Recent attempts have been made to bridge the gap that exists between basic toxicology and environmental chemistry and the potential ecological impact of pesticides in ecosystems. Van Straalen and van Rijn (1998) computed species sensitivity distributions for soil fauna (soil invertebrates that contribute to biogeochemical cycling), based upon laboratory test data for toxic effects in soil, and calculated the concentration that...
PESTICIDES, USES AND EFFECTS OF

TABLE V
Pesticide Toxicity Classes

<table>
<thead>
<tr>
<th>Toxicity rating</th>
<th>LD&lt;sub&gt;50&lt;/sub&gt; single oral dose for rats (mg/kg)</th>
<th>LD&lt;sub&gt;50&lt;/sub&gt; single dermal dose for rabbits (mg/kg)</th>
<th>Probable lethal oral dose for humans</th>
</tr>
</thead>
<tbody>
<tr>
<td>6, supertoxic</td>
<td>&lt;5</td>
<td>&lt;20</td>
<td>A taste, a gram</td>
</tr>
<tr>
<td>5, extremely toxic</td>
<td>5–50</td>
<td>20–200</td>
<td>A pinch, 1 teaspoon (3 ml)</td>
</tr>
<tr>
<td>4, very toxic</td>
<td>50–500</td>
<td>200–1,000</td>
<td>1 teaspoon (3 ml) to 2 tablespoons</td>
</tr>
<tr>
<td>3, moderately toxic</td>
<td>500–3,000</td>
<td>1,000–2,000</td>
<td>1 liquid ounce (30 ml) to 1 pint (470 ml)</td>
</tr>
<tr>
<td>2, slightly toxic</td>
<td>5,000–15,000</td>
<td>2,000–20,000</td>
<td>1 pint (470 ml) to 1 quart (940 ml)</td>
</tr>
<tr>
<td>1, practically nontoxic</td>
<td>&gt;15,000</td>
<td>&gt;20,000</td>
<td>&gt;1 quart (940 ml)</td>
</tr>
</tbody>
</table>

would be below the no-effect concentration (NOEC: the highest dose not to cause a specific effect) for reproductive inhibition for 99% of the species that are theoretically present within the soil arthropod community. Using a simple model for pesticide fate, based upon exponential decay, they then calculated for a series of compounds the time that each product would take to reach this “95% protection level.” This time, referred to as the “ecotoxicological recovery time,” can be used as a basis for calculating the minimum time for ecological recovery to take place after pesticide application. There is considerable scope for further development of this technique, built upon readily collected toxicological data sets. Such data are, however, surprisingly diffi-

TABLE VI
Examples of Pesticide Toxicities to Vertebrates from Laboratory Test Data Used in Registration

<table>
<thead>
<tr>
<th>Pesticide type and class</th>
<th>Pesticide technical name</th>
<th>Oral LD&lt;sub&gt;50&lt;/sub&gt; (rat, mg/kg)</th>
<th>Dermal LD&lt;sub&gt;50&lt;/sub&gt; (rat (a), rabbit (b))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insecticides</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organochlorine</td>
<td>Dieldrin</td>
<td>46</td>
<td>10 (a)</td>
</tr>
<tr>
<td></td>
<td>Endosulfan</td>
<td>80</td>
<td>339 (b)</td>
</tr>
<tr>
<td>Organophosphates</td>
<td>Acrophos-methyl</td>
<td>16</td>
<td>222 (a)</td>
</tr>
<tr>
<td></td>
<td>Chlorpyrifos</td>
<td>135</td>
<td>2000 (b)</td>
</tr>
<tr>
<td></td>
<td>Dichlorvos</td>
<td>36</td>
<td>73 (a)</td>
</tr>
<tr>
<td></td>
<td>Dimethoate</td>
<td>320</td>
<td>&lt;130 (b)</td>
</tr>
<tr>
<td></td>
<td>Malathion</td>
<td>1800</td>
<td>4100 (b)</td>
</tr>
<tr>
<td>Carbamate</td>
<td>Aldicarb</td>
<td>0.7</td>
<td>3 (b)</td>
</tr>
<tr>
<td></td>
<td>Carbaryl</td>
<td>830</td>
<td>&gt;2000 (b)</td>
</tr>
<tr>
<td>Pyrethroids</td>
<td>Deltamethrin</td>
<td>135</td>
<td>&gt;2000 (b)</td>
</tr>
<tr>
<td></td>
<td>Allethrin</td>
<td>930</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bioresmethrin</td>
<td>7070</td>
<td></td>
</tr>
<tr>
<td>Fungicides</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dithiocarbamates</td>
<td>Thiram</td>
<td>780</td>
<td>&gt;10000 (a)</td>
</tr>
<tr>
<td></td>
<td>Maneb</td>
<td>7900</td>
<td>&gt;50000 (a)</td>
</tr>
<tr>
<td>Dinitrophenol</td>
<td>Dinitropol</td>
<td>980</td>
<td></td>
</tr>
<tr>
<td>Herbicides</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inorganic</td>
<td>Sodium arsenite</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Biyphosphates</td>
<td>Paraquat</td>
<td>137</td>
<td></td>
</tr>
<tr>
<td>Traceines</td>
<td>Suxamone</td>
<td>3000</td>
<td>&gt;31000 (b)</td>
</tr>
</tbody>
</table>
TABLE VII  
Rankings of 10 Selected Pesticides According to Three Criteria

<table>
<thead>
<tr>
<th>Pesticide</th>
<th>Rank based on toxicity</th>
<th>Rank based on persistence</th>
<th>Joint assessment (recovery time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lindane (organochlorine insecticide)</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Dimethoate (organophosphate insecticide)</td>
<td>8</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Parathion (organophosphate insecticide)</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Chlorpyrifos (organophosphate insecticide)</td>
<td>1</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Carbofuran (carbamate insecticide)</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Carbaryl (carbamate insecticide)</td>
<td>2</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Methomyl (carbamate fungicide)</td>
<td>7</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Benomyl (benzimidazole fungicide)</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Atrazine (triazine herbicide)</td>
<td>10</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Fentin (organotin fungicide)</td>
<td>9</td>
<td>5</td>
<td>9</td>
</tr>
</tbody>
</table>

*The score 1 is given for the highest toxicity (the 95% protection level referred to in the text), the greatest persistence (soil half-life), and the longest predicted ecotoxicological recovery time (i.e., the time it takes the pesticide concentration to decline to the 95% protection level).*

cult to find in the published literature, and they are not routinely collected for suites of compounds in such a way that they may be exploited within this new approach as a decision-making aid.

Table VII, adapted from van Straalen and van Rijn (1998), summarizes the differences in ranking of toxicity that result if pesticide toxicity to soil organisms and chemical persistence (half-life) in soil are integrated to predict the time when recovery could take place by organisms. A compound with high intrinsic toxicity, such as chlorpyrifos, has a lower ranking once its limited persistence is taken into account, whereas benomyl climbs higher up the rankings because of its persistence. Analytical procedures of this form are beginning to forge a link between the large amounts of laboratory-obtained toxicological data that exist and the field, where functioning communities are exposed to the toxin, rather than individual species.

A. Farm-Scale Observations of Ecotoxicological Impact

A number of research projects investigating the ecology of farming systems were initiated throughout the 1980s and 1990s in Europe (Holland et al., 1994). Pesticide impacts were investigated in all of these studies, and comparisons between conventional and reduced pesticide inputs were made in at least 13 sites, in Germany, The Netherlands, Switzerland, and the United Kingdom (Table VIII). In all cases, beneficial nontarget invertebrate densities were higher in areas where the total regime of pesticide input had been reduced. In the larger scale studies (e.g., the Boxworth study, UK), with 5.6- to 15.7-ha treatment area sizes, some beneficial species were rendered locally extinct for the full 3-year treatment phase of the project, and recovery was subsequently slow. The level of impact on individual species was shown to be a function of life history attributes that affected pesticide exposure and capacity of that organism to reinvade the treated area. Predatory capacity was inhibited in the highest pesticide regimes, and there was evidence that this contributed to higher pest densities in some years. Although similar data have been obtained for the invertebrate community of rice systems in the tropics, data sets of this level of complexity are rare in the investigation of pesticide impacts in agroecosystems.

These investigations have revealed the subtler impact of second- and third-generation active ingredients that followed the organochlorines. They reinforce the need to measure ecological impacts on scales that reflect agricultural practice and pesticide use in the real world and that are tuned to the scale of dispersive movement of nontarget taxa. They have also triggered considerable
TABLE VIII
Summary of Data from Farming System Experiment in the United Kingdom (Boxworth, SCARAB, TALISMAN, RISC, LIFE), Germany (INTEX), The Netherlands (Nagele), and Switzerland (Third Way): Results of Integrated Farming, with Reduced Pesticide Inputs, Compared with the Conventional Levels of Use

<table>
<thead>
<tr>
<th>Project</th>
<th>Beneficial arthropods</th>
<th>Birds and mammals</th>
<th>Earthworms</th>
<th>Soil microorganisms</th>
<th>Soil minerals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boxworth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCARAB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TALISMAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RISC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIFE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lautenbach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTEX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third Way</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Key: +, increase; =, no change; o, variable result over the period of the study.

interest in the mechanisms that underlie long-term depletions, even local extinction of certain species. Both scale of treatment and the mode and rate of dispersal of arthropods influence rates of population recovery following pesticide use, and the proximity of local refugia from which recovery can occur, and landscape features conducive to movement and colonization are important factors underlying extinction risk. Modeling may provide an appropriate tool for testing our understanding of invertebrate population processes at the agroecosystem level, but it does not substitute for the need to undertake a far greater number of manipulative experiments and monitoring programs that examine the spatial dynamics of nontarget organisms in sprayed farming systems.

B. Avian Impacts

The requirement for large-scale, field-based monitoring and experimentation is not restricted to terrestrial nontarget invertebrates. Avian toxicologists have long recognized that field-based studies provide unique and indispensable data for interpretation of pesticide effects on birds (Taub, 1997). Many of the most important effects of pesticides on birds, including eggshell thinning as a result of the bioconcentration of organochlorines through food chains, were originally detected and documented as a result of field-based observations. These effects also include endocrine disruption, food supply reduction or alteration, variation in sensitivity between species, and the synergistic effects of exposure to multiple compounds.

C. Aquatic Systems

Similar arguments may be applied to the investigation of pesticide effects on aquatic organisms. The recent decision in the United States to reduce the requirements for ecological data from multispecies test systems in the regulatory process has been criticized because it will reduce the probability of detecting biologically significant effects (Taub, 1997). These effects include indirect trophic-level impacts, compensatory shifts within a trophic level, responses that are associated with seasonal trends in populations, chemical transformations effected by organisms in the exposed system, and impacts that result from long persistence of either the parent product or toxic breakdown products.

In conclusion, there is now abundant evidence that pesticide impacts can evolve at the agroecosystem scale and that this requires the development of appropriately scaled monitoring or experimental systems. Toxicological data are of fundamental value in the initial evaluation of pesticides, but in their real-world applications, pesticides may also elicit ecological effects that reverberate through the system, long after the chemical residues have become undetectable. Laboratory-based, single-species tests are limited in their predictive power for
impacts in the field, and further development of the theory and methodology associated with ecological effects is required.

D. Tools for Pesticide Management

It is questionable whether we apply our knowledge about pesticides effectively at any stage following the regulatory permission to use a product in a specified way. Given the need to combine knowledge of chemical properties, fate and behavior, environmental attributes such as soil type, toxicology, and ecology, the challenge is considerable. In each of these disciplinary areas, however, there are considerable databases of knowledge and predictive models that can be used to interpret how a pesticide will behave in a given set of field conditions. Attempts have been made to summarize clusters of pesticide properties in databases that may be used to compare the environmental and ecological risks posed by lists of candidate compounds for specific uses. These databases are not specifically tuned to local conditions; they do, however, combine many factors, including risks to wildlife, in the rankings that they generate, and enhance the capacity of the end user to make more informed decisions about particular uses. The most significant development in this field is the proposal to derive Environmental Impact Quotients (EIQ) for compounds from established databases of pesticide properties (Kovach et al., 1992). The quotient combines farm-worker effects (built from acute and chronic toxicity and plant surface half-life) with consumer effects (built from aquatic and terrestrial ecotoxicology). There is considerable scope for toxicologists to develop approaches such as this from first principles in order to make data available in a form that can be used to assist decisions in the field. Guided by EIQs, growers and their advisors could then plan pest control tactics that attempt to minimize nontarget impacts and track their progress with a rigorous and quantitative methodology that could also be explained to consumers.

See Also the Following Articles
AGRICULTURE INVASIONS • AGRICULTURE, INDUSTRIALIZED • ECOTOXICOLOGY • HERBICIDES • INSECTICIDE RESISTANCE

Bibliography
PHARMACOLOGY, BIODIVERSITY AND

Paul Alan Cox
National Tropical Botanical Garden

I. Evolutionary Perspectives
II. History
III. Modern Chemical Approaches to Drug Discovery
IV. Rational Screens of Biodiversity
V. Nature of Biodiversity-Derived Pharmaceuticals
VI. Biodiversity and Ethics

GLOSSARY

bioassay A test for pharmaceutical activity in substance conducted either in living organisms (in vivo assays) or in the test tube (in vitro assays).

coevolution A heritable response made by a species through evolutionary time to different species which is then reciprocated.

ethnobotany The study of the uses of plants by different groups of people, often indigenous peoples.

extremophile A organism that lives and thrives in a habitat characterized by extremes of temperature, acidity, alkalinity, salinity, light, or pressure that would prove lethal to most other organisms.

fractionation A method of obtaining a pure compound by extracting components from a mixture with solvents of different solubility.

herbal A compilation of medicinal plants and their properties often used in earlier times as a reference for physicians prescribing medical treatments based on plant therapies.

indigenous intellectual property rights Rights to intellectual properties belonging to indigenous peoples, including but not limited to iconographical representations, terminologies, names, phrases, legends, methods, and techniques of traditional cultivation, healing, identification, preparation, and use of biodiversity.

indigenous peoples Peoples who have resided in the same geographical area for many generations and who possess legends, proverbs, geneologies, languages, and other unique cultural features linking them to the land.

phylogeny A family tree of related organisms tracing their evolutionary history.

tvoucher A representative specimen of a plant or animal that is properly collected, prepared, and preserved in a herbarium or museum to facilitate expert identification of the species.

PHARMACOLOGY, the science of drugs, studies the ways in which natural products can be employed as medicines. Since ancient times, humans have recognized the pharmaceutical properties of certain compounds derived from plants and animals. Current research now also recognizes that biodiversity must be maintained in order for the environment to continue to be a source of such medicinal substances.

I. EVOLUTIONARY PERSPECTIVES

Life-forms have evolved a plethora of complex biochemistry to help mediate their interactions with their physi-
Some of the resultant molecules function as materials for membranes and cell walls, photosynthetic pigments, energy storage compounds, neurotransmitters (in animals) and hormonal signals (in plants as well as in animals), defensive compounds, a variety of protective pigments, and as structural materials. The resultant juxtaposition of similarity and diversity of the chemical compounds created by different life-forms is striking. For example, genetic inheritance in all known life-forms is based on nucleic acids; although some enzymes used in replication differ with regard to prokaryotes and eukaryotes, all known living organisms use RNA or DNA to transmit their genetic information. However, differences in other biochemical pathways are profound. These differences result in differential vulnerability to toxic compounds, with ranges in LD50s of specific compounds varying several orders of magnitude between different phylogenetic groups. For example, the concentration of penicillin that completely stops cell wall formation and growth in Streptococcus bacteria shows little effect when administered to a centipede or to a geranium, whereas the dose of 2,4-D sufficient to kill a geranium would have negligible impact on either centipedes or bacteria.

Throughout the history of life on Earth, such differences in biochemical vulnerability have generated abundant possibilities for chemical coevolution between different species. Even though life on Earth appears to share a common phylogeny, biochemical differences between species leave open the possibility that a mutation in the biosynthetic pathway of one organism may yield a compound toxic to another. Selection for such protective molecules from predation would likely be greatest in organisms that are unable to flee or lack alternative forms of defense, including sessile organisms such as corals, sponges, tunicates, and other marine invertebrates and nearly all plants. Although poisonous birds have been found in New Guinea, toxic frogs are prominent in the Neotropics, and brightly colored toxic insects are known to most amateur naturalists, such animals appear to sequester toxic agents and motile predators or parasites, a plant species generates compounds effective against a wide variety of plant enemies. The presence of toxic compounds in a plant cannot, however, be considered prima facie evidence of an evolutionary response to predation. Bioactive molecules occur in plants as secondary metabolites as well as evolved chemical defenses against predation, fungal attack, microbial invasion, and viral infection. It is also unlikely that plant compounds known to be toxic to humans represent evolutionary responses to anthropogenic harvesting; given the omnivorous, generalist foraging patterns of higher primates, it is unlikely (with the possible exception of agricultural weeds) that any plants have evolved a specific biochemical response to toxic compounds. Even though life on Earth appears to share a common phylogeny, biochemical differences between species leave open the possibility that a mutation in the biosynthetic pathway of one organism may yield a compound toxic to another. Selection for such protective molecules from predation would likely be greatest in organisms that are unable to flee or lack alternative forms of defense, including sessile organisms such as corals, sponges, tunicates, and other marine invertebrates and nearly all plants. Although poisonous birds have been found in New Guinea, toxic frogs are prominent in the Neotropics, and brightly colored toxic insects are known to most amateur naturalists, such biochemical protection against predation by motile organisms appears to be the exception rather than the rule.

A plant species under pressure from a predator or parasite may perish unless individuals in the plant population possess, by mutation or some other evolutionary accident, a compound toxic or unpalatable to its enemies. If the mutation is heritable, progeny of such plants will increase in time through population, replacing those individuals previously lost to predation. If the species of predator or parasite are highly selective in their choice of prey, they may in turn become imperiled by such chemical innovations, unless individuals in the predator or parasite population possess a means of detoxifying the defensive chemistry generated by the plants. The resultant coevolutionary race between two species, sometimes termed a “red queen race” after the never-ending race described by Lewis Carroll in Alice in Wonderland, has been documented in the interaction between Passiflora vines and Heliconius butterflies in the Neotropics by Larry Gilbert and coworkers at the University of Texas. Other such tight coevolutionary links are the topic of much scientific interest. More common than species–species coevolution is what might be termed diffuse chemical coevolution. Instead of evolving defensive chemistry against a single species of predator or parasite, a plant species generates compounds effective against a wide variety of plant enemies. The presence of toxic compounds in a plant cannot, however, be considered prima facie evidence of an evolutionary response to predation. Bioactive molecules occur in plants as secondary metabolites as well as evolved chemical defenses against predation, fungal attack, microbial invasion, and viral infection. It is also unlikely that plant compounds known to be toxic to humans represent evolutionary responses to anthropogenic harvesting; given the omnivorous, generalist foraging patterns of higher primates, it is unlikely (with the possible exception of agricultural weeds) that any plants have evolved a specific biochemical response to toxic compounds. Even though life on Earth appears to share a common phylogeny, biochemical differences between species leave open the possibility that a mutation in the biosynthetic pathway of one organism may yield a compound toxic to another. Selection for such protective molecules from predation would likely be greatest in organisms that are unable to flee or lack alternative forms of defense, including sessile organisms such as corals, sponges, tunicates, and other marine invertebrates and nearly all plants. Although poisonous birds have been found in New Guinea, toxic frogs are prominent in the Neotropics, and brightly colored toxic insects are known to most amateur naturalists, such biochemical protection against predation by motile organisms appears to be the exception rather than the rule.

The emerging picture of chemical interactions between sessile prey and motile predators is compelling: what might otherwise appear to be a quiet forest glen or a tranquil coral reef may be the evolutionary equivalent of a battlefield rife with chemical warfare. Sessile prey appear to sequester toxic agents and motile predators develop detoxification strategies as a prodigious rate through evolutionary time. Since many of these bioactive molecules have relevancy to human disease, a coral reef or forest glen could also be said to resemble a large pharmaceutical storehouse, perhaps one that is in great disarray. Tens of thousands of vials of pharmacologically active compounds litter the ground, but the labels identifying the contents and their therapeutic utility have been lost. How can one determine which molecules are useful and for which diseases? Eloy Rodriguez and coworkers at Cornell University have found evidence that some vertebrates self-dose with pharmacologically active compounds produced by plants. Many mammals, including primates, engage in
self-medication when ill. Ongoing pharmacological self-experimentation by animals may have been observed and mimicked by early humans. Early Polynesian legends indicate that kava, a beverage made from the roots and rhizomes of Piper methysticum rich in tranquillizing secoiridoides and kava lactones, was discovered by watching the calming effect the roots had on rats. Legends from India claim that in ancient times mongooses were observed to feed on Rattus rattus before engaging in combat with cobras. Copying the reputed feeding behavior of the mongoose, local people found that the shrub could serve as a potent antidote to snake-bite. The veracity of such legends is unclear, but it is clear that to trace the use of plant-derived bioactive molecules in medicine we must first examine their use in prehistory by indigenous peoples.

II. HISTORY

A. Early Understandings

All known indigenous groups have ethnomedical traditions based, at least in part, on pharmacologically active compounds found in plants and animals. It is therefore safe to assume that early humans also developed significant ethnopharmacopeias; this assumption is corroborated by archaeological finds of medicine pouches in ancient burial sites. Results of successful experiments with pharmacologically active plants and animals were likely passed from generation to generation. This accumulation of oral information resembles in part the results of a vast, uncontrolled, and unwritten human bioassay experiment. Later knowledge about medicinal plants was transmitted in written form. By the fifth century B.C. Hippocrates codified and disseminated plant knowledge. As can be seen in De Materia Medica, two pieces of information were disseminated concerning each plant: its purported healing properties and its identification. Agnes Arber (1953) pointed out that the necessity for precise identification led to an unusual feature of early botanical iconography: for nearly 2000 years, plants in herbals were pictured with the part most likely to be in commerce—their roots. Although these earlier herbals were lovingly copied throughout the ages, the quality of illustrations declined until, in the fourteenth century, it had become increasingly difficult to identify the plants. Innovators such as Fuchs and Mattioli began to inscribe illustrations drawn from living plants on wood blocks for printing. This was a major boon for botanists since images reproduced by wood block printing did not suffer the same image deterioration in successive copies as occurred in hand-inked illustrations. The new herbals also contained current information gleaned from European folk medicine.

In the first century Padanius Dioscorides, a Greek physician, wrote a seminal work on medicinal plants entitled De Materia Medica, which describes more than 500 medicinal plants and includes many drawings. Until the beginning of the Renaissance, Dioscorides’ work was the final word in medicinal plants for more than 1000 years. Such early compilations of folk wisdom concerning medicinal plants were not confined to the West. In 2000 B.C., the Chinese emperor Chi’en Nung compiled the Pen Tsao, which is perhaps the earliest known herbal, whereas in India, Sanskrit texts on medicinal plants are said to date to 3000 years ago. In the Americas, the Mayans also wrote extensive manuscripts on medicinal plants. Unfortunately, many of these codices were burned by the early Spanish conquistadors.

B. Herbals and Medicinal Plants

An increase of scholarly interest in herbal and medicinal plants began during the Renaissance, when the first physic gardens (gardens of medicinal plants for therapeutic use and study) were planted in Italy and Germany. Most early Renaissance herbals were largely based on the work of Dioscorides’ book, De Materia Medica, with additions made from the author’s own knowledge. As can be seen in De Materia Medica, two pieces of information were disseminated concerning each plant: its purported healing properties and its identification. The new herbals also contained current information gleaned from European folk medicine.

In the English-speaking world, the first herbal is an eleventh-century Anglo-Saxon codex known as the Herbaldum of Aelredus Paonius. The earliest printed English herbal is an anonymous quarto from 1525 improved by Richard Bancks (as quoted in Arber, 1953, p. 41): “Here beyynth a newe mater, the whiche shew-
Calibew and treateh of y vertues and proprytes of herbes, the whiche is called an Herball.”

A year later, a translation of a French herbal was published by Peter Treversi, and in 1538 William Turner published Leelhau de re Herbaria Nova. In 1531, Henry F. Lyte published a translation of Dodonaeus’s famous herbal, and in 1539 Anthony Askham published a herbal derived from the 1523 English work of Banckes. However, the most popular of all sixteenth-century herbals was that of John Gerard, published in 1597. Born in 1545, at the age of 32 Gerard was appointed superintendent of the gardens of Lord Burleigh at the Strand in London. Later, he was appointed curator of the Physic Garden of the College of Physicians of London. A catalog of plants in his own garden appeared in 1596, and in 1597 he published The Herball, or General Natural History of Plants, one of the most quoted botanical works ever published. Gerard’s Herball, with 1392 pages and 2200 woodcut images of medicinal plants, was greeted with tremendous enthusiasm by medical practitioners and the general public.

The importance of Gerard’s Herball in the history of discovery of novel bioactive compounds from biodiversity cannot be overstated. His description of the medicinal properties of Filipendula—now called Spirea, a genus in the rose family—led to the isolation of salicin from the plant and later found that Filipendula serpeniophora has long been used in India to treat snakebite, insomnia, and as a sedative for hyperactive children. In 1931, Indian chemists isolated a variety of molecules from the plant and later found that Filipendula powder lowered blood pressure. In 1949, R. J. Vakil published a clinical study of Filipendula in the British Heart Journal. Following up on Vakil’s research, Dr. Emil Schlittler and Hans Schwarz at CIBA extracted from Filipendula roots an alkaloid, reserpine, which at low doses shows strong activity in lowering blood pressure. CIBA soon introduced reserpine to commerce. Reserpine had a direct effect on the hypothalamus and recently has been prescribed in combination with other antihypertensive drugs such as hydralazine hydrochloride.

III. MODERN CHEMICAL APPROACHES TO DRUG DISCOVERY

Digitalis was discovered through the same sequence of steps used in modern ethnobotanical drug discovery programs: (i) folk knowledge accumulates concerning possible pharmacological activity of a plant, (ii) the plant is used therapeutically by a healer, (iii) the healer communicates this knowledge to a scientist, (iv) the scientist collects and identifies the plant, (v) plant extracts are tested with a bioassay (a preliminary screen for desired pharmacological activity), and (vi) a pure com...
increasingly sophisticated techniques of vouchering potential new pharmaceuticals have been facilitated by (Apocynaceae). Synthetic analogs of reserpine from Rauvolfia serpentina (Apocynaceae), indicating new modes of pharmacological action that allows complete synthesis of novel analogs (e.g., vincristine from Catharanthus roseus [Apocynaceae]), and (iii) by providing chemical "building blocks" used to synthesize more complex compounds (e.g., the synthesis of oral contraceptives from diosgenin derived from Dioscorea floribunda [Dioscoreaceae]), and (iv) by indicating new modes of pharmacological action that allow complete synthesis of novel analogs (e.g., synthetic analogs of reserpine from Rauvolfia serpentina [Apocynaceae]).

Regardless of their origin, recent discoveries of potential new pharmaceuticals have been facilitated by increasingly sophisticated techniques of vouchering and species identification, bioassay, fractionation, and structural elucidation of molecular entities.

A. Vouchering

Proper vouchering in pharmaceutical research has sometimes been overlooked with disastrous effects because accurate identification of a species of plant or marine organism has been rendered impossible due to a missing or inadequate voucher specimen. The importance of adequate herbarium voucher specimens cannot be overstated: Any subsequent question or dispute concerning the identity of the species involved can be unequivocally settled by expert examination of a properly collected voucher specimen. Vouchers for fish and marine invertebrates are usually stored as dried or fluid-preserved (FAA or aqueous ethanol) specimens deposited in zoological or natural history museums. Specimens of algae, lichen, and vascular plants are preserved as dried specimens in herbaria. Because of their scientific value for the future (sophisticated bioassays of the future may require only micrograms of material), voucher specimens should be deposited and preserved in well-curated herbaria or museums with duplicate specimens deposited in geographically distant institutions, including institutions in the countries in which the material was collected. Addresses and contact information for herbaria throughout the world are listed in Index Herbariorum, published by the International Society of Plant Taxonomists and the New York Botanical Garden.

The collection number of the specimen should be used to label all subsequent pharmacological fractions and residues so that any new discovery or question can be immediately referred to the original specimen. Herbarium specimens should consist of the leaf, fruit, flower, and other plant parts necessary for proper identification by a botanist. Plants are pressed in newspaper and dried in the field using plant presses and a heater. When it arrives at the herbarium, the dried plant material is glued to high-quality rag bond paper and stored in well-curated herbaria or museums with duplicate specimens deposited in geographically distant institutions, including institutions in the countries in which the material was collected. Vouchers for fish and marine invertebrates can be immediately referred to the original specimen. Vouchers for fish and marine invertebrates have been considered by expert examination of a properly collected voucher specimen. Vouchers for fish and marine invertebrates are usually stored as dried or fluid-preserved (FAA or aqueous ethanol) specimens deposited in zoological or natural history museums. Specimens of algae, lichen, and vascular plants are preserved as dried specimens in herbaria. Because of their scientific value for the future (sophisticated bioassays of the future may require only micrograms of material), voucher specimens should be deposited and preserved in well-curated herbaria or museums with duplicate specimens deposited in geographically distant institutions, including institutions in the countries in which the material was collected. Addresses and contact information for herbaria throughout the world are listed in Index Herbariorum, published by the International Society of Plant Taxonomists and the New York Botanical Garden.

The collection number of the specimen should be used to label all subsequent pharmacological fractions and residues so that any new discovery or question can be immediately referred to the original specimen. Herbarium specimens should consist of the leaf, fruit, flower, and other plant parts necessary for proper identification by a botanist. Plants are pressed in newspaper and dried in the field using plant presses and a heater. When it arrives at the herbarium, the dried plant material is glued to high-quality rag bond paper and stored in herbarium cases. Properly stored, under the correct conditions, herbarium specimens can retain their scientific value for centuries. In the field, it is also important to take careful notes to accompany the voucher specimens, including precise information on the collection location and other information on the plant, such as floral odor or color, that might be lost during drying. The collector's name and date of collection should always be included. For plants or animals used in folk medicine, ethnobiological data on the diseases treated, mode of formulation, and methods of
Bioassays vary with regard to type and precision. In vivo bioassays involve administering the test substance to a living organism to determine the substance’s pharmacological activity. One of the more common in vivo bioassays that has been used to great advantage by some investigators is a brine shrimp bioassay, in which the test substance is administered to a culture of brine shrimp, available from any aquarium supplier. The percentage death of the brine shrimp within a short period is recorded, with brine shrimp mortality yielding a crude measure of cytotoxicity. Other in vivo bioassays include the Hippocratic screen, a test protocol for oral or peritoneal administration of the test substance in rats or mice. A variety of different behavioral and physiological parameters are measured before and after administration, allowing type and mode of pharmacological activity to be adduced. The most important in vivo bioassays for pharmaceutical research are human clinical trials, which are almost always administered after reasonable expectations of safety and efficacy have been determined from animal trials. The power of in vivo assays is their ability to reveal unexpected or novel modes of pharmacological activity since the impact of the test substance on the whole organism can be deduced.

More common in modern pharmaceutical use, though, are in vitro assays—bioassays that do not depend on a living organism. Although some broad-based in vitro assays depend on living tissues or cells, such as the guinea pig ileum, frog sciatic nerve, or human cancer cell lines, far more common is the use of target-specific molecular assays. In such assays, a single point of action in a given biochemical pathway, such as the role of the enzyme phospholipase A2 in mediating cellular inflammation, has already been identified. Various substances are then screened to determine if they impact the enzyme, receptor, site, or rate-limiting step of interest. Modern in vitro assays are typically costly (given the immense research effort required to identify the best target along a given biochemical pathway), proprietary, and designed for high-throughput rapid testing of numerous samples for the desired pharmacological activity.

In a typical high-throughput bioassay, minuscule portions of a given sample are tested for an enzymatic color change when combined with the test system. These bioassays test by screening, using a computer-controlled camera, hundreds and even thousands of different samples deposited in the microwells of a glass dish that are manipulated by robotic arms. The advantages of modern high-throughput bioassays are their speed, efficiency, repeatability, and specificity. The number of false positives (initial indications of pharmacological activity that do not pan out in subsequent testing) in modern receptor-binding assays is low; however, such high specificity entails a research liability since (unlike tests on living animals) a molecular bioassay is unlikely to reveal any mode of pharmacological action other than the specific activity it is designed to detect.

The use of precise high-throughput bioassays designed by molecular biologists has resulted in some spectacular successes, such as the discovery of the cholesterol-lowering drug Zocor or the new generation of anti-inflammatory drugs based on COX-2 inhibition. However, such screens almost certainly miss many important modes of action. It is unlikely, for example, that the psychotropic action of LSD-25 or the use of Viagra to treat male impotency would have ever been discovered had their effects not been noted through inadvertent administration to living human beings.

C. Fractionation

Once a “hit” or pharmacological activity is detected in a test substance during a bioassay, the substance, particularly if it is a natural product of unknown chemical composition, must undergo fractionation. During fractionation, solvents of different polarity are used to produce increasingly refined chemical samples of the test substance. Each fraction is then tested against the original bioassay to determine its bioactivity. Fractions exhibiting significant bioactivity are further fractionated until eventually a purified compound is obtained. High-pressure liquid chromatography, gel electrophoresis, and other chemical techniques are also employed in addition to solvent fractionation to purify samples.
This approach to isolating pure bioactive molecules by tracing bioactivity through the purification process is called bioassay-guided fractionation. Typical yields of pure bioactive molecules under research settings are low: Depending on the nature of the active molecule, 1 kg of dried plant material can yield as little as 1–10 μg of pure compound. Thus, on detection of a hit in a bioassay, recollection of a bulk sample (10–40 kg.) is often required to produce sufficient pure substance for structural determination.

**D. Structural Determination**

Once a pure bioactive molecule is obtained through fractionation or chromatographic techniques, the next step is to determine its molecular structure. A definitive technique for determining the structure (if the purified molecule can be crystallized) is X-ray crystallography, but more common techniques involve the use of high-pressure liquid chromatography, gas chromatography (GC), mass spectrometry (MS), or, for novel structures, nuclear magnetic resonance (NMR) spectroscopy. Most modern GC-MS machines have a computer-based library documenting retention time of known molecules: Observed peaks on the GC-MS are quickly and easily compared to those in the computer library. This technique is particularly useful for forensic toxicologists who do not typically expect to find novel molecules in their work. Structural determination in most biodiversity studies, however, relies on NMR since novel rather than known structures are often discovered during the course of pharmaceutical research. The frequencies and resolution capabilities of NMR machines differ, as do the types of experiments possible on NMR equipment, but of particular interest to natural product investigators is the INADEQUATE experiment that, through supercomputer analysis, can resolve structures for molecules in substances that are parts of mixtures rather than pure samples.

**IV. RATIONAL SCREENS OF BIODIVERSITY**

Sometimes, such as in the case of penicillin, discovery of new pharmaceuticals from biodiversity is the result of serendipity. More often, through success in natural product research is the result of carefully designed screening processes. Such planned search strategies to select appropriate organisms for study are termed “rational screens,” of which there are several varieties: random, phylogenetic, ecological, and ethnobotanical.

**A. Random Screens**

Random screens, despite their name, involve a great deal of planning and design. To be successful, they must have a clearly defined disease target and a clear and unambiguous bioassay. One of the best examples of a random screen in recent times is the effort of the Natural Product Branch of the U.S. National Cancer Institute (NCI) to evaluate plants and animals for anticancer activity. Using a variety of in vitro cancer cell cultures as bioassays, the NCI screened more than 35,000 plant accessions and 6000 marine organisms for activity against human cancers. Although the screen was not truly random in that it was designed to sample a wide range of geographical and taxonomic diversity, it allowed the NCI (with a variety of contracting collecting institutions) to assemble a large archive of materials that could be retested as bioassays continued to evolve. The geographical reach of the NCI was truly broad and led to pioneering efforts to negotiate international protocols for biodiversity prospecting. The NCI “letter of intent” which collaborators used to obtain collecting permits in foreign countries pledges equitable sharing of data and economic benefits with host countries. This letter has emerged as a model for recent international agreements.

Although the low hit rate of the NCI random screens led to criticism of the approach, few can argue with the stellar success of the NCI program in producing what may be one of the most important chemotherapeutic agents discovered in the latter part of the twentieth century, taxol, which was initially isolated from the Pacific yew tree (Taxus brevifolia) and which has been licensed by the NCI to Bristol Meyers and approved by the Food and Drug Administration (FDA) for the treatment of ovarian and breast cancer. However, because of the low success rates and high costs of collecting random screens are currently suitable only for large institutions, even given the advent of high-throughput bioassays.

**B. Phylogenetic Screens**

Phylogenetic screens involve the pharmacological testing of related groups of organisms. Increased precision in elucidating phylogenies, largely due to rapid computer programs for cladistic analysis and the advent of molecular techniques for phylogenetic determination, facilitates the identification of relatives of any species showing pharmacological value. Phylogenetic screens, albeit in a crude sense, have long been utilized. For example, plants in the Apocynaeae, or milkweed family,
Ecological Screens

An ecological screen depends on the environmental setting of the organism for clues as to its possible pharmacetical value. Recently, such screens have focused on extremophiles—organisms that survive under extreme conditions of high temperatures, acidity, alkalinity, or salinity. For example, the polymerase chain reaction, invented by Kary Mullis at Cetus Corporation and widely used for genetic studies, medical diagnosis, or gene therapy, depends on an enzyme, Taq polymerase, derived from a bacterium discovered in the thermal springs of Yellowstone National Park by Thomas Brock of the University of Wisconsin. The bacterium that produces the enzyme, *Thermus aquaticus*, can withstand temperatures up to 95°C, and its enzymes are similarly heat resistant, an important feature for the rapid thermal cycles used in the polymerase chain reaction.

Ecological screens have also been conducted under the sea near hydrothermal vents where temperatures can reach in excess of 350°C. The enzyme Pfu polymerase, derived from the archaeon *Pyrococcus furiosus* and discovered by Karl Stetter of the University of Regensburg during exploration of undersea thermal vents, is of increasing interest because of its precision in polymerase chain reactions due to a built-in DNA repair mechanism. *Pyrococcus furiosus* can survive temperatures in excess of 100°C and is resistant to high pH and radioactivity. Enzymes derived from other organisms that survive conditions of extreme cold, alkalinity, acidity, or salinity are of increasing interest to the biotechnology industry (Madigan and Marrs, 1997). Other ecological screens focus on searches for antifungal compounds from plants that grow in moist tropical environments (successes in this area have been achieved by Alice Clark at the University of Mississippi) as well as searches for inexpensive mollusccidal compounds useful in the fight against schistosomiasis. Kurt Hostettmann at the University of Lausanne discovered saponin-like compounds in several African tree species that show intense toxicity to schistosomiasis-carrying snails, and he is investigating the possibility of supplying villagers with seeds and the simple techniques to prepare the crude solutions of these compounds which could be used to dose village bathing and washing areas.

D. Ethnobotanical Screens

An ethnobotanical screen is based on the belief that indigenous peoples through generations have accumulated useful information about pharmacetically active plants and animals. Since ethnobotanical leads have resulted in the discovery of the active compounds for 23% of all prescription drugs, there is little question as to the efficacy of ethnobotanical approaches to drug discovery. However, although there are strengths to ethnobotanical screens, there are also significant limitations.

In general, cultures with three characteristics seem most likely to have discovered pharmacologically active plants: (i) residency within an area of high biodiversity, (ii) an extended history of residence within the area, and (iii) a cultural mechanism for accurate transmission of ethnomedical information from generation to generation. Under the first criterion, the Kayapo Indians of Brazil merit more attention than the Alyuts in Alaska because of the greater biodiversity of tropical Brazil. Under the second criterion, the Aboriginal peoples of North Arnhem land who have been resident in Australia for thousands of years would be of more interest than the Pitcarin islanders who arrived in their island home only a few centuries ago. Under the third criterion, the Samoan islanders, with a precise matrilineal system of transmitting healing knowledge from generation to generation, would be of more interest that faith healers in the Philippines, who depend on dreams for their knowledge of healing plants.

In an ethnobotanical screen, healing information is recorded by a linguistically adept ethnomobotanist trained in both the techniques of ethnographic interviews and field botany. Often, there is little overlap between indigenous and Western disease concepts—for example, there is no Western equivalent of “susto” (a type of soul loss with diverse symptoms) which afflicts indigenous peoples in Central America—and so the ethnombotanist, often in working with a physician, must be adept at mapping signs and symptoms of indigenous disease states into concepts of illness understood by Western peoples. Those plants identified by the indigenous peoples as containing healing properties are collected and identified, with samples prepared for bioassay-guided fractionation. In addition, careful notes about possible adverse reactions, dosages, preparation, and application
than 100 having been commercially produced (Samuelsson, 1992). Only a single “natural” penicillin has ever been prescribed as a drug, but more than 100 semisynthetic penicillins have been produced by adding other carboxylic acids to the basic penicillin structure (Samuelsson, 1992). The structure of penicillin, a peptide with a thiazolidine ring, is similar to the structure of the cephalosporins, in which the thiazolidine ring is substituted by a dihydrothiazine ring. Bacillus brevis and Bacillus licheniformis have been used to produce gramicidin and bacitracin, respectively.

Many new antibiotics and other pharmaceutical products have been obtained from soil samples which are carefully cultivated in sterile petri dishes and fermentation vats. One of the more successful products to be produced from a soil culture is cyclosporin, a cyclic peptide produced from the fungus Cylindrocarpon lucidum, originally derived from a Norwegian soil sample. Cyclosporin, a powerful immunosuppressant, was initially seen to have little therapeutic value and remained archived as a refrigerated fungal culture at Sandoz Pharmaceuticals (now Novartis) in Basel, Switzerland until mycologist Michael Dreyfuss convinced management that the product had utility. Today, nearly every heart or kidney transplant patient in the world takes a daily dose of cyclosporin to prevent rejection of the transplanted organ.

Soil samples continue to yield new and interesting bacteria and fungi, whereas exploration of extreme habitats promises to yield important new useful compounds, including those produced by Archea, distant relatives of bacteria. Using recombinant DNA technology, microorganisms can be employed to produce pharmaceutical compounds which they do not ordinarily produce in nature. Genetic engineers can insert DNA sequences into microorganisms to produce a staggering array of pharmaceutically active compounds. Although most researchers hail such genetically modified organisms as a significant advance, there are public concerns, particularly in Europe, about the safety of food and pharmaceuticals produced by genetically modified organisms.

V. NATURE OF BIODIVERSITY-DERIVED PHARMACEUTICALS

A. Microbial Products

In 1929, Alexander Fleming discovered that a petri dish of Staphylococcus had become contaminated with a Penicillium fungus: around the fungus in the culture medium was a zone resistant to Staphylococcus growth. During World War II, the antibiotic penicillin was isolated. This led to dramatic treatment of infections from wounds and other trauma. Since that time, more than 6000 different antibiotics have been isolated, with more than 1000 having been commercially produced (Samuelsson, 1992). During World War II, the antibiotic penicillin was isolated. This led to dramatic treatment of infections from wounds and other trauma. Since that time, more than 6000 different antibiotics have been isolated, with more than 1000 having been commercially produced (Samuelsson, 1992).

B. Marine Organisms

Marine organisms, particularly invertebrates, are known to produce extraordinarily toxic compounds. Currently, three compounds isolated from marine resources are in preclinical development as possible anti-cancer agents by the NCI. The marine algal species Poricia hornemaii from the Philippines has yielded a drug called halomon, a species of the sponge genus Lissodendoryx from New Zealand has produced hal-
products derived from marine organisms. However, perhaps the most promising pharmaceutical product derived from marine biodiversity is bryostatin, a macrocyclic lactone derived from the Californian bryozoan Bugula neritina. Bryostatin has been evaluated in phase II clinical trials in which it has been found to inhibit melanoma, reticulum cell sarcoma, and lung carcinoma. Initial results with human patients have been encouraging but not spectacular. Nevertheless, because of its unique structure and immunomodulatory effects, it is likely that the discovery of bryostatin will lead to new therapeutic concepts in chemotherapy. Currently, 6,000 marine samples are archived at the NCI for testing, and there is continuing industrial interest in natural products derived from marine organisms.

C. Terrestrial Animals

Terrestrial animals have historically played an important role in the production of certain therapeutic hormones and steroids. Progesterone was first isolated from pig ovaries in 1934. Animals were eclipsed by plants, however, as a source of steroidal precursors when in 1940 Russell Marker discovered that Mexican yams of the genus Dioscorea could be utilized to produce diosgenin. There has also been a long-term research interest in venoms produced by snakes as sources of pharmaceuticals. Ancrod, a drug used to treat circulatory diseases, is isolated from the venom of the Malayan pit viper Agkistrodon rhodostoma. Recent research, such as that conducted by Francis Markland at the University of Southern California, suggests that proteins such as conrotrostatin found in certain snake venoms may be able to fight cancer tumors. The poisonous secretions of some toads and frogs contain bufotalin, which was traditionally used as a treatment for dropsy before Withering's discovery of digitalis. Ethnobotanist Wade Davis has also found that secretions of species of Bufo were used as psychotropic substances by indigenous peoples in Central America. Epibatidine, from the poison dart frog, is being developed as a possible analgesic. It is likely that other venoms and stings will produce new pharmacological insights.

D. Terrestrial Plants

Plant-derived drugs are used for a broad spectrum of diseases and include quinine (for malarial suppression), digitalis (for treatment of rapid atrial fibrillation), vincristine (for treatment of pediatric leukemia), tubocurarine (a muscle relaxant used in anesthesia), pilocarpine (used for the treatment of glaucoma), and y-strophantin (used for congestive heart failure) (Table I). However, in the latter quarter of the twentieth century, very few plant-derived drugs were released to market, even though there was an increase in consumer interest in North America, Europe, and Japan in “natural” plant medicines. Given that less than 1% of the world’s plants have ever been carefully studied for pharmacological activity, it is puzzling that pharmaceutical firms are not rigorously investigating plants as the sources of new pharmaceuticals.

Flowering plants and ferns have traditionally provided the bulk of biodiversity-derived pharmaceuticals, but recent advances in molecular biology and combinatorial chemistry have reduced the pharmaceutical industry’s ardor in studying plants. The popular mythology of industry-supported botanists combing the rain forests for new plant-based leads is simply not true; currently, there is not a single major pharmaceutical firm that lists among its employees a PhD botanist or ethnobotanist who works in a drug-discovery program based on plants, although some firms such as Pfizer and Merck have contracted with other institutions to conduct plant surveys. Despite the lack of interest by the pharmaceutical industry, both the promise and the impact of plant-derived pharmaceuticals is profound. The World Health Organization estimates that 85% of the world’s population depends directly on plants for medicine, and more than 25% of current prescription drugs have at least one active component derived from a flowering plant. The former statistic, derived principally from developing countries, has contributed significant opportunities to contribute to world health by assisting different nations to evaluate the safety and efficacy of their own pharmacopoeias. Currently, the People’s Republic of China, Mexico, Thailand, and Nigeria have decided at the national level to incorporate traditional plant-based medicine directly into primary health care. An example of the possibilities inherent in this approach has been demonstrated in Thailand, where a potent anti-inflammatory compound from the beach plant Ipomoea pes-caprae—long used by fisherman to treat stings from the Portuguese man-of-war—was isolated and purified using bioassay-guided fractionation and its structure determined using NMR spectroscopy. However, rather than testing a pure compound, the Thai researchers conducted a careful, double-blind controlled study of a crude tincture of the plant. It was found to be both safe and efficacious. The resultant tincture can now be bought in any Thai drugstore for a price far less than that of a synthetic drug. There are, however, several plant-based pharmaceuti-
### Table 1

<table>
<thead>
<tr>
<th>Drug</th>
<th>Medical use</th>
<th>Plant source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apomorphine</td>
<td>Heart arrhythmia</td>
<td>Rauvolfia spp.</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Analgesic, fever</td>
<td>Spirea ulmaria</td>
</tr>
<tr>
<td>Atropine</td>
<td>Ophthalmology</td>
<td>Atropa belladonna</td>
</tr>
<tr>
<td>Beetroot</td>
<td>Oral disintegrant</td>
<td>Styrax cinnamomatus</td>
</tr>
<tr>
<td>Caffeine</td>
<td>Stimulant</td>
<td>Canella sinensis</td>
</tr>
<tr>
<td>Camphor</td>
<td>Rheumatic pain</td>
<td>Cinnamonum camphora</td>
</tr>
<tr>
<td>Capsaicin</td>
<td>Purgative</td>
<td>Rhaetum pachyanum</td>
</tr>
<tr>
<td>Cocaine</td>
<td>Ophthalmic anesthetic</td>
<td>Erythroxytocus</td>
</tr>
<tr>
<td>Codeine</td>
<td>Analgesic, antimonial</td>
<td>Papaver somniferum</td>
</tr>
<tr>
<td>Colchicine</td>
<td>Gout</td>
<td>Colchicum autumnale</td>
</tr>
<tr>
<td>Demerol</td>
<td>Leukemia, lymphoma</td>
<td>Colchicum autumnale</td>
</tr>
<tr>
<td>Dorsolipope</td>
<td>Hypertension</td>
<td>Rauvolfia cuneanescens</td>
</tr>
<tr>
<td>Doxycycline</td>
<td>Thrombotic</td>
<td>Meldos officinalis</td>
</tr>
<tr>
<td>Ditosan</td>
<td>Arral fibrillation</td>
<td>Digitalis purpurea</td>
</tr>
<tr>
<td>Ditosol</td>
<td>Arral fibrillation</td>
<td>Digitalis purpurea</td>
</tr>
<tr>
<td>Emerine</td>
<td>Anemic dysentry</td>
<td>Cephalis ipecachuanha</td>
</tr>
<tr>
<td>Ephedrine</td>
<td>Bronchidilator</td>
<td>Ephedra sinica</td>
</tr>
<tr>
<td>Eugenol</td>
<td>Toothache</td>
<td>Syzygium aromaticum</td>
</tr>
<tr>
<td>Galtostine</td>
<td>Hemorrhoid suppressor</td>
<td>Hamamelis virginiana</td>
</tr>
<tr>
<td>Hyosophamine</td>
<td>Anticholinergic</td>
<td>Hyoscyamus niger</td>
</tr>
<tr>
<td>Ipecac</td>
<td>Emetic</td>
<td>Cephalis ipecachuanha</td>
</tr>
<tr>
<td>Ipratropium</td>
<td>Bronchodilator</td>
<td>Hyoscyamus niger</td>
</tr>
<tr>
<td>Methone</td>
<td>Analgesic</td>
<td>Papaver somniferum</td>
</tr>
<tr>
<td>Nicomine</td>
<td>Antimuscarine</td>
<td>Papaver somniferum</td>
</tr>
<tr>
<td>Papain</td>
<td>Antidepressant</td>
<td>Carica papaya</td>
</tr>
<tr>
<td>Papaverine</td>
<td>Antimuscarine</td>
<td>Papaver somniferum</td>
</tr>
<tr>
<td>Physostigmine</td>
<td>Glycoside</td>
<td>Physostigminum venenosum</td>
</tr>
<tr>
<td>Protonium</td>
<td>Barbataria antidiote</td>
<td>Anamaria cuculld</td>
</tr>
<tr>
<td>Procaine</td>
<td>Bronchodilator</td>
<td>Hyoscyamus niger</td>
</tr>
<tr>
<td>Prostagline</td>
<td>Bronchodilator</td>
<td>Hyoscyamus niger</td>
</tr>
<tr>
<td>Prostigline</td>
<td>Bronchodilator</td>
<td>Hyoscyamus niger</td>
</tr>
<tr>
<td>Protopine</td>
<td>Bronchodilator</td>
<td>Hyoscyamus niger</td>
</tr>
<tr>
<td>Quinidine</td>
<td>Bronchodilator</td>
<td>Hyoscyamus niger</td>
</tr>
<tr>
<td>Quinine</td>
<td>Bronchodilator</td>
<td>Hyoscyamus niger</td>
</tr>
<tr>
<td>Ruscuscale</td>
<td>Bronchodilator</td>
<td>Hyoscyamus niger</td>
</tr>
<tr>
<td>Sennoside</td>
<td>Aminine, B</td>
<td>Canella sinensis</td>
</tr>
<tr>
<td>Scopolamine</td>
<td>Motion sickness</td>
<td>Canella sinensis</td>
</tr>
<tr>
<td>Stegnoside</td>
<td>Steroidal precursor</td>
<td>Canella sinensis</td>
</tr>
<tr>
<td>Stephoside</td>
<td>Congestive heart failure</td>
<td>Canella sinensis</td>
</tr>
</tbody>
</table>

**Continued**

<table>
<thead>
<tr>
<th>Drug</th>
<th>Medical use</th>
<th>Plant source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temposide</td>
<td>Bladder neoplasms</td>
<td>Pedophyllum pellatum</td>
</tr>
<tr>
<td>THC</td>
<td>Anismetic</td>
<td>Cannabis sativa</td>
</tr>
<tr>
<td>Theophylline</td>
<td>Diuretics, asthma</td>
<td>Canella sinensis</td>
</tr>
<tr>
<td>Tofroline</td>
<td>Surgery, relaxant</td>
<td>Stephoside gracis</td>
</tr>
<tr>
<td>Toluritane</td>
<td>Muscle relaxant</td>
<td>Chondrodendron tomentosum</td>
</tr>
<tr>
<td>Vinblastine</td>
<td>Phlastours disease</td>
<td>Catharanthus roseus</td>
</tr>
<tr>
<td>Vinoreline</td>
<td>Pediatric leukemia</td>
<td>Catharanthus roseus</td>
</tr>
<tr>
<td>Xanthotoxin</td>
<td>Viscous</td>
<td>Ammi majus</td>
</tr>
</tbody>
</table>

**Continued**

tics that have been recently released or are under current preclinical or clinical evaluation. Taxol, discovered during the NCI random screen, has received FDA approval for treatment of breast and ovarian cancer. For AIDS, currently five plant-derived compounds are in preclinical development. Derived from the resin of Calophyllum languarium in Malaysia, calanolide A has shown intense activity against HIV-1, as has costatolide derived from a related tree Calophyllum tenuissim from the same region. Michellamine B, from the leaves of the Cameroon vine Ancistrocladus korupensis has also shown anti-HIV activity, as has conocurvone derived from Conospermum shrubs found in Western Australia. Because of its use in indigenous medicine, the fifth member of this anti-AIDS quintet, prostratin, perhaps deserves in-depth discussion as a case study of recent developments in this field.

For many years, Samoan healers have used the stem wood of a small understory rain forest tree, Homalanthus nutans, to treat a disease known as "fiva samasama," characterized by yellowing of the eyes, dark urine, jaundice, and fevers (determined to be hepatitis). To prepare the remedy, the wood of Homalanthus nutans is macerated into a clean cloth, which, like a tea bag, is immersed into boiling water. After steeping, the contents of the tea bag are discarded, and the resultant tea is drunk by the patient. Samples of the healer preparations and the plants were tested against an in vitro HIV-1 bioassay in a collaboration between Paul Cox and NCI researchers Kirk Gustafson, John Cardellina, John Beutler, Peter Blumberg, Gordon Craig, Michael Boyd, and others at the Natural Products Branch of the NCI (Gustafson et al., 1992). The healer mixture and the crude plant extracts were found to protect cells from death by HIV-1, even though there was no evidence of reverse transcriptase inhibition or other marked inhibition of the mixtures to the virus. Bioassay-guided fractionation
yielded a pure compound, prostratin. Although phorbols similar in structure to prostratin are potent tumor promoters, in vivo studies by Peter Blumberg at the National Institutes of Health show prostratin to be an antipromoter, even though it activates protein kinase C. The NCI has advertised prostratin as a potential new component of a combination therapy for AIDS. The NCI will require any pharmaceutical firm which desires to develop the compound to negotiate directly with the Samoan government for a fair and equitable return of a portion of any royalties. In addition, Brigham Young University honored an agreement negotiated between Cox and the Falealupo village in Samoa prior to discovery, promising a minimum of 30% of any royalty income from prostratin to be returned to the village. However, to date there has been little interest by the pharmaceutical community in prostratin. Although this is primarily due to the approval of effective proteases and reverse transcriptase inhibitors, the fact that prostratin is a phorbol and its discovery was linked to traditional medicine contributed to the wariness of pharmaceutical firms. However, as resistance to reverse transcriptase inhibition and other antiviral strategies continues to evolve in the AIDSs virus, compounds with a cellular mode of action such as prostratin may gain the interest of the pharmaceutical industry, as may other compounds with broad-spectrum antiviral activity. Thus, although prostratin currently languishes in the limbo of promising drug leads that have not been clinically investigated, in the future it may be developed as a potent antiviral compound.

VI. BIODIVERSITY AND ETHICS

A. The Convention on Biodiversity

The Convention on Biodiversity, often referred to as the Rio Treaty, provided a broad basis in international law for the discovery and development of pharmaceutical compounds from biodiversity. In order to promote the conservation and study of biodiversity, the parties to the Convention on Biodiversity agreed that each nation has sovereignty over its own biological resources. Collection and development of these resources can thus proceed only with express permission of the national government. What initially was seen as a mechanism to accelerate investigation and conservation of biological resources by facilitating international cooperation has been criticized by some as an impediment to research. Many countries, fearing uncompensated exploitation of their national biological resources, have either closed their doors to international scientific exploration or caused the process of obtaining research permits to be extraordinarily slow and difficult. Given the history of colonial exploitation of biological resources, with little thought given to equitable sharing of benefits of discoveries based on biodiversity, such reaction by developing countries is understandable. If the processes of biodiversity were static and extinction quiescent, closing national boundaries to scientific investigation would have little long-term consequence; however, given the rapid rate of extinction (the International Union for the Conservation of Nature estimates that one-eighth of the world’s plants are threatened or endangered), nations that close their doors to scientists may unwittingly ensure that their biodiversity treasure will vanish without any significant appraisal of its economic value. An interesting exercise would be to compare the relative difficulties of obtaining permits to collect and pharmacologically analyze 1-kg plant samples from various countries versus the difficulties in obtaining permits to clear-cut and export entire rain forests: too often, those who destroy biodiversity are facilitated while those who study it are hindered.

Given the relative lack of interest of the pharmaceutical industry in investigating biodiversity as a source of new pharmaceuticals, and the occasionally unreasonable expectations placed on pharmaceutical firms for large initial payments and high percentages of royalties, many pharmaceutical firms that once had nascent interest in natural product chemistry have retreated to computer modeling and combinatorial chemistry as the sole engines of pharmaceutical discovery. Although this situation may change if new, important drugs are discovered from biodiversity, it is crucial that investigators adhere to high ethical standards. Training local scientists, obtaining informed consent from village chiefs or elders prior to initiating research, depositing duplicates of all collected specimens with local or regional herbaria, and vigorously protecting indigenous intellectual property help promote understanding and collaboration. Investigators who fail to heed such standards or, worse, ignore local laws and international treaties imperil not only their own research programs but also conservation of the very biodiversity they purport to study.

An interesting, and perhaps unintended, consequence of the Convention on Biodiversity has been its impact on the rights of indigenous peoples. By granting sovereignty over all biological resources to national governments, the convention in a sense can disenfranchise indigenous peoples, particularly in countries with a history of mistreating them. Is it reasonable that a plant
that has been developed and studied through generations of indigenous peoples should be considered the sole property of a distant (and perhaps hostile) national government? However, by stating in article 8(j) that indigenous peoples bear some consideration, the architects of the Convention on Biodiversity provided indigenous peoples with one of the first broadly accepted international treaties that explicitly recognizes their existence and aspirations. Article 8(j) requires signatory nations to (i) respect, preserve, and maintain traditional knowledge; (ii) promote wide application of traditional knowledge; and (iii) encourage equitable sharing of benefits from traditional knowledge. Many indigenous groups hope that their voices will increasingly be heard and considered at meetings of the Parties to the Convention.

B. Bilateral Agreements on Pharmaceutical Research

There have been attempts to directly negotiate with national governments for rights to perform pharmaceutical research on biodiversity. One of the best known of these arrangements was a remarkable agreement entered into by the government of Costa Rica and Merck Pharmaceuticals. Under the IMBIO agreement, Merck received rights to survey the flora and fauna of Costa Rica in return for a share of the proceeds of any new discovery and for a payment of $1 million used to support the training of local investigators and development of a national inventory of Costa Rican biodiversity. Although no discoveries have yet been announced from Merck, the Costa Ricans have been very innovative in training local people in biological survey techniques. Touted "parataxonomists," these local people, often with profound folk knowledge of organisms in their environment, are trained in proper techniques of collection and vouchering. By devoting the bulk of the Merck funds to biodiversity research, the government of Costa Rica has guaranteed that the people of Costa Rica and indeed the world will benefit for many years.

C. Indigenous Intellectual Property Rights

In ethnobotanical approaches to drug discovery, it is important that the contributions of indigenous people be appropriately recognized and compensated. Often, indigenous peoples are considered at meetings of the Parties to the Convention. Informed consent of traditional leaders is crucial when discussions are held concerning fair and equitable returns for indigenous intellectual property. Traditional leaders cannot make informed decisions concerning biodiversity research unless they possess accurate information concerning the nature and value of the research. Rather than imposing Western values of fairness and equability on indigenous peoples, investigators should regard indigenous peoples as coequals in any negotiations and seek solutions that meet indigenous rather than Western concepts of fairness and equability. Often, traditional leaders are not familiar with Western culture or conversant in Western languages; instead, they have very distinct views about appropriate behavior and dress within a village. Respect can be communicated by completely informing such leaders and councils about the scope, nature, and possible impact of biodiversity research. Informed consent of traditional leaders is crucial when discussions are held concerning fair and equitable returns for indigenous intellectual property.

For many indigenous peoples, conservation is a keenly felt need, a means of both perpetuating their culture and demonstrating respect to both the earth and to their ancestors. Innovative approaches to assisting indigenous peoples in conservation can be successful if indigenous peoples truly control the conservation initiatives. The Terra Nova biomedical reserve in Belize or the Falealupo Rain Forest Reserve in Samoa are examples of indigenous conservation begun by Western scientists that are now completely controlled by indigenous peoples. Culturally appropriate acknowledgment of indigenous collaborators such as including publication of scientific abstracts and reviews in indigenous languages and coauthorship of papers and books with indigenous collaborators are examples of the best practices currently employed by ethnobotanists. Protection of indigenous financial interests often requires vigorous efforts on the part of Western scientists, who need
to assist indigenous leaders in dealing with lawyers, commercial firms, or governments that may not completely appreciate indigenous property rights. Entering into any collaborative project, including ethnobotanical research, with indigenous peoples imposes a responsibility on the part of scientists to protect indigenous rights. It should not be surprising, then, that some of the most persuasive Western advocates of indigenous rights have been scientists who have lived and worked closely with indigenous peoples.

D. Rights of the Sick and Afflicted

In an era in which life itself increasingly appears to be a patentable commodity, it might appear quaint to raise the possibility that stakeholders in biodiversity research include others than sovereign nations, pharmaceutical firms, and indigenous peoples. However, moral rights are not always harmonious with property rights: Most societies, for example, believe that small children have an inalienable claim on their parents for support, even though children do not possess legal title to any of their parents possessions, and in former eras the sick and afflicted were believed to have certain claims on healing plants and substances. Although it is unlikely in the modern market economy that the sick and afflicted will ever be granted a significant voice in biodiversity research, ill people remain often unseen but tangible stockholders in biodiversity-based pharmaceutical research. The sick and afflicted, and their families and societies that support them, stand to lose if the world’s biodiversity is thoughtlessly destroyed before it can be evaluated for pharmaceutical potential.

See Also the Following Articles

- BIODIVERSITY AS A COMMODITY • BIOPROSPECTING • COEVOLUTION • ECOTOXICOLOGY • ETHNOBIOLOGY AND ETHNOECOLOGY • INDIGENOUS PEOPLES, BIODIVERSITY AND • PLANT SOURCES OF DRUGS AND CHEMICALS

Bibliography

PHENOTYPE, A HISTORICAL PERSPECTIVE ON

R. J. Berry
University College London

I. Background
II. The Making of a Phenotype
III. Formal Analysis
IV. Genetic and Phenotypic Variability
V. Phenotypes and Evolution
VI. Taxonomy

GLOSSARY

allele (originally allelomorph) A form of variant (sometimes called a mutation) of a gene.
diploid Individual who has two sets of chromosomes, usually received from different parents, in contrast to a haploid organism such as most microorganisms, the gametophyte stage of higher plants and some parthenogenetic forms (e.g., male ants and honeybees). It may have the same form or allele of a gene on both members of a chromosome pair (in which case it is homozygous) in different forms (alleles) (in which case it is heterozygous).
gene Historically, the inherited factor which determines a trait. Tends to be used somewhat loosely; more strictly represents a place or locus on the chromosomes which codes for a particular function.
genome The total genetic composition of an individual.
genotype The allelic composition of an individual at a particular locus.
mutation Change in an allele, producing a different allele; rate of occurrence affected both physically (especially by ionizing radiation) and by many chemicals. It may also refer to changes in a chromosome (involving duplication, deletion, or inversion of a segment).
phenetics The study of phenotypes, usually describing the grouping of organisms into taxa on the basis of estimates of similarity.
phenotype The appearance (function or behavior) of an organism.

THE PHENOTYPE OF AN ORGANISM is its actual form or appearance to an observer. Usually, the term is used to describe all the features that make up the organism, but it may be used to distinguish a particular characteristic, such as whether a mouse is black or brown. In this case, the adjective refers to a specific characteristic of a mouse, implying that there is variation between individuals or groups of mice. A phenotype commonly refers to the physical appearance of an organism (e.g., a large brown mice with a short tail), but it can also describe nonvisual properties of the organism, such as physiology or behavior.

I. BACKGROUND

Phenotypes are the appearance and properties of real animals and plants—living and dying, reproducing and bearing fruit, and succeeding or failing—but in themselves they are the products of interactions between the inherited (genetic) material and the environment, both
The distinction between inherited constitution and external appearance (i.e., between genotype and phenotype) was one of the most important demonstrations of Gregor Mendel (1822–1884) in his series of breeding experiments with peas which laid the foundation for genetics, and in which he recognized that a character may be inherited in a dominant or a recessive manner. An organism manifesting a dominantly inherited trait (in Mendel's case, round as opposed to wrinkled pods or colored as opposed to white flowers; more familiar examples are brown versus blue eyes in humans or brown coat color versus albino in rabbits or mice) may carry both alleles for the dominant trait or one for the dominant and one for the recessive trait. That is, it will have the same phenotype but could be genetically homozygous or heterozygous. A distinction between genes and their manifestation was also implicit in August Weismann's (1834–1914) embryological division between germplasm (which gives rise to reproductive cells) and soma (or body). However, the formal nomenclature phenotype and genotype was devised by the Danish botanist Wilhelm Johannsen (1857–1927), who introduced the word gene for the material basis of an inherited character, and thence the terms genotype and phenotype. Johannsen set out to investigate the relationship between (phenotypic) variation and selection. The discovery of Mendel's work in 1900 led to a rift between the biometricians (notably Karl Pearson (1857–1936) and W. F. R. Weldon (1860–1906), who followed Darwin (1809–1882) in viewing small continuous variation as the raw material of evolution, and the 'mendelists' (or geneticists) led by William Bateson, (1861–1926), who believed that large discontinuous saltations (or mutations) were the main cause of variation. The problem was the maintenance of continuous variation. Darwin had postulated a continuous replenishment of variation in order for selection to act, but he did not know its source. His proposal of "pangenesis" was an effort to solve the problem. Weismann's demonstration of the early separation in development of the reproductive tissue from the rest of the body supported the general assumption at the time that continuous variation was produced by environmental effects. Intraspecific variation (i.e., subspecies or local races) was therefore regarded as environmentally caused; the species was viewed as monotypic in a way that had much in common with the pre-Darwinian ideal or Linnean type.

Johannsen experimented with a self-fertilized cultivar of the bean *Phaseolus vulgaris*. The implication of this was that all the descendants of a single individual would have the same genes and constitute what Johannsen called a "pure line." Although individual beans might be different (due, for example, to their place in the pod), the mean and variance of all the characters of plants in a pure line were the same and were not affected by attempts at selection, i.e., plants grown from both small beans and large beans produced beans of the same average weight as that of all the plants in the pure line. In contrast, plants grown from crosses between pure lines had different (usually intermediate) characteristics from the parents, although these characteristics remained constant in pure lines derived from each cross. Johannsen argued that selection on continuous variation was inevitably ineffective, and the only variation on which selection could work depended on new mutation. This strengthened the contemporary assumption that evolution was bound up with mutations and their rate of occurrence, and that individuals were largely genetically uniform (i.e., homozygous at most loci for 'wild-type' alleles).

This assumption became built into conventional population genetic theory so that when recessive traits manifested in inbred populations (usually in the laboratory or garden, but occasionally appearing under wild conditions) they were assumed to be recent mutations in the process of elimination by natural selection. Most mutations seemed to be deleterious to their carriers (which would be expected if they were random changes in a functioning organism), which meant that high mutation rates would inevitably impose a burden on a population. In 1950, at a time of acute concern about the genetic effects of atomic warfare, H. J. Muller (1890–1967) proposed the concept of 'genetic load' and showed that a doubling of the mutation rate in a slow-reproducing species such as humankind could lead to extinction through genetically caused death.

The work of Bateson, Johannsen, and other early geneticists (one of the most influential was the Dutchman Hugo de Vries, 1848–1933) led to the isolation of genetics from evolutionary studies, particularly as represented by paleontology. This was resolved by the theoretical work of Sewall Wright (1889–1988), J. B. S. Haldane (1891–1964), and especially R. A. Fisher (1890–1962). Fisher published significant papers in 1918 and 1922 describing the expected biometrical properties of a Mendelian (i.e., breeding) population.
and the effects of an allele substitution on a quantitative (i.e., continually varying) character. (The papers were published by the Royal Society of Edinburgh; the first one was rejected by the Royal Society of London on the advice of Karl Pearson and the geneticist R. C. Punnett.) He went on to argue that dominance (and recessivity) are traits that have evolved and this explains why the majority of new mutations are recessive, detrimental, and have a major effect (i.e., on the phenotype); most of the data available at the time were from laboratory breeding of Drosophila melanogaster. Fisher reasoned that there was no intrinsic reason for a mutation occurring for the first time to be either dominant or recessive; the greatest probability is that it will be intermediate, with an effect somewhere between its expression in double dose (i.e., when homozygous) and the unmodified condition. However,

1. Mutations occur repeatedly at virtually every locus. The rare (approximately 1 in 10^5) mutational events we observe are recurrences of something that has happened thousands of times in the past.

2. When a mutation occurs, it will almost always be present in the heterozygous condition. If an allele has become relatively common in a population so that a fresh mutation to it has a reasonable chance of occurring in an existing heterozygote, then mutation cannot be the only force influencing its frequency.

3. If a newly arisen allele has a beneficial effect on its carrier, combinations of it with other alleles that increase its effect will have a higher fitness than any which decrease it. This will repeatedly occur so that the architecture of the species will become modified to the extent that any new occurrences of that advantageous allele will always produce the maximum effect in its carrier; this will almost always be in the heterozygous condition. In other words, beneficial characters will be selected for dominance and will also spread to replace selected stock, and by the second generation (when the selected modifiers would have a chance of segregating independently) the original variable heterozygotes virtually indistinguishable from the homozygote in selection for the fewest individuals and ones most like the typical homozygote in the white selection line. In other words, he had changed the heterozygote from a position of no (or intermediate) dominance to complete dominance or recessivity respectively. Ford then crossed his modified heterozygotes with unselected stock, and by the second generation (when the selected modifiers would have a chance of segregating independently) the original variable heterozygotes reappeared: He thus showed that it was the response of the organism rather than the gene that had changed.

Laboratory experiments of this nature have been carried out on a variety of organisms and a range of characters. An experiment particularly informative with regard to genetical architecture in the wild was carried out by H. B. D. Kettlewell (1907–1979) using British and Canadian peppered moths (Biston betularia and Amphimallon solandri, respectively; these are fully interbreeding). Although there is a melanic form (swettaria) in the North American species, it is comparatively restricted in its distribution, and only pale (or typical) moths occur over vast tracts of Canada.
When a melanistic heterozygote and a typical homozygote of British origin are mated, the offspring are clearly dark or light. The melanistic character is a straightforward dominant. Even when the typical moths come from Cornwall in extreme southwest England where melanic have never been reported, only a slight loss of complete dominance occurs and that only after several generations of crossing melanics back to Cornish stock (i.e., back-crossing a melanistic from several generations of melanistic × Cornish cross with a “pure” Cornish parent). This modification consists of some white dots on the normally jet-black wings of the heterozygote. Perhaps significantly, this slight heterozygous expression of the gene gives specimens similar to those caught in the early days of the spread of peppered moth melanism in the mid-nineteenth century and which are now prized collectors’ specimens. This white speckling has long since disappeared in wild-caught British specimens, and modern heterozygotes are indistinguishable from melanistic homozygotes. Melanism has become fully dominant during the ensuing decades.

However, there has been no opportunity for dominance to evolve in Canadian peppered moths. When British melanics are crossed to Canadian stock (from areas where swettaria does not occur) the first-generation progeny segregate as dark or light in the same way as in a cross between British moths. In the first generation, the dominance modifiers in the British parent will be carried on the chromosomes in the same order as in British stock and produce dominance in the same way. In the next generation, the gametes contain chromosomes which have crossed over between the British and Canadian grandparents. Consequently, in the second generation, the “switch” between pale and black forms does not operate as efficiently. Kentewell crossed heterozygotes from a British × Canadian matings with Canadian moths and repeated this back-crossing for four consecutive generations, after which the heterozygotes ranged from black to pale—there was no sign of dominance. He then reversed the procedure and mated his “broken-down” melanics to British typicals. The dominance of the condition was immediately re-established. The architecture of the British chromosomes shaped a clear segregation between a dark heterozygote and a pale homozygote. Fisher’s theory has proved correct in many similar experiments in both plants and animals, and although it may not be universally operative it provided the genetic basis for the understanding between disciplines that was needed before the neo-Darwinian synthesis could occur.

As late as 1932, T. H. Morgan was asserting that “natural selection does not play the role of a creative principle in evolution,” but 10 years later all but a very few biologists were agreed on an evolutionary theory based firmly on Darwin’s ideas knitted with subsequent developments in genetics. This coming together was described by Julian Huxley as the “modern synthesis” in a book with the same name published in 1942. The synthesis can first be seen in three English books: R. A. Fisher’s *Genetical Theory of Natural Selection* (1930), E. B. Ford’s *Mendelism and Evolution* (1931), and J. B. S. Haldane’s *Causes of Evolution* (1932). It was consolidated in three works from America. Theodosius Dobzhansky’s *Genetics and the Origin of Species* (1937), Ernst Mayr’s *Systematics and the Origin of Species* (1942), and George Gaylord Simpson’s * Tempo and Mode in Evolution* (1944). As Mayr noted, it did not occur as a result of one side being proved right and the others wrong but rather from “an exchange of the most viable components of the previously competing research traditions.”

**II. THE MAKING OF A PHENOTYPE**

In the simplest microbial systems, there is a one to one relationship between gene action and phenotype: Change in a gene is likely to produce a change in its product, which will manifest directly in the organism as an altered character, perhaps the loss or modification of an enzyme. Such simple relationships exist in all organisms, however complex. There are more than 200 “inborn errors of metabolism” in humans, each resulting from changes in a particular gene, producing a change in phenotype in the whole organism. For example, albinism is due to the inability to synthesise melanin, which is made from tyrosine under the influence of tyrosinase, which is under the control of a gene on chromosome 11 in humans; “classical” hemophilia results from the absence of a protein (factor VIII) coded by a gene on the X chromosome. However, most traits are affected by many genes. For example, blood clotting in mammals is effected by a cascade of physiological reactions, each under the control of a different gene(s). Efficient clotting requires all the different stages to be operating, and defects in any stage (especially the genes directing
the relevant proteins and enzymes) will lead to a ‘bleeding’ phenotype. In blood clotting it is possible to detect where the error lies because we know the normal determinants of clotting in detail. For most traits we do not have information about the steps in their formation. One of the benefits of genome mapping is that the genetic determinants of complex characters will be analyzable in an orderly way. Currently, most phenotypic analysis is biometric rather than genetic.

For example, the tails of house mice may be shortened by mutations at approximately 40 loci. We know the action of many of these genes: Some affect the notochord, some cell division rates, and some inductive relations between endoderm and mesoderm. We also know that the tail length of mice varies between different inbred strains of mice (which are similar in the genetic sense to Johannsen’s pure lines) and that we can increase or decrease tail length by selecting wild-caught animals or the products of crosses between inbred strains. We do not know which genes are variable (or segregating) in any one population or which genes are being affected by selection. However, it is clear from many selection experiments that the genes affecting any complex trait are distributed throughout the genome (i.e., over many chromosomes) and that selection for any one trait may ‘unbalance’ the genome.

The concept of genomic balance (often called genomic or genetic architecture) is important. When a normal outbreeding species, such as maize, sugar beet, poultry, or Drosophila, is made to inbreed by manipulating its breeding system, the individuals characteristically decline in vigor and fertility until they stabilize at a stage before complete homozygosis is approached. The amount of this inbreeding depression varies from one line to another within any species. When two inbred lines are intercrossed, the F1 (i.e., the first-generation progeny) show a considerable increase in vigor and fertility, known as hybrid vigor or heterosis. In general, the F1 phenotype is similar to that in the population from which the inbred strains were derived. If F1’s are raised and inbreeding is resumed, inbreeding depression will again occur, although not necessarily to the same extent as in the original lines. Inbreeding depression is often reflected in increased variability not only between individuals but also among repetitive parts such as bilateral characters in animals and floral morphology in plants; fluctuating asymmetry, in which there are differences between the right and left sides of individuals, is commonly used as a rough indicator of inbreeding. F1’s tend to show decreased variability. Michael Lerner argued that this indicates an innate superiority of heterozygotes over homozygotes, perhaps because of a greater biochemical flexibility in the former. It is a phenomenon which has produced considerable research interest, particularly among marine biologists. Lerner argued that if artificial selection is suspended before much of the variation has been lost (as homozygosity increases) natural selection will tend to restore the character (and hence its genetic determinants) to an equilibrium value, with the mean of the character which was artificially selected tending to revert toward its original value; he called this genetic homeostasis.

Artificial selection tends to accumulate alleles which act on a character in a particular way, e.g., to increase its size of or the number of elements making up a repeated trait. If we make the reasonable assumption that any population needs to combine phenotypic uniformity in a stable environment with long-term flexibility should the environment change, the easiest way to do this is for the character in question to be controlled by alleles at different loci, some alleles acting to increase the expression of the character and others acting to decrease it. In such a case, Fisher showed that selection will favor linkage between the loci responsible, with the evolution of ‘balanced’ chromosomes containing ‘positive’ and ‘negative’ alleles. The simplest situation will involve two segregating loci, A,a and B,b with A,B acting in one direction and a,b in the other (i.e., additive genes). The intermediate type can be either the attraction or the repulsion heterozygote, AB/ab or Ab/aB. However, the latter will be favored since it will be less likely to produce zygotes giving the extreme phenotypes (Fig. 1). For the same reason, any mechanisms bringing about closer linkage between the loci concerned (such as a chromosomal inversion) will be favored.

This theoretical arrangement has been subjected to experimental analysis, mainly through selection experiments in Drosophila, and has been generally confirmed; it receives support from the distribution of ‘quantitative trait loci’ identified by molecular techniques. It also accounts for other properties:

1. Selection (natural or artificial) for any character almost invariably produces correlated responses in additional developmentally independent traits of the phenotype.
2. Gene loci affecting viability traits are interposed along the chromosomes with loci affecting other characters.
3. The highest rate of artificially induced new variation by mutation is many times less than that occurring spontaneously through recombination.
4. There is widespread occurrence of chromosomal inversions in nature.

5. Deleterious gene combinations (even to the point of lethality) may occur solely as a result of recombination.

There are a few cases in which the different genes contributing to complex variation have been identified. This has been done for the determinants of the pin thrum polymorphism in Primula vulgaris, in which different members of a linked group of loci control anther height, style length, pollen size, rate of pollen tube growth, and length of the papillae on the stigma; for mimetic patterns in the African swallowtail butterfly Papilio dardanus; and for color and banding patterns in the land snail Cepaea nemoralis. In house mice, approximately 16% of genes with an identified function are concerned with behavior and have been located on 19 of the 20 chromosomes of the species. However, in most cases it is only possible to conclude that different components of a character complex are inherited as a unit, i.e., that the gene complex is coadapted.

Coadaptation is presumably the reason why the genomes of interfertile species do not always merge when they meet. For example, races of dark- and light-bellied house mice (referred to as Mus musculus domesticus and Mus musculus, respectively, although they should perhaps be given full species status) meet in a narrow zone of intergradation across Jutland (Denmark) and south through Germany. Although the two forms readily interbreed in the laboratory, the hybrid zone has apparently been constant for at least 50 years. Gene frequencies on the two sides of the hybrid zone are very different. Remarkably, the frequencies in the light-bellied form in California, which are descended from the same stock as the light-bellied Danish animals, are more like those in the light-bellied Danish population than are the light-bellied Danish ones from their dark-bellied neighbors. A similar situation exists in deer mice (Peromyscus polionotus) in Florida and also in carrion and hooded crows in the same area in which the dark- and light-bellied mice occur in Europe. However, some species lose their identity wholly when brought into breeding contact. This has happened commonly in New Zealand, where much of the terrestrial biota has been
introduced. Hybrid zones are common. Presumably these occur when coadaptation has not evolved.

III. FORMAL ANALYSIS

In the past, it was suggested that the genes which control major or qualitative traits are different from those which affect quantitative ones (oligogenes and polygenes, respectively). This distinction is now rarely made. In other words, it is assumed that genes affecting quantitative traits (such as weight or size or physiological properties such as metabolic rate) follow Mendelian patterns of inheritance, may have multiple alleles, can mutate, change in gene frequency, show dominance, etc. Quantitative inheritance is merely a general case of the interaction of genes in which the interacting components are little or wholly known.

The number of genes which affect a trait can be estimated by the amount and speed of response in a selection experiment or by the mean and variance of the character as measured in a population. Using these techniques it has been calculated that human skin color may be determined by only 5 or 6 loci, whereas the number of genes affecting oil and protein production in maize may be as high as 54 and 122, respectively. However, such estimates are very dependent on the nature of interactions between the loci concerned and should be regarded as no more than suggestive of a large or small number.

The simplest assumption in multigenic trait determination is that all the loci affect the trait equally and therefore additively. However, detailed analysis has shown that in many (perhaps most) cases, a few genes have a major effect and many genes have a minor effect. For example, in the well-studied case of variation in sternopleural bristle number in D. melanogaster, approximately 10 loci account for 75% of the genetic variation in number. The situation is further complicated by pleiotropy: A gene may have a major effect on one character but minor effects on others. For example, phenylketonuria is an inborn error of metabolism producing severe mental retardation in humans and is produced by the nonfunctioning of phenylalanine hydroxylase, which is controlled as a recessive trait by a single gene on chromosome 12. The same enzyme is involved in melanin synthesis, and phenylketonurics have slightly paler hair and complexion than their normal sibs. The gene can therefore be regarded as having a major effect on intelligence but a minor effect on pigmentation.

However, a quantitatively inherited trait will be more likely than a qualitative one to be affected by environment. A group of individuals having identical genes for growth (i.e., a pure line or a clone) may show considerable variation in size due to differences in available nutrients. Although the same potential for size is present in the initial gene products, the manifestation of the phenotype will be limited by such factors as food availability. Conversely, a population of genetically heterogeneous individuals may grow to the same size if no gene–environment interaction is limiting (or if different interactions compensate for each other). Generalizing, we can express the phenotypic value $P$ for individual $i$ in environment $j$ as

$$P_i = G_i + E_j$$

where $G_i$ is the genetic contribution of the $i$th genotype and $E_j$ is the environmental deviation resulting from the $j$th environment.

A particular genotype may do well in a particular environment, implying a specific interaction between the two. In this case,

$$P_i = G_i + E_j + GE$$

In practice, there will be variance of these components so that

$$V_p = V_G + V_E + 2Cov_{GE}$$

where $V_p$, $V_G$, $V_E$, and $2Cov_{GE}$ are the phenotypic, genetic, and environmental variance and the genotype–environment covariance, respectively. The genotype–environment covariance is positive when genotypes with higher values are in better environments and poorer genotypes have poorer environments. This may occur in animals when one member of a litter is large because of its genes and gets more food from its parents or when an animal is socially dominant for genetic reasons and therefore has more resources in food and spaces. A plant genotype which grows faster may have a better environment because it is less likely to be shaded.

In controlled plant and animal breeding, efforts are made to randomize genotypes and environments, so that $Cov_{GE}$ can be neglected, and

$$V_p = V_G + V_E$$
Conventionally, 

$$V_g/V_P + V_e/V_P = h^2 + e^2$$

where $h^2$ and $e^2$ are the proportion of phenotypic variation due to genetic and environmental factors, respectively. The term $h^2$ is known as heritability in the broad sense,

$$h^2 = V_g/V_P$$

In practice the genetic variance is composed of a range of different interactions between loci, which may be additive, dominant, or epistatic. Variability in the narrow sense is defined as

$$h^2 = V_g/V_P$$

where $V_i$ is the variance due to additive genetic factors. It is an important statistic in determining the rate and amount of response to directional selection in breeding programs.

Herkatibility is a population-specific measurement. It does not measure an invariant property of a particular trait but only the relative contributions of genetic and environmental differences to phenotypic variation in a specific situation. If either genetic or environmental variation changes, heritability estimates will also change; heritability measures the proportion of phenotypic variation in a particular population due to genetic variation.

### IV. GENETIC AND PHENOTYPIC VARIABILITY

The assumption of the early geneticists that most populations carry little variation and that most individuals are homozygous at all but a few loci at which alleles are either recent mutants in the process of being eliminated by selection or maintained by opposing selection pressures or heterosis is clearly wrong. Evolutionary theory (requiring modifying genes to change phenotype expression) and artificial selection practice (revealing considerable potential for inherited response to selection) imply the existence of considerable genetic variation in and between populations. In the 1950s and 1960s assumptions about genetic load suggested there was a maximum amount of variation which could be tolerated in any population, but the application of protein and later DNA electrophoresis showed that load theory was too narrow and deterministic and led to reinterpretation so as to incorporate ecological factors (including heterogeneity in time and space and variable stress from biotic and abiotic agents). Empirical data have shown that even inbred organisms (such as obligate self-fertilizers) or ones living in an apparently constant environment (such as the deep sea) may still be heterozygous at a significant proportion of their loci (up to one-fourth for enzyme loci in plants and invertebrates and less for vertebrates).

The implications of load theory led to debate among population geneticists and evolutionists about the significance of the observed variation. Theoreticians impressed by the apparent rigor of load theory and biochemists not used to meaningful variation in their study of chemical pathways tended to dismiss the bulk of genetic variation as neutral and irrelevant to the organism, whereas evolutionary ecologists and practical breeders regarded it as potentially adaptive and as expected by Darwinian understanding. The latter emphasis has been shown to be correct: Far from being constrained by invariant genotypes, phenotypes can (and should) be treated as capable of rapid response to the environment and hence permissive rather than deterministic of survival. A phenotype can be interpreted as the consequence of developmental reaction norms, facilitated in most cases by the width of its underlying genetic base.

This concept is supported by a series of experiments carried out by Bruce Wallace to test if radiation-induced mutations in *D. melanogaster* were inevitably detrimental. He found that mutations induced in flies made homozygous by artificial breeding were often heterotic, although mutations occurring in flies with a heterozygous background were usually deleterious to their carriers, presumably since their gene expression had been adjusted by normal selection. This led Wallace to emphasize the importance of balancing (or stabilizing) selection rather than the directed or cleansing selection which removes unwanted variants.

The concept of reaction norm incorporates the idea of phenotypic plasticity championed by Bradshaw and Levins in the 1960s and provides a much-needed synthesis of allometry and ontogeny with ecological realism and fitness components. A particularly good example is the seasonal polyphenism of the African satyrine butterfly, *Bicyclus anynana*, studied by Paul Brakefield and colleagues. The wet and dry season forms of this species are phenotypically distinct in the size of the eyespots and the banding patterns on the wings. Seasonal shifts in rainfall are associated with changes in temperature and with butterfly behavior. In the cooler dry season,
the insects rest on dry grass or leaf litter and rarely fly; they do not breed at this time. In contrast, the adults fly actively in the warmer wet season, searching for mates and oviposition sites in the lush vegetation. Survival differences between the two forms suggest that cryptic matching is more important in the dry season and deception mediated by the false eyes in the wet season.

The wet season form arises through an acceleration of development which can result from an increased or fluctuating temperature, food quality, or hormonal influence. However, eyespot size and plasticity are affected by genetic variation as shown by variation between families in different temperature regimes and by directed selection. The switch between different wing spotting phenotypes in the butterfly Maniola jurtina studied over many years by E. B. Ford is probably also of this nature. Ford’s extensive field studies showed the interaction between founder population differentiation, natural selection in particular localities, and environmental determinants, although he did not interpret them in this way. Unlike the Bicyclus case, the changes in phenotype in Maniola are small and unlikely to be important in survival or fitness (although it would be wrong to be dogmatic about this conclusion). However, they serve as a marker of a variable developmental system which may well have a range of functional responses. There are many such examples of trivial (or even wholly cryptic) differences between individuals which have different functional properties (e.g., the possession of B chromosomes or resistance to a newly introduced pesticide or pathogen).

The complicated response system in Bicyclus (and other butterflies) is similar in principle to the much simpler situation of eye pigmentation in the aphid Gammarus chevreuxi studied classically by Julian Huxley and Ford. Here, the unpigmented, red-eye form is produced by a single mutant allele which has a slower rate of melanin synthesis than normal. However, pigmented eyes like the wild type arise if genetically red-eyed animals are raised at higher temperatures than normal or if they also carry a gene for small eyes which enables the small amount of pigment produced to cover the whole eye. Another example is the “himalayan” mutation of mammals frequently described in elementary genetics textbooks in which an unpigmented (white) coat becomes pigmented at the cold extremities (feet and ears), as in Siamese cats. Pigmentation may also develop in hair which grows in a shaved area which has a lower temperature than usual. The tyrosinase in this genotype is heat labile and has a maximum activity below normal body temperature. Rate genes of this nature have been very important in evolution and can be shown to produce major alterations in body form through various forms of allometry.

V. PHENOTYPES AND EVOLUTION

It is a truism that natural selection can only operate where phenotypic variation exists. Evolution results from changes in gene (strictly allele) frequencies, but these are the consequence rather than the cause of phenotypic differences. It follows that selection on local forms (or ecotypes) which are solely the environmental consequence of a particular environment (e.g., plants grown in sheltered or exposed conditions or subject to a particular dietary lack or a behavioral reaction in an animal) will not lead to adaptive adjustment unless the phenotype in question depends on different genotypes. However, a phenotypic response to environmental conditions may allow a genetically nonadapted population to survive long enough to accumulate variants (through mutation, recombination, or immigration) and then adapt genetically. This idea was put forward independently by Baldwin (1896), Osborn (1897), and Lloyd Morgan (1900): It is commonly known as the Principle of Organic Selection or the Baldwin effect. This has been claimed by some to bridge the directed response which is the basis of Lamarckism and the random source of inherited variation which underlies Darwinism. Certainly, it may mimic Lamarckism, although it is of course wholly Darwinian in its operation.

Most of the early studies of phenotypic plasticity were carried out by botanists: Gaston Bonnier in the Alps and Pyrenees, F. E. Clements in Colorado and California, Turesson in Sweden, and Clausen, Keck, and Hiesey in California. However, the evolutionary implications were not pursued until Gause and Schmalhausen in Russia and Waddington in Britain began to examine the relationships between genotype, phenotype, and environment in animal experiments. They were able to show not only that natural selection could lead to a character originally induced by the environment becoming an inherited character (Waddington called this “genetic assimilation”) but also that there is a causal connection between the environmentally induced change and subsequent genetic changes. They argued that since adaptability—the ability to acquire an adaptive variant during an animal’s lifetime—has a genetic basis, the genes underlying flexible adaptive variations may ultimately be responsible for the evolution of fixed adaptations to a new environment, i.e., the environment is much more than a sieve selecting
(or eliminating) chance mutations. These ideas have been criticized (notably by G. C. Williams) on two grounds: That most changes resulting from environmental challenges are not adaptive and that fixation of a genetic response results in a decrease in genetic potential because less (genetic) information is needed to specify a fixed, than a variable response. There is truth in both these objections, although the assumption that plasticity requires more inherited variation than a fixed state is debatable. However, there is no doubt that the link between stress-generated responses and subsequent adaptation requires more study. The problem historically is that ecologists have been primarily concerned with pattern (i.e., spatial relationships within communities), whereas evolutionists have concentrated on processes (i.e., temporal changes within communities), and both have used restrictive definitions of stress. This division between disciplines is not, of course, absolute, but it has led to a rift within population biology and a lack of understanding of coevolutionary possibilities and constraints.

The divergence between ecologists and evolutionists is seen in the models developed by each. In general, ecological models are self-contained because the dependent variables (numbers or density) and parameters (birth and death rates, dispersion differences, rates of predation, etc.) are all measurable by ecological methods. Ecologists have been particularly concerned with identifying criteria for stability or fluctuation when species interact, for invadability, and for extinction when competition occurs. Such models are limited because the parameters are all manifestations of the phenotypic properties of individuals and hence the product of evolutionary processes. Although lip service is paid to the fact that phenotypes may change, in practice they are assumed to be effectively constant in ecological time. Evolutionary models originally concentrated on examining possible rates of evolution; then they became concerned more with the factors involved in the maintenance of genetic polymorphism. Their variables are the frequencies of alleles and genotypes; their parameters are relative fitnesses and rates of mutation, migration, and recombination. For a long time, fitnesses were regarded by the model builders as virtually constant, and it is only comparatively recently that the dependence of selection coefficients on density and frequency has been incorporated. This false assumption of constancy led to the problems associated with genetic load and to controversies over neutralism.

Genetical (evolutionary) models have become increasingly realistic and treat relative fitnesses as capable of varying in time and space as well as with gene frequency and population density. Such models are both genetical and ecological, and therefore they are intrinsically more informative than purely ecological ones. However, they still suffer from a major problem in that they tend to assume a degree of genetical equilibrium or stasis which is unjustified.

The way forward will be to increase ecological reality of evolutionary models, taking into account the characteristics of the niche for particular populations and communities, particularly any genetic constraints. Such models obviously need to include behavioral (sociobiological) input and the notion of evolutionary “strategies.”

VI. TAXONOMY

Classification has become increasingly sophisticated with the increase in traits which can be used to characterize a group (or taxon). The problem is that conventional taxonomic (phenotypic) diversity may bear little relationship to genetic diversity. For example, the seaside sparrow (Ammodramus melba) is common on the eastern and southern coasts of the United States. Nine subspecies have been described, including a rare and recently extinct (1987) dusky form, originally regarded as a separate species. Molecular analysis (based on mitochondrial DNA sequences) showed that the dusky sparrow was indistinguishable from other Atlantic forms, but that there was a major and previously unsuspected distinction between Atlantic and Gulf Coast groups.

In contrast, New Zealand tuatara lizards (Sphenodon punctatus) are conventionally treated as a single taxon, but molecular (and morphological) criteria indicate they comprise at least three species. For conservation management, it is obviously important that each taxon be considered separately.

There is no clear correlation between genetic and taxonomic diversity. The problems of classification that have always challenged museum workers have been compounded by the possibility of using genetic factors as additional or substitute taxonomic criteria. Closely related species in the genetic sense may have very different niches than those of their near relatives, whereas genetically more distant forms may look alike and interact strongly. These are no clear rules to link phenotype with genotype or phenotypic variety with genetic variability. There is plenty of scope for better multidisciplinary understanding of phenotypes, this will have to involve genetics, development, behavior, biotic and abiotic environments, life history, and phylogeny. Notwithstanding-

...
ing case by case examinations of the determinants and plasticity of the phenotypes in particular species and species groups are still needed.

See Also the Following Articles

EVOLUTION, THEORY OF • GENES, DESCRIPTION OF • GENETIC DIVERSITY

Bibliography


I. Light Harvesting and Excitation Energy Transfer to Reaction Centers

II. Reaction Centers and Primary Photochemistry

III. Membrane-Associated Reactions Leading from Primary Photochemistry to ATP and NAD(P)H

IV. Use of NAD(P)H and ATP in CO2 Fixation and Other Reactions

V. Conclusions

GLOSSARY

ATP synthetase Membrane-associated protein complex that couples exergonic fluxes of H+ across the membrane to ADP phosphorylation, or vice versa.

carboxylase Enzyme that catalyzes the formation of a C=O bond between inorganic C (CO2 or HCO3-) and some organic molecule.

chemiosmotic energy coupling Exergonic (photo)-chemical redox reaction, or an exergonic hydrolysis of a phosphate anhydride or thioester, coupled to endergonic ion (normally H+) transfer across a membrane, or vice versa.

chromophore Organic molecule that absorbs photosynthetically active radiation and either carries out photochemistry or transfers excitation energy to another chromophore that carries out photochemistry. Always bound to a polypeptide in the photosynthetic apparatus.

reaction center (Bacterio)chlorophyll–protein complex that performs the primary photochemical reactions of photosynthesis, the earliest products being the oxidized reaction center (bacterio)chlorophyll at the outside/thylakoid lumen side and a reduced (bacterio)chlorophyll (RC1) or reduced (bacterio)phaeophytin (RC2) on the cytosol/stroma side of the photosynthetic membrane.
those in which primary photochemistry occurs. In O₂-evolving organisms, the basic photosynthetic equation is

\[
\text{C}_6\text{H}_12\text{O}_6 + 6\text{H}_2\text{O} + 6\text{CO}_2 \rightarrow 6\text{C}_6\text{H}_12\text{O}_6 + 6\text{O}_2 + 6\text{H}_2\text{O}
\]

1. LIGHT HARVESTING AND EXCITATION ENERGY TRANSFER TO REACTION CENTERS

For photochemical reactions to occur, photons must be absorbed by a pigment. In the majority of photosynthetic organisms, only one in 100 to one in 1000 pigment molecules can actually bring about useful (to the organisms) photochemical reactions. The photochemically inactive antenna pigments serve to absorb radiation and transfer the energy as excitation energy (i.e., not in stable chemical forms) to photochemically active pigments in reaction centers. This division of labor presumably relates to the generally rather “dilute” nature of solar radiation, such that each pigment molecule absorbs a photon on the order of a few times each second, yet the photochemical machinery and the immediate downstream reactions can react hundreds or thousands of times per second. Accordingly, the resource-expensive (energy, nitrogen, iron) reaction centers and the related catalytic machinery that is energetically and temporally downstream of the reaction centers are present as a small number of copies of the catalysts relative to the number of molecules of light-harvesting pigment; this arrangement is presumably related to optimal allocation of resources in natural selection.

A very small number of species, all of them members of the Archaea, do not accord with this generalization about antenna pigments and reaction centers. These Halobacteria use carotenoids in membrane-spanning proteins to catalyze their primary photochemical reaction, with each pigment molecule acting as a light-energized H⁺(bacteriorhodopsin) or Cl⁻(halorhodopsin) pump (see Section II).

By contrast, all other photosynthetic organisms are (e)Bacteria and their symbiotic eukaryotic descendants that use the antenna plus reaction center system. In all of these organisms the pigment involved in the primary photochemical reactions is a Mg- (very rarely Zn-) binding cyclic tetrapyrol, chlorophyll (in O₂-evolvers) and bacteriochlorophyll (in non-O₂-evolvers). Chlorophylls always occur in a functional photosynthetic apparatus in association with proteins in, or more rarely on, bilayer membranes. In all chlorophyll-containing organisms there are some chlorophyll molecules that act as antenna pigments. In many such organisms (e.g., all green algae and higher plants) the great majority of light harvesting is carried out by chlorophyll-protein complexes, in this case with chlorophyll a and b as the major light-absorbers with minor contributions from carotenoids. In the case of algae with chlorophyll a and, usually, chlorophyll c, carotenoids generally have a more significant role in light harvesting by pigment–protein complexes. In a final group of organisms (mainly cyanobacteria and red algae), the light-harvesting pigments in addition to chlorophylls are open-chain terpyrrol pigments covalently bound to proteins, attached to the outer (stromal in red algae, cytosolic in cyanobacteria) side of the photosynthetic membranes, or in the thylakoid lumen of cryptophyte algae. The variations in the chemistry and location of light-harvesting pigment–protein complexes are indicated in Table I.

A number of energetic and structural constraints determine the spatial relationship of the different pigments in light-harvesting complexes. The energetic constraint is that energetically efficient excitation energy transfer among antenna pigments, and from antenna pigments to reaction centers, involves an unavoidable energy loss during the transfer. Thus, the peripheral pigments in the antenna have a higher energy content of their excited states than do the inner antenna pigments, and those in turn have a more energetic excited state than the reaction center pigment. The situation is complicated by the occurrence of two readily accessible excited states in some photosynthetic pigments, and especially in the chlorophylls. Here the second excited state, produced by absorption of blue radiation, has a higher energy content than the first excited state, corresponding to the in vivo absorption in the red (chlorophylls) or infrared (bacteriochlorophylls) region of the electromagnetic spectrum. It is the first excited state that performs the primary photochemistry in (bacterio) chlorophyll-based photosynthesis. Thus, photons that excite (bacterio) chlorophylls to the second excited state are not directly used in photochemistry via this state, but only after they become converted to the first excited state; this also applies to excitation energy transfer among pigments.

These energetic requirements are reflected in the structure of the light-harvesting apparatus. Thus, the pigments are arranged spatially such that those absorbing shorter-wavelength (thus higher energy) photons are generally at a greater distance from the reaction center, whereas those absorbing at longer wavelengths...
PHOTOSYNTHESIS, MECHANISMS OF

TABLE I
Chemistry, Taxonomy, and Location of Light-Harvesting Pigment–Protein Complexes

<table>
<thead>
<tr>
<th>Chromophores</th>
<th>Taxonomy</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>O2-Evolvers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlorophylls a and b</td>
<td>Chlorophylla, Euglenophyta, Chlorarachniophyta, Chlorophyta, Eubacteriophyta, <em>&quot;chloroviruses&quot;</em>, cyanobacteria</td>
<td>Integral membrane protein of thylakoid</td>
</tr>
<tr>
<td>Chlorophylls a and (usually) c</td>
<td>Cryptophyta, Heterokontophyta (Bacillariophyceae, Chrysophyceae, Phaeophyceae, Trachelophyceae), Haptophyta, Dinophyta</td>
<td>Integral membrane protein of thylakoid</td>
</tr>
<tr>
<td>Chlorophyll a, peridinin</td>
<td>Dinophyta</td>
<td>Peripheral protein on stromal surface of thylakoid</td>
</tr>
<tr>
<td>Phycobilinogen</td>
<td>Cyanobacteria sensu stricto, Rhodophyta, Cryptophyta</td>
<td>Phycobilisomes on stromal/cytosol surface of thylakoid</td>
</tr>
<tr>
<td><strong>Non-O2-Evolvers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteriochlorophylls a and b</td>
<td>Proteobacteria: purple sulfur bacteria (Chromatiaceae), purple nonsulfur bacteria (Rhodospirillaceae)</td>
<td>Integral membrane protein of &quot;chromatophores&quot;</td>
</tr>
<tr>
<td>Bacteriochlorophyll g</td>
<td>Helioспорiaceae</td>
<td>Chlorosomes on the stromal side of plasmalemma</td>
</tr>
<tr>
<td>Bacteriochlorophyll e, d, and c</td>
<td>Chloroflexaceae, Chlorobiaceae</td>
<td>Integral membrane protein of plasmalemma</td>
</tr>
</tbody>
</table>

(more similar to that of the reaction center pigments) are close to the reaction center. Furthermore, the smaller-scale arrangement of individual chromophores in the pigment–protein complexes is also a function of the distance between chromophores over which effective excitation energy transfer can occur; these arrangements have been shown for a number of light-harvesting pigment–protein complexes using X-ray crystallographic techniques.

A related consideration is that of how many chromophore molecules can be involved in light harvesting for a given reaction center. A minimum estimate of the upper limit of this number comes from determining the ratio of pigment molecules serving a given kind of reaction center and dividing it by the number of the type of reaction center that they serve. This is clearly most readily done when there is only one kind of reaction center in the organisms concerned, as is the case in organisms that do not evolve O₂, but it is also applicable to O₂-evolvers, which invariably have two kinds of reaction center (see Section II). Such computations suggest that up to 1000–2000 chromophores occur per reaction center of a given type. Even more pigment molecules can be involved in excitation energy transfer to a given reaction center if pigment molecules are not all dedicated to a given reaction center but can be involved in excitation energy transfer to several nearby centers.

Another size-related aspect of light harvesting occurs at all scales, but becomes most obvious at a larger scale: this concerns the optical thickness of the cell or tissue in which the photosynthetic apparatus occurs. A small optical thickness involves a smaller physical thickness for a given pigment concentration per unit volume, or a lower concentration of pigment per unit volume if there is a fixed physical thickness, or a combination of the two. A large optical thickness means more self-shading of pigment molecules, so that the average specific absorption coefficient of each pigment molecule is lower. Furthermore, in a given radiation environment, it takes longer for a pigment molecule to absorb enough photons to cover the energy cost of its synthesis in a cell or organ with more self-shading. Finally, the cells or tissues with a small optical thickness, and hence smaller package effect, have a greater opportunity for the spectral diversity of pigments that they contain to
PHOTOSYNTHESIS, MECHANISMS OF

be manifest in the specific wavelengths of radiation that are absorbed. The ultimate in large optical thickness is found in the thalli of some seaweeds that, with more than one millimole of pigment per square meter, look black to the human observer (because of low reflectance or transmittance, with minimal wavelength dependence of these two processes due to high absorptance at all visible wavelengths) regardless of whether their pigments are of the kind found in green, brown, or red algae.

A final aspect of light harvesting concerns the occurrence of pigment molecules in photosynthetic cells that (of necessity) absorb radiation but that show little or no transfer of excitation energy to reaction centers. Examples of such pigments are β-carotene, and at least some portion of the other carotenoids. The pigments serve as photoprotectants, quenching the triplet excitation states of chlorophyll and singlet oxygen in the case of β-carotene, and as nonphotochemical quenchers of excess (singlet) excitation energy of antenna pigments in the case of phototransformable carotenoids such as violoxanthin–antheroxanthin–zeaxanthin and diadinoxanthin–diatoxanthin in certain O₂-evolvers.

II. REACTION CENTERS AND PRIMARY PHOTOCHEMISTRY

We have seen in Section I that the halobacteria, with their halorhodopsin or bacteriorhodopsin pigments, do not have separate antenna and reaction center pigments. These organisms have active transport of ions (H⁻ out of the cell or Cl⁻ into it) as the sole product of photochemistry external to the pigment–protein complex. The other photosynthetic organisms, with (bacterio)chlorophyll as their photochemically active pigment, do have reaction centers; these organisms are predominant in terms of number of species, habitats occupied, and quantity of energy and materials transformed.

In all of the (bacterio)chlorophyll-based reaction centers the primary photochemical event is the production of a reduced compound on the side of the membrane abutting the cytosol (prokaryotes) or chloroplast stroma (eukaryotes) and an oxidized compound on the side of the membrane adjacent to the external medium (many prokaryotes) or thylakoid lumen (almost all O₂-evolvers). The photochemical redox process thus involves the transfer of an electron from near one side of the membrane to near the other side of the membrane. Furthermore, the photochemically active pigment is bacteriochlorophyll a, b, or g, or chlorophyll a (or apparently, in one case, chlorophyll d), bound to protein (Table II).

While all (bacterio)chlorophyll-based reaction centers share this basic mechanism, such centers can be divided into two types termed RC1 and RC2 (Table II). In the RC1 type (e.g., the reaction centers of the Chlorobiaceae and Heliothriaceae, and PSII of O₂-evolving organisms), the (bacterio)chlorophyll catalyst of primary photochemistry that is oxidized generates an oxidant [oxidized (bacterio)chlorophyll] at a redox potential of +0.25–+0.45 V, while the primary reductant is another form of (bacterio)chlorophyll with a redox potential of −0.6–−0.7 V. In the RC2 type (e.g., the reaction centers of the Chloroflexaceae and Proteobacteria, and PSI of O₂-evolvers), the oxidized (bacterio)chlorophyll product of primary photochemistry has a redox potential of +0.45 (bacteria that cannot evolve O₂) to +0.1 V (O₂-evolving organisms), while the reductant [a (bacterio)pheophytin, i.e., a (bacterio)chlorophyll without Mg] has a redox potential of −0.8 V.

<table>
<thead>
<tr>
<th>Type of reaction center</th>
<th>Chromophore involved in primary photochemistry</th>
<th>Taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>O₂-Evolvers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC1 (as PSI)</td>
<td>Chlorophyll a (P700)</td>
<td>All O₂-evolvers</td>
</tr>
<tr>
<td>RC2 (as PSI)</td>
<td>Chlorophyll a (P680)</td>
<td>All O₂-evolvers</td>
</tr>
<tr>
<td>Non-O₂-Evolvers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC1</td>
<td>Bacteriochlorophyll g (P700)</td>
<td>Heliothriaceae</td>
</tr>
<tr>
<td>RC1</td>
<td>Bacteriochlorophyll a (P940)</td>
<td>Chlorobiaceae</td>
</tr>
<tr>
<td>RC2</td>
<td>Bacteriochlorophyll a (P940)</td>
<td>Chloroflexaceae</td>
</tr>
<tr>
<td>RC2</td>
<td>Bacteriochlorophyll a or b (P935 or P960)</td>
<td>Chromatiaceae and Rhodospirillaceae</td>
</tr>
</tbody>
</table>
end is much better understood, with the cytochrome
\( b-c \) (or plasto-
cyanin) and hence the primary oxidant of the photoca-
chemical reaction.

The bioenergetic role of cyclic electron transport is the
pumping of \( \mathrm{H}^+ \) from the cytosol (or stroma) to the
external environment (or thylakoid lumen) compart-
ments. Two \( \mathrm{H}^+ \) per electron are pumped via the quinone
and the cytochrome complex. If the NAD(P)/H dehydro-
genase is involved, then a further \( 2\mathrm{H}^+ \) per electron
could be transferred. A major role for these pumped
\( \mathrm{H}^+ \) is to move back across the membrane through the
ATP synthetase; \( 4\mathrm{H}^+ \) energize the phosphorylation of
one ADP. If all the absorbed photons are used to ener-
grize RC1, each photon could produce 0.5–1.0 ATP,
depending on the involvement of NAD(P)/H dehydro-
genase. An alternative use of the \( \mathrm{H}^+ \) gradient is to energize
secondary active transport coupled to exogenic \( \mathrm{H}^+ \) fluxes.

Another role for RC1, and apparently the major one
in almost every situation in which RC1 occurs, is to
generate a reductant at a low enough redox potential
to reduce \( \mathrm{CO}_2 \) to carbohydrate. As will be seen in Sec-

tion IV, in many cases the reductive step in \( \mathrm{CO}_2 \) fixation
involves an energetic input from ATP, and hence the
reductant is NAD(P)/H. The path from the reduced
protein-bound iron–sulfur involves ferredoxin and a
membrane-bound ferredoxin–NAD(P)/H reductase.
The reduction of ferredoxin and NAD(P)/H, and hence
the reduction of \( \mathrm{CO}_2 \) (or \( \mathrm{NO}_3^- \), or other electron ac-
ceptors), requires an input of (weaker) reductant at the
oxidizing end of the photosystenm. For members of the
Chlorobiaceae and Heliobacteriaceae, the reductant is
ultimately some exogenous organic or inorganic elec-
tron donor (other than \( \mathrm{H}_2\mathrm{O} \)) such as \( \mathrm{S}_2 \) \(_2\), which supplies
electrons to the oxidizing end of the photosystem via
cytochrome \( c \), with or without the involvement of the
cytochrome \( b-c \) complex. For the \( \mathrm{O}_2 \)-evolvers, PSI
(photoreaction I) is supplied with reductant by PSII
(photoreaction II) via plastoquinone, the cytochrome
\( f \) complex, and cytochrome \( c_1 \) or plastocyanin.

RC2 does not produce a stable reductant with a
redox potential that is sufficiently negative to reduce
ferredoxin or NAD(P)/H. For the photosynthetic Pro-
teobacteria and the Chlororibacteae, a major function of
the RC2 reaction center is in a cyclic electron transport
pathway. This uses the ubiquinol produced at the reduc-
ing end of the reaction center to reduce the cytochrome
\( b-c \) complex, and then cytochrome \( c_1 \), finally the
oxidant produced by the reaction center is reduced.
This mechanism pumps \( 2\mathrm{H}^+ \) per electron, and hence
per photon, whose excitation energy is transferred to
the reaction center. One use of the H\textsuperscript{+} gradient is, as for cyclic electron flow and H\textsuperscript{+} pumping by RC1, to generate ATP. Another is to use a weak reductant (e.g., S\textsuperscript{−}) to generate a strong reductant (NADH). Here electrons are transported against an energy difference using the energy from the flux of H\textsuperscript{+} from the medium back to the cytosol. This process is the reverse of the active H\textsuperscript{+} transport caused by the respiratory, exergonic, electron transport from NADH to S\textsuperscript{−} (generating S\textsuperscript{2−}). Similar uses of the H\textsuperscript{+} gradient can be envisaged in the Halobacterae.

The other, and quantitatively predominant, occurrence of RC2 in the biosphere is as PSII of O\textsubscript{2}-evolvers. Here the usual immediate sink for the electrons from reduced plastoquinone (=plastocyanin) is the cytochrome b\textsubscript{6f} complex, with subsequent sequential reduction of cytochrome c\textsubscript{1} or plastocyanin, PSI, ferredoxin, and NADP\textsuperscript{+} or, finally, to a smaller extent, some other oxidant such as NO\textsubscript{2} or O\textsubscript{3} (see the following). The corresponding electron donor to the oxidizing side of PSI, after the polypeptide-bound tyrosine Y\textsubscript{Z} and polypeptide-bound Mn, is the very weak reductant H\textsubscript{2}O. It appears that PSII does not engage in cyclic electron transport involving the cytochrome b\textsubscript{6f} complex in the manner of RC2 in the Proteobacteria and Chloroflexaceae, although it may perform cyclic electron transport via cytochrome b\textsubscript{559}, a redox catalyst with no analog in the RC2 of non-O\textsubscript{2}-evolvers. This cyclic electron transport probably occurs as a photochemical means of dissipating excess, and potentially damaging, excitation of PSI.

Putting PSII and PSI together, in terms of photon requirement, the transport of one electron from H\textsubscript{2}O to NADP\textsuperscript{+}, and hence CO\textsubscript{2}, and the associated H\textsuperscript{+} transport, requires the energy of two photons, one used by PSI and the other by PSII. The stoichiometry of H\textsuperscript{+} transport is that one H\textsuperscript{+} is left in the thylakoid lumen, and one H\textsuperscript{+} is taken up in the cytosol or stroma, per electron moved from H\textsubscript{2}O to NADP\textsuperscript{+} and then to CO\textsubscript{2}. In addition, for each electron moving through plastoquinone and the cytochrome b\textsubscript{6f} complex, two H\textsuperscript{+} are transferred from the cytosol or stroma to the thylakoid lumen. Thus, per electron moved from H\textsubscript{2}O to NADP\textsuperscript{+} (and thence CO\textsubscript{2}), 3H\textsuperscript{+} are moved from the cytosol or stroma to the thylakoid lumen:

\[
2\text{NADP}^+ + 2H_2^+O_2 + 12H_2O + 8\text{ photons} \rightarrow 2\text{NADPH} + 2H_2 + 12H_2O + O_2
\] (2)

where the subscript (s) means stroma (or cytosol of cyanobacteria) and the subscript (t) means thylakoid lumen.

With an H\textsuperscript{+}/ATP of 4 in ATP synthetase, the ADP phosphorylation corresponding to the recycling of 12H\textsuperscript{+} can be described by

\[
12H_2O + 3\text{ADP} + 3\text{P}_\text{i} \rightarrow 12H_2 + 3\text{ATP} + 3\text{H}_2O
\] (3)

Thus, each electron moved from H\textsubscript{2}O to NADP\textsuperscript{+} generates 0.75 ATP, so that 3ADP are phosphorylated per 2NADP\textsuperscript{+} (and hence 1CO\textsubscript{2}) reduced. This can be seen when Eqs. (2) and (3) are added together to yield

\[
2\text{NADP}^+ + 2H_2^+O_2 + 3\text{P}_\text{i} + 8\text{ photons} \rightarrow 2\text{NADPH} + 2H_2 + 3\text{ATP} + 3H_2O + O_2
\] (4)

This has important implications for the minimum photon cost of CO\textsubscript{2} fixation in O\textsubscript{2}-evolvers, which is 8 photons absorbed per CO\textsubscript{2} reduced to carbohydrate.

In addition to the reduction of CO\textsubscript{2} and NO\textsubscript{2} by PSI, there is also the reduction of O\textsubscript{2} to form, ultimately, H\textsubscript{2}O. The Mehler Peroxidase reaction involves two electrons from the reducing end of PSI passing to two O\textsubscript{2} to form two superoxide anions, which are then dismutated by the enzyme superoxide dismutase to form one H\textsubscript{2}O and one O\textsubscript{2}. The H\textsubscript{2}O, plus ascorbate, are then acted on by the enzyme ascorbate peroxidase, which converts one H\textsubscript{2}O and one ascorbate to two H\textsubscript{2}O and one dehydroascorbate, involving (via other enzymes) two more electrons from PSI to regenerate the ascorbate co-substrate from dehydroascorbate. Overall, the reaction involves the evolution of one O\textsubscript{2} at the oxidizing end of PSI and the uptake of one O\textsubscript{2} at the reducing end of PSI, so that there is no net evolution or uptake of O\textsubscript{2}. The reaction can be quantified by the use of \textsuperscript{18}O\textsubscript{2} tracer. The Mehler Peroxidase reaction is a means of disposing of active, and potentially damaging, oxygen species produced in thylakoid redox reactions and, to the extent that superoxide production by PSI can be increased above the unavoidable basal rate, a means of photochemical energy dissipation and of ATP generation not paralleled by NADP\textsuperscript{+} reduction at an alternative to cyclic photophosphorylation.

The photon cost of CO\textsubscript{2} fixation in terms of incident irradiance depends on the photon cost of CO\textsubscript{2} fixation in terms of absorbed photons (8 in the case mentioned here) and the fraction of the incident photons that are absorbed (see Section I). The catalytic capacity of the reactions generating NAD(P)H and ATP can, under many circumstances, constrain the rate of CO\textsubscript{2} fixation when photon supply is not limiting. However, in some
cases it is the supply of CO₂, or the capacity of CO₂ assimilation reactions (see Section IV), that constrains the CO₂ fixation rate at saturating light supply. The capacity for light harvesting relative to that for primary photochemical reactions and downstream thylakoid reactions and CO₂ fixation reactions (see Section IV) depends on the irradiance that the plant has encountered in its evolutionary history (genetic adaptation) and during the life of an individual (phenotypic acclimation). In both adaptation and acclimation, the response to low light is an increased capacity for light harvesting relative to downstream reactions, and vice versa for the response to high light.

In O₂-evolvers the rate and (incident) photon cost of CO₂ fixation at limiting incident irradiances, and the rate of CO₂ fixation at saturating irradiances, can be reduced by photoinhibition. This phenomenon has its basis in damage to the PSII reaction center by one in 10⁶ to 10⁷ of the photons whose excitation energy reaches the reaction center. In the absence of any protective mechanisms, this photodamage would occur more rapidly at higher incident photon flux densities, but would result in a greater fractional inhibition of photosynthesis when measurements on organisms exposed to high incident photon flux densities are subsequently made at lower (rate-limiting) irradiances. The observed phenomenon of photoinhibition, that is, a reduced rate of photosynthetic CO₂ fixation during, and following, exposure to (especially high) irradiances, is not solely a function of photodamage occurring faster than it can be repaired. Thus, the decreased photosynthetic rate can be a result of diversion of the products of PSII photochemistry to processes other than CO₂ fixation at high incident irradiances, for example, to cyclic electron flow round PSI or to the Mehler Peroxidase reaction, decreasing the photochemical unusable “excess” excitation of PSI. More commonly, excitation energy is prevented from reaching the PSII reaction center by dissociation of some of the light-harvesting pigment–protein complexes (Table II) from the reaction center, or by nonphotochemical quenching of excitation in the light-harvesting complexes by processes triggered by very low pH in the thylakoid lumen. This increased acidification of the lumen is a result of photochemical reactions at which H⁺ pumping in excess of the rate at which H⁺ is recycled through the ATP synthetase due to restricted ATP consumption and hence limited ADP availability, and can dissipate excitation energy via a xanthophyll cycle mechanism in all but cyanobacteria and red algae and, in all O₂-evolvers, via mechanisms not related to a xanthophyll cycle.

IV. USE OF NAD(P)H AND ATP IN CO₂ FIXATION AND OTHER REACTIONS

Most of the ATP and NAD(P)H generated in the thylakoid reactions in O₂-evolvers is used in CO₂ reduction to (CH₂O). Other uses for the ATP resulting from photochemical reactions include photoassimilation of exogenous organic compounds, as well as other transport and biochemical reactions. Photochemically produced reductant can be used for reductive biosyntheses, such as reduction of NO₂, NO₃, and SO₂⁻. These uses of ATP and reductant notwithstanding, the major global biogeochemical role of NAD(P)H and ATP generated in photochemical reactions is the conversion of CO₂ to carbohydrates; a net primary production of ~100 Pg carbon per year worldwide is carried out by O₂-evulators. However, some CO₂ is fixed by non-O₂-evolving autotrophs. In the Chlorobiaceae, the reverse tricarboxylic acid cycle is used, whereas the Chloroflexaceae use the hydroxypropionate cycle (Table III). These two processes each involve more than one carbon fixation reaction in converting CO₂ to carbohydrates. By contrast, the photosynthetic carbon reduction cycle used in the Proteobacteria and in all O₂-evolvers involves but a single carboxylase, that is, ribulose-1,5-bisphosphate carboxylase–oxygenase (RUBISCO), in converting CO₂ to carbohydrates, although some essential biosyntheses involve parallel, quantitatively minor, carboxylation reactions catalyzed by other carboxylases.

RUBISCO, as its name suggests, not only catalyzes a carboxylase reaction but also has an oxygenase activity. The oxygenase activity is competitive with the carboxylase activity. Thus, the oxygenase is not a significant reaction in proteobacterial photosynthesis in their typical high-CO₂, low-O₂ habitats where heterotrophic organisms dominate and organic carbon sedimentation into the habitat exceeds O₂ diffusion into it. Furthermore, the oxygenase activity would not have been significant in the high-CO₂, low-O₂ world prior to 2 billion years ago, that is, for the first 1.5 billion years or so of the occurrence of RUBISCO. When oxygenase activity is negligible, the conversion of CO₂ to the carbohydrate redox level by the photosynthetic carbon reduction cycle needs 2NAD(P)H and 3ATP:

$$2\text{NADPH} + 2\text{H}^+ + 3\text{ATP} + 3\text{H}_2\text{O} + \text{CO}_2 \rightarrow (\text{CH}_2\text{O}) + 2\text{NADP}^+ + 3\text{ADP} + 3\text{H}_2\text{O}$$  

Equation (5), showing the consumption of NADPH and ATP in CO₂ fixation, can be combined with Eq. (4), describing the production of NADPH and ATP by light-
### TABLE III
**Taxonomic Distribution and Energy Costs of CO₂ Assimilation Reactions in Photosynthetic Organisms**

<table>
<thead>
<tr>
<th>CO₂ assimilation reaction</th>
<th>Taxonomy</th>
<th>NAD(P)H cost per net CO₂ fixed</th>
<th>ATP cost per net CO₂ fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverse tricarboxylic acid cycle</td>
<td>Chlorobaccaeae</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>Hydroxypropionate pathway RUBISCO and photosynthetic carbon reduction pathway</td>
<td>Chloroflexaceae</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>(i) In high CO₂–low O₂, (present atmosphere or water in equilibrium with it)</td>
<td>Proteobacteriaceae</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>(a) Diffusive CO₂ entry to RUBISCO, with photosynthetic carbon oxidation cycle</td>
<td>Cyanobacteria; many algae</td>
<td>2.8</td>
<td>4.2</td>
</tr>
<tr>
<td>(b) Active transport of inorganic C to RUBISCO by a process that does not involve C–C bond formation</td>
<td>Cyanobacteria; many algae</td>
<td>2</td>
<td>4–5</td>
</tr>
<tr>
<td>(c) Active transport of inorganic C to RUBISCO by a process that does involve C–C bond formation</td>
<td>CAM higher plants</td>
<td>2</td>
<td>6–9</td>
</tr>
</tbody>
</table>

Driven processes in the thylakoid, to give the overall equation for photosynthesis [Eq. (1)].

While the reductant requirement per CO₂ assimilated is the same as for the reverse tricarboxylic acid cycle and the hydroxypropionate cycle, the ATP requirement is greater for the photosynthetic carbon reduction cycle (see Table III).

The greater energy requirement for the photosynthetic carbon reduction cycle than for the other carbon fixation cycles is exacerbated in the present atmosphere, or in natural waters in equilibrium with it. If CO₂ enters by diffusion (C₃ physiology) then there is significant oxygenase activity of RUBISCO, although there is significant variability among RUBISCOs of different organisms in selectivity for CO₂ relative to O₂. The oxygenase activity produces phosphoglycolate, and thence glycolate. Recycling this glycolate to carbohydrate has a significant energy cost; the excretion of glycolate in aquatic organisms imposes an even greater energy cost for net CO₂ fixation. The energy cost of the net fixation of 1CO₂ from the present atmosphere by a vascular land plant with C₃ physiology (CO₂ reaching RUBISCO by diffusion) is some 2.8 NADPH and 4.2 ATP. This contrasts with the 2NADPH and 3ATP required when CO₂ is saturating and, regardless of O₂ concentration, there is no oxygenase activity (see Table III). C₃ physiology occurs in the great majority of species of terrestrial algae, and some aquatic algae and embryophytes. Some red algal RUBISCOs have the highest reported selectivity for CO₂ over O₂, so that glycolate production may be minimal in red algae with C₃ physiology. The algae, lichens, and bryophyte gametophytes with C₃ physiology have limitations on photosynthesis by CO₂ diffusion that vary with the structure of the photosynthetic apparatus, whether the medium supplying inorganic carbon (CO₂/HCO₃⁻) is air or water, the CO₂ concentration in the medium, and the thickness of the diffusion boundary around the photosynthetic structure. In sporophytes of embryophytes in air the situation is as for gametophytes, whereas for sporophytes in air there is the additional complication of stomata. These variable conductance apertures connecting intercellular gas spaces to the atmosphere are crucial in maintaining land plants in a hydrated state when soil water availability is low and/or the atmosphere is very dry, and permit CO₂ fixation in the light when the plants can remain hydrated with open stomata. In most cases the C₃ vascular plant sporophytes on land fix CO₂ under conditions in which the rate of fixation is about one-third constrained by CO₂ diffusion and two-thirds constrained by biochemical reactions. In other words, were CO₂ diffusion constraints to be reduced by 10%, the rate of photosynthesis would increase by 3%; if the constraints from biochemistry were reduced by 10%, the rate of CO₂ fixation would increase by 7%. The fractional limitation of photosynthesis by CO₂ diffusion...
in aquatic plants with C₃ physiology is more variable than in the terrestrial vascular plant sporophytes with this photosynthetic physiology. In a number of species of O₂-evolvers the supply of CO₂ from the environment to RUBISCO involves some energized process that gives a higher CO₂ concentration, and CO₂/O₂ ratio, near RUBISCO than that in the medium. Such mechanisms suppress oxygenase activity, in some cases almost completely, and frequently increase the affinity of in vitro photosynthesis for CO₂ relative to the situation with C₃ physiology (diffusive supply of CO₂ to RUBISCO).

One group of these CO₂-concentrating mechanisms does not involve covalent bond formation between CO₂ and an organic compound before RUBISCO activity. These mechanisms have considerable diversity, and can involve active transport of CO₂, HCO₃⁻, or H⁺ as means of concentrating CO₂. Such processes occur in many aquatic O₂-evolvers, including all of the cyanobacteria, many of the algae, and some of the submerged vascular plants found in aquatic environments, and in a few terrestrial O₂-evolvers, including all of the cyanobacteria, many of the terrestrial algae and lichens, and, among embryophytes, certain hornworts. There is an energetic cost of operating the CO₂-concentrating mechanism; this has not been well defined for any mechanistic variant. In the case of cyanobacteria and dinoflagellates, the kinetic properties of RUBISCO mean that C₃ photosynthetic physiology is not an option for growth in an air-equilibrium solution. In some of the other organisms, the selective significance of a CO₂-concentrating mechanism is quantitatively less clear (see Table III).

A second group of CO₂-concentrating mechanisms does involve covalent bond formation between CO₂ and an organic compound in the delivery of CO₂ from the environment to RUBISCO. These mechanisms are essentially restricted to vascular plants, with C₄ metabolism occurring only in certain flowering plants, whereas CAM (Crassulacean Acid Metabolism) is found in a wider range of vascular plants. Here the initial carboxylation reaction preceding RUBISCO is almost invariably catalyzed by phosphoenolpyruvate carboxylase (PEPC) and an organic compound before RUBISCO activity. The core reactions of primary photochemistry (RC₁, RC₂ in Bacteria and eukaryotes; bacteriochlorophyll and halorhodopsin in Archaea) show relatively little variation in structure or overall function, although there are significant differences (e.g., RC₂ in O₂-evolvers can dehydrogenate water). Similarly, the quinones, the cytochromes b–c₁ and b–f, cytochrome c₅₅₃, and ferredoxin also show little variation, as do plastocyanin (alternative to cytochrome c₅₅₃ in some O₂-evolvers) and flavodoxin (alternative to ferredoxin in some phototrophs). ATP synthetase also shows great similarity among bacterial and eukaryotic phototrophs. There is more variation among the reactions upstream of the photosynthetic reactions that are involved in photon harvesting and excitation energy transfer.

The core CO₂ fixation reaction using RUBISCO in O₂-evolvers and Proteobacteria is supplemented in some O₂-evolvers by a diversity of CO₂-concentrating mechanisms. In many aquatic plants there are CO₂ pumps not based on preliminary carboxylation and decarboxylation, whereas these reactions are found in C₄ and CAM. Bacteria other than Proteobacteria fix CO₂ via reactions not involving RUBISCO.

V. CONCLUSIONS

The core reactions of primary photochemistry (RC₁, RC₂ in Bacteria and eukaryotes; bacteriochlorophyll and halorhodopsin in Archaea) show relatively little variation in structure or overall function, although there are significant differences (e.g., RC₂ in O₂-evolvers can dehydrogenate water). Similarly, the quinones, the cytochromes b–c₁ and b–f, cytochrome c₅₅₃, and ferredoxin also show little variation, as do plastocyanin (alternative to cytochrome c₅₅₃ in some O₂-evolvers) and flavodoxin (alternative to ferredoxin in some phototrophs). ATP synthetase also shows great similarity among bacterial and eukaryotic phototrophs. There is more variation among the reactions upstream of the photosynthetic reactions that are involved in photon harvesting and excitation energy transfer.

The core CO₂ fixation reaction using RUBISCO in O₂-evolvers and Proteobacteria is supplemented in some O₂-evolvers by a diversity of CO₂-concentrating mechanisms. In many aquatic plants there are CO₂ pumps not based on preliminary carboxylation and decarboxylation, whereas these reactions are found in C₄ and CAM. Bacteria other than Proteobacteria fix CO₂ via reactions not involving RUBISCO.
See Also the Following Articles
ARCHAEA, ORIGIN OF • BACTERIAL BIODIVERSITY • CARBON CYCLE • EUKARYOTES, ORIGIN OF

Bibliography
A phylogeny is a pattern of historical evolutionary relationship among species and higher level taxa that is often presented as a tree diagram, or phylogenetic tree. The term phylogenetics is often applied to the study of such relationships. Historically, phylogenetic trees were often produced by indirect methods and were not reproducible. Classifications often had little if any direct relationship to phylogeny. Modern phylogenetics utilizes cladistic methods to construct phylogenetic trees based directly on morphological and molecular data. For those who distinguish cladistics from phylogenetics, cladistics refers only to the methods by which the branching patterns are generated (e.g., parsimony or maximum likelihood) while phylogenetics refers to the interpretation of such diagrams as historical patterns. This is a useful distinction, since cladistic methods are neutral to the type of data and the resulting interpretations, and could be applied to nonphylogenetic problems (e.g., recovering or imposing hierarchic structure within any system of objects with shared variation). Phylogenies are analogous to genealogies on the scale of species and higher level taxa (e.g., genera and families). Phylogenies are usually presented as treelike branching diagrams, in which taxa that are on the same branch are thought to be more closely related to each other than taxa that occur on different branches. Interpreting such diagrams as historical patterns requires a basic understanding of hierarchy, and phylogenetic trees are often incorrectly assumed to support particular historical suppositions (e.g., one modern taxon is “primitive”) that are not indicated by the results. Phylogenetic trees are increasingly useful in a broad array of biological studies as a basis for experimental design as well as the framework on which to generalize results. Additional uses of phylogenetic information include
measures of phylogenetic diversity, which can be used in making conservation and habitat preservation decisions.

I. HISTORICAL OVERVIEW

In the late eighteenth century, taxonomy became stabilized with the work of Linnaeus and others, and the modern hierarchic system of classification (often called the Linnean system) became standard. Hierarchical classification systems in the late eighteenth and early nineteenth centuries were often illustrated with treelike diagrams, but in general these were not intended to show evolutionary relationship and should not be interpreted as such. Following Darwin and the ultimate acceptance of evolution as the explanation for the diversity and variation of life, efforts began to produce phylogenies (diagrams of higher level evolutionary divergence among taxa), which also were often presented as treelike diagrams. The methods by which these early phylogenies were produced varied, most often based on a combination of raw similarities among taxa and either implicit (or explicit) ideas about the way in which characters had evolved. For example, Bessey in the early twentieth century published a list of dicta indicating trends in morphological characters of flowering plants (e.g., flower parts primitively with many parts, becoming fewer) and used these dicta as an implicit means of deriving phylogenies. Anecdotal theories about character evolution have been utilized by many systematists as an explicit or implicit basis for major phylogenies even to the present day (e.g., Cronquist, 1981).

Early phylogenies were usually considered independently from systems of classification, or classifications were only partially based on the results of phylogenetic analysis. One of the most prominent features of classifications from this period that persists in many current classifications is the recognition of paraphyletic groups (see below).

In a modern discussion of phylogeny, there are three main aspects that need to be addressed. First is the issue of how the phylogeny is derived: the source of evidence and how that evidence is analyzed. Second, there is the nature of the resulting phylogenetic diagram, or pattern, and what information such a diagram conveys. Third, there is the relationship between a phylogenetic system (a phylogenetic tree) and a classification that may or may not be derived from such a phylogenetic pattern. All of these issues remain controversial at some level, with different workers taking different combinations of views on each topic. While some may view these controversies as overwrought, one healthy consequence for science is the competitive improvement of methods and the need for explicit theoretical justifications by all participants in these arguments.

A. Modern Phylogenetic Analysis

Beginning with the popularization of the works of Hennig in the late 1960s and early 1970s, the idea that raw similarity should be the basis for phylogenetic reconstruction was quickly overturned. One might refer to this as the "cladistic revolution." By the mid-1980s Hennig's notion that only special similarities, or synapomorphies, should be used to reconstruct relationships had become the dominant force in phylogenetic analysis. Although one still finds phenetic analyses published, these are often focused on intraspecific patterns where cladistic methods have not yet been fully developed. The vast majority of phylogenetic analyses in the late 1980s to the present were undertaken with cladistic methods, in particular parsimony (see below). Other related methods such as neighbor-joining and maximum likelihood have been embraced by a subset of workers, but for various reasons are still not as widely used as parsimony. Thus, the discussion here will focus primarily on parsimony.

II. ANALYTICAL METHODS

A. Phenetic Methods

Phenetics had its basis in the statistical literature that developed out of the "New Synthesis" and came to fruition with the advent of large and accessible computers in the late 1930s and early 1960s. The various methods of phenetic analysis are too many to enumerate here but generally can be classified into methods that produce treelike diagrams (clustering) and methods that produce scatter diagrams on two or more axes (ordination).

Phenetic methods do share the premise that all similarities, whether derived or primitive, should be used equally in an analysis. It is debatable as to whether the original proponents of phenetic methods actually intended the results of such analyses to be interpreted as phylogenies, but cluster methods produce treelike diagrams that were often directly interpreted as phylogenetic trees. Various justifications for phenetic methods have been proposed, and indeed, the methods as well as the theoretical underpinnings are eclectic. It became apparent in the 1970s and early 1980s that phenetic methods were not defensible as methods for...
recovering phylogenetic history, and parsimony analysis became the method of choice for most systematists interested in reconstructing patterns of divergence. Those interested in these debates can find entry into the literature by examining almost any issue of the journal Systematic Zoology from ca. 1970 to 1985. By the 1990s, phenetic methods had become rarely used except in particular subspecific and populational studies, and need not be considered further here.

B. Phylogenetic Methods

From the standpoint of methods of analysis, there is little to be gained by separating phylogenetics from cladistics although it should be noted that cladistics is the broader of the two. Cladistics is derived from the Greek term clado for ‘branch’ and was coined by opponents of such methods to refer to those who were more interested in the pattern of divergence than in the process and degree of divergence. Originally the term cladistics was mostly applied to parsimony analysis (and some related methods that have fallen out of favor) but at the present time cladistics might be more broadly interpreted to include at least two other methods, neighbor-joining and maximum likelihood. However, as with all such terminology, there are differences of opinion, and some workers would rather restrict the term cladistics to parsimony-based methods only. For the purposes of this discussion, we will use the term in the broader sense. Any method that results in a (nonreticulate) tree diagram that is interpreted as representing the evolutionary history and relationships of taxa can be considered phylogenetic. Because of problems in interpreting most phenetic diagrams as phylogenies, those methods cannot be considered phylogenetic.

1. Parsimony

Parsimony analysis remains the method of choice for analysis of character data by most morphological systematists, and perhaps still even for molecular systematists. Modern parsimony analyses utilize matrices of either molecular or morphological data, or a combination of the two, to produce cladograms. Although Hennig did not propose parsimony analysis as we now implement it, his works provided the basis for the mathematical and theoretical underpinnings of parsimony analysis that were largely developed by Farris and others from 1969 to 1981 in a series of papers published in Systematic Zoology and elsewhere. Parsimony is derived from Occam’s razor—the idea that the simplest explanation is always preferable. Parsimony analyses (when successful) result in cladograms (branching diagrams) that have the fewest number of steps (character transformations) when characters are optimally mapped onto the diagrams. Such trees are usually more or less directly interpreted as phylogenies, and character steps on these trees represent the minimum possible number of evolutionary changes to explain the distribution of the character. Such an interpretation does not require the assumption that evolution itself is parsimonious; indeed, most such trees based on real data have multiple origins for many, if not most, characters in the matrix. Parsimony simply eliminates the need for ad hoc explanations for multiple character origins by reducing them to the minimum possible number.

2. Parsimony Analyses

One major difficulty in parsimony analyses (and to an even greater degree in maximum likelihood) is the computational difficulty in finding most parsimonious solutions. There currently exist no methods for directly calculating most parsimonious trees. Because of this, the problem is directly related to the number of possible trees to be examined and is considered to be NP-complete. Because there is no direct way to calculate shortest trees, brute force methods that examine many possible trees must be used. These methods are generally implemented by ‘branch swapping’ on a starting tree and examining the length of each new tree generated by moving branches to different positions on the tree. For very small data sets (<15 taxa), it is possible to examine all possible trees and be certain that one has collected all of the most parsimonious trees. For slightly larger data sets (possibly up to 30 taxa, depending on the quality of data), it is possible to use a ‘branch and bound’ algorithm that reduces the number of trees examined but still guarantees that all shortest trees are found. However, for the vast majority of analyses, more than 30 taxa, and in recent analyses, as many as several hundred taxa (or “terminals”), are included. It is important to note here that the number of characters (e.g., the length of a DNA sequence) only affects computation in a roughly linear manner, while computation time is affected logarithmically as the number of taxa increases.

With larger analyses, branch swapping is used under various ‘quick’ strategies to find locally optimal trees and these trees are then used as starting points for additional searches. Recent advances in search strategies have resulted in programs that can easily analyze data sets of several hundred taxa with a very high confidence that shortest trees have been found (see Goloboff, 1999). Currently, for matrices larger than 1000 taxa, quick estimation methods such as the jackknife are used instead of directly searching for shortest trees.
a. Homology
The term homology, like the term monophyly, is used in various ways in the systematic literature. However, most recent workers accept the term to mean a shared similarity that is due to (explainable by) common descent. Thus, any similarity that is not explained by common ancestry is not homology. Characters that are initially thought to be homologous but are found by analysis to be nonhomologous are termed homoplasious. Homoplasy is best considered as error in homology assessment, which can be explained by parallelism, convergence, or other more complex processes such as hybridization and retained polymorphism. Homologous characters by definition are evidence of relationship. One theoretical basis for cladistic analysis using parsimony is that parsimony maximizes the homology statements presented in the data matrix. Thus, when we score two terminals (taxa, species, etc.) as having the same state for a character, we are asserting a hypothesis of homology (the “primary” homology of some workers). Parsimony then resolves conflict among such statements by picking the tree (cladogram) that maximizes the number of correct homology assessments (or minimizes the number of incorrect assessments). In so doing, it minimizes the number of ad hoc explanations that are necessary to explain similarities that appear to be homologous at first, but are not based upon the most parsimonious tree. Farris and others have demonstrated that the shortest trees maximize explanatory power as well as information content of a classification and are therefore preferable even outside of an evolutionary context. It is important to note that parsimony does not assume that evolution is parsimonious, only that parallelism and convergence are not correlated in such a manner as to be misleading. Given randomly distributed parallel and convergent false homologies, with enough “true” homologies parsimony will resolve the correct tree (the tree that most accurately reflects phylogenetic history). This assertion has been challenged recently by proponents of maximum likelihood, but this controversy remains unresolved at this time and it appears that all methods (parsimony and maximum likelihood included) fail under certain circumstances.

3. Neighbor-Joining (NJ)
Neighbor-joining is a method that is arguably cladistic and is favored by many molecular biologists because it is rapid and generally (at least as commonly implemented) results in a single tree. The basic algorithm of NJ is similar to the Distance Wagner method of Farris and constructs a tree by calculating distances among terminals and adding terminals in such a way as to minimize overall distance. While NJ is an expedient method for quickly obtaining an estimate of the most parsimonious tree, there seems to be no justification for preferring an NJ tree over a shorter tree obtained by parsimony analysis.

4. Maximum Likelihood (ML)
Maximum likelihood as a method of analysis is popular among many molecular biologists, particularly those who are more interested in models of evolution than in the actual phylogenetic pattern of taxon relationships. ML requires an explicit model of character transformation, with associated probabilities for each possible transformation from one state to another. Trees are searched in a manner similar to methods used in parsimony (e.g., branch swapping) but each tree is evaluated not by overall length, but instead by a measure of compound probability. Those trees with the highest compound probabilities (maximum likelihood) of character distribution are selected as best. Within certain theoretical frameworks, parsimony can be viewed as a particular form of maximum likelihood with reduced assumptions and infinite parameters. Under such circumstances, parsimony analyses and maximum likelihood analyses will give the same results. Under most circumstances, maximum likelihood and parsimony analyses of the same data sets have provided very similar results. However, at the present time maximum likelihood is not feasible for larger data sets due to massive computation times (at least with today’s hardware and software). The computational problems are unlikely to be resolved by the mere improvement of software that are probably not possible with current engineering capabilities in this field. However, it is possible that recent advances in algorithms for parsimony searches can be incorporated into maximum likelihood programs with similar relative levels of improvement. Because of the additional complexities of ML, even with such improvements, the ability to perform maximum likelihood analyses on the ever-larger data sets being produced with molecular techniques will lag behind parsimony for the foreseeable future.

III. MEASURES OF SUPPORT
A. The Bootstrap
Many workers wish to place some measure of statistical confidence on the phylogenetic trees that they construct. Because of theoretical issues, it is preferable to
referred to such statistical measures as support values as opposed to confidence intervals. The most widely used support measure in recent literature is the bootstrap. The bootstrap is calculated by permuting the original data set numerous times by sampling characters with replacement and recalculating the most parsimonious trees for each permuted data set. Usually, from 100 to 1000 separate permuted data sets are analyzed, and support for a particular clade can be evaluated as the number of separate replicates in which that clade occurs in the most parsimonious solutions. Thus, if a group is found in parsimony analyses of 90 of 100 permuted data sets, its bootstrap value would be 90. Theoretical issues surrounding the bootstrap include the assumptions about character distribution and sampling and the exact interpretation of the bootstrap values that are obtained for each branch in the tree. Because the relationship between the bootstrap values and the underlying "true tree" is not apparent, many workers prefer to view bootstrap values as an assessment of the strength of signal within a particular data set, and not as a statistical measure of confidence except in a very restricted sense. This has resulted in development of an alternative terminology, under which bootstrap (and jackknife) values are referred to as "measures of support.

B. The Jackknife

The jackknife is closely related to the bootstrap and is performed also by permuting the data set and counting the occurrence of groups (clades) among replicates. The jackknife differs from the bootstrap in that instead of resampling characters, characters are removed at a set level of probability—the probability of removal usually set to be the same as the probability of not being sampled in a bootstrap. The jackknife has some theoretical advantages over the bootstrap and is generally faster to calculate because of very fast software that is available. The jackknife has been used recently as the sole analytical method for some very large molecular data sets (e.g., Kallersjo's analysis of an rbcL data set for green plants with more than 2000 taxa). Indeed, some proponents of the jackknife have suggested its exclusive use in lieu of more intensive parsimony analyses, under which bootstrap (and jackknife) values are referred to as "measures of support.

C. Bremer Support

Another measure of support commonly used in conjunction with parsimony analyses is Bremer support, sometimes incorrectly termed the "decay index." Bremer support is measured by collecting suboptimal trees (trees of longer length than the shortest) and determining which groups are present. If, for example, a clade is present in shortest trees and all trees up to 2 steps longer, but is not present in all trees 3 steps longer, the Bremer support value for that group is 3 steps. While Bremer support is appealing because it utilizes parsimony directly, it is difficult to compare Bremer support values among different data sets. In general, the bootstrap and jackknife have been preferred in recent years as measures of support, and especially for large data sets, the former two are much easier to calculate than Bremer support.

IV. DATA SOURCES

Phylogenetic analyses have evolved from being entirely or largely based on morphological data to being more and more based on molecular data, in particular DNA sequence data. There are both theoretical and practical reasons for this shift. First, many believe that molecular data are more objective and that character definition and scoring of morphological characters is both more subjective and subject to error. Because the expense of sequencing DNA has dropped dramatically, it is arguable now that it is actually more cost-effective to extract and sequence DNA than to spend large amounts of time tediously collecting morphological data. One problematic aspect of DNA sequence data is the lack of such data for fossil taxa. Fossil taxa, when available and when well preserved, can often be included in morphological cladistic analyses and may provide additional insight into patterns of relationship. Occasionally, such fossils even may have dramatic effects on the position of particular taxa. A major caveat of the inclusion of fossil taxa in cladistic analyses is the large amount of missing or incomplete data for fossils, especially in "soft" features that do not preserve well. Terminals with large amounts of missing data have the potential to "move" on the most parsimonious trees, resulting in many equally parsimonious solutions and in essence "dersolving" the tree. Such taxa that move because of ambiguity in data are often referred to as "wildcards" because of the propensity for them to match multiple terminals in the analyses. In general, it is best not to include such fossils when there is evidence that they are not contributing relevant information to the analysis but instead are reducing the precision of the results.

In the early 1990s a debate developed about the best way to treat disparate data sets, such as different gene
sequences, or gene sequences and morphology. One side suggested that data sets should be analyzed separately unless it could be shown that they were congruent, while the other proposed that data sets should be combined as the best method to resolve such incongruence. Currently, it appears as though the latter position is more popular, since most multigene studies combine data sets in a simultaneous analysis regardless of whether tests indicate congruence.

The availability of vast quantities of molecular data across broad groups of taxa holds the promise of providing more or less complete large-scale phylogenies in the near future. Theoretically, larger analyses with more taxa should provide better estimates of relationship that are less sensitive to the addition of new data. Using parsimony jackknifing, Lipscomb and co-workers (1998) analyzed all of the eukaryotes in the Ribosomal Database Project (480 eukaryotes, 15 prokaryotes as outgroups). This analysis confirmed previous morphological and molecular studies that indicated that the protista and fungi were not monophyletic. However, this analysis showed that several of the startling and “new” findings of molecular systematics were probably the result of analysis of small data sets. In particular, there was no evidence for placing Giardia and its relatives as the most primitive eukaryotes nor any clear indication that the Dinoflagellates were sister to the Apicomplexa.

V. PHYLOGENETIC INTERPRETATION

One aspect of phylogeny and phylogenetic reconstruction that often presents difficulty is the interpretation of phylogenetic trees as evolutionary diagrams. Imprecision in both terminology and concepts has resulted in interpretations that are often inaccurate and unsupported. One of the most common mistakes in interpreting cladograms (or phylogenies) is the tendency to view the trees as providing a linear sequence of events, with a “first branch” that bears a more primitive group, followed by a “second branch,” etc. Such terminology generally confounds group size with time of origin, usually placing the smaller extant group of two concordant groups at the “base” of the tree. Thus, recent molecular analyses of the angiosperms have often been characterized as placing Amborella at the base or as the first branch of the angiosperms (see Fig. 1). However, the correct way to view such diagrams is in a dichotomous linear time scale, with each successive dichotomy younger than the preceding. In Fig. 1, time can be viewed as proceeding from left to right. Amborella is not itself the first branch anymore than the angiosperms minus Amborella is the first branch—the first two branches of the angiosperm tree bear a single species, Amborella, on one side and a few hundred thousand species (the remainder of angiosperms) on the other side. In terms of primitive or derived features, Amborella might be more derived than numerous persisting species on the other, larger branch of angiosperms and should not be viewed simply as the “most primitive” angiosperm. For instance, any one of several taxa near the base of the diagram (e.g., Nymphaea) might be more similar to the primitive angiosperm than is Amborella. However, Amborella might be considered to be the “most unique” angiosperm since it is the sole representative of one of the two basal clades (based on extant species alone) of angiosperms, and in that sense is the equivalent of the remainder of angiosperms. This simple and direct view of branching patterns in a phylogeny is not only the correct way to interpret phylogenetic diagrams, but it also provides the basic concept of phylogenetic diversity, or the measure of phylogenetic uniqueness, that is outlined below.

Another common mistake in dealing with phylogenies is the interpretation of polytomies, or multifurcations in cladograms, as indicating rapid radiation from a single ancestral plexus (see Fig. 2). Often, such patterns are referred to as “star phylogenies.” Although evolution may not always produce dichotomous divergences, it is not possible to positively identify multifurcations as rapid divergences. Instead, such patterns may simply reflect a lack of sufficient data to resolve more completely bifurcating trees. In general, polytomies re-
A portion of a phylogenetic tree based on a consensus tree from a parsimony analysis of a selected group of families of flowering plants. Note the large polytomy, or "star phylogeny" as such patterns are sometimes called. Such structures do not necessarily indicate rapid divergence from a common ancestor, but instead reflect lack of sufficient evidence, or conflicting evidence, regarding relationships. The graphic tree style was selected to more clearly indicate polytomies.

Another frequently misunderstood aspect of phylogenetic analysis is the role of *a priori* character polarization. In the 1970s, many of the computer programs used to calculate phylogenetic trees required *a priori* character polarity (the assignment of "primitive" and derived states). However, strict parsimony analysis will produce the same topology regardless of the ultimate character polarity of the characters; in fact, parsimony computer programs do not consider character polarity during tree searches. Beginning in the late 1970s, it became apparent that the correct method for rooting trees was to select one or more outgroups that were included in the analysis, and the resulting trees rooted between the outgroup(s) and the ingroup. Although the best choices for outgroup taxa are those that are closely related to the ingroup, virtually any taxa not part of the ingroup may be used. However, the more distant the outgroups, the less accurate the rooting is likely to be, and particularly with molecular data, if the outgroups are too far removed, virtually random patterns of rooting may emerge. Character polarity becomes a matter of reading character distributions from the resulting trees when the outgroup method is employed, not an *a priori* exercise. Not only is this far easier than attempting to assign polarities to characters directly, but it reduces a major source of subjectivity in the analysis of data. Maximum likelihood models that allow equal probability for changes in any direction are also calculated independent of *a priori* rooting, but any models that have asymmetric models of character change are sensitive to a choice of a root, or starting point.

**VI. PHYLOGENY AND CLASSIFICATION**

The issue of how to classify taxa into formal named systems may seem trivial to some biologists, but it has a profound impact on the way that we think about living things, and therefore does have an impact on the design and implementation of many biological studies. Statements such as "reptiles are cold blooded" or "reptiles have scales" or "the dinosaurs went extinct in the late Cretaceous" all have different interpretations depending on the classification in use. Characterization of paraphyletic groups on the basis of widespread and notable characters (e.g., the scales of "reptiles" or the naked seeds of "gymnosperms") may provide useful means of communicating, but in formal classifications paraphyletic groups often provide misleading implications about relationship. Thus, if we include crocodiles in the group reptiles, we might guess that crocodiles are more closely related to lizards than to birds; this in fact is not the case, based on several lines of evidence when analyzed cladistically that reveal that crocodiles are more closely related to birds than to other "reptiles"—in other words, crocodiles and birds form a monophyletic group that excludes other reptiles and mammals (see Fig. 3). The advantage of a phylogenetic classification lies in the information content; if we know that a classification adheres strictly to the rule that all named groups are monophyletic, then all members of the group will be more closely related to each other than to any taxa not included in the group. Traditional groups such as "reptiles" (or formally, Reptilia) do not meet such criteria, and unfortunately it is impossible to know from such a nonphylogenetic classification that crocodiles (when placed within the "reptiles") are more closely related to birds than to other "reptiles." This observation at once supports both the need to abandon paraphyletic groups in formal classifications, and in doing so, to adopt a system of classification that is based consistently on phylogeny—with named groups always...
A phylogenetic representation of relationships among major groups of vertebrates, based on several published studies of morphological and molecular data. Note that crocodilians are more closely related to birds than to other "reptiles."

corresponding to groups based on available evidence that are believed to be monophyletic.

Much of the difficulty that some systematists and many outside of systematics have with the abandonment of paraphyletic groups comes from the intuitive nature of such groups and their often easy characterization (e.g., reptiles are scaly, air-breathing, cold-blooded vertebrates that lay eggs). A classification that recognizes paraphyletic groups must do so based on primitive (plesiomorphic) features, as in the example just provided. The difficulty, and the arbitrary factor in such classifications, comes not in what to include in a paraphyletic group but instead in what to exclude. Because there is no objective way to determine what groups are significant enough to be excluded, the final decision is left to the authority, and in the strictest sense of the term, such classifications are authoritarian—and each authority may recognize different groups. Thus, removal of the birds from the group "reptiles" is based on a particular viewpoint about the importance of the characters (e.g., feathers) that the groups possess. However, another authority might just as easily argue to remove snakes (they lost those encumbering limbs) or turtles (the ultimate innovation in defensive armor). Decisions to exclude (and thereby elevate) groups such as birds are ultimately based on a view on the importance of particular characters, and such classifications do not accurately convey information about relatedness. Unfortunately, because paraphyletic classifications are based on arbitrary assignments of significance to characters, they are not even useful in the prediction of innovation or levels of anagenetic change (there is no measure of the amount of change involved in the removal of concordant groups, only its significance). The realization that paraphyletic classifications are at their heart arbitrary is the first step toward the acceptance of the need for classifications based solely on monophyly.

A. Ranks and Priority

The need for a useful classification to be based on phylogeny has been confounded in recent years with other aspects of classification, such as ranking and the priority of names. Recently, a movement has emerged that has been self-proclaimed as "Phylogenetic Systematics," which purportively embodies the ideals of Hennig and requires strict translation of phylogenies into named classifications. However, the need for maintaining only monophyletic named groups has been intertwined with other proposals, such as the abandonment of ranks, and proclamations that the existing Linnean system is incapable of accommodating a truly hierarchic system. Of course, the Linnean system is a strictly hierarchic system, and the claims that it must be abandoned are both premature and poorly reasoned. Unfortunately, many outside of this controversy have been misled to believe that the new proposals will provide greater stability in names, which is untrue. Stability in names in any phylogenetic system will come with stability in the underlying phylogenies on which the system of classification is based, and we are on the verge of achieving such stability with the vast quantities of molecular data and improved analytical algorithms that are becoming available.

VII. APPLICATIONS OF PHYLOGENY

A. General Evolutionary Context

In recent years, phylogeny has become an important aspect of many evolutionary and ecological studies. Phylogenetic studies focus on the pattern of history, but provide the foundation for both generalizations and specific conclusions about processes of evolution and diversification. Generalizations about evolutionary processes based on phylogenetic trees are more predictive because phylogenetic trees provide information about character distributions as well as relationships. Phylogenies also provide the ultimate test of particular ideas or theories about the course of evolution. By mapping character distributions in the most parsimonious possible manner on a well-supported phylogeny, it is possible to understand whether characters of interest have evolved once or more than once and whether particular
characters are correlated in a historical sense. Attempts in the past to accomplish these goals were often based on paraphyletic classifications that were misleading or based merely on ranked classifications with unknown or unstated bases. Without a clear understanding of phylogenetic pattern, these studies often miscounted the number of times evolutionary steps occurred within and among taxa and came to incorrect conclusions about character evolution. Precise phylogenies now provide a means to improve such studies.

B. Adaptation
Phylogenetic trees have become a standard tool in the study of adaptation, and such uses are often referred to as the “comparative method.” First, it is necessary to establish that a particular “adaptation” is distributed as an apomorphy within the group in question and then, if there are multiple origins, to determine if these origins are correlated with other characters and/or environmental variables. While numerous statistical approaches have been suggested for such studies, they all assume that multiple independent origins of characters correlated with environmental or historical factors are evidence of adaptation. Indeed, some workers maintain that it is only possible to discuss adaptation in a historical context, i.e., based on explicit phylogenetic trees. Undoubtedly continued work in these areas will result in improved statistical tests for adaptation based on character distributions on phylogenetic trees.

C. Vicariance Biogeography
Phylogenetics has also become the main tool for biogeographic studies, and indeed the field of biogeography is now almost synonymous with vicariance biogeography. Vicariance biogeography is rich in theory and a thorough explanation is beyond the scope of this article. The basic premise is that historical patterns of land connection and movement are recoverable through finding repeated patterns among the phylogenies of various plant and animal groups. Repeated patterns of clade distribution suggest a common history that is interpretable as supporting wider continuous distributions in the past. A well-known example of a broad vicariant pattern of distribution is found in the plant genus Nothofagus (Nothofagaceae) in the Southern Hemisphere, with species distributed in South America, New Guinea, Australia, New Caledonia, and New Zealand and fossils known from Antarctica. Similar patterns in other plants (e.g., Proteaceae) and animals (e.g., ratite birds) can be evaluated in a phylogenetic context to provide evidence of interchange among different “areas of endemism” and develop an “area cladogram” that presents a pattern of relationship among the land areas. These patterns can then be interpreted in the context of plate tectonics as indicators of historical fission of land masses and subsequent plate movements.

D. Phylogeny and Biodiversity: Phylogenetic Diversity
Phylogenies also provide a framework for alternative ways of looking at biodiversity. Most measures of biodiversity use species richness, either in a geographic (either broad or local) or ecologic sense. Other views of diversity may focus on the nature and breadth of adaptation. Such measures unfortunately require a subjective view of the importance of particular adaptations—e.g., the birds might be considered “diverse” because they include so many adaptations for use of the bill. However, phylogeny provides another perspective on biodiversity that allows an objective way to compare uniqueness and diversity of taxa. Although various specific measures of phylogenetic diversity have been proposed, most share a basic approach by which phylogenetic trees are used to evaluate species richness in concordant groups. It means little to say that “orchids are highly speciose” or “monotremes are species depauperate” unless we have some idea as to the relationships of the two groups being compared. Phylogenetic pattern provides the basis for such comparisons.

Currently, the use of phylogenetic diversity measures is largely limited to theoretical discussions and there have been few efforts to actually apply such measures to conservation. This is partly due to the relative paucity of high-quality phylogenies that are available across broad groups of taxa and partly because of a distinctly ecological bias in most studies of biodiversity. As molecular data sources provide better and more complete phylogenies for use by other workers, this is likely to change. It is probable that in the near future measures of phylogenetic diversity will become standard components, in combination with more traditional measures of ecologic uniqueness, species richness, and sensitivity, in the formulae that are used to evaluate conservation priorities for areas and endangered species.

VIII. CONCLUSIONS
An understanding of phylogeny is increasingly important as a basis for experimental design and for interpretation of results in a broad array of biological studies. New methods of analysis, improved computer hardware
and software, and new and improved sources of data, most notably DNA sequences, have revolutionized our ability to construct phylogenetic trees. It is likely that our understanding of the origins and diversification of most major lineages of plants, animals, and microorganisms will improve dramatically over the next century, and we will arrive at a stable and robust reference system for all of life on earth.

See Also the Following Articles

Biodiversity, Evolution and • Cladogenesis • Darwin, Charles • Evolution, Theory Of • Taxonomy, Methods Of • Vicariance • Biogeography

Bibliography


PLANKTON, STATUS AND ROLE OF

C. S. Reynolds
Freshwater Biological Association and NERC Institute of Freshwater Ecology

I. The Structure of Planktonic Communities
II. Habitat Constraints in the Plankton
III. Form, Function, and Selection in the Phytoplankton
IV. Form, Function, and Selection in the Zooplankton
V. Function in the Bacterioplankton
VI. Temporal Patterns in the Organization and Diversity of Planktonic Communities
VII. Mechanisms Promoting and Maintaining Diversity in the Plankton
VIII. Conclusions and Implications

GLOSSARY

autotrophy The ability of organisms to grow and reproduce independently of external sources of organic carbon compounds.
eukaryote An organizational state of cellular organisms in which the genome of the cell is stored in chromosomes enclosed in a membrane-bound nucleus; all protists (algae and protozoa), fungi, plants, and animals are eukaryotes.
euphotic The top layer of a water body through which sufficient light penetrates to support net photosynthetic gain and the growth of photosynthesizing organisms. Rarely more than 100 m in depth, the euphotic layer can be as little as 1 m in turbid waters.
heterotrophy The ability of organisms to grow and reproduce on organic carbon sources, taken in dissolved or particle form.
metazoan Literally, a multicelled animal.
mixotrophy The ability of a normally autotrophic organism to switch, circumstantially, to phagotrophy, or to support an otherwise meager food supply by resorting to the ingestion and assimilation of bacteria or their products.
pelagic The (open-water) part of the aquatic environment that is far from the shore and the bottom bed.
phagotrophy A type of heterotrophy that involves the consumption of protists, plants, or animals as food.
photoautotrophy A type of autotrophy in which organisms gather light energy in order to reduce carbon dioxide to organic carbon, characteristic of green plants, most algae, and some prokaryotes.
prokaryote Organizational state of cells lacking a membrane-bound nucleus and certain other organelles. Bacteria, including the Cyanobacteria, are typically prokaryotic.
picophytoplankton The smallest (<2 μm) size class of photoautotrophic plankton.

"PLANKTON" IS A COLLECTIVE TERM for organisms adapted specifically for a life passed mainly in suspension in the open waters (the pelagic zone) of the sea and of such inland waters as lakes, reservoirs, and rivers. Planktonic organisms include protists (allegedly simple, unicellular, or colony-forming algal primary pro-
The functional definition of plankton, ventured at the introduction to this chapter, has superseded the original, nineteenth-century allusions to plankton “floating” in water. Nevertheless, even this is still unsatisfactory, for its implication that the suspension is either complete or continuous is strictly erroneous. However, genuinely planktonic organisms—which include the planktonlike, chlorophyll-containing primary producers of the phytoplankton, the heterotrophic, decomposer microorganisms of the benthoplankton, and the more animal-like consumers of the zooplankton—are too small (often <20 mm) for their own intrinsic movements to be able to overcome, often or at all, the dispersive effects of water movement. Thus “embedded” within the tireless and unconstrained motion of open water, planktonic organisms broadly go wherever the flow takes them. In this way, the ecology of plankton is inextricably related to the physical properties of the medium, the extent and limits of its motion, and the environmental conditions set within these bounds (Reynolds, 1997b). This situation contrasts with that of most fish and other larger (>20 mm) animals of open water—the “nekton”—whose swimming strength is usually able to overcome normal movement of the water.

The older literature also promulgated a view that suspension in water was necessarily beneficial, supposing water to be something of an ideal habitat. Living in water does confer some notable positive advantages over terrestrial or aerial habitats. These include the mechanical (“Archimedian”) support water provides, as a consequence of its much greater density in comparison with air; its slow temperature fluctuations, as a consequence of its much higher specific heat than air; and its solvent properties, which maintain nutrients and metabolic gases in readily assimilable state.

In truth, however, the planktonic ways of life have evolved to accommodate several problems and drawbacks associated with living in open water. Dominant among these is the issue of turbulence. Water molecules experience strong mutual attraction, which makes the liquid relatively viscous when compared to other fluids. Seeing waves break on the shore, or watching “white” water plunging through riverine rapids, we may be casually impressed by the fluidity of water flow but, without the driving energy, calm is rapidly reestablished as viscosity overcomes the residual motion at the molecular level. What happens is that the introduced mechanical energy is dissipated through a cascade of propagating eddies, of diminishing size and velocity, until molecular attraction imparts order over chaos and the molecular movement is overwhelmed. This behavior is now measurable and it has been mathematically described (see, for instance, Mann and Lazier, 1991). What is of particular interest here is that, depending upon the intensity of persistent wind- or gravitational forcing, viscosity overcomes inertia within the range 0.2 to 3 mm (see Reynolds, 1997b, for examples). This means that the immediate environment experienced by organisms smaller than this (i.e., most of the phytoplankton, bacterioplankton, and the smaller components of the zooplankton) is wholly viscous: far from being fluid, the forces acting on the microorganisms are comparable to those experienced by a human immersed in treacle or unset cement. The organisms do not experience turbulence, neither are their delicate morphologies threatened with physical damage, but they remain entrained in the turbulent field and continue to be randomized throughout its spatial extent. Larger zooplankton (say >0.2 mm), though still too feeble to resist entrainment consistently, are sufficiently tough and flexible to tolerate the millimeter range of turbulence and to exploit it effectively in food gathering (Rothschild and Osborne, 1988).

Beyond the selective constraints imposed by the physical properties of pelagic, open-water environment, it is also necessary to recognize that, with respect to the obligate material components of living cells, the aqueous concentrations of some of these (especially carbon, nitrogen, phosphorus, iron, and fifteen or so micronutrient elements) are often so dilute that their availabilities place a severe constraint on the assembly of planktonic biomass. Moreover, despite its alleged transparency, the absorbance of solar energy by pure water (see, for instance, Kirk, 1994) is such that, at
Among the main groups are often still obscure. Distinctive metazoan phyla, though even there, affinities evasively “safe ground” is reached only among the more activities and their affinities at the molecular level. Relatively among most bacteria other than by their biochemical the cells of eukaryotes, it is still difficult to distinguish robust separation of their prokaryotic organization from and of certain other intracellular organelles provides a pri:

ria: whereas the lack of a membrane-bound nucleus and of certain other intracellular organelles provides a robust separation of their prokaryotic organization from the cells of eukaryotes, it is still difficult to distinguish among most bacteria other than by their biochemical activities and their affinities at the molecular level. Relatively “safe ground” is reached only among the more distinctive metazoan phyla, though even there, affinities among the main groups are often still obscure.

A. Phytoplankton

Taking the capacity for photoautotrophy, the ability to manufacture organic carbon compounds through the photosynthetic reduction of carbon dioxide, to be the sole distinguishing criterion for separating them from other planktonic organisms, the phytoplankton is still extremely diverse. More than 4000 species of marine phytoplankton have been named and described (Sournia et al., 1991). The total number recorded from inland waters is not certainly known, but it is estimated that there are quite 4000 of these as well (Reynolds, 1996). Few genera and still fewer species are common to both fresh and salt waters. Even if a fairly conservative view of their classification is adopted, the species are drawn from at least six distinct protist phyla and at least two major prokaryote subdivisions (see Table I). The Purple (Chromatiaceae) and Green Sulfur Bacteria (Chlorobiaceae) are represented in specialized, anoxic habitats. Of the planktonic genera of Cyanoprykaryotes (former ly classed as Cyanophyceae, or “blue green algae,” and now most commonly referred to as “Cyanobacteria”) most occur in lakes, though several are also common in the low-salinity (<11 parts per thousand) areas of the Baltic. “Sea sawdust” (Trichodesmium spp.) is found in low-latitude sea. The Cyanoprykaryotes are also well represented among the smallest marine and freshwater primary producers (the picophytoplankton: cells <2 μm in diameter; Waterbury et al., 1979). The more conspicuous components of the phyto-plankton of the open sea belong to the Pyrrophyta (including a wide variety of dinoflagellate species) or to the Chrysophytes. This large division is taken to include the large number of diatom species, drawn from one or other of the two main orders (the centric Biddulphi phaeae and the pennate Bacillariophyceae), as well as a divers ity of elaborate, scales-bearing Coccolithophorids. Besides diatoms and Cyanoprykaryotes, the more conspicuous components of the freshwater phytoplankton may be contributed from among the many chlo rophyte and chrysophyte orders. However, cryptomas nads, peridinians, chloromonads, and euagelids can all occur in large numbers at certain locations.

B. Marine Zooplankton

The number of species of represented in the zooplank ton of the sea is considerably enriched by the distinctive dispersal stages of many marine animals that spend their adult lives in the littoral or the benthos. To differing extents, these larvae (the amphipod larvae of sponges, the medusae and ephyrae of the cnidarian coelenterates, the pildia of nemertines, the trochos pheres of polychaetes, the cypess larvae of cirripedes, the phyflosomeae and zoae of the eucrid malacostraca, the veligers of the lamellibranch mollusks, the various auriculariae, bipinnariae, and plutei of the echinoderms and the appendicularian larva of the ascidians; see Table I I ) share the diminutive size ranges, membranous translucence, and feeble swimming movements charac
<table>
<thead>
<tr>
<th>Phytoplankton in Freshwater and Marine Systems</th>
<th>Freshwater phytoplankton</th>
<th>Marine phytoplankton</th>
<th>Freshwater phytoplankton</th>
<th>Marine phytoplankton</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prokaryota</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Division: Anoxyphotobacteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family: Chromatiaceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thiopedia, Thiodictyon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family: Chlorobiaceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlorobium, Pelodictyon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Division: Cyanoprokaryota (&quot;blue-green algae&quot;)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Chroococcales</td>
<td>Synechococcus, Microcystis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synechococcus, Microcystis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Nostoccales</td>
<td>Anabaena, Aphanizomenon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cylindrospermopsis</td>
<td>Glaciebacter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Oscillatoriales</td>
<td>Plantoscoleis, Lamenulaex,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudanabaena, Lyngbya, Phormidium</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eukaryota</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylum: Cryptophyta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Cryptophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryptomonas, Chlotomonas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhodomonas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylum: Pyrrophyta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Dinophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Peridiniales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pennales, Ophryostomata</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glaciebacter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cladophora</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Adinophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Peridiniales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procentrales, Pyrrocystis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylum: Raphidophyta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Raphidophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Chlormonadales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conostomium</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylum: Chrysophyta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Chrysophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Biocellales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biocellales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Chlormonadales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlormonadales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chrysophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Diatomophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Pedinellales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pedinella</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylum: Euglenophyta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Euglenophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Pedinelliales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylum: Chlorophyta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Prasinophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Chroococcales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chroococcales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylum: Euglenophyta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Euglenophyceae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Eugleniales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylum: Dictyochaeta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class: Dictyochaeta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: Didictyochae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE II

**Zooplankton in Marine and Freshwater Systems**

<table>
<thead>
<tr>
<th>Marine zooplankton</th>
<th>Freshwater zooplankton</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phylum: Mastigophora</strong></td>
<td><strong>Class: Branchiopoda</strong></td>
</tr>
<tr>
<td>Bodo, Peranema</td>
<td><strong>Order: Anostraca</strong></td>
</tr>
<tr>
<td>Phylum: Craspedomonadina (Choanoflagellates)</td>
<td>Phylum: Diplodonta (Cladocera)</td>
</tr>
<tr>
<td>Menusa</td>
<td>Diaphanosoma, Keratella, Diaphana, Cladocera, Keratella, Ceratophylax, Moina, Keratella, Brachionus, Daphnia, Beroe, Podon</td>
</tr>
<tr>
<td>Phylum: Rhizopoda</td>
<td><strong>Class: Copepoda</strong></td>
</tr>
<tr>
<td>Order: Amoeboidea</td>
<td><strong>Order: Cyclopoida</strong></td>
</tr>
<tr>
<td>Pelomyxa</td>
<td>Cerephylax, Eurytemora, Ceratophylax, Euphausia</td>
</tr>
<tr>
<td>Order: Radiolaria</td>
<td><strong>Order: Calanoida</strong></td>
</tr>
<tr>
<td>Acanthometra</td>
<td>Carchesium, Eurytemora, Ceratophylax, Euphausia</td>
</tr>
<tr>
<td>Phylum: Ciliophora</td>
<td><strong>Order: Pseudocyclopoida</strong></td>
</tr>
<tr>
<td>Order: Holotricha</td>
<td><strong>Order: Copepoda</strong></td>
</tr>
<tr>
<td>Colpoda, Nassula, Colpoda</td>
<td><strong>Order: Cyclopoida</strong></td>
</tr>
<tr>
<td>Order: Spirotricha</td>
<td>Carchesium, Eurytemora, Ceratophylax, Euphausia</td>
</tr>
<tr>
<td>Euplotes, Caenomorpha, Halteria, Metopus, Strombidium, Tintinnidium</td>
<td><strong>Order: Pseudocyclopoida</strong></td>
</tr>
<tr>
<td>Order: Peritricha</td>
<td><strong>Order: Calanoida</strong></td>
</tr>
<tr>
<td>Convolvulina, Epistylis</td>
<td>Carchesium, Eurytemora, Ceratophylax, Euphausia</td>
</tr>
<tr>
<td>Phylum: Porifera</td>
<td><strong>Phylum: Annelida</strong></td>
</tr>
<tr>
<td>[Amphiblastula larvae]</td>
<td><strong>Class: Polychaeta</strong></td>
</tr>
<tr>
<td>Phylum: Coelenterata</td>
<td><strong>Order: Tomopteris</strong></td>
</tr>
<tr>
<td>Subphylum: Cnidaria</td>
<td><strong>[trochosphere larvae]</strong></td>
</tr>
<tr>
<td>Order: Leptomedusae</td>
<td>Phylum: Platyhelminthes</td>
</tr>
<tr>
<td>Obelia</td>
<td><strong>Class: Polychaeta</strong></td>
</tr>
<tr>
<td>Order: Anthomedusae</td>
<td><strong>Class: Turbellaria</strong></td>
</tr>
<tr>
<td>Hydractinia</td>
<td>Convoluta, Microstomum</td>
</tr>
<tr>
<td>Order: Trachylina</td>
<td>Phylum: Nemertea</td>
</tr>
<tr>
<td>Limnocnida, Craspedacusta</td>
<td>[Pilidium larvae of certain Hoplonemertines]</td>
</tr>
<tr>
<td>Order: Siphonophora</td>
<td>Phylum: Nematoda</td>
</tr>
<tr>
<td>Velella, Physalia</td>
<td>[A few shelf-water species are described]</td>
</tr>
<tr>
<td>Phylum: Phyllophora</td>
<td>Phylum: Rotatoria</td>
</tr>
<tr>
<td>Velella, Physalia</td>
<td><strong>Order: Monogononta</strong></td>
</tr>
<tr>
<td>Subphylum: Scyphozoa</td>
<td>(Suborder: Flosculariacea)</td>
</tr>
<tr>
<td>A. Aurelia, Cyanea, Pelagia</td>
<td>Flinia, Conochilus, Brachionus, Kiefferia, Kellicottia, Synchaeta, Aplanchna</td>
</tr>
<tr>
<td>Subphylum: Ctenophora</td>
<td><strong>Phylum: Gastrotricha</strong></td>
</tr>
<tr>
<td>Class: Tentaculata</td>
<td>[Encountered in plankton of small water bodies]</td>
</tr>
<tr>
<td>Pleurobranchia</td>
<td><strong>Phylum: Annelida</strong></td>
</tr>
<tr>
<td>Class: Nuda</td>
<td><strong>Class: Polychaeta</strong></td>
</tr>
<tr>
<td>Beroe</td>
<td><strong>Order: Amphipoda</strong></td>
</tr>
<tr>
<td>Phylum: Phylum: Platyhelminthes</td>
<td><strong>Order: Eucarida</strong></td>
</tr>
<tr>
<td>Class: Turbellaria</td>
<td>(Suborder: Euphausiacea)</td>
</tr>
<tr>
<td>Convoluta, Microstomum</td>
<td><strong>Phylum: Arthropoda</strong></td>
</tr>
<tr>
<td>Phylum: Nemertes</td>
<td><strong>Class: Insecta</strong></td>
</tr>
<tr>
<td>[A few shallow-water species are described]</td>
<td><strong>[all larvae]</strong></td>
</tr>
<tr>
<td>Phylum: Phylum: Phyllophora</td>
<td><strong>Order: Megabactra</strong></td>
</tr>
<tr>
<td>Class: Turbellaria</td>
<td><strong>Phylum: Annelida</strong></td>
</tr>
<tr>
<td>Convoluta, Microstomum</td>
<td><strong>Order: Diptera</strong></td>
</tr>
<tr>
<td><strong>Phylum: Copepoda</strong></td>
<td><strong>Order: Diptera</strong></td>
</tr>
<tr>
<td><strong>Order: Cyclopoidea</strong></td>
<td><strong>Order: Schizophera</strong></td>
</tr>
<tr>
<td>Oithona, Mesocyclops</td>
<td><strong>Phylum: Mollusca</strong></td>
</tr>
<tr>
<td>Chondracanthus</td>
<td><strong>Class: Gastropoda</strong></td>
</tr>
<tr>
<td>Phylum: Phyllophora</td>
<td><strong>Order: Opisthobranchia</strong></td>
</tr>
<tr>
<td>Class: Branchiopoda</td>
<td><strong>Order: Euthecoidea</strong></td>
</tr>
<tr>
<td><strong>Order: Anostraca</strong></td>
<td><strong>Class: Branchiopoda</strong></td>
</tr>
<tr>
<td>Sialis</td>
<td><strong>Order: Arachnomorpha</strong></td>
</tr>
<tr>
<td>Pontomyia</td>
<td><strong>Order: Thoracica</strong></td>
</tr>
<tr>
<td>Chaoborus</td>
<td><strong>Order: Calanoida</strong></td>
</tr>
<tr>
<td>Phylum: Phyllophora</td>
<td><strong>Class: Copepoda</strong></td>
</tr>
<tr>
<td>Class: Branchiopoda</td>
<td><strong>Order: Oligostraca</strong></td>
</tr>
<tr>
<td><strong>Order: Calanoida</strong></td>
<td><strong>Order: Calanoida</strong></td>
</tr>
<tr>
<td>Calanus, Temora, Eudiaptomus, Eurytemora, Centropages</td>
<td><strong>Order: Copepoda</strong></td>
</tr>
<tr>
<td><strong>Class: Branchiopoda</strong></td>
<td><strong>Order: Copepoda</strong></td>
</tr>
<tr>
<td><strong>Class: Branchiopoda</strong></td>
<td><strong>Order: Copepoda</strong></td>
</tr>
<tr>
<td><strong>Class: Branchiopoda</strong></td>
<td><strong>Order: Copepoda</strong></td>
</tr>
<tr>
<td><strong>Class: Branchiopoda</strong></td>
<td><strong>Order: Copepoda</strong></td>
</tr>
<tr>
<td><strong>Class: Branchiopoda</strong></td>
<td><strong>Order: Copepoda</strong></td>
</tr>
<tr>
<td><strong>Class: Branchiopoda</strong></td>
<td><strong>Order: Copepoda</strong></td>
</tr>
</tbody>
</table>
teristic of the species which are planktonic throughout their lives. Among the smallest (<20 μm) examples of these heterotrophic protists are the nanoflagellates (some being closely allied to the photautotrophic phytoflagellates, classified here as phytoplankton) and choanoflagellates. The microzooplankton fraction (in the size range, 20–200 μm) includes the rhizopod foraminifers and radiolarians, and a range of ciliate and suctorian ciliphorans. More conspicuous (0.2–20 mm) in the marine zooplankton are the ctenophoran ‘comb jellies’ and sea-agoeberryes, the chaetognath ‘arrow worms’ (e.g., Sagitta), some specialized turbellarians (e.g., Convoluta, Microstomum) and polychaetes (e.g., Tomopteris), certain opisthobranch gastropods (such as Clione, Limacina), and the larvaceans (e.g., Oikopleura) and salps (e.g., Delichon).

The most prominent of the animals of the marine plankton, however, are crustaceans. Most of the species are copepods, including calanoids, like Calanus, Temora, and Centropages, and cyclooids, such as Oithona. There are also some representative cladoceran (e.g., Eudiadne, Podon) and ostracod (such as Gigantocypris) genera. The malacostracan orders are distinctively represented by organisms that remain planktonic in their adult stages. Mysidaeans (such as Leptomysis) are locally abundant and a few cumaceans (e.g., Dactylis) are remarkable for being nocturnally planktonic and diurnally benthic in coastal waters. Amphipods (e.g., Apherusa) and isopods (e.g., Eurydice) have planktonic representatives. Perhaps the most renowned component of the plankton of the high-latitude oceans, mainly for being one of the major food sources of the filter-feeding whales, is krill (Euphausia spp.). Elsewhere, the smaller euphausids form an important food for pelagic fish. Pyrosomes are one of the common genera in European coastal waters. Some of the ctenophalan jellyfish may be regarded as being the largest animals in the plankton (>20 mm, perhaps to 2 m). Though some of these are rather larger than some of the swimming organisms (‘nekton’ fish, cephalopods) that are excluded from the understanding of ‘plankton’ (discussed earlier), the large jellyfish qualify for their poor ability to control their own movements in the sea. The siphonophores, like Velella and the Portuguese man o’war, Physusis physalis, are little more than drifting “polyp colonies.” The true jellyfish, which move themselves by slow, rhythmic pulsation of the umbrella-like manubrium, include the distinctive Aurelia, Cyanophora, and Pelagia. Finally, the young stages of several species of pelagic fish are of such diminutive size and swim so feebly and with weakness of motility that, for the first part of their lives, they are reasonably included among the plankton.

C. Freshwater Zooplankton

The phyletic representation in the zooplankton of lakes is well known to be relatively much poorer than in the sea, but mostly this reflects the poorer representation of the animal phyla among fresh waters anyway. The summary in Table II shows considerable coincidence of representation at the higher phylogenetic levels but, as with phytoplankton, there is almost no commonality of species. Protists usually figure prominently in the plankton of inland waters, supposedly as a function of the organic matter available, including detrital particles and their associated bacteria. The smaller heterotrophic flagellates generally consume free-living bacteria, but many of the common planktonic microciliates, like Collops and Tintinnidium, feed also on small algae and flagellates. In doing so, they fulfill an important linkage in the so-called microbial loop (Azam et al., 1983): in many systems, the excretion of excess organic carbon fixed in photosynthesis, its assimilation by bacteria and its successive transfer through closely coupled flagellate-climate consortial linkages to copepod consumers and, eventually, planktivorous fish, demonstrably exceeds the carbon transfers along the conventional phytoplankton → zooplankton → fish trophic pathway.

Occasionally, the larger Amoebae and ciliates, such as Nassula, may dominate the zooplankton; this may owe to the prevalence of a particular food source, or to the fact that a crustacean plankton has not yet developed, or that some other factor (low oxygen concentration, for example) restrains potential competitors for the same food resource.

At least two freshwater coelenterate genera have planktonic medusae. The Gastrotrich phylum of mniute, wormlike but unsegmented animals is occasionally represented by planktonic specimens. The phylogenetically close phylum of the Rotatoria is prominently represented in the freshwater plankton by some two dozen genera, drawn mainly from the Order Monogononta. The most ecologically important genera include Asplanchna, Brachionus, Filinia, Keratella, Kellicottia, and Synchaeta. Some colonial rotifers also feature in the plankton (e.g., Conochilus). Some are specialist feeders; most of those mentioned browse or filter-feed on bacteria, detritus, and planktonic algae, generally within species-specific size ranges (Pourriot, 1977).

As in the sea, the most prominent planktonic animals
are crustaceans. The Branchiopods are represented by the anostracan "brine shrimps" (e.g., *Chirocephalus*), especially in temporary waters, and by the Cladocera, the familiar "water fleas." This grouping (its monophyletic origin is now doubted) includes the mainly herbivorous, filter-feeding species of *Daphnia*, *Cyclops*, *Moina*, *Simocephalus*, *Bosmina*, and *Halydus*, and the predatory *Bythotrephes* and *Leptodora*. Planktonic *On- troachoids* are noted from lakes in Southeast Asia (*Cypria javensis*) and from Laguna de Prerito, Guatemala (*Cypria petenensis*). The Copepods are particularly well represented by calanoids (such as *Eudiaptomus*, *Eurytemora*, *Boreocheile*) and by cyclopoids (e.g., *Mesocyclops*); some parasitic cyclopoids (e.g., *Argulus*) are dispersed through the plankton. The Branchiurans (e.g., *Argulus*), which are ectoparasitic on open-water fish, are certainly to be considered essentially planktonic. Though principally marine, the mysids are represented in the plankton of several high-latitude lake systems, where Myysis is regarded as a relict from the last glaciation. The Amphipod flag is carried by *Macrohectopus*, which is endemic to the plankton of Baykal Lake.

Among the arthropod insects, several genera of Dip- tera have larvae, which pass most of their time in the plankton. The most specialized of these are the juvenile chaoborines (e.g., *Chaoborus*) or phantom midges, whose transparent bodies reveal the internal provision of buoyancy-providing air sacs. The dispersal stages of aquatic larvae of other orders of insects sometimes show adaptations that are unmistakably planktonic: a striking instance is provided by the first instars of *Sialis* (*Mega- loptera*, see Elliott, 1996).

### D. Explaining Species Diversity

The foregoing passages confirm the richness of phyletic representation and the very large number of individual species that collectively contribute to the overall diver- sity of planktonic organisms. The challenge now is to account for the richness of the planktonic species and to explain the mechanisms by which it is maintained. At one level, we might be surprised that this problem arises at all. It should not be at all unreasonable to anticipate that, in a supposedly fluid and isotropic medium, fully accessible to suitably adapted species, Dar- winian selection should move the structure of the orga- nismic assemblage toward just a small number of specialists, each being the best-fit survivor in its key community role. Every less-fit competitor might be sup- posed to suffer progressive exclusion by the stronger species; overall, diversity should be suppressed. Thus, to confront the remarkable richness of planktonic plant- and animal-species surveyed previously is counter- intuitive and paradoxical (Hutchinson, 1961).

Neither is it clear for how long such richness has distinguished the biota of open waters. Owing to the facts that most microorganisms do not form robust fossils, that given freshwater habitats are, in geological terms, very transient features, and that, at the relevant evolutionary scales, even the present ocean floors are relatively young, it is difficult to be categorical about the origins of planktonic communities. However, they are likely to be old. By the beginning of the Cambrian period, some 600 million years before present, when some of the oldest fossiliferous sedimentary rocks were formed, most of the invertebrate phyla represented in modern plankton had already appeared. These were wholly aquatic, though not necessarily planktonic. The Cyanobacteria had been established long before this, with primitive oxygenic photosynthesis cocoid species already converting the reducing conditions of the early planetary environment to an oxic one, between 2 and 2.5 billion years ago. Many of the Protistan groups, including several of the eukaryotic algae, had also ap- peared by the Cambrian (Ragan and Chapman, 1978). Presumably, some of these were free living, in suspension in the water. The step to a truly planktonic exist- ence is supposed to be short and, with such a diverse phylogeny of modern plankton, it is reasonable deduc- tion that it was taken within each evolving group, per- haps several times. It is probable that the first planktonic communities began to come into existence about a bil- lion years ago, when many new opportunities for func- tional specialisms and the adaptive radiation of species were available.

Interestingly, the various Chrysophyte groups are considerably younger, there being no undoubted re- cords from before the Mesozoic period (Tappan, 1980). Both the diatoms and the chromulines expanded and diverged during the Cretaceous era (135–65 million years ago); the coccolithophorid *Haptophyceae* also ap- pear to have originated during the Mesozoic; it is their biominalized remains which predominate the chalk deposits that have lent their name to the Cretaceous. Some very significant changes in the chemical com- position of the sea must have occurred during that period.

While the same species may not have held sway throughout, or even for the past 100 million years, it is clear that the factors favoring a high planktonic biodiversity are recurrent, if not ongoing. Many theories have been advanced to explain Hutchinson’s (1961) paradox of the plankton, but the conundrum for long remained unsolved. Either the number of occupiable niches must be far more numerous than had been sup-
posed or the competition was somehow incomplete in its effects. Hutchinson himself suspected that the assumption of a homogeneous environment with steady-state properties was flawed. Collectively, the environments that planktonic populations inhabit are subject to huge variability in their chemical makeup and in their physical characters. They may be physically separated from each other (as are freshwater catchments), enjoying quite different climates or, even if contiguous (like the seas), their systems may be close to mutual isolation by currents and circulations. On the continents, lake basins are created and destroyed at, relative to evolutionary rates, high frequency. The skeletal understanding of the history of the world's oceans is that they have undergone large oscillations in metabolism and productivity associated with changes in the biospheric carbon cycles (Thierstein, 1989). Within the limits of habitat suitability that evolutionary specialization allows, the potential ranges of individual species should be distinct from those that have evolved separate specialisms. Superimposed on the longer-term changes are relatively higher frequencies of periodic forcing. Within these smaller habitat units, the higher frequency environmental oscillations might lead to alternations in the favored specialisms and consequent interspecific transfers of competitive advantage. Physical limitation of range (endemism) or otherwise (cosmopolitanism) would then represent the two extremes of dispersal efficiency across physical barriers and habitat preferences.

Current understanding of plankton ecology conforms to this general view in that it is certainly possible for the experienced observer to determine the broad habitat provenance of given planktonic assemblages: their species composition has an indicative value because similar assemblages characterize similar pieces of water. Students of the plankton have usually accepted a prevailing view that planktonic species are generally cosmopolitan and disperse freely, so that, on balance, most are able to establish quickly wherever suitable conditions arise. Dispersal mechanisms among the protists and prokaryotes are certainly effective (review of Kristiansen, 1996), and it is true that many conventionally identified morphotypes enjoy worldwide distributions. As more molecular information becomes available, the cosmopolitan nature of plankton is called increasingly into question. Besides, the notion that “everything is everywhere” is surely a matter of degree: ease and frequency of dispersal is not equal among planktonic species. Isolation and regional endemism certainly occur (Tyler, 1996), especially where long physical distances or significant, hostile barriers separate sites of suitable habitat. It often takes anthropogenic intervention to bridge these gaps, as the recent “breakout” of the lamellibranch, Dreissena polymorpha, and hitherto endemic mysids from their original Caspian-basin locations into the waterways of Europe and, now, North America graphically reminds us. Equally remarkable is the arrival of the first freshwater Cladocera on Easter Island in 1780, which had failed to bridge the distance from the next nearest lake (over 3000 km distant) prior to Captain James Cook's requirement to replenish his ship’s supply of drinking water (Dumont, 1990).

The same student of plankton will also be well attuned to the seasonality of the species composition of samples, as dominance moves frequently among the species of the assemblage. The differential responses of individual species of plankton to temperature or day length or nutrient resources are generally recognized to underpin what is perhaps the most familiar feature of planktonic communities—the so-called seasonal succession. As with other aspects of plankton ecology, the identity of its driving variables has been actively pursued and described by conceptual word models; the most compelling of these has been the PEG-model of Sommer et al. (1986): Like its less successful contemporaries, it is nevertheless founded on an implicit acceptance of the differences among planktonic species and the suitability of species-specific adaptations to particular habitat constraints. The successful pursuit of viable explanations for the origins of biodiversity and the mechanisms of its maintenance must acknowledge a distinction between “local diversity,” a measurable property of the temporal compositional fluctuations in response to local, within-patch variability, and the total species richness, or overall biological diversity, which is supported by the aggregate of patches and its continued ability to offer an adequate number of accessible habitats to satisfy the dynamics of survival of each species. The way to explain planktonic biodiversity is through a simultaneous recognition of the variables constraining habitat suitability and the adaptive specialisms and limitations of species for which the habitat constraints will select and those against which they will discriminate.

II. HABITAT CONSTRAINTS IN THE PLANKTON

The assumption of a uniform, hospitable, steady state is erroneous so far as most open-water habitats are
concerned. On the contrary, many are as hostile as their desert-like barrenness conveys. Relative infertility may be locally or regionally attributable to an inadequacy of light energy to sustain planktonic primary production and the import of organic carbon is too modest to support the heterotrophs. Light income may be scarce for reasons of latitude (short day length, low solar declination), or it may be severely “diluted” over a deep mixed layer, or its penetration may be abruptly curtailed by a high concentration of inert particles (turbidity). Commonly, the concentrations of assimilable sources of certain primary nutrients place a very low ceiling on the biomass that can be assembled (nitrogen, phosphorus, and iron are cited most frequently as capacity-limiting factors). Chronic nutrient deficiency also interferes with the production of microbial biomass, even supposing there to be an adequate flux of organic carbon. For the planktonic animals, the problem is to be able to forage sufficient food; the concentrations and the size distributions of potential food particles influence profoundly the nature and abundance of the consumer populations. A dearth of primary-producer plant biomass is no basis for the intense production of zooplankton, neither will it underpin large harvestable crops of prime pelagic fish. What we find is that the locations where significant net primary production is possible and that the occasions when it is exploitable by heterotrophic consumption are, in fact, strongly circumscribed.

Besides the spatial differentiation of planktonic habitats, there is usually a considerable temporal variability. Conspicuous are the seasonal changes consequent on latitude: lengthening spring days and higher flux densities are generic to all temperate ecosystems but for small organisms with short life spans, the changes are perceived by successive generations and not as some perennial amplitude of fluctuation registered by a forest tree, for example. Ambient temperature may increase too, incidentally raising the threshold of wind energy required to keep the water fully mixed. The onset of thermal stratification, as a consequence of a shifting balance between the buoyancy forces brought through surface warming and the kinetic dissipation of the work of the wind energy, precipitates a train of environmental effects, leading to enhancement of the segregation of the warm, insolated, aerated, and increasingly resource-depleted epilimnion from a colder, darker, and potentially less oxidizing hypolimnion. Neither is such seasonality confined to high latitudes: even quite small differences in wind prevalence, cloud cover, humidity, and hydraulic exchange result in seasonal variations in thermal stability and down mixing in the tropics (for examples, see Talling and Lemoalle, 1998).

So it is that the different kinds of water body (lake or pond, river, estuary, coastal shelf, upwelling, or the open oceans) each offer, to the most appropriately adapted species, or to those simply furnishing the largest inocula, varied and sometimes very transient opportunities to build their local populations. Furthermore, because the medium is fluid and the movement induced by wind or gravity is subject to fluctuations in strength, the persistence and vertical extent of an upper, differentiated layer is a highly variable character of the environment, which alters not just from season to season but from day to day and from hour to hour. Interest is also growing in the effect of year-to-year variations. At all these scales, the intensity and frequency of the environmental variability determine and, potentially, modify the critical habitat constraint and, hence, the attributes of organisms most likely to benefit dynamic performance.

It follows that selective advantage is likely to move among species. Environmental variability is a powerful influence on the assembly of planktonic communities and thus on their biologic diversity. The causal linkages between observable patterns and processes in the maintenance of species diversity in the plankton may be usefully explored through a diagrammatic representation of the two principal variables characterizing particular open-water environments, namely, the fluctuating fluidity or vertical extent of the uppermost water layer and the availability of the resources to support the assembly of biomass. The layout of Fig. 1 is descended from Margalef’s (1978) original scheme for marine phytoplankton, in which a “nutrient” axis is set against one of “turbulence”; the arrangement is amenable to tracking habitat variability and the changing species composition through time. The scheme has been developed for fresh waters to accommodate habitat conditions in which nutrient resources and the light availability in the surface-mixed layer are sufficient to sustain the fastest in situ algal growth rates (in the top left corner of Fig. 1) and to distinguish them from conditions that become either increasingly resource-constrained (moving downward in the matrix), or increasingly energy-deprived (moving rightward), or fall deficient in both (bottom right corner). The selection of given species or groups of species was found broadly to be correlated with the matrix space thus represented; high-performance, fast-growing algae would be favored in the relative paradise of the upper left-hand corner; algae favored by the second and third contingencies show some specialist adaptations, respectively, for op-
erating under conditions of low concentrations and remote sources of nutrients, or for harvesting light from low irradiances or from infrequent short periods in the light field. The fourth, low-nutrient, low-light contingency is unfilled and is something of a phycological desert, evidently untenable as a habitat for a photoautotroph (Reynolds, 1988b).

It was never imagined that such a two-dimensional matrix could describe the entire spectrum of freshwater variability, but it has proved to be a consistent template for predicting the structural responses of planktonic assemblages to the main dimensions of environmental variability. Latterly, effort has been invested to improve both the utility and the quantification of the axes. The horizontal axis in Fig. 1 describes the integral, $I^*$, which, as the product of the physical depth of mixing and the vertical attenuation of a finite input of light energy, is sensitive to the constraints both of deep mixing and of high turbidity; quantities diminish rightward on a logarithmic scale. The vertical axis in Fig. 1 accommodates another composite scale, $K^*$, being the quotient of the concentration of the critical nutrient in the medium (usually the surface mixed layer) and the gradient of concentration between the top and bottom of the entire trophogenic layer. This scheme differentiates sites that are chronically deficient in nutrients (points plotted well down the $K^*$ axis) from, on the one hand, those in which steep gradients develop as a consequence of near-surface uptake (points track from high to low on the $K^*$ axis) and, on the other, those in which the nutrient is scarcely exhausted (points located consistently toward the top of the $K^*$ axis).

Against these axes, it is possible to characterize the “signatures” of seasonal changes in various kinds of water body, including of small, “eutrophic” lakes, in which nutrient availability is seasonally reduced, of shallow, fertile systems, in which the turbidity is the predominant variable, and of deeper, oligotrophic systems, where the depth of mixing is the strongest seasonal variable. An analogous (and in many ways, a more self-evident) approach has been developed recently for the sea (Fig. 2). This distinguishes energy-limited, well-mixed, high-latitude oceans from the highly stratified, nutrient-segregated waters of the tropics and will represent processes and interactions in coastal and shelf waters, including frontal zones and major upwellings of deep circulation currents.

The premise to be developed is that the biodiversity of plankton may be fitted to such templates and, moreover, that environmental change and variability select for alternative species. Reverberations in the selection

\[ figure 1 \]

Two-dimensional representation of limnetic habitats, which, starting in the top left, distinguishes well-insolated, resource-replete water masses from (working down) long-stratified, resource-segregated water columns and from (working across) increasingly light-deficient environments.
III. FORM, FUNCTION, AND SELECTION IN THE PHYTOPLANKTON

A. The Basic Adaptations of Planktonic Photoautotrophs

The essential environmental requirements of planktonic photoautotrophs, to be able to grow in numbers and increase in total biomass and to be able to perpetuate and disperse their genes, do not differ fundamentally from those of plants in other ecosystems: access to water, exposure to adequate levels of photosynthetically active wavelengths of light, a source of assimilable carbon and adequate supplies of each of a score of other elements. Problems over the adequacy of the water supply to phytoplankton may be safely discounted (water relations can be adversely affected by salinity changes, where these result in the loss of cell water to the medium by osmosis). The main macronutrients (C, H, N, O, P, S, Na, K, Mg) involved in the synthesis of proteins, of cell protoplasm, and of the various organelles (all of which have to be replicated at each cell division), the key micronutrients (Fe, Mn, Mo, Cu, Zn, B, Va) mediating the assembly processes, and the elements (Ca, Si) condensed in the elaboration of calcareous and siliceous skeletal biominerals must be drawn into the cell from the bathing medium, mostly against significant concentration gradients. Elemental carbon, which constitutes nearly 50% of the dry mass of protoplasm, is usually distinguished as a separate resource to terrestrial plants.
because of the atmospheric source of carbon dioxide; for phytoplankton, the proximal source of carbon dioxide is that which is dissolved in the water and which, once again, has to be drawn into the cell against a steep concentration gradient, itself sometimes exacerbated by high organic demand. So it is, in the dilute world of the plankton, that diffusion alone is rarely able to supply the resource requirements of plankton. Adaptive mechanisms for gathering chronically deficient resources include the enhancement of uptake affinity and of the maximization of storage. Mechanisms enhancing access to remote reserves or for exploiting alternative sources of nutrients are valuable against a depleting resource base.

The biochemistry of the assembly processes is not the primary concern of the student of biodiversity, but it is important to understand the essence of the metabolic machinery. Every activity, from resource uptake, through internal transport, protein synthesis, organelle assembly, and operation, to the replication of new cells, requires the controlled expenditure of energy, which, as in just about all living things, is supplied by the respirational oxidation of carbohydrates. Fulfilling this basal metabolic requirement constitutes an energetic maintenance cost to the organism. In the case of the heterotroph, additional energy is consumed in the foraging of foods—the essential requirement is that the investment of energy yields a greater return of potential energy in the organic carbon thus derived. The distinguishing feature of the photosynthetic apparatus is not of primary concern here, but we should recognize that what constitutes an adequacy of exposure to photosynthetically active light depends partly on the efficiency of its interception and photochemical conversion to carbohydrate and partly on the fraction of photosynthate that is required to satisfy the basal metabolic costs of its operation.

These criteria assume considerable ecological relevance to phytoplankton where, because of the sharp underwater attenuation of downwelling sunlight, only the near-surface layer is readily supportive of net primary production. Even then, seasonal, diurnal, and weather-determined fluctuations in the intensity and duration of incident sunshine means that the spatial and temporal extent of this euphotic layer is highly variable. To operate habitually at more rarefied light levels clearly requires the provision of more, or more efficient, light-harvesting capacity. Moreover, because it is frequently the case that the depth of the wind-mixed layer extends beyond the euphotic layer, convective entrainment of phytoplankton often diminishes its aggregate exposure to the light field, at times leaving cells unable even to compensate their basic metabolic energy requirement. Here, too, great adaptive importance attaches to the sufficiency of the biomass-specific light-harvesting provision and its presentation to the light field. At the same time, the planktonic photoautotroph still needs a suite of biochemical and physiological defences to deal with the risk of overexposure of enhanced light-harvesting that the extreme variability that pelagic open water habitats present.

**B. Specializations among Planktonic Photoautotrophs**

Because the fundamental requirements of (allegedly) simple plants are more readily measurable than are those of more complex terrestrial ones, it is possible to quantify some of the generalized assertions about the selective value of species-specific adaptations to meet the stress of deficiencies in the supply of resources and the processing constraints set by operating in a fluctuating environment. Starting with the quantum yield of photosynthesis, the theoretical output of 1 mol of reduced carbon for the investment of 8 mol photons of photosynthetically active radiation harvested—or 0.125 mol C (mol photon PAR)$^{-1}$—is demonstrably approached in experiments with phytoplankton; typical measurements are 0.07–0.09 mol C (mol photon PAR)$^{-1}$ (Bannister and Weidemann, 1984). Over the temperature range, 10–30°C, and at the order of light intensities required to saturate chlorophyll-specific photosynthesis in phytoplankton (50–200 mol photon PAR m$^{-2}$ s$^{-1}$), photosynthetic yields are found to fall in the range, 2–15 g C (g chlorophyll)$^{-1}$ h$^{-1}$. The carbon fixation rate provides the capacity for assembling new producer biomass. Taking an average 90:1 carbon-to-chlorophyll ratio, exponential cell-specific growth capacities of 0.04 to 0.26 h$^{-1}$ may be deduced, which potentially might support biomass doublings from one every 17 h to one every 3 h. The maximum cell replication rates in laboratory measurements on the unicellular chlorophyte, *Chlorella*, one of the fastest-growing freshwater eukaryotes, are equivalent to biomass doublings every 19–4 h over the same temperature range (see Reynolds, 1997b), comfortably within the theoretical
carbon assimilation capacity. The shortfall is mainly attributable to consumption of fixed carbon during growth (respiration).

For such rates of cell replication to be realized and sustained, not only must the light levels and the carbon dioxide supply be upheld, but all the other chemical components resources have to be available in concentrations that will also saturate the uptake demand. For instance, the widely accepted stoichiometry of Redfield (1938) leads us to suppose that for every 42 mg C incorporated, the new generation would, ideally, also require roughly 1 g P and 7 g N if the normal cell stoichiometry was to be preserved. Taking the amount of new carbon dioxide that could migrate across the water surface to balance the consumption by primary producers (about 100 g C m\(^{-2}\) year\(^{-1}\); higher areal productivity in lakes and coastal waters are reliant on imports dissolved in inflows and the reuse of respiratory carbon dioxide), the potential new production also requires the supply of some 2.4 g P and 14 g N m\(^{-2}\) year\(^{-1}\). Moreover, the minimum direct PAR investment in biomass assembly would be not less than 100 mol photons m\(^{-2}\) year\(^{-1}\) (equivalent to a mean underwater photon flux of 10–20 μmol photon m\(^{-2}\) s\(^{-1}\) Reynolds, 1997b).

Such theoretical calculations of the stoichiometric capacities provided by the fluxes of each main resource are helpful to the identification of which of them is most likely to set the constraint on growth generally and, hence, that to variations in which will evoke the most sensitive changes in growth rate and supportable biomass. In this way, the factor group most likely to limit maximal attainment is revealed. In many kinds of lake and seas, the total annual area-specific loads (including recycled nutrient) of either nitrogen or phosphorus or both fall far short of the respective hypothesized saturation requirements of 14 g N m\(^{-2}\) and 2.4 g P m\(^{-2}\). So indeed the typically much lower standing crops of planktonic biomass are constrained by the rates of supply and reuse of critical nutrients. On the other hand, even in very clear water and under blue skies for 12 h per day, the depth of the surface mixed layer and its impact on the insolation of fully entrained algae continue to impose a severe capacity limitation on planktonic primary production (Reynolds, 1997b: Fig. 45). Cold, oligotrophic seas and lakes are not exclusively nutrient limited. However, it is where nutrients are abundant, or are rapidly cycled in a shallow mixed layer, that we expect to see a level of phytoplankton sufficient to challenge the carbon invasion rates, with an attendant rise in pH. In small lakes, internal recycling (carbon dioxide from community respiration, plus the carbon dioxide supplied in solution in inflowing streams) helps to offset the effects of capacity limitation by carbon.

The distribution of metabolic limitations can be fitted to the I**-vs-K** ordination of planktonic habitats (Figs. 1 and 2). In the energy- and resource-replete environments represented toward their upper left corners, the selective advantage resides with species adapted to the maximization of the opportunities for rapid resource exploitation and conversion to new biomass. Away from these areas, specialisms permitting continued operation under markedly subideal conditions are selected. In the downward direction, the most useful of these provide an advantageous measure of tolerance of resource stress—including the abilities to conserve nutrient resources, to search for them more efficiently, and to tap into other sources of resource. In the rightward direction, the derivation of harvestable energy is progressively more interrupted and truncated: adaptations for dealing with the disturbance to the processing ability—including the enhanced ability to capture photons from poor or fluctuating light fields—are increasingly demanded.

It is possible to show consistent morphological and adaptive traits among phytoplankton species that are variously exploitative, disturbance tolerant, and stress tolerant. It is also possible to quantify the impacts of adaptation and to demonstrate the satisfying coherence among the form, function, and ecology of planktonic algae (Reynolds, 1988b, 1995). For example, the species that appear in “new water bodies” (from rain puddles to tidal pools), in hydraulically (broad plain lakes after the river drops back) or seasonally refreshed systems (lakes at the onset of thermal stratification) are typically exploitative: they are characterized by ready and effective dispersal mechanisms and by a facility for rapid growth (>10 \(\times\) 10\(^{-3}\) s\(^{-1}\) at 20°C). Typical freshwater representatives include such algae as Chlorella, Anknya, Kellella, Chlamydomonas, Rhodomonas, Chrysochromulina, Monochrysis, Monodus, and Synecococcus. They are unicellular or form small coenobia (generally <10\(^3\) μm\(^3\)), offering a surface/volume ratio (>0.5 μm\(^{-1}\) favorable to rapid solute exchange and nutrient assimilation. The area (e.g., of light field) projected by the algal cell mass, usually a mark of its potential efficiency as a light-gathering antenna, is equivalent to at least 6.5 m\(^2\) (mol cell C)\(^{-1}\). Reynolds (1995) proposed that these traits were indicative of an “invasive” or, in the terminology of Grime (1979), C-type life-history strategy.

The adaptations that help species to survive developing nutrient stress include the physiological flexibility to overcome deficiencies in the supply of carbon
specific nutrient uptake and growth (fenestrated coenobia, protuberances, etc.) and so con-
by morphological complexity (filaments, attenuated or are not necessarily attained through small unit-size but
C)/H11022 to-volume ratios (generally of a large mass-specific surface area (this scale of pelagic disturbance through the projection to light. ''Attuning'' (Reynolds, 1995)—strategists tends to be restricted mainly to stress-tolerant—S-type, or (''acquisitive''; Reynolds, 1995)—strategists tends to be restricted mainly to warm, well-insolated shallow lakes and epilimnia.

The phytoplankton species tolerant of, if not depend-
ent on, near-continuous entrainment within deep or turbid mixed layers have well-developed capabilities for maintaining growth despite intermittent brief exposure to light. ''Attuning'' (Reynolds, 1995) strategists tolerate this scale of pelagic disturbance through the projection of a large mass-specific surface area (>8 m² (mol cell C)⁻¹) in the case of 1-mm threads of Planktothrix agardhii, ~30 m² (mol cell C)⁻¹ in the case of an 8-celled colony of Asterionella formosa). They have high surface-to-volume ratios (generally >0.5 µm⁻¹) though these are not necessarily attained through small unit-size but by morphological complexity (filaments, attenuated or fenestrated coenobia, protuberances, etc.) and so continue to benefit from reasonably rapid rates of cell-specific nutrient uptake and growth (≥10³ × 10⁻⁶ s⁻¹ at 20°C). The potential for energy conversion may be enhanced by an increased cell-specific chlorophyll content and accessory pigmentation to widen the spectrum of harvestable wavelengths. Other typical R species include larger diatom (e.g., Aulacoseira, Synedra), desmid (e.g., Closterium, Staurosira), and Chlorococcales genera (e.g., Pedoustrum) and solitary, filamentous members of the Oscillatoriales (like Limnothrix and Pseudanabaena). Not all phytoplankton species fit perfectly within one or other of the three categories but show properties intermediate between them. What is interesting, how-
ever, is that the intermediacy in morphological and physiological adaptation matches well the intermediacy in their ecologies. The ''space'' between the invasive and the stress-tolerant acquisitive species is occupied by genera such as Dinobryon, Dictyosphaerium, Sphaerocyclus, Genella, Psilobrya, Argulus, Aphanizomenon, and Gloeotrichia, which diminish in surface-area/volume ratio and maximum growth rate but increase in their abilities to exploit and conserve nutrient resources. Freshwater algae between the invasive and the attuning poles include species of Cyclotella, Scenedesmus, and Coelastrum, which could be said to be increasingly large, more convoluted, and increasingly disturbance turbidity-tolerant algae. The axis between stress- and disturbance-tolerance is occupied by rela-
tively large, acquisitive, but self-regulating species, which can persist for months to years on stable density gradients, notably Planktothrix rubescens, P. mougeotii, and species of Lyngbya and Phormidium. Cryptomonas shows traits almost intermediate between all three extremes.

C. Seasonality of Planktonic Autotrophs

The utility of the functional classification is best demon-
strated in relation to the complex issue of seasonal change. Despite its celebrated species richness, the phy-
oplankton is usually dominated by very few genera at a time (it has been suggested that 95% of the extant standing biomass will be incorporated in no more than eight species at any one time; often it will be in rather fewer, as few as one or two; Reynolds, 1997b). Yet it is well recognized that the dominance moves among different species through time and that, in a given sys-
tem, the sequence will be similar from one calendar year to the next. Moreover, similar patterns may be ob-
served in similar but often geographically remote lakes. Numerous such cycles have been described in the literature (Reynolds, 1984b; Sommer et al., 1986); it is sufficient to mention a single archetypal example. In a small, calcareous, temperate lake in Britain (Crose Mere), the diatom Asterionella dominates the early spring growth, increasing from a few tens to several thousand cells per milliliter, over a period of six to eight weeks of lengthening days and intensifying insolation. Then, in mid-April, when the work of the wind is no longer sufficient to discharge the increasing buoyancy of the heat flux to the surface, the lake will become thermally stratified. Asterionella settles out of a mixed
depth, which is too truncated to keep it in suspension, leaving the clear water open to the establishment of such algae as Rhodomonas and Monodus, Cryptomonas, and of the motile colonies of Eudorina. As the summer solstice approaches, these too are replaced, first by nitrogen-fixing Aphanizomenon or Anabaena and by the dinoflagellate Ceratium, which will dominate the annual biomass maximum. In the autumn, the shortening days and declining temperature lead to a weakening of the stratification, deeper wind mixing, and the restoration of the depleted nutrients: by the late autumn, diatoms, including Aulacoseira and Asterionella, are generally the most abundant algae, dominating the shrinking residual biomass.

No two years will be exactly the same, and the relative proportions of simultaneously dominant and co-dominant species will fluctuate. Yet the pattern is robust and is amenable to diagrammatic summary (Fig. 3A): against axes empiricized in terms of $I_{**}$ and $K_{**}$, or even analogized to mixed depth versus the biologically available concentration of the critical nutrient, the time trajectory of the changing coordinates traces the extent of the seasonally changing environment, from the well-mixed, nutrients-replete starting condition, through the slow resource depletion of the spring period, and the rapid depletion after the onset of stratification. Finally, the effect of enhanced mixing weakens the light income but gradually restores the resource base. The partial independence of the two axes is conveniently emphasized by the inclusion of the winter “loop” when the light income falls below a level that will support new growth, while the system may be accumulating new external resources, only after $I_{**}$ has increased to the point where net growth can be supported at the cost of net resource reduction.

The development continues by the imposition of the trajectory on the plot showing the distribution of the morphological traits of the algae (Fig. 3B); as these prove such useful predictions of physiological performance, which, in turn, anticipate their ecologies, the fit should not be surprising. Nevertheless, the correlation of function and the dynamic response to environmental change is an extremely satisfying one. It will

---

**FIGURE 3** (A) How growth conditions in a monomictic deep lake (mixed in winter) track through a calendar year, selecting sequentially for a low-light tolerant spring bloom (of R species), then opportunistic C species after the onset of stratification, leading to the biologically enhanced segregation of nutrient resources tolerated by S species, before autumnal downmixing again forces conditions back toward the top right: nutrients may be further increased (by inflow) in winter, when primary production is still increasingly light limited. (B) Some of the algal species that might be selected are shown in an analogous matrix that is based purely on shape- and size- characters of the algae concerned ($I_{**}$ versus $K_{**}$, where $I_{**}$ is the surface area of the alga in $\text{m}^2$, $V_{**}$ is its volume in $\text{m}^3$, and $m$ is the maximum length dimension, in $\text{m}$). A strong correlation exists between the morphologies of these algae and their functional ecologies. The algae shown are as follows: An = Anabaena flos-aquae, Ap = Aphanizomenon flos-aquae, Arh = Anabaena flos-aquae, Ank = Ankistrodesmus falcatus, Aph = Aphanizomenon flos-aquae, Ast = Asterionella formosa, Cer = Ceratium hirundinella, Chl = Chlorella, Cya = Cryptomonas ovata, Dim = Dinochrysis dyerianae, Eud = Eudorina elongata, Fra = Fragilaria crotonensis, Lim = Limnothrix redekei, Mel = Melosira varians, Mic = Microcystis aeruginosa, Mon = Monodus, Pla ag = Planktothrix agardhii, Per = Peridinium cinctum, Rh = Rhodomonas pusilla, Scq = Scenedesmus quadricauda,Tab = Tabellaria flocculosa, Vol = Volvox aureus.
provide us with one of the most important clues to understanding how local diversity is maintained in nature.

D. Traits in the Marine Phytoplankton

Before that, however, it is necessary to establish that the rest of the planktonic biota respond to environmental cues in broadly analogous ways. Marine phytoplankton would seem appropriate for first consideration. Evidence of a similar coherence among morphological form, physiological performance, and population responses to environmental forcing was presented in Margalef’s (1978) seminal discussion of plankton life forms, while the insights underpinning the “mandala” model (Margalef et al., 1979) continue to intrigue plankton ecologists. Nevertheless, the apparent functional and morphological correlations have scarcely been quantified. Preliminary considerations of the types of phytoplankton characterize the various subdivisions of the sea (Fig. 2), however, do reveal entirely parallel trends between morphology and performance. For instance, as days lengthen over the open seas of the temperate latitudes, an initial typical spring bloom of such chain-forming diatoms as Thalassiosira norden- skjoldii, Skeletonema costatum, and Chaetoceros (all manifestly having R-type, attuning characters) gives place in summer to relatively large, motile, S-like peridinians (e.g., Eucampia tulearensis) and ceratians (e.g., Ceratium tripos), which are often followed, in autumn, by chain-forming (Chaetoceros spp.) or needle-like (Rhizosolenia) diatoms with high surface-area-to-volume ratios and antennal enhancement. In the high-nutrient, high-energy waters of coastal margins, potential dominants are generally small, usually unicellular, centric diatoms (e.g., Cyclotella caspia, Thalassiosira withis- flogii), green flagellates (e.g., Dunaliella, Nannochloris), euglenoids (Eutreptia), and gymnodinoids (Gymnodi- nium), all qualify as fast-growing opportunists (C-) spe- cies. The estuarine procrustenoids are also conspicuous. Blooms of Phaeocystis may follow, before giving place to ceratians. In the highly stratified, tropical oceans, dominance passes quickly from diatoms to Emiliania and to the large Ceratium spp. and Orthoniscus. The big, busyancy-regulating, stress-tolerant, S-species of diatom (Ehodinodes) and dinoflagellate (Pyrocystis) are at the culmination of this sequence. Spatial differences—among latitudes or between neritic shell waters and the open oceanic systems—probably also attest to the functional differentiation of the phytoplankton. Areas of frontal activity and of surface upwellings of the deep oceanic circulation currents tend to combine intensive mixing with renewed re- sources and to select species in the C-R intermediate through much of the year: Gyrodinium spp., Alexan- drium tamarense, Thalassiosira leptophora can be typi- cal assemblage members. Away from the epicenters of the upwelling areas and where, indeed, the water is less mixed but still charged with nutrients, such dinoflagel- late taxa as Dinophysis, Gyrosigma and Gymnodinium catenatum provide prominent markers of the (R-S) tran- sition between stressed and disturbed conditions.

IV. FORM, FUNCTION, AND SELECTION IN THE ZOOPLANKTON

A. The Basic Adaptations of Planktonic Animals

In the terms of species success, the survival strategies, and selective biases, the ecological constraints governing the lives of phagotrophic zooplankton are wholly analogous to those acting on the phytoplankton. Spe- cific requirements differ substantially, as do the time-scales over which the organisms respond to environ- mental fluctuations. The essential requirements of the phagotroph are to be able to encounter and capture sufficient appropriate food to supply the elemental bases of its biomass, to maintain and defend that biomass and perpetuate its genetic instructions, and to secure the energetic demands of the relevant processes. Resource adequacy for nonphototrophs may be judged principally in terms of the availability of ingestible par- ticular biomass (living or detrital), being the only prac- tical source of reduced carbon, although the judgment of organic-resource quality must acknowledge the ca- pacity of its blend of proteins to fulfill the biochemical requirements for nitrogen, phosphorus, sulphur, iron, and so on. Resource gathering, whether essentially herbivorous, carnivorous or detritivorous, or whether ac- tively hunted or sedentarily filtered, carries a high en- ergy demand: “maintenance costs” of phagotrophs are proportionately high, not untypically accounting for some 90% of the resources consumed. The energy comes from the original photosynthetic investment in carbon bonds, whose controlled oxidation sets the prin- cipal processing constraint in animals, which is to sup- ply the tissues with sufficient oxidant.

In terms of the lives of planktonic animals, already contending with the problems of operating in a fluid, viscous medium, the key issues are still about staying alive, about finding adequate amounts of suitable foods, with a sufficiency of individuals reaching reproductive
that this imaginative perception is but a quite indepen-
dent assessment of the expression of "resource-stress
tolerance," "exploitative invasiveness," and "distur-
bance tolerance" among animals. In this way, it is possi-
bile to fit planktonic animals to a habitat template, analo-
gous to that for the phytoplankton, on the basis of a
wholly functional classification.

B. Life-History Strategies

The resident zooplankton of the open seas and of larger
oligotrophic lakes live in environments where the pro-
ducer biomass is chronically resource limited; the po-
tential food resources for zooplankton (planktonic al-
gae, bacteria, and particulate organic detritus) are
generally \(< 0.1 \text{ g C m}^{-3}\). To be successful under persis-
tent conditions of dilute resource, animals must blend
energy-efficient foraging with controlled resource ex-
ploration—they must adopt the "patient" strategy of
resource-stress tolerance, especially with respect to the
recruitment of juveniles. Investment in a small number
of relatively large eggs with a relatively long develop-
ment time is typical. Physical and behavioral defenses
reduce the vulnerability to predation. Such traits distin-
guish the life histories not only of Diaphanosoma but
of many of the calanoid copepods also. The sophisticated
browsing mode of feeding they adopt has been shown to
be sufficient to sustain natural populations and support
their reproduction at food concentrations in a range
equivalent to 0.01 to 0.08 g C m\(^{-3}\) in both the sea
(Huntley and Lopez, 1992) and lakes (Hart et al., 1996).
For reference, a producer biomass of 0.08 g C m\(^{-3}\) may
be supported by photosynthesis in a mixed layer of
\(\leq 60 \text{ m} (i.e., \leq 5 \text{ g C m}^{-3})\), so long as the microbial
loop is able to maintain a bioavailable nutrient base
exceeding 0.1 g P and 1 g N m\(^{-3}\)). Arguably, the ciliates
(e.g., Halteria, Strombidium), which constitute a sig-
nificant part of the diet of calanoids in oligotrophic
lakes, must be, similarly, stress-tolerant patient strate-
gists.

In the more productive waters of fertile lakes and
rivers, estuaries, nutrient-enriched coastal-shelf waters,
and oceanic upwellings, the capacity to support auto-
trophic biomass may, for some of the time at least, signifi-
cantly exceed 0.1 C g m\(^{-3}\). This is an approximate
threshold, above which the energetic return from filter-
feeding becomes steadily more favorable. Among lakes,
for example, Daphniids tend to become much more
abundant, absolutely, and relative to calanoids. Most
species of Daphnia, when adequately nourished, are able
to grow and mature rapidly, to increase egg production
and to recruit the next filter-feeding generation over a
few days (George and Reynolds, 1997). The 12-fold
increase in the Daphnia-dominated community-filtration rate in 13 days, measured by Reynolds et al. (1982; note, aggregate filtration capacity doubles every 3 to 4 days), conforms to the understanding of an exploitative, or “violent,” life-history strategy. Among the noncrustaceans, the ciliates of microaerophilous environments (e.g., Metopus, Caenomorpha) react in an analogous way to a food abundance in media from which other sorts of consumer are mainly excluded. Many noncrustacean members of the zooplankton take advantage of short generation times to track bursts of algal or bacterial abundance. Protists including Difflugia, Coleps, Tintinnidium, and Nassula and rotifers such as Keratella, Kellicottia, and Brachionus will often respond to an expansion in suitable algal foods, but this occurs before the Daphnids can ascend to dominance. Their disturbance-accommodating lifestyles are well suited to the direct exploitation of small phytoplankton and organic detritus in the middle reaches of large rivers, provided travel times permit (Viroux, 1997).

V. FUNCTION IN THE BACTERIOPLANKTON

Microbial populations constitute the third integral component of the planktonic communities of lakes and seas, where they are often numerous (typically within the range 10^5 to 10^7 cells ml^-1). Microbial diversity, however, is not yet well understood, principally because the bacterial taxa are still insufficiently distinguishable to be separated routinely to species level. For many years, bacteria have been identified as much by what they do as by their morphological affinities. Given that, for most of them, the utilization of dissolved reduced-carbon substrates provides the main source of energy, even the functional approach appears to have its limitations for bacterioplankton. "Organic carbon compounds" are derived from a very wide range of the breakdown products of biogenic materials, many originating from the land, as well as a host of very unnatural anthropogenic organic substances. The ability to break down particular classes of compounds or kinds of bonds is unlikely to be shared by all bacteria. Every such process presumably requires the action of one or more discrete enzymes, and the ability to produce each enzyme requires one or more genes dedicated to its production. As the genetic complement of microbes is relatively modest, it follows that the number of separate, substrate-specific bacterial strains is likely to be high.

Modern molecular approaches are slowly sorting out "who does what" but, for the present, the diversity of the bacterioplankton is anticipated to be high.

Among oxic freshwaters, the free-living bacteria are known to include members of the coccoid and rodlike Bacilli, the Flavobacteria, Pseudomonads, and Vibrios (Atlas and Bartha, 1993). All are heterotrophs, which exploit sources of biogenic organic carbon. These sources were always recognized to constitute the waste products and cadavers of the biont components of the aquatic system generally. As understanding of pelagic carbon metabolism has developed, it has become possible to separate the active role of microbial heterotrophs in transferring dissolved organic compounds derived from low-molecular weight, mainly algal photosynthetic intermediates (especially dissolved glycollate), as the food supply to the flagellate—ciliate—copepod trophic links, from that of degradation of biogenic particulates (organic detritus, fecal pellets, etc.), usually by sessile or stalked bacteria, such as Caulobacter, actinomycetes, and fungi, that attach to the particles and whose activities liberate carbon dioxide and mineral nutrients into the water. Both contribute substantially to the cycling of carbon and nutrients, but in contrasting ways. Oxic breakdown of abundant biogenic material releases resources to the autogenic synthesis of new mass. The phagotrophs of the microbial loop depend on the close coupling that the viscous scale facilitates, but most of the energy is consumed internally and the material transferred to higher trophic levels is absolutely small.

In smaller lakes and ponds, the nature and diversity of organic carbon sources are enriched by direct transport of biogenic materials from the catchment area. However, current understanding is also having to accommodate the recognition that often the largest fraction of organic carbon present in stream and lake waters comprises plant-derived humic and fulvic acids washed in from catchment soils (Wetzel, 1993). They are particularly abundant, of course, in brown and “black” waters draining forests and peatlands. The relative size of this fraction indicates its high resistance to microbial breakdown; it persists in measurable amounts in the largest oligotrophic lakes and even in the open oceans. Refractory dissolved organic carbon from terrestrial sources does break down slowly; some is rendered labile by exposure to ultraviolet radiation.

Away from aerobic mixed layers, freshwaters furnish other microorganisms exploitable by bacteria. In the microaerophilous environments of the bottom sediments and in the deep water of stratified, productive lakes,
anaerobic microbes are common and typically numerous; they include Clostridium, the sulphate-reducing Desulfovibrio, and the Archaea Methanogens. Interacting gradients of light and of redox superimposed on relatively stable density gradients may also provide vertical sequences of microhabitats, each niche potentially differentiated according to the microorganisms it supports and whose physiological adaptations it most suits. To be able to carry out photosynthesis in a reducing environment favors Rhodospirillum, but to be able to use sulfide or sulfur, as photosynthetic electron donor, can bias the selection in favor of Chromatiaceae and Chlorobiaceae. The involvement of specialist chemolithotrophs in nitrification (Nitrosomonas, Nitrobacter) and in the oxidation of sulfur (e.g., Thiobacillus, Beggiatoa) and iron compounds (Ochrobium) further adds to the microbial diversity of freshwater.

In the open waters of the sea, most bacteria are aerobic—relatively few anaerobes are ever found in the surface waters. Many of the same genera represented in freshwater include marine species: Pseudomonas and Vibrio are often found to dominate and species of Flavobacterium, Alcaligenes, and Cytophaga are found in high numbers (Atlas and Bartha, 1993). Sediments receiving organic inputs are bacteria-rich, moreover, where these fall anaerobic, marine sulfate-reducers and methanogens are present in substantial numbers. In both marine and freshwater, the vital contribution of the bacterioplankton to ecosystem function arises not only from the mineralization of organic carbon products but from the potential reutilization and renewed bioavailability of nitrogen, phosphorus, and the metals required in the assembly of new biomass. Most bacteria are supposed to be capable of rapid self-replication, with the potential to undergo several doublings per day. The relative ease of dispersal contributes to their apparent ubiquity and to their fidelity to appropriate habitat opportunities. Microbial biomass is, nevertheless, clearly subject to environmentally imposed rate and capacity limitations. In many instances, their requirements are not dissimilar from those governing the growth of photoautotrophs. However, small cell size and superior uptake kinetics favor the performances of bacteria. A greater collective versatility in the energetics and potential sources of reductant and oxidant opens up more of the aquatic environment to chemotrophs than to photothrophs. The distinguishing dependence is the one for appropriate organic carbon skeletons. Their availability remains the regulation on the numbers and dynamics of planktonic bacteria. The nature of the carbon compounds, together with the metabolic constraints set by the redox environment, are presumed to be decisive in species distributions.

VI. TEMPORAL PATTERNS IN THE ORGANIZATION AND DIVERSITY OF PLANKTONIC COMMUNITIES

A. Quantifying Structure in Planktonic Communities

This survey of the wealth of species to be found in the plankton serves to emphasize the collective breadth of the adaptive specialisms themselves but which, individually, provide advantage only to the certain species that possess them and only at the certain times when the specialism affords some operational benefit over others that do not have them. A reasonable deduction that may be made is that the collective diversity of planktonic species reflects the number of distinctive habitats, or niches, that it is possible to define. This is, roughly, the niche differentiation theory of biodiversity: every species has an optimum performance, which is facilitated when each of the component processes is at the species-specific optimum. When the conditions simultaneously satisfy all the species-specific optima of a given species, then that species is uniquely favored to outperform all the others for which the niche conditions are suboptimal. There can then be as many successful species as there are tangible niches, for each stands to be the fittest competitor under its favored blend of environmental conditions. This means that in any given niche location, the local diversity is likely to be actually quite low.

An alternative view might be moved to accommodate the observation that local diversity is sometimes actually quite high, even when the available niches seem to be few in number (this, it will be recalled, was Hutchinson’s paradox). It should also accommodate another of the observations emphasized in the foregoing sections—that is, there is a great deal of temporal environmental variability, sufficient, indeed, to move the selective advantage from one species to another, often before any has been able to assert its competitive dominance. Sometimes, moreover, the variability is plainly driven externally, beyond organisms’ power to regulate—the changes are brought about by physical forcing with sufficient magnitude to “disturb” the existing internal organization. At the appropriate intensity and fre-
quency, such disturbances might provide a mechanism for rotating the competitive advantage among species and for maintaining local diversity. This, indeed, is the essential basis to the main counterview to niche-differentiation theory.

Two further features of the present appreciation of planktonic communities that are attractive to the student of biodiversity have also been noted in the survey. One is the planktonic timescale: the frequency of generations allows the rates of internal compositional changes (competition, succession, predation) to be measured and the resistance to external forcing, as well as the resilience to recover from forcing, to be quantified (Reynolds, 1997a). The other is the quantitative information that is now available on the performance capacities of representative planktonic organisms—from their resource thresholds to their resource-regulated responses and the onset of their resource-saturation. These help us to distinguish internal structural changes from externally forced restructuring and to express them in energetic terms. In this section, we may deduce why the two main theories of biodiversity are not at all mutually exclusive and when we may presume either is valid. To do this, however, it is necessary formally to distinguish between the diversity of community structure at a given location and the richness of species present among a series of such locations at the regional or even global level.

B. Quantifying Diversity in Planktonic Communities

Diversity is a concept drawn from communication theory (Shannon, 1948), referring to the amount of discrete information in a particular location. In communities, it could refer to the different fragments of genetic information. A diverse community is one with many different kinds of genes (species) present simultaneously. This biotic diversity (Hd) can be quantified with great precision, using the specialized Shannon-Weaver function:

\[ H^* = -\sum \frac{b_i}{B} \log_2 \left( \frac{b_i}{B} \right) \]

Diversity increases with the number of species (s) each contributing biomass (b) to the total, B, in the sample of lake or seawater. The more species present, the greater is the diversity, until \( H^* = \log_2 s \). In practice, it is necessary to set a bar on the efficiency of search for the rarer species (more diligent searchers and more expert taxonomist-hunters would always score higher diversities). Nevertheless, the evenness of the interspecific distribution, \( E = H^*/H^*_\text{max} \), is itself a useful measure.

Local species richness (the sample, the pond, the lake) may run to a few tens of phytoplankton and, if the phagotrophic protists are included, perhaps 100 or so species of zooplankton. Regular sampling may raise this total considerably, as seasonal changes are encountered and the probability of encountering rare species is increased (Padišák, 1992); “rare” in this context may mean one individual per liter, but there could still be a billion such individuals in a small, 10-ha lake. As observed earlier, most (\( \approx 95\% \)) of the total biomass of phytoplankton and that of zooplankton will each be invested in eight or fewer species. Most of the consideration of evenness concerns how the total biomass is shared among those eight. Thus, whether local species composition is species rich or otherwise, the perception of its diversity is greatly affected by the relative evenness of the most common species.

Often without clear confirmation, the issue of “protecting biodiversity” generally extends to the totality of species, or the totality of species that could live in a particular habitat (Wilson and Peter, 1986). Unlike diversity, it has no formal mathematical expression. However, we may quickly recognize that if a single patch of the aquatic environment—say, the open water of a 10-ha lake—can furnish habitat sustaining viable populations of a few tens of species of planktonic species, then a great many such ponds must exist if all the species are to be supported, many of which will be common to a wide selection of the water bodies.

In this way, two questions need to be answered. The first is about the mechanisms by which within-patch diversity is upheld, against the tendency for the optimally adapted functional specialists to exclude its competitors. The second questions the mechanisms by which between-patch diversity is maintained, so that individual species have more opportunities to maintain stocks at viable and sustainable levels.

C. Assembly and Autopoiesis of Communities

The essential step in logic that separates the niche-diversification and the persistent-disturbance theories of species richness rests on an acceptance that the supportive capacity is not always filled; conversely, although the capacity of the resources is always finite, the underexploited resource base is, by definition, not “limiting.” Many students of the phytoplankton persist in comparing the quantities of nutrients in the water relative to the ideal composition of algal biomass, then...
judging that the nutrient in least supply is "the limiting factor." That phytoplankton concentration is ever able to increase at all is because, in reality, there are times when nutrient resources are available to support the additional biomass and unharvested light energy to fuel its assembly. Moreover, these opportunities are attributable to natural fluctuations in the resource base: the spring increase in temperate waters, the most conspicuous feature of lakes and seas at higher latitudes (Fogg, 1965; Sverdrup et al., 1942) is the consequence of lengthening days and strengthening insolation of waters, which have themselves demonstrably been recharged with nutrients during the winter months of hydraulic circulation and augmentation from inflow, at the same time as meeting low biological demands. The spring bloom, at least initially, is the response of the producer community to an expanding capacity of the resource and energy bases. Later in the year, the resource supply may well fail to meet the biological demand and severe constraints are then imposed on its further growth and maintenance. Alternatively, it is left to shortening days and convective mixing in the autumn months to erode the supportive capacity of the insolation flux. We will return later to the subject of these constraints but, for the moment, it is important to consider in more detail how the phytoplankton behaves when its production is not limited by nutrients and when the production of zooplankton leaves large amounts of phytoplankton relatively still unharvested. The organismic responses to the opportunities provided by a luxury of supply in all resources are instrumental in molding the structure of ascendent communities. The key feature of an expanding (or merely replete) resource base is that its exploitation is relatively straightforward: species do not compete for the resources in the sense that their garnering by the individuals of one species denies an adequacy of the nutrients simultaneously supplied to the individuals of a second or an nth species. The only question of ecological importance is "Can it obtain enough?" This is not to say that opportunistic "luxury uptake" and resource storage might not become important at some later juncture, only that species 2 to i can also perform to their capacities ever attain such outright dominance of the phytoplankton, leads to the deduction that there are many routes of internal change but few ultimate outcomes. This last point is taken to be a powerful indicator of the system's ability to self-organize, a measure of its autopoiesis. This is a property of all ecosystems (Jørgensen, 1992) but its mechanical basis is not well understood. However, the further examination of the processes of capacity-filling behavior of the phytoplankton yields a conceptual and preliminarily empirical view of internal organization. If the developmental progress of a producer-dominated pelagic system is considered against axes representing the standing crop (in units of organic carbon) and the harvest of photosynthetically active radiative flux (in W m⁻² or J m⁻² d⁻¹), then the coordinates corresponding to the "open stage" at the start of the summer stratification may be set close to the origin (Fig. 4A). The standing crop is far below the potential of the resource-limited carrying capacity (horizontal axis) and thus is collectively capable of
intercepting very few of the photons that penetrate the water surface before their energy is absorbed by the water. The notion of capacity filling is that the structure will move away from the origin and toward the upper right corner, whose coordinates are determined either where all of the least available of the resources has been incorporated into biological standing crops (the resource capacity) or where every available photon is intercepted by planktonic light-harvesting centers and is driving photosynthetic carbon reduction (the processing capacity). If we now imbue our producer community with the physiological performance of a Chlorella culture, the unimpeded biomass-specific light harvesting and carbon fixation rates are able to accumulate new cell carbon sufficient to double or possibly quadruple the plant biomass within 24 hr. It is possible to suppose, in these early developmental stages, that the doubling of the biomass will roughly double the area of light harvesting surface, so the coordinates representing the supportable biomass move upward and rightward. The new biomass incurs an increment in respiratory losses, so the coordinates representing the

FIGURE 4 (A) A simple plot to show the maximum light that can be harvested by a given biomass of planktonic algae relative to the energetic costs of its maintenance and loss rates; the difference is the exergy of the system. For a time, developing systems increase biomass and exergy, until the reduction in the energy cushion leaves them increasingly sensitive (B) to fluctuations in the harvest-determined carrying capacity. (C) resource limitation truncates the cushion but the “sail shape” (D) can be used to subsume the triangle from Figure 3.
maintenance costs—assumed, for simplicity, to be directly proportional to the biomass—diverge from the slope of the harvesting capacity. Given strong light and a steady water temperature, producer biomass begins to increase exponentially. It is predictable that, as accumulation proceeds, the light-harvesting centers are increasingly probable to be shaded (meaning that even with no diminution in the surface flux, individual light-harvesting centers are activated less frequently). The provision of additional biomass is no longer rewarded with proportionate energy harvest.

The limiting condition is approached when the maintenance of the expanded biomass consumes all the energy it is able to capture. This biomass, shown by the broken line in Fig. 4A, is the maximum that can be supported. The coordinates are proposed on the basis of photosynthetic and respirational properties solved for Chlorella and the highest daily aggregates of light income daily income of photosynthetically active radiation that can be realistically proposed (12.6 MJ m⁻² d⁻¹; 10.5 mol C m⁻²); through relationships developed in Reynolds (1997a).

D. System Exergy and Disturbance

The precise quantification of Fig. 4A may be misleading and not universally applicable. This is less important, however, than the form of the traces of the maximum light-harvesting curve and the maintenance diagonal and of the geometrical shape that they bound. It represents the only part of the plot in which population growth can be sustained. Above it is a void of underpopulated opportunity (exploitation of which requires the expansion in the light-harvesting capacity by increasing producer biomass in the rightward direction); below is an area representing unsustainable biomass in excess of the current energy income (far from being able to increase its total biomass, the system here must shed biomass, so that the maintenance of what remains is brought back within means of the current energy flux to sustain). In this way, the spring increase of phytoplankton is represented by a rightward response of biomass to an upward income in harvestable energy, with the increase in light harvesting capacity moving diagonally upward and rightward in consequence. By analogy, the annual decline in biomass is a leftward response to regain balance between now-high maintenance costs against a diminishing energy income. The excess of energy-harvesting capacity over the costs of its maintenance provides the existing crop with a second asset—a cushion of capacity that can absorb the impact of short-term variability in the energy income without resorting to a restructuring of the biomass.

Because we have defined this cushion in terms of energy exchanges, it is appropriate to refer to it in thermodynamic terms—it is the exergy of the system (Jørgensen, 1992). We may see at once a presumption that a high income flux relative to biomass maintenance favors exploitation by species that can build the highest levels of harvesting capacity for the biomass—that is, those which contribute most to a high level of exergy. However, high net fluxes are not reliable and energy harvesting is ultimately sensitive to changing day length and solar elevation and to the variability of cloud cover and atmospheric absorption and scatter, all of which affect the daily flux of harvestable energy penetrating the water. Moreover, atmospheric variability also includes the work of wind and the impacts of varying dissipation rates on the mixed-layer depth and, hence, the mean exposure of entrained plant biomass to harvestable energy.

The reality of biomass assembly and its additional energy-harvesting capacity is that they have to be conducted against a background of a stochastically changing energy income. Intriguingly, the components of the relationship are provisionally quantifiable in energetic terms, at least probabilistically, if the energy harvest is solved as the mean biomass-specific photon capture by plants entrained in the mixed-layer circulation (Reynolds, 1997a). In this way, the temporal fluctuations in the harvestable income can be represented as an irregularly oscillating time track (the broken line in Fig. 4B). High values correspond to intense photon fluxes into clear or shallow mixed layers, low points to overcast skies or strong wind-forcing and mixed-layer deepening. We may also recognize that the externally forced depression of the harvestable energy flux may still be sufficient to meet the biomass maintenance cost, with a little in hand to maintain positive increase—that is, the fall in income is absorbed by the exergy cushion; the structure survives and has the resilience to recover promptly to its maximum capacity when the harvestable energy flux is restored to an optimum and the capacity to expand the biomass is fully restored. On other occasions, however, especially once a substantial standing biomass has been put in place, the flux of harvestable energy falls below the minimum maintenance requirement. The cushion of exergy is exceeded; the energy needs of the system are now out of balance with its supplies and are fundamentally unsustainable. This situation cannot persist for long before there has to be a reduction in the standing biomass, back to a level that is energetically sustainable. There is abundant resort
to minimal metabolism, including the production of resting spores and propagules, through to mass mortalities of vegetative cells. The biomass response is sharp and severe, conforming to all conventional appreciations of an externally forced disturbance reaction.

Following the course of irregular forcing episodes, represented in Fig. 4(B), it is easy to appreciate that planktonic systems are far more liable to restructuring disturbances than they are to achieve their autopoeic potential of a low-diversity, competitively excluded climax to the succession. This maturation process is simply overridden by the frequency of externally forced disturbances that it rarely, if ever, proceeds to its logical outcome. It may also be seen that the intervention of weather fluctuations and events, superimposed on predictable climatic cycles, is a, if not the, principal agent resisting exclusion and local extinction of planktonic species and contributing to Hutchinson’s paradox of local species richness. Steady states are just so rarely achieved that species are retained, at least at the scale of plankton generations, in a nonequilibrium coexistence. Here, at least, is a candidate mechanism contributing to the maintenance of a high biodiversity in the plankton. We should pursue its workings a little further.

E. Resource Limitation and Diversity

Long before most pelagic systems can approach the attainment of a producer base at the capacity of the harvestable energy flux, biomass assembly will have been constrained by the capacity of the bioavailable nutrient resources. We can represent this constraint very simply by superimposing the vertical axis across the exergy cushion (Fig. 4C). Between the origin and the resource limit there remains an opportunity for harvestable energy flux, biomass assembly will have been constrained by the capacity of the bioavailable nutrient resources. We can represent this constraint very simply by superimposing the vertical axis across the exergy cushion (Fig. 4C). Between the origin and the resource limit there remains an opportunity for harvestable energy flux, biomass assembly will have been constrained by the capacity of the bioavailable nutrient resources. We can represent this constraint very simply by superimposing the vertical axis across the exergy cushion (Fig. 4C).

The geometric representation of environmental variability keeps resetting the coordinates of the environments, at least with respect to the chosen axes, the assembly opportunity remains strictly within the definition of exergy. Thus, the area of the plot corresponding to sustainability of growth (Fig. 4A) is finally shaped as in Figure 4D.

The new geometric figure is without a name—the best likeness I can think of is the sail of a windsurfer. Its periphery, however, readily corresponds to quantities established in sections III–V: the abundance of resources, relative to the consumptive demand, favors the advance of the most exploitative species, whose geometric range spreads from the origin along the line of the maximum biotic exergy flux. The closer this approaches the capacity of the resources, the greater is the stress of resource deficiency and the greater is the adaptation required to exploit it. Maintaining a high exergy flux remains the strongest driver of the planktonic succession. Plainly, the curved upper surface corresponds to the C-S axis of increasing structural development, increasingly subject to resource competition and powerful selection. In contrast, the maintenance axis is the lower boundary of the adequacy of the exergy flux to drive the assembly of the community. Thus, the lower boundary of the “sail” represents the extremes of disturbance tolerance permitted by the attuning R-strategy. The two straight boundaries could equally be scaled in terms of K** and P**, the rightward trend representing increasing resource limitation, the downward trend corresponding to carbon-processing limitations. The axes serve just as well in separating the animal analogues of food supply and resource stress and “explorer” opportunism provided by externally imposed disturbance. Subject to current methodological uncertainties, there seems to be every probability that analogous axes describe the availability of organic carbon sources and processing opportunities for microbial plankton too. Note also that the fluctuating coordinates of stochastically variable environments continue to fall both within and beyond the sail area defining positive growth responses; only when the track is kept firmly and persistently in the S or R areas of the plot is there likely to be fierce and ongoing competition leading to the progressive installation of a single species, dominating a low-diversity, low-equitability community. On the other hand, for as long as externally driven variability keeps resetting the coordinates of the environmental conditions, at least with respect to the chosen axes, the time track keeps moving freely and extensively across and beyond the body of the shape, signifying that conditions rarely exist for sustained competition to last either for long enough or in one direction for any of the species present to have sufficient chance to deploy its superior adaptations to the competitive exclusion of others.

VII. MECHANISMS PROMOTING AND MAINTAINING DIVERSITY IN THE PLANKTON

A. Diversity within Habitats

The geometric representation of environmental variability and of its impact on species selection may properly be pursued in relation to the exergy model, but it is actually easier to bring the concept of fluctuation tracking back to the habitat template (of Fig. 3A), because it is easier to relate to the environments it seeks to represent (Figs. 1 and 2) and because sufficient pre-
liminary knowledge exists about how the tracking thus represented actually selects for the preferred traits of species, at least of the freshwater phytoplankton (Fig. 3B). To be clear, the resource- and energy-replete conditions, wherein resources fully meet present organismic demands, are represented in the upper right-hand corner of the template. Of those present, the species most advantaged are the opportunistic, invasive C-type species (like *Chlorella*, see section III.B and Fig. 3B). Biomass growth creates a strain on the readily available resources, more elaborate resource gathering is required, and autopoesis favors a succession to more conservative, accumulative S-type dominants (like *Microcystis*) with a high tolerance of resource-supply stress and a resistance to disturbance. Large size becomes selectively valued but at the tradeoff in terms of surface-to-volume ratio (Fig. 3B) and at the price of slower metabolism and growth. In more nutrient-replete environments, the principal stress is imposed by having to operate on low or intermittent light doses. Survival prospects are enhanced by motility, combined with increased size, but the selective advantages of a high surface-to-volume ratio are not abandoned—the morphological attenuation among the acclimating R-type diatoms and *Planktothrix*-type filaments offers a high tolerance of mechanical disturbance and the resilience to recover from severe forcing events.

The organizational trends imposed by disturbance and stress press the selective bias toward the R or S apices of the triangular template, shown in Fig. 5A. We have already seen that seasonal trends track through the template matrix in broad, predictable ways (summarized as S-ward and R-ward trends in Fig. 5B), but substantial within-season variability will lead to almost chaotic short-term time tracks, such as the fragment included in Figure 5C.

Our deduction is again that the effect of environmental variability is to move any selective advantage among species at a faster rate than autopoesis narrows the opportunities or that competition can forge a low-diversity monoculture dominated by a single well-adapted species. We know that this theoretical outcome, correctly anticipated by Hutchinson (1961), certainly is achievable and good descriptions of local, competitively excluded community structures are described in the literature. These cover the sort of arrested "plagioclimaces" of overwhelming *Microcystis* dominance of tropical eutrophic Lake George (Gam and Viner, 1973) and imitated in several field-scale enclosure experiments (summarized in Reynolds, 1988a) of year-round *Planktothrix* dominance of exposed, continuously mixed hypereutrophic polder lakes in the Netherlands (first described by Berger, 1975), of the sustained dominance of nanoplancton in the continuously flushed, ground-water-flushed Montezuma's Well in Arizona (Boucher et al., 1984), and the unique dominance by *Chlorella* of a cooling gradient across the hot-spring fed Rotowhero, New Zealand (Jolly and Brown, 1975). No less impressive is the striking commonality of the autopoetic organization of vertically segregated layers (or "plates") of algal and microbial producers on the stable physicochemical gradients in permanently ice-covered lakes in Antarctica (e.g., Vincent, 1981), in tropical forest lakes (Reynolds et al., 1985), and in midlatitude karstic dolines (Vicente and Miraclée, 1988).

In contrast, however, the great majority of assem-

![Figure 5](image-url)
baffles that have been sampled, identified, and described in the literature are, simply, weakly organized. There need not be any paradox in this, once it is appreciated that in such accumulating, mid-successional assemblages most of the specific populations are not in any steady-state but rather in a state of flux, either increasing or decreasing in response to events in the recent past. Thus, most so-called communities, in fact, comprise populations which, at the given point in time, are in some stage between being dismantled and reassembled. The crucial questions about local diversity in the plankton should be “How long has it been since a restructurating was initiated by external forcing of a magnitude sufficient to create a disturbance?” and “At what rate is the restructurating taking place?” The field-scale experiments carried out in the large limnetic enclosures in Blelham Tarn, England, pointed to a progress to steady-state monocultural dominance requiring from 12 to 16 generations of the eventual dominant (Reynolds, 1988a). The time required for this could be as few as 35 to 60 days, provided growth rates could be sustained by water temperatures ≥20°C and a supply of carbon and nutrients of a capacity to support the last population doubling. Accordingly, limitations on growth rate imposed by low temperature, slow carbon renewal, or nutrient depletion extend the generation times reciprocally, delaying the onset of the climactic condition and slowing the exclusion of diversity. More likely, the intervention of external forcing (strong winds, heavy rain) would interrupt the successional maturation at an earlier date, and if of sufficient magnitude (exceeding the net structural exergy flux), selecting for alternative species with a new target outcome. It was found consistently that the community response, represented by two to four divisions of the newly selected species simultaneously with the dieback of the erstwhile-selected species, took some 5 to 15 days before it was clearly manifest. It was deduced that a disturbance frequency of this magnitude (two to four generations) is sufficient to maintain an optimal local species diversity (Reynolds, 1988a).

Analyses of temporal diversity fluctuations in small lakes in locations as far apart as Canada (Trimbee and Harris, 1983) and central Hungary (Padisák, 1993) confirm the validity of this deduction, showing diversities in phytoplankton composition peaking within 11 days of a recognized physical stimulus. These and numerous other case studies have been brought together to demonstrate this planktonic validation of Connell’s (1978) intermediate disturbance hypothesis (Padisák et al., 1993). Its articulation followed from a consideration of the celebrated diversity recognized in coral reefs and tropical forests, which in some ways, posed analogous questions to those raised by Hutchinson (1961) about the plankton. The essence of Connell’s hypothesis is that frequent disturbance excludes all but fast-maturing species; very infrequent disturbance allows competitive exclusion to reduce diversity, therefore, maximum diversity is maintained at intermediate frequencies. Recently, this simply stated insight has been the subject of a belated debate among ecologists. Setting aside those who contest whether others had not put forward the idea before Connell distilled it so elegantly and those who complain (mistakenly) of a lack of experimental evidence, the main line of argument has sought to distinguish the effects of disturbance intensity and of disturbance frequency. The debate is partly resolved by recalling the fact that disturbance is solely the response to an imposed forcing. The reference point of whether a disturbance has occurred at all and, if so, how intense it was, should be most usefully found in comparison of the forcing energy and the accumulated cushioning of the exergy flux. If structure has to be lost and regrouped as a consequence of an external force and if the recovery depends on the restoration of opportunity, local Shannon diversity can be shown consistently to benefit during the reconstruction phase, when competition is least. At both high and low frequencies of forcing, the numbers of intermediate species are reduced so that the inocula are no longer readily available to take advantage of conditions under which their growth might be favored. In contrast, local stocks of mid-successional species are enhanced by the sorts of opportunities that might be provided every two or three generations by the (intermediate) rejuvenation of appropriate growing conditions and with a marked alleviation of severe interspecific competition in the early post-disturbance sequel.

**B. Species Richness among Habitats**

It can be seen that the species representation contributing to the local diversity remains heavily biased toward those that have been well represented in the recent past, and which, of those whose growth should be favored by the onset of the appropriate environmental characteristics, remain potentially capable of seeding the largest inocula from standing stocks of vegetative cells and resting propagules. With the same kinds of environmental variability affecting Shannon-type species diversity in each tangible locality, there is still no clear explanation as to why the collective representation of planktonic organisms is so relatively species rich. We need a proposition for separate species representation...
in separated localities which recognizes that communica-
tion between localities is sufficient to maintain the high
level of apparent cosmopolitanism among plank-
tonic organisms. In other words, we need to be aware
of the roles of perennation and dispersal in relation to
the maintenance of biodiversity.

The survival of any species whose range of habitat
suitability (as defined in section II) is discontinuous,
in time as well as space, involves the separate develop-
ment of separated populations, but with a measure of
regenerative connection and gene renewal that resists
their permanent divergence. The metapopulation ecol-
ogy of plankton is not, formally, a well-rehearsed topic,
though there have been numerous studies to contribute
a general appreciation of the principles. Although the
existence of liquid water on the planet has a very long
geological history, individual bodies of fresh waters are
extremely transient. With the exception of the basins
formed by tectonic movements, a majority of these is
less than 20,000 years in age. The idea of lakes as islands
in a terrestrial sea is an easy one to assimilate, even if
we wish to extend the analogy to suggest they are like
volcanic eruptions that are shortly to be eroded back
under the sea. The actual oceans are physically and
temporally contiguous, but the patterns of global circu-
lation allow significant habitat differentiation to be
maintained at the scales at which planktonic organisms
live their lives. The principles of island biogeography
propounded by MacArthur and Wilson (1967) provide
a good model for plankton ecology.

The temporal discontinuities are most appropriately
bridged by the production of resistant, resting propa-
gules. Even the most elementary biological texts on
aquatic protists are memorably punctuated with refer-
ences to the production of cysts, to survive adverse
conditions. Certainly, among the freshwater phyto-
plankton, almost all the major groups represented pro-
duce some kind of physiological resting stage, if not a
discrete resting spore, cyst or akinete, some of which
can remain dormant for many years and still be fully
viable (see Reynolds, 1984a; over 60 years in the case
of some Anabaena akinetes in a dated lake sediment).
Eggs may fulfill a similar role for zooplankton, main-
taining banks of inocula, pending the restoration of
conditions favorable to growth and recruitment of suc-
cessive generations (Hairston, 1996). Produced in ade-
quately numbers, such dormant life-history stages pro-
vide a significant survival “hedge” through periods of
hostile conditions (low temperature, low light, poor
resources or food availability, drought, and so on). The
inocula potentially supplied by subsequent spore germi-
nation and egg hatching provide a powerful mechanism
for maintaining diversity in seasonally fluctuating envi-
ronments.

In many cases, the propagules facilitate spatial trans-
fers as well. Dispersal in water droplets, or in dust, or
in or on the bodies of animals have been shown to be
effective pathways of planktonic organisms (review of
Kristiansen, 1996) and to have been directly implicated
in the establishment of populations in new or isolated
bodies of water (Maguire, 1985, 1977). Planktonic spe-
cies are not equally amenable to dispersal, the trait being
related to other aspects of the life-history strategies;
for instance, effective perenniation and dispersion are
essential properties of species of temporary or tempo-
rally variable habitats. In general, the dispersability, or
invasiveness, of propagules (or, indeed, of vegetative
cells) is dependent on a raft of such species-specific features as their sizes, their resistance to desiccation,
and the numbers and frequencies with which they are
produced. The probability (p) of a given species (A)
being able to establish itself in another water body will
be determined as a function, partly of its relevant spe-
cies-specific traits (T_A) and partly of the problems posed
to all would-be invaders of the distance (d) and size
(a) of the new (or “target”) site from the existing (or
“source”). MacArthur and Wilson, 1967) population:

\[ p = f (1/d, a, T_A) \]

An important converse of this relationship also mani-
fests itself in relation to the ongoing suitability or effec-
tiveness of source and target sites as a habitat for the
given species. Either because the habitat changes in
consequence of autogenic properties or because species
A is prevented from completing its perennation in that
habitat, through the intervention of a facultative preda-
tor or pathogen, it is ultimately likely that the survival
of the species in that location is at risk. Repeated at
several sites, this process would threaten the survival
of the species in its entirety and bring about a diminu-
tion in total species richness (i.e., the biodiversity). The
resistance to that comes, quite literally, from the patch
dynamics of habitat distribution and their temporal
suitability to species A. Thus, the status of species A is
a function of the number of possible habitat patches
(N), the number that is occupied (O), the rate at which
they become excluded therefrom (e), and the probabil-
ity (p) of colonizing the (N-O) unoccupied habitats.
The rate of expansion (or loss) of the metapopulation
in time \( (dO/dt) \) may be the symbolized:

\[ (dO/dt) = p(O)(N-O) - eO \]
While the length of this "ecological memory" (Padisák, 1992) is not certainly quantified, we may be aware of the fact that it is generally the rarer species that contribute most to the total biodiversity of planktonic organisms. Whereas the metabolically significant planktonic biomass relies on the productivity of a relatively small number of species in a series of variable and renewable habitats, the majority of species is limited to habitats in which temporal variability is less extreme or exclusion is long protracted.

VIII. CONCLUSIONS AND IMPLICATIONS

There is scarcely a short answer to the question, "How is biodiversity maintained in the plankton?" The available evidence is that there is an operative blend of those mechanisms hypothesized to underpin diversity in other ecosystems. Global species richness is assisted by protracted isolation of populations and endemism of subsequently differentiated species to such locations as Lake Iyikal and small lakes in Australasia. Progressive functional adaptations and niche specialization among, for example, the photobacteria, the cyanoprokaryotes, and the planktonic crustacea contribute to the pervasion of selected species into sites offering their preferred habitat conditions for long time periods; even when those sites are sometimes mutually remote. However, the majority of species seem to be reliant on the periodic rejuvenation of a broad range of habitat conditions and the opportunity to fill, mainly noncompetitively, the spare capacity that is thus regenerated. On the basis that opportunities are simultaneously closing, there is an implicit dependence on the maintenance of viable seed banks and the ability to combine this either with efficient dispersal or with high survivorship. For the more familiar and (supposedly) more cosmopolitan planktonic forms, relative ease of dispersal is, manifestly, the dominant factor upholding the widespread occupancy among suitable habitats. For the majority of known species that are relatively rare (that is, there is a relatively low occupancy among the total number of patches), effective dispersal may be resisted by specific traits (T,.) or simply by low numbers. In this case, survival may be considered more tenuous. However, from the limited investigative evidence available, many of these rarer species have been numerous in the past and occasionally continue to perform relatively well on the inertia of persistent, viable propagules. While the length of this "ecological memory" (Padisák, 1992) is not certainly quantified, we may be aware of the fact that it is generally the rarer species that contribute most to the total biodiversity of planktonic organisms. Whereas the metabolically significant planktonic biomass relies on the productivity of a relatively small number of species in a series of variable and renewable habitats, the majority of species is limited to habitats in which temporal variability is less extreme or exclusion is long protracted.

This last distinction permits the differentiation of invasive and accumulative life-history strategies. They have equal merit because they exploit different frequencies and intensities of environmental fluctuation. Indeed, the full range of conservative mechanisms may be considered to have evolved to maximize the exploitative opportunities provided by the variability within and between the available habitats. Consequently and conversely, it is the variety of dimensions and scales of variability that upholds the diversity of species. The nearest that we can get to a short answer to the biodiversity question is that it is the means by which living systems cope with a nonequilibrium world.

The feature distinguishing plankton-based ecosystems from terrestrial ones is that the relevant temporal scales are mostly much shorter in open waters than on the land. Not only are the mechanisms of planktonic biodiversity conveniently observed but, provided that appropriate scaling factors are applied, the study of planktonic systems may be rewarded with insights into current issues about the importance and the protection of biodiversity.

Of the many propositions that have been advanced about the role of diversity in upholding ecosystem processes, current attention focuses mainly on four broad postulates (Lawton, 1997). In various ways, they relate to Darwin's (1859) proposition that a large number of species imparts a higher level of functional stability than does a small one. Thus, to lose species inevitably impairs the functional integrity of the ecosystem. The "redundant species hypothesis" counters that, because species are not equally represented, some are contributory much more than others to ecosystem function. Indeed, some "keystone species" (Paine, 1969) drive the main functions and have a much stronger influence than "passenger" species (Walker, 1992). Then, logic suggests, a minimal diversity is essential to adequate ecosystem functioning and most of the species are really redundant in their roles (Lawton and Brown, 1993; Walker, 1992). Almost diametrically opposed to this view is the "rivet hypothesis" (Ehrlich and Ehrlich, 1981), which accords to each species an essential contributory role, and which, if lost from the whole, like rivets lost from the structure of an aircraft, quickly lead to serious functional impairment and failure. A third hypothesis takes a much looser view, suggesting that function is modified by changes in the richness of species composition but in unpredictable ways ("idiosyncratic"; Lawton, 1994) because the contributions of individual species to system function are unequal. The fourth possibility is the null hypothesis that ecosystems are insensitive to changes in species composition. This
seems increasingly to be implausible and will not be discussed.

One of the important lessons from the plankton scale is that a distinction should be made between structural stability and functional stability. Taking the basic function of the plankton to be the material cycling of matter in open water, driven by solar energy invested in carbon bonds, the photoautotrophic, heterotrophic and phagotrophic roles are fulfilled, respectively, by any or all of a large number of species of phytoplankton, bacterioplankton, and zooplankton. Despite having vastly different and frequently changing species compositions, studies of the productivity of planktonic systems demonstrate a remarkable level of functional coherence (Schindler, 1990). In this way, pelagic fish may continue to feed planktivorously and to respire photosynthetically generated oxygen almost without reference to the planktonic structure, just so long as similar functions are maintained.

At the level of planktonic communities, a decline in the population of a dominant species will carry fewer implications if a second species is poised to substitute quickly. Frost et al. (1995) made the case that species complementarity made for functional compensation of the smaller scales of structural variability. The case is analogous to the arguments about the flexibility of communicating information through networks rather than along a single pathway (Pahl-Wostl, 1990), or what has been called the ‘‘World Wide Web’’ explanation of biodiversity (Reynolds, 1997b). This recognizes the idea that some species may contribute less to overall ecosystem function than other, more functionally creative keystone species (following Paine, 1969), or ‘‘ecosystem engineers’’ (Lawton, 1997), and by providing opportunities to others may also serve to promote and maintain the species richness. It also values the role of the reserve (richness) of complementary species able to fulfill the compensatory function. The versatile airplane still relies on having most of its rivers in place.

What clues does planktonic diversity give us about the approaches to upholding species richness in other ecosystems? The advice ‘‘Keep up the disturbance’’ is a little trite and somewhat counterintuitive to traditional conservation attitudes. Only at the global scale of aquatic systems is it possible to conceive the simultaneous existence of sufficient a range of habitats to accommodate the entirety of species. No single location, no matter how variable, offers habitat or refuge for more than a minority representation of the aquatic biota. Aquatic biodiversity has to be pursued at the scale of fluvial catchments and regional seas—every lake, pond, evolving flood plain, river estuary, coastal shelf, and ocean circulation is but a piece in the biospheric mosaic. The essential general aim should be to ensure that suitable habitat elements persist, each in a range of developmental states. Local extinctions are resisted by good between-locality communication of propagules. In this sense, managing biodiversity has to adopt the philosophies of patch dynamics, taking full account of the longevities and the invasiveness of species-specific populations in relation to the externally influenced availability and renewal of suitable patches.

See Also the Following Articles

BACTERIAL BIODIVERSITY • ESTUARINE ECOSYSTEMS • LAKE AND POND ECOSYSTEMS • OCEAN ECOSYSTEMS • REEF ECOSYSTEMS • RIVER ECOSYSTEMS

Bibliography


See Also the Following Articles

BACTERIAL BIODIVERSITY • ESTUARINE ECOSYSTEMS • LAKE AND POND ECOSYSTEMS • OCEAN ECOSYSTEMS • REEF ECOSYSTEMS • RIVER ECOSYSTEMS

Bibliography


PLANKTON, STATUS AND ROLE OF


PLANT–ANIMAL INTERACTIONS

Ellen L. Simms
University of California Botanical Garden

I. Types of Interactions
II. Antagonistic Interactions
III. Mutualistic Interactions
IV. Summary

GLOSSARY

Allee effect For population size to be regulated, it must exhibit a negative density dependence. That is, the population growth rate must decline as the population gets larger. However, under certain circumstances, some populations exhibit positive density dependence. This phenomenon, in which population growth rate increases as the population gets larger, is called an Allee effect. An Allee effect may generate a critical minimum population size, below which extinction will occur.

angiosperm Flowering plant. A lineage characterized by flowers, seeds enclosed in carpels, specialized conducting elements in the phloem (sieve tube members) and xylem (vessels), presence of endosperm, double fertilization, and tectate pollen.

diaspore A plant part distributed by dispersal, regardless of its developmental and morphological origins. A diasporae may be a naked seed, a seed enclosed in a fruit, or many seeds enclosed in a fruit. It may also mean bulbs or lengths of rhizomes. A good synonym is propagule.

fitness Relative contribution of offspring to the next generation. An individual, genotype, or phenotype whose progeny constitutes a large proportion of the succeeding generation has high fitness.

granivores Animals that eat seeds or achenes (grass fruits).

herbivores Animals that eat plants. Usually excludes instances when a single animal eats an entire plant, which is categorized as predation.

phylogeny The evolutionary relationships among taxa (groups of related organisms), often portrayed with some kind of branching diagram, with branches representing speciation events. Typically, the true phylogeny of a group is hidden deep in the past and evolutionary biologists must infer relationships. Various types of data and methods of analysis are used in this effort and there is considerable contention among groups of scientists as to which are most likely to estimate the true phylogeny.

symbiosis Interaction in which two organisms live in close proximity. Symbiosis can be antagonistic or mutualistic. Often, the larger individual is called the host, and its inhabitant is called the guest.

trophic level Position of a species in the food web (Fig. 1). Plants—autotrophs that convert solar energy to chemical energy and utilize mineral nutrients—constitute the first trophic level. Primary consumers—animals that feed on living plants—constitute the second trophic level. The third trophic level is composed of secondary consumers—animals that feed on primary consumers. Predation, parasitism, grazing, and herbivory are intertrophic level interactions; competition occurs among species in the same trophic level.
FIGURE 1 Trophic levels within a food web. Each trophic level is comprised of one or more species that consume individuals or resources at the next lower trophic level and are, in turn, consumed by species in the next higher trophic level. Arrows indicate the direction of flow of energy and resources in the food web.

BECAUSE PLANTS CAN OBTAIN nourishment and energy from inorganic sources, they are the foundation of most biotic communities. Animal consumption of plants is the primary conduit by which energy and resources enter the food web. However, the wide-ranging effects of plant–animal interactions on biotic diversity extend far beyond simple trophic links. This article will summarize the various types of plant–animal interactions and use a multidisciplinary approach to examine their implications for biotic diversity.

I. TYPES OF INTERACTIONS

Biotic interactions can be categorized by their effects on the interacting parties (Table I). An interaction may not affect a species, or may be beneficial or detrimental. Antagonistic interactions negatively influence one or both species. Some interactions are clearly antagonistic: When a vole consumes an oak seedling, the rodent benefits while the plant dies. In other cases, it seems clear that both parties are benefiting, as when a hummingbird obtains nectar while transporting pollen from one plant to another. These interactions are termed mutualistic. The implications of other interactions may be more ambiguous. For example, Clark’s nutcrackers consume pine nuts but are also important agents of seed dispersal, and it is unclear whether the interaction is mutually beneficial to both species or whether one species is benefiting at the expense of the other. The fitness costs and benefits to the parties involved in such an interaction may be conditional on the current environment. For example, during years of heavy seed production, birds may provide plants with a net benefit, whereas during poor seed years, the net effect of birds on the plant may be negative.

Ecologists seek to categorize interactions because they have distinct ecological and evolutionary implications. Within biotic communities, certain kinds of interactions increase species diversity whereas others reduce it. Similarly, the nature of interactions determines their impact on genetic diversity within populations.

II. ANTAGONISTIC INTERACTIONS

A. Plant Consumers—Herbivores and Granivores

1. Types of Plant Consumers

Animals consume plants in all kinds of habitats, including marine, terrestrial, and freshwater, and do so in wildly diverse ways. For some plants, there is at least one animal species devoted to consuming each type of organ. Some animals remove tissues by chewing, others suck plant sap. Grazing individuals feed from many different individual plants. Some animals live within plants, literally surrounded by food. These include borers, gallers, and leaf miners.

Many animals chew on plants, just as we do. However, another important mode of consumption is to use strawlike mouthparts to pierce and suck fluids from vascular structures such as xylem and phloem, which transport water, minerals, and other compounds throughout the plant. Aphids are common sap feeders. Spider mites are another type of plant-sucking arthropod that may be familiar to unhappy owners of house plants. The origin of this feeding mode probably extends back to the Carboniferous with the Paleodictyopteroidea, an assemblage of insects with sucking mouthparts.

Sap feeders share an interesting problem in common with many blood feeders: Neither blood nor plant sap provides a balanced or complete supply of vitamins

<table>
<thead>
<tr>
<th>Plant effect on animal</th>
<th>Animal effect on plant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herbivory</td>
<td>+</td>
</tr>
<tr>
<td>Carnivory</td>
<td>-</td>
</tr>
<tr>
<td>Mutualism</td>
<td>+</td>
</tr>
</tbody>
</table>

TABLE I Pairwise Ecological Interactions between Plants and Animals Categorized by Their Effects on the Fitness of Each Party
and amino acids. One solution to this problem among aphids has been to host intracellular microbial symbionts with the enzymes necessary to convert common nonessential amino acids into essential rare ones, much as Midas converted base metals to gold (Douglas, 1994). These intracellular guests inhabit special cells near the gut and oviducts and are transmitted by the mother to her eggs.

Perhaps the best known mode of plant consumption is grazing. All of us are familiar with picturesque scenes of cows grazing or deer browsing in verdant pastures. Far from peaceful, these are scenes of graphic violence. Hundreds, perhaps thousands, of plants are being eaten alive! Each grazer is eating photosynthetic organs from numerous living plant individuals. However, because the aboveground parts of most plants are constructed of repeating, renewable modules called shoots, a single episode of grazing or browsing rarely kills a plant.

Less well-known grazers include parrot fish, which maintain well-mown lawns of algae on coral reefs. When portions of the reef are protected from parrot fish, luxuriant algal growth can smother the coral. Sea urchins graze kelp forests and, when their natural predators are eliminated, may create vast barrens on the ocean floor. On land, many plant-eating insects, such as grasshoppers, katydids, and some beetles, are grazers. Some grazers feed on roots. These include tiny soil insects called springtails, relatives to silverfish, which move around in the air spaces between soil particles, nibbling on roots. Microscopic crustaceans in the water columns of oceans and freshwater lakes and ponds graze on tiny photosynthetic single-celled organisms called phytoplankton. In this case, however, the animals function more like predators because they must kill their prey to eat it.

In contrast to grazers, which move from plant to plant and eat only a portion of each plant, some animals feed entirely on one plant during their lifetime. In some cases, many generations of the consumer will occupy the same tree, evolving greater and greater specialization to that one individual. The rate of evolutionary change in a lineage is a negative function of generation time. Organisms such as long-lived trees, which have very long generation times, are at a distinct evolutionary disadvantage relative to their short-lived pests, which may be able to evolve very quickly because they go through multiple generations every year. Indeed, an important unanswered question in evolutionary biology is how long-lived plants such as trees avoid being destroyed by rapidly evolving pests.

Some consumers live inside their food. Perhaps the most intriguing are gall-forming arthropods—wasps, flies, and mites that induce plants to form elaborate structures within which the animal munches away on the host. Galls occur in many forms, including small red swellings on leaf blades, large globular swellings on branches, or bristly structures in odd places. In most cases, the female insect induces the gall when she injects her eggs into the plant tissues. Her young hatch and feed inside the gall, where they are relatively protected from enemies and the rigors of the physical environment. The relative frequency of gall-forming herbivores increases with increasing aridity of the environment, presumably because insects in galls are less vulnerable to desiccation. Plant-feeding nematodes, called vinegar eels, often live in galls on roots. The earliest known gall was produced by an insect feeding on tree fern fronds in the late Carboniferous.

Experiments show that the same species of gall-inducing insect produces very different looking galls on different plant species, suggesting that the plant determines gall form. Detailed work on the goldenrod ball gall, however, indicates that both insect and plant genes interact to determine gall size (Abrahamson and Weis, 1997). Although these insects appear to successfully manipulate plant development, which is a goal of many biotechnologists, relatively little research has been devoted to understanding the molecular mechanisms of this fascinating interaction.

Another group of consumers that lives inside plants are leaf miners. As their name suggests, the larvae of these flies and moths burrow within a leaf like a miner, eating the tissue they excavate and leaving behind their waste material, called frass. Some leaf miners snake along, leaving a long serpentine mine in their wake. Others remain in one place, creating a large blotch in the leaf. In many cases, leaf miners can be identified to species on the basis of host plant and the appearance of the mine. This has also made leaf miners easy to detect in the fossil record. Because they must live between the top and bottom surfaces of leaves, leaf miner larvae are very flattened from top to bottom. Some leaf miners pupate within the mine; others exit the mine and pupate in the soil beneath the host plant. In addition to leaf miners, some other insects mine plant stems, burrowing just beneath the stem surface and leaving similar-looking feeding galleries.

Bark beetles are economically very important plant consumers (Paine et al., 1997). Adults attempt to bore into tree trunks, but healthy trees generally can “pitch” them out, literally flooding the bore holes with sap. However, the sap attracts more adults of these gregarious beetles and an ailing tree’s defenses can be quickly overcome. Female beetles then excavate extensive galleries just beneath the bark and lay their eggs in the termini of the tunnels. The larvae feed on both the host
wood and tree-feeding fungi that adults transport in special structures called mycangia. Often, bark beetles do not kill their host but are indirectly responsible for its death because they introduce deadly fungal pathogens.

2. Evolutionary Responses by Plants

Plants have evolved numerous responses to their consumers. Primary among these is resistance. Plants resist consumers in three main ways. They may defend themselves chemically or mechanically, or they may escape damage by being difficult for consumers to discover.

a. Chemical Defenses

Plants are highly proficient chemists. There has been ongoing controversy as to why plants produce such a startlingly diverse array of chemicals, which, because they had no known function in primary metabolism, became known as secondary compounds. Early theories proposed that these compounds were waste products from “pathological overproduction of carbon.” But this explanation begged the question, “Why so many ways of throwing out the trash?”

By the 1950s phytochemists had begun speculating that secondary chemicals were important in plant defense, and this theory has dominated the latter half of the twentieth century. Nevertheless, competing hypotheses have held their own—notably that secondary chemicals are important in protecting plants from physical dangers such as ultraviolet light. Undoubtedly some plant chemicals do perform such functions, but it is certainly not clear why plants should have so many kinds of sun screen. Perhaps the most defensible alternative hypothesis is that secondary compounds are important in defense against microbial enemies such as fungal, bacterial, and viral pathogens. It is quite likely, in fact, that many compounds defend plants against both these agents and animal consumers.

Chemical defenses are divided into three categories: digestibility reducers, toxins, and repellents. Many plants produce digestibility reducers such as tannins, which are high molecular weight carbon-rich compounds that bind proteins and make them difficult to digest. Thus, these compounds do not directly kill the consumer, but make the plant less nutritious to eat. This indirect mechanism of defense leads to important questions about the selection pressures that might have led to their evolution. In particular, it is not always clear that individual consumers know the nutritional value of their host. Moreover, in many cases the individual that chooses the plant is not the one destined to feed on it. For example, when choosing among a population of the same plant species, female butterflies do not necessarily lay their eggs on the individuals on which their larvae can develop best. Another question is, “If the chemical does not kill the consumer, why would it be evolutionarily advantageous to plants possessing it?” This question becomes even more vexing in light of considerable evidence that, when protected from other causes of mortality, consumers can compensate for poor food quality by consuming more (not less!) plant tissue. The leading hypothesis to explain this paradox is that consumers feeding on a plant with a digestibility-reducing compound will grow more slowly, making them more vulnerable to enemies, such as predators, parasites, or diseases, and therefore die before consuming much plant tissue.

Toxins, as their name implies, are poisons that have direct negative effects on animals that consume them. Some plants are so toxic that apparently no animals will eat them. However, in many cases certain animals have evolved mechanisms that detoxify extremely potent poisons. Often, these animals are specialists on that host plant. There may be strong evolutionary benefits to feeding on such a previously unexploited resource. The most obvious benefit is that no one else is using the plant, which reduces competition for food. However, because a toxic plant harbors so few consumers, it may also provide “enemy-free space” in which consumers are less likely to be discovered by predators or parasites. Enemy-free space could also arise by a slightly different mechanism, which is nicely illustrated by small consumers on seaweeds. Small crustaceans called amphipods live on the algae they eat. Parrot fish also eat algae, but when they discover an amphipod, they snap it up like candy. Thus, amphipods that feed on very toxic species of algae are less likely to be consumed by herbivorous fish that are opportunistically omnivorous.

Perhaps because of these kinds of benefits, many animals that have evolved the ability to consume a toxic plant have also evolved specialization to that host. For example, specialist herbivores may use the toxic compound as a cue to find their host. Additionally, the compounds may stimulate feeding or egg laying. This level of specialization sets up an important evolutionary trade-off for toxic plants. A compound that previously killed all consumers has become an attractant to at least one consumer. Any individuals that have higher levels of the compound may be better defended against most consumers but may be more attractive to the specialist. Conversely, individuals with lower toxin concentrations will be less attractive to the specialist but may become vulnerable to other generalist consumers. This type of situation may create stabilizing selection and cause the plant population to evolve an intermediate
caterpillar to survive to adulthood. European starlings *Sturnus* will preferentially consume poison hemlock (*Conium*). But wooly bears infested with a lethal parasitoid avoid poison hemlock, the plant used to execute Socrates. For example, healthy wooly bear caterpillars secondary compounds (Rodriquez and Wrangham, 1993). For example, healthy wooly bear caterpillars secondary compounds from their food and use them for their own defense. The best known example being the monarch butterfly, which harbors cardiac glycosides from its milkweed hosts. Some animals even self-medicate with secondary compounds (Rodriguez and Wrangham, 1993). For example, healthy wooly bear caterpillars avoid poison hemlock, the plant used to execute Socrates. But wooly bears infested with a lethal parasitoid will preferentially consume poison hemlock (*Conium maculatum*), which can kill the parasite and allow the caterpillar to survive to adulthood. European starlings may protect their nestlings by lining their nests with fresh plant materials that inhibit arthropod hatching and bacterial growth. Animals may also use plant secondary compounds to preserve their food. Pikas, relatives of rabbits, live in burrows in talus slopes. To survive the winter in their alpine homes, they harvest and store enormous quantities of vegetation, which they store in haypiles. Pikas prefer to eat grass hay, but they often will harvest toxic herbs as well. Haypiles with these herbs are less likely to become moldy than those without them.

i. Animal Uses of Plant Defensive Compounds
Whatever the evolutionary reasons for these compounds, plant secondary chemicals have enormous value to humans. For example, synthetic pyrethroids, which were originally extracted from a species of chrysanthemum, are valuable insecticides because they are effective but degrade quickly and so do not accumulate in the environment. Plant secondary chemicals are the source of virtually all of the herbs and spices that make food interesting to eat. In many cases, these compounds were first important in preserving foods and keeping them safe to eat. Finally, plant secondary compounds are an important source of pharmaceuticals. In fact, 23% of the modern medical drug prescriptions (119 different chemical substances) written between 1959 and 1980 in the United States were pharmaceuticals derived from 90 different plant species (Farnsworth et al., 1985).

Other animals also use plant secondary compounds in interesting ways. Many consumers sequester plant compounds from their food and use them for their own defense, the best known example being the monarch butterfly, which harbors cardiac glycosides from its milkweed hosts. Some animals even self-medicate with secondary compounds (Rodriguez and Wrangham, 1993). For example, healthy wooly bear caterpillars avoid poison hemlock, the plant used to execute Socrates. But wooly bears infested with a lethal parasitoid will preferentially consume poison hemlock (*Conium maculatum*), which can kill the parasite and allow the caterpillar to survive to adulthood. European starlings may protect their nestlings by lining their nests with fresh plant materials that inhibit arthropod hatching and bacterial growth. Animals may also use plant secondary compounds to preserve their food. Pikas, relatives of rabbits, live in burrows in talus slopes. To survive the winter in their alpine homes, they harvest and store enormous quantities of vegetation, which they store in haypiles. Pikas prefer to eat grass hay, but they often will harvest toxic herbs as well. Haypiles with these herbs are less likely to become moldy than those without them.

b. Mechanical Defenses
Mechanical resistance to consumers may be obvious, as in the case of spiny cacti and thorny shrubs, or more subtle, as in the case of silica bodies that render grass leaves less palatable to many consumers. Some plants combine both chemical and mechanical defenses. For example, Wright's datura possesses both toxic alkaloids and leaf hairs, called trichomes. Moreover, some genotypes possess simple hairs (mechanical defense), whereas others have a sticky surface provided by glandular hairs that excrete a sticky exudate (mechanical and chemical defense). The sticky hairs defend plants against whiteflies but not mirid bugs; plants with simple hairs are better defended against the latter pests. It is generally supposed that mechanical defenses are more difficult for consumers to overcome.

When animals consume other animals, they eat tissue that has a composition similar to their own. However, plant tissues are generally richer in carbon and poorer in nitrogen than animal tissues. In part because of mechanical defenses, a high proportion of the carbon in terrestrial plants is devoted to structural molecules such as cellulose and lignins. Animals lack the enzymes necessary to digest these compounds. Microbes do have such enzymes, however, and herbivores sometimes have elaborate modifications of their guts that house microbial symbionts that can digest the fibrous fare.

c. Escape
In addition to defending tissues, plants may also escape consumers in time or space. Ephemeral plants, especially those with annual life cycles, may obtain temporal escape with a life cycle that does not sustain consumers long enough for them to complete their life cycles. Even if short-lived plants do not starve out their consumers, they may limit them to only a single generation. This constraint prevents the buildup of dense pest populations which short-lived consumers can attain on long-lived plants. Escape may also be achieved by appearing only during seasons when consumers are rare. For ex-
ample, many cool-season plants are relatively unaffected by insect consumers, which are far more abundant during warm weather. Highly dispersed plants may also escape consumption if the distances between them are greater than the average distance traveled by their consumers. This mechanism may be in part responsible for the astonishing diversity of certain tropical forests, in which only one individual of each tree species will be found in a large area. This concept is also embodied in the “resource-concentration hypothesis,” which states that dense concentrations of host plants will harbor the highest densities of consumers. If a plant species is both short-lived and rare, it may be so difficult to find that it can complete its life cycle before being found by consumers.

d. Tolerance and Compensatory Growth

Even if plants are damaged by consumers, they may evolve mechanisms that allow them to maintain fitness in the face of damage. Plants may tolerate damage through various compensatory mechanisms, including reallocating resources from undamaged plant parts to replace damaged tissues. Resources are usually allocated among plant parts in response to gradients between points of production (sources) and points of use (sinks). Thus, leaves, which produce photosynthate, typically function as sources, sending resources to metaxematic sinks where new growth is occurring. However, if consumers damage actively photosynthesizing leaves, the area around the damaged leaves may shift from source to sink, thereby attracting resources for compensatory growth. Tolerance of consumers may be especially important among fast-growing plants living in resource-rich environments.

3. Coevolution of Plants and Herbivores

In 1964, Ehrlich and Raven conceived a coevolutionary hypothesis to explain the magnificent diversity of plant chemistry. They postulated that herbivorous insects are a strong mortality agent for plants and that if any trait arose that protected a plant from its herbivorous insects, that trait would quickly spread in the plant population. Further, they argued that escaping its herbivores would create for the plant the opportunity for a period of rapid evolution. They postulated a similar process for the insects. For the herbivores, the newly evolved plant species represent unused resources. Any insect trait that allowed an insect to exploit these plants would likewise result in an adaptive radiation of herbivores. As Janzen later argued, the crucial characteristic of this process that distinguishes it from ordinary evolution by natural selection is its reciprocal nature. The plant evolves resistance to the herbivore, the herbivore then evolves a mechanism that negates the resistance, after which the plant evolves resistance to the herbivore, and so on, ad infinitum. With every crank of the coevolutionary process, new species arise through adaptive radiation.

This hypothesis has excited considerable controversy. Some authors argue that reciprocal coevolution is rare or nonexistent because insect herbivores do not impose sufficiently strong selection pressures on plants. Others argue that herbivores experience more selection from their natural enemies than from plants and that moving to new plants is driven by the adaptive advantage of enemy-free space. Further, Labandeira and Sepkoski have pointed out fossil evidence that indicates that the great radiation of modern insects began 245 million years ago and was not accelerated by the expansion of angiosperms during the Cretaceous period. However, neither insects nor plants have stopped evolving, and currently evolving systems provide the best tests of the coevolutionary hypothesis.

One way that scientists have tried to test the coevolutionary hypothesis is to compare the evolutionary lineages (phylogenies) of host plants and insect consumers that narrowly specialize on that group of hosts. Coevolution can lead to a pattern called cospeciation, in which the two phylogenies match, much like your fingers do when you place your palms together, fingers up. One hand represents the plant lineages, the other the insects. As the base of your palm is the ancestral species, the fingers represent various derivative lineages. An alternative phylogenetic pattern, in which lineages do not match, is produced by host switching. Host switching is the phenomenon whereby specialist consumers shift host species and then speciate on that new species, without any speciation by the host plant. The alliance of Hawaiian silverswords and the plant hoppers that live and feed on them provides a particularly exciting pair of phylogenies with which to test the coevolutionary hypothesis. The Hawaiian silverswords are derived from a pair of ancestral species in the rather prosaic group of California plants called the tarweeds. Over the past 5 million years the Hawaiian descendents have rapidly radiated into a stunningly diverse array of species. Some of these Hawaiian descendents are magnificent rosette plants with life histories much like century plants—they live long lives terminating in the production of giant flowering stalks. Others are multi-branched perennial shrubs. Most are attacked by members of the plant hopper genus, Neosydne, in the family Delphacidae. The plant hoppers are highly host specific;
each species feeds on one or a few closely related plant species. Phylogenetic analysis of molecular data from the plant hoppers and their hosts reveals a pattern of cospeciation: Each plant hopper species is most likely to use the host species that is most closely related to the host of its most close relative. This pattern is exciting, because it is expected to arise from reciprocal coevolution.

However, other mechanisms could produce the same pattern. For example, the plants could be evolving in response to some other selective pressure, with the insects following along behind. Matching phylogenies can also arise when both an insect and its host plant might speciate simultaneously in response to some external event as, for example, when they become geographically isolated from their main populations by a geologic event such as mountain building. Thus, phylogenetic comparisons alone cannot deduce whether coevolution has occurred. Other types of data must be examined. Careful observation of host use across hybrid zones suggests that the most likely explanation of the match between silversword and plant hopper phylogenies is that plant hosts are speciating in response to some external pressure and plant hoppers are tagging along behind plant host speciation.

4. Community and Ecosystem Effects of Plant–Consumer Interactions

Considerable evidence suggests that consumers can reduce the fitness of individual plants and thereby impose selection pressures that produce evolutionary changes in plant traits. However, it is less clear whether consumers influence plant abundance in the landscape. This issue is at the heart of a controversy that has raged among community ecologists for the past two decades over the relative importance to community structure of top-down control by consumers of their resource populations versus bottom-up control of consumer populations by resources at the base of the food web. The current view is that both types of forces interact in complex ways to structure biotic communities. However, several competing hypotheses aim to explain how these forces interact. Three of the major models are summarized here.

Donor control models predict that while plants are food to their consumers, consumers have little effect on plant abundance. Thus, biomass of organisms at one trophic level is a function of the productivity of their resource base at lower trophic levels. This means that adding resources to the base of a food web will trickle up the web, increasing biomass at all trophic levels. In contrast, consumer control models predict that each trophic level can be controlled by either its resources or its consumers, but not both. They further predict that the direction control moves depends on the trophic level being examined and the number of trophic levels in the ecosystem. In particular, plants, at the base of the food web, are expected to dominate in ecosystems with odd numbers of trophic levels whereas herbivores will dominate in ecosystems with even numbers of trophic levels. Thus, increasing abundance in a particular trophic level will cascade up and down the food web, alternately expanding or shrinking trophic levels. Finally, keystone predation models predict that the species composition at each trophic level modifies the relative effects of resources and consumers.

The last model is more complex than the previous two, incorporating aspects of each. It is of particular interest here because it incorporates information about the relative vulnerability to consumers (i.e., resistance) of different resource (e.g., plant) species. The keystone model predicts that species diversity in resource populations can be maintained if resource species exhibit trade-offs between their relative competitive abilities and their relative resistance to consumers. Further, the model predicts that when resources are scarce, consumer populations will be small and the plant community will be dominated by a few fast-growing, strong competitors that are highly vulnerable to consumers, which are therefore consumer controlled. Under nutrient-rich conditions, consumer populations will be large and the plant community will consist primarily of a few slow-growing, well-protected species that are resource controlled because of their heavy investment in resistance. At intermediate levels of nutrient availability, both types of plants can coexist because of the trade-off between competitive ability and resistance to consumers (Leibold et al., 1997). Consequently, species diversity will be greatest at intermediate levels of productivity. Other factors will determine whether these diverse communities are controlled by consumers (top down) or by resources (bottom up).

Leibold examined this prediction in planktonic communities of fishless ponds that varied in their level of mineral nutrient availability. These communities consist of photosynthetic planktonic algae (phytoplankton: single-celled green plants) that are grazed by herbivorous microarthropod zooplankton. Leibold found that algae in low-nutrient ponds consisted primarily of small, unprotected forms thought to be fast-growing but susceptible to grazing. Algae from more eutrophic (nutrient rich) ponds were larger and often sheathed...
or gelatinous forms thought to be slow-growing but resistant to grazers.

In another recent analysis, Chase and colleagues reviewed studies of temperate terrestrial grasslands to determine whether the effects of consumers on plant biomass fit the keystone herbivore model, which predicts that consumer control should be strongest at high levels of resource availability and decline with declining productivity. They reviewed the results of experiments that manipulated the presence or absence of large grazers. Because most temperate grasslands are water-limited, they sought evidence for a correlation between consumer effect and precipitation. As predicted, the proportional effect of consumers on plant biomass declined significantly with increasing precipitation. Schmitz found a similar relationship between plant productivity and the effect of insect herbivores on plant biomass. Further, Chase and colleagues found a turnover in species composition among plants along the precipitation gradient, as predicted by the keystone model.

### B. Plants as Consumers—Carnivorous Plants

The most ubiquitous interaction between plants and animals is the use by animals of plants as sources of material resources and energy. However, there are a few plants that turn the tables. In a world of plant-eating animals, carnivorous plants eat animals. To be considered carnivorous, a plant must have some mechanism to attract, capture, and/or digest prey and must be able to absorb nutrients from those prey (Givnish, 1989). Over 500 species in nine plant families have evolved the carnivorous habit.

#### 1. Mechanisms of Prey Capture

Carnivorous plants capture prey in several remarkable ways (Table II). The evolutionarily independent origin of carnivory is demonstrated by the many ontogenetic origins of the traps. A pitfall trap is a tubular structure, often containing liquid, which prey can enter but have difficulty leaving. Although five different plant families capture prey in some kind of pitfall trap, called tanks or pitchers, these traps may be comprised of leaf rosettes (e.g., *Brocchinia*), modified leaves (e.g., *Sarracenia*), or modified leaf tips (e.g., *Nepenthes*). Some plants have active mechanisms to trap prey. For example, the leaves of the Venus flytrap (*Dionaea muscipula*) function like a miniature steel-jawed trap when tripped by the hapless prey. There are even aquatic plants (e.g., *Utricularia*) with sophisticated underwater traps that slurp up prey unlucky enough to trip them.

#### 2. Costs of Carnivory

Carnivorous plants are usually restricted to sunny, moist, nutrient-poor habitats, such as bogs and fens. Their slow growth and restricted distribution suggest

<table>
<thead>
<tr>
<th>Genera (no. of species)</th>
<th>Attraction</th>
<th>Digestion</th>
<th>Type of trap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bromeliaceae</td>
<td>+</td>
<td>–</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Catopsis (1)</td>
<td>–</td>
<td>–</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Brocchinia (2)</td>
<td>+</td>
<td>–</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Eriocaulaceae</td>
<td>+</td>
<td>–</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Paepalanthus (1)</td>
<td>+</td>
<td>–</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Heliamphora (8)</td>
<td>Chemical and visual</td>
<td>–/–</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Darlingtonia (1)</td>
<td>+</td>
<td>–</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Sarracenia (5)</td>
<td>+</td>
<td>+</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Nepenthes (82)</td>
<td>+</td>
<td>+</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Cephalotus (1)</td>
<td>+</td>
<td>+</td>
<td>Pitfall</td>
</tr>
<tr>
<td>Dionaea (30)</td>
<td>–</td>
<td>+</td>
<td>Active Flypaper</td>
</tr>
<tr>
<td>Aldromida (1)</td>
<td>–</td>
<td>+</td>
<td>Steel trap</td>
</tr>
<tr>
<td>Dionaea (1)</td>
<td>+</td>
<td>+</td>
<td>Steel trap</td>
</tr>
<tr>
<td>Toxochytilus (1)</td>
<td>–</td>
<td>+</td>
<td>Passive Flypaper</td>
</tr>
<tr>
<td>Drosera (90)</td>
<td>+</td>
<td>+</td>
<td>Passive Flypaper</td>
</tr>
<tr>
<td>Cephalotus (1)</td>
<td>+</td>
<td>+</td>
<td>Passive Flypaper</td>
</tr>
<tr>
<td>Lentibulariaceae</td>
<td>–</td>
<td>+</td>
<td>Passive Flypaper</td>
</tr>
<tr>
<td>Pinguicula (35)</td>
<td>+</td>
<td>+</td>
<td>Passive Flypaper</td>
</tr>
<tr>
<td>Utricularia (280)</td>
<td>–</td>
<td>–</td>
<td>Bladder trap</td>
</tr>
<tr>
<td>Genlisea (35)</td>
<td>+</td>
<td>–</td>
<td>Lobster pot</td>
</tr>
<tr>
<td>Roridula (1)</td>
<td>–</td>
<td>+</td>
<td>Bladder trap</td>
</tr>
<tr>
<td>Byblidaceae</td>
<td>–</td>
<td>+</td>
<td>Passive Flypaper</td>
</tr>
</tbody>
</table>

* Modified from Gerzabek et al., 1984 (Gerzabek, 1989).

** Nutrient uptake apparently assisted by exudations of the kleptoparasitic bug *Pameridea roridulae.*
that there are fitness costs associated with carnivory. For example, it has been argued that leaves morphologically specialized for prey capture have compromised photosynthetic abilities, making carnivorous plants poor competitors. This hypothesis is supported by studies of the pitcher plant, *Sarracenia alata*, which produces two kinds of leaves, a “regular” leaf and one that is modified into a pitcher. The plant responds to competition from neighboring vegetation by diverting resource allocation from pitchers to regular leaves. Small stature and slow growth also make many carnivorous plants (e.g., *Pinguicula* and *Utricularia*) vulnerable to being buried by litter fall. Because of their compromised competitive ability, carnivorous plants generally respond poorly to addition of nutrients to their habitat, being easily outcompeted by plants that thrive under richer conditions. This characteristic creates important conservation concerns in the many parts of the world where atmospheric input of anthropogenic nitrogen is significantly increasing nitrogen availability in previously nitrogen-poor bogs and fens.

Another fitness cost of carnivory was identified by Zamora, who found that *Pinguicula vulgaris* tends to trap its own pollinators. He also found that reproduction in the plant is limited by pollen availability, indicating that feeding on its pollinators reduces plant fitness.

### 3. Benefits of Carnivory

As with so many plant novelties, these were noticed by the inquiring mind of Charles Darwin (1874), who demonstrated that the sticky traps of *Drosera rotundifolia* do indeed capture and digest animals. Darwin’s son, Francis (1878), first demonstrated experimentally that prey capture enhances the growth and reproduction of this species. Similar studies have subsequently found that in most circumstances, growth of carnivorous plants benefits from prey capture.

In many cases, the majority of a carnivorous plant’s nitrogen and phosphorus is obtained from prey. However, carnivorous plants appear to obtain only a small proportion of other necessary nutrients from prey. Comparisons of greenhouse and field studies suggest that plant growth is generally restricted by the rate of prey capture and that plants could utilize many more prey than they are able to catch. Carnivorous plants also tend to be frugal with their nutrients, practicing particularly efficient internal recycling of nitrogen and phosphorus.

Many elegant methods have been used to examine in more detail nutrient uptake from prey in carnivorous plants (Adamec, 1997). In particular, putatively carnivorous plants can be offered insects reared on media enriched in the stable nitrogen isotope $^{15}$N. If the plant tissues subsequently become $^{15}$N enriched, this indicates that their nitrogen supply has been supplemented by insect proteins. Using this method, Hanslin and Karlsson found that *Drosera rotundifolia* and several species of *Pinguicula* in a subarctic environment took up 20–45% of the nitrogen available in insect prey they were offered. Further, root uptake of nitrogen was stimulated by prey capture, an unanticipated additional benefit of carnivory. Experiments performed in glasshouse or laboratory environments generally reveal even greater uptake efficiencies.

Carnivorous plants show interesting developmental changes during maturation of their carnivorous organs. For example, using fluorescent dye tracers, Owen and colleagues found developmentally regulated bidirectional transport by leaf glands in the pitcher vine, *Neopenthes alata*. In mature leaves, the glands transport fluids directly from the pitcher fluid to the plant vasculature (internal plumbing system), apparently functioning in nutrient uptake. However, in immature, closed leaves, the glands secrete fluid from the vascular tissues into the pitcher, building up a supply of fluid in which to eventually trap prey. Galile and Chang examined developmental regulation of hydrolase expression in *Sarracenia purpurea* pitchers. Hydrolase is an enzyme involved in prey digestion. Hydrolase expression commenced when the pitcher first opened upon maturity, increased for several days, and would largely cease after 2 weeks without prey. However, adding prey-derived resources such as amino acids to the pitcher fluid could induce hydrolase expression in pitchers that had ceased expression due to lack of prey.

### 4. Non prey Guests—Inquiline Communities

Prey-digesting guests (inquilines) are very common in pitfall traps and provide for some carnivorous plants the sole means of benefiting from prey. In fact, the food webs of inquiline pitchers (e.g., *Sarracenia* spp.) have been the subject of numerous highly informative community and population ecology studies.

In one of the most arcane modes of nutrient uptake, *Roridula gorgonias* hosts a bug, *Pameridea roridulae*, that feeds on insects trapped on its sticky leaves. Although the plant has no known method of digesting its prey, stable-isotope studies indicate that it acquires $^{15}$N label from prey. It apparently derives nutritional benefit via exudations from the bug guest that has dined on the labeled prey. This example, however, highlights the vulnerability of insectivorous plants to kleptoparasites.
species, from 19 families, and up to 180 ant species have been described for about 250 plant regions alone, associations between plants and myrmecophytes have been recognized as widespread and ecologically important. In the neotropics and unusual, myrmecophytes (ant plants) are now recognized as the most developed myrmecophytes may also provide food in the form of specialized structures composed of lipids (Brevicarian bodies), proteins (Beltian bodies), glycogen (Multerian bodies), or some combination. In the most elaborate cases, plants may provide everything—room, board, and drink—to their ants.

III. MUTUALISTIC INTERACTIONS

In addition to antagonistic interactions in which one party feeds on the other, plants and animals may also interact in ways that can benefit both parties. Far from being pleasant affairs, however, mutualistic associations can be highly vulnerable to cheating, which often makes them evolutionarily uneasy truces between parties. The delicate evolutionary and ecological balances that can be achieved by these organisms are truly fascinating and lead to some common evolutionary issues.

One important question regards the degree of specialization between mutualists. Specialization is useful because it allows the evolutionary development of elaborate lock-and-key mechanisms that exclude cheaters and maintain the mutualism. However, specialization is an evolutionarily vulnerable position because extinction of one partner species can spell doom for the other. Further, in mutualistic interactions that must be reconstituted each generation, which is the case for all animal–plant interactions, specialization may doom individuals that cannot find the correct partner in the environment.

A. Plant-Protecting Ants and Plant-Feeding Ants

As described previously, plants have evolved a variety of mechanisms that defend them against herbivores. One of the strangest defenses, though, is provided when plants are guarded by ants. Once thought to be rare and unusual, myrmecophytes (ant plants) are now recognized as widespread and ecologically important. In many cases, these relationships appear to be quite casual. Visiting ants may rob nectar from flowers, but perform some guarding services in return. However, the term myrmecophyte is generally reserved for the more specialized case in which an ant colony resides in special structures provided by the plant.

1. Benefits to Ants—Costs to Plants

Tropical myrmecophytes display the most sophisticated development of this type of interaction. In the neotropical regions alone, associations between plants and myrmecophytes have been described for about 250 plant species, from 19 families, and up to 180 ant species from 5 subfamilies. These ant plants may provide ant housing in specially modified stems, hollow thorns, or specialized leaf pouches called leaf domatia. The trees or shrubs often feed ants with amino acid- or sugar-based solutions produced by extraloral nectaries. The most developed myrmecophytes may also provide food in the form of specialized structures composed of lipids (Brevicarian bodies), proteins (Beltian bodies), glycogen (Multerian bodies), or some combination. In the most elaborate cases, plants may provide everything—room, board, and drink—to their ants.

Food provision for ants may be quite costly to myrmecophytes and studies show that plants regulate production of food structures. For example, in Central American Cecropia (Moraceae) trees grown at intermediate nutrient availability, removing Mullerian bodies stimulates their production. In contrast, when Mullerian bodies accumulate, which would happen if ants were not present, plants cease production.

In many ant–plant interactions, however, there is an important third partner through which the ants obtain benefit: sap-sucking homopterans tended by ants. Ants derive benefit from these homopterans in two ways. They may keep “milk herds” of homopterans from which they obtain honeydew, or they may be in the “beef business” and eat the homopterans they tend. Rather than simply supporting ants, the plants in these situations must also support homopteran consumers. Homopterans increase plant risks as well. They are often important vectors of plant diseases. Further, colonization by ant queens is a relatively rare event; presumably colonization by both ant queen and homopterans would be even rarer.

The homopteran mode of ant benefit appears to provide few options for control by the plant. This problem is illustrated with the African myrmecophyte Loe- nardoxa africana, on which the same ant species (Apho- myrmex afer) may tend one or both of two different homopteran species. Gaume and colleagues found that homopteran identity influenced the costs and benefits to the plant of ant patrol. One homopteran, the pseudococcid, could support larger colonies of ants, leading to better plant defense. This homopteran was also more efficient at producing ant biomass, ants tending pseudococcids did not use other plant resources. However, when ants tended coccids, the other homopteran, they also used plant resources from the extraloral nectaries. Thus, when ants tended coccids, the only control that plants had over homopteran feeding was indirectly through nectar production. Plants that produced fewer nectaries supported fewer ants and fewer ants could tend fewer coccids. The plant could control pseudococ-
cid colony size more directly via domatia volume. Plants that provided a smaller total volume of the swollen stems used as domatia supported fewer pseudococcids and therefore fewer ants.

2. Benefits to Plants—What Motivates Ants?

Ants can benefit plants in three ways. First, they may patrol the plant and discourage or repel would-be herbivores. They also prune neighboring plants, thereby reducing plant competition for their host. Finally, some ants feed their host plant (myrmecotrophy).

Ants have frequently been observed killing and removing insect herbivores, and numerous experiments demonstrate the efficacy of this defense. For example, Fonseca observed four times as many herbivores on Tachigali myrmecophila plants from which he removed Pseudomyrmex concolor ants as on plants with intact ant colonies. Further, the daily rate of herbivory was about 10 times lower when ants were present, resulting in experimental plants without ants exhibiting about twice as much cumulative herbivore damage during the 18-month experiment. Leaf longevity was also substantially higher on plants with ants. It is interesting to note that these ants do not eat the herbivores they kill. Instead, they feed exclusively on catenococcid insects they tend inside the domatium, which is the hollow rachis of the compound leaf.

Arts are also effective deterents of mammalian herbivory. For example, the African myrmecophile Acacia drepanolobium possesses two kinds of thorns. The swollen thorns are domatia in which Crematogaster ants live and rear their broods. Stapley has shown that the unsowed thorns slow plant damage by browsing mammals, but that browsers may compensate by feeding longer. Ants were far more effective defenses. When a browsing mammal encountered and was stung by ants, it stopped feeding immediately and could not be induced to feed further on that tree.

A second benefit that patrolling ants may provide is in competition with neighboring plants. Ants will prune vines (lianas) and branches of neighboring trees, effectively preventing their host tree from being overgrown. The result of this vigilance is that the host tree occupies a dramatically open cylinder of space amid otherwise densely packed tree canopies. Although such pruning of neighbors clearly benefits the host tree, it also benefits the ant colony by reducing the number of directions from which it may be attacked by competing or predatory ants.

In certain circumstances, ants may harm their own host by pruning it rather than its neighbors. For example, Stanton and colleagues discovered a situation in which Crematogaster nigriceps so severely prunes its host tree, Acacia drepanolobium, that the tree cannot flower and is sterilized. In the habitat studied, four species of ants compete strongly for hosts, and C. nigriceps fares poorly in the violent conflicts over nest space. Instead of pruning neighboring trees, C. nigriceps prunes its own tree, apparently because it cannot prune neighboring trees occupied by competitively dominant ants. Indeed, careful observation of a large number of trees occupied by C. nigriceps revealed that these trees were always pruned in such a way as to avoid canopy contact with adjacent trees occupied by competing ant colonies. Canopy pruning of its own tree appears to be a defensive response by a C. nigriceps colony to competition with dominant ants that prevent it from pruning their trees.

3. Feeding Plants—Myrmecotrophy

Finally, a very different group of plants receives nutrients from the ants they house. These plants are also known as ant epiphytes. Epiphytes are plants that live on, but derive no nutritional benefit from, the branches of other plants. Myrmecotrophic epiphytes provide ant domatia in hollow or inflated roots, hollow rhizomes, or folded leaves. Ants then act as ‘mobile roots’ gathering food items for the nest, processing them, and then depositing the resulting waste and fecal matter within the plant. The best studied of these systems is Myrcyedula tuberosa (Rubiacae), an epiphytic shrub of Southeast Asia and northern Australia. This species also has elaiosome-bearing seeds, a typical feature of ant-dispersed seeds. The ants feed on these food bodies and then ‘plant’ the seeds along the walkways they create in the canopies of the trees their plant ‘homes’ inhabit.

B. Plant Pollinators

An important and obvious characteristic of plants is that, with few exceptions, they are rooted to the ground and cannot move. This poses a crucial problem for sexual reproduction: Immobile mates must exchange gametes, meaning that pollen must be moved to the ovule. Pollen can move passively with fluid flow in the physical environment. Wind pollination has been successful among the gymnosperms and water pollination is found among many aquatic angiosperms. However, as many allergy sufferers know too well, most of the pollen produced by wind-pollinated plants never reaches its intended target. Many land plants avoid this inefficiency by using animals as pollen vectors. Most
temperate angiosperms and almost all tropical angiosperms are animal pollinated.

Of course, animals will not move pollen around as a favor to the plant. An important evolutionary problem for plants, then, has been to attract pollen vectors. It seems quite clear that the need to attract pollinators has been a primary driver in the evolution of flower morphology. Animal-pollinated flowers often have large, brightly colored structures that function as “advertising” for the “goodies” available to the visitor. The rewards may be nectar, other more specialized chemicals, or even the pollen itself. As with any purveyor of delectables, the flower also must contend with thieves. For example, nectar-robining bees may drill through the flower wall and gain access to the nectar without transporting pollen. Other visitors, such as ants, may be too small to trip the elaborate pollen application mechanisms of some flowers.

The potential for cheating selects for specificity in pollinator attraction. Another advantage of specificity is that it can promote pollinator fidelity. Simply attracting a pollinator once is not sufficient. To ensure fatherhood, the plant must attract an animal that will visit other flowers of the same species. Moreover, to avoid inbreeding, the flowers must be on different individuals.

Thus, the twin needs to avoid exploitation by cheaters and ensure pollinator fidelity create strong evolutionary pressure for plant traits that promote pollinator specificity. As might be expected when sex is involved, response to this selection pressure has led many plants to develop elaborate and baroque relationships (Darwin, 1877; Grant and Grant, 1985). For example, male euglossine bees depend upon flowers of the euphorb and orchid families for fragrances that they convert to pheromones with which to attract mates. In another instance, male insects are tricked by orchids into thinking they have found a mate, when in fact all they have discovered is a cleverly shaped and scented mimic.

1. Role of Plant Pollinators in Plant Diversification

An important aspect of these highly specialized relationships is that they can prevent mating between individuals of otherwise very closely related populations; a phenomenon known as reproductive isolation. Evolution in a trait that promotes reproductive isolation can quickly lead to speciation. Traits that create the opportunity for rapid speciation by exploiting novel resources such as new pollinators are considered “key innovations.” Identifying such key innovations is a sticky problem in evolutionary biology. Circularity arises because the characteristic that best defines a group, and therefore allows it to be identified as speciose (having lots of species), is frequently also the character postulated to be the key innovation responsible for the radiation. One way out of this thicket is to identify a causal link between the putative key innovation and one of the processes that determines diversity.

a. Floral Nectar Spurs as a Key Innovation

One particularly persuasive example of a key innovation is floral nectar Spurs. These structures are critically involved in pollinator specialization because they hold nectar deep within the flower and make it available to only a narrow range of floral visitors. Animals must either be small enough to enter the spur or have sufficiently long and narrow mouthparts to sip nectar from the spur. Observing the 1½-in-long nectararies of a Madagascar orchid prompted Charles Darwin (1877, pp. 162–163) to predict correctly the existence of a moth with a sufficiently long proboscis to pollinate this fantastic flower. Nectar Spurs have evolved independently in several distantly related families and genera of flowering plants. They may be constructed from petals, sepals, or both, and genetic studies suggest that simple genetic differences can produce quite different shapes, which might favor different pollinators. Comparative studies suggest that, as would be expected were spurs involved in pollinator specialization because they hold nectar deep within the flower and make it available to only a narrow range of floral visitors. Animals must either be small enough to enter the spur or have sufficiently long and narrow mouthparts to sip nectar from the spur.

Detailed studies within genera also corroborate the putative link between nectar spur morphology and pollinator fidelity. Within two groups of orchids, experimental manipulation of spur length significantly decreased both pollinia removal by pollinators and fruit set. This observation demonstrates that spur morphology directly influences pollinator-mediated reproductive success. Another study examined the effect of spur morphology on pollinator attraction across a hybrid zone between two species of cumbines. Hybrids varied in floral characters, including spur length and orientation, and these morphologies differentially attracted either hummingbirds (primary pollinators of Aquilegia formosa) or hawk-moths (primary pollinators of A. pubescens), and thereby promoted reproductive isolation. On the other hand, the presence of a hybrid zone indicates that floral morphology has not prevented pollination “mistakes.”

b. Insect Pollination and the Angiosperm Radiation

The role of plant–pollinator interactions in reproductive isolation has also led to the much grander hypothe-
sis that insect pollination was a “key innovation” leading to the co-radiation of flowering plants (angiosperms) and anthophilous insects, which are those groups most involved in pollination, including certain bees and wasps (Hymenoptera), various families of flies (Diptera), and butterflies and moths (Lepidoptera). Fossil evidence suggests that the angiosperms diversified very rapidly, and many hypotheses have been advanced to explain this phenomenon. Like most hypotheses in the historical sciences, however, these have been very difficult to test. One necessary prediction of the insect-pollination angiosperm radiation hypothesis is concurrent diversification of the angiosperms and anthophilous insects.

Considerable controversy surrounds the dates of diversification of angiosperms and anthophilous insects. As early as the Carboniferous, seed ferns had large pollen that was probably too heavy for wind transport and may have been pollinated by paleodicotyptarian insects found in the same formations. The first direct evidence associating insects with plant pollen appears in the Lower Permian. However, the radiation of the insect groups that today are most strongly associated with angiosperm pollination probably occurred in the late Middle to early Upper Cretaceous, the period most commonly thought to have witnessed the radiation of the flowering plants. While not proving the codiversification hypothesis, these estimates at least do not rule it out.

2. Pollination Syndromes

Many plant species that share animal pollen vectors also share similar suites of floral traits, such as color, shape, symmetry, and scent, which appear to attract those kinds of animals. For example, hummingbirds fly by day, have excellent color vision, possess long beaks, and visit flowers while in flight. Correspondingly, hummingbird-pollinated flowers are day-blooming and tend to be brightly colored, often red, bilaterally symmetrical and tubular in shape, and frequently pendant. Floral traits also may correspond to the physiological needs of pollinators, as with hummingbird-pollinated flowers, which tend to produce copious sucrose-rich nectar that helps fuel the notoriously high metabolic rates of their pollinators. A number of these trait combinations, which are called pollination syndromes, are summarized in Table III (Howe and Westley, 1988).

Implicit in the concept of pollination syndromes is the assumption that floral evolution has been strongly entrained by interactions with specific classes of pollinators, leading to strong specialization. Specialization, however, may be an evolutionary dead end. A plant highly specialized and dependent upon one or a few species of pollinator is seriously vulnerable to extinction of its pollinators. Dependence upon specialized pollinators can also lead to Allee effects, which arise when populations become too sparse to persist. For example, individuals in small, isolated populations of the annual plant farewell-to-spring (Clarkia cocinea) were visited by so few pollinators that they could not produce enough seeds to replace themselves. If individual reproductive rates dip below replacement levels for very long, a population can dwindle to extinction.

Figs may provide a particularly impressive example of this problem. Fig trees are ecologically important components of tropical forests because the fruits they produce support frugivorous animals that are important seed dispersers for many other forest plants. Figs do not have large showy flowers to attract pollinators. Instead, fig reproduction is exquisitely dependent upon minute wasps that pollinate small flowers held tightly inside the closed fig. The wasp’s end of the bargain is met when the eggs it lays inside the fig hatch and its larvae feed on a few of the many developing seeds. Individual fig trees flower synchronously for a relatively short time but fig wasp populations cycle constantly. Maintenance of the wasp population requires that female wasps emerging from a fig find and lay eggs in figs of other trees of the same species that are flowering at times other than their host individual. Genetic evidence shows that, despite their minute size and short life span, these wasps routinely move pollen between trees 5 to 14 km apart. This interdependence argues that viable fig populations may require a higher density of trees than is generally assumed necessary for plants not involved in such a tight mutualism. The only study to test this contention, however, found that despite heavy forest fragmentation by humans, banyan tree (a type of fig) populations on the Cook Islands have not yet suffered from Allee affects.

As demonstrated in a comprehensive study of a group of Hawaiian shrubs by the laboratory of Weller and Sakai, plants can sometimes escape the grips of their pollinator addiction. Here, species derived from insect-pollinated ancestors have evolved wind pollination, apparently by passing through a transitional stage characterized by a relatively generalized pollination system. Indeed, several researchers have questioned the assumption that most plant–pollinator relationships are highly specific (Jordan, 1987; Waser et al., 1996). First, they argue that even specialization in insect pollinators is usually defined at the level of plant genus or family, not species. Further, insect species confined to
### TABLE III
Pollination Syndromes: Putative Characteristics of Flowers Associated with Particular Groups of Pollinators

<table>
<thead>
<tr>
<th>Animal</th>
<th>Opening time</th>
<th>Color</th>
<th>Odor</th>
<th>Shape</th>
<th>Symmetry</th>
<th>Nectar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entomophilous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beetles</td>
<td>Day/night</td>
<td>Dull or white</td>
<td>Fruity or amino</td>
<td>Flat or bowl-shaped</td>
<td>Radial</td>
<td>Often absent</td>
</tr>
<tr>
<td>Carnivorous dung flies</td>
<td>Day/night</td>
<td>Brownish or green</td>
<td>Fetal</td>
<td>Flat or deep; often traps</td>
<td>Radial</td>
<td>Rich in amino acids, if present</td>
</tr>
<tr>
<td>Bee flies</td>
<td>Day/night</td>
<td>Variable</td>
<td>Variable</td>
<td>Moderately deep</td>
<td>Radial</td>
<td>Hexose-rich</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Flat or broad tube</td>
<td>Radial or</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>bilateral</td>
<td></td>
</tr>
<tr>
<td>Hawkmoths</td>
<td>Night</td>
<td>White or pale green</td>
<td>Sweet</td>
<td>Deep, often with spur</td>
<td>Radial</td>
<td>Ample and sucrose-rich</td>
</tr>
<tr>
<td>Butterflies</td>
<td>Day/night</td>
<td>Variable, often pink</td>
<td>Sweet</td>
<td>Deep or with spur</td>
<td>Radial</td>
<td>Often sucrose-rich</td>
</tr>
<tr>
<td>Vertebrate-pollinated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bats</td>
<td>Night</td>
<td>Drab, pale; often green</td>
<td>Musty</td>
<td>Flat &quot;shaving brush&quot; or deep tube; often on branch or trunk; hanging, abundant pollen</td>
<td>Radial</td>
<td>Ample and hexose-rich</td>
</tr>
<tr>
<td>Birds</td>
<td>Day</td>
<td>Vivid; often red</td>
<td>None</td>
<td>Tube; often hanging</td>
<td>Radial or</td>
<td>Ample and sucrose-rich</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>bilateral</td>
<td></td>
</tr>
</tbody>
</table>

* Adapted from Howe and Westley, 1988.

One plant species in one geographical area often visit other plant species in other parts of their range (Thompson, 1994). They also argue that specialization by pollinators may be more a function of which plants are available. Short-lived insects are more likely to visit only one or a few plant species with which they temporarily co-occur whereas social insects, which have long-lived colonies and thus the opportunity to overlap with many plant species, often exhibit serial specialization on a large variety of plant species. For example, the only bee species on the Galapagos Islands is colonial and has been recorded visiting flowers of at least 60 plant species in 28 families. Finally, clustering of flowers into certain categories thought to be canalized through selection by pollinators may actually reflect physiological or morphological constraints in plants. For example, Chittka has argued that clustering of flower colors into particular narrow ranges may be a function of the physical and chemical constraints imposed by plant pigments rather than constraints imposed by the vision systems of different pollinators. If these arguments are true, the pollinator syndrome concept may be clouding our thinking about and study of plant-pollinator interactions. Researchers may be oblivious to, or fail to record, flower visitation by the "wrong" pollinators (Waser et al., 1996).

3. Competition for Pollinators
Pollinators are an important resource that plants may compete over. Whether plants compete for pollinators is determined by the factors that limit seed production. In many situations, plant reproduction is limited by mineral resources, in which case plants are unable to increase fruit set with increased pollinator availability. However, there are many examples of plants in which reproduction is limited by pollinator availability. Pollinator limitation has important implications for plant conservation. For example, prolific flowering by invading plant species may negatively impact native plant communities by depriving native species of pollination.
C. Fruit and Seed Dispersers

Many plants depend upon animal dispersal of seeds and fruits (diaspores). Unlike pollen dispersal, in which insects are major players, diaspore dispersal is dominated by vertebrates. Seed dispersal is important because plants generally cannot develop successfully in the shade of their mother. Seed dispersal may also be an important mechanism by which seeds escape the predator or pathogen populations that are well adapted to exploit them, having built upon and become adapted to their parents. Finally, seed dispersal is important for plants to colonize newly opened habitat.

1. Types of Dispersers
   a. Vertebrates
      i. Fish
         Perhaps the most unexpected and amazing diaspore dispersers are fish. This phenomenon has been most extensively studied along the Amazon River. However, it is likely to be important in many areas with extensive seasonal flooding. In the Amazon basin, where the timing of plant reproduction corresponds with the seasonal flooding, some fruits are adapted to passive dispersal by water (hydrochorous). However, even these fruits can derive facultative benefit from fish dispersal. Plants with heavy fruits or with seeds embedded within hard shells are apparently obligately fish dispersed. Fish differ in their efficiency as dispersers. Catfish are effective dispersers whereas characins are destructive and act largely as predators of all but the smallest seeds.
      ii. Mammals
         Many mammals are important seed dispersal agents. Primates and bats are the most important mammalian dispersers in tropical areas. Both types of animals can move quickly across the landscape, thereby dispersing diaspores long distances. Diaspore dispersal by bats is particularly important for forest regeneration after land abandonment in the neotropics. In temperate regions, diaspores may be dispersed by ungulates (e.g., antelopes, elephants, and zebras) and by many supposed carnivores. For example, black bears consume prodigious quantities of fruit, sometimes competing with humans for delectable berries (McCloskey, 1948).
      iii. Birds and Reptiles
         Birds are arguably the most important class of fruit and seed dispersers. The earliest known examples of animal-dispersed plant propagules include the fleshy seeds of cycad progenitors, which appear to have been consumed by ancient reptiles (Howe and Westley, 1988). Many dinosaurs were certainly important fruit and seed eaters and may have functioned as dispersal agents. However, except perhaps for birds, modern reptiles are only rarely important diaspore dispersers today.
         Birds can be hard on seeds. Beaks may break up the seed coat, rendering the embryo vulnerable to digestive acids and enzymes. Seeds may be ground up in the gizzard. However, the guts of frugivorous birds tend to be short and gentle (not highly muscular). In fact, many seeds require a trip through a bird's digestive tract to germinate successfully.
         Although large numbers of a broad range of birds feed on fruits, few depend solely on fruits. Even waxwings, perhaps the most specialized frugivores in the temperate region, also feed on insects when they are available. Nevertheless, fruits are an important resource with which many birds produce body fat prior to migration. Moreover, in the tropics, where seasonal constraints on fruit production may be weaker, several groups of birds depend almost exclusively on fruits (e.g., quetzals, toucans, and barbets).
   b. Invertebrates
      i. Ants
         The only major insect seed dispersers are ants. Myrmechorous seeds, those adapted to ant dispersal, often possess a starch- or lipid-rich body called an elaiosome attached to a tough and smooth seed coat that is difficult for ants to crack. Seed size is also constrained by selection by ants—large ants tend to carry larger seeds than small ants.
         In comparison with vertebrates, ants do not carry seeds very far. Nevertheless, ants can be important dispersers for many plants. Ants may either store seeds in the nest or remove the elaiosome and then discard seeds at the nest entrance or colony waste pile. Both locations tend to have well-aerated, nutrient-rich soil that can improve plant growth. Seeds collected by ants may also gain some protection from other seed predators through their association with active ant nests, which are generally avoided by most other animals.
         An important conservation issue in many areas has been loss of native seed-collecting ants to competition from invading Argentine ants (Iridomyrmex humilis), which are not strongly attracted to myrmechorous seeds. Red imported fire ants (Solenopsis invicta) do feed on seeds, though, and their arrival in an area may have a negative impact on myrmechorous plant populations by competitively excluding other more effective seed-dispersing ants.
      ii. Other Insects
         Occasionally seeds are dispersed by insects other than ants. For example, scarab beetles
### Characteristics of Fruits and Seed Dispersed by Different Animals

<table>
<thead>
<tr>
<th>Animal Dispersers</th>
<th>Color</th>
<th>Odor</th>
<th>Form</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>vertebrate dispersers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoarding mammals</td>
<td>Brown</td>
<td>Weak or aromatic</td>
<td>Indehiscent thick-walled</td>
<td>Seed itself</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>nuts</td>
<td></td>
</tr>
<tr>
<td>Arboral mammals</td>
<td>Yellow, white, green, or brown</td>
<td>Aromatic</td>
<td>Arillate seeds or drupes; often compound and dehiscent</td>
<td>Pulp protein, sugar, or starch</td>
</tr>
<tr>
<td>Bats</td>
<td>Pale yellow or green</td>
<td>Musky</td>
<td>Various; often hanging</td>
<td>Pulp lipid- or starch-rich</td>
</tr>
<tr>
<td>Terrestrial mammals</td>
<td>Often green or brown</td>
<td>None</td>
<td>Indehiscent nuts, pods, or capsules</td>
<td>Pulp lipid- or starch-rich</td>
</tr>
<tr>
<td>Highly frugivorous birds</td>
<td>Black, blue, red, green, or purple</td>
<td>None</td>
<td>Large drupes or arillate seeds; often dehiscent; seeds &gt;5 mm long</td>
<td>Pulp lipid- or starch-rich</td>
</tr>
<tr>
<td>Partly frugivorous birds</td>
<td>Black, blue, red, orange, or white</td>
<td>None</td>
<td>Small- or medium-sized drupes, arillate seeds; small barbs, hooks, or berries; seeds &lt;10 mm long</td>
<td>Pulp often sugar- or starch-rich</td>
</tr>
<tr>
<td>Feathers or fur</td>
<td>Undistinguishable</td>
<td>None</td>
<td>Barbs, hooks, or sticky hairs</td>
<td>None</td>
</tr>
<tr>
<td>Ants</td>
<td>Undistinguishable</td>
<td>None to toxic to humans</td>
<td>Elaiosome on seed coat; seed &lt;3 mm long</td>
<td>Oil or starch elaiosome with chemical attractant</td>
</tr>
</tbody>
</table>

*From Howe and Westley, 1988.*

### 2. Plant Adaptations to Diaspore Dispersers

#### a. Dispersal Syndromes

Animal dispersers impose selection on fruit and seed characters. As with the suites of floral traits ascribed to selection by particular groups of pollinators, biologists have also described "dispersal syndromes," suites of characters that appear to be shared by propagules sharing certain groups of animal dispersers or particular modes of transportation (Table IV). For example, most dog owners can describe the common characteristics of propagules dispersed on animal fur and feathers, which include barbs, hooks, and barbed hairs that cause these annoying passengers to attach firmly to socks as well as fur. However, as with pollination syndromes, considerable controversy exists over whether these suites of traits are the result of selection by particular groups of animal species or whether they instead reflect the evolutionary constraints imposed by the morphology and physiology of plant ancestors (Jordano, 1987).

#### b. Plant Adaptations to Frugivores

The ripe fruit, of course, functions as the attractant and reward for many seed dispersers. The problem for the plant in this case is to ensure survival of at least some of the seeds. First, the plant must ensure that the fruit remains on the plant long enough for the seed to develop and be provisioned by the mother. Consequently, immature fruits share with other plant organs various mechanisms that deter animal consumption. Unripe fruits often contain toxins and palatability-inhibiting compounds (Stiles, 1989). Perhaps for some readers the most memorable example of this phenomenon will be the inadvertent bite into an unripe persimmon.

A ripe fruit is advertising for dispersal of mature seeds. During ripening, fruits often change color, from an inconspicuous green that is poorly discernable amid the foliage to a contrasting color such as red, blue, yellow, or black, which is conspicuous to visually
searching frugivores. The ripe fruit may contain various nutritious goodies that attract prospective dispersers, including sugars, minerals, water, lipids, and proteins. These resources are costly for plants to provide, and some plants produce fruits or seeds that mimic more nutritious fruits but either lack any nutritional value or are much less nutritious than their model fruits.

Occupy the attractively nutritious fruit, seeds must possess traits that promote survival of fruit consumption and digestion. Sometimes seeds can be easily separated from the pulp and are discarded by frugivores prior to digestion (remember watermelon seed spitting contests). Other seeds are relatively indigestible, due either to hard seed coats, toxins, or simply the incompetence of the frugivore gut, and pass through unharmed.

Frugevores vary in where and how they deposit seeds. Those that discard seeds as they are eating often deposit seeds individually, but may also fail to move seeds any distance from the parent. Dispersers that regurgitate seeds may move them further and often deposit seeds individually. When seeds are defecated, however, they may be concentrated at high densities. If no secondary dispersal process spreads these seeds, they may experience extremely competitive growing conditions upon germination. Such dense concentrations of seeds may also attract seed predators or secondary seed dispersers. Ants, for example, often remove seeds from dung piles. Nevertheless, some seeds require passage through a vertebrate gut before they will germinate (Traveset, 1998).

Frugevores function as seed dispersers only if they move away from the parent plant before defecating, regurgitating, or otherwise discarding the seed. Fruits often contain mild toxins, which are thought to improve dispersal by deterring foraging frugivores from finishing all the fruits at a bonanza tree. Specialization on particular fruits is rare among frugivores, perhaps because no one species of fruit provides a complete diet. This, too, may be a mechanism that benefits plants because it encourages frugivores to move about to achieve a complete diet.

c. Seed Predators as Dispersers

Many diaspores lack a fleshy fruit or other enticement with which to attract frugivores. These seeds may benefit from dispersal by their predators. Many rodent and bird seed predators hoard seeds for future use. To the extent that they forget or lose buried seeds, they may be extraordinarily effective dispersal agents. Many members of the crow and jay family are important seed hoarders and dispersers. For example, Clark’s nutcrackers carry pinyon pine seeds up to 22 km and bury seeds in small clumps (scatter-hoards). Behavioral studies indicate that these birds are highly effective at finding hoarded seeds and will even dig into a meter of snow to find known hoards. However, individual birds will commonly store more than twice as many seeds as needed, perhaps to protect themselves against theft. Theft may be common, especially among social seed predators, such as pinyon Jays. Pinyon jays that observe other individuals burying seeds can also find and exploit those hoards.

Caching may have significant effects on the local ecosystem. For example, Clark’s nutcracker initiates forest succession after large fires by moving in limber pine seeds from long distances. Genetic structure of tree populations may also be influenced by the seed-caching behavior of Clark’s nutcracker.

Seeds dispersed by predators have interesting adaptations that promote offspring survival. Oaks, for example, are dispersed by their seed predators, which include birds and rodents. Both types of animals crack open acorns and feed on the cotyledons of the embryo. During germination, the embryo root (radical) emerges from the apical (pointy) end of the acorn. Several studies by Steele and colleagues have discovered that digestibility-reducing astringent tannins are concentrated at the apical end of the acorn. Further, all acorn consumers studied, including insects, birds, and rodents, preferentially consume the basal (cap) end of the acorn, which leaves the embryo intact and viable. Partially eaten acorns can germinate, sometimes at higher rates than found for intact acorns. Finally, Frost and colleagues have found that in the European oak Quercus robur, experimentally removing cotyledons from seedlings does not reduce their germination success when compared with seedlings with intact cotyledons. These authors suggest that the primary function of the cotyledons is to attract seed dispersers.

d. Evolutionary Dead Ends?

As with other forms of specialized dependence, specialization to a particular dispersal agent can be an evolutionary dead end. Perhaps the most famous example involves the fruits of the Calvaria major tree on the island of Mauritius in the Indian Ocean. This tree had not been observed to recruit young seedlings for over 300 years when Temple surmised that it was lacking its essential fruit disperser, which was the dodo, a large flightless pigeon that had gone extinct in the late 1600s. Temple was able to mimic dodo digestion by feeding the Calvaria seeds to turkeys.
Turkeys were not available to rescue Calvaria on Mauritius, but Janzen has argued that another tree has survived the loss of its dispersal agent in Central America through such a substitution. Huge, rare guanacaste (Enterolobium cyclocarpum) trees in Costa Rica produce fruits that are readily eaten by domesticated horses and cattle. Janzen speculated that these fruits were once dispersed by the native Pleistocene horse (Equus fraterus), which went extinct some 10,000 years ago. He demonstrated that most seeds pass unharmed through the guts of domesticated horses and went on to argue that these animals, which were introduced by Spanish conquistadors 500 years ago, have replaced the lost dispersal agent. Thus, adaptation to one seed dispersal agent preadapted the fruit to other similar agents.

IV. SUMMARY

Plant–animal interactions are ubiquitous and important. A common theme throughout the study of plant–animal interactions is the enormous effects that these interactions have on plant and animal evolution. There is strong evidence that the interaction between plants and insect pollinators is the primary driver of diversity in flowering plants and the groups of insects most involved in pollination. Selection by animal consumers has driven the evolution of numerous plant defense traits. These traits form the basis of many of the uses that we make of plants today. Plant-based fibers, pharmaceuticals, and flavorings all derive from plant evolutionary responses to consumers. The plant fitness trade-offs between these defensive traits and competitive ability also play an important role in determining the composition of biotic communities.

The primary benefit that plants obtain from animals is mobility. Many, perhaps most, plants depend upon animals to transport pollen and propagules. In many cases, the interactions between plants and their animal transportation providers are highly specialized and mutually beneficial. These specialized mutualisms can be quite vulnerable to extinction of either party, which is an important issue in both plant and animal conservation.

Finally, plant photosynthesis converts solar energy into chemical energy and thereby provides the energetic basis for most of the world’s life. Plants are therefore the foundation of the global ecosystem. Aside from the decomposition of plant litter through the microbial food chain, this energy flows into the global ecosystem via animal consumers of plants. Understanding how plant–animal interactions influence this process is crucial to understanding how intact ecosystems provide the goods and services upon which human endeavor, and indeed all life, depends.

See Also the Following Articles

ADAPTIVE RADIATION • COEVOLUTION • DEFENSES, ECOLOGY OF • FOOD WEBS • GRAZING, EFFECTS OF • PLANT–SOIL INTERACTIONS

Bibliography


I. INTRODUCTION

A. Defining Plant Biodiversity

Traditionally, the study of plants (kingdom Plantae), fungi (kingdom Mycota), and algae (kingdom Protista) is included under the umbrella of botany (the branch of biology dealing with plants), though this article focuses singly on the plant kingdom, including mosses, ferns, conifers, flowering plants, and related lineages. According to this circumscription, plants are characterized as autotrophs (including some derived heterotrophs), with complex multicellular structures. Their life cycles include an alternation of generations, with multicellular diploid and haploid phases, the sporophyte and gametophyte, respectively; although the rela-
isms are the planet's primary producers, harvesting light energy for their own growth, fixing carbon from the atmosphere into organic molecules, releasing oxygen to the environment, and ultimately sustaining life on earth. Globally, photosynthetic organisms fix approximately 100 billion metric tons of carbon annually and play a key role in maintaining the balance between fixed and atmospheric carbon. Human activities, most critically the use of fossil fuels, are liberating fixed carbon into the atmosphere at a rate in excess of the global carbon fixation rate, contributing to global warming. In light of this trend, it is increasingly important that the potential for carbon fixation by photosynthetic organisms be maintained at the very least. Many human activities that threaten plant biodiversity, such as deforestation and habitat destruction or deterioration, also may decrease plant biomass, further exacerbating the global warming trend. In consideration of global carbon cycles, vegetation is a central focus, yet the emphasis tends to be on biomass rather than biodiversity. While this perspective is reasonable given the role that plants play in these contexts, it is important to encompass all groups of taxa within the framework of biodiversity policy and practice. Therefore in the presentation that follows, the biological properties of each group of plants are presented, with emphasis on those features that may impact issues relevant to discussions of biodiversity.

II. PLANTS AND PEOPLE

As members of the global ecosystem, human life ultimately depends on plant life. In addition, we have dependencies on plants as sources of food, medicine, clothing, shelter, and fuel that predate the origin of the human species. The history of human civilization is inextricably linked to the history of agriculture, including the development of crops and livestock as food sources. It is estimated that only about 3000 of 235,000 flowering plant species have been used by people as food, and only 130 or so of these have been cultivated to any extent. Today, only 6 crops (wheat, rice, corn, potatoes, sweet potatoes, and manioc) directly or indirectly provide over 80% of the human population's calories. The world's more highly industrialized regions depend on high energy input agriculture, in which crops are largely grown as genetically uniform monocultures. Such cultivation practices render these pure stands susceptible to pest and pathogen attack, and it is estimated that one-third of all crops are destroyed by pests.

Most of the global human population, however, lives in tropical and subtropical regions where relatively low
input agriculture is most important. Further improvements in agricultural yields in these regions will have the greatest impact on global human welfare, where improving yields, nutritional quality, and disease resist-
tance could greatly increase overall productivity.

Sources of wild germplasm remain of critical impor-
tance for future crop development and improvements
required to feed the growing human population. Con-
serving the wild relatives of our crop plants has become
a primary concern of plant biologists the world over. Most plants are capable of interbreeding to some degree
with wild relatives, and thus, traditional breeding can
be used to introduce beneficial traits into cultivars. With
the advent of genetic engineering tools, the ability to
interbreed is no longer required to introduce traits into
cultivars, thus removing limitations on species that can
potentially be used as sources of genetic material. In
addition, given the small fraction of plants that have
been exploited as food sources by humans, the totality
of plant biodiversity can be viewed as a potential source
of new crops.

In addition to the major food crops described above,
plants also make their way into our diets and our cul-
tures as sources of beverages such as coffee and tea and
in the form of spices, condiments, and oils.

Another use of plants that has and will likely con-
tinue to affect human lives is as sources of medicinal
compounds. It is estimated that between 35,000 and
70,000 different species have been used as medicines
by various peoples of the world and 80% of the global
population derives almost all of their medicines exclu-
sively from plants. Plants are extraordinary chemical
factories: they produce a wide array of alkaloids, glyco-
sides, and saponins that are the principal source of
the action of a failing heart and in reducing a danger-
ously fast heartbeat. Plants are also sources of medicinal
agents such as quinine, used in the treatment of malaria, and analgesics, includ-
ing the alkaloids morphine and codeine, both derived
by various peoples of the world and 80% of the global
drug used to prevent and cure illness in both tradi-
and Western medicines. Modern Western phar-
macopoeia have derived approximately 7000 different
medical compounds from botanical sources, and 120
or so of these are presently widely prescribed. Included
among these are antimicrobial agents such as quinine,
which is the source of alkaloids used in the treatment of both Hodg-
son's disease and acute lymphocytic leukemia. More
recently, taxol, a drug made from extracts of the Pacific
cypress (Taxus brevifolia), has been introduced as a prom-
sing treatment for ovarian and breast cancer. These are
just a few of the thousands of species that have been
or are presently used to cure human diseases. There is
no doubt that more species have immense untapped
potential. Since tests for determining pharmaceutical
potential are both time-consuming and expensive, and
synthetic drugs continue to be unaffordable to a large
proportion of the global population, it is likely that the
dependence on local plant-based cures will continue to
rise and spread throughout the world.

The qualities of flexibility, durability, and strength, along with ready availability, make plant material such
as wood an ideal source of building material. Wood
and related products, including paper, are of enormous
commercial importance. In industrialized nations these
commodities account for a significant proportion of the
total national consumption of goods. Plant species such
as pine, spruce, hemlock, fir, redwoods, birch, beech,
and oak are used as building materials in the temperate
regions while mahogany, species of the family Dipter-
carpaceae, bamboo, and rattan are used in the tropics.

A wide variety of other materials are also derived
from plants, including the fibers used to weave cotton,
linen, and jute, as well as latexes, resins, perfumes, and
dyes. Finally, plants are also used by humans for their
aesthetic qualities. Ornamental plants in all their color
and shape add beauty to our surroundings and are used
in both public and private gardens, along roadsides in
our cities, and as houseplants. In an increasingly urban
environment, ornamental plants often constitute a rare
link with the natural world.

It is critical that while we exploit plants to satisfy
our needs we continue to conserve the natural habitats
from which they came, as such areas are undoubtedly
the homes for a wealth of other species with the poten-
tial to serve needs of a growing human population.

III. PLANT BIODIVERSITY DESCRIBED

A. Nonvascular Plants

Nonvascular plants (often referred to collectively as the
bryophytes) include three phyla: the mosses (Bry-
ophyta), approximately 9000 species; liverworts ( Hepat-
icophyta), approximately 8000 species; and hornworts (Anthocerophyta), approximately 330 species (Table
I). These groups are characterized by their small
stature, by the absence of specialized conducting tissues
(i.e., xylem and phloem), which occur in other groups
of plants, and by their life cycle. The conspicuous, dom-
inant phase of the life cycle of these plants is the gametophyte (haploid) generation. Gametes are formed in specialized structures on the gametophyte called antheridia (containing male gametes) and archegonia (containing female gametes). Antheridia and archegonia may occur on the same gametophyte or on separate male and female gametophytes. Male gametaphytes and nonvascular plants have motile sperm and are dependent on water for dispersal to the female gametophytes, which are nonmotile eggs, retained in the archegonia on the gametophyte. Dispersal of gametes is thought to be extremely localized (less than 1 m). After fertilization, the diploid phase of the life cycle, the sporophyte, grows out of the archegonium, where it remains attached throughout its development, dependent on the gametophyte for most or all of its nutrition throughout its existence. Meiosis occurs in the sporangia, specialized structures of the sporophyte, resulting in the formation of haploid spores that are dispersed primarily by air currents. Spores are extremely small, are tolerant of environmental extremes, and thus have the capacity to disperse over much greater distances than gametes. Many nonvascular plants also have the capacity to reproduce asexually as gametophytes, either by fragmentation or via specialized propagules called gemmae.

Ecologically, the three phyla of nonvascular plants share their dependency on water for sperm dispersal, and thus bryophytes tend to occur where water is at least seasonally available and/or tends to accumulate. Among the richest habitats for bryophyte diversity are temperate and tropical cool, moist forests and arctic and alpine habitats. Nonvascular plants grow on a variety of substrates, including soil and bare rock, and on other plants as epiphytes. On bare rock, they play an important role as initial colonizers. Peatlands are dominated by mosses in the genus Sphagnum. These unique and important ecosystems cover about 1% of the Earth's surface and represent a significant global carbon sink, containing over 100 billion metric tons of unavailable carbon.

More than 60% of bryophyte families have worldwide distributions, occurring on all continents. Bryophyte species also tend to have wide geographic distributions: species that occur on more than one continent are common. Two processes have contributed to establishing present-day distributions. First, much of the diversity that we see today may be ancient; and present-day distributions may be the result of vicariance, or the splitting of a larger, once continuous range into smaller isolates. Second, especially in taxa that have originated more recently, present distributions are likely the result of long-distance spore dispersal. The role of ongoing dispersal in bryophytes is relevant to biodiversity research, because it has broad implications for structuring of genetic variation in these species. As suitable habitats become fragmented or patchily distributed as a result, for example, of clear-cut logging on small and large scales, the potential for reconnection by extripated bryophytes depends critically on dispersal and establishment, which are poorly characterized in these groups. Although we know that spores are easily carried by air currents, only a handful of studies have examined patterns of gene flow in bryophytes. Another consideration in bryophytes that may be relevant to biodiversity is the relationship of taxonomic species described primarily on morphological and ecological criteria to biological species—units that are united by gene flow. The reduced morphology and absence of variation in traits related to gene flow are reasons to suspect greater decoupling of taxonomic species and biological species among bryophytes. As a consequence of our lack of understanding of evolutionary processes in these groups, a particular taxonomic species that may be considered common could comprise several cryptic, genetically isolated units, some of which may be critically endangered. Compounding this problem is the relative rarity of expertise in identification of bryophytes. While both trained professionals and amateur botanists contribute to our knowledge of flowering plant biodiversity, in many areas, no local experts in bryophyte identification, professional or amateur, are found. Although the bryoflora of a number of regions is well characterized, the bryophytes are often far less thoroughly studied than vascular plants in the same region. This shortage of expertise is reflected in the status of conservation initiatives in the bryophytes: to date, no comprehensive global listing of threatened bryophytes exists, although the International Union for the Conservation of Nature and Natural Resources (IUCN) has produced a listing for vascular plants. At the present time, the Red List for bryophytes consists of a list of 92 exemplary globally threatened bryophytes.

B. Seedless Vascular Plants

Seedless vascular plants include four phyla with living representatives: phylum Pterophyta (including only two genera, Polypodium and Tmesipteris), phylum Lycophyta (about 1000 species, including the club mosses and the genera Selaginella and Isoetes), phylum Sphenophyta (the horsetails in the genus Equisetum, represented by about 13 species), and the phylum Psilotophyta (the ferns, represented by about 11,000 species; Table 1). All of these groups share a number of characteristics.
They are the earliest set of lineages to have evolved vascular tissue (xylem and phloem). These tissues provide structural stability that allows vascular plants (seedless vascular plants, gymnosperms, and angiosperms) to grow much larger than the bryophytes. The seedless vascular plants also lack seed formation, which distinguishes them from gymnosperms and angiosperms considered below.

Like the nonvascular plants, the seedless vascular plants have a free-living gametophyte generation that forms antheridia and archegonia, in which gametes are produced. Gametophytes in some groups are subterranean, nonphotosynthetic, and associated with mycorrhizae (fungal symbionts that occupy the roots). The gametophytes of most seedless vascular plants are inconspicuous and ephemeral relative to the sporophytes, although in some lycopods, it may take 6–15 years for gametophytes to become fertile. Male gametophytes are flagellate sperm, and therefore, as with the nonvascular plants, they are dependent on water for transport to the stationary egg. After fertilization, the diploid sporophyte grows out of the archegonium on the gametophyte, but unlike the nonvascular plants, in these groups the sporophyte grows much larger than the gametophyte and eventually becomes the free-living, dominant form of the life cycle. Meiosis occurs in sporangia on the sporophyte, producing large numbers of wind-dispersed spores that settle and germinate to form the next generation of gametophytes.

The seedless vascular plants display a diversity of ecological tolerances. Even though the motile sperm remain dependent on moisture for transport to the egg, the dominance and perennial habit of the sporophytes appear to have somewhat lessened the strength of selection for regular water availability that appears to limit the range of habitats suitable for bryophytes. Living representatives of the Lycophyta are all herbaceous and mostly tropical. The two living genera of Psilophyta are tropical and subtropical. Approximately 75% of fern species are tropical, and about a third of these are epiphytic. A few species of ferns are known only as gametophytes, while others do not appear to form sporophytes near their range limits. It may be that when ecological conditions are no longer favorable for some ferns, sexual reproduction is not possible. This suggests that rates of sporophyte production could act as an indication of habitat quality in some species.

Among the phyla of seedless vascular plants, a number of families contain globally threatened species, according to the 1997 IUCN Red List for vascular plants. These small plants mostly occur in aquatic or wet terrestrial habitats in tropical and subtropical regions. They represent the nearest living relatives of ancient tree lycophytes that dominated the coal-forming swamps of the Carboniferous.

Among fern families, the Cyatheaceae, Ophioglossaceae, Loxomataceae, and Parkeriaceae all have in excess of 25% of their species listed as globally threatened. The Cyatheaceae is the family that includes the tree fern genus *Cyathea*. Tree ferns are conspicuous elements of tropical and subtropical montane forests. The Parkeriaceae includes only four species in the genus *Ceratopteris*, an aquatic genus of mostly tropical, edible floating ferns, some of which are cultivated for human consumption. The Ophioglossaceae includes three genera, among them the genus *Botrychium*. Members of the genus *Botrychium* have unusual population structures that remain poorly characterized. Species of *Botrychium*, most of which are of conservation concern, occur as rare, widely scattered sporophytes. These ferns are perennials and may not produce aboveground structures every year. Gametophytes, which are subterranean in this genus, are rarely found, and thus the dynamics of sexual reproduction have not been explored.

C. Gymnosperms

The gymnosperms are conspicuous and important components of many terrestrial ecosystems. They include four phyla with living representatives, including the conifers (phylum Coniferophyta; about 550 species), cycads (phylum Cycadophyta; about 140 species in 11 genera), the maidenhair tree, *Gingko biloba* (the only...
species in the phylum Gingkophyta, and the gnetophytes (phylum Gnetophyta, including three genera, Gnetum, Welwitschia, and Ephedra, totaling about 70 species; Table I). Together, the gymnosperms and angiosperms constitute a uniquely derived group, the seed plants. The four phyla of gymnosperms (as well as the angiosperms) all form seeds, defined as mature ovules that contain embryos. As such, they represent a further shift in the importance of the sporophyte relative to the gametophyte generation. In all seed plants, the gametophyte is enclosed within sporophytic tissue and, thus, is no longer free-living at any time during its existence. Male gametophytes are packaged and dispersed as pollen grains, while female gametophytes are retained on sporophytes in ovaries. Pollen may be trans-ferred to the vicinity of the ovules by wind or insects (e.g., some gnetophytes and cycads are likely beetle pollinated). Pollen tubes are formed, through which male nuclei are transported to the egg cells. After fertilization, the embryo develops inside the megasporangium wall, resulting in a seed that may be dispersed by wind or animal vectors. For example, the seeds of pinyon pine are dispersed by nutcrackers. Although the cycads and Ginkgo have flagellate sperm, these are released from pollen grains after they reach the vicinity of the ovule, and thus, the seed plants no longer have the requirement for water for gamete movement. This opens up yet another set of habitats for these plants. Among gymnosperm lineages, ecological requirements are highly varied. The Coniferae (conifers), the most species-rich phylum of gymnosperms, are also the most diverse ecologically. Conifers play an especially important role in temperate and boreal forest ecosystems, where they are often the dominant tree species. In these regions, conifer species are also of significant economic importance and are managed for timber and paper pulp production. Overexploitation of this resource has impacted these ecosystems in numerous ways, impacting not only narrow endemics such as the burrowing owl of the Pacific northwestern United States but also other resources such as native fisheries habitats, including the spawning grounds of Pacific salmon.

The 1997 IUCN Red List for vascular plants indicates that all four phyla of gymnosperms include families with globally threatened species. The conifers include seven families with more than 49% globally threatened species. The single species in the Gingkophyta, the maidenhair tree, Ginkgo biloba, is a relictual Asian species that was preserved in temple grounds in China and Japan and is not known to occur in the wild. A very large percentage of cycads (82.8%) are globally threatened (Table I). The cycads occur in a variety of tropical and subtropical habitats. These relictual plants are cultivated as ornamentals and are threatened in part due to over-collecting.

D. Angiosperms

With somewhere over 235,000 extant species, flowering plants represent over 80% of plant biodiversity. They essentially share the life cycle common to all seed plants, with the innovation of having reproductive structures contained in flowers and with further reduction in the gametophytic phase of the life cycle. The flowering plants have evolved a number of features that have been described as important in their radiation, including relationships with pollinators, which opened the way for diversification of floral structures that likely contributed to reproductive isolation among diverging lineages. Flowering plants display a wide array of variation in ecology, life history, growth form, breeding system, and other traits that make them impossible to characterize as a group. Ecologically, flowering plants are more diverse than all other groups of plants described. They occur in virtually all terrestrial habitats, and in many of these, they are the dominant elements in the landscape. A number of flowering plant species have become adapted to aquatic habitats, while others have evolved tolerance to arctic, extreme desert, and even marine conditions. The 1997 IUCN Red List for vascular plants lists nearly 300 families that include globally threatened species. Some generalizations about broad-scale patterns of diversity in flowering plants can be made, though regional, local, and taxon-specific patterns are also important in discussions of biodiversity. Because discussions of patterns of diversity most often consider all vascular plants (including seedless vascular plants and seed plants), these patterns will be considered in subsequent sections.

A number of features of flowering plants impact the way in which genetic variation is apportioned in nature, and thus these features impact not only the susceptibility of plant species to becoming threatened but also their potential for recovery. For example, the breeding systems of plants range from highly self-fertilizing to obligately outcrossing. Among outcrossing species, a variety of mechanisms exist to minimize the potential for self-fertilization. Chief among these are genetic self-incompatibility mechanisms, which are estimated to occur in as many as 50% of angiosperms. These mechanisms operate via the action of alleles at a single locus that must differ in paternal and maternal parents in order for matings to be compatible. Genetic self-incom-
Compatibility can have major effects on species conservation and recovery plans, because the number of self-incompatibility alleles that occur in a population can become depleted to the point of severely limiting the potential for successful matings. Highly selfing species tend to display much higher levels of differentiation among populations than many outcrossing species and thus may show a greater tendency for adaptation to locally varying conditions. While inbreeding is generally viewed as having a negative impact on threatened species, plants that habitually inbreed are well adapted to do so. Thus it is important to understand breeding systems in order to make well-informed conservation and management decisions.

IV. PATTERNS OF PLANT BIODIVERSITY

A. Measures of Biodiversity

Various measures are used to determine plant diversity. A simple measure of diversity is species richness. This measure refers to the number of different species in a given region. Although there are other more robust measures of diversity, plant diversity studies are often based on the measure of species richness. The most species-rich regions of the world are in the tropics and subtropics. Of the 270,000 or so currently known vascular plant species, approximately two-thirds are restricted to these regions. Species endemism is often used in discussions of species richness. Species can be categorized as broad or narrow endemics. Broad endemics are restricted to a specific region (which may be rather large) but are seldom rare. Narrow endemics, species that best fit the colloquial notion of rarity, are species that are highly localized and are found in small numbers in only a few localities. Measures of species richness and endemism are important criteria for estimating plant diversity as well as prioritizing conservation of a given region.

The extent of species endemism is also a useful measure in describing biodiversity hot spots (Myers, 1990). These are regions where exceptional concentrations of endemic species are facing exceptional threats of destruction of habitat. To qualify as a hot spot, a region must contain at least 0.5% or 1300 of the world’s 270,000 vascular plant species as endemics. Secondly, a hot spot should have lost 70% or more of its primary vegetation, described as the form of habitat that usually contains the most species, especially endemics.

B. Patterns of Endemism

As many as 133,149 species (+4% of all species of vascular plants) are confined to 25 hot spots comprising only 1.4% of Earth’s land surface. Fifteen of these hot spots contain at least 2500 endemic plant species. The remaining 10 contain over 5000 endemic species (Myers et al., 2000). Note that nonvascular plants are not included in these determinations because of the lack of the understanding we have of their taxonomy and global patterns of distribution. It is believed that even greater numbers of nonvascular species may be confined to these hot spots.

The 25 hot spots can be grouped into several habitat types. Predominant are tropical forests, appearing in 15 hot spots, and Mediterranean-type zones, in five. Nine comprise islands and almost all tropical islands fall into one or another hot spot. The five of the “hottest hot spots” are predominantly tropical—the tropical Andes, Sunderland (Indonesia, Malaysia, and Brunei), Madagascar, Brazil’s Atlantic forest, and the Caribbean. These five hot spots along with the Philippines, Mesoamerica, the Mediterranean Basin, and Indo-Burma contain endemics amounting to 30% of global totals for vascular plant species and represent a mere 0.7% of the Earth’s land surface.

The species-level analyses are complemented by an assessment of endemism among higher taxa such as plant families and genera. Madagascar and the surrounding islands possess 11 endemic families and 310 endemic genera, Cape Floristic Province has 6 endemic families and 198 endemic genera, and New Caledonia has 5 endemic families and 112 endemic genera. Such unique patterns of species richness and endemism make these areas a high priority for conservation.

V. THREATS TO PLANT BIODIVERSITY

A. Factors Affecting Susceptibility of Plant Populations to Threats

Human-induced disturbances relating to agriculture and urbanization, in addition to natural catastrophes such as droughts, floods, fires, disease, and hurricanes, have negative effects on plant diversity. Among the factors that affect the degree of susceptibility of a species to such threats are local population size, geographic range, habitat specificity, and the extent of dependence on other organisms.

Population size can be an important determinant of the fate of a species, especially when considered in
combination with the extent of a species geographic range. A species with a narrow geographic range and small population sizes will have an increased chance of all of its populations experiencing the same threat, and each small population will be at greater risk of extinction than in a species with a broader range and larger populations.

Specific biotic interactions can also be important for species persistence. Plants do not exist in isolation: in fact their survival often depends on other organisms. Many species of plants have evolved intimate relationships with a diverse array of organisms, without which they are unable to complete their life cycles. Symbiotic or mutualistic relationships with fungi, bacteria, insects, and birds are obligatory for the successful germination, growth, reproduction, and dispersal of propagules in many species of plants. For example, flowering plants that depend on one or a few species of animals for pollination and seed dispersal may become threatened with extinction in the face of a rapid decline in the population of these animals. Reduced seed set is the most direct effect of pollinator decline, yet indirect effects may also contribute to a decline in the plant population. For example, in the absence of pollinators, a higher percentage of seed may be set through self-pollination, decreasing heterozygosity and increasing the expression of deleterious traits often associated with a shift to inbreeding. Ultimately, loss of pollinators or disruption of pollination systems can cause plant extinction. A number of factors, including habitat alteration or fragmentation, grazing, introduction of alien pollinators, and the use of pesticides, can have drastic effects on pollinator populations. The plants most at risk from loss of pollinators are dioecious (species where male and female flowers occur on separate plants) or self-incompatible, those that have a single pollinator, and those that propagate only by seed. Once again, a plant species with a restricted range can be especially vulnerable to population fluctuations among its pollinators and seed dispersers.

B. Anthropogenic Threats to Plant Biodiversity

Anthropogenic phenomena pose the greatest threat to plant biodiversity. These threats include pressures such as habitat fragmentation resulting from agriculture, logging, and development as well as climate change, grazing, invasions by exotic species, and overharvesting of individual species. Agriculture has had exhaustive impacts on the planet. The search for a food supply has done more to decrease plant diversity and to physically alter our surroundings than any other human activity. Habitat fragmentation, often a result of agriculture, is a serious threat to plant diversity and is the primary cause of the present extinction crisis. Livestock grazing can also have drastic impacts on plant communities. The selective nature of grazing, combined with the limited tolerance of some plant species to grazing, can result in substantial shifts in species composition.

One of the most visible aspects of habitat degradation in natural landscapes is the spread of exotic, occasionally native, invasive vegetation that displaces native communities. Disturbances such as those described above increase opportunities for these invasions. The introduction and often widespread dissemination of alien species, such as Norway maple (Acer platanoides), kudzu (Pueraria lobata), purple loosestrife (Lythrum salicaria), and Japanese honeysuckle (Lonicera japonica), planted in environments where there are no natural controls or defenses, have been devastating.

Another rising threat to present-day plant diversity is global warming. The predicted warming of the atmosphere is a result of the increase in the levels of carbon compounds, especially carbon dioxide (CO2), produced as a result of many anthropogenic activities. A rise of 1–2% of atmospheric CO2 can double the present atmospheric levels (350 ppm) in about 50 years, with the resulting temperature increasing by 1–3°C. This increase in temperature can eliminate species confined to mountain peaks and northern temperate regions while some species could presumably move north, changing global patterns of plant diversity. Global warming can also induce other natural changes—changes in precipitation, evaporation rates, sea level, atmospheric chemistry (rising CO2), UV penetration of the atmosphere, and soil and water chemistry—which, singly or in combination, may differentially affect plant species and their communities, leading to drastic changes in patterns of global plant diversity.

The current wave of extinctions resulting from systematic pressures imposed by human activities is eliminating 27,000 species each year, making this the sixth greatest mass extinction in the history of our planet. The principal factor determining the rate at which species are becoming endangered and extinct is habitat destruction and deterioration, especially in the species-rich tropics. Of the 270,000 or so currently known vascular plant species, 170,000 are tropical or subtropical endemics. The rain forests, which harbor this amazing diversity of species, are subject to destruction by exploitation and land clearing on an unprecedented scale. It is suggested that 60,000 tropical plant species

H11034
will be at risk of extinction within the next 50 years. On a similar scale, of the 80,000 species of the temperate zone, about 8000 are threatened and several hundred endangered. At least 217 species of vascular plants and undoubtedly many species of nonvascular plants have gone extinct in North America over the past 300 years. Many extinctions have gone unnoticed because the species were not known to science. In the United States, old-growth forests of the Pacific Northwest, longleaf pine forests of the southeastern coastal plains, range-lands, grasslands and savannas, and wetlands—33% of the species listed under the Endangered Species Act are dependent on wetlands—are all considered endangered or threatened ecosystems due to one or more reasons ranging from logging, to conversion to agriculture, and secondarily, to fire suppression or overgrazing and subsequent invasion by exotics.

Eleven hot spots have already lost at least 90% of their primary habitats while three have lost 95%, making conservation a high and immediate priority. Some hot spots have their endemic species concentrated in exceptionally small areas, making these endemics highly susceptible to extinction. The Eastern arc, New Caledonia, and the Philippines are especially significant in this regard. The criteria used to define hot spots exclude some of the most species-rich areas of New Guinea, Amazonia, and the Congo Basin because these regions contain rich endemic floras (New Guinea has 15,000 endemic species while the other two regions amount to 30,000); they also retain at least 75% of their primary vegetation, disqualifying them for hot spot status. Further, regions such as the Ethiopian Highlands, southeastern China, and northern Rwanda, where exceptionally rich floras face exceptional threats, are not sufficiently documented to meet the hot spot criteria. Thus, while it can be argued that hot spot analysis may be used to prioritize conservation efforts to areas where efforts will have the greatest impact, other poorly characterized regions and areas of high endemism must also be considered.

C. Natural Threats to Plant Biodiversity

In addition to the devastating effects on diversity by anthropogenic activities, natural catastrophes can also have negative impacts on plant diversity. Natural catastrophes on a global scale are extremely rare but natural fires, floods, hurricanes, landslides, and droughts occur at various times in all parts of the world. However, regularly or sporadically occurring events such as fire and hurricanes are critical for the maintenance of some plant communities. Many plant species are not only adapted to such disturbances but also depend on them for successful growth and reproduction. In such cases, human interference that hinders or eliminates natural disturbances is likely to pose a threat to plant diversity. Such interference can often intensify these natural disturbances, as in the case of forest fires, resulting in drastic reductions in plant diversity. Hence, a knowledge of natural disturbance regimes is important in the conservation of plant diversity. Prairies, other grasslands, and ponderosa and longleaf pine forests often depend on frequent, yet low-intensity ground fires. Without these fires, these communities gradually change into other community types that are often less diverse.

D. Conservation of Plant Biodiversity

The consequences of loss of plant biodiversity for human welfare and global ecology are unpredictable and wholly irreversible. Hence, identification and immediate protection of sites of high conservation value (i.e., biodiversity hot spots) must be a highest priority. This underlines the vital necessity of increasing inventory and ecological and biogeographical information as a prerequisite to developing plans for conservation. Species are not distributed randomly across the planet nor do they occur in a uniform pattern. Rather, they respond to environmental gradients, including climate, topography, and substrate. They further reflect a long history of species colonizations and extinctions, plate tectonics, and other global processes. By paying attention to these processes, conservationists can identify areas of greatest importance for protection. There are several criteria used in assessing the conservation value of natural areas. Species richness (or diversity), endemism, rarity (extent of habitat), threat of human interference, amenity value, educational value, scientific value, and representativeness are some of those criteria often considered when declaring areas as protected.

Decreases in and fragmentation of natural areas is certain to lead to substantial increases in extinction rates. This is therefore a case for some form of selective...
program of ex situ (off-site) conservation although it is universally agreed that the most effective and efficient mechanism for the conservation of plant diversity is habitat protection or in situ conservation. It is also acknowledged that ex situ facilities can be critical in a comprehensive conservation program. Ex situ conservation programs supplement in situ conservation by providing for the long-term storage, analysis, testing, and propagation of threatened and rare species of plants and their propagules. They are particularly important for wild species whose populations are highly reduced in numbers, serving as an alternative to in situ conservation, as a source of material for reintroductions, and as a major repository of genetic material for future breeding programs of domesticated species. Methods of ex situ conservation can be classified according to the part of plant that is conserved—the whole plant, seed, tissues, or genetic material in culture. All these methods of ex situ preservation of live plant material require periodic regeneration and sexual reproduction of the stock. The latter requires knowledge of the breeding system and pattern of genetic variability of the species concerned.

Botanic gardens and arboreta have always played key roles in ex situ conservation. The most widely known function of these institutions is to assemble and maintain a diversity of plant species. They also conduct and facilitate botanical research, especially in plant taxonomy and systematics. Botanic gardens make their most important contribution to the conservation of plant diversity through education as well as their influence on public opinion.

VI. THE NEED FOR FURTHER STUDY

Representatives of all groups of plants contribute to the health of the global ecosystem and to human welfare. As such, any loss of biodiversity has the potential to negatively impact the future of the human population. Among the top priorities for efforts in plant diversity research are training of amateur and professional experts in the study and identification of poorly understood plant groups, including nonvascular plants as well as taxa that have restricted distributions. Special emphasis should be placed on characterizing species that are thought to have potential for medicinal use. In addition, regions of the world that are highest in diversity, especially those regions that are most critically threatened, must be the target of a significant proportion of resources available for conservation. This can best be accomplished by training of additional highly qualified personnel and increasing funding to conservation agencies, including ex situ conservation programs. All of these initiatives will be most effective if they are coordinated by an international body that relies heavily on recommendations made by scientists and regional experts. It is only through concerted efforts such as these that we will be able to slow the loss of plant diversity around the world. As plants play such a vital role in the health of our planet and our species, it is of the utmost importance that we value and protect this diversity for future generations.

See Also the Following Articles

PLANT-ANIMAL INTERACTIONS • PLANT COMMUNITIES, EVOLUTION OF • PLANT CONSERVATION, OVERVIEW

Bibliography


The patterns and processes within plant communities that underlie the fundamental mechanisms regulating and maintaining biological diversity are reviewed and evaluated. In particular, we focus on general processes or "rules" influencing the evolution of diversity within the framework of ecological communities. We do so by comparing ecological theory with patterns observed in the fossil record and in contemporary communities.

I. INTRODUCTION

In the past few decades, questions about the creation, maintenance, and role of biological diversity have stim-
ultated social, political, and scientific inquiry. However, the fundamental mechanisms responsible for biodiversity remain poorly understood. We review the patterns in contemporary and historical plant communities that we believe are critical to achieving a better understanding of the processes that regulate diversity. In doing so, we take an actualist (or pure uniformitarian) stance: We assume that the basic physical and biological processes that regulate diversity today are manifestations of natural “laws” that have been in operation throughout the existence of the earth.

Clearly, any realistic juxtaposition of the concepts derived from neo- and paleontological studies requires a clear exposition of the limitations and evidence on which they rest. Paleo- and neontological research programs each bring different strengths to bear on the question of how biological diversity organizes and is maintained. The greatest strength of neontology is the ability to achieve fine resolution. Much of what is known about the ecological dynamics of plant and animal communities is based on detailed, short-term, small-scale, neontological studies. These studies emphasize the role of local processes such as succession, predation, competition, mutualism, abiotic stress, and a host of other ecological phenomena in the regulation of biodiversity. On the other hand, the paleontological research program tends to emphasize the influence of evolutionary and physical changes on the species composition of communities. Despite their different strengths, the two disciplines often reach similar or complementary conclusions. For example, both paleontology and neontology concur that community composition is highly variable over space and time (Ricklefs and Schluter, 1993; Brown, 1995). Communities do not consist of a “tight-knit” assemblage of co-occurring species; instead, species respond individually to changing biotic and abiotic conditions.

Differences between the disciplines in the level of taxonomic, spatial, and temporal resolution have important consequences on how phenomena are defined and perceived. Within neontology, the definition of coexisting species can vary from very small sets to regional assemblages. In most cases, neo- and paleoecologists use the same vocabulary (e.g., community, taxonomic diversity, and stability), but the meanings of these words change as a function of the spatial and temporal scales used in the two fields. Despite these potential pitfalls, we believe that we can improve our knowledge of the important community processes by focusing on the complementary strengths of each discipline.

II. BRIEF OVERVIEW OF THE EVOLUTION OF TERRESTRIAL PLANTS

The fossil record clearly shows that the number of plant species has increased over geological time (Fig. 1; Niklas, 1997). The first land plants were diminutive, nonvascular, bryophyte-like organisms that inhabited the margins of the terrestrial landscape and provided resources and shelter to the first land-dwelling animals. These plants were restricted in their geographic coverage owing to their dependence on liquid water for vegetative survival and reproductive success (Iles et al., 1998).

The earliest vascular plants possessed better vegetative adaptations to life on land but were also tied to moist habitats for reproduction. They quickly diversified in morphological complexity, primarily during the Devonian, a period that saw the first plants with leaves, roots, lateral meristems, and an erect habit. These plants were larger than their nonvascular progenitors. By Late Devonian times, vascular plants evolved the capacity for secondary growth and with it the ability to form forests. These changes increased the complexity of the physical structure of plants and the diversity of resources available within terrestrial communities, creating opportunities that were subsequently exploited by the evolution of plants with vining, epiphytic, and understory lifestyles. Initial stages of the seed habit also evolved by the Late Devonian, allowing new ecological solutions within the swamp community and laying the groundwork for the subsequent establishment of communities in drier habitats in the Carboniferous. The basic physiognomies of modern plant communities were well established approximately 300 million years ago.

From the late Carboniferous through the early Mesozoic, gradual environmental changes associated with the formation of Pangaea resulted in extinctions of lowland taxa and their replacement by increasingly drought-tolerant taxa migrating from more mesic sites. This established an early mid-Mesozoic flora dominated by a host of seed-bearing clades collectively known as the gymnosperms. Pteridophytes survived but were a secondary element of Mesozoic vegetation.

Angiosperms became an important ecological element approximately 115 million years ago, in the Early Cretaceous. They initially evolved as disturbance-tolerant herbaceous and shrubby plants occupying early successional environments, but they quickly radiated. By the Late Cretaceous, angiosperms started to appear as forest trees in increasingly more stable sites, and they
were the most diverse terrestrial plant clade. However, it is not clear if they formed the dominant vegetation at this time. By the early Tertiary, angiosperms were represented by many living families and genera forming closed-canopy forests in temperate and tropical climates, and they dominated Earth’s vegetation. Cooling and drying climates in the later Tertiary resulted in the diversification of midcontinental grassland communities and the rearrangement of individual distribution patterns that created the biomes of the present. The rise of angiosperms established the full complement of structural and reproductive features (e.g., complex canopies, diverse life history, and reproductive strategies) observed today as well as created the most diverse terrestrial communities in Earth’s history. In their radiation, angiosperms did not displace equally diverse nonangiospermous seed plant communities. Rather, angiosperms both inserted themselves into existing communities and created totally new communities, often creating new opportunities for existing clades. The result has been the most speciose plant communities yet seen on Earth (Fig. 1).
III. NICHE DIVERSIFICATION AND COMPETITION

Understanding the evolution of plants in terms of niches provides the basis by which one can begin to assess the evolution of diversity within plant communities. The concept of the niche is useful because it provides a framework linking individual traits, such as morphology, physiology, and behavior, with their interactions with the biotic and abiotic environment. These features in turn influence population level traits such as growth, abundance, and distribution (Brown, 1995).

Both ecological theory and empirical evidence stress that competition is one of the important forces driving niche diversification and thus community diversity [Knoll (1986) as cited in Rosenzweig, 1993]. The increase in plant species diversity through time is paralleled by evolutionary shifts in resource requirements, body size, allocation strategies, and life history traits suggestive of niche diversification and thus the action of competition. Indeed, the degree of difference or similarity that is required for coexistence has not been quantified, and it has been hypothesized that two or more plant species with very similar resource requirements can coexist indefinitely in the same community as long as their reproductive units (e.g., spores, seeds, or fruits) are dispersed randomly such that adults do not consistently grow in close proximity (Hubbell and Foster as cited in Ricklefs and Schluter, 1993; Brown, 1995; Rosenzweig, 1993; Crawley, 1997).

Competition, however, is difficult to effectively demonstrate in the fossil record. At large temporal scales, such as those encountered in the fossil record, the evidence suggests that competition may be supplanted by other features such as shifts in the physical environment [Niklas et al. (1985) as cited in Ricklefs and Schluter, 1993]. The shift from a Pteridophyte-dominated late Paleozoic flora to a seed plant-dominated Mesozoic flora appears to have been caused by global climate change that selected against the reproductive modes of many pteridophyte lineages, resulting in extinctions in existing lowland communities. Opportunistically, seed plants migrated in from more mesic sites to replace the existing lowland communities. Neoecological observations indicate that mass-specific metabolism, age to reproductive maturity, relative growth rate, and population density are all inversely correlated with body size (Niklas, 1997; Enquist et al., 1999). Thus, although the total resource needs of larger individuals increase, their use per-unit-volume or per-unit-mass decreases. Body size also affects biological timing. For example, larger species have longer life spans and longer times until reproduction than do smaller species. Increases in size allow species to partition resources over time. Small species tend to be more sensitive to short-term environmental variation, whereas large species apparently cue to longer term fluctuations.

Individual clades within pteridophytes, gymnosperms, and angiosperms often display Cope’s rule. Derived taxa tend to become larger. An increase in size confers a competitive advantage in the struggle for the resources, such as light and nutrients, that are necessary for growth and reproduction [Harper (1977) as cited in Crawley, 1997]. Paleontological data indicate that Cope’s rule may also reflect the tendency for new clades to make their evolutionary debut in disturbed habitats [DiMichele et al., 1987; Niklas et al. (1985) as cited in Ricklefs and Schluter, 1993]. Such species often possess a small body size and its associated rapid vegetative growth and onset of sexual maturity (i.e., the classic r-selected species). Conversely, larger species within many plant clades appear to occupy more ecologically stable habitats.

Over long time spans there may be selection against larger sizes. Larger species tend to be more prone to extinction during periods of environmental change than their smaller counterparts [Bakker (1977), Stanley (1979), and Tiffney and Niklas (1985) as cited in Niklas, 1997; Brown, 1995]. Presumably this is because larger species take more time to reach reproductive maturity, tend to produce fewer progeny per reproductive cycle, and have lower population densities per unit.
area. This suggests that evolutionary processes operating on microevolutionary and macroevolutionary timescales may have conflicting outcomes. Under the limitations of physical and biological constraints, organisms benefit from competitive advantage by becoming larger until an unpredictable environmental change forces them to extinction.

A plant's form, or shape, is as important as its size in partitioning ecological opportunities. Many aspects of plant form and architecture can be interpreted as the result of natural selection acting to optimize biomechanical stability and the ability to procure resources from the environment while minimizing hydrodynamic costs of resource transport [Niklas and Kerchner (1984) as cited in Niklas, 1997; West et al. (1997, 1999) as cited in Enquist et al., 1999]. Only a relatively small number of "plant designs" appear to be functionally viable over the range of sizes within which vascular plants operate. It can be shown via mathematical modeling and computer simulation that most of these viable options have been exploited [Niklas and Kerchner (1984) as cited in Niklas, 1997] and that the evolution of plant size has been strongly limited by biomechanical constraints (Niklas, 1997; Bateman et al., 1998). In summary, although selection has apparently operated to increase plant size within clades, physical and biomechanical constraints have restricted maximum size (Fig. 2).

V. ALLOCATION AND REPRODUCTIVE STRATEGIES

Biomechanical limitations restrict niche partitioning involving body form. However, plants with different forms may vary the allocation or timing of energetic output, thus increasing the potential for coexistence. Separate species attain their "adult" size at different times by altering growth rates. Unless the total amount of energy available to the individual plant changes, such changes in growth rates must be accompanied by a trade-off in other activities requiring energy. For some species, changes in growth rate are compensated for by changes in wood density: Fast-growing arborescent species generally produce wood of a much lower density than do slower growing species (e.g., balsa versus mahogany; Enquist et al., 1999). This has obvious implications for relative mechanical stability, susceptibility to disease, and species longevity. Similarly, an energetic trade-off exists with regard to the ability to tolerate herbivory, parasitism, and disease (Crawley, 1997). By investing energy in the formation of secondary chemicals, plants can often avoid enemies, albeit at the cost of slower growth and delayed sexual maturity.

Perhaps the area in which different allocation strategies have the strongest effect is within reproduction.
Plants can alter both the proportion of resources used in reproduction and the amount of resources allocated to individual disseminules. Such adaptations are often linked to community structure. Large seeds, for example, tend to occur in forest communities, whereas smaller seeds are more common in open or disturbed habitats (Crawley, 1997). Variation in reproductive strategy provides plants with additional ways to partition environmental opportunities and coexist.

The fossil record suggests that reproductive adaptations have permitted the ecological exploration of previously poorly colonized terrestrial environments, starting with the invasion of land, the evolution of the seed, and finally the suite of adaptations involved in the origin of angiosperms (Niklas, 1997). Early land plants possessed pteridophytic reproduction, requiring freestanding water. With evolution of advanced forms of gymnosperm, this dependence was lost, but at the cost of vegetative reproduction and the seed habit are found to coincide with potentially protracted inclement environmental conditions. Dormancy, however, may also increase the probability of seed predation. It is only with the angiosperms that asexual vegetative reproduction and the seed habit are found to co-occur within an individual with either an herbaceous or woody growth habit. The possession of rapid sexual life cycles, the potential for an herbaceous habit, and in many cases the potential for vegetative reproduction predisposed angiosperms to outdiversify and potentially outcompete other lineages in multiple ways in response to environmental changes especially in the Cenozoic. These traits allowed for further invasion of new habitats, such as desert and arctic/alpine environments.

VI. PLANTS AND CLIMATE: PHYSIOGNOMY AND PHYSIOLOGY

The patterns of plant diversity and distribution are influenced by climate. Most biomes show a remarkable degree of convergence in plant architectures, size structure, and physiology, despite drastically different biogeographic and evolutionary histories of their component taxa (Leigh, 1975) and Mooney (1977) as cited in Brown, 1989; Muleski (1983) as cited in Ricklefs and Schluter, 1993. Many of the physiognomic features associated with plants found in communities within each of these biomes, such as the relative proportion of compound and simple leaves, root to shoot ratios, plant heights, leaf cuticle thickness, and leaf indument, all show consistent trends with variation in temperature, precipitation, humidity, and rainfall [Crawley, 1997; Rainhauck (1954) as cited in Crawley, 1997; Wolfe (1978, 1979) as cited in Ricklefs and Schluter, 1993]. Indeed, because of the tight association between vegetative physiognomy and climate, physiognomy is often used as a proxy for reconstructing past climates.

The tight link between climate and physiognomy is due mainly to physiological and stoichiometric trade-offs (Crawley, 1997). One of the most important is the trade-off between water and carbon balance. In order to assimilate atmospheric CO₂, plants must be able to exchange gasses through their stomata, but open stomata cause water loss via transpiration. In many terrestrial environments water stress can limit rates of production by reducing photosynthesis. The evolutionary history of terrestrial plants is marked by the evolution of adaptations that enhance the ability to increase productivity in the face of water stress. The initial radiation of vascular plants was most likely due to the evolution of a cuticle pierced by stomata whose aperture could be regulated (Niklas, 1997). These adaptations were honed through the subsequent evolution of sunken stomata, leaf pubescence, physiologically dormant, and the ability to mechanically position leaves in response to environmental change. Such morphological adaptations allowed for the colonization of more seasonal environments [Hutton et al. (1998) as cited in Battman et al., 1998].

Although it appears that changes in morphology allowed for the continued exploitation of new resources
and habitats, physiological processes appear to be highly conserved over geologic time (Raven [1993] as cited in Bateman et al., 1998). However, interpreting these features is difficult from fossils because the process of preservation alters chemical fossils. When physiological processes are mirrored in morphology, we infer physiological evolution. For example, changes in the density of leaf-horse stomata appear to correlate with geochemically predicted changes in atmospheric CO₂ content, at least back through the Mesozoic. Other morphological features suggest that basic plant physiology has changed little in 400 million years. C₃ photosynthesis appears the norm; CAM photosynthesis may have evolved by the Pennsylvanian. C₄ photosynthesis is, at the oldest, Late Cretaceous and probably younger, matching the radiation of angiosperms into hot and dry environments. The timing of observed increases in plant diversity does not appear to be associated with physiological changes.

VII. COEVOLUTION

The diversification of vascular plants is associated with an increasing number of coevolutionary relationships. Coevolution can be subdivided into two extreme forms: “tight” coevolution, in which the interactions between two closely linked organisms are intimately interdependent resulting in mutual evolutionary influences, and “loose” coevolution, in which the interdependence of two or more species is facultative and transient. Coevolution includes interactions in which all parties benefit and interactions in which one party benefits at the expense of the other.

The fossil record suggests tight coevolution is rare or short lived. For example, Abert’s squirrel currently feeds almost exclusively on ponderosa pine, which currently has a very large geographic range; the modern distributions of Abert’s squirrel and ponderosa pine closely overlap. However, Betancourt and Van Devender (as cited in Brown, 1993) suggest that this is a very recent feature since the geographic range of ponderosa pine was reduced to a handful of relict populations in the Pleistocene that probably could not have supported a viable squirrel population. Likewise, there are many angiosperm genera whose fruits or seeds are currently dispersed by specific mammals or birds. Although many of these genera had similar seed and fruit morphologies more than 30 million years ago, none of their modern dispersal agents existed during this time.

In fact, these plant lineages passed through two major turnovers of mammals during the Tertiary, indicating that dispersal coevolution must have been either loose or capable of rapid adaptation. Additionally, recent analyses indicate that the radiation of the angiosperms had little apparent influence on the diversification of insect families, although it is possible that coevolutionary response occurred below the family level. In fact, insects began radiating more than 100 million years before the ascendancy of angiosperms [Labandeira and Sepkoski (1993) as cited in Rosenzweig, 1993]. These observations suggest that many of the tight coevolutionary couplings seen in modern communities may be relatively transient phenomena.

Nonetheless, both neo- and paleontologists have shown that loose coevolutionary relationships have been an important part of plant evolution. Insect pollination is a basal feature among angiosperms and their closest relatives. Many floral features, including bilateral symmetry, prolonged calyx tubes, nectar and resin rewards, and sympatry, appeared by the Turonian and suggest specialization for bee pollination (Crepet, 1996). The obligate mutualisms between yuccas and yucca moths, which have been well documented by molecular data, indicate that the general relationship was likely established as long as 40 million years ago. Similarly, fossils indicate that the close association between figs and wasps is at least as old (Pellmyr, 1999). Coevolutionary relationships with pollinators and seed dispersers probably enabled the angiosperms to maintain broadly dispersed but rare or patchy populations by promoting genetic outcrossing and minimizing local inbreeding. This created an opportunity to colonize habitats that were inaccessable to older plant clades that were more reliant on wind and water for the dispersal of seeds and gametes. Further, the presence of mycorrhizal fungi in the earliest plant communities supports the hypothesis that fungi played a key role in allowing plants to spread over terrestrial habitats [Selosse and LeTacon (1998) as cited in Bateman et al., 1998].

As suggested in the Red Queen hypothesis, coevolution can result in an apparent “evolutionary arms race” in which both participants are evolving “as fast as they can” only to maintain their relationship relative to each other [Van Valen (1973) as cited in Brown, 1995]. Ehrlich and Raven (as cited in Ricklefs and Schluter, 1993), for example, suggested that advances in ant herbivory compounds by plants are matched by the evolution of more efficient detoxification mechanisms by herbivores. Fossil evidence for this kind of coevolution is difficult to obtain. One possible example involves seeds of Zanthoxylon (Rutaceae), which possess oils commonly interpreted as deterrents. Such seeds are commonly found pierced by a hole of consistent morphology from approximately 40 million years ago. In extant Zanthoxylon seeds, bruchid beetle larvae create...
similar holes. This suggests a pest–host relationship that has persisted for approximately 40 million years. Nevertheless, it is unknown if such an arms race of evolving strategy and counterstrategy occurred during this period.

The paleoecological perspective also provides evidence for a very broad coevolution between vertebrate herbivores and plants. Vertebrate herbivory appeared in the Pennsylvanian and became well established by the Permian. During this time a terrestrial food pyramid of plants and synapsid herbivores was established, lasting to the Early to mid-Triassic time. These herbivores became extinct during the Middle Triassic and, after a brief hiatus, were replaced by dinosaurs. From the Late Triassic through to the end of the Cretaceous, dinosaurian herbivores dominated. Some of these herbivores were small (on the order of 30 kg), but the dominant ones were nearly two orders of magnitude larger. At the Cretaceous–Tertiary boundary, average herbivore size plummeted with the extinction of terrestrial dinosaurs at the Cretaceous–Tertiary boundary. Together, these observations suggest a coupling between community composition and structure and the radiation of birds and mammals.

Although the forces driving turnovers in vertebrates are hotly debated, it is evident that changes in the composition and structure of floras accompanied changes in the composition of the vertebrate faunas. The Permian–Triassic marked a very broad global transition from floras dominated by pteridophytes and seed ferns to floras dominated by cycadophytes and conifers. This transition resulted in less digestible forage for herbivores, which the fossil record suggests generally favors the evolution of larger herbivores. The radiation of herbaceous and shrubby angiosperms and low-feeding ornithischian dinosaurs during the later Cretaceous again suggests a very broad-scale coevolutionary process. Dense angiosperm communities, spatially well suited to exploitation by small mammalian and avian herbivores, apparently became common only after the extinction of the terrestrial dinosaurs at the Cretaceous–Tertiary boundary. Together, these observations suggest a coupling between community composition and structure and the evolutionary history of large herbivores.

VIII. SUCCESSION

Succession is one of the most conspicuous processes in the change of community composition and structure over ecological time scales and is one of the most widely written about topics in plant ecology. Because succession involves local immigration and extinction coupled with changes in species relative abundance, it is conceptually tied to the notion of community stability and equilibrium. As a neoecological process, succession occurs during timescales defined by the lifetimes of individuals within species and the communities they form. It is also dependent on the frequency, duration, predictability, and magnitude of environmental change, the proximity of source areas, and the diversity of life history traits present in the species pool. These features are believed to dictate the apparent orderliness of succession and thus the abundance and diversity of species within a community at any given time (Crawley, 1997).

Succession was minimally important in the earliest land plant communities because they were patchy and composed of taxa with similar growth habits and physiological requirements. By the Late Devonian a variety of clades had evolved shrubs and trees, creating multilevel communities and ushering in light-controlled succession. Other features also affected succession, including nutrient availability, and physical disturbance. Fire-driven succession has been demonstrated for the arborescent lycopod Sigillaria, which occurs immediately above charcoal-rich layers in Pennsylvanian swamps. Succession in Mesozoic gymnosperm communities is more difficult to demonstrate. Retallack and Dilcher (as cited in Niklas, 1981) note that angiosperms and cycadeoids in the Late Cretaceous Dakota Formation tend to be associated with sedimentary features indicative of disturbance, whereas conifers and cycads are generally absent from such environments. Similarly, Hickey and Doyle draw attention to the fact that angiosperms in the Early Cretaceous Potomac Formation appear in otherwise conifer-dominated environments. Other features also affected succession, including the diversity of species within a community at any given time (Crawley, 1997). These features are believed to dictate the apparent orderliness of succession and thus the abundance and diversity of species within a community at any given time (Crawley, 1997).

Succession was minimally important in the earliest land plant communities, suggesting that some or all of the earliest flowering plant species were early successional specialists.

IX. STABILITY AND RESILIENCE

The stability and long-term diversity of living communities are features of considerable contemporary interest and debate. Discussion of these features, however, is often confused by matters of precise definition. Although recent neoecological debate has focused on the effects of diversity on ecosystem processes, such as nutrient cycling and biomass production, some studies have examined the persistence of taxonomic composi-
tion. With reference to taxonomic stability, some ecologists maintain that the more species that are in the community, the more interactions will occur and the more resilient the community becomes to perturbation. A similar argument is made for ecosystem processes. However, a similar number of interactions could occur within a community in which continuous species turnover was occurring, changing the species composition but maintaining species richness and ecosystem processes. Thus, two different measures of the community, taxonomic composition and ecosystem properties such as species richness, could demonstrate conflicting patterns (Haskell as cited in Brown et al., 2000). These features of ecosystems are complex and require careful identification of the appropriate models, definitions, and scales of investigation necessary to define stability and the methods used to measure it.

Within communities, species diversity ultimately reflects the dynamics of local colonization and extinction as influenced by the prevailing physical and climatic conditions, species life histories, and evolutionary clades present. As such, diversity (or species richness) is an emergent property of both ecosystems and evolution. It is the result of a myriad of complex interactions of abiotic and biotic factors. Species diversity per se does not dictate whether or not a community displays stability. However, increased biological diversity may confer an increased array of potential responses to varying environmental conditions. Thus, as diversity increases through increased niche diversification, the probability of including species with tolerances for new sets of environmental conditions also increases. During periods of drastic environmental change, the different capacities for productivity, mortality, reproduction, etc., may allow some community members to survive, or change their relative abundance within the community under the new environmental regime, and thus to produce a similar community under the new regime.

Community stability in the fossil record is primarily gauged in terms of taxonomic composition. However, as data are summed over increasingly long time periods, the level of taxonomic resolution generally decreases. Still, stability may be estimated by rates of extinction and origination, with high taxonomic turnover rates suggesting community instability. Using this gauge, late Paleozoic swamp communities appear to be very stable, changing slowly in the face of increasing global cooling leading to the threshold of the Permian-Triassic boundary (DiMichele and Phillips [1990] as cited in Ricklefs and Schluter, 1993). These communities contained large numbers of pteridophytic and early seed plant species, each of which offered a different “solution” to environmental change. With the transition to the generally continental climates of the early mid-Mesozoic, conifers, cycadophytes, and several other less diverse seed plant clades characterized by slow reproduction and growth rates replaced the Carboniferous swamp communities, but at a lower level of species diversity. However, despite their comparatively low diversities, these floras never experienced substantial extinction events, suggesting that taxonomic diversity and stability are not inevitably linked to one another. Similarly, even though the evolution of angiosperms led to arguably the most diverse communities in Earth history, there is no indication from origination/extinction rates that angiosperms form more or less stable communities than their predecessors.

Instead, it is possible that rates of compositional turnover may be more heavily influenced by rates of environmental or evolutionary change and independent of ecosystem function, taxonomic composition, or species richness. For example, climatically driven migrations observed in the late Pleistocene raise the possibility that species are constantly shifting in response to stimuli and may repeatedly reassort themselves into similar associations. The players may be the same over long periods of time, but the communities are not stable in the sense that we would discuss them in a neoecological context. The relative stability of communities observed from the Paleozoic to the present may suggest that most communities are relatively stable in the face of “normal” Earth change. It may take catastrophic change to destabilize the community, regardless of the diversity.

From the perspective of individual lineages, the fossil record suggests that those clades including species with a range of growth habits appear to be more long lived than clades possessed of but one growth habit. Additionally, large organisms tend to be more insulated from rapid changes in the environment and to respond to longer temporal oscillations in the abiotic environment that may be unavailable to smaller plants with shorter life spans. However, small organisms can respond rapidly to short-term environmental changes. If a clade is represented in a community by species with large and small body sizes, then that clade could persist in a community, despite short- or long-term changes in the environment, thus giving the community the appearance of taxonomic stability when gauged at the higher taxonomic levels.

X. DIVERSITY EQUILIBRIA

Many experiments have apparently demonstrated shortcomings in the original theory of island biogeography,
especially regarding the idea that island species richness is maintained at consistent equilibrium values by colo-
nization and extinction (MacArthur and Wilson, 1967; Rosenzweig, 1995). However, there is increasing evi-
dence that colonization and extinction processes may be important in maintaining an equilibrium in systems in which species composition is not severely limited by low colonization probabilities (Brown et al., 2000). For example, European pollen data show that there is no
overriding directional trend in plant family diversity throughout the past 10,000 years at 24 European sites located between 40 and 70°N latitude. Eighteen of the 24 sites examined show remarkably consistent diversity despite high rates of familial turnover at each site of
30–30% per 1000-year interval (Haskell as cited in
Brown et al., 2000).

Testing theories about diversity equilibria over the longer time spans of the pre-Quaternary record is fraught with potential sampling and taphonomic pitfalls [Alroy (1998) as in McKinney and Drake, 1998]. None-
theless, diversity equilibria models have been extended to evolutionary timescales (Sepkoski, 1991). In these models, diversity is hypothesized to be ultimately lim-
ited by rates of speciation and extinction. Equilibrium in this case is hypothesized to be due to "niche satu-
ration" and occurs when a clade has exploited all possi-
ble opportunities.

For several animal clades, at intermediate timescales, species origination rates appear to be richness depen-
dent, whereas extinction rates are not. Thus, as diversity increases, origination rates decline and diversity reaches an equilibrual number [Rosenzweig, 1995; Alroy (1998),
and McKinney (1998) as in McKinney and Drake,
1998]. A similar pattern is also observed in the plant fossil record. During the initial debut of each major land plant clade, speciation rates tend to be high and species longevity is short, but subsequent rates of origi-
nation tend to decrease while species longevities in-
crase (Fig 3). Furthermore, extinction rates have re-
mained essentially independent of standing diversity (Fig 4). Species diversity equilibria, however, are at-
tained only for short and intermediate intervals of geo-
logical time. Over longer periods, diversity has progres-
sively increased (Fig 3). It is not difficult to understand
why: Species equilibria require stable environmental conditions and an unchanging field of players. Over

![Figure 3](image-url)

**Figure 3**: Origination rates of vascular plants. Origination rates of pteridophytes, gymnosperms, and angiosperms are plotted as a function of geological time. Curved line denotes ordinary least squares regression for total standing diversity. Rates are
computed as the appearance of new species per unit time (relevant geological stage) per standing local diversity. Exceptionally high rates at the appearance of each group reflect low standing diversity of the group at first appearance (data from Niklas, 1997).
XI. LATITUDINAL GRADIENTS

An important macroecological pattern of diversity is the latitudinal diversity gradient, which is the increase in alpha and beta diversity from polar to tropical regions. This gradient is reflected by both the number of higher taxa and the number of species nested within those taxa [Ricklefs (1989) as cited in Ricklefs and Schluter, 1993]. For example, typically there are approximately 1–2.3 species per plant family and 12 plant families per 0.1-ha area plots within temperate tree communities. In contrast, tropical tree communities of the same size have between 2 and 4 species per family and as many as 38 families [Gentry (1988) as cited in Ricklefs and Schluter, 1993]. Evidence from the fossil record suggests that a latitudinal gradient in angiosperm diversity has existed since the origin of this clade. The Eocene floras of 40–50°N latitude were more diverse than those at 70°N latitude during this relatively warm geological interval. The same pattern is seen when comparisons are made among 20- to 16-million-year-old floras found in eastern or western North America and the Canadian Arctic.

Many hypotheses have been advanced to explain the latitudinal gradient in species diversity. The most promising propose multiple processes and consider both local- and global-scale patterns, invoking gradients of productivity and energy availability that correlate with changes in latitude [Brown, 1995]. Productivity ultimately controls biodiversity, and conditions that favor high productivity, such as warm temperatures and ample precipitation, are often associated with high diversity [Currie and Paquin (1987) as cited in Rosenberg, 1995].

Differential origination rates from the tropics to the poles may also play an important role in creating and maintaining the latitudinal diversity gradient, although a fundamental mechanism for this pattern has yet to be identified. In general, it appears that major evolutionary innovations and speciation events occur more frequently in lower rather than higher latitudes and that these innovations subsequently spread toward the poles over comparatively long periods of time. The hypothesis of the equatorial region as a source of novelty is sup-
ported by trends in the invasion of land, the origin of seeds and angiosperms, and even the invasion of land by tetrapods, all of which appear to comply with a "tropics to the poles" trend (Jablonski, 1993).

XII. ABUNDANCE

As ubiquitous as the latitudinal diversity gradient but possibly more mysterious are the patterns of abundance that are found in both modern and historical communities. Although it is generally agreed that increasing resource availability can increase diversity, very little is understood about how these resources are divided among members of a community. Abundance patterns in extant communities tend to conform to a "canonical" pattern: Most species are rare, whereas very few are common [Preston (1948, 1962) as cited in Brown, 1995; Rosenzweig, 1995; Rosenzweig (1998) as cited in McKinney and Drake, 1998].

Unfortunately, examination of past abundance patterns is confounded by preservational features. The plant fossil record consists of plant parts (dispersed leaves, fruits, wood, etc.) that are preserved in a manner that imperfectly reflects the abundance of species in communities. A deciduous tree, for example, within an otherwise evergreen forest might appear to dominate due to excessive production and shedding of organs. However, several Tertiary floras and a few Carboniferous ones are both geographically extensive and widely

![Graph of floristic composition over geological time.](image-url)

**FIGURE 5** Changes in floristic composition. Taxonomic composition of representative floras (expressed as the percentage of each major group of vascular land plants) are plotted against geological time. Rhyniophytes and taxonomically "problematic" groups of plants dominate the first land plant floras, angiosperms dominate the most recent fossil plant floras.
sampled, and they suggest that canonical abundance distributions may be a common feature in historical communities.

Dominated within a community, as reflected by abundance, appears to be a transient feature for an individual species. Indeed, the fossil record indicates that most species that were once common typically become rare with time [Knoll (1986) as cited in Rosenzweig, 1995]. At a higher taxonomic level, however, certain groups do exhibit long-term dominance (e.g., the Mesozoic is the age of conifers and cycadophytes). This dominance likely reflects the increased genetic, developmental, morphological, and reproductive diversity inherent in pooling larger numbers of species within higher taxa. When members of a clade breach a developmental barrier, radically new solutions to life may arise that may allow for increased abundance via the colonization of new environments and/or the ability to outcompete with older clades. This dynamic may account for major transitions in dominance such as that between gymnosperms and angiosperms (Fig. 5).

XIII. SUMMARY

Examination of the major patterns of plant diversity from both contemporary and historical communities holds the promise of illuminating the fundamental mechanisms responsible for the origin and maintenance of biological diversity. Here we have outlined several “rules” and patterns that appear to have been important in the evolution of plant communities. In many situations, the contemporary and historical evidence clearly complement each other. In other cases, such as the relative importance of competition, they appear to conflict. The latter cases may reflect the inherent research limitations of paleontology and neeology, the real effects of temporal scale, or the operation of different processes.

Of particular note to ecologists interested in diversity should be identification of those factors that cause diversity to change at both short and long timescales. Dramatic climate change and evolutionary innovation appear to be the most important features in the historical record. It is interesting to note that the current diversity—stability debate has considered these variables not as cause but as effect, examining the influence of diversity on community and ecosystem processes rather than vice versa. We live in the most diverse terrestrial communities in the history of Earth, but they function as the result of the same basic rules of interaction, energetics, and stoichiometry as did their Paleozoic predecessors. When events, human or otherwise, create an extraordinary disturbance, biodiversity may plummet and modern communities, as we recognize them, may cease to exist. However, the evidence of 400 million years strongly suggests that the survivors will reorganize into communities and continue to function following the same basic physical, chemical, and ecological rules that created the world around us. They have done so before, and there is good reason to assume that they will continue to do so as long as this planet harbors life.

Acknowledgments

This article was completed as part of the Body-size Working Group supported by the National Center for Ecological Analysis and Synthesis, a center funded by NSF (Grant DEB-94–21335), the University of California at Santa Barbara (UCSB), and the state of California. We thank J. Damuth, W. DiMichele, F. A. Smith, and J. H. Brown for commenting on initial versions of this article. BJE was supported by a NSF Postdoctoral Fellowship JPH was supported by NSF Grant DEB–9717406 to J. H. Brown and by the National Center for Ecological Analysis and Synthesis. BHT acknowledges research support provided by the College of Creative Studies, UCSB.

See Also the Following Articles

COEVOLUTION • DIVERSITY, COMMUNITY/REGIONAL LEVEL • HABITAT AND NICHE, CONCEPT OF • LATITUDE, COMMON TRENDS WITHIN • PALEOECOLOGY • PLANT BIODIVERSITY, OVERVIEW

Bibliography


PLANT CONSERVATION

Mike Maunder
Royal Botanic Gardens

I. A History of Plant Conservation
II. Distribution and Loss of Plant Diversity
III. Human Influences on Plant Diversity
IV. Plant Extinctions: How Many and Where?
V. Responses to Biodiversity Loss
VI. Ex Situ Conservation
VII. Toward an Integrated Approach
VIII. Facilities and Skills for Plant Conservation

GLOSSARY

ark paradigm The concept that threatened species can be preserved in or at a special facility; a term derived from the Biblical account of Noah’s Ark saving every extant species on earth during a great flood.
ex situ Literally, away from the site or location; in this context referring to conservation efforts elsewhere than the natural habitat; e.g., in botanical gardens.
hot-spot An area indentified as having an unusually large number of plant species, a high proportion of endemic (localized) species, and a threat of habitat destruction.
in situ Literally, in or at the site or location; in this context referring to conservation efforts within the natural habitat.
in vitro Literally, in glass; i.e., in a test tube; more broadly, in a laboratory or other artificial setting rather than in nature.
living dead A term for a species in which scattered adult individuals still survive but the capacity for reproduction has been lost; thus extinction will occur when the present generation dies out.

PLANT CONSERVATION is the management of plant resources to maintain current levels of plant diversity and to avoid population and taxonomic extinctions. Historically, plant conservation has been overshadowed by the politically more pressing issues of wildlife conservation. This article provides a historical review of plant conservation issues and provides an overview of present approaches and activities.

I. A HISTORY OF PLANT CONSERVATION

Human societies have historically managed plant resources to ensure dependable access to timber, fruits, roots, and medicinals. However, for both rural communities and larger political dynasties, this management has been driven largely by utility. The cedars of Lebanon, Cedrus libani, now surviving as scattered and isolated groves in Lebanon, Syria, and Turkey and covering probably less than 5% of their historical area, have been subject to fluctuating periods of management and overexploitation for over 2000 years. Indeed the marker stones to Roman forest reserves can still be found on Lebanese hillsides. The concept of human-caused habitat degradation and species loss was established only
during the eighteenth century, as a reaction to catastrophic habitat changes following colonial occupation of oceanic islands such as Mauritius (Grove, 1996). One of the earliest ex situ conservation attempts can be attributed to Governor Byfield of St. Helena, who in the early eighteenth century collected two seedlings of the then very scarce St. Helena redwood, Trochodendron erythroxylon, for cultivation in his garden on the island. The recorded extinction of Franklinia alatamaha in the United States established the concept of plant extinctions. During the late nineteenth century and early twentieth century, a small number of species were assumed to have become extinct in the wild, most notably Ginkgo biloba and Amherstia nobilis. Both species were considered to be extinct in the wild and only surviving in Asian temple gardens. However, they appear to have been treated as isolated novelties and did not prompt any general concern from the botanical community. In contrast, animal extinctions attracted significant attention, as early as the 1980s the National Zoo, Washington, DC, was proposed as a ‘home and a city of refuge for the vanishing races of the continent.’

The broader environmental movement can be traced, in part, to the late nineteenth century protected area movement in the United States (the Yellowstone National Park Act of 1872) and the wilderness-inspired literature of Aldo Leopold. The 1960s saw the start of a public concern for the environment and the beginning of international and national structures for conservation. This concern led to a public and political environmental awareness of the 1970s as manifested through the establishment of Earth Day on April 22, 1970, in the United States and the Endangered Species Act of 1973. The Russian agricultural botanist N. I. Vavilov, during the 1920s and 1930s, promoted the value of crop relatives and wild species in supporting agriculture. The present concerns about plant conservation, as both a political and scientific issue, can be traced to a series of international conferences on plant genetic resources during the late 1960s and 1970s. The erosion of these resources was recognized as an urgent problem at a Food and Agriculture Organization (FAO) conference in 1961 and subsequently discussed by a joint FAO and International Biological Programme (IBP) meeting in 1967, where the term “genetic resources” was introduced. The FAO Panel of Experts on Plant Exploration and Introduction led the debate on plant resources from the mid-sixties to seventies, with the creation in 1974 of the International Board for Plant Genetic Resources (IBPGR, now International Plant Genetic Resources Institute [IPGRI]).

A widely recognized initial impetus to conservation was derived from the United Nations Conference on the Human Environment held at Stockholm in 1972. The Threatened Plants Committee (TPC) of the International Union for the Conservation of Nature (IUCN) was established in 1974 as a direct product of the initial Red Data Book for Threatened Plant Species. This network was designed to deal specifically with plant species outside of the remit of the FAO. The TPC acted as the template for the subsequent evolution of the IUCN’s Species Survival Commission (SSC).

The discussions in the early and mid-1970s set the stage for botanical gardens to take up plant conservation as a serious responsibility. The ‘Ark Paradigm,’ the idea that ex situ facilities would hold stocks of threatened plants during a period of habitat degradation, the “demospheric winter” sensu Soule (1991), was established as a working objective by botanical gardens in the 1970s. This attitude is manifest in the 1978 Red Data Book (Lucas and Synge, 1978); for instance, in the data sheet for Dracaena ombet, it was stated that “it seems too late for such a proposal [in situ conservation] to be worthwhile. Great efforts must now be made to bring the ombet into cultivation and maintain it safely in the botanical gardens of the world.” A notable conference in the late 1970s was “Extinction is Forever” (Prance and Elias, 1977); it provided a regional overview of American plant and habitat conservation issues with both a geographical and taxonomic focus. This conference attempted to assess levels of species decline and loss in the region and subsequently set the scene for later developments in tropical forest inventory. Two conferences hosted by the Royal Botanic Gardens, Kew, in 1976 and 1978, established the agenda for botanical garden activities over the next two decades (Simmons et al., 1976; Synge and Townsend, 1979). As a result of the 1976 conference, the Threatened Plants Committee of the SSC of the IUCN was requested to establish a Botanic Gardens Conservation Co-ordinating Body “to find out which threatened plants are in cultivation and where, and to keep botanical gardens informed of current conservation activities.” This need was reiterated at the First International Botanic Gardens Conservation Congress in Las Palmas, Canary Islands, in 1985 with the recommendation of the establishment by IUCN of the Botanic Gardens Conservation Secretariat (BGCS), later to become Botanic Gardens Conservation International (BGCI). In parallel with discussions on botanical gardens, a number of broader conservation references were published in the early 1980s that paved the way for the 1992 UNCED meeting. These included The Brandt Report (Independent Commission on International Development Issues, 1980), the Global 2000
Report (U.S. Council on Environmental Quality and U.S. Department of State, 1980), and the World Conservation Strategy (IUCN/UNEP/WWF, 1980). All three reports served to emphasize ongoing loss of species and habitats and the linkages to human welfare and development.

The Convention on Biological Diversity (CBD) is the major legislative influence on the conservation of global biodiversity; directly influencing national activities through the requirement to produce and implement Biodiversity Action Plans.

The main objectives of the CBD are:

• The conservation of biological diversity (Articles 6–9, 11, and 14)

• The sustainable use of its components (Articles 6, 10, and 14)

• The fair and equitable sharing of the benefits arising from the use of genetic resources, including by appropriate (1) access to genetic resources (Article 15), taking into account all rights over those resources; (2) transfer of relevant technologies (Articles 16 and 19), taking into account all rights to technologies; and (3) funding (Articles 20 and 21).

Outlined under Article 6 (General Measures for Conservation and Sustainable Use) is the requirement for each Contracting Party to develop a national biodiversity strategy, action plans, or program. The primary function is to make specific recommendations for national action through identifying areas for action, iden-
FIGURE 3 Encephalartos woodii. This cycad is extinct in the wild and only survives as a male clone in botanic garden collections. It is an example of a cultivated member of the "living dead."

FIGURE 4 Restoration of a degraded island habitat on Round Island, Mauritius. A collaborative exercise between the government of Mauritius and the Mauritian Wildlife Fund with support from the Durrell Wildlife Conservation Trust, Royal Botanic Gardens, Kew, and Fauna and Flora International is restoring the original forest after control of introduced goats and rabbits.

Identifying obstacles, such as national capacity, finances, technology, inadequate legal mechanisms, etc.; identifying relevant government sectors and affected constituencies; identifying cost-effective solutions; and assigning tasks. In general, this study will be undertaken through a National Biodiversity Unit.

The past 20 years has seen a revolution in the available technology for supporting plant conservation. Fundamental to this has been the explosion in information technology allowing the collection, exchange, and analysis of data pertaining to the wild status of species. The past two decades has seen the move from record keeping in a ledger and the largely verbal transmission of associated information toward the increased use of computerized records and geographical information systems (GIS). Based upon the scientific advances attained in the fields of crop genetic resources, there have been major advances in seed storage, cryopreservation, and in vitro propagation. The FAO Panel of Experts on Plant Exploration and Introduction drafted standards and procedures for the long-term storage of seed. Recent advances have included protocols for the storage of in vitro cultures. An increasing number of technical manuals are available to guide professionals (e.g., Given, 1994). The development of polymerase chain reaction (PCR) based molecular technologies has opened the door to identification and management of both evolutionary lineages and conservation units. The utility of these techniques is increasing as both levels of resolution increase and costs decrease. In addition, as the understanding of major phylogenetic relationships advances, it will become increasingly feasible to focus conservation activities on evolutionarily significant clades.

II. DISTRIBUTION AND LOSS OF PLANT DIVERSITY

Botanical diversity is concentrated into areas of unusual richness and exhibits gradients on both a global and regional scale. The taxonomic diversity of higher plants broadly correlates to latitude and is generally highest in the tropical and subtropical regions. Species richness and endemism were commonly used attributes for the identification of biodiversity priorities since they presumably reflected both the complexity and uniqueness of ecosystems. Accordingly, early discussions identified tropical rain forests and oceanic islands as conservation priorities.
In 1986, the International Union for the Conservation of Nature (IUCN) and the Worldwide Fund for Nature (WWF) initiated a project to identify major centers of plant diversity. The objectives of the project were to identify which areas around the world, if conserved, would safeguard the greatest number of plant species, to document the many benefits, economic and scientific, that conservation of those areas would bring to society and to outline the potential value of each for sustainable development; and to outline a strategy for the conservation of the areas selected. The project (Davis et al., 1994, 1996, 1997) has identified 234 areas of especially high species diversity, and correlates strongly with the listed 12 megadiversity countries (sensu Mittermeier et al., 1998) (Table I).

Myers (1988, 1990) identified conservation ‘hot spots’ areas, defined as areas having an exceptional concentration of plant species, having high endemism, and threatened with destruction. Myers initially identified 10 hot spots, all in tropical rain forest areas. The analysis was subsequently expanded to include four additional rain forest areas and four Mediterranean climate regions, including the Cape Floristic region.

The ‘megadiversity countries’ concept, developed by Mittermeier (1988) and expanded in Mittermeier et al. (1998), is based on three premises: (1) biodiversity is not evenly distributed among the world’s countries; (2) although international conservation priorities should be based on scientific information on biodiversity, it is governments that develop conservation policies and programs; and (3) a very small number of species; to document the many benefits, economic and scientific, that conservation of those areas would bring to society and to outline the potential value of each for sustainable development; and to outline a strategy for the conservation of the areas selected. The project (Davis et al., 1994, 1996, 1997) has identified 234 areas of especially high species diversity, and correlates strongly with the listed 12 megadiversity countries (sensu Mittermeier et al., 1998) (Table I).

Myers (1988, 1990) identified conservation ‘hot spots’ areas, defined as areas having an exceptional concentration of plant species, having high endemism, and threatened with destruction. Myers initially identified 10 hot spots, all in tropical rain forest areas. The analysis was subsequently expanded to include four additional rain forest areas and four Mediterranean climate regions, including the Cape Floristic region.

The ‘megadiversity countries’ concept, developed by Mittermeier (1988) and expanded in Mittermeier et al. (1998), is based on three premises: (1) biodiversity is not evenly distributed among the world’s countries; (2) although international conservation priorities should be based on scientific information on biodiversity, it is governments that develop conservation policies and programs; and (3) a very small number of countries, lying wholly or partly within the tropics, contain a high percentage of the world’s species. The first phase of analyses identified 12 centers of diversity to be mostly in the tropical regions. This initial focus on the megadiverse tropical countries overlooked the diversity of Mediterranean and subtropical floras; for instance, the five Mediterranean climate regions of the globe have been transformed, managed, or utilized by man (Turner et al., 1990); nearly 40% of the world’s plant diversity (Mittermeier et al., 1998).

III. HUMAN INFLUENCES ON PLANT DIVERSITY

Historically, about half of the ice-free surface of the globe has been transformed, managed, or utilized by man (Turner et al., 1990); nearly 40% of the potential terrestrial net primary productivity of the Earth is used by mankind or lost as a result of land change (Vitousek, 1994). Hannah et al., (1994) concluded that 73% of the Earth’s land surface, other than rock, ice, and barren land, is either human dominated (36.3%) or partially disturbed (36.7%), with only 27% undisturbed. Mackinnon and Mackinnon (1986a, 1986b) estimate that 65% of the original ecosystems south of the Sahara have been subject to major ecological disturbance and that 68% of the natural habitat has been lost in the Indo-Malayan nations. This trend will continue as human populations and their demands grow.

The extensive conversion of wild habitats has resulted in changes to both the ecological and taxonomic composition of these areas. In many parts of the world the change is almost complete; taxonomically and ecologically diverse wildlands have been converted to agricultural landscapes dominated by a small range of domesticates, e.g., cereals. In addition, the modernization of agriculture in traditional landscapes has led to profound environmental changes and high levels of plant species extinctions following Polynesian settlement. One example, the Pacific island...
of Rapa Nui (Easter Island), underwent massive environmental degradation as long ago as 1200 to 800 B.C. In the Caribbean, Atlantic islands, and Mascarenes, environmental degradation can be attributed to European colonial administration and the development of unstable and fragile economies based upon plantation products. Colonial settlement would often focus on the clearance and agricultural development of the lowlands, with the retention, planned or otherwise, of upland habitats, albeit in a modified and fragmented form. Examples of islands with severely modified lowlands include Mauritius, the Gulf of Guinea islands, and New Caledonia. Although the island of Sao Tome, Republica Democrata de Sao Tome and Principe, has over 90% forest cover, the remaining primary forest is restricted to high ground and steep slopes. While the lowlands of Gran Canaria, Islas Canarias, were converted to plantations, the uplands were decimated as a source of both firewood and topsoil. On Gran Canaria less than 1% of the original plant biodiversity hot spots survives. Some islands have suffered the virtual complete loss of original habitats (Box 1). This can result from periods of over-exploitation over centuries. On Rapa Nui, no original plant communities have survived the Polynesian settlement and later colonial sheep ranching. On St. Helena, only fragments of gumwood forest and montane tree fern thickets have survived, covering less than 1% of the island’s land area (Maunder et al., 1995).

### Table 1: Major Plant Diversity Countries

<table>
<thead>
<tr>
<th>Number of megadiversity centers of plant diversity</th>
<th>Number of vascular plant taxa</th>
<th>Biodiversity hot spots with number of endemic plant species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>Megadiversity country ranking</td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Colombia</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>China</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Mexico</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Australia</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Indonesia</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Peru</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Ecuador</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Malaysia</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>India</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Zaire</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Madagascar</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>United States</td>
<td>n/a</td>
<td>8</td>
</tr>
<tr>
<td>Philippines</td>
<td>n/a</td>
<td>6</td>
</tr>
<tr>
<td>Republic of South Africa</td>
<td>n/a</td>
<td>6</td>
</tr>
<tr>
<td>Turkey</td>
<td>n/a</td>
<td>6</td>
</tr>
</tbody>
</table>

*This table illustrates the concentration of plant diversity in 16 countries holding half of the world’s centers of plant diversity. The 12 megadiversity countries are ranked in column two. Total number of centers of plant diversity in the 16 countries is 140/234 c. 60%. The 16 countries contain (entirely or in part) 18/24 of the most important plant biodiversity hot spots. Data are derived from McNeely et al. (1990), Myers (1990), and Davis et al. (1994, 1996, 1997).
total transformation of an oceanic island can result from modern industrial exploitation, for instance, phosphate mining on Nauru in the Pacific.

As a result of deliberate introductions and as artifacts of the agricultural and horticultural industries, an increasing number of exotic species are growing in new territories and habitats (Box 1; Fig. 5). While such introductions may in the short term increase local biodiversity, in the long term such introductions are having disastrous impacts on both species diversity and ecological processes. Disastrous invasions include the impact of guava in Mauritius, miconia on the Polynesian Islands, and water hyacinth in the East African Rift Valley. In addition, introduced diseases, such as chestnut blight in the United States, and pathogens, such as the New Zealand flatworm in the United Kingdom, are impacting upon plant diversity. Human activities, notably the combustion of fossil fuels and the subsequent release of carbon dioxide, are altering the composition of the atmosphere. It is predicted that the historic concentration of carbon dioxide.

---

**Box 1**

**Case Study: The South Atlantic Island of Ascension—Too Late for Habitat Restoration?**

On discovery in 1501, the island was a bare volcanic landscape, with vast seabird colonies supporting only 25 vascular plant species; of these 11 are endemic. However, as a result of nineteenth century garrison settlement, a large number of exotic plants were introduced for both agricultural and horticultural purposes (Fig. 5). The effects of exotic species have transformed the original bare volcanic landscape of Ascension into a semiforested one. Successive importations of plants from British and South African botanic gardens represented massive introductions of both plant species and associated invertebrate and fungal diversity. Consequently, the original upland Marattia fern thicket has been transformed into an exotic woodland dominated by an introduced bamboo. In the middle altitudes, introduced Casuarina, Araucaria, and Juniperus bermudiana are forming open woodlands. The xeric lowland lava fields have been transformed into open woodland of Prosopis juliflora with introduced African grass species. The Prosopis is now encroaching on the turtle (Chelonia mydas) nesting beaches and threatening to expand onto the unique breeding colonies of Sooty Tern (Sterna fuscata). The transformation of terrestrial ecology has been massive and habitat restoration is no longer feasible; the priority should be to maintain and promote specific species and communities, e.g., seabird colonies, through the control of introduced predators (cats and rats) and herbivores (sheep and donkeys). The net effect has been an increase in total botanical diversity and primary productivity and the creation of a new synthetic ecosystem.
Botanic gardens in the tropics are playing a real role in the conservation and sustainable use of plant resources. The Entebbe Botanic Garden, Uganda, is working with traditional healers to catalog and propagate medicinal plants.

In the Earth's atmosphere may double in the next century and change global climate. This could result in an increase in mean global temperature of approximately 1.5–4.5 °C during the twenty-first century (IPCC, 1992). Botanical diversity is expected to change as species react through both local extinctions and the colonization of new areas. Ecological processes will be modified as fire and hydrological systems alter. It is expected that these changes will cause the extinctions of plant species as populations die out and are unable to migrate between isolated habitats (Morse et al., 1995). It is likely that invasive weed species will dominate many of these new and unstable plant communities.

All of these damaging impacts are driven by one fundamental force, the increase in human population. By the year 2025 the world population is projected to total about 8.3 billion people, compared with the current estimate of 5.7 billion. The patterns of increase will vary regionally; European populations may decline slightly, Asia could grow from 4.96 billion, while the population of Africa will possibly double between 1995 and 2025, from 728 million to 1.49 billion (WRI/UNEP/UNDP/World Bank, 1996).

As populations increase, there will be continued demand for both new and traditional plant products; for instance, between 4000 and 6000 products of medicinal plants are traded internationally. One new wild product is derived from Prunus africana, harvested from montane African forests for its bark, from which alkaloids are extracted for the treatment of prostate disorders. The annual over the counter trade is worth over $200 million per year.

IV. PLANT EXTINCTIONS: HOW MANY AND WHERE?

Plant resources are being exploited to the point of both economic and biological extinction. The exhaustion of the Caribbean mahogany (Swietenia spp.) and the Mauritian ebony (Diospyros spp.) are examples of island resources taken to economic extinction, while the Juan Fernandez endemic sandalwood (Santalum fernandezianum) has been exploited to biological extinction. Horticultural exploitation has resulted in a small number of species being eradicated in the wild and surviving in cultivation, an example being the Chilean crocus Tecophilae cyanocrocus.

Since 1600, it is estimated that 654 flowering plant species have become extinct (WCMC, 1992). This is certainly an underestimate as many areas of the world (e.g., the Mediterranean Basin) have undergone massive levels of habitat destruction, with presumed species loss prior to scientific inventory. The most comprehensive survey of species decline undertaken on behalf of the IUCN indicates that about 33,400 plant species are threatened with extinction, equating to about 10% of the world's 250,000–300,000 plant species (Walter and Gillett, 1998). For the largest part of the planet, there is no clear consensus on the rate of species and population loss. It could be argued that the plant extinctions logged by the IUCN and World Conservation Monitoring Centre (WCMC) reflect, in part, the geographical distribution of botanical knowledge and monitoring, rather than actual rates of species loss; for instance, all the recorded sub-Saharan plant extinctions (45) are from the Republic of South Africa alone.

Koopowitz et al. (1994), estimating both historical and prevailing rates of extinction based upon rates of habitat conversion and distribution of restricted endemic plant species for tropical Latin America, produced figures exhibiting a dramatic lack of congruency with the WCMC/IUCN records. This is particularly notable for Brazil, where WCMC/IUCN has recorded only 3 extinctions and Koopowitz estimates a loss since 1950 of 2261 species. It is evident that the data pertaining to plant extinctions are not sufficient to identify in advance which plant species are at greatest risk of extinction; the paucity of field survey work and rapidity of habitat loss mean that many species' extinction will be identified only in retrospect. A few attempts have been made to identify extinction-prone tropical forest species in advance (Martini et al., 1994).

Apart from the well-publicized estimates of loss from the tropics, it is evident that the European nations are
Plant conservationists face a dilemma in dealing with the living dead, species that are surviving as scattered, nonreproductive, individuals isolated from their original habitats (Fig. 3). Oceanic islands having suffered massive levels of habitat loss contain examples of the living dead. Both *Knobia cookei* from Hawaii and *Ramosomunia rodriguesiana* from Rodrigues survive only as sterile single clones. The endemic palm, *Hyophorbe amarissana*, survives only as a single tree in the Curepipe Botanic Gardens, southwest Mauritius. The single specimen has persisted within the botanical garden for over 30 years with no regeneration recorded. The specimen produces flowers and fruits regularly, but attempts to grow these using both micropropagation and conventional horticultural techniques have not been successful.

## V. RESPONSES TO BIODIVERSITY LOSS

Conservation actions are implemented to mitigate or reverse the damaging impacts of human social, demographic, and economic change on plant diversity. Management actions aimed at the conservation of biodiversity will take place at various levels of the biodiversity hierarchy, a nested hierarchy of spatially defined units with often ill-defined boundaries (Soule, 1991). These efforts vary in the scale of spatial, capital, and scientific investment, encompassing single-species management through to wilderness retention. Increasingly, the traditional schism between on-site activities (in situ conservation) and off-site activities with a species or genetic focus (ex situ conservation) is eroding. Conservation activities focused on any one layer of the biospatial hierarchy should take into account the linkages to other levels. An integrated approach to plant conservation is being promoted encompassing species recovery programs and habitat management (Falk et al., 1996). A parallel development has been an increasing trend toward ecosystem-level conservation in both the temperate and tropical regions (Wecks, 1996). In addition, the linkages between biodiversity management and sustainable development have been given due recognition (Barzen, 1993).

The majority of the world’s species will be retained through the “coarse filter” approach of habitat conservation; this potentially could conserve all levels in the biodiversity hierarchy and their interactions. However, many protected areas will require increasing management because of external influences impacting on ecological processes and promoting changes in both community structure and composition. Protected area borders are permeable to disease, invasive species, poaching, civil unrest, climate change, etc. Protected areas have been established with the assumption that environmental conditions and community patterns/composition have been relatively stable for long periods in the past and will continue to be stable into the future. It could be argued that habitats are loosely organized collections of species whose coexistence is dependent on their individual limits and subsequent distribution along environmental gradients. On a geological time scale, they could be viewed as relatively transient assemblages. Accordingly, a “fine filter” approach will be required as a backup to catch those species not secured through the priority action of habitat conservation.

The surviving major wilderness areas (Hannah et al., 1994; Mittermeier et al., 1998), large relatively undisturbed natural areas, offer the best opportunities for retaining ecosystem and evolutionary processes. Beyond these areas, plant conservation will depend upon a number of core skills: (1) the protection and active management of habitats to maintain plant diversity and ecological processes, (2) the management of individual plant populations to retain viable populations, and (3) dealing with the human context of plant conservation.

**Box 2**

**Case Study: The “Living Dead”**

Plant conservationists face a dilemma in dealing with the living dead, species that are surviving as scattered, nonreproductive, individuals isolated from their original habitats (Fig. 3). Oceanic islands having suffered massive levels of habitat loss contain examples of the living dead. Both *Knobia cookei* from Hawaii and *Ramosomunia rodriguesiana* from Rodrigues survive only as sterile single clones. The endemic palm, *Hyophorbe amarissana*, survives only as a single tree in the Curepipe Botanic Gardens, southwest Mauritius. The single specimen has persisted within the botanical garden for over 30 years with no regeneration recorded. The specimen produces flowers and fruits regularly, but attempts to grow these using both micropropagation and conventional horticultural techniques have not been successful.
Increasingly, ex situ approaches are being used in species conservation. A Species Recovery Plan provides a forum to bring together all required expertise to ensure a balanced integration of ecological, taxonomic, genetic, and social studies. Such a program should have clear, numerically based objectives utilizing, where appropriate, information from ecological, taxonomic, genetic, and social studies. A Species Recovery Plan provides a forum to bring all required expertise together to ensure a balanced integrated approach to species conservation.

VI. EX SITU CONSERVATION

Increasingly, ex situ collections are promoted as integral components of conservation programs for both wild taxa and crop genetic resources. The world’s ex situ plant resources encompass the traditional garden plots of the tropics, the intensive allotments, farms and gardens of Europe, and the global networks of seed banks and botanical gardens. The world’s network of plant genetic resource collections (e.g., seed banks and field gene banks) contains relatively few species, but with high levels of infraspecific sampling. The network maintains some 6 million accessions, representing mostly agricultural crop cultivars and landraces. The most important 150 major crop gene banks are maintained, as part of the global network of the International Agricultural Research Centers (IARCs), under the umbrella of the Consultative Group on International Agricultural Research (CGIAR). Important centers include the International Rice Research Institute in the Philippines and the International Wheat and Maize Improvement Centre (CIMMYT) in Mexico. The total annual cost of maintaining all accessions currently in gene banks is estimated at $300 million per annum, with many facilities suffering a chronic lack of resources.

The largest collections of noncrop wild species are maintained by the world’s botanical garden networks; currently they are estimated to maintain 4 million accessions, representing 80,000 taxa. The majority of these collections are maintained as living plants in mixed collections serving a wide range of purposes (Maunder, 1994). There are approximately 1700 botanic gardens in 148 countries; however, over 40% of these are concentrated in western Europe and North America. These collections cultivate a skewed representation of the world’s wild diversity, but can be relatively comprehensive at the generic level for attractive and horticulturally amenable families such as the palms, cacti, bromeliads, and orchids. The levels of infraspecific genetic diversity are low since most species are represented by relatively few individuals, often from a limited number of introductions. These collections contain specimens of threatened species, including species now extinct in the wild, e.g., Sophora toro-miro (Maunder et al., 1999) (Fig. 2). Botanic gardens should be recognized as vitally important resources of horticultural skills and conservation biology research that can support the imperatives of in situ management of wild populations and conservation of economically useful plants (Fig. 7), and as readily accessible venues for public education. Botanic gardens have a fundamental role in public education as accessible and often urban venues for introducing the public to plant diversity. Large collections of horticultural plants, mostly cultivars, are maintained by commercial nurserymen and amateur horticulturists. In Europe, the collections closely overlap with botanical gardens at the taxonomic level for hardy herbaceous and woody taxa, but in contrast with botanic gardens, the level of documentation is often poor. Some countries have established national collections of garden plants; such networks exist in Australia, France, and the UK. The National Council for the Conservation of Plants and Gardens (NCCPG) in the UK coordinates over 600 collections, maintained by professional and amateur horticulturists, containing 13,000 species and 39,000 cultivars. Such networks can also serve noncommercial crop resources; the Seed Savers Exchange in the United States exchanges over 3000 crop varieties.

VII. TOWARD AN INTEGRATED APPROACH

Current trends of both habitat loss and conversion have promoted discussion on the most effective responses to retaining both the taxonomic and ecological components of biodiversity. Conservation biologists have realized that plant conservation management needs can be targeted at distinct, but interlinked, levels of the
biodiversity hierarchy, namely, whole systems at the landscape or ecosystem level, habitats, species, populations, and genes. Superimposed upon these patterns will be an increasing knowledge about plant phylogeny, allowing conservation issues to be guided by evolutionary perspectives.

Traditionally, conservation professionals have operated within isolated and distinct hierarchies; for instance, the ecosystem-level conservationists (protected areas) did not interact greatly with ex situ practitioners with a focus on the population/gene level (plant genetic resources and botanic gardens). Accordingly, plant conservation practice and debate have traditionally been polarized between in situ activities and ex situ activities. Recently, a number of authors have recognized that plant conservation strategies need to utilize a variety of complementary techniques, “integrated strategies” sensu Falk et al. (1996). Recovery Planning for threatened species is a relatively recent development, with the first United States plant recovery plan initiated in 1979. The later production of regional plant conservation strategies reflects this integrated approach, for instance, the Federal Native Plant Conservation Initiative of the United States and the Australian Network for Plant Conservation.

Habitat conversion as the major threat to biodiversity produces an “extinction debt,” a pool of species destined for extinction unless the habitat is repaired or restored (Tilman et al., 1994). While stocks of such taxa could be maintained ex situ for a future reintroduction, the physical and technical capacity does not currently exist to hold large numbers of taxa or individuals over time. In addition, the longer material is held in cultivation ex situ, the more likely it is that genetic modification will take place (Fig. 1). The restoration of those threatened habitats, particularly areas adjacent to protected areas, will provide a vital means of enabling a significant number of species to recover.

VIII. FACILITIES AND SKILLS FOR PLANT CONSERVATION

Plant conservation has often been marginalized as a peripheral activity for governments agencies involved in protected area or forestry management, often as a subset of “wildlife” management. The establishment of national plant conservation networks is playing an important role through (1) promoting an integrated approach to plant conservation, utilizing and promoting the available professional skills; (2) developing collaborative relationships with protected area networks, government agencies, parastatals, and NGOs; (3) establishing a network of research/conservation facilities in different climatic/vegetation zones; (4) developing in-country...
Wild plants will continue to be an integral component of rural life for millions of people. Traditional resource management practices will need to be protected and given a contemporary economic and legislative relevance, e.g., the cooperative government and NGO management of habitats in Brazil, Belize, and India. However, these activities must be built upon sound science and, particularly, an understanding of how plant populations respond to change. The future challenge is to maintain plant populations, as evolutionary lineages, as ecological components of functioning landscapes, and as valued economic resources, within a changing ecological and political climate. For instance, the 1980s saw the collapse of one of the world’s most effective protected area and plant genetic resource networks during Russia’s transition from communism to a free-market economy. The increased need for species and habitat management will require greater continuity of information between successive management regimes.

Plant conservation, encompassing the needs for managing increasingly fragmented habitats and populations within human-dominated landscapes (Box 3; Fig. 6), will face a number of challenges: (1) Sound conservation solutions must include strong elements of social science, resource economics, and commercial practice. (2) An “theoretical” awareness of the need for sustainable-use practices alone does not necessarily change practices. (3) Many important plant habitats are subject to increasing levels of encroachment by settlement and cultivation as pressure for land forces people further into protected areas. These areas are increasingly becoming ecological islands, and many of their enclosed species’ populations will increasingly face issues of viability. (4) As land between protected areas is progressively modified by human activities and its suitability for sustaining original levels of biodiversity will decline, the opportunities for linkage between prime habitats in parks will vary. However, it is evident that modified habitats can contribute to maintaining connectivity between protected areas, and increased research is urgently needed on this topic. (5) While there have been efforts to “soften” the edge between parks and the lands outside them, there is a continued need to develop schemes by which protected area resources and revenues are shared with neighboring communities.

Plant conservation could easily become focused on salvaging the lost (Box 2). Future investment in scientific and practical activity should focus increasingly on the retention of viable habitat areas and the management of socially and economically important plant resources. While a proportion of the world’s plant diversity can be saved as “threatened species,” funded as abstracted and emotionally charged conservation projects, the vast ma-

A number of international networks have established for plant conservation the plant genetic resources network under the International Plant Genetic Resources Institute (IPGRI) and the wild species network under the Species Survival Commission of the World Conservation Union (SSC/IUCN). While there is continued need to further consolidate international networks, a more urgent need is the establishment of local and regional networks to support and direct local frontline initiatives. Examples include the Korean Plant Specialist Group of the SSC, a voluntary network that links government, NGO, and academic interests in plant conservation.

The Center for Plant Conservation (CPC) in the United States can be viewed as one model for a national network. The center is based at the Missouri Botanical Garden, St. Louis, MO, and was established in 1984 as a national network of collaborating botanic gardens with the clear focus on conserving threatened native flora. The CPC has promoted the utilization of both in situ and ex situ techniques. Fundamental to the effectiveness of the CPC has been the availability of quality data on distribution and status of botanical diversity; this has been heavily dependent upon the activities of the Nature Conservancy’s Heritage Programs. In addition, CPC has produced two keystone references for plant conservation, namely, Genetics and Conservation of Rare Plants (Falk and Holsinger, 1991) and Restoring Diversity: Strategies for Reintroduction of Endangered Plants (Falk et al., 1996). This has resulted in the development of working collaborations between protected area authorities, government agencies, and the adoption of population genetics as a working tool for botanic garden conservation activities. National networks for plant conservation are now established in a number of countries, including Australia, Indonesia, and Canada.

There is an urgent need to incorporate plant conservation into national conservation infrastructures and in particular foster integral links with protected area agencies. The recent implementation of Biodiversity Action Plans is a great opportunity to consolidate the plant conservationist’s national role. Plant conservation needs to identify specific national and local incentives and sanctions that can be used as linkages between the central government institutions (the legal infrastructure) and the public sector (the financial infrastructure). Wild plants will continue to be an integral component of national and the public sector (the financial infrastructure).
jority of plant diversity will depend on a pragmatic bal-
ance between conservation and local utilization. This
will be dependent upon establishing effective monitor-
ning and conservation tools that can be locally adopted
and modified.

Bibliography
Development in Latin America and the Caribbean. IUCN (Protected
Areas Programme), Gland, Switzerland.

Species: Conceptual Issues, Planning and Implementation. Cam-
bridge Univ. Press, Cambridge, UK.

Burgess, M. A. (1999). Cultural responsibility in the preservation of

forests of Eastern Africa: Status, endemism patterns and their

Cox, A., and Caballero, J. (1996). Traditional management and
morphological variation in Leucaena leucocephala (Fabaceae: Mimo-
sideae) in the Meso region of Guerrero, Mexico. Econ. Bot. 50(2),
167–181.

Centres of Plant Diversity: A Strategy for Their Conservation. Europe,
Africa, South West Asia and the Middle East, Vol. I. IUCN/WWF,
Gland, Switzerland.

Centres of Plant Diversity: A Strategy for Their Conservation. Asia,
Australia and the Pacific, Vol. 2. IUCN/WWF, Gland, Switzerland.

Centres of Plant Diversity: A Strategy for Their Conservation. The
Americas, Vol. 3. IUCN/WWF, Gland, Switzerland.

Falk, D., Miller, C. L., and Powell, M. (Eds.) (1996). Restoring Diver-
sity: Strategies for Restoration of Endangered Plants. Island Press,
Washington, D.C.

Timber Press, Portland, OR, and Chapman & Hall, London/ New
York.

Grégoire, A., de Brein-Guilmot, F., Synge, H., McNeely, J. A., and
Diversity: Environment Policy and Law Paper No. 30. IUCN,
Gland, Switzerland.


Cambridge Univ. Press, Cambridge, UK.

on genetic diversity and mating system in a tropical tree, Pachyca-

Hamdan, L., Lohe, D., Hutchinson, C., Cari J., L. and Lankawati,

and the conservation of biological diversity: In Conservation Biology:
The Theory and Practice of Nature Conservation, Preservation and
Management (P. L. Fiedler and S. K. Jain, Eds.), pp. 197–237.


model for the prediction of biodiversity losses based on habitat

IUCN, Morges, Switzerland.

biodiversity conservation with particular reference to alien
invertebrates. In Islands: Biological Diversity and Ecosystem Function
Springer-Verlag, Berlin/Heidelberg.

Arms System in the Indo-Malaysiaan Region. IUCN, Gland, Swit-
zeland.

Arms System in the Indo-Malaysiaan Region. IUCN, Gland, Swit-
zeland.

predict which Amazonian tree species may be threatened by log-
ging activities. Environ. Conserv. 21(2), 152–162.

Mann, M. (1999). Botanic garden: Futures challenges and respon-
sibilities. Biodiversity Cons. 3, 97–103.

Mann, M., Cullum, A., Bowles, A., Allangumale, J., and Wilm-
son, M. (1999). Genetic diversity and pedigree for Sophora inter-
media (Leguminosae): A tree extinct in the wild. Mol. Ecol. 8, 725–736.

Saint Helena Sustainable development and conservation of a
highly degraded island ecosystem. In Islands: Biological
Diversity and Ecosystem Function (P. M. Vitousek, L. L. Loope,
York.

studies from Brazil, Madagascar and the importance of megadiver-
National Academy Press, Washington, D.C.

Mittermeier, R. A., Myers, N., Trouer, M. J. B., da Fonseca, G. A. B.,
and Oliver, S. (1998). Biodiversity hotspots and major tropical
wilderness areas: Approaches to setting conservation priorities.

of climate change on North American flora. In Our Living Re-
sources: A Report to the Nation on the Distribution, Abundance,
and Health of U.S. Plants, Animals and Ecosystems (E. T. Lutton, G. S.
Farr, C. E. Funkenst, P. D. Doran, and M. J. Mad, Eds.) U.S. De-
partment of the Interior, National Biological Service, Wash-
ington, D.C.

Environmentalist 8, 1–20.

Myers, N. (1990). The biodiversity challenge: Expanded hotspot anal-
ysis. Environmental 10, 245–255.


Handbook for Prairies, Savannas and Woodlands. Island Press,
Washington, D.C.

IUCN, Gland, Switzerland.

IUCN, Gland, Switzerland.

The Reconstitution of Fragmented Habitats: Global and Regional
Perspectives. Surrey Beatty & Sons, Ltd., Chipping Norton, Aus-
tralia.
I. Plant Hybridization
II. Plant Hybrid Zones
III. Genetic Basis of Hybrid Plant Resistance
IV. Communities on Plant Hybrids
V. Population and Evolutionary Consequences
VI. Conservation of Hybrid Plants
VII. Conclusions

GLOSSARY

allopolyploid (amphiploid) Species with chromosome sets derived from interspecific hybridization and doubling of chromosomes of the sterile hybrid, which restores fertility.
dependent community The group of species that requires a particular host plant to complete some or all of their life cycle.
diploid hybrids Species formed from hybridization where chromosomes from species of equal chromosome numbers become stabilized through recombination.
hybrid An individual that is heterozygous (intermediate) for one or more heritable characters that distinguish two or more populations, including individuals that are F1s, F2s, and the set of all backcrosses. Hybrids are initially formed when gametes from two species, subspecies, or races combine to form F1 plants.
hybridization Interbreeding of individuals from two or more populations of species, subspecies, or races, which are distinguishable by one or more heritable characters.
hybrid swarm Mixture of hybrid genotypes, including F1s, F2s, and backcrosses, due to hybridization between two or more species that co-occur in a locality.
hybrid zone Areas of overlap or points of contact between two populations that are distinguishable based on one or more heritable characters where viable or partially fertile hybrids are formed.
introgression Permanent transfer of genes from one or more species, subspecies, or race into another species, subspecies, or race via hybridization and backcrossing.
reticulate evolution Pattern of speciation where new species arise from interspecific hybridization between two species coupled with establishment of reproductive isolation.

PLANTS HAVE PERVERSIVE effects on biological diversity. The diversity and structure of plants are used to define all ecological communities. But it is the biological diversity of organisms that live on plants, their herbivores, pathogens, pollinators, dispersers, root and leaf mutualists, and the natural enemies of these organisms that contributes the most to total biological diversity. Interspecific hybridization of plants contributes to biological diversity in at least four fundamental ways. First, new species or subspecies of plants are formed from hybridization via several mechanisms. Second, hybrid-
Plant hybridization leads to persistent hybrid populations that exist in a variety of hybrid zones. Third, introgression could lead to the loss of diversity. Fourth, biological diversity is increased by plant hybridization and introgression through its influence on the abundance and diversity of herbivores and their natural enemies, and their pathogens. This article will consider the formation of plant hybrids, the structure of hybrid zones, the effects of introgression on genetic diversity, and the distribution of hybrid plants in nature. However, the primary focus of the article will be on the effects of plant hybrids on the biological diversity and community structure of higher trophic levels, herbivores, their natural enemies, and pathogens. The conceptual model proposed in this article suggests that plant hybridization contributes to genetic diversity of plant populations, that genetic diversity increases biological diversity, and that biological diversity creates new opportunities for interaction diversity (Fig. 1). The proposed cascade of effects flowing from plant hybridization is discussed in the following sections.

I. PLANT HYBRIDIZATION

A. Occurrence of Hybridization in Plants

Interspecific hybridization is a common and important evolutionary mechanism in plants. Allopolyploid or homoploid (diploid) speciation gives rise to reticulate phylogenies. Estimates of the proportion of angiosperm species derived from interspecific hybridization via polyploidy, of which allopolyploidy far predominates, are between 47 and 52% for angiosperms, 43% of ferns, and up to 99% of all pteridophytes (Grant, 1981). Species formed in this way occupy new habitats and can have expanded geographical ranges compared to the parental species. For example, Tragopogon mirus and T. miscellus, tetraploid species, have expanded ranges compared to their diploid parents.

Hybrid species may originate multiple times. Among polyploid hybrid species investigated using molecular means, the minimum number of recurrent origins of hybrid species ranges from 2 to 13. Thus, multiple origins appear to be the rule rather than the exception in polyploid species. This generates substantial genetic diversity, thereby providing greater opportunity for subsequent evolutionary change. Genetic mechanisms of gene silencing (sometimes leading to diploidization of polyploid genomes), gene diversification (divergence of duplicate genes), and genome diversification (repatterning of chromosomes) operate in polyploid species, but knowledge of how widespread these mechanisms are in polyploid-derived species is unknown.

The proportion of families and genera where interspecific hybrids have been documented has been estimated for the floras of the British Isles, Scandinavia, the Great Plains, the Intermountain West of North America, and Hawaii. The proportions of families and genera with hybrids ranged from 16 to 34% and from 6 to 16%, respectively, and the numbers of hybrids identified in these floras ranged from 134 to 642. These values indicate that hybridization is common and naturally occurring hybrid zones are numerous.

Within these floras, certain families and genera have disproportionate numbers of hybrids. The families Asteraceae, Cyperaceae, Onagraceae, Poaceae, Rosaceae, Salicaceae, and Scrophulariaceae and the genera Bidens, Carex, Cystandra, Euphrasia, Rosa, and Salix, among others, have large numbers of hybrids between various species relative to their representation in the flora. Traits that appear to promote the occurrence of hybrids in families and genera are perennality, outcrossing, and vegetative or clonal reproduction. Outcrossing increases the likelihood of interspecific crossing and the other traits ensure that hybrids persist long enough to reproduce.

B. Hybrid Complexes

Groups of species within a region that are connected by degrees of hybridization are called syngameons. Syngameons have been identified in Betula, Geum, Iris,

Juniperus, Melandrium, Nothophagus, Pinus, Quercus, and Salix. The extent of hybridization between pairs of species in syngameons varies, with some species having abundant gene flow and other pairs having limited gene flow (Grant, 1981). The web of hybridization provides the opportunity for species that do not hybridize directly to share genes via hybridization with a third species. A syngameon between irises along the Pacific coast of the United States is illustrated in Fig. 2.

C. Types of Hybrids

Hybrids are formed when gametes from two species, subspecies, or races combine to form F₁ plants. F₁s, if they are sterile, may be the only hybrid genotypes found in some hybrid zones (e.g., the sedges *Carex canescens* × *C. mackenziei* in Scandinavia). They may persist by establishment of clones or by recurrent formation. In other populations where hybrids are common, formation of F₁s may be a rare event. In Louisiana irises, formation of F₁s between *I. fulva* and *I. hexagona* is rare (none are found in nature, though F₁s can be readily made artificially) (Arnold, 1997), but the rare F₁s serve to produce backcross hybrids that have high fitness and are common in some habitats. Often, F₁ hybrids are intermediate in morphological characteristics to parent species.

F₁ hybrids have greater phenotypic variability than do F₂ hybrids due to recombination that occurs in the formation of gametes. Combinations of parental traits present in F₁ hybrids are dissociated in formation of F₂ hybrids. The resulting phenotypic and genotypic variation spans the range between each parent's phenotype, with many F₂s resembling intermediate forms. This variation influences the fitness of these hybrids and the responses of herbivores and pathogens.

Progeny formed from crosses between F₁s and a parent species (P) are known as first backcrosses (BC₁) (crosses of BC₁/P give rise to BC₂, etc.). Backcrosses may be formed between hybrids and one or both of the parental species, resulting in unidirectional or bidirectional backcrossing, respectively. Unidirectional backcrossing is known in hybridization of *Populus fremontii* and *P. angustifolia*, with *P. angustifolia* being the recurrent parent (the parent species to which the F₁ crosses). Preferential gene flow occurs from *Iris fulva* to *I. brevicaulis* through backcrossing, although backcrosses in both directions occur. In the willows *Salix sericea* and *S. eriocephala*, backcrosses occur to both parents. Backcross hybrids contain more of the recurrent parent's genetic material and thus tend to resemble that parent. Recombination creates substantial phenotypic and genetic variation among backcross progeny.

Hybrid swarms occur when extensive hybridization occurs within and among hybrid genotypes and parent species, creating a broad range of hybrid genotypes and phenotypes. Location of a putative hybrid within the range of variation between parents is usually made using a hybrid index based on plant morphology, and often this shows continuous variation in plant traits between parental species. Hybrid index methods based on morphology assume that parental traits are independent of each other (there is linkage equilibrium and lack of pleiotropy) and that traits are additively inherited. However, traits are not always independent and dominance of parental traits and transgressive morphologies make this method questionable. Molecular methods provide many more markers that are independent for creating hybrid indices that are useful for ecological studies (Rieseberg and Ellstrand, 1993).

Repeated backcrossing of hybrids with a parental species can permanently transfer genes from one species into the genome of another species, a process called introgression. Introgression may be localized or dispersed (Rieseberg and Brunsfeld, 1992). In localized introgression, gene transfer between species is restricted to the area of hybridization between the two species (e.g., the thistles *Carduus nutans* × *C. acanthoides*). In dispersed introgression, gene flow occurs beyond the
range of hybridization. Evidence of dispersed introgression has been found for Pinus banksiana × P. contorta and Iris fulva × I. hexagona. However, localized introgression appears to be more common than dispersed introgression (Rieseberg and Brunsfeld, 1992). One outcome of introgression is the formation of new species or races. Molecular evidence has revealed that several plant species and races have originated from stabilized introgressants. The Fort Davis race of Salix taxifolia has cpDNA of S. interior and nuclear markers of S. taxifolia, S. interior, and as many as three other willow species. Rieseberg and Brunsfeld (1992) document numerous other cases of introgression.

The transfer of genes between plant species can generate substantial genetic variation that can be acted on by natural selection. In the evolutionary novelty model (Arnold, 1997), hybridization coupled with selection is hypothesized to favor transfer of adaptive traits with two possible outcomes. The first is the formation of new species via diploid speciation, which incorporates novel adaptations derived from both parental species. Transfer of similar chromosome segments between species has been demonstrated in three different pedigrees of hybrids between the sunflowers Helianthus annuus and H. petiolaris, and the resulting chromosome structures resemble the diploid hybrid species, H. anomalous. This strongly suggested that selection governed the formation of H. anomalous and that coadapted parental gene combinations were maintained by natural selection. The second outcome is adaptive transfer of traits between species, which increases genetic diversity. Thus, hybridization coupled with selection by abiotic and biotic factors is a creative evolutionary process that generates genetic diversity in plant populations.

II. PLANT HYBRID ZONES

A. Types of Hybrid Zones

Hybrid zones are locations where hybrids between species, subspecies, or races are found. Typically, hybrid zones are described as clines, spatial gradients in traits or alleles across a geographic transect where two taxa meet. The balance between selection on hybrids and dispersal of genes determines cline width. Clines are expected to be narrow if there is strong selection against hybrids, if gene flow is limited, or if there are steep environmental gradients. Clines will be wider if selection is weak, gene flow is more extensive, or environmental gradients are gradual. Cline shape is predicted to be a smooth, sigmoid curve if selection acts on single genes or on quantitative traits, but linkage disequilibrium combined with dispersal of genes can distort the smooth shape of the cline, creating a stepped cline. In stepped clines most of the change in allele frequency or trait expression occurs in a narrow range in the middle of a cline. Allele frequency, linkage disequilibrium with other alleles or traits, and gene flow can be used to infer the form by which selection acts in these clines. A clinal hybrid zone is one type of spatial pattern of plant genetic diversity that can influence biodiversity and spatial distribution of animal and pathogen species.

Narrow hybrid zones 10 m wide are found between the sedges Carex canescens and C. machaerophila at the edge of water along the Bothnian coast in northern Sweden. Oak hybrid zones approximately 30 m wide between the oaks Quercus depressipes and Q. rugosa are found on steep slopes in northern Mexico. A much wider hybrid zone extending about 20 km occurs in the same area between Q. coccolobifolia and Q. rugosa. In contrast to the traditional clinal model of hybrid zones, mosaic hybrid zones occur when habitats of the hybridizing taxa are patchily distributed. Hybridization may occur where the different habitat patches abut, leading to the patchy distribution of hybrids. In contrast to the more or less discrete location of hybrid zones geographically in the clinal hybrid zones, mosaic hybrid zones are as widely distributed as the distribution of habitats and parental species. Therefore, the impact of mosaic hybrid zones on the distribution of plant genetic variation and its effects on diversity of communities of herbivores and pathogens are geographically more widespread.

Louisiana irises Iris fulva, I. brevicaulis, and I. hexagona fit the mosaic model since species and hybrid genotypes are associated with different, interspersed habitats. For example, I. fulva is associated with maple-dominated forest habitats, I. brevicaulis is associated with black oak forests, and I. hexagona is found at the edge of freshwater marshes. Hybrid genotypes are either not strongly associated with specific habitats or occupy intermediate or novel habitats. More dispersed mosaic hybrid zones occur where broadly sympatric species hybridize, either occasionally or extensively, where their habitats mix. The sunflowers Helianthus annuus and H. petiolaris exemplify this type of hybrid zone. These sunflower species have overlapping ranges and form local hybrid swarms in the western United States. Hybrids between Salix sericea and S. eriocephala also fit the mosaic model, with hybrids being found throughout the sympatric range of these species in eastern North America.
B. Models of Hybrid Zone Dynamics

Clinal hybrid zones have been of particular interest to evolutionary biologists and a number of models have been proposed to explain them. These models differ in the causes of selection on hybrids and how it varies across the cline. The hybrid disadvantage or tension zone model proposes that a balance exists between gene flow across a cline and endogenous selection against hybrids due to genetic incompatibilities in hybrid individuals. The environmental gradient model proposes that a balance of gene flow exists across a gradient of environmental conditions (exogenous selection), favoring different alleles (from the two species) at opposite ends of the cline. The hybrid advantage or bounded hybrid superiority model predicts that within a narrow ecotone separating the parental species, hybrids have higher fitness. The advancing wave model proposes that selection favors one parental species over the other, so that hybrids have higher fitness than one parent but lower fitness than the other parent species. As the name implies, the zone of hybridization will shift under this model until the inferior species is eliminated or endogenous or exogenous selection results in a balance suggested by one of the previous models. Finally, the neutral diffusion model suggests that neutral alleles (alleles with equal fitness effects) mix after secondary contact between parental species but that endogenous or exogenous selection does not act on these alleles in the cline.

C. Fitness in Hybrid Zones

Analysis of fitness of hybrids across parental and hybrid zones can reveal which of the models of hybrid zone dynamics is operating. The traditional view, derived mostly from studies of animal hybrids, is that hybrids are unfit compared to parental species. The presence of some degree of unfitness of naturally occurring or artificially created hybrids and the presence of concordant and concordant clines of traits are used to conclude that endogenous or exogenous selection acts against hybrids (Arnold, 1997). Lower viability or fertility of some hybrid genotypes is commonly found, suggesting endogenous selection against some hybrids. Higher susceptibility of hybrids to herbivores and pathogens or inferior ability to cope with environmental stress suggests exogenous selection against hybrids (Fritz et al., 1990).

Support for the hybrid superiority model was found for intersubspecific hybrids of sagebrush, Artemisia tridentata spp. tridentata and A. t. spp. vaseyana. Reciprocal transplants of seeds and seedlings across the narrow environmental gradient that separates the subspecies in Utah showed that hybrids had higher total fitness in the "hybrid" habitat and that parental subspecies had highest fitness in the parental habitats. Likewise, analysis of fitness of hybrid and parental Iris partially supported this model (Arnold, 1997). Fitness of Ipomopsis aggregata was greater than that of I. tenuituba and their interspecific hybrids across a narrow hybrid zone in the Rocky Mountains of Colorado due to strong hummingbird pollinator preference for I. aggregata floral traits, supporting the advancing wave model.

A review of fitness of hybrids by Arnold (1997) showed that frequently hybrids do not have lower fitness than parent species, rather that hybrid fitness may be equal to or greater than that of their parents. The genotype of hybrids affects their fitness, but the conclusion that hybrids are uniformly unfit is not generally supported by empirical studies.

III. GENETIC BASIS OF HYBRID PLANT RESISTANCE

The influence of hybrid plants on biological diversity depends on the resistance of hybrids to phytophages. Resistance is the ability of a plant genotype to avoid attack or prevent the development of a specific herbivore or pathogen relative to other genotypes. Resistance is specific to herbivore or pathogen species and resistance to different herbivores or pathogens are separate traits that may or may not be correlated. Resistance determines the abundance of each species of phytophagous organism on hybrid plants; resistant plants have lower amounts of damage resulting from lower densities of herbivores, whereas susceptible plants experience higher amounts of damage resulting from higher densities of herbivores. Consequently, the resistance of hybrid plants to their phytophages contributes to the biodiversity, community structure, and interactions among phytophages.

A. Theory

Several hypotheses concerning the resistance of hybrid plants compared to parental species assume that resistance is a polygenic trait, influenced by many genes, and that quantitative genetic models best describe the patterns of herbivore and pathogen response to hybrid plants. Some models specify that hybrids are F1s, but other models do not specify the genetic composition
of the hybrids. Box 1 describes the hypotheses that would be supported by comparisons of F1 hybrids to one or both parents. Providing that tests of resistance are performed in a uniform environment, the outcome of these comparisons is used to infer the genetic basis of resistance in hybrid plants. The additive pattern suggests that resistance genes act in a dosage-dependent manner (Fig. 3A). Dominance of either susceptibility or resistance suggests dominant effects of genes from one parent species (Fig. 3B). Dominance of susceptibility implies that hybrids have similar attractant or performance traits as the equally susceptible parent. Dominance of resistance implies that hybrids have similar repellent or antibiosis traits as the resistant parent. Hybrid susceptibility (Fig. 3C) indicates that traits that condition plant resistance fall below a level that deters herbivores or pathogens, a response called hybrid breakdown. Hybrid resistance (Fig. 3D) occurs when hybrids are more resistant than either parent species, a response called heterosis. For this pattern, different resistance genes from each parent may be dominantly inherited and act together in their effects on herbivores.

Box 1

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>Abundance of an herbivore species on hybrids is intermediate between herbivore abundances on the two parental species.</td>
</tr>
<tr>
<td>Dominance</td>
<td>Abundance of an herbivore species on hybrids is equal to that of one parental species and is significantly different from the other parental species. If the abundance on hybrids is similar to that of the more susceptible, parent, the pattern is dominance of susceptibility. If the abundance on hybrids is similar to that of the more resistant parent, the pattern is dominance of resistance.</td>
</tr>
<tr>
<td>Susceptibility</td>
<td>Abundance of an herbivore species on hybrids is significantly higher than on each individual parental species.</td>
</tr>
<tr>
<td>Resistance</td>
<td>Abundance of an herbivore species on hybrids is significantly less than on each individual parental species.</td>
</tr>
<tr>
<td>Partial dominance</td>
<td>Abundance of an herbivore on hybrids is intermediate compared to both parents, but there is a significant deviation toward one of the parents.</td>
</tr>
<tr>
<td>No difference</td>
<td>Abundance of herbivores on hybrids and parents do not differ significantly.</td>
</tr>
</tbody>
</table>

Formation of F2 or backcross plants can have further effects on plant resistance. Backcrosses may be intermediate between the F1 and recurrent parent (Figs. 3A, 3C, and 3D). This would suggest that resistance traits act additively. Genetic recombination may lead to hybrid breakdown in these hybrid classes (higher susceptibility) compared to F1 and parents, as shown in backcross to Species A in Fig. 3B. Such patterns imply the breakdown of coadapted gene complexes (several resistance genes evolved to work together in a species). Breakup of coadapted gene complexes may be one cause of hybrid susceptibility in F1 or BC plants. Backcrosses could resemble the recurrent parent (Species B in Figs. 3C and 3D), suggesting recovery of resistance or susceptibility factors above a threshold level. As with backcrosses, F2 resistance needs to be compared to parent and F1 progeny to document hybrid breakdown or some other pattern.

Regression analyses are useful in examining the patterns of resistance where continuous variation in hybrid genotypes exists, such as in hybrid swarms. Measures of herbivore abundance are plotted against hybrid indices and the best fitting regression models are determined. Significant linear, quadratic, and cubic regressions support the additive, susceptibility and resistance, and dominance hypotheses, respectively.

When resistance is due to single genes that have a major effect, segregation of discrete resistance phenotypes may occur. Gene-for-gene theory of the disease resistance of plants was based on the recognition of discrete infection types in the flax–flax rust (Linum usitatissimum–Melampsora linartica) pathosystem, but discrete resistance phenotypes also frequently characterize plant parasites in natural plant–pathogen interactions. Mendelian hypotheses of simple inheritance can be tested when distinct phenotypic classes of resistance are present in controlled crosses. F1 progeny of a controlled
cross may be either all susceptible or all resistant, matching the “dominance to susceptible” and “dominance to resistance” hypotheses, respectively. Segregation for resistance to *Venturia populina* and of resistance to the rust *Melampsora medusae* has been reported in *F*<sub>1</sub> progenies of *Populus trichocarpa* × *P. deltoides*. Likewise, segregation of recessive resistance can occur in the *F*<sub>2</sub> progeny of two susceptible parents. Resistance of *Populus trichocarpa* × *P. deltoides* to stem canker caused by *Septoria musiva* segregates in *F*<sub>1</sub> progenies. *P. deltoides* is resistant, but its *F*<sub>1</sub> progeny are always susceptible. However, when two susceptible *F*<sub>1</sub> clones are crossed, resistant *F*<sub>2</sub> individuals are seen. Conversely, resistance to *Melampsora occidentalis* is dominant, so those interspecific hybrids *F*<sub>1</sub>s are all resistant. Yet, *F*<sub>2</sub>s from crosses between these resistant *F*<sub>1</sub> progeny can produce some susceptible progeny (Fritz et al., 1999). If intermediate resistance phenotypes can be identified, then crosses of *F*<sub>2</sub>s and analysis of segregation ratios can be used to determine the number of genes responsible for resistance. Alternatively, correlation of resistance phenotypes with molecular markers can be used to identify the numbers and importance of resistance genes.

**B. Evidence**

1. **Resistance Traits**

Inheritance of resistance traits in hybrids will determine the effects of hybridization on susceptibility to herbivores and pathogens. Chemical traits of hybrids are usually found in one or both parental species. Concentrations of phenolic glycosides (derived from *Salix sericea*) and condensed tannins (predominant in *S. eriocephala*) were found in intermediate concentrations in hybrids between these species (Fig. 4). Some chemical traits in hybrids are inherited as dominant traits, but other resistance mechanisms are recessive in hybrids. Pod abscission, a defensive trait of palo verde (*Cercidium microphyllum*) appears to be recessive in hybrids.

with *C. floridum*, whereas seed coats of *C. floridum* are chemically defended against bruchid weevil attack and this trait is also recessive in hybrids. However, chemical traits of parents may be missing in hybrids, and hybrid plants may produce novel chemical traits (chemicals found in neither parent species nor parental chemical traits expressed in novel plant tissues). Thus, hybrids and hybrid-derived species can have unique or intermediate resistance characteristics (Rieseberg and Ellstrand, 1993).

**TABLE I** Numbers (N) and Percentages (P) of Tests That Support Hybrid Resistance Hypotheses from Plant–Herbivore and Plant–Pathogen Studies

<table>
<thead>
<tr>
<th>Plants/herbivores</th>
<th>Field census</th>
<th>Common garden</th>
<th>Laboratory</th>
<th>Plants/fungi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis</td>
<td>N (P %)</td>
<td>N (P %)</td>
<td>N (P %)</td>
<td>N (P %)</td>
</tr>
<tr>
<td>No difference</td>
<td>17 (30.4)</td>
<td>17 (34.7)</td>
<td>3 (14.3)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Additive</td>
<td>8 (14.3)</td>
<td>8 (16.3)</td>
<td>12 (57.1)</td>
<td>6 (20)</td>
</tr>
<tr>
<td>Dominance of susceptibility</td>
<td>7 (12.5)</td>
<td>16 (32.7)</td>
<td>4 (10.0)</td>
<td>10 (33.3)</td>
</tr>
<tr>
<td>Resistance</td>
<td>7 (12.3)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>4 (14.3)</td>
</tr>
<tr>
<td>Partial dominance</td>
<td>1 (1.8)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Susceptible</td>
<td>11 (19.6)</td>
<td>8 (16.3)</td>
<td>2 (9.3)</td>
<td>8 (26.7)</td>
</tr>
<tr>
<td>Resistant</td>
<td>5 (8.9)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>2 (6.7)</td>
</tr>
<tr>
<td>Total</td>
<td>36 (69)</td>
<td>40 (80)</td>
<td>21 (57)</td>
<td>30 (50)</td>
</tr>
</tbody>
</table>

*From Fritz et al., 1999*
garden studies of herbivores, but in laboratory tests, only 14.3% of the cases fit this pattern. The additive pattern was found in 14.3 and 16.3% of the field and common garden studies, respectively, but a much larger proportion of the laboratory studies (57%) found this pattern. The next most common pattern was dominance of susceptibility, which accounted for 12.9, 32.7, and 19% of the three types of studies. This pattern indicates that an herbivore has equal abundance to the more susceptible parent species. Susceptibility, where the herbivore is more abundant than on either parent, accounted for 19.6, 16.3, and 9.5% of the three types of studies. Resistance of dominance was equal to dominance of susceptibility in frequency in the field studies (12.9%), and the resistance pattern occurred in 8.9% of the field studies. Cases of resistance and dominance of resistance did not occur in the common garden and laboratory studies.

The predominant patterns found for plant–pathogen studies of hybrids are dominance of susceptibility, susceptibility, and additive. Dominance of resistance and hybrid resistance seem to be more common in plant–pathogen systems, compared to common garden and laboratory studies of herbivores.

Studies performed in controlled environments point to the fundamental influence of genetics on hybrid plant resistance. No difference, additive, dominance of susceptibility, and hybrid susceptibility to insects are common patterns seen in these studies. Among these common patterns, only hybrid susceptibility would suggest that hybrid plants would have lower fitness compared to both parents, assuming the herbivores lower fitness of their host plants. The absence of dominance of resistance or hybrid resistance patterns indicates that resistance to herbivores is not inherited as a dominant trait in most systems. In contrast, plant–pathogen systems demonstrate several cases of hybrid resistance or dominance of resistance (20% of all studies).

Hybrid susceptibility is illustrated by dramatic increases in herbivore abundance. The bud gall mite, *Aceria paraparapulsi*, was over 800 times more abundant on F1-type hybrids of *Populus fremontii* and *P. angustifo\fa005lua* than on either parent species. *Pemphigus betae*, a gall-inducing aphid of *Populus*, is found at much higher abundances (8- to 119-fold) in six poplar hybrid zones in river drainages throughout western North America. Severity of infection by the smut pathogen *Anthracnosea fischeri* was 30- to 80-fold higher on F1 *Carex canescens* × *C. macch
diennis* hybrids than on pure species. Usually, differences in herbivore abundance on hybrids and parents are much less pronounced. Hybrid resistance effects on upper surface leaf and blotch miners in two oak hybrid zones (*Quercus depressipes* × *Q. rugosa*, *Q. coccolobifolia* × *Q. emoryi*) resulted in 3–8 times as many miners on parents compared to the hybrid oaks, and less than twofold higher populations of *Phyllonycter salcifoliella* occur on willow hybrids.

Even for the same plant hybrid system, several different hypotheses are supported. On the *Salix sericea* × *S. eriocephala* hybrid system, five different hypotheses were supported by different herbivore species. Figure 3 shows the different responses of three herbivores and a leaf pathogen on *Salix sericea*, *S. eriocephala*, and F1 hybrid willows. Moreover, a guild of leaf gall inducers, three in the same genus, supported four different hypotheses in this system in a single year. This highlights the point that herbivore or pathogen resistance traits are independent traits of the plant that require separate consideration in determining patterns of plant resistance to particular herbivores.

3. Genetic and Environmental Effects on Plant Resistance

Studies of plants in the field combine the effects of plant genotype with the range of environmental factors experienced by the plant. Together, genetic and environmental factors can affect plant resistance. The role of environmental variation in determining plant resistance is important since some hybrid zones occur across steep environmental gradients, hybrids may occur at the limits of ranges of each parent species, and some hybrids are found in unique habitats where abiotic conditions differ. Plant stress, induced by environmental extremes, is known to make plants more susceptible to herbivores; thus the stress hypothesis proposes that elevated levels of herbivory on hybrid plants are a direct result of stress rather than due to direct genetic effects on plant resistance traits (Whitham et al., 1994; Fritz, 1999). Mortality of pines, regardless of whether they were hybrid or parental, in the hybrid and overlap zones between *Pinus edulis* and *P. californianus* was greater than in the allopatric parental populations, suggesting stress as a factor. F1 hybrids of *Salix sericea* and *S. eriocephala* are less tolerant to drought than are the pure species. The response of hybrid plants to stress is likely a consequence of their hybridity, and the consequent effects on resistance to herbivores or pathogens is therefore a genotype-by-environment interaction.

4. Generalist and Specialist Herbivores

If specialists are defined as herbivores that use only one parent species in a hybrid zone and generalists are defined as using both parent species, then these categories of herbivores are predicted to vary in how they respond to hybrid plants. Hybrid plants, which have some of the traits of the nonhost, should be less
FIGURE 5 Abundances of three herbivores and infection score of a pathogen on 1-year-old seedling willows of *S. sericea*, *S. eriocephala*, and their *F₁* hybrids growing in a common garden: (A) *Phyllocnistis* sp.; (B) leaf folder–*V*; (C) *Phyllocolpa terminalis*; (D) *Melampsora epitea*. (A–C from Fritz et al., 1998; D from Roche, B. M., and Fritz, R. S. (1998). Effects of host plant hybridization on resistance to willow leaf rust caused by *Melampsora sp.* Eur. J. For. Pathol. 28, 259–270.)

suitable for specialists, whereas hybrids should be more suitable for generalists, which are already adapted to traits of both species. Fitting these predictions, 92% of generalists (11 of 12 taxa), but only 50% of specialists (14 of 28 taxa), were more abundant on hybrid than parental taxa in the *Eucalyptus amygdalina × E. risdonii* hybrid zone. However, there is less consistency among three other generalist species in other systems: dominance of susceptibility was found for a bagworm on sagebrush, dominance of resistance was found for the Japanese beetle on willow, and an additive pattern was found for the spruce budworm on spruce.

Specialist herbivores are predicted to have higher abundance on backcross hybrids that are genetically more similar to their host species when backcrosses are present in populations. Among the 28 specialists of the *E. amygdalina × E. risdonii* hybrid zone, 68% (19) were most abundant on the class of hybrids most similar to their host species, 28% (8) were equally abundant on all hybrid classes, and 4% (1) was most abundant on the least similar hybrid category. A gall-forming cynipid, *Andricus californicus*, occurred with equal abundance on parental (*Quercus dumosa*) and all hybrid genotypes with *Q. engelmannii*, but not on pure *Q. engelmannii*, contrary to the prediction.

5. Conclusions

Several conclusions are apparent from these studies. First, resistance traits are inherited in several ways in hybrid plants. Chemical traits may be dominant, additive, or resistant. Moreover, novel plant chemicals or expression of chemicals in new plant tissues can occur in hybrid plants. Chemical traits may be dominant, additive, or resistant. Moreover, novel plant chemicals or expression of chemicals in new plant tissues can occur in hybrid plants. Second, there is not a single predominant effect of hybridization on plant resistance to herbivores and pathogens. Diversity in the resistance mechanisms of plants to different herbivores or pathogens exists, with dominance of susceptibility, susceptibility, additive, and the no difference being the patterns most frequently reported. Third, genetic effects of hybridization on resistance are well documented and widespread, but the role of environmental variation in modifying the genetic effects is poorly understood. Finally, specialist and generalist insects vary in predictable ways in their abundance on hybrid plants.
IV. COMMUNITIES ON PLANT HYBRIDS

A. Diversity of Herbivore Species

Species richness on hybrids will be greater than on either parent under two conditions that should commonly apply: parent species have species-specific herbivores and hybrids are not completely resistant to these herbivores. Given these conditions, hybrids will have herbivore communities derived from both parent species.

The effects of hybridization on species richness will be small if parental species share a large proportion of herbivores. This is the case for hybrids between *S. sericea* and *S. eriocephala*. Most of the gall formers, leaf miners, and leaf tiers, occur on both plant species and on hybrids. Several herbivores occur on one species and hybrids but not on the other parent. For example, *Pontana gracilis* occurs on *Salix sericea* and hybrids and *P. pomum* occurs on *S. eriocephala* and hybrids. Likewise, the stem gallers *Rabdophaga rigidae* and *R. salicisbrassicoides* occur on *S. sericea*, but *R. strobiloides* occurs on *S. eriocephala*. These herbivores are responsible for the higher diversity on hybrid willows compared to the two parents.

Extensive surveys of herbivore diversity on Eucalyptus hybrid systems have included comparisons of F1-like hybrids and putative backcrosses to each parent species. Species richness on F1 hybrids and backcrosses to *E. amygdalina* were significantly greater than on either parent or backcrosses to *E. risdonii* (Fig. 6A). Additional species numbered 7–8 more than on *E. amygdalina* and 11–12 more than *E. risdonii* in their pure zones, respectively. Backcrosses to *E. risdonii* had higher diversity than on pure *E. risdonii* in the hybrid zone or in the pure zone. Parent species in pure zones had even lower diversity than parent species in the hybrid zone. Analysis of another Eucalyptus hybrid zone found greatest species richness on putative backcross hybrids than on either parent or F1-type hybrids (Fig. 6B). In both of these examples, relative abundances of herbivores were higher on hybrids than on parent species. This suggests that hybrid breakdown may create a more permissive environment for herbivores on backcross plants.

A significant number of species occurring on hybrids may be unique in that they are not found on either parent species. On spruce hybrids in two locations in Michigan, 12.2 and 15% of herbivore species were unique to hybrids. About half of species are restricted to one or the other parent species and 35 and 41.5% of herbivores on hybrids were found on both parent species (Fig. 7).

**FIGURE 6** Species richness in a *Eucalyptus amygdalina* × *E. risdonii* hybrid zone (40 insect and fungal taxa species) (A) and in a *Eucalyptus obliqua* × *E. leucoxylon* hybrid zone (25 insect and fungal taxa species) (B). Bars with different letters differ significantly at P < 0.05. (A from Whitham et al. (1994); B from Morrow, P. A., Whitham, T. G., Potts, R. M., Ladiges, P., Aichele, D. H., and Williams, J. B. (1994). Gall-forming insects concentrate on hybrid phenotypes of Eucalyptus. In The Ecology and Evolution of Gall-Forming Insects (P. Price, W. Matson, and Y. Baranchikov, Eds.), pp. 121–134. North Central Forest Experiment Station, Forest Service, USDA, St. Paul, MN.)

B. Community Structure

1. Shape of Communities

The combination of altered resistance of hybrid plants compared to parents, species-specific responses of herbivores to hybrid plants, and new combinations of coexisting phytophages on hybrid plants dictates that community shape (the relative representation of different species) will differ between hybrids and parent species.
and among genetic classes of hybrids. The ecological consequences of the novel communities of phytophages will depend on species abundances and how species interact with each other. However, novel interactions (competition, mutualism, predation, and parasitism) are likely to result.

Community shape differs on *S. sericea*, *S. eriocephala*, and their hybrids. A discriminant function analysis suggests that the relative abundances of herbivore species varies among parents and hybrid field plants and among full-sib families derived from crosses between pure female *S. sericea* and male *S. eriocephala* (Figs. 8A and 8B). However, herbivore communities are not completely distinct between these taxa. For both analyses, some plants from each taxon are closely spaced, indicating that communities do not differ between some plants. A more distinct pattern of community structure was seen between *Populus fremontii*, F1 hybrids, and backcross hybrids (Fig. 8C). But even for this example, some individual plants fall within the space occupied by other taxa. The use of herbivore species to categorize hybrid and parental genotypes provides similar levels of discrimination of hybrids as the use of plant morphology.

2. Keystone Species
Species that have large effects on populations of other species in a community are called keystone species. Removing them from the community and measuring the responses of other species can reveal their indirect effects on community structure. Two such cases of indirect effects of keystone species on diversity and relative abundance of other species on hybrids were found for the gall aphid, *Pemphigus betae*, and the leaf beetle, *Chrysomela confluens*, in the *Populus angustifolia*–*P. fremontii* hybrid zone (Whitham et al., 1999). Removal of *P. betae* from hybrid trees reduced species richness by 32% and relative abundance by 53% of other herbivores in the system. The negative effects on diversity were due to elimination of the parasites and predators of the aphids and reduction of other herbivores, which benefit from aphid modification of the host plant. In contrast, removal of the beetle, which preferentially feeds on leaves of young plants, increased species richness of other herbivores by 120% and relative abundance by 75%. Defoliation by the beetle reduces the available foliage for several other herbivores, an intense competitive effect. Thus, when plant hybridization strongly alters resistance to even one herbivore, there may be a cascade of indirect effects on other species.

3. Third Trophic Level
The third trophic level is the group of parasitoids, predators, parasites, and pathogens of herbivores or plant pathogens. These species typically outnumber herbivores severalfold, and therefore their diversity is an important part of the biological diversity supported by hybrid plants. Few studies have considered the effects of parasitoids or predators on herbivores occurring on hybrid and parental plants, but variation in abundance and fitness of herbivores on hybrid plants may be explained in part by the impact of the third trophic level. Decreased levels of parasitism or predation could explain the higher abundance of some herbivores on hybrid plants compared to parental species. Higher parasitism of herbivores on hybrids may reduce the density and fitness impact of herbivores on hybrids. Parasitism by chalcid parasitoids did not differ among galls of *Andricus californicus* on *Quercus dumosa* and a wide
FIGURE 8. Community shape based on densities of herbivore and pathogen species on parents and hybrids. Plots of the first two canonical variables from discriminant function analyses using densities of 12 gall-inducing, leaf mining, and leaf-tying herbivores of *S. sericea*, *S. eriocephala*, and their hybrids. (A, B) *S. sericea*, *S. eriocephala*, and clones in the field (A), and F1 hybrids and parents in a common garden (B). (C) *Populus fremontii*, and hybrids with *P. angustifolia* in the field. (From Floate, K. D., and Whitham, T. G. (1995). Insects as traits in plant systematics: Their use in discriminating between hybrid cottonwoods. *Can. J. Bot.* 73, 1–13.)

A range of hybrid genotypes. Mortality of aphids on hybrids of *Alnus incana* and *A. glutinosa* was intermediate between that of the two parents (Table II). However, parasitism rates of a seed-eating herbivore of two cattails (*Typha latifolia* and *T. angustifolia*) and their F1 hybrid were significantly greater on the hybrid, although densities of the seed predator were significantly lower on the hybrids. In this case, hybridization seems to directly affect herbivore susceptibility to parasites. Parasitism of the leaf miner *Phyllonorycter* sp. on oak hybrids was lower than on parental species, and lower parasitism of *Phyllonorycter salicifoliella* on willow hybrids was found in the field, but was not repeated the next year or on F1 hybrid plants in a common garden experiment (Table II) (Fritz et al., 1999). Predation on beetle larvae on willow hybrids was intermediate between that on the parental species. There is a wide range of effects of parasitoids and predators on herbivores of hybrid and parental plant species. The outcome of these interactions is likely to influence the population dynamics of herbivores of hybrids and therefore their effects on hybrid plant fitness. One consequence of unique communities on hybrid plants is that related herbivores that share natural ene-
Summary of the Effects of the Third Trophic Level on Herbivores on Hybrid and Parental Species

<table>
<thead>
<tr>
<th>Taxa Herbivore</th>
<th>Enemy</th>
<th>Mortality relative to parent(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quercus dumosa × Q. engelmannii</td>
<td>Andricus californicus</td>
<td>Chalcid</td>
</tr>
<tr>
<td>Quercus depanneux × Q. rugosa</td>
<td>Phyllonorycter sp.</td>
<td></td>
</tr>
<tr>
<td>Salix sericea × S. eriocephala</td>
<td>Phyllonorycter salicifoliella</td>
<td>Parasitoids</td>
</tr>
<tr>
<td>Salix viminalis × S. dasyclados</td>
<td>Galerucella lineola</td>
<td>Various predators</td>
</tr>
<tr>
<td>Alnus incana × A. glutinosa</td>
<td>Phyllonorycter sp.</td>
<td>Predators and parasitoids</td>
</tr>
<tr>
<td>Typha latifolia × T. angustifolia</td>
<td>Lymnaecia phragmitella</td>
<td>Four parasitoids</td>
</tr>
</tbody>
</table>

**4. Other Indirect Effects**

The interactions of plants with other species extend beyond herbivores and their natural enemies. Hybridization often results in novel plant architecture. F1-type hybrids in the cottonwood hybrid zone between *P. fremontii* and *P. angustifolia* have greater numbers of lateral branches than parent species. Compared to parent species and backcrosses to *P. angustifolia*, these hybrids supported significantly greater numbers of bird nests. Northern orioles (*Icterus galbula*) and black-billed magpies (*Pica pica*) were the most common species using hybrid cottonwoods as nest sites, but other common species using these trees were the American robin (*Turdus migratorius*) and the warbling vireo (*Vireo gilvus*) (Whitham et al., 1999).

Plants have mutualist or commensalistic interactions with endophytic fungi, which may have detrimental effects on insect and mammalian herbivores. In the *Quercus grisea* and *Q. gambelii* hybrid zone, total endophyte frequency and frequency of *Gnomonia cerastis* was highest on *Q. gambelii* and was positively associated with hosts resembling *Q. gambelii*, but the frequency of *Coccochorella quercifolia* was most frequent on *Q. grisea* and on hybrids resembling *Q. grisea*. While direct effects on the density of a leaf mining moth (*Phyllonorycter sp.*) were not found, mortality of the tissue-feeding stage of *Phyllonorycter sp.* was positively associated with the presence of *G. cerastis*. This study suggests the possible interaction between endophytes and an insect herbivore mediated by hybrid genotype.

Mycorrhizal fungi are ubiquitous mutualists of plants that commonly facilitate nutrient uptake. How hybridization affects these mutualists is unknown, but the effects of high levels of herbivory that occur on some hybrid plants could be detrimental to mycorrhizae. Herbivory on pinyon pine, *Pinus edulis*, by a stem-boring moth reduced mycorrhizal levels by as much as 33%, but when moths were removed, mycorrhizae recovered. Thus, interactions between herbivory on hybrids and root mutualists may be antagonistic, resulting in more harmful effects on plant fitness.

**V. POPULATION AND EVOLUTIONARY CONSEQUENCES**

**A. Population Dynamics of Phytophages on Hybrid Plants**

The observation that most of the population of the gall-forming aphid (*P. betae*) was found on hybrid poplars rather than on pure parental species suggested the possibility that susceptible hybrids could limit the abundance of this parasite on parental poplars by attracting all of aphids (sink hypothesis). An alternative hypothesis is that hybrids are a reservoir for parasites that disperse and colonize parental species at higher rates than would occur in the absence of hybrids (source hypothe-
E. amygdalina was consistently lower than that of risdonii a third of the parents in response to the same phytophages. In contrast, fitness of hybrids between mularia sp. cause greater losses on these hybrids compared to parents. Insect hybrids. Studies of gens on hybrid plants can affect the fitness of hybrids, even death of plants. Effects of herbivores and pathogens affect the fitness of plants. High levels of herbivore or pathogen attack typically result in slower growth, diminished reproduction, or result in slower growth, diminished reproduction, or even death of plants. Effects of herbivores and pathogens on hybrid plants can affect the fitness of hybrids, the distribution of hybrid zones, and selection among hybrids. Studies of Eucalyptus hybrids have revealed two effects of phytophages on fitness of hybrids. Hybrids of E. melanophloia × E. crebra have lower fitness, measured by seed production, than either parent. Insect predispersal seed predators and the seed pathogen Ranularia sp. cause greater losses on these hybrids compared to parents. In contrast, fitness of hybrids between E. populnea × E. crebra was intermediate between that of the parents in response to the same phytophages. In a third Eucalyptus hybrid system, fitness of hybrids and E. amygdalina was consistently lower than that of E. risdonii. Fitness of the F1-type hybrids was lower than that of either parent in 1980, but in 1990 and 1992, hybrid fitness was similar to that of E. amygdalina. In all years, fitness was inversely related to herbivore species richness and abundance, suggesting a direct role of phytophages on fitness. A number of studies have found that absolute and relative abundances of herbivores and pathogens are higher on hybrid genotypes than on parental species (Whitham et al., 1994). If the higher load of herbivores causes more damage and more damage reduces plant fitness, then hybrid plants may be selected against. Two outcomes could result given this scenario. First, selection against hybrids could maintain narrow hybrid zones and favor reinforcement of species isolating mechanisms. Second, recombination of resistance traits can create genetic variation among hybrid progeny. Selection by herbivores or pathogens among the recombinant hybrids can favor certain hybrid genotypes over others, setting the stage for evolution of adaptive hybrid combinations and the formation of hybrid taxa.

2. Introgression of Plant Resistance

Backcrossing and selection can lead to introgression of plant resistance between plant species. Selection by grazers favored introgression of defensive traits from unpalatable Carex canescens and C. maritensis influenced by the presence of hybrids. Only in populations where hybrids were present was the smut fungus found on parent species; here hybrids clearly act as a source.

**B. Evolutionary Outcomes**

1. Effects of Phytophages on Plant Fitness

Herbivores and pathogens affect the fitness of plants. High levels of herbivore or pathogen attack typically result in slower growth, diminished reproduction, or even death of plants. Effects of herbivores and pathogens on hybrid plants can affect the fitness of hybrids, the distribution of hybrid zones, and selection among hybrids. Studies of Eucalyptus hybrids have revealed two effects of phytophages on fitness of hybrids. Hybrids of E. melanophloia × E. crebra have lower fitness, measured by seed production, than either parent. Insect predispersal seed predators and the seed pathogen Ranularia sp. cause greater losses on these hybrids compared to parents. In contrast, fitness of hybrids between E. populnea × E. crebra was intermediate between that of the parents in response to the same phytophages. In a third Eucalyptus hybrid system, fitness of hybrids and E. amygdalina was consistently lower than that of E. risdonii. Fitness of the F1-type hybrids was lower than that of either parent in 1980, but in 1990 and 1992, hybrid fitness was similar to that of E. amygdalina. In all years, fitness was inversely related to herbivore species richness and abundance, suggesting a direct role of phytophages on fitness. A number of studies have found that absolute and relative abundances of herbivores and pathogens are higher on hybrid genotypes than on parental species (Whitham et al., 1994). If the higher load of herbivores causes more damage and more damage reduces plant fitness, then hybrid plants may be selected against. Two outcomes could result given this scenario. First, selection against hybrids could maintain narrow hybrid zones and favor reinforcement of species isolating mechanisms. Second, recombination of resistance traits can create genetic variation among hybrid progeny. Selection by herbivores or pathogens among the recombinant hybrids can favor certain hybrid genotypes over others, setting the stage for evolution of adaptive hybrid combinations and the formation of hybrid taxa.

VI. CONSERVATION OF HYBRID PLANTS

The value of hybrid plants to biological diversity depends on the attributes of the hybridizing taxa and on the effects on dependent communities. Hybridization of widespread, common plant species can produce genetically diverse hybrids that can contribute to evolution of plant taxa and support unique herbivore communities with novel interactions. On the other hand, interspecific hybridization threatens plant taxa that are rare, have isolated populations, or are endemic to islands. Decisions about conserving plant hybrid populations depend on the effects that hybridization are likely to have on the persistence of the parental species and on their dependent communities. Taxonomic, genetic,
A. Threats to Diversity from Plant Hybridization

Hybridization can threaten rare, endemic, or isolated plant species. Genetic assimilation via hybridization can threaten small populations of species occurring on islands or in isolated, relict populations when larger populations of related taxa are present. Most offspring of these rare taxa may be hybrid individuals, rather than pure conspecifics. A further threat is that these hybrid individuals may suffer outbreeding depression, reduced fitness relative to nonhybrid individuals, leading to a loss of their alleles from the population. Finally, inasmuch as hybrid individuals may not be recognized taxonomically, protection afforded to the threatened parent may not include hybrid individuals that possess a significant portion of the remaining genetic diversity.

The rarest plant in California, Cercocarpus traskiae, is found in a single canyon on Santa Catalina Island. A more common species, C. betuloides, also occurs on the island and hybridizes with C. traskiae. Five pure C. traskiae and two hybrid trees were discovered, and all of the seedlings found in the area were F1 hybrids of just two of the pure C. traskiae individuals (Rieseberg, 1991). Thus, genetic assimilation and the loss of pure individuals threaten the remaining C. traskiae population. Rieseberg (1991) recommended that the C. betuloides individual near the remnant population of C. traskiae be destroyed, that isolated populations of C. traskiae be established in new locations, and that non-native mammals be eliminated. In other systems, hybrid species are difficult to identify, being confused with hybrid swarms or with putative species. Furthermore, hybrid species are sometimes found in novel and restricted habitats (e.g., Helianthus paradoxus) where they are threatened by habitat destruction.

Plants threatened with extinction by hybridization also have their dependent communities of herbivores and pathogens threatened. Hybridization can eliminate the only habitat of specialist herbivores or pathogens.

B. Benefits of Hybrid Plants to Biological and Interaction Diversity

Hybridization between widespread and common species commonly creates stable hybrid zones or hybrid swarms that do not threaten the population or genetic integrity of parent species. Rather, these hybrid populations are sites of substantial genetic diversity where plant evolution occurs and to which communities of herbivores, pathogens, mutualists, and commensals respond.

Hybrid populations are centers of biological diversity of dependent phytophages. Not only do more species of herbivores and pathogens coexist on hybrid plants but some have much higher densities on hybrids and may be largely restricted to hybrid zones. This is the case for Pemphigus betae and Chrysomela confluens on poplar hybrids, but other hybrid zones have similar patterns. While a higher density of herbivores is not the most common response to hybrid plants, higher species richness on hybrids and in hybrid zones may be ubiquitous. Independent herbivore response leads to unique communities on hybrid plants. These communities support greater interaction diversity and provide the environment for evolution of interactions between species. Competitive, parasitic, and mutualistic interactions in these unique species assemblages can influence the adaptations of herbivores to plants. Little experimental evidence exists for this assertion, but community patterns are suggestive.

The combinations of higher plant genetic diversity, higher herbivore and pathogen diversity, and more opportunities for interaction diversity strongly suggest that hybrid zones deserve conservation protection. The conservation status of hybrid plants is ambiguous at present. Since hybrids are not taxonomically distinct or may not even be recognized in nature when they occur, they usually do not receive consideration for protection. Because hybrid plants may be the major hosts of rare or threatened insect species, some hybrid populations may deserve special protection.

C. Hybridization of Crops and Native Plants

Interspecific hybridization can transfer traits from crop plants to their native relatives. Gene flow is known for radish (Raphanus), squash (Cucurbita), sorghum (Sorghum), sunflower (Helianthus), and many other crop-native plant species pairs. Concerns over gene transfer are that crop genes may irreparably contaminate smaller populations of native plant species, that crop-native plant hybrids may make more aggressive weeds that disrupt native plant communities, and that selected or genetically altered traits may disrupt interactions with communities of dependent herbivores and pathogens. For gene flow to be evolutionarily important, F1s produced between crops and native species must be par-
tially fertile so that later generation backcrosses can occur. However, even production of $F_1$s can result in outbreeding depression, perhaps leading to reduced populations of native species.

Transgenic crop plants have specific traits, such as insect resistance and herbicide tolerance, introduced using molecular genetic methods. When these crops are grown for agriculture, hybridization with wild conspecific or heterospecific relatives can lead to formation of hybrids possessing the genetically engineered trait. Subsequent backcrossing can carry the transgenic trait into populations of the wild relative. Transgenic traits that provide a fitness advantage may spread throughout the population. Evidence of $B. campestris$ ($\approx B. rapa$)-like plants with the transgenic herbicide-tolerance trait BASTA (glufosinate)-tolerant were produced experimentally and were found to occur in the field. Since these plants had high fertility, the rapid spread of herbicide tolerance is possible. The spread of transgenic herbicide tolerance or insect and pathogen resistance via hybridization could be a threat to biological diversity.

VII. CONCLUSIONS

Hybridization is a naturally occurring process that contributes to biological and genetic diversity of plant species. Hybridization affects the resistance of plants to herbivorous insects and plant pathogens, typically affecting resistance differently for different phytophages. Consequently, dependent communities of herbivores and pathogens are unique assemblages on hybrid hosts compared to parent species. This presents the opportunity for novel interspecific interactions among the phytophages, creating interaction diversity on hybrid plants. The novel community can affect the population dynamics and the evolution of hybrid plants. Conservation of naturally occurring hybrids and hybrid zones can maintain biological diversity of herbivore and pathogen species.

See Also the Following Articles

CONSERVATION EFFORTS, CONTEMPORARY • EX SITU, IN SITU CONSERVATION • LATENT EXTINCTIONS: THE LIVING DEAD • PLANT BIODIVERSITY, OVERVIEW • PLANT INVASIONS

Bibliography


I. Concepts and Terminology
II. The History of Alien Plant Invasions
III. The Current Extent of Alien Plant Invasions
IV. Invasions Processes
V. Invasiveness and Invasibility
VI. Modeling Plant Invasions
VII. Managing Plant Invasions

GLOSSARY

alien plants Plant taxa in a given area whose presence there is due to intentional or accidental introduction as a result of human activities. Synonyms include "exotic plants," "nonindigenous plants," and "nonnative plants."

environmental weeds Invasive alien plants that impact on natural or seminatural ecosystems, for example, by eliminating native organisms or altering ecosystem functioning (also known as "wildland weeds"). Native species can be environmental weeds, especially when the disturbance regime or resource levels have been altered.

invasibility The properties of a community or ecosystem that render it susceptible (or resistant) to invasion by alien plants.

invasive alien plants Naturalized plants that produce reproductive offspring, often in very large numbers, at considerable distances from parent plants, and thus have the potential to spread over a considerable area.

invasiveness The delimitation of features of an organism (e.g., life-history traits) that enable it to invade (i.e., to overcome various barriers to invasion).

naturalized plants Alien plants that reproduce consistently and sustain populations over many life cycles without the input of resources or direct intervention by humans; they often recruit offspring freely, but mainly very near adult plants, and do not necessarily invade natural ecosystems (cf. invasive alien plants).

TECHNOLOGICAL INNOVATIONS, driven by rapid increases in the global human population and requirements for diverse products and services from nonnative plants, have facilitated the large-scale (in terms of numbers of individuals, taxa, and foci of introduction) movement of plants to regions far beyond their natural dispersal range. This, together with widespread changes to disturbance regimes and the development of novel anthropogenic plant communities, has led to many plants (many, but not all, of which are "colonizers" in their natural habitats) becoming "weeds" (in agricultural systems, inside or outside their natural range) or "invaders" (in natural and seminatural systems outside their natural ranges). Plants carried to new habitats far from their natural ranges are very often introduced, intentionally or accidentally, as seeds and arrive without key natural enemies that limit their performance in their natural range. For many reasons, most introduced species do not spread from planting sites, and they
cause no damage in the receiving environment. Many alien plants provide us with food, fuel, and timber, beautify our cities, and serve many other useful purposes without invading. Those species that do become invasive are a major component of habitat transformation in many parts of the world. This chapter focuses primarily on the ecology of those species that have invaded natural and seminatural systems.

I. CONCEPTS AND TERMINOLOGY

Terms such as colonization (or colonizer) and invasion (or invader) are frequently applied with reference to plants that reclaim their distributions rapidly following disturbance, or that quickly occupy new sites adjoining their original range after alterations to factors that previously limited their ranges. Within any assemblage, species vary in their ability to persist under conditions of environmental change, and also with respect to their ability to invade new sites. Most species occupy a reasonably fixed position on a continuum from "colonizers" or "pioneers" to late-seral species. The former typically undergo large fluctuations in population size whereas the latter are specialized for more stable environments where rapid population growth is unimportant. Such strategies are associated with distinctive suites of life-history traits.

Episodes of range expansion and contraction mediated by natural disturbances, interactions with co-occurring organisms, and restricted by natural barriers have driven the diversification of floras and the delimitation of the world's phytogeographic zones. Humans have influenced the distribution of plants for millennia by, among other things, moving plants to new habitats (breaching geographic barriers) and altering opportunities for persistence, regeneration, and spread. Such actions have resulted in thousands of plant species spreading into areas which would, without human intervention, have been either unfavorable for the persistence of such species, or outside their normal dispersal range. There is some confusion in the literature regarding concepts relating to the different categories of "weedy" plants (Fig. 1).

The study of "weeds" (taxa in areas 1, 2, 4, and 5 in Fig. 1) has for centuries been directed at solving problems (how can weeds be killed or how can their effects on crops be reduced?). Colonization dynamics have been systematically explored by plant ecologists for about a century. Large-scale problems with the invasive spread of plants introduced to natural and seminatural ecosystems outside their natural ranges have become a recent phenomenon, and "invasion ecology" has only recently begun to emerge as a field of specialization in ecology. Milestones in the study of such invasions have been the publication, in 1938, of Charles Elton's book titled *The Ecology of Invasion by Animals and...*

Humans have influenced the geographic ranges of plants for millennia, by shuffling their distributions and altering their abundances, both within and outside their natural ranges. The extent to which such movements prior to the age of European colonialism shaped current distributions in Europe and Asia is often overlooked. The extent to which these range changes represent “invasions” can be debated since many changes were within or adjoining the natural ranges of such species. This is one reason for the blurred distinction between the categories “colonizers” and “invaders” (Fig. 1) in the ecological literature. Examples of human-induced changes in the distribution of pines (Pinus spp.; Pinaceae) serve to illustrate the different categories of range expansion. The distributions of species such as Pinus brutia, P. halepensis, P. pinaster, and P. pinea have changed dramatically due to human activities over many centuries. For example, P. brutia subsp. eldarica currently occurs in scattered locations from Azerbaijan to Pakistan. Genetic studies have shown that it occurs “naturally” only on a single mountain in Azerbaijan; the rest of its current range is due to humans who spread and managed this pine as a semidomesticated landrace. Most of its known locations follow ancient trade routes along the path taken by Alexander the Great (356–323 B.C.). In North America and other parts of the world that currently face the most severe problems with invasive alien plants. To cite but one instructive example: Plant introductions from Europe and Asia to South Africa between 1650 and 1800 reflected the need of Dutch colonists to cultivate a wide variety of agricultural and horticultural crops, mainly from their homeland in Europe and from the Dutch possessions in the East. Fifty or more crop plants were introduced within the first few years of European settlement, including many present-day weeds of agriculture. Subsequent waves of human immigrants, especially between 1800 and 1870, led to a surge in plant and animal introductions. Most of the plant species that now cause the greatest problems as invaders in fynbos and other South African biomes arrived between 1825 and 1860.

The magnitude and composition of the influx of alien plants to a given region during the colonial era was determined by many factors, including the geographical location of the area, its cultural links with colonial powers and other colonies, and special features of the region that influenced decisions on what types of plants to introduce. For example, the four regions of the world with a Mediterranean-type climate outside the Mediterranean Basin (California, Central Chile, southern and southwestern Australia, and the southwestern tip of South Africa) all received the bulk of their alien floras from the Mediterranean Basin. Smaller components came from other regions with which they had colonial ties, and even fewer from other regions. Current problems with plant invasions in these regions clearly reflect the dynamics of biological exchange during the colonial era.

The rapid globalization of European trade routes starting in the eighteenth century led to a dramatic increase in the magnitude and diversity of the transcontinental movement in plants (Fig. 2). Trends affecting plant invasions in the nineteenth and twentieth centuries are complex and region specific. There are important differences in patterns between Northern and Southern Hemispheres, between eastern and western parts of the Northern Hemisphere, and between what are now known as “industrialized” and “developing” countries. The numbers of alien plant species that are

II. THE HISTORY OF ALIEN PLANT INVASIONS

Humans moved many plant species beyond their natural ranges long before the dawn of the Age of Discovery and European colonialism. The frequency and pervasiveness of such early translocations were too limited to have resulted in widespread, damaging invasions. The voyages of exploration and the gradual settlement of the East and the New World by European powers (Britain, France, Germany, the Netherlands, Portugal, and Spain) heralded the start of a huge wave of translocations of plants and animals. This phase of “ecological imperialism” shaped the alien floras of many parts of the world that currently face the most severe problems with invasive alien plants. To cite but one instructive example: Plant introductions from Europe and Asia to South Africa between 1650 and 1800 reflected the need of Dutch colonists to cultivate a wide variety of agricultural and horticultural crops, mainly from their homeland in Europe and from the Dutch possessions in the East. Fifty or more crop plants were introduced within the first few years of European settlement, including many present-day weeds of agriculture. Subsequent waves of human immigrants, especially between 1800 and 1870, led to a surge in plant and animal introductions. Most of the plant species that now cause the greatest problems as invaders in fynbos and other South African biomes arrived between 1825 and 1860.

The magnitude and composition of the influx of alien plants to a given region during the colonial era was determined by many factors, including the geographical location of the area, its cultural links with colonial powers and other colonies, and special features of the region that influenced decisions on what types of plants to introduce. For example, the four regions of the world with a Mediterranean-type climate outside the Mediterranean Basin (California, Central Chile, southern and southwestern Australia, and the southwestern tip of South Africa) all received the bulk of their alien floras from the Mediterranean Basin. Smaller components came from other regions with which they had colonial ties, and even fewer from other regions. Current problems with plant invasions in these regions clearly reflect the dynamics of biological exchange during the colonial era.

The rapid globalization of European trade routes starting in the eighteenth century led to a dramatic increase in the magnitude and diversity of the transcontinental movement in plants (Fig. 2). Trends affecting plant invasions in the nineteenth and twentieth centuries are complex and region specific. There are important differences in patterns between Northern and Southern Hemispheres, between eastern and western parts of the Northern Hemisphere, and between what are now known as “industrialized” and “developing” countries. The numbers of alien plant species that are
The human-orchestrated movement of plants throughout the world often involved movement from the native area as well as further dispersal from initial foci of establishment. A good example of this pattern is the history of human-aided dispersal of the *Lantana camara* (Verbenaceae) complex, now one of the most widespread and damaging plant invaders in the tropics and subtropics. The pattern reflects the major importance of European trade routes in disseminating plants.

Although new arrivals originate from a wider source area than the early introductions, they tend to arrive via more direct routes than in the past.

Among the fundamental factors that have affected invasions globally, especially in the twentieth century, have been (a) the improvement of transportation systems, both at an intercontinental scale, with the emergence of aircraft transportation and massive growth in shipping, and within regions (roads, railways, internal navigation canals, etc.); (b) massive habitat transformations and changes to prevailing disturbance regimes, greatly increasing the extent of ruderal habitats; (c) the introduction of large numbers of livestock to temperate grasslands dominated by caespitose grasses that lack the ability to withstand grazing pressure; (d) widespread afforestation using alien tree species (the area of tree plantations—comprising mainly alien species of *Pinus, Eucalyptus, Tectona* and *Gmelina*—doubled between 1980 and 1995 to cover 180 million ha; and most developing countries with large forest estates plan to double their plantation area by 2010); and (e) the increasing demand for alien plants for use in horticulture (North American seed and nursery catalogs offer more 59,000 plant taxa for sale to national and international markets). Developments in the twentieth century (especially the last quarter) that are particularly pertinent to plant invasions have included the increasing attention given to scouring floras of previously underexplored areas for “new” useful plants, and the growth of agencies to facilitate the rapid transfer of germplasm of many “multipurpose” tree species for agroforestry and other purposes.

Most of the world’s worst invasive plants reached the stage where they cause major damage only in the past four decades. It is now well established that plant and animal invasions are a significant component of human-caused global change.

### III. The current extent of alien plant invasions

Invasive alien floras include representatives of all main growth forms of higher plants (annuals, perennials, aquatic plants, epiphytes, geophytes, grasses, herbs, shrubs, trees, etc.). Alien plant invasions have affected aquatic ecosystems (esmearine, freshwater, marine, wetlands) and most types of terrestrial ecosystems throughout the world, including arid systems, coastal systems, forests (boreal, temperate, tropical), grasslands, savannas and woodlands (including mangroves) and shrublands.
Plant invasions in natural and seminatural ecosystems involve the following fundamental phases: introduction to the region by humans, establishment, population growth (sometimes accompanied by genetic adjustment), spread to new areas within the region (and often also outside the region via further dispersal by humans; see Fig. 2), interaction with the local biota and disturbance regime, and displacement of native elements. There are many ways of conceptualizing the various processes involved in invasion and interactions with biotic and abiotic features of the new environment. One may depict the various potentially limiting factors as a series of ‘barriers.’ The simplest representation of such a model shows (a) a geographic barrier, which must be overcome by dispersal; (b) a habitat barrier, which requires preadaptation or genetic adjustment to the conditions of the new environment; and (c) a biotic barrier, which integrates the forces of predation, herbivory, competition, and interference that must be overcome in the new habitat, or the new mutualistic relationships that must develop. Additional complexity can be added by, for example, splitting the geographic barrier into two components (to account for factors limiting introduction to the region and dispersal within the region, respectively), by adding a reproductive barrier (to account specifically for factors that potentially limit seed set), or by splitting the biotic barrier into components (e.g., to isolate the role of mutualisms) (Fig. 3).

A. Stages of Invasion

1. Dispersal to a New Area

Important reasons for the intentional widespread translocation of plants include agriculture and forestry and agroforestry, botanical gardens, horticulture (including the commercial trade in seeds, bulbs, and cuttings and urban gardeners using seed exchanges), and soil stabilization. Many plants have been moved around the world inadvertently, notably in ship ballast, with military transport, and as contaminants in fertilizers, hay and straw, grains, wool, and cotton.
2. Establishment and Naturalization

Empirical evidence shows that the chance of becoming established, naturalized, and later invasive increases markedly with an increase in the number of propagules introduced, and with multiple introductions (including introductions at different times and from different source populations). More propagules reduce the likelihood of extinction and increase the chance of long-distance dispersal. Multiple introductions allow the incipient invader to sample a greater range of sites over space and time in the new environment and (since different introductions often originate from different source populations) improve the likelihood of introducing a genotype closely suited to local conditions. Also, multiple introductions increase the likelihood of forming novel genotypes that facilitate invasion. Successful establishment entails dealing with numerous physical, chemical, and biotic barriers (Fig. 3). Many introduced plants are initially grown in small populations that are inherently susceptible to extinction due to chance events. To establish and persist, a population must exhibit $dn/dt > 0$ when $N$ is small (the “invasion criterion”). If an introduced plant can deal with various reproductive barriers, it becomes naturalized.

3. Spread

Invasion involves dispersal within the new area and population growth. Invasive alien floras show a wide range of adaptations for dispersal. Many species are dispersed by “passive” agents such as water or wind. A large proportion of the world’s most widespread and damaging invaders are dispersed by birds and mammals (both native and introduced). The rapidity with which these mutualistic seed-dispersal interactions establish suggests that vertebrate-dispersed plants have converged into generalized dispersal syndromes regardless of phylogenetic and geographical origins. Epizoochorous dispersal, mainly by cattle and sheep, facilitates the spread of many (mainly herbaceous) invaders. Vegetative reproduction is also important.

The dynamics of range expansion and population growth of an invasive alien plant typically follow the pattern shown in Fig. 4. There is frequently a time lag between the arrival of an alien plant in a new habitat and the start of widespread invasion. Examples include Thlaspi cornutum (Brassicaceae), which was cultivated at Oslo Botanical Garden in Norway since 1814, was first collected as an escapee in 1874, spread slowly until 1900, then expanded rapidly until it reached most of its present range in about 1945; Mimosa pigra (Fabaceae), which was virtually confined to small areas around Darwin in Australia for 80 years before exploding; and Melaleucaquinuifera (Myrtaceae), which showed no invasive tendencies for its first 30 years in Florida (United States). A recent analysis of the history of woody species introduced to Brandenburg in Germany revealed an average time lag between introduction...
PLANT INVASIONS

and invasive spread of 147 years (170 years for trees; 131 for shrubs)! Such lags, frequently alluded to but very seldom adequately explained in the invasion literature, are probably usually due to one or more of the following factors: (a) the founder population may maintain a stable, small population until genetic adjustment occurs or essential mutualists (seed dispersers, pollinators, mycorrhizal fungi) arrive (the extent of genetic alteration preceding or accompanying invasion has been studied for a few widespread invaders, e.g. *Ailanthus altissima* in the United States); (b) some lags are probably explained by an initial shortage of “safe sites,” which have become more abundant as human-induced disruption of ecosystems has increased (increasing time also improves the chance of a potential invader encountering a “safe site” formed by a rare event such as a flood); (c) populations spread slowly at their periphery but only show accelerated growth rates when there are many foci of growth.

The lag phase is followed by a phase of sudden growth during which populations increase at an exponential rate (and generally become noticed as invaders). One reason for the increased growth rate (besides those mentioned earlier that may prevent its realization) is the typical two-phase pattern of spatial expansion. This involves the densest recruitment of offspring close to founder populations (“neighborhood diffusion”) and the establishment of isolated colonists through long-distance dispersal. As satellite foci grow in size through diffusion, often coalescing with each other and the founder metapopulation, more propagules become available for additional jump dispersal. With increasing numbers of growth foci, population growth and range increase rapidly. The rate of spread is frequently augmented by intentional or accidental movement of plants within the invasion arena by humans, thus creating additional nascent foci. This process, termed “stratified diffusion,” has been documented for many plant invasions involving disparate plant taxa and environments and is evident at scales ranging from global (Fig. 2),

![Diagram](image)

FIGURE 5 The simulated spread of alien pine trees in South African fynbos over 100 years in an area of 1.5 km × 1.5 km to illustrate the process of stratified diffusion. All simulations start with 75 plants; A shows spread from the edge of a plantation (at bottom left); B shows spread from a clump in the center, and C shows spread from plants randomly arranged across the invasion arena. White pixels show areas occupied by invading pines; black areas are free of pines. Details of the modeling procedure are described in Higgins and Richardson (1998).
PLANT INVASIONS

The end of the exponential phase usually occurs when most optimum sites for invasion are occupied (or when successful control is instituted). Alien plant invasions may also follow linear trajectories (e.g., in the case of spread along rivers or coastal foredunes). Invasions in these habitats usually proceed as a wave front.

Spread rates reported for invasive alien plants vary greatly. Some examples are 5000 m yr$^{-1}$ for Bromus tectorum (Poaceae) in North America; 4000 to 13,000 m yr$^{-1}$ for Heterotheca latifolia (Asteraceae) in the Georgia piedmont; 970 m yr$^{-1}$ for Fraxinus ornus (Oleaceae) along a river in France; 76 m yr$^{-1}$ for Mimosa pigra (Fabaceae) in wetlands in northern Australia; 21 to 33 m yr$^{-1}$ for Acacia cyclops (Fabaceae) and Pinus pinaster (Pinaceae) in South African fynbos; and 14 m yr$^{-1}$ for Ammophila arenaria (marram grass; Poaceae) on coastal dunes in California.

V. INVASIVENESS AND INVASIBILITY

A major thrust of studies on plant invasions in recent decades has been the search for traits that separate invasive species from noninvasives and features that distinguish invaded systems from those that are apparently able to resist invasion. Although the two concepts are intimately linked, many studies have addressed the issues separately.

A. Invasiveness

Numerous studies have described the features of individual alien plant species that have enabled them to invade given areas. For example, the "recipe for success" for Lantana camara (Verbenaceae) includes the following ingredients: (a) dispersal of seeds for long distances by many native and introduced birds; (b) toxicity of its fruit for many mammals, which limits damage by herbivory; (c) its ability to resprout vigorously following damage (e.g., by trampling); (d) its ability to invade a wide range of habitats, including sites disturbed by alien mammals on islands; (e) production of allelopathic substances, which improves its competitive ability; and (f) its ability to flower profusely for long periods, thus attracting pollinators and ensuring copious seed set.

Attempts have been made to define invaders for particular regions (usually biomes or countries) or the whole world and to partition groups of plants into invaders and noninvasives. A notable early attempt to delimit traits of plants that enable them to become weeds or invaders was Herbert Baker's description of "the ideal weed." Briefly stated, Baker's hypothetical superweed is a plastic perennial that germinates in a wide range of physical conditions, grows quickly, flowers early, is self-compatible, produces many seeds that disperse widely, reproduces vegetatively, and is a good competitor. Many of the world's most widespread invaders and weeds possess many "Baker characters," but none has all, and many notable invaders possess very few, if any, of these traits. Are there an optimum number of Baker's traits that confers invasiveness? The work of Mark Williamson and coworkers on British annual plants suggests that optimum weediness is realized in plants with an intermediate number of Baker characters (4 or 5 of 7).

There is evidence that plants that have performed best as alien invaders are not evenly distributed among taxonomic groups. Among the angiosperms, two families that are clearly overrepresented among the most widespread and damaging invaders of natural communities are Fabaceae and Poaceae. Other families that seem particularly prone to contribute invaders of natural areas are Myrtaceae, Rosaceae, Salicaceae, and Tamaricaceae. The Asteraceae, the largest family of flowering plants, contributes many species to alien floras in most regions, but the family is not significantly overrepresented.

Many studies have attempted to gain an understanding of the life-history attributes of plants that makes them successful invaders by comparing traits of native and alien species or invasive and less invasive alien plants. Rejmanek (1998) considered perspectives from many of these studies and added many new insights to arrive at a general model of seed plant invasiveness based on recognized and assumed causal and correlative relationships between a wide range of traits. This model is probably near the limit of what can be achieved in defining invasiveness at a global scale, and further advances will require the development of models based on plant-environment interactions (discussed later). The most consistent, and practical, predictor of invasiveness is still whether the species has become invasive elsewhere (discussed later).

B. Invasibility

Elton's The Ecology of Invasions by Animals and Plants (which, it should be noted, is heavily biased in favor of evidence on animal invasions) put forward the view...
that undisturbed native communities are not susceptible to invasion by introduced species. Recent debate on this topic has been rather futile since so few communities on earth are unaffected by human activities. What has become abundantly clear over the past few decades is that disturbance, be it naturally occurring or human induced, is a fundamental driver of plant invasions. Irrespective of other factors that facilitate or limit invasions, the susceptibility of communities to invasion by alien plants increases with increasing disturbance up to a threshold after which disturbance acts as a barrier. Useful insights on the types and magnitude of disturbance required to initiate and sustain invasions emerge from examining cases of invasion of pine trees (Pinus species) in many parts of the Southern Hemisphere. Most records document pine invasions in areas where the natural disturbance regime has been noticeably altered by humans. Changes to disturbance regimes have involved mainly increased or decreased herbivore pressure (grazing, browsing, and trampling), altered fire frequency, or mechanical clearing of vegetation. Different vegetation types differ in the type and magnitude of disturbance required to facilitate invasions. All records of pine invasions into native forests showed that severe disturbance is required to facilitate seedling recruitment and population growth. Invaded grasslands in Australia, New Zealand, and South Africa invariably have markedly changed disturbance regimes. The same applies for most invaded shrublands. The ultimate cause of increased invasibility caused by disturbance in these cases is clearly the reduced competition from resident plants through the reduction in ground cover. Several pine species have spread into South African fynbos where the only disturbance is the naturally occurring fire; intense fires and the dearth of vigorous plants immediately postfire combine to create opportunities for invasion. In general, intermediate levels of disturbance clearly make sites more vulnerable to invasion by pines, whereas severe disturbance usually prevents invasion. Naturally occurring disturbances where vegetation cover is removed is caused by factors such as volcanicity, slope instability, wind, and flooding; these are also important determinants of invasibility for pines. "Correlative" evidence from natural experiments such as this is very useful for defining the main factors that affect invasibility (Table I). This example shows that "invasibility" needs to be context specific; this pattern of invasibility relates to pines and probably not to other growth forms or to alien species in general.

Empirical evidence shows that certain communities are more susceptible to invasion by a wide range of alien plants than others. For example, the greater susceptibility to invasion of island communities (the more isolated, the more invasible) compared to mainland ones has long been noted. Islands where established alien species make up 40% or more of the flora include Bermuda, Hawaii, Ascension, Rodrigues, Tristan da Cunha, Lord Howe, and New Zealand (figures for continental areas are much lower). Among the many reasons advanced to explain this are (a) the species richness of islands is low (theoretically reducing the likelihood of an introduced species encountering a close competitor); (b) island floras have evolved in isolation, often without adaptation to high levels of competition, grazing, trampling, or regular burning; (c) the small size of islands means that the history of human intervention is "concentrated" (more intense disturbance); and (d) many islands were colonized very early and have been the crossroads of intercontinental trade (and thus were exposed to greater numbers of potential invaders).

Among other systems that seem particularly susceptible to alien plant invasions are riparian zones. In virtually every region, these zones are among the most severely invaded habitats. An important reason for this is the frequent disturbance from water-level fluctuation, including floods, which disperses propagules and provides favorable sites for establishment.

There has been much debate on the link between the number of species in a community and its susceptibility to invasion. Some "correlative" studies report a negative correlation between plant species richness and invasibility, whereas others report a positive correlation. In many cases, such confusion is undoubtedly attributable to the overriding importance of disturbance (e.g., several studies point out that species-rich riparian zones are highly prone to invasion). Systematic experiments to explore the role of species numbers and identities have only begun in the past decade. Results to date emphasize the importance of considering spatial scale and productivity. For productive, small-scale grassland communities, for example, the identity of species is more important than the number of species.

C. Linking Plant Traits to Environmental Features

There have been important recent advances in our understanding of the determinants of invasiveness and invasibility. There is now a reasonable understanding, at least for some severely invaded regions, of the plant traits that are correlated with success as invaders and also the environmental features that mediate invasions.
### TABLE I
Determinants of Invasibility for Pines (*Pinus* spp.; Pinaceae) Introduced to the Southern Hemisphere

<table>
<thead>
<tr>
<th>Factors</th>
<th>Facilitating feature(s)</th>
<th>Limiting feature(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species attributes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seed mass</td>
<td>Small seeds with large wings</td>
<td>Large seeds with small wings</td>
</tr>
<tr>
<td>Juvenile period</td>
<td>Short (&lt;10 years)</td>
<td>Long (&gt;10 years)</td>
</tr>
<tr>
<td>Interval between large seed crops</td>
<td>Short (&lt;3 years)</td>
<td>Long (&gt;5 years)</td>
</tr>
<tr>
<td>Ability to survive moderate browsing levels</td>
<td>Good</td>
<td>Poor</td>
</tr>
<tr>
<td>Residence time</td>
<td>Long (&gt;30 years)</td>
<td>Short (&lt;50 years)</td>
</tr>
<tr>
<td>Extent of planting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total area</td>
<td>Large</td>
<td>Small</td>
</tr>
<tr>
<td>Boundary: total area ratio</td>
<td>Large</td>
<td>Small</td>
</tr>
<tr>
<td>Ground-cover characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic vegetation structure</td>
<td>Bare or sparsely vegetated ground, shrubland, grassland</td>
<td>Forest</td>
</tr>
<tr>
<td>Vegetation cover</td>
<td>None-low (&lt;30%)</td>
<td>High (&gt;80%)</td>
</tr>
<tr>
<td>Latitude (%)</td>
<td></td>
<td>&lt;30</td>
</tr>
<tr>
<td>Disturbance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>Low-moderate</td>
<td>Very low/very high</td>
</tr>
<tr>
<td>Human-induced types</td>
<td>Moderately increased herbivore pressure (grazing, browsing, trampling) or equivalent</td>
<td>Greatly increased or greatly reduced herbivore pressure or equivalent</td>
</tr>
<tr>
<td>Contributing factor</td>
<td>Decreased competition from ground-cover</td>
<td>Increased competition from ground-cover or physical elimination of pines (e.g., by mechanical clearing)</td>
</tr>
<tr>
<td>Natural types</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resident biota</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composition of plant community</td>
<td>Naturally invadable community (e.g., <em>Dacrydium laabeanum</em> shrubland, Chionochloa tussock grassland in New Zealand)</td>
<td>Naturally resistant community (e.g., <em>Eucalyptus</em> Hakea forest in SE Victoria, <em>Poa</em> tussock fynbos in South Africa)</td>
</tr>
<tr>
<td>Indicators of invasibility</td>
<td>Conditions unsuitable for C₄ photosynthetic pathway and nutrient-poor soils, paucity of vigorous herbs</td>
<td>Conditions suitable for C₄ photosynthetic pathway and nutrient-rich soils, abundance of vigorous herbs</td>
</tr>
<tr>
<td>Role of mammals other than humans</td>
<td>Removal of competing vegetation (e.g., through grazing)</td>
<td>Destroying pine seedlings (browsing, trampling)</td>
</tr>
<tr>
<td>Role of fungi</td>
<td>Presence of appropriate mycorrhizal symbionts</td>
<td>Absence of mycorrhizal symbionts (no longer limiting following intentional widespread dissemination of spores)</td>
</tr>
<tr>
<td>Influence of pathogenic fungi</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Derived from the assessment of invasion histories of 19 *Pinus* species in Argentina, Australia, Brazil, Chile, Madagascar, Malawi, New Caledonia, New Zealand, South Africa, and Uruguay. Factors should be viewed as loadings—the greater the number of facilitating features, the greater the chance of invasion for any taxon/site combination. Modified from Richardson et al. (1994).

Further progress, certainly at scales of resolution that have value in management, will rely on the development of models that link plant traits with critical environmental features. Such models show that interactions between the various determinants of invasive success are sometimes at least as important as the main effects. For example, in the case of pines invading forests, shrublands, and grasslands in the Southern Hemisphere, interactions between basic features of the environment, the features of the disturbance regime, and plant traits...
explain the different spread rates observed in different areas.

VI. MODELING PLANT INVASIONS

The use of models has greatly improved our understanding of invasions and how to deal with them. The many types of models that have been applied to plant invasions may be crudely grouped in three categories. Simple demographic models include exponential, logistic, logistic-difference, and stochastic models; they predict the future number of individuals in a population by making assumptions about the nature of population growth and by estimating demographic parameters regarded as being important in determining population dynamics. Spatial-phenomenological models describe plant-environment interactions using empirical data (invoking no ecological mechanisms); they include regression models, geometric models and Markov models. Spatial-mechanistic models are based on independent estimates of ecological parameters affecting invasions; they include reaction-diffusion models, population dynamic metapopulation models, and individual-based cellular automata models.

A particularly influential model in the study of invasions was that published by J. G. Skellam in 1951. He used a diffusion equation, combined with estimates of population growth, to show that invasion fronts advance at a constant velocity. Skellam's model, on which most subsequent invasion models have been based, did not incorporate long-distance dispersal, now known to be critically important for modeling the rate and variability of an expanding population front. Recent advances in mathematical modeling have facilitated the development of spatial-mechanical models (notably individual-based, cellular automata mixture models), which allow much more accurate modeling of all stages of plant invasions. Besides offering exciting opportunities for exploring critical processes in invasion, thus contributing to invasion theory, these models have important applications in management.

VII. MANAGING PLANT INVASIONS

There are three basic options for managing invasive alien plants: (a) prevention, (b) removal ("eradication"), and (c) ecological management. The appropriateness of different control approaches depends on the ecology of the invader and the invaded system, and on a myriad of socioeconomic issues.

A. Prevention

Despite pessimistic prognoses in the 1980s regarding the ability of ecology to predict plant invasiveness and invasibility at scales useful for management, important advances have been made. Various methods have been developed for screening introduced plants to assess the risk of them invading particular habitats. As mentioned earlier, a good predictor of invasiveness is whether a species has invaded other (similar) areas where it has a longer history as an alien. For example, 90% of invasive plant species in Australia are also invasive in other locations. A major problem with applying this concept to management or regulation is the time lag inherent in many (most) invasions—at what stage should the "performance" of a species be scored with regard to invasiveness? Also, many species are currently being introduced directly from their native ranges to many different areas simultaneously.

B. Removal

The number of options for controlling an invasive plant decline rapidly when the phase of exponential increase (Fig. 4) is reached. Management is clearly most effective when directed at removing invaders while they still occupy a small range and occur in small populations. Many potentially widespread invaders are effectively kept in check by various means of control at this stage. Removal is much less feasible once the invaders have spread over large areas, but is still attempted in some cases, usually when the damage the invaders cause over-rides economic constraints and when other control measures are unavailable or impractical. For example, expensive, large-scale clearing of invasive Hakea and Pinus spp. in South African fynbos is warranted by estimated costs of reduced water production in the absence of control.

C. Ecological Management

Experience has shown that satisfactory control of plant invaders is usually only achieved when several complementary methods, including biological control, improved land management practices (e.g., through pre-
Invasive plants are being effectively managed in many parts of the world, albeit at a great, and increasing, cost. The dimensions of the problem are changing rapidly, and management strategies need to be flexible enough to deal with exigencies. Among the main reasons for changing trajectories of invasions are the following:

- The increasing magnitude and pervasiveness of global trade and free trade agreements.
- The increasing availability and demand for alien plants for a wide range of purposes, especially for agroforestry and horticulture.
- Global change (including climate change, changed nutrient regimes, elevated CO₂, fragmentation); although little is known of how all these factors affect invasions, especially in concert, there is consensus among ecologists that these changes greatly exacerbate problems with invading alien plants.
- Genetic engineering (GMOs, transgenic plants); genetically modified plants could acquire novel traits that may make them (better) invaders.
- Conflicts of interest; many alien plants are essential crops in some parts of the landscape/region/country and damaging invaders in other parts. Good examples are many tree species used in forestry and agroforestry, and Echium plantagineum (Boraginaceae), an important dry-season forage and a honey plant in South Australia, but an important weed of pastures in New South Wales. A major challenge is to develop objective methods for assessing the costs and benefits associated with our increasing dependence on alien species.

See Also the Following Articles

DISTURBANCE, MECHANISMS OF • INTRODUCED SPECIES, EFFECT AND DISTRIBUTION • MIGRATION • PLANT-SOIL INTERACTIONS

Bibliography


I. Introduction
II. Small-Scale Processes: Micron to Millimeter Scales
III. Mesoscale Processes: Centimeters to Meters
IV. Large-Scale Interactions: Stands and Biomes
V. Implications for the Management and Conservation of Ecosystems

GLOSSARY
aerenchyma Porous root tissue, especially well developed in wetland plants, that allows diffusive flux of oxygen from above-ground tissues to root tips. This tissue both supports the respiratory demand of the root tissues and allows oxygen to leak into the surrounding soil.
mycorrhizae A root tip that is infected with fungi in a mutually beneficial partnership. The fungal hyphae explore large volumes of bulk soil, absorbing nutrients and transferring them to the plant; the plant supplies the organic carbon necessary for growth and energy production to the fungus. Different groups of fungi form vesicular-arbuscular mycorrhizae (fungal hyphae invaginate into the plant root cells) and ecto-mycorrhizae (fungal hyphae grow between plant root cells and form a thick sheath over the root tip, but they do not invaginate). Several other forms are specific to particular plant families (Ericaceae, Orchidaceae).

net mineralization/immobilization The net result of microbial decomposition of organic matter is either the incorporation of nutrient elements (particularly nitrogen) into the microbial biomass, rendering it unavailable for plant uptake (immobilization), or their release into the soil solution (mineralization) after microbial demand for each element has been satisfied.
nutrient uptake capacity The instantaneous rate of nutrient acquisition, usually measured in brief (1–2 hr) incubations. Uptake capacity reflects the abundance of transport sites on the root cell membranes and their affinity for nutrient ions.
rhizodeposition The mixture of sloughed cells, mucilages, and small-molecular-weight sugars, amino acids, and other compounds leaked from root cells, which are deposited in the soil adjacent to the surface of fine roots. Exudation takes place from the root tip back to the zone of suberization. The chemical quality and quantity of the exudate is altered by the presence of mycorrhizae.
rhizosphere Volume of soil adjacent to, and strongly influenced by, a plant root. The rhizosphere is usually considered to extend about 2 mm from the root surface, and includes the "rhizoplane," or soil directly in contact with the root surface.
plasticity Ability of a plant to respond to temporal changes or spatial variation in environmental conditions by altering the size or the distribution of plant parts. These are phenotypic, rather than genetic changes.
siderophore. Chemicals secreted by roots (primarily non-protein-forming amino acids), which complex with insoluble metal ions bringing them into solution and permitting their transport to and uptake into the root.

soil aggregate. A crumb-sized unit of soil, composed of aggregated soil minerals, microbes, and soil microfauna, which are cemented together by a combination of biological materials such as polysaccharide secretions, fungal hyphae, and chemical substances such as precipitated carbonates or silicates. Aggregates are classified by size and stability in water (dissintegrating versus retaining their structure and integrity).

soil organic matter. Organic substances, including a wide variety of carbohydrates, proteins, lipids, waxes, phenolic, and humic compounds, which accumulate in soil as a result of both plant and microbial growth. These compounds include small-molecular weight materials, which are rapidly decomposed to carbon dioxide; larger compounds, which may be slowly decomposed over years to decades; and large, complex, aromatic substances, which may be stable within the soil for millennia. Soil organic matter affects all aspects of the soil’s biology, chemistry, and physics.

soil texture. The relative abundance of sand (50 μm < \( \phi < 2 \text{ mm} \)), silt (2 μm < \( \phi < 50 \text{ μm} \)), and clay (\( \phi < 2 \text{ μm} \)) particles in the soil (USDA criteria). Particle size distribution determines the distribution of pore sizes, which in turn strongly affects the behavior of water in the soil.

PLANT-SOIL INTERACTIONS include a wide range of biological, chemical, and physical effects exerted by soil on plant survival, growth, and reproduction, and reciprocal effects of plants on soil formation, soil physical structure, and the activities of the soil biota.

I. INTRODUCTION

Plants are often assumed to be passive recipients of the environmental conditions that happen to exist at the location in which they grow. Soils supply mineral nutrients and water, and they provide anchorage, but they also set limits on plant growth. Plants, however, provide the organic substrate that forms the basis for all biological activity in the soil and that forms the soil organic matter; their growth drives the formation of soil from bedrock and alters its chemistry and physics. This chapter reviews the two-way street that constitutes the plant-soil system and examines the implications of these mutual effects for the management and conservation of biodiversity.

Soil-plant interactions take place over a wide range of spatial and temporal scales (Fig. 1). At a scale of microns, roots etch microscopic channels of weathering within mineral grains and provide a variety of different habitats for microbial growth. These microscopic phenomena cause millimeter-scale differences in the abundance and diversity of microbes and the soil micro- and mesofauna. Nutrients flow toward roots over distances of millimeters to centimeters; active uptake causes sharp concentration gradients of nutrient elements, moisture, and acidity. These modifications spread out from the root surface to form a cylinder of differentiated, rhizosphere soil that may be millimeters thick. Mycorrhizal hyphae extend from the root surface into the soil over distances from centimeters to meters, they bind soil particles together into aggregates and contribute to the long-term accumulation of soil organic matter. The quantity and quality of such millimeter-scale effects

FIGURE 1 Concept of three-dimensional structure of plant-soil interactions. Processes occurring across dimensions of time (minutes to millennia) and space (microns to hundreds of kilometers) occur with respect to the species pool—both the numbers of species and the kinds of species (see text).
vary among plant species, and so patches of soil filled by roots of different species may become chemically and physically different at scales of centimeters to tens of meters, depending on plant size. These species-level effects may translate into stand-level effects. At these larger spatial scales, however, local and regional differ-

tiation of soils due to differences in topography, geo-

logical substrate, and geological history may override the effects of plant species and constrain the nature of the plant-soil interactions. But at these scales, plant communities also exert control over the long-term accumulation of organic matter.

Plant-soil interactions occur over an equally great range of temporal scales. At the level of the individual plants, uptake of nutrients and water occurs more or less continuously over hours, but rates vary over sec-

onds or minutes, depending on environmental condi-
tions. The aggregate effects of this uptake over periods of months (seasons) or years reflects the continuity of plant community composition. Similarly, the exudation of substances used by the microbiota, and the response in microbial growth and metabolic activity, occurs over timescales of minutes to hours, but the aggregate effect on soil properties and on plant growth extend over seasons and years. Over the long term (decades to cen-
turies or millennia), different plant communities cause different types of soil to develop.

Finally, these patterns of interaction themselves de-
pend on the size and diversity of the available pool of plant species. In highly diverse communities, the effects of individual species may be masked by the variability among those effects, or by the overlapping nature of root and shoot systems. In communities that are species-poor, either because of human manipulation (e.g., forest plantations of a single species) or because of biogeographic and climatic constraints (e.g., boreal for-
ests with few species), the individual species-specific effects seen at small scales may translate into effects seen over hectares or square kilometers.

Within each of the temporal and spatial scales of inter-
action, the mechanisms creating the interaction are sub-
ject to feedback, both positive and negative (Fig. 2). In-
deed, the presence of feedback in the soil-plant system is one of its salient properties: the results of a particular mechanism can alter the environment so as to modify (increase or decrease) the continued activity of that mechanism. Feedbacks have been recognized as crucial in mediating such disparate processes as the supply of mineral nutrients through decomposition, and the activi-
ties of soil pathogens and mycorrhizae.

The complex interplay of soil and plants has com-
manded attention for a long time. Gilbert White, in "The Natural History of Selbourne" in 1778, observed that "In a district so diversified with such a variety of hill and dale, aspects, and soils, it is no wonder that great choice of plants should be found." Darwin pointed out that plant growth is modified by the presence of soil fauna, especially earthworms; he noted that Gilbert White had also observed this a century earlier. The German scientist Justus Liebig (1876) founded the scient-
ific basis for agriculture by emphasizing the role of soil in controlling vegetation and crops through the supply of mineral nutrients. Indeed, the perception of soil as the ultimate control on the development of plant communities has been a central paradigm of plant ecol-

ogy for at least a century. In a parallel development, the role of plants in shaping soils was recognized with the beginnings of soil science (Dolomieu, 1879; cited in van Breeman 1995). Hans Jenny, in his classic work "Factors of Soil Forma-
tion" in 1941, identified plants (and animals) as one of the primary forces in the genesis of soil from bedrock, and acknowledged that soil scientists extending back to the nineteenth century had perceived plants and soils to be a "coupled system." Conversely, the critically important role of plants in causing soils to form from weathered and unweathered rock was elaborated in de-
tailed studies of primary succession on newly exposed substrate by Crocker and Major in 1955. In the past several decades, the multiple forms of interaction be-

tween plants and soil have become a major focus for research.

Plant-soil interactions play a critical part in human interactions with the biosphere. Humans create plant
emerging new lateral roots ("exudate"). These materials leaked from intact cells and cracks caused by secretions of mucilage from these cells, and soluble cells sloughed from the root cap and the epidermis, root into the rhizosphere. This material includes whole position, the transfer of organic substances from the plant, and degree and type of mycorrhizal infection from the root tip, root age, physiological stress to the changes longitudinally along the root with distance from the root surface; thus, there is a presence on the soils decreases rapidly with distance from the root surface. Not only is the rhizosphere most strongly affected by rhizodeposition, the transfer of organic substances from the root into the rhizosphere. This material includes whole cells sloughed from the root cap and the epidermis, secretions of mucilage from these cells, and soluble materials leaked from intact cells and cracks caused by emerging new lateral roots ("exudate"). These materials mix with polysaccharide secretions of the adhering bacteria. The mucilage, or "mucigel," lubricates the root surface as it pushes past soil particles and helps to maintain the physical continuity between the root and the bulk soil. In addition, the mucigel may adsorb nutrient ions, especially cations, and prevent their removal from the root surface by leaching. Table I illustrates the diversity of compounds released from various species. The relative proportions of each class of compound may vary with position along the root, for example, amino acids are released in higher amounts near the root tips than farther back.

Studies in a variety of plants have shown that approximately 10 to 20% of the carbon fixed during photosynthesis is released to the soil through these mechanisms. This rhizodeposition accounts for 40 to 90% of the total amount of carbon transferred to the roots from the shoots, the remainder is respired by the root tissues or is used to construct new root biomass. Commonly cited values of 10 to 100 mg g$^{-1}$ root dry weight for soluble substances and 100 to 250 mg g$^{-1}$ root for insoluble materials give some indication of the amounts released. The amounts of carbon transferred from the shoot to the root system and the fraction of that carbon that is released to the soil vary greatly with plant species and soil conditions (Table II). mycorrhizal infection, soil microbial abundance, and soil physical conditions. The total amounts of carbon involved in the transfer can be quite large (an estimated 1.2 t C ha$^{-1}$ yr$^{-1}$ to 7.3 t C ha$^{-1}$ yr$^{-1}$). The effects of this input of carbon and nitrogen to the rhizosphere are usually in the range of 10$^3$ to 10$^5$ cells cm$^{-3}$, compared to 10$^2$ cells cm$^{-3}$ in bulk soil. Figure 3b illustrates a typical distribution of microbes with respect to the root surface. Rhizosphere microbial cell densities: bulk soil cell densities range from $<$1 for microalgae to $>$24 for many bacteria; for most microbial groups, ratios are $>$5. Successions of different microbial species occur (a) over time, as newly produced root tips extend into the soil and are colonized, (b) over the length of the root (from root cap to zones of suberization and secondary thickening), and (c) with distance out from the root surface. Not only is the population of microbes in the rhizosphere larger than in the bulk soil, but it differs in composition. Bacteria are more abundant than fungi, while the reverse is often true in the bulk soil, and more species of bacteria may be found in the rhizosphere. Several genera of bacteria capable of fixing atmospheric N$_2$, including Enterobac-
PLANT-SOIL INTERACTIONS

FIGURE 3  (a) A schematic diagram of the rhizosphere, illustrating different zones of the root tip (root cap, zones of cell elongation, and epidermal cell lysis), different types of mycorrhizal fungal growth (endomycorrhizal and ectomycorrhizal), and rhizosphere soil. The diagram also illustrates the zone of decreasing root influence with distance from the soil. (b) Distribution of microorganisms with distance from the roots, based on Bazin et al. in Lynch (1990); different species of microorganisms utilize substrates directly derived from roots or secondary materials produced by other microbes in adjacent zones.

er, Klebsiella, Azotobacter, Azospirillum, Bacillus, Pseudomonas, and Clostridium, are only found in the rhizosphere; these nonsymbiotic bacteria may (but not always) contribute to greater plant growth. Species of Pseudomonas are particularly important, as many have been found to benefit plant growth. The composition of the rhizosphere biota changes rapidly (within months) in response to changes in the species composition of the plant community, demonstrating the close coupling of the rhizosphere community to the species—specific morphology and physiology of root systems. Variation in the abundance and diversity of plant pathogens also reflects the differences in rhizodeposition among plant species.

Structural complexity in the zone of root/soil contact is provided by extensions of the root surface into the soil matrix. Most plants produce extensive root hairs from cells just distal to the root cap and zone of elongating cells. The zone of root hair development extends only a short distance along the root (1–4 cm), as the epidermal cells producing the hairs are sloughed off in more basal areas (Fig. 3a). Production of root hairs is itself affected by soil chemical conditions (e.g., pH, ion concentrations), texture, and water content. This zone provides increased structural complexity (e.g., a greatly expanded root surface area that is in direct contact with soil particles). Structural complexity also results from mycorrhizal infection, particularly by ectomycorrhizae, as fungal hyphae spread out from the root surface into the soil (Fig. 3a). Ectomycorrhizal fungi, in particular, produce large quantities of extra-radical mycelium, hyphae may extend for several meters from the host plant. These hyphae greatly expand the contact zone between the root and the soil.
TABLE I
Compounds Identified in Root Exudates, Classified by General Type

<table>
<thead>
<tr>
<th>Type</th>
<th>Compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple sugars</td>
<td>hexoses (glucose, fructose, galactose), pentoses (arabinose, xylose, rhamnose), di- and oligosaccharides (wide variety), uronic acids, amino sugars</td>
</tr>
<tr>
<td>Amino acids</td>
<td>α-α-Alanine, aspartic acid, phenylalanine, glutamic acid, leucine, serine, threonine, α- and γ-amino butyric acids, γ-aminobutyric acid, γ-aminoadipic acid, many others</td>
</tr>
<tr>
<td>Aliphatic acids</td>
<td>formic, acetic, oxalic, citric, malonic, tartaric, malic, succinic, lactic, palmitic, stearic, oleic, linoleic, linolenic, others</td>
</tr>
<tr>
<td>Aromatic acids</td>
<td>p-Hydroxybenzoic, ferulic, o-cumaric, gallic, vanillic, sinapic, shikimic, trans-cinnamic</td>
</tr>
<tr>
<td>Complex carbohydrates</td>
<td>Polygalacturonic acid, other hydroxy-, ket-, di- and tricarboxylic acids</td>
</tr>
<tr>
<td>Enzymes</td>
<td>Several phosphatases, peroxidase, invertase, urease, α-glucosidase</td>
</tr>
<tr>
<td>Vitamins</td>
<td>B group, biotin, pantothenic, niacin, pyridoxine, riboflavin</td>
</tr>
<tr>
<td>Proteins and peptides</td>
<td>Unspecified</td>
</tr>
<tr>
<td>Plant hormones</td>
<td>Auxins, gibberellins, kinetins, ethylene</td>
</tr>
<tr>
<td>Other</td>
<td>indoles, indole-3-carboxylic acid, organics, hormones, terpenoids, lectins, phytoalexins, alcohols</td>
</tr>
</tbody>
</table>

* Commonly occurring and/or representative specific compounds are listed; in all cases, a very large number of specific compounds have been identified in each chemical class. From Vancura and Kunc (1988).

B. Nutrient and Water Uptake: The Function of the Root/Soil Interface

Nutrient supply to plants, at the scale of the root surface and the immediately adjacent soil, is a complex function of the chemistry of individual nutrient ions. The availability of nutrients other than nitrogen is ultimately determined by the chemistry of the soil minerals present, which in turn reflects the geology of the parent material. Weathering releases these nutrients from the crystalline lattice of the minerals into solution, making them available for uptake; however, both physical and chemical weathering rates are themselves strongly affected by vegetation. Roots penetrate cracks and fractures in bedrock, widening them, allowing water to penetrate, and promoting chemical weathering. Chemical weathering depends on the supply of protons for the dissolution and transformation of soil minerals. Plant-derived sources supply most of acidity involved in weathering; these include CO₂ from root and microbial respiration, which forms H₂CO₃ in the soil solution, organic acids in root exudates and the products of microbial decomposition of plant tissues, and hydrogen ions released from roots in exchange for nutrient cations. Soil genesis also results from the chelation and mobilization of iron and aluminum by the organic acids. As different species of plant generate different types and amounts of organic acid formation, the composition of the plant community strongly affects the nature of the weathering and pedogenic processes that take place. Thus, processes at the scale of the root surface result in broad-scale patterns of correlated vegetation and soil development.

Plant-mediated weathering and pedogenic reactions generate a supply of nutrient cations; their availability for uptake depends on a different set of intersecting plant-soil processes. This includes the diffusion rates, exchange, adsorption, and precipitation chemistry of each element, patterns of water movement through the soil, the morphology of the root system, the amount and diversity of mycorrhizal fungi, and the physiology of uptake by the roots. Nutrients are brought to the root surface in the mass flow of water and through diffusion; roots encounter nutrients as they grow through the soil. The relative importance of each path-
way varies with the different elements (Fig. 4). The availability of sparingly soluble elements, such as iron and manganese, may be enhanced by siderophores in the exudate. These are non-protein-forming amino acids, which complex with these metals, facilitating their mobility and uptake. Organic acids help to mobilize metal nutrients by lowering pH, and help to solubilize P by preferentially binding to the metals. For diffusion-limited nutrients such as P, zones of depletion develop within the rhizosphere (Fig. 5a); this does not occur with highly mobile ions such as NO₃⁻ (Fig. 5b). The size of the depletion zone reflects the balance of supply (diffusion and mass flow) and demand (uptake rate).

Unlike the nutrient cations derived from soil minerals, the availability of nitrogen depends on the microbiologically mediated decomposition of plant-derived organic matter. It is often assumed that the supply of readily metabolizable carbon from root exudate will promote increased nitrogen mineralization because it relieves carbon limitation, this is not always the case. Theoretical studies have suggested that the net effect of exudates on nitrogen dynamics will depend on the C:N ratio of the exudates relative to that of the indigenous organic matter of the soil, so that balance between the release of N through mineralization versus immobilization in new microbial biomass will vary with the quality of the exudates. Several studies with herbaceous plants have found that the presence of rhizosphere microbes can increase availability of N to the plants, but the effect varies with the availability of N from the soil and the

---

**FIGURE 4**

**FIGURE 5**
(a) Observed patterns of P depletion with distance from the surface of roots of maize and rape: the zone of low P concentrations adjacent to the root is an indicator of the distance that root hairs grow outward from the root surface. (b) Calculated depletion curves (as percent of bulk soil concentration) of three nutrients: the depletion rate out from the root surface depends strongly on the chemical mobility (combination of diffusivity and likelihood of adsorption, precipitation, or other immobilizing chemical processes). From Jungk, in Waisel et al. (1991), “Plant Roots: The Hidden Half,” by courtesy of Marcel Dekker Inc.
kind of plant grown. Alternatively, N may be immobilized within the microbial communities in response to root exudation. Thus, the effects of root exudation on the supply of mineral N to plants is a complex function of the kind of plant, the nature of the organic matter in the soil, and the amounts and composition of the exudate. Root exudate may decrease or increase N availability, even in different horizons of the same soil.

Not least among the effects of roots on the rhizosphere soil is the alteration of soil pH. In addition to the excretion of organic acids and CO₂, roots exchange protons and hydroxyl ions for nutrient ions, particularly ammonium (NH₄⁺) and nitrate (NO₃⁻). Because nitrate is rapidly taken up when it is present, a net imbalance of anion and cation uptake occurs. Maintenance of electroneutrality during this requires that an anion be extruded in exchange for the nitrate ion; this anion is usually hydroxyl (OH⁻). Conversely, hydrogen ions (H⁺) are excreted during uptake of ammonium. Soil pH can rise as much as two units within the rhizosphere when nitrate uptake is rapid, whereas it may decrease by up to two units when all uptake is in the form of ammonium.

C. Physical Structure of Soil

Plants modify the physical arrangement of soil particles in a variety of ways. In most soils, the grains of mineral material and the soil organic matter are bound together in aggregates. In the rhizosphere, the mucigel produced jointly by the plant and the resident microbes bind clay minerals together into microparticles of about 50 μm diameter. Fungal hyphae can bind together these microparticles into larger aggregates of 1 to 2 mm diameter. The products of root growth (and root-supported mycorrhizal fungal growth) are especially important in stabilizing the aggregates over time. Moreover, different plant communities create different quantities and sizes of soil aggregates, depending in part on the morphology of the root system and the extension and mass of mycorrhizal hyphae.

Pore size distributions are also affected by roots through several other mechanisms. Alternate wetting and drying of the soil induced by transpiration can affect cracking patterns, especially in clay-rich soils. Roots also preferentially grow into larger pores, or zones of lower penetration resistance, and may thereby enlarge them. The death and decay of roots also creates macropores or channels of highly porous organic matter; these large voids provide preferential pathways for the flow of water. Finally, the degree of contact between the root surface and the soil varies with root age, amount of exudation, and root physiological status (e.g., hydration). Roots also lose molecular oxygen to the soil, as air diffuses from leaves and stems through lacunae in the cortex tissues and through cracks in the surface of the root epidermis. This "radical oxygen loss" is particularly large in wetland plants, in which the root cortex develops extensive air spaces ("aereenchyma") through the degradation of cell walls. Root oxygen loss can be sufficiently high to alter the reduction-oxidation potential of the surrounding bulk soil (Table III). As a result, a zone of 2 to 3 mm of oxidized soil (often apparent as a coating of orange-red iron oxide grains) may develop around the root surface under anoxic, waterlogged conditions. This zone of oxidized iron minerals acts as a trap for other metal ions in the soil, such as cadmium, copper, lead, and zinc, which become enriched in the surface concentrations as much as 10-fold over the bulk soil concentrations. Moreover, by maintaining a higher redox potential in the soil, denitrification and methano-

---

**TABLE III**

<table>
<thead>
<tr>
<th>Plant species</th>
<th>Growth form</th>
<th>No plant</th>
<th>With plant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acer rubrum</td>
<td>Tree</td>
<td>+280</td>
<td>+320</td>
</tr>
<tr>
<td>Rosa palustris</td>
<td>Shrub</td>
<td>+220</td>
<td>+240</td>
</tr>
<tr>
<td>Salix viminalis</td>
<td>Shrub</td>
<td>+200</td>
<td>+230</td>
</tr>
<tr>
<td>Ranunculus cymbalaria</td>
<td>Herb</td>
<td>+290</td>
<td>+310</td>
</tr>
<tr>
<td>Myriophyllum verticillatum</td>
<td>Herb</td>
<td>+170</td>
<td>+270</td>
</tr>
<tr>
<td>Isoetes lacustris</td>
<td>Herb</td>
<td>+100</td>
<td>+120</td>
</tr>
<tr>
<td>Icacin aemula</td>
<td>Herb</td>
<td>-100</td>
<td>-120</td>
</tr>
<tr>
<td>Spatula alterniflora</td>
<td>Herb</td>
<td>-50</td>
<td>+150</td>
</tr>
</tbody>
</table>

For each study, the maximum observed difference between planted and unplanted soils are reported.

changes in growth rate, differential growth of above-ground tissues of resource allocation by whole plants, including that occur over scales of centimeters to meters. Strategies of resource allocation by whole plants, including changes in growth rate, differential growth of above-ground and below-ground tissues, changes in root longevity and turnover, and physiology (i.e., in changing metabolic capacity for nutrient uptake) affect the vertical and horizontal extent of root penetration through the soil, thereby affecting the spatial extent over which mesoscale plant-soil interactions occur. At this larger scale, the deposition of litter from the above-ground processes becomes as important as the patterns of root growth within the soil; indeed, litter input may dominate the biological processes within the soil. However, spatial variation in soil properties, driven by topography, bedrock geology, and other exogenous factors, also affects root system morphology and growth, and condition the fate of the litter.

III. MESOSCALE PROCESSES: CENTIMETERS TO METERS

The mesoscale interactions between root surfaces and the adjacent soil are part of a larger set of interactions that occur over scales of centimeters to meters. Strategies of resource allocation by whole plants, including changes in growth rate, differential growth of above-ground and below-ground tissues, changes in root longevity and turnover, and physiology (i.e., in changing metabolic capacity for nutrient uptake) affect the vertical and horizontal extent of root penetration through the soil, thereby affecting the spatial extent over which mesoscale plant-soil interactions occur. At this larger scale, the deposition of litter from the above-ground processes becomes as important as the patterns of root growth within the soil; indeed, litter input may dominate the biological processes within the soil. However, spatial variation in soil properties, driven by topography, bedrock geology, and other exogenous factors, also affects root system morphology and growth, and condition the fate of the litter.

A. Nutrient Acquisition and Root Growth: Individual Plants

The acquisition of nutrients and water is strongly affected by the architecture and size of the root system, a function of the allocation of photosynthate below ground. Conversely, the availability of nutrients and water molds the growth of the root system, thus setting up a feedback interaction. The metabolic capacity to absorb nutrients not only varies among plant species, but also varies within an individual as the availability of nutrient elements changes. Nutrient uptake capacity is often inversely related to availability, because transport systems in the root cell membranes increase in activity as nutrients become more limited in supply. For example, barley plants growing in nitrogen-limited cultures have uptake rates of both ammonium and nitrate more than 200% greater than plants growing with excess N; similarly, phosphate uptake rates increase by 900% in phosphorus-limited cultures, and sulfate uptake rates increase by almost 900% in sulfur-limited culture. Similar inverse relationships between availability, as measured in soil solutions or extracts, and short-term uptake rates have been found for a wide range of wild plants, including both herbs and trees. However, while relative uptake capacity varies within a species inversely with the availability of nutrient elements, absolute rates of uptake of plant species from low-nutrient environments may be less than, equal to, or greater than plant species found in high-nutrient environments. Thus, the effects of nutrient availability on root physiology depend on both the species of plant and the variation present in the soil. Nutrient acquisition is also affected by the abundance and distribution of the roots. Different kinds of plants are characterized by differences in the relative allocation of biomass to roots and shoots (Fig. 6). In
general, grasses tend to have higher relative allocation to roots than do woody plants, although trees and shrubs in highly nutrient-limited environments, such as alpine tundra or bogs, have higher root/shoot ratios than do woody plants from more nutrient-rich environments such as tropical or temperate deciduous forests.

Root systems also adjust to varying nutrient availability by changes in morphology (length, diameter, branch density, etc.). Root length density (the length of roots per unit volume of soil) commonly increases in zones within the soil that have higher concentrations of limiting nutrients, as has been frequently demonstrated in experiments (Fig. 7). However, such responses are not seen in all species, and the response to variations in nutrient availability depends on a plant’s tendency to produce thick versus thin and fast-growing versus slow-growing roots. Increases in root length in one area, in response to a patch of high nutrient availability, are often compensated for by decreases in other parts of the root system. Furthermore, nutrient concentrations in plant tissues do not necessarily increase, and plant growth also often does not show an overall increase, suggesting that plants compensate for increased growth in one part of the root system in response to increased nutrient availability by decreasing growth elsewhere.

The response also depends on the plant’s tendency to produce herringbone versus dichotomous patterns (Fig. 8). The optimal response depends in part on the plant’s ability to allocate carbon to the root system (versus shoots, leaves, or reproductive structures), the size of an area of increased nutrients, and the chemistry (i.e., the diffusivity) of the nutrient at stake. Research by A. Fitter has shown that herringbone patterns are most efficient for mobile nutrient ions, but herringbone and dichotomous patterns are equivalent for immobile ions. Others have shown that herringbone patterns are more common in roots growing in infertile soils; the lower amount of branching allows the root system to explore a larger volume of soil. In fertile soils and in nutrient-rich microsites, branching patterns become more dichotomous; this pattern presumably permits more intensive and complete exploitation of the soil resource.

The heterogeneity of root systems in response to soil conditions reflects the high degree of heterogeneity in soil conditions at scales of centimeters to tens of meters. In agricultural fields, for example, chemical and physi-
FIGURE 7 The results of a classic experiment by Drew (1975, Comparison of the effects of a localized supply of phosphate, nitrate, ammonium and potassium on the growth of the seminal root system, and the shoot, in barley. New Phytol. 75, 479–490) in which the roots of barley plants grow through zones of nutrient-enriched soil. Control plants received all nutrients throughout the root system; the other plants received high concentrations of the target nutrient only in the middle of the root system.

Cal soil properties vary significantly between points that are only tens of centimeters apart, despite the apparent uniformity that might be expected from the nearly flat topography, plowing and planting of a single species of crop plant (Fig. 9). Extreme patchiness in one soil resource may not be paralleled by similar patchiness in other resources in the same site; the distribution of mobile, limiting elements like nitrate can vary greatly among points only 3 cm apart while immobile, nonlimiting elements (P, K) do not vary, even over large distances.

In natural communities, significant variation in nutrient concentrations and nutrient supply rates (e.g., nitrogen mineralization rates) have been documented in soils beneath different co-occurring species of grasses and trees. Spatial variation in soil properties also is generated by topographic variation. Even small changes in elevation or geomorphic position (e.g., hill crest, upper and lower slope positions, valley bottoms) are associated with differences in soil profile development, accumulation of organic matter, soil texture, the mineralization rates of nutrient elements, and gaseous losses through microbial processes such as denitrification. At scales of tens to hundreds of meters, differences in plant community composition are clearly correlated with differences in soil properties related to topography.

Spatial heterogeneity of soils is particularly strongly developed in plant communities in which either plants...
FIGURE 8 Alternative patterns of root growth. Dichotomous branching systems have more internal segments of the root system, compared to herringbone branching systems. The internal segments, or links, tend to be larger in diameter and therefore are more "expensive" to construct and maintain (in units of carbon fixed during photosynthesis). Plastic responses of roots to heterogeneous soil conditions involve changes in branching pattern, branch number and length, branching angle, root diameter and tissue density (observed as changes in specific root length and area, or length and area per unit root mass), and density of root hairs. Based on A. Fitter, in Waisel et al. (1991) and other papers.

are patchily distributed over otherwise barren ground (typical in many arid ecosystems) or in which differing life-forms (e.g., trees and grasses) are interspersed (e.g., savanna ecosystems). In deserts, the scattered plants form "islands of fertility" because of their ability to trap both plant litter and wind-blown fine soil particles. The soils within and beneath the scattered shrubs or tussock grasses in these systems have higher amounts of nitrogen, phosphorus, organic matter, silt or clay-sized particles, and moisture than the intervening bare areas. Similarly, patches of woody vegetation within grasslands accumulate higher amounts of nutrients and organic matter, again leading to patches of higher fertility within the grassland landscape. Plasticity of root system morphology and physiological capacity allows plants and communities to respond to this spatial variation in soil properties.

B. Nutrient Acquisition and Root Growth: Stand Scale

The response of individual plants to variation in soil properties manifests itself at the level of the community as changes in root biomass and the production and mortality rates of the roots. Many studies have shown that root biomass tends to be lower in more nutrient-rich sites, and experimental studies in which nutrients (particularly nitrogen) are added often confirm these observations. Similarly, it has been noted that plants typical of infertile sites have a higher ratio of root biomass to above-ground biomass than do plants from fertile sites. In a comprehensive review of available literature on nutrient cycling and root production, Vogt et al. (1986) found that the mass of fine roots in forest ecosystems is higher in forests with larger amounts of nitrogen in the forest floor, implying that the biomass

FIGURE 9 Spatial patterning of physical and biological soil properties in a soybean field in Michigan (United States). Units: (a) net N mineralization, mg N g⁻¹ d⁻¹; (b) microbial biomass, mg C g⁻¹ soil; (c) peak plant biomass, g m⁻²; (d) nitrate, mg N g⁻¹ soil; (e) elevation, m. From Robertson et al. (1997). Soil resources, microbial activity, and primary production across an agricultural ecosystem. Ecol. Applic. 7, 136–171.
of roots increases when nitrogen is sequestered in compounds that decompose slowly and accumulate in thick forest floors. This supports the idea that root biomass is greater in infertile than in fertile soils.

Rates of production and mortality of fine roots also respond to nutrient status, although experimental studies have documented both increases and decreases in response to nutrient additions. For example, Vogt et al. (1986) suggested that the rate of root turnover (input of new roots per year) is higher in forests with smaller amounts of nitrogen input in litterfall (an index of nitrogen limitation), but experimental addition to N to northern hardwood forests caused a decrease in mortality rates of fine roots. The contradictory data may reflect differences between the “normal” growth patterns of roots relative to the inherent fertility of the soil and the response of existing root systems to large, sudden changes in fertility induced by experimental nutrient additions.

### C. Competitive Interactions between Plants and Microbes

As the plant community and the soil microbiota require the same resources (nutrients and water), there may be competition between the microbiota and plants. Microbial demand for nitrogen is closely tied to the availability of degradable carbon, as carbon is often considered to be their most limiting resource. In this context, the likelihood of competitive interactions with plants depends on the C:N ratio of the available carbon resources: root exudates tend to have low C:N ratios (about 13:1), litter normally has high C:N ratios (>30:1, and >100:1 for woody plant tissues), and soil organic matter ratios are in the range of 10 to 20. Theoretical analyses have suggested that the relative mineralization versus immobilization of N derived from root exudates varies in relation to the relative availability and quality of the soil organic matter. Where “quality” refers to the metabolic capacity of the microflora to degrade the components of the organic matter. When N is limiting (i.e., when C:N ratios are high), microbes may take up mineral N from the same sources (soil solution, exchange surfaces) that plants use, resulting in competition.

Experimental additions of mineral N (NH₄⁺, NO₃⁻), to forest soils have shown that the microbiota captures a substantial proportion, often a majority, of the added N, but that nitrification rates are decreased when plants are present. These results suggest that while the microbiota competes successfully for mineral N in solution, plants can limit the availability of NH₄⁺ to the ammonia oxidizers (nitrifiers). However, long-term studies suggest that although microbes may outcompete plants for mineral N in the short term, the long-term activity of root systems, compared to the shorter, more localized bursts of activity of the microbes in response to changing temperature and moisture regimes, may allow plants to compete successfully. In this regard, the influence of plants on the soil moisture regime, through uptake of water and loss in transpiration, may indirectly affect competitive interactions for N by causing limitation of microbial activity due to soil drying. Furthermore, competition for N may be modified by differential use of the inorganic and organic forms of N in the soil.

### D. Litter Inputs to the Soil: Effects of Plant Tissue Chemistry

Perhaps the most widely studied and best-known interaction of plants and soil is the supply of organic matter to the soil through the deposition of litter. Litter deposition includes “fine litter,” the leaves, herbaceous stems, and smallest woody twigs, “coarse woody debris,” the large twigs, branches, and holes of shrubs and trees, “greenfall,” of leaves and leafy twigs clipped by herbivores such as squirrels, and root litter, the dead roots of all size classes. However, most measurements of ‘litter’ refer only to fine litter, as this is most easily measured. Litter is the largest source of organic matter supplied to the microbiota of the soil, and thus it fuels all of the microbial processes responsible for decomposition, nutrient cycling, and reduction-oxidation reactions of soil minerals and also is the ultimate source of the soil organic matter.

Litter inputs vary widely in quantity and timing. Tropical rain forests may have >2,000 g m⁻² yr⁻¹ litterfall, whereas tundra ecosystems may have <200 g m⁻² yr⁻¹. Deposition of fine litter is seasonal in temperate zones and continuous in the tropics, but the formation of coarse woody debris may be sporadic (i.e., large inputs following major storms, or infrequent inputs at any time of year). Amounts of coarse woody debris are comparable (about 120 to 3000 Mg m⁻² yr⁻¹) but much more patchily distributed. Greenfall production is also sporadic, as it reflects the population dynamics of the herbivorous insects or mammals that cause it. The timing of root death is quite variable among ecosystems; in some forests, root growth and death are pulsed, but in other sites, production and mortality may occur simultaneously and not in synchrony with seasons. Assuming that root populations are at steady state, so that root production rates can be used as an estimate of root litter deposition rates, the amounts of root litter...
production similarly range from roughly 100 g m⁻² yr⁻¹ to 1,700 g m⁻² yr⁻¹.

The ratio of above-ground to below-ground litter inputs is also quite variable among ecosystems (Table IV). Root litter inputs can be very high in grassland ecosystems as well, with ratios of below-ground : above-ground litter mass much greater than 1. The variability of the relationship between above-ground and below-ground litter production was clearly demonstrated by Raich and Nadelhoffer (1992), who found no statistically significant relationships between the amount of fine root production and the amount of above-ground litter production over a wide geographic range of locations and forest types.

As important as the quantity and timing of input of litter is, the chemical composition of the plant tissues is perhaps more important as the ultimate control on the activity of the soil microbiota. The ability of microbes to decompose the litter material depends on (a) the ratio of carbon to nitrogen and (b) the chemical forms in which the carbon occurs. Carbon occurs in plant tissues in several forms, including "nonstructural carbohydrates," which include soluble and membrane-bound contents of the cytoplasm and vacuoles. Cell walls contain cellulose (an unbranched chain of glucose residues), hemicelluloses (branched chains of glucose residues), and pectins; these materials are variably bound to lignin, a complex phenolic material that provides strength to the tissues. Because of its molecular structure, lignin is extremely resistant to decomposition;

### TABLE IV
Above- and Below-Ground Inputs of Organic Matter to the Soil in Forest Ecosystems

<table>
<thead>
<tr>
<th>Ecosystem</th>
<th>Root litter input (kg ha⁻¹ yr⁻¹)</th>
<th>Above-ground input (kg ha⁻¹ yr⁻¹)</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical broadleaf evergreen</td>
<td>n.d.</td>
<td>9.438 ± 1.104</td>
<td></td>
</tr>
<tr>
<td>Warm temperate deciduous</td>
<td>5.732 ± 1,070</td>
<td>9.390 ± 790</td>
<td>0.61</td>
</tr>
<tr>
<td>Warm temperate evergreen</td>
<td>9.053 ± 493</td>
<td>4.452 ± 234</td>
<td>2.04</td>
</tr>
<tr>
<td>Cold temperate deciduous</td>
<td>2.250 ± 920</td>
<td>3.854 ± 213</td>
<td>0.59</td>
</tr>
<tr>
<td>Cold temperate evergreen</td>
<td>6.152 ± 1,077</td>
<td>3.144 ± 194</td>
<td>2.93</td>
</tr>
<tr>
<td>Boreal evergreen</td>
<td>9.08 ± 208</td>
<td>2.428 ± 204</td>
<td>0.44</td>
</tr>
</tbody>
</table>

From Vogt et al. 1986, averages of values reported for several studies for each community type.

### TABLE V
An Example of the Role of Plant Tissue Chemistry in Regulating Decomposition Rates

<table>
<thead>
<tr>
<th>Plant</th>
<th>% N</th>
<th>% lignin</th>
<th>%C</th>
<th>%C : N</th>
<th>k</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>White spruce</td>
<td>9.11</td>
<td>89.4</td>
<td>47.0</td>
<td>-0.751</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Douglas fir</td>
<td>10.3</td>
<td>86.6</td>
<td>46.7</td>
<td>-0.890</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balsam poplar</td>
<td>13.6</td>
<td>76.9</td>
<td>44.6</td>
<td>-0.907</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspen</td>
<td>14.1</td>
<td>71.1</td>
<td>45.3</td>
<td>-1.331</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grass</td>
<td>13.2</td>
<td>56.3</td>
<td>45.8</td>
<td>-1.391</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Douglas fir</td>
<td>13.1</td>
<td>35.2</td>
<td>41.3</td>
<td>-1.706</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rose</td>
<td>3.4</td>
<td>38.7</td>
<td>44.3</td>
<td>-1.841</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cow-parsnip</td>
<td>7.0</td>
<td>30.3</td>
<td>39.6</td>
<td>-2.350</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Taylor et al. 1986. Nitrogen and lignin content as predictors of litter decay rates: a microcosm test. Ecology 70, 93–104. Tissues were allowed to decompose in laboratory microcosms for 4 months. Regression analyses of the data showed that the best predictors of mass loss were the C : N ratio, %N and lignin %, in that order. Nitrogen content was more important early in decomposition; lignin became more important as decomposition proceeded.
position rates, and in describing the relationship between decay rate and lignin:N as linear or nonlinear, almost all identify some subset of these characteristics as critical.

Decomposition not only produces CO₂, but also results in (a) the production of microbial metabolic products that are stable in the soil, and form the passive or long-term component of soil organic matter, and (b) the mineralization of nutrients such as nitrogen, sulfur, and phosphorus, which are bonded to carbon. Release of these nutrients into the soil solution is, again, a function of the availability of the nutrients in the plant litter substrate relative to the demand by the microbiota, as modified by the efficiency of assimilation of the microbes. Thus, the chemical composition of the plants that compose a community will affect the rates of accumulation of organic matter in the soil, as well as the rates of mineralization of limiting nutrients for plant growth. Alterations of plant community structure, through disturbance, succession, or management, can precipitate changes in nutrient cycling rates and carbon sequestration rates that create feedbacks to the plant community. For example, extensive browsing by moose on birch trees in boreal forests allows white spruce to grow up in their place; the switch from readily decomposable birch leaves to recalcitrant spruce needles causes decreases in the availability of nitrogen, which in turn favors the slow-growing spruce over nitrogen-demanding birch. This positive feedback maintains spruce-dominated forests until stand-destroying fires permit birch to become reestablished.

IV. LARGE-SCALE INTERACTIONS: STANDS AND BIOMES

The small and medium-scale interactions described earlier contribute to larger-scale patterns that reflect, on the one hand, the structure and composition of the whole plant community and, on the other hand, geographical effects on soil structure. These patterns may change over time, in response to successional changes in the plant community and in response to the geomorphological evolution of the landscape.

A. Geographic Patterns of Soils in Relation to Vegetation

In the classic formulation of Hans Jenny, the formation of soils reflects not only climate, bedrock geology, time, and local topography, but also vegetation. Over the broad geographic scale of biomes, regular patterns of soil characteristics are associated with patterns of plant community composition (Table VI). In polar regions, extreme climatic conditions constrain the development of vegetation; without organic matter inputs, microbial activity is low, weathering processes are extremely slow, and soils are barely more than bedrock material. In boreal and temperate forests, differences in the kinds of plant species result in differences in the rates of decomposition and accumulation of soil organic matter and the production of organic acids; these processes in turn affect the kinds of chemical weathering and pedogenic processes that occur. In grasslands, the high production rate of fine, nonwoody roots introduces both organic matter and the root-associated processes described earlier to depths of several meters; rapid decomposition and incorporation of surface litter combined with the pattern of root production create soils with very different properties than those of forest soils. Indeed, many studies have examined soils beneath forests that developed following long-term management of an area as grassland or cropland; rapid changes in the amounts, spatial distribution, and chemical quality of the soil organic matter cause changes in cation and metal chemistry associated with the changed chemistry of organic acid production and correlated changes in nitrogen availability and microbial activity.

B. Variations in Soils within Biomes: Effects of Species Composition

Within biomes, soil properties vary geographically with the species composition of the vegetation. The differential effects of particular tree species on soil properties result in parallel differences among stands when different species are dominant. These effects are clearly shown when single-species plantations are compared (Fig. 10). The effects of different plant species on soil acidity is particularly important, as changes in pH and acidity affect the mobility of cationic nutrients and can result in large losses through leaching. The differences are due both to differences in the amount of cations stored in the vegetation and also to differences in the loss of cations in water leaching through the profile. Cations lost through leaching may cause a permanent change in the quality of the soil, unless weathering rates are equal to loss rates.

Successional changes in plant communities may in part be driven by plant-caused changes in soil structure and chemistry. In many communities, high rates
TABLE VI

Regional Characteristics of Soils and Vegetation

<table>
<thead>
<tr>
<th>Vegetation</th>
<th>Climate</th>
<th>Soil order</th>
<th>Speed of Pedogenesis</th>
<th>Weathering processes</th>
<th>Horizons present</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold desert</td>
<td>Extremely dry,</td>
<td>Entisols</td>
<td>Extremely slow</td>
<td>Precipitation of carbonates only</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tundra</td>
<td>Dry, cold</td>
<td>Inceptisols</td>
<td>Slow</td>
<td>Organic acid formation clay</td>
<td>Thin O, A, B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>weathering Fe-hydroxides</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boreal forest</td>
<td>Very moist, cold</td>
<td>Spodosols</td>
<td>Slow to moderate</td>
<td>Organic acid leaching Fe-hydroxides clay weathering</td>
<td>Thin O, A, leached E, differentiated B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temperate forest</td>
<td>Moist, mesic</td>
<td>Inceptisols</td>
<td>Moderate</td>
<td>Organic acid formation clay</td>
<td>Thin O, A, variable E, deep, differentiated B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>allisols</td>
<td></td>
<td>weathering Fe-hydroxides</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grassyland</td>
<td>Moist to dry</td>
<td>Mollisols</td>
<td>Moderate</td>
<td>Nonmobile organic acids,</td>
<td>O absent, deep A, thin B, little differentiation</td>
</tr>
<tr>
<td></td>
<td>mesic</td>
<td></td>
<td></td>
<td>CaCO₃ precipitation, Fe-hydroxides</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Savannah</td>
<td>Moist to dry</td>
<td>Vertisols,</td>
<td>Fast</td>
<td>Nonmobile organic acids, clay</td>
<td>No O, deep A, differentiated, very deep B</td>
</tr>
<tr>
<td></td>
<td>warm</td>
<td>Unisols</td>
<td></td>
<td>weathering, variable CaCO₃,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fe-OH formation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Desert</td>
<td>Dry warm</td>
<td>Aridisols</td>
<td>Slow</td>
<td>CaCO₃ precipitation Fe-OH formation</td>
<td>No O, A, shallow B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tropical rainforest</td>
<td>Wet, very warm</td>
<td>Oxisols</td>
<td>Very fast</td>
<td>Mobile organic acids Fe, Al-OH</td>
<td>Thin O, A thin E, very deep, differentiated B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>formation strong clay weathering</td>
<td></td>
</tr>
</tbody>
</table>


of nitrogen cycling (mineralization and nitrification) are observed early in succession, in association with herb- or shrub-dominated communities, but these rates decrease markedly as trees come to dominate the community. These decreases in nitrogen cycling rates are associated with changes in litter chemistry (more recalcitrant material, lower nutrient concentrations) and in root quantities and distributions. A good example of this process is demonstrated by boreal ecosystems. Early in succession, N-fixing alders (Alnus spp.) in shrubby communities support high rates of nitrogen mineralization and nitrification; these rates decline rapidly when the alders are replaced by poplars (Populus balsamifera) and white spruce (Picea glauca). The decline in nitrogen cycling rates that accompanies the successional shift from alders to poplars is driven not only by differences in the N content of the litter, but also by the presence of high concentrations of tannins and phenolics in the spruce needles. The tannins inhibit microbial activity, while the phenolic compounds are used as a carbon source by the microbiota; however, both compounds result in N immobilization.

Invasions of trees into grasslands are associated with a wide range of changes in soil properties. For example, water content of the soils changes due to (a) differences in inputs (stem flow and drip from canopy leaves accentuate inputs, thick litter layers promote infiltration, but high rates of interception by the canopy can reduce total precipitation reaching the surface, and interception within the litter layer can prevent moisture from reaching mineral soil), (b) differences in soil porosity (large pores and channels associated with the growth of woody roots, and the growth of such roots to greater depths than the grasses), and (c) differences in losses (higher transpiration rates by trees, but shading decreases soil temperature and therefore evaporation rates, as does the accumulation of thick litter layers). Large withdrawals of cationic nutrients occur, which are sequestered in the larger amounts of perennial biomass of woody plants. Nitrogen becomes increasingly unavailable, as recalcitrant, lignin-rich litter accumulates. Finally, greater acidity of leaf litter tissue, greater production of organic acids from litter, and protons released through cation and ammonium uptake all drive acidification of the soil profile, with consequent changes in both leaching loss of cations and weathering rates.

Just as changes in species composition may drive changes in soil properties, soil properties can affect successional patterns. Fertile soils are thought to support faster rates of succession (i.e., rates of species replacement), but with lower plant diversity than infertile soils. However, a wide variety of patterns have been observed, and many of the studies involve applications of nutrient fertilizer to experimental plots. Additions of nutrients have promoted the growth of early successional annual species, or the persistence of perennial herbs at the expense of woody pioneers, but opposite results have also been observed. Successional processes have also been related to the mechanism of competition among plant species for soil resources. Tilman (1988) has shown that competitive ability is related to the capacity of a plant to obtain nutrients from the soil solution and thereby reduce concentrations below the minimum needed for competitors; this ability is balanced by the need to compete for light, so that the outcome of competition (and thereby the composition of a community) reflects the tradeoffs in competing for different resources. Plant growing on infertile soils tend to have lower nutrient requirements (low uptake rates but compensating lower growth rates) and higher rates of nutrient conservation (higher efficiency of nutrient use) compared to plants typically found on fertile soils. The tradeoffs among characteristics promoting competitive success on infertile soils (strong competition for nutrients) versus success on fertile soils (competition for light being more important) ultimately determine the successional process on a given site.

Clearly, as soil properties themselves are dynamic, and change with changes in species composition, interplant interactions such as competition will change over successional time as the soil-plant system itself evolves. Such a situation has been observed in the Netherlands, where high rates of atmospheric nitrogen deposition has driven a successional change from heath-dominated (Erica tetralix) communities to grass-dominated communities (Molinia caerulea). E. tetralix is characterized by low growth rates, low rates of biomass loss to shed tissues, herbivory, and so on, and poorly decomposable litter, which supports low rates of mineralization but high rates of sequestration in thick organic horizons. In contrast, Molinia is characterized by high growth rates, high rates of tissue loss to shedding and herbivory, and rapidly decomposable litter. As expected, in experiments specifically testing competitive ability, Erica is a superior competitor when nutrient supplies are low, whereas Molinia is the superior competitor when nutrient supplies are high. Thus, the feedback relationships between plants and soils can drive successional changes in community composition.
V. IMPLICATIONS FOR THE MANAGEMENT AND CONSERVATION OF ECOSYSTEMS

Clearly, the complex, multiscale interactions of plants and soil affect all aspects of both the biology of plants and the properties of soils. This implies that both purposeful and inadvertent human effects on plant-soil systems will have diverse, perhaps unexpected ramifications, as changes in soils affect plants and vice versa. Following are brief discussions of four major areas of current concern, in which plant-soil interactions are likely to play a major role.

A. Forestry, Agriculture, and Ecosystem Restoration: Human Creation of Ecosystems

Humans create plant communities for a wide variety of purposes, from the production of food, fiber, and fuels to the restoration of “natural” ecosystems and the creation of beauty for its own sake (in gardens and parks). The choice of plant species is often conditioned by soil properties, certain crops cannot be grown in certain soils, or conversely particular soils may be ideally suited to particular crops. Conversely, the choice of crop plant or tree can alter soil properties, in some cases perhaps irreversibly. As described earlier, the substitution of coniferous trees for deciduous trees, a widespread practice because of the rapid growth and high economic value of many conifers, results in acidification, the loss of nutrient cations through leaching, and the creation of thick litter and humus layers. In some cases, foresters use mixtures of tree species that are specifically designed to take advantage of differences in plant-soil interactions among the species. For example, larch (Larix spp.) and lodgepole pine (Pinus contorta) are often interplanted with Sitka spruce in plantations in the British Isles, because larch and pine promote higher rates of nitrogen cycling and therefore faster growth of the spruce. The increased nitrogen availability results, in part, from more widespread roots of the interplanted pines or larches, which reduce waterlogging in the organic horizon and thereby stimulate mineralization. Nitrogen-fixing plants are often introduced into both agricultural and forest plantations in order to alter the cycling of N through the soil. Ecosystem restoration similarly involves the deliberate establishment of particular plant species and species mixtures. The choice of species is most commonly made on the basis of replicating the “natural” community, rather than explicitly with respect to soil properties. However, plant communities are sometimes specifically designed to alter soil properties: the introduction of fast-growing grasses to introduce organic matter in barren soils, and the parallel introduction of N-fixing species into such soils, the use of nutrient-demanding, fast-growing species to remove nutrients (through repeated harvest) from overfertilized fields being restored to native grasslands or from wetlands being used for wastewater treatment are examples. Conversely, soil properties may unexpectedly affect the course and success of restorations. The potential for such effects is clearly illustrated in efforts to restore coastal wetlands in San Diego, California. Coarse-textured sediments were used to construct the wetland, which were unlike the fine-textured sediments found in naturally occurring marshes. Although the introduced salt marsh plants survived and spread, they only attained about half the height of plants in undisturbed marshes; the cause of the problem was the low storage rate of organic matter in the coarse sediments and the concomitant low intrinsic supply rate of nitrogen from the sediments.

B. Exotic Species Invasions—Inadvertent Human-Caused Changes in Species Composition

The spread of exotic plant species around the world is recognized as one of the major threats to biodiversity. Plant communities are being altered through invasions of species that in some cases become part of the community, but in other cases exclude and eliminate the native species. These invasions often cause economic problems, as well as challenges for conservation of native communities, as the invading species degrade grazing lands, choke stream channels, or promote wildfires. These changes in plant community composition are likely to alter soil properties, especially when the exotic species is of a different growth form than the natives or when its litter chemistry is different. Recent studies have shown that, for example, invasions of cheat grass (Bromus tectorum) into desert grasslands alters the abundance and species composition of fungi, protozoa, nematodes, and soil invertebrate communities, decreases the abundance of N-fixing lichens, and thereby decreases the amount of nitrogen in the soil. Invasions of exotic shrubs (Berberis thunbergii) and grasses (Microstegium vimineum) into deciduous forests stimulates increases in nitrification and soil pH. Invasions of the hawkweeds (Hieracium spp.) into pastures in New...
Zealand causes the accumulation of nitrogen, but decreases soil pH. Inversions of grasses (Melitus minuti- flaora) into Hawaiian shrub lands stimulates increases in net nitrogen mineralization rates. It is likely that such changes are widespread and may affect the ability of resource managers to remove the exotics and restore the native plant communities.

C. Nutrient Pollution—Inadvertent Human-Caused Changes in Soil Properties

The profound and far-reaching changes in atmospheric chemistry that are accompanying the spread of industrial society around the world include greatly increased inputs of several nutrients to soil-plant systems, through the deposition of particles and the dissolution of soluble ions in precipitation. The most important changes include greatly elevated deposition of nitrogen and sulfur. As these are both important plant nutrients and as their chemistry is intimately linked with other chemical and biological processes within the soil, a cascade of interactive changes to the soil-plant system are taking place in many places around the world.

Deposition of sulfates at rates below the level that directly injures plants still causes problems through soil acidification. Leaching of sulfate in soil water carries with it nutrient cations, and in soils with low base saturation and a parent material that supplies few cations through weathering, cations can become seriously depleted. Calcium, an essential element for the integrity of root function, is particularly strongly affected. Hardwood forests in New Hampshire (United States) that have received large inputs of sulfate for several decades have prematurely stopped growing, and their tissues show abnormally low concentrations of calcium. Depletion of nutrient cations is accompanied by increases in the concentration of aluminum ions in the soil solution and on exchange sites, and the aluminum can be directly toxic to plants. Depletion of calcium also leads to increased sensitivity to frost damage of needles, due to change in the structure of cell membranes, as has been demonstrated for declining red spruce in the eastern United States.

Nitrogen deposition to ecosystems is now occurring at double the estimated rates prior to the industrial revolution. The nitrogen originates from agriculture (applications of fertilizers and manures, and the planting of N-fixing crops), combustion (automobiles, power plants, and also forest burning), and waste disposal (e.g., sewage). Although nitrogen is considered the most important limiting resource in most terrestrial ecosystems, an excess does not simply alleviate stress on plants. As was discussed earlier, plants growing on nutrient-poor soils have a variety of characteristics, including patterns of both root and above-ground morphology and physiology, characteristics of life history, and intrinsic growth rates, which adapt them to such habitats and which contrast strongly with the characteristics of plants found on nutrient-rich soils. Added nutrients often cannot be absorbed or assimilated by such plants, nor can they respond with increased growth. Rather, species adapted to nutrient-rich soils become capable of competitively displacing these species. Thus, many ecosystems that had been subject to limiting supplies of nitrogen have undergone profound changes in species composition. The substitution of grass for heather in European heathlands, as discussed earlier, is a case in point. Similar changes in native plant communities are occurring elsewhere in the industrialized world. And because communities found on infertile soils tend to have high diversity of plants, including many rare species, the increasing dominance of nitrophilous species may threaten many of them.

Nitrogen additions to soil not only affect the plant community through competitive displacements and substitutions, but also affect the internal nitrogen dynamics of the soil. Evidence from a variety of ecosystems has shown that N additions stimulates intrinsic mineralization processes, thus accentuating the effects of added N. Moreover, changes in the microbiota under conditions of high N inputs can result in the development of acid-tolerant nitrifiers; the increased levels of NO3–N. Not only cause changes in plant community composition, but cause leaching of cations and thus changes in soil chemistry. As plant community composition changes, the chemistry, amount, and timing of litter inputs will change, as will all other components of the plant-soil system. The long-term effects of this chain of linked changes can only be guessed.

D. Changing CO2 in the Atmosphere—Inadvertent Human-Caused Changes to the Soil-Plant System

The rising concentration of CO2 in the atmosphere is expected to affect virtually all aspects of the plant-soil system. The topic can be touched on only briefly here, simply to note the major processes and interactions that are anticipated. Plants are expected to respond to increasing CO2 by increasing their overall rate of production (particularly C3 plants), increasing the allocation of carbon to below ground (including root growth, root turnover rates, root respiration rates, and
rhizodeposition rates). Water use efficiency and nutrient uptake rates also often increase in experimentally enriched plants. Changes in litter chemistry (increased C:N ratios, increases in lignin, phenolics, or tannins) are anticipated but have yet to be clearly demonstrated. However, individual species responses have proven to be highly variable; studies of the growth response of a wide range of species have shown that changes in growth rate can range from a 25% decrease to a 200% increase. However, it is clear that virtually all aspects of plant morphology and chemistry that have been shown to affect soil properties are likely to change, precipitating changes in the soil in accordance with the processes discussed in this chapter. Figure 11 illustrates the complexity of the possible responses. Although a large amount of research is currently being conducted on these linkages, the complexity of the problem and the variability in response among different plants and in different soils make the discovery of general patterns and predictability difficult and yet to be achieved.

E. Biodiversity and Ecosystem Function

The effects of changing plant community diversity on the structure and function of soil are among the most important current topics of research. Several studies utilizing experimental grasslands have shown that such ecosystem processes as nitrogen mineralization, carbon accumulation and mineralization, nitrogen concentrations, soil respiration, and litter decomposition rates are altered (often decreased) by decreases in plant diversity, but other studies, including studies in forested ecosystems, have presented contradictory results. In many cases, particular species or particular classes of species (for example, nitrogen-fixers) have large effects. In addition, responses of soil processes to changes in plant diversity often are largest in comparing small numbers of species; once the plant community reaches a moderate number of species (three or four), there is no further change as plant species diversity increases. Reciprocal effects of diversity within the soil microbiota and soil fauna on the plant community are largely unknown. The effects of changes in diversity within the plant community to the biology, chemistry and physics of soils, and vice versa, will be among the most important questions to be addressed in the future.

See Also the Following Articles

FOREST ECOLOGY • GREENHOUSE EFFECT • PLANT INVASIONS • RESTORATION OF BIODIVERSITY • SOIL BIOTA, SYSTEMS AND PROCESSES • SOIL CONSERVATION
Bibliography


PLANT SOURCES OF DRUGS AND CHEMICALS

William H. Gerwick, Brian Marquez, Ken Milligan, Lik Tong Tan, and Thomas Williamson
Oregon State University

I. Introduction
II. Examples from Marine Plants Illustrating Plant Sources of Drugs and Chemicals
III. Examples from Terrestrial Plants Illustrating Plant Sources of Drugs and Chemicals
IV. Conclusions

GLOSSARY

anticancer agent A chemical substance that is capable of bringing about a remission or cure of the family of diseases known as cancer.
biosynthesis The biochemical process by which a plant produces molecules of utility or adaptation.
cytotoxicity The ability of a chemical compound to kill cells in an experimental system, such as in a petri dish.
secondary metabolite A chemical substance produced for a reason other than basic life requirements and with some adaptive or defensive value.

NATURAL PRODUCTS from plants have formed the basis for many of our useful pharmaceuticals and agricultural chemicals. This chapter presents examples of useful plant-derived pharmaceutical agents, which illustrate the complexity of the drug discovery process and the importance of maintaining biological diversity in order to preserve this inherent chemical diversity and its genetic origins.

I. INTRODUCTION

The intent of this chapter is to illustrate by example the enormous diversity of plant-derived natural products and their tremendous importance to human society. It is precisely this diversity of structure that has been, and continues to be, of such incredible value to the pharmaceutical and agricultural industries. Ultimately, this chemical diversity is rooted in the inherent biodiversity of our planet. Sadly, space limitations for this chapter relegate this treatment to only a sampling of interesting examples. These have been chosen to exemplify several concepts; various chemical adaptations typical of particular plant groups, classes of compounds of utility to society, older examples deriving from ethnobotanical information, and modern examples resulting from methodical broad-based screening programs.

It cannot be debated that the premier importance of plants to human existence is as a source of food and oxygen. Related to their serving a food function is the use of plant-derived materials as food additives to impart desirable taste or textural properties. Examples include the use of a class of biopolymer from red marine algae, the carrageenans, to stabilize emulsions in such...
diverse products as ice cream and beer, and the use of the sugar alcohol sorbitol, a product of the mountain ash Sorbus aucuparia, as a noncaricogenic sweetener with humectant (wetting) properties in toothpaste. However, this chapter will focus on another property of plants; their ability to produce unique molecular entities with potent pharmacological effects in mammalian systems. Why do plants make such compounds? Given their complicated, often exotic, structures, they certainly expend considerable biochemical energy in their production. While some hold that such compounds are vestigial in nature (previously served a function that is no longer in evidence), there is excellent experimental evidence to support the idea that most, if not all, are produced with powerful adaptive functions, such as defense against potential predators and inhibition of the growth of competing species.

Primitive peoples throughout the world make use of their indigenous flora as a source of medicines. Through the process of trial and error, these cultures have examined and discovered many plants that produce unique molecular entities with valuable biological properties. The field of "ethnopharmacology" or "ethnomedicine" seeks to obtain new pharmaceutical leads from a study of the native medicines of these primitive peoples, an approach that has been highly successful and resulted in such compounds as aspirin from willow bark (Salix), the antitumor alkaloid vincristine from the Madagascar periwinkle (Catharanthus), and the antimalarial sesquiterpene artemisinin (qinghaosu) from the Chinese herb Artemisia annua.

Beginning in the 1950s, comprehensive evaluations of the unique constituents of plants were undertaken. In the United States, these earliest efforts were coordinated largely by the National Cancer Institute, and so necessarily had a focus on anticancer properties. Assays for potential antitumor activity at that time were run in intact animals (in vivo evaluations). While even the most modern drug developments require evaluation of a drug for efficacy and toxicity in animal systems, this usually occurs only after a candidate pharmaceutical has shown the requisite property in isolated biochemical and cell-based screens. Consequently, current in vivo evaluations only occur on a very small fraction of the most promising candidate pharmaceuticals. In contrast, the initial screen of plant extracts in early efforts used in vivo techniques; consequently, they were very slow, expensive, and, perhaps worse, used test animals in large quantities.

In the late 1970s, and gaining broad acceptance in the 1980s, a major shift away from whole animal primary screening was achieved with the introduction of cell-based screening. In the cancer effort, this largely had the endpoint of cytotoxicity to cancer cells grown in small petri dishes or, later on, 96-well trays. Ultimate expression of this idea has been realized at the National Cancer Institute wherein extracts and compounds are evaluated for their level and profile of toxicity to 60 different cancer cell lines. The cell lines have been chosen so as to represent cancers affecting nine different organ types and have been quite useful in detecting new cytotoxins as well as giving information about their molecular mechanism of action.

The past 10 years has seen a growing departure from even cell-based assays. Over the past 30 years, there have been tremendous advances in understanding at a molecular level the causes of many diseases, including cancer (Shu, 1998). This has translated into assays that screen for compounds or extracts that interfere with a specific enzymatic reaction or protein-protein interaction that has been shown to underlie a particular human disease. With the reduction in screening format from whole animal to cell to isolated protein, it is now possible to screen hundreds of thousands of compounds or extracts in a few weeks time using high throughput screening (HTS) technologies. Here, a target protein is raised in large quantities by molecular biological means in a surrogate organism, placed in a 96-well or even larger formats, and compounds or extracts evaluated for inhibitory effects using robotics that work around the clock. It is hoped that screens of this design and scope will uncover a new generation of pharmaceuticals that will have highly specific actions, targeting the underlying causes of disease, with very high potency, and leaving normal or undiseased cells untouched.

II. EXAMPLES FROM MARINE PLANTS ILLUSTRATING PLANT SOURCES OF DRUGS AND CHEMICALS

A. Anticancer Agents from Marine Cyanophyta (Cyanobacteria)

The marine "plants" generally include a number of groups, including microscopic as well as macroscopic forms. The major subdivisions of the microscopic forms include dinoflagellates, cryptophytes, chrysophytes, cyanophyta (= cyanobacteria), and prochlorophyta. The major groups of macroscopic marine plants include the Rhodophyta (= red algae), Chlorophyta (= green algae), Phaeophyta (= brown algae), and a number of marine angiosperms (sea grasses); the latter represent...
terrestrial species that have “readapted” to life in the ocean. While interesting and significant molecules have been isolated from all of these groups, secondary metabolites of special note have been obtained from marine cyanobacteria (Jaspar and Lawton, 1998), dinoflagellates, and red and green macrophytes. Examples from these four groups follow.

1. Curacin A and “Chemotype” Concept

Cyanobacteria take on many different physical forms in the oceans, from gelatinous encrustations and tufts of 20 cm hairlike filaments, to microscopic unicells free in the water column. One species of pan-tropical distribution, *Lyngbya majuscula* (family Oscillatoriae), has been particularly plentiful in its production of structurally unique and biologically important secondary metabolites. Such an example is given by a collection made in Curacao, in the southern Caribbean, which was reported to possess an extract with fish, snail, brine shrimp, and cancer cell toxicity (Orjala et al., 1996). Bioassay guided isolation efforts led to the isolation of several structurally unrelated substances, which were each responsible for only some of the observed biological properties. The cancer cell and brine shrimp toxicity was principally due to a unique thiazoline-containing lipid, named curacin A (Fig. 1). Its mechanism of cytotoxicity was shown to involve inhibition of microtubule formation, crippling the ability of cells to properly segregate chromosomes at mitosis (Gerwick et al., 1994). Several very valuable anticancer agents that are in common use in the clinic today work by this same essential mechanism (e.g., see discussion for taxol, presented later). Curacin A has been shown to interact with microtubules at the same site as the antigout drug, colchicine. Unfortunately, curacin A was found difficult to work with in vivo because of solubility and stability problems. Recent efforts have overcome at least some of these problems, and work continues to develop an anticancer drug patterned on this structural concept.

![Figure 1: Structure of the antimitotic lipid “curacin A,” obtained from a Caribbean collection of the blue-green alga (cyanobacterium) *Lyngbya majuscula*](image1)

Sources of curacin A for continued drug testing have largely been from materials freshly collected in Curacao, although the living alga does produce curacin A in laboratory culture. Interestingly, when collections of *Lyngbya majuscula* were made from approximately 20 sites along the leeward coast of Curacao, only two chemotypes, both quite limited in extent, were found to make curacin A. If continued investigation of this phenomenon substantiates these findings, then a very intriguing conclusion must be drawn, which is of enormous significance to “biodiversity preservation.” It is widely held that in order to preserve the valuable genomic potential of a particular “species,” it is sufficient to preserve a few individuals or a limited population. This work with Curacao collections of *L. majuscula* suggests that in order to accomplish the larger goal, it is actually necessary to consider preservation of the many, almost innumerable, “chemotypes” of a given species.

2. Cryptophycin

Another very promising anticancer treatment derives from the freshwater cyanobacterium *Nostoc* sp. While one of the active compounds, cryptophycin A, was first isolated as an antifungal substance, it was only several years later that, upon re-isolation, its anticancer potential was realized (Fig. 2). In common with curacin A, cryptophycin also blocks microtubule assembly processes, although at a different drug-binding site (the vinca alkaloid site). In animal testing, cryptophycin has effected actual “cures” of some tumor types. An enormous effort has been devoted to the total chemical synthesis of cryptophycin and several hundred analogs with the intent of ensuring that (a) a ready supply of the drug is available for clinical testing and (b) that advanced cancer treatment evaluations utilize the most effective molecule in this drug series. Testing of the drug in humans has recently advanced to phase II trials, and there is great hope that this agent will provide an important new tool in cancer treatment.

![Figure 2: Cyclic peptide structure of the anticancer metabolite “cryptophycin.”](image2)
B. Toxins of Impact and Utility from Dinoflagellates

Dinoflagellates represent a very diverse group of organisms, made up of both freshwater and marine varieties. Dinoflagellates are protists, which are not strictly considered plants or animals although they share characteristics of both. Some dinoflagellates are photosynthetic whereas others are parasitic to fish or to other protists. They can be very small or as big as 2 mm in diameter, such as in the case of Noctiluca sp. A curious feature of some dinoflagellates is that they are bioluminescent.

Dinoflagellate blooms in the world's oceans have had tremendous impact on human society; despite this importance, many aspects of these organism's chemistry and biology are poorly understood. Hundreds of people suffer from the effects of seafood poisoning and millions of fish and other marine life are killed each year as a direct consequence of large blooms of dinoflagellates. Some of the coastal species can “bloom” during the warmer months and can even make the water appear red or golden colored, hence, the term “red tide.” For many years seafood poisonings and fish kills have been associated with these dinoflagellate blooms but it is only within the past decade that the chemistry behind these effects has been unraveled.

1. Ciguatoxin and Maitotoxin

A prime example of a dinoflagellate toxin with huge impact on society is given by Ciguatera Seafood Poisoning (CSP), which is quite prevalent in tropical areas around the globe (Yasumoto and Murata, 1993). CSP is more common than any other illness associated with consumption of tainted seafood. Generally, the poisoning results from ingestion of coral reef fish that have accumulated the toxins through their diet. Symptoms include, but are not limited to, memory loss, joint pain, miosis, erethism, cyanosis, and prostration. One of the more interesting symptoms associated with ciguatera is a neurological disturbance leading to reversed hot and cold sensations. Most of the toxic effects of CSP are attributed to ciguatoxin, which is a metabolically oxidized derivative of a substance produced by the dinoflagellate, Gambierdiscus toxicus. The structure of ciguatoxin was deduced after a monumental 15-year effort (Barker, 1997), there is much controversy surrounding research on this topic. This particular organism is different from the other known species of dinoflagellates in that it may have a stage in its life cycle where it becomes dormant but, in the presence of large numbers of finfish, can become active and in effect ambush the fish as a predator. While research into this behavior is still in the early stages, all evidence gathered thus far seems to support this hypothesis. Less is known about the toxins produced by this organism except that another toxin that should be noted here is maitotoxin (Fig. 4). Probably produced by dinoflagellates related to those that produce the ciguatoxin precursor, this compound is even more toxic than ciguatoxin. It is by far the largest compound ever isolated that is not a biopolymer (e.g., protein or carbohydrate). Its structure was also elucidated by the laboratories of Yasumoto and Murata.

2. Gymnodinium breve

For many years it was thought that seafood poisonings caused by dinoflagellates were limited to tropical regions of the world. However, in 1987, 14 dead hump-backed whales were discovered in Cape Cod Bay, Massachusetts. Examination of the animals showed that they had been well until just before their deaths, appearing to be well nourished and showing no signs of ill health. A few months later, fishermen and beach-goers along the North Carolina coast began to complain of respiratory problems and eye irritation. Soon after, seafood consumers began to complain of diarrhea and dizziness. In later years and on a periodic basis, these same symptoms have been reported all along the east coast and the Gulf of Mexico. Most of these outbreaks have been connected with the dinoflagellate Gymnodinium breve. This dinoflagellate frequently forms large blooms, or “red tides,” off the Florida coast, which result in many tons of dead finfish washing up on the beaches. This, in turn, repels many tourists and has forced many fishers to be closed. In the early 1980s the toxic constituents produced by this organism were found to be brevetoxin A and B, along with other structurally related compounds (Fig. 5).

3. Pfiesteria piscicida

A recent development in the economic and environmental impact of dinoflagellate blooms is the newly discovered dinoflagellate, Pfiesteria piscicida. Outbreaks of this organism have been reported in coastal North Carolina and Maryland. Because these blooms are possibly linked to pollution from swine and chicken farms, plus the interest generated by the publication of a popular book (Barker, 1997), there is much controversy surrounding research on this topic. This particular organism is different from the other known species of dinoflagellates in that it may have a stage in its life cycle where it becomes dormant but, in the presence of large numbers of finfish, can become active and in effect ambush the fish as a predator. While research into this behavior is still in the early stages, all evidence gathered thus far seems to support this hypothesis. Less is known about the toxins produced by this organism except that
FIGURE 3 Polyether structure of ciguatoxin, the active poison in Ciguatera Shellfish Poisoning.

Ciguatoxin $R_1 = -\text{CH(OH)}-\text{CH}_2\text{OH}; R_2 = \text{OH}$
Pre-ciguatoxin (CTX-4B) $R_1 = -\text{CH}_2\text{CH}_2; R_2 = \text{H}$

FIGURE 4 Structure of maitotoxin, the most complex nonpeptide structure ever determined.

Brevetoxin A

FIGURE 5 Brevetoxin A, the causative fish poison produced by the red tide organism Gymnodinium breve.
there appears to be two: one water soluble and one soluble only in organic solvents. The symptoms resulting from contact with this dinoflagellate are quite severe, ranging from serious memory loss, loss of motor function, to death.

C. Kainic Acid and Domoic Acid, Neurotoxic Anthelminthics from Red Algae

Two Japanese red algae, *Digenea simplex* and *Chondria armata*, have been employed for more than 1000 years in Japan for their potent anthelminthic properties; that is eliminating intestinal worms, such as parasitic roundworms (*Ascaris lumbricoides*), whip worms (*Trichuris trichiura*), and tape worms (*Taenia spp.*). Two closely related compounds, domoic acid and kainic acid (Fig. 6), have been isolated from these red algae and are responsible for these healing effects. These compounds were traditionally ingested by consumption of the producing algae, their names deriving from the Japanese names for these seaweeds, *domoi* and *kaininso*. Kainic acid has become an important tool in neurobiological research because of its potent neurotoxic effects on neurons (Brown and Nijjar, 1995). On the other hand, domoic acid has been identified as the active poison in amnesiac shellfish poisoning (ASP), which has recently struck both east and west shores of North America. The levels of the drug that are ingested for anthelminthic properties are relatively small when compared to the amount of compound observed in the recent outbreaks of ASP. However, the macrophytic red algae were determined to not be the culprit in the recent ASP outbreaks; rather, a planktonic pennate diatom, *Nitzschia pungens* f. *multiseries*, now renamed *Pseudonitzschia australis*, is the producer in this latter case.

The mechanism of action of these compounds is very similar, as is suggested by their closely related chemical structures. Both are analogs of glutamic acid, a well-known neurotransmitter found in the brain of mammals. The algal-derived substances are known as excitotoxic amino acids as they can stimulate neurons through the release of endogenous glutamate. The release of high doses of L-glutamate causes damage or death to various neuronal cell types. Domoic acid is two to three times more potent than kainic acid and 100 times more potent than L-glutamate. Although these compounds (domoic and kainic acid) are environmental neurotoxins to be avoided, they have shown considerable utility in biomedical research. For example, the resulting extensive neuronal loss seen with the appropriate dose of kainic acid is very similar to that observed for Huntington’s disease and hence provides an excellent model for the study of this disease in humans. Kainic acid has also been used in the study of epilepsy.

A problem of epidemic proportion that occasionally arises from these excitotoxic amino acids is amnesic shellfish poisoning (ASP) (Todd, 1993). It has been shown that blooms of the planktonic diatom *P. australis* result in a buildup of domoic acid in shellfish. Three highly publicized ASP outbreaks have occurred in the past 10 years in North America. The first was caused by the ingestion of mussels containing high levels of domoic acid from Prince Edward Island, Canada. The second was a large pelican kill in Monterey Bay, California. The domoic acid was found to accumulate through the food chain from diatoms to small fish to anchovies to pelicans. The last reported outbreak was in razor clams and Dungeness crabs in Oregon and Washington. The discovery of high concentrations of domoic acid led to the shutdown of these seafood harvests for several years, causing substantial economic hardship for al-

![Glutamic acid, Kainic acid, Domoic acid](image-url)
As a result of these ASP outbreaks, strict guidelines have been developed for the maximum allowed domoic acid content in harvested shellfish (20 ppm).

D. Green Algal Metabolites

Coral reef environments are habitats of high speciation and biodiversity ranging from microorganisms to top predators. This biodiversity inherently creates intense competition between species, including predator-prey relationships such as between herbivores and their algal diets. Methods that algae use to protect against predation include calcification, production of spiny appendages, and the biosynthesis of chemically noxious deterrent molecules that are either toxic or deter feeding.

The green alga Halimeda sp. employs both calcification and chemical defense strategies. As a result, Halimeda species constitute the largest macroalgal biomass in many tropical reef systems. In feeding preference studies, and analyses of reef fish stomach contents, Halimeda spp. have been shown to not be a food source for generalist herbivores. While calcification can provide algae with a strong structural protection against potential predators, some herbivores have coevolved specialized feeding structures that allow them to macerate even very tough materials. However, the dominance of Halimeda spp. in reef systems is enhanced through its production of a potent defensive and antifouling substance, a diterpenoid natural product named halimedatrial (Fig. 7). Halimedatrial displays a broad spectrum of activity against marine bacteria, a marine fungus, the division of fertilized sea urchin eggs, and the motility of sea urchin sperm (Paul and Fenical, 1988). Halimedatrial was also toxic to common reef fish as well as being inhibitory of reef fish feeding at ecologically relevant doses.

In the alga, halimedatrial exists in a “pro-toxin” form, a protected tetraacetate named halimedatetraacetate. Field experiments with Halimeda sp. have shown that halimedatetraacetate is converted to halimedatrial upon wounding of the alga. This conversion is rapid and specific to the area of herbivory or wounding and does not take place simply upon introduction to air but is an enzymatic conversion of a “protoxin” secondary metabolite to a highly potent antipreditory substance. This activated defense of Halimeda represents the first to be described in the marine environment. Interestingly, a similar activated defense system that produces feeding deterrents has previously been described in Russulacean mushrooms.

III. EXAMPLES FROM TERRESTRIAL PLANTS ILLUSTRATING PLANT SOURCES OF DRUGS AND CHEMICALS

A. Anticancer Agents from Higher Plants

1. Taxol

The story of the discovery and development of taxol as an anticancer agent is a superb example of the opportunities and problems inherent in plant-derived medicines (Wall and Wani, 1994). A National Cancer Institute-sponsored plant collection program in the Pacific Northwest region in 1960 gathered 650 plant species for screening through various cell-based and in vivo anticancer evaluations. The twigs and bark of the Pacific Yew, Taxus brevifolia, showed very good activity in a cell assay and was subsequently subjected to bioassay-directed isolation of the active compound by Wall and Wani at the Research Triangle Institute. Taxus species have a rich history of use by native peoples for treatment of various illnesses, including cancer. Painstaking efforts were involved in the isolation of the active component, taxol, which was shown to be a new taxane deriva-
FIGURE 8 Structure of the anticancer natural product "taxol" from the tree Taxus brevifolia.

Despite its intriguing structure and impressive activity in cell assays (but not in antileukemic in vivo assays!), research on taxol languished for nearly a decade due to a combination of short supply of the drug, difficulty in working with this lipid soluble substance, and lack of research funds.

Interest in the drug was rekindled when Susan Horwitz at the Albert Einstein Institute showed that taxol had the opposite effect on microtubules compared to all other known classes of antimitotic anticancer agents. By binding to a unique drug-binding site on the β-subunit of tubulin, taxol promotes the stabilization of microtubules whereas other classes of antitubulin agents promote their destabilization. This mechanistic difference fueled additional evaluations of taxol, and once formulation problems were overcome, taxol emerged as the most significant new anticancer drug of the past 20 years. Currently, taxol has shown good efficacy to ovarian, breast, and lung cancers, as well as several others. However, its development was greatly impaired from severe supply shortages, a common issue for natural product derived drugs. It has been estimated that it takes the recoverable taxol content of four full grown 70 year-old trees to treat 1 patient for 1 year. At its peak, natural collections of 60,000 pounds of Pacific Yew bark were made in 1988 and 1989, which threatened the survival of this relatively common tree. Subsequently, supply of the drug has been made plentiful by harvest of the needles from cultured trees (Taxus baccata), extraction of the essential core of the taxol molecule, and its conversion by chemical synthesis into taxol. The 25-year development period from isolation to clinical use is typical of modern pharmaceutical development wherein many basic scientific, regulatory, and safety issues must be answered before drugs reach the market. The taxol supply problem has subsequently initiated a national program to explore how to resupply a natural product drug in large scale during its development as an anticancer agent.

2. Camptothecin

The same group that discovered taxol made a second major contribution in the fight against cancer. This latter compound, called camptothecin, comes from the wood and bark of the tree Camptotheca acuminata (Pantozis et al., 1996). Monroe Wall, who was working at the eastern regional Research laboratory of the USDA in Philadelphia, Pennsylvania, discovered it in the 1950s. Previously, this group had been screening plant extracts for steroids that could be used as cortisone precursors. In 1958, samples from this entire collection of extracts were sent to the National Cancer Institute (NCI) for testing against certain cancer cell lines. Only one extract, that from Camptotheca acuminata, showed activity in this particular biological assay.

Camptothecin was eventually isolated through the use of the Craig Countercurrent Distribution method, which takes advantage of different compounds having varying solubilities in dissimilar solvents. In 1966 the structure of camptothecin, and its corresponding sodium salt, were published on the basis of a large amount of spectroscopic and chemical degradation data (Fig. 9). In early animal trials, camptothecin showed significant antileukemic activity to the L1210 cell line at concentrations as low as 0.2 mg/kg body weight of mice. By 1970, camptothecin had progressed to phase I clinical trials at NCI and soon progressed to phase II clinical trials. Encouraged by these results, the NCI decided to introduce the water soluble sodium salt of camptothecin into phase II clinical trials. The results of these trials were not very promising and camptothecin lost much of the interest it had attracted in its earlier phase I trials. It was later discovered that the sodium salt of camptothecin was only one-tenth as active as the parent compound against the P388 cancer line.

In 1985, it was found that camptothecin acts by a novel mechanism and interest was rekindled. These later studies showed that camptothecin and various water-soluble analogs worked by inhibiting the action of mammalian topoisomerase I (T-I, an enzyme in mammalian topoisomerase I (T-I, an enzyme in mammalian topoisomerase I (T-I, an enzyme in mammalian topoisomerase I (T-I, an enzyme in mammalian topoisomerase I (T-I, an enzyme in mammalian topoisomerase I (T-I, an enzyme in mammalian topoisomerase I (T-I, an enzyme in mammalian topoisomerase I (T-I, an enzyme in mammalian topoisomerase I (T-I, an enzyme in
malian tissues responsible for unwinding DNA during the replication process) by a mechanism never before observed. This coincided with later discoveries that T-I is overexpressed in a variety of advanced cancerous tumors.

In 1991 two water-soluble analogs of camptothecin, topotecan and irinotecan, were introduced into advanced phase II clinical trials in the United States. Within the next few years topotecan was also approved for use in Japan and other analogs were approved for phase I clinical trials in Europe. Subsequently, topotecan and irinotecan were approved for use in certain types of ovarian and breast cancers in the United States.

B. Bioactive Alkaloids: Tropane Alkaloids

Alkaloids are a very diverse and abundant class of natural product. Their chemical structures range from relatively simple to very complex; but all contain one important theme, a basic nitrogen atom, usually heterocyclic, within their structural formula. Alkaloids are not only diverse in chemical structure, but are also diverse in the organisms that produce them. Alkaloids are divided into different categories depending on the different moieties within their chemical structures. One of the best known subclasses are the tropane alkaloids (Tyler et al., 1988). The name is derived from the inclusion of a nitrogen-containing bicyclic appendage, tropane. Historically, two of the most important drugs in this class are atropine and cocaine (Fig. 10). The tropane alkaloids continue to be vigorously investigated by many laboratories throughout the world.

Atropine was originally characterized from the poisonous plant Atropa belladonna. Aside from its poisonous properties when ingested, early usage of this plant involved placing the juice from the berry into the eye, causing dilation of the pupil. Women of the 16th century thought this to be a beauty-enhancing quality, hence the plant name ‘belladonna’ (‘beautiful lady’ in Italian). In modern times, atropine has been used to counteract the effects of cholinesterase inhibitors. These inhibitors, such as physostigmine and organophosphate insecticides, act as competitive inhibitors of acetylcholine at the muscarinic receptor. Numerous synthetic routes have been developed to produce atropine, however, none are economically feasible. At the present time, the most viable economic means by which to obtain this useful alkaloid is still by isolation from the raw plant material. Atropine can be isolated from a number of different plant sources (e.g., Hyoscyamus muticus).

Another tropane alkaloid, cocaine, was first isolated from Truxillo coca in 1860 with its intentional use in western medicine as an anesthetic. In Incan civilization this shrub was known as the ‘The Divine Plant.’ Natives would chew the leaves (mixed with lime so as to solubilize the alkaloid substituents) in order to endure long journeys or extensive manual labor without experiencing the inherent fatigue associated with great exertion. In present-day times, cocaine’s highly addictive stimulatory effect has given rise to many social problems, in addition to retaining its importance as an anesthetic. Chemically, cocaine is a tropane alkaloid containing a methyl ester appendage. The primary mechanism of action of cocaine is that of a dopamine re-uptake inhibitor. In essence, this leaves dopamine at the synapse for a longer time, giving rise to the euphoric sensation. Research on cocaine is still very actively pursued across the globe, from synthetic derivatives to effects on prenatal exposure to the drug.

C. Other Bioactive Classes: Steroidal Glycosides—Digitalis

Original therapeutic usage of Digitalis sp., commonly known as foxglove, can be found dating back to tenth century England, Wales, and Europe where it was used as an expectorant and treatment for epilepsy, swelling, and sore throats. Officially entering the London Pharmacopeia in the mid 1600s, and subsequently the pharmacopeia of most other countries by the eighteenth century, the late 1700s saw the emergence of Digitalis as a remedy for ‘dropsy.’ This disease condition is characterized by a swelling that results from inadequate heart function and is known today as congestive heart failure. The utility of this plant was introduced to Western medicine by William Withering in 1775, who was led to the medicinal value of Digitalis through conversations with a Welsh woman known for healing dropsey using herbal remedies containing foxglove. The active components of foxglove are classified as cardiac glycosides, two of which give rise to the drugs digoxin and
Cardiac glycosides are naturally occurring steroids that are functionalized by sugar and lactone moieties (Trease and Evans, 1983). The most active of these possess an unsaturated lactone ring connected to C-17 of the steroid nucleus. Optimal cardiac activity also requires cis stereochemistry of the A/B and C/D ring junctures. Structure-activity studies on the cardiac glycosides have demonstrated that the sugar moiety aids in bioavailability and solubility, while functionalization of the steroid nucleus with hydroxyl, methyl, and lactone ring groups is essential for pharmacological activity. The unsubstituted aglycone, or steroidal portion, is less active than the glycoside (sugar attached).

In treating heart failure, the cardiac glycosides of Digitalis function by increasing the strength and efficiency of ventricular contraction, thus shortening the length of contraction and allowing the heart muscle a longer relaxation time between contractions. This results in recovery of the myocardium, decreased heart rate, and improved renal function through enhanced circulation.

Nearly 30 different cardiac glycosides have been identified in Digitalis purpurea and more than 70 from Digitalis lanata. Cardiac glycosides have been observed in nine other Digitalis species. Interestingly, none of the major glycosides of D. lanata are identical to those isolated from D. purpurea, with the most common alteration in the molecule being acetylation of hydroxyl groups in the D. lanata representatives. This fact indicates great chemical, biochemical, and metabolic diversity among Digitalis species.

Historically, a drawback of Digitalis treatment is its toxicity. This toxicity led to its being eschewed by most medical practitioners until its reappearance in relatively modern treatments. The fact that no fewer than nine prescription products from Digitalis species are used to treat congestive heart failure, atrial fibrillation, and atrial and ventricular tachycardia demonstrates the human need and utilization of not only the innocuous plants, but also the noxious varieties, such as Digitalis.

1. **Ginseng**

Perhaps the most venerable of traditional Chinese herbs is 'ginseng' or 'Korean ginseng,' also known by its Latin name *Panax ginseng*. Usage of *Panax ginseng* has a long and rich history in China dating back to the Han dynasty some 2000 years ago. The earliest description of its application appeared in the oldest Chinese pharmacopoeia, believed to have been written in the first century during the late Han dynasty. It is the root of *Panax ginseng* that is used extensively in Chinese medicine as an effective tonic for enhancing stamina as well as the capacity to endure fatigue and physical stress. *Panax ginseng* was originally found in northeastern China, Korea, and eastern Siberia. Other congeners of ginseng, such as the *P. quinquefolius* (American ginseng) are also used widely as medicines. *Panax quinquefolius* was initially reported growing wild in the northeastern United States. Another common member is the Siberian ginseng (*Eleutherococcus senticosus*), which belongs to the Araliaceae family and contains distinctly different glycosides. The main source of *Panax ginseng* used in the United States and Europe comes from Korea where it is cultivated extensively. Two main forms of ginseng are currently being used; white ginseng, which is processed from the air-dried roots of five to seven year old plants, and red ginseng, which is white ginseng that has been steam treated for 2 to 4 hr.

Research on the chemistry of *Panax ginseng* suggests that the major active components of the roots are the ginsenosides (glycosides), which are derivatives of the triterpene ‘dammarane.’ To date, more than 20 ginsenosides have been reported from the roots, leaves, and flower buds of ginseng (Fig. 12). Common sugars found attached to the protopanaxadiol core of these ginsenosides include glucose, maltose, fructose, and saccharose. In spite of the discoveries of these bioactive glycosides in ginseng, the precise pharmacological mechanisms of ginseng’s actions are unclear. However, there are extensive reports of the effects of ginseng on the function of the neuroendocrine system, the central nervous system (memory, learning, and behavior), carbohydrate and lipid metabolism, the immune system, and the car-
diovascular system. For instance, ginsenoside Rh1 was demonstrated to have CNS-sedative, tranquilizing, and hypotensive actions while ginsenoside Rg1 was shown to be CNS-stimulating, hypertensive, and possess anti-fatigue properties. As this example demonstrates, many of the reports on the biological activities of ginseng’s metabolites are contradictory; this may be due to the difference of ginsenoside content in ginseng root or root extracts.

2. Echinacea

The application of Echinacea as a medicinal plant is derived from its use by Native North Americans. Traditional use of Echinacea involved the treatment of numerous conditions, such as external applications to wounds, burns, inflammation of lymph nodes (mumps), and insect bites. Internal uses of the plant included stomach cramps, headache, coughs, chills, measles, and gonorrhea. The root of Echinacea was the most important part of the plant used in these native treatments. There are also reports of the use of the juice or a paste of macerated fresh plant by Native North Americans. In view of the multitude of beneficial uses of Echinacea by these people, it was likely one of their most important medicinal plants.

The first pharmacological studies on Echinacea extracts began in the early 1950s and were mainly concerned with its nonimmunological actions. However, we now know that Echinacea acts mainly via stimulation of the nonspecific immune system. The main immunostimulatory components of Echinacea extracts were found to be lipophilic alkylamides as well as cichoric acid. Cichoric acid (2,3-O-dicaffeoyltartaric acid) was first isolated from Echinacea purpurea and found to cause significant stimulation of phagocytic activity in an in vitro granulocyte bioassay (Fig. 13). In addition, polysaccharides have also been suggested to be active components in Echinacea juice and aqueous extracts. A purified mixture of polysaccharides from the roots of E. purpurea were found to increase in vitro phagocytosis by macrophages 23 to 32% at concentrations of 0.01 and 0.001 mg/mL. It appears that the complete immunostimulatory properties of Echinacea extracts are due to the combination of activities from several different classes of compounds.

IV. CONCLUSIONS

The examples given here clearly demonstrate the pivotal role that plants have played in the development of much of our current pharmacopeia (Shu, 1998). Indeed, it has been estimated that as much as 37% of all pharmaceutical sales are for compounds that derive, either

![Structure of cichoric acid](image)
wholly or in part, from natural products. For anticancer drugs, the percentage deriving from natural products is even higher (ca. 60%). Nevertheless, the enormous task that we require of pharmaceuticals, to have potent and selective action for a particular disease state, to be orally active and distribute with high efficiency to the target organ, and to be nontoxic to normal cells and tissues of the body, makes the development of any new pharmaceutical a long-term affair. On average, it is estimated that it requires nearly 15 years from the time of discovery to the time of clinical application of a new drug entity, a process estimated to cost approximately $650 million U.S. dollars in 1998.

Recognition of the enormous long-range value of plant-derived natural products, and the genetic sequences encoding for them, is one of the most compelling pragmatic arguments for biodiversity preservation. Indeed, the International Cooperative Biodiversity Group program in the United States has the simultaneous objectives of natural product drug discovery, biodiversity preservation, and developing short- and long-term economic benefit from ecologically preserved habitats, mainly in the tropics. One program funded through this mechanism has the innovative strategy of fostering the transmission of the folklore uses of plants and animals within a primitive society by artificially enhancing the “status” associated with the village “shaman.” This is being accomplished through paying modest “salaries” to shamans and their assistants, which in turn gives these individuals a unique and desirable status with their peers.

As our understanding of the molecular basis for various diseases increases and defines new drug targets, we will need an ever-increasing source of molecular diversity to act as starting points for developing new generations of more potent and more selective drug agents. Nature’s storehouse is still the largest and most chemically diverse source of unique molecular architectures. Preservation of this resource, and of the genes that encode for it, should be an overarching societal priority.

See Also the Following Articles

Biodiversity as a Commodity • Bioprospecting • Economic Value of Biodiversity, Overview • Ecosystem Services, Concept of • Pharmacology, Biodiversity and

Bibliography

I. What Is Pollination?

The relationship between the plants and pollinators is commonly assumed to be mutualistic, with the plants benefiting from the transfer of pollen and the pollinators receiving a nutritional or other reward. However, there are also instances in which the plants appear to provide no reward and cases in which pollinators may also be seed predators. This rich diversity of relationships has been fertile ground for investigation by pollination biologists for hundreds of years. The importance of pollination for agriculture has also been a powerful stimulus for the study of plants and pollinators.

II. The Diversity of Pollinators

III. Coevolutionary History of Plants and Pollinators

IV. Pollinators in Natural Systems

V. Pollinators in Agriculture and Their Economic Value

VI. Conservation Biology and Pollination
There is an important distinction to be made between “flower visitor” and “pollinator.” Visitors to flowers (anthophilous animals) cannot be assumed to be pollinators because in reality they may be nectar or pollen thieves which, owing to a mismatch in morphology or an unusual behavior, do not pollinate. For example, “base workers,” or insects that remove nectar from between the petals of a flower with an unfused corolla, or insects too small to contact the reproductive parts of a flower would not be pollinators despite the fact that they may spend much time harvesting nectar. Similarly, a small insect that collects pollen from anthers but never contacts stigmas is an example of a pollen thief that does not pollinate. Simple experiments, such as collecting stigmas of flowers to confirm pollen deposition, or the observation of transfer of a pollinium (a packet of pollen, characteristic of flowers in the Orchidaceae and Asclepiadaceae), can help to distinguish between visitors and pollinators. Unfortunately, many animals observed on flowers are probably best categorized as flower visitors because this kind of confirmation of pollination has not been conducted.

II. THE DIVERSITY OF POLLINATORS

The list of species of animals that serve as pollinators is long and diverse. The largest group of pollinators is insects, but both flying and non-flying mammals, birds, and at least one reptile have been recorded as pollinators. Table I provides a list of pollinator classes for the world’s wild flowering plants (the angiosperms; approximately 240,000 species) and estimates of the number of species in each class. Much work remains to be done to confirm the activities of these taxa as pollinators, but these numbers are the first estimates available.

A. Invertebrate Pollinators

Although flower visitation has been observed in species from at least 16 orders of insects, only 4 include many species that regularly pollinate flowers and that seem to have been involved in coevolutionary interactions with plants: the beetles (Coleoptera), flies (Diptera), butterflies and moths (Lepidoptera), and ants, bees, and wasps (Hymenoptera). In addition, the thrips (Thysanoptera) include some pollen-eating species that may be pollinators, some stoneflies (Plecoptera) and true bugs (Hemiptera) visit flowers and eat pollen and nectar, and some lacewings (Neuroptera), scorpion-fly (Mecoptera) and caddisfly (Trichoptera) species eat nectar. Additional work is needed to confirm whether most of these lesser known flower visitors are indeed pollinators.

1. Beetles

Beetles have been abundant since at least the Mesozoic, and it is likely that some of them have been flower visitors since the origin of the earliest angiosperms. For example, the current association of beetle pollination (cantharophily) with primitive woody angiosperms (e.g., Magnolia) probably dates back to the evolutionary origins of both groups. Beetle pollination is considered to be the most primitive type of pollination by animals and is not very important in cool temperate regions. It is most common in the moist tropics and to a lesser degree in and areas. Beetles constitute the largest order of insects, and some of this diversity is thought to have arisen through the same evolutionary radiation of flowers and insects during the Tertiary that led to the origin of the other major orders of flower-visiting insects.

2. Flies

Flies are probably the second most common order of flower visitors (after Hymenoptera). Fly visitors from

---

**TABLE I**

Pollinator Classes for the World’s Wild Flowering Plants
(Approximately 240,000 Angiosperm Species)

<table>
<thead>
<tr>
<th>Pollination category</th>
<th>Estimated pollinator taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind (shunt)</td>
<td>20,000</td>
</tr>
<tr>
<td>Water</td>
<td>130</td>
</tr>
<tr>
<td>All insects</td>
<td>289,166</td>
</tr>
<tr>
<td>Bees</td>
<td>40,000</td>
</tr>
<tr>
<td>Hymenoptera (bees and wasps)</td>
<td>43,245</td>
</tr>
<tr>
<td>Lepidoptera (butterflies and moths)</td>
<td>19,110</td>
</tr>
<tr>
<td>Diptera (flies)</td>
<td>14,126</td>
</tr>
<tr>
<td>Coleoptera (beetles)</td>
<td>211,955</td>
</tr>
<tr>
<td>Thysanoptera</td>
<td>900</td>
</tr>
<tr>
<td>All vertebrates</td>
<td>1,221</td>
</tr>
<tr>
<td>Birds</td>
<td>923</td>
</tr>
<tr>
<td>Bats</td>
<td>165</td>
</tr>
<tr>
<td>Mammals other than bats</td>
<td>133</td>
</tr>
</tbody>
</table>

*The numbers are “the number of species comprising the invertebrate and vertebrate genera, families, and orders in which there are more known effective pollinators than there are pollen or nectar cheaters, robbers, or avoiders” (p. 214) (reproduced with permission from Nabhan and Buchmann, 1997).
at least 71 families of Diptera to flowers in 137 plant families have been recorded in the literature (compilation by honey). The families Syrphidae (hoverflies), Bombyliidae (bee flies), and Muscidae are especially common as flower visitors. In tropical areas, the diversity of Diptera can rival or exceed that of the Hymenoptera. For example, 4,896 species of Diptera have been recorded from Australasia from flower-visiting families compared to approximately 2,570 bees (superfamily Apoidea), whereas for the neotropics the estimates are 2940 species for Diptera and 5630 species for bees (Roubik, 1995). In some parts of the world, including the Faroe Islands, New Zealand, and Australia, where bees are nonexistent or relatively rare, flies have filled that ecological niche of pollinators.

Pollination by flies (myiophily) is economically important, in tropical areas flies are the primary pollinators of cacao and they also pollinate mango, cashew, and tea. Roubik (1995) lists pollinators of 785 species of cultivated plants in the tropics, and 26–33 of these plants are apparently pollinated only by flies, 32 or 33 by flies as the primary pollinators, and 87–101 by flies as secondary pollinators. In Europe, flies are used commercially to pollinate protected crops of onion, chive, carrot, strawberry, and blackberry. In both tropical and temperate areas, flies appear to be more important as pollinators than has been generally recognized, and there is much need for additional research on them.

In addition to the more specialized flower-visiting flies, carrion and dung flies are pollinators of some plants in a category of pollination called sapromyophily. The basis for attraction of the flies is floral fragrances that mimic rotting flesh or dung, and typically the flies do not receive any nutritional reward for their visits. Blood-feeding Diptera such as mosquitoes and biting flies (Tabanidae; hordesflies and deerflies) are also sometimes pollinators, attracted to flowers by the nectar they offer.

3. Butterflies and Moths

The Lepidoptera (butterflies and moths), as a consequence of their two distinct lifestyles as primarily herbivorous larvae and commonly flower-visiting adults, have dichotomous relationships with plants. In at least one case, the pollination of yucca, the act of pollination is accompanied by oviposition and the larval yucca moths are seed predators. Although not all Lepidoptera feed as adults, most if not all that do depend on nectar as a source of both sugars and amino acids. Some of the longest-lived of all butterflies, the Neo-tropical Heliocidias species, also feed regularly on pollen by collecting it on their proboscis and regurgitating nectar on the pollen mass to leach out amino acids. The floral rewards sought by Lepidoptera are not always nutritional; some species are attracted to flowers by the pyrrolizidine alkaloids that they produce, which the male butterflies require for the production of pheromones, for mating success, and for the establishment of multi-species leks. Lepidoptera, as agents of natural selection, may be responsible for much of the diversity of defensive chemicals in plants as well as the diversity in flowers and the rewards they offer.

Two families of moths stand out as pollinators: the hawkmoths (Sphingidae; also called sphinx moths) and the noctuid moths (Noctuidae; the largest family in the order Lepidoptera). The hawkmoths are strong and agile fliers that maintain very high body temperatures (e.g., as high as 40°C) while in flight. Some diurnal species are easily mistaken for hummingbirds (e.g., the European hummingbird hawkmoth, Macroglossum stellatarum). Some hawkmoths are migratory and (in successive generations) may move hundreds of kilometers during a summer. Hawkmoths include the pollinators with the longest known tongues; most of these extraordinary species are found in Madagascar. Darwin measured nectar spurs in some orchids from Madagascar that were 29 cm long, and concluded that there must be moths there that have proboscides of the same length. Forty years after his prediction, a hawk moth with a proboscis that reaches more than 24 cm in some individuals was found and given the name Xanthopan morgani praedicta. There is actually a large guild of long-tongued hawkmoths and flowers with matching corolla lengths in Madagascar.

Some noctuid moths are long-distance migrants, flying distances as far as 1800 km. One species is reported to travel as much as 1000 km in a day when migrating from the Baltic states to Britain. This presents the opportunity for some very long-distance gene flow on the part of their food plants. Some species migrate vertically, moving up in altitude during some times of year to avoid unfavorable seasonal weather. Before these migrations the moths may store substantial quantities of fat, and at least one such species was an important part of human diets; the bogong moth (Agrotis infusa) was collected while aestivating in the Snowy Mountains by aboriginal Australians.

4. Ants, Sawflies, Wasps, and Bees

The Hymenoptera (ants, sawflies, wasps, and bees) are the largest and most diverse group of pollinators. Ants
are only rarely recorded as pollinators, although they are not uncommon as flower visitors collecting nectar. One suggestion for why there are not more species of plants adapted for pollination by ants is that the antibi-
otic and other secretions on their exoskeletons are detrimen-
tal to pollen survivorship and growth. Sawflies are more common as pollinators. They have herbivorous larvae, and in some cases the adults restrict their flower visits to the larval host plant. The adults may also eat flower parts in addition to nectar and pollen. Flower visitation by wasps has been studied in part because of interest in attracting and sustaining populations of parasitoid wasps (which prey on pest insects) in agricultu-
ral situations. However, predatory and social wasps have also been recorded as flower visitors and a few species of plants are specialized for pollination by pseu-
docopulation by male wasps (see Section IV,B2). Some of the most specialized relationships between plants and pollinators are those between species of figs and the wasps that pollinate them; many of these relation-
ships are reported to involve single species of wasps pollinating each fig species.

There are fewer species of bees than beetles, but a much higher proportion of bees are pollinators than are beetles. Almost all species of bees are dependent on flowers for pollen and nectar as nutritional resources as both adults and larvae, and of all the insects bees are the mostly highly adapted for flower visitation and pollination. Their behavior, morphology, and senses of vision and smell all appear to be adapted for finding and collecting floral resources. For example, the trichomatic color vision of bumblebees and honeybees includes ultravio-
let, which allows them to see the ultraviolet reflec-
tance patterns that many flowers use as nectar guides or a petal color. Humans cannot see these patterns without the help of cameras. Although pollen (primarily for feeding larval bees) and nectar (both a larval and adult resource) are the most common resources collected by bees, some flowers offer oil as a reward. The oil is collected by bees in the family Anthophoridae; females of these species have absorbent brushes for holding the oil and sharp edges on their legs that are used to squeeze the oil out so it can be mixed with pollen as food for larvae. Some bees collect resin as a floral reward, which they then use in nest construction. Pollination by bees (melittophily) is very important for agriculture (see Section V). The social bees are the best known crop pollinators and are the species most commonly managed for pollination or honey produc-
tion. These include honeybees (Apis), bumblebees (Bombus), and stingless bees (Trigona). The presence of large quantities of honey (concentrated nectar), pollen, and larvae makes social bee nests an attractive target for insects and vertebrates that prey on nests and adult bees. Bee-eaters and honey guides are examples of Old World birds that specialize on such resources.

B. Vertebrate Pollinators

1. Mammals

Although invertebrates are the most common pollina-
tors, mammals, birds, and lizards have also been docu-
mented as pollinators. Mammalian pollinators include both bats, some of which eat primarily nectar and pol-len, and nonflying species such as some Australian mar-
supials. Most bat pollination occurs in the tropics, but some migratory bats are important pollinators in tem-
perate North America to about 30° north latitude. In the Old World, species of the family Pteropidae are
important pollinators, whereas in the New World
flower-visiting bats are confined to the Phyllostomidae. Although pollination by bats (chiropterophily) is geo-
graphically widespread, the other mammalian pollina-
tors are more restricted and less common as pollinators, including lemurs in Madagascar and several species of Australian marsupials such as sugar gliders, honey possums, and some marsupial mice. There are some reports of other mammals, including mice and giraffes, that may serve as pollinators in unusual cases. Verte-
brate taxa other than mammals or birds are only rarely reported as pollinators; there is one report of a lizard that appears to be a pollinator.

2. Birds

Bird pollination (ornithophily) is common in many parts of the world. Several families of birds are primarily nectarivorous and undoubtedly include important pol-
inators. These include hummingbirds (Trochilidae; New World), honeyeaters (Meliphagidae; Australia, New Zealand, and parts of Asia), sunbirds (Nectarini-
da; Africa and southwest Asia to the Philippines), sug-
arbirds (Petroicidae; South Africa), flowerpeckers (Dicaeidae; Asia and Australasia), and Hawaiian honey-
creepers (Drepanididae; Hawaii). Other families that include some nectarivores or for which there are a few records of flower visitation include the Thraupidae (honeycreepers and some tanagers; New World), Icteridae (orolos; New World), white-eyes (Zosteropidae; Africa, Asia, and Australia), and Pittacidae (lorrels and hanging-parrots; Southeast Asia and Australia). Many flower-visiting birds are recently extinct in the
Pollinators have evolved a diverse range of relationships with plants, some even taking over the role of pollination when native species are extinct or introduced. In the South Pacific islands, for instance, a species has taken over the pollination duties of endemic plants pollinated by extinct birds.

**III. COEVOlUTIONARY HISTORY OF PLANTS AND POLLINATORS**

The “abominable mystery” to which Darwin referred, the origin and relationships among flowering plants, is still being revealed. Beetle pollination is thought to be the primitive condition in flowering plants, but by 100 million years ago butterflies and moths had joined beetles as important pollinators. Recent fossil evidence indicates that the origin of the angiosperms (flowering plants) may date back as far as the late Jurassic, suggesting that the origin of the coevolutionary relationship between flowers and pollinators may also be older than previously thought. Much of the tremendous biodiversity found in the flowering plants is thought to have arisen through evolutionary interactions with pollinators, and certainly much of it is maintained through their flower-visiting activities that result in production of seeds. The remarkable diversity of flower size, shape, odor, rewards, color, and pollination mechanisms is the result of these coevolutionary relationships.

**IV. POLLINATORS IN NATURAL SYSTEMS**

*A. Obligate and Facultative Mutualisms*

The relationships between plants and pollinators range from obligate to facultative in nature. Although one study in Brazil found that 43% of plants studied were visited by only a single species of pollinator, other studies have found tens of visitor species to a single plant species. In only a few cases have researchers quantified the relative importance of different species in cases with multiple pollinators of a single plant species.

A common theme of multiyear studies is variation, both temporal and spatial, in the abundance of these different groups of pollinators, which can include diverse taxonomic groups including both vertebrate and invertebrate pollinators (e.g., hummingbirds and bumblebees) visiting a single plant species. By having several groups of potential pollinators, it is more likely that these plants will not suffer from pollen limitation (a shortage of pollination); if one group of pollinators is unusually low in abundance at a particular site or during a particular year, it is likely that another group will not be low in abundance. Thus, it can be important for there to be a diversity of pollinators available.

The obligate nature of some plant-pollinator relationships extends to pollinators because some insects appear to visit flowers of only a single species of plant. In such cases there must be a good match in the phenology of life cycles of both plant and insect to ensure success for both partners, and a concern about global climate change is the potential for such phenological relationships to become mismatched. There are similar cases of plant species visited by only a single species of pollinator. Although a plant or pollinator may thus have an obligate relationship with a single partner, the partner may be involved in relationships with multiple species.

Although we assume that most plant-pollinator relationships are mutualistic in nature, some examples appear to involve a mixture of positive and negative interactions from at least the plant's perspective. Several studies have demonstrated seed predation or other forms of herbivory by insects that may also serve as pollinators. These examples include beetles and palms, moths and yuccas, wasps and figs, as well as flies and moths and some flowers that they pollinate. Some of these relationships (e.g., yucca flowers and their moths) can be further complicated by the presence of closely related species of insects that are herbivores but not pollinators. Much detailed study is needed to dissect the complicated interactions in such relationships.

There is no doubt that long-lived or highly mobile pollinators must interact with a large number of plants because their life spans or travels extend beyond the temporal or spatial availability of single plant species. Thus, a group of plant species, even if they are not sympatric, could be considered mutualists if in concert they sustain and share a particular pollinator. This kind of multispecies relationship has significant conservation implications because it may be necessary to conserve plant species that flower widely separated in time and space to conserve a (e.g., migratory) pollinator.

*B. Cheaters in Pollination*

1. pollen and nectar robbers:

The nutritional rewards that pollinators find in flowers—pollen and nectar—are sometimes harvested by flower visitors that may pierce or bite holes in the flowers to obtain the resources “illegitimately.” Some species of birds (Diglossa, the flower-piercers) and bees (Xylo-
livestock food or for the essential oils they contain. The as crops, or gathered from the wild, for human and/or plants include at least 264 species that are cultivated grown in the European Union are unknown. These requirements of about one-third of the crop species Our knowledge of the pollination biology of crop spe-

process pollinate (pseudocopulation). Sex pheromones) sufficiently resemble female insects that in both morphology and odor (production of insect learn to avoid the flowers. Other orchids have flowers nothing from visiting (and pollinating), and quickly In such species, it appears that the pollinators gain pollen that is not accessible or attractive to pollinators. Species of orchids have flowers with no nectar and tily is perhaps best known in this regard because many advertising opportunities for feeding or sex. The orchid family is perhaps best known in this regard because many species of orchids have flowers with no nectar and pollen that is not accessible or attractive to pollinators. In such species, it appears that the pollinators gain nothing from visiting (and pollinating), and quickly learn to avoid the flowers. Other orchids have flowers that in both morphology and odor (production of insect sex pheromones) sufficiently resemble female insects that males attempt to copulate with them and in the process pollinate (pseudocopulation).

V. POLLINATORS IN AGRICULTURE AND THEIR ECONOMIC VALUE

Our knowledge of the pollination biology of crop spe-
cies is surprisingly sparse. For example, the pollination requirements of about one-third of the crop species grown in the European Union are unknown. These plants include at least 264 species that are cultivated as crops, or gathered from the wild, for human and/or livestock food or for the essential oils they contain. The best known pollinators of cultivated food plants are honeybees, which have been widely introduced throughout the world (see Section VI). However, honeybee populations in the United States have been deci-

mated in recent years by the introduction of two species of parasitic mites and are also being threatened by the introduction of a hive beetle. Most feral colonies of honeybees in the United States have disappeared in the past decade and the number of managed colonies has declined to less than half of its peak. This decline has resulted in a growing appreciation of the roles of other, native, species in the United States and elsewhere as pollinators of crops. These problems have also prompted some attempts to quantify the economic value of pollination services provided by both managed and wild pollinators. A recent model estimates that U.S. consumers realize $1.6–5.7 billion in annual social gains that would be lost if honeybee services for 62 crops were reduced. The potential value of non-honeybee pollinators in the U.S. agricultural economy is estimated at $4.1–6.7 billion each year. The global value of pollination services has been estimated to be $117 billion per year.

A. Introduced Species of Pollinators

Undoubtedly the largest-scale introduction of pollina-
tors has been the global traffic in the European honey-
bee (Apis mellifera). Such continental-scale introduc-
tions have a long history. For example, European colonists in North America imported honeybees as early as 1641. This species went on to establish feral colonies throughout much of North America and probably through subsequent introduction throughout much of the neotropics. They were also introduced to Australia. Probably the second largest introduction of pollinators has been the spread of bumblebees (Bombus). The first introduction of bees anywhere specifically for pollina-
tion was that of European bumblebees to New Zealand in 1885, and subsequently they were also introduced to Australia. The second successful introduction of a pollinator was a fig-pollinating wasp (Blastophaga psenes) into California in 1899. In these cases, the incentive for introducing bees or wasps was for pollination of introduced species of plants that were not attractive to or could not be pollinated by native pollinators (e.g., clover in New Zealand and figs in California). Recently, international transport of bumblebees has been the result of their use in pollina-
tion of greenhouse crops such as tomatoes. In the 1980s, techniques were developed for the year-round commer-
cial-scale propagation of bumblebee colonies (which...
POLLINATORS, ROLE OF

A. Threats to Plant–Pollinator Mutualisms

Pollinators face a large variety of threats of anthropogenic origin, including habitat fragmentation, a variety of effects of agriculture, pesticides and herbicides, and the introduction of both pollinators and plants. Fragmentation of habitats is likely to affect nonflying pollinators most strongly because it may be difficult for them to locate and visit patches of flowers that are isolated by areas of different, perhaps nonpreferred, habitats. However, even winged pollinators may be reluctant to leave a preferred habitat, such as undisturbed forest, to fly across a newly created pasture to another patch of forest. If the fragments are small enough to affect the size of pollinator populations, there may be negative genetic consequences for the smaller populations.

Modern agriculture can have a variety of negative impacts on pollinators. The creation of large fields that are disturbed regularly by plowing can prevent ground-nesting bees from establishing populations, and if only a single plant species is available there may not be enough pollen and nectar resources to support the life cycle of a pollinator species. The use of pesticides to control agricultural pests can have a negative impact on pollinators, and herbicides may remove species that could provide resources for the pollinators. Pesticide use in nonagricultural areas, such as spraying of large tracts of forest to control lepidopteran herbivores, has also been shown to have strong negative impacts on nontarget pollinator populations.

The introduction of exotic plants, such as weedy plants, could have potential implications for the pollination of native plants. If the exotic is close relatives of native species, there is potential for hybridization. Also, if the exotics are prolific producers of nectar or pollen, they may draw pollinators away from native species that may then suffer a deficit in seed production. The potential for these kinds of consequences is not well understood, but this area is beginning to attract attention from researchers.

The ecological consequences of widespread introductions of A. mellifera and Bombus for native plants and pollinators are not well-known because of a lack of data from before the introductions. We can speculate that the introduction of honeybees to North America may have had significant consequences for species of bumblebees that have the same proboscis length be-

B. Pollinators as Vectors and Victims of Engineered Genes

As genetic engineering of crop plants has become more common, and as these crops are approved for cultivation in the field, concerns have arisen about the potential for gene flow through pollen transport between the engineered genotypes and wild plants. One management tool to prevent, or at least minimize, this potential is to plant a buffer zone of unengineered plants around a field of engineered plants in hopes of intercepting pollinators carrying pollen before they can visit related native plants.

There is also the potential for pollinators to become victims of engineered genes. For example, laboratory studies of monarch butterfly caterpillars showed that if they eat significant quantities of corn pollen from plants engineered to have the BT toxin in their leaves, this can kill them. Corn is wind pollinated, and a likely scenario is that large quantities of corn pollen could be distributed onto leaves of milkweed plants at the margins of corn fields. The potential detrimental effects of this on the monarch butterfly and other herbivores remain to be studied.

VI. CONSERVATION BIOLOGY AND POLLINATION

VI. CONSERVATION BIOLOGY AND POLLINATION

normal life cycle of a few months). Commercially produced colonies of bumblebees have now largely replaced the tedious hand pollination of tomato flowers with electronic vibrators (tomato flowers produce no nectar but are visited by bees that "buzz" the flowers to release pollen from poricidal anthers so that they can collect it for feeding their larvae, and during this process they pollinate the flowers) and the use of chemical sprays to induce fruit set. In the early 1990s a European species of bumblebee (Bombus terrestris) was introduced to Japan for use in greenhouse pollination; it has subsequently escaped and is now established in the wild. Although an effort is being made to exterminate the species in the wild (and to develop native species for use in greenhouse pollination), it is likely to be unsuccessful.

Other species of pollinators have also been introduced. The wasp pollinators of figs have been introduced to New Zealand and the United States, where they pollinate naturalized figs. The Japanese white-eye, an introduced passerine in Hawaii, has taken over the role of pollinator of two native plants that lost their original pollinator to extinction. An Asian bee is now the primary pollinator of alfalfa plants grown for seed production in the United States, and a Japanese bee is used in the United States and Canada for pollination of apples.

There is also the potential for pollinators to become victims of engineered genes. For example, laboratory studies of monarch butterfly caterpillars showed that if they eat significant quantities of corn pollen from plants engineered to have the BT toxin in their leaves, this can kill them. Corn is wind pollinated, and a likely scenario is that large quantities of corn pollen could be distributed onto leaves of milkweed plants at the margins of corn fields. The potential detrimental effects of this on the monarch butterfly and other herbivores remain to be studied.
cause the perennial honeybee colonies have many more workers than the annual bumblebee colonies. The spread of *A. mellifera scutellata*, the African subspecies ("Africanized honeybees"), from its point of release in Brazil to as far north as the United States was much more recent and has provided opportunity for some studies of their impact on native Neotropical pollinators (relatively minor, it appears).

Species from every major group of pollinators, invertebrate and vertebrate, have been classified as endangered, and recent extinctions have been documented for others. Thus, pollinators are just as susceptible to the current human-induced mass extinction as any other group of organisms.

### B. Potential Management Solutions

A first step in solving problems in the conservation of plant–pollinator relationships is gathering information about both the nature of the relationships and the problems plants and pollinators face. Probably the best way to conserve pollinators is to preserve habitat that includes their food plants and nest sites. Given that the growing human population is resulting in rapid habitat destruction, the establishment of protected areas is an important conservation tool. In agricultural situations, management techniques as simple as changing the timing of pesticide application, leaving buffer zones around fields where bees can nest and food plants can grow, or providing suitable artificial nesting sites can make a significant difference. Domestication of wild bees may also play a role in their maintenance in agricultural situations.

If pollinators are locally extinct, it may be possible to reintroduce them once the factors that caused the original extinction are addressed. If the species is globally extinct, plants that were dependent on that pollinator could be maintained through hand pollination or the introduction of a suitable replacement. Although the bird was not introduced for this purpose, the presence of the Japanese white-eye in Hawaii has had the effect of replacing an extinct endemic avian pollinator. Removal of introduced pollinators, such as the honeybee, may also help to preserve populations of native pollinators.

We have taken some important steps toward conservation of pollinators and thus the relationships they have with plants. However, much basic knowledge about pollination and pollinators remains to be learned and it remains to be seen how many of these species can be conserved.

See Also the Following Articles

**BEETLES** • **BUTTERFLIES** • **COEVOLUTION** • **FLIES, GNATS, AND MOSQUITOES** • **HYMENOPTERA** • **MOTHS**

### Bibliography


I. Introduction

II. Perturbation of the Carbon Biogeochemical Cycle

III. Perturbation of the Nitrogen Biogeochemical Cycle

IV. Perturbation of the Sulfur Biogeochemical Cycle

V. Trace Pollutants of Global Importance

VI. Conclusion

GLOSSARY

acid deposition Precipitation or dry fall-out with a pH of less than 5.0.
anthropogenic Related to the activities of human beings.
biodiversity Total inventory of genes, organisms, species, populations, communities, and their habitats.
biogeochemical cycles Pathways and storage of chemical elements through the biota and geologic resources, including the atmosphere, hydrosphere, and lithosphere.
biomagnification The increase in body burden of persistent pollutants in organisms at higher trophic levels in food webs.
climate change Variability of weather over time.
ecotoxicology The study of chemical stressors on ecological resources.
food web The pathways of energy and nutrient movement between organisms in an ecosystem.
hydrologic cycle The movement of water between the surface of the earth and the atmosphere.
nitrogen saturation Nitrogen levels in excess of the biological nitrogen needs of an ecosystem potentially causing adverse ecological effects.
ozone Tri-atomic oxygen naturally present (and beneficial) in the stratosphere but present in the troposphere (due to human activities) in amounts potentially toxic to the biota.
pollution Materials or chemicals in excess of natural levels caused by the activities of humans.

BIODIVERSITY INCLUDES THE total inventory of genes, organisms, species, populations, communities, and their environments, and it integrates all of their complex interactions. This chapter focuses on the approximately 1.4 million species of living organisms known to presently exist on earth while recognizing that this total may represent fewer than 15 percent of the actual number (Raven and Wilson, 1992).

I. INTRODUCTION

Over geologic timescales, species numbers have been very dynamic. Massive species disappearances have occurred as follows: dinosaurs and marine reptiles in the Cretaceous (66 million years ago), reptiles and cono-
dents in the Triassic (210 million years ago), trilobites and marine animals in the Permian (250 million years ago), marine reef plants and animals in the Devonian (365 million years ago), and nautiloids and small ocean animals in the Ordovician (460 million years ago). Their mass extinctions are presumed to have resulted from earth impacts (asteroid collision?), climatic change, or large-scale alteration of biogeochemical cycles.

Changes in major element biogeochemical cycles over geologic time have been responsible for the appearance and disappearance of life forms. Changes in the oxygen cycle, for example, which allowed increased concentrations of oxygen gas in the troposphere and ozone in the stratosphere, allowed the evolution of life on earth to proceed. Major episodes of volcanic activity, resulting in marine eruptions and the release of sulfur containing and particulate pollutants, has periodically destroyed life.

Contemporary losses of species are recognized to result from physical stressors (e.g., habitat loss, harvesting), biological stressors (e.g., exotic organism introductions), and chemical stressors (changes in biogeochemical cycles, pollution). This chapter focuses on the latter group of stressors. Presently humans are recognized to have the ability to alter global and regional biogeochemical cycles via air, water, and soil pollution. As a result, the impact of pollution on biological diversity is a contemporary concern of major dimension (Barker and Tingey, 1992; Peters and Lovejoy, 1992).

II. PERTURBATION OF THE CARBON BIOGEOCHEMICAL CYCLE

Mineral carbon is located in a limestone (CaCO3) reservoir from which it may be leached into a mineral solution as dissolved hydrogen carbonate (HCO3). Carbon exists as carbon dioxide (CO2) in the atmosphere, in the stratosphere, and in the troposphere. Another fraction of carbon is ultimately fixed as coal, petroleum, and natural gas. This fraction may return to the atmosphere as CO2 via combustion.

In recent decades, human activities have accelerated the return of CO2 to the atmosphere via increased deforestation, biomass burning (Nepstad et al., 1990), and fossil fuel combustion (IPCC, 1996). The accelerating input of CO2 to the atmosphere from human deforestation and fossil fuel combustion (very approximately 8 billion metric tons annually) (Fig. 1) is a significant driver of global climate change.

In addition, the incomplete combustion of fossil fuels generates a variety of toxic and precursor compounds that contaminate the atmosphere and ultimately may cause adverse effects on plant and animal populations. Both of these stressors (i.e., climate change and air pollutants) have significant potential to reduce biological diversity.

A. Climate Change

Climate is defined as the time-averaged value of meteorological quantities. Over time, climate, like an ecosystem, is characterized by change not constancy. In geological terms, the climate of the earth is most typically characterized by extended “moderate periods” with equable weather the year round, lack of ice caps, and generally warm seas. Humans evolved after the last “moderate period” and our development has been in a period of climatic revolution. This period of revolution ary climatic alteration has been characterized by a complex of “cycles within cycles.” Over the past 3000 years, for example, the general evidence suggests that the northeastern portion of North America has become cooler and more moist. Over the past several hundred years, however, particularly during the first half of the present century, there is evidence for a moderating trend. There is general agreement that there has been a systematic fluctuation in recent global climate characterized by a net worldwide warming of approximately 0.5°C between the 1880s and the early 1940s. Warming has generally continued over the past three decades. The regulation of global climate is complex and incompletely appreciated. Numerous hypotheses have been proposed to explain the forces responsible for the variability of climate. The most plausible of these include variations of the solar constant, changes in solar activity, passage of the solar system through an interstellar gas-dust cloud, variation in the velocity of the earth’s rotation, gigantic surges of the Antarctic ice sheet, changes in the earth’s orbital parameters, and alterations in the interactions between glaciers and oceans. Climate may respond rapidly and dramatically to small changes in these independent variables.

Added to this complexity and uncertainty is the suggestion that the activities of human beings, particularly land use activities and atmospheric contamination, are currently influencing global and regional climates. Numerous trace gases of the atmosphere, including water vapor, carbon dioxide, methane, nitrous oxide, halocar-
POLLUTION, OVERVIEW

FIGURE 1  The global carbon cycle. The storage (numbers without arrows) and the annual flows (numbers with arrows) are shown for the global system. The units are petagrams of carbon (\( \times 10^{15} \text{gC} \)). From Schlesinger (1992).

bonds, and ozone (greenhouse gases), have strong infrared absorption bands. As a result, these gases can have a significant effect on the thermal structure of the atmosphere because they absorb within the 7- to 14-\mu m atmospheric window, which transmits most of the thermal radiation from the surface of the earth and troposphere to space. Presumably a primary result of more carbon dioxide and other “greenhouse” gases in the atmosphere will be warming. While incoming solar radiation is not absorbed by carbon dioxide and these trace gases, portions of infrared radiation from the earth to space are. Over time, the earth will become warmer. While the forces controlling global temperature are varied and complex, as suggested, the increase of 0.5°C since the mid-1800s is generally agreed to be at least partially caused by increased carbon dioxide.

General circulation models, widely used to predict future climates, are numerical models of the earth-atmosphere system that solve the basic equations for atmospheric motion and provide boundary conditions of the earth and ocean. General circulation models divide the landscape into a series of regions called grid boxes. Each grid box consists of a series of layers, which represent land, ocean, and layers of the atmosphere. Global weather is simulated using difference equations, which detail the physics and dynamics of energy and material movement among the boxes. To evaluate the climatic change hypotheses, models are initially run with current atmospheric conditions. They are then run with altered atmospheric trace gas concentrations (for example, carbon dioxide) to simulate a new climate. The difference between these two equilibrium climate simulations provides an estimate of how the climate system may change. Most general circulation models predict warming and overall intensification of evaporation and precipitation in the warmer climate regimes of doubled atmospheric carbon dioxide.

Using a mid-range scenario of driving variables, an increase in global mean surface air temperature relative to 1990 of about 2°C by 2100 has been estimated. Under any scenario, the average rate of warming will probably be greater than any seen in the past 10,000 years (IPCC, 1996). General circulation model predictions of the change in the hydrologic cycle with doubling of atmospheric carbon dioxide are not clear. While precipitation is generally estimated to increase, warming will intensify evaporation. As a result, soil moisture may not increase with warming; it may even decrease. The movement of water into and out of the soil-vegetation system has a large influence on the hydrologic cycle.
and is very variable regionally. Models have predicted that in the United States, portions of the Northwest could become wetter, portions of the Southern Great Plains could become somewhat drier, and portions of the Northeast could become considerably drier.

**B. Ecosystems and Climate Change**

The consequences of a warmer global climate, with even very modest changes in regional temperature and hydrologic patterns, can have profound effects on ecosystems. Natural ecosystems, as opposed to artificial ecosystems designed and intensively managed by humans, may be the most vulnerable to climate change. We are least able to intervene and facilitate adaptations of natural ecosystems because of their expansiveness, our limited knowledge about fundamental ecological processes, and the absence of effective global institutions to assume responsibility for adaptations.

A shift in climate zones may exceed the ability of vegetation, for example grassland, savanna, or forest systems, to adapt through migration. If climate zones shift hundreds of miles over the next century, rates of vegetative dispersal and colonization, which are on the order of tens of miles in a century, may be inadequate to maintain present-day eco-region patterns. With regard to terrestrial ecosystems, climate change may also adversely influence habitats by changing patterns of major stressor groups including fire regime, drought regime, and insect and disease relationships.

Alteration of ecosystems could lead to a change in species diversity, including a loss of species. The potential for migration of plants and animals to new suitable habitats is not known, but barriers such as water bodies, roads, or development could impede migration. Isolated (‘island’) species may find themselves in climate zones that are no longer suitable for their survival.

Montane species, unless they can migrate to higher elevations, may face extinction or extinction. A study of a 3°C warming in the Great Basin National Park in eastern Nevada concluded that it would cause 20 to 50 percent of species in individual mountain ranges to go extinct (Murphy and Weiss, 1992).

1. Forests

Tropical montane cloud forests are unique systems due to their dependence on regular cycles of cloud formation for their existence. Still et al. (1999) have simulated environmental conditions for such forests at CO2 doubling and found that a combination of reduced cloud contact and increased evapo-transpiration could threaten the persistence of these systems. Pounds et al. 1999 have, in fact, provided evidence that alterations in bird, reptile, and amphibian communities in the highland forests at Monteverde, Costa Rica, have occurred and may be related to recent increases in air temperature.

The vulnerability of forest ecosystems to climate change is a function of the forest location, biology, and human management activity. Forests at greatest risk will generally be those located in the northern portions of temperate latitudes. Altitude, proximity to continental margins, and distance from large water bodies will be important influencing factors. Plant species that are able to use increased CO2 efficiently will have an advantage over other species. Similarly, species that distribute seeds widely may fare better under climate change. A mixed-species forest may tolerate a wider variety of changes than would a single-species forest. Individual plants with a low tolerance for climate fluctuations would be least adaptable to ongoing climate change. Individual populations with little genetic diversity among plants might prove to be at greatest risk of long-term decline. The most intensively managed industry and private forest land may be least at risk of catastrophic loss or long-term decline because mitigation efforts to reduce such effects will be undertaken. Many private forest managers have both the financial incentive (and resources) and the latitude to protect against extreme loss from climate-related threats. They can use several available techniques: short rotations to reduce the length of time a tree is exposed to an unsuitable climate; planting better-adapted varieties developed through selection, breeding, or genetic engineering programs to reduce vulnerability; and thinning, weeding, pest management, irrigating, improving drainage, or fertilizing to improve general health (Table 1). The first observable effects of climate change on ecosystems will not be so much climate related as weather related. The near-term effects of climate change will be driven by changes in weather extremes and mediated through those stressors that have always been the primary controllers of ecosystem structure and health: insects, disease, wind, and fire. Even in regions where productivity may be ultimately improved, the transition period could be extended and punctuated by sudden dieback and decline. Forest ecosystems are complex, long-lived systems that can only slowly adjust to climate but that can suddenly be threatened by weather-related stresses.

The near-term response of forest systems to climate change will involve complex reactions to new averages, new patterns, and new extremes in weather variables. Some forest species that are specialized to current cli-
POLLUTION, OVERVIEW

TABLE I
Forest Ecosystem Vulnerability to Climate Change Caused by Alteration of the Global Carbon Cycle

<table>
<thead>
<tr>
<th>Forest location</th>
<th>Forest biology</th>
<th>Forest management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher latitude</td>
<td>Small, fragmented range</td>
<td>Fragmented forests</td>
</tr>
<tr>
<td>Higher elevation</td>
<td>No or few migration corridors</td>
<td>Low-diverse forests</td>
</tr>
<tr>
<td>Continental interior</td>
<td>Low genetic variance</td>
<td>High stand density</td>
</tr>
<tr>
<td>Maritime sites</td>
<td>Low species diversity</td>
<td>Inappropriate species</td>
</tr>
<tr>
<td>Forest-range boundaries</td>
<td>Genetically specialized to site</td>
<td></td>
</tr>
<tr>
<td>Low-productivity sites</td>
<td>History of widespread dieback</td>
<td></td>
</tr>
<tr>
<td>High-fire-risk sites</td>
<td>Heavy seed</td>
<td></td>
</tr>
<tr>
<td>Drier sites</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Smith et al., 1993.

climate conditions may not thrive. Altered patterns of exposure to high and low temperatures could mean that winter chilling requirements will not be met. Flowering, seed-formation, and seed-dispersal processes could be disrupted, especially if pollinators do not adjust to changing conditions. With longer growing seasons, trees might add more “light” earlywood relative to the “dense” latewood that forms at the end of the growing season. This would mean a lower-quality wood for structural lumber and higher costs for pulp mills. Changes in early growing season weather conditions, particularly moisture and frosts, may affect the establishment of seedlings. Warmer and moister weather might favor the spread and boost the significance of certain fungal diseases and insect pests. Elsewhere, the drying effects of higher temperatures could be especially damaging, especially where the frequency of drought is increased. Associated with droughts would be higher risks of secondary threats from forest fires and insects. Insect damage may increase, for example, if insect pests produce more generations or persist longer during the tree-growing season.

The initial effects of climate change will not at first be easily recognized as distinct from the effects of the normal regulators of forest growth and development. The potential initial effects of climate change can be illustrated by current weather-related stresses on selected highly valued tree species in several regions of the United States. These potential vulnerabilities, by region, are presented in Table II.

2. Coastal-Marine
Coastal zones are very productive ecosystems, very rich in species, and very perturbed by human activities. In addition, they may be at substantial risk from global climate change. These risks stem from changes in water column temperature or other water quality parameters, change in ocean current dynamics, and alterations of sea level. Widespread extinctions are not likely, but expansive alteration of community distributions and composition are possible. Most marine organisms, in contrast to long lag times discussed for forest ecosystems, have high mobility, large ranges, high fecundity, and rapid growth rates. As a result, they may be able to adapt to climate change more efficiently than terrestrial systems (Ray et al., 1992).

Coral reefs are the most species rich of marine ecosystems. These reefs are largely restricted to tropical waters, but generally do not occur where temperatures exceed 30°C for extended periods. As a result, even small increases of 1 to 2°C for the surface waters of tropical oceans could have important implications for reef ecosystems.

Temperate latitude estuarine and tidal wetland and marsh ecosystems will be at substantial risk if sea level increases over the next several decades. For each cm of rise in sea level, beaches may erode 1 m landward. Storm surges, a major eroding force, will increase, especially in areas of modest shore slopes. For every 10 cm rise, saltwater wedges in estuaries and tidal rivers may advance 1 km. Sea-level rise will also increase salinity intrusion into coastal freshwater aquifers (NAS, 1987).

3. Fresh Water
Freshwater ecosystems will present challenges for native biota if global climate continues to warm. Changes in precipitation and evapotranspiration can combine to offset surface-water and groundwater quantity and quality, low and high flow conditions, and drought and
TABLE II
Forest Vulnerabilities of Selected United States Forest Systems to Changes Resulting from Climate Alteration

<table>
<thead>
<tr>
<th>Region</th>
<th>Tree</th>
<th>Potential stressors/key climate factor and vulnerability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast/North Central</td>
<td>Maple</td>
<td>Insect defoliators/warm, dry weather</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Amorilia</em> (root death)/drought</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Deep root freezing/lack of winter snow cover</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Vulnerability:</strong> High potential for damage with warmer temperatures, drier conditions</td>
</tr>
<tr>
<td>Southeast</td>
<td>Loblolly pine</td>
<td>Southern pine beetle/prolonged hot, dry weather</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fungus (fusiform rust)/warm, moist weather</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fire (favoring longleaf or shrub)/warm, dry weather</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Storm damage/increase in coastal storms</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Vulnerability:</strong> Potential for much warmer weather (with increase or decrease in precipitation) to reduce productivity.</td>
</tr>
<tr>
<td>Rocky Mountain/Pacific Southwest</td>
<td>Ponderosa pine</td>
<td>Borers, bark beetles/drought</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fire/drought or lightning</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Vulnerability:</strong> Resistant to near-term climate change, though productivity may decrease</td>
</tr>
<tr>
<td>Pacific Northwest (Coastal)</td>
<td>Douglas fir</td>
<td>Most stressors not strongly weather related</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Vulnerability:</strong> Resistant to near-term climate change, though productivity may decrease</td>
</tr>
<tr>
<td>Alaska</td>
<td>Spruce</td>
<td>Spruce beetle/warm weather (speeds insect development), moisture problems, erratic freezes</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Vulnerability:</strong> High potential for rapid effects because climate plays pervasive role</td>
</tr>
</tbody>
</table>

From Smith, 1993.

III. PERTURBATION OF THE NITROGEN BIOSOCHEMICAL CYCLE

The atmosphere is 78 percent by volume elemental nitrogen (N₂). The N₂ molecule is extremely stable and only very specialized pathways allow this nitrogen to be made available to the biota. Elemental nitrogen is incorporated into chemically bound forms (fixed) by biochemical processes mediated by microbes. High-energy lightning discharges can produce nitrogen oxides. The nitrogen incorporated in the biota is mineralized to the inorganic form during the decomposition of biomass. The production of gaseous N₂O and N₂O3 by microorganisms and the evolution of these gases to the atmosphere completes the nitrogen cycle through a process called denitrification (Fig. 2). While elemental nitrogen and elemental oxygen (O₂) do not react at ambient temperature and pressure conditions in the troposphere, they do react at the elevated temperatures of combustion processes. As a result, all anthropogenic activities involving combustion, from industrial and power generation combustion to automobile and truck engine combustion, function to inject large amounts of nitrogen oxides (NO and NO₂) to the lower atmosphere. These oxides are extremely important pollutants and have substantial potential significance for biological diversity. Nitrogen oxides present risks to the biota directly in the form of acid rain and nitrogen saturation and indirectly as they act as precursors for the formation, along with hydrocarbons, of ozone in the troposphere.
A. Acid Rain

Acid rain is appropriately described as an old environmental problem with a new image. Acid rain, more than any other environmental contaminant, has focused societal concern on ecosystem toxicology. Natural rain, including precipitation in relatively clean or unpolluted regions, is naturally acidic, with a pH in the range of 5.0 to 6.0. This natural acidity results from the oxidation of carbon oxides and the subsequent formation of carbonic acid. Formic and acetic acids, originating primarily from natural sources, may also contribute minor amounts of acidity to precipitation.

In regions downwind from electric-generating power stations employing fossil fuels, industrial regions, or major urban centers, precipitation can be acidified below pH 5.0. Precipitation with a pH less than 3.0 is designated acid rain. This human-caused acidification of precipitation results primarily from the release of sulfur dioxide (SO₂) and nitrogen oxides (NO and NO₂) from smokestacks and tailpipes. The sulfur and nitrogen oxides are subsequently oxidized to sulfate (SO₄²⁻) and nitrate (NO₃⁻), hydrolyzed, and returned to earth as sulfuric (H₂SO₄) and nitric (HNO₃) acids. This additional acidification of precipitation by human activities can readily reduce the pH of precipitation in downwind regions to between 4.0 and 5.0 on an annual average basis. Individual precipitation events that have a pH in the 3.0 to 4.0 range are not uncommon.

The atmosphere deposits acidity onto the landscape both during and in between precipitation events. In the latter case, termed dry deposition in contrast to wet deposition, the acids are delivered in the gas phase or in association with fine particles (aerosols). Acid deposition is a term that includes acid delivery in the form of precipitation (rain, snow, fog, and cloud moisture) plus dry deposition. In view of the importance of both wet and dry deposition in acid transfer from the atmosphere to the biosphere, acid deposition is a much more appropriate descriptor than acid rain.

B. Aquatic Ecosystems

The pH range for undisturbed lakes and streams with full compliments of native biota, especially selected finfish species important to humans, typically is within...
the range of 6 to 8. Lakes and streams with a pH < 6 do occur naturally. Land use practices, for example, conversion from agriculture to forest in the watershed of a lake can, over time, also cause lake acidity to increase. Aquatic resources acidified from the input of acid rain are also well documented, particularly in the northern United States, eastern Canada, and Scandinavia.

Fortunately, most surface waters in the United States have sufficient buffering (acid-neutralizing) capacity to be resistant to acidification via atmospheric deposition. In the instance of high acid deposition inputs and poorly buffered aquatic systems, however, declines in water column pH can be caused by acid rain. In the mid-1980s a National Surface Water Survey was conducted in areas of the United States known to contain lakes and streams with little capacity to neutralize acids. This survey identified the Adirondacks and mid-Atlantic Highlands region as having both sensitive aquatic resources and high levels of acid deposition.

The adverse response of aquatic biota to acidification is very well documented (NAPAP, 1991). Sensitive species may be stressed or lost at small increases in acidity. Acid-sensitive mayfly and stonefly species can be impacted at pH levels near 6.0 and sensitive finfish species, for example, fathead minnows, can be lost at pH 5.6 to 5.9. Acid-sensitive phytoplankton, zooplankton, and benthic macroinvertebrate species may also be lost. Acid-tolerant species may increase significantly with acidification. Survival of fish in acidic waters is primarily related to pH level, inorganic chemistry of aluminum, and the concentration of calcium. Other relevant considerations include food web effects, spatial and temporal variation in exposure to acidity and aluminum at various life stages, and behavioral avoidance. Critical pH levels for selected aquatic organisms are presented in Fig. 3.

C. Forest Ecosystems

The adverse impact of acid deposition on forest systems is focused on the montane forests of eastern North America. Prevailing wind patterns and the importance of cloud moisture delivery of acidic materials places these high-elevation forests at special risk. Both morbidity and mortality of large numbers of forest trees are

![Figure 3: Critical pH levels for selected aquatic organisms where the American toad, wood frog, leopard frog, and spotted salamander are shown in the embryonic stage, and the crayfish, mayfly, clam and snail are selected species. From NAPAP (1991).]
stresses on aquatic species. Excess nitrogen loading of surface waters can impose levels of nitrate in stream waters of forest catchments. Nitrogen saturation of forest soils may lead to high nitrification, and possibly tree morbidity and mortality. Imbalances in foliar nutrient concentrations, enhanced forest systems and lead to decreases in foliar biomass, indicate that nitrogen saturation can be detrimental to saturated. If sustained, evidence has been presented to the system to enter a condition designated nitrogen vegetation and microbial nitrogen demand and cause nitrogen availability may reach a level that exceeds the nitrogen in acid rain and exhibit increased growth and productivity.

If high levels of nitrogen deposition persist, however, nitrogen availability may reach a level that exceeds the level of atmospheric input and normal dynamics of internal nitrogen cycling, forests can be fertilized by nitrogen in acid rain and exhibit increased growth and productivity.

For forests in temperate latitudes, the input of nitrogen via atmospheric deposition may approximate 0.5 kg N ha/yr in clean atmospheric conditions. This low input combined with acidic organic soils in cool climates with high lignin content and slow rates of decomposition and, therefore, slow internal nitrogen turnover cause most forest systems to be nitrogen deficient. Numerous forests provided with supplemental nitrogen exhibit enhanced growth rates indicating that they were nitrogen limited prior to fertilization. In forest locations with less than clean atmospheric conditions (acid rain), forests may receive up to 27 kg N ha/yr if both dry and wet deposition sources are inventoried. At elevated levels of atmospheric input and normal dynamics of internal nitrogen cycling, forests can be fertilized by nitrogen in acid rain and exhibit increased growth and productivity.

When hydrocarbons, aldehydes, or other reactive atmospheric constituents are present, however, they can form peroxy radicals that oxidize the nitric oxide back to nitrogen dioxide. With reduced nitric oxide available to react with the ozone, the latter may accumulate to relatively high concentrations.

High concentrations of ozone are toxic to vegetation. Morbidity and mortality of agricultural and forest plants caused by ozone are well documented. California provides numerous examples of ozone stress on wildland ecosystems. In southern California the predominant native shrub land vegetation consists of chaparral and coastal sage scrub. The former occupies upper elevations of the coastal mountains, extending into the North Coast ranges, east to central Arizona, and south to Baja California; the latter occupies lower elevations on the coastal and interior sides of the coast ranges from San Francisco to Baja California. Westman (1979) applied standard plant ordination techniques to these shrub communities to examine the influence of air pollution. The reduced cover of native species of coastal sage scrub documented on some sites was statistically indicated to be caused by elevated atmospheric ozone. Sites of high ambient ozone were also characterized by declining species richness.

Ponderosa pine is one of five major species of the "mixed conifer type" that covers wide areas of the western Sierra Nevada and the mountain ranges, including the San Bernardino Mountains, in southern California. Mortality from elevated nitrogen deposition has been extensive. White fir has suffered slight damage, but scattered trees have exhibited severe symptoms. Sugar pine, incense cedar, and California black oak have exhibited only slight foliar damage from oxidant exposure. A 233 ha study block was delineated in the northwest section of the San Bernardino National Forest in order to conduct an intensive inventory of vegetation present in various size classes to evaluate the healthfulness of the forest. Ponderosa pines in the 30 cm diameter class or larger were more numerous than other species of comparable size in the study area. These pines were most abundant on the more exposed ridge crest sites of the sample area. Mortality
POLLUTION, OVERVIEW

of ponderosa pine ranged from 8 to 10 percent. The loss of a dominant species in a forest ecosystem clearly exerts profound change in that system. Miller (1989) concluded from his investigations that the lower two-thirds of the study area will probably shift to a greater proportion of white fir. It was judged that incense cedar will probably remain secondary to white fir. Sugar pine was presumed to be restricted by lesser competitive ability and dwarf mistletoe infection. The rate of composition change was deemed dependent on the rate of ponderosa pine mortality. The upper one-third of the study area, characterized as more environmentally severe due to climatic and edaphic stress, supports less vigorous white fir growth. Following the loss of ponderosa pine in this area, sugar and incense cedar may assume greater importance. Miller judged, however, that the natural regeneration of the latter species may be restricted in the more barren, dry sites characteristic of the upper ridge area. California black oak and shrub species may become more abundant in these disturbed areas. Generally, the data support the hypothesis that forest succession toward more tolerant species such as white fir and incense cedar occurs in the absence of fire. In the presence of fire, pine may be favored by seedbed preparation and elimination of competing species.

The changes in forest composition caused by ozone in this southern California forest have created a management concern, as well as ecological change, because the forest is intensively used as a recreational resource and the loss of ponderosa pine is judged to reduce the aesthetic qualities of the forest.

IV. PERTURBATION OF THE SULFUR BIOGEOCHEMICAL CYCLE

The natural sulfur cycle is complex in that it involves several gaseous species, poorly soluble minerals, and several species in solution (Fig. 4). Sulfur enters the atmosphere from natural sources as hydrogen sulfide (H₂S) from active volcanoes and the decay of organic matter in anaerobic environments (swamps, tidal flats), sulfur dioxide (SO₂) from active volcanoes, and particles of sulfate salts (e.g., ammonium sulfate) from sea spray. Approximately one-third of all sulfur compounds and 99 percent of the SO₂ reaching the troposphere come...
from human combustion activities. The combustion of coal and oil for the production of electricity and the smelting of metal-bearing ores have historically been major sources of SO₂ to the atmosphere.

The industrial and energy sector release of SO₂ over multiple decade timescales has reduced biological diversity in local environments surrounding point source facilities in many parts of the world.

In North America, for example, we can cite the cases of Ducktown, Tennessee, and Sudbury and Wawa, Ontario. The Ducktown, Tennessee, area consists of approximately 243 km² (60,000 acres) and was originally covered with southern deciduous forest. Mining operations were initiated in the basin in 1850 and smelting operations were most active between 1880 and 1895. By 1910 gross forest simplification resulting from excessive sulfur dioxide had created three new vegetative zones surrounding Ducktown. In a 27 km² (10.5 mile²) area closest to the source vegetation was devastated and largely eliminated. All trees and shrubs were destroyed, and only a few, isolated islands of sedge grass occurred in the outer portions of this zone. A belt of grassland ecosystem, 68 km² (17,000 acres) in size, surrounded the barren zone. The principal grassland species was broomsedge. A transition zone of somewhat indefinite boundary and consisting of approximately 120 km² (30,000 acres) was located beyond the grassland. Few trees were located along the inner edge of the transition zone. Sassafras, red maple, sourwood, and post oak were common in the middle of the transition forest. The uninfuenced forest beyond the impact of the smelter consisted principally of mixed oaks, hickory, dogwood, sourwood, black tupelo, and some eastern white pine. The distance of vegetative impact extended 19 to 24 km (12 to 15 miles) to the north and approximately 16 km (10 miles) to the west of the smelter. Eastern white pine damage was recorded 32 km (20 miles) from the industry.

Sheet and gully soil erosion has been excessive in the acutely damaged inner zone. Micrometeorological changes in the inner zone relative to the surrounding forest have been substantial; summer air temperature averages are 1 to 2°C higher, while the winter air temperature averages 0.3°C lower, the soil temperature is 1°C higher in the summer; the wind velocities are 5 to 15 times higher and rainfall is consistently lower.

Reforestation of Copper Basin has not been rapid due to severe erosion, low soil nutrients and moisture, and high winds. Recent reforestation efforts have employed both pine and hardwood species (Smith, 1990).

In the Sudbury area, a century of sulfur dioxide fumigation, copper and nickel particulate deposition, fire, soil erosion, and increased frost action have interacted to create 10,000 ha (25,000 acres) of barren land and 36,000 ha (89,000 acres) of stunted, open birch-maple woodland. Three large nickel and copper smelters have historically discharged several thousand tons of sulfur dioxide daily into the surrounding atmosphere. Sulfur dioxide emissions from this area have historically approximated 10% of the North American sulfur dioxide total and 25% of the smelter total. Extensive simplification of the mixed boreal forest ecosystem surrounding this region has occurred primarily via the mortality of eastern white pine throughout a 1865 km² (720 square mile) area to the northeast of the Falconbridge, Copper Cliff, and Coniston smelters. Acute impact on balsam forest has been recorded in excess of 40 km (25 miles) from the source of sulfur dioxide. Red oak, red maple, and red-berried elder are more tolerant and may exist in disturbed forests as close as 1.6 km (1 mile) to the smelters. Mortality and mortality of forest trees in the Sudbury region continued to spread at least until the construction of the world’s tallest smokestack of 403 m (1250 feet) at Copper Cliff in 1972 and the closing of the Coniston smelter also in 1972. Soil erosion has followed the destruction of surrounding forest ecosystems. Rainfall has been made highly acidic, commonly less than pH 3.0 in the early 1970s. Elevated nickel and copper concentrations in soils have been recorded to distances of 50 km (31 miles). Eroded sediment, contaminated with trace metals, has contaminated area lakes and water resources. This contamination, along with acidification of surface waters, has resulted in numerous aquatic species extirpations.

Forest destruction in the vicinity of an iron smelter in Wawa, northern Ontario, exhibits a more discrete pattern, one smelter relative to three; and a shorter history of impact resulting in less equilibration of vegetative response than in Sudbury. The Wawa smelter initiated operations in 1939 (significantly expanded in 1949) and has released as much as 100,000 tons of sulfur dioxide annually to the surrounding atmosphere. Negative impact is primarily confined to a strip north-east from the plant in the direction of the prevailing wind. Symptoms of sulfur dioxide damage may be observed for at least 32 km (20 miles) to the northeast. The mixed boreal forest in the Wawa area consists mainly of white spruce, black spruce, balsam fir, jack pine, white cedar, larch, and white pine in the dominant layer. Mountain maple and Pyrus decora are common in the understory.
V. TRACE POLLUTANTS OF GLOBAL IMPORTANCE

Persistent organochlorine compounds, including the polychlorinated biphenyls (PCBs) and organochlorine pesticides, are among the most important anthropogenic contaminants with global status. These pollutants biomagnify in food webs and have resulted in acute and subtle ecotoxicological effects in aquatic and terrestrial organisms. Despite increased regulation and reduced use, urban and industrial areas remain a significant source of PCB pollutants. The insecticide DDT has seen reduced or terminated use in Canada, the United States, and western Europe, but increased use in Mexico, Central America, India, and eastern Europe. Both PCBs and DDT (including its breakdown products DDD and DDE) can enter the atmosphere via volatilization, or in combination with aerosols, and be transferred from the atmosphere to the biosphere via both dry and wet deposition. Long-distance atmospheric transport and subsequent deposition have been observed to contaminate numerous ecosystems remote from primary sources including eustarine and lacustrine systems. Less research has been directed to documenting contamination of terrestrial systems, especially forests.

Northern and boreal forests develop a thick organic forest floor and in recognition of the tendency of trace organic pollutants to bind to the surface (adsorb) or penetrate (absorb) into soil particles, it is especially important to examine these organic soils for trace pollutants. Such an examination of forest floors of remote northern New England hardwood and montane forest soils has revealed the presence of numerous PCBs and DDT and its breakdown products (Smith et al., 1993). While the soil and water column concentrations of these pollutants are relatively low, their low water solubility but high lipid solubility allow the processes of bioaccumulation and bioconcentration to elevate these potential toxics to levels sufficient for adverse biological effect (Allen-Gil et al., 1997a, 1997b). These elevated concentrations can cause change in community composition (Ferraro and Cole, 1997).

VI. CONCLUSION

Human beings have—through their efforts to generate electricity, manufacture things, and transport themselves and their goods—managed to alter major biogeochemical cycles of the earth. In addition, the manufacture and use of a variety of environmentally persistent chemicals have resulted in the global circulation of materials with potentially biologically toxic properties. Perturbation of the carbon cycle via increased combustion and deforestation and the resulting global warming presents a high risk to biological diversity. Continued global warming will result in species extinctions and possible extinctions as we progress to the middle of the twenty-first century. Changes in the nitrogen cycle caused by human activities are less impressive than the alteration of the carbon cycle but still have the extended-term potential to change both terrestrial and aquatic ecosystems and to result in species losses. Changes in the sulfur cycle, while substantially less widespread than alterations in the carbon cycle, have proven historically to have been sufficient to cause numerous local extirpations in the vicinity of point sources.

The atmosphere, while highly structured (layered) horizontally, has no vertical partitions. As a result, the injection of chemically persistent toxics anywhere in the world means they may begin a journey of global circulation and eventual deposition. Accumulation of these toxics via environmental media or food-web exposure can place certain species at risk of loss from receiving environments. Pollutants of global importance and the ability of humans to alter fundamental biogeochemical cycles must be recognized as important mechanisms for the loss of biological diversity. Efforts to monitor the trends of both pollutants and major chemical cycles along with species at risk must be made a high priority for contemporary environmental management. The loss of species and communities must be recognized as a primary reason for increased international efforts to reduce the production and release of pollutants.

See Also the Following Articles

ACID RAIN AND DEPOSITION • CARBON CYCLE • CLIMATE CHANGE AND ECOLOGY • SYNERGISM OF • GREENHOUSE EFFECT • NITROGEN AND NITROGEN CYCLE

Bibliography


POPULATION DENSITY

Brian H. McArdle
University of Auckland

I. Measuring Population Density
II. Spatial Patterns
III. Determinants of Spatial Density Variation
IV. Consequences of Spatial Density Variation
V. Temporal Patterns in Density
VI. Determinants of Temporal Density Variation
VII. Consequences of Temporal Density Variation
VIII. Transspecific Patterns
IX. Conclusions

GLOSSARY

autocorrelation Densities from adjacent samples are likely to be more similar than are ones far apart.

compound Poisson distribution A family of probability distributions of numbers per area where some areas have a higher expected density than others.

Poisson distribution The probability distribution of numbers per area if the organisms are distributed completely at random in space.

POPULATION DENSITY can be simply defined as the number of organisms per unit area/volume. Its importance to ecologists is that it relates directly to population size (multiply by the total area/volume and one obtains total numbers). Its importance to organisms is that it defines (to some extent) the number of individuals with whom an individual can (or has to) interact. However, despite such a simple definition, it remains a difficult concept. This is clearly evident when we consider why people measure it.

I. MEASURING POPULATION DENSITY

Generally (although inevitably not exclusively), population density of a species of organism is measured either as a feature of a natural (or manipulated) system that responds to the environment or as a predictor (natural or manipulated) or causal agent of some response of the ecosystem.

A. As a Response Variable

Population density is often used as a simple relative measure of how an organism responds to local conditions. If conditions are not good for the species, the density will be low (organisms will have died or moved out of the sampled area), whereas if conditions are good the density will be high (organisms will have reproduced and/or immigrated into the area). In this way, changes in density can provide insight into the natural history of the preferences and tolerances of individuals of the species. Of course, if the species is regulated by density-dependent processes (e.g., mortality or emigration) then the relationship of density with the attractiveness of the environment can be obscured. Even though the environment changes in a positive way, there may be no increase in density.
Sometimes, density can be used as an explicit proxy for population size, which of course is what many ecologists want to know about. This is particularly true in applied ecology (e.g., conservation and fisheries science). Unfortunately, the link between population density and population size is not always direct. Therefore, definitions of rarity that use either population density or species range are likely to be misleading compared with a definition that uses the product of the two.

The main problem lies in defining the area to be sampled. If it includes the entire population of interest, then the density multiplied by the area gives total population size. However, if the area does not include the whole population then this simple calculation does not work. However, it will perhaps give a relative measure of the population so that changes in the population size will be reflected in changes in the density. Unfortunately, density-dependent processes can weaken this link. If, as the population increased in size, the population was unable to expand the area it inhabited, then the density would increase in proportion to the size. However, if the population was able (or driven by density-dependent migration) to expand its range, then the density could remain constant while the population grew. Since range expansion appears to be common, density should only be used as a proxy for population size when the range is constrained, as on islands. In most studies, therefore, density simply gives the number of organisms present in some defined study area. This will seldom correspond to a biological population.

B. As a Causal Agent

In many studies, changes in density are seen as causing changes in the population of the same or other species. For example, many experiments are performed each year in which the density of organisms is manipulated (or simply observed) to investigate the response of individuals in the same species (e.g., intraspecific competition) or other species (e.g., predation or interspecific competition). In such studies, the biological effect of density is generally determined by the number of direct or indirect interactions between individuals (of the same or different species). Density in this case is often acting as a proxy for the number or probability of interactions. There are problems associated with this use of density. In particular, the area has to be defined carefully. If the area for which the density is being estimated includes habitat unsuitable to the study species, then the zeros measured from such sites are not due to sampling error (sampling zeros) but reflect a genuine inability of the species to live there (structural zeros). Including structural zeros in the density estimate would lead to an underestimation of the density that individuals of the same species actually experience. Of course, if it is the response of other species that is of interest, then these areas of habitat may be relevant (other areas may not). Again, the purpose for which the density is to be used determines how one defines the area over which it is to be measured.

There are additional problems. It is well-known that even if one defines the extent of the study area, estimating the number of organisms can be difficult. Perhaps less well appreciated is that natural surfaces can be generally considered as fractal (or at least approximately so). The effective area of a given area of the earth's surface experienced by an elephant is considerably less than that experienced by a mite. The area used to calculate the density of the organism is nearly always anthropocentrically biased. The more similar the organism is to humans with regard to size and habit, the more accurate the estimate of effective area is likely to be, and therefore the more relevant the estimate of density.

Of course, a simple estimate of density can still be used as a relative measure of the effective density. However, it must be remembered that organisms are seldom (if ever) uniformly distributed over a study site. Typically, some parts of the site have a high density and others a low density. If most of the organisms are located in one small part of the site, the average density for the site will reveal little about the density most experience. The marked nonlinearity of most ecological dynamics means that inferences based on the average density are unlikely to be particularly relevant. The situation is similar though more complex when one is interested in the density of one species experienced by another (e.g., in predation or interspecific competition studies). Population density is therefore nearly always used as an approximate proxy for the real abundance measure of interest. The degree to which it can be regarded as relevant in a particular study depends to a large extent on issues of scale.

II. SPATIAL PATTERNS

Within a species' range, we can expect the density to vary with the habitat from high (optimal) to low (suboptimal). The effect of this on the ground depends on the scale at which one is sampling. For example, at the largest scale it has often been noted that the density of a species tends to be highest in the core of its range and lowest at the periphery. However, much local variation can still exist within the species range at a number
of scales down to that of the individual organism (spatial pattern often depends on spatial scale). Similarly, the range and regions of optimal conditions may be temporally variable. For small animals such as insects, the vagaries of wind and weather can lead to short-lived increases in different geographic regions at different times (Fig. 1). Other species (e.g., habitat specialists) tend to be far more predictable. Of course, temporal stability can be dependent on the scale. The spatial pattern might be relatively stable at one scale (e.g., geographic region) but not at another (e.g., locally within habitat patches).

It is worth noting that although we may often refer to the density as patchy, the pattern of variation is often more continuous. The pattern is better shown with contour maps rather than with a tiling or mosaic. The edges of so-called patches may not be real discontinuities but rather areas of lower density. The density within the patches may not be constant. Therefore, although we may often talk in terms of patches, we must think in terms of contours. Of course, thinking in terms of patches is an implicit acknowledgment that the presence/absence data are often more reliable and easily interpreted than abundance or density data. This is because the organism counts produced by most sampling programs are subject to considerable statistical sampling error.

A. Statistical Patterns in Spatial Variation

One of the most common ways of investigating population density is to count organisms in small representative portions (e.g., quadrats) of a larger study area. Then, one estimates the mean density or total numbers in the study site. One can also deduce many characteristics by examining the frequency distribution of the counts per sampling unit.

The most obvious feature of real data is that the frequency distribution of these counts is long tailed and heavily skewed. (Fig. 2: Low densities are common, but very high values sometimes occur. As a general rule, a relatively small number of sampling units contain most of the organisms, and many of the sample units are empty. This indicates a fundamental difference between total number of organisms and the typical density experienced by members of another species (e.g., a predator or competitor) or the typical density experienced by members of the population.

Thus, the total number of animals in the whole area is estimated by the arithmetic mean for total area. However, the typical number of animals found in a quadrant, i.e., the density of the species most commonly experienced by other species (e.g., predators) searching at the same scale as the quadrats under study, is given by the geometric mean, antilog(mean (log [count])), or more commonly by the pseudogeometric mean, antilog[mean (log [count] + 1)]. The number 1 is added to each count to overcome problems with zeros. However, the density most commonly experienced by members of the target species is neither of these. Since most of the animals are in a few quadrats, most experience very high densities. Thus, the best estimate of the effective conspecific density, the average density experienced by a member of the population at the spatial scale of the quadrat, is

\[ \sum N_i / \sum N_i \]

which gives a density that is very different from the other two average densities.

Statistically, there are many distributional models that could fit these kinds of data. The most commonly used are members of the compound Poisson family (e.g., negative binomial and Poisson log-normal). That these long-tailed distributions fit in no way suggests that the mathematical processes that generate them have anything to do with the biological processes that generated the numbers. However, there are a few interesting ideas that can be derived from these distributions that may have biological relevance in some situations, and which when combined provide some understanding of one of the few ecological laws discovered to date (discussed later).

The simplest way of viewing these models suggests that the expected density of organisms at any point in space derives from some long-tailed continuous distribution (a gamma for the negative binomial, and a log-normal for the Poisson log-normal), but that the number actually recorded is from a Poisson with that expected value. There are many ways of parameterizing these distributions, but perhaps the most useful is with the mean (m) and shape parameter (k). The shape parameter defines the relationship between the variance and the mean. The relationship is often displayed as:

\[ \sigma^2 = \mu + (1/k) \mu^2 \]

indicating that the variance of the counts increases approximately as the square of the mean, with the rate determined by the shape parameter. A small 1/k leads to small variances for a given mean, whereas a large 1/k leads to large variances. Thus, large 1/k values suggest high aggregation (some quadrats containing high
numbers of organisms and some quadrats containing low numbers), and low \(1/\hat{k}\) implies low aggregation. Therefore, \(1/\hat{k}\) and related statistics are commonly used as aggregation indices. If the statistical basis of the compound Poisson was in fact correct (i.e., expected density at a point varying spatially, and the number actually collected a Poisson variable), then \(1/\hat{k}\) separates the Poisson sampling error from the underlying distribution of expected density. Indeed, it is the coefficient of variation squared of the underlying distribution of expected density, i.e., the ratio of the variance of the gamma or log-normal with the squared mean:

\[
1/\hat{k} = \frac{(\mu^2 - \mu^2)/\mu^2}{\sigma^2/\mu}
\]

Compound Poisson distributions have a biologically relevant property. If a population of organisms is distributed in an aggregated fashion and if some of them randomly die, the mean density changes but the shape or aggregation parameter of a compound Poisson distribution does not. This is of course as it should be: Random mortality leaves the degree of aggregation unchanged. Therefore, if the compound Poisson family of distributions is appropriate, the basic relationship between mean and variance of animal densities (counts) should be

\[
\sigma^2 = \mu + \frac{1}{\hat{k}} \mu^2
\]

if mortality is operating randomly in space. Of course, there is no need for this to be true (indeed, it might be considered unlikely). Real data might be expected to show a general relationship similar to this, but the degree of aggregation will change with the mean density (possibly due to density-dependent processes). This is the case for virtually all sets of field data. Therefore, the compound Poisson models are not necessarily always appropriate, but they often fit the available data.

### B. Taylor’s Power Law

One of the few ecological laws that have some universality was discovered in 1961 by L. R. Taylor. He worked with populations that had been sampled at many sites on many occasions. He showed that the variance of the densities over the sites on an occasion when plotted against the mean density over the sites on that occasion resulted in a very tight power curve. A plot of log(variance) against log(mean) gave a generally straight line [Taylor’s power plot (TPP)]. This implies that the formula is \(\sigma^2 = \mu^2/\hat{k}\) (Taylor’s power law). As might be expected from the compound Poisson model, the slope of this line is often close to 2.

This plot and the formula are not just applicable to organism abundances. They are commonly used by statisticians searching for transformations to remove
the variance–mean relationship and therefore variance heterogeneity, a problem for many traditional methods of statistical inference. They try to find a transformation that gives a slope of zero to the log(variance)—log(mean) relationship.

Real data are seldom quite so obligingly simple as Taylor's original model implies. Indeed, the dominant pattern is linear in the log–log plot but curved at low densities. Alternative formulae have been suggested to accommodate this. The modified compound Poisson (or Nelder) model \( \sigma^2 = \mu + a\mu^b \) generally performs well, as does a split straight line, although the threshold that the latter implies seems unlikely to some biologists. Since some populations are demonstrably not Poisson even at the lowest densities, alternatives such as \( \sigma^2 = \gamma \mu + a\mu^b \) and \( \sigma^2 = \gamma + \mu + ab^2 \) have also been used.

Although there may be a variety of models, the basic pattern in the data is clear, but the interpretation is not. In all these models, the parameters \( \alpha \) and \( \beta \) have a simple statistical interpretation, but what they imply about the population dynamics is less clear. The parameter \( \alpha \) clearly represents the variance (and therefore the level of aggregation) when the mean density is 1. In the modified compound Poisson (or Nelder) model it represents the value of \( 1/b \) when the mean is 1. If \( \beta = 2 \), then this suggests that the level of aggregation is constant at \( \alpha \) at all densities. Thus, contrary to much that has been written about the use of the TPP, it is \( \alpha \) that tells us about aggregation and not \( \beta \). The slope of the relationship (\( \beta \)) indicates how rapidly aggregation changes with the mean and not the level of aggregation.

To make this clear, we can rewrite the modified compound Poisson model as

\[
\frac{1}{b} = \frac{\alpha^2 - \alpha}{\mu^2} = c a \mu^{2-b}
\]

Aggregation is therefore related to density by a power relationship (i.e., a linear relationship in a log–log plot). Thus, \( \beta - 2 \) indicates how aggregation varies with the mean. Of course, if \( \beta = 2 \) the level of aggregation arguably does not change with density. If \( \beta < 2 \) then aggregation decreases as density increases. This is consistent with density-dependent migration out of higher concentrations into lower ones, evening out density variation as the overall mean density increases. If \( \beta > 2 \) then the organisms are becoming more concentrated—for example, if population growth was restricted to areas of higher density with no diffusion. Of course, it would be very unwise to link any value of \( \alpha \) or \( \beta \) to a particular dynamic process. One reason is that \( \alpha \) and \( \beta \) depend on the particular context (habitat, sampling method and scale, and even the age structure of the population) of the study and not, as suggested by Taylor, on a characteristic slope \( \beta \) for a species and an intercept \( \alpha \) depending on the sampling situation. This makes it difficult to interpret TPPs with any confidence. However, with the spatial variance \( \times \) multiple times data sets originally studied by Taylor, the patterns of variation implied by \( \beta \) can help us understand the biology of the species at least for the site and scale studied.

Interpretation is much more difficult if the TPP is generated by a spatial variance \( \times \) multiple sites sampling scheme. Here, the variance is over samples (e.g., quadrats) taken within a site, but each point on the graph is given by a different site. Since habitat is likely to vary over sites, we are likely to be looking mainly at patterns of habitat variation and utilization. A \( \beta > 2 \) might imply that sites with higher means had greater local concentrations of resources than sites with lower means. A \( \beta < 2 \) might suggest that where the resources were abundant, rather than being locally concentrated they were distributed more uniformly. Of course, the concept of resources is a broad one that includes predator- or competitor-free space and other dynamic features of the environment.

### C. Density–Biomass Thinning Laws

In this article, population density has been defined as the number of organisms per unit area. However, there is no reason why density should not be measured using other properties of the population. Probably the most common alternative to abundance is biomass. It should therefore not be surprising that there is a relationship between the two measures of density. However, biologists might be surprised by the relationship that seems to exist in plants.

1. **Plants:** \(-3/2\) Thinning Rule (Yoda's Law)

In 1963, K. Yoda and coworkers noted that if a cohort of plants were monitored over the course of a growth season, there was a clear negative relationship between the biomass and numerical density as the plants "self-thinned." This relationship seems to be well modeled by a power relationship:

\[
R = c N^{-b}
\]

where \( R \) is the biomass density, \( N \) is the density, and \( c \) and \( b \) are constants. The value of \(-b \) has generally
been thought to be between $-0.3$ and $-0.8$ with an ideal value at $-0.5$. It is generally assumed to be the result of density-dependent mortality and allometric patterns of growth interacting with the way in which trees pack together in the canopy. It is called the $-3/2$ thinning rule because it is commonly formulated in terms of $w$, the biomass per individual, as

$$w = cN^{-\alpha}$$

where $\alpha = b - 1$. The exponent is now between $-1.3$ and $-1.8$ with an ideal value of $-1.5$.

In recent years, its generality has come into question. There does seem to be a generally negative relationship between density and biomass in most investigated species, but there is little evidence that there is any ideal value of $a$, let alone one of $1.5$.

2. Animals: $-4/3$ Thinning Law
In contrast to the empirical origin of the plants’ thinning law, zoologists derived an analogous rule for animals from theory. The postulated relationship is again a power function $w = cN^{-\alpha}$, where $-\alpha = -4/3$ rather than $-3/2$ as in plants. To date, it has been most convincingly shown in salmonid fish but it seems unlikely that all animal species have the same slope.

III. DETERMINANTS OF SPATIAL DENSITY VARIATION
Although the existence of spatial variation in the population density of a species is almost trivially self-evident, the processes that can produce it are many and varied. This reinforces the problems of inferring process from pattern. The mere existence of spatial variation at any given scale is seldom enough to suggest the processes that generated it. There are three main classes of processes that generate spatial variation in density.

A. External Forcing
The most obvious (and at many scales probably the most common) cause of spatial variability in population density is variation in the quality of the available habitat. If the environment varies spatially at any scale, so will the density of many species. A patchy environment usually produces a patchy distribution of organisms. Of course, some species will be affected more than others. There is enormous variation in the tolerance of organisms to variation in environmental factors.

Another external influence (which might be viewed as a special case of environmental variation) might be the distribution or spatial response of predators or competitors. Predator- or competitor-free space can be as significant a resource as food or space with suitable environmental conditions. Other things being equal, higher density can be expected in such areas.

B. Lifestyle Characteristics
Of course, interactions among conspecifics can also lead to spatial variation in density. Social aggregations of many kinds are common in animals and at some scales can lead to variation in density that is independent of the environmental conditions. Conversely, random variation in density at local scales can be reduced by spacing behaviors (such as territoriality).

Some of the possible conspecific interactions are not strictly social. Many species modify the local environment, which makes it more attractive to recruits (e.g., larval settlement). This can also result in the formation of aggregations.

Other biological processes result in strong aggregations. Some species have no choice regarding where they will live (e.g., most plants). Those that start life close to their parent (e.g., many forms of vegetative reproduction, nondispersing seeds, and colonial animals) will stay there. Aggregations can hardly be avoided.

C. Population Dynamics
Perhaps one of the most topical areas of research concerns the possibility that spatial patterns of density variation could be the result of intrinsic dynamic processes—an emergent property of the dynamic system of which the population is a part. Arguments are often taken from complexity or chaos theory and often refer to the behavior of cellular automata models (such as the Game of Life referred to in most popular texts on complexity and chaos theory).

In fact, as one might expect, the tendency of spatially distributed dynamic systems to produce a spatial pattern is related to a very simple property of the system which in ecology has a simple biological interpretation. In 1992, M. J. Phipps postulated that spatial pattern will only be produced in cellular automata models if areas neighboring high-density sites are more likely to achieve high density than other areas. He called this...
the neighborhood coherence principle. In ecology, this predicts that only species that attract conspecifics either by recruitment (e.g., vegetative growth or limited dispersal from place of birth) or by movement (e.g., flocking) or that somehow reduce the death rate in areas near high densities are likely to show intrinsically generated spatial patterning that is independent of environmental forces. All the sophisticated theory results in a trivial prediction that social and other aggregative behaviors are the main reason why spatial pattern emerges in populations for which the environmental factors have little influence. Generally, the more important question is the following: Given that the biology of the animal predisposes it to aggregation, why is it not more obvious? Density-dependent forces such as predation and emigration that tend to reduce high densities will tend to prevent major concentrations. It is the balance of these ecological forces that produces the observed patterns. This is not to say that there may not be other system dynamics that can produce spatial patterning, but they will almost certainly be similarly spatially nonrandom in their action or somehow be relatable to the forcing processes described previously and, above all, they will be biologically obvious. There is nothing mysterious about the action of nonlinear dynamics in producing spatial patterns. It depends on the properties of the system and in a biological context seems generally to agree with common sense.

IV. CONSEQUENCES OF SPATIAL DENSITY VARIATION

A. Population Dynamic Consequences

The existence of spatial variation in population density implies consequences at both the individual and the population levels (and other scales in between). The most obvious is that the state of the individual is frequently influenced directly and indirectly by the surrounding density of conspecifics (intraspecific density-dependent processes). The direct effects may include behavioral changes (e.g., increased propensity for emigration), physiological changes (e.g., increased stress hormones and reduced or enhanced sexual activity), or even morphological changes (e.g., the social phase of locusts). It must be appreciated, however, that the density perceived by an individual will seldom be that measured by researchers. Again, the issue of scale is crucial. Measuring density averages counts over areas and reduces variation at smaller scales. These may, however, be the scales at which the individual's perception operates. Local high densities determine the perceived density per individual and the resulting consequences. Similarly, the indirect consequences of local high density can be very different than expected if one simply examines the average density at a larger scale. The impact of predators, parasites, and pathogens can be very different if the organisms are concentrated rather than uniformly distributed. Spatial variation in density can be as important as differences in overall average density. These considerations have led to an increasing appreciation that population dynamic models should be spatially explicit. A move in this direction started with optimal foraging models, but recently entire populations have been modeled in this way. One class of such models is those that describe metapopulations. A metapopulation is a population of subpopulations. Although this is usually an oversimplified spatial pattern—all organisms are in discrete, fixed clusters or in transit between them—the behaviors are quite different from those in a simple uniform population model. Rate constants (e.g., birth or death rates) averaged over a spatially distributed population will seldom adequately reproduce the behavior of a population. It is worth noting that models that use discrete partitions of space do not behave in the same way as those based on continuous space (i.e., on differential calculus). Therefore, although ecologists are currently attempting to produce more realistic models by incorporating space, the way it is incorporated may influence the conclusions.

B. Genetic/Evolutionary Consequences

Although the population dynamic effects of density variation may be important at a variety of scales, the evolutionary consequences have been considered only at larger scales: A subpopulation of organisms becomes spatially isolated from the rest and allopatric speciation may then occur. Work by D. S. Wilson on multilevel selection theory has incorporated the spatial organization of organisms into groups at different scales to expand the range of evolutionary forces. One genetic consequence of low density and high aggregation in less mobile organisms is the possibility of genetic drift and inbreeding. Such populations are possibly more vulnerable to disease and can show depression of reproduction rates if the aggregations are stable and isolated.

V. TEMPORAL PATTERNS IN DENSITY

When population density is plotted against time, a logarithmic scale is often used because populations can...
fluctuate so widely, over many orders of magnitude, that the arithmetic scale is inadequate. Locusts are an obvious example. However, for many species the magnitude of the fluctuations depends on the temporal and spatial scale over which the density was determined. For example, insect species with a single synchronized generation per year will fluctuate enormously within the year—huge numbers of eggs and first-instar larvae decline throughout the year to much smaller numbers of adults—but might not fluctuate much between years. If one averaged or totalled density over the year, much of the variation would be smoothed out. Similarly, sampling at the same time each year would also not provide information about within-year variability. The spatial scale over which the density is averaged is also crucial. Variation in the density of starlings through time will appear very different when measured at the scale of the bird table, the garden, or the suburb.

A. Shape of Distribution

Although the nature of the distribution of organism counts sampled in space is well-known, far less effort has gone into characterizing the distribution of counts in time. Part of the reason, of course, is the difficulty in obtaining long time series of suitable data. Also, often the data are not independent; rather, they are autocorrelated. The value at one time will be more similar to recent values than those from later times. This makes characterizing the distribution difficult. At this stage all that can be sensibly stated is that, like the spatial distributions, they are discrete and long-tailed, and the members of the compound Poisson family could be used as an initial approximation. However, the autocorrelation often leads to multimodality in real, short time series.

B. Red-Shifted Variance

One observed characteristic of such time series of density data is that, generally, the variability (measured, for example, as $\text{SD}[\log(N)]$) increases with the length of the series. There are at least three phenomena which can produce this effect, and they are not mutually exclusive.

1. Long-term trends: If density is consistently increasing or decreasing over time (as commonly occurs due to human impact and climate change), then its variance will naturally increase as long as the trend continues.
2. Internally generated autocorrelation in the series: Autocorrelation (or serial correlation) is present when the density at the next sample occasion is similar to the current value, perhaps because some of the same organisms are still living (interval between samples $< \text{generation time}$) or because the number living at the next time point depends on the current breeding population. In this situation, short time series cannot vary widely, but as the series becomes longer the variance gradually increases to some asymptotic level. This is of course particularly common with long-lived species that, naturally, we sample at a frequency convenient to us but not necessarily relevant to the species.
3. Externally forced autocorrelation: If the physical environment fluctuated in an autocorrelated fashion then the population’s density might vary in response. Such autocorrelated time series are characteristic of “red-shifted” spectra of environmental variables (characteristic of climatic factors).

C. Taylor’s Power Law

If time series of densities are produced for many sites, then a plot of the variance over time against the mean over time for the sites will provide a close relationship, as occurs for spatial variances and means. Essentially the same models describe this relationship as those used for the spatial TPP. Temporal variance increases with the mean as does spatial variance. Again, the slope is usually approximately 2. The intercept $\alpha$ measures the variability when the average density at a site is 1, and therefore it describes a sort of baseline temporal variability at this spatial and temporal scale. The slope $\beta$ describes how this variance changes for sites that have higher or lower densities. A slope of 2 suggests that there is essentially no change in variability at higher densities. Values higher than 2 indicate that the species is more variable at denser sites. This is a behavior characteristic of species subject to local explosions in numbers at certain sites. A good site is sometimes good and sometimes bad. A bad site is always bad. A slope of less than 2 means that sites with generally high densities tend to be less variable than poor sites. This is consistent with density dependence, although other habitat-driven explanations are possible.

VI. DETERMINANTS OF TEMPORAL DENSITY VARIATION

Although the processes that produce temporal variability in density are similar to those that produce spatial...
variability, they are not identical. The effects of choices of scale tend to be greater. However, the same three classes of factors are discussed.

A. External Forces

Fluctuations in the environment will drive temporal variability of density in many species. Invertebrates in particular are at the mercy of the elements, although vertebrates are demonstrably not immune. The extreme winter of 1962 and 1963 in the British Isles resulted in a marked decrease in the densities of many bird species that took years to recover. The densities of many marine organisms are driven by El Niño-like phenomena (as are many terrestrial organisms). The threshold-like effects of many environmental variables can lead to large fluctuations in organism densities. For example, a species may not breed until the temperature is above some value (which may only happen every few years) or unless rain falls (especially in desert species).

Of course, the consequences for a species of such fluctuations in the physical environment may be indirect. Food, competitors, disease, and predation may also be driven variables, in many cases amplifying or extending the direct effects. Clearly, an organism’s lifestyle (food/nutrient requirements, vulnerability to disease or predation, etc.) can also influence its susceptibility to such forces.

B. Life History Characteristics

Life history characteristics, such as synchronized generations or life stages (perhaps dormant) that are not available for sampling (e.g., the seed bank or the nymphs of the 21-year cicada), can lead to apparently major changes in density. Arguably, the variation is the result of a particular definition of density. Whether such a definition is useful will depend on the question under investigation.

Some species have reproductive strategies that amplify fluctuations in density. Physiological and/or morphological changes (usually triggered by habitat characteristics) can lead to dramatic increases in population growth, e.g., the shift from sexual to asexual reproduction when conditions are good.

An organism’s mobility (or its ability to disperse propagules) can also influence its temporal variability. If it has an aggregated, patchy, spatial distribution (e.g., metapopulation structure) the degree of connectedness between the parts of the population and the degree to which they fluctuate independently can influence the extent of the variation of the whole population.

C. Population Dynamics

By definition, density-dependent processes should be major determinants of the degree of temporal variation. In fact, the extent to which such processes influence density is largely unknown. The problem is twofold:

1. It has proved easier to study the apparent consequences rather than the process. That is, it is easier to show that changes in density over a period depend on densities prior to that period (the statistical detection of density dependence) than it is to show the ecological process that produced the relationship. However, there is little doubt that internal density-dependent processes cause many species to fluctuate less than they might otherwise.

2. The second process is more subtle. Complexity and chaos theory predict that certain combinations of high growth rate and strong density-dependent dynamics can lead to chaotic behavior in the resulting population densities. Since this can lead to apparently random fluctuating densities, it is difficult to separate the effects of density dependence from the results of external forces. Indeed, since these driving variables may be the result of chaotic processes (e.g., the weather or market forces), it is difficult to determine how such density-dependent effects can be distinguished without much more data about environmental factors than is generally available.

VII. CONSEQUENCES OF TEMPORAL DENSITY VARIATION

A. Increased Probability of Extinction

The most widely discussed consequence of low density is the increased probability of extinction. Although it is difficult to demonstrate empirically, it seems obvious that a population whose size (not necessarily density) varies widely is more likely to go extinct (all things being equal) than a population whose size varies less. Once a population reaches low densities within a restricted range, chance events can determine whether it has a future. If the numbers are sufficiently low a single death or failure to breed can lead to eventual extinction. It is worth reiterating that such stochastic extinctions are the result of low population sizes and not densities.

The probability of extinction can (occasionally) be influenced by low density per se. For example, regardless of the population size, at very low densities Allee effects may increase the chances of extinction. That is,
if the density of a sexually reproducing species becomes so low as to affect the probability of locating a mate, then the population can start a slide to extinction from which it cannot escape. Of course, if the population responds by aggregating, then the distinction between local and regional density becomes crucial. Ultimately, it will be the fate of the aggregations, and therefore the total number of organisms, that determines the destiny of the population.

B. Cascade Effects
Of course, fluctuating densities of one species can have knock-on effects on other species. Patterns of density variation affect man’s exploitation of the natural world. Outbreaks of pests or periods of low fish density have economic consequences, and assessing the risk of such events is a lucrative job for well-paid consultants. Such an anthropocentric view ignores the cascade of effects that are likely to propagate up and down the food chain, for example, the changes consequent on fluctuations in the density of kelp on the west coast of North America. Such effects are usually difficult to demonstrate unequivocally.

C. Evolutionary Consequences
Temporal variation in density can produce genetic bottlenecks with consequent loss of genetic variability. However, if the population rebounds rapidly the reduction in heterozygosity may be minimal.

VIII. TRANSSPECIFIC PATTERNS
A. Comparing Densities across Species
It should be apparent that both the measure and the scale of a density must be appropriate for the processes or patterns of interest. In particular, the appropriate area might not be the same for different species. If perceived density of conspecifics is required, the appropriate area would probably depend on the movement patterns of individuals of the different species and would likely differ among species. If perceived density by a predator species is required, then the appropriate area would be defined for the predator and would likely be the same for most prey species. Since the biological relevance of a density depends on the processes being investigated, it is likely to vary between species. As a result, it is extremely difficult to interpret patterns of simple arithmetic density measured across species at some arbitrary scale. This is not to say that such patterns are not interesting (indeed they are sometimes tantalizingly so); however, interpreting the causes and consequences of these patterns will be more difficult because the measure and scale chosen will be to a greater or lesser extent irrelevant to the dynamics of most of the species.

In addition, many other factors complicate such transspecific comparisons, e.g., phylogenetic relatedness which leads to statistical nonindependence and possible artifactual detection of pattern. Transspecific patterns of population density must therefore not be taken at face value. Their meaning is seldom as simple as it appears.

B. Distribution of Density over Species
It has long been observed that organism abundances are approximately lognormally distributed over species, regardless of the size of the sample area. Preston (1962) suggested that this distribution, which he named the canonical lognormal, had a remarkably constant property, $(2(SD_{\log2})^2)^{1/2} = 0.2$ (where $SD_{\log2}$ is the standard deviation of the abundances after transforming using log base 2). This translates into a $SD_{\log10} \approx 1$. In fact, subsequent work has demonstrated that nature is not so kind, but the $SD_{\log10}$ does appear quite consistent. An explanation based on the law of large numbers was given by R. May in 1975, who suggested that this phenomenon is inevitable when large numbers of species are involved. Alternative models have been suggested, notably MacArthur’s broken-stick model and the geometric log series. Although the lognormal model does not imply any ecological generating process, the other models usually suggest such a process. These (and other more complex niche partitioning models) often fit as well as a lognormal, but there is a danger that workers will infer a process from an adequate fit. The amount of information in a single sample frequency distribution is clearly not sufficient to distinguish between generating processes, especially given the difficulties of defining and comparing densities across species.

C. Transspecies $\sim 3/2$ Thinning Law
The $3/2$ thinning law has a transspecies extension (although the word thinning seems inappropriate). If we plot the maximum recorded density of species against their biomass density we should get a classic allometric power curve $w = cN^{a}$ with an exponent $a$ of close to $-0.5$. Enquist et al. (1998) suggested that for the curve $a$ is closer to $-0.333$ ($-4/3$ thinning law
like that suggested for the intraspecific relationship in animals). They suggest a theoretical support for this value.

The existence of a negative relationship between body size and density in animals has long been accepted. However, research suggests that there are two main shapes to the relationship on a log–log plot: linear negative (Fig. 3a) and polygonal (Fig. 3b). Other forms, usually nonlinear with a generally negative slope, have occasionally been suggested for other groups, but currently it is difficult to assess their generality.

The way in which the densities are estimated clearly influences the form of the relationship; in the more than 500 studies investigated by Blackburn and Gaston (1997), the scale of the area considered ("local" versus "regional") was the most important factor. This reinforces the idea that until we know and understand the nature of the density that is being measured we cannot interpret any relationships observed with which it is involved.

The linear negative relationships had highly variable slopes and there was no tendency for them to cluster around \(-0.333\). Indeed, all the slopes (including manifestly polygonal ones) had a mean of \(-0.51\) and a mode near \(-0.75\).

D. Relationships with Other Traits

1. Breeding Systems
Charactertically low-density species (especially plants) rely less on outcrossing and sexual reproduction in general than do more common species.

2. Reproductive Investment
It has been suggested that rare species (restricted range or low density) produce smaller numbers of propagules per unit time, although they often have longer life spans.

3. Dispersal Ability
In general, studies have shown that rare species (with restricted ranges and lower regional density) have poorer dispersal abilities than more common ones.

4. Homozygosity
As expected from theory, there is ample evidence that less dense species have lower genetic diversity than more common species.

5. Resource Usage
Since density, as previously noted, is often driven by environmental factors, it is not surprising that low-density species are frequently associated with restricted or uncommon conditions and resources.

6. Trophic Status
One of the oldest ecological concepts, the ecological or Eltonian pyramid (Elton, 1927), predicts that the higher a species is in the food chain, the lower its density. There is ample evidence that this is usually true; but there are enough exceptions to remind us that few ecological "rules" have universal generality.

7. Latitude
For many years it has been suggested that density for a species tends to be higher in higher latitudes. (This surely has to be true for blood-sucking insects.)
part, this is a prediction of the density compensation hypothesis, which predicts that in communities of low diversity (e.g., islands) density will be higher. Although there is evidence to support this prediction for islands, the evidence for latitude is more equivocal.

8. Geographic Range

The relationship between abundance and range was first noted (publicly) by Darwin and has occupied many workers since. That it exists is not in doubt (Fig. 4), but whether it implies anything of biological interest is debatable. The following is the central question: Does such a relationship imply anything about differences in life history between rare and common species? It is easy to show that in an environment that varies spatially in its attractiveness or ability to support species (as in the compound Poisson model described earlier) but in which all species have similar properties besides mean density (or at least whose properties are independent of their average density), there must be a positive relationship between density and the proportion of sites at which they are found. Such a relationship therefore does not imply anything about the life history or lifestyle of common compared to rare species.

In order to demonstrate that such differences do exist, it is necessary to compare the observed distribution of organisms with an appropriate null such as that previously discussed. This model is sometimes called the ‘sampling’ or ‘artifact’ model and has been shown to fit some data very well. However, it is not simply a sampling phenomenon, disappearing if a sufficient number of areas are sampled. It is predicted by many population dynamic processes in which rate parameters (e.g., birth rate and death rate per individual) are independent of local density. Efforts to investigate the density–range relationship in the past have been made more difficult by the use of density at sites at which the species is actually present—local density (treating all zeros as structural and therefore ignoring them).

In fact, there is an easier approach to the problem. Since the probability of being found at a site depends on the degree of aggregation as well as on the mean, the null (the artifact model) assumes that aggregation at the chosen scale is constant (or at least that common species have the same distribution, and therefore mean, of aggregation as rare species). A trend showing that common species were less aggregated (lower $1/k$) than rare ones would show that common species were found at more sites than the null suggests, i.e., they have wider ranges than would be expected if they had the same life history characteristics as the rare ones. Such a trend might be considered interesting (although there might be many explanations—some trivial—for such a pattern). A trend showing that they were more aggregated would be even more interesting. It is worth noting that most of the studies performed to date use either snapshots of the spatial pattern of species or maps accumulated over some extended period of time. Both of these ignore the temporal variation of range (and mean density) that is typical of many groups of organisms. If such temporal information is available, examining each species separately and demonstrating life history phenomena for each and then searching for a trend for those over mean density would be more productive. I discuss this in more detail in the next section.

9. Transspecific TPPs

Although Taylor's power plots were originally defined for single species, it is possible to plot variance mean relationships across species, with each species contributing a single point. Such plots are clearly difficult to justify if the area over which the data are collected is not equally relevant to all species and the means and variances used are not truly representative of the species. However, if these problems are ignored, then the transspecific TPP can address the density–range problem by plotting the level of aggregation against mean density. The null described in the previous section now becomes equivalent to a TPP slope ($\beta$) of 2 or a horizontal line ($\beta = 2$) on a log($1/k$) vs log(mean) plot. If $\beta < 2$ then aggregation appears to decline as mean density increases; the range of common species is greater than the null would suggest. However, before
accepting an alternative of differences in the life history characteristics of rare and common species, another null must be considered and rejected. If all the species have individual slopes $\beta < 2$ (a very unlikely event in most groups), then a transspecific TPP will tend to have a slope of less than 2 as well, even if the rare and common species have the same life history characteristics. In most situations, it will be impossible to check this since if there were enough information to produce spatial variance TPPs for each species, it would be unnecessary to produce a transspecific plot in which each species was represented by a single point. It would be more informative to fit separate TPPs and search for trends in the parameters.

IX. CONCLUSIONS

The concept of density is an amorphous one. It has to be specifically determined for a particular purpose. The area used for both the sample unit and the study area is crucial and can only be determined after considering what aspect of density is relevant to the current study, and it may even then be difficult to determine since it depends on information about the species’ biology that may not be available (e.g., degree of aggregation and movement patterns). This makes comparisons across species especially difficult to interpret (particularly from published studies). That there is significance in all these patterns seems obvious, but interpreting them rigorously is a considerable (possibly insuperable) challenge.

See Also the Following Articles

*MEASUREMENT AND ANALYSIS OF BIODIVERSITY* • POPULATION DYNAMICS • POPULATIONS, SPECIES, AND CONSERVATION GENETICS

Bibliography


I. Origins and Detection

Populations are usually defined ecologically or genetically. The ecological entity is called a demographic unit, a group of individuals whose population dynamics are not significantly influenced by migration from nearby conspecific groups. In other words, the fluctuations in population size of one group are largely independent of those of other groups. The genetic entity is called a Mendelian population, a genetically distinguishable group of individuals evolving independently of other groups. Demographic units may be Mendelian populations and vice versa, but the two are not necessarily the same. Moreover, populations are not always clear, discrete units (as is the case when defining any taxonomic boundaries). Although the majority of individuals in a population breed with members of the same population, some individuals may move and interbreed with individuals from other populations, making a clear cutoff between most populations impossible. Here, a genetic definition of a population is used—that is, the groups must be significantly genetically different from one another.

Just as populations can be defined in two general ways, so can population diversity. Population diversity can refer to the amount of genetic divergence between two populations. Thus, two populations that are situated nearby and exchange genetic material frequently are not as diverse as two populations that are completely isolated from one another, perhaps because of a geographic barrier. Population diversity can also be defined

GLOSSARY

allele One of several forms of a gene.
chromosome A long, threadlike structure in the nucleus of a cell composed of DNA and protein.
conspecific Of the same species.
gene flow The incorporation of genes from one group of individuals into the gene pool of another group.
gene pool The total of all genes in a group of individuals.

EVERY SPECIES IS COMPOSED OF ONE OR MORE POPULATIONS—localized groups of individuals of the same species. Population diversity is the variety of these groups. This article surveys the origin and detection of genetic differences between populations, discusses the importance of populations to humanity, and reviews estimates of the extent of global population diversity and human-driven modifications of this diversity.

Jennifer B. Hughes
Stanford University

I. ORIGINS AND DETECTION

Populations are usually defined ecologically or genetically. The ecological entity is called a demographic unit, a group of individuals whose population dynamics are not significantly influenced by migration from nearby conspecific groups. In other words, the fluctuations in population size of one group are largely independent of those of other groups. The genetic entity is called a Mendelian population, a genetically distinguishable group of individuals evolving independently of other groups. Demographic units may be Mendelian populations and vice versa, but the two are not necessarily the same. Moreover, populations are not always clear, discrete units (as is the case when defining any taxonomic boundaries). Although the majority of individuals in a population breed with members of the same population, some individuals may move and interbreed with individuals from other populations, making a clear cutoff between most populations impossible. Here, a genetic definition of a population is used—that is, the groups must be significantly genetically different from one another.

Just as populations can be defined in two general ways, so can population diversity. Population diversity can refer to the amount of genetic divergence between two populations. Thus, two populations that are situated nearby and exchange genetic material frequently are not as diverse as two populations that are completely isolated from one another, perhaps because of a geographic barrier. Population diversity can also be defined
as the number of populations in an area. Thus, one can consider population diversity at many scales, for instance, the number of populations in a habitat, on a continent, or on the entire planet. This definition of population diversity—the total number of populations in an area—is used here.

The following sections review the processes by which conspecific groups diverge into genetically distinct populations and how genetic differences are detected.

A. Variability within Species

Much biodiversity exists within a species. This variation occurs at three different levels. First, there is genetic variation within individuals of sexually reproducing species. Individuals carry multiple copies of most genes (two copies in diploid species such as humans), and these copies (alleles) may be different. Second, there is genetic variation between individuals. Except in the case of identical twins, no two individuals will have the same alleles for every gene. Finally, there are genetic differences among populations. A population may contain alleles that other populations do not have and vice versa. In addition, two populations may differ in the proportion of individuals that carry alleles shared by both groups.

B. Processes of Divergence

Individuals are not usually spread out evenly within a species’ range but are clustered into localized groups. When these groups have limited genetic exchange, they may diverge into genetically distinguishable populations. (Methods to detect genetic divergence between populations are discussed in the next section.) There are many mechanisms by which populations may diverge, including mutation, genetic drift, and natural selection. These agents of microevolution can result in genetic differentiation by acting on populations at different rates or in different directions.

1. Mutation

Mutations change the gene pool of a population by substituting one allele for another. Generally, mutation is not a strong force in genetic divergence between populations because it occurs so rarely at any particular gene. In other words, frequencies of alleles shared between groups are not likely to change dramatically through mutation. If a new allele is produced by mutation and increases in frequency in a group, however, then genetic divergence between groups results from the presence of the unique allele in one group and not in another.

2. Genetic Drift

Gene frequencies may change from one generation to the next simply because of chance. Particularly in a small population, random sampling error in the reproduction of alleles may contribute to changes in allele frequencies from one generation to the next. For instance, an allele that is only represented in a few individuals may be completely lost in the next generation if those few individuals by chance fail to reproduce or pass down that particular allele to their offspring.

3. Natural Selection

Natural selection is the differential reproductive success of genetically different individuals resulting from the interaction of the individuals’ inherited traits and their environment. Thus, different environmental conditions or biotic communities in a habitat can result in different selection pressures on populations. Some alleles may be beneficial to the reproduction of the individual carrying them in one population, whereas in another population the same allele may be relatively neutral or harmful to the carrier. In other words, selection in these two groups operates at different rates or directions, contributing to their genetic divergence.

4. Factors Influencing Divergence

Two factors, population size and dispersal, greatly influence the rate and importance of the previously mentioned agents of divergence. The smaller the population size, or number of individuals in a group, the greater the effect of genetic drift. The probability of gene frequencies in the next generation being representative of the previous generation decreases as the number of reproducing individuals decreases. Dispersal, or the migration of individuals from one group to another, affects the amount of gene flow between groups. When an individual successfully reproduces in another group, the genetic composition of the new group becomes more similar to the group from which the immigrant originated. Thus, gene flow has the effect of homogenizing genetic composition, thereby decreasing the rate of divergence resulting from the three agents discussed previously.

C. Detection of Divergence

In general, the degree of genetic divergence between groups of organisms is measured by comparing the amount of variability within a group to the variability
between groups. Thus, the genetic diversity embodied by the individuals in each group must be assessed in order to detect genetic divergence. Genetic diversity may be estimated indirectly by measuring morphological characteristics and migration rates or directly by examining the genetic material. A survey of some of these methods is provided in the following sections.

1. Morphological Variation

The variation that is observed among individuals is the result of their genetic composition, their environment, and the interaction between these factors. Thus, morphological variation will in some part represent underlying genetic variability, but it will also be intertwined with variation caused by environmental factors. Nevertheless, in the absence of direct genetic methods, morphological variation of a suite of traits is often used to resolve the population structure of a species. There have been attempts to tease apart genetic and environmental factors by using breeding experiments. In these studies, the environment is kept constant and morphological traits of parents and offspring are measured. One can then estimate the heritability, or the variation resulting from the genetic background, of a trait in the experimental conditions.

2. Migration Rates and Population Size

The migration rate of reproductive individuals into a group gives an estimate of the amount of gene flow, or exchange of genetic material, with outside groups. The lower the gene flow into a group, the greater the impact of genetic drift. The other critical factor affecting the degree of genetic drift in a population is its size: The smaller the population, the greater the expected drift. Thus, a combination of migration rate and population size estimates gives an indirect estimate of the amount of genetic divergence that occurs resulting from genetic drift.

3. Protein Variation

A relatively direct method of detecting genetic divergence is to examine a sample of proteins (allozymes) from individuals between groups. The composition and structure of a protein are determined by the sequence of nucleotides (the subunits that make up the DNA molecule) comprising that gene. Alleles of the same gene have differences in their sequences, and each allele gives instructions for making a particular protein. The proteins can be distinguished by their rate of movement through a gel medium that is exposed to an electrical field. The movement rates of proteins differ because of different sizes and net charges of the molecules. By sampling many individuals in a group, one can assess the number of unique alleles and their frequencies. Thus, the genetic divergence between groups can be estimated from the variation of a random sample of proteins. The greater the differences in the proteins and their frequencies among groups, the greater the genetic divergence.

4. Chromosomal Variation

A direct method of surveying the genetic material of an individual is to examine the chromosomes. Under a microscope, differences in the number, size, or shape of the chromosomes of an individual can be observed. Differences in the morphological features of chromosomes reveal differences in genetic composition. Below the species level, however, most variation in DNA will not be apparent from chromosome morphology; in other words, the chromosomal variation among populations of the same species is relatively small. Thus, genetic divergence between populations often cannot be detected with this method.

5. DNA Variation

Specific sections of an organism's DNA can be isolated and reproduced using the polymerase chain reaction method. This section of DNA can then be examined by a variety of methods that vary in their resolution. For instance, restriction fragment length polymorphism analysis offers a broad examination of DNA differences. The isolated DNA is exposed to proteins that break the DNA wherever a particular sequence of nucleotides occurs. The length of these fragments can be distinguished by their rate of movement through a gel medium that is exposed to an electrical field (similar to the method used to detect protein variation). The greater the difference in the pattern of fragment lengths, the greater the difference in the piece of DNA. Determining the exact nucleotide sequence of an isolated piece of DNA offers the greatest resolution of genetic composition. The genetic divergence between groups can then be estimated by comparing the similarity of the sequences.

D. Indices of Divergence

A variety of indices are used to evaluate the population structure or the genetic divergence within a species. One of the most common indices reported is Sewall Wright’s $F_{ST}$. This statistic can be calculated from the allele frequencies of a gene or genes and describes the variation among groups relative to the variation within the groups. The greater the variation that can be attrib-
used to differences between groups rather than within groups, the greater the $F_{ST}$ value. Other measures of divergence describe the genetic distance between groups. These measures are indices of similarity or dissimilarity and are calculated using any of the types of data described previously, including protein frequencies and DNA sequences. Groups that are more isolated and for longer periods of time are expected to be more genetically distant.

II. IMPORTANCE

Many of the benefits that biodiversity confers on humanity are delivered through population diversity. These benefits include aesthetic enjoyment, species conservation, discovery and improvement of pharmaceuticals and agricultural crops, replenishment of stocks of economically valued species, and delivery of ecosystem services.

Again, population diversity is defined here as the number of populations in an area. Certainly, characteristics of a population, such as the area it occupies and the number of individuals it includes, will affect the benefits provided by a single population. In general, however, as the number of populations of a species increases, its area occupied and number of individuals increase. Thus, the following discussion focuses on numbers of populations and ignores population characteristics.

A. Aesthetic Value

Populations of organisms, their physical environments, and the interactions between them make up natural ecosystems. People value the aesthetic benefits of the populations in these habitats. For instance, bird watchers enjoy bird populations, hikers appreciate shade trees, and divers seek out reef fishes. Others simply enjoy the scenery created by the sum total of populations in an area.

B. Species Conservation Value

By definition, populations are essential to the conservation of species. Specifically, the number and size of populations influence the probability of species extinction; a species with many large populations is less susceptible to extinction than a species with a few small populations. Migrants from other populations can prevent the extinction of a population by contributing individuals when numbers are low or supplying genetic variation needed to adapt to changing environmental conditions. If local extinction does occur, individuals from other populations can found a new population in the area. The threat of rapid global climate change makes population diversity within a species especially critical; a species with many populations is more likely to include the variation necessary to adapt to new conditions.

C. Genetic Value

Populations of the same species may produce different types or quantities of defensive chemicals, compounds that may be of medicinal or agricultural value to humans. The story of the development of penicillin provides a clear example of the importance of population diversity to pharmaceuticals. The widespread use of penicillin as a therapeutic drug did not occur until 15 years after Alexander Fleming's discovery of the compound in common bread mold. One reason for this delay was a search to find a population of the mold that produced greater quantities of penicillin than did the original population.

Population diversity among wild crop relatives supplies genetic material to agricultural strains. The world's three main crops (rice, wheat, and maize) are widely planted in genetically uniform strains. The emergence of a new disease or pest can therefore threaten large fractions of the harvest at one time. Wild crop relatives contain genetic variation that does not exist in the agricultural populations, including genetic resistance to a variety of diseases and pests. Thus, when an agricultural strain is susceptible to a disease or pest, researchers search for genetic resistance in wild crop relatives. Thousands of populations may have to be tested until one is found that carries the genetic resistance that can be used to protect the crop. For example, when the grassy stunt virus emerged as a serious threat to the rice crop in Southeast Asia in the late 1960s and 1970s, a search for resistant rice varieties was conducted. Only one strain of a wild rice at the gene bank of the International Rice Research Institute was found to resist the virus. Genetic variation in wild populations will also be crucial in providing genetic material to sustain crop yields in the face of human-induced changes in growing conditions.

D. Direct Economic Value

The reduction of economically important species often has direct consequences for local peoples. For instance, the decline of fish harvests leads to the loss of income.
These services include water purification by forests in the region where the services are to be provided. It is necessary not only that there be populations of many species, but that these species exist in diversity in the region of interest. For these services, higher diversity in one area would not be as important as the population diversity at global, regional, and local scales affects the quality and quantity of the services provided. (This article focuses on how the number of populations affects ecosystem services, although the size and density of populations are also important factors.)

1. Global Population Diversity

Higher global population diversity probably enhances the delivery of ecosystem services such as purification of air and water, detoxification and decomposition of waste, generation and maintenance of soil fertility, pollination of crops and natural vegetation, and pest control. Populations deliver these services, and therefore population diversity at global, regional, and local scales affects the quality and quantity of the services provided. (This article focuses on how the number of populations affects ecosystem services, although the size and density of populations are also important factors.)

2. Regional Population Diversity

For many ecosystem services, global population numbers will not be as important as the population diversity in the region of interest. For these services, it is necessary not only that there be many populations somewhere in the world but also that they exist in the region where the services are to be provided. These services include water purification by forests and wetlands and mitigation of floods and droughts by forests. Loss of these services occurs when the populations in forest and wetland habitats are destroyed in one area, regardless of the continued existence of these populations elsewhere.

New York City provides an excellent example of the value of regional population diversity for water purification. The city was renowned for its clean water, which came from the Catskill Mountains 100 miles to the north. Natural purification processes, performed by populations of plants and soil organisms, were sufficient to cleanse the water for most of the city's history, but in recent years land development and related human activities reduced the effectiveness of these processes. In 1996, city water officials floated an environmental bond issue to purchase land, freeze development on other lands, and subsidize the improvement of septic tanks. It is expected that these actions will restore and safeguard the local populations of the organisms that filter and purify the water. If so, an investment of $8–10 billion in natural purification services will have saved city taxpayers $6–8 billion, the additional avoided cost (over 10 years) of building a water treatment plant.

Regional population diversity is also essential for pest control. This function of populations is easy to take for granted, but it is dramatically illustrated when one kind of organism is transplanted to a new environment that lacks populations of predators that usually keep its numbers in check. A classic case is the importation of prickly pear cacti (Opuntia) into Australia by early settlers. In the absence of their normal predators, the cacti spread over approximately 25 million ha in New South Wales and Queensland. About half of the area was so densely covered with the cactus that the land could not be used for farming or ranching, and the costs of poisoning or removing the Opuntia were more than the value of the land. The problem was solved by importing a cactus-eating moth, Cactoblastis cactorum, from the South American homeland of the Opuntia. Once regional populations of the moth were established, the Opuntia populations were decimated. Although the cactus can still be found in Australia, it now occurs only as scattered clumps.

The importance of regional populations of native pollinators for agriculture is made clear by their decline (chiefly a result of pesticides and habitat destruction). For example, for more than 60 crops grown in the United States, farmers have to pay honeybee keepers to bring hives to the fields or orchards to be pollinated. This service is estimated to cost farmers more than $60 million a year and the federal government more than $80 million in subsidies. Problems within the bee-keep-
3. Local Population Diversity
The number of populations at a particular location—the local diversity of species—affects local ecosystem function. In many greenhouse and field experiments, plant productivity has been found to increase with the diversity of plants. Higher diversity also seems to be associated with greater stability of plant productivity. For instance, more diverse grasslands appear to be more resistant to drought and grazing disturbance than less diverse areas. Thus, local population diversity likely influences the amount and variation of services provided by an ecosystem. Also, because regional and global services are performed by an assemblage of local ecosystems, this scale of population diversity will affect the delivery of larger scale services as well. For example, global biogeochemical cycles are probably enormous, information on their diversity is scarce).

Further, traits that affect local ecosystem functioning may differ among populations. Treseder (2000) studied multiple populations of two ecologically important tree species in Hawaii. She found genetically based differences among populations of Metrosideros polymorpha in the decomposability of leaf litter and among populations of Acacia koa in the potential to fix nitrogen. Thus, these populations are not interchangeable if they were replaced with other populations of the same species, the functioning of the native ecosystem would be altered.

III. EXTENT
Although populations are critical to humanity, little is known about the extent of population diversity. Hughes et al. (1997) made a rough estimate of the total number of populations on the planet. They used a Mendelian population definition and restricted the estimate to eukaryotes (although the diversity of bacteria and viruses is probably enormous, information on their diversity is scarce).

The evaluation of global population diversity involved three steps. First, Hughes et al. (1997) estimated the average number of populations per unit area from literature on population differentiation. Second, they calculated the average range size of a species from published range maps. The product of these two numbers yields an approximation of the average number of populations per species. Finally, they multiplied this product by the number of species on Earth to arrive at an estimate of global population diversity.

A. Populations per Area
To quantify the number of populations per area for an average population, Hughes et al. (1997) searched 15 journals for genetic studies on population differentiation. They found 81 articles that provided appropriate data to quantify the number of populations per area. Most of the species were vertebrates (35), followed by plants (23), arthropods (19), mollusks (4), and a flatworm.

For each species, Hughes et al. (1997) determined whether the sampling locations in the studies were for separate populations or a single population. If the genetic differentiation between localities was statistically significant, then it was assumed that all of the samples were taken from one population. If the size of a population was the size of the sampling area. For many species, an intermediate amount of differentiation was found. For example, although not all the samples were significantly different, some clusters of sampling locations were significantly different from others. In this case, the number of significantly diverged clusters was used as the number of populations.

Comparing genetic differentiation across studies is difficult because each study uses slightly different methods for detecting genetic differentiation. Thus, the estimates of populations per area from each study were rounded to the nearest order of magnitude. The average of these estimates yields an estimate of one population per 10,000 km² for an average species. Note that the arithmetic average of orders of magnitude is equivalent to the geometric mean of the untransformed estimates. Therefore, all else being equal, the estimate of populations per area is conservative.

There are several notable biases with this method. First, the species represented give a taxonomically biased sample of the earth’s species diversity. For example, arthropods make up about 65% of the planet’s species, whereas birds account for probably less than 0.01%. However, in this estimation arthropods account for only 20% of the species, whereas birds account for more than 11%. Second, the sampling intensity of each study limits the number of populations that can be estimated. This constraint makes the estimate conservative since in many cases additional sampling in the
study area may have revealed further differentiation. Finally, the molecular markers used in the studies may not always reveal notable differences between groups, again causing the estimate to be conservative.

B. Average Range Size
The average range size of a species was calculated from more than 2400 species range maps for birds, mammals, fish, and butterflies (Hughes et al., 1997). Weighting each of the taxonomic groups equally, the mean range size of a species is 2.6 million km$^2$. The average range size of the most species-diverse group, the butterflies, is 2.2 million km$^2$ and this was conservatively used as the estimate of the average range size of a species.

The shaded areas of species distribution maps almost always include habitats in which populations do not occur. Therefore, estimating species’ ranges from these maps will probably inflate the population diversity estimate. In addition, the majority of species included occurred in temperate regions; however, it is estimated that two-thirds of species diversity exists in the tropics. Because species’ range sizes in some taxa tend to increase toward the poles, the bias toward temperate areas may inflate the estimate of average range size. In contrast, the range maps that were used were restricted to one continent; therefore, the range sizes of intercontinental species were probably underestimated.

C. Species Diversity
The product of the average number of populations per area and the average range size of a species gives an average of 220 populations per species. Multiplying this number by three estimates of global species richness (5, 14, and 30 million) yields estimates of 1.1, 3.1, and 6.6 billion populations globally. This calculation assumes that a species’ range size and its number of populations per area are independent. If these two factors are strongly correlated, however, the estimate may be quite inaccurate. However, it is not known whether a correlation exists, let alone whether the correlation is positive or negative.

IV. HUMAN-INDUCED CHANGES
With the expansion of the human enterprise, human beings are having a significant impact on population diversity. Populations are being both created and destroyed on a large scale as a direct result of human activities. As species are introduced to nonnative habitats, new populations are formed. At the same time, habitat disturbance and destruction are driving the extinction of populations. Although most public and scientific attention to biodiversity is centered on species, a few recent studies have attempted to evaluate the extent to which humans are modifying population diversity.

A. Introductions and Extinctions in Western Australia and California
Hobbs and Mooney (1998) collected data on the deletions and additions of populations in Western Australia and California. They gathered records of species that are threatened, extinct, or introduced in these areas. The number of regional extinctions gives a very conservative estimate of population extinction since most, if not all, of the species originally had more than one population in the region. Similarly, it is reasonable to assume that the threatened species have already lost populations. These species are endangered in large part because of range contractions—that is, the extinction of populations, thereby reducing the geographic distribution of the species. The number of introduced species is also a lower bound estimate of the number of new populations.

For California, Hobbs and Mooney (1998) gathered data for butterflies, amphibians, reptiles, birds, mammals, and plants. Of 8274 total species, 1109 are introduced (13.4%), 71 endangered, and 49 extinct. The endangered and extinct categories include subspecies, which are most certainly genetically distinct populations. For Western Australia, the study reported information for plants, mammals, birds, reptiles, amphibians, and freshwater fishes. Of more than 12,000 plants, 1032 have been introduced, 232 are threatened, and 52 have gone extinct. The percentage of all introduced species is 7.3%, which is lower than that in California. In California, no extinctions have been reported for reptiles and amphibians, and in Western Australia no extinctions have been recorded for birds, reptiles, amphibians, or freshwater fishes.

As mentioned previously, however, these results mask most of the population extinctions. In the groups for which no extinctions were recorded, other studies report significant species’ range contractions. For instance, the loss of native vegetation in Western Australia has resulted in range contractions and population extinctions of species that rely on this habitat.
B. Extrapolating Population Extinctions from Tropical Deforestation Rates

Most estimates of species extinction rates are based on species–area relationships and the rate of habitat loss due to deforestation. The most commonly used model of the species–area relationship is shown in Fig. 1. This model is $S = cA^z$, where $S$ is the number of species, $A$ is the area (or habitat size) in which the species are found, and $c$ and $z$ are constants estimated from empirical studies. This model is in agreement with a multitude of empirical studies on a variety of taxa and scales. With this formula, the rate of species extinction in an area can be approximated from the rate of habitat loss. Figure 1 also illustrates a widely cited rule of thumb: A decrease in habitat area by 90% should result in a decrease in species diversity by approximately 50%.

1. Population–Area Relationship

There is no comparable work describing a population–area relationship. Hughes et al. (1997) reason that population numbers and area probably correspond approximately in a one-to-one manner. The basis for the difference between the species– and population–area relationships is their size: a population occupies a small area relative to a species. When a substantial amount of habitat is destroyed, populations that were encompassed by the area will be lost. In contrast, many fewer species will go extinct because other populations of the species exist elsewhere. Thus, the rate of population extinction will be faster than that of species extinction. If the one-to-one relationship does hold, then when 90% of an area is destroyed about 90% of the populations in the original area will be lost, in contrast to only 30% of all species (Fig. 1).

2. Extinction Rate

The rate of population extinction in tropical forest regions can be estimated with the population–area relationship in Fig. 1. Assuming that tropical deforestation is occurring at 0.8% per year, global population diversity is 3 billion, and two-thirds of all populations exist in tropical forest regions, Hughes et al. (1997) calculate that 16 million populations per year, or approximately 1800 per hour, are being exterminated in tropical forest alone. This rate is three times higher than conservative estimates of species extinction.

C. Impacts of Changing Population Diversity

Impacts resulting from human-caused extinction and introduction of populations are already evident and widespread. The extinction of population diversity results in the loss of the benefits outlined previously, and the establishment of populations of invasive species often disrupts native communities, altering ecosystem functioning or causing the extinction of native species. For instance, approximately 20% of all endangered vertebrates are thought to be threatened by invasive species in some way. Furthermore, it is unclear how often the addition of introduced populations (or “weedy” populations) will compensate for the loss of native population diversity. For services such as pest control, there is abundant evidence that such compensation will be rare. Pesticides destroy populations of natural enemies as well as the intended pest. As a result, obscure organisms, released of predation pressure, often are promoted to pest status when no new weedy species fill the role of the natural predators. Similarly, weedy species are not compensating for the loss of pollinator populations in the United States. For other services such as flood control and soil retention, the potential for substitution by weedy populations, at least in the short term, may sometimes be high. Over the long term, however, the ability of weeds to maintain services is unknown. For example, weedy ground cover on bare slopes may prevent soil erosion, but it may be poor at maintaining soil fertility. Moreover, the capacity for large-scale technological substitu-
tion of ecosystem services by human-made communities of populations appears limited. The Biosphere 2 project is a case in point. Despite hundreds of millions of dollars invested in development and operating costs, scientists failed to engineer a system that could support eight people with food, air, and water for 2 years. The failure dramatically illustrates humanity’s dependency on the life-support services that populations in natural ecosystems provide for free.

D. Ameliorating Changes in Population Diversity

There are many ways in which human modification of population diversity can be slowed and its impacts alleviated.

1. Controlling Invasions

In general, the impact of newly introduced populations on an ecosystem is highly unpredictable. In fact, most invasions do not produce noticeable effects. The invasive species that do have significant impacts, however, can cause widespread ecosystem disruption and economic loss. In theory, removing invasive populations can restore the native ecosystem (unless species have gone extinct). In practice, however, complete removal or even partial control of invasives is rarely successful. One of the few examples of a successful program to control invasive populations is the Opuntia cactus case described previously.

There are three broad methods to control invasive populations: biological, chemical, and manual. One reason that introduced species can be very harmful to an ecosystem is that the species are free from their natural enemies and therefore overabundant. Biological control involves reducing the invasive species’ numbers by introducing a predator or parasite from its native habitat. Chemical controls are poisons (usually synthetic) such as pesticides and herbicides. Manual control entails the physical removal or killing of individuals. All three of these fixes are usually expensive and in the case of biological and chemical controls risk doing more harm than good to the ecosystem. An introduced predator might attack native species, and chemicals may harm native species or human beings. Thus, the best control is to prevent invasive populations from establishing in the first place. Unfortunately, this is an extremely difficult challenge—most introductions are accidental. Furthermore, with the accelerating movement of products and people around the world, the rate of introductions will certainly increase in the future.

2. Preservation

Protecting intact habitat from destruction or degradation is the most straightforward method to prevent the extinction of populations. If population diversity and habitat area are indeed related in a linear manner, then the more land that is preserved, the more population diversity that is protected. In most cases, however, limiting almost all human access to a piece of land is usually not feasible because of competing uses, such as agriculture and urbanization. Nevertheless, these competing uses do not eliminate all prospects of conserving population diversity and the benefits they provide. In fact, in some parts of the world, many populations appear to be surviving in the ‘countryside’—that is, in human-dominated areas (Daily, 2000). Thus, a promising approach involves managing the countryside to prevent further degradation (and perhaps some restoration) of population diversity in these areas.

3. Restoration

Unlike species, populations can be restored and, along with them, at least some of the benefits that they once provided. Although genetic variation unique to the extinct populations will be lost, groups of individuals can be reintroduced to an area and in a relatively short time may evolve significant genetic differences from other populations. Reestablishing populations of a native species is not a simple task, however. Often, the native ecosystem is degraded and little is known about the biological requirements of the species.

See Also the Following Articles

DIVERSITY, COMMUNITY/REGIONAL LEVEL • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • ECOSYSTEM SERVICES, CONCEPT OF • EXTINCTIONS, MODERN EXAMPLES OF • HUMAN IMPACT ON BIODIVERSITY, OVERVIEW

Bibliography


I. Introduction

The term population refers to the members of a single species that can interact with each other. Thus, the fish in a lake, or the moose on an island, are clear examples of a population. In other cases, such as trees in a forest, it may not be nearly so clear what a population is, but the concept of population is still very useful.

Population dynamics is essentially the study of the changes in the numbers of a species at a single location. Understanding the dynamics of a population requires incorporating both within species and between species interactions. This chapter presents a conceptual approach to the study of population dynamics with brief illustrative examples, beginning with single species approaches and finishing with interactions between two species.
will provide indications of what features of changes in numbers are important and what measurements are critical to make, and they will help determine what the cause of changes in population levels might be.

To understand the dynamics of biological populations we start with the simplest possibility and determine what the dynamics of the population would be in that case. Then deviations in observed populations from the predictions of that simplest case would provide information about the kinds of forces shaping the dynamics of populations. Therefore, in describing the dynamics in this simplest case it is essential to be explicit and clear about the assumptions made. We would not be arguing that the idealized population we describe would ever be found, but that focusing on the idealized population would provide insight into real populations, just as the study of Newtonian mechanics provides understanding of more realistic situations in physics.

II. DENSITY-INDEPENDENT DYNAMICS OF A SINGLE SPECIES WITHOUT AGE STRUCTURE

The very simplest description of population dynamics looks only at a single population of a single species. Thus, complications due to the effect of one species on the growth or dynamics of another are initially ignored. Any differences among individuals within the species are assumed insignificant. Thus, differences in age, sex, genetics, spatial location, and other individual characteristics are all ignored. Moreover, we focus on a single population, so any effects of immigration or emigration are ignored as well. We also ignore any effect of randomness, which might arise either from the effects of small population sizes or environmental variability.

We still need to make one more fundamental assumption when describing the dynamics of a population in order to consider what effect members of the population have on each other. The simplest case is density independence—in which the demography of the population does not depend on the size (in numbers) of the population. We assume that the per capita birth rate or numbers of births and of deaths do not depend on the size of the population. This is not the same as assuming that the overall birth and death rates do not depend on the population size, but it is equivalent to assuming that the overall birth and death rates are proportional to the population size.

To describe these dynamics under these assumptions in detail, we need to split our discussion into two cases, depending on the way reproduction occurs. For some species, such as humans or many microorganisms, births and deaths occur essentially continuously through time. For other species, such as many butterflies or annual plants, organisms reproduce once each year and then die. We will treat these two possibilities separately and then turn to more complex cases.

For the case of overlapping generations and continuous reproduction, we can write our description of population dynamics using the verbal equation

\[ \text{Rate of change of population} = \text{Rate of births} - \text{Rate of deaths}. \]

With our assumption of density independence, we can write the rate of births as \( bN \), where \( b \) is the per capita birthrate and \( N \) is the population size. Similarly, the rate of deaths is \( mN \), where \( m \) is the per capita death (mortality) rate. The verbal equation can then be written as the mathematical equation

\[ \frac{dN}{dt} = bN - mN \]

or, letting \( b - m = r \), the intrinsic rate of increase (or "little \( r \)") as

\[ \frac{dN}{dt} = rN. \quad (1) \]

We can determine an explicit solution to this model expressing the population size at any future time \( t \) in terms of population size at time 0 and the intrinsic rate of increase, \( r \). The solution of the model (1) is exponential growth, namely

\[ N(t) = N(0)e^{rt}, \quad (2) \]

where \( e \) is the base of the natural logarithms, \( N(0) \) is the population at time 0, and \( N(t) \) is the population at time \( t \). We will turn to ecological implications of this formula after we discuss the alternate case of growth in discrete time.

For the case of nonoverlapping generations, appropriate for annual plants, we can write down a verbal model as follows:

Numbers next year = Per capita reproduction times numbers this year

In mathematical terms, this equation becomes

\[ N(t + 1) = RN(t), \quad (3) \]

where \( N(t) \) is the population size in year \( t \), and \( R \) is the number of individuals in the population next year.
produced by each individual in the population this year. This model, too, can be solved explicitly,

\[ N(t) = R^t N(0), \]  

which is geometric growth. In fact, if the growth rates in the two cases were matched so that \( R = e^r \), or \( \ln R = r \), the two population descriptions (2) and (4) would provide exactly the same numerical values (at the discrete times \( t = 1, 2, 3, \ldots \)).

The implications of the solutions (2) and (4) are not that we would expect any real population to grow at these rates, but, as emphasized perhaps first by Malthus several centuries ago, that exponential or geometric growth is very rapid. If \( r \) is positive or \( R \) is greater than 1, after relatively short times, the predictions of exponential or geometric growth are that populations will become very large. Thus, exponential growth cannot last, and the fundamental question of population dynamics becomes what prevents exponential growth and what are the consequences for population dynamics of these forces that prevent exponential growth.

Long time density independent growth in natural systems obviously does not occur, with the closest examples following introduction of nonnative species. For example, the number of collared dove in Great Britain increased exponentially from less than 10 to more than 10^6 from 1955 to 1963, but the population essentially stopped growing shortly thereafter. Other introduced species have behaved similarly.

III. DENSITY-INDEPENDENT AGE STRUCTURE

There are many ways to extend our understanding of the dynamics of a single population from the simplest cases of density independent growth we have just covered, and to explore what prevents exponential growth. There are two basic ways to extend our model for a single species (and remain within the context of looking at a single species). We could look at the effects of members of the population on each other and how they affect the dynamics. We could also look at the effect of differences among members of the population on the dynamics of the population. We will first look at the effect of differences among individuals, focusing on differences due to age. Other differences are also important and we focus on age because it is a critical factor, has been well studied, and is easy to understand and measure. Does including the effect of age prevent exponential growth?

It is clear that within many populations some individuals are too young to reproduce, while others are too old. Thus, just looking at the numbers of individuals in the population is not enough to predict the future numbers. When looking at the effects of age structure, we also face the issue that the proportion of females and males may affect the growth of the population.

Since we will still ignore the role of density dependence, we will take the approach of only looking at the female members of the population, and therefore only looking at births of females to females. We thus implicitly assume that males do not affect the dynamics of the population, which may be valid for some natural populations, such as moose, but not all, such as birds with parental care.

The information we want to include is the probability of births (or rate of births in continuous time) to a female of a given age, and the probability (or rate) of death at each age. From this information, and from the current distribution of females of different ages in the population, we should be able to predict the future population. Once again, there are different formulations for this model, depending on whether the model is phrased in continuous or discrete time.

In discrete time, one way to write this model is known as the Leslie matrix model. We first decide on a census interval, for example five years in humans. The population is described by a vector (which is just a vertical list of numbers) in which each entry is the number of females in each age class, 0 to 5, 6 to 10, 11 to 15, 16 to 20, and so on. The length of the age classes and the census times have to be the same. Then from data we determine the average number of births to, and probability of survival of, a female in each age class, over a 5-year interval. These data are presented in a life table, which summarizes the relevant demographic information. Life tables have been constructed from numerous species, ranging from insects, to small mammals, to long lived reptiles, to plants, to humans. The data in a life table can be used as the entries in a matrix, known as the Leslie matrix, in which the first row are the birthrates and the entries under the diagonal are the survival probabilities.

Analysis of this model reveals information about both the rate of growth of the population and the distribution of individuals of different ages. The key concept is the stable age distribution—the unique proportions of individuals of different ages that is preserved through time in the model. Thus, if the population starts at the stable age distribution, it will remain in the stable age distribution. Also, for a population in the stable age distribution, the proportional change in the numbers in each age class, and therefore the population as a
whole, is the same at each time step. This is simply geometric growth once again.

What happens if the population is not in the stable age distribution? It is clear that then the growth rate may change through time and will depend on the ages of the individuals—for example, a human population of all 0–5-year-olds would have no births, while one composed only of 20–25-year-olds could have many births. Yet, under some simple assumptions (essentially that reproduction is not confined to a single age class), if the population does not start in the stable age distribution, it will approach the stable age distribution through time. The result is that geometric or exponential growth occurs again, but only in the limit.

The analogous model in continuous time provides similar results, though there is a difference in the conceptual development. In this case, which can be traced back several centuries to the work of the Swiss mathematician Euler, the focus is typically on the rate of births through time rather than the individuals of all ages at a single time, through the use of a renewal equation. In words, the dynamics are described as

Rate of births at time $t$

$= \sum \text{birthrate at age } j \times \text{probability of surviving to age } j \times \text{birthrate at age } j$.

Using the assumption of exponential growth this equation can be reduced to an equation for the growth rate of the population, the Euler-Lotka equation. This equation (or its discrete time approximation) is the basis for most estimates of the growth rate of populations.

The demography of populations and the ways in which population growth rates are calculated are worth emphasizing because they play a central role in conservation biology. A calculation of the growth rate of a population is often used to determine if the population is endangered. By looking at the effect of changes in different demographic parameters on the growth rate, one can determine which conservation efforts may have the largest effect on the growth rate and therefore the survival of the species in question.

IV. DENSITY DEPENDENCE

A. Continuous Time

We now need to continue our search for processes that will prevent populations from growing exponentially and what the consequences of these limiting processes are for population dynamics. The simplest answer is that as the density of a species increases, either the birthrate goes down or the death rate goes up, because the members of the species compete with each other (due to the fact that the food supply or some other resource is limited). This kind of intraspecific competition is one of the fundamental reasons why populations may not grow exponentially, and we will explore its consequences both for species with overlapping generations and those that reproduce once per year. In both of these cases and all that follows we will be ignoring age structure, for simplicity.

One of the earliest formal descriptions of these dynamics was given by Verhulst, who produced what is now known as the logistic equation. In this equation, the growth rate of the population is described by the product of the exponential growth term as in model (1) and the term representing reduction in per capita reproduction due to interactions with other individuals. In modern terms, the model is written as follows:

$$\frac{dN}{dt} = rN - \frac{N}{K}$$

where the term $K$ is called the carrying capacity of the population.

The logistic equation can be solved explicitly (Fig. 1), but the most important conclusion is that any population reaches the carrying capacity, $K$. This qualitative conclusion also holds for similar assumptions. The search for the mechanisms leading to density dependence, and even direct evidence for density dependence in natural populations, is difficult.

Another aspect of density dependence is to reduce population growth rate at low population levels, perhaps through difficulties in finding a mate. This concept, the Allee effect, provides a lower limit to popula-

![Figure 1](image_url) Explicit solution of the logistic equation with $r = 1$, $K = 100$, and $N(0) = 1$. 
tion size. If the population is too small, it may actually decline and thus can go extinct.

B. Discrete Time

An alternate formulation based on discrete time leads to very different conclusions. Here the idea is that the number of individuals next year will simply be a function of the number of individuals this year, with density dependence. The history of models like this began with examples from fisheries, such as the Ricker model:

\[ N_{t+1} = N_t e^{r(1 - x_t)} \]

The dynamics and behavior of this model are captured in the behavior of a much simpler model, the quadratic model, usually written as follows:

\[ x_{t+1} = r x_t (1 - x_t) \]

where \( x_t \) is a scaled measure of population size and \( r \) is a measure of the rate of population growth. The density-dependent or carrying capacity parameter is absorbed in the scaling of \( x_t \). Beginning with work of Robert May in the early 1970s, much of the interesting behavior of this model and its implications for population dynamics were elucidated.

The dynamics of density dependence in discrete time differ greatly from those in continuous time because of the presence of the implicit time delay imposed by the discrete time framework. As the population tends toward equilibrium, it can overshoot the equilibrium, with striking consequences. The overall dynamics of the quadratic model as a function of the parameters is as follows. If the growth rate is too low, \( r < 1 \), the population will die out. If \( 1 < r < 2 \), the population monotonically approaches an equilibrium.

For larger values of the growth rate, the dynamics differ from the continuous time model. If \( 2 < r < 3 \), the population approaches an equilibrium, but in an oscillatory fashion. For \( r > 3 \), different dynamics emerge. At first the population exhibits stable cyclic behavior of alternating low and high years. For still larger values of \( r \), the population is typically chaotic, meaning that the population fluctuates and exhibits very sensitive dependence on initial conditions.

The search for this kind of behavior in natural populations has proven very difficult, but the attempts have been shown to behave chaotically. Further study of this phenomenon is discussed later in the context of disease dynamics.

V. ROLE OF STOCHASTICITY

The discussion so far has assumed that no chance, or stochastic, factors play a role in the dynamics of the population. In many cases, this is far from the truth. For very small populations, demographic stochasticity, resulting from the chance that many births or deaths in a row might occur, has a large effect on the dynamics. In small populations, extinction may occur by chance even if the expected outcome would be for the population to grow.

A second important source of variability is the influence of the environment. Random fluctuations including year-to-year variability in weather, can cause large changes in the growth rate of a population. These effects need to be taken into account to understand the dynamics of natural populations. Historically, ecologists phrased this as a dichotomy, asking whether density independent (e.g., weather) or density-dependent factors were responsible for controlling populations. The more modern version of this question is discussed later in the context of disease dynamics.

VI. THE SIMPLE TWO-SPECIES INTERACTIONS

A. Overview and Definitions

Thus far, we have focused only on effects of the environment or within a single species, intraspecific interactions, on population dynamics. A fundamental question is, what is the role of interspecific interactions on population dynamics? The consequences of the interactions among many species fall within the realm of community ecology, but we can understand many of the simple effects of species interactions by taking a population dynamic approach, starting with the interactions between two species. This study will begin to highlight the important issue of which interspecific interactions act to prevent exponential growth of natural populations. The study of the interactions between two species has a long history in ecology, beginning with the seminal theoretical work of the American Lotka and the Italian Volterra, and the experimental work of the Russian microbiologist Gause.

The first step in a study of two species interactions
is to classify the interactions, primarily according to the effects each species involved has on the other. We will then outline the dynamical effects of each interaction. An interaction where one species eats another, such as wolves preying upon moose, is called a predator-prey interaction. The predator gains a benefit and the prey is harmed. Interactions between herbivores and plants also are similar, so these interactions can more generally be called exploiter-victim or consumer-resource. Interactions in which one species is harmed and one gains we will treat separately because their dynamics are different or special. One of these is the interaction between disease-causing organisms and their hosts, and the other is the interaction involving insects (parasitoids) that lay their eggs inside developing stage of other insects (hosts).

A second important two-species interaction is between two species that share a resource, such as two species that both eat the same food items. This is a competitive interaction, where each species reduces the growth rate of the other. The resource that is the object of competition can also be nonbiological, such as light or water for plants or nutrients. Other competitive interactions might be more active and involve the production of chemicals, which would harm the competitor. A third kind of two-species interaction is one in which both species benefit, as might occur between a pollinator and the species it pollinates. The pollinator gains, typically, a food resource, and the species being pollinated is aided in its reproduction. This is known as mutualism.

This classification does not exhaust possible two-species interactions, but these are the major interactions that have been the subject of most study. Many of the dynamical consequences of other interactions can be deduced by considering the cases we have outlined.

B. Predation

We now begin with a discussion of the dynamical effects of each interaction, beginning with the predator prey interaction. The description of these dynamics began with the work of Lotka and Volterra, and we start here with the simplest case they considered. We assume that in the absence of the predator, the prey population grows exponentially, in accordance with our examination of the single species case earlier. In the absence of the prey, we assume that the predator population declines exponentially. To complete the description of the dynamics, we need only describe the interaction between the predator and the prey and the consequences for the dynamics of both populations. The simplest case is to assume that encounters between predator and prey are random, so they occur at a rate proportional to the size of each population. We also assume that there is a fixed probability that the prey is killed in such an interaction, and that each prey individual that is killed contributes a fixed amount to the rate of growth of the predator population. These assumptions lead to the classical equations,

\[ \frac{dH}{dt} = rH - bHP \]
\[ \frac{dP}{dt} = cHP - kP \]

where \( H \) is the size of the prey population, \( P \) is the size of the predator population, \( r \) is the intrinsic rate of increase of the prey in the absence of the predator, \( b \) is the death rate of the predator population in the absence of the prey, the encounter rate is proportional to \( b \), and \( c \) measures the conversion of prey deaths to increases in the predator population.

The analysis of the model shows that there is an equilibrium, but that the equilibrium is neutrally stable. Thus, all other solutions cycle. From this analysis we draw the important conclusion that predator-prey interactions lead to cycles, but the model is unsatisfying because the solution depends completely on the initial conditions. The cycles essentially result from delayed feedbacks. Modifying the model changes the conclusions from the unsatisfying one of neutrally stable cycles. If the equations are modified so that there is density dependence in the prey population, the system approaches an equilibrium in an oscillatory fashion. Changing the description of the consumption of prey by predators can then produce stable cycles. This presence of cyclic behavior is the most important conclusion from the study of predator-prey interactions.

There are numerous examples of interactions between predator and prey in natural systems, which at least exhibit a tendency toward cyclic behavior, ranging from the classic example of hare and lynx in Canada to wolves and moose on Isle Royale. The observations in many of these cases are somewhat equivocal because it is often difficult to determine what is really causing the cycles, partly because it is difficult to isolate a single predator-prey pair within a larger community.

C. Host-Parasitoid Dynamics

By contrast, there is another kind of exploiter-victim relationship that does have more tightly coupled species interacting. The overwhelming majority of animal spe-
cies are insects, and approximately one-fifth to one-sixth of all insect species are parasitoids; they have a life cycle where they lay their eggs in the developing stages of other insects, the hosts, and consume them from within so the host is killed and one or more adult parasitoids emerge. Many insect pests of crops have important parasitoids. The interaction between a host and a parasitoid is essentially another version of the interaction between a prey and a predator, but the approach and the questions are somewhat different, so we treat this case separately. Here, the typical approach is discrete in time, which is appropriate for host insect species with single generations per year. The host species is assumed to grow exponentially (or geometrically) in the absence of the parasitoid. Unlike the predator-prey interaction, the parasitoid life cycle is tightly coordinated with the host because the parasitoid lays its eggs in the developing host. Thus the dynamics of the interaction are described by the probability that an individual host is parasitized, which could depend on the population density of the host and of the parasitoid.

The equations describing this interaction were first developed by Nicholson and Bailey, but these initial equations predicted that the populations would be unstable with cycles of ever-increasing amplitude. Thus, efforts since then have been focused on determining what would allow host and parasitoid to persist in a stable fashion. Suggestions like aggregation of host or parasitoid, or density dependence have been studied in models. Yet, as recent careful work by Murdoch and his colleagues on red scale and its parasitoids in California shows, the search for the stabilizing feature in natural systems is quite difficult. This is a system where a host and its parasitoids coexist stably, yet the search for the mechanisms that stabilize this interaction has been able to reject many of the conventional explanations.

D. Disease Dynamics

Another very important interaction that is similar to the predator-prey system is the interaction between diseases caused by bacteria or viruses and their hosts. However, in this case the focus is very different, since we do not count the number of the disease organisms and instead count only the number of diseased hosts. These systems have become very important subjects of study in population biology partly because of the very good records of the number of diseased individuals, in particular for human diseases.

The classic models in this area, which go back to the early work of Kermack and McKendrick in the 1920s, break up a population into three categories: susceptible, those individuals that can become infected; infectives, those individuals which are infected and infect others; and removed, those individuals which no longer can infect others. The simplest case is to look over relatively short times, so that overall population sizes are constant, and look at epidemics. The models, called SIR models, are framed in terms of differential equations where the key parameters are the contact rate, the recovery rate, and the total population size. The rate of infection is assumed to be equal to the product of the contact rate, the number of susceptibles, and the number of recessives. The recovery rate for individuals is assumed to be just a constant. Then the dynamics depend on a single nondimensional parameter, the reproductive number, \( R_0 \), which is the number of new infectives produced by each infective, when the population is essentially composed of all susceptible individuals. If \( R_0 \) is less than one, the epidemic dies out, and if \( R_0 \) is greater than one, the epidemic will grow. This simple model fits many epidemics very well, as shown by the original work of Kermack and McKendrick looking at the 1905–1906 Bombay plague epidemic.

Disease dynamics have become central to some of the most important current questions in population ecology—namely, the relative importance of endogenous (within the system) versus exogenous (outside the system, often called stochastic) forces in determining the dynamics of population. This question, which is central to understanding the importance of any of the interactions outlined in this chapter, has been dealt with most clearly within the realm of childhood disease (chicken pox, measles, whooping cough, etc.) dynamics because the quality of the data is so good and because the models are more mechanistic.

The data, such as that for measles, have several important features, exhibiting both seasonal behavior driven in part by the school cycle and more complex multiyear cyclic behavior. There have been many different attempts to use extensions of the SIR model to understand the dynamics in a quantitative fashion. The approach taken is essentially to explicitly include the stochastic forces as an additional term in the model and to find the best fit (in a statistical sense) of the model to the data. This procedure illustrates the importance of using these parameter estimation approaches since the standard SIR model with constant infectivity does not provide a very good fit to the data, implying that the model is inadequate in some way. Further investigation shows that making the infectivity seasonally variable, corresponding to the role of school vacations in reducing the likelihood that children infect each other, greatly
improves the fit of the model. Moreover, the rather complex dynamical patterns in the disease dynamics are shown to be generated by a combination of exogenous and endogenous forces: both the stochasticity and the nonlinearity (and seasonality) in the SIR model are critical to producing the observed dynamics.

E. Competition

The simplest models of competition are essentially phenomenologically based and start with the logistic model for a single species as their basis. The effect of each species on the other is to reduce the per capita growth rate. From this simple model of competition, several important ecological paradigms have emerged. The basic question to ask is what allows two species to coexist, rather than having a superior competitor eliminate an inferior one. The outcome of competitive interactions ends up being determined, in part, by the relative importance of interspecific effects as compared to intraspecific effects. Coexistence requires that the interspecific effects be weaker than the intraspecific effects.

In contrast to the predator-prey interaction, the dynamics of competition view is a monotonic approach to equilibrium without any cycles. However, if interspecific effects are strong enough, there can be a kind of priority effect. The species that wins in competition may depend on the initial densities in that case.

A more mechanistic approach to competition sheds some light on the issue of coexistence. With interference competition, in which the two species might, for example, affect each other by the production of allelopathic chemicals, interspecific effects may be stronger than the intraspecific ones. With exploitation competition, in which the two species use common resources, the relative strength of interspecific effects and intraspecific effects depends on the differences in resource use—for example, differences in the size of seeds used. From this view comes what was an extraordinarily important principle, the competitive exclusion principle of Gause, which stated that competing species have to differ in order to coexist.

The search for evidence of competition in natural populations is also difficult, but several large-scale surveys have demonstrated that there is relatively good evidence for competition. Classic work in this area includes careful studies of lizards in the Caribbean, which has shown competition for food and avoidance of competition by different species perching at different heights in trees. Competition for space has also been well documented in intertidal systems, in which removal of one species allows another to expand the range of habitat it uses. Recent work has also focused on apparent competition, in which one species “competes” with another by allowing a common predator to increase its density, which then suppresses the competitor.

VII. CONCLUSIONS

Population dynamics is one of the fundamental areas of ecology, forming both the basis for the study of more complex communities and of many applied questions. Understanding population dynamics is the key to understanding the relative importance of competition for resources and predation in structuring ecological communities, which is a central question in ecology.

Population dynamics plays a central role in many approaches to preserving biodiversity, which until now have been primarily focused on a single species approach. The calculation of the intrinsic growth rate of a species from a life table is often the central piece of conservation plans. Similarly, management of natural resources, such as fisheries, depends on a population dynamics approach to determine the largest sustainable yield.

See Also the Following Articles

DISEASES, CONSERVATION AND • PARASITOIDS • POPULATION DENSITY • POPULATION GENETICS • PREDATORS, ECOLOGICAL ROLE OF • SPECIES INTERACTIONS

Bibliography

I. Variation within Populations
II. Deterministic Population Genetics
III. Random Genetic Drift
IV. The Interaction of Drift with Deterministic Forces
V. Conclusions

GLOSSARY

allele frequency The frequency of a variant form of a genetic locus within a population.

genetic drift Evolutionary change caused by random sampling of genotype frequencies in a finite population.

genotype The state of an individual with respect to a defined genetic locus or set of loci.

heritability The proportion of the variance in a trait that is due to additive genetic effects.

inbreeding Matings between close relatives.

mutation rate The frequency with which new mutations arise per generation.

neutral mutations Mutations whose effects on fitness are either nonexistent or so small that their fate is controlled by genetic drift rather than selection.

phenotype The state of an individual with respect to a trait of interest.

polymorphism The existence at intermediate frequencies of two or more variants at a locus within a population.

selection The differential survival or reproductive success of individuals, associated with differences in phenotype or genotype.

DARWIN’S THEORY OF “DESCENT WITH MODIFICATION” implies that all of the stupendous diversity of life on Earth is ultimately traceable to genetic diversity within populations. The study of the nature and causes of within-population variation, and of the mechanisms by which it is transformed into differences between populations over space and time, is the province of population genetics. The subject involves both theoretical modeling of evolutionary processes, based on knowledge of the mechanisms of inheritance, and the testing of these models using data on variation and evolution in natural and artificial populations.

I. VARIATION WITHIN POPULATIONS

A. Types of Phenotypic Variation

Since evolutionary change depends on the existence of genetic variation within populations, measurement of the extent of such variation is crucial (Lewontin, 1974). Variation at the level of externally visible phenotypes can be divided into three categories.

1. Discrete Variation

This involves traits which can be divided into a small number of discrete categories, such as eye color in hu-
mans or shell color and pattern in the land snail *Cepaea* (Ford, 1975). It is often controlled by one or a few genes, and usually it involves relatively superficial traits such as color patterns. Only a relatively small proportion of the phenotypic variation of interest to evolutionists is of this kind.

2. Quantitative Variation

Quantitative variation is all-pervasive. This can involve either meristic traits, such as bristle number in *Drosophila*, in which there is a large number of discrete categories, or continuously varying metrical traits such as body size. Variation in typical quantitative traits is known to be under the joint control of environmental effects, accidents of development, and sets of genes whose individual effects are small relative to the total range of variation in the traits (Falconer and Mackay, 1996). Statistical methods that utilize the degree of resemblance between close relatives enable the determination of the proportion of the total phenotypic variation that is contributed by additive genetic causes—the heritability, which controls the rate of response to selection on a trait (see Section II,D). Heritabilities for quantitative traits typically are between 20 and 80%, corresponding to the fact that artificial selection is highly effective in changing the mean value of almost every trait that has been examined (Falconer and Mackay, 1996).

3. Concealed Variability

A more subtle form of phenotypic variation is concealed variability, i.e., variability that is only exposed when homozygous genotypes are produced by close inbreeding. This is responsible for the increased variation among inbred lines when a set of such lines is made from a random-bred base population and for inbreeding depression, which is the decline in the mean values of fitness-related traits such as viability and fecundity with inbreeding (Falconer and Mackay, 1996). Both of these phenomena reflect, at least in part, the widespread occurrence of recessive or partially recessive rare alleles in random-mating populations (see Section II,F), whose phenotypic effects are only fully exposed when they are made homozygous and are therefore not evident in randomly mating populations.

In *Drosophila*, special breeding methods involving the use of genetically marked chromosomes with inversions that suppress crossing over have shown that up to 30% of haploid genomes carry recessive lethal genes. These contribute about half the inbreeding depression manifested when fully homozygous genotypes are produced; a similar magnitude of inbreeding depression is caused by genes of small effect (Crow, 1993). The net fitness of fully inbred *Drosophila* is only a few percent of that of outbred flies. Even more extreme effects of complete inbreeding are likely in vertebrates, which have much larger genomes. The deleterious fitness effects of inbreeding have probably played a major role in promoting the evolution of mechanisms of inbreeding avoidance, such as the self-incompatibility loci of flowering plants.

4. Interpreting Phenotypic Variation

The previously mentioned basic facts were established by the early 1950s and led to an active debate concerning the causes of natural variation (Lewontin, 1974). In one view, championed by H. J. Muller, variation is mostly due to rare deleterious alleles maintained by mutation pressure at a large number of loci; the coexistence of alleles at a locus at intermediate frequencies (polymorphism) is characteristic of only a small number of loci. In the other view, advocated by Dobzhansky, polymorphism is the norm, and it reflects variation that is actively maintained by selection. In the absence of any means of identifying loci without the prior existence of genetic variability, no unbiased survey of the extent of genetic variation at individual loci was possible with the methods of classical genetics, and so this question could not be answered.

B. Molecular Variation

1. Protein Electrophoresis

The previously mentioned situation was transformed by the development of molecular genetics. Gel electrophoresis of soluble enzymes and proteins provides a rapid and simple method for surveying populations for variants affecting the structure of a large number of different proteins and hence genes (Lewontin, 1974; Hartl and Clark, 1997). The results of such surveys reveal that a high fraction of loci coding for soluble proteins are polymorphic in the sense of having at least one rare variant whose frequency exceeds 5%; the average individual from a randomly mating population is typically heterozygous for a significant fraction (several percent) of such loci. Despite some biases in the methodology, particularly the inability of electrophoresis to detect many types of amino acid sequence changes and the restriction of the method to soluble proteins, it is clear that protein polymorphism is not an exceptional situation.
2. Measurement of Variation in DNA Sequences

The introduction of recombinant DNA technology has meant that population geneticists can now study variation at the level of the nucleotide sequence. Surveys of within-species DNA sequence variation of nuclear genes have been most intensively carried out in Drosophila, but comparable results are emerging from other species (Li, 1997). The basic conclusion is that variants due to single nucleotide changes are the most abundant source of variation in natural populations. For silent substitutions in third codon positions, which do not change the amino acid sequence, and for changes in introns and flanking sequences, the probability that two randomly chosen alleles from a Drosophila population differ at a given site (the nucleotide site diversity) is typically of the order of 1% or a few percent, depending on the species. The level of this type of variability is about 10 times higher in the bacterium Escherichia coli and one-tenth as high in humans (Li, 1997). For most genes, diversity is much lower for replacement changes, which alter the amino acid sequence. In addition to single nucleotide polymorphisms, DNA variability is contributed by small insertions and deletions of sets of nucleotides and by insertions of transposable elements, mostly in noncoding regions. Other types of variability include variation in the sizes of tandem arrays of microsatellite and minisatellite loci, which are often highly polymorphic and provide useful genetic markers (Bruford and Wayne, 1993; Hartl and Clark, 1997). The density of such loci, however, is low in relation to the total size of the genome. The total level of variability at the level of DNA sequences is about two orders of magnitude greater than that revealed by electrophoresis because of the high degree of variability at silent and noncoding nucleotide sites relative to replacement sites.

3. Interpreting DNA Sequence Variation

Although variation at the level of DNA sequences must underly heritable phenotypic variation, it is difficult to relate the two, except for intensively studied human genetic diseases. The abundance of variation in both protein and DNA sequence might seem to vindicate Dobzhansky’s view of the causes of natural variation. However, it is likely that much of the silent and noncoding variability is close to neutrality with respect to effects on fitness (see Section III.C). Nevertheless, there is a real possibility that selection also frequently influences variation and evolution in protein and regulatory sequences (Hartl and Clark, 1997; Li, 1997). The role of deterministic forces in variation and evolution within populations will thus be considered next.

II. DETERMINISTIC POPULATION GENETICS

A. Allele and Genotype Frequencies

If we focus on a given nucleotide position, the basic descriptor of the state of a population is the set of frequencies of the four alternative states, A, T, G, and C. If recombination within a gene is ignored, we can consider the set of all nucleotide sequences observed at a locus as alternative alleles, whose frequencies characterize the state of the population with respect to this locus. Mendelian inheritance implies that this state is not changed in the absence of evolutionary forces (the occurrence of intragenic crossing over and gene conversion at low frequencies means that in practice this is a good approximation rather than an exact description). This is enshrined in the Hardy–Weinberg principle, which states that the frequencies of diploid genotypes in a random mating population with a set of $n$ alleles with frequencies $p_1, p_2, \ldots, p_n$ rapidly reach equilibrium values given by the multinomial expansion of $(p_1 + p_2 + \ldots + p_n)^n$. The importance of this result is that existing natural variation is preserved by Mendelian inheritance. This removes Darwin’s difficulties over the rapid loss of variation under blending inheritance, which led him to adopt a theory of the inheritance of acquired characteristics for which there is no empirical foundation (Fisher, 1930).

B. Mutation

1. Types of Mutations and Their Rates

The ultimate source of natural variation is known to be spontaneous mutation, defined as a heritable change in the genetic material and which occurs without reference to the adaptive utility of the phenotypic consequences of the change in question. The most abundant mutations are nucleotide substitutions, but small deletions and insertions due to slippage during DNA replication are relatively common as well (Drake et al., 1998). Insertions of transposable elements, large deletions and duplications, and chromosome rearrangements such as inversions and translocations also occur and contribute to evolutionary changes in genome structure. Rates of spontaneous mutation in organisms with DNA genomes are extremely low due to the operation of complex enzymatic systems which repair lesions in DNA; the rates of nucleotide changes per site per cell generation
in DNA-based microbes are between $1 \times 10^{-8}$ and $5 \times 10^{-7}$. Similar values apply to higher eukaryotes such as Drosophila and humans, but the rate per organism generation is between 10 and 50 times higher owing to the many cell divisions that occur during the production of germ cells. Per locus rates of mutation to alleles with major phenotypic effects are substantially higher—on the order of $10^{-3}$ per generation in Drosophila, mice, and humans (Drake et al., 1998). This is not surprising given the large number of nucleotides in the coding and regulatory regions of typical loci. Rates of change in copy numbers in microsatellite and minisatellite loci in mammals are much higher, however, up to $10^{-1}$ per locus per generation (Bruford and Wayne, 1993). Mutation rates in viruses with RNA genomes are also extremely high due to their lack of repair mechanisms (Drake et al., 1998).

2. Evolution under Mutation Pressure

Given the low rate of mutation at the nucleotide level in DNA-based organisms, the time scale of mutational change in the frequencies of the four alternative states at a nucleotide site is very large—on the order of 1 billion generations. Mutation pressure at this level is thus an extremely weak force and is easily opposed by other evolutionary factors. For most purposes, therefore, mutation can be regarded simply as a source of new variation and as unimportant as a cause of directed evolutionary change (Fisher, 1930). This statement needs to be qualified when the aggregate effects of mutations affecting a particular phenotype are considered; the numbers of loci affecting a single quantitative trait are sufficiently large that increases in variability due to mutation are detectable in stocks that are initially genetically uniform. The rate of increase in variance per generation is typically on the order of $10^{-3}$, relative to the nongenetic variance in the trait (Falconer and Mackay, 1996). Given that fitness-related traits are affected by many genes, and that most phenotypic changes caused by mutations are harmful to the organism, there is a tendency for the mean values of fitness components to decline under mutation pressure when selection is relaxed by experimental manipulations (Crow, 1993; Drake et al., 1998).

3. Mutation and Selection

These considerations imply that a relatively weak force of selection, far smaller than is measurable experimentally, can prevent the spread of deleterious mutations at a locus. This explains the fact that amino acid polymorphisms are usually much less frequent than silent or noncoding polymorphisms. The same observation applies to comparisons of sequences between different species (Kimura, 1983). A very important form of selection is thus purifying selection, whereby deleterious mutations are constantly being eliminated from the population (see Section II,F). Similarly, quantitative traits are often subject to stabilizing selection, such that individuals with extreme trait values are less fit than individuals with intermediate values (Falconer and Mackay, 1996). Variability in quantitative traits is maintained, at least partially, by a balance between the input of new variation by mutation and its elimination by stabilizing selection (Falconer and Mackay, 1996).

C. Selection at a Single Locus

1. The Basic Model

A large body of theory has been developed to describe the action of natural selection on Mendelian variation. In the simplest case, a single locus with a pair of alternative alleles, A and a, is postulated. A randomly mating, infinitely large population is assumed. Ignoring sex differences, the relative fitnesses of the three possible genotypes AA, Aa, and aa in a diploid can be written as 1, 1 − $hs$, and 1 − $s$, respectively, where $s$ is the selection coefficient and $h$ is the dominance coefficient. $s$ measures the strength of selection ($s > 0$, when a is disfavored by selection); $h$ measures the extent to which the fitness of the heterozygote is reduced by the presence of a. The fitnesses are assumed to be constant over time.

Fitness in this context is most easily understood in terms of viability selection in a population with discrete generations, where the relative probabilities of the three genotypes surviving from egg to adult are equivalent to the three fitnesses. In general, selection may involve many different aspects of the life history, especially female fecundity and male mating success. More elaborate models have been developed to study these cases, including extensions of the theory to populations with overlapping rather than discrete generations, but the basic conclusions are similar (Crow and Kimura, 1970; Ewens, 1979; Hartl and Clark, 1997).

If the frequencies of the two alleles in one generation are $p$ and $q$, respectively, the change in frequency of A over one generation is

$$\Delta p = qsp \left[ q + h(1 - 2p)q / \bar{w} \right], \quad (1s)$$

where $\bar{w} = 1 - 2pqhs - q^2s$ is the population’s mean fitness.

An interesting, equivalent form, from Wright...
Equation (1b) implies that gene frequency change is in the direction of the gradient of mean fitness with respect to gene frequency. A more detailed analysis of the dynamics of a single locus with an arbitrary number of alleles, and arbitrary but constant fitnesses, shows that mean fitness increases monotonically as allele frequencies change so that stable equilibria in gene frequency space correspond to local maxima in mean fitness. This is known as Wright’s adaptive landscape (Crow and Kimura, 1970; Wright, 1977; Ewens, 1979; Hartl and Clark, 1997).

2. Directional Selection

Regarding the case of a pair of alleles, when $0 \leq h \leq 1$ (so that the fitness of the heterozygote is bounded by the fitnesses of the homozygotes), $A$ is favored by selection and will progress to fixation. This case is referred to as directional selection. If $A$ is initially rare so that second-order terms in $p$ can be neglected, Eq. (1) can be approximated by $\Delta p = (1 - h)p$. The initial rate of change of log gene frequency ($=\Delta p/p$) is therefore proportional to the fitness difference between the genotypes $AA$ and $aa$. This reflects the fact that, with random mating, rare alleles are overwhelmingly carried in heterozygotes (frequency $2pq$) since the frequency of homozygotes ($p^2$) is negligible. If $A$ is completely recessive so that $h = 1$, Eq. (1a) for rare $A$ is approximated by $\Delta p = sp$ so that the logarithmic rate of increase in frequency of $A$ is proportional to $sp$, which tends to zero with decreasing $p$. This reflects the extreme rarity of the favored homozygotes, and it implies that rare recessive alleles are only weakly selected in randomly mating populations (Haldane, 1932).

3. Survival of Favorable Mutations

The previous conclusion is reinforced by calculation of the probability of survival of new mutations. Even in very large populations, a new mutation is likely to be represented initially in only one or a few individuals. Since reproduction is subject to random variation, described by the probability distribution of the numbers of surviving offspring per mated individual, there is a finite chance in each generation that all carriers of a rare mutant gene will fail to transmit it (Fisher, 1930; Haldane, 1932). The chance that a single copy of a new favorable mutation ultimately survives random loss from a large population of constant size is approximately $2x(1 - h)$, assuming a Poisson distribution of offspring number (Haldane, 1932; Crow and Kimura, 1970). Most favorable mutations are thus likely to be lost from the population a few generations after they arise; on average, 148 occurrences of a mutation with a heterozygous selective advantage of 1% are needed for a 95% chance that one will survive. This implies that there is a considerable random element to adaptive evolution at the genetic level. If there are several different loci at which mutations that provide adaptations to a given pressure of selection can occur, it may be a matter of chance which locus actually responds to selection in a given population. The same pressure of selection can therefore result in the divergence of isolated populations at the level of the genotype, even if the same phenotype is evolving. The numerous different genes involved in the adaptation of different human populations to malaria parasites illustrate this principle (Hill and Weatherall, 1998).

The formula for survival probability implies that a recessive favorable mutation has a zero chance of survival in an infinite population; calculations based on diffusion theory (see Section IV,D) show that the probability for a randomly mating population of size $N$ in this case is approximately $0.8e^{-2N}$ (Crow and Kimura, 1970). This means that recessive autosomal mutations are unlikely to become established by selection in randomly mating populations of even moderate size. As first pointed out by Haldane (1932), it is thus no accident that nonrecessive alleles have been established by selection in cases of recent adaptation involving genes of major effect, such as industrial melanism and pesticide resistance, despite the fact that most spontaneous mutations are recessive with respect to their effects on the phenotype (Haldane, 1932; Ford, 1975).

4. Time Course of Gene Frequency Change

Once a favorable allele rises to a sufficiently high frequency that random loss is unlikely, its progress in a large population can be calculated by integration of the differential equation which approximates Eq. (1) when selection is weak. This procedure yields expressions for the time needed to change gene frequencies by a given amount (Haldane, 1932; Crow and Kimura, 1970; Hartl and Clark, 1997). These can be used to estimate selection intensities in experimental and natural populations by comparing observed trajectories of gene frequency change with the theoretical predictions. This method is particularly useful for microbes, which have short generation times and can be grown in very large artificial populations. Selection coefficients on the order of 0.5% can be measured in microbial experiments (Hartl and Clark, 1997).

The results of these calculations show that the time
required to cause a given amount of change in gene frequency is inversely proportional to \( s \) and depends only logarithmically on the initial frequency of the favorable allele, except when it is recessive. This implies that selection on nonrecessive alleles can cause evolutionary change on a timescale of the order of a few multiples of \( 1/s \), almost regardless of the initial conditions. Even a very weak force of selection can thus transform a population in a period of time that is negligible in geological terms (Fisher, 1930; Haldane, 1932). This conclusion is historically very important because it removes many of the objections to natural selection as a potent force in evolution.

D. Selection on Quantitative Traits

1. Predicting the Response to Selection

This question can also be considered in relation to quantitative traits, which are more relevant to phenotypic evolution than traits controlled by single genes. The standard model of quantitative genetics (which involves several simplifying assumptions) implies that the response to selection on a single quantitative trait is governed by the equation

\[
R = h^2 S
\]

where the selection response, \( R \), is the difference in mean value between one generation and the next; \( h^2 \) is the heritability of the trait (see Section I,A,2) (there is no connection with the dominance coefficient introduced previously); and \( S \), the selection differential, measures the intensity of directional selection applied to the trait and is equal to the difference in mean between the parental individuals who have survived selection and the population mean before selection (Fig. 1). Given the fact that \( h^2 \) is usually substantially greater than zero, this implies that populations can respond rapidly to directional selection on quantitative traits, as is indeed the case (Falconer and Mackay, 1996).

2. Rates of Change under Selection

In a sexually reproducing population, the ability of recombination to put together new combinations of alleles means that selection can eventually produce genotypes with trait values that far exceed the range of variation that existed in the original population, simply by causing the fixation of alleles which increase trait value and that were segregating in the original population. In the absence of new variability created by mutation, the selection response will eventually cease when all favorable alleles have been fixed by selection. Given the evidence for a significant input of new variation by mutation for typical quantitative traits, a relatively gentle pressure of selection in a large population will never fully deplete variation and therefore sustained responses to selection are possible (Falconer and Mackay, 1996). The breeder can produce substantial changes in a single trait over less than 100 generations, comparable in magnitude to geologically rapid changes that have taken thousands of generations in nature. It is therefore unlikely that lack of genetic variation is often a serious constraint on the ability of populations to undergo phenotypic evolution, unless there is intense selection for a novel combination of traits, which cannot immediately be produced by recombination among existing genotypes (Wright, 1977).

3. Selection on Multiple Characters

In general, natural selection acts on suites of characters, not on a single trait in isolation. A multivariate generalization of Eq. (2) has been derived by Lande (1988), and it is useful for the interpretation of data on selection on multiple traits:

\[
\Delta \mathbf{x} = G \mathbf{V} \ln \mathbf{w}
\]

where \( \Delta \mathbf{x} \) is the change per generation in the vector of mean values for the set of traits, \( \mathbf{V} \ln \mathbf{w} \) is the gradient vector of log mean fitness with respect to the set of trait means, and \( G \) is the matrix of additive genetic variances and covariances among the set of traits. The similarity to Eq. (1b) is evident.

Equation (3) shows that, in general, we can only predict the effect of selection on a trait if we know the extent to which it is genetically correlated with other traits that are also the target of selection. Short-term evolutionary changes in a given trait may be due, at
least in part, simply to the fact that it covaries with another trait that is the target of selection. However, provided that the G matrix is nonsingular and fitnesses are constant in time, a stable equilibrium state corresponds to a maximum in the surface of mean fitness with respect to the mean trait values, indicating that selection will ultimately bring the population close to the optima with respect to each individual trait unless there are strong constraints on the structure of genetic variation. This implies that the concept of an adaptive landscape can be applied to quantitative traits (Lande, 1988).

E. Maintenance of Variation by Selection

The role of selection in preserving variation rather than destroying it, called balancing selection, is now discussed. This form of selection was unknown to Darwin, and its discovery is one of the most important contributions of population genetics.

1. Heterozygote Advantage

The simplest case is when the heterozygote at a locus with two alleles has a higher fitness than the two homozygotes, which was first investigated by Fisher in 1922. Let the fitnesses of AA and aa relative to Aa be written as \( s = s^* = 1 - c \), Eq. (1b) now yields the result that there is a unique stable equilibrium at which the frequency of A is \( p^* = 1/s + 1 \), to which the population converges from any starting point other than fixation for A or a (Crow and Kimura, 1970; Hewens, 1979; Hartl and Clark, 1997). The classic example of heterozygote advantage is the polymorphism for the \( \beta \)-globin variant in humans that causes sickle-cell anemia when homozygous, a disease which is effectively lethal under natural conditions. The maintenance of this allele in human populations subject to severe malaria infections is due to the selective advantage of heterozygous carriers, conferred by their resistance to malaria, as suggested by Haldane in 1948. Several human globin gene polymorphisms, as well as some other human polymorphisms, are also maintained by resistance to malaria (Hill and Weatherall, 1998).

2. Frequency-Dependent Selection

Balancing selection can also be caused by negative frequency-dependent selection, in which the relative fitness of a genotype decreases with its frequency in the population. This obviously acts to inhibit the fixation of an allele which may initially have a selective advantage over other alleles when introduced into a population. Genetically controlled resistance to parasitic disease is subject to this form of selection since the abundance of a parasite decreases as the number of susceptible hosts diminishes, thereby reducing the selective advantage to a host allele that causes resistance to a particular parasite genotype (Li, 1997). A similar frequency dependence may affect alleles controlling virulence in the parasite population. Genes thought to be involved in disease resistance, such as the major histocompatibility complex (MHC) loci of vertebrates, are often highly polymorphic and segregate for large numbers of alleles; molecular analyses show clear evidence for balancing selection (see Section IV.C). The role of frequency-dependent selection in these cases remains to be established, however. Frequency dependence is inherent in the dynamics of the self-incompatibility loci of flowering plants, which also have very high allele numbers (Hartl and Clark, 1997). Similarly, Batesian mimicry, in which an edible species mimics a distasteful model, exhibits frequency dependence since a predator is more likely to mistake a rare mimetic form for the model than a common one (Fisher, 1930; Ford, 1975). It should be noted that frequency-dependent selection does not necessarily lead to polymorphism; its outcome depends on a delicate balance of the selective parameters, and dynamic complexities such as limit cycles or even chaos are possible. With frequency-dependent selection, maxima in mean fitness do not necessarily correspond to stable equilibria.

3. Temporal Variation in Fitnesses

Temporal fluctuations in relative fitnesses can also lead to the maintenance of polymorphisms by selection, in the case of two alleles at a locus, a sufficient condition is that the geometric mean fitness of the heterozygote over generations exceeds that of both homozygotes (Crow and Kimura, 1970; Hartl and Clark, 1997). Similarly, spatial variation in the direction of selection can maintain variation. This can happen in two ways. The first requires strong density-dependent regulation within different environmental patches so that the number of adults emerging from a patch is largely independent of the genetic composition of the eggs laid in that patch. In this case, opposing directions of selection in different patches can maintain polymorphism even if there is complete random mating among patches (Crow and Kimura, 1970). Directional selection in opposite directions on alleles in males and females is a special case. Second, restricted migration among populations subject to different directions of selection can result in polymorphism within each population accompanied by genetic differentiation among populations. Clinal variation in allele frequencies or quantitative trait values...
results from geographic gradients in selection pressures, coupled with the smoothing effect of migration, as in the case of Bergmann’s rule, which states that the mean body sizes of mammal populations increase with higher latitudes (Harril and Clark, 1997).

4. Meiotic Drive

Antagonism between the effects of selection at different levels may also maintain variation. The best studied cases involve the phenomenon of meiotic drive or segregation distortion, in which one allele (D) at a locus when heterozygous causes the destruction of gametes carrying the alternative allele (d). In animals, this is usually found to occur only in males. Provided that the fertility of Dd males is affected less than linearly by the destruction of the d sperm, D will gain a transmission advantage. It will spread through the population unless there is a countervailing selective disadvantage at the level of individuals (Harril and Clark, 1997). In the best studied cases of this kind, the SD system of Drosophila melanogaster and the t-haplotype system of house mice, the primary disadvantage seems to come from sterility of DD males.

F. Mutation–Selection Balance

The balance between recurrent mutations to variants that impair the functions of gene products and selection against them is probably a major factor in maintaining genetic variation, given the fact that higher organisms have tens of thousands of genes (Drake et al., 1998).

The large number of changes in a coding sequence that can impair the function of a gene product implies that the process of mutation to deleterious alleles at a locus can be regarded as effectively irreversible, provided that the wild-type allele predominates in the population. If the deleterious alleles at a locus are completely recessive, with selection coefficient \( s \) and a rate of origination by mutation from wild-type of \( u \), their equilibrium frequency in a randomly mating population (assuming \( s > u \)) is \( q^* = \sqrt{u/s} \). Even lethal alleles (\( s = 1 \)) can thus reach appreciable frequencies if they are completely recessive, with \( u = 10^{-3} \); for example, \( q^* = 3 \times 10^{-2} \) . However, experimental studies of lethal mutations in Drosophila show that they usually impair the viability of heterozygotes with wild-type by 2 or 3%; detrimental alleles with more minor effects (\( s \) on the order of a few percent) appear to be much less recessive, with \( h \) values of approximately 0.25 (Crow, 1993). With random mating, the much greater frequency of heterozygotes than mutant homozygotes means that selection on heterozygotes controls the frequencies of mutations; in this case, \( q^* = u/(hs) \). Indirect experimental evidence indicates that the mean of \( h \) s for detrimental \( s \) is on the order of 1% in Drosophila so that \( q^* \) is approximately \( 10^{-1} \) with \( u = 10^{-5} \). The frequency of heterozygous carriers is approximately \( 2q^* \) so that the mean number of heterozygous detrimental mutations per individual in a genome such as that of Drosophila with approximately 15,000 genes is \( 30,000 \times 10^{-1} = 30 \) with these assumptions. The total rate of mutation to lethal mutations per haploid genome in Drosophila is 0.01, assuming \( hs = 0.02 \) the mean number of heterozygous lethals per diploid individual is on the order of 1, in agreement with the direct estimate mentioned previously.

G. Genetic Load

1. General Considerations

The previous discussion leads to the consideration of the effect of selection on the fitness of the population as whole; if there is a large amount of variability with respect to loci under selection, it is obvious that the mean fitness of the population must be much less than that of the best genotype. If fitnesses are measured relative to a value of 1 for the optimal genotype in the system under consideration, the reduction in fitness can be conveniently measured by the genetic load, defined as \( L = 1 - \bar{w} \). In the case of viability selection, no genotype can have a survival probability greater than 1, so \( L \) provides an upper bound to the probability that a zygote survives to maturity. More generally, \( L \) measures the proportion of the population that dies or fails to reproduce as a result of selective differences among genotypes. The effects of multiple loci on mean fitness can be calculated by assuming that different loci have independent effects, so that the fitness of a multilocus genotype is given by the product of the fitnesses of all the single-locus genotypes which contribute to it. If \( L \) is the load contributed by the \( i \) th locus, the mean fitness with respect to \( m \) independent loci, relative to the value for the optimum genotype, is:

\[
\bar{w} = \prod_{i=1}^{m} (1 - L_i) \approx \exp - \sum_{i=1}^{m} L_i \quad (4)
\]

2. Mutational Load

With nonrecessive deleterious alleles maintained by mutation, the load for a single locus at equilibrium is approximated by \( 2q^*hS = 2u \). The total load is \( 1 - \exp - U \), where \( U \) is the mean number of new deleterious mutations in a diploid individual. Lethal
mutations contribute relatively little to this since their total mutation rate is very low, but detrimental have a major effect. Assuming 15,000 genes with a mutation rate of $10^{-8}$, the mutational load for a *Drosophila* population would be about 0.26 (Drake et al., 1998). For a genome of 80,000 genes, as in mammals, the load would be 0.80, a considerable burden of selective loss.

The existence of a large mutational load suggests that there is an adaptive advantage to a reduction in the mutation rate; this can be studied theoretically by calculating the rate of spread of a rare modifier gene that reduces $U$ by a small amount $rU$. If the modifier recombines freely with autosomal loci subject to mutation and selection, it has a selective advantage of $hsU$. This raises the question of why mutation rates are not closer to zero; this probably reflects the fact that there is a fitness cost to the necessary repair systems so that $U$ is adjusted to a level at which the costs and benefits of increased repair balance (Drake et al., 1998).

3. Segregational Load

Similar calculations can be performed for models of balancing selection, yielding estimates of the segregational load. In the case of heterozygote advantage, the load due to a single locus is $s/d(s + 1)$ (Crow and Kimura, 1970). Equation (4) can be used to determine the segregational load contributed by a large number of polymorphic loci with independent effects. This can be considerable, even if selection is weak. For example, 10,000 loci each with $s = t = 0.001$ would yield a mean fitness of only 0.00007. This is so low that only a very high fecundity species would be able to produce the two surviving offspring per adult needed to maintain itself. This implies that either most molecular variation has very slight or no effects on fitness or the assumption of multiplicative fitnesses is unrealistic.

An extreme alternative to multiplicative fitnesses is truncation selection. Genotypes at a set of loci are assumed to be ordered with respect to their fitnesses as determined by the multiplicative fitness model; a fixed proportion of the population, containing the set of genotypes with the highest fitnesses, is allowed to survive. This is equivalent to assuming that individuals compete for a limiting resource, and that only the fittest succeed. Under these conditions, a much larger number of loci can be exposed to selection for a given total $L$ than with multiplicative fitnesses, for the same selection intensity per locus (Crow and Kimura, 1970). Less extreme forms of departure from multiplicativity can have similar but smaller effects on the total load.

4. Substitutional Load

Genetic loads also apply to adaptive evolution. Consider the case of a biallelic locus, where $A$ is initially deleterious and held at a low frequency. If there is a change in the environment so that $A$ becomes favored by selection, it will start to increase. However, in any generation before it reaches fixation, the mean fitness of the population will be reduced below its final value of 1 because of the presence of the disfavored $a$ allele. There is thus a load associated with the substitution of a by $A$, reflecting the fact that natural selection cannot instantly transform a predominantly a population into one which is fixed for $A$. The sum of the loads for each generation over the course of a gene substitution is Haldane's cost of selection, $C$, if the population size is $N$, the total number of individuals eliminated by selection is $CN$. Providing selection is not too strong, $C$ is proportional to minus the logarithm of the initial frequency of $A$; a typical value is 30 (Crow and Kimura, 1970).

The effect of changes at multiple loci can be derived as follows. Assume that there is a steady rate of change in the environment so that each generation $K$ loci start to experience gene substitutions of this kind. $K$ is the rate of gene substitution in the genome as a whole such that after a long period of evolutionary time, $T$, the population will differ from its ancestral state by $KT$ substitutions. If a gene substitution takes $t$ generations to complete, $Kt$ loci will be segregating in any given generation, each associated with an average load of $Ct$ relative to a population which is fixed for the favorable alleles. The mean fitness under multiplicative fitness, relative to a population that is fixed for favorable alleles at all currently segregating loci, is then given by $(1 - Ct)^{50} \approx \exp -CK$. Data on rates of protein evolution suggest that $K$ for an average amino acid site is about $1.5 \times 10^{-7}$ per year in mammals (Kimura, 1983). With 80,000 loci coding for proteins with average size of 300 amino acids, $K$ for the genome is 0.036. With $C = 30$, the mean fitness is 0.34, assuming one generation per year. A much higher load would be found if changes at silent and noncoding sequences are also taken into account. This finding of a high substitutional load associated with molecular evolution was one of the main motivations for the development of the neutral theory of molecular evolution, which asserts that most evolution at the molecular level is caused by the random sampling of alleles in finite population size, genetic drift, and not by natural selection (Kimura, 1983). The substitutional load can be considerably reduced by modifications to the assumption of multiplicative fitnesses, such as truncation...
H. Multiple Loci

1. No Selection

So far, it has tacitly been assumed that evolutionary change involving more than one locus can be modeled by assuming that alleles at different loci are distributed independently of each other in the population and have independent effects on phenotypes and fitness. Although this may be a good approximation for many purposes, it is necessary to examine the consequences of relaxing these assumptions. Deviations from independence among loci in randomly mating populations can be described by linkage disequilibrium parameters, which measure the extent to which the frequencies of the different multilocus gamete types or haplotypes depart from the frequencies expected by randomly combining alleles at different loci. In the simplest case of a pair of loci, each with two alleles (A and a and B and b), there are four haplotypes: AB, Ab, aB, and ab. Let the frequencies of these be \(x_1, x_2, x_3,\) and \(x_4\). If the allele frequencies at the two loci are \(p_A\) and \(p_B\), we can write the haplotype frequencies as 

\[
D = x_3 p_A + D, \quad (1 - p_A) - D, \quad (1 - p_B) - D, \quad (1 - p_B)(1 - p_A) + D, \quad \text{respectively,}
\]

where \(D\) is the coefficient of linkage disequilibrium. It is easily seen that \(D = x_3 p_A - x_4 p_B\). If the frequency of recombination between the two loci is \(c\) (\(0 \leq c \leq 0.5\)), the value of \(D\) in the next generation in an infinitely large, randomly mating population with no selection is \(D (1 - c)\).

In the absence of evolutionary forces other than recombination, the extent of nonrandom association between a pair of loci thus decays exponentially at a rate that is determined by the frequency of recombination. This result can be generalized to associations between multiple loci (Crow and Kimura, 1970). Unless populations are far from equilibrium, departures from linkage equilibrium at a set of loci require the operation of forces tending to generate nonrandom associations between alleles at the different loci. Their magnitude will be at least on the order of the recombination frequencies among them. One such force is genetic drift (see Section III.C), which can cause randomly generated linkage disequilibrium among loci for which \(c\) is on the order of the reciprocal of the population size (Ewens, 1979).

2. Selection on Several Loci

Another possible force causing linkage disequilibrium is epistatic selection, in which the difference in fitness between genotypes at one locus varies according to the genotypes at the other loci in the system. If the fitness effects of different loci combine additively, there is no epistasis, and it can be shown that polymorphic equilibrium under random mating exhibits no linkage disequilibrium (Ewens, 1979). With epistasis, selection tends to preserve haplotypes which contain favorable combinations of alleles, whereas recombination breaks them down. Linkage disequilibrium is not necessarily present if linkage is sufficiently loose in relation to the strength of epistasis. Multiple alternative stable equilibria may occur in multilocus systems so that the fate of a population can be affected by the initial conditions from which evolution starts. Stable equilibria do not necessarily correspond to maxima in mean fitness in the space of haplotype frequencies, another violation of the adaptive landscape principle. However, if epistatic selection is weak in relation to the frequency of recombination, populations tend to converge to trajectories where linkage disequilibrium is nearly constant (quasi-linkage equilibrium), and mean fitness increases monotonically at a rate approximately equal to the additive genetic variance in fitness according to Fisher’s fundamental theorem of natural selection (Ewens, 1979). This has provided a very useful tool for the analysis of the dynamics of multilocus systems (Barton and Turelli, 1991).

3. The Evolution of Close Linkage

There are two biologically important features of systems with strong epistatic selection. The first is that such selection may impose strong constraints on the degree of linkage between polymorphic loci. Suppose that the population is initially segregating for alleles A and a at one locus but is initially fixed for B at a second locus. If a mutation B arises at this locus which interacts with the alleles at the first locus, such that AB is selectively favored but aB is disfavored, B may be unable to invade the population unless \(c\) is below some threshold value. Only mutations at loci that are sufficiently closely linked to the first polymorphism in the system will be able to establish subsequent polymorphisms. This process has probably been important in the evolution of some of the classic examples of supergenes (systems of very closely linked loci held in strong linkage disequilibrium by selection), such as Batesian mimicry in butterflies (Ford, 1975) and sex chromosomes. Similarly, if ab and AB are both fitter than Ab and aB, a population fixed for ab may only evolve a two-locus polymorphism if there is a double mutation to AB and if \(c\) is sufficiently small. This probably occurred in the evolution of meiotic drive systems, which require combinations of alleles at several loci that are individually disfavored.
Second, there is a selective advantage to modifier alleles that reduce the frequency of genetic recombination between the two loci once a two-locus polymorphism has been established. If suitable genetic variation in recombination rates is available, this will eventually lead to such close linkage that the system has the appearance of a single locus (Fisher, 1930). This principle has wide generality; analysis of the conditions for spread of rare modifiers of recombination rates has shown that randomly mating populations under epistatic selection generating linkage disequilibrium at a system of loci will always tend to evolve closer linkage (Barton and Charlesworth, 1998). Since genetic recombination is a near-universal feature of living organisms, these findings have led to the search for situations that promote rather than repress recombination; these involve forces such as mutation and environmental change that perturb populations away from equilibria under selection (Barton and Charlesworth, 1998).

III. RANDOM GENETIC DRIFT

The discovery that random sampling of allele frequencies in finite populations may be a significant factor in evolution is another major contribution of population genetics. This process has two aspects: The first is the tendency for a population of finite size to become genetically uniform, owing to the fact that there is an increasing tendency as time passes for all the copies of a gene at a locus to be descended from a single ancestral allele (Fig. 2). The second is the tendency of isolated populations to diverge in allele frequencies over time, since independent trials of a population with the same initial state will arrive at different allele frequencies by chance.

The first process is closely related to the increase in homozygosity that accompanies the inbreeding of close relatives; both are conveniently studied by means of the concept of identity by descent. Two alleles at the same locus drawn from a population are said to be identical by descent if they trace their ancestry back to a single ancestral allele. The extent to which a population has progressed toward genetic uniformity can be measured by its inbreeding coefficient, defined as the probability that a pair of randomly sampled alleles are both derived from the same allele in the previous generation.

The simplest Wright–Fisher model of genetic drift assumes a discrete-generation, randomly mating population of N hermaphroditic individuals with no selective differences among genotypes at the locus under consideration. New individuals are formed by random sampling (with replacement) of gametes produced by the parents. With diploid inheritance, such that there are 2N gene copies at a locus among the breeding adults, the following recursion relation for \( f_t \), the inbreeding coefficient in generation \( t \), is obtained:

\[
1 - f_t = \left(1 - \frac{1}{2N}\right)(1 - f_{t-1}).
\]

This follows from the fact that the probability that a pair of randomly sampled alleles are both derived from the same allele in the previous generation is \( 1/(2N) \), in which case they are identical by descent. The probability that they come from two different alleles is \( 1 - 1/(2N) \); their probability of identity is \( f_t \).

Equations (5) shows that \( f_t \) tends asymptotically to 1; if \( N \) is moderately large, \( 1 - f_t \) decays exponentially with a half-life of \( 1.4N \) generations. This result can be generalized to more realistic types of breeding system by means of the concept of the inbreeding effective population number \( (N_e) \) (Crow and Kimura, 1970). This utilizes the fact that genetic drift in these more general cases can be described by matrix equations or
higher order difference equations, such that a constant rate of decay of \( 1 - f \) is reached asymptotically, yielding an expression of the same form as Eq. (5). The asymptotic decay constant can thus be equated to \( 1 - 1/(2N_u) \).

For example, a population with \( N \) breeding females and \( N \) males has \( N_u = 4N_N/(N + N_N) \). In general, \( N_u \) is less than the census number of breeding individuals, often considerably so (Crow and Kimura, 1970).

**B. Differentiation of Populations**

The effect of drift in causing genetic differentiation among populations is described by the variance in allele frequencies among the set of populations in question. In the Wright–Fisher model, the sampling of allele frequencies at a neutral biallelic locus with alleles \( A \) and \( a \) and frequencies \( p \) and \( q \) will generate a binomial distribution of the new frequency of \( A \), with mean \( p \) and variance \( pq/(2N) \). Repetition of this sampling process results in a probability distribution of allele frequencies with steadily increasing variance (Crow and Kimura, 1970). The variance of allele frequency, \( \sigma_A^2 \), can be related to the inbreeding coefficient as follows. Assume that we have an infinitely large set of completely isolated populations, all founded with the same initial frequency, \( p_0 \). At some arbitrary time, the set of populations will have a mean gene frequency \( p \) (since drift does not change the mean) and variance \( \sigma_A^2 \). If \( N \) is reasonably large, the genotype frequencies within each given population will be in Hardy–Weinberg proportions; the mean frequencies of \( AA \), \( Aa \), and \( aa \) over the set of populations are equal to \( p^2 + \sigma_A^2 \), \( 2pq \), and \( q^2 + \sigma_A^2 \), respectively.

This is the Wahlund effect: Genotype frequencies averaged over a set of populations that are individually in Hardy–Weinberg proportions show an excess of homozygotes and a deficiency of heterozygotes compared with Hardy–Weinberg expectation, whose value is determined by the variance in gene frequencies among the populations. This is a purely algebraic result, independent of the causes of the variation. It can be related as follows to the effects of genetic drift in causing increased homozygosity. Assume that a diploid individual is formed by sampling two random alleles from the same population. The probability that the alleles are identical by descent is \( f \); they are then both \( A \) in state with probability \( p \). The probability that they are non-identical is \( 1 - f \), in which case the probability that the two alleles are both \( A \) is \( p^2 \). The net probability that the individual is \( AA \) is thus \( fp_0 + (1 - f)p^2 \). Comparison with the previous expression shows that

\[
\sigma_A^2 = fpq. \tag{6}
\]

This establishes that the variance in gene frequency is proportional to the inbreeding coefficient under the Wright–Fisher model, and therefore its change over time is governed by Eq. (5). This is often but not always true under more general models of population structure, with \( N_i \) replacing \( N \) in the binomial formula for the variance conditional on the current gene frequency. In some circumstances, particularly when the population size changes in time, the conditional variance in gene frequency requires a different denominator in order to be represented by the binomial formula. In this case, a variance effective population number is computed (Crow and Kimura, 1970).

**C. Molecular Evolution and Variation**

1. The Neutral Theory

These simple models of genetic drift can readily be applied to the study of molecular evolution and variation, assuming selective neutrality at the loci in question. Neutral theory allows for the possibility that many mutations are subject to purifying selection and are rapidly eliminated from the population (see Section IV,D,3); it is claimed that the fate of the bulk of the mutations that are not removed by purifying selection is determined by drift rather than selection (Kimura, 1983). This theory thus constitutes a useful null hypothesis, which can be tested against data on molecular evolution and variation by means of predictions of several types.

2. The Rate of Neutral Evolution

Consider first the rate of molecular evolution as measured by the rate of gene substitution, \( K \) (see Section II,G,4). In a population of \( N \) breeding adults, there are \( 2N \) allele copies at an autosomal locus. As shown in Section III,A, Eq. (5) implies that the population tends to homozygosity with probability 1. This means that the remote descendants of the current population will all trace their ancestry back to just one of these \( 2N \) alleles. Under neutrality, the probability that a given allele is the ancestor is thus \( 1/(2N) \). It follows that the probability of fixation of a new neutral mutation in a population of size \( N \) is \( 1/(2N) \); the probability that it is lost is \( 1 - 1/(2N) \). If the rate of mutation to neutral variants is \( u \) per generation, the expected number of new mutations that enter the population is \( 2Nu \), of which only \( 1/(2N) \) are destined for ultimate fixation; the expected number of mutations that ultimately become fixed is \( 2Nu/(2N) = u \).
If the process of mutation and drift has reached a stationary state, so that the expected number of new substitutions must equal the number of substitutions that go to completion each generation, we obtain the fundamental equation of neutral molecular evolution:

$$K = \bar{u}$$  \hspace{1cm} (7)

This relation holds for any level in the genetic hierarchy, from the nucleotide site through the locus to the genome as a whole, provided that $K$ and $\bar{u}$ are defined appropriately. Since $K$ can be determined by comparisons of DNA sequences among species with known divergence time (see Section II,G,4) and $\bar{u}$ values are known from molecular genetics, Eq. (7) can be used to test the neutral theory.

One prediction is that genomic regions whose sequences are essentially functionless, such as pseudogenes and the internal parts of introns, should evolve at the mutation rate since they are necessarily unconstrained by selection. If adaptive Darwinian evolution is a minor factor in molecular evolution, the rate for sites in these regions should be much higher than that for functionally significant regions, where selection is expected to eliminate most mutations [see Equation (10)]. These regions do indeed evolve at the rates expected from mutation rates, and other regions evolve more slowly (Kimura, 1983; Li, 1997). This does not, however, rule out a role for positive selection in fixing variants in selectively constrained regions; it could still be true, for example, that most replacement substitutions (see Section I,B,2) are deleterious, but a minority are advantageous rather than neutral, so that changes that are fixed in evolution are adaptive rather than neutral. There are also some exceptional cases of higher rates of replacement versus silent substitutions in coding regions; this is strong evidence for a positive role of selection on the amino acid sequences concerned (Li, 1997).

Another prediction of Eq. (7) is that the rate of molecular evolution should be constant over long periods of time since it depends only on the mutation rate. In contrast, the rate of evolution under natural selection is expected to be highly variable since the theory described earlier suggests that populations will tend to adapt quickly to a new environment (or go extinct), after which change will be slow or nonexistent. This is borne out by the observations of comparative biology and paleontology, which show that evolution at the external phenotypic level is generally highly episodic and triggered by ecological opportunities such as the occupation of vacant niches. Studies of DNA and protein sequence evolution suggest that the rate of evolution of a given molecule is much less variable among different lineages, or within the same lineage at different times, than is true for the external phenotype, especially when noncoding or silent substitutions are considered (Kimura, 1983; Li, 1997). This has generated the concept of a molecular clock, which is used to estimate the times of divergence of taxa when paleontological data are absent. However, there is evidence for more variability in rates of evolution than predicted by the simplest form of the neutral theory, especially for amino acid sequences. This suggests a role for selection, although the interpretation of rate variability is controversial (Li, 1997).

3. Neutral Polymorphism

The process of fixation of neutral variants is a slow one; calculations based on diffusion theory (see Section IV,D) show that the mean time to fixation of a new neutral mutation in a random mating population (conditioned on ultimate fixation) is approximately $4N_e$ generations. While the variant is on the way to fixation, it causes a polymorphism. Similarly, variability is contributed by the large fraction of new mutations that are destined for ultimate loss; the mean time to loss is $(N_e/N) \ln (2N_e)$ generations, a much shorter time than the time to fixation. In the neutral theory, polymorphism is simply a phase of molecular evolution. Although there is a constantly shifting set of variants at any one locus, the mean amount of variability can be determined assuming a statistical equilibrium between drift and mutation, using the following argument (Kimura, 1983).

The simplest version of this applies to a single locus, which is assumed to have no recombination. Thus, these are many possible sequences; new neutral mutations are assumed to occur with probability $u$ per generation, and each mutation represents a sequence that has not been observed before. This is the infinite alleles model. Let $h_i$ be the probability that two randomly sampled alleles are distinct in sequence in generation $t$. An argument similar to that leading to Eq. (3) shows that, neglecting terms in $u/(2N)$, $h_i$ for a Wright–Fisher population obeys the equation

$$h_i = \left(1 - \frac{1}{2N} h_{i-1}\right) + 2u(1 - h_{i-1}). \hspace{1cm} (8a)$$

At equilibrium, rearrangement of this equation yields
the relation
\[ h = \frac{4Nu}{4Nu + 1}. \]  

More generally, \( N_i \) can be substituted for \( N \) in Eq. (8b). The parameter \( \theta = 4Nu \) thus controls the equilibrium level of variability under the neutral model. A large amount of neutral variability can be maintained, provided it is sufficiently large, e.g., with \( N_i = \theta \) and \( u = 2 \times 10^{-5}, \theta = 0.08, \) and \( h = 0.074. \) This value of \( h \) is similar to the mean per locus heterozygosity for electrophoretic alleles in mammalian populations (Lewontin, 1974; Kimura, 1983).

This model can be modified to predict the equilibrium level of diversity per nucleotide site by assuming that the units of observation are the individual sites, not the entire locus. If we assume that \( \theta \) is much less than 1 so that at most one variant is segregating in the population at each site, Eq. (8) can be applied to yield the equilibrium value of \( \pi \), the nucleotide site diversity (see Section III,B,2), such that \( \pi = \theta \) to the assumed order of approximation. This is Kimura's infinite sites model, which is widely used in the interpretation of data on molecular variation.

D. The Coalescent Process

1. General Considerations

The growing body of data on DNA sequence variation within populations has stimulated interest in the development of statistical tests of the agreement of observed patterns of variation with the predictions of the neutral theory. In order to conduct such tests, it is essential to have predictions concerning the properties of statistics describing samples of alleles from populations and not merely of properties of the populations from which the samples are drawn since these cannot be observed directly. A powerful method for deriving properties of samples has recently been developed and is known as coalescent theory (Hartl and Clark, 1997; Li, 1997). It is based on the following principle. Consider a pair of alleles at an autosomal locus sampled from a Wright–Fisher population. As shown in Section III,A, there is a probability 1/2(2N) that they are derived from a common ancestral allele in the previous generation, i.e., that they coalesce. If they fail to coalesce, which has probability 1 – 1/2(2N), they have a probability 1/2(2N) of coalescing in the next generation back, and so on.

There is thus a geometric distribution of the time back to the common ancestral allele, such that the probability of time \( t \) is \( (1/2N)(1 – 1/2N)^{t – 1}. \) From the well-known properties of this distribution, the mean time to the common ancestor is \( 2N \) and the variance is \( (2N)^2. \) Assume that mutations occur at rate \( u \) per site in the gene, such that each mutation that arises in the line of descent connecting the two sampled alleles is at a different site (the infinite sites assumption, see Section III,D,3). If there are \( m \) nucleotide sites in the sequence in question, the number of mutations conditioned on \( t \) has a mean and variance of \( 2mu \) since the total time separating the alleles is \( 2t \), and the conditional number of mutations follows a Poisson distribution. The mean and variance of the number of differences between the two alleles are thus \( m\theta \) and \( m\theta + 0.5(m\theta)^2 \), respectively. The result for the mean corresponds to that derived from Eq. (8) since the nucleotide site diversity is simply the number of differences between the pair of alleles divided by \( m \).

This can be extended to a sample of \( n \) alleles from a population. If \( N \) is sufficiently large, the chance of more than one coalescent event per generation can be neglected. There are \( n(n – 1)/2 \) possible pairwise allelic combinations so that the probability of a coalescent event is \( n(n – 1)/(4N) \). The time to the first coalescent event is therefore distributed approximately exponentially, with a mean of \( 4N/n(n – 1) \) and variance equal to the square of the mean. The time from this to the next coalescent is also exponentially distributed, replacing \( n \) with \( n – 1 \), and so on. The process can be represented by a genealogical tree, whose nodes with \( h \) and \( k – 1 \) alleles are separated by a time \( t \), that is exponentially distributed with mean \( 2/(k(k – 1)) \) on the coalescent timescale of \( 2N \) generations (Fig. 3). As before, \( N \) replaces \( m \) for more general models of breeding structure. The rate of coalescence evidently diminishes as the number of nodes decreases.

This representation can be used to derive many important results, and it also provides a rapid method of simulating genetic processes, since time can be rescaled to the coalescent timescale and the properties of a sample of alleles represented by generating genealogical trees from samples drawn from the relevant exponential distributions, with mutations scattered randomly over the branches of the tree. The expected pairwise difference between all \( n(n – 1)/2 \) pairs of alleles on the infinite sites assumption is readily seen to be \( m\theta \). Its variance is also known (Li, 1997). This statistic provides the obvious means of estimating the nucleotide site diversity in the population (\( \theta \) by equating the observed mean pairwise difference to its expectation. However, another statistic, the total number of segregating sites in a sample of \( n \) alleles (\( S \)), has better statistical properties under the infinite sites model. These can be obtained as follows. The total size of a tree is the sum of \( k_0 \), over the entire tree. Application of the properties of the
Figure 3: Two types of gene genealogies. (A) The genealogy of a set of alleles in a population of constant size. The tips of the tree represent six alleles sampled from the population. Each node corresponds to a coalescent event. Each cross line represents a new mutation in the DNA sequence of the gene in question, which has arisen since the common ancestor of the sample; mutations in internal branches of the tree give rise to at least two variants in the sample. (B) The genealogy of a set of three alleles in a population that experienced a recent reduction in size to one allele, followed by rapid expansion with no opportunity for coalescent events.

The exponential distribution shows that the expectation of the tree size in units of coalescent time is simply $A = 2\left(\frac{1}{N} + \frac{1}{N} + \frac{1}{N} + \cdots \frac{1}{N}\right)$. The number of sites in the sample that are segregating for nucleotide variants is the number of mutations that appear on the tree. Its expectation is $2\frac{NmuA}{H}$; therefore, $\theta$ can be estimated by dividing $S$, by mA. The variance of this estimator is substantially lower than that based on pairwise diversity.

2. Tests of Neutrality
Several tests for departures from neutrality have been devised based on the properties of the coalescent process (Li, 1997). One widely used method is Tajima’s D test, which uses the fact that departures from neutrality have different effects on the pairwise difference estimate of $\theta$ than on the $S$ estimate. For example, a population bottleneck of reduced size that completely removes variability at a locus will be followed by a long period of recovery, during which new mutations are all at low frequencies. This means that the pairwise diversity, which is weighted by the frequencies of variants, will be much lower in relation to the observed number of segregating sites than in an equilibrium population, reflecting the fact that the genealogical tree is like a “star phylogeny” in this case (Fig. 3). Purifying selection on the variants in question will have a similar but much smaller effect. The opposite pattern is expected if there has been a partial bottleneck, in which rare variants have been lost from the population, or if variation is affected by balancing selection. The Tajima test computes the ratio of the difference between the two estimates of $\theta$ to the standard deviation of their difference (calculated using the neutral equilibrium model) and compares the result with critical values obtained from coalescent simulations. Other tests have been derived from related principles (Li, 1997).

IV. THE INTERACTION OF DRIFT WITH DETERMINISTIC FORCES
A. Population Subdivision
1. General Considerations
One important complication of the neutral theory is population subdivision; species in nature are not simple randomly mating populations but instead are usually distributed over wide geographic areas, with limited migration between localities. Random genetic drift can thus produce significant genetic differentiation among local populations if migration is sufficiently restricted. This creates two problems: How to describe data on allele or nucleotide site variant frequencies when there is differentiation among populations, and how to relate models of the underlying evolutionary processes to the data.

2. Partitioning Variability
The basic method used to summarize data collected from a set of populations sampled from the same species is to partition variability into a within-population component that describes the mean level of variability within a sample and a between-population component that measures the mean difference between alleles sampled from different populations. This can be done either using data on allele frequencies, as in the case of electrophoretic loci or microsatellite loci, or using nucleotide site variant frequencies. In many cases, it is also possible to organize the populations sampled into a hierarchy of reduced size that completely removes variability at a locus will be followed by a long period of recovery, during which new mutations are all at low frequencies.
measures each lower level measure of between-population variability relative to variability at the next higher level of the hierarchy, generating a set of Wright’s $F$ statistics or the related $G$ statistics of Nei (Harrl and Clark, 1997).

The procedures will be illustrated here for the case of DNA sequence variation in a population divided into a set of equal-ranking subpopulations. The total nucleotide site diversity, $\pi_s$, is defined as the fraction of nucleotides that differ between a pair of alleles drawn randomly from the set of samples as a whole. The within-population nucleotide site diversity, $\pi_w$, is the mean fraction that differ between a pair of alleles drawn from the same subpopulation. The amount of between-population differentiation can be measured by the between-population component of nucleotide site diversity, $\pi_{w\rightarrow s} = \pi_s - \pi_w$. Alternatively, a measure of nucleotide site divergence, $\pi_D$, can be defined as the mean fraction of nucleotide sites that differ between pairs of alleles drawn from a pair of different subpopulations, and another measure of between-population differentiation is defined as $\pi_D = \pi_w - \pi_s$. Two normalized measures of differentiation can be defined as $G_s = \pi_s/\pi_D$ and $F_{st} = \pi_{w\rightarrow s}/(\pi_w + \pi_s)$. Both measures are widely used in the literature and are often similar numerically, especially if many subpopulations are sampled.

3. The Island Model

Although the previously mentioned measures of population differentiation are useful as descriptive tools, they can also be used to estimate the evolutionary parameters that determine the extent of population differentiation. This requires the development of models of the joint effects of genetic drift, mutation, and migration, one of the most complex problems in theoretical population genetics. The simplest and most widely used model is the island model, which assumes that the species is divided into $n$ distinct subpopulations or demes, which each behave according to the Wright–Fisher model with population size $N$. After reproduction has occurred within each deme, a fraction $m$ of each deme’s genes are replaced with genes drawn randomly from the other $n-1$ demes. Coalescent theory can be used to determine the mean coalescence times of pairs of alleles sampled from the same population ($\langle \tau \rangle$) or from different populations ($\tau_0$), $\tau_0 = 2N$ and $\tau_0 = 2Nn/(1 + n - 1)/[\langle \tau \rangle N]$. In the infinite sites model, the mean fraction of nucleotides that differ between a pair of alleles is equal to the product of the mutation rate per site and twice their coalescence time (see Section III.D) so that the expected nucleotide site differences between alleles can be derived directly from the corresponding coalescence times.

An important and somewhat counterintuitive conclusion is that the within-population nucleotide site diversity, $\pi_w$, is equal to $4Nm$, i.e., it depends on the total number of individuals in the set of populations in the same way as does the diversity in a panmictic population under the infinite sites model, and it is independent of the migration rate (with the proviso that $m > 0$). As expected, the other diversity measures depend inversely on $Nm$, with large between-population divergence being possible only when $Nm < 1$; for large $n$, both $F_{st}$ and $G_s$ are approximately equal to $1/(1 + 4Nm)$. Values of these statistics that are close to zero are generally taken to indicate relatively little population differentiation, whereas values close to one imply considerable differences among local populations relative to within-population variability.

4. Other Models of Population Structure

Although simple to analyze, the island model is not very realistic biologically since dispersal is limited in most species so that local populations are most likely to acquire immigrants from nearby. Attempts to generate useful results from more realistic models have taken two directions. The first involves maintaining the assumption of a set of discrete demes but allowing for differences in deme sizes. Migration is described by a migration matrix $M$, such that $m_i$ is the probability that an allele in deme $i$ originated in deme $j$ in the previous generation. However, it is difficult to obtain transparent general results for such a model. Under the infinite sites model, however, the result that $\pi = 4Nm$ still holds if migration is conservative, i.e., migration does not change the sizes of local populations (Maruyama, 1977; Nagylaki, 1986).

Results on genetic diversity within and between populations are only available for special cases, such as the stepping-stone model, in which demes of size $N$ in a linear or planar array receive migrants with probability $m$ only from immediately adjacent populations (Maruyama, 1977). This model also permits the analysis of the dependence of degree of genetic differentiation between populations on the distance between them. The results show that extensive differentiation between populations is possible with a one-dimensional array of populations, even if there is a considerable amount of migration, whereas the results for a two-dimensional array are similar to those for the island model.

The second approach assumes that individuals are distributed over a one or two-dimensional spatial continuum, with density $\rho$. Migration is represented by the probability density that an individual moves a given distance between birth and reproduction. If migration follows a random walk, the variance of the migration...
B. Effects of Directional Selection at Linked Loci

Another complication in interpreting data on DNA sequence variation is that even neutral variation may be affected by selection at linked sites. The classic example is hitchhiking, whereby a new favorable mutation arises as a unique event and spreads through the population. In the absence of recombination, variants at linked loci will be dragged to fixation as the favorable allele sweeps through the population so that variability at these loci is eliminated. With recombination, the magnitude of this hitchhiking effect decreases with the ratio of the frequency of recombination between the selected and neutral locus (c) to the selection coefficient (s) at the selected locus, and it is negligible when this ratio is on the order of 1. The effects of hitchhiking by favorable mutations are thus only likely to be manifest at sites that are closely linked to the target of selection. There are several natural examples of such hitchhiking events, including an increased frequency of a DNA sequence variant linked to the sickle-cell anemia mutation.

However, hitchhiking effects can also be caused by deleterious mutations, a process termed background selection (Hartl and Clark, 1997; Li, 1997). A neutral variant which is tightly linked to a deleterious mutation is in the process of elimination from the population so that variability at these loci is eliminated. With recombination, the magnitude of this hitchhiking effect decreases with the ratio of the frequency of recombination between the selected and neutral locus (c) to the selection coefficient (s) at the selected locus, and it is negligible when this ratio is on the order of 1. The effects of hitchhiking by favorable mutations are thus only likely to be manifest at sites that are closely linked to the target of selection. Selection at linked sites can also cause increased variability at neutral sites if selection is balancing rather than directional in nature. This can be understood in terms of a single locus with two alleles, A and a, maintained by strong balancing selection in a randomly mating population. This means that the population is effectively divided into two subpopulations represented by the two allelic classes. If the two alleles are equally frequent, a neutral site linked to this locus with recombination frequency c has a probability \( 0.5 c^2 \) of crossing over from one subpopulation to the other, this is equivalent to the migration rate m in the case of an island model with just two islands. This leads to an expansion of the coalescent time at the neutral site by a factor of \( 1 + 1/(4Nc^2) \), where N is the effective population size. This implies that the maintenance of variability by selection is accompanied by a corresponding increase in nucleotide site diversity at sites that are very closely linked to the target of selection, as seen at the Adh locus in D. melanogaster and at the MHC loci in mammals (Li, 1997). Increases in neutral variation among subpopulations of a species may also occur at linked sites when there is local selection, causing large among-population differences in allele frequencies in different populations and at sites closely linked to the targets of selection. Variation in the behavior of neutral variability along a chromosomal region can thus provide valuable evidence on the action of selection.

D. Diffusion Equations

1. General Considerations

A full treatment of genetic drift must deal with properties other than summary statistics of allele frequency distributions of the type considered so far, particularly if selection is to be studied. The difficulties involved in exact analytical treatments of the properties of genotype frequency distributions are great, so resort is usually made to approximations involving the use of diffusion equations, which treat genotype frequencies and time as continuous variables and assume that all evolutionary forces are sufficiently weak that second-order terms in their effects on frequencies are negligible. The standard approach is to assume a single biallelic locus. Using the
continuity assumptions, the state of the population can be described by either of the following two partial differential equations, writing \( \phi(x, p, t) \) for the probability density of the frequency, \( x \), of allele \( A \) at time \( t \), given initial frequency \( p \):

\[
\frac{\partial \phi}{\partial t} - \frac{1}{2 \sigma^2} \frac{\partial^2 \phi}{\partial x^2} - \frac{\theta}{2 \sigma^2} \frac{\partial \phi}{\partial x} = 0
\]  

(9a)

and

\[
\frac{\partial \psi}{\partial t} = \frac{V_s}{2 \theta} \frac{\partial \phi}{\partial p} + M_b \frac{\partial \phi}{\partial x} 
\]  

(9b)

where \( M_b \) is the expected change in gene frequency and \( V_s \) is the variance of the change in gene frequency, both conditioned on \( x = M_b \), can be equated to the change in gene frequency in an infinite population, under random sampling of allele frequencies, \( V_s = x(1 - x)/(2N) \).

Equation (9a) is the forward Kolmogorov equation and Eq. (9b) is the backward Kolmogorov equation (Crow and Kimura, 1970; Ewens, 1979). Multidimensional versions of these equations describe systems with multiple alleles and multiple loci but are usually difficult to use.

2. Stationary Distributions

Equation (9a) is most useful for describing the probability density function. However, even for the simple case of a biallelic locus, a full general solution of this equation to yield the density as a function of time has been obtained only for some special cases, such as no selection. It is most useful for studying the properties of stationary distributions of gene frequency, when drift comes into statistical equilibrium with mutation, migration, and selection.

The study of such distributions has led to some important conclusions, most notably that selection is effective at countering the effects of drift in a randomly mating population when \( Ns \approx 1 \) but is ineffective when \( Ns \ll 1 \). When the first condition holds, there is little scatter around the mode of the gene frequency distribution; when the second condition holds, there is a high probability that the population is close to fixation, or fixed, for the disfavored allele. If \( Ns \) is on the order of 1, both drift and selection are significant forces.

3. Fixation Probabilities

The previous conclusion can also be derived by consideration of the probability of fixation of an allele, \( P(p) \), for which a general formula was found by Kimura using the backward equation (Crow and Kimura, 1970; Ewens, 1979). With intermediate dominance, such that \( h = 0.5 \) in Eq. (1), and assuming that a single copy of the mutation is present initially (\( p = 1/(2N) \)), this formula simplifies as follows, first discussed by Fisher (1930):

\[
P \left( \frac{1}{2N} \right) = \frac{1 - \exp \left( -\frac{N}{N_s} \frac{x}{2} \right)}{1 - \exp (-2N_s x)}
\]  

(10)

When \( s > 0 \), the fixation probability tends to \( (N/N_s) \) as \( N \) tends to infinity. When \( N = N \) (so that the Wright–Fisher model applies), this is equivalent to the branching process result for the survival probability of a favorable mutation in a very large population (see Section II, C, 3). The asymptote is approached when \( Ns \gg 1 \) so that a selection coefficient larger than \( 1/N_i \) is sufficient to ensure that a new favorable mutation behaves as though the population is infinite. Conversely, a selection coefficient of this order ensures that a deleterious mutation (\( s < 0 \)) is almost certain to be eliminated from the population.

This led Fisher to conclude that random genetic drift is unlikely to be effective as an evolutionary force on the grounds that (i) most species have numbers of breeding individuals at least in the tens of thousands, and usually in the hundreds of thousands or even millions, and (ii) it is unlikely that a gene would have such a small effect on the phenotype that its average effect on fitness over evolutionary time would be on the order of \( 10^{-6} \) or less. Although a compelling argument with regard to genetic changes that affect the phenotype, this view has been challenged by the neutral theory of molecular evolution. As already noted, the causes of protein sequence evolution remain controversial, but studies of the statistical properties of between- and within-species patterns of silent substitutions at third coding positions support the idea that this is affected by both drift and selection in bacteria, yeast, and Drosophila. There is a tendency for certain triplets that code for the same amino acid to be favored by selection, but there is also evidence that this selection is so weak that disfavored nucleotide changes can drift to fixation within a species. Estimates of the intensity of selection can be obtained by comparing observed distributions of frequencies of silent-site variants with those predicted by solutions of the diffusion equations (Harl and Clark, 1997).

E. Muller’s Ratchet

Selection may become ineffective if recombination is restricted among the loci subject to selection as a
result of linkage disequilibrium generated by random drift (the Hill–Robertson effect; Barton and Charlesworth, 1998). One process that has been much studied in this context is Muller's ratchet, which operates in a finite population subject to mutation to deleterious alleles at many loci. Consider, for example, the case of a haploid asexual population in which mutations occur exclusively from wild-type to deleterious alleles but not in the opposite direction (see Section II,F) and where $N_e > 1$. If the selective effects of mutations at different loci are identical, a population can be characterized by the frequencies of genomes containing 0, 1, 2, … mutations. If the frequency of the mutational class containing the lowest number of mutations (the least-loaded class) is sufficiently small, it will be lost from the population after a finite number of generations. Given the assumed irreversible nature of mutation and the lack of opportunity for genetic recombination, the least-loaded class cannot be reconstituted and will be replaced by the class with one more mutation. This class is now vulnerable to stochastic loss in the same way. There is thus a repetitive process of loss of successive least-loaded classes, in which the loss of each class can be regarded as a turn of Muller's ratchet. The ratchet can only operate at an evolutionarily significant rate if the equilibrium number of individuals in the least-loaded class in an infinite population ($n_\text{e}$) is relatively small (e.g., $< 100$); if it does operate, there will be an approximately exponential decline in the mean fitness of the population. Given the results discussed previously, which suggest that a typical Drosophila individual may carry 30 or more detrimental alleles so that the frequency of mutation-free individuals is approximately exp $(-30) = 9 \times 10^{-14}$, even a very large asexual population of a higher organism is likely to be vulnerable to the operation of the ratchet, leading to its eventual extinction. Asexual prokaryotes, with their much smaller genomes and enormous population sizes, are unlikely to suffer this fate. This may account for the fact that very few species of higher organisms, with their large genomes, are asexual, whereas prokaryotes and mitochondria have very low levels of recombination (Barton and Charlesworth, 1998). It may also contribute to the evolutionary degeneration of Y chromosomes, which are usually largely devoid of active genes.

**F. Group Selection**

Another possibility that violates the supremacy of individual selection occurs when species are subdivided into local populations, among which migration is so restricted in relation to the reciprocal of local population size that drift can overcome its homogenizing influence (see Section IV,A). An allele that is deleterious within its local population may drift to fixation locally in opposition to selection since $N_e$ for the local population is much smaller than for the species as a whole. This raises the possibility of group or interdeme selection. If the allele in question causes its carriers to be altruistic in the sense of conferring increased fitness on the members of their deme at the expense of suffering a loss in fitness to themselves, demes in which the allele reaches high frequency or fixation will achieve a higher mean fitness. This may render them less susceptible to extinction and more able to contribute to the pool of migrants or to found new demes. Selection among demes can therefore result in the spread of an allele that causes a loss in individual fitness but benefits the population as a whole (Haldane, 1932; Hartl and Clark, 1997).

Although this is a theoretically viable mechanism, there are reasons to doubt that it is widely applicable to evolution in nature. First, studies of molecular genetic variation within and among populations indicate that many species lack extensive differentiation among local populations (see Section IV,A,2), suggesting that migration is so effective that the necessary conditions for group selection are often not met. Second, it does not seem capable of producing a stable evolutionary outcome: If an altruistic allele becomes fixed in a set of populations, a “selfish” opponent that is reintroduced by migration or mutation will have a good chance of spreading through the species since the outcome of the conflict between group and individual selection is probabilistic and not deterministic.

**G. Kin Selection**

Altruistic behavior, such as the sterility of the worker castes in social insects, is generally believed usually to be due to kin selection. This is based on the fact that an altruistic genotype that aids relatives may experience a selective advantage, even in a large, randomly mating population, if the fitness benefit $b$ to the recipients of the altruism is sufficiently large in relation to the cost $c$ to the altruist (Fisher, 1930; Haldane, 1932; Hartl and Clark, 1997). This is because the genetic similarity of related individuals means that the relatives of an altruist have a greater chance than average of carrying an allele that promotes altruism; according to Hamilton's rule, there can be an increase in the frequency of an altruistic allele
when \( br > c \), where \( r \) is a measure of the relatedness of the altruists to the recipients. There is an extensive theoretical literature on the correct way to calculate \( r \) under different conditions; this result has played an important role in the evolutionary interpretation of social behavior in animals.

**H. The Shifting Balance Theory**

The second objection to group selection is overcome by the related model of Wright, the shifting balance theory of evolution (Wright, 1977; Hard and Clark, 1997). Wright postulated that epistatic interactions in fitness among alleles at different loci are widespread, resulting in multiple stable equilibria under selection. The simplest case is a haploid two-locus, two-allele model, with \( A \) and \( B \) both fitter than \( ab \) and \( AB \). Fixation for \( A \) or \( B \) is stable against introduction of \( ab \) and \( AB \), whereas fixation for \( AB \) or \( ab \) is unstable to the introduction of \( A \) and \( B \). With constant fitnesses and loose linkage, locally stable equilibria are approximated by the peaks in the surface of mean fitness as a function of the allele frequencies at the loci concerned (see Sections II.C.1 and II.H.2). A population will approach the peak that is the nearest attractor rather than the highest peak in the landscape. Genetic drift can cause a local population to travel down the adaptive valley separating the current equilibrium to the zone of attraction of a neighboring peak, and selection can then bring it to the new equilibrium. If this is associated with a higher mean fitness than the surrounding peaks, the process of interdeme selection can cause the species as a whole to acquire the genotype associated with this peak, and hence improve in mean fitness. In contrast to the group selection model of altruism, the new equilibrium is dynamically stable, and so there is only a low probability of reversal. This process has the attractive feature that it allows the species to acquire adaptively superior genotype or character combinations that would require multiple simultaneous mutations to be produced in a large population. Its drawback is that it requires a delicate balance between restricted migration, local population size, and the nature and intensity of selection if it is to operate with any frequency. It is difficult to distinguish the end products of the shifting balance process from ordinary individual selection, which does not have such stringent requirements, and it is unclear to what extent it has played an important role in adaptive evolution (Kimura, 1983).

**V. CONCLUSIONS**

This article necessarily omitted many important topics. It focused on the basic general principles governing evolutionary change in populations; space did not permit more than a brief mention of the application of these principles to wider biological problems, including life history evolution, the evolution of genetic and sexual systems, the evolution of social behavior, speciation, and the interpretation of macroevolution. Very little useful can be said about natural variation and evolution at almost any level without taking population genetic concepts into account.

*See Also The Following Articles*

**DIVERSITY, MOLECULAR LEVEL • EVOLUTION, THEORY OF • INBREEDING AND OUTBREEDING • PHENOTYPE, A HISTORICAL PERSPECTIVE**

**Bibliography**


POPULATION STABILIZATION, HUMAN

Alene Gelbard
Population Reference Bureau

I. History of Human Population Growth
II. Causes and Effects of Human Population Growth
IV. Responses to Human Population Growth
V. Conclusion

GLOSSARY

family planning  The conscious effort of couples to regulate the number and spacing of births through artificial and natural methods of contraception. Family planning connotes conception prevention to avoid pregnancy and abortion, but it can also refer to efforts of couples to induce pregnancy.
fertility  The actual reproductive performance of an individual, a couple, a group, or a population measured by a variety of rates, including the crude birth rate (number of births per 1000 population) and total fertility rate.
life expectancy  The average number of additional years a person could expect to live if current mortality trends were to continue for the rest of that person’s life; most commonly cited as life expectancy at birth.
mortality  Deaths as a component of population change. Measures include the crude death rate (number of deaths per 1000 population).
population composition  The distribution of a population, usually by age and sex, measured by the number and proportion of males and females in different age groups.
population momentum  The tendency for population growth to continue beyond the time that replacement-level fertility has been achieved because of the relatively young age structure of the population.
replacement-level fertility  The level of fertility at which a couple replaces itself in the population. In low mortality countries, a total fertility rate of 2.1 is considered replacement because some women will die before the end of their childbearing years.
stable population  A population with an unchanging rate of growth and an unchanging age composition as a result of age-specific birth and death rates that have remained constant over a sufficient period of time. Zero population growth is a special case of population stabilization in which no growth occurs because the number of births and deaths are the same.
total fertility rate  The average number of children that would be born to a woman during her lifetime if she were to pass through her childbearing years conforming to the age-specific fertility rates of a given year.

HUMAN POPULATION CHANGE has three components: births, deaths, and migration. On a global level,
only births and deaths determine changes in population size. In the 1990s, more than 70 million people were added to the world’s total population each year and it took only 12 years for the world’s population to increase from 5 to 6 billion people. This growth will continue into the twenty-first century. How much growth will occur and how quickly, and when or whether this growth will stabilize, depends on a host of factors. In 1994, the world community agreed to the need to stabilize global population growth. This article describes the history of global population growth, factors associated with this growth, and how countries view this growth. It concludes with a discussion of prospects for the stabilization of population growth in the future.

I. HISTORY OF HUMAN POPULATION GROWTH

For much of our history, humans have struggled to survive. By 1 A.D., perhaps 300 million people lived on Earth, a paltry total after millions of years of human existence. For most of the next 2000 years, population growth was exceedingly slow. High birth rates were often offset by frightful mortality from wars, famines, and epidemics. The bubonic plague, for example, reduced the populations of China and Europe by one-third in the fourteenth century.

Despite dramatic spikes in mortality rates, the number of births exceeded the number of deaths during the seventeenth and eighteenth centuries and population growth proceeded at a slightly faster pace. As shown in Fig. 1, world population was about 790 million in 1750 and reached 1 billion in approximately 1800.

During the next century, in Europe and in a few other areas throughout the world better hygiene and public sanitation reduced the incidence of disease. Expanded commerce made food supplies more widely available and improved nutrition. The wild fluctuations in mortality of previous centuries began to recede, and life expectancy began a slow rise. Population grew more quickly and more steadily. Total world population was nearly 1.7 billion by the beginning of the twentieth century and reached 2 billion within the next 30 years.

The nineteenth-century surge of population growth occurred primarily in the more developed countries. The population of Europe more than doubled between 1800 and 1900, whereas the population of North America increased nearly 12 times—fueled by immigration from Africa and Europe. In 1800, about one-fourth of the world population lived in the now more developed regions of Europe (including Russia), Japan, and North America, but this share increased to about one-third by 1900.

Less developed countries grew more slowly than more developed countries in the nineteenth century, but they already held the bulk of the world’s inhabitants. Asia, dominated by China, had 62% of the world population in 1800, and Africa had 11%. Latin America and the Caribbean accounted for only about 2% of the world’s population. Like North America, Latin America would see most of its population growth in the twentieth century.

A. Demographic Transition

The improvement in human survival and the consequent explosion of population growth marked the beginning of the shift from high to low mortality and from high to low fertility that is known as the “demographic transition.” This shift occurred throughout Europe, North America, and many other areas in the nineteenth and early twentieth centuries. It gave rise to the dominant model of demographic change, which most demographers assume applies to all countries. In the classic demographic transition, the trend of high birth and death rates (and minimal population growth) is disrupted by a long-term decline in mortality. Mortality rates eventually stabilize at low levels. Birth rates also begin a long-term decline and decrease to about the same level as mortality rates. With birth and death rates at similar low levels, the equilibrium of slow population growth is regained.

The pace of change in a country will vary depending
on its culture, level of economic development, and other factors. As countries pass through the various stages of the transition, population growth from natural increase (birth rate minus death rate) accelerates or declines depending on the gap between the birth rates and death rates. Not all countries will follow the same path to low fertility and low mortality as did European countries. Also, there may be additional stages of transition that have not been identified, e.g., long-term population decline. However, the demographic transition theory provides a useful framework for assessing demographic trends and projecting future population size.

B. Population Change: 1900–1950

When the twentieth century began, more developed countries were entering a new stage of the demographic transition. In 1900, life expectancy at birth was 47 years in the United States and between 45 and 50 years in Europe, Japan, and Australia, which was a slight increase from an average of about 40 years during the nineteenth century. A revolution in health had already begun, and life expectancy would reach unimaginably high levels by mid-century. These improvements in health reflected scientific advances of the previous century. Louis Pasteur, Robert Koch, and others had identified disease-causing “germs,” and Joseph Lister introduced antiseptic practices that were eventually adopted by hospitals. Mortality was also declining because of better personal hygiene and public sanitation projects that removed garbage and sewage from city streets and provided safer drinking water. Death rates for infectious diseases began to decrease well before vaccines and antibiotics were widely available.

Infants and young children benefited most from this health revolution. In the more developed countries, the infant mortality rate (IMR; number of deaths to infants less than 1 year of age per 1000 births) was about 200 in the 1800s—about 2 of every 10 babies died before their first birthday. In the early 1900s, the IMR decreased below 100 in the United States and many European countries. It was lower than 50 in nearly all these countries by the 1950s. United States life expectancy at birth increased to 56 years by 1920 and to 68 years by 1950. Average life expectancy was even higher in some European countries by 1950.

Although birth rates had decreased during the latter part of the nineteenth century, women were still having relatively large families in 1900. According to the United Nations, an American woman had four or five children on average; a European woman had somewhat fewer. Fertility decline quickened after 1900. During this same period, most of Africa, Asia, and Latin America were still in the predemographic transition stage of high mortality and high fertility.


The second half of the century brought many new demographic trends and patterns. The more developed countries completed their transition to low mortality and low fertility. Population growth slowed and even turned negative in a few countries. Populations grew older. The more developed countries also experienced sometimes disruptive changes associated with baby booms and baby busts, crises in health, and waves of immigrants and refugees.

In less developed countries, the second half of the century brought decades of rapid population growth and swelling streams of migrants from rural to urban areas. Some countries appeared to be rushing through the various stages of the demographic transition, whereas others appeared to be following a new path of demographic change.

D. Mortality, Fertility, and Natural Increase

In Europe, population growth accelerated as countries recovered from the devastating effects of World War II. The rapid decline in death rates of the early part of the century slowed considerably, in part because infant and childhood mortality had already decreased to very low levels. By 1975, the IMR was 10 in Japan, 16 in the United States, and 15 in much of Europe. According to the U.S. Census Bureau, U.S. life expectancy increased by more than 10 years in the second half of the century, from 68 years to 76 years, after increasing by more than 20 years during the first half. After World War II, “baby booms” were commonplace in Europe, although they were more modest than the baby boom that occurred in the United States between 1946 and 1964. By the mid-1970s, however, total fertility rates (TFRs; which reflect the average number of children per woman) in many European countries had decreased below 2, the level at which a couple replaces itself in the population. A TFR must be slightly higher than 2.0 (about 2.1 in low-mortality countries) to reach replacement level because some women will die before the end of their childbearing years. When the TFR remains lower than 2 for a prolonged period, populations may experience a natural decrease because deaths will outnumber births.

European fertility had decreased during the 1930’s Great Depression, but in the mid-1980s TFRs sank to
record low levels and showed little sign of recovery. By the late 1990s, the TFR was 1.2 or less in Belarus, Bulgaria, the Czech Republic, Estonia, Italy, Latvia, and Spain.

The fertility decline began in Western Europe during a period that saw delayed marriage, more divorce, high inflation, and an increase in the percentage of women attending college and working outside the home. These same social and economic factors favored lower fertility in the United States, in which the TFR reached an all-time low in 1976 at 1.7 children per woman. Below-replacement fertility also occurred in Eastern Europe and the former Soviet Union after 1990.

Two decades of low fertility have halted population growth in nearly all of Europe and Japan. In many cases, a decline in population was avoided only by the flow of immigrants from abroad. In the late 1990s, 14 European countries experienced natural decrease, or fewer births than deaths each year.

Natural decrease will spread to other countries as low birth rates drastically reduce the number of people entering the childbearing ages. Although some countries have a net population gain from immigration, this is not expected to generate enough growth to stave off eventual population decline if fertility levels continue to remain low. It is possible that total fertility rates will increase again in some countries as recent trends toward later marriage and childbearing stabilize and couples who delayed childbearing complete their families. However, at the end of the twentieth century, not one major industrialized country had fertility above replacement level.

Among the more developed countries, only a few traditional immigration countries (Australia, Canada, New Zealand, and the United States) can expect significant long-term population growth. They have younger age structures and more immigration than do Europe and Japan, which contributes to momentum for continued growth.

Fertility and mortality patterns have been very different among less developed countries in the past 50 years. Gains in life expectancy accelerated after 1950. The average life expectancy at birth in less developed countries increased from 41 to 62 years between 1950 and 1995, according to United Nations (UN) estimates. The IMR decreased from 178 to 68 deaths per 1000 births during the same period.

Average life expectancy increased to more than 60 years in east Asia and Latin America by the early 1970s and to about 70 years by the late 1990s. The IMR decreased to about 29 in east Asia and 36 in Latin America by 1998.

Progress has been much slower in sub-Saharan Africa and south central Asia. In the 1950s, about 180 infants died per 1000 births in these regions. By the 1990s, the IMR was still close to 100 in sub-Saharan Africa and was nearly 80 in south central Asia.

The pace of mortality decline in some areas has been slowed by the spread of HIV/AIDS, and many experts predict dramatic declines in life expectancy in some countries of sub-Saharan Africa. Worldwide, nearly 14 million people have died from HIV/AIDS since the beginning of the pandemic in the 1980s. An additional 33 million are infected with the virus. Most will die within the next decade. The UN agency that tracks the HIV/AIDS pandemic, UNAIDS, estimates that there are nearly 16,000 new infections daily, 1600 of which are to children.

E. Population Growth

The general reduction in death rates after 1950 led to explosive population growth in many less developed countries. For the world as a whole, growth rates peaked during the 1960s and early 1970s at about 2% annually. The population total for less developed countries increased from 1.7 to 4.7 billion between 1950 and 1998. Population growth would have been even higher if fertility rates had not started to decrease in less developed countries. The pattern and pace of decline varied tremendously, depending on economic and social development, government policies, family planning use, and other factors (see Box 1).

Box 1

The Reproductive Revolution

The “reproductive revolution” was one of the most remarkable events of the second half of the twentieth century. The development of family planning methods such as the pill and the IUD, simpler sterilization techniques, and contraceptives that can be injected or implanted under the skin made it easier and safer for women to avoid unintended pregnancies. Increased access to these methods and socioeconomic changes that motivated couples to limit their family size drove the fertility declines of the past few decades. Family planning use increased from less than 10% for married women of childbearing age in the 1960s to more than 50% of this group of women in the 1990s (Fig. 2).

Before 1960, women’s choices of family planning methods were limited to such methods as withdrawal, rhythm, diaphragms, foams or jellies,
or ineffective methods such as herbal medicines or douching. Women’s options improved immensely when the pill and the modern IUD became available after 1960. In the 1990s, about 20% of women worldwide relied on one of these two methods. New contraceptives, including injectables and implants, became available in many countries in the 1980s. They have become popular methods in some African countries. Female sterilization has been widely adopted in Asia and Latin America and is the most popular method worldwide. An estimated 17% of married women ages 15–49 rely on female sterilization to prevent pregnancy.

The dramatic increase in family planning use caused fertility to decline much more rapidly in the less developed countries than it had during the fertility transition in the more developed countries. Organized family planning programs and government promotion of family planning use were important components of this phenomenon. Some experts credit family planning programs for 40–50% of the fertility decline in less developed countries since the 1960s (Ronhaar, 1995).

An estimated 120 million couples worldwide want to delay or prevent another pregnancy but are not using family planning (Miller et al., 1999). If unmarried sexually active women were included, the number would be much higher, according to survey data (Shane, 1997).

Family planning use varies widely throughout the world. Less than 10% of women use family planning in Mali, for example, and less than 20% use it in Pakistan. However, more than 60% of married women use family planning in Brazil, Mexico, Thailand, and many other less developed countries.

The expansion of family planning services has been controversial in some countries. There have also been many obstacles to their use. Many women report that they fear adverse health effects from specific methods. Others want to practice family planning but are dissuaded by their husband’s disapproval, their limited decision-making powers, or family pressures to have more children. Some methods are opposed for religious reasons. Difficulties in obtaining and transporting supplies and a shortage of trained medical personnel have also restricted access to family planning services.

Political and cultural barriers have limited access to family planning, especially for young people. In some countries, unmarried adolescents are denied access to family planning services on the assumption that such access would promote promiscuity. However, about 40% of girls in less developed countries give birth before age 20. The pace of fertility decline in Africa, south Asia, and other high-fertility regions will be affected by whether young couples delay their first birth until they are in their 20s. This delay lengthens the interval between generations and lowers average fertility. Health analysts estimate that women age 15–19 face twice the risk of dying from pregnancy and child-birth as do women in their 20s. In many countries, children born to mothers under age 20 are 1.5 times more likely to die before their first birthdays than are children born to mothers in their 20s (Shane, 1997).

A majority of less developed countries provide family planning services. In many countries, family planning methods are also widely available from pharmacies and private health clinics. Not all women have easy access to family planning, but the expansion in the choices of methods and availability of services throughout the world during the past 40 years has been truly revolutionary.

During the 1950s and 1990s, childbirth levels declined significantly in many regions. Figure 3 illustrates that declines were greatest in Asia and Latin America and much smaller in sub-Saharan Africa. Despite these declines, all developing regions had TFRs above replacement level in 1998 and more than 100 countries had TFRs of 3.0 or higher.
F. Changing Age Profiles

Fertility, mortality, and migration trends are reflected in the age and sex profiles of the world’s countries. The decades of high fertility rates in the less developed countries meant ever-increasing numbers of young people, illustrated by the broad base of the age–sex pyramid shown in Fig. 4. Improvements in infant mortality have also contributed to the expanding youth population. Children under age 15 comprised one-third of the population in 1998 and even greater proportions in some regions. In sub-Saharan Africa, children comprised nearly one-half (45%) of the population. Elderly people ages 65 or older comprised only 3% of the population in all less developed countries and 3% of the population in sub-Saharan Africa.

The base of the population pyramid for less developed countries shows some narrowing—the result of declining fertility in many countries beginning in the 1980s. However, even with declining fertility rates, the young age structure creates considerable momentum for future growth because the population reaching childbearing ages continues to expand. Women currently have fewer children than women did in the past, but today there are more women having children.

II. CAUSES AND EFFECTS OF HUMAN POPULATION GROWTH

The demographic processes of fertility and mortality are influenced by biological, cultural, economic, geographic, political, and social factors. These factors affect demographic processes directly and indirectly through a web of interdependent variables. Cultural traditions that encourage girls to marry at a young age, for example, can contribute to high fertility rates because women will spend more years exposed to the risk of becoming pregnant. Early marriage can also lead to higher mortality because health risks to the infant and mother are greater when childbearing starts in adolescence.

In the 1980s, demographer John Bongaarts identified four variables that account for most differences in fertility rates. These “proximate determinants” of fertility are (i) the proportion of women married or in a sexual union; (ii) the percentage of women using contraception; (iii) the proportion of women using contraceptives that are effective; and (iv) the likelihood that a pregnancy will result in a live birth.
tion; (iii) the proportion of women who cannot conceive a pregnancy, especially during the infertile period following childbirth (postpartum infertility); and (iv) the level of abortion.

The importance of each proximate determinant depends on cultural, economic, health, and social factors within a population. The proportion of women in a sexual union is partly determined, for example, by the age at marriage, the proportion of women who never marry, and levels of divorce. Cultural mores about sexual activity and childbearing outside marriage also play a role.

In societies in which women marry young, and in which nearly all childbearing takes place within marriage, changes in the age at marriage can significantly affect fertility. In the Arab countries of the Middle East, for example, an increase in the average marriage age for women led to significant fertility declines in some countries (Rashad and Khadr, 1999).

The length of postpartum infertility usually depends on how long women breast-feed their babies. Breast-feeding releases hormones in the nursing mother that can prevent her from becoming pregnant. Postpartum infertility is not a significant factor in such countries as the United States, in which women usually breast-feed their babies only for a few months, but it is important in sub-Saharan Africa and other traditional societies in which women commonly breast-feed their babies for 2 years. In most populations, contraceptive use and abortion are the primary determinants of fertility levels.

Education and poverty are among the most important influences on the proximate determinants and consequently have a strong indirect effect on fertility. Low levels of education and poverty are related, and they are also related to health and to levels of economic development, urbanization, and environmental conditions.

A. Education

Although researchers cannot determine all the reasons why, education is associated with lower fertility and mortality. A formal education may act as a catalyst for changes in values and behavior. Education may make people more receptive to new ideas, such as family planning. Social scientists note that education does not have the same effect in all cultural settings, and that many other factors (such as women's status) may explain much of the association (Riley, 1997; Jeffrey and Basu, 1996; Knodel and Jones, 1996).

More educated women have higher rates of family planning use, smaller families, and healthier children than other women in the same society. Where educational levels are high, women are likely to postpone marriage until they finish secondary school or college.

Married women are more likely to use family planning if they have some formal education. A 1998 survey in the Philippines indicated a contraceptive use rate of 50% for married women of reproductive age who had at least some secondary education. Only 1% of their counterparts with no formal education used a contraceptive method.

In most societies, total family size declines as education increases. In the early 1990s, Peruvian women with at least some secondary education had nearly four fewer children, on average, than women with no formal education. A similar gap was recorded in a 1998 survey in Togo, West Africa. Togolese women with a secondary or higher education had 2.7 children on average, whereas women with no education had an average of 6.5 children.

Education usually expands employment options, and educated women may delay marriage and childbearing to earn income. Also, school may introduce young women to new ideas or values that could influence the number of children they want and also their use of family planning.

Women's education is also associated with better child health. Children of mothers with some education have fewer risk factors for infant mortality. Infants are at a higher risk of dying if they are born to adolescents or to mothers older than age 40, if they are born into large families, or if they are born less than 2 years after an older sibling.

By delaying marriage and childbearing, education reduces high-risk births to teenage mothers. In Indonesia, for example, 32% of women ages 20–24 who completed less than 7 years of education had a baby by age 20, whereas only 13% of those with 7 or more years of education had a baby by age 20. The gap was less pronounced in Kenya, as shown in Fig. 5, but greater in Peru and Egypt in the early 1990s. Women who have completed some formal education tend to wait longer between pregnancies and births and to stop childbearing at a younger age than do less educated women. Consequently, they have smaller families and have fewer births after age 40.

In most societies, children of mothers with some education have a lower risk of dying than children whose mothers had no education. As shown in Fig. 6, in Zambia, the IMR was 133 for the children of mothers with no education, whereas it was 82 for children of women with a secondary or higher education. The
twentieth century brought enormous improvements in literacy and educational levels. The recent improvements in literacy rates reflect the expansion of educational services throughout the world. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) reported that 77% of people over age 15 were literate in 1995 compared with only 56% in the 1950s. Basic literacy is nearly universal among populations of Europe, North America, and other industrialized regions, but the range is substantial throughout the rest of the world. In 1995, an estimated 50% of the populations of south Asia were literate, as were 57% of the populations in sub-Saharan Africa and the Middle Eastern Arab states. More than 83% of the populations are literate in east Asia and Latin America and the Caribbean.

Rapid population growth in some countries is undermining improvements in educational attainment. In the sub-Saharan African countries of Angola, Benin, and Togo, for example, economic difficulties and burgeoning numbers of young people caused school enrollment ratios to level or decrease in the 1980s and 1990s (Bledsoe and Cohen, 1993). In the mid-1990s, about 67% of girls and 81% of boys in sub-Saharan Africa were enrolled in primary school, according to UNESCO estimates.

B. Economic Development and Environment

In most societies, poor families have higher mortality and fertility than affluent families. Some of the association between poverty and population reflects the lower educational levels and rural residence of poor households. However, the relationship among demographic variables, poverty, and affluence is highly complex, and it is tied to the broader question of how population size and the pace of population growth are linked to economic development. The issue is further complicated by questions about whether economic growth and human activity are causing irreversible damage to the natural environment.

The research into these questions has yielded contradictory results. The extremes of these differences are characterized by two opposing camps: “pessimists” and “cornucopians” (Marquette and Bilsborrow, 1997). The theoretical foundation of the pessimist view can be found in writings published in 1798 by the economist Thomas Malthus. He suggested that the potential population size is limited by the amount of cropland, and therefore food, available for human consumption. Malthus assumed (based on his observations of eighteenth-century English society) that if population growth continued unchecked, population would outstrip the food available and cause widespread famine and death. He also described a natural feedback mechanism: When the population grew too large for the available food supply, elevated mortality would reduce the population to the level that could be sustained by the amount of food produced.

A neo-Malthusian view of the relationship between population, economic growth, and resources gained
credence between the 1940s and the 1960s, a period of unprecedented population growth and economic development. In a landmark study in the 1950s, Ansley Coale and Edgar Hoover found that population growth slowed economic development and held down per capita incomes. These researchers also assumed that the supply of some natural resources and capital was fixed, or that supply would grow more slowly than population. Other researchers during this period expanded the idea that rapid population growth will eventually reach some absolute limit on resources and that population growth accompanied by the environmental stresses associated with economic development could cause irreversible damage to the basic natural systems that sustain life. Such concerns were popularized by books such as The Population Bomb by Paul Ehrlich, The Limits to Growth by Dennis Meadows, and also by other authors.

The ideological basis of the “cornucopian” approach owes much to the writings of Ester Boserup in the 1960s and 1970s, who argued that the need for more food coupled with the synergy created by the concentration of intellects and flow of ideas in dense settlements can stimulate the adoption of better farming techniques or sharing of higher yield plant varieties. Economist Julian Simon, in The Ultimate Resource and other writings, also rejected the idea that population growth was a threat to the welfare of humans or the environment. The scientific evidence concerning the effects of population size and growth on economic development was still inconclusive in the early 1990s at the national and regional level, but it was less ambiguous at the individual level. An extensive review of research completed in 1994 by the Overseas Development Council observed that “the clearest evidence of negative effects of population growth under high fertility are at the individual and household levels.”

In the late 1990s, several studies provided a clearer picture of the relationship between population and development at the national level and the links between poverty and demographic factors at the household level. Several studies suggested that a rapid transition from high to low fertility contributed to the economic miracles in South Korea and other east Asia countries (Mason, 1997). The rapid fertility decline increased the share of working-age people in the population relative to young and old dependents, which created a “demographic bonus.” The working-age population adds more to the economy than it consumes in services and generates taxes and savings that can be invested in education and further economic growth. This demographic bonus may last several decades; it recedes as the bulge of working-age men and women reach retirement age and the proportion of the dependent population relative to the working-age group population increases again.

Research shows that countries can benefit from this bonus only if they increase the value of their human capital—especially the youth entering the labor force—through education, and if governments adopt policies favoring international trade and industrialization. The newly industrializing Asian countries capitalized on their demographic bonus by making these investments. As researcher Andrew Mason noted, these Asian countries “raised millions of people from abject poverty and transformed some of the poorest economies in the world to some of the richest.” Research on the relationship between population change, economic development, and environmental systems remains much less clear, plagued by the complexity of the relationships and the difficulty of measuring such factors as environmental quality. At a symposium on population change and economic development in 1998, scientists assigned highest priority to research on population–development–environment linkages.

III. POPULATION PROSPECTS: 2000–2050

In the past century, the world’s population has undergone a sweeping change in both its total numbers and its distribution across regions. The twenty-first century is likely to see the second phase of this transformation—lower fertility and an even more dramatic redistribution of population among the more developed and less developed countries. Nearly all future world population growth will take place in less developed countries. In other words, the earth is reinventing itself demographically.

Because mortality is relatively low, fertility levels and trends will determine future population size. A common issue (and a common assumption) is when, or whether, a country will reach the “magic” replacement-level TFR of about 2.1 children per woman. With fertility at replacement level, a population eventually will cease growing and “stabilize” at a given size. Every 2 years, the UN Population Division produces a set of population projections for every country. By 2050, the UN suggests that total world population will increase to between 7.3 and 10.7 billion. In the high projection, world population will still be increasing in 2050; under the low projection series, it will have begun
a gradual decline. All three projections assume continuing declines in fertility. They differ in terms of fertility levels attained in 2050, from 1.57 for the low projection to 2.5 for the high, illustrating the major differences in total population size that can result from seemingly very small differences in childbearing during the next 50 years.

Regardless of the projection used, the UN indicates that at least 1.3 billion people will be added to the world’s population during the next 25 years. There are three reasons for this inevitable growth. First, fertility in less developed countries is on average twice as high as in more developed countries. Second, the young age structure of less developed countries constitutes momentum for population growth for several decades, no matter what future fertility trends may be. Third, continuing improvements in mortality will contribute to additional growth, particularly in countries in which life expectancy remains comparatively low.

IV. RESPONSES TO HUMAN POPULATION GROWTH

Anxiety about the negative effects of rapid population growth and excessive population numbers has a long history. Long before Malthus, ancient Greeks and Egyptians voiced concern about “overpopulation” in lean times. They also promoted population growth in times of plenty.

In the 1930s and 1940s, scientists and intellectuals in some less developed countries, such as Egypt, India, and Mexico, began to express concern that rapid population growth would hinder development in their countries. Widely publicized food shortages and famines in certain less developed areas in the 1960s were also linked to rapid population growth.

These concerns sparked many actions throughout the world directed at lowering fertility and slowing population growth. India initiated a national policy to slow population growth in 1952. The International Planned Parenthood Federation, the largest private sector organization devoted to family planning, was founded in the same year. United Nations involvement in population issues also increased. The first UN meeting on global population was convened in 1954, in collaboration with the International Union for the Scientific Study of Population. United Nations agencies, including UNICEF and the World Health Organization, incorporated reproductive health services into their missions. In 1969, the UN Fund for Population Activities became a separate entity. Beginning in the 1960s, governments of some wealthier countries, most notably the United States, supported efforts to strengthen family planning programs in less developed countries.

A. Population Policies

The idea that couples should limit their family size went against cultural mores in many societies, and some governments were loathe to support a potentially unpopular policy. Many governments embraced the more acceptable idea that fertility would decrease and that population growth would slow as living standards increased through economic development. This view was expressed at the 1974 UN World Population Conference when an Indian delegate declared that “development is the best contraceptive.”

During the late 1970s and 1980s, concern about the negative effects of population growth on economic development broadened. Increasing numbers of countries accepted the idea that government actions could slow population growth.

National efforts to influence population growth include incentives to have more or fewer children and disincentives for having more than a given number of children. These efforts have met with mixed success. Some argue that China’s population policies initiated in the 1970s were a success from a demographic perspective. China’s TFR decreased from about 6.0 in the 1960s to less than 2.0 in the 1990s, in part because of government policies and programs. However, China’s stringent “one-child family” policy introduced in 1979 was widely criticized for violating human rights. Between 1973 and 1977, Indira Gandhi’s government in India promoted male sterilization campaigns that sometimes led to coercion. Public outrage about the reported abuses contributed to the downfall of Gandhi’s government and created a backlash against family planning programs in India that took years to overcome.

Women’s rights activists and others have generally opposed a demographic rationale for family planning as an infringement on individual rights. They have argued that women’s rights and well-being should take precedence over national interests. Many have criticized the family planning programs’ lack of integration with other health services.

During the 1970s and 1980s, women throughout the world began forming small nongovernmental organizations (NGOs) to lobby for improvements in their social, economic, and political circumstances. By the 1990s,
women's NGOs in less developed countries were advocating for improvements in family planning programs by better informing clients about various contraceptive methods, expanding the range of methods available, and encouraging service providers to treat clients with greater respect.

As of 1997, 135 countries subsidized family planning services, and 68 stated explicitly that they wanted to slow their population growth. In Africa, the world's fastest growing region, 40 countries viewed their fertility levels as too high and 36 had policies to lower fertility. A few countries, in contrast, viewed their fertility rates as too low and welcomed faster population growth. In 1997, 23 countries reported to the UN that they had explicit policies to increase birth rates. Many governments in Europe and the former Soviet Union are concerned that continued low levels of fertility will cause rapid population aging and an eventual decline in population size. Some small oil-rich countries in the Persian Gulf also want to increase, or at least maintain, current levels of population growth. They view population growth as a way to spur socioeconomic development and reduce their reliance on foreign labor.

B. A New Vision: The 1994 International Conference on Population and Development

In 1994, the world community redefined the world's view of population growth and the best way to address this growth. The Program of Action adopted at the International Conference on Population and Development (ICPD) in Cairo, Egypt, agreed to the need to stabilize global population growth. One hundred and eighty countries agreed to a strategy placing population within the context of sustainable development and calling for investments in human development, especially improvements in women's status, as key to stabilizing population growth. It rejected the use of demographic targets by family planning programs and it integrated family planning into a broader women's health agenda.

Participation in the meeting by NGOs was a critical factor in achieving such widespread consensus. Twelve hundred NGOs participated as delegates and observers, and for the first time conference deliberations on global population were informed by a wide range of interests, from the grassroots to the highest levels of government. Women's groups were a driving force behind the strong emphasis on women's empowerment as part of human development. This focus was also driven, however, by research from the past 30 years that linked fertility declines with reductions in infant mortality, increased use of family planning, and improvements in women's education and other aspects of women's status contributing to their empowerment.

Despite the consensus, the ICPD engendered dissent and debate. Ideological and religious tensions characterized discussions before the conference, deliberations during the conference, and follow-up after the conference. Abortion generated the most highly publicized ideological differences. There was also debate regarding definitions of reproductive health and family and adolescent reproductive rights and responsibilities. None of the 179 nations rejected the central premises and goals of the ICPD, despite the range of political structures, cultures, and religions they represented. This marked the first time in the history of UN conferences that no official delegation rejected the entire document.

The final ICPD document defined reproductive health to encompass a broad range of services, including family planning, prenatal and postnatal care, medical attention at birth, cancer screening, and protection from sexually transmitted diseases. It also supported access to safe abortion where it is legal, but it stated that abortion should not be used as a method of family planning.

The ICPD Program of Action specified five goals for 2015 to improve individual and family well-being and enhance women's status, including universal access to family planning and other reproductive health services; universal access to primary school education; increased access by girls and women to secondary and higher education; and reductions in infant, child, and maternal mortality. The ICPD document also called for government and private sector actions to alleviate poverty, protect the environment, encourage greater male involvement in the family, and address the specific health and education needs of adolescents.

A unique feature of the agreement was that it called for specific financial commitments to achieve these goals. Countries agreed that less developed countries would pay for two-thirds of the costs of reproductive health programs, including family planning, and that more developed countries would pay for the remaining costs, estimated to be $17 billion per year by the year 2000. The historic agreements reached at the ICPD were reaffirmed at subsequent UN conferences in the 1990s, including the World Summit for Social Development, in Copenhagen in 1995; the Fourth World Con-
V. CONCLUSION

Population change has been one of the most significant events of the twentieth century. Since 1900, the world population has more than tripled in size and average life expectancy has increased by two-thirds. Declines in childbearing and shifts in population distribution were more striking than at any other time in history. Along with these population changes, the world has witnessed extraordinary improvements in technology, communication, education, and agriculture. These changes have undermined the dire predictions of Thomas Malthus and his successors that population growth would lead to worldwide famine and disease. However, these predictions may come true for some areas of the world. According to the UN, more than one-fifth of today’s population lives in poverty, surviving on less than $1 a day. The HIV/AIDS pandemic threatens the health and well-being—and the very survival—of large portions of the population in many countries.

Under all likely scenarios, the twenty-first century will see continued population increases, at least during the first few decades. This is because of the built-in momentum of growth associated with the very young age structures of most less developed countries. The growth will also be fueled by childbearing levels that are still higher than replacement levels in most areas of the world. Not all countries will experience this growth, but they will all be affected by it.

At the end of the twentieth century, the world community made financial and program commitments to continue investments in family planning and other health programs, in education, and in greater social and economic opportunities, especially for women, arguing that these are key to future population stabilization.

See Also the Following Articles

ECONOMIC GROWTH AND THE ENVIRONMENT • ENERGY USE (HUMAN) • HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW • SUSTAINABILITY, CONCEPT AND PRACTICE OF

Bibliography


I. Genetic Variation and Its Significance
II. Levels of Interest to Conservation Geneticists
III. Evolutionary Processes of Interest to Conservation Geneticists
IV. Genetic Management: Examples
V. Genetics, Evolvability, and the Future of Biodiversity

GLOSSARY

evolution  Biological and physical processes leading to heritable change in characteristics of populations or species over time. Key agents of evolutionary change are mutations, genetic drift, gene flow, and natural selection. The ability of a species to adapt to environmental changes is controlled genetically and is here termed evolvability.

extinction  The final disappearance of a species unable to evolve, to adapt to changes in its local physical or biological environment, or to shift its geographic range to avoid such changes. Extirpation is the local extinction of a population on a habitat patch that may subsequently be recolonized by dispersal from other populations of the same species. Risk of extinction is codified, and vulnerable, endangered, and critical imply a significant probability of extinction within years or decades rather than centuries; threatened is used here as a general term.

fitness  Life time reproductive success of an individual with a particular genotype relative to another individual with a different genotype. Natural selection typically favors “survival of the fittest” and thus facilitates the future evolvability of a population.

genetic enhancement  Management actions taken to increase the genetic variability and viability of a population, includes translocations and reintroductions.

genetic erosion  A population viability-threatening process in small, isolated populations whereby random genetic drift and inbreeding diminish the population’s innate genetic variability. Increasingly, it is associated with habitat fragmentation.

N  Symbol for the genetic effective size of a population, an important indicator of future evolvability or extinction risk; usually much less than the observed census size (N0).

variation  The abundant normal genetic differences between individuals in a population. Such variation is ultimately due to changes at the DNA base pair sequence level but is typically monitored at higher levels of expression. Overall genetic variation is associated with population viability and future evolvability. Conservation genetics is focused mainly on the protection and maintenance of genetic variation.

CONSERVATION GENETICS is concerned with population genetic variation, population viability, and the future evolution of species. Conservation genetics, ecol-
I. GENETIC VARIATION AND ITS SIGNIFICANCE

Until the 1960s, it was widely thought that genetic variation was unusual. Biologists regarded populations as being composed of many very similar “wild-type” individuals and a few rare mutant individuals. Following the introduction of allozyme electrophoresis it quickly became clear that most plant and animal populations were highly variable. Subsequently, it has become possible to examine DNA sequences directly and large numbers of base pair mutations have been discovered in functional genes of most living organisms studied. Even higher levels of variation occur in the introns and highly repetitive stretches of noncoding “junk DNA” that lie between eukaryote genes. High levels of variability are usually within populations, whereas a dinucleotide repeat microsatellite may have 10–15 alleles (differing in the number of repetitions) in the same individuals. A polymorphic allozyme may have 2 or 3 alleles segregating in a population, whereas a dinucleotide repeat microsatellite may have 10–15 alleles (differing in the number of times the motif is repeated) in the same individuals. Biochemical and molecular genetics have shown that most living organisms are richly variable. Estimates of variability in a population vary with the feature examined and analytical method used. Karyotypic variation is usually low within a species, allozyme variation is high, and DNA sequence variation may be very high in some parts of the genome. The principle of high natural variability is perhaps best illustrated by the published surveys of allozyme patterns in plants and animals (Schonewald-Cox et al., 1983). Geneticists typically sample a population of 20 individuals and determine each individual's genotype at 10–20 loci. This allows them to calculate the mean number of alleles per locus (A), the mean individual heterozygosity (H), and the proportion of loci in the population that are polymorphic (P). A few thousand species have been examined to date and variation is typically in the following range: $H = 0.03–0.15$ and $P = 0.20–0.50$. Still higher levels of natural variation are found at the DNA sequence level, especially in areas of the genome which are apparently free to mutate. Short repeat sequences (microsatellites) of nuclear DNA, for example, are often 10 times more variable than allozyme loci. A polymorphic allozyme may have 2 or 3 alleles segregating in a population, whereas a dinucleotide repeat microsatellite may have 10–15 alleles (differing in the number of times the motif is repeated) in the same individuals. Biochemical and molecular genetics have shown that most living organisms are richly variable.

Much of the cryptic variation in natural populations that has been discovered in the past 30 years appears to be selectively neutral or near neutral in its effect on the phenotype. That is, the individuals carrying these various allelic variants appear perfectly healthy. It is rare to find a single locus genetic trait in which a deleterious condition such as albinism is controlled by a single allelic variant. In fact, there is strong circumstantial evidence that much of this genetic variation is actually beneficial. Although the relationship between genetic variability and individual fitness is not well understood, it is clear that variability is associated with viability. Experiments and field observations with a few species have shown that there is a positive relationship between genetic variability and individual growth rate, size at maturity, symmetry of body parts, fecundity, and health as measured by parasite load. Extrapolating from the individual level to the population level, it is a fundamental maxim of evolutionary biology that genetic variation is positively related to adaptability or evolvability. Fisher's fundamental theorem holds that additive genetic variation in fitness is positively related to a population's ability to respond to natural selection. Evolutionary success, the ability of a species to persist despite changes in climate and exposure to new diseases, predators,
and competitors, is somehow related to innate genetic variability. In a world of unpredictable change, alleles that are selectively neutral for thousands of generations may suddenly become lifesavers for the individuals that carry them. If evolution is largely dependent on genetic variability, then the conservation of species will sooner or later depend on the conservation of their innate genetic variability (Woodruff, 1989).

Of course, there are many apparently successful genetically invariant plants and animals. There are many known clonal organisms whose descendents are genetically identical to their mothers. Such clonal plants, snails, fish, and lizards are genetically invariant even though they may exhibit great ecological success. However, they are more likely to go extinct when their environment changes than are closely related sexually reproducing species. Although meiosis ensures the maintenance of high variability in the sexual species, their asexually reproducing daughter species gradually lose their initial low variability and over time become evolutionary dead ends.

There are many different ways of measuring genetic variability, including the examination of karyotypes and single-locus markers (allozymes, mitochondrial and chloroplast DNA restriction fragment length polymorphisms), DNA minisatellite fingerprints, random amplified polymorphic DNA, mtDNA sequences, nuclear DNA sequences, and nDNA microsatellites (simple sequence repeats). Benirschke and Kumamoto (1991) and Smith and Wayne (1996) provide numerous examples of the application of karyotypes and molecular genetic markers, respectively, to conservation. The various methods differ in their resolution of pedigree, population, and species-level questions and there is no single correct technique. One method, allozyme electrophoresis, is informative with mammals but frustratingly uninformative for birds. The methods also differ greatly in their cost and technical difficulty. Some relatively inexpensive methods are satisfactory for one-time analyses but the results cannot be built upon in subsequent studies. Other methods, in contrast, provide genotypes which can be archived in expandable permanent databases such as GenBank for future comparative study of samples collected at other times or places. Although a given method will give comparable results in a study of closely related populations or species, results across unrelated taxa may not be comparable.

Currently, nuclear and mitochondrial sequence data are the most informative methods for characterizing variability at or above the level of populations. For studies of variation within populations, hypervariable nuclear microsatellite loci are ideal markers. Sequence data and microsatellite genotypes can now be determined directly from minute DNA samples amplified many times by the polymerase chain reaction. Both methods require fully equipped laboratories, trained personnel, and considerable time and money for developing the synthetic DNA primers to amplify the gene sequences of interest in species that have never been studied before.

Noninvasive genotyping methods involving the extraction of DNA from shed hair and feathers were introduced in 1989 and are now widely used (Morin and Woodruff, 1996). Noninvasive (shed tissues, feces, urine, and scent markings) and nondestructive (toe, tail, and ear clips and fish scales) genotyping permits the study of wildlife populations that previously were almost impossible to sample. Not surprisingly, the DNA extracted from some types of samples may be degraded and very difficult to work with. Nevertheless, with technical care and patience it is possible to genotype some animals without actually seeing or handling them.

Conservation geneticists can also genotype museum specimens and determine patterns of variability over periods of decades and sometimes centuries. Several species that have gone extinct recently, including the dodo, moa, thylacine, and quagga, have been characterized genetically. DNA can also be extracted and sequenced from some fossil remains, but DNA degradation rates are such that fossils more than a few million years old are unlikely to yield reliable sequences. Although enormously interesting to evolutionary biologists, such ancient DNA cannot be used to recreate extinct organisms. Conservation geneticists must concentrate on saving existing biodiversity; they cannot fall back on the idea of being able to recreate extinct species from the tiny fragments of DNA (typically less than 0.000001% of an organism’s genome) currently under scrutiny.

Genetic data, once acquired, are used by conservation geneticists to quantify within- and between-population variability. Variability within populations is used to establish pedigree relations, mating system, sex ratio, and genetic effective population size (N). Interpopulation comparisons reveal spatial structuring and historical patterns of gene flow. Geographic variation is normal within a species and the new field of phylogeography uses genetic information to infer the historical relationships among populations. Single and multilocus genetic differences between kin, populations, subspecies, and species are expressed as genetic distances. These metrics and their interpretations are beyond the scope of this review, but it is important to note that the absolute values of genetic distances will vary between different
groups of plants and animals and increase over geological time. Within a group of related species a major difference in the genetic distances observed within and between taxa can be used to define species and other evolutionarily significant units for conservation management purposes.

Although the previous discussion focused on molecular genetic variation, the growing field of quantitative trait genetics promises to provide a new means of measuring evolutionarily significant population variables (Storfer, 1996). Quantitative genetics is concerned with characters such as morphology, behavior, parasite resistance, and physiology that are controlled by several to many genes that work additively, in dominance/recessive relationships, or epistatically. Such oligogenic and polygenic control, involving quantitative trait loci (QTL), is characteristic of many traits of interest to conservationists—those that affect long-term population persistence and evolvability. QTLs affecting body size, hatching date, and predator avoidance behavior (escape speed), for example, are ecologically important and are arguably of greater significance to conservationists than allozyme polymorphism. Conservationists are therefore interested in the heritability of such traits that have a direct impact on fitness. High heritabilities of a QTL indicate that a population has great potential for trait evolution and low heritabilities indicate a more limited ability to respond to environmental change. Unfortunately, such heritability is difficult to measure because it requires pedigree-level studies conducted over several generations or long-term manipulative experiments such as controlled garden plots. Heritability (h²) is the ratio of the variance of the genetically inherited proportion of a trait (the additive genetic variance, Vₐ) to the total phenotypic variance V_p measured in the population. Vₐ is one component of genetic variance (V_G) which also includes nonadditive genetic variance due to dominance (V_D) and epistasis (V_E); only Vₐ responds to directional selection. Estimating Vₐ is further complicated by the need to estimate variance due to the environment (V_e) as well as the other genetic components. Nevertheless, the preparation of studbooks for captive populations of endangered animals and the comparison of laboratory-raised seedlings of rare plants to their parents in the wild have provided opportunities for the first limited applications of QTL analyses to conservation. Future technical advances may permit the inclusion of QTL in the conservation geneticist’s tool kit of predictors of a population’s risk of extinction.

A major issue requiring resolution is the validity of the widely held relationship between molecular genetic variation and ecological viability and evolutionary potential. Are single gene markers (those typically surveyed by geneticists) useful indicators of variation at quantitative traits? Are multilocus allozyme surveys unreliable as predictors of a population’s viability since heterozygosity may be only weakly correlated with the additive genetic variation associated with QTLs? In desert upminnows of the genus Poccilopus, Vrijenhoek (in Avise and Hamrick, 1996, pp. 367–397) found that rapid losses of heterozygosity in small, isolated populations were associated with a decline in fitness that was manifested as poor competitive ability, growth rate, developmental stability, and resistance to parasites. O’Brien (1994), in a remarkable series of genetic studies of captive and wild cheetahs (Acynonyx jubatus), demonstrated a powerful association between low levels of genetic variability and susceptibility to viral diseases. There is no “normal” level of variation for a population—even as determined with a specific method. Cheetah, northern elephant seal (Mirounga angustirostris), and European badger (Meles meles) are ecologically successful despite low absolute levels of genetic variation. Different types of genetic variability will respond differently to natural selection, inbreeding, population collapse, and range fragmentation. Conservation geneticists can identify cases in which variation has or is being lost, establish the causes of the loss, and make recommendations to counter its ultimate effects. Because currently available genetic markers are only proxies for fitness determinants, this underscores the importance of using different markers. It is not the purpose of this review to discuss the technical problems associated with each method or the applicability of the different methods to different groups of plants and animals, but it is important to emphasize that each method has unresolved technical problems (e.g., null alleles, allelic dropout, pseudogenes, and nonreplicable patterns). Until molecular and QTL genotyping become routine, conservation geneticists must guard against the natural tendency to overstate the statistical power of their results.

II. LEVELS OF INTEREST TO CONSERVATION GENETICISTS

A. Genes

Long before the term conservation genetics was coined, the phrase “genetic conservation” was introduced to describe the science of managing specific genes or phenotypic traits in crop plants, land races and cultivars, bacteria and fungi used in food production, and domes-
ticated animals. Genetically modified organisms (GMOs) are simply special cases that require even more intensive management for their perpetuation. The methods of gene discovery and artificial selection developed for managing microorganisms, plants, and animals are relevant to the far more broadly focused field of conservation genetics, but very few wild species have received such intensive effort.

Discussions of the need to save something or other from extinction or the inevitable loss of variation at other potentially important loci. Some alleles are common and others are rare. Deleterious alleles (e.g., alleles responsible for albinism or other genetic "defects") are typically very rare and have frequencies of less than 0.0001. Conservation geneticists are often asked to devise breeding plans that will further reduce or even eliminate such alleles from a population. On the other hand, it has been argued that conservation geneticists should strive to save rare alleles in threatened populations because they may prove vital for a population's adaptation to future environmental changes. Although this argument is reasonable, the maintenance of desirable rare alleles, even if they were identifiable, requires very large population sizes (N_e > 5000) and is simply not possible in most management programs. Rare alleles contribute very little to variation in fitness among individuals and are less likely than alleles at relatively high frequency to be the basis of adaptive response to environmental change.

It has been suggested that conservationists should focus on saving diversity at major histocompatibility complex (MHC) genes because they play a role in recognition of infectious agents, disease susceptibility, and defense. This recommendation was well intended but rejected because the functional (fitness-related) significance of the large number of alleles at each of the many MHC genes is unknown. Managing them as a single linkage group would require very large populations or the inevitable loss of variation at other potentially important loci.

B. Populations

Interest in the concept of minimum viable populations (Gillpin and Soulé, 1986) spawned the development of new methods of pedigree analysis and population viability analysis. Populations have both a census size (N) and a genetic effective population size (N_e); the latter is one of the most important concepts in theoretical conservation biology and is defined in Section III.G and by Lande and Barrowclough (1987). The characterization of genetic variation within and among populations enables geneticists to help set conservation priorities. Comparative levels of variation and gene flow (or lack of either) provide clues to a population's viability and extinction proneness. Data on genetic relationships among populations guide translocation decisions and identify well-defined clusters of populations for management as separate entities.

Pedigree analysis refers to a suite of genetic models for understanding processes in small populations. First developed for assessing management practices in captive populations, pedigree analysis is also applicable to wild populations of individually monitored organisms. It is used to establish kinship and individual founder contributions, to identify genetically desirable and undesirable individuals and their descendents, to minimize inbreeding, to describe population structure and mating system, and to choose individuals for reintroduction or translocation. Examples of pedigree analysis with the Gene Drop computer program include Haig's (1998) study of the red-cockaded woodpecker, Picoides borealis. Pedigree management programs based on mean kinship or equalizing founder contributions seek to minimize inbreeding in local subpopulations and in the metapopulation as a whole. Captive breeding programs have been successful in slowing the loss of genetic variation and preventing inbreeding depression.

Population viability analysis (PVA) is the methodology used to assess the ecological and genetic risks facing a wild or captive population and to develop a conservation management plan. PVA refers to a suite of mathematical models that seek to predict the probability of a population's extinction by some time in the future, e.g., 20 or 100 years. Early models considered demography (growth rate, present population size, and birth rate) and environmental stochasticity, but Gilpin and Soulé (1986) broadened the definition to include genetic factors. Genetic factors, including genetic drift and fixation of deleterious mutations, are expressed through demographic factors that affect population dynamics. The genetic factors thus contribute to extinction probabilities through a very complex and little understood series of interactions affecting fitness. Geneticists can use population variation to provide estimates of the various parameters of interest to modelers and managers. Unfortunately, genetic models with links to ecological factors are still insufficiently developed to yield the type of statistically powerful PVA managers seek. Most PVAs, in fact, have not included genetic parameters, but this is changing as the significance of genetics in the long-term survival of small populations becomes more widely appreciated (Beissinger and McCullough, 2000).
Metapopulations are populations of subpopulations within some defined area, in which dispersal from one local population (subpopulation) to at least some other habitat patches is possible. There is significant turnover of local populations, local extinction, and recolonization by dispersal. The metapopulation concept is central to much ecology and conservation theory and single- and multiple-species metapopulation dynamics are reviewed by Hanski and Gilpin (1998). The genetic effective size of a metapopulation is affected by the carrying capacity of the habitat patches, the rates of extirpation and recolonization, the number and source of the founders, the number of local populations, and the rate of gene flow between patches. It is difficult to establish metapopulation effective size using genetic data because it is strongly affected by extirpation and recolonization dynamics. As in the case of genetic effective population size, the metapopulation effective size is 10–100 times less than the census size in many systems. Metapopulation dynamics, with frequent local extinction and recolonization of habitat patches by few founders, can reduce deep phylogenetic subdivisions typically within a species or occasionally, as in the case of local endemics, the entire species, and MRIs reflect shallower subdivisions within a species which for practical reasons become the focus of management activities. Translocations between ESUs are typically undesirable.

C. Subspecies

Subspecies or local geographic races have been the focus of intense conservation efforts. Genetics is useful in establishing whether such groups of populations are sufficiently different to warrant separate conservation efforts. The conservation of every local race or subspecies is difficult to justify if they are genetically almost identical. Although we may recommend trying to preserve every variety of wild tomato or maize in a seed bank, it is difficult to use the same justification to try to conserve all local variants of geographically widespread organisms.

Many biologists have abandoned the subspecies concept and the associated trinomial nomenclature. To them, the subspecies is an evolutionarily insignificant artificial taxon typically based on a few superficial morphological features. There are numerous cases in which genetic studies provide no support for the traditional subspecific taxonomy. Nevertheless, there is much to be learned from the geographic patterns of variation in nature. Subspecies were typically defined as geographically isolated races with allopatric or parapatric distributions. The observation of significant hybridization without introgression, in the latter cases, led to the development of the semispecies concept which is of relevance to conservation. Semispecies are typically parapatrically distributed and hybridize where their ranges meet, but they show very limited introgression. They are effectively isolated groups of populations evolving independently of one another. Collectively, a group of semispecies comprise a superspecies and each semispecies is treated taxonomically as a full species. In this case, geographically defined taxa that used to be treated as subspecies are actually independent evolutionary lineages and therefore as worthy of conservation as other "good" species.

Conservation geneticists are often asked to advise on proposals to pool individuals from allopatric populations on the argument that it is better to save a genetic species than no species. Although poorly differentiated subspecies may often be mixed without genetic harm, the pooling of individuals of well-differentiated semispecies or species is likely to have negative genetic consequences. Unfortunately, management decisions are of-
ten made before the appropriate genetic data are available.

Geneticists are often involved in two other situations involving subspecies and local variants. The first involves populations on either side of national or state boundaries that may be assigned to different subspecies and can receive radically different levels of protection; a species-wide conservation plan may allocate resources differently and be biologically preferable. The second situation occurs when peripheral populations of a widespread species become the focus of conservation activities. Although such peripheral populations may be at high risk of local extinction, their conservation may not be warranted, especially in cases in which reintroduction is practical. On the other hand, some peripheral populations may be critical to a species’ long-term survival. Some peripheral populations may be better adapted to changing climatic conditions than central populations even though the latter may be more genetically variable. At a time of global warming, poleward peripheral populations of a species may be more important than those closest to the equator.

D. Species

As it is thought that half the species of larger vertebrates are at risk of extinction in the next 100 years, most discussions of the conservation of biodiversity focus on species. Species are fundamental units of evolution and classification (taxonomy). However, despite their centrality to the field, conservation geneticists rarely work at the level of whole species and concentrate instead on infraspecific levels of organization.

Ernst Mayr introduced the biological species concept in 1942; species were defined as groups of actually or potentially interbreeding natural populations that are reproductively isolated from other such groups. This concept stimulated an enormous amount of research in the second half of the twentieth century. Despite its impact, problems with Mayr’s original definition, with its overemphasis on reproductive isolation and its focus on sexually reproducing outcrossing populations over short time spans, led to the development of at least 18 alternative concepts by the end of the century. Although some of these have limited operational utility, three newer biological species concepts are of particular relevance to conservation geneticists. Wiley’s evolutionary species concept, Templeton’s cohesion species concept, and Mallet’s genetic cluster species concept. All seek to recognize discrete groups of populations with a shared evolutionary future. All three concepts, like Mayr’s, seek to capture the essential genetic relatedness within, and the genetic distance between, biological species. All allow that the absolute values of observed relatedness and distance will vary with the geological age of a species and will vary in different groups of plants and animals.

Mayr’s biological species concept forced researchers to search for reproductive isolating mechanisms between species in nature and to investigate their potential significance in laboratory and greenhouse hybridization experiments. Such work typically took years and the results were often compromised by methodological limitations. It became clear that morphology was not always a good indicator of species boundaries and not surprisingly that traditional taxonomy was often a poor guide for conservation decision making. The introduction of protein electrophoresis and molecular genetic methods of measuring genetic variation has dramatically changed the approach to defining species. It is now possible to quickly establish whether populations exchange genes or whether they are effectively reproductively isolated. It is possible to estimate genetic distances between groups of populations and gauge their significance in comparison to within-population variation. It is possible to estimate the time since a speciation event and the historical patterns of gene flow within and between taxa. Our newly found ability to characterize and recognize species genetically does not diminish the value of field and laboratory studies of behavioral ecology, but it does permit geneticists to make powerful contributions to the practice of conservation.

Although it is fortunately still unusual for geneticists to work on the conservation of entire species, there are an increasing number of cases in which every individual in a species has been genotyped to some degree for management purposes. Such cases include Przewalski’s horse (Equus przewalskii), San Clemente loggerhead shrike (Lanius ludovicianus mearnsi), whooping crane (Grus americana), and Catalina mahogany (Cerocarpus traskiae).

E. Communities

Although it is unusual to think of geneticists working at higher levels of organization than species, regional multispecies phylogeographic surveys are useful in defining the historical interactions of whole communities of organisms. Studies of regional phylogeographic structure, as for example in the southeastern United States or savanna ecosystems of Tanzania, are relevant to the design and maintenance of biodiversity sanctuaries (Avise, 2000). Elsewhere, populations in ecoregions have been found to have higher gene diversity...
and are thus recommended for higher conservation priority (Smith and Wayne, 1996).

III. EVOLUTIONARY PROCESSES OF INTEREST TO CONSERVATION GENETICISTS

A. Mutation

Mutations, the occurrence of heritable changes in the genetic material, are typically very rare processes that ultimately provide the raw material of genetic variation upon which the other agents of evolution operate. Mutations span a wide array of phenomena; from single base pair changes in the genetic code to accidental doublings of the number of chromosomes in a gamete. Many mutations are deleterious or lethal, some are near neutral, and a few may be beneficial to the carrier. The vast majority of mutations are completely invisible in the phenotype and can only be detected with an array of genetic techniques.

Mutations become of concern to conservationists in a couple of circumstances. First, the presence of a normally very rare allele of major effect in a remnant or closed captive population can have serious consequences. Deleterious traits discussed by Ryder and Fleischer (1996) include hairlessness in red-ruffed lemurs (Varecia variegata rubra), funnel chest in black-and-white ruffed lemurs (Varecia variegata variegata), and congenital diaphragmatic hernia in golden lion tamarin (Leontopithecus rosalia). Second, artificial selection for rare alleles is sometimes the goal of captive breeding programs. White tigers (Panthera tigris), homozygous for a recessive allele, are beautiful but suffer genetic disease with metabolic, nervous, and developmental consequences. Selection for such traits in the context of the conservation of an endangered species is unjustified.

As most populations of conservation concern lose genetic variability, the question arises as to whether new mutations will replace variability lost by population extirpation and genetic erosion. The answer is yes and no. Given that mutation rates are typically on the order of one per $10^7$ cell divisions, the time for the accumulation of new variants in a population is on the order of tens of thousands of years. Therefore, conservation geneticists are more concerned with the deleterious effects of mutations in small populations than with their very long-term benefits.

Extinction due to genetic causes is almost unknown, but their contribution to the process should not be ignored. Although natural selection purges deleterious alleles from populations almost immediately, mildly deleterious, near-neutral mutations will gradually increase in frequency and become serious problems when their frequencies exceed $0.05$ or $1/(2N_e)$. Fortunately, this process takes hundreds of generations in all but the smallest isolated populations. Therefore, although such mildly deleterious mutations are rarely considered by wildlife conservationists, they will ultimately diminish the long-term viability of many threatened taxa. On the other hand, managers of captive populations have to recognize this threat from the outset. If the goal of a breeding program is to return a captive population to the wild, then managers should maximize genetic variation including mildly deleterious mutants. Alternatively, if a population cannot be returned to the wild and must be sustained in captivity for many generations, then managers will need to purge deleterious mutations as they are identified.

Mutation rates at near-neutral genes controlling quantitative characters set a lower limit on the population size necessary for future evolution. Harmful mutation rates set lower limits for population sizes for avoiding inbreeding depression and for preventing genetic erosion of fitness by the accumulation of mildly deleterious mutations. The suggestion that small populations ($N_e < 100$) may decline in fitness with the accumulation of mildly deleterious mutations, termed mutational meltdown, is under theoretical and experimental study.

B. Mating System

Although much population genetic theory is premised on “random mating,” such behavior is rarely observed in nature. Even related species may have very different mating systems with very different genetic consequences. Self-fertilization in hermaphroditic organisms and obligate outbreeding in dioecious organisms are the two extreme modes. Conservation managers have to be aware of these differences if they are to mimic a species’ natural history.

The most extreme examples of conservation problems involving mating systems involve cases in which the last surviving individuals in a sexually reproducing population all belong to the same sex. The last passenger pigeon (Ectopistes migratorius) was a female; the last member of one of the 11 surviving subspecies of Galapagos tortoise (Geochelone elephantopus) is a male, Lonesome George. It is likely that cloning technologies currently under development will be applicable to saving such lineages in the future.
Inbreeding refers to the mating of close relatives in species that are normally outbreeding. Matings between father and daughter, brother and sister, or first cousins are examples of inbreeding. Many species of plants and animals have evolved devices to minimize close inbreeding. Species vary greatly in their tolerance of inbreeding; some trees and dioecious plants are obligate outcrossers. The genetic underpinnings of inbreeding depression are best understood in Drosophila, in which recessive lethal mutations and mildly deleterious mutations are major causes. Gradual inbreeding permits natural selection to purge the former but the partially recessive near-neutral mutations continue to increase in frequency and significance. Outcrossing populations that suddenly decline in numbers usually experience reduced viability and fecundity known as inbreeding depression. Inbreeding produces increased homozygosity of recessive partially deleterious mutants and by chance in small populations these alleles become fixed. In the simplest genetic situation of a trait under the control of a lethal recessive allele, there is an increased risk that the offspring of two related healthy but heterozygous individuals will inherit the harmful allele from each parent and die. Although the risk in this case is only one in four, this amounts to a powerfully strong fitness differential on which natural selection will act. Generalizing from this simplest single-locus case, geneticists speak of inbreeding depression as the manifestation of the whole genomic effects of mating of close relatives. These effects may involve outright genetic disease (congenital abnormalities) but are more often subtle and appear as decreased growth rate, behavioral abnormalities, and reduced fertility and fecundity. Therefore, inbreeding is rare in typically outbreeding populations but becomes a serious problem in small isolated populations. In small fragmented populations in nature and in captive populations, inbreeding depression can threaten population viability. Animal breeders learned these lessons from centuries of experience with artificial selection, and they limit inbreeding rates to less than 2% per generation. The consequences of very close inbreeding are well illustrated by experience with inbred strains of laboratory mice; the majority of inbred lines die out within 10 generations. There is abundant evidence that captive wildlife populations suffer inbreeding depression. Ralls (in Schonewald-Cox et al., 1983, pp. 164–184) was the first to show that even well-intended captive breeding programs subjected small populations to inbreeding depression. She reviewed empirical records for 40 species, mainly ungulates, in zoos and found inbreeding to be a problem in most cases. Inbreeding is also associated with decreased growth rate and blindness in a captive wolf population in Sweden. In the wild, the final decline of the heath hen (Tympanuchus cupido) on Martha’s Vineyard island in 1932 involved inbreeding effects. Other better documented cases (Hedrick in Smith and Wayne, 1996, pp. 459–477; Lacy, 1997) involving declining or threatened wild populations involve the middle spotted woodpecker (Dendrocopos medius) in Sweden, the Florida panther (Puma concolor coryi), Barrow Island populations of the black-footed rock wallaby (Petrogale lateralis), common shrews (Sorex ananus) in England, deer mice (Peromyscus polionotus), and Glanville fritillary butterfly (Melitaea cinxia) in the Aland Islands, Finland (Saccheri et al., 1998).

In 1980, Franklin and Soule independently showed, based on theory and experiments, that inbreeding depression can be avoided in the short term if \( N_e > 50 \). The inbreeding coefficient \( F \) increases by \( 1/(2N) \) per generation, and centuries of animal breeding experience shows that a 1% increase in \( F \) per generation is tolerable; thus, an \( N_e = 50 \) is necessary to avoid inbreeding depression. Franklin and Soule further concluded that an \( N_e > 500 \) was necessary to enable a population to continue to evolve in the long term. Although this 500 number has been revised upwards, the simplicity of the Franklin–Soule numbers caught the attention of managers and legitimized the role of genetics in conservation. The theory behind the 50 number is still accepted (Lande, 1999); but it is important to realize that its derivation was based on controlled laboratory experiments; larger \( N_e \) are required in nature, where environmental fluctuations are more severe and stressful.

### D. Outbreeding

Outbreeding, or the crossing of unrelated individuals, is widespread in nature. It is widely believed that sexual reproduction evolved in part because chromosomal crossing over and recombination facilitated by outbreeding produces more genetic variability than do other mating systems. Many species of plants and animals have effective immunological and behavioral mechanisms to favor outbreeding. The latter include sex-based dispersal of young adults from their natal population and elaborate courtship behaviors. Even in plants with both male and female flowers, outbreeding is ensured by asynchronous maturation of male and female gametes and the evolution of various self-incompatibility systems.

Outbreeding depression occurs when very distantly...
related conspecific individuals are mated or when members of two different but related species hybridize. The male and female genomes are sufficiently different to produce a hybrid with genetic disorders. Conservation geneticists encounter outbreeding depression in inadvertently mixed captive populations. Sterility, or partial sterility in one sex, and high neonate mortality are commonly observed manifestations.

E. Hybridization

Outbreeding depression occurs in nature in some hybrid zones between semispecies and species of plants and animals. Hybrids are interesting because they show that the evolution of many groups of plant and animal species involves both lineage splitting and lineage anastomosis. Fertile interspecific hybrids permit gene flow between species (introgression). Hybrids call into question species definitions that overemphasize reproductive isolation. The notion that “species” are somehow “purebred” and always reproductively isolated from their close relatives is not borne out by observations of some animal groups and many plants, in which low rates of hybridization between congeners often occur in nature.

In the past, it was argued that hybrid populations did not qualify for legal protection under the U.S. Endangered Species Act. However, hybrids are very much a normal part of nature. Rare or very rare in some groups, and more common in others, they present conservationists with a dilemma because their occurrence appears to diminish the value of a taxon. Should one save Florida panthers if they are known to harbor genes of introduced South American panthers, a different subgroup? Do Texas red wolves (Canis rufus) merit conservation if they are gray wolf–coyote hybrids? Should one save the remaining Przewalski’s horses if it is shown that historical mismanagement resulted in a large fraction of the surviving animals being tainted by the genes of domestic horses, a karyologically distinct species? Whether hybrids should be afforded the same priority as nonintrogressed populations or species will remain controversial; in the previous cases, the answer was yes and geneticists contributed to pedigree management.

Habitat disturbance can result in increased opportunities for hybridization between species that would not normally interbreed. Fragmentation of the recently continuous Pacific Northwest old-growth forest has led to hybridization between the declining northern spotted owl (Strix occidentalis) and the barred owl (S. varia), which favors disturbed sites. Hybridization is more common in plants than in animals; therefore, not surprisingly it is in plants that there are numerous examples of rare species being hybridized to extinction (genetic assimilation) by hybridization with a more common sympatric congener (Solís and Gitzendanner, 1999). This is the case for the Catalina Island mahogany, in which 3 of the remaining 11 adult trees are actually hybrids with the more common mountain mahogany (Rieseberg and Swensen in Avise and Hamrick, 1996, p. 305–334). Other cases involve plants (Asteracea: Argyranthemum) in the Canary Islands. The Simien jackal (Canis simensis) of Ethiopia is at risk of being introgressed into extinction by hybridization with domestic dogs. It is now recognized that restocking rivers with genetically uniform hatchery-bred salmon has contributed to the collapse of the Pacific Northwest salmon runs. Hatchery fish show reduced fitness in the wild (they are not locally adapted) and compete and hybridize harmfully with the remaining wild salmon (Lande, 1999).

F. Gene Flow

Gene flow is a fundamental agent of evolution based on the dispersal of genes between populations of a species. It involves the active or passive movement of individual plants, animals, gametes, or seeds. Gene flow involves not just dispersal but also the successful establishment of the immigrant genotypes in the new population. Gene flow is often confusingly referred to as migration, but the latter term is best reserved to describe dispersal behaviors involving a seasonal or longer term round-trip. Gene flow tends to homogenize linked populations and lack of gene flow permits interpopulation differentiation. It is of interest to geneticists and managers in that to conserve a population one needs to establish the historical patterns and rates of gene flow. This is typically estimated from allele frequency data and reported in terms of the number of “migrants” per generation. In theory, one migrant per generation between two populations will ensure that they remain genetically homogeneous. Inbreeding depression can be ameliorated by the artificial translocation of one reproducing migrant per generation between populations.

Gene flow is often gender biased and limited to certain phases of the life cycle. It may be accelerated under certain climatic conditions that occur at frequencies of many years or at irregular intervals many years apart. Interspecific gene flow results in introgressive hybridization (discussed previously). The translocation of individual organisms results in gene flow if they reproduce at the release site. In the future, genetically depauperate populations will be enhanced by transloca-
tion of individuals from more secure areas. Unfortunately, such genetic enhancement carries risks associated with the introduction of pathogens that could harm the target population or completely unrelated species. Furthermore, the introduction of individuals from genetically well-differentiated source populations may result in outbreeding depression in the threatened population of conservation concern (discussed previously). Gene flow can thus erode the genetic basis of adaptation to local conditions.

If previously continuous populations become fragmented, historical patterns of dispersal and gene flow may be disrupted with potentially serious consequences for population viability. For example, if young female chimpanzees can no longer emigrate from their natal social group because of habitat destruction in the surrounding countryside, their isolated natal population will experience increased inbreeding.

G. Genetic Drift

Genetic drift involves the loss of alleles from a population by chance. Random fluctuations in allele frequencies in small populations reduce genetic variation, leading to increased homozygosity and loss of evolutionary adaptability to change. The rate at which alleles are lost from a sexually reproducing population by genetic drift can be predicted. Sewall Wright (1969) developed the basic theoretical model in 1931 and showed analytically how the rate varies with population size. Actually, it is not the census size (N) that is important but rather the genetic effective population size (N_e). This parameter takes into account the fact that closely related individuals will share alleles by common descent. Monozygotic twins are genetically identical and therefore should be counted as one individual rather than two. Sibs share half their genes with each other and half with each of their parents and are therefore not equivalent to two genetically unrelated individuals. The genetic effective number of individuals in a population is therefore almost always less than the number of individuals counted by an ecologist. N, can, under some breeding systems, be one or two orders of magnitude less than N. Consider, for example, the number of adults in a sexually reproducing population. In a monogamous species the census count of adults is useful, but in a harem species only 1 of the 10 males may be contributing to N. N, can be variously defined in terms of unequal sex ratios among breeders, fluctuations in population size over several generations, and variance in family size (Lande and Barrowclough, 1987).

Wright (1969) defined the variance effective population size (N_e) as the number of individuals in an ideal population that would experience genetic drift at the same rate as the actual population. N, can be defined and estimated in various ways using temporal ecological data, DNA sequences, and various methods of estimating migration rate. Some methods of estimation have theoretical value but little operational utility—it is almost impossible to determine the values that some algorithms require. Nevertheless, by estimating N, one can assess the effects of different population management strategies. Unequal numbers of males and females, increased variance in family size, and temporal fluctuations in N all cause N, to be much less than the census size, N. In many endangered populations N, is only 10–30, and at such levels genetic variation becomes significant for a population’s viability.

H. Population Bottlenecks

Sudden population declines followed by recovery in numbers are referred to as population or demographic bottlenecks. They can have an immediate impact on viability at molecular genetic loci as genetic drift robs the population of its innate variation. The evidence of a bottleneck may persist for hundreds of thousands to millions of generations in low levels of variation at allozyme and molecular genetic marker loci. However, large populations that are almost isogenic at such loci may maintain high heritable variance in QTL, low inbreeding depression, and high heterozygosity for simple repetitive microsatellite DNA because variation at QTL may return to outbred levels in 10^3 to 10^4 generations. Furthermore, bottlenecks can actually result in a short-term increase in population variation because epistatic variation (due to interactions among genes controlling a trait) is converted into additive variation. Whether such release of previously hidden variation is beneficial or harmful to population viability is unknown.

Sudden reduction in N results in a loss of fitness unless there is a rapid and sustained recovery. Gradual reduction, on the other hand, permits natural selection to purge recessive lethal mutations and avoid a substantial part of inbreeding depression. The best advice a geneticist can give the manager of a collapsed population is to increase N as fast as possible and then worry about genetics.

Very low variability is known for many sexually reproducing species whose currently large populations have recovered from one or repeated brushes with extinction. If a large variable population collapses, then
the few individuals that survive the catastrophe carry only a fraction of the original population's genetic variability through the demographic bottleneck. By chance, some individuals and the alleles they carried are lost to genetic drift. Only one of six mtDNA haplotypes survived the severe bottleneck (N = 14) in the whooping crane (Grus americana) in 1930. Genetic drift becomes a significant agent of evolutionary change in small populations. Drift may account for the very low levels of variability observed in African cheetahs and northern elephant seals. Cheetahs were not known to be genetically less variable than other cats or at genetic risk until half of a large captive breeding colony died soon after being exposed to a common domestic cat coronavirus (feline infectious peritonitis virus). Although the northern elephant seals are known to have recovered after having been overhunted to near extinction in the late nineteenth century, the low levels of variation in the cheetah may be attributable to metapopulation dynamics rather than a classic population collapse. Metapopulation structure, with frequent extinction and recolonization of subpopulations, can reduce metapopulation N, orders of magnitude below the census populationsize and mimics the genetic effects of a demographic bottleneck.

I. Genetic Erosion

Genetic erosion, the decrease in population variation due to random genetic drift and inbreeding, is both a symptom and a cause of endangerment of small isolated populations. Population genetic theory shows that variation will be lost by genetic drift with an almost clocklike regularity (Wright, 1969). In closed populations, in the absence of factors promoting genetic variation (mutation and gene flow) the expected rate of loss of heterozygosity, or rate of loss of genetic variance in quantitative characters or selectively neutral variation, is 1/2N per generation. Little variation is lost in any one generation but small N sustained for several generations can severely deplete variability. Most variability is lost within 2N generations. An effective population of 100 is predicted to lose heterozygotes five times faster than a population of effective size 1000; 50% of its heterozygosity will be lost in approximately 20 generations. Therefore, in theory, small isolated populations have a higher rate of loss of heterozygosity and are expected to have lower levels of genetic variation than large continuously distributed populations. Because variability is inherently related to evolvability, genetic erosion in small, recently fragmented populations may contribute to their endangerment. Barrett and Kohl (1991) and Young et al. (1996) review the growing literature of the population genetic consequences of habitat fragmentation for plants. There are numerous examples of a positive relationship between N_e and population genetic variation at allozyme loci for remnant populations. Outbred and colleagues conducted experimental investigations of genetic erosion with Scabiosa columbaria and Salvia pratensis plants under common garden conditions and found positive correlations between variance for adaptive traits related to growth rates and population size. Recently, others studied Clarkia pulchella and found an increased probability of extinction associated with decreased N_e. Such experimental studies indicate the potential significance of genetic erosion in natural populations.

The phenomenon of genetic erosion has long been understood in terms of population genetic theory, but the critical early stages of the process in nature have gone undocumented because the changes are rapid and difficult to monitor. I developed a new molecular genetic method for monitoring genetic erosion and provided evidence for its commencement in mammal populations isolated recently in small forest fragments (Srikwan and Woodruff, 2000). An opportunity to study genetic erosion empirically arose when 163 km^2 of lowland rain forest were flooded in 1987 following construction of a hydroelectric dam on Khlong Saeng, southern Thailand. Former hilltops became permanent islands in Chiew Larn reservoir and retained their original fauna of 12 species of small mammals. During years 3–8 postfragmentation, the demographic collapse of these communities and genetic erosion in three common species [a forest rat (Maxomys saviri), a tree mouse (Chirodipomys glonioides), and a tree shrew (Tupia glis)] whose populations were effectively isolated on some islands were monitored. Nuclear microsatellite markers are sufficiently variable to be used to monitor the process of genetic erosion in nature. As expected, small populations lost variability by genetic drift faster than large populations, and allelic variability is a better indicator of the onset of genetic erosion than heterozygosity. Interestingly, in one of the three species studied, genetic erosion commenced before detectable demographic decline. This demonstration that the process of genetic erosion can be monitored in free-ranging natural populations provides numerous research opportunities because habitat fragmentation is a very ubiquitous phenomenon. Furthermore, the method can easily be upscaled to larger mammals of conventional concern to conservationists. Although monitoring genetic ero-
sion in long-lived species may not be practical, much can be learned immediately by comparing isolated pop-
ulations to those still more continuously distributed.

The importance of such rapid genetic erosion on population viability remains unclear because there are so few studies of the process in nature. Two larger questions remain to be answered: At what point (in terms of \( N \) and \( N_e \)) does genetic erosion threaten a population's viability? and What level of natural or artificial gene flow can protect a population from the negative effects of genetic erosion? Answers to such questions may emerge from the 35-year study of the decline and assisted recovery of an isolated population of greater prairie chicken, \( Tympanuchus cupido \) (Westemeier et al., 1990). Unfortunately, there are very few studies of this duration.

Genetic enhancement, the introduction of selected individuals into a threatened population with the intent of maintaining or increasing its genetic variability and hence viability, is a conservation method in its infancy. Although the genetics may seem straightforward, the translocation of new individuals carries a significant risk of introducing diseases into the threatened population. Furthermore, ill-planned genetic enhancement may lead to a breakdown of local adaptation (outbreeding depression) and actually decrease a threatened population's viability.

J. Natural Selection and Adaptation

Natural selection, the differential survival and reproduction of some genotypes over others, is the major agent of microevolutionary change. It is of interest to conservation geneticists for two reasons. First, human activities can radically alter selection coefficients in both natural populations and managed ones. Such human-influenced evolutionary change is termed artificial selec-
tion whether or not it is intentional. Intense harvest-
ing based on size or gender can cause rapid changes in behavior and natural history and reduce fitness. Exam-

ples include reduced body size in game fish and the impact of hunting only male horned or tusked mammals on social behavior. Second, one of the major challenges facing geneticists this century will be assisting species to adapt to ongoing global climatic changes. In the past, in the absence of humans, natural selection favored individuals adapted to change and many species shifted their ranges to accommodate major changes. Unfortu-
nately, in the twenty-first century, the pace of environ-
mental alteration and destruction is too fast for many species to respond. Conservation managers will have to intervene on behalf of many species if they are to survive.

Quantitative characters are typically under stabiliz-
ing selection and show some optimum phenotype from which they can evolve in response to environmental changes. This optimum balances current fitness with the need for future flexibility or adaptability. The main-
tenance of this variability imposes a fitness cost or gen-
etic load on a population—the price for long-term evolvability.

The rate of directional selection that a population can manage, in response to some environmental change, is determined in part by its innate variability. Rapid anthropogenic changes, such as those associated with global warming, place a premium on genetic variation and adaptability, especially in fragmented populations. To maintain variability in quantitative characters (long-
term adaptability), the Franklin–Soule\' number for an effective population size of at least \( N_e \approx 500 \) is often cited. Revisionary work by Lande (1999) has shown, however, that an upward revision to \( N_e \approx 5000 \) is re-
quired. Such numbers are larger than those found in many endangered and threatened populations and underscore the need for genetic vigilance in their manage-
ment and the importance of keeping numbers as high as possible.

The risk of extinction due to fixation of mildly deleterious mutations is comparable in importance to envi-
rionmental stochasticity and could substantially de-
crease the long-term viability of populations with \( N_e \) of less than a few thousand. The current recovery goals for many threatened species are inadequate to ensure their long-term viability if this requires \( N_e \approx 10,000 \) individuals. Genetic and demographic factors, acting synergistically, require that minimum viable popula-
tions be \( >10^3 \).

IV. GENETIC MANAGEMENT:

EXAMPLES

Until a textbook on conservation genetics is written, the reader must consult conference proceedings and the primary literature for examples of the application of genetics to conservation management. The journals Conservation Biology and Molecular Ecology are espe-
cially useful in this regard. Without making reference to the specific methods employed (because these con-
tinuously change), the following examples, in addition to those mentioned previously, are illustrative of the type of contributions conservation geneticists have
made. Details may be found in Loeschcke et al. (1994), Olney et al. (1994), Frankham (1995), Avise and Hamrick (1996), Smith and Wayne (1996), Lacy (1997), Landweber and Dobson (1999), and the specific references cited in the following sections.

A. Pedigree Management in Very Small Populations

Olney et al. (1994) describe examples of multi-institutional breeding plans with genetic components for numerous species, including Przewalski’s horse, Peter David’s deer (Elaphurus davidianus), Hawaiian goose (Branta sandvicensis), California condor (Gymnogyps californianus), and Tahitian Parula tree snails, all of which were extirpated in the wild. Cases of more intensive genetic management, including the establishment of relationships, founder representation, and breeding to maximize N, include captive populations of lion-tailed macaque (Macaca silenus) (San Diego Zoo), Speke’s gazelle (Gazella spekii) (St. Louis Zoo), Waldrapp ibis (Geronticus eremita) (Zurich Zoo), Guam rial (Rallus owstoni), Micronesian kingfisher (Halioclysin cinnamomina), and Mauritius pink pigeon (Nesoenas mayeri). Genetic sex determination of juvenile California condors enabled recovery program managers to pair birds efficiently.

Geneticists have identified low genetic variability as a concern in wild and captive populations of many species, including cheetah, Californian Channel Island fox (Urocyon littoralis), Newfoundland black bear (Ursus americanus), Git Forest Asian lions (Panthera leo), southern koalas (Phascolarctos cinereus), European bison (Bison bonasus), Arabian oryx (Oryx leucoryx), Peter David’s deer, and Torrey pine (Pinus torreyana). A loss of self-incompatibility alleles may pose a threat to reproduction in plants with genetically determined self-incompatibility systems such as the rare lakeside daisy (Hymemenys acaulis) in Illinois.

Geneticists identified inbreeding as a probable cause of reproductive failures in populations of Ngorongoro lions (Panthera leo), Florida panther (Puma concolor coryi), Barrow Island black-footed rock wallaby ( Petrogale lateralis), big horn sheep (Ovis canadensis), Puerto Rican parrot (Amazona vittata), and the Isle Royale gray wolf (Canis lupus).

B. Mating System

Geneticists have discovered that the mating systems of many species differ from expectations based on direct observations, often with profound implications for captive management and for management of “wild” populations. In some species, females preferentially mate with males outside their social group, e.g., Atlantic salmon (Salmo salar), blue tits (Parus caeruleus), and pilot whales (Globicephala sp.). Preferential interpod mating in long-finned pilot whales indicates the importance of conserving as many pods as possible. Genetics shows that highly gregarious black vultures (Coragyps atratus) are in fact monogamous and that Galapagos hawks (Buteo galapagoensis) are polyandrous. Despite behavioral observations suggesting a high frequency of matings between sibs in Australian splendid fairy wrens (Malurus splendens), genetics shows that outcrossing is the norm. In other birds, including stripe-backed wrens (Campylorhynchus nuchalis), geneticists found that subordinate males reproduce. The mating system of wild gray wolf, chimpanzees, and other species have been established and used to improve captive breeding programs or management in reserves. Establishing the mating system of the red-cockaded woodpecker led to improved estimates of N, and changes in the recovery program for a small population in South Carolina.

C. Problems of Hybrids

Geneticists have elucidated the hybridity of some taxa in both the wild and in captivity, including Przewalski’s horses (many of which are domestic horse hybrids), Asian lions (most animals in Western zoos were hybrids between Asian and African lions and were removed from the breeding program), and the red wolf of Texas [shown to be primarily a natural coyote (Canis latrans) × gray wolf hybrid]. Genetic data revealed the threat by hybridization to indigenous Scottish deer (Cervus elaphus scoticus) by introduced Japanese sika deer (C. nippon). Genetics was used to identify hybrids among the remaining Catalina Island mahogany and establish seedlings of pure Cercocarpus taxatione at several sites to protect the species from extinction by introgression. Genetic markers are being used to conserve remnant cutthroat trout populations by identifying and removing populations introgressed by hybridization with nonnative species.

D. Genetic Censusing

Because every individual in most species is genetically distinct, it is possible to census populations by counting unique multilocus genotypes in an area. Geneticists have censused very difficult to count animals such as...
African forest elephants and the largely fossorial northern hairy-nosed wombats (Lasiorhinus krefftii) by individually genotyping dung samples. Fecal genotyping (including sexing) has also been used to provide more accurate census data of animals such as coyotes and bears than could be obtained from long-term ecological surveys. The genetic data also provided information on home ranges and pedigree relatedness without requiring that the animals be seen or disturbed.

E. Phylogeography, Gene Flow, and Population Structure

Avise (2000) reviews many examples of phylogeographic studies that provide managers with essential data on population structure and on the characterization of MUs and ESUs. Geneticists have also provided estimates of historical gene flow between populations that would be impossible to obtain by direct observation (e.g., chimpanzees, humpback whales, and green turtles). Comparative phylogeographic studies of four sympatric species of East African bovids underscore the dangers of extrapolating results from one species to another and support conservation efforts that take species-specific differences into account.

F. Defining Species, ESUs, and MUs

Karyological differences distinguish Bornean and Sumatran orangutans (Pongo pygmaeus) and captive breeding programs are now managed to prevent hybridization of the two. Dik-dik species (Madoqua) can also be distinguished karyologically and neonatal mortality in captive populations has been reduced following the sorting of animals by karyotype. MitDNA sequences have been used to sort sibling species and subspecies of gibbons (Hylobates sp.) of unknown geographic origin into correct ESUs, to distinguish sibling species of chimpanzees (Pan troglodytes and P. verus) that have been mixed in captivity, and to identify subspecies of black tinamoceros, Diceros bicornis. Genetics was used to show that the conservation of the living fossil tuatara (Sphenodon) of New Zealand depends on the management of not one but two genetically recognizable species. There are many cases in which genetic data justify conservation efforts for isolated subspecies or varieties that are shown to be genetically well differentiated—for example, Darwin’s fox (Dusicyon f. puchi) of Chile, Kemp’s ridley turtle (Lepidochelys kempi), San Clemente Island loggerhead shrake, and several “subspecies” of Hawaiian Amalolululu home creeper (Hemignathus vires）。The Mexican wolf (Canis lupus baileyi) was found to be a genetically distinct ESU, untainted by hybridization with gray wolf, coyote, or dog.

The endangered San Clemente Island shrake is illustrative of several of the points made previously because it is technically a subspecies of a widespread mainland bird and might accordingly be written off as a peripheral population, local variant, or “just a subspecies.” Taxonomic practice and field observations notwithstanding, my genetic survey revealed no evidence that it has hybridized with the neighboring mainland subspecies since 1913 despite repeated opportunities. It is genetically differentiated and apparently reproductively isolated and merits management as a separate ESU (Mundy et al., 1997).

Geneticists have also provided data questioning the justification of other taxon-focused conservation efforts. The 27 original subspecies of leopard (Panthera pardus) were found to be referable to only 8 genetically defined subspecies or ESUs. Similarly, genetic variation provides no justification for conserving all 30 described subspecies of puma (Puma concolor). The dusky seaside sparrow (Ammodramus maritimus nigrolineatus), after considerable and unsuccessful efforts to save it, was shown to be only a marginally distinct local race of a common widespread species.

Taxonomic status (species or subspecies) does not automatically lead to the justification of conservation efforts. Proposals to reduce the eastern Pacific black sea turtle, Chelonia agassisi, to a subspecies of the widespread green turtle, C. mydas, are supported genetically but do not justify the abandonment of this taxon as a high conservation priority.

A phylogeographic survey of 14 subspecies of the songbird bananaquit (Coereba sp.) from 19 West Indian islands and the mainland of South and Central America showed that if it were necessary to restock the small and vulnerable populations on the northern Lesser Antilles Islands, the birds from nearby Puerto Rico or from Jamaica would be genetically inappropriate. Similar studies of orioles (Icterus sp.) are relevant to the conservation of the Montserrat oriole, I. oberi, which is under threat of catastrophic (volcanic) extinction.

Phylogeographic analyses have also helped to define natal homing patterns in marine turtles on foraging grounds. It was found that both green turtle and loggerhead turtle rookeries are demographically autonomous and that low levels of interrookery matrilineal exchange suggest that extirpated colonies are unlikely to recover by natural recruitment of nonindigenous females. Similar methods have been used to define stocks in whales and dolphin and elucidate the migratory strategies of...
different groups of humpback whales intermingled at the breeding ground near Hawaii.

Hierarchical phylogenetic analyses have been used to suggest conservation priorities among species of cranes. Within a group of related species the evolutionarily oldest monophyletic clades are considered to represent a greater genetic heritage than recently originated clades (Forey et al., 1994). Similarly, it has been argued that areas with a disproportionate number of evolutionary ancient genotypes are more valuable than areas populated by recent colonists.

G. Reintroductions, Translocations, and Genetic Enhancement

Two genetic generalizations complicate the prospects for successfully moving organisms around in the wild or for returning them to the wild after a period in captivity. First, the chances for successful reintroduction are diminished by evidence for rapid genetic adaptation to captivity in fish, plants, and Drosophila. The same problem applies to wildlife brought into captivity for reintroduction at some future time. Second, in plants, cryptic local adaptation results in the fitness of transplants being about half that of residents even when environments are apparently similar. This makes it difficult to reestablish populations once they are extirpated.

Genetic criteria were used to chose founders for a new population of the extirpated Guam rail on the island of Rota, for sea otters (Enhydra lutris), and for a Western Australian shrub (Corrigan grevillea) reduced to 27 plants. Genetic data were also used to influence the choice of source population for a Gila topminnow reintroduction program (Vrijenhoek in Loeschcke et al., 1994, pp. 37–33). Fish from a population with high allozyme diversity were selected to successfully replenish the diversity and viability of a declining and nearly monomorphic population rather than fish from a less variable but adjacent population.

Genetic study of red-cockaded woodpeckers in the southeastern United States led to the recommendation that translocations be made between nearby populations rather moving birds over great distances. In another study, it was shown to be genetically appropriate to use gray wolves from British Columbia as a source of animals for reintroduction into Yellowstone National Park. Genetic variation and multilocus genetic similarity were used to justify the introduction of panthers from Texas (traditionally regarded as a different subspecies) to counter severe inbreeding depression (low heterozygosity, poor sperm quality, and cryptorchidism) in the remnant Florida panther population.

Genetic criteria have also been used to argue against certain types of translocations. For example, it was found that Tasmanian eastern barred bandicoots (Parmales gunni) should not be used as a source population for enhancing the endangered mainland Australian population. South African wild dogs (Lycaon pictus) were genetically inappropriate for reintroduction into Kenya. Isolated northern and southern populations of Brazilian muriqui (Brachyteles arachnoides) are well differentiated genetically, so threatened northern populations should perhaps not be translocated to larger southern reserves.

An example of the hazards of implementing a transplantation program without first considering genetic factors involves the endangered Hawaiian silversword. In this case, the outplanted individuals were all descendents of only one or two maternal plants and therefore retained only a small fraction of the genetic diversity of the remaining populations of the species they represented. Furthermore, because they were so closely related to one another, they had significant reproductive problems associated with self-incompatibility and a seed set of <20% (Reiseberg and Swensen in Avise and Hamrick, 1996, pp. 303–334).

H. Conservation Management

Genetic methods are used in forensic identification of tissues of endangered species in illegal or mistreated trade (e.g., abalone, "caviar," cage birds, and primates including chimpanzees). Misrepresentation occurs when a wild-caught parrot or falcon is claimed to be legally captive bred. Geneticists showed that some whale meat legally on sale in Japan was actually meat of endangered and allegedly protected species including humpback whale. Sequence data revealed that loggerhead turtles from Caribbean nesting beaches are threatened by a Mediterranean fishery off Spain.

Geneticists are playing an increasingly important role in the management of mixed stocks of threatened and commercially harvested fish. Consider the conservation management of salmon (Oncorhynchus) in the U.S. Pacific Northwest. Salmon with their precise homing behavior present a major problem because each local spawning population should be managed as a genetically distinct taxon. Should managers give every natal stream-adapted salmon stock equal priority? Geneticists have developed allelic frequency marker systems for stock identification that are sufficiently sensitive to permit real-time regulation of mixed stock fisheries.
involving both hatchery and wild salmon. The latter can be partially protected because they return from the sea to the rivers a few weeks later than the former and the fishery can be terminated when they are detected. Using the same approach, geneticists developed markers to identify endangered and protected winter-run chinook salmon in the Sacramento River in California. Fall- (hatchery) and spring-run chinook salmon do not enjoy protection, but the fishery could not be managed from a conservation perspective until the different races could be identified during the downstream migration of smolts to the ocean.

Genetic tracking of movements of migratory birds permits sorting arctic and Mexican falcons (Falco peregrinus) that mix when the former reach their wintering grounds. Similar genetic tracking permitted the identification of the wintering grounds of several species of declining arctic shorebirds and led to changes in conservation focus from the breeding grounds to the wintering grounds.

Finally, geneticists have been able to monitor the loss of variability in translocated populations of wild turkey (Meleagris gallopavo), white-tail deer (Odocoileus virginianus), and alpine ibex (Capra ibex).

1. Other Applications

Space does not permit presentation of all the caveats, corrections, and revised recommendations made in most of the previous cases as more data became available and as various tests were repeated. Many of the classic examples are less clear-cut than originally proposed. Space does also not permit mention of all the examples in which genetics showed populations were not genetically depauperate or sufficiently distinct to warrant priority conservation efforts. Such contributions are of equal importance to biodiversity conservation because they free up limited resources for other investigations.

Although most of the previous examples involved captive or wild populations of threatened species, much useful conservation genetics can be done using laboratory animals as model organisms. Valuable experimental tests of conservation genetic principles have been completed using Drosophila, Tribolium, mosquitofish (Gambusia holbrooki), the butterfly (Bicyclus anynana), Mus, and Peromyscus (Frankham, 1995; Leberg in Smith and Wayne, 1996, pp. 87–103). For example, using laboratory populations of Drosophila it has been shown that equalizing family size can double \( N_e/N \), and more intensive pedigree management can increase the \( N_e/N \) ratio 40-fold. Similarly, in most of the examples discussed previously, investigators also examined the genetics of a close relative of the taxon of interest. It is standard practice to develop genetic methods using a common relative as a surrogate before commencing work on a highly endangered taxon. This approach also has the advantage of providing comparative data useful in interpreting the results of a study of an endangered taxon.

V. GENETICS, EVOLVABILITY, AND THE FUTURE OF BIODIVERSITY

The magnitude of the task facing conservation geneticists is daunting. There are on the order of 10 million species living on the planet today and about 2 million of these are recognized and named in a formal taxonomic sense. Describing a new species requires little more than that a scientist know what it looks like and where it is found; unfortunately, this constrains our state of knowledge for most of biodiversity. Closer to 10^6 than to 10^8 species have been characterized ecobehaviorally, and only on the order of 10^3 species have been examined by geneticists. If sound conservation is based on a knowledge of ecology, behavior, and genetics, then we must admit that we are currently capable of scientifically managing the evolution of less than 1000 species. However, the number of species requiring individual management to prevent their extinction in the next 100 years is in excess of 10,000.

Conservation geneticists have devoted their efforts disproportionately toward the charismatic megavertebrates. Whales and cats have received more attention than bats and rats. Given the enormous number of species requiring attention, one might inquire as to how priorities are set. First, most research has gone into species that were favored for utilitarian reasons; they provide us with food, clothing, medicines, recreation, or companionship. Most of this research has been aimed at stock improvement rather than whole genome or species conservation. Second, as already noted, the charismatic megavertebrates and a few groups of flowering plants have received inordinate attention. Perhaps not surprisingly, rare species have also been studied more than common ones. The same applies to phylogenetically unique species, living fossils, and evolutionary relics. However, because the real goal is to save functional ecosystems, conservation geneticists are rethinking their priorities. Although some rare species clearly merit genetic management, it would be better to focus more attention on ecological keystone species whose
activities are critical to the maintenance of entire communities. We also need to know more about the genetics of ecologically successful colonizing species and of clades of species that have evolved very recently (the cichlid flocks of Rift Valley lakes) because their study may teach us how to manage apparently less successful taxa.

Conservation geneticists rarely advocate bringing plants and animals into captivity ‘‘to save them.’’ Species are typically better managed in their natural communities than in isolation. Existing institutions concerned with conservation, however, are not equipped to deal with the magnitude of the task they face (Woodroffe, 1990). Parks and wildlife reserves are the preferred approach to both species and community conservation. Zoos and botanical gardens are extremely limited in what they can accomplish and can at best serve only to shelter a few critical cases that require intensive care: germline frozen storage systems are valuable adjuncts for researchers, but no credible geneticist has yet proposed that we will be able to awaken these frozen tissues and recreate the animals from which they were derived. Although frozen tissue banks are extremely valuable for geneticists, the revitalization of mammoths, quaggas, thylacos, and dodos is still science fiction.

In its first two decades, conservation genetics was perceived by some wildlife ecologists as an unnecessary intrusion into their field. It was argued that demography and behavior are far more important than genetics in saving endangered species. Others have argued that the genetic threat to population viability has been overstated (Lande, 1988). Genetics was viewed as too theoretical and contributing too little and too slowly to the day-to-day efforts to save populations in nature. Furthermore, molecular genetics studies, which are relatively expensive, compete for the limited funds available to the traditional conservationists. Some of the criticisms were justified and some were not. It is easy to disparage the potential contribution of genetics to saving a particular population or species if genetics is defined very narrowly as, for example, the determination of heterozygosity in a remnant population. In the case of cheetahs in Africa, it is clear that predation by the lions and humans is more significant today than low viability. Similarly, in captivity, different husbandry practices in different zoos are more significant than poor sperm quality. However, if one takes a longer term view, the answer is different: Genetics is and will be increasingly important. As this review shows, geneticists have a great deal to offer managers. It is incorrect to suggest that ecology and genetics are alternative approaches. Although there are clearly times when genetic studies will be lower priority in a multifaceted conservation strategy, it is undeniable that increasingly more populations will need genetic management. Genetics, ecology, and behavior are all necessary parts of biodiversity conservation.

Although conservation geneticists focus on populations and species, their ultimate goal is the conservation of things but of a process, evolution, that produced them. The ultimate goal of conservation biology is to preserve the processes of organic evolution—to maintain the ability of populations and species to evolve and communities to function and provide ecosystem services. The basic science is still not equal to the task conservation geneticists are expected to perform. The relationship between genetic variation and ‘‘genetic health’’ is illusive and needs clarification. Society’s expectations of conservation geneticists also need to be specified or we will forever be accused of treating the symptoms and not the causes of the biodiversity crisis.

Typically, species are not afforded legal protection until their populations have fallen into the hundreds, 10–100 times below the level at which their genetic integrity and viability are reasonably secure. Geneticists need to point out that current standards of endangerment are far too low, that recovery from previous mass extinctions took on the order of 10 million years, and that we have not thought through the global implications of a 50% decrease in the number of remaining large plant and vertebrate species (Myers, 1996).

See Also the Following Articles

Biodiversity, Evolution and Conservation Biology, Discipline of; Genetic Diversity; Inbreeding and Outbreeding; Population Diversity, Overview; Population Viability Analysis (PVA); Species Diversity, Overview; Subspecies, Semispecies

Bibliography


Forrey, P. L., and Humphries, C. J., and Vane-Wright, R. I. (Eds.)
I. What Is Population Viability Analysis?

II. Stochastic Population Processes and Extinction

III. Why Do PVA?

IV. Methods and Tools for PVA

V. Testing PVA

VI. Criticisms of PVA: What Makes a Good PVA?

GLOSSARY

stochasticity A stochastic process is one in which the state of the system cannot be precisely predicted given its current state and even with a full knowledge of all the factors affecting that process. In a population context, we may know the detailed life history parameters of a species. However, various unpredictable (stochastic) processes, such as the chance nature of birth and death (demographic stochasticity), year-to-year variation in climate (environmental stochasticity), and catastrophes, means that accurately predicting the precise size of a population in the future is not possible. Despite this, we can use probability theory and simulation models to make probabilistic predictions. A stochastic population model is one in which each possible future population size has an associated probability. This approach to prediction is the same as stating that the chance of getting a head with the next toss of a fair coin is 50%. Our prediction is accurate but we cannot say if the outcome will be a head or a tail.

viability (and related concepts) A viable population is one deemed to have a reasonable chance of long-term persistence. Such a definition begs two questions: What is reasonable and what is long-term? It has been stated that a viable population is one that has a 99% chance of persisting 1000 years. Different managers, researchers, and policymakers use different measures that range from a 99% chance of persisting 1000 years to a 95% chance of persisting 100 years. The population size that just meets the definition of being viable is called a minimum viable population (MVP). One of the key questions of reserve design is how big does a reserve, or network of reserves, need to be to contain an MVP. This size is called a minimum viable habitat area for a particular species.

POPULATION VIABILITY ANALYSIS (PVA) is a process in which the extinction probability of a population is assessed. This article discusses the processes that interact to cause extinction, and this discussion is then used as the basis of explaining how the extinction process can be modeled. An emphasis is placed on recent attempts to provide quality control for PVA and use the models as a decision support tool for conservation planning.
I. WHAT IS POPULATION VIABILITY ANALYSIS?

Traditionally, population viability analysis (PVA) is a process in which a stochastic population model is used to assess the viability of a population. The population may be a species, subspecies, metapopulation, or an isolated subpopulation of a single species. Classical PVA uses a demographic population model to derive a result such as the following: Given current circumstances the probability of extinction of species \( X \) over the next \( Y \) years is \( Z \)%.

For example, in his seminal 1983 paper, Shaffer used a demographic population model to determine the probability of extinction of grizzly bears, \( Ursus arctos \), in Yellowstone National Park within the next 100 years.

Most PVAs use a demographic population simulation model which incorporates all the processes likely to affect the dynamics of the population. PVA differs from other population modeling exercises in its focus on predicting the probability of extinction and the inclusion, typically, of stochastic processes that are believed to have a strong impact on the probability of extinction. For convenience, we categorize these stochastic processes into four types; each can be thought of as a way in which variability influences the dynamics of a species. Given that these processes are central to our thinking about PVA, they are described in detail.

During the past decade, we have realized that the simple view of PVA as a process using a demographic simulation model to assess extinction risk is too narrow. First, the restriction of PVA to the use of demographic population modeling software ignores other tools for assessing viability. We now view PVA as a broader process that includes various sorts of theory, analytic and computer modeling, and forms of data analysis. Second, our ability to accurately assess the extinction probability of a species is questionable. The broader roles of PVA are explored later. It is this management-based approach to PVA that is now broadly accepted and considered practically useful.

II. STOCHASTIC POPULATION PROCESSES AND EXTINCTION

A. Types of Demographic Population Fluctuation

1. Demographic Stochasticity

Demographic stochasticity is the chance nature of birth and death. It causes populations to fluctuate because populations are composed of individuals that are units. Each unit counts as one and is born and dies as a unit, so populations can only increase and decrease on the set of integers. Each individual has a probability of dying and a probability, each year, of finding a mate and producing offspring. The dynamics of five separate populations, fluctuating as a result of demographic stochasticity alone, are shown in Fig. 1.

Each population has its own trajectory even though they all had the same life history parameters and started from the same population size. In a sense, demographic stochasticity is the most fundamental phenomenon that causes populations to fluctuate because it occurs in all populations. In Fig. 1 each population should, on average, increase 5% per year, but one becomes extinct. The importance of demographic stochasticity for extinction and PVA diminishes in larger populations because the minor “wobbles” caused by the unitary nature of birth and death are overshadowed by the more dramatic environmental stochasticity and catastrophes (Fig. 2).

A rarely discussed type of demographic stochasticity is the chance nature of sexual determination in organisms with more than one sex. For example, a population may be known to have six individuals, which may appear to have some chance of persisting if conditions are good. However, there is a 1 in 32 chance that all the individuals are identical and the population is doomed if each has an equal chance of being male or female. When populations are small, a skewed sex ratio can have a major impact on the chance of extinction.

2. Environmental Stochasticity

As biotic and abiotic conditions vary from year to year, there will be variation from year to year in vital rates. This sort of variation tends to affect the entire population, therefore, there are years in which breeding suc-
FIGURE 1  Five sample trajectories for a population in which demographic stochasticity is operating. For each individual, each year there is a 30% chance of dying. Those that survive have a 50% chance of giving birth to one more individual. Despite identical conditions the populations diverge rapidly, with one becoming extinct after 15 years and another increasing to more than 40 individuals. With these parameters, in the absence of stochasticity, the population should on average increase at a rate of 3% (~70% of 1.5%) per year.

FIGURE 2  The impact of environmental stochasticity on the dynamics of five populations. For trajectory A, the parameters are the same as those used for Fig. 1 with just demographic stochasticity, except the population started with 100 animals. Note the fairly constant increase of about 5% per year with minor fluctuations caused by demographic stochasticity. For the other four populations, the death rate is not fixed at 30% but varies randomly between 20 and 40% from year to year. Also, the probability of an offspring is not fixed at 50% but varies from 40 to 60% from year to year for the whole population. Note the much wider fluctuations in population size, even when the populations are larger than 100. Also note the differences between the trajectories and dynamics even though the parameters are the same on average—the consequence of stochasticity.
cess may be good or bad. Because this variation tends to operate on the entire population, it causes more dramatic fluctuations in population size (Fig. 2).

3. Catastrophes

Some natural events cause precipitous declines in the sizes of populations, such as fire, flood, disease, and frost. Although these events can be seen as an extreme form of environmental variation, they are often considered separately as catastrophes. We define a catastrophe as being a single event that causes the death of a substantial fraction of a population. Recent analysis implicated catastrophes in the extinction, or near extinction, of many species. The Puerto Rican parrot was almost wiped out by a hurricane, the Guadeloupe oriole by a volcanic eruption, black-footed ferrets by canine distemper, and a population of heath hen by fire. Species with populations that are spatially restricted, such as those on islands, are particularly vulnerable to catastrophes. However, disease appears to be able to decimate continental populations over huge areas. Disease is implicated as a factor contributing to declines of once common species, such as the passenger pigeon and many Australian carnivorous marsupials such as the Tasmanian tiger. Figure 3 shows the trajectory of four populations for which there is a 2% annual probability of 90% of the individuals being killed in a single catastrophe.

Including catastrophes is an important part of any PVA (Mangel and Tier, 1993) but a difficult component to include in models because of limited data on the frequency and impact of catastrophes. Although we often have good data on catastrophes such as fire, and these have been successfully included in models of Australian forest fauna (Lindenmayer and Possingham, 1996), we rarely have good data on catastrophes such as disease that are more difficult to detect but probably just as important.

B. Genetic Stochastic Processes

Genetic stochastic processes change the frequencies of genes within a population. When a population is small and isolated, genetic variation is typically lost from generation to generation. This process is called genetic drift. As frequencies of alleles change at random with each successive generation, genetic variability is reduced because rarer alleles are lost by chance through the stochastic nature of birth and death. A consequence of genetic drift is increased homozygos-
ity, which can in turn lead to juvenile mortality and lower rates of fecundity and growth (inbreeding depression). Inbreeding depression is defined as mating between individuals that are related by ancestry and is more likely in populations that are, or have been, small. Inbreeding depression results in the selective removal of inbred animals and the genes carried by such animals (Lacy, 1993). Morphological defects in threatened species such as the Florida panther (*Felis concolor coryi*) and lack of breeding success in the Puerto Rican parrot (*Amazona vittata*) appear to be a result of inbreeding depression. The restoration of genetic variability is slow and believed to be between 1 and $10^4$ genes per generation per 100 years. Thus, genetic mutations have only a minor effect in reversing the loss of genetic variation during the time frame typical for PVA. A longer term consequence of loss of genetic variability is a reduced ability of a population to adapt to environmental changes. The long-term consequences of lost genetic variability are not well understood.

C. Overview of the Role of Stochasticity on Time to Extinction

Each kind of stochasticity has quite distinct effects on the relationship between extinction probability and population size (or carrying capacity). Analytic and simulation studies have shown that although the mean time to extinction, and hence the viability of a population, grows nonlinearly with population carrying capacity when only demographic stochasticity is operating, the other two forms of stochasticity can make even large single populations quite vulnerable to extinction (Fig. 4). With only demographic stochasticity the mean time to extinction increases very rapidly above species-specific thresholds (usually approximately 50–100 individuals). Therefore, in the absence of environmental variability, catastrophes, and genetic stochastic processes, quite small populations could persist for very long times. Adding environmental variation means that the benefits of increasing population size are diminished and there are no simple thresholds above which populations are likely to be completely secure. Because catastrophes can reduce a population to a small size very quickly, a larger population size brings diminishing returns in terms of viability. When this occurs, the best way to improve population viability may be to ensure that a single catastrophe cannot affect the whole population—a risk-spreading approach. This issue will be discussed later.

D. Deterministic Extinction Processes and Extinction Vortices

A major contribution of PVA to conservation biology is that it focuses attention on stochastic extinction processes in small populations. This should not distract our attention from equally important, more deterministic processes that lead to extinction (Caughley, 1994). Habitat loss and associated fragmentation, increases in mortality rates from introduced predators, and reduced reproduction through habitat modification all have significant impacts on the carrying capacity and rates of increase of populations. Although the distinction between deterministic and stochastic extinction processes is useful conceptually, it is worth noting most important impacts on species have deterministic and stochastic components.

The deterministic and stochastic extinction forces need to be integrated. We can think of the interaction of some of these processes as extinction vortices, in which they nonadditively contribute to the demise of a species. For example, a catastrophe may take a population close to extinction, resulting in a period of inbreeding depression which keeps the population sufficiently low so that demographic stochasticity finally causes extinction. Another example is the way in which habitat loss causes population fragmentation where each isolated population is small enough to fall victim to the interaction of genetic and demographic stochastic processes.

Understanding the process of extinction is funda-
mental to our ability to model it. The obvious question arises: What use is an extinction model?

III. WHY DO PVA?

A. Applications of PVA

Originally, PVA had two general purposes. Shaffer's (1981) original intent was to assess the viability of a population. This assessment of extinction risk is useful in that it informs us of the effectiveness of conservation strategies. In the case of the grizzly bear in the Yellowstone region of the United States, the assessment implied that the species was unlikely to persist in the long term. In the following years, PVAs were tied very closely to the idea of a minimum viable population (MVP). Scientists and managers used PVA to determine if the strategies for conserving a population were adequate. The goal was to use PVA to derive a target population that was large enough to be viable and then to recover the species to that level as quickly as possible.

When one of the primary tools for managing a threatened species includes setting aside an area of suitable habitat for that species, the idea of an MVP extends to that of a minimum viable habitat area (MVHA). Scientists and managers used PVA to determine if the strategies for conserving a population were adequate. The goal was to use PVA to derive a target population that was large enough to be viable and then to recover the species to that level as quickly as possible.

When one of the primary tools for managing a threatened species includes setting aside an area of suitable habitat for that species, the idea of an MVP extends to that of a minimum viable habitat area (MVHA). A MVHA is a useful concept in which the main threatening process is habitat destruction. However, it has limited value when threats to a species are not simply removed by reservation, such as the continued activities of a predator (e.g., the western swamp tortoise in Western Australia). A MVHA can be predicted using a PVA by predicting extinction probabilities as a function of reserve area. Typically, the relationship between extinction probability and area conserved is sigmoidal (Fig. 5). The MVHA is simply derived by choosing an acceptable level of extinction probability and then conserving the corresponding area. In some cases, the possibility that a single habitat patch ensures viability is unlikely because of catastrophic events, and PVA models that allow for multiple populations (metapopulations) are necessary to determine an adequate reservation strategy. Such a strategy depends on the number, size, and spatial distribution of reserves and the nature of the catastrophic events.

Another application of PVA that relies on the accurate assessment of extinction probability for a species is that of ranking threatened species. The International Union for the Conservation of Nature has explicit definitions of the different categories of threatened species: critically endangered, endangered, vulnerable, rare, insufficiently known, and extinct. These criteria are becoming increasingly complex and they vary from country to country. One means of classifying species according to their level of threat would be to do a PVA on every species in the list. Although this has the merits of apparent objectivity, PVAs are rarely used for classifying threatened species because there is insufficient data to parameterize PVAs for so many species. One approach to dealing with the uncertainty associated with extinction probabilities is to take a decision analysis, or fuzzy set, approach in which each species is allocated a probability of being in a particular threatened species category.

In the United States, the ideas of MVPs, MVHAs, and species classifications are given a litigious dimension because of the federal Endangered Species Act. The Endangered Species Act and associated recovery plans and habitat conservation plans state in a variety of ways that species should be managed for viability. Indeed, any development should not compromise the viability of an endangered species. The idea of habitat conservation planning is to allow development where parties agree to the level of mitigation necessary to ensure viability. Because the notion of viability is central to this law and the planning processes that have arisen from the law, PVA has entered the courts in land-management and planning disputes.

B. PVA for Making Management Decisions

During the 1990s, scientists questioned the traditional uses of PVA and also questioned the ability of PVA
to accurately predict extinction risk for most species (Possingham et al., 1993; Taylor, 1995; Heissinger and Westphal, 1998). Workers have found that the probability of extinction of an organism is very sensitive to small variations in key life history parameters, especially adult mortality rates. There appear to be adequate data to make very accurate estimates of extinction risk for only a very small fraction of threatened species.

An alternative role for PVA is as a decision support tool to help make management decisions. This role focuses on the robustness of different management decisions to deliver acceptable conservation strategies. PVA can be used to make decisions regarding the following:

1. Determining if, and how much of, a population should be brought in to captivity
2. Translocation strategies between populations or to new sites
3. Reintroduction strategies
4. Deciding how to set priorities between species
5. Determining optimal habitat management, including disturbance regimes such as fire and logging
6. Control strategies for the predators, competitors, and prey of threatened species
7. Optimal size, distribution, and spatial pattern of reserve systems

Probably the most widespread and frequent application of PVA is through the activities of the Captive Breeding Specialist Group (CBSG). This group aids zoos and conservation agencies throughout the world to set conservation priorities and actions by workshopping species groups region by region.

Some argue that PVA has been most useful in providing a framework within which experts can operate. The task of running a PVA draws the experts together and focuses their attention on estimating population dynamics and identifying threats. This is certainly true for the CBSG workshops and users must view PVA as a decision support tool and not as the decision maker.

C. Other Measures of Viability
The probability of extinction of a population is not the only measure of population viability. Other common measures are the mean and median of the time to extinction. When the extinction probability for a species is fairly constant, the expected life span of the species can be thought of as being negatively exponentially distributed, so the rate of extinction is simply the inverse of the mean time to extinction. All these measures are fairly closely related.

Because of the uncertainties associated with modeling the dynamics of very small populations (Allee effects, genetic effects, etc.), one prudent approach to providing a measure of risk is to introduce the notion of quasi-extinction, which is the risk of the population decreasing in number to some preset size. For example, although the probability of extinction is the chance of decreasing to zero individuals, a manager may argue that a species has failed long before this and prefer to set a threshold population size of, for example, 20, below which the population is quasi-extinct. This approach eliminates some of the problems associated with modeling very small populations.

An even more general approach to assessing extinction risk that extends the idea of quasi-extinction is to develop risk curves. A risk curve (Fig. 6) shows the probability of a population decreasing below any population size, from zero (extinction) to the carrying capacity of the population. With a risk curve, the manager is free to pick whichever quasi-extinction threshold he or she wishes, or to assess alternative management strategies by examining changes in the whole risk curve.

IV. METHODS AND TOOLS FOR PVA
A. Demographic Models for PVA
Generally, a PVA is a demographic model that includes some or all of the extinction processes discussed previously. There are several ways of constructing these models, and there are several publicly available simulation "packages" that people can use to do PVA.
general categories of PVA model are discussed in the following sections.

1. Matrix Population Models

Basic demographic projection is one of the most powerful and important parts of PVA, although sometimes it is bypassed. By constructing an age- or stage-structured population projection matrix, one can determine both the overall growth rate of the population and the sensitivity of that growth rate to variation in life history parameters by finding the dominant eigenvalue of the matrix. More simply, the population can be projected forward in time using a simple spreadsheet. If the dominant eigenvalue of the population is less than 1, then we know the population is doomed to extinction (deterministic extinction). If there is uncertainty about the value of key life history parameters, then the risk of deterministic extinction can be calculated by carrying out a sensitivity analysis on the dominant eigenvalue. Although this approach does not allow for any of the stochastic forces described previously, it gives the manager an idea of how the population is faring in the absence of stochasticity. If the dominant eigenvalue is likely to be less than 1, then our sensitivity analysis allows us to target those life history attributes, such as birth rate or adult death rate, most likely to generate a positive long-term growth rate. Matrix population models can include stochasticity (indeed, RAMAS Age is based on the application of projection matrices for its predictions), but without a computer package the mathematics is quite complex.

2. Analytic Methods and Approximations

Although including stochasticity normally requires that we proceed to a computer simulation, there are analytic methods for calculating the probability of extinction, and in particular the mean time to extinction, of populations. One approach uses Markov chain theory to model a basic birth and death process. This model includes only birth and death and can be used to generate analytic expressions for the mean time to extinction.

Another analytic approach involves approximating the stochastic population dynamics with differential equations called “diffusion” equations. Diffusion equations can be analyzed to generate explicit expressions for the mean time to extinction of a population under different sorts of stochasticity; indeed, this approach was used to generate the relationships shown in Fig. 4.

Although analytic approaches have elegance and generality, they are restricted in the processes which can be included and they can be quite difficult to parameterize from field data. They have made significant contributions to our general appreciation of the role of stochasticity in extinction and underpin a general theory of extinction, but they are rarely used in practice.

3. Time Series Analysis Methods

A relatively new method for exploring extinction risks utilizes analysis of time series of population abundance. During the past decade, a considerable amount of time series data on a wide variety of organisms have been obtained. If these time series of population size are sufficiently long (at least 15 years), then it is possible to statistically “fit” different models of stochastic population growth (such as a Ricker model or logistic model) to these time series. Such models are necessarily simple but have the advantage that they do not require any life history data—only a series of counts. Given the best fit model, it is a straightforward task to simulate the dynamics into the future and calculate extinction risks. Although this approach will prove increasingly useful for making extinction risk predictions, and hence helping with tasks such as prioritizing species, the lack of mechanisms in these models limits their application for management purposes.

4. Numerical Methods

The approach of using a Markov chain (or stochastic population projection matrix) can be expanded to include other forms of stochasticity besides simply demographic. If transitions from any population state to any other population state can be estimated as a function of life history attributes and biotic and abiotic influences, then increasing these matrices to increasingly higher powers allows us to project the probability distribution of possible population sizes into the future. This method provides the foundation of the RAMAS extinction risk software.

The numerical method has the advantage of some mathematical elegance, and it is certainly powerful when coupled to front-end software such as that available in RAMAS. However, most people who build models from scratch find the mathematics involved too complex and tend to use the Monte Carlo simulation method.

5. Monte Carlo Simulation Models

The most common way of doing PVA is to use a Monte Carlo simulation model. These simulation models are either written from scratch or the user uses one of the available packages (Lindenmayer et al., 1995). The most frequently used packages are VORTEX and RAMAS (several versions/types); however, others are available (e.g., ALEX and GAPPS). The basic structures of all
these models are fairly consistent and are best visualized in a flow chart (Fig. 7). However, the details of how each component is modeled and which processes can be included in the models vary considerably from model to model and species to species. Each has its strengths and weaknesses. For detailed application to a particular species it is advisable, although expensive and time-consuming, to construct models specific to that species that focus on relevant processes and management options.

For example, several different kinds of PVA models were built to determine efficient and adequate conservation strategies for the northern spotted owl (*Strix occidentalis caurina*). These ranged from simple demographic models to very detailed landscape models with habitat dynamics and very specific biological factors such as predation by another owl. The diversity and number of models for this species largely reflect the controversy regarding it: The spotted owl requires large tracts of forested land which are very valuable for timber production. For species that have far less economic importance, there is neither the time nor the money to embark on detailed modeling exercises and so one of the existing packages is used.

Monte Carlo simulation models run on a PC have the advantage that they can incorporate many different parameters and processes; their complexity is limited only by the imaginations of the biologists and managers. However, the inclusion of detail and complexity can give the illusion of accuracy, whereas some consider that adding more poorly understood processes will make the results less accurate. The factors driving the extinction estimates from complex models are often very difficult to isolate.

### B. Sensitivity Analysis of PVA Models

Sensitivity analysis is an important component of modeling because one can use it to systematically investigate the complex interactions of a model. Sensitivity is usually measured by varying a parameter by a small amount from its estimated value. The resulting change in the state variable (e.g., the risk of extinction) provides an index of the sensitivity of the model to that parameter. Sensitivity analysis provides practical information for model builders and users by highlighting parameters that have the greatest influence on the results of the model. It can highlight model parameters that should be most accurately measured so as to maximize the precision of the model, give a general indication of the reliability of the model predictions, and highlight parameters and interactions that have the largest influence on the population to help determine effective management strategies.

The sensitivity of deterministic matrix population models can be determined analytically by eigen analysis. Extension of these techniques to PVA is not appropriate for most PVA models except for matrix models for at least three reasons: (i) PVA models are often complex, so obtaining solutions analytically is difficult if not impossible; (ii) in PVA, the result of interest is the risk of population decline and not the deterministic growth rate; and (iii) interactions between variables are largely ignored.

Any method of sensitivity analysis should be clearly defined, interactions between parameters should be distinguishable from single parameter effects, and the method should account for variability associated with parameter estimates. The simplest approach to sensitivity analysis of PVA models is to vary the model parameters in turn and investigate the effect on the risk of
population decline. Alternatively, such a sensitivity analysis can determine whether the relative efficacy of different management decisions changes with different parameter values. PVA models may be complex with numerous parameters, particularly when individuals are modeled, and understanding the relative importance of different parameters and interactions between parameters may be difficult. For example, if there are 10 parameters and three levels for each parameter are to be investigated, then 21 different parameter combinations would be necessary to assess each parameter independently. However, $3^{10} = 59,049$ different combinations are required to test all possible interactions.

When the risk of population decline is the important state variable, logistic regression may be useful for summarizing the effects of different parameters and interactions. Data for the regression are generated by using numerous parameter combinations. For each parameter combination, the PVA model is simulated to obtain a limited number of predictions of the incidence of decline. The regression analysis uses the model parameters as explanatory variables and the incidence of decline as the dependent variable. The regression equation provides a simple expression to approximate how the probability of decline is influenced by the model parameters.

C. Problematic Issues in PVA Models

Although all models of populations are necessarily false because they cannot include all the factors and processes that influence population dynamics, we believe that some processes and factors in PVA modeling are important but are difficult to incorporate.

1. Genetics in PVA Models

Early work on PVA paid considerable attention to the role of genetics, although the significance of genetics in short-term threatened species management has declined in the past decade. Two genetic issues are of particular concern: inbreeding depression and loss of heterozygosity. Inbreeding depression is of particular concern for very small populations, especially captive populations. Estimating the likelihood and effect of inbreeding depression on vital rates is problematic. The loss of genetic heterozygosity is a long-term concern. Populations with an effective size greater than 50 are generally thought to experience low losses of heterozygosity. VORTEX is a generally available package that allows the user to include genetic factors, hence its popularity with zoos and reintroduction programs.

Genetic data can be difficult and expensive to obtain and one needs to consider how the data will influence one's decision making. A useful role of molecular genetic data in PVA is to provide estimates and plausible bounds for demographic attributes that are otherwise impossible to obtain (such as social structure and dispersal rates).

2. Density Dependence in PVA Models

Density dependence at both low and high densities is important for PVA. At low densities the possible consequences of positive (or negative) density dependence are critical to the ability of a species to recover. For populations with limited habitat, the way in which vital rates change as populations approach their carrying capacity, and whether the carrying capacity fluctuates greatly from year to year, is known to influence estimates of viability. Unfortunately, information on density-dependent processes for most threatened species is extremely limited. Different PVA models use different forms of density dependence, from logistic growth and a population “cap” to no density dependence, whereas other models allow the user to choose the appropriate form.

3. Space in PVA models: Metapopulation Models and Risk Spreading

Space is an essential component of any PVA model for which the intention is to make statements about what management actions should take place at which particular geographic locations. For example, in determining habitat reconstruction plans for threatened species (linking corridors and new habitat patches), it is generally essential to represent space explicitly in the models.

Space can be incorporated into a PVA model either implicitly or explicitly. The implicit use of space is typical of many metapopulation models in which the interest is in the number of patches that are occupied, with no particular regard to exactly which patches are occupied. These models can be used in a general sense to report on the values of different reserve sizes or corridors. For detailed management for which the question of exactly where a corridor needs to be located, or which patch needs to be protected, might need to be answered, we normally require models in which real landscapes are explicitly represented.

4. Habitat Dynamics

Early PVA models, and almost all of the analytic approaches, assume a static habitat. Given that habitat loss is one of the primary causes of species extinction,
the introduction of habitat dynamics in PVA models is of increasing interest. Even in the absence of habitat loss, most habitats are dynamics and some recent PVAs have focused as much on the dynamics of the landscape as on the dynamics of the population. This reflects an increasing interest in population ecology—not just in the role of space but also in the role of a changing habitat mosaic in population dynamics.

Combining habitat dynamics with spatial complexity has been facilitated by faster computers and the integration with geographic information system (GIS) software. Future wildlife managers will certainly have access to GIS systems that enable them to interactively explore the consequences of different habitat management actions on extinction probabilities. These complex new-age GIS/PVA tools will include the tools inherent in GIS systems, habitat suitability models, management models, stochastic catastrophes, and detailed population dynamics. To some this will be a powerful and compelling tool and to others a dangerous cascade of complex untested models interacting in ways that the user will never understand.

D. Other Methods of Assessing Viability

There are many measures of viability and many ways of assessing viability—from a wild guess to a complex computer model, and from a simple analytic model to extrapolation from empirical data. So far, we have discussed more traditional model-based methods. Here, we briefly mention two other ways of assessing the viability of a population

1. Historical Data

The occurrence of populations in isolated habitat, especially on oceanic islands, gives us an idea of how large patches of habitat need to be for the long-term persistence of certain species. For example, in the Caribbean reptiles of certain species rarely occur on islands less than critical sizes, but they are consistently present on larger islands. This sort of island data can help us set MVHAs for particular taxa. As is known from island biogeography theory, islands are not isolated and their species complement is a consequence of local extinction and colonization processes. In this context, the presence of a species on an island may not reflect its long-term persistence, only its recent arrival. However, where species have persisted on islands long enough to show variation from mainland populations, we can assume they have persisted for long enough to be termed viable. Although this general approach to PVA has been fruitful, the species that occur on these islands are usually not the same species for which we are trying to design reserve systems on continents. The question arises, how relevant is it to use data on MVHAs from one species for another species?

2. Extrapolation from Similar Species

Although we are unlikely to have data on the viability of, or MVHAs for, threatened species, it is possible that such information on similar, more common species can be used to assess viability. For example, we may know much about the fate of many small populations of a common species and use this information to determine the MVHA of a similar threatened species.

Expanding this line of thinking, some workers have begun to develop rules of thumb that relate extinction probability of MVHA to key life history attributes of a species, such as body size or home range size. We do know enough to say that predators can persist at much lower densities than herbivores of a certain size, and that larger animals and plants appear to be able to persist for longer periods in small populations than smaller animals or plants. These fundamental patterns may enable us to draw broad conclusions about the viability of species and strategies for their protection with limited data and no predictive modeling. This crude approach may only be appropriate when data, time, and money are limited.

V. TESTING PVA

Uncertainty about parameter estimates and the structure of PVA models means that their predictions may be inaccurate. As with any model, PVA is an imperfect description of reality, and model testing with real field data helps to identify the limits of its accuracy. The primary prediction of many PVA models is the risk of extinction of an endangered species over several decades or centuries. In such circumstances, we do not have the luxury to test the models by monitoring numerous populations until they become extinct. A sufficient number of populations rarely exist, and endangerment demands immediate management decisions.

Instead, we can test PVA models by predicting population attributes prior to extinction. The kinds of predictions that can be tested include the annual population growth rate, the population size in a series of years, the local population sizes in patches, and the probability of local extinction of patches. It is not sufficient to test the models by comparing mean predictions with
observations because this ignores variability, which of-
ten is a critical factor influencing the risk of population
decline. Tests of PVA models should investigate the
distribution of the predictions—not only the average
predictions but also factors such as the variance and
predicted extremes such as local extinction. Replicates
are necessary for such tests and may be obtained from
annual observations over several years or from observa-
tions from several different locations. These tests rely
on appropriate data obtained from the field, which may
require a substantial commitment to field surveys. As
noted previously, we may need to test PVA models on
more common species and then, by analogy, assume
that they will work for threatened species. This requires
a certain leap of faith.

Correspondence between the predictions of the
model and observed data does not imply that the model
is correct. Tests of PVAs may not be sufficiently powerful
to identify the errors, or multiple compensating errors
may exist. Regardless of how many times a PVA model
has been tested, new data or a new test will expose the
model to a chance of being proved wrong. This is similar
to the scientific process of hypothesis testing.

VI. CRITICISMS OF PVA: WHAT MAKES
A GOOD PVA?

There is a healthy skepticism about the value of PVA.
Although almost all authors acknowledge that PVA is
valuable and a cornerstone of the science of conserva-
tion biology, PVA, like any tool, has been abused.
The following are major criticisms of PVA:

1. Not enough data
2. Black box model that no one understands
3. Sensitivity to parameters
4. Missing processes; not enough ecology

Because of the "black box" nature of PVA, there is
concern that PVA should be more explicit and detailed
in their reporting of how the model works and exactly
how the parameters were chosen. This is just part of
good science—that is, whatever we do should be repeat-
able and transparent. Despite the diversity of PVA mod-
els, there is consensus on what constitutes a good PVA.
Attributes of a good PVA include

1. A description of why the PVA is being done (a mo-
tivation)
2. A fundamental understanding of the species' ecol-
yogy, including what constitutes suitable habitat and
the ability of the species to disperse between patches of
habitat
3. Detailed demographic data and the way in which
demographic rates change with time and habitat: Care
needs to be taken in describing exactly what parameters
in the model do and exactly how they were estimated
from data
4. An understanding of the response of the species to
threats
5. An assessment of the current state of the popula-
tion, acknowledging inaccuracies in the estimate
6. A statement about how the PVA is likely to help
managers make decisions
7. A description of constraints on decision variables
and management options
8. An understanding of the environmental distur-
bances that directly threaten (or indirectly threaten
through habitat alteration) a species
9. A detailed description of the model and data so
that the PVA process can be repeated and extended
10. An honest assessment of uncertainties in the
data, processes that have been ignored, and possible
flaws in model construction and parameter estimation
11. A sensitivity analysis.

Acknowledgments

This work was conducted as part of the Biological Diversity Working
Group supported by the National Center for Ecological Synthesis and
Analysis, a center funded by NSF (Grant No. DEB-94-21535), the Uni-
versity of California at Santa Barbara, and the State of California. Work
by all three authors was also supported by an ARC SPIRT grant to HPP
and DBL, partly funded by Environment Australia. We thank our nu-
merous colleagues who have worked on PVA with us, especially Mark
Burgman, Sandy Andelman, Bob Lacy, Drew Tyre, and Ian Ball.

See Also the Following Articles

CARRYING CAPACITY, CONCEPT OF • FRAMEWORK FOR
ASSESSMENT AND MONITORING OF BIODIVERSITY •
MEASUREMENT AND ANALYSIS OF BIODIVERSITY •
POPULATION DENSITY • POPULATION DYNAMICS
Bibliography


POVERTY AND BIODIVERSITY

Madhav Gadgil
Indian Institute of Science

GLOSSARY

artificial ecosystems Ecosystems dominated by human artifacts and human induced material and energy flows.

biosphere people People enjoying access to resources garnered from the entire biosphere and made available through markets.

biosphere reserve An international conservation program aiming to protect representative ecosystems throughout the world in ways compatible with development efforts.

ecological footprints Spatial coverage of ecological impacts consequent on the activities of a given group of people.

ecological refugees People that have lost access to their traditional base of natural resources yet have very limited access to resources through markets.

ecosystem people People meeting the bulk of their resource requirements from a limited area near their habitation through gathering or low input agriculture and animal husbandry.

resource catchment Locality from which the resources consumed by a group of people are derived.

WEALTH—AND ITS OBVERSE, POVERTY—is today reckoned in terms of the capacity of a person or groups of persons to obtain goods and services through exchanges in the market-place. The bulk of such goods and services are partly or wholly products of intensively managed or artificial ecosystems; crop fields, plantations, shrimp ponds, mechanical fishing fleets, factories, towns, and cities. Most of these intensively managed or artificial ecosystems tend to harbor low levels of biodiversity. Rich people are then those with extensive access to produce of managed or artificial ecosystems supporting low levels of biodiversity; poor people have very limited access to such produce. Ecosystems harboring high levels of biodiversity are, on the other hand, natural or semi-natural ecosystems with low levels of human demands for their produce. Rich people rarely live in close vicinity of such ecosystems, though they may visit them for recreational purposes. Groups of poor people may, on the other hand, permanently live in their vicinity. In some cases, the poor control and serve as the stewards of such biodiversity rich ecosystems. More often though many such biodiversity rich ecosystems are subjected to overexploitation, primarily
to meet the large resource demands of the rich, often living far away, with the local poor serving as agents of the resultant destruction of biodiversity. Large numbers of poor also live permanently in the vicinity of biodiversity poor, intensively managed, or artificial ecosystems. There is then no simple relationship between poverty and biodiversity, the equations vary greatly from context to context.

I. OF ECOSYSTEMS AND PEOPLE

To appreciate these complexities, one must tease out the many contexts. These may be viewed along two axes: one of the extent to which ecosystems have been transformed through human interventions, and the second of the manner in which people relate to the world of nature, to the world of manufactured artifacts, and to each other. The terrestrial and aquatic ecosystems may be classified into four major categories: (a) natural ecosystems, which are subject to very low levels of human demands because of their inaccessibility; (b) natural and semi-natural ecosystems, including low-input agricultural systems subject to higher levels of human demands; (c) ecosystems managed intensively for biological production; and (d) largely artificial ecosystems dedicated to industrial production and organized services. Building on Dasmann's (1988) pioneering work, we may, for our purpose, classify people into four major, largely inclusive categories; (a) autonomous ecosystem people, (b) subjugated ecosystem people, (c) biosphere people, and (d) ecological refugees. Ecosystem people have limited access to sources of energy other than human and livestock muscle power and to the more sophisticated artifacts. They gather or produce most of the resources they consume from their immediate surroundings, from the forest, scrub, rivers, or seas and from low-input cultivation. In some of the more inaccessible corners of the world, the ecosystem people are still autonomous. However, over most of the earth they have been subjugated by the biosphere people and have very limited control over their own resource base of natural and semi-natural ecosystems. They gather and produce little that can fetch value in markets and therefore have very limited access to produce of intensively managed and artificial ecosystems. The biosphere people owe their dominant position to their extensive control over artifacts and additional sources of energy. They engage in energy-intensive agriculture, animal husbandry, aquaculture, or in organized services or industrial production and generate much of value in the markets. They have large ecological footprints thanks to their substantial purchasing power; their resource catchments are vast encompassing all of the biosphere, bringing to them goods and services from all over the earth. This confers on them the ability to take over resources in demand by the ecosystem people, catalyzing transformation of natural and semi-natural ecosystems into those managed intensively to meet their own demands or, in some special cases, conserved as natural or semi-natural ecosystems for recreational purposes. In the process, they often deprive ecosystem people of access to their traditional resources; this may convert them into ecological refugees. Ecological refugees are then people with attenuated access to resources of natural and semi-natural ecosystems, with little purchasing power to access products of intensively managed and artificial ecosystems. They often end up constituting the unorganized labor force in tracts of intensive agriculture and urban settlements (Gadgil, 1993; Gadgil and Guha, 1995).

II. AUTONOMOUS PEOPLE

I propose to explore the relation between poverty—or wealth—and biodiversity in this framework of different categories of ecosystems and of people. Table I summarizes the relationship and notes the examples that are discussed later at some length. All of these derive from my own fieldwork in India; I use them purely because I know them well. However, they do represent broader patterns encountered in other parts of the world. Not all 16 cells of the matrix are occupied; moreover our focus is on poverty, and we shall therefore not elaborate on the relationship of the biosphere people with biodiversity, except as they impact on the ecosystem people or ecological refugees. There are relatively few examples of truly autonomous ecosystem people, people fully in control of largely natural ecosystems with very light human demands, and in consequence with high levels of biodiversity. Inhabitants of Sentinelese island (11°30' N lat. 92°15' E long.) in the Andaman-Nicobar chain in the Bay of Bengal provide one such example. These relatively inaccessible islands harbor tropical rain forest biota with high levels of endemism (i.e., of species restricted to this chain of islands). They were inhabited by a number of hunter-gatherer tribal groups, without knowledge of metal tools, with the exception of Shompens of Nicobars who had a more advanced fishing economy. British colonized the islands in mid-nineteenth century, an attempt that was strongly resisted by the indigenous people. By 1870s the resistance was overcome and as a consequence many of the tribal populations either drastically declined or were extermi-
TABLE I

Variety of People-Ecosystem Contexts, with Specific Examples

<table>
<thead>
<tr>
<th>Ecosystems</th>
<th>Autonomous ecosystem people</th>
<th>Subjugated ecosystem people</th>
<th>Biosphere people</th>
<th>Ecological Refuge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inaccessible natural/seminal natural ecosystems</td>
<td>Poor in modern economic sense, with access to high levels of biodiversity, e.g., Sentinelese Islanders</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Accessible natural/seminal natural ecosystems</td>
<td>Poor in modern economic sense, serve as agents for destruction of biodiversity, e.g., Panshet peasants, Gangtes of Maldharis of Gir Manipur</td>
<td>Visitors for recreational/commercial use</td>
<td>—</td>
<td>Poor in modern economic sense, serve as agents of destruction of biodiversity, e.g., Pancher peasants, Maldharis of Gir</td>
</tr>
<tr>
<td>Ecosystems managed intensively for biological production</td>
<td>—</td>
<td>—</td>
<td>Well off in modern economic sense, promote low biodiversity production systems</td>
<td>Poor in modern economic sense, alienated from biodiversity, e.g., Jharkhand tribals</td>
</tr>
<tr>
<td>Artificial ecosystems dedicated to industrial production/organized services</td>
<td>—</td>
<td>—</td>
<td>Well off in modern economic sense, promote high biodiversity systems for recreational purposes</td>
<td>Poor in modern economic sense, alienated from biodiversity, e.g., Jharkhand tribals</td>
</tr>
</tbody>
</table>

During British times the islands primarily served as a convict colony, with many of the released convicts settling down to agriculture (Superintendent of Government Printing, 1909). After independence in 1947 the islands were used to create agriculture based settlements of many people displaced from then East Pakistan (now Bangladesh), as well as to supply forest resources for rapidly growing forest based industries, especially plywood (Saldanha, 1989). As these were progressively overused and exhausted leading to substantial erosion of biodiversity, pressure has built up to overcome the resistance of the tribal groups still holding out to commercial forest exploitation. Two of the tribal groups, however, do continue to retain hold over their territories; these are Jarwas and Sentinelese. Jarwas live on the larger South Andaman and Baratung islands (12°0’ N–12°20’ N lat. 92°45’ E–92°55’ E long), about two-thirds of which has been colonized by immigrants and subjected to forest exploitation. Their territory is therefore easily accessible; it is, however, stoutly defended by Jarwas with their bows and arrows. There are ongoing attempts to overcome this with the aid of a major road passing through their territory. This Jarwa territory remains much richer in biodiversity than the rest of the islands; Jarwas are, however, poorer in a modern economic sense with access to very few and simple artifacts. The Sentinelese are also very poor in a modern economic sense. Presumably Jarwas and Sentinelese do not themselves have concepts of wealth and poverty or biodiversity.
The vast majority of the ecosystem people of the world are, however, no longer isolated in this fashion. In no case do they seem to have voluntarily sought to integrate with the larger society dominated by the biosphere people. Rather the biosphere people have sought to integrate them in order to access resources of their territories, as well as to utilize their labor. The process of such integration tends to proceed through several stages involving political subjugation, followed by economic subjugation. Such subjugation is accompanied by a reduction in levels of access of the people to the biodiversity resources of their own localities, though simultaneously they may have enhanced access to other resources through market channels, thereby implying a reduction in the levels of their poverty in modern economic terms. The process is also generally accompanied by an erosion of biodiversity driven by overexploitation of natural resources to meet outside demands with the subjugated ecosystem people often serving as agents of such erosion.

These processes are the dominant form of interplay of poverty and biodiversity in the world today, and we propose to illustrate them through a series of case studies. The most detailed case to be presented would involve Gangtes, one of the Kuki tribes of Churchandpur district (23°4’–23°59’ N lat. and 93°–94°57’ E long.) of the northeastern state of Manipur on the India-Myanmar-China border, since this illustrates the complete sequence from autonomy through political and then economic subjugation (Gadgil, Hemam, and Reddy, 1997; Hemam, 1997). Like many other Kuki and Naga groups, Gangtes continued head hunting well into nineteenth century. At that time, there were no roads so that people had to walk for several days to reach markets where they could exchange hand-woven cloth or honey for iron tools or rice. These journeys were hazardous as they could involve passage through territories of alien tribes. Each settlement was in consequence a largely self-sufficient and self-governing entity. Their base of subsistence was shifting cultivation with long fallows of about 15 years. There were no fixed village boundaries and individual social groups presumably shifted around from time to time. The settlements were located on hill-tops; the valleys were considered unsafe and more vulnerable to raids from alien groups. Among Gangtes, hereditary village chiefs determined through primogeniture made all group-level decisions in consultations with a council of elders. They assigned lands for cultivation and annually received about 16 kg of grain from each family in the group as a tribute for use in entertaining any visitors. In the course of cultivation the densest, tallest forest tracts were left alone as requiring too much labor for conversion to fields; the more recent fallows were also avoided. The people worshipped many natural elements including mountain peaks, streams, plants, and animals. They strictly protected patches of sacred groves called gakhal as also other spirit possessed lands called nungens. Also protected were bamboo groves called mavuak from which bamboo may be extracted only for house construction, but the shoots, much reduced as food, were left alone. This preserved a luxuriant vegetation, which led Captain Pemberton to remark in 1835: "I know no spot in India, in which the products of the Forests are more varied and magnificent but their utility is entirely local" (Pemberton, 1835).

The process of political subjugation of Gangtes began in the late nineteenth and early twentieth centuries with the development of incipient links with mainstream India, then ruled by the British. Initially, the British did not so much want access to resources of Churchandpur as of the much richer Myanmar. Hence they wanted to ensure safe movements of British troops in the hill tracts of Manipur bordering Myanmar. A people without fixed, well-defined villages and with traditions of killing aliens coming into their territories were an obvious threat. The British therefore concentrated on fixing village boundaries and assigning land ownership. Given the Gangte system of hereditary village chiefs having a major role in all decisions including assigning land for shifting fields, they decided to confer on the chiefs all ownership over land, converting others into tenant farmers from a legal perspective. This move then formally introduced wealth and poverty to the Gangte society. However, during the British reign this made little operational difference in most areas, with no acceptance of private property in land by Gangtes, except in the Churchandpur town area beginning 1930s. Elsewhere Gangte community members continued to pay the village chief a tribute of 16 kg of grain as before. The late nineteenth and early twentieth centuries also
witnessed the gradual spread of Christianity. Christianity rejects the attribution of sacred qualities to elements of nature and, consequently, taboos against killing of certain animals or felling of certain patches of forests like gamkhal. Conversion to Christianity therefore began to slowly erode the traditional belief system underlying these conservation practices.

In this phase of political subjugation there was little change in the economy. The road network remained extremely limited and the Gangte communities remained almost totally self-sufficient in terms of resource use. The fallow cycles for shifting cultivation were still long, and substantial areas of forest retained protection in forms of sacred groves. However, by introducing the legal concept of private ownership and a religion that questioned attribution of sacred qualities to nature, this phase did set the stage for the radical changes that followed independence; but unlike on the mainland, this political subjugation had limited impact in the absence of an access to markets and therefore of any attempt at economic subjugation.

IV. MARKET FORCES

The British were prompted by the march of the 'Free India Army' fighting in collaboration with the Japanese into Manipur through Myanmar to strengthen the road network of Manipur on a war footing, laying the foundation for the rapid development of transport, communication, and commodification that began on independence in the 1950s. This was consistent with the policies of economic development and national integration adopted at the time of independence. These development policies also encouraged forest based industries by offering them many resources, including wood, at highly subsidized rates (Gadgil and Guha, 1992, 1995). As the production of these industries grew at a rapid pace, the demands outstripped supplies leading to severe overexploitation. Such overexploitation was facilitated by the fact that there was no social group sufficiently motivated to ensure a more sustainable pattern of resource use with a secure enough control over the resource stocks. Access to markets meant that a community no longer suffered an inevitable shortage of resources if those from their own vicinity were overused and exhausted. Neither would such a difficulty be experienced by an industry drawing resources from a larger spatial scale. Market access therefore reduced the motivation of the various parties concerned, local communities, such as those of Gangtes, as well as other consumers such as plywood industry for restraints on levels of harvests. At the same time conversion to Christianity eroded the conservation ethos of local communities grounded in attribution of sacred qualities to various natural elements. Linkages to the larger national society also diluted the control of any one agency over any given resource base by bringing a variety of actors into play. Thus in the autonomous phase where a person might even be killed while passing through alien territory, there was a clear-cut control by a local community. In the phase of political subjugation this was legally recognized as the control in terms of ownership by the local chieftains. But in the post-independence phase of economic subjugation the state Forest Department has stepped in to claim control over forest lands and forest resources. How this claim is to be reconciled with the claim over land ownership by the Gangte chiefs has not been clarified. This uncertainty has affected the security of resource control by both parties. At the same time, the traditional authority structure of a Gangte village community headed by a hereditary chief with a council of elders has also been affected by the institution of a democratically elected village council. There are continuing contradictions in this system as well since the hereditary chief remains at the head of the elected village council. Under these circumstances, forest resources of Gangte villages have been affected by demands greatly exceeding sustainable yield levels. These demands are no longer local, but reflect the markets not only of remaining parts of the state of Manipur, but the rest of the country as well. Indeed market pulls come from even beyond national borders as Indian plywood was over some periods exported in substantial quantities to the mid-east.

The traditional Gangte society is largely egalitarian. In this phase of political subjugation there was little change in the economy. The road network remained extremely limited and the Gangte communities remained almost totally self-sufficient in terms of resource use. The fallow cycles for shifting cultivation were still long, and substantial areas of forest retained protection in forms of sacred groves. However, by introducing the legal concept of private ownership and a religion that questioned attribution of sacred qualities to nature, this phase did set the stage for the radical changes that followed independence; but unlike on the mainland, this political subjugation had limited impact in the absence of an access to markets and therefore of any attempt at economic subjugation.

IV. MARKET FORCES

The British were prompted by the march of the 'Free India Army' fighting in collaboration with the Japanese into Manipur through Myanmar to strengthen the road network of Manipur on a war footing, laying the foundation for the rapid development of transport, communication, and commodification that began on independence in the 1950s. This was consistent with the policies of economic development and national integration adopted at the time of independence. These development policies also encouraged forest based industries by offering them many resources, including wood, at highly subsidized rates (Gadgil and Guha, 1992, 1995). As the production of these industries grew at a rapid pace, the demands outstripped supplies leading to severe overexploitation. Such overexploitation was facilitated by the fact that there was no social group sufficiently motivated to ensure a more sustainable pattern of resource use with a secure enough control over the resource stocks. Access to markets meant that a community no longer suffered an inevitable shortage of resources if those from their own vicinity were overused and exhausted. Neither would such a difficulty be experienced by an industry drawing resources from a larger spatial scale. Market access therefore reduced the motivation of the various parties concerned, local communities, such as those of Gangtes, as well as other consumers such as plywood industry for restraints on levels of harvests. At the same time conversion to Christianity eroded the conservation ethos of local communities grounded in attribution of sacred qualities to various natural elements. Linkages to the larger national society also diluted the control of any one agency over any given resource base by bringing a variety of actors into play. Thus in the autonomous phase where a person might even be killed while passing through alien territory, there was a clear-cut control by a local community. In the phase of political subjugation this was legally recognized as the control in terms of ownership by the local chieftains. But in the post-independence phase of economic subjugation the state Forest Department has stepped in to claim control over forest lands and forest resources. How this claim is to be reconciled with the claim over land ownership by the Gangte chiefs has not been clarified. This uncertainty has affected the security of resource control by both parties. At the same time, the traditional authority structure of a Gangte village community headed by a hereditary chief with a council of elders has also been affected by the institution of a democratically elected village council. There are continuing contradictions in this system as well since the hereditary chief remains at the head of the elected village council. Under these circumstances, forest resources of Gangte villages have been affected by demands greatly exceeding sustainable yield levels. These demands are no longer local, but reflect the markets not only of remaining parts of the state of Manipur, but the rest of the country as well. Indeed market pulls come from even beyond national borders as Indian plywood was over some periods exported in substantial quantities to the mid-east.

The traditional Gangte society is largely egalitarian. In this phase of political subjugation there was little change in the economy. The road network remained extremely limited and the Gangte communities remained almost totally self-sufficient in terms of resource use. The fallow cycles for shifting cultivation were still long, and substantial areas of forest retained protection in forms of sacred groves. However, by introducing the legal concept of private ownership and a religion that questioned attribution of sacred qualities to nature, this phase did set the stage for the radical changes that followed independence; but unlike on the mainland, this political subjugation had limited impact in the absence of an access to markets and therefore of any attempt at economic subjugation.
### Table II

Resource Ownership, Sharing, and Utilization Patterns among Different Sections of Gangte Tribe at Different Levels of Accessibility and Modernization

<table>
<thead>
<tr>
<th>Shifting cultivation area</th>
<th>Settled agriculture area</th>
<th>Town area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low accessible</td>
<td>Family accessible</td>
<td></td>
</tr>
<tr>
<td>Ownership</td>
<td>Village chief</td>
<td>Village chief</td>
</tr>
<tr>
<td>Commercialization of forest and current forest condition</td>
<td>Little commercial exploitation and loss degradation</td>
<td>Under heavy commercial exploitation and degraded condition</td>
</tr>
<tr>
<td>Onset of commercial forest exploitation</td>
<td>Still unmanaged, except for domestic use</td>
<td>Early 1970s</td>
</tr>
<tr>
<td>Commercialization of forest and current forest exploitation</td>
<td>Little exploited, good forest cover</td>
<td>Heavily exploited, degraded forest</td>
</tr>
<tr>
<td>Collection of timber</td>
<td>Free access</td>
<td>Pay royalty to the chief</td>
</tr>
<tr>
<td>Collection of fuelwood</td>
<td>Free access</td>
<td>Pay royalty to the chief</td>
</tr>
<tr>
<td>Maintenance of VFR and Bamboo reserve</td>
<td>Still intact</td>
<td>No more</td>
</tr>
<tr>
<td>Traditional rights of resource sharing and use</td>
<td>Still followed</td>
<td>All have disappeared except for free use of shifting fields and fuelwood collection for domestic use</td>
</tr>
<tr>
<td>Collection from forest</td>
<td>Fuelwood, solid vegetables, timber, and other NTFP</td>
<td>Fuehwood, wild vegetables, timber, and other NTFP</td>
</tr>
<tr>
<td>Land use</td>
<td>Shifting cultivation</td>
<td>Shifting cultivation, terraced cultivation</td>
</tr>
<tr>
<td>Commercial crop plantation</td>
<td>Chilli and other vegetables</td>
<td>Banana, chilli and vegetables</td>
</tr>
<tr>
<td>Energy use</td>
<td>Human muscle power, fuel-wood, and little use of fossil fuel</td>
<td>Human muscle power, fuel-wood, and little use of fossil fuel</td>
</tr>
<tr>
<td>V. LAND USE CHANGES</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This process of over-exploitation of forest resources has also shaped patterns of use of land. Since the forest resources bring in highest levels of net profits closest to the market town of Churchandpur, all tree stocks have been exhausted in this area. In response, settled cultivation on terraced fields with very low levels of crop diversity such as monocultures of pineapple has replaced the much more biologically diverse shifting cultivation. In more remote areas too, commercial monoculture plantations have been taken up of species like Agor (Aquilaria aquausta), the highly valuable wood that was the first to be exhausted even from areas far from market.

The traditional land-use pattern of Gangtes included leaving aside substantial areas as gamkhalas, or so-called village forest reserves (VRF), as well as bamboo reserves for collection of fuelwood, while further afield they have free access to even more valuable timber.

---

1 Village Forest Reserve (VRF).
2 Nontimber Forest Produce (NTFP).
and other spirit lands. These practices have been completely abolished in villages close to market towns where land and its produce have acquired high commercial value. In some of the more remote villages the practice of protection of a VFR encircling the settlement has been revived following recognition of its value as a firebreak preventing the spread of fire to the settlement during the slash and burn operations (Gadgil, Hemam, and Reddy, 1997). The traditions of protection of such forest patches still continue in other more remote areas.

VI. VALUE APPROPRIATION

These cash requirements are met from harvests of forest produce, either through sale or through wages earned as laborers in tree-felling operations taken up by contractors. With the assertion of land ownership and demand for royalty by chiefs, part of the profits go to them. However, the chiefs are unfamiliar with the skills of development of infrastructure such as roads, organizing transport, and marketing of forest produce. As a result outside contractors carry out most of the timber harvesting operations, retaining a bulk of the profits for themselves. Thus in the early 1990s in Nalwan village the contractor paid Rs. 50,000 (US $ 1600) for rights to harvest timber from an area of over 1000 ha. A conservative estimate of the value of timber thus accessed is Rs. 60 million given a standing stock of 100 tonnes/ha and price of timber at Rs. 600/tonne. The chief was then paid less than 1.0% of the timber value. Of course, the contractor would have to invest in labor charges, making roads, and transport. Nevertheless, it is clear that bulk of the cash income flows to the contractor, a part to the chief, and a very small fraction to members of Gangte community. The Gangte use this cash to meet their routine requirements. None of it is saved or invested productively in ventures such as developing horticultural plantations to replace shifting cultivation fields. In the long run then the whole process of forest resource use taking place today is leading to exhaustion of these resources with little gains for the Gangte community members.

Today the Gangtes have greater access than they had earlier to goods on the markets and are in that sense richer, though this access is still extremely limited and by all criteria they are quite poor. However, at an earlier time when they were poorer in terms of market access, they had more secure and more equitable rights over land and substantial access to biodiversity and other natural resources; this has drastically reduced. The erosion of biodiversity this transformation has accompanied has in part been driven by the desire of Gangtes to acquire cash; however, the real driving force behind the loss of biodiversity lies in the demand for forest produce by the biosphere people. Much of the wood of Manipur has been used for manufacture of plywood, and all the value added in this process has been appropriated by those already rather well off, controlling trade and industry of mainland India. The transformation has also witnessed a loss of the traditions of conservation of biodiversity, such as those of protection of sacred groves, by relatively poor people like Gangtes.

Gangtes thus provide an excellent case study of the different stages and processes involved in the transformation of autonomous societies of ecosystem people with access to high levels of biodiversity, but with little access to manufactured artifacts into societies that have lost control over their base of living resources, often turning into instruments of depletion of these resources. Thus subjugated, the ecosystem people may end up with some, though often insignificant, enhancement in their access to artifacts, but with significant losses in their access to biodiversity.

VII. ECOLOGICAL REFUGEES

Gangtes have come under the influence of the state and the market economy relatively recently. True, most of them have lost all legal claims over land or forest resources, and the claims of Gangte chiefs over large tracts of land are being contested by the Forest Department; nevertheless, in fact, Gangtes still control extensive tracts of forested land. This level of autonomy is unusual on the Indian subcontinent where state authority has
been holding sway for several centuries. One such unusual area is that of Maharashtra Western Ghats, where the local peasants had retained a considerable measure of autonomy by virtue of their active role in the Maratha armies of seventeenth and eighteenth centuries. The British were reluctant to offend them on the defeat of Marathas in early nineteenth century. As a consequence they were permitted to continue their shifting cultivation, which was banned over much of the remainder of the subcontinent. Individual peasants, rather than any chiefs were assigned land ownership among the Marathas. So while poor in modern economic terms because their hilly terrain permitted only very restricted access to the markets, these peasants lived in tracts rich in forests and biodiversity with full rights over the resources till the time of independence. The tract was also rich in water resources with annual precipitation in excess of 3000 mm.

On independence the Indian nation state geared itself to tap these rich water resources to promote economic growth with the help of irrigation and hydroelectric projects. One such irrigation project was launched in the Panshet valley to the west of Pune, the old capital of Marathas (18°30′ N lat. –73°40′ E long.). The Indian state has assumed extensive powers of acquisition of land for public purposes and all the more fertile low lying paddy land of the Panshet valley was taken over at low rates of compensation. The peasants still retained ownership over their hill lands above the submersion level. This had been traditionally under long fallow shifting cultivation during which mango and Terminalia chebula trees, both producing fruit of considerable value, were left intact. However, as roads reached Panshet valley to facilitate the dam construction work, these trees caught attention of charcoal merchants. These merchants worked in collusion with the government officials in charge of land acquisition to persuade the peasants to sell the trees at exceedingly low prices with the assurance that the peasants would soon be resettled with alternative farmland far from Panshet valley. Most peasants sold the trees, and there was a massive spurt of deforestation in 1950s during the construction of the dam, the trees being primarily used to produce wood charcoal for consumption in Pune. The peasants, however, were never provided alternative agricultural land elsewhere and have continued staying on hill slopes above the dam for the past 40 years, while some of them have migrated to Pune, primarily to work as unskilled labor. The entire catchment has been devastated by the initial deforestation in 1950s followed by continual, ever more intensive cultivation of hill slopes by the impoverished peasants. Today the peasants are poor in every sense and live in an environment depleted of biodiversity. They have served as agents of this destruction of biodiversity that has only benefited the charcoal merchants of Pune and provided citizens of Pune with a relatively inexpensive fuel for a period of few years. Thus subjugated, these ecosystem people may be said to have been turned into ecological refugees (Gadgil, 1979).

Our next case study is that of the tribal populations of the so-called Jharkhand tracts of the states of Bihar and West Bengal (22°0′–24°0′ N lat. 84°0′–87°40′ E long.). This region adjoins the old, thickly settled Gangetic plains that were the nucleus of the 2000 year old Magadha empire extending over much of the Indian subcontinent. Emperor Ashoka of Magadha is famous for his conquest of and subjugation of the tribals in the extensive hilly tracts that border the Gangetic plains (Gadgil and Guha, 1992). This part of the country also fell rather early to British colonialists, who quickly established a hold over the region during the second half of the eighteenth century, far more firm than that over the Maratha kingdom discussed earlier. The forest of Jharkhand are rich in sal (Shorea robusta), an important timber for the production of railway sleepers. This was a resource much valued by the British who therefore acted firmly to ban all shifting cultivation by the tribal people in early nineteenth century. The forests were also taken over either as a property of the British government or that of private landlords. Where the tribal villages were a part of the government-controlled forest lands, the tribals were totally deprived of any rights to forest resources. Unlike with Gangtes, the landlords too did not come from among the tribals themselves. They came from other castes at a higher level in the hierarchy. These landlords then exploited the lower caste or tribal tenants on their lands quite ruthlessly. The “Jharkhand” tribals were therefore subjected to extreme political as well as economic subjugation as early as the first half of the nineteenth century. Their only earnings came from work as poorly paid tenant farmers or as forest laborers along with sale of some minor forest produce at extremely low rates. They then have a long history as poor people serving as agents of destruction of biodiversity.

These Jharkhand tribes are among the most striking example of ecological refugees of the country. They have served as the most mobile source of very inexpensive, unskilled labor in many contexts. One of the major economic enterprises under the British rule was the development of tea estates replacing the rain forest of Brahmaputra valley in northeastern India. This involved wholesale destruction of biodiversity. While the planta-
tion owners were British, the laborers responsible for actual destruction of biodiversity were mostly Jharkhand tribes working under conditions that have been described as being close to slavery (Gadgil, 1942).

Elsewhere, these tribes were resettled in so-called forest villages. The primary focus of the plantations they worked on was to replace the natural sal-dominated, fairly diverse humid forest with monocultures of teak (Tectona grandis). Teak is excellent timber resistant to termite and fungal attacks, highly valued for ship building and gun carriages in nineteenth and for furniture and house construction in twentieth century. But teak is hard wood, little used traditionally; it also does not provide any other product of local utility. On the other hand, sal leaves are used to make plates and sal seed has value as an oil seed. Many associates of sal, such as mahua (Madhuca indica) and tendu (Diospyros melanoxylon) have great local utility. So replacement of natural sal forest by teak plantations deprives local people of access to a diversity of biological resources of value. Uprooted impoverished tribals have served over years as primary agents of such an erosion of biodiversity (Gadgil and Guha, 1995).

Transformations such as these are particularly common today in tropical developing countries, transformations that are actively encouraged by nation states as well as transnational development agencies as development efforts. These development efforts then have adverse impacts both on biodiversity and on the ability of the weaker sections of the population to access biodiversity. Indeed, they also have adverse impacts on the biodiversity conservation practices of the ecosystem people, practices such as protection to sacred plants and animals and sacred sites such as sacred groves and ponds.

VIII. COSTS OF CONSERVATION

Conservation practices of the ecosystem people with their limited resource catchments are organized on limited spatial scales. Thus, the sacred groves have sizes ranging from a fraction of a hectare, to tens, at best hundreds of hectares. The biosphere people have their own conservation practices. In keeping with their larger resource catchments, they tend to protect much larger areas as nature reserves (Gadgil, 1996). Agrarian states had their tradition of hunting preserves of aristocracy, ranging in size from few hundreds to thousands of hectares. Such hunting preserves of nobles are described in Kautilya's Arthashastra, the 2000-year-old Indian manual of statescraft (Kangle, 1969). They are also known from medieval Asia and Europe. The hunting preserves were aimed at protecting major game animals such as lions or antelopes and their habitat for the entertainment of aristocracy. Their main focus was therefore on exclusion of access, especially to hunting in these areas, by all the commoners. In this, they would have tended to accentuate the poverty of people living within and adjacent to the hunting preserves.

The modern-day nature reserves are even larger in spatial scale, in keeping with the even larger resource catchments of modern-day societies of biosphere people. Thus the largest national parks of the day extend over hundreds of thousands of hectares. Many of them, as in India, have their roots in the hunting preserves of the aristocracy. Two notable examples of these are the national parks at Gir, the last stronghold of the lion in India, and at Keoladeo, Ghana, a wetland that shelters enormous nesting colonies of herons, storks, pelicans, and cormorants during monsoons and huge populations of migrating waterfowl in the winter. The focus at both these reserves is on exclusion of resource access by local peasants and herders, leading to major controversies.

At Gir (21°48' N lat. 70°30' E long.), the controversy centers on relocation of traditional colonies of buffalo keepers, the Maldharis outside the national park. Their free-ranging buffaloes were earlier an important prey for the lions, especially the males. On being located outside the quality of grazing for the buffaloes has declined, and despite some reduction in the mortality of buffaloes from lions, the Maldharis are worse off. Within the national park the herbivore populations have increased with reduction of competition for grazing by buffaloes, but the lion population has not gone up. In fact, male lions continue to hunt Maldhari buffaloes who are now outside the limits of the national park. While Maldharis were earlier willing to accept some buffalo kills by lions as a compensation for access to grazing within the national park, they are no longer willing to do so. So they poison carcasses of buffaloes killed by lions; as a consequence the lions suffer (Johns, pers. comm.)

In Keoladeo national park (27°15' N lat.–77°35' E long.), famous for its water birds too, modern conservation attempts have involved exclusion of grazing in the extensive shallow wetlands. When grazing was thus banned in early 1980s without any alternatives being made available, there were protests leading to police firing, with some deaths. But the ban was implemented, forcing local peasants to substantially reduce their livestock holdings, impoverishing them to a significant extent. But this cessation of grazing has also had an ad-
verse impact on wetlands as a water bird habitat. This is because the wetlands have now become choked by an excessive growth of a grass, Paspalum. This particular conservation measure has thus tended to impoverish people as well as the bird life, which was at the center of the conservation concern (Vijayan, 1987).

Jharkhand tribals too have suffered from being excluded in the interest of nature conservation. One of the major recreation areas for the wealthy from Calcutta city is the Betla National Park (23° 40′ N lat. – 84° 40′ E long.) in Daltonganj district of Bihar. The tribals who were earlier settled in may forest villages in what now constitutes this national park are facing serious difficulties. With the constitution of the national park, the forestry operations employing them have been halted. Further restrictions are imposed on their collection of forest produce. Moreover, the tribals are now being asked to move out of the forest villages altogether, without any provisions for alternative livelihoods. So the tribals were earlier instruments of destruction of the forests and the present situation is the result of the same policies.

A notable development accompanying these protests is the program of participatory management of forest resources. Under this program villagers are encouraged to organize village forest committees (VFCs) to promote natural regeneration of forests. The VFCs are given certain authority to control forests, an authority that was earlier monopolized by the Forest Department, and are expected to assume responsibility for forest protection. In return they are given full rights over non-timber forest produce such as sal leaves and oil seeds, as well as a share of any timber that may be harvested. This program, which was initiated in 1973 in the tribal villages of Midnapore district (22° 20′ N lat. – 87° 20′ E long.) of West Bengal, has now spread over many other parts of this tribal belt, as well as elsewhere in the country. The program has by and large been a success. It is an important instance of the poor getting organized to promote ecological restoration, and in the process enhancing their own standard of living. Such ecological restoration also tends to enhance levels of biodiversity in conjunction with production of a variety of forest produce of value to local people (Pollnauerberger and McGean, 1996).

The Jharkhand political struggle also incorporates an issue with significant implications for biodiversity conservation. The tribal religious practices throughout the country included protection of sacred plants, animals, ponds, and forests, as was mentioned earlier in the Gangte case study. The system of sacred groves, called “saranas,” has been an important ingredient of the traditional tribal religion of Jharkhand. These saranas are variable in size—mostly small, about 0.1 ha—but very numerous and serve to protect samples of original natural plant life. These were treated with contempt by Christian missionaries who have a significant influence in this tribal region, as also by proponents of high-caste Hinduism. Now protection of these saranas has emerged as an element of the move to restore tribal self-respect and to empower them (Gokhale et al., 1998). Indirectly then, conservation of indigenous biodiversity is beginning to emerge as a part of the agenda of the poor to pull themselves out of the poverty trap.

UNESCO’s Biosphere Reserve program is another endeavor to combine efforts at conservation of biodiversity with alleviation of poverty. However, it has met with limited success largely because of inadequate understanding of local ecology and people’s needs. Thus, Nandadevi (30° 30′ – 30° 40′ N lat. and 79° 44′ – 79° 58′ E long.) in Western Himalayas is one of India’s eight Biosphere Reserves. In this region, the traditional livelihoods depended on summer grazing in these pastures supplemented by collection of medicinal herbs. This grazing has been banned in the core zone of the biosphere reserve, along with mountaineering expeditions, another important source of cash incomes. As a result, local people feel significantly impoverished. The few ecodesvelopment programs such as fuelwood plantations that have been brought in as part of the Biosphere Reserve activities have failed to create sufficient levels of additional incomes. At the same time local people believe that in the absence of grazing the diversity of medicinal herbs in the pastures are now being replaced by extensive growth of a few species such as Rumex (Negi, 1990).
The philosophy behind biosphere reserves, that of combining conservation with development efforts, also motivates a number of development programs that go by the name of “Integrated Conservation and Development Programs” (ICDP) and the many projects funded by Global Environment Facility (GEF). However, there has been little genuine progress so far in actually realizing these goals. As a review of a GEF program in New Guinea documents, the programs remains ineffective as they continue to be designed by outsiders committed to conservation but with little understanding of what would motivate local people, especially the poor, to participate in conservation efforts (Global Environment Facility, 1998).

There are also attempts to bring to local people financial benefits either through ecotourism or through sustainable harvesting of biodiversity resources. There are limited success stories of combining the two in parts of Zimbabwe where substantial incomes are generated though high-priced hunting licenses for large mammalian game animals, with levels of such harvests kept well within sustainable limits (McNeely, 1995). These incomes are shared with local, relatively poor people. Other attempts at combining poverty alleviation with biodiversity conservation have focused on promoting enterprises such as processing of medicinal plants or wild fruits. Programs in India involving processing of fruit of Phyllanthus emblica, a moist deciduous forest tree species by local Solliga tribals in B. R. hills of Karnataka or processing and marketing of wild mango tree species by local women’s groups in the Kangra district of Himachal Pradesh have met with varying degrees of success (Chopra, 1998). Providing alternative means of livelihoods to communities depending on a threatened biological resource has also been attempted with success in the Indian ocean island of Seychelles where turtle shell artisans were provided financial and technical support to shift to occupations like coconut souvenir carving (Global Environment Facility, 1997).

**X. THE CHALLENGE AHEAD**

Poverty thus relates to biodiversity in many, complex ways. But, by and large, the poor are today instruments of decimation of biodiversity in their attempts to earn livelihoods. There has been little success in addressing this problem though designing programs that would combine conservation and sustainable utilization of biodiversity with alleviation of poverty. This then is a major challenge that must be addressed in years to come.

See Also the Following Articles

**Biodiversity as a commodity** • **Biodiversity-rich countries** • **Deforestation and land clearing** • **Economic growth and the environment** • **Economic value of biodiversity, overview** • **Ethnobiology and ethnotechnology** • **Indigenous peoples, biodiversity and land-use issues** • **Market economy and biodiversity** • **Social and cultural factors**

**Bibliography**


Kangle, R. P. (1993). *The Philosophy behind Biosphere Reserves, that of Combining Conservation with Development Efforts, also Motivates a Number of Development Programs that Go by the Name of “Integrated Conservation and Development Programs” (ICDP) and the Many Projects Funded by Global Environment Facility (GEF) However, There Has Been Little Genuine Progress So Far in Actually Realizing These Goals. As a Review of a GEF Program in New Guinea Documents, the Programs Remains Ineffective as They Continue to Be Designed by Outsiders Committed to Conservation but with Little Understanding of What Would Motivate Local People, Especially the Poor, to Participate in Conservation Efforts (Global Environment Facility, 1998)*.

**See Also the Following Articles**

**Biodiversity as a commodity** • **Biodiversity-rich countries** • **Deforestation and land clearing** • **Economic growth and the environment** • **Economic value of biodiversity, overview** • **Ethnobiology and ethnotechnology** • **Indigenous peoples, biodiversity and land-use issues** • **Market economy and biodiversity** • **Social and cultural factors**

**Bibliography**


Kangle, R. P. (1993). *The Philosophy behind Biosphere Reserves, that of Combining Conservation with Development Efforts, also Motivates a Number of Development Programs that Go by the Name of “Integrated Conservation and Development Programs” (ICDP) and the Many Projects Funded by Global Environment Facility (GEF) However, There Has Been Little Genuine Progress So Far in Actually Realizing These Goals. As a Review of a GEF Program in New Guinea Documents, the Programs Remains Ineffective as They Continue to Be Designed by Outsiders Committed to Conservation but with Little Understanding of What Would Motivate Local People, Especially the Poor, to Participate in Conservation Efforts (Global Environment Facility, 1998)*.


PREDATORS, ECOLOGICAL ROLE OF

James Estes,* Kevin Crooks,† and Robert Holt‡
*University of California Santa Cruz, †University of California San Diego, and
‡The University of Kansas

I. Introduction
II. Background and Definitions
III. Case Studies
IV. Summary and Synthesis of Case Studies
V. Theoretical Studies of Predation and Biodiversity
VI. Generality
VII. The Future

GLOSSARY

apex predator An organism that occupies a food web’s highest trophic level.
bottom-up forces Population-regulating processes based on the availability of food, nutrients, and energy.
carnivore An organism that consumes other animals.
competitionism The view that competition regulates populations.
food web The interconnections among organisms based on diet.
herbivores Organisms that feed on plants and other photosynthesizers.
keystone species A strong interactor that is relatively rare.
Lotka–Volterra model An early equation relating rate of population change to the interplay between competition and predation.
mesopredator A small to mid-sized predator.
nutritionalism The view that bottom-up forces regulate populations.
phytoplankton Microscopic primary producers that live in water column habitats.
piscivores Predators that consume fish in aquatic habitats.
planktivores Predators that consume zooplankton in aquatic habitats.
top-down forces Population-regulating processes that originate from consumer limitation.
trophic cascades A chain reaction of top-down interactions across multiple trophic levels.

PREDATORS occur in all of the planet’s ecosystems and initiate top-down forces and trophic cascades in many of these. Although evidence for trophic cascades is strongest for aquatic systems, there is increasing evidence that they occur in a variety of terrestrial ecosystems as well. Trophic cascades result in weak or strong plant–herbivore interactions when the respective number of trophic levels is odd or even. Indirect food web effects of trophic cascades, known for lakes and kelp forests, are unstudied in most systems. The loss of large, apex predators from several terrestrial systems has resulted in mesopredator release—the proliferation of moderate-sized predators that commonly reduce or eliminate the smaller vertebrate species. Many dysfunc-
tional ecosystems have developed because of the loss of apex predators. Thus, reserve design and other conservation strategies must be adequate to preserve the apex predators.

I. INTRODUCTION

The science of ecology has undergone a succession of paradigms on the nature and importance of species interactions, including those between predators and their prey. The earliest view (henceforth termed nutritionalism) was that bottom-up forces (i.e., primary production and the efficiency of energy and material transport upward across trophic levels) regulate populations. Ecosystem ecology was built around this view of nature, which implicitly holds that apex predators, as the end points of energy and material flux, are of minor consequence to ecosystem function. Beginning in the late 1950s and early 1960s, the focus on species interactions changed to competition. In contrast with nutritionalism, competitionism holds that lateral forces within trophic levels regulate population abundance. By this view, predators are no more or less important than any other species. Recently, top-down forces have captured the attention of ecology, thereby legitimizing predators as important ecological entities. Ecologists now recognize that important species interactions follow all three pathways (Fig. 1), often simultaneously and sometimes interactively. Thus, although our focus in this article is on top-down forces generated by apex predators, understanding the ways in which predators influence biodiversity requires a more eclectic view of food webs and species interactions than simply “bottom-up vs top-down.”

We begin with a discussion of who the predators are and how they affect populations, communities, and ecosystems. We then present a series of case studies demonstrating the wide range of systems in which predation is an important organizing process, including examples of the unifying concepts and explanations of how they were discovered. This discussion is followed by a theoretical exploration of predation and biodiversity. Next, we discuss the levels of biological organization at which predation can influence biodiversity and develop a conceptual model for how apex predators might influence the location and strength of bottom-up and competitive forces in systems under top-down control. We conclude by considering the needs and opportunities for further research on predators and how predators are likely to figure into the future of conservation biology.

II. BACKGROUND AND DEFINITIONS

A. What Are Predators?

Broadly defined, all consumers are predators, thus including all living things except photo- and chemosynthesizers. The carnivorous plants add a minor wrinkle to this dichotomy. Nonetheless, predators would be represented by such diverse functional groups as herbivores, parasites (including microbes and parasitoids), and the immense diversity of invertebrate and vertebrate consumers that hunt and kill their prey. Attempts to define or classify predators based on trophic status, consumer–prey size relationships, or just about any other scheme are similarly problematic. For example, herbivores exist at one end of the trophic-status spectrum and microbes at the other, but it is uncommon for consumers in either group to kill their prey outright. Plants may even benefit from being eaten and a single act of predation by most microbes is of virtually no consequence to their prey because of the prey’s immensely greater relative body size. Although herbivores, microbes, and parasites are predators in this broad sense, our focus is on those that kill their prey. Even this restricted definition includes a vast array of species.
B. Species Interactions

The influence of predators on biodiversity depends first and foremost on direct predator–prey interactions. However, the consequence of predation to communities and ecosystems has less to do with these direct interactions than it does with their indirect effects. Although the direct effects of predation, by definition, are limited to impacts on prey populations, the nature of indirect effects is almost limitless, thus potentially causing populations to increase or decline anywhere in the food web. As will be shown in several of the case studies presented later, such indirect effects may involve long interaction chains, which in turn have broad impacts on associated ecosystems.

C. Indirect Effects of Predators

An awareness of the indirect effects of predators can be traced back at least to the writings of Charles Darwin, who described an interaction chain leading from cats to mice to bumble bees to clover. Similar early examples were provided by such well-known ecologists as Charles Elton and G. E. Hutchinson. Hairston, Smith, and Slobodkin’s (1960) now-classic paper (hereafter HSS) was perhaps the first effort to mold the indirect effects of predation into a conceptual model of trophic interactions and population regulation. HSS recognized four trophic groups—producers, decomposers, herbivores, and predators—and argued that although herbivores are commonly limited by predators, plants, decomposers, and predators are ordinarily limited by resources. The HSS model has weathered the test of time, along the way setting the stage for several conceptual advances, including the importance of top-down forces in population regulation and community organization, the ideas of keystone species and trophic cascades, and a generalized theory of food chain dynamics. Each of these is briefly explained in the following sections.

D. Top-Down Forces

Bottom-up forces are those passing from producers to consumers, whereas top-down forces are those passing from consumers to producers. As previously mentioned, recognition of top-down regulation dates back to at least Darwin, although it was HSS that introduced the idea in ecology. A Special Features section in the journal Ecology, published in 1992, stimulated further interest in the issue, in part by pointing out that top-down and bottom-up forces need not be competing processes, even though bottom-up forces are necessary for the function of all ecosystems. This realization freed ecologists to imagine a broad potential for the role of predation in nature.

E. Keystone Species

HSS was followed in the mid-1960s by Robert Paine’s highly influential paper on food web complexity and species diversity. Paine argued that predators often selectively consume and thus limit competitively dominant species, thus enhancing species diversity by releasing their subordinates from competitive exclusion. This argument was based on three essential premises: (i) Predators selectively consume the competitively dominant prey; (ii) in so doing, populations of the competitively dominant species are reduced; and (iii) in the absence of predation, the prey guild is limited by interspecific competition. Paine’s work captured the interest of community ecologists because it linked the influence of predators to species diversity, and (perhaps most important at the time) it was supported by results from field experiments. His empirical studies of predation by sea stars on mussel bed assemblages were done in the temperate rocky intertidal zone where competition for space can be extreme. This also led to two important developments in ecology: the idea of keystone species and the intermediate disturbance model of species diversity.

The intermediate disturbance model, further refined and generalized by Joseph Connell, holds that species diversity is influenced by the intensity of disturbance (either physical or biological; Fig. 2). When the intensity of disturbance is very high or very low, species diversity is low because the most vulnerable species are eliminated in the former instance and excluded by their competitive dominants in the latter. These limiting conditions are relaxed at intermediate levels of disturbance, thereby elevating species diversity. The notion of keystone species, as envisioned by Paine, applied to cases in which predators were the agents of disturbance. Although the definition of keystone species has broadened, on the one hand, to include other kinds of interactions and grown more restrictive, on the other hand, to exclude the effects of common species, this idea is rooted historically with the indirect effects of predators.

F. Trophic Cascades

A trophic cascade is the progression of indirect effects by predators across successively lower trophic levels.
FIGURE 2 The intermediate disturbance model of species diversity. This model is based on the assumption that competitive exclusion occurs in benign systems. Thus, when the intensity of disturbance (including predation) is low, strong competitive interactions by the dominant species reduce species diversity. When the intensity of disturbance is high, species diversity is again low because those species that cannot cope are eliminated. Maximum species diversity occurs at intermediate levels of disturbance—strong enough to prevent competitive exclusion but not so strong as to directly eliminate species.

HSS’s proposed relationship between predators, herbivores, and producers was a generalized trophic cascade. Stephen Carpenter and James Kitchell popularized this idea based on the striking influences of predatory fishes on the essential components of lake food chains—from minnows (the predatory fishes’ prey) to zooplankton (prey of the minnows) and to phytoplankton (prey of the zooplankton).

G. Generalized Food Web Theory

A generalized food web theory was developed by Stephen Fretwell to show how predation, trophic cascades, and food chain length combine to predict the strength of plant–herbivore interactions (Fig. 3). To understand this theory, first imagine an ecosystem with producers but no consumers. Lacking consumers, the producers are limited by competition for resources. Adding herbivores creates a two-trophic level system in which the plant populations become limited by herbivory. Adding predators limits herbivore populations, thus releasing the producers from limitation by herbivory and returning them to limitation by resource competition. The progressive increase of trophic complexity cascades downward through the food chain such that plant–herbivore interactions switch from being weak to strong as the respective number of trophic levels alternates between odd and even.

Next, we summarize many case studies that provide empirical evidence for these theories and concepts. Our presentation is organized around the various systems in which the work was done.

III. CASE STUDIES

A. Rocky Shores

Studies of rocky seashores furnish the earliest and some of the most compelling evidence for the effects of predation on communities and ecosystems. The first well-known experimental studies were done by Joseph Connell in Scotland. Connell’s work focused mainly on competition between the two barnacles (Cthamalus stellatus and Balanus balanoides) and predation on these species by the whelk (Thais latifilis). This research showed that the upper shore limit of Cthamalus was set by physical factors (weather) and the lower limit by competition for space with Balanus and predation by Thais.

Shortly thereafter, Paine began his studies of predation by the sea star (Pisaster ochraceus) in mussel beds along the outer coast of Washington. Paine hypothesized that sea star predation limited the lower distribution of mussels in the mid-litoral zone. This was subsequently confirmed by downward expansion of the mussel bed when the stars were removed. California mussels (Mytilus californianus) are the competitive dominants in this system. Predation by sea stars prevents mussels from controlling primary space (the rock...
(iii) the power of experimental evidence. Subsequent research has confirmed a similar role for Pisaster elsewhere in western North America and for other species of mussels and sea stars elsewhere in the world.

Other predators also influence rocky intertidal communities. Work by Philip Hockey and colleagues demonstrated a trophic cascade among African black oystercatchers (Haematopus moquini), herbivorous limpets, and intertidal algae. Succeeding studies of oystercatchers and limpets have confirmed similar interactions in South America, Australia, and western North America. Research in central and southern California further demonstrated how humans perturb the trophic cascade by exploiting owl limpets (Lottia gigantea, a large, territorial species) and by causing oystercatchers to abandon their breeding territories. The former effect causes a competitively subordinate guild of small limpets to replace owl limpets as the principal herbivore. The latter effect, induced simply by large numbers of humans being present along rocky shores, transforms the intertidal community from a three- to two-trophic level system. Small limpets come to dominate such areas, in turn reducing the algal cover. These human-caused perturbations probably are responsible for much of the modern-day character of rocky shores in central and southern California.

A final example of predation on rocky shores concerns the loco (Concholepas concholepas), a large, munsid gastropod that consumes intertidal mussels and is exploited by humans in central and southern Chile. Juan Carlos Castilla excluded humans from a small stretch of shoreline at the Las Cruces Marine Laboratory near Santiago in order to better understand their influence on this system, and as expected loco abundance greatly increased. The more surprising result was a whole scale shift in the intertidal landscape, from one dominated by extensive mussel beds to one largely devoid of mussels. This particular example is noteworthy because it demonstrates (i) how humans can perturb predator-mediated interaction chains with landscape-level consequences, (ii) that reserves can be used effectively both to demonstrate and to mitigate such effects, and (iii) the power of experimental evidence.

B. Kelp Forests

Kelp forest communities provide several examples of the ecological role of predators. One is that of the sea otter, which was hunted to near extinction in the Pacific maritime fur trade. Following protection in the early 1900s, the process of recovery created a fragmented population distribution within what had been a continuously occupied range. Contrasts between areas with and without sea otters revealed striking differences in kelp forest communities. Areas with sea otters supported lush kelp forests, whereas those without otters were extensively overgrazed by sea urchins, the otter’s principal prey. These patterns result from a trophic cascade, driven by sea otter predation on sea urchins, thus releasing kelp beds from sea urchin grazing.

In addition to contributing early empirical support for HSS, the sea otter–kelp forest system provides evidence for a wide range of predator-driven effects beyond those expected from simple trophic cascades. Sea otters influence numerous species by enhancing kelp abundance, thereby providing three-dimensional habitat and fueling increased primary production. This process is especially noteworthy because it shows how bottom-up processes can be altered by the top-down forces of apex predators. Other known or suspected consequences of sea otter predation in kelp forests are summarized in Fig. 4.

Understanding of the sea otter–kelp forest system has several interesting historical dimensions. Faunal remains in Aleut kitchen middens show that sea urchin size distributions during most of Aleut prehistory were similar to those of modern systems lacking sea otters, thus suggesting that aboriginal humans, by limiting sea urchins, influenced coastal ecosystems long before modern humans arrived on the scene. Paleontological and biogeographical data provide an even longer time perspective. Because the distribution of sea otters and their recent ancestors was limited to the North Pacific basin, their influence on the evolution of plant–herbivore interactions has been inferred by contrasting plant defense and herbivore resistance between North Pacific and Australasian kelp forests. Australasian kelp forests apparently lacked predators of comparable influence to the sea otter, at least since the Pleistocene–Pliocene. In contrast with North Pacific kelp forests, strong coevolutionary forces between marine plants and their herbivores in Australasia were thus expected, thereby facilitating an arms race between plant defense and herbivore resistance. Marine algae use secondary chemicals as their principal defenses against herbivory, and for this reason the evolutionary hypothesis was put to an initial test by measuring the secondary chemistry of North Pacific and Australasian seaweeds. Phlorotanins (the principal chemical defenses in brown seaweeds) concentrations were approximately an order of magnitude higher in kelp forests without sea otters than in/or.
FIGURE 4 A conceptual representation of the direct and indirect effects of trophic cascades in the sea otter–kelp forest system. See the legend to Fig. 3 for explanations of circles and arrows. Some of the known or suspected indirect effects of these two alternate states of kelp forest community organization are shown around the periphery of the central box. See Estes (1996) for further discussion of specific cases (reproduced with permission from Peterson and Estes, 2000).

magnitude greater in Australasian algae and North Pacific herbivores were more strongly deterred by these compounds than were their Southern Hemisphere counterparts. These evolutionary responses to predation probably explain why Northern Hemisphere kelp forests have been so devastated by sea urchin grazing following decimation of their predators.

The sea otter–kelp forest system changed remarkably in recent years as killer whales entered the coastal ecosystem and began preying intensively on sea otters after their normal prey populations (seals and sea lions) declined. Since the early 1990s, killer whale predation has driven otter numbers downward by approximately an order of magnitude across large areas of western Alaska. The consequent reduction in sea otter predation has caused sea urchin numbers to increase and kelps to decline (Fig. 5). This example illustrates that predator–prey interactions, acting through trophic cascades, influence herbivore–plant interactions in a manner consistent with the predictions described earlier for odd-vs even-numbered food chains (Fig. 4). It further indicates a role for predators in linking ecosystems over large areas.

Sea urchins have deforested kelp beds in the Gulf of Maine. Early reports attributed this to the overfishing of American lobsters (Homarus americanus), a purported ecological analog of the sea otter. However, this explanation is in doubt because the lobster fishery is apparently at an all-time high. Atlantic cod (Gadus morhua) also prey on a variety of benthic species, including urchins and lobsters, and the well-known collapse of cod populations may have influenced both lobsters and kelp forests in the Gulf of Maine.

Predation by sheephead (Semicossyphus pelcher, a benthic feeding fish) and spiny lobsters (Panulirus interruptus) is thought to limit sea urchins in warm-temperate kelp forests of southern California. Sea otters also occurred in this system, but deforestation events in this area did not occur until long after the otter’s demise. The alternative urchin predators may explain the difference. As humans have progressively depleted these predators in recent decades through commercial and recreational fisheries, deforestation has become an increasing problem.

In the Southern Hemisphere, predation by rock lobsters (Jasus lalandii) in South Africa limits predatory whelks, in turn releasing subtidal mussel beds from limitation by whelk predation. A remarkable predator–prey role reversal occurred in this system following the extirpation of lobsters from several small islands. Whelk populations increased substantially in the lobsters’ absence, thus transforming the reef from a mussel bed...
into a kelp forest. In an effort to reestablish lobsters and their associated role as the system’s dominant predator, a large number of lobsters were relocated to one of the islands. However, the whelks had become so abundant that they attacked the lobsters in mass, killing all of them within hours of the translocation. This surprising case study demonstrates how a density-dependent role reversal between predator and prey can generate alternate stable-state communities.

These several examples from kelp forest systems...
have shown that predators can shape populations and communities on ecological timescales and life history characters on evolutionary timescales. Comparative studies of sea urchins in tropical and temperate reef systems also suggest that predators influence their prey's behavior in complex ways. In warm temperate/tropical systems, sea urchins commonly retreat to protective cracks and crevices within the reef during daylight hours in order to avoid being eaten by diurnally active benthic predatory fishes. Fishes quickly attack urchins removed from their refuges during the day and placed on exposed habitats. Similar patterns have been shown for a variety of warm temperate and tropical systems in which benthic predatory fishes occur. However, the nature of urchin behavior appears to differ between species whose evolutionary histories are rooted in tropical vs temperate environments. Tropical species tend to display diel sheltering as a fixed behavior, regardless of ecological context, whereas the sheltering behavior is plastic in temperate species depending on whether predatory fishes are present or absent. The explanation for this difference in plasticity may lie in the fact that tropical urchins have long been subject to predation by diurnally active fishes, whereas temperate urchins have come into contact with benthic predatory fishes more recently, and then only at the warm margins of their geographical ranges.

C. Lakes

Studies of freshwater lakes provide some of the clearest and best known evidence for trophic cascades. There are two main reasons for the quality of this evidence. Lakes, as discrete and recurrent entities, are well suited for comparative and experimental studies. Furthermore, the producers and herbivores (especially phytoplankton and zooplankton) have very short generation times, thereby making population-level responses to perturbations rapid enough for scientists to observe and document.

Some of the earliest evidence for the influence of predation in lake systems comes from Brooks and Dodson's analysis of New England lakes. These researchers showed that in the absence of planktivorous fishes, zooplankton assemblages were dominated by species with large body size because of their increased foraging efficiency and competitive superiority over small species. In lakes with planktivorous fishes, the composition of the plankton shifted toward small body size due to the influence of size-selective predation. This example was followed by Zaret and Paine's report on the cascading influences of introduced peacock bass (Cichla ocellaris) to Lake Gatun, Panama. Peacock bass, a cichlid native to the Amazon River, was first introduced to Lake Gatun in 1963 for sport fishing and consumption. These introduced predators are voracious piscivores and they caused a remarkable series of food web effects as the bass population grew and spread across Lake Gatun. The immediate influence was a rapid and extreme reduction of planktivorous minnows, thus causing zooplankton populations, including that of mosquito larvae, to increase. This example added two interesting dimensions to the understanding of lake systems. One is the strength of influence by an exotic predator on naıve prey, with broad-ranging indirect effects across the lake and surrounding terrestrial systems. In addition to the top-down effects described previously, the reduced populations of planktivorous minnows negatively impacted other apex predators, including several species of aquatic birds and predatory fishes. Another dimension is the potential impact on human health, in this case resulting from an increased threat of malaria because of increased mosquito populations. Similar examples of broad-ranging influences by exotic predators are known for many other lake systems throughout the world.

Numerous reports from various lake systems throughout the world show that altered populations of apex predators result in altered food webs. The essential players in these lake systems include four main groups of organisms: phytoplankton, herbivores, planktivores, and piscivores. The relationship of the first three of these to piscivore abundance, explained by cascading trophic interactions, is shown in Fig. 6. The evidence for these interactions comes from a variety of areas and approaches. Early insights were provided by contrasts among lakes in which piscivore populations varied serendipitously but for unknown reasons. There are many such examples from tropical and temperate lake systems in both the New and Old Worlds. Additional evidence that these patterns are caused by trophic cascades has come from the results of microcosm experiments, by tracking changes associated with the fortuitous extinction or reintroduction of piscivores into particular lakes through time, and recently by whole-lake experiments in which the piscivores were purposely added or removed. Although the details vary depending on such factors as food chain length and the nature of particular species, the overall view of food web dynamics in lake ecosystems is remarkably uniform, especially the importance of apex predators and trophic cascades.
D. Rivers and Streams

Experimental work in rivers and streams has demonstrated important influences of predation on both food web structure and the life history of prey populations. The structure and dynamics of river food webs are grossly similar to those described for lakes, the main differences being that rivers are episodically disturbed by changes in water flow and they depend less on waterborne phytoplankton and zooplankton. Both fishes and birds are important apex predators in river food webs, and like lakes, many river food webs are strongly influenced by trophic cascades. The experimental exclusion of these predators from a variety of tropical and temperate river systems by Mary Power and colleagues provided several novel dimensions to the understanding of predation and trophic cascades. The manipulation of predatory fishes and other consumers provided consistent evidence for top-down forces and exclusion of birds has revealed depth-related gradients in the outcomes of trophic cascades. These findings show that the influence of predators on food webs can be strongly influenced by prey refuges, which in turn can vary across habitat gradients. Another important contribution of the riverine studies is that they have been done in systems that deviate in food chain length, thus providing the first experimental evidence that the strength of plant–herbivore interactions varies predictably between odd- and even-numbered food chains. Work by Sih and colleagues in streams of the eastern United States also shows how the risk of predation can influence the interplay between feeding and reproductive behavior in several prey species.

The pioneering work of Endler and Reznick on Trinidadian guppies provides some of the strongest and most comprehensive evidence for the selective effects of predation. The risk of predation to guppies by various larger fish species varies within and among streams, and a variety of life history characters, color patterns, and features of guppy mating systems covary accordingly. By manipulating predator populations and translocating guppies among habitats, these researchers demonstrated rapid selective responses to altered risks of predation.

E. Oceanic Systems

Although the oceans dominate our biosphere and provide critical ecosystem services in such diverse forms as food production and climate control, little is known about the role of apex predators in the open sea. One reason is that the open sea and its associated seafloor habitat present serious logistical challenges to studies of any kind. Furthermore, many ocean ecologists still embrace the view that bottom-up processes are the main drivers of biological pattern in ocean ecosystems. Although bottom-up forcing in the sea is clearly important, this does not preclude top-down effects, which might be expected for several reasons. One is that strong predator-induced effects occur broadly in lakes and the general structures of ocean food webs (from phytoplankton to zooplankton to planktivores–piscivores) are similar to those of lakes. A second is that nowhere else on the planet are predators so abundant, as witnessed by the vast schools of marine mammals, seabirds, and predatory fishes. Despite this, we are aware of but one example of an oceanic trophic cascade. Pink salmon (Oncorhynchus gorbuscha) populations in the North Pacific fluctuate on a 2-year cycle. During years when pink salmon are abundant, zooplankton are depressed and phytoplankton are abundant, whereas during years when pink salmon are rare, zooplankton are abundant.
and phytoplankton are relatively rare. Another potential example of predation in the open sea resulted when the blue, fin, sei, and minke whales were decimated by the whaling industry. This reduction in the great whales may have released Antarctic krill populations from limitation by predation, in turn elevating the carry capacities of other krill-feeders—pinnipeds, penguins, and perhaps additional groups of consumers. Increased growth rates and reduced age of first reproduction of seals and whales after the depletion of great whales from Antarctica have been interpreted as evidence for such effects.

Predation by gray whales (Eschrichtius robustus) and walruses (Odobenus rosmarus) has important effects on seafloor systems. Gray whales influence these systems by resuspending sediments and consuming amphipod amphipods. Furrows formed by the whales in the soft benthos are colonized by scavenging bivalve amphipods, serve to accumulate detritus, and thus facilitate a local detritus-based food web. Walruses further impact these systems by consuming clams and other large infauna, in turn attracting predatory and detritivorous sea stars.

Although evidence from food web dynamics for a role by apex predators in the open sea is spotty at best, behavioral patterns of various prey species suggest strong predator-prey interactions. For example, krill and other large zooplankton form dense swarms, which probably reduce their likelihood of being consumed by predators that must search for and capture individual prey. Pagophilic (ice-loving) pinnipeds in the Arctic and Antarctica also provide a commanding case. In the Arctic, where polar bears and humans are both important predators, pinnipeds flee from the ice to water at signs of danger. In Antarctica, where the threat of predation is much greater in the water (from killer whales and leopard seals) than it is on the ice, the pinnipeds do not display such extreme flight behavior and often are nearly oblivious to potential disturbances when hauled out.

F. Boreal/Temperate Forests

Although terrestrial biotas of the New World once contained numerous large mammalian carnivores, the potential ecological significance of these predators was unknown until recently. There are at least five reasons for the prolonged state of ignorance. One is that the largest of these creatures—gray wolves and grizzly bears—were all but exterminated in the United States and Mexico well before modern ecological research had taken form. Second, even if large carnivores have been able to persist in the face of direct persecution, they are extremely difficult animals to study due to their low densities, nocturnality, secretive habits, aggressive behavior, and wariness of humans. Third, just as ocean ecologists have downplayed the importance of predators in the open sea, many wildlife ecologists have tended to be skeptical about the importance of predation in population regulation, and this topic has been hotly debated in the wildlife literature. In a famous example, Rasmussen, Leopold, and then HSS attributed an irruption of mule deer on the Kaibab Plateau, subsequent overgrazing, and the eventual mass starvation of the deer herd to the extermination of gray wolves and other large predators, but Caughley later attempted to debunk this explanation. Fourth, the long generation times of key players (decades to centuries for trees; multiple years to decades for ungulates and carnivores) and the large areas required for the measurement or manipulation of their representative populations make rigorous study of the top-down effect of apex predators very challenging. Finally, political, social, ethical, and legal issues have dissuaded many scientists from studying large mammals.

Despite these difficulties, there are indications of top-down effects by large predators in boreal forests. McLaren and Peterson used historical information on wolf and moose abundance, together with growth ring measurements from balsam fir, as evidence for a trophic cascade at Isle Royale in Lake Superior (Fig. 7). Wolf numbers have fluctuated substantially throughout the twentieth century, apparently in large measure because of demographic factors related to their small population size. Inverse changes in moose numbers followed wolf population fluctuations, thus suggesting regulation by wolf predation. Direct measures of herbivory were unavailable. However, the distance between annual tree rings in balsam fir showed that sapling growth rates were lower when moose were abundant than when moose were rare.

G. Fragmented Coastal Scrub Habitats

Although experimental manipulation of terrestrial carnivores is exceedingly difficult, fragmented habitats can provide valuable, large-scale, ecological experiments to rigorously explore the top-down effects of mammalian predators. Large carnivores are particularly vulnerable to local extinction in fragmented landscapes due to large ranges and resource requirements, low population
PREDATORS, ECOLOGICAL ROLE OF

FIGURE 7: The trophic system on Isle Royale, reconstructed for 1958–1994. (A) Wolf abundance calculated from aerial surveys; (B) moose abundance calculated from skeletal remains and aerial surveys; mean ring-width indices for balsam fir from the west (C) and east (D) ends of Isle Royale; and (E) actual evapotranspiration from April to October, a measure of water availability during the growing season. The shaded areas highlight intervals of forage suppression that the authors believe are closely tied to periods of elevated moose density, which in turn follow periods of low wolf density (reproduced with permission from McLaren and Peterson, 1994).

densities, and direct persecution by humans. Larger carnivores can depress populations of smaller mammalian carnivores, or “mesopredators,” through direct predation, resource competition, and interference competition, including spatial and temporal avoidance. Consequently, the decline and disappearance of dominant carnivores in fragmented systems may lead to the ecological release of smaller predators that in turn threaten birds and other vertebrates. This process has been labeled “mesopredator release” and the phenomenon has been implicated in the extinction of prey species worldwide.

In coastal southern California, intensive urbanization during the past century has destroyed most of the native coastal sage scrub and chaparral habitats, creating a “hot spot” of endangerment and extinction in the region and leaving undeveloped canyons dissecting coastal mesas to function as habitat islands immersed within a matrix of inhospitable urban habitat. Michael Soule proposed the mesopredator release hypothesis as a possible mechanism to explain the rapid disappearance of scrub-breeding birds from the habitat fragments. He predicted that the decline of the most common large predator (coyote) would result in the ecological release of native (striped skunk, raccoon, and gray fox) and exotic (domestic cat and opossum) mesopredators, and that increased predation by these particularly effective avian mesopredators would result in higher mortality and extinction rates of scrub-breeding birds.

To test this prediction, Crooks and Soule exploited a serendipitous ecological experiment—spatial and temporal variation in the distribution and abundance of coyotes among these urban habitat fragments—to investigate direct and indirect effects of this top predator on community structure. In accordance with the mesopredator release hypothesis, lower visitation rates of coyotes in small, isolated remnants resulted in elevated numbers and activity of urban mesopredators. Coyotes directly preyed on some mesopredator species; for example, domestic cats were found in approximately 20% of coyote scats in the fragments. Mesopredators temporally avoided coyotes as well. In fragments that coyotes visited episodically during the course of the study, mesopredator activity increased when coyotes were absent. As predicted, scrub bird diversity was lower in fragments with fewer coyotes and more mesopredators, even after accounting for the positive effect of fragment area and the negative effect of fragment age on bird persistence.

The top-down effect of coyotes on cats seems to have had the strongest impact on the decline and extinction of scrub-breeding birds in the urban fragments. Unlike wild predators, domestic cats are recreational hunters. Maintained well above carrying capacity by nutritional subsidies from their owners, they continue to kill even when prey populations are low. Using data on cat densities and predation rates, Crooks and Soule estimated that cats surrounding a moderately sized fragment return approximately 840 rodents, 523 birds, and 595 lizards to residences per year. Such high levels of predation appear to be unsustainable for many small vertebrate populations. For example, existing population
sizes of some birds do not exceed 10 individuals in small to moderately sized fragments, so even modest increases in predation pressure from mesopredators may quickly drive native prey species to extinction. Extinctions of scrub-breeding birds are frequent and rapid; at least 73 local extinctions may have occurred in these fragments during the past century. Overall, this example illustrates that trophic cascades generated by the disappearance of an apex predator can combine with other fragmentation effects to influence species diversity in terrestrial systems.

H. Midcontinent North American Prairies

The top-down effect of coyotes is evident in other systems as well. Historically, coyotes were generally confined to open plains and arid regions of western North America. The eradication of wolves from most of the continental United States in the late 1800s and early 1900s likely facilitated the expansion of coyote populations. Currently, predator control efforts in the United States are directed primarily at coyotes, and lethal and nonlethal measures have resulted in at least temporary coyote declines in some areas. In the Prairie Pothole Region of the North American midcontinent, coyote population reduction is considered one of the principal causes of increases in red fox populations starting in the 1930s; other carnivore species have been shown to increase in areas of coyote control as well. Although coyotes will kill red foxes, spatial and temporal avoidance, behavioral exclusion, and territorial shifting are likely the primary mechanisms by which coyotes reduce fox populations.

Red foxes are the most important predators on nesting ducks and their eggs and offspring in the prairie region. The ecological release of red foxes following coyote control has therefore resulted in increased predation on ground-nesting dabbling ducks, most notably mallards. Predation accounts for more than 70% of nest failures in these duck species, and in some areas intense predation on eggs, ducklings, and hens has been sufficiently intense to depress recruitment below replacement levels. Predation has resulted not only in population declines of duck species in the Prairie Pothole Region but also in altered population composition and skewed sex ratios.

Interestingly, since the mid-1970s, coyote populations have begun to rebound in parts of the Prairie Pothole Region due to restrictions in control and fur harvest methods and to reduced commercial value of fur. In these areas, expanding coyote populations have contributed to reduced red fox activity and higher duck nest success; overall nest success in sites where coyotes are the principal canid is nearly twice as high as that in areas where red foxes dominate. Overall, the excellent series of studies by Sargeant and colleagues in the Prairie Pothole Region highlight the changes in the canid predator assemblage in North America during the past century and emphasize the top-down community-level effects generated by these dynamic canid–canid interactions.

I. Tropical Forests

Some of the most dramatic evidence for top-down control by apex predators comes from research by John Terborgh and colleagues in New World tropical forests. Terborgh’s vision of top-down control in this system stems from a contrast between Barro Colorado Island, in the Panama Canal, and Cocha Cashu Biological Station in Peru’s Manu National Park. Although the two sites are similar in climate and native biota, Barro Colorado Island currently supports notably higher densities of herbivorous mammals, such as agoutis, coatimundis, sloths, and howler monkeys, than does Cocha Cashu—differences attributed to the loss of predators from Barro Colorado Island. These ideas are now being put to a more rigorous test by using recently formed habitat fragments—the islands of Lago Guri in Venezuela—as a large-scale ecological experiment. The Caroni Valley of east-central Venezuela supports notably higher densities of herbivorous mammals, such as agoutis, coatimundis, sloths, and howler monkeys, than does Cocha Cashu—differences attributed to the loss of predators from Barro Colorado Island. These ideas are now being put to a more rigorous test by using recently formed habitat fragments—the islands of Lago Guri in Venezuela—as a large-scale ecological experiment. The Caroni Valley of east-central Venezuela supports notably higher densities of herbivorous mammals, such as agoutis, coatimundis, sloths, and howler monkeys, than does Cocha Cashu—differences attributed to the loss of predators from Barro Colorado Island. These ideas are now being put to a more rigorous test by using recently formed habitat fragments—the islands of Lago Guri in Venezuela—as a large-scale ecological experiment. The Caroni Valley of east-central Venezuela supports notably higher densities of herbivorous mammals, such as agoutis, coatimundis, sloths, and howler monkeys, than does Cocha Cashu—differences attributed to the loss of predators from Barro Colorado Island.

Terborgh and colleagues in New World tropical forests. Some of the most dramatic evidence for top-down control by apex predators comes from research by John Terborgh and colleagues in New World tropical forests. Terborgh’s vision of top-down control in this system stems from a contrast between Barro Colorado Island, in the Panama Canal, and Cocha Cashu Biological Station in Peru’s Manu National Park. Although the two sites are similar in climate and native biota, Barro Colorado Island currently supports notably higher densities of herbivorous mammals, such as agoutis, coatimundis, sloths, and howler monkeys, than does Cocha Cashu—differences attributed to the loss of predators from Barro Colorado Island. These ideas are now being put to a more rigorous test by using recently formed habitat fragments—the islands of Lago Guri in Venezuela—as a large-scale ecological experiment. The Caroni Valley of east-central Venezuela supports notably higher densities of herbivorous mammals, such as agoutis, coatimundis, sloths, and howler monkeys, than does Cocha Cashu—differences attributed to the loss of predators from Barro Colorado Island.
points. First, predators often exert crucial roles in maintaining local species diversity. Terborgh’s data indicate that the majority of tree species will eventually be lost from these islands systems, largely because of the loss of predators. This example, together with the Crooks–Soulé study of fragmented coastal scrub habitats, also serves to remind us of the power of ecological experiments and the importance of scale (space and time) in designing studies on the role of apex predators in nature. It is highly unlikely that any amount of study of unperturbed terrestrial systems could have demonstrated the magnitude and breadth of effect by relatively rare apex predators. However, anthropogenically disturbed systems, such as fragmented landscapes, offer unique opportunities to understand the complex trophic interactions generated by large carnivores.

**J. Exotic Predators on Islands**

Perhaps nowhere is the top-down effect of predation on biodiversity so apparent as with the introduction of nonnative predators onto islands. Islands typically support few large predators and grazers, and apex predators such as mammalian carnivores are often absent. Consequently, insular endemic species regularly evolve in the absence of predation and thus lack adequate antipredator defenses. For example, many insular animals exhibit tame or fearless behavior that increases their vulnerability to introduced predators, and some island birds are flightless ground nesters. Similarly, many island plants do not produce the same noxious chemicals or physical defenses found in their closest mainland relatives—features that discourage herbivory. Consequently, the introduction of nonnative predators can be catastrophic for sensitive island communities, to the point of driving insular prey species to extinction. The examples are numerous. Worldwide, of all the species that have gone extinct since 1600, 90% of the 30 species of reptiles and amphibians, 81% of the 65 mammal species, and 93% of the 176 species of birds each year. On subantarctic Marion Island, 5 cats were introduced as pets in 1949 and by 1975 about 2000 cats were killing 450,000 burrowing petrels annually and were suspected in driving another petrel species to local extinction. In the mid-1900s about 5 cats were introduced to Kerguelen Island in the sub-Antarctic and their descendants have killed more than 3 million petrels per year and are responsible for the extinction of several bird populations. In the most infamous and perhaps most extreme example known, the lighthouse-keeper’s pet cat on Stephen Island, a tiny island off New Zealand, arrived in 1894 and within a single year this one cat exterminated the flightless Stephen Island wren. The indirect effects of these changes, although likely important in some cases, are largely unstudied. The effects of introduced predators extend far beyond their prey species and can include modification of ecosystem-level processes. For example, the New Zealand flatworm (Aratiopousia triangulata) was accidentally introduced to the British Isles in the early 1960s and, with no natural predators, has spread rapidly. The flatworm is a voracious predator of native earthworm species. Earthworms, through their burrow excavation and casting activity, provide an invaluable ecosystem service by shaping the structure and hydrological patterns of soils. Flatworm infestations and the consequent depletion of earthworms alter both soil structure and hydrology. The ramifications are far-reaching and directly impact human welfare. By depleting native earthworms, the introduced New Zealand flatworm increases at least 15, 39, and 53 insular bird species, respectively, and introduced rats are thought to be responsible for 54% of insular bird extinctions caused by predators. For example, the introduction of black rats on Big South Cape Island, New Zealand, in 1906 resulted in the rapid extinction of 5 species of land bird, 1 species of bat, and an unknown number of invertebrates, including a species of large flightless weevil. On the Galapagos Islands, introduced black rats have reduced populations of the giant tortoise and the dark-rumped petrel by preying on eggs. Mongoose and domestic cats have also been introduced to islands, at times deliberately to control nonnative rodents. Unfortunately, they instead often eradicate native prey species. On Hawaii, introduced mongoose had little impact on rodents but decimated flightless rail populations. Domestic cats have been accidentally or deliberately introduced to at least 65 island groups and are thought to be responsible for 26% of insular bird extinctions caused by predators. Incredibly, 375 cats on Macquarie Island near Australia killed an estimated 56,000 rabbits and 58,000 ground-nesting seabirds each year. On subantarctic Marion Island, 5 cats were introduced as pets in 1949 and by 1975 about 2000 cats were killing 450,000 burrowing petrels annually and were suspected in driving another petrel species to local extinction. In the mid-1900s about 5 cats were introduced to Kerguelen Island in the sub-Antarctic and their descendants have killed more than 3 million petrels per year and are responsible for the extinction of several bird populations. In the most infamous and perhaps most extreme example known, the lighthouse-keeper’s pet cat on Stephen Island, a tiny island off New Zealand, arrived in 1894 and within a single year this one cat exterminated the flightless Stephen Island wren. The indirect effects of these changes, although likely important in some cases, are largely unstudied. The effects of introduced predators extend far beyond their prey species and can include modification of ecosystem-level processes. For example, the New Zealand flatworm (Aratiopousia triangulata) was accidentally introduced to the British Isles in the early 1960s and, with no natural predators, has spread rapidly. The flatworm is a voracious predator of native earthworm species. Earthworms, through their burrow excavation and casting activity, provide an invaluable ecosystem service by shaping the structure and hydrological patterns of soils. Flatworm infestations and the consequent depletion of earthworms alter both soil structure and hydrology. The ramifications are far-reaching and directly impact human welfare. By depleting native earthworms, the introduced New Zealand flatworm increases
the risk of surface runoff and therefore the potential of soil erosion, agrochemical pollution, and flooding.

Threatened prey species often recover, sometimes rapidly, when alien predators are controlled or eradicated on islands. However, the removal of alien predators can also yield unexpected results. For instance, a recent eradication of introduced black rats on Bird Island in the Seychelles has resulted in a population explosion of exotic crazy ants (Anoplolepis longipes). Ironically, these ants are now threatening those bird colonies that the rat eradication was intended to protect. Furthermore, simulation models predict that on islands colonized by both cats and rats, elimination of cats may release rat populations, and that increased numbers of rats may actually increase predation pressures on island birds. In essence, this cascade represents another example of the mesopredator release phenomenon, with cats on islands as top predators and rats as the mesopredators. On a variation of the theme, controlling cats on islands that also support exotic rabbits may result in more rabbits, excessive grazing by these prolific herbivores, and severe impacts to insular vegetation and associated animal species; this example, therefore, directly follows the predictions from the models of HSS and Fretwell. Clearly, alien predators on islands represent a complex, unpredictable, and occasionally dramatic example of the relationship between predators and biodiversity.

IV. SUMMARY AND SYNTHESIS OF CASE STUDIES

It is abundantly clear from the preceding examples that the manifestations of predation in nature are dramatic and diverse, occurring at organizational levels ranging from the behavior of individuals to the dynamics of ecosystems and on timescales ranging from ecological to evolutionary. Numerous studies show or suggest that predators influence the abundance, distribution, and population structure of their prey. Indirect effects of predation are less appreciated by scientists and the public, despite the fact that they occur broadly in nature, are important to ecosystem function, and often result in processes that benefit human welfare. Trophic cascades are the most common of known indirect effects. These may be nearly as ubiquitous in nature as the transfer of material and energy upward through food webs. In any case, ecologists should be more surprised by the absence of such top-down effects than by discoveries of new ones.

Despite extensive evidence for trophic cascades from many of the planet's major ecosystems, we know little else about their influences on overall food webs. The lake studies provide an important exception because here a relationship between the top-down effects of predation and the bottom-up effects of production has been shown. These findings indicate that predators may fuel production in odd-numbered food chains, and that maximum production across all trophic levels should be realized at intermediate intensities of predation. A corollary to this hypothesis is that the length of food chains under top-down control is necessarily limited not by production and the efficiency of energy transfer but by the population constrictions that occur at consumer-regulated trophic levels. Such constrictions may be common in nature because intermediate intensities of predation probably occur rarely, except in highly managed ecosystems.

The theory of trophic cascades also provides guidance on where in a food web one might expect to find strong competitive interactions. Competition should be most strongly manifested within trophic levels in which populations are resource rather than consumer limited. These occur at the odd trophic levels in odd-numbered systems and even trophic levels in even-numbered systems. Although this prediction requires further analysis, it provides hope for an integrated theory of what previously has been viewed and treated as the largely unrelated processes of top-down, lateral, and bottom-up species interactions.

What humans perceive as "dysfunctional ecosystems" are often consequences of the recently altered roles of predators (e.g., losses of native species or introductions of exotics). Local extinctions and invasions are increasingly common, but unless these changes are observed or known from historical records, their significance to extant ecosystems may be difficult to understand. This point is illustrated by the following example. In a recent essay, Paine argued that the HSS model is generally correct for herbivorous insects but incorrect for herbivorous mammals—a contention supported by numerous examples of plant damage by mammalian grazers. The purported difference between insects and mammals as agents of herbivory has at least three possible explanations: that vegetation is intrinsically more vulnerable to mammalian than insect herbivores, that mammalian herbivores are less vulnerable to predation than their insect counterparts, or that the predators of mammals have been lost in disproportionately large numbers. The known reductions of large carnivores in North America make the latter mechanism a strong possibility.
assessed? As seen in some of the preceding examples, predation often leaves a mark on species-level characters, especially behavior, life history, and morphology. Vermeij’s analysis of shell damage and morphological change (ostensibly from crushing predators) in Mesozoic marine gastropods provides a good example of one such record. The sudden increase in crushing predators was responsible for what Vermeij termed the “Mesozoic marine revolution.” Using this perspective, Richard Palmer confirmed that shell structures indeed reduce the incidence of attack from crushing predators by experimentally removing spires from gastropod shells. Behavioral patterns also provide clues about the role of predators, as described in the case studies for sea urchins, pagophillic pinnipeds, and North American pronghorn. Other examples could be cited, but those listed are sufficient to make the point that morphology and behavior, when thoughtfully and cautiously interpreted, frequently reflect evolutionary response to predation.

Although these examples provide insight into species-level responses to predation on historical timescales, they afford little insight into the food-web effects of predators. For this purpose, one might profitably examine the characteristics of producers. Plants can deter herbivores by modifying their morphology, demography, and chemistry. The degree to which these defensive characters exist among plant species and populations sometimes indicates the intensity of herbivory on historical timescales. Examples of this approach are provided by studies of differences in meristem location in steppe vegetation between the eastern and western slopes of the northern Rocky Mountains, variation in the resistance of birch trees to insect herbivores in boreal forests, the susceptibility to grazing damage in marine algae across coral reef habitats, and the evolution of reduced chemical and physical defenses in insular plants in the absence of herbivory. Mismatches in extant communities between the intensity of herbivory and the degree of plant resistance sometimes can be taken as evidence for recent changes in top-down regulation, as suggested from the previously described contrast between Northern and Southern Hemisphere kelp forests. Similar approaches might be taken to discern the evolutionary importance of apex predators in other ecosystems.

V. THEORETICAL STUDIES OF PREDATION AND BIODIVERSITY

The topic of predation and biodiversity involves interactions among multiple species at different trophic levels that interact in nonlinear, complex ways. A full understanding of this topic requires the analysis of mathematical and computer models, which permit one to keep track of multiple forces influencing the dynamics of interacting populations. A rich theoretical literature exists exploring impacts of predation on the dynamics and structure of ecological communities. Here, we summarize highlights of this work, emphasizing conceptual insights rather than mathematical details.

A predator can influence whether or not a particular species is present in a community either by facilitating its persistence (i.e., predators can enrich species composition) or by preventing it from invading (i.e., predators can constrain species composition). If removing an apex predator greatly increases the abundance of a particular prey, and this prey is a predator on species at lower trophic levels, predator removal could indirectly lead to shifts in competitive interactions and thus persistence of species several trophic levels removed (as in a trophic cascade). Even if a predator does not dramatically affect composition, it may strongly influence relative abundances of resident community members. Finally, predators can influence the existence and magnitude of temporal fluctuations in abundance. Mathematical models help one understand all these effects.

To analyze mechanisms influencing species composition, we consider the growth of species when rare. If each species in a community increases when rare, diversity is maintained in the face of perturbations. In theoretical studies, one writes equations describing the dynamics of each species and then analyzes this set of equations (e.g., with and without a top predator). To illustrate the basic approach, we discuss a simple model in detail and then briefly discuss results from other models. Theoretical studies suggest that there is no single relationship between predation and biodiversity but instead many relationships, depending on numerous contingent details of systems.

A. Predation as a Density-Independent Mortality Factor: Effects on Biodiversity

As discussed previously, predator removal can unleash competition among prey and induce a wave of additional extinctions. To understand this effect, we express the growth rate of each species as a function of three factors: $dN/dt = [\text{inherent growth}] - [\text{effect of resident competitor}] - [\text{mortality from resident predator}]$, where $N$ is abundance. Predation may both facilitate coexistence (e.g., by reducing the abundance and impact of competitors) and hamper coexistence (e.g., by direct mortality).
The most basic effect of predation is increased prey mortality. The simplest form of predation is that described by a fixed, density-independent mortality term. Generalist predators can occasionally act in this manner. Assume that two species experience strong, direct competition, described by the classic Lotka-Volterra competition model, with added mortality due to predation:

\[
\frac{dN_1}{dt} = N_1 r_1(K_1 - N_1 - a_12 N_2)/K_1 - m_1 N_1,
\]

where \(N_1\) is the density of species 1, \(N_2\) is that of species 2, \(r_1\) and \(K_1\) respectively are the intrinsic growth rate and carrying capacity of species 1, and the competition coefficient \(a_{12}\) is the effect of an individual of species 2 on species 1 (compared to the effect of 1 on itself). The second term expresses predation as density-independent mortality at a constant per capita rate \(m_1\). (A comparable equation for species 2 completes the model.)

Using the criteria that each species should increase when rare, after some algebra the following condition for species coexistence emerges:

\[
\frac{1}{a_{12}} \frac{K_1(1 - m_2/r_2)}{K_2(1 - m_1/r_1)} > a_{21},
\]

where the term \(K_1(1 - m_2/r_2)\) is the effective carrying capacity of species 1 in the face of predation. This inequality implies several interesting conclusions. If \(a_{21} > 0\), no pattern of imposed, density-independent mortality leads to coexistence. (This in effect says that interspecific competition is stronger than intraspecific competition.) Likewise, if competition coefficients are unity, one will not observe coexistence, regardless of the pattern of mortality. If the two competitors have the same intrinsic growth rate, and predation is uniform (or more generally, \(r_i/m_i = r_j/m_j\)), mortality drops out, so there is no effect of predation on coexistence.

However, if \(a_{21} > 0\) and the two competitors differ in the ratio \(m_i/r_i\), predation can occasionally facilitate coexistence. Consider a case of a competitive hierarchy, such that \(a_{21} > 0\) but \(a_{12} = 0\) and species 2 is competitively excluded. In the absence of the predator, species 2 is excluded if \(K_1 < K_2 a_{21}\), with the predator, coexistence is permitted if \(K_1(1 - m_2/r_2) > K_1(1 - m_1/r_1)a_{21}\). Comparing these two inequalities, it can be seen that the predator facilitates coexistence only if \(m_2/r_2 < m_1/r_1\). In other words, the dominant competitor must either experience higher mortality or have a lower intrinsic growth rate. However, if predation is too low (low \(m\) for both species), there will still be competitive exclusion. Moreover, if predation is too intense, there will not be coexistence because one (or both) species is directly eliminated by predation.

This model illustrates several general points that are applicable to a wide range of models. First, predation will not always facilitate coexistence. Second, if a key-stone predator effect is possible, the effect occurs only within a particular range of population parameter values. Typically, predator-mediated coexistence requires an intermediate level of predation. Very intense, generalized predation will almost always reduce species richness. This is particularly likely in systems in which prey species have not had a shared evolutionary history with predators (e.g., the brown rat snake as a predator on birds at Guam), the physical structure of the environment makes it easy for predators to encounter prey (e.g., no refuges, as in open lakes), or the species in question have low intrinsic growth rates. Third, predator-mediated coexistence requires a tradeoff. The species that is the superior competitor needs to be more vulnerable to predation. When this occurs, predator removal will endanger the persistence of inferior competitors (as in Paine's Pisaster). Finally, the appropriate measure of a species' vulnerability to predation combines mortality rates \((m)\) and the ability to replenish losses \((r)\). High, uniform rates of mortality tilt the balance of competitive interactions toward species with high intrinsic growth rates.

The previous model structure assumes that predation influences competition via changes in abundance. However, prey can also show behavioral shifts when faced with predators—for instance, spending more time in refuges and less on foraging. Such 'higher order interactions' can either make coexistence more difficult or weaken competition, depending on the detailed nature of the behavioral changes.

The Lotka-Volterra model most literally applies to systems with strong, direct interference interactions, in which the only dynamical variables are each competing species density. In models of exploitative competition for a single limiting resource, predation that leads to density-independent mortality can influence which species wins, but it will not lead to coexistence. Predation is more likely to promote (or occasionally to destroy) coexistence when mortality rates are dynamical variables responding to prey abundances. We next explore several modifications of this model which illustrate the rich repertoire of dynamical behaviors made possible when predation varies dynamically in response to prey abundances.
B. Numerical and Functional Responses

The rate of mortality imposed by a predator reflects both predator numbers and the attack rate per predator. The total rate of mortality is \( mN = fP \), where \( P \) is predator density and the parameter \( f \) is the number of prey (of the focal species) captured per individual predator. Because predators consume prey, the rates of the demographic parameters (birth, death, and movement) of predators will often vary as a function of the abundance of prey. Thus, the number of predators should depend on prey numbers. This is the numerical response. The rate at which an individual predator captures prey of a given species should also depend on the number of prey that are available, typically (although not always) increasing with prey abundance but saturating at high prey numbers. This is the functional response. It is useful to express the functional response as \( f = aN \), where \( a \) is the attack rate per predator per prey—the risk of mortality an individual prey face from an individual predator. Different predators have distinct numerical and functional responses and therefore will have different impacts on species coexistence.

C. Specialist Predators and Biodiversity

Specialist predators (whose diets are restricted to single prey species) typically reduce abundance of their favored species, freeing up resources for nontarget species. This can facilitate coexistence if dominant competitors attract more specialist predators or parasites than do subordinates. This diversifying effect of specialized predation can be mimicked in Eq. (1) by letting the attack rate on each species be a function of the abundance of a specialist predator, \( mN = aP_N \), where \( P \) is the abundance of predator species \( i \), which is specialized in its foraging just to prey species \( i \). To complete the model, we need an equation for the dynamics of each predator: \( dP_i/dt = P_i[g(N_i, P_i)] \), where \( g \) is a function which increases with \( N_i \) (e.g., because predators convert prey consumption into births) but may decrease with \( P_i \) (e.g., because predators interfere with each other). For any biologically reasonable system, at low numbers of their required prey specialist predators must decrease (i.e., \( g < 0 \)). Hence, the growth rate of prey species \( i \) when it is rare and other prey are resident will involve only the intrinsic growth of prey \( i \), discounted by competition with the resident species. Because the residents continue to be attacked by their own specialist predators, their numbers will be depressed below carrying capacity, reducing competition imposed on the rare focal species. Thus, specialized predators are expected to facilitate prey coexistence.

An interesting twist is that specialist predator–prey interactions are often unstable when the predator is effective at limiting prey numbers because of time lags in the numerical response of predators to prey. This has two consequences for species coexistence. First, as shown by Peter Abrams and others, with saturating functional or numerical responses instability tends to depress average predator numbers and thereby increase average prey numbers. This increases competition, making coexistence more difficult. Second, with unstable dynamics between a resident specialist predator and its prey, there will be times when that prey is rare and an inferior competitor can invade and temporarily persist, only later to be competitively excluded when the predator is rare and the dominant competitor has rebounded in numbers. Thus, unstable specialist predator–prey dynamics induces instability in community composition as well.

These effects are believed to be particularly important when considering impacts of insect herbivores on plant communities. However, most predators in the examples discussed in this article tend to have generalized diets.

D. Generalist Predators and Biodiversity

1. Switching and “Enemy-Free Space”

The expectation that specialist predation helps competing species to coexist depends on the very general and reasonable assumption that specialist predators will have numerical responses to their prey. Generalist predators can have a wide range of effects on prey communities. A single, effective generalist can act like a whole suite of specialists in promoting prey coexistence if the predator ignores whichever prey species is temporarily the rarest, concentrating attacks on common prey—the mode of foraging behavior called switching. As shown by Roughgarden and Feldman, switching predators can readily prevent competitive exclusion. In our formulation for mortality due to predation, for example, on prey species \( i \), \( mN_i = a_iP_N \), we can represent switching with an attack rate \( a_i = a(N_i, N_i, \ldots) \), where \( a_i \) declines toward zero as \( N_i \) approaches zero but increases if other \( N_i \) decrease. In effect, a prey species may persist because of a refuge in relative rarity, as defined by the predator’s behavioral responses. Because the predator is also reducing the abundance of potential competitive dominants, this is a potent mechanism for predator-mediated coexistence. If such a predator is removed, numerous prey species may risk extinction.
This idea is intuitively appealing, but there are surprisingly few demonstrations of it in natural systems. Those that do seem to involve prey species that are found in different habitats. A behavioral rule that predators leave patches with few prey, and aggregate in patches with numerous prey, leads to switching and will tend to foster prey coexistence. If prey species are spatially segregated, they are not likely to be strongly competing in any case. Other potential cases of switching seem to involve prey species with very different strategies for blending into the background, different activity times, or different behavioral tactics for escaping predation. John Lawton suggested that such species differences promoting coexistence be viewed as partitioning of enemy-free space, an aspect of niche differentiation.

In any case, several examples of species exclusion caused by predation discussed elsewhere in this article show that predators often do not tend to ignore rare prey species but rather continue attacking even to the point of extinction.

2. Saturating Functional Responses

All predators have a maximal capacity for attacking prey that is set by limited time or gut capacity. Time or effort expended in attacking one prey will be unavailable for attacking other prey. If predator numbers are fixed, an increase in abundance of one prey species may reduce attacks on another. As with switching, we can represent this as $a = a(N_1, N_2, \ldots)$, but now attacks decline with the abundance of each species. This inverse density dependence has two consequences. First, considering the effect of a prey species on itself, its mortality will decline with increasing abundance. This inverse density dependence can lead to alternative stable states for a given prey species—one at low and another at high densities. The low-density equilibrium may even be at zero density. Second, an increase in any prey species can reduce attacks on other prey species; this in effect makes alternative prey indirect mutualists. Such mutualism is most likely when considering predators that are constrained in their numerical responses (e.g., due to long generation lengths relative to those of their prey).

Unlike switching, saturating functional responses are universal. The existence of indirect mutualisms arising because predator functional responses can be swamped may explain many natural phenomena, such as herding, mixed-species flocks of birds and schools of fish, and synchronized mass emergences and migrations. One consequence of importance to biodiversity is that if some prey species are reduced in abundance, there is an immediate increase in predation on other, alternative prey, which thus risk extinction.

3. Numerical Responses and Apparent Competition

The potential for indirect mutualism via the functional response is often offset by a kind of indirect competition via the numerical response. Just as consumer species can reduce each other's abundance by depleting a shared limiting resource, alternative prey species can indirectly depress each other's abundance by increasing the abundance of a shared predator. This indirect interaction is known as apparent competition. It is particularly likely if predator population growth rates increase with the abundance of each prey type in the predator's diet and predator numbers are not strongly limited by other factors (e.g., territoriality or higher order predators). It is also more likely for predators with short generation lengths, not greatly exceeding those of their prey, or for predators which are highly mobile and can quickly aggregate into habitats with unusually high prey densities.

When any given prey species is rare, an increase in predator abundance will usually lead to an increase in its mortality. Predator abundance is expected to increase with the productivity and availability of alternative prey. The rate of predation on a focal prey is determined by the indirect, cumulative impact of alternative prey, sustaining the predator population at densities higher than allowed by the focal prey. Particularly when prey species are not strongly competing, the negative indirect interaction of apparent competition can limit prey species diversity.

When a prey species is rare, its rate of population growth can be represented as $\frac{dN}{dt} = rN - aP$. As noted previously, often the presence of alternative prey reduces the attack rate on a rare species. However, these same prey determine the magnitude of $P$. The net effect of prey on each other can only be determined by analyzing specific models. However, some general points are worth making. A given prey species is excluded by predation if $0 < r < aP$. Prey species with low $r$ are particularly vulnerable to exclusion by shared predation, as are species which have high attack rates. Another way of stating the exclusion criterion is that the maximal predator density which this prey can tolerate is $\frac{r}{a}$. All else being equal, prey with low values for $\frac{r}{a}$ are vulnerable to exclusion (note that a prey species with a high value for $\frac{r}{a}$ can sustain a high abundance of a generalist predator, which can then with impunity overexploit alternative prey with lower $\frac{r}{a}$). The upshot of these observations is that there is a tendency toward
If different prey species occupy different habitats, and predators have limited mobility, this too can prevent exclusion via shared predation. However, mobile predators can be sustained by productive prey in one habitat and with impunity can overexploit prey in low-productivity habitats. A serious effect of habitat fragmentation is that it exposes species in habitat remnants to predation from generalist predators sustained by alternative prey in the surrounding landscape. Most examples of dramatic prey limitation by predation seem to depend on the availability of alternative prey, which permit predator numbers to remain high. The brown rat snake on Guam can persist on a diet of rats and lizards, which permits it to eat out of existence every native bird species.

Previously, we discussed the cascading effects of top predator removal. Such removals shift the factors regulating intermediate mesopredators or herbivores, which will increase and become more regulated by food availability than they were in the past. This can unleash strong apparent competition effects at lower trophic levels. Prey species harmed by polyphagous predators via “mesopredator release” are victims of apparent competition.

E. Generalist Predators and Community Stability

Generalist predators can have many different effects on the overall stability of communities. Here, we discuss a few interesting effects.

Richard Vance examined a Lotka–Volterra model akin to Eq. (1) but with a predator showing numerical responses to each prey species. Even if the pairwise interactions were all stable, the entire ensemble could show large-amplitude cycles, or even chaotic dynamics. If generalist predators are mobile and seek out patches with high prey abundance, this can lead to switching. If these responses are rapid, then generalist predators help stabilize prey dynamics. However, there can be time lags in these responses, which in turn can be destabilizing. Recent theoretical explorations suggest that switching behavior in patchy environments leads to systems which persist but which have bounded oscillations. Moreover, in changing landscapes, mobile generalist predators can concentrate in habitat remnants, leading to transient spikes of high predation and extinction risk for prey species residing in these remnants.

Generalist predators usually have a mixture of a few strong interactions and many weak interactions with the species of prey in their diet. Theoretical studies have recently demonstrated that weak interactions can help reduce inherent instabilities in strong predator–prey interactions. However, the effect depends on detailed assumptions made about predator–prey feedbacks; for instance, if prey flow into a given habitat and contribute to the diet of a resident predator, this predation does not feed back to influence prey numbers in the source habitat. If the predator instead is highly mobile, its ability to feed in multiple habitats may permit severe overexploitation of prey in some habitats.

Given saturating functional responses, in unstable systems apparent competition can be reduced relative to that in stable systems because the nonlinearity in the functional response means that predators are harmed more by times of low prey abundance than benefited by times of high prey abundance. This reduces the number of predators which can be sustained, thereby weakening apparent competition effects. The dominant effect may then be competitive exclusion (if the prey are strong competitors) or indirect mutualism (with noncompeting prey and saturating functional responses).

The theoretical studies summarized here suggest that there is no universal effect on predation on biodiversity but rather many effects. More important, they help emphasize potential surprising effects of predator removal and highlight the range of information a conservationist needs to gauge the likely impacts of management alternatives. Most theoretical studies of predation have been limited to interactions across two trophic levels. More complex theoretical studies of trophic interactions are needed to investigate such phenomena as trophic cascades.

VI. GENERALITY

There is much evidence for the influences of predators on species, populations, communities, and ecosystems. However, how predictable and widely occurring are
PREDATORS, ECOLOGICAL ROLE OF

these effects? To what degree do predators regulate the structure and function of the planet’s ecosystems relative to other biological interactions and physical forces?

These important questions can be asked at two levels—within and between ecosystems. Within systems, the question of generality usually concerns the geographical extent of an effect that has been demonstrated at one or several sights. Michael Foster raised the issue for the influences of sea stars in mussel beds and sea otters in kelp forests, claiming that the interactions were less common than generally believed. Bruce Menge has begun to deal with the problem for sea stars and mussel beds by identifying some of the conditions along rocky shores under which sea star predation effectively limits or regulates mussel populations. Not surprisingly, sea star predation is important at many sites but not everywhere. For sea otters and kelp forests, Estes and Duggins approached the question by evaluating how consistently predictions of the otter–kelp forest paradigm played out at many sites with and without sea otters. In this case, the predictions held up (i.e., otter-dominated sites supported kelp forests and otter-free sites were deforested by sea urchin grazing) nearly everywhere they examined from southeast Alaska to the western Aleutian archipelago (Fig. 8), and similar results have been obtained from British Columbia. However, as mentioned previously, these patterns are less general in southern California kelp forests.

Across ecosystems, the question of generality for top-down influences of apex predators becomes one of both relative importance and variation in process. We know that trophic cascades are not limited to aquatic systems, as earlier suggested by Donald Strong. Nonetheless, understanding whether ecosystem function is controlled by trophic cascades involving a few key species (the HSS and keystone species models) or by a greater complexity of food web interactions within and across systems (a view espoused by Polis and Strong) remains one of ecology’s most daunting challenges.

VII. THE FUTURE

Although predators have long concerned conservationists and resource managers, the future will bring heightened attention to this group. As a general rule, apex predators are more vulnerable to local extinctions than are lower trophic-level species. The rapid demise of predators has occurred in part because they have been (and in many areas still are) treated as competitors of humans for fish, wildlife, and agricultural resources; thus, they have been persecuted rather than conserved.

Habitat fragmentation has also hastened local extinction rates of large apex predators because their typically low densities and large home ranges render the smaller fragments incapable of maintaining viable populations. These facts, together with the increasing realization that predators are often essential for maintaining native ecosystems, is leading to a paradigm shift in conservation biology, especially in the area of reserve design. Earlier approaches to conservation planning focused on preserving representative habitats. Many ecologists now believe that this approach, although necessary, is not
sufficient. Large reserves, or a series of smaller connected reserves, are necessary to maintain viable populations of predators, which in turn are essential for maintaining the functional integrity of these systems. This view has important implications to restoration ecology, which until recently has focused mainly on the reintroduction of native plants and the elimination of exotic species.

If the maintenance or restoration of native predators is important to conservation biology, so is the elimination of exotic predators. Exotic predators have devastated many natural biotas, both because of their ability to reduce or exterminate native prey species and the many indirect effects of these prey throughout their food webs. As discussed previously, island biotas provide the most obvious and poignant examples of such effects. The removal of exotic species is usually expensive, time-consuming, and wrought with technical challenges. Nonetheless, exotic predators, because of their great mobility and low population density, are often easier to remove than other invasive species.

The conservation and management of predators requires more and better information than is currently available. As recently noted by George Schaller, field studies have been conducted on less than 15% of the species of mammalian carnivores. Although this is indeed a feeble record, there are even fewer successful efforts to understand the ecological importance of this group.

A final need is for greater understanding of the full range of food web effects by predators. Most prior studies of predator effects focused on predator–prey interactions, thus creating the overly simplistic mind-set still held by many wildlife and fisheries managers. Most studies of indirect effects take this perspective only a step further by focusing on trophic cascades. However, as discussed previously, trophic cascades are expected to mediate competitive interactions and influence the strength of bottom-up forces by altering production levels at the base of the food web. Little is known about such interactions, although several examples of long and complex chain reactions among species show that they can be tremendously important to ecosystem function.

For the most part, food webs and ecosystem dynamics have been studied by one of two approaches. The oldest of these involves descriptions of food web structure and estimating the transfer of materials and energy through their various linkages, the critical assumption being that flux rate reflects interaction strength. The fallacy of this assumption is evident in the fact that strong top-down forces necessarily reduce prey abundance to such low levels that the interaction may no longer be apparent in a static food web. The second approach is to observe these dynamics directly by perturbing the system. Unfortunately, most experimental studies have been limited to processes acting on small spatial and temporal scales (for obvious logistical reasons) and to invertebrates and heterothermic fishes (for social and political reasons). Our general lack of understanding of the larger apex predators (especially large...
mammals) is due in large part to the almost complete absence of experimental evidence.

Creative approaches are needed to put the ecological influences of large apex predators into a dynamical context. The use of historical information and opportunities provided by "natural experiments" holds promise in this regard. The historical record is a rich source of information that has barely been utilized. Assessments of the wolf/moose/boreal forest system from tree ring analysis and the sea otter–kelp forest system from faunal remains in Aleut kitchen middens illustrate the utility of this approach. Fortuitous change is another potential means of assessing the ecological roles of large apex predators. Studies by Terborgh on the recently formed islands of Lago Guri, by Crooks and Soulé on mesopredator release in urban habitat fragments, and by Estes and Palmisano on sea otters and kelp forests demonstrate the value of this approach. As these and other examples show, ecologists and resource managers must remain mindful of appropriate scales of time and space for both learning about predators and applying that knowledge to the conservation of biodiversity.

Acknowledgment

We are deeply grateful to Stacey Reese for helping prepare the manuscript.

See Also the Following Articles

CARNIVORES • COMPETITION, INTERSPECIFIC • FOOD WEBS • KEYSTONE SPECIES • PARASITISM • POPULATION DYNAMICS • SPECIES INTERACTIONS • TROPHIC LEVELS

Bibliography


At the beginning of the twenty-first century, we contemplate environmental changes that may be wrought by global warming and ozone layer depletion and brace for mass extinctions of plants and animals predicted by many of the world’s leading biologists. As primatologists, we contemplate the subjects of lifelong study and wonder how many of the world’s prosimians, monkeys, and apes will survive into the next millennium since efforts to stem the loss of their tropical forest habitats have met with only minimal success.

I. INTRODUCTION

During the past few decades, interest in nonhuman primates has increased significantly on many fronts and has helped build support for conservation. Pioneering and long-term field studies of the great apes (Schaller, 1963; Goodall, 1968; Fossey, 1983; Galdikas, 1995) have dispelled age-old myths about mankind’s closest living relatives, greatly narrowing the gaps between these species and our own, and provided new insights into human origins and behavior. The continuing search for drugs to treat global maladies such as malaria, cancer, and AIDS has required large numbers of nonhuman primates as experimental subjects (Mack and Mittermeier, 1984). For some species, this use has contributed to serious declines of wild populations and ultimately forced issues of conservation and captive breeding as part of the long-term strategy for biomedical research. In other cases, little known and formerly obscure primate species of the neotropics, Africa, and Asia have emerged as prominent “flagships” for conserving their tropical forest habitats, which biologists agree are the richest natural terrestrial ecosystems on the planet. Although interest in nonhuman primates is increasing, the threats to their survival persist. Varying combinations of habitat destruction, hunting, and live capture have driven dozens of primate species to the brink of extirpation.

GLOSSARY

anthropoids “Higher” primates such as marmosets, tamarins, monkeys, and apes.
diversity The variety of species within a given taxonomic group.
endemic Unique to or occurring only in a specified geographic location or region.
prosimians “Lower” primates such as lemurs, lorises, and galagos.
extinction, to the point that several taxa now number only in the low thousands and a few no more than a few hundred individuals. Such populations are doomed without long-term protection, monitoring, and a heightened understanding of their plight by local human populations.

In the face of continuing threat, however, primate conservationists can also reflect upon the past century and realize with some degree of pride that, to the best of anyone’s knowledge, not a single primate taxon went extinct during that period. The next few years could conceivably witness the loss of such primates as the Tonkin nub-nosed monkey (Rhinopithecus avunculus) of Vietnam or Msas Walldron’s red colobus (Procolobus badius walldoni) of West Africa, both critically endangered and the latter has not been located during several recent surveys. In the new millennium, the survival of these and other threatened taxa will depend on the continuation of existing conservation programs and the establishment of new ones according to a global strategy for the most endangered. Fortunately, support for global primate conservation increased in the latter half of the 1990s after having suffered something of a dry spell, and the expertise is at hand to direct available resources to the highest priority species, habitats, and projects.

II. OVERVIEW OF GLOBAL PRIMATE DIVERSITY AND CONSERVATION

The order Primates represents one of 21 mammalian orders that together total at least 4675 species. It includes 13 families, 63 genera, and approximately 630 taxa worldwide. Two suborders of primates are recognized, the Prosimii (prosimian or lower primates) with 7 families and the Anthropoidea (higher primates) with 6 (Table I).

Living prosimians occur only in the Old World, despite the fact that North America once represented a major center of their early evolutionary history. Of the

<table>
<thead>
<tr>
<th>Suborder Prosimii</th>
<th>Family</th>
<th>Common name(s)</th>
<th>Distribution</th>
<th>Genera</th>
<th>Species</th>
<th>Taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cheirogaleidae</td>
<td>Dwarf, mouse, and fork-marked</td>
<td>Madagascar</td>
<td>5</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Megaladapidae</td>
<td>Sportive lemurs</td>
<td>Madagascar</td>
<td>1</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Lemuridae</td>
<td>Ring-tailed, gentle, brown,</td>
<td>Madagascar</td>
<td>4</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>crowned, red-bellied,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>mongOOSE, and rufified</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Indriidae</td>
<td>Indri, avahi, and sifakas</td>
<td>Madagascar</td>
<td>3</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Dusphotemusidae</td>
<td>Aye-aye</td>
<td>Madagascar</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Loriniidae</td>
<td>Angwantibos, lorises, pottos,</td>
<td>Africa, Asia</td>
<td>8</td>
<td>20</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and galagos</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tarsiusidae</td>
<td>Tarsiers</td>
<td>Asia</td>
<td>1</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Suborder Anthropoidea</th>
<th>Family</th>
<th>Common name(s)</th>
<th>Distribution</th>
<th>Genera</th>
<th>Species</th>
<th>Taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Callitrichidae</td>
<td>Marmoset, tamarins, and</td>
<td>Neotropics</td>
<td>3</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Cebidae</td>
<td>Owl, titi, square, tamarin,</td>
<td>Neotropics</td>
<td>11</td>
<td>62</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td></td>
<td>capuchins, squirrel, sae,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>boarded sae, howling,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>spider and woolly monkeys, and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>tamarins and mur-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ceropithecidae</td>
<td>Macaques, baboons, mangabeys,</td>
<td>Africa, Asia</td>
<td>19</td>
<td>101</td>
<td>273-283</td>
</tr>
<tr>
<td></td>
<td></td>
<td>genets, colobus, leaf</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>monkeys, langurs, and shub-nosed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>and proboscis monkeys</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hylobatidae</td>
<td>Gibbons and siamangs</td>
<td>Asia</td>
<td>1</td>
<td>11</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Pongidae</td>
<td>Chimpanzees, pygmy chimpanzees,</td>
<td>Africa, Asia</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gorillas, and orang-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>utans</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hominoidea</td>
<td>Man</td>
<td>Global</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

| Total                |                         |                                 |              | 40     | 273     | 622–632 |
seven extant prosimian families, five (Lemuridae, Cheirogaleidae, Megaladapidae, Indridae, and Daubentoniidae) occur naturally only in Madagascar, where they are represented by 14 genera and at least 33 species and 51 taxa. The family Tarsiidae is only slightly more geographically widespread, represented in Indochina and the Philippines by 1 genus, 3 species, and 11 taxa. The 8 genera, 20 species, and 44 taxa of the Lorisoidea are distributed throughout mainland Africa, India, and Southeast Asia.

The anthropoid primates are a much more diverse group than the prosimians, having almost twice the number of genera and four times the number of species and total taxa. Of the six anthropoid families, two (Callitrichidae and Cebidae) occur only in the New World tropics, two (Cercopithecidae and Pongidae) occur throughout much of Africa and Asia, one (Hylobatidae) is restricted to Asia, and the last (Hominidae), represented by our own species, is global in its distribution. New World nonhuman primates comprise 16 genera, 99 species, and 205 taxa. By comparison, the Old World monkeys (subfamilies Cercopithecinae and Colobinae), although much more widely distributed, are only somewhat more diverse, comprising 19 genera, 101 species, and 273–283 taxa overall. The lesser apes, which include the gibbons and siamangs, are a relatively small group that includes a single genus, 11 species, and 28 taxa. The great apes are even less diverse with 3 genera, 5 species, and only 9 taxa overall. Finally, the human primate is the sole representative of a morphologically and ecologically variable but monotypic family. Accordingly, for the remainder of this article, we use “primate” only in reference to nonhuman taxa unless otherwise specified.

Examining global primate diversity from a regional perspective, it can be seen that it is by no means evenly distributed. For example, the remaining tropical forests of Madagascar are dwarfed by the extensive tropical forests of Africa, Asia, and Central and South America, but Madagascar has by far the densest concentration of primate diversity anywhere on Earth within an area of 587,045 km². Although usually considered part of the Ethiopian zoogeographical region, Madagascar is in many respects a zoogeographic region by itself in that the entire region harbors a significant 38% of global primate diversity at the species level and an even more impressive 54% of higher order (beta) diversity at the generic level.

Within the neotropics, primates occur from southern Mexico through Central America and northern South America and as far as southern Brazil, northern Argentina, and Paraguay (but not Chile). Native primate populations also occur on Trinidad; there are several introduced populations of African primates on the islands of St. Kitts, Nevis, Barbados, and Grenada; and both New World and Old World species have been introduced to the island of Puerto Rico. Sixteen genera and 203 taxa are found within the neotropics—amounts comparable to those for Africa. However, the 99 primate species of the neotropics are the most for any single major region, account for 36% of global primate diversity, and are approximately equivalent to the number of species inhabiting both Africa and Madagascar.

Asian primates are found mainly in the Oriental zoogeographical region and in the southeastern portion of the Palearctic and in Wallacea (the transition zone between the Oriental and Australian regions). In southern Asia primates are widely distributed from the Indian subcontinent and the island of Sri Lanka, throughout Southeast Asia as far as the Philippines and the Indone- sia islands of Halmahera and Sulawesi, to central and northern Asia from Afghanistan through southern China (including the islands of Hainan and Taiwan) to Japan. In contrast to Africa (excluding Madagascar) and the neotropics, where primates are basically continental, Asian primates are found in large numbers on islands as well. The region is home to 13 genera, 71 species, and 176 taxa, a slightly lower level of generic diversity by comparison to the neotropics and Africa but a level of species diversity approximately equal to that of Africa.

Ninety-two of the world’s 192 sovereign nations have wild primate populations; the 7 countries with the highest primate species diversity are listed in Table II. Brazil, with 77 species, is by far the leader and accounts for slightly more than three-fourths of all Neotropical primate species. Together, the top four countries (Brazil, Democratic Republic of the Congo, Indonesia, and Madagascar), which also represent the world’s four major primate habitat regions, account for 181 or 182 species, or approximately two-thirds of all living primates.

Furthermore, three of the top four countries for primate diversity (Madagascar, Brazil, and Indonesia) also are at the top of the list of the world’s top countries.
TABLE II
World’s Top Seven Countries for Primate Species Diversity (≥30 Species)

<table>
<thead>
<tr>
<th>Country</th>
<th>Species</th>
<th>Genera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>37</td>
<td>16</td>
</tr>
<tr>
<td>Indonesia</td>
<td>36</td>
<td>9</td>
</tr>
<tr>
<td>Madagascar</td>
<td>34</td>
<td>14</td>
</tr>
<tr>
<td>Democratic Republic</td>
<td>33–34</td>
<td>18</td>
</tr>
<tr>
<td>of the Congo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cameroon</td>
<td>31</td>
<td>18</td>
</tr>
<tr>
<td>Colombia</td>
<td>31</td>
<td>12</td>
</tr>
</tbody>
</table>

Furthermore, if we extend the analysis to consider sub-
species, we find that other countries will rise to the top
of the priority list. Consider India for example. With
only 15 primate species and three endemics, it is well
behind many other countries on the previous lists. How-
ever, of India’s 34 or 35 primate taxa, 22 are endemic,
and the additional 19 endemic subspecies represent
important wild populations that should be considered
when establishing conservation priority rankings.

III. THREATS TO PRIMATES

Threats to the survival of nonhuman primates are easily
divided into three major categories: habitat destruction,
hunting for food and a variety of other purposes, and
live capture for export or local trade (Mittermeier et al., 1986). The effects of these threats vary significantly
from species to species and from region to region and
are influenced by the extent of remaining habitat, hu-
man population densities, the presence of roads and
large rivers, the nature and degree of human activity
within the range of a particular species, local hunting
traditions, the size and desirability of different species
as food items or as sources of other products useful to
man, the demand for a given species in research or the
pet trade, enforcement of existing wildlife laws, and
regulation of commercial animal dealers. However, one
or more of the three major threats affect almost all
primate populations.

A. Habitat Destruction

On a global scale, habitat destruction is the principal
factor contributing to the disappearance of wild primate
populations. The continuing growth of the human pop-
ulation and its ever-expanding need for natural re-
sources have contributed greatly to the destruction or
alteration of natural habitats on an almost unimaginable
scale, and nowhere has this problem been more acute
than in the tropical regions of the world. More than 90% of
all nonhuman primates inhabit the tropical forests of
Africa, Asia, and South and Central America, and these
forests are being cut at a rate of more than 10 million
ha per year (Bryant et al., 1997).

The immediate effects of habitat destruction on non-
human primates vary significantly from one region to
the next. For example, in Madagascar and the Atlantic
forest region of eastern Brazil, so little suitable forest
habitats remains that any further loss constitutes a grave
threat to primates and other wildlife. In contrast, in the
vast forest regions of Amazonia and the Congo Basin,
which along with the island of New Guinea represent two of the three remaining major tropical wilderness areas of the planet, the effects of habitat destruction are only starting to be felt.

**B. Hunting for Food and Other Purposes**

The hunting of primates by human populations occurs for a variety of reasons, but by far the most important is to acquire food. Although primate hunting is prohibited by law in many countries, enforcement of such protective legislation is typically rare and sometimes nonexistent in the remote areas where this activity almost always occurs.

Hunting of primates as a source of food is a significant threat in at least three parts of the world: the Amazon region of South America, west Africa, and central Africa. In each region, primates are among the animals most frequently hunted and they are regularly sold in markets, except where this is prohibited by law. However, even in areas in which primate hunting is common, it by no means affects all species equally. In Amazonia, for example, the larger monkeys, such as Lagothrix, Ateles, Alouatta, and Cebus, are heavily hunted and among the most desirable food species, whereas smaller monkeys such as Saguinus and Saimiri are rarely shot for food because they barely provide enough meat to recompense the hunter for the cost of his or her shotgun shell. As a result of better law enforcement in recent decades in countries such as Brazil, bushmeat hunting is no longer considered a serious threat to wild primate populations. The situation in central and west Africa, in contrast, has quickly reached a critical stage because bushmeat hunters in these regions are being paid by logging companies to shoot large quantities of primates and other forest wildlife to feed work crews.

In areas where the hunting of primates for food is common, it can sometimes represent a threat even more severe than forest destruction. For example, in some parts of Amazonia there are large tracts of primary forest remaining in which populations of Lagothrix, Ateles, Alouatta, and Cebus have effectively been exterminated by excessive hunting (Mittermeier and Coimbra-Filho, 1977, Sotini, 1982). This is sometimes referred to as the “empty forest syndrome” and appears to be exactly what is currently happening in central and west Africa, where it includes both monkeys and apes. Of course, in areas where food hunting and deforestation are both prevalent, populations of all forest primates and other game species can disappear quickly.

It is important to note that, in some parts of the world, religious restrictions or other cultural factors prohibit (or inhibit) the killing and eating of primates. In India, for example, primates are rarely hunted for food because they are linked to the monkey god Hanuman, which occupies an important role in the Hindu religion, whereas in strictly Muslim countries primates are not eaten because their flesh is considered unclean and unfit for human consumption. Indeed, in India, Hindu people refuse to kill rhesus monkeys and resist translocating them even when populations become so high that they constitute a menace to humans. In other countries, such as Madagascar, local taboos may exist against eating certain primates (e.g., Indri), whereas other species (e.g., Eulemur and Varecia) may be the most popular food items for a given tribe or village.

Primates are also hunted to supply many other products in addition to food: traditional medicines; bat, body parts for ornamentation, and trophies. Primate hunting to supply medicinal products may be nothing more than a by-product of food hunting in most cases, and it usually involves the use of specific body parts for their supposed medicinal value. In south India, for example, the meat of the Nilgiri langur (Presbytis johnii) and the endangered lion-tailed macaque (Macaca silenus) is regarded as an aphrodisiac and thought to contain other medicinal properties. The blood of leaf monkeys, such as Phayre’s langur (Presbytis phayrei) in Thailand, is believed to impart vigor to the drinker, especially when mixed with local whiskey. Also, in various South American countries, drinking from the cup-shaped hyoid apparatus of an adult male howling monkey (Alouatta) is reported to cure goiters and stuttering as well as to ease a mother’s labor pains during childbirth. Although the hunting of primates for medicinal purposes is considered a relatively minor factor overall in the global decline of wild primate populations, when it involves endangered species, such as in the case of India’s lion-tailed macaque or some of the Southeast Asian colobines (Trachypithecus and Pygathrix), it can be a serious threat.

Primates are also shot to provide bait for capturing and killing other animals, mainly in remote areas of the Amazon region. There, spotted cat hunters preferentially shoot larger monkeys such as Lagothrix and Ateles to bait crude wooden traps set for jaguars and ocelots. Any number of Amazonian primates may also be shot for fish or turtle bait, and in Sri Lanka monkeys often serve as bait for crocodiles. Although the use of primates for bait is a relatively minor threat, it can and does add to the pressures faced by overexploited, large-bodied taxa such as Lagothrix and Ateles.

In some countries, primates may be killed for their...
soups or to provide other body parts used in ornamentation. Perhaps the most striking case of this is in Africa, where the skins of black-and-white colobus (Colobus guereza) and related species have been used to fashion cloaks and headdresses for native African peoples but have also figured significantly in the international fur trade. For example, in 1899 a reported 223,999 monkey skins were auctioned in London, and at least 2.5 million were probably exported to Europe between 1860 and 1900, especially to Germany, where they were used to make capes, mufffs, and rugs. As recently as the early 1970s, colobus monkey rugs were common in East African tourist shops and colobus coats were being sold in Europe and Japan.

Throughout much of Amazonia, tourist shops still offer stuffed monkeys, monkey skulls, monkey-skin hats, monkey-tail dusters, and necklaces fashioned from monkey teeth, bones, hands, feet, and tails. However, these products are typically available on a small scale and almost always as a by-product of hunting for food.

Nonetheless, the demand for primate body parts for sale to tourists can be a serious matter if it involves endangered species. The most striking example of this is the slaughter of mountain gorillas (Gorilla gorilla beringei) in Rwanda and the Democratic Republic of the Congo (formerly Zaire) which produces hands and skulls for sale to European tourists (Fossey, 1983). Although rare, this practice still occurs despite effective, long-term conservation programs in this region.

Hunting primates for sport is fortunately rare and a minor threat to wild populations. It appears to be most prevalent near logging camps and within military zones in remote areas of developing countries, where arms are plentiful and law enforcement is basically nonexistent. Children armed with slingshots and air rifles are often among the worst offenders. More prestigious trophy hunting has also played a role (albeit a minor one) in primate decline. Species such as the gorilla were especially desirable quarry for nineteenth-century and early twentieth-century trophy hunters, and the tales of their exploits are recounted in many books. On the whole, however, sport hunting must be considered a very minor factor unless an endangered species is involved, in which case the activity is almost always illegal as well.

A final reason for hunting primates considered here is because they are sometimes considered agricultural pests; for some African and Asian species, this can represent a significant drain on wild populations. The most striking example is that of government-sponsored “monkey drives” that were common in Sierra Leone several decades ago. Eleven of the country’s 14 primate species were routinely shot or driven into nets and clubbed to death during such drives; only 3 species were considered harmless to farm crops. According to government records, approximately 230,000 monkeys were destroyed in such drives between 1949 and 1952, and these were only the ones actually counted. Bounties were paid for primate heads or tails, and there was no control over the species killed.

The major crop raiders are usually the more adaptable and widespread species, such as the savanna baboons (Papio spp.) in Africa and the macaques (Macaca spp.) in Asia, but there are also instances of orangutans being killed for raiding fruit trees and gorillas being killed for destroying crops. The only Neotropical species regarded as agricultural pests are the capuchins (Cebus spp.), whose common names sometimes reflect their crop-raiding habits. For example, the common name for the tufted capuchin (Cebus apella) in Colombia is maicero and one of the Surinamese names for the weeper capuchin (Cebus olivaceus) is nyan-karu mongi, both of which translate as “corn-eater.”

It is difficult to assess how much damage primates actually do to crops in different parts of the world. It is equally difficult to determine how effective pest control efforts have been or to what degree they have contributed to the decline of wild primate populations. However, as primate habitats continue to be encroached upon, resulting in shortages of other food sources, it is likely that the more adaptable primate species will continue to raid crops and perhaps become more dependent on them as a food source. This, unfortunately, will likely result in increased conflict between man and nonhuman primates.

C. Live Capture of Primates

Primates have routinely been captured alive for export (the international trade to supply zoos and for biomedical research and pharmaceutical testing) or to serve local pet trades. The height of the international primate trade began at the end of the 1950s and continued through the early 1960s, during which time hundreds of thousands of monkeys were taken from the wild each year (Mack and Mittermeier, 1984). The trade consisted largely of rhesus macaques (Macaca mulatta) exported from India and used in laboratory tests as part of the effort to develop a vaccine for polio and squirrel monkeys (Saimiri sciureus) imported by the United States from several Amazonian countries. Subsequently, the imposition of export bans by habitat countries, import restrictions by user countries, and a decreased demand from biomedical research and zoological parks contribu-
uted to a significant decline in the international traffic in primates. In 1982, in recognition of the serious effect that live capture for export can have on wild primate populations, the International Union for Conservation of Nature and Natural Resources (IUCN)/Species Survival Commission (SSC) Primate Specialist Group prepared a Policy Statement on the Use of Primates for Biomedical Purposes, which includes the recommendation that endangered, vulnerable, or rare species be considered for use in biomedical research projects only if they are obtained from existing, self-sustaining captive breeding colonies. This policy statement was subsequently adopted by the World Health Organization and the Ecosystem Conservation Group of the United Nations, which includes UNESCO, the Food and Agricultural Organization (FAO), UNEP, and IUCN. It is still valid today.

IV. CONSERVATION STATUS OF PRIMATES

The most recent, comprehensive conservation status assessment of the world’s primates is included in the 1996 IUCN Red List of Threatened Animals (Baillie and Groves, 1996), a collaborative effort of the IUCN Species Survival Commission, the World Conservation Monitoring Centre, and BirdLife International. This document differs significantly from past red lists in its use of new categories and criteria for threat. All primate taxa were included in this assessment and have been identified as Threatened (a designation which includes the categories Critically Endangered, Endangered, and Vulnerable), Lower Risk: Conservation Dependent, Lower Risk: Near Threatened, Lower Risk: Least Concern, Extinct, Extinct in the Wild, and Data Deficient. In general terms, a taxon is defined as

1. Critically Endangered if the extent of its occurrence is estimated to be less than 100 km², if its population is estimated to be less than 250 individuals, and/or if quantitative analysis indicates the probability of extinction in the wild is at least 50% within 10 years or three generations.

2. Endangered if the extent of its occurrence is estimated to be less than 5000 km², if its population is estimated to number less than 2500 individuals, and/or if quantitative analysis shows the probability of extinction in the wild is at least 20% within 20 years or five generations.

3. Vulnerable if the extent of its occurrence is estimated to be less than 20,000 km², if its population is estimated to number less than 10,000 individuals, and if quantitative analysis shows the probability of extinction in the wild is at least 10% within 100 years.

As a result of this assessment, 204 (approximately one-third) of the world’s approximately 620 primate taxa are currently considered Critically Endangered, Endangered, or Vulnerable. Of these, 104 taxa (approximately 17%) are listed as Critically Endangered or Endangered—34 in the neotropics, 29 in Africa, 17 in Madagascar, and 24 in Asia (Tables IV and V). Nine genera (Allocebus, Varecia, Indri, Daubentonia, Leontopithecus, Brachyteles, Simia, Pan, and Gorilla) are considered Endangered, as is the monotypic family Daubentoniidae.

V. PRIMATE CONSERVATION DURING THE PAST 20 YEARS

The IUCN was established in 1948 to promote and carry out scientifically based action for the conservation and sustainable use of living natural resources. IUCN enrolls sovereign states, governmental agencies, research institutions, and nongovernmental organizations to conserve the world’s natural heritage. The SSC was founded in 1949 and is the largest of IUCN’s six commissions, with more than 7000 volunteer member scientists, field researchers, government officials, and conservation leaders from 88 countries. SSC works principally through its more than 100 specialist groups, of which the Primate Specialist Group is one of the largest.

The founding mission of the Primate Specialist Group is to maintain the current diversity of the order Primates, with dual emphasis on (i) ensuring the survival of endangered and vulnerable species wherever they occur and (ii) providing effective protection for large numbers of primates in areas of high primate diversity and/or abundance.

Although activities under way in many parts of the world make it inevitable that a portion of the world’s forests and primates living in them will disappear, the role of the Primate Specialist Group is to minimize this loss wherever possible by

1. Setting aside special protected areas for critically endangered, endangered, and vulnerable species

2. Creating national parks and reserves in areas of high primate diversity and/or abundance
## Critically Endangered Primates

<table>
<thead>
<tr>
<th>Critically Endangered</th>
<th>Endangered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harry-eared dwarf lemur (Lepilemur microdon)</td>
<td>White-collared lemur (Eulemur fulvus albocollaris)</td>
</tr>
<tr>
<td>Sylvester's sifaka (Propithecus verreauxi corniculatus)</td>
<td>Black-and-white ruffed lemur (Varecia variegata variegata)</td>
</tr>
<tr>
<td>Northern brown howling monkey (Alouatta fusca fusca)</td>
<td>index (Indo mspo)</td>
</tr>
<tr>
<td>Golden bamboo lemur (Hapalemur griseus)</td>
<td>Diademed sifaka (Propithecus diadema diadema)</td>
</tr>
<tr>
<td>Broad-nosed guenon (Cercopithecus ascanius)</td>
<td>Nilhe-Edward's sifaka (Propithecus diadema edwardsi)</td>
</tr>
<tr>
<td>Red-ruffed lemur (Varicocapucus rubicundus)</td>
<td>Copper's sifaka (Propithecus verreauxi capucinus)</td>
</tr>
<tr>
<td>Silky sifaka (Propithecus diademum candidus)</td>
<td>Aye-aye (Daubentonia madagascariensis)</td>
</tr>
<tr>
<td>Perrer's sifaka (Propithecus diadema perrieri)</td>
<td>Bully-tailed-eared marmoset (Callithrix aurita)</td>
</tr>
<tr>
<td>Tattersall's sifaka (Propithecus tattersallii)</td>
<td>Bully-headed marmoset (Callithrix jacchus)</td>
</tr>
<tr>
<td>Crowned sifaka (Propithecus verreauxi corniculatus)</td>
<td>Golden-hooded lemur tamarin (Lutreolum personatum)</td>
</tr>
<tr>
<td>Black-faced lion tamarin (Lagothrix lagotricha)</td>
<td>Bicolored lemur (Saguinus oedipus)</td>
</tr>
<tr>
<td>Black lion tamarin (Lagothrix lagotricha)</td>
<td>Cotton-top tamarin (Saguinus oedipus)</td>
</tr>
<tr>
<td>Northern muriqui (Brachyteles arachnoides)</td>
<td>Colobus Island howling monkey (Alouatta satanas satanas)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta satanas satanas)</td>
<td>Night monkey (Aotus lousianus griseomammatus)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>White-bellied spider monkey (Ateles belzebuth bueri)</td>
</tr>
<tr>
<td>Northern brown howling monkey (Alouatta satanas satanas)</td>
<td>Hybrid spider monkey (Ateles belzebuth hybrida)</td>
</tr>
<tr>
<td>Brown-headed spider monkey (Ateles geoffroyi geoffroyi)</td>
<td>Grizzled spider monkey (Ateles geoffroyi griseus)</td>
</tr>
<tr>
<td>Azuero's spider monkey (Ateles geoffroyi geoffroyi)</td>
<td>Panamanian spider monkey (Ateles geoffroyi panamensis)</td>
</tr>
<tr>
<td>Southern muriqui (Brachyteles hypoxanthus)</td>
<td>White-whiskered spider monkey (Ateles geoffroyi geoffroyi)</td>
</tr>
<tr>
<td>Northern Bahian brown titi (Cacajao calvus novaesi)</td>
<td>Bald uakari (Cacajao calvus calvus)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Bald uakari (Cacajao calvus novaesi)</td>
</tr>
<tr>
<td>Red-headed howling monkey (Alouatta belzebul ululata)</td>
<td>Red uakari (Cacajao calvus rubicundus)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Red uakari (Cacajao calvus rubicundus)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Red uakari (Cacajao calvus rubicundus)</td>
</tr>
<tr>
<td>Red-faced lion tamarin (Lekontopitheicus chrysomelas)</td>
<td>Black-and-white ruffed lemur (Varecia variegata variegata)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Red-handed howling monkey (Alouatta belzebul ululata)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Corb Island howling monkey (Alouatta satanas satanas)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
<tr>
<td>Golden lion tamarin (Leontopithecus rosalia)</td>
<td>Coquerel's sifaka (Propithecus diadema coquereli)</td>
</tr>
</tbody>
</table>
3. Maintaining parks and reserves that already exist and enforcing protective legislation in them
4. Determining ways in which human and nonhuman primates can coexist in multiple-use areas
5. Establishing conservation-oriented captive breeding programs for threatened taxa
6. Ending illegal and otherwise destructive traffic in primates
7. Ensuring that research institutions using primates are aware of conservation issues and the status of species they use, that they use primates as prudently as possible, and that they make every attempt to breed most or all of the primates they require
8. Creating public awareness of the need for primate conservation and the importance of primates as a natural heritage in the countries in which they occur

In late 1977, the chairman of the Primate Specialist Group, in collaboration with group members, wrote the Global Strategy for Primate Conservation (Mittermeier, 1978). This document was an attempt to organize primate conservation activities based on the highest international priorities and to ensure that limited funds available for primate conservation were put to the best possible use. The first draft of the Global Strategy included 65 projects in Africa, Asia, and South and Central America. Each project was categorized as highest priority, high priority, priority, and desirable based mainly on the status of focal species and how likely the project would be to bring about the desired conservation action. The Global Strategy quickly led to a substantial increase in funding for primate conservation activities, and in 1979 it led to the establishment of a special Primate Program and Primate Action Fund by the World Wildlife Fund–U.S. In addition to major projects supported as a result of this program, the Primate Action Fund provided rapid support for small primate conservation projects (ranging from $500 to $3000). The Primate Action Fund functioned for more than a decade, contributing several hundred thousand dollars to more than 100 projects. Other key institutions that contributed significantly to primate conservation during this period include the New York Zoological Society (now the Wildlife Conservation Society), the Fauna and Flora Preservation Society (now Fauna and Flora International), the Rare Animal Relief Effort, the Jersey Wildlife

<table>
<thead>
<tr>
<th>Region</th>
<th>Taxa</th>
<th>Threatened</th>
<th>%</th>
<th>Critically Endangered and Endangered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neotropics</td>
<td>202</td>
<td>69</td>
<td>34.2</td>
<td>34</td>
</tr>
<tr>
<td>Africa</td>
<td>190</td>
<td>41</td>
<td>21.1</td>
<td>29</td>
</tr>
<tr>
<td>Madagascar</td>
<td>51</td>
<td>35</td>
<td>66.6</td>
<td>17</td>
</tr>
<tr>
<td>Asia</td>
<td>178</td>
<td>60</td>
<td>34.1</td>
<td>24</td>
</tr>
</tbody>
</table>

Almost a decade after the Global Strategy was launched, the first regional primate conservation action plans were prepared by the IUCN/SSC Primate Specialist Group. First to be published was the Action Plan for African Primates 1986–90 (Oates, 1986), which was quickly followed by the Action Plan for Asian Primates Conservation: 1987–91 (Euey, 1987) and several years later by Lemurs of Madagascar: An Action Plan for Their Conservation: 1993–1999 (Mittermeier et al., 1992). The last plan to appear was African Primates: Status Survey and Conservation Action Plan (Oates, 1996), which is an update of the 1986 document. These action plans have effectively focused conservation activities in three or four major regions in which primates occur, and they are useful measures with regard to the success of proposed strategies.

The first vehicle for regular and effective communication among the world’s primate conservationists was the IUCN/SSC Primate Specialist Group Newsletter, which was launched in 1981. Changed to Primate Conservation in 1985, it has appeared on more or less an annual basis ever since. In addition, the four regional sections of the Primate Specialist Group subsequently began publishing their own periodic newsletters to meet the increasing need for more timely information. Asian Primates appeared in 1991, Neotropical Primates and Lemurs News in 1993, and African Primates in 1995. In combination, these newsletters have significantly increased the amount, quality, and timeliness of information available to primate conservationists throughout the world.

VI. OUTLOOK FOR THE FUTURE

In the future, there is a need to sustain conservation activities based on recommendations of the original Global Strategy for Primate Conservation and subsequent regional action plans as well as to increase the focus on those primate taxa most seriously threatened with extinction. The twentieth century ended without a single primate taxon being lost—an enviable record indeed considering the number of reptiles, birds, and other mammals known to have disappeared during this period. However, several species and subspecies are in serious jeopardy.

Much of the groundwork for developing a more focused conservation strategy has already been done. The 1996 IUCN Red List of Threatened Animals provides a good starting point for identifying the highest priority taxa. To those that are considered Critically Endangered and Endangered, we expect that the Primate Specialist Group will add several more species and subspecies and then consider factors such as taxonomic uniqueness to establish priority rankings for conservation action. The results of this analysis are expected to be presented in the Year 2000 in the form of a global action plan for the world’s most endangered primates.

With such a plan in hand, serious work can begin to amass both the human and the financial resources needed for implementation. Fortunately, several new sources of support for primate conservation have materialized during the past decade. Although the World Wildlife Fund–U.S. Primate Program no longer exists, many other traditional nongovernmental sources still offer grants for field, captive, and laboratory programs; academic institutions continue to provide funds for primate field studies that have significant conservation impact; government-supported efforts such as the Indo–U.S. Primate Project provide excellent models for international cooperation; and an increasing number of zoos have joined forces to focus on regional primate faunas, generating funds not only for captive breeding programs but also for in situ projects. In addition, at least two new significant sources of philanthropic support dedicated to primates were established in the 1990s: Primate Conservation, Inc., and the Margot Marsh Biodiversity Foundation. Together, these organizations and agencies represent the core of funding necessary to move ahead with a global action plan for the most endangered primates. It is hoped that such a plan will help uncover new sources of support as well.

See Also the Following Articles

CAPTIVE BREEDING AND REINTRODUCTION • CONSERVATION EFFORTS, CONTEMPORARY • ENDANGERED MAMMALS • MAMMALS, CONSERVATION EFFORTS FOR • NATURAL RESERVES AND PRESERVES • ZOOS AND ZOOLOGICAL PARKS

Bibliography


PROPERTY RIGHTS AND BIODIVERSITY

Susan S. Hanna
Oregon State University

I. The Context of Property Rights
II. Forms of Property Rights
III. Functions of Property Rights
IV. Evolution of Property Rights
V. Property Rights and Biodiversity Protection
VI. Conclusions

GLOSSARY

discount rate Rate at which future benefits and costs are discounted.
economic rent Payment to resources in excess of costs.
ecosystem services Contributions of ecosystem components through genetics, information, and reproduction.
externality An effect—either a cost or a benefit—external to the generating activity.
free rider One who enjoys the benefit of a good or service without paying the cost.
institutions Organizational structures that shape human interactions.
path dependence The influence of previous actions on the present state.
prisoner’s dilemma Uncertainty about others’ behavior leading to choices that are individually rational but collectively irrational.
property rights systems Bundles of property rights with their associated rules and obligations.

PROPERTY RIGHTS to natural resources define privileges and responsibilities in the use of environmental goods and services. They specify the way people are to behave toward one another as they use environmental resources. This chapter describes the form and function of property rights in general and discusses the relation of property rights to biodiversity in particular. This discussion summarizes what is known about the potential and limitations of property rights to protect biodiversity. It also examines the considerable uncertainty that exists with respect to the design of property rights for biodiversity protection.
I. THE CONTEXT OF PROPERTY RIGHTS

Property rights define the conditions that guide and control the human use of the natural environment. They establish the terms under which people use and sustain the capacity of the environment to generate a continuing flow of goods and services. Property rights are a means by which people interact with their environment and control their behavior toward one another. They embody the expectations people have about natural resources and influence the way people make resource use decisions. Property rights link humans to each other and to natural systems through these expectations and decisions.

A. Scope

Property rights to natural resources control both use and conservation. Their scope may be either individual species or areas of land and water in which species live. The scope of property rights also includes different types of use. At present, most systems of property rights are designed for direct uses, such as catching fish for food, but they may also include indirect uses, such as the right of the public to enjoy populations of whales or to protect endangered species like bald eagles. Property rights are almost never defined for unused species or for communities of species. The idea of using property rights for the protection of not only single species but also biodiversity is a new and broader application of their accustomed use.

B. Values

If resources are used, they have value. The values of ecosystems that derive from direct uses are most easily recognized—for example, the harvest of shellfish and seaweed from intertidal marine ecosystems. Their value is expressed in markets or in the subsistence they provide. But indirect use also reflects values that, although not well quantified, are important and well articulated as economic concepts. Resources have option values for the future in terms of their potential uses—for example, the future food or pharmaceutical uses of marine species. These potential uses are valued as insurance against the unknown.

Resources also have quasi-option values in future uses that may exist through preservation or be lost by irreversible actions, such as the localized genetic information in subraces of salmon that are lost as habitat is destroyed.

Finally, resources have existence values in their aesthetic or spiritual attributes, such as the enjoyment of watching an osprey catch a fish or the satisfaction of just knowing that ospreys are nesting nearby. Existence values can also be thought of as bequest values, in that the unused goods and services are available to be passed on to future generations.

An important contributor to the option value of ecosystems is ecological resilience. Resilience is the elasticity of the ecosystem; the ability of an ecosystem to absorb perturbations and continue its essential functions. Although much remains unknown about the necessary conditions for resilience, what is known is that there are keystone species that require protection and conservation to enhance both species diversity and critical ecosystem functions.

The option values of direct uses of ecosystems are relatively easy to understand and measure, because they are linked to tangible goods like seafood and timber that are traded through markets. Considerable uncertainty exists, however, about the value of services or goods that are not currently used. Measuring and accounting for the option values of ecosystem services and unknown goods is difficult because there is no market in which people can express their willingness to pay for them or be compensated for their loss.

C. People and Biodiversity

Biodiversity can be viewed in a number of dimensions, including the diversity of species, the diversity of genetic material, and the diversity of functional roles in the ecosystem. Many of the benefits of biodiversity are public goods, in which value accrues to all. The question for humans and biodiversity is one of managing the direct and indirect effects of activities in ecosystems in order to maintain the diversity of genetic, species, and functional components.

One challenge is how to provide appropriate incentives so that people find it in their interest to promote and maintain the public good of biodiversity. A second challenge is how the rights, rules, and responsibilities that constrain resource use can be expanded from single species or physical areas to multiple species that may be distributed over wide areas. A third challenge is how to make the transition from traditional single-species commodity production types of use to new types of use that accommodate the protection of species diversity through the maintenance of ecosystem services.

D. Ownership

The idea of ownership is well understood for goods produced by ecosystems. Overall, ownership is better
TABLE I  
A Classification of Pure Forms of Property Rights

<table>
<thead>
<tr>
<th>Property rights</th>
<th>Rights holder</th>
<th>Privileges</th>
<th>Responsibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open access</td>
<td>None</td>
<td>Capture</td>
<td>None</td>
</tr>
<tr>
<td>State property</td>
<td>Citizens</td>
<td>Designate management authority</td>
<td>Stewardship</td>
</tr>
<tr>
<td>Common property</td>
<td>Collective</td>
<td>Exclude nonmembers</td>
<td>Participation, maintenance</td>
</tr>
<tr>
<td>Private property</td>
<td>Individual</td>
<td>Socially acceptable users, control of access, transfer</td>
<td>Compliance with laws</td>
</tr>
</tbody>
</table>

The idea of owning ecosystem services, in contrast, is less familiar. At present, ecosystem services are either considered the property of the public or of the owner of the area in which they exist. Ecosystem services are public goods, but in neither land-based nor marine ecosystems are rights to services specifically defined. One particular ecosystem service—the genetic information embodied in species—is particularly important to the question of biodiversity. Property rights to the informational properties of resources are in only the very early stages of development and application.

II. FORMS OF PROPERTY RIGHTS

Property rights take many forms. Differences in form derive from the scope of the rights, the type of rights owner, and the privileges and responsibilities of the rights owner. The most common categories are private, public, state, and common (Table I). The types of property rights are ordered loosely along a spectrum, where the owner ranges from an individual person to no one.

A. Open Access

Open access, res nullius, is the absence of property rights and has no ownership assigned. It is open to all, with resources becoming owned only at the point of capture. The absence of rights is often, but mistakenly, referred to as “common property.” The “tragedy of the commons” metaphor is a description of the outcome of open access. It describes the use of a common-pool resource from which it is too costly to exclude people from use. In the tragedy of the commons, individuals make choices about resource use based on their own private costs and benefits. And while the benefits of their actions accrue to them alone, the costs, in terms of the effect on the resource, are social costs spread among all users. Eventually and inevitably without some control over access, the failure to account for social costs locks people into behavior that leads to resource degradation and overuse.

B. Common Property

Although the “tragedy of the commons” is a metaphor that is widely used, it does not really describe common property. Common property, res communes, is owned by an identified group of people who have the right to exclude nonowners, the duty to participate in decisions about use, and the responsibility to act as resource stewards. For example, nearshore fishing territories may be owned and managed by the residents of the adjoining coastal community. Community members as owners decide how many people have rights to use the resource, what kind of rights these may be, and what objectives they have for resource productivity.

C. State Property

State property, res publicae, is owned by citizens who assign management authority and stewardship responsibility to an agency of the state. For example, U.S. fish and wildlife resources are the property of the citizens of individual states or of the country as a whole. The citizens designate the fish and wildlife agencies of individual states and the federal government to manage these resources in their name. The agencies may grant rights of use to individual users or communities of users, but the resources remain owned by the citizens at large.

D. Private Property

Private property, res privatae, assigns ownership to named individuals including legal individuals, such as corporations, guaranteeing to these owners a bundle of rights about access and use. Although individuals have the greatest autonomy under private property, this form of property rights is also stinted by prohibitions against
unacceptable uses, such as activities that pollute. For example, individuals may own land and the fish and wildlife on that land, but the types of use to which they can put either the land or the wildlife may be constrained by law.

E. Property Rights over Goods
The property rights listed in Table I have one attribute in common: they have, in practice, been applied almost exclusively to environmental goods. We think of resources as natural capital that create value through the size of standing stock and through the flow of resources from that stock. We are familiar with property rights over the stock and flow components of natural capital. For example, in marine ecosystems rights can be assigned to the resource only on capture (competitive fishing), to the standing stock itself (territories for sedentary species), or to the flow of goods from the standing stock (individual fishing quotas). These are the tangible components of ecosystems.

F. Property Rights over Services
There are other intangible components of ecosystems that do not fall under these definitions of rights. For example, species provide services to the ecosystem and to humans through genetic information, reproduction, and contribution to critical ecosystem functions through roles such as predator or prey. Species and groups of species also provide additional value to humans through the information they contain and the pleasure they provide.

Property rights are poorly defined for ecosystem services such as genetic information. The idea of intellectual property rights over discoveries leading from information and ideas is a familiar one. Nations have patent and copyright systems that protect rights of ownership over goods and services that flow from the application of intellectual capital; these include designs, published works, and medicines. But the focus of intellectual property rights is on the discovery or creation that results from information, not on the information itself.

The lack of systems of property rights for genetic information is a critical issue for its protection. The information stored in biodiversity could lead to discoveries and patented products, but for the potential investor, the necessary protections to encourage investment are not in place. Since rights to the use of genetic information are unspecified, incentives for its protection are also absent. Biodiversity is treated as an open access resource, with the familiar “tragedy of the commons” result.

G. Optimal Forms of Property Rights
Biodiversity is a public good at both national and international scales. Developing the means to protect its value is a critical challenge to property rights. Some of the same conclusions that apply to property rights and the protection of single resources apply as well to multiple resources and biodiversity. The evidence of the applied property rights literature indicates that no single type of property rights can be a remedy for all needs of resource protection. It depends on the resource management objectives and on the context.

People often advocate a particular form of property rights as being best suited to manage and sustain natural resources. For example, arguments have been made for private property rights to natural resources to provide incentives for maintaining the flow of resource goods and services into the future. The idea is that private property rights would keep people from the “prisoner’s dilemma,” in which ignorance over the behavior of others leads to the collectively irrational outcome of resource overuse. The argument for universal private property ignores the differences in context in which they might be applied. It also disregards possibilities for cooperation and collective action that provide assurance about others’ behavior and that exist under other types of rights systems.

In recent years counter-arguments have arisen for the superiority of collective ownership over private ownership. These arguments are often framed in the context of community-based resource management. The idea behind this argument is that collective resource ownership and management is superior in terms of providing incentives for stewardship, providing collective access to resource benefits, and fairness in resource outcomes.

Disputing each of these arguments for the superiority of a single type of property right is a large body of research that has demonstrated that there is no particular form of property right that is superior in all cases. Each type of property right—with the exception of open access, which is the absence of rights—can either succeed or fail in sustaining resources. What matters is how well the property right system fits within the ecological, economic, and social context and how well the form of the right reflects the type of use.

The fundamental problem for biodiversity protection is that much of the biodiversity value falls outside the realm of direct use. Direct uses are easiest to monitor
and measure, control and trade. Indirect uses such as
genetic information and aesthetic appreciation are less
visible because their contribution is more diffuse over
space and time. Because these resource values are sel-
dom owned, bought, or sold, they are often at a competi-
tive disadvantage in a market economy, leaving them
undervalued and overused. This leads to the common
approach of protecting biodiversity not through prop-
erty rights but by removing them from human influence
in sanctuaries, refuges, or reserves.

III. FUNCTIONS OF PROPERTY RIGHTS

Property rights have several functions. They delineate
the population of legitimate owners. They specify the
allowable actions of these owners and their associated
responsibilities so that expectations are consistent and
enforcement of rules is possible. And by setting consist-
ten expectations they reduce uncertainty about others’
behavior. In the larger sense, they connect the pieces
of the natural system to the pieces of the human system.

The basic functions of managing resources—
coordinating users, enforcing rules, and adapting to
changing environmental conditions—cannot be met
without a system of property rights. The way that prop-
erty rights function in any particular context determines
whether the natural and human system will interact in
compatible or conflicting ways. To be compatible with
changing environmental conditions—cannot be met
without a system of property rights. The way that prop-
erty rights function in any particular context determines
whether the natural and human system will interact in
compatible or conflicting ways. To be compatible with
long-term biodiversity protection, systems of property
rights must function in a way that deals with uncer-
tainty, externalities, transactions costs, and scale.

A. Uncertainty

Uncertainty is endemic in natural systems. We lack
knowledge about ecosystem structure and ecosystem
condition. We often lack assurance about how others
will behave, or we lack confidence about the future.

For example, there are large gaps in basic knowl-
edge about biodiversity; about threshold levels of pro-
tection, the measurement of ecosystem function, and
the definition of goals for biodiversity. In marine ecosys-
tems relatively little is known about ecosystem compos-
tion, links between species, food, and reproductive
requirements, critical ecological processes, and trade-
offs among species.

These uncertainties affect how well property rights
perform. Uncertainty about ecosystems may limit the
ability to define goals and objectives for biodiversity.
Uncertainty about the behavior of others will encourage
intensified use to capture as many benefits as possible
while they are available. Uncertainty about the future
will shorten the time horizons over which decisions are
made, removing the incentive to invest in long-term
protection. All of these forms of uncertainty work
against the protection of biodiversity.

In an uncertain environment, decision making takes
place through trial and error. The uncertainty creates
a natural tension between the individual and the group,
and between people and ecosystems. Property rights
address some, but not all, of the components of uncer-
tainty. They define and sanction ownership, resolving
the question of future access. They provide assurance
about the behavior of others, because they define appro-
priate types of use. They do not in themselves increase
knowledge about the ecosystem, although they may
provide an incentive for owners of property rights to
produce this information.

B. Externalities

Another function of property rights is to resolve the
problem of externalities, in which one action affects
another. When property rights to resources do not exist
or are incomplete, people do not take full account of the
costs of their actions because there is no corresponding
“owner” to lay counter claims. In open access fisheries,
for example, people compete for resource benefits by
fishing as fast as they can and catching as much as they
can. The result is that one person’s behavior affects the
amount of fish available to others over time, unless
their rights are correspondingly protected. Similarly,
destructive fishing practices can eventually affect the
functioning of the ecosystem if rights to habitat or non-
fished species are not in place.

In some cases property rights may be defined but
unenforceable, and the lack of enforcement then be-
comes equivalent to removing the right. For example,
if fishing rights are expressed in terms of areas and
people without rights are not excluded from those areas,
their encroachment will render the right meaningless
and externalities will continue. The outcome of missing
or unenforceable property rights is biodiversity loss, as
only some components of the ecosystem are protected,
leaving others vulnerable to external effects.

The idea behind using property rights to protect
biodiversity is to assign claimants to the full spectrum
of ecosystem components and services, so that external
effects are accounted for, or “internalized.” If the owners
of biodiversity are citizens, the state can then act in
their behalf to protect biodiversity from various dam-
ages over space and time. The ability of the state to
represent claims over biodiversity depends on a clear
definition of the goods and services of ecosystems, an articulation of the relation between ecosystem components and ecosystem function, and an active constituency for protection.

C. Transactions Costs

How well a system of property rights functions both affects and is affected by transactions costs. Transactions costs are the costs of doing business, which in the resource management context include costs of gathering information, coordinating users, organizing decision making, and enforcing rules. Some transactions costs remain fixed regardless of the type of process used to make decisions. Others vary with the way decisions are made—the amount of data collected, analyses done, and the process used to make decisions.

Transactions costs are also influenced by the condition of the ecological system. As resources become depleted, a system of property rights must account for more and more externalities that increase the costs of management program design and enforcement. It is possible to create a system so costly to design or enforce that potential benefits are outweighed by the costs. This cost effect, particularly in consideration of actions to change property rights, is a particularly important factor in the consideration of the transition to property rights for protecting biodiversity.

D. Scale

The functioning of property rights is also affected by the scale of the area over which they apply. There is a disconnect between the geographic scale over which species are distributed and the scale at which species habitat is owned. Property rights to goods are relatively easy to define and enforce because they are associated with geographic space and their value is embodied in the goods themselves. But property rights to services are much more elusive. For example, rights to the information value of species would involve several owners, even nations, over wide geographic space.

For rights to resources to have meaning in terms of providing incentives for their conservation, those rights must have value to the owners. The value rests in part on their uniqueness and exclusivity. When the resource is the genetic information encoded in a plant and the plant species is distributed over a range larger than that areas encompassed by the property right, the information embodied in that plant is also available to others, leaving the rights owner without the power to exclude or to capture the potential financial value from the information. This scale mismatch limits the potential for property rights to provide the appropriate incentives for investing in conserving the genetic information.

The scale question also means that biodiversity protection is often an international issue. There are global benefits to biodiversity that are unaccounted for in national systems of property rights. Species distributions are independent of national boundaries. In addition, the loss of biodiversity is often an international externality, where the impact of one nation’s actions is felt by another.

IV. EVOLUTION OF PROPERTY RIGHTS

Property rights to resources tend to evolve incrementally over time in response to changes that alter the costs and benefits of particular forms. The current impetus for considering property rights for the protection of biodiversity is the scarcity resulting from declines in the number, range, and diversity of species that enhances biodiversity’s value. The evolutionary path to this point, marked by a gradual expansion of property rights over ecological goods and services, is one that has reflected the relative changes in the benefits and costs of property rights as conditions of resource use change.

A. Single-Species Use

When resources exist in surplus to human needs, conservation actions are unnecessary and so rights, rules, and responsibilities have little meaning. In an environment of surplus, the costs of developing protective mechanisms exceed the benefits gained. As resource abundance declines, either from human pressures or natural events, the increasing scarcity raises the benefit of protection. When the benefit of protection exceed their costs, property rights in some form develop. The first controls are on how the resource is taken or how much is taken. For example, an open access, no-rules fishery becomes restricted by how people can fish—for example, by seasons or by limits on gear—and how much fish they can take. Eventually, property rights of access to the fishery are developed to restrict who can fish.

B. Ecosystem Use

As ecosystems are further exploited they reach a point where the focus on extracting and conserving individual species erodes the functioning of the ecosystem as a whole. Exploitation of species is interconnected, the
sources of impacts are diffuse, and effects of exploitation are system-wide. Species richness declines, community interactions are altered, and resilience erodes. It is at this point that the focus of managing human effects shifts from single species to ecosystems.

This shift in focus takes place against a background of using resources for the production of goods: lumber from trees and seafood from fish. The rights, rules, and responsibilities are all directed toward this end. They are not designed to control effects on unused ecosystem components or to protect the flow of ecosystem services. The system of property rights is poorly adapted to these interactive effects.

Biodiversity decline creates scarcity in services provided by species richness and genetic information. But the same scarcity in ecosystem services that provides an incentive to create property rights also presents a difficulty to developing ecosystem-level property rights. Ecosystem management is typically proposed when exploitation levels of single species are too high and showing signs of stress. Human competition and conflict over access to individual species complicate the shift to a broader focus and cause users to challenge the legitimacy of an ecosystem approach.

C. Expectations about Use

Property rights systems, once in place, create expectations about what is normal. They start in motion a new path of development. At each stage in the path, the condition of the resource, the definition of rights, and the expectations about those rights influence action. The further along the path the more embedded are the property rights and the more vested are people in their continuance. When faced with the necessity of expanding those rights to the protection of biodiversity, it must be done against the legacy of the path to this point.

A property rights transition to the protection and conservation of biodiversity is affected by resource conditions at the intervention point. Protecting biodiversity means losing access to some of the goods and services of the ecosystem, which, under the old property rights rules, were at the disposal of the owner. Many resources have competing uses with known market values. Unknown future uses of these resources will be hard pressed to compete against known present uses.

V. PROPERTY RIGHTS AND BIODIVERSITY PROTECTION

A number of emerging issues illustrate the challenges to the use of property rights to protect biodiversity. Four are particularly relevant: uncertainty, exclusivity, distribution of benefits, and the alignment of private and social goals.

A. Uncertainty

Protecting biodiversity requires more than a set of rules and responsibilities for property rights set within institutions that accommodate the attributes of the ecosystem and the people who use it. How to design property rights and institutions for this accommodation is what is at issue. Although there is general understanding of the importance of biodiversity to the stability, function, and sustainability of ecosystems, there is poor understanding of specifically what to protect. For example, knowledge is generally lacking of the thresholds at which biodiversity loss irreversibly changes ecosystems. Knowledge of the role played by individual species in contributing to critical ecosystem functions is also absent. These uncertainties create a corresponding uncertainty in the objectives and design of property rights to reflect the full range of values involved.

B. Exclusivity

A problem with many property rights that apply to natural resources is that they do not specify claims to the full range of goods and services provided by an ecosystem. In failing to fully specify property rights claims, they fail to protect exclusive use. If property rights were defined for all components of an ecosystem, users and decision makers would take all the consequences of their actions into account. This would be the first step in making biodiversity conservation profitable. But under current systems of property rights, this is rarely the case. The history is to apply property rights to natural resources as commodities but not to the services they provide or to their existence value. The lack of full specification means that it is unclear who can claim and control rights of use.

C. Distribution of Benefits

How benefits will be distributed is the core debate of natural resource policy. The same debate is at the heart of biodiversity protection. Biodiversity values have the potential to be protected through a number of different types of property right. But whatever type of right is used will create winners and losers. It is difficult to design an effective system of property rights without addressing the questions of what the objectives are, how progress toward those objectives will be measured,
and the time frame over which they will be met. How to promote equitable distributions is a difficult question in all resource policy, but it is particularly so for the complex issue of biodiversity. Equity questions extend into the international arena, where the distribution of benefits between rich and poor nations is at stake.

Effective resource protection depends on legitimacy—on the acceptance of rules and procedures by participants. Many natural resource systems are difficult to monitor, and the possibilities for circumvention of rules are many. When people doubt the legitimacy of the system of property rights because they cannot accept its distributional outcomes, their incentives are to undermine rather than promote its evolution to a new form. Scarcity compounds the erosion of legitimacy by creating greater incentives and opportunity for rent seeking that is characteristic of resource competition.

D. Alignment of Private and Social Goals

For property rights to bring private preferences into line with public preferences for biodiversity conservation, the tensions between private and social goals must be resolved. Individuals have private goals for productivity of ecosystems goods that may not be compatible with social goals for biological production. In addition, biodiversity is a public good, so it is subject to the potential for free riders to enjoy the benefits without paying the costs. Biodiversity protection must be accomplished in concert with existing rights and rules that favor direct uses of ecosystem goods.

How to realign private and social incentives to conserve biodiversity is an important and to a large extent unresolved question. Most current systems of rights, because they are still unspecified for ecosystem services, favor the conversion of ecosystems into goods. Options for change include the expansion of the scope of property rights, payment of compensation to owners for conserving rather than using resources, or developing prohibitions against certain uses.

VI. CONCLUSIONS

Property rights outline the conditions under which resources can be used and the interactions of people in using them. They exist in particular contexts that are combinations of ecosystems and people. There are many forms and permutations, none of which is superior to others. What determines the effectiveness of property rights is the extent to which they are able to perform the functions of reducing uncertainty, removing externalities, containing transactions costs, and accommodating scale.

Property rights evolve in response to external conditions. Scarcity is the driving force for their existence, and their contribution of benefits in excess of costs is the means of their continuance in a particular form. Despite the increasing scarcity of biodiversity, particular challenges face the evolution of systems of property rights to meet these conservation needs. To protect biodiversity, a transition must be made between current systems of rights that protect direct uses of ecosystems to expanded systems of rights that also protect indirect ecosystem services such as genetic information and aesthetic enjoyment.

The application of property rights to protect biodiversity is challenged by forces that are both internal and external to ecosystems. Internally, the increasing scarcity brought on by overexploited resources means that the range of options has declined. Fewer adaptations are possible. Formidable information needs for keystone species and critical functions of ecosystems exist.

Both internally and externally, forces are competing to shape the time horizon of resource management. Biodiversity protection and notions of sustainable use—values held by the public at large—are long-term concepts, requiring a long time horizon for management decision making. At the same time, the overexploited levels of many resources combined with uncertainty about their sustainability leads to internal pressures to shorten the time horizon and make decisions for the short run.

As these external and internal pressures illustrate, the very conditions in ecosystems that require biodiversity protection are also those which are causing difficulties in its implementation. Increased knowledge of ecosystems is needed as are innovative designs for property rights that work within and across national boundaries.

See Also the Following Articles

Biodiversity as a Commodity • Commons, Theory and Concept of • Economic Value of Biodiversity, Overview • Environmental Ethics • Land-Use Issues

Bibliography


PROTOZOA

Bland J. Finlay
Centre for Ecology and Hydrology—Windermere

I. Functional Roles
II. The Nature of Protozoan Species
III. Protozoa and Ecosystem Function
IV. Biological Diversity and Global Species Richness

GLOSSARY

benthos The substrate at the bottom of the sea or fresh waters. Most benthic protozoa live in the spaces between sediment particles.

commensal Of a protozoan—one that is loosely but not obligately associated with another organism (e.g., the ciliates attached to the external surfaces of crustacean zooplankton).

endosymbiont An organism living in a long-term association inside a host organism to their mutual benefit (e.g., the endosymbiotic methanogenic bacteria that live inside anaerobic protozoa and utilize waste H₂ produced by the host).

eukaryote An organism with membrane-bounded nuclei in its cells. Protozoa are unicellular eukaryotes.

heterotrophic Mode of nutrition in which carbon is obtained from the organic compounds made by autotrophic organisms.

phagotroph An organism that ingests solid food particles (e.g., bacteria).

planktonic Living in the water column of lakes and/or the sea.

PROTOZOA are microscopic organisms with animal-like features. Each protozoan typically exists as a single, independent cell, and all free-living protozoa fit within the definition of phagotrophic microbial eukaryotes. In some species, independent cells coalesce to form plasmodia (e.g., slime molds) or unite to form colonies (e.g., colonial choanoflagellates). There has never been unanimous agreement as to where the boundaries of the protozoa should lie, largely because of the extraordinary diversity of their lifestyles. Many species cause disease (e.g., malaria). Others thrive as commensals in the digestive tracts of ruminants and wood-eating insects, and many plant-like flagellates (e.g., Euglena) have at various times been referred to as protozoa, algae, or plants. Here, we are concerned principally with free-living protozoa, and the definition of this group is one that is based on its key function in the natural environment: Protozoa are capable of phagotrophy—the ability to capture and ingest food particles. Many also have functional chloroplasts or endosymbiotic algae, and the resulting consortia are known as “mixotrophs” because they are capable of both phagotrophy and phototrophy. Like most microorganisms, protozoa have very large population sizes, and they are the most abundant group of phagotrophic organisms in the biosphere. Biodiversity at the level of protozoa has characteristics that are not shared by macroscopic animals and plants. Most protozoan species are probably globally ubiquitous, so the global number of species is relatively small. A significant proportion of local protozoan species richness, at any moment in time, is rare or cryptic and awaiting
the arrival of conditions suitable for growth and reproduction.

I. FUNCTIONAL ROLES

All of the important functional roles of free-living protozoa derive from their small size. The smallest flagellates are 2–4 μm and most are <20 μm, most amoebae are <30 μm, and most ciliates are <200 μm. Exceptionally, some amoeboid protozoa, such as the radiolarians and foraminifers, and the agglutinated foraminifers of the deep-sea benthos, may reach 2 mm or more, especially if they have spiny extensions. Because protozoa are so small, most suitable prey items are other, smaller microbes. Protozoa are the principal consumers of the immense natural resource of bacteria and other microbes, and because they have population growth rates that are similar to those of the microbes on which they feed (doubling times in the order of 1 to several days), they are usually able to control microbial abundance. Flagellated protozoa can probably consume all bacterial production in the plankton. In the benthos, protozoa overlap in their niche requirements with nematodes, rotifers, tardigrades, turbellarians, and gastrotrichs, but because of their great abundance protozoa are indeed quantitatively the most important grazers in the fresh-water and marine (including deep-sea) benthos. Also, just as microbes achieve astronomical abundance on a global scale, so too are the protozoa that graze on them represented by species populations with proportionately smaller but still unimaginably large global abundances.

Protozoan grazing on microbes also stimulates activity of the whole microbial community, in both aerobic and anaerobic environments. The process involved is not fully understood, although it may operate by increasing the rate of turnover of essential nutrients that would otherwise remain “locked up” in bacterial bio mass. The net effect is that grazing by protozoa stimulates the rate of decomposition of organic matter. The variety of shapes, sizes, and relative abundances of microbial food items has driven the evolution of a comprehensive suite of methods to capture them and a considerable diversification of protozoan morphologies. In general, the size of a protozoon relative to its prey dictates the most efficient food-capturing mechanism. Where the predator is typically much larger than its prey, filter-feeding prevails, and where the size difference is less the protozoon is more likely to be a raptorial feeder—one that seeks out relatively large individual food items. Thus, the planktonic choanoflagellate feeding on a dilute suspension of tiny bacteria 20 times smaller than itself does so with a very fine filter, and the ciliate feeding on dinoflagellates half its size seeks out and captures each one individually. There are, of course, many exceptions, such as the marine heterotrophic dinoflagellates that use a feeding veil to trap and digest diatoms much larger than themselves.

The third main feeding type in protozoa is termed “diffusion feeding.” This is particularly common in planktonic amoeboid protozoa (radiolarians, foraminifers, and heliozoans) and in the suctorian ciliates. It works when prey items collide with the sticky spines, tentacles, or axopods that radiate from the protozoon. Unlike the other two main modes of feeding, the protozoon simply waits for the arrival of its prey, much as a spider waits for an insect to be snared in its web. Thus, there is a close link between protozoan morphology (especially of the food-capturing organelles) and the way in which a protozoon functions as a grazer. Therefore, when we classify the free-living protozoa into broad morphological groups, they are simultaneously allocated to broad functional groups. The three broadest morphological–functional groups are the amoeboid, the flagellated, and the ciliated protozoa, and each has its own strengths as a phagotroph. Representatives of all three may feed on the same type of microbes in the same place (e.g., in an aquatic sediment), but they will differ in the mechanics and efficiency of capture of any particular food particle. A filter-feeding flagellate will have a relatively large filter area, a high volume-specific clearance and competitive superiority over filter-feeding ciliates when grazing on planktonic bacteria. A helioflagellate and a suctorian ciliate will both practice diffusion feeding, but the former will be adapted for snaring bacteria, whereas a diffusion-feeding suctorian will specialize in trapping flagellates and ciliates. Many protozoa are microaerophilic. They seek out habitats with a low level of dissolved oxygen that is just sufficient to drive their aerobic respiration and low enough to exclude metazoan competitors and predators. Microaerobic habitats are common in aquatic sediments and in oceanic “oxygen minimum zones.” These are zones in which the raw materials for microbial growth arrive from opposite directions (e.g., where oxygen and light arriving from above meet carbon dioxide and sulfide from below) and in which there is therefore an elevated abundance of microbial food. Therefore, microaerophily is an adaptive behavior: It brings protozoa into contact with high abundances of microbial food. It also stimulates the growth of nutritional symbionts such as sulfide-oxidizing bacteria and endosymbiotic algae. Many microaerophilic protozoa are also tem-
porary anaerobes, but unlike the "true" anaerobes that live permanently in the absence of oxygen, their metabolism is fundamentally aerobic. The true anaerobes—those that complete their entire life cycle in the absence of oxygen—live principally in aquatic sediments. There are many species, but none is ever abundant. Most use hydrogen-evolving fermentations for energy generation, and the hydrogen is used by anaerobic bacteria, especially endosymbiotic methanogens. Thus, methane is released from these protozoan consortia. The anaerobic protozoa are probably the only phagotrophic organisms capable of living permanently in the absence of dissolved oxygen.

The real diversity of symbiotic associations involving protozoa is poorly known. In some cases, complex interactive behaviors have evolved between the partners. In the marine, sand-dwelling ciliate *Kentrophoros*, the entire dorsal surface of the ciliate is a coat of sulfide-oxidizing bacteria that can grow only in the narrow layer within the sediment where oxygen and sulfide overlap. The ciliate host's innate microaerobic behavior enables it to seek out the habitat that the symbiotic bacteria need for growth. The ciliate then invaginates its dorsal surface and digests the bacteria because it does not have a mouth, and this is its only source of nutrition.

It is clear that many of these symbiotic consortia involving protozoa represent tightly integrated functional units; indeed, the symbioses may be almost as deeply embedded functionally in the consortium as the protozoan's other organelles. Two points must be noted. First, it is the combined phenotype of the consortium, rather than that of any individual consortium partners, on which natural selection will operate, and there are examples of how the fitness of a protozoan in a particular habitat can be improved by the acquisition of endosymbionts (e.g., the algal symbionts of ciliates living in the metalimnia of freshwater lakes). Second, the biodiversity of protozoa, when quantified simply in terms of protozoan species richness, will fail to take account of the large supplementary microbial diversity with which the protozoa are necessarily associated.

Therefore, the diversity of free-living protozoa may be classified into broad morphological–functional groups: amoeboid, flagellated, and ciliated protozoa. Almost any free-living protozoan can be placed without difficulty in one of these groups. It must be stressed, however, that these groups are not concordant with any system of classification of protozoa published in recent years; nor are they in most cases aligned with the independent lineages that are emerging in the molecular phylogenies (e.g., those based on sequence variation in ribosomal RNAs) which reflect the main episodes in the history of eukaryotic evolution. Heterotrophic flagellate groups, such as the diplomonads and trichomonads, appear in the early emerging lineages, but other flagellates (e.g., chonoflagellates, Euglenida, chrysomonads, dinoflagellates, and haptomonads) are classified within recently diverging lineages. The amoeboid protozoa too are scattered across many lineages. The naked amoebae without mitochondria (the pelobionts) diverge early and close to the diplomonads, whereas the valvulifera, slimes molds, and various other groups of naked and testate amoebae appear in other independent lineages that are evolutionarily quite distant from each other. The morphological–functional group of the ciliates is the only one which remains intact, as a monophyletic group, in current molecular phylogenies.

The process of distilling the vast quantity of molecular information generated in recent years has generated some entirely new phyletic assemblages, including the stramenopiles, a group containing organisms as morphologically and functionally dissimilar as chrysomonad flagellates and diatoms, and the alveolates—a group that embraces the dinoflagellates, the ciliates, and a large group of exclusively intracellular parasites (the apicomplexans). In the next section, I focus on the broad morphological–functional groups, define what is meant by species, and quantify the species within these groups.

II. THE NATURE OF PROTOZOA SPECIES

There is no universal agreement on what constitutes a protozoan "species." The most widely used concept is the morphospecies because it is relatively easy to discriminate a great diversity of protozoa using body form alone. This concept is especially useful because morphology is closely related to ecological function. In many protozoa, the structure of the feeding apparatus and the size and shape of the cell determine the way the protozoan functions in the natural environment; therefore, the form largely determines the ecological niche that the protozoan occupies. Discriminating species on the basis of form might then be equivalent to discriminating them according to the ecological niches they occupy. We could say that if a protozoan looks the same in different places, then it is the same in different places. However, there are problems with such a simple concept. It is known that a morphospecies can be com-
posed of ecologically distinct populations, such as those adapted for maximum growth rate at different tempera-
tures. Therefore, although it is possible to collect appar-
ently identical representatives of a morphospecies in
different corners of the world, and one might be encour-
aged by the belief that they were filling exactly the same
niche in these different places, in reality they could be
phenotypically quite different. The same problem
applies with respect to genetic differences. It is possible
to find genetic differences, especially sequence differ-
ences in ribosomal RNAs, in morphologically identical
isolates collected from different places. The significance
of these genetic differences is not known; nor is it
known if genetic and phenotypic divergence are
correlated. It appears, however, that there is no correla-
tion between genetic divergence and geographic dis-
tance.

Protozoa spend most of their time as asexual organ-
isms, but some do reproduce sexually, if only periodi-
cally, and in some well-studied cases (e.g., the ciliate
Tetrahymena and Paramecium) morphologically
indistinguishable biological species (reproductively iso-
lated gene pools, also known as sibling species) have
been studied thoroughly. Different sibling species
within a morphospecies may be genetically identical to
each other or extremely divergent (at least with respect
to ribosomal RNAs), and there is no apparent correla-
tion between genetic isolation and genetic divergence.
Again, there is no good evidence for biogeography of
protozoan sibling species, and members of the same
sibling species can be found on different continents. It
is possible that sibling species carry unique phenotypic
traits that equip each species for a particular niche, but
this has not been demonstrated. The real problem with
using a biological species concept for protozoa is that
it is not practical for all but a few easily cultivated
protozoa. The majority of protozoan species have never
been cultivated, and neither the frequency nor the char-
acter of their sexual behavior (if any) are known or
are ever likely to be known. Most people who study
protozoa in the natural environment use the morpho-
species concept because it is practical, it embodies the
close link between form and function, and it is the
morphospecies that fills the niche, if not always the
minutest niches, that can be discerned by the human
observer. The morphospecies may contain much ge-
netic variation and be capable of expressing a wealth
of phenotypic variation, but it is the best tool that is
available for ordering the diversity that lies within
the protozoa.

One particular problem that is particularly acute
when dealing with organisms the size of protozoa is
that the quality of one's perception of morphology is
very definitely a function of the tools that are available.
Thus, better tools (e.g., quality microscopes) enable the
discrimination of finer detail. The logical extension of
this is that even individual protozoa in a population
might be discriminated one from the other and could
therefore be referred to as separate morphospecies.
Modern optics have undoubtedly lent impetus to the
business of describing many new species, some of which
are undoubtedly legitimate, whereas many have been
established on morphological criteria that are trivial or
have no conceivable functional significance.

An additional problem with the morphospecies is
that it is often difficult to decide exactly where a species
begins and ends. Although most observations are made
of the vast number of individuals that cluster around
the central tendency of any population, the individuals
at the tail ends of the distributions abut and often over-
lap those of other species. Morphological variation then
appears to be continuous across many species. Exam-
pies of this have been described: In some of the large
sp сентяоге spumellarian radiolarians, the major features
(e.g., skeletal and cytoplasmic morphology) that are
used to discriminate species intergrade to such an extent
that it is not possible to unambiguously ascribe individ-
ual radiolarians to nominal species.

However, the evolutionary process in radiolarians
and other protozoa probably works in the same way
as it does for other organisms such that it maintains
phenotypically discrete species in niche space that is
essentially continuously variable. The phenotypic traits
that sustain these discrete species in protozoa are pre-
sumably quite diverse in character and possibly not
totally comprehensible. Among the more accessible
of these traits are the morphological characters that we
use to separate species. However, the likely limitation
of these is that they enable us to perceive only the fairly
closely resolved, sometimes overlapping entities that
we call morphospecies.

III. PROTOZOA AND
ECOSYSTEM FUNCTION

Protozoa are abundant. One gram of soil typically con-
tains 10^5–10^7 naked amoebae, 10^5 planktonic foramin-
iferans can often exist beneath 1 square meter of oceanic
water, and almost every milliliter of fresh water or sea-
water on the planet supports at least 100 heterotrophic
flagellates. When expressed in global terms, these num-
bers are very large, and an inevitable consequence of
the persistence of such a large number of very small organisms is that migration rates will be relatively high. It follows that rates of speciation and extinction must be low, as will the consequent global number of species. It also follows that protozoa are unlikely to have biogeographies, and “endemic” species probably do not exist. We might expect that the local diversity of protozoa would account for a significant proportion of global diversity, even if at any moment in time much of this diversity is represented by rare or inactive individuals (e.g., cysts awaiting the arrival of suitable conditions). There is indeed good evidence for the global distribution of protozoan species, including the morphologically distinctive flagellate *Rhynchomonas nasuta* that has been found in most aquatic and terrestrial environments worldwide; marine foraminifers and ciliates found living in slightly salty water of desert oases, hundreds of kilometers from marine coasts; the same radiolarian species living in high northern and southern oceanic latitudes; the same pond-dwelling ciliates living in Australia and northern Europe; and the cosmopolitan distribution of the same species of agglutinated foraminifers in the deep-sea benthos. In general, protozoan morphospecies are ubiquitous and apparently cosmopolitan if the habitats to which they are adapted are distributed in different parts of the world. In accordance with this, the global number of protozoan species is indeed relatively modest (Table I), and the number of species that can be retrieved from a local area (e.g., a pond), in both “active” form and from a “passive” state is a significant proportion (usually at least 10%) for various morphological-functional groups of the global total. This fact may not be obvious from short-term ecological sampling programs because only a limited

![Table 1: Estimates of Global Species Richness of Extant Free-Living Protozoa](https://example.com/table1.png)

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Marine</th>
<th>Non-Marine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amoebozoa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slime molds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dictyostelids</td>
<td>60</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Myxomycetes</td>
<td>530</td>
<td>530</td>
<td></td>
</tr>
<tr>
<td>Rhizopod amoebae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naked</td>
<td>180</td>
<td>220</td>
<td>400</td>
</tr>
<tr>
<td>Testate</td>
<td>220</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>Foraminiferae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planktonic</td>
<td>45</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Benthic, shallow</td>
<td>4000</td>
<td>4000</td>
<td></td>
</tr>
<tr>
<td>Benthic, deep sea</td>
<td>250</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>Actinopod amoebae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acantharians</td>
<td>1300</td>
<td>1300</td>
<td>2600</td>
</tr>
<tr>
<td>Radiolarians, solitary</td>
<td>750</td>
<td>750</td>
<td></td>
</tr>
<tr>
<td>Radiolarians, colonial</td>
<td>50</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Holotricha</td>
<td>120</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>Flagellated protozoa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Excluding heterotrophic dinoflagellates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marine plankton</td>
<td>420</td>
<td>420</td>
<td></td>
</tr>
<tr>
<td>Marine benthos</td>
<td>330</td>
<td>330</td>
<td></td>
</tr>
<tr>
<td>Freshwater and soil</td>
<td>350</td>
<td>350</td>
<td></td>
</tr>
<tr>
<td>Heterotrophic dinoflagellates</td>
<td>900</td>
<td>300</td>
<td>1200</td>
</tr>
<tr>
<td>Other mixotrophic flagellates</td>
<td></td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>Ciliated protozoa</td>
<td>1400</td>
<td>1660</td>
<td>3060</td>
</tr>
<tr>
<td>Total</td>
<td>11890</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Compiled from numerous published and unpublished sources. The more problematic estimates are highlighted.

*There is considerable uncertainty attached to these estimates. The figure of 4000 is generally accepted as a working figure for extant species but is probably inflated by synonyms, especially those of shallow-water benthic species. There is no firm information for species richness of deep-sea benthic foraminifers. The estimate of 250 is approximately double the typical figure for local richness of species, most of which may be cosmopolitan, but there are probably many undescribed deep-sea soft-shelled foraminifers that have in the past been ignored by geologists.

*The total given here for heterotrophic flagellates is 1100 species. This includes some synonyms and mixotrophs. It is believed by some that the real global total is closer to 3000 species.

*Assuming there are 1800 marine and 220 freshwater species, and that 30% of these are heterotrophs.

*This estimate is simply derived by doubling some recent estimates. Note that these species may be only temporarily phagotrophic.
number of microbial niches are available at any moment in time.

Protozoa and other microorganisms have other special properties. Microbial activities interact strongly with physical and chemical factors in the natural aquatic environment (e.g., light transmission or the concentrations of essential nutrients) to create a continuous turnover of microbial niches. These niches are quickly filled from the locally available diversity of rare and dormant microbes, and the activities of the latter create further reciprocal interactions. Therefore, the diversity of active protozoan species in a pond, at any moment in time, is the result of preceding reciprocal interactions involving many biological and nonbiological factors, and the biodiversity of protozoa and other microbes is an integral part of ecosystem functions such as carbon fixation and nutrient cycling.

IV. BIOLOGICAL DIVERSITY AND GLOBAL SPECIES RICHNESS

A. Amoeboid Protozoa

Slime molds may be regarded as protozoa because they spend most of their active lives as amoebae or as naked amoeboid (and often macroscopic) plasmodia (Fig. 1). They are known by many names, including mycetozoa ("fungus animals"), because although they never de- 

IV. BIOLOGICAL DIVERSITY AND GLOBAL SPECIES RICHNESS

A. Amoeboid Protozoa

Slime molds may be regarded as protozoa because they spend most of their active lives as amoebae or as naked amoeboid (and often macroscopic) plasmodia (Fig. 1). They are known by many names, including mycetozoa ("fungus animals"), because although they never de- 

IV. BIOLOGICAL DIVERSITY AND GLOBAL SPECIES RICHNESS

A. Amoeboid Protozoa

Slime molds may be regarded as protozoa because they spend most of their active lives as amoebae or as naked amoeboid (and often macroscopic) plasmodia (Fig. 1). They are known by many names, including mycetozoa ("fungus animals"), because although they never de-
Some naked amoebae (especially those that form extensive anastomosing plasmodial networks) resemble hybrids of shell-free foraminiferans and slime molds. Some can reach several centimeters in diameter. They are typically very slow moving and apparently perfectly adapted for grazing extensive areas of soil surface (Leptomyxa) or diatom mats in the marine benthos (Coralloomyxa). Some free-living naked amoebae are anaerobic and do not have mitochondria. As a group, these are referred to as pelobionts, rhizomastigids, archamoebae, or karyoblasteans (e.g., Mastigamoeba, Mastigella, and Pelomyxa). They are relatively common in organically enriched anoxic freshwater and marine sediments. The very large (up to several millimeters in diameter) Pelomyxa supports several types of endosymbiotic bacteria, including methanogens, and probably lives exclusively
in freshwater sediments. It appears to ingest any organic particles with which it makes contact. It can be locally abundant and the dominant grazer in some anaerobic lake sediments. Most species of naked amoebae have been described from fresh waters and soil, and most marine species have been collected from inshore areas. It is not clear to what extent freshwater species can also live in seawater. Some genera (e.g., Mayorella and Acanthamoeba) are well represented in both environments.

Foraminifera live in marine or brackish water environments, and most have calcareous shells. Cytoplasm emerges through the aperture and pores in the shell to form a repeatedly branching and anastomosing reticulopodal net in which prey are trapped. The "spinose" planktonic species (e.g., Globigerimoides ruber) have calcareous spines, and familiar Cretaceous chalk deposits (e.g., the white cliffs of Dover, UK) consist largely of the sediments of these shells together with the coccoliths of haptomonad flagellates. Some naked amoebae that produce similar pseudopodial networks live in fresh waters (e.g., Biotumex) or soil (e.g., Reticulomyxa).

The shells of planktonic foraminifera are generally 0.1–1 mm in diameter, but the pseudopodial network can reach 10 mm and the spines may increase the size to more than 20 mm, making them larger than many marine metazoans. Planktonic species are open-water organisms. They are most abundant in tropical and subtropical waters, in which there is typically one per liter (or 10^-3 m^-3, depth integrated to 150 m). Each planktonic foraminifera produces a mass of reticulopodia (the "halo") that radiates outward from the shell. This is used to snare prey and to support algal symbionts. Spinose species appear to prefer zoo-plankton prey, and nonspinose species (e.g., Globorotalia menardii) prefer phytoplankton. These preferences may affect the local distribution of species. In the Red Sea north of 20°N, the dominating spinose species appear to reflect the greater local abundance of zoo-plankton.

Many planktonic foraminifera, especially the spinose species (e.g., G. ruber), contain algal symbionts that are carried along the spines by cytoplasmic streaming. There are probably only two types of symbionts in planktonic species: the dinoflagellate Gymnodinium beti, which is morphologically identical in many host species, and a small yellow-green (cryptomonad-like) symbiont. The symbionts are protected from digestion by the host. The carnivorous foraminifera Hastigerina pelagica envelops itself in a cytoplasmic bubble capsule which is probably a flotation device but is also used for trapping and digesting prey. Hastigerina pelagica does not have symbionts but commensal dinoflagellates (Pyrocystis spp.) that live in the capsule. The role of symbionts and commensals is unclear. High rates of photosynthesis are maintained within the consortium, the symbionts probably transfer significant amounts of fixed carbon to the host, their photosynthetic activity may be necessary for calcification of the host's shell, and the symbionts may use the host's (nitrigenous) waste.

Benthic foraminifera have a greater diversity of symbiotic partners, including diatoms (Nitzschia), dinoflagellates (Symbiodinium microadriaticum), red algae, and chlorophytes (Chlamydomonas). Some also sequester chloroplasts from ingested algae. Although many benthic foraminifera species have been described, information on their life cycles and ecology is available for only about 20 shallow-water species (e.g., Allogromia, Rosalina, Spiroloculina, and Sorites). These typically drag themselves slowly over surfaces, continually extending reticulopodia to trap prey (diatoms, algae, and bacteria). The abundance of benthic species is correlated with local biological productivity. In the fertile Mississippi Delta, numbers may reach several thousand in an area 10 cm^-2. Close to sewage outfalls, they are typically very abundant but low in species diversity.

The spatial distribution of all benthic species is very patchy. This may be related to assexual reproduction (known only for benthic species) and the consequential limited range of migration of offspring from the parent. Benthic foraminifera become relatively more important with increasing water depth, possibly due to a competitive advantage they have over metazoans in their ability to rapidly exploit seasonally deposited "phytodetritus." In oxygen minimum zones along continental margins of tropical oceans, they can, because of their relatively small size, tolerate low oxygen conditions better than can metazoans. Also, certain taxa (e.g., Bulinina and Uvigerina) seem to be adapted for life in oxygen-depleted waters.

Foraminifera in the deep-sea benthos are often abundant (typically hundreds per 10 cm^-2) and diverse (species richness may exceed that of nematodes), and
they are now regarded as major components of the fauna in this zone. Because they consume bacteria and detritus, they probably act as a link in the food chain to higher trophic levels in the benthos. Very small species (30–63 μm) are common, but the dominant fauna >300 μm in central oceanic regions (and particularly at depths greater than the carbonate dissolution depth of ~4000 m) are large agglutinated foraminifers, especially the komokiaceans, which have complex anastomosing networks of tubes. In the benthos underlying the oligotrophic waters of the central North Pacific, the biovolume of komokiacean tests may greatly exceed the oligotrophic waters of the central North Pacific, the biovolume of metazoans. However, the amount of protozoan biomass in komokiacean tests is relatively small, and it may be even less than that of the metazoans that also occupy the test. Edgertonia floccula consists of a network of branching agglutinated chambers buried in a large (2-8 mm) inorganic mudball that it shares with cohabiting metazoans. This foraminiferan is one of the most abundant meiofaunal animals in some areas of the abyssal northeast Atlantic. Other deep-sea agglutinated foraminifera (e.g., Bathysiphon and Hyperammina) may even reach several centimeters in length, and some species stand vertically in the sediment and project into the overlying water, in which they presumably feed. Many of the common deep-sea foraminiferal species are probably cosmopolitan in their distribution.

The xenophyophoreans are another deep-sea group about which relatively little is known. They live in agglutinated, often fragile tests (typically 1–10 cm in length) and they too probably feed on sedimented detritus. Abundances of hundreds per square meter have been reported. The multinucleate protoplasm is enclosed within a labyrinthine organic tube encrusted with foreign particles. The organism also contains numerous crystals of barium sulfate, which have no obvious function. Growth is episodic, with intervening periods of months of inactivity. They have no confirmed fossil record, although they often use the sedimented shells of the planktonic foraminiferan Globigerina to strengthen their tests (as in Homogamma maculosa). Foraminiferan fossils are useful for biostratigraphy because (i) their sedimented remains are abundant (e.g., in the “Globigerina ooze” which covers much of the abyssal plain beyond the continental shelves), (ii) individual shells are identifiable to species level, and (iii) their remains allow determination of sea surface “paleotemperature.” Seawater and calcite differ in their δ18O/δ16O ratio. This difference increases with temperature at the time the calcite was deposited in the fossil foraminiferan shell. This estimate can then be compared with the established correlation between the temperature of a water mass and the extant species it typically supports. Much of this work is supported financially by the oil exploration industry.

Actinopod amoebae all have axopods, which are stiffened ray-like pseudopodia that radiate from the cell body. They also have a peripheral network of filopodia, which are long, thin, sticky pseudopodia used for trapping food organisms that make chance contact. There are three principal groups: (i) radiolarians, consisting of the polycystines with perforated spiny lattice skeletons of silica and the phaeodareans with skeletons of hollow tubes and spines; (ii) acantharians with skeletal spines of strontium sulfate radiating from the center of the organism; and (iii) heliozoans, which often lack an internal mineral skeleton (although some species produce skeletal spicules which are either organic or siliceous) and superficially resemble polycystine radiolarians. All actinopod protozoa, apart from a large number of heliozoans, are exclusively planktonic in deep oceanic water.

Radiolarians and acantharians range in size from ~30 μm to several millimeters in diameter. Colonial radiolarians (e.g., Collozoa longiforme), which occur as gelatinous cylinders with diameters of several centimeters, can reach 3 m in length. Typical food items include diatoms, tintinnids and other ciliates, crustacean larvae, and other zooplankton, all of which are trapped in the peripheral network of axopods and filopodia. The smallest species may eat bacteria. Dinoflagellate (Symbiodinium), chrysomonad, and prasinomonad symbions are associated with many polycystine radiolarians. They probably contribute to host carbon metabolism because the consortium can last many weeks without exogenous food. The symbions are carried about in the host’s cytoplasmic streaming, gathering in the peripheral filopodia at dawn and withdrawing inside the host shell at sunset. The phaeodarians that live near the ocean floor do not have algal symbions. Radiolarians are found at all water depths down to ~8000 m, although they are most abundant and diverse between 200 and 2000 m. No benthic species are known. Like planktonic foraminifers, each radiolarian species inhabits an oceanic water mass that lies within a particular temperature range. Moreover, because of the immense depths to which living radiolarians sink, those species living in surface waters give a misleading impression of radiolarian “biogeography.” It is believed that radiolarians may be transported in very deep oceanic waters.
waters between high northern and high southern latitudes.

Radiolarian abundance and productivity increase in warmer waters. Abundance may be almost nil in the surface waters of the Antarctic, up to 50 per m³ in the Caribbean and Gulf Stream, and up to several thousand per m³ in the Gulf of Mexico. Colonial radiolarians are much less abundant. In the Sargasso they are considered abundant if represented by more than one colony per 100 m³. Acantharians can reach relatively high abundances (up to 25–35 per liter) and they are often the most abundant of the planktonic shelled amoebae, but relatively little is known about living organisms because as soon as they are collected the fragile cell membrane breaks and the strontium sulfate skeleton dissolves. Radiolarians are important stratigraphic tools, and skeletal morphology is the main characteristic used to identify species; hence, much information is available on the diversity of radiolarian skeletons.

The heliozoans are predominantly a freshwater group. They resemble functionally the radiolarians, especially with respect to their "diffusion" feeding. They span a wide size range. Some (e.g., Actinopharynx) can be >1 mm, but most species are —40 μm (Actinophrya). Depending on their size, different species feed with little apparent selectivity on algae, flagellates, ciliates, and rotifers. Small heliozoan species account for most of the biomass of amoeboid protozoa in the freshwater plankton. Most, however, are attached to or loosely associated with sediment and other submerged surfaces.

B. Flagellated Protozoa

There is little consensus on how to classify the flagellates, and they are here divided into broad functional groups (Fig. 2). Heterotrophic flagellates are fundamentally important because they are abundant (there are seldom less than 1000 per milliliter, even in the plankton) and because their grazing activities are largely responsible for controlling the abundance of bacteria in aquatic environments. In some taxonomic groups, all species are exclusively heterotrophic (e.g., choanoflagellates and bodonids); others contain many mixotrophs (e.g., the euglenids and chrysomonads), whereas the haptomonads and cryptomonads are dominated by phototrophs and only a minority are capable of phagotrophy. In the past 15 years, a large diversity of heterotrophic flagellates has been discovered. Most of these are choanoflagellates, chrysomonads, euglenids, or bodonids. Some of the more easily recognized species (e.g., Rhynchomonas nasuta) have been recorded from a wide range of habitat types in marine, freshwater, and terrestrial environments.

Many heterotrophic flagellates are anaerobes, including intestinal parasites of man (e.g., Giardia intestinalis), but free-living "diplomonads" (Hexamita and Trypano- nus) and Retortamonas are relatively common bacteria feeders in organically enriched anoxic waters and sediments. Some anaerobic flagellates have hydrogenosomes, the anaerobic derivatives of mitochondria. Almost all of these are endosymbions or internal parasites of animals. They are also known as parabasalans and include many of medical or economic importance, e.g., the human parasite Trichomonas vaginalis, and Trichom- iotis, which degrades cellulose in the hindgut of termites and other wood-eating insects and also supports endosymbiotic methane-producing bacteria (termites may be responsible for a globally significant flux of methane to the atmosphere).

Most choanoflagellates are small (<10 μm), with a feeding filter that forms a "collar" around the single anterior flagellum. The flagellum creates the water current that flows through the filter. Choanoflagellates are exclusively phagotrophic and either solitary (e.g., Actinophrys). They include many of medical or economic importance, e.g., the human parasite Trepomonas vaginalis, and Parvichorbdia. Euglenids usually have two flagella which emerge from a small anterior pocket. This is a large group of planktonic and benthic protozoa, common in both fresh and marine waters. Species associated with surfaces typically keep one trailing flagellum in contact with the substrate, whereas the other pulls the cell forward. Many species are green and phototrophic (e.g., Euglena), but there is a great diversity of phagotrophic (e.g., Astasia, Petalomonas, Entosiphon, and Notoso- leus). These are especially common in marine shallow-water sediments, in which they graze on bacteria. All bodonids are small flagellates with characteristic heterodynamic flagellae (e.g., Bodo and Rhynchomonas). They are almost always found in organically polluted (even anoxic) waters or soils that are rich in bacteria, on which they feed. Many closely related ("kinetoplastid") flagellates are parasites: Ichthyobodo causes "costiasis" in freshwater fish, and the "trypanosome" species are important parasites of man (Trypanosomosis brucei causes sleeping sickness). In heterokont flagellates, each organism has two flagella—one hairy and one smooth, with the former creating the water current that brings mainly heterotrophic food to the (filter-feeding) flagellate. They include the bicosoceans that live in a lorica (Bicosoeca) or are naked (Cafeteria and Pseudodhobodo), that may be stalked.
or colonial, and which resemble the mixotrophic loricate chrysomonads (e.g., Dinobryon). In the phagotrophic chrysomonads (e.g., Paraphysomonas and Spumella), the chloroplast is essentially vestigial (and the cell colorless). Functional types range from heterotrophs (e.g., Paraphysomonas) to mixotrophs (e.g., Dinobryon and Uroglena) and those that are predominantly phototrophs (e.g., Synura and Mallomonas). Chrysomonads are the most abundant heterotrophic flagellates in the freshwater plankton. The helioflagellates (also referred to as “pedinellid stramenopiles,” e.g., Actinomonas, Peridinomona, and Ciliophyta) are filter or diffusion feeders that superficially resemble heliozoa and/or choanoflagellates.

The haptomonads are typically photosynthetic, biflagellated organisms. They include the coccolithophorids and the prymnesiophytes, and some (e.g., Emiliania) are often abundant in the marine euphotic zone. They have an additional flagellar appendage known as the “haptonema.” In the marine Chrysochromulina, this is used in food uptake, but in Prymnesium, which has a very short haptonema, ingestion of food (dinoflagellates, chrysomonads, and green algae) is by means of a posterior pseudopodium. Prey organisms may be immobile by toxins prior to ingestion (toxic bloom-producing species of Prymnesium are responsible for fish kills). The scale and importance of phagotrophy in freshwater haptomonads is unclear.

Most cryptomonads are yellow-brown and phototrophic, with two flagella of similar length (e.g., Cryptomonas). Colorless species can be relatively abundant, especially when associated with benthic detritus. The heteromitids and cyathobodonids represent a
varied assemblage of small, poorly studied flagellates (e.g., Apusomonas, Heteromita, Cercomonas, Cyatho-
boidea, and Kathabophileta), some of which (e.g., Cerco-
monas) produce pseudopodia to ingest bacteria. Others secrete a stalk (Cyathobodo) to attach the cell to the 
substrate. Many are common in soil. Others feed mainly 
on bacteria in the freshwater plankton (e.g., Kathabophile-
aris, which also feeds on small algae) and on sediment 
surfaces. The percolozoans are a diverse collection of 
bacteria-feeding organisms, including amoebophagel-
lates (e.g., Naegleria), the filter-feeding flagellate Percol-
omonas, and anaerobic flagellates with hydrogenosomes 
and endosymbiotic methanogens (Puillitriomonas).

Dinoflagellates are best known as a large group of 
photosynthetic flagellates and for the ability of some 
species to cause "toxic blooms" and "red tides" (Protozoa 
nyaulax causes paralytic shellfish poisoning). However,
about half of all known marine species lack chloro-
plasts. Some of these (in the genera Gymnodinium and 
Amphipleura) feed on cryptomonads and sequester 
their chloroplasts, thus transforming themselves into 
mixotrophs. Moreover, many typically photosynthetic 
species can become mixotrophs through their ability 

to ingest particulate food. The non-photosynthetic and 
mixotrophic species are referred to collectively as het-
erotrophic dinoflagellates. These can be quantitatively 
important in marine food webs, especially as consumers 
of diatoms. Some species are benthic, but little is known 
about these. The diversity of heterotrophic species in 

fresh waters is much less, and it is known mainly from 
the genera Katidinum, Peridinium, Gymnodinium, and 
Ceratium. This may be due in part to the relative rarity in 

fresh waters of large, chain-forming diatoms—a typical 
food item of many marine dinoflagellates. These large 
food items are trapped and digested externally in a 
pseudopodial feeding veil or "pallium." This explains 
why it has often been difficult to recognize ingested 
organisms within dinoflagellates. Other heterotrophic 
dinoflagellates use a feeding tube known as a peduncle. 
The common freshwater dinoflagellate Peridinopsis be-
rotinensis uses such a tube to ingest the fluid contents 
from injured and dying protozoa and small metazoans.

The number of marine dinoflagellate species with a 
recorded capacity for phagotrophy is increasing rapidly. 
These include thecate species (e.g., the peridinoids) 
previously considered to be incapable of phagotrophy: 
It is widely believed that most dinoflagellates—at least 
in marine environments—may be capable of phagotro-
phy. Heterotrophic dinoflagellates (e.g., Oblea, Polyh-
ribos, Heterocapsa, and Protoperidinium) can dominate 
the protozoan biomass in coastal and oceanic waters, 
in which they feed on bacteria, flagellates, diatoms, 
other dinoflagellates, and ciliates. Their biomass can be 
approximately the same as that of ciliates, and the two 
groups may compete with each other for food (although 
the size range of dinoflagellates is slightly greater— 
from 3 or 4 μm in Gymnodinium simplex to 2 mm in 
Northula, with the majority in the size range 20–200 
μm). Also, most planktonic ciliates are incapable of 
ingesting the large diatoms that are consumed by dine-
flagellates.

Almost all stony and stinging corals (Cnidaria) in 
shallow tropical waters harbor photosynthetic dinoflag-
ellates (especially Symbiodinium) as symbionts. These 
consume carbon dioxide and bicarbonate. This 
promotes calcium deposition in the external skeletons 
of their hosts and may enhance the rate with which 
coral reefs are built. In the marine plankton, outbursts 
of parasitic dinoflagellates are sometimes associated 
with the collapse of zooplankton populations. The para-
site Ichthyodinium ichthyodoris destroys the eggs of sar-
dines, Blastodinium sp. in the copepod gut castrates its 
host, and Gymnodinium catenella (a dinoflagellate that 
can cause red tides) is parasitized by another dinoflagellate 
(AMobyphrya ceratii). The fossil record of dinoflagel-
late cysts extends over the past 220 million years, and 
they have a role as biostratigraphic tools.

Dinoflagellates in the marine plankton exhibit "lati-
tudinal cosmopolitanism," such that a morphospecies 
occurring in a circumglobal belt within fairly broad latitu-
dinal limits which correspond to a specific temperature 
range in the surface waters. The recorded species rich-
ness is greatest in tropical marine waters.

**C. Ciliated Protozoa**

This diverse and distinctive group uses cilia for locomo-
tion and feeding (Fig. 3). They demonstrate a consider-
able adaptive radiation of feeding mechanisms and cell 
morphologies (e.g., the ribbon-shaped forms Trachelo-
aphis and Gelesa adapted for life in the marine intersti-
tial). The smallest species tend to feed on bacteria-sized 
particles and the larger species on unicellular algae, 
filamentous cyanobacteria, other protozoa, and even 
rotifers and other microzooplankton. The raptorial 
feeders (e.g., Protospon) use a simple mouth to catch diatoms, dinoflagellates, and other large food items in-
dividually; some (e.g., Chilodinium "hoover") diatoms 
and other elongate food particles from surfaces. The 
filter feeders (e.g., *Cyclidium*) use fine-mesh filters to 

digest and ingest the large diatoms that are consumed by 
dinoflagellates. In many ciliates (e.g., *Oxytricha*, 

...
Aspidisca, and Strombidium), a row of membranelles generates a water current and acts as a relatively coarse feeding filter, and some of these ciliates (e.g., Euplotes) feed most efficiently if they are raised on cirri (fused cilia). Many ciliates are typically sessile and aligned perpendicular to the substrate (e.g., Stentor and Vorticella). Diffusion feeders (e.g., Podophrya and other Suctoria) catch swimming prey (usually other protozoa) that collide with their sticky tentacles.

More is probably known about the biodiversity and ecology of free-living ciliates than any other protozoan group. There are several reasons for this, including the
distinctive and immediately recognizable ciliate morphology and swimming behavior and the relative ease with which many species can be cultured. They are a group of predominantly free-living protozoa. A few are parasites (e.g., _Ichthyophthirius_ infects fish and there is one human endoparasite of minor importance—_Balantidium coli_), but most species are either free-living or harmless commensals of aquatic invertebrates (the ciliate epifauna of marine crustaceans is particularly diverse). Free-living species are known from all natural aquatic habitats in which temperatures are <45°C, including oceanic sinking detritus, freshwater and marine sediments, anaerobic municipal landfill sites, and sewage treatment plants. They are abundant (up to ~10^6 per milliliter) in activated sludge plants, in which they consume bacteria and also flocculate bacteria and other suspended particulate matter. These activities aid the clarification of the effluent and the formation of sludge. About a dozen ciliate species are frequently recorded in used-water treatment plants, although more than 200 ciliate species have been recorded. Ciliates also have a role as indicators of the level of organic pollution in river water.

Most ciliates are in the size range 0.02–2 mm, so they are generally larger than the heterotrophic flagellates and other nanoplankton (<0.02–0.02 mm) on which many of them feed. Planktonic ciliates are relatively abundant (1–100 per milliliter) and important grazers of nanoplankton in marine and fresh waters. They are probably key grazers within the ‘microbial loop’ which is responsible for the rapid remineralization of organic matter in the water column. The diet of planktonic metazooans includes ciliates, although the quantitative significance of this link is unclear. Benthic ciliates are often abundant (>1000 per milliliter) and usually the most important grazers in freshwater sediments (especially in lakes) and marine sandy sediments of inshore waters.

Many ciliate species harbor prokaryotic and/or eukaryotic symbionts. At the oxic–anoxic boundary in the water column of lakes, most ciliates may harbor sufficient endosymbiotic algae (Chlorella) to render the consortia capable of net photosynthesis in dim light. The marine interstitial ciliate _Kentrophoros_ carries ectosymbiotic chemolithotrophic sulfide-oxidizing bacteria, and most anaerobic ciliate species harbor methanogenic bacteria that act as a sink for (potentially inhibitory) H_2 produced by the ciliate. The endosymbiotic nonsulfur purple bacteria in the anaerobic ciliate _Strombidium purpuratum_ use waste H_2 from the ciliate as a reductant for photosynthesis. Most marine anaerobic ciliates carry ecosymbiotic sulfate-reducing bacteria. Many ciliate species are microaerobic, and they seek out the (microbe-rich) oxic–anoxic boundary in sediment or the water column. Some of these species are facultative anaerobes (in the genera _Cyclidium_, _Euplotes_, _Strombidium_, and _Paranophrys_). There are approximately 70 known species of anaerobic free-living ciliate species, mainly in the genera _Metopus_, _Caenomorpha_, _Suprodinium_, _Eupaxilla_, _Trimyema_, and _Plagiopyla_. It is likely that all these contain hydrogenosomes.

Anaerobic ciliates also live as endocommensals in the enlarged forestomach (rumen) of ruminants and in the cecum of other mammals with postgastric fermentation. It is unlikely that any of these ciliates (e.g., in the genera _Dasytricha_, _Entodinium_, and _Polyploea_45–50, 105–106 per milliliter of rumen liquor). They consume bacteria and microscopic fragments of grass and other plant material. Some species are capable of degrading cellulose and other structural carbohydrates, and the endosymbiotic methanogenic bacteria in some species contribute to the methane emitted from ruminants. The economic significance of such large numbers of ciliates living in the rumen is unclear.

Ciliates also live in soil, in which all species are probably capable of producing desiccation-resistant cysts (e.g., _Colpoda_). The abundance of active ciliates in soil is extremely variable and reflects repeated cycles of cyst formation and excystment in response to fluctuating physical factors such as the level of soil moisture. Many species found in soil are frequently found in other (predominantly freshwater) habitats, although only about 100 ciliate species have been described from soil.

D. Others

The following are often referred to as protozoa. Because they are endocommensals or parasites, or incapable of phagotrophy, they do not fall within the definition of protozoa used here (i.e., free-living unicellular phagotrophs). They are mentioned briefly for the sake of completeness but are excluded from Table 1. _Chytridiomycetes_, _Oomycetes_, and _Hyalomyxymycetes_ have the typical fungal characteristics of saprotrophic (absorptive) nutrition and cell walls in the vegetative state. Some are economically important (the _oomycete_ _Phytophthora infestans_ causes potato blight). The _plasmodiophorids_ are cell wall-free endoparasitic slime molds ( _Plasmodiophora brassicae_ causes club root disease in cabbages). The _labyrinthulids_ (e.g., the ‘slime net’ _Labyrinthula_) feed saprotrophically, especially on marine algae, and form complex branching colonies of
spindle-shape cells that move through slime channels. All five groups reproduce by means of unicellular flagellated zoospores (hence, they are also known as "zoosporic fungi").

The opalinids superficially resemble ciliates, but all are mouthless endocommensals living in the hindgut of amphipods and some fish. The microsporidians are a large group (~800 species) of intracellular parasites of animals and other protozoa. They extrude a tubular filament through which the spore is injected into the host cytoplasm. They are believed to be highly deriv'd fungi rather than early diverging eukaryotes. The haplosporidians are a small group of parasites of aquatic invertebrates, notably marine mollusks, and the parahaplosporidians (anaerobic flagellates with hydrogenosomes, e.g., Trichomonas) are all parasites, with one or two possible exceptions (Pseudotrichomonas).

The organisms now referred to as apicomplexans used to be an important component of the "sporozoans." All are intracellular parasites of animals, and many cause life-threatening diseases of man. They share the distinctive morphological feature of an "apical complex" at certain stages in the polymorphic life cycle. The degree of host specificity of many apicomplexans is not known. They include the "gregarines" that infect insects, marine polychaetes, and other invertebrates (e.g., Monocystis in earthworms), and also the "coccidians" parasites including Toxoplasma gondii (the domestic cat is the final host, but intermediate infection in the human can cross the placenta) and about 1050 species are known that infect the digestive tract of mammals, chickens, and other farmed birds. Cryptosporidium is an apicomplexan, and 8 species are known that infect the digestive and respiratory systems of vertebrates. Cryptosporidium parvum infects humans and cattle, and waterborne transmission of oocysts is thought to complete the fecal–oral cycle. Infection can cause life-threatening diarrhea in immunocompromised humans. Cryptosporidium parvum may be sufficiently different from other coccidians to warrant creation of the new higher taxon "cryptosporidiosis." Plasmodium, the causative agent of malaria, is an apicomplexan, as is Pneumocystis, which is an enigmatic parasite that can persist in the human lung for long periods. Pneumocystis carinii hominis causes Pneumocystis pneumonia in AIDS patients. In the early 1970s, the organism was considered to be a fungus; in the early 1980s it was considered to be an apicomplexan, and by the late 1980s it was again considered to be a fungus, possibly an ascomycete. Confusingly, it can resemble Cryptosporidium when viewed with the light microscope. There are approximately 4000 nominal species of apicomplexans.

The Myxozoa are a large group (~1200 species) of parasites, of which the myxosporidians are important parasites of farmed fish (Myxobolus cerebralis causes "tortion disease" in trout). During their life cycle they show both protozoan and metazoan characters. Recent gene sequence data (16S-like DNA and 18S) indicate that myxozoa are greatly modified and simplified metazoans.

Acknowledgments

See Also the Following Articles
EUKARYOTES, ORIGIN OF • MICRORGANISMS, ROLE OF • PLANKTON, STATUS AND ROLE OF • PREDATORS, ECOLOGICAL ROLE OF

Bibliography
PSYCHROPHILES, ORIGIN OF

Richard Y. Morita* and Craig L. Moyer†
*Oregon State University and †Western Washington University

I. Definitions and Historical Background
II. The Environment and Its Microflora
III. Biodiversity of Psychrophiles
IV. Evolution of Psychrophiles
V. Physiology of Psychrophiles (Metabolic Activities)

GLOSSARY

barophile Pressure-loving bacteria.
cryobiosis Anabiosis (latent life) due to freezing.
endolithotrophic Living inside rocks, usually sandstone.
homeophasic adaption The adaptation of the membrane to maintain the bilayer phase.
meltwater Ice or snow melted by radiant energy in the polar regions.
permafrost Ground (ice, bedrock, and soil) that remains frozen below 0°C for more than 2 years.
thermocline In the stratification of warm surface water over cold, deeper water, the transition zone of rapid temperature decline between the two layers.
upwelling Transport of water from the deep ocean to the surface, replacing the surface water that has moved offshore.

PSYCHROPHILES are cold-loving bacteria, whereas psychrotrophs are cold-tolerant bacteria. This article focuses on their environment, biodiversity, and physiology.

I. DEFINITIONS AND HISTORICAL BACKGROUND

Psychrophilic bacteria are defined as cold-loving bacteria. Specifically, their cardinal temperatures are 20°C for maximal growth, 15°C or lower for optimal growth, and 0°C or lower for minimum growth (Morita, 1975), and this definition is accepted by most microbiologists. The old definition of psychrophiles applied to those organisms that produced a visible colony in 1 week at 0°C. However, it is important to recognize that there is a continuum of cardinal temperatures for the diverse microbes in nature. From an ecological standpoint, psychrotrophs and psychrophiles are both found in cold environments, but psychrophiles are not found at temperatures higher than 20°C. This temperature was selected as the maximum temperature for growth based on the fact that laboratory temperatures in the United States are approximately 21 or 22°C. In higher organisms, the cold-loving organisms are known as cryophiles.

Microorganisms capable of growing at 5°C or lower are psychrophiles, regardless of the optimum temperature for growth. The psychrotrophs are cold-tolerant bacteria, but their maximal growth temperature ranges above 20°C and in many cases their optimal growth temperature is also above 20°C. A better term for these organisms that withstand cold temperatures is psycho-
tolerant. However, because of common usage this term should be retained. For example, there are a few texts employing the term “psychrophile” in their title and this term is widely used by industry (mainly dairy and food). e.g., the normal souring of milk is due to psychrotrophs. Furthermore, precedence should be adhered to in keeping the name.

Bacteria capable of growing at 0°C were first reported by Foster in 1887 and 1892. The source material for these bacteria were from fish, natural waters, foods, wastes, rubbish, soil surface, and intestines of fish. The lowest temperature at which bacteria can grow remains to be determined definitely and −12°C is the lowest temperature reported. The term “psychrophile” was first used in 1902 by Schmidt-Nielsen. In 1903, Müller objected to this term because the organisms described actually grow well at higher temperatures. As a result, cold-tolerant bacteria were called cryophile, Glaciale Bakterien, rhizophile, psychrotolerant, psychrocartericus, psychrobe, thermophilic bacteria, facultative psychrophile, obligate psychrophile, and psychrotrophic. A review of the literature indicates that the organisms described were actually psychrotrophs (as defined previously) with the possibility of one psychrophilic exception. For many years, it was thought that there were no bacteria that could be termed psychrophiles, only yeasts and certain algae. The lack of refrigeration equipment in the early days added much to this confusion. Thus, the term psychrophile was considered a misnomer until true psychrophiles were isolated by three different laboratories in 1964. The confusion was ended when Morita (1975) defined the term psychrophile (as noted previously). For further details, Morita (1975) should be consulted.

II. THE ENVIRONMENT AND ITS MICROFLORA

Eighty percent of the earth’s biosphere is permanently cold and the average temperature of the earth is 15°C. Permafrost occupies 20% of the earth’s surface: 80% of Alaska, 50% of Canada, 20% of China, and 50% of Russia are covered by permafrost. However, the amount of microbiological research done on psychrophiles is extremely low compared to that done on thermophilic bacteria.

The polar regions comprise about 14% of the earth’s surface. Approximately 71% of the earth’s surface is ocean, and more than 90% (by volume) of the oceans are 5°C or colder. Other cold environments include caves, the tops of mountains, certain rivers and streams, the upper atmosphere (10°C or less at 1000 m, decreasing as the altitude increases), snow and ice, and the water below the thermocline of freshwater lakes; each has its own microbial flora (Baross and Morita, 1978).

A. Upper Atmosphere

Air samples collected near the earth’s surface up into the stratosphere exceeding 27,000 m have shown the presence of viable bacteria, viable fungi, pollen, and other microscopic particles. A high incidence of viable bacteria has been reported within the troposphere at approximately 10,000 m, where the temperature may be lower than −45°C. The highest incidence of bacteria appears to be at altitudes slightly higher than 10,000 m, where the temperature is lower than 10°C. Although this temperature is well within the cardinal temperature range of psychrophile and psychrotrophs, there are no reports of psychrophilic bacteria in the upper atmosphere. Furthermore, there is organic matter (cobalamin, biotin, and niacin) in the air but the concentration is very low. Thus, these organisms have the ability to resist starvation, drying, freezing, and radiation. These microbes can be the nucleation site for the formation of ice and snow.

However, in the lower atmosphere of the polar regions, nutrient agar plates were media exposed to the air and showed the presence of viable bacteria. Unfortunately, there are no reports of seawater-requiring psychrophilic bacteria from air samples. In all probability, the atmosphere, especially above the polar regions, does contain psychrophiles and psychrotrophs.

B. Caves

There are many glaciated and subterranean caves in which the permanent temperature is about 10°C to below freezing. In addition to the low temperature, there is an absence of light, low levels of organic material, and relatively high moisture. Generally, there is an absence of psychrophiles in these caves, but psychrotrophs are found.

C. Arctic

The similarities of the Arctic and Antarctic polar regions are the continuous sunlight during the summer and the total lack of it during the winter, the presence of sea ice during much or all of the year, and the cold temperature. On the other hand, they differ in that the Arctic has major riverine inputs, but the Antarctic does not.
They also differ in their landmasses, topographical features, large-scale water transport features, and the magnitude of nutrient supply. Relative microbial activities, as measured by the incorporation and respiration of 14C-labeled glucose and glutamic acid (called the heterotrophic potential method by marine microbiologists and substrate-induced respiration by soil microbiologists), are comparable in both the Arctic and Antarctic nearshore water. These measurements were made by the same laboratory working at different times in both regions.

D. Antarctic

About 98% of the surface of Antarctica is ice, leaving 2% of the continent ice free. The lowest temperatures on Earth occur on this continent and it has the lowest precipitation and relative humidity levels, making it the driest area on Earth. As a result, dry valleys and ice-free areas occur in addition to many other types of environments (e.g., temporarily and permanently ice-covered lakes, transient rivers and streams, oligotrophic lakes, and moss and peat bogs), with all sharing low temperatures. Freezing, like drying, will produce hyperthermophiles (as high as 100 cells/g) were reported in Arctic permafrost (peak cores 10,325 years old). Viable microorganisms were reported from Holocene sedimentary deposits and Pliocene deposits (3–5 million years old). Viable fungi, spore-forming bacteria, and non-spore-forming bacteria were in the Antarctic ice sheet up to a depth of 300 m (11,600 years old) but not at 320 m (13,000 years old), whereas in a 107,000- to 200,000-year-old layer of glacier ice, viable organisms were found (Friedmann, 1993). In permafrost material millions of years old, thermophiles, mesophiles, and psychrophiles have been found (Vorobyova et al., 1997). This ancient material does not have sufficient energy in it to keep the organism metabolizing for long periods of time; hence, one has to ask where the energy is coming from. In most cases, the energy supplied has been utilized since any ecological niche has a large number and consortium of microbes that utilize the energy source leaving the more recalcitrant material. This recalcitrant material, if used metabolically, will require exoenzymes to degrade it to its monomers so that it can be transported into the cell. In addition, many substrates that the cell utilizes require an active transport mechanism which also consumes energy. Because of this situation, microbes in ancient material, such as the deeper strata of permafrost, are in an anabolic stage. Specially, the anabiotic stage is cryobiosis, brought about by the low temperature. However, the cells are also in a starvation-survival mode. Both processes generate stress proteins, some of which are shared. The Russian microbiologist Omelyanski (1911) attributed the existence of microbes in ancient cold environments to anabiosis, a term not employed by most microbiologists. Nevertheless, one always must ask ‘where is the energy?’

For further information concerning viable microorganisms in permafrost, Gilichinsky (1994) should be consulted.

E. Permafrost

The discovery of microorganisms in permafrost was initiated in the 1930s in the Trans-Baikal and North Ural regions, Central Yakutia, and Arctic islands (Gilichinsky, 1995). Permafrost cores yielded numerous microbes. Microbes have been reported in all the Arctic and Antarctic permafrost environments except in the lower strata of permafrost ice of Lake Vostok in Antarctica. This exception may be due to the techniques employed. Unfortunately, the early microbiologists did not recognize the abnormal thermostability of psychrophiles, and as a result psychrotrophs and mesophiles were mainly demonstrated. In the early literature, thermophiles (as high as 100 cells/g) were reported in Arctic permafrost (peak cores 10,325 years old). Viable microorganisms were reported from Holocene sedimentary deposits and Pliocene deposits (3–3 million years old).

About 98% of the surface of Antarctica is ice, leaving 2% of the continent ice free. The lowest temperatures on Earth occur on this continent and it has the lowest precipitation and relative humidity levels, making it the driest area on Earth. As a result, dry valleys and ice-free areas occur in addition to many other types of environments (e.g., temporarily and permanently ice-covered lakes, transient rivers and streams, oligotrophic lakes, and moss and peat bogs), with all sharing low temperatures. Freezing, like drying, will produce hyperthermophiles (as high as 100 cells/g) were reported in Arctic permafrost (peak cores 10,325 years old). Viable microorganisms were reported from Holocene sedimentary deposits and Pliocene deposits (3–5 million years old). Viable fungi, spore-forming bacteria, and non-spore-forming bacteria were in the Antarctic ice sheet up to a depth of 300 m (11,600 years old) but not at 320 m (13,000 years old), whereas in a 107,000- to 200,000-year-old layer of glacier ice, viable organisms were found (Friedmann, 1993). In permafrost material millions of years old, thermophiles, mesophiles, and psychrophiles have been found (Vorobyova et al., 1997). This ancient material does not have sufficient energy in it to keep the organism metabolizing for long periods of time; hence, one has to ask where the energy is coming from. In most cases, the energy supplied has been utilized since any ecological niche has a large number and consortium of microbes that utilize the energy source leaving the more recalcitrant material. This recalcitrant material, if used metabolically, will require exoenzymes to degrade it to its monomers so that it can be transported into the cell. In addition, many substrates that the cell utilizes require an active transport mechanism which also consumes energy. Because of this situation, microbes in ancient material, such as the deeper strata of permafrost, are in an anabolic stage. Specially, the anabiotic stage is cryobiosis, brought about by the low temperature. However, the cells are also in a starvation-survival mode. Both processes generate stress proteins, some of which are shared. The Russian microbiologist Omelyanski (1911) attributed the existence of microbes in ancient cold environments to anabiosis, a term not employed by most microbiologists. Nevertheless, one always must ask ‘where is the energy?’

For further information concerning viable microorganisms in permafrost, Gilichinsky (1994) should be consulted.

F. Oceans

The area below the thermocline is approximately 3°C or colder. Thus, the deep sea is permanently cold. Grad-
ualy, as the higher latitudes are approached, the depth of the thermocline decreases until it is at the surface of the ocean. In the deepest portions of the oceans (trenches and deeps), barophiles can be isolated that are also psychrotrophic or psychrophilic. At the edge of the ice where ice is melting or being formed, there is a sea ice environment with an associated microbial community. All types of microorganisms can be demonstrated in waters below the thermocline, mainly because there are many different types of habitats available.

G. Sea Ice

Where the seawater is being frozen, a sea ice microbial community resides. The freezing process helps concentrate the dissolved nutrients, and when the ice melts the concentrated nutrients in the ice are released to permit the growth of this microbial community. Thus, this community is influenced mainly by the onset of spring and winter. Within this sea ice community a variety of bacteria are found displaying numerous morphological types. Seventy percent of the bacteria in the sea ice community are free living, and ice nucleation bacteria can also be found. Various pigmented and gas vacuolate bacteria are also present and 65% of the epiphytic bacteria are found associated with the diatom Amphiprora. The best growth response of the sea ice microbial community occurs between 5 and 10°C.

III. BIODIVERSITY OF PSYCHROPHILES

The first known species of psychrophiles described taxonomically are Vibrio (Moritella gen. nov.) marinus MP-1 and Vibrio (Colwellia gen. nov.) psychroerythrus, both isolated in 1964. The biodiversity among psychrophiles in the various cold environments has yet to be studied extensively. Nevertheless, the various species within the genera Achromobacteria, Alcaligenes, Alteromonas, Aquaspirillum, Arthrobacter, Bacillus, Bacteroides, Brevibacterium, Cleistidium, Colwellia, Cytophaga, Flavobacterium, Gelibacter, Methanococcales, Methanogenium, Methanosarcina, Microbacterium, Micrococcus, Moritella, Octadecabacter, Phormidium, Photobacterium, Polari-
methanogenium frigidum, a psychrophilic, slightly halophilic, \( H_2 \)-using methanogen, was isolated from the perennially cold, anoxic hypolimnion of Ace Lake, Antarctica (Franzmann et al., 1997). Although it may be too early to state that this may be an indication of the evolutionary processes affecting both hyperthermophiles and psychrophiles, this does demonstrate that members of the domain Archaea are also capable of psychrophily across the domain Bacteria.

Incredibly, as much as 30% of the marine picoplankton (planktonic organisms with an average diameter of 0.2–2.0 \( \mu \)m) from both polar and temperate coastal waters are also Archaea (i.e., archaeoplankton) and the majority of these are associated with the Crenarchaeota. Despite this cosmopolitan distribution in the world's oceans, little is known about the physiological properties of these archaeoplankton, with the exception that they are hypothesized to potentially be psychrophilic based on the marine sponge symbiont Cnarcharum symbium (Preston et al., 1996). However, similar archael phylotypes have been located at a deep-sea hydrothermal vent system (Meyer et al., 1998) in waters with an environmental temperature of 15–30°C. Many hyperthermophiles are members of the Archaee that can utilize \( H_2 \) as an energy source, and recently Methano-
nogenium frigidum, a psychrophilic, slightly halophilic, \( H_2 \)-using methanogen, was isolated from the perennially cold, anoxic hypolimnion of Ace Lake, Antarctica (Franzmann et al., 1997). Although it may be too early to state that this may be an indication of the evolutionary processes affecting both hyperthermophiles and psychrophiles, this does demonstrate that members of the domain Archaea are also capable of the psychrophilic lifestyle and that further research is needed. The main deterrent is the lack of isolation of psychrophiles, mainly because interest in this thermal group was lack-
ing for many years. Yet to be isolated from the meltwater
in the Antarctic are psychrophilic bacteria that are also alkaliphilic. Each body of water in the Antarctic has its own cations and anions and the salinity may be very high due to freezing of the water.

In any cold environment in which microbes have been isolated, many of the isolates are psychrotrophs. When all things are equal, the psychrophiles will outgrow the psychrotrophs at low temperature.

IV. EVOLUTION OF PSYCHROPHILES

The common approach used in microbiology is to apply phylogenetic analysis to establish evolutionary relationships among organisms and to use this as a framework for making inferences about community structure, making inferences about genetic and thereby inferred organismal diversity, and (to a lesser degree) to infer physiological adaptation when applicable. This approach is possible due to the detailed theory of evolutionary relationships among the domains Bacteria, Archaea, and Eucarya that has emerged from comparisons of ribosomal RNA “signature” sequences. Most researchers currently infer that the earliest common ancestor to all three domains was a H$_2$-metabolizing hyperthermophile based on the deepest branching lineages found in the domains Bacteria and Archaea (Pace, 1991). If we accept this premise, then the thermophiles must have evolved from the hyperthermophiles, followed by the mesophiles and finally the psychrophiles. This hypothesis has been challenged by criticism that the hot origin of life scenario is not compatible with the RNA world hypothesis (because of the instability of RNA at near water boiling temperatures) and/or by the argument that hyperthermophiles are not primitive but have instead successfully adapted to temperatures higher than the limits imposed on other life-forms (Forterre, 1996).

In addition, dibiphytanyl ether lipids used in the formation of lipid “monolayer” cell membranes in hyperthermophilic Archaea were found in nonthermophilic members of the kingdom Crenarchaeota, the Euryarchaeota, the Flexibacter–Cytophaga–Bacteroides group, the gram-positive Bacteria group, and Proteobacteria. The majority of individual isolates are located in lineages contained either by the Flexibacter–Cytophaga–Bacteroides group or by the Moritella, Colwellia, and Shewanella subgroups of the γ-Proteobacteria.

V. PHYSIOLOGY OF PSYCHROPHILES (METABOLIC ACTIVITIES)

Psychrophiles were not isolated until 1964 mainly because of the fact that previous investigators did not recognize the abnormal sensitivity of this thermal class. Although it was known that most marine bacteria would not withstand the plating temperature of agar, microbiologists working in the Arctic and Antarctic did not take the precaution of keeping their sample material, media, pipettes, loops, and needles cold before use. Even for some psychrophiles, 10°C is too warm and the organisms will expire. Many assume that growth at low temperature is slow. This assumption was shown to be incorrect when Moritella marina was reported to have a doubling time of 17.9 h at 3°C (DeLong et al., 1986). Recently, a nonhyperthermophilic earliest common ancestor was hypothesized based on the correlation between the G + C nucleotide content of ribosomal RNA sequences with the optimal growth rate temperature found in prokaryotes (Galtier et al., 1990). If the nonhyperthermophilic hypothesis is presumed to be correct, it then implies that mesophiles were most likely the first to evolve with thermophily and psychrophily adaptations followed thereafter. Psychrophily undoubtedly has evolved multiple times, thereby giving rise to polyphyletic origins (i.e., psychrophilic members are not contained within lineages spanning the divergence of divisions and domains).

A phylogenetic tree generated exclusively from SSU rDNA sequences of cultivated obligate psychrophiles demonstrates five major lineages spanning the prokaryotic domains of Bacteria and Archaea (Fig. 1): the Crenarchaeota, the Eurarchaeota, the Flexibacter–Cytophaga–Bacteroides group, the gram-positive Bacteria group, and Proteobacteria. The majority of individual isolates are located in lineages contained either by the Flexibacter–Cytophaga–Bacteroides group or by the Moritella, Colwellia, and Shewanella subgroups of the γ-Proteobacteria.
FIGURE 1  Radial phylogenetic tree using the neighbor-joining distance method demonstrating the evolutionary relationships among cultivated psychrophiles. The tree was constructed using complete SSU rDNA sequences contained in the Ribosomal Database Project from the 99/99 release of Version 7.1 (http://www.cme.msu.edu/RDP/html/index.html), with the additions of Cenarchaeum symbiosum and Moritella sp. ANT-300. The scale bar represents 0.10 fixed mutations per nucleotide position.

A. Enzymes

Unlike thermophiles, research on the physiology of psychrophiles has been neglected over the years. The few studies dealing with enzymes (and cytoplasmic proteins) from psychrophiles indicate that they are more thermostable compared to their counterparts in mesophilic bacteria—their maximum, minimum, and optimum temperatures for activity are lower. However, most of these enzymes will operate several degrees above the maximum growth temperature of the psychrophiles. Malic dehydrogenase, the first enzyme from a psychrophile to be studied, obtained from cells of M. marinus was found to be stable between 0 and 15°C and inactivation occurred between 15 and 20°C. This organism’s optimum growth temperature is 15°C and its maximum is 20°C. If the enzyme is partially purified, inactivation at higher than 20°C becomes very pronounced. Phosphofructokinase/glyceraldehyde-3-phosphate dehydrogenase complex, lactic dehydrogenase, and aldolase from this same organism...
lost activity when incubated for 1 h to 35–40°C. Surpris-
ingly, glyceraldehyde-6-phosphate dehydrogenase was
found to be stable at 20°C, however, at 36°C for 1 h,
it lost 90% of its activity. Trisphosphate isomerase,
a glycolytic enzyme, isolated from Moritella sp. strain
ANT-300 (maximum growth temperature of 12°C) is
extremely heat sensitive and has a half-life of heat inact-
vation of 520 s at 25°C. Other enzymes that have been
examined from psychrophiles include α-amylase, li-
pase, protease, alkaline phosphatase, and β-lactamase.
As a general rule, the enzymes isolated from psychro-
philes have not adapted entirely to correspond with the
cardinal temperatures of the respective psychrophile
strain (Russell, 1992). Membranes of psychrophiles and
psychrotrophs contain more polyunsaturated short
chains and branch and/or cyclic fatty acids than do
mesophiles. Therefore, the ability to alter
the fatty acid composition may be restricted to psychro-
philes and not psychrotrophs.

It should be noted that when the psychrophile Morit-
ella sp. strain ANT-300 was starved, there was induced
an increase in fatty acids (e.g., fatty acid 16:1 increased from 42 to 62.5%). When
starved, cells of Moritella sp. strain ANT-300 (as with
many other bacteria) underwent fragmentation (i.e.,
reductive division), resulting in ultramicrocells. The
starvation-survival state is the normal state of most
microbes since most of the biosphere is oligotrophic.
Thus, it appears that various species of psychrophiles
evolved from their mesophilic component and they in-
habit permanently cold environments. Their upper
and lower temperature limits are 20 and −12°C, respec-
tively.

See Also the Following Articles
ANTARCTIC ECOSYSTEMS • ARCHAEA, ORIGIN OF • ARCTIC ECOSYSTEMS • BACTERIAL BIODIVERSITY • MICROBIAL BIODIVERSITY • THERMOPHILES, ORIGIN OF

Bibliography


RAINFOREST ECOSYSTEMS, ANIMAL DIVERSITY

Gregory H. Adler
University of Wisconsin at Oshkosh

GLOSSARY

allopatric speciation The evolutionary development of new species in the presence of a geographical barrier which reduces gene flow and promotes genetic divergence.

neotropics A biogeographic region that includes the New World tropics, extending from southern Mexico south through the Southern Cone of South America to Tierra del Fuego. Many different ecosystems are found here, including tropical rainforest.

species diversity This has two connotations. In a broad sense, it simply refers to the number of species of a particular taxonomic group living within a given area and is used synonymously with species richness. In a narrow sense, it refers to the number of species within a given area while simultaneously taking into account their relative abundances. In this article, species diversity is used in the broad sense and is used interchangeably with species richness.

species richness The number of species of a particular taxonomic group living within a given area.

TROPICAL RAINFORESTS contain more species of animals than any other ecosystem on Earth. This article reviews the distributions and types of rainforests and the diversity of selected groups of animals in tropical rainforests. The article concludes with a discussion of global and regional patterns of rainforest animal diversity, some hypotheses that have been developed to explain the high diversity, and the future of rainforest animal diversity.

It is indeed ironic that Linneaus, cataloging specimens collected by early naturalists for the museums of Europe, considered the tropics to be of low diversity. However, the first naturalists collected solely in villages that contained second-growth vegetation and where hunting was intense. It was not until naturalists penetrated the dense border of vegetation between villages and undisturbed forest that the high diversity became evident. Single collecting expeditions routinely returned to Europe with thousands of undescribed species of plants and animals. Discoveries of new species of animals continue into the twenty-first century. It has been estimated that several million species of insects alone are yet to be described from tropical rainforests. Even larger and presumably more conspicuous animals,
such as birds and mammals, are being discovered every year in the world’s tropical rainforests. For instance, in the 1990s new species of monkeys were found in Brazil and two new species of deer and a relative of the goat were discovered in Laos and Vietnam.

I. DEFINITION AND GEOGRAPHICAL CONTEXT

Tropical forests cover much of the land area that lies between the Tropics of Cancer and Capricorn (23°30’ N and S latitude). Tropical forests also extend outside of the two tropics in southeastern Brazil and northeastern Australia. The lowland tropics have a relatively high constant temperature, where the mean temperature of the coldest month is at least 18°C. However, not all warm areas in the tropics support tropical forest. Rainfall, in concert with high stable temperature, is a crucial factor determining the development of forest. In many areas, closed-canopy forest begins to develop where there is at least 800–1000 mm of rain per year. Although the term rainforest has been applied loosely and sometimes inappropriately to virtually any type of tropical forest, a more precise definition of rainforest has been developed. Tropical rainforests occur in areas of high, relatively constant temperature (below approximately 1000 m in elevation) and where rainfall exceeds 2000 mm per year. A dry season may occur but generally does not exceed 4 months in duration. Forests growing under such conditions are composed primarily of evergreen trees (<13% of the trees are deciduous) and have a closed canopy of at least 25 m in height. Larger trees, called emergents, often protrude above the canopy and may reach heights of up to 30 m. Thus, rainforests are defined based on the temperature and precipitation regimes under which tropical forests develop and also on the structural characteristics of these forests.

An understanding of the geographical context of rainforests is necessary to a discussion of animal diversity in these forests. For historical reasons (i.e., evolutionary origins of taxa in specific geographical regions and subsequent inability to disperse to other regions), many groups of animals are geographically restricted to single tropical regions. Vertebrates illustrate this point. The species-rich frog family Leptodactylidae is restricted to the New World and reaches its greatest diversity in rainforests. Iguanas (Iguanidae) and anoles (Polychrotidae), two species-rich families of lizards with greatest diversity in rainforests, are restricted to the New World and reaches its greatest diversity in rainforests. Iguanas (Iguanidae) and anoles (Polychrotidae), two species-rich families of lizards with greatest diversity in rainforests, are restricted to the New World, except for iguanas in Fiji that stem from New World colonizers. Families of birds such as toucans (Ramphastidae), motmots (Momotidae), jacamars (Galbulidae), puffbirds (Buconidae), woodcreepers (Dendrocolaptidae), ovenbirds (Furnariidae), antbirds (Thamnophilidae), broadbills (Eurylaimidae), sunbirds (Nectariniidae), white-eyes (Zosteropidae), and pittas (Pittidae) are restricted largely to the tropics of the Old World (with a few species found also in adjacent subtropical regions). Birds of paradise (Paradisaeidae) and bowerbirds (Ptilonorhynchidae) are restricted to the tropics of Australia and New Guinea. Marsupials, which reach their greatest diversity in rainforests and surrounding areas, are restricted to the New World and the Australopapuan region. Among primates, two families of monkeys (Callitrichidae and Cebidae) are restricted to the New World tropics, whereas Cercopithecidae is restricted to the Old World and Hyl注册资本 is restricted to Southeast Asia. The lemurs, a group of primates composed of five families, are found solely in Madagascar. Rainforests are found in three major regions of the world. The rainforests of the neotropics cover the greatest area (approximately 4 million km²) and extend from southern Mexico to southeastern Brazil. The neotropics contain four fairly distinct regions that are isolated from each other by mountain ranges, savannas, or scrub vegetation. Trans-Andean rainforests are distributed from Chiquitos and Valdivian, Mexico, to the Pacific slope of Colombia and Ecuador. Venezuelan coastal rainforests are found in northern Colombia, Venezuela, and the Guianas. Amazonian rainforests occupy most of the Amazon basin and constitute the largest contiguous area of tropical rainforest in the world. Atlantic rainforests formerly covered a long narrow swath of coastal southeastern Brazil but have been reduced to remnant patches by human activity. Only about 10% of the Atlantic rainforest remains. The Asia-Pacific tropics, also known as Malesia, extend from China’s Yunnan Province southward to Australia. Included in this region are large parts of Southeast Asia (China, Thailand, Malaysia, Indonesia, Singapore, Cambodia, Laos, and Vietnam) and the Australopapuan region (New Guinea, northeastern Australia, and tropical Pacific islands such as New Caledonia, the Bismarck Archipelago, Solomon Islands, Fiji, and Vanuatu). Asia-Pacific tropical rainforests also occur far to the west in the Western Ghats of India. This discontinuous and largely insular block of rainforest is second in areal extent only to the neotropics, covering approximately 2.3 million km². The Aus-
II. TYPES OF RAINFORESTS

Researchers frequently delimit lowland tropical forests based on annual amounts of rainfall such forests receive and the temporal distribution of rainfall. These delimitations are somewhat arbitrary, and it is important to recognize that contiguous forest often occurs along a rainfall gradient where one type of forest grades into another with little obvious demarcation. Tropical scrub forest and tropical dry forest typically receive less than 2000 mm of rain per year, most of which falls in a few months. Severe dry seasons of up to 8 or 10 months punctuate the short rainy seasons, and many trees are deciduous and shed their leaves during the dry season to conserve water. Such forests, particularly scrub forests, often do not have a closed canopy. If we adhere to our definition of rainforests given previously, then these tropical forests do not constitute true rainforests (although they frequently abut rainforests and share many species of animals), and this article will not discuss their animal diversity. The remaining forest types that will be discussed are sometimes referred to as semievergreen and evergreen forests and conform to the definition of tropical rainforests. Tropical moist forests receive from about 2000 to 4000 mm of rain per year. These forests frequently have a pronounced dry season of up to 4 months, but heavy rains may occur during the dry season. Deciduous trees may occur in such forests, but they are much less frequent than in the preceding two types of forests. Tropical moist forests occupy the greatest areal extent of rainforests and occur throughout much of the three major geographical rainforest regions. Tropical wet forests typically receive more than 4000 mm of rain per year and frequently have no extended dry season. The wettest of these forests (those forests receiving more than 8000 mm of rain per year) are designated as pluvial forests and are essentially aseasonal. The "driest" months in these forests typically receive more rain than the wettest months in tropical moist forests.

Tropical forests may also be defined based on the elevations at which they occur. In mountainous areas, contiguous forest may be found from near sea level to the tree line. However, forest characteristics change dramatically with elevation. The upper elevational limit of lowland rainforests is frequently considered to be 1000 m, where premontane forest often occurs, although this limit is arbitrarily chosen. Lower montane and upper montane forests, which do not fall within the operational definition of lowland rainforests, occur at increasingly higher elevations. In some areas, montane forests may occur at elevations lower than 1000 m, depending on latitude and stature of mountain ranges. For instance, small mountain ranges and lower spurs from larger ranges often support montane forests at lower elevations, which is termed the Massenerhebung effect. This article will focus on animal diversity in lowland rainforests (i.e., tropical moist and wet forests lower than 1000 m).

III. OVERVIEW OF ANIMAL DIVERSITY

Tropical–temperate comparisons of diversity are commonly made to highlight the high diversity of the tropics, and this latitudinal diversity gradient has attracted considerable attention from ecologists and evolutionary biologists. These comparisons show that many major groups of animals reach their greatest diversity in tropical rainforests. Conspicuous among groups of animals that reach their greatest diversity in such forests are beetles, butterflies, frogs, birds, and mammals. This article will focus on these groups because they best exemplify the diversity of rainforest animals, and tropical–temperate comparisons will be included to highlight rainforest diversity.

It should be noted, however, that other major groups of animals, particularly among invertebrates, have been poorly sampled in tropical rainforests. It is evident, however, that other such groups also exhibit high rainforest diversity. For example, 498 species of spiders in 33 families were collected during 6 weeks at Pakitza in Manu National Park, Peru. Many of these specimens represented new species. Several weeks of sampling at the same site yielded 224 species of caddisflies (order Trichoptera), 60% of which could not be identified confidently. Sampling of dragonflies and damselflies (order Odonata) in the same area yielded 117 lowland rainforest species at Tambopata Reserve in Peru, 135
species of ants were found, and one individual tree harbored 43 species, which is approximately the total number of ant species found in the entire British Isles or in all of Canada. The 1500-ha La Selva reserve in Costa Rica contains an estimated 4000 species of moths. Thus, when invertebrates have been sampled in tropical rainforests, diversity is usually very high, and new species invariably are found.

IV. DIVERSITY OF BEETLES

Approximately 1.4 million species of organisms have been described, about 400,000 of which are beetles. Beetles therefore constitute one-fourth of all living species that have been cataloged by taxonomists. Beetles have been poorly sampled in rainforests, and any discussion of their diversity will be incomplete. However, it is clear from even preliminary sampling that beetle diversity is extremely high and that beetles no doubt are the most species-rich group of rainforest animals, comprising an estimated 40% of all arthropods. Indeed, the majority of rainforest animal species may be still undescribed beetles.

There has been considerable conjecture over how many species of beetles occur in rainforests. Rainforest beetles reach their greatest richness in the canopy, in which standardized sampling efforts have been concentrated. Such sampling involves the deployment of insecticidal sprays into the canopy, a process known among coleopterists (scientists who study beetles) as fogging. Sampling of beetles (excluding weevils) by fogging 19 individuals of a single species of rainforest tree (Lueheea semmanni) in Panama produced 955 species, 163 species of which were found only on this 1 species of tree. This preliminary sampling was used to estimate the total number of arthropod species in tropical rainforests (Erwin, 1982). The estimate was based on the number of beetle species restricted to a single species of tree, the approximate number of tropical tree species, the proportion of arthropods that are beetles, and the contribution of canopy species to the overall arthropod species pool. Based on this reasoning, estimates of 30–50 million species of insects and 7 million species of beetles have been derived, which indicate that insects may constitute about 97% of all species of living organisms on Earth and that tropical rainforest beetles contribute the greatest proportion of species. These estimates of global biodiversity based on rainforest beetles stimulated biologists to reassess their more traditional estimates of biodiversity.

Sampling of ground beetles (Carabidae) at Pakitza in Manu National Park, Peru (not all species of which are confined to the ground), has revealed this 4000-ha site to have the greatest species richness of this family ever recorded. Indeed, nearly as many species have been found in this small area as have been found in all of New Guinea. More than 600 species have been recorded from Pakitza, and not only many new species but also new genera have been discovered. It should be noted that sampling was incomplete and that dozens of additional species are expected.

Comparison of beetle species richness among sites is difficult because of differences in sampling methodology and intensity and the paucity of sufficiently rigorous studies. However, several studies using the standardized fogging methodology allow researchers to calculate species richness on equal footing as the number of species/m³/m. Among canopy and subcanopy beetles, estimates range from 0.02 species/m³/m in an Australian rainforest tree and in rainforest in Sulawesi to 1.5 species/m³/m in a Peruvian rainforest. Other estimates for rainforest sites are 0.29 species/m³/m in New Guinea, 0.32 species/m³/m in Brunei, and 1.17 species/m³/m in Panama. The estimate for Panama was derived from the same data set from which global biodiversity estimates were made.

V. DIVERSITY OF BUTTERFLIES

Unlike most groups of insects, butterflies have been studied so extensively that about 90% of all living species have been described. An estimated 13,750 species of butterflies, excluding skippers (family Hesperiidae), occur in the world. With skippers included, the number is approximately 17,900 species. Because of our thorough understanding of butterfly systematics, they are perhaps the most appropriate insects for examining rainforest diversity.

A tropical–temperate comparison of butterfly diversity highlights the great richness of tropical rainforest butterflies. Table I shows the numbers of butterfly species in four families in temperate forests of West Virginia and tropical rainforests of Costa Rica. These two areas are of comparable size, and the butterfly faunas in the four families are well known. Butterflies in two additional families (Lycaenidae and Hesperiidae) are too poorly known to make a valid comparison with those in West Virginia. Included in the species tally for West Virginia are those species known to occur in all forest types below 1000 m, whereas the species tally for...
TABLE I

Approximate Numbers of Butterfly Species in Four Families in Lowland Rain Forests of Costa Rica Compared with Lowland Temperate Forests of West Virginia

<table>
<thead>
<tr>
<th>Family</th>
<th>West Virginia</th>
<th>Costa Rica</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papilionidae</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>Pieridae</td>
<td>4</td>
<td>31</td>
</tr>
<tr>
<td>Riodinidae</td>
<td>1</td>
<td>208</td>
</tr>
<tr>
<td>Nymphalidae</td>
<td>16</td>
<td>234</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>498</td>
</tr>
</tbody>
</table>

Costa Rica includes only those species that are known to occur in lowland rainforests. Thus, butterflies occurring in both dry and wet temperate deciduous forests are included in the tally for West Virginia, but habitats such as tropical scrub and dry forests and mangrove forests are excluded from the tally for Costa Rica.

It is important to note that species tallies will likely vary depending on definitions of forest types and the compilation of results from additional sampling and natural history studies. Many species of butterflies live in disturbed areas adjacent to rainforest but are not included in the species tally because they are not true denizens of the rainforest.

Regardless of such qualifications, it is clear that rainforest butterfly diversity is much higher than in temperate forests. Costa Rican rainforests harbor nearly 20 times as many species of butterflies in the four families. Of particular note is the extremely high relative diversity in two Costa Rican families, Riodinidae and Nymphalidae, which together constitute nearly 90% of Costa Rica’s rainforest butterfly fauna. Although the majority of West Virginia forest species are nymphalids, only one species of riodinid is found in West Virginia’s lowland forests, and the two families together constitute only 65% of the total butterfly fauna.

VI. DIVERSITY OF FROGS

Frogs are conspicuously species rich in tropical rainforests, and their richness is particularly evident when breeding. At a single site in the Amazon rainforest in Santa Cecilia, Ecuador, 81 species of frogs have been recorded, which to date is the highest tally of frog species for any single site in the world. The number of species at this site equals the total number of frog species in the United States. High frog species richness has also been recorded from other tropical rainforest sites. In Panama, 59 species of frogs have been recorded on the 1500-ha Barro Colorado Island. More than 40 species have been recorded from La Selva, Costa Rica, 38 species from Makourou, Gabon, 25 species from Pasoh, Malaysia, and 23 species from Gogol, Papua New Guinea. On the island of Borneo, 33, 51, and 46 species of frogs have been recorded at three different lowland rainforest sites, respectively, and 29 species of larval frogs have been recorded from rainforest streams at Nanga Tekalit in Sarawak. At Sakaerat, northern Thailand, 19 species of frogs have been recorded within rainforest, with an additional 5 species found in adjacent dry forest.

VII. DIVERSITY OF BIRDS

Tropical rainforests typically harbor the greatest diversity of birds compared with other ecosystems. Costa Rica’s bird fauna is well-known, and a comparison of the number of resident bird species in that country’s lowland rainforests with resident birds in all forest types in West Virginia illustrates this point. Table II gives the approximate number of bird species that are resident in Costa Rica’s lowland rainforests compared with the number that are resident in lowland forests of all types in West Virginia. Excluded from this list are migratory...
<table>
<thead>
<tr>
<th>Order and Family</th>
<th>West Virginia</th>
<th>Costa Rica</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Species</td>
<td>Species</td>
</tr>
<tr>
<td>Tinamiformes</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Ciconiiformes</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Ardeidae</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Threskiornithida</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Falconiformes</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Cricetidae</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>Rallidae</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cracidae</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Columbiformes</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Columbidae</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Picae</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Picidae</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Tityridae</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Tyrannidae</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

Costa Rican rainforests contain more than eight times as many species of birds as do all lowland forest types in West Virginia. Much of this greater diversity stems from two taxonomic sources. First, Costa Rican rainforests harbor more families of birds with resident species (42 compared with 18 in West Virginia). There are 29 families of birds that have representatives that are resident in Costa Rican rainforests but not in West Virginia's forests, and these families contribute 218 spe-
cies to the total rainforest avifauna. In contrast, only 7 families have resident species in West Virginia’s forests but not in Costa Rican rainforests, and these families contribute only 10 species to West Virginia’s resident forest avifauna. Second, several families that have resident representatives in both areas are much more diverse in Costa Rican rainforests. For example, approximately 21 species of hawks are rainforest residents, compared with only 4 resident forest species in West Virginia; 12 species of pigeons and doves are rainforest residents, but only 1 species is resident in West Virginia’s forests; 44 species of tyrant flycatchers are rainforest residents but only 1 is a resident in West Virginia (although many species are migratory); and 14 species of wrens are rainforest residents but only 2 species are resident in West Virginia. This comparison shows that bird diversity is higher not only at the species level but also at higher taxonomic levels and highlights a common theme in tropical–temperate comparisons of diversity: There is often greater representation at higher taxonomic levels such as genus and family in tropical rainforests, and there are generally several higher level groups that are particularly rich in species in rainforests.

High bird species richness is typical of tropical rainforests, and Costa Rica’s rainforests may be considered about average with respect to number of species. Approximately 410 species of birds have been found at La Selva in Costa Rica, an area of approximately 1300 ha. However, not all of these species are truly rainforest birds. If only rainforest birds are included (i.e., birds frequenting open water, marshes, cleared areas such as pastures, etc. are excluded), the total list is approximately 338 species. Approximately 314 rainforest species have been found on Barro Colorado Island and adjacent mainland areas, with a list of all species nearing 400. By contrast, about 354 species of birds have been recorded from Cocha Cashu, Peru, with nearly 460 truly rainforest species. To date, 575 species of birds have been found adjacent to the Explorer’s Inn Reserve, an area of 5500 ha in the Amazon rainforest of southern Peru. The majority of these birds are rainforest species.

Other rainforest regions also have high bird diversity. Khao Yai National Park, Thailand, harbors 318 species, approximately 216 of which are true rainforest species. A single plot of <50 ha in southern Vietnam contains 164 species of forest birds. Makoukou, Gabon (2000 km²), has 342 species of birds; 212 species have been recorded from Pasoh, Malaysia (800 ha), and 162 species from Gogol, Papua New Guinea (1000 ha). Although not all species recorded from these last three sites are true rainforest residents, the majority can be considered rainforest species.

VIII. DIVERSITY OF MAMMALS

Table III highlights the high diversity of rainforest mammals by comparing the approximate number of species that are found in Costa Rica’s lowland rainforests with that of lowland forests of West Virginia (including species that were extirpated following European colonization). Costa Rican rainforests contain nearly three times as many species of mammals as do forests in West Virginia. A comparison of families shows the same trend with birds. There are 31 families represented in Costa Rica’s rainforest mammal fauna and only 17 represented in West Virginia forests. The 20 families that have rainforest representatives in Costa Rica but not in West Virginia forests contribute 97 species. In contrast, 6 families contributing only 8 species have lowland forest species in West Virginia but not in Costa Rican forests. Particularly noteworthy are bats (order Chiroptera). Only 11 species of bats, all from a single family, are found in West Virginia forests, but approximately 89 species of bats in 9 families are found in Costa Rican rainforests. A single family, Phyllostomidae, contains 56 rainforest species. However, not all higher level taxonomic groups are more species rich in rainforests. For instance, the order Insectivora contains 9 lowland forest species in West Virginia but only 2 in Costa Rican rainforests. These 2 species are confined to the upper elevational limits of rainforest in Costa Rica and reach greater richness in montane forests. Similarly, there are 18 species of carnivores in West Virginia forests (4 of which have been extirpated from the state) and only 16 in Costa Rican rainforests.

Mammal diversity at single sites within rainforests is concomitantly high. However, species lists for most sites are certainly low because of inadequate sampling. In La Selva, Costa Rica, the mammal list comprises 117 species, with as many as 138 species expected to occur there based on geographical distributions of species that overlap the area but have not been recorded. As with birds, Costa Rica’s rainforest mammal diversity can be considered average. Barro Colorado Island’s list comprises 113 species, with as many as 144 species expected. Several study sites located in Amazonian rainforests have species lists of >130 species, with nearly 203 species expected from these sites. Makoukou’s list...
of mammals is 199 species, Pasoh’s is 89 species, and Gogol’s is 27 species. By far the most species-rich order of mammals in these rainforests is bats (Chiroptera). Even in well-studied sites, though, bat communities are incompletely known because of difficulties in sampling species such as those that forage above the canopy. New techniques that allow researchers to identify bats based on ultrasonic recordings are adding substantially to bat inventories at sites that have been sampled for decades by more traditional means such as mist netting.

IX. GEOGRAPHICAL PATTERNS OF ANIMAL DIVERSITY

Researchers have searched for both global and regional patterns of animal diversity. In particular, researchers have attempted to identify the major rainforest region that harbors the greatest diversity. The answer depends to some extent on the taxonomic group under consideration. At higher taxonomic levels (e.g., class and order), however, it is clear that the neotropics are generally richest in species. As noted for beetle species richness, Peru and Panama have the greatest numbers of recorded species per volume of canopy and subcanopy foliage, with the Asia-Pacific tropics, particularly Australia, having the lowest richness. Among butterflies, the neotropics again exceed other tropical regions with respect to numbers of species. Comparing species lists of countries or regions of approximately comparable size that occur in the tropics and that have substantial rainforest highlights this point. Panama in the neotropics has approximately 1550 species of butterflies, Liberia in the African tropics has about 720 species, and the Malay Peninsula in the Asia-Pacific tropics has 1031 species. Although not all species in these lists are true rainforest species,
the trend of higher butterfly species richness in the neotropics is clear.

Among frogs, a similar trend is evident, with the Neotropical rainforests having the highest species richness. The greatest single-site frog species count is in Ecuador, with other Neotropical sites in Peru and Central America having nearly as many species. Sites in the African and Asia-Pacific tropics have considerably lower frog species richness, with the lowest being in insular sites in the Australopapuan region and rainforests of mainland Asia such as Thailand. However, some sites on the island of Borneo have nearly the same levels of frog diversity seen in some parts of the neotropics.

Ornithologists have long referred to South America as the bird continent because of its high bird diversity; approximately one-third of all species of birds occur on this continent. Within single rainforest sites, the highest bird diversity clearly occurs in the neotropics of Amazonia in Peru and surrounding countries on the eastern slope of the Andes. Neotropical rainforests are followed by the African and finally the Asia-Pacific rainforests. Lowest rainforest bird diversity on a per site basis is found in the Australopapuan region, particularly on smaller islands.

Panglobal comparisons of mammal diversity provide a somewhat different view. On a per site basis, mammal diversity is comparable in the neotropics and African tropics. However, bats are better represented in the neotropics, whereas primates are considerably more species rich in the African tropics.

On a regional scale (i.e., within one of the three major rainforest regions), researchers have noted substantial variability in species richness. For instance, with respect to butterflies, a belt of high species richness apparently extends across the neotropics from southern Colombia and the border of Peru and Bolivia at the base of the Andes eastward into the Brazilian states of Rondonia and Acre. However, even within this belt of high butterfly diversity lie areas with apparently lower diversity. Within the Asia-Pacific tropics, lowest diversity of most taxa is found in the Australopapuan region, particularly in smaller and more isolated islands such as New Caledonia and those of the Bismarck Archipelago, Solomon Islands, Fiji, and Vanuatu. Many of these islands, particularly those in Fiji, New Caledonia, and Vanuatu, lack entire groups such as native rodents and amphibians that have been unable to disperse long distances over open ocean. Other taxa such as butterflies, lizards, and birds, although represented, are depauperate because of island isolation and small island area and a subsequent lack of opportunity for substantial allopatric speciation within an island. Only New Guinea, by far the largest island in the region, harbors taxa that have undergone substantial allopatric speciation.

Soils have a major effect on forest characteristics and productivity and consequently on animal diversity. Although soils of tropical rainforests are notoriously poor in nutrients, there is substantial variability. Where soils change abruptly, differences in forest structure and animal species richness are often evident. For example, the most species-rich sites in the world, those at the eastern base of the Andes, are situated on relatively richer soils. In contrast, the soils of the Guyana Shield in northern South America are particularly poor, and faunas are notably depauperate. Similarly, the sandy soils in parts of the Orinoco drainage of Venezuela are also very poor in nutrients, and animal diversity is often lower than that in surrounding areas with more fertile soils.

X. HYPOTheses of HIGH SPECIES RICHNESS

Explaining the high species richness of tropical rainforests has preoccupied ecologists and evolutionary biologists since naturalists first recognized the tropical–temperate latitudinal richness gradient and began to appreciate such richness. Many hypotheses have been advanced, but it is important to realize that none of these hypotheses has been adequately tested. It is also important to note that the great richness may be due to a combination of factors, and invoking a single-factor explanation ignores the complexities and intricacies of evolutionary and ecological processes and the long history of geological and climatological changes that certainly influenced diversity. Although the high species richness of animal consumers in tropical rainforests no doubt is linked with the high species richness of the plants that act as primary producers, invoking high plant diversity to explain high animal diversity is largely inadequate because it does not explain the high plant diversity. In other words, the root cause of higher tropical diversity is left unanswered by such circular reasoning. Researchers have sought explanations that would account for high diversity of all taxa. Several of the more important hypotheses are briefly reviewed here and may be divided into biotic and abiotic hypotheses. Biotic hypotheses emphasize interactions among species to account for high species richness in the tropics.
relative to temperate and polar regions. Abiotic theories of species richness rely on climatological and geological processes that may promote greater speciation in the tropics. Several hypotheses are specific to terrestrial ecosystems, but tropical–temperate richness gradients are also evident in marine organisms. Ecologists and evolutionary biologists must therefore grapple simultaneously with the phenomena of high species richness in both terrestrial and marine systems to develop explanations that will apply to both systems.

The first of the biotic theories is the spatial heterogeneity theory, which states that the tropics are more heterogeneous in space and structurally more complex, thereby providing more niches for animals to exploit and thus a greater number of species. This theory relies on the development of greater vegetational complexity in tropical forests but does not explain adequately how this vegetational complexity arose.

The competition theory states that because of a more benign climate in the tropics, organisms live closer to the carrying capacity. Because organisms are closer to the carrying capacity, interspecific competition for limiting resources is more intense, and such competition promotes morphological and ecological divergence and specialization on a narrower range of resources. Specialization then allows more species to coexist, thereby promoting greater diversity. This theory is based on the presumption of a more benign climate, but quantifying how benign a climate is from each organism’s perspective is not possible. It is also not possible to measure how each organism in the tropics lives closer to its respective carrying capacity.

The predation theory is contrary to the competition theory and states that predation and parasitism are more intense in the tropics, which keep organisms below their carrying capacities. Because organisms are kept below their carrying capacities, more resources are available to support a greater number of species. It has not been determined whether or not predation rates are indeed higher in the tropics (at least not for the vast majority of animal species), and again it is not possible to determine whether or not most animals are below their carrying capacities.

The animal pollinators theory states that there is less wind in the tropics, and therefore plants have evolved a greater reliance on animals to pollinate their flowers. This greater reliance promotes a closer relationship with animals and greater specialization to enhance the mutualistic relationship, which therefore promotes diversity. Establishing a definitive link between a greater reliance on animal pollinators to less wind in the tropics is not possible.

The first of the abiotic theories is the ecological time theory, which states that the tropics are older because they have not been subjected to the devastating effects of glaciation that occurred in north temperate regions and that periodically covered entire ecosystems under the great ice sheets. A corollary of this theory is that tropical forests underwent a series of contractions and expansions with cyclic changes in precipitation as water was tied up in the expanding ice sheets and subsequently was released with global warming and contraction of the ice sheets. As the tropical forests contracted, they became isolated as refugia within a matrix of drier savanna and scrub ecosystems. Isolation into refugia promoted allopatric speciation and greater species richness. This corollary was an attractive explanation for pockets of endemism found in organisms such as birds in Amazonia but has become controversial.

The area theory states that the tropics occupy a greater area and that more species occur there because of the species-area relationship. This theory also states that there are more geographical barriers, such as mountain ranges and large rivers, in the tropics, which isolate populations, thereby reducing gene flow and promoting genetic divergence and allopatric speciation.

The energy theory states that there is greater incident solar radiation striking the tropics and that this solar radiation is more equitably distributed throughout the year. Primary production is consequently greater, which provides more resources for consumers, and more species can consequently coexist. This theory does not address the origin of the greater diversity of species.

The evolutionary speed theory states that generation times are shorter in the tropics and more mutations occur. More mutations give natural selection a greater genetic base on which to operate, and speciation rates are consequently higher, which leads to an accumulation of species. This theory and most of the preceding abiotic theories do not explain how more species can coexist in the tropics and consequently do not explain the maintenance of high species richness in the tropics.

XI. THE FUTURE OF RAINFOREST ANIMAL DIVERSITY

Much has been written about the rapid rate of tropical rainforest clearance by humans and subsequent impacts on animal diversity. Indeed, virtually every volume that has been written in the past two decades on rainforests addresses the potential for a catastrophic decline in global biodiversity. Several biologists have argued that
the earth is on the verge of another mass extinction event of the magnitude of the Cretaceous event that eradicated the dinosaurs. Although such arguments are persuasive, it is not possible to predict accurately the magnitude of an impending extinction event, particularly because current knowledge of rainforest biodiversity is very limited and no precise estimate of numbers of species is available. However, with approximately 10% of the earth’s entire butterfly fauna and untold thousands of species of other animals found within a single 200-ha tract of tropical rainforest, it is not difficult to surmise that the clearing of large tracts of such rainforests indeed will lead to a major decline in global animal diversity.

See Also the Following Articles

AMAZON ECOSYSTEMS • BIODIVERSITY-RICH COUNTRIES • FOREST CANOPIES, ANIMAL DIVERSITY • RAINFOREST ECOSYSTEMS, PLANT DIVERSITY • RAINFOREST LOSS AND CHANGE • TROPICAL ECOSYSTEMS

Bibliography

RAINFOREST ECOSYSTEMS, PLANT DIVERSITY

I. M. Turner
Singapore Botanic Gardens

I. INTRODUCTION

The tropical rainforests grow in conditions of high rainfall (usually more than 2000 mm per year) with few, if any, dry months, where a month would be considered dry if less than 100 mm of rain fell. The temperatures in the lowlands generally average approximately 27°C, with little variation throughout the year. The day course of temperature with peaks in the afternoon and lows before dawn is frequently the major signal in long-term records—the so-called tropical diurnal climate. Tropical rainforest will grow under climatic conditions of relatively short dry seasons, or even quite long ones if groundwater is available to the trees. As the length and severity of the seasonal drought increases, the proportion of deciduous species in the canopy of the forest tends to increase. When more than about one-third of the canopy trees are drought-deciduous, the forest is no longer tropical rainforest by many definitions.

The plant species of the lowland tropical rainforest are generally intolerant of freezing and frequently sensitive to chilling temperatures. Many major tropical plant families have few members outside the tropics (e.g., the nutmegs, Myristicaceae), possibly because they have not developed cold tolerance.

The tropical rainforests constitute a belt of evergreen vegetation that until relatively recently almost provided a continuous link across the landmasses near the equator. Only in East Africa is rainforest naturally rare. The largest contiguous areas of rainforest were the basins of two major rivers, the Amazon in South America and the Congo in Africa. These essentially continental rainforests can be contrasted with the largely insular forests of the Asia-Pacific region.
II. OVERVIEW OF TROPICAL RAINFOREST BIODIVERSITY

Biodiversity can be assessed at many different levels ranging from the genetic diversity within populations to landscape or regional heterogeneity. The scale of assessment most widely studied and best understood (though still poorly known) is that of species diversity, which is the focus of this review. The meaning of the term species remains a contentious issue in biology, but in practice in the tropics species are defined by the judgment of taxonomists based largely on the gross morphology of specimens.

There can be little doubt that tropical rainforests are the most species rich of terrestrial ecosystems. Naturally they would cover about 6% of the earth’s land surface. Therefore, the high species diversity is not merely a reflection of the tropical rainforests covering an enormous area. The high diversity, relative inaccessibility, and remoteness from major centers of taxonomic research (mostly situated in Europe and North America) have compounded to slow progress in completing the basic inventory of the vascular plants of the humid tropics. Most tropical areas are studied by ongoing projects to provide floras, but progress on most of these is very slow. The Flora of Tropical West Africa is the only major tropical flora to have been completed. Others, such as Flora Neotropica and Flora Malesiana, have published accounts for relatively small proportions of the floras they cover despite decades of work. At this pace, it may take centuries to complete the flora production. However, the state of affairs may not be as bad as it seems because much information is available in monographs, local floras, and checklists. Nevertheless, some tropical areas, notably Central Amazonia, are still poorly known and estimates of plant diversity in terms of named species must be considered as provisional and tentative. Of extant living organisms, vascular plants are among the most completely described, but there remain many undescribed taxa. Exactly how many species remain unnamed cannot be estimated readily, but I suspect that the floristic inventory of the tropical forests has reached a stage at which the number of recognized species will change relatively little because descriptions of new species are heavily compensated for by synonymization among overdescribed taxa.

In Table I, I have attempted to estimate the total plant diversity (in terms of species richness) of the tropical rainforest regions of the world. It is difficult to assess the accuracy of this estimate, and until a species list is compiled it will remain uncertain. However, the

<table>
<thead>
<tr>
<th>Area</th>
<th>No. of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neotropics</td>
<td>93,300</td>
</tr>
<tr>
<td>African tropics</td>
<td>20,000</td>
</tr>
<tr>
<td>Mainland Africa</td>
<td>16,000</td>
</tr>
<tr>
<td>Madagascar</td>
<td>4,000</td>
</tr>
<tr>
<td>Asia-Pacific tropics</td>
<td>61,700</td>
</tr>
<tr>
<td>Malesia</td>
<td>63,000</td>
</tr>
<tr>
<td>Indo-China and adjacent areas</td>
<td>10,000</td>
</tr>
<tr>
<td>Southern India</td>
<td>4,000</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>1,000</td>
</tr>
<tr>
<td>Australia</td>
<td>700</td>
</tr>
<tr>
<td>Pacific Islands</td>
<td>1,000</td>
</tr>
<tr>
<td>Total</td>
<td>175,200</td>
</tr>
</tbody>
</table>

derived data provide some interesting results. First, the vascular plant species richness of the tropical rainforests amounts to more than half of the estimated global total of 240,000. This clearly indicates the important contribution of tropical rainforests to global biodiversity. Second, on the broad geographic scale the data allow comparison of the three main rainforest blocks: the neotropics, tropical Africa (including Madagascar), and the Far Eastern tropics from India into the Pacific. Africa evidently has far fewer species than the other two areas, and the neotropics are the most diverse. To a first approximation, a 3:2:1 ratio of species richness between tropical America, tropical Asia-Pacific, and tropical Africa can be considered, with an approximately equal proportion of diversity between the Old and New Worlds. At higher taxonomic levels the Neotropical dominance tends to disappear. For instance, it has been estimated that 292 higher plant families are represented in the neotropics, but Malesia, with fewer than half the number of species, has 310 families in its flora (Prance, 1994). Southern Indo-China is the region of the world most diverse in angiosperm families (Williams et al., 1994).

III. THE RAINFOREST COMMUNITY

Any visitor to a tropical rainforest can readily appreciate the botanical diversity they contain because of the many forms of plants to be observed ranging from tiny herbs to gigantic trees. There have been numerous attempts to classify plant species by their life-forms. Although there is no universally accepted or universally applicable system, most of these classifications use a similar set of criteria for determining groups within the structure. This set, at its simplest, generally includes plant mature stature and degree of woodiness, need for mechanical support from other plants, and information on trophic requirements. A classification of plant life-forms found in the tropical rainforest, based on these criteria, is given in Table II.

An area of lowland forest in Brazil would have relatively few plant species in common with a site in Costa Rica and very few indeed in common with sites in Congo or Indonesia, but all would be classified as lowland tropical rainforest. Despite this strong geographic dissimilarity in species composition, there are floristic affinities throughout the world. These occur at higher taxonomic levels, particularly at the rank of family, but also include some strong generic similarities in makeup. Tropical florists tend to be more similar to each other, in terms of species abundance in common families, than to subtropical and temperate florals (Turner, 1997). Table III summarizes floristic data by life-form and geographic area. Gentry (1988, 1993), in particular, emphasized the consistent patterns of family abundances, in terms of numbers of species represented in inventories of forest areas, across the tropics. Among woody stems, families such as Leguminosae, Rubiaceae, Euphorbiaceae, and Myristicaceae are nearly always among the most speciose at a lowland tropical rainforest site. Pteridophyta, Arecaceae, and Zingiberaceae are frequently the most common ground herbs, and Orchidaceae and Pteridophytes almost invariably dominate the epiphytes. Table III lists 17 very large pantropical genera that are well represented in all regions, again emphasizing the floristic similarity of diverse tropical areas. However, there are also differences between regions. The "dipterocarp substitution" has been noted as a major anomaly in the Asia-Pacific region (Gentry, 1988). That is, in this region the Dipterocarpaceae substitute for Leguminosae as the dominant woody family. The Dipterocarpaceae are a pantropical family, but the vast majority of its species occur in the region from India east to Wallace's line. In this area, dipterocarps form a majority of the large trees in lowland forests. However, this does not translate into dominance in terms of species abundance in the complete flora, or even the tree flora, unless a very large stem diameter limit is chosen. Legumes are not as numerically important as in American or African florals, but nevertheless they are still well represented, so the dipterocarp substitution is not as important a phenomenon as might be thought. Additionally, East Malesian rainforests have only a small representation of Dipterocarpaceae. Table III indicates other regional variations in representation of more important families. Perhaps the most notable of which are the virtual restriction of the important ground herb and epiphyte family Bromeliaceae to the neotropics; the importance of Bignoniaceae and Sapindaceae as climbers in the same region; the paucity of palms, particularly understory species, in Mainland Africa (although not Madagascar); and the abundance of Dicotyledonae. There are many other smaller families with restricted ranges in the tropics. Because of the range of spatial scales that can be considered and the relative inclusivity of the sampling, many approaches are available for assessing plant species richness of tropical forests. These include plot, transect, and plotless sampling at small spatial scales.

1 Pteridophyta are frequently accorded "family status" in these analyses.
### TABLE II
Classification of Vascular Plant Life-Forms in the Tropical Rainforest and a Summary of Their Taxonomic Composition

<table>
<thead>
<tr>
<th>Life-form group</th>
<th>Definition</th>
<th>Examples from tropical rainforest</th>
<th>Subdivisions of the group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees</td>
<td>Autotrophic, woody, and mechanically independent (self-supporting) plants</td>
<td>Many taxa within the dicotyledons, palms among the monocots, a few conifers and gymnosperms, some cycads and tree ferns.</td>
<td>Mature size and form have often been used to subdivide the tree category. Unfortunately, no consistent definition has been employed. The giant herbs, largely from the Zingiberales, can be distinguished by size, and in ecological terms may have more in common with trees than the other herbs.</td>
</tr>
<tr>
<td>Herbs</td>
<td>Autotrophic, mechanically independent plants with little or no lignification and secondary thickening of stems or roots</td>
<td>Many dicot and monocot taxa, some characteristically epiphytic species are facultatively terrestrial when they fall from the trees.</td>
<td></td>
</tr>
<tr>
<td>Climbers</td>
<td>Autotrophic, woody, or herbaceous plants, rooted in the soil and dependent on other plants for support, at least in later stages</td>
<td>A wide range of taxa including aloids, monocots (e.g. palms and aroids), gymnosperms (Gnetum spp.), and ferns.</td>
<td>Woody climbers (lianas) are sometimes distinguished from herbaceous ones, means of climbing can also be employed</td>
</tr>
<tr>
<td>Epiphytes</td>
<td>Autotrophic, woody, or herbaceous species habitually growing on other plants and not rooting in the soil</td>
<td>Epiphytes are dominated by three groups—orchids, pteridophytes, and aroids; bromeliads are an almost exclusively Neotropical epiphyte group.</td>
<td>Division into herbaceous and woody forms might be possible. Often, a distinction is made between species on their relative preference for crown and hide locations on their host trees.</td>
</tr>
<tr>
<td>Hemiepiphytes</td>
<td>Plant, mostly woody, that begin life as epiphytes but may eventually grow roots down to the ground</td>
<td>Strangling figs (Ficus spp.) are the largest group of hemiepiphytes.</td>
<td></td>
</tr>
<tr>
<td>Plant parasites</td>
<td>Plants that directly parasitise other plants</td>
<td>Direct plant parasitism is confined to the dioecious, and with the exception of Casytha (Lanaceae) is restricted to probably advanced groups. Muscioides (Loranthaceae and Viscaceae) and related rose parasitic groups (Santalaceae and Olaaceae) and the mostly herbaceous Scrophulariaceae form the two largest groups.</td>
<td>The parasites can be divided into hemiparasites that are chlorophyllous but possess haustorial connection to host xylem (e.g., muscioids), holoparasites that are achlorophyllous and entirely heterotrophic (e.g., dodders, broomrapes, and Balanophoraceae), and endoparasites that live inside the host and are only visible externally as reproductive structures (Rafflesiaceae).</td>
</tr>
<tr>
<td>Mycotrophs</td>
<td>Heterotrophic plants that derive carbohydrates from a fungus, the fungal mycorrhizas may derive its carbon from dead material or from another living plant</td>
<td>All orchids are mycotrophic as seedlings, but most become autotrophic as they develop. The small proportion that remain entirely heterotrophic represent the largest group of mycotrophic species. Other monocot mycotrophs include members of the families Primulaceae, Triuridaceae, Iridaceae, Burmanniaceae, and Convolvulaceae. Ericaceae, Gentianaceae, and Polygalaceae are the only dicot families with mycotrophic species.</td>
<td>Detailed investigations can distinguish the &quot;parasitic&quot; from the &quot;saprophytic&quot; species but these have been conducted for relatively few of the tropical species.</td>
</tr>
</tbody>
</table>

...that may include all plant taxa present, larger plot sampling that considers only trees and climbers exceeding a certain lower stem diameter limit, and local or regional floristic accounts based on intensive collection within the area concerned. Relatively few efforts have been made to identify all the individual plants present within an area of rainforest largely because few people possess the botanical expertise to identify the enormous amount...
### Table III

Summary of the Floristic Composition of Tropical Rainforest, with a Breakdown by the Major Geographic Regions

<table>
<thead>
<tr>
<th>Regional Breakdown</th>
<th>All tropics</th>
<th>Neotropics</th>
<th>Africa (including Madagascar)</th>
<th>Asia-Pacific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Important tree families</td>
<td>Leguminosae</td>
<td>Vochysiaceae</td>
<td>Dichapetalaceae</td>
<td>Dipterocarpaceae</td>
</tr>
<tr>
<td>Lauraceae</td>
<td>Bignoniaceae</td>
<td>Myrtaceae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annonaceae</td>
<td>Myrtaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rubiaceae</td>
<td>Bignoniaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monocotyledons</td>
<td>Vochysiaceae</td>
<td>Lecythidaceae</td>
<td>Myrtaceae</td>
<td></td>
</tr>
<tr>
<td>Myristicaceae</td>
<td>Eucalyptaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sapotaceae</td>
<td>Fraxinaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meliaceae</td>
<td>Bombacaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palmae</td>
<td>Euphorbiaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Euphorbiaceae</td>
<td>Pseudophyta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Araceae</td>
<td>Bromeliaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zingiberaceae</td>
<td>Heliconiaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyperaceae</td>
<td>Gramineae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urticaceae</td>
<td>Zingiberaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orchidaceae</td>
<td>Gesneriaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Important herb families</td>
<td>Asclepiadaceae</td>
<td>Compositae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acanthaceae</td>
<td>Annonaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convulvulaceae</td>
<td>Dilleniaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leguminosae</td>
<td>Sapindaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Araceae</td>
<td>Nepenthaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apocynaceae</td>
<td>Euphorbiaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cucurbitaceae</td>
<td>Euphorbiaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Important climber families</td>
<td>Asclepiadaceae</td>
<td>Compositae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acanthaceae</td>
<td>Annonaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convulvulaceae</td>
<td>Dilleniaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leguminosae</td>
<td>Sapindaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Araceae</td>
<td>Nepenthaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apocynaceae</td>
<td>Euphorbiaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cucurbitaceae</td>
<td>Euphorbiaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Important epiphyte families</td>
<td>Orchidaceae</td>
<td>Bromeliaceae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudophyta</td>
<td>Cyrtandra</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Araceae</td>
<td>Castaneae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Piperaceae</td>
<td>Asclepiadaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melastomataceae</td>
<td>Rubiaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gnetaceae</td>
<td>Gnetaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Important genera (&gt;300 spp)</td>
<td>Aplectania</td>
<td>Anthurium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Begonia</td>
<td>Asteranthus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buxaceae</td>
<td>Erythroxylum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clusia</td>
<td>Eufrae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton</td>
<td>Euphorbia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cynanchum</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dioscorea</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ficus</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hakea</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impatiens</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Justicia</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phyllanthus</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Piper</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychotria</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schleichera</td>
<td>Eucalyptus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selaginella</td>
<td>Selaginella</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solanum</td>
<td>Selaginella</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Names in bold indicate families entirely restricted to a particular geographic region, or nearly so. Names underlined indicate families notable for their scarcity or absence from a particular region. Names in bold and underlined are entirely absent from a particular region, or almost so.*
of juvenile and sterile plant material that such total inventories generate. Recently, an attempt was made to enumerate all the plant species occurring in 1 ha of forest in lowland Amazonian Ecuador. A total of 942 species were encountered in the 100 × 100-m plot (Balslev et al., 1998). Lower montane forest at 1500 m in Java was estimated to contain 333 species in 1 ha (Meijer, 1959). A single plot of 10 × 10 m in Costa Rica contained 233 species (Whitmore et al., 1985), more than any among eight plots of the same size recorded in Gabon (74–130 spp.; Reitsma, 1988). Dui- venvoorden (1994) enumerated all the vascular plant species in a series of ten 0.1-ha plots in Colombia. The most species-rich plot contained 313 species. Gentry and Dodson (1987a) recorded 365 species from 0.1 ha at Rio Palenque, Ecuador, but this was from 10 noncontiguous subplots. Two other sites in Ecuador, sampled with the same methods, were considerably less diverse, with 169 and 173 species recorded (Gentry and Dodson 1987b).

Most estimates of plant diversity in tropical forests involve the enumeration of stems above a certain diameter (or girth) at breast height (1.3 m) or above obvious outgrowths, such as buttresses or stilt roots. Many different sampling methods and size limits have been employed in such investigations, but fortunately a few approaches have been applied relatively consistently and allow intersite comparisons of diversity. Three commonly employed methods of assessing plant diversity in tropical forests produce scales of species-richness values that are comparable, at least in the broad view (Table IV). I used these data to define categories of

<table>
<thead>
<tr>
<th>Species-richness class</th>
<th>No. of species</th>
<th>Examples from different geographic regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Less than 100</td>
<td>Barro Colorado Island, Panama; San Carlos, Venezuela; Belem, Brazil</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kibale, Uganda; Kafue, Ghana; Lope, Gabon</td>
</tr>
<tr>
<td>Intermediate</td>
<td>100–199</td>
<td>Many, including La Selva, Costa Rica; Makeni, Gabon; Korup, Cameroon; Owie, Gabon</td>
</tr>
<tr>
<td>High</td>
<td>200–249</td>
<td>Coca, Ecuador; Jatun Sacha, Ecuador; Mohana, Peru</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nosy Mangabe, Madagascar; Ilombe, Malagasy; Lamtibe, Malaysia; CMBRS, Papua</td>
</tr>
<tr>
<td>Very high</td>
<td>250 or more</td>
<td>Cotubanamo, Ecuador; Yanomama, Peru; Raya Cahuna, Colombia; Trombendi, Colombia; Manaus, Brazil</td>
</tr>
</tbody>
</table>

a Species richness can be measured in one of three ways: (x) species represented among the trees and lianas of 10 cm dbh (or 30 cm gbh) or greater on a 100 × 100-m (1 ha) plot; (y) species represented among the first 50 tree and liana individuals of 10 cm dbh (or 30 cm gbh) or greater on a contiguous plot; and, (z) species represented among the trees and lianas of 2.5 cm dbh or greater on a 0.1-ha sample consisting of ten 50 × 2-m transects.

1 Valencia et al. (1994) (x).
3 Ferreira and Rankin-de-Merona 1998 (x).
4 Phillips et al. (1994) (y).
5 Condit et al. (1996) (x).
6 Valencia et al. (1998) (x).
7 Gentry (1993) (z).
10 Black et al. (1990) (z).
11 Cao et al. (1990) (x).
12 Newbery et al. (1992) (x).
13 Wright et al. (1997) (x).
14 Proctor et al. (1993) (x).
relative diversity for lowland tropical forests. Forests of very high diversity have thus far only been reported from the New World. They are known from the lowland forests of the Pacific coast of Colombia in Chocó Province and the Upper Amazon Basin stretching in a belt below the eastern slopes of the Andes from Ecuador to Peru. The occurrence of 230 species in a 1-ha plot near Manaus in the center of the Brazilian Amazon (Ferreira and Rankin-de-Mérona, 1998) countered the impression that the forests of the middle and lower Amazon were of uniformly lower diversity than the uppermost reaches. High-diversity forests are well represented in the South American tropics and are of similar species richness as the lowland dipterocarp forests of West Malesia and at least some of the lowland forests of New Guinea. To date, only Madagascan rainforests have been shown to attain high-diversity status among those from Africa. The most diverse mainland African rainforests occur in Cameroon and Gabon in West Africa, but these are substantially less rich than many forest sites in the neotropics or the Asia-Pacific region and on a global scale could only be classed as of intermediate diversity. Many sites where detailed ecological research has been conducted, such as Barro Colorado Island, Panama, San Carlos de Río Negro, Venezuela, and Kibale, Uganda, are low-diversity rainforests. These geographic patterns in local species richness appear to mirror quite closely the tropical regions of the map of plant species diversity per 10,000 km² (Barbault et al., 1996; see http://www.botanik.uni-bonn.de/biodiv/phytodiv.htm).

An important point is that the range of values for diversity within a locale, i.e., an area of approximately identical climate, can cover much of the general range exhibited across wide geographic and climatic gradients. For example, at Caquetá in Colombia total vascular plant diversity in ten 0.1-ha plots was 40–313 species (Duivenvoorden, 1994), at Tambopata in Peru five 1-ha plots had 60–173 species for stems of 10 cm dbh or larger (Phillips et al., 1994), and at Gunung Mulu National Park in Sarawak four plots varied between 73 and 223 species for the same plot and tree sizes as those of the previous example (Proctor et al., 1983). In all these cases the lowest diversity plots were ones on predictably harsher substrates, such as permanently waterlogged soils (Caquetá and Tambopata) or limestone karst (Gunung Mulu). Soil physicochemical properties are highly influential in the composition of the forest community and clearly affect species richness. However, there are also other environmental factors that can lead to gradients in plant diversity (Table V). Rainfall appears to be the major determinant of diversity when old-growth sites on freely draining, nonextreme soils are compared. A relatively strong positive correlation between average annual precipitation and diversity has been observed in many tropical data sets (Givnish, 1999), although good data for total plant diversity is lacking so that only estimates for large woody stems can be used. This correlation is evident over the range of true lowland tropical rainforest and it is not merely a comparison of diversity between drier forest formations and proper rainforest. Rainfall seasonality (the presence and severity of a dry season) is of key importance, but it is difficult to quantify this from standard meteorological data, and total annual rainfall is a good predictor of diversity in most cases. It is not immediately evident why further increasing the wetness of a wet climate should continue to increase the local species diversity, although highly ombrogenous sites may favor epiphytes. As can be seen from the data presented in Table VI, the proportion of the flora that epiphytes represent increases quite sharply with average annual rainfall.

### IV. Why are Tropical Rainforests so Diverse?

Lowland tropical rainforests undoubtedly represent the zenith of terrestrial biodiversity, probably with more than half the world’s biota on less than a twentieth of the land surface. High plant diversity may well be a major contributory factor to the overall abundance of species in the rainforest through the specialization of a myriad of invertebrate species to each plant. However, why are there so many plant species in the tropics? In evolutionary terms, high species diversity must imply many speciation events and relatively few extinctions of species because diversity is the balance between these two processes. Therefore, the tropics may be richer in species than other areas of the world because there has been more rapid speciation here and/or a slower rate of extinction. It is possible that the benign and equable climate of the tropics makes it less likely that a species...
TABLE VI
Representation of Major Life-Form Groups in the Floras of Different Tropical Rainforest Sites

<table>
<thead>
<tr>
<th>Site</th>
<th>Mean annual rainfall (mm)</th>
<th>Total No. spp.</th>
<th>Percentage of species in major life-form groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Trees</td>
</tr>
<tr>
<td>La Selva, Costa Rica</td>
<td>3950</td>
<td>1450</td>
<td>39</td>
</tr>
<tr>
<td>Rio Palenque, Ecuador</td>
<td>3000</td>
<td>1055</td>
<td>25</td>
</tr>
<tr>
<td>Barro Colorado Island, Panama</td>
<td>2650</td>
<td>966</td>
<td>43</td>
</tr>
<tr>
<td>Manus floodplain, Peru</td>
<td>2130</td>
<td>1217</td>
<td>57</td>
</tr>
<tr>
<td>Singapore (forest flora)</td>
<td>2000</td>
<td>1673</td>
<td>36</td>
</tr>
<tr>
<td>Jammeh, Ecuador</td>
<td>1830</td>
<td>608</td>
<td>28</td>
</tr>
</tbody>
</table>

a Sites are listed in descending order of mean total annual rainfall.
c Gentry and Dodson (1987b).
d Foster and Hubbell (1990).
e Foster (1990).
f Turner (1994).

will become extinct than in the harsher climates elsewhere. The positive correlation between forest diversity and annual precipitation within the tropics also indicates a link between environmental adversity and reduced diversity.

Switching from an evolutionary to an ecological perspective, tropical rainforest diversity maintains a perplexing fascination. A commonly held view of natural communities is that they consist of populations of species in which intraspecific competition is stronger than interspecific competition. In other words, species occupy separate niches within the community and have reached a competitive equilibrium. An equilibrium model of the structure of a lowland tropical rainforest plant community requires that all the species present, including the many species of similar life-form and basic ecology such as canopy trees, or understory treelets, occupy separate niches. Given the simple requirements of plants—light, water, and nutrients from the soil—it is not immediately obvious where so many niches could come from to support the observed diversity. Differences in requirements for reproduction and regeneration may allow more axes from which to partition niches. These niches could be extremely narrow because of the constancy of the tropical environment allowing persistence of small or sparse populations.

I suspect, however, that few tropical forest ecologists believe that all the plant species in a forest are ecologically distinct. A more plausible hypothesis is that most species have ecologically equivalent or similar species with which they cooccur. It is worth noting that a high proportion of the species present in a rainforest exist as populations intermingled with those of congeners. For instance, of the 814 tree species recorded from 50 ha at Paoib, Malaysia, 82% had a congeneric present in the plot and 70% had a congeneric present in the same broad stature class (Manokaran et al., 1992). For descriptive convenience, ecologists may group species into guilds of ecologically similar entities such as pioneer trees or hole climbers, but it must be remembered that these are not necessarily discrete groupings in terms of niche space in the community. They may represent arbitrary divisions of whatever axes of differentiation are being considered.

If competitive equilibrium has not been obtained then something must be preventing competitive exclusion within the community from occurring. One possible process by which this could happen is through frequency-dependent mortality operating within the community setting an upper limit on the abundance of a species and thus making it unlikely that it can totally exclude other species from the landscape. This compensatory mortality, particularly among tropical trees, is often referred to as the Janzen–Connell effect after the two researchers who independently put forward the hypothesis that species-specific predators could produce the density-dependent mortality required to promote the coexistence of many plant species in a forest. Invertebrate seed and seedling predators, and plant pathogens such as fungi, have frequently been observed to cause severe mortality of germinants and juveniles occurring in high densities near parent trees. Detailed analyses of spatial patterns of sapling mortality on Barro Colorado Island, Panama, showed evidence of compen-
satory mortality in recruitment for 67 of 84 of the most common tree species (Wills et al., 1997). Rare species, which make up the bulk of the diversity in the forest, may not reach sufficient adult densities to suffer compensatory mortality. However, they may just be too rare to provide statistically significant results with the available data sets. Givnish (1999) hypothesized that compensatory mortality may increase in intensity with rainfall because pest and pathogen populations are less likely to be checked by a dry season in an ever-wet forest. This may explain, at least partially, the observed increase in forest diversity with precipitation.

An alternative nonequilibrium view of the community is that chance is the major determinant of community composition and relative abundance. The majority of species are so infrequent that they very rarely come into direct competition with other ecologically equivalent, similarly sparse species. Chance, and not competitive ability, is an important determinant of which species establishes an adult successfully at a given site because most species are unable to produce enough seeds, disperse them well enough, or survive the early onslaughts of pests and pathogens to ensure that they have a candidate for any vacancy in the community (Hubbell et al., 1999). Many potential episodes of interspecific competition may be won by default, and by competitively inferior species, because no adversaries were present at the right time and place.

At present we have no simple answer to the question of why there are so many species in the rainforests and how they manage to coexist. Nor does it seem likely of why there are so many species in the rainforests and how they manage to coexist. Nor does it seem likely that there is a simple answer. Probably all the factors mentioned previously, and possibly others not yet considered, are operating within any rainforest but probably not affecting all members of the community simultaneously.

V. THE VALUE OF TROPICAL RAINFOREST PLANT DIVERSITY

If tropical rainforests are to be conserved, conservationists need to convince a wide range of people that tropical rainforests have sufficient value to merit conservation. In attempting to do so, conservationists face a series of dilemmas. The first is whether to consider value as an economic parameter or not. The global economy hardly has a good record for preventing extinction or supporting sustainable exploitation, but the reality is that people in the tropics exploit the rainforest so that they can improve their standard of living, even if only in the short term. Conservationists can promote the development of a truly sustainable global economy, but in the meantime they must address the immediate problem of impending extinctions. Purely economic analyses generally favor rapid liquidation of forest resources, although a notable exception was a comparison of exploitation of minor forest products, such as fruits and latexes, with logging and timber tree plantations in Peru (Peters et al., 1989) which showed that the relatively sustainable nontimber production generated greater revenues and thus put the highest net present value on the resource. However, the study has been criticized as being atypical in having a forest with a very high density of fruit and latex trees relatively close to a major population center providing a ready market for the products.

Are there any economic benefits to be gained from leaving rainforest undisturbed? A group of people for which the answer to this question is generally in the affirmative are traditional forest dwellers. Few of the tropical rainforest regions are without people who live in the forest, either making their living by hunting and gathering in the forest or by farming on a small scale, usually as some form of traditional shifting agriculture. Large-scale forest clearance dispossesses the traditional forest dwellers of their entire resource base and often much of their social fabric.

Tourism potential is perhaps the most economically lucrative opportunity offered by virgin tropical rainforest, but fortunately we are not yet in a situation in which every remaining area of forest can attract enough paying visitors to make conservation profitable. However, at a local level, ecotourism probably offers the best hope of generating an income from unexploited forest. Plant biodiversity, in addition to providing the glorious backdrop, is rarely likely to be the focus of ecotourism, although Rafflesia may be a notable exception.

The value of the tropical rainforest as a storehouse of potentially useful genetic diversity is probably the best claim for a high intrinsic value to be placed on biodiversity. For vascular plants the potential uses are largely of three main types: medicinal, novel crops, and novel genes for improvement of existing crop varieties. A new drug that is superior to other available medicines can be worth millions of dollars, not to mention that it can have tremendous humanitarian value in the reduction in suffering. New or improved crops also have substantial economic value. Given the massive concentration of biodiversity in the tropics, it is almost certain that superior and economically rewarding new products await discovery in the rainforest. However, searching the haystack for the few golden needles is a difficult,
expensive, and often fruitless procedure. Island biogeography theory predicts that a 90% loss of habitat equates to a 10% reduction in diversity. Therefore, the potential economic returns from undiscovered rainforest products seem unlikely to provide strictly economic arguments for preventing the destruction of tropical forests until the area left is very small. Forests are important for their so-called “ecosystem services”—the amelioration of atmospheric composition, water quality, and local climatic regime among a variety of beneficial emergent properties of the rainforest community. Despite considerable speculation and some very simplistic experimentation, there is no solid evidence that species richness per se has a major influence on these community-level processes. Most species in the rainforest are very rare, so it seems unlikely that the random removal of a certain proportion of the flora would consistently influence oxygen production, carbon sequestration, nutrient cycling, and transpiration.

In the prevailing international financial framework it is difficult to find a sound, strictly economic, argument for conserving tropical rainforest. When environmental and ecological sustainability become the key goals of the global economy, conservation will make economic sense. The biodiversity of the tropics has a unique and irreplaceable intrinsic value as well as cultural, spiritual, educational, and humanitarian worth to millions of people, many of whom have never been in a tropical rainforest.

VI. THE FATE OF THE TROPICAL RAINFOREST

The human population of the tropics has already destroyed the rainforest occupying half the area it would be expected to cover in today’s climatic regime. Much of the remaining forest is fragmented into small remnant patches, damaged, burnt, or defaunated. Extensive, relatively pristine forests are largely confined to Amazonia, the Congo Basin, and New Guinea. Most tropical rainforest nations have little remaining forest, and although they generally have designated national parks, forest reserves, or other protected area status, the actual protection that these sites receive is often inadequate to prevent deterioration. There are of course exceptions, and excellent tropical rainforest nature reserves do exist.

There is no sign that the global rates of deforestation in the tropics are declining. Tropical forests are cleared for a variety of purposes and the relative importance of these varies between regions (Bawa and Dayanandan, 1997). In Africa and parts of South America, the demand for land by a large population of poor people who need to subsist on what they can grow leads to deforestation. Clearance for cattle ranching is also common in Latin America. Establishment of plantations for commercial tree crops, such as oil palm (Elaeis guineensis) and rubber (Hevea brasiliensis), is a more destructive force in the Asian tropics. Human populations continue to grow rapidly throughout the tropics, making it unlikely that the pressure for land for food production or cash crops will decline.

There is increasing evidence that the tropical rainforest community, particularly its botanical component, is more resistant to human disturbance and fragmentation than has been feared. Studies have shown that selective logging does not reduce tree diversity very much (Cannon et al., 1998) and long-isolated forest fragments maintain relatively high proportions of their original floras (Turner and Corlett, 1996). This information is met with some ambivalence by conservationists, who have tended to stress the fragility of tropical forest ecosystems. It is positive and hopeful in that it provides some scientific support to the aspirations for sustainable exploitation of tropical forests and the persistence of the native biota in nations in which only fragments remain. However, there is a real concern that knowledge of this relative robustness will be used as an excuse by those with political power to allow greater levels of exploitation and deforestation, possibly including inside currently protected areas.

From a pragmatic viewpoint, it is obvious that a multiplicity of approaches will be required to conserve the tropical rainforest biota. Where there are still large areas of relatively undisturbed forest with low human population densities, efforts to establish large in situ parks are warranted, but in areas with a landscape dominated by human activity the remnant patches of the original forest have an important conservation value in their own context. These fragments may not contain all the original species, but they still support a rich diversity that otherwise might be completely lost. In highly degraded tropical landscapes, which are very common, forest fragments may also play a vital role in the reforestation of these derelict lands by acting as foci for tree establishment and sources of planting material.
RAINFOREST ECOSYSTEMS, PLANT DIVERSITY

ANIMAL DIVERSITY • RAINFOREST LOSS AND CHANGE • TROPICAL ECOLOGIES

Bibliography


RAINFOREST LOSS 
AND CHANGE

K. D. Singh 
Harvard University

I. Assessment Techniques
II. The Historic Rainforest Area
III. The Current Rainforest Area and the Rate of Loss
IV. Types of Changes within the Forest Land
V. Implications for Biological Diversity

GLOSSARY

change matrices A way to relate area of land use classes at the beginning with that at the end of a period. From such matrices, it is possible to follow the exact path of change viz. how much land was transferred from a given use to other uses, how much came into that use, and what remained unchanged during the period.
deforestation The change from forest to other land uses such as agriculture or ranching; or depletion of forest crown cover to less than 10% density. Because shifting cultivation involves a change of land use, it is considered deforestation.
forest change processes A collective term for all activities on the forest land which affect the stand or site and, in particular, the biological diversity. These processes include forest exploitation, forest fragmentation, and establishment of new plantations.
historic rainforest All land areas with a potential to support the tropical rainforest as determined by the climatic and physiographic conditions whether currently forested or not. The concept is interesting because it provides a baseline to calculate the rainforest loss in a country or region over a historic period.
rainforest In this context, the "tropical rainforest formation" which occurs approximately within a latitudinal belt of 10° on either side of the equator. It may be noted that temperate and sub-tropical zones have forest formations structurally similar, but biologically not as diverse as the tropical rainforest.

THERE HAS BEEN GROWING WORLDWIDE CONCERN ABOUT THE FATE OF THE TROPICAL FORESTS and, in particular, the rainforest. Incomplete and often contradictory information has added to the uneasiness of the international community. Widely differing estimates of loss have appeared in the media and the issue has also taken a political overtone. This article presents the reasons for differing estimates and discusses how recent international initiatives have improved the information base. It then presents the most reliable data on the historic and current area of the tropical rainforest, the current rate of loss, and the process of ongoing changes. Finally, some implications for biological diversity are presented at the conclusion.

I. ASSESSMENT TECHNIQUES

A detailed review of the problems associated with assessment of deforestation has been published by the
Food and Agriculture Organization of the United Nations (FAO) in Forestry Papers 112 and 130 released 1993 and 1996, respectively. According to these reports, the assessment of global deforestation or that within a country is complicated due to several reasons:

- There is no globally accepted definition of forest or deforestation. Some include only primary forest in the term “forest,” whereas others include all forests whether primary or disturbed, closed or open.
- Assessing deforestation requires a minimum set of two consistent observations over time. In many countries or regions, there are no observations due to a weak capacity for conducting forest inventories.
- Rainforests are located in inaccessible terrain costly to survey. Perpetual clouds in some equatorial zones prevent acquisition of cloud-free images or the taking of aerial photographs. Radar images, due to their cloud-penetrating capabilities, hold promise but have only limited information to offer.
- Deforestation has very uneven and patchy spatial distribution. Special statistical techniques are required for a cost-effective assessment.
- Sometimes, inappropriate technology has been employed and data have been reported for a country without the reporting of the associated error. Fernside (1993) presents an account of varying estimates of deforestation reported for the Brazilian Amazonia.
- Even in countries with a tradition of forest inventory, the techniques used have not been very appropriate for monitoring changes and do not enable a statistically sound comparison of estimates on two dates.

FAO, as a part of Forest Resources Assessment (FRA) 1990 project, organized a comprehensive study on tropical deforestation as a collaborative effort of a team of more than 100 national and international experts. Most of the statistics presented in this article originate from this project. A study by FAO for the Year 2000 is currently in progress and reports are expected to be published by the end of 2000 or beginning of 2001.

The FAO FRA 1990 project made use of the following data sources and procedures to estimate tropical deforestation:

1. The existing reliable information on tropical forests since 1960 was compiled at a subnational level in the form of a database called FORIS (Forest Resources Information System). All repeated observations for the same area were recorded in the form of a time series. The database includes information relating to forest area, volume/biomass, forest management status, and forest utilization.

2. Maps were made showing the current and historic distribution of tropical forests (the latter map is also termed the ecological zone map) using forest-type maps of countries, available remote sensing data, and ecological information related to forest-type distribution such as climate, physiography, and soil. This major work was accomplished in cooperation with the Laboratoire d'écologie Terrestre in Toulouse, France. Subnational boundaries for all tropical countries were georeferenced and corresponding time series of demographic data since 1960 compiled in the form of a database.

3. A model was developed relating the time series data on forest cover with the time series data on demography by subnational-level unit stratified by ecological zone.

4. A pantropical sample survey was conducted using pairs of high-resolution satellite images, one taken in approximately 1980 and the other in approximately 1990, with a goal to produce change matrices at the sample, regional, and global tropical levels. The survey divided the tropical zone into 10 geographic and 3 forest cover strata using the forest cover and ecological zone maps. All 117 satellite images, each showing approximately 3 million ha, were randomly selected and distributed to various strata. The study, covering nearly 10% of the area of the tropics, was conducted under a very controlled environment. The images of each sample site were of comparable quality and taken during the same season and interpreted by the same person with a centralized quality control and training. The (interdependent) image-to-image interpretation procedure, specifically developed to capture deforestation, enabled assessment of changes with a much lower standard error compared to that of the existing method of independent assessments at two dates.

II. THE HISTORIC RAINFOREST AREA

Rainforest is the natural vegetation of the equatorial zone. It is associated with a very uniform temperature throughout the year (26–28°C) and wet conditions with precipitation exceeding evaporation at least for 10 months. The historic distribution of rainforests, as described in Singh (1974), is as follows:

South and Central America: The Amazon basin extending west to the lower slopes of the Andes and east to the coast of Guyanas, south to the Gran
Chaco, and north along the eastern side of Central America.

Africa: The Congo basin (mainly Zaire and Congo), Gabon, and Cameroon. It extends westward into Nigeria, Ghana, Ivory Coast, Liberia, and Guinea. On the eastern side it extends to Uganda.

Asia: mainly Southeast Asia, including the Malay Peninsula, the islands of Sumatra, Borneo, Sulavesi, and New Guinea, part of India, Sri Lanka, Burma, Thailand, and the Philippines.

Although climate in the rainforest regions is relatively uniform, there are local variations caused by changes in soil, topography, and geology. These site variations give rise to a corresponding change in vegetation forms. Some typical forest formations in the equatorial zone are the following: The rainforest proper (also called climax rainforest) usually occurs on well-drained soils and below elevations of 500–800 m. This is generally what the common person means when he or she talks of the rainforest. Other types are mangrove forest, swamp forest, periodically flooded forest, riparian forest, hill forest (usually at altitudes higher than 800–1000 m), forest on tropical podsols (a soil type), and forest on coastal sand.

Ecological studies show evidence of striking similarities among the rainforest formations across continents even though they occur in widely separated areas. According to Richards (1952), they are composed of plant species with comparable characteristics and great similarity in their spatial arrangement, life form, and physiognomy. The fundamental pattern of structure is thus the same through the whole extent of the rainforest.

According to the FAO ecological zone map, the potential area of the tropical lowland rainforest worldwide was 935 million ha, with 127 million ha (13.5%) in Africa, 308 million ha (33.0%) in Asia, and 500 million ha (53.5%) in South America. South China and Australia also have small areas of tropical rainforests that were not included in the previous total.

III. THE CURRENT RAINFOREST AREA AND THE RATE OF LOSS

Based on FAO FRA 1990, Table I gives the regional distribution of the lowland rainforest in 1990 and the rate of loss during 1980–1990. An approximate estimation of the historic rainforest area, based on the FAO ecological zone map, is also given to make comparisons with the current area. Accordingly, the worldwide area was 715 million ha, with 87 million ha in Africa, 178 million ha in Asia, and 451 million ha in America. A comparison with the historic areas shows a global decline of 220 million ha, with Africa contributing nearly 40 million ha (32%), Asia 130 million ha (42%), and South America 50 million ha (~10%). These data are very approximate but still interesting. The area losses by continent, as discussed in Section V, have varying significance for biological diversity.

The time series data of FORIS permit a study of trends in the rainforest loss (Singh et al., 1995). The curve shows an acceleration of loss from 1960 to 1980 followed by some stabilization from 1980 to 1990. Assuming this trend will continue, the rate of loss during 1990–2000 is projected to be approximately 50 million ha. The worldwide loss during 1960–2000 is approximately 180 million ha. Compared to the historic loss of 220 million ha, this suggests that most of the rainforest loss has happened during 1960–2000.

Some discussion of the reliability of FRA 1990 data sources may be useful in interpreting the previous estimates. For the purposes of global assessments, FAO continuously searches all published forest inventory

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>126,561</td>
<td>86,604</td>
<td>467</td>
</tr>
<tr>
<td>Asia</td>
<td>307,979</td>
<td>177,837</td>
<td>2235</td>
</tr>
<tr>
<td>America</td>
<td>500,060</td>
<td>450,891</td>
<td>1900</td>
</tr>
<tr>
<td>Total</td>
<td>934,600</td>
<td>715,332</td>
<td>4602</td>
</tr>
</tbody>
</table>

reports and maps. The published reports on country inventories are reviewed, relevant statistics are archived in the database, and the survey year and quality of estimates are noted. In general, the reference date and definition of data vary from country to country. However, data quality in FORIS is usually quite good and mostly derived from the interpretation of high-resolution satellite images by local professional staff. The FRA 1990 staff apply an adjustment for the definition and date to make all country estimates comparable on a global basis. These comparable estimates are termed standard data in contrast to source data.

A study was done by FAO (1996) on the differences between estimates derived from FORIS and the remote sensing-based sample survey. Means of forest cover and rate of deforestation were calculated by the two methods for each of the 10 forest cover strata. The correlation coefficient (R²) between the forest cover area was 0.94 and that between deforestation rates 0.64. A lower value of correlation for the deforestation is expected because the estimation is subject to many sources of errors as explained previously. Skole and Tucker (1993) and INPE (1992) studied the rate of deforestation in the same area (legal Amazonia), during the same period (1978–1988), and for comparable data sets. The first reported deforestation rate was 1.5 million ha and the second was 2.1 million ha, a difference of 30% in a geographic area of 5000 million ha. This highlights the complexity associated with the assessment of deforestation.

IV. TYPES OF CHANGES WITHIN THE FOREST LAND

The statistics presented in Section III mainly relate to the transfer of land from forest to nonforest. Changes within forest land are of greater interest from an ecological as well as an economic perspective. However, representative data on these changes, on a global basis, are very scarce or not available. The only source of regionally consistent information is the FRA 1990 Survey of the Tropics, which provided information on change processes in a matrix form (Table II). The matrices are ideal for tracking changes but not very easy to read. To give a bird’s-eye view of the change information, FRA 1990 developed a pictorial presentation called a flux diagram which shows area changes on the x-axis and associated average biomass changes on the y-axis. Thus, the figure as a whole shows the total biomass change in the form of a rectangle. Such a presentation is easy to read and informative for studies on climate change.

The FAO change matrices have nine rows and nine columns that represent the full range of land use and forest cover changes important for a study of biological diversity and climate change. The transition among the first four classes relates to changes within the forest domain and represents various grades of density or forest cover changes (closed and open), spatial disturbance (fragmented), and temporal disturbance (long fallow). The transition among the next four classes relates to transfers from forest to nonforest domain including shrub, short fallow, other land cover, and water. The last class, plantation, could be in the forest or nonforest domain (e.g., forest or agricultural plantation). These domains are grouped as one because it is not easy to separate them with certainty on satellite images.

A distinction needs to be made between the long fallow and short fallow categories, both of which are affected by shifting cultivation. The distinction between long and short has been made with the purpose of dividing the total area under shifting cultivation into “predominantly forested,” no matter how much degradation has occurred, and “predominantly nonforested” (but woody) on the basis of estimated intensity of cultivation or, specifically, of the rotation cycle. This distinction brings some clarity to the controversy about “shifting cultivation” which, without further subdivision, is sometimes considered nonforest (i.e., classified as “other wooded land” according to the FAO–FORIS definition) and sometimes forest (i.e., according to the definition used by some countries).

The information appearing in the flux diagram (Fig. 1) clearly shows that most of the changes result in the loss of the rainforest. Of the total rainforest lost, about 31% goes to permanent agriculture, 45% to short and long fallow (viz. shifting cultivation), 7% to fragmented forest, 3% to open forest, 11% to plantations, and 3% to other land uses. Most of these changes have far-reaching implications for biological diversity because they involve the introduction of fire and an external gene pool in the forest ecosystem. Logging is also a major cause of changes within forest land (Table III). Asia is the leader in terms of the area logged and the timber extracted, followed by Africa and South America.

V. IMPLICATIONS FOR BIOLOGICAL DIVERSITY

What is the implication of the previously mentioned changes for biological diversity? As expected, forest area lost and the condition of the remaining forests
<table>
<thead>
<tr>
<th>Classes as at year 1990 (million ha)</th>
<th>Total 1990</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed forest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open forest</td>
<td>685.36</td>
<td>0.96</td>
</tr>
<tr>
<td>Long fallow</td>
<td>11.94</td>
<td>0.16</td>
</tr>
<tr>
<td>Fragmented forest</td>
<td>0.27</td>
<td>0.01</td>
</tr>
<tr>
<td>Shrub fallow</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Short fallow</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Other land</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Water</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Total 1990</td>
<td>801.76</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*The number of sampling units was 26; land area sampled was 52.54 million ha. Sampling results have been expanded to the total land area covering 801.76 million ha.
(including degradation and fragmentation) jointly determine the loss of biological diversity. Overall, Asia has lost a major part of its historic rainforest area (about 42%). The current rate of deforestation is the highest among the three regions in absolute area terms. The area logged annually and the intensity of logging are also the highest compared to those of the other two regions. In Asia, only Indonesia still has some large areas of tropical rainforests left; the rainforests in the other countries are in a highly fragmented form, with species diversity under a high risk of loss (FAO, 1993).

Africa initially had a small share of the world's rainforest and had lost nearly 32% of the total by 1990. In particular, the west African region has undergone a very high loss of the historic rainforest area, with the remaining forests occurring in a highly fragmented manner like islands in a sea of forest fallow (FAO, 1993). The ongoing changes in the central African region were observed to be relatively slow during 1980–1990. However, the recent flux of population at the border of Zaire and Rwanda must have affected the rate of loss and biological diversity, but this effect is not yet known. Second, the ongoing forest area changes in Africa are from closed forest to short fallow, which is an extensive and destructive form of land use change.

![Diagram showing area transition on the x-axis and biomass transition on the y-axis. Each rectangle represents total biomass change associated with the transition (reproduced with permission from FAO, 1996).](image)

**TABLE III**

<table>
<thead>
<tr>
<th>Continent</th>
<th>Logged area (ha)</th>
<th>Logging intensity (m³/ha)</th>
<th>Log production (million m³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>91</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Asia</td>
<td>215</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>America</td>
<td>258</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>564</td>
<td>19</td>
<td>54</td>
</tr>
</tbody>
</table>

from the biological diversity perspective. The pattern of transfer is different from that of South America or Asia, where the land use change is more planned and intensive (Table IV). This implies that more land would be needed in Africa to feed the same number of people as are fed in other regions. With a high rate of population growth, such land use changes are a cause of concern from a forestry and a biological diversity perspective.

---

Table IV

<table>
<thead>
<tr>
<th>Region</th>
<th>Total area transferred</th>
<th>Transfer to nonwooded areas (mainly agriculture)</th>
<th>Transfer to other wooded areas (mainly short follow)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ha</td>
<td>ha %</td>
<td>ha %</td>
</tr>
<tr>
<td>Africa</td>
<td>2,036 100</td>
<td>32 3</td>
<td>1,984 97</td>
</tr>
<tr>
<td>S. America</td>
<td>11,510 100</td>
<td>7,775 68</td>
<td>3,736 32</td>
</tr>
<tr>
<td>Asia</td>
<td>9,941 100</td>
<td>3,824 38</td>
<td>6,117 62</td>
</tr>
<tr>
<td>Total</td>
<td>23,487 100</td>
<td>11,651 49</td>
<td>11,837 51</td>
</tr>
</tbody>
</table>

The land use changes in South America and the Amazon, in particular, are of more recent origin compared to those in Asia and Africa. The total area of the rainforest is still very high and the total area lost, therefore, is relatively low. However, a state-wise distribution of the remaining forest cover and the rate of deforestation provides a different view (Fig. 2). Only the states of Amazonas, Roraima, and Amapa have large tracts of inaccessible forests; other states, such as Maranhao, Tocantins, Para, Mato Grosso, Rondonia, and Acre, are undergoing larger than average forest area changes per capita compared to other regions of the world. Cattle ranching seems to be the main cause of deforestation, which is quite different from the causes of deforestation in the other regions of the tropics.

The rainforests are fragile ecosystems. The time series data on forest cover and population changes, collected by the FRA 1990 project from all regions of the tropics, show a more rapid forest area loss with an increase of the population pressure in the rainforest region compared to other ecological zones (Fig. 3). The rainforest, per unit area, is also the most biodiverse. The two facts imply a larger loss of diversity by human settlement and hence the need for a long-term perspective and care in planning of land use and forestry changes in the tropical rainforest zone.

See Also the Following Articles

DEFORESTATION AND LAND CLEARING • RAINFOREST ECOSYSTEMS, ANIMAL DIVERSITY • RAINFOREST ECOSYSTEMS, PLANT DIVERSITY • TROPICAL ECOSYSTEMS

Bibliography

Instituto Nacional De Pesquisas Espaciais (INPE) (1992). Deforestation in Brazilian Amazonia. INPE.
RANGE ECOLOGY, GLOBAL LIVESTOCK INFLUENCES
J. Boone Kauffman* and David A. Pyke†
*Oregon State University and †USGS Forest and Rangeland Ecosystems Science Center

I. Rangelands of the World
II. Biological Impoverishment Associated with Livestock Grazing
III. Conclusions: Human Institutions and Livestock-Caused Degradation

GLOSSARY
animal unit month (AUM) Quantity of forage needed to sustain one 1000-lb cow or five sheep for one month.
biological soil crusts Collective term referring to the combination of nonvascular and non-seed-forming plants that commonly occur on the soil surface of many arid and semiarid ecosystems. These crusts may consist of mosses, lichens, liverworts, green algae, and cyanobacteria. Synonyms: cryptogams, microfloral, cryptophytic, cryptogamic, microbiotic, organogenic, biogenic, biotic, and microphytic soil crusts.
browse That part of leaf, twig, and reproductive growth of shrubs, woody vines, and trees available for animal consumption. The act of consuming leaves, twigs, or reproductive parts of shrubs, woody vines, and trees.
degradation Decrease in ecosystem productivity or structure, and/or declines in native species diversity (sometimes with concomitant increases in exotic species) due to land use practices. Directly related to declines in biodiversity, ecosystem degradation encompasses soil impoverishment (e.g., compaction, erosion, salinization, loss of biological soil crusts) and hydrological alterations that diminish water availability. Similar to desertification.
desertification Process by which an area becomes more arid through the decline in primary productivity, elimination of biological cover, shifts in plant diversity, or degradation of soils due to human uses whose impacts exceed the sustainability of the landscape. Similar to degradation.
forbs Non-graminoid herbaceous plants.
graminoids Grasses (family Poaceae) and grass-like plants, mostly sedges (family Cyperaceae) and rushes (family Juncaceae).
graze Act of consuming herbaceous plants, including grass, grass-like plants, and forbs.
overgraze Consumption of forage to the extent that declines in productivity or desirable species composition are probable (i.e., degradation). Overgrazing is excessive use in terms of the amount, duration, and frequency of plant utilization.
pasture Lands in which the native vegetation has been removed in favor of cultivated grasses or other forage plants.
range condition Current productivity and composition of plants on a rangeland relative to its ecological potential. Often range condition is classified into arbitrary classes (i.e., excellent, good, fair, and poor).
rangeland Land on which the potential native vegetation is predominately grasses, grass-like plants, forbs, or shrubs. Rangelands include prairies, marshes, tun-
dra, wet meadows, savannas, shrubland steppe, chaparral, desert grasslands, and woodlands. Riparian zone. Interface between streams and terrestrial uplands. These zones are three-dimensional areas of direct interaction between the terrestrial and aquatic ecosystems. They often contain unique species and high levels of biological diversity.

Ruminant. Grazing mammal with a complex four-chambered stomach (including a rumen). Common ruminants include cattle, sheep, goats, antelopes, and giraffes.

Savanna. Ecosystem with a more or less continuous herbaceous layer dominated by graminoids and broad-leaved herbs with an overstory of trees or shrubs that covers less than 10% of the area. In contrast, grasslands are typified by a pure graminoid or herbaceous layer with no (or very few) tree or shrub elements that rise above the grass layer.

Woodland. Ecosystem in which the mature overstory tree or tall shrub cover makes up 10 to 60% of the area. Typically, a well-developed herbaceous understory is present. Note: forests typically have a mature tree cover >60%.

Livestock have played a pivotal and positive role in human development. Yet this has come at a heavy cost to the biological diversity of the world’s rangelands and forests. Livestock grazing is the most widespread land use occurring on Earth. In many regions of the world, grazing has reduced the density and biomass of many plant and animal species, reduced biodiversity, aided in the spread of exotic species and disease, altered ecological successions and landscape heterogeneity, altered nutrient cycles and distribution, accelerated erosion, and diminished both the productivity and land use options for future generations. Rangelands, a general term for grasslands, savannas, semi-arid shrublands, and woodlands, are a dominant terrestrial land cover type in all continents (except Antarctica). They contain a significant portion of the world’s biodiversity. Ecosystem degradation or desertification associated with livestock grazing is a significant problem throughout these rangelands. In addition, the majority of tropical forests in South and Central America are being deforested at alarming rates primarily for conversion to cattle pasture. In effect, the most biologically diverse ecosystems on Earth are being converted to simple pastures dominated by exotic grasses. To halt this degradation or to restore native ecosystems, innovative approaches to land use and livestock management need to be implemented. Ecosystems should be managed such that natural ecological and physical processes are allowed to occur with minimal disruption. Governmental institutions and policies are among the most significant barriers to land use change.

I. Rangelands of the World

Rangelands are defined as those lands where the native vegetation is predominantly grasses, shrubs, or open woodlands. Rangelands include grasslands, shrublands, savannas, open woodlands, and most desert, tundra (arctic and alpine), meadow, wetland, and riparian ecosystems. Occurring from the tropics to polar regions, rangelands cover more land area than any other type on Earth. Because of climatic or edaphic limitations, they tend to be lands that are incapable of growing marketable timber or agronomic crops without irrigation or some other modification (i.e., they are over poor-quality soils and/or are too wet, too dry, or too cold for cultivation). The composition, structure, productivity, and diversity of these ecosystems are governed by a combination of climate, geography, topography, and geology, including soil development. In addition, rangelands are used by a large number of vertebrate and invertebrate herbivores, including a diverse combination of domestic or native ungulates.

A. Tundra

Tundra ecosystems are dominated by perennial grasses, forbs, shrubs, and biological soil crusts consisting of cyanobacteria, lichens, and mosses (Fig. 1). These species can tolerate the climatic conditions of high latitudes or altitudes. Tundra ecosystems are generally set apart from forested ecosystems by a climatic tree line defined by the 10°C isotherm for the mean temperature of the warmest month. Tundra plants are low in stature with growth buds near the soil surface, where temperature is less variable. The morphology of herbaceous plants are often cushion or rosette growth forms and lichens are usually a foliose form. Shrubs are low-growing with short internodes between leaves. Temperate alpine tundra communities have short growing seasons similar to those in arctic tundra. In contrast, tropical alpine tundra communities of Africa, South America, and Oceania exist in conditions of freezing or near-freezing temperatures each night, with daytime temperatures allowing for active plant growth throughout the year.

The important herbivores of the Arctic are rodents and ungulates. Lemmings (Dicrostonyx spp. and Lem-
FIGURE 1 The arctic tundra landscape of the Brooks Range, Arctic National Wildlife Refuge, Alaska. (Photo by Boone Kauffman.)

Lemmings (Dicrostonyx spp.) can have great impacts on arctic ecosystems during years of high population densities. Lemmings exhibit different seasonal patterns of use within various parts of their habitats. During the winter, they use wetlands and graze the soil-level parts of plants while discarding the upper portions of the plants. This grazing pattern results in piles of graminoid litter known as lemming hay. During spring, lemmings move to south-facing slopes, where solar insolation results in the earliest available green biomass. During peak population cycles when densities are high, lemmings can consume between 20 and 70% of the available plant mass.

The dominant native ungulates in the Arctic include caribou or reindeer (Rangifer tarandus) and musk ox (Ovibos moschatus). Domestication of reindeer and musk ox has led to their introduction throughout the arctic tundra and to reindeer introduction into northern Antarctica. Alpine tundra provides seasonal habitat for wild ungulates, including elk, deer, red deer, mountain sheep, and mountain goats. Domestic livestock breeds adapted to cooler environments are also present in various arctic or alpine tundra communities of the world (e.g., cattle, sheep, goats, horses, llamas, and yak).

B. Grasslands and Savannas

Natural grasslands, also known as campos, llanos, pampas, plains, prairie, steppe, and veld, are frequently associated with semi-permanent high air pressure systems and high amounts of solar radiation. Annual precipitation typically ranges from 160 to 1700 mm with distinct wet and dry seasons. Grasses or graminoids dominate the vegetation composition, with forbs being a secondary component and short-stature shrubs contributing a lesser amount. The grass (or graminoid) dominance of grasslands is the result of current climatic conditions and/or natural disturbance processes such as fire. With human degradation such as overgrazing or fire suppression, shrubs often increase in abundance.

In the tropics and in temperate locations with summer precipitation, C₄ grasses (warm-season grasses) will predominate because of their optimal growth at warmer temperatures. Northern temperate locations, where moisture is most available when temperatures are cooler, favor C₃ grasses (cool-season grasses). Biological soil crusts frequently occupy the soil surface below or between the plant cover, particularly in drier desert grasslands.

In equatorial regions, tropical savannas are a dominant rangeland ecosystem. They form a transitional zone along a precipitation gradient from tropical rain forests to arid deserts. Dense grasslands with occasional trees and shrubs characterize tropical savannas. As in grasslands, distinct wet and dry seasons typify the climate of tropical savannas. Disturbances are a prevalent feature of savannas; during the dry season, lightning-caused fires may occur. Human-ignited fires are very common throughout the world and maintain the open
nature of many tropical savannas. Fire-return intervals in savannas are among the most frequent of any landscape on Earth. It is estimated that 50–75% of the humid savannas of Africa and South America burn each year.

Tropical savannas are highly productive. Grass biomass in humid savannas (>700 mm precipitation yr$^{-1}$) averages about 6.6 Mg ha$^{-1}$ in Africa and South America, and 4.9 Mg ha$^{-1}$ in tropical Asia. In savanna-woodlands, aboveground biomass may be as high as 25–61 Mg ha$^{-1}$. Roots and other subterranean tissues account for the majority of the plant biomass in savannas and grasslands (de Castro and Kauffman, 1998).

Tropical savannas are also very species-rich. For example, the largest savanna type in South America is the Brazilian cerrado, which often contains over 400 species ha$^{-1}$ (Fig. 2). Except for tropical rain forest, this is among the richest vascular plant assemblages on Earth. For example, about 3500 plant species and 400 bird species have been identified in the cerrado of the 5800-km$^2$ Federal District surrounding Brasilia, the capital of Brazil.

Savannas can experience rapid changes in species composition and structure following the introduction of an anthropogenic perturbation. Changes in the patterns of fire are one example; fire suppression results in an increase in the dominance of tree and shrub vegetation. Conversely, increasing the occurrence of fire enhances grass abundance. Linkages between fire, grazing animals, disease, and vegetation structure were described by Sinclair (1979). He related how the elimination of a ruminant disease, rinderpest, was a perturbation that changed the structure and composition of landscapes within the Serengeti of Africa. The immediate effect of rinderpest elimination was an increase in wildebeest numbers. On the plains, increased grazing pressures reduced grasses and increased forbs. Grant’s gazelle increased along with the forbs, thus leading to population increases in cheetah. In woodlands, grazing reduced the grasses, which in turn reduced the fuel to support fires. The reduction in grasses resulted in a reduction in buffalo. Tree numbers then began to increase, which favored an increase in the giraffe populations.

Plants within natural temperate grasslands and tropical savannas vary in their tolerance to large numbers of grazing animals. The images of grasslands and savannas for many people include large herds of ungulates, such as bison (*Bison bison*) in the Great Plains of North America and wildebeest in East Africa. In these ecosystems, large herds of ungulates were a keystone feature that strongly influenced composition and structure. Plants are well adapted to herbivory in these grass-dominated ecosystems. However, large herds of ungu-

---

**FIGURE 2** Overview of the Brazilian cerrado. This is a mosaic of grasslands, savannas, and woodlands that have a high level of biotic richness. Note the fire in the background, which is a common disturbance in this ecosystem. Land conversion to cattle pastures, soybeans, and other croplands is occurring at alarming rates in this ecosystem. (Photo by Boone Kauffman.)
lates are not a natural feature in all grasslands and savannas (e.g., the Intermountain and Palouse grasslands of the United States and the Brazilian cerrado). In these grasslands, rodents and insects are the dominant native herbivores.

The most widely distributed wild ungulates in North American grasslands and woodlands are pronghorn (*Antilocapra americana*), elk (*Cervus elaphus*), and deer (*Odocoileus* spp.). However, only bison comprised large herds within the Great Plains until they were decimated in the late nineteenth century by hunters. Bison also occurred in other North American grasslands, but not in large herds. Rodents and lagomorphs are also common grazers in North America. Population eruptions can result in years when the vegetation is heavily grazed by these animals in local areas, but within a few years disease, predators, competition, or other density-dependent factors reduce their numbers.

The common ungulates in South American grasslands and savannas include tapirs (*Tapirus* spp.), peccaries (*Catagonus, Dicotyles, and Tayassu* spp.), camels (*Lama* spp.), and deer (*Odocoileus* and *Ozotoceros* spp.). Some of these are primarily forest or gallery forest (riparian) dwellers, but they also use savannas or grasslands. Only one native lagomorph (*Sylvilagus brasiliensis*) occurs in South America. The dominant vertebrate grazers in South American savannas are a diverse group of rodents. The largest rodent, the capybara (*Hydrochoerus hydrochaeris*), is widely distributed, but prefers wetland savannas such as the Brazilian Pantanal. Capybara graze in herds as large as 1500 animals.

The African grasslands and savannas maintain the greatest species richness of wild ungulates of any range-land ecosystem in the world (Fig. 3). Rodents comprise a lower herbivore biomass than ungulates in most regions of Africa. The African ungulates consist of three taxonomic orders: Proboscidea (1 species; elephant); Perissodactyla (6 species; rhinoceros and zebras); and Artiodactyla, with five families. The families include Suidae (3 species; pigs), Hipposomatidae (2 species; hippopotamus), Tragulidae (1 species; water chevrotain), Giraffidae (2 species; giraffe), and Bovidae (79 species; buffalo and antelope). Although a few of the African ungulates are forest-specific species, most have distributions that include grasslands, savannas, and woodlands. Riverine and wetland locations with high rainfall in East Africa maintain populations of hippopotamus, waterbuck, and buffalo, whereas midheight grass regions with intermediate moisture support zebra, giraffe, and hartebeest. Native ungulates have been extirpated or driven to extinction in many regions of Africa. In the Fété Oé savanna of the Sahel, 11 species of native ungulates are now extinct.

Northern Asia has few native ungulates within its grassland/savanna ecosystems. Common ungulates include gazelle (*Procapra or Gazella* spp.), wild horse (*Equus przewalskii*), and the Bactrian camel (*Camelus bactrianus*), and none of these form large herds. Rather,
In the Australian grasslands, large grazing animals are represented by marsupials (euros and kangaroos, Macropus spp., wallabies, Petrogale spp.). The red kangaroo (Macropus rufus) grazes in the arid and semi-arid zones, whereas the antilopine kangaroo is restricted to tropical grasslands of the north. Introduced herbivores, such as the European rabbit (Oryctolagus cuniculus) and the Timor water buffalo (Bubalus bubalis), compete for herbage with native grazers in southern Australia. In Australia as throughout the world, the high productivity of grasslands and savannas has led to their widespread use by domestic cattle, sheep, and goats.

C. Temperate or Cold-Desert Shrublands and Semi-deserts

Deserts or desert shrublands can be broadly classified into two different types—hot deserts of subtropical and tropical latitudes and cold deserts of temperate latitudes. Temperate deserts and semi-deserts occur at higher latitudes than hot deserts (between 30° and 30°N and 5°S), which allows for moisture from oceans to move into these regions by way of prevailing circulation patterns. Most of the temperate deserts and semi-deserts are characterized by sparse vegetation that is usually dominated by low shrubs and herbaceous plants. Total vascular plant cover rarely exceeds 30%, and biological soil crusts commonly occupy interspaces between vascular plants.

Cold deserts are characterized by cold winters and hot summers. Most occur in North and South America (Great Basin, Intermountain West, and Patagonia) and Eurasia (Kazakh-Dzungarian deserts and semi-deserts). In North America and western Eurasia, most of the annual precipitation arrives as snow in the winter, while summers are dry. However, eastern Eurasia receives most of its moisture during the summer from frontal or cyclonal storms that move west from East Asia, while the winter remains dry. Patagonian semi-deserts receive low amounts of precipitation throughout the year, with strong winds exacerbating arid conditions during the summer from elevated evaporation.

The importance of native ungulates in cold deserts varies widely among these geographic regions. Large herds of ungulates, largely saiga antelope (Saiga tatarica), zheiran gazelle (Gazella subgutturosa), and wild ass (Equus hemionus), once roamed Eurasian ecosystems (until 100–200 years ago), but are now largely restricted to reserves. Wild sheep are common to both Northern Hemisphere ecosystems, with Ovis ammon cycloceros present in Afghanistan and Iran and Ovis canadensis in North America. Common native ungulates in North American cold deserts include pronghorn, deer, and elk. Feral horses and burros are now locally common in areas. Although bison were once found in this region, they never attained large numbers as found on the Great Plains and were largely absent by the time European exploration began. Only one native Patagonian ungulate exists, the guanaco (Lama guanicoe). Today domestic herbivores are predominate throughout all of these regions.

D. Hot Deserts

Hot deserts (or hot desert shrublands) are the most arid form of rangeland in the world. The hot deserts generally occur between latitudes of 15° and 30°N and S, the belt of the subtropical anticyclones that create near permanent high-pressure systems. Most hot deserts receive less than 120 mm of rain annually. The position within the high-pressure belts also results in a wide variation in the amount, location, and timing of rainfall. A wide spacing of drought- and heat-adapted shrubs is characteristic of hot desert shrubland vegetation. Perennial vascular plants normally cover less than 10% of the land area within hot deserts. Hot desert vegetation must be adapted to the uncertainty in the timing of adequate growing conditions that are characteristic of this ecosystem. The vascular plants persist in the variable and unpredictable growing conditions by using a host of adaptations to avoid dry seasons while quickly responding to moisture when it becomes available. Perennial plants often have persistent belowground organs that facilitate survival during the dry season. When moisture is available, some perennial plants can respond with a rapid production of adventitious roots and leaves. Many desert species (the Cactaceae, some Euphorbiaceae, and others) have photosynthetic stems that may allow growth during sporadic
storm events without the cost of leaf production. Plant reproduction is limited to years with sufficient moisture. Ephemeroid annuals survive in deserts by germinating and completing their life cycles during wet periods and then persisting in the soil as a dormant seed during dry periods. In locations where fog is the major source of water (coastal deserts of Chile, Peru, and Namibia), fog desert vegetation forms, with biological soil crust species often being the dominant ground cover.

Native ungulates are virtually missing from most hot deserts except along major rivers and wetlands. They may occupy the fringes of the hot deserts and migrate into the deserts during favorable conditions. An example is the desert bighorn sheep in the North American Mojave Desert. Though populations are found in adjacent semi-deserts where greater food and water are available, use of the Mojave Desert by bighorn only occurs during wet periods. Australian marsupials follow a similar pattern, with most only using the arid desert areas during favorable times. A few wallabies are found in rocky arid areas, but never in large numbers. Herbivores are found in low densities in the North African (Sahel and Sahara) and Middle Eastern deserts. Native ungulates of these deserts include gazelle, ibex, oryx, wild asses, and wild boar. The exception to a low density of ungulates in hot deserts is found in the Karoo and Kalahari Deserts of southern Africa. Here, springbok, gemsbok, and burchell's zebra disperse in small herds or treks during dry periods. Once moisture produces aboveground growth of grass, these animals congregate into larger herds. Populations of these animals 200 years ago were larger than those today. Although forage is scarce and unpredictable, domestic livestock (cattle, sheep, and goats) grazing remains common in hot deserts of the world.

II. BIOLOGICAL IMPOVERISHMENT ASSOCIATED WITH LIVESTOCK GRAZING

Domesticated animals (livestock) have played prominent and largely beneficial roles in human society for thousands of years, providing food, fuel, fertilizer, transport, and clothing. Yet livestock have had a dramatic negative impact on global biodiversity. The influence of livestock on global biodiversity is of concern simply because cattle and all other ruminant livestock graze about one-third to one-half of the planet's total land area. Along with pigs and poultry, they eat feed and fodder raised on one-fourth of the cropland. Livestock grazing is the most ubiquitous human activity on Earth and occurs on more area than any other land use. The impacts of livestock are far-reaching and often not immediately apparent (Table I). In many regions of the world, grazing has reduced the density and biomass of many plant and animal species, reduced biodiversity, aided in the spread of exotic species and disease, altered ecological succession and landscape heterogeneity, altered nutrient cycles and distribution, accelerated erosion, and decreased both the productivity and land use options for future generations.

Given the widespread degradation caused by livestock on the diversity and structure of native rangelands, one would assume that rangelands are the source of a large proportion of the feed utilized by the grazing animals on Earth. Yet it is estimated that three-quarters of the world's 3 billion domestic ruminants are raised in conjunction with farming and are fed crop residues, hay, and other forages such as alfalfa. Roughly 38% of the world's grain, especially corn, barley, sorghum, and oats, is fed to livestock. In the United States, livestock accounts for 70% of the domestic grain consumption; in contrast, India and sub-Saharan Africa provide about 2% of their cereal harvest to livestock. A 1983 U.S. Department of Agriculture report stated that ≈103 million ha of fertilized pastures and forage crops provided 84% of the U.S. livestock forage. The ≈312–319 million ha of U.S. rangelands provided only 16% of the forage. Two-thirds of these rangelands are privately owned, and the rest are publicly owned. In 1984, permits issued by the U.S. Bureau of Land Management accounted for less than 4% of the forage consumed by the U.S. herds.

Rangelands of the world are important reservoirs of biological diversity as well as water, natural beauty, and other plant, animal, and mineral resources. For example, native rangelands provide essential habitat for a wide variety of wildlife and plant species. An estimated 84% of U.S. mammal species and 74% of bird species are associated with rangeland ecosystems (particularly using riparian zones within these semi-arid landscapes). Public rangelands in the United States are home to more than 3000 wildlife species, including about one-third of the nation's threatened or endangered species. Compared to forests, rangelands have received comparatively less attention with respect to environmental degradation in general, and losses in biodiversity in particular.

Because of the great differences in composition, soils, climate, and historical and current land uses among rangelands of the world, grazing effects will vary (see Table I). Degradation occurs when the frequency, intensity, or season of use exceeds the biological thresholds
TABLE I
Some Direct and Indirect Influences of Domestic Livestock on Biological Diversity

- Desertification or degradation—the impoverishment of native plant and animal ecosystems.
- Deforestation/forest conversion to pasture.
- Contribution of greenhouse gases to the atmosphere—directly from methane emissions and carbon losses due to forest conversion.
- Dispersal of exotic species and/or the creation of conditions suitable for their increase.
- Spread of disease to native herbivores.
- Geomorphic influences on stream channels (widening and incision) that disrupt floodplain/stream interactions.
- Water diversions for livestock forage production.
- Development of permanent water sources in areas where they traditionally did not occur.
- Erosion/sedimentation or fecal inputs into aquatic ecosystems.
- Influences on biogeochemical cycling.
- Control of wildlife species perceived as competitors or predators.
- Applications of herbicides and pesticides to control undesirable plants and insects.
- Altering natural fire regimes—elimination of fires in natural grasslands and shrublands or frequent introduction in pastures converted from tropical forests.
- Social causations—overpopulation, government incentives, culture, and customs.

A. Effects of Livestock on Terrestrial and Aquatic Ecosystems

Figure 4 presents a simple model depicting how livestock influence biological diversity and ecosystem integrity. The direct or primary livestock influences are immediate and readily observable in the field. These direct influences include: (1) removal of vegetation through grazing; (2) the trampling of soils, vegetation, and biological soil crusts; (3) the redistribution of nutrients via forage removal, defecation, urination, gaseous loss, and animal gain; and (4) the dispersal of exotic plant species and pathogens. The cumulative effects of these four primary influences lead to a suite of physical and biotic responses or adjustments within the ecosystem or landscape. These are the secondary influences and they include alteration in disturbance regimes (fire cycles), accelerated erosion, altered hydrology (e.g., runoff, infiltration rates, and water-holding capacity), altered competitive relationships among organisms, and changes in plant and animal reproductive success and establishment patterns of plant seedlings. Tertiary effects include ecosystem persistence or recovery. Plant communities that evolved with large herds of ungulates may have adaptations that allow persistence with high levels of defoliation by domestic livestock. Conversely, those rangelands that evolved with low densities of large herbivores may be susceptible to alterations following the introduction of even low levels of domestic livestock impacts. Unsustainable grazing and ranching practices result in the loss of forests, grasslands, shrublands, savannas, and the indigenous species that inhabit them. As the most pervasive land use, livestock grazing has had widespread and dramatic ecological impacts, including loss of native species, changes in species composition, soil deterioration, degradation of fish and wildlife habitat, and changes in ecosystem structure and function. The loss of biological diversity and productivity of rangelands due to overgrazing is a concern because it diminishes both human and ecological welfare. However, livestock influences on biodiversity are not limited to rangelands. The range of environmental effects of livestock encompasses such varied land uses as conversion of tropical rain forest to pasture in the Amazon Basin, water pollution from manure associated with large livestock factory farms in Europe and the United States, and the desertification of semiarid rangelands in the African Sahel, western United States, and interior Australia.
FIGURE 4 A conceptual model of the influences of livestock on ecosystem structure and function. Direct influences are those that are immediate and readily observable in the field. The synergistic effects of these influences lead to the secondary influences, the physical and biological adjustments in ecosystems. The tertiary influences are the long-term alterations in landscapes due to the introduction of livestock. Tertiary influences are the cumulative effects of the primary and secondary influences of livestock and are reflective of varying degrees of desertification or degradation. The degree of alteration would be affected by unique characteristics of the ecosystem (i.e., soils, topography, climate), as well as the intensity and frequency of livestock use.

The effects of herbage removal depend on a number of environmental, animal, and land management factors. The quantity of plant mass consumed and season, duration, and frequency of grazing are important determinants of how herbivory affects ecosystems. Overgrazing and degradation of ecosystems are a combination of excesses in all of these factors. The season of use refers to the phenological stage during which the plants are grazed. Typically, grazing when plants are dormant is less detrimental than grazing during active growth. Season of use may also influence the reproductive success of plants if flowers and seed heads are not allowed to form or are eaten. In addition, the degree of harm is dependent on the duration and amount of grazing. Declines in plant vigor will occur if grazing is of such duration or intensity that plants are not allowed to regrow and replenish carbon and nutrient reserves. For shrubs and trees, grazing more than the current year's growth will reduce the size of the plant and, depending on the plant's morphology, can result in the death of the plant. Finally, the frequency of grazing is an important determinant. Plants, other organisms, or soils in ecosystems in which grazing is interspersed with periods of rest have a greater opportunity for recovery than in ecosystems that are grazed continuously or annually.

Trampling damage occurs via the mechanical compression and compaction of soils or the physical destruction of biological soil crusts or vegetation as ani-
Riparian Zones: Hot Spots of Biodiversity on Rangelands

Riparian zones are the unique environments adjacent to rivers and streams and comprise assemblages of plant and animal communities whose presence can be either directly or indirectly attributed to factors that are stream-induced or stream-related (Kauffman and Krueger, 1984). Functionally, they are three-dimensional zones of direct interaction between terrestrial and aquatic ecosystems. Boundaries of riparian zones extend outward to the limits of flooding and upward into the canopy of streamside vegetation (Gregory et al., 1991). These zones perform a variety of ecosystem functions, including influences on water quality, flood attenuation, and provision of valuable fish and wildlife habitats.

Riparian areas are hot spots of biodiversity. This is particularly true in arid and semiarid environments, where riparian zones may be the only tree-dominated ecosystem in the landscape. The presence of water, increased productivity, favorable microclimate, and unique disturbance (flood) regimes combine to create a disproportionately higher biological diversity than the surrounding uplands. In the Intermountain West and Great Basin of the United States, about 85% of the wildlife species are dependent on riparian zones for all or part of their life cycles. In these same riparian zones, well over 100 plant species commonly can be found on a single gravel bar (<50 m in length along the streambank).

Livestock grazing is among the most significant land uses affecting the structure and productivity of riparian zones. Because of the high plant productivity, close proximity to water, favorable microclimate, and level ground, these areas are preferred habitat for livestock, and overgrazing commonly occurs. Grazing changes the composition and structure of the riparian zones, which lead to dramatic effects on the large numbers of wildlife that use the areas for feeding, breeding, or hiding cover. Of particular importance are the effects of livestock grazing on streamside forests of cottonwood, aspen, and willows. Long-term overgrazing can eliminate the woody vegetation that provides essential breeding bird habitat. In addition, trampling in riparian zones decreases infiltration and breaks down streambanks, resulting in channel incision or widening. These influences on streambank structure sever the linkages between the riparian and aquatic zones through the elimination of overbank flooding. The removal or loss of riparian vegetation results in the decline of organic inputs into the stream and the loss of overstory vegetation shade, which can dramatically affect the aquatic biota, such as salmon and trout.

Because of the high level of biodiversity and other ecosystem values of riparian zones, particularly in semiarid and arid environments, their restoration yields many positive benefits. Riparian zones are naturally an environment with frequent disturbances (floods), and because of this, the adapted biota often exhibit considerable resilience following the cessation of land uses that cause degradation or that prevent recovery. In rangelands, this most often includes abusive grazing practices. Thus, the cessation of overgrazing is the first important step in riparian restoration and should be implemented prior to determining if other active restoration measures need to be taken. For lasting riparian restoration, a larger effort should be aimed at the recovery of the entire watershed (Kauffman et al., 1997).


mals move across the land. The break-up of soil crusts (both physical and biological) can increase soil loss via wind and water erosion. The susceptibility of soils to compaction is related to soil texture (sand, silt, and clay content), as well as the season of grazing. Clay soils are more susceptible to compaction than sandy soils and wet soils are more susceptible than drier soils. Soil compaction results in reductions in soil porosity, water-holding capacity, and rates of infiltration. All of these result in less water available for plant growth and more water lost via surface runoff.

Another direct impact of livestock is the influence on nutrient inputs and redistribution. The effects of herbivory on litter reduction, trampling, and soil compaction include direct influences on nutrient cycles.
Trampling of biological soil crusts can reduce nitrogen inputs while increasing nitrogen losses via erosion. Urination and defecation are forms of nutrient redistribution. These inputs can enhance soil fertility, but can also have serious effects on water quality and aquatic organisms. Concentrated nutrient inputs can increase bacterial and protozoan pathogens, promote algal growth, and alter water chemistry—particularly the depletion of dissolved oxygen.

The direct effect of livestock grazing is the dispersion of undesirable exotic organisms (as well as native weeds). Seeds of exotic species may be transported on the fur or hooves of animals or deposited in the feces. As livestock move through an area, they may deposit exotic plant seeds while simultaneously preparing a seedbed via trampling. A grazing preference for desirable native species may further improve the competitive advantage of the less desired weeds, thus increasing their probability of establishment and dominance. Animals do not graze indiscriminately. Different animals have preferences and select some plant species over others. This will confer a selective advantage on less palatable species over those preferred by livestock. Often exotic species are less palatable and more grazing tolerant than native species. Livestock grazing in range-lands where native species are not well adapted to large herbivores will confer a selective advantage to exotic plant species.

Livestock also spread exotic diseases to native herbivores. Numerous herds of bighorn sheep in the United States have been eliminated following disease transmission from domestic sheep.

2. Indirect Effects: Secondary and Tertiary Influences

The direct or primary influences of livestock elicit a number of feedback responses that affect additional species, functions, and processes. The outcomes of these far-reaching effects are described as secondary influences and include changes in landscape disturbance cycles (e.g., fire regimes), accelerated rates of erosion, alterations in hydrology and plant available water, and alterations in successional patterns due to changes in competition and reproduction of both the native and exotic species (see Fig. 4).

The cumulative impacts of livestock grazing are apparent in many desertified or degraded ecosystems throughout the world. The secondary influences of livestock overgrazing on ecosystems are the characteristic endpoint of desertification or land degradation (Fig. 4). The tertiary effects include dramatic alterations to the biotic structure, composition, and productivity of ecosystems. These are closely linked to simultaneous degradation of hydrological and soil properties. Declines in the cover of native perennial plants and biological soil crusts and concomitant increases in bare ground, unpalatable shrubs, or noxious weeds and annuals are examples of the long-term degradation of ecosystems. Tertiary effects of livestock are often difficult to separate from other significant influences on ecosystems. Degraded landscapes are often the cumulative result of inappropriate livestock management in concert with other poor land management practices, including logging, road building, inappropriate agronomy, fuel-wood harvesting, mining, water diversions, and over-harvesting of plant and/or animal resources (i.e., hunting, poaching).

B. Livestock Grazing and the Desertification of Native Rangelands

Desertification of rangelands has had strong negative effects on Earth’s biodiversity and the continued rate of desertification represents a major global threat. Desertification is manifested in a combination of declines in native species diversity, ecosystem productivity, loss of topsoil, changes in nutrient cycles, and alterations in hydrology and microclimate. For example, Flather et al. (1994) reported that 40% of all U.S. federally listed Threatened and Endangered (T&E) species were associated with rangelands. More T&E plant species occur on rangelands in the United States than any other land cover type. Trends in biodiversity loss are likely similar in other rangelands of the world. This is particularly true for large native herbivores, which are now largely restricted to small preserves or parks. Nearly all of Earth’s rangelands have been degraded by human activities, including livestock grazing, the introduction of exotic species, fuelwood harvesting, alteration of natural fire cycles, wildlife depredation, and conversion to cropland or urbanization.

Estimates by the United Nations Environmental Programme (UNEP, 1990) indicated that 73% of the world’s 3.3 billion ha of dry land is at least moderately desertified, having lost more than 25% of its productive capacity. It has been estimated that ≈84% of the world’s rangelands are at least moderately desertified (i.e., conditions in which plant productivity has declined by ≥25%; or indications that accelerated wind or water erosion has occurred due to land use; or soil salinity has reduced crop yields 10–30%). Of the 3.1 billion ha of rangeland in the world, Mabbutt (1984) estimated that as many as 1.3 billion ha are severely degraded (i.e., productivity loss exceeds 50%).
The UNEP study on soil degradation reported that over the past 45 years about 11% (1.2 billion ha) of Earth’s vegetated soils have become degraded to the point that their original biotic functions were damaged; reclamation would be costly or in some cases impossible. Most of these degraded soils are suffering from “moderate” degradation, that is, productivity is greatly reduced but can still be used for agriculture. The soil’s original biotic functions (its capacity to process nutrients into a form usable by plants) have been partially destroyed and only with major improvements can productivity be restored. A smaller proportion (~3% of Earth’s surface) showed severe degradation, where original biotic functions are largely destroyed and may be reclaimable only with major financial and technical assistance. Nine million hectares of land were classified as extremely degraded, that is, un reclaimable and beyond restoration. On this land the original biotic functions are fully destroyed. At continental scales, land areas with soil degradation classified as moderate to severe ranged from 4% in North America, to 14.4% in Africa, to 24% in Central America and Mexico.

Throughout the world, accelerated wind and water erosion is the process that causes the majority of soil degradation, and overgrazing is the most ubiquitous cause of the acceleration in the erosion process. Globally, 33% of the soil degradation that has occurred since 1945 was attributed to overgrazing (UNEP, 1990). Thus overgrazing is the most pervasive cause of soil degradation, affecting 679 million ha. In Africa and North America, grazing was the principal cause for 49 and 30% of the soil degradation, respectively. Soil erosion resulting from the decrease in plant cover and trampling may require thousands of years for recovery, particularly in semiarid and arid ecosystems.

Perhaps the largest data set showing trends in the ecological condition and biological impoverishment of rangelands at large scales (country or continental scales) are those of public lands in the western United States. Determining the pristine state of rangelands is difficult, and current approaches are being challenged as to their usefulness for truly ascertaining ecological condition. This is because few areas of undisturbed or pristine rangeland exist (i.e., areas with no history of livestock grazing), unlike the case with old growth or primary forest (i.e., never deforested). In addition, comparison of the degree of departure from a single hypothetical climax composition is specious at best. Keeping in mind these uncertainties associated with ecological interpretations of rangeland condition, only about 4% of nonarctic rangeland in the United States is considered to be in excellent condition (i.e., covered with native vegetation and suffering no loss in productivity). Most U.S. rangeland is in fair or good condition (i.e., slight to moderate degrees of degradation), but a significant portion (13%) is categorized as poor or severely degraded. In 1990, the U.S. Bureau of Land Management reported that only 33% of its holdings were in good or excellent condition.

Range trends, whether the rangelands are improving, degrading, or static, are a measure of how well rangelands are being managed. The results vary depending on the parameters examined. The U.S. land-managing agencies suggest that about 86% of rangelands are either improving or static (i.e., exhibit no discernable trend), and that 14% are degrading. Yet portions considered static or stable may be in such a degraded ecological state that further decline would be difficult to discern. For example, many western rangelands are completely dominated by exotic annuals or weeds, and others have experienced extensive shrub encroachments during the latter half of the twentieth century. These patterns make it difficult to interpret landscape-scale changes in ecological condition.

Though many scientists believe that many U.S. rangelands are currently in the best ecological condition of the last 60 years, others point out that more than 50% of the rangelands contain less than half of their potential natural plant composition after more than 50 years of “modern range management” (i.e., directed toward recovery and sustained use). Because of the loss of soil and native species, coupled with ongoing invasions of exotic species, many suggest that desertified North American rangelands will not likely regain their former diversity in time frames of less than a century. Research and innovative approaches for the restoration of biological diversity of rangelands are needed.

C. Conversion of Tropical Forests to Livestock Pasture

Livestock impacts are not limited to Earth’s rangelands. Rain forests are among the most biologically diverse ecosystems in the world. Tropical forests cover about 720 million ha and contain 40–90% of the world’s species. The dramatic rate of conversion of tropical forests to cattle pastures is one of the most deleterious environmental impacts on global biodiversity in the latter half of the twentieth century. Conversion of tropical forest to pasture not only results in the extinction of native rain forest plants and animals, but also alters the microclimate, hydrology, and fire cycles, and is a significant source of greenhouse gases.
Box 2

Biological Soil Crusts: Their Role and Susceptibility in Rangeland Ecosystems

Biological or microphytic soil crusts consist of a combination of nonvascular plants (also known as cryptogams; these include cyanobacteria, mosses, lichens, liverworts, and green algae). Biological soil crusts serve a number of ecosystem functions that make them an important component of the biological diversity of many of the semiarid and arid environments of the world. These soil crusts exist in nearly all ecosystems where vascular plant material (live or dead) covers less than 100% of the ground surface; they occupy the interspaces between vascular plants. Spatial coverage within some ecosystems may be greater than that of vascular plants, ranging from 10 to 100% coverage in some undisturbed temperate and arctic ecosystems. Biological soil crusts are adapted to extreme environmental conditions of extended drought and extreme high and low temperatures. These species are metabolically active only after hydration. Some are able to hydrate under conditions of saturated air, such as fog, dew point, or high vapor pressure; others require liquid water.

The functional role of biological soil crusts in semiarid and arid ecosystems is multifaceted. Some species are involved in the breakdown of humus and in the release of nutrients. However, a more significant role in some ecosystems may be their direct contribution of increasing nitrogen availability to other plants and organisms within the ecosystem through biological nitrogen fixation. Nitrogen is often an important limiting factor for ecosystem productivity. Biological soil crusts (cyanobacteria and cyanolichens) can be important sources of nitrogen for many semiarid ecosystems, such as the North American Great Basin and Colorado Plateau. Fixed nitrogen is immediately released by these organisms into the surrounding soils, however; the amount taken up by vascular plants has not been widely studied. Biological soil crusts can diminish the rate of wind and water erosion from rangelands. Soil stability is enhanced by these soil crusts through the formation of exudates of polysaccharides that bind soil particles into aggregates. They also protect the soil from raindrop erosion (the dislodging of soil particles by individual raindrops, also known as splash erosion) by absorbing the kinetic energy of raindrops. The potential to diminish soil erosion varies with biological crust composition, increasing from algal, to lichen, to moss-covered crusts. Crusts create a surface roughness that slows the surface runoff of water and may trap soil particles arriving via wind deposition. The phototropic nature of the crust species is an adaptation that contributes to their upward growth and soil entrapment when soil buries the crusts. They are important contributors to dune stability when dunes consist of sufficient amounts of fine soil particles (sands and clays).

Complex direct and indirect mutualistic relationships exist between biological soil crust and vascular plants. In addition to nitrogen fixation and soil stability, the presence of soil crusts is positively correlated with floristic diversity in several ecosystems around the world. This positive correlation may be related to the increased seedling establishment and survival of vascular plants when grown in association with biological soil crusts. The surface variation that crust species provide may also contribute to greater variation in suitable sites for successful seedling germination and establishment.

Throughout the rangelands of the world, the introduction of large numbers of domestic ungulates has damaged biological soil crusts via herbivory and trampling. Herbivory by ungulates is generally restricted to locations where domestic reindeer have overgrazed lichen-dominated winter habitat in arctic tundra ecosystems. Globally, the most prevalent ungulate impacts are related to trampling, which breaks the dry, brittle lichens into small pieces that are then blown from the site. In many rangelands, intense grazing by livestock around watering places results in the elimination of lichens near the water, as distance from water increases, livestock impacts on lichen abundance decrease. Biological soil crusts growing on soils with low aggregate stability (i.e., sandy soils) are more susceptible to trampling damage during dry periods. The disruption of the dry soil surface breaks the bonds between the biological soil crust and the soil. Both the soil particles and the crust species are then susceptible to wind and water erosion.

Following disturbance or loss, biological crust recovery time depends on the severity of the disturbance and the environmental characteristics of the ecosystem. Primary recovery begins with cyanobacteria within one year after disturbance.
Belnap (1993) estimated that some ecosystems in the U.S. Southwest may require 30–40 years for the replacement of sheath material in the soil, 45–85 years for lichen diversity to recover, and over 250 years for moss cover to return if left undisturbed. Recovery of crust species is more rapid in locations with higher effective precipitation and finer-textured soils, and if inoculating material is present. Clearly, ecosystems in which biological soil crusts are a major component need appropriate grazing management or protection to maintain or restore biodiversity and to prevent degradation, erosion, and desertification. 


Most of the deforestation in the tropical rain forests, moist forests, and dry forests of Latin America is for conversion to cattle pasture. In addition to soil, climatic, and vegetative characteristics, socioeconomic status of the landowner is a determinant of the patterns of land use. For example, on large ranches in the Brazilian Amazon, tropical forests are converted directly to cattle pasture. In contrast, small-scale subsistence farmers may use the land for one or two shifting cultivation cycles before ultimately converting to cattle pasture.

Slashing and burning tropical forest for cattle production converts some of the most structurally and biologically diverse ecosystems on Earth to simple pastures dominated by exotic grasses (Fig. 5). Rain forests frequently have over 100 canopy tree species per hectare and millions of highly interdependent organisms living within the forest and soil. In contrast, the only trees, if any, in pastures are cultivated fruit and nut trees or highly invasive second-growth trees. Rather than the diverse array of birds, mammals, insects, reptiles, amphibians, and other native tropical fauna, the pastures are dominated by the cattle and the few other animals that thrive in human-modified environments.

Deforestation affects biological diversity in three ways: (1) destruction of habitat; (2) fragmentation of formerly contiguous forest habitat; and (3) edge effects within the forest that is immediately adjacent to pastures. In fragments and edge forests, the microclimate is altered such that temperatures and wind speeds are higher, and relative humidity is lower. Higher wind speeds and drier conditions result in increased tree losses due to windthrow and maybe increased water stress. The drier conditions, in concert with increased quantities of dead wood, increase the susceptibility of fragmented and edge forests to wildfires. Ignition...
sources for wildfire are widespread in a tropical landscape mosaic of fragmented forest and pastures because fire is frequently used to maintain the pastures. Because native flora and fauna are poorly adapted to fires in tropical rain forests, even light fires result in high rates of mortality. Low-intensity fires that spread from cattle pastures to adjacent edge forest have been found to completely kill the overstory canopy, while only \( \approx 40\% \) of the species had the capacity to sprout after burning (Kauffman, 1993).

Data from the Brazilian space agency (INPE) and the U.S. Space Agency (NASA), as well as that from other Brazilian and U.S. scientists, indicate that the Amazon Basin is experiencing the world’s highest absolute rate of forest destruction. The rate of deforestation averaged 15,000 km\(^2\) yr\(^{-1}\) from 1978 to 1988, while the rate of habitat degradation (i.e., the sum of cleared and fragmented land) was 30,000 km\(^2\) yr\(^{-1}\). The rate of deforestation has accelerated in recent years from about 11,000 km\(^2\) yr\(^{-1}\) in 1991 to just over 20,000 km\(^2\) yr\(^{-1}\) from 1995 to 1997. At the rates experienced in the late 1990s, 5479 ha of forest are lost every day; this is equivalent to about 3.8 ha destroyed every minute.

It is difficult to determine the total area of tropical rain forest that has been converted to cattle pasture. Fearnside (1993) reported that by 1991 the area of forest cleared in the Amazon Basin had reached 426,000 km\(^2\). This estimate did not take into account forest areas affected by fragmentation and edge effects. Cumulatively, areas in which biodiversity has been affected by forest clearing (primarily for cattle pasture) may be over twice that of deforestation alone (Skole and Tucker, 1993). Because about 13% of the region has now been deforested (INPE, 1998), the total area affected by deforestation and fragmentation could comprise a third of the entire Amazon Basin.

Rain forests tend to receive the most attention, but tropical dry forests have been deforested to an even greater degree. These are also the most abundant type of tropical forest. Few areas of intact (i.e., uncut or primary) tropical dry forest still exist in Latin America today (e.g., the Brazilian caatinga and tropical deciduous forests of Central America and Mexico).

Although scientists and economists have questioned both the sustainability and economic viability of cattle pastures in areas of former tropical forest, the deleterious consequences to biological diversity are evident. If cattle pasture production is neither sustainable nor economically viable, then why are rates of deforestation in the Americas not a response to international demands for beef nor a response to increasing human populations (hood-and-mouth disease is prevalent in Brazilian cattle, which precludes importation into the United States, and the Amazon Basin supports only \( \approx 3\% \) of all cattle production in Brazil). Deforestation largely occurs because it is a way to obtain title to the land, or to profit from land speculation, government financial incentives, and subsidies. In addition, labor costs on ranches are quite low compared to intensive agriculture or agroforestry systems. The environmental limitations in many tropical landscapes, coupled with governmental and social forces, combine to promote a land use that produces little food and little direct monetary return, but tremendous environmental degradation (Hecht, 1983).

D. Livestock Grazing and the Greenhouse Effect

Livestock production and related land uses are significant sources of many greenhouse gases. In addition, changes in climate due to the greenhouse effect are predicted to have strong feedbacks on the environments that provide feed and water for livestock. Among the most significant influences of the livestock industry in relation to the greenhouse effect is slash burning, which occurs during the conversion of tropical rain forest to pasture. When slashed tropical rain forest burns, as much as one-half of the aboveground carbon (76–112 Mg ha\(^{-1}\)) may be released as CO\(_2\), CO, and other radiatively active gases. As much as 800–1600 kg nitrogen ha\(^{-1}\) are also released by burning forest (Guild et al., 1998). When tropical dry forest is slashed and burned, the proportion of biomass consumed is even greater—as much as 95% of the aboveground carbon pool may be released as greenhouse gases during slash fires. The burning of the African savannas, most often initiated by herders for forage enhancement, is another important source of CO\(_2\). Savanna fires contribute as much as 16% of the total annual CO\(_2\) arising from agricultural sources. However, much of this will be absorbed the following year by regrowing grasses. The same is not true for forest that is converted to pasture; the biomass of pasture grasses will be about 3% of the forests that it replaced.

Methane is a powerful greenhouse gas with a radiative absorption capacity that is 21 times that of CO\(_2\). The global increase in methane results from human activities such as livestock production, manure management, rice production, landfills, and the production and use of oil, gas, and coal. Livestock and manure management contribute about 16% of the total annual methane production (about 87 million Mg out of 550...
In addition to the effects previously mentioned, there are many other livestock-related land use activities and industries that affect global biological diversity (see Table 1). Grazing on native grasslands and shrublands supports a dwindling share of the livestock forage supply of the world. As intensive livestock production has expanded and the importance of grazing native rangeland has declined, there have been large conversions of rangelands to seemingly more lucrative land uses. The highly diverse Brazilian cerrado, a mosaic of savanna, grasslands, and open evergreen woodlands, covers about 2 million km² in Brazil. The cerrado is undergoing a rapid conversion from species-rich savanna/woodland to planted pastures, soybeans, and other annual crops. This region has experienced more land cover change than the Amazon forests (i.e., about 600,000 km² of the cerrado had been cleared as of 1991, compared to about 400,000 km² for the Brazilian Amazon; Klink et al., 1994).

Many rangelands of the U.S. Southwest and southern Great Plains were plowed into croplands when government subsidies made it cost-effective to utilize groundwater for irrigation (i.e., the Ogallala aquifer). Often, these water sources were rapidly depleted and groundwater became prohibitively costly to pump. This resulted in the abandonment of lands that are now in a desertified state—a depauperate cover of native vegetation coupled with high rates of soil erosion.

2. Exotic Species and Pathogens

Invasions of exotic species are among the greatest threats to global biodiversity and livestock play significant direct and indirect roles in facilitating their establishment. Livestock influence exotic plant distribution through seed dispersal in fur and dung. Establishment of exotic species is enhanced via creation of seedbeds through trampling and forage removal. In those regions where large herbivores were not a strong selective pressure on the native flora (i.e., much of the Pacific Northwest, Intermountain West, and California, in the United States), exotic annuals now dominate areas formerly dominated by perennial bunchgrasses. Indirectly, the livestock industries also facilitated introductions and spread of exotic plant species through cropland plantings (fodder and feed grains) that contained weed seeds.

In contrast to unanticipated shifts of native ecosystems to exotic species dominance, many native rangelands and forests have been converted to exotic grass dominance through purposeful seeding. The goals of these so-called “range improvements” were to increase forage for livestock, decrease soil erosion, and provide a desired alternative to weeds. However, there was scant consideration for influences on biological diversity. For example, in the United States millions of hectares formerly occupied by shrub/bunchgrass ecosystems have been seeded with grasses of exotic origin (e.g., crested wheatgrass, Agropyron cristatum, and Lehmann lovegrass, Eragrostis lehmanniana).

3. Livestock, Water Developments, and Aquatic Ecosystems

Livestock and related management activities have strong effects on the biodiversity of aquatic ecosystems. These not only include direct effects such as consumption of streamside or wetland vegetation, trampling of streambanks, and fecal inputs, but also indirect effects such as irrigation withdrawals, water developments, and wetland draining. These activities can alter aquatic diversity through degradation of aquatic habitats as well as water quality (temperature, chemistry, and microbial composition). Numerous studies have documented accelerated streambank erosion and losses of streambank structure due to livestock grazing. The changes and the accrual of sediment in the channel degrade spawning and reproductive habitats for fishes and aquatic insects. Overgrazing results in the simplification of stream channels, which may include loss of channel sinuosity, increases in channel widths, increased channel incision, and decreases in deep pools. Incised and simplified channels result in the elimination of important linkages between the floodplain and stream channel that positively influence biodiversity.

Water diversions for pasture or forage crop irrigation are a widespread influence on the biodiversity of rivers and their associated riparian zones. Diversions
range from the de-watering of small streams for the irrigation of a single pasture or farm field to huge dams that span major rivers. Regardless of the scale, when water is diverted or dammed, the natural hydroperiod of the river or stream is altered (i.e., changes in the timing and magnitude of peak and low flows throughout the year). This can have dramatic influences on channel-forming processes and riparian community development. Because many aquatic and riparian organisms are adapted to cycles of floods and flood effects, these hydrological alterations can completely alter their environment, leading to species losses. Throughout the western United States, livestock-related agriculture is a dominant user of irrigation water stored in federally-constructed dams. Significant detrimental effects on native fishes, as well as on biologically diverse riparian plant and animal communities, occur both upstream and downstream of the dams or diversions.

Land rehabilitation and forage enhancement efforts have been attempted throughout degraded rangelands of the world. However, these have often had disappointing results. Because of misinterpretations of ecosystem needs or ignorance of ecological consequences, rehabilitation efforts have sometimes exacerbated the deterioration and degradation that they were intended to reverse. A striking example is the establishment of permanent water sources in dry season grazing areas in Africa (similar developments have also been implemented in the U.S. West). Water sources were developed in areas so distant from surface water that they were not traditionally used by livestock during the dry seasons. Following developments such as the drilling of wells or construction of small catchment basins, lands around these African water sources were often severely overgrazed and trampled. In addition to losses in biodiversity, these water sources have fostered the growth of herds beyond the carrying capacity of the land, resulting in range deterioration and livestock deaths during droughts. During droughts, livestock did not die from lack of water but from starvation. The ultimate consequences include degradation of the native ecosystem coupled with increased human suffering and hardship.

Water pollution is a common problem associated with extensive livestock production in relatively confined spaces. Manure is a valuable organic fertilizer and soil builder in modest amounts, but it is a dangerous environmental hazard when waste production exceeds the absorptive capacity of land and water. In fecal-polluted waterways, pathogens detrimental to both the native biota and humans are present. In addition, increased nitrogen and phosphorus concentrations can result in the eutrophication of waterways, in which algae blooms rapidly consume oxygen to the point where fish kills occur.

4. Biogeochemical Cycling

Livestock grazing can affect nutrient cycling in numerous ways. The most obvious would arise from the dramatic effects associated with accelerated erosion, trampling of soils, and the replacement of perennial plants (and often biological soil crusts) with annuals or shrubs and bare ground. The direct consumption of vegetation changes the patterns and distribution of litter, which affects decomposition and nutrient cycling. Consumption of herbage resulting in less litter also increases the proportion of bare ground, thereby increasing susceptibility to erosion. Accelerated erosion results in the loss of nutrients, soil organic matter, and the capacity of the soil to store water and retain nutrients. In riparian zones, diminished litterfall can also affect the aquatic biota. For many headwater streams, the nutrients and energy that drive in-stream aquatic ecosystems are derived from terrestrial inputs. It is particularly deleterious to aquatic ecosystems when long-term grazing results in the elimination of shrub and tree overstories in streamside environments.

The synergistic effects of forage removal and trampling damage by livestock have been described in Oregon riparian meadows that were grazed for over a century. Grazed meadows had lower rates of nitrogen mineralization and lower soil organic matter than areas where grazing had been halted for 9–12 years. These indicators of decreased productivity were associated with dramatically lower water infiltration rates, higher soil bulk density, lower levels of residual litter, and lower root biomass in the grazed sites.

Desertification often results in the alteration of natural desert grasslands and savannas into shrublands with bare soil in the intershrub spaces. Under these scenarios, nutrient and water resources become concentrated under shrubs. Livestock grazing can also concentrate nutrients in areas where livestock may congregate through the concentration of urine and feces (i.e., resting places, water holes, or streambeds).

5. Attempts to Control Undesirable Species

Other activities associated with livestock management that affect biodiversity include predator control (e.g., eliminating carnivores via trapping, poisoning, or shooting) and other forms of wildlife control for species that are perceived as competitors with the livestock industry. These activities, called “animal damage control,” are often carried out without consideration of their effects on biodiversity, ecological processes, or...
ecosystem functions. Techniques of animal damage control are often not specific to the "target species" and other "nontarget" species can be eliminated as well. Poison bait traps and many pesticides often kill nontarget animals (and insects). The black-footed ferret (Muscari nigripes) was almost driven to extinction because of the decline of the prairie dog (Cynomys spp.), its major prey. Prairie dog populations have precipitously fallen due to habitat loss by agricultural development and from purposeful elimination because they eat the same forage as livestock. Many large carnivores have been eliminated in regions because of their real or perceived threats to livestock. The livestock industry interest groups provide the most vehement opposition to the reintroductions of large carnivores to extirpated ranges (e.g., wolves to Yellowstone National Park).

As with predator control, the use of petrochemicals (herbicides and pesticides) to control unwanted species of plants and insects has sometimes resulted in undesirable effects on non-target species. In animal, plant, and insect control programs, scant attention has been paid to the functional role that these organisms may play in ecosystems and possible feedback responses to their removal. In addition, questions regarding the underlying reasons for the increase in pest or weed damage frequently were not asked or left unanswered. In the western United States, practices such as mechanical and chemical removal of shrubs and trees were implemented for the benefit of increased cattle forage, but at the expense of other human and ecological values. For example, widespread rangeland herbicide applications eliminated traditional food and medicinal plants used by indigenous peoples on many western U.S. rangelands. This is because herbicides applied to control unwanted plant species also kill a suite of other native broad-leaved species. Species with limited capacities for reinvansion may be locally extirpated from sites after a single herbicide application. Similarly, insecticides may lack selectivity, such that a suite of species performing a multitude of ecosystem functions are eliminated. Many of these chemicals act as environmental estrogens and their elevated concentrations within secondary consumers (predators and scavengers) have resulted in reproductive and physiological maladies.

6. Altering Fire Regimes

In many grasslands and savannas of the world, fire is a dominant disturbance feature of the landscape. The high productivity of flammable grasses, climatic conditions with long dry periods, and a prevalent ignition source (lightning or humans) result in the frequent occurrence of fires. Fires in grasslands and savannas are important in cycling nutrients and in influencing biotic structure and composition. Both the flora and fauna of grasslands and savannas are adapted to, and often dependent on, fire for their continued existence. Altering the fire regime can lead to a vastly different structure and composition. Livestock alter fire regimes by removing the fuels (i.e., the grasses) that carry the fire in rangelands and savannas. For example, throughout the U.S. West, livestock grazing has resulted in the diminution of fires in rangelands with a concomitant increase in shrub species that are not as fire-adapted. Increases in the abundance of juniper (Juniperus spp.) and mesquite (Prosopis spp.), as well as declines in aspen (Populus tremuloides), are related to the synergism of livestock and the decline of fire occurrence. Long-term changes in ecosystem structure due to grazing and fire suppression have resulted in a scenario in which simple livestock removal would not completely remedy the ecosystem decline; vegetation manipulations coupled with the reintroduction of fire would also be necessary components of ecological restoration.

III. CONCLUSIONS: HUMAN INSTITUTIONS AND LIVESTOCK-CAUSED DEGRADATION

Livestock create an array of environmental problems, not because cows, sheep, goats, and other grazers are hazards in themselves, but because human institutions have forced animal farming out of alignment with the ecosystems in which they are practiced. The root causes of land degradation, including deforestation and desertification, lay not in the animals, but in the social and political institutions and population pressures placed upon the environment. Sociopolitical causations must be addressed if degradation of biodiversity by livestock is to be halted and ecosystems restored.

In many countries there persists government-based subsidies and economic incentives that result in the expansion of livestock production that accelerates environmental degradation. Government policies on pricing, taxing, and land titling incentives affect resource use by influencing decisions about type of use, inputs, technology adoption, and investments in development. Policies have misguided livestock development through the subsidized pricing of inputs and products that induced the non-sustainable use and degradation of natural resources. Resources have been priced too low to reflect their true environmental costs, thus passing the


I. Types of Recombination
II. Basic Features of Reciprocal Homologous Recombination (Crossing-Over)
III. Genetic Control and Mechanisms of Recombination
IV. Biodiversity of Recombination Systems
V. Evolutionary Effects of Recombination
VI. Evolution of Recombination
VII. Applied Aspects

GLOSSARY

crossing-over Complex interaction of homologs within bivalents at pachytene resulting in reciprocal exchange of genetic material between non-sister chromatids.

chiasma A-like configurations.
gene conversion A process of nonreciprocal (unidirectional) transfer of genetic information.
synaptonemal complex A three-part proteinaceous structure that affects the number and distribution of crossovers and converts crossovers into functional chiasmata.

THE GENERAL NOTION “RECOMBINATION” includes a range of genetic phenomena. The common dominator of various recombination processes is reassociation of pieces of genetic information resulting in the formation of new combinations that differ from the parental ones. Systematic recombination studies initiated by T. Morgan at the beginning of the twentieth century developed into one of the most fruitful branches of classical and modern genetics. The main types of genetic recombination include recombination of whole chromosomes (the basis of Mendel’s law of independent assortment of genetic factors), reciprocal exchange between homologous chromosomes (crossing-over), non-reciprocal exchange (gene conversion), site-specific recombination occurring between homologous sequences, illegitimate recombination that involves nonhomologous sequences, transformation, and transduction. Genetics has accumulated abundant evidence on common features and peculiarities of these processes in various organisms, their genetic control, and molecular mechanisms. In fact, the achieved understanding is a product of combined efforts of genetics, cytology, molecular biology, biochemistry, population genetics, and evolutionary biology. It appears that recombination plays a key role in a remarkably broad spectrum of genetic processes related to DNA repair, sexual reproduction, immunity, genetic diversity, adaptation and speciation, evolution of genes, gene families, chromosomes, and the entire genome. Recombination is the basis for traditional breeding, genetic mapping, recombinant DNA technology and genetic engineering, gene targeting and gene therapy, developmental biology, and...
cancer genetics. However, despite a long history and thousands of studies, recombination continues to be a puzzle with respect to its mechanisms, diversity of genetic and evolutionary effects, and factors determining its evolution. The importance of recombination as a crucial factor determining the dynamic balance between the stability and flexibility of genetic organization as well as the interplay between heredity and environment justifies the keen interest in this process within the “biodiversity” paradigm. As such, in demonstrating recombination-related phenomena I use examples from a wide range of species in addition to traditional Drosophila, fungi, and humans.

I. TYPES OF RECOMBINATION

There are various ways to classify different types of recombination. No matter how logical a classification system may seem, the simplest way is to compare alternative aspects (modes) of the general phenomenon of recombination. These may be based on types of participating cells (meiotic or mitotic and male or female), cell compartments (nuclear or mitochondrial), chromosomes and DNA sequences (homologous or nonhomologous), and types of exchanges (reciprocal or nonreciprocal).

A. Crossing-Over and Chiasmata

Two types of cell division are known for eukaryotic organisms: mitotic and meiotic. Recombination takes place in both. In cells that enter meiosis, the chromosomes have already undergone DNA replication so that each chromosome consists of two sister chromatids connected by a shared centromere. During the meiotic prophase, pairing of homologs brings about bivalent formation; each bivalent combines four chromatids. Complex interaction of homologs within bivalents at pachytene results in reciprocal exchange of genetic material between non-sister chromatids, a process referred to as crossing-over (Fig. 1). Observation of corresponding meiotic stages (diplotene–diakinesis) reveals characteristic X-like configurations called chiasmata. These configurations are interpreted as cytological manifestation of crossing-over, although their precise 1:1 correspondence is still debated (Nilsson et al., 1993; Tease and Jones, 1995). Chiasmata were observed and interpreted as a result of an unknown process of rejoining genetic material at the end of nineteenth century, a decade before the discovery of partial linkage in plants and two decades before Morgan’s studies on crossing-over on Drosophila.

B. Nonhomologous Chromosome Segregation

After crossing-over occurs, chromosomes of each bivalent segregate independently relative to chromosomes of other bivalents in two consequent meiotic divisions. These two processes, crossing-over and recombination of nonhomologs, are characteristic of all sexual eukaryotes. With independent segregation, the ratio of parental and nonparental combinations of corresponding alleles is 1:1, although a departure from independent assortment of unlinked genes, termed quasi-linkage, was observed in both plants and animals, especially in interspecific hybrids (Korol et al., 1994).

C. Mitotic Recombination

Exchange of genetic information can also occur during mitotic cell division, but it does so with a frequency of a few orders of magnitude lower than that of meiotic ones. The major exclusion from this rule is mitotic recombination resulting in hyper-mutation of somatic cells of the vertebrate immune system. Mitotic recombination may occur in the germline cells, e.g., in Drosophila males in which meiotic recombination is normally suppressed. Some genetic factors increase the rate of mitotic recombination manyfold. For example, homozygosity for an inserted mobile element may result in
FIGURE 2 Segmental homozygosity resulting from mitotic crossing-over. (a) A diploid heterozygote for a marker locus (M/m); (b) after DNA reduplication each homolog is represented by two sister chromatids (the position of mitotic crossing-over between nonsister chromatids is marked by $\times$); (c) the structure of the homologs after crossing-over; and (d) the daughter cells resulting from mitosis are homozygous for the entire segment distal to the point of exchange.

male (premeiotic) recombination rates in Drosophila comparable to that in female meiosis (Sved et al., 1991). Mitotic crossing-over between the centromere and any locus results in two daughter cells homozygous for the entire segment distal to the site of exchange (Fig. 2). This may have a strong negative impact if the initial cell was heterozygous for some detrimental alleles (see Section VII).

D. Nonreciprocal Exchange

Normally, meiotic and mitotic crossovers are based on reciprocal physical exchange of genetic material between the parental chromosomes. Studies of recombination in fungi (Sordaria, Neurospora, and Asco- bolus) have revealed a process of nonreciprocal (unidirectional) transfer of genetic information called gene conversion that accompanies reciprocal exchange (crossing-over). Gene conversion is based on DNA repair of mismatched strands along the "recombination tract." The demonstration of this process was due to the possibility of classifying marker segregation (spore color) of ordered spores in asci resulting from individual meiotic cells in hybrids between colored and wild-type strains. In addition to the expected regular 4:4 (colored: noncolored) octads, aberrant ratios 2:6, 6:2, 3:5, and 5:3 are also observed, reflecting nonreciprocal exchange (see Section III).

E. Ectopic Recombination

Crossing-over occurring between homologous sequences of homologous chromosomes generates new genetic variation without changing genome structure. However, crossing-over may also involve homologous sequences of nonhomologous parts of the genome. This process is called ectopic recombination. In particular, numerous families of repeated elements that are spread in eukaryotic genomes may participate in ectopic recombination. Especially active can be retrotransposons, either from nonhomologous chromosomes or from different sites of the same chromosome (giving rise to inter- and intrachromosomal ectopic recombination), resulting in chromosome rearrangements. In fact, the frequency of such events is very low and it is unclear which mechanisms reduce this danger. One possibility may be a strong bias of recombination interaction in favor of gene conversion if such ectopic contacts occur. Ectopic recombination may also occur between such partners as chromosome and a plasmid, such as a two-micron plasmid in yeast harboring the gene for site-specific recombinase Flp and the tumor-inducing T-DNA segment of the Ti plasmid of Agrobacterium tumefaciens interacting with the host plant chromosome.

F. Illegitimate Recombination and Horizontal Transfer

Unlike homologous and site-specific ectopic recombination, illegitimate recombination is a process of joining DNA molecules that have only a small sequence homology. Illegitimate recombination is common in nature, especially in interactions between retroelements with host genome DNA and in some host–parasite systems. One of the most common examples is the previously mentioned interaction between bacteria, such as A. tumefaciens, and their host plants mediated by the bacterial plasmid transformation. Although this process can be considered as a "physiological" rather than a genetic transfer, it may also have evolutionary consequences as one of mediators of horizontal gene transfer. Illegitimate recombination is involved in many situations of foreign DNA integration in direct transformation experiments aimed at obtaining genetically engineered (transgenic) organisms.

G. Extrachromosomal Recombination, Exchanges between Cell Compartments

Recombination was proven to affect the organization of mitochondrial DNA. Convincing evidence of wide-
spread recombination in mitochondrial genomes has been obtained for plants. In fact, a specific substrate for frequent mitochondrial recombination was discovered and sequenced in a few plant species. Especially interesting from the evolutionary viewpoint are putative exchanges between different cell compartments: nuclear–mitochondrial, nuclear–chloroplast, and mitochondrial–chloroplast. Many such sequences demonstrating “chloroplast → mitochondrial” transfer were obtained for cereals.

II. BASIC FEATURES OF RECIPROCAL HOMOLOGOUS RECOMBINATION (CROSSING-OVER)

A. Recombination Rate and Map Distance

The average number of crossovers in a given segment represents its genetic length. Correspondingly, genetic distance between two loci can be defined as the average number of recombination events occurring in the segment. Therefore, for loci flanking a unit length segment one exchange per meiosis is expected on average, resulting in 50% recombinant gametes. This notion of genetic distance serves as a basis for gene mapping. Two fundamental facts make possible genetic mapping: (i) linear organization of genetic material in chromosomes and (ii) increasing recombination rate between loci with their physical distance. Map distance is of additive nature: For any subdivision of a segment, the segment length is the sum of lengths of its subintervals. A limitation of map distance as a measure is that it cannot be estimated directly: An odd number of exchanges in a segment marked by two loci leads to the appearance, in the progeny, of the same recombinant genotypes as those resulting from one exchange, whereas an even number results in no observable recombinants. Recombination frequencies (RFs) estimated from the observed proportions of recombinant and nonrecombinant individuals coincide with map distances only for short segments when multiple exchanges can be ignored. Therefore, transformation of experimentally estimated RF values into map distances (x) is needed. Relations of the form \( RF = RF(x) \) are referred to as mapping functions.

B. Genetic Interference and Mapping Functions

Multiple exchanges are relatively rare among eukaryotes. These vary between 1 and 3 for the majority of higher organisms, with 5 or 6 occurring seldomly, although values up to 12 have been reported (Korol et al., 1994). In Drosophila the frequency of tetrads with four or more crossovers is extremely low (usually much less than 1%).

1. Interference: Its Measurement and Basic Properties

Consider two adjacent marked segments \( m_1–m_2 \) and \( m_2–m_3 \) with recombination rates \( RF_1 \) and \( RF_2 \). If exchanges in adjacent segments occur independently, the probability of a double exchange within \( m_1–m_3 \) can be determined as \( RF_{12} = RF_1 RF_2 \). Under the same assumption, recombination frequency \( RF_3 \) between loci \( m_1 \) and \( m_3 \) is given by

\[
RF_3 = RF_1 (1 - RF_3) + RF_2 (1 - RF_3) = RF_1 + RF_2 - 2RF_3
\]

Much evidence indicates that the observed frequency of double crossovers usually differs from the expected one, a phenomenon termed genetic interference. The degree of interference is measured by the coefficient of coincidence (c):

\[
\epsilon = \frac{RF_{12}(observed)}{RF_{12}(expected)} = \frac{RF_1 + RF_2 - 2RF_{12}}{RF_{12}}
\]

The existence of interference was shown both genetically and cytologically. It is ubiquitous in eukaryotes. To date, only in three species (\( Shizosaccharomyces pombe \), \( Aspergillus\) \( \text{nidulans} \), and \( \text{Ascobolus immersus} \)) has recombination been found to proceed without interference (Egel-Mitani et al., 1982). If the occurrence of crossing-over in one segment reduces the probability of exchange in an adjacent segment (\( c < 1 \)), then it is positive interference; when one exchange increases the frequency of another (\( c > 1 \)), it is negative interference. Two types of interference have been distinguished: chiasma interference, in which an exchange occurring in a segment affects the probability of another one occurring in an adjacent segment, and chromatid interference, which determines the pattern of distribution between successive exchanges along a chromosome. The amount of interference within a segment depends on its distance from the centromere. Near the centromere, interference...
between exchanges in the adjacent segments of the same chromosome arm is maximal. It is generally believed that no interference exists across the centromere, although many exclusions have been registered.

2. Mapping Functions
Positive interference \( (c < 1) \) is a widespread phenomenon. It was found that in many cases the approximation \( c(RF) = 2RF \) (Kosambi interference) fits the data well (i.e., \( c \approx 0 \) at small distances and \( c \approx 1 \) at large distances). The corresponding mapping function

\[
RF = 0.5 \tanh(2x) \text{ or } x = 0.25 \ln\left(1 + 2RF/(1 - 2RF)\right)
\]

appeared to be a good approximation for Drosophila, rice, mouse, and other eukaryotes. The formula for RF values in adjacent intervals, corresponding to Kosambi's function, takes the following form:

\[
RF_i = (RF_i + RF_j)/(1 + 4RF_iRF_j)
\]

Situations with no interference can be approximated by the mapping function proposed by J. Haldane:

\[
RF(\chi) = \frac{1}{1 - \exp(-2\chi)}
\]

Some results indicate that negative interference (i.e., an excess of double exchanges over the level expected with independence assumption), may also occur. Such data were observed for Drosophila and some plants, fungi, and viruses. One should take into account that values of \( c > 1 \) can also result from data pooling, owing to heterogeneity of individual RF values.

C. Distribution of Crossovers within Eukaryotic Chromosomes
The efficiency of crossing-over strongly depends on between- and within-chromosome patterns of exchange distribution in the nucleus. Here, two different (albeit related) problems are generally distinguished: (1) crossover distribution in relation to major components of the chromosome, i.e., the centromere, telomeres, heterochromatin (intensely stained cytological segments of chromosomes—gene-poor regions), and euchromatin (faintly stained gene-rich regions), and (ii) relative positions of crossovers with respect to one another. These questions were thoroughly studied based on scoring recombination between multiple markers complemented by cytological observations. A wide range of spatial variation in exchange frequency was revealed—from long stretches of heterochromatin in which crossing-over is almost blocked to short segments containing recombination hot spots (see Section III.E). Heterochromatic blocks affect the distribution of crossovers in the entire genome. Recombination nonuniformity of the chromosomes was first established in Drosophila and then confirmed in a wide range of organisms. Thus, in the B genome of wheat the physical length of DNA per exchange varies more than 130-fold along the chromosome (Łukaszewsky and Curtis, 1993). In the centromeric region of the human X chromosome the recombination rate is at least eight-fold lower than the average rate of female recombination on the X chromosome. The centromere also seems to inhibit gene conversion.

D. The Effects of Sex
By the Haldane–Huxley rule, if crossing-over is strongly suppressed or lacking in meiosis of one of the sexes, then that sex is invariably the heterogametic one. The effect of sex on RF has been established in a variety of animal species, with this rule holding in all cases of alternative (all-or-none type) sex differences. Provided crossing-over occurs normally during the formation of both male and female gametes, then RF in the heterogametic sex may be either lower or higher than that in the homogametic one, with the sign and magnitude of sex difference in RF being segment specific. Sex differences in recombination have also been revealed
by cytological analysis. Recent years have seen rapid accumulation of data on human recombination owing to the use of molecular markers. The total length of female maps is nearly twice as long as that from male maps; in some regions (mainly subtelomeric ones), the opposite tendency is observed. Most of the maize genomic regions examined exhibit higher male recombination or no difference; the extent of the difference (if it exists) may strongly depend on the genotype. A substantially higher crossing-over rate in male compared to female meiosis has been found in Arabidopsis (Korol et al., 1994). This is also true for some animals exhibiting a higher degree of linkage and/or more localized pattern of exchange distribution in the homogametic sex than in the heterogametic one, such as marsupials (Trivers, 1988).

The mechanisms responsible for differences in RF between male and female meiosis are unknown. In most animals, sex is determined by a certain chromosome. Thus, it can be suggested that the difference between male and female meiosis is controlled by genes located on this chromosome. Alternatively, sex chromosome can canalize the development of general physiological (biochemical) sex differences (e.g., in the hormonal status) in a toggle-like manner, with specific recombination features of male and female meiosis being a secondary effect of the main regulator. The latter suggestion is supported by data from the fish Oryzias latipes in which the heterogametic sex is represented by XY males (Yamamoto, 1961). When males XY were transformed into functional XY females by a hormonal treatment, RF increased by a factor of 5 compared to the level in normal males. In humans, it was found that the distribution of sex difference in RF is related to genomic imprinting manifested in region-specific differential modification (methylation) of parental alleles.

E. The Effects of Environmental Conditions

The environment can affect recombination in two ways. First, it can affect recombination indirectly when changes in the frequency of alleles at recombination controlling loci depend, through a feedback mechanism, on changes in the frequency of the selected (fitness-related) loci (see Section VI). Second, it can affect recombination directly when in response to environmental factors the organism modifies (within genetically determined limits) the value of RF, i.e., when not RF but the norm of reaction with respect to RF is genetically determined. The fact of direct influence of the environment on recombination has long been known in genetics. As early as 1917, H. Plough observed in Drosophila a U-shaped dependence of RF on temperature: Recombination was minimum at optimum temperature and it increased with the departure of temperature from the optimum. Similar data indicating an increase in recombination level with deteriorating environment have been reported in other works (Korol et al., 1994). Extreme conditions also tend to reduce interference. Moreover, variation of recombination parameters in response to stressful factors can be modulated by the organism's fitness: Higher fitness provides a better preservation of gene complexes formed in previous generations. Experimental evidence for the last principle has been obtained for Drosophila and tomato (Korol et al., 1994). In Drosophila, the increase in male recombination that is induced by a heat treatment is negatively correlated with flies' resistance to high temperatures. Results reported by Laurençon and co-authors (1997) indicate that this may also apply to DNA repair and transposition of mobile elements. Therefore, recombination is not absolutely random relative to an individual's adaptation to environmental stresses affecting recombination rate and distribution. The foregoing feedback mechanism may assist the selection process in preserving adaptive gene combinations, thereby reducing the recombination load.

III. GENETIC CONTROL AND MECHANISMS OF RECOMBINATION

A. Selection for Changed Recombination, Genetic Modifiers of Recombination

Selection for altered recombination proved an important tool in analyzing genetic control of recombination. Such experiments have been conducted on dozens of organisms, including Drosophila, flour beetle Tribolium castaneum, grasshopper Schistocerca gregaria, silkworm Bombyx mori, lima bean, and fungi Neurospora and Schizophyllum (Korol et al., 1994). These experiments complemented the numerous findings on the existence of a considerable amount of genetic variation in RF in natural and laboratory populations (see Section IV.B) available for selection to act on. The effectiveness of directional selection for altered RF has been shown to depend on the segment under study, the size, structure, and origin of the start population; the breeding system; the estimation procedure; and the intensity and duration of selection. Genetic analysis of the accumulated differences between selected lines enabled the question of the genetic basis of variation in recombination to be addressed. In as few as 10 generations, divergent selection for RF in the p-1 region of silkworm chromo-
some 2 succeeded to produce lines with RF = 37–39 and 5–7%, starting from 25.6%. The obtained evidence suggests that recombination (frequency and distribution) in eukaryotic genomes is under complex control of polymorphic modifiers with small to moderate effect. The rate of recombination in a given region may depend on both linked modifiers and genes located on other chromosomes.

B. Meiotic Mutants as an Analytical Tool in Recombination Studies: Overlapping of DNA Repair, Recombination, and Segregation Systems

The first point mutation affecting recombination was Drosophila gene c(3)G (discovered in 1922), which almost completely blocks crossing-over in females. During subsequent years, many other genes with strong effect on different meiotic steps and recombination were isolated in eukaryotes (mainly in fungi, Drosophila, maize, and recently Arabidopsis and the nematode Caenorhabditis elegans). The availability of diverse mutants affecting the coordinated steps of meiosis and recombination allowed a detailed cytogenetic characterization of meiosis. This includes commitment to meiosis, regulation of major steps in chromosome conjugation and formation of the synaptonemal complex (see Section III.C), DNA breaks formation and strand exchange, chiasma maintenance, and chromosome disjunction. The broad network of recombination–meiosis mutants identified in yeasts and other fungi and in Drosophila serves as a basis for cloning recombination genes and studying the molecular mechanisms of recombination.

Numerous results were obtained on this rich material showing a deep overlapping in pathways related to DNA recombination and DNA repair. Thus, mutations for loci controlling meiosis and recombination also tend to exhibit disturbed repair functions. Such a duality seems to be an ancient phenomenon. In prokaryotes, the best example is the RecA protein of Escherichia coli that is involved in early stages of recombination and, simultaneously, in induced SOS repair. Recombination is also strongly affected by mutations in genes controlling other enzymatic steps in DNA metabolism, e.g., replication. Genes for DNA repair and recombination exhibit a high degree of sequence, structural, or functional similarity throughout life. This conservation was established using a combination of genetic and cytological methods with molecular cloning, sequencing, and genetic transformation. Thus, defects in the recombination–repair system of one species may be complemented by a gene transferred from another species, sometimes a distant one (e.g., from a prokaryote to eukaryote). However, this does not exclude high interspecific differences in many important features of recombination (Korol et al., 1994).

Comparison of recombination in meiotic mutants and in normal genotypes provides an interesting conclusion. It appears that genomic distribution of exchanges in normal meiosis is more restricted, less random, than in meioses altered by mutations, indicating that these restrictions are largely a result of evolutionary adjustment of recombination. Meiotic mutants tend to lessen the restrictions, and regions normally excluded from exchanges become involved in crossing-over. Simultaneously, a reduction of interference is also observed despite a general tightening of linkage. In meiotic mutants, the correspondence between physical and recombination length of chromosome segments becomes closer. In many cases, mei mutations referred to as hyper-rec can significantly enhance crossing-over (and not only in specific regions) or increase the rate of multiple exchanges, intragenic recombination, and/or meiotic crossing-over many-fold.

The meiotic system manifests a close connection between homologous conjugation, crossing-over, and regulation of chromosome segregation. This association is invariable for normal meiosis and served as a basis for the view that regulation of chromosome segregation is one of the main functions of crossing-over. Meiotic mutants were isolated where the two phenomena, segregation and recombination, are not coupled. For instance, crossing-over is normally almost completely suppressed in chromosome 4 of D. melanogaster. Some mei mutations with reduced recombination and/or altered exchange distribution within large chromosomes exhibit crossing-over within chromosome 4 and disturbed segregation of this chromosome (Sandler and Szauter, 1978).

C. Ultrastructure of Recombination: Synaptonemal Complex and Recombination Nodules

Electron microscopy reveals meiosis-specific organization of chromosome associations at the meiotic prophase I. A three-part proteinaceous structure, called synaptonemal complex (SC), is initiated at the zygonote stage, maintains the paired homologs together during pachytene, and disappears at the diploneme. Short stretches of DNA of the homologs are involved in the central part of SC, where intimate molecular pairing takes place. SCs are considered as structures that affect the number and distribution of crossovers and convert
crossovers into functional chiasmata. Small electrondense bodies, recombination modules (RN), were detected within the central part of the SC. Early RNs that appear before pachytene are more or less evenly distributed along the bivalent and are supposed to participate in synopsis. Late, or pachytene, RNs seem to represent recombination enzymes associated with exchanging DNA molecules. The distribution of late RNs corresponds well to chiasma distribution and is believed to represent the sites for recombination. Numerous meiotic mutants were tested with respect to their ultrastructural and molecular phenotype—the morphology and biochemistry of SC. An interesting fact is that only in three organisms was meiotic recombination normally found without SC (S. pombe, A. nidulans, and A. immer- sus). However, in many cases SC was found in achiasmatic meiosis (with no recombination). In the three fungi with recombination occurring without SC, the frequency of exchanges per bivalent is high, and especially interesting is the absence of interference (Egel-Mitani et al., 1982). Thus, the absence of SC may allow for more reshuffling of the homologs. Analysis of SC appeared to be an important tool in determining the causes of altered fertility in farm mammals.

D. Molecular Mechanisms of Recombination

When considering molecular mechanisms of recombination in eukaryotes, a broader spectrum of basic processes involved may also be addressed, including chromosome conjugation, recombination, and segregation. This is partly due to the prevailing experimental methodology based on analyzing cytogenetic and molecular effects of numerous mutants defective for meiosis—recombination and DNA repair. Historically, the studies of molecular mechanisms of recombination in prokaryotes served as the main source of models to be tested in eukaryotes. The basic enzymatic activities involved in prokaryote recombination include those of endonucleases, strand transfer, DNA unwinding, ligation, and resolving Holliday’s heteroduplex structure. The best known recombination signal is the octamer 5’GCTGGTGG-3’ of E. coli referred to as the mei- W68 gene and yeast Spo1I (McKim and Hayashi-Hagihara, 1998). Eukaryotic analog were found and (are to be) characterized in the best studied eukaryotes, such as yeast, Drosophila, lily, and mammals. Clearly, eukaryotic recombination systems are supposed to be more complex due to additional problems related to meiosis, such as homologous pairing and segregation.

The classical concept of molecular recombination included the following steps: (i) intimate pairing of homologs, (ii) formation of single-stand DNA nicks at homologous sites and local denaturation of double-stranded DNA, (iii) formation of recombination intermediate (Holliday junction) by reciprocal strand transfer and renaturation, its migration from the initiation site, and heteroduplex formation (with local mismatched DNA due to heterozygosity of the parental chromosomes); and (iv) resolution of the Holliday structure, either after or before inactivation. In the first case, the content from both sides of the considered regions remains unchanged, whereas in the second case it results in reciprocal exchange. In both cases, the mismatched positions should undergo correction. Depending on the strand used as a template in the repair process, two alternatives will be observed: recovering the initial state and nonreciprocal information transfer (gene conversion).

A new understanding of eukaryotic recombination mechanisms resulted from the discovery of the leading role of double-strand DNA breaks (DSBs) in yeast Saccharomyces cerevisiae. It was demonstrated that meiotic recombination is closely associated with the repair of DSBs. Details on different pathways related to this process are being studied in yeast, mammalian cells, Drosophila, and plants based mainly on recombination—repair mutants. Another strategy is to study specially engineered genetic models with a modified recombination system (Shalev et al., 1999). Homology of recombination genes in one species to genes controlling DSB-dependent recombination in yeast may be indicative for a shared DSB mechanism of recombination initiation, as exemplified by the Drosophila min-1 gene and yeast Spo1I (McKim and Hayashi-Hagihara, 1998).

E. Recombination Signals, Hot Spots of Recombination, and Transcription

Although recombination occurs all over the genome, the accumulated data points to the existence of hot spots and cold spots of recombination. In certain cases, highly specific interactions were found between recombination—repair mutations and “signal” DNA sequences resulting in a sharp increase in RF within defined genomic segments. The best known recombination signal is the octamer 5’GCTGGTGG-3’ of E. coli referred to as the mei-w68 element. An example of a short signal acting as a meiotic recombination hot spot is the M26 heptamer 5’ATGACGT-3’ of S. pombe. Likewise, in eukaryotes, mobile elements may act as recombination enhancers. Thus, with identical locations on homologous chromosomes, a pair of P elements of D. melanogaster can induce recombination at a frequency of 20% in males in which recombination is normally absent (Sved et al., 1991). This effect serves as a basis for an efficient method of fine genetic mapping, in which P elements are used to promote recombination at their insertion
sites. Hot spots may represent specific signals such as χ or M26; certain genes or target sequences within these genes, or more extended chromosomal segments such as gene-rich regions. In the maize \textit{bronz}\textit{e} locus, recombination rate per physical length is 100-fold higher than the average level in the genome (Dooner and Martinez-Ferez, 1997). A few recombination hot spots have been characterized in mammals, with the major histocompatibility complex being one of the best known examples.

In the 1960s and 1970s, it was speculated that initiation of recombination is attached to transcription promoters (Whitehouse, 1966). In moving from lower to higher organisms, the recombination density over the genome is reduced by three or four orders of magnitude with no marked trend in structural genes. Thuriaux (1977) attributed this paradox to preferential localization of recombination in structural genes. It was suggested that the requirement for common specific changes in chromatin organization may cause an overlapping in control of transcription, recombination, repair and mutation (Whitehouse, 1966; Thuriaux, 1977; Korol et al., 1994; Laurencos et al., 1997). The evidence from many eukaryotes indicates that regions with high RF correspond to gene-rich regions (Lichten and Gold- man, 1999; Gill et al., 1996), although an opposite trend was found in \textit{Caenorhabditis elegans}. The results obtained during the past few years mainly for yeast (S. cerevisiae and \textit{S. pombe}) strongly support the idea of a possible recombination–transcription relationship (Ni- colas, 1998). In general, the available information shows high intragenomic heterogeneity and nonran- domness of recombination (Korol et al., 1994). These are caused by a hierarchical control system that involves interaction of a complex web of recombination–repair genes with peculiarities of macro- and micro-organization of the genetic material (Table I). The important role of gene distribution in this regulation indicates its adaptive role in the evolution of genetic systems.

### IV. BIODIVERSITY OF RECOMBINATION SYSTEMS

#### A. Macroevolutionary Trends in Genome-Wise Recombination Rate

A considerable (by several orders of magnitude) decrease in RF per unit physical length of DNA is observed when moving from prokaryotes and lower eukaryotes to higher organisms (Thuriaux, 1977). Even among eukaryotes, the density of recombination events within the genome varies widely (Table II). The genetic length of the genome varies less and strongly depends on haploid chromosome number. Higher “regularity” of recombination due to frequent outcrossing, the obligatory occurrence of at least one exchange per bivalent, and increasing genome size are considered to determine the evolutionary importance of recombination in higher sexual eukaryotes. In lower organisms, recombination was supposed to play a less important role due to a slight chance of sexual contacts between individuals (Maynard Smith et al., 1991). Consequently, natural populations of such species should manifest a clonal structure, with strong correlation [linkage disequilib- rium (LD)] between alleles at different loci. However, this may not be the case in some protozoan species, such as malaria \textit{Plasmodium falciparum} (Conway et al., 1999). The revealed population pattern of \textit{P. falciparum} includes high polymorphism and a rapid decay of LD with distance (reaching independence between poly- morphic sites of 0.3–1.0 kb) that looks like that of higher sexually reproducing eukaryotes.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>Hierarchical Control of Recombination Events*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System effects</strong></td>
<td><strong>Hierarchical Control of Recombination Events</strong></td>
</tr>
<tr>
<td>Control of the nucleus as a whole, manifested in the classical interchromosomal effect of arrangements, as well as effects of cis- and aneuploidy and supernumerary (B) chromosomes</td>
<td>Hierarchical Control of Recombination Events</td>
</tr>
<tr>
<td>Effect of chromosome size</td>
<td></td>
</tr>
<tr>
<td>Regulation based on cytoskeletal interactions</td>
<td></td>
</tr>
<tr>
<td>Sex differences in recombination frequency and distribution</td>
<td></td>
</tr>
<tr>
<td>Effect of environmental conditions (stress) and age</td>
<td></td>
</tr>
<tr>
<td>Position of euchromatin (as main target for recombination) relative to the centromere, telomeres, and heterochromatic blocks</td>
<td></td>
</tr>
<tr>
<td>DNA sequence organization of the target region</td>
<td></td>
</tr>
<tr>
<td>Heterogeneity for micro- and macroreversions, deletions, and translocations</td>
<td></td>
</tr>
<tr>
<td>Microsite organization of DNA sequences including distribution of specific regulatory sequences such as recombination hot spots and micro- and minisatellites</td>
<td></td>
</tr>
<tr>
<td>Distribution of epigenetically modified (e.g., methylated) DNA stretches</td>
<td></td>
</tr>
<tr>
<td>Effects of mobile genetic elements</td>
<td></td>
</tr>
<tr>
<td>Distribution of gene-rich islands</td>
<td></td>
</tr>
<tr>
<td>Genes controlling DNA metabolism</td>
<td></td>
</tr>
<tr>
<td>Genomic-, chromosome-, and segment-specific effects of major recombination genes of the “coarse control” system affecting the basic steps in recombination mechanisms (pairing, DNA strand exchanges, and recombination repair)</td>
<td></td>
</tr>
<tr>
<td>Segment-specific regulation of crossover frequency by genes of the “fine control” system with relatively small effects of individual components</td>
<td></td>
</tr>
<tr>
<td>Genes encoding active transposases</td>
<td></td>
</tr>
</tbody>
</table>

* Interaction of factors both within and between the groups plays an important role in observed patterns of recombination frequency and distribution within the genome and its variation within and between species.
TABLE II
Recombination Characteristics of Eukaryotic Genomes

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome length (cM)</th>
<th>Haploid number (n)</th>
<th>Chromosome length (cM)</th>
<th>kb cM ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnaporthe grisea</td>
<td>600</td>
<td>7</td>
<td>85</td>
<td>80</td>
</tr>
<tr>
<td>Pyricularia oryzae</td>
<td>1700</td>
<td>12</td>
<td>140</td>
<td>230</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>600</td>
<td>5</td>
<td>120</td>
<td>200</td>
</tr>
<tr>
<td>Pinus taeda</td>
<td>1900</td>
<td>10</td>
<td>190</td>
<td>280</td>
</tr>
<tr>
<td>Common bean</td>
<td>1200</td>
<td>11</td>
<td>110</td>
<td>530</td>
</tr>
<tr>
<td>Phaseolus vulgaris</td>
<td>1200</td>
<td>7</td>
<td>120</td>
<td>4000</td>
</tr>
<tr>
<td>Lycopersicum esculentum</td>
<td>1400</td>
<td>12</td>
<td>120</td>
<td>900</td>
</tr>
<tr>
<td>Fruitfly</td>
<td>300</td>
<td>4</td>
<td>75</td>
<td>570</td>
</tr>
<tr>
<td>Salivary gland</td>
<td>900</td>
<td>28</td>
<td>30</td>
<td>300</td>
</tr>
<tr>
<td>Hordeum spontaneum</td>
<td>600</td>
<td>10</td>
<td>60</td>
<td>350</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>5000</td>
<td>26</td>
<td>150</td>
<td>50</td>
</tr>
<tr>
<td>Zebralish</td>
<td>2900</td>
<td>25</td>
<td>120</td>
<td>900</td>
</tr>
<tr>
<td>D. ananassae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicken</td>
<td>3000</td>
<td>28</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td>1600</td>
<td>20</td>
<td>80</td>
<td>2000</td>
</tr>
<tr>
<td>Cat</td>
<td>3300</td>
<td>19</td>
<td>110</td>
<td></td>
</tr>
</tbody>
</table>

B. Intraspecific Variation for the Rate of Recombination

The analysis of genetic variation in recombination parameters allows a deeper insight into patterns, mechanisms, and the evolutionary role of recombination. The genetic differences in crossing-over rate seem to have been originally established by Sturtevant in 1913, who discovered modifiers sharply reducing crossing-over during oogenesis in heterozygous females of *Drosophila*. Later, he suggested these to be inversions, a prediction that was subsequently confirmed by cytological analysis of salivary gland chromosomes. Among point mutations affecting recombination, the first to be discovered was c(3)G, which almost completely blocks crossing-over in *Drosophila* females. Additional studies demonstrated high genetic variation in RF in *D. melanogaster* and other *Drosophila* species (Korol et al., 1994; see Section III.B). Recombination studies in *D. ananassae* males revealed high variation in crossing-over rate in the second and third chromosomes; individual RF values in the ml-ru segment of chromosome 3 varied between 2.2 and 42.8% (i.e., 20-fold). Differences between maize plants in RF in the sh-su-wx segment increased as the degree of relatedness decreased and there was correspondence between micro- and macrosporogenesis. Polymorphism for B chromosomes and heterochromatic knobs and their interaction play a significant role in the variation of RF in maize. B chromosomes affect pairing and crossing-over in many plants and animals.

Abundant evidence for intraspecific genetic variation in crossing-over rate and/or chiasma frequency is available for many eukaryotes, both lower (*Coprinus, Schizopyllum, Neurospora, Saccaromyces,* and *Saccharomyces*) and higher (pearl millet, pea, jute, ryegrass, barley, wheat, tomato, snails, silkworm, grasshoppers, locusts, flour beetle, lizards, mouse, rat, cattle, and human) (Korol et al., 1994). This variation may be of primary importance in better understanding the genetic basis of biodiversity in natural systems, optimizing genetic conservation strategies, designing experiments on genetic mapping of complex traits, and map-based cloning.

C. Recombination and Domestication Evolution

Two questions are of major interest when discussing the interface domestication–recombination: (i) How have recombination features of the progenitors affected the pattern of selection response and selection advance? and (ii) How has selection for domestication affected the recombination system? Despite many theoretical studies on the interaction between selection and recombination, the very limited available evidence concerns only a few organisms out of a few hundreds domesticated plant and animal species. Burt and Bell (1987)
studied the relationship between male chiasma frequency and features of reproductive strategy, longevity, body size, etc. of more than 30 mammalian species. They found that domesticated species exhibit higher rates of exchanges than might be expected based on a general relationship among these parameters. The authors attributed this difference to the advantage of increased recombination that facilitated to overcome unwanted correlations between characters in the course of selection. At the stage of primitive selection, recombination was a major factor of genetic variation allowing for a deep and very fast (on the evolutionary scale) reorganization of domesticated organisms. Consequently, this selection pressure could become a driving force of evolution of increased recombination, at least during the first phase of domestication (thousands of generations). However, at the second phase, after regular agriculture had become established, with more stable environmental conditions, stabilizing selection had to play an ever-increasing role. This might reverse the direction of recombination evolution, at least in some species. For example, Williams et al. (1995) established a higher chiasma frequency in primitive forms of maize compared to modern industrial lines.

Polyploid formation is considered a unique evolutionary pathway exploited by both animal and plant organisms. A few of the most important crops, such as wheat, cotton, or potato, are allopolyploids. The recombination–domestication interface may have very interesting aspects when related to polyploids. These include the evolution of diploid-like chromosome behavior in meiotic pairing, recombination and segregation, the effect of various forms of recombination on the coevolution and sequence polymorphism of genomes sharing one nucleus, and the effect of selection on the foregoing processes. Detailed mapping and DNA sequencing efforts recently initiated on the foregoing crops combined with comparative evidence on their diploid relatives and/or ancestors may be considered an important way of highlighting the role of recombination in evolution and domestication.

V. EVOLUTIONARY EFFECTS OF RECOMBINATION

A. Recombination and Genetic Variation for Adaptation to Varying Environment

Producing variation is considered one of the major evolutionary functions of recombination that is posed to assist in adaptation to both spatial and temporal environmental changes. Abundant genetic variation manifested by natural population is a result of a complex interaction between a few factors: mutation, selection, population subdivision, breeding system, and recombination in its various forms. All other things being equal, genetic variance for a sexual population with free recombination may be several times that for a population with no recombination (Charlesworth, 1993). Classical estimates from Drosophila species indicated that up to 25–40% of variation in fitness in natural populations is regenerated by crossing-over in one or two generations from a randomly drawn chromosome pair. Thus, a temporary arrest of the mutation process would not result in a significant decrease in variation over many generations. Such a view corroborates recent studies showing the presence of polymorphism in small populations in nature and recovery of variation after severe bottlenecks (Nevo et al., 1997).

An important question is how recombination affects genetic variation. In a population at linkage equilibrium (LD = 0), changes in recombination rate do not affect the amount of variation. However, this does not mean that the level of variation is independent of RF. Indeed, the stability of polymorphism varies with RF. The general trend is a reduction of polymorphism stability with increased recombination. This will occur when one models the dynamics of a diploid population subjected to stabilizing or directional selection with all selected loci linked within a block. Stability of polymorphism for these loci inversely depends on recombination within the block. However, the effect of linkage may be opposite for configurations with multiple blocks of selected loci linked within a block. Stability of polymorphism for these loci inversely depends on recombination within the block. However, the effect of linkage may be opposite for configurations with multiple blocks of selected loci (Nevo et al., 1997). Another example is selection (either diploid or haploid) in a fluctuating environment (Korol et al., 1996). The effect of recombination on the amount of variation may also result from a dependence of the steady-state LD on RF: LD ≠ 0 is achievable (if at all) only at tight linkage and/or strong selection.

B. Recombination as a Source of Evolutionary Novelties

Through gene duplications (by unequal exchanges), recombination creates prerequisites for increasing variability. In fact, it is a general belief that the evolutionary formation of a new gene by mutation–selection interaction is preceded by the duplication of the genetic material. Local duplications may result from
unequal crossing-over, unequal sister chromatid exchange, polymerase slippage, or replicative transposition. There is also evidence (from both pro- and eukaryotes) suggesting that structurally (and in many cases functionally) related genes are linked to one another. These may include tandemly repeated gene families, gene clusters (with arbitrary spacing and orientation of the individual genes), and dispersed distribution of the family members across the genome. Different mechanisms were proposed for allele diversification after duplication. For example, a new allele may be produced by intrachromosomal homologous recombination mediated by DNA sequences within the duplicated elements. The creative role of different forms of recombination (unequal crossing-over, gene conversion, reciprocal exchange, and ectopic and illegitimate recombination) is well documented for different organisms.

Recombination was recognized to play an important role in the evolution of gene families that are usually organized in the genome as arrays of tandem repeats residing in one or a few chromosomes. Coordinated change of individual elements of such families is referred to as concerted evolution. The diversification of the elements by mutations is opposed to homogenization processes caused by unequal crossing-over and gene conversion. This process is opposite the scenario of multigene families, selection presumably favors the most homogenous arrays. Different variants of homogenized arrays may coexist within a population, as shown on the rDNA cluster in Drosophila. This implies a low recombination between different arrays combined with high intrachromosomal recombination within arrays.

**C. Recombination and Sequence Polymorphism**

During the past decade, new evidence became available on the association between recombination and DNA sequence polymorphism in natural populations. The first group of data came from Drosophila: Begun and Aquadro (1992) compared variation in RF per physical DNA length in different regions of D. melanogaster genome with polymorphism of DNA sequences of these regions. Positive correlation between recombination and sequence polymorphism was established. Similar results were obtained in different Drosophila species, humans, and some plants. Recombination seems to play an important role in sequence variation of natural populations for the mitochondrial genome.

Two opposite though not mutually exclusive hypotheses were proposed as explanations for the association between recombination and sequence polymorphism. The first, referred to as selective sweeping, is based on selection of new favorable mutations. In genomic regions with a low recombination rate, the process of fixation of positively selected mutations will result in fixation of closely linked neutral or even slightly deleterious sequence variations. The second explanation is based on the theory of background selection (Nordborg et al., 1996) that considers the consequences of purifying selection against deleterious mutations. This process can also reduce sequence variation at closely linked sequences. Combination of these two mechanisms seems to explain the main pattern of polymorphism–recombination association, although the remaining difficulties may point to the importance of additional factors.

**D. Horizontal Gene Transfer**

Numerous mechanisms exist that normally prevent the exchange of genetic material among distant species. Nevertheless, if this process is possible even at a low frequency, it may have important consequences on the long-term evolutionary scale. Only 20 years ago such a possibility was considered an unrealistic one. Now, rich evidence is available showing that the opposite is likely to be true. Illegitimate recombination seems to be the main responsible mechanism. Extensive similarities have been revealed between house-keeping genes of Archaea and Eubacteria, although for the majority of informational genes (involved in transcription and translation) the difference is very large. An interesting fact is that some of the foregoing genes showing high similarity between the kingdoms are clustered in the genome that may result from horizontal transfer events.

If horizontal transfer is indeed an evolutionary significant form of genetic recombination, one could consider ecological interaction between species a replacement of sexual interaction (vertical transfer). In other words, ecologically interacting species are the most probable exchanging partners. Viral DNA integration in the host nuclear genome is considered a relatively frequent event (on the evolutionary scale) for animal hosts and recently also for plants. Likewise, there is evidence of an opposite, host → parasite flow of genetic information. Many transfers of signaling domains of
pathogenesis-related plant genes were detected in bacterial genomes.

VI. EVOLUTION OF RECOMBINATION

A. Theoretical Models of Recombination Evolution

Artificial selection experiments suggest that almost every population has enough stored genetic variability to ensure response to direct selection for changed recombination frequency (see Sections III.A and IV.B). The observed polymorphism at loci affecting recombination could be either balanced (selected) or transient. Theoretical analysis shows that under a stable environment a panmictic population should evolve toward a minimum possible level of recombination. This can be formulated in terms of the fate of a selectively neutral modifier locus affecting recombination.

Understanding the forces maintaining sex and recombination is considered one of the most challenging problems in evolutionary theory (Michod and Levin, 1998). Shared genetic control and molecular mechanisms of DNA recombination and DNA repair across life (see Sections III.B and III.D) indicate their common evolutionary origin and functional overlap in extant organisms. It could probably be supposed that repair functions played the leading role at early evolutionary stages, having provided opportunities for a large increase in the genome size and the transition from haploidy to diplody. The latter offered the possibility of recombination repair of two-strand DNA lesions that is impossible in haploid systems. Some authors hypothesize that the subsequent stages in the evolution of recombination and sex were associated with repair alone. These explanations of repair functions were called "physiological" (Maynard Smith, 1978). Another physiological explanation of recombination in sexuals is its association with chromosome segregation (see Section III.B), although numerous examples are known, such as male Drosophila and female silkworm, in which normal segregation is associated with achiasmatic (without crossing-over) meiosis.

Combinative, or generative, hypotheses consider the main function of sex and recombination in shuffling genes. This removes negative correlation between favorable alleles at different loci, thereby increasing the efficiency of natural selection. Generative models can be classified according to the source of linkage disequilibria between selected loci: stochastic or deterministic, caused by new mutations or variation of external conditions (Kondrashov, 1993). In the 1930s, Fisher and Muller proposed that sex may be advantageous by combining beneficial mutations randomly occurring in different individuals (Otto and Barton, 1997). A complementary version of the stochastic–mutation explanation considers the role of recombination in selection against deleterious mutations (Muller, 1944). According to Muller, deleterious mutations tend to be fixed in a finite asexual population due to random drift despite purifying selection (Muller's ratchet), whereas recombination helps to stop this process. In the 1980s, a few deterministic models of selection against deleterious alleles were proposed, with the evolutionary advantage of recombination being dependent on linkage disequilibria resulting from synergistic interaction between harmful mutations produced at a high rate (Kondrashov, 1988). Other models of deterministic sources of linkage disequilibria include (i) adaptation to temporarily or spatially varying environment, (ii) antagonistic species interaction (the Red Queen hypothesis), and (iii) intraspecific competition (sib competition or tangled bank hypothesis) (Hamilton et al., 1990; Kondrashov, 1993; Feldman et al., 1997; Barton and Charlesworth, 1998; Otto and Michalakis, 1998). In most of these models, the conditions favoring increased recombination are associated with negative linkage disequilibria among selected loci owing to stringent conditions for epistasis. Some mechanisms (e.g., selection of beneficial mutations) may favor recombination in the absence of epistasis, with linkage disequilibrium caused by finite population size (Otto and Barton, 1997).

B. Testing the Theoretical Assumptions

Despite voluminous literature, the problem has been studied little experimentally or by observations in natural populations. In the classical period of genetics it was proposed that sex and recombination may facilitate selection for advantageous and against deleterious mutations and assist in adaptation to changing environment. The major question is whether these processes in turn are able to promote evolution of the recombination system. Are the resulting changed patterns of recombination detectable experimentally? Is it possible to find corresponding changes in nature? Theoretical models answer positively to these questions. However, do these models (and/or their parameter values) fit the real world? Unfortunately, there is not enough evidence to answer this question. Two different and complementary approaches can be considered (Barton and Charlesworth, 1998; Korol, 1999).
1. Testing the Initial Assumptions

The aim is to test whether the real estimates of the main parameters underlying the proposed mechanism(s) fit the expectations. These include mutation rate, effective population size, mode of interaction of deleterious or beneficial mutations, relative role of stabilizing and directional selection, and cost of sex and meiosis. The major advantage of this approach is that it allows the accumulation of data on basic parameters affecting the evolution of recombination. The drawback is that it is difficult to believe that even the main factors were taken into account so that the predicted direction of recombination evolution is determined correctly. As a recent example of a successful application of this approach, competition experiments with yeast *S. cerevisiae* (Greig *et al.*, 1998) are worth mentioning. Using sexual and asexual strains it was found, contrary to the expectations following from the concept of costs of sex or meiosis, that sex can confer significant selective advantage to its carriers. Direct tests have also been conducted with the nematode *C. elegans* demonstrating the possible role of recombination in purifying selection (Zetka *et al.*, 1987). It was shown that increased RF owing to enhancer Rec-1 results in a higher fitness of genotypes carrying mutator gene mut-6 compared to analogous genotypes with a normal recombination rate. Therefore, the presence of a gene for increased recombination, which has no effect on mutation, reduces the genetic load caused by the mutator system. However, recent studies of normal (wild-type) genotypes of *C. elegans* give very low estimates of mutation rate, which is inconsistent with the hypothesis of selection against deleterious mutations.

2. Searching for Changed Recombination

The second approach is to test the "final effect" by comparing rates and genomic distribution of recombination in populations subjected to contrasted selection regimes during many generations. The advantage of such an analysis is that the target effect (changed pattern of recombination) is estimated directly (see Sections VI.D and VI.E), although one cannot ensure that it is causally connected only with the presumed process.

C. Recombination System and Life History Traits

Rasmusson (1927) was possibly the first to discuss the genetic basis of putative differences in recombination systems of outbreeding and inbreeding species. On the basis of extremely high variation in the recombination rate in peas that is usually not found in *Drosophila*, he advanced a hypothesis of possible differences in the system of genetic control of recombination between outbreeders and sellers. The species origin and life conditions determine the strategy for adaptation and hence the "optimum" amount of recombination. Evidence of variation of chiasma frequency in nature and its correlation with ecological and life historical traits is available for different organisms and may be employed to compare alternative theoretical models (Burt and Bell, 1987; Koella, 1993). In mammals with low reproduction rates, longer life cycles, and small progeny sizes, recombination is higher than in species with short generation times and large progeny sizes (Burt and Bell, 1987), fitting the Red Queen model. Koella (1993) found that in plants with animal-dispersed seeds recombination is higher than in other plants, and perennials show a higher recombination than annuals when compared among genera but an opposite tendency is displayed on the species level.

D. Recombination System and Species Ecology

Attempts to correlate life history traits and the level of recombination to justify theoretical concepts have occasionally encountered serious criticism. The fact that interspecific variation usually includes a broad complex of traits does not allow clear conclusions to be reached when testing theoretical predictions. One may take advantage of intraspecific comparisons, especially when testing hypotheses of recombination maintenance due to its role in adaptation to adverse environments. In particular, it is of great interest to compare, within the same species, the rate of recombination in populations inhabiting stressful versus mild conditions. Recombination rate was evaluated in *D. melanogaster* males from the opposing slopes of ‘Evolution Canyon’ on Mount Carmel (Israel) manifesting strong microclimatic differentiation, mainly in temperature and humidity (Korol, 1999). A significant difference in RF was found in chromosome 3 in flies derived from the opposing slopes. In particular, higher (four-fold) RF was manifested by the sub-population of the more stressful south-facing slope compared to the north-facing slope. Parallel patterns have been found in Evolution Canyon in the soil fungus *Sordaria fimicola* in the rates of mutation, crossing-over, and gene conversion,
all being higher on the south-facing slope (Lamb et al., 1998).

E. Experimental Evolution of Recombination

1. Changes in Recombination Caused by Selection for Adaptive Traits

Theoretical models indicate that directional or variable selection for multilocus traits may promote evolution toward increased recombination if the selected population is polymorphic for recombination-modifying loci (Charlesworth, 1993; Korol et al., 1994; Otto and Barton, 1997). Only a few experiments seem to have a direct bearing on this question. The following are examples with regard to Drosophila (Korol 1999). First, Clegg and coworkers (1979) attempted to test Fisher’s (1930) prediction that stable conditions may favor reduced recombination. The experiment started with a large cage population of D. melanogaster with maximum linkage disequilibrium between allozyme loci of chromosome 3. No general trend in RF for 20 generations was revealed, which was explained as an indication of a low epistatic component of selection within the marked interval. Second, Flexon and Rodell (1982) established a synthetic population of D. melanogaster and subjected it to selection for DDY resistance for 300 days. By the time of estimation, resistance increased 10- to 100-fold relative to the control level. The control and treatment populations were compared with respect to recombination rate across the genome. Increased RF between marker loci on chromosomes 2 and 3 but not X was found in the selected population. Third, most experimental studies on recombination evolution have employed one-way selection for fitness traits. The limitation of this scheme is the danger of a change in RF caused by initial linkage disequilibrium between selected genes and recombination modifiers. Therefore, Korol and Iladi (1994), in an attempt to assess the effect of selection for geotaxis on recombination in D. melanogaster, employed two-way selection. Forty generations of selection for altered geotaxis resulted in an increment in recombination for the genome of 78 cM for geo and 66 cM for geo compared to non-selected control. Selection for negative geotaxis did not affect RF in chromosome 2, whereas selection in the opposite direction caused a 4-fold increase in RF in the b-cn segment spanning the centromere of chromosome 2. The observed change in RF, caused by two-way selection for geotaxis, was attributed to the advantage conferred by selection on recombinants.

2. Experimental Evolution of Recombination Caused by Adaptation to Stress

I review a series of studies with D. melanogaster conducted in my laboratory and aimed to demonstrate that population adaptation to adverse conditions may promote evolution for increased recombination (Korol et al., 1994; Korol, 1999). The experiments were performed on four large heterogeneous populations with the objective of simulating the process of population adaptation to stress and evaluating its effect on recombination. At the start, each population was subdivided randomly into two parts, control (C) and treatment (T); T variants were then subjected to daily fluctuating temperature with amplitude increasing with generation (from 23–27°C at the beginning to 12–32°C at the end of the experiment). C variants were maintained at 25°C.

a. Increased Recombination Resulting from Adaptation to Stressful Conditions

The control and treatment populations appeared to diverge genetically with respect to thermotolerance. Changes in recombination were evaluated using multiple marked lines. All four populations showed one pattern—a segment-specific increase in RF, resulting from selection for thermoadaptation. Those that reacted most were the near-centromeric regions of chromosomes 2 and 3 and the left arm of chromosome 3 (Fig. 3). The T and C variants were also compared with respect to the frequency of double exchanges (interference). Significant differences in the values of coefficients of coincidence have been found for adjacent and non-adjacent intervals of large autosomes. The general tendency was a reduced interference in T variants, corroborating the results for recombination rate.

b. Reduction of Recombination Rate under Optimal Stable Conditions

The described divergence over time with respect to recombination rate resulted from two independent trends. An increase in RF was observed in populations adapting to adverse conditions, and simultaneously tightening of linkage in control populations that were maintained under optimal temperature. The proportion of these two effects varied in different “treatment-control” pairs corresponding to the structure of the starting population. These results can be considered as the first experimental verification of Fisher’s (1930) hypothesis that in a constant environment selection favors tighter linkage. Indirectly, these results also indi-
Adaptation to daily fluctuating temperature results in a segment-specific increase in recombination rate in large autosomes of *D. melanogaster* (based on data from Korol et al., 1994).

cate that selection against harmful mutations alone seems to be insufficient for maintenance of the normal level of recombination, at least in the genetic material that was tested.

VII. APPLIED ASPECTS

A. Recombination as a Tool of Genetic Mapping

Recombination-based mapping as a method of studying the genetic material topography has been a major analytical approach in genetics for nearly 80 years. However, until recently this technique was of little applied interest because of a lack of easily scorable markers. The situation changed dramatically with the appearance of DNA markers. Establishing high-quality maps is considered one of the most important objectives in the Human Genome Project, which also includes a few model organisms. Mapping efforts allow a better understanding of the organization of genetic material and also allow one to conduct far-reaching comparative mapping. This approach is a power tool of ever-increasing importance for studies of genome structure, functioning, and evolution, and for diverse medical and commercial applications of modern genomics. Fine genetic mapping is a basis of positional cloning, an approach proved successful in many organisms. Its efficiency strongly depends on the kb : cM ratio in the region harboring the target gene. If the target region is very rarely involved in recombination, then map-based cloning is impractical. Induced recombination may be a nontraditional approach for such situations. The availability of dense maps facilitates mapping of genetic disease genes in humans with subsequent cloning, sequencing, and medical applications. An important application of mapping is genetic dissection of complex traits, or quantitative trait loci mapping, especially in the framework of new marker-assisted strategies in plant and animal breeding.

B. Recombination as a Major Source of Genetic Variation in Breeding

When viewed at the level of phenotypic traits, genetic innovations produced by recombination upon artificial hybridization can arbitrarily be divided into three main types: (i) transgression for individual traits with the range of trait values in the segregating progeny exceeding the respective parental ranges; (ii) formation of new trait combinations of the crossed components; and (iii) the appearance of new traits ("anomalous" variation) as a result of recombination in genetic complexes with strong nonallelic interactions. These forms of recombination variability play a significant role in breeding programs, providing the raw material for selection. The importance of controlling recombination increases with the necessity of involving new genetic resources for breeding purposes, especially in light of ever-decreasing homeostasis and tolerance to abiotic and biotic stresses of elite animal breeds and plant cultivars. New genes are supposed to be introgressed from exotic germplasm based on recombination. This problem is complicated by reduced recombination on distant crosses and linkage drag (close linkage between the target gene and undesirable genes). Hence, manipulating recombination may become an efficient tool of modern breeding.

C. Recombination and Cancer

There are many aspects of anomalous chromosome and cell behavior and cancer related to DNA metabolism and recombination–repair. A few are discussed here.
1. Mitotic Recombination
A surprisingly large proportion of human neoplasms appear to result from loss of function of tumor-suppressing genes, for example, due to loss of heterozygosity of tumor suppressors. In retinoblastomas caused by loss of heterozygosity, about half of the homozygosity cases appear to be the result of mitotic recombination.

2. Specificity of Chromosome Translocations
Many types of human cancers are associated with non-homologous recombination events producing specific translocations. This results in an enhanced expression of oncogenes residing in the vicinity of the breakpoint. Some of these oncogenes are transcription factors participating in normal development. It seems reasonable to assume that interaction between recombination and transcription (see Section III.E) may also involve ectopic recombination resulting in the specificity of some tumorigenic translocations (Barr, 1998).

3. Defects in DNA Repair–Recombination
A few cancer-related genetic disorders of this type have been characterized, including Bloom syndrome, which manifests mutator and hyperrecombination phenotypes; Ataxia telangiectasia mutation, which increase V(D)J-mediated chromosome rearrangement in T lymphocytes; and Omenn syndrome, a disorder caused by mutations of recombination–activation genes \textit{Rag1} and \textit{Rag2} [related to V(D)J recombination] resulting in severe immune deficiency combined with a complete lack of circulating T and B cells. One of the most important tumor-suppressing genes controlling cell apoptosis, \textit{p53}, is involved in repair–recombination processes.

D. Recombination Shuffling of Genes: A New Technology of Artificial Sequence Optimization
Recombination combined with selection has recently become a basis for a new technology of “molecular breeding” aimed at improving protein functions via directed evolution in vitro. It includes reshuffling of a few copies of the coding DNA by homologous or site-specific recombination with subsequent screening of the resulting sequences for functional (enzymatic) activity. The reshuffling is provided by template switching in self-priming polymerase chain reaction (sexual PCR). The efficiency of such artificial evolution may be enhanced by increasing the initial genetic variation involved in recombination, e.g., by mutagenesis of the target sequences based on error-prone PCR. Another possibility is to include a wider spectrum of parental sequences using the natural diversity of homologous genes of the same or related species or even distant taxa (Crameri et al., 1998). This approach is especially promising for the improvement of molecules that provide resistance against toxic agents or increase survival on minimal media. Experiments including transformation of the engineered sequences into single-cell organisms (E. coli) and selection of the transformants showed dozens-to-hundreds- and even thousands-fold increases in efficiency of the target molecules obtained after recombination–selection cycles. DNA shuffling can improve the targeted pathway by complex mechanisms, even though it may be impossible to predict the optimized sequence by a rational design.

E. Targeted Gene Replacement
Gene targeting is an experimental approach aimed at producing site-specific changes in the genome based on genetic recombination. Predetermined changes can be introduced in the target site, in both somatic and germ-line cells. Usually, the targeting system involves a gene mediating site-specific recombination and a specific target locus. The best known examples are \textit{Cre} and \textit{Flp} recombinases (of bacteriophage P1 and yeast) with \textit{loxP} and \textit{FRT} targets, respectively. The basic idea is to first transform the target sites flanking a selectable marker into the recipient genome. Subsequent introduction of the recombinase activity leads to the excision of the marker. Many important applications have been proposed and many others are expected using the powerful strategies of targeted gene replacement. These include controlled expression of transformed genes, disruption of a chosen gene in vivo for a better understanding of its role in physiological and pathological processes, disruption of genes to rechannel the biosynthesis and achieve overproduction of certain molecules, site-specific targeting of DNA into embryonic stem cells for production of gene-modified model organisms, targeted delivery of specific constructs aimed at producing anti-tumor effects or complementing a mutated gene, and testing of gene expression in donor versus recipient tissues in biomedical transplantation studies.

Acknowledgments
I thank E. Nevo, A. Kondrashov, A. Levy, and three anonymous referees for useful comments and suggestions.


REEF ECOSYSTEMS: THREATS TO THEIR BIODIVERSITY

James W. Porter and Jennifer I. Tougas
University of Georgia

I. Coral Reef Biodiversity
II. Coral Biology
III. Anthropogenic Causes of Coral Decline
IV. Coral Disease
V. Coral Reefs and Global Climate Change

GLOSSARY

biodiversity Refers to the diversity of life, including genetic biodiversity (diversity within a species), species biodiversity (diversity among species), and ecosystem biodiversity (diversity among ecosystems).

bleaching The loss of symbiotic zooxanthellae from corals. Bleaching is usually caused by elevated sea surface temperatures, but it can also be caused by sedimentation, salinity variation, or bacterial infection.

calcification The deposition of calcium carbonate skeletons by aquatic plants or animals. In reef-building corals, calcium is deposited in its aragonitic mineral form.

Cnidaria The marine invertebrate phylum containing the reef-building corals.

disease Any impairment of the normal physiological functions of an organism. While disease normally refers to infection by bacterial, fungal, protozoan, or viral pathogens, technically bleaching could also be classified as a disease based on its physiological effect.

epizootic Disease outbreaks among animal populations (as distinguished from an epidemic in human populations).

eutrophication Nutrient enrichment, typically in the form of nitrates or phosphates, most often from human sources such as agriculture, sewage, or urban runoff from land.

extinction Extinction is said to occur when a species is not definitely located in the wild during the past 30 years.

global climate change Refers to a suite of changes in the Earth’s climate, including phenomena such as global warming, severe storm frequency and intensity, and glacial melting. Increasingly, scientists believe that global climate change is being accelerated by anthropogenic inputs of CO₂.

gonochoric A mode of reproduction in which individuals of the species are either male or female and produce either eggs or sperm within a single colony.

hermaphroditic A mode of reproduction in which individuals of the species produce both eggs and sperm within a single colony, sometimes within the same polyp.

hermatypic Reef-building; more recently, this term has been replaced by the term zooxanthellate to refer to those coral species with symbiotic algae.

nematocysts Harpoon-like stinging cells found in the tentacles of all cnidarians. They are used to pierce, immobilize, and capture zooplankton food.

oligotrophic Low in nutrients and low in primary production. Coral reefs grow in oligotrophic water.

planula A coral larva. This ciliated planktonic stage...
CORAL REEFS ARE the oldest and most diverse communities on Earth. With 32 of the 34 presently known animal phyla, reef ecosystems are vastly more diverse than tropical rain forests, which support only 9 free-living phyla. There are many close analogies between coral reefs and tropical rain forests. Both exhibit high species diversity, both have high topographic complexity (trees in the rain forest, corals on the reef), and both have a high proportion of their organic material resident in the living biota rather than in organic-rich soils or sediments. However, it is probable that no other ecosystem on Earth has, or ever had, as many higher level taxa as are present on modern-day coral reefs.

To a certain extent, coral reefs are enigmas: on the one hand, they are the most luxuriant ecosystems on Earth, supporting high diversity and high biomass, and yet on the other, they achieve this status in the least fertile waters on Earth. Corals solve this problem by tight recycling and high efficiency. The flesh of corals is a symbiotic association between algae, called zooxanthellae, and cnidarians (10% plant, 90% animal). Corals are primary producers, herbivores, and carnivores all at the same time. This tightly knit symbiosis produces trophic efficiencies as high as 90%. Furthermore, filter-feeding invertebrates, which create and cover the topographically complex three-dimensional structure of the reef, capture and retain a high proportion of the material that moves over them.

From a geological perspective, reefs may be defined as masses of carbonate limestone, built up from the seafloor by the accumulation of the skeletal material of many coral reef plants and animals. For every gram of carbon dioxide fixed into organic (living) material by coral photosynthesis, an equal amount of carbon dioxide is deposited into inorganic material (limestone) by calcification. Reef growth has shaped the face of the Earth by creating limestone structures over 1.3 km thick (Enewetak Atoll) to over 2000 km long (Great Barrier Reef). Depending on their proximity to land, coral reefs are classified as either fringing reefs (paralleling the coast line at a distance of <1 km from shore), barrier reefs (paralleling the coast line >3 km from shore), or atoll reefs (mesoceanic reefs without any relationship to continental or island land masses). Reefs can be further subdivided into back-reef, patch-reef, or offshore-reef habitats.

Coral reefs flourish on stable substrates within a very narrow range of physical parameters. These requirements include shallow depths (0–50 m), normal oceanic salinities (32–38 parts per thousand), warm sea surface temperatures (mean annual values of 22–28°C), high ambient light levels (100–2000 \( \mu \text{E m}^{-2} \text{s}^{-1} \) at solar noon), high water clarity (transmittance values above 90%), high oxygen concentrations (near 90% full saturation), and extremely low nutrient concentrations (<1.0 \( \mu \text{M} \) dissolved inorganic nitrogen; <0.1 \( \mu \text{M} \) soluble reactive phosphorus). Although some coral reefs can exist under conditions slightly suboptimal to these, such reefs are never the richest, fastest growing, or most diverse. As a result of these requirements, coral reefs are restricted to the tropics, generally between 25° north and south latitude, and predominantly on the western boundaries of the world’s oceans in the Caribbean and the Indo-Pacific. Tropical coastal zones cover 9.8 × 10^6 km^2, or 1.9% of the Earth’s surface; coral reefs are thought to occupy only 0.6 × 10^6 km^2, or slightly less than 0.1% of the planet. Humans have a special responsibility and a special challenge to preserve these environments as they house the fullest expression of the evolution of life on Earth.

RARELY LASTS FOR MORE THAN 1 OR 2 WEEKS PRIOR TO SETTLEMENT.

\( P/R \) The ratio between photosynthetic and respiratory rates of the combined coral host and zooxanthellate symbiont. A ratio greater than 1 (\( P/R > 1 \)) indicates a net gain of energy that is then available for growth and reproduction.

Scleractinia The taxonomic order of cnidarians that includes the reef-building corals.

sedimentation Particulate material falling out of the water column onto the seafloor.

trophic efficiency The percentage of material or energy that moves, without loss, from one trophic level to the next. Most food chains have trophic efficiencies around 10%. Through tight internal recycling, corals routinely achieve trophic efficiencies in excess of 90%.

trophic level Position within the food chain, e.g., primary producer, herbivore, and carnivore. Corals, however, with their symbiotic algae and their ability to feed on zooplankton, exist at all three trophic levels simultaneously.

turbidity Particulate material suspended in the water column that reduces water clarity, light penetration, and hence photosynthesis.

zooxanthellae Symbiotic dinoflagellate algae in corals and other tropical marine invertebrates.

‰ The oceanographic symbol for salinity, or the salt content of seawater in parts per thousand. 

REEF ECOSYSTEMS
I. CORAL REEF BIODIVERSITY

A. Phyletic Diversity

Coral reefs harbor extraordinary biodiversity. At the phyletic level, a level that more accurately tallies the diversity of evolved life forms in an ecosystem, 32 of the 34 described phyla are found on coral reefs. In contrast, only 9 are found free-living in the tropical rain forest (Table I). Even if freshwater and parasitic forms are included in the count, the rain forest total rises to 17 phyla, approximately half of the phyletic diversity of coral reefs.

This observation raises important concerns in the conservation of biodiversity. Whereas most biologists focus on issues pertaining to species loss, geologists frequently examine extinction patterns in higher level taxa (Veron, 1995). Ninety percent of the 83 described animal classes are marine. Almost all of these are found on coral reefs, and some, such as the class Scleractinia, are exclusively tropical. If coral reef habitats worldwide become significantly degraded, then it might be reasonable for ecologists (as well as geologists) to contemplate the loss, over the next century, of some of the Earth’s higher taxa.

B. Species Diversity

The species diversity of coral reefs greatly exceeds that of any other marine environment. Of the roughly 1.86 million plant and animal species described, 274,000 are thought to be marine and more than half of these are tropical (Table II). At present, there are thought to be 93,000 described species of coral reef plants and animals. Almost 66,000 of these are macroscopic invertebrates. Specific examples of this extraordinary diversity exist in the disparate coral reef literature; a few of these remarkable numbers are listed in Table III. At present, no fully comprehensive all-taxa biodiversity inventory has ever been conducted on a coral reef (Ormond et al., 1997), but it is obvious that were this to be done, the total biodiversity would be extremely high.

As in the rain forest, estimates of coral reef species diversity based on the number of described species are considered to be a gross underestimate of the actual number of species there. Also, as in the rain forest, the tiniest members of the community (insects in the rain forest and microinvertebrates on the coral reef) are thought to be the most diverse, and least well described, component of the fauna. Reaka-Kudla (1997) has pointed out that most of the diversity and most of the biomass of coral reefs reside within the cryptofauna,
Biodiversity Patterns Suggest That, as with Terrestrial Organisms, Species Diversity among Marine Organisms Is Higher in the Tropics Than in the Temperate or Arctic Zones

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of described species (to nearest 1,000)</th>
<th>Percentage of total described species (1.87 million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed species diversity</td>
<td>1,986,000</td>
<td>—</td>
</tr>
<tr>
<td>Total marine species, all taxa</td>
<td>274,000</td>
<td>14.7</td>
</tr>
<tr>
<td>Total macroscopic marine species</td>
<td>200,000</td>
<td>10.7</td>
</tr>
<tr>
<td>Total animals</td>
<td>193,000</td>
<td>10.3</td>
</tr>
<tr>
<td>Macroinvertebrates</td>
<td>100,000</td>
<td>9.6</td>
</tr>
<tr>
<td>Total algae</td>
<td>8,000–8,000</td>
<td>0.2–0.4</td>
</tr>
<tr>
<td>Total described tropical coastal species</td>
<td>193,000</td>
<td>10.4</td>
</tr>
<tr>
<td>Total described coral reef species</td>
<td>93,000</td>
<td>5.0</td>
</tr>
<tr>
<td>Total macroscopic coral reef species</td>
<td>66,000</td>
<td>3.6</td>
</tr>
<tr>
<td>Animals</td>
<td>66,000</td>
<td>3.5</td>
</tr>
<tr>
<td>Algae</td>
<td>2,000–3,000</td>
<td>0.1–0.2</td>
</tr>
</tbody>
</table>

Expected species diversity

| Total expected coral reef species | 618,000 | 34.3 |
| Most conservative estimate       | 948,000 | —    |
| Intermediate estimate            | 9,477,000 | —  |

Despite the paucity of data on marine biodiversity, it also appears (1) that most of the biodiversity of coral reefs has not been described and (2) that many species may already have gone extinct. The data are summarized from Reaka-Kudla (1997).

Briggs (1999) has argued against such extrapolations, pointing out that statistical errors are compounded unrealistically when small sample sizes are increased by several orders of magnitude, e.g., from 93,000 observed species to 9,477,000 expected species (Table II). After an exhaustive review of the literature and advice from marine systematists, Poore and Wilson (1993) argue that only 1 in 20 marine species have been described, producing a conservative estimate for tropical marine biodiversity of 1,870,000 species. This 1.87 million estimate suggests that the number of species to be found on a coral reef equals all of the currently described life forms on our planet (Table II). If we accept as fact that tropical marine biota is almost certainly more poorly described than temperate biota, that is, that the ratio of undescribed to described species is greater than 20 to 1, then there appear to be somewhere between 1.86 and 9.47 million species on coral reefs. Regardless of the estimating technique used,
**TABLE III**

Examples of the Extraordinary Biodiversity of Coral Reefs

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of species</th>
<th>Sampling unit</th>
<th>Location</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organisms &gt; 0.2 mm (all groups)</td>
<td>309</td>
<td>Columns of the coral Oculina arbuscula</td>
<td>Florida</td>
<td>McCloskey, 1970</td>
</tr>
<tr>
<td>Scleractinian corals</td>
<td>800</td>
<td>10 m&lt;sup&gt;3&lt;/sup&gt;</td>
<td>Australia</td>
<td>Poore and Wilson, 1993</td>
</tr>
<tr>
<td>Boring sponges</td>
<td>350</td>
<td>10 m&lt;sup&gt;3&lt;/sup&gt;</td>
<td>Aldabra Atoll</td>
<td>Hughes and Gamble, 1977</td>
</tr>
<tr>
<td>Polyzoans</td>
<td>138</td>
<td>6 liters of sediment</td>
<td>Oahu, HI</td>
<td>Butman and Carlton, 1993</td>
</tr>
<tr>
<td>Star corals</td>
<td>103</td>
<td>One colony of living coral</td>
<td>Heron Island, Australia</td>
<td>Grasse, 1973</td>
</tr>
<tr>
<td>Motile cryptofauna</td>
<td>776</td>
<td>One reef flat</td>
<td>Moorea</td>
<td>Peyrot-Clausade, 1983</td>
</tr>
<tr>
<td>Mollusks</td>
<td>637</td>
<td>Milne Bay, Papua</td>
<td>New Guinea</td>
<td>Werner and Allen, 1998</td>
</tr>
<tr>
<td>Boring cryptofauna</td>
<td>220</td>
<td>Dead coral</td>
<td>Solomon Islands</td>
<td>Gibbs, 1971</td>
</tr>
<tr>
<td>Cheilostome bryozoans</td>
<td>46</td>
<td>Hard substrates</td>
<td>Jamaica</td>
<td>Jackson, 1984</td>
</tr>
<tr>
<td>Infaunal invertebrates</td>
<td>350</td>
<td>10 m&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Aldabra Atoll</td>
<td>Hughes and Gamble, 1977</td>
</tr>
<tr>
<td>Polycheates</td>
<td>158</td>
<td>6 liters of sediment</td>
<td>Oahu, HI</td>
<td>Butman and Carlton, 1993</td>
</tr>
<tr>
<td>Mollusks</td>
<td>637</td>
<td>Milne Bay, Papua</td>
<td>New Guinea</td>
<td>Werner and Allen, 1998</td>
</tr>
<tr>
<td>Boring cryptofauna</td>
<td>220</td>
<td>Dead coral</td>
<td>Solomon Islands</td>
<td>Gibbs, 1971</td>
</tr>
<tr>
<td>Cheilostome bryozoans</td>
<td>46</td>
<td>Hard substrates</td>
<td>Jamaica</td>
<td>Jackson, 1984</td>
</tr>
<tr>
<td>Infaunal invertebrates</td>
<td>350</td>
<td>10 m&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Aldabra Atoll</td>
<td>Hughes and Gamble, 1977</td>
</tr>
<tr>
<td>Polycheates</td>
<td>158</td>
<td>6 liters of sediment</td>
<td>Oahu, HI</td>
<td>Butman and Carlton, 1993</td>
</tr>
<tr>
<td>Mollusks</td>
<td>637</td>
<td>Milne Bay, Papua</td>
<td>New Guinea</td>
<td>Werner and Allen, 1998</td>
</tr>
<tr>
<td>Boring cryptofauna</td>
<td>220</td>
<td>Dead coral</td>
<td>Solomon Islands</td>
<td>Gibbs, 1971</td>
</tr>
<tr>
<td>Cheilostome bryozoans</td>
<td>46</td>
<td>Hard substrates</td>
<td>Jamaica</td>
<td>Jackson, 1984</td>
</tr>
<tr>
<td>Infaunal invertebrates</td>
<td>350</td>
<td>10 m&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Aldabra Atoll</td>
<td>Hughes and Gamble, 1977</td>
</tr>
<tr>
<td>Polycheates</td>
<td>158</td>
<td>6 liters of sediment</td>
<td>Oahu, HI</td>
<td>Butman and Carlton, 1993</td>
</tr>
<tr>
<td>Mollusks</td>
<td>637</td>
<td>Milne Bay, Papua</td>
<td>New Guinea</td>
<td>Werner and Allen, 1998</td>
</tr>
<tr>
<td>Boring cryptofauna</td>
<td>220</td>
<td>Dead coral</td>
<td>Solomon Islands</td>
<td>Gibbs, 1971</td>
</tr>
<tr>
<td>Cheilostome bryozoans</td>
<td>46</td>
<td>Hard substrates</td>
<td>Jamaica</td>
<td>Jackson, 1984</td>
</tr>
<tr>
<td>Infaunal invertebrates</td>
<td>350</td>
<td>10 m&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Aldabra Atoll</td>
<td>Hughes and Gamble, 1977</td>
</tr>
<tr>
<td>Polycheates</td>
<td>158</td>
<td>6 liters of sediment</td>
<td>Oahu, HI</td>
<td>Butman and Carlton, 1993</td>
</tr>
</tbody>
</table>

and regardless of how fully we accept Briggs’s caveat, the gathering impression is that, with the exception of species in a few showy classes and orders, the vast majority of coral reef species are as yet undescribed.

**C. Control of Scleractinian Coral Biodiversity**

Geography, age, and temperature appear to control biodiversity patterns in reef-building corals. The Indo-Pacific region, with its vastly greater age and geographic extent is richer by far than the Caribbean (Veron, 1995).

For instance, whereas 362 species of coral are found on the eastern end of Papua New Guinea, only 53 are found in Jamaica (Table III). The Indo-Pacific has a “species-generating” topography: tens of thousands of isolated islands scattered across vast spaces. Despite the recent discovery of sibling species of corals even within the relatively well known Caribbean genus Montastrea (Knowlton et al., 1992), the numerical disparity between the two regions will persist as the list of described species from both oceans lengthens.

Figure 1 shows the widespread distribution of coral genera throughout the Caribbean. The dense packing of generic diversity isopleths along the eastern coast of

---

TABLE IV
Examples of Undescribed Biodiversity among Several Tropical Marine Invertebrate Faunae from Familiar and Easily Accessible Marine Habitats (Merrell, 1995)

<table>
<thead>
<tr>
<th>Site</th>
<th>Taxon</th>
<th>Number of undescribed species out of the total collected in the taxon</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Guinea</td>
<td>Corals</td>
<td>14 of 362</td>
<td>Werner and Allen, 1998</td>
</tr>
<tr>
<td>New Guinea</td>
<td>Fish</td>
<td>3 of 1030</td>
<td>Werner and Allen, 1998</td>
</tr>
<tr>
<td>New Guinea</td>
<td>Snails, sea slugs</td>
<td>310 of 364</td>
<td>Gosliner, 1993</td>
</tr>
<tr>
<td>Philippines (one island, multiple sites)</td>
<td>Snails, sea slugs</td>
<td>153 of 320</td>
<td>Gosliner, 1993</td>
</tr>
<tr>
<td>Hawaii</td>
<td>Marine polychaete worms</td>
<td>112 of 158</td>
<td>Dutch, 1986</td>
</tr>
<tr>
<td>(one island, 6 liters of coral reef sediment)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Great Barrier Reef (two islands)</td>
<td>Marine flatworms (Polyclads)</td>
<td>123 of 134</td>
<td>Newman and Cannon, 1994</td>
</tr>
<tr>
<td>Gulf of Mexico</td>
<td>Copepods (Harpacticoids)</td>
<td>10–27 out of 20</td>
<td>Merrell, 1995</td>
</tr>
</tbody>
</table>

a This table has been arranged from larger to smaller body size and suggests that, as with fauna everywhere, especially in the tropics, the smaller the body size, the higher the percentage of undescribed species in the sample.

Florida correlates with the frequency and intensity of cold water disturbances (Birkeland, 1996). Cold temperature limits the distribution of coral reefs northward in the Northern Hemisphere and southward in the Southern Hemisphere.

From a biodiversity perspective, Bermuda (with 14 hermatypic scleractinian coral genera) may have more in common with the coral reefs of the Florida Middle Grounds (12 genera) and the Flower Garden Banks (13 genera) than with coral reefs of the Florida Keys (24 genera; Fig. 1). The absence of the family Acroporidae from all three of these northern localities is probably more a function of winter cold kills than of limits to dispersal due to geographic isolation. These three reef localities demonstrate that temperature tolerances of individual Caribbean species are probably more important than distance in determining which species are present at a given location.

Local environmental conditions can also influence coral diversity. The loss of coral diversity along the northern and eastern coast of South America is probably due to sedimentation, not temperature. The presence of extensive terrigenous beaches and sediments transported from tropical rivers, such as the Orinoco and the Amazon, diminishes coral reef survival in these locations. It would be interesting to see if the species and genera waning as one travels south along the coast of South America (where low temperature is not a factor) are the same as those disappearing as one travels north along the coast of Florida (where low temperature is the controlling factor).

D. Species Loss

Both IUCN (International Union for Conservation of Nature and Natural Resources) and CITES (Convention on the International Trade in Endangered Species) define extinction as occurring when a species is not definitely located in the wild during the past 50 years. With this strict definition, and in the complete absence of monitoring efforts at the appropriate temporal and spatial scales, extinctions in the marine environment in general, and on coral reefs in particular, are almost impossible to prove. An example of this kind of difficulty can be seen in the announcement of the extinction of an eastern Pacific coral species due to severe El Niño conditions, and an almost immediate retraction when it was subsequently rediscovered alive (Glynn and Feingold, 1992). Nevertheless, Carlton (1993), in his review of modern marine invertebrate extinctions, includes one tropical species in his list, the Indo-Pacific mangelrove periwinkle, _Littoraria flammea_, which was last seen in the mid-1800s. There are several methodological reasons why marine extinctions might be especially difficult to detect. As pointed out by Ray (1988), “The last fallen mahogany...
FIGURE 1  Patterns of generic scleractinian coral diversity in the Caribbean reveal no endemism within the region, but rather broad-scale distribution followed by rapid faunal diminution north and south. Coral loss northward along the coast of Florida is due to cold temperature limitations; faunal loss southward along the coast of South America is probably due to the influence of river sediments pouring into the coastal zone. While the northward distribution of corals in the Caribbean is due to historical patterns of global temperature regimes, the sediment load of coastal environments in South America is increasingly influenced by anthropogenic upland management practices in the coastal zone.
would lie perceptibly on the landscape, and the last black rhino would be obvious in its loneliness, but a marine species may disappear beneath the waves unobserved, and the sea would seem to roll on the same as always.” In addition, there is the perception that marine species are somehow less susceptible to extinctions. Lamarck, in his 1809 *Philosophie Zoologique*, states that “Animals living in the waters, especially the sea waters . . . are protected from the destruction of their species by man. Their multiplication is so rapid and their means of evading pursuit or traps are so great, that there is no likelihood of his being able to destroy the entire species of any of these animals.”

While this argument may carry some validity, Reaka-Kudla (1997) points out that most species on coral reefs are small and that these smaller species also have much smaller geographic ranges. This leads to the conclusion that most species on coral reefs may be much more vulnerable to extinction than has been widely assumed.

The few clear examples of marine extinctions have in common a vulnerable, extinguishable habitat. Coral reefs, especially those located near population centers, fall into this category. The most interesting perspective on the complex, worrisome, but poorly researched topic of marine extinctions may belong to Carlton (1993), “At the end of the 20th Century, one of the major crises in global marine invertebrate conservation is not so much that invertebrates are becoming extinct at a rapid rate (although they may be)—the crisis is that we do not know.”

II. CORAL BIOLOGY

A. Anatomy

Corals are benthic marine invertebrates belonging to the phylum Cnidaria, which is characterized by two distinct tissue layers, the inner endoderm and outer ectoderm, separated by an amorphous collection of cells called the mesoglea. A single coral polyp has a central mouth cavity surrounded by tentacles armed with stinging cells called nematocysts. Corals can be solitary, consisting of a single large polyp, or colonial, consisting of thousands of interconnected polyps. Colonies form through budding—one polyp produces a daughter polyp that is genetically identical to the original.

B. Reproduction and Recruitment

Corals can reproduce asexually through fragmentation or self-generation of brooded larvae. This form of reproduction restricts genetic diversity of coral populations. In contrast, sexual reproduction through fertilization of gametes originating from genetically distinct colonies increases the genetic diversity of coral populations.

Sexual reproduction in corals occurs in one of two ways: either through mass spawning, in which thousand of gametes (eggs and sperm) are released simultaneously into the water column where fertilization takes place, or by brooding, in which sperm are released into the water column and are taken inside the maternal coral polyp to fertilize the eggs stored there. Depending on the species, a given colony may be hermaphroditic, producing both eggs and sperm, or gonochoric, producing either eggs or sperm. In both instances, ciliated planulae larvae are produced (Birkeland, 1996).

The coral larva spend between 3 days and 3 weeks in the water column, during which time they disperse. They may travel only a few meters away on the same reef or to entirely different reefs kilometers away. Dispersion maintains gene flow in coral populations. After dispersion, larvae settle onto relatively clean, hard surfaces on the reefs, metamorphose into polyps, and begin to form new colonies through asexual budding (Birkeland, 1996).

Coral recruitment is favored by nutrient-poor conditions with high light availability, low sedimentation rates (Rogers, 1990), limited competition for space by algae, and decreased predation by fish, sea urchins, and starfish. The patterns of settlement, survival, and growth of coral recruits directly influence the structure and function of coral communities and associated reefs.

C. Calcification

Common to all scleractinian (stony) corals is the ability to secrete calcium carbonate. The shape of the resulting skeleton is species specific at the polyp level, but the overall shape of the colony is influenced, within limits, by environmental conditions. Colony morphologies aid in the removal of trapped sediments (Rogers, 1990) and the capture of food and influence both zooxanthellate and host physiology (Gehes, 1994).

As a chemical process, deposition of CaCO₃ is influenced by the ambient concentration of CO₂, which is directly related to temperature, pressure, and concentrations of other dissolved materials. As a biological process, calcification is driven by photosynthesis and is closely controlled by temperature (Dubinsky, 1990). Under optimal conditions, growth rates of branching corals, such as the Caribbean coral *Acropora cervicornis*, can exceed 10 cm per year. However, local variables
such as nutrient concentrations and sedimentation rates reduce realized growth rates (Birkeland, 1996).

Calcification by thousands of colonies over hundreds of thousands of years produces the complex, three-dimensional structure of modern reefs essential to the maintenance of reef biodiversity. For example, topographical features are important for the distribution, survival, and resulting abundance of many reef fishes and invertebrates (Sebens, 1994).

D. Photosynthesis

While corals can capture prey with their tentacles, many scleractinian corals rely on endosymbiotic algae for nourishment. Known as zooxanthellae, the algae are located within the ectoderm of the coral. Depending on the species, corals may host a variety of zooxanthellae within a colony through space and time (Rowan et al., 1997). Photosynthesis by the zooxanthellae provides nutrients required by the coral for growth and reproduction and drives calcification and subsequent reef formation. As a result, the bathymetric distribution of reef-building corals is largely restricted to high light environments, typically less than 50 m depth, which can sustain this symbiotic relationship (Dubinsky, 1990).

Depending on the clarity of the water, ultraviolet light penetrates the ocean to about 5 m. Ultraviolet radiation inhibits photosynthesis and is damaging to many organisms, including corals and zooxanthellae. However, some coral species have developed protective pigments that allow the transmission of visible light while blocking ultraviolet radiation (Dubinsky, 1990). As not all corals have this ability, the distribution of corals is also influenced by the presence of ultraviolet radiation.

If the relationship between the coral and its symbiotic zooxanthellae is disturbed through increased temperatures or exposure to elevated UV light, bleaching may occur. The term “bleaching” describes the condition in which the zooxanthellae exit, or are expelled from, the coral, thus showing the white skeleton beneath the coral tissue. Without the symbiotic algae, corals lose their vital source of nutrition, slow their growth rates, stop reproducing, and sometimes die (Birkeland, 1996). When environmental conditions return to normal, zooxanthellae repopulate the coral. Susceptibility to bleaching is influenced by the species of coral in question and the species of zooxanthellae it hosts (Rowan et al., 1997). Consequently, two colonies of the same species may have dramatically different bleaching responses to the same stresses.

E. Physiological Limitations

While availability of light limits the depth distribution of corals, temperature limits the latitudinal and longitudinal distribution and is one of the best predictors of coral diversity (Veron, 1995). Optimal temperature for coral growth and reproduction ranges from 22 to 26°C, depending on geographic location and species in question. Corals generally do not grow in waters in which minimum temperatures drop below 18°C, and such a thermal barrier also limits dispersal of larvae. A few corals survive in temperatures above 30°C, such as those found in some locations in the Middle East. To some extent, corals are able to adapt to ambient conditions; consequently, upper lethal temperatures for a species in the tropical zone will be higher than those of the same species in a subtropical zone (Dubinsky, 1990).

Salinities can also influence the distribution of corals. Corals grow well in water that has a constant salinity of 32–36‰. Low salinity (<20‰), due to increased freshwater flow from localized flooding or exposure to heavy rainfall during low tides, limits coral distribution and reduces diversity. High salinities (>38‰) can also inhibit coral growth, particularly in the Persian Gulf (Dubinsky, 1990).

III. ANTHROPOGENIC CAUSES OF CORAL DECLINE

A. Benefits from Coral Reefs

Humans benefit from both the resources and recreation that coral reefs provide. Coral is used for building materials in areas where there is no viable alternative. In fact, many inhabited tropical islands around the world were, at one point, coral reefs themselves. In addition, coral reefs reduce coastal erosion by protecting coastlines from severe storms (Hoegh-Guldberg, 1999). This is particularly important in tropical waters where hurricanes and tropical storms occur frequently (Richmond, 1993). Indeed, entire islands have been washed into the sea when their surrounding living coral reefs were removed.

Coral reefs are important for the development of local economies (Birkeland, 1996). The reefs support valuable fisheries for local consumption and for the aquarium trade. Throughout the Caribbean and Indo-Pacific, local diets derive nearly 60% of their intake of protein from these reefs. The life cycles of many commercially important fish and shellfish are depen-
dent on the presence of healthy mangrove swamps, coral reefs, sea grass beds, and coastal lagoons.

A multibillion dollar tourism industry is supported by tropical coral reefs (Hoegh-Guldberg, 1999), and is a critical source of income, particularly for small island nations with few alternative resources to exploit. For the 1992–1993 fiscal year, attendance at the Coral Reef State Park on Key Largo, Florida, had the highest visitation of any state park in Florida that year (Fig. 2). Tourism can only be a viable option for economic development if reefs are healthy.

Finally, pharmaceutical companies have discovered naturally occurring bioactive compounds among the organisms found on coral reefs (Birkeland, 1996): anti-tumor compounds have been found in the mucus of corals, anti-inflammatory agents have been isolated from soft corals, and coral has successfully been used as a bone substitute in reconstructive surgeries.

B. Coastal Urbanization

Despite their importance, coral reefs around the world have been declining at an alarming rate. At the core of this decline are human activities (Fig. 3), spurred by population growth (Table V). Nearly 13% of the human population lives within 100 km of coral reefs (Hoegh-Guldberg, 1999). The geographical locations of highest coral diversity also coincide with large human populations. More than 100 countries have coral reefs within their territorial boundaries (Birkeland, 1996). Most of these are developing nations and have, by far, the fastest growing populations due to advances in medicine, technology, and public health services. Coastal cities are growing rapidly by attracting immigrants from country interiors to these bustling centers of trade and commerce. Influxes of tourists can substantially increase the effective population of an area and place additional demands on potable fresh water, power, and sewage systems. For example, in 1990, the population of the Florida Keys was 80,500, but there were over 2,000,000 visitors per year to this tourist destination.

As populations continue to expand, human pressures on coral reefs will increase. There is a direct correlation between reef degradation and proximity to urban centers. Activities associated with urbanization include waste disposal and power and desalination plant operation. Rapid urbanization has outpaced sewage treatment capacities in several regions and has caused eutrophication in coastal zones as raw sewage is often discharged directly into nearshore waters (Richmond, 1993; Sebens, 1994). The effluents from operating power and desalination plants are up to 5 or 6°C warmer than ambient temperatures.

Industrialization often accompanies urbanization and is encouraged by economic demands for hard currency and international commerce. Effluents from some
Coral reefs are subject to terrestrial, atmospheric, and oceanic influences. Sediments, nutrients, and toxins, released from activities such as deforestation, agriculture, and industry, are hydrologically transported to coral reefs through local rivers. CO₂ buildup in the atmosphere increases CO₂ concentrations in the ocean and alters climate patterns. Finally, diseases are circulated by ocean currents. Reefs located near human population centers are subjected to multiple stresses simultaneously and so suffer losses in diversity and cover [adapted from Wilkinson, C. R., and Buddemeier, R. W. (1994). Global Climate Change and Coral Reefs: Implications for People and Reefs. Report of the UNEP–IOC–ASPEI–IUCN Global Task Team on the Implications of Climate Change on Coral Reefs. IUCN, Gland, Switzerland].

Industries, such as rum distilleries and fertilizer plants, contribute to coastal eutrophication and heavy metal contamination. Ores are one of the few available resources for economic development in the tropics and mining activities can be a significant source of sediments. Furthermore, heavy metals readily bind to clays and are transported by terrigenous sediments to the reefs. Mining activities throughout the Indo-Pacific and Caribbean deliver thousands of tons of toxin-laden sediments to coral reefs each year.

Another land-based activity that affects coral reefs is deforestation. Nearly 70% of all tropical hardwood products originate from Southeast Asia. In the Philippines, forests have been reduced to 25% of their original cover. Upland areas of French Polynesia are cleared for residential and commercial construction as well as for agriculture and hydroelectricity. Deforestation, particularly of coastal mangroves (Rogers, 1990), increases erosion and the amount of soils transported from the land to the reefs. Erosion can be severe when heavy rains fall on logged areas (Birkeland, 1996).

Agricultural activities often take over land cleared for timber production. Millions of hectares of mangrove forests have been reclaimed for aquaculture and agriculture. In Southeast Asia, farming has become universally dependent on the use of agrochemicals. As a result of agricultural activities, nearshore waters are subjected to increased nutrients, sediments, and agrochemicals (Richmond, 1993). Heavy metals found in corals from Panama and Costa Rica were common components of agricultural pesticides (Guzman and Jimenez, 1992).

Upland and coastal ecosystems on land are intimately linked with coral reefs in the sea (Porter et al., 1999). At the organismal level, corals subjected to land-based pollution undergo metabolic changes that lead to bleaching, reduced growth and reproduction rates, and, on occasion, death (Richmond, 1993). Coral recovery after natural disturbances is inhibited by pollution. At the ecosystem level, these effects lead to losses in coral diversity, coral cover (Edinger et al., 1998), and shifts in dominant benthic organisms (Lapointe, 1999). The causal agents include sedimentation, eutrophication, altered temperatures, and altered salinities.

C. Sedimentation

Sedimentation influences coral communities through lethal and sublethal mechanisms, depending on the sediment load and the life cycle of the marine organisms. While increased sedimentation causes direct mortality of corals by smothering them, most effects are sublethal. Corals remove sediments by secreting copious amounts of mucus that trap the sediments. These mucus sheets, which are moved off of the coral through ciliary
### TABLE V

Disturbances to Reef Ecosystems: Their Sources and Consequences

<table>
<thead>
<tr>
<th>Disturbance</th>
<th>Effect of disturbance</th>
<th>Source of disturbance</th>
<th>Cause of source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sedimentation</td>
<td>Lethal effects</td>
<td>Deleterious effects on coral</td>
<td>Human population growth</td>
</tr>
<tr>
<td>Smothering</td>
<td>Infrasound construction</td>
<td>Road construction</td>
<td>Migrations to cities</td>
</tr>
<tr>
<td>Reduced coral cover</td>
<td>Logging in the watershed</td>
<td>Economic demands</td>
<td></td>
</tr>
<tr>
<td>Sublethal effects</td>
<td>Clearing for agriculture</td>
<td>Introduction of new technologies</td>
<td></td>
</tr>
<tr>
<td>Decreased water clarity</td>
<td>Clearing mangroves</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shift toward shallower community</td>
<td>Industry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreased photosynthesis</td>
<td>Mining</td>
<td>Economic demands</td>
<td></td>
</tr>
<tr>
<td>Increased respiration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased mucus production</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced coral recruitment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eutrophication</td>
<td>Lethal effects</td>
<td>Waste management</td>
<td>Human population growth</td>
</tr>
<tr>
<td>Overgrowth by macroalgae</td>
<td>Migrations to cities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced coral cover</td>
<td>Agriculture</td>
<td>Introduction of new technologies</td>
<td></td>
</tr>
<tr>
<td>Sublethal effects</td>
<td>Fertilizer application</td>
<td>Food needs</td>
<td></td>
</tr>
<tr>
<td>Decreased water clarity</td>
<td>Ranching (raising pigs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shift toward shallower community</td>
<td>Industry</td>
<td>Introduction of new technologies</td>
<td></td>
</tr>
<tr>
<td>Decreased photosynthesis</td>
<td>Fertilizer plant operation</td>
<td>Economic demands</td>
<td></td>
</tr>
<tr>
<td>Reduced coral reproduction</td>
<td>Rum distillery operation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced coral recruitment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toxic contamination</td>
<td>Lethal effects</td>
<td>Agriculture</td>
<td>Economic demands</td>
</tr>
<tr>
<td>Heavy metals</td>
<td>Death</td>
<td>Pesticide application</td>
<td>Food needs</td>
</tr>
<tr>
<td>Pesticides</td>
<td>Increased bacterial infections</td>
<td>Herbicide application</td>
<td>Introduction of new technologies</td>
</tr>
<tr>
<td>Sublethal effects</td>
<td>Industry</td>
<td>Human population growth</td>
<td></td>
</tr>
<tr>
<td>Herbicides</td>
<td>Increased mucus production</td>
<td>Mining</td>
<td>Migrations to cities</td>
</tr>
<tr>
<td>Increased respiration rates</td>
<td>Fertilizer plant operation</td>
<td>Economic demands</td>
<td></td>
</tr>
<tr>
<td>Decreased photosynthesis</td>
<td>Power plant operation</td>
<td>Introduction of new technologies</td>
<td></td>
</tr>
<tr>
<td>Decreased growth</td>
<td>Desalination plant operation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreased reproduction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bleaching</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Altered temperatures</td>
<td>Lethal effects</td>
<td>Industry</td>
<td>Human population growth</td>
</tr>
<tr>
<td>Increases</td>
<td>Bleaching, leading to death</td>
<td>Power plant operation</td>
<td>Migrations to cities</td>
</tr>
<tr>
<td>Decreases</td>
<td>Desalination plant operation</td>
<td>Tourism</td>
<td></td>
</tr>
<tr>
<td>Sublethal effects</td>
<td>Altered hydrology</td>
<td>Fossil fuel consumption</td>
<td></td>
</tr>
<tr>
<td>Bleaching and recovery</td>
<td>Global climate change</td>
<td>Human population growth</td>
<td></td>
</tr>
<tr>
<td>Increased respiration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreased photosynthesis</td>
<td>Introduction of new technologies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced reproduction</td>
<td>Urban development</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

continues
### Disturbance Effect of disturbance Source of disturbance Cause of source

<table>
<thead>
<tr>
<th>Disturbance</th>
<th>Effect of disturbance</th>
<th>Source of disturbance</th>
<th>Cause of source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altered salinity</td>
<td>Lethal effects</td>
<td>High salinities</td>
<td>Human population growth</td>
</tr>
<tr>
<td></td>
<td>Bleaching, leading to death</td>
<td>Desalination plant operations</td>
<td>Migration to cities</td>
</tr>
<tr>
<td></td>
<td>Decreased coral cover</td>
<td>Reduction of freshwater input</td>
<td>Tourism</td>
</tr>
<tr>
<td></td>
<td>Sublethal effects</td>
<td>Low salinities</td>
<td>Introduction of new technologies</td>
</tr>
<tr>
<td></td>
<td>Bleaching and recovery</td>
<td>Increased freshwater runoff from deforestation of watersheds</td>
<td>Urban development</td>
</tr>
<tr>
<td></td>
<td>Decreased photosynthesis</td>
<td>Decreased respiration</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reduced fertilization</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decreased coral cover</td>
<td>Reduction of freshwater input</td>
<td>Tourism</td>
</tr>
<tr>
<td></td>
<td>Decreased coral diversity</td>
<td>Introduction of new technologies</td>
<td>Urban development</td>
</tr>
<tr>
<td>Disease</td>
<td>Lethal effects</td>
<td>Largely unidentified pathogens</td>
<td>Increased susceptibility to disease by multiple stressors</td>
</tr>
<tr>
<td></td>
<td>Tissue death</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sublethal effects</td>
<td>Decreased photosynthesis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decreased growth</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decreased reproduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storms</td>
<td>Lethal effects</td>
<td>Global climate change</td>
<td>Fossil fuel consumption</td>
</tr>
<tr>
<td>Increased frequency</td>
<td>Scouring</td>
<td></td>
<td>Human population growth</td>
</tr>
<tr>
<td>Increased intensity</td>
<td>Fragmentation</td>
<td>Introduction of new technologies</td>
<td>Urban development</td>
</tr>
<tr>
<td></td>
<td>Sublethal effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fragmentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Increased sedimentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Increased turbidity</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Increased nutrients</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

action, constitute a tremendous energy drain for the corals and cause a decrease in the \( P/R \) by increasing respiration (Rogers, 1990). Despite this removal process, sediments tend to accumulate in depressions on large, massive colonies and cause death to those patches. Consequently, there is a positive correlation between the amount of terrigenous sediments and the amount of coral injury.

Water turbidity, which increases when sediments are suspended in the water column, decreases the amount of light available for photosynthesis. As photosynthetic rates decrease, so do growth and reproduction rates. Because of the reduced availability of light, the maximum depth at which corals can grow decreases and the coral community compresses into shallower environments (Dulvyński, 1990).

Adult corals are more tolerant to sedimentation stresses than juveniles. Coral larvae are not able to settle on loose sediments (Rogers, 1990). Consequently, if a fine layer of sediments covers the reef benthos, then coral settlement patterns shift toward vertical surfaces (Rogers, 1990) and successful recruitment drops dramatically (Richmond, 1993).

### D. Eutrophication

The effect of eutrophication varies according to the quantity and quality of the nutrient source, as well as the hydrographic regime in the area, and becomes especially apparent when high nutrients are present for an extended period of time. On naturally oligotrophic reefs, tight nutrient cycling between the coral host and zooxanthellate symbionts affords a competitive advantage to the coral: corals are able to flourish and outcompete many other primary producers on the reef. When nutrients are added to the system, the competitive edge shifts to faster growing macroalgae (Lapointe, 1999; Richmond, 1993) and filter feeders. The algae proceed to overgrow the corals and effectively shade them until the corals die. Coral recruitment is reduced because algae occupy space on the reef and prevent coral larvae from settling (Lapointe, 1999). Finally, the growth of...
boring organisms is promoted, which weakens the reef structure itself and increases the probability of storm damage (Richmond, 1993). Eutrophication is also associated with increased turbidity and a concomitant decrease in light availability (Richmond, 1993), largely due to an increase in phytoplankton densities. When photosynthesis decreases, growth rates and reproduction diminish. This leads to decreases in coral diversity and coral cover (Birkeland, 1996). Furthermore, vertical zonation becomes truncated under decreased light availability (Dubinsky, 1990).

**E. Heavy Metals and Toxins**

Howard and Brown (1984) reviewed the effects heavy metals have on corals. Corals are able to directly absorb soluble metals from seawater. Alternatively, they may ingest metals directly by catching particulate matter in mucous nets or indirectly as a result of feeding on copepods, which accumulate metals in their chitinous skeletons. Some metals may be deposited directly into the skeleton and become immobilized. Others remain in the coral tissues and cause dramatic physiological responses. These include excessive mucous production, increased bacterial infections, bleaching, decreased skeletal deposition, which decreases vertical growth rates, decreased skeletal deposition, which decreases vertical growth rates, reduced reproduction, and death (Howard and Brown, 1984). Similarly, when corals are exposed to agrochemicals, responses include increased respiration, decreased photosynthesis, increased mucous production, increased planulae release (a common stress response for brooding species), and decreased larval settlement (Birkeland, 1996).

**F. Altered Temperatures**

Although the warmth of tropical water may seem benign to humans, reef-building corals live much closer to their upper lethal temperature (the temperature that will kill or disable them) than to their lower lethal temperature. In fact, a rise in the water temperature of only 2–3°C above the normal summertime average is much more stressful physiologically than a drop of 2–3°C below this value. There are two ways in which elevated temperatures affect coral: increased respiration and decreased photosynthesis. Under severe thermal stress, bleaching occurs. Anything that acts to increase temperatures has the potential to adversely affect the health and survival of coral reefs. Coral respiration increases with increasing temperatures (Porter et al., 1999). Bleaching under increased temperature is correlated with increased respiratory rates and a decrease in photosynthesis (Porter et al., 1999). Under both of these conditions, the $P/R$ ratio decreases for the coral, and growth and reproduction decrease. If exposed for an extended period of time to temperatures above the average maximum temperature they are accustomed to, bleaching can occur and the coral colonies can die (Birkeland, 1996).

Reproductive success decreases with increased temperatures and is far more sensitive to temperature fluctuations than growth rates. Consequently, healthy adult corals could live in environments unsuitable for reproduction. Nonetheless, as temperature-sensitive species die, or fail to reproduce, the composition of coral communities will change (Hoegh-Guldberg, 1999).

**G. Altered Salinity**

Hyposalinity results from increased discharge or runoff associated with deforestation, particularly of mangrove forests, and from urban development. On the other hand, hypersalinity is associated with power and desalination plant effluents as well as large-scale reductions of freshwater flow from land. Responses of corals to altered salinities vary according to species and region. In Florida, the coral *Siderastrea siderea* can grow in areas where salinity fluctuations are prevalent. Changes in salinity of up to 10‰ away from the mean produces little response in the coral. Beyond 10‰ above the mean, respiration and photosynthesis decreased and some bleaching was observed. In contrast, *Porites* species have demonstrated a narrower tolerance to salinity fluctuations: an increase of 10‰ causes corals to contract their polyps, shed copious amounts of mucus, and bleach (as cited in Porter et al. (1999)).

In Kaneohe Bay, Hawaii, widespread coral death has accompanied increased freshwater runoff. Low salinities also inhibit fertilization and larval survival (Richmond, 1993). Fertilization of mass-spawning species takes place at the water surface, where eggs and sperm mix. Once the eggs have been fertilized, the resultant larvae float near the water surface for several days. Freshwater also floats on seawater. Therefore, the gametes and larvae could be exposed to lowered salinities if mass spawning occurs during heavy rainfall. In one study, fertilization rates and larval survivorship dropped by more than 50% when the salinity dropped to 28‰. These results demonstrated that terrestrial runoff can have a major influence on reproductive success (Richmond, 1993).
Water from Florida Bay naturally flows through channels between the keys and out over the reef tract. Extensive channelization of water for use in Miami and agricultural areas and for flood control has decreased the amount of freshwater entering the bay. Consequently, salinities in Florida Bay rose dramatically in the 1980s, particularly during drought years. Because Florida Bay is shallow, temperatures fluctuate with the seasons. Warm, hypersaline waters originating from Florida Bay have been recorded at depth on reefs along the Florida Keys (Porter et al., 1999).

In a study of six sites along the Florida Keys reef tract, between 1984 and 1991, all six sites lost coral diversity, and five out of six sites lost coral cover. Looe Reef, the southernmost and hardest hit reef, lost 43.9% of its coral cover (Porter et al., 1999). Porter et al. (1999) suggested that the reef degradation observed could result from the influence of poor water quality originating from Florida Bay. Another potential source of stress from Florida Bay water is eutrophication (Lapointe, 1999).

### TABLE VI

<table>
<thead>
<tr>
<th>General disease category*</th>
<th>Common name</th>
<th>Pathogen</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black line disease</td>
<td>Black band</td>
<td>Phormidium corallyticum</td>
<td>Rutzler and Santavy, 1983</td>
</tr>
<tr>
<td></td>
<td>White pus</td>
<td>Unknown</td>
<td>Porter et al., in press</td>
</tr>
<tr>
<td>White line diseases</td>
<td>White band, Type II</td>
<td>Unknown</td>
<td>Ritchie and Smith, 1998</td>
</tr>
<tr>
<td></td>
<td>White plague, Type I</td>
<td>Unknown</td>
<td>Richardson et al., 1986b</td>
</tr>
<tr>
<td></td>
<td>White plague, Type II</td>
<td>Unknown</td>
<td>Richardson et al., 1990a</td>
</tr>
<tr>
<td>Other diseases</td>
<td>Yellow blush</td>
<td>Unknown</td>
<td>Santavy et al., 1999</td>
</tr>
<tr>
<td></td>
<td>Dark spot</td>
<td>Unknown</td>
<td>Goran et al., 1998</td>
</tr>
<tr>
<td></td>
<td>Ridge mortality</td>
<td>Unknown</td>
<td>Goran et al., 1998</td>
</tr>
<tr>
<td></td>
<td>Red band</td>
<td>Oscillatoria?</td>
<td>Goran et al., 1998</td>
</tr>
<tr>
<td></td>
<td>Rapid wasting</td>
<td>Fungal/predation</td>
<td>Goran et al., 1998</td>
</tr>
<tr>
<td></td>
<td>Necroplasta</td>
<td>Cancer?</td>
<td>Goran et al., 1998</td>
</tr>
</tbody>
</table>

* Criteria for disease designation: active tissue mortality, tissue necrosis, bared skeleton, mucus production, bisected or partial polyps.


### IV. CORAL DISEASE

#### A. Identification of Diseases

Coral reefs are no exception to the truism that, even in healthy ecosystems, disease is part of the natural environment. Diseases in the ocean, however, are poorly understood because of the conceptual and methodological challenges in studying ephemeral phenomena in an alien environment. This explains why most coral reef pathogens are unidentified (Table VI). For instance, of the twelve scleractinian coral diseases easily recognized by their symptoms, only two have been positively identified (Table VI). Nonetheless, an increase in either the frequency or severity of disease epidemics, called epizootics in animal populations, can be cause for legitimate concern.

While there is a perception that the incidence of coral disease has increased (Harvell et al., 1999), it is easy to dismiss these accounts as either biased by heightened environmental concern or unfounded in the
absence of baseline data. However, evidence to the contrary is mounting. White band disease in the Caribbean inflicted heavy losses in both St. Croix and Belize (Aronson and Precht, 1997). Paleontological evidence has demonstrated that disease outbreaks in Belize have no historical precedence over the past 5000 years (Aronson and Precht, 1997), lending credence to the idea that disease outbreaks on the present scale are a recent phenomenon.

To date, most of the well-documented epizootics are from the Caribbean, but it is not clear if this represents a real difference between the Caribbean and the Indo-Pacific or merely a difference in observational coverage. The need to know more is urgent. Only multidisciplinary teams will be able to provide the ecological information necessary to devise appropriate management strategies.

B. Effects of Diseases on Diversity

Coral reef scientists are coming late to the realization that disease may exert a major control on diversity. In his review of factors explaining the biological diversity of coral reefs, Connell (1978) does not mention disease. The Environmental Protection Agency’s Coral Reef Monitoring Project in the Florida Keys has been collecting information on coral disease since 1996 (Fig. 4). Because these data have been collected systematically, they allow one to resolve whether coral diseases are more widespread now than in the past. For the 3-year period covered by the survey, these data show significant increases in all disease parameters measured, including the number of stations and the number of species with diseases present. Of the 160 stations surveyed, the number of stations with diseased corals rose from 26 in 1996 to 131 in 1998, an increase of 404%. Over the same period, the number of species affected by disease rose from 11 to 35, an increase of 218%. Many of the rarest corals disappeared from the study sites due to disease.

C. Ecosystem Effects of Disease

When diseases dramatically affect populations of a single species, the effects can influence whole ecosystems. The first documented coral reef epizootic occurred between 1982 and 1983 when almost all of the black-spined sea urchins, Diadema antillarum, in the Caribbean died from an unknown pathogen. From its point of origin near the Atlantic terminus of the Panama Canal, this disease spread throughout the Caribbean as a waterborne agent moving at the same speed and in the same direction as well-mapped Caribbean oceanic currents (Lessios et al., 1984). Diadema is a major herbivore on Caribbean coral reefs, and its loss led to an increase in algal abundance, especially on reefs with reduced herbivorous fish populations due to overfishing.
Coral disease stress model. While natural background levels of disease are expected even in healthy ecosystems, a variety of stresses could lead to the suppression of the immune and disease defense systems in coral. The consequence of reduced health would be an increase in the number of pathogenic organisms, susceptible species, locations, and mortality rates. All of these results have been observed in the Florida Keys, and while they do not prove the accuracy of this model, this hypothesis is at present the only one that explains all of the observations.

(Hughes and Connell, 1999) or on reefs with elevated nutrient levels from coastal eutrophication (Lapointe, 1999).

Corals themselves have experienced mass mortalities due to epizootics. A new disease, white pox, has inflicted high mortality among *Acropora palmata* stands on some Key West coral reefs (Table VI). For some white pox and white-band outbreaks, coral mortality rates were as high as coral losses during the worst crown-of-thorns starfish "plagues" in the Indo-Pacific (Birkeland, 1996). In the Florida Keys, the most dramatic change linked to coral disease can be seen in the loss of living coral exhibited at the deep site (18 m) on Carysfort Reef in the Upper Keys (Fig. 4). Sixty percent of the living coral there died in 2 years, mostly due to disease. Clearly, Floridian coral reefs cannot survive if these mortality rates continue.

Because corals grow slowly and live for decades or centuries, epizootics will have far-reaching impact on coral reefs on geological time scales. When deadly diseases decimate coral populations to this extent, geological rates of carbonate deposition in the Caribbean may actually be affected (Aronson and Precht, 1997). It is clear that disease epidemics can have a real impact on coral reefs.

**D. The Coral Disease Model**

We propose a coral reef disease model (Fig. 5) that depicts how changes in environmental conditions alter the interactions between hosts and pathogens and subsequently enable disease outbreaks. Stress factors (Porter et al., 1999) are considered relevant, even for corals whose immune systems are not well known, because the ability to resist infection is a function of the host’s overall health. Compromised immune systems result in increased susceptibility to disease. One of the most striking aspects of the disease patterns seen in the Florida Keys is the simultaneous increase in all disease parameters measured (Fig. 4). Only a hypothesis that addresses environmental quality will explain the simultaneous increase in the number of diseases, the number of species affected, and the rates of coral mortality throughout such a large geographic area. If this model is correct, then the incidence of disease would be expected to be higher near polluted population centers.
In 1997 and 1998, severe coral bleaching episodes were caused by dramatically elevated sea surface temperatures worldwide (see also Fig. 8). A vast majority of elk horn corals (Acropora palmata) on Looe Key, in the Florida Keys, bleached stark white. It is not known whether these colonies would have recovered because on August 27, 1998, Hurricane Georges removed nearly all branching corals from this reef, including the bleached colonies shown here (photograph by James W. Porter).

A model of sea surface temperatures based on greenhouse gas concentrations and El Niño Southern Oscillation events predicts temperatures will exceed normal thresholds for many reefs in the very near future. The horizontal lines indicate the temperature thresholds at which corals begin to bleach. As the twenty-first century proceeds, a higher percentage of time is spent above this line [Hoegh-Guldberg, O. (1999). Mar. Freshwater Res. 50, 839–866].
FIGURE 8  The burning of fossil fuels (coal, oil, and natural gas) and the destruction of a majority of the Earth’s forests by an increasing human population have caused an increase in atmospheric CO₂ concentrations over the past half century. It is highly likely that these elevated atmospheric CO₂ concentrations have also caused the incontrovertible increase in the Earth’s temperature over the same time period. Even if population growth shows some signs of slowing down as we enter the twenty-first century, energy consumption shows no such sign of declining. An increased reliance on coal, especially toward the latter half of the twenty-first century, could exacerbate rising CO₂ levels considerably [Houghton, J. T., Meira Filho, L. G., Callander, B.A., Harris, N., Kattenberg, A., and Maskell, K. (1996). Climate Change 1995. The Science of Climate Change. Cambridge Univ. Press, Cambridge, UK]. The low temperature value in 1976 resulted in coral death in the Florida Keys [Porter, J., Battey, J., and Smith, G. (1981). Proc. Natl. Acad. Sci. USA 79, 1678–1681].

FIGURE 9  Recently, the frequency, intensity, and geographical extent of bleaching episodes have increased. During the strongest bleaching event to date, 1998, bleached corals were recorded for the first time in many provinces [Hoegh-Guldberg, O. (1999). Mar. Freshwater Res. 50, 839–866].

for example, or following bleaching events (Fig. 6), both of which might be expected to compromise the coral’s immune system.

E. The Human Connection

Oceanic diseases (Harvell et al., 1999), and wildlife diseases in general (McCallum and Dobson, 1995), appear to have increased. It is not premature to ask whether or not these disease outbreaks are caused by, or influenced by, humans. At present, the historical novelty of the outbreaks is a suggestive, but not a definitive, answer to this question (Aronson and Precht, 1997). Recently, however, the disease link to human activities has been strengthened by an examination of a fungal pathogen, Aspergillus sydowii, of sea fans (Harvell et al., 1999). These authors have proposed that this marine pathogen is a terrestrial fungus that has secondarily invaded the marine environment via sediment runoff from land.
FIGURE 10  The potential interaction between coral bleaching and disease can be seen in this montage of images from 1997 to 1998. Healthy colonies of Montastrea annularis (left, May, 1997) on Looe Key, in the Florida Keys, bleached in late summer due to elevated sea surface temperatures (middle, October, 1997). This colony also contracted black band disease (middle, lower part of the image). By May, 1998 (left), most of the colony had recovered, but the black band damaged tissue did not.

V. CORAL REEFS AND GLOBAL CLIMATE CHANGE

A. CO2, Temperature, and Human Population Growth

The Earth is warming. Data from analyses of tree rings, sea ice extent, and ice cores, as well as direct measures of air and sea surface temperatures in both the Atlantic and the Pacific (Hoegh-Guldberg (1999), Fig. 7) demonstrate that the Earth is warmer now than a half century ago. The temperature rise closely parallels human population growth and the growth of atmospheric CO2 inputs from the burning of fossil fuels (Fig. 8). CO2 is one of the greenhouse gases, and the general consensus is that the buildup of this gas in the Earth’s atmosphere is causally related to the measured temperature increases.

The frequency and intensity of major storms, such as hurricanes, are expected to increase with increasing temperatures. These storms cause direct physical destruction of corals by increased wave action and scouring (Birkeland, 1996). Indirect effects include increased sedimentation and turbidity and release of nutrients from dying tissues. Some species are more resistant to storm damage than others, so the frequency with which storms strike could influence the diversity of corals present on a reef (Birkeland, 1996). There is an ongoing debate as to whether storms increase coral diversity (Connell, 1978); and there is substantial evidence for both sides of the argument (Sebens, 1994). Ultimately, the effects of storms will depend on the ability of the reef corals to recover from this disturbance.

Their recovery ability may be severely compromised in areas subjected to strong anthropogenic influences (Sebens, 1994).

B. Coral Bleaching and Elevated Sea Surface Temperatures

Coral bleaching is the loss of the symbiotic algae and is caused by elevated temperature. All marine organisms harboring zooxanthellae lose their symbiotic algae when exposed to high temperatures. Temperature-induced bleaching occurs in one of two ways, either by brief exposure to moderately increased temperature (1.5–2.0°C above average summertime temperatures for several days) or by prolonged exposure to slightly elevated temperature (only 1.0–1.5°C above normal for 3–4 weeks beyond the end of the typical summer warm season). Bleached corals appear white and lifeless (Fig. 6). The ghostlike appearance is deceptive. The chalky coloration is not due to the coral’s death but instead due to the fact that, in the absence of pigmentation conferred by the symbiotic algae, the flesh of the coral is transparent. The white limestone skeleton of the coral is visible underneath its tissue. If very high temperatures persist for a few weeks, or even if moderately high temperatures persist for more than a month, the coral will die.

FIGURE 12 Projected changes in reef calcification rates are depicted as a percent of conditions from 1880 [reprinted from Kleypas, J. A., Buddemeier, R. W., Archer, D., Gattuso, J-P., Langdon, C., and Opdyke B. N. (1999). Science 284, 118–120 © 1999 American Association for the Advancement of Science]. This model suggests that oceanic conditions in the year 2100 will be substantially less optimal for coral growth than in the nineteenth century.

The evidence suggests that coral reefs are at serious risk from high temperatures. Over the past 20 years, there has been a dramatic increase in the number of reef provinces bleaching (Fig. 9) and in the severity of these bleaching episodes. During the 1982–1983 bleaching event, Glynn and Feingold (1992) documented up to 95% loss of corals in the Galapagos Islands. Mass mortalities have also been reported recently for Australia and the Indian Ocean (Hoegh-Guldberg, 1999). Unfortunately, arguments over the cause of high temperatures have clouded the unambiguous connection between rising temperatures and increased coral mortality. As the earth warms, more corals will die. It remains to be seen whether corals can evolve genetic
Based on predicted increases in sea surface temperatures, models indicate bleaching events per decade will increase [Hoegh-Guldberg, O. (1999). *Mar. Freshwater Res.* 50, 839–866]. Resistance fast enough to survive this coming thermal stress, or alternatively, whether human energy policy can evolve fast enough to prevent corals from the necessity of having to do so.

Bleaching represents a life-threatening stress to corals. The disease model presented in Fig. 5 suggests that bleaching should promote increased susceptibility to disease. This hypothesis has not been tested experimentally, but anecdotal observations from the Florida Keys suggest it may be correct (Fig. 10).

**C. Coral Calcification and Elevated CO₂**

Since the ocean is in equilibrium with the atmosphere, rising CO₂ concentrations will cause an immediate increase in the amount of carbon dioxide dissolved in seawater. While this increase is not expected to modify the highly buffered pH of the ocean, it will alter the ocean’s chemistry (Fig. 11). Tropical surface waters are supersaturated with dissolved calcium carbonate. Corals exploit this supersaturation to manufacture their calcium carbonate skeletons at a substantially reduced metabolic cost. Over the next century, grossly elevated atmospheric CO₂ concentrations are expected to reduce this supersaturation and reduce coral growth (Fig. 11). Kleypas et al. (1999) argue convincingly that this reduction in coral reef calcification has already begun (Fig. 12). The end point of this global experiment is not known, but it is extremely worrisome.

**D. Global Climate Change and Coral Reef Survival**

Climate change models predict that tropical sea surface temperatures will continue to rise (Figs. 7 and 13). If these scenarios are correct, then bleaching will be (a) more frequent, (b) more prolonged, and (c) more lethal (Hoegh-Guldberg, 1999). These predictions are not for the distant future, but for the near future, only a few decades away. It is also becoming clear that although coastal zone management practices are critical in protecting the well-being of some coral reefs, especially those near population centers, over the next century, global climate change, and how humans mitigate this anthropogenic stress, will determine the long-term survival of the most diverse environment on Earth.

**Acknowledgments**

We thank G. P. Schmahl (Sanctuary Program, NOAA) for coral biodiversity information about the Flower Gardens Bank, Kathryn L. Patterson (University of Georgia, Athens), Stephanie Borrett (University of Georgia, Athens), DeeVon and Craig Quirolo (Reef Relief, Key West) for editorial suggestions.

**See Also the Following Articles**

- CRUSTACEANS
- EUTROPHICATION AND OLIGOTROPHICATION
- GRAZING, EFFECTS OF
- INVERTEBRATES, MARINE, OVERVIEW
- MARINE ECOSYSTEMS
- PLANKTON, STATUS AND ROLE OF

**Bibliography**


I. Need for Reforestation
II. Methods of Overcoming Degradation
III. Can Restoration Ever Be Achieved?
IV. Methods of Reforestation
V. Ecosystem Recovery after Reforestation
VI. Choosing Areas to Reforest
VII. Specialized Conditions or Sites with Particular Problems
VIII. Socioeconomic Issues

GLOSSARY

degradation A loss of forest structure, productivity, and native species diversity. A degraded site might still contain trees (i.e., a degraded site is not necessarily deforested) but it has lost at least some of its former ecological integrity.
reclamation To recover productivity at a degraded site using mostly exotic tree species. The original biodiversity is not recovered although the protective function and many of the original ecological services may be reestablished.
reforestation The reestablishment of trees and understory plants at a site previously occupied by forest cover.
rehabilitation To reestablish the productivity and some, but not necessarily all, of the plant and animal species thought to be originally present at a site. For ecological or economic reasons, the new forest might also include species not originally present at the site. The protective function and many of the ecological services of the original forest may be reestablished.
restoration To reestablish the presumed structure, productivity, and species diversity of the forest originally present at a site. The ecological processes and functions of the restored forest will closely match those of the original forest (see text for alternative definition).

Biodiversity and productivity are being lost as natural ecosystems are fragmented and degraded. Reforestation is used to reverse this process but this process can take a variety of forms. In some cases, reforestation can restore significant amounts of plant and animal biodiversity to degraded sites. In other cases, only productivity is restored. True ecosystem restoration is very difficult and more modest goals are often all that is possible, particularly over large areas. Successful reforestation usually requires solving social and economic problems as well as just ecological problems. These socioeconomic issues are often the most difficult to resolve.

I. NEED FOR REFORESTATION
Large areas of the world’s forested land have been cleared to enable agricultural development. Estimates
by FAO (1995) suggest 163 million ha of forest were lost between 1980 and 1990, of which more than 90% was in the tropics.

Even if the rate of the world’s population growth is reduced and human populations are stabilized, it is likely that much more of the world’s remaining natural ecosystems will be cleared. Many of these cleared lands are now used for food or fiber production but many have been degraded through misuse and have since been abandoned. In the face of this process of biological simplification, much effort has gone into developing a comprehensive, representative, and adequate series of nature reserves to protect the biodiversity remaining in the intact ecosystems still unaffected by these changes. But in many places there are doubts about whether this network can be sufficient since it is based on a residual series of plant and animal communities left behind in the landscape after the original ecosystems have been fragmented. This problem might be diminished if some of these degraded lands could be reforested and recover some of the biodiversity they once contained. Then they, too, might also contribute to the conservation of biodiversity in this developing reserve network.

II. METHODS OF OVERCOMING DEGRADATION

Degradation is a subjective term. A newly cleared area of forest might be regarded as prime agricultural land by a farmer but as degraded wildlife habitat by a bird enthusiast. To some extent, degradation is in the eye of the beholder. On the other hand, many would agree that once-forested land that has lost both its structure and diversity and is not used in any productive way is now, in some sense, “degraded.” The two main components of degradation are shown in Fig. 1. One component is the proportion of the original biodiversity still present and the other is the proportion of the original structure or productivity remaining. The original undisturbed ecosystem is at point A and point B represents the same site in a degraded state. The degraded area now has less biodiversity and is at a lower level of productivity or structure. The degraded site might degrade even further, perhaps because of recurrent fires (point C) or, if not disturbed further, might gradually recover, unaided, back to point A. The circumstances under which this recovery could occur are described further below.

The process of recovery can be accelerated by reforestation but there are several ways in which this might be done. Each leads to a different outcome. Restoration describes the process in which the aim of reforestation is to restore the ecosystem back to its former condition (point A) containing the original complement of plant and animal species. At this point the ecosystem will have regained its original structure and productivity and the ecological processes that sustained the system will have been reestablished. (This definition may be too exclusive in some situations and the notion of “restoration” is discussed further below.) Reclamation, on the other hand, attempts to restore structure and productivity but not the original biodiversity (point D). Reforestation may be carried out using a single species and that species might be an exotic tree species planted for commercial purposes. The structure and productivity may be partially regained (point D1) or even exceed that of the original ecosystem because of the use of fertilizers or other management inputs (point D2).

Between these two approaches lies a midway position that might best be described as rehabilitation. This may recover the original structure and productivity but may not recover all of the original biodiversity (point E).

The new ecosystem may contain a mix of native and exotic species. Over time, the new systems at D and E
may gradually drift toward point A as some of the original species recolonize from intact forest nearby or they may remain in a partially restored state if no recolonization takes place.

These three forms of reforestation are a necessary simplification of the variety of methods of reforestation that might take place on degraded lands. They differ in the degree to which biodiversity is recovered but they also share certain common attributes. These include the fact that a new stable and productive land use is achieved and that at least some of the ecological services and protective functions of the original forest have been recovered.

III. CAN RESTORATION EVER BE ACHIEVED?

Some have questioned if reforestation can ever achieve complete ecological restoration and recover all the biodiversity once present at a degraded site. For example, restoration implies that the original species complement at a site is known and that this composition was static. But the original state of many long-degraded sites is often poorly known and successional development is more common than any static "climax" condition. Even without degradation, changes might have occurred to the original condition in both space and time. And what of places such as the Mediterranean Basin, where some of the original species are no longer present in the landscape following a long process of deforestation and fragmentation. Or where other exotic species have become naturalized over a long period, because of particular cultural practices, and are now impossible to eradicate? These difficulties may mean that "restoration" is sometimes an uncertain target that might be difficult to achieve in practice. For these and other reasons, the Society for Ecological Restoration defines ecological restoration as "the process of assisting the recovery and management of ecological integrity. Ecological integrity includes a critical range of variability in biodiversity, ecological processes and structures, regional and historical context, and sustainable cultural practices."

Ecological restoration may also be difficult for other reasons. In some severely degraded sites the numbers of species to be restored may be too high or the magnitude of the changes such as exposure, topsoil erosion, or increases in salinity may be too great that restoration is too difficult to achieve even if the technical means were available. The costs of attempting full restoration of the original system might simply be too high. In some situations social constraints may also apply. Some traditional landowners or managers may be unwilling to agree to restoration of degraded sites they are not currently using because it is not a goal they share or because they believe restoration might somehow lessen their rights to its future use. In such cases outside intervention to achieve restoration is unlikely to be successful.

None of these constraints mean that reforestation to improve the biodiversity across degraded landscapes is not worth attempting. Rather, they simply point to the fact that different ecological and social situations will require different approaches.

IV. METHODS OF REFORESTATION

In some situations degraded forest ecosystems can recover unaided. A well-known example is the reforestation that has occurred over the past 100 years on much of the northeastern United States following the abandonment of some of the land previously cleared for agriculture. The former farm sites are now occupied by deciduous woodlands with a structure and diversity probably similar to that which was originally present. This same natural recovery process can be found in many other temperate and tropical forests. The rate at which such successions take place varies widely but most are usually slow. Natural recovery after degradation is not an invariable process and it only takes place under certain conditions. In all cases the disturbing agents that were the causes of degradation must have been removed and the original topsoils must have remained more or less intact. Further, remnant communities of the original forest species must have persisted on the landscape to act as a source of propagules and colonists for the new succession.

A particular problem with such slow recovery processes is that it increases the risk that sites might be damaged again by disturbing agents such as fire or grazing that degrade the site once more (and take it back to point B or C on Fig. 1). Alternatively, new land users may appear and think a site has been abandoned because they are unaware it is being reforested. More active reforestation programs may be needed to accelerate the process and overcome these risks. These programs can take a variety of forms but all require that the causes of degradation (fire, grazing, firewood collection, etc.) be removed and that sites be actively protected. The subsequent reforestation process can then be carried out by direct seeding or by planting seedlings.
A. Direct Seeding

The advantages of direct seeding are that a costly nursery stage is avoided and seed can be broadcast in the field comparatively cheaply. Seed can be sown by hand or vehicle or can be distributed from the air and large areas can be covered relatively cheaply (Fig. 2). The disadvantage of direct seeding is that only a small proportion of the seed distributed is likely to be established as seedlings as many are removed by insects or fail to germinate under field conditions or are outcompeted by weeds. This means the method may require large volumes of seed to achieve a desired tree density and it may be impossible to acquire such volumes for many of the less common species being reestablished at a particular site. Situations where direct seeding has been most useful have been mine sites where topsoil has been respread after mining has ceased and where competing weeds have not yet reestablished. Direct seeding is also possible in areas where competing weeds can be removed by plowing or weedicide. In such circumstances a cocktail of species might be reintroduced as seed although the diversity of species in such a mix can be limited by the availability of seed for some less common species and these might need to be added to the site as seedlings.

B. Planting Seedlings

Planting of seedlings raised from seed (or cuttings) in a nursery is a more expensive operation but it may be necessary for species with very small seeds or particular germination or mycorrhizal requirements. It also necessitates other management inputs such as site preparation (ripping or plowing), weed control, or fertilizing which improve survival and foster more rapid growth of the planted seedlings. In plantation monocultures being established for commercial reasons (i.e., reclamation), it is common to plant trees in regularly spaced rows. This also permits the use of planting machines pulled behind tractors which can lead to very rapid planting rates. In a restoration operation a more irregular planting design might be used with a random distribution of species.

C. The Choice of Species

A key issue in such restoration plantings is the number of species to use. It is rarely the case that all of the original species can be planted and usually some subset must be chosen. The choice of which species to use will be guided initially by the objectives of the reforestation program. If the goal of reforestation is commercial production, then a fast-growing, exotic species might be chosen. Such species usually have readily available seed, well-established nursery and planting methods, and defined silvicultural prescriptions to govern their management. If, on the other hand, ecological restoration of a former ecosystem is the goal, then it is likely that a larger number of native species will be used. In these cases the seed of many species may not be as readily available and much less might be known about how to raise large numbers of seedlings of these species or establish these in the field. At some heavily degraded areas it might be necessary to adopt a two-stage operation and use a tolerant exotic species to first alter the environment (e.g., reduce exposure, improve soil fertility, or lower water tables) before reestablishing the original native species that are not tolerant of these changed site conditions.

Complete restoration of plant diversity by sowing seeds or planting seedlings is rarely possible except...
perhaps in the most simple of ecosystems. In most situations plantings can only initiate a new succession and further ecosystem development is dependent on the remaining species colonizing the site from intact forests remaining elsewhere on the landscape (Case Study 1). In these cases particular attention might be paid to facilitating conditions suitable for the return of wildlife since many of these are mutualists and able to foster the seed dispersal needed for future successional development.

Attributes of some species that might make them attractive to establish early in any reforestation program (by direct seeding or as seedlings) are shown in Table I. The actual numbers of species to use are discussed further below.

### D. Planting Densities and Species Richness

The choice of planting density and species will be determined by the objectives of reforestation and the resources available. Commercial tree growing in well-watered temperate or tropical locations usually requires moderately close initial spacings (perhaps 1100 trees per hectare) to exclude weeds and ensure rapid site occupancy. Close spacings also reduce side branching and promote rapid height growth. In most commercial plantations a single species is commonly planted. Where ecological restoration is the objective, close spacings might also be used. Again, the dense planting excludes weeds and ensures rapid site occupancy. But high planting densities are costly and many of these seedlings may die as competition develops. In drier areas it may be preferable to use lower density plantings to ensure adequate water is available for each tree as it develops.

1. **Plant Density**

Low-density plantings may also be useful if large areas of degraded land must be restored but resources are limited. Observations of isolated trees in vacant fields suggest these trees often attract seed-dispersing birds. The birds frequently deposit seeds around the base of the tree that germinate and grow into a cluster of seedlings around each tree. Planting isolated trees or tree clusters to act as bird perches could simulate this process and foster successional development over large areas, albeit at the expense of much slower rates of recovery.

2. **Species Richness**

The numbers of species that should be included in these plantings can vary. In landscapes where large numbers of forest remnants remain close by (i.e., the dispersal

<table>
<thead>
<tr>
<th>Type of species</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native species</td>
<td>To restore the original communities once present at the site; where possible the seed of these species should be from local sources</td>
</tr>
<tr>
<td>Exotic species</td>
<td>Since only exotics can tolerate the conditions now present at the degraded site or because they are fast growing and provide a commercial return</td>
</tr>
<tr>
<td>Commercially useful species</td>
<td>For example, providing timber, fruit, oil, and nuts, financial return needed to justify funding of reforestation</td>
</tr>
<tr>
<td>Traditionally or culturally useful species</td>
<td>For example, those providing fuel, food, or medicines used by local communities</td>
</tr>
<tr>
<td>Tolerant species</td>
<td>Able to tolerate the adverse conditions likely to be found at degraded sites such as low soil fertility, exposed windy conditions, soil salinity, or occasional fire; in some cases, only exotic species may have the required attributes</td>
</tr>
<tr>
<td>Species with fleshy fruit</td>
<td>To attract seed-dispersing frugivores to the site</td>
</tr>
<tr>
<td>Poorly dispersed species</td>
<td>Species with large fruit or those with high-specialized animal dispersers unable to tolerate the conditions now present at the degraded site</td>
</tr>
<tr>
<td>Nitrogen-fixing species</td>
<td>To improve soil nitrogen levels</td>
</tr>
<tr>
<td>Fast-growing species</td>
<td>To rapidly occupy the site, shade out weeds, and improve microclimatic conditions</td>
</tr>
<tr>
<td>Understory species</td>
<td>To provide the structural complexity and habitat conditions required by some wildlife species</td>
</tr>
<tr>
<td>Multipurpose species</td>
<td>Trees that fulfill more than one requirement such as providing timber and food</td>
</tr>
<tr>
<td>&quot;Special&quot; species</td>
<td>Species required for some special purpose, e.g., rare or endangered trees requesting enlarged populations; &quot;keystone&quot; species producing food required by wildlife at certain times of the year, etc.</td>
</tr>
</tbody>
</table>
distances are short), only a modest number of species might need to be planted to initiate the succession because natural colonization by other species will take place over time and enrich the planted area. In these circumstances a ‘dense’ planting of a short-lived early colonizer may be sufficient to accelerate the successional process and only a small number of species might need to be used. Likewise, if the number of canopy species in the natural forest is low, it may be possible to replant most of these species and rely on successional development to restore understory diversity. In both cases the recovery process is initiated by planting but is completed by natural successional processes.

A different approach is needed where remnant vegetation is not close by or seed dispersal might be limited. In such cases a much higher proportion of the total plant diversity may need to be brought to the site to achieve any degree of biodiversity within a useful time period. Under these circumstances true restoration (as defined earlier) might be difficult to achieve.

A common difficulty in all of these circumstances is that seed dispersers such as birds may distribute weed species as well as native species. All restoration programs, but particularly those relying on successional development to achieve ecosystem diversity, may therefore need substantial weed control before full site occupancy by the original species is achieved. Even then, tending may be needed for some years until canopy cover is achieved and the new community is less prone to colonization by weeds.

V. ECOSYSTEM RECOVERY AFTER REFORESTATION

Random plantings of a variety of trees will not necessarily restore the ecosystem that was once present at a site. These species must be able to reproduce and regenerate in the formerly degraded site and the animal community must be introduced as well. Some plant species may require particular soil or microclimatic conditions to germinate and grow and will not become established from seed dispersed into a newly forested area until these conditions are met. Other species may need particular regeneration niches to reproduce after they have been planted at a site. For Species A to successfully colonize or reproduce at a site, Species B may need to be established first. In most ecosystems such ‘rules of assembly’ are only poorly understood. What is clear, however, is that many animal species (including insects) are part of this development process because of the mutualistic partnerships they often form with plants that enable processes such as pollination and seed dispersal. Species A (a plant) may require Species B (a bird) for pollination, that is, also needs Species C (another plant flowering and supplying food at a differing time of the year). Ecosystem restoration is dependent on the formation of these complex trophic processes and interactions.

It is commonly assumed that animals will colonize new plant communities once successional processes have created appropriate habitat conditions. Like plants, animals may be classified as early colonists, tolerant of a broad range of conditions, or later colonists that have more specialized habitat requirements and are only able to enter the succession when it has reached a relatively mature phase. The extent to which animals actually join the new forest community will depend, of course, on the extent to which the appropriate habitat conditions develop. A commercial plantation established using a single exotic tree species is likely to be much less attractive to many wildlife than a spatially and structurally complex new forest of native tree and undergrowth species. Likewise, the structural features of an old-growth forest such as thick litter layers, hollow trees, or rotting logs on the forest floor may take years to develop following restoration. These might need to be artificially simulated to foster the reentry of particular wildlife (e.g., by dragging logs into a restored area or by using nest boxes).

Wildlife from early successional stages are often able to recolonize restored forests relatively quickly. But two factors can limit the extent to which recolonization occurs more generally. Some wildlife, especially those with habitats in mature or old-growth forests, are unable to cross degraded landscapes lacking trees. Restored forests not linked with natural forest remnants may therefore be permanently deprived of such animal species unless some means can be found to overcome the blockage. This topic will be discussed further below.

A second limiting factor is that small forest remnants in degraded landscapes often contain only a subset of the original species formerly present. Vertebrates appear to disappear more rapidly from fragments than do plants while top-carnivores and large-bodied animal species in particular have large area requirements, making them especially sensitive to fragmentation. This means that recolonization may be impossible unless these species are brought in from more distant geographic locations. Wildlife recolonization may also be limited by predatory behavior or competitive relations between various animal species. Naturalized exotic species such as foxes or cats represent a particular problem.
Threatened or endangered wildlife species may require rather special habitat conditions. Where the threat to these species is from naturalized exotic predators, these new habitats might be most easily restored on offshore islands, which ensures the exotics are excluded (that is, the “isolation” effect noted above becomes an advantage). However, a dilemma can then emerge over whether the specialized habitats required by the endangered species means other species are discriminated against. Is “ecological restoration” the goal or is a more specialized and restrictive objective such as maximizing the population of a particular species more appropriate? Answers to such questions obviously depend on local circumstances. This topic is discussed further below.

All natural ecosystems are subject to disturbances and the well-known intermediate disturbance hypothesis suggests ecosystem diversity reaches an optima when the frequency of disturbances reaches some “intermediate” level. This means that the initial protection offered to reforestation projects to prevent further degradation must be relaxed at some stage to enable normal ecosystem processes to occur. There are few guidelines to describe how this should be carried out. The disturbance regime that produced degradation is different from that fostering diversity but the nature of these differences is rarely understood. For example, at what point can fire be allowed to occur? Indeed, when should fire be actively encouraged? And how intense should this fire be? These issues will not be important in reclamation projects where enhanced productivity is the objective but might be crucial to the ultimate success of many restoration projects.

VI. CHOOSING AREAS TO REFOREST

Reforestation is expensive and takes time. It is rarely possible to quickly reforest large areas at once and strategic planning is needed to identify priority areas where the early benefits from reforestation can be maximized. When reforestation is primarily a reclamation process and is carried out for commercial reasons, the location of a reforestation area will be dictated by markets and soils. The best locations will be those where transport conditions are shortest and where productivity is highest.

When the purpose of reforestation is to restore biodiversity to a degraded site, the choice is more complex. Possible locations are outlined in Table II. Common targets are degraded areas within forested nature reserves or the buffer strips around such reserves. In these areas the task is to restore the ecological integrity of a particular site and hence that of the larger reserve area as well. Since colonization distances are likely to be short, there is a high probability of success if disturbing agents can be removed and weeds eradicated.

Another common target is to establish corridors linking forest fragments still remaining in a formerly forested landscape. The purpose of this is to foster connectivity between forest fragments as well as to increase the area of habitats for species in these landscapes. Some have argued that narrow corridors are more likely to be areas of increased predation rather than a means by which wildlife species movement across the landscape and genetic exchange are enhanced. But others have argued that any increase in spatial heterogeneity should

### TABLE II

<table>
<thead>
<tr>
<th>Location</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degraded areas in or around nature reserves</td>
<td>To remove weeds and restore ecological integrity to the reserve</td>
</tr>
<tr>
<td>Corridors between forest fragments or reserves</td>
<td>To increase the overall habitat area and to foster spatial heterogeneity across a landscape</td>
</tr>
<tr>
<td>Habitats of endangered or threatened species</td>
<td>To increase the populations of these species</td>
</tr>
<tr>
<td>Offshore islands</td>
<td>Because islands may be areas free from naturalized weeds or pests that would otherwise preclude restoration on a mainland</td>
</tr>
<tr>
<td>Riparian areas</td>
<td>To protect streambanks and because riparian areas are often centers of wildlife activity in many drier landscapes</td>
</tr>
<tr>
<td>Upper watersheds</td>
<td>To stabilize catchments, prevent erosion, and improve downstream water quality</td>
</tr>
<tr>
<td>Lower watersheds</td>
<td>To lower water tables or reduce salinity</td>
</tr>
<tr>
<td>Mine sites or other severely degraded areas</td>
<td>To overcome the extreme degradation that may have occurred and to prevent these areas from becoming erosion sources or sources of toxic leachates</td>
</tr>
<tr>
<td>General landscape matrix between forest fragments</td>
<td>To restore biodiversity or productivity across the general landscape wherever it was formerly forested</td>
</tr>
</tbody>
</table>
be beneficial and the advantages would seem to out-
weigh any disadvantages provided corridor widths are 
not too narrow.

Key targets for many restoration projects are the 
habitats of particular species such as those whose popu-
lations have declined or are endangered by other preda-
tory species. This presumes, of course, that these habitat 
requirements are known. In the latter case, offshore 
islands have been especially useful areas to use because 
they are of sufficient size to contemplate eradicating 
such predators and to be assured that recolonization 
will not occur. Encouraging results have been reported 
from New Zealand (Case Study 2).

Many reforestation programs are carried out because 
of the improvements in ecological services rather than 
because they restore biodiversity. For example, stream-
sides are often degraded by grazing cattle that cause 
bank erosion and slumping. Reforestation can stabilize 
these areas and the vegetation can act as a filter by 
trapping topsoil erosion. Such reforestation can also 
have conservation benefits, of course, since these areas 
are often the focal points within drier landscapes for 
much wildlife activity. Reforestation of upper or lower 
watershed areas may be necessary to improve waterlog-
ging or salinity and, again, indirectly improve landscape 
biodiversity. Some of the most innovative and impres-
sive restoration has been carried out at mine sites. These 
areas are intensively degraded and can cause serious 
pollution problems (e.g., heavy metals, acid drainage, 
salinity) in large areas of the catchment below the mine 
site even though the size of the actual mine itself may 
be small. Mine site restoration usually requires careful 
stockpiling and respraying of topsoils but the large 
expenses of doing this are commonly only a small per-
centage of total mine expenses.

VII. SPECIALIZED CONDITIONS OR 
SITES WITH PARTICULAR PROBLEMS

Some sites have particular problems. This might be 
because of the degree of degradation suffered (com-
pacted or infertile soils, saline areas, waterlogged sites) 
or because of the stressful environmental conditions 
present (e.g., unusable soils, windsewer or cold moun-
tain areas, desert areas). A variety of techniques have 
been developed to deal with many of these conditions, 
including ripping and plowing, fertilizing, using drains 
or mounds to allow more tolerant species to become 
established at waterlogged or saline sites, etc. In some 
cases mycorrhiza and soil fauna such as earthworms 
must be reintroduced as well as plants to promote nutri-
ent cycling. Sites in extreme environments such as 
mountainous or exposed areas are particularly difficult. 
Sometimes tolerant exotic species can be planted to 
facilitate the later establishment of native species. But 
the survival rates of planted seedlings in especially se-
vere environments can be very low. This raises the 
question of whether it is worth even attempting to 
restore severely degraded ecosystems or whether a bet-
ter return on the resources to be invested might come 
from restoring only “moderately” degraded landscapes?

In such cases the balance between what is biologically 
possible is confronted by what is economically feasible.

VIII. SOCIOECONOMIC ISSUES

Ecological restoration can be difficult because it re-
quires detailed knowledge of the autecology of a num-
ber of species and of the processes of successional devel-
opment. But it is frequently less complex than the 
socioeconomic issues surrounding many reforestation 
projects. Degradation is commonly caused by some 
kind of socioeconomic problem and it can only be over-
come if these problems are also dealt with. That is, 
successful reforestation projects invariably involve re-
solving problems concerning ecological, social, and eco-
omic systems. Our understanding of the interactions 
between these three systems is usually poor.

Many reforestation projects directly or indirectly in-
volve a number of stakeholders. These might be land-
owners, leaseholders, or people who have some tradi-
tion claim to the land although that claim is not 
recognized by the state. Others may be land users down-
stream from the land or simply the general public. As 
noted earlier, these various groups may view degrada-
tion in quite different ways and many may not want 
change because they perceive themselves being disad-
vantaged by the change (e.g., they lose access to grazing 
land or food-producing areas). The planning stage of 
reforestation projects must resolve these differences if 
a project is to succeed. Experience in many places sug-
gests this is best done by involving all stakeholders in 
the planning and management process, including those 
in a weak bargaining position and those with little polit-
ical power. In this way, all interests can be protected 
from the beginning. Some form of educational program 
may also be necessary to allow the various stakeholders 
to see the reasons for the project and the benefits it 
may have for themselves (e.g., perhaps in employment 
opportunities) as well as for the community as a whole. 
If these collaborative interactions do not take place or
The degree to which biodiversity is recovered can vary with the resources invested. Unlimited resources may regain most of the original biodiversity (1), while a small resource expenditure (A) is likely to restore a smaller proportion of the original biodiversity than a larger expenditure (B). Further recovery may take place over time (i.e., time substitutes for resources invested on the x axis). Further recovery may be small (2) or larger (3), depending on the extent of subsequent recolonization from nearby intact forest.

If the concerns of stakeholders are not resolved, there may be a high risk of continued disturbances and further degradation (Case Study 3). A particular problem is cost. In most situations there is a resources–response relationship of the kind shown in Fig. 3. Complete restoration usually requires substantial financial and other resources and the more resources that can be utilized, the greater the degree of restoration that might be achieved and the faster this is likely to take place. For example, restoration is obviously faster if more species are planted at the commencement rather than waiting for recolonization to occur. Likewise, planted seedlings are more likely to survive if weed control is carried out than if it is not. However, these large resources are rarely available, except, perhaps, in some postmining restoration projects. In such cases species recovery becomes dependent on time and this replaces resources on the x axis of Fig. 3 (at point A if limited resources are available or point B if more are available). But time may be an incomplete substitute because nearby remnants do not contain the species needed or because colonization is impeded. In those cases a much lower level of “restoration” is achieved (with little further recruitment producing outcome 2 or only modest levels producing outcome 3 rather than the complete restoration of outcome 1). These several factors mean that complete ecological restoration is a difficult task to achieve and that sometimes more modest goals might be all that is possible, at least in the short term.

All of this raises the question of how to finance the restoration of diversity over large areas. There is no doubt that promising techniques are now available to accelerate restoration over small areas. How can these methods be expanded to cover the large areas of degraded land now being created each year? One way of doing this might be to modify other commercially attractive forms of land use such as timber plantations or agroforestry and use these to enhance biological diversity. Most commercial timber plantations, for example, are single-species monocultures that contribute little to landscape biodiversity. But these might be broken up by a network of species-rich buffer strips (natural or restored). Such strips would have benefits as firebreaks or as streamside filters and they could substantially increase landscape heterogeneity. Or the plantation might use more than one species planted in a mosaic of monocultures across the landscape, with each species being targeted to fit a particular site type (e.g., plantings would vary with soil type, slope position, or aspect). Or species mixtures might be planted instead of monocultures. Or, finally, advantage might be taken of the fact that many plantations acquire a diverse understory over time, especially if intact forest is nearby (Fig. 4). The value of such plantations as wildlife habitats is likely to be substantially enhanced by this understory.

In the same way, agroforestry also has the capacity to improve regional biodiversity while providing eco-
nomic benefits to land users. The traditional home gardens of Indonesia and Sri Lanka, for example, are well-known agricultural systems utilizing a very rich mix of species. Perhaps even more striking are the man-made forests in parts of Indonesia based around trees that produce fruit, gums, spices, and timbers. These are managed for the tree products plus a variety of agricultural goods but they also contain a substantial degree of indigenous plant and animal biodiversity.

Neither timber plantations nor agroforestry will achieve complete ecological restoration. In the case of the timber plantations, this biodiversity is temporarily lost when each plantation unit is harvested. But both situations can potentially recover some of the former biodiversity across large areas of landscape and reverse the trend toward biological simplification that forest plantations and agriculture usually generate. And they do this while improving the socioeconomic conditions of the land users.

Much more work needs to be done to explore new forms of reforestation and to redesign timber plantations and agroforestry systems to suit the various ecological and socioeconomic situations present in the increasing areas of degraded forest landscape likely to be found in the future.

A. Case Study 1: Reforestation after Mining Tropical Forest Lands in Brazil

Mining causes complete degradation of an ecosystem and perhaps represents one of the greatest challenges to those interested in restoration, particularly if the site was previously occupied by a biologically rich tropical rain forest. The Trombetas bauxite mine in Amazonia has an annual reforestation program of 100 ha. Following mining, the sites are leveled and stockpiled topsoil is spread back over the area to a depth of 15 cm. The site is then ripped and replanted using seed, stumps, and seedlings. About 70 species are planted at a density of around 2500 trees per hectare.

Studies carried out in an area that had been replanted 10 years earlier found forest development had been rapid. A number of plant species had colonized the site from the adjoining rain forest. After 10 years these represented 60% of the total number of woody species present and the site contained around 50% of the species present in the adjoining, undisturbed forest. Some areas had rather sparser forest cover than others and the colonization process was less vigorous in these, most probably because of the presence of grasses in the understory. The density and composition of the understory developing beneath the planted trees were related to the overstory basal area and topsoil depth but negatively related to the distance from the undisturbed forest. Primary forest species were able to colonize up to 640 m although most colonists had smaller diaspores. Most of the colonists in the more open sites with grass were early pioneer species.

Surveys of wildlife found low densities of birds but high densities of bats. Most of the bird species were representative of secondary forests and are unlikely to have carried many seeds from the undisturbed forest nearby. On the other hand, the bats included several important seed dispersers which feed on a variety of tree species and are able to carry seed over large distances. A number of other mammal species (e.g., opossums, deer, agouti, and armadillo) were also observed within the study site.

It appears that a dynamic and sustainable successional environment has been established in a much shorter time than it would otherwise take to occur. This was due to the respersing of topsoil which contained a number of important tree species and the choice of trees planted in the original planting program. The program was less successful where tree survival or growth was poor, and perhaps represents one of the greatest challenges to those interested in restoration, particularly if the site was previously occupied by a biologically rich tropical rain forest.

B. Case Study 2: Restoration of Habitats of Endangered Biota in New Zealand

The island (or minicontinent) of New Zealand has a unique biota but many of these have become endangered since the arrival of humans. In the case of birds, for example, 48% of species have been lost in the past 200 years because of the introduction of predatory and browsing mammals. Even on protected areas, attempts to maintain populations of some presently endangered species are difficult because of the difficulty in controlling predatory species over large areas. New Zealand has many offshore islands that were once part of the mainland (and consequently contain many mainland species). Although a number have been deforested or degraded, they are potential safe havens for endangered species since it may be feasible to eradicate predators and browsers over these smaller and more isolated locations. This has fostered a shift in attitude from trying to simply preserve species to restoring their habitats.

The first step has been to mount a vigorous campaign to eradicate exotic species such as rodents, cats, goats, or cattle from the islands. While the larger grazing animals are comparatively easy to remove, it has only been in more recent years that techniques for eradicat-
ing rodents have become successful. In many situations this alone has led to improvements in habitat quality although the responses of many resident species is often unpredictable. On islands where deforestation has been severe, it has been necessary to replant trees and shrubs to reestablish appropriate habitat conditions. Similarly, translocations have been necessary in the case of some animals to establish new breeding populations.

The objective of this program has been quite specific; it is not to necessarily restore the original ecosystems (because of difficulties in deciding which previous time period to target) but rather to seek to reintroduce sufficient elements of the original ecosystem to enable survival of nominated species, communities, or ecosystems. This goal acknowledges that restoration may never be complete because some exotic organism may always be present or because knowledge of the system may always be incomplete. In the New Zealand context, however, the goal is a way of ensuring the protection of a significant proportion of the country’s biological heritage.

Further Reading: Towns and Ballantine (1993)

C. Case Study 3: Reforestation of Degraded Lands in Nepal

Forests are an integral part of the farming system in Nepal, with leaves and grass being used for fodder and animal bedding. As well, such material is also used in compost to maintain productivity of agricultural lands. Timber is regularly harvested for fuel.

Some parts of the middle hills of Nepal have been highly degraded over time because of intensive harvesting. However, externally sponsored attempts to encourage land users to reforest have not always been successful. If there is forest near villages, there is commonly little interest in forest protection or tree planting although indigenous management systems (primarily to define user access rights) for these forests may exist. By contrast, where there is a severe shortage of forest products and any remaining forests are more distant, there is likely to be genuine interest in forest development activities and well-developed indigenous management systems for the remaining forests. These systems are likely to have biological objectives (e.g., rotational cutting) as well as simply defining user access rights. That is, interest in reforestation depends on the “perceived need” of land users and not on an external perception that forest degradation has occurred.

Indigenous management systems can foster reforestation under certain conditions and some areas of private farmland in Nepal now have more forest cover than in earlier years. However, this does not invariably occur. For example, larger trees may be protected but grazing beneath may lead to a failure of regeneration. Likewise cutting of smaller trees for firewood may be tolerated while the larger trees are saved. A particularly difficult situation can be where a cash market for forest products develops. In such cases poor people can succumb to the temptation to cut firewood, often at night, to sell the next day. Indigenous management systems must be very strong to resist such pressures.

Tree planting has been carried out at particularly degraded sites. The most successful reforestation has been with pines, although broad-leaved species are more valued by villagers as leaf fodder. Manipulative trials with young pine plantations have found that many broad-leaved species can also then regenerate from suppressed stumps and root stocks or seed once the sites are protected. This natural successional development increases the silvicultural options available, giving local communities greater flexibility in determining their preferred forest management strategy. It is also possible to thin the young pine stand and enrich it using other broad-leaved species that might otherwise be difficult to establish. This, too, increases the options available.

Reforestation of degraded sites and the development of indigenous management systems may not always take place without some outside intervention. Once benefits start to accrue to land users, however, neighbors are often quick to pick up the ideas and follow the example. At this point it is important that government agencies who might legally “own” the land are cautious in imposing rigid bureaucratic controls within which the new forests must fit. For example, allowing someone from outside the user group to harvest resources from the new forest could easily unleash a spurge of unregulated harvesting, leading to renewed degradation.

Further Reading: Gilmour (1990), Gilmour et al. (1990)

See Also the Following Articles

Deforestation and Land Clearing • Forest Ecology • Land-use Issues • Restoration of Biodiversity, Overview • Timber Industry • Tropical Ecosystems

Bibliography


Bridgwater, A. D., and Chadwick, M. J. (1980). The Restoration of
Land: The Ecology and Reclamation of Derelict and Degraded Land. 

and Agriculture Organization of the United Nations.

Gilmore, D. A. (1990). Resource availability and indigenous manage- 

Silviculture of plantation forest in central Nepal to maximize 

Ecology: A Synthetic Approach to Ecological Research. Cambridge 
Univ. Press, Cambridge, UK.

Lamb, D. (1988). Large-scale ecological restoration of degraded tropi- 
cal forested lands: The potential role of timber plantations. Restor. 
Ecol. 6, 271–279.

Hall, London.

Majer, J. D. (Ed.) (1989). Animals in Primary Succession: The Role 
of Fauna in Reclaimed Lands. Cambridge Univ. Press, Cam- 
bidge, UK.

Munshower, F. F. (1994). Practical Handbook of Disturbed Land Reveg- 
etation. CRC Press, Boca Raton, FL.

velopment of floristic diversity in 50 year old restoration 
forest on a bauxite mined site in Amazonia. For. Ecol. Manage. 
99, 21–42.

Reconstruction of Fragmented Ecosystems: Global and Regional 

flas.ufl.edu/ser/SERhome.html

Towns, D. R., and Ballantine, W. J. (1993). Conservation and restora- 
tion of New Zealand island ecosystems. Trends Ecol. Evol. 8, 
492–497.
RELIGIOUS TRADITIONS
AND BIODIVERSITY

Fikret Berkes
University of Manitoba

I. The Context: Religion for Encoding Ethics
II. Monotheistic Traditions
III. Asian Traditions
IV. Pantheistic Traditions
V. Conclusions

GLOSSARY

animism  Belief in spiritual beings. The term is associated with the anthropology of E. B. Tylor, who described the origin of religion and primitive beliefs in terms of animism in *Primitive Culture* (1871). Tylor considered animism a minimum definition of religion and asserted that all religions, from the simplest to the most complex, involve some form of animism. From Latin *anima*, "breath" or "soul."

ethics  Codes that exert a palpable influence on human behavior. Embedded in worldviews, ethics provide models to emulate, goals to strive for, and norms by which to evaluate actual behavior.

monotheism  Belief in the unity of the Godhead, or in one God, as opposed to pantheism and polytheism. Monotheism is a firm tenet of Islam and Judaism. Christianity, with its concept of Trinity, alone among the three monotheistic religions, dilutes monotheism. From Greek *mono*, "one," and Greek *theos*, "god."

pantheism  The doctrine that identifies the universe with God. In Western thought, the term is associated with the Dutch philosopher Baruch Spinoza. His view represents an important criticism of the "orthodox" view of a god whose reality is somehow external to the reality of the world. From Greek *pan*, "all," and Greek *theos*, "god."

religion  Human recognition of superhuman controlling power, and especially of a personal God or gods entitled to obedience and worship; the effect of such recognition on conduct or mental attitude.

dx traditional ecological knowledge  A cumulative body of knowledge, practice, and belief, evolving by adaptive processes and handed down through generations by cultural transmission, about the relationship of living beings (including humans) with one another and with their environment. A subset of indigenous knowledge, which is local knowledge held by indigenous peoples or local knowledge unique to a given culture or society.

dx traditional societies  Groups in which knowledge, practice, and belief are handed down through generations largely by cultural transmission. Tradition itself evolves by adaptive processes, but not all tradition is necessarily adaptive.

worldview  The larger conceptual complex in which ethics are embedded. A. N. Whitehead called it the conceptual order, or one's general way of conceiving the universe, which supplies the concepts by which one's observations of nature are invariably interpreted. In general, world-views limit and inspire human behavior, shape observations, and perceptions. A. Toynbee's *Weltanschauung*.

Copyright © 2001 by Academic Press. All rights of reproduction in any form reserved.
This article is about religions and attitudes toward the natural environment as relevant to biodiversity conservation. Religious traditions have little to say specifically about biodiversity, but they provide the values, worldviews, or environmental ethics that shape the way in which different societies interact with biological diversity and nature in general. In this sense, religion can be part of the problem or part of the solution. Roy Rappaport suggests a general theory of religion for encoding information and for involving human emotions. Thus, religion can encode adaptive strategies for resource management and biodiversity use, and supply emotionally powerful beliefs to put these strategies into practice. The anthropologist Eugene Anderson observes that all traditional societies that have succeeded in keeping their resources productive over time have done so in part through religious or ritual representation of resource management. The key point, he says, is not religion per se, but the use of emotionally powerful cultural symbols to help maintain a sense of sacred respect.

I. THE CONTEXT: RELIGION FOR ENCODING ETHICS

Different groups of people in different parts of the world perceive and value sacredness differently. Religion is a general term that has become established in the Western world for a whole class of cultural codes and rituals. In the West, there often is a distinction between a person’s religion and environmental ethics. In many other societies, religion is part of a way of life as well as a worldview. For example, in Chinese there is no one word that translates as “religion” in the Western sense. In some traditional societies, including many American Indian groups, religion, worldview, and environmental ethics and practice are inseparable. Therefore, it is necessary to define religion in its broader sense of “human recognition of superhuman controlling power,” inclusive of both monotheistic and pantheistic traditions.

Religions provide a central organizing myth and include cultural symbols for a moral code. Conceived that way, religions can be thought to include a wider variety of beliefs, within the definition of a superhuman controlling power. For example, Rodney Dobell refers to the “religion of the market.” Peter Timmerman uses the term “econotheism” as the adoption of market economics “to bring the rest of the earth under new management.” Econotheism, argues Timmerman, “contains within it all the elements of a religion: a priesthood, scriptures, catechisms (Econ 100 textbooks), an explanation of happiness and misery, a way to get to heaven, and a core model of the human.”

Anderson holds that religion is best regarded as something providing an emotionally powerful way to “sell” a moral code. The content of the moral code is negotiable and variable; it is not in itself a part of the religion. One can have a moral code without a religion, as in secular humanism, but at the risk of losing the emotional content supplied by religion.

Religion is not a particularly popular topic among those involved in biodiversity conservation. The index of the 1,140-page Global Biodiversity Assessment contains a grand total of only two minor entries on religion. One problem with religion is that it has often been bent to serve all possible ends, including those destructive of biodiversity and cultural diversity. Thus, the role of religion in biodiversity conservation, and more generally in environmental ethics, is controversial.

There are several cautions in interpreting the discussion on religion and biodiversity. First, generalizations are always risky. As Donald Worster put it in The Ends of the Earth, “not a few scholars have fallen into the trap of speaking of ‘the Buddhist view of nature’ or ‘the Christian view’ or ‘the American Indian view,’ as though people in those cultures were all simple-minded, uncomplicated, unanimous; and totally lacking in ambivalence. Every culture, we should assume, has within it a range of perceptions and values, and no culture has ever really wanted to live in total harmony with its surroundings.”

Second, there is invariably a gap between “the ideal” and “the actual” in making sense of how societies deal with biodiversity. Ethics do not describe how people actually behave but rather set out how people ought to behave. There often is a discrepancy between belief and practice. Some scholars argue that one could spend too much time on values and beliefs, neglecting to examine behavior or the actual effects of these values on human practices regarding the environment. The field of environmental history studies the historical record of changes in the landscape as a way of evaluating the actual effects of the belief-and-practice complex of a culture.

A third caution is the intangible nature of traditions and their record in the literature. Not only do traditions change and adapt all the time, but the literature is based largely on outsiders’ accounts of various religious traditions, especially those that do not have a written record. Since traditional societies are, by definition, those cultures in which much knowledge is transmitted orally, available documentation on such groups is al-
most always suspect. This is because insiders do not normally record their teachings in writing, and outsiders do not always interpret these teachings correctly and tend to add their own interpretations, as in Black Elk Speaks.

II. MONOTHEISTIC TRADITIONS

A. Debates over the Judeo-Christian Tradition

The controversy about environmental values in Christianity and Judaism has centered on the appropriate interpretation of the relationship between God, “man,” and nature as set out in the Book of Genesis 1:26–28:

26 And God said, Let us make man in our image, after our likeness: and let them have dominion over the fish of the sea, and over the fowl of the air, and over the cattle, and over all the earth, and over every creeping thing that creepeth upon the earth.

27 So God created man in His own image, in the image of God created He him; male and female created He them.

28 And God blessed them, and God said unto them, Be fruitful, and multiply, and replenish the earth, and subdue it: and have dominion over the fish of the sea, and over the fowl of the air, and over every living thing that moveth upon the earth.

There are two widely held interpretations of these relationships. The “mastery” interpretation is that Genesis gives humans a unique status among the species by virtue of creation in the image of God, and awards humans a God-given right to exploit nature without moral restraint, except where it affects human welfare. God, according to this view, seems to have intended humans to be his viceroy on earth. Humans are to the rest of earth’s species as God is to humans. Humans are given dominion over the earth, and are expressly asked to subdue (Hebrew קַבָּשׁ, stamp down) the earth, as if nature needed humans to put it in order.

The “stewardship” interpretation also relies on the verses above, as well others before and after it, and emphasizes responsibility. Since humans are created in the image of God, they have not only rights and privileges but also special duties and responsibilities. Foremost among these is the human duty to rule the dominion of nature wisely. To degrade nature and destroy God’s other species would violate the trust God has placed in his viceroy. Thus, far from giving a free hand to exploit and degrade nature, according to this view, God expects humans to exercise stewardship in the wise use of nature.

The five Declarations of Assisi are the addresses by religious leaders to their own faithful in the Buddhist, Christian, Hindu, Jewish, and Moslem worlds. The Declarations were the highlight of an interfaith ceremony at the World Wildlife Fund’s 25th anniversary celebrations (WWF News, Nov/Dec 1986). The three monotheistic traditions represented in Boxes 1, 2, and 3 (Christian, Jewish, and Moslem) make it clear that all three strongly support the stewardship interpretation of Genesis. It should be noted that Islam’s holy book, the Qur’an, is similar in content to the Old Testament, and gives humans dominion over nature, but at the same

Excerpts from the Five Declarations of Assisi:
The Christian Declaration by Father Lanfranco Serrini

Because of the responsibilities which flow from his dual citizenship, man’s dominion cannot be understood as license to abuse, spoil, squander or destroy what God has made to manifest his glory. That dominion cannot be anything else than a stewardship in symbiosis with all creatures. On the other hand, his self-mastery in symbiosis with creation must manifest the Lord’s exclusive and absolute dominion over everything, over man and over his stewardship. At the risk of destroying himself, man may not reduce to chaos or disorder, or, worse still, destroy God’s bountiful treasures.

For St. Francis, work was a God-given grace to be exercised in that spirit of faith and devotion to which every temporal consideration must be subordinate: uncontrolled use of technology for immediate economic growth, with little or no consideration for the planet’s resources and their possible renewal; disregard for just and peaceful relations among peoples; destruction of cultures and environments during war; ill-considered exploitation of natural resources by consumer-oriented societies; unmastered and unregulated urbanization; and, the exclusive preoccupation with the present without any regard for the future quality of life.
Excerpts from the Five Declarations of Assisi:

The Jewish Declaration by Rabbi Arthur Hertzberg

The encounter of God and man in nature is conceived in Judaism as a seamless web, with man as the leader, and custodian, of the natural world. Even in the many centuries when Jews were most involved in their own immediate dangers and destiny, this universalist concern has never withered. Now, when the whole world is in peril, when the environment is in danger of being poisoned, and various species, both plant and animal, are becoming extinct, it is our Jewish responsibility to put the defence of the whole of nature at the very centre of our concern.

. . . Man was given dominion over nature, but he was commanded to behave towards the rest of creation with justice and compassion. Man lives, always, in tension between his power and the limits set by conscience.

Our ancestor Abraham inherited his passion for nature from Adam. The later rabbis never forgot it. Some 20 centuries ago they told the story of two men who were out on the water in a rowboat. Suddenly, one of them started to saw under his feet. He maintained that it was his right to do whatever he wished with the place which belonged to him. The other answered him that they were in the rowboat together—the hole that he was making would sink both of them (Vayikra Rabbah 4:6).

Excerpts from the Five Declarations of Assisi:

The Moslem Declaration by Dr. Abdullah Omar Nasseef

Unity, trusteeship and accountability, that is tawheed, khalifa and akhrah, the three central concepts of Islam, are also the pillars of the environmental ethics of Islam. They constitute the basic values taught by the Quran. It is these values which led Muhammad, the Prophet of Islam to say: "Whoever plants a tree and diligently looks after it until it matures and bears fruit is rewarded," and "If a Moslem plants a tree or sows a field and men and beasts and birds eat from it, all of it is charity on his part," and again. "The world is green and beautiful and God has appointed you his stewards over it." Environmental consciousness is born when such values are adopted and become an intrinsic part of our mental and physical makeup.

Moslems need to return to this nexus of values, this way of understanding themselves and their environment. The notions of unity, trusteeship and accountability should not be reduced to matters of personal piety; they must guide all aspects of their life and work. Shariah should not be relegated just to issues of crime and punishment, it must also become the vanguard for environmental legislation. We often say that Islam is a complete way of life, by which it is meant that our ethical systems provides the bearings for all our actions.

B. Searching for a New Religious Base for Environmental Ethics

The debates over the interpretations have been going on since the 1970s. Many environmentalists do not see monotheistic stewardship as a sufficient solution. The critics include the historian Lynn White, Jr., who started the debate over Judeo-Christian environmental ethics with his 1967 paper in Science, "The historical roots of our ecologic crisis." According to White, Christianity is the most anthropocentric (human-centered) religion the world has ever seen, especially in its Western form. In pre-Christian times every tree, every spring, every stream, every hill had its own genius loci, its guardian spirit. These spirits were accessible to men, but were very unlike men; centaurs, fauns, and mermaids show their ambivalence. Before one cut a tree, mined a mountain, or dammed a brook, it was important to placate the spirit in charge of that particular situation, and to keep it placated. By destroying pagan animism, Christianity made it possible to exploit nature in a mood of indifference to the feelings of natural objects.

Since the roots of the environmental crisis "are so largely religious." White continues, "the remedy must
also be essentially religious." He briefly entertains and then rejects Zen Buddhism as a solution: "Zen is as deeply conditioned by Asian history as Christianity is by the experience of the west, and I am dubious of its viability among us." Seeking instead an alternative Christian view, White proposes Saint Francis of Assisi as a patron saint for ecologists. "The key to an understanding of Francis is his belief in the virtue of humility—not merely for the individual but for man as a species. Francis tried to depose man from his monarchy over creation and set up a democracy of all of God's creatures."

White may be taken to recommend a Franciscan theology as the basis for a Christian environmental ethic. Others have provided other Christian theologies. For example, Rene Dubos in A God Within recommends St. Benedict, as the Benedictines were the original environmental managers of Europe who drained swamps and made the countryside both productive and beautiful. The theologian John B. Cobb, Jr., casts his net wider, beyond European cultures, in looking for a religious solution in Is It Too Late? A Theology for Ecology.

Accepting White's view that Christianity is largely responsible for the environmental crisis, Cobb examines in particular American Indian and Chinese worldviews. He rejects the American Indian view because, he thinks, Indians did not respect human life sufficiently and because the Indian way of life cannot support the existing North American population. Turning to the Chinese tradition (represented by Taoism), Cobb argues that it was unable to prevent deforestation and other ecological ills. He concludes that it is more prudent for the West to fix the Western tradition than to find a non-Western alternative. Rejecting Francis as too radical (Saint Francis had also preached poverty to the Catholic Church), Cobb finally opts for Albert Schweitzer's Christianity and his reverence-for-life ethic.

The historian Arnold Toynbee differs from the above-mentioned commentators in that he sees no fundamental problem in developing a new environmental ethic that can be integrated with Asian traditions. As a classically educated scholar, he sees a historical continuity from pantheistic traditions ("the original religion of all mankind") to the Graeco-Roman religion and to Asian religions.

By the experience of the west, and I am dubious of its viability among us." Seeking instead an alternative Christian view, White proposes Saint Francis of Assisi as a patron saint for ecologists. "The key to an understanding of Francis is his belief in the virtue of humility—not merely for the individual but for man as a species. Francis tried to depose man from his monarchy over creation and set up a democracy of all of God's creatures."

White may be taken to recommend a Franciscan theology as the basis for a Christian environmental ethic. Others have provided other Christian theologies. For example, Rene Dubos in A God Within recommends St. Benedict, as the Benedictines were the original environmental managers of Europe who drained swamps and made the countryside both productive and beautiful. The theologian John B. Cobb, Jr., casts his net wider, beyond European cultures, in looking for a religious solution in Is It Too Late? A Theology for Ecology.

Accepting White's view that Christianity is largely responsible for the environmental crisis, Cobb examines in particular American Indian and Chinese worldviews. He rejects the American Indian view because, he thinks, Indians did not respect human life sufficiently and because the Indian way of life cannot support the existing North American population. Turning to the Chinese tradition (represented by Taoism), Cobb argues that it was unable to prevent deforestation and other ecological ills. He concludes that it is more prudent for the West to fix the Western tradition than to find a non-Western alternative. Rejecting Francis as too radical (Saint Francis had also preached poverty to the Catholic Church), Cobb finally opts for Albert Schweitzer's Christianity and his reverence-for-life ethic.

The historian Arnold Toynbee differs from the above-mentioned commentators in that he sees no fundamental problem in developing a new environmental ethic that can be integrated with Asian traditions. As a classically educated scholar, he sees a historical continuity from pantheistic traditions ("the original religion of all mankind") to the Graeco-Roman religion and to Asian religions.

III. ASIAN TRADITIONS

The religious traditions of South Asia and East Asia, with their many gods and non-dominant relationship to nature, are similar in many ways to the pre-
monothestic traditions of Europe, and fundamentally different from the monothestic tradition, as Toynbee points out. “Confucianism and Taoism and Shinto, counsel man to respect nature even when he is applying his human science to coax nature into bestowing her bounty on man.” Elements of this thinking can be seen in the Buddhist and Hindu declarations in Assisi. The statements, “we should be wary of justifying the right of any species to survive solely on the basis of its usefulness to human beings” (Box 4) and “the divine is not exterior to creation, but expresses itself through natural phenomena” (Box 5)

Box 4
Excerpts from the Five Declarations of Assisi:
The Buddhist Declaration by The Venerable Lungrig Nmgyal

Buddhism is a religion of love, understanding and compassion and is committed towards the ideal of non-violence. As such it also attaches great importance towards wildlife and the protection of the environment on which every being in this world depends for survival. The underlying reason why beings other than humans need to be taken into account is that, like human beings, they too are sensitive to happiness and suffering. We should therefore be wary of justifying the right of any species to survive solely on the basis of its usefulness to human beings.

We are told that history is a record of human society in the past. From existing sources there is evidence to suggest that for all their limitations, people in the past were aware of this need for harmony between human beings and nature. They loved the environment. They revered it as the source of life and well-being in the world.

We regard our survival as an undeniable right; as coinhabitants of this planet, other species too have this right for survival. And since human beings as well as other non-human sentient beings depend upon the environment as the ultimate source of life and well-being, let us share the conviction that the conservation of the environment, the restoration of the imbalance caused by our negligence in the past, be implemented with courage and determination.

Box 5
Excerpts from the Five Declarations of Assisi:
The Hindu Declaration by Dr. Karan Singh

Not only in the Vedas, but in later scriptures such as the Upanishads, the Puranas and subsequent texts, the Hindu view-point on nature has been clearly enunciated. It is permeated by a reverence for life, and an awareness that the great forces of nature—the Earth, the sky, the air, the water and fire—as well as various orders of life including plants and trees, forests and animals, are all bound to each other within the great rhythms of nature.

The divine is not exterior to creation, but expresses itself through natural phenomena.

The Yajurveda lays down that “no person should kill animals helpful to all. Rather, by serving them, one should attain happiness.” This view was later developed by the great Jain Tirthankara, Lord Mahavira, who regenerated the ancient Jain faith that lives down to the present day. For the Jains, Ahimsa, or non-violence, is the greatest good, and on no account should life be taken. This philosophy was emphasized more recently by Mahatma Gandhi, who always spoke of the importance of Ahimsa and looked upon the cow as a symbol of the benign element in animal life. All this strengthens the attitude of reverence for all life including animals and insects. The Hindu tradition of reverence for nature and all forms of life, vegetable or animal, represents a powerful tradition which needs to be reenforced and reapplied in our contemporary context. India, the population of which is over 80% Hindu, has in recent years taken a special interest in conservation.

signal a worldview that is fundamentally different from the monothestic one.

Teachings of Hinduism can be traced in the environmental literature on species preservation and animal rights. The philosopher Arne Naess adopted the concepts of “identification” and “self-realization” from Hindu thought and used them as central ideas in Deep Ecology. Zen Buddhism (a form of Buddhism) is one of the few major Asian religions to put firm roots in contemporary Western culture through the writings of D. T. Suzuki, Alan Watts, Peter Timmerman, and the poet Gary Snyder. However, “pop Zen” or “West Coast Zen,” as Callicott and Ames call it, has been used to
reinforce selfish individualism already endemic in Western culture, instead of the self-discipline, simplic-ity, and enlightenment of true Zen. Buddhism has also influenced some environmental writing. The most radical chapter in E. F. Schumacher's Small Is Beautiful is entitled “Buddhist economics.” West is rich in technological and economic means, Schumacher points out, but it has little clue about the ends that are most appropriate for the use of those means. Buddhist economics proposes the solution that appropriate means should be used to achieve appro-priate ends. The discovery of these appropriate ends, in turn, can be guided through our place as caretakers of the world around us, and the appropriate means will then become obvious.

Taoism, one of the Chinese religious traditions, con-siders nature through the concept of tao, the way, and emphasizes living in harmony with nature of “flowing with nature.” Taoism has influenced Western culture and environmental thinking through such books as The Tao of Physics by Fritjof Capra. In a Taoist story told by Timmerman, a philosopher falls asleep and dreams that he is a butterfly. From that moment on, he is never sure if he is a philosopher dreaming that he is a butterfly, or a butterfly dreaming that he is a philosopher. In Taoist thinking, the crucial point is that there is no right answer: the moment an answer is attempted, the original symmetry is broken and one crystallizes out as a philosopher or a butterfly—but not both. Capra, a physicist, argues that quantum physics works in much the same way as electrons move out of a probability sphere into determined existence.

Romantic notions of Asian religions are debunked by the geographer Yi-Fu Tuan, who pointed out the discrepancies between attitude and behavior. “Western humanists commonly show bias in favor of Taoist and Buddhist traditions,” says Tuan. But the people of China, through their long history, have transformed and degraded the landscape in a major way. Specifically, Taoism, with its vaguely anti-urban and anti-humanis-tic stance, according to Tuan, represents little more than a hermit’s point of view in the larger scheme of environmental destruction. Tuan’s exposition of epi-sode after episode of land abuse by the ancient Chinese, long before they were influenced by the West, leads to the conclusion that Chinese religions are unlikely to provide a solution for all. However, this is not to deny that there existed a tradition of natural philosophy in China consistent with contemporary ecological ideals. Rather, Tuan’s critique points out that there often exist “glaring contradictions of professed ideal and actual practice.”

IV. PANTHEISTIC TRADITIONS

A. Traditional Peoples, Indigenous Knowledge

Some traditional societies retain elements of pantheism, the “original religion” of all humankind. In these socie-ties, religion, worldview, and environmental practice are often intertwined. Religious sanctions may be in-voked in two ways in direct support of biodiversity conservation: through the prohibition of areas or of species. ‘Sacred groves’ or sacred forests occur throughout the world, especially in India, Indonesia, South America, and parts of Africa (Box 8). Even small sacred groves may be surprisingly effective in conserving bio-diversity. A botanical survey in a Nigerian sacred grove yielded 330 plant species as compared to only 23 in surrounding non-protected areas (Warren and Pinkston in Linking Social and Ecological Systems). In the case of species taboo (the word is borrowed from the Polynye-sian tabu), a study by Colding and Folke showed that about 30% of taboo species identified worldwide prohibited the use of species that also happened to be threatened by the IUCN Red Data Book. Traditional views may also help conserve biodiversity indirectly through the emotional involvement of people with their land and living things. This may en-tail, as in some American Indian traditions, a commu-nity-of-beings worldview in which animals and plants are considered persons—not human persons but per-sons nonetheless. Land may also provide a strong sense of place or sense of identity for a social group, as seen with American Indians and Inuit, Southeast Asian indig-enous peoples, and cultures in Oceania.

The report Our Common Future identified indigenous peoples as “the repositories of vast accumulations of traditional knowledge and experience that link human-ity with its ancient origins.” Agenda 21 of the 1992 Rio Conference encouraged the use of this indigenous knowledge to develop new adaptive strategies for con-servation. Who are these traditional societies and indig-enous peoples? According to international criteria, four characteristics distinguish indigenous peoples from others. They are descendants of groups inhabiting an area prior to the arrival of other populations; they are politically not dominant; they are culturally different from the dominant population; and they identify them-selves as indigenous.

Indigenous peoples are found in many parts of the world. Callcott’s book on environmental ethics and religious traditions provides a representative survey of
Box 6
Sacred Groves and Traditional Forest Conservation in Nigeria

Among the Yoruba of southwestern Nigeria, there were a number of categories of traditional sacred forests, although these have fallen into disuse in recent years. According to the Nigerian environmental study *Nigeria’s Threatened Environment* (1991), land was regularly set aside as hunting forests, religious groves, and isolation and quarantine forests, and to serve as the abode of fairies and spirits.

*Igbo ode* (hunting forest): Lands located at some distance from settlements and devoted to game. *Igbo egan* (high forest), abandoned secondary forest, when put to use for game-hunting activities becomes *Igbo ode*. Lands are often named according to their wildlife inhabitants, e.g., *Igbo erin* (elephant forest) and *Igbo efon* (buffalo forest). Only brave hunters dare use such specialized forests.

*Igbo oro* (religious groves). Places set aside for religious worship of many of the elements of the physical environment. They are not extensive (usually less than a quarter of a hectare) and are uncultivated forests located on the borders of a settlement and in as many separate locations as there are families of the deities.

*Igbo egbee* (religious groves “land of sorrows”): Reserved forests for the burial of people whose deaths are considered mysterious. Such lands, isolated further away from settlements, were never put under cultivation in the past when diseases were rampant and sudden deaths were attributed to the anger of the gods. Only brave hunters dare enter such lands.

indigenous ethics from various parts of the world, with major sections entitled “Polynesian Paganism,” “American Indian Land Wisdom” (with major cases of Lakota shamanism and Ojibwa totemism), “South American Eco-eroticism” (cases of Tukano systems theory and Kayapo agro-ecology), “African Biocommunitarianism” (cases of Yoruba anthropo-theology and San etiquette of freedom), and “Australian Aboriginal Conservators.”

The globalization of Western culture has meant, among other things, the globalization of Western modes of production (e.g., monocultures) and resource conservation (expert-knows-best positivist science). Has ancient knowledge become irrelevant, or perhaps simply swamped by Western science and practice? Conversely, are there useful lessons that can be learned from indigenous knowledge and practice? These are some of the central questions in a growing body of literature on traditional ecological knowledge.

B. The Context of Traditional Ecological Knowledge

Berkes and colleagues define traditional ecological knowledge as a cumulative body of knowledge, practice, and belief, evolving by adaptive processes and handed down through generations by cultural transmission, about the relationship of living beings (including humans) with one another and with their environment. As a knowledge–practice–belief complex, traditional ecological knowledge includes the religious traditions of a society. It is both cumulative and dynamic, building on experience and adapting to changes. It is an attribute of societies with historical continuity in resource use on a particular land. By and large, these are non-industrial or less technologically oriented societies, many of them indigenous or tribal, but not exclusively so. Traditional ecological knowledge is a subset of indigenous knowledge, generally defined as local knowledge held by indigenous peoples or local knowledge unique to a given culture or society.

Any discussion of traditional ecological knowledge and indigenous conservation needs to be qualified. The question of the relevance of traditional societies for biodiversity conservation is confounded by what has been called the “noble savage” syndrome. On the one hand, there is the romantic, Rousseauian view of “the primitive,” intrinsically attuned to nature and somehow living “in balance” with the environment. Such noble savages tend to become “fallen angels” when they come into contact with the dominant society. Lest they become a threat to the very ecosystem in which they live, they should continue to live as primitives. On the other hand, there is the view that primitive peoples are not noble savages but tend to be ignorant and superstitious. Historically, they lived as biological populations at the mercy of natural forces and supernatural beliefs, not as organized communities with their own knowledge systems and ecologically adaptive practices. They had a tendency, even as primitive hunters, to cause massive species extinctions, as in New Zealand, Polynesia, Madagascar, and perhaps the Americas and Australia.

As with all myths, there are elements of truth at
the basis of these views, and many people, including conservationists and indigenous peoples themselves, no doubt believe parts of one or more of these views. An alternative view, however, is to start with the recognition that traditional ecological knowledge is not mere tradition but a set of adaptive responses that have evolved over time. All societies, pre-scientific and scientific, strive to make sense of how the world behaves, and to apply this knowledge to guide practice. Because people were dependent for their survival on resources in their immediate environments, there were strong incentives for them to use resources sustainably. They could not mask this dependence with fossil fuel subsidies or capital markets in a globalized economy. Thus, the ability to nurture and sustain biodiversity was a selective pressure on these societies. That is, there was a survival value in conserving and augmenting local biodiversity.

Pre-scientific, traditional systems of management have been the main ways by which societies have managed their natural resources for millennia. In many cases, the main reason we still have any biodiversity to speak about is because of these systems of management. Within this context, biodiversity conservation is the indirect outcome, rather than the objective, of traditional practices, but the practices themselves had adaptive value—biodiversity conservation was often a matter of survival.

Such a probabilistic and evolutionary view clearly rejects the notion that all traditional peoples always manage resources well. What the literature suggests is that groups can make mistakes but can also learn from their mistakes. They are capable of responding to resource collapses and to environmental change, along the lines suggested by the books Linking Social and Ecological Systems and Sacred Ecology. Detailed accounts of traditional ecological knowledge do not support the noble savage. What they do is something very different: they provide a documentation of extremely detailed, pragmatic, empirical knowledge, represented in religious traditions and rituals, bound up with an emotional tie to the land and living beings.

C. Conserving Biodiversity: Traditional Rules of Thumb

How indigenous knowledge of nature and human social behavior is translated into resource use practices that tend to promote sustainable use of the biota and conservation of biodiversity is little known. Better known, many traditional peoples exhibit resource use restraints that promote conservation. But knowledge, belief, and practice leading to restraint tend to be intermingled, making it difficult to trace linkages among them. Conservation researchers often ignore social restraints on grounds that they are not considered rational in our worldview, often involving beliefs in supernatural forces, for example, as in the case of taboos. To arrive at an appropriate set of social restraints is an order of magnitude more complicated than to employ knowledge of habitat preference and behavior for efficient hunting strategies. To implement a set of social restraints is complicated also because it requires continued cooperation of a large number of individuals.

The dealings of traditional societies with nature are often hedged by prescriptions as to what, when, and how much is to be left undisturbed. These prescriptions become part of a culture and are mediated by religious traditions. Madhav Gadgil and colleagues identified four kinds of widely used “rules of thumb” as social restraints leading to indigenous biodiversity conservation practice:

1. Provide total protection to some biological communities or habitat patches. These may include pools along river courses, sacred ponds, sacred mountains, meadows, and forests. For example, sacred groves were once widely protected from Africa to China, and in fact, throughout the Old World. In the tribal state of Mizoram in northeastern India, they continue to be protected even after conversion to Christianity. It is now called a “safety forest,” while the village woodlot from which regulated harvests are made is called the “supply forest.” Ecological theory suggests that providing such absolute protection in “refugia” can be a very effective way of ensuring persistence of biological populations.

2. Provide total protection to certain K-selected species. Trees of all species of the genus Ficus are protected in many parts of the Old World. It is notable that Ficus is considered a keystone genus significant to the conservation of overall biodiversity. Local people seem to be often aware of the importance of Ficus as affording food and shelter for a wide range of birds, bats, and primates, and it is not difficult to imagine that such understanding was converted into widespread protection of Ficus trees at some point in the distant past. Taboos with apparent functional significance may also be placed on some less obvious species within the ecological community. For example, some Amazon fish species considered important for folk medicine are taboo and are avoided as food.

3. Protect critical life history stages. In south India, fruit bats may be hunted when foraging, but not at daytime roosts on trees that may be in the midst of villages. Many waders are hunted outside the breeding season, but not at heronaries, which may again be on
trees lining village streets. Cree Indians of James Bay in the subarctic hunt Canada goose, a major subsistence resource, but rarely kill or even disturb nesting geese. The danger of overharvest and depletion of a population is clearly far greater if these vulnerable stages are hunted and the protection afforded to them seems to be a clear case of ecological prudence.

4. Mandate local stewards to supervise resource use

Traditional resource harvesting systems in diverse parts of the world rely on the guidance of a traditional expert to organize the harvest, control access, supervise local rules, and generally act as a “steward.” This practice also ensures the proper use and transmission of knowledge. Further, in some societies, major events of resource harvest are carried out as a short-term, prescribed group effort. Thus, many tribal groups engage once a year in a large-scale communal hunt. Such a group exercise may also serve the purpose of assessing the status of prey populations and their habitat and may help to adjust resource harvest practices to sustain yields and conserve biodiversity.

Considering that contemporary scientific prescriptions for biodiversity conservation are little more than “rules of thumb,” these traditional prescriptions add up to a reasonably good set of practices. Thus, traditional knowledge, with its time depth of observations, can help us arrive at practical, locally tailored prescriptions for resource use relevant to conservation needs.

D. Ecosystem People, Local Knowledge, and Sacred Respect

Many practices used by indigenous peoples contribute to the conservation of biodiversity through the use of more varieties, species, and landscape patches than do modern agricultural food production systems (which tend to rely heavily on the monoculture of a few varieties). The chapter by Berkes, Folke, and Gadgil in Biodiversity Conservation documents many cases. Gary Nabhan’s work in Arizona and Mexico provides an illustration of biodiversity enhancement due to traditional multi-species agriculture, in this case, as compared to a protected area. Investigating two similar oases in the Sonoran Desert, Nabhan found 32 species of birds in the U.S. Organ Pipe Cactus National Monument on the Arizona side of the desert. The other oasis, 50 km away on the Mexico side of the border and still being farmed in the traditional Papago style by a group of Indian villagers, supported 65 species. A possible explanation was that the farmed oasis provided greater diversity and amounts of food available for the birds, as compared to the protected oasis.

There is evidence that indigenous practices based on good local observations and natural history can provide the capability to respond to feedbacks from the environment and to ecosystem change. The book Linking Social and Ecological Systems documents examples of multi-species management, resource rotation, succession management, landscape patchiness management, and other ways of responding to and managing pulses and ecological surprises. Even assuming that biodiversity conservation is the indirect outcome of these practices, for the most part, the amount of ecological understanding suggested by these practices is considerable. Social mechanisms behind these practices include adaptations for the generation, accumulation, and transmission of knowledge and the use of local institutions and rules for social regulation. Religious traditions are important for the cultural internalization of traditional practices and for the development of worldviews and cultural values appropriate for them.

Some traditional societies have ecosystem-like concepts. Ancient conceptualizations of ecosystems exist in several American Indian, Asia-Pacific, European, and African cultures. Among some indigenous peoples of the North American Subarctic, land is more than a physical landscape; it encompasses the living environment, including humans. For example, the term used by the Dene groups of the Western Subarctic, ndé (ndëk), is usually translated as land but its meaning is closer to ecosystem because it conveys a sense of relations of living and non-living things on the land. However, it differs from the scientific concept of ecosystem in that ndé is based on the idea that everything in the environment has life and spirit.

Table I provides a summary of traditional ecosystem-like concepts in which the unit of nature is often defined in terms of a physical boundary (such as a watershed) and living and non-living elements are considered interlinked. For example, among the Gitksan Indians of the Pacific Northwest, tribal chiefs describe their land boundaries as “from mountain top to mountain top” and orient themselves by two directional axes within this watershed framework: vertically up and down from valley bottom to mountain top, and horizontally, upstream and downstream. Detailed land use maps of the kinship-based house groups (wilps) of the Gitksan show that there is a close correspondence between watershed areas and wilps or clusters of wilps. Clearly, these are not merely political boundaries but watershed-ecosystems-as-territories.

In Oceania, there was a wealth of ecosystem-like...
concepts. Examples include the ancient Hawaiian ahupua'a, wedge-shaped land units controlled by local chiefs, encompassing entire valleys stretching from the top of mountains to the coast and shallow waters. The variations of the Hawaiian system may be found in the Yap tabinau the Fijian vanua, and the Solomon Islands pua'a. The common point in each is that the term refers to an intimate association of a group of people with land, reef, and lagoon and all that grows on or in them. It is the "personal ecosystem" of a specific group of people. In the Solomons, for example, a pua'a is a named territory consisting of land and sea, and it includes all lands and resources associated with a kinship group.

These examples provide insights into what Raymond Dasmann (in The Ends of the Earth) calls ecosystem people—people who depend heavily on natural resources of their own localities, and hence develop a very detailed understanding of their environment, as well as a spiritual intimacy with the land. Such indigenous ecological understanding is different from that of the scientist. These traditional views tend to depict ecosystems not as lifeless, mechanical, and distinct from people, but as fully alive and encompassing humans. In some cases, traditional concepts of land also incorporate spirits of animals and other natural objects (as among the Dene of Northern Canada and Alaska), and spirits of human ancestors, as among some African groups and among the Australian aborigines (the concept of Dreamtime).

In monotheistic traditions, humans have dominion over the earth, they have a God-given right to subdue the earth, and they are stewards of God. By contrast, pantheistic traditions do not have a dominant relationship with nature. Asian religious traditions and traditional European traditions also have a non-dominant view of nature. The record of different religious traditions in conserving biodiversity is mixed. This is because human behavior is conditioned not only by religious beliefs but also by many other forces.

Traditional knowledge—practice—belief systems, tapping the wisdom of many traditional cultures with pantheistic traditions, offers a number of lessons. These systems are characterized by a similarity of concepts of nature, in which humans are part of nature. As well, these knowledge systems are characterized by similarity in design. An example is shifting cultivation, developed apparently independently, by tropical forest peoples in Africa, South America, South Asia, and New Guinea. At the same time, traditional knowledge systems are characterized by a remarkable diversity in practice, even in adjacent areas. For example, as Robert Johannes and others have shown, taboos used in the social control of resources in Oceania can vary greatly from one island to another, and even from one part of an island to another. Such diversity in resource use practice contrasts with the relatively uniform conservation prescriptions of government agencies, and highlights the need for conceptual pluralism (Richard Norgaard).

One important lesson from traditional ecological knowledge is that values and beliefs are important in encoding ethics, including the ethics of conservation. As Rappoport and Anderson point out, the use of emotionally powerful cultural symbols is important to implement a moral code. If this is true, the incorporation of values and beliefs into biodiversity conservation efforts is more likely to succeed than the use of purely scientific arguments or purely economic incentives. A number of contemporary concepts appear to be exploring the combination of emotions and ideas that may help restore a sense of sacred respect. They include Naess’ deep ecology, E. O. Wilson and Stephen Kellert’s biophilia hypothesis (love of living beings), bioregionalism (with its combination of local self-reliance and sense of belonging), the related notion of “sense of place,” and the Gaia hypothesis of James Lovelock, which is the contemporary version of the Mother Earth idea. Each provides an approach to the understanding of reciprocal ties that bind humans with the natural

<table>
<thead>
<tr>
<th>System</th>
<th>Country/region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watershed management of salmon rivers and associated hunting and gathering areas by tribal groups</td>
<td>American Indians of the Pacific Northwest</td>
</tr>
<tr>
<td>Delta and lagoon management for fish culture (samshik in Java), and the integrated cultivation of rice and fish</td>
<td>South and Southeast Asia</td>
</tr>
<tr>
<td>Vauva (in Fij), a named area of land and sea, seen as an integrated whole with its human occupants</td>
<td>Oceania, including Fiji, Solomon Islands, and ancient Hawaii</td>
</tr>
<tr>
<td>Family groups claiming individual watersheds (moana) as their domain for hunting, fishing, and gathering</td>
<td>The Ainu of northern Japan</td>
</tr>
<tr>
<td>Integrated floodplain management (daso) in which resource areas are shared by social groups through reciprocal access arrangements</td>
<td>Melanesia, Africa</td>
</tr>
</tbody>
</table>

world, and these ties invariably have a spiritual or religious aspect.

See Also the Following Articles

CONSERVATION MOVEMENT, HISTORICAL • ECOSYSTEM, CONCEPT OF • ENVIRONMENTAL ETHICS • HISTORICAL AWARENESS OF BIODIVERSITY • INDIGENOUS PEOPLES, BIODIVERSITY AND • LITERARY PERSPECTIVES ON BIODIVERSITY • STEWARDSHIP

Bibliography


I. INTRODUCTION

It is now widely recognized that biodiversity is a multidimensional and multiscale phenomenon encompassing different organization levels (populations, species, functional groups, communities, ecosystems, and landscapes) and a wide range of spatial scales (from microhabitat heterogeneity to global-scale patterns). Identifying patterns of biodiversity and their causal factors is therefore an enormous task that requires (1) mapping and monitoring of biological patterns across different spatial and temporal scales and (2) analysis of such patterns with respect to diverse aspects of the physical and human environment. Remote sensing is one of the best tools available for coping with this challenge. Information obtained from remote sensing is intrinsically multidimensional (horizontally, vertically, across large spatial scales. Data derived from remote sensing provide information on landscape characteristics that influence biodiversity, structural and functional properties of ecosystems, spatial distribution of different components of biodiversity, patterns of natural and human-induced vegetation changes, and impacts of various disturbances and ecological interactions. Remotely sensed data can be integrated with information on the physical and human environment for analysis and modeling purposes. When combined with field surveys, the potential contribution of remote sensing to studies of biodiversity is literally infinite.
temporally, and spectrally) and may cover spatial scales ranging from a few centimeters (Fig. 1) to entire continents (Fig. 2). Data derived from remote sensing provide information on all levels of biological organization, including species, functional groups, ecosystems, and biomes. These capabilities make remote sensing an extremely valuable tool for studies of biodiversity.

The aim of this article is to draw attention to the potential and actual contribution of remote sensing to studies of biodiversity. The article consists of two main sections. The first section introduces fundamental concepts of remote sensing. The second section provides selected examples of how remote sensing is being used for mapping, monitoring, and modeling patterns of biodiversity. The examples were chosen to represent different aspects of biodiversity as well as a wide range of data sources and analytical methodologies. The article is concluded by a brief summary that points to some current and expected future developments in this field.

II. FUNDAMENTALS OF REMOTE SENSING

A. An Overview of the Remote Sensing Process

Remote sensing can be broadly defined as "the technique of obtaining information about objects through the analysis of data collected by special instruments that are not in physical contact with the objects of investigation" (Avery and Berlin, 1992). A refined definition of remote sensing that better fits the scope of this article is: The practice of obtaining information about the Earth's surface, using images acquired from airborne or spaceborne vehicles by measuring reflected or emitted electromagnetic radiation.

This narrower definition restricts our discussion to applications in which the general target of the observation is the Earth's surface, the measured energy is electromagnetic radiation, the sensors are positioned on airborne or spaceborne platforms, and the recorded data are available as two-dimensional images. While some techniques of remote sensing do not fit this formalization (e.g., laser, acoustical or sonar technologies), the above definition covers most of the methods that are currently applied for the documentation, monitoring, and mapping of biodiversity.

Figure 3 schematically illustrates the basic elements of the remote sensing process. Any remote sensing application consists of two distinct processes: data acquisition (detection and recording of electromagnetic radiation), and data analysis (extraction of information from the recorded data). The first process can be performed either photographically (i.e., by a photographic camera) or electronically (by electronic sensors). A photographic camera uses optical lenses and a light-sensitive film to detect electromagnetic radiation. The film acts
as both the detector of the radiation and the recording medium. Electronic sensors convert electromagnetic radiation into electronic signals that can be stored as digital images. The end product of this stage of remote sensing is therefore either a photograph or a digital image.

The second step of the remote sensing process is data analysis (Fig. 3). Remotely sensed data, either photographs or digital images, must be analyzed in order to provide useful information about the observed features. Analytical techniques available for the analysis of remotely sensed data are diverse, ranging from traditional methods of visual interpretation to sophisticated computer-based image processing. The final product of both visual interpretation and computer-based image processing is usually a map showing the spatial distribution of the variables of interest (Figs. 1–3). Maps generated from remotely sensed data can be used directly, or in combination with other data sources, for a variety of applications, including research, decision making, planning, and management.

B. Electromagnetic Radiation and Its Measurement

As already defined, remote sensing is concerned with the measurement of electromagnetic radiation reflected and emitted from objects and features on the Earth’s surface. The main source of radiation reaching the Earth is the sun. Some of the solar radiation is reflected from the Earth’s surface and can therefore be detected by cameras or electronic sensors. Other components of the solar radiation are absorbed at the Earth’s surface and are reradiated as thermal energy. This form of electromagnetic radiation can also be detected and recorded by appropriate sensors. As different objects emit and reflect different types and amounts of radiation, images...
capturing such differences contain much information on the nature of the Earth's surface.

1. The Electromagnetic Spectrum
A basic characteristic of electromagnetic radiation is its wavelength. Solar radiation consists of a wide and continuous range of wavelengths, stretching from less than $10^{-11}$ m to more than $10^{-5}$ m. This continuum is subdivided into several divisions called spectral bands, which share similar characteristics (Fig. 4a). The main bands used in remote sensing are ultraviolet, visible, infrared, and microwave.

The ultraviolet (UV) band lies between the X-rays and visible light with wavelength limits of 0.01 and 0.4 μm (Fig. 4a). Wavelengths shorter than 0.3 μm are unable to pass through the atmosphere and therefore only the 0.3- to 0.4-μm wavelength interval, or near-UV, is available for remote sensing (Fig. 4b).

The visible band stretches from 0.4 to 0.7 μm (Fig. 4a). This spectral range constitutes an extremely small portion of the full electromagnetic spectrum and its boundaries are defined by the wavelength limits of human vision (Fig. 4b). The visible spectrum is further subdivided into three equal segments: blue (0.4–0.5 μm), green (0.5–0.6 μm), and red (0.6–0.7 μm).

The infrared (IR) band extends from 0.7 to 1000 μm (Fig. 4a). In remote sensing this wide band is divided into two major components: the reflected IR band (from 0.7 to 3.0 μm) and the emitted, or thermal, IR band (3.0–1000 μm). The first component is essentially solar radiation reflected from the Earth's surface. This component is further subdivided into near-IR (0.7–1.3 μm) and mid-IR (1.3–1000 μm). Radiation in the near-IR behaves, with respect to optical systems, in a manner analogous to radiation in the visible spectrum. It can therefore be detected and recorded by films and cameras similar to those used for the visible spectrum. The thermal component of the IR spectrum represents heat energy that is continuously emitted by all objects on the Earth's surface. It ranges from 3 to 1000 μm, but wavelengths beyond 14 μm are largely absorbed by the atmosphere and are therefore not available for remote sensing (Fig. 4b).

The microwave band falls between the infrared and the radio bands (Fig. 4a). It ranges from about 1 mm to 1 m and contains the longest wavelengths used in remote sensing. An important feature of microwave radiation is its ability to pass through clouds, precipitation, and tree canopies.

2. Types of Remote Sensors
The sensors used in remote sensing can be divided into four main groups: (1) photographic camera, (2) electro-optical, (3) passive microwave, and (4) radar.
Photographic cameras can detect wavelengths ranging from about 0.3 to 0.9 μm (Fig. 4b). This includes the near-ultraviolet band (0.3–0.4 μm), all the visible light (0.4–0.7 μm), and the relatively short wavelengths of the near-IR (0.7–0.9 μm). This spectral range, which is nearly twice as wide as the human vision, can be recorded directly into film and is therefore referred to as the photographic spectrum.

Electro-optical sensors measure radiation in narrow wavelength ranges that can be located at various points along the near-UV, visible, reflected IR, and thermal IR spectral bands (Fig. 4b). This spectral range (0.3–14 μm) is termed the optical spectrum because wavelengths in this range can be reflected and refracted using optical devices such as lenses and mirrors. An important characteristic of electro-optical sensors operating at the thermal IR band is their ability to collect data day and night without being sensitive to dust and haze.

Passive microwave sensors are designed to measure microwave radiation that is naturally emitted from the Earth’s surface. These sensors operate at a spectral range of 0.15–30 cm (Fig. 4b). Radar systems also operate at the microwave region. However, in contrast to passive microwave sensors that measure naturally emitted radiation, radars are “active” sensors that transmit artificially produced energy to the Earth’s surface and record the reflected component of this radiation.

Three forms of remote sensing can therefore be distinguished: the detection of solar radiation reflected from the Earth’s surface, the detection of radiation emitted from the Earth’s surface, and the detection of radiation transmitted from the sensor itself.

C. Sources of Remotely Sensed Data

1. Aerial Photography

Aerial photography is the oldest form of remote sensing. Despite the increasing availability of more sophisticated imaging systems, aerial photographs remain one of the most reliable and widely used sources of remotely sensed data. In addition, aerial photographs provide some advantages that cannot be achieved using alternative sources of image data, particularly high spatial resolution and long periods of documentation. High spatial resolution enables one to detect and analyze patterns at very small spatial scales (Fig. 1, Case Study 4). The availability of historical aerial photographs makes it possible to measure and analyze long-term ecological processes (Case Study 8).

Both black-and-white and color photographs are used in remote sensing. Black-and-white photographs can be taken with either panchromatic film or infrared-sensitive film. Panchromatic film is characterized by an aggregate spectral sensitivity that includes the near-UV, blue, green, and red wavelengths (0.25–0.7 μm). It provides a tonal rendition that closely approximates the brightness of the scene being photographed and can be used to distinguish between objects of truly different color. Infrared-sensitive film has spectral sensitivity that extends from 0.25 to about 0.9 μm. This range encompasses the near-IR band in addition to the near-UV and visible wavelengths. Thus, infrared-sensitive film operates beyond the confines of the visible spectrum and can be used to detect information that is unavailable for human vision. Infrared-sensitive films are especially valuable for vegetation classification because differences between vegetation classes are often more distinct in the near-infrared than in the visible range (Fig. 3).

While the information contained in black-and-white photographs is expressed by differences in gray levels, color films offer the additional qualities of hue (dominant wavelength), chroma (color strength), and value (color intensity). This added information greatly improves interpretation ability. Two kinds of color films used in aerial photography are normal color films and color infrared films. Normal color films are sensitive primarily to wavelengths of the visible spectrum and provide a color rendition that approximates the original scene as it would be viewed by the human eye. This facilitates photo interpretation because objects appear in their natural colors. Color infrared films are sensitive to green, red, and near-IR radiation (0.5–0.9 μm). Since maximum reflectance from vegetation occurs in the near-IR spectral region, such films are highly valuable for studies of vegetated landscapes.

![Generalized spectral reflectance curves for deciduous (broad-leaved) and coniferous (needle-bearing) trees. Note that both reflectance intensity and differentiation ability are highest in the near-IR region.](image)
Remote sensing applications based on aerial photography are presented in Case Studies 4 and 8.

2. Satellite Imagery

Data acquired by spaceborne imaging systems are usually recorded in digital form and are therefore available as digital images. A digital image is composed of a two-dimensional array of picture elements (pixels), with each pixel corresponding to a particular area on the ground. The radiance measured over the ground area represented by each pixel is translated into a digital number (DN). This number (also referred to as brightness) is a positive integer expressing the average intensity of radiance measured from the particular ground area. Two important characteristics of digital images are spatial resolution and radiometric resolution. Spatial resolution refers to the size of the ground area represented by each pixel in the image. For example, a spatial resolution of 10 × 10 m means that each pixel in the image corresponds to a ground unit of 10 × 10 m. The term radiometric resolution refers to the number of DN values, or brightness levels, used for recording differences in radiation intensity. Thus, a higher radiometric resolution permits the detection of finer differences in radiation intensity. Both spatial and radiometric resolution have important consequences for the amount of information available in remotely sensed data.

Satellite images suitable for mapping, monitoring, and modeling patterns of biodiversity have usually been derived from multispectral sensors measuring reflected or emitted radiation in the optical spectrum (0.3–14 μm). Imaging systems operating at this spectral range have been used in a variety of national and international Earth observation programs. Three of these programs, namely, Landsat, SPOT, and NOAA AVHRR, are of particular importance for studies of biodiversity. The main characteristics of these programs are outlined below.

a. Landsat Program

The Landsat program is the longest running project for acquisition of repetitive multispectral data of the Earth’s surface. The first Landsat satellite was launched in 1972 and the most recent one (Landsat 7) in 1999. Landsats 1–3 covered the Earth every 18 days. Landsats 4, 5, and 7 have a coverage cycle of 16 days (Landsat 6 has failed to achieve orbit). Currently (1999), the only active satellites in this program are Landsats 5 and 7. However, data recorded by earlier Landsat missions are still being used in many remote sensing applications.

Table 1 compares the characteristics of data acquired by the two main sensors used in the Landsat program, the Multispectral Scanner (MSS) and the Thematic Mapper (TM). The MSS collects reflected solar radiation in four contiguous spectral bands, two in the visible spectrum at 0.5–0.6 μm (green) and 0.6–0.7 μm (red), and two in the near-IR at 0.7–0.8 and 0.8–1.1 μm. This imaging system was used by Landsats 1–5 (note that the same spectral bands had different numbering systems in Landsats 1–3 and Landsats 4 and 5). Data collected by the MSS are framed into individual scenes that cover a nominal ground area of 185 × 185 km at a spatial resolution of 79 m in Landsats 1–3 and 82 m in Landsats 4 and 5.

The Thematic Mapper (TM) was designed to improve the imaging capabilities of Landsat satellites. A major difference between the TM and the MSS is in the acquisition of data in seven bands instead of four, with new bands in the visible (blue-green), mid-infrared, and thermal portions of the spectrum (Table 1). Also, based on experience with MSS data and extensive research, the wavelength range and location of the TM scanner have been modified to facilitate the spectral differentiability of major Earth surface features. The radiometric resolution of the TM is higher than that of the MSS (256 vs 64 digital numbers) and its spatial resolution is also higher (30 m for all bands except the thermal band, as compared to about 80 m in the case of the MSS). These overall capabilities significantly improved the spectral coverage, radiometric resolution, and spatial resolution of Landsat images.

<table>
<thead>
<tr>
<th>Band</th>
<th>Wavelength range (μm)</th>
<th>Nominal spectral location</th>
<th>Spatial resolution (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSS</td>
<td>4 (1) 0.5–0.6 Green</td>
<td>79 (82)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 (2) 0.6–0.7 Red</td>
<td>79 (82)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 (3) 0.7–0.8 Near-IR</td>
<td>79 (82)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7 (4) 0.8–1.1 Near-IR</td>
<td>79 (82)</td>
<td></td>
</tr>
<tr>
<td>TM</td>
<td>1 0.45–0.52 Blue-green</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 0.52–0.60 Green</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 0.63–0.69 Red</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 0.70–0.90 Near-IR</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 1.35–1.75 Mid-IR</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 10.4–12.5 Thermal IR</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7 2.08–2.35 Mid-IR</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

* MSS bands of Landsats 1, 2, and 3 are numbered 4–7. The same bands are numbered 1–4 in Landsats 4 and 5. Spatial resolution is 82 m in Landsats 1–3 and 79 m in Landsats 4 and 5.
Applications based Landsat data are presented in Case Studies 1, 6, and 9.

b. SPOT Program
SPOT (Systeme Pour l’Observation de la Terre) is an international Earth observation program initiated by the French government in the late 1970s. The first SPOT satellite was launched in 1986, and the most recent one (SPOT 4) in 1998. All satellites have an Earth coverage cycle of 26 days. The sensors of SPOT 1, 2, and 3 consist (SPOT 4) in 1998. All satellites have an Earth coverage satellite was launched in 1986, and the most recent one the French government in the late 1970s. The first SPOT international Earth observation program initiated by SPOT (Systeme Pour l’Observation de la Terre) is an

Case Studies 1, 6, and 9.

b. SPOT Program
SPOT (Systeme Pour l’Observation de la Terre) is an international Earth observation program initiated by the French government in the late 1970s. The first SPOT satellite was launched in 1986, and the most recent one (SPOT 4) in 1998. All satellites have an Earth coverage cycle of 26 days. The sensors of SPOT 1, 2, and 3 consist (SPOT 4) in 1998. All satellites have an Earth coverage satellite was launched in 1986, and the most recent one the French government in the late 1970s. The first SPOT international Earth observation program initiated by SPOT (Systeme Pour l’Observation de la Terre) is an

The sensors of SPOT 1, 2, and 3 consist (SPOT 4) in 1998. All satellites have an Earth coverage cycle of 26 days. The sensors of SPOT 1, 2, and 3 consist (SPOT 4) in 1998. All satellites have an Earth coverage satellite was launched in 1986, and the most recent one the French government in the late 1970s. The first SPOT international Earth observation program initiated by SPOT (Systeme Pour l’Observation de la Terre) is an

The National Oceanic and Atmospheric Administration (NOAA) of the United States operates national weather programs that rely to a large extent on data collected by meteorological satellites. Imaging systems of such satellites are usually designed to have a relatively low spatial resolution, but very high temporal resolution. Satellites of the NOAA series, one of the longest weather satellite programs, observe the complete Earth twice a day at a ground resolution of 1.1 km at nadir. These satellites carry a multispectral scanner known as the Advanced Very High Resolution Radiometer (AVHRR). The AVHRR was designed primarily for meteorological applications, but it has been applied successfully for many land-oriented applications. Details of the specific spectral coverage vary with the specific mission, but in general AVHRR sensors collect data in four or five bands falling in the visible, near-IR, and thermal IR spectral regions. Images acquired by AVHRR are available at a full 1.1-km resolution (local area coverage, or LAC) and at a subsampled resolution of 4 km (global area coverage, or GAC). Applications based on AVHRR data are presented in Case Studies 2, 3, and 10.

D. Analysis of Remotely Sensed Data
Data obtained from remote sensing must be interpreted in order to become usable information. The outcome of the interpretation process is usually a map showing the spatial distribution of the objects or variables of interest. Such maps can be analyzed with respect to other sources of information using geographical information systems (see Section D.2).

1. Interpretation of Remotely Sensed Data
In the case of aerial photographs, the interpretation process is usually based on visual examination: the interpreter systematically examines the image (and possibly supporting materials such as maps or field reports) and tries to identify the nature of the objects and phenomena seen in the image. This form of image interpretation relies on the ability of the human mind to qualitatively evaluate spatial patterns in an image. In most cases, it involves the identification and delineation of discrete areal units throughout the image. For example, in studies of vegetation variation, the interpreter outlines the boundaries between areas covered by different vegetation types.

Digital images are usually interpreted using computer-based image processing. Recall that the brightness of each pixel in a digital image is expressed by a numerical integer (its DN value): the general objective of image processing is to “translate” the original image into a more informative image by mathematically manipulating the recorded DN values. Practically this is done by subjecting the original image, pixel by pixel, to appropriate mathematical operations and storing the results of the computation as a new image. The mathematical algorithms used in image processing are literally infinite, but most kinds of algorithms can be categorized into one of three main operations: image restoration, image enhancement, and thematic classification.

Image restoration involves various operations of spectral and geometric corrections. Digital images are noisy and suffer from different kinds of spectral and geometric distortions that reduce their quality and limit their use. For example, the spectral characteristics of objects detected in a digital image are distorted by differences in sun angle, topographic variation, atmospheric scattering, and sensor characteristics. All of these factors may influence the intensity of electromagnetic radiation measured by the sensor. Other factors may introduce positional errors to digital images. Radiometric corrections are mathematical operations that compen-
sate for various sources of spectral distortion in the data (Case Study 6). Geometric corrections attempt to correct for positional errors and to transform the original image into a new image that has the geometric characteristics of a map.

Image enhancement operations attempt to improve the detectability of objects or patterns in the image. For example, contrast stretching algorithms increase the contrast between features of interest and thus improve interpretation ability. Edge enhancement algorithms are designed to automatically detect boundaries between different kinds of features (Case Study 3). Other algorithms operate on multispectral data. For example, vegetation indices (VI) represent a group of spectral operations that attempt to enhance vegetation mapping and monitoring. The most widely used VI is the normalized difference vegetation index (NDVI) defined by the general equation

\[
NDVI = \frac{\text{near-IR band} - \text{red band}}{\text{near-IR band} + \text{red band}}
\]

This index can be computed using data acquired by a variety of imaging systems (e.g., Landsat, SPOT, and NOAA AVHRR) and has proved to be a sensitive indicator of the presence and condition of green vegetation (Case Studies 5 and 10).

Thematic classification is a process of assigning all pixels of an image into land cover classes (themes) based on their DN values. In applications such as land cover or vegetation mapping, thematic classification is the main object of the analysis. The outcome of the classification process is a new image where each pixel represents a particular land cover type. Such an image can be interpreted as a map depicting the spatial distribution of the selected land cover types within the scene (Figs. 1, 2, and 6).

Two standard approaches for image classification are unsupervised classification and supervised classification. In unsupervised classification the pixels of the image are numerically classified into a number of spectral classes based on the likeness of their DN values. The land cover identity of each spectral class is then determined using reference information such as field data, aerial photographs, or published maps (Case Study 2). In supervised classification, the analyst first selects areas of known cover type in the image. These ‘training sites’ are analyzed statistically to identify the spectral characteristics of each cover type. With appropriate algorithms, the DN values of each unknown pixel

![FIGURE 6](image_url)
in the image are compared with those of the training sites, and the pixel is assigned to the land cover type to which it is most similar (Case Study 6). It is important to emphasize that classification of digital images is never perfect. The success of image classification depends on many factors and may vary considerably from one case to another. Therefore, any classification of remotely sensed data must be followed by some form of accuracy assessment.

2. Integration of Digital Images within a GIS Environment

Geographical information systems (GIS) are computer-based systems capable of storing, manipulating, and visualizing geographically referenced data. Basically, any kind of information that can be presented by standard maps can be converted into digital form and handled by GIS. An important benefit of GIS is the ability to spatially integrate multiple types of data stemming from different sources of information. For example, in Case Study 2, GIS was used to integrate field-based data on the distribution of buzzard nests in northwestern Scotland with remotely sensed data of vegetation conditions and a set of topographic maps derived from a digital terrain model. By integrating these data in the GIS, it was possible to identify factors affecting the spatial distribution of buzzard nesting sites.

The foregoing example illustrates a typical GIS operation called overlay analysis. Many other kinds of spatial operations can be performed with GIS. For example, aggregation is an operation by which detailed map categories are aggregated into less detailed categories (e.g., aggregation of detailed vegetation units into a few broad vegetation types). Buffering is an operation by which a zone of user-defined width is created around map objects. Case Study 7 exemplifies the use of buffering as a tool for testing the effect of distance to road on deforestation rates in the Philippines. Data merging is an operation by which different grid maps of the same area are coregistered into a uniform, multidimensional database. This operation is commonly applied to combine digital images representing different wavelengths, different times of observation, or different sensors. For example, monitoring seasonal variation in NDVI requires the coregistration of images taken at different times (Case Study 7). Low-resolution images (e.g., AVHRR) can be merged with images of higher resolution (e.g., LandSat TM) for calibration or interpretation purposes (Case Study 2). Using the merging capability of GIS, one is able to coregister digital images with “ancillary” data (e.g., elevation and slope) that correspond to each pixel in the image. Once such data are spatially registered, information available in the GIS can be used to improve image classification. Thus, the interaction between remote sensing and GIS is two way in nature; remote sensing is used to generate digital maps that can be fed into GIS, while GIS data can be used to interpret and classify remotely sensed data. GIS can also be used for modeling purposes. For example, by overlaying vegetation maps derived from remote sensing with corresponding maps of climatic and topographic variables, one is able to develop regression models that predict vegetation characteristics from information on climate and topography (e.g., Case Study 1).

III. REMOTE SENSING APPLICATIONS

The previous section has focused on general concepts and elements of remote sensing. Its main purpose was to provide the scientific and technical background required to understand how remote sensing technology can be used for collecting, analyzing, and visualizing images of Earth’s surface. This section presents 10 case studies that were chosen to illustrate the wide spectrum of biodiversity issues that can be studied with remote sensing.

Three main kinds of applications are exemplified: (1) mapping of land cover and vegetation patterns, (2) measurement of temporal changes in land cover and vegetation characteristics, and (3) modeling of species distribution patterns. Each case study is described with respect to its scientific questions, data sources, and methodologies. The case studies also differ from each other considerably in their spatial and temporal scales. For example, Case Study 3 focuses on continental patterns of vegetation changes (spatial resolution 4 km), while Case Study 4 focuses on individual trees within a small forest stand (spatial resolution 0.1 cm). This trade-off between spatial scale and spatial resolution is a common characteristic of remote sensing applications. Temporal intervals between successive images of the same system range from 1 month in Case Study 5 to four decades in Case Study 8. These differences point to the diversity of spatial and temporal scales that can be investigated using remotely sensed data.

A. Mapping Vegetation and Land Cover Patterns

Mapping vegetation and land cover patterns has been the most common application of remote sensing. It may contribute to studies of biodiversity both directly (by documenting the distribution of different components
of biodiversity) and indirectly (by providing knowledge on landscape and habitat characteristics that influence patterns of biodiversity). The case studies described here were chosen to illustrate four different levels of land cover and vegetation mapping. The first study was designed to provide a general map of land cover types. The second study attempted to quantify variation in the percentage cover of one particular vegetation type (conifer forests). The aim of the third study was to develop a method of counting and mapping tree densities, while the fourth study was designed to develop methods for mapping individual tree crowns. The four studies also differ from each other in their data sources: Landsat TM in Case Study 1, AVHRR in Case Study 2, SPOT panchromatic in Case Study 3, and color IR aerial photographs in Case Study 4.

1. Case Study 1: Producing a Land Cover Map of Wyoming

The objective of this project was to produce a detailed land cover map of Wyoming for purposes of biodiversity conservation and management (Driese et al., 1997). The map was based on Landsat TM data from 1988 to 1993. Classification was performed using visual interpretation of color composites produced from TM bands 3 (red), 4 (infrared), and 5 (infrared). The color composites were displayed on a computer screen and boundaries between areas interpreted as different land cover types were delineated using on-screen digitizing. A total of 14,690 polygons were digitized, and each polygon was classified into a particular land cover type based on image interpretation and supporting information from existing maps, literature, and extensive field surveys. Boundaries between polygons classified into the same land cover type were dissolved using tools available in the GIS. The final map included 6167 polygons representing 41 land cover types (Fig. 6).

With standard GIS tools, the total area occupied by each land cover type was calculated. It was found that the two dominant land cover types in Wyoming are Wyoming big sagebrush (Artemisia tridentata ssp. wyomingensis) shrubland (33.4% of the area) and mixed grass prairie (17.5% of the area). Which of these types is more likely to occur in a particular site was found to be a function of the amount of summer rainfall (Fig. 7). By merging the Wyoming land cover map with digital topographic data, it was possible to test the hypothesis that treeline elevation decreases from north to south. The results supported this hypothesis and indicated that upper treeline decreases at an average rate of 0.5 m/km from north to south. Lower treeline showed a corresponding decrease of 1.3 m/km. Treeline also decreased in elevation by 0.35 m for every degree of aspect away from south.

This case study shows two aspects of remote sensing that are relevant for studies of biodiversity: (1) the ability to produce detailed land cover maps and (2) the ability to test ecological hypotheses by integrating such maps with other types of data (e.g., climate and topography) within a GIS environment. Land cover maps such as the one produced for Wyoming are often used for making predictions on patterns of species distribution and species richness (see Case Study 10).

2. Case Study 2: Estimating Percentage Cover of Closed Coniferous Forests in Oregon

This project was designed to test the hypothesis that AVHRR data can be used to estimate the percentage cover of closed conifer forests in the Pacific Northwest (Ripple, 1994). The motivation for this project was legislation, approved by the U.S. Congress, that such forests should be protected in order to facilitate the long-term survival of the northern spotted owl, an endangered species that needs large stands of old-growth forests to survive extinction.

An area of about 2600 km² in the western Cascade Mountains of Oregon was chosen for the study. A sub-scene of NOAA-9 AVHRR data (from July 19, 1988) covering this area was resampled to $1 \times 1$ km resolution and was registered to UTM coordinates. Each pixel in the scene was characterized by its visible (0.58–0.68 μm) and near-IR (0.725–1.1 μm) band values. Landsat MSS images of the same area (from August 31, 1988) were resampled from the original 79-m resolution to a
50-m pixel size and were merged to the AVHRR scene such that each pixel in the AVHRR image corresponded to 20 × 20 pixels of the MSS image. Using unsupervised classification, the MSS data were classified into 100 spectral classes, and these classes were further classified into two broad categories: "closed-canopy conifer forest" and "other areas." By counting the number of MSS pixels classified as closed canopy conifer forest within each AVHRR pixel, it was possible to evaluate the relationships between the spectral characteristics of the AVHRR data and the proportion of closed conifer cover. Such analyses indicated that both visible and near-IR values of the AVHRR data decrease linearly with the proportion of closed conifer cover within a pixel (Fig. 8).

A regression model calibrated from these results was applied to a larger set of AVHRR data in an attempt to create a statewide map of the percentage cover of closed conifer forests. The resulting map (Fig. 9) was used to identify areas characterized by highly fragmented vs spatially continuous patches of closed conifer forests as well as to locate forest corridors that may link different regions within Oregon.

3. Case Study 3: Estimating Tree Density in Oak Woodland of Southern Spain

This project was designed to test the hypothesis that SPOT panchromatic data can be used to estimate tree density in savanna-like vegetation systems (Joffre and Lacaze, 1993). The spatial resolution of SPOT panchromatic data (10 m) is not much larger than the canopy of mature trees. Since trees differ in their reflectance in the visible spectrum from herbaceous plants, it was expected that SPOT panchromatic data would discriminate between pixels occupied by trees and pixels representing herbaceous vegetation. It was also expected that edge enhancement algorithms would facilitate image classification by amplifying the contrast between trees and their surrounding herbaceous vegetation. These hypotheses were tested using data obtained for a landscape characterized by oak woodland in southwestern Spain. Figure 10 shows the relationships between actual tree density in a set of representative test sites and predictions of tree density based on two analytical methods. The first method was based on the original DN values of the panchromatic data. The second method was based on the thresholded Laplacian index, a measure that was calculated by integrating two types of edge enhancement algorithms. Both methods showed reasonable fit between predicted and observed densities. However, as expected, the application of edge enhancement algorithms significantly improved classification accuracy. Figure 11 presents a map of tree density classes derived from the enhanced data. Maps of this sort can be used to predict spatial patterns of tree density over large geographic areas.

4. Case Study 4: Delineation of Individual Tree Crowns in a Temperate Forest

In this study, a multiple-scale approach based on a series of image processing operations (image smoothing, edge detection, curvature estimation, and segmentation) was developed for automatically delineating individual tree crowns in high-resolution images (Brandtberg and Walter, 1998). The method was developed and tested using color IR aerial photographs of a temperate forest dominated by mixed and pure stands of Scotch pine (Pinus silvestris), Norway spruce (Picea abies), and Aspen (Populus tremula) trees. The photographs were scanned and the resulting images were resampled to a pixel size corresponding to 10 cm on the ground. Figure 12 shows a section of an original image and the result of the

delineation algorithm. Accuracy assessment based on 45 ground truthing plots indicated that, on average, 70% of the trees were identified correctly. This was almost equivalent to visual interpretation, suggesting that this method can be used for mapping large forest areas at the individual tree level.

B. Measuring Vegetation Changes

Identifying patterns, scales, and trends of vegetation changes is important for understanding the dynamics and structure of both plant and animal communities. As with mapping applications, measurements of vegetation changes based on remotely sensed data may provide direct evidence for changes in biodiversity (e.g., changes in the relative frequency of different vegetation types) as well as information on environmental changes that may influence biodiversity patterns (e.g., deforestation and forest fragmentation). The case studies presented here illustrate four different aspects of vegetation changes: interannual fluctuations in vegetation cover (Case Study 5), forest succession (Case Study 6), deforestation (Case Study 7), and tree demography (Case Study 8). The respective data sources are AVHRR, Landsat MSS, SPOT multispectral, and black-and-white aerial photographs. Time spans analyzed in these case studies range from 10 to 54 years.

1. Case Study 5: Interannual Vegetation Changes in Sub-Saharan Africa

This study was designed to evaluate the magnitude of interannual changes in the vegetation of sub-Saharan Africa and to identify the spatial and temporal characteristics of these changes (Lambin and Ehrlich, 1997). It was based on daily global area coverage (GAC) data acquired by NOAA AVHRR from 1982 to 1991. Two measures were calculated from the AVHRR data: the normalized difference vegetation index (NDVI) and an index of surface temperature ($T_s$) calculated from the thermal bands of the AVHRR. Theoretical considera-
Figures 10 and 11

**Figure 10** Relationships between actual tree density in representative sites of oak woodland and predictions of density based on two kinds of measures derived from SPOT panchromatic data: the original DN values (above), and the thresholded Laplacian index, a measure that integrates two types of edge enhancement algorithms (below). Predictions based on edge enhancement were more accurate than those based on the original data. Adapted from Joffre and Lacaze (1993). Estimating tree density in oak-savanna-like ‘dehesa’ of southern Spain from SPOT data. *Int. J. Remote Sens.* 14, 685–697. Copyright 1993, with permission from Taylor & Francis (http://www.tandf.co.uk/journal).

**Figure 11** A map showing the distribution of five classes of tree density, produced for a savanna-like ecosystem in southwestern Spain using SPOT panchromatic data. Reproduced from Joffre and Lacaze (1993). Estimating tree density in oak-savanna-like ‘dehesa’ of southern Spain from SPOT data. *Int. J. Remote Sens.* 14, 685–697. Copyright 1993, with permission from Taylor & Francis (http://www.tandf.co.uk/journal).

Tensions and empirical data have indicated that the arc-tangent of the ratio between $T_s$ and NDVI (arctan($T_s$/NDVI)) is negatively and linearly correlated with the amount of vegetation cover. This index was therefore computed for each monthly period for the 10 years of observation. The status of the vegetation in a given pixel in a given year was determined by subtracting the vector of the monthly values of arctan($T_s$/NDVI) recorded for the relevant pixel during the year of observation from a corresponding vector representing the ‘best’ conditions recorded for that particular pixel during the study period. Images of this standardized vegetation index were constructed for every year from 1982 to 1991.

Figure 13a presents an image of the standardized vegetation index produced for 1983. The brightness of a given pixel in this image indicates how much vegetation conditions in 1983 were different from the best possible conditions recorded for this particular pixel during the whole study period. A comparison of this image with a map of 1983 drought events produced by the U.S. Department of Commerce Climate Impact Assessment (the areas outlined by dashed lines) indicates that areas that experienced droughts in 1983 were characterized by relatively high brightness values. Analysis of similar images constructed for other years confirmed the conclusion that the standardized vegetation index can be used effectively to identify patterns of vegetation changes caused by rainfall fluctuations. Interestingly, while differences between images of successive years revealed a high degree of interannual variability in the vegetation, only small parts of the continent exhibited continuous trends of vegetation changes (Fig. 13b). This result contrasts the ‘marching desert’ prospect and similar views concerning the extent of deforestation and land degradation in Africa.

The overall results of this project indicate that spectral indices derived from AVHRR data can be used effectively for monitoring and interpreting continental-scale patterns of changes in vegetation cover.
2. Case Study 6: Forest Succession in Northeastern Minnesota

In this study, Landsat MSS data were used to investigate succession processes in boreal forests of northeastern Minnesota (Hall et al., 1991). The analysis was based on a comparison of two images: one from July 3, 1973 (acquired by Landsat 1), and the second from August 18, 1983 (acquired by Landsat 4). Training sites located on the 1983 image were used to define the spectral characteristics of five major states of forest succession: clearing, regeneration, broadleaf, mixed, and conifer. Figure 14 shows the spectral ranges of these states in a two-dimensional space defined by the green and the near-IR MSS bands. The spectral boundaries defined by the training sites were used to classify all unknown pixels in the two images. The 1973 image was subjected to radiometric corrections prior to the classification process in order to compensate for differences in growing season, atmospheric conditions, and sensor response characteristics between the two acquisition dates.

To quantify succession rates, the two images were coregistered and the frequencies of all possible state transitions (clearing to clearing, clearing to regeneration, clearing to broadleaf, etc.) were summarized in the form of a transition probability matrix (Table II). Each element in such a matrix indicates a probability of transition from one state to another within a certain time period (10 years in this study). Examination of the values obtained for the diagonal elements of the matrix indicates that late successional states were more stable during the study period than earlier states (i.e., had higher probabilities of staying in the same state). It can also be seen that the primary succession pathway (defined by the largest transition probability in each row) was from clearings to regeneration, to broadleaf, to mixed, to conifer. This result was consistent with ecological expectations.

A transition matrix like the one computed in this study can be used to project future changes in the vegetation, to estimate the steady-state composition of the system, and to calculate dynamic characteristics such as the recurrence time of various successional states. Remote sensing is the only tool available for studying such aspects of succession over large spatial and temporal scales.

3. Case Study 7: Deforestation in the Philippines

This project was designed to investigate factors leading to deforestation in the Philippines (Liu et al., 1993). Previous studies have suggested that road building may increase deforestation rates by improving accessibility to undisturbed forest stands. Other studies proposed
FIGURE 13 Vegetation changes in Africa as obtained from the analysis of AVHRR data. (a) An image of a standardized vegetation index based on monthly NDVI and $T_s$ values produced for 1983. The brightness of each pixel in the image indicates how much vegetation conditions in 1983 were different from the best conditions recorded for this particular pixel between 1982 and 1991. Thus, bright areas indicate relatively unfavorable vegetation conditions. Dashed lines indicate areas that experienced droughts in 1983. Note that such areas are characterized by relatively high values of the standardized vegetation index. (b) Areas showing a continuous gain (green) and continuous loss (red) in vegetation cover over a period of at least 6 years between 1982 and 1991. Only small parts of sub-Saharan Africa experienced a continuous change in vegetation cover during this period. Adapted from Lambin and Ehrlich (1997), with permission from Elsevier Science.

that small forest patches and patches with long perimeter per unit area (i.e., more edge exposed to humans) are more likely to be disturbed than large patches. It was therefore expected that (1) deforestation rates in the Philippines would decrease with increasing distance to the closest road and (2) forest patches characterized by a high perimeter-to-area ($P/A$) ratio would show higher rates of forest loss than patches characterized by a low $P/A$ ratio.

These hypotheses were tested using GIS tools by integrating three types of data: a land-use map of the Philippines derived from 187 SPOT multispectral images recorded during 1987–1988, a land-use map from 1934 based on survey data, and a road map of the Philippines from 1941. Due to differences in classification systems between the two land-use maps, all land cover types in both maps were aggregated into two main groups: forest and nonforest. The effect of roads on deforestation rates was analyzed using a buffering approach. Buffer zones were created at 1.5-km intervals around the 1941 road network. The smallest buffer included areas at a maximum distance of 1.5 km from the closest road, the second buffer covered areas up to 3 km from the closest road, etc. The largest buffer (22.5 km) covered almost the whole land area of the Philippines. These buffers were overlaid in the GIS on the forest/nonforest maps, and the percentage change in forest area between 1934 and 1988 was determined for each buffer ring (0–1.5 km, 1.5–3 km, 3–4.5 km, etc.). The results (Fig. 15) supported the hypothesis that road building accelerated deforestation. The percentage of forest loss was highest (78%) for areas within 1.5 km of roads and decreased linearly to about 40% at areas located 15 km from the closest road. Further analyses of the data indicated that forest loss was concentrated in patches characterized

FIGURE 14 Spectral ranges of different states of boreal forest succession in a two-dimensional space defined by the green and the near-IR MSS bands. Boundaries between successional states were determined by analyzing the spectral characteristics of data obtained from a set of representative training sites. Adapted from Hall et al. (1991), with permission from the Ecological Society of America.
by a high \( P/A \) ratio (Fig. 16). These results are consistent with the hypothesis that distance from road and the \( P/A \) ratio of forest patches are important determinants of deforestation processes in the Philippines.

4. Case Study 8: Demography of *Acacia* Trees in the Negev Desert

This study was designed to investigate the long-term demography of *Acacia* trees in the Negev desert, Israel (Lahav-Ginott and Kadmon, unpublished data). *Acacia* trees function as keystone species in this ecosystem by providing food and hospitable habitat conditions for many organisms, including insects, reptiles, birds, and mammals. Recent field studies have documented large-scale mortality of *Acacia* trees in the Negev and pointed to the importance of evaluating the long-term viability of these populations. This study applied a demographic approach based on the analysis of historical aerial photographs, to evaluate the viability of *Acacia* populations in the Negev.

Long-term changes in the density of *Acacia* populations were studied using image processing of black-and-white aerial photographs from 1956 and 1996. The photographs were scanned and the resulting images were corrected for geometric distortions and resampled into a spatial resolution of 30 cm. Since canopies of trees growing in this desert environment rarely overlap, it was possible to derive maps of individual tree distribution from the classified images (Fig. 17a). These maps were analyzed using a GIS in an attempt to (1) estimate changes in density of *Acacia* populations in two different sites, (2) calculate corresponding rates of mortality and recruitment, and (3) test for differences in these demographic rates between habitats characterized by high vs low input of runoff water.

The results (Table III) revealed a general trend of increase, rather than a decrease, in tree density. The magnitude of this increase varied between the two sites as well as between the two habitat types. Mortality and recruitment rates were determined by overlaying the 1956 and 1996 images in the GIS and distinguishing between three states of trees: trees that were present in 1956 but absent in 1996, trees that were present in both years, and trees that were present only in 1996 (Fig. 17b). The results of these GIS operations revealed relatively high rates of mortality and recruitment (Table III) and pointed to some patchiness in the distribution of mortality and recruitment events (Fig. 17b). The overall results of this project indicate that image processing of historical aerial photographs may provide important insights into the long-term dynamics of desert tree populations.

---

**TABLE II**

Transition Matrix Showing Probabilities of Transition between Different Successional States in a Temperate Northern Minnesota Forest As Calculated from Landsat MSS Data

<table>
<thead>
<tr>
<th>From</th>
<th>Clearings</th>
<th>Regeneration</th>
<th>Broadleaf</th>
<th>Mixed</th>
<th>Conifer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clearings</td>
<td>0.17</td>
<td>0.46</td>
<td>0.17</td>
<td>0.15</td>
<td>0.05</td>
</tr>
<tr>
<td>Regeneration</td>
<td>0.01</td>
<td>0.31</td>
<td>0.17</td>
<td>0.37</td>
<td>0.10</td>
</tr>
<tr>
<td>Broadleaf</td>
<td>0.01</td>
<td>0.20</td>
<td>0.47</td>
<td>0.28</td>
<td>0.04</td>
</tr>
<tr>
<td>Mixed</td>
<td>0.01</td>
<td>0.07</td>
<td>0.11</td>
<td>0.38</td>
<td>0.23</td>
</tr>
<tr>
<td>Conifer</td>
<td>0.01</td>
<td>0.04</td>
<td>0.02</td>
<td>0.31</td>
<td>0.38</td>
</tr>
</tbody>
</table>

* Probabilities are calculated for a time step of 10 years (1973–1983). Adapted from Hall et al. (1991), with permission from the Ecological Society of America.
C. Modeling Species Distribution and Species Richness

Obviously, most plant and animal species cannot be detected directly by remote sensing. However, the integration of data obtained from remote sensing with field-based observations of plants and animals enables one to develop statistical models that predict distribution patterns of individual species from information on the environment. A similar approach can be applied to model and predict patterns of species diversity. Examples for these two types of applications are presented in the following case studies.

1. Case Study 9: Spatial Distribution of Buzzard Nesting Sites in Scotland

This study used vegetation cover data derived from satellite imagery and topographic variables derived from a digital terrain model to identify factors affecting the distribution of buzzard (Buteo buteo) nesting sites in upland Argyll, Scotland (Austin et al., 1996). A vegetation map of the study area was produced using supervised classification of a Landsat 5 image (TM bands 3, 4, and 5) recorded in May 1990. A grid at a resolution of 500 m was superimposed on the vegetation map, and each grid cell was classified based on whether it was known to contain a buzzard nest or not. With GIS tools, the relative area of different vegetation categories was calculated for circles of radius 500, 1000, and 1500 m about the center of each grid cell. The length of boundaries between different vegetation categories was also determined for each radius, to give an index of habitat heterogeneity. Finally, a set of topographic indices was extracted from a digital terrain model to describe the topography within, and around, each grid cell. Overall, 165 habitat variables were derived from the satellite and topographic data as potential predictors of nesting site locations.

Analysis of this huge data set indicated that buzzard nesting sites could be predicted with an accuracy level of about 90% using regression models based on only four variables: (1) area of heathland within 500 m, (2) length of borders between prethicket forestry and open land within 500 m, (3) total length of borders between major land cover categories within 1500 m, and (4) median altitude within 500 m. The ecological significance of these results was interpreted in terms of food availability, food preference, territoriality, and foraging and hunting behavior of the buzzard. A probability map of buzzard nesting sites based on model predictions showed a close agreement with observed patterns of nesting site distribution (Fig. 18). These results indicated that satellite imagery may help to predict the location of buzzard nesting sites over large geographic areas.

2. Case Study 10: Vertebrate Species Richness in Wyoming

This study used data derived from AVHRR and Landsat TM imagery to investigate factors affecting vertebrate species richness in Wyoming (Fraser, 1998). Previous studies of North American vertebrates have pointed to
FIGURE 17. Analysis of Acacia tree populations in the Negev desert using image processing of aerial photographs. (a) A section of an aerial photograph from 1996 with a corresponding map of individual tree distribution derived by computerized delineation of tree canopies. Only trees with a canopy >6 m² were delineated. (b) A section of an aerial photograph from 1956 with data on Acacia tree distribution. Trees larger than 6 m² were classified into three states: trees that were present in 1956 but absent in 1996 (died, open circle), trees that were present in both years (old, black circle), and trees that were present only in 1996 (recruited, gray circle). Note the magnitude of turnover in tree density and the patchiness in mortality and recruitment events. S. Lahav-Ginott and R. Kadmon, unpublished data.


density of energy as a factor controlling continental-scale patterns of species richness (the “species-energy” hypothesis). In contrast, data from local-scale studies have usually emphasized the importance of habitat heterogeneity as a determinant of species richness. This study focused on intermediate spatial scales, and its main purpose was to evaluate the relative importance of energy availability vs environmental heterogeneity as factors affecting mesoscale vertebrate species richness.

In general, parameters of energy availability (e.g., maximum July NDVI and forest biomass, both derived from AVHRR data) were less correlated with species richness than parameters of environmental heterogeneity (e.g., NDVI variability, forest biomass variability, and the number of land cover types as derived from Landsat TM data). Figure 19 shows the relationships between species richness and the most strongly related energy (maximum July NDVI) and heterogeneity (land cover variety) parameters. These results indicate that environmental heterogeneity may be more important than energy availability in determining mesoscale patterns of species richness.

It is interesting to compare these results with a study focusing on bird species diversity in Senegal (Nohr and Jørgensen, 1997). As with the Wyoming study, patterns...
FIGURE 19 Relationships between vertebrate species richness in Wyoming and two environmental indices based on satellite imagery: maximum July NDVI (derived from AVHRR data) and land cover variety (derived from Landsat TM data). Adapted from Fraser (1998), with permission from Blackwell Science Ltd.

of species richness were analyzed with respect to parameters of vegetation biomass (derived from AVHRR data) and landscape heterogeneity (derived from Landsat data). A regression model integrating the effects of biomass and landscape diversity (Fig. 20) accounted for 36% of the observed variation in bird species diversity (as expressed by Simpson's diversity index). Based on this model, a predictive map of bird species diversity was produced for an area of about 30,000 km² in northern Senegal (Fig. 21).

These studies demonstrate that information derived from satellite imagery can be used to investigate factors affecting large-scale patterns of species diversity.

D. Other Applications

The case studies presented above exemplify various aspects of biodiversity that can be investigated using remote sensing. Three additional aspects, not covered by these case studies, are mentioned here to provide a more complete picture of the subject. These are fire, grazing, and canopy gap dynamics.

Fire is a major source of disturbance in many kinds of ecosystems. Studies of fire distribution in space and time may therefore contribute to our understanding of the dynamics and structure of ecological communities. Remotely sensed data have been used to reconstruct fire histories, to estimate spatial and temporal characteristics of fire disturbances (e.g., patch size and frequency), to assess fire severity, to evaluate the susceptibility of different vegetation types to fire, and to measure rates of vegetation recovery following fires. Patterns of biomass burning were mapped using remotely sensed data for spatial scales ranging from local sites to the entire Earth. As one example, Fig. 22 presents a fire map constructed for western Africa by integrating data acquired by two imaging systems: the Defense Meteorological Satellite Program (DMSP) administered by the U.S. Air Force and the NOAA AVHRR radiometer. This map was used to evaluate the role of fire as a determinant of regional vegetation changes in Africa (Ehrlich et al., 1997).

Another aspect of biodiversity that has been studied extensively with remote sensing is grazing. Data obtained from remote sensing were used to investigate vegetation responses to both natural and domestic grazing. Some studies have focused on spatial patterns (e.g., differences between fenced and nonfenced areas), while other studies were designed to quantify temporal vegetation changes caused by grazing (e.g., biomass removal, replacement of herbaceous vegetation by woody species, or shrub encroachment). A study of vegetation responses to grazing is exemplified in Fig. 23. This study investigated the effect of cattle and goat grazing on long-term vegetation dynamics in a disturbed Mediterranean ecosystem. By integrating data on changes in vegetation cover derived from image processing of historical aerial photographs (Fig. 23), with corresponding maps of grazing type (cattle vs goats) and intensity (high, moderate, or low), it was possible to evaluate the effect of the grazing regime on long-term changes in the vegetation. Such analyses indicated that both cattle grazing and goat grazing reduce the rate of tree regeneration in this ecosystem and that the negative impact of goats on tree regeneration is stronger than that of the cattle.

Canopy gaps caused by the death of one or more trees are a major disturbance mode in many temperate and tropical ecosystems. Such disturbances may have considerable effects on the diversity of plant and animal communities. Traditionally, studies of canopy gap dynamics have focused on simple “static” parameters such
FIGURE 20  A predictive map of bird species diversity (expressed by Simpson's index) constructed for an area of about 30,000 km² in Senegal using the regression model presented in Fig. 20. Reproduced from Noth and Jorgensen (1997), with permission from Kluwer Academic Publishers.

FIGURE 21  Bird species diversity in Senegal plotted against an index combining information on landscape diversity (derived from Landsat TM data) and vegetation biomass (derived from AVHRR data). Adapted from Noth and Jorgensen (1997), with permission from Kluwer Academic Publishers.

as gap size distribution and total gap area. An alternative (and more direct) approach for studying gap dynamics is to analyze time sequences of remotely sensed data. Such an approach is exemplified in Fig. 24. This figure shows a map of canopy gaps generated for a temperate forest in Japan using digital analysis of aerial photographs from 1981 and 1986 (Tanaka and Nakashizuka, 1997). Data derived from such maps can be used to parameterize models of gap dynamics and to predict the equilibrium structure of the forest canopy.

IV. SUMMARY

Remote sensing in general, and Earth observation satellites in particular, have brought a new dimension to our understanding of the Earth’s surface. Since the launch of the first Landsat satellite in 1972, millions of images of the Earth have been acquired from spaceborne imaging.
systems. Earth observation satellites have provided images over spatial scales ranging from local regions to global cover, at spatial resolution of a few meters to tenths of a kilometer, and at spectral wavelengths ranging from near-ultraviolet to microwave radiation. Some of these data have been acquired systematically and repetitively over many years, with temporal resolution ranging from hours to a few weeks. In addition, most of the Earth observation satellites were designed to provide multispectral images of the Earth's surface. This enormous amount of information has been utilized extensively for studies of biodiversity. As demonstrated in this article, data obtained from remote sensing have been used for mapping various components of biodiversity, measuring different ecosystem characteristics (e.g., standing biomass and primary productivity), analyzing factors affecting the distribution of plant and animal species, monitoring natural and human-induced vegetation changes (e.g., succession and deforestation), and evaluating the impact of various disturbances (e.g., fire) and ecological interactions (e.g., grazing). The case studies presented in the previous section exemplify some of these capabilities.

It should be noted, however, that this review focuses on the two most common forms of remote sensing, namely, aerial photography and electro-optical satellite imagery. Two other forms of remote sensing that should be mentioned here are hyperspectral scanners and radars. Hyperspectral scanners are imaging systems that acquire multispectral images in many, very narrow, contiguous spectral bands throughout the visible, near-IR, and mid-IR portions of the spectrum. This spectral capability permits discrimination among objects that have fine diagnostic reflection characteristics that are "lost" within the bands of conventional electro-optical scanners. For example, AVIRIS (Airborne Visible/Infrared Imaging Spectrometer), one of the main hyperspectral scanners now in use, collects data in 224 contiguous bands between 0.4 and 2.45 μm.

Imaging radars have been used for remote sensing since the late seventies. However, most of these applications were of experimental nature. For example, the Shuttle Imaging Radar (SIR) carried by the space shuttle Columbia in its second flight was used to collect data from different continents in an attempt to construct a library of images representing a wide variety of environments differing with respect to climate, geology, vegetation, and other qualities. More recent studies have evaluated the potential of radar data for vegetation mapping, detection of deforestation, determination of vegetation biomass, and estimation of various forest parameters. Such applications have led to a better understanding of the capabilities of radar imagery and demonstrated that this form of remote sensing may complement many characteristics of images based on the optical spectrum.

There are many reasons to believe that in the near future much more information on biodiversity will be obtained from remote sensing. First, recent imaging systems are equipped with more advanced sensors. For example, the imaging system of SPOT 4 (launched on March 24, 1998) has been improved by adding a 20-m resolution band in the mid-IR portion of the spectrum.
FIGURE 23  Maps of changes in the percentage cover of herbaceous vegetation (a) and trees (b) between 1964 and 1992 in a Mediterranean ecosystem. Data on vegetation cover were obtained using image analysis of black-and-white aerial photographs. Changes in vegetation cover were calculated as \((\frac{1992 \text{ cover}}{1964 \text{ cover}}) - 1\). Boundaries of plots subjected to different intensities of cattle and goat grazing are superimposed on the maps. Y. Carmel and R. Kadmon, unpublished data.

SPOT 4 also carries a new imaging instrument called VEGETATION, which was designed to provide daily, global data of the continental biosphere. Landsat 7, the most recent satellite of the Landsat program (launched on April 15, 1999) carries a new, eight-band sensor called the Enhanced Thematic Mapper Plus (ETM+) that includes an additional panchromatic band (0.52–0.90 \(\mu\)m) with 15-m spatial resolution. Another advantage of the ETM+ is an increase in the spatial resolution of the thermal IR channel from 120 to 60 m. During the writing of this article (September 24, 1999), an imaging satellite called IKONOS was launched and opened a new generation of imaging capabilities. This satellite will provide panchromatic images at a spatial resolution of 1 m and multispectral images at a resolution of 4 m. These levels of spatial resolution are much higher than those of any other satellite data available today and approximate the resolution obtained by standard aerial photography. In parallel to these advances in space technology, new analytical
methodologies are being developed for the analysis and interpretation of high-resolution and hyper-spectral images. We can therefore expect a considerable growth in the contribution of remote sensing to studies of biodiversity during the next decade.

See Also the Following Articles
COMPUTER SYSTEMS AND MODELS, USE OF • MEASUREMENT AND ANALYSIS OF BIODIVERSITY • NATURAL RESERVES AND PRESERVES

Bibliography
REPTILES, BIODIVERSITY OF

F. Harvey Pough
Arizona State University West

1. Extant Evolutionary Lineages
2. Ancestral and Derived Characters
3. Characteristics of Groups
4. Conservation Issues

GLOSSARY

ancestral  Describes a character or character state of the organism being considered that retains the primitive condition for its evolutionary lineage.
derived  Describes a character or character state of the organism being considered that has changed from the primitive condition for its evolutionary lineage.
ectothermy  Deriving the energy needed to raise body temperature from sources outside the body.
dermatothermy  Deriving the energy needed to raise body temperature from within the body—i.e., from metabolic heat production.
paraphyletic  A taxonomic grouping of animals that does not meet the cladistic criterion of including the most recent common ancestor and all its descendants.
sister group  The evolutionary lineage most closely related to the one being discussed.

THE WIDESPREAD ADOPTION of phylogenetic systematics (cladistics) as the basis for taxonomic designations has had more impact on our understanding of reptiles than of any other group of terrestrial vertebrates. As our knowledge of ancestral and derived characters of reptiles has grown and been applied over the past decade, our perception of evolutionary relationships of lineages within the Reptilia has changed, and some lineages have even moved in and out of the Reptilia. Our current understanding of the relationships of extant reptiles recognizes three major monophyletic lineages (Pough et al., 2001): Turtles, crocodilians and birds, and lepidosaurs (with tuatara, lizards, and snakes as the extant representatives). Birds are derived from the lineage that includes crocodilians as well as dinosaurs, pterosaurs, and several other groups of extinct reptiles (Dingus and Rowe, 1997), but the anatomical and physiological specializations seen at the avian grade (most notably endothermy, insulation, and flight) make birds functionally different from other reptiles. As a result, the pre-cladistic allocation of nonavian reptiles to the field of herpetology and avian reptiles to ornithology is useful, and will be followed here. Nonetheless, the evolutionary origins of the characters of birds provide a useful context for considering the present and past diversity of other reptiles.

I. EXTANT EVOLUTIONARY LINEAGES

Reptiles, exclusive of birds, include about 7150 living species (Pough et al., 1990). The vast majority of these
(about 6850 species) are lepidosaurs (tuatara, lizards, and snakes). There are approximately 260 living species of turtles and 22 crocodilians. Species diversity is greatest in the tropics, but reptiles are not limited to warm regions—temperate deserts in the Holartic, Africa, and Australia have rich lizard faunas, and temperate aquatic habitats in North America and Asia are home to many species of turtles. Reptiles occur on every continent except Antarctica, occupy habitats ranging from freshwater swamps to desert salt flats, and exploit ecological zones from subterranean to arboreal.

A. Turtles (Testudines or Chelonia)

The shell that makes turtles instantly recognizable is formed by bone overlain by horny epidermal scales called scutes. The ribs, vertebrae, and parts of the pectoral girdle are fused to the dorsal shell (carapace), which is connected to the lower shell (plastron) by a bony bridge. The scutes and/or bone are secondarily reduced in some lineages of turtles, most notably the freshwater soft-shell turtles (Trionychidae) and the marine leatherback turtle (*Dermochelys coriacea*).

The earliest fossils of turtles are from the late Triassic, and the two major lineages, Pleurodira and Cryptodira, can be distinguished by the early Jurassic. Pleurodires have the common name side-necked turtles because they bend the neck horizontally when they retract their heads, whereas cryptodires bend their necks vertically. The geographic occurrence of pleurodires is currently restricted to the Southern Hemisphere (although fossil pleurodires are known from the Northern Hemisphere), and all of the approximately 65 species are aquatic. Pleurodires are the only turtles native to Australia and New Guinea and the only aquatic turtles in sub-Saharan Africa. The nearly 200 species of cryptodires have a worldwide distribution and include specialized marine and terrestrial forms as well as aquatic and semi-aquatic species. Tortoises (Testudinidae, about 40 species with a worldwide distribution in temperate and tropical regions) are the most terrestrial turtles. Most tortoises have domed carapaces, and many have forelimbs that are modified for digging. Semi-aquatic and aquatic turtles generally have low, streamlined shells and webbed feet. Most of the 33 species of emydid turtles fit this description. Emydids are found primarily in the New World, with one European species, the pond turtle *Emys*. The paraphyletic group *Bataguridae* is the Old World equivalent of the emydids, with about 50 species in Europe and Asia and one genus (*Rhinoclemys*, 9 species) in the Americas. The sea turtles (Cheloniidae [7 species] and Dermochelyidae [1 species]) are still more specialized for swimming, with forelimbs that are modified as flippers and short necks that cannot be retracted. Sea turtles have a worldwide distribution, primarily in tropical and subtropical regions although the range of the leatherback sea turtle extends well into the North Atlantic and South Pacific.

B. Crocodilians (Crocodylia)

Crocodilians are part of the archosaurian lineage that includes dinosaurs and birds as well as a number of other groups. Most crocodilians are large and have bony plates in the skin. These characteristics have contributed to an extensive fossil record that extends back to the early Jurassic. The history of crocodilians includes terrestrial species, a marine group (metriornithochids) with the forelimbs modified as paddles and a lobed tail, and species as large as tyrannosaur dinosaurs. Crocodilians are the smallest group of reptiles, with only 22 living species. Crocodilians are primarily inhabitants of tropical and subtropical regions, but the American and Chinese alligators (*Alligator mississippiensis* and *A. sinensis*) are found in temperate areas. All living crocodilians are aquatic; they swim with lateral undulations of the tail, which is somewhat flattened laterally. The nostrils of crocodilians are at the tip of the snout and are slightly elevated, as are the eyes. Alone among nonavian reptiles, crocodilians have developed a secondary palate that separates the passageway for air from the oral cavity. This combination of characters allows crocodilians to float with only the eyes and nostrils exposed.

Three lineages of extant crocodilians are recognized: The Alligatoridae includes seven species of freshwater crocodilians. All except the Chinese alligator occur in the New World. *Paleosuchus* (dwarf caimans) are among the smallest crocodilians at 1.7 m total length, and *Caiman* (6 m) is one of the largest. Alligatorids have relatively broad snouts and the teeth of the lower jaw fit into pits in the upper jaw and cannot be seen when the mouth is closed. Crododylids have narrower snouts than alligators, and the fourth tooth of the lower jaw of a crocodile is visible when the mouth is closed. The 13 species of crocodiles have a worldwide distribution, primarily in tropical and subtropical regions. Several species of crocodiles, especially the American crocodile (*Crocodylus acutus*) and the saltwater crocodile (*C. porosus*), regularly make long sea journeys, and the saltwater crocodile has colonized islands 1000 km from the nearest mainland. The saltwater crocodile is the
largest species, growing to a length of more than 7 m. The African dwarf crocodile, Osteolaemus tetraspis (2 m), is the smallest member of the family. The Gavialidae includes two species from Asia, the gharial (Gavialis gangeticus) and the false gharial (Tomistoma schlegeli). Both are specialized fish eaters with long, extremely narrow snouts.

C. Lepidosaurs (Lepidosauria)

More than 95% of living reptiles are lepidosaurs, and unlike turtles and crocodilians, this lineage displays enormous diversity in body form and ecology. Lepidosauria are covered by scales (the name means "scaly reptile") and shed the skin at regular intervals, they have a transverse cloacal slit (rather than the longitudinal slit seen in crocodilians and turtles), and they exhibit caudal autotomy that is facilitated by fracture planes within the vertebrae of the tail. Lepidosaurs include two lineages of extant forms, the Rhynchocephalia (the tuatara of New Zealand) and the Squamata (snakes and lizards).

Like their sister group the Archosauria, lepidosaurs are diapsids (i.e., primitively they have two fenestrae defined by bony arches on the sides of the skull). Modifications of the diapsid condition help to define subgroups within lepidosaurs and contribute to functional differences in prey handling.

1. Rhynchocephalia

The Rhynchocephalia, as it is currently understood (Gauthier et al., 1988), includes several extinct forms dating back to the Triassic and a single extant lineage, the Sphenodonidae, represented by two species of tuatara (Sphenodon punctatus and S. australis) found only on islands off the coast of New Zealand. (Tuatara is a Maori word meaning "spines on the back" and does not add an s to form the plural.) Tuatara are lizard-like in body form and reach a total length of about 30 cm.

Unlike their sister group the Squamata, tuatara retain the fully diapsid skull condition, with a lateral opening that is bounded dorsally by the postorbital and supratemporal bones and ventrally by the supratemporal and jugal. These bony connections impart a rigidity to the skull that allows tuatara to deliver a forceful bite. The single row of teeth on the lower jaw fits between two rows of teeth in the upper jaw—an outer row on the maxilla and an inner row on the palatines—and prey is chewed to a pulp before it is swallowed. Tuatara lack the paired hemipenes (intromittent organs) that characterize squamates.

2. Squamata

A large number of anatomical features characterize squamates. Prominent characteristics are reductions in the bony connections forming the arches in the skull, allowing a degree of kinesis unknown in other nonavian reptiles, and the presence in male squamates of paired hemipenes that develop as outgrowths of the posterior wall of the cloaca and are housed in the base of the tail. The hemipenes, which are ejected in use, are held in the female’s cloaca by spines, and sperm passes along a lateral groove.

Reduction or loss of limbs, usually accompanied by elongation of the trunk and tail, is a recurrent theme among squamates that reaches its zenith in snakes. Phylogenetically snakes are included within lizards, but the abundance and diversity of snakes combined with their morphological, ecological, and behavioral specializations make it helpful to discuss lizards and snakes separately.

a. Lizards

Fossils of lizards are known from the middle Jurassic, and the major lineages were distinct by the end of the Mesozoic. Lizards are found on every continent except Antarctica and are well represented on oceanic islands because large species, such as iguanas, can survive long journeys from the mainland on floating vegetation and small species (most notably skinks and geckos) often travel as stowaways with humans.

The more than 3000 species of extant lizards can be assigned to 19 families, three of which are very large: Scincidae (more than 1000 species), Iguanidae (more than 900 species), and Gekkonidae (870 species). Skinks and geckos have worldwide distributions, whereas only three genera of iguanids are found outside the New World (Brachylophus on Fiji, and Oplurus and Chalarodon on Madagascar).

Most species of iguanids are diurnal, visually oriented lizards whose activities and social behaviors can readily be observed, and they occur in areas accessible to biologists. As a result, iguanids figure prominently in ecological and behavioral studies of lizards. The family as it has traditionally been defined included several genera of large herbivorous lizards (the Iguanidae) and a variety of smaller, primarily insectivorous species that were placed in other subfamilies (Eutherid and de Queiroz, 1988). A review of anatomical characters elevated the subfamilies of Iguanidae to family status, leaving only the large herbivores in the Iguanidae (Frost and Etheridge, 1989). That reclassification had a broad impact in the biological literature, because the newly
Phylogenetically, snakes are included within lizards, and the most primitive snakes, aniliids and uropeltids, lack many of the derived characters that are responsible for the distinctive ecology and behavior of advanced snakes. Even primitive snakes have greatly increased the number of precloacal vertebrae, however, and have lost the bony connection between the postorbital and supratemporal bones, freeing the supratemporal to rotate on its fl at articulation with the parietal. This anatomical change allows the supratemporal and quadrates to act as part of the lower jaw, and these bones show varying patterns of elongation in advanced lineages of snakes. The brain of snakes is enclosed in a rigid box formed by ventral extension of the frontal and parietal bones and their articulation with the sphenoid bone.

Snakes are clearly derived lizards, but the sister group of snakes and the conditions that led to their evolution are still controversial topics. Snakes focus images on the retina by moving the lens, whereas lizards focus images by changing the shape of the lens. This difference, as well as differences in the photosensitive cells in the retinae of lizards and snakes, has long been interpreted as indicating that snakes are derived from a subterranean lineage of lizards with degenerate eyes. Other hypotheses of the origins of snakes are plausible, although they do not specifically account for differences in the eye. For example, legs are not particularly useful to small squamates that live in dense vegetation such as clumps of grass, and this is the habitat occupied by many legless lizards. This epigean adaptive zone might provide an intermediate stage in the evolution of snakes. Alternatively, aquatic habits could have been an intermediate stage in the evolution of snakes. Pachyrhachis problematicus, a recently described elongate aquatic lepidosaur from Cretaceous marine deposits, was over a meter long and had more than 100 presacral vertebrae and a skull that shows many of the derived features of snakes, including envelopment of the brain by bone (Caldwell and Lee, 1997).

Snakes were apparently widespread by the late Cenozoic, and about a dozen families are usually recognized, containing more than 2300 species. Primitive snakes, including boas and pythons, can be compared to more advanced forms, the Colubroidea, which include elapids (cobras, coral snakes, sea snakes, and their relatives), vipers (pit vipers and true vipers), and colubrids (the largest family, containing about 1700 species, which is 70% of the total number of extant species of snakes).

II. ANCESTRAL AND DERIVED CHARACTERS

The functional ecology of reptiles is shaped by a combination of ancestral and derived characters, and an appreciation of these features and the ways in which they interact is central to understanding how reptile function as organisms. Important derived characters include reproduction via a shelled egg (with the secondary development of viviparity in many lineages), excretion of nitrogenous wastes as urate salts, and a dry skin that is relatively impermeable to the passage of water. Retained ancestral characters include ectothermal thermoregulation (with the accompanying low metabolic rate and the absence of an insulating layer of hair or feathers) and an anatomically three-chambered heart (except in crocodilians) that permits intracardiac blood shunting.

A. Temperature and Water Relations

With the exception of a few very large species, nonavian reptiles are ectotherms. That is, they obtain the energy they need to raise body temperatures to activity levels from external sources, either directly from the sun or via contact with surfaces that have been heated by the sun. Reptiles employ a mixture of behavioral and physiological mechanisms of thermoregulation. Moving between sun and shade, orienting the body to maximize or minimize interception of solar radiation, and concentrating or dispersing melanin to change reflectivity adjust heat exchange in relation to body temperature. These mechanisms are supplemented by intracardiac shunting of blood and adjustments of the peripheral circulation to accelerate or retard exchange of heat with the environment. In contrast, birds and mammals are endotherms, changing the insulation provided by hair or feathers and adjusting metabolic rate to match rates of heat production and heat loss. Ectothermy imposes some limits on reptiles. When
heat sources are not available—at night for example, in water, or beneath the closed canopy of a tropical forest—reptiles must either accommodate their temperature requirements to the ambient temperature or cease activity. In the absence of environmental constraints, however, ectothermal thermoregulation can produce body temperatures that are as high and as stable as those of birds or mammals. Lizards and snakes, especially species that live in open habitats, generally control their body temperatures within activity temperature ranges of 2 or 3°C. Terrestrial turtles, too, control their body temperatures by moving between sun and shade during periods of activity. The large body sizes of some turtles and crocodilians stabilize their body temperatures. Leatherback sea turtles, which weigh up to 1000 kg, are found off the coasts of New England and Nova Scotia in the summer in water temperatures as low as 8°C. A countercurrent exchange mechanism in the flippers conserves the heat produced by muscular activity as the turtles swim, and body temperatures can exceed water temperature by at least 18°C (Spotila and Standora, 1985).

The dry skin of reptiles is an essential component of their ectothermal thermoregulation. In contrast, the high rate of evaporation from the moist skins of extant amphibians limits their capacity to maintain body temperatures above ambient levels as well as their ability to be active during the day when evaporative stress is high. Reptilian water balance is additionally facilitated by enzymatic pathways that convert nitrogenous waste to uric acid, which is excreted as a complex mixture of urate salts. The shelled eggs of reptiles absorb water from the substrate during development, and the availability of water in the nest environment influences the body size and, probably, the viability of newly hatched reptiles.

B. Metabolism and Energetics

Because ectotherms do not depend on metabolic heat production for thermoregulation, their resting metabolic rates are low—one-seventh to one-tenth those of endotherms of the same body size. The influence of low resting metabolic rates can be seen in the energy use, body size, and the exercise physiology and foraging behavior of reptiles.

1. Energy Requirements and Conversion Efficiency

The daily energy budget for a lizard is only 3% of the energy requirement for a mammal of the same body size (Bennett and Nagy, 1977). Low energy use by the lizard results partly from its low resting metabolic rate during the day when it is active and partly from a reduction in body temperature (with a consequent further reduction in metabolic rate) at night when it is inactive. In contrast, a mammal maintains a higher metabolic rate during the day and may increase its metabolic rate above resting levels at night to counteract the effect of falling ambient temperature.

Low metabolic rates translate to low energy requirements for reptiles. As a result, reptiles can cope with environments in which food is limited, either episodically or permanently. Areas of low primary production, such as deserts, and even regions where there is no primary production, such as desert salt flats and mobile sand dunes, are home to a variety of lizards.

Efficient conversion of food to biomass is a second consequence of the low metabolic rates of reptiles. Endotherms use most of the energy they ingest to stay warm and be active—19 species of birds and mammals have an average net conversion efficiency of 1.4% (range 0.5–3%), whereas 9 species of reptiles have an average net conversion of 46.6% (range 18–86%) (data from Pough, 1980).

2. Body Size

Mass-specific metabolic rates (W/g body mass) are related to body mass by an exponent that approximates −0.25, although statistically significant variation from that value is associated with both phylogeny and ecology (Hayssen and Lacey, 1985; Andrews and Pough, 1985). The consequence of that negative allometry is a mass-specific energy requirement that increases steeply as body size decreases (Fig. 1).

The low metabolic rate of reptiles makes extremely small body size energetically feasible, and lizards, in particular, include very small species. The modal body mass of lizards falls in the range 1–10 g, and nearly 20% of lizards have adult body masses less than 1 g. In contrast, there are no mammals with adult body masses less than 1 g, and fewer than 5% of mammals have body masses below 10 g (see Pough et al., 1990, p. 483). More than half the living species of lizards occupy a zone of body sizes in which there are practically no mammals, and the smallest lizards have their major competitive and predatory interactions with invertebrates. In contrast to lizards, snakes are relatively large reptiles—nearly 80% of snakes are larger than 20 g. Thus, snakes and lizards illustrate two different ways to succeed as a squamate in a world that is largely dominated by endotherms: Small body size is a morphological specialization of lizards, just as elongation is a morphological specialization of snakes.
FIGURE 1 The mass-specific energy requirements of animals increase at small body sizes. The metabolic rates of reptiles are one-seventh to one-tenth those of birds or mammals of the same body size, and the smallest reptiles have adult body masses less than 1 g, whereas very few birds and mammals are smaller than 10 g as adults. (Modified from Pough, 1980. © 1980 by The University of Chicago.)

3. Exercise Physiology and Behavior

The low resting metabolic rates of reptiles limit their aerobic metabolic capacity—i.e., the maximum rates of oxygen consumption they can achieve during strenuous activity. Both reptiles and mammals increase resting rates of oxygen consumption about 10-fold when they run at maximum speed, so the differential in metabolic rates of ectotherms and endotherms persists during activity. Most reptiles have limited aerobic metabolic capacity, and they derive most of the ATP used during high levels of activity from anaerobic pathways. Glycogen, the metabolic substrate for anaerobic metabolism, is stored in the muscle cells and a brief period of activity can exhaust the muscles’ glycogen supply. As a result, anaerobic metabolism is an effective way to support bouts of intense muscular activity, but it is not suitable for sustained behaviors such as long-distance locomotion. The activity patterns of many reptiles are built on this relationship—they employ sit-and-wait modes of foraging, they escape predators by fleeing to a nearby shelter, and their social behaviors do not include sustained high levels of activity.

C. Active Reptiles and the Evolution of Endothermy

In an ecological context, the generalization that reptiles are ectotherms with low metabolic rates and low levels of activity applies to extant species, although some modifications of that characterization will be addressed in subsequent sections. In an evolutionary context, however, the generalization is patently false because birds are reptiles and they are endotherms with high metabolic rates. Clearly the reptilian lineage has a capacity for endothermy that is barely expressed in nonavian reptiles. An examination of the evolution of endothermy explains that dichotomy and emphasizes how tightly anatomical and physiological characteristics are linked to thermal ecology.

Endothermy is the ancestral condition for vertebrates, and the derived condition of whole-body endothermy has evolved at least twice, in mammals and in birds. In a broader view, regional endothermy has evolved independently in sharks, tunas, and billfishes, and whole-body endothermy may have been characteristic of pterosaurs (flying archosaurian reptiles of the Mesozoic) and some lineages of dinosaurs. Endotherms and endotherms have very different relationships to their physical environments: Ectotherms rely primarily on behavioral thermoregulation to raise their body temperatures because they have low metabolic rates, and the absence of insulation facilitates uptake of heat from the environment. In contrast, endotherms use internal heat production from high metabolic rates to regulate body temperature, and they require insulation to retain metabolic heat in their bodies.

The evolutionary transition from ectothermy to endothermy is impeded by a Catch-22—adding insulation to an ectotherm impedes its behavioral thermoregulation, but in the absence of insulation any heat produced by increasing its resting metabolic rate is lost to the environment. The solution to this paradox lies in finding a basis for the evolution of insulation or the evolution of an increased metabolic rate that does not depend on the preexisting occurrence of the other character. Mammals are the sister group of reptiles (including birds) and the common ancestor of mammals and reptiles was ectothermal. Thus, the evolution of endothermy in the mammalian lineage may provide a model for the evolution of endothermy among reptiles. Anatomical changes seen in the fossil record of predatory synapsids, the sister group of mammals, strongly support the hypothesis that the initial step in the evolution of mammalian endothermy was selection for increased locomotor capacity. These changes include the evolution of a cursorial body form, changes in the rib cage that suggest the presence of a diaphragm, and increased surface area in the nasal passages to warm and humidify large volumes of air. Increasing levels of locomotor
activity require an increase in metabolic rate, and internal heat production would create a selective value for insulation (see Pough et al., 1999, Chapters 8 and 19, for details and references).

Some features of the fossil record of birds suggest that a similar scenario can be applied to the evolution of avian endothermy, but others appear to contradict that interpretation. Like the predatory synapsids, the small maniraptoran dinosaurs that form the sister group of birds appear to have been fleet-footed predators that pursued their prey. That interpretation suggests that these dinosaurs may have evolved the metabolic capacity for endothermy just as synapsids did, and if the recent report of feathers in fossil dinosaurs is correct, it would support that interpretation. However, two lines of evidence cast doubt on the hypothesis that the dino-saurian precursors of birds had high metabolic rates. Examination of an excellently preserved specimen of the small dinosaur Sinosauropteryx suggests that it had simple septate lungs that were ventilated by a pistonlike movement of the liver like those of living crocodilians. Lungs of this sort would not support high rates of oxygen consumption. That interpretation is supported by CAT scans of the nasal passages of dinosaurs that reveal no trace of modifications of the nasal passages to warm and humidify large volumes of air (see Pough et al., 1999, Chapter 13, for details and references).

This evolutionary perspective emphasizes intricate interconnections among the anatomical and physiological characters of extant reptiles and their ecology and behavior, as well as the evolutionary capacity for breaking those links. An examination of living reptiles reveals additional connections among anatomy, physiology, ecology, and behavior.

### III. CHARACTERISTICS OF GROUPS

Evolution proceeds in a mosaic fashion; all descendant lineages inherit ancestral characteristics, and each lineage develops a set of derived characters that interact with ancestral characters and with each other to shape the way animals live. The following sections focus on combinations of characters found in specific groups of reptiles that are important in shaping their ecology and behavior. Temperature-dependent sex determination is widespread among turtles, for example, and crocodilians show extensive parental care. To avoid repetition, examples from other lineages are included when a topic is discussed. Additional details, examples, and references can be found in Pough et al. (1999), Pough et al. (2001), and the sources cited.

#### A. Turtles

**1. The Nest Environment and the Size and Sex of Hatchlings**

All turtles lay eggs in holes they excavate with their hind legs. Physical conditions within the nest determine both the size and sex of hatchlings of many species of turtles (Packard and Packard, 1988). In general, embryos in nests constructed in moist substrates metabolize most of their yolk before they hatch and grow larger than embryos in drier nests. Hatchlings of snapping turtles (Chelydra serpentina) from moist nests run and swim faster than the smaller hatchlings that emerge from dry nests. This difference in locomotor performance, which persists for at least a year after hatching, might result in higher first-year survival of hatchlings from moist nests.

Temperature-dependent sex determination is widespread among both pleurodira and cryptodires. For most species of turtles, low temperatures during incubation produce males (Pattern Ia), but in a few species males are produced at high temperatures (Pattern Ib), and some species show a biphasic response in which females develop at both low and high temperatures and males are produced at intermediate temperatures (Pattern II). Environmental sex determination can produce both male and female hatchlings from the same nest, because the shift between sexes typically occurs within a range of only 1 or 2°C. Thus, eggs from the top of a nest can develop into females while eggs deeper in the same nest produce males, or clutches laid early in the season when nest sites are exposed to the sun can produce females, whereas clutches laid in the same sites later in the summer produce males because the nests are now shaded by vegetation.

Temperature-dependent sex determination is universal among crocodilians and tuatara and is known for some iguanid, agamid, gekkonid, and lacertid lizards. No snake has been shown to have temperature-dependent sex determination.

#### 2. Gigantism

Giant tortoises are among the most spectacular reptiles. The largest species of living tortoise are found on the Galápagos and Aldabra Islands and reach carapace lengths of more than a meter and body weights exceeding 100 kg, but large tortoises are not limited to islands. Extant members of the genus Geochelone have a pantropical distribution, with mainland representatives in Africa, India, Asia, and South America as well as additional island species on Madagascar, Sri Lanka, and Indonesia (Ernst and Barbour, 1989; Pritchard, 1996). The largest living mainland species, the African
spur-thighed tortoise *G. sulcata*, reaches a carapace length of 76 cm, and fossil tortoises with carapace lengths from 60 cm to 1.5 m are known from Pleistocene deposits on all of the continents except Australia and Antarctica. Giant tortoises were present in North America when humans arrived, and a subfossil carapace from Florida contains a spear point and is charred as if it had been placed upside down on a fire and cooked.

**B. Crocodilians**

1. Reproduction and Parental Care

All crocodilians are oviparous, laying eggs either in nests that the female excavates in the ground or in mounds of vegetation that she builds. Crocodilians in some areas are reported to lay eggs on floating mats of vegetation. Parental care is probably universal among extant crocodilians, although the reproductive behavior of some species is unknown and parental care appears to have been reduced in populations that have been hunted extensively (Lang, 1989). Females are the primary caregivers, but males often participate. A female crocodilian remains near her nest while the eggs are incubating. As the young approach hatching, they start to vocalize within the eggshells, and these vocalizations signal the female to begin opening the nest. The male may participate in opening the nest, and both parents pick up hatchlings in their mouths and transport them to water. Adult crocodilians show remarkable dexterity in this process, sometimes picking up an intact egg and cracking the shell to release the hatching embryo. Parental care is virtually unknown among turtles, but about 100 species of squamates exhibit some degree of parental care (Shine, 1988; Gans, 1996). These behaviors range from defense of the nest site to brooding the eggs. In an exception to the generalization that reptiles are ectotherms, females of some species of python gather their eggs into a clam enclosure by coils of the body and use muscular contractions to keep the temperature of the eggs close to 32°C. The frequency of contractions, and thus the python's rates of oxygen consumption and heat production, increases as ambient temperature falls. A female Indinan python (*Python molurus*) can maintain her egg mass at 31–32°C at ambient temperatures between 23 and 32°C, and during this period she is functioning as an endotherm.

**C. Lepidosaurs**

1. Tuatara

Two species of tuatara are known, *Sphenodon punctatus* and *S. guntheri*, but the existence of *S. guntheri* was largely forgotten after it was described in 1877 until it was rediscovered by a molecular analysis more than a century later (Daugaard et al., 1990). The entire population of *S. guntheri* consists of about 300 individuals living on 1.7 hectares of scrub on the top of North Brother Island. *S. punctatus* is represented by about 30 populations on small islands off the coast of New Zealand.

Tuatara have a number of unusual ecological and physiological characteristics. They are crepuscular and forage on cool, foggy nights, with body temperatures as low as 6°C. During the day tuatara bask in the sun and raise their body temperatures to 28°C. Tuatara feed on invertebrates and nestling shearwaters. Prey is crushed as the lower jaw closes with an initial vertical movement and is then sheared by the interdigitating rows of triangular teeth as the lower jaw slides forward. Tuatara, unlike other lizards, lack hemipenes. Sperm is transferred by apposition of the cloacas of the male and female. In late spring, female tuatara excavate shallow depressions in which they deposit 6–15 eggs. Embryonic development requires 12–15 months (Saint Girons, 1985).

Tuatara are the sister group of squamates, and their anatomy shows characters that are probably ancestral for squamates, including the connection between the jugal and supratemporal bone forming the lower border of the lateral arch of the skull. Some authors have used the phylogenetic position and primitive anatomical characters of tuatara as a basis to suggest that their ecology and physiology also reveal ancestral conditions for squamates, but that is probably not the case. New Zealand is a cold place, and the native geckos and skinks are active at body temperatures well below those of geckos and skinks in warmer parts of the world. It seems most likely that the low activity temperatures of tuatara and the slow development of their eggs are derived characters that reflect current ecological conditions.

2. Squamata

Lizards have proven to be ideal study animals. Many lizards are diurnal and have limited home ranges, so their activities can readily be observed. As a result of their ease of study, species of *Anolis, Sceloporus, Pogona, Urosaurus*, and *Uta* have played seminal roles in the fields of animal behavior, ecology, and evolutionary morphology and physiology (Greer, 1989; Vitt and Pianka, 1994; Wainwright and Reilly, 1994). Most snakes, in contrast, are secretive in their behavior and direct observation is much more difficult. Nonetheless,
the combination of perseverance with radiotelemetry has recently produced several important field studies (Shine, 1991; Seigel and Collins, 1993; Greene, 1997).

a. Lizards

Classifying species of lizards in terms of the foraging modes they employ provides a framework that organizes diverse aspects of their biology into a series of useful generalizations (Table I). Like other mobile animals, lizards display a spectrum of foraging behaviors that extends from species that perch motionless on lookouts and make short dashes to capture prey to species that are in nearly continuous motion as they forage, searching for hidden prey under leaf litter and in holes and cavities. Species at the inactive end of this spectrum are called sit-and-wait predators, those in the middle are cruising foragers, and the species that move continuously are called active foragers.

### Table I
Ecological and Behavioral Characteristics Associated with the Foraging Modes of Lizards

<table>
<thead>
<tr>
<th>Character</th>
<th>Sit-and-wait</th>
<th>Cruising forager</th>
<th>Widely foraging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foraging behavior</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Movements/hour</td>
<td>Few</td>
<td>Intermediate</td>
<td>Many</td>
</tr>
<tr>
<td>Distance traveled/hour</td>
<td>Small</td>
<td>Intermediate</td>
<td>Large</td>
</tr>
<tr>
<td>Primary sensory mode</td>
<td>Vision</td>
<td>Vision and olfaction</td>
<td>Olfaction</td>
</tr>
<tr>
<td>Exploratory behavior</td>
<td>Low</td>
<td>Intermediate</td>
<td>High</td>
</tr>
<tr>
<td>Prey and prey</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Types of prey</td>
<td>Mobile, large</td>
<td>Intermediate</td>
<td>Sedentary, often small</td>
</tr>
<tr>
<td>Types of predators</td>
<td>Widely foraging</td>
<td>?</td>
<td>Sit-and-wait and widely foraging</td>
</tr>
<tr>
<td>Risk of predation</td>
<td>Low</td>
<td>?</td>
<td>High</td>
</tr>
<tr>
<td>Mode of escape</td>
<td>Cryptic, flee a short distance</td>
<td>?</td>
<td>Conspicuous, may flee a long distance when attacked</td>
</tr>
<tr>
<td>Physiological characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprint speed</td>
<td>High</td>
<td>?</td>
<td>Low</td>
</tr>
<tr>
<td>Endurance</td>
<td>Low</td>
<td>?</td>
<td>High</td>
</tr>
<tr>
<td>Aerobic metabolic capacity</td>
<td>Low</td>
<td>?</td>
<td>High</td>
</tr>
<tr>
<td>Anaerobic metabolic capacity</td>
<td>High</td>
<td>?</td>
<td>Low</td>
</tr>
<tr>
<td>Heart mass</td>
<td>Low</td>
<td>?</td>
<td>High</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>Low</td>
<td>?</td>
<td>High</td>
</tr>
<tr>
<td>Energetics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily energy expenditure</td>
<td>Low</td>
<td>?</td>
<td>High</td>
</tr>
<tr>
<td>Daily energy intake</td>
<td>Low</td>
<td>?</td>
<td>High</td>
</tr>
<tr>
<td>Social behavior</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social system</td>
<td>Territorial</td>
<td>?</td>
<td>Not territorial</td>
</tr>
<tr>
<td>Home range</td>
<td>Small</td>
<td>Intermediate</td>
<td>Large</td>
</tr>
<tr>
<td>Body form</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trunk</td>
<td>Stocky</td>
<td>Intermediate</td>
<td>Slim, elongate</td>
</tr>
<tr>
<td>Tail</td>
<td>Often short</td>
<td>?</td>
<td>Often long</td>
</tr>
<tr>
<td>Color and pattern</td>
<td>Often bleached, may match background, rarely colorful</td>
<td>?</td>
<td>Often solid or striped, sometimes colorful</td>
</tr>
<tr>
<td>Reproduction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mass of clutch relative to mass of female</td>
<td>High</td>
<td>?</td>
<td>Low</td>
</tr>
<tr>
<td>Viscarity</td>
<td>Often</td>
<td>Sometimes</td>
<td>Rare</td>
</tr>
</tbody>
</table>

*Foraging modes are presented as a continuum from sit-and-wait to widely foraging species. In most cases, data are available only for species at the extremes of the spectrum. (Modified from Pough et al., 1990.)
Sit-and-wait lizards rely on vision to detect moving prey and may be selective in deciding which prey to attack. Widely foraging lizards examine places that may be hiding; olfaction is an important sensory mode for these species, and they usually take any prey items they encounter.

Sit-and-wait lizards are usually cryptic and are most likely to be detected by widely foraging predators, whereas widely foraging lizards are conspicuous because of their movement and are consequently vulnerable to sit-and-wait predators. Thus, there may be an alternation of predatory modes at successive levels in a food chain—widely foraging predators find sedentary prey and are eaten by sit-and-wait predators. In the American Southwest, for example, whiptail lizards (Cnemidophorus) forage widely and locate sedentary insects such as termites. Leopard lizards (Gambelia wislizeni) wait motionless beneath bushes and capture whiptail lizards when they come within range. In turn, leopard lizards are captured by coachwhip snakes (Masticophis flagellum), which move from bush to bush in search of prey.

The quick dash that a sit-and-wait lizard makes to capture prey requires high sprint speed, whereas the continuous movement of a widely foraging predator requires endurance. These characteristics are reflected in the physiology of lizards—sit-and-wait lizards can sprint rapidly but they rely primarily on anaerobic metabolism and have limited endurance. Widely foraging lizards cannot sprint so fast, but they have relatively larger hearts and higher hematocrits than sit-and-wait species, so they have greater aerobic metabolic capacities and greater endurance.

The colors of sit-and-wait lizards often match the backgrounds on which they rest and their patterns obscure the outlines of the body, making them cryptic when they are motionless. Widely foraging lizards move across a variety of backgrounds and do not necessarily match any one. They are often solid colored or striped, and these patterns may confuse predators when a lizard is moving. Sit-and-wait lizards are exhausted by brief periods of rapid locomotion, and when they are attacked, they sprint to shelter. Widely foraging lizards often run from bush to bush, changing direction when they are out of sight and traveling a long distance before they stop. The generally long tails of widely foraging lizards may facilitate caudal autotomy as a last-ditch means of escape.

Sit-and-wait lizards can survey their surroundings from their perches whereas widely foraging lizards can see little beyond their immediate horizon. Sit-and-wait species have small home ranges and are often territorial, whereas widely foraging species have larger home ranges and are rarely territorial. The bulk of eggs within the oviducts is probably a handicap for a widely foraging lizard, and widely foraging species have smaller ratios of clutch mass to maternal mass than do sit-and-wait species. Vitruparity is more common among sit-and-wait lizards than among widely foraging species.

These generalizations organize a large number of empirical observations into a coherent whole, but assigning evolutionary cause-and-effect interpretations to the correlations may confound phylogeny with ecology. Most of the sit-and-wait species that have been studied are in the iguanian lineage, whereas the widely foraging lizards are scincophids. Thus, we cannot separate the effects of phylogeny (iguanian versus scincophid) from those of ecology (sit-and-wait versus widely foraging). However, a study of several closely related species of lacertid lizards identified sit-and-wait and widely foraging species and detected many of the same behavioral and physiological correlations that are found in the broader comparison (Huyn et al., 1984).

Because many species of lizards use behavioral mechanisms to regulate their body temperatures within narrow limits while they are active, the thermal environment plays a central role in defining the niche of a species. The thermoregulatory characteristics of habitats can be defined as high cost or low cost, depending on the behavioral and ecological trade-offs a lizard must make to maintain a body temperature that is different from ambient temperatures. Lizards that live in habitats in which basking sites, perches, and escape sites are abundant and close to each other have low ecological and behavioral costs of thermoregulation and are likely to maintain body temperatures within a narrow range during activity. In contrast, lizards that live in habitats where basking sites are scarce or ephemeral, such as the understory of a tropical forest, have high costs of thermoregulation. These lizards often have body temperatures indistinguishable from ambient temperature.

The costs of thermoregulation in a particular habitat are not the same for all lizards because body size profoundly affects the thermoregulatory options available to a lizard. An example is provided by three species of Anolis that live in microsympathy on the Osa Peninsula of Costa Rica. The three species partition the structural habitat on the basis of the radiant environment. Adults of the smallest species (A. quadrifrenata) weigh about 10 g and forage in short vegetation in full sun, whereas the middle species (A. festiva) weighs about 32 g and forages in areas where taller vegetation provides broken shade. The largest species (A. leiptophrys) weighs...
Pseudechis and blacksnakes (lengths that are 10–14 times the maximum trunk circumference) have more than 400 precaudal vertebrae. Even generalized snakes—rat snakes (Elaphe)—have head-plus-trunk lengths that are 10–14 times the maximum trunk circumference. In contrast, elongate lizards (Cnemidophorus, Egerinia, Lacerta, and Eumeces) have length/circumference ratios of 2.0–2.5.

In a functional sense, repackaging the mass of a lizard into a long, thin body form means that a snake must feed a large body through a small mouth. Lizards and snakes almost always swallow their prey whole; many lizards manipulate prey items in their mouths until they have been softened by repeated bites, whereas nearly all snakes swallow prey without any oral processing. Most legless lizards eat large numbers of small prey items; they retain the connection between the postorbital and supratemporal bones on the side of the skull, and this structure gives the skull sufficient rigidity to crush the exoskeletons of insects. In contrast, snakes have lost the postorbital-supratemporal connection and have replaced the body symphysis at the front of the lower jaw with a flexible ligament. These changes, which increase skull kinesis and allow the lower jaws to move independently, also reduce the ability of snakes to exert pressure on food in the mouth. A few snakes have secondarily evolved modifications of the skull and jaws that allow them to eat hard-bodied prey, but most snakes feed on soft prey.

Lizards normally consume several small prey items on a daily basis, whereas most snakes eat larger prey items less frequently. The prey of snakes is large relative to the predator, and many snakes have specializations that allow them to immobilize prey before it can escape or injure the snake. Snakes often swallow prey items with diameters larger than the snake’s head, and these snakes are capable of large gapes. Prey size must be evaluated relative to the size of the snake in two respects—mass and diameter. That is, a prey item can be light or heavy compared to the snake in two respects—mass and diameter. Insects are Type I prey, and this is the prey type that is characteristic of most lizards and probably of the earliest snakes.

• Type I prey have both small mass and small girth ratios. Insects are Type I prey, and this is the prey type that is characteristic of most lizards and probably of the earliest snakes.

b. Snakes

Elongation of the body is the immediately distinctive feature of snakes, and specializations associated with feeding form the core of their diversity. The body form of snakes results primarily from an increase in the number of vertebrae in the neck and trunk, and the longest species have more than 400 precaudal vertebrae. Even generalized snakes—rat snakes (Elaphe), gopher snakes (Pituophis), Australian whipsnakes (Demansia), and blacknakes (Pseudechis)—have head-plus-trunk lengths that are 10–14 times the maximum trunk circumference.
snakes, *Cylindrophis* and *Anilius*) are limited to Type II prey.

- Type III prey are heavy and bulky and have large mass and girth ratios. Frogs, mammals, and some lizards are Type III prey. These prey require both immobilization and large gaps.
- Type IV prey are bulky and light, with small mass ratios and high girth ratios. Birds (which are bulky like their feathers) are Type IV prey and require gape specializations but not immobilization. Boas and pythons take Type III and IV prey, as do most of the advanced snakes (elapids, vipers, and colubrids).

The foraging behavior of snakes extends from active searchers that use visual or olfactory cues to find prey to mobile ambushers that move to places they are likely to encounter prey. Rattlesnakes (*Crotalus*), for example, coil beside trails followed by rodents and wait for prey, and tropical racers (*Mastigodryas*) wait beside sunspots on the forest floor to capture lizards (*Anolis*) that pause to bask.

The exercise physiology of snakes appears to parallel differences in their foraging behavior. Widely foraging species, such as North American whipsnakes (*Masticophis*), have high aerobic metabolic capacities and high endurance, whereas species that rely on ambush (*Licha-nata*) have lower aerobic capacities and endurance. Widely foraging species probably eat more often than ambushers and take smaller prey, but even a frequent feeder such as the coachwhip snake (*Mastigodryas*) waits beside sunspots to capture lizards (*Anolis*) that pause to bask.

Type III prey (heavy and bulky) can pose a risk of injury to a snake, especially those species that swallow prey while it is still struggling. Constriction and venom are specializations that allow snakes to subdue prey before they are swallowed. Constriction is characteristic of primitive snake lineages and of boa and pythons. Apparently it evolved as a method of controlling elongate prey and subsequently proved effective with bulky prey as well. Constriction requires enclosing a prey item with coils of the snake’s body. Little pressure is exerted on the prey, and friction between adjacent coils holds them in place. Death may result from asphyxiation (because the snake takes up the slack each time the prey exhales) or because increased internal pressure interferes with and ultimately stops the action of the heart. The short-radius curves required for constriction require short vertebrae and trunk muscles that span only a few vertebrae. This morphology is not compatible with rapid locomotion, which requires large-radius curves formed by long vertebrae and muscles that span many vertebrae. The evolutionary appearance of advanced snakes in the Miocene coincides with the spread of grasslands. Constrictors (largely small boas) dominated the snake fauna at the beginning of the Miocene, but by the end of that epoch the snake fauna was composed largely of colubroids. These were fast-moving snakes, but their locomotor specializations precluded constriction. Although some lineages of colubroids secondarily evolved constriction, venom is the distinctively colubroid method of immobilizing prey with minimum risk to the snake.

Venom evolved early in the history of colubroids as part of the feeding apparatus. The defensive use of venom is a secondary development, enhanced in some cases by morphological specializations. The venom channel in the fangs of spitting cobras, for example, makes an abrupt right-angle turn near the tip of the fang that enables the cobra to emit a spray of venom that travels a meter or more. The snake aims for the predator’s eyes, and accounts by people who have been sprayed confirm the powerful deterrent effect of this defense. Attacking a venomous snake can have severe consequences for a predator, and the warning mechanisms of venomous snakes (such as the rattle of rattlesnakes, the hood of cobras, and the distinctive colors of coral snakes) foster both learned and innate avoidance behavior by potential predators (Greene, 1988; Pough, 1988).

Duvernoy’s gland, which lies at the rear of the maxilla and produces venom that immobilizes prey, is a derived character of colubroids. The embryonic development of Duvernoy’s gland is linked to development of the posterior pair of maxillary teeth. These rearmost teeth are enlarged and grooved in many colubrids. Although colubrids are often called ‘the nonvenomous snakes’ in popular literature, many colubrids use venom to subdue prey, and a few species known as rear-fanged snakes (notably the African boomslang, *Dispholidus typus*) can deliver defensive bites that are lethal to animals as large as humans. In the two groups popularly known as venomous snakes, elapids (cobras, coral snakes, sea snakes, and their relatives) and vipers (vipers and pit vipers), the maxilla has shortened, thereby moving the enlarged fang on the maxilla to the front of the mouth. Elapids have short fangs in a relatively immobile maxilla, whereas vipers have long fangs and the maxilla rotates so the fang lies horizontally when the mouth is closed.

The venoms of colubrids and elapids attack the nervous system and rapidly immobilize prey. In contrast,
the venoms of vipers and pit vipers have high proteolytic activity and cause tissue destruction. The proteolytic activity of venom speeds digestion because the venom hydrolyzes the body tissues of the prey from the inside while the snake's digestive enzymes attack from the outside. Some vipers, especially species of the African genus Bitis, appear to combine proteolytic venom, injected deep into the prey by very long fangs, with modifications of the body form to feed on very bulky prey. Large species of Bits are remarkably heavy-bodied—length/circumference ratios of the puff adder (B. arietans) and Gaboon viper (B. gabanica) are 4–5. The thick bodies of these snakes accommodate bulky prey without disrupting locomotion or interfering with the function of internal organs.

The physiology of the digestive system of snakes is as specialized as their trophic anatomy (Secor and Diamond, 1998). Large meals taken at infrequent intervals pose a challenge—the physiological cost of digesting a meal is high, but occasional periods of high activity are separated by long periods during which the digestive system is inactive. Within 24 h after feeding, snakes display remarkable factorial increases in rates of oxygen consumption; the masses of organs in the cardiovascular, respiratory, digestive, and excretory systems; and activities of digestive enzymes. Between feedings, the sizes of organs and activities of enzymes regress to a resting state. This cyclic hypertrophy and regression, which is interpreted as a mechanism that conserves energy compared to the alternative of maintaining the physiological systems continuously at high levels, emphasizes the extent to which snakes have developed suites of trophic specializations not seen among lizards.

IV. CONSERVATION ISSUES

Many of the causes of declining populations of reptiles are depressingly familiar and affect other kinds of animals and plants as well. Habitat loss as a result of human activities is probably the major reason for shrinking populations, although mortality on roads may be more significant than is generally recognized. Road kills are considered the major form of human-induced mortality for Australian reptiles and amphibians, exceeding the effect of habitat loss by nearly 70% (Ehmann and Cogger, 1985). Some European countries have built drift fences and constructed tunnels beneath roads to reduce mortality of reptiles and amphibians. Introductions of exotic species (especially on islands) and pollution (potentially including global warming, [Dunham, 1993]) are additional factors that affect animals and plants on a large scale. Other threats are limited to particular groups and result from combinations of human activities and specific biological characteristics.

A. Turtles

Many species of turtles are long-lived and display a suite of coevolved traits, including delayed sexual maturity and high annual survivorship of adults (Congdon et al., 1994). Adult turtles are often nearly immune from nonhuman predators, and a turtle that survives to adulthood can expect a long reproductive period. When adult mortality is increased by human activity, as is the case for sea turtles which are drowned at sea in the nets of shrimp trawlers and killed for food when they come ashore to nest, a crucial component of the life history strategy is destroyed. Sea turtles are the most visible examples of excessive adult mortality resulting from human activities, but automobiles kill adult turtles and tortoises all over the world.

Well-intentioned efforts at management can founder on aspects of the biology of a species. Sea turtles, for example, exhibit temperature-dependent sex determination—males are produced at low temperatures and females at higher temperatures. Conservation activities at the nesting beaches of sea turtles usually include moving the eggs from natural nests to a protected site where they are buried in the sand to complete incubation. Under these circumstances all of the eggs are exposed to similar temperatures, and as a result the hatchlings will be predominantly one sex. Even worse, in some cases the incubation conditions used by management programs produced high proportions of male hatchlings.

B. Crocodilians

The skins of crocodilians have long been used for expensive leather goods, and hunting has depleted many populations. The recovery of the American alligator, Alligator mississippiensis, following passage of protective legislation shows the potential for managing and restoring crocodilians. Success requires enforcement of laws and control of commerce, and these are difficult goals in many areas where crocodilians occur.

Crocodilians may be especially susceptible to some forms of pollution as a result of their position near the top of aquatic food chains. Pollution of Lake Apopka by the pesticides dicofol and DDT and their metabolites has reduced the hatching success and hatching viability of alligators in the lake and has interfered with normal sexual development (Guillette et al., 1994). Dicofol and
DDT have estrogen-analogue properties and have af-
fected the development of both sexes. Six-month-old
female alligators from Lake Apopka have abnormal
ovarian morphology and higher plasma levels of estradiol-17β than females from the control site. Male
hatchlings from Lake Apopka had abnormally small
penises and plasma testosterone levels only one-quarter
those of males from the control site.

C. Lepidosaurs

Most extant species of lepidosaurs are relatively small.
Of course, small size does not protect a species, but it
does tend to make its problems less conspicuous—and
hence less appreciated—than those of animals like sea
turtles and crocodilians. With a few exceptions, which
have earned special status by being unique (e.g., tuatara,
Sphenodon) or colorful (the San Francisco garter snake,
Thamnophis sirtalis tetrataenia), the status of popula-
tions of lepidosaurs is unknown.

Larger species of snakes and lizards are killed for
food and for their skins. Professional hunters in Central
and South America ship iguanas (Iguana) to food mar-
kets in the cities and collect tegu lizards (Salvator) with those biological characteristics can withstand high
wild populations, but it seems unlikely that animals
with long adult life spans and are high in the food
chain. We have little information about the status of
the United States, Italy, and Spain. The lizards and
skins were Indonesia, Singapore, Thailand, and Argen-
tina, and the major importing countries were Singapore,
Thailand, and South America ship iguanas (Iguana),
and the major importing countries were Singapore,
the United States, Italy, and Spain. The lizards and
snakes hunted for the leather trade are large animals
with long adult life spans and are high in the food
chain. We have little information about the status of
wild populations, but it seems unlikely that animals
with those biological characteristics can withstand high
rates of human predation indefinitely.

See Also the Following Articles

AMPHIBIANS, BIODIVERSITY OF • BIRDS, BIODIVERSITY OF • ENDANGERED REPTILES AND AMPHIBIANS • FISH,
BIODIVERSITY OF • MAMMALS, BIODIVERSITY OF • VERTEBRATES, OVERVIEW

Bibliography

56, 214.

Bennett, A. F., and Nagy, K. A. (1977). Energy expenditure in free-
ranging lizards. Ecology 58, 691.

marine Cretaceous of the Middle East. Nature 386, 705.

Demographics of common snapping turtles (Chelydra serpentina):
Implications for conservation and management of long-lived or-

Neglected taxonomy and ongoing extinctions of tuatara
(Sphenodon). Nature 347, 177.


change. In Biotic Interactions and Global Changes (P. M. Kareiva,
J. G. Kingsolver, and R. B. Huey, Eds.), pp. 93–119. Sinauer,
Sunderland, MA.

sonian Institution Press, Washington, D.C.

Etheridge, B., and de Queinter, R. (1988): A phylogeny of the iguani-
dae. In Phylogenetic Relationships of the Lizard Families, Essays
Commemorating Charles L. Camp (R. Estes and G. Pregill, Eds.),


Guillette, L. J., Jr., Gross, T. S., Mason, G. R., Matter, J. M., Percival,
of the gonad and abnormal sex hormone concentrations in juve-
niles alligators from contaminated and control lakes in Florida.
Environ. Health Perspect. 102, 680.

Hayssen, V., and Lacey, R. C. (1985). Basal metabolic rate in mam-
mals: Taxonomic differences in the allometry of BMR and body

32, 31.


Evolutionary shifts in three major structural features of the mito-

of reptilian eggs and embryos. In Biology of the Reptilia (C. Gans


RESOURCE EXPLOITATION, FISHERIES

John Beddington
Imperial College of Science, Technology, and Medicine

I. History
II. The Current Situation
III. Fisheries Science
IV. The Future of Commercial Fisheries

GLOSSARY

fishing effort: The level of fishing activity quantified in terms of the number and power of vessels and duration of fishing.
maximum sustainable yield: The maximum level of catch that can be removed continually from a fish population without depleting the population.
mesh size: The size of the holes in a fishing net.
recruits: The young fish that are first caught by a fishery.
stock recruitment relationship: The relationship between the adult breeding stock and the number of recruits produced by that stock.
total allowable catch: The level of permitted catch established by fishery regulations.

FISHERIES INVOLVE THE active removal of individual organisms from aquatic ecosystems using a variety of different techniques. Commercial fisheries occur in all major aquatic ecosystems. This article briefly reviews the history of fisheries and the current global situation, sets out in simple terms the scientific basis for fisheries management, and examines the possibilities for the future.

I. HISTORY

Fishing as a commercial activity has been conducted since the very earliest of times. Early records indicate that it was occurring in both Egyptian and Indian societies as long ago as 5000 B.C., and fisheries in the Mediterranean were already significant as long ago as 1000 B.C.

Since the early years, fishing has increased largely as a result of the changes that have occurred in the human population and its activities. In Fig. 1a, the UN estimates for world population from 1800 are shown. This shows an increase from around 1 billion in 1800 to approximately 6 billion at present. The major increase in population has occurred in the past 100 years or so, and particularly since 1950 the rate of increase has been very high.

In Fig. 1b, estimates of fish catch since around 1800 are shown. The figure indicates a catch of around 2 million tonnes in 1800, increasing to around 20 million tonnes in 1950. Since that time, catches have increased to a current level of around 120 million tonnes. The striking similarity between Fig. 1a, which shows the world population, and Fig. 1b, which shows the global fish catch, results from the fact that fisheries to a large extent have mimicked the change in population and...
the changes in human activity reflected by increasing industrialization.

Fishing as a commercial activity can usefully be divided into two elements: that involved with catching and that involved with preserving and processing the catch.

A. Methods of Catching Fish

There are four main methods:

Hooks and lines: This is perhaps the simplest method of fishing, from the earliest times handlines have been used, but current commercial longlines can be up to 100 km in length.

Gill nets: In this operation, nets are hung from vessels or floating buoys and are largely invisible to the fish, which are caught as they try to pass through the net. Such fishing methods can catch many species other than the main target species and large high seas gill nets are now banned under international treaties.

Trawls: This type of fishing involves the vessel dragging a large net (or trawl). This can be done either through the water, close to or along the bottom of the sea (demersal trawling), or in the water column (pelagic trawling).

Seines: The principle of this operation is that a school of fish is encircled by a net which is then tightened, reducing the circumference of the net until the school is caught. In modern tuna purse seineing, a fast motorboat is used to encircle a shoal of tuna and the net is then pulled in from the main vessel.

Although the basic methods of fishing are little changed since earliest times, the development of motor vessels to succeed sailing or manpower and the increasing power of engines and gear to locate fish has dramatically increased the fishing power of vessels over the years.

B. Processing the Catch

Original methods for processing or preserving fish involved sun drying, smoking, salting, and pickling. Subsequent developments as a result of industrialization have involved preserving fish on ice, freezing and processing on board the vessel, and reduction to meal. Modern factory vessels are highly industrialized operations that catch, process, and package fish ready for delivery directly to markets.

C. Changes Since 1945

The increased industrialization prior to World War II increased the pressure on fish stocks, but the war provided a brief respite as commercial fishing activity was reduced, for obvious reasons. Analyses of this period indicate significant recoveries in major fish stocks that had been heavily exploited prior to the start of the war.

Since then, total fish catch has increased by almost a factor of 5. Much of the increase has come from the coastal regions, where fishing activity has expanded with increases in population, creating more demand for fish products. However, in this period an activity of global significance has been the development of distant-water fishing by fleets primarily, but not exclusively, of the USSR and Japan. These fleets consisted of large vessels capable of operating for many months away from the home port and with substantial industrial freezing capacities. In the maritime regimes applying in the 1950s and 1960s, these vessels fished the rich continental shelf waters off the coast of many countries. In this period, the catch of distant-water fleets increased from around 500,000 tonnes in 1950 to a peak of 8.5 million tonnes in 1972. The change in the maritime regime brought about by the UN Conference on the Law of the Sea and the consequent UN Convention on the Law of the Sea has altered this in a dramatic way. Coastal states have set up Exclusive Economic Zones (EEZs) and restricted fishing on their continental shelf waters. This has resulted in a decline in the proportion of the world fish catch coming from distant-water fleets from 15% in 1972 to 5% currently.

D. Fisheries Management

The need to manage fisheries as an activity has been recognized since early times, and as long ago as the fifteenth-century, Dutch vessels operating in the North Atlantic had a closed season and restricted gear types in order to reduce pressure on fish stocks. Probably with an analogy to farming, it was also recognized that in addition to restrictions on season and types of gear, there was merit in protecting young fish that had yet to breed or fish that were currently breeding. However, regulations were comparatively few and were largely unenforceable except via peer pressure at a local level.

The need for management was recognized at an international level after World War II, and a number of international fishery commissions were set up. These varied from the International Whaling Commission, with its focus on whaling worldwide, to commissions specifically aimed at dealing with fisheries in a particular area,
which were the most typical. These commissions were set up via international conventions agreed between the coastal states of the region with fishing interests and states with distant-water fishing fleets operating in the area.

It is fair to say that with rare exceptions these international fishery commissions were largely ineffective in directly regulating the level of fishing pressure on fish stocks either via restricting the amount of fishing (effort) or by regulating the catch. The main difficulty was that the commissions typically found the necessary degree of consensus difficult to achieve among states with differing economic imperatives. Most agreements reached concerned readily enforceable regulations such as restrictions on gear types and on fishing seasons. Despite these difficulties, the fisheries management in these bodies was dependent on proper scientific assessment (discussed below) and routinely collected statistics on the catches of species under exploitation. Such data have proved invaluable in assessing the history of exploitation of particular fish stocks and are still of use today.

E. Post-UN Convention on the Law of the Sea

Since the Law of the Sea convention came into force, the need for international regulation of fisheries by international commissions has been reduced. Most states now have an Exclusive Economic Zone which extends 200 miles from their coast. The reason for 200 miles is that this is an approximation to the extent of the continental shelf where, with very rare exceptions, the vast majority of fish are caught, the nutrient-rich environment providing food for plankton and the fish communities dependent on it. Management in these EEZs has had mixed success, but they do in principle provide the opportunity for a single entity, the coastal state, to directly regulate the fisheries in the area.

II. THE CURRENT SITUATION

The UN Food and Agriculture Organization (FAO) has been collecting comprehensive data on fisheries since around 1950. Their figures indicate that the total world catch of fish is dominated by a small group of very large states, or those states that have particularly rich fish resources or a high level of distant-water fleet activities.

<table>
<thead>
<tr>
<th>Country</th>
<th>Catch (×10^6 tonnes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>14.2</td>
</tr>
<tr>
<td>Peru</td>
<td>9.5</td>
</tr>
<tr>
<td>Chile</td>
<td>6.7</td>
</tr>
<tr>
<td>Japan</td>
<td>6.0</td>
</tr>
<tr>
<td>United States</td>
<td>5.0</td>
</tr>
<tr>
<td>Russian Federation</td>
<td>4.7</td>
</tr>
<tr>
<td>Indonesia</td>
<td>3.7</td>
</tr>
<tr>
<td>India</td>
<td>3.5</td>
</tr>
<tr>
<td>Thailand</td>
<td>3.1</td>
</tr>
<tr>
<td>Norway</td>
<td>2.6</td>
</tr>
<tr>
<td>Korea Republic</td>
<td>2.4</td>
</tr>
<tr>
<td>Iceland</td>
<td>2.0</td>
</tr>
</tbody>
</table>

In fact, 12 countries take around 70% of the total world catch (Table I).

In the context of biodiversity, the world fish catch is dominated by a relatively small number of species: some 30% of the total world catch consists of only 11 species (Table II) and the many hundreds of other species caught only form a very modest proportion of this total catch. Nevertheless, fisheries for certain highly valued species such as squid, shrimp, and tuna are extremely important in global terms despite the volume of catch being small.

The development of commercial fisheries in the last three decades of the twentieth century has led to such pressure on fish resources that the paramount question is...
is whether the stocks can continue to sustain these levels of exploitation. Such questions have been posed for a number of decades and it has been recognized for several centuries that overexploitation can occur unless proper management action is taken. However, the current situation at a global level is clearly of serious concern. In a recent study by FAO, it has been calculated that of those important commercial fish stocks where sufficient information was available to allow judgment to be made, 44% are heavily or fully exploited, 16% are overfished, 6% are depleted with no evidence of any recovery, and 3% are depleted with some evidence of recovery. In other words, almost 70% of the key commercial fish stocks in the world are fully exploited or overexploited and 25% have been depleted to an extent that their capacity to deliver a sustainable yield has been seriously eroded.

In another study, FAO characterized three different phases in the fishery development process. This analysis has been based on a detailed historical record of the changes in fishery catches over the history of some 200 key fish stocks. The summary of these ideas is set out in Fig. 2, which shows the four phases of a fishery. These comprise an initial underdeveloped phase where exploitation is low and catches are low, a developing phase where the fishery is growing and catches increase, a mature phase where catches reach their peak, and finally a senescent phase where catches decline as the fishery is overexploited and the capacity of the resource to provide a sustainable yield has been eroded. Of the stocks analyzed in this way, 35% are senescent, 25% are mature, 40% are developing, and 0% have low exploitation. Clearly, the 0% is not a representative sample of all fish stocks, but is indicative of a low figure.

That these two sets of results are similar is of significant importance as they have been derived in rather different ways. The first explored the underlying scientific assessment of stocks and the second used a simple analysis of catch trends. Both show that there is a major problem of overexploitation and that some limit appears to be being reached to the overall global level of catches. The total catch of commercial fish stocks appears to be stabilizing around some 95 million tonnes. It is important to recognize that this apparent stabilization is not an intrinsic characteristic of the biological productivity of the oceans but rather a combination of that productivity and the plethora of individual fisheries that exploit it. Indeed, productivity varies quite significantly from area to area in the world.

A. Economic Issues

The historical changes that have led to the current level of overexploitation of the world’s fisheries have been paralleled by a somewhat unsatisfactory economic situation. Studies on changes in fishing capacity are relatively few as data are often difficult to come by and changes in overall levels of fishing are difficult to quantify, particularly where diverse regimes and types of fishing vessel operate. However, a recent attempt at synthesis has been made by FAO by referring to the number of decked vessels which are reported as being licensed by states. The distinction between a decked and an undecked vessel is a reasonable one between

![FIGURE 2 Schematic representation of fisheries development (after FAO: World Fisheries Resources 1997).](image-url)
primarily commercial vessels, which have a reasonable level of construction, and simple undecked vessels, canoes, etc. Such data that are available indicate that the numbers of undecked vessels primarily operating in subsistence fisheries in Africa and Asia have remained constant over the past two decades. By contrast, since 1970, the number of decked vessels has doubled. It is worth noting that the vast majority of these increases in vessel numbers have come from one state alone: China. In the period between 1980 and 1997, FAO reported that the number of decked vessels in China increased from 60,000 to 460,000.

This significant increase in the capacity of fleets as measured by the number of vessels also conceals the phenomenon that, on the whole, vessels are getting larger and more technologically sophisticated. Their costs are also correspondingly higher and thus the value and operating costs of the vessels of the world fleet in recent times are likely to be significantly more than those of a few decades ago—certainly more than the simple doubling indicated by the statistics.

In a recent study, Garcia and Newton analyzed the behavior of the world fleet in a very simple but informative way. They calculated that the total revenue obtained by the world’s fishing fleets (essentially the total first-sale value of the world catch) was less than the costs of operating the fleets by a very large amount. They quote a figure of some $54 billion for the operating loss in the year 1989.

How such operating losses are being sustained is an obvious question, to which there is an equally simple answer. These losses have got to be sustainable by either public, i.e., state, or private subsidies of the fishing operations. Given that most fishery operators are small, owning at most a few vessels, it is quite clear that such losses could not be sustained by the private sector and hence it is state subsidies that are underwriting the continuing losses of the global fleet.

That this is occurring has two unfortunate consequences. First, it means that substantial world resources are being devoted to the continuation of a failing economic activity. Second, vessels are being subsidized to continue in a fishery when normal economic forces would have driven them out much earlier. Hence the fishing pressure on already heavily exploited resources has been continued artificially.

This issue of subsidy in world fisheries has provoked major concern since the work of Garcia and Newton. A recent examination by Milazzo for the World Bank indicated the scale of the problem and showed how substantial and ubiquitous these subsidies are. Clearly, an immediate removal of subsidies from the fishing fleets around the world would lead to quite unacceptable levels of social deprivation and loss of food supplies. However, the overcapacity of fishing fleets has serious consequences for the continuation of sustainable fishing. States have recognized this and in some areas a reverse process is happening. For example, within the European Union there is a program, the Multi Annual Guidance Programme, which is aimed at ensuring, by fiscal means, a reduction in the European fleet. This and other mechanisms for reducing fishing effort are going to be essential if fisheries are to be managed sustainably in the future.

B. Ecosystem Considerations

1. Fishing Down the Trophic Web

Much of the discussion in the review so far has focused on the general characteristics of fisheries and it has treated individual fish resources as if they were in isolation from the ecosystem in which they are embedded. However, the level of exploitation that has been occurring has been accompanied, as might be expected from simple ecological principles, by a change in the species composition of the ecosystems that are being exploited. In a recent review, Pauly and coauthors have shown that the average trophic level of catches has been declining over the past five decades. In other words, fish catches are coming from lower in the food chain: fewer predators and more prey. It is too simplistic to consider this as a simple fishing down of the trophic web in which first the predators are removed followed by an increase in abundance of their prey. Fish species interact within an ecosystem in complex ways, often at different trophic levels and hence such a simplification is inappropriate. Nevertheless, it indicates that there is a secondary effect where levels of exploitation are high.

2. The Issue of Bycatch

It is well known that certain commercial fisheries operate inefficiently, in that the species targeted are not the only ones to be removed, damaged, or killed. Recently, Alverson and co-workers estimated that the annual level of bycatch in commercial fisheries is on average some 27 million tonnes (a very significant portion of the overall total world catch). This bycatch occurs primarily in fisheries where the gear, almost by definition, is unselective. The elimination, of such bycatch by gear change and regulation and/or the utilization of the catch for human consumption offers a significant potential for increasing overall productivity.
III. FISHERIES SCIENCE

Although work in the late nineteenth century anticipated some developments, the main pioneers of fishery science worked in the first two decades after World War II. Of the many scientists involved in the development of fisheries as a science, five individuals stand out: Schaefer, an American working at the Inter American Tropical Tuna Commission; Ricker, a Canadian working at the Nanaimo Laboratory in Canada; and three British scientists, Beverton, Gulland, and Holt, all working at the Fisheries Laboratory in Lowestoft. It is their work that set the scene for fisheries science as it is today. All of this work was developed prior to the availability of modern computing power and the scale of their achievements, set in this context, is impressive.

A. Maximum Sustainable Yield (MSY)

The basic question of fisheries science is how exploitation, the removal of individuals from a population, affects the dynamics of that population: the processes of birth, growth, and death. A central concept in this exploration is sustainable yield. The idea is simple but it requires an explanation in the context of population ecology. In its unexploited state, a population will on average be in balance with its environment, the processes of birth and growth will be offset by death, and the population will be at equilibrium. A simple differential equation, commonly used in other branches of ecology, expresses this idea well. The rate of change of the biomass \( B \) of a population over time is given by

\[
\frac{dB}{dt} = rB(1 - \frac{B}{B_0}). \tag{1}
\]

When the biomass is equal to \( B_0 \), the rate of change is zero and the population stays at its equilibrium (unexploited) level \( B_0 \). When \( B \) is very small compared to \( B_0 \), the population will grow exponentially at a rate \( r \). At intermediate levels, the rate of change will decline from its maximum at \( B = 0 \) as \( B \) increases to zero when \( B = B_0 \). It is possible to plot the right-hand side of Eq. (1) to show how the increase in biomass of a population is affected by its size.

This is done in Fig. 3a. The figure shows that there is a maximum increase in biomass which occurs at a level where the biomass is equal to \( B_0/2 \). Some simple calculus can be used to show that this maximum is given by the expression \( rB_0/4 \). If exploitation is removing biomass from the population, then as long as the removals are less than the maximum, the removals will be replenished by the population. These removals (the yield) are thus sustainable as long as they do not exceed this maximum, which is called the maximum sustainable yield (MSY). The MSY will vary with the abundance of the fish stock concerned (\( B_0 \)) and with its capacity to increase (\( r \)).

The exploitation of the population can be modeled by a simple modification of Eq. (1). Let \( p \) be the proportion of the population that is removed by exploitation. The equation for exploitation is then

\[
\frac{dB}{dt} = rB(1 - \frac{B}{B_0}) - pB. \tag{2}
\]

This is shown by a simple modification of Fig. 3a in Fig. 3b.

The expression for the removals \( pB \) is a simple straight line with slope \( p \). The larger the proportion removed, the steeper the line. Where the line intersects the curve showing the increase in biomass, the removals are exactly replenished by the increase in biomass. Simple calculations can show that where \( p = r/2 \), the maximum sustainable yield is reached. Where \( p \) is greater than this level, there will be a sustainable yield, but it will be less than the maximum and the biomass of the population will be reduced below the level where it produces the maximum sustainable yield. Increasing the intensity of harvesting then will actually reduce rather than increase the yield. By contrast, if \( p \) is less than \( r/2 \), increasing the intensity of harvesting will increase the yield.

In the practical situation, the proportion of the fish stock removed will be determined by the level of fishing effort: the number of boats, their size, the type of fishing gear, the duration of fishing, etc. If the level of fishing effort is less than that required to take the maximum sustainable yield, increasing the effort will result in an increased yield. If, however, the level of fishing effort is higher than that required to take the maximum sustainable yield, further increases in that effort will result in a decrease in sustainable yield. Such a situation is often termed overcapacity as removing fishing effort will actually increase yield in the longer term. It should be emphasized that the idea of MSY developed above is an oversimplification and that environmental variability and other factors render its simple application problematic.

The simple models used to illustrate the ideas of sustainable yield and fishing capacity do not allow a proper exploration of some of the other key areas of fisheries science. To explore these requires a deeper understanding of the way in which fish populations behave. The idea can be developed simply using as a
subject an idealized fish population; here young fish develop from their larval form and become part of the population at a particular time of the year. A year later they will have grown and some of the cohort will have died, either through predation or other natural causes. The process continues and a typical temperate fish population will consist of a number of age groups, with fewer fish surviving to greater and greater ages. This is illustrated in Fig. 4a.

Exploitation of this population can in principle start at any age, but in practice will be determined by fishing practices. For example, the size of the net will determine what age groups are caught, smaller and younger fish can escape through the mesh. Figure 4b shows how this will affect the age structure of the population. In the exploited population, fewer fish survive to greater ages and the heavier the exploitation, the more extreme will be the reduction in older fish. However, the biomass of the population is affected not just by natural deaths and fishing but also by growth. As fish get older, they increase in size according to some relationship. Figure 5 illustrates this for a typical fish.

In a comprehensive work first published in 1957, Beverton and Holt explored how the yield could be affected by changing the age at which fishing commenced and the intensity of fishing. Their results show how the interplay between these two parameters, which can be changed by fishing practice, can alter the level of sustainable yield. A centrally important result is that in many cases delaying the age at which fishing starts to operate can improve yields. This was in part the basis for many national regulations and international agreements to regulate the mesh sizes of nets and the type of fishing gear permitted.
B. Stock and Recruitment

One further element is missing from the description of the dynamics of an exploited fish population: reproduction. Typical fish are enormously fecund, laying as many as several hundred thousand eggs at a time. The eggs and the larvae into which they develop become part of the plankton, where they are subject to high levels of mortality from predators and from starvation. To illustrate the enormous mortality involved, a male and female fish might produce 100,000 eggs. For the population to replenish itself, neither grow nor decline, two eggs need to survive to adulthood. There are two issues here. The first is conservation, for clearly sufficient adult fish must be permitted to survive to breed to allow population renewal. The second is variability, for it is hard to conceive of a natural process that can provide such control: if on average less than two individuals survive, the population is failing to replenish itself; if more than two survive, it will increase. Unsurprisingly, observations on fish recruitment, as it is termed, show high variation.

C. Management

The management implications of this analysis are important. If fishing starts on age groups that are not yet mature, then fishing intensity must be controlled to ensure sufficient adults survive to breed so that declines in recruitment do not occur. If fishing does not start on age groups until after breeding has occurred, the conservation issues become less intense, although care is still needed. For example, larger, older fish produce more eggs.

An overexploited fish population will thus exhibit characteristic symptoms. First, there will be few age classes in the population and hence the average age/size of the catch will be lower than in less intensively exploited times. Second, the recruitment of young fish will be in decline, reflecting overexploitation of the adult stock. In such a situation, the management action needed is straightforward, a reduction in fishing intensity and an increase in the age of first fishing. However, depending on the severity of overexploitation, fishing may need to cease altogether.

Although the management action needed to address overexploitation is well understood, once severe overexploitation has occurred, there is no guarantee that
reducing or ceasing fishing will reverse the process and that the stock will recover. There are plausible mechanisms that can prevent this, of which the most important are likely to be increased competition or predation by other species. Accordingly, the best fisheries management practice is aimed at preventing overexploitation from occurring.

D. Monitoring

Clearly, a key to managing fish stocks is the ability to monitor what is happening to the stock under exploitation. There are direct methods of surveying and monitoring stocks using research vessels which can work well in many cases. However, in other situations such methods are either inefficient or economically unjustifiable. Regular monitoring of information from the catches of the commercial fleets permits a number of different methods to be used. These methods use this statistical information to reconstruct the behavior of the stock. One common method involves returning to the model of Eq. (2). The yield or catch from the stock was given by the equation

\[ \text{Catch} = pB. \]  

(3)

As noted earlier, \( p \) depends on the level of fishing effort. The definition of fishing effort \( (E) \) is dependent on circumstances, hence for a particular definition of fishing effort, e.g., hours fished:

\[ p = qF. \]  

(4)

where \( q \) is a scaling factor (known in the literature as the catchability coefficient)

\[ \text{Catch} = qFB. \]  

(5)

Statistics can provide information on catch for known effort and hence catch per unit of effort (CPUE) is proportional to stock biomass, with the constant of proportionality being the catchability coefficient. A variety of statistical methods exist to estimate the catchability coefficient and as a consequence the stock size. Time series of information on catch and effort thus provide direct insight into the way in which the stock has responded to exploitation.

A second widely used method uses information not just on the catch but on the age composition of that catch. This extra information is used in a method known as virtual population analysis (initially developed by Gulland) to reconstruct the changes in the population under exploitation. The method is simple in conception, although computationally quite laborious. In essence, a simple accounting procedure is used: a catch of 1000 eight-year-olds this year implies that at least 1000 seven-year-olds were alive in the previous year. Adjustments for natural deaths and fish catches then permit the changes in population to be estimated.

E. Fisheries Economics

It is reasonable to enquire why fisheries are in such a poor state when the science of fisheries is reasonably well developed, methods of monitoring stocks exist, and management actions needed to address overexploitation are understood. The answer lies primarily in the domain of social science.

Fisheries as an economic activity has one fundamental difference from other industrial activities. In manufacturing industry, increases in supply develop in response to demand by increasing inputs of labor, raw materials, capital, etc. In fisheries (and in other industries that exploit renewable resources) there is a limit to supply which is determined by the maximum sustainable yield. Increasing the inputs to a fishery—more fishermen, better gear, more vessels, etc.—actually reduces the yield if the stock is already being exploited at or above the level of effort that produces the maximum sustainable yield. In this situation, increases in demand reflected in higher prices can lead to increases in fishing effort, lower catches, and hence higher prices. Depending on the situation, vessels may still be profitable or covering costs even though the effort being expended is well above that required to take the catch and the price paid by consumers well above that which would prevail if effort were lower and the catch higher. In a situation where there is no control of the fishery, this process can continue until resources are severely overexploited and individuals in the industry barely cover their operating costs. The issue is one of ownership; many fisheries are still (and most were) common property resources. In such cases, access is unrestricted and additional effort enters the fishery as long as the return is better than in other activities. Because fishing is a specialized activity (fishing gear and vessels have few alternative uses), other alternative activities offer very low returns. Hence most fisheries without management tend to be overcapitalized and economic returns are poor.

Most government action in the past few decades has been the reverse of what would be economically efficient. Fishing operators have been subsidized and have stayed with the industry rather than leave, thus
accentuating the overexploitation problem. Clearly, social issues play an important role here; it is obviously problematic to let fishing communities face economic collapse, but the scale of subsidy involved is substantial (see above).

**F. Fishery Management in Practice**

The management of fisheries involves some simple principles based on sustainable yield ideas yet its practice is often fraught with difficulty. Management measures typically control fishing by regulations that impose limits on catch (quotas, or total allowable catches (TACs)) or limits on access or effort (licensing, closed areas, and seasons) or by indirect controls such as minimum size of catch or restrictions on types of gear.

It is fair to say that such methods have not been entirely successful, for a variety of reasons. Control and policing of regulations are difficult and can be extremely costly. Some regulations, TACs in particular, provide an economic motivation for misreporting of catches and hence undermine the scientific basis of the process. Political issues, particularly when there is significant overcapacity, can override fisheries management concerns. Often in the face of uncertainty, regulations are set with an optimistic view of the stock, which subsequent scientific information reveals to be unfounded. Even when regulation has succeeded in its conservation goals, the fishing industry remains overcapitalized and continues to need assistance.

One development where there appear grounds for some optimism is the construction by certain states of regimes based on the allocation of property rights to operators: Australia, Iceland, and New Zealand have been particularly active. In these regimes, operators are allocated the rights to a particular level of catch (or proportion of some fixed catch). These rights, known as Individual Transferable Quotas, can be bought and sold. In this way the inefficient operations can be bought out by the more efficient ones, which can adjust their capacity to the appropriate level for exploiting the fish stock. The attraction of such a regime is that it is in the interests of the operators (owners) to ensure that the stock is exploited sustainably.

**G. New International Measures**

Although most states now operate 200-mile zones, in which they control fishing, in certain areas the continental shelf extends beyond 200 miles and exploitation of the same stock can be possible both within the 200-mile zone and beyond it on the high seas. A recent international agreement, the UN Agreement on Straddling Stocks, offers a new international framework for addressing problems of this sort. Its aim is to facilitate the creation of regional bodies composed of both coastal states and those with distant-water fleets operating in the area. It is too early to assess whether this will be successful.

Certain types of species, tuna in particular, are highly migratory and are exploited over very large areas of ocean. Management regimes exist under International Treaty to regulate these fisheries. The latest commission of this sort to be set up is the Indian Ocean Tuna Commission, which held its first meeting in 1997. These regimes rely on considerable political goodwill. A recent development (August 1999) involved Australia and New Zealand taking Japan to the International Tribunal for the Law of the Sea to seek judgment for Japan to reduce fishing on southern bluefin tuna. The three states are in the Commission for the Conservation of Southern Bluefin Tuna but have proved unable to agree on management action within the Commission. The Tribunal made an interim judgment effectively asking Japan to reduce catches pending a full arbitration on the merits of the case. Whether similar cases will be brought and the Tribunal becomes the international arbitrator of fishery management regimes on the high seas remains to be seen.

**H. Ecosystem Management**

Fish stocks exist in aquatic ecosystems, forming the prey of some species and being predators of others, yet most fisheries science and almost all fisheries management regimes tend to focus on individual species. The regime that was set up to deal with the Southern Ocean, the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR), had a different focus. This regime was specifically oriented to the Southern Ocean ecosystem and regulations were aimed at ensuring that each element of the ecosystem was protected from the effects of exploitation. The primary reason for this markedly different approach was that the large baleen whales of the Southern Ocean had been grossly overexploited by distant-water fleets this century. A significant development of a fishery for their common prey, krill, could hinder, or indeed stop, the recovery of the populations following the cessation of whaling. To date, CCAMLR has been reasonably successful in its operation, setting precautionary levels of catch for krill as well as regulating other fisheries.

However, better scientific understanding of the implications of exploitation of groups of interacting spe-
cies is needed. Current understanding is really little more than educated commonsense—e.g., a predator cannot be exploited at its MSY level if its prey is simultaneously harvested at a high level.

IV. THE FUTURE OF COMMERCIAL FISHERIES

The increase in the world population over the next two decades will clearly produce an increase in basic demand for food. This is likely to be enhanced in the case of fish as it is a preferred protein source and increasing wealth tends to be associated with increasing fish consumption. Additionally, much of the population increase is estimated to be occurring in Asia, where fish is a preferred source of animal protein.

Statistical analyses of trends in capture fisheries point to a maximum harvest level in the region of 95 million tonnes, a value close to recent average landings. However, more recent analysis by FAO indicates that there is a potential for increasing these catch levels if the full potential of underexploited areas and species were to be realized. Similarly, there is scope, as discussed earlier, for significant improvement in the reduction of unwanted bycatch (around 27 million tonnes is estimated to be discarded each year).

Improved fishery management could in principle add significantly to the global production—recall the high proportion of overexploited stocks. However, the extent to which species interactions may mean that the recovery of one stock may lead to lower catches from another is not well understood.

FAO has concluded that a substantial increase in catch levels of around 20 million tonnes could be achieved if (a) degraded resources are rehabilitated, (b) underexploited resources are successfully managed to increase yields, (c) fully exploited resources are not degraded, and (d) discarding and wastage are reduced. This is the challenge for fisheries science and management in the new century.

See Also the Following Articles

AQUACULTURE • FISH BIODIVERSITY, OVERVIEW • FISH CONSERVATION • FISH STOCKS

Bibliography

RESOURCES ARE ESSENTIAL to the life of every organism. Species constantly compete with one another for these resources but may be able to coexist in the long term if their resource requirements are sufficiently different, known as “resource partitioning.” Maintenance of species diversity within a community is contingent on this ability of species to coexist together. In this article, we discuss the evidence that differences in plant species’ resource requirements are sufficient to permit their coexistence. We emphasize the importance of considering species’ responses to multiple, interacting resource axes and to variation in resource availability in both space and time. Species do partition the resource environment in multiple ways, but in plant communities of high diversity, e.g., tropical rain forest, there is considerable overlap in species’ resource requirements. In such systems, other diversity mechanisms might play a more significant role.

I. INTRODUCTION

The scientific community and the public at large are expressing great concerns about the global loss of bio-

GLOSSARY

coeexistence Ability of species to persist together in an ecological community indefinitely in the absence of any major environmental change.

eheterogeneity Variation in environmental conditions through space (spatial) or time (temporal).

niche Pattern of response of an individual, a population, or a species to the physical and biological gradients of its environment.

niche differentiation Differential resource use or response that results from long-term, consistent reciprocal selection.

reciprocal selection Evolutionary change as a result of species acting as selective agents on each other through interspecific interactions.

resource A consumable or depletable substance, such as a nutrient, water, or light, that is required by plants for maintenance, growth, and reproduction.

trade-off Constraint on expression of a species’ trait behavior as a result of expression of a correlated trait behavior.

RESOURCES ARE ESSENTIAL to the life of every organism. Species constantly compete with one another for these resources but may be able to coexist in the long term if their resource requirements are sufficiently different, known as “resource partitioning.” Maintenance of species diversity within a community is contingent on this ability of species to coexist together. In this article, we discuss the evidence that differences in plant species’ resource requirements are sufficient to permit their coexistence. We emphasize the importance of considering species’ responses to multiple, interacting resource axes and to variation in resource availability in both space and time. Species do partition the resource environment in multiple ways, but in plant communities of high diversity, e.g., tropical rain forest, there is considerable overlap in species’ resource requirements. In such systems, other diversity mechanisms might play a more significant role.

I. INTRODUCTION

The scientific community and the public at large are expressing great concerns about the global loss of bio-

GLOSSARY

coeexistence Ability of species to persist together in an ecological community indefinitely in the absence of any major environmental change.

eheterogeneity Variation in environmental conditions through space (spatial) or time (temporal).

niche Pattern of response of an individual, a population, or a species to the physical and biological gradients of its environment.

niche differentiation Differential resource use or response that results from long-term, consistent reciprocal selection.

reciprocal selection Evolutionary change as a result of species acting as selective agents on each other through interspecific interactions.

resource A consumable or depletable substance, such as a nutrient, water, or light, that is required by plants for maintenance, growth, and reproduction.

trade-off Constraint on expression of a species’ trait behavior as a result of expression of a correlated trait behavior.
logical diversity. Currently, biological diversity is threatened worldwide by the widespread destruction of forests and by impacts of global environmental change, such as nitrogen deposition and global warming. In addition, ease of travel, the dramatic increase in the number of people who travel, and deliberate introductions of organisms into various habitats can all lead to the modification of ecosystems and the loss of biological diversity. Invasions are now rampant and may lead to a great deal of homogenization of the Earth's biota. This situation has led the renowned biologist Gordon Orians to call this epoch the "Homogenee."

If we are to understand and predict the consequences of loss of species diversity for the functioning of the biosphere, we need to ascertain what factors control patterns of species richness both within and across ecological communities. The notion of community diversity has its roots in the writings of Henry Gleason and Frederic Clements. These American ecologists differed sharply in their interpretation of the plant community. The Clementsian conception of the community assumed it to be a cohesive active unit. The notions of reciprocal selection and niche differentiation are both imbedded in this idea, although not explicitly. In contrast, Gleason thought of the community as a chance assemblage of species with overlapping physiological requirements. This view may assume that there is no interaction between species or that these interactions are incidental. More recent discussion of why there are so many species was rekindled by G. E. Hutchinson in 1959 when he asked this question in his now very famous "Homage to Santa Rosalia" paper.

Currently, there are many hypotheses about the generation and maintenance of plant diversity (see discussions in both Tilman and Pacala (1993) and Huston (1994)). Much like the Clements–Gleason debate, discussion of the maintenance of species diversity often focuses on the distinction between theories that assume evolution has driven species to partition resources so as to prevent competitive exclusion (equilibrium) and those that explain diversity through the action of demographic dynamics and do not assume a past evolutionary history between species (nonequilibrium). Ecologists commonly favor one mechanism for maintenance of diversity over others, but it is more likely that the strength of any cause differs under different locations and circumstances. It is now critical that we ascertain what characteristics of species and ecosystems lead to the dominance of any one diversity mechanism, so that we can develop a predictive framework for determining causes for maintenance of species diversity across all ecological communities.

II. ASSESSMENT OF DIVERSITY AND ITS FUNCTIONAL SIGNIFICANCE

Presently, there is no agreed quantitative way to assess diversity, except to use the taxonomic species concept. However, the use of species as the primary unit of biological diversity is not always appropriate for the situation. We have previously argued that we should focus on different components of ecological diversity depending on the patterns and processes being studied. Indeed, there is now an increasing awareness that biological diversity should encompass the full range of organizational units in ecology, from genes to ecosystems. If we wish to incorporate this broader definition of diversity into our study and stewardship of ecological communities, we must present a clear framework for the accurate assessment of diversity. Developing such a framework involves explicitly outlining the different kinds of biological diversity that should be recognized and how the level of diversity can be quantified.

In many cases, the confusion arises from not recognizing the appropriate scale for the ecological phenomenon of interest. The need for clarification of scale was first recognized by Robert H. Whittaker in 1972, when he addressed the importance of distinguishing within- and between-patch diversity (known as \(\alpha\)-diversity and \(\beta\)-diversity, respectively). The temporal scale of ecological dynamics may be as relevant to our discussion of biodiversity assessment as the spatial scale. This notion is well illustrated by the need in certain systems to distinguish between apparent diversity and total diversity, which may be quite different from one other. Many ecological communities depend on an intermittent availability of resources. For example, in deserts, when...
Although plant communities are only theoretical constructs (see Bazzaz, 1996), much discussion has gone into the notion of biological diversity and its impact on the functioning of ecosystems. A number of recent, high-profile experiments have addressed how many species must be present for a given ecosystem to function appropriately, e.g., the Ecotron experiment (Shahid Naeem and others), field studies on California serpentine grassland communities (David Hooper and Peter Vitousek) and on midwestern tallgrass prairie communities (David Tilman and others), and the BIODEPTH experiment carried out at eight grassland sites across Europe. Current data in these grassland ecosystems seem to suggest that the number of species for greatest ecosystem productivity hovers close to nine. Unfortunately, there is still considerable uncertainty about what mechanisms are driving these diversity–productivity patterns (see discussion in recent BIODEPTH paper: Hector et al., 1999). Some ecologists currently believe that complementarity in species’ resource use could make a multispecies community more effective at utilizing available resources (to produce higher net productivity), while others argue that more diverse communities are merely more likely to contain more productive species (known as the sampling effect). As we currently lack a complete mechanistic understanding of the relationship between species diversity and ecosystem function and as relevant data exist for few other ecosystem types, we cannot yet know how general these results will turn out to be. In these particular experiments, the species are all somewhat similar in their behavior and no ‘keystone’ species exist. In communities containing a species that performs a special function (such as nitrogen fixation), the removal or addition of such a species may severely impact ecosystem function, e.g., Myrica faya invasion in Hawaii. We must also pay close attention to species that supply unique goods and services to humans.

III. RESOURCES AND NICHE DIFFERENTIATION: THEORETICAL CONSIDERATIONS

Maintenance of species diversity is often thought to be a result of niche differentiation. This idea was implicit in much early writing on community structure that recognized that species could not coexist with one another unless they had sufficiently different patterns of resource use. As plants are predominantly sessile organisms, it is assumed that plant species that are compo-
nents of a community have coexisted and competed for a long time. In this way, they are assumed to have shaped each other’s responses to the multitude of critical physical and biological resources in their local environment through reciprocal selection, which could result in niche differentiation in the long term (Connell, 1980). This idea is similar to character displacement in animals. However, requirements for niche differentiation are stringent and may not always be met. Long-term species co-occurrence is needed and the frequency of encounter between plant species will depend on dispersal patterns (see Section V.C). In addition, reciprocal shaping of species’ responses to their environment is not easy to accomplish, unless initial traits related to fitness are somewhat independent of each other. To avoid assuming that niche differentiation has taken place in the past, Bazzaz (1996) suggested previously that it is more tractable to consider copresence rather than coexistence, especially in communities with a poorly known history. Studies on the postglacial migrations of temperate tree species during the Holocene in North America have clearly demonstrated how species have moved individually across the landscape and have not necessarily coexisted in the same community for long periods of time (see papers by Margaret Davis). Here, we consider communities where species are found together without regard to whether or not they co-occurred over evolutionary time and therefore have shaped each other’s responses to the environment.

While niche differentiation has been touted as a major cause of the generation and maintenance of biodiversity, particularly in late successional stable ecosystems, it was unclear for many years how the concept applied to plant communities. The notion of resource partitioning was originally developed for animal communities, where species can potentially utilize a wide range of food resources and thus can easily partition up their resource environment. This idea was harder to apply to plant communities, as plants all have relatively similar resource requirements (Connell, 1978). As a result, much ecological research has focused on trying to determine how species partition themselves along resource axes. Progress on understanding factors controlling plant community diversity has been achieved by considering species’ responses to multiple resources in the environment. David Tilman (1982) tackled this issue by using a mathematical model to show that species coexistence was feasible if there were trade-offs in species’ requirements for different ratios of resources. This approach clearly demonstrated the potential for a resource-based approach to contribute to explaining plant species diversity. Often, however, species’ responses to environmental factors are not independent of each other. Thus, it may be more informative to develop a multidimensional response surface that approximates the response of the species to its environment (Fig. 2) (Bazzaz, 1996). Gradient techniques from vector calculus now allow us to quantify these response surfaces.

In this multidimensional view of the plant niche, maintenance of diversity depends on species separating themselves along multiple resource axes, such that they can coexist for a long time. It is assumed that there must be a minimal overlap between members of a community. In 1973, Sir Robert May attempted to quantify the minimum allowable overlap of the species for their stable coexistence and concluded that, if the difference between the means of their responses is more than twice that of the spread of each individual curve, stable coexistence should be possible. If species are highly specialized (narrow response curves), they do not need to position themselves very far apart on one particular resource axis to prevent competitive exclusion, whereas species that are more generalist in their resource use (wide curves) must be further separated in niche space. This kind of analysis has to be extended to multidimen-
RESOURCE PARTITIONING

177

We still do not know, even theoretically, what the minimum overlap to allow coexistence should be when we consider species' responses to more than one critical environmental gradient. These minimum overlap conditions for coexistence may differ for contrasting community types. For example, it is reasonably well established that plants of early successional communities have broader responses to resource gradients than do plants of late successional communities. This pattern alters the length and intensity with which species interact with one another, and thus the degree to which species partition the environment. However, it has not been determined whether this pattern results from or leads to the lower species diversity of early successional communities relative to late successional ones. Early successional plants have a near-geometric type of distribution (each species takes approximately the same proportion of resources to produce the same amount of biomass). As succession proceeds, there is a shift toward a more log-normal distribution (Fig. 3) and then finally, without major disturbance, a return to a near-geometric distribution. This variation in patterns of resource use between species of different successional position creates a hump-backed relationship between diversity and successional time, as predicted by the intermediate disturbance hypothesis (Connell, 1978).

We have seen that mathematical models of community structure predict that diversity is largely a function of how species partition themselves along multiple environmental resource axes. Theoretical considerations suggest that trade-offs in plant species' responses along multiple environmental gradients could permit multispecies coexistence. But, do we see such trade-offs in nature and are they sufficient to explain long-term species coexistence? In the next section, we discuss the evidence for niche differentiation in plant communities and consider whether or not the degree of resource partitioning in these communities is adequate to explain species diversity in different plant communities. Resources may be partitioning both spatially and temporally and we present evidence for both kinds of niche differentiation among plant species.

IV. RESOURCES AND NICHE DIFFERENTIATION: EMPIRICAL EVIDENCE

A. Spatial Heterogeneity

Most resource partitioning arguments proposed to explain species diversity assume that the environment is heterogeneous. Coexistence is possible because species have different resource requirements and are specialized to succeed on particular patch types. Diversity is maintained through the presence of an array of patch types. As our ability to measure and quantify environmental variation has improved, we have developed clear evidence that the environment is heterogeneous with regard to resources that are critical to plants. To date, however, we have not critically assessed if the degree of heterogeneity recorded by our instruments is representative of the resource environment perceived by the plants themselves (Bazzaz, 1996). Despite this uncertainty, there is now clear evidence that, at least on the broad scale, species do partition themselves along resource axes. Studies that relate species' distributions to multiple environmental factors using either direct or indirect gradient analysis have identified numerous soil-related axes of specialization, e.g., R. H. Whittaker's 1956 work on the vegetation of the Great Smoky Mountains. Research on experimentally produced plant communities also supports the notion that species may be differentiated along soil resource axes (see studies discussed in Bazzaz, 1996).

Although plants themselves can influence their local environment and thus produce intrinsic heterogeneity within the community, much spatial resource heterogeneity is generated by extrinsic factors, the most common of which is disturbance. In fact, disturbance is judged both by change in availability of resources and by per-
cept of that change. Disturbance events create heterogeneity in resource availability at multiple spatial and temporal scales, and these different scales can allow species to partition the environment in more ways. Awareness of the importance of disturbance events for understanding the maintenance of plant species diversity was heightened when Peter Grubb (1977) proposed that regeneration traits of species should provide a critical axis for differentiation, the so-called regeneration niche. Much of the research on partitioning along this regeneration axis has focused on the broad-scale effects of canopy disturbance, addressing differences in conditions and plant responses between gaps and closed forest. Gap creation generates variability in the light environment and has been advanced as a major factor in increasing diversity at the stand level. Broad guilds of plant species that respond similarly to gap formation (light demanding vs shade tolerant) have been identified in numerous forest communities where vertical canopy stratification creates a wide range of light micro-environments (Whitmore, 1989). More recently, sophisticated mathematical techniques have been used to classify both physiological and demographic responses of tree seedlings to light availability more precisely (see SORTIE papers of Steve Pacala, Richard Kobe, and others). Much of this research into species’ responses to total light quantity has found considerable overlap in response, and species’ partitioning of the light micro-environment is not usually sufficient to explain maintenance of species diversity, particularly in communities with high numbers of co-occurring species, such as lowland tropical rain forest.

An additional mechanism for the maintenance of diversity arose from the gap partitioning hypothesis, originally developed by Ricklefs (1977) and applied specifically to the tropical rain forest by Julie Denslow (1980). Here, finer-scale variation in resource availability across the gap–understory continuum was suggested as a possible mechanism for coexistence in multispecies communities. Unfortunately, this hypothesis has yet not been validated in tropical rain forests, where we lack a good understanding of the maintenance of species diversity. Experimental work by Tim Sipe and F. A. Bazzaz in temperate forests in eastern North America, in contrast, provides some support for the gap partitioning hypothesis. In experimentally created gaps at Harvard Forest in central Massachusetts, we found clear micro-environmental differences between gaps of different sizes (especially with regard to light). We compared morphological and physiological responses of three co-occurring maple (Acer) species to this light variation and found clear evidence that the species differed in their response to gap size. Many response variables showed significant differences between large gaps, small gaps, and understory plots, and these differences often varied between species, creating distinct species’ preferences for canopy gap environment (Fig. 4). We also observed substantial variation in light availability between different parts of canopy gaps, due to seasonal and diurnal trends in solar patterns. This variation, however, was rarely reflected in differences in species’ responses to position within the gap.

B. Temporal Heterogeneity

As resources vary in both space and time, species could partition their resource environment temporally as well as spatially. Temporal heterogeneity in resource availability is often not considered in discussions of maintenance of diversity, but could prove central to our under-
standing of species coexistence. Differences in the timing of plant developmental events, such as germination, flowering, and fruit drop, could play a critical role in mechanisms of species coexistence. In some communities that have existed for a long time, there is evidence for clear separation in flowering between groups, e.g., Peter Ashton’s work on staggered mast-flowering in the Dipterocarpaceae in the tropical rain forests of Southeast Asia. In the tallgrass prairie of the American Midwest (communities that have a long evolutionary history and great species diversity), we have demonstrated clearly how these species separate their life history events along the entire growing season, i.e., phenological separation. One can recognize three fairly distinct groups of flowering species (Fig. 5). Also, in these grasslands one can observe the separation of the phenology of the introduced species, *Poa pratensis*. Although it is now considered by many investigators to be part of the tallgrass prairie, it is an introduced species (from European grasslands) that grows well in the cool portion of the season. Thus, it is able to grow at two distinct times of the year: before and after the main growing season of its competitors, the native, warm-season grasses of the community (Fig. 6). In this way, this species separates its growth activity from that of the native species and reduces competition for critical resources.

This kind of phenological separation is also evident in some early successional communities where species generally have broad niches (see earlier discussion in Section IV.A). In old-field communities in the American Midwest, *Chenopodium album* delays its growth and reproduction much later than all other species in the community (Bazzaz, 1996), thereby reducing competition during the main portion of the growing season. In this case, it is not clear that competition has shaped these phenological responses, as this in early successional community have not necessarily been present together over evolutionary time. Many arguments about coexistence assume that competition has shaped the niches of co-occurring species in the past (see earlier discussion). However, in communities of unknown history, the question remains whether a species has modified its phenology by competition with the native species or was preadapted to function in these communities. An example of such preadaptation is found in the shifted daily flowering behavior of the native *Erigeron*
annus and the introduced Lactuca scariola. In winter annual communities of the midwestern United States, Lactuca opens its flowers in the morning hours 7:00 to 10:00 a.m. and requires a large number of pollinators. After 10:00 a.m., however, Erigeron starts to open and Lactuca closes its flowers (Fig. 7). As a result, the pollinators shift their foraging to Erigeron. Thus, while both species use the same pollinators, the daily separation in the opening of the flowers ensures that both are pollinated. At first glance, this may appear to be a convincing example of niche differentiation. However, when examined carefully, one finds that the introduced Lactuca developed this flowering behavior in its native habitat in Europe and not by competition for pollinators in the successional fields of the Midwest.

Temporal differences in species’ traits may extend to very early stages of the life cycle. The formation of seed banks that can persist in the soil for a long time may also lead to maintenance of species diversity. A persistent seed bank may buffer species’ responses to short-term variation in environmental conditions and prevent irreversible population crashes. Weedy species are known to form persistent seed banks. In good years for a particular species, many seeds are produced and enter the soil seed bank. In bad years, in contrast, few seeds are produced. The species can, therefore, partition the time axis to contribute the most to future generations in years when conditions are favorable for growth and reproduction. The development of this sort of persistent seed bank has been mathematically modeled and is often called the storage effect (see work by Peter Chesson). Coexistence is promoted if species have differing sensitivities to temporal variation in environmental conditions.

Further niche axes may potentially be created through the combination of temporal heterogeneity in multiple resources (see discussion in Section III). Because plant responses to a particular resource are contingent on prevailing local environmental conditions, e.g., temperature and relative humidity, the time course of resource availability patterns could critically determine a plant’s performance. If resource availability is high when other conditions are favorable (i.e., resources are congruent), then plants may be able to make good use of that resource. If, however, conditions are unfavorable and plant activity is inhibited to some extent, then plants may not be able to take full advantage of the resource in ample supply. Species might have different levels of tolerance for the extent of resource congruency in a given plant community (Bazzaz, 1996). Additional research by our laboratory at Harvard Forest has demonstrated that, for seedlings in a canopy gap where environmental conditions show distinct temporal patterns, resource congruency could play an important role in permitting species coexistence (see papers by Peter Wayne and Gary Carlton).

C. Resource Partitioning: Is This Sufficient to Understand Species Diversity?

There is now good empirical evidence for substantial resource partitioning between species in the same community. Underlying variation in soil resources and disturbance-mediated changes in the aboveground microenvironment provide primary axes for species specialization. Further axes may be produced by responses to multiple environmental gradients, e.g., resource congruency, and by substantial temporal separation of developmental events, such as germination and flowering. It is not clear, however, that these differences are adequate to prevent competitive exclusion in the long term, as the theoretical considerations for community-level consequences of niche differentiation in multidimensional space have not been adequately explored (see Section III). In transient, early successional communities, this discussion may not be relevant, but in communities with a long-term evolutionary history, we certainly need to consider whether resource partitioning plays a central role in community diversity. Communities characterized by long-term species’ interactions, such as the tropical rain forest and the tallgrass prairie, typically contain a high number of species, and
so it is critical that we unequivocally establish the role of niche differentiation in maintenance of species diversity. To date, experimental evidence suggests that resource partitioning is not sufficient to fully explain the high species richness of some such communities, e.g., tropical rain forest. Work by Hubbell et al. (1999) on Barro Colorado Island in Panama demonstrated that gaps did not explain variation in species richness. Gap formation did increase seedling establishment, but this effect was nonspecific and unpredictable.

V. ADDITIONAL MECHANISMS FOR MAINTENANCE OF SPECIES DIVERSITY

Resource partitioning is clearly not the only diversity mechanism operating in plant communities containing a large number of species. Species' specialization on different portions of multidimensional resource space prevents competitive exclusion by reducing the degree of interspecific interactions. However, other processes within plant communities might also reduce the extent or slow the rate of competitive exclusion. In this section, we discuss three additional mechanisms that have been suggested as causes of maintaining plant diversity.

A. Frequency-Dependent Effects

Mathematical models of competition clearly demonstrate that stable coexistence is possible if intraspecific interactions are stronger than interspecific ones. Weak interspecific interactions commonly arise from reciprocal selection acting over evolutionary time to produce niche differentiation. However, other intraspecific density-dependent effects could act to effectively reduce interspecific interactions. A mechanism for such density-dependent effects in species-rich plant communities like tropical rain forest was first suggested by Joseph Connell and Daniel Janzen. They both argued that the most common and dominant species in the community would be most commonly attacked by herbivores and pathogens, resulting in community-level compensatory mortality. In this way, no single species would be able to dominate the community and reduce overall species diversity. Host-specific herbivores or pathogens could act in a similar compensatory manner, increasing the average distance between parent plants and effective offspring recruitment. Currently, however, empirical support for frequency-dependent mortality in tropical rain forests has been variable. Broad-scale examination of size-specific population dynamics of tropical trees has demonstrated negative density-dependent effects in some cases, e.g., the work of Cam Webb and David Peart in Borneo, but not others (Hubbell and Foster, 1986).

B. Chance

The strongest alternative theory to the hypothesis that resource partitioning determines community diversity proposes that plant communities are dynamic entities containing a random assemblage of species not organized by reciprocal selection (see Clements vs Gleason dichotomy earlier). According to this view, higher diversity communities result from larger regional species pools rather than from species dividing up the resource environment more finely. Under these so-called lottery model scenarios, coexistence is possible as all individuals have an equal chance of reaching a particular location and whichever individual reaches the spot first will successfully establish there.

Finding little evidence for compensatory mortality in tropical rain forests (see above), Hubbell and Foster (1986) have argued that these lottery models may well represent tropical rain forest dynamics and explain their high species diversity. In these communities, beyond the basic early vs late successional dichotomy, many tropical forest species are essentially ecologically identical. Fluctuations in numbers result simply from chance events, and these fluctuations lead to a random walk in population densities with no stabilizing tendencies at all. Species coexistence is possible because the time for competitive elimination is even longer than the rate of new species formation.

Can we reconcile the niche differentiation vs chance perspectives on tropical rain forest diversity? It might be that the relative importance of deterministic and stochastic processes in community structure depends on the characteristics of the particular community. Chance could play a more important role in community diversity when communities are dynamic and have a fast rate of change, e.g., the tropical forest on Barro Colorado Island. In contrast, in older communities, where the rate of change is much slower and where there is likely to have been considerable sympatric speciation, e.g., the forests of Southeast Asia, niche differentiation could be the more dominant process maintaining species diversity. Until we develop a broad conceptual model of plant communities that incorporates both deterministic and stochastic processes, it will be difficult to test these ideas explicitly. Currently, the relative importance of niche differentiation vs chance...
and dependent on resource conditions (soil moisture, As the density of the understory is spatially variable act as filters for seedlings trying to ascend to the canopy. the eastern United States. These dense understory layers are dominated by a specific species, such as understory is a critical consideration for plant diversity, and the way that biodiversity is interpreted will depend on the goals of the study. Understanding the functional significance of biological diversity for ecosystem pro-
RESOURCE PARTITIONING

FIGURE 8 Variation in seedling demography in experimental fern plots: ferns removed (light shading) and ferns present (dark shading). Redrawn from George and Bazzaz (1999), Ecology, 80, 833–846, 846–856.

cesses has become particularly relevant for assessing the full impact of species loss on the functioning of the biosphere.

Any attempts to determine the underlying causes for global diversity patterns rely on a mechanistic understanding of the processes driving species diversity in ecological communities. We have evaluated the extent to which we can understand species diversity from a resource-based approach. Theory suggests that long-term species coexistence is possible if species’ responses to their environment are sufficiently separated in multi-dimensional resource space. Empirical evidence from a wide range of plant communities supports the notion that species separate themselves along spatial and temporal resource axes. As yet, however, we have not been able to combine this information with theoretical models to establish if the degree of separation is sufficient to favor long-term coexistence and maintenance of high species diversity. Certainly, in communities with very high species diversity such as tropical rain forest, it becomes more challenging to understand coexistence in terms of resource partitioning. Other mechanisms for maintenance of diversity such as frequency-dependent mortality, chance, and spatial structure may play more significant roles in these systems. Now that we have established likely contenders for mechanisms of species coexistence, it is important that we develop a comprehensive, integrative conceptual model that can incorporate these distinct factors into a cohesive framework of biological diversity.

See Also the Following Articles

DIVERSITY, TAXONOMIC VERSUS FUNCTIONAL • HABITAT AND NICHE, CONCEPT OF • SPECIES COEXISTENCE • SPECIES DIVERSITY, OVERVIEW • STABILITY, CONCEPT OF

Bibliography


RESOURCE PARTITIONING


RESTORATION OF ANIMAL, PLANT, AND MICROBIAL DIVERSITY

Edith B. Allen,* Joel S. Brown,** and Michael F. Allen†
*University of California, †University of Illinois, and ‡Center for Conservation Biology, University of California

I. Restoration Goals: Conservation and Biodiversity
II. Restoring a Diverse Plant Community
III. Restoring Animal Diversity
IV. Promoting Microbial Diversity
V. Conclusions

GLOSSARY

active restoration Requiring manipulation by humans for successful colonization and/or establishment of organisms and ecosystem functioning.

functional diversity Having all of the functions required for maintenance of ecosystem processes, but not necessarily all of the species richness.

functional redundancy, functional similarity Having species that may be substituted because their contributions to ecosystem processes are similar or overlapping.

passive restoration Relying on natural successional processes for restoration after the stresses that caused the disturbance have been removed.

reclamation A revegetation or land management goal that includes a lower diversity of species and may include substitutions by introduced species.

reference area An undisturbed or natural area chosen to compare with a restored site to determine the success of restoration.

rehabilitation Creation of an alternative ecosystem following a disturbance, different from the original and having utilitarian rather than conservation values.

restoration The manipulation of organisms and ecological processes to create self-organizing ecosystems that resemble predisturbance structure and functioning and promote conservation of biodiversity.

CONSERVATION OF BIODIVERSITY is the central goal of most restoration efforts and ranges from reintroductions of individual species of rare plants and animals to efforts to reintroduce a high diversity of species. Restoration may be defined as the manipulation of organisms and ecological processes to create self-organizing, sustainable, native ecosystems as integral parts of the landscape, as much as possible as they existed before disruptive human disturbances. In this article, we will examine the possibilities and the limits to restoration of biodiversity.

I. RESTORATION GOALS: CONSERVATION AND BIODIVERSITY

Conservation of biodiversity is the central goal of most restoration efforts and ranges from reintroductions of
individual species of rare plants and animals to efforts to reintroduce a high diversity of species (Jordan et al., 1987, 1988; Bowles and Whelan, 1994; Falk et al., 1996). Restoration may be defined as the manipulation of organisms and ecological processes to create self-organizing, sustainable, native ecosystems as integral parts of the landscape, as much as possible as they existed before disruptive human disturbances. Where propagules of native organisms are remnant, restoration may require reintroducing ecosystem functions such as fire or hydrologic regime to enable natural recolonization and recovery processes. Re-creation of a close replica of a previously existing ecosystem type is increasingly difficult in a world with a growing human population, considering fragmentation and limits to dispersal of organisms; global invasions of exotic species that cause large-scale replacement and even extinction of native species; air, water, and soil eutrophication; and habitat loss by land conversion to urbanization and agriculture. Even nature reserves and parks suffer from visitor overutilization and other impacts, and restoration coupled with more careful management is required to preserve the original flora and fauna.

There are limits to restoration whether biodiversity or functioning of the ecosystem is concerned, but restoring all of the species richness that was originally present on a site is usually more difficult than restoring similar functioning. The reasons are that many rare species are difficult to propagate, their basic biology has often not been studied, they have lost genetic diversity, and their propagules are limited due to habitat loss. They are also less likely to be chosen for many kinds of restoration because they are more expensive to reintroduce and they contribute relatively little as individual species to ecosystem functioning. On the other hand, abundant species are generally better known ecologically, and in some cases, individual abundant species will regulate ecosystem functioning. Restorationists more often focus on reintroducing the abundant plant species or the “matrix” species to imitate succession and recovery of disturbed lands. Reintroducing dominant species enables a recovery of major functions and most of the vegetation vertical and horizontal structure. For instance, John Ewel showed that low-diversity replanted (but not restored) tropical forest in Costa Rica may have similar levels of soil nutrients, organic matter, and nutrient loss, compared to natural forest, but will not have the same conservation value for species preservation.

Additional limits to restoration of diversity are economic, political, and social. Restoration for the intrinsic value of nature, with its complement of biodiversity, often includes participation by laypersons who wish to see nature returned to a state they remember from years ago or that was documented in historical texts or anecdotally. Lesser goals than restoration include reclamation or rehabilitation (National Academy of Sciences, 1974), which have utilitarian values and less emphasis on conservation and biodiversity values. Reclamation may include a less diverse mix of species and exotic substitutions, while rehabilitation is simply to make the land useful again and may produce an alternative ecosystem, such as turning a forest into a pasture.

In this article, we will examine the possibilities and the limits to restoration of biodiversity. Different species have different limits and require different approaches for restoration. We divide the discussion into restoration of plants, animals, and soil microorganisms. This is admittedly an overly simplistic approach, as they all interact. However, points of overlap among the groups are brought out numerous times within each section. The division is logical in that the three groups require different degrees of active or passive restoration, depending on the level of disturbance of a site (Fig. 1). The kinds of disturbances that require restoration are varied. They range from drastic alteration of the ecosystem, such as surface mining or other construction projects that remove the topsoil, to abandonment from agriculture in the northeastern United States, passive restoration may suffice. For animals, habitat may be created by manipulating the structure of the plant community, with the hope that the animals will recolonize. This has come to be called the “build it and they will come” hypothesis (see article by Palmer et al. in Restoration Ecology, Vol. 5, 1997).

However, recolonization may be restricted by fragmentation, lack of corridors, or lack of propagules, so re-introduction of the animal into restored or intact vegetation may be needed.

For microorganisms, passive restoration is generally the rule in contrast to higher organisms that may be purposefully reintroduced. Very few species of microorganisms that are members of natural ecosystems are
RESTORATION OF ANIMAL, PLANT, AND MICROBIAL DIVERSITY

FIGURE 1

A comparison of active and passive restoration. Active restoration requires various degrees of human intervention, depending on the degree of disturbance, while passive restoration only requires removal of the stresses that caused the initial disturbance, such as grazing or pollution, followed by natural recovery processes.

The goals for restoration may vary for these three groups of organisms. For plants and animals, species diversity concerns are high, but for soil microorganisms, functional diversity is more often expressed as the concern. The limits and possibilities for restoration of diversity vary for the three groups, and the discussion below will expand upon these topics.

II. RESTORING A DIVERSE PLANT COMMUNITY

To improve plant species diversity, the kind of revegetation or vegetation management is determined by the degree of disturbance. These form a gradient of disturbance types and recovery possibilities that require the entire range of active to passive restoration. Heavily disturbed sites such as surface mines need to have all cultured, so they are not even available for restoration purposes. The exceptions are symbions (N-fixing bacteria, mycorrhizae) that are used routinely in agriculture and reforestation, but less often in restoration. All the other taxa that contribute to belowground functioning of native ecosystems are never or rarely cultured, including soil saprotrophic fungi, bacteria, nematodes, and other soil micro- and mesofauna. For all of these groups, we must learn how to manage the soil to promote their recolonization and to understand the distances from which they may be able to recolonize. In addition, there may be 1000s of species of microorganisms even in 1 cm³ of soil. So the argument must be made whether all these species can or should be reintroduced. Many ecologists argue that there is functional redundancy among so many microbial species and that a minimum number of functional groups rather than a minimum number of species are required for restoration of ecosystem functioning.
species reintroduced, whereas slightly disturbed sites may be missing certain species that are not adapted to the new postdisturbance conditions, such as alteration of fire regime. Other sites may require weed control to increase native species diversity. There are no examples of severely disturbed sites that have been restored to their original complement of species. This may be related to the high costs involved, the changed environmental conditions, and the sheer impossibility of artificially reintroducing a large number of species. As richness increases in more productive habitats, the probability of reintroducing even a majority of the species diminishes. One of the best examples of attempts to restore a diversity of species comes from bauxite mining in southwestern Australia, where 80 species or more may be included in the seed mix. However, the adjacent undisturbed jarrah (Eucalyptus marginata) forest may contain more than 200 species. In Wyoming sagebrush–grasslands, mining reclamation regulations are not as strict, and typically only 3–10 native species are planted in an area that may have 50 or more naturally occurring (Fig. 2). The dominant rather than the rare species are typically chosen for revegetation, and the long “tail” of rare species that typifies a dominance–diversity curve for natural areas is missing (Fig. 2). Reclamation and restoration both generally create com-

![Figure 2](image_url) Dominance–diversity curves for reclaimed and natural vegetation in Wyoming sagebrush–grassland. The long tail of inabundant species found in natural vegetation is seldom reestablished in revegetated areas.
ommunities of a few abundant species but few rare species, whereas undisturbed communities may have the same few abundant species, but in addition will have many rare species.

A. Limits to Restoring a Diverse Plant Community

The reasons for omitting the many species that form the rare species tail in Fig. 2 are many. Seeds or other propagules must be collected locally for the restoration to reflect the local genetic populations, but these are typically not available unless the project is planned well in advance, usually two growing seasons, and the seeds are collected specifically for that project. Collecting native seed is becoming a large industry but it is unusual to have all the species available on the open market at the time they are needed, and even rarer to have the seed collections from local populations (see chapters in Falk et al., 1996).

Loss of genetic diversity and lack of local ecotypes are also a limitation to restoration. Where local extirpations have occurred, the nearest populations may no longer have all the genetic diversity of the original, and restoration in the true sense of restoring genetic as well as species diversity is no longer possible. Locally selected ecotypes have the “home team advantage” of being better adapted to local conditions and also avoid problems of outcrossing and hybrid depression with remnant native individuals (see article by Montalvo et al. in Restoration Ecology, Vol. 5, 1997). Additional discussion of genetic issues is in the section on animal restoration.

Once the appropriate local seeds or propagules have been collected, there is little information on seed dormancy and propagation of most species. The emphasis on conservation biology in the past decade has resulted in increased concern for research on rare species, and information on propagation, microenvironment requirements, reasons for disappearance, interactions with pollinators and other species, and so forth have enabled restorationists to include rare species in revegetation plans. More often, restorationists work with relatively unknown species and must begin research anew for each species, as, for example, the work on restoring the endangered Amsinckia grandiflora (see chapter by Pavlik in Falk et al., 1996). Lack of biological information on species translates into practical economic limitations. Most often, only the species that are best understood with the most available propagules are used.

Once germination and propagation requirements are understood, the plants must go into a field setting that presents a whole new set of problems. Different species germinate at different times and have different growth rates, so some will never emerge from certain seed mixes. The northern Great Plains of the United States are dominated by Bouteloua gracilis in the coal mining regions, but this has proven to be one of the most difficult species to reestablish for mining reclamation. It is a slow-growing, late-germinating, warm-season grass, but when it is seeded as part of a mix of native species, the cool-season grasses germinate first, grow quickly, and dominate. Reclamationists have devised numerous methods to reestablish shortgrass prairie, such as alternating seed drill rows or planting Bouteloua seeds a year earlier. However, the most important consideration for companies that have spent millions of dollars on earth-moving is to stabilize the soil, rather than to establish high levels of diversity that are not required by law in any case. Preventing soil erosion is the first goal, and the mines of this region use a fast-growing native plant mix that reduces the establishment of slow-growing species. The goals of soil stabilization using productive plant species and establishing a diverse mixture are often at odds (Fig. 3).

Reestablishing the full complement of species may require reintroduction in stages, as shown in examples from the tallgrass prairie. When Robert Betz began prairie restoration in the late 1960s at the Fermi Lab prairie in Illinois, he quickly learned that the dominant native grasses could be readily established from seed. He used the grasses to form the matrix of vegetation, followed by later introductions of the less common forbs. These forbs could not colonize naturally into the restored grassland because of the high density of grasses, but could be hand planted. Thus Betz was able to restore 116 plant species to the Fermi Lab prairie, but only by expensive hand labor. Similarly, the Curtis Prairie in Wisconsin, the oldest restored prairie, was initially planted during the 1930s, but plants have been continually introduced since then (see chapter by Cottam in Jordan et al., 1987). Their survival has been monitored, and they now contribute to the most diverse tallgrass prairie anywhere. The Curtis Prairie has over 300 species in 10 acres, more than any remnant natural prairie.

Competition from invasive plant species is another important limit to restoration of diversity. The invasives may be weeds that are part of the successional process, but a more difficult problem than early seral weeds are those invasive plant species that persist and cause vegetation type conversions. In this case, the major restoration activities involve mowing, fire, or selective weeding to remove the offensive species, planting or managing for regrowth of natives, and then continual management to keep the invasives from becoming dom-
FIGURE 3  The relationships among soil fertilizer level, plant species richness, plant productivity, and soil erosion. The trade-off for using fertilizer to promote high initial rates of plant productivity to reduce soil erosion is a loss in diversity.

imant once again (see chapters in Allen, 1988; Falk et al., 1996). Such activities have been undertaken in Florida, where Schinus terebinthifolium and Melaleuca quinquenervia dominate wetlands, Hawaii, where exotic perennial grasses and Myrica faya have replaced native vegetation, California, where Mediterranean annual grasses have replaced native perennial grasslands, and in many other sites worldwide (Lee, Richardson in Encyclopedia of Biodiversity).

Poor, toxic, eutrophied, or erosive soils also impair the conservation value of restored communities. Any practice that changes the soil from its original physical or chemical state will also change the plant species composition. An example comes from the china clay mines in England, where the mining refuse is high in sand. Local plants are poorly adapted to the droughty and nutrient-poor sand, so strand plants are imported for revegetation (Bradshaw and Chadwick, 1980). As expected, these artificial strand communities are poor in diversity. However, Bradshaw has documented some relatively high species diversity following natural colonization onto abandoned waste sites, although not as high as surrounding native areas. There is a tendency during revegetation to overfertilize soils to agricultural levels when fertilizer is applied, with expected poor results for native plant diversity. The reason for overfertilizing is usually economic, to promote vegetation establishment and therefore prevent soil erosion, but may also be aesthetic, to please the paying public as rapidly as possible. As shown by the experiments of Tilman (Tilman in Encyclopedia of Biodiversity), overfertilization, especially by nitrogen, will promote the productiv-

B. Improving Plant Species Diversity

Restorationists have used many techniques to increase richness of plant communities when the seed and propagule sources are limiting. Taking advantage of the existing seedbank is a primary one, where the seedbank still exists. Mined land reclamation laws in developed nations often require retopping with fresh topsoil, and many studies document the importance of this source. Certain species still need to be supplemented, as in the case of late seral species absent from the seedbank of jarrah forest in southwestern Australia (see chapter by Bell in Allen, 1988). Jarrah seed is never found in the seedbank, and it reproduces vegetatively after fire in natural communities. It must be reintroduced as nursery transplants for revegetation to be successful. While the seedbank is an important source of diversity in many kinds of restorations, it is also notori-

ous for harboring a large complement of early succes-

sional and weedy species. Bradshaw and Chadwick
(1980) have recommended against using topsoil just for this reason, if appropriate weed-free subsoil can be found and amended to support plant growth.

Dispersal of propagules into the restored site is another way to increase diversity that depends upon landscape structure such as proximity of source populations or existence of corridors. In a study of 2- to 18-year-old revegetated, untopsoiled roadsides in San Diego County, dispersal from adjacent mediterranean shrubland more than doubled the richness from 12 planted species to a maximum of 16 colonizing native species. However, dispersal of native species only occurred when the shrublands were adjacent to the highway, and occurred rarely in urban areas of the highway. Reclaimed surface mines in Wyoming that are surrounded by native sagebrush–grassland have a higher richness of colonizing species than, for instance, a re-vegetated landfill in Staten Island that is surrounded by suburbia (see articles by Bell and by Ehrenfeld in Restoration Ecology, Vol. 5, 1997). While dispersal is often limited by vectors in terrestrial habitats, aquatic and riparian habitats typically are recolonized rapidly (National Academy of Sciences, 1992). Water is a very effective medium for propagule dispersal, and these habitats are often not even revegetated, unless the soil needs to be stabilized rapidly in riparian edges or the dispersing propagules include unwanted exotic species. Vegetation may be replanted to attract animals that are seed dispersers or pollinators and thus may create a positive feedback on the future reproduction and diversity of a site. When a limited suite of species is chosen, the selected species become especially critical. Synthetic grasslands that simulated tallgrass prairie had sufficient diversity and especially structure to attract birds and small mammals (Howe and Brown, 1989). Shrub islands were planted on a surface mine in Wyoming to increase the movement of animals and microorganisms onto a site that was otherwise dominated by grasses (Allen, 1988). The feedback of these shrub species and patterns determined both their mortality and recolonization patterns of additional plant species. In Costa Rica, Karen Holl and Daniel Janzen have shown that tree and shrub ‘islands’ within pastures are critical for attracting animals that disperse seeds and then increase the diversity of the plant community.

C. Single-Species Plant Introductions

Individual plant species have been reintroduced where they were extirpated, largely because of laws protecting endangered species. Mitigation laws often require transplantation of a rare species from a site that is about to be destroyed to a safe site. For instance, Howald (see chapter in Falk et al., 1996) documents 40 instances where rare species were transplanted for habitat mitigation purposes in California. Of these, only 2 were considered successful, 7 had limited success, 13 were not successful, and others were unknown or ongoing and too recent to evaluate. The reasons for lack of success were varied but included moving plants to sites where they did not exist previously, having different environmental conditions in the transplant site, or using poor horticultural techniques. Primack (chapter in Falk et al., 1996) planted seeds of 41 species into sites in Massachusetts where they were once known to occur. Of these, only 10 produced seedlings, and only three produced a second generation. In this study, the reasons for failure may have been due to exotic species, changes in natural disturbance regime, and other changes such as anthropogenic nitrogen deposition that causes inappropriately high levels of soil fertility. The relatively few successful transplantations indicate that mitigation transplantation in general is poor policy because the kind of planning and postintroduction management that are required is seldom done in practice. Often little consideration is given to the quality of the habitat into which organisms are being reintroduced. Overall, the reintroduction of one species does little to improve the diversity of a site, but it may be the only option to avoid extinction, provided restorationists and land managers use their best practices.

While many single-species introductions focus on rare species, they may alternatively focus on abundant species. The reintroduction of a keystone species or ecosystem engineer (see Palmer et al. in Restoration Ecology, Vol. 3, 1997) is one way to ‘jump start’ natural successional processes in a community. Shrubs and trees were removed from rangelands for many years in an effort to increase forage production for domestic animals, but this was short-lived and in the long run promoted degradation of the community because the deep-rooted shrubs changed the nutrient and moisture balance of the entire stand. The reintroductions of Pistacia lentiscus in Israel, oak trees in California, or Artemisia tridentata in mined lands in Wyoming enable the shrub or tree island effect to reinitiate, building up soil and biotic resources around the transplanted shrub (Allen, 1988). This could potentially result in further increases in diversity as the shrubs enable recolonization of additional species.

D. Succession on Restored Lands

Restoration depends upon succession to complete the process that was started by revegetation, but succession alone does not always bring us to the goal we wish to
achieve. Revegetated lands are not static and can be expected to undergo change in species density, richness, and relative composition. The degree to which they undergo succession varies, depending upon the amount of colonization, initial species complement, and site conditions. Little change in species composition occurred after bauxite mining in the Australian jarrah forest. The site typified Egler’s initial floristics model of succession, where late seral dominants colonize in the earliest stages of succession, because very few species that did not occur in the original seed mix became dominant later on. Sites that were 16 years old appeared structurally different from sites that were 1–2 years old in that woody plants were larger and more evident in the older sites, but most of the same species were present in sites of all age classes. A few surprising species did colonize the older sites, such as a native orchid. In Anthony Bradshaw’s experiments on plants colonizing industrial waste heaps in England, most plants that did not establish appeared to be limited by dispersal and establishment, rather than poor substrate. The sites were too distant from propagule sources or had dense stands of initially colonizing species that prevented establishment of a diversity of species. The initial floristics model can also be applied to Wyoming, Pennsylvania, and Czech surface mined lands, where the native grasses dominated early on allowed little colonization of native trees, shrubs, or forbs after 10–30 years. Even when soil conditions are optimal and sources of propagules for local colonization are available, the initial revegetation treatments will affect long-term succession. The initial floristics of succession, where late seral dominants colonize in the earliest stages of succession, because very few species to colonize later. From these examples, and many others in the literature, it is apparent that succession may promote an increase in richness if the planted stand is not so aggressive that it precludes establishment of colonizers. Most often, additional interventions are needed if the destined diversity is to be achieved.

III. RESTORING ANIMAL DIVERSITY

A. An Introduction to Animal Restoration

Like plant restoration, animal restoration involves reintroducing or encouraging the return of native species to an area or region from which they have been lost. An animal restoration project may involve just a single species, as in the reintroduction of wolves into Yellowstone National Park, or an entire community, as in Augrubes Falls National Park in South Africa, where over 10 species of large mammals have been reintroduced (including black rhinoceros, eland, giraffe, and springbok). Animal and plant restorations will generally involve the same ecological considerations, and most differences between the two concern a matter of degree and emphasis (Jordan et al., 1988). Save for insects (which represent a challenging form of restoration more in line with promoting microbial diversity), plant restorations generally call for a larger number of species. In terms of ecosystem function, animals have significant roles in restoration as pollinators and herbivores of the plants. Animal roles in ecosystem function also occur when herbivores alter nutrient cycles and where animals may be “ecological engineers” of the physical environment, such as the contributions to soil function from the digging and foraging activities of earthworms or burrowing mammals (Jordan et al., 1987). In general, animals are more mobile than plants and while plants may disperse widely and unexpectedly as seeds and pollen, a single animal may range over scales that are much larger than that of the restoration project and site. Some of John Launder’s radiocollared mountain lions from south central Idaho have appeared as far away as Yellowstone National Park in Wyoming and mountain ranges in central Nevada. As a consequence of mobility, restoration sites from the perspective of animals will be open rather than closed systems. In animal restoration, the environs of the restoration site may deserve equal attention. Much attention has been given by all stakeholders to the possible fates and consequences of wolves wandering in and out of their Yellowstone restoration site.
The ecology of animal restoration draws less heavily from theories of ecological succession and more from the ecology of invasions and the ecology of small population sizes (Major, 1989). Plant restorations, as succession, start with an inoculum of plants that initiate a process of changes in species composition that eventually leads to the desired natural community in terms of structure, function, and appearance. This aspect of succession is absent from most animal restorations except in the case of the mostly passive development of an insect and soil invertebrate community that fits and complements the plant community. With most animal restorations, each species enters the community at a very small population size either as an invader or as a reintroduction. Small and alien describes the starting population. Most random introductions of animals to a community fail. Native animals fall to predators more easily, compete less successfully for resources, often fail to develop or find suitable denning or nesting sites, and sometimes attempt to emigrate from the site. In the absence of preexisting burrows, reintroduced prairie dogs simply roam far and wide in search of colonies. Small animal populations face twin genetic and demographic threats. Small, sexually reproducing populations may suffer inbreeding depression and produce a higher proportion of young with genetic defects. Demographically, small populations may teeter on the edge of extinction. An accidental death, a missed breeding opportunity, or a chance skewing of sex ratios or age distributions may compromise the population irrecoverably.

Animal restoration begins with knowledge of why the species or animal community is currently absent or threatened at the site. Next comes an assessment of whether the site is currently suitable for the animals and an evaluation of what site preparations are necessary to promote success. Then the current status of the animals receives attention to determine the best and likeliest sources of individuals. The animal(s) may be present on site but at low numbers, they may be present near the site or far from the site, or they may only exist on site but at low numbers, they may be present near the site or far from the site, or they may only exist as captive populations. Finally, the passive or active restoration of the animals begins with appropriate considerations of how to manage and monitor the project’s success. We will briefly examine each of these steps.

B. Reasons for the Absence of an Animal or Animal Community from a Site or Region

Habitat fragmentation or changes in land use may reduce or eliminate particular animal species, subsets of species, or entire taxonomic groups of animals from an area. For instance, most animals associate with particular types and structures of vegetation. When altered, by livestock grazing or agriculture for instance, some of the original animals disappear as they cannot find food, nesting, or denning sites or they may be excluded by changing intensities of competition or predation from other animals more suited to the new circumstances. In rivers and lakes, pollution, erosion, and sedimentation drastically change species composition. Everything from invertebrate larvae to fish may die off or be replaced by other species more tolerant of the polluted or modified waterways.

Hunting, poaching, and commercial harvesting may be so intense as to extirpate an animal or group of animals from an area or region. The African gray parrot and South American macaws face threats from those supplying the demand for pets. Incidental mortality in fishing nets threatens many of the world’s populations and species of sea turtle. Overharvest of commercially valuable species has created a litany of crashes such as California sardine, North Sea herring, blue whales, kaluga sturgeon, and southern bluefin tuna. As a valuable source of meat and a denizen of potential rangeland and farmland, the pronghorn antelope of the western half of the United States, like the American bison, faced extirpation by the 1900s.

Exotic species, those species accidentally or intentionally transported by humans into new places, often eliminate native animals via predation, competition, or keystone effects on structure and function of the ecosystem. The introduction of mosquitoes into Hawaii brought bird malaria for which the birds were no more prepared for than the peoples of the Americas were for the introduction of smallpox from Europe. An inoculum of larvae in ship ballast introduced zebra mussels from Europe into the Great Lakes of North America. In the early 1990s, the mussel’s spread was spectacular, and with it has come dramatic declines in the abundances of phytoplankton and the fish and invertebrates relying on the phytoplankton. Competition from introduced North American gray squirrels has eliminated the native European red squirrel from much of its former range in England and Wales.

With respect to restoring their wildlife, Australia and Israel provide interesting contrasts. Both have seen their native mammals ravaged. In Israel this has gone on dramatically for millennia whereas it is more recent in Australia. In Israel, overexploitation and habitat modification have been the bane of most animals like lions and crocodiles, both extirpated long ago. Deforestation, reaching its nadir under the Turks just prior to World War I, nearly extirpated the subspecies of European...
red squirrel and jay from Israel's wooded habitats. In Australia, wildlife habitat is much more available and overexploitation generally more benign than in the Middle East. However, introduced species (red foxes, feral house cats, and European rabbit) of animals and plants have dictated the decline in numbers and range of comparable Australian mammals. Having contributed to their extirpation on the mainland, cats now occupy Kangaroo Island off of southern Australia and threaten the last remaining population of Tamar wallaby. Feral goats compete with and threaten the dwindling population of yellow-footed rock wallabies in the Flinders Range. Red foxes prey heavily upon Australia's native small mammals and consume the eggs of ground-nesting birds.

The reasons for the absence of animal species from their former range are manifold but fall roughly into the categories of habitat change, overexploitation, or exclusion by animal species nonnative to the particular location. Understanding the reason for the absence of an animal or community of animals is the essential starting point for animal restoration.

C. Site Assessment and Habitat Restoration

All species have the capacity to grow exponentially under ideal conditions, but no population can grow exponentially forever. There are limits to growth, and population interactions such as competition and predation from other species can exclude a species from a community. The goal of site assessment and habitat restoration is to ensure near ideal conditions for the single species or the community of animals under restoration. Ideal conditions means ample food, space, shelter, and safety from predation. Limits to growth invites a consideration of specific factors in the environment that might currently or ultimately limit the growth and size of the animal's population. Considerations of competition and predation focus on which current species at the site, possibly exotic species, might preclude a successful introduction of the desired species to the site.

Site assessment evaluates whether the area offers an appropriate combination of food, safety, shelter, and space for a small population of animals to grow exponentially and prosper. If conditions are not appropriate, then habitat restoration must precede animal restoration. If the vegetation of a site has been heavily degraded, animal restoration may await plant restoration. At Midewin National Grassland south of Chicago, long-term hopes include first restoring the tallgrass prairie and then reintroducing elk and bison. The site is currently a mix of abandoned munitions facilities, farm fields, pastures and assorted oldfields, woodlots, and groves. Wetland restorations often require establishing the aquatic and surrounding vegetation and then passively or actively reintroducing the associated aquatic invertebrates (usually all passive), vertebrates (usually passive but nest boxes or nesting platforms may be used to encourage the return of waterfowl, egrets, or herons), and fish (sometimes passive or active). Different species have different needs, and reconstructing those habitats that fill an animal's needs is central to attracting wildlife or to introducing wildlife. The possible reasons for the animals' absence from a site focuses site assessment. Often several factors combine to explain an animal's absence. When habitat alteration or fragmentation changes and reduces the diversity of animals, the first step is to determine whether the site has been, or needs to be, revegetated to its former state. This condition of the site does not have to be exact with respect to its former state, but rather it need only include the salient environmental factors that favor the animal's ecological requirements and aptitudes. For instance, many bird species are most responsive to the structure rather than the exact composition of the vegetation. Hence, bird community restoration can commence with promoting the appropriate vegetation structure rather than being particular to its composition (which may be the goal of a corresponding plant restoration). Habitat fragmentation offers unique challenges because the remaining area may simply be too small to successfully support the desired animal or animal communities. The site assessment may examine the need for more space or wildlife easements, the need for corridors to connect habitat fragments, or the need to enhance the quality of a habitat fragment beyond its natural state to permit the successful persistence of animals in a smaller, more confined space. At Jackson Hole, Wyoming, a wintering elk population is maintained on supplemental hay within a fenced pasture surrounded by the extensive human developments that have robbed the elk of most of their wintering habitat. In spring, the elk return to their natural mountain pastures. In Israel and elsewhere, feeding stations stocked with animal carcasses or offal have been used to facilitate the maintenance and restoration of such animals as wolves, griffon vultures, and eagles.

The California condor provides an illustration of a complex site assessment. Reasons for the decline of condors that culminated in their removal from the wild in the late 1980s included too small a population (demographic threats), habitat fragmentation, lack of carcasses, and lead poisoning from carcasses containing...
lead shot or bullets. The restoration plan called for captive breeding, gazetting of habitat and habitat corridors, restrictions on development, and for reasons beyond just the condor, the banning of lead shot for hunting. The actual restoration took several approaches to evaluating the complexities of habitat suitability and the ability to introduce a small population of naive animals. First, less rare Andean condors were introduced as a surrogate species. They permitted tests of the reintroduction techniques and indicators of ranging patterns and habitat suitability. The actual reintroductions have used the original site north of Los Angeles but also additional sites, including one near the Grand Canyon. The California release site has more recently had condors but suffers much greater degrees of habitat fragmentation and development. The alternative sites have seen condors less recently, but have the advantage of offering extensive tracts of original and natural habitat.

In response to habitat degradation or fragmentation, habitat restoration may be as simple as adding nest boxes, as in the case of bluebirds at some sites in north-eastern Illinois, or as involved as adding proper soil, microbes, and restoring an entire vegetation community, as in the case of toxic waste sites or former mining operations. Ceasing the use of pesticides or release of pollutants may be all that is necessary to restore a habitat. Peregrine falcons in North America have benefited from the ban on DDT. Portions of fish communities and whole invertebrate communities can recover just from preventing sewage and pollution discharges into waterways. The structure of the environment with respect to shrub cover in arid lands, the mixture of ages and types of trees in forests, and the availability of salt licks and waterholes for wildlife all become considerations for particular habitat restorations. Koalas in Australia require particular species of eucalyptus trees as food, whereas gravel beds at particular depths provide spawning grounds for trout of the Great Lakes.

When a species absence is the consequence of over-exploitation, success often depends less on the availability of suitable habitat and more on the cessation of hunting, poisoning (including mortality from hazardous chemicals and pollutants), harvesting, or poaching. Wolves were exterminated throughout the western half of the United States through individual hunting and trapping and by explicit eradication programs. Their reintroductions into Yellowstone National Park, Wyoming, and the White Mountains of Arizona involve presumably moderate to high quality habitat. The success of the programs probably depends most on the cessation of poaching, which at present is a minor problem for Yellowstone and a major threat to the Arizona reintroduction. In 1989, it was discovered that stock assessments of cod within the Canadian North Atlantic were wildly optimistic. The stock had crashed. Subsequent quotas follow from the premise that habitat quality remains high and that reduced harvest will be sufficient to promote recovery. In Southeast Asia, edible-nest swiftlets face decline and extirpation. The birds nest in caves, use saliva for nest building, and face true “nest predation.” Humans gather the nests even as the nests may have eggs or nestlings. Even with complete bans on nest harvesting, places like Sarawak, Malaysia, still face intense nest destruction from local people, pirates from countries such as Indonesia, and organized collecting groups. Policing thousands of cave entrances in often remote places throughout the country is impractical. But, until poaching relaxes or ceases, restoration and recovery cannot be achieved. Having reduced poaching of black rhinoceros throughout the 1980s, the Kenyan Wildlife Service has begun the recovery and restoration of current and former populations.

Habitat restoration for overexploited species requires the first step of controlling the harvesting. However, the absence of wildlife from an area may coincide with other changes to the habitat, some which at first may seem subtle. “Nature abhors a vacuum” is a saying recognizing that unused food or opportunities in an environment often become filled by alternative species or by exotics. Habitat restoration may require controlling the abundance of exotic species or those species that compete or prey. Factors contributing to the decline of red squirrels in England include an increase in oaks and a decline in conifers. Both factors assist the exotic gray squirrel in outcompeting the red squirrel. Habitat restoration may require both changes in forest composition and active control measures of gray squirrels. The draining of reservoirs, netting, or poisoning of lakes infested with exotic fish such as carp has preceded several fish restorations in North America.

D. Sources of Animals for Restorations

Small remnant populations, dispersal from other areas, or active reintroductions provide the sources for animal restorations. Ideally, animal restoration should begin while a remnant population still occupies the site. Such a population, while often small, has the advantages of already being established and acclimated to the site. This avoids problems associated with naive animals unfamiliar with the site. In general, once habitat restoration has assured a quality site, resident populations recover faster than immigrants from other areas, which
recover faster than reintroduced populations from captive stock. When remnant populations exist, restoration activities can immediately focus on improving vegetation, improving food availability, improving removal of exotics or competitor species, and/or ceasing harvesting and pollution. Success is measured by the recovery of the species or the expansion of the species into previously unoccupied parts of the site.

When a restoration relies on immigrants, there are the twin concerns of: “Is the site suitable for the establishment of immigrants?” and “Will the immigrants find the site?” When a site such as an oldfield, mine tailings, or forest clearcut is restored, there will usually be successful invasions of insects, soil invertebrates, aquatic invertebrates, mammals, and birds from the environs. For instance, in an oldfield subject to small experimental prairie restorations (Howe and Brown, 1999), there was an invasion of two small mammals (prairie vole and white-footed mouse) and some red-wing blackbirds have shifted their breeding from a nearby pond to the prairie plots. Mourning doves preferentially feed on rather than off of the prairies. Inversions of insects, birds, and mammals ease the restoration effort but sacrifice complete control of the resulting composition of animals. Species that are already well established and abundant in the region of a restoration are the most likely invaders. Rare and remote species are the least likely. Target species may require coaxing in the form of habitat corridors connecting the restoration site with existing populations. Temporary augmentation with food or nesting boxes can encourage reestablishment. In Tsavo West National Park, Kenya, rhinos have been corralled into a sanctuary (ca. 70 km²) that is much smaller than the whole park (ca. 9000 km²). The sanctuary is fenced to keep the rhinos in, patrolled to protect the animals from poaching, and supplied with piped water to maintain permanent water holes.

Animal restoration may require the active reintroduction of animals (Bowles et al., 1994). This is best done by translocating individuals from wild populations, but may require captive breeding. Wild populations provide experienced individuals, whereas captive animals may be particularly naive and unfamiliar with the wild. The reintroduction population should have the appropriate balance of sexes and age classes and sufficient genetic diversity to preclude serious consequences of inbreeding. When the reintroduction is very expensive, is politically sensitive, or is from a very small source population, great care goes into ensuring that the site offers close to ideal conditions and that the animals are prepared for the new environment. When animals are relatively cheap and available, the reintroductions can be more numerous and there is room for experimentation. Wild asses (onagers) in the Negev Desert of Israel, pronghorn antelope in Arizona, lynx in Switzerland, white rhinoceros in Lake Nakuru of Kenya, and peregrine falcons in various midwestern cities all provide examples of reintroductions from captive or wild source populations. In the case of the Guam rail, this bird had been driven extinct on Guam by the introduction of the brown tree snake. Rather than attempting the impossible task of eradicating the snakes on Guam, the rail has been reintroduced on a neighboring, snake-free island.

IV. PROMOTING MICROBIAL DIVERSITY

A. Microbial Diversity and Functional Redundancy

Compared to plants and animals, microorganisms are highly diverse and offer a special challenge to understanding biodiversity and to assuring successful restoration. Microbes should be the underpinnings of any discussion of biodiversity as they constitute the vast majority of the diversity of any ecosystem at any location, yet are rarely even mentioned in terms of maintaining diversity. Microbes should be a focus, not an afterthought, for restoring disturbed lands. Without animals and most species of plants, ecosystems would stabilize and most ecosystem functions would be performed. Without microbes, the ecosystem would cease to function.

Estimates of microbial diversity range up to 10⁵ prokaryotic taxa per gram of soil, although 10³ is generally from more accepted value. In a study of restoration of a Wyoming coal surface mine, 57 different fungal taxa were found out of 135 colonies randomly chosen from approximately 12,300 colony sources (spores, hyphal fragments, bits of organic matter). These taxa were evaluated within a 4-cm² area (using 5 g of soil). Given the numbers, it is not surprising that a large fraction of these organisms remain undescribed. Recent estimates suggest that less than 10% of soil organisms remaining from bacteria to spiders have been described. And, a few efforts that extracted DNA directly from soils have even found new kingdoms of prokaryotes. Despite the difficulty in estimating diversity, every study published has shown a loss in the richness of taxa with severe disturbance. This ranges from soils in the Pumice Plain of Mount St. Helens, which was sterilized, to burned areas in which only the aboveground material and soil
surface were directly affected. Because of the extremely high microbial diversity, the concept of microbial functional redundancy has been raised. Among those thousands of species, how many need to be restored to maintain ecosystem functions? This is the single largest challenge for studying soil microbes and restoration.

There are three critical issues for evaluating microbial diversity and restoration. The first issue is defining the spatial and temporal distribution of species and functional groups and their relationship to ecosystem processes. The second concern is assessing the richness of organisms within the different functional groups. The third is the system-level capacity for dispersal and natural reestablishment versus the need for artificial introduction of microbial inoculum.

B. Spatial and Temporal Arrangement

Just as important as the richness of organisms are the changes in their spatial and temporal distributions. Unfortunately, few studies have evaluated microbial communities using species increment curves or overlap estimates. Several types of analyses are critical to understanding biodiversity and restoration. These include species × area, species × time, and dispersion relative to ecosystem processes. Unfortunately, few data sets are available to evaluate microbial recolonization on this basis.

Microbial richness estimates tend to be taken on a per-sample basis. However, plants and animals tend to be analyzed on an area basis. This makes comparative studies difficult but opens an important area of research. Nevertheless, evaluating the spatial array is absolutely critical. Fungi, for example, exist as a network of hyphae (a mycelium) extending from a few millimeters (such as a Trichoderma colony occupying a single fern petiole) to tens of meters in diameter (fairy rings, or the giant Wisconsin Armillaria, for example). For the same Wyoming data set, the species increment rate was the same up to the size of a 400-cm² patch of disturbed as well as reference area (Fig. 4). However, in the disturbed site, as one expanded outward, the species increment rate declined whereas it continued to increase in the reference area. In the reference area, new species were added as the habitat changed. In the disturbed area, the habitat was rather uniform across the site (in this case, mixed, respread topsoil on a surface mine). Thus, one conclusion is that microbial activity and composition become more diffuse and repetitive across scales in severely disturbed areas, and overall landscape diversity is lower than in native undisturbed areas.

Frequency of sampling over time is also crucial for describing microbial biodiversity. Many microorganisms are only identified based on sporulating structures. However, these organisms may be present continuously but can only be found periodically. For example, macrofungi are spread widely in the mycelial stage and live for many years. Several continuous years of observations are needed for the right conditions to occur before a sporocarp forms. In many cases, fruiting times may occur over successional time. For example, on Mount St. Helens, establishing the first ectomycorrhiza found on the Pumice Plain took 5 years. We never saw a sporocarp to identify the fungus that formed on the ectomycorrhizal root tips of conifers. Development of techniques for DNA fingerprinting will eventually allow us to identify more of these organisms even when they do not sporulate.

Samples are often taken at the wrong time, leading to erroneous conclusions about microbial diversity. For example, soil animals migrate vertically in response to soil moisture conditions. Soil samples are normally taken from surface soils. Thus, the mesofauna may only be detected when they return to the surface. Usually soil organisms are sampled at the convenience of the investigator, but microbial populations are often event
driven (e.g., precipitation). Thus, frequent or a deterministic sampling regime is needed to detect their presence.

Spatial relationships are just as crucial as temporal ones. Ecosystem processes (e.g., decomposition, mineralization, and immobilization) are not uniform across an undisturbed area. Microbial-regulated processes tend to be highly patchy and organized to optimize production. However, possibly the greatest impact of disturbance on ecosystem dynamics is spatially mixing soils and creating relatively uniform conditions across a site. In fact, this led to an oldfield view of succession that still largely dominates restoration practices, where a relatively uniform aboveground community is planted. However, succession may be a patchwork of starts and stops, with a few initial colonists acting as islands that become the nuclei for future colonists. Succession and microbial composition and activity are tightly coupled to the developing patchwork. In restored or recovering ecosystems, these patch recovery patterns are evident. In many abandoned, disturbed sites, no spatial recovery is detectable.

C. Diversity and Functional Groups

The biodiversity of types of microbes in ecosystems is daunting. However, to a certain degree, maintaining or recovering the functional groups of microbes is the first critical task in restoration efforts. Microbes play every ecosystem role at every site. In fact, microbes alone can, and do, form fully functioning ecosystems without higher plants or larger animals. In the most extreme environments of the Sahara desert and the uplands of the Antarctic Dry Valleys, microbes are the only living organisms, existing on aeolian-deposited or ancient carbon inputs. In many extreme environments, microbes make up the entire ecosystem. These range from the simple endolithic (inside rocks) communities of the Dry Valleys of Antarctica to the thermal pools of New Zealand and Yellowstone geysers. As one proceeds to more favorable environments, more and more types of microbes emerge, subdividing the processes of primary production and decomposition. At all sites, microbes undertake primary production (bacteria, cyanobacteria, algae). The relative contribution of the microbes to the overall proportion of net primary production tends to range from high in more extreme environments (such as deserts and tundra) or situations with dispersed nutrients at low concentrations (open oceans) to low in conditions highly favorable such as tropical rain forests.

In addition to directly fixing C, microbes also catalyze the nutrient cycling processes that transfer elements directly to plants or convert unavailable nutrients into forms that can be taken up and utilized. Thus, they are indirectly linked to carbon fixation by providing limiting resources. Although these organisms are generally modeled as “microbial mass,” they often live symbiotically with plants. Mycorrhizal fungi probably have the largest biomass within this group. These fungi form mutualisms with plants and transfer from soil to plant a range of soil resources, from water to N to P. Importantly, they can also make unavailable soil resources, such as bound P, available, by producing organic acids and phosphatases, and Fe with siderophores. Other prokaryotes fix atmospheric N\textsubscript{2}, ranging from free-living forms such as cyanobacteria and Arthrobacter, to symbionts such as Frankia and Rhizobium. Other microbes catalyze almost every other nutrient transformation that is biotically important to the sustainability of ecosystems, from N and S transformations, to Fe state transitions, to immobilization of heavy metals and bioremediation of toxic organics. In the case of mutualistic symbionts, many studies have demonstrated that an increasing diversity of species and genotypes can be critical to establishing and maintaining a diversity of plants.

Microbes are the dominant decomposers. Higher animals only take a small fraction (1–10%) of the NPP; the remainder of the energy goes to microbes. The animals themselves constitute a source of a slightly different C source from plant material, making a new type of C resource. Microbes then utilize almost all of the remaining plant material, thereby releasing the nutrients immobilized in plant tissues. Only a small fraction of C remains, as highly complex plant constituents or recalcitrant microbial compounds. These are critical in that this forms the organic matter essential to recovery of all sites.

1. Free-Living Saprobos

In every study, microbial diversity even of disturbed lands continues to increase with increasing sampling. While the actual slope of species increment curves may be lower than for undisturbed areas, it still remains very high. It is not clear if the reduced diversity of microbes is a factor in these detrimental responses. However, it is clear that if the environmental conditions for free-living microorganisms are present, a high diversity of species and a high density of individual cells will reestablish. Thus, restoration of free-living microbes is largely a matter of management of the soils, rarely by inoculation with bags of “beneficial” microorganisms. To date, there is no evidence that biodiversity per se
of free-living microorganisms limits saprobic microbial activity in restored lands. Free-living saprobes form the bulk of the microbial diversity in both functional pathways and the diversity of taxa. As we look at the known studies, those processes catalyzed by free-living microbes always occur, sometimes in detrimental levels. For example, Thiobacillus ferrooxidans uses Fe^{2+} in pyrite, which results in the release of sulfuric acid, detrimentally reducing the pH of streams. Immediately following disturbance, there is a rapid mineralization of N, resulting in N leaching and denitrification. Reduction in some of these microbial-catalyzed processes often is an important restoration task.

2. Soil Animals and Food Webs
Microbes consist of prokaryotes and fungi. These are capable of immobilizing nutrients such as N in the presence of excess C. Soil formation is also dependent on mixing of surface organic matter down through the horizon. These activities are undertaken by a food web of enormous complexity. Soil animals generally invade rapidly, either dispersing directly or by moving with soils or other materials. Soil food webs are generally characterized using functional groups as the richness of species is simply too high to characterize in detail. Food web analyses indicate that there are distinct channels (Fig. 5) that can be affected by the soil conditions and the composition of the microbes. Undertaking detailed studies of the role of biodiversity in these food webs is a critical future task. We currently do not know if species changes really matter to the recovery process.

3. Symbionts
Symbiotic microorganisms are much less diverse and clearly play critical roles in the establishment and persistence of vascular plants and plant composition. These roles are basically of two types, pathogens that inhibit plant growth, and mutualists that extract resources and exchange those resources with plants for energy or provide protection in some form, again in exchange for energy.

Plant pathogens are of two basic types for our purposes, specialists and generalists. Specialist pathogens are those that are associated with only a single species or group of host species. They tend to be highly diverse. Generalist pathogens tend to be widely spread across plant groups. Specialist pathogens are known to be devastating in agricultural ecosystems. However, they tend to be much less of a problem in restoration efforts. This probably results from the efforts made to restore a diversity of plants, making it more difficult for a pathogen to find a host and build up adequate inoculum densities. Exceptions exist when there is a high prevalence of a single species coupled with an exotic introduction. Generalist pathogens may be another matter. These are highly diverse organisms that often live as saprobes except when conditions prove favorable to a parasitic lifestyle. For example, there are a wide variety of Fusarium and Phymoza fungi found in virtually all soils. These can destroy a wide variety of plants under appropriate conditions. Phytophthora cinnamomi is responsible for loss of plants ranging such as the eucalyptus in the Jarrah forest in Australia. Often, these are almost undetectable except for very short times. In Wyoming, snow mold reduced sagebrush densities up to 60% and reduced growth in the survivors. This "mould" was a complex mix of fungi, not Typhula sp. found in the snow mold diseases of wheat. The disease was opportunistic and only found during El Nino years of high autumn rains and locations of high snowfall accumulation. It was found only one year and only in locations of high snowfall. Plant parasitic nematodes are always found in soils. They are responsible for high levels grazing, but, remarkably, rarely can nematode damage be observed in a restoration project.

Thus, despite the examples where disease was present, there are remarkably few demonstrations where diseases were highly diverse or markedly changed the outcomes of a restoration effort. Even under rather
optimum conditions, such as tropical seasonal forests, we have observed few instances of root or shoot disease and then, it tended to be single root tips or individual leaves, but not widespread across a site. This supports the need to establish a diverse plant community. Clearly, it is generally not desirable to restore pathogens to a restoration site. Mutualistic symbionts are relatively diverse but that diversity may play unique roles in restoration. Probably the best known are N-fixing prokaryotes. Legumes tend to be important early colonizers as N often limits primary production. In croplands, nodulation tends to be highly specific. This led to the all too common practice of using commercial inoculum. However, at sites ranging from glacial outwash in Alaska to Mount St. Helens to a seasonal tropical forest, we have planted or observed invading legumes. In no case were legumes present. In well-established hot deserts, Bradyrhizobium appears to be largely associated with plasmids and not a nuclear genomic component. Thus, limitation in nodulation is likely not a function of the presence of rhizobia, but of the conditions of the site. The effectiveness of the nodulation, however, could be an important question. In well-established hot deserts, Bradyrhizobium was an efficient bacterium stimulating high rates of N fixation. However, it was slow-growing and deep in the soil profile. Rhizobium was fast-growing and found near the surface. It rapidly colonized plants but was not an effective fixer. In pasture soils, different rhizobia are distributed in patches scattered across a site. Thus, while the presence of rhizobia is likely not limiting to restoration, having a diversity of populations capable of acting with a range of plants under a range of conditions may be important.

We know far less about Frankia, although there is a wide diversity of associated plants. In bioassays of respread cold-desert soils, we found that the plants failed to become nodulated. However, the soils had high N concentrations, which may have restricted activity. Alternatively, invading species such as Russian Olive has nodules even in areas where it has previously not been found. Unfortunately, beyond just a few observations of groups and N fixation rates, there are no studies of the diversity of N-fixing species or genotypes in restored areas of these critical groups. Mycorrhizal fungi have been studied in much greater detail. Their diversity is highly variable. In desert sites, we have found as few as two or three species. Alternatively, in forests, there can be hundreds of species and thousands of genotypes. These fungi are often eliminated by the disturbance event. However, even when they are not completely eliminated, the diversity of species is often radically altered. Moreover, many species depend on a mycelial network that can extend up to many meters across. This spatial structure is always broken up, providing opportunities for new taxa to invade. The resulting pattern is an increase in the intraspecies diversity with more, smaller clones. As these clones expand, some die and disappear while others continue expanding into the open habitats. Thus, intraspecific diversity initially increases as many propagules arrive and then declines as fewer colonies come to predominate.

Recovery of symbionts is a critical limiting step in restoration. There are two limiting steps: first, invasion of propagules, and second, establishment on site. Invasion is by physical or biological vectors. The most notable physical vector is wind. Wind has been shown to move organisms as large as mycorrhizal fungal spores up to 2 km. However, there are important limitations. Spores larger than 70–100 µm in diameter are rarely wind-dispersed. In those cases, animals are the vectors for microbes. Many animals feed on microbial spores. This can occur directly. For example, the diet of many rodents can be predominantly fungal sporocarps. This was the major means for mycorrhizal recolonization on Mount St. Helens following the eruption. Other propagules are transported unintentionally. Ungulates and rabbits feed on forbs and grasses, but in doing so, they tear plants from the soil, bringing fungal hyphae and internal spores and vesicles. Animals such as gophers preferentially feed on the nodules of legumes in addition to mycorrhizal fungi. The microbes are adapted to pass through the guts and are deposited across restored sites. Thus, just as for plants and animals, a key factor in restoration is the proximity of the source areas (Fig. 6).

D. Microbial Establishment

Different microbial species have different abilities to reestablish on a disturbed site. Because of their remarkable diversity, we are unable to artificially return even a small fraction of the microbes necessary for successful restoration. Thus, dispersal from surrounding areas is critical. In all of these invasions, two factors emerge as critical: distance and directionality for the appropriate vectors, and a suitable site. Adjacent source areas are important for reinvasion. At Mount St. Helens, for example, disturbed areas within or adjacent to surviving patches were rapidly recolonized. This took several years for sites at a greater distance to recover (Fig. 6). This pattern can be found in many other areas. Both physical and biotic vectors travel along specific pathways. These
When topsoil is limiting, it can be stored for a short while and respread. However, mining studies have repeatedly demonstrated that although mine moves, the newly stripped topsoil can be immediately replaced into the newly restored area. This facilitates microbial recovery rapidly. Salvaged transplants would have the same effect of introducing microbes into a disturbed soil.

In circumstances where soil microorganisms are completely lost, it may be necessary to inoculate. We know of no cases where inoculation of bacteria or saprophytic fungi (primarily imperfect fungi and bacteria) came in almost immediately. Symbiotic nitrogen-fixing bacteria established saprophytic fungi (primarily imperfect fungi and bacteria) came in 2–5 years. Area A was largely invaded from vectors moving within 2–5 years. Area A was largely invaded from vectors moving 2–5 years. Area A was largely invaded from vectors moving 2–5 years. Area A was largely invaded from vectors moving 2–5 years. Area A was largely invaded from vectors moving 2–5 years.

A. Assessing Restoration Success

Some protocols require the addition of symbiotic mutualists, but all others assume that microbes recover just fine and will “do their jobs.” In fact, soil microbes probably never “stabilize.” Their short individual life spans, coupled with the ability of some members of each functional group to invade and establish, makes assessment of composition and activity difficult. Inoculation can be an important practice in conditions where little or no inoculum for an entire functional group remains and has little chance for reinvasion. However, restoration requires management of soils, plants, and animals to encourage natural migration, patch structure for concentration of resources, and a complex structure that facilitates spatial and temporal diversity in ecosystem processes. If these conditions are met, it is likely that microbes will be capable of taking care of themselves quite well.

V. CONCLUSIONS

These examples of restoration of plants, animals, and soil microorganisms all show the difficulties and limitations of restoration of the entire richness of a prior existing community. While dominant plants and animals may be reintroduced, microorganisms are all expected to recolonize naturally. The resultant lower diversity restored communities indicated that if preservation of biodiversity is the goal, then conservation prior to disturbance is the preferred alternative, rather than restoration after disturbance. In focusing on species richness, we have placed little emphasis on ecosystem functioning, even though restoration of functioning is one of the major goals of restoration. Natural ecosystems provide ecosystem services, such as water supply, oxygen, soil stability, natural products, and so forth for free. Reclamation or rehabilitation is usually sufficient to provide these basic services, without the necessity of reintroducing all of the original biodiversity.
Measurements of both structure and functioning are used to assess restoration success. Restoration success is usually assessed by comparing the restoration site to a reference area, a native site with structure and functioning that are predetermined as the restoration goal. Measurements are compared between the reference and restoration sites. Structural measurements, such as the richness, density, and relative composition of species, are easier to measure than functional measures such as decomposition, nutrient cycling, erosion rate, or biological functions such as species reproduction and mortality or food web energy throughput. Yet it is the functional measurements that we need to determine whether the restored land has really stabilized, not simply the relatively easy measurements that require species counts. Measurements of restoration success are often not legally required, so many restorations/reclamation efforts receive no assessment at all. When they are, a species count, density, or percent cover is often all that is required to declare success.

B. Designer Ecosystems

Preservation of certain rare species may require manipulation of the ecosystem to stabilize their populations, possibly to the detriment of associated species. Such actions are already taken in numerous situations. For instance, wetland parks for shore birds have been diked, dredged, and damned to create aquatic habitat for bird species with different water depth requirements. Pastures in Europe receiving high anthropogenic nitrogen deposition are mowed at critical times of the year to reduce the growth of nitrophilous-dominant plants and promote survival of rare plant species. James MacMahon has termed these “designer ecosystems” because they are highly managed ecosystems that have a specific conservation goal, compared to the ecosystem where the species in question may occur naturally. Biodiversity has become highly manipulated in many areas where human populations are dense and where the remnant landscape is managed to promote as high a diversity of species as possible. Virtually all restored communities are missing species, so in one sense restoration may be considered an unintended experiment to determine the impacts of rare or other missing species on community and ecosystem functioning. This will require more research and monitoring than has been done in the past. One aspect of ecological restoration that has not been emphasized in this article is the general lack of data. Many sites are restored that have never received any kind of monitoring or research, or the data are simply not available. The generalization that restoration will not return the original diversity holds for the limited number of sites that have been studied, but as more data become available, we will understand more about how to manipulate ecosystems to maximize diversity.

Acknowledgments

We thank Karen Hall and Zev Naveh for reviewing the manuscript. The research reported here was supported by grants from the Division of Environmental Biology, National Science Foundation, and the National Research Initiative, U.S. Department of Agriculture.

See Also the Following Articles

CAPTIVE BREEDING AND REINTRODUCTION • CONSERVATION BIOLOGY, DISCIPLINE OF • CONSERVATION EFFORTS, CONTEMPORARY • MICROBIAL BIODIVERSITY • PLANT CONSERVATION, OVERVIEW • REFORESTATION • RESTORATION OF BIODIVERSITY, OVERVIEW

Bibliography

RESTORATION OF BIODIVERSITY, OVERVIEW

Joy B. Zedler, Roberto Lindig-Cisneros, Cristina Bonilla-Warford, and Isa Woo
University of Wisconsin

1. Introduction
2. Multispecies Approaches
3. Single-Species Approaches
4. Novel Opportunities for Restoration of Biodiversity
5. Conclusion

GLOSSARY

assembly rules Constraints that one community imposes on subsequent configurations of that community.

habitat creation Construction of one habitat type from another, often a disturbed upland excavated to make a wetland.

matrix species Dominant plant with broad coverage.

nonindigenous species A species occurring beyond its natural range or potential natural dispersal range. Synonyms used here are exotic and alien species.

nurse plant A plant that shelters and facilitates the growth of others.

propagules Reproductive units (spore, seed, bulb, cyst, egg, bud, larva, etc.) that give rise to new individuals.

restoration The return of an ecosystem to its condition prior to disturbance, or to that of a nearby reference system (representative, little-disturbed ecosystem).

HABITATS THAT HAVE LOST populations of native species have potential for biodiversity restoration, that is, the return of species-rich conditions. Attempts to restore biodiversity involve multispecies efforts (e.g., sowing seed mixtures or using restoration tools such as fire to encourage colonization or spread of native species) and single-species reintroductions. Genetic issues in biodiversity restoration involve the potential to reduce intraspecific diversity, especially in reintroduced populations of clonal plants and captive-reared animals. The deliberate introduction of nonindigenous species or species not known to occur naturally at a site does not constitute biodiversity restoration; however, the area formerly occupied by a taxon is rarely known in detail, making it difficult to distinguish reintroductions from deliberate range expansions. Spatial scale is also an issue, as a taxon may be restored to a region without being restored to its exact historical location. Several assumptions underlie biodiversity restoration efforts, namely, that ecologists know what biodiversity elements historically occurred at the restoration site, that the site can support more taxa, that dispersal and environmental constraints can be overcome, and that suitable plant and animal propagules can be found on site or nearby for expansion or re-introduction. Tests of these assumptions are limited and rarely consider more than a single species or single site. In this article we review examples of multispecies and single-species approaches, reiterate concern that community-level resto-
The desire to restore biodiversity stems from a long history of biodiversity loss as described elsewhere in this encyclopedia. The result of habitat destruction and degradation in the United States is that 1175 native plant communities are rare (Grossman et al., 1994). Restoring biodiversity is thus a huge task—a part of the growing effort to protect and restore ecosystem structure and functioning that is being undertaken by governmental and nongovernmental organizations around the globe.

Ecologists are still struggling to understand what leads to biologically diverse communities at various spatial scales. Because our knowledge of how natural ecosystems support high species richness is inadequate, it is no surprise that there are few guidebooks for restoring diversity (although see Packard and Mutel, 1997). Yet the hypotheses that predict causes of richness are useful in identifying constraints at restoration sites. Natural communities that have high species richness share several features: great age (they have had time for speciation and time for many species to arrive by dispersal); heterogeneous habitats that can support more habitat specialists; large area (cf. species–area curves, island biogeography models); low-stress environments (e.g., tropical climates); and/or intermediate disturbance regimes. Few restoration sites have any or all of these attributes. Additional constraints on species richness in a natural community concern the sequence of species introductions, called assembly rules. For example, in greenhouse experiments several grassland plant species increased their germination rates when sown after another species (Eriksson and Eriksson, 1998), indicating the need for a facilitator. Many species require or benefit from the presence of facilitators (nurse plants, host plants, or mycorrhizae). Once established, species may be inhibited or extripated by competitors or predators. The concept of assembly rules is particularly attractive for restoration, but the development of rules has been limited to specific cases.

Ecologists are also unsure what dictates the sequence in which species are lost as habitats become degraded. In prairie remnants, legumes are readily eliminated as prairies become fragmented. Because small prairies rarely burn (and less nitrogen is volatilized), the soil can accumulate nitrogen, reducing the competitive advantage of nitrogen-fixing legumes. As indicated below, nutrient-rich sites often support fewer species of plants than nutrient-poor sites.

In every natural community, some species are common and others are more rare. There is little guidance on which species that drop out of a degraded habitat will be restored most easily. Rarity derives from several situations, including a requirement for rare microsites or dependence on other species that are rare. Species with high habitat specificity (e.g., host-specific insects that rely on rare plants) are more likely to drop out as habitats are lost. Panzer et al. (1993) identified 236 insects from the Chicago area that were dependent on habitat remnants and absent in more disturbed habitats.

An exhaustive study of this same region's vegetation showed that over two-thirds of the native plant species were remnant-dependent, with one-third having broader distributions. Thus, a minority of species occupies a majority of the landscape. Plant and animals species that are not remnant-dependent are considered adaptable and disturbance-tolerant.

There are concerns that the adaptable species will dominate restored habitats and landscapes and that restoration efforts will result in "generic communities" (Zedler, 1999). Habitat restoration involves disturbance, such as contouring and grading that creates topography lacking in heterogeneity. The species attracted to restored habitats are likely to be adaptable, disturbance-tolerant generalists. In prairie restoration efforts, the sowing of species-rich mixtures of grasses and forbs often results in dominance by a few species of grasses. Such low-quality restoration efforts have a floristic quality index of 2–5, compared to that of the original prairie's value of 20 or higher (Packard and Mutel, 1997). [Floristic quality indices consider both the number of species and their fidelity to high-quality habitats.]

The paucity of habitat specialists in restoration sites has many explanations. Some forbs may require specific germination sites, such as open patches disturbed by small mammals. Recently, the absence of diverse microflora (mycorrhizae) has been indicated as a cause of the low diversity of old field sites (van der Heijden et al., 1998). The scarcity of specific microsites and rarity of facilitators both limit reestablishment of rare species.

Efforts to restore biodiversity take place in a broad range of contexts and involve minimal or major effort. In a recent overview of international case studies in wetland restoration, Zedler (1999) described the diversity of contexts and efforts as comprising an ecological...
restoration spectrum (Fig. 1). The effectiveness of restoration efforts was hypothesized to relate to two major variables, the degree of degradation of the site and the degree of effort expended in restoration. Where biodiversity is relatively intact, less restoration effort is necessary, and outcomes may more likely match expectations. Where massive damages have occurred, however, the challenge is daunting, and many actions are required, including restoration of the hydrology and soils and reintroduction of the microbes, vegetation, and fauna. Regrettably, comprehensive reviews of case studies have not been done; hence, we focus on various constraints to restoring biodiversity that we identified from case studies.

An emphasis on constraints is appropriate because biodiversity restoration involves the removal of conditions that limit species establishment and persistence. We describe multispecies approaches, single-species approaches, and concerns about intraspecific genetic diversity in biodiversity restoration.

II. MULTISPECIES APPROACHES

A. Restored Habitats That Became Species Rich

Establishing a community that has a known composition of native species from local reference sites constitutes biodiversity restoration. Four “success stories,” one for plants and three for animals, show that species richness can be restored, although the factors responsible are not entirely clear. All four cases occur in habitats that were severely modified, but only the first two had deliberate introduction of the propagules of the species assessed. For the others, highly dispersible species were evaluated. A feature that is common to all four case studies is that propagules were readily available from plantings or dispersal from nearby natural habitats, allowing colonization and development of species-rich ecosystems.
1. Pine forests: An area of pygmy pine forest in New Jersey (an Air National Guard Weapons Range) was severely damaged by military exercises, which eliminated the woody vegetation. Planting of different species mixtures and the use of soil fertility amendments (fertilizer, sewage sludge, or shredded bark mulch) produced diversities comparable to those of reference sites after two growing seasons. Although species richness was similar, a shrub component was lacking. The restoration sites were near or adjacent to nondisturbed areas (Fimbel and Kuser, 1993).

2. Dipterans: Streever et al. (1996) found comparable numbers of dipteran species in created and natural wetlands, and the 20 most common species had similar densities. The wetlands were constructed following phosphate mining by contouring depressions, importing stockpiled sediments, and mulching with substrate from nearby wetlands. The mulch would likely have contained insect propagules. Both created and natural wetlands were within a 33-km radius of central Florida, where agricultural land uses dominated the landscape.

3. Epibenthic invertebrates: Salt marsh invertebrates readily colonized restoration sites in San Diego Bay (Scatolini and Zedler, 1996) and the North Carolina coast (Sacca et al., 1994) but in both cases, densities were lower than in reference sites. Lower soil organic matter may have limited densities without precluding the presence of many species. An aggressive exotic mussel (Musselica senhuassui) was more abundant in the constructed than natural channels of San Diego Bay. The native and exotic species were all present in the adjacent natural wetlands. Tidal channels connected the sites and allowed larval dispersal, as no transplants were done (Scatolini and Zedler, 1996).

4. Fishes: Fish species composition in the San Diego Bay wetlands (discussed above) was similar to reference sites. Fish densities were higher in constructed channels, which were deeper and broader than reference sites; however, the occurrence of an exotic fish and an exotic shrimp caused concern. As with the invertebrates, fish dispersal occurred via tidal channels that connected the natural and constructed habitats.

Without many more examples of biodiversity equivalence between restored and natural ecosystems, it is difficult to generalize the features that are essential. Questions remain about the natural diversity levels of the communities under concern (are they naturally species-poor or dominated by habitat generalists?) and the taxa assessed (are they the best dispersers and do they have the broadest ranges of tolerances?).

B. Restoration Cases That Identified Abiotic and Biotic Constraints on Biodiversity Restoration

Most assessments of biodiversity restoration identify one or more shortcomings of ecosystem structure or function. We list potential constraints from case studies that document outcomes with less than natural biodiversity.

1. Excess nutrients: In a series of experiments to establish species-rich grasslands in a reclaimed opencast coal site in Northumberland, UK, Chapman and Younger (1995) used a commercial seed mix but did not compare outcomes with natural grasslands. They found that fertilizers decreased diversity while other treatments (e.g., grazing) increased diversity (Chapman and Younger, 1995). Several strategies were examined to control the exotic grasses Agropyron cristatum and Bromus inermis in an old field in Saskatchewan, Canada. Just introducing seeds of native species did not restore grasslands. Low nutrient levels, especially for nitrogen, favored native species restoration. In addition, spraying herbicide on nonindigenous grasses prior to planting increased the number of natives that established (Wilson and Gerry, 1995).

2. Insufficient nutrients: Soils that are very low in nutrients can support too few species to satisfy biodiversity restoration goals. In the Appalachians, open-pit coal mine reclamation is constrained by low nitrogen and by the precipitation of phosphorus as these iron-rich mineral soils weather, oxidize, and form complexes that remove phosphorus from solution (Daniels and Zipper, 1995).

3. Lack of seed banks or propagules and limitations to dispersal: Prairie pothole wetlands in North America that were restored by eliminating drainage structures (tiles, ditches) had fewer plant species than natural wetlands. Three years after restoration, restored wetlands had a mean of 27 species and natural wetlands had a mean of 46 species. This is a consequence of poor representation of wetland species in the seed bank and low dispersal rates (Galatowitsch and van der Valk, 1996).

4. Improper timing of site preparation: Efforts to increase the botanical diversity following bauxite mining in Eucalyptus marginata forests in southern Australia often produce lower plant diversity than natural refer-
...ence sites, even though overburden and topsoil are returned as part of the restoration process. Rainfall is highly seasonal and the timing of rippling, cultivation, and seed sowing were suspected as important to the outcomes. Early site preparation should maximize the number of species that establish.

5. Improper timing of propagule introduction: Some species may be difficult to reestablish because recruitment may occur only at particular times of the year or following unusual events, such as forest fires, unusually rainy periods for desert perennials, or hurricanes or other disturbances that uproot trees and expose the soil (Primack, 1996).


7. Excessive fragmentation and insufficient connectivity: Results of experiments in which up to 34 species were added to 2 × 2 m plots of grassland communities show that dispersal and recruitment limitations have a strong influence in species abundance. Thus, processes that disrupt natural dispersal dynamics, such as habitat fragmentation, should lead to changes in species abundances. Processes that overcome natural dispersal barriers can allow novel species to invade habitats and consequently cause marked changes in community composition, diversity, and functioning.

8. Unfavorable pH: Exposed fen soils in The Netherlands became acidic upon exposure to oxygen, and calcophilic species dropped out of the community. Attempts to raise pH by liming failed. Because acidification was not reversed, native species did not reestablish, even after 5 years (van Duren et al., 1998). Also in The Netherlands, acidic heath, degraded by atmospheric nutrient deposition, was depleted in species and several methods were employed to restore the native plant community. Both liming (to increase pH) and cutting the sod to the mineral soil layer (to reduce nutrient concentrations) were employed. The highest increase of species occurred with a combination of both treatments (De Graaf et al., 1998).

9. Lack of nurse plants: Existing vegetation provides suitable conditions for some rare plant species. In limestone grasslands of northern Switzerland, drought and substrate-heaving limit plant establishment. Arabis hirsuta and Primula veris were unable to establish in gaps; they were found to depend on the shelter of neighboring plants for seedling recruitment (Ryser, 1993). Under experimental restoration efforts in an arid region, cluster plantings provided better conditions for establishment than more dispersed ones because they facilitated organic matter accumulation, soil water accumulation, and higher rates of formation of mycorrhizal associations (MacMahon, 1998).

10. Lack of mycorrhizae: Experimental evidence has recently been published showing that a more diverse mycorrhizal community will sustain more diverse plant communities. This work concerned grassland species in Europe and old-field assemblages in Canada (van der Heijden et al., 1998).

11. Insufficient safe sites: Many rare species require specific places/conditions for seeds to germinate and establish (Primack, 1996). In an alpine community, above timberline, large-scale removal of vegetation and mechanical disturbance to the substrate exacerbate an already severe environment. In revegetating graded areas of a downhill ski run, increasing biodiversity was dependent on two factors, immigration from nearby populations and the provision of safe sites for seedling establishment. The extant vegetation was an important source of propagules and seedling establishment was aided by use of fibrous mats (Urbanska, 1995).

12. Lack of fire: Fire is commonly reintroduced to combat exotic plants and slow the invasion of woody species (Wheeler et al., 1995). Where habitats are small or where controlled burns are a hazard to people or structures, the elimination of fire is a constraint on biodiversity restoration (Packard and Mutel, 1997).

13. Inhibition by nonindigenous species: Introduced weeds strongly impact native plant populations, both in their occurrence (Welch et al., 1996) and in reestablishment. Restoration sites are particularly susceptible to dominance by nonindigenous plants because of disturbances before and during restoration and because of the small size and isolation of many restoration projects. For example, many prairie and savanna restorations have a high proportion of edge, allowing ready access to propagules of nonindigenous species, just as is true in natural ecosystems. In wetland plant microcosms with varying levels of water, nutrients, and leaf litter cover, Bidens cernua, a native annual, dominated all 24 treatments in the first year. After 3 years, most of the microcosms were becoming dominated by the exotic purple loosestrife (Lythrum salicaria), with only a few of the original species remaining. Flooding regimes and soil fertility were the main factors influencing establishment and growth of Lythrum, with high water and low fertility limiting its dominance (Weiher et al.,...
Habitat loss and competition from nonindigenous species are thought to be the most important causes of the imperilment of rare and endangered species in the United States (Wilcove et al., 1998). As with multispecies efforts, attempts to reintroduce single species to a habitat it formerly occupied have suggested both abiotic (e.g., substrate type, fire, habitat size, and contiguity) and biotic (competitors, predators, disease organisms) factors that reduce population viability and limit the restoration of biodiversity. Hall (1987, in Primack, 1996) reviewed 15 plant reintroduction projects and identified five key elements of a successful project: (a) the appropriateness of planting techniques and effective execution, i.e., matching microsite conditions for each species; (b) effective site selection (safe sites); (c) complete documentation of the reintroduction project; (d) maintenance of good growing conditions; elimination of competitors, etc.; and (e) long-term monitoring, with success determined as expansion of the newly established population. Griffiths et al. (1989) reviewed animal reintroduction attempts, focusing on efforts in the United States following passage of the Endangered Species Act. Only 7% of the 93 species of birds and mammals treated in their review were rare species; most of the rest were game species not protected under the Act. These latter species were more easily reintroduced than endangered or threatened species. Features identified as facilitating reintroduction of animal species were (a) selection of quality habitat; and (b) consideration of life history features (see item 9), introduction where competitors are lacking, use of wild-caught animals, and transplantation of large numbers of individuals.

We offer examples of these and other traits that constrain single-species efforts to restore biodiversity.

1. Inappropriate substrate: Prior to constructing habitat to attract nesting by Caspian terns (Sterna caspia) adjacent to Lake Ontario, substrate preferences were determined in a field experiment. The terns preferred sand substrates over other experimental substrates or the preexisting compacted ground. Experimental results were used to design a larger nesting site using the preferred substrate, which was successfully colonized by the birds (Quinn and Sirdevan, 1998).

2. Elimination of natural fire regimes: Changes in the frequency, intensity, and timing of fire have reduced rare species populations, and reintroduction of fire is beneficial to ecosystems that burned naturally. For example, fire improved the survival of the endangered annual plant Amsinckia grandiflora, which was introduced to habitat dominated by exotic annual grasses (Pavlik, 1996).

3. Landscape context: A key factor in restoring single species is proximity to large natural habitats that support the same species and its matrix community. For example, in New Jersey, a major utility company breached dikes that restricted tidal flows to former salt marshes. Rapid restoration of Spartina alterniflora occurred, in part attributed to the landscape context—some 32,000 ha of fully tidal wetlands surrounded the ~1000 ha of diked wetlands that were breached to allow tidal action and natural colonization by the target clonal grass (John Teal, Woods Hole Oceanographic Institute, personal communication).

4. Behavioral rigidity: Animal species that cannot tolerate human influences are unlikely to persist upon reintroduction to human-dominated landscapes, while those that can take advantage of modifications may flourish. For example, laboratory-reared and released Mauritius kestrel (Falco punctatus) changed their behavior to make use of agricultural areas and to prey on exotic species. Their numbers increased from 6 individuals in the 1970s to at least 30 mating pairs in 1992 (Crade and Jones, 1993).

5. Presence of aggressive predators: The unsuccessful reintroduction of a marsupial, the quokka (Setonix brachyurus), to a field station near Perth, Australia, was attributed to competition for food by rabbits and predation by foxes, both nonindigenous species. In general, invasive predators, particularly foxes and cats, challenge reintroduction efforts of many macropod spe-
cies in Australia (Short et al. 1992). The Parrula snail, a native to Moorea, French Polynesia, was driven to extinction by the introduction of the predatory snail Achatina fulica, another introduced snail raised for escargot. Parrula snails were reintroduced in enclosures with a physical and chemical barrier to keep out the predatory snail, but in the wild, the reintroduced species is still vulnerable to predatory snails (Mace et al., 1998).

6. Abundant competitors: Weedy vegetation (nonindigenous annual grasses) reduced germination, growth, and survival of an endangered annual plant, Amsinckia grandiflora; fire and herbicides were effective in facilitating its reintroduction. Griffith et al. (1989) did not highlight competitors as a factor in constraining animal reintroduction efforts.

7. Method of reintroduction: Hall (1987, in Pri-mack, 1986) listed the use of appropriate planting techniques, effective execution, and appropriate maintenance of good growing conditions, including watering, as important to reintroduction efforts. Similar concerns were expressed by Griffith et al. (1989) for animal reintroductions (see item 9).

8. Insect outbreaks: Plant species that are naturally rare may not be restorable when planted in dense populations. Such was the case for a tropical woody plant that endured high mortality from insects when planted as a monotype but grew well as isolated individuals (P. Kagayama, University of Sao Paulo, personal communication).

9. Animal species life history: In their survey of reintroduction efforts for birds and mammals, Griffith et al. (1989) found that carnivores and omnivores were less readily translocated than herbivores, and late breeders with small clutches were slightly less likely to establish in new surroundings.

10. Scarcity of pollinators: An endangered annual plant was restored to a San Diego Bay salt marsh by sowing seeds on a small island where most of the habitat was not suitable for some of its pollinators, ground-nesting bees, that require upland, not intertidal, habitat. Insufficient pollination was shown to limit seed production in this species even when introduced to marshes next to disturbed upland.

11. Improper canopy structure: Habitats created for the endangered light-footed clapper rail (Rallus longir-ovis levipes) supported the right plant species, the cordgrass Spartina foliosa, but not the tall canopy that this rail species requires for nesting and camouflage. The ultimate cause of the short vegetation was the coarse soil, a sandy dredge spoil, which was much coarser than salt marsh soils and which failed to supply or retain nitrogen.

Restoration efforts begin with eliminating known constraints, and work will be most efficient if the reasons for a species’ decline or extirpation are known. If limiting factors are unknown, then experimental approaches with a suite of management tools are advisable. From detailed field experiments, Pavlik (1996) determined that at least two restoration methods were needed to reestablish Amsinckia grandiflora to grassland in northern California, namely, fire and control of nonindigenous annual grasses. But even with his considerable understanding of the constraints, one reintroduced population spread without much treatment, while another required both fire and removal of nonindigenous grasses. His work suggests that outcomes of reintroduction may be quite site-specific.

Where experimental approaches are not feasible, knowledge of the species natural history can be employed in designing the reintroduction program. In re-storing an endangered annual plant (Cordylanthus maritimusspp. maritimus) to San Diego Bay, seeds were sown in microhabitats thought to be optimal for this hemiparasite: small openings in the canopy of a favored host plant. Additional limiting factors were later identified (canopy closure and loss of regeneration sites, drought, and hypersalinity), but none prevented the re-introduced population from persisting in the short term (8 years to date). A major decline and loss of seed banks occurred in response to a dry year, so a series of years of low rainfall could reduce the population to nonviable proportions. As in most restoration projects, it is unlikely that monitoring will continue long enough to discover the link between environmental constraints and persistence/extirpation.

B. Genetic Diversity Issues

Genetic diversity is of great concern in single-species efforts to restore biodiversity, because

1. Propagule sources may be far from the reintroduction site. Managers (e.g., M. Kenney and B. Collins, US FWS, personal communication) have recognized several alternatives: (a) importing material from distant donor populations; (b) mixing gene pools from several potential donor populations; and (c) selecting the nearest donor population. The issues have been discussed, but no standard guidelines are available (Bowles and Whelan, 1994). Tijuana Estuary was the nearest donor site for an attempt to restore an endangered salt marsh...
210       RESTORATION OF BIODIVERSITY, OVERVIEW

plant (Cordylanthus maritimus ssp. maritimus) to San Diego Bay. Seeds from that one site were sown repeatedly at the recipient site in order to ensure that the annual hemiparasite would become established. Although thousands of seedlings were present and reproducing after several years of seeding, the resulting population had low genetic diversity and was considered at risk from genetic drift.

2. Reintroduction efforts involve small populations, and genetics are very important in small populations. For example, the Florida panther is experiencing inbreeding depression; males have low sperm quality and quantity and are cryptorchid (having only one descended testicle) (Lacy, 1994). While inbreeding depression can lead to the decline of population viability, there are also risks with outbreeding. Outbreeding depression can result in an intermediate form of hybrid that is adapted to neither parent’s habitats. Also outbreeding can break up beneficial coadapted gene complexes. In restoration, if genes can be selected for an additive effect, then heterotic (hybridizing) may outweigh outbreeding depression. However, if populations are locally adapted through the evolution of coadapted gene complexes, then outbreeding and the mixing of alleles can possibly decrease fitness.

3. Captive breeding is employed to increase animal population sizes before reintroduction to the wild. In captivity, species can go through genetic changes that compromise their ability to survive in the wild. For example, captive-bred tamarins were released in Brazil, but these individuals moved slowly, were confused by scarce food, and were unwary of predators; preconditioning and training and some harsh experience helped other released tamarins integrate with natural populations (Lacy, 1994). Lacy (1994) recommends a population size of at least 50 for short-term captive breeding to help avoid immediate deleterious effects of inbreeding. For long-term efforts, a population size of at least 500 is desirable to maintain genetic variability. The desired number of individuals would depend on demographics, sex ratio, nonrandom production of offspring, and fluctuations of population across generations. If possible, individuals from the wild should be introduced to the captive population at different breeding stages to maintain diversity.

4. Restoration of matrix species that reproduce vegetatively (e.g., Spartina species, Zostera species) are easily propagated, but when restorationists draw on small areas for propagules, the reestablished population may be primarily one genotype. Williams and Davis (1996) documented reduced genetic variability in Zostera marina in Mission Bay relative to more natural stands in nearby San Diego Bay. Whether or not genetic diversity is important to the long-term functioning of restored habitats is another issue. Seliskar (1995) showed that three different genotypes of Spartina alterniflora planted to a salt marsh restoration site in Delaware differed in many structural and functional attributes, including support of animal populations.

IV. NOVEL OPPORTUNITIES FOR RESTORATION OF BIODIVERSITY

Numerous opportunities exist for partial restoration of regional diversity. Human-influenced landscapes include many patches of habitat that currently support nonindigenous species, either intentionally or unintentionally. Highway edges and privately owned properties could support some or many native plants. Even if grown in horticultural arrangements, gene pools would be sustained by planting propagules native to the region. We found papers describing two suitable places to introduce native plants that once occurred within the region, if not on the exact site:

A. Tree Plantations

Plantations can be managed to provide a diversity of understory plants. In New Zealand, more indigenous plants and animals occur in forest plantations that are older and closer to natural habitats (Norton, 1998). In European boreal forests, biodiversity is being restored by managing fire regimes, e.g., privately owned plantations in Sweden (Angelstam, 1998). In northern Australia, plantations of native and exotic tree species span a gradient from higher to lower species richness of understory natives. For similar sites, the natives Flindersia bryleyana and Araucaria cunninghamii support more diverse understory vegetation than the exotic Pinnus caribou (Kreman et al., 1997).

Planting a diversity of trees can also be beneficial to forest production by reducing chances of fungal (Armillaria spp.) infection. An experiment in Minnesota varied the mixtures of tree species (six conifers and four hardwoods) and planting densities. Species were differentially susceptible to infection, and mortality was highest when the conifers were most numerous in the plots. Seedling mortality was correlated with the species planted, the density of planting, and the proportion of conifers (Gerlach et al., 1998).
B. Utility Line Rights-of-Way

An opportunity exists for restoration of vegetation native to areas where power lines have been installed. Since tall trees cannot be allowed beneath high-voltage lines, tree growth is controlled by cutting and use of herbicide. These methods of tree control result in communities low in species diversity and wildlife. Brown (1995) suggested that the planting of competitive cover crops, including grasses, legumes and forbs, could inhibit the establishment of trees. *Dactylis glomerata* (a nonindigenous grass) reduced survival of two tree species (*Fraxinus pennsylvanica* and *Acer saccharum*) planted into experimental plots as desired, but biomass and diversity of forbs was also reduced. This alternative approach can produce communities that require less physical and chemical maintenance, as well as greater species diversity and habitat for wildlife.

V. CONCLUSION

Community-level approaches to biodiversity restoration often produce systems that are less diverse than reference systems. Prevention of natural habitat loss is thus the preferred conservation approach over allowing damages and attempting to reverse losses. More understanding of factors responsible for a species being common or rare in the region would assist restoration efforts. The relationship between a species commonness, adaptability, the breadth of its tolerance range, and its ability to be reintroduced needs to be known before the outcomes of biodiversity restoration efforts can be more predictable.

Many rare species have been the subject of reintroduction, and studies of such efforts have yielded more information about individual species requirements, especially populations that have variable reestablishment in space or over time. Comparisons of conditions associated with greater and lesser recruitment have identified many abiotic and biotic constraints on species reintroductions. Restoration efforts need to begin with an understanding of what factors were responsible for greater diversity prior to species declines.

Little work has been done on the restoration of animal assemblages, particularly terrestrial species, although wildlife ecologists have considerable experience introducing selected game species and endangered animals. In tidal wetlands, fishes readily invade sites with suitable hydrology; many marine invertebrates are introduced with greater and lesser recruitment have identified many abiotic and biotic constraints on species reintroductions. Restoration efforts need to begin with an understanding of what factors were responsible for greater diversity prior to species declines.

V. CONCLUSION

Community-level approaches to biodiversity restoration often produce systems that are less diverse than reference systems. Prevention of natural habitat loss is thus the preferred conservation approach over allowing damages and attempting to reverse losses. More understanding of factors responsible for a species being common or rare in the region would assist restoration efforts. The relationship between a species commonness, adaptability, the breadth of its tolerance range, and its ability to be reintroduced needs to be known before the outcomes of biodiversity restoration efforts can be more predictable.

Many rare species have been the subject of reintroduction, and studies of such efforts have yielded more information about individual species requirements, especially populations that have variable reestablishment in space or over time. Comparisons of conditions associated with greater and lesser recruitment have identified many abiotic and biotic constraints on species reintroductions. Restoration efforts need to begin with an understanding of what factors were responsible for greater diversity prior to species declines.

Little work has been done on the restoration of animal assemblages, particularly terrestrial species, although wildlife ecologists have considerable experience introducing selected game species and endangered animals. In tidal wetlands, fishes readily invade sites with suitable hydrology; many marine invertebrates are ready invaders, but they may be accompanied by nonindigenous species, and their restored populations may have lower numbers of individuals. Genetic diversity is a concern for reintroduced populations because small populations may not have sufficiently diverse gene pools and because large populations of matrix species are sometimes too homogeneous genetically due to vegetative propagation from a few clones.

Tree plantations and utility rights-of-way offer novel opportunities for biodiversity restoration. As landscapes become more fragmented and open habitat becomes more limited, sites managed for vegetation will become more important in conservation, and using such sites to establish diverse, native plant populations will aid overall biodiversity efforts.

Acknowledgments

We thank Kimberly Hamblin Hart for assistance in manuscript preparation. This review was supported in part by a grant from the National Science Foundation, Award DEB 96-19875, to J.Z., J.C., and G.S.

See Also the Following Articles

CONSERVATION EFFORTS, CONTEMPORARY • DIVERSITY, COMMUNITY/REGIONAL LEVEL • GENETIC DIVERSITY • RESTORATION, CHARACTERISTICS AND REQUIREMENTS

Bibliography


RIVER ECOSYSTEMS

K. E. Limburg,* D. P. Swaney,† D. L. Strayer‡
*State University of New York, †Cornell University, and ‡Institute of Ecosystem Studies

I. Introduction
II. Anatomy of Rivers
III. Energy and Matter Transformation (Biogeochemistry)
IV. Different Conceptual Models of Riverine Ecosystems
V. Riverine Biota
VI. Threats to Biodiversity
VII. Assessment and Management

GLOSSARY

allochthonous Referring to the production of organic matter outside of the ecosystem; in streams and rivers, this would be inputs from the upstream watershed.

autochthonous Referring to the production of organic matter within the ecosystem, for example, primary production by aquatic plants.

cultural eutrophication The acceleration of nutrient overenrichment caused by humans; can be caused by direct, point-source pollution (sewers) or by diffuse, non-point-source pollution (such as fertilizer runoff from farm fields).

hyporheic zone The area of saturated soils beneath a stream or river channel; interacts with the stream through the processes of hydraulic upwelling and downwelling.

riparian zone The vegetated areas on either side of a stream or river, with saturated soils usually underlaying aerated soils.

river continuum concept The first of a series of conceptual, unifying models to explain and predict the structure and function of river ecosystems.

stream order A numbering system to denote stream size within a network of streams. There are several stream order systems. Stream order is also dependent on the scale of observation.

watershed (catchment) A unit of landscape in which precipitation drains to a common stream or lake; alternatively, a watershed is the divide between catchments.

RIVER ECOSYSTEMS ARE NETWORKS of streams that drain the landscape. They are thus composed of a hierarchical series of fluvial channels, beginning with small headwater streams and enlarging, ultimately, to estuaries that meet the sea. This article introduces concepts of stream flow within the catchment area, or watershed, drained by the stream network. We describe several conceptual models, which attempt to provide unifying concepts about the connections of rivers with the landscape in terms of ecosystem properties such as processing of energy and matter, habitat, biodiversity, and resilience in the face of disturbance. The main groups of riverine biota include organisms that live in the water column (such as the free-floating plankton, or fish), organisms (the benthos) that live in the sediments,
organisms that use the shoreline area (the littoral zone), and organisms that dwell within adjacent areas (hyporheic and riparian zones) but which are connected to the stream in some fashion. Many human activities pose threats to the well-being of river ecosystems, including the shifting of land from forests, grasslands, and wetlands into urban or agricultural uses, construction of dams, pollutant loadings, alteration of natural drainage characteristics, introduced species, overharvesting, and climate change.

I. INTRODUCTION

Rivers are the sinuous conduits of overland flow that drain the world’s continents. They displace very little of the world’s land area and their standing stock of water is small relative to other hydrologic storages. Of the estimated 1,385,984,610 km$^3$ of global water, the average volume of rivers composes about two ten-thousandths of one percent, or 0.006% of the freshwater supply (Maidment, 1993). Nevertheless, rivers are of disproportionate importance in terms of geomorphology, ecology, and biodiversity, for they serve as transport vectors for sediments, organic matter, nutrients, and species. In contrast to the image of a river as a pipeline from source to mouth, river networks form bifurcating, fractal patterns through the landscape which are among the most complex in nature. Along any reach of the river channel, a fully three-dimensional world lies between the riverbanks; therein are found habitats in which organisms form ecological communities and process energy and matter in various ways, depending upon where the reach is located. Because of the intimate connections of rivers with the landscape, the burgeoning pressures of human-accelerated environmental change are becoming particularly acute in these systems.

II. ANATOMY OF RIVERS

To a hydrologist, the basic unit of the landscape is the catchment or watershed, in which water organizes itself into flowpaths. Watersheds are formed by the various forces of continental uplift and wearing down (weathering) of the bedrock by water and wind. Geomorphologically, vegetation within watersheds plays a dual role: it aids in weathering and conversion of rock to soils, and it retards surface runoff, diverting some of it to gaseous water vapor by evapotranspiration. Watershed boundaries, or divides, are those more resistant zones where water flows either one way or the other into alternative drainages. The Continental Divide in the United States separates the western drainages (which flow into the Pacific Ocean) from those (principally the Mississippi) which drain east or south. Divides at the continental scale can often limit the flow of species, but the process of species flow is affected at all scales of watersheds.

Watersheds and their accompanying flowing waters are organized hierarchically; that is, smaller watersheds usually form into larger watersheds (unless a physical boundary prevents this), which in turn may form parts of regional drainages. Hydrologists in the 1940s and 1950s recognized that streams and rivers form drainage networks and defined topological metrics to describe these. The most common metric is stream order. Though there are various definitions of order, Strahler (1952) considers that a first-order stream has no tributaries, a second-order stream is formed at the meeting of two first-order tributaries, a third-order stream is formed at the meeting of two second-order tributaries, etc. Shreve’s link magnitude system (Shreve, 1960) defines the magnitude of a stream, downstream of the meeting of two streams, as the sum of the magnitudes of those streams. It should be noted that order is inevitably related to the scale of the map on which the stream and its tributaries is plotted (typically, in the United States, the smallest streams delineated at 1:24000 scale USGS topographic maps are considered “first-order” streams, but higher resolution maps produce higher order streams). This is a characteristic of the fractal nature of drainage networks (Rodríguez-Iturbe and Rinaldo, 1997). Another metric is the bifurcation ratio, or the ratio of first-order to second-order, second-order to third-order, etc. streams within a drainage network. The drainage density, or ratio of cumulative length of streams (L) to watershed area (A), is a third measure of network organization. High bifurcation ratios might indicate the presence of many small streams and, in conjunction with a high drainage density, also describe a highly dissected landscape. More complex models, which take into account differences in elevation, geology, and soils, provide even more detail for water flow and habitat creation.

Early models of watershed hydrology assumed that precipitation events saturated the soils, producing overland flow, beginning from the base of a drainage and moving uphill, much like filling a bathtub. However, in the 1970s the variable source area model, in which flow originates from source areas that expand and contract in different parts of the watershed in response to the dynamics of precipitation and drainage characteristics, gained acceptance. Much overland flow during a
storm is actually water that had been stored in the soil, which is displaced by the new water coming in. From an ecological perspective, the multitude of geomorphologic forms and flowpaths gives rise to a similar multitude of habitats and ways for matter to be transformed, at scales ranging from the microscopic to the entire basin. A typical cross-sectional view of a stream is given in Fig. 1. In this figure, it is clear that the stream is in contact with several areas: a zone of saturated sediments below, called the hyporheic zone; an area with emergent mud or gravel bars overlying saturated soils, called the parafuvial; and beyond that a zone of bankside vegetation, also in contact with saturated soils, the riparian zone. In upland streams, where the terrain is steep and rugged, these zones may be compressed around the stream channel; but farther downstream the zones may spread out across the floodplain that was carved out over geological time by the flowing water.

Hynes (1975) was among the first stream ecologists to point out the intimacy of connection between rivers and their watersheds. As systems, rivers and streams are inherently open and serve as connectors, transporters, and processors of material. Among the biological connections of streams and watersheds are the inputs of organic matter from vegetation: leaves, insect frass, and partial or entire trees (which form debris dams). Water does much of the physical work in carving channels and carrying sediment; but organisms mediate this by regulating the rates of mineral use and transport, fixing organic matter and passing it along the food chain, and even altering landforms so that hydrology is affected (e.g., the work of beavers).

**III. ENERGY AND MATTER TRANSFORMATION (BIOGEOCHEMISTRY)**

From a biogeochemical standpoint, rivers differ from other ecosystems in at least three important ways: first, they are flowing systems that provide a source of kinetic energy, as well as externally produced nutrients, to resident organisms. Second, they are spatially complex and extensive systems, with a relatively large interfacial connection to contiguous terrestrial ecosystems. Such a large "buffer" zone has significant implications for processing nutrients and other fluxes into the river. Finally, as discussed above, they are transitional systems, changing in character from small, shallow, flashy first-order streams that are closely connected to the landscape to large, relatively deep and steady rivers that may be tidally influenced as they become estuarine in character near their mouths. The interplay between light, nutrients, and turbulent energy and circulation structures the biotic community, and this structure changes along the river from sources to mouth.

Riverine ecosystems are fueled by inputs of nutrients, including carbon, nitrogen, phosphorus, and silicon. Nutrients that are available to primary producers must generally exist in inorganic (mineral) forms. In pristine watersheds, rock weathering and decomposition of organic material in forests, grasslands, and soils provide the sources of these nutrients. In most watersheds today, atmospheric deposition of nutrients from distant industrial and automotive sources occurs via rainfall and dust. Within the watershed, fertilizer from agricul-
and urban applications enters rivers in surface runoff and groundwater; discharges of sewage treatment plants and industry provide other sources. Frequently, nutrients also enter rivers in various organic forms (that is, as byproducts of other organisms, such as soil particles, plant and animal remains, and feces). Sources of organic nutrients associated with human activity include untreated sewage and manure runoff from farms. Organic nutrients entering a river segment from the watershed upstream augment the local primary production of biomass within it to fuel heterotrophic organisms. Such external sources are termed allochthonous as opposed to the autochthonous production occurring within the river. Regardless of whether nutrients enter rivers in organic or inorganic forms, excessive loading can result in accelerated riverine metabolism in which organic matter is consumed in respiration, with a consequent increased biological oxygen demand (BOD). When BOD exceeds the local oxygen supply in a reach of the river, available oxygen levels may decline precipitously (resulting in the "oxygen sag curve" downstream of nutrient sources familiar to environmental engineers), with catastrophic results to riverine organisms. Such anoxic zones typically resemble the popular image of "polluted waters": smelly masses of decaying algae, fish kills, and greatly reduced biodiversity.

Because organisms require nutrients in particular ratios (in oceanography, these are termed Redfield ratios after their discoverer Alfred C. Redfield (1890–1983)) to maintain their metabolism, deviations of environmental concentrations from these ratios can result in species shifts. For example, the shift of the available nitrogen:phosphorus (N:P) ratio below the Redfield ratio of 16:1 (mole of N per mole of P) in freshwater is often associated with the onset of a bloom of cyanobacteria (blue-green algae), which are capable of fixing nitrogen from the atmosphere. Depletion of silicate below the Si:N molar ratio of 1:1 has been observed to shift the dominant species of algae from diatoms, which require silicate to synthesize their siliceous "skeletons," to species that do not utilize silicate. Such species shifts are sometimes associated with nuisance algal blooms. It must be emphasized that the dominant physical controls of riverine ecosystems change dramatically as stream order increases. The ecosystem of the shallow, well-mixed, intermittent stream is controlled almost completely by weather and the character of the contiguous terrestrial ecosystem—geology and soils, nutrient processing in the riparian zone, light levels established by the degree of canopy closure, etc. Near its mouth, the river is deeper and broader and influenced by the sea. Relatively dense seawater meets the lighter freshwater, sometimes resulting in stratification, which, together with the effects of the tide, yields complicated patterns of circulation. The cumulative sediment load of the watershed results in decreased water clarity and increased sediment deposition (the deltas of many large rivers). The combined workings of internal circulation and turbidity structure the light and nutrient fields of the water column as well as the distribution of biological communities in ways that are not completely understood at present.

IV. DIFFERENT CONCEPTUAL MODELS OF RIVERINE ECOSYSTEMS

Beginning in the 1970s, synthetic conceptual models of rivers began to be developed as part of the ongoing movement to understand the world's ecosystems, under the auspices of the International Biological Programme (IBP). One of the fundamental insights that came to the fore was the necessity of integrating stream ecosystem dynamics over a range of spatial and temporal scales.

A. River Continuum Concept, or RCC

The RCC (Vannote et al., 1980) originated as a theoretical concept to organize and predict the structure and function of riverine ecosystems from source to sink. The premise is that river ecosystems are dominated by the physical factors that create the drainage network and that ecological communities tend to be selected in response to fluvial geomorphic processes, which along with hydrology force a more or less predictable template onto the biotic components of the system. According to the RCC, the kinds of producers, consumers, magnitude of organic inputs, and degree of heterotrophy or autotrophy should be predictable along the physical gradient. Ecological communities should be in dynamic equilibrium with the stream, and the openness of streams, connecting upstream and downstream processes, is made manifest in the RCC.

Figure 2 depicts the RCC. The first thing to notice is that the stream is organized as a gradient, from an idealized headwater stream to a large river (stream order is on the left of the diagram, and it suggests a logarithmic scale). The figure also attempts to unify biological communities with functional and ecosystem-level properties.

The RCC authors divide streams into headwaters (orders 1–3), medium-sized streams (orders 4–6), and
large streams (orders >6). They point out that many headwater streams are dominated, in terms of the energy coming in, by the allochthonous inputs from out-of-channel, because the tree canopy tends to close off the stream from direct sunlight. In this case, the P:R ratio (ratio of primary production to ecosystem respiration) is less than 1, and most of what enters the stream is coarse particulate organic matter (CPOM, matter >1 mm), which, in the case of tree boles, can be very coarse indeed. The aquatic food chain here is dominated by organisms that can deal with CPOM, mediated by microbial processing—the so-called guilds of shredder, grazers, collectors, and their predators.

Moving downstream, the relative contribution from allochthonous inputs declines, and autochthonous production (from submerged macrophytes, phytoplankton, and epiphytes) and processed inputs from upstream (now fine particulate organic matter, or FPOM, 0.005- to 1-mm grain size) become more important. Someplace in this midzone, P:R reaches a maximum and is greater than 1, indicating this is the most productive (in situ) part of the continuum.

The large river end of the gradient is dominated by FPOM inputs and primary production, but the latter can be limited due to turbidity (due in part to inorganic sediment, not discussed explicitly in the RCC). Turbidity attenuates the light in the water column, so phytoplankton production may be limited and P:R is predicted to decrease. Because large rivers increase in depth, they become more lake-like, with a deep water column and a muddy benthic zone, so that many of the stream insects drop out of the system.

The RCC also makes predictions about where in the system maximum diversity of organisms should occur (in the medium-sized reaches, or areas with high physical variation, because these are arguably the most diverse of habitats) and that the biotic components of the system will serve to enhance the stability of the ecosystem (referring to a long-ongoing debate as to the importance of biodiversity to ecosystem stability).

One final prediction, or observation, in the RCC is that rivers have both spatially and temporally redundant groups of organisms that tend to perform the same kinds of ecosystem functions, i.e., breaking down the organic matter inputs and remineralizing nutrients; they also argue that one should not expect to see the development of successional stages of biotic communities in rivers, „„ because the communities in each reach have a continuous heritage rather than an isolated temporal composition within a sequence of discrete successional stages” (Vannote et al., 1980, p. 135). Later on we shall see that scientists dispute this idea.

B. Resource (Nutrient) Spiraling

“All land represents a downhill flow of nutrients from the hills to the sea. This flow has a rolling motion. Plants and animals suck nutrients out of the soil and air and pump them upward through the food chains; the gravity of death spills them back into the soil and air. Mineral nutrients, between their successive trips through this circuit, tend to be washed downhill. Without the impounding action of soils and lakes, plants and animals would have to follow their salts to the coast line.” (Leopold, 1941)

The famous conservationist Aldo Leopold makes important points in his elegant prose: first, there is a tendency for elements to run downhill from the uplifted land to the sea; and second, biota play a key role in slowing down this transfer. Third, he speaks of the
“rolling motion” of mineral transfer across the landscape and waterscape. Webster (1975) used the term spiraling to describe the combined downhill movement and cycling aspects of nutrients and organic matter, and Elwood et al. (1983) derive a specific model to describe the dynamics. The concept describes how an atom of a resource (a nutrient like N, P, or C) moves into a stream and how it is retained, processed, and released downstream. The atom may travel downstream, be picked up in a part of the system, cycled there many times, buried for some time, or released relatively quickly. The spiraling model provides an analytic framework to quantify the relationships between biogeochemical processing of materials and hydrodynamic transport, proposing the spiraling length as a measurable index of this interaction.

Spiraling of nutrients in streams is slowed by biotic retention (through sorption and ingestion) of dissolved and particulate forms. Recycling of nutrients in situ further slows downstream transport. Taking a steady-state, averaged approach for a first derivation, the total downstream flux of a nutrient \( F_T, \text{ g/s} \) is related to the standing stock of nutrient available per unit length of stream \( N, \text{ g/m} \) and the average downstream travel speed (velocity) of the nutrient \( v, \text{ m/s} \):

\[
F_T = Nv
\]

As the biota take up the nutrient (postulated to happen mostly on the stream bottom), the average downstream velocity of the nutrient slows. The channel waters are assumed to be highly advective and the channel bottom retentive.

A second aspect of spiraling is the rate at which nutrients are cycled in the stream. The use rate of a nutrient \( N \) is defined as

\[
U = kN
\]

where \( k \) (1/S) is a constant of proportionality (the fraction of the standing stock of \( N \) that is recycled per unit time, time being seconds in this case). A nutrient atom on average completes a cycle in \( T = 1/k \).

By dividing the first equation by the second one, a ratio of retention to cycling is obtained:

\[
Nv/Nk = v/k = vT = F_T/U = S
\]

The parameter \( S \) is the average downstream distance traveled by a nutrient atom during a cycle (between bottom and flowing water) and is called the spiraling length. If the value of \( k \) is large (high cycling rate within the stream section) and \( v \) is low ("resistance to stream transport"), then \( S \) will be short, and a greater retention of the nutrient is implied.

Figure 3 illustrates several predictions for how the spiraling length will be affected by different retention and biological processing rates. Ecosystem stability and response to nutrient loading (a disturbance) are also predicted as a consequence of spiraling.

The nutrient spiraling model can be elaborated for any number of compartments, and different nutrient spirals can be measured simultaneously as well. Carbon is one exception which is treated separately, because carbon moves so readily in and out of the dissolved and gaseous phases.

Another important point about spiraling is that it potentially identifies fast and slow components of the system. Slow components (with tight nutrient spirals and short \( S \)) are postulated to be relatively resistant to perturbation, implying that these help to stabilize a system in the face of disturbance.

C. Serial Discontinuity Concept

This heuristic model is really an early corollary to the RCC, developed by two river scientists (Ward and Stanford, 1983) from the western United States, where many rivers do not flow freely. Whereas the RCC is intended to describe more or less pristine systems, Ward and Stanford were concerned about rivers that contain dams and impoundments. They argued that such regulating structures "reset" the river continuum, although not always in the low-order to high-order direction. The key feature of serial discontinuity is that it can cause a stream reach to "behave" or function like reaches that the RCC would predict should occur in a different stream order. Figure 4 illustrates the general concept.

The serial discontinuity concept proposes that in regulated rivers, two measurable parameters of interest are the discontinuity distance (i.e., how far upstream or downstream a given stream parameter becomes displaced) and the parameter intensity (how much has the maximum value, duration, etc. of the parameter been changed?).

D. Flood Pulse Concept

In contrast to the RCC, which predicts a downstream decreasing influence of the riparian zone, the flood pulse concept (Junk et al., 1989) deals with rivers that interact strongly with the floodplain by rising out of the channel bed. This theory describes the effect of floods on river channel and floodplain in large, unregu-
RIVER ECOSYSTEMS

FIGURE 3 Predicted effects of interactions of downstream transport and measures of biological activity, and postulated responses of the river ecosystem to disturbance, here in the form of nutrient addition. The last column refers to streams in a study (by state) and their approximate stream order. From Minshall, 1983, reproduced in Cummins et al., 1995.

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Effect on Nutrient Cycling</th>
<th>Ecological Response to Nutrient Addition</th>
<th>Ecological Stability</th>
<th>Compaction of Nutrient Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. HIGH</td>
<td>HIGH</td>
<td>CONSERVATIVE</td>
<td>HIGH</td>
<td>MI 2,3</td>
</tr>
<tr>
<td>B. HIGH</td>
<td>LOW</td>
<td>STORMY</td>
<td>HIGH</td>
<td>ID 1,2</td>
</tr>
<tr>
<td>C. LOW</td>
<td>HIGH</td>
<td>INTERMEDIATE CONSERVATIVE</td>
<td>LOW</td>
<td>MI 5, 4</td>
</tr>
<tr>
<td>D. LOW</td>
<td>LOW</td>
<td>EXPORTING</td>
<td>LOW</td>
<td>ID 3, 4</td>
</tr>
</tbody>
</table>

FIGURE 4 Main features of the serial discontinuity concept, describing the effect of an impoundment on a river ecosystem. DD, discontinuity distance is the distance that a parameter is displaced; PI = parameter intensity, or degree to which a parameter is altered.
will evolve to take advantage of the pulse because it amplifies resource availability. The flood pulse is postulated to enhance diversity and productivity by structuring the dynamics of plants, nutrients, detritus, and sediments.

During floods, matter that was remineralized in the floodplain is quickly dissolved and made available to aquatic primary producers, and production in many of these systems (including the Amazon, Bayley, 1995) exceeds decomposition. As the floodwaters stop rising,
decomposition in deeper, slower-moving parts of the stream increases, often becoming anoxic. Terrestrial plants, in the meantime, recruit into the sites of re-treating waters and begin to build up terrestrial biomass again.

The “moving littoral zone” is used by many fish as a nursery for young and by aquatic invertebrates that can utilize shallow environments. Many fish have adapted to the pulse by spawning just before the floods, so their offspring can be on hand when the pulse triggers off the high primary production with associated high secondary production.

E. Telescoping Ecosystem Model (TEM)
A recent paper by Fisher et al. (1998) presents another way of looking at the holistic, biogeochemical functioning of river systems and their environs, and bears some discussion. In contrast to the RCC that visualizes riverine systems as continua, and in contrast with models such as nutrient spiraling, which assumes rivers are simple, uniform chemostats, these authors emphasize heterogeneities in space and time:

“The landscape is a patchwork of many patch sizes and shapes. Materials follow a large number of routes from uplands to oceans. The path is stochastic and often downhill because the predominant transporting vehicle is water. Because these hydrologic linkages are driven by weather, the path is episodic and jerky. Within patches, characteristic rates of uptake, transformation, and release occur: thus, as the vehicle (water) travels, its load (elements) is adjusted in response to patch-specific material dynamics.” (p. 20)

This perspective corresponds to the understanding of variable source flow in watershed hydrology and takes into account the patchiness and position of patches in the landscape.

As connectors of landforms with lakes and oceans, rivers with relatively fast throughput link land and water units with considerably larger residence times; thus, the connection of land–river–lake or land–river–ocean is (to some extent) slow–fast–slow in terms of processing length change and will be greatest away from the center of the “telescope.”

1. When a stream system is disturbed, the different subsystems will alter their processing lengths, but differently. The resistance of the system, or how little it changes when perturbed, is inversely proportional to processing length change and will be greatest away from the center of the “telescope.”

2. The resilience of the stream telescope (how fast
the system returns to the previous state) decreases away from the center.

These two emergent properties of the TEM are shown in Fig. 6. However, cross-links (interactions between the subsystems) are proposed to enhance resilience of the system. Thus, the TEM is a model that can predict the rates of system change, but also maintains a hierarchical structure.

**F. Geomorphic–Trophic Hypothesis and Other Geomorphological Controls on Ecology**

Recent work in relatively undisturbed ecosystems has demonstrated how topography and landscape position can influence not only the biogeochemistry of waters within a catchment but even the invasibility of aquatic habitats by organisms, which sets the stage for subsequent ecological interactions. At a Long-Term Ecological Research (LTER) site in northern Alaska, researchers have found that the fish communities of lakes, connected by streams, are determined by such factors as basin steepness (which determines waterfalls that may not be passable by certain fish species), lake depth and size, etc. (Hershey et al., 1999). This results in a limited number of fish communities, and the species present appear to regulate the presence and abundance of other organisms (invertebrates and phytoplankton). This would be much more difficult, if not impossible, to observe in temperate systems subject to human alteration.

**V. RIVERINE BIOTA**

**A. Microbiota**

Microbiota are widely recognized as being the most abundant and diverse organisms on Earth. One cubic centimeter of water can hold over $10^9$ bacteria. Microbes are responsible for most of the biogeochemical processing of nutrients and organic matter in rivers. Some important functional groups of bacteria are nitrifiers (which convert organic N into nitrate and nitrite), the denitrifiers (which convert nitrate and nitrite back into atmospheric N), sulfate reducers, and methanogens (which convert carbohydrates into methane). Microbes can be free-swimming or attached to particles; they are found in the hyporheic and riparian zones as well as in the water column itself. Microbial diversity, although recognized as vast, is difficult to assess, but new techniques involving DNA are improving the process of identification and classification.

**B. Plants**

Riverine plants fall into two general groups, the littoral, mostly rooted vegetation (macrophytes, or, literally, large plants), and the microscopic, free-living phytoplankton. Macrophytes can be further divided into those
that send their leaves up above the water surface (emergent macrophytes) and those that photosynthesize within the water column (submersed macrophytes). A number of macrophytes are also free-floating, such as the water hyacinth (*Eichhornia crassipes*), which has enlarged, buoyant leaf stems. Some algal species are also macrophytic.

Aquatic vascular plants have several special adaptations for dealing with the conditions imposed by the aquatic environment. Many plants have roots that extend into the water-saturated sediments, and as a result, are faced with an anoxic environment. To avoid root anoxia, many species have evolved air spaces (aerenchyma) in their roots and shoots that conduct oxygen down from the leaves. An example is the American water celery (*Vallisneria americana*), which translocates so much oxygen to its roots that it actually oxidizes the surrounding sediments, which in turn causes iron to precipitate and form concretions of iron oxides on the roots.

Macrophytes that grow in the brackish or salty parts of rivers have more freshwater within their cells than in the surrounding water, so the osmotic concentration gradient tends to force water out of the plant into the water. To avoid dehydration in this manner, vascular plants have a number of adaptations for maintaining their osmotic balance, including specialized cells that either serve as barriers to salt intrusion or glands that actively excrete salt. In mangroves (*Rhizophora spp.*, *Avicennia spp.*), for example, the root endodermis is specialized glands in its leaves. The emergent salt-marsh plant, Spartina, excretes salt crystals from specialized glands in its leaves.

Phytoplankton, the microscopic plants that float in the water column, rely on turbulent flow to remain high enough in the water column to photosynthesize. In this photic zone, phytoplankton can fix carbon and reproduce, but below this zone, they can only respire away their carbon. In shallow rivers, this is usually not a problem, as the entire water column may be photic, but larger rivers usually become turbid due to the downstream transport of sediments, organic detritus, etc. The main groups of phytoplankton are the green algae (colonial greens, diatoms, desmids, and others), blue-green algae, or cyanobacteria, and dinoflagellates (which are actually both autotrophic and heterotrophic). Diatoms have hard exteriors, called frustules, which are composed of silicate, so they cannot persist if this mineral is in short supply. They are an important taxonomic group in rivers, and often the riverine populations are maintained by inputs from tributaries. Algae in small streams often grows directly on rocks (epilithic algae) and can be an important source of organic matter in open (not covered by tree canopy) river segments. Blue-green algae, such as *Anabaena* or *Oscillatoria*, can fix nitrogen from the atmosphere by means of specialized cells called heterocysts, they will do this if there is an excess of available phosphorus, such as from cultural eutrophication.

Macrophytic algae are attached to hard substrates and are common in streams or along the littoral zones of larger rivers. Some, such as *Cladophora*, grow as filaments, while others (*Ulothrix, Oedogonium*) grow in tufts; Chara and Nitella are branched and their cell walls reinforced with calcium carbonate. The filamentous green alga *Cladophora glomerata* is a cosmopolitan species that does particularly well in the presence of elevated nutrients, and therefore large colonies serve as bellwethers of cultural eutrophication.

In a review of phytoplankton diversity from 67 rivers around the world, Rojo et al. (1994) found that average species richness was 126, and about half of the species were reported as sporadic. Average species richness was higher in temperate rivers (130 species) than in the Tropics (73 species). The total numbers of species enumerated, by family and zone, are given in Table I. Although the tropical rivers may have been underrepresented in the survey, nevertheless there are some clear trends. Diatoms are the most diverse taxon and predominate in temperate systems, whereas desmids are of greater importance in tropical rivers.

<table>
<thead>
<tr>
<th>Family</th>
<th>Total number of species</th>
<th>Photic (%)</th>
<th>Eutrophic (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlorophyceae</td>
<td>385</td>
<td>28</td>
<td>27</td>
</tr>
<tr>
<td>Chrysophyceae</td>
<td>43</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Cryptophyceae</td>
<td>29</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Cyanophyceae</td>
<td>141</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Dinophyceae</td>
<td>408</td>
<td>40</td>
<td>29</td>
</tr>
<tr>
<td>Dinoflagellata</td>
<td>27</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Euglenophyceae</td>
<td>112</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Xanthophyceae</td>
<td>11</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Zygophyceae (desmids)</td>
<td>221</td>
<td>9</td>
<td>29</td>
</tr>
</tbody>
</table>

*Family and tropical percentages indicate the taxon representation in temperate and tropical planktonic assemblages, respectively (source: Rojo et al., 1994).*
C. Invertebrates

The invertebrates that live in rivers can be divided into two groups, the zooplankton (animals that live suspended in the water) and the zoobenthos (animals that live in or on surfaces such as the river bottom, plants, etc.). River currents usually are much faster than the swimming speeds of the zooplankton, so washout downstream is a major problem for river zooplankton. Consequently, zooplankton are common in rivers only where their growth rates (or inputs from nearby lakes and wetlands) exceed loss rates from washout and other losses (e.g., predation). Thus, river zooplankton usually is dominated by rotifers and small cladoceran crustaceans, which have high growth rates. Zooplankton densities often are highest when the water is warm and low, when growth rates are high and washout rates are low. The biomass of zooplankton in the main channel of rivers usually is much lower than that in lakes, but off-channel habitats such as oxbow lakes may support a rich zooplankton. River zooplankton often includes a few dozen species of rotifers, cladocerans, and copepods, fewer than 10 of which are abundant. At least some of these species probably are sustained by inputs from nearby habitats that are more hospitable to zooplankton, such as backwaters, impoundments, or lakes. Most riverine zooplankton are neither specialized nor endemic: the same species and genera of zooplankton that live in rivers also are common in lakes and occur in rivers worldwide. Despite its low density and richness, river zooplankton plays an important role as food for young of many species of riverine fish as well as adults of a few specialized zooplanktivorous fish such as paddlefish, Polyodon spathula.

By comparison with the zooplankton, the riverine zoobenthos often is rich and dense. A typical large river contains several hundred species of benthic invertebrates representing a biomass of 1–100 g/m² (ash-free dry matter). Dominant groups include insects (chironomid midges and many others), bivalves, prosobranch (gill-breathing) snails, tubificid oligochaetes, and (in tropical rivers) crabs and prawns. As is the case with zooplankton, most of these species are not associated with the open waters of the main channel, but with shorelines, backwaters, vegetation, and other structurally complex habitats along the river margin (Table II). Unlike the zooplankton, many zoobenthic species are found only in rivers and are endemic to particular river systems. Thus, the rivers of southeastern North America contain several hundred species of mollusks and crayfish that are found neither in American lakes nor in rivers elsewhere in the world. Local density and species composition of riverine zoobenthos are controlled by the character of the bottom (particle size, organic content, stability), flow conditions (current speed, frequency and size of floods and droughts), the availability of food supplied by phytoplankton, rooted vegetation, attached algae, and the watershed, and biological interactions such as predation and competition. According to the river continuum concept, large rivers should be dominated by filter-feeding animals (both benthic and planktonic) and burrowing deposit-feeders. Data to test this idea in large rivers have not been collected, but it appears that large rivers actually contain a rich variety of feeding types, including filter-feeders, deposit-feeders, predators, algal scrapers, and herbivores that eat rooted plants.

Like the zooplankton, the zoobenthos is an impor-

---

### TABLE II

<table>
<thead>
<tr>
<th>Taxonomic group</th>
<th>Main channel</th>
<th>Dammed backwaters</th>
<th>Forested backwaters</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porifera (sponges)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Hydrozoa (hydroids)</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Turbellaria (flatworms)</td>
<td>9</td>
<td>8</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>Mollusca (snails and mussels)</td>
<td>32</td>
<td>42</td>
<td>53</td>
<td>62</td>
</tr>
<tr>
<td>Annelida (worms)</td>
<td>20</td>
<td>57</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>Crustacea</td>
<td>11</td>
<td>10</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>Insecta</td>
<td>224</td>
<td>234</td>
<td>536</td>
<td>733</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>306</td>
<td>354</td>
<td>663</td>
<td>800</td>
</tr>
</tbody>
</table>

Riverine fish species are adapted to flow regimes that are considerably more turbulent than lakes or oceans. In the riffles and rapid of lower order streams, many fish spend much of their time in "dead zones," or locally quiet waters, such as under rocks, behind boulders or tree roots, or along the riparian banks, emerging briefly to feed. The larvae of freshwater lamprey (amycoecotes) actually live under the stream, in sediments, for up to 6 years. In higher order streams, current velocities are slower, and fish may swim more freely in the water column. An exception is in tidal estuaries, where currents at peak flood and ebb may be as high as a meter per second or more. Some fish apparently "ride the tides" selectively, entering a tidal current to move either up- or downstream, then moving out of the flow to remain in place.

A fairly small, but important, group of fish undertakes large-scale migrations between freshwater and marine water. These so-called diadromous fish number about 200 species, most of which are anadromous, or freshwater spawning. This group includes many of the salmonids, the Southern Hemisphere galaxiids, and many herrings. These species use the freshwater environment to spawn; the young rear in the relatively protected riverine or estuarine environment and subsequently migrate to sea to feed, grow, and mature. A few species, notably the freshwater eels (Anguillidae), have the opposite strategy (catadromy), spawning at sea and then moving into freshwater to mature. Diadromous fish are important in many fisheries, as their spawning migrations tend to concentrate their numbers and make them relatively easy to catch.

Studies of riverine fish diversity on a global scale reveal that species richness increases with the size of the drainage basin and with river discharge (note: discharge and drainage size are closely correlated in unimounded rivers). Tropical rivers also tend to be more speciose, and this is in part due to warmer water temperatures, which stimulates primary productivity (Fig. 7). The Amazon River, being both the largest river and tropical, supports close to 4000 species. Some of the high-diversity groups in tropical rivers include the catfishes (Siluriformes), the minnows (Cypriniformes), and characins (Characiformes).

E. Other Large Animals

Although the numbers of other animal species that use rivers is far smaller, they can be important components of these ecosystems. Among the amphibians and reptiles...
are frogs and water snakes, many turtles, and the Croco-
dylia (alligators, crocodiles, and allles). Shoreline bird
species are numerous, including herons, ibises, cranes,
and top predators such as eagles and ospreys. Mammals,
such as beaver (Castor spp.), otter (Lutra spp.), and
hippopotamuses (Hippopotamus amphibius) function as
keystone species in many systems, either by virtue of
their physical alteration of the habitat (beavers, hippos)
or trophic influence (otters). Beavers build dams, which
create ponds that change the flow characteristics, and
hence the processing of organic matter, in small-
to
medium-sized (first to fifth order) streams in North
America and Europe. Hippos, by creating "aquatic high-
ways" between their daytime resting areas and nighttime
foraging grounds, can cause drainage alterations that
can be seen in satellite imagery (McCarthy et al., 1998).
Some tropical Asian rivers support freshwater dolphins
and sharks. Further, many large mammals, including
many of the threatened or endangered "charismatic
megalaua," occur in riparian zones of rivers; for exam-
ple, flooded tropical grasslands are an important ripar-
ian habitat for such animals.

VI. THREATS TO BIODIVERSITY
Humans have always modified rivers and their water-
sheds, both purposely and inadvertently. Human settle-
ment and development transform the landscape from
its natural patterns of vegetative cover to networks of
roads, drainage works, farms, and urban areas aimed
at efficient transportation, food production, and sanita-
tion. In the last century, technological advances have
enabled transformation of the landscape on an unprece-
dented scale. Now, the effects of human activities even
threaten to modify climatic factors. Effects of these
transformations on the resident species of the watershed
have been underestimated in the past, largely due to
ignorance. Table III reviews some examples of effects
of human activities on riverine ecosystems.

A. Dams
In many areas, dams are constructed for flood control,
water supply, and hydroelectricity. In order to keep up
with world demand for water and energy, dams are
### TABLE III
Threats of Biodiversity in Rivers

<table>
<thead>
<tr>
<th>Human impact</th>
<th>Proximate effects</th>
<th>Consequent short-term change in</th>
<th>Species richness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Damming</strong></td>
<td>Reduction and change of temporal patterns of flow,</td>
<td>Decreases downstream associated with nutrient reduction</td>
<td>Probable decline, particularly in migratory species</td>
</tr>
<tr>
<td></td>
<td>sediment flux, and turbidity</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Stream channelization</strong></td>
<td>Increase in flow velocity; disruption of benthos,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hyporheic zone</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nutrient loading</strong></td>
<td>Fertilization of productivity; fueling of increased</td>
<td>Increase in GPP, respiration;</td>
<td>Declines</td>
</tr>
<tr>
<td></td>
<td>respiration; BOD</td>
<td>BOD</td>
<td>associated with</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ammonia, potential collapse of some communities</td>
</tr>
<tr>
<td><strong>Toxic substance loading</strong></td>
<td>Increased mortality of resident species; reproductive</td>
<td>Probable reduction in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>failure</td>
<td>productivity in response to toxicity</td>
<td></td>
</tr>
<tr>
<td><strong>Introductions of alien species</strong></td>
<td>Colonization of ecosystem by new competitors</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Potentially dramatic shifts if the introduction is successful</td>
<td></td>
</tr>
<tr>
<td><strong>Land use change</strong></td>
<td>Typically increased sediment and nutrient load;</td>
<td>Probable shift toward</td>
<td>Potential</td>
</tr>
<tr>
<td></td>
<td>hydrological alterations; possible increases in</td>
<td>heterotrophy in response to nutrient additions and light reduction</td>
<td>decreases</td>
</tr>
<tr>
<td></td>
<td>loading of toxicants</td>
<td></td>
<td>associated with</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>the first four impacts above</td>
</tr>
<tr>
<td><strong>Overharvesting of species</strong></td>
<td>Depletion of target species; perturbations in predator</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>and prey populations</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Climate change</strong></td>
<td>Changes in temperature, precipitation, evaporation,</td>
<td>Increases or decreases,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and atmospheric CO2</td>
<td>depending upon location</td>
<td></td>
</tr>
</tbody>
</table>

Being constructed at an unprecedented rate. Approximately 40,000 large dams have been constructed in the past 50 years, and the total area of land flooded by large dams exceeds 400,000 km²—an area the size of California. In the United States alone, there are nearly 2,000,000 small dams, and 68,000 dams two stories (6 m) or higher.

The most obvious ecological effect of a dam is to act as a barrier for water flowing downstream. Sediments are no longer transported downstream, but rather accumulate behind the dam. For example, prior to the construction of the Aswan Dam in the 1960s, the Nile River in Egypt transported over 120 million tons of sediment to the Mediterranean annually, and 10 million tons were dispersed in the floodplain and delta, providing fertile soils which were farmed for millennia. The dam now retains 98% of the sediments, cutting off the vital mechanism of soil renewal, causing a decline in floodplain agriculture, and also resulting in coastal erosion.

In ecological terms, a dam is a most effective barrier for migrating fish and other species, and the presence of dams is known to eliminate or drastically reduce populations of these species unless mitigating structures (e.g., fish ladders) are put in place. Even these are only partially successful in many rivers with hydroelectric dams; the passage of fish to upstream spawning grounds is a “one-way ride;” for most of these fish are killed as they attempt to move downstream and pass through the turbines.

Less obvious is the effect of dams on flow regime and biogeochemistry. Dams alter the temporal pattern...
of river discharge as well as the flow pattern of groundwater. Nutrient and sediment fluxes are typically diminished downstream because of settling or biological uptake in the impoundment. Resulting changes in elemental ratios (for example, the decrease in available silicate compared to nitrogen and phosphorus; Humborg et al., 1997) downstream have been associated with species shifts, reductions in productivity, and (it is claimed in some cases) decreased yields in the downstream fishery.

B. Stream Channelization
Stream channelization is designed to improve the navigability of rivers or to reduce flooding potential in streams. In the former case, the river channel is deepened and widened by dredging, which destroys benthic habitat. In the latter, the streambed is sometimes straightened and “paved” to increase the capacity of the stream to transport water downstream. These processes change the flow regime of the stream, favoring species that tolerate faster, turbulent currents, and excluding others. Changes to the streambed can affect the conductivity of water through the hyporheic zone, which affects nutrient processing (see Section IV).

C. Nutrient Loading
Nutrient loading, as discussed above, fuels the ecosystem. Excessive loading results in catastrophic collapse when the resulting increase in metabolism exceeds the available oxygen supply of the river. Cultural eutrophication of rivers is severe to dire in some parts of the developing world, in particular in India, Southeast Asia, and Africa, as well as around many urban centers worldwide.

D. Toxic Substances
Associated with pesticide application in agriculture (or suburban gardens), industrial effluents, and chemical spills, toxic substances have a spectrum of effects, all deleterious to organisms. Some herbicides dramatically reduce productivity by direct mortality of aquatic plants. Other chemicals kill sensitive species higher in the food chain. Sublethal effects include behavioral changes (due to endocrine-disrupting compounds) in predators or prey species, which can shift the structure of food chains, increase susceptibility to disease, and cause infertility. What complicates this story is that different organisms do not respond uniformly to exposure to a particular toxic substance, so that the effect on a particular riverine system is determined largely on the mix of species present. Chronic exposure to a stream of toxic material will gradually shift the species present to those that can tolerate it, generally resulting in a decline in species richness.

E. Overharvesting Species
In addition to humanity’s indirect impacts on biota, humans threaten biodiversity directly through the overharvesting of species. In the United States and Canada, a number of river-associated fisheries are currently in decline, including those for several species of Pacific salmon, Atlantic salmon, American shad, sturgeons, and oysters. Worldwide, fishing pressures in rivers are enormous. Many river fisheries in developing countries are artisanal, providing food for a local community, and little is known about harvesting pressures. However, responses of fish communities to overharvesting are becoming understood. For instance, in one West African river, catch per unit of fishing effort remained relatively constant from the 1950s through the 1990s, but the trophic structure of the harvested fish went from large piscivores to small planktivores, with loss of food web complexity (Fig. 8; Welcomme, 1999). This phenomenon is called “fishing down food webs” and was first noted in marine fisheries (Pauly et al., 1998). At the population level, overharvesting often selectively removes older age classes, pushing the age of first reproduction down, and often skews the sex ratios in the population (for example, female sturgeon are selectively sought for their roe from which caviar is produced).

F. Introduced Species
Humans are also introducing new species into ecosystems at an unprecedented rate. Some of these introductions are intentional: for example, salmon, trout, and bass are stocked into rivers and lakes all around the world for sport fishing, and grass carp are introduced to control aquatic weeds. Many introductions are unintentional, arising from the pumping out of ballast water in large tankers, overland transfer of boats with attached organisms from one drainage to another, or accidental spills. Although most organisms that are introduced into a novel environment perish, the increase in commercial boat traffic particularly increases the likelihood of a successful introduction.

Rates of exotic species introductions are difficult to quantify, in part because many of the species are too small to notice. However, Mills et al. (1996) documented the rate of species introductions into the Hud-
son River, in New York State, since 1840, the rate has been about one successful introduction per year—much higher than natural rates of species flow.

Direct perturbations of the mix of species, either by introduction of new species or by removal of extant species by overharvesting, may cause dramatic changes in relations of the food web, thereby also affecting predators and prey. Some species additions affect their physical and chemical environment as well: in North America, zebra mussels, which were introduced into the Great Lakes in 1988 and which spread rapidly through the Mississippi and eastern drainages, have been documented to increase the clarity of the water column by virtue of their filter feeding and also carpet large areas of the benthos, where they have successfully colonized.

VII. ASSESSMENT AND MANAGEMENT

River ecosystems are clearly important and at risk. Their direct economic importance to societies includes their use in transportation, water supply, energy, and provision of harvestable products. Their indirect importance, sometimes termed essential ecosystem services, includes their fundamental role as biogeochemical transformers.
of energy and matter, physical transformers of the landscape (shaping the land through fluvial processes), and the provision of a wide variety of ecological habitats along the river continuum.

Historically, the assessment of river ecosystems was hindered by two factors. First, on a practical basis, it is difficult to adopt a single methodology by which to assess the state of a large river system from headwaters to mouth. Many of the methods that were developed for small-stream study are not easily adapted to the higher order parts of the system, and at the other end of the continuum, oceanographic methodologies may be designed for scales of study that are larger than rivers. Second, until the development of scale-spanning river paradigms such as the river continuum concept, ecological studies of rivers tended to consider a river in isolation from its watershed; as is now widely recognized, upstream processes in the watershed (including human-accelerated land use change) have major downstream effects on rivers. (A third difficulty may be due to the somewhat arbitrary tendency of researchers to specialize on broadly different parts of the system: there are many more small stream and estuarine scientists than there are ‘river scientists.’)

Recent revival of the concept of ‘ecosystem health’ has proven to be a useful point of departure for current study and management of river ecosystems (see, for example, Karr and Chu, 1999; and Scow, 1999). There are many more small stream and estuarine scientists than there are ‘river scientists.’

See Also the Following Articles
ENDANGERED FRESHWATER INVERTEBRATES • ESTUARINE ECOSYSTEMS • FISH, BIODIVERSITY OF • LAKE AND POND ECOSYSTEMS • INVERTEBRATES, FRESHWATER, OVERVIEW • MARINE ECOSYSTEMS • WETLANDS ECOSYSTEMS

Bibliography


I. Classification

II. Life Histories

III. Special Adaptations

IV. Historical Status and Overview of Decline

V. Habitat, Harvest, Hatcheries, and Hydropower as Major Threats to Salmonid Diversity

VI. What Is the Value of Salmonid Biodiversity and Can We Rescue Salmon from Extinction?

GLOSSARY

alvin Newly hatched salmon with unabsorbed yolk sac.
anadromous Aquatic organisms that spawn and undergo early development in freshwater, migrate to sea to grow and mature, and return to freshwater to reproduce.
fry Juvenile salmon that have absorbed their yolk sac and emerged from the gravel.
grilse Atlantic salmon that spends only one year at sea before returning to spawn.
iteroparous Aquatic species that can spawn more than one time.
jack A sexually precocious male salmon that spends one winter or less at sea before returning to its natal stream to spawn.
kelt A spawned out Atlantic salmon or steelhead that is returning to the ocean to recuperate, sexually mature, and spawn again.
parr Older juvenile salmon that are distinguished by parr marks—prominent oval spots on their sides.
semelparous Aquatic species that die after reproduction.
smolts A downstream migrating salmon that has begun the physiological transition to seawater.

THE PACIFIC AND ATLANTIC SALMON are a diverse group of fish distributed throughout the Northern Pacific and Northern Atlantic Oceans and their adjacent freshwater coastal habitats. The Pacific Salmon include seven species in the genus Oncorhynchus: pink (O. gorbuscha), chum (O. keta), sockeye (O. nerka), coho (O. kisutch), and chinook (O. tshawytscha) salmon are found in North America and throughout the Pacific rim; masu (O. masou) and amago (O. amago) salmon are found only in Asia. In addition, there are two trout members of the genus Oncorhynchus that have anadromous forms: steelhead (O. mykiss, the anadromous form of rainbow trout) and the sea-run cutthroat trout (O. clarkii). There is a single species of species of Atlantic salmon (Salmo salar). This chapter focuses on the anadromous salmonids of North America.

I. CLASSIFICATION

Over the past two centuries, fish have been the subject of extensive systematic research and countless taxonomic revisions—and among the some 24,000+ named species of fishes, the salmonids have been no exception.
Currently, the salmonids are members of the class Os
treichytes (the bony fishes), the subclass Actinopterygi
(the ray-finned fishes), the division Teleostei, and the
super order Protacanthopterygi. The Protacanthoptery-
gi, which contains some 310-plus species, is divided
into three orders: the Salmoniformes (the salmonids),
the Esociformes (the pike and mudminnow), and the
Osmeriformes (the smelts). There are about 70 species
in the family Salmonidae (order Salmoniformes), which
can be divided into three subfamilies: Salmoninidae
(salmon and trout), Coregoninidae (whitefishes), and Thy-
nallinae (graylings).

According to Behnke (1992), the earliest fossil rec-
ord of the salmonids is Eosalmo driftwoodensis, which
was discovered in Eocene deposits some 40 to 50 mil-
lion years ago in British Columbia. The separation of
the Atlantic Ocean group (Salmo) and the Pacific Ocean
group (Oncorhynchus) likely occurred by the mid-Mio-
cene about 15 million years ago. By the end of the
Miocene (about 5 million years ago), the ancestral form
of Oncorhynchus in North America evolved into two
distinct lineages, one leading to the Pacific salmon and
the other to the Western trout. The evolution of Pacific
salmon is still the subject of scientific debate. Particu-
larly intriguing is the question of whether salmon evolved
first in freshwater, only later developing the
ability to migrate to the ocean, or whether they evolved
first as marine species that later developed the ability
move into freshwater refugia to spawn and rear. No
matter which case, the currently recognized species are
generally believed to have differentiated only
500,000 to 1 million years ago, which represents an
extremely short time period for such extensive specia-
tion to occur. Rapid speciation was probably favored by
the geological and climatic history of North American
landscape—constantly being shaped and reshaped by
volcanism, alternating periods of warm and cold cli-
mate, alternating arid and pluvial periods, and re-
arrangement of major drainage basins.

The classification of North American salmonids does
not stop with the traditional Linnean system of classifi-
cation for plants and animals. In particular, salmon
have recently been classified into evolutionary signifi-
cant units or ESUs. An ESU is any group of populations
that is distinct according to two criteria: (a) It must
intensively exchange genes with other conspecific
population units, and (b) It must represent an impor-
tant component in the evolutionary legacy of the species
(Waples, 1995). The concept of the ESU as a fundamen-
tal unit of conservation under the Endangered Species
Act has been reviewed and endorsed by the National
Research Council (1995) and is an important landmark
in evolutionary thinking applied to conservation. This
is because the ESU emphasizes genetic structure and
life history adaptation as a focus for protection as op-
posed to more typological definitions of species or sub-
species based on appearance.

In Western North America, 36 distinct salmonid ESUs have been identified, with 23 of them currently
listed as threatened or endangered under the U.S. En-
dangered Species Act.

II. LIFE HISTORIES

Salmon and anadromous trout display an amazing vari-
ety of life history patterns, both among and within
species. However, many features are held in common.
The life cycle begins with females depositing eggs in
neests or redds in gravel bottoms of rivers and lakes.
The young emerge from the gravel as alevins or fry to
rear in fresh water for periods varying from a few days
to several years. As juveniles or parr approach the time
for their seaward migration, they undergo physiological
metamorphosis or smoltification in which, among other
changes, they develop the ability to osmoregulate in
the hypertonic marine environment. In the ocean, they
often undertake extensive migrations, covering many
thousands of miles, while they grow and mature. After
periods that range anywhere from a few months to 4
or more years, adult salmon return with remarkable
fidelity to their natal river and tributary to spawn and
continue the cycle.

The key characteristic that distinguishes the Atlantic
salmon and anadromous trouts from the five species of
Pacific salmon is the fact that Pacific salmon species are
semelparous and die after spawning only once, whereas
anadromous trout and Atlantic salmon are iteroparous
and may spawn multiple times. As is evident in the
ensuing sketches of species-specific life history traits,
diversification among salmon primarily has involved
permutations that alter the schedule and timing of
freshwater versus ocean habitat use (Busby et al., 1997;

A. Pink Salmon (Also Known as Humpback Salmon)

Pink salmon, which are distributed throughout the
North Pacific and Bering Sea north of 40°N, are the
most abundant of the Pacific salmon. In North America,
pink salmon spawn from August to November and gen-
erally return at smaller sizes than other salmon—
ranging between 1.0–2.5 kg. Returning adults undergo only modest migrations, often spawning in intertidal areas. After emerging from the gravel, they swim directly to sea, where they migrate and feed extensively in the North Pacific. After 18 months at sea, they return to their river of origin to spawn and die. This fixed 2-year cycle leads to the genetic isolation of what are referred to as odd- or even-year runs. Typically, different geographic regions are dominated by odd- or even-year runs, but not both. The tendency of having only one type of run (odd or even) in each river may exist because the rivers are not productive enough to support spawning by pink salmon every year.

B. Chum Salmon (Also Known as Dog Salmon)

Chum salmon are second only to pink salmon in abundance, but because of their large size they comprise up to 50% of the marine biomass of salmon in the North Pacific. Historically ranging from the San Lorenzo River in California to the Mackenzie River on the Beaufort Sea to the north, returning adults can weigh as much as 20 kg. Like pink salmon, chum salmon tend to spawn low in river systems and migrate quickly to estuarine and marine waters, often spending extended periods rearing in estuarine habitats. Once offshore, chum salmon are widely distributed throughout the North Pacific, typically spending 2 to 5 years at sea, but occasionally as long as 7 years. There are two generally recognized races of chum salmon, which are distinguished by their time of return to fresh water: early and late stocks, or summer- and fall-runs, with the fall-run populations far more prevalent.

C. Sockeye Salmon (Also Known as Red Salmon)

Sockeye salmon are the third most abundant of Pacific salmon, comprising 17% by weight and 14% by number of the salmon total. They display a wide variety of life history patterns, including a subspecies that spends its entire life in fresh water—the kokanee. The vast majority of sockeye salmon populations spawn either in tributaries associated with lakes or directly in lakes on gravel shoals. After emerging from the gravel, the juvenile sockeye salmon spend from 1 to 3 years rearing in lakes. There are, however, rare populations that spawn in the rivers not associated with lakes and populations that migrate directly to sea shortly after emergence from the gravel. Like the other salmon species with extended freshwater rearing, smolts tend to move quickly downstream, spending little time in the estuarine and near-shore habitats and are found widely distributed throughout the North Pacific Ocean. Adults typically spend 1 to 4 years at sea, before returning to spawn in their natal lakes and tributaries. Maturing adults enter fresh water from late spring to early summer, with spawning occurring from late July through January.

D. Coho Salmon (Also Known as Silver Salmon)

Coho salmon are the most widely distributed of the Pacific salmon, but they are the least abundant. The vast majority of the populations have a 3-year fixed cycle—spending about 18 months in fresh water and 18 months at sea. In North America, they range from Monterey Bay, California, in the south to the Kukpuk River in Alaska. Coho salmon typically enter fresh water from August to January and quickly move upstream to spawn in coastal streams and the smaller tributaries of large river systems. After emergence from gravel, juveniles rear in the smaller tributaries, and overwinter an additional year before beginning their seaward migration during the spring of their second year of life. At sea, coho salmon tend to stay on the continental shelf, feeding and growing in marine waters at roughly the same latitude as their stream of origin.

E. Chinook Salmon (Also Known as King Salmon)

Chinook salmon are the largest of the Pacific salmon—returning at weights up to 45 kg. In North America they range from central California to Kotzebue Sound, Alaska, and spawn in diverse habitats ranging from tidewater to small tributaries located as far as 3200 km inland. Locally, runs tend to be identified by the time at which adults enter fresh water and begin their spawning migrations, hence the colloquial names spring-, summer-, fall-, late fall-, and winter-run chinook salmon. However, these designations do not capture the true biological diversity of this species. Along with sockeye salmon, chinook salmon display the greatest within-species variation in life histories of all the Pacific salmon, including variation in age at seaward migration, variation in duration of freshwater, estuarine, oceanic residency, variation in ocean distribution and oceanic migratory patterns, and variation in age and season of spawning migration. This variation is derived from two sources, one racial and the other
environmental. The racial differences result from geographic and hence genetic isolation that divides chinook salmon populations into one of two behavior forms, stream-type and ocean-type. Stream-type populations spend one or more years in fresh water before seaward migration, undertake extensive offshore migrations, and return to their natal streams during the spring or summer, often holding in fresh water for several months before spawning in late summer. In contrast, populations of ocean-type chinook salmon migrate to sea prior to age 1, utilize estuarine and nearshore habitats for extended periods, often spending their entire ocean residence on the continental shelf, and return to their natal stream in the fall immediately before spawning.

The second major source of life history variation in chinook salmon is expressed as run-timing and age-at-spawning differences within races, within populations, and sometimes even within cohorts. The adaptive value of such variation is well known in evolutionary theory—by spreading the timing of spawning among years, these fish are effectively “spreading the risk” and reducing the likelihood of total failure if any single year or even week is disastrous (the argument works at a variety of temporal and spatial scales). Uncertainty in juvenile survival and marine productivity is enormous for salmon, and this is reflected by the fact that year-to-year variability in spawner counts for salmon (see Fig. 1) have coefficients of variation typical of pest insects and other notoriously fluctuating species.

F. Atlantic Salmon

Atlantic salmon are the only anadromous salmonid native to the Atlantic coast of North America. Historically found in nearly every major coastal river north of the Hudson River, the extinction of several U.S. populations by the end of the 19th century has shifted the southern extent of the species distribution over 200 kilometers northward.

The majority of adult Atlantic salmon enter fresh water to spawn during the spring (late May to early June primarily), spending most of the summer holding in fresh water and then spawning in fall. In general, the more northern populations arrive earlier, and spawn earlier. However, a few populations do not enter fresh water until late summer or early fall, bypassing the summer holding period and consequently spawning directly. Juvenile Atlantic salmon emerge from the gravel in the spring and typically spend from 2 to 4 years rearing in fresh water (although some populations spend as little as 1 year and others as much as 8 years in fresh water). Outmigration of juveniles usually occurs from April to mid-June, with the smolts spending little time in the estuarine and nearshore zone. Atlantic salmon are widely distributed in the North Atlantic, with American and European stocks often concentrating and mixing in the waters off Greenland—a distribution which enabled the development of highly efficient

![Image of salmon life history graphs](image-url)
commercial fisheries that nearly drove the species to extinction in the 1960s. The majority of Atlantic salmon spend two winters at sea before returning to spawn.

G. Steelhead

There are two major genetic groups of *O. mykiss* presently recognized in North America: the inland group and coastal group, with the distinguishing character being their use of freshwater habitat. As the name implies, the coastal form populates the wet coastal regions of western North America, whereas the inland form is found in the more arid interior regions of western North America. Both forms occur in British Columbia, Washington, and Oregon; Idaho has only the inland form, and California has only the coastal form. Both forms display a wealth of life history variation, including an anadromous form that is known as steelhead. Although steelhead can be observed entering fresh water to spawn throughout the year, they usually display striking seasonal peaks in spawning activity. This has resulted in the use of seasonal run timings to distinguish different populations (fall-run, summer-run, etc.). Steelhead can be divided into two basic reproductive ecotypes based on their state of sexual maturity at the time of river entry and the duration of spawning migration. The stream-maturing type (commonly known as fall steelhead in Alaska, summer steelhead in the Pacific Northwest and northern California) enters fresh water in a sexually immature state and spends several months maturing before they spawn. In contrast, the ocean-maturing type (spring steelhead in Alaska and winter-run elsewhere) enters fresh water in a sexually mature state and spawns shortly thereafter. While many rivers contain populations of both ecotypes, coastal streams tend to be dominated by winter steelhead, whereas inland reaches tend to be predominately the summer steelhead. Despite the major differences in migration timing of the adults, both ecotypes of steelhead generally spawn from December through March. After emerging from the gravel in late spring, juveniles spend up to 7 years in fresh water, with 1 to 3 years being the norm. Most populations in Alaska and British Columbia tend to smolt and migrate to sea at 3 years of age, whereas the more southern stocks tend to smolt at 1 or 2 years of age. Most North American steelhead spend 2 years at sea; however, some of the more southern stocks characteristically spend only 1 year at sea. An unusual variant in life history found among selected steelhead populations in Southern Oregon and Northern California is that of the half-pounder. Half-pounders are juvenile fish that migrate to sea, only to return to fresh water in an immature state after only 2 to 4 months. The biological basis for this unusual behavior is unknown. As an iteroparous species steelhead can spawn multiple times, with the frequency varying both within and among populations. The more northern populations tend to have a lower proportion of repeat spawners than those further south.

H. Sea-Run Cutthroat Trout

Cutthroat trout display a complex suite of life histories—some exclusively fresh water and others involving a period of marine residency. While it is the latter life history strategy that justifies their inclusion in this chapter, the reader should be aware that the interrelationships between the different life history variants are not fully understood, and sea-run forms probably cannot be treated as isolated from the freshwater forms (Trotter, 1989).

In North America, sea-run cutthroat trout are distributed along the Pacific Coast from the Eel River in California to Prince William Sound in Alaska. Adults typically spawn in the small tributaries of low gradient streams and the lower-gradient downstream reaches of large river systems from late winter to late spring. Juveniles rear in fresh water for 2 to 4 years, generally utilizing slower moving waters and pools. Those populations that migrate into the protected waters of large estuaries and sheltered bays tend to smolt at age 2, whereas those populations migrating into the nearshore ocean often delay smoltification until age 3 or 4. Sea-run cutthroat trout seldom spend more than a few months in marine waters and stay well within the nearshore zone, typically returning to fresh water in late summer, fall, or winter of the same year they migrate to sea. Although fish returning to fresh water to spawn return to their natal stream with high fidelity, nonmaturing fish are known to leave the ocean and overwinter in nonnatal streams. Like steelhead and Atlantic salmon, sea-run cutthroat are iteroparous, and multiple spawning is quite common.

III. SPECIAL SALMON ADAPTATIONS

A. Anadromy

Anadromous fishes are those that spawn in fresh water, migrate to the ocean to forage and mature, and return to fresh water to spawn and begin the cycle again. As anadromous species, the salmonids pass through multiple habitat transitions and face major physiologi-
cal challenges—the most critical adaptation for this lifestyle is the ability to osmoregulate in both freshwater and marine environments, and to do so at precisely prescribed times. The evolution of anadromy has intrigued biologists for as long as salmon have been studied. Although many theories have been put forth to explain the evolution of anadromy, such a major life history character must, on balance, provide an overall advantage to individual fish. The potential advantages of anadromy include (a) a mechanism for dispersal and rapid recolonization of regions suddenly made available to fish (such as the retreat of a glacier), and (b) being able to get the “best of two worlds” (freshwater and marine). The “best” of the freshwater world involves reduced predation on eggs and juvenile fish because the embryos develop in the protected environment of gravel until they emerge at a large enough size to actively avoid stream-dwelling invertebrate predators. The “best” of the marine world involves higher growth rates for larger maturing fish in the oceans, made possible because in upwelling areas and years of high productivity the regions of the ocean frequented by salmon provide an abundance of larger prey items that far exceeds the prey base of most freshwater ecosystems.

B. Homing

The legendary ability of Pacific and Atlantic salmon to home with high fidelity to their stream of origin is one of the more spectacular migrations in nature. Beginning with the pioneering work of Hasler, Quinn, and others, it is growing increasingly clear that this ability involves a combination of magnetic detection, solar navigation, and olfactory cues. The salmon’s ability to navigate the high seas appears to be based on the ability to detect inclination and declination of the earth’s magnetic field, combined with celestial orientation and an “endogenous circannual rhythm, synchronized by day length or rate of daylight change to sense latitude” (Quinn, 1982). Once in the vicinity of their home stream, it appears that salmon homing is facilitated by following unique gradients of odors that are learned during the downstream migration of juveniles. Despite uncertainty about the exact mechanism of homing, the fact that salmon return to their natal streams is ecologically quite significant. From a very practical perspective, it assures that salmon return to an area with favorable spawning and rearing conditions. In addition, although often considered a failure in homing, the small proportion of salmon in a population that stray to neighboring streams play an important role in the colonization of vacant habitat. Such a nearest neighbor strategy is more likely to be successful in matching salmon with habitat than a random pattern of straying, and hence it is a more efficient means of reestablishing sustainable populations.

IV. HISTORICAL STATUS AND OVERVIEW OF DECLINE

Historically, spawning populations of Atlantic salmon were abundant in most rivers and their tributaries throughout Europe, the British Isles, and the Baltic North, and in all the major North America rivers north of the Hudson River along the Atlantic seaboard. Likewise, the thousands of rivers and streams dissecting the coastal lands surrounding the North Pacific Ocean supported major populations of Pacific salmon and anadromous trout. Today, however, these once plentiful species are greatly reduced in both abundance and distribution. Due largely to the actions of humans and the encroachment of “civilization,” populations of Atlantic salmon throughout much of Europe and in New England have all but disappeared from the landscape. Similarly, in the Pacific Northwest, only a fraction of the streams and rivers that once supported robust salmon populations continue to do so. Although there are some obvious differences, there are also some striking similarities in the causes and patterns of decline of Pacific and Atlantic salmon—in Europe the onset of the Industrial Revolution, in New England the colonization by European settlers, and in the Pacific Northwest the arrival of the Euro-American settlers. The 1991 review of Nehlsen et al. perhaps best captures the status of salmon in the Pacific Northwest by noting that of 214 individual populations of native naturally spawning populations of salmon in California, Oregon, Idaho, and Washington, 101 were at high risk of extinction, 58 were at moderate risk of extinction, and 54 were of special concern. The situation has become so dire that the 25 currently endangered and threatened salmonid ESUs span almost every major freshwater ecosystem from the Sacramento River northward, with vast political and social consequences for the western United States (if the Endangered Species Act is enforced with vigor).

V. HABITAT, HARVEST, HATCHERIES, AND HYDROPOWER AS MAJOR THREATS TO SALMONID DIVERSITY

Many factors, both natural and human caused, are major sources of mortality among populations North
American salmon. While there is a handful of examples of a single factor causing populations to go extinct (e.g., a dam completely blocking access to spawning and rearing habitat), more often than not there is a constellation of factors that place salmon at risk of going extinct. An extremely thorough review of the current status of Pacific Northwest salmon and the factors contributing to their decline was recently published by Lichatowich (1999). In Salmon without Rivers, Lichatowich chronicles the methodical extirpation of one salmon population after another as the Euro-American settlers and their industrial economy replaced what he refers to as the "gift economy" of the native Indian tribes. Given the widespread destruction of habitat and a "modern" harvest philosophy built on the false premise of unlimited productivity, it is indeed amazing we still have viable populations of salmon in the Pacific Northwest.

The human-caused threats to salmon are often described in a framework of the four H's: habitat, harvest, hatcheries, and hydropower. This convenient framework is followed here, but it is also a little misleading in how it categorizes risk factors in tidy compartments. For example, while it has been traditional to view habitat threats as fresh water and affected only by human destruction, this view grossly neglects the importance of estuarine and marine habitats and the role they play in salmon productivity. Further, while there is a tendency to dismiss the ocean as a homogeneous "black box" that is unaffected by human activities, there is likely to be a strong interaction between some of the H's such as hatcheries and ocean conditions. For example, the industrial-scale production of hatchery fish in large river systems such as the Columbia River Basin may not be problematic except during periods of low primary and secondary productivity in the nearshore ocean. Hatchery and naturally produced smolts may compete only during periods of low ocean productivity. In addition, the ill effects of individual fitness compromise by gene flow from hatchery fish may be felt only if poor ocean conditions.

A. Habitat

The destruction of freshwater habitat is perhaps the single most significant factor affecting Pacific and Atlantic salmon in North America. Logging, mining, and agriculture (including grazing, irrigation withdrawals, soil erosion, and chemical contamination) have been major causes of salmon decline. While many of the impacts of these activities are similar, some are unique. For example, the logging practices of the past two centuries has caused widespread destruction of riparian vegetation along stream corridors (producing increased in-stream temperatures) and has caused landslides due to poor road-building practices (resulting in sedimentation and destruction of spawning and rearing habitat). Other impacts, such as the creation of thousands of acres of impermeable surfaces, are unique to urbanization. While it is beyond the scope of this chapter to describe in detail the nature of all of the different forms of habitat destruction, it is important to note that most are clearly the product of human actions and hence controllable.

B. Harvest

Pacific and Atlantic salmon have been the target of harvest by the indigenous peoples of North America for many thousands of years. However, it was not until the early- to mid-1800s that the Euro-American settlers (with their advanced methods of capture and preservation) began to so overfish salmon that extinction became a possibility. Particularly destructive have been offshore fisheries than indiscriminately harvest mixtures of plentiful hatchery fish and dramatically less abundant wild fish. Although harvest is currently regulated to some extent, it still continues at substantial rates, even on endangered and threatened ESUs. For example, fall chinook salmon from the Snake River are currently harvested at an annual rate that exceeds 20%, and some listed ESUs in the Puget Sound area are harvested at annual rates that exceed 40%. Fish taken by harvest represent individuals with maximum reproductive value and the impact of their mortality prior to spawning is unarguably large. It may be that the models guiding our harvest policy are themselves flawed, as suggested by recent theoretical papers which suggest that harvest models built with the concept of an optimum percentage take and maximum sustained yield often fail in highly variable environments (Lande et al., 1997; Ludwig, 1998). The good news is that simple harvest models built with the concept of an optimum percentage take and maximum sustained yield often fail in highly variable environments (Lande et al., 1997; Ludwig, 1998).

The destruction of freshwater habitat is perhaps the single most significant factor affecting Pacific and Atlantic salmon in North America. Logging, mining, and agriculture (including grazing, irrigation withdrawals, soil erosion, and chemical contamination) have been major causes of salmon decline. While many of the impacts of these activities are similar, some are unique. For example, the logging practices of the past two centuries has caused widespread destruction of riparian vegetation along stream corridors (producing increased in-stream temperatures) and has caused landslides due to poor road-building practices (resulting in sedimentation and destruction of spawning and rearing habitat). Other impacts, such as the creation of thousands of acres of impermeable surfaces, are unique to urbanization. While it is beyond the scope of this chapter to describe in detail the nature of all of the different forms of habitat destruction, it is important to note that most are clearly the product of human actions and hence controllable.

B. Harvest

Pacific and Atlantic salmon have been the target of harvest by the indigenous peoples of North America for many thousands of years. However, it was not until the early- to mid-1800s that the Euro-American settlers (with their advanced methods of capture and preservation) began to so overfish salmon that extinction became a possibility. Particularly destructive have been offshore fisheries than indiscriminately harvest mixtures of plentiful hatchery fish and dramatically less abundant wild fish. Although harvest is currently regulated to some extent, it still continues at substantial rates, even on endangered and threatened ESUs. For example, fall chinook salmon from the Snake River are currently harvested at an annual rate that exceeds 20%, and some listed ESUs in the Puget Sound area are harvested at annual rates that exceed 40%. Fish taken by harvest represent individuals with maximum reproductive value and the impact of their mortality prior to spawning is unarguably large. It may be that the models guiding our harvest policy are themselves flawed, as suggested by recent theoretical papers which suggest that harvest models built with the concept of an optimum percentage take and maximum sustained yield often fail in highly variable environments (Lande et al., 1997; Ludwig, 1998). The good news is that simple harvest models built with the concept of an optimum percentage take and maximum sustained yield often fail in highly variable environments (Lande et al., 1997; Ludwig, 1998).
FIGURE 2. Analysis of the benefits of a temporary moratorium on harvest for Snake River fall chinook salmon. A simple age-structured model iterates salmon populations forward 20 years, using estimates of variability and recruitment from 1980 onward. Two alternative simulations are run: simply continuing to harvest at the current rate versus halting harvest for 10 years and then reinitiating harvest for years 11 through 20, but at 75% of the current rate. Note that if there is a 10-year moratorium followed by harvest at 75% of the current rate, the population rebuilds and total cumulative harvest after 20 years (even including 10 years without any harvest) ends up being higher than if current practices are continued. One may be able to have fish and harvest, if harvest is temporarily halted.

C. Hatcheries
The building of hatcheries and the production and release of millions of hatchery salmonids in the Pacific Northwest closely tracked the wholesale destruction of habitat and overfishing. Indeed, hatcheries were widely viewed as the solution—in fact, a solution that allowed the extraction of natural resources to proceed unchecked. It was not until the past several decades that biologists began recognizing that such use of hatcheries were, by themselves, a potential problem. Beyond the obvious creation of harvest opportunities and destructive mixed-stock fishing, there is now much greater recognition that the release of large numbers of hatchery fish can have detrimental genetic and ecological impacts on natural populations. Interbreeding of hatchery and wild fish can reduce the fitness of wild fish through the disruption of local adaptation; and increased interactions between hatchery and wild fish can lead to competition for food and space, as well as the spread of infectious diseases.

D. Hydropower
Dams block access to spawning habitats, create reservoirs that flood spawning habitats, and may stress fish migrating through them to such a large extent that females cannot replace themselves due to depressed physiological fitness. Dams affect both upstream adult spawning migrations and downstream juvenile migrations. Primary mechanisms for the negative impacts of dams involve increased mortality of juveniles associated with passage through turbines, atmospheric gas supersaturation, and decreased flow and turbidity. On the other hand, engineering improvements in dams can mitigate some of these effects, although certainly not to the extent that rivers return to the equivalent of their free-flowing natural state. Because dams are the most visible impediment to salmon survival, getting fish around, over, or through dams has been the focus of much research. Unfortunately, the understandable focus on hydropower as a risk factor for salmon has left the ‘‘other H’s’’ less well studied. Recovery of salmon is likely to often require more than simply removing dams—we may also need improvements in spawning habitats and in ocean and estuarine environments, further reductions in harvest, and perhaps major modifications of hatchery programs. Finally, it is worth pointing out that although removal of dams is often championed in terms of ‘‘returning to a normative river’’ (ISG, 1999), most proposed dam removals still leave many dams untouched and our rivers so altered by other factors that they remain a long way away from the ideal of a free-flowing ‘‘wild river.’’

E. Natural Environmental Variability and Climate Change
Although the tendency in recent years has been for resource managers to focus on the human-caused fac-
tors that have contributed to the decline of salmon, natural variation in environmental conditions, both freshwater and marine, are also major factors affecting the abundance and distribution of salmon. Operating on seasonal, annual, and decade-long timescales, events such as regime shifts, El Ninos, and changes in seasonal patterns of coastal circulation (particularly upwelling) all can have a profound effect on oceanic survival of salmon. Particularly noteworthy has been the research documenting correlations among broad patterns of productivity, and major physical features of the ocean-atmosphere interaction. For example, Hare et al. (1999) found that harvest (and presumably abundance) of Pacific salmon in Alaska has varied inversely with catches along the U.S. West Coast over the past 70 years. In exploring various explanations for this observation, they found a striking correlation between switching pattern of north-south abundance and the Pacific Decadal Oscillation (PDO; a recurring pattern of pan-Pacific atmosphere-ocean variability). While there are several ocean-atmosphere indices such as the PDO, they all, in concept, attempt to capture the key physical forces that drive decadal-scale basin-wide climate variability. For example, since the last major climate shift of 1976–1977 in the North Pacific Ocean, oceanographers have documented decreased mixed-layer depths in the Subarctic Gyre and increased sea surface temperatures along the west coast. The biological response to these physical changes were profound, including major changes in primary and secondary production, and micro- and mesoscale changes in zooplankton biomass and species composition. In addition to salmon, a large number of higher trophic level fish, birds, and mammals show similarly dramatic shifts in abundance and distribution coincident with these changes in ocean conditions.

Operating on an even grander scale, the impacts of long-term climate change will undoubtedly have profound consequences for all animal and plant species. Salmon, because of their anadromous nature, will suffer the dubious distinction of being subject to impacts in both the marine and freshwater habitats—a distinction that may well challenge their ability to adapt and survive over much of their current range. Of the many studies designed to predict the consequences of climate change on aquatic species, one of the more dramatic is the recent work of Welch et al. (1998). To explore potential marine impacts on salmon, Welch and colleagues compiled an extensive data set on distribution of salmon at sea and found that, for sockeye salmon, marine distribution was limited by sharp thermal boundaries. Using a climate model, and assuming a doubling of greenhouse gases over the next 50 years (which is consistent with the current rate of input), it was predicted that sockeye salmon would be excluded from most of the Pacific Ocean and restricted to a relatively small area of the Subarctic Pacific for much of the year. Similar analyses revealed marked reduced ocean habitat for other species of Pacific salmon as well. Climate change is also expected to depress oceanic productivity by disrupting upwelling patterns in a way that reduces the delivery of nutrients (Bakun, 1990). To the extent that salmonid population fluctuations are determined by bottom-up forces, the possibility of reduced nutrient upwelling could have dire consequences for these and many other fish species.

VI. WHAT IS THE VALUE OF SALMONID BIODIVERSITY AND CAN WE RESCUE SALMON FROM EXTINCTION?

Two arguments dominate the mainstream ecological literature regarding the “value of biodiversity”: (a) diversity fosters enhanced productivity by allowing a more complete exploitation of the environment via a variety of different forms, each adapted to different habitats or resources, and (b) diversity fosters enhanced stability by spreading the risk and providing redundancy in the face of unpredictable catastrophes. Both hypotheses regarding the value of biodiversity apply well to salmon. First, different salmonid ESUs spawn in different portions of the freshwater environment (inland versus coastal, mainstream rivers versus streams, lakes versus rivers, and so forth) and migrate to different regions of the ocean. Any loss of salmonid diversity would mean that the productivity of these ecosystems would be reduced, with minimal likelihood that any other species could take the place of salmon (since the anadromous lifestyle is so unique). Second, due to climatic fluctuations, different months of the year will be favorable for migration; in fact, recent data indicate that even a shift of a few days with respect to the time at which juvenile chinook salmon arrive at the ocean can alter survival rates by a factor of two. If salmon diversity is reduced, then there will be less variety in run timing and life history schedule and consequently much less insurance against the vagaries of climatic fluctuations. The precipitous decline of Pacific and Atlantic salmon throughout much of their range raises a number of important questions: Can we save the salmon? Should we save the salmon? And what do we need to do to save the salmon?
A. Can We Save the Salmon?

The answer to this question is an unequivocal yes. A species complex that has survived and flourished under the wide range of conditions from which salmon have evolved has almost certainly weathered numerous survival bottlenecks in the past, caused by environmental perturbations ranging from volcanic eruptions to the total loss of huge blocks of habitat to glaciers, fire, and floods. Perhaps the key attribute that has allowed the salmonids to survive these events is the tremendous diversity of life histories and patterns of habitat use. The salmonids represent a plastic species complex with a robust evolutionary potential that should facilitate survival under diverse conditions.

That said, many populations in the Pacific Northwest and Northeast North America are already extinct, and many others have never been at lower levels of abundance. Saving Pacific and Atlantic salmon will clearly require immediate action that will include economic sacrifice, societal discipline, and a commitment to a science-based recovery strategy—all at a very significant cost to the people of the affected regions.

B. Should We Save the Salmon?

Because of the societal implications, this question perhaps extends well beyond the boundaries of what should be the subject of this volume. However, from strictly an ecological perspective, the salmonids are a vital component of the North Pacific and North Atlantic ecosystems and the associated coastal habitats. The ecological importance of salmonids as predators, prey, and nutrient recyclers is obvious to even the most casual observers of nature. While as scientists we can marvel at the complexity of an ecosystem and strive to understand species and environmental interactions in the broadest sense, we must also acknowledge that such complex and dynamic systems will never be fully understood. Nonetheless, the loss of such a key species complex as salmonids over such a broad geographic area would no doubt be a biological event without precedent.

C. What Do We Need to Do to Save the Salmon?

As noted earlier, the keys to solving the salmon problem are societal commitment, discipline, and a science-based recovery strategy. While societal commitment and discipline are not scientific issues, they certainly are not devoid of science. To make a decision on whether and how to proceed, society must have knowledge—knowledge of the complexity of the problem and the certainty (or lack thereof) associated with various courses of action. This knowledge is vital to the economic decision-making process.

In contrast, the design of recovery strategies is largely a science issue. At the core of every science-based recovery strategy should be an analytical framework that allows (a) the methodical determination of survivorship in each of the life history stanzas of salmon, (b) the identification of actions and opportunities to increase survivorship, (c) the assessment of biological (and societal) feasibility of each action or suite of actions, and (d) a rigorous monitoring and evaluation plan. Whatever models are used should be clear and understandable to the average citizen, and the data used as model inputs must be available to all.

Clearly, society cannot continue the extraction-based economy of the past 200 years and expect even remnant populations of salmon to exist over even a fraction of their former range. A major shift in environmental decision making will be required. In particular, we must abandon the habits of the past two centuries, which have involved almost universally approaching conservation through a “reduction from the status quo” perspective. For example, rather than arguing about whether we should reduce harvest by 10, 20, or 50% from current levels, any harvest above 0% should be viewed with great skepticism. Likewise, the issue should be whether any timber harvest is allowed in salmon watersheds, not whether the current acreage of timber harvest should be reduced. When populations face a high probability of extinction, modest adjustments of the status quo are not enough. Unless society seriously acts to rebuild salmon populations, the future of salmon is not assured, and moderate actions ought to be recognized as optimistic gambles (as opposed to reasonable moderation).

See Also the Following Articles

AQUACULTURE • ENVIRONMENTAL IMPACT, CONCEPT AND MEASUREMENT OF • FISH, BIODIVERSITY OF • MIGRATION • RESOURCE EXPLOITATION, FISHERIES

Bibliography


I. Biodiversity as a Measurable Quantity

A. Scalable Quantities

Biodiversity, no matter how it is measured, is a quantity with units on one of several types of measurement scale.

IT IS A FAMILIAR experience that taxonomic diversity increases as more area is examined, as more time is spent watching an area, as more organisms are collected, and as a wider range of body sizes are accumulated. These increases in diversity do not scale directly with area, time, collection size, or body size. More than 50 years ago, C. B. Williams found that, on average, a 10-fold increase in area will increase the number of plant species by a factor of 2 rather than 10. Williams ascribed divergence from a 1:1 scaling to the vagaries of sampling (at the scale of plots), to habitat diversity (at the scale of biogeographic regions), and to evolutionary history (at the scale of continents). Subsequent work has confirmed this empirical scaling, developed a wealth of empirical scalings by area, developed alternatives to scaling by area, and tested theoretical explanations for these scaling relations. Biodiversity now includes genetic and habitat diversity in addition to taxonomic diversity. This article presents the quantitative basis for scaling of biodiversity based on rapid development of scaling theory in ecology in the past decade.
Consequently, some care is needed in defining the measure of diversity at hand to obtain accurate scaling relations with area or time.

A well-defined measure of biodiversity has five parts: (i) a name, (ii) a symbol, (iii) a procedural statement that sets forth the method and conditions for measurement, (iv) a set of numbers generated by that procedural statement, and (v) units on one of several types of measurement scales. A scaled quantity is conveniently represented as a symbol equated to a set of numbers arranged inside brackets, multiplied by the unit of measurement. An example is the number of cichlid fish species found in each of six African lakes, as reported by Ricklefs and Schluter (1993):

<table>
<thead>
<tr>
<th>Procedural statement</th>
<th>Name</th>
<th>Symbol</th>
<th>Numbers</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ricklefs and Schluter (1993)</td>
<td>Species</td>
<td>( S_j )</td>
<td>200</td>
<td>( \times ) lake (^{-1} )</td>
</tr>
<tr>
<td>(p 358)</td>
<td>richness</td>
<td>( s_i )</td>
<td>220</td>
<td></td>
</tr>
</tbody>
</table>

In this example the lakes are listed in order from large (Lake Victoria = 69,486 km\(^2\)) to small (Lake Edward = 2130 km\(^2\)).

Measurements occur on one of one types of scale: nominal (yes/no); ordinal (ranks); interval (arbitrary zero points such as longitude in degrees), and ratio (true zero such as distance from the equator in kilometers). The mathematical rules for working with a ratio scale quantity (such as number of species per island) differ from those for working with quantities on other types of scales (such as ranking of lake area from large to small in the previous example). If our interest is in scale and its effects, then we need to work with the species-abundance distribution (on a ratio type of scale) rather than the rank-abundance distribution (on a rank type of scale).

B. Taxonomic Diversity Distributions

Diversity studies typically employ three basic quantities: \( N \) is the number of organisms, \( A \) is the area, and \( T \) is time or duration of measurement. A collection of \( N \) organisms from a unit area \( A \), during a unit period \( T \), is sorted into \( S \) groups (typically species) as shown in Fig. 1. Any taxonomic level can be used in sorting, but for the sake of clarity species will be used to illustrate taxonomic diversity distributions. The number of organisms \( n_i \) is recorded for each species. A familiar characterization of this taxonomic diversity is the curve of abundance across species, ranked from high (largest \( n_i \)) to low (smallest \( n_i \), usually one organism per species). The information in the rank-abundance curve can be re-expressed (see Fig. 1) as a frequency distribution \( s(n_i = n) \). The symbol \( s(n_i = n) \) represents the number of species at abundances ranging from \( n_i = 1 \) upward to the abundance \( n_i \) of the most common species in the collection. This frequency distribution shows, in full detail, the information on species diversity in a single collection of size \( S \), taken from unit area \( A \), in unit time \( T \). If our interest is in the effects of scaling, then we will need to work with the frequency distribution \( s(n_i = n) \), which is on a ratio type of scale, rather than with the rank abundance curve (Fig. 1), which is on an ordinal type of scale.

Following practice in the literature on diversity, we can reduce the information in the species abundance distribution \( s(n_i = n) \) to an index with a single value. The most common index is the total number of species \( s \), obtained by counting classes or equivalently by summing the distribution \( s(n_i = n) \) across abundance classes \( n_i \). We might also compute Shannon Weaver diversity or the Simpson index for the collection. Both go beyond gross measures of diversity such as species number \( s \), but fail short of describing all the information in the species diversity distribution \( s(n_i = n) \) for the collection \( N \).

If we obtain a second collection, again from an area \( A_i \), during a period \( T_i \), we expect a somewhat different number of species \( n_i = 2 \), a new set of species \( s_i \) (many the same as in the first collection), and a new diversity distribution \( s(n_i = n) \). With several collections \( s(n_i) \), now labeled \( j = 1 \) through \( J \), we have a fuller characterization of diversity. We can construct the combined distribution \( S(n_i = n) \) across some number of collections, \( J \). The total number of species \( S \) across the combined collection will exceed the number of species \( s \) in any one collection.

If the collections were always drawn by the same procedural statement, they can be viewed as samples from a set of all possible collections having unit area \( A \), and duration \( T \). The diversity distribution for this hypothetical set is \( S(n_i = n) \). This is the species abundance distribution for the community, provided all collection sites have an equal (or at least known) chance of appearing in the sample. The total number of species \( S \) and total number of organisms \( N \) are computed respectively as \( S = \sum_s(n_i = n) \) and \( N = \sum_s n_i \). Integral signs \((\int)\) replace the summation signs for distributions based on continuous values of population size \( n_i \) (May,
SCALE, CONCEPT AND EFFECTS OF

1975). Following May (1975) we can compare an observed distribution \( s(n_i = n_k) \) to a mathematical function, allowing us to compute species numbers at unobserved values of abundance \( n_k \). An impressive number of methods (Bunge and Fitzpatrick, 1993) have been devised to estimate \( S \) from several collections \( s(n_i = n_k) \), taken to be samples.

When making statistical estimates from taxonomic distributions, it often proves convenient to use the cumulative distribution \( s(n_i \leq n_k) \) rather than the distribution \( s(n_i = n_k) \). The cumulative distribution rises monotonically from left to right, as it records the number of species with \( n_k \) or fewer organisms per species, rather than recording the number of species at each value of \( n_k \). Both distributions express exactly the same information. The cumulative distribution is less lumpy and therefore easier to use.

C. Other Diversity Distributions

Genetic variability within a species can be characterized in the same fashion as taxonomic diversity. For a collection of \( N \) individuals the usual single-locus measure is the proportional presence of each type of allele \( q_i = n_i/2N \). This information can be expressed as a rank-abundance distribution of allelic proportions from common \( (q_i \text{ large}) \) to rare \( (q_i \text{ small}) \). For convenience, the information in this distribution is reduced to an index.
such as the number of classes (alleles) or expected heterozygosity \( H_e = 1 - 2q^2 \). The information in several collections can be expressed as a rank-abundance curve for each collection and a rank-abundance curve for the combined collection. This information can in turn be reduced to an index, such as \( F_{\alpha} \), which compares expected heterozygosity for the entire collection \( H_e \) to the average over \( k \) collections \( H_e = k^{-1} \sum H_e \). This approach, applied to genetic diversity at the single-locus level, can be extended to coarser classifications, such as recognizable genotypes in a population (Mallet as cited in Gaston, 1996, p. 16). If we are interested in the effects of scale on genetic diversity, then the logic of working with ratio scale quantities will compel us to take the (highly unusual) step of re-expressing the rank-abundance curve as the genetic diversity distribution curve for the combined collection. This information can be summarized in any given study. Starting with Whittaker, the diversity did not come into wide use. In practice, differentiation diversity was one individual \( (n_1 = 1) \), two individuals \( (n_2 = 2) \), etc. This can be converted to a cumulative frequency distribution \( q(n_1 \leq n) \) to again smooth away the inconvenient lumpiness of the distribution \( q(n_1 = n_2) \).

We can construct a diversity distribution to summarize habitat diversity. This requires defining a collection area \( A \) divided into \( N \) units of size \( A_n \), which are sorted into habitat classes. From this we can construct the usual rank-abundance curve, running from the most common habitat \( (n_1 \text{ large}) \) to rarest \( (n_1 \text{ small}) \). The same information can be plotted on a ratio scale as a habitat diversity distribution \( h(n_1 = n_2) \), expressing the frequency of habitats found in one unit \( (n_1 = 1) \), in two units \( (n_1 = 2) \), etc.

D. Alpha, Beta, and Gamma Diversity

In 1960, R. H. Whittaker distinguished taxonomic diversity within a habitat (alpha diversity) from diversity within larger geographic units composed of several habitats (gamma diversity). Examples of these larger units include islands, lakes, landscapes, and coastal embayments. Whittaker defined beta diversity as the among-habitat diversity, or diversity change going from a single habitat to the larger geographic unit. Later (Box 1) Whittaker added two more spatial scales or "levels"—that of point diversity within a habitat and that of regional (epsilon) diversity for broader geographic units containing smaller units, each characterized by its own gamma diversity. Whittaker presented these as convenient labels, recognizing that point, alpha, gamma, and epsilon diversity were arbitrary points on a continuum, and that the size of these units might be defined differently in any given study. Starting with Whittaker, the most effective use of these labels has been in distinguishing diversity patterns at the scale of habitats from patterns at larger scale. The labels are also useful in discussing the very different processes responsible for changes in diversity at habitat, regional, and global scales. Alpha, beta, and gamma diversity have remained in wide use during the past 40 years, point and epsilon diversity did not come in to wide use. In practice, gamma diversity is sometimes dropped in favor of two

<table>
<thead>
<tr>
<th>Box 1</th>
<th>Alpha, Beta, and Gamma Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>In an evolutionary context, Whittaker (1977) defined inventory diversity at four spatial scales or levels: the point sample, the habitat, the landscape, and the region. Differentiation diversity is defined as the change in diversity between these four levels.</td>
<td></td>
</tr>
<tr>
<td><strong>Point diversity</strong> The number of species for a small or microhabitat sample within a community regarded as homogeneous. Also called internal alpha or subsample diversity.</td>
<td></td>
</tr>
<tr>
<td><strong>Beta diversity.</strong> Change going from one point to another within a habitat. Also called internal beta diversity.</td>
<td></td>
</tr>
<tr>
<td><strong>Within-habitat or alpha (( \alpha )) diversity The number of species in a sample representing a community regarded as homogeneous (despite its internal pattern). It measures the number of potentially interacting species.</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Pattern diversity</strong> Change along a climatic gradient or among the different communities of a landscape. Developed by Whittaker (1960) as a measure of packing of competing species along a gradient, it was defined as the ratio of regional (gamma) diversity to average within-habitat (alpha) diversity. Wilson and Shmida (1984#) review other measures of beta diversity.</td>
<td></td>
</tr>
<tr>
<td><strong>Geographic differentiation or delta (( \delta )) diversity</strong> Change along a climatic gradient or between geographic areas.</td>
<td></td>
</tr>
<tr>
<td><strong>Regional or epsilon (( \epsilon )) diversity</strong> The number of species in a broad geographic area including differing landscapes.</td>
<td></td>
</tr>
</tbody>
</table>
labels—within-habitat (alpha) and among-habitat (beta) diversity.

It is of note that similar sets of arbitrary labels to distinguish pattern and process over a range of spatial scales have been developed in both marine (Haury et al., 1978; Box 2) and terrestrial ecology (Delcourt and Delcourt, 1988; Box 2). Habitat features persist longer in terrestrial than marine systems, and hence the time-scales associated with any given spatial scale are greater for terrestrial than marine systems (Steele, 1991; Box 2).

II. CONCEPT OF SCALE

A. Multiscale Analysis

Multiscale questions naturally arise for biodiversity: How does diversity change with collection size? How does it change with duration of collection? These questions can be addressed quantitatively by examining the change in a biotic diversity distribution with change in collection size $N$, in collection area $A$, and in collection duration $T$. For habitat diversity similar questions of scale arise: How does the habitat diversity distribution $h(n = n_i)$ change when the collection area $A$ is doubled? As with taxonomic diversity, we expect genetic and habitat diversity to be greater in large collections than in small collections, greater in widely separated samples than in closely placed samples, and greater in large areas than in small areas. We do not expect that a doubling in area will simply double the number of habitats or the number of alleles.

There are several ways to compare diversity at large and small scales. One of the more venerable is to rate (or plot) diversity relative to island size (Fig. 2). Rating diversity against area is equivalent to rating against collection size if organism density is the same on large and small islands. Another venerable method is to make a series of collections and then plot a measure of diversity against effort, measured as the accumulated number of collections (Fig. 2). These "collector's curves" are equivalent to plots against accumulated collection size $N$ if organism catch per unit effort remains constant across collections. Figure 3 shows the increase in diatom species number with increase in effort (number of slides). A third method, currently popular but less venerable, is to fix the unit size, measure the separation between units (in time or space), and then plot some measure of difference in diversity between units as a function of separation or lag (Fig. 2). An additional method is to construct a diversity distribution from a large area and then take the average diversity distribution for each half of the area, each quarter, etc. This method of coarse graining (Fig. 2) is better known in...
geophysics (Rodríguez-Iturbe and Rinaldo, 1997; Bar- enhart, 1996) than in biology. Coarse graining expresses the same information as lagging, an equivalence that goes unrecognized. Both lagging and coarse graining assume that, on average, organism number N does not depend on distance separating samples, and hence diversity differences are due to spatial scale (separation) rather than spatial gradients in organism number N.

These various methods of multiscale comparison express the same information in principle but not in practice. Lagging and coarse graining (in space) express the same information if the spatial layout is the same. In practice, somewhat different information is obtained because coarse graining is applied to contiguous units, whereas lagging is applied to separated units. Collector’s curves introduce time-scales. In a seasonal environment, collections made monthly for a year will match change in diversity relative to change in spatial scope if these are distributed over an area. Accruing units will increase the scope if these are distributed over an area. Accruing can be by adjacency (along a transect or concentric rings around a point) or by random sampling within a larger area (in statistical jargon, within the frame of all possible units). With accrual, the scope is $A_{new}/A_{min}$. The scope is the ratio of the largest to smallest spatial unit, $A_{new}/A_{min}$. For the lake example, the spatial scope is $(69,484 \text{ km}^2)/(2,150 \text{ km}^2) = 32$. Accruing units will increase spatial scope if these are distributed over an area. Accruing can be by adjacency (along a transect or concentric rings around a point) or by random sampling within a larger area (in statistical jargon, within the frame of all possible units). With accrual, the scope is $A_{new}/A_{min}$. For the lake example, the spatial scope is $(69,484 \text{ km}^2)/(2,150 \text{ km}^2) = 32$. Accruing units will increase spatial scope if these are distributed over an area. Accruing can be by adjacency (along a transect or concentric rings around a point) or by random sampling within a larger area (in statistical jargon, within the frame of all possible units). With accrual, the scope is $A_{new}/A_{min}$.

In comparing diversity within or among studies we would like to have a measure that does not depend on choice of units. To accomplish this, the scope of a quantity is defined as the ratio of its maximum to minimum value. For the cichlid example, the scope in species number is $(200 \text{ species lake}^{-1})/(7 \text{ species lake}^{-1}) = 29$. For non-cichlid species the scope in number per lake is different: $(111 \text{ species lake}^{-1})/(17 \text{ species lake}^{-1}) = 6$. The scope of a quantity has units that cancel and hence are expressed as dimensionless numbers, independent of units of mass, length, or time. However, the units that form the dimensionless ratio should be kept in view to avoid drawing conclusions outside the scope of measurement.

Care is required in comparing quantities with scopes that differ procedurally. One can increase the scope of collection size by increasing area from unit area $A_s$ to total area $A_{tot}$. The scope is $\Sigma A/A_s$. This generates accumulation curves relative to area: accumulation of organisms N(A), of species S(A), of genotypes Q(A), and of habitats H(A). One can increase the scope of collection size by repetitive collection at one location, resulting in accumulation curves relative to time: N(T), S(T), Q(T), and H(T). There is no reason to expect that change in diversity relative to change in spatial scope will match change in diversity relative to change in temporal scope.

Spatial scope can be expanded in several ways. Searching out a wide range of isolated units—lands, in an ocean and lakes on a landscape—will increase the scope. With this method the scope is the ratio of the largest to smallest spatial unit, $A_{new}/A_{min}$. For the lake example, the spatial scope is $(69,484 \text{ km}^2)/(2,150 \text{ km}^2) = 32$. Accruing units will increase spatial scope if these are distributed over an area. Accruing can be by adjacency (along a transect or concentric rings around a point) or by random sampling within a larger area (in statistical jargon, within the frame of all possible units). With accrual, the scope is $A_{new}/A_{min}$. For the lake example, the spatial scope is $(69,484 \text{ km}^2)/(2,150 \text{ km}^2) = 32$. Accruing units will increase spatial scope if these are distributed over an area. Accruing can be by adjacency (along a transect or concentric rings around a point) or by random sampling within a larger area (in statistical jargon, within the frame of all possible units). With accrual, the scope is $A_{new}/A_{min}$. For the lake example, the spatial scope is $(69,484 \text{ km}^2)/(2,150 \text{ km}^2) = 32$. Accruing units will increase spatial scope if these are distributed over an area. Accruing can be by adjacency (along a transect or concentric rings around a point) or by random sampling within a larger area (in statistical jargon, within the frame of all possible units). With accrual, the scope is $A_{new}/A_{min}$.
FIGURE 3  Cumulative diversity distributions from areas of increasing size (one, two, three, and four slides). Note the linear relation (on a logarithmic scale) between species number \( S \) and slide area. However, species number \( S \) represents only part of the information in the cumulative diversity distributions from areas of increasing size. Data from Patrick (1968).

\[ \sum T = \text{number of units.} \] Temporal scope also increases by fine graining down to the smallest time units \( T \) within a continuous period \( T \). The scope is \( T/T \).

C. Complete and Incomplete Similarity

The classical solution to the problem of change in scale has been to define two systems (such as a model boat and full-scale prototype), define groups of similar measurement units (typically mass, length, and time), and then work out the scaling from model to prototype with reference to groups of similar units called dimensions [mass \( M \), time \( T \), Euclidean lengths \( L \), areas \( L^2 \), and volumes \( L^3 \)]. If the model and prototype are completely similar (Barenblatt, 1996) then we expect that a 10-fold increase in length of prototype relative to model will result in a \( 10^3 \)-fold increase in volume. We also expect a 1000-fold increase in mass if model and prototype are constructed of the same material and have the same density. The scaling relation is mass \( \propto \) volume \( \propto \) length\(^3\). Dimensional analysis, based on the principle of complete similarity, has been notably successful in engineering and several areas of physics, including geophysical fluid dynamics. It has been notably unsuccessful with natural rather than engineered objects, or with dynamics that “lurch” rather than ticking along in a clock-like fashion.

In ecology, the failure of complete similarity was addressed by comparative methods of “hierarchy theory,” which recognizes that large-scale spatial structure or dynamics cannot be computed directly from local scale structure or rates. Using Herbert Simon’s concept of “nearly decomposable levels,” the hierarchical method describes structure or computes dynamics at two (or more) levels and then compares the results. For example, in the biodiversity literature it is customary to distinguish local or within-habitat diversity from regional (among-habitat) diversity. One widely recognized finding (Holt as cited in Ricklefs and Schluter, 1993, p. 77) is that the relation of species number to area at the habitat level (i.e., within a habitat) does not
match that at the ecosystem level (i.e., among larger areas having multiple habitats).

When completely similar scaling fails, a property (e.g., coastline length) at a large scale can often be computed from the average property at a smaller scale by a power law expressing incomplete similarity (Barenblatt, 1996). The most familiar example of incomplete similarity is that of a fractal relation between coastline length and Euclidean length $L$. If we take large steps along a coastline and then count these as if on a straight line, we obtain a coastline perimeter of length $P(L_{\text{large}})$. If we take small steps and then count these as if in a line, we obtain a length $P(L_{\text{small}})$ that exceeds $P(L_{\text{large}})$.

The relation of the two measures of coastline length is

\[ \frac{P(L_{\text{large}})}{P(L_{\text{small}})} = \left( \frac{L_{\text{large}}}{L_{\text{small}}} \right)^\alpha, \]

In briefer form, $P(L) = c \times L^\alpha$. In even briefer form, $P(L) = L^\alpha$. The exponent $\alpha$ is not an integer, as it would be for completely similar Euclidean objects. Mandelbrot (1977) coined the term "fractal" to describe non-Euclidean objects where $\alpha$ is a non-integral constant.

More generally, incomplete similarity relates some property $S(x)$ to some measure $x$ according to a power law:

\[ S(x) = c \times x^\alpha \]

This relation states that when $x$ is rescaled (e.g., by a factor of 2), the property $S(x)$ is still proportional to $x^\alpha$. In the example of African lakes, how does species number scale with area? A rough scaling is obtained by applying the principle of homogeneity of scope (Schneider, 1994), which requires that all terms in an equation have the same scope. An exponent is applied to equate the scope of diversity with that of lake area:

\[ \frac{S(2A)}{S(A)} = \left( \frac{2A}{A} \right)^\alpha = 2^\alpha \]

\[ \frac{200}{40} = \left( \frac{59,484}{2150} \right)^\alpha \]

Solving for the exponent, we obtain the scaling for incomplete similarity

\[ \alpha = \log(200/40)/\log(69484/2150) = 0.463 \]

This is a rough estimate that could be improved by using regression to include more information in the scaling. Based on the rough scaling, a doubling in lake area from, for example, 2130 km$^2$ to 4300 km$^2$ is expected to amplify cichlid species number by $2^{0.463} = 1.4$ rather than by a factor of 2.

Incomplete similarity applies to cumulative frequency distributions (Fig. 3) as well as to single indices such as species numbers. The scaling based on incomplete similarity is

\[ P[X \geq x] = c \times x^\alpha \]

where $x$ is again some measure of interest, such as length, area, or time. For some property (such as species per unit area) the probability distribution relative to the measure is

\[ S(x) = c \times x^{-\alpha} \]

The exponent that quantifies incomplete similarity can be estimated by selecting isolated units, as in the cichlid example. The exponent can be estimated via autocorrelation techniques, which measure correlation of a quantity with itself at increasingly large spatial separations. Scaling exponents can be estimated by fine graining (Milne, 1997) or equivalently by coarse graining (Fig. 2). Estimates via fine or coarse graining are equivalent to those from lagging (Fig. 2), a fact that is well recognized for time series analysis but largely overlooked in the literature on spatial structure. Fine graining (or equivalently, coarse graining) has been used to quantify the fractal scaling exponent for habitats as diverse as bare soil, leaf perimeter, seaweed, and coral heads; it has been used to quantify coastline nesting habitat of eagles, pine nesting habitat of woodpeckers, and foraging habitat of rabbits (Milne, 1997).

The dynamics of systems that are incompletely similar at small and large scales are also examined via renormalization and coarse graining (Barenblatt, 1996). Renormalization has been used to investigate the dynamics of landscapes (Rodriguez-Iturbe and Rinaldo, 1997) and population interaction (Levin and Pacala as cited in Tilman and Kareiva, 1997, p. 271). Renormalization is an appropriate technique for examining the dynamics of biodiversity, which do not proceed by evenly paced transitions, but instead jump and lurch by episodes of invasion, local extinction, speciation, and anthropogenic extinction.

Incomplete similarity provides an alternative to hier-
architectural (usually two-level) analysis of structure and dynamics in ecology. Currently, the hierarchical approach is more familiar and remains the dominant approach when analyzing ecological interactions (Tilman and Kareiva, 1997) or biodiversity patterns at multiple scales (Ricklefs and Schluter, 1993; Rosenzweig, 1995).

D. Scaling Relations for Biodiversity

The development of scaling relations for biodiversity has been vigorous for taxonomic diversity. It has just begun for habitat diversity and does not yet exist for genetic diversity. For taxonomic diversity, the history of theoretical development can be summarized briefly as a sequence of scaling relations:

1. $S \approx A^{3/4}$
   Species number scales incompletely with area. This is the oldest and best known scaling relation in ecology, with a rich history (Ricklefs and Schluter, 1993; Rosenzweig, 1995). The scaling coefficient was recognized to be less than 1 for most situations, but values were completely empirical and could not be generalized beyond the data used to estimate the coefficient.

2. $S \approx N^{0.6}$
   Species number scales incompletely with numbers according to a $3/4$ power law. This relation, due to Preston, was discussed in detail by May (1975), who provided a list of models. It is worth noting that the canonical value ($3/4$) depends on the assumption that the taxonomic diversity distribution has a lognormal form.

3. $S \approx N^{0.5} \approx A^{1}$
   Species number scales with total numbers and hence with area according to a $3/4$ power law. Preston's canonical theory and MacArthur and Wilson's theory for isolated communities attribute the scaling to equilibrium conditions (May, 1975), but it is now known that scaling laws such as this can arise in non-equilibrium conditions (Barenblatt, 1996). This scaling assumes that organism number $N$ scales in a $1:1$ fashion with area.

4. $N \approx A^{1}$
   Numbers scale completely with area. This scaling will always hold for coarse graining—we can always compute the average numbers in smaller areas from numbers in larger areas. This scaling will hold on average for accumulation if we take multiple sequences of organism counts. However, this scaling is usually untrue for any one sequence of accruals because of patchiness. Often, the rate of accrual will appear to accelerate because often the densest site will be sampled later rather than early during accrual. Complete scaling of numbers $N$ with area $A$ is easier to assume than demonstrate for islands or other isolated units.

5. $S \approx A^{0}$, where $\beta_{\text{island}} > \beta_{\text{continent}}$
   The scaling exponent is steeper for islands (0.25 < $\beta_{\text{island}}$ < 0.33) than for isolated blocks of land on continents (0.13 < $\beta_{\text{continent}}$ < 0.18). Rescue effects, where small populations are maintained by frequent migration from surrounding areas on continents, readily explain the shallow scaling for continental communities (Rosenzweig, 1995). Milne (1997) describes a correction factor to account for scaling effects in comparing the area of Euclidean blocks to the area of islands with fractal shapes.

6. $S \approx A^{0}$, where $\beta_{\text{island}} > \beta_{\text{continent}}$
   Biogeographic provinces, isolated at evolutionary timescales (Rosenzweig, 1995), have scaling exponents that exceed blocks within a province. Estimates of $\beta_{\text{island}}$ fall closer to unity (Rosenzweig, 1995) than to typical values for $\beta_{\text{continent}}$.

7. $S(n_{i} \leq n) \equiv S(A_{i} \leq A) \equiv H(A_{i} \leq A)$
   Species diversity, as a distribution, scales with habitat diversity, again as a distribution. Substantial qualitative support exists as correlations of species richness with several habitat variables (Wright et al. as cited in Ricklefs and Schluter, 1993, p. 73; Rosenzweig, 1995), including fractal measures of habitat complexity (Milne, 1997). An analytic review, similar to that of May (1975), is needed to integrate the taxonomic diversity literature with the literature on the structure and dynamics of landscapes (Rodriguez-Iiturbe and Rinaldo, 1997).

III. EFFECTS OF SCALE

A. On Estimates of Diversity

Estimates of diversity and of rate of extinction are required at both local and global scales, but measurements are restricted to small scales. The principal effect of incomplete similarity on estimates of diversity and extinction rate is that it precludes intuitive scalings or extrapolations. If species number, or any other index of taxonomic diversity, scales completely with area or with number of organisms, then scaling could be handled with intuition-based 1:1 scalings. A survey could be used to estimate diversity in a larger area by knowing only the area of interest relative to the area surveyed to make the extrapolation. Extinction rates could be estimated over large areas by knowing only what percentage of the area had been examined directly. Complete similarity of taxo-
nomic diversity with any other quantity has yet to appear in the literature, which is vast and growing (Gaston, 1996; Ricklefs and Schluter, 1993).

Incomplete similarity poses little problem if the scaling exponent is used to interpolate within the scope at which it was estimated. Extrapolation to larger scales poses a problem because of uncertainty in choice of exponents. Extrapolation becomes possible if the scaling exponent increases in a predictable way with increase in the degree of isolation (Rosenzweig, 1995). Choosing an incomplete scaling exponent, however, remains fraught with uncertainty.

Problems of non-intuitive (incomplete) scaling and uncertainty in choice of scaling exponent attend any effort to estimate change in taxonomic diversity. For habitat diversity, the problems are less acute because remote sensing can provide maps at high resolution over large areas. Consequently, scaling exponents can be estimated over a scope wide enough to be useful, from large areas of interest down to the small scale of plots and quadrats amenable to experiment and direct measurement of population and community dynamics. The number of variables that can be sensed remotely is limited. Hence, there is need for scaling relations that connect remotely sensed variables to population variables, such as density, biomass, production, movement, recruitment, and mortality. There is also a need to connect remotely sensed variables to community variables such as extinction rate.

B. On Conservation Planning

Conservation planning is carried out at far larger scales than measurements of the underlying dynamics. The problem is inescapable but not solvable by intuitive modes of thinking based on complete similarity. One cannot execute a field study, and apply it with certainty to conservation issues at larger scales, using the intuitive notion that rates (e.g., per capita mortality) are independent of spatial scale (i.e., scale as $4^3$). Nor can one proceed on the intuitive notion that a variable will scale completely with area on a 1:1 relation. The idea of incomplete similarity has already started to take its place in conservation planning. An example is the "single large vs several small" debate over refuge size—a debate that recognizes that doubling area does not leave rates unchanged nor does it double quantities such as the number of species. Several other examples are provided by Milne (1997). The next step will be to develop incomplete scalings that are reliable and based on biologically based theory rather than empirical coefficients. One particularly exciting topic, which may soon yield to theoretical development, is the scaling of taxonomic diversity to habitat diversity. The basis in biological theory already exists. Species lists increase by invasion or speciation, for which adaptation to habitat via natural selection is important. Species lists shrink by local or global extinction, for which change in habitat is important. There is thus the biological basis for quantitative scaling of taxonomic diversity with habitat diversity. It is a promising topic for theoretical development, for which reliable computations are needed to address urgent questions of local and planet scale change in biodiversity.

See Also the Following Article

MEASUREMENT AND ANALYSIS OF BIODIVERSITY

Bibliography


Rosenzweig, M. L. (1993). The scaling of taxonomic diversity to habitat diversity. The basis in biological theory already exists. Species lists increase by invasion or speciation, for which adaptation to habitat via natural selection is important. Species lists shrink by local or global extinction, for which change in habitat is important. There is thus the biological basis for quantitative scaling of taxonomic diversity with habitat diversity. It is a promising topic for theoretical development, for which reliable computations are needed to address urgent questions of local and planet scale change in biodiversity.


I. The Origin and Evolution of Seagrasses

Seagrass originated early in the evolution of angiosperms, approximately 100 million years ago, from ancestors that have been hypothesized to be freshwater plants or mangrove- or salt-marsh-like plants. The origin of seagrasses appears to be polyphyletic, and recent analyses based on molecular phylogeny have supported the development of seagrass species from three clades of primitive plants (Les et al., 1997). There are two main
TABLE I

<table>
<thead>
<tr>
<th>Species</th>
<th>Biogeographic membership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amphibolis antarctica</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Amphibolis griffithii</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Cymodocea angustata</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Cymodocea nodosa</td>
<td>Mediterranean flora</td>
</tr>
<tr>
<td>Cymodocea rotundata</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Elphidium acutum</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Halodule jassoides</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Halodule uninervis</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Halodule wrightii</td>
<td>Caribbean flora</td>
</tr>
<tr>
<td>Halophila baillonis</td>
<td>Caribbean flora</td>
</tr>
<tr>
<td>Halophila beccarii</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Halophila capricornii</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Halophila decipiens</td>
<td>Caribbean and Indo-Pacific floras</td>
</tr>
<tr>
<td>Halophila engelmannii</td>
<td>Caribbean flora</td>
</tr>
<tr>
<td>Halophila minor</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Halophila ovalis</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Halophila spinulosa</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Halophila stipulacea</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Heterozostera tasmanica</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Phyllospadix iwatensis</td>
<td>Temperate W. Pacific flora</td>
</tr>
<tr>
<td>Phyllospadix japonicus</td>
<td>Temperate W. Pacific flora</td>
</tr>
<tr>
<td>Phyllospadix scouleri</td>
<td>Temperate E. Pacific flora</td>
</tr>
<tr>
<td>Phyllospadix torreyi</td>
<td>Temperate E. Pacific flora</td>
</tr>
<tr>
<td>Posidonia australis</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Posidonia oceanica</td>
<td>Mediterranean flora</td>
</tr>
<tr>
<td>Posidonia caspica</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Posidonia sinuosa</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Posidonia angustifolia</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Posidonia coriacea</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Posidonia dierii</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Posidonia fruticosa</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Posidonia rostrata</td>
<td>S. Australian flora</td>
</tr>
<tr>
<td>Syringodium filiforme</td>
<td>Caribbean flora</td>
</tr>
<tr>
<td>Syringodiumiddyf forsythianum</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Thalassodendron hemprichii</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Thalassia hemprichii</td>
<td>Caribbean flora</td>
</tr>
<tr>
<td>Thalassodendron ciliatum</td>
<td>Indo-Pacific flora</td>
</tr>
<tr>
<td>Thalassodendron pseudobulbinum</td>
<td>Caribbean flora</td>
</tr>
<tr>
<td>Zostera capricornio</td>
<td>S. Atlantic flora</td>
</tr>
<tr>
<td>Zostera capricornio</td>
<td>S. Atlantic flora</td>
</tr>
<tr>
<td>Zostera marina</td>
<td>Temperate W. Pacific flora</td>
</tr>
<tr>
<td>Zostera noltii</td>
<td>N. Atlantic and Mediterranean floras</td>
</tr>
<tr>
<td>Zostera novaezelandica</td>
<td>New Zealand flora</td>
</tr>
</tbody>
</table>

*After Phillips and Menéndez (1998) and Kirkman and Walker (1990).*
seagrasses is far from settled, and there is considerable uncertainty not only at the level of species but also at the level of genera. Indeed, the total number of species described could, if uncritically accepted, exceed 60. Although species descriptions are often based on reproductive structures, these characters are, in practice, not always useful as diagnostic criteria to separate them. In practice, the identification of species in field studies is often dependent on vegetative characters to discriminate between genera, whereas the discrimination of species within genera is often based on the biogeographical segregation between species because congeneric species are often present in different floras. The diagnostic characters most commonly used to classify seagrasses into genera and species include differences in the venation, the presence of tannin cells, denticulation, and the tip shape of leaves (den Hartog, 1970). Although differences between reproductive structures are likely to be robust indicators of the broad genetic differences expected from species, differences in leaf anatomy are not as reliable. Leaf characteristics are remarkably plastic within seagrass species, even within adjacent shoots, rendering their diagnostic value questionable. Classification systems based on differences in leaf characteristics are therefore conducive to an ever-increasing number of species, particularly in the small and very plastic genus *Halophila*.

It is clear that a major revision of the taxonomy of seagrass species is needed, and that this revision cannot be based on anatomical or morphological characters alone. Experiments to test for phenotypic plasticity and the more parsimonic methods of molecular taxonomy must play a critical role in this revision. The use of biogeographic separation as diagnostic criteria for species and genera is also questionable because the current distribution of seagrasses may result from local extinctions isolating previously continuous distributions. Indeed, populations of some species with a wide distribution are genetically isolated across biogeographical regions (e.g., *Zostera marina*), although there is no evidence that such isolation may have resulted in sufficient genetic changes as to render them different species. In fact, the description of some seagrasses as separate species on the basis of their biogeographic separation (e.g., *Zostera* and *Phyllospadix* species as either side of the Pacific) has yet to be validated by robust genetic analyses. Indeed, examination of seagrass phylogeny using molecular markers has led to the postulation that there is no basis to separate *Zostera* and *Heterozostera* species into different genera (Les et al., 1997) and has generated a new, more parsimonic, tree of genetic relationships.
between the species (Table II) that represents a useful departure point for the needed reassessment.

Independently of current uncertainties regarding the exact size of the seagrass flora, it is clear that seagrasses are minor contributors to the richness of plant species in the sea, where they represent \(<0.5\%\) of the oxygen-evolving photosynthetic species. This minor contribution cannot be accounted for by the comparatively much longer evolutionary history of marine algae because seagrasses also represent \(<0.02\%\) of the angiosperm flora, despite an early origin in the angiosperm evolution. Although extinctions of some genera are documented in the fossil record, it is unlikely that past seagrass floras ever comprised a substantially greater number of species than does the extant flora.

The reason for the paucity of speciation in seagrasses remains a subject of intense debate, involving many possible constraints which operate collectively to result in the remarkably low membership of the seagrass flora.

The evolution of the land angiosperm flora is closely associated to a coevolution with insect pollinators. The fact that there are no insects in the sea has been used to argue that the very low extent of speciation in the seagrass flora is attributable to the inefficiency of hydrophilous pollination (Van der Hage, 1996). The stress associated with submerged life in a highly saline, high-energy environment may also further constrain the number of species which have been capable of undergoing the needed physiological, anatomical, and architectural (all seagrass species are rhizomatous) adjustments. No single constraint can be invoked in isolation to account for the low number of marine angiosperm species. The constraints imposed by hydrophilous pollination are not likely to differ much from those associated with the fertilization of red algae, the most species-rich phylum of marine algae. The adaptation of angiosperms to underwater life has led to approximately 500 species in fresh waters, and hundreds of angiosperms have adapted to life in saline environments. In addition to these factors, the greater possibilities for exchange of genetic material between distant populations in the sea, which reduces allopatric isolation, may have played an important role in accounting for the small number of marine angiosperms when compared with the apparent shorter-range exchange between angiosperms on land or in fresh water (Ackerman, 1998).

B. Seagrass Biogeography and Species Richness

The continuity of the marine environment across large spatial scales is best portrayed by the vast distributional ranges of many seagrass species, as exemplified by Zostera marina, which extends from subtropical to subartic waters. Basically, only nine floras, including some areas of overlap, can be identified (Table I and Fig. 2) in the shallow, coastal areas inhabited by seagrasses between polar waters (Hemminga and Duarte, 2000). Six of these species assemblages are composed of temperate species, and only two are represented by exclusively tropical (or subtropical) species.

The biogeographic patterns of seagrass floras are parallel to those followed by hermatypic corals, tropical fish, and mangrove species, all of which reach their highest diversity in the Indo-Pacific region, centered at Malaysia, which is believed to represent the center of origin of these organisms. Indeed, the number of seagrass species declines sharply with increasing distance from the region of highest diversity along the major oceanic currents in the region (Mukai, 1993).
The boundaries between seagrass floras are often sharp, corresponding to frontal areas separating different water masses, such as the Almería-Oran density front separating Atlantic from Mediterranean surface waters. Oceanographic fronts act as physical barriers for the dispersal of organisms, preventing the dispersal of seagrass propagules. The importance of major oceanographic fronts as boundaries between different seagrass floras indirectly validates the perception that the low number of seagrass floras and species result from the continuity of the marine environment over large scales.

Geographically isolated floras, however, have many common characteristics, such as the presence of congeneric species assemblages in distant floras such as the twin assemblage of *Thalassia*–*Halodule*–*Syringodium* in the Caribbean and the Indo-Pacific regions and *Zostera*–*Phyllospadix* species in both sides of the temperate Pacific zone (Fig. 2). Single genera are also common to different floras, such as *Posidonia* present in the Mediterranean and Southern Australian floras (Table I and Fig. 2). The occurrence of twin species suggests an even wider early distribution of the species, separated by the rearrangements of continental mass and local extinctions. The high species richness in the Indo-Pacific region (Fig. 2) suggests that this may have represented an area of particularly high speciation acting as a focus for radiative dispersal of the species or that, alternatively, extinctions have been fewer there.

### III. DISTRIBUTION AND CONTROLS OF SEAGRASS SPECIES DIVERSITY

#### A. Distribution and Controls at the Global Scale

The biogeographic factor constraining the distribution of floras is the major determinant of species richness at the global scale because it sets the maximum species richness possible at any one site. This maximum, however, may never be realized. For example, in Western Australia, the flora contains many species which often grow segregated, forming mostly monospecific stands. The maximum species richness reported at any one stand is 12 species, but the majority of meadows studied are monospecific, with an exponential decline in the number of meadows with increasing number of species (Fig. 3). The mean richness of 596 descriptions of seagrass meadows encountered in a thorough revision of the published literature is approximately two seagrass species per meadow (mean ± SE = 1.99 ± 0.07; Hemmings and Duarte, 2000). The resulting diversity, as calculated from the biomass distribution of the seagrasses using the Shannon–Weaver index, is closely linked to the species richness. The majority of seagrass meadows are monospecific and therefore have the minimum diversity index of 0. The highest Shannon–Weaver diversity index for a seagrass meadow was 1.56 for a
The number of species present in 596 seagrass meadows reported in the literature. Seagrass meadow containing seven seagrass species in Bolinao (Pangasinan Province, the Philippines), and indices have only been reported for meadows in the Indo-Pacific region, although the available estimates for such comparisons are few. These values compare poorly with diversity indices calculated for most other marine taxa, which generally oscillate between 1 and 4 for marine algae, invertebrates, and fish assemblages (Margalef, 1980).

The low species richness of seagrass stands, with most meadows being monospecific or containing two species, extends to areas occupied by floras with high numbers of species, such as Western Australia, in which species tend to be segregated. Mixed meadows, however, are common in the Red Sea and the Indo-Pacific region (Fig. 4), with the highest species richness found in Philippines waters. As a result of these patterns, seagrass species richness shows a clear latitudinal pattern, with the average number of seagrass species being greatest in meadows near the equator and least in meadows in cold, temperate areas (Fig. 5). The seagrass species richness in the subtropics in the Southern Hemisphere is somewhat greater than that for similar latitudes in the Northern Hemisphere (Fig. 5), reflecting the high species richness in the Australian continent and the Indian Ocean. The pattern toward species richness declining from the tropics to the cold temperate zone is comparable to those described for terrestrial taxa.

B. Distribution and Controls at the Regional Scale

The species richness encountered at any one location is constrained by the size of the flora present in the particular biogeographic region and the suitability of growth conditions in the area to support the different species. The study of the habitat requirements of seagrasses has received considerable attention and is relatively well delineated. Seagrass life is restricted to marine waters, extending to estuarine areas with low salinity, such as those encountered in the Baltic Sea. Seagrasses are further constrained by exposure to desiccation and physical disturbance by waves or, in cold seas, ice scouring in shallow waters, and the downslope limit is imposed by light, with seagrasses penetrating to a depth receiving, on average, 1% of the surface irradiance (Duarte, 1993). Within these limits, seagrasses require appropriate substratum which, for most species—except Phyllospadix, Amphibolis, and Posidonia, which are able to grow attached to rocks—is sandy to silty sediments. Highly silted sediments, which are also associated with reduced sediment conditions and accumulation of phytotoxins, are not appropriate to support seagrass life, which is restricted to sediments with low organic contents, redox potentials above 200 mV, and sulfide concentrations ≤100 µM. Although these growth requirements define habitats appropriate to support seagrass life, different species are expected to have different tolerances to these stress factors, which should lead to differences in species richness among locations at the regional scale.

Species-specific growth requirements are poorly documented for seagrasses, but some regional patterns are emerging. Seagrass species richness declines along gradients of siltation in Southeast Asian coastal waters, with a steep reduction in species richness as the silt content of the sediments exceeds 13%, and only monospecific (E. aequatorialis) meadows occur at highly silted locations (Fig. 6). That these patterns derive from species-specific differences in the resistance to disturbance has been predicted by theory and verified experimentally (Duarte et al., 1997).

C. The Maintenance of Local Species Diversity

Seagrass species are closely related and therefore often engage in competitive interactions in the resource-poor environments which they inhabit. Competitive interactions, if strong, could lead to species replacement and
the dominance of one species following a period of coexistence. Indeed, disturbance experiments have documented how a period of high diversity during intermediate stages of colonization is followed by the dominance by the climax species, which is typically slow growing. That the occurrence of such monospecific stands of climax species may result from strong resource limitation has been demonstrated by long-term fertilization experiments of a monospecific *Thalassia testudinum* meadow, which was colonized by the pioneer species *Halodule wrightii* following experimental nutrient additions (Fourquean et al., 1995).

High disturbance is also expected to lead to species-poor meadows which contain only small, pioneer species capable of fast colonization and growth, often arrested in a permanent stage of colonization. This is best illustrated by the occurrence of meadows containing only *Halophila* species in areas visited by dugongs in Southeast Asia. The grazing mode of the dugong uproots the seagrass, creating small gaps (feeding tracks about 2 m long and 30 cm wide) that must be recolonized in between successive visits. Only the small *Halophila* species, able to grow 9 m annually, are sufficiently fast colonizers to withstand this pressure. Because *Halophila* leaves are also tender, low in fibers, and nutrient rich, dugongs have been postulated to be actively farming these seagrass species (Preen, 1995).

Hence, seagrass species diversity is expected to comply with the intermediate disturbance hypothesis, in which the highest diversity is expected at moderate levels of disturbance. Multispecific seagrass meadows would therefore develop in the presence of moderate disturbance levels. The maintenance of the high species diversity in the most diverse meadow reported to date, located in the Philippines, has been attributed to the disturbance induced by the activity of burrowing shrimps, which maintain small-scale (<cm²) patchiness by generating gaps that allow the proliferation of pioneer species in meadows otherwise dominated by large, climax species (Duarte et al., 1997).

Although the preceding discussion highlights the role of competitive interactions, there is also evidence of positive interactions whereby the presence of some species may facilitate the occurrence of other species,

---

**FIGURE 4** Contour isolines of seagrass species richness generated from data on 596 seagrass meadows reported in the literature.
FIGURE 5 The mean (±SE) species richness in seagrass meadows growing in different latitudinal ranges. Compiled from data on 596 seagrass meadows encountered in a thorough review of the literature.

thereby maintaining high seagrass diversity. This role has been assigned to “engineering species” (Duarte et al., 1997) able to modify the growth conditions to improve their suitability to maintain plant growth. The occurrence of some seagrass species in Southeast Asia has been reported to be associated with the presence of the dominant species, Thalassia hemprichii, which is also linked to high biomass development in the meadows (Terrados et al., 1997). Seagrasses are able to alter the sedimentary conditions in which they grow, injecting oxygen into the sediments through their roots, thereby avoiding the accumulation of phytotoxins produced under anoxic conditions. Through this effect, seagrass species with extensive root systems, such as T. hemprichii, may maintain sediment conditions tolerable for seagrass growth where the organic loading derived from the seagrass growth would otherwise render sediments highly reduced and unsuitable for seagrass growth.

IV. SEAGRASS GENETIC DIVERSITY

Seagrasses are clonal organisms, so the genetic diversity present in meadows may be low. Indeed, vegetative reproduction is by far the dominant mode of propagation of seagrasses; therefore, the hypothesis that entire meadows may be composed of one or a few organisms cannot be rejected without serious consideration. Some aspects of seagrass biology, such as the occurrence of mass mortality across large spatial scales, have been ascribed to the dominance of vegetative reproduction in seagrasses and an assumed low genetic diversity to resist diseases. However, examinations of seagrass genetic diversity are still few and have lagged well behind studies of genetic diversity in other angiosperm groups.

The introduction in the mid-1990s of molecular techniques to quantify the extent of genetic diversity of seagrass populations is now providing the needed
V. OBVIOUS DIVERSITY IN THE SEAGRASS HABITAT

As indicated previously, seagrass ecosystems are important reservoirs of biological diversity in the sea. Seagrass ecosystems tend to contain more species than the adjacent unvegetated bottoms, and the species common to both seagrass meadows and unvegetated areas tend to be more abundant within the meadows. The enhanced biological diversity and richness within seagrass meadows result from their importance as habitat for (i) species that use seagrasses, and the rich detritus products as food sources; (ii) species that use seagrass surfaces as substratum; and (iii) species that seek refuge within seagrass meadows. These species encompass a broad taxonomic spectrum from fungi to mammals.

A. Microbial Diversity

The abundance of microbial communities appears to be enhanced within seagrass meadows compared to adjacent bare sediments. These microbial communities benefit from the large amounts of detritus seagrass produces as well as the efficient retention of carbon deposited from the water column in vegetated sediments. As a result, there seems to be clear links between bacterial activity and seagrass production and nutritional status, and exploratory analyses have shown that seagrass meadows host highly diverse bacterial communities. The association between seagrass meadows and fungi is even clearer because seagrass rhizomes and roots are heavily colonized by fungi, although mycorrhizal-type associations do not appear to have been developed in the sea. In fact, some of these fungal species may be pathogenic; in particular, the pathogenic Labyrinthula species have been found to be widespread in seagrass stands. Labyrinthula species were associated with large-scale seagrass (Z. marina) declines in the Atlantic in the 1930s, although the role of these pathogens on the decline remains hypothetical. Recent analyses have also reported a rich ciliate community to be associated with seagrass leaves. This ciliate community encompasses a wide diversity of taxa, and their role in the seagrass ecosystems has yet to be elucidated. These microbial communities may rely directly on detritus produced by seagrasses or benefit from the trapping of particles within the quiescent waters within seagrass canopies.

The best studied microbial community associated with seagrass meadows is the epiphytic microalgal community that develops on the seagrass leaves. This community is largely represented by large benthic diatoms, which colonize newly produced seagrass leaves and are replaced, as the age of the leaf substratum increases, by macroalgae and sessile invertebrate communities. Hence, there is a gradation in epiphytic community structure from the younger leaf base toward the leaf apex, particularly in species with long leaf life spans such as the Mediterranean seagrass P. oceanica, the leaves of which last for about a year. The epiphytic community is an important component of the ecosys-
tem metabolism within seagrass meadows, often supporting a primary production comparable to that of the host plants; it also makes an important contribution to heterotrophic processes. In addition, the epiphytic community is responsible for high deposition rates of carbonates on seagrass surfaces, particularly in warm waters, and is therefore responsible for the high carbonate deposition in seagrass sediments.

B. Macrophytes

Macrophytes often live in association with seagrasses, prominently as epiphytes on the leaf apices of species with long leaf life spans. The number of epiphytic macrophytes exceeds 330, including all major taxonomic divisions, in which blue-green algae are relatively rare. Competitive interactions and other negative effects are particularly intense in eutrophic environments because the abundance of opportunistic algae (Ulua, Enteromorpha, and Chaetomorpha) is promoted under high nutrient loading. The growth of these opportunistic algae can be so extreme that they can develop thick blankets over the seagrass meadows, which may suffocate or uplift (from the buoyancy due to gaseous release) the plants. In addition, competitive interactions may also be established with the siphalonales, notably species of the genus Caulerpa, the rhizoids of which confer the ability to extend horizontally over soft sediments and exploit the sediment nutrient pools. Seagrass meadows in Southeast Asia often contain many Caulerpa species. Introduction of Caulerpa species (C. taxiformis and, recently, C. racemosa) to the Mediterranean poses an important threat, where invasive expansion occurs, to the seagrass meadows (e.g., P. oceanica).

C. Invertebrates

The invertebrate fauna associated with seagrasses is very rich, including a multitude of organisms living epiphytically on their leaves and on their rhizomes and associated with seagrasses on the sediments. Epiphytic fauna comprise more than 173 species. Meiofauna include nematods, rotifers, tardigrades, copepods (planktonic and epibenthic harpacticoid species), and ostracods. Copepods and other free-living zooplankters often swarm within seagrass beds. The leaves support a rich community of epiphytic filter-feeding sessile organisms, including hydrozoans, ascideans, and bryozoans. Sponges grow attached to the seagrass rhizomes exposed aboveground and sometimes epiphytically on their leaves (e.g., E. acoroides). Polychaetes are abundant within seagrass sediments and also reach high densities on exposed seagrass rhizomes. Biter polychaetes have been reported to cause damage to seagrass leaves and rhizomes. Gastropods rank among the most abundant organisms on seagrass leaves, on which they graze the epiphytic algae and sometimes erode the leaf epidermis, damaging the seagrasses. The large gastropod Strombus gigas, the queen conch, is an important consumer in Caribbean T. testudinum meadows, but heavy fishing pressure has decimated its populations. Seagrass beds are important habitats for bivalves including scallops, which show behavioral adaptations toward a preferential recruitment to seagrass beds. Bivalves may experience an increased growth rate within seagrass beds, as reported for Mercenaria clams and mussels, due to the particle trapping by seagrasses and may also be less prone to predation there. These bivalve populations engage in mutualistic relationships with the plants, such as those reported for blue mussels (Mytilus edulis) and temperate seagrasses, because their enhanced activity within seagrass beds provides inorganic nutrients to the surrounding waters and sediments which promote seagrass growth. An outstanding association between seagrasses and bivalves is that by Pinna nobilis, a very large (up to 80 cm tall) bivalve that is largely restricted to P. oceanica beds. Pinna nobilis is a protected Mediterranean species and is particularly endangered because, in addition to direct threats from collection and predation, its populations are declining as a result of the widespread decline of P. oceanica across the western Mediterranean. Furthermore, P. nobilis is the host of a crustacean species (Pontonia pinnoophilus) which lives in mutualistic association inside its large shell and is also an endangered species because of the decimated populations of the host bivalve. This nested association between P. oceanica, P. nobilis, and P. pinnoophilus has been termed a “biodiversity Russian doll” (Richardson et al., 1997) and used as a paradigm of how the threat to landscape components, such as seagrass meadows, involves a threat to the many species therein contained. Although nudibranchs are not particularly prominent components of seagrass beds, sea hares (Aplysia) have been observed to graze on seagrass leaves, particularly when gathering in large numbers during the reproductive season. Isopods and amphipods are major consumers of both seagrasses and their epiphytes. In particular, the isopod Idotea plays an important role as a link between seagrasses and higher trophic levels in food webs. Amphi-
pods rank among the most abundant organisms within seagrass meadows. Thalassiniid shrimps are important components of seagrass meadows, particularly in tropical waters in which their burrowing activity generates gaps within the seagrass and affects the growth of the adjacent shoots. Thalassiniid shrimps can also be major grazers in seagrass meadows, removing up to 30% of their production. Thalassiniid shrimps clip leaves of seagrass species adjacent to their burrows and transport them into the burrows to decompose prior to consumption. In addition, their burrowing activity and the active and passive ventilation of their burrows appear to have a major effect on the sediment conditions, attenuating the reduced stage of the sediment while increasing nutrient availability, which may promote seagrass growth. Hence, Thalassiniid shrimps are important engineering species within (particularly tropical) seagrass beds, in which their disturbance effect has been equated to that of typhoons and is responsible for generating small-scale, highly dynamics gaps that are believed to be essential for maintaining multispecific seagrass beds (see Section III). Seagrass beds are also important nursery areas for prawns such as tiger prawns (e.g., Penaeus semisulcatus and P. esculentus in Australian waters), which have been found to feed on seagrass seeds, and caridean shrimps are abundant in subtropical and tropical seagrass meadows. Decapod crustaceans such as Callinectes spp. can also be important predators of seagrass seeds. Fiddler crabs (Uca spp.) are also very abundant in tropical and subtropical intertidal beds, in which they use their developed claw to remove the epiphytes of the seagrass leaves for consumption. Detritivore echinoderms, such as holothurids and ophiurids, and herbivores such as sea urchins are abundant within seagrass meadows and are among the dominant invertebrate components therein. Sea urchins can graze heavily on seagrass leaves—to the point that sudden population outbreaks may defoliate seagrass meadows over vast areas.

**D. Vertebrates**

The fish communities within seagrass beds typically comprise between 12 and 50 species in any one meadow. Fish communities tend to be more abundant within seagrass meadows than in adjacent bare bottoms, and they seem to seek refuge from predation there or benefit from the higher prey densities often encountered within the seagrass bed. Only approximately 20 species worldwide graze on seagrass. Some species, however, graze on seagrass and/or the epiphytes present on their leaves, such as Boops boops (a herbivorous species in the Mediterranean), parrot fish, and rabbit fish (Siganus sp.). Cryptic species such as stonefish are often encountered within seagrass meadows, and gobids are also abundant in seagrass meadows, sometimes in association with Thalassiniid shrimps. Active predators are often observed visiting seagrass meadows to prey on the abundant animal resources present therein. Other species, such as rays, can act as important disturbance agents by uprooting the plants when hiding under the sediment. Demersal fish production has been found to correlate with seagrass biomass, likely reflecting the higher prey production within these ecosystems.

The main reptile users of seagrass beds are adult sea turtles, which are important consumers of seagrass meadows. The most studied species is the green turtle, Chelonia mydas, which grazes on the seagrass. The association of turtles with seagrass meadows is so strong as to be reflected in some vernacular names (e.g., turtlegrass for the Caribbean species *T. testudinum*). Their importance as grazers in seagrass beds, however, has decreased due to the decline of the populations. In Southeast Asia, sea snakes are often observed foraging on seagrass beds.

In temperate climates, intertidal seagrass beds (mostly *Zostera* sp.) often support migrant goose and swan populations. They are the most important vertebrate grazers in cold, temperate waters, in which low water temperature would hamper the digestibility of seagrasses by poikilotherm grazers. Goose and swans feed on both leaves and rhizomes, and they may play an important role in the dispersal of the plant seeds. Many bird species, including waders, shore birds, gulls, pelicans, and diving birds, feed on seagrass meadows, particularly on intertidal ones. The syrenidae or marine cows are large marine herbivores that feed exclusively on seagrasses and, in the case of manatees (*Caribbean* waters), also on freshwater plants growing in rivers. Dugongs (*Dugong dugong*) prefer fast-growing, small seagrasses, especially *Halophila*, which in Southeast Asia develop large stands in intertidal flats (Fig. 7), have low lignin and cellulose contents, and regenerate fast. Dugongs feed by absorbing the sediments through their snouts and filtering out the seagrass while dropping the finer sediment particles. Their feeding activity leaves behind distinct feeding trails, usually approximately 20–30 cm wide and 1 or 2 m long, that are bare of seagrasses. In contrast, manatees graze only on the leaves and do not demonstrate a digging behavior. Dugongs and manatees are threatened and protected throughout their ranges. Al-
though the manatee population is still significant, dugong (Southeast Asia to Australia) populations have disappeared from most of their original range, and sizable populations have only been reported from the Great Barrier Reef. Although active hunting of sea cows has been abandoned, their loss to the point of extinction of local populations continues due to mortality induced by fishing gear, dynamite and cyanide fishing, and accidental collisions. As a result, the dugong is now among the most endangered marine mammals at great risk of extinction.

VI. THE FUNCTION AND MANAGEMENT OF SEAGRASS DIVERSITY

Seagrass meadows are one of the ecosystems that provide most valuable services for the biosphere. They produce approximately 0.6 Gt C year⁻¹ and are responsible for one-sixth of the total net CO₂ uptake by oceanic biota (Duarte and Chiscano, 1999). Seagrasses provide habitat for animals, many of which are economically important or endangered, and also play an important role by stabilizing sediments in coastal areas (Hemminga and Duarte, 2000). These functions are closely linked to seagrass diversity because there is evidence indicating that mixed meadows can support considerably more biomass and be much more productive than less diverse meadows within regions (Terrados et al., 1997). High seagrass diversity may also be linked to a more effective retention of sediments because the more complex canopies of mixed meadows will achieve a greater dissipation of energy by acting at different scales of the energy spectrum than the simpler canopies of monospecific meadows. They are also expected to host a larger number of animal taxa than monospecific ones because of their diversity of substrate and food types. However, the possible functional roles of seagrass diversity remain largely untested, as is unfortunately the case for most other marine taxa.

The importance of seagrass meadows as components of marine diversity does not stem primarily from the diversity of the angiosperm but rather, as indicated previously, from the high diversity that these ecosystems harbor. The role of seagrass meadows as sites of enhanced marine diversity was first addressed in the study of commercially important species, which often use seagrass meadows as nurseries. These studies were later extended to the study of endangered marine organisms, such as marine turtles and dugongs, which rely on seagrasses for food. Indeed, marine biodiversity presents a nested structure within seagrass meadows, which contain endangered species that often act as hosts for species that are necessarily rare, such as the P. oceanica—P. nobilis (mollusca)—P. pinnophyla (decapoda) association described in Mediterranean waters (Richardson et al., 1997). The pattern toward high biological diversity within seagrass meadows has been further strengthened by the demonstration of high microbial diversity, with the leaves of seagrass species containing complex microbial assemblages (fungi, bacteria, microalgae, ciliates, and flagellates) which benefit from the substratum offered by the plants as well as their release of detrital carbon to support these active microbial food webs. The plants also benefit from the activity of epiphyte grazers that prevent an excessive accretion on seagrass leaves, alleviating the negative effects of moderate levels of eutrophication for the seagrasses. Seagrass ecosystems also support food webs distant from the meadows; exported seagrass material has been documented as the carbon basis for some deep-sea communities, and seagrass material accumulating on shore, which may form mounds up to 3 m in height, is used by insects and other terrestrial animals.

A. Managing Seagrass Diversity: Threats

The limited number of seagrass species in the ocean implies that threats to one or a few species may have devastating consequences for the overall diversity of these plants. Unfortunately, catastrophic seagrass decline has been reported for many seagrass species (Short and Wyllie-Echevarria, 1996), both locally and across entire basins, such as the wasting disease that decimated Z. marina in the Atlantic Ocean in the 1930s and the
widespread decline of *P. oceanica* in the Mediterranean Sea (Marbá et al., 1996). These losses often result from direct human-induced disturbance, primarily reduced water clarity derived from increased nutrient loading, but they also involve diffuse sources likely associated with changes in climate (Marbá et al., 1996) which may increase the occurrence of diseases (Short and Wyllie-Echevarria, 1996). Human disturbance leading to seagrass loss includes many factors in addition to eutrophication, such as sitation derived from deforestation (Teradas et al., 1997), mechanical damage from boating, dredging, filling, dynamite fishing and trawling, and altered sediment dynamics after coastal constructions (Short and Wyllie-Echevarria, 1996). Remarkably, seagrasses have considerable resistance to the release of toxic compounds (Short and Wyllie-Echevarria, 1996).

The seagrass area lost cannot be quantified due to lack of a reliable baseline information, but it is certainly large since even large meadows may be lost within a few years (Marbá et al., 1996). Concern about the loss of seagrasses, and the associated loss of ecosystem functions, biological diversity, and resources, has been translated into legislation aimed at protecting seagrasses in many countries (e.g., USA, Spain and France), and particular emphasis has been placed on the importance of preserving seagrass meadows at the International Convention of Biological Diversity (i.e., the Rio Convention). The effectiveness of these measures to promote seagrass recovery, however, is uncertain because, although disturbed seagrass can recover, the current capacity to predict this process is meager. Models of seagrass recovery timescales predict vast differences in recovery time among species, with small, fast-growing species being able to recover within a few years provided that adjacent reproductive populations exist, and the recovery of slow-growing, long-lived species such as *P. oceanica* involving timescales of centuries (Duarte, 1995). Therefore, restoration efforts have been designed to speed up recovery, but these efforts have a high cost and can only be implemented in the countries with the strongest economies.

Hence, the management of seagrass biodiversity must be based on prevention measures aimed at achieving sustainability targets. In particular, management practices should be based on the known habitat requirements of seagrasses to maintain water quality targets (particularly water transparency) and sedimentary fluxes (particularly to prevent erosion and reduce burial rates) compatible with seagrass life. Deterioration of these properties will lead to reduced seagrass diversity, species loss, and the loss of the associated seagrass functions and biological diversity in the meadows they form.

See Also the Following Articles

Grazing, Effects of *Marine Ecosystems*

Bibliography

Ackerman, J. D. (1998). Is the limited diversity of higher plants in marine systems the result of biophysical limitations for reproduction or evolutionary constraints? Functional Ecol. 12, 975 – 980.


SLASH-AND-BURN AGRICULTURE, EFFECTS OF
Stefan Hauser* and Lindsey Norgrove†
*International Institute of Tropical Agriculture and †King’s College London

1. Reasons for Slash-and-Burn Agriculture
2. The Importance of Fallow Length: Long and Short Fallow Slash-and-Burn Systems
3. Geographic Distribution and Characteristics of Slash-and-Burn Agriculture
4. Effects of Clearing at Field Establishment
5. Consequences of Burning
6. Effects during the Cultivation Phase
7. Effects during the Fallow Phase
8. The Scale Dependence of the Effects Caused by Slash-and-Burn Agriculture
9. Alternative Systems to Slash-and-Burn Agriculture

GLOSSARY
bulk density The mass of the undisturbed dry soil per unit volume.
cation exchange capacity The capacity of a soil to adsorb cations such as calcium, magnesium, and potassium on the surfaces of soil particles and organic matter.
dormant/dormancy Condition under which a seed will not germinate, even if all environmental requirements for germination are met.
fallow A phase in which no crop is on the land and the volunteer vegetation regrows.
leaf area index Single-sided total leaf area of the vegetation per unit area of land.
mulch Retained slash covering soil surface; this can be in situ slashed or imported biomass.
pegging The penetration of the fertilized groundnut flower into the soil.
penetrometer resistance Energy required to push a probe of defined size and shape into the soil.
transpiration Amount of water taken up by the vegetation and released through the leaves to the atmosphere.

SLASH-AND-BURN AGRICULTURE is a generic term for agricultural systems in which the fallow vegetation is manually slashed, left to dry and cleared from the field by burning before crop cultivation. "Swidden" is an English dialect word for a burned clearing, thus swidden agriculture is a synonym for slash-and-burn agriculture. After a cropping phase, the land is abandoned to a fallow phase. Later, the cycle is repeated. Only systems that alternate between crop and fallow phases are included in this definition. Multistory tree gardens, home gardens, and cocoa plantations, where crops are permanently cultivated, are excluded. With the exception of labor, slash-and-burn farmers use few or no external inputs. Implements such as machetes and hoes are most commonly used. Systems in which machinery is used for clearance and irrigated systems are excluded from this definition. Not considered in this context are the systems such as the anakua of the western Highlands in Cameroon, the shula of the Tanza-
nian grasslands, and the gy of Ethiopia, in which vegetation is slashed, gathered, covered with soil, and then burned inside the soil mounds.

I. REASONS FOR SLASH-AND-BURN AGRICULTURE

Most slash-and-burn farmers are poor (section III.B). Unlike many people in developed countries, they have little choice in how to live their lives. Often the only resource available to them is land. Thus, farming, whether subsistence or market oriented, might be their only option. However, even cash-crop production may not provide sufficient income for farmers to change to other economic pursuits. Thus, farmers remain. Farming methods used reflect the economic constraints of the farmers and tend to minimize labor requirements (section III, C).

A. Removal of Biomass

The amounts of biomass debris on the soil are extremely variable ranging from less than 10 Mg ha\(^{-1}\) to more than 500 Mg ha\(^{-1}\) dry matter. Large amounts of biomass make planting difficult and would require too much labor to remove manually. Thus, farmers use fire to remove the biomass, giving easier access to the soil for planting.

B. Release and Addition of Nutrients from Biomass

Where the fallow vegetation has a high proportion of aboveground biomass as wood, the nutrients bound in it are not easily available to annual crops, as wood does not decompose rapidly. Burning the biomass releases the nutrients.

C. Weed Avoidance and Weed Suppression

Slash-and-burn agriculture farmers have three principal methods to suppress weed infestation: site selection, the timing of the burn, and tree retention. In forested areas, farmers might choose older secondary or primary forest for clearing because the arable weed seed bank is depleted or absent because the fallow phase has exceeded the viable period of arable weed seeds. If fields get too weedy, farmers may abandon them and clear a new field as the labor requirement for a new clearing may be less than weeding an old field. In short fallow rotations, where the fallow length is insufficient to deplete the weed seed bank, some farmers delay the burn until a flush of weeds has germinated. These weeds are destroyed in the burn.

D. Pest and Disease Avoidance

As pesticides are not available to most slash-and-burn farmers, they attempt to avoid crop pests and diseases by appropriate site selection. This option is limited to regions where long fallow phases would withdraw the hosts of pests and diseases for a sufficiently long time to eliminate them. However, pests such as rodents and monkeys would not be affected.

E. Declining Crop Yields

The combination of nutrient depletion, increasing weed, and pest and disease pressures leads to declining crop yields. After 1 to 4 years the field is abandoned to a fallow phase usually because of crop yield declines, or high weed biomass decreasing labor productivity to the point where clearing a field requires less labor than maintaining an old one. In South America, farmers often abandon their fields when they cannot expect the yield of the subsequent crop to be more than half that of the first crop. However, superimposed on this is the social custom of caring less for crops that are expected to yield less, which accentuates and accelerates any decline. When a cash crop is grown, reasons for abandonment can be external. Farmers in Indonesia reportedly abandoned their slash-and-burn pepper farms before crop harvest, when the market price for pepper fell.

II. THE IMPORTANCE OF FALLOW LENGTH: LONG AND SHORT FALLOW SLASH-AND-BURN SYSTEMS

Two main types of slash-and-burn agriculture are distinguished, which differ in their effects on the environment: long fallow systems (shifting cultivation) and short fallow systems.

The fallow is the successional vegetation that follows the cropping phase. The timing of the burn, and tree retention. In forested areas, farmers might choose older secondary or primary forest for clearing because the arable weed seed bank is depleted or absent because the fallow phase has exceeded the viable period of arable weed seeds. If fields get too weedy, farmers may abandon them and clear a new field as the labor requirement for a new clearing
slash-and-burn agriculture, effects of

pest or disease buildup, or a combination of these. The
fallow can act as a weed-, pest-, or disease break and
fallow vegetation accumulates nutrients in the
aboveground biomass, improves soil physical properties
through root penetration, and usually permits the re-
covery of soil macrofauna. Thus, the sustainability of
slash-and-burn agriculture depends primarily on the
rate of degradation during the cropping phase, regener-
ation during the fallow phase, and their relative time
allocation.

A model of this is detailed in Fig. 1. However, the
diagram is a simplification and the rates of change are
unknown. At low human population density, even on
nutrient-poor soils, fallow phases may exceed 20 years;
thus the system is a case 1 type "long-fallow system." 
Labor is the main factor limiting production. For a
cropping phase of time a, fallow length \( \frac{b}{a} \) may
exceed the optimum. The optimum fallow length, b,
is defined as the minimum fallow length required to
maintain and maximize crop yields per unit area in the
long term and thus is just when the system has recov-
ered enough to permit this. When population densities increase, more land is
required and fallow phases are shortened, up to the
minimum length required to restore soil fertility (case
2). However, where human population densities in-
crease to the extent that more land needs to be cropped
than in case 2, fallows are further shortened and system
recovery is not possible. If no additional inputs are
made, this can cause soil fertility, weed, pest, and dis-
ease problems and lead to lower crop yields (case 3
type short fallow system).

III. GEOGRAPHIC DISTRIBUTION AND
CHARACTERISTICS OF SLASH-AND-
BURN AGRICULTURE

A. History

Slash-and-burn agriculture is probably the oldest
method of land preparation for planting crops. It was
used as a technique in China for establishing rice fields
as early as the late Stone Age, 8000 to 10,000 years
ago, and in Mexico at least 5500 years ago, concurrent
with the domestication of maize. It was used in central

---

FIGURE 1

Degradation and recovery in slash-and-burn systems as related to the cropping interval and
when no agronomic adjustments are made. Modified after Guillemin (1976).
SLASH-AND-BURN AGRICULTURE, EFFECTS OF

Europe, along the Danube River, to penetrate the post-glacial European forests, for cereal production and, later, in the boreal forests of Scandinavia.

B. Geographic Distribution

Today, slash-and-burn agriculture is practiced in the moist savanna, transitional, and forest zones of the tropics and subtropics of Central and South America, Africa, Asia, and the islands of Australasia. The majority of countries in these regions have weak economies. Slash-and-burn agriculture is also practiced in other nonindustrial areas, such as Bhutan. An estimated 36 million

C. Labor Requirements in Slash-and-Burn Agriculture

In areas with low population density, long-fallow systems predominate. Long fallow systems in forested areas have large land, yet low-capital and low-labor requirements. Soil fertility, weed, pests, and disease problems are avoided, rather than managed, by shifting to a new field. In the Amazon basin, only 8% of the human energy input to cultivate a cassava field, including post-harvest processing, is used for slashing and burning. The energy or labor efficiency of long fallow slash-and-burn systems is the highest among all agricultural systems in terms of energy invested versus the amount of energy gained with the crop yield. This is largely due to the absence of fossil fuel use and chemical inputs. However, the destruction of the biomass and the release of the accumulated energy and carbon (section V) in the burn is not considered in such calculations.

In long fallow systems, crop diversity within one field can be high. A tuber species is often the dominant crop, which is intercropped with other roots and tubers, grains, vegetables, and herbs and spices. In the early 1980s, researchers counted an average of 10 crop species per 25 m² in 3-month-old fields of the Maring people of Papua New Guinea. The dominant crop was taro (Colocasia esculenta). Other crops included yams (Dioscorea spp.), sweet potato (Ipomoea batatas), maize (Zea mays), beans (Phaseolus spp.), sugarcane (Saccharum officinarum), and summer squash (Cucurbita pepo). Taro monocrops were also common. In the Colombian and Peruvian Amazon, fields may be dominated by cassava, constituting 80% of crop numbers. In Pará, in the Brazilian Amazon, monocrop maize (Zea mays) fields are common and farmers may use fallow from 3 to 30 years old or even primary forest.

D. Typical Fallow/Crop Sequences in Slash-and-Burn Agriculture

In West Kalimantan in the mid-1990s, fallow lengths averaged 17 years after a 1-year cropping phase, thus less than 10% of land was cultivated at any time. Such areas are characterized by low availability of labor and a lack of infrastructure, leading to very limited exchange of products and thus predominantly subsistence farming. In 1996, in southern Cameroon, 32% of fields cultivated had a previous fallow length of 8 or more years or had not been previously cultivated. In the early 1980s, in north Sierra Leone in the forest-savanna transition zone, most fields had been established from falls of 30 years or more. Most farmers planted upland rice (Oryza sativa) for 1 year before abandoning the land. A minority of farmers followed rice by a groundnut then millet (Setaria spp.) sequence before abandonment.

In long fallow systems, crop diversity within one field can be high. A tuber species is often the dominant crop, which is intercropped with other roots and tubers, grains, vegetables, and herbs and spices. In the early 1980s, researchers counted an average of 10 crop species per 25 m² in 3-month-old fields of the Maring people of Papua New Guinea. The dominant crop was taro (Colocasia esculenta). Other crops included yams (Dioscorea spp.), sweet potato (Ipomoea batatas), maize (Zea mays), beans (Phaseolus spp.), sugarcane (Saccharum officinarum), and summer squash (Cucurbita pepo). Taro monocrops were also common. In the Colombian and Peruvian Amazon, fields may be dominated by cassava, constituting 80% of crop numbers. In Pará, in the Brazilian Amazon, monocrop maize (Zea mays) fields are common and farmers may use fallow from 3 to 30 years old or even primary forest.

In areas with low population density, long-fallow systems predominate. Long fallow systems in forested areas have large land, yet low-capital and low-labor requirements. Soil fertility, weed, pests, and disease problems are avoided, rather than managed, by shifting to a new field. In the Amazon basin, only 8% of the human energy input to cultivate a cassava field, including post-harvest processing, is used for slashing and burning. The energy or labor efficiency of long fallow slash-and-burn systems is the highest among all agricultural systems in terms of energy invested versus the amount of energy gained with the crop yield. This is largely due to the absence of fossil fuel use and chemical inputs. However, the destruction of the biomass and the release of the accumulated energy and carbon (section V) in the burn is not considered in such calculations.

In areas with higher population densities and consequently shortened fallow periods, economic conditions become more suitable for market-oriented, commercial farming as the higher population densities naturally create markets. While labor is available, small-scale farmers have scarce financial resources to purchase agricultural inputs. Due to infrastructural, economic, and soil-related problems of pesticide and fertilizer use, high-input, intensive agriculture is rarely practiced. With reduced fallow length, the labor requirement increases as additional field work, such as weeding and tillage, become necessary (section VI). In short fallow systems of southern Cameroon, land preparation, including slashing, burning, and cleaning the soil surface of unburned debris and weed stumps, was 10 to 13% of the total labor required to establish and harvest a groundnut/maize/cassava intercrop.

While population densities are increasing in many tropical areas and slash falls shorten, in some situations the reverse has occurred. There were decreases in the population density of the Mayan lowlands and this was followed by a transition from intensive agriculture to long-fallow slash-and-burn systems.

In West Kalimantan in the mid-1990s, fallow lengths averaged 17 years after a 1-year cropping phase, thus less than 10% of land was cultivated at any time. Such areas are characterized by low availability of labor and a lack of infrastructure, leading to very limited exchange of products and thus predominantly subsistence farming. In 1996, in southern Cameroon, 32% of fields cultivated had a previous fallow length of 8 or more years or had not been previously cultivated. In the early 1980s, in north Sierra Leone in the forest-savanna transition zone, most fields had been established from falls of 30 years or more. Most farmers planted upland rice (Oryza sativa) for 1 year before abandoning the land. A minority of farmers followed rice by a groundnut then millet (Setaria spp.) sequence before abandonment.

In long fallow systems, crop diversity within one field can be high. A tuber species is often the dominant crop, which is intercropped with other roots and tubers, grains, vegetables, and herbs and spices. In the early 1980s, researchers counted an average of 10 crop species per 25 m² in 3-month-old fields of the Maring people of Papua New Guinea. The dominant crop was taro (Colocasia esculenta). Other crops included yams (Dioscorea spp.), sweet potato (Ipomoea batatas), maize (Zea mays), beans (Phaseolus spp.), sugarcane (Saccharum officinarum), and summer squash (Cucurbita pepo). Taro monocrops were also common. In the Colombian and Peruvian Amazon, fields may be dominated by cassava, constituting 80% of crop numbers. In Pará, in the Brazilian Amazon, monocrop maize (Zea mays) fields are common and farmers may use fallow from 3 to 30 years old or even primary forest.
In southern Cameroon, four of the most common slash-and-burn field types are as follows (Fig. 2):

- The essep or long fallow or primary forest conversion field, which is used to gain title to the land. This is approximately 0.25 ha and a few larger forest trees are retained. Relatively shade-tolerant crops, particularly tannia (Xanthosoma sagittifoli-

um), egusi melon (Cucumisops manni), and plantain (Musa spp. AAB), are grown with many minor crops. After the plantain harvest, the fields are either abandoned for 20 or more years or are burned and recultivated after 1 to 4 years as a groundnut/maize/cassava field.

- The groundnut/maize/cassava field (Arachis hypoga-

era, Zea mays, Manihot escalenta) with plantains (Musa spp. AAB) and leafy vegetables as minor components. Fallow length is 8 to 10 years near the border of Equatorial Guinea yet 4 years near the capital city, Yaoundé.

- Monocrop horticultural systems, particularly toma-

toes (Lycopersicum lycopersicon) or maize for fresh consumption. These fields are most common around Yaoundé and external inputs, both inorganic and organic, are used.

- Musa spp. (banana and plantain) monocrops. These are usually established after long-fallow forest clearing.

Clearly, slash-and-burn systems may have high or low crop diversity. This does not depend on ecoregion or population density, but rather on the preferences of the farmers.

### IV. EFFECTS OF CLEARING AT FIELD ESTABLISHMENT

Farmers in different parts of the world use slash-and- burn techniques yet their goals may vary. These range from completely and irreversibly removing the initial vegetation to retaining a rich repository of indigenous and useful plant species while introducing crops. A number of elaborate clearing and tree-felling techniques were developed by different indigenous cultures to facilitate the drying and burning of biomass. These techniques are summarized by Peters and Neunsch- 

wander (1988).

The types and numbers of trees retained vary consider- 
elably, depending on the traditions of the farmers and their intentions. However, usually very large high-canopy trees are not felled as it is not worth the considerable effort involved. It would require paying for the labor of a skilled chainsaw operator, and the felled tree would take up a considerable part of the plot. Other trees retained include those producing fruits, nuts, or medicine, or those believed to maintain or enhance soil fertility or have a commercial value. Such trees are a source of seed, although such species cannot always establish in an environment imposed by slash and burn. Furthermore, these trees can harbor a large number of species of animals and possibly host microsymbionts. The retention of trees in slash-and-burn fields can facilitate the return of the initial vegetation, depending on the stand density of trees retained.

Reductions in tree densities, through slash-and-burn or other forms of clearance, may affect densities and species numbers of arboreal and litter-dwelling ants. Yet these species may also have a role in pest control. In areas cleared of forest in the Amazon, it was found that isolated trees can retain considerable ant diversity, although the species composition was not identical to that of the nearby forest.

When slash-and-burn farmers in the forests of West and Central Africa clear for a short-fallow groundnut/maize/cassava field, they completely clear the fallow vegetation and remove all trees so that there is abundant light for these crops, which do not tolerate shade. Additionally, as any mulch layer would impede groundnut growth, the farmer attempts a complete burn. The soil is tilled by hoeing after burning and weeding is frequent to reduce competition. In contrast, the root, tuber, and leafy vegetables of the essep field can be planted even if trees are retained, and if the slash is not completely burned, the soil is not tilled and weeding is neither as intensive nor as frequent as in the short-fallow fields.

Partial or complete removal of vegetation has differ- ential effects on the microclimatic and hydrological conditions after clearing. Soil temperatures are usually higher in cleared areas and even higher when the slash is burned. The presence of tall vegetation leads to a different redistribution of rain, as less or no rain will be intercepted in the canopy. Raindrops reach the soil surface with their full impact, potentially contributing to compaction and crust formation at the surface. Surface compaction can lead to temporarily ponding water at the surface and, if on sloping land, can cause erosion. Furthermore, crust formation can impede the gas ex- change between soil and atmosphere resulting in deteri- orating conditions for the soil fauna and flora.

Slash-and-burn fields have lower biomass than for- ests and crops may not cover the land for the entire year. The leaf area index may not be as high as in a
forest, thus transpiration is lower. With reduced transpiration, water flux through the soil is increased, which can increase mineral losses due to leaching. The root systems of crops have to grow for each cropping cycle while in the forest the root system of the trees remains in place. Access to and uptake of water at the start of the rains is thus immediate in forest. Rooting depths of crops are restricted to the upper layers of the soil, while forest vegetation roots may penetrate deeper layers drawing on water during dry seasons, not accessible and available to food crops. The more permanent canopy and root system of forest contributes to less strong fluctuations of most climatic and hydrological factors.

V. CONSEQUENCES OF BURNING

A. Heat Evolved during the Burn

Few studies have measured the heat evolving from the burning of slashed biomass. The amount of heat evolved depends on the amount of biomass and its water content. Climatic conditions during the burn, such as wind speed, topographic features such as aspect, and slope of the field, have an impact on the intensity of the burn and the maximum temperatures attained. Slash-and-burn agriculture in savanna regions or in short fallow systems, where small amounts of biomass are burned, does not cause significant heat-related effects on the soil. Temperatures at the soil surface are relatively low and heat penetration into the soil is limited to 1 to 2 cm depth and the exposure time to the heat is short. However, even fast burns, which do not attain high temperatures, can kill seeds lying on the soil surface. There are no reports of light burns affecting seeds buried in the soil.

Burning of cleared forest biomass, depending on its water content, can reach 800°C near the soil surface, and these temperatures will be maintained for a longer period. If temperatures reach 400°C, soil organic matter may be lost by combustion, resulting in decreases in soil organic C and N, and clay particles may fuse, altering the soil texture and its cation exchange capacity. The spatial variability of temperatures is higher in a cleared forest than in short fallow or savanna systems. Tree trunks may burn for more than a week, evolving large amounts of heat and depositing large amounts of ash in a small area. In southern Cameroon, on an Ultisol, the maximum temperatures attained during burning of 3000 Mg ha\(^{-1}\) of wood, equivalent to a trunk of approximately 0.65 m diameter at a wood density of 0.6 Mg m\(^{-3}\), were 788°C at the soil surface, 225°C at 5 cm, and 172°C at 10 cm and 105°C at 20 cm depths. The burn lasted for approximately 24 hr. Such severe burns usually kill plant seeds even in deeper layers and reduce living microbial biomass in the soil, including mycorrhizae, rhizobia, and other microsymbionts. Further, all soil meso- and macrofauna are affected, depending on the temperature reached and the soil depth to which it penetrated.

B. Soil, Soil Nutrients, and Soil Physical Properties

Depending on the level of disturbance, local consequences range from nutrient inputs from the atmosphere such as sulfur dioxide and flying ash to changes in microclimate and increased variation in rainfall pattern, and influxes of pests, diseases, and invasive weeds. Many soils in forested areas of the tropics have a low nutrient status, aluminium toxicity problems, high levels of phosphorus fixation by iron oxides, and a low cation exchange capacity due to the dominant kaolinitic
clays, which are also susceptible to structural collapse under mechanization. Where the fallow vegetation has a high proportion of aboveground biomass as wood, the nutrients bound in it are not easily available to annual crops, as wood does not decompose rapidly. Burning the biomass releases the nutrients, and with the exception of nitrogen and sulfur, a large proportion will remain in the ash on the soil surface. While the nitrogen is almost entirely lost to the atmosphere, the sulfur, as sulfur dioxide, can be redissolved as sulfuric acid near the burn, if the humidity of the air is high. When the fuel (biomass debris) is very dry, the burn is fast and hot and strong upward air currents result, which carry away the ash particles. Windy conditions during the burn will increase such losses. In the Amazon, element transfer to the atmosphere due to ash particle transport and volatilization have been reported as up to 98% C, 98% N, 33% P, 31% K, 24% Ca, and 43% Mg of the initial amounts in the fuel. Sulfur losses to the atmosphere were estimated between 69 and 70% of that in the original biomass.

Wood ash contains calcium, magnesium, and potassium in the form of phosphates, carbonates, and silicates and also other elements. The nutrients in the ash are partially water soluble and will be released into the soil solution with the rain. Thus, slash-and-burn agriculture makes nutrients available to crops. While farmers realize the fertilizing effect of the ash, they are more concerned to achieve complete burns to remove the biomass. The longevity of increases in calcium, magnesium, and potassium concentrations in the topsoil depends partially on the cation exchange capacity. While the increased calcium concentrations are likely to be long term because of low calcium mobility and solubility, magnesium and potassium may be prone to leaching. Burning usually reduces soil organic matter content. On kaolinitic soils, much of the cation exchange capacity (CEC) is on organic matter thus a reduction in organic matter will reduce CEC and increase the risk of leaching. In southern Cameroon, it could be shown that increases in exchangeable potassium were measurable down to 50 cm within 6 months of burning forest biomass.

However, with the rain dissolving the ash, soil pH usually increases rapidly to above pH 8 and remains high for an extended period. Topsoil pH usually increases after burning. In soils where aluminum toxicity problems occur, burning reduces extractable aluminum concentrations and increases phosphorus availability.

It has been reported in the literature that the burn loosens the soil and facilitates tillage and planting by reducing bulk density and penetrometer resistance. Direct measurements in southern Cameroon could not confirm these reports. Farmers experience the forest floor to be difficult to penetrate with tools before the burn, probably because of the dense network of small but woody roots between the litter layer and the soil surface. The litter layer and the roots will be destroyed in the burn and will no longer be an obstacle. Furthermore, attempts to work the soil before burning are on dry soil, which is harder, while planting after the burn usually happens after some rains, softening the soil. However, where burns are hot and lengthy, such as under a log, the resultant ash is white. Soil color changes to bright red and compaction and a collapse of its soil structure can be observed. This is probably due to a relative excess of monovalent cations, which leads to a breakdown of cation bridges between clay minerals, the combustion of soil organic carbon and thus substances binding soil particles, and the elimination of soil macro and microfauna mixing soil particles and organic materials to form stable aggregates. Slash-and-burn farmers recognize such areas in their fields and usually do not plant crops there.

In savanna regions with small amounts of biomass, the effects of burning upon soil are not discernible. Ash inputs do not cause a measurable increase in pH or cation concentrations. Neither soil organic carbon nor total soil nitrogen is affected by such light burns. In grass fallows, the amounts of biomass may not be an obstacle to planting. However, the quality of the slashed biomass is low. Biomass retention as mulch would have adverse effects on crop growth, because soil nitrogen might be immobilized in the decomposition process, causing nitrogen deficiency for the crop.

Most slash-and-burn agriculture systems use only the in situ slashed vegetation. There are, however, systems such as the chitemene system in southern Africa in which wood from a larger cleared area is collected and concentrated in a smaller area where it is burned. Farmers seek nutrient augmentation or concentration in the topsoil by piling large amounts of wood on the soil surface. The wood is transported to the field from the surrounding miombo woodland. The area cleared may be up to 20 times larger than the field in which the wood is burned, adding large quantities of nutrients. Although this system appears to be designed to concentrate nutrients, its efficiency in retaining these nutrients remains low. In an experiment conducted in Zambia, 84% of the phosphorus contained initially in the vegetation was accounted for by increases in the top 50 cm of soil at 40 days after burning. However, 57% of this phosphorus was at 20 to 30 cm depth, in a region already inaccessible to any growing crop. For potassium, only about 10% of the original input was retained.
SLASH-AND-BURN AGRICULTURE, EFFECTS OF

by the top 50 cm soil at 40 days after burning. The remainder presumably had leached below 50 cm depths.

In other systems, only part of the biomass is burned while a certain portion is used for other purposes such as fencing the field against animals, as fuelwood or for construction.

C. Soil Bacteria and Fungi

Burning can reduce living microbial biomass in the topsoil and potentially their decomposing activities. However, there are reports that burning increased microbial biomass and activity in tropical savanna and in a tropical plantation forest. It has been shown in Kenyan soils that total bacteria counts dropped after burning. Nitrogen-fixing bacteria are very susceptible to pH changes after burning. Such effects of the burn on microorganisms can subsequently affect the composition and productivity of the vegetation as certain species will depend on symbionts.

Many crops have mycorrhizae and they are able to access organic phosphorus using the enzyme phosphatase in a hydrolysis reaction. Thus, although burning can increase phosphate availability, if burning reduces mycorrhizae, overall effects could be negated, depending on whether the crops grown are mycorrhizal and whether the planting material is infected. Studies in India comparing vesicular-arbuscular mycorrhizal (VAM) infection of plants between burned and non-burned areas of a tropical forest found that infection was reduced in burned areas. However, studies from wet tropical forests in Costa Rica found that mycorrhizae survived burning.

D. Invertebrates: Soil Fauna

The soil fauna are only directly affected by the burn if their active stages are present in the topsoil during the burn. As fields are usually burned toward the end of the dry season, groups such as earthworms are inactive and thus unaffected. After burning large amounts of biomass, the ash is dissolved in the first rains. This solution is very high in pH (up to 10). Such solution infiltrating the soil might kill certain species that used to live under rather acidic conditions under the previous forest cover (pH of around 4.0 to 5.5 in most tropical forests).

E. Weeds

Severe burns usually kill plant seeds even in deeper layers and can reduce the weed seed bank. Although burning is believed to reduce weed pressure, it can facilitate weed invasion. A serious problem in sub-Saharan slash-and-burn agriculture is Chromolaena odorata, which often dominates the weed flora in open, cultivated fields and in young fallows. C. odorata seeds have higher germination rates when they are exposed to light on the soil surface. Thus, airborne seeds can colonize newly burned fields. On the contrary, where slash is retained, seeds fall onto the slash, which dries out quickly preventing germination. Once established, C. odorata plants can survive burning as their woody stumps may remain unaffected in the upper soil layers. Burning destroys seedlings of other species and thus can reduce competition from other weeds on C. odorata.

F. Other Plant Species

Slash-and-burn agriculture in a tropical dry deciduous forest in Mexico eliminated 29% of resprouting species, particularly those that were present at low abundance. This reduction was exacerbated by repeated burning.

G. Vertebrates

The effect at field establishment and specifically at burning depends on the species and its mobility. While mammals may escape the slash and the burn by moving into surrounding unaffected vegetation, some reptiles and amphibians are killed when forest is cleared and burned, as they cannot escape quickly. Depending on the season and the temporal pattern of activity, animals that hibernate through the dry season may escape the fire if burning happens before larger rains have fallen. Where vertebrate densities decrease after slash-and-burn agriculture, the potential for fallow recovery may be reduced where vertebrates are responsible for seed dispersal.

H. Global Effects

The destruction of natural vegetation and its replacement by crops in slash-and-burn agriculture systems is accompanied by the release to the atmosphere of carbon as carbon dioxide during the burn. Decomposition rates of soil organic matter may increase during the cropping phase due to tillage and soil chemical changes, releasing additional amounts of carbon. Furthermore, food crops will assimilate less carbon from the atmosphere than the initial vegetation would have assimilated, thus reducing the C-fixation in biomass and contributing to a larger net release of carbon.
VI. EFFECTS DURING THE CULTIVATION PHASE

In traditional shifting cultivation with long fallows, the cropping phase lasts 1 to 4 years. As the length of cultivation increases, the potential of land to recover is reduced. Operations such as tillage and weeding frequently disturb the soil and the viability of seeds of the initial vegetation is reduced. Regrowth from stumps is slashed more frequently, increasing the risk of dieback. There are changes in microclimatic and edaphic conditions such that species of the initial vegetation cannot survive.

A. Tillage

The traditional shifting cultivator does not use many tools to work the soil. Most crops are planted with a simple planting stick or by opening small planting holes with a machete or hoe. Most crops, such as cassava, manioc, taro, plantain, melon, and maize, do not require tillage to succeed. If slash and burn is practiced in short fallow systems in degraded land, crops such as melon and plantain are not grown as they are considered not to produce well, while other crops can perform well but require tillage. The soil is tilled either to mix the ash with the soil, to bring the seed to the required planting depth, or to concentrate nutrient-rich topsoil around the crops. Most intensive soil disturbance is caused by tillage for crops such as yams (Dioscorea spp), which are mounded, groundnuts (Arachis hypogaea) for which the soil is tilled at planting and a second time before pegging, potatoes (Solanum tuberosum), and sweet potato (Ipomea batatas) for which ridges or small mounds are made. Farmers throughout have made their own observations and know which crops require tillage. However, tillage is a labor-intensive operation and is more frequently used where labor availability is high.

Tillage disrupts the natural layering of the soil and breaks up soil aggregates. Naturally formed pores (earthworm channels, termite and ant galleries, root channels) are interrupted and often are filled with loose fine material. Soil fauna dwelling in such pores will be negatively affected by tillage, as their habitats are destroyed or the number of potential niches is reduced. Tillage changes the water regime in the topsoil and might lead to a more rapid drying out of the top layer, causing stronger moisture and temperature fluctuations. Soil mites and springtails respond to such fluctuations by withdrawing into deeper layers, which may be less rich in substrate and thus do not permit the same level of activity and population density.

When land is cleared, slashed, burned, and tilled, earthworm density, diversity, and activity are reduced. This has been shown in southwest Nigeria (Fig. 3), where the earthworm fauna is dominated by Hyperodrilus africanus and the epigeic Eudrilus eugeniae, in the Peruvian Amazon, in southern Cameroon and in southeast Mexico (Figs. 4 and 5). Earthworms are classified functionally into epigeic, anecic, and endogeic categories. Epigeics live in the litter layer. Anecics feed on litter and soil, dragging litter into their vertical burrows. Endogeics live in the soil feeding on soil organic matter and dead roots. Anecic and epigeic are the groups most immediately affected by slash, burn, and tillage due to the loss of current and future substrates and physical habitat disruption. These groups will take longer than endogeics to recolonize the recovering fallow.

Tillage affects the seed bank in the soil. This is less important in long fallow slash-and-burn systems as seeds will normally be on the soil surface where they are most likely to be killed by the burn. In younger fallow, with light burns, not all seeds on the soil surface may be killed. Where the land has been previously cultivated with tillage or any soil-moving operation (harvest of root and tuber crops) has been conducted, a seed bank, protecting the buried portion of seeds from the effect of subsequent burns, may already exist in the soil. Buried seeds can become dormant and survive for several years. Tillage after the burn will move such seeds closer to the soil surface, increasing their chance of germination.

B. Weeding

Most weeding operations are conducted with the same or similar tools as tillage. Therefore, in addition to the tillage, weeding affects the soil, soil inhabitants, and the weed seed bank. It is difficult to separate the terms “seed bank” and “weed seed bank” as the same species may be both: a weed in the cropping phase and a desired plant in the fallow phase of a slash-and-burn rotation (cycle). To simplify, all nonplanted species will be treated as weeds in the cropping phase, although in some slash-and-burn systems farmers tend certain “weeds” as they produce useful items, spices, or food. The weeding of a field, established after forest clearing, will eliminate seedlings and reduce stump regrowth of forest species. Depending on the cutting height and the frequency of weedicings, these species may eventually die. When the land reverts to fallow, these species will...
be unable to reestablish immediately. The fallow may thus consist only of weeds from the cropping phase.

C. Pest and Disease Suppression

Certain pests and diseases of crops may be avoided by choosing land that was under fallow for a long time. The efficiency of pest and disease avoidance depends on the distances between fields, the mobility of pests, and the vectors by which pests and diseases are spread. Soil-borne diseases might be affected by the dramatic soil chemical changes after the burn and the associated increase in soil pH and the changed microclimate. One important, yet not directly “slash-and-burn” related mechanism in pest and disease suppression is the complex mixture of crop species in slash-and-burn fields. Up to 40 species might be planted or tended within a single plot. Pests and disease buildup is not possible or is strongly delayed if the organisms are confined to a few isolated plants. The large number of pests and diseases with their different ecological requirements does not permit a general statement of the effects of slash-and-burn agriculture on their presence and severity in crops.

Radopholus similis is a plant-parasitic nematode, cosmopolitan in the tropics. It is the greatest cause of yield loss in bananas and plantains worldwide. Groundnut and maize are also hosts, as are Commelina benghalensis and Fleurya aestuans, minor components of weed and fallow communities in West and Central Africa. Radopholus similis does not survive in the soil for more than 6 months when host roots and corm pieces are absent. Thus, even a single year fallow can prevent infection of the following crop, if C. benghalensis and F. aestuans are uprooted and the new planting material is pest-free. The Maring people of Papua New Guinea plant banana suckers into their field during the burn. Heat would penetrate the outer layers of the banana sucker and would be likely to kill nematodes infecting the planting material. Other peoples roll banana suckers in ash before planting, believed both to kill nematodes and have a fertilizing effect.

Certain crop diseases are affected by the degree of shade in the plot. For example, Mycosphaerella fijiensis fungus is the causal agent of black sigatoka disease of plantains and bananas. Mycosphaerella musicola is the less virulent yellow sigatoka, dominant at higher altitudes. Workers in the Caribbean, Central America, and Africa have noted that when trees are eliminated from slash-and-burn fields, these diseases destroy larger areas of the leaves and yields are reduced by up to 40%.

Nowadays, short fallow slash-and-burn farmers have
SLASH-AND-BURN AGRICULTURE, EFFECTS OF

FIGURE 4 Forest clearing in southern Cameroon after burning. Note the large tree stumps and remaining trunks on the field. The wood may be piled and burned again or allowed to rot, depending on the crop to be planted. Photo by S. Hauser.

a limited choice of land, but in some areas they have access to pesticides. With a broad range of crops and more recently intensive vegetable production, the use of pesticides has become more common in many developing countries. Slash-and-burn farmers do resort to pesticide use. While any pesticide may be targeted against a specific pest or disease, it usually affects other organisms as well and indirectly organisms relying on the affected ones. With often little knowledge of the effects of pesticides on slash-and-burn agriculture, cultivators may destroy organisms that are important to the ecosystem’s functioning. Thurston (1992) provides a very comprehensive description of pest and disease control in such low-input cropping systems.

D. Management of Crop and Weed Residues

Some effects of slash-and-burn land preparation are aggravated by crop management. In the natural vegetation, in both long and short fallows, litterfall will accumulate on the soil surface, contributing to reduced soil temperature and providing nutrient and substrate inputs. During the cropping phase, some crops and weeds do produce litter. Yet mulching or simple retention of such residues is not a common practice amongst slash-and-burn farmers. In most cases, litter and weeds are removed, thrown outside the field, or piled on tree stumps. Retaining residues may create niches in which some of the fauna may survive the cropping phase. Yet slash-and-burn farmers are aware that mulch can also provide potential niches for pests.

E. Nutrient Export through Crop Harvesting

In addition to unproductive losses (combustion, wind, leaching, erosion, transfer to unavailable forms), the harvested crops contain nutrients that are exported from the field. As slash-and-burn crop yields are usually low, crop export losses are considered marginal compared to other losses, especially in situations in which old fallows are cleared. However, in short fallow systems, thus with more frequent cropping or prolonged cropping phases, the crop export losses will become more important in the overall nutrient balance of a field.

F. Effects of Prolonged Cropping Phases

Prolonged cropping phases permit the invasion of weeds suited for conditions in a clearing. Such weed species die out if the field is abandoned after a short
cropping phase. However, if the cropping phases are prolonged, competition from these invasive weeds further reduces the chances of recovery of the initial vegetation, particularly if the plot is burned repeatedly. Repeated burns kill tree stumps and seeds and can select for grasses (Poaceae), which establish from seeds but often propagate by rhizomes in the soil and thus are less affected than other species by the burn. The result is that the plot may revert to “arrested succession” grassland and be rendered unsuitable for cropping in the medium term. This may be a goal of a farmer who wishes to create pasture or may be an unintended consequence. In southeast Asia, prolonged cropping after forest clearing has facilitated the invasion of the land by Imperata cylindrica.

Generally, climax forest species have seeds that remain viable for a short time. Thus, extended cropping phases will compromise the regeneration potential of the fallow. However, where the intended purpose is to recrop after short fallow phases, the absence of viable seeds is academic, as they would have no possibility to establish on that site.

VII. EFFECTS DURING THE FALLOW PHASE

The effects of slash and burn on biodiversity in the fallow phase are strongly related to the conditions prevailing during the cropping phase. In long fallow systems with short cropping phases, the transition into a fallow dominated by pioneer trees is fast, without a phase in which arable weeds dominate. In the highlands of Papua New Guinea, the Maring people tend the re-sprouting trunks of the pioneer tree rokunt (Saurauia sp.) in their fields near the end of the cultivation phase, even though these compete with their crops. Seedlings of Casuarina algodon, a nitrogen-fixing tree, are transplanted to the field at this time to enhance soil fertility recovery during the fallow. Such reestablishment of the fallow vegetation will probably ensure the reestablishment of the soil micro-, meso-, and macrofauna. With an increased length of the cropping phase, fields tend to weediness as species not existing in the previous fallow might have established in the cultivation phase. These are likely to form at least part of the first fallow succession community. In such a situation, the establishment of a fallow vegetation as before the slash-and-burn cycle is delayed, in some cases impossible (Fig. 6). Biomass production and nutrient uptake and cycling are the primary processes of soil fertility restoration. The amounts of biomass and nutrients accumulated are a function of the fallow length. The minimum fallow length to completely restore soil fertility (Fig. 1, case 2) will vary depending on soil type, the types of weeds, crop pests and diseases present, and the growth rates, composition, and succession of the fallow species.

Aboveground biomass accumulation rates of forest fallows vary between sites and may vary between fallow ages. Most data from 4-year-old fallows range from 24 to 56 Mg ha\(^{-1}\), for 8-year-old fallows ranging from 40 to 150 Mg ha\(^{-1}\), and with older fallows up to 500 Mg ha\(^{-1}\). Biomass accumulates for at least the first 8 years. Some authors have reported that half the biomass in a 20-year-old fallow had already accumulated by year 4. In contrast, others showed that accumulation rates of approximately 5.5 Mg ha\(^{-1}\)yr\(^{-1}\) remained constant from the 4th to the 15th year.
Generally soil organic matter levels increase with time under fallow. During 4 years of bush fallow in southwest Nigeria, soil organic carbon content increased by 46% and total nitrogen by 99%. A 15- to 20-year-old fallow had soil organic matter levels of 3.8% compared with 1.1% in cultivated plots. In the drier region of northeast Brazil, a 10-year-old fallow was sufficient to restore soil carbon and nitrogen levels to those under undisturbed vegetation. Many authors have reported that soil organic matter and soil nitrogen concentrations do not increase substantially after 8 to 10 years of fallow.

Occasionally farmers use indicators to judge when to recrop a fallow. In southwest Cameroon, farmers stated that there is a definite minimum fallow time to regain fertility, expressed in years but judged by the type and maturity of the vegetation growing there. The Turumbu of Yangambi in eastern Zaire judged when to break fallow by the girth of umbrella trees (*Musanga cecropioides*) and the biomass of the undergrowth. They cleared when they could walk easily beneath the trees. These secondary considerations coincided with soil fertility regeneration (Jurion and Henry, 1969). In southwest Nigeria, farmers claim to use the appearance of the vegetation cover, the presence of particular plant species, and the presence of earthworm casts to decide whether a fallow is ready for recropping.

In Papua New Guinea, bird and reptile species diversities were low in young fallows but increased as succession proceeded. Reptiles were less affected than birds. The bird communities in young, grassy fallows were dominated by obligate granivores, whereas the forest supported specialist feeders (frugivores, nectarivores, and branch gleaners). Butterfly species richness showed a similar pattern.

In southern Cameroon, species richness of birds in *Chromolaena odorata* fallow was only a fifth of that in secondary or primary forest. Butterflies and termites had similar magnitudes of declines. Canopy ants and canopy beetles were completely absent in such fallows, given the near absence of trees. The species richness of flying beetles, leaf-litter ants, and soil nematodes was not different between forest and young fallow. Presumably, the young fallow, located within a mosaic of recovering fallows and secondary forest, provided sufficient food and conditions for their continued survival. For more details, see Lawton et al. (1998).

In southern Cameroon, earthworm activity, indexed by surface cast production declined rapidly after forest clearance, burning, and cropping and did not recover.
during 2 years of fallow. At the same site, cast produc-
tion decreased after clearing, burning, and cropping a
4-year-old Chromolaena odorata bush fallow, but did
recover again during the fallow phase. The change in
living conditions from forest to cropland was too drastic
for the existing earthworm community to maintain its
activity level. A new community structure comprising
species capable to cope with the new conditions could
not establish within 2 years of fallow. On the contrary,
the community in bush fallow had probably already
undergone some shifts thus recovering as soon as bush
conditions were reestablished. The response of termites
was different. In both forest and bush clearings, termite
densities dropped to about 1% of densities in the re-
tained forest. There was no recovery during the fallow
phase. The number of species in the cleared area was
only half as high as in the forest.

VIII. THE SCALE DEPENDENCE OF THE
EFFECTS CAUSED BY SLASH-AND-
BURN AGRICULTURE

A. Topographic Preferences
and Fragmentation

Slash-and-burn farmers have topographic preferences.
Long-fallow farmers often prefer gently sloping land as
this facilitates the burn. They avoid slopes that are too
steep as this increases the risk of soil erosion. Yet their
choice is limited by the type of terrain available. In
other parts of the tropics, farmers prefer the plateaux
and avoid cropping even on the gentle slopes. Valley
fringes, although difficult to access and waterlogged
during the rainy season, are under greater pressure from
farmers for dry season cropping of high value market
crops. In contrast, riverine forests and swamps are
avoided and therefore remain undisturbed.

Such preferences result in fragmentation. Ultimately,
in areas of high-population density, fragments of the
previously contiguous vegetation will only remain in
agriculturally unsuitable areas. Not all species of plants
and animals are represented in such fragments of vege-
tation and thus a recovery of the surroundings to the
initial condition is not possible. However, if slash and
burn in particular topographic positions occurs in low
population areas where systems have sufficiently long
fallow to restore the forest, it does not have major effects
on biodiversity.

B. Temporal and Spatial Scales

Traditional systems at low population densities, with
small fields (0.25–3 ha) in 1-year cropping phases and
long fallows of 20 years, use 5% of the land at any given
time (Fig. 7). Fields are surrounded by fallow land at
various stages of recovery. Even if the fields are cropped
for up to 4 years, not more than 20% of the land will
be cultivated. Recolonization of a cleared plot by species
from the surrounding fallow occurs rapidly after farm-
ers abandon the land. The rate of recolonization by
animals depends on the species' mobility. Plant species'
invansion or recolonization of the site depends on the
species' type of seed and the method of seed dispersion.
Unless species are extremely immobile, or in the case
of plants, if the fallow length is too short for species to
reach the reproductive phase, this system can maintain
the initial biodiversity.

With increasing human population densities, slash-
and-burn agriculture is practiced in shortened bush or
grass fallow cycles, with prolonged cropping phases.
Both processes lead inevitably to a larger proportion of
land being under cultivation. This occurs rapidly when
fallow are shorter than 10 years in systems cropping
for more than 1 year and when fallows are shorter than
5 years in systems cropping for 1 year (Fig. 7).

Topographic preferences and an increasing propor-
tion of land cropped at any given time lead to fragmenta-
tion, with areas of the initial or recovering vegetation
isolated by cropped land and early stages of fallow.
Spatial fragmentation can compromise (a) the ability
of the remaining undisturbed or recovering habitat frag-
ments to maintain diversity, which determines the po-
tential to recolonize the disturbed habitats, and (b) the
ability of species to be dispersed to and establish in
disturbed areas. Both aspects depend on the species

![Figure 7](image-url)
investigated. Generally, with increasing size of animals, spatial requirements for foraging and hunting increase and with increasing fragmentation their chances of survival diminish. A number of additional factors, such as mating habits and the ability to cross disturbed areas to reach other undisturbed fragments, will modify the likeness of the species’ maintenance. Smaller animals, largely invertebrates, might not be affected as severely by fragmentation as larger animals. On isolated trees retained in pastures in the Amazon, a considerable diversity of ants was found, with a large proportion of species usually found in the forest. Ant diversity was positively related to the epiphyte load of trees.

Fragmentation of forest by pastureland such as in the Amazon basin negatively affects certain tree species retained in the forest fragments. Some species need cross pollination to produce seeds. If fragments become too small or too distant from each other, the tree density might fall below a threshold required for pollen to reach other trees. Such a process, observed in mahogany (Swietenia macrophylla), reduces or eliminates seed production and may lead to local extinction.

The potential to regenerate degraded land to the initial vegetation and habitat depends on seed dispersion and mobility of animals. Seed dispersion by birds, bats, and monkeys has been shown to be important in disturbed and undisturbed areas. However, in Uganda, it was shown that bats and birds did not considerably contribute to the dispersal of forest species’ seeds to degraded grassland areas. The vast majority of seeds moved by these animals were of species dominantly found in disturbed areas or of species unable to establish in the grassland. Similarly, seed dispersion by gibbons confined to small forest fragments after large-scale slash-and-burn agriculture in Kalimantan was found to be uncertain as individuals were extremely hesitant or unable to cross cleared areas to reach other fragments.

The effects of slash-and-burn agriculture on biodiversity are various. However, generally, the shorter the cropping phase and the longer the fallow phase, and the smaller the proportion of cropped land, the greater the possibility of recovery of the previous vegetation. However, these criteria limit the productivity of slash-and-burn systems. To maintain their livelihoods, the choice for many farmers whose land allocation is fixed is either to prolong the cropping phase or to shorten the fallow phase. Clearly research should also compare and contrast the effects of these options on biodiversity, rather than focusing exclusively on comparisons with undisturbed ecosystems, which are increasingly unrepresentative.

IX. ALTERNATIVE SYSTEMS TO SLASH-AND-BURN AGRICULTURE

A. Plantations of Perennial Crops and Trees

Plantation systems include monocrop oil-palm, fast-growing timber species, coconut, rubber, tea, coffee, and cocoa. Most plantation systems are established by slash and burn. Timber plantations are often established in forests by selective felling or poisoning of undesired trees, underbrushing, and sometimes light burning of the understory. In such systems, the forest habitat is only slightly altered or reestablishes quickly. In southern Cameroon, timber plantations had similar levels of termite species richness to secondary forests and more species of leaf-litter ants than both secondary and near-primary forest (Lawton et al., 1998). However, the dominance of one tree species may affect the underbrush regrowth and the conditions for soil fauna through a narrower range of substrate. Many plantations, such as oil palm, coconut, rubber, tea, and other species that demand full sun light, are planted after clear-cut felling. In contrast, cocoa plantations in southern Bahia, Brazil, are established in old secondary or primary forest. Cocoa is shade tolerant, so while the lower-canopy trees and herbaceous components are slashed, many upper-canopy trees are retained. These include many valuable timber species, rosewood (Dalbergia nigra), pau Brasil (Caesalpinia Espinata), and cedro (Cedrela odorata), that have been nearly eliminated elsewhere by loggers. Consequently associated fauna have been maintained, including rare primate species, for example, the sagui (Callithrix Killari) and the lion tamarin (Leontopithecus rufus).

B. Pastures

In large parts of the Amazon basin and in Central America, forest is clearcut to establish pastures. While in tree and palm plantation systems a forest-like microclimate will be reestablished after some years, pastures will remain unshaded. Any forest regrowth is regularly slashed and in some situations pastures are regularly burned to reduce forest regrowth and promote grasses.

C. Improved Short Fallow Systems

Slash-and-burn agriculture is demanding on the natural resource base, irrespective of the fallow type (forest or
SLASH-AND-BURN AGRICULTURE, EFFECTS OF

short fallow. It requires large areas for cultivation as yields are low. In many situations the natural regrowth in short fallow systems is not capable to restore soil fertility because of the shortened period of time available in which the volunteer regrowth cannot produce sufficient biomass and accumulate sufficient amounts of nutrients. Improved fallow systems seek to replace the, usually undesired, volunteer regrowth by species that produce more biomass faster, contain more nutrients, or produce a type of biomass that is of a different physical structure than the natural regrowth. Such species are expected to outcompete the weed flora quickly and not to contribute to propagation or maintenance of crop pests and diseases. Often legumes fixing atmospheric nitrogen are used to augment the nitrogen resources of the soil. Depending on the plant type used for improved fallows (trees, shrubs, or herbaceous), the biomass is slashed and burned or only slashed and retained as mulch. Systems using herbaceous legumes, such as Pueraria phaseoloides, reduce the number of plant species surviving in the fallow rather drastically. However, depending on the crop planted after the fallow, the slashed mulch can be retained as a mulch layer, with minimal disturbance to soil invertebrates. The tree- and shrub-based systems are more likely to be managed in a similar way as any other slash-and-burn system, because the woody proportion of the biomass is more easily cleared by burning. However, these systems are more productive per unit area or restore soil fertility faster than systems relying on natural regrowth. Therefore less land is used for agricultural production and consequently more land remains undisturbed, retaining biodiversity.

See Also the Following Articles

AGRICULTURE, TRADITIONAL • CARBON CYCLE • DEFORESTATION AND LAND CLEARING • FIRES, ECOLOGICAL EFFECTS OF • INDIGENOUS PEOPLES, BIODIVERSITY AND • POVERTY AND BIODIVERSITY • SOIL BIOTA, SOIL SYSTEMS, AND PROCESSES

Bibliography


SOCIAL AND CULTURAL FACTORS

Jeffrey A. McNeely
IUCN—The World Conservation Union

I. Religion and Biodiversity
II. Language and Biodiversity
III. Indigenous Peoples, Resource Management, and Biodiversity
IV. Urban Versus Rural Perceptions
V. Conclusions

GLOSSARY

adaptation Any behavioral pattern that makes a human or a cultural group more fit to survive and reproduce in comparison with other individuals or cultures.
carrying capacity The maximum population that can be sustained indefinitely in a given area without changing the ecosystem in ways that will eventually reduce the sustainable population. This balance between population and resources is a dynamic one that involves changes in technology and other factors.
community A social unit consisting of members who are in direct interaction with each other and who have a collective identity, however defined. Relationships in such a group are principally primary rather than secondary in nature, and conformity to group norms is achieved mainly by peer pressure.
culture All capabilities and habits acquired by people as members of society.

indigenous peoples Those groups of human beings who share and preserve direct, everyday connections to their distinguishable distinctive cultural roots (even though they may have willingly migrated or been forcefully moved from their homelands, including moves to cities). These indigenous peoples consciously or unconsciously draw on specific knowledge and strategies developed and tested by past generations to address current problems. Cultural traditions emerge and are maintained in a dynamic process of creative invention and reinvention, as well as borrowing and adaptation from other subgroups and cultures.
language A system of conventional spoken or written symbols that enable human beings to communicate as members of a social group and participants in its culture.
management The efforts of humans to select, plan, organize, and implement programs designed to achieve specified goals; activities can range from protective measures to ensure that human influences on natural resources are minimized to greater interventions required to maintain diversity, install facilities, control populations, and eliminate unwanted elements.
myth A story, usually of unspecified origin and at least partially traditional, that ostensibly relates actual events to explain some practice, belief, institution,
or natural phenomenon and that is frequently associated with religious beliefs; myths provide models for human behavior and often emphasize behaviors that support conservation of biodiversity. Religions order the spiritual and physical relationships of people with other humans and with their environment. Approaches to conserving biodiversity that are based on cultural and religious values are often much more sustainable than those based only on legislation or regulation. While religions offer important guidance on the relationship between people and the rest of biodiversity, all religions are characterized by a wide gap between their philosophy and the practices of the people who have accepted the philosophy.

Religion frequently has had the effect of preventing excess human demands from outstripping the environmental resources that are required to support them. For example, in Roman times, high priests dictated that the well-being of humanity depended on thousands of animals being sacrificed every year, thereby keeping the size of herds within the carrying capacity of the grazing lands. More fundamentally, religions provide a complex holistic view of how to use natural resources, supporting this knowledge with an ethical perspective that is built on the implicit social contract that enables communities to function.

All of the world’s major religions are today sensitive to the importance of biodiversity, though of course their historical writings do not use today’s conservation vocabulary. The following is a brief summary of some major religious belief systems and how they relate to modern biodiversity concerns.

Animism is a term applied to a wide variety of beliefs that focus on numerous spiritual beings concerned with human affairs and capable of either helping or harming people. Animism is a view of the world found among many traditional peoples who conceptualize an intrinsic spiritual connection between humans and nature. This connection has been described by Harvard University biologist Edward O. Wilson as “biophilia,” an innate human need for contact with a diversity of life-forms. For example, in the North American desert, the Yaqui Indians believe that close communication exists among the plants, animals, birds, fish, rocks, and springs that inhabit the Sonoran Desert. Many traditional approaches to conservation are supported by religious beliefs, often based on various kinds of animism that have the effect of fostering respect for plants and animals. In many parts of the world, people have established sacred sites on the basis of inherent spiritual or religious significance. Such sacred sites, based ultimately on animistic beliefs, are often sanctuaries for biodiversity. They may well survive substantial cultural changes; for example, the Parliament of Kenya (a very modern institution) voted in 1990 to protect all of the country’s remaining sacred forests, known as kaya. Many animistic systems of belief are accompanied by the idea of taboo, that which is forbidden; breaking a taboo can bring sanctions such as illness, social ostracism, or even death. Taboos often applies to certain sets of natural resources that are particularly vulnerable to overexploitation. Animism is also often associated with totemism, a complex of ideas and practices based on the belief in a mystical relationship (often kinship) between people and certain animals or plants; these relationships often include reverential and genealogical relationships between social groups or individuals and the totems. Totems normally are associated with taboos of avoidance, so an Amazonian Indian hunter within a social group that has the peccary as a totem may be forbidden from hunting peccaries. Thus totemism also helps to restrict exploitation of harvestable resources.

Buddhism, with a total of about 300 million prac-
titioners found in many Asian countries, teaches that a behavior has a natural relationship to its resulting consequences in the physical world. All Buddhist teachings and practice come under the heading of Dharma, which means truth and the path to truth. Based on the teachings of the Buddha, who lived in Nepal more than 2500 years ago, Buddhism teaches that people are responsible for their actions and go through a cycle of rebirths before finally reaching nirvana. Right actions lead to progress toward nirvana while negative actions, such as killing animals, lead to regression from that goal. Committted to the ideal of nonviolence, Buddhism also attaches great importance to wildlife and the protection of biodiversity (though of course not using such modern terms in its ancient teachings). Respect for life in the natural world is essential, and by living simply one can be in harmony with other creatures and learn to appreciate the interconnectedness of all that lives. This is not to claim that nature is unchanging; on the contrary, Buddhism recognizes change as the very essence of nature. The Buddha taught that all things are interrelated and do not have an autonomous existence, so the health of the whole is inseparably linked with the health of the parts and the health of the parts is inseparably linked with the whole.

Christianity, with some 1.6 billion members, is the dominant religion in Europe, sub-Saharan Africa, the Pacific, and the western hemisphere. It has numerous denominations, all based on the belief in a single all-powerful God who created nothing unnecessarily and omitted nothing that is necessary. Christians believe that Jesus Christ was the son of God and a great teacher whose message was that all of creation is the loving action of God, who continues to care for all aspects of existence. The very nature of biodiversity considered in itself, without regard to humanity's convenience or inconvenience, is considered to give glory to the Christian God. Christianity teaches in effect that humanity may not disorder biodiversity or destroy God's creations, at the risk of destroying itself, in the Holy Bible, Ecclesiastes 3:19 says: "For that which befalleth the sons of men befalleth beasts... as the one dieth, so the other... a man hath no preeminence above a beast." Human acts of irresponsibility toward creatures are contrary to the divine wisdom, which sustains and gives purpose to the interdependent harmony of the universe. Modern Christian teachings are that humanity, both individually and collectively, should perceive the natural order as a sign and sacrament of God, recognizing that all creatures and objects have a unique place in God's creation. On the other hand, Christianity also recognizes a special role of humanity within creation. The Holy Bible contains Christianity's main teachings. It provides strong support to the protection of natural resources, including passages mandating the preservation of fruit trees (Deuteronomy 20:19, Genesis 19:23–25), agricultural lands (Leviticus 25:2–4), and wildlife (Deuteronomy 22:6–7, Genesis 9). The Bible often refers to the impressive intelligence of wild creatures, such as in Jeremiah 8:7–8, Proverbs 6:6–8 and 30:24–28, Numbers 22:22–35, and Isaiah 1:3.

Hinduism, the dominant religion in India with about 700 million followers, teaches the all-encompassing sovereignty of the divine, manifesting itself in a graded scale of evolution. While the human race is currently at the top of the evolutionary pyramid, it is not seen as something apart from the earth and its biodiversity. Hinduism is permeated by a reverence for life and an awareness that the great forces of nature—earth, sky, air, water, and fire—as well as biodiversity are all bound together within the great rhythms of nature. Hindus recognize that all lives play their fixed roles and function together so that no link in the chain of life is lost. The divine is not exterior to creation but expresses itself through natural phenomena. Hindus consider at least some forests and groves as sacred and associate various plants and animals with gods and goddesses in the Hindu pantheon. The most important teachings of Hinduism are contained in the Bhagavad Gita, a dialogue between Sri Krishna and Arjuna, which has a clear description of ecology as a cycle of life dependent on everything from bacteria to birds. Like animists, Hindus believe that all plants and animals have souls, and that people must do penance even for killing plants and animals for food. Hinduism involves sacrifice, forms of ritual worship that are designed to protect life through reinvigorating the powers that sustain the world by securing cosmic stability and social order. Hindus recognize certain rivers and mountains as sacred, because they give and sustain life.

Islam, with about a billion adherents, is the dominant religion in North Africa, the Middle East, and many Asian countries. Its teachings are very similar to those of Christianity in relation to biodiversity. The entire universe is God's creation; Allah makes the waters flow upon the earth, upholds the heaven, makes the rain fall, keeps the boundaries between day and night, creates all of biodiversity, and gives it the means to multiply. Tawheed is the principle of Oneness of Allah, teaching that "there is no God but God and Mohammed is His messenger." This testifies to the unity of all creation and the interlocking grid of the natural order of which people intrinsically are parts. The Holy Koran provides a set of principles that define the relationship of man...
to God and of God to the environment in its totality. However, humans are considered a very special creation because they alone have the power to reason and think, giving them the potential to do great good, or great harm. For Islam, the role of people on earth is that of a Khalifa, or trustee of God, so humans are entrusted with the safekeeping of earth and its biodiversity. People are answerable for their actions, including maintaining the unity of Allah’s creation, the integrity of the earth, and its biodiversity. The Islamic relationship to biodiversity is that each generation uses and makes the best use of biological resources, without compromising the interests of future generations. The Koran says, “With it we have produced diverse pairs of plants each separate from the others. Eat (for yourselves) and pasture your cattle; verily, in this are signs for men endowed with understanding” (Sura 20 Aya 53). While humans have the right to utilize and subjugate natural resources, this involves a commitment to conserve them both quantitatively and qualitatively. In the Koran, God says, “There is not an animal (that lives) on the Earth, nor a being that flies on its wings, but (forms part of) communities like you” (Sura 13 Aya 15). Thus Islam looks upon biodiversity as an expression of God’s wisdom and omnipotence and as support for human development. This leads to the necessity of conserving and developing biodiversity for its own sake as well as for the benefit of humans. And the Prophet Mohammed said, “There is a reward in doing good to every living thing.” Such beliefs can lead to biodiversity being conserved. For example, in Pakistan, representatives of original tree species still persist in old Moslem graveyards, because of a taboo against cutting these trees. And the only surviving population of a freshwater turtle, Trionyx nigerians, survives in a sacred pond dedicated to a Moslem saint in Bangladesh.

Jainism is one of the oldest living religions, beginning in India at least 2800 years ago. Jains believe that all living beings have an individual soul (jiva), which occupies the body until it dies, then leaves the body and immediately takes birth in another. Attaining nirvana and thereby terminating this cycle of birth and death is the goal of Jain practice, much like Buddhism. Jainism is based on the principle of ahimsa (nonviolence) toward human beings and all of nature. Jains teach that no human quality is more subtle than nonviolence and no virtue greater than reverence for life. While biodiversity often is affected negatively by people, the intention to harm is what makes an action violent, and without violent thought no violent action is recognized. Jain cosmology recognizes the fundamental natural phenomenon of symbiosis or mutual dependence, with all aspects of nature belonging together and bound in a physical as well as a metaphysical relationship—an ancient perspective that is reflected in modern ideas about biodiversity.

Judaism, which originated in the Middle East but whose 18 million practitioners today are thinly spread around the world, teaches that God created the world, making order out of primal chaos. The sun, the moon, the stars, plants, animals, and ultimately humanity were each created with a rightful and necessary place in the universe. The encounter of God and man in nature is conceived in Judaism as a seamless web with man as the leader and custodian of the natural world. Judaism maintains that the earth is the arena that God created for man, half beast and half angel, to prove that he could behave as a moral being. Judaism views God as the divine “Giver of the Torah,” which sets out a series of ethical obligations including many relevant to biodiversity. It teaches that the relationship between man and nature is one of ownership, though limited by good sense; man was commanded to behave toward the rest of biodiversity with justice and compassion. But humanity inevitably lives in tension between its power and the limits set by conscience. The Bible contains much environmental wisdom, for example allowing agricultural fields to go fallow one year out of seven (Leviticus 25:1–5). The Bible also warns that even in time of war, it is forbidden to destroy fruit-bearing trees of one’s enemies (Deuteronomy 20:19).

Shinto is the system of indigenous religious beliefs and practices of Japan, first appearing in written form around 1400 years ago, after Buddhism had been introduced into Japan. Shinto is based on beliefs concerning the nature and attributes of Kami, sacred power, which is found in all individual things. Shinto temples are often established on sites that have particular spiritual integrity and force, often with large groves of trees (totaling nearly 120,000 ha in Japan). Shinto is based on numerous sacred texts, collectively known as Shiten. Shinto is strongly based in rural agricultural practices, involving various ceremonies and festivals that guide the relationship between people and nature.

Sikhism, a modern religion that began in India the late fifteenth century, builds on the message of the oneness of a universe created by an Almighty God, who is master of all forms in the universe and the source of the birth, life, and death of all beings. God has created the natural beauty, which exists and can be found in all of biodiversity. Without his hukum (order) nothing exists, changes, or develops, but having brought the world into being, God sustains, nourishes, and protects it. Biodiversity exists under God’s command and with God’s grace. Sikhism teaches against a life of conspicuous consumption, emphasizing mastery over the self.
II. LANGUAGE AND BIODIVERSITY

As an essential part of human cultures, languages provide the categories to conceive the natural and social world. A species of plant or animal that is important to a community gets named, and by learning that name, the people within the culture also learn what is vital to know in their natural environment. Thus language helps to organize the world and frees energy for other tasks, using words like pegs on which to hang the meanings stored in the storehouse of the human mind and providing a framework that binds together the details into a meaningful whole. Words for objects and phenomena help us learn the connotations, associations, emotions, and value judgments of a culture. Thus knowledge, beliefs, and values are linguistically encoded, thereby helping to promote a diversity of adaptive ideas and support various forms of biodiversity.

Many scientists argue that the preservation of the world's linguistic diversity (totaling more than 6000 languages) and the distinct forms of local knowledge that local languages encode must be incorporated as an essential goal in bioculturally oriented diversity conservation programs. Indeed, the diversity of human languages may be the best available indicator of human cultural diversity.

Many indigenous peoples believe that the external reality or environment is no different than the description of this reality or environment in a linguistic ecology. Variation in humans is adaptive, and language helps to support cultural diversity in humans. Human culture is a powerful adaptation tool and language both enables and conveys much cultural behavior.

Many countries with high biodiversity also have high cultural diversity, as indicated by their large number of languages. Languages have multiplied and thrived in places where natural selection has produced a rich variety of landscapes, animals, and plants. Researchers have found that linguistic diversity in sixteenth-century North America was greatest in areas with the greatest diversity of habitats, irrespective of the latitude. Ten of the 12 "megadiversity countries" are among the top 25 countries for endemic languages. This correlation between biodiversity and linguistic diversity may have been fostered by a process of coevolution of small-scale human groups with their local ecosystems, in which humans interacted closely with the environment, modified it as they adapted to it, and acquired intimate knowledge of it. This knowledge was encoded and transmitted through language, supporting the contention of some linguists that life in a particular human environment depends on the ability of people to talk about it.

In many parts of the world, anthropologists have found a strong correlation between biodiversity, linguistic diversity, and ethnobiological knowledge. More generally, landscapes are anthropogenic not only in the sense that they are physically modified by human intervention, but also because they are symbolically
brought into the sphere of human communication by language—by the words, expressions, stories, legends, and songs that encode and convey human relationships with the environment and that inscribe the history of those relationships onto the land. When people name places, they are identifying where things happen within the local environment, thus providing an entry in a mental encyclopedia that helps to describe the ecological niche occupied by the local peoples.

For example, certain American Indian groups in the Pacific Northwest have an extensive vocabulary for talking about salmon and other fish, vegetation, streams, and trails, along with extensive development of grammatical devices that express the direction, distances, and relative positionings of these valuable features. Thus language helps to enable these seminomadic hunter-gatherers to be very clear about where they were and where they needed to go to find food and other necessities. Many cultures in the far north have numerous names for various types of snow, and pastoral peoples have detailed terms for various types of livestock. The richness of language thereby reflects the many ways people relate to their environment.

III. INDIGENOUS PEOPLES, RESOURCE MANAGEMENT, AND BIODIVERSITY

Humans earn a living from the earth and its biodiversity in four main ways: hunting and gathering, pastoralism, shifting cultivation, and permanent agriculture. Only the latter is sufficiently productive to enable large cities to flourish, and the very high productivity enabled by irrigation may have been needed for civilization. Historical evidence indicates that traditional activities related to agriculture, fishing, and livestock husbandry sometimes have led to sustainable systems that were in the self-interest of the people involved, at least at their current level of population and technology. But all four forms of using biological resources carry with them the dangers of overexploitation. Broadly speaking, the combination of technology and population growth potential requires human cultures to develop effective ways of living in a sort of balance with the resources available with a given technology. The extinction of numerous cultures over time indicates how frequently the limits have been exceeded. On the other hand, many cultures have developed belief systems and practices that served to limit overexploitation and to foster greater biodiversity (even if these benefits sometimes were unintended).

For most of our history as a species, humans were hunters and gatherers, living from harvesting a very wide range of plants and animals; a few such groups still survive today, usually where agriculture is impossible or impractical. Hunter-gatherers were usually nomadic and fit ecologically into their habitat in much the same way as other omnivorous species. An important discovery that interrupted this presumed harmony and may have had major ecological impacts on biodiversity was the deliberate use of fire for driving game and clearing forest. This cultural activity produced major modification in vegetation patterns and fauna distribution in at least some regions. The invention and use of spears, bows and arrows, harpoons, and nets also must have affected the interrelationships between human beings and populations of other animal species. The effects of culture on humans themselves in the hunter-gatherer phase included impacts on techniques of food gathering, especially hunting. Nevertheless, with the exception of the use of fire in some situations, it is difficult to accurately determine the impacts on biodiversity of these changes.

Given the very great variability in approaches to resource management taken by various groups of hunting peoples, it is not surprising that various hypotheses have been developed to explain why some groups of indigenous peoples appear to have been able to live in better balance with their resources than other groups. A common observation is that ethnic groups that have lived in an area for extensive periods of time and appear most dependent on their locally available resources are most likely to have developed sustainable forms of resource use. Many of the population processes that link human hunters and their prey occur over timescales that elude both ethnographic and archaeological fieldwork. The critical factor is the distinctive features of the foraging economy and their outcomes for resource population ecology, as a matter of behavior with practical consequences.

Fishing is a specialized form of hunting and gathering. People who depend primarily on fish for their protein typically have developed both technology and management techniques to ensure sustainable yields. For example, Amazonian Indians may avoid certain parts of the river during breeding seasons; eastern Indonesian fishermen have complex systems of taboos, known as sasi, which control how and when the fishing grounds are to be used; and Polynesian fishermen had various taboos that served the function of managing their fisheries. The development of domestication of plants and animals about 10,000 years ago marked the beginning of an entirely new era in the interplay between human
Society and biological systems. The controlled breeding of plants and animals, using a small sample of the world’s biodiversity, established a more reliable and expandable resource base than subsistence hunting and gathering, enabling humans to reduce substantially the space required for sustaining each individual and allowing the human population to increase. All early farming activities consisted of redistributing plant and animal species in a given area by deliberately increasing the local concentrations of the species humans valued and decreasing concentrations of species that competed with the favored ones. In terms of biodiversity, the selective breeding of species also expanded the range of human impact from habitats and species to genes. The 40 or so species of animals that have been domesticated have often greatly increased their genetic diversity, range, and populations due to human management, as have the 100 or so main domesticated plant species and the thousands of other plants selected for use as food, spices, medicine, decoration, and construction. But the gains for these species typically have been at the expense of other species which people found less easy to mold to their vision of the desirable.

Domestication of livestock apparently arose after domestication of plants and remains the most significant use of arid and semiarid lands by people. The dangers of overgrazing have led to many beliefs about human relationships, livestock management, and range management. For example, many Middle Eastern pastoralists had complex ways of sharing grazing lands, including the idea of hema, a traditional grazing reserve. At least some species of trees typically are considered sacred by pastoral peoples and are not destroyed except in times of dire need; conservation of such vegetation thereby provides an emergency resource reserve (as well as conserving biodiversity).

Fallowing was the agricultural technique originally followed in most places; it remains a component of the shifting cultivation systems in many parts of Asia, Africa, and Latin America today. Such systems often use fire as a means of clearing the land, so they are sometimes called “slash-and-burn” systems. Such a system involves cultivating a plot of land—known as a “swidden”—for a few years and leaving it to develop into successional stages for a much longer period. This process creates a mosaic of vegetation, much of which is useful for foods, medicines, building materials, and other useful products (such as dyes, colorings, ceremonial objects, etc.). Many species of wildlife, including birds and terrestrial mammals, are attracted to the fallow swiddens because of their high productivity of nutritious vegetation. Furthermore, part of the land was often left out of cultivation. Thus in northeast Indian states like Manipur, as much as 10 to 30% of the land was permanently maintained under natural mature vegetation in the form of sacred groves. Historically, this would have ensured the persistence of almost all the natural elements of biodiversity, coupled with stimulation of overall productivity by favoring faster-growing early successional species in the patchwork of successional stages covering 70 to 90% of the land. Selection for adaptation to highly heterogeneous local environments also promoted considerable genetic variation in the cultivated species.

As an example of how shifting cultivation affects biodiversity in the Ecuadorian portion of the Amazon forest, Runa Indian swiddens resemble agroforestry systems rather than the slash and burn that merely results in temporary clearings in the forest canopy. Compared to unmanaged fallows, management actually increases species diversity in 5-year-old fallows. Between 14 and 35% of this enhanced species diversity is attributed to direct planting and production of secondary species. Thus Runa agroforestry can be seen as a low-intensity succession management system that alters forest composition and structure in the long term.

Among the most diverse of agricultural systems known are the home gardens in the humid tropics of tropical Asia, the result of long historical development of technology designed to meet the needs of local agricultural communities. In West Java, the typical home garden appears as a crowded assemblage of trees, shrubs, climbers, herbs, and creeping plants that are used for fruit, vegetables, starchy food crops, spices, ornamentals, medicines, fodder, fuel, and building material, and involve over a hundred species useful to people. The Kantu of Kalimantan, Indonesia, plant at least 44 varieties in one area with an average of 17 per household. A gene pool of potatoes of some 3000 varieties representing 8 species is traditionally under cultivation in the Andes. In Papua New Guinea, as many as 3000 varieties of sweet potato are under cultivation, with as many as 20 varieties being planted in a single garden. Thus indigenous peoples often maintain high biodiversity, at least among those species useful to them.

A. Traditional Belief Systems and Resource Management

The way of life of the world’s tribal or indigenous peoples depends closely on biodiversity, often using cultural and religious beliefs to avoid disastrous overexploitation. Notions of sustainability of use are inherent...
in the value systems of many traditional societies, usually manifested in some notion of intergenerational equity. For example, the Iroquois tribe in North America would plan for the seventh generation when making their decisions, the life span, incidentally, of the dominant tree in their region. And the Koyukon Indians believe that future events will depend on the way people behave today and that the world can be nurtured by prudent use or harmed by unrestrained abuse, but equally important, the natural world will respond to gestures of respect given by those who recognize its sensitivity and awareness and humble themselves to its power. Plains Indians generally believed that their relations with other species were regulated by expectations and obligations similar to those that governed relations between kin or allies; such relations could vary from beneficial to harmful and could be mediated by ritual specialists.

In Colombia, the forests of the northwest Amazon basin harbor the world’s most diverse array of plants and animals. This region is often considered part of the world’s greatest remaining tropical wilderness. But the Tukano people who live there perceive their “wilderness” environment to be anthropogenic, transformed and structured in the past by the symbolic meaning their ancestors gave to resources and the knowledge they obtained about the plants and animals that enabled people to survive. Their forest is a system of resources in which the energy produced is directly proportional to the amount of energy it receives, a very modern perspective. They know that they cannot harvest more than the forest can produce, and they apply sophisticated knowledge of individual species and their uses. Their myths tell of animal species that were punished by the spirits for indulging in gluttony, boastfulness, improvidence, and aggressiveness. These myths serve as object lessons to human society, in which animals are metaphors for survival. By analyzing animal behavior, the Tukano find an order in the physical world within which human activities can be adjusted.

The constraints on harvesting may involve protection to keystone species that may support the persistence of a range of other species. Thus fig trees belonging to the genus Ficus are recognized as important sources of fleshy fruits available in seasons when no other species are producing such fruits in the tropical forest communities. Ficus thereby promote persistence of a number of insect, bird, bat, squirrel, and primate species for which they serve as a critical resource in a period of fruit shortage. All species of Ficus are even today to some extent protected as sacred trees through much of tropical Asia and Africa, where the local communities are aware of this ecological role. In other parts of the world, other species of useful trees may also be sacred, such as date palms in desert areas, baobab trees in Africa, palmyra in Brazil, and kapok among the Mayas. Harvesting restraints also include protection of critical stages in the life history of species that are especially vulnerable to overharvest. Thus a nomadic hunting tribe of Western India has the tradition of releasing any pregnant does or fawns of antelope or deer caught in their snares, and egrets, storks, herons, pelicans, ibises, and cormorants at their colonial nesting colonies are given immunity from hunting over most of India, although these birds are hunted in the nonbreeding season. In many Asian villages, fruit bats are not hunted at their daytime roosts but may be killed at a distance from the roost during the night. Numerous such examples can be found in all parts of the world.

Some traditional communities have developed detailed resource regulation systems. For example, many traditional communities—whether “indigenous” or “tribal” like the Taran’s Dayaks of West Kalimantan or the nontribal, such as the ribereños of Amazonian Peru—establish community reserves and formulate rules about how the species therein can be exploited, with the expressed intention of preserving these resources for the existing community and for future generations. Therefore, in some cases reserves may be protected by religious sanctions and in others by a more overtly “social” contract and enforced by strong economic and social sanctions; combinations of approaches are also found.

Biological diversity may well have been higher in rural areas in former times, when large numbers of farmers of many different cultures had long-term stakes in the land they farmed and had control over their own technology. These historical systems of land management were highly variable, following a range of different rules to take into account specific attributes of the physical systems within which they were found, cultural views of the world, and the economic and political relationships that existed in the setting. Despite their great diversity, such systems often had characteristics such as clearly defined boundaries, specific rules on the harvesting of different products, involvement of the affected people in these collective choices, a system of monitoring the use of resources, cultural sanctions for those who violated the operational rules, inexpensive local mechanisms for resolving conflict, and ways of organizing these activities so that different types of decisions were taken at different levels.
IV. URBAN VERSUS RURAL PERCEPTIONS

About half the world’s population today lives in urban settings far removed from the agricultural systems on which their lives depend. While rural peoples typically have a very good practical understanding of biodiversity, those living in cities increasingly have no more than a theoretical understanding, perhaps reinforced by an occasional visit to a national park or by a television program about nature. The growing numbers of nature-based films, protected areas, and forms of tourism based on idealized visions of nature indicate that urban people still feel a need to connect to biodiversity. Of course, different cultures have different perceptions, and the diversity of human societies is reflected in the diversity of perceptions urban people have of their environment. Any classification of such cultural values will necessarily mask some of this diversity, but rural people who live in close contact with the realities of nature typically have very different attitudes about biodiversity from urban people who have only a distant relationship with nature. Examples of these contrasting attitudes are presented in Table I.

These attitudes are ideal types, and will not necessarily be found in their pure forms in any rural or urban community. Rather, societies both over time and over space can be placed on a continuum between these opposites, as human attitudes change due to both internal social dynamics and external influences. However, it is clear that such attitudes greatly influence how humans use natural resources and therefore the impacts they have on biodiversity.

In marked contrast to the rural areas where people still live in a close relationship with biodiversity, urban dwellers are able to gain the benefits from the biodiversity of the entire globe. Their connections through trade and communications enable urban dwellers to have rapid and unprecedented access to biodiversity. Ironically, the expanding access to more biodiversity has tended to weaken the feedback between human welfare and the way biological resources are managed; the urban dwellers typically have only the very faintest notion of how their consumption might have impacts on biodiversity in the countryside of a distant country.

V. CONCLUSIONS

Establishing a connection between specific cultural practices and conservation or enhancement of biological diversity is by no means a simple matter, for the overtly declared purpose of a practice that seems to help conserve biological diversity may in fact be quite different. Thus, in South Asia many sacred ponds have helped conserve the indigenous fish fauna. But people may leave these ponds alone out of respect for some deities, not with an expressly declared purpose of conserving fish diversity. It is then quite possible that many cultural or religious practices that seem to promote conservation may have originated from different motivations, while others declared as promoting conservation may in reality achieve something very different.

Traditional societies have often protected parts of the natural landscape they live in, or left untouched some of its elements. Most such societies, for instance, have considered certain sites as sacred, where most or all human activities are prohibited. Most societies have also considered certain species as sacred, with elaborate myths and folk tales about how humans originated from such species, or how these species are incarnations of gods and deities, or in some way associated with them, or how they obtained magical powers. These could be wild species that are left undisturbed (and therefore have survived in even the most densely populated and radically altered human landscapes, such as rhesus and bonnet macaques in Indian cities) or domesticated plants and animals whose utilitarian value is intertwined with spiritual values (such as the cow in Hinduism). Such belief systems can have the effect (although perhaps unintentional) of regulating resource use.

<table>
<thead>
<tr>
<th>Rural attitudes</th>
<th>Urban attitudes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans as one part of nature's biodiversity</td>
<td>Humans not part of biodiversity</td>
</tr>
<tr>
<td>All living creatures considered equal</td>
<td>Humans superior to all creation</td>
</tr>
<tr>
<td>Culture/nature as a continuum (no such thing as wilderness)</td>
<td>Culture/civilization human, and nature considered wild</td>
</tr>
<tr>
<td>Natural way is right, and human activities should be molded along nature's rhythms</td>
<td>Human technology is superior, needs to mold nature to suit human needs</td>
</tr>
<tr>
<td>Biodiversity has an integral set of multiple values (cultural, spiritual, material)</td>
<td>Biodiversity is predominantly a material (economic) resource</td>
</tr>
</tbody>
</table>
A rich literature of traditional conservation practices indicates that they have been a common feature for many cultures over many years. This literature indicates that such practices serve group interests of communities and remain viable only so long as (a) local communities continue substantial levels of dependence on resources harvested from their immediate vicinity, (b) local communities have full control over the local resource base, and (c) local communities retain a sufficiently high level of internal cohesion. In most cases, these conditions are no longer fulfilled when outside state or corporate bodies establish control over natural resources.

When local people are part of a local ecosystem, their behavior directly affects their own survival. But cultural mechanisms that have been developed as adaptations to the environment over tens or hundreds of generations are quickly cast aside when trade or new technology frees people from traditional ecological constraints, changing them from "ecosystem people," who are adapted to their local ecosystem, into "biosphere people," who often live in cities and can draw from the resources of the entire world.

While changes in traditional attitudes toward biodiversity can be a result of internal dynamics (e.g., population increase), it is perhaps more often a result of outside influences: interaction with a modern culture, intrusion of the market, a technical innovation imported from another culture, and so on. Some have argued that traditional resource use patterns may be sustainable only under conditions of low population density, abundant land, simple technology, and limited involvement with a market economy. When confronted with market pressures, higher population densities, new technologies, and increased opportunities, few indigenous peoples appear able to maintain the integrity of their traditional methods, even when these are reinforced by religious beliefs.

See Also the Following Articles

- AESTHETIC FACTORS
- TRADITIONAL FACTORS
- ETHNOBIOLOGY AND ETHNOECOLOGY
- HISTORICAL AWARENESS OF BIODIVERSITY
- HUNTER-GATHERER SOCIETIES
- ECOLOGICAL IMPACT OF INDIGENOUS PEOPLES, BIODIVERSITY AND LITERARY PERSPECTIVES ON BIODIVERSITY

Bibliography


SOCIAL BEHAVIOR
Daniel I. Rubenstein
Princeton University

I. Introduction
II. From Social Behavior to Social Organization: Patterns and Mechanisms of Formation
III. From Social Organization to Ecological Impacts

GLOSSARY
affiliative behavior Behavior that is supportive and brings individuals together.
agonistic behavior Behavior that is aggressive, threatening, combative, and submissive.
altruism Behavior that enhances the recipient’s reproduction while reducing the actor’s.
musth Annual reproductive period in male elephants characterized by extreme aggressiveness and secretions from temporal glands.
polyandry Mating system in which a female mates with more than one male.
polygyny Mating system in which a male mates with more than one female.
protandry Breeding system in which individuals change sex from male to female.
protogyny Breeding system in which individuals change sex from female to male.

SOCIAL BEHAVIOR characterizes the interactions that occur among individuals. These can be aggressive, mutualistic, cooperative, altruistic, and parental. When individuals interact repeatedly, social relationships develop and these can form among strangers, relatives, members of the same or opposite sex, and members of the same or different generations. Sets of consistent social relationships produce social systems or social organizations that can be variations on monogamous or polygamous themes and involve various types of helpers. The nature of any social system is ultimately determined by ecological and social circumstances, demography, and kinship.

I. INTRODUCTION
No animals are completely asocial. All must come together at some point to breed. Few, however, are highly social coming together and interacting repeatedly. For those that do, social interactions—both affiliative and agonistic—occur sometimes with relatives, sometimes with strangers, sometimes with members of the same sex, sometimes with members of the other sex; and sometimes with members of the same generation while at other times with members of other generations. The interactions themselves can be aggressive, cooperative, or even altruistic and can develop into strong relationships among particular individuals. Depending on the nature of these relationships and with whom they form, a variety of social systems can develop. Some may be made up mostly of kin or mostly of nonkin, some may be based on territorial separation or on the aggregation...
of competitors, some will exhibit monogamous as opposed to polygamous relationships between the sexes, and some will rely on the help of nonmates in the rearing of offspring. This seemingly bewildering pattern of social diversity is shaped ultimately by ecological circumstances and patterns of demography. But the very nature of a population’s social structure itself will in turn affect the population’s demography and its place in the biological community. It is these dual connections between social activities and the environment and the resulting feedbacks that mandates understanding both the causes and consequences of animal sociality so that effective and efficient management or conservation policies can be developed for protecting endangered species and for preserving biodiversity and ecosystem integrity.

II. FROM SOCIAL BEHAVIOR TO SOCIAL ORGANIZATION: PATTERNS AND MECHANISMS OF FORMATION

Social animals form groups. While some are temporary, others are more permanent. Given that animals in groups incur automatic costs of increased disease and parasite transmission as well as intensified competition, groups will only form when there are sufficiently large benefits to offset these costs. Benefits largely come in three forms. First, animals can develop forms of social behavior specific to stable groups that compensate them for the costs of group living. One such example is forming mutual grooming partnerships as do olive baboons (*Papio anubis*) to lower disease and parasite transmission. Second, by forming groups animals can enhance foraging by being better able to find, acquire, or defend food. Examples include colonial cliff swallows (*Hirundo pyrrhonota*) that transfer information about the locations of rich but ephemeral feeding sites or troops of monkeys who drive smaller troops away from feeding trees. And third, animals in groups can reduce their risk of being preyed upon by either increasing the likelihood of detecting predators, diluting their personal risks, or by decreasing the likelihood that predators can make a kill; confusion and cooperative defense are mechanisms that provide such antipredator benefits. Examples include the scattering of fish in schools, or the gathering of young inside a ring of adult musk oxen (*Ovibos moschatus*) facing outwards toward approaching predators with upturned horns.

Depending on the nature of the social relationships that develop, groups take on particular organizational forms. As Fig. 1 illustrates, these relationships are shaped by the features of an individual’s physical and social environment. Particular distributions of food, water (bottom-up factors) and predators (top-down factors), in conjunction with the physiological demands of individuals differing in body sizes or reproductive states, will determine the frequency and magnitude of competitive and cooperative interactions that occur. The outcome of these repeated interactions will shape overall time budgets and activity patterns. Because females maximize their reproductive success relative to other females by their ability to raise offspring to the age of independence, females are forced to efficiently solve the three ecological problems listed earlier. What particular associations and distributions develop will depend on the particular ecological and physiological circumstances that females experience. These associations and distributions in turn will shape male associations and distributions, because male fitness is mostly determined by their ability to acquire mating access to disproportionate numbers of females. The mating

<table>
<thead>
<tr>
<th>External Constraints</th>
<th>Needs</th>
<th>Internal Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niche/ Habitat Distribution</td>
<td>Food, Water, Safe Sites</td>
<td>Body Size, Reproductive State</td>
</tr>
<tr>
<td>Female Distribution &amp; Abundance</td>
<td>Male Distribution &amp; Abundance</td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 1** Model of ecological and physiological pressures influencing both intra- and intersexual relationships and showing how they shape social systems. Arrows depict social relationships. What type of society develops depends on details and perceptions of environmental conditions.
system that develops becomes the core of a species' social system.

A. Ecological Challenges and Phenotypic Constraints

Typically, in environments in which resources are abundant, especially when they are evenly distributed, competition is low enough to permit females to aggregate. If sufficiently large foraging or antipredator benefits can be derived by females that aggregate, and the groups that form are not too large, then these groups can be defended by single males and so-called harem defense polygyny results. If resources, however, are more patchily distributed so that competition among females intensifies periodically, then female group sizes will vary and female associations become more transitory. Rather than defending unstable groups of females, males instead attempt to defend resource patches sought by females. Typically, in these systems of "resource defense polygyny" the most able males defend the best patches and thus gain access to the largest number of females for the longest periods of time. If resources are not only patchily distributed, but the patches are large, widely separated, and fluctuate seasonally in abundance, then competition among females becomes so low that the formation of large groups is even more likely, provided that females can range widely and follow the shifting locations of peaks in food abundance. Males will thus be forced either to follow these large groups competing for, and then tend, one reproductive female at a time ("wandering") or to position themselves at the intersection of female migratory routes and wait for females to visit them ("lekking"). In either case, intense male-male competition generates a mating system based on "male dominance polygyny," and in the latter case females are afforded the exquisite opportunity to simultaneously compare many males before choosing with which one to mate! Whenever resources are sparsely but somewhat evenly distributed, high levels of competition prevent females from forming groups. As a result, individual females defend territories thus ensuring a regular supply of a renewing resource. Since solitary individuals searching for members of the opposite sex will face increased predation risk, pairs often share territories and monogamy results.

This model accounts for the diversity of mating systems for many different groups of animals including insects, fish, reptiles, rodents, and many varieties of birds, ungulates, carnivores, marine mammals, and pri-
equids the close association between food and water enable horse (Equus caballus) and plains zebra (Equus burchelli) females of different reproductive states to associate permanently. Thus males are able to defend such groups and so-called harems form. When these two resources are widely dispersed, as for Grey’s zebra (Equus grevyi) and the Asian wild ass (Equus hemionus), females of different states are precluded by metabolic constraints from foraging together. As a result, males compete for territories along traveling routes that take females from feeding areas to watering points.

In felids, females remain separate when food is scarce and the habitats are densely wooded because in such circumstances individual prey can be caught and consumed before competitors can intervene. In more open habitats and where both prey and competitors are large and much more numerous, coalitions of females form to help hold on to kills until they are completely devoured. If these female coalitions are themselves large, then there is pressure for males to aggregate to control reproductive access to females. Thus the only highly social felid is the savanna-living lion (Panthera leo), yet even in this plentiful landscape the leopard (Panthera pardus) remains solitary because it can safely cache its large prey in trees.

For canids monogamy is the rule. But variations do occur with small-bodied foxes (Vulpes spp.) sometimes exhibiting polygyny and large-bodied hunting dogs (Lycaon pictus) and timber wolves (Canis lupus) developing polyandry. Typically canids need help from nonbreeding foragers to help nourish lactating mothers. For the mid-sized jackals, such as the silverback jackal (Canis mesomelas), the helpers are young from previous litters that cannot themselves find successful breeding locations. For the smallest species, however, prepartum investment in young is relatively small for a female and thus her mate provides all the help that is needed. In years when prey resources are very high, competition between mothers and their soon-to-be-fecund daughters is low and their daughters are not forced to disperse. As a result they provide neighboring males with additional mating opportunities. For the largest bodied species, however, female prepartum investment is very high and they need all the help they can get. To enlist the support of other adult males that must hunt cooperatively to capture large prey, a dominant female not only kills the offspring of other females in the group, she mates with their mates so that these males behave as if they are the sires of the dominant female’s young. Hence, depending on size-determined metabolic investments and needs, social systems of canids can vary from polygyny to polyandry and sometimes use the services of juvenile or adult nonbreeders to help rear offspring.

B. Kinship and Demography

Although the abundance and distribution of key ecological resources are important in determining the particular pattern of sociability exhibited by a population, other factors such as kinship and demography also play significant roles. Ever since W. D. Hamilton formulated the concepts of “inclusive fitness” and “kin selection,” the magnitude of the costs and benefits associated with particular social behaviors, and that ultimately determine which strategy is best, had to be adjusted by degree of relatedness between social partners. Altruistic or cooperative behavior between relatives should be favored whenever Benefits > Costs × r where r is the degree of relatedness. For both parents and offspring and among full siblings, r = ½; for grandparents and grand-offspring and among half-siblings, r = ½ 2; and among first cousins, r = 1 4. Hence, the stronger the degree of relatedness among relatives, the more likely altruists are to enhance the reproductive opportunities of kin while incurring costs associated with diminished personal reproduction. Thus in the examples presented here, it is not surprising to find that the coalitions that form among lonelines when protecting kills and among male lions when protecting mating opportunities with females are formed most often among full siblings; nor is it surprising that the helpers jackals recruit to rear additional offspring are themselves the full siblings of the offspring being raised. In general, strong kinship lowers the threshold for the appearance of altruistic and cooperative behavior and may substantially affect the costs of living in groups.

But as both these cases illustrate, high population density and the intensified competition it engenders for finding suitable territories with sufficient food and habitable burrows is the factor ultimately responsible for favoring the establishment of coalitions or the recruitment of helpers. Thus demographic factors, such as population density as well as sex ratio and age structure, that result from differences in phenotype-specific vital rates also shape patterns of social organization. And since these features of populations are often altered by human activities, understanding how they shape, and are shaped by, patterns of sociability is essential if conservation assessment and planning is to be effective. Hypothetically, if mortality, for example, were age specific and higher for prereproductive females than males, then high breeding sex ratios (males/females)
would result and mating relationships would generally become more polygynous. If the sex-specific patterns of mortality were reversed, however, then polyandry would become more common. Since some of these mortality concerns were responsible for the variation in social structure described earlier, these mortality schedules can have important consequences. Similarly, if mortality rates were size specific and happened to be greater for larger males rather than smaller ones, then discrete size polymorphisms could arise in populations. This appears to be the case regarding the maintenance of so-called alternative male mating strategies. Often the typical strategy adopted by older and larger males of defending harems or resources yields the most reproductive gains but it often incurs the highest costs as well. Because displaying, fighting, attracting the attention of predators, and delaying reproduction while growing are all costly activities, males adopting less successful but also less costly tactics can also flourish. The maintenance of such alternative mating patterns is common among many species of insects, fish, amphibians, birds, and mammals. For example, in bluegill sunfish (Lepomis macrochirus) males typically defend nest sites where they display, attract females with which to mate, and then fertilize and guard the eggs they lay. Since only the largest males have the ability to defend such nest sites, they must delay breeding in this fashion often for more than seven years. As a result, smaller and younger males have evolved various cuckolding strategies that may in fact be equally successful evolutionarily. In one, the so-called sneak begins breeding at two years of age. Although very small, such sneak males are virtually all testes and because of their inconspicuousness can dive from the surface just as a mating pair release sperm and eggs. By exuding large volumes of sperm, some make it to the eggs and fertilize a few. In the other, males delay breeding for as long as do females. By being the same size and color as females, these males join the mating pair and apparently fool the displaying male into thinking that he is courting two females. Then as the original pair releases their gametes, he does too and thus fertilizes some of the females’ eggs. For these two strategies to be equally successful alternatives, the costs and benefits of each must vary inversely with frequency. Although they sometimes do, as in the case of the bluegill sunfish, in other species they do not.

A variant on this theme is associated with sex change. For many species, especially among the fishes, individuals begin life as one sex and end life as another. In bluehead wrasse (Thalassoma bifasciatum), for example, most individuals begin reproducing as females while residing with one male. If he disappears, then the largest female changes sex and becomes the harem tending male. Alternatively, in anenomefish (Amphiprion spp.), some individuals change from male to female. In either protandrous (male first) or protogynous (female first) species, the sex that is last to the one whose reproductive success is most influenced by body size, and this ultimately depends on the nature of the social system. Clearly, a variety of environmental features influence the patterns of sociality that populations of animals develop. In many cases these patterns are flexible and species can vary in the system of social organization they exhibit depending on environmental conditions. Thus although female burros and asses typically live in transitory groups whose membership changes, when populations move from arid to mesic areas, social relationships can change. As was found on an island off the coast of Georgia, with food and water both more abundant and less separated, females with differing needs can and do coalesce into permanent groups. Males, which in arid areas are forced to establish territories along routes to and from water, respond by deferring these groups much like males of horses, their close kin, and harem-like societies emerge.

III. FROM SOCIAL ORGANIZATION TO ECOLOGICAL IMPACTS

Despite the fact that some species exhibit social flexibil-ity and that there exists significant ‘‘environmental de-
terminism’’ in the development of patterns of sociality, knowledge is only now emerging about how changes in social systems impact important vital rates, alter re-
terspecific competitive abilities, and shape a popula-
tion’s growth rate and genetic structure. Never before have the social systems of so many species been forced to respond to human activities, most of which lead to reductions in population size, through overharvesting, fragmentation of landscapes or habitats, and changes in the earth’s climate. How are we able to predict which species will respond and how will a species’ response impact its long-term growth and the health of the eco-
system in which it resides? When we intervene will we do so with a sufficient understanding of what the consequences of changing social behavior and social systems is likely to be? For example, will we know when a species switches from a monogamous to a poly-
ygynous social system and whether its growth rate or genetic diversity will increase or decrease? Will we
know whether it will be more or less able to withstand selective harvesting or habitat fragmentation? Such consequences of differing patterns of sociality must be understood before endangered species or their habitats can be protected.

A. Consequences of Social Change

Many examples are now appearing of human induced changes in species’ social systems. One population of feral horses inhabiting a barrier island off the east coast of the United States has been impacted twice by human activities and both times major changes to social behavior and social organization resulted in large-scale changes in the population’s vital rates. The first instance occurred when dredging of surrounding ship channels altered currents and sediment deposition around the island. Before the emergence of hundreds of hectares of a new continuous sward of grassland, the island’s horses exhibited two types of social organization.

Where the vegetation was abundant and evenly distributed, females formed permanent membership groups and associated with a single male. Such harem groups were stable and males protected their females from harassment by intruders and rarely herded their own females. Where the preferred high-quality vegetation was broken up into patches of variable size by extensive ridges of sand dunes, females were not able to aggregate into permanent membership groups. As a result, males wandered in search of reproductively receptive females and, when doing so, harassed females and limited their ability to forage. Not surprisingly, the body condition and reproductive success of the harassed females was much lower than that of females bonded to males. After the emergence of the new sward of grassland in the region where females lived in the fission-fusion type of society, females were able to aggregate and males were able to establish harems and the growth rate of the entire population increased markedly. Thus when the abundance and distribution of key ecological resources alter social systems so that they more closely resemble a species’ typical system, vital rates and the overall health of the population improve.

When the perturbation moves a population away from its normal pattern, as happened during the feral horse population’s second brush with human activity, vital rates decline. As the horse population approached its carrying capacity, population control measures were instituted. During the initial round of management actions, over 85% of the harem stallions were removed from the island. The resulting social disruptions were massive; many females were separated from their males and young, inexperienced bachelor males took over. Extremely high levels of harassment led to low reproductive rates despite the fact that overall density had also been reduced and food supplies were expanding. As these two disruptions illustrate, human activities that alter existing patterns of social relationships dramatically alter reproductive rates.

Similar outcomes occurred when large-scale El Niño induced climatic changes resulted in major adjustments to time and activity budgets of Alaskan red foxes (Canis aureus). Before the 1982–1984 El Niño, many of the foxes on an island in the Bering Sea bred polygynously and produced litters of sizes equivalent to, and often greater than, those of monogamous pairs. After the El Niño and the reproductive failure of the large seabirds, the foxes shifted their diet to smaller, less abundant, and harder to catch prey. Dietary shortages were common, fewer foxes bred and polygyny disappeared. Moreover, the litter sizes of these monogamous pairs were smaller as well. Clearly, changes in social behavior in response to altered environmental conditions enabled some individuals to make the best of a bad situation. Overall, however, the health of the population suffered. Major changes in the breeding patterns of elephants (Loxodonta africana) could result if climate changes alter the patterning of grasslands in East Africa. Ordinarily group sizes and compositions change seasonally in response to changing abundances of vegetation. After the rains, when grasses grow rapidly on the extensive upland plains, small family groups of elephants aggregate into large maternal assemblages. With the flush of new vegetation many come into estrus and become reproductively active. At this time only the most dominant bulls come into musth, a heightened sexual and aggressive state. By aggregating, females incite male-male competition thus ensuring that they mate only with the best males. During the dry season, however, after the grasses on the plains stop growing and have been consumed by elephants and other ungulates, the elephants retreat to the swamps where other grasses and browse remain. The patchy nature of the habitat, however, prevents large groups of females from re-maintaining together and the huge herds fission into smaller family units. In addition, few females remain to be mated and the dominant bulls go out of musth. Subordinates become sexually active at this time, but reproductive opportunities are limited. Thus, under a normal environmental regime two evolutionary consequences result: most males do not mate, but those that do are the most fit, and because generic effective population size is affected by the number of males relative to the number of females mating, the effective population size
of elephants is much smaller than the census population size. If global warming increases the duration and intensity of dry seasons, then selection for the best males will be relaxed but the effective population size of the population will increase, provided that elephants can sustain their high metabolic levels and maintain their high fertility and survival rates. With global changes in climate appearing in many different habitats, other examples of social readjustments that either change a species’ evolutionary potential or that act as demographic shock absorbers in the short run, but are unlikely to provide compensatory relief in the long run, are likely to become more common.

B. Consequences of Demographic Change

Many demographic changes leave their mark by altering critical aspects of social behavior or by disrupting the development of important social relationships. As the elephant example illustrates, changing environmental conditions can change operational sex ratios thus altering a population’s ability to maintain genetic diversity. But by changing sex ratios in other ways, humans can put populations at risk by making it difficult for them to grow in size. In lions, for example, trophy hunting, if it removed pride defending males and nomads in proportion to their abundances, would not severely affect population growth and effective population sizes would be maintained at normal levels. But if such hunting focused on pride males, then increased turnover would foster increased infanticide by take over males. While such diversification of the gene pool might increase effective population sizes, it would certainly lower population growth rates because infant survival, which is already low, would plummet. Much clearly depends on whether or not the population is polygynous or monogamous and whether the perturbation will accentuate or reverse the pattern. Large-scale climate changes could reverse and ameliorate the normal pattern as is likely to be the case in elephants, but selective hunting or poaching, as in the case of lions, could exacerbate an already critical situation.

Selective poaching could even make it difficult for populations to recover once the poaching has been eliminated. Since hunting elephants for ivory meant that harvesting in fish might be accelerating already rapid population declines because of disruptions to normal social interactions. When harvesting is size selective and reduces the number of large females in a population, fecundity might be disproportionately reduced since ovary volume scales allometrically with linear body size dimensions; larger females have disproportionately larger ovaries and high fecundity. And as sex ratios increase, male harassment of the remaining females may further reduce recruitment and population growth rate. When such size selective harvesting removes more males than females, sex ratios are reduced and females may become less choosy and more aggressive when selecting mates. While such changes in the intensity of sexual selection might not alter population growth rates, they might alter the genetic diversity of the population. Yet if males provide the majority of parental care and large males are removed from the population as in often the case in sports fisheries, then recruitment will be markedly reduced as unguarded eggs are cannibalized.

The implications of overzealous size-selective harvesting in sex-changing fish or in fish that exhibit alternative male mating behaviors appear to be equally severe. For species such as anemonefish in which the largest males change to females, removal of the largest females will force males to change sex at ever smaller sizes, thus producing fewer eggs. Population growth rates will be curtailed and the species will find itself at risk all because of overzealous aquarium traders. Migratory Atlantic salmon (Salmo salar) provide perhaps the clearest example of where human activities are changing the balance of male mating strategies within a population and hence the patterns of sociality. Atlantic salmon typically are born in fresh water and develop for 1 to 2 years before smolting and heading to sea to grow and fatten by feeding on a rich supply of marine invertebrates. Once they attain a certain size, they become sexually mature and return to rivers. There they travel long distances upstream until they reach clear and cool breeding grounds. Some individuals, however, never head to the sea. Instead, they remain in their natal streams and mature sexually at young ages and at small sizes. Such individuals are called parr and they never go through the smolting process that adapts them to a marine lifestyle. Under pristine environmental conditions, the fraction of the population that becomes parr is small since the reproductive gains of such a strategy are low. When competing for mates with larger more aggressive males, parr fare poorly. What matings result, population growth rates could remain low for historically heavily poached populations.

Harvesting in fish might be increasing already rapid population decline because of disruptions to normal social interactions.
they obtain are derived by ‘sneaking’ among a mating pair and releasing milt at just the right time. Such events are rare and the mixing of milt is poor so such ‘sneak’ matings result in few young being sired. But the survival prospects of parr are high since they don’t incur the risks of going to sea and, moreover, they begin breeding at a very early age. Thus over a lifetime, reproductive success is moderate. But as human fishing increases and the netting of the older larger salmon intensifies, the relative lifetime reproductive success of parr is improving. Since the costs of migrating to sea and back again have increased dramatically, the long life span of parr give them a relative advantage. And given that competition with the larger males for mates is also being reduced, the chances of parr securing matings are also improving. Thus it is not surprising that the composition of the population is changing as parr increase markedly in abundance. The impact on the long-term stability of the population is unclear. But the long-term impact on the balancing of mating morphs and sexual relationships is.

C. Consequences of Inappropriate Interventions

When humans intervene and develop management and conservation plans they sometimes do so without accounting for important aspects of social behavior. As a result some disastrous consequences have ensued.

Ignorance about the mating system of sperm whales ( Physeter macrocephalus ) has led to the implementation of harvesting plans that have almost led to their extinction. By using data from previous harvests in which the counts of males were much fewer than those of females, sperm whales were thought to exhibit a harem breeding system. Such a conclusion could only have been derived by assuming that sperm whales lived in permanent, closed membership groups. By making this assumption, the International Whaling Commission concluded that most males sighted would be superfluous thus disproportionate hunting of males would not limit the growth rate of the population, most of the males caught would not be necessary for fertilizing females. Unfortunately, the social relationships of sperm whales are very different from those assumed. Females are the core of the society and do appear to live in permanent membership groups. But these groups often merge when diving for fish and divide labor between fishing and tending the young that cannot dive deeply. Males are forced to leave their natal group when reaching sexual maturity, but they cannot accompany the breeding associations to the tropics. Instead they remain at the higher latitudes foraging in the cooler more productive waters. Thus when males are hunted on the breeding grounds it is breeders and prereproductives that are taken. With breeding males severely reduced in abundance and the younger, subadult upstart located thousands of kilometers away, many females are going unmated. As a result, the population is failing to recover and it will be years before the next generation of males is ready to start the population, already handicapped by such a low recruitment potential, on the road to recovery. Knowledge about sperm whale social relationships could have prevented the mistaken overharvesting of mature males.

Another species in which ignorance about important features of social behavior is problematic is the black rhinoceros ( Diceros bicornis ). During the past 25 years, numbers have declined from 65,000 to fewer than 2,500 because demand for their horns is so great. In an attempt at making rhinos less desirable, some nations have implemented a strategy of dehorning. Other nations have opted to translocate rhinos to extremely well protected areas. Unfortunately, horns, and thus the value of rhinos, regrow quickly. But what is more troubling about the dehorning treatment is that it appears to disrupt effective maternal antipredator behavior. Although dehorned rhinos were no more likely to flee from predators than intact rhinos, the disproportionate disappearance of offspring being reared by dehorned mothers as opposed to intact mothers in areas with large numbers of lions and spotted hyenas suggests that horns play a vital parental and protective function. Dehorning, although initially thought of a cure for the disappearance of adults via poaching, became a contributing factor to the disappearance of young via predation. Although dehorning might enhance adult survival in the short run, it would limit a population’s growth potential in the long run.

A third example where ignorance hindered, but did not harm, a conservation program involved the reintroduction of Asiatic wild asses ( Equus hemionus ) to habitats within its historical range. A goal of the Israeli government is to repopulate Judea and Samaria with biblical animals. Since the Palestinian race of the wild ass was extirpated by the Ottoman Turks at the beginning of the twentieth century, the onager was a prime candidate for translocation. The first reintroduction took place in 1982 with subsequent additions throughout the 1980s and increased the number of breeding females to 14. By the mid-1990s, however, the population had hardly grown and the number of breeding females totaled 16. What was not known at the time was that breeding success is bolstered by being reared
in the wild and that onagers could facultatively adjust the sex ratio of the offspring they produced. In the Negev Desert, free-ranging onagers gave birth disproportionately to sons.

Trivers and Willard were the first to suggest that differences in the ability to invest in the successful rearing of offspring should lead to individual differences in the primary sex ratios of offspring. They argued that females with sufficient resources should invest in offspring of the sex with the higher variance in reproductive success. By producing well-endowed offspring of this sex, mothers would increase the number of grand-offspring they were likely to have. In polygynous species like the onager, males are usually the sex with the greatest variance. Since onager females compete more on the basis of resource exploitation, or utilization efficiency, rather than via direct conflicts settled by social status, sex biases in offspring production were not expected. However, because age and experience affect utilization efficiency, differences in bodily condition and ability to invest in offspring were expected to exist among females in this onager population. Thus it was heartening to find that middle-aged females—those in the best physical condition—gave birth to more sons than daughters. Conversely, those breeding for the first or second time gave birth to daughters as did older females nearing the end of their reproductive lives.

The strategy the Israelis employed of translocating middle-aged females to the desert was sensible in economic terms since high risks were attached to releasing very young females and low future expectations were associated with releasing older females. But had the Israelis released old females with yearling daughters they would have quickly accelerated the growth of the population. At the time, an understanding of how environments shape patterns of onager social behavior and the impacts that these behaviors would have on population dynamics was unknown. By understanding the relationships between environments and social behavior and between social organization and population processes, we should be in a better position to understand the problems endangered species are likely to face and what interventions to protect them are likely to be effective.

See Also the Following Articles

MUTUALISM, EVOLUTION OF • POPULATION DENSITY • SALMON • SPECIES COEXISTENCE

Bibliography


SOIL BIOTA, SOIL SYSTEMS, AND PROCESSES

David C. Coleman
University of Georgia

I. Soils as Components of Ecosystems
II. Soils as Organizing Centers in Ecosystems
III. Major Soil Processes
IV. Biodiversity in Soils
V. Conclusions

GLOSSARY

domains The major divisions of the biota on earth, namely: Bacteria, Archaea, and the Eucarya.
ecosystem engineers The concept whereby members of the macrofauna (e.g., termites and earthworms) are actually moving parts of the soil volume for their own uses (e.g., making macropores, which permit flow of large amounts of water rapidly through the soil).
imobilization The process wherein nutrients are taken up or immobilized in litter and other organic detritus until later (usually weeks to months) in the decomposition process.
microfauna Animals that have high turnover rates, and live in water films in soils. Small mesofauna, such as nematodes, are also water-film dwellers.
larger Mesofauna and macrofauna live in pores in portions of soil profiles.
mineralization The availability of inorganic nutrients in the decomposition of organic detritus, occurring after the immobilization phase; see above.
organizing centers Soils are centers of history and activity in terrestrial ecosystems. See the legacy concept in forest ecology.

SOIL BIODIVERSITY is an intriguing, largely unappreciated facet of global biodiversity. There are many phyla, even “domains,” within soils, which are largely unseen, making use of the uniquely diverse physicochemical complexity of soils, which is an intersection of mineral, organic, aquatic, and aerial habitats. Organisms have evolved in soils literally since pre-Cambrian times (more than 600 million years ago). They are still largely undescribed, and this is particularly true for the prokaryotes, which have awaited the development of new techniques to characterize them. By linking several organismal groups to major processes in global biogeochemistry, it is proving possible to appreciate the wide array and diverse nature of soil organism functions in the biosphere.

I. SOILS AS COMPONENTS OF ECOSYSTEMS

A. Soil-Forming Factors

Soils are an intriguing, relatively thin (often <1 m depth) zone of physical-chemical and biological weathering of the earth’s land surface. Soils are formed by
an array of factors, namely climate, organisms, parent material, the extent of slope, and aspect (relief) operating over time (Fig. 1). These factors affect major ecosystem processes, such as primary production, decomposition, and nutrient cycling, which lead to the development of ecosystem properties unique to that soil type, as a result of its previous history. For example, a deep loess soil in Iowa, with a very fertile and deep surface or "A" horizon, containing considerable amounts of organic matter, will be very different from an "A" horizon developed in the Nebraska sandhills, with much greater porosity and lower water retention due to the nature of the sandy surface material. As noted in the soil-forming factors diagram (Fig. 1), the array of biota—namely microbiota, vegetation, and consumers (herbivores, carnivores, detritivores)—is influenced by soil processes and in turn has an impact on the soil system.

B. Poly-Phasic Nature of Soils, Influence on the Biota

Soils are perhaps the ultimate in interface media, located at the intersection of four principal entities: the atmo-

FIGURE 1
II. SOILS AS ORGANIZING CENTERS IN ECOSYSTEMS

Soils may be viewed as the organizing centers for terrestrial ecosystems. Major functions such as ecosystem production, respiration, and nutrient recycling are controlled by the rates at which nutrients are released by decomposition in the soil and litter horizons and transported to the photosynthetic layers of the ecosystem. This is particularly true for less heavily managed, near-natural ecosystems, many of which occur on soils of relatively poor nutrient status. In these systems, mycorrhizas are often obligate partners in the obtaining of adequate nutrients for the growing plants. Mycorrhizas are known to be efficient at extracting nutrients from both mineral and organic sources, enabling plants to thrive in habitats that are considered poor in nutrients. We need to be aware of these and other mutualistic associations between microorganisms and roots, such as rhizobia and actinorhizas, various root-associated symbiotic bacteria that facilitate the nitrogen fixation process on which the entire ecosystem often depends. These associations have arisen in soils over evolutionary time and are key to an understanding of ecosystem function.

III. MAJOR SOIL PROCESSES

A. Decomposition: Immobilization and Mineralization

A very large proportion (greater than 90%) of the terrestrial net primary production is returned to the soil as dead organic litter. This litter, consisting of leaves, roots, and wood from trees and organic residues from agricultural fields, is decomposed on or in the soil, and the nutrients contained within it recycled for further use. The decomposition process drives complex food webs in the soil, with numerous interactions between the initial agents of decomposition, the bacteria and fungi, and the fauna which in turn feed on them.

Decomposition is the catabolism of organic compounds in plant litter and other organic detritus. Decomposition is principally the result of microbial activities; few soil animals have cellulases in their guts, which allows them to hydrolyze the celluloses in plant residues. The decomposition of organic residues involves the activities of a variety of soil biota, including both microbes and fauna, which interact conjointly in the process. For example, the initial breaking up of plant litter usually is conducted by the chewing and macerating action of both large and small animals. This initial breaking into smaller pieces, or "communation," is a process that benefits the fauna, which derive nutritional benefit from the litter or microbes initially colonizing the plant material. The increased surface area and further inoculation of the smaller pieces enhances the microbial access to, and breakdown of, these tissues.
B. Nitrogen Cycle: Major Processes

Nitrogen enters the ecosystem via nitrogen fixation, in which the dinitrogen molecule ($N_2$) is separated into two nitrogen atoms, with considerable expenditure of energy and the assistance of the nitrogenase enzyme, to break the triple covalent bond. The atoms are ammonified and then used in the production of amino acids and proteins in the plant. Another avenue for nitrogen entry into soils is by lightning fixation, in which the extensive high-voltage energy in the lightning charge ruptures the dinitrogen molecule, hydrogens are attached, and then the ammonium is brought in by rainfall. As shown in Fig. 3, nitrogen is lost from the system via harvest and erosion of organic forms of N, it can be ammonified in decomposition, and then undergoes nitrification to nitrate ($NO_3^-$), whereupon it can be taken up by biota, either plant roots or into microbial tissues. If there is adequate energy and low amounts of oxygen present, there can be denitrification, in which the nitrogen is lost as either nitrogen gas ($N_2$) or $N_2O$, nitrous oxide. For further details, consult textbooks on ecology or ecosystem studies.

The nitrogen cycle is of critical importance to biodiversity considerations, because key points in the cycle are dependent on relatively species-poor assemblages of microbes, including the nitrogen fixation and nitrification steps. There are only a few species of nitrogen fixing rhizobia, in the genera *Rhizobium* and *Bradyrhizobium*. The other principal nitrogen-fixing symbiont, the bacterium *Frankia* (Actinomycetales) forming the actinorrhiza (literally actinomycete-root), contains only a few species in the genus. However, approximately 194 plant species in eight families and four different subclasses of flowering plants have been identified as hosts. These plants share the general tendency to grow in marginal soils and play an important role as pioneer species in early successional habitats.

In the nitrification steps, noted earlier, there are only a few genera and species of nitrifiers. Most of them are autotrophic and quite sensitive to changes in soil pH. This means that these organisms may be unusually prone to being diminished or eliminated in regions where there is considerable acid rain.

IV. BIODIVERSITY IN SOILS

A. Evolutionary History

Soils, as we know them, with well-differentiated profiles, probably developed concurrently with the origin of a land flora in the early Devonian era, about 425 million years ago. The microorganisms that inhabit the soils, particularly the prokaryotic microbes such as the cyanobacteria, originated perhaps 3 billion years ago.
B. Diversity of Biota

Biodiversity is an inclusive concept, including a wide range of functional attributes in ecosystems in addition to being concerned with numbers of species present in the system. This differentiates it from the concept of species diversity, which is concerned with the identity and distribution of species in a given habitat or region.

Soil biodiversity is best considered by focusing on the groups of soil organisms that play key roles in ecosystem functioning. Spheres of influence (SOI) of soil biota are recognized, such as the root biota, the shredders of organic matter, and the soil bioturbators. These organisms influence or control ecosystem processes and have further influence via their interactions with key soil biota (e.g., plants). What is the extent of redundancy within functional groups within these SOI? Some soil organisms, such as the fungus and litter-consuming microarthropods, are very species rich. For example, there are up to 170 species in one Order of mites, Oribatida (members of the Arachnida, eight-legged arthropods), in the forest floor of one watershed in western North Carolina. The soil biota considered at present to be most at risk are some of the species-poor functional groups, such as specialized bacteria, that is, nitrifiers and nitrogen fixers (see diagram). Others include fungi forming mycorrhizae (literally fungus-root), a symbiotic association that benefits both plant and fungus, with the plant supplying high-quality carbon to the fungus, and the fungal hyphae exploring a greater volume of the soil, obtaining scarce mineral nutrients, particularly phosphorus. Other species-poor functional groups include macrofaunal shredders of organic matter (e.g., millipedes) and bioturbators of soils, which includes various types of earthworms and termites.

C. Three Great “Domains” of Organisms on Earth

All of life exists in three great “urkingdoms,” or domains. These domains are (a) the Bacteria (eubacteria), which are the bacteria as generally considered; (b) Archaea (archaebacteria), which include the methanogens (methane-producers), most extreme halophiles (ones living in hypersaline environments), and hyperthermophiles (ones living in volcanic hot springs, and in mid-sea ocean hot-water vents); and (c) Eukarya (eukaryotes) (Fig. 4). The first two domains are prokaryotes, which are unicellular organisms, lacking a unit membrane-bound nucleus and other organelles, usually having their DNA in a single circular molecule. Eukaryotes, in comparison, consist of all of the organisms that have a unit membrane-bound nucleus and other organelles, such as mitochondria. Eukaryotic organisms are often multicellular. This scheme is based on an increasing body of evidence from ribosomal RNA (rRNA) phylogenies, that the archaeabacteria are worthy of the same taxonomic status as eukaryotes and bacteria. As shown in Fig. 4, the universal rRNA tree develops from a postulated “cenancestor,” leading to the relative positions of the three great domains.

1. Number of Species of Prokaryotes

Recent estimates of the number of prokaryotic species range from 100,000 to 10 million. Interestingly, the number of described species of bacteria in soil amounts only to about 4000. This discrepancy is due largely to the fact that only a small proportion, usually less than 1%, of the bacteria present in soil or any other medium are amenable to culturing and subsequent microscopic observation.

It should be noted that, on the basis of the accepted criterion for separating taxa in microbial studies, which is a greater than 70% DNA homology, a mouse and a human would be considered as being in the same species. This leads to complications, as we shall see, in discussing the total amount of genetic diversity of all organisms, including the as-yet largely unknown diversity of Archaea and Eubacteria. The latter now are estimated to have an array of 36 kingdoms, which are genetically as diverse as the Kingdoms Animalia, Plantae, and Fungi in older classification systems.

2. Biomass and Numbers of Bacterial Species on Earth

This figure is vastly underestimated. We are just now delineating the overall genetic makeup of isolates taken from soils, which are determined by the use of molecular probes. The total numbers of bacteria on earth in all habitats is truly staggering: $4 \times 10^{30}$ cells, or 350 to 3500 petagrams of Carbon. One petagram is $10^{15}$ g, or one billion metric tonnes. The amount of the total that is calculated to exist in soils is approximately $2.6 \times 10^{23}$ cells, or about 5% of the total on earth. A majority of bacteria exist in oceanic and terrestrial subsurfaces, especially in the deep mantle regions, extending several kilometers below the earth’s surface.

3. Viruses as Quasi-Organisms

Viruses are quasi-organisms, not included in the three domains. Viruses are RNA or DNA molecules contained within protein envelopes. Viral particles are metabolically inert, carrying out neither biosynthetic nor respiratory functions. They multiply only within host cells, by inducing a living host cell to produce the necessary
viral components. Once assembled, the replicated viruses escape from the cells. Viruses infect all sorts of animals, plants, and microbes. Viruses parasitizing bacterial cells are commonly called bacteriophages, or simply phages. Although little is known about the ecology of viruses, they can persist in soils for many years and decades. Some research on viruses in deserts showed that they were inactivated in soils at acid pH levels between 4.5 and 6. There is little information on the overall species diversity of viruses in soils. Current estimates are 5000 species known and perhaps 130,000 in existence.

4. Numbers and Biodiversity of Eukaryotes

a. Fungal Diversity

Fungi are multicellular eukaryotes that are found in many habitats worldwide. They have long, ramifying strands (hyphae), which can grow into and explore many microhabitats, and are used for obtaining water and nutrients. The hyphae secrete a considerable array of enzymes, such as cellulases, and even lignases in some specialized forms, decomposing substrates in situ, imbibing the decomposed subunits and translocating them back through the hyphal network. Fungi are very abundant, particularly in undisturbed forest floors in which literally thousands of kilometers of hyphal filaments will occur per gram of leaf litter.

Fungi are still little-described, with possibly less than 5% of them known to Science (69,000 described; perhaps 1,500,000 in existence (Table I)). This is largely because of the fact that so many fungi are associated with tropical plants and animals, and these in turn have not been described.

As noted earlier, the roles of mycorrhizas in soil systems are being increasingly viewed as central to much of terrestrial ecosystem function. The total number of mycorrhizal species may be just 1000 or 2000, but they are essential to the growth and reproduction of numerous families of plants. Recent experimental studies have noted that species richness, namely with large versus small numbers of species of Arbuscular

![Diagram of a universal rRNA tree showing the relative positions of evolutionary pivotal groups in the domains Bacteria, Archaea, and Eucarya. The location of the root (the cenancestor) corresponds to that proposed by reciprocally rooted gene phylogenies. The question mark beside the Archezoa group Microsporidia denotes recent suggestions that it might branch higher in the eukaryotic portion of the tree. (Branch lengths have no meaning in this tree.) From Brown and Doolittle (1997).]
TABLE I

<table>
<thead>
<tr>
<th>Group</th>
<th>Known species</th>
<th>Estimated total species</th>
<th>Percentage known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vascular plants</td>
<td>220,000</td>
<td>270,000</td>
<td>81</td>
</tr>
<tr>
<td>Bryophytes</td>
<td>17,000</td>
<td>25,000</td>
<td>68</td>
</tr>
<tr>
<td>Algae</td>
<td>40,000</td>
<td>60,000</td>
<td>67</td>
</tr>
<tr>
<td>Fungi</td>
<td>69,000</td>
<td>1,500,000</td>
<td>5</td>
</tr>
<tr>
<td>Bacteria</td>
<td>3,000</td>
<td>30,000</td>
<td>10</td>
</tr>
<tr>
<td>Viruses</td>
<td>5,000</td>
<td>130,000</td>
<td>4</td>
</tr>
</tbody>
</table>


mycorrhiza, has a positive impact on plant primary production in macrocosms of North American old fields (fields undergoing succession and not intensively managed).

b. Microfauna

The unicellular eukaryotes, or Protoctista, include a wide range of organisms, which are more often called protozoans. These include the flagellates, naked amoebae, testaceae, and ciliates (Fig. 5). These organisms range in size from a few cubic micrometers in volume to larger ciliates, which may be up to 500 micrometers in length and 20 to 30 micrometers in width. Protozoa are quite numerous, reaching densities of from 100,000 to 200,000 per gram of soil. Bacteria, their principal prey, often exist in numbers up to 1 billion per gram of soil. All of these organisms are true water-film dwellers and become dormant or inactive during episodes of drying in the soil. They can exist in inactive or resting stages for literally decades at a time in very arid environments.

About 40,000 extant protozoan species have been described, but many more undoubtedly are awaiting scientific discovery. Foissner (1997) notes that about 360 protozoan species per year are being discovered. In an extensive survey of soils from Africa, Australia, and Antarctica, in some cases nearly half of the total species described were new to science. This was particularly true in Africa, where of 507 species identified, 240 of them, or 47%, were previously undescribed. Even in a more extensively investigated region, Australia, 43% of the total of 361 species were new to science. In Antarctica, 93 species were described, with only 14, or 15%, being unknown.

Because many habitats have been uninvestigated yet, and the isolation procedures are still imperfect, from 70 to 80% of all soil ciliates may yet be unknown. This high proportion may hold true for the other protozoan groups as well.

c. Mesofauna

i. Nematodes

Nematodes feed on a wide range of foods. A general trophic grouping is bacterial feeders, fungal feeders, plant feeders, and predators and omnivores. Anterior (stomal or mouth) structures can be used to differentiate general feeding or trophic groups. The feeding categories are a good introduction, but feeding habits of many genera are complex or poorly known. For example, some genera in immature phases will feed on bacteria and then become predators on other fauna once they have matured. Because of the wide range of feeding types and the fact that nematodes seem to reflect ages of the systems in which they occur (e.g., annual versus perennial crops, or old fields and pastures and more mature forests), they have been used as indicators of overall ecosystem condition. This is a growing area of research in soil ecology, and one in which the intersection of community analysis and ecosystem function could prove very fruitful. Current species described total some 5000, and upward of 20,000 may exist.

ii. Collembola

Collembolans, or "springtails," are primitive Apterygote (wingless) insects. They are called "springtails" because many of them have a spring-like lever, or furcula, which enables them to move many body lengths away from predators by use of it, in a springing fashion. Collembolans are ubiquitous members of the soil fauna, often reaching abundances on 100,000 or more per square meter. They occur throughout the soil profile, where their major diet is decaying vegetation and associated microbes (usually fungi). However, like many members of the soil fauna, collembolans defy placement in exact trophic groups. Many collembolan species will eat nematodes when those are abundant. Some feed on live plants or their roots. One family (Onychiuridae) may feed in the rhizosphere and ingest mycorrhizae or even plant pathogenic fungi. Eight families of collembolans occur in soils. Many collembolans are opportunistic species, capable of rapid population growth under suitable conditions. Eggs are laid in groups. Collembolans become sexually mature with the fifth or sixth instar, but they continue to molt throughout life. Although many species are bisexual, some of the common species are parthenogenic, consisting of females only. Collembolan "blooms" are a phenomenon of late winter or early spring, when some species may appear in large numbers on the surface of snow banks, on the surface ice of pond water, or on
lichen-covered granite outcrops. There are some 6500 described species and possibly more than 10,000 in existence.

iii. Mites (Acari) The soil mites, Acari, are chelicerate arthropods related to the spiders. They are often the most abundant microarthropods in many types of soils. A 100-g sample may contain as many as 500 mites representing nearly 100 genera. This diverse array includes participants in three or more trophic levels, with varied strategies for feeding, reproduction, and dispersal.

Four suborders of mites occur frequently in soils: the Oribatei, Prostigmata, Mesostigmata, and Astigmata. Occasionally, mites from other habitats are extracted from soil samples. These include, for example, plant mites (also called spider mites), predaceous mites normally found on green vegetation, and parasites of vertebrates or invertebrates. The most numerous ones...
are the true soil mites. The orbibatid mites (Ornibatidae) are the characteristic mites of the soil and are usually fungivorous or detritivorous. Messostigmatid mites are nearly all predators on other small fauna, although a few species are fungivores and may become numerous at times. Astigmata mites are associated with rich, decomposing nitrogen sources and are rare except in agricultural soils. The Prostigmata contains a broad diversity of mites with several feeding habits. Very little is known of the niches or ecological requirements of most soil mite species, but some interesting information is emerging. For further details on the life-history characteristics of these interesting animals, refer to Coleman and Crossley (1996). About 20,000 species have been described and possibly in excess of 80,000 exist.

d. Macrofauna

i. Termites Termites (Isoptera) are one of the major ecosystem "engineers" particularly in tropical regions. Termites are social insects with a well-developed caste system. By their ability to digest wood, they have become economic pests of major importance in some regions of the world. Termites are arranged in five different families. The termites in a more primitive family, the Kalotermitidae, possess a gut flora of protozoans, which enables them to digest cellulose. Their normal food is wood that has come into contact with soil. Many species of termites construct runways of soil, or along root channels, and some are builders of large, spectacular mounds. Members of the phylogenetically advanced family Termitidae possess a formidable array of microbial symbionts (bacteria and fungi, but not protozoa), which enable them to process and digest the humified organic matter in tropical soils and to grow and thrive on such a diet. Although termites are mainly tropical in distribution, they occur in temperate zones and deserts as well. Termites are often considered the tropical analogs of earthworms since they reach large abundances in the tropics and process large amounts of litter. Termites parallel earthworms in ingestive and soil turnover functions. The principal difference is that earthworms egest much of what they ingest in altered form (that enriches microbial action), whereas termites can transfer large amounts of soil/organic material into building nests and mounds (carbon sinks). More than 2000 species of termites have been described, and probably up to 10,000 exist.

ii. Earthworms The earthworm fauna of North America is surprisingly poorly known, given the importance of these animals to soil processes and soil structure. Much of the evidence for earthworm effects on soil processes comes from agroecosystems and involves a small group of European lumbricids (family Lumbricidae) that are commonly known of the niches or ecological requirements of most soil mite species, but some interesting information is emerging. For further details on the life-history characteristics of these interesting animals, refer to Coleman and Crossley (1996). About 20,000 species have been described and possibly in excess of 80,000 exist.

V. CONCLUSIONS

It is apparent that a large proportion of the biota associated with soils are as yet undescribed, with the most extreme cases being the bacteria and fungi. However, even somewhat more extensively studied groups, such as Orbited lumbricids, more than half remain unknown to science. Therefore, it is premature to give even a rough estimate of the total numbers of species that occur in many of these taxa, as such large percentages of the total number of organisms are still unknown. It is incumbent on the rising generation of ecologists and biologists to develop more innovative ways to describe, catalog, and understand the myriad patterns and processes in the biosphere, which are due in large part to the actions of the biota. It is hoped that some of the observations in this chapter, plus the insights offered...
SOIL BIOTA, SOIL SYSTEMS, AND PROCESSES

by the references cited in the bibliography, will encourage this effort.

See Also the Following Articles

ARCHAEA, ORIGIN OF • BACTERIAL BIODIVERSITY • EUKARYOTES, ORIGIN OF • FOOD WEBS • FUNGI • NITROGEN AND NITROGEN CYCLE • SOIL CONSERVATION

Bibliography


SOIL IS AN IMPORTANT natural resource supporting plant, animal, and human populations and is a habitat for a diversity of species. Productive soils throughout the world are being degraded rapidly due to human activities. Declining soil quality over the past century has manifested itself in loss of agriculturally productive lands, forests, and wetlands. Soil conservation efforts aim for preserving, sustaining, or improving the quality of soils. This chapter discusses soil quality and the many methods for conserving, sustaining, and restoring soil ecosystems.

I. DEFINITION OF SOIL

A. A Complex and Dynamic System

Soil is composed of living and nonliving components organized vertically, in a profile of horizontal layers or horizons. Soil is the habitat for a great abundance and diversity of living organisms, many of which are microscopic (e.g., bacteria, fungi, protozoa, nematodes, mi-

GLOSSARY

agroecosystem Agricultural ecosystem (e.g., crop field and grazing pasture).
biocontrol Control of agricultural pests with the use of predators and other beneficial organisms (e.g., control of turf grass crickets with insect parasitic nematodes).
eutrophication The process by which a body of water becomes enriched in dissolved nutrients (nitrogen, phosphorous) that stimulate growth of aquatic plant life (e.g., algae), usually resulting in the depletion of oxygen from water. This frequently creates unfavorable conditions for fish and other biota.
herbivory Animals feeding on plants.
mineralization Process of transforming from organic to inorganic form.
soil food web Representation of all feeding interactions among organisms in the soil (who eats whom).
soil structure Spatial arrangement of soil particles.
soil texture Percent composition of clay, silt, and sand in soil.
trophic Describing feeding habits or the kind of nutrition used by a group of organisms.

water-holding capacity Capacity of soil to hold water (e.g., sandy soils have very low water-holding capacity).
TABLE I
Examples of Some Groups of Soil Fauna Ordered by Body Width and the Number of Described Species in Soil

<table>
<thead>
<tr>
<th>Soil fauna</th>
<th>Estimated species number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microfauna (1–130 μm body width)</td>
<td></td>
</tr>
<tr>
<td>Protozoans</td>
<td>1,500</td>
</tr>
<tr>
<td>Nematodes</td>
<td>5,000</td>
</tr>
<tr>
<td>Mesofauna (80 μm–4 mm body width)</td>
<td></td>
</tr>
<tr>
<td>Mites</td>
<td>20,000–30,000</td>
</tr>
<tr>
<td>Springtails</td>
<td>6,500</td>
</tr>
<tr>
<td>Diplurans</td>
<td>660–800</td>
</tr>
<tr>
<td>Pot worms</td>
<td>600</td>
</tr>
<tr>
<td>Termites</td>
<td>1,600</td>
</tr>
<tr>
<td>Macrofauna (1–35 mm of body width)</td>
<td></td>
</tr>
<tr>
<td>Isopods</td>
<td>5,000</td>
</tr>
<tr>
<td>Centipedes</td>
<td>2,500</td>
</tr>
<tr>
<td>Millipedes</td>
<td>10,000</td>
</tr>
<tr>
<td>Earthworms</td>
<td>3,600</td>
</tr>
<tr>
<td>Ants</td>
<td>8,800</td>
</tr>
<tr>
<td>Flies (larvae)</td>
<td>60,000</td>
</tr>
</tbody>
</table>

Only about 1 to 5% of soil biota have been described to species level. Modified from Wall and Virginia (1999).

croarthropods) (Table I). The nonliving component of soil consists of the solid phase, or weathered parent geological material, which contributes to physical (e.g., soil texture, structure, density, porosity [the space within and between aggregates]) and chemical (e.g., fertility, moisture, acidity) properties. The formation of soil is a function of climate, parent material, biota, and topography, which through geologic time has formed a multitude of natural soil types of varying properties across geographic landscapes (Jenny, 1980). Hundreds to thousands of years may be required to form just a centimeter of soil.

**B. Role of Soil**

Fertile, productive soils provide a basis for the economic wealth of a nation by providing food, fiber, and fuel. Soil supports human civilization by supplying nutrients for growth of plants, including agricultural crops, by regulating the water flow from rainfall to groundwater, and by filtering and transferring damaging substances that might enter the atmosphere or groundwater. Water quality is largely dependent on the filtering capacity of the living and nonliving components of soil. Soil as a natural resource, therefore, contributes to the foundation of our social and industrial infrastructure.

Belowground systems, as fundamental constituents of all terrestrial ecosystems, influence and are influenced by the functioning (the performance) of the aboveground systems. For example, soils affect aboveground biodiversity, ecosystem nutrient and energy cycles and fluxes, and even certain atmospheric components (e.g., global cycles of C and N). Soils are one of the largest reservoirs of global carbon and are the major terrestrial reservoir of dead organisms—plants, animals, and microorganisms. The quality (chemical composition) and quantity of dead material (organic matter) received from both aboveground and within the soil determine the primary nutrient and energy base for the soil biota. Plant litter (dead leaves, twigs, roots) and other dead organic matter is decayed by soil micro- and faunal assemblages (organized in detritus food webs) during decomposition, a process of transforming organic matter to inorganic compounds (e.g., nitrate, ammonium, and phosphate), which supplies nutrients for plant growth. Thus, factors affecting the quality and quantity of soil organic matter (SOM), such as climate, cultivation, invasive species, and atmospheric nitrogen (N) inputs (e.g., acid rain), can alter the diversity of above- and belowground organisms, the rate of decomposition, the structure of soil, and the availability of nutrients needed for plant growth.

**II. SOIL QUALITY**

A. Attributes of High-Quality Soils

In the broader sense, the quality of soil refers to an ability to sustain biological productivity and the diversity of plant, animal, and human populations, and to maintain quality of water and air. High soil quality also implies an ability of soils to maintain high fertility, productivity, and resist erosion. Natural differences in the quality of soils indicate different capacities of soils to resist stress whether of natural (e.g., wind, fire, and rainstorms) or anthropogenic (e.g., plow, invasive species, and pesticides) origin. In the narrower sense, soil quality relates to the inherent combination of biotic, physical, and chemical properties that allows soils to have long-term productivity.

A number of factors are involved in defining the quality and productivity of soils: parent material (the geologic weathering of rock), climate, soil organic mat-
ter, soil structure (aggregation of soil particles), soil stabilization, and soil biota (including roots) (Coleman and Crossley, 1996). One of the most important factors is soil organic matter. The amount and type of soil organic matter and the products of decomposition affect several soil properties. A higher content of SOM results in higher cation exchange capacity (CEC), higher water-holding capacity, higher infiltration rates, better soil aeration, and increased soil particle aggregation, all leading to improved moisture infiltration and retention, reduced runoff of nutrients, and less soil erosion. Cation exchange capacity refers to the ability of soil to store nutrients (e.g., calcium, magnesium, and potassium) for future plant uptake. Organic matter, soil biota abundance and diversity, plant roots, and water and air movement affect soil structure formation and aggregation (Hartel, 1998; Paul and Clark, 1996). Soil stabilization, which is the ability of soil to maintain its structural integrity when subjected to natural or anthropogenic stress, occurs when there is some degree of aggregation of soil particles. The soils are then more resistant to soil degradation (e.g., erosion, loss of fertility, reduced filtering, and buffering capacity).

### B. Erosion

Erosion is a general term for the removal of the surface of the earth by abrasive actions of wind, water, waves, and glaciers. There are two types of erosion: geological and accelerated. The Grand Canyon in the United States is an example of the effects of geological erosion that occurred over millions of years. Geologic erosion in natural ecosystems is a slow process, typically occurring at a rate slower than the rate of soil formation. This is because the soil is protected by vegetation. For example, aboveground vegetation can reduce wind speed and roots help anchor the soil. Accelerated erosion happens when the rate of soil loss is higher than the rate of soil formation. This type of erosion occurs when the lighter individual particles of soil aggregates are detached and transported by wind or water. These particles can be blown for great distances as dust (the Dust Bowl of the 1930s in the United States) and be deposited to form new soils or washed into streams, rivers, and oceans. The risk of soil erosion depends on the natural conditions (climate, slope, vegetation cover, and soil) and land use (e.g., removal of the protective cover of vegetation). Erosion results in a deterioration of soil quality and imposes hazards to humans and organisms in terrestrial and aquatic ecosystems (e.g., loss of soil fertility, reduced plant productivity, loss of water and air quality, flooding, mudslides, and sedimentation).

Erosion is the primary factor degrading soils on a global scale (Fig. 1). In many areas of the world eroded land is no longer productive and is often abandoned. A recent estimate indicates the amount of arable land on earth is about 1.44 billion hectares ("Conserving Land") (Fig. 2). In many areas, continuous intensive agricultural cultivation leads to soil erosion, salinization, and desertification and often to the loss of the land from agricultural production. It is estimated that globally 25 billion metric tons of soil erodes each year from agricultural land (17 tons per cultivated hectare, or 4.3 tons per person) (Food and Agriculture Organization, 1992). To accommodate the rapidly increasing human population, agriculture will have to increase globally the quantity or quality of food produced on a given area or increase cropland area into marginal areas of lower soil quality. Thus, global efforts to conserve soils are of immediate importance for human well-being.

Water is estimated to be responsible for more than one-half the global soil degradation, followed by wind, chemical (e.g., salinization, acidification, pollutants, atmospheric nitrogen deposition, excessive fertilizers, pesticides and manures), and physical (soil compaction, water logging, subsidence) degradation. However, whether a region is affected more by wind or water is dependent to a large degree on its climate. Wind erosion is a more serious problem for agricultural lands in the arid and semiarid regions of the world (e.g., North Africa, the Near East, parts of Asia, Australia, northwest China, southern South America, and North America). Salinization, the increased concentration of salts in the topsoil, can occur in irrigated lands of the world due to increased evaporation, susceptibility of soils to salty groundwater, and invasion of seawater. These forms of soil degradation continue to decrease the availability of productive land for future food, fiber, and fuel production.

As high-quality and agriculturally productive land diminishes, other ecosystems (e.g., forests, wetlands, and meadows) become vulnerable. Within the past 3 centuries, 2.2 billion ha of forests (down from 4–6.2 billion ha worldwide) have been converted to agriculture. Forests generally have a high biodiversity and are a reservoir of hundreds of billions of tons of carbon stored in trees and soils. On a global scale, conversion of forests to croplands releases CO₂ into the atmosphere, contributing to global warming. Conservation of soil is
C. Effects of Conversion of Natural Systems on Soil Quality

In natural systems, the amount of SOM is maintained by high (generally greater than 90% of primary production) litter inputs, but in agricultural and grassland grazed systems due to harvesting the yield or herbivory, only about 50% of primary production contributes to SOM. In agriculture, the lower amounts of organic matter input to SOM results in less substrate (energy and nutrients) for the soil food web and fewer nutrients available to plants over time. In intensive agriculture, nutrients removed in crops or lost with erosion of topsoil have to be replaced with synthetic fertilizers in order to maintain high productivity.

Human activity changes soil chemical and physical properties, soil biodiversity, and soil quality. Agriculture (including forestry and grazing), urbanization, and industrialization are the most significant modifiers of soil quality. Tillage (plowing), one of the most successful agricultural practices, has been used for centuries to reduce weeds, aerate soils, and break up compacted soils. However, tillage along with other intensive agricultural practices, such as application of fertilizers and pesticides, use of heavy machinery, irrigation, and monocropping have resulted in degradation of soils in many regions of the world. Intensive tillage destroys soil aggregates, redistributes and enhances SOM turnover by increasing rates of decomposition, decreases the quality and quantity of organic matter, changes soil climatic conditions, and decreases soil moisture, soil food web complexity, soil fertility, and soil stability (Paustian et al., 1997). These changes increase microbial activity, organic matter oxidation, soil compaction, and water runoff. The timescale of agricultural soil degradation varies from decades to a few years (in tropical soils) to a single rainstorm. The loss of topsoil depends on the original status of soil quality, farming practices, climatic events, and the amount of topsoil (Jenny, 1980).
Construction (e.g., building of dams, urban development) and deforestation also contribute to an increased rate of erosion and loss of soil quality. Local effects of soil degradation are manifested on a landscape scale (watersheds, rivers, wetlands, oceans) through surface and groundwater pollution and soil sedimentation, with resulting impacts on the quality of water and air, and economic impacts on tourism and the fishing industry.

III. CONSERVATION AND RESTORATION

A. Methods of Soil Conservation in Agriculture

The rate of soil degradation presently exceeds the rate of soil formation in systems impacted by humans (Jenny, 1980; Jenny, 1984). All vegetative productivity depends to a great extent on the nutrient availability in the soil, but in managed systems, such as intensive agricultural systems with continuous monocultures, a disproportionate amount of nutrients are removed from the system as crop yield. Although synthetic fertilizers renew the nutrient pool necessary for plant growth, they do not regenerate the quantity or the heterogeneity in the quality of soil organic matter. Thus, slower decomposing organic materials (e.g., fertilizers such as manures) are used increasingly to replenish soil organic matter. Crop harvest also leaves the soil bare and exposed to erosion by rain and wind. Thus, the diminishing quality of soil resources in natural and managed ecosystems, and the increasing demand for food, fiber, and fuel supply, has resulted in a reevaluation and selection of land management strategies that enhance longevity and quality of soil ecosystems. These alternatives focus on maintenance of soil organic matter, soil fertility, soil biodiversity, soil structure, and soil stabilization, and a reduction of soil erosion as a means of sustaining ecosystems for the long term.

Methods to reduce soil erosion are directed at pro-
without protecting the surface of the soil (topsoil) and include tillage practices (reduced tillage, conservation tillage, no till, minimum till or zero tillage—techniques that use specially designed machines and herbicides for minimal impact on the soil); diversified farming or cover crops; polycultures; adding organic composts or mulches; crop rotations; plowing techniques, such as contour farming (plowing at right angles to the land to create ridges to hold water); terracing (leveling areas on a slope to prevent water runoff); and timing of plowing.

Methods of agricultural soil conservation have proven valuable whether used alone or in combination. In the United States, soil conservation methods have decreased soil erosion rates so that only one-third of the agricultural lands are eroding faster than the average rate of soil formation. Although this is a considerable improvement over the Dust Bowl days of the 1930s in the United States, soil conservation still appears to be necessary. A recent summary of long-term experiments has evaluated the methods of conservation tillage intensity and crop management used alone or in combination on soil carbon storage. The combination of reduced tillage, bare fallow, increased inputs of crop residue, and crop rotations (with use of perennial vegetation) was more efficient at restoring soil organic matter and soil carbon and reducing levels of soil erosion than any of these management methods used separately. Rates of erosion decreased in Creek Basin, Wisconsin, over the past 1-40-year period (historical and current data) because of improvements in local agricultural land management in 1975 through 1993. These studies provide evidence supporting the efficacy of soil conservation practices.

1. Reduced Tillage

One of the most important factors decreasing soil quality is tillage. Reduced tillage practices that incorporate crop residues into the soil are among the best alternatives to conventional tillage and have increased globally over the past 10 years. Although reduced tillage practices increase SOM and soil moisture and contribute in many cases to an improved soil food web, they may require a greater use of herbicides due to the establishment and spread of weeds and soil pathogens. In some situations, crop yields may decrease due to the presence of the weeds and pathogen diseases.

Conventional and reduced tillage systems appear to be economically comparable, particularly if crop rotation is used in the reduced tillage system. The reduced tillage system, however, promotes greater long-term benefits for ecosystems locally and globally. Soil biotic complexity is generally positively affected by the retention of crop residues, and the rate of decomposition generally changes from a microbial to a fungal pathway resulting in a slower pulsed release of nutrients for plant uptake. The retention of soil residues also results in a reduction in CO2 emissions to the atmosphere compared to conventional tillage, and a long-term conservation of beneficial soil chemical and physical properties.

In conventional high-input agroecosystems, high crop productivity is achieved by application of synthetic fertilizers (mainly nitrogen, potassium, and phosphorus), not by decomposition of soil organic matter. As SOM diminishes in conventionally tilled systems, N and P retention is reduced and significant amounts of added fertilizers (nitrogen and phosphorus) are lost from the system. Fertilizer runoff and nutrient leaching reduces surface and groundwater quality. At a regional and global scale, nutrients from agricultural fields are lost to the atmosphere or lead to eutrophication of coastal waters with effects on coral reefs, estuaries, and fisheries. For example, in the Mississippi River, the nutrient levels have doubled or tripled since the 1950s, resulting in large algal blooms that deplete oxygen from the water, killing fish and shrimp, as well as other organisms.

In contrast, in most reduced tillage systems, the quantity of nitrogen applied as fertilizer is reduced, and there is an increase in the amount of applied nitrogen that becomes immobilized in the soil's organic matter. The nitrate leaching potential is reduced and subsequently the risk of water pollution is lower. However, increased earthworm populations co-occurring with reduced tillage practices can result in increased soil porosity and higher nitrate leaching.

2. Diversified Cropping

Organic matter can also be conserved in soils by growing cover crops, termed diversified or multiple cropping, during fallowing. Advantages of cover crops are numerous. Cover crops not only reduce soil nitrogen leaching (nitrates become immobilized in plant biomass), but they also provide a relatively effective means for weed control. They also decrease erosion since the plants and not the soil intercept the raindrops. Cover crops are selected based on several attributes, including plant species that improve soil (e.g., depth of roots, type of organic matter produced, legumes that fix atmospheric nitrogen), their effect on plant pathogens and their predators, or for their economic benefit. Although multiple cropping requires an extensive knowledge on the type of crop, and in what sequence to grow cover crops, the benefits have been recognized for centuries. Multiple cropping systems are increasingly being used...
as a tool in modern agriculture. Today’s research provides scientific support for diversified agricultural systems as they maintain soil quality (SOM, soil fertility), low insect and disease occurrence, and high plant productivity.

3. Organic Amendments

Organic materials or mulches offer in many instances, a long-term option for maintaining soils with higher soil organic matter for long-term agricultural production. Composts provide sources of nutrients that promote plant productivity and soil quality. Other benefits include improved soil aggregation, soil aeration, water-holding capacity, and cation exchange capacity. Examples of organic amendments include green manures (herbaceous crops plowed under while green), chicken and cow manure, pig slurry, urban grass cuttings, and homeowner garden composts of leaves, grass, and food remains. When conformed to quality standards set up by national environmental agencies, composts such as those derived from municipal and industrial wastes have the potential to improve soil quality and at the same time reduce the conversion of land to landfills (land conservation). Currently, however, the supply of compost does not meet demands, and with the higher cost of compost as compared to synthetic fertilizers, the use of organic amendments is not an economical option for many growers in industrialized nations. An increase of consumer interest in the purchase of organically grown food products may increase future use of composted organic materials.

4. Bioremediation of Pollutants

Soil organic matter, soil structure, fertility, and stability—and thus soil quality—can be affected by pollutants. Chemicals may enter the soil system purposefully (e.g., fertilizers and pesticides) or accidentally (via spills or failures of technological processes). Depending on the final concentration, the presence of many hazardous and toxic chemicals can have an immediate or long-term effect on soil quality. Restoration of polluted soils through bioremediation not only avoids the risk of human and animal health hazards, but restores, to some degree, otherwise degraded land. Bioremediation, unlike other methods, takes advantage of soil biota or plants to detoxify contaminants in the soil.

Soil is a natural habitat to a diversity of microorganisms and other soil biota that assist in the reduction of soil contaminants. Microorganisms, for example, may use the pollutant as a substrate or energy source for their metabolism and change the composition of the compound to a less harmful chemical. Pollutants can be changed from harmful to environmentally safe compounds through natural bioremediation (use of indigenous microflora), bioaugmentation (addition of nutrients to soil to stimulate activity of indigenous microflora), bioventing (addition of gases such as oxygen or methane to stimulate activity of microflora), biosorption (inoculation of soil with exogenous microorganisms), landfarming (mixing healthy and toxic soil), and phytoremediation (use of plants) (Skipper, 1998).

Bioremediation offers an interesting and economical alternative to the conventional way of cleaning soil contamination. The entire treatment takes place at the polluted site as opposed to more traditional method of moving large amounts of soil to treatment facilities and then back to the field. Bioremediation is often cheaper, resulting in a savings on energy used to remove and transport soil, and typically has a lower environmental impact, as it depends on natural processes with no hazardous byproducts (Skipper, 1998). The major constraints of bioremediation apply to the nature of the pollutants, as microorganisms cannot access many new synthetic compounds. Bioremediation can be further complicated by soil contaminants that are mixtures of pollutants, degradation of which may require the presence of several types of organisms each capable of transforming a different chemical in the mixture.

5. Manipulating the Biotic Community

Another aspect of soil quality relates to the diversity of biotic communities. Native land transformation, agricultural management, and pollution significantly affect, in most cases negatively, the abundance and diversity of organisms (Fossness, 1999; Wall and Virginia, in press). The array of multiple plant species (trees, shrubs, grasses) seen in native systems contribute to soil stabilization because the heterogeneity of rooting distribution, rooting depth, and plant chemical composition help to maintain soil structure and a diverse biotic community. In conventional agricultural systems, monocropping or cropping of one plant species (if not one variety) provides homogeneous root morphology, root depth, and litter quality across the landscape. The differences between the belowground heterogeneity of natural systems and the belowground homogeneity imposed by monoculture agriculture must be a consideration when developing long-term soil conservation plans. Using polycultures or intercropping is an agricultural practice that mimics, to some degree, the belowground heterogeneity of natural systems. An unresolved question is whether creating or restoring the structure of soil by this practice will actually recreate and restore the function provided by the soil biota.
The abundance and diversity of microflora and fauna in soil provides beneficial services to humans and ecosystems, including, but not limited to, nutrient mineralization, biological control, nitrogen fixation, soil aggregation, and soil stabilization. In natural systems, the decomposition of organic matter (plant litter, dead roots, animals) to inorganic chemicals necessary for plant growth involves many different groups of soil organisms. Microflora (bacteria and fungi) influence the amount of C stored in the soil and are responsible for the global cycling of many minerals (e.g., N, C, S, P) (Gregorich et al., 1997). Microflora play a critical role by aggregating soil particles, which helps to prevent soil erosion. For example, they produce chemicals (e.g., polysaccharides) that bind soil particles, or have morphological structures (e.g., fungal hyphae) that connect soil particles. Microfauna (e.g., protozoa, nematodes) affect soil fertility by trophic interactions with microflora that increase available N and P for plants. Nematodes, for instance, by grazing on bacteria can alter bacterial abundance and activity and thus significantly affect the rate of organic matter turnover and nutrient availability. Invertebrates transport microflora throughout the soil. In addition to functions similar to microfauna, mesofauna (e.g., mites and springtails) feed on fungi and other small fauna (Brussaard et al., 1997). Macrobioflora (e.g., earthworms, termites, ants, and snails) comminute and redistribute organic matter within the soil profile and by burrowing affect soil physical properties (Lee and Foster, 1991). They have been termed ‘‘soil engineers’’ for their role in mixing soils and microflora. Tillage, monocropping, and additions of fertilizers and pesticides often reduce the diversity of many components of soil biotic communities, limiting at the same time their potential positive effects on many ecosystem processes. Improvement of soil quality will depend on integrated management practices that reduce the artificial energy inputs (fertilizers) and exploit the food webs of soil organisms. Maintenance of or exploitation of soil biodiversity for beneficial human services may require a significant change from intensive conventional tillage. Reduced tillage, for instance, promotes more diverse decomposer communities (often with a higher fungal proportion) and lower mineralization rates (reduction of N losses). A long-term addition of organic matter (compost or no till) promotes microbial and faunal activities through which restoration of nutrients and soil organic matter is possible. However, considerable research is needed to determine how these practices will affect the food web. Just as all soils differ in their composition, food webs vary with soil, climate, and the quality and quantity of organic matter. Therefore, different organic composts cannot be expected to create the same food web in all soils.

Another way to maintain soil biodiversity and belowground heterogeneity is development of agricultural mosaics. Traditional agriculture (still common in the tropics and many European countries) promotes a patchy landscape with the use of strip weed margins, hedgerows, and shelterbelts. These latter two types create vertical structures to reduce wind, and also provide microhabitats and refugia for many beneficial soil organisms. A huge area with a monocropping system increases the risk of pest outbreaks. Among many characteristics of agroecosystems with low pest potential are diversified crops in time and space, crop rotations, a structural mosaic of cultivated and uncultivated lands, the presence of perennial crops, and high crop genetic diversity. All these elements not only stimulate development of favorable habitats for biological control agents, but also they provide a reservoir for recolonization of organisms involved in decomposition, detoxification, and other renewal processes.

6. Conservation of Urban Soils

Human population growth has been followed by urbanization. Conversion of native land for housing, parking, roads, industry, landfills, mining, and so on not only reduces potentially arable land but also affects global nutrient cycles (C and N). Additional consequences of urbanization include an increased use of fertilizers and pesticides in home lawns and gardens. These contribute to soil and water pollution, decreased soil biotic complexity, and increased nitrogen runoff. With urban development it is important to adopt urbanization strategies that are safe environmentally. Proposed methods of environmental urbanization, including vertical rather than horizontal sprawl, smaller house acreage, and preservation of open spaces between developed areas, are already being socially accepted and implemented. On a smaller scale, planning home gardens based on the local climate and natural vegetation is becoming more popular.

B. Assessment and Monitoring of Soil Quality

The goal of providing future generations with an opportunity for a high quality of life cannot be detached from the preservation or improvement of environmental quality. In order to evaluate the status or change of the environment, assessment, monitoring, and regulation
programs are being developed. On the basis of monitor-
ing data, farming strategies, urban development tech-
niques, and industrial technologies are being modified
to prevent further degradation of a particular system.
Typically, to evaluate the quality of the soil system,
or its state of sustainability, the system attributes are
compared against native or minimally disturbed refer-
ence sites. This procedure can provide an initial esti-
mate of human impact (e.g., agriculture, urbanization,
and pollution) and through monitoring (an assessment
of a soil ecosystem attributes through time) gives us a
perspective on the positive or negative aspects of a
management strategy. To be effective, monitoring has
to be connected with the process of management deci-
sions; therefore adjustments or modifications of man-
agement strategies should accompany these decisions
(a process termed “adaptive management”).

1. Indicators

Indicators are measures that tell us about the status of
the environment over time. Monitoring of, for example,
air and water quality, as well as CO₂ concentrations,
are officially regulated by many countries. There have
been many indicators proposed for monitoring soil
quality status, but it has been difficult to select a single
indicator that would aid policy makers.

The quality of soil can be defined in terms of
many variables or combinations of variables. These
variables include items such as biodiversity, levels of
specific mineral elements and pollutants, levels of
primary productivity, or profitability. Some aspects
of high-quality soils may contradict each other. For
instance, higher levels of nitrogen in the soil may
stimulate primary production but may suppress bio-
diversity. The choice of an indicator is further compli-
cated by the spatial heterogeneity of soils. Different
soil systems may require a different suite of indicators.
In addition, soil has many functions and maintaining
those functions may have different impacts at different
scales. For example, continuous cultivation may in-
crease local crop productivity, but may deplete carbon
stored in the soil, contributing to global elevated
atmospheric CO₂. Theoretically, some optimal condi-
tions may have different impacts at different
scales. For example, continuous cultivation may in-
crease local crop productivity, but may deplete carbon
stored in the soil, contributing to global elevated
atmospheric CO₂. Theoretically, some optimal condi-
tions may have different impacts at different
scales.

In terms of chemical attributes, the major functions
of soil are to store and supply sufficient amounts of nutri-
ents to sustain an ecosystem’s primary productivity and
immobilize or detoxify hazardous compounds that are
toxic to plants and animals. Among the chemical soil
properties most useful as indicators of soil quality are
mineralogy, organic matter content, cation exchange
capacity, salinity, and pH (acidity or alkalinity). Soil
mineralogy determines bioavailability (adsorption and
precipitation) and mobility of nutrients in the soil solu-
tion and is dependent on the content and type of silicate

1. Indicators

Indicators are measures that tell us about the status of
the environment over time. Monitoring of, for example,
air and water quality, as well as CO₂ concentrations,
are officially regulated by many countries. There have
been many indicators proposed for monitoring soil
quality status, but it has been difficult to select a single
indicator that would aid policy makers.

The quality of soil can be defined in terms of
many variables or combinations of variables. These
variables include items such as biodiversity, levels of
specific mineral elements and pollutants, levels of
primary productivity, or profitability. Some aspects
of high-quality soils may contradict each other. For
instance, higher levels of nitrogen in the soil may
stimulate primary production but may suppress bio-
diversity. The choice of an indicator is further compli-
cated by the spatial heterogeneity of soils. Different
soil systems may require a different suite of indicators.
In addition, soil has many functions and maintaining
those functions may have different impacts at different
scales. For example, continuous cultivation may in-
crease local crop productivity, but may deplete carbon
stored in the soil, contributing to global elevated
atmospheric CO₂. Theoretically, some optimal condi-
tions may have different impacts at different
scales. For example, continuous cultivation may in-
crease local crop productivity, but may deplete carbon
stored in the soil, contributing to global elevated
atmospheric CO₂. Theoretically, some optimal condi-
tions may have different impacts at different
scales.

In terms of chemical attributes, the major functions
of soil are to store and supply sufficient amounts of nutri-
ents to sustain an ecosystem’s primary productivity and
immobilize or detoxify hazardous compounds that are
toxic to plants and animals. Among the chemical soil
properties most useful as indicators of soil quality are
mineralogy, organic matter content, cation exchange
capacity, salinity, and pH (acidity or alkalinity). Soil
mineralogy determines bioavailability (adsorption and
precipitation) and mobility of nutrients in the soil solu-
tion and is dependent on the content and type of silicate
clay minerals and the content of Fe and Al oxide and hydrous oxide minerals.

Soil organic matter (or soil carbon) is usually the most important factor affecting soil quality and productivity. SOM has a direct influence on physical and chemical soil characteristics and an indirect effect on plant production, and it is the major energy source for the abundance and diversity of soil organisms. Soil organic matter increases water retention and decreases runoff by preventing the sealing of the soil surface and promoting infiltration (and less erosion). When bound to soil particles as aggregates, it improves soil structure. The upper part of the soil typically has the most soil organic matter, generally about 1 to 10 percent, compared to 1 percent or less at greater depths. Mineral soils with high SOM are generally the most productive, and the carbon contained in the soil organic matter is an important component of the global carbon cycle. Because of the positive correlation between high levels of organic matter and desirable attributes of mineral soil, it has been suggested as a reliable measure of soil quality. Even though changes in SOM are slow and vary locally and regionally, SOM is reliable as a long-term, rather than short-term, indicator of change in soils.

Cation exchange capacity is another useful indicator of changes in soil quality. In general, a higher content of soil organic matter and clay particles has a positive effect on CEC. Soil pH can also significantly influence nutrient availability. For instance, Ca, Mg, and K deficiencies are often present in acidic soils, with Fe and Zn deficiencies appearing in alkaline soils. Trace elements and heavy metals may become more bioavailable in acidic soils. Plant productivity not only depends on the storage and supply of nutrients, but also on the ability of soil to detoxify pollutants. Phytotoxicity and bioaccumulation may occur in areas under strong influence of human activities (e.g., fuel burning, mining, and industry) particularly in soils deficient in organic matter and low in CEC and pH.

c. Biological Attributes of Soil as Indicators

The environmental impact of human activities can be indicated, as outlined earlier, by changes in the physical and chemical properties of the soil ecosystem. Despite their usefulness in characterization of the general status and quality of the soil, physical and chemical indicators have limitations. Do we know the biologically important thresholds of soil chemicals? Do we know how different sets of physical soil factors affect the bioavailability of soil chemicals? Are they stable enough to permit detection of environmental change? Are they easily interpretable? How useful are the indicators across regional and global scales? These and other questions have led many scientists to examine other indicators.

As biological attributes of the soil system directly relate and respond to physical and chemical attributes, soil biota have been investigated as indicators for soil quality. Many organisms are sensitive to changes in the physical and chemical properties of the soil-air or soil-water interfaces. Research shows that disturbances affect the soil biota differently depending on their physiology and life histories. Specific groups of organisms are associated with the soil at spatial (vertical versus horizontal, rhizosphere versus bulk soil) and temporal scales. Moreover, biota vary in their geographic distribution, therefore those groups of organisms that occur in all soils (nematodes, protozoa, microbes) are promising as indicators on a global scale, while those geographically limited groups (e.g., earthworms) may be more valuable on a local scale. The use of soil invertebrates as bioindicators might be problematic because the taxonomic and life history diversity of soil fauna requires extensive knowledge and training. As with soil physical and chemical indicators, no single bioindicator has been yet proposed.

Microbes, protozoa, nematodes, mites, and earthworms are the most noted indicators for soil quality monitoring. Changes in community structure of soil biota can be based on a direct analysis of the dynamics of all identified taxa, or on data derived from taxonomic or ecological indices. Measures for comparisons of soil communities include species, genus, family, or functional group level. Because identification of soil organisms to species or genus level is difficult, a higher level of taxonomic hierarchy or functional group level is often preferred. To relate the status of soil quality to ecosystem change, the ecosystem processes driven by soil biota (e.g., decomposition, mineralization, and respiration) can be measured.

Many of the smaller organisms in soil share a number of attributes that are necessary for useful indicators. These include high abundance for ease of monitoring, occurrence in all soils (microbes, fungi, protozoa, nematodes), sensitivity to changes in soil chemistry and pollutants (microbes, fungi, protozoa, nematodes, mites, earthworms), sensitivity to soil physical changes (most taxa), and representation of a wide range of groups in the soil food web (nematodes, mites). Both protozoa and nematodes have been used to monitor the effects of pesticides and heavy metals. Nematode community structure has also been used to illustrate the effects of natural ecosystem succession, environmental disturbance, and land management practices. Limita-
tions to both protozoa and nematodes as indicators include the difficulties in enumeration and identification of species and their applicability at larger geographic scales. This pattern of local scale specificity seems to apply to most soil biota. Although common indices for abundance, diversity, richness, or rate of a particular process for different ecosystems may not exist, the applicability of biomarkers at a local scale might provide a powerful long-term tool for soil conservation efforts.

2. Policies and Regulations

Adoption of soil conservation measures by farmers may not always be an easy choice. Surveys of farmers reveal that a conservation ethic is a less effective motivation for adoption of soil conservation practices than demonstration of economic benefits. However, farmers knowledgeable about erosion and soil degradation are more likely to adopt soil conservation practices than are uninformed farmers. Thus, education and access to information are important aspects of any conservation effort, as are regulations and policies that conserve or sustain the land.

When private incentives differ from societal incentives, the influence of government policies and regulations can have a major impact on managing the natural resources. In the United States, 25 percent of arable land is regulated with beneficial effects for soil conservation. Conservation tillage in the United States was part of the 1983 Food Security Act (FSA), which encouraged farmers to take erosion-control measures by providing farm subsidy payments. The U.S. Food, Agriculture, Conservation, and Trade Act of 1990 (amendment to FSA) established financial penalties and ineligibility for most farmer program subsidies to farmers who produced agricultural crops on wetlands that were converted after enactment. This act also established the Conservation Reserve Program, providing an opportunity for farmers to take highly erodable lands out of production by receiving annual rental payments from their 10-year contracts with the Department of Agriculture. The Federal Agriculture Improvement and Reform Act of 1996 introduced “planting” flexibility, giving farmers entering commodity programs freedom to choose crops on the contracted acreage. Under the same act, soil erosion control and wetland restoration regulations were improved. These few examples illustrate that government policies can slow land degradation.

The effects of land cultivation and soil erosion expand beyond the border of an agricultural field, state, or even country. Separate agricultural ecosystems are connected via a network of groundwater, streams, and rivers. Silt, sediments, nutrients, and other agricultural pollutants that are transported to streams, rivers, and ultimately marine systems can restrict possibilities for navigation, irrigation, food production, and fisheries, and can affect water and air quality. Thus, the local ecological impacts of agricultural practices have been recognized globally. This has resulted in national and international policy frameworks including such organizations as the Food and Agriculture Organization (FAO), the International Geosphere and Biosphere Program—Global Change in Terrestrial Ecosystems (IGBP—GCTE), the International Union of Soil Science (IUSS), Global Assessment of Soil Degradation (GASOD), and Global Change International Panel on Climate Change (IPCC), all of whom have been active in either research, communication, or assistance to countries requiring immediate soil conservation measures and agricultural improvement.

There are also many international agreements whose policies include or are directed at ensuring soil sustainability. Agenda 21 is an entire chapter to sustainable agriculture and rural development both at national and international levels. Among many others, it aims for policies on land reform, less environmentally destructive developments, and conservation and rehabilitation of soil resources. Another example is the Convention on Biodiversity, which encourages agricultural practices that sustain biodiversity and ecosystem functioning. As our knowledge increases about the components of soil and their complexity, we realize that soil is a rare natural resource that should be used wisely.

IV. SUMMARY

Soils provide the basis for the world’s food, fiber, and fuel production, as well as for the functioning of global ecosystems. They provide a habitat for a diversity of species comparable to the diversity of life above ground. Human activity changes the soil chemical and physical properties and biodiversity of soil. Agriculture, urbanization, and industrial development are the primary activities that are accelerating the loss of soil quality. Because ecosystems are so interconnected, degradation of soils has detrimental effects on all life. Practices exist that can be selected to conserve and allow the sustainable use of soils.

The most effective methods improving the quality of agricultural soils are reduced tillage, cropping mosaics,
multiple cropping, crop rotations, and application of organic amendments. Not only do they prevent further loss of soil organic matter and soil carbon (an important property of soil environments), but also loss of other favorable soil physical and chemical soil characteristics. To determine the status of soil environment, assessment and monitoring programs are being developed. Soil quality can be estimated with the use of various indicators reflecting physical, chemical, or biological soil attributes. The collected information serves as a basis for evaluation and modification of current management methods. Different soil ecosystems may require different suites of indicators.

When private incentives differ from societal, government policies and regulations can have a pronounced impact on managing soil resources. Since the problem of soil degradation crosses spatial scales (local, regional, and global), development and implementation of national or international policies is essential.

See Also the Following Articles

AGRICULTURE, SUSTAINABLE • AGRICULTURE, TRADITIONAL • ECOLOGY OF AGRICULTURE • GREENHOUSE EFFECT • POLLUTION, OVERVIEW • SOIL BIOTA, SOIL SYSTEMS, AND PROCESSES

Bibliography


SOUTH AMERICA,
ECOSYSTEMS OF

Luis A. Solórzano C.
The Woods Hole Research Center

I. Physical Environment
II. Biogeography
III. Major Ecosystems of South America

GLOSSARY

caatinga  Collective name assigned to the semiarid ecosystems of eastern South America. Some authors use this term as a general name for a type of thin Amazonian forests. In this chapter however, the term is reserved for the xeric caatinga region of northeastern Brazil.
cerrado  Name assigned by phytogeographers to the tropical savannas of the central Brazilian shield. Cerrado vegetation is further divided in four categories: (a) campo limpio (i.e., clean field), (b) campo sujo (i.e., dirty field or grasslands with scattered shrubs), (c) campo cerrado (i.e., closed fields or grasslands with numerous trees and shrubs), and (d) cerradão (i.e., when the vegetation is dominated by a closed canopy of trees).
chaco  Region located between the Paraná basin, the central Brazilian shield, and the Andes. The ecosystems present in the Chaco region include temperate grasslands, savannas, and arid and semiarid environments.
igapo  Floodplain of the blackwater rivers of the Amazon basin.
mangal/mangle  Name assigned to the vegetation of mangrove ecosystems.
morichal  Plant community characteristic of tropical savannas, it is seasonally flooded and the presence of the palm *Mauritia flexuosa* is conspicuous.
llanos  Region of tropical South America, in the east side of the Andean range of Colombia and Venezuela. The llanos landscapes occupy lower elevations (2–300 m), and the landforms vary from flat to rolling terrain. The dominant vegetation is savanna with areas of dry forest, gallery forest, and morichales.
pantanal  Largest wetland of South America drained by the Paraguay River and its tributaries.
páramo  Ecosystem of the high Neotropical mountains, found in South America above tree line in the northern Andean mountains of Colombia, Venezuela, Ecuador, Peru, and Bolivia. The dominant vegetation is similar to an alpine meadow or grassland; cacti and giant rosette plants are conspicuous.
puna  Ecosystem of the high Andean mountains found above tree line from Bolivia to southern Peru. The Puna comprises four distinct ecological regions: wet, dry, thorny, and desert Puna. The dominant vegetation types vary from grassland/shrub lands to desert.
varzea  Floodplain of the whitewater rivers of the Amazon basin.
SOUTH AMERICA CONTAINS a diversity of ecosystems from tropical, subtropical, alpine, and temperate environments. The geology, climate, and biogeographic history of the continent have important consequences in shaping the current geographic distribution of ecosystem types and their structural attributes and functioning. The stability of the old Precambrian shields and the recent orogenic evolution of the Andean range have been the principal factors that determine geomorphologic, edaphic, and continental climatic patterns that correlate with the geographic distribution of major ecosystem types. The Andean range causes an altitudinal zoning of different environments and dramatically increases the diversity of ecosystems that occur in the continent. Most of the area of South America lies in tropical regions and therefore tropical ecosystems dominate the landscapes. The main tropical ecosystems include tropical rain forest, dry forest, cloud forest, savannas, shrub lands, xeric formations, and high Andean ecosystems. Ecosystems of the southern temperate regions include temperate grasslands, the Mediterranean Matorral, and temperate forest and deserts. There are extensive wetland ecosystems across the continent occupying areas along the floodplains of major rivers, lakes, estuarine areas, and seasonally flooded savannas. Mangrove ecosystems occur along the tropical coastlines of both the Atlantic and Pacific oceans.

I. PHYSICAL ENVIRONMENT

South America occupies an area of approximately 18 million square kilometers and extends from about 11°N to 56°S, and from about 35°W to 81°W. The continent has been connected to Central and North America by the land bridge of the Panamanian Isthmus for approximately 3.5 million years and is bounded on the east by the Atlantic Ocean, on the north by the Caribbean Sea, and on the west by the Pacific Ocean. A general overview of the geology and physiography of the continent will help to explain the distribution and general properties of its major ecosystems.

The basic geomorphic features of South America are (a) the high Andean Mountain chain on the western border of the continent spreading from Venezuela to Chile, (b) the low plains of the piedmont that occupy the eastern side of the Andes, (c) the very large sedimentary lowlands of the Amazon valley, and (d) the continental shields outcropping on the east (Fig. 1). These major geotectonic categories of the continent have important differences in the origin, thickness, and stage of weathering and composition of the underlying materials, which are of essential importance to determine the type, depth, and physical composition of soils. Geomorphologic processes such as soil formation and erosion have produced characteristic surface features, which identify discrete areas that developed under different climatic conditions.

The central and eastern part of the continent is made of three cratonic areas: the Guyana Shield, the central Brazilian Shield, and the Coastal Brazilian Shield (Fig. 1). These cratonic areas have not been geologically disturbed since the Silurian (≈408.5 × 10^6 years ago). Large plateaus and rolling highlands occupy the most eastern part of the continent covering most of the Guyana and Brazilian shields. From remote geological times, in the east side the stable Precambrian shields of the continent have held positions above sea level and have been affected by simple deformations. Landscapes in these areas, therefore, have evolved continuously under conditions of direct exposure to the atmosphere, and this has resulted in the widespread development of old erosion surfaces. The shields consist primarily of crystalline basement complexes from the Precambrian period and the dominant rocks are granites, gneisses, and mica schists. The waste mantle may be very thick on the old erosion surfaces of the shields. The materials that form these rocks are at an advanced stage of in situ weathering and can be millions of years old. A shallower layer of waste material covers the landscapes of residual relief. The valley bottoms and lowlands are covered with drift material that can attain considerable depths.

Along the Andean belt (Fig. 1) there are Precambrian rocks masked by different younger deposits. Since the Paleozoic, the ocean has invaded pre-Andean areas several times, and extensive marine deposits mixed with metamorphic and intrusive rocks occur all along the range from Venezuela to Chile. One important consequence of marine transgressions (i.e., ingressions and regressions) is the inland deposition of rich marine sediments. Marine regressions disrupt the deposition...
of marine sediments on land and further erosion may increase the loss of previously deposited layers. In the early Devonian (408–362 × 10^6 years ago), the largest of all marine transgressions covered the Amazon and Parnaiba basins (Fig. 1), which remained submerged until the middle Devonian. In the upper Paleozoic (290–245 × 10^6 years ago) only the Parnaiba basin received mixed marine and continental sediments. In the early Mesozoic (245–× 10^6 years ago),olian and fluviatile sediments were deposited in the Parnaiba and Amazon basins. The Triassic sediments (208–× 10^6 years ago) came from continental areas and were deposited in an arid environment. During the Jurassic (208–145 × 10^6 years ago), the products of major volcanic activity in central and western Argentina were deposited in the Parana, Sao Francisco, Parnaiba, and Amazon basins (Fig. 1). During the late Cretaceous, marine depotsions also occurred over large areas of eastern South America.

Most of the current physical features of South

FIGURE 1 Major geotectonic components of South America. Adapted from Harrington (1973). 1 = Andean Belt; 2 = Amazon Basin; 3 = Central Brazilian Shield; 4 = Goiana Shield; 5 = Coated Brazilian Shield; 6 = Llanos; 7 = Parnaiba Basin; 8 = Sao Francisco Basin; 9 = Parana Basin; 10 = Chaco; 11 = Pampas; 12 = Pampean Massif; 13 = Paragonien Massif; 14 = Deseado Massif; 15 = Lowlands of Iquitos, Acre, and El Beni.
region and, therefore, influences the structure and function of ecosystems. Figure 2 illustrates the major South American climate types according to the classification of Köppen. The shape of South America is of paramount importance in determining its large-scale climatic patterns, which in turn influence the geographical distribution of ecosystems. South America has the shape of an acute triangle with most of its area situated in tropical latitudes and its southern portion extending well into high latitudes. Although the southern portion of South America reaches more than 20° farther south than the tip of Africa, the continent is so narrow at the southern end that the landmass of this region rather constitutes a continental peninsula. Because most of South America occupies tropical regions and the continental area decreases at higher latitudes, the property of climatic continentality is absent. Accordingly, the more important factors that determine the diversity of climatic regimes observed across South America are the major global atmospheric circulation patterns, the proximity to oceans, the Andean relief, and other coastal and inland topographic variations.

In general, tropical temperatures dominate the northern part of the continent, declining smoothly toward the south. Major variations in temperature are due to the Andean Mountains and ocean currents. The presence of the Andes causes climatic diversity over the continent due to altitudinal variations that change wind circulation patterns and to adiabatic responses of air masses along the altitudinal gradients of the mountain chain. South American environments can range from hot deserts and warm rainy climates in the lowlands to cold deserts, temperate, and iced polar climates in the Andean highlands (Fig. 2). Two ocean currents are particularly important to the regional temperatures: the cold Humboldt current that brings lower temperatures from the south up north along the West Coast and the warm Brazilian current that brings warmer temperatures down south along the eastern part of the continent.

Due to the lack of continentality, the distinct climates of South America are often characterized by differences in precipitation. Rainfall across most of the northern and central parts of the continent depends on the inter-tropical convergence (ITC). The ITC occupies its most northerly position from June to September, when is located between 5° and 9°N. Its advance and retreat migration does not take place parallel to the equatorial line. When the ITC drifts south, a prevalent High Pressure Cell in the Atlantic Ocean restrains its movement in the eastern part of the continent causing an arch of precipitation that reaches southeast Brazil in February-March, while dry conditions persist in the northeast.
This Atlantic high cell also causes higher precipitation in most of the East Coast north of Patagonia. The Andean chain also has a significant effect on the continental patterns of precipitation. Due to adiabatic cooling, the Andean windward slopes remove moisture from the air masses moving from the Pacific causing a large rain shadow that extends from east of the Andes from Bolivia to Tierra del Fuego (Fig. 2). Conversely, in the windward slopes of the Pacific coasts of Colombia and Chile the Andes cause exceptional orographic rainfall levels, which in the Choco region of Colombia can surpass 10,000 mm/year (Fig. 2). A similar effect is observed in the eastern Andean slopes of central South America where along the western edge of the Amazon basin moist winds rise and deliver higher precipitation levels than those register in lower elevations of the basin. In the north of Venezuela, the northern Andean ranges block the easterly trade winds and create a tropical savanna climate that extends from northern Colombia to Surinam (Fig. 2). In the western coast of the continent, between 27° and 10°S, the edge of the western Pacific high pressure cell combined with the cold Humboldt current produce an arid zone called in Chile the Atacama Desert (Fig. 2). The structural complexity
of the Andean relief includes inter-Andean cordilleras and valleys (e.g., the Eastern, Central, and Western Colombian cordilleras) and high plateaus (e.g., the Altiplanos of Peru and Bolivia) that further create many local climates according to variations in wind exposure and altitude. Other extra-Andean topographic features that produce orographic climatic modifications of rain shadows or rainfalls are the highlands of the Brazilian shield that block onshore winds between Porto Alegre and Bahia; the uplands of northeast Brazil that cause a dry rain shadow in the interior northeast region; the Guiana Highlands that block the trade winds and cause a dry winter effect in their leeward slopes; and the Sierras de Cordoba and Patagonian Plateau in Argentina, which affect air temperature and can collect moisture from humid air masses crossing Patagonia (Fig. 2).

II. BIOGEOGRAPHY

A. Early History and Associations

Biogeographers include central and northern South America in the Neotropical Kingdom, which also contains the Antilles and most of tropical Central America. Until approximately the early Cretaceous, South America, Africa, Australia, and Antarctica were joined together composing the supercontinent Gondwana, thus allowing for a continuous interchange of their ancient biotas. Africa and South America share many plant families like Annonaceae, Myristicaceae, Cecropiaceae, Sterculiaceae, and Bombacaceae. Some plant genera are very diverse in one continent and barely represented on the other; for instance, Mayua (eight species in America and one in Africa), Duvernoya (35 species in Africa and 3 in America), and Hyptis (400 species in America and 2 in Africa).

The southern cone of South America is included within the Holantarctic Kingdom, which includes the temperate mesic-adapted floras of southern South America, New Zealand, and southwestern Australia. The genera of trees Araucaria, Podocarpus, and Nothofagus are typical although not exclusive to the southern cone. Fossil records have also shown strong ancient faunal associations between South America, Africa, Antarctica, and Australia. For instance, some of the existing species of fishes of South America, as well as their parasites, show affinities with African groups. There are also close phylogenetic relations among groups of South American, African, and New Zealand invertebrates including, among others, arachnids and mollusks. After the separation of South America from the rest of Gondwanaland in the early Cretaceous, the biota of the continent was isolated for millions of years. Because of their isolation, South American flora and fauna are extraordinarily rich in endemic groups and species. Among the endemic plant families are the Bromeliaceae, Marcgraviaceae, Nolaniaceae, Cactaceae, Tropoeolaceae, Quinaceae, Lacistemataceae, Bixaceae, Brunellaceae, Krameriaeae, Cyclanthaceae, and palm tree genera such as Jubaea and Mauritia. It is estimated that about 7% of the total world's superior plants are endemic species from Brazil, 6.8% from Colombia, 3.2% from Venezuela, 2.1% from Peru, and 2% from Ecuador.

Brazil is the country with more mammal species reported in the world (324 of which 131 are endemic), Colombia the fourth (356), and Peru the ninth (344 of which 96 are endemic). Endemic faunal families include the armadillos Dasypodidae, the anteaters Myrmecophagidae, and the monkeys Cebidae. Particularly important are the marsupials that in South America are represented by 2 genera, 2 families, and 87 species. There are no fewer than 3000 bird species with 2 orders and 30 endemic families. The hummingbirds, which are endemic of the New World, have more than 250 species in South America and populate ecosystems from the Amazonian lowlands to the high Andean Mountains. The diversity of endemic reptiles is exceptional; there are 520 endemic species reported for Colombia, 468 in Brazil, 374 in Ecuador, 298 in Peru, and 293 in Venezuela, including species of snakes, turtles, and iguanas—which other than in South America only occur in the Fiji and Tonga Islands of the Pacific. Brazil is also considered the country with highest diversity of freshwater fish species in the world with >3000 species reported, Colombia the second (>1,500), Venezuela the fourth (>1,200), and Peru the seventh (>800). The diversity of invertebrates is also outstanding; for example, there are 350 endemic species of butterflies reported for Peru, 300 for Colombia, 200 for Brazil, and 200 for Bolivia.

B. Late Tertiary

Although it is likely that some exchange occurred between North and South America during the Eocene and Miocene, the land bridge of the Panama Isthmus connected permanently South, Central, and North America about 3.5 million years ago. The completion of the Panamanian land bridge in the early Pliocene was perhaps one of the most influential biogeographical events that have occurred in the continent and dramatically changed the ancient biota of South America.
C. Quaternary

During the past 2 million years, climatic fluctuations of dry glacial and wet interglacial periods affected both the extents of temperate and tropical areas. Palynological sequences indicate that at least during the Pleistocene major climatic changes affected speciation and biogeographical patterns in South America. The modern geographical distributions of many endemic taxa resulted from Quaternary speciation and redistribution of species overlaid with previously existing patterns. The explicit mechanisms that generated current biogeographical patterns are not fully understood, in part due to their inherently historical nature that is difficult to reconstruct.

Paleoecological studies initially assumed that the richness and complexity of plant communities was a consequence of continuous favorable growing conditions persisting over long time periods. Many areas have been identified as centers of endemism and proposed as tropical forest refugia during the period spanning 18,000 to 13,000 years ago (Haffer, 1969). According to this hypothesis, during the last four glaciating cycles of the Pleistocene, tropical forest contracted during dry glacial periods causing most of the Amazon Basin, as well as other forested areas, to be occupied by savanna vegetation. Continuous but isolated islands of forest (refugia) might have contained populations of plant and animal species that underwent speciation while they remained isolated. During mesic interglacial periods similar to currently existing conditions, the forest expanded again and new geographic patterns appeared including the signal of speciation processes that occurred within the refugia themselves. Paleoecological data have also shown that in the Andes the tree line has migrated down and up-slope by as much as 1,500 m, matching the frequency of each northern hemispheric ice age. The estimated cooling experienced during the glacial maxima at the Sabana de Bogota is 7 to 9°C. A consequence of the lowering of the tree line would be the connection of some of the high Andean páramo ecosystems, which are currently isolated on top of mountains. The modern distributions of several plant and animal species appear to agree with the refugial hypothesis.

An alternative explanation for the distribution of centers of endemism proposes changes in temperature rather than precipitation as the driving force of vegetation re-arrangements during glacial periods. It has been proposed that the areas of refugia were not “isolated islands of stability” but rather of “maximal disturbance” due to cooler glacial temperatures, reduced atmospheric CO₂, and moderate reductions of precipitation (Bush, 1994; Colinvaux, 1989). Far from being static places of refuge, under this hypothesis the regions of endemism would be the dynamic edges between the forest below and the cool-tolerant vegetation above.

The uplifting of the Andes during the Quaternary has also contributed to the isolation or dispersal of some taxa. While the mountain range has served as a dispersal corridor from north to south, it also represents a dispersal barrier (like the Atacama Desert) to some plants and animals. For instance, the floristic individuality of the High Andean deserts is attributed to the short history of these environments that appeared after the recent and fast uplifting of the cordillera. The age of these ecosystems might have not allowed enough evolutionary time for many species to adapt and disperse under those new conditions. There is evidence also that during interglacial dry periods the southern Andes cone experienced a cold and arid climate that served as an effective dispersal obstacle for rodents and plant species from the northern arid ecosystems.

An alternative explanation for the distribution of centers of endemism proposes changes in temperature rather than precipitation as the driving force of vegetation re-arrangements during glacial periods. It has been proposed that the areas of refugia were not “isolated islands of stability” but rather of “maximal disturbance” due to cooler glacial temperatures, reduced atmospheric CO₂, and moderate reductions of precipitation (Bush, 1994; Colinvaux, 1989). Far from being static places of refuge, under this hypothesis the regions of endemism would be the dynamic edges between the forest below and the cool-tolerant vegetation above.

The uplifting of the Andes during the Quaternary has also contributed to the isolation or dispersal of some taxa. While the mountain range has served as a dispersal corridor from north to south, it also represents a dispersal barrier (like the Atacama Desert) to some plants and animals. For instance, the floristic individuality of the High Andean deserts is attributed to the short history of these environments that appeared after the recent and fast uplifting of the cordillera. The age of these ecosystems might have not allowed enough evolutionary time for many species to adapt and disperse under those new conditions. There is evidence also that during interglacial dry periods the southern Andes cone experienced a cold and arid climate that served as an effective dispersal obstacle for rodents and plant species from the northern arid ecosystems.

Not only long-term events like Pleistocene glaciations or orogenic processes such as the uplifting of the Andes have had significant effects on the ecosystems of South America, but short-term climatic episodes like El Niño cause important impacts as well. For example, changes in water temperatures of the western Pacific during El Niño years trigger profound modifications in rainfall patterns across all of South America. Major large-scale ecosystem disturbances have been associated with El Niño: fires in savannas and forests of Amazonia; droughts in the llanos of northern South America, in the Chaco region and in northeastern Brazil; and floods in Peru, Ecuador, and Colombia. There is evidence that these perturbations affect the dynamics of populations and ecosystems, which over evolutionary time could modify the composition and geographic patterns of their biota.
III. MAJOR ECOSYSTEMS OF SOUTH AMERICA

Although it is possible to make some generalizations about the nature of the biomes of the continent and their major ecosystem types, it is important to bear in mind that across several spatial scales they are in fact complex mosaics of microhabitats. The categories of ecosystem types applied in this summary are very broad and based on the classification system developed by FAO-Unesco (1971), the satellite-based map of Stone et al. (1994), and the ecoregions study of the World Bank/World Wildlife Fund (Dinerstein et al., 1995).

In this review, very general categories are used and ecosystems are geographically delimited based on broad vegetation formations. Only some important features of these major ecosystem types are briefly discussed.

A. Tropical Forest

Tropical forests are among the most complex ecological communities of the world and offer a variety of niches that support an immense diversity of plants and animals. South America contains over 40% of the tropical forest of the world.

1. Distribution and Structure
   a. Tropical Rain Forest
      Tropical wet forests occupy the drainage of the Amazon and southern Orinoco basin, the piedmont of the Andes in northern and western South America, the Choco-Darien region between the Andes and the Pacific Ocean in western Colombia, and the southwest and Atlantic coast of Brazil (Fig. 3). Among other types of tropical rain forests are Babassu palm forests, which occupy the southeastern border of Amazonia, and the liana forests, which are found in southern Amazonia between the rivers Xingu, Tapajos, and Tocantins. The liana forest is an open forest with well-spaced trees, rich in woody climbers, and often completely entwined by lianas. It occurs on all soil types, but there are no detailed studies of this vegetation, nor explanation of the factors that cause its wide distribution.
   b. Tropical Semideciduous and Dry Forests
      These ecosystems occur part of the eastern edge of the Andes in Bolivia and Brazil, the inter-Andean valleys of Colombia, the northern edge of the Venezuelan Llanos, and the western regions of Ecuador and Peru (Fig. 3). Most of the South American regions originally covered by dry forests are today populated areas, and much of the primary vegetation has been cleared. Tropical semideciduous forests usually occur on alluvial lowlands, submontane, and mountain terrains up to 1,000 m. Between the Amazonian rain forest and the cerrado (discussed later) of central Brazil, there is a so-called transition forest, which comprises a large number of plant communities. Transition forests are generally less species rich than moister forest in terms of plant and vertebrate species; however, plant physiognomy diversity seems to be greater in dry than in wet forests. Both transition dry forest and savannas can be found under similar climatic conditions. In eastern South America, transition forests run along most of the boundary between Amazonia and the Guyana and Brazilian shields, and they vary in width from a few kilometers to over 100 km in some areas. The transition forest is taller and with a more closed canopy than the closed cerrado, cerrado (discussed later), but is lower and with a more open canopy than the typical Amazonian rain forests.
   c. Mountain Cloud Forests
      These ecosystems occur in the humid tropical slopes of the Andes from Venezuela to northwest Argentina and in the Santa Marta Massif in Colombia. They occupy defined altitudinal belts with the lower belts intermixing with lowland forests and the upper belts limiting with the high Andean grasslands at the tree line about 3,000 m above sea level. Fog is persistent in these ecosystems and the vegetation is luxuriant with an extraordinary variety and abundance of epiphytic forms and mosses.
   d. General Patterns
      Structural variations between tropical forests are mainly governed by differences in the amount and seasonality of precipitation and the temperature regimes. In general, the proportion of deciduous woody components increases along a gradient of rainfall, as the amount of annual precipitation decreases below 2,000 mm. Although exact boundaries between ecosystems are somewhat arbitrary—because vegetation is not always so sharply delimited—the transition from arid to rain forest seems to be primarily determined by continental climatic patterns, especially rainfall intensity and seasonality. When non-evergreen trees are present in wet tropical forests, they never shed all their foliage at the same time. A large number of woody species, of which most have evergreen foliage, characterize these...
ecosystems and the most frequent tree species rarely represent more than 15% of all species present.

In general, tropical forest have a vertical structure with several strata where the tallest trees form a discontinuous layer of emergent individuals, which in South America can reach heights of 30 to 40 m. A denser canopy grows below and consists of two or more layers of shorter trees, small palms, tall herbs, ferns, and shrubs. Lowland forest trees are generally shallow rooted (but see the discussion presented later on deep rooting) and many of them develop buttresses at the base of their trunks, an adaptation thought to provide mechanical support. The epiphytes and vines are conspicuous in tropical forests, with the epiphytic Orchidaceae and Bromeliaceae are usually prominent. The barks of lowland tropical rain forest trees are usually thin and smooth, and this morphological characteristic makes difficult the anchorage of epiphytic forms. In the Andean cloud forests, the heights of trees decrease at increased altitudes and their shapes become irregular.
2. Tropical Rain Forests

A mean monthly temperature of >26 and mean annual rainfall usually exceeding 1,800 mm characterizes the Af climate that coincides with the geographic distribution of tropical rain forest (Figs. 2 and 3). Some areas of tropical rain forests are the result of orographic precipitation. Isolated mountains (e.g., Sierra Nevada de Santa Marta in Colombia) or coastal escarpments (e.g., the Atlantic Brazilian Shield and the lower slopes of the Andes in the Pacific coast of Colombia and Ecuador) act as barriers to airflow, forcing moist air to ascend upslopes where adiabatic cooling generates clouds and precipitation. These geophysical conditions conduct to forested landscapes with a physiognomy similar to the Amazon wet evergreen forest. In areas of lowland tropical forest (e.g., Amazonia), minimum changes in temperature between summer and winter and continuous heating coupled with the high evapotranspiration rates of vegetation create a predictable pattern of daily convective rain that falls during the middle afternoon. It is estimated that up to 80% of the precipitation falling in Amazonia is retained within the hydrological cycle of the basin.

The geological material underlying the hydrographic network also influences the development of forest ecosystems. In the Amazon lowlands, large areas of forest are subject to inundation (varzea) and these ecosystems differ from those that occupy high terrain (terra firme). Rivers running from the Andes carry large amounts of silt and have a light brown color (whitewater). These large loads of silt are deposited in areas of low relief and form broad floodplains of high fertility. During the flooding season, these areas remain under several meters of water. In contrast, rivers running through the old Precambrian Shield are poor in nutrients and minerals and their waters are clear. The Shields have been exposed to millions of years of leaching and in order to conserve nutrients plants protect their foliage with tannins. The tannins leached into the rivers give a tannin coloration to the water (blackwater). The Rio Negro and other blackwater rivers also have floodplains, but they are different because of the lack of sediments and because plant communities are adapted to acidic conditions (igapo).

Because of the high temperatures and abundant rainfall of tropical forest ecosystems, there is intense chemical action on bedrock, and soils and all soluble components continuously leach, producing characteristic Ultisol and Oxisol soils. These soils are rich in iron, manganese, and aluminum, which stay behind after all other soil soluble constituents have been leached. Large amounts of these minerals arrange in stratified layers forming laterites.

Recent studies have shown that tropical forest exploit a larger volume of soil than previously thought. Soil depth is therefore quite important on water balance variability because it allows evergreen forest to keep evapotranspiration during dry periods by absorbing water from soil depths of more than 8 m (Nepstad et al., 1994). Although the most rapid changes in water availability take place at the soil surface, where it is depleted by plants, the process of hydraulic lift can take water from deep soil layers and discharge it into the dryer upper layers. During the day, the water potential (Ψ) gradient causes water to move from the ground into the roots, shoots, and through stomata to the atmosphere. During the night when stomata are closed, Ψ in the plant equals Ψ in the deep soil and a water potential gradient can move water from the deeper moist soil and plant interior into the dryer upper soil layers.

Decomposition and nutrient dynamics proceed rapidly under the warm, moist conditions of tropical forest ecosystems. High temperatures also favor an intense bacterial activity on upper soil layers and therefore there is no accumulation of humus. Many nutrients are stored in the forest biomass where they are kept from leaching. Nutrients are partitioned in different compartments of the forest and, on a weight basis, leaves contain the highest concentrations. Most of the phosphorus is stored in the leafy biomass and potassium and calcium in stem tissues. Nutrient losses in intact forests are generally low because of the high concentration of roots in the upper soil layers. The presence of mycorrhizal associations in roots enhances their nutrient uptake. Epiphytes also reduce the loss of nutrients washed away from standing biomass. Over long successional periods, the forest biomass is progressively stored in the woody components and nutrient accumulation in leaves decreases.

Invertebrates can account for up to 60% of the animal biomass of tropical forest ecosystems. About 19% of the woody fraction of the biomass is consumed by termites, beetles, and the larvae of insects. Litter is the main source of food for the animal community and leaves, flowers, and fruits are consumed mainly by insects, birds, bats, and small mammals. Chemical defenses are
common among tropical forest plants and many species have developed mechanical defenses by incorporating silica, lignin, and fibers in their tissues.

b. Tropical Semideciduous and Dry Forests
Along gradients from tropical rain forest to dry forest, the structural and functional attributes of the ecosystems change. Tropical deciduous forests correlate with climates with two well-determined seasons: a rainy period followed by a dry one. More than 50% of the tree species shed their foliage during the dry period. The predominant characteristic of the habitat of these forests is the prolonged seasonal drought, causing desiccation of the topsoil and lowering atmospheric humidity. The length of the dry season determines the degree of divergence in physiognomy and structure of the seasonal forest. From wet evergreen to dry deciduous forests, prolonged drought correlates with a reduction in physiognomy and floristic composition and the lowering of canopy height. At the same time, the degree of deciduousness of the trees of the overstory increases. The phenology of these ecosystems is determined by a seasonal climate with a period of intense summer rainfalls followed by a dry winter. In the southern dry deciduous forests, the winter season brings also changes in mean temperature down to 15°C. Although between 20 to 50% of the woody species found in these ecosystems are deciduous, these forests are still dominated by Amazonian genera widely distributed across South America, such as Parapiptadenia, Peltophorum, Cariniana, Leucaena, Tabebuia, Astronium, and others of lesser physiognomic importance.

Although tropical rain forests contain the highest diversity of species measured in terrestrial ecosystems, life-form diversity is higher in dry forest probably because of habitat heterogeneity. Woody plants in tropical rain forests tend to converge to a small number of life-forms. In drier more seasonal environments, the proportion of deciduous trees and shrubs increases, the presence of epiphytes decreases, and vines become more frequent. The presence of epiphytes is associated with high air humidity and the presence of dew. Along seasonality gradients succulent plants conspicuously increase, including those with Crassulacean acid metabolism (CAM). Succulent-stemmed plants have very stable water relations and are drought resistant. Evergreen woody plants dominate at both extremes of the gradient, and although they all belong to the C3 photosynthesis type, there are differences in their leaf structure and drought resistance, which allow them to function in both mesic and arid environments.

B. Savannas, Grasslands, and Shrub Lands
1. Distribution and Structure
Savannas and grasslands occupy nearly 25% of South America in both tropical and subtropical regions.

a. Tropical Savannas
These ecosystems are characteristic of the warm lowland tropics and occur in areas with a strongly seasonal rainfall regime and a dry period lasting from 4 to 7 or 8 months. Although a herbaceous cover consisting mostly of bunch grasses and sedges is dominant in savannas, the term ‘savanna’ embraces a variety of vegetation types found in tropical latitudes. An important characteristic of tropical savannas is a clear seasonality in their phenology and a period of lower activity associated with times of water stress. According to Sarmiento (1984), tropical savannas can be divided into four functionally distinct types: (a) semiseasonal savannas, with weak water stress conditions; (b) seasonal savannas, with a distinct rainfall seasonality and common fires during the dry season; (c) hyperseasonal savannas, where in addition to a marked dry season there is also a period of water flooding; and (d) ‘esteros,’ which are areas without a clear dry season, but an excess of soil water for most of the year.

The largest extension of savanna in South America occupies the old Precambrian Brazilian shield and has been dubbed by phytogeographers “Cerrados.” Savanna ecosystems are also found in the Orinoco Llanos of Colombia and Venezuela, in Suriname, Guyana, in some Amazonian regions of Brazil, and in Paraguay, Bolivia, Argentina, and the northern Chaco region (Fig. 3). The vegetation of the Guayan savannas is more related floristically to the llanos than the cerrado. Perennial grasses herbs and shrubs are the dominant life-forms in savanna ecosystems. They are well adapted to the seasonal climate, soils poor in nutrients and the fire disturbance regime that dominate in these environments. At least five phenological groups of grasses have been recognized in savanna ecosystems: perennial with a seasonal semidormant period, annual ephemeral with a short cycle, annual with long cycle, perennial with a seasonal dormant period, and of continuous flowering and growth. Tree species found in savannas have adapted to these conditions by increased allocation to underground biomass. Trees in woody savannas can reach 25 to 30 m in height, but their diameter rarely exceeds 50 cm. Some savanna tree species are described as subterranean trees because their roots can penetrate to depths of 18 m. Other species like Caratella ameri-
communities (smaller streams and the seasonally inundated palm vanna landscapes are the gallery forest along rivers and depths of 6 m. Other conspicuous components of sacana have shallow roots that can grow more than 20 m horizontally and secondary roots that can reach more or less recognizable stages in this continuum receive vernacular names. Dry grasslands without shrubs or trees are called “campo limpio” (i.e., clean field); grasslands with shrubs scattered are called “campo sujo” (i.e., dirty field); grasslands with numerous trees and shrubs are called “campo cerrado” (i.e., closed field); when the vegetation is dominated by a closed canopy of trees it is called “cerrado” (i.e., the vegetation is closed). The latter is a woodland composed of trees often 8 to 12 m or even taller, with ground vegetation reduced because of the shade. Many factors probably determine which of these forms of cerrado vegetation occurs in a given locality, among them toponymy and soil texture seem to have the main causal effect. Cerrados have a markedly seasonal climate and possesses a large characteristic flora of fire-resistant plants, including about 800 species of trees and large shrubs, and many times that number of herbs and subshrubs. The vast majority of these species are endemic to the Cerrado, an ancient vegetation formation dating back perhaps 50 million years.

b. Temperate Grasslands

These ecosystems occupy the eastern part of southern South America in Argentina, Uruguay, Chile, and southern Brazil. They expand across the pampas of Rio de la Plata in Argentina, the Chaco region, the semiarid region west of the humid pampa (Monte), and reach the edges of desert and semidesert areas of the Patagonian region (Fig. 3). The dominant landforms are large plains with a few table-shaped outcroppings no more than 300 m above the plains. The dominant climate across the large grassland region is characterized by a mean temperature of 10 to 20°C with an annual precipitation ranging from 400 to 1,600 mm. The dominant soil types are young and include Mollisols, Alfisols, Vertisols, and Entisols.

Vegetation in the Pampean temperate grasslands is composed of about 1,000 species and vary from grass-dominated communities in the east, to xeric woodlands and semidesert communities in the west. In the Argentinean Pampas, plant communities are dominated by species of Agrostis, Bouteloua, Elyonurus, Festuca, Pan-
b. Temperate Grasslands

There are different hypotheses that attempt to explain the physiognomic origin of the temperate grasslands. An initial hypothesis proposed that pre-Columbian peoples burned an original scrub forest. Another hypothesis points to an adaptive dominance of grasses due to a climatic effect that maintains a negative water balance during part of the year. Although there are significant regional differences, edaphic factors such as the fine texture of soils are proposed as major causes for the dominance of grasses. A lack of air spaces in the soil combined with summer dry periods would favor the competitive ability of grasses over trees. A variety of root growth habits is observed among both grasses and shrubs and the stratification of root biomass along the soil profile minimizes competition for water and nutrients. The dominant species of grasses are perennial, live for 3 to 4 years, and are short and leafy.

The soils of the southern grasslands usually present bad drainage conditions, relatively acid pH, and the humus layer can be 30 to 40 cm deep. Deeper soil layers are less organic, more alkaline, and composed mostly of clay or calcareous sediments. From west to east, soil texture of soils are proposed as major causes for the regional differences, edaphic factors such as the fine texture of soils are proposed as major causes for the dominance of grasses. A lack of air spaces in the soil combined with summer dry periods would favor the competitive ability of grasses over trees. A variety of root growth habits is observed among both grasses and shrubs and the stratification of root biomass along the soil profile minimizes competition for water and nutrients. The dominant species of grasses are perennial, live for 3 to 4 years, and are short and leafy.

The soils of the southern grasslands usually present bad drainage conditions, relatively acid pH, and the humus layer can be 30 to 40 cm deep. Deeper soil layers are less organic, more alkaline, and composed mostly of clay or calcareous sediments. From west to east, soil texture of soils are proposed as major causes for the regional differences, edaphic factors such as the fine texture of soils are proposed as major causes for the dominance of grasses. A lack of air spaces in the soil combined with summer dry periods would favor the competitive ability of grasses over trees. A variety of root growth habits is observed among both grasses and shrubs and the stratification of root biomass along the soil profile minimizes competition for water and nutrients. The dominant species of grasses are perennial, live for 3 to 4 years, and are short and leafy.
c. High Andean Grasslands

The most important functional attributes of these ecosystems are associated with the extraordinary amplitude of daily fluctuations in temperature and humidity, which are more pronounced than the seasonal ones. In these environments, high daytime temperatures follow nocturnal frost causing significant physiological stress in both plants and animals. Plants have evolved mechanisms that allow them to grow during the whole year. The frailejon giant-rosettes, for instance, at night protect their buds in the center of the rosette by adjusting the position of their leaves. In addition, these plants have large intercellular air spaces and excrete mucilage that contributes to their heat-storage capacity. The discontinuous areas of paramo ecosystems are isolated by montane wet (cloud) forests that grow at lower altitudes and as a result there is a large number of endemic species. Nutrient dynamics in high Andean ecosystems are not fully documented. In general, low temperatures and moist acid soils slow decomposition rates, and a thick layer of organic humus usually accumulates in sites where the relief allows deposition.

C. Xeric Formations

In South America there is a variety of xerophytic ecosystems, all marked by an intensive and prolonged period of seasonal drought during which the vegetation suffers water deficit.

1. Distribution and Structure.

A mean annual precipitation of 400 to 1,200 mm and more extreme daily changes of temperature characterize the BS climate. Soils are brown and yellowish, reflecting a higher content of humus. The BS climate coincides with the geographic distribution of semiarid environments and generally represents transitions between desert environments of the type BW and savanna environments of the type Aw (Figs. 2 and 3). In South America, xeric ecosystems are represented by the xeric scrubs of La Guajira peninsula in Colombia and Venezuela, the Paraguaña xeric scrub of Venezuela, the dry forest of west Ecuador, the Sechura and Atacama deserts in Chile, the Gran Chaco region in southern Bolivia, west Paraguay and north Argentina, and the Caatinga xeric scrub of northeastern Brazil (Fig. 3). The dry highland plateaus of the Puna region and the thorn woodlands of Patagonia are usually included within the dry formations of the continent. There is also a xeric formation called “agreste,” which forms a narrow strip between the Caatinga and the seasonal forest of eastern Brazil. Although not a strictly xeric ecosystem, the Matto Torral of Chile is also usually included as a semiarid type of vegetation (Fig. 3).

Deciduous forest trees that severely restrict their transpiration during the middle of the day—even during the rainy period—characterize the thorny woodlands of Colombia, Venezuela, and northeastern Brazil. Significant areas of the Guajira region consist of sand, moving dunes. Only a few trees are evergreen, but along the periodically dry channels of streams, riparian vegetation retain their leaves and reflect the presence of groundwater. The mean canopy height of the woody components is 3 m, sometimes reaching 7 m. The most important genera in these plant communities are Acacia, Bulnesia, Barlera, Carapaua, Capparis, Croton, Jacquinia, Opuntia, and Prosopis.

The caatinga ecosystems in the larger “Sertão” depression of northeastern Brazil are dominated by the genera Canavileslea, Chorisia, and Schinopsis. Communities on the dry tablelands are dominated by the genera Spondias, Commiphora, Cnidoscolus, Aspidosperma, and Mimosa. The driest parts of the Sertao are dominated by the genera Mimosa, Acacia, Combretum, and Aristida. The Agresta formation between the Caatinga and seasonal forest of Brazil has an open canopy of xerophytic trees (mostly Leguminoseae and Myrtaceae) with a very scanty understory of palms (e.g., Copernicia cerifera) and cactus.

The ecosystems of the Gran Chaco region are structurally diverse reflecting the varied environmental changes observed across the Chaco range. Precipitation increases from the center of the region, both to the east and west. The landforms are flat in the west and north and during the rainy season, flooding is common. The thorn woodlands of Patagonia occupy the west margin of the Pampa region, where annual precipitation is lower (300–500 mm) and the woody components of the vegetation (Acacias and palms) are characteristically undersized.

Due to the effect of the Andes and climatic circulation patterns (see section 1B), an arid zone occupies the zone from the coasts of Peru and Chile across the mountains into the Patagonian steppes (Fig. 3). The Pacific coastal area is a desolated desert where rain can be absent for years. The Atacama Desert is the driest climatic region of South America and occupies the eastern Pacific coast of the continent from about 35°S to 26°S (Figs. 2 and 3). In many areas there is no fog to provide moisture for vegetation and the cover is dominated by lichens and algae that depend on water condensed as dew. In those areas where fog and clouds develop, bringing a few millimeters of annual precipita-
tion, there are some short-cycled annual plants from the genera Aristida, Loasa, Maleshoria, Nolana, and Tropelion. On the west-facing slopes, between 100 and 150 m, moisture from fog carried by the Humboldt Current allows the presence of some vegetation. Some of the common species in these ecosystems are Prosopis juliflora, Distichilis spicata, Capparis angulata, Sporobolus indicus, Acacia macrocarpa, and the giant cactus Neoraimondia gigantea. In the high Andean slopes, the lowland forests and the grassland Puna limit the dry ecosystems of the Andean Peruvian valleys. The dominant genera there are Acacia, Bombaxes, Bursera, Caesalpinia, Cereus, Prosopis, Puya, and several cacti.

The Mediterranean Matorral occupies the coastal west-facing slopes of the Andes in central Chile (Fig. 3). Plant communities in these ecosystems are composed of evergreen shrub species that reach 3 to 5 m in height and have characteristic sclerophyllous leaves (e.g., Lithaea caustica and Fluorencia thurifera), while succulents and herbs cover the ground. Some Acacias and Nothofagus are present in communities of higher areas and those found further inland. Despite its latitudinal position, the flora of the Chilean Matorral is very rich, with more than 2,000 species reported.

2. Functional Aspects
There are only a few studies on the physiology and phenology of South American xeric ecosystems. Most of these ecosystems are characterized by at most 3 to 5 months of rainy season with no more than 400 to 800 mm/year. In tropical lowland areas, temperatures are 23 to 27°C year round. Relative humidity is stable around 50% and the potential evapotranspiration is usually 1,500 to 2,000 mm/year. Plant communities are affected by water deficit, which varies greatly from region to region, and so do the physiognomic and physiologic responses of the vegetation. Some annual plants germinate, grow, and reproduce only during periods of available water (e.g., most herbsaceous). Other plants evade stress conditions by restricting growth to periods of available water (e.g., most herbaceous). Other plants evasive mechanisms to control evapotranspiration. In addition, many species from these ecosystems produce leaves with chemicals that inhibit the germination of other plants. Crassulacean acid metabolism (CAM) is common and succulent perennials use this type of photosynthesis to endure the scarcity of water.

Annual precipitation of about 300 to 900 mm and a dry summer climate characterize Mediterranean ecosystems. The dry, hot weather favors the occurrence of fires, which are the main disturbance in these ecosystems. Topographical conditions are also important in Mediterranean ecosystems, where erosion and continuous rearrangement of surface materials are common. In Chile, the duration and intensity of the drought period increases southward and Matorral communities become dominated by deciduous evergreen shrubs and succulent species. The climatic conditions allow all shrub species to use C₃ photosynthesis to fix carbon and there are not species with C₄ metabolism. As already pointed out, the availability of soil resources is important in determining the investment that plants make on belowground biomass. In the Chilean Matorral it has been estimated that shrubs have an average root to shoot ratio of 1.5 and the roots spread in a radius at least two times the area of their crowns. Most of the fine absorbing roots, however, are found under the crown where they absorb nutrients from decomposing leaves and debris. Water availability appears to be the most important controlling factor of productivity in Matorral ecosystems. Annual productivity in these ecosystems range from 2 to 6.5 kg m⁻².

D. Coastal Ecosystems and Wetlands
1. Mangroves
In South America, mangroves occur along fragments of the Atlantic coastlines of Colombia, Venezuela, Guyana, Surinam, French Guyana, and Brazil, and the Pacific coastlines of Colombia, Ecuador, and Peru (Fig. 3). Mangroves develop well in areas where the ocean temperature remains above 24 to 27°C. Because of the effect of ocean currents (see section I.B.), the southern limits of mangroves in South America occur at about 25°S in the Atlantic coast of Brazil and 4°S in the Pacific coasts of Ecuador and Peru. These woody communities can grow up to 30 m. The mangle is best developed in coastal areas where high rates of sedimentation are enhanced by the mangroves themselves. Local topography, runoff channels, and sediment stability determine distinct zonal patterns. The vegetation of mangrove ecosystems (mangal) is adapted to salinity and is dominated by Rhizophora mangle, Avicennia sp. and in high ground by Laguncularia racemosa. Plant communities directly influenced by seawater but occupying high grounds are characterized by genera typical of seashore environments: Remirea, Salicornia, Canavalia, Atriplex, and Vigna. Mangrove areas are sometimes associated with river deltas and coastal swamps that have habitats in which the soils are inundated for at least part of the year. In the mouth of the Amazon, however, mangroves do not grow well. According to the quality of the water, it is
possible to distinguish freshwater swamps and brackish water swamps. To the first type belong the freshwater swamp forests of the Atlantic coast between the Orinoco delta and the mouth of the Paranhita River. Most of the other mangrove swamp forests belong to the second type. Other conspicuous coastal ecosystems are the Restingas of the Brazilian coast. These ecosystems are characterized by sand dunes where low scrubby vegetation grows sparsely. Their flora is rich in endemic species.

2. Wetlands

In South America, there are many dispersed wetland ecosystems. Some savanna areas of the Llanos region in the Orinoco watershed and the Rio Branco in Brazil have bad drainage and are seasonally inundated. In Ecuador, west of the Daule River, there is also a region of wetlands. The complex of grassland and woodland savanna vegetation occupies flat plains with tropical seasonal climate. The different types of vegetation are closely associated with the duration of the period of inundation. Areas with a long inundation period maintain communities dominated by the genera Euterpe and Mauritia. The genera Typha, Cyperus, and Juncus dominate swampy areas. In well-drained floodplains, the dominant genera are Panicum, Paspalum, and Thalia.

In western and eastern Amazonia there are also some areas of flooded grasslands but larger wetland ecosystems usually occupy the floodplain of large rivers (e.g., varzea and igapo in Amazonia) and the structure of the plant communities depends on the level and duration of the flooding period (see section III.A.2). The larger wetlands of the continent are in the Gran Pantanal region in Brazil Bolivia and Paraguay, and the Paramo flooded savannas in Argentina (Fig. 3). The Pantanal occupies the upper watershed of the Paraguay basin and is one of the world’s largest wetlands, which is flooded in more than 80% of its area during the rainy season (from May to December). The geomorphology is essentially flat and small high-ground islands separate hundreds of small lagoons. Cerrado tree species and palms occupy the high-ground islands of the Pantanal. The region accommodates large seasonal migrations of wildlife including aquatic birds and caimans.

E. Temperate Forests

In South America, temperate forest have a limited distribution and are restricted to relatively small areas. These ecosystems occur in the southern Andean areas of Chile and Argentina (Fig. 3). Evergreen temperate forest occupies areas of frost-free oceanic climate. These ecosystems are dominated by tall trees, some reaching heights up to 30 m, and composed of species such as Eucryphia cordifolia and Laurelia philippiana and some conifers like Araucaria araucana.

In the western humid slopes of the Andes, temperate rain forests are divided into the Valdivian and Magellanic forests. The Valdivian rain forest has an upper limit of 900 m and occurs from 45°S to 49°S where it becomes the Magellanic forest. A lush vegetation rich in evergreen species, abundant epiphytes, and lichens characterizes Valdivian ecosystems. The Valdivian forests can support an extraordinary biomass, in part because the climate is very humid and the mean annual temperature ranges from 10 to 12°C. In areas with excessive humidity, coniferous trees (Fitzroya patagonica)—whose stands can reach heights of 50 to 60 meters—replace common species of the rain forests. Coniferous forests also grow well at higher altitudes up to the tree line, which in the southern cone occur at about 1,600 to 1,900 m above sea level. Between the Valdivian forests and the sclerophyllous vegetation of central Chile there are limited extensions of deciduous forests composed of Nothofagus obliqua, which are found up to 1,200 m above sea level.

The Magellanic forests occupy the western slopes of the Cordillera Patagonica in southern Chile. In this region, the coastline is very rocky, curved by fjords and with an abrupt slope. Abundant rainfall, low evapotranspiration, and strong winds characterize the climate. The Magellanic forests are poor in species. As precipitation decreases rapidly to the east, in the leeward slopes deciduous forests occur and are composed almost of pure stands of Nothofagus with an understory of small evergreen trees. These forests grow to heights of about 30 m, but they become shorter as altitude increases. The tree line falls from 1,200 to 1,600 m in the north to about 600 m in the region of Ushuaia.

See Also the Following Articles

AFRICA, ECOSYSTEMS OF • CENTRAL AMERICA, ECOSYSTEMS OF • EUROPE ECOSYSTEMS OF • NEAR EAST ECOSYSTEMS • NORTH AMERICA, PATTERNS OF BIODIVERSITY IN

Bibliography

SOUTH AMERICAN NATURAL ECOSYSTEMS, STATUS OF

Philip M. Fearnside
National Institute for Research in the Amazon (INPA)

I. Original Extent of Terrestrial Ecosystems
II. Present Extent of Terrestrial Ecosystems
III. Human Use of Converted Areas
IV. Human Use of Remaining Natural Habitats
V. Threats to Remaining Natural Habitats
VI. Status of Protected Areas
VII. Priorities for Conservation

GLOSSARY

bioregion One of six biogeographic divisions of South America consisting of contiguous ecoregions. Bioregions are delimited to better address the biogeographic distinctiveness of ecoregions.
ecoregion A geographically distinct assemblage of natural communities that share a large majority of their species and ecological dynamics, share similar environmental conditions, and interact ecologically in ways that are critical for their long-term persistence.
ecosystem A set of interacting living and nonliving components in a defined geographic space. Ecosystems include both plant and animal communities and the soil, water, and other physical elements of their environment.
major ecosystem type Groups of ecoregions that share minimum area requirements for conservation, response characteristics to major disturbance, and similar levels of diversity (i.e., the rate of species turnover with distance).
major habitat type Groups of ecoregions that have similar general structure, climatic regimes, major ecological processes, diversity, and flora and fauna with similar guild structures and life histories.

THE TERM ECOREGION, as used in this article, refers to "natural" ecological systems, or terrestrial and aquatic areas as they were when Europeans first arrived in the New World. The original extent of natural ecoregions is presented, grouped by bioregion, major habitat type, and major ecosystem type. The definitions of these terms, given in the Glossary, are taken from Dinerstein et al. (1995); the rating codes are given in the footnotes to the table. Indications of the extent of remaining natural ecosystems, the threats to their continued existence, and the status of protected areas are discussed, together with priorities for conservation.

I. ORIGINAL EXTENT OF TERRESTRIAL ECOSYSTEMS

Ecosystems can be classified in many ways, making the number of categories vary widely depending on the use intended. Here, the system adopted by Dinerstein et al. (1993) is used. This divides the continent into 95 terrestrial "ecoregions," exclusive of mangroves. These are grouped into four "major ecosystem types": tropical broadleaf forests, conifer/temperate broadleaf forests, grasslands/savannas/shrublands, and xeric formations.
Within each of these categories are varying numbers of "major habitat types," such as tropical moist broadleaf forests. These are further divided into nine "bioregions." For example, the Amazonian tropical moist forests, for example, is a bioregion.

The 95 ecoregions, with their hierarchical groupings, are presented in Table I. Also included are the ratings for conservation status, biological distinctiveness, and biodiversity priority derived by Dinerstein et al. (1995). This study made a systematic survey of the status of natural ecosystems in Latin America and the Caribbean (LAC) and applied a uniform methodology to assigning priorities to these ecosystems for conservation efforts. The work was done for the United States Agency for International Development (USAID) by the WWF-US Biodiversity Support Program (BSP). The document is based on three workshops, plus consultations with relevant organizations and individual experts (the list of contributors contains 178 names).

The classification system is hierarchical, starting with four "major ecosystem types" (e.g., Tropical Broadleaf Forests), which are divided into 10 "major habitat types" (e.g., Tropical Moist Broadleaf Forests). These are crossed with 9 bioregions (e.g., Amazonia) and divided into 95 ecoregions (e.g., Rondônia/Mato Grosso moist forests). The system allows the priority of some ecoregions to be promoted upward based on uniqueness and regional representation, even if indicators of diversity and vulnerability are not so high.

The effort was unusual in emphasizing protection of areas with high beta diversity (a measure of the turnover of species along ecological gradients), as well as the more commonly used alpha diversity (species diversity within a habitat). In the case of mangroves, the diversity assessed is ecosystem diversity, including aquatic animal life. This avoids mangroves receiving the unfairly low diversity ratings that tend to result when assessments are restrained to terrestrial organisms, especially trees.

Although the ecoregions identified in Table I refer to "natural" (pre-Columbian) ecosystems, it should be emphasized that these had already been subject to millennia of influence by indigenous peoples prior to the arrival of Europeans. This influence continues today, together with much more rapid alterations from such activities as deforestation and logging done by nonindigenous residents. "South America" is taken to include the three Guianas (different from usage by the Food and Agriculture Organization of the United Nations (FAO)) and to exclude Panama (however, in the case of ecoregions that extend into Panama, the area estimates in Table I include the Panamanian portions).

The ecoregions are mapped in Fig. 1. The ecoregion numbering corresponds to Table I and also to the report by Dinerstein et al. (1995); the numbering presented here is not continuous, since the report also includes ecoregions in Mexico, Central America, and the Caribbean. Extensive bibliographic material on the delimitation of the ecoregions and on the state of knowledge about them can be found in Dinerstein et al. (1995).

Mangroves occur along the coasts of Brazil, the three Guianas, Venezuela, Colombia, Ecuador, and northern Peru. Dinerstein et al. (1995) divide them into five complexes: Pacific South America, Continental Caribbean, Amazon–Orinoco–Maranhão, Northeast Brazil, and Southeast Brazil. Each complex is further subdivided into 2–3 units, corresponding to distinct segments of coastline. Mangroves are essential to maintaining populations and ecological processes in surrounding marine, freshwater, and terrestrial ecosystems.

II. PRESENT EXTENT OF TERRESTRIAL ECOSYSTEMS

Unfortunately, information is not available on the present extent of each of the 95 ecoregions listed in Table I. Information on the extent of tropical forests in approximately 1990 is available from the FAO Tropical Forest Resources Survey (FAO, 1993). These data are tabulated by country in Table II. Nontropical areas are covered by a variety of national surveys (Harcourt and Sayer, 1996). National data are important because decisions regarding land-use policies and conservation are taken at the national level—not at the levels of bioregions or ecosystem types. Over half of the South American continent is represented by a single country: Brazil (Fig. 2).

An idea of the extent of existing ecosystems can be gained from measurements of land cover in 1988 made using 1 km AVHRR data. These are tabulated in Table III. It should be emphasized that many ecosystems can be heavily disturbed by logging and other activities without the change being evident on satellite imagery. This is true for Landsat TM imagery (30 m resolution) used for deforestation estimates in Brazil, and the limitations are much greater for 1 km AVHRR data. Brazil is the country with the most extensive satellite information on forest cover and its loss. Unfortunately, information on nonforest vegetation types such as cer-
rado is much less complete. Considerable confusion arises between the FAO (1993) classification and others such as the one adopted here because FAO classifies cerrado, caatinga, and chaco as “forests.”

Brazil’s Legal Amazon region originally had 4 million km² of forests, the rest being cerrado and other types of savannas. Agricultural advance was slow until recent decades because of human diseases (especially yellow fever and malaria), infertile soil, and vast distances to markets. These barriers have progressively crumbled, although a range of limiting factors restricts the extent and the duration over which many uses of deforested areas can be maintained (Fearnside, 1997a). Deforestation in the region has been predominantly for cattle pasture, with critical contributions to the motivations for the transformation coming from the role of clearing as a means of establishing land tenure and in allowing land to be held and sold for speculative purposes (Fearnside, 1993).

The Atlantic forests of Brazil (ecoregions 54 and 55) have been almost completely (≥95%) destroyed, mainly for agriculture, silviculture, and real estate development. Most of what remains of this extraordinarily rich ecosystem is in protected areas, but unprotected areas continue in rapid retreat. These forests are recognized as major “hot spots” of biodiversity (Heywood and Watson, 1995; Stotz et al., 1996).

In Andean countries, clearing by small farmers has predominated in driving deforestation, in contrast to the predominant role of medium and large cattle ranchers in Brazil. Migration from densely populated areas in the Andean highlands (altiplano) has led to settlement in lowland forests areas, with consequent upsurges in clearing (e.g., Rudel and Horowitz, 1993).

Savanna ecosystems have suffered heavy human pressure. The pampas of Argentina and the Uruguayan savannas of Uruguay and southern Brazil (ecoregions 120 and 121) have largely been converted to agriculture. The Brazilian cerrado, originally covering 2 million km², is the largest ecoregion in South America, as well as holding the largest number of species of any of the world’s savannas. The cerrado was largely intact until the mid-1970s. Clearing, especially for soybeans and planted pasture, reduced the cerrado to 65% of its original area by 1993 according to Landsat imagery interpreted by Brazil’s National Institute for Space Research (INPE). The advance of clearing has proceeded at an accelerating pace, speeded by infrastructure projects and an array of government subsidies.

III. HUMAN USE OF CONVERTED AREAS

Conversion of natural ecosystems to agroecosystems and secondary forests creates landscapes that maintain biodiversity to varying degrees. “Shifting cultivation” as practiced by indigenous peoples and by traditional nonindigenous residents (caboclos) in Amazonian forests maintains a substantial part of the original biodiversity. This contrasts with the effect of the vast expanses of cattle pasture that have replaced this, either directly or following a phase of use in pioneer agriculture by small farmers who have recently arrived from other places.

In densely settled areas along the coast of Brazil and in the southern portions of the country, agricultural use has gone through a series of “cycles,” such as sugarcane and coffee. The productivity of many areas has been damaged by soil erosion and other forms of degradation. Cattle pasture is often the land use replacing these crops. Since the 1970s, plantation silviculture (which now covers over 70,000 km²) and soybeans (130,000 km²) have made large advances.

In Argentina and Uruguay, cattle ranching and wheat and rice farming are major land uses. Natural vegetation is better represented in areas with little agricultural potential, such as mountain and polar areas and arid and semiarid zones.

IV. HUMAN USE OF REMAINING NATURAL HABITATS

Areas that remain under natural vegetation cover, rather than being converted to other land uses through clearing, are also subject to human use and alteration. Selective logging in tropical forests, for example, leaves much of the basic structure of the ecosystem intact, but also can lead to significant changes that can set in motion a sequence of events leading to complete destruction of the ecosystem. Logging leaves a substantial amount of dead biomass in the forest, including the crowns and stumps of harvested trees and all of the biomass of the many additional trees that are killed by damage
<table>
<thead>
<tr>
<th>Major ecosystem type</th>
<th>Major habitat type</th>
<th>Bioregion</th>
<th>Ecoregion name</th>
<th>Ecoregion number</th>
<th>Countries</th>
<th>Original area (km²)</th>
<th>Conservation status</th>
<th>Biological distinctiveness</th>
<th>Biodiversity priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical broadleaf forests</td>
<td>Orinoco Delta swamp forests</td>
<td>Orinoco Delta</td>
<td>Orinoco Delta swamp forests</td>
<td>18</td>
<td>Venezuela, Guyana</td>
<td>31,608</td>
<td>4</td>
<td>3</td>
<td>III</td>
</tr>
<tr>
<td>Guianan moist forests</td>
<td>20</td>
<td>Venezuela, Brazil, Guyana</td>
<td>248,018</td>
<td>5</td>
<td>2</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tepuis</td>
<td>21</td>
<td>Venezuela, Brazil, Guyana, Suriname, Colombia</td>
<td>49,157</td>
<td>5</td>
<td>1</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Napo moist forests</td>
<td>22</td>
<td>Peru, Ecuador, Colombia</td>
<td>369,847</td>
<td>4</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazonian tropical moist forests</td>
<td>Macarena montane forests</td>
<td>23</td>
<td>Colombia</td>
<td>2,386</td>
<td>3</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaguar/NEGRO moist forests</td>
<td>24</td>
<td>Colombia, Venezuela, Brazil</td>
<td>718,351</td>
<td>5</td>
<td>1</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uramba moist forests</td>
<td>25</td>
<td>Brazil, Venezuela, Guyana</td>
<td>288,128</td>
<td>4</td>
<td>3</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amapa moist forests</td>
<td>26</td>
<td>Brazil, Suriname</td>
<td>149,120</td>
<td>4</td>
<td>3</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guiana moist forests</td>
<td>27</td>
<td>Venezuela, Guyana, Suriname, Brazil, French Guiana</td>
<td>457,017</td>
<td>4</td>
<td>3</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paramaribo swamp forests</td>
<td>28</td>
<td>Suriname</td>
<td>7,760</td>
<td>3</td>
<td>3</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ucayali moist forests</td>
<td>29</td>
<td>Brazil, Peru</td>
<td>173,927</td>
<td>2</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Western Amazonian swamp forests</td>
<td>30</td>
<td>Peru, Colombia</td>
<td>8,313</td>
<td>4</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southwestern Amazonian moist forests</td>
<td>31</td>
<td>Brazil, Peru, Bolivia</td>
<td>534,316</td>
<td>4</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jurua moist forests</td>
<td>32</td>
<td>Brazil</td>
<td>384,059</td>
<td>5</td>
<td>2</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Va´rza forests</td>
<td>33</td>
<td>Brazil, Peru, Colombia</td>
<td>193,129</td>
<td>3</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purus/Madidi moist forests</td>
<td>34</td>
<td>Brazil</td>
<td>501,763</td>
<td>4</td>
<td>4</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rondônia/Maroa</td>
<td>35</td>
<td>Brazil, Bolivia</td>
<td>649,089</td>
<td>3</td>
<td>2</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region / Biome</td>
<td>Country(s)</td>
<td>Area (ha)</td>
<td>Level</td>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------</td>
<td>-----------------------------------</td>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beni Swamp and gallery forests</td>
<td>Bolivia</td>
<td>31,329</td>
<td>4</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tapajós/Xingu moist forests</td>
<td>Brazil</td>
<td>630,905</td>
<td>3</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tocantins moist forests</td>
<td>Brazil</td>
<td>279,419</td>
<td>2</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chocó/Darién moist forests</td>
<td>Colombia, Panama, Ecuador</td>
<td>82,079</td>
<td>3</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eastern Panamanian montane forests</td>
<td>Panama, Colombia</td>
<td>2,905</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northwestern Andean montane forests</td>
<td>Colombia, Ecuador</td>
<td>32,937</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Western Ecuador moist forests</td>
<td>Ecuador, Colombia</td>
<td>40,248</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Casca Valley montane forests</td>
<td>Colombia</td>
<td>32,412</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magdalena Valley montane forests</td>
<td>Colombia</td>
<td>49,392</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magdalena/Urubá moist forests</td>
<td>Colombia</td>
<td>73,060</td>
<td>2</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cordilleran Oriental montane forests</td>
<td>Colombia</td>
<td>66,712</td>
<td>3</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eastern Cordilleran Real montane forests</td>
<td>Ecuador, Colombia, Peru</td>
<td>84,442</td>
<td>3</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santa Marta montane forests</td>
<td>Colombia</td>
<td>4,707</td>
<td>3</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venezuelan Andes montane forests</td>
<td>Venezuela, Colombia</td>
<td>16,658</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catatumbo moist forests</td>
<td>Venezuela, Colombia</td>
<td>21,813</td>
<td>1</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peruvian Yungas</td>
<td>Peru</td>
<td>188,753</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bolivian Yungas</td>
<td>Bolivia, Argentina</td>
<td>72,317</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andean Yungas</td>
<td>Argentina, Bolivia</td>
<td>35,457</td>
<td>3</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazilian Coastal Atlantic forests</td>
<td>Brazil</td>
<td>233,266</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazilian interior Atlantic forests</td>
<td>Brazil</td>
<td>803,908</td>
<td>2</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orinoco tropical dry forests</td>
<td>Venezuela</td>
<td>44,177</td>
<td>2</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazonian tropical dry forests</td>
<td>Venezuela, Brazil</td>
<td>136,814</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazonian tropical dry forests</td>
<td>Venezuela, Brazil</td>
<td>136,814</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazonian tropical dry forests</td>
<td>Columbia, Peru</td>
<td>188,753</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazonian tropical dry forests</td>
<td>Bolivia, Argentina</td>
<td>72,317</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazonian tropical dry forests</td>
<td>Argentina, Bolivia</td>
<td>35,457</td>
<td>3</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central Andean tropical moist forests</td>
<td>Peru</td>
<td>188,753</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bolivian Yungas</td>
<td>Bolivia, Argentina</td>
<td>72,317</td>
<td>2</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andean Yungas</td>
<td>Argentina, Bolivia</td>
<td>35,457</td>
<td>3</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazilian Coastal Atlantic forests</td>
<td>Brazil</td>
<td>233,266</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazilian interior Atlantic forests</td>
<td>Brazil</td>
<td>803,908</td>
<td>2</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orinoco tropical dry forests</td>
<td>Venezuela</td>
<td>44,177</td>
<td>2</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazonian tropical dry forests</td>
<td>Venezuela, Brazil</td>
<td>136,814</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazonian tropical dry forests</td>
<td>Venezuela, Brazil</td>
<td>136,814</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

continues
<table>
<thead>
<tr>
<th>Major ecosystem type</th>
<th>Major habitat type</th>
<th>Bioregion</th>
<th>Ecoregion name</th>
<th>Ecoregion number</th>
<th>Countries</th>
<th>Original area (km²)</th>
<th>Conservation status</th>
<th>Biological distinctiveness</th>
<th>Biodiversity priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Andean tropical dry forests</td>
<td></td>
<td></td>
<td>Cauca Valley dry forests</td>
<td>77</td>
<td>Colombia</td>
<td>9,130</td>
<td>1</td>
<td>4</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Magdalena Valley dry forests</td>
<td>78</td>
<td>Colombia</td>
<td>13,837</td>
<td>1</td>
<td>4</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Patia Valley dry forests</td>
<td>79</td>
<td>Colombia</td>
<td>1,291</td>
<td>1</td>
<td>4</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sini Valley dry forests</td>
<td>80</td>
<td>Colombia</td>
<td>39,473</td>
<td>1</td>
<td>4</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ecuadorian dry forests</td>
<td>81</td>
<td>Ecuador</td>
<td>22,271</td>
<td>1</td>
<td>1</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tumbos/Puera dry forests</td>
<td>82</td>
<td>Ecuador, Peru</td>
<td>64,388</td>
<td>2</td>
<td>1</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Marathon dry forests</td>
<td>83</td>
<td>Peru</td>
<td>14,921</td>
<td>2</td>
<td>3</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Maracacibo dry forests</td>
<td>84</td>
<td>Venezuela</td>
<td>31,471</td>
<td>2</td>
<td>4</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lana/Falcon dry forests</td>
<td>85</td>
<td>Venezuela</td>
<td>18,178</td>
<td>2</td>
<td>4</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bolivian montane dry forests</td>
<td>86</td>
<td>Bolivia</td>
<td>30,308</td>
<td>1</td>
<td>3</td>
<td>II</td>
</tr>
<tr>
<td>Central Andean tropical dry forests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conifer/temperate broadleaf forests</td>
<td>Temperate forests</td>
<td></td>
<td>Chilean winter rain forests</td>
<td>87</td>
<td>Chile</td>
<td>24,937</td>
<td>2</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Valdivian temperate forests</td>
<td>88</td>
<td>Chile, Argentina</td>
<td>166,248</td>
<td>3</td>
<td>1</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Subpolar Nothofagus forests</td>
<td>89</td>
<td>Chile, Argentina</td>
<td>141,120</td>
<td>3</td>
<td>3</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Brazilian Araucaria forests</td>
<td>109</td>
<td>Brazil, Argentina</td>
<td>206,459</td>
<td>1</td>
<td>3</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Orinoco grasslands, savannas, and shrublands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Llanos</td>
<td>110</td>
<td>Venezuela, Colombia</td>
<td>355,112</td>
<td>4</td>
<td>3</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Guianan savannas</td>
<td>111</td>
<td>Suriname, Guyana, Brazil, Venezuela</td>
<td>128,378</td>
<td>4</td>
<td>3</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Amazonian savannas</td>
<td>112</td>
<td>Brazil, Colombia, Venezuela</td>
<td>120,124</td>
<td>4</td>
<td>3</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Beni savannas</td>
<td>113</td>
<td>Bolivia</td>
<td>165,449</td>
<td>2</td>
<td>3</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Chaco savannas</td>
<td>114</td>
<td>Brazil, Paraguay, Bolivia, Argentina, Paraguay, Bolivia, Brazil</td>
<td>1,982,249</td>
<td>3</td>
<td>1</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Eastern South American grasslands, savannas, and shrublands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Humid Chaco</td>
<td>116</td>
<td>Argentina, Paraguay, Uruguay, Brazil</td>
<td>474,340</td>
<td>3</td>
<td>4</td>
<td>IV</td>
</tr>
<tr>
<td>Region</td>
<td>Type</td>
<td>Country</td>
<td>Area</td>
<td>Rank</td>
<td>Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------</td>
<td>--------------------------</td>
<td>----------</td>
<td>------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cordeliers montane savannas</td>
<td>Argentina</td>
<td>93,700</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Argentine Monte</td>
<td>Argentina</td>
<td>197,710</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Argentine Espinal</td>
<td>Argentina</td>
<td>207,034</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pampeas</td>
<td>Argentina</td>
<td>426,377</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Uruguay savannas</td>
<td>Uruguay, Brazil, Argentina</td>
<td>336,846</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Orinoco flooded grasslands</td>
<td>Venezuela</td>
<td>6,403</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Orinoco wetlands</td>
<td>Venezuela</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Western Amazonian flooded grasslands</td>
<td>Peru, Bolivia</td>
<td>10,111</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eastern Amazonian flooded grasslands</td>
<td>Brazil</td>
<td>69,353</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>São Luis flooded grasslands</td>
<td>Brazil</td>
<td>1,681</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Guayaguei flooded grassland</td>
<td>Ecuador</td>
<td>3,617</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pantanal</td>
<td>Brazil, Bolivia, Paraguay</td>
<td>140,927</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pantanal flooded savannas</td>
<td>Argentina</td>
<td>36,452</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Santa Marta paramo</td>
<td>Colombia</td>
<td>1,329</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cordillera de Mérida paramo</td>
<td>Venezuela</td>
<td>3,318</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cordillera Central paramo</td>
<td>Ecuador</td>
<td>38,806</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cordillera Central paramo</td>
<td>Peru, Ecuador</td>
<td>14,128</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Central Andean puna</td>
<td>Bolivia, Argentina, Peru, Chile</td>
<td>183,808</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Central Andean wet puna</td>
<td>Chile</td>
<td>188,911</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Central Andean dry puna</td>
<td>Argentina, Bolivia, Chile</td>
<td>232,958</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Southern Andean steppe</td>
<td>Argentina, Chile</td>
<td>198,643</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Patagonian steppe</td>
<td>Argentina, Chile</td>
<td>474,757</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chilean grasslands</td>
<td>Argentina, Chile</td>
<td>39,383</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Central Andean matorral</td>
<td>Chile</td>
<td>141,643</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Continued...**
<table>
<thead>
<tr>
<th>Major ecosystem type</th>
<th>Major habitat type</th>
<th>Bioregion</th>
<th>Ecoregion name</th>
<th>Ecoregion number</th>
<th>Countries</th>
<th>Original area (km²)</th>
<th>Conservation status</th>
<th>Biological distinctiveness</th>
<th>Biodiversity priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deserts and xeric shrublands</td>
<td>Orinoco deserts and xeric shrublands</td>
<td>168</td>
<td>La Costa xeric</td>
<td>168</td>
<td>Venezuela</td>
<td>64,379</td>
<td>2</td>
<td>4</td>
<td>III</td>
</tr>
<tr>
<td>Deserts and xeric shrublands</td>
<td>Araya and Para xeric scrub</td>
<td>169</td>
<td>Venezuela</td>
<td>169</td>
<td>9,424</td>
<td>2</td>
<td>3</td>
<td>II</td>
<td></td>
</tr>
<tr>
<td>Northern Andean deserts and xeric shrublands</td>
<td>Galapagos Islands xeric scrub</td>
<td>170</td>
<td>Ecuador</td>
<td>170</td>
<td>9,122</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>I</td>
</tr>
<tr>
<td>Central Andean deserts and xeric shrublands</td>
<td>Guajira/Barranquilla xeric scrub</td>
<td>171</td>
<td>Colombia, Venezuela</td>
<td>32,404</td>
<td>2</td>
<td>3</td>
<td>II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central Andean deserts and xeric shrublands</td>
<td>Patagana xeric scrub</td>
<td>172</td>
<td>Venezuela</td>
<td>172</td>
<td>15,987</td>
<td>2</td>
<td>3</td>
<td>II</td>
<td></td>
</tr>
<tr>
<td>Central Andean deserts and xeric shrublands</td>
<td>Sachura Desert</td>
<td>173</td>
<td>Peru, Chile</td>
<td>173</td>
<td>189,928</td>
<td>3</td>
<td>3</td>
<td>III</td>
<td></td>
</tr>
<tr>
<td>Central Andean deserts and xeric shrublands</td>
<td>Atacama Desert</td>
<td>174</td>
<td>Chile</td>
<td>174</td>
<td>103,841</td>
<td>3</td>
<td>3</td>
<td>III</td>
<td></td>
</tr>
<tr>
<td>Eastern South American deserts and xeric shrublands</td>
<td>Guaramo</td>
<td>175</td>
<td>Brazil</td>
<td>175</td>
<td>782,606</td>
<td>3</td>
<td>3</td>
<td>III</td>
<td></td>
</tr>
<tr>
<td>Restingas</td>
<td>Northern Andean restingas</td>
<td>176</td>
<td>Venezuela</td>
<td>176</td>
<td>13,987</td>
<td>2</td>
<td>3</td>
<td>II</td>
<td></td>
</tr>
<tr>
<td>Restingas</td>
<td>Amazonian restingas</td>
<td>177</td>
<td>Brazil</td>
<td>177</td>
<td>10,248</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Restingas</td>
<td>Eastern South American restingas</td>
<td>178</td>
<td>Brazil</td>
<td>178</td>
<td>8,740</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

* Data source: Dinerstein et al., 1995.
* Conservation status codes: 1, critical; 2, endangered; 3, vulnerable; 4, relatively stable; 5, relatively intact.
* Biological distinctiveness codes: 1, globally outstanding; 2, regionally outstanding; 3, bioregionally outstanding; 4, locally important.
* Biodiversity priority codes: I, highest priority at regional scale; II, high priority at regional scale; III, moderate priority at regional scale; IV, important at national scale.
sustained during the logging process. Openings created in the canopy allow sunlight and heat to penetrate to the forest floor, drying out the fuel bed more quickly than in unlogged forests. Climatic variations such as those provoked by the El Niño phenomenon make logged forests especially susceptible to entry of fires. Ample opportunities for fires are provided as fields are burned to prepare land for planting and as cattle pastures are burned to control invading weeds. The fires burn slowly through the understory, charring the bases of trees as they go. Many of these trees then die, leading to a positive-feedback process whereby more dead biomass and canopy openings are provided and subsequent fires begin with greater ease, killing still more trees. This can degrade the entire forest within a few years (Nepstad et al., 1999).

Tropical forests are also used for “extractivism,” or the collection of nontimber forest products (NTFPs) such as rubber and Brazil nuts. This does relatively little damage to the forest, although extractivists do have an impact through hunting and through clearing for subsistence crops. The extractivist population can also play a protective role in defending the forest against encroachment by more aggressive actors such as ranchers and loggers. This is the basis of the extractive reserve system in Brazil (see Anderson, 1990).

Savannas are often grazed by cattle without cutting trees. Cerrado (ecoregion 114), “lavrado,” or Guianan savannas (ecoregion 111), the Pantanal wetlands (ecoregion 133), and the llanos of Venezuela (ecoregion 110) are among the savannas often used in this way. Increasing fire frequency, virtually all a result of human-initiated burning, can lead to shifts in species composition and to a drain of nutrients.

Aquatic ecosystems are traditionally exploited by fisheries. This alters the relative abundance of the species present. Use of watercourses as recipients for sewage and other pollutants also affects aquatic life in many ways.

FIGURE 1  Ecoregions for pre-Columbian vegetation of South America. Numbers correspond to Table 1. (Adapted from Dinerstein et al., 1995)
**TABLE II**

Area of Tropical Forest Present in 1990 (km²)

<table>
<thead>
<tr>
<th>Country</th>
<th>Tropical rain forests</th>
<th>Moist deciduous forest</th>
<th>Dry deciduous forest</th>
<th>Very dry forest</th>
<th>Desert</th>
<th>Hill and montane forest</th>
<th>All forests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolivia</td>
<td>0</td>
<td>355,820</td>
<td>73,460</td>
<td>0</td>
<td>40</td>
<td>63,850</td>
<td>493,170</td>
</tr>
<tr>
<td>Brazil</td>
<td>2,915,970</td>
<td>1,970,820</td>
<td>288,630</td>
<td>0</td>
<td>0</td>
<td>435,650</td>
<td>5,611,070</td>
</tr>
<tr>
<td>Colombia</td>
<td>474,830</td>
<td>41,010</td>
<td>180</td>
<td>0</td>
<td>0</td>
<td>24,900</td>
<td>540,640</td>
</tr>
<tr>
<td>Ecuador</td>
<td>71,500</td>
<td>16,690</td>
<td>440</td>
<td>0</td>
<td>0</td>
<td>31,000</td>
<td>119,620</td>
</tr>
<tr>
<td>French Guiana</td>
<td>79,910</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>79,910</td>
</tr>
<tr>
<td>Guyana</td>
<td>133,310</td>
<td>31,870</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19,120</td>
<td>184,360</td>
</tr>
<tr>
<td>Paraguay</td>
<td>0</td>
<td>60,370</td>
<td>67,940</td>
<td>0</td>
<td>0</td>
<td>270</td>
<td>128,390</td>
</tr>
<tr>
<td>Peru</td>
<td>403,580</td>
<td>122,900</td>
<td>100</td>
<td>2,690</td>
<td>1,840</td>
<td>147,770</td>
<td>679,060</td>
</tr>
<tr>
<td>Suriname</td>
<td>114,860</td>
<td>33,280</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>147,860</td>
</tr>
<tr>
<td>Venezuela</td>
<td>196,020</td>
<td>194,850</td>
<td>2,220</td>
<td>1</td>
<td>0</td>
<td>103,900</td>
<td>493,910</td>
</tr>
<tr>
<td>Total</td>
<td>4,389,320</td>
<td>2,787,330</td>
<td>433,050</td>
<td>2,691</td>
<td>1,880</td>
<td>828,460</td>
<td>8,440,870</td>
</tr>
</tbody>
</table>


* Includes cerrado, caatinga, and chaco.
FIGURE 2 Locations mentioned in the text.
<table>
<thead>
<tr>
<th>Country</th>
<th>Closed tropical forest</th>
<th>Recently degraded TMF</th>
<th>Closed forest</th>
<th>Degraded forest</th>
<th>Woodlands</th>
<th>Degraded woodlands</th>
<th>Savanna, grasslands</th>
<th>Degraded savanna, grasslands</th>
<th>Scrublands, desert, bare soil</th>
<th>Water</th>
<th>Snow, rock, ice</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>1.2</td>
<td>0.0</td>
<td>96.8</td>
<td>0.6</td>
<td>643.4</td>
<td>13.7</td>
<td>734.8</td>
<td>232.8</td>
<td>804.8</td>
<td>37.9</td>
<td>34.0</td>
<td>31.4</td>
<td>35.7</td>
</tr>
<tr>
<td>Bolivia</td>
<td>323.5</td>
<td>12.7</td>
<td>409.2</td>
<td>24.6</td>
<td>343.1</td>
<td>102.2</td>
<td>87.7</td>
<td>86.2</td>
<td>4.8</td>
<td>16.5</td>
<td>11.9</td>
<td>0.1</td>
<td>1.1</td>
</tr>
<tr>
<td>Brazil</td>
<td>3522.3</td>
<td>519.7</td>
<td>3686.0</td>
<td>1692.2</td>
<td>1559.9</td>
<td>330.0</td>
<td>740.0</td>
<td>179.4</td>
<td>0.0</td>
<td>0.0</td>
<td>80.9</td>
<td>0.0</td>
<td>124.0</td>
</tr>
<tr>
<td>Chile</td>
<td>0.0</td>
<td>0.0</td>
<td>134.1</td>
<td>29.1</td>
<td>75.2</td>
<td>29.8</td>
<td>101.1</td>
<td>14.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>3.8</td>
</tr>
<tr>
<td>Colombia</td>
<td>981.6</td>
<td>5.4</td>
<td>622.8</td>
<td>19.4</td>
<td>116.3</td>
<td>14.5</td>
<td>239.5</td>
<td>84.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.1</td>
<td>0.0</td>
<td>22.8</td>
</tr>
<tr>
<td>Ecuador</td>
<td>113.3</td>
<td>1.7</td>
<td>121.0</td>
<td>1.7</td>
<td>33.7</td>
<td>4.3</td>
<td>41.9</td>
<td>13.3</td>
<td>3.2</td>
<td>2.5</td>
<td>0.6</td>
<td>0.0</td>
<td>0.8</td>
</tr>
<tr>
<td>French Guiana</td>
<td>78.8</td>
<td>0.0</td>
<td>79.8</td>
<td>2.4</td>
<td>0.6</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Guyana</td>
<td>139.4</td>
<td>2.0</td>
<td>171.6</td>
<td>2.4</td>
<td>3.4</td>
<td>0.3</td>
<td>18.4</td>
<td>1.5</td>
<td>0.0</td>
<td>0.0</td>
<td>1.2</td>
<td>0.0</td>
<td>3.7</td>
</tr>
<tr>
<td>Paraguay</td>
<td>0.3</td>
<td>0.0</td>
<td>8.9</td>
<td>0.2</td>
<td>209.1</td>
<td>50.7</td>
<td>104.0</td>
<td>26.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>1.1</td>
</tr>
<tr>
<td>Peru</td>
<td>620.8</td>
<td>19.1</td>
<td>654.7</td>
<td>10.1</td>
<td>88.0</td>
<td>78.8</td>
<td>130.0</td>
<td>97.4</td>
<td>64.3</td>
<td>8.8</td>
<td>3.7</td>
<td>0.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Suriname</td>
<td>126.0</td>
<td>2.9</td>
<td>128.9</td>
<td>10.0</td>
<td>0.3</td>
<td>0.3</td>
<td>1.2</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
<td>1.1</td>
<td>0.0</td>
<td>3.3</td>
</tr>
<tr>
<td>Uruguay</td>
<td>1.4</td>
<td>0.0</td>
<td>2.1</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
<td>134.1</td>
<td>11.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.0</td>
<td>0.0</td>
<td>5.9</td>
</tr>
<tr>
<td>Venezuela</td>
<td>379.1</td>
<td>0.2</td>
<td>413.3</td>
<td>9.9</td>
<td>33.9</td>
<td>40.2</td>
<td>243.5</td>
<td>82.0</td>
<td>27.2</td>
<td>0.0</td>
<td>11.4</td>
<td>0.0</td>
<td>8.4</td>
</tr>
<tr>
<td>Unclassified</td>
<td>5909.9</td>
<td>565.4</td>
<td>6330.7</td>
<td>1803.7</td>
<td>3109.8</td>
<td>666.9</td>
<td>2642.0</td>
<td>808.5</td>
<td>1080.5</td>
<td>331.7</td>
<td>163.2</td>
<td>48.9</td>
<td>217.2</td>
</tr>
<tr>
<td>Continent (%)</td>
<td>53.4</td>
<td>3.2</td>
<td>56.9</td>
<td>10.2</td>
<td>17.6</td>
<td>3.8</td>
<td>14.9</td>
<td>4.6</td>
<td>6.1</td>
<td>1.9</td>
<td>0.9</td>
<td>0.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Category (%)</td>
<td>8.7</td>
<td>21.6</td>
<td>17.7</td>
<td>23.4</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V. THREATS TO REMAINING NATURAL HABITATS

A. Terrestrial Ecosystems

1. Deforestation

Deforestation is the dominant transformation of forested ecosystems that threatens biodiversity. In Brazil, which holds most of the continent's remaining forests, ranching is the dominant use for land once deforested. In the 1990s, soybeans began to enter forested regions, representing a new force in this process (they had already been a major factor in transformation of the cerrado since the 1970s). The most important effect of soybeans is not loss of forest directly planted to the crop, but the extensive infrastructure of waterways, railways, and highways that are built to transport soybeans and the inputs needed to grow them. The cycle of deforestation that has repeatedly occurred along Amazonian highways can be expected to accompany these new access routes.

Population growth is a fundamental contributor to deforestation and other forms of natural habitat loss. In recent years, however, the redistribution of population through migration has overshadowed the impact of absolute growth in population size. These include migrations from the semiarid Northeast of Brazil to Amazonia, from Parana to Rondonia, from the highlands of Bolivia, Peru, and Ecuador to the Amazonian lowlands and, in the case of Ecuador, to the Pacific lowlands as well.

2. Logging and Charcoal Manufacture

Logging is an increasingly important factor in Amazonia, and the catalytic role of this activity in increasing the flammability of the logged forest gives it potential impact far beyond its direct damage. So far, logging in Brazil has been dominated by domestic demand for sawn wood, plywood, and particleboard, which is almost entirely supplied from tropical forests rather than from silvicultural plantations (which produce wood for pulp and, to a lesser extent, charcoal). However, global markets for tropical timber are presently dependent on supplies from Asian forests that will soon come to an end if current rates of exploitation continue. In the 1990s, Asian logging companies began buying land and/or obtaining concessions in such countries as Brazil, Guyana, and Suriname, and pressure from global timber markets can be expected to increase in the future. Asian loggers are also the principal forces in clear-cutting the Valdivian and Nothofagus forests of Chile (ecoregions 88 and 89).

In eastern Amazonia, demand for charcoal for pig-iron smelting in the Carajas area is a potential threat to forests. Carajas, with the world's largest deposit of high-grade iron ore, is expected to be mined for 400 years at the present rate of exploitation. Wood from native forests is inherently cheaper as a source of biomass for charcoal production as compared to plantation-grown sources. Charcoal manufacture has an impact on the forest both through direct removal (including officially sanctioned forestry management systems) and by increasing the profitability of logging and deforestation (see Anderson, 1990).

Deforestation impacts are magnified by fragmentation and edge effects (Laurance and Bierregaard, 1997). This division of the remaining natural habitat into many small islands surrounded by cattle pastures or other highly modified land uses, together with forming edges with increased entry of light, wind, and foreign organisms, results in many changes in the remaining natural ecosystems. Most of these changes are forms of degradation, such as greatly increased mortality in the trees that provide the dominant component of forest structure. Vine loads on trees near edges also increase, leading to further increase in mortality and susceptibility to windthrow.

3. Other Threats

Climate change represents a major long-term threat to many South American ecosystems. The Intergovernmental Panel on Climate Change (IPCC) has prepared detailed reviews of potential climatic impacts on South America in its 1998 Special Report on Regional Impacts (Chapter 6) and its 2000 Third Assessment Report (Working Group II, Chapter 14).

Removal of fauna through hunting is a virtually universal consequence of proximity of human settlements to natural habitats. The removal of fauna can affect seed dispersal, pollination, and other processes needed for maintaining plant and animal communities. Introduc

Mangrove ecosystems are subject to some unique threats. Shrimp culture in mangrove areas has had severe impacts on the coast of Ecuador. Mangroves in Maranhão have been subject to pressure for charcoal manufacture. In São Paulo state mangroves have often suffered from oil spills and are also losing ground to real estate development. This has also affected restingas (ecoregions 176-178).
B. Aquatic Ecosystems

1. Dams

Hydroelectric dams have major impacts on river ecosystems by blocking fish migration, by eliminating rapids and replacing well-oxygenated running water with reservoirs that usually have anoxic water in their lower layers. The composition of fish present changes radically and undergoes a succession of changes as reservoirs age. Anoxic water released through the turbines severely reduces fish and freshwater shrimp productivity in the rivers downstream of the dams.

In Brazil, the 2010 Plan, released in 1987, listed over 300 dams for eventual construction in Brazil, independent of the expected date of completion. Of these, 65 dams were in the Amazon region. Economic difficulties have caused projected construction dates to be successively postponed, but the ultimate number of dams has not changed. Most contentious is the Babaquara Dam on the Xingu River, which would flood over 6000 km² of forest, much of it in indigenous areas. This has been renamed the “Alamira Dam” and appears in the current decennial plan for construction by 2013.

In Chile, the dams planned and under construction on the Bio-Bio River are expected to have major environmental impacts. The Ralco Dam is particularly contentious. In Uruguay, at least five major dams are planned for construction in the next few years.

2. Waterways

Industrial waterways, known as hidrovias in Brazil, greatly alter aquatic habitats. No less than seven waterways are under construction or planned for soybean transport on barges: the Paraguay–Paraná (Hidrovia do Pantanal), the Madeira River waterway, the Tocantins–Araguáia waterway, the Teles Pires–Tapajós waterway, the Capim River waterway, the Mamoré–Guaporé waterway, and the Rio Branco and Rio Negro–Orinoco waterways. Waterway construction involves blasting rock obstructions, cutting sharp curves, and dredging sediment from the river beds. The Corumba–Caceres stretch of the Hidrovia do Pantanal, if built, would lower the water level in the Pantanal wetlands (ecoregion 133), threatening one of the world’s most renowned concentrations of wildlife.

3. Other Threats

Other threats to aquatic habitats include sedimentation from soil erosion and landslides. This is severe, for example, in rivers draining steep areas of former Atlantic forest in the coastal mountains of Brazil. Mining for gold, tin, and diamonds in Amazonia can also inject large amounts of sediment into streams and rivers. Destruction of varzea forest (ecoregion 33) in Amazonia can affect aquatic life through loss of important fish breeding areas and food sources for fruit- and seed-eating fish. Destruction of varzea lakes and overfishing represent additional threats.

VI. STATUS OF PROTECTED AREAS

The choice and design of reserves depend on the financial costs and biodiversity benefits of different strategies. In Brazil, rapid creation of highly protected “paper parks” has been a means of keeping ahead of the advance of barriers to establishment of new conservation units, but emphasis must eventually shift to better protection of existing reserves (Fearnside, 1999).

Creating reserves that include human occupants has a variety of pros and cons (Kramer et al., 1997). Although the effect of humans is not always benign, much larger areas can be brought under protection regimes if human occupants are included. Additional considerations apply to buffer zones around protected areas. A “fortress approach,” whereby uninhabited reserves are guarded against encroachment by a hostile population in the surrounding area, is believed to be unworkable as a means of protecting biodiversity, in addition to causing injustices for many of the human populations involved.

VII. PRIORITIES FOR CONSERVATION

Indigenous peoples have the best record of maintaining forest, but negotiation with these peoples is essential in order to ensure maintenance of the large areas of forest they inhabit (Fearnside and Ferraz, 1995). The benefits of environmental services provided by the forest must accrue to those who maintain these forests. Development of mechanisms to capture the value of these services will be a key factor affecting the long-term prospects of natural ecosystems.

In the case of deforestation in Amazonia, a variety of measures could be taken immediately through government action, including changing land tenure establishment procedures so as not to reward deforestation, revoking remaining incentives, restricting road building and improvement, strengthening requirements for environmental impact statements for proposed development projects, creating employment alternatives, and, in the case of Brazil, levying and collecting taxes that discour-
age land speculation. A key need is for a better informed process of making decisions on building roads and other infrastructure such that the full array of impacts is taken into account.

Environmental services represent a major value of natural ecosystems, and mechanisms that convert the value of these services into monetary flows that benefit the people who maintain natural habitats could significantly influence future events in the region (Fearnside, 1997b). Environmental services of tropical forests include maintenance of biodiversity, carbon stocks, and water cycling. The water cycling function, although very important for countries in the region, does not affect other continents as the first two services do. At present, avoiding global warming by keeping carbon out of the atmosphere represents a service for which monetary flows are much more likely to result from international negotiations. Activities under the United Nations Framework Convention on Climate Change (UN-FCCC) are at a much more advanced stage of negotiation than is the case either for the Biodiversity Convention or for the “Non-Binding Statement of Principles” and possible future convention on forests.

In the case of carbon, major decisions regarding credits for tropical forest maintenance are likely to be taken at the sixth Conference of the Parties (COP-6) to the Kyoto Protocol, at the end of 2000, considering the IPCC Special Report on Land Use, Land-Use Change and Forestry (SR-LUCF), released in May 2000. Regardless of what is decided at COP-6, global warming is a permanent consideration that can be expected to receive increasing weight in decision making. The threats to natural ecosystems in South America are many, and recognition of the multiple environmental services provided by them is a key factor in ensuring that substantial areas of each of these ecosystems continue to exist, thereby maintaining their biodiversity.

Acknowledgments

I thank Eric Dinerstein and the World Bank for permission to publish Fig. 3 and Table I, and Tom Stone and the American Society for Photogrammetry and Remote Sensing for permission to publish Table II. Brazil’s National Council of Scientific and Technological Development (CNPq) (523980/96-5) and National Institute for Research in the Amazon (INPA PPI 1-3160) provided financial support. S. V. Wilson and two anonymous reviewers made helpful comments on the manuscript.

See Also the Following Articles

AMAZON, ECOLGY OF • DEFORESTATION AND LAND CLEARING • FIRES, ECOLOGICAL EFFECTS OF • GRAZING, EFFECTS OF • INDIGENOUS PEOPLES, BIODIVERSITY AND • LOGGED FORESTS • RAINFOREST LOSS AND CHANGE

Bibliography


SOUTHERN (AUSTRAL) ECOSYSTEMS

Robert S. Hill* and Peter H. Weston†
*University of Adelaide and Royal Botanic Gardens, Sydney

I. Introduction
II. History
III. Patterns
IV. Processes
V. Conclusion

GLOSSARY

continental drift The movement of continents as a result of plate tectonic processes. This was especially important for Southern Hemisphere continents.
disturbance A set of processes that lead to an ecological point crisis, that is, an event or events that are not part of the normal environment for the organisms that are influenced by it and that can be catastrophic in its impact.
Eucalyptus The dominant tree genus in Australia today.
Gondwana The southern supercontinent, named after an ancient kingdom in India. Gondwana rifted apart over many millions of years, giving rise to the current Southern Hemisphere landmasses.
Nothofagus A genus of Southern Hemisphere trees with a classic “Gondwanic” distribution in southern South America, southeastern Australia, New Zealand, New Caledonia, and New Guinea.

IN THIS ARTICLE we describe the biogeographic patterns on which the concept of southern ecosystems is based. We outline how several different research traditions have interpreted those patterns, briefly discussing the strengths and weaknesses of each approach. Finally, we discuss the evolutionary, geological, and climatic processes that might have caused or constrained the patterns. It is easy to show that the living biotas of these landmasses are the result of their post-Gondwanic history, and by comparing them, we demonstrate processes that influenced the entire hemisphere during this phase leading to the modern world. We concentrate mostly on Australia, where both the extant biota and the fossil record are relatively well documented, comparing the other landmasses to it. The impact of European settlement in the Southern Hemisphere is not considered, since despite its magnitude it is well documented elsewhere.

I. INTRODUCTION

A large number of plant taxa are shared by New Zealand, southern South America, and southeastern Australia (including Tasmania) but are found nowhere else. Joseph Hooker was so impressed by this pattern that in 1893 he wrote in his Introductory Essay to the Flora Novae-Zelandiae:

“… many of the peculiarities of each of the three great areas of land in the southern latitudes are
representative ones, effecting a botanical relationship as strong as that which prevails throughout the lands within the Arctic and Northern hemisphere zones, and which is not to be accounted for by any theory of transport or variation, but which is agreeable to the hypothesis of all being members of a once more extensive flora, which has been broken up by geological and climatic causes.”

Thus was born the concept of “southern (austral) ecosystems.” Hooker knew nothing of the geological basis for the patterns that he observed but he thought that they demanded explanation. We now know that the “once more extensive flora” about which he speculated did exist and that the “geological and climatic causes” that he hypothesized were the consequences of continental drift. “Southern (austral) ecosystems” thus comprise an historical entity rather than an ecological class as one finds in a volume of Ecosystems of the World.

The Southern Hemisphere is dominated by oceans, and the development of these oceans is critical in understanding the evolution of terrestrial biodiversity of this part of the world. All major southern landmasses, along with some that are now mostly in the Northern Hemisphere, were once part of the supercontinent Gondwana. The breakup of Gondwana, and the consequent movement of the continents that were once part of it, produced vicariant distribution patterns, as observed by Hooker, but it is also had other profound effects on the Gondwanan biotas.

II. HISTORY

In 1859, Joseph Hooker further emphasized the close relationship between the floras of the southern landmasses in his Introduction Essay to the Flora of Tasmania and insisted in maintaining that this must be the relic of a continuous, ancestral flora. His friends and colleagues soon disagreed. In 1859, Charles Darwin in his Origin of Species cited Hooker’s austral floristic pattern but not his geological speculations, explaining the pattern primarily as the result of repeated, independent episodes of long-distance dispersal across the Southern Ocean. In 1880, Alfred Russell Wallace in his Island Life, while noting that a number of animal taxa, such as the ratite birds, also showed an austral distributional pattern, chose to explain this as the result of waves of southward dispersal from the Northern Hemisphere, followed by extinction in the north. He agreed with Darwin that the austral floristic relationship must be the result of long-distance dispersal. Neither had any time for geological instability on the scale that Hooker had proposed.

These pioneering biogeographers spawned different, competing research traditions that remain alive and distinguishable today. Hooker inspired an intellectual lineage of phytogeographers to search for repeated distributional patterns and to postulate general explanations for them, invoking dispersal and vicariance of entire floras. This approach was applied in extraordinary breadth by Croizat in his “panbiogeographic” analyses of numerous plant and animal distributions. Croizat’s work produced not only a general distributional pattern (“generalized track”) crossing the Southern Ocean, consistent with Hooker’s analysis, but plenty of other general patterns too, including tracks spanning all ocean basins. After 100 years, the Hookeran tradition was still looking for an explanatory geological theory.

A competing (and for long, dominant) research tradition, inspired by Darwin and Wallace, flourished among zoogeographers, paleontologists, and many phytogeographers. The primary aim of this approach was to construct plausible scenarios describing the centers of origin and dispersal histories for particular taxa, based chiefly on their patterns of occurrence in the fossil record, knowledge of their phylogeny, the ecology of extant species, and knowledge of geological history. An example of this tradition is the work of Darlington (1965).

Corroboration of the theory of continental drift through the discovery of plate tectonics in the 1960s vindicated Hooker’s approach by providing the former land connections that he had predicted, in the form of the ancient supercontinent Gondwana. The transformation from “austral flora” to “Gondwanan biota” did, however, involve a substantial expansion in scope. Its potential boundaries shifted northward to include the whole of South America and Africa, Madagascar, India, all of Australia plus southern New Guinea, and New Caledonia. Refinements in geological reconstruction progressively added more fragments to Gondwana, such as eastern Sulawesi, the “Outer Melanesian Arc” (Solomons to Vanuatu and Fiji), and, eventually, large parts of eastern Asia.

The immediate effect of this geological revolution on biogeography was to flip the conventional null hypothesis in historical biogeography from long-distance dispersal to dispersal over land. “Darwinian” biogeographers changed the paleogeographic maps on which they plotted dispersal routes but continued to use the same meth-
ods as before. A new question was added to their task list: “To what extent can extant distributions be explained by continental drift?”

Different problems exercised the minds of “Hookerian” biogeographers. While a number of Croizat’s tracks were corroborated in surprising detail as vicariant patterns, others, such as his trans-Pacific tracks, remained geologically anomalous. It could be argued that all approaches to historical biogeography had been shown by continental drift to be inadequate in some way or other.

The heuristic success of Croizat’s track method and the emergence of new cladistic techniques for reconstructing phylogeny inspired the development of a new approach, cladistic (or vicariance) biogeography, in the 1970s and 1980s (see Humphries, this volume). This is based on the idea that historical biogeographic relationships can be summarized in the same way as phylogenetic relationships: in the form of tree diagrams. However, cladistic biogeographic trees connect areas of endemism occupied by taxa, rather than the taxa themselves. Methods of analysis have been developed that reconcile “area cladograms” for different taxa to produce “general-area cladograms,” which summarize episodes of biotic dispersal and fragmentation. The optimistic expectation of this approach is that general patterns will dominate such analyses, producing a small number of general-area cladograms. Several cladistic biogeographic analyses of austral biotas have been published, most of which reproduce expanded, cladistically resolved versions of Hooker’s austral biota.

Another new approach requiring detailed phylogenetic knowledge is based on the assumption of “molecular clocks.” If a set of homologous DNA sequences, sampled from a clade of biogeographic interest, has accumulated site mutations at a constant rate, then it can be used to calculate the relative age of speciation events. Such a clock can be “calibrated” by reference to the minimum age of one or more clades, as indicated by the fossil record. This method requires copious amounts of sequence data and few applications relevant to the minimum age of particular taxa derived from molecular clocks. Finally, biogeographers of all persuasions have sought to explain patterns of occurrence of austral taxa in the fossil record.

A. Distributional Patterns in Relation to Means of Dispersal, Climate, and Area

Darwinian biogeographers emphasized a number of patterns that they thought outweighed Hooker’s austral pattern and these were reviewed by Darlington (1965). He pointed to the low number of vertebrate taxa that belong to the austral biota: marsupials, chelid turtles, leptodactylid and hylid frogs, and galaxiid fishes, and of these, only the galaxiids occur in New Zealand. If these dispersed over land between Australia and South America, then why did no other vertebrate groups follow them? Why did hardly any mammals, reptiles, or amphibians walk to New Zealand? Darlington considered plants and insects, which comprise the great majority of austral taxa, to have more effective means of dispersal than most terrestrial vertebrates.

Darlington also argued that climatic patterns were highly significant in explaining austral distributions. The southern parts of South America, Tasmania, and
New Zealand are cold and their western sides very wet because of the strong prevailing westerly winds. The austral biota (as circumscribed by Hooker) is largely restricted to these areas. Moreover, the strong westerly winds blow directly over parts of Tasmania, New Zealand, and southern South America. The area of land in the Southern Hemisphere tapers off markedly to the south, greatly restricting the amount of land available to the austral biota. This suggested to Darlington that the “southern end of the world” would be a poor evolutionary center. Biological diversity also tapers off to the south. The extreme southern parts of Australia, New Zealand, and South America all have lower biological diversity per unit area than the less climatically extreme areas further north.

B. General-Area Cladograms

Do Gondwanic landmasses share a common ancestral biota to the exclusion of Laurasian landmasses? What are the cladistic interrelationships between different Gondwanic biotas? These are the kinds of questions that cladistic biogeographers are inclined to ask about the Southern Hemisphere. Among the first studies to focus primarily on these questions was that of Humphries (1981), which was based on cladograms for Nothofagus and 24 other taxa. Most of these cladograms were derived from precladistic taxonomic treatments and all relied entirely on morphological evidence. Nevertheless, he discerned two general cladistic patterns, summarized as:

1. ((((Australia, New Guinea, Tasmania, New Zealand, New Caledonia) South America) Africa) (North America, Europe))
2. ((Australia, New Guinea) (South America, North America, Europe))

He concluded that “the two positions for South America ... are probably due to the fact that it is a huge composite area and shouldn’t be treated as a single area of endemism” (Humphries, 1981, p. 205).

More recent attempts to resolve a general-area cladogram for Gondwanic landmasses have tended to subdivide South America, and sometimes other continents and islands such as Australia, New Zealand, and New Caledonia, into two or more areas for purposes of analysis. They have also been able to use more rigorously produced cladograms, some of which have incorporated molecular data.

For example, Crisci et al. (1991) analyzed 17 cladograms (8 insect, 8 angiosperm, 1 fungal) and found that southern South America, New Zealand, Tasmania, Australia, New Caledonia, and New Guinea consistently grouped together as an austral biota, to the exclusion of northern South America, North America, and Africa. Interrelationships within the austral group, however, remained ambiguous.

Muona (1991), on the other hand, intuitively examining cladograms for 84 genera of Eucnemid beetles, concluded that “fifty-seven genera shared a pattern coinciding with the traditional model of Laurasia–Gondwana breakup” (Muona, 1991, p. 165). He did add, however, that “twenty five groups showed an anomalous feature within the Gondwanan pattern having an Indomalesian clade as the sister group of the Australia–New Guinea clade” (Muona, 1991, p. 174).

A feature common to all of these general-area cladograms is the clustering of most, but usually not all, Gondwanic landmasses. Africa, Madagascar, and northern and eastern South America often group with Laurasian areas such as Europe and North America. Within the austral cluster, relationships tend to be unstable. These oddities probably reflect some of the limitations of cladistic biogeographic methods, which handle hierarchical vicariant patterns elegantly, but which require ad hoc adjustment to cope with reticulate patterns. The latter may be due to long-distance dispersal or secondary mixing of biotas caused by geological processes such as continental collision.

C. Molecular Clocks

The obvious limitation of the molecular clock is its reliance on an assumption that is known to be false: a constant rate of molecular evolution. Molecular rates differ within and between genes and among lineages but the idea of a biological dating technique analogous to radioactive decay is so attractive that many are prepared to tinker with the assumption or with their sampling in order to apply it to even a limited extent.

Waddell et al. (1999), for example, applied relative rate tests to mitochondrial DNA sequences for mammals and birds in order to select the largest subset of sequences consistent with the assumption of rate equality. Unfortunately, this procedure consigned their only marsupial sequence to the dustbin. However, they did manage to get two ratite sequences into their bird analysis and concluded that the Rhea and Ostrich are too similar for their most recent common ancestor to have walked between Africa and South America.

Bremer and Gustafsson (1997) took a different ap-
proach. They relaxed the assumption from rate homoge-
neity to minimal heterogeneity in their analysis of rbcL
sequences for the plant order Asterales. They concluded
that this clade is of mid to late Cretaceous age. This
implies, for instance, that the austral distribution pat-
tern of the family Carpodetaceae may be of Gondwanic
origin while those of the austral daisy genera are all
due to long-distance dispersal.

Manos (1997) was more cautious in interpreting rela-
tive levels of molecular divergence in Nothofagus, find-

Manos (1997) was more cautious in interpreting rela-
tive levels of molecular divergence in Nothofagus, find-
ing that rate heterogeneity was significant within and
between subgenera. He concluded (p. 1148) “a com-
bined gene approach using multiple substitution rates
and various calibration points taken from the fossil and
geologic record may be able to statistically support or
refute long-distance dispersal as a biologically realistic
event in the biogeographic history of Nothofagus.”

Molecular clocks look as though they will be useful
in rejecting vicariance as an explanation for very re-
cently dispersed, disjunct taxa but much less decisive
in detecting ancient dispersal events.

D. Fossil Record

The patterns of occurrence of taxa in the fossil record
provide evidence of the minimum ages of taxa and
often indicate that they previously had more extensive
distributional ranges than today. In most cases, absence
of fossils of a taxon from deposits of a particular age
or geographic location cannot reasonably be construed
as evidence that the taxon really did not exist at that
time or place. Absence of fossil evidence can also be
due to incompleteness of the fossil record. However, a
few taxa fossilize so readily and abundantly and are so
reliably identified as fossils that their absence from the
record has to be taken seriously as an indication of
either true absence or extreme rarity. The angiosperm
genus Nothofagus falls into this class (see below).

The fossil record has other limitations, the most
restrictive of which is the relatively low proportion of
taxonomic characters that are preserved. Vertebrates,
for example, are usually only represented by skeletal
remains. Their “soft parts” and DNA are rarely pre-
served. Similarly, many plant taxa are represented only
by their fossilized pollen grains. While these sometimes
display several distinctive synapomorphies diagnostic
of particular clades, most often they do not. Moreover,
pollen grains are such simple structures that the proba-
bility of convergent evolution of a distinctive pollen
type in the ancestor of an extant clade and in a different,
extinct taxon, while not being high, is not negligible.

Despite these limitations, the fossil record is a rich
source of historical information. Rather than inade-
quately reviewing the paleontological literature per-
taining to all Gondwanic taxa, we concentrate on exam-
ining all of key angiosperms.

Probably no taxon has been as strongly associated
with the concept of the austral biota as Nothofagus. In
1853, Hooker included two subgroups (as Fagaceae)
in his list of Antarctic plant groups, Darlington (1965)
came close to defining his “southern end of the world”
on the basis of its distribution, and van Steenis consid-
ered it a key genus for plant geography. Nothofagus has
an exceptionally detailed fossil record, with both pollen
and macrofossils relatively common. At least five dis-
tinct types of fossil Nothofagus pollen are recorded and
they include all the extant subgenera, Lophozonia, Fus-
cospora, Nothofagus, and Brassospora, respectively.

Nothofagus pollen is produced in abundance and fossil pollen is extremely common. If Nothofagus pollen
is not present in sediments then the genus probably
was not present in the immediate vicinity. Therefore,
the only limitation to determining the early bioge-
ographic history of the genus is the lack of sediments at
critical times and places. To a certain extent, dating
the earliest appearances of Nothofagus in early Campanian
sediments of southern Gondwana (Dettmann et al.,
1990) by pollen assemblages alone could involve some
circularity, since they are sometimes indicator species
for the pollen stratigraphy. However, in both southern
Australia and the Antarctic Peninsula, where the earliest
Nothofagus pollen has been found, the grains are pre-
served in marine sediments that are independently
dated using dinoflagellate and invertebrate fossil assem-
blages. Earliest records of Nothofagus pollen date from
the early Campanian, making Nothofagus one of the
oldest extant angiosperm genera in the fossil record.
Records of this pollen are from (in the order they ap-
pear) Australia, West Antarctica, New Zealand, and
South America (Dettmann et al., 1990). By the early
Maastrichtian, the Nothofagus clade had diversified into
all four modern subgenera, first appearing in West Ant-
arc tica, shortly thereafter in South America and subse-
quently in Australia some 12–13 million years later
(Dettmann et al., 1990). According to the pollen rec-
ords, the last of the modern subgenera to appear in
Australia was either the most basal subgenus Lopho-
zonia or the advanced Brassospora. New Caledonia
and New Guinea are devoid of fossils other than subgenus
Brassospora, the current earliest record being from the
Miocene in New Guinea (Dettmann et al., 1990).

In contrast to Nothofagus, the eucalypts (Myrtaceae:

SOUTHERN (AUSTRAL) ECOSYSTEMS

365
Eucalyptus alliance) have only been proposed as a Gondwanic clade relatively recently. Like many other plant taxa, the pollen grains of most eucalypts are not particularly distinctive and may be confused with those of other myrtleaceous groups such as Syncarpia and Metrosideros (Hill, 1994). They are pollinated by insects, birds, and mammals and produce fewer, less widely dispersed pollen grains than wind-pollinated Nothofagus.

Phylogenetic studies (e.g., Udovicic et al., 1995) have resulted in the recognition of seven genera of eucalypts: Eucalyptus, with more than 500 species, nearly all of which are endemic to Australia (several species occur in New Guinea, one of which extends west and north-west to Sulawesi and Mindanao, several occur in the lesser Sunda Islands as far west as Flores); Corymbia, with 113 species, restricted to Australia and New Guinea; Angophora, with about 20 species restricted to eastern Australia; Arillastrum with one species endemic in New Caledonia; Eucalyptopsis with one species endemic in New Guinea; Allosyncarpia, with one species restricted to Arnhem Land in northern Australia; an unnamed species most closely related to the previous two genera and surviving as a single population in the Queensland wet tropics. Unfortunately, fossil names for eucalypts have not yet been changed to match this new classification, and so all fossils mentioned here are referred to Eucalyptus.

In form the species range from low shrubs to the tallest flowering plants in the world. Eucalypts have been estimated to contribute 75% of Australian vegetation, and at the wetter margins of the continent, they dominate nearly all vegetation except rain forest and allied mesic types. Only in the arid Australian interior are eucalypts generally lacking in dominance.

Given their diversity and current dominance, the eucalypts would be expected to yield a complex and challenging fossil record. That this is not the case has been well documented (Hill, 1994). The oldest reliably dated and described Australian macrofossils associated with Eucalyptus include tree stumps, leaves and umbels about 20–22 million years old from several sites in the southeast of Australia (Hill, 1994).

The early pollen record of Eucalyptus is enigmatic mostly because of the difficulty in identifying eucalypt pollen. However, probable eucalypt pollen is now recorded from the late Paleocene of the Lake Eyre Basin in inland Australia (Hill, 1994).

An intriguing aspect of the fossil record of Eucalyptus is its apparent presence in both South America and New Zealand. A group of three fruits from probable Eocene sediments in Patagonia have been described as Eucalyptus patagonica, although there is some doubt as to whether they belong to the Eucalyptus alliance (Hill, 1994). Leaves from early Miocene sediments in the South Island of New Zealand are very similar to many Eucalyptus species and occur with eucalypt-like fructifications. This record is far more reliable at present than that from South America, especially when it is considered in conjunction with the pollen record, since Eucalyptus pollen in New Zealand occurs from the Miocene–early Pleistocene, suggesting a relatively recent extinction.

IV. PROCESSES

The complex patterns described here for the Southern Hemisphere biota suggest complex processes, and we still have much to learn about the way in which the Southern Hemisphere biota was molded into its living form. In the following section we briefly mention some of the more likely processes that must have had a significant impact.

A. Continental Drift

Although the precise detail of the breakup of Gondwana is still being refined, our broad understanding is robust and very important for the history of the biota. When continent-sized landmasses rift apart there must be profound direct and/or indirect consequences for the biota.

The most obvious direct consequence is the physical separation of once continuous populations of species, which must have often led to allopatric speciation as conditions on the resultant landmasses changed relative to one another. Indirect consequences include changes in climate as the various land masses move into and out of different climatic zones and the altered relative position of the continents can also influence climate, for example by altering ocean currents on a massive scale.

B. Long-Distance Dispersal

The Southern Hemisphere is dominated by oceans, which provide formidable barriers to dispersal. However, it seems likely that many organisms have successfully crossed these ocean barriers and formed successful populations at the other end. This is likely to have been in the direction of prevailing winds. While long-distance dispersal is a fact of life, it is very difficult to provide clear evidence for it and so it is largely recorded.
on a narrative basis. However, the methodological inadequacy of long-distance dispersal as a general explanation should not be taken to imply its unreality as a process.

Another example of long-distance dispersal that may have occurred predominantly overland involves north to south dispersal in more recent times. As the Gondwanic landmasses moved close to their present positions, the opportunity for relatively simple, predominantly land-based dispersal opened in three major regions: from Southeast Asia into Australia, from North America to South America, and from Europe to Africa. There are many compelling examples of such dispersals.

C. Climate

The climate history of the Southern Hemisphere is relatively well known and has clearly undergone major change during the rifting of Gondwana. This has been due to a combination of effects, including altered ocean currents, changing atmospheric CO2 levels and shifting continental positions. The most obvious place where this has had an impact is Antarctica, and this makes a useful case study to demonstrate the potential impact of climate change.

Antarctica has a long history as a continent close to the South Pole. Throughout the rifting of Gondwana, Antarctica has been one of the more stable landmasses, and although its history is complex, it can be considered as a long-standing high-latitude region. The idea that Antarctica was a source of plants and animals for surrounding high-latitude landmasses such as southern South America, southern Australia, and New Zealand is central to Southern Hemisphere biogeography. However, Antarctica today not only is geographically isolated from the other southern landmasses, but is also far too climatically hostile to support the biota concerned.

Antarctica now has only a very sparse covering of plants near the coastline, only two of which are vascular. Most of the continent is without plant cover, and where terrestrial plants occur, they are usually lichens and mosses. The Antarctic vertebrate fauna is sea-based, using the land only temporarily. In contrast, there is abundant evidence from Antarctica and adjoining landmasses to demonstrate the presence of complex biotas near the South Pole at times in the past. The almost total lack of vascular plants on Antarctica today testifies to major extinctions on this continent. This rates as one of the most complete regional extinction events in the Earth’s history, and it is important for us to understand it now more than ever, since it was indisputably a climate-based extinction.

There is now abundant fossil evidence for complex Cretaceous and Cenozoic forests on Antarctica that contain taxa that often still occur (or at least their descendants do) in high southern latitudes in South America, Australia, and New Zealand (Hill and Scriven, 1993). A combination of plate tectonics and other less well defined factors can be used to explain a very different climate at high latitudes in the past, and physiological research on the extant flora has demonstrated that diverse plant life was possible in Antarctica in the past without other physical changes.

During the Cretaceous at high southern latitudes massive speciation occurred, and there is no doubt that lineages were produced that today occupy niches from the equator down to the farthest southern latitudes at which the biota can survive.

The Cenozoic Antarctic vegetation is known from only limited data, but it appears that it continued to show basic similarities to Paleogene floras from Australia and South America. The separation of Australia from Antarctica began in the late Cretaceous, although opportunities for floristic interchange across restricted water gaps probably persisted well into the Cenozoic. Opening of the Drake Passage began in the late Oligocene, but details are lacking concerning the extent of dry land, probably as islands, between the Antarctic Peninsula and South America prior to that date.

The question of when the Cenozoic vegetation of Antarctica was eliminated by increasing cold and ice cover has not yet been answered by the fossil record. This is largely because of the confusion introduced into the spore and pollen record by the recycling of plant microfossils once processes of glacial erosion and sedimentation had begun. However, there is reasonable evidence for a cover of temperate forest in the Eocene, perhaps over much of Antarctica. Truswell (1991) noted that sea-surface temperatures about Antarctica were close to 0°C after the end of the Eocene, allowing the formation of sea ice; temperatures on land were probably lower, and it is unlikely that anything but a highly specialized, restricted flora would have persisted for long in such conditions. However, sparse Oligocene palynofloras containing relatively abundant Nothofagus (at least three subgenera) and a low diversity of other vascular plants have been described from Antarctica (Hill and Scriven, 1993). The data indicate a coastal Nothofagus forest that was diverse enough to suggest an extension of temperate forest in the region beyond the Eocene.

Milder sea-surface temperatures adjacent to the lower latitude regions of the northern Antarctic Penin-
southern latitude flora still exist today, and their ecolog-

cosystem nature of these angiosperms.

disturbance level in these environments and the early

early angiosperms, and this may be because of the high

Dettmann (1989) has suggested that rift valley systems at high

Gondwana by the close of the Cretaceous formed the

Woody plants no longer grow at very high southern

E. Plant Growth at High

Southern Latitudes

Woody plants no longer grow at very high southern

lata probably allowed the persistence there of some

Some prominent angiosperm taxa from the early high

SOUTHERN (AUSTRAL) ECOSYSTEMS

D. Disturbance

The angiosperms (flowering plants) play a critical role

in the current Southern Hemisphere biota, providing a

substantial component of the biodiversity and also act-

ing as an important part of the food chain and niche

requirements for many other organisms. Angiosperms

probably evolved in the early Cretaceous and radiated

worldwide. Early angiosperms probably had general-

ized pollination and dispersal strategies which made

them ideal for long-range dispersal by pioneering the

fresh sedimentary surfaces of coastal deltas, lagoons,

tidal flats. During the early Cretaceous, these pio-

neering early angiosperms or angiosperm ancestors dis-

persed along coastlines out of the rift valley of West

Gondwana (between Africa and South America).

Angiosperms were established generally in Gond-

wana by Barremian–Aptian times (Dettmann, 1989),

and those lineages that were widespread in southern

Gondwana by the close of the Cretaceous formed the

foundation for the living austral floras. Dettmann

(1989) has suggested that rift valley systems at high

southern latitudes acted as migrational pathways for

early angiosperms, and this may be because of the high

disturbance level in these environments and the early

successional nature of these angiosperms.

Some prominent angiosperm taxa from the early high

southern latitude flora still exist today, and their ecol-

ogy is well understood. A good example is Nothofagus,

which has been studied extensively. South American

Nothofagus does not usually regenerate continuously

at low altitudes, especially at lower latitudes, where

climatic conditions favor more complex forests and No-

thofagus seedlings cannot establish. Instead, Notho-

agus relies on catastrophic disturbance, a relatively com-

mon phenomenon in the highly unstable Andes

mountain chain and the Southern Alps of New Zealand,

to provide a fresh, cleared substrate which is readily

colonized by seedlings. This regeneration behavior is

probably very similar to that which occurred early in

Nothofagus history, when it either occupied dense for-

est on unstable sites and was unable to regenerate in

the absence of disturbance, or was in more sparse forest

on less equable sites where its seedlings could survive

in the understory. Hill (1994) hypothesized that the

large reduction in Nothofagus diversity at high latitudes

today, especially away from the Andes, may be due to

the relative modern stability of the landscape, which,

coupled with the density of the forest cover, does not

provide the right environment for the continued pres-

ence of many Nothofagus species.

Thus, at high southern latitudes during the Creta-

ceous–Paleogene, unstable habitats provided both a

pathway for the migration of early angiosperms and

also the potential for peripheral isolation of populations

of widespread species, which may have been a critical

factor for evolution in the region (Hill and Scriven,

1995). Therefore the evidence available suggests pro-

gressive extinction from an earlier vegetation in which

Nothofagus was very diverse. It is important to note

that the Sirius Formation fossils are well inland in

Antarctica and were deposited at some altitude (al-

though this is difficult to quantify). In that case it

can be speculated that there may have been other,

more diverse vegetation present at lower sites closer

to the coast as recently as the late Pliocene.

A more recent form of disturbance has been the

widespread impact of fire. In Australia, fire history as

a major impact on the biota can be traced back to

the mid Cenozoic at least, but its influence has been

particularly profound only in the past 100,000 years or

so. Fire is now very significant in areas of the Southern

Hemisphere with hot, dry summers.
ton of the fossil record in general that behavioral traits such as dormancy rarely leave a clear imprint (Truswell, 1991). Nevertheless, information on physiological adaptations to life at high latitudes can be obtained, especially from fossil wood. Cretaceous gymnospermous wood from paleolatitudes probably higher than 70° shows a consistent growth pattern, with increments, probably annual in origin, clearly delineated, and indicating a pronounced seasonal influence (Hill and Scriven, 1995). The amount of wood added annually was usually large, even when compared with modern low-latitude trees, but there is considerable variation in ring widths from year to year, suggesting that the trees were highly sensitive to environmental fluctuations (Truswell, 1991). A fossil forest at one early Cretaceous site had trees spaced 3–5 m apart and the tallest preserved trunk height is 7 m. Hill (1994) used these data to explain the reduction in wind pollination in understory plants during the very early Paleogene in southern Australia. With the sun at a low angle in the sky during summer and tracking an almost circular path around the horizon during the day, he concluded that it was likely that the forest structure was quite different to that observed in heavily forested regions at lower latitudes today. Relatively widely spaced, conical trees probably dominated and this may have continued until at least the late Cretaceous, and thus wind pollination was a viable strategy for understory plants below the open canopy. As today’s forest-bearing landmasses moved to lower latitudes and the sun angle increased, a closed forest structure became possible, due to the increase in incoming solar radiation. This may have been critical for the loss of wind pollination as a viable strategy in the associated understory plants.

The presence of a relatively dense plant cover, including large and rapidly growing trees, in regions south of the Antarctic circle, was for a long time viewed as difficult to explain, given that today growth at high latitudes is considered to be constrained by characteristic day-length patterns and by the temperatures associated with a regime of long rigorous winters and short cool summers (Truswell, 1991). However, it is now generally accepted that the light energy at high latitudes was sufficient, at present values of axial obliquity, to assume that the processes in the two hemispheres were at all similar.

Enhanced atmospheric CO2 levels may have been influential in producing the high annual increments of growth observed in Cretaceous trees. The overall effect of the high predicted levels of CO2 in the Cretaceous on plant growth are difficult to quantify, since experiments carried out on living plants examine the instantaneous effects of large changes in CO2 concentration on plants which have evolved to a particular ambient level. When plants have evolved to accommodate higher levels (as in the Cretaceous) the result may have been even more spectacular (Hill and Scriven, 1995).

V. CONCLUSION

The biota of the Southern Hemisphere is extremely diverse and is the result of a complex history. However, clear biotic linkages occur among the major landmasses. These linkages have historical significance, with vicariance and long-distance dispersal contributing to the geographic spread of organisms and climate change and disturbance being among the major processes that have led to the living species that occur there. Our knowledge of the history of the Southern Hemisphere is incomplete, but we have enough information from interrelationships among living organisms and the fossil record to be able to reconstruct at least part of this history with reasonable accuracy. The most important event in the history of the Southern Hemisphere as it influences the living biota was the rifting of Gondwana. Not only did this separate the biota on to smaller landmasses that subsequently had very different histories, but it also influenced other processes such as climate change and disturbance regime. This distinguishes the Southern Hemisphere biota very clearly from that of the Northern Hemisphere, and it is important not to presume that the processes in the two hemispheres were at all similar.

See Also the Following Articles

ANTARCTIC ECOSYSTEMS • AUSTRALIA, ECOSYSTEMS OF • SOUTH AMERICA, ECOSYSTEMS OF

Bibliography


SPECIATION is the process by which new species are formed. Understanding the origin of species and their contribution to the origin of biodiversity requires an understanding of the process of speciation and an appreciation of the problem of how to define a species.

I. SPECIATION AND SPECIES

A. Speciation

The ultimate source of the earth’s biodiversity is speciation, a process that occurs when gene flow is reduced sufficiently between sister populations to allow each to become irrevocably committed to different evolutionary lineages. Speciation results in the splitting of a lineage into two or more species (cladogenesis). However, when a species is transformed over time (anagenesis) by the acquisition of phenotypic and genetic modification, there is no increase in the number of species and thus no speciation event (Fig. 1a).

B. The Species Problem

Although the end product of speciation seems clear, defining when and under what conditions two popula-
Speciation, Process of

Figure 1. Major modes of speciation. See text for details on the processes involved in divergence and speciation.

T1
a. Anagenesis
b. Dichopatric
c. Peripatric

Time

Transformation

T2
R

T3
R'

A

Non-allopatric Speciation

T1
A

d. Parapatric
e. Sympatric
f. Hybridization

Time

Allopatric Speciation

 Generally not recognized as species. Recent molecular studies on closely related sympatric animal and plant taxa recognized as distinct, closely related sister species (the products of one speciation event) reveal that gene flow between them may still occur (Feder et al., 1995; Taylor et al., 1997). Such neospecies have acquired sufficient genetically based differences to maintain independently genotypes and distinct phenotypes but are still capable of sustaining some gene flow as they diverge. Several other species concepts have been proposed (Futuyma, 1998), but all suffer from problems of circularity, inconsistencies, and untestable qualities.

For this reason, the genetic-based species concept of Mallet (1995) provides a useful, less assumption-laden, operational species definition for identifying sister species. Mallet proposed that sister species be recognized only when they maintain distinctly different sym-
PATRIC GENOTYPIC CLUSTERS. DEFINING A SPECIES BASED ON GENOTYPIC CLUSTERS REPRESENTS THE LEVEL OF DIFFERENTIATION MOST INTERESTING TO THOSE STUDYING SPECIATION. GENERALLY, SYMPATRIC SISTER TAXA THAT MAINTAIN DISTINCT GENOTYPIC CLUSTERS EXPLOIT DIFFERENT HABITATS OR HOSTS. THE WIDE USE OF MODERN DNA ANALYSIS IN EVOLUTIONARY STUDIES MAKES THIS A USEFUL APPROACH PARTICULARLY WHEN COMBINED WITH SOUND BIOLOGICAL STUDIES. HOWEVER, AS WITH THE BSC, USING GENOTYPIC CLUSTERS TO DETERMINE THE BIOLOGICAL STATUS OF GEOFISOLATED SISTER POPULATIONS WITHOUT A "TEST OF SYMPATRY" IS STILL A SUBJECTIVE ALTHOUGH SOMEWHAT MORE QUANTITATIVE CALL.

II. REPRODUCTIVE BARRIERS TO GENE FLOW AND THE EVOLUTION OF MATE RECOGNITION SYSTEMS

A. Reproductive Barriers

The free exchange of genes between populations adapting to different niches inhibits the evolution of independent genetic systems. As soon as adaptive gene combinations are formed, they are broken up by recombination through interbreeding. There are several biologically based reproductive barriers or "isolating mechanisms" that evolve during the course of speciation which limit gene flow between sister species (Table I). These biological barriers constitute components of a mate recognition system that promotes assortative mating between individuals adapted or adapting to shared environmental and reproductive conditions.

III. TWO MODES OF SPECIATION

There are two primary modes of speciation recognized by most evolutionary biologists (Table II). Allopatric (geographic) speciation occurs when sister populations are isolated for a period by a physical barrier such as a mountain range or an expanse of uninhabitable terrain or water. In the absence of gene flow, such isolated populations inevitably accumulate unique mutations and, over time, may diverge genetically by genetic drift and in response to divergent selection pressures. If sufficient genetic divergence occurs during isolation to establish and maintain distinct genotypic clusters, particularly when their ranges later overlap, then the taxa are regarded as species.

Non-allopatric (ecological) speciation occurs in the absence of physical barriers to gene flow when sister populations diverge genetically and become ecologically and reproductively isolated as they adapt to different habitats. New sister species thus evolve within the dispersal range of the offspring from a single deme. Several different patterns of geographic and ecological speciation are recognized based on the factors involved in promoting their divergence, such as population:

<table>
<thead>
<tr>
<th>TABLE I Classification of Reproductive Isolating Mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prezygotic isolating mechanisms: mechanisms that prevent interspecific mating</td>
</tr>
<tr>
<td>Temporal isolation (potential mates have overlapping ranges but reproduction occurs at different times)</td>
</tr>
<tr>
<td>Habitat isolation (potential mates have overlapping ranges but reproduction occurs in different habitats)</td>
</tr>
<tr>
<td>Ethological isolation (potential mates meet but do not mate)</td>
</tr>
<tr>
<td>Mechanical isolation (potential mates attempt to copulate but no sperm is transferred)</td>
</tr>
<tr>
<td>Mechanism that prevents fertilization</td>
</tr>
<tr>
<td>Gametic mortality (sperm is transferred but egg is not fertilized)</td>
</tr>
<tr>
<td>Gametic incompatibility (gametes meet but fertilization is not completed)</td>
</tr>
<tr>
<td>Postzygotic isolating mechanisms: mechanisms that prevent the development of interspecific hybrids</td>
</tr>
<tr>
<td>Zygote mortality (eggs fertilized but zygote dies)</td>
</tr>
<tr>
<td>Hybrid inviability (zygote produces F1 hybrid of reduced viability)</td>
</tr>
<tr>
<td>Hybrid sterility (F1 hybrid is unable to partially or completely sterile or produce deficient F2 hybrid)</td>
</tr>
<tr>
<td>Coupling of interaction or cytoplasmic interactions (individuals from a population infected with an endoparasite or with a particular cytoplasmic element are fertile with each other, but fertility and/or viability break down when matings occur between infected and uninfected individuals)</td>
</tr>
</tbody>
</table>
TABLE II
Classification of Modes of Speciation

<table>
<thead>
<tr>
<th>Mode of Speciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allopatric speciation</td>
</tr>
<tr>
<td>Dichopatric speciation (vicariant division of wide-ranging species by an insurmountable barrier or extinction of intervening populations)</td>
</tr>
<tr>
<td>Peripatric speciation (by evolution in a very small isolated colony)</td>
</tr>
<tr>
<td>Nonallopatric speciation</td>
</tr>
<tr>
<td>Parapatric speciation</td>
</tr>
<tr>
<td>Sympatric speciation</td>
</tr>
<tr>
<td>Autopolyploid speciation</td>
</tr>
<tr>
<td>Spontaneous thelytokous parthenogenetic speciation</td>
</tr>
<tr>
<td>Hybrid speciation</td>
</tr>
<tr>
<td>Homoploid hybrid speciation</td>
</tr>
<tr>
<td>Introgressive speciation</td>
</tr>
<tr>
<td>Recombinational speciation</td>
</tr>
<tr>
<td>Polyploid speciation</td>
</tr>
<tr>
<td>Allopolyploid speciation</td>
</tr>
<tr>
<td>Direct and reticulate allopatric speciation</td>
</tr>
<tr>
<td>Symbiont-induced speciation</td>
</tr>
</tbody>
</table>

Structure, habitat specialization, hybridization, and polyploidy.

A. Geography, Geographic Races, and Allopatric Speciation

The process of geographic speciation (often referred to as allopatric speciation) may happen in two ways. Dichopatric or vicariant speciation (Fig. 1b) occurs when a widespread species becomes geographically divided into two or more large subpopulations by an insurmountable or impassable barrier. A second mode of geographic speciation, usually called peripatric or founder event speciation (Fig. 1c), occurs when a geographically isolated population is established by a single fertilized female or a very small number of founding individuals.

However, determining the actual biological status of sister taxa that have arisen in isolation following either a vicariant or a founder event is difficult and controversial. Species can be recognized when artificial hybridization of allopatric taxa in the laboratory reveals post-mating reproductive isolation. If no post-mating incompatibility is noted and mating occurs, then it is necessary to determine if there are factors important in pre-mating isolation not provided in the laboratory environment of the tests. Whether or not such populations are species or geographic races may require an unequivocal “test of sympathy” in which the two taxa coexist without fusing. Only when previously isolated taxa have the opportunity to interbreed under natural field conditions can the species status of closely related taxa lacking post-mating reproductive isolation be confirmed. If they maintain distinct sympatric genotypic clusters, they should be recognized as distinct species. Because such tests cannot often be performed, deciding on whether isolated taxa represent species or geographic races (i.e., subspecies) is a subjective decision. However, it should be stressed that when two taxa overlap naturally (i.e., are parapatric) and do not interbreed, this does not necessarily mean that they speciated sympatrically. This is because it is often impossible to establish if the overlap represents a case of secondary contact following allopatric speciation or if the taxa evolved as the result of non-allopatric divergence.

1. Dichopatric Speciation

When populations of geographically isolated sister taxa remain large during and following geographic subdivision, they will slowly diverge generically by genetic drift and as they adapt to local conditions in isolation. Over time, they may accumulate sufficient genetic differences to cause negative pleiotropic gene interaction and hybrid incompatibility among genes responsible for proper mate recognition and genome integration. Reproductive isolation therefore occurs by chance as the by-product of genetic divergence in isolation rather than the outcome of natural selection acting directly to promote reproductive isolation.

a. Dichopatric Race Formation and Speciation in Salamanders

The long-term study of evolution in the plethodontid salamander genus Ensatina by David Wake and colleagues (Wake and Schneider, 1998; Wake, 1997) provides an excellent example of dichopatric speciation and incipient species formation. The Ensatina eschscholtzii complex represents an ancient lineage that has undergone several instances of range contraction, isolation, and expansion and divergence of populations along broad ecological gradients. It is composed of seven congeneric subspecies wrapped in ring-like fashion around the Central San Joaquin Valley of California (Fig. 2). Historically, populations of Ensatina slowly differentiated into highly distinct forms from the northern subspecies E. e. argenesis in color and pattern as they expanded southward. Blotched forms occur in the Sierra Nevada and mountains of southern California, whereas unblotched forms are found in the coastal and northern region of their range. The animals, which are long-lived (10–15 years) and take 4 years to mature, are...
sedentary and disperse only short distances (home range 10–20 m).

Hybridization and intergradation occur between adjacent subspecies to varying degrees in all except one contact zone. The exception occurs in the Cuyamaca Mountains of San Diego County, in which *E. e. eschscholtzii* and *E. e. klauberi* overlap but rarely interbreed in sympathy and are regarded by some biologists as distinct species. In other areas, the complex includes many geographically and genetically distinct entities, some of which are near the species level of differentiation.

---

A limited narrow secondary hybrid zone also occurs in Calaveras County between the upland *E. e. platensis* and a coastal, lowland subspecies, *E. e. xanthoptica* that established an outpost in the Sierran foothills sometime in the past. *Ensatina e. platensis* lives in the cool and moist, closed canopy forest at higher elevation than those of *E. e. xanthoptica*, which prefers warmer and drier lower elevations. Hybridization between the pure subspecies is rare, with most hybrids representing progeny from backcrosses or from crosses between hybrid individuals. Hybrid individuals appear to be at a selective disadvantage and thus subject to strong negative selection. They have reduced success in competing for preferred habitats and mates and seem to be more prone to predation. Because little or no gene flow occurs across the narrow hybrid zone, the coastal and Sierra taxons behave as incipient species or taxa in the final stages of speciation. Four other hybrid zones in central California involving *E. e. xanthoptica* have been characterized in detail. The *Ensatina* complex presents a full array of differentiation from geographically variable populations or races to well-marked species. It provides an excellent example of how speciation has progressed slowly by adaptive divergence and stochastic processes during periods of geographic isolation.

2. Peripatric Speciation and the Founder Effect Principle

There are several examples in which species appear to have evolved in small, isolated populations at the periphery of the range of a sister species. Unlike the slow rate of evolution in dichopatric speciation, peripatric speciation is postulated to take a relatively short time. How and how often this divergence occurs is controversial (Barton, 1990; Hollocher, 1996).

Three models, each based on a founder event in which a population is established by only a few individuals, have been proposed to account for the rapid evolution of species observed on islands and elsewhere (Fig. 1c). The founder effect principle was developed by Mayr (1954). It is based on the assumption that reproductive isolation from the parent species can evolve rapidly in a population established by a very small number of founding individuals (i.e., 2–10). He postulated that in such populations a genetic revolution could take place as a by-product of inbreeding, selection, drift, and genome reorganization. While the population is small, these genetic changes may promote substantial morphological and ecological shifts.

A modification of the founder effect principle was proposed by Carson (1975). In his founder-flush model of speciation, isolated populations undergo a series of population expansions and drastic contractions to a very small number of individuals. Carson believes that founder-flush speciation can occur only in certain cross-fertilizing diploid organisms with “open” genetic systems. The genome in such organisms represents a clique of harmoniously collaborating or coadapted genes united by strong epistatic interactions. Their genomes also have abundant pleiotropic interacting genetic polymorphisms and share a high recombination index. Carson hypothesized that these attributes provide great genetic flexibility that predisposes such organisms to speciate by the founder-flush process.

In Carson’s view, the drastic events required to reorganize the original genetic system and restore new balances that are incompatible with ancestors are accomplished during cycles of the founder-flush process. Selection, which is relaxed during the flush phase of population expansion, is greatly intensified as the population crashes. Repeated disorganization and reconstitution of the genome results in the rapid evolution of reproductive isolation.

Templeton (1980) proposed a third modification of founder-induced speciation that involves a genetic transilience (Carson and Templeton, 1984). It is similar to Carson’s founder-flush speciation but requires changes in only one or a few segregating units, commonly with epistatic modifiers responsible for reproductive isolation that occurs when a population rapidly passes through an extremely unstable intermediate genetic state. As in the case of the founder-flush model, a genetic transilience involves strong inbreeding and large variance in population size. The critical trigger that initiates a transilience requires a reweighting of fitness components owing to drift-induced shifts in allele frequencies at one or more major loci that have pleiotropic effects on reproductive isolation.

It seems clear that speciation may occur rapidly following the geographic isolation of a small population. It is not clear, however, whether speciation results from a founder-flush or transilience process or solely from natural selection in response to factors such as runaway sexual selection or rapid adaptation to divergent ecological and reproductive conditions. Because conditions required for stochastic transitions are severe and well-documented cases are lacking in which drastic genetic changes caused by founder effects result in speciation, it appears that the process of speciation in small, peripheral populations is the same as that which occurs in dichopatric speciation. Although certain kinds of epistasis can promote strong reproductive isolation, divergence occurs by selection and not random genetic drift (Barton, 1996).
Reproductive isolation between sister populations in cases of allopatric speciation arises as a by-product of genetic changes that originate and accumulate independently in each population during periods of geographic isolation. Speciation may or may not be accompanied by an ecological transformation. In contrast, nonallopatric or ecological speciation inevitably involves a shift to a new niche at the time of speciation, and adaptation to different habitats is the driving force initiating and sustaining genetic divergence (Figs. 1d–1f). Divergent selection on existing genetic variation results in the evolution of reproductive isolation as different suites of genetic variants are favored in each habitat. The emergence of reproductive isolation is also facilitated when mating occurs within a preferred habitat or if mate preference that is based on a phenotypic variant of a sexually selected trait becomes associated with a preferred habitat (Johnson et al., 1996).

A misconception concerning nonallopatric speciation is that it requires a process called reinforcement, which involves the divergence of the mate recognition system between populations during the speciation process as an outcome of selection against hybrids (Burris, 1987). Because theoretical models suggest that reinforcement will not allow reproductive reinforcement to develop, critics have argued that parapatric and sympatric speciation are unlikely to occur in nature. These models, however, fail to factor in the role of habitat preference and specialization in speciation and the effects of ecological divergence, which always accompanies sympatric speciation. When habitat choice is taken into consideration, conditions for sympatric speciation are greatly relaxed and its occurrence is probable (Johnson et al., 1996).

Three modes of nonallopatric speciation are generally recognized in sexually reproducing organisms. Parapatric speciation (Fig. 1d) takes place when sister species evolve while adapting to contiguous habitats or ecotones across a narrow contact zone (Bush, 1994). Sympatric speciation (Fig. 1e) occurs in the absence of geographic segregation when sister species evolve within the dispersal range of the offspring of a single deme. During the course of sympatric speciation, the probability of mating between two individuals depends on their genotypes, and divergence occurs between populations adapting to alternate habitats within the “cruising range” of each other. Parapatric and sympatric speciation actually represent extremes of a continuum in the pattern and extent of habitat and geographic-imposed spatial segregation and gene flow reduction that occurs during nonallopatric divergence. Hybridization (Fig. 1f) is a third mode of nonallopatric speciation. New species rapidly evolve from a mating between individuals of two closely related species by a variety of different processes. It is a frequent mode of speciation in plants but appears to be rare in animals.

1. Parapatric Speciation

This mode of speciation occurs when sister species evolve while adapting to contiguous habitats (or ecotones) along a zone of contact. Examples from nature involve situations in which individuals from a species adapted to one habitat invade and colonize an adjacent habitat. Individuals bearing novel genetic recombinants capable of exploiting and reproducing in the new habitat are the first to invade and colonize the narrow zone where the two habitats meet. As adaptation proceeds, the new colony expands throughout the range of the new habitat. Only along the adjoining borders between the original and new habitat is gene exchange possible between the two populations. As the new population adapts to conditions imposed by the new habitat, divergent selection promotes the evolution of reproductive isolation and eventually parapatric species.

a. Parapatric Speciation in Mimulus

An example of rapid parapatric adaptation and speciation of a plant to a newly established vacant niche is that of the monkey flower Mimulus cupriphilis in California (Macnair and Gardner, 1998). This species grows only on relatively dry and toxic mine tailings of two small copper mines in California. It grows in close proximity and is recently derived from the widespread hydrophylic M. guttatus.

The two species differ in many ways. Mimulus cupriphilis, an obligate annual, flowers earlier and produces many small flowers that differ in shape and color from those of M. guttatus. Mimulus cupriphilis also has higher fitness on the dry tailings of the copper mines and flowers early when pollinators are rare. Because it is self-fertilizing, it is reproductively isolated from the outcrosser M. guttatus, which blooms later when its larger flowers are fertilized by bumblebees. Genetic studies have revealed that the species differ in a few major genes controlling flowering time, flower size, corolla spot number, and general size. Because all the genetic systems are recessive with the exception of flowering time, recessive alleles would be spread by natural selection in the original outbreeding parent spe-
cies. In inbreeders, there is no difference between domi-
nant and recessive alleles.

Mimulus guttatus has evolved recently since the cup-
er copper mines are less than 150 years old. The shift
to the new soils created by the mines involved the
development of a copper-tolerant ecotype with a pri-
mary semidominant adaptive mutation that shifted
the flowering time earlier in the plants growing in dry habi-
tats. In the absence of pollinators early in the season,
selection favored alleles for self-fertilization. This was
accomplished by flower size reduction that brought the
stigma and anthers in close proximity. Reduction in
size of corolla and associated structures may have also
freed up resources for seed production. The result is a
new selling species that has evolved locally in a very
short time and that is well adapted to the unique condi-
tions of the mine tailing fields.

There is no evidence of a genetic revolution or reduc-
tion in genetic diversity in the new species and F2 prog-
eny. Nor is there evidence of significant breakdown in
so-called gene complexes or major epistatic interac-
tions. Therefore, it is concluded that speciation has
been achieved by new selection pressures on normal
M. guttatus as it colonized a new and unusual habitat.
Major genes of large effect that play important roles in
reproductive isolation have also been demonstrated in
other sister species of Mimulus species (Bradshaw et al.,
1993). In such cases, new species have the potential to
appear nearly full-blown in relatively few evolutionary
steps because of changes in a few essential genes.

2. Sympatric Speciation
It is now apparent that sympatric speciation, once
thought to be rare in animals and plants, occurs more
frequently than previously realized. Because two closely
related sympatric sister species cannot share the same
resources, nonallopatric speciation is inevitably accom-
panied by the shift and exploitation of a new habitat
by the daughter species. Such habitat or resource shifts
reduce competition between sister species. It is for this
reason that sympatric speciation is often referred to as
ecological speciation.

a. Sympatric Speciation in Cichlid Fish
The important role of ecology in sympatric speciation
is exemplified by the study of Schleven et al. (1994).
In a molecular phylogenetic study they discovered that
two endemic cichlid fish species flocks (9 and 11 spe-
cies) in two small, ecologically monotonous volcanic
lakes (4.15 and 0.6 km²) in Cameroon speciated sym-
patrically. Because each crater is isolated from rivers
and lacks internal structure, past lake levels are not
responsible for physically isolating sister populations
during speciation. Field observations, stomach content
analyses, and the presence of specialized morphological
features related to feeding confirm that the species
within each lake have different benthic or pelagic tro-
phic and reproductive ecologies. Although all species
nest on the bottom and are capable of encountering
one another at moderately high frequencies along eco-
tones, no hybrids are found because the species mate
assortatively and do not interbreed.

Phylogenetic trees were constructed based on an
analyses of a 340-base pair fragment of mtDNA cyto-
chrome b and an additional 350 base pairs from the
rapidly evolving mitochondrial control region for all
20 species and all tilapine species in neighboring river
systems and lakes. This analysis confirmed that the
lake species are monophyletic with respect to the river
species, i.e., each flock evolved within each lake after
a single colonization event.

b. Sympatric Host Race Formation and
Speciation in Rhagoletis Fruit Flies
The importance of habitat and resource shifts in the
different species of Mimulus, colonized apples in approximately
flowering time of the adult in summer, and host recognition
and acceptance. In addition, strong frequency differ-
ences in alleles at several loci coding for proteins are
maintained by strong selection and greatly reduced gene
flow between the races. The apple and hawthorn races
now behave as semispecies or species. All seven mem-
ber of the R. pomonella species group are sympatric in
eastern North America and speciation in this species
group has always been accompanied by a shift to a new
host plant family or genus.
C. Other Modes of Nonallopatric Speciation

1. Spontaneous Thelytokous Speciation
This mode of sympatric speciation occurs when a unisexually reproducing taxa arises spontaneously from an unfertilized egg of a diploid bisexual species. Subsequent reproduction in taxa originating in this way produces only females from unfertilized eggs. Several good examples are provided by White (1978).

2. Autopolyploid Speciation
Occasionally, polyploidization of a diploid species may occur spontaneously in one or more individuals. Because autopolyploid individuals have three or more chromosome sets, each chromosome has more than one homologous pairing partner. During meiosis, multivalents are produced leading to unbalanced gametes and zygotes, sterility, and other problems. Only rarely does autopolyploidy result in the origin of new species, such as in the common potato (Solanum tuberosum) and its relatives (Grant, 1981). These usually originate from crosses between races whose chromosomes differ only slightly.

3. Speciation by Interspecific Hybridization
A new species can arise two ways by interspecific hybridization (Fig. 1f). Homoploid hybrid speciation results in a diploid-derived species, whereas polyploid hybrid speciation produces a species that combines a complete set of chromosomes from each hybridizing parental species. Hybrid species occupy habitats different from those of the parental species, thus reducing competition and the level of gene flow between them. Hybrid speciation, which is far more common in plants than in animals, can occur in at least four ways.

a. Introgressive Hybrid Speciation
Individual gene exchange among closely related species provides recombinant offspring that shift to and exploit a new habitat not utilized by either parental species. The hybrid species may be interfertile with one or both parental species, but it is reproductively isolated from them by premating barriers to gene flow. This mode of speciation has been reported in several plant species (Grant, 1981), but confirmation of this mode of speciation remains controversial and requires definitive experimental and analytical studies.

b. Recombinational Hybrid Speciation
A far more common mode of hybrid speciation involves the formation and establishment in the progeny of a chromosomally sterile or semisterile species hybrid of a new, structurally homozygous recombination type. Individuals are fertile within the line but isolated from other lines and from the parental species by a chromosomal sterility barrier. It is most likely to occur when the hybrid interface is long and the organisms involved are predominantly selling, relatively fertile, and possess few structural chromosome differences between the parental species.

i. Hybrid Speciation in Wild Sunflowers
A molecular study of hybrid speciation in the wild sunflowers Helianthus by Rieseberg et al. (1995) revealed that F1 hybrids of H. annus and H. petiolaris are semisterile with pollen viabilities less than 10% and seed set less than 1%. F1 pollen viability is highly variable, ranging from 13 to 97%. The two species are distinguished by several morphological and chromosomal features, and based on chloroplast DNA and nuclear ribosomal DNA variation they occur in divergent clades. Although the species are sympatric throughout much of the western United States, they have different ecological requirements. Helianthus annus is restricted to heavy, clay soils, whereas H. petiolaris predominantly inhabits dry, sandy soils. Helianthus anomalus is a rare endemic to xeric habitats in northern Arizona and southern Utah. It is well within range of parental species and is a recombinational hybrid resulting from a cross between H. annus and H. petiolaris. The F1 hybrids with parental species are partially sterile because chromosomal structural differences enhance reproductive isolation. A preliminary survey of 126 loci in natural populations of the parental species indicated that H. anomalus has loci derived from both H. annus and H. petiolaris. Some blocks of markers, possibly protected from recombination, are transmitted intact. Helianthus anomalus combines rDNA repeat units and allozymes of H. annus and H. petiolaris as predicted for diploid hybrid species, although individuals possess chlDNA haplotypes of H. annus and H. petiolaris rather than a unique haplotype. Genetic linkage maps generated for all three species using random amplified polymorphic DNA markers reveal loci distributed onto 17 linkage groups corresponding to the haploid chromosome number of the three species. Although levels of polymorphisms vary from 212 in H. annus to 600 in H. petiolaris, map density is similar among species. By comparing genomic location and linear order of homologous markers, chromosomal structural relationships were inferred among the three species. Even though 6 linkage groups showed no changes...
in all three species, the remaining 11 linkages were not conserved in gene order. The parental species differ from *H. anomalus* by at least 10 separate structural rearrangements, 3 inversions and a minimum of 7 inter-chromosomal translocations. The genome of *H. anomalus* is thus extensively rearranged relative to its parents. All 7 novel rearrangements in *H. anomalus* involve linkage groups that are structurally divergent in parental species, suggesting that structural differences may induce additional chromosomal rearrangements upon recombination.

c. Allopolyploid (Amphiploid) Speciation

Interspecific hybridization can also result in combining two or more complete chromosome sets. F₁ hybrids produced between two established related species are often sterile because chromosomes lack sufficient homology to pair well at meiosis. Fertility is restored if hybrids persist long enough by asexual reproduction until somatic doubling of the chromosomes can occur in a flower, or until there is a rare union between two unreduced gametes. A new sexually reproducing species is then established that is "instantaneously" isolated from both parental species (Grant, 1981).

i. Allopolyploid Speciation in Spartina anglica

The recent natural rapid evolution of the amphiploid perennial salt marsh grass, *Spartina anglica*, provides an example of allopolyploid speciation (Raybould et al., 1991). This species originated on the south coast of England at the end of the nineteenth century. It arose as a result of chromosome doubling in *S. x townsendii*, a hybrid between the native British *S. maritima* and the North American *S. alterniflora*, introduced by shipping (Fig. 3). *Spartina anglica* is now widespread along the English coast and is highly successful.

Although more than half of all plant species are directly or indirectly the by-products of allopolyploid speciation, allopolyploid speciation is relatively rare in animals (White, 1978).

d. Direct and Reticulate Allogenous Speciation

There are two modes of allogenous speciation (i.e., combining the genomes of two distinct species). In the case of direct allogenous speciation, hybridization between two bisexual, closely related species combines the genomes of two distinct parental species giving rise to a new, unisexual species (Bullini, 1994). The hybridization event produces either an allotriploid or an allotetraploid unisexual species that acquires clonal (parthenogenesis) or hemiclonal (hybridogenesis; i.e., it must mate with males of a bisexual parental species) modes of reproduction. In most such clonal and hemiclonal organisms, the heterozygous genetic structure of the parental species is retained.

In the case of reticulate allogenous speciation, individuals of unisexual hybrid taxon hybridize with a bisexual relative giving rise to new, unisexual species, often with a higher ploidy level than that of the parental species (Bullini, 1994). Most have highly heterozygous genomes that display heterosis. Clones produced by direct and indirect allogenous speciation exhibit heterosis and demographic advantage over both parental species and individuals.

IV. SUMMARY

After 140 years of speciation research, it is now clear that animal and plant species can originate in a variety of ways. Rapid autopolyploid speciation is common in
plants and accounts for many plant species. In animals, this mode of speciation is relatively rare. Dichopatric speciation gives rise to many, possibly the majority, of the species in some animal groups such as the land vertebrates. In fish, both allopatric and nonallopatric speciation have been reported. The majority of living organisms, however, are insects. Some authorities estimate there are as many as 30–40 million species, and all agree that there are more than 10 million. Approximately 70–75% are highly specialized parasites that feed in or on plant and animal tissue. Although many of these insects probably originated by nonallopatric speciation, it is not clear what percentage have done so. The same is true for the many mites and nematode species that also may number in the millions. Because the mode of speciation in only an extremely small number of animal and plant species has been established, it is impossible to estimate how often each mode of speciation occurs in any particular group of organisms. Such estimates must await research on a great many more taxa.

See Also the Following Articles

CLADOGENESIS • SPECIATION, THEORIES OF • SPECIES, CONCEPTS OF • SPECIES DIVERSITY, OVERVIEW

Bibliography


SPECIATION, THEORIES OF

Hope Hollocher
Princeton University

I. Theoretical Considerations
II. Defining Species
III. Evolutionary Forces Involved in Speciation
IV. Common Modes of Speciation
V. Genetic Patterns of Species Differentiation
VI. New Directions for Studying Speciation
VII. Conclusions

GLOSSARY

assortative mating The tendency of like to mate with like whether it is based on similarities in genotypes or similarities in phenotypes.
coadapted gene complexes Genes that have been selected to work in a coordinated fashion to confer high fitness.
epistasis An interaction between different alleles at two or more loci such that the phenotype differs from what would be expected if each locus acted independently.
evolution Change in allele frequencies over time.
fitness The average number of offspring produced by individuals with a certain genotype relative to the number produced by individuals with other genotypes as a result of differences in survival and reproductive success.
gametogenesis The developmental process for the formation of functional male and female haploid reproductive cells that combine at fertilization to produce the zygote.
gene pool The collective set of genes in a population at a particular time.
genetic architecture Characterization of the number and types of genes and their interactions that underlie a particular trait.
genetic recombination The physical exchange of genetic material between a pair of chromosomes during meiosis.
genotype The state of an individual with respect to a specific genetic locus or set of loci.
heterogamety Having two different sex chromosomes.
hybrid inviability The phenomenon whereby one or both sexes of progeny produced in crosses between two different species are unable to survive.
hybrid sterility The phenomenon whereby one or both sexes of progeny produced in crosses between two different species are able to survive, yet are unable to reproduce.
monogamy One male mates with one female exclusively.
morphometrics A statistical procedure by which morphological traits are quantified and analyzed.
phenotype The state of an individual with respect to a specific trait.
pleiotropy One gene affects more than one trait.
polygamy Males and females both have several different mates.
reproductive isolation Intrinsic barriers to the production of offspring.
SPECIATION IS THE PROCESS by which new species are formed. Thus, its central importance to the study of evolution and issues of biodiversity cannot be overstated. The seemingly simple definition of speciation as the origin of species belies the fact that the actual process of speciation is almost as complex and varied as the diversity of life itself. No one mechanism of speciation is sufficient to describe the origination of all the diversity we see in the world and more often than not several speciation mechanisms come into play simultaneously in the formation of a single new species. The study of speciation attempts to analyze these competing mechanisms to determine the relative importance of different evolutionary forces promoting species divergence under particular circumstances. By looking for recurring patterns across a wide spectrum of individual case studies, we are devising models that help explain speciation more generally. This knowledge then serves as a framework for making policy decisions regarding the preservation of biodiversity, for the ultimate goal of conservation biology is not just to preserve current species, but also to preserve the natural processes that help generate new ones.

I. THEORETICAL CONSIDERATIONS

Although speciation has been studied intensively for more than 140 years, surprisingly very little is understood about the patterns of genetic change that occur during the process. This lack of knowledge is a true stumbling block for describing speciation using a population genetic framework. A key aspect of all population genetic models of speciation is the assumptions made about the underlying genetic architecture—the number and types of genes and their interactions—of traits that have changed during speciation. The underlying genetic architecture influences directly how traits change during speciation even when the same evolutionary forces are known to be in operation (reviewed in Hollocher, 1998). Therefore, the examination of the genetic basis of species differences reveals information about the specific evolutionary processes that were involved in creating them.

Equally important to theories of speciation is determining which traits are most influential for promoting species divergence. We recognize species because they are different from one another, be it morphologically, behaviorally, or reproductively. In all cases, a sequence of events has led to the eventual fixation of these observed differences. Knowing whether certain traits are more prone to change than others during speciation or whether certain traits have a larger impact on future patterns of divergence directly informs our models of speciation. This knowledge improves our ability to target groups of organisms at the nascent stage of divergence and to focus our attention on specific traits for future studies of speciation that will bring a better understanding of the process.

II. DEFINING SPECIES

A. Why Is There More Than One Definition of Species?

One would think that by now evolutionary biologists would agree on what constitutes a species. However, it turns out that species concepts are almost as varied as the types of researchers that work on the question. To date there is no single, universally accepted definition of a species, although certain frameworks for thinking about species have come in and out of favor over the years (see "Species, Concepts of," by Mallet in this volume; also see Ereshefsky, 1992; Ridley, 1996; Howard and Berlocher, 1998). Most definitions of species serve a useful research purpose and are a function of the types of questions that are being addressed (e.g., whether species designations should serve a strictly taxonomic purpose as in systematics or represent the active evolutionary unit of diversification as in population genetics). This is not to imply that the definition of species is purely a semantic issue, but rather that the diversity of definitions reflects very different perspectives as to what factors should be emphasized in describing species. Beyond these purely theoretical issues, species concepts have an additional problem of not always being easy to apply in real-life situations.

As a result of these conflicting purposes of species concepts, every subdiscipline within speciation research has formulated its own particular species concept. However, the problems surrounding species definitions are more fundamental than simple differences in perspective. There is an inherent problem of trying to categorize groups of organisms that are undergoing a continual process of change. Speciation is a process, not an instantaneous event. It is also not caused by a single unified mechanism, but by a collection of different mechanisms of divergence affecting different traits at potentially different rates. The formation of natural groups does eventually result from the ongoing action of each of these processes of divergence and at some point in time these natural groups can and should be delineated as separate species. However, because of the
fluidity of the process, the boundaries that circumscribe these natural groups are often at best blurry given the fact that each underlying mechanism of divergence can potentially delineate a separate natural group. In some instances, this creates no problem and these different boundaries will actually coincide with each other. In these cases, several different species concepts will correspondingly give the same species designations. In other instances, these boundaries may not correspond well at all. Hence, depending on which process is being emphasized, the actual definition of a particular species may also vary in these circumstances. In most cases, conflicting species boundaries are not so much a problem of the competing species definitions as they are a reflection of the complicated nature of the underlying biological process that governs speciation. For a population geneticist interested in the process of speciation, it is those cases where the boundaries of divergence based on different criteria do not correspond well that often prove to be most interesting to study.

B. Most Commonly Used Species Concepts

Because of space constraints, a thorough critique of competing species concepts is beyond the scope of this article (for a more complete review, see “Species, Concepts of,” by Mallet in this volume; also see Ereshefsky, 1992; Ridley, 1996; and the chapters by De Queiroz, Harrison, Shaw, and Templeton in Howard and Berlocher, 1998). However, a general overview of the most commonly used species concepts will be presented here, paying particular attention to how each concept relates to the elucidation of underlying speciation processes.

1. Morphological and Phenetic Species Concepts

Intuitively, using morphological criteria to get a first approximation of species boundaries works rather well. Different groups of organisms look different from each other. This simple “morphological species concept” was used by the early naturalists of the eighteenth and nineteenth centuries and is still often used today by field biologists and paleontologists. This concept is also used by most evolutionary biologists as a starting point from which other criteria are then later added (the fact that species appear morphologically distinct is usually taken for granted by most practicing evolutionary biologists).

The “phenetic species concept” embellishes on this first approximation by incorporating specific morphometric analyses to describe explicitly the phenotypic groupings. From a functional point of view, groups of organisms that are morphologically distinct most likely represent different groups genetically as well.

Using morphological criteria alone, however, carries with it the risk of categorizing groups of organisms as the same species when indeed there are reproductive barriers between them (as is the case for cryptic species, which are morphologically identical but nevertheless represent distinctly different gene pools) or categorizing groups of organisms as separate species when in fact the difference is not genetically based, but solely a phenotypic response to varying ecological conditions (such is the case for some plants that can change the morphology of their leaves in response to varying light levels). In addition, the categorization of species based solely on morphological criteria without the incorporation of an evolutionary framework is generally unsatisfactory to the majority of people working in the field of speciation.

2. The Genotypic Species Cluster Definition

An improvement over most morphological approaches to defining species has recently been derived by Mallet (1993) using genetic criteria to cluster groups of organisms rather than strictly morphological traits (termed the “genotypic species cluster” definition). In this case species are defined as genetically distinguishable groups of individuals that exhibit few or no intermediate genotypes (both for single and multiple loci) when in contact. This is clearly an advantage over strict morphological criteria because it gets right to the heart of speciation—genetic differentiation. It is also very satisfying as an operational definition of species groups, because genetic clustering is easily diagnosable for all taxa, whether sexual or asexual. However, this definition is only easily applicable to the situation where species are in direct contact with each other and cannot be generally applied. More importantly, it is still a phenetic approach to species definitions and lacks any inherent evolutionary framework for the interpretation of species groups from a mechanistic point of view.

3. The Biological and Recognition Species Concepts

Until recently, the most enduring species concept has been the “biological species concept,” first outlined in detail by Dobzhansky (1937) and subsequently popularized by Mayr (reviewed in Mayr, 1982). The concept defines species as interbreeding groups of organisms that are reproductively isolated from other groups. In other words, species are groups of organisms that share
a common gene pool. Speciation, under this concept, is the development of a biological barrier to gene flow and the subsequent differentiation of the two resulting gene pools. Biological barriers to gene flow were termed isolating mechanisms and covered the full gamut from prezygotic barriers (including ecological or habitat isolation, seasonal or temporal isolation, sexual or ethological isolation), mechanical or physiological barriers to mating, isolation by different pollinators, and gametic isolation (blocking fertilization) to postzygotic barriers (including hybrid sterility, hybrid inviability, and other forms of reduced fitness in progeny resulting from crosses between differentiated populations). The biological species concept is firmly rooted in speciation processes and encompasses a number of different mechanisms that can serve to reduce the exchange of genes between different populations and thus allow them to evolve independently.

The "recognition species concept" is very similar to the biological species concept in that species are defined by interactions among its members (Paterson, 1985). In addition, both concepts try to pinpoint traits that are biologically relevant in species divergence. In the case of the recognition concept, species are defined as groups of organisms that share a common system of fertilization or specific mate recognition system; i.e., the emphasis is placed on what holds species together as a cohesive group rather than on what may separate them. The same general concept is embedded in the biological species concept, but the emphasis on mate recognition systems versus isolating barriers does influence how questions about speciation mechanisms are formulated and places more importance on evolutionary forces that positively influence the specific mating interactions between members of a species rather than on the more negative effects associated with preventing matings from occurring between members of different species. In either case, both of these concepts are mechanistically based and represent a prospective view of species distinctions by giving some indication of how isolated groups are now and will continue to be in the future.

4. Phylogenetic, Evolutionary, and Genealogical Species Concepts

Other more recently derived species concepts rely more heavily on a retrospective view by defining a species in a strictly historical sense as a separate evolutionary lineage that is internally connected through time (i.e., the "phylogenetic species concept" (Cracraft article in Otte and Endler, 1989; Nixon and Wheeler, 1990), the "evolutionary species concept" (Wiley, 1978), and the "genealogical species concept" (Baum and Shaw, 1993); for a good overview, see Harrison article in Howard and Berlocher, 1998). These species concepts are less oriented toward process and identifying specific biological traits that maintain cohesion within species or promote divergence between species and instead are more oriented toward the final evolutionary result—lineage divergence. That is not to say that process cannot be usefully inferred from looking at historical patterns of lineage sorting and splitting (see The Cohesion Species Concept below and Section VI.A), but the emphasis in these particular species concepts is clearly on pattern rather than on trying to incorporate a variety of biological processes into the species definitions themselves.

5. The Cohesion Species Concept

Perhaps the most inclusive species concept that has been proposed to date is the "cohesion species concept" (see Templeton in Otte and Endler, 1989; see also Templeton in Howard and Berlocher, 1998). This single species concept combines key elements of the biological, recognition, phylogenetic, evolutionary, and genealogical species concepts in order to form a comprehensive framework for defining species and studying speciation mechanisms. Like the recognition species concept, the emphasis in this case is on defining species based on mechanisms that provide cohesion among members of a population. Unlike the recognition concept, cohesion involves not just shared fertilization systems promoting gene flow (termed genetic exchangeability in this model) but also shared historical fates (i.e., belonging to a common evolutionary lineage) as well as shared evolutionary tendencies (i.e., experiencing common selective regimes), both of which will maintain species cohesion (through demographic exchangeability) as well as shared evolutionary tendencies (i.e., experiencing common selective regimes), both of which will maintain species cohesion (through demographic exchangeability) in the absence of gene flow. With this definition, the repertoire of traits important to speciation is broadened beyond traits determined primarily by their effects on reproduction and includes any traits involved in ecological adaptation, regardless of whether or not they directly affect the probability of genetic exchange. Although some would argue this more inclusive, pluralist species concept obfuscates the ability to clearly define species (if anything goes, then what should be the object of study?), in reality, the concept serves as a better starting point for speciation studies precisely because it is not narrowly focused on any one particular speciation mechanism to the exclusion of all others. The functionality of the cohesion species concept has been improved of late by incorporation of historical patterns of genetic variation to use as an explicit framework for both testing hypotheses of species boundaries and investigating spe-
speciation processes by modeling how different evolutionary forces pattern this genetic variation under different geographical scenarios (Templeton in Howard and Berlocher, 1998; see also Section VI.A).

III. EVOLUTIONARY FORCES INVOLVED IN SPECIATION

Because genetic change lies at the heart of speciation, the formation of new species involves the action of the same evolutionary forces known to cause genetic changes within populations (see Charlesworth in this volume for an in-depth description of population genetics). The principal evolutionary forces most commonly involved in species divergence are natural selection, genetic drift, sexual selection, and mutation. All of these forces act to change allele frequencies within populations over time and there is no question that they operate during speciation as well. The challenge of speciation studies comes from understanding how these evolutionary forces operate alone and in conjunction with each other within very specific geographical, ecological, behavioral, and genetic contexts. Change any one of these contexts and the same evolutionary force can give rise to very different evolutionary outcomes, creating very different patterns of speciation.

A. Natural Selection

Darwin (1872) was the first to popularize the idea that speciation could readily occur through the prolonged action of natural selection, the process by which genetic variants that are better suited to the natural environment increase in frequency while variants that are not well suited decrease. In Darwin’s model of speciation, competition to survive was thought to be extremely intense. Therefore, the modification of descendants through natural selection served to increase the adaptation of any one population to the environment while also acting to decrease the intense competition between populations by promoting diversification. Over time, speciation would result automatically under these conditions.

Because of the general acceptance of Darwin’s idea that natural selection is a potent driving force in species divergence, it is often erroneously considered the only evolutionary force important in species formation. However, natural selection can act to cause change (e.g., through directional or disruptive selection as imagined originally by Darwin) as well as to prevent change (e.g., through balancing selection—see Charlesworth’s article on population genetics in this volume, Section III.E). Because of the dual nature of selection, speciation studies need to distinguish between those circumstances under which natural selection is promoting speciation versus those where natural selection is a force that actually helps populations remain similar to each other and needs to be overcome in order for change to occur.

Natural selection’s role as a powerful agent for stasis rather than change comes into play in speciation studies in many instances. One particular instance that has been the focus of a prolonged debate in speciation studies surrounds the issue of coadapted gene complexes. Organisms are required not only to adapt to the external environment but also to their own internal genetic environments. Organisms develop and survive as a result of the coordinated control of a complex network of genes selected to function together properly. All possible genetic changes that may help an organism survive in a particular external environment must also function within the constraints posed by this internal genetic environment in order for these changes to be propagated through the population. Genes that have been selected to work particularly well together are called coadapted gene complexes. Once coadapted gene complexes have evolved, it can be very difficult for natural selection to deviate from the established optimum (termed a fitness peak) in order to reach a new optimum that may ultimately be better for the population, but would require the population to experience a temporary loss of fitness (through suboptimal gene interactions) during the transition.

Sewall Wright (1932) coined the phrase fitness landscape or adaptive landscape to describe this complicated behavior of fitness as a function of gene interactions. Different genotypic combinations plotted in two dimensions create a complex fitness landscape in the third dimension, much like how a mountainscape appears from an airplane. Mountain peaks correspond to regions of high fitness (for particularly good combinations of genes) separated by valleys of low fitness (for not so good combinations of genes). Natural selection, being somewhat myopic, would work to drive populations up a local peak, but would not be able to take populations across valleys to neighboring peaks of fitness even if those peaks represented greater overall fitness.

Because the true nature of adaptive landscapes is a point of contention for competing models of divergence (for a review, see Hollocher article in Grant, 1998; see also Coyne et al., 1997; and Wade and Goodnight, 1998), much work in population genetics has been geared toward understanding what the fitness landscape...
really looks like under a variety of circumstances (for a review, see Whitlock et al., 1995). In spite of its importance, it has been relatively difficult to get a firm handle on the topology of adaptive landscapes partly because at any given moment natural populations tend to occupy a single peak. Advances have been made more recently by studying microbial and viral populations, which can be followed for tens of hundreds of generations (Lenski and Travisano, 1994; Burch and Chao, 1999), and through in vitro studies of the evolution of new functions for specific classes of nucleic acids (for example, see Huynen, 1996), both of which provide more direct information about the actual shape of adaptive landscapes. Different species themselves represent the successful transition from one optimal state to another that has been frozen in time. Characterizing the genetic transitions that distinguish closely related species gives us clues to the actual shape of the fitness landscape from a complementary perspective. It also allows us to generate genetic models of speciation that describe how different species are able to make this shift from one fitness peak to another.

B. Genetic Drift
Given the possible constraints posed by selection acting on coadapted gene complexes, an additional mechanism that would facilitate the movement of a population from one fitness peak to another is often incorporated into models of speciation. Genetic drift is the random change in allele frequencies caused by sampling across generations in a finite population (see Charlesworth’s article on population genetics in this volume, Section IV). The actual change in allele frequency caused by genetic drift is random for any given generation; however, the effects of drift will accumulate over time. In the absence of the introduction of new variation through mutation or immigration from another source, genetic drift will serve to decrease the amount of genetic variation present in a population. Because this process is entirely random, how one population responds can differ from how another population responds just by chance, and drift can actually serve to differentiate two populations in the complete absence of natural selection.

In Wright’s original model describing adaptive landscapes, genetic drift was one mechanism by which populations could move from one adaptive peak to another. For small population sizes, drift effects can be so strong as to overcome the effects of natural selection. If genetic drift could free the population from the bonds of natural selection holding it on one particular fitness peak, then the population could randomly explore the adaptive landscape and possibly find another fitness peak better than the original one. Individually, genetic drift and natural selection can be very effective for bringing about change. Together they may be even more effective. Although speciation is not specifically addressed in Wright’s original model, it does form the theoretical basis for Founder Effect Speciation, which will be discussed in more detail later in this article (see Section IV-B).

C. Sexual Selection
Adapting to the natural environment is only one manner in which organisms can increase their ability to leave more offspring for future generations. Another method is to improve their ability to secure mates through the action of sexual selection. Sexual selection arises from differences in reproductive success caused by competition over mates (Darwin, 1872; Andersson, 1994). Because variance in reproductive success can often be quite high (especially for males), adaptation to the mating environment can sometimes be even more effective at promoting change than adapting to the natural environment and has resulted in the rapid divergence of traits related to reproduction, such as the morphology of male genitalia, sperm composition and abundance, and male secondary sexual characteristics (Eberhard, 1985; Grant and Grant, 1997).

The pressure for sexually reproducing organisms to have compatible reproductive systems ensures that male and female reproductive systems will coevolve. Reproductive systems include behavioral traits related to mate recognition, courtship, and mating as well as physiological traits related to gametogenesis, fertilization, and the production of offspring. Reproductive systems are generally subject to several counteracting evolutionary forces that drive the system in different directions; therefore, mating systems within any particular population tend to reach a stable equilibrium (Lande, 1981; Iwasa and Pomiankowski, 1995). Although the population may be evolving directionally through time, individuals who deviate too much from the norm at any given point along the evolutionary trajectory will generally be disadvantaged either because of mate discrimination or because of natural selection pressures. Although stabilized with respect to any one population, reproductive systems can be very fluid with respect to the number of stable equilibria that are theoretically possible. Therefore, across several populations, traits related to reproduction are particularly suscepti-
The pressure to maintain sexual compatibility (positive male–female coevolution) within sexually reproducing populations is only one important evolutionary consideration for the role of sexual selection in speciation. In addition, there is a second evolutionary force operating that involves antagonistic male–female coevolution which is also very effective at bringing about change, especially in the case of internal fertilization (see Rice in Howard and Berlocher, 1998). The reproductive interests of males and females are not always identical, especially when there is any deviation from strict monogamy. In these cases, evolutionary conflicts can easily arise. For example, females may mate with multiple males to maximize their own reproductive success, but that creates a battleground for sperm competition among males. To combat the intensified selection for sperm to perform better under these circumstances, the chemical composition of male seminal fluid may evolve to include proteins that act to kill other sperm. These proteins in turn may also have negative pleiotropic effects on the female reproductive system itself, which is then selected to combat these negative effects. In this scenario, polygamy sets the stage for continual antagonistic selection to occur between male and female reproductive systems. As seen earlier, within any one particular population, antagonistic coevolution will force the male and female reproductive systems to track each other very closely. However, the effect is very localized and populations that are not in contact with each other have the potential to diverge rapidly with respect to their reproductive physiology (Wu et al., 1996; Rice article in Howard and Berlocher, 1998).

D. Mutation

Neither selection nor drift can operate in the absence of genetic variability. Although quite a bit of variability can be generated just through genetic recombination and the random assortment of chromosomes during reproduction, ultimately spontaneous mutation is the primary source of genetic variation (see Charlesworth’s article on population genetics in this volume, Section III.B). Mutation is considered random with respect to the environmental challenges faced by organisms, although that is not to say it is without bias. Because mutations are important in the context of their effects on the phenotype of organisms, certain phenotypic classes of mutations can be more common than other classes of mutations. In this respect, most spontaneous mutations tend to be deleterious to the fitness of an organism. In addition, mutations are constrained by the chemical nature of DNA or particular properties of the DNA sequence itself, resulting in certain types of changes being more common than others. This can lead to certain classes of base substitutions being more common and to recurring mutation producing the same allele several times independently as is seen for mutations involving slipped DNA replication of tandemly repeated sequences.

An important consideration for speciation studies is whether new mutations tend to have large or small phenotypic effects (i.e., does divergence occur in leaps and bounds or through the accumulation of numerous small changes?). In order to model speciation, it is also important to know whether mutations tend to have dominant or recessive effects with respect to fitness—the dynamics of how allele frequencies change over time are dependent on whether or not alleles at low frequencies (as is always the case for new mutations) are readily subject to natural selection or are masked by the effects of other alleles.

IV. COMMON MODES OF SPECIATION

The crux of speciation studies involves understanding the processes by which a single population splits to form two diverging lineages. Speciation involves the weakening of forces that hold populations together (such as the presence of gene flow, shared mating systems, recurring mutation, and natural selection in common environments) in favor of forces that drive them to diverge (such as the action of genetic drift, mutation, and natural and sexual selection in different environments in the absence of gene flow). Most modes of speciation incorporate the action of all the evolutionary forces discussed earlier (natural selection, genetic drift, sexual selection, and mutation), but emphasize the operation of certain forces over that of others. Several of these modes of speciation are discussed in another article in this volume (see “Speciation, Processes of,” by Bush in this volume; see also Ridley, 1996); therefore, I will focus my discussion here on a subset of all possible modes of speciation, concentrating on the specific interaction of different evolutionary forces operating within the context of different geographical and ecological situations.

A. Classical Allopatric Speciation

Perhaps the simplest conceptual framework for speciation is the allopatric model. Intuitively, it is easy to
recognize that if a physical barrier prevents populations from exchanging genes, they will be free to follow independent evolutionary trajectories. Extensive geographical variation exists (reviewed in Mayr, 1963); therefore, any complete barrier to gene flow most likely means the operation of natural selection will differ to some extent on either side of the barrier. Even in the absence of varying environmental pressures, the continual action of mutation, genetic drift, and sexual selection (see Section III) can change the populations relative to each other over time, if they are completely isolated from each other and no longer actively exchanging genes.

Given enough time, speciation will most likely occur under this scenario of no gene flow between isolated populations. Therefore, the interesting issues to explore in allopatric speciation involve examining the relative rates of divergence caused by the action of natural selection, sexual selection, mutation, and drift as well as determining which traits are most easily affected. There are no general rules that can be presented for what happens during allopatric speciation; examples for all possible scenarios (where each of the evolutionary forces has been shown to play a role to a varying degree) can be cited. The specific outcome is very much dependent on the circumstances surrounding the isolation. It has also been shown well that changes in the reproductive system (both prezygotic behavioral changes in mating propensity and postzygotic physiological changes in the ability to generate offspring) occur rather commonly when populations are separated (Rice and Hostert, 1993; Ridley, 1996).

B. Peripatric or Founder Effect Speciation
Peripatric speciation represents a variation on allopatric speciation. In this case, a small population forms at the periphery of a larger population. This type of geographical isolation can happen anywhere, but it is most easy to visualize in the situation of founders colonizing oceanic islands or, more generally, isolated pockets of habitat. Because the founding of peripheral populations can sometimes involve the movement of only a few individuals (or even a single gravid female), this type of speciation has also become known as founder effect speciation. The emphasis here is on the interaction between genetic drift and natural or sexual selection that occurs during the early stages of speciation.

If a new population is founded by a small number of individuals, just by chance the genetic composition of the founding population may differ significantly from that of the original source population because of genetic drift (see Section III.B). The population need not remain small for very long in order for this sampling effect to influence the future evolutionary trajectory of the population. In addition to this immediate genetic change, oftentimes small populations founded in peripheral habitats or on islands also experience changed environments (both the physical environment, including such things as the quality of the habitat, the distribution of resources, or the presence of competitors or predators, as well as the mating environment, represented by a shift in the distribution of available mating types and preferences), creating new selection regimes. Even in the complete absence of new selective environments, the shift in allele frequencies alone can potentially have a profound effect on how the population will respond to selection because of the changed internal genetic environment that results from drift. The combination of shifting gene frequencies by drift and the presence of potentially new selection regimes under this scenario has led some researchers to propose that this type of speciation can occur more rapidly than the more standard form of allopatric speciation which generally involves populations of larger size and less drastic changes in the physical and mating environment upon isolation (for a review, see Hollocher article in Grant, 1998; see also Ridley, 1996).

The theoretical framework used to justify the conclusion that speciation would be accelerated during founder effect speciation stems directly from Wright's model of an adaptive landscape (see Sections III.A and III.B). A major underlying genetic assumption that enters into the idea that the random sampling of alleles during the founder event can have a profound effect on the evolutionary trajectory of a population is that epistasis (where interactions between alleles at different loci produce phenotypic effects that are not predicted by the action of the individual allelic effects considered alone) and pleiotropy (where a single locus can directly influence more than one phenotypic trait) are quite common. It is under the assumptions of this type of genetic architecture that fitness peaks of varying heights will exist in the adaptive landscape and where random shifts in allele frequencies can have profound effects (for example, see Gavrilets and Hastings, 1996; for a review, see Hollocher in Grant, 1998). If allelic effects are more additive (where interactions between alleles at different loci are minimal), then allele frequency changes do not greatly affect the action of natural selection.

Much of the debate surrounding the likelihood of founder effects accelerating the process of speciation has focused on the specific influence drift alone would have on the probability of shifting from one fitness peak...
to another (for a review, see the Barton and Hollocher articles in Grant, 1998). What has emerged from these theoretical studies has been the idea that the actual size of the founding population does not play as crucial a role in determining the probability of shifting from one fitness peak to another as does the underlying genetic architecture of fitness. On the basis of these theoretical results, researchers have begun to shift their focus to evaluating the genetic architecture underlying traits that change during speciation to see how often epistasis is an important component (see Section V). Although it has been shown that drift alone is not likely to cause populations to shift from one fitness equilibrium to another (see Barton article in Grant, 1998), it is becoming clear that genetic drift acting in concert with natural selection can facilitate speciation when divergence involves traits characterized by extensive epistasis (see Hollocher article in Grant, 1998).

In addition to the genetic architecture influencing rates of change, the actual nature of the trait itself can affect the type of response that is expected under founder effect speciation. Reproductive isolation (both prezygotic and postzygotic) can be particularly susceptible to rapid change under this scenario because of the tight coevolution of male and female traits that normally occurs via sexual selection (see Section III.C). The random sampling of individuals during a founder event can easily move the population away from the stable equilibrium that characterizes the reproductive system in the original population. Reestablishment of a new equilibrium can often involve a radical shift in the mating system of the new population relative to the ancestral one. For sexually selected traits, random genetic drift coupled with sexual selection can act as a particularly powerful mechanism for driving speciation (Lande, 1981; Kaneshiro, 1989; Grant and Grant, 1997).

C. Parapatry and Speciation through Hybridization

Parapatric speciation occurs when new species evolve in contiguous, yet spatially segregated habitats. Unlike allopatric speciation, the populations that are diverging during parapatric speciation maintain a zone of contact and do not cease the exchange of genes completely. In this case, a balance is achieved between continual gene flow and strong natural selection to maintain divergent populations at the two ends of the contiguous habitats. The zone of contact between the two diverging populations, where hybridization between the differently adapted types takes place, is called a hybrid zone (for more general information on hybrid zones; see Hewitt article in Otte and Endler, 1989; Harrison, 1993; Butlin article in Howard and Berlocher, 1998). There can be a single zone of contact along a linear environmental gradient or several points of contact between habitats that are more patchily distributed. In both cases, the basic dynamics of hybrid zones are similar. In practice, it is generally impossible to distinguish between the situation in which the two populations continually maintained contact during the process of divergence (in which case the hybrid zone would then be considered a primary zone of contact) and the scenario in which the populations were actually allopatric at some point during divergence and then more recently came back into contact (in which case the hybrid zone would be considered a secondary zone of contact).

In either event, hybrid zones are particularly compelling to population geneticists because of the opportunity they present for examining the rate of exchange of genes that may be under different selection pressures in the two habitats. For most stable hybrid zones, the point of contact between the two species represents a semipermeable barrier to gene flow. Examining the distribution of different alleles sampled along a transect through the two habitats can provide insights into the nature of the selective forces operating during divergence. The distribution of alleles at loci that have absolutely no effect on fitness will be governed entirely by mutation and drift. Those that are more generally powerful for driving speciation (Dobzhansky, 1940; Butlin article in Harrison, 1993). In either event, hybrid zones represent natural laboratories for determining whether there is speciation by reinforcement. If two species coming into contact exhibit some degree of postzygotic isolation, it is then theoretically possible for selection to act on mating behavior (increasing assortative mating) to eliminate the production of hybrids, thus reinforcing the divergence that has already occurred and completing the speciation process (Dobzhansky, 1940; Butlin article in Howard and Berlocher, 1998). Reinforcement is a very appealing concept because it allows postzygotic isolation to play
an active role in driving speciation rather than simply being a pleiotropic consequence of divergence. Although undoubtedly reinforcement is a selection pressure that does occur in nature, many theoretical arguments have narrowly limited the range of circumstances under which it is likely to occur and the general consensus is that reinforcement is possible, but probably not prevalent, in hybrid zones.

As is clear from the above discussion, hybridization generally creates zones of tension between the operation of natural selection and gene flow. Speciation is thought to be proceeding through the constant action of natural selection serving to increase adaptation at the two extremes of the zone in the face of a continuous influx of genes that hamper adaptation. Intermediate types that form at the zone of contact are often inferior in fitness and disfavored by natural selection. The actual situation is far more complicated than this simple scenario suggests, even for hybrid zones that generally operate in this fashion (see Hewitt article in Otte and Endler, 1989; Harrison, 1993). More importantly, it has become increasingly clear that hybridization does not always play such a negative role in speciation. Instead, it has been shown repeatedly that hybridization can actually provide an important arena for evolutionary innovation (reviewed in Arnold article in Howard and Berlocher, 1998). In this case, hybridization presents the opportunity for the formation of unique genetic combinations through the mixing of different gene pools. In certain circumstances, these unique gene combinations end up performing better than either parental type in particular habitats and can result in the establishment of novel independent evolutionary lineages and the formation of new species.

D. Sympatric Speciation

Sympatric speciation represents the extreme opposite of allopatric speciation by being wholly independent of geographical context. New species form well within the dispersal range of the ancestral species through divergent natural selection to adapt to alternative habitats. Although ecological adaptation is thought to be the most important process in most cases, sympatric speciation can also occur through the sole operation of sexual selection (Higashi et al., 1999). Because gene flow is known to have a powerful homogenizing effect on populations, it was generally thought that sympatric speciation would be an impossibility. More recently, modeling and several important case studies have revealed this skepticism to be unwarranted (see Bush in this volume, “Speciation, Processes of”; see also the article by Johnson and Gullberg as well as that by Feder in Howard and Berlocher, 1998). Most well-studied examples of sympatric speciation involve shifts from one host to another that occur in the same geographical region. The genetic adaptations involved in host shifts do not appear to be simple and often involve several different loci that interact to influence a wide variety of traits associated with fitness on a particular host. Often adaptation to one host will preclude the ability to do well on the other, especially if changes in life history are needed in order to track the host species more closely. Host-specific fitness trade-offs will then result, creating the opportunity for divergent selection to cause different host-adapted genotypes to segregate in the population. This type of divergent selection alone will probably not result in speciation, but if it is combined with a certain level of host fidelity in which individuals who are reared on a particular host and then return to feed and mate on that same host as adults, then there is a much greater probability that different populations specifically adapted to one host or another will evolve and eventually give rise to new species.

V. GENETIC PATTERNS OF SPECIES DIFFERENTIATION

It is clear from the above discussion that the tempo and mode of speciation not only are affected by geographical and ecological considerations but also can be directly influenced by the nature and genetic architecture of the traits that are diverging. Whether genetic drift can act as a facilitator of speciation during founder effect speciation, or whether specific adaptations generally result in fitness trade-offs that lead to divergence in sympatric speciation, is partially dependent on how commonly epistasis and pleiotropy underlie traits subject to change during speciation. In addition, specific models to explain genetic patterns of postzygotic reproductive isolation (see Section V.A) predict very different outcomes depending on whether isolation involves a few or many genes and whether they act recessively or not. To begin to sort through these different possibilities, it is necessary to evaluate the genetic basis of traits that have diverged during speciation to see if any common genetic patterns begin to emerge.

Ultimately, assaying the genetic basis of traits having diverged between species could be helpful in determining whether genetic patterns of change inform us about the evolutionary forces themselves that played a major role during divergence. If this is true, then genetic patterns of change alone could serve as genetic road maps for navigating the process of speciation. Although we
are just in the beginning stages of research that examines speciation from a strictly genetic point of view, population genetic models investigating what the properties and distribution of these genetic changes would look like under different evolutionary scenarios have begun to be formulated (articles by Hey and Templeton in Schierwater et al., 1994; Templeton article in Howard and Berlocher, 1998; Orr, 1998).

A. Postzygotic Reproductive Isolation and Haldane’s Rule

A remarkable and repeated pattern in speciation is the tendency for interspecific hybrids to be either sterile or inviable (collectively known as postzygotic reproductive isolation). This phenomenon is so striking and common that it forms the central premise of the biological species concept (see Section II B) and has been the focus of intense genetic study over the past several decades. A central organizing principle for the study of the genetics of postzygotic reproductive isolation has been Haldane’s rule. Haldane (1922) observed that “When in the F1 offspring of two different animal races one sex is absent, rare or sterile, that sex is the heterogametic sex.” As Haldane’s rule is obeyed when males are the heterogametic sex as well as when females are the heterogametic sex, the genetic mechanisms that explain Haldane’s rule are thought to be fundamental to speciation in all taxa.

Several explanations for Haldane’s rule have been proposed and an almost equal number have been rejected (reviewed in Laurie, 1997; Orr, 1997; Hollocher, 1998). Part of this controversy stems from the difficulties inherent in performing genetic analyses on such traits as hybrid sterility and hybrid inviability. Part also stems from the fact that most researchers naturally sought a single, universally applicable genetic mechanism to account for all cases conforming to Haldane’s rule. The general consensus today is that Haldane’s rule requires separate explanations depending on whether or not hybrid inviability or hybrid sterility is being considered.

For hybrid inviability, Haldane’s rule results because genetic incompatibilities causing inviability that evolve between species tend to act recessively. Given this recessivity, genes causing hybrid inviability will be expressed in the homogametic sex while remaining masked in the heterogametic sex, thus generating Haldane’s rule. For hybrid sterility, the explanation is more complicated and involves the joint action of several different processes. As is the case for hybrid inviability, Haldane’s rule for hybrid sterility results because genetic incompatibilities causing sterility also tend to act recessively. In addition, however, when males are the heterogametic sex, the evolution of hybrid male sterility is accelerated, most likely due to the additional action of sexual selection driving the rapid divergence of male sexual traits (see Section III C; Wu et al., 1996; Rice article in Howard and Berlocher, 1998). Interestingly, much detailed work on the characterization of the genetic basis of postzygotic reproductive isolation (namely male hybrid sterility) has revealed that many genes of small effect are involved and that epistasis is of primary importance for this trait (reviewed in Wu and Palopoli, 1994; see Wu and Hollocher article in Howard and Berlocher, 1998; Hollocher, 1998); therefore, it is conceivable that the general evolution of postzygotic reproductive isolation may occur so rapidly simply because it involves epistasis, regardless of whether or not drift or selection is the primary force driving the change.

B. Patterns of Genetic Divergence for Other Species Traits

Postzygotic reproductive isolation is only one trait that has generally diverged between species. In order to formulate a broader picture of the genetic patterns of divergence, it is necessary to compare the genetic architecture of this trait with what is seen for other traits that diverge between species (reviewed in Hollocher, 1998). The traits that have been looked at in some genetic detail include interspecific mate discrimination and interspecific differences in secondary sexual characteristics. Overall, the general pattern that emerges is that these traits, too, are governed by many genes of small effect. However, in contrast to what was found for the genetic basis of postzygotic reproductive isolation, epistasis does not play a dominant role in governing the evolution of these traits.

Comparison of patterns of genetic variation within species versus patterns of genetic variation between species for the same traits can be useful for gaining insights into the evolutionary mechanisms that may have played a role during species divergence (reviewed in Hollocher, 1998). If these within-species versus between-species comparisons reveal strong similarities, then generally it can be concluded that speculation proceeded through the same general action of evolutionary forces (in terms of type, direction, and strength) normally operating on these traits within species. In contrast, strikingly different patterns of within-species versus between-species comparisons could reveal the operation of a different set of evolutionary forces operating during divergence of species than what normally occurs within species.

Interestingly, within- and between-species patterns
of genetic variation for mate discrimination and secondary sexual characteristics are very similar, indicating that divergence of these traits probably reflects the direct extension of the same evolutionary forces (most likely directional sexual selection in this case) that operate on these traits within species. In contrast, within- and between-species genetic patterns of sterility and inviability do not show similar patterns at all. Not only is the role of epistasis drastically different between the two comparisons, the relative frequency of genes that affect sterility versus inviability is completely reversed depending on whether within- or between-species patterns are considered. This disjunction between the genetic patterns observed within species versus those observed between species suggests that the evolution of postzygotic reproductive incompatibilities may result from the accumulated action of relatively rare evolutionary events happening over long periods of time. Such rare events may include periodic episodes of random genetic drift happening alone or in combination with natural and sexual selection working on these traits with varying intensity or changes in direction over the course of evolution.

VI. NEW DIRECTIONS FOR STUDYING SPECIATION

A. Incorporation of Genealogical Models

As discussed above, the classic approach for determining patterns of genetic change between species has been to cross species and analyze the genetic basis of their differences. Most detailed analyses have focused on the genetic bases of postzygotic and prezygotic reproductive isolation. The emphasis of this approach is on understanding how specific biological properties of species change as they diverge from one another. Therefore, the method involves starting with groups of organisms that show interesting differences in evolutionarily important traits and working down to the genetic level to gain insights into speciation mechanisms that may involve these traits specifically. A very different approach is to use gene genealogies to trace the evolutionary history of genetic variation without regard to the phenotypic effects (in fact, assuming that the genes being investigated are neutral). Population genetic theory involving gene genealogies was originally developed to study the patterns of genetic variation within and among populations of a species to investigate microevolutionary processes such as gene flow, genetic drift, and natural selection. This theory is now being extended from populations up to the species level to investigate how speciation affects patterns of variation within and between species (Hey and Templeton articles in Schierwater et al., 1994; Templeton article in Howard and Berlocher, 1998; Wang et al., 1997). This approach examines the genetic structuring of neutral variation within and between species to evaluate the role such things as random genetic drift, continuous gene flow, and different geographical patterns play in speciation.

The two approaches described above supply very different information about speciation. Each approach in isolation has contributed substantially to formulating hypotheses about speciation. Using a combined approach could potentially be even more powerful for investigating important issues in speciation. Future studies of speciation would benefit from continuing to use traditional genetic methods to identify candidate genes for phenotypic traits that may have been critically important in promoting divergence and then analyzing the genealogical structure of these gene regions in the context of the structuring of genes presumed to be neutral to test specific hypotheses about the role the candidate gene regions may have played during speciation. The evolutionary history of genes that contribute directly to phenotypic differences between species is embedded in the evolutionary history of all genes. A combined approach would allow us to tease apart the population dynamics of genes directly involved in speciation from the dynamics of neutral genes. In other words, the genealogical structuring of neutral genes can serve as a baseline measure of population history (supplying information on such things as the geographical patterns of divergence) to compare to the genealogical structuring of genes (thought to contribute more directly to the speciation process itself (which would supply information on selective forces important in divergence). In taking this dual approach to understanding the genetics of speciation, we will be able to move one step closer to inferring the evolutionary process of speciation based on the examination of the genetic patterns of divergence.

B. Incorporation of Developmental Genetics

For a complete understanding of the process of speciation, it is not enough simply to understand in broad terms the number of genes involved in species divergence and the general distribution of their effects. As this article has revealed, this approach has been very helpful so far to give us new insights into genetic mecha-
The study of speciation is a rich and exciting field. Despite being the object of intense study for the past 140 years, speciation studies continue to capture the imagination of evolutionary biologists and offer a wide range of important questions ripe for investigation. The continued vibrancy of the field comes from its ability to give rise to new species. Although the very roots of evolutionary biology are firmly anchored in the field of development, it is only recently that attempts have been made to meld these two disciplines (for example, see Dickinson et al., 1993; Stern, 1998). The union of these two fields is important not only from a mechanistic point of view for understanding how genes are translated into phenotypes but also for investigating how the developmental systems themselves can serve as direct conduits for the action of natural selection in generating phenotypic diversity.

VII. CONCLUSIONS

The study of speciation is a rich and exciting field. However, ultimately it will be necessary to identify the specific products of genes that have changed during speciation in order to understand how the changes in these gene products have specifically led to the changes in the phenotypes that we observe between species. To achieve understanding at all these different levels, it will be necessary to incorporate developmental biology into speciation theories in order to model how specific genetic changes are translated through the developmental program of an organism to give rise to new species. Although the very roots of evolutionary biology are firmly anchored in the field of development, it is only recently that attempts have been made to meld these two disciplines (for example, see Dickinson et al., 1993; Stern, 1998). The union of these two fields is important not only from a mechanistic point of view for understanding how genes are translated into phenotypes but also for investigating how the developmental systems themselves can serve as direct conduits for the action of natural selection in generating phenotypic diversity.

See Also the Following Articles

- BIODIVERSITY, EVOLUTION AND PHENOTYPE, A HISTORICAL PERSPECTIVE
- POPULATION GENETICS
- SPECIATION, PROCESS OF SPECIES, CONCEPTS OF SPECIES DIVERSITY, OVERVIEW

Bibliography

- Orr, H. A. (1998). The population genetics of adaptation. The distri-
bution of factors fixed during adaptive evolution. Evolution 52, 933–948.


Species–area relationships

Edward F. Connor* and Earl D. McCoy†

*San Francisco State University and †University of South Florida

I. Introduction and Underlying Mechanisms
II. Sampling and Statistical Practice in Description of Species–Area Relationships
III. Functional Form of the Species–Area Relationship
IV. Interpretation of the Parameters of Species–Area Models
V. Use of Species–Area Curves in Conservation Biology

GLOSSARY

relative abundance distribution The frequency distribution depicting the number of species in a community as a function of the number of individuals comprising each species.
species–area curve A graphical depiction of the dependence of species richness on area.
species–area model A function used to describe species–area curves.
species–area relationship The dependence of the number of species in a sample region on the area or size of the region.

A SPECIES–AREA RELATIONSHIP is simply the observation that the number of biological species found in a region is a positive function of the area of the region. Species–area relationships are depicted graphically as a bivariate plot of species richness on the ordinate and area on the abscissa, a species–area curve (Fig. 1). Species–area relationships appear to be ubiquitous, having been observed for a wide array of taxa ranging from diatoms to fish, insects, birds, vascular plants, and mammals and for geographical entities such as islands, political entities, woodland, grassland, and cropland habitat patches, lakes, river drainages, and artificial substrates from microscope slides to synthetic sponges and slates.

I. INTRODUCTION AND UNDERLYING MECHANISMS

H. G. Watson first described the species–area relationship in 1835 by remarking that as the area of a county in England increases by a factor of 10 the number of plant species found in that county increases by a factor of 2. The long history of discourse on species–area relationships has evolved from a focus prior to the 1960s on its empirical utility in determining optimal sample size and sample number for community description, for determining the minimum area of a community, and in extrapolating predictions of species richness to areas larger than those sampled to a focus in the 1960s and 1970s on the mechanisms underlying the species–area relationship, in finding the best mathematical
model for species–area relationships, and in explaining species–area relationships in the context of the equilibrium theory of island biogeography (Connor and McCoy, 1979). Since 1980, the primary focus of discussion involving species–area relationships has been their use and application in conservation biology to determine the optimal design of nature reserves and to project the expected loss of species richness from a region undergoing specified levels of area reduction (habitat loss). We review the discourse on species–area relationships beginning in the 1960s and focus our synthesis on the application of species–area relationships in conservation biology.

Three biological mechanisms have been proposed to account for species–area relationships: (1) the habitat diversity hypothesis, (2) the area per se hypothesis, and (3) the passive sampling hypothesis.

A. The Habitat Diversity Hypothesis

The habitat diversity hypothesis (Williams, 1964) proposes that the increase in species richness in large areas relative to small areas arises because large areas have a greater variety of habitats than small areas. This greater variety of habitats permits species that are only found in specific habitats to occur in large areas and permits species that require multiple habitats to persist in large areas, resulting in higher species richness in large areas than in small areas and the existence of species–area relationships. The habitat diversity hypothesis views area as affecting species richness indirectly because of its association with habitat diversity rather than any direct effect of area on the ability of species to colonize or persist in larger areas. Studies that demonstrate positive correlations between the number of species and the number of habitats on an island have been viewed as supporting the operation of the habitat diversity hypothesis.

B. The Area Per Se Hypothesis

The area per se hypothesis (Simberloff, 1976) is based on the assumptions that the abundance of each species in a sample region varies as a positive function of that region’s area and that the probability of each species going stochastically extinct in that area is a negative function of abundance, and therefore of area. Given these assumptions, large areas would have more species than small areas because more species would persist (not go locally extinct) in large areas. The area per se hypothesis suggests that even in a group of patches consisting of a single type of habitat, one would observe a species–area relationship. Correlations between species richness and area in studies that purport to examine a single habitat type and experimental studies that examine the consequences of reducing patch area are viewed as evidence supporting the operation of the area per se hypothesis.

C. The Passive Sampling Hypothesis

The passive sampling hypothesis (Connor and McCoy, 1979) conjectures that larger areas are more likely to receive more colonists than small areas and that these colonists are likely to represent a wider array of species than the pool of colonists arriving on small areas. Therefore, purely as a result of the higher abundance of colonists expected for large areas and independent of any increase in habitat diversity or reduction in extinction probabilities, one would expect more species to arrive on large areas, leading to species–area relationships. Several authors have examined the number of colonists arriving at their study sites and conclude that passive sampling accounts in part for the species richness of invertebrates on intertidal boulders, fish in stream pools and riffles, and birds on forest habitat islands. Each of these studies shows that for habitats that are colonized seasonally, those patches receiving more colonists have higher species richness.
Two other area-dependent factors that affect species richness have more recently entered discussions concerning species–area relationships, particularly for habitat patches: the resource concentration hypothesis and edge effects.

D. The Resource Concentration Hypothesis

The resource concentration hypothesis attempts to explain the phenomenon that habitat patches with large amounts of resources (e.g., monocultures, areas of high plant density, or large patches) have higher densities of insects. Therefore, the resource concentration hypothesis conjectures that population density should be positively correlated with patch area. In 1973, R. B. Root conjectured that the higher density of animals in larger patches might be solely a consequence of movement behavior (the movement hypothesis); herbivores are more likely to find, and remain in, large, monospecific stands of their host plant than in small or heterogeneous patches. If many species have higher population densities in large patches because of the resource concentration hypothesis, then extinction probabilities should be even lower than expected from the area per se hypothesis, and hence resource concentration could contribute to observed species–area relationships. E. F. Connor and co-workers have recently reviewed the existing literature on the relationships between animal population density and patch area and found for insects and birds that positive area–density relationships are common, but not so for mammals.

E. Edge Effects

Edge effects, or habitat edge dependent changes in abundance or risk of mortality, have been reported for species in a variety of taxa in habitat patches. Given that the proportion of a patch that occurs within any fixed distance from its edge is inversely related to area, edge effects could lead to species–area relationships even within a single habitat type. Such an edge effect on species richness would be mediated by a reduction in the abundance of a species on small patches because of a larger amount of “edge habitat,” leading to higher probabilities of local extinction. Therefore, part of the dependence of species richness on area that has previously been attributed to area per se may actually be caused by edge effects.

F. Multiple Causes of Species–Area Relationships

Habitat diversity, area per se, passive sampling, edge effects, and resource concentration are not mutually exclusive mechanisms and may operate individually or in combination to cause species–area relationships (Connor and McCoy, 1979). Experiments on invertebrate colonization on artificial substrates of varying size and habitat diversity clearly demonstrate the joint contribution of habitat diversity and either the area per se or passive sampling hypothesis, or both, to the generation of species–area relationships. Studies that subsampled islands using a constant-sized quadrat on each island in an attempt to eliminate the effect of habitat diversity found strong effects of habitat diversity, but species richness remained a positive function of island area even for data derived from constant-sized quadrats. Using path analysis, D. D. Kohn and D. M. Walsh concluded that area had both a direct effect on species richness (area per se and/or passive sampling) and an indirect effect mediated by the effect of area on habitat diversity.

The evidence that would allow one to clearly partition the causes of a specific species–area relationship among these underlying mechanisms is exacting. Habitat diversity is difficult to measure, and edge effects might make it impossible to separate the effects of increasing area from the effects of area on habitat diversity. The area per se, resource concentration, and passive sampling hypotheses act by affecting the abundances of species on an island. Area per se and resource concentration do so by reducing extinction probabilities via allowing larger populations (or more dense populations) to persist on large islands, and passive sampling does so by proposing that greater numbers of colonists arrive on large islands. To demonstrate an effect of area per se or resource concentration, one must ultimately show that species on average have larger population sizes (or in the case of resource concentration, higher population densities) and lower extinction probabilities on large islands or habitat patches. To demonstrate an effect of passive sampling, one must show that the arrival rate of colonists on large islands or habitat patches is greater than for small patches and that the colonizing individuals arriving on large patches comprise a greater number of species than those arriving on small patches.
To demonstrate that edge effects contribute to species–area relationships would require evidence that species absences from small patches could be uniquely attributed to edge effects and not to the alternative mechanisms discussed above.

II. SAMPLING AND STATISTICAL PRACTICE IN DESCRIPTION OF SPECIES–AREA RELATIONSHIPS

A. Sampling Practice

Two main sampling schemes have been used to generate data on the relationship between species richness and area. The most widely used approach has been to sample physically separated areas such as islands or habitat patches or to sample adjacent or abutting areas of continuous habitat as independent, nonoverlapping replicates (Fig. 2). For physically separated areas, sampling from the observed natural range of areas generates a range of sample areas. For continuous habitat, the region is divided arbitrarily into a series of nonoverlapping subregions to generate a range of sample areas. The alternative approach, which has been used widely by plant ecologists, is to enumerate the locations of species within a larger region and generate a sample of areas by subsampling a range of areas within the larger region. These subsamples may be overlapping or not, depending on the choice of the researcher (Fig. 2).

1. Independent Areas

Physically independent sample areas have been widely used because species–area data could be readily generated for natural geographical units, such as islands, by combining published monographic species lists with published data on island areas. The majority of species–area curves were published in the 1960s and 1970s by scientists gleaned existing data from the monographic literature and combining data on species richness with published estimates of area. Species–area curves based on gleaned data could be produced relatively quickly via a trip to the library. Another advantage of using physically independent sample areas is that they are

![Image of sampling designs](https://via.placeholder.com/150)

**FIGURE 2** Independent and nested sampling designs used in species-area studies. Independent areas are physically separate. Nested areas may be serially self-contained as in A or abutting as in B. If abutting areas are used, then the sampled region is partitioned into a series of equal-sized, adjacent areas, and this process is repeated for the range of different sized areas. Only two areas are depicted in B, one with six replicates and one with three.
arguably also statistically independent. The statistical independence of separate, nonoverlapping areas allows use of the methods of statistical inference and hypothesis testing associated with ordinary least-squares (OLS) regression. The application of OLS regression permitted the estimation of the parameters (and their standard errors) for any of the proposed functional forms of the species–area relationship that could be made sufficiently linear by transformation of the data and an assessment of the statistical significance of the fit of a particular model. Furthermore, the use of statistically independent areas allowed the rigorous comparison of parameter estimates between studies via analysis of covariance or other simpler techniques.

2. Nested Areas

Plant ecologists have generated data for many years on the relationship between species richness and plot or quadrat area for a variety of plant communities. Much of this work was done to determine the optimal plot size to use when describing plant communities. Presumably because of the saving in sampling effort, plant ecologists used serially self-contained or nested quadrats rather than physically independent quadrats to determine optimal quadrat size. However, nested quadrats are not statistically independent so it is inappropriate to use OLS regression to fit a linear model to such data and assess the fit of the model or estimate its parameters. Therefore, use of nested quadrats to examine species–area relationships in plant communities was largely discontinued by the end of the 1970s. Recent research programs aimed at understanding the structure and dynamics of tropical forests have once again led to studies of species–area relationships using nonindependent data. However, Condit et al. (1996) partition their largest quadrat into as many equal-sized quadrats as possible and use the mean for all nonoverlapping quadrats to estimate species richness for that quadrat size. They repeat this process of partitioning their large quadrat (50 ha) into the maximum number possible of nonoverlapping subquadrats for a variety of quadrat sizes. Condit et al. (1996) then plot the mean and standard error of species richness for each quadrat size as a function of quadrat area to produce a species–area curve. This approach leads to appropriate estimates of species richness and its standard error for each individual quadrat size, since within a size category the quadrats used to estimate the mean and standard error of species richness are independent. However, since the same sampled area is used to estimate species richness for each quadrat size, these estimates are not statistically independent and Condit et al. (1996) appropriately refrain from using regression techniques that require independence to fit statistical models to species–area data.

B. Statistical Practice

OLS regression has been used to fit models to species–area data. By applying OLS regression, estimates of the model's parameters and their associated standard errors can be obtained readily. If one makes the additional assumptions that the model's errors are independent, homoscedastic, and normally distributed and that species richness or the logarithm of species richness is a linear function of area or its logarithm, then rigorous statistical inferences about the parameters may be made. In recent years, nonlinear regression procedures have become widely available so that inherently curvilinear species–area models may be directly fit to data. Nevertheless, OLS regression continues to be used widely. P. J. Vincent, J. M. Haworth, and M. R. Williams point out that the assumptions of normality and homogeneity of error variances are unlikely to be met by species–area data. First, species richness is unlikely to be normally or log-normally distributed and probably should be treated as a discrete, rather than a continuous, variable. Second, the variance in species richness is likely to be a function of the mean species richness. These properties arise in part because on very small areas both the mean and variance of species richness must be nearly zero. Vincent and Haworth suggest analyzing species–area data using a generalized linear model and treating species richness as a Poisson-distributed variable. Williams also suggests analyzing species–area data using a generalized linear model but recommends that species richness be treated as a binomially distributed variable (see discussion of Williams' extreme value model below). In both cases, parameters and their standard errors can be estimated and rigorous statistical inferences can be made. Unfortunately, few authors have explored the treatment of species-richness data as Poisson or binomial variables in studies of species–area relationships.

III. FUNCTIONAL FORM OF THE SPECIES–AREA RELATIONSHIP

When species richness is plotted as a function of area, the resulting plot is a curve that may be linear, concave downward, concave upward, or sigmoid (Fig. 3). The shape of species–area curves appears to be a function of the particular range of areas studied, and this may,
in part, explain the variety of transformations used to linearize species–area curves.

Historically, much of the discussion of species–area relationships focused on the specific functional form of the relationship between species richness and area. In 1921, O. Arrhenius proposed that species–area relationships followed a power function (Fig. 3B):

$$ S = cA^z $$

which, for statistical convenience, has commonly been approximated by the log–log transformation:

$$ \log S = \log c + z \log A $$

with $c$ and $z$ as constants. The double-logarithmic transformation linearizes the power-function model, so that species–area curves are linear on a double-log plot (Fig. 1). In 1922, H. A. Gleason championed an exponential model of the species–area relationship (Fig. 3C):

$$ S = \log c + z \log A $$

because he observed the power-function model to predict impossibly large numbers of species for large areas. The exponential model produces species–area curves that are linear on a semilogarithmic plot. Both of these models were supported because they seemed to fit specific data sets using independent areas reasonably well.
and because it was argued that they could be derived as a consequence of assuming that the distribution of individuals among species (the relative abundance distribution) was either log-normal or log-series, respectively. Other ecologists reported that when a wide range of areas were sampled, neither the power-function model nor the exponential model fit their data since species–area curves appeared sigmoid in shape (Fig. 3D).

Connor and McCoy (1979) examined the fit of the power-function model, the exponential model, and two other models to data from 100 species–area studies. They found that while the power-function model often fit the data well, it was not found to be the best fit model substantially more often than other models. Based on this analysis, the observation that the log-log transformation can linearize a wide range of curves, and because of its widespread use, Connor and McCoy (1979) recommended continued use of the power-function model of species–area relationships.

Recent discussions of the functional form of the species–area relationship have continued to promote use of the power-function model but have continued to justify its use because of its hypothesized connection to log-normal relative abundance distributions (Rosenzweig, 1995). Given a common log-normal relative abundance distribution from which the abundances of species on all sites (islands or habitat patches) arise, it is possible to derive species–area data that are reasonably well fit by a power-function model. However, simply because a particular set of species–area data is reasonably well fit by a power-function model does not imply that this fit derives as a consequence of sampling from a common log-normal relative abundance distribution (Hanski and Gyllenberg, 1997). Many island archipelagos are colonized from multiple source regions rather than from a single source pool, and individual islands within such archipelagos may receive differing proportions of their colonists from each source. On the other hand, habitat patches may be more accurately viewed as units whose species composition is determined by sampling from a source pool with a single relative abundance distribution.

Continued debate about the functional form of the species–area relationship assumes that underlying the variability we observe in empirically estimated species–area curves lies a single true function that will characterize this relationship. However, most data on species–area relationships using independent areas have small sample sizes, cover a narrow range of areas, and may be confounded by area-dependent sampling effort, among other problems. Given the quality of the data available, it will be deviously difficult to discern the functional form of the species–area relationship, even if a single form exists. Furthermore, it is possible that differences among taxa or spatial scales could generate species–area relationships of different functional forms.

A. Self-Similarity and the Power-Function Model of Species–Area Relationships

J. Harte and co-workers have recently proposed an alternative derivation of the power-function model of the species–area relationship. This derivation comes from examining nested areas and making the assumption that successive partitions of a continuous area have the fractal property of "self-similarity." That is, the distribution of species in successive partitions is independent of spatial scale. Beginning with a large continuous area $A_0$ with $S_0$ species, bisections of $A_0$ such that the rectangles comprising the two halves of $A_0$ have area $A_i = A_0/2$ each will have on average $S_i$ species, where $i$ indicates the $i$th bisection of $A_0$. The assumption of self-similarity requires that a species known to be in $A_i$ will have probability $a$ of being found in at least a specific one of the two rectangles of area $A_{i-1}$ created by bisection, and therefore that the fraction of species found in $A_i$ that are found in a specific one of the $A_{i-1}$ rectangles equals the same constant $a$ for all $i$ bisections. The constancy of $a$ for all $i$ bisections follows from the assumption of self-similarity since $a$, the probability of occurrence in a half-patch under bisection, must be independent of spatial scale. Harte et al. (1999) show that the assumption of self-similarity leads directly to the power-function model of the species–area relationship: $S_i = a^i$, with the self-similarity parameter $a = 2^\zeta$. Furthermore, since under successive bisections of an initial continuous area the value of $a$ must lie between 0.5 and 1, the relationship between $a$ and $\zeta$ dictates that the slope of the power-function model, $\zeta$, must lie between 0 and 1. Hence, the power-function model of the species–area relationship is derivable without recourse to any assumptions about the underlying relative abundance distribution of species.

Harte and colleagues have extended the assumption of self-similarity to generate an expectation for the relationship between the number of endemic species and area and the expected spatial turnover between two patches isolated by a known distance, $d$. These extensions of the assumption of self-similarity provide a basis for estimating the expected loss of species from areas undergoing habitat reduction, for calculating $\zeta$ from spatially separated sites, and for estimating species richness at spatial scales larger than commonly possible in previous studies of species–area relationships. However, as Harte and colleagues point out, the assumption...
of self-similarity is not likely to hold for a wide range of spatial scales, so care must be taken to test the assumption of self-similarity for the taxa and spatial scales of inference. While Harte and colleagues have published a few examples where the assumption of self-similarity seems reasonable over a specific range of spatial scales, Plotkin and colleagues show that tropical forest trees are not self-similar in distribution over spatial scales from 1 to $10^5$ m$^2$.

B. Extreme Value Model and Random Placement

An alternative model of species–area relationships based on ‘random placement’ was first proposed by B. D. Coleman in 1981 (Coleman et al., 1982) and extended by Williams (1995). The random placement model of species–area relationships derives the expected number of species on a site, $s(\ln A)$, of area $A_k$ as a consequence of placing the $n_i$ individuals of the $i$th species on sites independently and at random with probability $\alpha_i = A_i/A_k$, where $A_i$ is the area of the $i$th site, $A_k$ is the combined area of all sites, and $S$ is the total number of species among all sites:

$$\tilde{s}(\alpha_i) = S \sim \sum_{i} (1 - \alpha_i)^{y}$$

As presented by Coleman et al. (1982), data on the total abundance of each species combined among all sites were required to estimate the expected species–area curve under the random placement model. Because the random placement model required data that ecologists seldom have, censuses of the abundances of each species at each site, and because this model does not yield fitted parameters comparable to other species–area models, few authors attempted to fit the random placement model to their data. Williams (1995) extended and adapted the random placement model to be approximated by an extreme value function which permits model fitting within the context of the generalized linear model and requires data only on species richness and area, not the abundances on the individual species. The extreme value function model of species richness, $\tilde{s}$, in log $A$ is then

$$\tilde{s} = P[1 - \exp(-\exp(y \log A + \log d))]$$

with $P$ being the number of species in the species pool, and $y$ and $\log d$ the slope and intercept of the model, respectively. Williams (1995) outlines a method for estimating $P$ if the species composition of the biota is unknown but suggests that the best estimate of $P$ is the total number of species found. The extreme value function model has been fit to a limited number of data sets but appears sigmoid when species richness is plotted as a function of log(area). The random placement/extreme value function model of species–area relationships is appealing because it is derived under a hypothesis of independence within and between species. However, as Williams (1995) points out, discriminating between the power function and the extreme value function models with most existing data sets derived from sampling independent areas will be difficult.

IV. INTERPRETATION OF THE PARAMETERS OF SPECIES–AREA MODELS

The widespread use of the power-function model of the species–area relationship coupled with specific numerical expectations for its slope parameter, proposed by F. W. Preston in 1960, led to numerous attempts to infer biological significance to values of this parameter. The slope parameter measures the rate at which species are added as area increases, and the intercept parameter has been considered a function of taxon-specific attributes and environmental variation. Only a limited attempt has been made to offer biological interpretations of the intercept parameter, partly because in the power-function model the absolute value of $c$ depends on the units in which area is measured (Connor and McCoy, 1979; Rosenzweig, 1995). The parameters of other models of the species–area relationship have generally been treated as fitted statistical constants, with no attempt to interpret them biologically.

The slope of the power-function model of the species–area relationship and patterns of variation in this parameter have been subjected to considerable analysis and interpretation. Connor and McCoy critiqued many of these interpretations in their 1979 review. We briefly examine a few of these interpretations and touch on those proposed more recently.

A. Canonical Slope Values

F. W. Preston proposed that isolated islands in equilibrium that sample colonists from a common log-normal relative abundance distribution with parameter $\gamma = 1$ (Preston’s canonical hypothesis) will have a slope of 1.
0.262 in the power-function model. He subsequently broadened the range of slope values that he expected from isolates to values between 0.17 and 0.33. R. M. May showed in 1975 that using a wide range of log-normal parameter values leads to power-function slope values in the 0.13–0.39 range. Many authors have generated slopes from the power-function model of the species–area relationship and interpreted values in the range of 0.2–0.4 to be consistent with Preston’s idea that species–area curves arise because the species richness of islands results from sampling from an underlying log-normal relative abundance distribution. However, Connor and McCoy (1979) challenged Preston’s idea of a canonical range of slope values by showing that slopes in the 0.2–0.4 range are expected purely as a consequence of the tendency to publish studies that show a high correlation between species richness and area and because the variance in species richness will always be less than the variance in area. Connor and McCoy (1979) concluded that slope values from the power-function model of species–area relationship in the 0.2–0.4 range could not be used as evidence for the existence of an underlying log-normal relative abundance distribution.

B. Island–Mainland Differences in Slope Values

Preston extended his idea that islands or isolates should have power-function slopes in the 0.2–0.4 range to project that nonisolated or mainland areas should have lower slope values than isolated areas. His rationale derived from his belief that nonisolated areas would sample from a truncated relative abundance distribution with a higher ratio of species to individuals. In their monograph “The Theory of Island Biogeography,” MacArthur and Wilson (1967) modified Preston’s idea, suggesting a specific range to be expected for slopes derived from nonisolated areas, 0.12–0.19, and explaining the lower slopes as arising because of the “transient hypothesis.” The transient hypothesis suggests that more transient individuals will be encountered in small, nonisolated areas than in small, isolated areas, which, in turn, will lead to more species being encountered on small, nonisolated areas than on small, isolated areas. The greater number of species found on small, nonisolated areas would depress the slope of the species–area curve for mainland areas relative to that expected for islands. The available evidence, while limited, is consistent with the idea that power-function slopes for mainland areas are lower than those for islands.

Hanski and Gyllenberg (1997) develop dynamical models of species incidence that generate predictions about the slopes of species–area curves as a function of the moments of the relative abundance distribution and the ratio of species’ colonization and extinction rates. Hanski and Gyllenberg (1997) develop two models, one in which sites are colonized from an external source, their “island–mainland model,” and one in which the sources of colonists are internal to the system of sites, their “metapopulation model.” They find that species–area slopes are lower for their metapopulation model and claim that such a model is analogous to mainland areas, while their island–mainland model is analogous to truly insular situations. Hanski and Gyllenberg (1997) suggest that their models imply that the observation of lower slope values on mainland sites is a result of metapopulation dynamics rather than the transient hypothesis.

C. The Effect of Isolation on Slope Values

MacArthur and Wilson extended their transient hypothesis to predict that the slopes of species–area curves should be lower for distant archipelagos of islands than for islands located close to mainland areas. However, T. W. Schoener illustrated with species–area curves for birds in 23 archipelagos that exactly the opposite pattern occurs. Schoener’s result suggests that the slope of species–area curves depends on the size of the source pool of colonizing species, which would be smaller for distant than near island groups. In other words, the low slopes for species–area curves observed for isolated archipelagos tell us no more than that isolated biotas are depauperate. Hanski and Gyllenberg (1997) suggest that Schoener’s observation of lower slope values for isolated archipelagos arises because isolated archipelagos behave according to their metapopulation model, not as an island–mainland system.

D. Other Interpretations of Slope Values

Many other efforts have been made to explain variation in the slope of the power-function model of the species–area relationship. These efforts include (1) attempts to equate the slope of the species–area curve with β, or between-habitat diversity, and the intercept with α, or within-habitat diversity; (2) attempts to predict patterns in the latitudinal dependence of the species–area relationship (both slope and intercept); and (3) attempts to identify taxonomic and trophic group differences in species–area curves, among others. Connor and McCoy (1979) critiqued many of the attempts to interpret the
parameters of the power-function model of the species–area relationship and recommended that these parameters be viewed as fitted constants, with no specific biological interpretation.

E. Conclusions
Species–area relationships represent a pattern expected in nature that may arise from the colonization and development of quasi-independent biotas on islands or habitat patches (island-colonization model) or from the loss, reduction, and fragmentation of a previously continuous or widespread biota into remnant habitat patches (mainland-vicariance model). Historically, inquiry into the mechanisms underlying the species–area relationship and its functional form has been biased toward the island-colonization model. However, models based on self-similarity more closely represent the mainland-vicariance model and provide a basis to unify species–area relationships with other patterns in the geographical distribution of species (e.g., distribution and abundance relationships, compositional similarity among sites, and relative abundance distributions). Continued empirical study of the species–area relationships and inquiry into the functional forms expected under specific biological and sampling models will continue to improve our understanding of spatial patterns of species richness.

V. USE OF SPECIES–AREA CURVES IN CONSERVATION BIOLOGY
What can an understanding of the species–area relationship contribute to the preservation of biodiversity? The question has interested ecologists for more than 30 years since the pioneering work of N. W. Moore, E. Maarel, and others and has led, almost from the beginning, to a bewildering confusion of missteps and dead ends (see Shafer (1990) for an overview). Implicit in this question are the important assumptions that species richness is the primary object of preservation and that area is the primary influence on species richness. Here, we shall explore two separate, but similar, aspects of the question. We shall see what ecologists have been able to conclude about the loss of species accompanying area reduction and what they have been able to conclude about the best way to slow the loss.

A. Loss of Species from Area Reduction
As we have illustrated, the species–area relationship is an extraordinarily common pattern in nature, and it has been documented numerous times. It would seem, therefore, that calculating the loss of species accompanying a certain amount of area reduction would be a rather straightforward exercise. All that one would need to know to perform the calculation would be the original number of species ($S_{original}$), the amount of area reduction ($A_{reduced}/A_{original}$), and the slope of the species–area relationship ($z$). For example, if one assumes a power-function model of the species–area relationship ($S = cA^z$), then $S_{original} = c(A_{original})^z$, $S_{reduced} = c(A_{reduced})^z$, and, therefore, $S_{reduced}/S_{original} = (A_{reduced}/A_{original})^z$ (Fig. 4).

Although calculations of this sort are reasonably com-
mon, they require at least five assumptions. First, area reduction completely eliminates species that were originally present (but some species may survive on relict fragments or disturbed lands). Second, all of the species that were originally present were distributed homogeneously (but most species are unlikely to be distributed in this manner). Third, an appropriate model has been selected to describe the species–area relationship. Fourth, the slope of the species–area relationship is accurate and a constant (but it is not likely to be, for a number of reasons). Fifth, the loss of species is a direct consequence only of area reduction (but it is not likely to be, for a number of reasons). The last two assumptions have received the most attention from ecologists. We shall discuss both of them further. \[ \text{Correct choice of the slope of the species–area relationship is critical.} \]

Clearly, if one knew the original number of species and area and the reduced number of species and area, then the slope would be determined precisely. The point of the calculations presented above, however, is to estimate potential species loss, before it happens, so the slope of the species–area relationship also must be estimated, from theory and empirical study. A great deal of uncertainty accompanies such estimation. First, one must decide which model of species loss is appropriate. For example, if the relationship between species and area were assumed to be linear (Fig. 3A), then a much different rate of loss of species would be predicted to occur than if the relationship were assumed to be a power function (Fig. 3B). Second, one must assume that the slope of the species–area curve is constant between the spatial scales over which species loss is to be estimated. A nonconstant slope value would suggest that an appropriate species–area model has not been completely specified and could lead to substantial misestimation of species loss. Third, one must decide if the reduced area better represents an isolate (a “true” island) or simply a subsample of the original area. Typically, slopes of species–area relationships are greater in the first case (approximately 0.20–0.40, on a log–log scale) than in the second (approximately 0.12–0.19, on a log–log scale). Furthermore, it may be that the reduced area could come to function less like a subsample and more like a true island over time, so that the range of potential slopes could expand dramatically over time. Fourth, even if one managed to choose the more accurate of the two ranges of slope values, an impressive amount of variation would still remain. For example, with the equations presented above, if area is reduced by 80% and \( z = 0.20 \), then species loss is estimated at 27%, whereas if area is reduced by the same percentage but \( z = 0.30 \), then species loss is estimated at 38%. Fifth, even if one knew precisely the correct slope of the species–area relationship in one location at one time, Connor and McCoy (1979) and others have shown that slope is not likely to be amenable to translation across locations, latitudes, or time.

One empirical means for focusing the choice of slope values is the body of literature documenting decline in species richness accompanying area reduction. J. M. Diamond termed this process “relaxation.” A number of studies, for example, have examined the relaxation in the species of mammals that may have occurred on nature reserves, both in North America and in Africa. The results of these studies clearly illustrate the difficulty in choosing a slope of the species–area relationship. For African savanna reserves, R. East has shown that estimates of species loss accompanying a 100-fold reduction in area vary from about 40% to more than 90%, depending on the required minimum viable population size (i.e., the population size that ensures persistence with a certain probability for a certain length of time). W. D. Newmark has shown that similar high variability exists for western North American national parks.

At least three potential reasons underlie the poor predictability generally seen in studies of relaxation. The first two reasons, the presence of statistical or sampling errors and the lack of sufficient time since area reduction took place to allow full relaxation, could be addressed by careful experimentation. The protocol is to document the species richness of a relatively large area, reduce the area, and then document the species richness of the reduced area. Unfortunately, most of these kinds of experiments have been carried out at such a small scale that their general value at the scale of typical nature conservation efforts is questionable. Slope values derived from these experiments are, therefore, not likely to be of much practical value. A few reasonably large scale experiments in area reduction have been undertaken, most notably The Minimum Critical Size of Ecosystem project of D. H. Janzen, T. E. Lovejoy, R. O. Bierregaard, and others. These relatively large scale experiments suggest that the species–area relationship alone does not predict species loss particularly well. Slope values derived from these experiments, therefore, also are not likely to be of much practical value. The third reason, lack of the ecological (i.e., species-specific ecological) information needed to infer that the loss of species is attributable directly to relaxation, cannot be addressed as readily by experimentation.

Recall that one of the assumptions necessary to pre-
 dictates loss of species accompanying area reduction was that the loss of species is a direct consequence only of area reduction. Loss of species actually may occur only as an indirect consequence of area, because of some factor that happens to be correlated with area reduction, such as level of disturbance, complexity of habitat structure, or—probably most importantly—fragmentation. It seems clear that in virtually all real-world cases, area reduction is likely to be accompanied by fragmentation. That is, one relatively large area is not likely to be reduced to a single relatively small area, but rather to an archipelago of relatively small fragments. For example, Simberloff (1992) has shown that fragmentation effects, as well as size-dependent habitat changes, are at least as important as area reduction in predicting species loss in the relatively large scale experiments mentioned previously. It has been known for quite some time, at least since the pioneering work of N. W. Moore and others, that fragmentation can have adverse effects on organisms, independently of the effects of area reduction. The species–area relationship predicts that each of the fragments will support fewer species than the original area, but it alone cannot predict the cumulative number of species in the entire archipelago of fragments. As D. Simberloff and L. G. Abele have clearly shown, we also must know the degree of overlap in the species compositions of the fragments. If, for example, the compositions of the fragments were identical, then the archipelago would support fewer species than the original area. If, on the other hand, the species compositions of the fragments did not overlap at all, then the archipelago could support as many—or, potentially, even more—species than the original area. Assuming, realistically, that compositions overlap to an intermediate degree, calculating the loss of species by the method we have just illustrated could yield either an underestimation, if many species were confined to but a small subset of fragments, or an overestimate, if they were not. Fragmentation, therefore, adds more uncertainty—the degree of overlap in species compositions—to any prediction of the loss of species accompanying area reduction. We could predict the degree of overlap in the species compositions of fragments better if we had detailed information on the habitat requirements and minimum viable population sizes of individual species and on spatial variation in species richness, but as Simberloff and others have shown, such information is notoriously difficult to obtain and, therefore, is in very short supply.

Recently, a method has been suggested by Harte and co-workers whereby the slope value for species–area relationships among fragments might be obtained more easily. This method is based on the idea of self-similarity that we discussed previously. Essentially, the method requires only measurement of the degree to which species lists in separated, small, censused patches overlap as a function of the distance between patches and the area of the patches. Self-similarity, then, allows translation of the resulting slope value across scales, so that censusing of prohibitively large areas or obtaining extensive autoclinal information becomes, in theory, unnecessary. A. P. Kinzig and J. Harte have also developed a procedure for estimating species loss from habitat reduction that derives from the assumption of self-similarity and from examining the relationship between the number of “endemic species” and area, an endemics–area relationship (EAR). In the examples examined thus far, the estimated species loss from area reduction is less using Kinzig and Harte’s EAR approach than from the traditional approach based on the species–area relationship outlined above. Whether these approaches to estimating species–area slopes and calculating species loss will lead to yet another dead end remains to be determined but will depend on how reasonable it is to assume self-similarity over a range of spatial scales.

B. Slowing the Loss of Species

In the previous section, we explored what the species–area relationship was able to tell us about species loss accompanying area reduction. In this section, we shall explore what the relationship can tell us about slowing that loss. In doing so, we shall treat area reduction with concomitant loss of species as an accomplished fact and concern ourselves solely with maximizing the remaining species richness. We shall, in turn, focus on maximizing species richness by the judicious choice, or design, of nature reserves.

The serious study of the relationship between the species–area relationship and the design of nature reserves began in the mid-1970s. At that time, several papers by J. Terborgh, J. M. Diamond, E. O Wilson, E. Willis, and others came out more or less simultaneously touting the overarching importance of largeness of nature reserves. Some of these papers maintained further that the species–area relationship indicates that nature reserves should be subdivided as little as possible, that is, that a single intact reserve of a certain area will support more species than a series of individually smaller reserves that are cumulatively of the same area. Diamond, for example, said “many species that would have a good chance of surviving in a single large reserve would have their survival chances reduced if the same area were apportioned among several smaller reserves.” The basic problem with these ostensibly reasonable proposals—call them largeness and singularity—is that they may be too simplistic for real-world conservation
SPECIES–AREA RELATIONSHIPS

409

Challenges. They may be too simplistic because of the substantial set of assumptions that accompanies them. We shall address perhaps the most basic of these assumptions in detail. This assumption is that the importance of largeness and singularity follows directly from the species–area relationship.

Most of the early papers based their arguments for largeness and singularity on R. H. MacArthur and E. O. Wilson's theory of island biogeography. This theory of island biogeography takes a dynamic view of the maintenance of species richness on 'islands,' as a balance between colonization and extinction. The resulting species richness, therefore, is considered an 'equilibrium number of species.' If relatively large areas support relatively large populations, then relatively large areas also should have relatively low extinction rates and, consequently, support relatively high equilibrium numbers of species. Accordingly, any subdivision of a relatively large area essentially creates smaller areas, each with higher extinction rates, so that the equilibrium number of species on each of the areas necessarily will fall ('relaxation'). The early papers also made a variety of proposals to enhance colonization, reflecting the potential importance of clustering and corridors and of shape, but we shall not address these proposals. Almost immediately, the arguments for the overarch- ing importance of largeness and singularity were challenged. Simberloff and Abele showed that the species–area relationship did not unambiguously favor a single intact reserve over an archipelago of smaller reserves. For example, if one assumes a power-function model of the species–area relationship \( S = cA^{z} \) and that \( A_{\text{tot}} = A_{\text{large}}/2 \), then the number of species in the large reserve is \( S_{\text{large}} = (A_{\text{large}})^{c} = c(2A_{\text{small}})^{c} = 1.2S_{\text{small}} \) (Fig. 4). This equation could yield \( S_{\text{small}} < 1.2S_{\text{large}} \) in a variety of realistic circumstances. Subsequently, R. W. Rafe elaborated the mathematical underpinnings of the comparison of the relative abilities of a single large reserve (SLR) and two reserves of half the area (THR) to support species. R. W. Rafe, A. J. Higgs, M. B. Usher, and others have shown that relatively small fragments could harbor unusually high population densities of species or disproportionate representations of rare species. Finally, this statement implies that relatively small fragments often are not simply random samples of larger habitat units, but rather may represent 'special' places that have been left either inadvertently or purposely. For example, R. T. T. Foreman, J. Hanski, E. D. McCoy, and others have shown that relatively small fragments may have ‘worth’ despite their small sizes. Relatively small fragments are not simply random samples of larger habitat, but rather may represent ‘special’ places that have been left either inadvertently or purposely. For example, R. T. T. Foreman, J. Hanski, E. D. McCoy, and others have shown that relatively small fragments may have ‘worth’ despite their small sizes. Relatively small fragments are not simply random samples of larger habitat, but rather may represent ‘special’ places that have been left either inadvertently or purposely. For example, R. T. T. Foreman, J. Hanski, E. D. McCoy, and others have shown that relatively small fragments may have ‘worth’ despite their small sizes. Relatively small fragments are not simply random samples of larger habitat, but rather may represent ‘special’ places that have been left either inadvertently or purposely. For example, R. T. T. Foreman, J. Hanski, E. D. McCoy, and others have shown that relatively small fragments may have ‘worth’ despite their small sizes. Relatively small fragments are not simply random samples of larger habitat, but rather may represent ‘special’ places that have been left either inadvertently or purposely.
likely to be poorly known, if known at all. This lack of information also plagued earlier proposals for estimating needed area, such as E. D. McCoy’s “minimum refuge area,” T. E. Lovejoy and D. C. Oren’s “minimum critical size,” and S. T. A. Pickett and J. N. Thompson’s “minimum dynamic area.”

C. How Incidence Functions and Nestedness Fit In

Recall that the SLOSS debate was generated largely by one of the early nature reserve design principles: a single intact reserve of a certain area will support more species than a series of individually smaller reserves that are cumulatively of the same area. For this principle to apply universally, the minimal assumptions are that the habitats included on areas of different sizes are more or less uniform, that the population densities of species on areas of different sizes are similar, and that the process of relaxation largely is deterministic. Subsequent ecological research has made it clear that none of these assumptions always holds. However, do these assumptions hold in enough instances or to such a degree that it is possible to know the species composition of various-sized fragments with a high degree of certainty? We shall address two, related, methods developed to address this question.

J. M. Diamond developed “incidence functions” as a tool for determining minimum nature reserve area. The incidence function, J, equals the number of fragments (“islands”) of a certain size harboring a certain species divided by the total number of fragments of the same size. Supposedly, the higher the value of J at a particular size, the higher the probability that the species can persist in fragments of that size. E. F. Connor, D. Simberloff, M. Williamson, and others questioned the validity of these conclusions, but, if they do have any validity, even in a general way, then certain patterns of incidence functions could indicate that species disappear from increasingly fragmented habitats in a predictable manner.

B. D. Patterson and W. Atmar developed a sophisticated extension of incidence functions, the “nestedness” of species’ geographical distribution. Species’ distributions are nested when the species on the most species-poor (and, likely, smallest) fragment comprise a subset of those on the next most species-poor fragment, which, in turn, comprise a subset of those on the next most species-poor fragment, and so on. K. B. Jones and co-workers supplied the first explicit evidence for nested distributions in the mid-1980s, although implicit evidence can be found in papers by N. W. Moore, M. D. Hooper, R. T. T. Foreman, and others a decade earlier. Jones and co-workers attributed the pattern to selective extinctions in the absence of colonizations, as have most subsequent authors, although D. T. Bolger, M. V. Lomolino, and others have shown that the pattern actually can result from a number of causes. Nestedness could have important implications for nature reserve design. If the species on an archipelago of habitat fragments were a perfect subset of the species in a single large unit of habitat, then the archipelago could never contain more species than the single large unit. Furthermore, if fragments of similar size were to harbor similar suites of species, because species loss with area reduction is deterministic, then the archipelago must contain fewer species than the single large unit. The temptation is great, therefore, to assume that simultaneously significant species–area and nested subset relationships must indicate that smaller fragments have the fewest species, but recent evidence supplied by D. Doak, E. D. McCoy, W. J. Boecklen, and others suggests that this assumption is not a good one.

D. Conclusions

Regardless of the reason for it, poor predictability of the effects of area reduction can stall effective decision making. A good example is the controversy over the amount of species loss accompanying tropical deforestation. For more than 30 years, ecologists have sought “shortcut methods” to substitute for the autecological information that is needed to improve predictability but is so difficult to obtain. To date, they have not been particularly successful. An interesting additional example that involves species–area relationships is the controversy over the amount of species loss accompanying global warming discussed by K. A. McDonald, J. H. Brown, R. W. Skaggs, W. J. Boecklen, T. E. Lawlor, and others. So, we are saddled with a seemingly unsolvable dilemma: we need more information, but we cannot afford to wait to get it. Faced with this dilemma, some biologists, for example R. East and M. Kent, have suggested that we view the species–area relationship as perhaps the best tool available for making conservation decisions and have promoted the idea of retaining large units of habitat as a matter of general course. If one is focusing solely on predicting loss of species with area reduction and refrains from becoming overly specific, then this may be a reasonable strategy. Others, for example C. F. Mason and E. D. McCoy, have suggested that reliance on species–area relationships may lead to undesirable conservation decisions and have promoted a “save-all-the-pieces” strategy. If one is focusing on
forestalling loss of species, then this may be a reasonable strategy (which has been termed SLATS ("several little all too small") by E. D. McCoy and H. R. Mushinsky) for habitats suffering from extreme area reduction and fragmentation. The SLATS strategy is likely to become increasingly relevant for most habitats, because the amount of habitat needed to allow species to persist, let alone to flourish or to evolve, appears to be much larger than humans are willing to grant.

See Also the Following Articles

DIVERSITY, COMMUNITY/REGIONAL LEVEL • HABITAT AND NICHE, CONCEPT OF • ISLAND BIOGEOGRAPHY • METAPOPULATIONS • POPULATION DENSITY • POPULATION DIVERSITY, OVERVIEW

Bibliography


SPECIES COEXISTENCE

Robert D. Holt
Museum of Natural History and Center for Biodiversity Research, University of Kansas

I. General Issues
II. Constraints on Coexistence: Examples from Introduced Species
III. Coexistence and Exclusion: Messages from Bottles
IV. Constraints on Coexistence: Rules of Dominance
V. Mechanisms of Coexistence
VI. Traditional Approaches to Coexistence
VII. Mechanisms for Local Coexistence: Current Approaches
VIII. Temporal Niche Partitioning
IX. Coexistence in Open Communities
X. Future Directions
XI. Conclusions

GLOSSARY

coexistence The state of two or more species being found in the same place at the same time.
community The assemblage of species found in a defined area, in which these species can interact.
competition The most widely used definition: use or defense of a resource by an individual which reduces resource availability to other individuals. Alternative definition: a reduction in one species' growth rate because of the effects of another species.
food web The pattern of feeding relationships among organisms.

Lotka-Volterra competition model A model describing the per-capita growth rates of two competing species as linear, declining functions of the abundances of each species.
mechanistic models of population dynamics Models which are explicit about resource consumption, the relationship of consumption to demography, and mortality factors.
microcosm studies Laboratory experiments involving population dynamics of small organisms with short generation lengths.
natural enemy Any species which consumes or parasitizes another species; a general term that includes predators, herbivores, parasites, pathogens, and parasitoids.
niche (Grinnellian/Hutchinsonian) The range of resources and conditions within which populations of a species are expected to persist. (Eltonian) The role of a species in a community.
open communities Communities which receive immigrants from external sources and may export emigrants as well.
permanence A property of communities that ensures long-term coexistence of species because community trajectories (variation in numbers through time) have no species approaching very low numbers.
zero-growth isocline In a graph with axes describing factors important to population growth (e.g., resource abundance), a line along which a population's growth rate is zero.
THE OXFORD ENGLISH DICTIONARY states that to “coexist” is “to live together in the same place, at the same time, with another.” The topic of “coexistence” focuses on how biological species are organized into communities, in space and through time. Identifying the factors that influence coexistence is fundamental to understanding biodiversity. Important factors that influence coexistence include interspecific interactions, spatial and temporal scales, and historical contingencies. Insights into coexistence come from laboratory and field experiments, historical reconstructions and observational studies, and theoretical explorations. The topic of coexistence has had a long history in community ecology, some appreciation of which is necessary to understand contemporary perspectives. After a few general remarks about how scale, interactions, and the notion of permanence are related to coexistence, I provide a survey of the “phenomenology” of coexistence and exclusion, and lessons derived from field and classic lab studies. I then discuss key ingredients in both historical and current theoretical interpretations of species coexistence, emphasizing conceptual generalities rather than model details, and conclude with suggestions regarding significant unanswered questions about coexistence.

I. GENERAL ISSUES

A. The Importance of Scale

In some sense, all the species in the teeming plenitude of life present on Earth coexist because they are found in one place (Earth) at one time (now). However, from the perspective that matters most to an individual organism such as a vampire bat, a tulip, or a flatfish—the perspective defined by the space that an individual occupies or moves within over its life span—populations of any given species co-occur with only a tiny fraction of all living things. The same holds for community samples, for instance, those gathered in the quadrats beloved of plant ecologists or the towed plankton nets of limnologists; the number of species recorded is always much less than the potential number. That is, what counts as “coexistence” varies with the scale of ecological inquiry. Viewed at small spatial scales, tailored to the ecology of individual organisms or seen through the lenses of conventional field methods, most species in the global flora and fauna do not coexist. However, even at this scale many (and sometimes very many) species can coexist. Species coexistence is also temporally bounded. In the long term most species face extinction, and even in the short term coexistence may be transient rather than permanent.

B. The Importance of Interactions

Species (like ships passing in the night) may either coexist or fail to do so for reasons having nothing to do with each other. A null model of community organization is that communities arise from the independent responses of species to the environment. However, coexistence often reflects the impact of interactions among species. Familiar categories of interactions include competition for resources, natural enemy–victim interactions (e.g., predator–prey and host–pathogen interactions), commensalisms and mutualisms (e.g., plant–pollinator relationships), allelopathy (poisoning), and a broad range of environmental modifications (collectively called “ecological engineering”). In the study of coexistence, competition has received by far the most attention, but there has also been considerable work on how natural enemy–victim interactions influence coexistence. Interactions are not fixed properties of species, or pairs of species, but can vary in their strength and pattern as a function of the physical environment and many contingent details of community structure. If species A affects species B, and species B in turn influences species C, then one says there is an indirect interaction between A and C. A complete understanding of coexistence requires one to consider indirect and direct interactions. Interspecific interactions may be strongly asymmetrical, and indeed, interactions frequently cause species exclusion (constraints on coexistence). Analyzing mechanisms of coexistence and exclusion forms the conceptual core of the discipline of community ecology.

C. Coexistence and Permanence

Historically, coexistence is an outcome of the assembly of communities by colonization from species pools at larger spatial scales. If rules describe the historical process of community assembly, these rules must in large measure reflect the importance of species’ interactions in determining coexistence rather than extinction as assembly occurs. The paleontological record shows that species’ distributions are not stable but move in concert with climate change, fluctuations in sea level, and the stately movement of the continents. Moreover, for many species, chance vicissitudes of dispersal provide a small trickle of potential colonists “testing the waters” outside their normal range. Local coexistence at the very least re-
quires species’ geographical ranges to overlap. Many species do not coexist together today in local communities because of historical factors preventing them from ever encountering one another.

Within a species’ geographical range, for that species to be a resident member of a local community, at one time it must have colonized, increasing from very low numbers. Once present, given the vagaries of climate and fluctuations in the abundance of other species, it is likely that resident species experience times of low abundance and thus extinction risk. To persist, a species must be able to rebound from these dangerous troughs of low population size. These observations suggest that a robust form of coexistence is ensured if each species can increase when rare. If all trajectories describing fluctuations in abundance in a community are bounded away from zero, all species in the community will coexist. This criterion for coexistence is called “permanence” in the jargon of mathematical ecology. Examining conditions for increase when rare provides a natural protocol for experimental studies of coexistence and analysis of mathematical models of interacting species. Permanence is always assessed relative to a certain spatial and temporal scale. A given pair of species may not coexist indefinitely in any single site but nonetheless coexist at larger scales in ensembles of sites.

II. CONSTRAINTS ON COEXISTENCE: EXAMPLES FROM INTRODUCED SPECIES

In recent times, humans have greatly accelerated movements of species into novel habitats, both via deliberate introductions (e.g., most of the lowland birds found in Hawaii) and as incidental “hitchhikers” tracking human transport (e.g., most benthic invertebrates in San Francisco Bay are exotics released as by-products of ship ballast water dumps). Analyses of introductions suggest species interactions can constrain coexistence.

A. Competition

Classical biological control involves deliberate introductions of species that one hopes will control a pest species. Between 1947 and 1952 the Hawaii Agriculture Department released parasitoid species to control the oriental fruit fly, an economically significant pest. (Parasitoids are insects, e.g., braconid wasps, whose larvae live within and ultimately kill their host, such as caterpillars.) Three wasps in the genus Opius were considered to be potential control agents. The first species established, Opius longicaudatus, parasitized approximately 20% of the fruit fly hosts. The second species introduced, O. vandenboschi, was more effective, parasitizing about 30% of the hosts; as O. vandenboschi increased, O. longicaudatus decreased toward extinction. The third species, O. oophilus, was even more effective, parasitizing up to 80% of the hosts and replacing in turn O. vandenboschi. This system provides an example of competitive exclusion in exploitative competition, where species dependent on a single limiting resource (here, host insects) cannot coexist. Increasing parasitism reduces the availability of unattacked hosts. The competitive dominant in this pattern of competitive displacement is the parasitoid species that persists at the lowest host abundance—an example of a simple rule of dominance in resource competition. Understanding how multiple species can persist on a limited resource base is a perennial theme in the study of coexistence.

B. Natural Enemies

Very effective control agents can, after introduction, eliminate the target species (and thus themselves) over broad areas. The floating fern Salvinia molesta from Brazil escaped from a garden in Sri Lanka and became a serious aquatic pest in much of the Old World wet tropics. The beetle Cyrtobagous salviniae was introduced in Australia and proved highly successful at limiting the fern, which (with the beetle) is now found in only a few scattered populations. This example illustrates another species coexistence problem; that is, how effective natural enemies manage to coexist with their prey without overexploiting them and thus driving themselves to extinction.

C. Community and Spatial Contexts

Many game departments routinely introduce game species. Failed introductions reveal constraints on species coexistence. Caribou were once common in the Maritime Provinces but declined during the 1800s to extinction by 1913. Attempted reintroductions have failed because the caribou pick up a widespread nematode parasite (carried by smalls and incidentally consumed as the caribou forage on vegetation). The nematode lodges in the caribou brain and is fatal within a few months. The nematode population is sustained by another host, white-tailed deer, which tolerate parasitism. Deer are abundant because land practices by European settlers in the nineteenth century increased the area of
the second-growth habitats that deer favor; this may explain the disappearance of caribou. Caribou herds wander very widely, which makes it likely that they will encounter nematode-laden snails. Another ungulate species, moose, coexists with deer, even though infected moose also quickly die. Moose and deer overlap in diet and can compete; in contrast, caribou and deer do not overlap in diet. Nonetheless, it is caribou, not moose, which have been indirectly excluded by deer. Moose are relatively sedentary (for their body size) and can occupy patches of highland forests with deep winter snows, where deer do not penetrate. At the coarse scale of North America, caribou and white-tailed deer coexist but with little spatial overlap among populations. Moose and deer, in contrast, overlap geographically, coexisting at the level of landscapes, but at a finer scale they have considerable spatial segregation in habitat use.

This example illustrates several points. First, natural enemies (a term that refers to any species which consumes a species of interest) can prevent coexistence. Second, coexistence or exclusion may arise from complex webs of multispecies interactions. The nematode directly causes exclusion of caribou, but viewed more expansively exclusion is caused by the entire vegetation–snail–deer ensemble, which collectively governs nematode abundance and hence caribou infection rates. Third, movement patterns can influence interspecific interactions. Moose and white-tailed deer coexist at the landscape scale because moose have spatial refuges from infection, whereas caribou do not and are thus excluded.

III. COEXISTENCE AND EXCLUSION: MESSAGES FROM BOTTLES

The previous examples strongly suggest the importance of species interactions in determining coexistence but are not conclusive. Without detailed observational studies tied to parameterized models or rigorous manipulative field experiments, it is difficult to persuade skeptics that other hypotheses are not also plausible. As an alternative, one simple but illuminating approach to the study of coexistence is to put together a few species in a confined setting and see what happens. The study of "nature in a bottle" simplifies the world in important respects (e.g., a few species in closed and temporally constant environments), and such experiments are viewed skeptically by some ecologists. Despite these limitations, a wealth of important messages about species coexistence emerged from classic bottle experiments—messages which have proven robust when applied to natural communities. Microcosm studies of ecological processes are now enjoying a renaissance in community ecology.

A. Competition Experiments

The term "competition" usually refers to an interaction among individuals (within or among species) that arises because they seek a resource in short supply. If this involves direct harm, the interaction is referred to as "interference competition." If instead competition involves depletion of a resource, one refers to "exploitative competition."

1. Protozoa

Most ecology textbooks (Hutchinson, 1978) recount experiments by the Russian ecologist G. F. Gause, who as a young man in the 1930s put mixed cultures of protozoa into vials full of liquid media to study species coexistence. Gause's famous competition experiments compared populations of the ciliate protozoans Paramecium aurelia and P. caudatum grown separately and together on a nutritive medium containing their essential resource (bacterial food). Both species thrived when alone, but P. aurelia usually displaced its congener in joint cultures within 30–50 generations. This outcome was reversed if the medium was completely replenished with fresh nutrient on a regular basis. Gause argued that metabolic by-products were building up in the experiments, and that part of the dominance of P. aurelia may involve its resistance to the chemical by-products of metabolic activity as well as its superior ability to exploit the food base. Here, competition combines environmental modification and exploitation of a limiting food resource.

In other experiments Gause found that P. aurelia could coexist with another species, P. bursaria, even in the confines of a closed culture. Paramecium bursaria contains symbiotic algae, which release oxygen in photosynthesis. In incompletely mixed cultures, bacteria accumulate on the bottom, creating a zone slightly depleted of oxygen. The protozoan with the algae in effect carries its own oxygen supply into this anoxic habitat and so can use a food source unavailable to the other, competitively superior species. Here, coexistence depends on both the availability of different habitats and differential species' abilities to utilize those habitats.

2. Beetles

Many ecologists have examined the dynamics of mixtures of beetle species competing for grain stored in
containers. L. C. Birch examined several species combinations that led on wheat or maize under different conditions of temperature and moisture. Usually, one species won. The criterion for dominance was simple: the winner always had the highest intrinsic growth rate (the maximal rate of population growth when a species is rare and growing alone). However, the identity of the winner depended on both resource type and physical conditions. Moreover, competitive exclusion sometimes required many generations; in one experiment, the apparent loser survived until the experiment had to be terminated.

Experiments by T. Park with a different set of stored grain pests (Tribolium castaneum and T. confusum) also showed that the species identity of the winner varied with temperature and humidity. However, for some abiotic conditions the outcome was indeterminate, either species could win, with the winner tending to be the species initially most abundant. The net interaction between these beetles combines several mechanisms, including exploitation of a shared resource and cannibalism (both within and between species). In one set of experiments, Park intriguingly observed that when a coccidean parasite, Adelina sp., was present, the normal dominant T. castaneum became a weak competitor, thus permitting the persistence of T. confusum (and even its dominance).

Sometimes, coexistence in microcosms reflects very subtle biological differences between species. Crombie raised two grain beetles (Rhizopertha dominica and Oryzaephilus surinamensis) in vats of cracked wheat and found indefinite persistence, despite the fact that there seemed to be just a single resource. The main difference in the two species seemed to be in larval feeding habits; larvae of Oryzaephilus lived and fed from outside the wheat grains, whereas larvae of Rhizopertha were sufficiently small to live and feed from within the grains. This slight difference sufficed for the competitors to coexist.

3. Key Insights

Several insights emerged from “bottle” studies of competition. First, species competing in a closed homogeneous medium for a single limiting resource typically do not coexist. Second, the winner depends on environmental conditions; and the outcome can change if the environment is altered—there is no universally superior competitor. Third, exclusion is not instantaneous and may require many generations; this leads to the possibility of transient coexistence. Fourth, indefinite coexistence occurs but requires heterogeneity in the environment as well as differences in species’ responses to this heterogeneity. However, the species’ differences permitting coexistence can be quite subtle. Fifth, pairs of species can interact in many distinct mechanistic ways (e.g., via impacts on resource levels or on levels of pollution from metabolic waste). It is difficult to generalize among studies if one has not clearly identified and characterized specific mechanisms of interaction. Finally, the outcome of competition can vary due to effects of other species, including parasites or mutualisms.

B. Predation

In classic literature, compared to studies of competition, somewhat less attention was given to predator-prey and host-parasite interactions (and essentially none to mutualism). Gause did carry out experiments with the predatory protozoan Didinium nasatum and a prey protozoan, P. caudatum. The predator quickly overexploited its prey and went extinct. To achieve coexistence, Gause concluded that there needed to be a spatial refuge for prey, inaccessible to the predator, or recurrent immigration. A qualitatively similar message emerged from experiments by C. Huffaker and colleagues, who studied an intrinsically unstable interaction between a voracious predatory mite (Typhlodromus) and a prey mite (Eotetranychus) living on oranges. In single oranges, the predator rapidly drove its prey to extinction. In contrast, in a large “universe” of oranges laid out in a grid, separated by barriers to dispersal, the interaction persisted over many generations, with complex patterns of spatial occupancy for the two species. Luckinbill reexamined the unstable predator-prey pair studied by Gause but reduced the encounter rate between predators and prey by adding methyl cellulose (a thickener) to the medium so as to slow predator movement. Reducing movement rates in effect expands the size of the spatial arena of the interaction (scaled by the distance a predator moves per unit time). This led to stable coexistence. The qualitative message of these microcosm experiments on predation is that patchiness, localized dispersal, and spatial heterogeneity (e.g., refuges) may facilitate the coexistence of effective specialist predators and their vulnerable prey.

IV. CONSTRAINTS ON COEXISTENCE: RULES OF DOMINANCE

A. The Competitive Exclusion Principle

The experimental observation that in homogeneous well-mixed environments it was often difficult to
achieve coexistence between similar species became enshrin ed in ecology as Gause's principle, or the "competitive exclusion principle." Another way to state this principle is to note that, to coexist, different species must have distinct ecologies.

Assume we are examining a local community defined by a spatial scale in which all individuals can reach all sites over a single generation. The basic logic of the competitive exclusion principle is impeccable (Levin, 1970): Consider two species with continuous generations, where both species respond to the same environmental factors, denoted by E. The quantity E could be many things (e.g., resource availability, predator abundance, or a weighted sum of the competitors' own numbers). The growth rate of species i is dN_i/dt = N_i(E), where N is the density of species i. We make three assumptions: (i) There is a single limiting factor, (ii) the species interact in a closed habitat (i.e., no immigration), and (iii) each species when alone settles down to an equilibrium at constant densities (i.e., the environment is temporally constant).

For species i, there will be some value of the environmental factor, E^*, at which that species equilibrates. It is a biological truism that any two species will almost surely differ in some way. Hence, it is very improbable that they will have exactly the same value of E at which they reach equilibrium. In other words, there should be no long-term persistence of two species limited by a single factor in a constant, closed environment. For this argument to work, there should be some effect of the species themselves on the magnitude of E, leading to either direct or indirect density dependence in demographic parameters such as birth or death rates. For example, if two species are consuming a single resource, then consumption should depress resource levels.

1. Exploitative Competition: The R* Rule

What counts as a "limiting factor" needs to be interpreted quite broadly and is often quite difficult to identify in practice. Nonetheless, sometimes one can observe a single, simply characterized limiting factor, which then defines a rule of species' dominance. For instance, the Hawaiian parasitoids fit the "R* rule" proposed by David Tilman; that is, the dominant species, given exploitative competition for a single resource, is the species persisting at the lowest resource level. (The asterisk denotes equilibrium, and R* is the equilibrium resource level at which a given consumer species is in demographic equilibrium.) A virtue of this dominance rule is that one can measure R* of each species when alone and then predict the outcome of competition. A species may have a low R* and be competitively superior either because it is efficient at resource consumption or because it has low morality (e.g., it can escape predation).

The R* rule successfully predicts the outcome of competition among phytoplankton competing for nutrients in microcosms and also characterizes competition in some natural systems. For example, in the nitrogen-poor soils of Cedar Creek, Minnesota, plant species compete for nitrogen. Wedin and Tilman showed that species with low R* for nitrogen won in pairwise competition experiments. If two species had similar values for R*, the rate of competitive displacement was greatly reduced, as expected by theory.

2. Apparent Competition: The P* Rule

Analogous "rules of thumb" arise in other situations. For instance, Sharon Lawler carried out a microcosm experiment in which a predatory protozoan (Euplotes patella) coexisted with either of two prey protozoans (Tetrahymena pyriformis and Chilomonas paramecium) grown alone. However, when all three species were together, Chilomonas was driven extinct. In the single-prey cultures, Tetrahymena sustained four times the number of predators as did Chilomonas cultures. In the mixed cultures, the latter species suffered higher predation than it could sustain. The limiting factor is the abundance of a shared predator, which can respond numerically to its prey. Let P* be the abundance of the predator sustained (and tolerated) by a prey species. A "P* rule" now describes dominance: The winning prey species is the one with greater P*. This form of indirect competition between species arising from shared natural enemies (including parasites) is called apparent competition (Holt and Lawton, 1994); the word 'apparent' is used because the interaction has the same consequences for coexistence as does classical exploitative competition for resources but may occur even between species with totally different resource requirements. Dominance in apparent competition may occur because of different vulnerabilities to a natural enemy (as in the caribou example) or because one prey is highly productive and sustains an abundant enemy population (as in Lawler's experiment).

V. MECHANISMS OF COEXISTENCE

The previous discussion emphasized constraints on species coexistence arising from interspecific interactions. Rules of dominance are important conceptual tools which quantify these constraints and help identify biological traits leading to dominance. However, even in simple microcosms coexistence can occur, and most
natural communities are rich in species. To understand coexistence, the competitive exclusion principle is reconsidered (Crawley, 1997; Grover, 1997; Tokeshi, 1999). Species may coexist when any of the assumptions leading to the competitive exclusion principle are violated. This suggests three classes of mechanisms promoting species coexistence of potentially competing species in a local community:

1. Species may coexist in a closed, temporally constant world if they experience different limiting factors at the spatial scale of the local community; this includes classical niche partitioning of resources, as well as mechanisms involving predation and parasitism, and direct interference.

2. Species may coexist, even though they experience the same limiting factor, if the environment is temporally variable and species respond differently to this temporal variation (temporal niche partitioning).

3. Species may coexist if the environment is spatially open; this includes spatial niche partitioning at scales broader than the local community and mechanisms such as colonization–competition tradeoffs in metapopulations.

From the 1950s to the mid-1970s, stimulated largely by G. E. Hutchinson and his brilliant student Robert MacArthur, most community ecologists emphasized classical niche partitioning in studies of species coexistence. In recent years, the balance of attention has shifted markedly to a broader range of coexistence mechanisms. Ecologists now believe that maintenance of diversity—coexistence writ large—often depends on spatial dynamics in open communities, food web interactions (including predation and parasitism), and nonequilibrium dynamics reflecting either extrinsic temporal variation or the endogenous instability of complex ecological system.

VI. TRADITIONAL APPROACHES TO COEXISTENCE

A. Classical Niche Partitioning

Competitive exclusion is expected if growth rates of two (or more) species are determined by a single limiting factor. Species may coexist, even in an unvarying and spatially confined bottle, given multiple limiting factors, such that each species is limited more strongly by its own distinct set of factors. When species coexist, one sensible approach to begin to understand this coexistence is to map their niche requirements against the spectrum of limiting factors present in the environment. This basic methodology has been a source of significant insights for much of the history of ecology and continues to be fruitful. The classic paper in this genre is the 1958 study by Robert MacArthur of wood warblers in a New England boreal forest. Five warbler species in the genus Dendroica occurred in the same tract of forest; all five species have similar body sizes and eat the same range of insect taxa. MacArthur found that the species segregated with respect to microhabitat (with one species feeding at tree-top, another in low branches, and so on). He argued that because of this microhabitat partitioning, each species consumed an independent pool of insect prey, thereby reducing the potential for competitive exclusion. Following publication of this study, there was a proliferation of field studies of niche partitioning patterns.

B. Lotka–Volterra Competition Model

MacArthur’s study was motivated in part by Gause’s experiments on competitive exclusion, which in turn were stimulated by theoretical models of interacting species explored by the Italian mathematician Vito Volterra. In the usual textbook formulation (the famous Lotka–Volterra equations), the competitive effect of one species on another is expressed as a direct, density-dependent reduction in abundance:

$$\frac{dN_1}{dt} = r_1N_1(K_1 - N_1 - a_1N_2/K_2);$$

where $N_1$ and $N_2$ are the abundance of competing species 1 and 2, respectively, $K_1$ is the carrying capacity of species 1, $r_1$ is its intrinsic rate of increase, and $a_1$ is a competition coefficient (the equation for species 2 is the same, with the subscripts 1 and 2 switched). The quantity $a_1$ measures the effect an individual of species 2 has on reducing the per capita growth rate of species 1 compared to the effect of an individual of 1 on its own species. Both species increase when rare and hence coexist if

$$\frac{1}{a_1} > \frac{K_2}{K_1}.$$

The outer inequalities imply $1 > a_1R_1$. This necessary condition for coexistence states that one (or both) of the competing species experiences stronger intraspecific than interspecific competition. If the inequalities are reversed, either species can exclude the other if it is initially sufficiently abundant (as in Park’s experiments with Tribolium). If the competition coefficients are near
unity (i.e., density dependence occurs uniformly within and between species), the species with higher carrying capacity wins.

The Lotka–Volterra model usefully describes competition, and multispecies extensions to it have been the focus of a rich body of theoretical work. However, the model is difficult to use predictively because competition coefficients and direct estimates of density dependence can only be measured during competitive trials rather than a priori. Moreover, taken literally, the model best describes systems in which competition is due to direct interference. Recognizing these limitations led to the development of a wide range of mechanistic models of resource–consumer and other interspecific interactions beginning in the late 1960s.

C. Whither Limiting Similarity?

This development was concordant with interest in the important concept of “limiting similarity,” which is the notion that there might be a quantifiable limit to how similar species could be in their utilization of resources and still coexist. [To gain an understanding of the idea of limiting similarity, note that in Eq. (2) similar species should have competition coefficients near 1, so coexistence will not occur unless carrying capacities are finely balanced.] For a period in the history of the discipline, the goal of quantifying limiting similarity seemed to permit predictions of the maximal number of species which could persist on a defined resource base, providing a basic tool for understanding biodiversity. These studies were based on the Lotka–Volterra model coupled with assumptions about resource use. In some limiting cases, the Lotka–Volterra model emerges as a reasonable approximation of more complex resource–consumer interactions, and one can directly map niche overlap onto measures of competition. With simplifying assumptions about symmetry in resource use, the role of environmental variation, and other factors, a limiting similarity of competing species could be calculated and then compared against observed similarities. For instance, using MacArthur's warbler data, it was assumed that spatial overlap in foraging was directly related to competition coefficients. Substitution into the model showed that observed overlap was consistent with long-term coexistence. More broadly, if niche overlap were proved to be closely related to the strength of competition, observational studies of overlap and niche partitioning would provide a powerful link between descriptive analyses of community patterns and the dynamical forces of species interactions.

Further exploration has tempered the initial flush of enthusiasm for this approach to coexistence. After considerable grinding of mathematical gears, theoretical ecologists have concluded that limiting similarity, if it occurs at all, can only be characterized for models tailored to specific situations. Instead of a single general theory of limiting similarity and coexistence, there are many special theories. Moreover, there is a real sense in which increasing species similarity can facilitate, rather than hamper, coexistence. Equation (2) describes bounds on the permissible differences in the two species' carrying capacities. If the $a_i$ are less than unity, coexistence is more likely if carrying capacities are nearly equal rather than very different. The more similar species are in their demographic responses to the environment, the more similar their carrying capacities are likely to be. All else being equal, this kind of similarity promotes the coexistence of competing species.

D. Manipulative Field Experiments

Recognizing the limitation of observational studies of niche overlap and partitioning as evidence for competition, ecologists have turned to manipulative field experiments, typically removing one species and monitoring changes in the abundance of others. Reviews of such experiments show that when species are suspected to compete (e.g., because of overlap in resource requirements and habitat use), they often in fact do compete. For instance, Hairston noted that in the Great Smoky Mountains two species of Plethodon salamanders had altitudinal ranges which were nearly mutually exclusive, with only a narrow range of overlap. In contrast, in the Balsam Mountains altitudinal overlap was extensive. Hairston hypothesized that this was due to stronger competition in the Smokies, and using reciprocal transplants and removals he demonstrated that competition (due to behavioral aggression) was indeed much stronger in the narrow overlap zone in the Smokies.

Unfortunately, relatively few field experiments have been directly tied to mechanistic models of competition, which makes it difficult to generalize among studies. Moreover, such experiments tend to focus on species which already coexist, at least to a degree; removals then assess the magnitude of competition, given coexistence, rather than coexistence or exclusion per se. One study which directly addressed coexistence was performed by Bengtsson, who studied three species in the zooplankton genus Daphnia (which compete exploitatively for algal food) on the coast of the Baltic Sea.
Observational studies suggested that usually just one, or more infrequently two, species was present in any given pool. Bengtsson added all possible combinations of the three species to artificial pools and found that there were no extinctions in the single-species pools. Yet high extinction rates occurred in the two- and three-species sets. This directly demonstrates the importance of competition in constraining coexistence in natural conditions.

VII. MECHANISMS FOR LOCAL COEXISTENCE: CURRENT APPROACHES

A. Mechanistic Models of Multiple Limiting Factors

Most theoretical studies of competition today focus on models in which the mechanisms of interaction are clearly described. Detailed analyses of mechanistic models of competition are often mathematically challenging, but important insights can often be gleaned from simple graphical analyses. A generalization of the competitive exclusion principle is that "coexistence of \( n \) species requires \( n \) limiting factors." Mathematically, one expresses the growth rate of each species as a function of a vector of limiting factors (e.g., \( \frac{dN_i}{dt} = N_i(E_1, E_2, \ldots) \)). For illustrative purposes, I discuss one example in detail (Tilman, 1982; Grover, 1997; Pacala as cited in Crawley, 1997).

Consider two species competing for two limiting resources. Assume exploitation depresses resources and that resource–consumer dynamics tend toward a stable equilibrium. For each species, there will be some combination of resources that allows equilibrium (with births matching deaths). On a graph with axes of resource abundance, assume this combination can be portrayed as a straight line, as in the line marked \( 1 \) in Fig. 1a (for species 1) (MacArthur, 1972). This line is the "zero-growth isocline" of species 1. The linear form of the isocline implies that resources are qualitatively substitutable so that a sufficient supply of one compensates for low abundances in the other. [Tilman (1982) examines exploitative competition for a broader range of resource types.] At equilibrium, resource abundances should be along this line; if resources lie outside this line, species 1 should increase, depressing resource levels (with reverse dynamics inside the line). An equilibrium with coexistence requires resource levels at which both consumers have zero growth; graphically, the isoclines must cross. The isoclines do not intersect (Fig. 1a) species 1 has a lower \( R^* \) for each resource. If species 1 is resident and at equilibrium, and a few individuals of species 2 are introduced, the introduction will fail due to competitive exclusion. If isoclines cross (Fig. 1b) species 1 has a lower \( R^* \) for resource 1 than does species 2, and the converse is true for resource 2. Coexistence is now possible.

Crossing isoclines reflect differences in species' ecologies and the existence of two distinct limiting factors (here, linear combinations of resources). However, such niche differences do not suffice for coexistence. As in the Lotka–Volterra model, there must be broad similarities in how the two species respond to the environment. In Fig. 1a, exclusion occurs because the species are too different in their overall requirements for resources. Coexistence, in contrast, is permitted because overall resource requirements are approximately the same for the two species (Fig. 1b).

Having isoclines that cross does not guarantee coexistence. If species 1 is alone, and resources equilibrate at point \( a \), species 2 invades. However, if resources instead equilibrate at point \( b \), species 2 is excluded. If species 2 is resident, species 1 invades if resources are at point \( c \) but not at point \( d \). Comparing these points to each species' \( R^* \) (for each resource) suggests a necessary requirement for coexistence: Given two resources, each competitor must have the greater impact on that resource for which it has the lower \( R^* \). In effect, each species must limit itself (via resource consumption) more strongly than it limits the other species. Whether or not this occurs depends on both the intrinsic renewal rates of the resources and the rate of consumption of each resource. (A complete analysis of the conditions for coexistence requires a model with equations for dynamics of both consumers and each resource.)
This graphical model illustrates the important insight that coexistence depends on a balancing of overall similarities and differences in species' niche requirements, as well as differences in species' impacts on their environments. The former involves species' intrinsic properties, whereas the latter depends on the system in which the interaction is embedded, including ecosystem processes (e.g., resource renewal rates).

### B. Food Web Effects on Coexistence

This general approach can be extended in many ways, for instance, by including interactions among species at multiple trophic levels. With predation, herbivory, or parasitism inflicted on competing species, one greatly increases the number of potential limiting factors. This is a large and complex topic, and I briefly discuss some highlights here.

Specialist predators and parasites typically reduce the abundance of their favored species, freeing up resources for other nontarget species. This can facilitate coexistence if dominant competitors tend to attract more specialist predators or parasites than do subordinates. Generalist predators can have the same effect via differential fixed preferences for dominant competitors or if they reduce attacks on whichever species is temporarily rarest. However, predators or pathogens which attack two or more species can also hamper coexistence (as in the examples discussed previously). This is particularly likely if the prey species under attack do not strongly compete, and if some prey species are sufficiently productive to sustain high abundances of the predator or parasite (Holt and Lawton, 1994).

Whenever species interact via multiple mechanisms (providing different limiting factors), coexistence may occur. For instance, if two species compete for a single limiting resource, the species with higher $R^*$ may nevertheless persist if it can also consume the superior competitor. This mixture of predation and competition is called intraguild predation.

### C. Local Habitat Heterogeneity

Traditional ecological models of interacting species such as Eq. (1) assume that populations are spatially well mixed and average over local environmental variation. Relaxing this assumption often promotes coexistence. As shown in Gauff's experiments, even in the confines of a microcosm there can be spatial heterogeneity in abiotic conditions that influences coexistence. Assume that within the microcosm each species has a set of conditions in which it is superior. If competitive interactions are sufficiently localized (e.g., because of limited movement or habitat selection), and each species spends more time in the microhabitat in which it is superior, one can readily generate coexistence. This simple mechanism for coexistence via local habitat partitioning is very important. A review by Schoener (1974) of resource partitioning studies following the footsteps of MacArthur’s warbler study revealed the ubiquity of habitat differences among potential competitors; subsequent years have not altered this basic message.

Competition for light is clearly important in plant communities, but it is not well understood. There is an asymmetry in that tall plants shade small plants but not the reverse. Because light absorption is imperfect, some light penetrates any canopy. Theoretical studies of light competition suggest that two (or more) plant species can coexist, given appropriate tradeoffs between size and the ability to use light at various levels. The physical structure of plants absorbing a directional flux of light almost inevitably leads to local gradients in light availability and quality (due to differential absorption of different wavelengths), which provides the opportunity for plant niche differentiation and coexistence.

In animals, habitat selection can be an important mechanism promoting coexistence with local spatial heterogeneity. If individuals of a species (when rare) can discriminate among local microhabitats and spend more time in those which provide the greatest fitness rewards, the rate of increase at low $N$ will be increased, relative to that of a species which utilizes habitats at random. If local habitat heterogeneity permits each species to have a habitat in which it is competitively dominant, habitat selection will sharpen habitat partitioning and thereby make coexistence more likely. Habitat selection can permit subordinate species to withstand superior competitors. Recent studies in east Africa suggest that the fleetness of the cheetah helps it persist in the face of direct aggression by lions and hyenas (who kill cheetah cubs) by allowing it to seek out areas with low lion and hyena densities.

### D. Consequences of Individual Discreteness

Traditional ecological models also assume that populations are sufficiently large that abundances can be treated as continuous variables. The growth in computing power in recent decades has stimulated ecologists to analyze population dynamics by tracking the fates of individuals. This reveals novel mechanisms of coexistence. For many organisms (particularly terrestrial
plants), an individual occupies a small site and interacts with only a few neighbors during its lifetime. Probabilistic events of individual life histories (birth, death, and dispersal) lead to spatial variance in competition, even among sites with homogeneous physical conditions. Moreover, dispersal typically occurs over small spatial scales. Combining these two general facts together in models leads to nonuniform spatial patterns in which competition is typically stronger within species (because they tend to occur in clumps) than between species (which tend to become segregated spatially) (Pacala as cited in Crawley, 1997; Pacala and Levin as cited in Tilman and Kareiva, 1997). These effects are particularly strong for species that are relatively similar to each other. This mechanism may go a long way toward explaining the puzzling coexistence of large numbers of similar coexisting species in many plant communities.

VIII. TEMPORAL NICHE PARTITIONING

Outside the controlled confines of the lab, environments are rarely constant. Temporal variation occurs on scales ranging from diurnal cycles to climatic changes over millennia. Many ecologists have intuitively argued that temporal variability by itself weakens negative interspecific interactions and thereby facilitates coexistence. Theoretical studies (reviewed in Chesson and Huntly, 1997) have conclusively shown that this is not the case. Instead, temporal variation can provide a rich arena for differentiation among species in responses to the environments, which can promote coexistence. The basic idea is that if species A does better when the local environment is in state A' and species B does better when the local environment is in state B', the environment alternates between A' and B', coexistence is possible. Two plausible scenarios are presented in the following sections.

A. Nonlinear Consumption

Assume two consumer species exploit a single limiting resource and have saturating, nonlinear relationships between feeding rates and resource availability. Species A has a higher rate at low resource levels, whereas species B enjoys a higher feeding rate at high levels. In a constant environment the resource level will equilibrate at a constant level, and the species with lower R* will win. However, because of large-scale seasonal variation in resource supply rates, resource levels may fluctuate between high and low levels. Such variation can permit coexistence of these two competitors on a single resource. This mechanism hinges crucially on nonlinearity in species’ responses to the shared limiting factor; if feeding rate were to increase linearly with resource availability in both species, then even in a variable environment one would observe competitive exclusion.

B. The Storage Effect

In a desert, rain falls sporadically and at different times in different years, sometimes early in the spring when the air is cool and at other times later in the hot summer. At each rain, a mixture of plant species appear in abundance from the resting seed bank and compete for water and soil nutrients. These plants are short-lived as adults, often completing their entire aboveground life cycle in just a few weeks. Some species do disproportionately better following warm rains, whereas others do better following cold rains. The seeds each species produces enter the seed bank and germinate gradually during the following years (like a time-release capsule of past recruitment). The "storage" of good years of recruitment in a long-lived seed bank can facilitate coexistence of competitors. Again, it is not variability alone which allows coexistence. One also needs (i) niche partitioning, with different species being superior at different times; (ii) nonlinear effects of the environment on responses to competition; and (iii) correlations between the varying environmental factors and the strength of competition (Chesson and Huntly, 1997). The existence of distinct life stages (e.g., long-lived seeds) can induce the needed nonlinearity in responses to temporal variation that promotes the coexistence of competing species.

IX. COEXISTENCE IN OPEN COMMUNITIES

In contrast to laboratory microcosms, many natural communities are open, coupled to the external environment via dispersal. Such coupling influences local coexistence of species in many distinct ways.

A. Autecology and Population Size Effects

A consumer species which relies on a sparse or sporadic resource, or extracts different essential resources in different habitats, may need to be mobile to persist. If absolute population size is small, extinction is risked even in favorable environments. Specialist consumers are likely to go extinct if their resources are rare. All
these problems are aggravated when there are barriers to dispersal and habitable area is small. These considerations may help explain why food chains are often short on oceanic islands or isolated habitat patches. Moreover, as one increases the number of species which coexist deterministically on a fixed resource base, the abundance per species typically declines. Small population sizes have an increased risk of extinction; this can put a loose limit on coexistence.

B. Colonization–Extinction Dynamics

It was previously mentioned that strong specialist predator–prey interactions are prone to local extinctions. Coexistence may require dispersal among habitats, with prey dispersing sufficiently fast so that they can find and reproduce within empty habitat patches before being discovered by predators. Similarly, if a guild of competitors utilizes a single resource, species with successively lower values of $R^*$ should eventually colonize and displace species with higher $R^*$. If there are extinctions, asynchronous among patches, inferior competitors may have a temporary window of opportunity during which they can occupy habitat patches and reproduce sufficiently to colonize other patches. This metapopulation mechanism for coexistence can be promoted by a tradeoff between competitive ability and colonization ability, and in principle it can promote the coexistence of many competing species (Tilman and Lehmann as cited in Tilman and Kareiva, 1997). Even if the inferior competitor is not a superior colonizer, it may be able to persist if it has a lower basal extinction rate when it occurs in patches alone compared to that of the superior competitor.

C. Landscape Heterogeneity

1. Spillover Effects

If a species can persist in one local community, then with emigration it can also be found in other nearby communities in which otherwise it would be excluded by local interactions. To model this “spillover” effect, assume the species when rate declines at rate $r < 0$ because of competitive exclusion but immigrates from an external source community at rate $I$. Its local dynamics are thus described by $dN/dt = I + rN$, which implies $N^* = I/r$ at equilibrium. A species which should be absent, considering only local interactions, may not just persist but even be abundant if (i) there is a large rate of immigration (e.g., from a productive source habitat into an unproductive habitat) and (ii) the rate of exclusion is slow. In this scenario, the answer to the species coexistence problem ultimately requires analyzing population dynamics at the appropriate spatial scale, larger than the local community, spatial niche partitioning at this larger scale is responsible for coexistence.

2. Landscape Mechanisms of Exclusion

Spatial coupling can also generate novel mechanisms for exclusion. For instance, a predator sustained by a prey species in a productive habitat may move through another habitat and inflict mortality at a sufficiently high rate to exclude a prey species there. Likewise, a species which is a superior competitor in one habitat may be excluded by a species which is inferior there but sufficiently abundant elsewhere so that via dispersal it can “swamp” the local habitat. For instance, Ted Case and associates recently described impacts of an invasive ant species, the Argentine ant, on an entire community of ants in coastal southern California. In contrast to many ants, the Argentine ant shows little intercolonial hostility and so has a high carrying capacity; it competes exploitatively for food but also preys on the juvenile life stages of some ant species. The Argentine ant readily becomes established in disturbed habitats. It can then spread into fragments of natural habitat resulting from intense development. Case suggests that in these fragments many other ant species have declined to the point of extinction because of the sheer force of numbers. These effects are particularly dramatic in small fragments, and at the edges of large fragments, because of incursions from surrounding human-disturbed habitats in which the Argentine ant is abundant.

I now return to the lab and field case studies of competition discussed previously. In light of what we now know about mechanisms of coexistence, it is easy to see why competitive exclusion was often observed in the classic bottle experiments with protozoa and grain beetles. The design of these experiments included most of the elements assumed in the syllogism leading to the competitive exclusion principle. The lab environments were climatically controlled, and the culture bottles were spatially closed, precluding two of the major classes of coexistence mechanisms. The culture media and conditions (e.g., stirring) were set up to have a restricted number of limiting resources (ideally, one) and little within-culture spatial heterogeneity in limiting food resources, the communities are very simple, so there is little opportunity for complex food web interactions. The most interesting result of these experiments may be that coexistence is observed at all. In the field study of introduced parasitoids, these species were specialized to the same host species, and the environment was relatively constant in time and homogeneous.
in space (the lowlands of tropical islands such as Hawaii have very stable climates, and the habitat is a deliberately homogenized landscape—plantations with regularly spaced fruit trees). It is likely that the communities are simple (islands tend to be low in species richness, particularly in heavily disturbed habitats dominated by introduced species). Most natural systems, in contrast, are temporally variable and spatially heterogeneous, and they harbor rich, complex communities; therefore, it should not be surprising that it is often difficult to document competitive exclusion in the wild.

X. FUTURE DIRECTIONS

There are many important themes relevant to species coexistence which have not been addressed in-depth by ecologists but are likely to receive considerable attention in the near future.

A. Transient Coexistence

Ecological theory has traditionally emphasized equilibrium community states and the development of criteria for exclusion and indefinite persistence. However, exclusion takes place over some timescale. If this is long (e.g., relative to climate change), communities may exhibit transient dynamics and be far from equilibrium. This is particularly likely when considering interactions among species which are very similar in their niche requirements and environmental effects; their abundances can vary through time in an essentially random fashion (a process known as community “drift,” championed by Steve Hubbell). An important task for future ecological theory and empirical work is to derive a deeper understanding of transient dynamics and the drift hypothesis.

B. Allee Effects and Coexistence

Most ecological theory applies literally only to clonal, asexual organisms. However, most species of concern to ecologists are sexual. At very low population sizes, outcrossed sexual species should have depressed growth rates simply because the two sexes have to get together to reproduce—an “Allee” effect (positive density dependence in growth rates at low densities). Allee effects can influence coexistence. For instance, consider two sexual species with identical ecologies in a landscape with many habitat patches. If colonization is infrequent, whichever species arrives first in a patch should be able to dominate there because the first colonist increases until births just match deaths. When the second species appears, its potential birth rate will then also just match its death rate (because its ecology is the same as that of the resident), but its realized birth rate will be depressed because it has difficulties in pair formation at low numbers. Hence, in this patch the second species to arrive will be excluded. In the landscape as a whole, the two species may coexist if by chance each species is the first to occupy a subset of patches. In some patches, both species may by chance appear initially in sufficient numbers so as not to experience the Allee effect; in these patches, abundances should drift through time. Over the entire landscape, patterns in abundance will not be correlated with any discernible factors in the external environment. The magnitude of the Allee effect varies greatly among species, depending on details of the mating system and mate-finding strategies (e.g., selfing in plants should greatly weaken Allee effects). This influence of mating ecology on coexistence may provide an as-yet poorly understood source of variation among taxa in community structure.

C. Speciation Mode and Coexistence

Over long timescales, all species originate from other species. Understanding the mode of speciation may provide insight into coexistence. For instance, sympatric speciation occurs when a lineage diverges within a single community. Most models of sympatric speciation depend on substantial ecological differentiation being present from the beginning, as speciation unfolds, ecological differences between the two emerging species must be sufficient to withstand cross-mating. In other words, with this mode of speciation, mechanisms of coexistence (e.g., substantial habitat or resource segregation) are built into the very branching pattern of the phylogenetic tree. In contrast, allopatric speciation requires geographical isolation. This can occur without any ecological difference arising between the daughter lineages (although speciation should occur more rapidly if correlated with ecological divergence). Dispersal can later bring the daughter species together, with essentially any degree of ecological difference being possible at this stage. Speciation may also often involve changes in sexual selection and mating systems, which can occur with no change in ecological requirements. After speciation, when these species’ ranges begin to overlap, relative abundances should be particularly prone to community drift.

XI. CONCLUSIONS

In conclusion, understanding the factors which promote or constrain the coexistence of species is an ongoing en-
enterprise at the intellectual core of the study of biodiversity. Community ecologists have a rich smorgasbord of hypotheses to explain both species coexistence and exclusion. In most natural communities it is likely that many mechanisms operate simultaneously. A challenging problem is to ascertain the relative contribution of these ecological mechanisms of coexistence to explain major patterns in biodiversity in space and time. There is much work to be done in analyzing mechanisms of coexistence in the context of food web dynamics and metapopulations. Moreover, very little mention was made in this article about commensalism, mutualism, and non trophic interactions, all of which can be crucial to coexistence. An improved understanding of the factors governing species coexistence is needed to address many applied problems, particularly the conservation of natural communities. Given the importance of spatial openness for coexistence, the anthropogenic alteration and simplification of landscapes pose particular dangers to the continued coexistence of many species. Indeed, the basic unstated problem of conservation is how to structure our activities and modify our impacts so that most of the world’s biota can manage to coexist (at some spatial scale and, it is hoped, over long timescales) with just a single species, namely, ourselves.

Acknowledgments

I thank many friends for conversations during the years about species coexistence, particularly Mark McPeek, Peter Chesson, John Lawton, David Tilman, Peter Abrams, Peter Morin, and James Grover.

See Also the Following Articles

CoEvolution • Food Webs • Mutualism, Evolution of • Introduced Species, Effect and Distribution • Landscape Diversity • Parasitism • Predators, Ecological Role of • Population Dynamics • Scale, Concept and Effects of

Bibliography


Crawley, M. J. (1997). Plant Ecology, 2nd ed. Blackwell, Oxford. [See Chapters 8 (Tilman), 14 (Crawley), 15 (Pacala), and 16 (Grover and Holt)].


SPECIES, CONCEPTS OF

James Mallet
University College London

GLOSSARY

cladistic A classification based entirely on monophyletic taxonomic groupings within a phylogeny; taxonomic units that are paraphyletic or polyphyletic are rejected. A cladist is one who practices cladistics, usually in the sense of using parsimony to adjudicate between data from multiple characters in the construction of a cladogram, which is an estimate of the true phylogeny.

cohesion The sum total of forces or systems that hold a species together. The term is used especially in the interbreeding and cohesion species concepts. Cohesion mechanisms include isolating mechanisms in sexual species as well as stabilizing ecological selection, which may cause cohesion even within asexual lineages.

disruptive selection Selection acting to preserve extreme phenotypes in a population. Speciation usually involves disruptive selection, because intermediates (hybrids between incipient species) are disfavored (see also stabilizing selection).

gene flow Movement of genes between populations, usually via immigration and mating of whole genotypes, but sometimes single genes may undergo horizontal gene transfer via transfection by microorganisms.

gene pool The sum total of the genetic variation within a reproductively isolated species population; this term is mostly used by supporters of the interbreeding species concept.

isolating mechanisms The sum total of all types of factors that prevent gene flow between species, including premating mechanisms (mate choice) and postmating mechanisms (hybrid sterility and inviability). Modern authors deny that these “mechanisms” have necessarily evolved to preserve the species’ integrity as originally assumed, though this may sometimes be the case in reinforcement of premating isolation. Isolating mechanisms are a subset of the factors that cause cohesion of species under the interbreeding and cohesion species concepts.

monophyletic A grouping that contains all of the descendants of a particular node in a phylogeny. Monophyly is the state of such groupings. Compare paraphyletic and polyphyletic. Butterflies (Rhopalocera) and birds (Aves) are examples of two groups thought to be monophyletic.
paraphyletic: A grouping that contains some, but not all, of the descendants of a particular node in a phylogeny. Paraphyly is the state of such groupings. Compare monophyletic and polyphyletic. Moths (Lepidoptera, excluding butterflies) and reptiles (amniotes, excluding birds and mammals) are examples of two groups thought to be paraphyletic.

phenetic: A classification or grouping based purely on overall similarity. Pheneticists use matrices of overall similarity rather than parsimony to construct a phenogram as an estimate of the phylogeny. Examples of phenetic methods of estimation include unweighted pair group analysis (UPGMA) and neighbor joining. Cladists reject phenetic classifications on the grounds that they may result in paraphyletic or polyphyletic groupings.

phylogenetic: Pertaining to the true (i.e., evolutionary) pattern of relationship, usually expressed in the form of a binary branching tree, or phylogeny. If hybridization produces new lineages, as is common in many plants and some animals, the phylogeny is said to be “reticulate.” Phylogenies may be estimated using phenetics, parsimony (cladistics), or methods based on statistical likelihood.

polyphyletic: Groupings contain taxa with more than one ancestor. Polyphyly is the state of such groupings. Compare paraphyletic and monophyletic. “Winged vertebrates” (including birds and bats) give an example of a polyphyletic group.

real, reality: Two tricky words found frequently in the species concept debate. “Reality” is typically used to support one’s own species concept, as in: “The conclusions set forth above . . . lead to a belief in the reality of species” (Poulton, 1904); similar examples can be found in the writings of Dobzhansky, Mayr, and especially phylogeneticists. The term reality in this sense is similar to an Aristotelian “essence,” a hypothetical pure, albeit obscure, truth that underlies the messy actuality; unfortunately, in everyday language “real” also means “actual” (curiously, a “reality” in the first sense may be “unreal” under the second?). By rejecting the “reality” of species, one can therefore send very mixed messages: some readers will understand the author to be a nominalist who merely believes useful terms require little theoretical underpinning; others assume the author is nonsensically using some definition that does not apply to actual organisms. Here, when I discuss the reality underlying a species concept, I mean it in the sense of a hypothesized truth. Many authors of species concepts and some philosophers of science argue that definitions must be underpinned by a theoretical justification or reality. Other philosophers such as Wittgenstein and Popper agree that terms need no such definition to be useful.

speciation: The evolutionary process of the origin of a new species.

specific mate recognition systems (SMRS): Fertilization and mate recognition systems in the recognition concept of species, the factors leading to premating compatibility within a species. See also cohesion, which is similar to SMRS, but includes postmating compatibility as well.

sibling species: A pair of closely related, morphologically similar species (usually sister species).

stabilizing selection: Selection that favors intermediate phenotypes.

SPECIES ARE CRUCIAL: In many biodiversity issues: much of conservation, biodiversity studies, ecology, and legislation concerns this taxonomic level. It may therefore seem rather surprising that biologists have failed to agree on a single species concept. The disagreement means that species counts could easily differ by an order of magnitude or more when the same data are examined by different taxonomies. This article explores the controversy on species concepts and its implications for evolution and conservation.

I. WHAT ARE SPECIES CONCEPTS FOR?

Individual organisms can usually be recognized, but the larger units we use to describe the diversity of life, such as populations, subspecies, or species are not so easily identifiable. Taxonomists further group species into genera, families, orders, kingdoms, etc., while ecologists group species into higher structures such as communities and ecosystems. The justification for these group terms is utility, rather than intrinsic naturalness, but as far as possible we attempt to delimit groups of organisms along natural fault lines, so that approximately the same groupings can be recovered by independent observers. However there will be a virtually infinite number of different, albeit nested, ways of classifying the same organisms, given that life has evolved hierarchically.

Darwin (1859) felt that species were similar in kind to groupings at lower and higher taxonomic levels; in contrast, most recent authors suggest that species are
more objectively identifiable, and thus more “real” than, say, populations or genera. Much of ecology and biodiversity today appears to depend on the idea that the species is the fundamental taxon, and these fields could be undermined if, say, genera, or subspecies had the same logical status.

Species concepts originate in taxonomy, where the species is “the basic rank of classification” according to the International Commission of Zoological Nomenclature. The main use of species in taxonomy and derivative sciences is to order and retrieve information on individual specimens in collections or data banks. In evolution, we would like to delimit a particular kind of evolution, speciation, which produces a result qualitatively different from within-population evolution, although it may of course involve the same processes. In ecology, the species is a group of individuals within which variation can often be ignored for the purposes of studying local populations or communities, so that species can compete, for example, while subspecies or genera are not usually considered in this light. In biodiversity and conservation studies, and in environmental legislation, species are important as units which we would like to be able to count both regionally and globally.

It would be enormously helpful if a single definition of species could satisfy all these uses, but a generally accepted definition has yet to be found, and indeed is believed by some to be an impossibility. A unitary definition should be possible, however, if species are “real,” objectively definable, and fundamental biological units. Conversely, even if species have no greater objectivity than other taxa, unitary nominalistic guidelines for a definition could still be found, perhaps after much diplomacy, via international agreement among biologists; after all, if we can adopt meters and kilograms, we should be able to agree on units of biodiversity in a similar way. In either case, knowledge of the full gamut of today’s competing solutions to the species concept problem will probably be necessary for a universal species definition to be found. This article reviews the proposals currently on the table and their usefulness in ecology, evolution, and conservation.

II. STATEMENT OF BIAS

I am of the opinion that the “reality” of species in evolution and in ecological and biodiversity studies over large areas has been overestimated. In contrast, it is clear to any naturalist that species are usually somewhat objectively definable in local communities. It is my belief that confusion over species concepts has been caused by scientists not only attempting to extend this local objectivity of species over space and evolutionary time but also arguing fruitlessly among themselves as to the nature of the important reality that underlies this illusory spatiotemporal objectivity. To me, agreement on a unified species-level taxonomy is possible, but will be forthcoming only if we accept that species lack a single, interpretable biological reality over their geographic range and across geological time.

Just as Marxist theory may be wrong, yet remains a convenient tool for studying political history, I hope that my own views can provide, even for the skeptic, a useful framework on which the history of proposals for species concepts can be compared. A variety of other outlooks can be found in Mayr (1982), Cracraft (1989), Ridley (1996), Claridge et al. (1997), and Howard and Berlocher (1998).

III. DARWINIAN SPECIES CRITERIA

A. Darwin’s Morphological Species Criterion

Before Darwin, it was often assumed that each species had an Aristotelian “form” or “essence” and that variation within a species was due to imperfections in the actualization of this form. Each individual species was defined by its essence, which itself was unvarying and inherently different from all other species essences. This mode of thought of course precluded transformation of one species into another and was associated with belief that each form was separately created by God. Darwin’s extensive travels and knowledge of taxonomy led to a realization that the distinction between intraspecific and interspecific variation was false. His abandonment of the essentialist philosophy and its species concept went hand in hand with his appreciation that variation itself was among the most important characteristics of living organisms, because it was this variation which allowed species to evolve.

Darwin guessed (correctly) that essentialist species would be hard to give up: “...we shall have to treat species in the same manner as those naturalists treat genera, who admit that genera are merely artificial combinations made for convenience. This may not be a cheering prospect, but we shall at least be freed from the vain search for the undiscovered and undiscoverable essence of the term species” (Darwin, 1859, p. 483). He argued that species were little more than varieties which acquired their claim to a greater reality only
when intermediates died out leaving a morphological gap. ... I believe that species come to be tolerably well-defined objects, and do not as any one period present an inextricable chaos of varying and intermediate links” (Darwin, 1859, p. 177). This morphological gap criterion, which seems to have been accepted by most early evolutionists (e.g., Wallace, 1865; Robson, 1928), has been called a “morphological species concept” because Darwin used the gaps in morphology to delimit species; however, it would be easy to extend his species criterion to ecology, behavior, or genetics (see Section VIII.D).

B. Polytypic Species

A major revolution in zoological taxonomy occurred around 1900. As the great museum collections became more complete, it became obvious that apparently distinct “species” found in different areas frequently intergraded where they overlapped. These replacement series were usually combined as subspecies within a “polytypic” species (see Subspecies, Semispecies, Superspecies), an idea suggested for “geographical varieties” by early systematists and Darwinists such as Wallace (1865). The taxonomic clarification that followed, which allowed identifiable geographic varieties to be named below the species level as subspecies, was conceptually more or less complete by the 1920s and 1930s. At the same time, other infraspecific animal taxa such as local varieties or forms were deemed unnameable in the Linnean taxonomy. These changes are now incorporated into the International Code of Zoological Nomenclature (see also Subspecies, Semispecies, Superspecies). Similar ideas were promoted in botany by G. L. Stebbins (see Mayr, 1982), although local varieties and polymorphic forms remain valid and nameable taxa in the International Code of Botanical Nomenclature.

IV. THE PHILOSOPHIZATION OF SPECIES: THE INTERBREEDING SPECIES CONCEPT

In January 1904, E. B. Poulton read his famous presidential address—“What is a species?”—to the Entomological Society in London. Following up some ideas raised (but immediately dismissed) by Wallace (1865), Poulton proposed “syngamy” (i.e., interbreeding) as the true meaning of species. Poulton and Wallace were both particularly knowledgeable about swallowtail butterflies (Papilionidae). In swallowtails, there are strong sexual dimorphisms: the female color pattern often mimics unrelated unpalatable butterflies while the male is nonmimetic. The females themselves are often polymorphic, each female form mimicking a different distasteful model. Under a morphological criterion each form could be designated as a different species, whereas mating observations in the wild showed that they were part of the same interbreeding group. Similar ideas were promoted by the botanist J. P. Lotsy, who termed the interbreeding species a “syngameon.” In the 1930s, T. Dobzhansky studied morphologically indistinguishable sibling species of Drosophila fruit flies and concluded that Lotsy’s approach had some value. A species will rarely if ever interbreed with its sibling; each chooses mates from within its own species. Dobzhansky proposed his own interbreeding species concept, later popularized by Mayr as the “biological species concept,” so named because interbreeding was considered the single true biological meaning or reality of the term species (reviewed by Mayr, 1970, 1982).

A short definition of the biological species concept is: “Species are groups of interbreeding natural populations that are reproductively isolated from other such groups” (Mayr, 1970). This concept was not so much new as a clarification of two distinct threads: (1) a local component, the Poulton–Dobzhansky interbreeding concept, and (2) a global component which extended the interbreeding concept to cover geographical replacement series of actually or potentially interbreeding subspecies (Mayr, 1970), as in the preexisting idea of polytypic species (see also Subspecies, Semispecies, Superspecies).

This extended interbreeding concept was, until recently, almost universally adopted by evolutionists. The species concept problem appeared to have been solved; species were interbreeding communities, each of which formed a gene pool reproductively incompatible with other such communities. The new concept answered both perceived problems of Darwin's morphological approach: (1) that under morphological criteria, distinct mutants and polymorphic variants within populations might be considered separate species, and (2) that sibling species might be misclassified morphologically as members of the same species. The new approach was promoted in a long series of books and articles by Dobzhansky, Mayr, and their followers. Mayr in particular was highly influential by justifying the taxonomic application of the polytypic species criteria in terms of the new concept of gene flow.

In order to adopt this change, it was necessary to see species in a new post-Darwinian light. Instead of species being defined simply, using manmade criteria
based on demonstrable characters such as morphology, species became defined by characteristics important in their own maintenance, that is, by means of their biological function (Mayr, 1982). Significantly, the philosophical term "concept" came into vogue along with these ideas about species, and the term "species problem," which hitherto referred to the problem of how species arose (Robson, 1928), became instead the problem of defining what species were. The important features of species defined by the "biological concept" were that they were protected from gene flow by what Dobzhansky termed "isolating mechanisms," including prezygotic factors (ecological, mate choice, and fertilization incompatibilities) and postzygotic factors (hybrid inviability and sterility caused by genomic incompatibilities). By going beyond simple character-based identification of species, the concept could not apply to all living things; for example, Dohzhansky simply concluded that asexuals (between which no interbreeding is possible) could not have species.

Poulton, Mayr, and Dobzhansky emphasized that their new concept was based on the reality that underlay species, rather than being merely a criterion useful in taxonomy; in this new philosophical approach, taxonomic criteria and conceptual issues of species became separate, with taxonomic criteria taking a more minor role. The concept was true from first principles, and was therefore untestable: difficulties such as hybridization, intermediates, or inapplicability to many plants and asexuals caused taxonomic problems, but did not disprove or even challenge the underlying truth of the concept. These imperfect actualizations of species' true reality were expected in nature. Mayr claimed that the biological concept would do away with "typology" (his term for species definitions based on a fixed, unvarying type or Aristotelian essence), but in many ways it can be seen that the biological concept reverts to a new kind of essentialism, where evolutionary maintenance via interbreeding is the underlying reality of species.

V. ALTERNATIVE SPECIES CONCEPTS

It is interesting that exactly this kind of search for the essence of species had been criticized by Darwin (1859). In his chapter "Hybridism," he specifically argued against using hybrid sterility and zygote inviability as a cut-and-dried characteristic of species (although he made little mention of premating factors). Oddly, Mayr (1982, p. 269) claimed both that Darwin treated species "purely typologically [i.e., as an essentialist] as characterized by degree of difference" and also that Darwin "had strong, even though perhaps unconscious, motivation . . . to demonstrate that species lack the constancy and distinctiveness claimed for them by the creationists." Whether it is reasonable to criticize Darwin in such a contradictory way can be debated, but it is clear that Mayr's proposition that interbreeding is the true essence or reality of species immediately laid itself open to debate. Although the interbreeding concept had a long run (and still does), proposals for different kinds of biological reality of species were eventually forthcoming. By proposing a single reality for species, Poulton, Dobzhansky, and Mayr opened the Pandora's box of alternative essences deemed more important by other biologists.

A. Ecological Species Concept

Asexual organisms such as the bdelloid rotifers can clearly be clustered into groups recognizable as taxonomic species (Hutchinson, 1968). On the other hand, distinct forms such as oaks (Quercus), between which there are high rates of hybridization, can remain recognizably distinct even where they co-occur. This suggested to L. van Valen and others that the true meaning of species was occupancy of an ecological niche rather than interbreeding. This ecological idea became known as the "ecological species concept." It became clear to Mayr during the 1970s also (see Mayr, 1982) that gene flow could not unite every population in a polytypic, biological species' range and that stabilization of phenotype might be effected by ecologically mediated stabilizing selection (see also Sections VII.B and VIII.B) rather than purely because of gene flow.

B. Recognition Concept of Species

An important attack on the biological species concept came from H. E. H. Paterson in the early 1980s. His claims were twofold: first, the Dobzhansky–Mayr term "isolating mechanisms" implied reproductive isolation was adaptive, which Paterson felt was unlikely; second, the true reality underlying species was proposed to be prezygotic compatibility, consisting of mating signals and fertilization signals. According to Paterson (1983), this compatibility is strongly conserved by stabilizing selection, whereas isolating mechanisms such as hybrid sterility or inviability are nonadaptive and can be argued to be a result rather than a cause of species separateness. To Paterson, the true reality of species must be adaptive. He termed his idea of species the "recognition concept" versus Mayr's "isolation concept," and its important characteristics "specific mate recognition systems"
(SMRS) instead of isolating mechanisms. Species were
defined as “that most inclusive population of individual
biparental organisms which share a common fertilization
system” (Paterson, 1985). The idea is generally recognized as a useful critique
and has gained strong currency in some circles. How-
ever, it has been pointed out that SMRS are more or
less the inverse of prezygotic isolating mechanisms and
that the recognition concept therefore differs from the
biological species concept mainly by focussing on a
subset of isolating mechanisms used within the inter
breeding concept. The interbreeding concept had al-
ways stressed a common gene pool and compatibility
within a species, as well as isolation between species.

VI. SPECIES CONCEPTS BASED
ON HISTORY

A. Monophyly
The rise of cladistic methods revolutionized systematics
by proposing that all classification should be based on
the idea of monophyly. This new systematics formalized
the principle that paraphyletic and polyphyletic taxa
were unnatural groupings which should not be used in
taxonomy. It was logical to attempt to apply this idea
throughout systematics, all the way down to the species
level, leading to a monophyly criterion of species, a
type of “phylogenetic species concept” (Hennig, 1968;
see also the diagnostic definition below). Species were
seen as forming when a single interbreeding population
split into two branches or lineages that did not exchange
 genetic material. In a somewhat different formulation,
the “cladistic species concept,” species are branch seg-
ments in the phylogeny, with every branching event
leading to a new pair of species (Ridley, 1996). Other-
wise, if only one of the two branches were recognized
as new, the other branch would become paraphyletic.
Perhaps the main criticism of this idea is that it
could, if applied in taxonomy, cause great nomencla-
tural instability. Monophyly exhibits fractal self-similar-
ity and can exist at very high or very low levels of the
phylogeny, so the precise level at which species taxa
exist becomes unclear. Supposing a new monophyletic
form is discovered overlapping with, but remaining dis-

tinct from, a closely related local form in the terminal
branches of an existing species. Recognition of this
taxon as a species would leave the remaining branches
within the original species paraphyletic. Many other
branch segments would then need to be recognized
at the species level, even if they interbreed and have
reticulate, intermingling phylogenies. Many phyloge-
netic systematists therefore adopt a different phyloge-
netic concept, the diagnostic concept (see VI.C below),
which can allow paraphyly at the species level.

B. Genealogy
Another problem with a monophyly concept is that
a single, true phylogeny of taxa may rarely exist: an
organismal phylogeny is in fact an abstraction of the
actual genetic history, consisting of multiple gene gene-
alogies, some of which may undergo genetic exchange
with other taxa. There is now good evidence that occa-
sional horizontal gene transfer and hybridization may
selectively transfer genetic material between unrelated
forms. Furthermore there are multiple gene lineages
within any population, so that, if such a population
were to become geographically or genetically split into
two distinct forms, it would be some time before each
branch became fixed for different, reciprocally mono-
phyletic gene lineages at any single gene. The idea of
monophyly for whole genomes then becomes hard to
define, especially near the species boundary. However
annoying, phylogenetic methods and evolutionary the-
ory must face up to these facts (Avise and Bull, 1990;
Maddison, 1997). It has therefore been suggested that
species should be defined when a consensus between
multiple gene genealogies indicates reciprocal mono-
phylly. This is called the “genealogical species concept”
(Baum and Shaw, 1995).

Critics argue that this idea has many problems in
common with other monophyly concepts of species
(Davis, 1997). Geographic forms that have become iso-
lated in small populations or on islands, say, could
rapidly become fixed for gene lineages and become
viewed as separate species without any biologically im-
portant evolution taking place. On the other hand,
clearly distinct sister taxa such as humans and chimpan-
zees still share gene genealogy polymorphisms at some
genes such as the HLA complex involved in immuno-

 logical defense and might therefore be classified as
the same species under genealogical considerations.

C. Diagnostic Species Concept
The motivation for the diagnostic concept, usually
called the “phylogenetic species concept” by its adher-
ents, was again to incorporate phylogenetic Hennigian
thinking into species-level taxonomy. There are many
cases of hybridization between taxa on very different
branches of species level phylogenies, which suggest
that interbreeding and phylogenetic realities conflict.
Cracraft (1989) also noted that many bird taxa normally thought of as subspecies were far more recognizable and stable nomenclaturally than the polytypic species to which they supposedly belonged (see also Section VIII.D). Cracraft therefore argued that the polytypic/interbreeding species concept should be rejected, and in its place we should use a diagnostic criterion in the form of at least one fixed difference at some inherited character. "A phylogenetic species is an irreducible (basal) cluster of organisms, diagnosably distinct from other such clusters, and within which there is a parental pattern of ancestry and descent" (Cracraft, 1989). According to Cracraft, species defined in this way are the proper basal, real taxa suitable for phylogenetic analysis and evolutionary studies.

Of course, if diagnostic criteria are applied strictly, rather small groups of individuals, or even single specimens might be defined as separate species. Cracraft recognized this and argued that such diagnosable groups have no "parental pattern of ancestry and descent," i.e., they are not proper populations. However, this qualification appears similar to an interbreeding criterion of species, whereas the whole approach of using diagnostic characters was an attempt to get away from interbreeding.

Most evolutionary biologists balk at the idea of speciation being merely the acquisition of a new geographic diagnostic character, a DNA base pair or color pattern change perhaps (Harrison, 1998). "Speciation" is only a different, or special kind of evolution if the new "species" is a distinct population which can coexist locally with its sibling or parent population without losing its integrity.

Characters used to diagnose phylogenetic species may not be shared derived characters; they may be primitive (pleisiomorphic) characters, or they may have evolved several times. Therefore, phylogenetic species need not be monophyletic, and could presumably be paraphyletic and perhaps polyphylectic. Cracraft appears confused on this matter: on the one hand, he claims that phylogenetic species "will never be nonmonophyletic, and could presumably be paraphyletic and perhaps polyphylectic." Consequently, this qualification appears similar to an interbreeding criterion of species, whereas the whole approach of using diagnostic characters was an attempt to get away from interbreeding.

In other words, Simpson combined the idea that species were historical lineages with the concept of their evolutionary and ecological role. The key essence here appears to be "evolutionary independence." This concept appeals to phylogenetic systematists and paleontologists alike, because of its historical dimension, and to neontologists because of its acknowledgment that biological mechanisms are what make the species real. De Queiroz (1998) is perhaps the most recent reviewer to propose that a single concept, which he calls "the general lineage concept," under which "species are segments of population-level lineages," underlies all other species concepts. According to De Queiroz, apparently competing species concepts merely emphasize different characters or criteria for species definition, but all acknowledge implicitly or explicitly that evolutionary

VII. COMBINED SPECIES CONCEPTS

A. Evolutionary and Lineage Concepts

Faced with the problem of studying the evolution of species through time, the paleontologist G. G. Simpson (1951) proposed his "evolutionary species concept," in which a species is "a lineage (an ancestral-descendant sequence of populations) evolving separately from others and with its own unitary evolutionary role and tendencies." In other words, Simpson combined the idea that species were historical lineages with the concept of their evolutionary and ecological role. The key essence here appears to be "evolutionary independence." This concept appeals to phylogenetic systematists and paleontologists alike, because of its historical dimension, and to neontologists because of its acknowledgment that biological mechanisms are what make the species real. De Queiroz (1998) is perhaps the most recent reviewer to propose that a single concept, which he calls "the general lineage concept," under which "species are segments of population-level lineages," underlies all other species concepts. According to De Queiroz, apparently competing species concepts merely emphasize different characters or criteria for species definition, but all acknowledge implicitly or explicitly that evolutionary
separateness of lineage is the primary concept. This is a nice ideal, but evolutionary independence has little logical force in its application to actual forms that hybridize or undergo genetic exchange.

B. Cohesion Concept of Species

In similar vein, Templeton's (1998) "cohesion concept" fuses a number of competing ideas of species. Templeton argues that a combination of ecological and reproductive cohesion are important for maintaining a species' evolutionary unity and integrity, thereby incorporating components of the evolutionary, ecological, recognition, and interbreeding concepts. As well as applying to asexual taxa ("too little sex"), Templeton's idea also applied to species like oaks that undergo frequent hybridization and gene flow ("too much sex" for the interbreeding concepts). He further argues that separateness of genealogy is another important characteristic of species.

We are perhaps nearing the apogee of the species debate with these combined concepts. By incorporating evolutionary and phylogenetic origins together with every possible biological means by which species are currently maintained, these combined concepts "cover all the bases." One can acknowledge that species evolve and are maintained as cohesive wholes by all of these multifarious processes, yet at the same time argue that species can, and perhaps should be seen as separate from their histories of origin and from current reasons for their integrity. If groups with very different and conflicting biological and evolutionary characteristics are all considered species, there should exist a simpler criterion that unites them. It can also be argued that to confine the origin and evolutionary role of a taxon with the definition of that taxon itself may lead to circularity, particularly in conservation or ecological studies, or when investigating speciation.

VIII. DISSENT: MAYBE SPECIES ARE NOT REAL

Throughout the history of the species debate, starting with Darwin, there have been some who argue that species are not individual "real" objects, but should instead be considered merely as manmade constructs merely useful in understanding of biodiversity and its evolution. These people are not necessarily nihilists, who deny that species exist: they simply argue that actual morphological and genetic gaps between populations would be more useful for delimiting species than inferred processes underlying evolution or maintenance of these gaps. By their refusal to unite these ideas under a single named concept, this biologist "silent majority" have rarely found a common voice.

A. Taxonomic Practice

Taxonomists are on the front line of the species battle, because it is they who ultimately decide whether to lump or split taxa, and at what level to name them as species. If the objectivity and individuality of species as the primary taxon exists, taxonomists' activities have not been made any easier; and many taxonomists have simply ignored or denied belief in the evolutionary reality of species. In general, it is probably true to say that at least 10% of taxonomic species are subject to revision because of these practical difficulties in delimitation.

For this reason, since the rise of the polytypic/interbreeding species concept, there has been little impact of the postwar species concepts on practicing taxonomists, even while the debate raged around them. Procedure, at least in zoology, was more or less as follows: geographic variants which blended (or were thought to be able to blend) together at their boundaries were united within a single, polytypic species, unless morphological or genetic differences were so great that it seemed necessary to recognize two species. On the other hand, whenever two divergent forms differing at several unrelated traits, overlapped spatially, they were recognized as separate species even if a few intermediates suggested some hybridization or gene flow. Some taxonomists regarded subspecies as artificial taxa to be avoided, and may either have ignored geographic variation, or elevated subspecies of polytypic species to the rank of full species. But good taxonomic practice on species remained broadly similar across most branches of systematics, and involved careful analysis of multiple, chiefly morphological character sets tested in large samples of specimens collected from as many geographic regions as possible.

This view on species and subspecies has led in zoology to a steady reduction in the numbers of recognized species, as more and more dubiously separated taxa, previously ranked as species, became inserted as subspecies into larger and larger polytypic species. Recently, however, the diagnostic version of the phylogenetic species concept (Section VI.C) has been making strong inroads into zoological nomenclature, with the result is that counts of species on continents are again climbing as former subspecies are reelevated to the spe-
cies level, in spite of intergradation at their boundaries. However, the situation could get much worse; many Heliconius butterflies, for example, have over 30 geographic subspecies per species, all of which can be diagnosed easily. The numbers of bird and butterfly species could easily increase 2–10 times in some groups if the diagnostic criterion were generally adopted. The one reality that is clear in species-level taxonomy is that the species is not real enough to remain at the same phylogenetic level while taxonomic fashions change. This is good evidence that actual species have been and still are purely manmade taxonomic units lacking in any objectively determinable underlying essence, even if such an essence exists.

B. Populations Are Evolutionary Units, Not Species

Botanists deal with geographically variable organisms with low powers of dispersal, and have therefore never been very happy with the polypotypic/interbreeding concept applied with such apparent success in zoology. Meanwhile, the strong surge in experimental population genetics and evolutionary studies that followed the books by Dobzhansky, Mayr, and Stebbins has led to a greatly improved understanding of gene flow in natural populations. Gene flow in quite mobile animals, such as birds or butterflies, may not unite even adjacent local populations into a common gene pool. If local populations only rarely exchange genes, then gene flow across the range of a continental species is clearly insufficient to explain species integrity, because it would be outweighed easily by weak local patterns of adaptation or genetic drift (Endler, 1977).

This increasing input of population biology into systematics and evolution led to the proposal by Ehrlich and Raven (1969), Levin (1979), and others that species are not real biological units at all; instead, local populations are the only real groupings united by gene flow within a common gene pool, and which adapt to local conditions, compete, and so on. Any homogeneity of ecological niche or genetics over the range of a species might be due either to simple evolutionary inertia or to similar stabilizing selection everywhere. To these authors species exist and are real in local communities; but it is fallacious to treat distant populations in the same way (see also Section VIII.D).

This viewpoint is generally understood and respected by population biologists, but curiously has not been incorporated explicitly into current thinking on species in systematics and evolution. Perhaps there is a sneaking suspicion that even very weak levels of gene flow may explain species integrity over wide areas.

C. Phenetic Species Concept

In the 1960s and 1970s, a major systematics movement proposed numerical methods in taxonomy now usually referred to as phenetics. Pheneticists, as they were called, argued that taxonomy and systematics should be based on multivariate statistical analysis of characters rather than on underlying evolutionary or biological process information. If taxa were defined by nonevolutionary criteria, studies of evolution would be freed from the tautology of testing hypotheses about processes, when those same processes are used as assumptions in the definitions of taxa under study. Species, like other taxa, would be defined in numerical taxonomy on the basis of multivariate statistics, as clusters in phenotypic space (Sokal and Crovello, 1970).

Phenetics is reviled by those who believe that classifications should be phylogenetic. However, the approach is closely similar to the intuitive methods adopted by most actual taxonomists, who use multiple morphological or genetic characteristics to sort individual specimens into discrete groups between which there are few intermediates (see Section V.A). Some large areas of practical taxonomy are based purely on this phenetic approach. Bacterial systematists, for instance, use multiple biochemical tests to assign microbes to species taxa. The usefulness of this taxonomic method is attested by its success in hospitals for predicting pathogenicity and antibiotic sensitivity.

Phenetic classifications based on morphology introduce the danger that, if convergent characters are used as data, one may group unrelated forms into paraphyletic or even polyphyletic taxa. In addition, single gene polymorphisms and sexual dimorphism can affect multiple morphological characters. This could lead to recognition of multiple species within polymorphic populations. Sibling species, on the other hand, could be lumped into the same species using a phenetic approach, unless a set of highly diagnostic characters could be found.

D. Genotypic Cluster Criterion

For morphological or genetic gaps to exist between species, gene flow between species (if any) must be balanced by an opposing force of disruptive selection. In my own work, I had studied hybrid zones between geographic forms of butterflies, and I attempted to show that a practical statistical definition of species could
be constructed using morphological and genetic gaps alone, rather than using phylogenetic or evolutionary processes that caused the gaps to exist.

However, to define species by means of the gaps between them requires consideration of the nature of the gaps to avoid falling into the trap of defining polyphyletic forms as separate species or of lumping sibling species. Rather than merely using external morphology, in difficult cases I proposed that we could consider the genetics as well: DNA has a digital, rather than analog code, so there is a genetic gap between virtually any pair of individuals. Clearly, then, we cannot use just any discreteness at the genetic level to define species. Separate sexes and polymorphic female forms of mimetic Papilio butterflies also have gaps between them in exactly this way. A genetic element, which may be a single base pair, an allele at a gene, the entire mitochondrial genome, a chromosomal rearrangement, or perhaps a sex chromosome, may determine the genetic or morphological differences between such polymorphic forms.

To be considered part of a single local population, and therefore part of the same local species, we expect that polymorphic genetic elements like mimicry genes and sex chromosomes will be mixed in with polymorphisms at genetic elements found on other chromosomes or extrachromosomal DNA. Each individual may be a distinct multilocus genotype, but we recognize groupings of these genotypes because a polymorphism at one genetic element is more or less independent of polymorphisms at another. Conversely, if alleles at one locus are strongly associated with alleles at other, unlinked elements, we have evidence for two separate populations or species.

I therefore proposed a "genotypic cluster criterion" of species. Species are recognized by morphological and genetic gaps between populations in a local area rather than by means of the phylogeny, cohesion, or reproductive isolation that are responsible for these gaps (Mallet, 1995). In a local area, a single species (the null hypothesis) is recognized if there is but a single cluster in the frequency distribution of multilocus phenotypes and genotypes. Separate species are recognized if there are several clusters separated by multilocus phenotypic or genotypic gaps. These gaps may be entirely vacant, or they may contain low frequencies of intermediate genotypes, or hybrids (Fig. 1). The definition is useful because it avoids tautological thinking; hypotheses about speciation or phylogeny become independent of assumptions about the nature of reproductive isolation or phylogeny underlying the taxa studied.

Genotypic clusters are neither profound nor original; I trace the idea to Darwin (1859, see Section III.A), although it undoubtedly goes further back since its use does not require acceptance of evolution. Many similar proposals have been made (Simpson, 1937; Hutchinson, 1968; Sokal and Crovello, 1970; Avise and Ball, 1990; Cohan, 1994; Smith, 1994); in fact, the approach is similar to the practice used on morphology by most taxonomists (see Section VIII.A), and indeed by those who attempt a practical application of the biological species concept (see below). This widespread use of direct genotypic or morphological data, as opposed to reproductive or phylogenetic inferences made from such data, has apparently lacked support because its practitioners were loath to label it as a "concept." I intended merely to justify the Darwinian and practical taxonomists' species definition statistically and in terms of genetics, rather than to enforce the use of genotypes instead of morphology to define species. Most genotypic cluster species can be recognized morphologically; for example, minor pattern elements in Papilio can be used to unite the various polymorphic forms, however, with abundant molecular marker data we could easily use the criterion to sort actual specimens.

There is also nothing wrong with concluding after seeing a male butterfly mating with an unlike female that they belong to the same species, but, because hybridization does occur occasionally between forms normally thought of as different species, one is not so much using the mating behavior itself to define species as inferring that the mating behavior is a common event, so that, if we were to analyze their genomes, the two forms would have similar genetic characteristics apart from those determining sexual dimorphism, i.e., they
would belong to the same genotypic cluster. Instead of reproductive compatibility being the primary criterion of species, we can turn the argument on its head and infer from limited data on reproductive compatibility that a single genotypic cluster may result.

Asexual forms, unclassifiable under the interbreeding concept, and arbitrarily definable at any level under concepts depending on phylogeny, can be clustered and classified as genotypic clusters in exactly the same way as sexual species. The precise taxonomic level of species clustering is potentially somewhat arbitrary, as in the phylogenetic concepts, but at least the method acknowledges this arbitrariness rather than pretending that some higher evolutionary principle is being used. However, many asexual forms such as bdelloid rotifers have easily distinguishable species taxa (Hutchinson, 1968). In bacteria, competition is thought to structure promiscuous, but largely asexual, populations into recognizable genetic clusters (Cohan, 1994; see also Section V.A).

Critics have argued that the genotypic cluster criterion in sexual species is nothing other than a gene flow concept of species under a different guise. This is true for one specialized interpretation of “gene flow.” If we define gene flow as the successful or effective as opposed to actual input of genes, then it is easy to see a “gene flow criterion” becomes similar to the “genotypic cluster” criterion: to find whether a hybridization or gene flow event is successful, we must either follow every gene through all possible descendants for all time, or we may examine the genotypic state of a population and determine if genes from one form are mixed randomly with genes from another form. It seems clear that looking for random association of genes within genotypes in the genotypic cluster approach will be methodologically the same as genotypic state analysis to determine whether a population is interbreeding, but the latter requires additional assumptions. The genotypic cluster criterion in sexual species could be looked upon simply as a practical application of the biological species concept. However, I distinguish the genotypic cluster criterion from the interbreeding concept, if only because its name emphasizes that the definition is character-based, rather than actually based on interbreeding, and is thus applicable to asexuals as well as sexual species. If a single geographic race, which previously intergraded at all its boundaries with other geographic races, were to split into two forms that coexist as separate genotypic clusters, we could have the situation that the original polytypic species became paraphyletic. The new species has been derived from only one of the component subspecies. Thus paraphyly of species must be recognized as a possibility under this definition, as with both interbreeding and diagnostic concepts.

E. The Unreality of Species in Space and Time

Geographic races often form clusters differing at multiple loci from other races in the same species. The interbreeding concept or genotypic cluster criterion can be used to justify a classical polytypic species if the various geographic races are separated by zones which contain abundant intermediates (hybrids). We sample multilocus genotypes or phenotypes in local areas of overlap and determine whether a single peak (one species, i.e., abundant hybrids), or two peaks (two different species, i.e., rare hybrids) are evident in the local genotypic distributions (Fig. 1). Hybridization may occur, but if it is rare so that character and genotypic distributions remain distinctly bimodal in zones of overlap, we usually classify them as separate species, even under the interbreeding concept.

Although this spatial extension of the local species is practical to apply to any pair of forms in contact, it is unlikely to lead to general agreement. The problem is that hybrid zones can be very narrow and may separate forms that are highly distinct at multiple characters or loci, in spite of complete unimodal blending in local areas of overlap. Even adherents of the interbreeding concept are reluctant to lump such geographic forms within the same species. Examples include North American swallowtail butterflies (Papilio glaucus/P. canadensis; see Hagen et al., 1991) and European toads (Bombina bombina/B. variegata; see Szymura, 1993).

An even worse problem is found in “ring species,” which form a continuous band of intergrading subspecies, but whose terminal taxa may be incompatible, and overlap without intergrading. A commonly cited example are the herring gulls and lesser black-backed gulls (Larus argentatus complex; Mayr, 1970). Similarly, while most hybrid zones between European Bombina are unimodal, the same pair of taxa may have bimodal genotypic distributions in other zones of overlap (Szymura, 1993). Thus, geographic forms may be apparently conspecific in some areas, but overlap as separate species in other areas. Finally, if distinct populations are geographically isolated and there is no area of overlap, one cannot disprove the null hypothesis of “same species” under interbreeding or genotypic cluster criteria, but biologists are reluctant to unite such populations if they are very divergent. Laboratory hybridization could be tried, but many overlapping species are known to hybridize freely in captivity, while remaining separate.
in nature. There are good examples in the great apes, for instance, the bonobo (Pan paniscus) versus the chimpanzee (P. troglodytes) and the gorillas (Uchida, 1996), but similar decisions must be made in almost any animal and plant group.

The problem for spatial extensions of species criteria is in fact due to the evolutionary divergence of spatially separated lineages: time and space are correlated. Paleontologists face a similar temporal problem when classifying fossils in different strata. Evolutionary rates may vary, but all lineages must ultimately be continuous, so there is no very logical place to put a species boundary in time any more than there is in space. Paleontologists typically use operational species on the basis of morphological gaps between taxa from the same and different time periods (Simpson, 1937; Smith, 1994).

These difficulties show that there is no easy way to tell whether related geographic or temporal forms belong to the same or different species. Species gaps can be verified only locally and at a single time. One is forced to admit that Darwin’s insight is correct: any local reality or integrity of species is greatly reduced over large geographic ranges and time periods (see also Mayr, 1970, and Section IX.D).

IX. THE IMPORTANCE OF SPECIES CONCEPTS FOR BIODIVERSITY AND CONSERVATION

A. Traditional: Species as Real Entities

Different species concepts seek to define species in mutually incompatible ways. Thus, a monophyletic species concept seems not very useful to evolutionary biologists because of difficulties with multiple gene genealogies and paraphyletic remnants. In contrast, the interbreeding concept and other combined concepts incorporating biological processes of species maintenance suffer in the eyes of phylogenetic systematists because they lack phylogenetic coherence and produce paraphyletic taxa, or worse. If we were to allow the basal unit of our taxonomy to incorporate paraphyly, it would be harder to justify a strict adherence to monophyly at other taxonomic levels. It is beyond the scope of this article to resolve these difficult issues, but these conceptual conflicts fuel the continuation of the debate, and also highlight the fact that if species are indeed real, objective biological units, their unifying reality has been extremely difficult to verify.

Many ecological and biodiversity studies of actual organisms ignore these difficulties, and assume that species are objectively real basal units. Thus in ecology, we have theories of global species diversity. In conservation, we have the Endangered Species Act in the United States, which prescribes the conservation of threatened taxa we call species. Populations not viewed as species, particularly putative hybrid taxa (like the red wolf, Canis rufus, of the southeastern United States), are seen as less valuable, even if rare. How do we recognize that a taxon is hybridized? Obviously, to be a hybrid, it must be a mere intergrade between two, real, objectively identifiable entities. The Endangered Species Act views species as important real conservation units and hybrids as unimportant. It did this because it incorporated the species concept in vogue at the time of its enactment, i.e., the biological species concept, in which hybridization is seen as a “breakdown in isolating mechanisms” (Mayr, 1970).

B. Alternatives: Genetic Differences More Valuable Than Species Status

If this Act were to be rewritten today, what would it say? There is undoubtedly a greater realization today that other levels in the taxonomic hierarchy are important elements of biodiversity. The diagnostic concept of species, while claiming to support the basal, objective nature of species, can at least have the beneficial effect of allowing its basal unit of biodiversity to be recognized at a lower level, in this case as subspecies within polytypic species. Some molecular geneticists have advocated conservation of “evolutionary significant units,” “management units,” or “stocks” (a fisheries term) defined on criteria of continuous genetic differentiation at molecular markers (Moritz, 1994) as being more important than worrying about the species level at all. The true reality of spatiotemporally extended species eludes us, if it exists. If this is so, then it seems best to adopt some other measure of conservation value that relies purely on the degree of genetic differentiation, for instance, at molecular genetic markers.

C. Species Differences as Ecologically Important Markers

However, there are many who oppose this view. Species within a local area such as a nature reserve are, for the most part, easily and objectively identifiable using morphology, behavior, genetics, or phylogeny. A pair of similar species must usually be ecologically different to coexist. To remain distinct, sexual species will need some prezygotic isolation, so their mating behavior
must also be different. Thus, counting species in a local area makes ecological sense, and conserving species diversity in a local area would conserve actual ecological and behavioral diversity. Behavioral and morphological differences that cause speciesness are evolutionarily more valuable than potentially neutral genetic differences at molecular markers.

D. Biodiversity in Space and Time

As we have seen, this local view breaks down when we try to apply the term "species" over large areas or geological time scales. In some cases, there is excellent homogeneity over large areas; for example, the painted lady butterfly (Vanessa cardui) and the barn owl (Tyto alba) have a virtually worldwide distribution and look nearly identical everywhere. Other species are not so homogeneous: the familiar mallard group of ducks (Anas platyrhynchos) has become highly differentiated into some 18 or so forms in far-flung outposts of the world, but exactly how many are good species, and how many are races, or indeed, how many races there are in total, is a matter of taste. Current authorities recognize about 10 species, but there might easily be five or 15 in alternative treatments. One of the forms, the Mexican duck, A. platyrhynchos diazi, is threatened with hybridization by its mallard relative, Anas platyrhynchos platyrhynchos, which has been expanding from the north, and the American black duck (Anas rubripes) also hybridizes with the mallard but appears to resist hybridization somewhat better than the Mexican form—hence its species status.

Faced with these difficulties, should we worry about the species level when conserving endangered taxa over large areas? Whatever the answer to this question, it does not seem very sensible to rely too much on the spatiotemporal reality of species for an answer. We might rename the Mexican duck as a separate species instead of a subspecies, but ideally this should not affect its conservation value since there has been no actual change in the knowledge of biological characteristics that affect conservation value. Most conservationists now agree that the former fetish for species-level legislation was a mistake: conservation and legislation should now recognize that living, evolving populations form fractal continua over time and space, rather than attempting a division into spurious "fundamental" units.

Species are certainly fundamental units of local biodiversity, but they have this clarity only in a small zone of time and space, and so species counts become less and less meaningful as larger and larger areas are covered. Taxonomists might come to nominalistic agreements on a case-by-case basis, but even this shows little sign of happening yet. Ecological theory, as well as conservation and biodiversity studies must recognize that species counts over large expanses of space and time represent only a sketchy measure of biodiversity, a measure which owes more to taxonomic and metaphysical fashion than to science.

See Also the Following Articles

SPECIES, CONCEPTS OF

See Also the Following Articles

BIODIVERSITY, DEFINITION OF • CLADISTICS • DARWIN, CHARLES • EVOLUTION, THEORY OF • PHYLGENY • SPECIATION, PROCESS OF • SPECIES DIVERSITY, OVERVIEW • SUBSPECIES, SEMISPECIES • TAXONOMY, METHODS OF

Bibliography


Faced with these difficulties, should we worry about the species level when conserving endangered taxa over large areas? Whatever the answer to this question, it does not seem very sensible to rely too much on the spatiotemporal reality of species for an answer. We might rename the Mexican duck as a separate species instead of a subspecies, but ideally this should not affect its conservation value since there has been no actual change in the knowledge of biological characteristics that affect conservation value. Most conservationists now agree that the former fetish for species-level legislation was a mistake: conservation and legislation should now recognize that living, evolving populations form fractal continua over time and space, rather than attempting a division into spurious "fundamental" units.


SPECIES DIVERSITY, OVERVIEW

A. Ross Kiester
U.S. Department of Agriculture Forest Service

GLOSSARY

clique A group of organisms in a food web in which any pair of species in the clique shares at least one prey species. Thus, it is a group of species that have similar diets and are different from other such groups.
gap analysis A biodiversity policy tool which compares the actual distribution of species and vegetation classes to areas managed for the long-term protection of biodiversity and other classes of land management. Species and classes that are poorly represented in protected areas are gaps in the protection of all biodiversity. Such species and classes are candidates for proactive protection.
lognormal distribution A statistical distribution which is normal or Gaussian ("bell curve") when the logarithm of the original data is used. Species abundance data from many communities fit this distribution well. The lognormal distribution often occurs as the result of the law of large numbers in which many small factors interact multiplicatively to produce a result such as species diversity.
Simpson diversity index An index of diversity based on species abundance data. If $p_i$ is the frequency of species $i$ in a community of $n$ different species, then
\[
\text{Simpson diversity} = \left(1 - \sum_{i=1}^{n} p_i^2\right)
\]
and is equal to the probability that two randomly chosen individuals from a given community are of the same species. This index of diversity is statistically unbiased, making it especially useful in practice.
species diversity The variability of a group of species classified in some way. For example, individual species can be classified by size and then a species size diversity measure can be developed which expresses the variability with respect to size. The most common classification is by abundance or population numbers. Then species diversity measures the variability of the number of individuals of each species. The Simpson diversity is an example of such an index. Many such indices have been developed for species diversity.
species richness The number of species occupying a particular area (such as an island) or biological entity (such as a branch of a tree) without regard to any other properties of the species. Species richness may
SPECIES DIVERSITY is one of the most fundamental aspects of biodiversity. However, there are many ways to think about it and ever more ways to measure it. The most basic idea is that of species richness, which is simply a count of the number of species inhabiting a given area or habitat. Species richness may also be thought of as the list of species that were counted. Patterns of species richness in space and time are some of the most foundational data in ecology, biogeography, and paleontology. These patterns both suggest and test theories of species richness dynamics. In addition to species richness, there are many concepts of species diversity. Usually species diversity is formed from species richness by further classifying the species by some attribute, such as abundance, size, or ecological role. When species are classified by abundance or population size, species diversity is the variability in the distribution of individuals into species. This species abundance relation and other measures of species diversity link species richness to many ecological processes such as population dynamics and competition. They also link to evolutionary processes such as adaptive radiation and the evolution of phenotypic plasticity. Species diversity is the building block for the diversity of higher taxa and for the diversity of ecological associations such as communities and biomes. Species diversity is an important policy tool because species diversity is relatively easy to measure, and it is possible to frame management goals in terms of species diversity. As a practical matter, it is frequently necessary to use species diversity as a surrogate for other levels of biodiversity that may be much more difficult to measure.

Here, we initially focus on species richness (the number of species in a given area) and then discuss other approaches to species diversity. Finally, we discuss the roles of species diversity in biodiversity policy.

II. SPECIES RICHNESS

The simplest form of species diversity is species richness, which is the number of species which occur in a given area. For example, the species richness of hippopotamuses in Africa is two. However, even within this simple definition there are many important issues. For example, species richness is thought of as a number, but corresponding to each number is also a list of species. The list [Hippopotamus amphibius, Choeropsis liberiensis] corresponds to the richness number of 2 for living African hippos.

Computing species richness requires choosing a concept of species and a higher taxon for which a list of species exists. There are many species concepts available, but most are grounded in the theory of evolution and have a theoretical basis. That is, they are entities for which there is a theory that provides methods of determining similarity and difference. The theory is what is used to resolve differences between various workers classifications. Species are thus in contrast to many other biodiversity entities such as higher taxonomic categories or ecological associations such as communities and ecosystems. These entities are more creations of humanity and their classifications are more

I. INTRODUCTION

The study of biodiversity often begins with species diversity because it is the most familiar aspect of biodiversity as a whole. Popular interest in biodiversity such as bird watching usually begins with learning to distinguish species, and this leads to a direct appreciation of species diversity. Later, other aspects of biodiversity, such as morphological variants or subspecies, the phylogenetic grouping of species, or the habitats and landscapes associated with particular species, are added to the initial view. Species diversity also lies at the heart of much of evolution and ecology, the sciences most concerned with biodiversity. Each discipline, in its own ways, seeks to explain as much as possible about species diversity and its patterns in time and space. Both sciences use the patterns of species diversity found in nature as tests for ideas about various ecological and evolutionary processes. The relation of species diversity to all other elements of the biodiversity hierarchies of taxa and ecological assemblages (such as communities and ecosystems) is also an important focus of these sciences.
Species richness is most commonly expressed in terms of area. It may be expressed in terms of a political subdivision of the earth or an ecological subdivision of the earth, or it may be calculated by reference to an equal area grid placed arbitrarily with regard to political boundaries. In all cases, a process of discretization occurs. A species is counted simply as present or absent within a given area. Attributes such as the abundance of the species in the area or the percentage of the area occupied are not taken into account. For equal area grids we obtain an approximation of the underlying continuous variation in the species richness in such a way that we are able to quantify richness. Figure 1 shows a map of species richness for the reptiles (crocodilians, turtles, lizards, and snakes) of the continental United States. The grid cells are 100 miles on a side or 10,000 square miles in area. Figure 2 shows hand-drawn isolines of species richness which give a sense of the underlying continuous distribution.

However, species richness may also be expressed in terms of other kinds of aggregates of species. We may determine the species richness of the parasite load of a single animal or we may determine how many species of epiphytes live on a given tree.

As with any process of discretization, the patterns of species richness revealed are very much a function of the underlying scale of the discretization. Figures 3 and 4 show species richness of terrestrial vertebrates for the state of Oregon at two scales of discretization. It is clear that the pattern of species richness is scale dependent. Any inferences made from a map of species richness must clearly take the effect of scale into account. It is worth noting that finer resolution maps are not necessarily better for all purposes. There is an uncertainty tradeoff between accuracy and comparability in the choice of grid size. A smaller grid size is more accurate, but a larger grid size may produce cells that are more comparable since the effect of data of varying accuracy is smoothed out. For example, a coarser resolution map may provide better insights into large-scale biogeographic patterns because ecological noise may be averaged out. Furthermore, a minimum scale of discretization must be chosen that is appropriate given the accuracy of the individual range maps from which species richness is calculated. The data for Fig. 1 were compiled from the Peterson Field Guides to Amphibians and Reptiles (Conant and Stewart). The range maps in these books show a great deal of cartographic generalization, so a relatively large-scale grid is appropriate.

In general, it is a good idea to make the grid cell size several times larger than the spatial scale of the assumed errors in the individual maps. Most work has been done with species richness as a number because the manipulation and analysis of large lists have only become possible with relatively recent advances in computer science. However, there is clearly more information in the list of species richness than the simple number. Two areas may easily have the same number of species but quite different lists. In general, the further apart two areas are, the fewer species they will have in common. Species richness thus has two components: the absolute number of species in an area and a relative measure of the uniqueness of one area compared to another. This measure is frequently computed by comparing the lists of species from the two areas and noting the number of species that are unique to each area and the number that are common to both areas.

The maps shown in Figs. 1, 3, and 4 are typical of species richness maps derived from the mapped ranges of individual species on which they depend. Calculation of species richness from a set of individual maps assumes that the maps are relatively consistent with regard to accuracy and underlying scale. This in turn requires that the taxonomy of the group under consideration be relatively complete and stable. This requirement is true only for certain taxa and for some parts of the world. In the United States and Europe, vertebrates, butterflies, and trees are generally sufficiently well-known to allow the construction of species richness maps. For the entire planet, only a few groups, such as birds, turtles, and milkweed butterflies, are sufficiently well-known to allow global species richness maps to be constructed. It is difficult or impossible to calculate species richness for poorly known taxa that have many undescribed species or whose taxonomy is controversial. Even for very well-known groups, it is not clear how up-to-date the range maps are since they are often necessarily based on collections made over a long period of time and almost never on a recent species-wide sam-
FIGURE 1  Species richness of North American reptiles; raw richness on a grid scale of 10,000 square miles per cell (reproduced with permission from Kiester, 1971).

FIGURE 2  Isolines of North American reptile species richness. The lines were hand interpolated from the data of Fig. 1 (reproduced with permission from Kiester, 1971).
FIGURE 3 Terrestrial vertebrate (amphibians, reptiles, breeding birds, and mammals) richness for Oregon; raw richness on a grid scale of 71 km² (reproduced with permission from Kennelly, 1998).

FIGURE 4 Terrestrial vertebrate (amphibians, reptiles, breeding birds, and mammals) richness for Oregon; raw richness on a grid scale of 213 km² (reproduced with permission from Kennelly, 1998).
plunging program. For example, the data used to make the maps of Oregon in Figs. 3 and 4 were based on locality records during a 30-year period in order to have relatively complete coverage. Therefore, data for species whose ranges have changed drastically in the past two decades will be out-of-date and consequently the richness maps are in some sense always out-of-date.

There are two ways in which species richness data are accumulated. The first is the slow process of accumulation of the description of new species and the second is a more rapid method in which species richness of a given area is estimated from a statistically designed sampling effort.

As science progresses, our knowledge of species richness increases. This pattern, which is usually an increasing function of time, is called a species accumulation curve. We can think of species as being accumulated on several scales of effort and geography. For example, at the largest scale, that of all scientific efforts taken together for the entire world, we can see a historical accumulation of species. In 1758, Linnaeus recognized 11 species of turtles, in 1889 Boulenger listed 201, and in 1992 Iverson listed 237. New species of turtles are still being described even though they are among the best known animals. Turtle taxonomists estimate that by 2005 the number will be more than 300. The number of turtle species has not increased due to any organized or statistically designed effort. The work has simply proceeded as a haphazard attempt at exhaustive enumeration.

In contrast, several attempts have been made to statistically estimate species diversity at a particular location. Usually these efforts determine both the number of species and the number of individuals. In such efforts the number of species caught is a biased estimator since rare species are often necessarily missed. The contrast between haphazard enumeration and statistical estimation is that between more complete knowledge without error estimates and less complete knowledge with error estimates. The best method of synthesizing these two approaches via meta-analysis is an outstanding research problem.

For many taxa and most localities throughout the world, species richness is known only for certain restricted localities such as refuges and research field stations. In some cases, the exact area of knowledge is known, but in others the site may be only generally described. These point locality data are sometimes used to extrapolate species richness over larger areas. This is a new area of research reviewed by Colwell and Cody (1994), who found many unanswered questions. However, extrapolation may be the best hope of obtaining an estimate of global diversity in the foreseeable future.

III. PATTERNS OF SPECIES RICHNESS

The most basic pattern of species richness is that between the number of species and the size of the area considered. In general, the larger the area, the more species that will be found in it. This species–area relationship frequently takes the form

\[ S = cA^z \]

where \( S \) is the number of species, \( A \) is the area considered, and \( c \) and \( z \) are estimated constants. By taking logarithms of both sides of this equation, we find that the logarithm of species richness is a linear increasing function of area. However, many empirical studies have found that the data scatter about a wide band of values of \( z \) and that other mathematical forms of the relationship may be equally likely. The analysis of the mechanisms by which area determines species richness is a major line of enquiry for ecologists and biogeographers. Species richness is very unevenly distributed over the entire Earth. The most important general pattern is that richness increases as one goes from the poles to the equator. Figure 2 shows that this latitudinal gradient in species richness can be quite striking. The average number of species of reptiles increases from 10 at the Canadian border to more than 60 at the Mexican border (Fig. 2). Of course, many taxa are exceptions (penguin species diversity is highest closer to the south pole and there is only one species at the equator), but overall the pattern is strong. Again, studies of the latitudinal gradient in species richness have been a major line of enquiry for ecologists and biogeographers. Another pattern is the peninsula effect whereby species richness decreases toward the end of a peninsula. For example, the number of species of mammals and reptiles decreases in Florida and Baja California as one moves toward the end of these peninsulas. This peninsular pattern can be seen for reptiles on the Florida peninsula in Fig. 2.

Species richness varies strikingly in time. The paleontological record provides good evidence that overall species richness increases with time but that there may be massive extinction events that drastically reduce species richness in a relatively short time. At ecological timescales species turnover is common so that the list of species is often changing. The total number of species
usually changes less and usually in response to a change in the environment.

IV. DETERMINANTS OF SPECIES RICHNESS

Over the continuum of scales of time and space, the sciences of ecology, biogeography, and evolution all strive to understand the processes which determine species richness. Indeed, concern for species richness unites these fields (as well as bridging fields such as macroecology and paleoecology). One can easily imagine that they are all part of a single science concerned with species richness. It was recently proposed that this field be called the study of biodiversity dynamics (McKinney and Drake, 1998) and that it should include all kinds of diversity, not just species diversity.

The first question to be addressed is the role of history and the question of equilibrium versus nonequilibrium dynamics. Are the patterns of diversity that we see a result of the vicissitudes of history or of the action of relatively general mechanisms? For example, is the latitudinal gradient in species richness in North America fully recovered from the end of the last ice age. For certain taxa with good dispersal abilities, such as birds or butterflies, it is very likely that they have finished responding to that episode of climate change and are near equilibrium. Other groups, such as salamanders, with very poor dispersal capabilities may still be in the process of moving into areas opened up at the end of the glaciation and hence their latitudinal gradient may not yet be at equilibrium. Another way of viewing this question is to ask to what extent are there general mechanisms that determine species richness in particular cases and to what extent are there only particular cases. Some ecologists believe that there are only case studies because the details of any particular case overwhelm any instance of a potential general pattern. Others are much more sanguine that general rules and even laws will be able to account for species richness. In any event, for all of the biodiversity sciences the relation between the general and the particular is a difficult problem.

One useful way of dealing with the issue of the general and the particular was developed by E. E. Williams (1969), who introduced the distinction between the distant and close views in island biogeography. In the distant view is typified by MacArthur and Wilson’s (1963) theory of island biogeography that accounts for the patterns of species richness seen on islands that areof different size and different distances from a source mainland. This model predicts an equilibrium number of species as a function of rates of colonization and extinction. These rates are statistical composites formed from an ensemble of many species and are the mechanisms of the distant view. Here, only the number view of species richness is used. The equilibrium number of species on an island is predicted, but the actual list of species is not. This is not viewed as a problem because the theory predicts that species are constantly turning over and so the list is always changing. The close view is taken by Williams (1969) in his work on the lizards of the genus Anolis of the Caribbean or by David Lack (1976) in his study of the land birds of Jamaica. In this approach, the focus is on a list of species and the details of the biology of each species examined are considered important. Mechanisms such as colonizing ability, competition, ecological plasticity, and the details of the interactions of species and available habitats are studied on a species by species basis. Adaptive radiation is often considered an important mechanism for generating diversity on a given island. The distant view has the strength of numbers and interprets broad patterns well. It does not account for the actual biology of any particular species. The close view gives detailed accounts of the biology of particular species but generally is not able to predict the number of species. Thus, the distant view focuses on the number version of species richness, whereas the close view is concerned with the actual list of particular species. The close view is not contained within the distant view. Rather it is complementary to it.

V. A DIVERSITY OF SPECIES DIVERSITIES

Beyond species richness are an enormous number of species diversities. Generally, these take the form of a species richness of species classified by some property of the species. Species may be classified by abundance (population size), body size, trophic position, taxa, membership in various ecological assemblages, and in many other ways. As with species richness, the actual number of species in a given class or the list of species in a given class may be analyzed. Because there may be several classes to which individual species are assigned,
it is possible to develop an enormous variety of species diversity indices which summarize in various ways the pattern of distribution of species into the classes of a particular classification. The study of these diversity indices, their sampling properties, and the inferences about ecology and evolution that can be drawn from them have been the subject of major scientific activity in the past few decades.

The most common extension of species richness is to the species abundance relation. Here, a set of species is classified by the abundance or population size of each species. Some species are rare and others common; the distribution of all species abundances is called a species abundance curve. Figure 5 shows the species abundance curve for a classic data set of macro-Lepidoptera (mostly large moths) collected by light trap at Rothamsted in England in 1935. A total of 6814 individuals of 197 species were collected. Thirty-seven species were represented by a single individual, whereas the most common species was represented by 1799 individuals. In this example, the number of classes of the classification is just the number of different integer valued numbers of abundance. All abundances greater than 65 were represented by single individuals. As with many data sets of species abundance, this one fits a lognormal distribution well. The many attempts to understand these data both as empirical distributions and as data to test predictions made by theories of community structure have been reviewed by May (1975) and Magurran (1988). Many of the indices fall on a continuum from emphasizing rare species to emphasizing common species. Simple species richness is at one end of this continuum since it weights all species equally. At the other end is the Simpson index, which gives the most weight to the species with the most individuals. Of the many empirical indices of species diversity that combine species abundance data, Lande (1996) found that only the Simpson index is statistically unbiased and leads naturally to measures of similarity between multiple communities. Using both species richness and the Simpson index seems to provide a useful characterization of the species abundance properties of a community.

Living organisms vary enormously in size and it is commonplace that there are fewer large species than small species. Hutchinson and MacArthur (1959) pioneered an ecological approach to understanding the species diversity of body sizes. Figure 6 shows the patterns of species diversity by body size for mammals over three scales in North America. At the scale of the continent, the distribution is unimodal centered on medium and small sizes. At the finest scale (ponderosa pine community of the Oregon Cascade Mountains) the distribution is near flat and shows distinct clumps. This area of research has many competing hypotheses which have been reviewed by Kelt and Brown (1998). Among these hypotheses, a basic idea is that there is more environmental variation at smaller size scales. The world looks more complex to a small beetle than to a larger grazing mammal. This environmental complexity in turn provides a great number of potential niches at smaller scales, allowing coexistence of more species. A provocative analysis by Holling (1993) presented evidence that species sizes of birds and mammals in individual ecological communities are clumped into a small number of discrete, separate size groups. He interpreted this pattern to be the result of a small number of structuring processes ranging from vegetative growth to disturbance dynamics and geomorphological processes which form a discontinuous temporal hierarchy and generate the discontinuous size distributions.

The familiar concept of a food web leads to a classification of species by their position in the web. Here, the basic pattern is that there are fewer species higher in the food web. There are fewer species of top predators than herbivores. Beyond these simple trophic level gen-
One intriguing result concerns the relationship between the species diversity of a web and the number of dominant cliques in the web. A dominant clique of a food web is a group of species in which the predator species share at least one prey species in common and in which the groups are isolated from other such groups and do not share any of their prey species. Figure 7 shows this relatively strong relationship. This pattern may be interpreted to mean that food webs are organized into functional units (groups of interacting predators and prey) and that species diversity increases with the number of such groups.

The species diversity of higher taxa ranging from genera to phyla varies a great deal. The phylum Uniramia, which contains the myriapods and insects, has well over 1 million described species "but many millions more probably await scientific discovery" (Barnes, 1998, p. 247), whereas the phylum Placozoa apparently has only a single species—Trichoplax adhaerens (p. 179). Other ranks in the taxonomic hierarchy vary similarly in their diversity. We are just beginning to understand the patterns of species diversity of higher taxa and clearly this requires a synthesis of evolutionary and ecological theories. Why orchids and beetles are more diverse than many comparable groups is a function of evolutionary plasticity and ecological opportunity. Evolutionary plasticity is a result of both genetic diversity and developmental constraints and opportunities, whereas ecological opportunity is influenced by structural diversity and energy availability. Currently, we lack a detailed understanding of both of these concepts. However, recent paleontological work has demonstrated some interesting patterns of the temporal structure of species diversity of higher taxa. In particular, as origin and extinction probabilities of higher taxa are followed from the Paleozoic to the present there appears to be an accumulation of higher taxa with lower extinction probabilities. Certain groups seem to have greater staying power than others and so they are more prevalent today. These patterns offer new possibilities for testing ideas about the evolution of higher taxa species diversity.

At the other end of the taxonomic hierarchy, species may be classified by their internal variability. This variability has something of a hierarchical structure of its
own: subspecies, geographic structure, population structure, and genetic structure. Subspecies represent the largest scale of variation within species and are usually thought of as being composed of a major portion of a species range within which several characters are concordant. Geographic variation may also be continuous as with clines or nonconcordant with different characters varying with different geographic patterns. Population structure ranges from continuous to highly fragmented and island-like with many varieties of meta-population structure in between. Genetic structure follows in part from population structure and provides variability beyond that. The degree of genetic polymorphism shown by local populations provides the most basic variation shown by species. There have been few attempts to quantify species diversity of within-species diversity, although such analyses should lead to further integration of ecological and evolutionary approaches to species diversity.

VI. SUMMARY: STRUCTURE OF TAXA AND ECOLOGICAL ASSEMBLAGES

Almost any classification of species results in the fact that species richness of the classes is unequal and often shows interesting patterns. The description of these patterns and their interpretation in terms of evolutionary and ecological processes is a cornerstone of biodiversity science. Two of the most important kinds of classification are those that classify species by membership in higher taxonomic categories (which, it is hoped, represent phylogenetic groups as well) and into various ecological assemblages such as food webs, communities, landscapes, and biomes. These patterns of species diversity function both to generate and to test hypotheses about the phylogenetic and ecological processes. A great deal of effort has been devoted to analyzing some classifications such as abundance. However, it is clear that attempting to explain the pattern produced by a single classification will inevitably be incomplete. Abundance, body size, and trophic position are strongly related—large top predators are uncommon—and rather than try to explain these characteristics separately, work now focuses on attempting to explain the whole syndrome. The methodology of progressive synthesis outlined by Ford (2000) offers one way to counteract the usual pattern of fragmenting questions. The ultimate challenge to students of biodiversity is to devise better ways of understanding how the phylogenetic and ecological classifications of species diversity interact.

VII. SPECIES DIVERSITY AND BIODIVERSITY POLICY

Of the many levels of the hierarchy of biological diversity, from genes to the world, species diversity historically has played the dominant role in attempts to maintain and conserve biodiversity. Generally, this has led to a species by species approach such as is exemplified by the Endangered Species Act. This approach has usually been reactive and has been analogized to emergency room medicine. However, recent efforts such as the Gap Analysis Program have focused on sets of species and attempted to lay out proactive approaches. These approaches have in turn been analogized to public health medicine. Generally, it is less expensive to protect a species from extinction while it is relatively common rather than to wait until it is endangered. Both of these approaches use species diversity as a key level of biological diversity (gap analysis also uses vegetation communities as an important level). This focus on species diversity is due to several factors. First, the idea of conserving species is well understood by many people who have a good intuitive understanding of the concept. Second, as mentioned previously, species are scientifically well-defined entities. Finally, species diversity is apparently a good surrogate for many other levels in the hierarchy, such as genes, which are much more difficult to measure in nature. Therefore, it is likely that species diversity will remain a central focus of conservation biology even as other levels of biodiversity become better understood.

See Also the Following Articles

DIVERSITY, COMMUNITY/REGIONAL LEVEL • DIVERSITY, MOLECULAR LEVEL • DIVERSITY, ORGANISM LEVEL • DIVERSITY, TAXONOMIC VERSUS FUNCTIONAL • LANDSCAPE DIVERSITY • SPECIES-AREA RELATIONSHIPS • SPECIES, CONCEPTS OF SUBSPECIES, SEMISPECIES

Bibliography


I. Introduction
II. Pairwise Interactions
III. Interactions and Evolution
IV. Multispecies Interactions
V. Human Impacts on Interactions
VI. Summary

GLOSSARY

community The collection of all species which occur together in space and time.
intrinsic rate of growth A parameter, often designated as the variable $r$, reflecting the rate of change in population size due to births and deaths.
population A collection of individuals of the same species in a defined geographic area.
population dynamics The change in the number of individuals in a population over time.
resource Any necessity that is used or consumed for organismal growth, reproduction, or survival, including food, space, shelter, and nutrients.
species A group of individuals that is capable of interbreeding; a species is composed of one or more populations.
trait A quality possessed by an individual organism that affects its ability to survive and reproduce.

SPECIES INTERACTIONS, including competition, predation, mutualism, and parasitism, are associations among individuals of two or more different species such that each species' presence has an influence on the population dynamics of the other.

I. INTRODUCTION

The world can be divided into two parts—the living and the nonliving. Scientists refer to the nonliving component of the earth as the abiotic environment, and all living organisms are labeled biotic. Organisms survive and reproduce under many abiotic and biotic influences. Abiotic nutrients and elements are needed for survival and reproduction, and nutrient limitation is a common factor regulating limits to growth. Biotic factors also influence how an organism lives and grows. When organisms come into contact, they can facilitate, control, harm, use, or even help to propagate each other. Biotic associations that involve members of two or more different species are called species interactions.

Because the world contains a tremendous diversity of species, interactions among the biota are common. Many of these interactions are conspicuous and have a strong influence on the organization of biological communities. Studies of species interactions have produced a large body of data in a great variety of ecosystems, and these studies comprise an entire discipline called community ecology. The treatment of species interactions given here will briefly discuss the types of interactions found in nature and the simple mathematical representations of their effects, discuss the role evolution plays in shaping interactions, and provide some empirical examples. Because an increasing number of
community ecologists are examining the effect of human influences on species interactions, the intersection of community ecology and conservation biology is also briefly discussed.

II. PAIRWISE INTERACTIONS

Species interact in a tremendous variety of ways, and associations can involve two, several, or many species. First, the various kinds of interactions that include only two species are discussed. A discussion of multiple species interactions is provided in Section IV. When individuals of two species interact, this association is called a pairwise interaction. Pairwise interactions, like all interactions, are described by the direction and the strength of their effects. Direction refers to whether the impact on each associating organism is positive or negative, and interaction types are classified according to the direction of their effect (Table I). Interaction strength refers to how much an interaction controls species dynamics.

More specifically, the strength of an interaction is a measure of how much one species affects the population size and growth rate of another species. Strength is often detected as the effect on the population of one species by removing the second species. Imagine two interacting species, species A and B. If we would like to test the effect of B on A, we could remove B and measure the resulting change in population size of A. If species A responds to the removal of species B with a large change (increase or decrease) in abundance or population growth rate, we say that B has a strong effect on A. If little effect on the population of species A is seen by removing species B, the interaction plays a small role in the population dynamics of species A, and the influence of B on A is weak. In turn, the effect of A on B can be tested by removing A and observing how the population size or growth rate of B responds.

Many factors determine how strongly two species interact. For example, the population size of either species, the behavior and age of the interacting individuals, the abundance and availability of abiotic resources, and the patchy nature of the environment all influence how strongly two species affect one another. In other words, understanding the nature of an interaction is more complicated than simply identifying that two species come into contact. In fact, even a simple experiment of interaction strength such as that described previously may not provide enough information on exactly how organisms interact with one another. Furthermore, the strength of an interaction can change at different times or in different places in the species’ habitat. In the long term the strength of pairwise interactions can change due to the forces of evolution, but first the types and effects of pairwise interactions in ecological time are discussed.

A. Competition

Competition is characterized by the common use of resources by two or more species, and the effect of competition on all participating species is negative (Table I). Members of the same or different species can compete for resources. Within-species competition is known as intraspecific, and between-species competition is called interspecific. Of these, interspecific competition meets the definition of a species interaction, and intraspecific competition will not be discussed. Interspecific competition can be further divided into two categories: interference and exploitative. Interference competition is the more direct form of competition, in which individuals of one species actively dominate a resource, preventing or decreasing the access of another species to those resources. Exploitative competition, on the other hand, is less combative. It is the diminishment of a limited resource by a species so that the other species is not able to utilize as much of the

<table>
<thead>
<tr>
<th>Direction of effect on species 1</th>
<th>Direction of effect on species 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>Predation/parasitism</td>
</tr>
<tr>
<td>−</td>
<td>Competition</td>
</tr>
<tr>
<td>None</td>
<td>Commensalism</td>
</tr>
<tr>
<td></td>
<td>Amensalism</td>
</tr>
</tbody>
</table>

TABLE I

Pairwise Interaction Types
resource as it would in the absence of the exploitative competitor.

We can predict how the population dynamics of two competing species will be affected by their interaction with the use of a mathematical model. A mathematical theory of competition began with the work of Lotka and Volterra in the late 1920s and early 1930s. The Lotka–Volterra model quantifies the influence of competition by expressing the population dynamics of each species in an equation that includes both growth in the absence of competition and the influence of competition itself. In other words, Lotka–Volterra represents the change in population size through time as a function of the growth rate of each species and the detrimental effect of competition on each population. Remember that the presence of competitors causes harm. This harm reduces the total number of individuals in each population that would be possible in the absence of competitors.

Briefly, I discuss the components of the Lotka–Volterra model. Two equations are needed to represent the population dynamics of competition (one equation for each species). The pair of equations describes the rate of change in population size with time for the two species. Subscripts represent a variable’s species designation; hence, $N_1$ refers to the population size of species 1, and $N_2$ refers to the population size of species 2. The relative effect of each species on the population size of the other is determined by the value of the competition parameter, $\alpha$. This competition parameter can also be interpreted as the amount of overlap between the two species in the resources they require. The effect of one member of species 2 on one member of species 1 is represented by the parameter $\alpha_{12}$, and the effect of species 1 on species 2 is the value $\alpha_{21}$. The values of $\alpha_{12}$ and $\alpha_{21}$ can range from 0 to 1. If they take on a value close to 1, then the effect of one species on the other is very strong. If they take on a value close to 0, then the effect of one species on the other is weak. The values of $\alpha_{12}$ and $\alpha_{21}$ can be different from each other because it is not necessary that the two species affect each other equally. If the values of $\alpha_{12}$ and $\alpha_{21}$ are the same, this special case is called symmetric competition.

If both $\alpha_{12}$ and $\alpha_{21}$ are zero, then the two species do not compete at all and grow according to a simple expression known as logistic growth. In logistic growth, a population grows from zero to some carrying capacity, $K$, according to an intrinsic growth parameter, $r$, as shown in Fig. 1. The hallmark of logistic growth is its density dependence. In density-dependent growth, the expansion of a population slows as the number of indi-

![FIGURE 1](Sample plot of logistic growth. N is the number of individuals in a population of a single species, and t is time. In logistic growth, the number of individuals increases quickly when the population size is small, but the rate of growth decreases to zero as N approaches the carrying capacity. Here, the carrying capacity is 2000 individuals.)
individually in the population approaches the carrying capacity or the maximum number of individuals that a habitat can support. All of the terms in the following equation for species 1, except the component \( c_1N_1 \), represent the logistic growth of species 1. The remaining term \( a_2N_1 \) reflects the slowing or reduction in population growth of species 1 that results from competition with species 2. The same is true for the equation of species 2; the term \( a_1N_2 \) represents the negative effect of competition with species 1 on population 2’s otherwise logistic growth.

\[
\begin{align*}
\text{Species 1:} & \quad \frac{dN_1}{dt} = r_1N_1\left(1 - \frac{N_1}{K_1}\right) - a_2N_1N_2 \\
\text{Species 2:} & \quad \frac{dN_2}{dt} = r_2N_2\left(1 - \frac{N_2}{K_2}\right) - a_1N_1N_2
\end{align*}
\]

These equations can predict the impact on \( N_1 \) and \( N_2 \) for different values of \( r_1 \), \( K_1 \), \( a_{12} \), \( r_2 \), \( K_2 \), and \( a_{21} \). For some values of \( r \), \( K \), and \( a \), one species will drive the other to extinction. In other cases, however, each species can maintain a positive population size in the presence of each other through time (\( N_1 > 0 \) and \( N_2 > 0 \)). One such case is shown in Fig. 2, in which the population size of both species is plotted against time. To produce this graph, one begins with population sizes just larger than zero for both species and uses the competition equations to predict the change in population sizes \( (N_1(t) \) and \( N_2(t) \) as time passes. Note that in Fig. 2, both species eventually reach a stable and constant population size.

What conditions produce an outcome in which two competing species can coexist? It can be shown that coexistence depends on the relative value of the carrying capacities of the two species in comparison to the values of the competition coefficients, \( a \). In short, coexistence is possible whenever the value of the competition coefficients is less than the ratio of the carrying capacities (if \( a_{12} < K_2/K_1 \) and \( a_{21} < K_1/K_2 \)). This makes intuitive sense: Two species are able to persist on the same resource if they have a relatively small effect on each other. 

---

**FIGURE 2** Sample plot of the population growth of two competing species according to the Lotka-Volterra competition equations. \( N(1) \) is the number of individuals in one species, and \( N(2) \) refers to the number of individuals in the other species. In this diagram, species 1 and species 2 coexist because the abundance of both populations is positive through time. The growth of each population appears logistic (see Fig. 1), but because of competition the level at which each curve stabilizes is less than the carrying capacity for each species. In other words, the number of individuals in a population in a competitive environment is smaller than the total number of individuals that is possible in the absence of a competitor.
one another. In other words, there is a constraint on the amount of resource overlap (\(a_{ij}\) and \(a_{ji}\)) between two competing species that is given by ratios of their carrying capacities.

The idea that two species, based on the Lotka–Volterra equations, cannot significantly overlap in resource use is known as the competitive exclusion principle. Generally stated, no two species can occupy the same niche. (A niche can be imprecisely defined as a species' role in a community or the entire set of resources that a species requires.) If the same niche were filled by two species, or if two species used exactly the same resources, the species would compete strongly, their \(r\) values would approach 1, and one of the two would be driven to extinction.

Additional factors not captured in the Lotka–Volterra equations influence coexistence, however, and species may coexist in situations in which classical theory would predict otherwise. The role that these factors play in coexistence reduces the general applicability of the competitive exclusion principle. Spatial and temporal heterogeneity and variability of resources, for example, may enable a greater diversity of species to persist in an area than simple theory would predict. In other words, the patchy nature of resources in a habitat may enable two species to reduce their overall competitive overlap. Alternatively, species may compete only when resources are in short supply.

Controversy regarding the role of competition in natural communities has resulted in a tumultuous debate in the ecological literature. Some authors argue that competition is a critically important factor regulating community composition. Others contend that competition is overemphasized at the expense of other types of interactions that are equally (if not more) important in determining the number and kinds of species that can persist in a community. This debate as to the importance of competition in controlling community structure is ongoing, but there is little argument that competitive interactions can be found and quantified.

Gause was the first to study competition in 1934 using experiments with two Paramecium species that shared a single resource. Gause found that one species always went extinct when both species were put together, suggesting that, as the Lotka–Volterra theory predicted, coexistence of two competitors is possible only in special cases. In 1946, however, Cromie placed two beetle competitors in a mixed environment of flour (i.e., resources) and small pieces of fine glass tubing (i.e., environmental complexity) and found that the species coexisted. Interestingly, in the absence of tubing coexistence was not possible, and one species always drove the other to extinction. Cromie's experiments suggest that some factors or influences that are not found in the simple Lotka–Volterra model may be important in facilitating coexistence. Since the time of Gause and Cromie, many other ecologists have studied the role of competition in natural ecosystems. In the early 1980s, Schoener and Connell each reviewed the literature for the prevalence of competition and found that a majority of published field studies successfully detected the impact of competition on the population dynamics of species.

### B. Predation

Predation is the consumption of one species by another. Consumption leads to a negative effect on the population of one of the two interacting species, known as the prey, and a positive effect on the second population or species, known as the predator (Table I). Also as in competition, predation has a rich history of mathematical theory, beginning with the work of Volterra. The predator–prey interaction is modeled with two equations—one for the prey (species 1) and one for the predator (species 2). Again, the components of these equations are discussed without going into great detail.

\[
\frac{dN_1}{dt} = rN_1 - aN_1N_2 \\
\frac{dN_2}{dt} = bN_1N_2 - dN_2
\]

According to this model, in the absence of a predator, species 1 would grow exponentially according to the \(rN_1\) portion of the first equation; unlike the Lotka–Volterra competition equations, no carrying capacity limits growth in this model. This type of growth is called density independent. In other words, the rate of increase in the prey population remains constant through time and is independent of the number of individuals in the two populations. (Again, \(r\) represents the intrinsic rate of growth.) The number of prey removed from the prey population by predation is represented by the term \(aN_1N_2\), where \(a\) describes the fraction of encounters between predator and prey that ultimately results in kills. Predators consume prey for a single purpose—to produce new predators. The number of predators born (i.e., new predators added to the \(N_2\) population in the second equation) is a function of how efficient predators are at producing offspring from consuming prey. The variable \(b\) describes this efficiency, and the term \(bN_1N_2\)
thus represents births into the predator population. In this model, prey die only from predation, but predators die from other causes, and this death rate is the variable $d$. The number of individuals removed from the predator population due to death is the term $dN_p$ and the fraction of predators removed is constant through time.

Several outcomes are predicted from this simple model. First, predators can drive prey to extinction by consuming them faster than prey can reproduce. Second, the predators can go extinct if their death rate exceeds their ability to produce new predators from the consumption of prey. Lastly, coexistence is possible.

However, unlike the competition situation discussed previously, these predator–prey equations do not reach a stable and persistent number of predators and prey in cases of coexistence; instead, the abundance of each species cycles through time. These cycles are called oscillations, with the predator abundance “lagging” behind that of the prey. As the number of prey increases, it is followed by an increase in the number of predators, driving a decrease in prey, in turn driving a decrease in the number of predators. Once the predators decrease, the pressure of predation declines and prey can increase again, and the cycle repeats. The cyclic dynamics of predator–prey can be seen in Fig. 3.

To determine whether the oscillations predicted by the Volterra theory reflect the true nature of predator–prey interactions, researchers have sought evidence of population cycles in nature. The most well-known example of predator–prey oscillations was shown in lynx feeding on hare in Canada. Recent evidence, however, calls into question whether or not the oscillations observed in the lynx–hare system are in fact caused by a predator–prey interaction or by some nonpredation factor affecting hares such as seasonal food shortage.

There is agreement that, oscillating or not, predator–prey interactions are common and are often strong. For example, following the extinction of many important large forest predators, in conjunction with human alteration of the forested landscape, eastern U.S. deer populations expanded to extraordinarily high levels. This suggests that predators once played a large role in suppressing the numbers of deer. A second example is the introduction of the brown tree snake to the island.

**FIGURE 3** Sample diagram of a predator–prey system as predicted by the Volterra equations. $N$ is the number of individuals $\times 10^3$ and $t$ is time. The solid line, $N(1)$, is the number of individuals in the prey population, and the dashed line, $N(2)$, is the number of individuals in the predator population. Both predator and prey consist in this diagram because neither goes extinct. The abundance of predator and prey is not stable through time but goes through oscillations or cycles. Notice that increases and decreases in predator population size follow those of the prey.
of Guam. In this case, the snake decimated the island's bird community, driving most of its prey to local extinction.

Several factors that are not included in the Volterra model can influence the dynamics of a predator–prey interaction, increasing the likelihood of coexistence and dampening cycles. Environmental spatial heterogeneity in abiotic and biotic factors can benefit either predators or prey. For example, in some portions of a species' habitat, it may be more difficult (or easier) to find or capture prey. It is also possible that the Volterra model does not reflect the true nature of a specific predator–prey interaction. For example, it may be unrealistic to model some prey species as density independent. Researchers have modified the Volterra model in many ways to increase its realism, including versions that incorporate density-dependent growth of prey.

C. Parasitism and Herbivory

Several other types of interactions have effects similar to those of predation, where one species benefits to the detriment of the other. Herbivory, the consumption of plants by animals, is a subtype of predation but differs in some important ways. First, most herbivores consume only a part of their "prey," and consumption frequently results in death. Second, plants cannot escape their predators as many animal prey do, changing the nature of the predator–prey interaction. Just as predators can greatly affect prey, herbivores play an important role in the dynamics of their host plants. For example, the introduction of the gypsy moth to the eastern United States dramatically reduced the population size of many forest tree species, especially oaks and aspen. And the effects of herbivorous agricultural pests on crops are the impetus for widespread application of pesticide.

An interaction known as parasitism is also distinct from simple predation. Parasites are usually smaller than their prey (also called a "host") and often feed on their hosts without killing them. When parasitic interactions are lethal, parasites kill their host after using or feeding on it for a period of time. Many human diseases such as malaria (a protozoan), intestinal roundworms, and disease-causing trematodes such as liver flukes are parasitic. Viruses can also be considered a type of parasite. One interesting form of parasitism is a phenomenon known as brood parasitism in which one species of bird lays its eggs in the nest of another, leaving the offspring to be raised by a surrogate parent. In this case, the parasite does not consume the host but steals some of its parental care. Brood parasites, like all parasites, reduce the success and reproductive output of their hosts. An often fatal form of parasitism in many insect populations is infection by parasitic wasps or flies called parasitoids. These wasps lay eggs in the body cavity of their host; the parasitic larvae then feed on the host during development, killing it upon emergence. Some types of parasitic interactions may have dynamics similar to those predicted by Volterra predator–prey theory. Recent evidence shows, for example, that cyclic red grouse populations in Britain may be driven by the effects of a nematode parasite on host reproductive ability.

D. Mutualism and Symbiosis

When both participants in an interaction benefit from their association, the interaction is known as mutualism (Table I). Many types of mutualisms can be found in nature, and increasing evidence indicates that positive interactions may be more important than many ecologists previously considered in determining who can coexist in a community. There is no well-developed theory as to how mutualism promotes coexistence or how mutualistic species impact the population dynamics of one another. There is a large and developing empirical literature, however, that documents the presence of mutualism in nature. Many of the common interactions among organisms, such as animal pollination of plants and some animal dispersal of seeds, are in fact mutualisms. Fruit-eating birds, for example, get nutrition from a plant in exchange for dispersing the plant's seeds across a wide area.

Mutualistic interactions can be facultative or obligatory. Facultative interactions are those in which neither species requires the other to persist but each benefits from the interaction when the opportunity arises. The association between a fruit-bearing tree and its avian visitors is facultative. Many species of lycænid butterflies have mutually beneficial, and facultative, associations with ants. The ants are thought to protect the larvae from predators while lycænid larvae offer the ants a nutritious secretion in return.

Obligatory interactions are those in which each species requires the other for survival. These are typically highly specialized species associations. A classic pollination-based obligatory mutualism is that of the fig and fig wasp. Nearly every species of fig is pollinated exclusively by a specialized species of fig wasp, which lives inside the fig fruit. The figs rely on the wasps for pollination and the wasps rely on the fig as a habitat in which to execute their life cycle. Several plant species also have obligate mutualistic interactions with fungi. For example, many species of orchid require a fungal...
association for seed germination. The fungus, in turn, derives nutrients from the orchid. A third example is an obligatory interaction between some species of algae and fungi. The algae and fungus live together, creating an “organism” known as lichen.

The lichen example also illustrates a critical vocabulary point. Lichen is an example of a symbiotic mutualism. The terms “mutualism” and “symbiosis” are often used interchangeably, but they actually differ subtly. The term symbiosis stresses that two species live close to one another in a prolonged association. While many symbiotic interactions can also be mutualistic, this need not be the case. For example, you and the bacteria inhabiting your stomach which aide in your digestion are symbiotic and mutualistic, but you and the flukes feeding in your liver form a parasitic symbiosis.

E. Asymmetrical Interactions

One species can harm or help another species without any benefit or detriment in return. Commensalism refers to the benefit of one species, species A, from the presence of another species, species B, while B experiences no effect from the presence of A (Table I). Conversely, amensalism refers to the detrimental effect of species B on A while B experiences no effect of A in return (Table I). Far less research has focused on commensalism and amensalism than on other types of interactions, and the strength of commensalism and amensalism is generally thought to be weak. A famous example of commensalism is an association between cattle egrets and cattle. The egrets eat insects flushed by the cattle. The presence of the egrets, however, has no measurable effect on the cattle. Amensalism often occurs as the incidental damage to one species from the presence or activity of another. For example, in the cattle–egret example, some ground-dwelling insects suffer incidental mortality from the cows that step on them.

III. INTERACTIONS AND EVOLUTION

Previously, the direction of interaction effects and differences in interaction strength were discussed. It was also mentioned that interactions can change through time and space. In ecological time, interactions between species affect population size. In the longer term, however, interactions can lead to evolutionary change in the two participating species, either reducing or increasing the strength of the association. For example, in competitive and predatory interactions, one species may evolve mechanisms to avoid strong interaction with others in response to the pressures of natural selection.

In mutualistic interactions, on the contrary, specialized behaviors and morphology can be selected through time to promote efficiency, and species can grow to be more dependent on one another. Evolution typically results from a pressure known as natural selection. Natural selection acts in the following way: Among a collection of organisms in a population, those individuals that are best suited for their environment will be the most successful and will produce a greater fraction of a population’s offspring than will other individuals. Successful individuals will genetically pass their beneficial traits to their offspring if these traits are heritable. Through many generations of enhanced survival and reproduction of individuals with beneficial qualities, directional change in physiological, morphological, or behavioral traits at the population level can result. Natural selection can act in a variety of ways. For example, natural selection may favor traits that confer efficiency in food digestion or nutrient acquisition. Interactions between species can also be a force of natural selection for either or both species involved. When natural selection results in reciprocal trait evolution in two interacting species, it is called coevolution. In other words, if morphological, physiological, or behavioral traits determined by genes in two interacting species evolve in response to an interspecific interaction, coevolution has taken place. The pressure of natural selection (called selection pressure) in coevolution stems from the positive or negative effects that species have on each other. Depending on the type of interaction driving coevolution (competition, predation, mutualism, etc.), many trait changes are possible.

A. Evolution and Competition

Remember that competition is thought to limit how similar two coexisting species can be. If, as seen in the Lotka–Volterra model of competition, resource overlap can drive one of two competitors to extinction, then competition can also be a strong selection pressure. Individuals with traits that enable them to compete less with other species will have an advantage compared to other members of their population. Over many generations of selection for minimizing overlap, populations can evolve away from interaction with their competitors.

The evolutionary change in a trait in response to competition is called character displacement. Character displacement results from selection for traits that reduce the overlap in resource use with a similar species. If
two or more competing species evolve in response to one another, character displacement can be considered a type of coevolution. For example, in the Galapagos Islands, a group of coexisting finches have distinctly different body and beak sizes as well as beak shapes. Size specialization allows each finch to exploit a different food resource. The seemingly well-organized categories of finches suggests that there was once a strong influence of competition on these species and that competition drove the finches to evolve different strategies and diverge. In fact, the type of pattern seen among finches on the Galapagos is widespread. Many species of similar type in a single habitat (e.g., desert rodents or island lizards) appear divided into distinct body size categories. This division suggests that selective pressure to reduce resource overlap was once (and may still be) an important influence.

It should be noted that the pattern of body size classes often seen in nature need not necessarily evolve character displacement as an explanation. An alternative hypothesis is the idea of selective retention of invaders. This hypothesis states that as a community is being filled by new arrivals to the system, those invading species that are able to persist are ones whose niche is currently vacant. In other words, invaders who find their niche already occupied are not able to persist in the system and go extinct shortly following invasion. This hypothesis is distinct from the character displacement hypothesis in that no evolutionary adaptation is necessary to explain the morphological differences seen among similar species.

B. Evolution and Mutualism

Change in morphologies in response to competitive pressure as in the Galapagos finches is one type of interaction-driven evolution. A second type can occur in mutualistic interactions. In mutualism, natural selection may favor those individuals that benefit more from an interaction than do other members of the same species. If the partner species in a mutualism is under analogous selective pressure, reciprocal, genetic changes in both species may result. Over many generations, coevolution increases the benefit each species gains from its interaction with the other.

Many plant–pollinator systems appear to have experienced some coevolutionary change. For instance, it may be advantageous for plants to attract pollinators that visit only a few or a single plant species, and many plants attract pollinators with specialized flashy petals and nutritious rewards. Because the next flower visited by such a specialist pollinator is more likely to be a member of the same plant species, a plant that attracts specialists enjoys an increased rate of successful pollination and a decreased expense of excess pollen production. Many pollination associations seem to be coevolved in this way, including hummingbirds with long curved bills feeding on species with long curved flowers and orchid species dispersing fragrances that attract specialized bee visitors.

Other coevolved mutualisms between animals and plants are also known. For instance, some acacia trees have mutualistic associations with ants. An ant colony resides in the tree and uses tree secretions called “nectaries” for food. The acacia uses the ants for protection from other insects and plants. Ants can attack herbivores that land on the plant, and in some cases the ants even remove competing plant material from around the acacia’s base. Studies have shown that mutualistic acacias perform very poorly without ant inhabitants. In effect, acacia success and survival are dependent on the ants. Similarly, several species of ants are found only on acacia, indicating that for some, acacia is an obligatory ant resource.

In both the acacia–ant and plant–pollinator examples, it is difficult to imagine how such specialized associations could evolve without multiple steps of reciprocal morphological and behavioral change. It is thought that specialized interactions evolved from generalized interactions, where survival and reproduction benefits of increased mutualism brought two select species closer together. Such increased association ended with, in some cases, obligatory dependence.

C. Evolution and Predation

Predators can place selection pressure on their prey to avoid capture, and the benefits of efficient catch and consumption of prey can drive the evolution of predators. The evolution of a trait in prey that makes predation more difficult can be followed by the reciprocal evolution of a trait in predators that reduces the prey’s new advantage. Conversely, traits that confer advantages to predators can lead to increased natural selection on prey to evolve predation avoidance traits. Again, if two or more predators and prey evolve in response to one another, this is called coevolution.

Through evolutionary time, many species of prey have evolved defenses against predation, including the ability to flee, avoid detection, or be difficult to eat. Predators have also evolved techniques and physical characters to improve their efficiency or probability of successful capture of prey. Predators can become less conspicuous, faster, or more skilled at capturing and
killing prey. This phenomenon was first discussed by Darwin in the late 1800s; he wrote of a constant struggle between prey to escape their enemies and predators to consume their prey. Today, we refer to the “constant struggle” as an evolutionary “arms race,” with the analogy to nations that reciprocally accumulate weapons to maintain an even balance of defense.

One frequently cited example of a coevolutionary arms race exists among plants and their herbivores. Plants must defend themselves against their predators since they cannot flee, and they often do so with chemicals called secondary compounds. These compounds either have a negative impact on the physiology of herbivores or decrease the ability of herbivores to consume the plant. Studies have shown that many plants increase the tissue concentration of secondary compounds in response to herbivory, suggesting that such compounds are indeed defensive. In addition, the presence of secondary compounds has been shown to effectively reduce the number of herbivores feeding on a plant. Some herbivores, however, have overcome the negative influence of secondary compounds by adapting ways to metabolize or resist the effects of the plant’s defenses. Some herbivore species even use defensive compounds as a cue for identifying the plant as a food source. The pressure to evolve chemicals to escape herbivory, as well as the pressure to overcome antiherbivory compounds in order to sequester food, may have driven the degree of specialization seen in some herbivore–plant relationships.

A second example of a predation-like coevolutionary process can occur between brood parasites and their hosts. Recall that brood parasites lay their eggs in the nests of other species; parasite offspring hatch and are raised by the surrogate parent. Brood parasites place a selective pressure on their host to identify parasitic eggs and either remove them or abandon the nest. However, as hosts become better at recognizing parasitic eggs, parasites are also selected to produce eggs that appear more similar to the host species. Parasites can become specialized on an individual host, producing eggs resembling only a single host species.

IV. MULTISPECIES INTERACTIONS

Limiting the treatment of interactions to two species helps us to classify associations and consider the effect an interaction might have on the evolution of participants. In actuality, however, all organisms live in communities composed of more than two species. This reality is not reflected in the simple theories of competition and predation. Unfortunately, the addition of multiple species to simple models can quickly make them complex and difficult to manipulate. Despite these difficulties, a large body of empirical and theoretical work considers multispecies interactions and the role of multispecies interactions in driving evolutionary change.

A. Food Webs and Trophic Levels

One way to characterize interactions in a community is to document who eats whom. A food chain is a series of prey species and their predators (Fig. 4). Food chains begin with plants (known as primary producers), followed by the animals that consume plants (primary consumers or herbivores), the animals that eat herbivores (secondary consumers), and so on. The collections of all food chains in a community comprise a food web. One can think of a food web as a diagram that describes how all species or groups of species in a community use one another for food (Fig. 5). The positions (producer, secondary consumer, etc.) within a food chain or food web are known as trophic levels. Food webs and trophic levels are useful paradigms for thinking about species interactions; not only do they show the relationships among predators and prey but they also suggest when species might compete or interact in other ways. Because organisms live in a world of multispecies interactions, food webs enable us to count interactions and consider the relative importance of multiple interactions on a single species.

One focus of research in the study of food webs and community composition is the issue of food web...
stability and resistance to disturbance and change. Stability refers to how a food web will respond to the removal, addition, or a large change in the population size of one or more community members. In general, if the extinction of one community member is followed by the extinction of several other members, the web is unstable. If the web adjusts only slightly in response to extinction or large population shifts, it is stable. The issue of stability in food webs is important because the environment is very dynamic, and over long periods of time such as large changes in resources or natural disasters are likely to occur. If a community persists through time, it must be resistant to such extreme events.

In 1972, May was the first to propose that webs with many members and many links were unstable. Using a mathematical model, May showed that the more interactions in a community, the greater the potential for a small perturbation to affect multiple web members and the structure of the web. This result, however, seems counterintuitive because many of Earth’s communities are diverse with many species and many links. Also using mathematical theory, other researchers have shown that in certain cases diversity can lead to stability. For example, food web stability may be related to the asymmetry or symmetry of its interactions or to the location or arrangement of strong interactions in the web. To date, little consensus on the issue of community and food web stability has emerged.

B. Keystone Species

A species that exerts a strong influence on the dynamics of several members of its community is called a keystone species. A keystone species’ influence is one that is larger than we might expect based on its abundance alone. In other words, a keystone species must exert its influence based on the important interactions it has in the community and not simply on its dominance or high population size.

For example, the sea star, Pisaster ochraceus, is a classic keystone species in Oregon and Washington. It preys on several mussel species, including the mussel Mytilus californianus. Mussel species compete strongly for places to adhere on intertidal rocks. Mytilus californianus is highly successful at dominating intertidal spatial resources, blocking out space needed by other mussel species. By feeding on M. californianus, P. ochraceus opens up rock space, resulting in an increase in the diversity of the community. In the absence of P. ochraceus, several species of mussel would be driven to extinction by the dominant competitor M. californianus, but in the presence of P. ochraceus the number of mussel species that can persist in the community is increased. Table II shows some experimental results of removing P. ochraceus from the intertidal zone in Washington.

Other predatory keystones include the sea otter and the bass; however, not all keystone species are found at the tops of food chains. For example, nitrogen-fixing bacteria may be keystone members of forest communities if they increase diversity by reducing the effects of nitrogen limitation. Beavers influence species diversity in their communities by altering the habitat structure of streams and ponds. Because keystones maintain the biodiversity of their communities, identifying and preserving them is an important conservation goal. To date, however, there is no general theory that predicts which members of a community will be keystones. The only way to identify a keystone species is to study its community in detail.

C. Trophic Cascades and Indirect Interactions

Keystone species exert their strong influence on community composition by a combination of direct and indirect effects. Each of the interactions previously discussed in the context of pairwise interactions were di-
### TABLE II
Species Composition in the Presence and Absence of *Pisaster ochraceus*

<table>
<thead>
<tr>
<th>Survey date</th>
<th>July 1963</th>
<th>April 1973</th>
<th>August 1966</th>
<th>June 1971</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plot type</td>
<td>Control</td>
<td>Control</td>
<td>Removal</td>
<td>Removal</td>
</tr>
<tr>
<td>% cover by species</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Balanus cariosus</em></td>
<td>10</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Balanus glandula</em></td>
<td>12</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Chthamalus tissus</em></td>
<td>15</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Pellicipes polypeam</em></td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td><em>Mytilus californianus</em></td>
<td>5</td>
<td>2</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td><em>Endocladia maricata</em></td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Corallina tenuicrustis</em></td>
<td>20</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Lithothamnium sp.</em></td>
<td>16</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Hedophyllum sessile</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Halichondria panacea</em></td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total % cover</td>
<td>89</td>
<td>89</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

* Data from Mukkaw Bay, Washington (reproduced with permission from Paine, 1974).
* Plots were of two types: those in which *P. ochraceus* was removed ("removal") and those in which *P. ochraceus* remained ("control").
* All species are sessile invertebrates or algae that live on the rocks of the intertidal zone.
* Numbers do not necessarily total 100%; remaining space was either unoccupied or held by a few rarer species.

---

rect—two species in contact and each influences the dynamics of the other. An indirect interaction occurs when species influence one another through their interactions with other species. According to a review of indirect interactions by Wooton (1994), five types of indirect interactions have been identified. I focus on three here: apparent competition, interaction modifications, and trophic cascades.

Apparent competition is the case in which two prey species share a single predator. If one of the two prey species were to increase in abundance, an increase in predator population size might follow. An increase in predators can then lead to a decline in both prey species. For example, an increase in the number of one of two grasshopper species has been shown to increase the number of parasitoids infecting both host species. Interaction modification is the case in which one species changes the interaction of two directly interacting species. For example, the amount of phytoplankton (small green plants) in the water column of a lake may affect a predatory fish's ability to find and capture a prey species.

A trophic cascade refers to the indirect effects on many species in a food chain to a change in the population size of a single chain member. Trophic cascades were first discussed by Hairston et al. (1960), and the effect of cascades on population sizes at each level in a food chain was thereafter named the HSS hypothesis. HSS predicts that predators indirectly control the abundance of primary producers by reducing the numbers of herbivores through predation and releasing producers from some herbivory pressure. Predation on herbivores keeps herbivore abundance low, and a smaller number of herbivores consumes fewer producers than they would in the absence of predators. Some data exist to support the HSS claim that predation can limit the consumption of producers. As mentioned previously, following the human-caused extinction of large predators, some herbivore densities (such as that of eastern deer) have greatly increased. The increase in herbivores has in turn increased the amount of herbivory damage to forest plant species and has changed forest plant dynamics.

Trophic cascades can also be expanded to four trophic levels. For example, in lakes with piscivorous (fish-eating) fish, there is a decrease in the number of planktivores (plankton-eating fish), an increase in the number of herbivores, and a decrease in the amount of green phytoplankton in comparison to lakes without piscivores. This alternating suppression and release of predatory pressure might explain why some lakes are green and others are clear; water clarity (i.e., the amount
of phytoplankton) may be determined in part by the number of trophic levels in the lake food web.

D. Multispecies and Evolution

Just as direct and pairwise interactions can influence evolution, species assemblages may influence selection pressures as well. Multiple and indirect effects, however, can select traits in complicated ways. In multispecies assemblages, many interactions can simultaneously place evolutionary pressure on a species. Where these pressures counterbalance, no change in morphology or behavior may result. When the pressures coincide, evolution by natural selection may occur. In general, a trait will be selected if an organism benefits from possessing the trait over the sum total of all its interactions. There are many examples of evolution in the balance of multiple pressures. For example, a mutualistic interaction between a lycaenid butterfly and its associated ants may have arisen in response to the pressures of parasitic infection of larvae by parasitoids.

V. HUMAN IMPACTS ON INTERACTIONS

Many important services and goods used by humans result from species interactions. Services include pollination of plants and the maintenance of soil fertility, and the products of these services provide goods such as agricultural crops, timber, and clean water. Increasing evidence suggests that human influences on the environment are impacting the very interactions that create these useful goods and services. Influences such as global warming, alteration of the earth’s biogeochemical cycles, habitat loss, overharvesting, and pollution can cause disassociations among once-interacting species or change the nature of a service-providing association. Human stresses on the environment can drive local extinction, and via species interactions single extinctions can impact multiple community members. Stresses such as global warming are changing the abiotic environment in a systematic way, forcing local extinctions in some areas and population expansion in other areas. The responses of individual species to these changes are unique, shifts in species ranges can result, and the overlap of species can break down.

For example, historical data show that vegetation responds very slowly to regional temperature change, and rates of plant species migration will be slow in response to global warming. Some insects, on the other hand, have the potential for long-range dispersal and will respond to climatic changes more quickly than will plants. If insects shift faster than plants, some plant populations may be left without pollinators. Furthermore, even if interacting species, such as plants and insects, were able to shift at the same pace, other stresses such as habitat fragmentation stand in their way. If one species is able to shift over a fragmented landscape and another species is not, disassociation can result.

Humans can alter the timing of interactions as well as the ability of species to persist in the same geographic location. For example, evidence suggests that human influences on biogeochemical cycles, including increases in anthropogenic nitrogen deposition, can impact the growth rate of plants, affecting their flowering and development timing (i.e., phenology). Changes in plant phenology may affect the temporal overlap between herbivores and their food plants, possibly leading to herbivore population decline.

Lastly, human domination of the earth has greatly increased the distribution of select species. Many organisms travel with people as they disperse throughout the world, and some of these species have established themselves in locations other than where they originally evolved. These species are often dominant competitors; they spread quickly and displace other organisms. Because they drive native species extinct or harm native populations, such introductions have significantly altered native interactions in many parts of the world. For example, the introduction of European grasses to California has excluded native plants from all but very small relic patches throughout the state. The loss of native plants has in turn affected the viability of many herbivores that are specialized to forage on native California plants.

An alarming number of Earth’s species are going extinct at the hand of human civilization. Identifying and conserving keystone species, as well as predicting and preventing extinction cascades, will be critical to successful mitigation and evasion of further biodiversity loss. To preserve the interactions and communities we need, community ecologists will be asked to provide predictive and restorative information on a great diversity of community and interaction types. Ecologists also will be asked to increase the generality of their findings because species are going extinct more quickly than we can identify taxa and describe communities.

VI. SUMMARY

Species interactions are classified by the direction of their effects and are divided into the direct, pairwise categories of competition, predator–prey, mutualism,
commensalism, and amensalism (Table 1). The strength of an interaction, or the amount an interaction affects the population size of its participants, can be determined by experimentally removing one species and observing the population response of the second species and vice versa.

A simple mathematical theory for competition predicts that the amount of resource overlap between two coexisting, competing species must be small (Fig. 2). A simple mathematical theory for predation predicts that for coexisting predators and prey, population sizes will oscillate through time (Fig. 3). In both cases, theory may oversimplify the opportunities for species coexistence. Factors such as variation in resources, space, and time, as well as the complexity of a habitat, can minimize resource overlap among competing species or dampen predator–prey oscillations. Mutualistic or positive interactions do not have a simple theory for predicting when and where they will occur, but they are common in nature.

Pairwise interactions between species can select for changes in morphological, physiological, and behavioral traits through evolutionary time. Coevolution occurs when there is reciprocal genetic change between two interacting species. Coevolution in competitive systems can lead to character displacement. Coevolution in predator–prey systems can lead to an arms race of capture and escape. And coevolution in mutualistic systems can lead to highly specialized associations between mutually benefiting species.

Communities are composed of many species, and indirect as well as direct interactions control community composition (Fig. 3). Indirect interactions include trophic cascades, apparent competition, and the modification of a direct interaction. A keystone is a species that strongly influences the biodiversity of its community via indirect and direct effects (Table II). Keystones are often, but not always, found at the top of a trophic chain or food web. Both indirect and direct interactions can influence the evolution of community members. Humans are causing widespread extinction of species, and the loss of species is leading to the decline of many goods and services provided by species interactions in natural systems. Human stresses on the environment, such as climate change, alteration of geochemical cycles, and destruction of habitat, can break down species associations because individual species, not communities, respond to changes in the environment.

See Also the Following Articles
COEVOLUTION • COMPETITION, INTERSPECIFIC • MUTUALISM, EVOLUTION OF • NEST PARASITISM • PARASITISM • PREDATORS, ECOLOGICAL ROLE OF • TROPHIC LEVELS

Bibliography
STABILITY, CONCEPT OF

Clarence L. Lehman
University of Minnesota

I. Background
II. Basic Theory of Stability
III. Stability under Changing Conditions
IV. Other Kinds of Stability

GLOSSARY

community The collection of all species living together in a given area. Alternatively, some designated subset, such as the avian community or the vascular plant community.
deterministic system A dynamical system whose detailed future behavior can be predicted, in principle, for all time, assuming perfect knowledge of the system at the present.
dynamical system A set of rules defining how certain variables change with time. Ecological models are dynamical systems representing what are believed to be (vastly more complex) dynamical systems in nature.
dynamical variable A quantity in a dynamical system that changes with time according to the rules of the system (contrast with parameter).
ecological model A vastly simplified mathematical representation of an ecological system intended to capture the full system’s essence. Qualitatively, ecological models are to natural ecological systems as line drawings are to full-color photographs.
equilibrium A condition of stasis in some dynamical variable.
parameter A quantity in a dynamical system that is fixed as part of the rules of the system (contrast with dynamical variable).
perturbation A temporary change in one or more dynamical variables or parameters of a dynamical system due to external factors.
population A collection of individuals of the same species, often interacting ecologically and genetically.
stochastic system A dynamical system whose detailed future behavior cannot be predicted, even in principle, due to random forces inherent in the system.

STABILITY may be defined broadly as the tendency of a system to return to its former state after some disturbance. In the natural world, the term can be applied to the capacity of an ecosystem to resist environmental disturbances. Given the fact that human activity now produces such disturbances on a vast global scale, the ability of ecosystems to remain stable has become an issue of great significance. Current investigators study the interrelationship of stability and biodiversity; i.e., the effect that the biodiversity of an ecosystem has on its stability, and the corresponding role of stability in maintaining the biodiversity of the system.

Falling down once makes a building unstable.
—GERALD WEINBERG (1975)
I. BACKGROUND

A. The Meaning of Stability

Something is said to be stable if its condition tends to remain unchanged despite external influences. The population sizes of two different species may both be relatively constant in a steady, unchanging environment, but such constancy by itself discloses nothing of the stability of the two populations. If both populations were somehow perturbed—for example, by a spring flood eliminating half of each population—and if the first population subsequently recovered to former levels while the second declined to extinction, then the first population is said to have been stable while the second was unstable. Stability is not mere constancy—it also implies an ability to recover from perturbations. The condition that recovers from perturbations need not be as simple as a constant population value. Populations of predators and prey, for example, may have an intrinsic tendency to cycle repeatedly through high and low values, but the cycle may be stable. Suppose the populations of a predator and its prey were both maintained at fixed levels by artificial management, such as by hunting. If the cycle returned after management ceased, then the cycle could be said to be stable, even though neither population is constant during the cycle. The cycle would be an abstract condition to which the populations return after a perturbation. In fact, the condition that recovers from perturbations can be any recognizable property of the system. It could be the number of species in the system, the collective biomass of all species together, the length of the food chain, the average quantity of nutrients leached from the soil, the degree of susceptibility to disease, or many others.

In the natural world, the stability of an object is closely tied to our perception of the very existence of that object. Any property that is unstable will not retain its condition after a disturbance, by the definition of stability. Given that the world is filled with continual disturbance, any property that is unstable is therefore likely to soon change to something else and hence not be observed. For example, if food chains beyond a certain length tend to be unstable in diverse communities, then a very long food chain, should it ever appear by chance combination of species in an ecosystem, will eventually collapse to a length that is stable.

As human domination of global biogeochemical systems increases—including eutrophication of habitats, changes in atmospheric gases, fragmentation of habitats, distribution of toxic organic compounds, translocation of species, and a general reduction of biodiversity—the resulting perturbations test the stability of ecosystems. Also, as global changes take effect, induced changes in the systems alter the stability of their parts. Stability in ecosystems is therefore not just an abstract concept but something deeply connected to the persistence of the services on which all living beings, humans included, depend.

B. The Diversity–Stability Question

In the 1950s and 1960s, investigators such as Charles Elton argued that ecosystems containing more species would be more stable—less subject to fluctuations due to the myriad forces acting on them—owing to such factors as the greater number of pathways for energy to flow through them. This is the diversity–stability hypothesis. The hypothesis was generally accepted in the 1960s, although not without controversy.

Theoretical studies on the topic in the 1970s shook this general acceptance. Working with mathematical models of ecological communities, Robert May and others showed that such communities with more species were less stable: Populations of individual species returned to normal after a disturbance less rapidly as the number of species in the community increased. Although May (1974) later pointed out that certain ecosystem properties, such as total biomass of the community, can be more stable than biomasses of the constituent species, and despite objections by McNaughton (1977) and others to rejecting an accepted hypothesis based solely on theory, the acceptance of the diversity–stability hypothesis waned. During the 1970s and 1980s, ecologists generally expected that greater biodiversity would reduce stability, or at least that there was no consistent connection between biodiversity and stability.

By the 1990s, experimental evidence had accumulated that seemed to support both views simultaneously, depending on the level of focus. As the number of species increased, population densities of individual species became more variable from year to year but ecosystem parameters such as total productivity and nutrient leaching became less variable. In other words, the system seemed less stable when viewed in detail, from the perspectives of individual species, but more stable when viewed in total, from the perspective of the ecosystem (Tilman, 1999) (see Section III.A). At the end of the twentieth century, issues concerning the effect of biodiversity on stability, and of environmental stability on biodiversity, were still being resolved, but greater biodiversity could be seen to stabilize certain important properties of the ecosystem as a...
C. Aspects of Stability

The seemingly simple idea of whether a system returns to its former condition after a disturbance has several aspects. Which is relevant depends on the ecological question at hand:

1. Does the system return to its former state after a disturbance, given unlimited time without further disturbance? This is the basic idea of stability, and if the system does return from all small disturbances, that state is said to be asymptotically stable or, commonly in the ecological literature, simply stable.

2. How much does the system fluctuate under variable conditions? This is called variability and it is typically measured statistically, for example, by the standard deviation of a time series, by the standard deviation divided by the mean, or by the reciprocal of that quantity. (The standard deviation being the square root of the average squared deviation from the mean.)

3. How small does the disturbance have to be in order that the system return? The set of all possible disturbances that allow the system to return to its former condition is called the domain of attraction. As the domain of attraction becomes vanishingly small, the system becomes unstable. Small domains of attraction correspond to fragile systems, whereas large domains of attraction correspond to robust systems.

4. How much is the system changed by a given disturbance? The more a given condition is changed, the lower the resistance associated with that condition. For example, suppose population values in a given area do not change much when average temperature increases, but may change considerably when rainfall does so. Then the system would be said to be resistant to temperature changes but not to rainfall changes.

5. How fast does a perturbation decay? Resilience is the amount of time needed for a perturbation to be reduced to a specified fraction of its initial size. There has been variation in the use of this term in the ecological literature. Before the 1980s, resilience was often simply referred to factors that reduced the chance that one or more species would become extinct (e.g., the size of the domain of attraction and the populations densities in that domain).

6. How long does the system take to return after a given disturbance? This is related to both how far the system is perturbed by the disturbance (resistance) and how fast the perturbation decays (resilience). The lower the resistance and resilience, the longer it takes to return.

7. How long can the condition be expected to last? The longer on average it lasts, the greater its persistence. The idea of persistence applies to systems subject to random variations, where the chance of encountering dangerously large variations during a time interval increases as the interval grows longer. The longer the time interval, for example, the greater the chance that one or more species will be driven extinct by random events.

8. Does the condition remain intact as the parameters of the system change slightly? If it does, the system is structurally stable with respect to those parameters. For example, if precise birth and death rates of the constituent species change slightly occasionally, does the condition remain?

9. Does the condition remain intact under evolutionary pressures? If it does, the condition is evolutionarily stable. A condition may appear stable over short time, ignoring evolution, but may change on longer timescales. Indeed, over the course of geological time, no ecosystem is invariant.

10. Does the condition depend on spatial scale? What is stable in a small area may be unstable over a larger region and vice versa.

Pimm (1984) enumerated five definitions of stability in the ecological literature at three levels of complexity and three levels of organization—potentially 45 different nuances to the meaning of stability. Hence, care is necessary in defining and applying the concept.

II. BASIC THEORY OF STABILITY

A. Equilibrium

The basic concepts of stability become precise when considered in the abstract. The many complexities in natural ecological communities, including factors that cannot be completely understood, mean that any abstraction of an ecological system is necessarily a vast reduction. To start this reduction, suppose the population increases solely by births and decreases solely by deaths. That is, processes such as immigration are not active. Suppose the rate of birth is highest and death is lowest when conditions are least crowded. As the population increases, resources become more limited—less light, reduced nutrient levels, less food, and so
STABILITY, CONCEPT OF

Figure 1. Geometric interpretation of stability. Solid and open circles represent stable and unstable equilibria, respectively. (A) Per capita birth and death rates are functions of population size; (B) growth of the entire population as a function of population size; (C) populations approaching carrying capacity with the passage of time, for the functions of A and B. The horizontal axis is time; the vertical axis is the population at the corresponding time. Cross symbols at the far left indicate the starting points. Curve 1 starts from a small population near zero. Curve 2 starts above the carrying capacity. Curve 3 starts below the carrying capacity but well above zero. (D) Growth of a population subject to the Allee effect, where 0 is a stable equilibrium.

Forth; the birth rate decreases and the death rate increases. At some point, the population will increase enough so that births just balance deaths. This occurs at the carrying capacity, at which an individual, on average during its lifetime, just replaces itself with a single offspring. Above this carrying capacity, deaths exceed births, so an individual on average does not quite replace itself. In the simplest case, births decrease and deaths increase along straight lines as the population increases (Fig. 1A). The net rate of change per individual (the difference between births and deaths) is largest at the far left in Fig. 1A (r) but decreases to zero when the population size reaches the carrying capacity (K) and becomes negative when the population size is greater than the carrying capacity.

The rate of change of the entire population is simply the rate of change per individual times the number of individuals in the population. This is graphed in Fig. 1B, which shows the growth rate of the entire population (vertical axis) as a function of the population (horizontal axis). If the population is nonexistent (zero), then of course population growth is zero; with no individuals there are no births and can be no deaths. If the population is at carrying capacity, population growth is also zero; births balance deaths. These are two population equilibria—places where the population remains constant.

Between these two equilibria, the population grows. In this simplest case, in which the birth and death rates are straight lines, the population reaches the greatest rate of growth at half its carrying capacity. (In economic models, this is called the point of maximum yield.)

B. Geometric Interpretation

1. Single Species

To understand the stability of an equilibrium, the properties of the growth rate are first examined in the near neighborhood of the equilibrium. Consider the equilibrium at the carrying capacity (solid circle in Fig. 1B at position K). The graph representing population change crosses the horizontal axis precisely at the carrying capacity, which means the population neither grows nor declines there. At that point the growth curve slopes downward to the right (the slope is negative). If the population for any reason falls below the carrying ca-
pticy (e.g., point b in Fig. 1C), the growth function is above the axis, meaning that population growth is positive or that births exceed deaths. Thus, the population increases toward the carrying capacity. On the other hand, if the population is above the carrying capacity (e.g., point c in Fig. 1B), the growth function is below the axis, meaning that population growth is negative or that deaths exceed births. Thus, the population decreases toward the carrying capacity. In both cases, an external change moving the population away from its carrying capacity induces population growth or decline in exactly the right way to counteract the external change. The population returns to the carrying capacity. That equilibrium is stable.

Now consider the equilibrium at 0 (open circle in Fig. 1B). As before, the graph representing population growth touches the horizontal axis precisely at 0. However, at this point the curve slopes upward toward the right (the slope is positive). Therefore, if the population ever increases, however slightly, above zero (e.g., point a in Fig. 1B)—due, for example, to the arrival of propagules from some outside source—the growth function is above the axis, meaning that population growth is positive or that births exceed deaths. Thus, the population increases away from 0. The zero equilibrium is unstable. Thus, internal dynamics carry these populations toward their carrying capacity (which is called an attractor) and away from 0 (called a repellor). If the population starts near zero, but not precisely at zero, it may linger a long time at low values and then begin a rapid growth phase before leveling off toward its carrying capacity (Fig. 1C, curve 1). If the population starts above the carrying capacity, it rapidly decays back (Fig. 1C, curve 2). If it starts below the carrying capacity but well above zero, it can increase to the carrying capacity without the long-lag period (Fig. 1C, curve 3).

2. Local versus Global Stability

In the previous example, the carrying capacity is said to be a global attractor or globally stable because almost all replicate populations eventually arrive there (all but those starting at zero in this case). Because the equilibrium at 0 is unstable, the population has a level of permanence; if driven to low values, but not completely to zero, it can recover spontaneously. This contrasts with populations operating under an Allee effect, wherein the equilibrium at 0 is an attractor. Figure 1D depicts the growth function in such a situation. Very low densities inhibit reproduction, for example, by reducing chances of encountering a mate. There are three equilibrium points—at 0, A, and K. Slopes both at 0 and at K are negative, meaning that these equilibria are stable, whereas the slope at A is positive, meaning that equilibrium is unstable. If such populations are driven to a low enough level (below A), they will spontaneously become extinct.

In this case, both 0 and K are local attractors (locally stable), but neither is a global attractor (neither is globally stable). If the system is pushed a small distance away, it will return to its former state. However, if pushed too far (beyond the domain of attraction of that equilibrium), the system will switch to another state. This is a bistable system. The region between 0 and A is the domain of attraction of the 0 equilibrium, the region between A and K and the entire region above K make up the domain of attraction of the carrying capacity. There is no domain of attraction for the equilibrium at A because this equilibrium is unstable (it is a repellor).

3. Multiple Species

The geometric example of Fig. 1 represents a single species in isolation. In this case, geometric stability arguments are direct and intuitive. The geometric arguments are similar for more than one species, but instead of curves representing growth rates, surfaces in multidimensional space are used. Each additional species requires another dimension. Unfortunately, such surfaces are difficult to visualize for more than two species.

The growth surface for a single species will in general cut through the plane of zero growth (analogous to the horizontal axis in Figs. 1B and 1D) along a curve. This curve is called the zero net growth isocline, and each species in general will have its own such curve. If the zero net growth isoclines for many species meet at a single point, then this is a multispecies equilibrium point. For this equilibrium to be stable, however, it is not sufficient that the growth surfaces of all species, considered separately, have negative slopes there. Interactions among the species must be accounted for, and this is best done algebraically.

C. Algebraic Methods

The geometric interpretation of stability, as depicted in Fig. 1B, has a direct algebraic interpretation in terms of derivatives and eigenvalues, which correspond geometrically to slopes. Where an algebraic description of the ecological system is available (i.e., in ecological models), the algebraic method is widely used and can be followed as a recipe. The algebraic method is essential for theoretical work in ecology, but the material in
Equilibrium is negative (i.e., if \( \frac{dN}{dt} < 0 \)). This may be measured in individuals, biomass, or other units. The rate of change of this population is represented by \( \frac{dN}{dt} \), which tells how large an increment occurs in the population during a small increment of time (\( dt \)). Assuming that this rate of change is a function of the population size, as in the previous discussion, then the system will be described by \( \frac{dN}{dt} = f(N) \). For example, in Fig. 1, \( f(N) = rN(1 - N/K) \), which is the well-known logistic equation described in most introductory ecology texts.

Equilibria are commonly designated by symbols like \( N^* \) (pronounced "hat \( N \)"). They occur at such \( N \) where \( f(N) = 0 \). If the slope of the function \( f(N) \) at any such equilibrium is negative (i.e., if \( df/dN|_{N^*} < 0 \), as shown by the solid circles in Figs. 1B and 1D), then the equilibrium is stable. If the slope is positive (as shown by the open circles in Figs. 1B and 1D), then the equilibrium is unstable. If the slope is zero, then further information is needed to determine stability. In the case of zero slope, the equilibrium may be stable, unstable, or neutrally stable, depending on the exact shape of the function near the equilibrium. In the ecological literature, neutral stability means that if the system is perturbed from equilibrium a small amount, it neither returns nor moves further away but rather maintains its new value. (Note that there is a difference between the way ecologists use the word "stable" and the way mathematicians do, as summarized in (Table 1).

### TABLE 1

<table>
<thead>
<tr>
<th>Term</th>
<th>Ecological literature</th>
<th>Mathematical literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotically stable</td>
<td>Decreases</td>
<td>Decreases</td>
</tr>
<tr>
<td>Stable</td>
<td>Decreases</td>
<td>Does not increase</td>
</tr>
<tr>
<td>Unstable</td>
<td>Increases</td>
<td>Increases</td>
</tr>
<tr>
<td>Neutrally stable</td>
<td>Remains unchanged</td>
<td>(Typically not used)</td>
</tr>
</tbody>
</table>

This section is not essential for an intuitive understanding of the concept of stability.

1. **Single Species**

Consider an ecological system of a single species in which the symbol \( N \) represents the population size. This may be measured in individuals, biomass, or other units. The rate of change of this population is represented by \( \frac{dN}{dt} \), which tells how large an increment occurs in the population during a small increment of time (\( dt \)). Assuming that this rate of change is a function of the population size, as in the previous discussion, then the system will be described by \( \frac{dN}{dt} = f(N) \). For example, in Fig. 1, \( f(N) = rN(1 - N/K) \), which is the well-known logistic equation described in most introductory ecology texts.

Equilibria are commonly designated by symbols like \( N^* \) (pronounced "hat \( N \)"). They occur at such \( N \) where \( f(N) = 0 \). If the slope of the function \( f(N) \) at any such equilibrium is negative (i.e., if \( df/dN|_{N^*} < 0 \), as shown by the solid circles in Figs. 1B and 1D), then the equilibrium is stable. If the slope is positive (as shown by the open circles in Figs. 1B and 1D), then the equilibrium is unstable. If the slope is zero, then further information is needed to determine stability. In the case of zero slope, the equilibrium may be stable, unstable, or neutrally stable, depending on the exact shape of the function near the equilibrium. In the ecological literature, neutral stability means that if the system is perturbed from equilibrium a small amount, it neither returns nor moves further away but rather maintains its new value. (Note that there is a difference between the way ecologists use the word "stable" and the way mathematicians do, as summarized in (Table 1).

### 2. Multi-Species

Multiple species are represented by multidimensional dynamics, in general described by \( m \) equations of the form \( \frac{dN_i}{dt} = f(N_1, N_2, N_3, \ldots, N_m) \). One equation for each species. Equilibria are multidimensional points \( (N_1, N_2, N_3, \ldots, N_m) \) that cause all \( dN_i/dt \) to be simultaneously zero. Some of the \( N_i \) may be zero, in which case fewer than \( m \) species coexist at that equilibrium.

The slope of a system in one dimension generalizes to eigenvalues of the Jacobian matrix in multiple dimensions. (In one dimension, the slope is a special case of an eigenvalue.) To determine stability of an equilibrium, (1) construct the Jacobian matrix of the system, \( J = \{ \frac{df}{dN_i} \} \), (2) substitute the equilibrium value \( (N_1, N_2, N_3, \ldots, N_m) \) into the matrix of step 1, and (3) determine the eigenvalues of this matrix. Numerical software libraries like MATLAB or Mathematica can be used to calculate eigenvalues for full details.

The resulting eigenvalues are complex numbers, and as such may contain both real and imaginary parts. Only the real parts of the eigenvalues affect stability. If the real parts of all eigenvalues are negative, then equilibrium point is stable. If one or more real parts is positive, then equilibrium point is unstable. If none is positive but one or more real parts are zero, then further information is needed to determine stability. In some cases, it is possible to determine stability without further information. (That is, an eigenvalue has the form \( a + bi \), where \( a \) is the real part, \( b \) is the imaginary part, and \( i = \sqrt{-1} \). The real part determines the rate of approach to or retreat from equilibrium after perturbation. In one dimension, the imaginary part is always zero, but in higher dimensions it need not be. When it is not zero, the approach to or retreat from equilibrium follows a spiral path.

Most ecological models are more naturally defined on a discrete time axis. For example, insects may emerge together at a specific time each year or seeds may be set once per year in the fall. In such models, time is better represented not as a continuum but rather as a series of integers (1, 2, 3, ...), for example, representing a series of years. In simplest form, discrete-time models have the structure \( N(t+1) = f(N(t)), N_i(t+1) = f(N_1(t), N_2(t), \ldots, N_m(t)) \), where \( N_i(t) \) is the population level of species \( i \) at time \( t \). To determine stability of an equilibrium in a discrete-time system, the three steps outlined previously are followed, but the resulting eigenvalues must be interpreted differently. Again, the eigenvalues are complex numbers. In the discrete-time case, it is
the absolute values of the eigenvalues that determine stability. That is, it is the positive square root of the sum of the squares of the real and imaginary parts. (For the eigenvalue \( \lambda = a + bi \), \( |\lambda| = \sqrt{a^2 + b^2} \).) If the absolute value of all eigenvalues is less than 1, then the equilibrium is stable. If the absolute value of one or more eigenvalues is greater than 1, that equilibrium point is unstable. If none is greater than 1 but one or more absolute values are precisely equal to 1, then further information is needed to determine stability, as in the case of zero eigenvalues discussed previously.

4. Strength of Return

The magnitude of the eigenvalues (or, equivalently, the steepness of slope) determines the resilience, or the strength of return to equilibrium. If the eigenvalues are all very negative, the corresponding slopes will be very steep. Small deviations from equilibrium will result in large rates of growth or decline, and the system will return rapidly. On the other hand, if any of the eigenvalues is negative but close to zero, then the corresponding slopes will be shallow. Small deviations from equilibrium will not lead to rapid recovery. Of course, given a sufficiently long time, and given the absence from disturbance, the system will return regardless of the eigenvalue’s magnitude, given that it is negative. However, over short times, or with repeated disturbances, the magnitude of the eigenvalues is an important part of stability.

III. STABILITY UNDER CHANGING CONDITIONS

A. Biodiversity and Temporal Stability

If a system is subject to repeated disturbances, variations in a given property of that system over time will be related to the asymptotic stability of that property. Because asymptotic stability is an abstract concept, well defined in ecological models but difficult to determine in natural systems, various measures of fluctuations are often used instead to quantify stability of natural systems.

Suppose two populations are fluctuating through time about their individual mean abundance levels due to stochastic effects in the environment and possibly due to internal dynamics as well. Suppose fluctuations in the first population are on average relatively small compared to the first population’s mean abundance, whereas fluctuations in the second population are noticeably larger compared to the second population’s mean abundance (Fig. 2). Independent of the idea of asymptotic stability, the first system can be defined to be more stable with respect to constancy in abundance. Precisely how this fluctuational stability is related to the concept of asymptotic stability discussed in Section II is a separate theoretical issue.

The average fluctuation can be quantified in alternative ways; for example, as the average absolute value deviation from the mean or as the standard deviation of the mean. In any case, if fluctuations are large relative to the mean, the system is likely to obtain values very far from the mean. Also, if the underlying dynamics have a domain of attraction that is responsible for maintaining the mean, then large fluctuations are more likely, on rare occasions, to combine in the wrong direction and push the system out of that domain, inducing a switch to a new mode of behavior. Thus, the relative amount of fluctuation can carry information on the long-term persistence of the system.

MEASURES PROPORTIONAL TO THE COEFFICIENT OF VARIATION—THE STANDARD DEVIATION OVER A TIME SERIES DIVIDED BY THE MEAN OF THAT TIME SERIES—HAVE BEEN USED AS A MEASURE OF FLUCTUATION. THESE ARE ACTUALLY PROPORTIONAL TO INSTABILITY BECAUSE LARGER VALUES IN THE COEFFICIENT OF VARIATION CORRESPOND TO LOWER STABILITY. THE RECIPROCAL OF THE COEFFICIENT OF VARIATION (I.E., THE MEAN DIVIDED BY THE STANDARD DEVIATION, \( \mu/\sigma \)) HAS BEEN TERMED TEMPORAL STABILITY (TILMAN, 1999). IT CARRIES THE SAME INFORMATION, BUT LARGER VALUES OF TEMPORAL STABILITY CORRESPOND TO GREATER STABILITY OF THE SYSTEM. IN THIS FORM (\( \mu/\sigma \)), IT IS REMiniscent OF SIGNAL-TO-NOISE RATIO IN ENGINEERING. CLEARLY, TEMPORAL STABILITY IS GREATER IF THE MEAN IS GREATER (\( \mu \), THE NUMERATOR), IF THE STANDARD DEVIATION IS
smaller (or, the denominator), or both. Therefore, any forces that tend to increase the mean or decrease the variation will increase temporal stability. Temporal stability is maximal (infinite) when the standard deviation is zero—when there is no variation at all.

Evidence from both experiment and theory suggests that the temporal stability of certain community characteristics of competitive plant communities tends to increase as the number of species they comprise increases (Tilman, 1999). Along a long-term nitrogen fertilization gradient, the number of species varied, changing from diverse prairie communities at low nitrogen levels to only a few agricultural grass species at high nitrogen levels. Individual species abundances varied from year to year due to many causes, both environmental and ecological. As the number of species increased, (i) the total biomass of the community tended to increase and (ii) the total biomass of the community became less variable. Simultaneously, the biomass of an individual species tended to become more variable relative to the mean biomass for that species embedded in the community. In other words, increased biodiversity appears to stabilize the community while destabilizing the individual species.

Given this pattern from experimental systems, a theoretical question immediately arises: Do standard ecological models predict similar phenomena? This is a rapidly developing area, but initial results appear to indicate that they do (Tilman, 1999). In a model of competition for a single resource, and in other more general models, temporal stability of total community biomass increased steadily and linearly as the number of species increased (Fig. 3A). Temporal stability of individual species biomass decreased sharply as biodiversity increased from one to about five species, and then it leveled off at higher diversities (Fig. 3B). In other words, simple ecological models appear to predict the effects observed along the nitrogen gradient—increased biodiversity will stabilize the community but destabilize individual species.

The importance of various possible causes of stabilization of community properties remain under discussion. When abundances of individual species are perturbed by complex and effectively random forces, some abundances may be perturbed to higher levels and others to lower levels. On statistical grounds, the variation in the total biomass will likely contain both positive and negative perturbations. Hence, the total is likely to fluctuate less, relative to its mean value. This has been called the portfolio effect (Tilman, 1999) by analogy to economics, wherein a portfolio diversified over many investment instruments will fluctuate less than one containing only a few. It depends on species responding differently to perturbations, if only for statistical reasons. Effects of competition allow species to expand in abundance when a competing species is suppressed by external forces. This expansion, in turn, buffers any change in total biomass. This is called the covariance effect because it is evident when there is strong negative covariance between pairs of species. Finally, increased biodiversity also often leads to increased total abundance, and this increases stability by making any given level of fluctuations smaller in proportion. This is called the overyielding effect.

Regardless of the source of the stabilization, however, it appears that certain composite community properties may be stabilized by increased biodiversity. More direct experiments that will eventually be able to clarify the effects of biodiversity on stability and other ecosystem properties have been established in many countries (Hector et al., 1999; Tilman, 1999). Results of this coordinated set of experiments will emerge during the
first decade of the twenty-first century and bear watching for information they will provide on the diversity–stability question.

B. Temporal Niches

Temporal stability, as defined previously, certainly applies to a system continually perturbed from equilibrium. However, equilibrium was not part of the definition—only the mean and the variation about the mean participated in the measure of temporal stability. Temporal stability therefore applies to cases in which equilibrium, stable or otherwise, may not even exist.

Changing environmental conditions, such as variation in average rainfall from year to year, can favor different species at different times. This, in effect, can partition resources among species, setting up temporal niches and permitting long-term coexistence of species that greatly outnumber the resources they consume (Chesson, 1994). Without changing conditions, biodiversity would be reduced—coexistence would be limited by the number of resources available.

For a given regime of environmental fluctuations, and for a given community, each species will have some pattern of fluctuations in response to the fluctuations in the environment and those of other species in the community. Again, the issue with respect to biodiversity is how the relative size of the fluctuations—the temporal stability—of each species depends on the number of species in the community and how community-level properties such as total productivity or nutrient leaching depend on the number of species.

Coexistence among species in the absence of a multispecies equilibrium can result from stochastic variations in the environment, as explained previously. However, it is not necessary that the variations be stochastic, nor that their source be the external environment. The variations could result from some regular, well-determined periodic changes in the environment, or they could result from population cycles set up by dynamics of the ecological community (Armstrong and McGehee, 1980; Huisman and Weissing, 1999). Whether the effects are external or internal, stable coexistence at increased biodiversity can exist under fluctuating conditions.

C. Emerging Deterministic Stability

Stable deterministic characteristics such as those described in Section II can emerge from stochastic systems. Individuals may live and die, populations may fluctuate in abundance, and some species may become extinct while others appear; however, amid all this complexity, simple patterns of biodiversity can emerge. One of the early successful theoretical explanations of biodiversity concerned island biogeography (MacArthur and Wilson, 1967). On islands, simple deterministic characteristics emerge amid complex ecological change. It is a conspicuous fact that oceanic islands have fewer species than adjacent mainlands. This effect can be attributed, to a large extent, to a stable balance between local extinction of resident species and immigration of new species.

A new individual arriving on the island may be a member of a species already resident on the island, or it may represent a novel species. If the island is devoid of life, the new individual is certain to be a member of a novel species. As the island becomes more populated with mainland species, the chance that the individual is a member of a novel species decreases. Similarly, the chance that an existing species vanishes from the island increases as the number of resident species increases.

Such considerations lead to immigration and extinction curves such as those in Fig. 4. Notice the qualitative similarity of the immigration and extinction curves in Fig. 4 to the birth and death curves, respectively, in Fig. 1. The immigration and extinction curves intersect at an equilibrium point, where resident species becoming extinct are balanced by new species arriving. Note that at this equilibrium, neither population values nor
community composition are constant. In fact, the individual species making up the community are constantly in flux. It is simply the biodiversity that remains constant.

Is this biodiversity equilibrium stable? The net rate of increase or decrease in the number of species is simply the difference between the immigration and extinction rates, which is the solid curve in Fig. 4. Unlike the curves of Fig. 1A, this net rate represents the entire island, it is not a per capita or per species rate. Therefore, its slope at the equilibrium directly corresponds to stability—it need not first be multiplied by the number of species on the island. Because the slope is negative, by the arguments of Section II this equilibrium is stable.

D. Stability Amid Chaos

Consider a population that changes deterministically according to the rules of Box 1—a straight-line relative of the logistic equation of Fig. 1. Populations oscillate chaotically between high and low values, but except for certain infinitely rare starting conditions they never return to a previous state (Fig. 5). This system is chaotic in the sense that slight deviations are always magnified. In almost all cases, two slightly different starting populations, no matter how nearly identical, grow increasingly different with time.

However, amid such complete chaos can be stability. After sufficient time, population values fall into a pair of disjoint intervals, shaded in Fig. 5. This pair of intervals is an attractor. The population then alternates regularly between the intervals, though its position within either interval cannot be predicted for very long from measurements of finite precision. This is a deterministic system with dynamics that appear superficially stochastic, but among all its instabilities arises another level of stability.

Box 1

\[
N_{t+1} = \begin{cases} 
2N_t & \text{if } N_t < a \\
2a(1-N_t)(1-a) & \text{if } N_t \geq a
\end{cases}
\]

Piecewise linear discrete-time population growth.

If \( a < 1/3 \), there is a stable equilibrium at \( N = 2a(1-a) \). If \( a > 1/3 \), there are no stable equilibria or cycles (see Fig. 5).

IV. OTHER KINDS OF STABILITY

A. Structural Stability

In the cases discussed previously, the ecological system did not change. Dynamical variables such as population levels were perturbed, but ecological parameters such as birth rates remained fixed. Ecological parameters, however, constantly change: Global temperatures increase, glaciers retreat, and spruce forests give way to pine and then hardwoods. Rainfall, soil substrates, and a host of environmental conditions alter the parameters and the structure of ecological systems. Hence, not only may the dynamical variables such as population abundances be perturbed but also the very structure of the ecological system may change.

If a given property of a system persists under small changes in the system itself, then the system is said to be structurally stable with respect to that property. The equilibrium carrying capacity in the system of Fig. 1, for example, is structurally stable. Changes in the slopes of the birth and death rates, or small changes that make the birth and death rates curves rather than straight lines, still leave a carrying capacity that all nonzero initial populations eventually reach, provided that the net per capita growth declines smoothly with increasing population. The precise size of the carrying capacity may change, but the fact that it exists and is stable does not.

1. Classical Predator–Prey Systems

Some ecological models are on a razor edge of structural instability. In the simplest form of the classical Lotka–
STABILITY, CONCEPT OF

Volterra predator–prey system, included in most introductory ecology texts, the prey population is limited by predation and the predator population is limited by availability of prey (first pair of equations in Box 2). Predator and prey populations oscillate indefinitely about an equilibrium, with the predator population lagging behind that of the prey (Fig. 6A). This is a common characteristic of predator–prey systems or, more generally, of producer–consumer systems. However, this particular system has a peculiar property for an ecological system. It possesses a “memory” of past events. If something perturbs either population, or both populations, to a new level, that new level will be revisited on each subsequent cycle. This system has no asymptotically stable behavior; instead, everything is neutrally stable.

Such neutral stability would not necessarily be pathological in an idealized physical system, such as a perfect harmonic oscillator, but here it is pathological. All abstractions are simplifications, and here one of the simplifications is the assumption that the prey population is limited only by the predators; in the absence of predators, prey can increase in numbers without bound. Inclusion of any carrying capacity for the prey, no matter how large or small (as in the second pair of equations in Box 2), changes the dynamics completely. The equilibrium becomes stable, oscillations die out, and memory of past perturbations fades with passing time (Fig. 6B). Another simplification is that predators are never satiated. They consume all the prey they encounter. Inclusion of a more realistic predator response can make the equilibrium unstable. The oscillations converge to a stable cycle of fixed amplitude, again with memory of past perturbations fading.

Thus, this simplest Lotka–Volterra formulation is structurally unstable with respect to important ecological factors. Structurally unstable ecological systems do not commonly appear in nature, so conclusions drawn from structurally unstable models must be used with caution.

2. Related Effects of Eutrophication

The change in structure from stable equilibrium to stable limit cycle, as described previously, may be induced by changes to parameters of the ecosystem. Eutrophication is enrichment by high levels of nutrients, such as increased phosphorus in a lake or nitrogen in the soil. Eutrophication is one of the principal effects of human domination of ecosystems, and it may affect both their biodiversity and their stability.

Graphical analysis by Rosenzweig and others in the 1970 (Rosenzweig, 1990) showed that stable equilibria in producer–consumer systems were favored when the carrying capacity of the consumer was relatively small. If the carrying capacity of the consumer were higher, the equilibrium would be less stable (less resilient). At a sufficiently high carrying capacity, the system would pass through a structural instability and the equilibrium would lose its stability entirely. Oscillations in producers and consumers would then occur. At even higher
carrying capacities, the size of the oscillations would increase, driving both producer and consumer periodically to low population levels, thereby increasing the chance of extinction.

Now, carrying capacity is directly related to the level of resources available. Eutrophication increases these resources, thereby increasing the carrying capacity of the producer. Ironically, the act of providing more food or resource to the producer can lead to its destruction as a result of induced instability. This effect has been called the paradox of enrichment. A common result in observed fertilized terrestrial systems is a reduction in the number of species. Although there can be many reasons for this, loss of stability upon enrichment has been suggested as a contributor (Rosenzweig, 1990).

B. Spatial Stability

A property that is stable at one spatial scale need not be stable at larger or smaller scales (Levin and Segel, 1985). This was decisively demonstrated in the mid-twentieth century by the mathematician Alan Turing. Turing’s simplest example was a (nonecological) dynamical system operating in two separate cells, optionally with some migration between the cells (Box 3). If the two cells are completely disconnected, then there is an asymptotically stable equilibrium that both cells approach. Thus, if the density in either cell is perturbed, it will return to its equilibrial value (Fig. 7, times 0–4).

Box 3

Turing’s basic example of diffusive instability. Variables $x_i$ and $y_i$ are densities in cell $i$. The remaining symbols are constant coefficients, $a = 5$, $b = 6$, $c = 7$, $h = 1$, $m_x = 1/2$, and $m_y = 9/2$ (see Fig. 7).

However, if there is sufficient migration between the cells, then the equilibrium, though it still exists, becomes unstable. Any random deviation that causes a difference between cells, however slight, unbalances migration and causes a sustained and accelerating net transport toward the cell with the higher concentration (Fig. 7, times 4–6). This is called a diffusive instability or a Turing instability. An equilibrium that is stable at a point of space can become unstable when its components diffuse to neighboring parts of the space. This resulting instability can lead to variations and patterns in biodiversity over the landscape, with the resulting patterns being stable.

The opposite can also hold. A system that is unstable in a point of space can become stable, or at least persistent, when extended over a region. Predators clearly coexist with their prey and parasites coexist with their hosts for long periods in nature. However, such coexistence has been difficult to scale down to the size of experiments. Many of the early experiments on predator–prey and host–parasite systems found them to be unstable at small scales. Confined to a small area, the predators captured all the prey, and then themselves disappeared from the experiment for lack of food (or, analogously, the parasites infected all the hosts and then themselves disappeared). However, as the spatial extent of the experiment was increased, persistence of the system increased dramatically. Ecological models of such systems show a similar behavior (Hassel and
Wilson, 1997). Although each individual cell quickly runs to extinction, migration ‘rescues’ empty cells, and the host and the parasite can have a stable average density over a large group of cells. Thus, an equilibrium that does not exist or is unstable in a point of space can be replaced by a stable equilibrium averaged over the entire region. Something similar happens in ecological systems described by metapopulation dynamics (Hanski, 1997). Individual populations of a species may be separated by distances or barriers that inhibit movement between populations. If movement is strongly inhibited, individual populations may be driven extinct locally by stochastic effects, only to be restored later by propagules from some other population. Despite individual populations going in and out of existence, the portion of local sites occupied at any time can approach a stable equilibrium. This locally nonpersistence system can be both persistent regionally and have a stable regional equilibrium.

C. Evolutionary Stability

On the ecological timescale, ecological parameters such as birth and death rate are considered fixed. They are taken to be immutable characteristics of the species under consideration. However, ecological systems have many layers of complexity beyond our simple abstractions of them. On the evolutionary timescale, the parameters are malleable through the process of mutation and natural selection. Moreover, species formerly isolated may come into contact either as a result of natural causes or, now with great frequency, through the actions of humans. Both invading species and mutant phenotypes present new parameters that test the evolutionary stability of the system. A condition that is stable with ecological parameters fixed need not be stable when these parameters can change.

How the parameters might change can be seen by examining the growth rate of potential invaders, initially at negligible densities, entering the community at equilibrium. If the growth rate of any phenotype is positive, then the system can be invaded by that phenotype. The resident species, and hence the parameters of the system, are evolutionary stable only if the growth rate of every potential invader is negative. Originally developed for behavioral systems, this idea is being applied to ecological communities (Geritz et al., 1998), and such work promises to shed light on the properties of invasions and ultimately on the evolution of biodiversity.

See Also the Following Articles

CARRYING CAPACITY, CONCEPT OF • DISTURBANCE, MECHANISMS OF • EUTROPHICATION AND OLGOTROPHICATION • HABITAT AND NICHE, CONCEPT OF • ISLAND BIOGEOGRAPHY • POPULATION DYNAMICS • SPECIES, CONCEPTS OF • STRESS, ENVIRONMENTAL

Bibliography


STEWARDSHIP, CONCEPT OF

Peter Alpert
University of Massachusetts at Amherst

I. Etymology
II. Rationales
III. Contexts
IV. Types of Stewards

GLOSSARY

community-based conservation Conservation-oriented management of communal or public property by local residents.
ecosystem management Management of an ecologically defined geographical area as an integrated whole for both sustainable yields of natural resources and maintenance of ecological processes and species populations.
integrated conservation and development project Project that links biological conservation with human development in a local area or set of areas.
land ethic Philosophy of proper treatment of lands and the natural communities on them.
natural resources management Management of natural systems to maintain or provide yields of wood, water, and other resources for people.
nongovernmental organization An organization that is not an agency or part of a government, typically one that is not-for-profit and not owned by individuals.

STEWARDSHIP is taking care of something for someone else. Originally, a steward was a person who managed household affairs for a landowner. Today, the concept of stewardship extends to the management of intangible things, such as beliefs, and to management of things legally owned by the manager, as long as he or she is acting on behalf of interests greater than just his or her own. In natural resource management, stewardship often refers to voluntary actions taken by private landowners to promote ecological goals on their own lands. Stewardship may have an instrumental rationale, such as the human need for natural resources; or an intrinsic rationale, such as the moral rights of species to exist. Stewardship of biodiversity occurs within a context of ecological, economic, cultural, and political factors that help determine the forms stewardship can take and how well it will work. For example, the large scale of some ecological processes requires forms of international stewardship. Governmental agencies, nongovernmental organizations, local communities, classes of citizens, companies, private individuals, and partnerships of these different types of groups have all undertaken stewardship of biodiversity. Their roles may be just to protect species and habitats against intrusion, or to actively maintain population sizes and ecological processes in habitats too isolated or small to retain
I. ETYMOLOGY

The term “steward” has existed for approximately 1000 years. According to the Oxford English Dictionary, its early form “stegward” first appeared in manuscripts in the eleventh century. The prefix “steg-” is probably the Old English word for “house or some part of a house.” “Weard” is an Old English word for keeper. A steward is literally a “household-keeper.” This gives the word an interesting kinship with the word “ecology,” coined eight centuries later from Greek words for “house” or “household” and “study.” The steward manages what the ecologist studies.

The essence of the idea of a keeper is one who manages something for someone else. For example, the Old Teutonic root “warda,” presumably related to “weard,” gave rise to the word “warden.” This word was used in a conservation sense during the Renaissance. A royal act under Edward I in 1294 proclaimed that “The wardens shall keep and susteyne the lands without makynge dysruption of any thynge.” In thirteenth-century France, wardo became “garde” and then “guardian” in fifteenth-century England, meaning “one to whom the care and preservation of any thing is committed.” Like his Latin equivalent, a “custodian,” a “guardian” could manage but not rule: “From that yere . . . were al custodies & gardeyns and no mayres” (Robert Fabyan, The Newe Chronicles of Englande and of France, 1516).

The “household” in “steward” may at first have meant exactly that. The first meaning of “steward” in the Oxford English Dictionary is “an official who controls the domestic affairs of a household, supervising the service of his master’s table, directing the domestics, and regulating household expenditures.” However, usage of “stewardship” soon extended beyond the kitchen and pantry. As the “royal households” in England and Ireland became more exalted, “steward” became a title of high office. In early Scotland, the Lord High Steward was the first officer of the king. One such official, Robert the Steward, claimed the throne. As the new ruler of the royal household, he was no longer its steward, and his epithet changed, founding the royal house of Stuart.

On the other hand, a king threatened with the loss of his crown might find it expedient to claim to be in the employ of a still higher authority and therefore to be a steward. Shakespeare had Richard II make this case: “Show us the hand of God, That hath dismiss’d us from our stewardship.” The formal usages of stewardship now include the management of any property owned by someone else.

Some contemporary common usages of “steward” adhere to these formal meanings. People today probably encounter the word steward most often in the person of an airline or ship’s steward, someone in charge of food service on an airplane or ship owned by someone else. Organizations such as The Nature Conservancy employ “land” or “reserve stewards” to manage natural areas. These stewards are paid to manage part of an estate for the owner, which is the organization.

Other common usages of “steward” include the management of intangible things, such as rights or beliefs. A “shop steward” is a union official charged with protecting the rights of a group of employees; a “religious steward” may be a volunteer charged with helping to maintain a faith or church. Stewardship can also refer to the management of things owned by the steward, as long as he or she is acting on behalf of interests greater than his or her own. Currently, the most common use of “steward” in relation to the conservation of biodiversity is for private individuals or companies that agree to abide by special practices designed to protect or restore natural species or ecology on their own lands. A wide variety of national, state, and local governmental and nongovernmental agencies and organizations sponsor programs that recruit private landowners to volunteer as such “ecological,” “environmental,” or “habitat stewards.”

Despite these extensions of the original meaning of “steward,” the concept of stewardship remains distinct from the concept of belonging, on the one hand, and from the concept of domination, on the other hand. Belonging implies being an equal or subordinate part of a thing rather than managing it. The first premise of stewardship is the competence and right to act on the thing in one’s charge so as to produce certain results. Domination, or proprietorship, implies controlling or managing a thing in one’s own interests. The second premise of stewardship is that the desired outcomes of management are designed to satisfy interests greater than those of the steward alone. Stewardship is a three-way relationship between a manager, a thing being managed, and the principle or party in whose interests the manager acts.

II. RATIONALES

Stewardship can have a variety of rationales. These rationales can be classed as instrumental, if management is viewed as a means to an end, or as intrinsic if manage-
sustainable development as "development without growth." Its objective, as stated in the 1987 "Brundtland Report" of the World Commission on Environment and Development, is "meeting the needs of the present without compromising the ability of future generations to meet their own needs." On a more humble scale, this is like passing on the family farm in better shape than you found it, and one of the most resonant arguments for sustainable development is that is essential for the welfare of one's children. Author Peter Mather has appealed for stewardship of biodiversity because "indifference to the loss of species is, in effect, indifference to the future, and therefore a shameful carelessness about our children." When Margaret Jacobson and Garth Owen Smith asked Himba pastoralists in northwestern Namibia in the 1980s and 1990s why they wanted to conserve local wildlife, a common reply was so that their children could see them.

Ecosystem management focuses on natural resources management rather than on human development but makes the same argument for environmental stewardship as does sustainable development—that it is beneficial to people in the present and in the future (Johnson et al., 1990). Steven L. Yaffee of the University of Michigan noted that the term "ecosystem management" covers a range of approaches to natural resources management, from "anthropocentric" plans for multiple use of natural resources through "biocentric," whole-ecosystem management to "ecocentric," natural regionwide plans. The first approach was codified as early as 1906 in the Multiple Use Sustained Yield Act of the U.S. Congress under Theodore Roosevelt. Whole-system management was in principle implemented by the National Park Service in 1963 when it accepted the recommendation in the "Leopold Report" of its Advisory Board on Wildlife Management, chaired by Starker Leopold, to manage its lands as "biotic wholes." The statutes that govern the Oregon Board of Forest Lands prescribe a regionwide plan to manage for "greatest permanent value," defined in a 1999 administrative rule as "healthy, productive and sustainable forest ecosystems that over time and across the landscape provide a full range of social, economic, and environmental benefits to the people of Oregon." There is evidence that people in the United States find instrumental rationales for ecological stewardship more convincing than intrinsic ones. As part of a study funded by the U.S. Forest Service Office of Communications, 30,000 on-line news stories about national forests published from 1992 to 1996 were searched by co-
put for references to benefits in four categories: "recreational," "commodity," "ecological," and "moral/spiritual/aesthetic." Recreational benefits accounted for about 40% of the citations, commodity benefits for 30%, ecology for 20%, and moral benefits for only about 10%. Recreational benefits tend to translate into local economic benefits and are important not only to people who recreate but also to those who sell to them or tax them. A 1990 report from the U.S. Department of Agriculture on population and economic growth in nonmetropolitan counties in the contiguous United States from 1970 to 1996 found that growth was strongly dependent on local natural amenities, as measured by an index based on weather, topographic variation, and water surface area. Populations grew by 120% on average in the counties that scored high on the amenity index and only by 1% in the counties that scored low. High-scoring counties also had a threefold higher increase in number of jobs compared to low-scoring ones.

B. Intrinsic Arguments for Stewardship of Biodiversity

Moral philosophy has provided many intrinsic arguments for the stewardship of species. Concepts of "moral status" have been held to require varying degrees of solicitude for different species depending on their ability to make decisions, be self-aware, feel desire or pleasure or pain, and strive to exist. A being has moral status when it has "moral duties directed to it" by "moral agents," those "able to deliberate and act in a responsible and answerable way" (Wetlesen, 1999). Some philosophers, such as Plato and Kant, assign moral status only to beings with free will, meaning normal humans. Thomas Regan has argued for moral status for beings that experience emotions and have a sense of their own individuality, meaning normal mammals more than 1 year old and possibly other animals such as birds. Animal rights advocate Peter Singer has proposed sentiment, the ability to have conscious feelings of pleasure or pain, as the qualification for moral status.

In addition to the issue of who or what should be accorded moral status, there is the issue of how equal this status should be. One might have equal moral duties toward all things with moral status or greater duties toward some of these things than toward others. Many ethical philosophers accept that persons have equal moral status. As signatories to the Universal Declaration of Human Rights of the United Nations, many nations accept this as well: "All human beings are born free and equal in dignity and rights. They are endowed with reason and consciousness and should act towards one another in a spirit of brotherhood." Ethical philosophers generally assign nonpersons a lesser moral status. For example, Mary Anne Weaver reasoned that people have a moral responsibility to accept the equal rights of all persons, to avoid cruelty to sentient beings, and to respect all life. Wetlesen (1999) proposed "equal value for moral persons and agents and lesser values for nonpersons, in proportion to their similarity with moral persons."

Stewardship of things that have no moral status can be morally justified on the basis of their importance to things that do have moral status. For instance, it could be held to be wrong to shatter a stalactite because it will deprive other people of the enjoyment of seeing it or wrong to divert water from a pool because it will endanger a fish. In most minds, this provides the only intrinsic justification for stewardship of nonliving things. However, thinkers such as Baird Caldicott elaborate reasons for attributing intrinsic values to wholes, which could apply to ecological systems.

In addition to its instrumental schools, the history of conservation contains a tradition of intrinsic arguments for stewardship, documented in the writings of John Muir, Henry David Thoreau, David Brower, and others. In 1904, George Perkins Marsh wrote that "Man has too long forgotten that the earth was given him for usufruct alone, not for consumption." Toward the end of his career in 1949, Aldo Leopold proposed that "A thing is right when it tends to preserve the integrity, stability and beauty of the biotic community. It is wrong when it tends otherwise" and that "The last word in ignorance is the man who says of an animal or plant, 'What good is it?'" Ecologist Charles S. Elton ventured in 1998 that the real reason for conservation is that "animals have a right to exist and be left alone, or at any rate that they should not be persecuted or made extinct as species." Robert H. Nelson summed up how environmentalists who justify biological conservation on moral grounds migh view things: "To argue for building a dam would be like arguing for the institution of slavery because it was economically efficient."

Intrinsic arguments for the stewardship of biodiversity have found their way into legal philosophy and animal rights legislation. Roderick Nash (1989) made a case for regarding legal rights for nonhumans as a logical continuation of the progressive enfranchisement of different economic classes, races, and genders of people. Christopher D. Stone raised the question of legal rights for nonhumans in his 1972 essay, "Should Trees Have Standing?" [Southern California Law Review 45(2), republished in 1974 by William Kaufmann,
Los Altos]. Some natural lands have been set aside for use by wild animals only. Several areas in the Sierra Nevada of California that have bighorn sheep are closed to human entry during most of the year.

Does it make a difference whether the rationale of stewardship of biodiversity is instrumental or intrinsic? Two ways that it might are if one type of rationale is more likely to withstand arguments against conservation or if the two types lead to different ways of managing. It has been widely argued that only instrumental justifications for biological conservation will lead to political action or stand up to conflicting economic pressures. The plant most often cited as an example of why it is important to conserve biodiversity is the rosy periwinkle of Madagascar. Chemicals extracted from this wild plant were discovered to cure childhood leukemia. Others argue that only a belief in the right of species to exist will motivate stewardship of biodiversity since most species have no demonstrable human use or known ecological significance. One interesting possibility is that it takes intrinsic arguments to generate instrumental ones. Officials of the United Nations Commission on Economic Development interviewed in the early 1990s by P. P. Craig and colleagues confirmed that although they made only instrumental arguments for environmental conservation in their professional work, they privately believed that the real arguments were intrinsic.

C. Counterarguments against Stewardship as a Model for Conserving Biodiversity

Without necessarily rejecting the idea that it is important to conserve biodiversity, many groups and individuals have rejected the concept of stewardship of biodiversity. Objections have been made to each of the two basic premises of stewardship: that management of biodiversity is needed and justified and that people should care for biodiversity on behalf of interests greater than their own.

One objection to the idea of managing biodiversity is that it is self-defeating because the essence of the biological world is its wildness. In an essay on “the etiquette of freedom,” Gary Snyder (as cited in Oelschläger, 1992, pp. 21–39) contrasts the meanings of “natural” and “wild.” He argues that managed things may be natural but not wild and that wildness is the world itself. Paul Taylor distinguished the realm of “bioculture,” in which husbandry and agriculture may take place, from that of wilderness, in which “noninterference ethics” should prevail. George Sessions (as cited in Oelschläger, 1992, pp. 90–130) rejects the model of stewardship for wilderness and admits only that “perhaps some ecologically enlightened version of the ‘stewardship’ model is appropriate for the bioculture.”

A second objection to the idea of managing biodiversity is that it is foolhardy because we do not know enough about wild species and ecological systems to know which actions to take. This objection is supported by past management errors. Human suppression of fires in western North American parks and national forests in which natural fires occur regularly has inhibited reproduction of plant species and reduced the diversity of some forests. Alston Chase argued that ecosystem management has altered wildlife processes in Yellowstone National Park.

These objections to the concept of stewardship of biodiversity can be viewed as cautions about what stewards should do rather than reasons to abandon stewardship altogether. Those who argue against managing wilderness do not necessarily preclude protecting wilderness areas from development. Those who argue that humans need more knowledge to manage ecological systems successfully may also argue that the best course is to try to acquire this knowledge through “adaptive management,” in which the results from each management prescription are analyzed and used to plan the next. Jack Ward Thomas said that “an ecosystem is not more complicated than we think, it is more complicated than we can think.” However, as then head of the U.S. Forest Service, he probably did not mean by this that ecosystem management should be abolished but rather approached with humility. Richard L. Knight has likewise proposed that an appreciation of the complexity of nature and a willingness to learn from mistakes and modify techniques are essential components of good stewardship of biodiversity.

A counterobjection to the notion that human management of biodiversity is foolhardy is that humans now have no choice but to manage natural systems because we already dominate them, foolishly or not. Globally, humans are said to consume two-fifths of the earth’s photosynthetic productivity and to have doubled the rate at which nitrogen is being supplied to plants (Vitousek et al., 1997). On some islands and some continents, human influences have been judged to now determine the distributions of species and the risks of extinction of animal populations more than natural factors. Many remaining areas of habitat are too small to maintain their biodiversity naturally. In southern Africa, where elephant populations have been largely restricted to movement within reserves, elephants can multiply to the point at which they deforest large areas, leading managers to kill elephants to conserve trees.

Globally, humans are said to consume two-fifths of the earth’s photosynthetic productivity and to have doubled the rate at which nitrogen is being supplied to plants (Vitousek et al., 1997). On some islands and some continents, human influences have been judged to now determine the distributions of species and the risks of extinction of animal populations more than natural factors. Many remaining areas of habitat are too small to maintain their biodiversity naturally. In southern Africa, where elephant populations have been largely restricted to movement within reserves, elephants can multiply to the point at which they deforest large areas, leading managers to kill elephants to conserve trees.
other systems, humans have eliminated key elements and may have to replace their actions artificially. Short periods of grazing by livestock have been found to promote biodiversity in interior North American grasslands, possibly because the grazing simulates the effects of the great herds of buffalo that formerly wandered the grasslands.

A second counterobjection to the charge of foolhardiness is that humans are an integral part of some ecological systems. Aboriginal burning of savanna in what is now Kakadu National Park in Australia may have been part of the fire regime for 50,000 years. This is clearly long enough for native species to have evolved in response to human-caused fire frequency.

A third objection to human management of biodiversity is that we have no right to be managers, that we stand in relation to other species, not as stewards, but as siblings. The native American writer Black Elk represented his people’s belief that “We are of Earth and belong to You. Every step that we take upon You should be done in a sacred manner.” In his book The Green Fuse, biologist John Harte explicitly rejected the idea of humans as stewards of ecological systems; it is more accurate, he said, to think of nature as the steward of humanity and to call upon humans to be stewards of their own impulses. “Self-Stewardship” is the title of the chapter that deals with stewardship in Vice President Al Gore’s book Earth in the Balance.

The second premise for the stewardship of biodiversity is that persons should manage it for the sake of interests greater than their own. Many people reject this in favor of the premise that humans rightfully dominate nature and are the rulers and not the stewards of biodiversity. Captain Vancouver is said to have remarked upon seeing the virgin temperate rainforests of his eponymous island that here were enough masts to supply the British Navy forever. The belief that man should rule over nature figures prominently in Judeo-Christian tradition. The Book of Genesis in the Bible records that humans shall “have dominion over the fish of the sea, and over the fowl of the air, and over every living thing that moveth upon the earth.” Confusingly, some recent Christian models called stewardship models still portray humans as dominant over other living things. The medieval Jewish philosopher Moses Maimonides wrote, “It should not be believed that all the beings exist for humanity. On the contrary, all the other beings too have been intended for their own sake and not for the sake of something else.” According to the contemporary Catholic theologian Thomas Berry, “The earth belongs to itself and to all the component members of its community.” In The Ecology of Eden, Evan Eisenberg traces these contradictory theological viewpoints back to the biblical injunction to Adam and Eve to both work and protect the Garden of Eden, that is, to have dominion over the organisms but safeguard the divine creation.

A recent legal embodiment of the belief that people can own other life is the extension of intellectual property rights to cover species or aspects of the genetics or metabolism of a species. This principle, upheld by U.S. courts and by international governmental trade organizations, has been touted as an essential economic incentive for the development of biological and medical technology and as a way to ensure that developing countries and local residents derive benefits from biodiversity. In 1971, the first patent on a genetically engineered organism was granted in the United States to General Electric and its employee, Anand Mohan Chakrabarty, for a Pseudomonas bacterium into which plasmids from three other bacterial strains had been introduced. The General Agreement on Trade and Tariffs that governs much of world trade contains a Trade Related Intellectual Property Rights agreement that can cover wild species and their products. Even aspects of human biodiversity can become private property. The government of Iceland has debated transferring the intellectual rights to a national database of citizen’s genetic information to a corporation.

Throughout history, many societies have held that humans can own individual living things as property, including other humans. Real property rights often include the rights to plants, animals, and minerals on the property. However, it is also usual to recognize that there are multiple interests in real property, that public and private interests are commingled on almost all lands, and that property owners have responsibilities to avoid using their property in ways that harm their neighbors or society. The interests of landowners in the natural resources on their property can be legally balanced against the recreational and environmental interests of the public. On a more moral plane, Aldo Leopold wrote in “The Ecological Conscience” in 1947 that “when a farmer owns a rarity he should feel some obligation as its custodian and a community should feel some obligation to help him carry the economic cost of custodianship.”

A final counterargument against stewardship as a model for the conservation of biodiversity is that the most effective basis for conservation is love, which does not involve acting in interests greater than one’s own.
To explain this argument, it is useful to return to the example of a guardian and a child. Instead of taking care of a child to produce an adult or out of duty or principle, a guardian might do so out of love for the child. Love is neither an instrumental nor an intrinsic rationale; it is not rational and involves no third party or principle in whose interests one acts. Just as many will argue that love of a child leads to the best child care, Wendell Berry has proposed that “love of place” leads people to care for species and habitats. The land ethic propounded by Aldo Leopold (1966) also contains an element of love: “When we see land as a community to which we belong, we may begin to use it with love and respect” and “affection based on utility alone leads to the same pitfalls and contradictions in land as in people.”

One practical implication of this view is that the best way to promote the conservation of biodiversity is to take people to see it. One cannot love without at least first sight. In southern Africa, many programs are dedicated to taking children to see wildlife. Courses for “decision makers” run by the Organization of Tropical Studies in Costa Rica discuss arguments for conservation but also take their prominent students to watch sea turtles lay their eggs in the moonlight on a tropical beach.

III. CONTEXTS

The concept of stewardship has been applied to the conservation of biodiversity in a multitude of forms and with varying success. This is partly due to the ecological, economic, cultural, political, and historic contexts in which stewardship of biodiversity has been undertaken. These contexts include the mobility of animal populations, disequilibrium in ecological systems, economic incentives and disincentives for conservation, poverty and industrialization, the scale of trade, traditional views of the place of humans in nature, legal systems of land and resource tenure, and the displacement or migration of peoples. Some of these contexts for stewardship are briefly discussed in the following sections.

A. Ecological Factors

Stewardship requires that the scale of management match the scale of the thing being managed. In the phrase of Kai Lee, there is often a “mismatch of scales” between ecological processes and human societies. For example, many species and ecological processes cross the jurisdictional boundaries that constrain managers. More than 100 species of Neotropical migratory birds breed in North America and winter in Central or South America. Declines in North American breeding populations have been associated with forest losses in Central and South America. Marine organisms with planktonic larvae often disperse widely on ocean currents and depend on long-distance dispersal to maintain local populations. The establishment of new juveniles in marine reserves may depend almost entirely on dispersal from outside reserves. Aerial nitrogen deposition, which occurs on a global scale (Vitousek et al., 1997), is thought to be responsible for replacement of heathland by grassland in The Netherlands. Due to the “openness” of ecological systems, reserves may be, as Mark H. Carr and colleagues suggest, “necessary but not sufficient for conservation.”

Ecological fluxes and connectedness complicate the notions of ownership and management. Many species and the resources on which they depend fall into the category of “common pool resources,” which have been defined as “those from which multiple independent users cannot be excluded and in which consumption by one user detracts from another’s consumption.” These include species outside national boundaries, as in international waters, and species on communally owned lands. Models for the stewardship of biodiversity in commons range from international legal agreements such as the United Nations Convention on Marine Law, which assigns states bordering oceans and seas specific rights and responsibilities for their use, to the customary land tenure systems of African pastoralists.

Land stewardship may be more difficult in habitats in which resources are relatively scarce or where natural conditions are unpredictable or rapidly changing. E. Steen observed that cultures in the relatively dry habitats around the Mediterranean have tended to exhaust soil fertility except where there are reliable pulses of water with low salt and high alluvial content, such as along the Nile. Frequent erosion and deposition of soil maintain plant species diversity in highly dynamic habitats, for example, along rivers and on coastal sand dunes, making it necessary to manage for land instability. The prevailing view of ecological systems has shifted from a notion of dynamic equilibrium, in which systems are expected to tend to reach and return to one steady state, to one of disequilibrium, in which systems are expected to change directionally over time. This makes biodiversity a moving target for management. Some ecological systems switch between alternative states: They may show a nonlinear response to human use, remaining relatively stable until use reaches a certain
intensity or cumulative level and then change relatively rapidly, or they may not change back even when use discontinues. This makes it difficult for managers to learn from experience. In habitats in which rare events strongly affect the distribution and abundance of species, managers may need to be “stewards of catastrophe,” preserving or mimicking the potential for major disturbance.

B. Economic Factors

Among the many economic factors that impinge on stewardship are disincentives and incentives for conservation, degree of industrialization, and trade across ecological boundaries. When land is taxed based on its most lucrative potential use instead of its current use, landholders are more likely to sell or develop their land. Many laws, such as the Williamson Act in California, have been passed to remove this economic disincentive for farmers. Often, farmers agree not to develop their land in exchange for current use valuation of the land for tax purposes. Laws have also been passed to create economic incentives for stewardship of biodiversity. One widely used instrument is the conservation easement, an agreement under which a landowner agrees not to use his or her lands in certain ways in exchange for relief from taxation. The Federal Agriculture Improvement and Reform Act of 1996 created or redirected a set of U.S. Department of Agriculture programs that provide economic encouragement to private landowners to protect wildlife habitats. The Wildlife Habitat Incentives Program provides cost-sharing for habitat improvement, the Conservation Reserve Program pays farmers to fallow cropland under a vegetation cover for a decade, the Wetlands Reserve Program provides conservation easements and cost-sharing for restoration and protection of wetlands for 30 years or more, and the Environmental Quality Incentive Program combines education and technical and financial assistance to help landowners adopt practices that reduce environmental problems.

Economic conditions in developing countries have been linked to poor stewardship of biodiversity in several ways. Poor people in rural areas are likely to depend heavily on local natural resources, and a large proportion of land area is likely to be settled by farmers. Over time and with population growth, use tends to become unsustainable. Governmental agencies are likely to lack the resources to enforce protection of species in nature reserves. Wood and charcoal are a main source of fuel in cities, leading to deforestation for hundreds of kilometers into the countryside, as around Kinshasa in Congo. On the other hand, it is by no means clear that economic development improves stewardship. In central Africa, “de facto” conservation of biodiverse areas by their remoteness and local economic poverty has probably been as effective as intentional stewardship in the region. One hypothesis is that development will tend to improve stewardship of biodiversity where subsistence resource use is the main cause of its loss, as is probably the case for forests in Madagascar, but not where commercial exploitation is the main cause, as in the forests of Cameroon.

The differences between economic conditions in developing and industrialized countries are often held to require different approaches to stewardship. Conservationists in developing countries largely argue that local rural communities must receive economic benefits from protected areas. This is the basis for the spread of integrated conservation and development projects, which generally link biological conservation within a protected area to human development in the surrounding communities (Wells and Brandon, 1997). These projects are mainly limited to developing countries.

The globalization of commerce has to some extent reversed the mismatch of scales between ecology and society. It is increasingly the case that trade occurs across ecological boundaries. Global trade is often held to threaten stewardship of biodiversity because it makes it more likely that parties using or profiting from resources will live outside the region from which the resources come, and that they will therefore be less concerned about stewardship or constrained by environmental regulations. This is one basis for “bioregionalism,” advocated by Sale Kirkpatrick and others, which counsels that people use only the natural resources from within their ecological region.

C. Cultural and Political Factors

Human cultures probably differ in their attitude toward stewardship. For example, D. R. Given argued that the Maori have an environmental ethic that emphasizes guardianship and stewardship, in contrast to a European tradition of dominance over nature. He predicted that laws giving the Maori greater control over resource management would help develop a multicultural, ecocentric biodiversity ethic in New Zealand. Others contend that economic factors tend to supersede cultural traditions. Rodrigo Sierra found no recent differences in impacts between three indigenous and three nonindigenous populations in northwest Ecuador, although there had been differences in the past when economic conditions were different. He concluded that the degree
of impact of different peoples on natural systems in this region was primarily associated with economic conditions. Cultural and political factors can result in "incidental stewardship" of biodiversity, achieved as a by-product of management for other purposes. Some of the most effective stewardship of natural habitat in both the United States and developing countries has been incidental. Lands reserved by the U.S. Army for military bases have become Nature Conservancy preserves and parts of national wildlife refuges and National Park Service units. In the western United States, the military was politically able to completely exclude resource extraction over large tracts of land, some of which now constitute the most pristine examples of semiarid habitats in the region. In some areas of Africa and Asia, small groves held sacred and reserved for religious purposes are among the least disturbed forests.

Local political authority over resource and land use has been associated with good stewardship. Al Gore has proposed that "freedom is a necessary condition for an effective stewardship of the environment" and called to witness the environmental degradation in eastern Europe and hazardous waste sites in poor communities with relatively little political power. Secure land tenure is widely regarded as a precondition for effective land stewardship by individuals and communities. In African countries, governmental ownership of wildlife on private or communal land has been blamed for lack of interest in wildlife conservation. However, a given political arrangement may have opposite effects on stewardship in different economic contexts. The establishment of wildlife management agencies or funds as "parastatals"—governmental bodies that are economically independent of the rest of the government—has promoted wildlife conservation in Kenya and Zambia. In these countries, tourism in some parks and game management areas was generating more revenues for the central treasury than the treasury was providing to the parks largely because the tourists were almost all foreign and could afford to pay large amounts relative to the local economies. Once given political authority to raise and disburse its own revenues, the Kenya Wildlife Service was able to retain gate receipts at Amboseli National Park and share them with the Maasai around the park. During the wet season, the large ungulates in the Amboseli system rely largely on the Maasai lands for grazing, making the cooperation of the Maasai essential for conservation of biodiversity in the park. In Zambia, governmental revenues from European and North American sports hunters were channeled to the park and wildlife service for management, providing hundreds of jobs for local residents in game management areas. In contrast, requiring parks to be economically self-sufficient in the United States, in which the national park service receives more governmental revenues than it generates, would probably lead to a decrease in resources for park management.

IV. TYPES OF STEWARDS

Almost every possible type of social or political group has undertaken, or purported to undertake, the management of species and habitats on behalf of interests greater than those of the group. The possible stewards of biodiversity include governmental agencies, nongovernmental organizations (NGOs), corporations, local communities, and private individuals. Governmental agencies and NGOs include international, national, and local bodies. They have worked separately and in partnerships of different types of groups. Through legal and political action, classes of citizens have also acted as stewards. The following sections provide examples of these types of stewards and discuss some of their apparent strengths and weaknesses: Are certain types of stewards more effective or trustworthy as guardians of biodiversity?

A. Governmental Stewards

International governmental agencies exercise stewardship of nature in the form of conventions, declarations, and moral and financial support for conservation efforts. Through trade agreements and support for development, international bodies have both negative and positive effects on biodiversity. International governmental bodies are uniquely able to enforce stewardship of global ecology such as upper atmospheric conditions and of common pool resources such as marine fisheries and the Antarctic continent. They can finance and lobby for the protection of local sites that are internationally prized. The most visible international governmental steward of biodiversity has been the United Nations (UN). In 1982, its General Assembly adopted a Charter for Nature that recognized an intrinsic basis for the stewardship of biodiversity:

Every form of life is unique, warranting respect regardless of its worth to man, and to accord other organisms such recognition, man must be guided by a moral code of action. . . . Nature shall be
respected and its essential processes shall not be disrupted.

The vote was 114 for, 17 abstaining, and 1 (the United States), opposed. UNESCO, an organization within the UN, designates areas of special cultural and conservation interest as Man and the Biosphere Reserves and World Heritage Sites. Most Biosphere Reserves center on existing protected areas, and designation as a reserve is largely symbolic. However, it may have an important educational value by affirming the preciousness of natural places and help inspire support from other sources. The National Natural Landmark program of the U.S. National Park Service is an example of a similar program on a national scale. The United Nations Environmental Programme directly funds conservation efforts and projects.

The “Montreal Protocol” is often cited as a successful example of international governmental cooperation to protect the environment. This agreement banned most release of chlorofluorocarbons into the air. These chemicals can lower the concentration of ozone in the upper atmosphere and increase the amount of cell-damaging ultraviolet radiation that reaches the earth’s surface. Important scientific analysis to support this political agreement was supplied by an international nongovernmental body, the Scientific Committee on Problems of the Environment of the International Council of Scientific Unions. This is an association of professional scientific associations such as the American Association for the Advancement of Science.

International governmental trade and financial institutions such as the World Bank do not have ecological stewardship as their primary goal. The World Bank has supported projects agreed to have been environmentally harmful, notably the Aswan Dam in Egypt. However, World Bank policies formulated during the 1980s explicitly called for conservation of tropical forests. The bank set up an internal environmental review process for its projects and funded the Global Environmental Facility, which gave unusually large amounts of money for its projects and funded the Global Environmental Facility. These agencies are stewards inasmuch as they manage the resources on behalf of groups outside the agency. The U.S. Fish and Wildlife Service refers to the National Wildlife Refuges it administers as “stewardship lands.” However, the primary charge of these agencies is often to maximize the economic returns from resource use rather than to conserve biodiversity. Georgia M. Mace characterized the role of natural resource managers in governmental agencies as generally aiming to maximize continuing yields and collect extensive data on a few species rather than assessing management risks to a wide diversity of species. This would tend to make them poor stewards of biodiversity. For example, maximum productivity is generally negatively associated with maximum species richness. On the other hand, governmental agencies have significant advantages as land stewards. They often command relatively extensive resources and manage a large proportion of the natural lands in a country, including most of the large natural areas. Depending on how their professional incentives are structured, managers within agencies may be largely free from personal conflicts of interest with conservation.

Many governmental resource agencies emphasize resource use over biodiversity conservation for both economic and political reasons. In many countries, including the United States, the operating budgets of national forests are tied to the fees paid by companies for logging concessions. This provides a powerful incentive to manage forests for wood production and to focus on a few economically valuable species. The charter of the U.S. National Park Service calls upon it to protect parks for the enjoyment of visitors. Parks experience pressure to manage for maximum continuing yields of visitors and
therefore to develop lodging and shops and to focus on species that are appealing and easy to see, the “charismatic megafauna.”

Research institutions and research departments within resource management agencies operate natural reserves throughout the world. The approximately 30 reserves of the University of California comprise the largest system of any single university. The U.S. Forest Service has designated hundreds of Natural Research Areas, often prime examples of distinctive forest types, that are open to research use only. One goal of the sites funded under the Long-Term Ecological Research program of the U.S. National Science Foundation is to provide baseline data on ecological function and community structure in diverse habitat types that are as little disturbed by humans as possible. An advantage of research reserves is that they are generally completely protected from consumptive use and largely protected from tourism. However, the reserves are mostly small and subject to experimental manipulations such as animal exclusions and vegetation removal and nutrient addition in plots.

B. Nongovernmental Organizations and Communities as Stewards

Conservation land trusts, NGOs that purchase land or the development rights to land for the express purpose of maintaining it in an undeveloped state, are likely to have the conservation of species or habitats as their primary goal. Another advantage of land trusts relative to governmental agencies seems to be that they are able to act quickly to acquire habitat in immediate threat of development. The Nature Conservancy is the largest private land trust for natural lands in the United States, with more than 10 million acres of reserves throughout the country. One of its mottos, “the last of the least and the best of the rest,” epitomizes its dual objectives to prevent the extinction of rare species and to conserve whole natural communities and ecological systems. The first objective requires a large number of reserves, some of which may be very small, for example, if the species of concern is a grass. The second requires large reserves, which have to be fewer in number for reasons of cost and availability. Local land trusts have also played important stewardship roles. Starting in the early 1900s, the Save-the-Redwoods League purchased a significant proportion of the remaining old-growth forests of the coast redwood, *Sequoia sempervirens*, which the trust eventually deeded to the government for parks. A disadvantage of land trusts as stewards is that they are rarely able to purchase large areas and generally control only a small proportion of the natural lands in an area. According to the Land Trust Alliance, local land trusts owned about 0.002% of land in the United States in 1999.

A highly controversial type of steward of biodiversity is the local community, the set of human residents in a place. “Community-based conservation” is an important option in societies in which a large proportion of lands are owned or customarily used by communities as a group. It is viewed as an alternative both to the establishment of protected areas that exclude consumptive uses of natural resources by local residents and to natural resources management by governmental agencies. During the 1980s and 1990s, many projects were established to encourage local communities to assume or resume the major responsibility for local environmental stewardship, especially in the developing countries of Africa and Asia. Sets of these projects have been reported on in books edited by Western et al. (1994), Mark Pollenburger, and others and in networking organs such as the Forest, Trees and People Newsletter (www-trees.slu.se). Two of the best known examples in Africa, the CAMPFIRE project in Zimbabwe and the ADMADE project in Zambia, have given varying degrees of authority for local wildlife management to communities. Community-based forest management has become particularly widespread in India. In the late 1990s, the Van Gujjars, a forest-dwelling pastoral tribe in Uttar Pradesh, proposed to resolve their objections to the establishment of the proposed Rajaji National Park by taking responsibility for forest management in the park. Reasons cited for turning to community-based conservation are that it is morally right, that local people have expert, “indigenous knowledge” of local species and habitats, that local people have lived in harmony with local nature for centuries, that traditional indigenous populations use a wide variety of wild species and thus have a practical incentive to conserve biodiversity, and that management by governmental agencies or in nature preserves in the area has failed to prevent habitat degradation. The Dayak groups of central Kalimantan in Borneo are reported by W. de Jong and others to provide successful examples of ethnoconservation; the Dayaks have manipulated the forest to increase productivity of species they use without reducing species diversity or simplifying structure.

One of the main reasons for doubting the efficacy of communities as stewards is the world history of environmental degradation. O. M’Hirit concluded that peoples around the Mediterranean have generally harmed their local forests and that only in the twentieth century has there been effective forest conservation in
the regions. In 1957, A. Starker Leopold took a dark view:

It is surprising that in the long history of man's conquest of the earth there is no evidence of sustained effort on the part of any people to preserve native landscape for its own sake, until our national park system began to take form late in the nineteenth century.

The consensus at the end of the 1990s was that local communities sometimes do and sometimes do not make effective stewards of biodiversity. Communities of recent immigrants in tropical rainforest tend to cause extensive deforestation because they want cleared land for crops and livestock. Even long-settled, traditional societies do not always use sustainable practices; a recent study concluded that hunters from the Piré people in the Peruvian Amazon make no attempt to conserve species vulnerable to overhunting. Practices may be sustainable only as long as human population density remains low and may change when economic and social conditions change. In a 1999 essay on community-based conservation projects in Africa, J. D. Hackel concluded that they tended to lack clear rules for the protection of local ecology and plans for what to do if conservation goals were not met. In both developing and industrialized countries, local communities may have a limited capacity to ask key questions about conservation values or to provide potential answers and limited time or will to bring issues before the public and decision makers. Others believe that community-based conservation has rarely been properly tested because governments have failed to devolve the authority over resources to communities at the same time that they have decentralized the responsibilities for resource conservation.

A different definition of a community is the set of people that have interests in a thing. In this sense, class actions by citizen groups and lobbying and public education by conservation organizations on behalf of their members are also forms of community-based conservation. The political and legal systems in the United States provide a direct route for stewardship by citizen groups through referenda and class action suits. An example of stewardship by class action is the work in the 1990s of the Mono Lake Committee in California to prevent Los Angeles County from diverting water from the streams that flow into this large desert lake. Diversion threatened, among other things, to allow coyotes to walk out to the nesting sites of birds on islands in the lake. The Sierra Club is credited with having dissuaded the U.S. federal government from creating a reservoir in the Grand Canyon in Arizona. The club's public advertising campaign asked "Would you flood the Sistine Chapel to get a better view of the ceiling?" Duck hunters are said to be largely responsible for motivating governmental protection of wetlands, which are now, under the RAMSAR convention, the first globally protected habitat type.

Such "community-of-interest-based conservation" can have the advantage of a longer economic view but the disadvantage of pitching legal and sometimes corporeal battles. The controversy over protection of old-growth Douglas fir forests in the Pacific Northwest, in the name of the northern spotted owl, pitted local logging interests against national conservation ones. According to one estimate, the protection plan that was adopted should result in economic gains of about $110 billion versus losses of only $32.3 billion. However, the losses are concentrated within the region, whereas the gains are spread throughout the country.

To avoid such battles, a broader community-based approach to conservation has been advocated based on the notion of "stakeholder rights." Stakeholders are all parties with interests of any sort in a given thing, be they economic, political, cultural, moral, or personal. One formulation of this "community-of-different-interests-based conservation" is to provide a forum for stakeholders to produce a common management plan for an area. For instance, the town library in Quincy, California, became a neutral meeting place for local environmentalists and loggers, who produced a plan for an adjacent area of national forest. To the surprise of many outside the "Quincy Library Group," the plan called for relatively intensive harvests, and its adoption by the Forest Service was challenged in court. Truman Young argued that many major conservation successes of recent decades, such as the international bans on whaling and ivory, have essentially rejected current economic stakeholder rights.

C. Private Companies and Individuals as Stewards

Private companies have also been engaged as stewards. One strategy has been to set up certification programs for commercial products or services. Companies or their taxes pay an independent body to certify that a product has been produced in an environmentally sound way. The expected benefit to the company is that environmentally concerned consumers will buy the product. A well-known certification program is run by the Forest Stewardship Council, an international NGO founded in
and residential development was at least partly respon-
sible for population losses in 35% of the 1880 imperiled
plants and animals in the United States. According to
Richard L. Knight, “unchecked conversion of U.S. pri-
ivate, open lands to human-dominated development is
causing a simplification of our native biodiversity.” In
some regions, such as northern New England in the
United States, most of the lands with natural or semina-
tural vegetation are privately owned.

An objection to relying on private landowners as
stewards is that private landholdings are generally small
compared to public ones and cannot accommodate
large-scale ecological needs. Remote imaging of land
use has documented that habitat patches may tend to
be smaller on private than on public lands. This objec-
tive can in some cases be partly met by cooperation
between landholders. In about 1980, local landowners
in southern Vermont formed the Newfane Wildlife Hab-
itat Improvement Group to plan together for wildlife
management. In 1999, the group included more than
50 properties and 7000 acres in three towns. However,
the total amount of land in the United States under the
protection of voluntary agreements by private landown-
ers is also small—just 0.02% of total land according to
a 1999 article by Keith Wiebe and colleagues.

As the previous examples of stewardship on private
lands suggest, stewardship is often undertaken through
partnerships between public agencies and private par-
ties. This is just one sort of “cross-boundary steward-
ship”—coordinated land management by multiple
stewards, often of different types (Knight and Landres,
1998). Many integrated conservation and development
projects are organized as partnerships among govern-
mental agencies, NGOs, and local communities (Wells
and Brandon, 1997).

A major advantage of cross-boundary stewardship is
that it allows for land use to be “not separate but not
equal,” to have a diversified portfolio of stewardship
types arranged side by side within ecological regions.
Arguments for zoning as an alternative to multiple use
are not new (Sessions as cited in Oelschlaeger, 1992,
p. 90–130). In 1971, Eugene Odum and JohnPhillips
proposed a landscape of urban-industrial, production,
and protection zones. The Biosphere Reserve concept
promoted by UNESCO since 1980 envisions protected
areas surrounded by an inner zone of human use un-
likely to impact the protected core and an outer zone
of unrestricted use. Some species do require protected
areas in which there is virtually no human impact. A
recent analysis of conservation versus human land use
by Kent Redford and B. Richter indicated that only
extremely limited and largely nonextractive use will
protect all components of biodiversity. However, as Sara
Vickerman argued, biodiversity cannot be conserved on reserves and with regulation alone (Defenders of Wildlife, 1998). A managed landscape can support important elements of biodiversity while meeting human needs. Significant components of biodiversity can and must be under the stewardship of private landowners.

The relative effectiveness of different types of stewards in conserving biodiversity seems to increase with the total amounts and individual sizes of the land areas they manage, the economic and technical resources they command, their independence from conflicting economic pressures and political constraints, the degree of authority they can exercise, and how close their primary objective is to conserving biodiversity. No one type of steward has greatest effectiveness in all these regards. Since it is also true in many landscapes that no one type of steward is responsible for enough land to ensure the conservation of biodiversity, it appears that a mix of stewardship types, arranged side by side in a landscape, is necessary and desirable.

See Also the Following Articles

- COMMONS, THEORY AND CONCEPT OF
- ECOSYSTEM SERVICES, CONCEPT OF
- ETHICAL ISSUES IN BIODIVERSITY PROTECTION
- HISTORICAL AWARENESS OF BIODIVERSITY
- LITERARY PERSPECTIVES ON BIODIVERSITY
- RELIGIOUS TRADITIONS AND BIODIVERSITY
- SOCIAL AND CULTURAL FACTORS
- SUSTAINABILITY, CONCEPT AND PRACTICE OF
- WILDLIFE MANAGEMENT

Bibliography


STORAGE, ECOLOGY OF

Caroline M. Pond
The Open University

GLOSSARY

adipocytes Large cells unique to vertebrates that take up and release fatty acids, produce receptors and respond to circulating and locally produced agonists, and secrete a variety of protein and lipid informational molecules.
adipose tissue, brown Thermogenic tissue unique to mammals that consists of small adipocytes containing numerous mitochondria and many small lipid droplets plus vascular and neural tissue. Thermogenesis is controlled by the degree of uncoupling of ATP synthesis in the mitochondria and can be very high in neonates and adults emerging from hibernation.
adipose tissue, white Storage tissue unique to vertebrates and best developed in tetrapods that consists almost entirely of expandable adipocytes (q.v.) that contain a large droplet of lipid plus vascular and neural tissue.
blasbber Specialized superficial adipose tissue found only in pinnipeds, cetaceans, and sirenians that serves as thermal insulation as well as providing energy storage.
fat body Storage tissue of arthropods that stores lipids and also has endocrine and immunological functions.
fatty acid Any of hundreds of different aliphatic hydrocarbons with an acid group. The main differences are the number and position of one or more double bonds and the presence of substitutions.
glycogen An insoluble polymer of glucose, the main storage carbohydrate in animals.
polyunsaturated fatty acid Fatty acid with two or more double bonds in the chain of carbon atoms. They occur in distinct families defined by the position of the double bonds, of which the most common are n-3 and n-6.
saturated fatty acid A fatty acid with the maximum complement of hydrogen atoms and no double bonds.
storage organ Any structure that sequesters storage materials (usually the carbohydrates, starch or glycogen, or triacylglycerol) for export to other tissues that utilize it.
triacylglycerol Ester of glycerol and three fatty acids; the most common lipid storage molecule in animals and green plants.

MOST ORGANISMS STORE LIPIDS and/or carbohydrates for energy production when food is unobtainable.
Carbohydrates are bulky but easily synthesized and transported. Much more energy per unit mass can be stored as lipids, and because they are less dense than water they aid buoyancy; however, long-chain fatty acids are insoluble and must be transported as lipoproteins. Vertebrates and higher arthropods have tissues specialized for storage and management of lipids for provisioning the rest of the body and for gamete formation. Adipose tissue can comprise 50% of the body mass before migration or breeding fasts with superficial depots expanding the most, especially in large animals. Up to 85% of adipocyte volume can be triacylglycerols and occur in various intra-abdominal, intermuscular, or superficial sites in all tetrapods and some fish; their histological appearance is similar but biochemical studies reveal that some depots have site-specific properties consistent with specialized local functions.

I. WHAT CAN BE STORED?

Few organisms have continuous access to nutrients derived from feeding or generated by photosynthesis; therefore, some storage is necessary. The role requires molecules that are insoluble, compact, and nontoxic to the holding cell, even when present in large quantities. Thus, oxygen, which is only slightly soluble in water, is stored in substantial quantities only when bound to special pigments in blood or muscle. Vertebrates and a few invertebrates and micro-organisms can hold oxygen bound to the iron-based pigment hemoglobin. Vertebrate cardiac and skeletal muscles contain the closely related oxygen-binding molecule myoglobin. However, even specialized diving animals such as seals and whales, whose muscles are dark red because of the high concentration of myoglobin, can store enough oxygen to support strenuous activity for at most about 1 h. Stores held by sedentary invertebrates and microorganisms may last longer, but storage capacity for oxygen is still very minor compared to that of other nutrients.

Many vitamins and essential minerals, including iron, are too toxic to be stored in more than small quantities, and then only when bound to specialized proteins, usually in the liver. In many kinds of animals, the skeleton can serve as a store of calcium and phosphate; in mammals, bone is laid down before breeding and withdrawn during lactation as the minerals passing to the offspring in the milk. However, the quantities thus stored are limited by the extra weight of the dense minerals and the need to maintain normal functioning of the mother’s skeleton. The materials stored by the widest variety of organisms and in the largest quantities are those that can be broken down to provide metabolic energy.

Energy-providing molecules that animals can safely store in large quantities are the complex carbohydrates, glycogen, and triacylglycerols—esters of glycerol and three fatty acids. The complex carbohydrates generally require less energy for synthesis but they are denser and less compact than lipids because they are associated with large quantities of water. Lipid molecules have such high affinity for each other that they form almost pure droplets that are significantly less dense than other cell components, but their synthesis and reclamation are biochemically quite complex and usually slower than those of carbohydrates. Although carbohydrates (and the glycerol component of lipids) can be broken down by aerobic or anaerobic metabolism, long-chain fatty acids must be oxidized aerobically; therefore, storage lipids can only fuel prolonged, strenuous activities in animals in which the uptake and transport of oxygen are highly efficient.

II. STORAGE IN INVERTEBRATES

Many invertebrates are capable of surviving long periods without food, but there is little detailed information about how these fasts are sustained. Many unicellular organisms and lower invertebrates have intracellular storage molecules and organelles that contain sufficient energy-generating material to sustain them for hours or days. In prolonged starvation, the body fabric, including structural proteins, is broken down to produce metabolic energy. The first to be reabsorbed are usually
reproductive tissues, early embryos as well as gametes, but at least in less complex invertebrates, skeletal, locomo-
tory, and digestive tissues can all be utilized if neces-
sary. Some animals can shrink to an impressive extent while remaining capable of normal life functions and can rebuild themselves when food is again available. For example, after 4 months of starvation, the medusae of the hydrozoan jellyfish _Aurelia aurita_ shrink to less than one-fourth of their original diameter but they continue to swim and can regrow if adequate feeding is restored. The ability to retain more or less normal function after prolonged starvation may be one of the advantages of the simple cnidarian body plan. More complex animals that have hard skeletons cannot lose as much body fabric without serious impairment of function.

Vertebrates and higher invertebrates have specialized storage tissues that sequester energy-providing molecules at much higher concentrations than is normally possible when materials are stored in cells with other primary functions. Annelids, especially oligochaetes, have variable quantities of chloragogenous tissue (so called because cell inclusions confer a green or yellow color) in the coelom around the intestinal wall. It has several functions, including that of storing lipids and glycogen, and may be ancestral to the arthropod "fat body" or "hepatopancreas" as it is known in crustaceans. The insect fat body is the most thoroughly studied invertebrate storage organ. It is present in larval and adult insects, and although originating in the abdomen it can grow to be massive, extending into the thorax in which it surrounds the major locomotory muscles. The stored lipids confer a white or yellow color when replete, as in locusts before migration and many other species before metamorphosis, but the proportion of lipid is rarely as high as it can be in adipose tissue of laboratory rodents. Storage lipids reach the muscles and other tissues as diacylglycerols or triacylglycerols in lipoproteins called lipophorins. Several hormones involved in energy metabolism and activity were shown to be secreted from the fat body in the 1950s and 1960s, leading to the suggestion that the insect fat body had gland-like functions comparable to those of the vertebrate liver. Recent interest in invertebrate immunity has revealed that it also contributes to innate immunity. When stimulated by bacterial extracts, _Drosophila_ fat body cells produce at least seven antimicrobial peptides. Like vertebrate adipose tissue, invertebrate storage organs are much more than just repositories; they play an active role in energy metabolism and defense against disease.

### III. Plants

The aerial structures of many temperate-zone and arctic plants die back in winter when cold or lack of light make photosynthesis inefficient. Regrowth in spring is supported by storage materials distributed through roots and surviving parts of the stem, which may be enlarged to form corms or tubers in some species. These structures are not severely limited in volume and density, so the storage material is nearly always starch with much associated water. Only a few higher plants, notably sugar beet (_Beta vulgaris_), store large quantities of disaccharides or monoosaccharides. Such concentrated nutrients are often protected from herbivores by stout casings or by the presence of toxic secondary compounds. Removing or inactivating these plant defenses by cooking (e.g., potatoes, carrots, and onions) or by leaching (e.g., cassava) was among the major technological advances of prehistoric humans and greatly increased the range and nutritional quality of plant foods available to them.

#### A. Energy Storage in Seeds

The main function of energy stores in seeds is to supply the germinating seedling until it grows its own leaves and can photosynthesize on its own. The huge range of sizes of seeds is due mainly to the chemical composition and abundance of their storage material. The storage materials of the seeds of plants native to cool or cold climates, such as oats, rye, acorns, beechnuts, and hazelnuts, are usually starches, as are those of grasses such as rice and wheat. Seed starches generally contain less water and therefore are more compact than those found in vegetative storage tissues. Like these storage tissues, seeds are often protected from herbivory by toxins. The need to compress much energy into a small space makes synthesizing triacylglycerols from primary photosynthetic products worthwhile, and the storage materials of many large, longer-lasting seeds that are dispersed by animals are lipids or mixtures of carbohydrates and lipids.

Green plants have a special biochemical mechanism, not known to be present in animals other than a few kinds of parasitic worms, that permits the incorporation of fatty acid carbon atoms into precursors for the synthesis of glucose, which is the starting point for the formation of the structural carbohydrates that build the vegetative structures that enable them to synthesize sugars in sunlight. By converting the fatty acid components of triacylglycerols into carbohydrate, plants can...
use seed lipids in ways that animals cannot use their lipid stores. Because they can be used in this way, instead of being oxidized in mitochondria or incorporated into phospholipids, seed lipids contain a much wider variety of fatty acids than do animal triacylglycerols.

The triacylglycerols in seeds are always dispersed into tiny compartments called oil bodies, each only a few micrometers (thousandths of a millimeter) across, that are enclosed in membranes. At least some of these membranes contain lipases that enable plant storage lipids to be mobilized quickly. Such structures contrast with the (relatively) huge droplet of triacylglycerols in vertebrate adipocytes that may be more than 100 μm (0.1 mm) in diameter or up to 1 million times the volume of oil bodies in seeds. This arrangement explains why nuts do not feel or taste as "greasy" as do vertebrate adipose tissue and purified oils even though they may be more than 50% lipid. Mobilization of seed storage materials begins at germination with the uptake of water that swells the proteins and starches. Preformed enzymes are activated and new ones synthesized, but, as in the case of animal systems, water-based enzymes cannot easily attack large droplets of lipid.

B. Chemical Composition of Plant Lipids

To take advantage of a short growing season, the seeds germinate early in the spring while the soil is still cold. The temperature at which major physiological processes occur seems to be the most important determinant of the composition of plant lipids. Most biochemical reactions occur in solution so the reactant molecules must be dissolved for metabolic processes to proceed. Unsaturated fatty acids make phospholipid membranes more fluid, so other molecules can move within and across them more readily, and also lower the freezing temperature of triacylglycerols. Complete breakdown of unsaturated fatty acids releases slightly less metabolic energy (about 1 or 2% less for each double bond) than the saturated equivalents with the same numbers of carbon atoms. Triacylglycerols containing saturated fatty acids also pack more neatly (i.e., they have higher melting points) than those containing mono- and polyunsaturates, and homogeneous assemblages solidify at higher temperatures than do mixtures. Therefore, in general, more energy can be stored in a small space by means of saturated triacylglycerols, the more homogeneous, the better.

All the plants in which saturated fatty acids are abundant as storage lipids are native to the warm, equable climates of tropical lowlands. Plants such as coconut, oil palm, nutmeg, and cacao, which live only in hot climates, do not need polyunsaturated fatty acids to keep their storage lipids fluid. They have the luxury of being able to maximize the energy obtainable from the smallest volume of storage organ because they have triacylglycerols that consist mostly of saturated fatty acids. Those of the seeds of the cinnamon tree, for example, are up to 95% lauric acid (C12:0), although the fatty acids in the lipids of the fruits of the same species and of its relative, the avocado pear, are approximately 90% oleic acid and 2% palmitic acid, with very few medium-chain saturates—a typical composition for an oily rather than a fatty fruit. Large seeds such as oil palms that contain saturated triacylglycerols provide their seedlings with supplies that last weeks or months. Plant oils rich in polyunsaturated fatty acids are found in species adapted to temperate climates. Olive trees, with fruits and seeds containing mostly monounsaturated fatty acids, and nuts such as walnuts and pistachios, which contain both storage lipids and starch, grow in cooler climates than that required by oil palm or cacao but are still not worth cultivating for their oil outside areas that have long, hot summers. Lipids containing mixtures of different kinds of fatty acids remain liquid at lower temperatures than do those in which all the fatty acids are similar, therefore, those of nearly all temperate-zone and polar species are mixtures of many different fatty acids. However, several possible mixtures of fatty acids in triacylglycerols may have similar physical properties and each plant may be adapted to germinate under slightly different conditions, so the fatty acid composition of seed lipids may differ between species growing in similar climates. Some species can make major alterations to the composition of their seed lipids according to the climate in which they are growing. The fatty acids of the seeds of the sunflower (Helianthus annuus) are 44–72% linoleic acid (18:2n-6), with the remainder being mostly oleic acid (C18:1). Those growing in the coolest conditions contain the most linoleic acid, but their total yield is smaller and the crop takes longer to mature. The seeds of the jojoba bush (Simmondsia chinensis) are almost unique in that the storage material is a wax and not a triacylglycerol. Jojoba oil is liquid at normal temperatures (in contrast to beeswax, which is used by bees as a solid) because both the fatty acid and the alcohol components of the wax are long-chain molecules. The fatty acids are mostly (approximately 74%) C20:1, with smaller quantities of C22:1 and oleic acid (C18:1), and the alcohols also contain 20 or 22 carbons. Although the proportion of lipid in the seeds varies widely between strains and according to the conditions under which the plant is grown, the fatty acid composi-
tion of the wax, in contrast to that of triacylglycerols, is remarkably constant. Jojoba oil, like other waxes, is indigestible to most animals but the plant is cultivated widely in California and other warm, semidesert areas of the United States, where the extracts are used as lubricants and in cosmetics. Jojoba oil would also be suitable as fuel, so the crop has been proposed as a renewable substitute for gasoline.

C. Protection from Seed Predators

The lipids and starches concentrated in grains, nuts, and seeds provide a rich source of food for animals including humans. The partitioning of seed lipids into tiny oil bodies makes them very digestible to animals because less mechanical emulsification is required. Seed-eating mammals such as rodents gnaw or crush seeds with their hard, continuously growing teeth, whereas parrots, finches, and other seed-eating birds use their powerful, finely controlled beaks to shred nuts and seeds into tiny fragments before swallowing them. Nuts are digestible, highly nutritious food that is especially suitable for flying birds, which cannot afford to carry around heavy gut contents for long periods of slow digestion. Seeds such as walnuts, almonds, cacao (chocolate) beans, and coconuts protect their nutritious contents with a hard shell (in many species, the nutshell is much harder than the tree's wood) that only the most powerful beaks or persistent gnawing teeth can open. Other nuts, such as peanuts, form underground, avoiding seed predators that cannot dig to reach them. Small seeds such as those of wild strains of flax, sesame, and sunflower scatter as they fall, making it difficult for animals to collect them, and their coats are very tough relative to their size. Many plants have evolved chemical and mechanical means of deterring predation on their seeds.

Some seeds are rendered poisonous to seed predators by the presence of toxic amino acids, proteins, and alkaloids and/or because the storage materials are poisonous to animals. The beans of the castor oil plant, Ricinus communis, are protected from herbivory by the poisonous protein ricin and by the fatty acid ricinoleic acid, which can account for more than 90% of the triacylglycerol fatty acids. Animal lipases cannot easily hydrolyze triacylglycerols containing ricinoleic acid, so most pass undigested through the gut. Eaten in small quantities, castor oil acts as a lubricant and slight irritant, thus relieving constipation in people and dogs and cats. Oil from unimproved oilseed rape (genus canola oil) is produced from plants selectively bred, and recently genetically engineered, to form seed triacylglycerols that contain as little as 3% erucic acid. Insects as well as people find the artificially modified plants much more appetizing than the wild forms, and more pesticides are needed to protect such crops from herbivore damage.

The caterpillars of the cabbage white butterfly feed exclusively on the leaves and flowers of cabbage and related brassicas, including cauliflower, broccoli, and oilseed rape, but they have not acquired more of a taste for erucic acid than have mammals. Even when present in quite high concentrations in the diet, the unusual fatty acid is absent from phospholipids and triacylglycerols in the insects' tissues. Either it is not absorbed through the gut or not esterified into triacylglycerols or, less likely, it is all used as fuel at once, before it has a chance to become incorporated into structural or storage lipids. Erucic acid seems to be toxic even to the plants. Highly specific enzymes direct it into seed triacylglycerols only. None is found in phospholipids or any other non-storage lipid, suggesting that its presence in such molecules would disrupt membrane function.

IV. VERTEBRATE ADIPOSE TISSUE

Proper adipose tissue is unique to vertebrates. It occurs sporadically among fish, often in a greatly modified form, but it is most extensive in mammals, birds, reptiles, and amphibians. Its distinctive feature is a unique type of cell called an adipocyte, which can accommodate much larger quantities of triacylglycerols than any other kind of cell. Almost all the adipocytes' stores of lipid are not for their own use but for export to other tissues as required. Most physiological studies of adipose tissue are aimed at controlling obesity or diabetes, and the primary purpose of most studies of plant storage tissues is food production. Consequently, the basic principles governing the cellular structure, gross anatomy, and metabolism of these tissues are only slowly coming to light.

Compared to other animal cells, replete adipocytes are huge, thousands of times larger than red blood cells, most brain cells, and the cells of the immune system that protect the body from disease. In Fig. 1, only the top layer of adipocytes is in the plane of focus and because the tissue is not stained, the nucleus, cyto-
plasm, and lipid droplet cannot be distinguished. Unless severely depleted of lipid, adipocytes are spherical, or very nearly so, which is an unusual shape for functionally mature animal cells (other than eggs). Adipocytes are packed closely in a tight mesh of collagen, so together they occupy most of the volume of the adipose tissue.

The cells' most distinctive feature is that most of the cytoplasm is occupied by lipid droplets, a single large one in white adipocytes and many smaller ones in brown adipocytes. Comparative studies of angiosperm seeds and fruit, yeasts, and various animal cells show that the intracellular mechanisms of the formation of lipid droplets are essentially similar in all eukaryote cells; however, the process is greatly exaggerated in adipocytes. Lipid droplets arise from specialized regions of the endoplasmic reticulum and are controlled by specific enzymes and other associated proteins called oleosin in seeds and perilipin in adipocytes.

The compartmental arrangement of plant storage lipids is not necessary in white adipocytes because such cells do not metabolize significant amounts of their own triacylglycerols. However, brown adipose tissue oxidizes its own lipids to generate heat, although when such supplies run low it can, like most other animal cells, take up metabolites from the blood circulation. The similarity in size and general appearance between plant oil bodies and brown adipocytes may arise from the fact that in both cases storage lipids are utilized in the cells in which they are sequestered.

A. Size and Number of Adipocytes

Among mammals and birds (there are no data for other classes), the adipose tissue of larger species consists of fewer, larger adipocytes. Adipocyte number scales to $(\text{body mass})^{0.75}$, and the adipocytes range in volume from 0.01 nl in bats and shrews to up to 4 nl in well-fed baleen whales. Carnivorous mammals and ruminants (which utilize mainly fatty acids rather than glucose) have about four times more adipocytes than do non-ruminant herbivores (whose energy metabolism is based mainly on glucose) of the same size and fatness; however, they are not fatter because the adipocytes are smaller. By coincidence, the adipocytes of rats, small non-ruminant herbivores, are about the same size (0.1–1 nl) as those of humans, who are large omnivores on a high-fat diet.

Thorough studies of many specimens of the same species in the wild reveal much interindividual variation in the total number of adipocytes in the body as a whole that cannot be attributed to age, sex, or any obvious feature of dietary history. For example, most Svalbard reindeer, which accumulate enough fat during summer and autumn to sustain them during the long winter when only low-quality dry vegetation can be reached by digging through the snow, have only two or three times more adipocytes than would be expected in temperate-zone and tropical mammals of similar size.

The number of adipocytes (relative to total body mass) proved to be even more variable in carnivores. Some arctic foxes (Alopex lagopus) and wolverines (Gulo gulo) have up to five times as many adipocytes as expected from their overall size, but a significant minority actually have fewer than the expected number. At the time the specimens were collected, there was no consistent relationship between the total number of adipocytes and fatness, measured as the relative mass of adipose tissue in the body. The number of adipocytes in...
the adipose tissue is not a major determinant of the capacity for fattening. If fewer adipocytes are present, each must expand to a larger volume to accommodate as much storage lipid as that in specimens that, for whatever reason, have proportionately more adipocytes.

In humans, and also in laboratory animals kept under highly artificial regimes of diet and exercise, the number of adipocytes increases with age after sexual maturity, but all these effects are slight, generating at most a doubling of the adipocyte population. They cannot be induced in naturally obese wild mammals such as dormice (Glis glis), and there is almost no evidence for an increase in adipocyte number after growth of lean tissue stops in any wild mammal investigated to date. Until recently, biologists believed that once fully formed and functional, adipocytes did not die during the lifetime of their owner; however, adipocytes and preadipocytes, like most other cells, have been found to disappear by apoptosis, an orderly and often energy-consuming process in which the cells synthesize enzymes that destroy their fabric. Apoptosis is rarely observed because such large cells do not die very often and death is swift, probably only a few hours from the first signs of decline to total disintegration.

In people and sedentary domestic animals, the proportion of lipid in whole adipose tissue is rarely less than 40% and it increases with fattening, reaching as much as 85% of the total mass of the tissue in middle-aged men who are 44% by weight adipose tissue. The rest of the tissue is mostly protein and water. The proportion of lipid is always lower in wild animals and does not change significantly with fatness. Dwarf hamsters (Phodopus) naturally become very fat during the summer, but chemical analysis shows that their adipose tissue never exceeds 40% lipid, even when it amounts to 35% of the body mass. The proportion of lipid in adipose tissue is almost constant (42–66% depending on the depot) in arctic foxes living wild in the high Arctic, whose body composition ranges from 3 to 33% of the body mass. The proportion of lipid in whole adipose tissue is rarely less than 40% and it increases with fattening, reaching as much as 85% of the total mass of the tissue in middle-aged men who are 44% by weight adipose tissue.

Mammalian adipose tissue is always partitioned into a few large and many small depots. Its comparative anatomy is based on measurements of the size and number of the adipocytes. Figures 2 and 3 show the patterns of distribution of larger and smaller cells for various mammals. Larger adipocytes do not always make larger depots. Some of the smallest depots, such as the popliteal behind the knee, consist of relatively large cells. Nor do adjacent depots, such as those on either side of the forelimb, necessarily consist of cells of similar size.

In the adult camel (Fig. 3A), more than one-third of all adipose tissue is in the humps, compared with 4–8% in the corresponding site of hares (Fig. 2B), up to 4% in squirrels, and a maximum of 1% in stoats (Fig. 2C). However, less than 1% of the camel’s adipose tissue is in the inguinal depot, on the front of the thighs and the sides of the abdomen. This depot contains more than one-third of the total adipose tissue in stoats and squirrels, more than one-fifth of that of the lion (Fig. 3C), and one-eighth of that of horses (Fig. 3B). There are also contrasts in the proportions of adipose tissue located on the dorsal wall of the abdomen, from nearly half in squirrels and hares (Figs. 2A and 2B) to one-eighth or less in the large animals (Fig. 3). Although the range of sizes of adipocytes differs greatly between species and between specimens, partly because some species were faster than others, the anatomical pattern of relative sizes of adipocytes is similar in all terrestrial mammals. What differs between species is the relative abundance of adipocytes in each depot. The relative size of their adipocytes forms the same general pattern. Taxonomically related species have similar patterns of distribution of adipose tissue.

Although much is known about the biochemical steps by which adipocytes differentiate and mature, almost all of the information was derived from the study of cells in culture so that it cannot be explained in detail why adipose tissue forms where it does or what determines the species differences in the relative sizes of the depots. The gross anatomy must have implica-
STORAGE, ECOLOGY OF

...ions for the animals’ habits and capabilities because there are site-specific differences between adipocytes in many physiological properties. Some depots seem to be equipped mainly for taking up lipids from the blood after a meal, whereas others seem to be equipped for quick responses to the onset of strenuous exercise. Many of the minor depots that enclose lymph nodes may be mainly or entirely devoted to supporting immune processes and contribute little to whole-body energy storage. They are among the first to form and the last to be depleted in starvation. In very lean wild animals such as rabbits, the only replete adipocytes may be those associated with lymph nodes.

Relatively minor depots may be almost undetectable in small animals, especially if the specimens are lean. Therefore, perhaps it is not surprising that larger animals (Fig. 3) appear to have more “extra” adipose depots for which there are no exact equivalents in the smaller species (Fig. 2). Some are just too large to be thus explained away. Up to one-third of all adipose tissue of horses (Fig. 3B) and donkeys forms a sheet, sometimes several centimeters thick, on the inner ventral wall of the abdomen. There is a small quantity of adipose tissue in the corresponding depot of large carnivores and camels (Figs. 3A and 3C), but it is absent in most other species.

In primates, the most conspicuous such depot is the “paunch” that arises from the midline on the outer wall of the abdomen and expands laterally as it thickens, sometimes becoming very thick at the midline and extending as far as the chest and the hip bones. In humans, the paunch adipose tissue is centered around the navel and is usually slightly thicker above it in men and below it in women. The paunch depot seems to accumulate fat selectively in primates. It may be minimal in lean (but not emaciated) specimens, but it often becomes very massive in humans and also in monkeys and apes that become obese in captivity, whereas the other depots expand more slowly with increasing fatness or, in the case of the intermuscular depots, expand hardly at all. An abdominal paunch with these properties seems to

FIGURE 2
The organization of adipose tissue in various wild mammals. (A) The gray squirrel (Sciurus carolinensis): Data are the means of four specimens; body mass, 0.38–0.67 kg; 3.1–14.4% dissectible adipose tissue. (B) European hare (Lepus timidus scotus): Data are the means of five specimens; body mass, 2.4–3.5 kg; 0.3–3.2% fat. (C) Stoat (Mustela erminea): Data are the means of five specimens; body mass, 0.19–0.4 kg; 3.2–6.3%. Top sets of data are the mean volume of adipocytes in nanoliters (10⁻¹¹ liters); bottom sets of data are the proportions of the total dissectible adipose tissue in the depot. The intermuscular depots are shaded (data from Pond, 1998).
be a unique and distinctive feature of this group, and
it is as conspicuous in lemurs as in apes, indicating
that it appeared early in their evolutionary history. De-
spite having this extra depot, with the exception of
modern humans, primates are generally no fatter than
other mammals when living under natural conditions
(although some do become obese in zoos).

There is no trace of this depot in hares, squirrels, or
any related small mammals, but Carnivora have a
paunch depot on the outer ventral wall of the abdomen.
The depot can become thick in cats and their relatives
but it is relatively small in canids, ursids, and mustelids.
It does not expand disproportionately any more than
any other superficial depot, even in very obese wild
carnivores such as arctic foxes. Very little is known
about the physiological basis for its special properties
because the paunch depot is completely absent in ro-
dents so it cannot easily be studied in the laboratory.

C. Measuring Fatness in Wild Animals

Because adipose tissue is the most variable component
of the body, the simplest and most widely used measure
of fatness is body mass, sometimes refined to include
a term for body length. In small animals, the guts repre-
sent a large and variable fraction of the total mass that
such methods are not very accurate or reproducible. If
samples of adipose tissue can be obtained by biopsy or
post mortem, adipocyte volume is also used to estimate
fatness, but such measurements are not as reliable for
most wild animals as experiments on laboratory rodents
might suggest. It is essential to correct for the effects
of body size and natural diet on the cellular structure of
adipose tissue, for site-specific differences in adipocyte
volume, and for variability in the cellular structure of
the adipose tissue.

Figure 4 shows data on adipocyte volume from wild
polar bears. Females (but not males) stop growing when
they become sexually mature and fatten enormously
during pregnancy, producing a predictable relationship
between body mass and adipocyte volume. The correla-
tion is weak for cubs and males that are still growing
because adipocyte number is also increasing. However,
these data illustrate some important aspects of bear
ekology. Cubs still with the mother are generally fatter
than those of similar size that have been abandoned to
forage alone, and there is great variation in the fatness
cubs and adult males, especially among the larger
specimens, probably reflecting differences in hunting
success.

For children and young adult humans, measuring
skinfold thickness with calipers is widely used to assess
fatness, but the skin of most mammals is too thick and
tough and the underlying adipose tissue too thin and
patchy for skinfold thickness measurements to be of
much use. As wild mammals become fatter, superficial
adipose tissue expands more rapidly than that inside
the abdomen. The depots shown in Figs. 2 and 3 become
thicker and expand laterally to cover a greater area. In
very fat specimens, including many humans, adjacent
superficial depots overlap and merge so that they be-
come difficult to distinguish, but measurements of adi-
pocyte volume can reveal their identity. Since ex-
panding depots spread out as well as thicken, thickness
alone is only a crude measure of the total adipose tissue,
even if the site from which the measurement is taken
is kept strictly constant. The fatness of small birds is
often assessed by inspection and/or palpation of the
superficial depots around the neck or leg. This simple
method takes account, at least in a qualitative way,
of changes in both the thickness and the area of the
adipose tissue.

A high-tech version of skinfold thickness measure-
ments is the ultrasound scanner, which can measure
fat thickness fairly accurately in large mammals, espe-
cially if, as in pigs, the hair is sparse. The method
depends on the fact that sound vibrations travel at dif-
f erent speeds through chemically different materials
and are reflected, producing an echo, at the interfaces
between the two tissues. A pulse of high-frequency
sound is directed into the body from a source placed
on the skin, and the timing and intensity of the echo
are recorded. With a resolution of only about 1 mm at
best, ultrasound scanners are not accurate for animals
smaller than dogs and for any animal in which the
superficial adipose tissue is very thin.

Lipids are much poorer conductors of electricity than
the watery components of cells and body fluids. This
contrast has been exploited in the development of an
apparatus that measures whole-body conductance of
electric current as an indicator of the proportion of
lipid present. More fat produces lower whole-body con-
ductance. The apparatus has to be used carefully be-
cause hair, feathers, and the pockets of air they enclose
also have high resistance to electric currents. However,
it is a useful technique, especially for small birds and
mammals, and it works equally well on living or freshly
dead specimens.

Because these new techniques are impractical for
most large animals, dissection is still widely used. For
more than 30 years, assessment of lipid reserves in
wild vertebrates in vivo and post mortem has depended
mainly on empirically derived “indices” which are based
on the dimensions of the whole body or that of the
thickness or mass of one or a few adipose depots. The sample sites are chosen at random or for some technical reason such as simplicity of dissection. The enormous diversity of sample sites and indices is cumbersome and makes comparisons between studies very difficult and imprecise. One of the favorites is the kidney fat index, the mass of the perirenal adipose tissue as a percentage of that of the kidneys. This index works fairly well for lean deer, antelope, and similar ungulates, unless the mass of the kidneys changes seasonally, which can happen in some drought-adapted species. It is less accurate in carnivores (see Figs. 2 and 3), in which the perirenal depot represents only a small fraction of the total adipose tissue and is indistinguishable from that on the rest of the dorsal wall of the abdomen.

Wildlife biologists also try to estimate total fatness of mammals from the lipid content of the marrow in the long bones, thereby eliminating the need for a complete dissection. However, “narrow fat indices” only work on species such as ruminants in which the narrow adipocytes are depleted in starvation. In many other mammals, these adipocytes fail to respond to fasting, probably because they are specialized to support the immune functions of the red marrow.

An additional problem with using the mass of one or a few depots as an index of fatness is the inherent variability of the partitioning of adipose tissue between depots. In all species for which there are sufficient data, there is substantial inter-individual variation in the relative masses of depots that cannot be attributed to age, sex, or lean body mass, even in genetically homogeneous wild populations and in animals bred and maintained under carefully controlled laboratory conditions. Of the intra-abdominal depots, the dorsal wall of abdomen depot (which includes the perirenal) undergoes the largest changes, especially in deer, sheep, cattle, antelope, and other ruminant animals. It expands greatly in extreme obesity and shrinks to almost nothing after prolonged starvation. Its anatomical position means that, like the fat bodies of lower vertebrates, such large and sometimes rapid changes in size have minimal impact on the shape or size of any adjacent organs. Depots that contain lymph nodes undergo the least change, so they are relatively massive (compared with those elsewhere in the body) in lean specimens and relatively small in obese ones.

V. ADIPOSE TISSUE FUNCTION

A. Adipocyte Metabolism

Molecular traffic into and out of adipocytes is regulated by a variety of receptors, carrier molecules, enzymes, and other substances, most of which are produced by the cells. As recently as a decade ago, adipocytes were known to produce only a few kinds of proteins. However, techniques to separate, identify, and synthesize proteins and other biological molecules are rapidly becoming cheaper and more efficient, so biochemists are constantly adding to the list of messenger molecules secreted from adipocytes, receptors that enable them to respond to such signals from each other and more remote tissues, and transport proteins and enzymes that regulate uptake of materials that at least some adipocytes can produce.

The genetic instructions for synthesizing such proteins are located in the nucleus, which is relatively small and is displaced from its usual position near the center. The lipid droplet (see Fig. 1) has no internal structure because it is almost pure liquid triacylglycerol. The thin rim of cytoplasm between the outer membrane and the lipid droplet contains the biochemical apparatus for...
building and storing receptors, messenger molecules, and enzymes plus a few mitochondria in which the fuel that powers their synthesis is produced. Leptin (formerly known as Ob protein) is among the best known secretions of adipose tissue, identified in the early 1990s from the study of a spontaneous mutation that causes severe obesity in mice. This small protein passes from adipocytes (and, it has recently been discovered, certain other tissues, especially those involved in reproduction) into the blood, in which it is found in concentrations of approximately $10^{-10}$–$10^{-9}$ M, and then to the brain, in which it binds to specific receptors on many neurons in the hypothalamus, the region of the brain that controls appetite (and many other drives and emotions). This small protein seemed to have all the properties of the long-sought signal that regulates appetite in proportion to the level of fat reserves in adipose tissue. It has been identified in rats, humans, and Sminthopsis macroura (Marsupialia: Dasyuridae) and possibly in fish. In addition to regulating appetite, leptin adjusts energy expenditure and the physiological response to cold. Obesity could be caused by disruption of these chemical signals from the adipose tissue to the brain, leading to chronic overeating and/or very low rates of energy expenditure. Excess lipids cannot be excreted; therefore, unless they can be oxidized to produce energy, they have to be deposited in the storage tissues. Most tissues oxidize lipids, but the largest consumers are cardiac and skeletal muscle, the liver, and, especially in neonatal mammals and hibernating species, brown adipose tissue. ATP production in mitochondria is never 100% efficient, a fraction of the energy used appears as heat. But heat production in brown adipose tissue can be greatly increased by "uncoupling" of fuel utilization from the formation of ATP in the mitochondria, implemented by a special protein. Until the late 1990s, uncoupling protein was believed to be unique to mammals and restricted to brown adipose tissue, but similar proteins have been found in a variety of tissues and organisms including green plants, although it is not clear that they all act to promote heat production and lipid utilization. A large fraction of the volume of brown adipocytes is occupied by mitochondria so that there is not much space for storing lipid, which is present as numerous tiny droplets instead of a single large droplet in storage white adipose tissue. Active brown adipose tissue is only about 10% lipid by weight. Like white adipose tissue, brown adipose tissue is richly innervated by the sympathetic nervous system, and in both cases nor-epinephrine stimulates lipolysis. In brown adipose tissue, the fatty acids produced remain in the cell and are consumed for heat production, whereas those of white adipocytes are released into the blood, in which other tissues can take them up and use them. When fully activated (e.g., during rapid warming of the body at emergence from hibernation), brown adipose tissue uses fuels and oxygen at up to 10 times the rate that muscles use them. More than 90% of the chemical energy it consumes is released as heat instead of being used to synthesize ATP, producing enough heat to raise the body temperature by more than 2°C per hour. In the newly born and in awakening hibernators, much of the brown adipocytes' fuel comes from white adipose tissue. The latter's stores therefore determine the neonate's ability to maintain its temperature until it can find food (mother's milk or a rich, digestible meal).

Under artificial conditions, brown adipose tissue can often be transformed into white and vice versa, and at least in certain species, it does so naturally. Many of the sites in which it actively generates heat in newborn mammals contain normal-looking white adipose tissue in the adult. In small mammals that live in cold climates, such as dwarf hamsters, at least some depots consist of a mixture of brown and white adipocytes throughout life. The proportion of space occupied by the brown ones increases when the animal is placed continuously in a cold environment, especially if the light–dark cycle is adjusted to resemble the long, dark nights of winter. In large mammals such as reindeer, brown adipocytes in newborn calves become white adipocytes within days, and cannot be reconverted back to brown.

### B. Fatty Acid Composition of Storage Lipids

Animals can synthesize the common saturated (palmitic, stearic, and myristic) and monounsaturated (oleic) fatty acids and many can insert an additional double bond into long-chain fatty acids, although this capacity is minimal in higher vertebrates including mammals. Animal desaturases work slightly differently from the corresponding enzymes in plants and protocists; therefore, fatty acids that originate in animals are distinct from those synthesized by primary producers. Except in omnivorous species (in which the ruminant microbomes alter the fatty acids before they can be absorbed), most fatty acids pass unaltered from plant to herbivore and from prey to predator. Therefore, fatty acids originally synthesized by microbes, protocists, or plants often appear unaltered higher up the food chain.

The proportions of the various fatty acids, and their arrangement in the triacylglycerol molecules of most
animal fats, are such that they are fluid at the higher body temperatures of birds and mammals. The lipids of cold-blooded animals generally contain a higher proportion of unsaturated fatty acids than those of warm-blooded animals, with species living in colder regions having the most polyunsaturates.

Recent studies suggest that the lipid composition of the diet may determine the habits and habitats of poikilothermic animals. The blue-tongued skink or shingle-backed lizard, Tiliqua rugosa, is quite common in the deserts of Western Australia. Its natural diet is mainly plants (at least when an adult) but also includes carrion and small prey; therefore, it can be maintained on an artificial diet in captivity. After a few weeks on experimental diets, those given food containing sunflower oil (rich in polyunsaturated fatty acids) chose to spend their days (and nights) in places that were up to 3°C cooler than those fed on mutton fat (containing mostly saturated fatty acids). These experiments show that diet can affect important aspects of their ecology, including the times of day at which they are active and where they forage.

The sea is always cold compared with the warm bodies of mammals and birds. Marine algae generally have more highly unsaturated fatty acids than terrestrial plants, and the animals that eat them both incorporate the algal fatty acids unchanged into their own tissues and elongate and/or desaturate them further. Thus, most plants cannot build fatty acids with more than 18 carbons, but many marine animals can elongate oleic acid to form gadoleic acid (C20:1 \( n \)-9) and various 22- and 24-carbon fatty acids. Gadoleic acid is common in the lipids of many kinds of fish, especially those living at high latitudes. Other fatty acids are also readily elongated, but those of the n-6 series derived from linoleic acid are even more unstable (in the sense that they readily participate in chemical reactions, especially oxidation) than those of the n-3 series; therefore, long-chain fatty acids are in more common marine organisms.

Algae are the basis of nearly all marine food chains, so the storage and structural lipids of almost all animals that feed in or from the sea—e.g., ragworms, mussels, oysters, shrimps, crabs, fish, seals, whales, polar bears, and penguins—are rich in n-3 polyunsaturated fatty acids. Such mixtures of triacylglycerols rich in polyunsaturates are often called “fish oils,” although they are not unique to fish, and the basic fatty acids from which they are derived are not even produced by fish. The lipids originate in the algae but may be modified by animals as they pass up the food chain, eventually reaching top predators such as large fish, seabirds, and marine mammals. Among the fish used as human food, cool-water species such as herring, sprats, capelins, mackerel, and halibut usually store more lipid than their tropical relatives because at high latitudes the supply of most kinds of food changes seasonally and many species routinely live on their fat reserves for weeks or months. The triacylglycerol fatty acids of large cold-water species include more monounsaturates than polyunsaturates because they are derived from waxes that are in the planktonic organisms that these species eat. Fish such as sardines, pilchards, and anchovies live in warmer waters in which the planktonic invertebrates rarely contain waxes, so their lipids are a rich source of n-3 polyunsaturated fatty acids for human consumers.

In addition to being chemically distinct, lipids of marine origin differ in the isotopic composition of the carbon atoms. Slightly different proportions of \(^13\)C and \(^12\)C are taken up by marine algae that obtain their carbon in solution as bicarbonate and terrestrial plants that take in carbon dioxide straight from the atmosphere; therefore, organic materials of marine origin contain about five to seven parts per thousand more of the heavier isotope \(^13\)C than similar molecules from animals that have eaten terrestrial foods. This effect can be used to investigate the dietary habits of individual animals and populations. Storage fatty acids and cholesterol are particularly suitable because they pass intact up the food chain, are chemically quite stable, and are only slowly degraded by microbes; therefore, they can be detected in excreta and animal remains as well as in biopsy samples. The isotopic and chemical composition of triacylglycerol fatty acids can indicate the proportion of the diet that is derived from marine or terrestrial sources. This technique is particularly useful for widely ranging species with broad, locally variable diets, such as arctic foxes, that are not easily observed in the wild. Material from the marrow of burned human and other animal bones and in coral fragments has been used to study natural diets and food chains.

Some of the food of freshwater animals is derived from aquatic algae, but detritus from terrestrial plants and the animals that feed on them also makes a substantial contribution; therefore, fish and insects living in lakes and rivers usually have more fatty acids of the n-6 series than do oceanic animals. The fatty acid composition of predatory fish follows that of their diet: “Farmed” salmon fed on chow made from scraps of slaughtered livestock and other materials originating from the land acquire a correspondingly “terrestrial” fatty acid composition.

Mammalian adipocytes can selectively take up and
release certain fatty acids (there are no data on those of other vertebrates), making possible seasonal accumula-
tion of triacylglycerols with physical properties appro-
piate to hibernation. Recently, site-specific differences
have also been described. The adipocytes around lymph
nodes, and to a lesser extent elsewhere, in node-con-
taining depots sequester more polyunsaturates that are
essential to the nutrition of the immune system than
do large nodeless depots. Such variation must be consid-
ered when selecting tissue samples for the study of diet.

C. Adipose Tissue as Thermal Insulation

The notion that the superficial adipose tissue insulates
the body against heat loss is one of the most firmly
established of all dogmas in twentieth-century biology,
but there is very little evidence that supports it. The
hypothesis can be tested by determining whether its
distribution and anatomical arrangement have evolved
to carry out such a role more efficiently in animals
that need more thermal insulation because they remain
active during winter in cold climates or habitually swim
in very cold water. The mammalian order Carnivora
occur in a wide variety of habitats from hot tropics
(e.g., desert foxes and jackals) to the Arctic (e.g., polar
bears, wolverines, and arctic foxes) and almost all are
predators, at least to some extent. Carnivora occur in
an enormous range of sizes, from stoats and weasels to
bears and tigers that can be several thousand times
larger, and they are therefore suitable for investigating
whether species differences in the relative abundance
of superficial adipose tissue are determined by habits
and habitat or by body size.

Figure 5 shows some measurements of the masses
of all the intra-abdominal and superficial adipose tissue
from various temperate, tropical, and arctic carnivores.
To ensure comparability, the sample included only
specimens that were of similar fatness to the bears.
There is a fair amount of inter-individual variation in
the partitioning of adipose tissue between internal and
superficial depots, but overall the mass of superficial
depots increases in simple proportion to body mass. In
contrast, the intra-abdominal depots become propor-
tionately smaller in larger mammals, no matter whether
the animals are native to the Arctic or to warmer hab-

Iturritia brown bears

Otter (temperate)

Otter (cold)

Otter (arctic)

Carnivora

MAX

MIN

FIGURE 5 Allometric comparison of the masses of superficial (solid
symbols) and intraabdominal (open symbols) adipose depots in vari-
ous species of the order Carnivora. Each point represents a single
individual. Squares, Mustelidae (mink, badgers, and wolverines); dia-
monds, Felidae (cats, jaguars, tigers, and a lion); circles, Canidae
(arctic foxes); triangles, Ursidae (inverted triangles, brown bears).
The lines are fitted to all the data except those from the polar bears
(triangles) and fit the following: superficial depots
\( \frac{\text{Mass of adipose depots}}{\text{Lean body mass}} = 1.05, r^2 = 0.867 \), intraabdominal depots
\( \frac{\text{Mass of adipose depots}}{\text{Lean body mass}} = 0.81, r^2 = 0.873 \) [data from Pond and Ramsay (1992) and additional
sources].

The measurements from the wild polar bears are
very near to the lines fitted to the data obtained from
tropical and temperate-zone carnivores, indicating that
the partitioning of adipose tissue between internal and
superficial sites arises entirely from the fact that polar
bears are very large and produce large quantities of fat,
and it has nothing to do with adaptation to thermal
insulation. There is no evidence for reorganization of
adipose tissue in the smaller species of mammals that
have evolved to become aquatic; for example, otters
have many adaptations of the trunk, tail, and sense
organs to their life as predators in lakes and rivers, but
the arrangement of their adipose tissue is not different
from that of their terrestrial relatives, stoats, weasels,
wolverines, and badgers. Otter fur is thicker and less

a smaller area to cover, the superficial depots are thicker
in larger animals, even if the specimens are not fatter
(i.e., if the ratio of the masses of lean and fat tissues is
unchanged). The superficial adipose tissue is 10 times
thicker in a 500-kg bear than in a 500-g rat or weasel
from this effect alone, quite apart from the fact that in
larger animals, a greater proportion of the adipose tissue
is superficial and bears are often fatter than rat-sized
animals.

The measurements from the wild polar bears are
very near to the lines fitted to the data obtained from
tropical and temperate-zone carnivores, indicating that
the partitioning of adipose tissue between internal and
superficial sites arises entirely from the fact that polar
bears are very large and produce large quantities of fat,
and it has nothing to do with adaptation to thermal
insulation. There is no evidence for reorganization of
adipose tissue in the smaller species of mammals that
have evolved to become aquatic; for example, otters
have many adaptations of the trunk, tail, and sense
organs to their life as predators in lakes and rivers, but
the arrangement of their adipose tissue is not different
from that of their terrestrial relatives, stoats, weasels,
wolverines, and badgers. Otter fur is thicker and less
mammals feed normally while they are pregnant and accumulate lipid in preparation for lactation. The reproductive organs and the gut cannot be excluded from the abdomen, but adipose tissue can be excluded. The subcutaneous depots may simply be the most convenient place to put a lot of adipose tissue. In pinnipeds, cetaceans, and sirenians, the cellular and gross anatomy of adipose tissue is substantially altered to form blubber. Heat conducts through adipose tissue at only slightly less than half the rate of that through muscle or stagnant water and much faster than through fur or feathers. Therefore, blubber has to be thick to insulate thoroughly, and it is satisfactory as the sole form of insulation only for large animals with habits and habitats for which weight is not a major problem. Consequently, most small and medium-size seals retain fur as additional insulation, and extant sirenians and the smallest cetaceans, porpoises, and dolphins are restricted to temperate and tropical regions. Many of the great whales (e.g., humpbacks and fin whales) that feed in polar seas migrate to warmer waters to give birth to calves that are only one-twentieth of their size. Many such places provide little or no food for adult whales, so the mothers fast for several weeks while lactating, drawing on their lipid reserves. The whales do not return to cooler, food-rich waters until their calves have grown sufficiently large and fat for blubber insulation to be effective. Seals are born on beaches or ice floes and most species do not go to sea until they are several weeks old. Neonates have thick coats of fluffy fur but little adipose tissue, although they fatten rapidly while suckling their mothers' rich, creamy milk. At about the time the seal pups are weaned, the neonatal fur is replaced by shorter, stiffer fur. However, fur of some sort continues to make an essential contribution to insulation in all seals except the adults of the largest species, the elephant seals (Mirounga spp.). In walruses (Odobenus rosmarus), the hair is very sparse throughout life, and at birth the single pup is very large compared to its mother, up to twice the size of neonatal elephant seals, although as adults the latter are much larger.

Both these very large pinnipeds occur in cold seas, Odobenus in the Arctic Ocean and one species of Mirounga in the northern Pacific including near Alaska and the other in the Southern Ocean. The skin of walruses hauled out on beaches basking in sunshine is noticeably pink due to the profuse flow of blood through it and the underlying blubber, which can be up to 15 cm thick. However, walruses are a ghostly pale gray when in the sea or exposed to cold weather because the blood is
withdrawn from the skin and blubber to the muscles and other warm internal tissues. Blubber can work as an adjustable insulator in this way because blood flow through it and the skin can be almost completely shut down for long periods without ill effects.

If necessary, marine mammals can be very fat. As in the case of penguins, body bulk is only a slight impediment to swimming, and seals and whales can maintain near neutral buoyancy over a wide range of body compositions by adjusting the volume of air in the lungs. The maximum amount of blubber that whales or seals can have without risking overheating during vigorous swimming depends on body shape, the temperature of the water in which they are living, and how much heat they generate. Marine mammals dissipate excess heat by increasing blood flow through the blubber-free skin of the flippers and tail. On hot days, seals on land sometimes hold up a flipper to the breeze to accelerate cooling. A thickness of about 20 cm seems to be the upper limit. Blubber of that thickness apparently does not hold enough lipid to meet the energy storage needs of the very largest whales, fin whales, blue whales, and their relations. They have adipose tissue in the mesentry that supports the gut and around the kidneys and intermuscular adipose tissue in certain muscles, just as do terrestrial mammals.

VI. ENERGY STORAGE IN THE LIFE HISTORY

A. Optimizing Energy Stores

Healthy animals living wild usually have approximately 4–8% dissectible adipose tissue under ideal nutritional conditions but may transiently be much fatter just before some special event such as breeding, migration, or hibernation. Many animals that have less than 4% dissectible adipose tissue are probably having trouble finding enough food but not yet in danger of starvation; for others, however, having much less adipose tissue seems to be normal. Some common wild animals, such as rabbits, hares, moles, and foxes, seem to live and grow satisfactorily with less than 1% of the body mass being adipose tissue.

The concept of costs and benefits to accumulating fat stores has led to much theorizing but little experimentation about the ecological conditions that might define optimum fatness and how they change during the animal’s life history. Some computer models incorporate data about the relationship between body composition and energy utilization and simulate real situations quite accurately, predicting how much adipose tissue is appropriate for birds engaged in certain activities under certain conditions. However, direct measurements of many species in the wild show that there is often much variation in fatness among members of the same species at the same times of year which cannot easily be linked to habits or the ecological conditions in which they live.

Polar bears are top carnivores living in the Arctic, where the climate is not only harsh but also extremely variable from time to time and place to place. These bears are among the very few large carnivores to have huge home ranges but no proper territories. They travel long distances between areas in which ice conditions make seal hunting feasible, eating large but highly irregular meals. Moderate levels of obesity seem to be essential to such habits. A sample of young adult bears studied in the Canadian Arctic in early November, when the sea ice was just beginning to freeze over and seal hunting was again becoming efficient, indicated that the bears comprised between 10 and 21% dissectible fat, which amounted to 73 kg of adipose tissue for the largest bear whose total body mass was more than 400 kg. Every year, a few individuals, adults as well as newly weaned juveniles, are found in a severely emaciated condition, sometimes attributable to an injury but sometimes apparently just following a run of bad luck in catching seals or finding carrion. Their superficial adipose tissue, which over the rump might be more than 10 cm thick in a successful hunter, has shrunk to a flimsy sheet of watery tissue. Their muscles are wasted, so the hips are narrow and the legs thin, and at first glance they look more like a dog than a bear.

Average fatness can be very different in different localities (and probably also from year to year, although there is less information on this topic), even for wild animals of the same species. For example, the common red fox is usually lean where food is available regularly but predators abound, even when, as now often happens, it frequents towns and eats the remains of people’s fatty meals. Populations living in Scandinavia and parts of Canada, in which the food supply is very unpredictable and predators and potential competitors are scarce, are substantially fatter than those elsewhere in Europe. Birds that are fat for only part of the year just before migration may have habits that reduce the risks associated with being heavy and sluggish. They may live and feed in flocks or herds in which the “safety in numbers” principle applies or spend time only in places where predators are few or absent. Small birds such as great tits that feed only by daylight become measurably heavier toward evening and lose weight during the hours of
darkness, especially if the night is cold or prolonged. The amount of fat that they carry from day to day correlates with the likelihood of obtaining food, which depends on the presence of suitable prey, being senior enough in the “pecking” order to get sufficient access to it, and the prevalence of predators. Great tits in Britain became significantly heavier when their principal predator, the sparrowhawk, disappeared in the late 1950s and 1960s as a result of poisoning from agricultural insecticides. When improved farming practices allowed sparrowhawks to re-establish large populations in western Britain in the 1970s, the mean body mass of great tits declined, although that of wrens, a species rarely caught by hawks, remained unchanged.

B. Migration
In contrast to walking, running, and swimming, the muscle power required to stay airborne by active flapping flight increases disproportionately with increasing body mass. Other factors being equal, smaller birds can generate substantially more power than is needed to keep airborne, so they can carry proportionately more fuel, up to 50% of the body mass if necessary. Larger wings make flying energetically more efficient, so the longest nonstop journeys are undertaken by medium-sized birds, such as sandpipers, knots, turnstones, curlews, and godwits, with a body mass of approximately 100–800 g.

Small birds with a body mass of approximately 100 g that set out with adipose tissue triacylglycerols amounting to about 20–80% of the body mass as fuel) can fly for 3 or 4 days nonstop or 3000 to 4000 km (depending greatly on wind direction and other weather conditions). Flying time can be prolonged by slowing down, or it can be shortened by speeding up or by flying into a headwind. In general, larger birds can fly faster, and so arrive at their destination sooner, but because they can carry less fuel they can take off with. If they encounter bad weather or adverse winds, they land on the sea and wait, sometimes for 30 h or more, until traveling conditions improve enabling them to continue their journey.

C. Hibernation
Hibernation and its hot weather equivalent, estivation, are periods of inactivity and seclusion, usually accompanied by cooler than normal body temperature, which enable animals to pass through seasons when food is scarce or inaccessible. Hibernation was among the first physiological states in which adipose tissue of wild mammals, reptiles, and amphibians was studied thoroughly. They depend on their fat reserves while their body temperature is too low to allow the collection and digestion of food, but reclaiming the stores is not entirely straightforward.

The enzymatic processes involved in fasting and starvation are essentially similar to those of slow exercise, but there is a crucial difference. During exercise, the body is warm, often slightly warmer than when sedentary, but in hibernation the body temperature is low, sometimes close to 0°C or 35°C below normal. Enzymes do not work on solidified fats any more than they function in frozen water. Animals must still be able to metabolize the triacylglycerols in their adipose tissue, albeit much more slowly than when they are fully active. The fatty acid composition of triacylglycerols is largely irrelevant to their role as fuel when animals are warm, but it is crucial for their use during hibernation. Experiments on captive chipmunks (Eutamias amoenus) and golden-mantled ground squirrels (Spermophilus lateralis) show that they enter hibernation more readily,
remain cooler for longer, and are better able to survive long winters when plenty of polyunsaturated lipids are included in their diet during the weeks preceding hibernation than when they are fed saturated fats of similar caloric value. Unsaturated fatty acids may lower the melting point of the triacylglycerols and the membrane phospholipids, enabling them to remain more fluid, and hence retain their proper affinity for carrier molecules and enzymes, at cooler temperatures. The chemical composition of the storage lipids, together with other aspects of adipose tissue, is thus adapted to the physical conditions and its role in whole-body metabolism.

Squirrels obtain many such unsaturated fatty acids from the seeds and other plant parts that they eat. Hibernation is an active, physiologically controlled process, and metabolic preparations can be identified days or weeks before the animal actually allows its body to cool. As the weather becomes cooler and the days shorten in autumn, squirrels and other hibernatory rodents actively seek nuts and other foods that contain these lipids. The woodchuck or marmot (Marmota flaviventris) selectively retains linoleic acid (C18:2) before hibernation. The saturated fatty acids are released and oxidized by the muscles, liver, etc. while the animal is warm and active, but the polyunsaturates remain in the adipose tissue for use when the body is cold. Like other mammals, the squirrels cannot add more than one double bond to most kinds of long-chain fatty acids, so they depend on the increased availability in autumn of seeds rich in polyunsaturated lipids. Failure of a seed crop could prevent successful hibernation and thus lead to death from cold or starvation, even if plenty of other foods were available. Recent measurements from biopsies of gonadal and inguinal adipose tissue from alpine foodies of Marmota marmota (Marmota flaviventris) throughout the year indicate that selective release of certain fatty acids allows active regulation of the composition of storage triacylglycerols. The storage lipids remain in an appropriate state of fluidity both during deep hibernation and in the euthermic state by maintaining a high proportion of monounsaturates by retention of these fatty acids from the diet and, if necessary, by synthesis.

Sleeping undisturbed in a cool, secluded place uses storage materials only very slowly. Small tortoises may lose less than 5%, and rarely lose more than 13%, of the body mass during 5 months of hibernation at about 5°C, and much of this loss is water that is lost through evaporation from the lungs. Many reptiles and amphibians emerge from hibernation with a surprisingly large proportion of their lipid reserves still remaining and regain what they lost during the winter in a few weeks of feeding. These stores not only come in handy during spells of cold weather when food is scarce but also are often important for fueling mating and egg production. The less lipid a hibernator uses during the winter, the more it has left to fuel breeding the following spring.

D. Reproduction

The formation of large, yolky eggs that nourish the embryo until it is at an advanced stage of development is a very ancient means of reproduction among vertebrates. Most sharks, skates, and rays, whose ancestors were abundant in Devonian seas 600 million years ago, produce such eggs, which hatch into miniatures of the adults capable of feeding for themselves. One-third of the fresh mass of the yolk of birds’ eggs is storage fats, a proportion that may represent the maximum that can form a stable emulsion, holding the lipids in a state that the embryo or its mother can manage. The relative sizes of the yolk and the “white” part of eggs differ greatly between species, being lowest in species that feed their young in the nest for some time after hatching, such as pelicans (17% of the mass of the egg), gannets (18%), and crows (19%). Yolk content is highest in species whose relatively large eggs hatch to produce large, mature chicks that can walk well and feed themselves. The yellownest eggs are those of kiwis (6%), megapodes (also known as brush turkeys or moundbuilders) (more than 51%), ducks (40%), and of course species of the order Galliformes, which includes turkeys, pheasants, guineafowl, quail, and domestic poultry.

In developing domestic chickens, the cells lining the gut start “eating” droplets of yolk on approximately the 12th day of incubation and pass its lipids into the blood as lipoproteins. Simultaneously, the adipose tissue matures and starts to produce lipoprotein lipase in large quantities, enabling it to take up the yolk lipids. Both the number and the mean size of adipocytes increase until at least the 19th day of incubation and continue to increase after hatching on the 21st day. The adipose tissue serves as more than just a temporary store. As a cellular tissue rather than just a sack of yolk, its adipocytes and the lipoprotein lipase they produce discriminate between triacylglycerols of different fatty acid composition and thereby “ration” the embryo’s irreplaceable lipid provisions to ensure that it develops properly.

Lipids are the main agents of transfer of energy from mother to offspring in most vertebrates and many other kinds of animals, including insects. Mammals are some-what unusual in that their eggs are very small and...
STORAGE, ECOLOGY OF

contain almost no yolk because soon after fertilization they attach themselves to the lining of the uterus, forming a placenta through which the embryo obtains all the nutrients it needs. Lipids cross the placenta and enter the fetus's blood, but most seem to be deployed in structural roles, particularly the formation of membranes of the brain and nervous system, rather than being broken down for energy production. The situation changes abruptly at birth, when the neonate starts to suckle. The milk of most mammals is rich in triglycerides but contains only small quantities of carbohydrates as lactose. During suckling, carbohydrates and lipids almost swap the roles they had in the fetus. Suckling mammals convert lactose into glucose, most of which is incorporated into structural materials, often as “finishing touches” to proteins destined to move between cells or to attach to cell surfaces, and lipids become the main source of fuel.

As soon as the mother gives birth, an array of changes to hormones, their receptors, and enzymes gives the mammary gland preferential access to circulating lipids. The hormones that prompt the maturation of the mammary gland during the final stages of gestation and trigger milk secretion after birth also act on adipose tissue, disabling its uptake of triglycerides from the blood and prompting it to release its fatty acids. These changes direct fatty acids from the adipose tissue to the mammary gland, in which they are incorporated into milk triglycerides. Although storage lipids contain only long-chain fatty acids, the enzymes of the mammary gland readily esterify medium-chain and even short-chain fatty acids into triglycerides. The precursors of milk of adequately fed animals come straight from the gut without passing via the adipose tissue so they may include fatty acids that never appear in storage triglycerides. However, during fasting, the adipose tissue becomes the major source of fatty acids, so the milk triglycerides from starving animals contain more long-chain fatty acids.

The composition of fatty acids in milk triglycerides is thus quite variable in many mammals, depending on what the mother has just eaten and how much is derived from her adipose tissue or obtained directly from the diet. The contribution of storage lipids is important for wild animals that come into contact with lipid-soluble pollutants such as DDT (dichlorodiphenyltrichloroethane), widely used as an insecticide from the 1940s to the 1970s, and PCBs (polychlorinated biphenyls, formerly used as components of large batteries and other electrical apparatus). Such contaminants eaten during lactation may go straight into the milk, and in fasting mothers such as bears, those that have accumulated in adipose tissue during the previous months are released with the fatty acids and thus transferred into the milk. These effects can produce alarmingly high concentrations of toxins in the offspring's tissues, permanently impairing growth and development.

For some mammals, food supplies are too erratic, or obtaining them requires too many long absences from the next, for lactation to be fueled simply by eating more. Milk production has to draw on the body's reserves of lipid from adipose tissue, calcium, and other minerals from the skeleton and protein from it, probably mainly muscle and liver. Therefore, many female mammals fatten during pregnancy and deplete their adipose tissue lipids during lactation.

The diet of most animals that are large as adults changes as they grow. Many lizards and snakes eat insects, slugs, and other small prey when young and, for obvious reasons, defer tackling larger prey until they are large and experienced. However, the composition of the diet of weanling mammals changes more radically and abruptly. The adjustment is just one of the many physiological adaptations of both mother and offspring to this uniquely mammalian means of nurturing the young. The lactation habit has the enormous advantage that it emancipates mammals from the limitation of breeding only where and when suitable food is available for the young and permits rapid growth of the neonate.

Lipid storage in adipose tissue is central to this breeding strategy, enabling mothers to store enough materials to synthesize milk at a high rate and the young to take full advantage of the nutrients their mother provides during the often brief period during which they are together. Its role is enhanced in large species that live in highly variable habitats: Polar bears (Figs. 4 and 5) can carry a large amount of storage tissue because they have no natural predators (other than humans) and hunt by stealth and skill rather than by speed and agility. The bears' massive adipose tissue enables females to separate feeding and mothering to a greater extent than can almost all other warm-blooded animals, thereby enabling them to breed successfully despite the vagaries of food supplies in the Arctic.

Marine mammals can afford to be fatter, at least for short periods, than can terrestrial species because they do not attempt to move far or fast on land and therefore are less constrained by the mechanical limitations of fatness. There are no large terrestrial predators in the Antarctic, so seals breeding on beaches or on sea ice can afford to suckle their young for several weeks. However,
those living in and around the Arctic are obliged to share their habitat with polar bears, and they have evolved ways of transferring lipids from mother to young at a spectacular rate. Harp seals (Phoca groenlandica) that breed on pack ice throughout the Canadian and Russian Arctic are only 3% lipid by weight at birth but reach 47% lipid at weaning 13 days later. The single pup drinks an average of 3.7 kg of milk each day, gaining weight at an average rate of 2.3 kg per day. Transfer of lipid from mother to offspring increases as the young matures. The lipid content of the milk increases from 36% at the start of lactation to 57% just before weaning.

Lactation in the hooded seal (Cystophora cristata), which breeds on temporary ice floes in the North Atlantic, is even more compressed. The single pup weighs about 22 kg at birth, which is large relative to adult size but is only slightly fatter than harp seal pups. It almost doubles its body mass to approximately 43 kg in only 3–5 days of suckling. More than 70% of the increase in body mass is blubber, which holds enough lipid to sustain the pup during a postweaning fast that can last several weeks. During the 4 days of lactation, the mother’s body mass declines from 170 to 150 kg, with more than 80% of the loss from the blubber. She thus synthesizes milk at 2.5–6 times the rate of other seals that have been studied, which is probably a record for any mammal. Such haste may be necessary to minimize the risk that mother and pup become separated by storms and currents that move ice floes around and that of predators.

Most mammals, especially carnivores and other species for which skill and experience are essential for finding food, are weaned with storage tissues replete. Much of the triacylglycerols transferred to suckling seals are stored in their blubber and utilized during the long period between weaning and the time at which the pups become proficient at finding food for themselves.

See Also the Following Articles

CARNIVORES • DIAPAUSE AND DORMANCY • MIGRATION

Bibliography


I. What Is Environmental Stress?

II. Types of Environmental Stress

III. Stress Assessments

IV. Future Trends

GLOSSARY

acute An exposure to an environmental stress that is brief in relation to the temporal scale of the biological system exposed.
chronic An exposure to an environmental stress that is comparable in duration to the temporal scale of the biological system exposed.
ecosystem services The structures and functions of natural biological systems which directly or indirectly support human life.
environmental stress An action, agent, or condition that impairs the structure or function of a biological system.
function The performance of a biological system as a rate.
receptor A biological system that is exposed to the environmental stress.
response A particular structure or function of the receptor that is changed by exposure to the environmental stress.
structure The number, kinds, and arrangement of component parts at one point in time.
threshold The point at which a response begins to be produced.

uncertainty Imperfect knowledge concerning the current or future state of a system under consideration; a component of risk resulting from imperfect knowledge of the degree of hazard or of its spatial and temporal pattern of expression.

ENVIRONMENTAL STRESS is an action, agent, or condition that impairs the structure or function of a biological system. The terms 'disturbance' and 'perturbation' often are used to describe this concept. Examples of environmental stresses on biodiversity include floods, fire, drought, hurricanes, volcanic activity, climate change, land-use changes, introduction of exotic species, and chemical pollution.

I. WHAT IS ENVIRONMENTAL STRESS?

Three components are involved in the relationship that defines environmental stress. First, there is the environmental stress itself as defined previously. However, the environmental stress can only be defined in reference to its interaction with some biological system. Therefore, there must be a receptor—a biological system that is exposed to the environmental stress. Finally, there must be an adverse response—a particular structure or function of the receptor that is changed by exposure to the environmental stress to the detriment of that system. If the survival of that biological system (or another) is
not threatened by the change, then there is no environmental stress.

II. TYPES OF ENVIRONMENTAL STRESS

A. Natural vs Anthropogenic

Environmental stress can be either natural or anthropogenic (i.e., resulting from human actions) in origin. Many environmental stresses, such as most hurricanes, droughts, floods, and fires, are a periodic feature of life on Earth. In contrast, environmental stresses such as the production and release of new chemical compounds and large-scale land-use changes result directly from human action. Ironically, the suppression of natural environmental stresses such as fires and floods can also be a source of stress to biological systems resulting from human actions. Some species have adapted to periodic disturbance and cannot continue without them. For example, some seeds will germinate only after exposure to the high temperatures of a fire. However, such adaptations take time to evolve. Other stresses, such as the meteor that probably wiped out 70% of Earth’s species 65 million years ago, occur too quickly and intensely to result in adaptation.

Natural and anthropogenic stresses often have common components. For example, both hurricanes and wood harvesting result in downed trees. However, as a result of recent major increases in human population, technological capabilities, and standard of living globally, the amount of anthropogenic environmental stress has increased greatly. Anthropogenic environmental stress existed even 50,000 years ago, when fires set to aid hunters are thought to have altered the landscape in central Australia (Flannery, 1999). Since that time, human population increased slowly to 1 billion people in 1804 and then rapidly to 6 billion people in 1999, and the population is expected to increase by another 1 billion people every 12–15 years. The cumulative environmental stresses resulting from these exponential increases in human population are exacerbated by technological changes that expand the character and scope of changes humans can make to their environments. Both the agricultural revolution (about 10,000 years ago) and the industrial revolution (about 200 years ago) expanded the types of anthropogenic stresses on the environment.

In addition, increased affluence for people throughout the world increases the environmental stress on natural systems by increasing the per capita human use of natural resources. Humankind’s collective ecological footprint (i.e., the amount of Earth’s surface required to produce the resources used and to assimilate the wastes produced) is rapidly increasing (Rees, 1996) at the same time that productive land is decreasing through erosion, salinization, and unsustainable land-use changes. The ecological footprint for an average person can range from 0.4 ha required to provide for the lifestyle of one person in India, where the level of affluence is quite modest, to 5.1 ha for a person in the United States. If the lifestyle of every person living in 1996 was elevated to that of a typical North American, an additional two Earths would be needed to provide the surface area required (Rees, 1996).

B. Characterization of Environmental Stress

Environmental stresses of both natural and anthropogenic origin can be characterized on the basis of their spatial distribution, temporal distribution, intensity, and novelty (Kelly and Harwell, 1989). Spatial distribution of a stress describes its geographic extent and pattern. One basic concern is the size of the area affected by the stress, i.e., is the stress local, regional, or global in extent? Some kinds of environmental stress have an intrinsic spatial scale. A poorly dispersed chemical spill may be quite local in its effects, whereas air pollution can affect entire regions. Some stresses will be consistently spread over an area, whereas others will occur irregularly in patches. Cumulative environmental stresses, in which many individual small patches can join together to have larger impacts at a larger scale, have been documented. In one example at the local level, the cumulative loss of small wetland areas within a watershed had demonstrable adverse effects on water quality (Johnston et al., 1988). The magnitude of the biological responses to a stress will often be modified by the spatial distribution of environmental stress, for example, whether key features in the environment such as riverbanks or fencerows are affected or whether similar habitat patches nearby are left unaffected. The field of landscape ecology deals with these factors.

The temporal distribution of a stress describes its frequency and duration. Some stresses, such as chemical spills, are one-time occurrences. Others, such as the winter season or wildfires, can be expected to reoccur either at predictable or at unpredictable intervals. Some stresses rapidly ameliorate; others remain for long periods of time. The terms “acute” and “chronic” are used to describe the duration of a stress. Acute refers to stresses of short duration, whereas chronic refers to stresses that last longer in relationship to the duration of the biological system they affect. In some cases, the
timing of the stress in relationship to other biological events may modify the magnitude of the biological response. For example, the spawning season for amphibians in northern North America in spring corresponds with the greatest thinning of the ozone layer and ultraviolet (UV) light penetrations. The developing eggs of some of these species have been shown to be affected adversely by UV light (Blaustein and Wake, 1995). This temporal factor may be contributing to observed declines in the numbers and kinds of some amphibian species present in those areas.

The intensity of an environmental stress describes its relative ability to evoke a response from the receptor. With increasing intensity, the impact may progress from a few, slightly affected, particularly sensitive components of the biological system to most components being grossly affected. Small changes in the histology or physiological state of individual species can be expected to occur at lower stress intensity and chronologically before changes in survival at a similar temporal and spatial scale. Similarly, at the community level, changes in species composition can be expected to occur at lower stress intensity and chronologically before changes in community functions. The intensity of hurricanes is routinely ranked from category 1 to category 5. In the case of chemical pollution, the intensity of the environmental stress can be described by the concentration of the chemical in the environment. Thus, copper is a natural, background constituent of water in a river or stream. Intake of copper is essential for both animal and plant life. However, concentrations higher than 10 μg/liter in river water can be expected to eliminate a few sensitive species and change the age class distributions in others by changing reproductive success. Higher concentrations can be expected to affect more components in more obvious ways. For example, in Shayler Run, Ohio, 120 μg/liter of copper caused fish kills at some times of the year, avoidance of the stream reach by other fish, declines in macroinvertebrate community species richness, and other gross responses (Geckler et al., 1976).

The novelty of an environmental stress will determine whether or not biological systems will have mechanisms in place to deal with it. Environmental stresses that resemble naturally occurring stresses in their mode of action will be dealt with by the system in the same ways. Novel stresses may be more devastating because no mechanisms have evolved to cope with them. For example, when human harvests of wood products mimic relatively frequent natural events such as treefall or windthrow in their spatial extent, mechanisms are in place in the biological system to recover from this event. When human-created gaps are larger and more intense than historical disturbances, these same mechanisms may not help. Similarly, some natural systems can break down, render biologically unavailable, or disperse low levels of some chemical materials that are naturally occurring or that resemble naturally occurring substances without detectable disruption. The ability of a natural system to receive materials at some concentration, including anthropogenic wastes, without being degraded is its assimilative capacity. However, overloading a system with too much waste destroys both the structure and function of the ecosystem and its future assimilative capacity.

C. Receptors and Responses

An action, agent, or condition can be a stress to one biological system while simultaneously not affecting many others. Because environmental stress is defined by the observation of an impaired biological system, the probability of identifying an environmental stress is related to how thorough the search for impairment has been. Although there are an almost unlimited number of receptors and responses that could be affected by any particular environmental stress, it is generally impractical to monitor more than a small sample. Also, experience has shown that it is easy to overlook a response that may be important. For example, DDT caused eggshell thinning in some birds, although routine toxicity tests failed to identify this response.

The most useful responses to examine for studies that aim to influence decisions about environmental management tend to be those that are clearly related to stated environmental goals. These tend to be responses that are both biologically and socially relevant and that can be measured reliably. Sometimes, the responses that society cares most about cannot be measured directly. Other, presumably related, responses can be measured. In addition, responses that occur earlier in a chain of events, and which lead to an ecologically relevant event, may be useful as early warnings of conditions that have the potential to cause unacceptable damage. Table 1 summarizes some responses that have been used to evaluate environmental stress and guide environmental management.

As the goals in environmental protection have changed over time, so have the responses that are monitored. Most early tests of environmental stress were designed to protect one species—humans. This objective spurred tests that monitored the physiology of species used as human surrogates. Gradually this protection was extended, first to domesticated animals and
TABLE I

Examples of Receptors and Responses Used in Studies of Environmental Stress

<table>
<thead>
<tr>
<th>Level of biological organization</th>
<th>Structural responses</th>
<th>Function responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>Condition</td>
<td>Growth</td>
</tr>
<tr>
<td></td>
<td>Fat stores</td>
<td>Feecundity</td>
</tr>
<tr>
<td>Population</td>
<td>Occurrence</td>
<td>Yield</td>
</tr>
<tr>
<td></td>
<td>Abundance</td>
<td>Gross morbidity</td>
</tr>
<tr>
<td>Community</td>
<td>Species richness</td>
<td>Production</td>
</tr>
<tr>
<td></td>
<td>Trophic structure</td>
<td>Respiration</td>
</tr>
<tr>
<td></td>
<td>Proportion of exotics</td>
<td>Extinction rate</td>
</tr>
<tr>
<td>Ecosystem</td>
<td>Neutrinic pool size</td>
<td>Materials cycling</td>
</tr>
<tr>
<td></td>
<td>Biomass</td>
<td>Materials export</td>
</tr>
<tr>
<td>Landscape</td>
<td>Habitat proportions</td>
<td>Regional production</td>
</tr>
<tr>
<td></td>
<td>Patch size</td>
<td>Materials export</td>
</tr>
<tr>
<td></td>
<td>Perimeter-to-area ratio</td>
<td>Resistance to stress</td>
</tr>
</tbody>
</table>

plants and then to commercially valuable wild species. These additional tests monitored the survival of populations of ‘important’ species. Currently, goals extend beyond the protection of individual species and include the protection of biodiversity and ecosystem services (i.e., those structures and functions of natural biological systems that directly or indirectly support human life). Assessments should reflect these new goals because millions of species in the environment must be protected. Each species cannot be examined individually. In practice, a few species must serve as surrogates for many others, and a few systems serve as surrogates for many others.

Environmental responses may be characterized by type and scale. Responses are either structural (e.g., describing the number and kinds of components, such as the macroinvertebrate community structure) or functional (e.g., describing performance or flux, such as biological oxygen demand or primary production). Also, unique responses may occur at many distinct spatial and temporal scales and levels of biological organization (e.g., cells, tissues, organs, organisms, populations, communities, ecosystems, landscapes, biomes, and the world). Some attributes at higher levels of biological organization are not present at lower levels; for example, energy flow and nutrient spiraling are properties of ecosystems but not of organisms. Other attributes are present in some form at many levels; for example, one can measure the diversity of phenotypes at the population level and the diversity of species at the community level. Environmental goals can be stated on many of these levels, but tests of environmental stress are largely limited to those levels that are more accessible to human observation.

An awareness of scale provides two contrasting approaches to studying environmental stress. Top-down methods start with observed damage to a biological system of interest and investigations move down through hierarchical levels. Component structures and functions are examined in order to diagnose the causative agent and plan remedial actions. At the outset, the damage has already been done, so the relevance of the changes is known. However, the causative agent and the chain of events leading to unacceptable damage are not known. Bottom-up methods start with an environmental stress, and the effects of that stress on biological systems are determined through designed experiments because experiments on small and quick biological systems at lower scales are generally less expensive, these experiments are most common. In bottom-up assessments, the causative agent is known at the outset, but the importance of ultimate changes at any ecologically relevant higher scale is not known.

Microcosms and mesocosms are attempts to increase the spatial and temporal scales and level of complexity in biological systems that can be used in designed experiments of environmental stress. Microcosms and mesocosms simulate important attributes of natural systems in laboratory or outdoor conditions. As the names indicate, the main difference is in size. Microcosms are sometimes small enough to hold in one’s hand; mesocosms may cover 1 acre. Neither is an exact reproduction of any real ecosystem, but they do enable studies of environmental stress in ways that avoid damaging natural systems. On rare occasions, environmental stress may be studied in designed experiments using entire ecosystems. For example, one whole system manipulation was carried out in the Hubbard Brook drainage basin in New York State (Bormann and Likens, 1979). Such efforts are of great value in calibrating models.

D. A General Environmental Stress Syndrome

A threshold is defined in Webster’s Third International Dictionary as “the point at which a physiological or psychological effect begins to be produced.” Moving upwards in biological systems, this effect can be generalized to the point at which a response begins to be produced. Woodwell (1974) asked the question, “Is
it reasonable to assume that thresholds for effects of disturbance exist in natural ecosystems or are all disturbances effective, cumulative, and detrimental to the normal functioning of natural ecosystems? Thresholds may be artifacts of testing procedures, reflecting the power of particular test designs rather than a feature of the system being studied. However, perhaps the more important question is “Can humans detect those environmental changes that are important to their own quality of life?” As is the case with human health, the gradient in environmental systems may be extensive between robust health and collapse in some cases, but an abrupt transition from health to collapse may occur in others. By reviewing information available about the behavior of ecosystems under stress, several researchers have tried to outline general ways in which ecosystems respond to various types of stress (Barrett et al., 1976; Odum, 1985; Rapport et al., 1985; Schindler, 1990). An environmental general stress syndrome at the ecosystem level may include the features listed in Table II; however, experience is continually modifying this list. These efforts to derive an environmental stress syndrome are important because they define a progression of impact in which some minor changes precede other more serious ones. By recognizing changes early in the progression of impact, remediation could begin and crises could be averted. However, the challenge of finding one general description for widely varying systems, challenged by widely varying combinations of stress, is daunting.

Stressed ecosystems often recover once the stress has been removed. However, sometimes human assistance is required, and this process is called ecological restoration or rehabilitation. Restoration has as its goal the return of an ecosystem to a close approximation of its condition prior to the stress and the recreation of a functioning, self-regulating system that is integrated into the ecological landscape in which it occurs. The practice of ecological restoration often involves the reconstruction of physical conditions present prior to the stress, chemical cleanup, and biological manipulation, including revegetation and the reintroduction of native species.

### III. STRESS ASSESSMENTS

Studies of environmental stress can have different purposes. In some studies, the purpose is accounting, i.e., what is the existing condition of this biological system? This question can be important for the purposes of disclosure, national environmental accounting, prioritization, and remediation. Studies of environmental stress can also be used for prediction, i.e., will this action cause a problem or which action is better? Predictive studies are used to register chemicals, rank risks, design processes, etc. Another distinct purpose for studies of environmental stress is to provide early warning of conditions that, if left unchecked, will result in damage significant to human quality of life. By detecting damage before it is of a magnitude that is unacceptable, crises can be averted.

#### A. Appraisal

Studies of environmental stress can assess the condition of biological systems that exist at a particular point in time. When repeated over time, trends in condition can be assessed. Appraising the condition of a biological system can confirm that environmental quality is adequate or can serve to define an existing environmental problem. Many countries are undertaking a national

---

**TABLE II**

Responses Expected in Stressed Ecosystems

<table>
<thead>
<tr>
<th>Category</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Energetics</strong></td>
<td></td>
</tr>
<tr>
<td>Community respiration</td>
<td>Increases</td>
</tr>
<tr>
<td>Maintenance cost</td>
<td>Increases</td>
</tr>
<tr>
<td>Importance of auxiliary</td>
<td>Increases</td>
</tr>
<tr>
<td>Exported or unused</td>
<td>Increases</td>
</tr>
<tr>
<td>Nutrient cycling</td>
<td></td>
</tr>
<tr>
<td>Nutrient turnover</td>
<td></td>
</tr>
<tr>
<td>Horizontal transport</td>
<td></td>
</tr>
<tr>
<td>Vertical cycling of</td>
<td></td>
</tr>
<tr>
<td>nutrients decreases</td>
<td></td>
</tr>
<tr>
<td>Nutrient loss increases</td>
<td></td>
</tr>
<tr>
<td>Community structure</td>
<td></td>
</tr>
<tr>
<td>Proportion of r-strategist</td>
<td>decreases</td>
</tr>
<tr>
<td>Size of organisms</td>
<td>Decreases</td>
</tr>
<tr>
<td>Life spans of organisms</td>
<td>Decreases</td>
</tr>
<tr>
<td>Food chains shorten</td>
<td></td>
</tr>
<tr>
<td>Species diversity</td>
<td>Decreases</td>
</tr>
<tr>
<td>Dominance increases</td>
<td></td>
</tr>
<tr>
<td>Redundance decreases</td>
<td></td>
</tr>
<tr>
<td>General system level</td>
<td></td>
</tr>
<tr>
<td>Ecosystems become</td>
<td></td>
</tr>
<tr>
<td>Autogenic successional</td>
<td></td>
</tr>
<tr>
<td>Efficiency of resources</td>
<td>Decreases</td>
</tr>
<tr>
<td>Parasites increases</td>
<td></td>
</tr>
<tr>
<td>Mortality decreases</td>
<td></td>
</tr>
<tr>
<td>Functional properties</td>
<td>more robust than structural properties</td>
</tr>
</tbody>
</table>

*After Odum (1985).*
accounting of the health of their ecological systems. For example, the Canadian State of the Environment Reports and the Environmental Monitoring and Assessment Program in the United States measure the condition of rivers, lakes, forests, wetlands, arid lands, and agroecosystems. Although some of these programs measure common sources of environmental stress, as well as biological response, others focus solely on response. Once a problem is found, additional studies to diagnose the problem would include measures of environmental stress.

B. Prediction

Often, the easiest way to maintain environmental quality is to prevent damage before it occurs. This strategy requires prediction of the future. Such predictions can come from observations of the effects of similar stress on similar systems or from extrapolations or models from the effects of dissimilar stresses or the effects on dissimilar systems.

The ability to extrapolate the measured effects of environmental stress at one level to consequences at a higher hierarchical level depends on the use of mechanistic models that describe the interaction of component parts. The model is then calibrated, i.e., compared to observed behavior of a system under environmental stress and adjusted to maximize the accuracy of its predictions. Currently, there are a few calibrated models for large-scale predictions. Also, it is unlikely that calibrated models will be available for some processes because their spatial and temporal scales make testing impractical or unethical. All hypotheses and theories are more readily accepted if they have withstood rigorous testing. Predictions of environmental stress are no exception. As a general rule, multiple lines of evidence published in peer-reviewed professional journals whose contents have been reviewed by respected professionals result in acceptance both by the person carrying out research on environmental stress and by mainstream science. Generally, validation occurs in two primary ways: (i) a designed test of a hypothesis derived from a theory, especially by those having nothing to do with its development, and (ii) consensus with other well-accepted and tested theories. Predictions of environmental stress are particularly likely to be challenged outside the scientific community because taking precautionary measures to avoid conditions estimated to cause stress often requires changing societal and industrial practices and sometimes engenders costs. Environmental stress associated with global warming is a good illustrative example. Limiting greenhouse gases in the atmosphere will affect the lives of almost every person on the planet, as would significant global warming. In this case, the entire planet is the experimental unit, making designed tests of impact at the same hierarchical level as that of the environmental problem impossible since there is no “control” planet available. As a consequence, much uncertainty about the probable effect of increased greenhouse gases will persist. However, management decisions about such gases must be made and must be based on the best available information; managers must act even though there is uncertainty accompanying any estimate of risk. Fortunately, most forms of environmental stress, such as exposure to potential toxicants, are much more easily validated.

C. Early Warning

Monitoring is a systematic and orderly gathering of data to ensure that previously established quality control conditions are being met. Biomonitoring applies this activity to the detection of environmental stress. In this case, the goal is to provide an early warning that unacceptable levels of environmental stress have occurred. As is the case in an intensive care room in a hospital when heart and respiration rates are monitored and unacceptable conditions are detected, immediate action is mandatory. In any form of quality control, the more rapidly the information becomes available, the more quickly corrective action can be taken. Extensive use of information technology has made complete automation of some monitoring systems, including triggering the remedial action, possible. However, in the absence of carefully selected goals and objectives related to the decisions to be made, they can also generate huge amounts of unnecessary and inappropriate data.

Setting the corrective action threshold low to ensure early detection of deleterious change seems prudent, but it can also produce false-positive readings. A false positive is an indication that some deleterious effect has occurred when in fact none has occurred. Emergency response team response to eliminate stress can be quite expensive and unpopular with management. False positives are usually most numerous when monitoring is in the early developmental stages and decrease substantially as experience is gained. Avoiding false positives by setting the action threshold well beyond the response threshold will probably result in false negatives. A false negative is information that no deleterious effects have occurred when in fact some have occurred. Thus, ecosystem damage occurs because no corrective action alert is produced. These false signals are clearly a matter of prime importance in the design of all monitoring
IV. FUTURE TRENDS

In view of unprecedented growth in human population and the desire to raise the standard of living above the subsistence level for most of the world's people, there are new challenges for those studying environmental stress. The level of environmental stress from food production and energy usage may increase. Simultaneously, there is great pressure to increase food production to feed the increasing human population, and the need to protect intact biological systems and ensure their robust functioning so that they can continue to provide necessary ecosystem services will become more pressing. Some approaches to these problems are of great interest.

The World Commission on Environment and Development (1987) of the United States published Our Common Future, which focuses attention on the future condition of the planet, arguably more so than any publication that preceded it. The commission defined sustainable development as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs." This presents a curious combination of words because "sustain" means to continue and "development" is usually associated with growth. However, infinite growth on a finite planet is clearly not possible.

In December 1989, the General Assembly of the United Nations (UN) attempted to address the problems identified in Our Common Future (World Commission on Environment and Development, 1987) by organizing the UN Conference on Environment and Development (popularly known as the Earth Summit), which was held in Rio de Janeiro in June 1992. This conference resulted in a heightened awareness of the interrelatedness of environmental degradation and stress, population development, and depletion of natural resources. Previously, all had been viewed as separate problems, but now attempts are being made to address them as an interactive system. The resulting Rio Declaration on Environment and Development endorsed the following principles: (i) Nations should not cause damage to the environment of other states and areas beyond their borders; (ii) eradicating poverty and reducing disparities in worldwide standards of living are indispensable requirements for sustainable development; (iii) the pollutant, in principle, should pay the cost of pollution; (iv) states should discourage or prevent transboundary movements of activities and substances that endanger health or environment; and (v) scientific uncertainty should not be a reason for postponing urgent measures to prevent environmental degradation.

Sustainable use of the planet will not be possible if excessive environmental stress impairs the ecological life support system that provides necessary ecosystem services. As a consequence, keeping environmental stress at tolerable levels has a direct bearing on the quality of life for humans. Events beyond human control, such as a large extraterrestrial object striking Earth, make it impossible to guarantee that reducing environmental stress will ensure sustainability.

The theory of "weak sustainability" asserts that human society is sustainable provided that the aggregate stock of manufactured and natural assets is not decreasing. Thus, the loss of the whaling industry would not impair sustainability if the proceeds of liquidation are invested in industries of comparable income-producing potential. Pearce and Atkinson (1993) dispute the assumption that natural and human-made capital are sustainable in this context. They assert that strong sustainability requires that natural capital stocks be held constant, independently of human-made capital. A weak sustainability scenario would permit considerably more environmental stress than a strong sustainability scenario. Conventional monetary analyses are biased against strong sustainability. For example, at a discount rate of 5%, the current value of ecological services for an American life span (about 76 years) from the present on is approximately 2.5¢. Using this approach, the farther into the future one projects, the less valuable natural systems appear, thus diminishing the significance of environmental stress.

Reducing environmental stress requires the cooperation of all of society. One of the most important components is the industrial system. Fortunately, the field of industrial ecology (IE) is developing worldwide (Hawken, 1993). The goal of IE is to reduce environmental stress at all stages: (i) extraction of raw materials, (ii) processing, (iii) disposal of manufacturing wastes, (iv) packaging, and (v) reincorporation into the environment at the end of the product's life in a nonstressful way, ideally in a way that enhances ecological integrity. Industrial ecology's primary goals are to (i) reuse materials as much as possible, (ii) reduce energy consumption per unit produced, and (iii) design both processes and products so that they can be reincorporated into the environment with minimal stress. Books such as Engineering within Ecological Constraints (Schulze,
1996), which was produced by the National Academy of Engineering, are directed toward achieving this goal.

See Also the Following Articles

DISTURBANCE, MECHANISMS OF, ECOLOGICAL FOOTPRINT, CONCEPT OF, Ecosystem Services, CONCEPT OF, Human Effects of Ecosystems, OVERVIEW, STABILITY, CONCEPT OF, SUSTAINABILITY, CONCEPT AND PRACTICE OF

Bibliography

I. A Brief History of Subspecific Taxonomy

A. Variation below the Level of Species

Since the invention of binomial nomenclature by Linnaeus, there has been a conflict between “splitters,” who named more or less well-defined local populations as separate species, and “lumpers,” who ignored geographic variation and united local variants into a single species. The problem was compounded by early systematists’ belief that species had an Aristotelian “essence,” each fundamentally different from similar essences underlying other species. To Linnaeus’ followers, it seemed important to decide which level of variation was fundamental. The terms “genus” and “species” both result from Aristotelian philosophy, and although Linnaeus is usually credited with establishing the species as the basal taxonomic unit, he confused matters, after recognizing that some plant species were of hybrid origin, by suggesting that genera were a more important taxonomic level (i.e., a separately created kind) than species.

Once evolution was accepted, it became clear that variation at all levels in the taxonomic hierarchy was due to more or less similar causes, the only difference between variation above the level of genus or species and below was one of degree. Darwin realized that species could evolve from intraspecific varieties. Darwin (1874) used the term species in a new and nonessentia-list sense:

The complete absence, in a well-investigated region, of varieties linking together two closely-allied forms, is probably the most important of all the criterions of their specific distinctness. . . . Geographical distribution is often brought into play unconsciously and sometimes consciously; so that forms living in two widely separated areas, in which most of the other inhabitants are specifically distinct, are themselves usually looked at as distinct; but in truth this affords no aid in distinguishing geographical races from so-called good or true species.
Darwin showed convincingly that there was no essential difference between species and “varieties”; species were simply varieties which had diverged more. However, with his term varieties, Darwin did not clearly distinguish between polymorphic variants within populations and the identifiable geographic populations that are today normally considered geographic “races” or “subspecies.” To Darwin, the distinction was unimportant because polymorphic variants, clinal variation, geographic races or subspecies, and “good” species formed a continuum. Darwin demonstrated that this continuum was excellent evidence for an evolutionary origin of the taxa we call species.

**B. The Trinomial Revolution**

Many systematists wished to preserve the purity of the simple genus–species binomial nomenclature, but by the 1850s there were enormous stresses. It began to be realized that many clearly identifiable geographic replacement forms were an important intermediate stage between insignificant local variants and good species. Some lumped these replacement forms as varieties within species, whereas others continued describing these geographic forms as separate species. Practices varied widely, leading to considerable confusion. Although some Europeans had long advocated naming geographic races or subspecies, the accumulation of major North American museum collections during the great push of colonization and railway construction westwards was probably the most important catalyst of a revolutionary new systematics. In this new approach, nomenclature consisted of a trinomial, genus–species–subspecies, which still is the dominant taxonomic practice today. The maxim was “intergradation [at the boundary between two geographic replacement forms] is the touchstone of trinomialism.” Examples from commonly observed birds which intergrade are, in North America, the eastern rufous-sided towhee (*Pipilo erythrophthalmus erythrophthalmus*) replaced in the west by the spotted towhee (*Pipilo erythrophthalmus maculatus*), and in Europe the carrion crow (*Corvus corone corone*) found in the south and west replaced by the hooded crow (*Corvus corone cornix*) in the north and east. Among ornithologists responsible for this revolution in North America were Elliott Coues, who published a catalog of American birds in 1872 incorporating an early version of this trinomial nomenclature in which subspecies were prefixed by “var.,” and Robert Ridgway, who dropped the “var.” in his 1881 summary of North American bird nomenclature. The American Ornithologists’ Union soon adopted this policy, and the idea then spread to Europe, particularly England, where Walter Rothschild began amassing his vast collection of birds and butterflies and hired excellent and productive staff (the ornithologist Ernst Hartert and the entomologist Karl Jordan) to curate and describe the new material. Jordan was particularly important in spreading the idea of trinomial nomenclature to entomologists, and he was regarded by the Rothschilds as the “clever” member of the staff. He published important papers on the theory of systematics, justifying trinomial nomenclature and the recognition of the “subspecies” as a valid, identifiable taxon in its own right. Both Jordan and Hartert were Germans who contributed to and read German as well as English journals, and in Germany a similar revolution was taking place. Thus, these systematic ideas were able to spread to the rest of Europe during a time when science was often highly parochial. The standard trinomial nomenclature for subspecies soon became established in the International Code for Zoological Nomenclature and has remained there ever since.

**C. Theories of Divergence**

It is difficult to imagine the diversity of ideas by which the systematists of 100 years ago explained geographic variation. At that time, evolution by natural selection was far from generally accepted; in fact, many believed it had been disproved. One of the most influential ornithologists of the time was Otto Kleinschmidt, who believed that all species suddenly came into being long ago and thereafter remained completely separate. Replacement forms or subspecies developed via natural selection from the main species but, in Kleinschmidt’s view, subspecies could never evolve into new species as the Darwinians supposed. To distinguish his new species concept from the older one in which geographical replacements might be named as separate species, Kleinschmidt called his theory of variation the Formenkreis (ring of forms) theory. The Formenkreis theory fitted neatly with, and indeed promoted, the new practices of naming subspecies and trinomial nomenclature.

In those times, there were many somewhat peculiar competing explanations for geographic variation and speciation, including Kleinschmidt’s non-speciation theory, J. P. Laut's hybridization theory, mutationism, inheritance of acquired characters, and natural selection. In Britain, Jordan and Rothschild argued eloquently and influentially against any new terminology (including Formenkreis) that had theoretical implications, and they proposed incorporating as little evolu-
tionary theory into taxonomy as possible in view of the lack of agreement among scientists at the time. Although Rothschild and Jordan (1899, 1903), supported by Hartert, agreed both with the nomenclatural practice of naming subspecies and that subspecies were valid real taxa, they argued that the Linnaean term species should be retained for the whole group of races and that the geographic forms were not true species—they were simply subspecies or incipient species.

However, others believed that the term species was too emotive to be used in the new, multiple-subspecies sense. Some scientists continued following the Formenkreis doctrine and had begun to name quite distinct taxa, which did not intergrade at their boundaries, as subspecies. This situation led in the 1920s and 1930s to the neo-Darwinian ornithologist Bernhard Rensch scrapping the term Formenkreis because of its theoretical limitations and instead substituting two new terms, Rassenkreis (circle of races) and Artenkreis (circle of species). Rassenkreisen were again considered to be equivalent to species, composed of races or subspecies. However, now there was an additional layer in the taxonomy composed of groups of Rassenkreise that replaced another geographically—the Artenkreise. Thus, an Artenkreis could consist of multiple Rassenkreise. Rensch and many others believed that the subspecies was an incipient species, of which the geographic replacement species Artenkreise were a further development, until finally divergence was sufficient to allow complete geographic overlap, whereupon new Rassenkreise could again form.

These terms did not catch on, and most people came to the conclusion that the Rassenkreise were equivalent to the species referred to by Linnaeus and Darwin. Probably a major reason that we do not use these multiple taxonomic terms is due to the prolific work published in English by another German, Ernst Mayr. Mayr had worked for Walter Rothschild and knew Hartert. After Walter Rothschild was blackmailed by a lover, his enormous bird collection of 280,000 skins was sold in 1932 to the American Museum of Natural History, where Mayr was hired as curator. Mayr's experience of ornithology, contact with the European literature, and friendship with the geneticist Theodosius Dobzhansky (who helped to convince him of the lack of evidence for inheritance of acquired characters) resulted in a unique opportunity to influence the course of systematics and evolutionary biology. Mayr did not waste this opportunity. Mayr used the ideas underlying Rensch's new terms but renamed them in English. The Rassenkreise became simply the true species or "polytypic" species, with its geographic races being subspecies, whereas the Artenkreise became the "superspecies" and its component parts "semispecies," i.e., not very divergent true species. Mayr successfully blended the local species concept of Poulton and Dobzhansky based on interbreeding with the geographic Rassenkreis idea of species and renamed this combination of ideas "the biological species concept," a term which has remained strongly associated with Mayr's name. His many influential articles and books promoted a new program of species study, a science of the species still with us today. Central to Mayr's system was the belief that discrete taxa such as species or subspecies would normally diverge in "allopatry," i.e., in complete geographic isolation.

II. THE SUBSPECIES TODAY

A. Modern Views of Subspecies, Semispecies, and Superspecies

The view of Darwin, Wallace, Rensch, and Mayr that geographic replacement forms, subspecies and semispecies, which form a continuum with species, were in fact incipient species has few critics today. Most geographic replacement species or semispecies which do not intergrade when they meet must indeed have evolved from previously interbreeding subspecies. Modern genetic data have done nothing to cast doubt on this idea.

However, taxonomists were now required to describe subspecies, which has never been viewed as a particularly honorable or worthwhile activity in comparison with describing species, especially recently. A strong attack on the subspecies was mounted by Wilson and Brown (1953). Both were systematists working on ants, a group particularly riddled with poorly conceived trimorals at the time. Wilson and Brown argued that subspecies rarely if ever could be justified on the basis of multiple characters, and that therefore they were not "real taxa." The only real taxa were species, which in a sense were self-defining because interbreeding prevented divergent genes from flowing from one species to another. Subspecies which interbred at their boundaries, on the other hand, were not so endowed; therefore, genes and morphological characters could flow between them. Good examples were put forward of subspecies which undoubtedly would be difficult to justify on multiple-character grounds. This single article was enormously influential on systematics in the United States, and generations of systematists trained at Harvard and Cornell (where Wilson and Brown
worked), their own many intellectual descendants, and their students’ students in turn have eschewed the practice of naming subspecies.

Through genetic studies we now know, however, that many subspecies separated by hybrid zones differ at multiple morphological, behavioral, and genetic characters (Barton and Hewitt, 1985). For instance, the toad *Bombina bombina* meets its relative *Bombina variegata* across a broad front in Europe and differs strongly in call, morphology, skin thickness, the sizes of water bodies used, and egg size as well as in mitochondrial DNA and protein sequence. Their levels of differentiation suggest that the *Bombina* have evolved separately for many millions of years. (The two forms hybridize freely in the contact zone—although the hybrids can be shown to suffer some inviability—and so should be classified as members of the same polytypic species under the polytypic or biological species concept, but it has always seemed natural to place such well-defined forms in separate species despite the fact they have not truly “speciated.”) This situation of multiple character changes has been shown to be true across many examples of hybrid zones, and gene flow can be shown to be almost completely blocked by hybrid zones such as these, despite abundant hybridization. Thus, although many named subspecies undoubtedly merited Wilson and Brown’s (1953) scorn, genetic evidence shows that there are plenty of local replacement forms which hybridize at their boundaries but which do form “real” identifiable taxa and are valid subspecies under the Wilson and Brown criteria.

B. Subspecies, Species, and Conservation

Opposition among modern taxonomists to the subspecies can be traced as an influence on the recent “diagnostic” version of the polygenetic species concept. The adherents of this view of species, led by the ornithologist Cracraft (1980), proposed a radical species concept so that even a single fixed character difference may define a geographic form as a separate species; multiple-character justification is not considered necessary by them even at the species level. The practical result of this new concept is that many local forms are again being recognized as species. In birds and butterflies, which often have many morphologically or genetically distinct subspecies, this could easily result in a 2- to 10-fold increase in the number of species, or even more in some groups.

It is probable that the revision of geographic forms upwards to the level of species is being driven not only by theoretical considerations but also by existing legislation, which proposes that “endangered species” are the valuable units to be conserved. If an area contains a taxon recognized as a species rather than just a local race, it may be viewed as more valuable for conservation purposes. The potential consequences for biodiversity and conservation of the continued instability of the term species are detailed elsewhere. Here, I only mention that today’s conservationists are reducing emphasis on species conservation and are becoming increasingly aware of biodiversity at all the levels of the hierarchy of life, including well-marked subspecies. Thus, the legislative need for differentiating local races as species may ultimately become less of an impetus provided that future legislation falls more into line with prevailing biological thought.

III. FURTHER READING

Much of the historical overview in this article is discussed in the excellent reviews of Stresemann (1936, 1975), Mayr (1982), Rothschild (1983), and other sources cited previously.

**See Also the Following Articles**

**NOMENCLATURE, SYSTEMS OF • SPECIATION, THEORIES OF • SPECIES, CONCEPTS OF • SYSTEMATIC, OVERVIEW • TAXONOMY, METHODS OF**

**Bibliography**


I. The Subterranean Domain
II. Sources of Energy to Subsurface Environments
III. Adaptations to Subterranean Life
IV. Colonization and Speciation in Subterranean Environments
V. Taxonomic Survey of Subterranean Life
VI. Some Representative Subterranean Communities
VII. Geography of Cave Biodiversity
VIII. Conservation and Protection of Subterranean Habitats

GLOSSARY

amphibites Species that require both surface waters and subterranean waters in order to complete their life cycle.
anchialine (or anchihaline) Haline water, usually with restricted exposure to open air, and always with subterranean connections to the sea.
chemolithotrophy The process, utilized by some microbes, of obtaining energy from the breaking of inorganic chemical bonds.
epikarst Upper part of the percolation zone in karst.
hyporheic zone Interstitial space within the sediments of a streambed; a transition zone between surface water and permanent (phreatic) groundwater.
interstitial The spaces between sediment particles, especially in alluvial deposits.
karst A landscape in which the primary geomorphic agent is solution rather than erosion; typically formed in carbonates; landforms include caves, sinkholes, blind valleys, and large springs.
MSS The interstitial spaces deep in the soil and mantle/bedrock interface, as found in glacially fragmented zones. By origin, the milieu souterrain superficiel, or mesovoid shallow stratum, thus the acronym MSS.
stygobite Aquatic species that are obligate subterranean dwellers.
stygophile Aquatic species that can live and reproduce in both subterranean and surface environments.
troglbite Terrestrial species that are obligate subterranean dwellers; sometimes used for aquatic species as well.
trogloomorphic Pertaining to the morphological, behavioral, and physiological characters that are convergent in subterranean species.
troglophile Terrestrial species that can live and reproduce in both subterranean and surface environments; sometimes used for aquatic species as well.

SUBTERRANEAN ECOSYSTEMS include a wide variety of aphotic, resource-poor sites, including caves, aquifers, and the underflow of rivers. This article discusses the kinds of ecosystems, the unique organisms that inhabit them, the evolution and adaptation of these organisms, and the energy flows through them.
I. THE SUBTERRANEAN DOMAIN

A. Caves

The subterranean domain, literally the domain below the earth, includes both air- and water-filled underground habitats. The most celebrated of these are caves. Most caves are formed by the dissolution of rock by acidic waters, but caves can also result from mechanical action, especially the flow of lava. Nearly all of the caves formed by the action of acidic waters are in carbonates (particularly limestone) and evaporites (particularly gypsum). Landscapes in which the primary agent molding the landscape is dissolution rather than erosion are called karst landscapes. Comprising approximately 1% of the earth’s surface, karst represents 75% of the land area of Cuba, 40% of Slovenia, 23% of France and Italy, and 40% of the United States east of Tulsa, Oklahoma. Caves are present in the oldest Cambrian rocks to Holocene limestone less than 10,000 years old. Lava tube caves are common on volcanic islands such as Hawaii and in the northwestern United States. Although caves have a reputation as both exotic and rare habitats, they are actually quite common. More than 100,000 caves are known from Europe, and nearly 50,000 are known from the United States. Caves come in many sizes and varieties, ranging from a few meters long to Mammoth Cave in Kentucky with a surveyed length of more than 500 km. All caves share two important characteristics: complete darkness away from the entrance and reduced environmental variability relative to surface conditions.

Many organisms spend some or all of their life cycle in caves, particularly cave entrances. The moist rock of cave entrances provides habitat for many mosses and ferns, such as the Hart’s tongue fern (Phyllitis scolopendrium). The entrance and twilight zone of a cave are refuges from temperature extremes of the surface. Some species, such as the spider Meta menardi, are specialized for the surface–subsurface ecotone at cave entrances. The entrance and twilight zones of caves are relatively predator free, at least for vertebrates. The neotropic oil bird Steatornis caripensis nests in caves; many other birds do so as well, albeit on a less regular basis. Many cave entrances in the eastern United States have one or more phoebe (Sayornis phoebe) nests. The best known visitors to caves are bats. Depending on the species, bats use caves as maternity colonies, as hibernacula, and as temporary roosts during the warmer months of the year. More than half of the species of bats found in the eastern United States use caves all or part of the year. About 95% of the more than 1 million gray bats, Myotis grisescens, hibernate in eight caves. Many bat species are at risk because of the vulnerability of their hibernating and maternity sites to disturbance.

Many species spend their entire life cycle in caves. In the case of terrestrial species, troglobites have an obligate dependence on caves and must complete their life cycle in caves. Troglophiles can complete their life cycle in caves, but they can also complete their life cycle in surface habitats. The equivalent terms for aquatic species are stygobites and stygophiles. Another useful term is troglomorphic, which denotes the morphological syndrome associated with life in caves—loss of eyes and pigment and the elaboration of extraoptic sensory structures. Most but not all troglobites are troglomorphic. Those that are not have likely not been isolated in caves for a long enough time for troglomorphic characteristics to evolve.

The remarkable cave amphibian, Proteus anguinus, was the first stygobite to be mentioned in scientific writing. Writing in 1689, Valvasor reported that local residents thought that this amphibian, common in caves in Slovenia, was the larvae of dragons. Its scientific description was accomplished by Laurenti in 1768. The first troglobite, the cave beetle Leptodirus hochenwartii (Fig. 1), was described by Schmidt from the same region in 1832. Currently, more than 4000 troglobites and 2000 stygobites have been formally named, and at least several times that number probably exist.
B. Other Shallow Subsurface Habitats

Terrestrial subsurface habitats also occur outside of caves. These include cracks and small voids in volcanic rocks and a superficial zone of rock fissures and debris slopes in schists, gneiss, and granite. Generally occurring at a depth of few meters, this habitat was called milieu souterrain superficiel (MSS) in French, or meso-void shallow substratum in English. The MSS is generally found in mountains in temperate zones but apparently not in the tropics, where voids are usually filled with laterites and clay. Several dozen troglobites have been found in the MSS. For the most part, species in the MSS have been sampled by baits placed several meters deep in holes drilled for the purpose. In karst, there is a similar habitat in the upper part of the rocks in the upper part of the percolation zone with numerous fissures and solution pockets—the epikarst. Either water or air filled, it also contains troglobites and stygobites. Although in many cases species found in the epikarst are also found in caves, there are species predominantly found in epikarst. Although the epikarst has occasionally been sampled directly by means of shallow wells and holes, most of the collections of the fauna of epikarst are from caves. Species collected in dripping water, in temporary pools, and in other habitats near the surface are denizens of the epikarst. Caves provide a window to this fauna.

A very different kind of subterranean aquatic habitat also contains an interesting fauna—the interstitial habitat. These habitats are aphotic and widespread. They include sands of marine and freshwater beaches—the psammmon. The fauna in these habitats lives in the narrow voids between sand grains and belongs to the meio-fauna, a size class of aquatic organisms too small to be considered macroinvertebrates and too large to be microorganisms. In addition to the absence of light, low oxygen levels are frequently encountered. The unconsolidated sediments of running water are also an important interstitial habitat. Because alluvial sediments are transported and deposited by currents, they have larger grain sizes and spaces than do sand beaches of lakes and oceans. Consequently, the fauna is larger and in many cases overlaps with the stygobitic fauna of caves. The hyporheic is the interstitial space within gravel deposits beneath streams channels. It is an ecotone between surface waters and permanent (phreatic) groundwaters. The hyporheic usually ranges from several centimeters beneath the channel to several meters away from the channel. The habitat is very complex, with areas of upwellings and downwellings and low oxygen and high oxygen often occurring within meters of each other.

Beneath the hyporheic and to the sides, the phreatic zone also contains an interesting fauna. In addition to stygobites and stygophiles, alluvial aquifers also have amphibites, species that require both surface waters and subterranean waters in order to complete their life cycle. Early sampling of alluvial aquifers was done by digging a hole near a stream and filtering the water that entered the hole through a plankton net. During the past 30 years, this technique has been supplemented by a small pump (Bou-Rouch Pump) attached to a pipe driven 1 or 2 m into the sediments. In addition to sampling in wells, sampling of the interstitial fauna of alluvial sediments has been done for only a few rivers, most of which are in Europe.

C. Deep Subsurface Habitats

Wells and artesian springs in karst areas often connect with water-filled underground voids tens to hundreds of meters deep that are part of the permanent water table—the phreatic zone. Few such sites have been sampled because of the difficulty in biological sampling in deep wells. However, the few sites that have been sampled have yielded a diverse array of macroscopic invertebrates. The artesian spring at San Marcos, Texas, which is connected to the Edwards Aquifer, has a diverse array of invertebrates, many of which are endemic, as well as several stygobitic fish and salamanders. Perhaps the most exotic subsurface habitat is that of deep groundwater in porous aquifers. The water within both sedimentary and igneous rock contains an array of Eubacteria and Archaea. The presence of bacteria in the deep subsurface was first suggested by Edwin S. Bastin in the 1920s. He suggested that the curious presence of hydrogen sulfide and bicarbonate in water extracted from oil was the result of the activity of sulfate-reducing bacteria at depth. It was nearly 60 years before the technology was developed to sample bacteria at depth. To do so required the solution of several formidable technical problems. The first was to obtain samples uncontaminated by surface microbes, and the second was to characterize the microbes. Characterization of the samples is difficult because of the unusual environmental conditions in which the microbes live, making standard culturing techniques of limited value. Much of our knowledge of deep groundwater microbes is derived from sampling wells near former nuclear weapons production sites in Washington State and in South Carolina. Bacteria and Archaea have been found at depths of more than 2 km. Their presence in sedimentary rock may be the result of the slow infiltration of water from the surface or they may have been present at the time...
of the deposition of the rock. The possibility that they were present at the time of deposition is plausible given the depth of the water and its age based on stable isotope ratios. Microbes present in igneous rock must have been deposited after it formed since it was too hot at the time of formation. In these habitats, chemolithotrophs, which are forms that use inorganic carbon as a carbon source and obtain energy from the oxidation of reduced inorganic chemicals, predominate. The possible great age of some of these microorganisms combined with their ability to survive in the absence of organic matter and sunlight makes it likely that they are more representative of some of the earliest forms of life on the planet than are surface-dwelling microorganisms. They may also be useful analogs for understanding what life on other planets, such as Mars, might be like. More than 9000 strains of organisms have been cataloged from subsurface habitats.

II. SOURCES OF ENERGY TO SUBSURFACE ENVIRONMENTS

All subsurface environments share an absence of light and hence an absence of photosynthesis. With the exception of a few caves and possibly most deep-ground-water habitats, there is no significant primary production. With these exceptions, all subsurface food webs rely on the import of surface organic matter in one form or another. For the most part, subsurface habitats are energy poor. The famous Romanian cave biologist Emil Racovitza stated that many cave organisms are “carnivores by predilection but saprophages by necessity” (translation by author). Also, the famous American cave biologist A. S. Packard noted more than a century ago that “cave animals, even the carnivorous species, take remarkably little food.” However, for most subsurface habitats, organic matter is brought in by the action of water. Dissolved organic matter (DOM) and fine particulate organic matter (FPOM) are important energy sources both in interstitial habitats and in many caves. In caves, water percolates through the soil into the cave, where it collects in pools or streams. In many caves, sinking streams (swallets) bring in additional organic matter—not only DOM and FPOM but also coarse particulate organic matter, including leaves and twigs. In a sense, many cave streams are not very different from a low-productivity stream on the surface. In many surface streams, allochthonous input greatly exceeds autochthonous production. The organic matter brought in by streams is also an important food source for terrestrial cave animals. The stranding of organic matter by flooding of streams is the primary source of food for many terrestrial cave communities.

For many subsurface terrestrial communities, the main source of energy is the result of the movement of animals in and out of the cave. In many North American caves, cave crickets in the subfamily Rhaphidophorinae periodically leave caves to forage and deposit their eggs in sandy substrates in the cave. These eggs are in turn the major part of the diet of some troglobitic carabid beetles. The fecal material of the cave crickets and the occasional dead cave cricket body are an important food supply for other troglobites. Fecal matter left by other cave visitors is also important in many caves. The most important source of fecal material is bats. In most temperate zone caves, only a few bats use the cave, and their guano is the basis for a group of species whose food base is transitory organic matter. In those temperate zone caves that serve as maternity sites or hibernacula and in many tropical caves, large bat colonies produce enough guano for there to be a community specialized on the guano. Most of these species are guano specialists rather than cave specialists.

There are other allochthonous sources of food. In a few caves, such as the lava tubes on the Canary Islands and some MSS. habitats, the major source of organic matter is wind-carried particles and debris, which penetrate into the subterranean habitats through cracks and fissures. In lava tubes and some shallow limestone caves, roots penetrating through the ceiling are an important source of fecal material. In some temperate zone caves, only a few bats use the cave, and their guano is the basis for a group of species whose food base is transitory organic matter. In those temperate zone caves that serve as maternity sites or hibernacula and in many tropical caves, large bat colonies produce enough guano for there to be a community specialized on the guano. Most of these species are guano specialists rather than cave specialists.
thermore, this autotrophic production is the basis not only for the aquatic food web but also for the terrestrial food web—the only documented case of a chemoautotrophic terrestrial ecosystem. Several other sulfidrich caves, including Frasassi Caves in Italy, Villa Luz Cave in Mexico, and Cesspool Cave in Virginia, also exhibit many features of chemosynthetic production. In Villa Luz Cave and Frassati Caves, a rich but not stygobitic and nontroglobitic community occurs. The explanation for the rarey of troglobites and stygobites lies in their closer connection to surface waters, the penetration of light into many parts of the cave (Villa Luz and Cesspool Caves), and the relatively young age of the caves (Frasassi and Cesspool Caves).

It is likely that many phreatic deep systems are chemoautotrophic (actually chemolithoautotrophic). A variety of chemical reactions are used by Bacteria and Archaea to provide the energy for biosynthesis. Among the reactions occurring in deep groundwater is methanogenesis:

\[ \text{CH}_4 + \text{CO}_2 \rightarrow \text{CH}_4 + 2 \text{H}_2 \text{O} \Delta G = -32.5 \text{ kcal} \]

The methanogenic reaction uses hydrogen gas as the energy source and inorganic CO2 as the carbon source. Hydrogen derives from the reaction of the oxygen-poor water with iron-bearing minerals. Methanogens can persist indefinitely without any supply of carbon from the surface.

III. ADAPTATIONS TO SUBTERRANEAN LIFE

Compared to related surface-dwelling species, subterranean animals are characterized by the absence or reduction of a series of features, particularly eyes and pigment. Subterranean animals are also characterized by an elaboration of extraoptic sensory structures. This pattern of elaboration and reduction is of course extremely general when comparing two groups of organisms from different habitats. What makes subterranean animals particularly interesting in this regard are the obvious and unusual reductions, often called regressive evolution. Unlike some aspects of evolution, such as the evolution of sex and the origin of evolutionary novelty, eyelessness and regressive evolution in general have always appeared easy to explain. To an orthogenetist such as Albert Vandel writing in the mid-twentieth century, eye loss was a classic example of the role of disuse. To a neo-Darwinian, eye loss in subterranean animals was the result of an evolutionary tradeoff resulting from natural selection for elaborated nonoptic sensory structures. To neutral mutationists, eye loss in cave animals was easily explained by the accumulation of selectively neutral, structurally reducing mutations made possible by the lack of stabilizing selection. In a modern context, theories that account for regressive evolution in subterranean cave animals are classified in two groups: those that involve natural selection directly or indirectly and those that involve neutral mutation and genetic drift.

One of the best studied organisms in this regard is the amphipod Gammarus minus which occurs in springs and caves in much of the central Appalachians in West Virginia and Virginia. Populations independently isolated in caves show a strongly convergent morphology of appendage elongation and eye and pigment reduction. There was a strong correlation between the fitness components of mating and egg number with relative appendage length and eye size. Furthermore, this correlation was opposite for cave and spring populations. However, natural selection is not the whole explanation for eye and pigment reduction in G. minus. The rate of change of reduced structures is consistently greater by an order of magnitude than the rate of change of elaborated structures (Table I). Elaborated structures changed as a result of both natural selection and the accumulation of structurally reducing, selectively neutral mutations. The increased variability of eyes in recently isolated cave populations is also most easily explained as a result of neutral mutation. The genetics of eye loss are best understood in the Mexican cave fish Astyanax fasciatus. Like G. minus, it has apparently been isolated in caves for less than 1 million years. Defects in developmental control genes are primarily responsible for eye reduction. Few if any of the sequences of opsin proteins and a structural lens protein α-A-crystallin are altered in cave populations. The master control gene of eye development Pax-6 was not altered in cave populations. The defect causing eye degeneration is assumed to be in the regulation cascade responsible for eye development below the Pax-6 and above the α-A-crystallin gene.

Among subterranean organisms, adaptations have been most thoroughly studied in cave animals. After eye and pigment reduction, the most characteristic trog-
### TABLE I

Evolutionary Rates of Divergence for Different Morphological Characters of a Cave Population of the Amphipod *Gammarus minus* from Its Ancestral Spring Population

<table>
<thead>
<tr>
<th>Morphological character</th>
<th>∆</th>
<th>Relative rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (HL)</td>
<td>4.0 x 10⁻²</td>
<td>0.02</td>
</tr>
<tr>
<td>Eyes (EN)</td>
<td>2.4 x 10⁻²</td>
<td>1.00</td>
</tr>
<tr>
<td>EA</td>
<td>1.5 x 10⁻²</td>
<td>0.82</td>
</tr>
<tr>
<td>Antennae P1</td>
<td>8.2 x 10⁻³</td>
<td>0.03</td>
</tr>
<tr>
<td>L1</td>
<td>4.1 x 10⁻²</td>
<td>0.02</td>
</tr>
<tr>
<td>N1</td>
<td>1.1 x 10⁻²</td>
<td>0.05</td>
</tr>
<tr>
<td>P2</td>
<td>1.1 x 10⁻²</td>
<td>0.06</td>
</tr>
<tr>
<td>L2</td>
<td>8.9 x 10⁻³</td>
<td>0.04</td>
</tr>
<tr>
<td>N2</td>
<td>1.4 x 10⁻³</td>
<td>0.06</td>
</tr>
</tbody>
</table>

* The measure of divergence, , is independent of the size of the character and depends only on its variability. The two populations compared are Organ Cave and Organ Cave spring, about 2 km apart in southern West Virginia. Time since divergence was set to 5 x 10⁵ years. Traits were head length (HL), eye ommatidia number (EN), surface area of compound eye (EA), peduncle length of first antenna (P1), flagellum length of first antenna (L1), and number of flagellar segments of first antenna (N1) and the second antenna analog (P2, L2, and N2). In the last column, all rates are compared to that of ommatidia number, which was set to 1. [Adapted from Culver et al. (1995) by permission of the publishers. © President and Fellows of Harvard College.]

Table 1: Evolutionary Rates of Divergence for Different Morphological Characters of a Cave Population of the Amphipod *Gammarus minus* from Its Ancestral Spring Population

- **Size (HL):** The measure of divergence, , is independent of the size of the character and depends only on its variability. The two populations compared are Organ Cave and Organ Cave spring, about 2 km apart in southern West Virginia. Time since divergence was set to 5 x 10⁵ years.
- **Eyes (EN):** Eye ommatidia number.
- **Surface Area of Compound Eye (EA):**
- **Antennae P1:** Peduncle length of first antenna.
- **L1:** Flagellum length of first antenna.
- **N1:** Number of flagellar segments of first antenna.
- **P2, L2, N2:** Peduncle length of second antenna analogs.

In the last column, all rates are compared to that of ommatidia number, which was set to 1. [Adapted from Culver et al. (1995) by permission of the publishers. © President and Fellows of Harvard College.]

### FIGURE 2

![Brain morphologies of four fishes in the family Amblyopsidae:](image)

- (A) *Chologaster cornutus,* a surface-dwelling species.
- (B) *Chologaster agassizi,* a stygophile found in both springs and caves.
- (C) *Amblyopsis rosae,* a stygobite.
- (D) *Speoplatyrhinus poulsoni,* a stygobite.

O.L., olfactory lobe; T, telencephalon; O.L., optic lobe; C, cerebellum; E.G., eminentia granulare; C.C., cristae cerebelli; M, modiolus oblongus; O, otoliths. Note the relative hypertrophy of extraoptic parts in stygobites. [Adapted from Poulson (1963) with permission.]

Figure 2: Brain morphologies of four fishes in the family Amblyopsidae: (A) *Chologaster cornutus,* a surface-dwelling species; (B) *Chologaster agassizi,* a stygophile found in both springs and caves; (C) *Amblyopsis rosae,* a stygobite; and (D) *Speoplatyrhinus poulsoni,* a stygobite. O.L., olfactory lobe; T, telencephalon; O.L., optic lobe; C, cerebellum; E.G., eminentia granulare; C.C., cristae cerebelli; M, modiolus oblongus; O, otoliths. Note the relative hypertrophy of extraoptic parts in stygobites. [Adapted from Poulson (1963) with permission.]

Lomorphomorphisms are the elaboration or elongation of extraoptic sensory structures. Among arthropods, the elongation of antennae is especially noticeable. A striking example is the first described troglobite, the Slovenian cave beetle *L. hochenwarti* (Fig. 1). Vertebrates such as the North American cave fish in the Amblyopsidae display a striking hypertrophy of the lateral line system. These changes are also reflected in the central nervous system. Areas associated with vision decrease in size and areas associated with the other senses increase (Fig. 2). The potential advantage of elaboration and increase in extraoptic sensory structures is to compensate for the absence of light by increasing tactile and chemical sensory sensitivity. In the resource-poor environment of caves, metabolic efficiency is often increased, particularly among species at or near the top of the food chain. Activity is often reduced, as is metabolic rate. The frequently observed decline in intraspecific aggressive behavior of troglobites and stygobites may also in part be an evolutionary response to food scarcity. Food scarcity is also apparently a strong selective force in life history patterns of cave animals. Cave animals live longer, begin reproduction at a later age, produce fewer offspring, and have larger eggs—all adaptations to a relatively constant, food-poor environment. Perhaps the most striking example of life history modifications is that of a population of the crayfish *Orconectes australis* in Shelta Cave, Alabama. In an extensive mark-recapture study lasting more than 6 years, John E. Cooper found that the age at first reproduction was approximately 35 years, and that the crayfish lived to an age of more than 100 years. Animals in other subterranean habitats also show many of the same adaptations since they share in common with cave animals an aphotic, resource-poor environment. However, there are two other environmental...

---

*Note:* The table and figure are placeholders as the text does not provide specific details about the data. The table and figure are included to demonstrate the format in which the text is presented. The table contains evolutionary rates of divergence for different morphological characters, and the figure illustrates brain morphologies of four fishes in the family Amblyopsidae. The text discusses the evolution of extraoptic sensory structures in subterranean ecosystems, highlighting examples of how these adaptations occur, including the elongation of antennae and the hypertrophy of the lateral line system in cave fish. The text also references an extensive mark-recapture study in Shelta Cave, Alabama, which provides evidence of long lifespans and delayed reproduction in crayfish populations. The text concludes with a discussion of other environmental modifications observed in subterranean habitats.
factors that exert strong selective pressures not present in cave environments: intragranular space size in interstitial habitats and oxygen levels in aquatic interstitial habitats. Interstitial animals tend to be both thin and small since they are constrained by interstice diameter. The morphological contrast between cave and interstitial animals can be quite striking. A terrestrial example comparing a cave with an epikarstic and deep-soil-dwelling Collembola is shown in Fig. 3. The small size of many aquatic interstitial species may occur as a result of progenesis, which is the precocious sexual maturation of an organism resulting in an adult descendant exhibiting the juvenile morphology of its ancestor. This process also results in reduced extremities, spines, and bristles (Fig. 4). More than cave dwellers, interstitial stygobites must be able to survive under conditions of low oxygen. Many interstitial crustaceans are able to regulate respiration rate, varying the ventilating activity of their pleopods according to oxygen availability. Metabolic adaptations of deep-groundwater microorganisms are just beginning to be understood. Unlike the situation with eukaryotic organisms, the basic tenet of microbial ecology, Beijerinck’s principle, holds in the deep subsurface: “Everything is everywhere, the environment selects.” The environment in this case is an extremely food-poor one, with little or no organic matter. The major characteristic of phreatic microorganisms is the widespread occurrence of facultative or obligate chemoautotrophy. The Gibbs free energy available in chemoautotrophic responses is much less than that of photosynthetic organisms, reinforcing the concept of resource scarcity from a physical chemical standpoint.

IV. COLONIZATION AND SPECIATION IN SUBTERRANEAN ENVIRONMENTS

The terms troglobite, stygobite, troglophile, and stygophile are an ecological classification which attaches great importance to the isolation of species in caves. Troglophiles and stygophiles, species which can complete their life cycle in both surface and subterranean habitats, are often important components of cave communities. In many cases, especially in the tropics, they outnumber troglobites and stygobites. Troglobites and stygobites are typically but by no means always troglomorphic. The evolution of troglobitic (stygobitic) species that are troglomorphic (stygobitic) has received the most attention.

There are two components to evolution in subterranean habitats: colonization and subsequent isolation of the subterranean populations usually as a result of the extinction of the surface-dwelling populations. Colonization may be active or it may be passive, i.e., “colonization under constraint.” An example of active colonization is the movement of surface-dwelling Homoptera in lava fields in Hawaii and elsewhere into lava tubes and cracks in the lava in which root sap is an abundant resource. In some cases, speciation has proceeded without the extinction of the surface populations. The mechanism of speciation is either parapatric or sympatric. Active colonization may be particularly important for interstitial organisms. The classic case of colonization under constraint is that of many troglobites in north temperate caves. Many of the surface-dwelling ancestors of these troglobites were leaf-litter-dwelling inverte-
brutes. During Pleistocene interglacials, as warming occurred, many of these species were either forced into caves or up mountain slopes to cooler conditions. As warming continued, surface populations became extinct. For the beetle genus Psamaphagus, various speciation events can be associated with particular interglacials (Fig. 3). For most cases of colonization under constraint, the factor that forced animals into caves, such as Pleistocene interglacials or the Messinian salinity crisis in the Mediterranean, was also the factor that resulted in the extinction of surface populations.

By whatever mechanisms organisms colonize subterranean environments and by whatever geological or climatological event surface populations become isolated in subterranean environments, subsequent dispersal is generally extremely limited. Most subterranean species have very restricted ranges. For example, more than 60% of the subterranean species described from the United States are known from a single county, and about half of these are known from a single locality. Species with large ranges, at least for subterranean species, have on closer examination been shown to be complex of morphologically identical but genetically distinct species. For example, the European cave salamander Proteus anguinus ranges more than 100 km from Italy to Croatia but actually represents several invasions of a common surface-dwelling ancestor that is extinct. For some species, there is little gene flow at distances of less than 1 km. This makes subterranean species excellent markers of biogeographic events. Once species are isolated in subsurface habitats, which are often a refuge from the vicissitudes of surface climatic change such as temperature and salinity, their distribution may leave a record of ancient events, especially continental drift. Examples abound in the Crustacea. The distributions of the order Syncarida and the amphipod family Fidgiellidae are Pangean and predate the pre-Triassic breakup of Pangea. The distributions of the amphipod family Crangonyctidae and the isopod family Asellidae are Laurasian; their distributions of the amphipod family Crangonyctidae and predate the pre-Triassic breakup of Pangaea. The distributions of the order Syncarida include H2, N/H4 acceptor among subsurface autotrophs. Electron donors also considerable diversity in the electron donor and cycle found in Archaea and a few bacteria. There is a wide diversity of metabolic pathways among autotrophs, most important in the carbon fixing pathways: the Calvin cycle in bacteria, the acetyl-coenzyme A pathway in Archaea, and the reductive citric acid cycle found in Archaea and a few bacteria. There is also considerable diversity in the electron donor and acceptor among subsurface autotrophs. Electron donors include H2, N/H4, S, and Fe2+; acceptors include O2, SO4²⁻, and NO3⁻. Except for a few protozoans, the Eukarya is usually not found in deep groundwater. A wide taxonomic array of organisms are stygobitic or troglobitic, including the following phyla: Porifera, Hydrozoa, Platyhelminthes, Nemertina, Nematomida, Mollusca, Rotifera, Annelida, Onychophora, Arthropoda, and Chordata. However, for many of these groups, there are only a handful of subterranean species. Furthermore, many of these ‘oddball’ species are found only in the Dinaric karst of Bosnia and Herzegovina, Croatia, and Slovenia.

Among stygobites, the Crustacea predominate. At least half of the described species are Crustacea. Among the Crustacea, the orders Remipedia, Spelaeographacea, and Thermosaenacea are exclusively stygobitic. Numerically, crustacean stygobites are dominated by Crustacea: Among the Crustacea, the orders Remipedia, Spelaeographacea, and Thermosaenacea are exclusively stygobitic. Numerically, crustacean stygobites are dominated by Co- pepoda (especially in the interstitial) and Amphipoda (especially in caves). Some Amphipoda genera are particularly specious. For example, nearly 175 stygobitic species of the amphipod genus Stygobromus are known from North America. Other important groups include the planarians (Platyhelminthes: Tricladida), snails (Mollusca: Gastropoda), salamanders (Chordata: Cau-
Phylogenetic hypothesis of evolution in the Ptomaphagus hirtus species group of beetles in the southeastern United States. Numbers refer to characters, transforming from the ancestral state (H17040) to the derived state (H17039). State abbreviations with species names indicate distribution. Lineage splits and character origin are shown for the latest likely times for the events (adapted with permission from Peck, 1984).
data), and fish (Chordata: Pisces). The surface-dwelling ancestors of these species were largely benthic, and many species were adapted to environments with dim light and a detrital food base.

Among troglobites, the Arachnida, Diplopoda, and Insecta predominate: At least 90% of the species are in these groups. Among the Arachnida, spiders (Araneae), false scorpions (Pseudoscorpionida), and harvestmen (Opiliones) are common in temperate caves and in the tropics. Other groups, such as Schizomida, Ricinulei, and Amblypygi, are also important. Among the Insecta, springtails (Collembola) and bristletails (Diplura) are common, but the beetles (Coleoptera) predominate. More than 2000 troglobitic beetles are known. One genus in the United States (Pseudanophthalmus) accounts for nearly 250 species. The surface-dwelling ancestors of many of these species occurred in leaf litter in forests and are believed to have colonized caves during climatic vicissitudes of the Pleistocene.

Although bats are not troglobites since they spend part of their life cycle outside of caves foraging for food, their dependence on caves is every bit as profound. Among the nearly 1100 described species of bats, at least one-third utilize caves as day roosts, hibernacula, or maternity sites. For many species, especially those in the Vespertilionidae and Molossidae, there is an obligate dependence on caves. Some species form giant hibernating colonies. More than 90% of Myotis griseus found east of the Mississippi hibernate in three caves in which numbers may reach more than 1 million. The Mexican free-tailed bat, *Tadarida brasiliensis*, forms large summer maternity colonies, with up to 20 million individuals in a single cave. This species declined in numbers due to the widespread use of DDT.

### VI. SOME REPRESENTATIVE SUBTERRANEAN COMMUNITIES

#### A. Mammoth Cave, Kentucky

The longest cave in the world with more than 500 km of passage, Mammoth Cave harbors 28 troglobites and 15 stygobites. Most of the cave lies within the boundaries of Mammoth Cave National Park. The long length of the cave is due in part to a sandstone caprock that reduces erosion and loss of upper level passages. In the large upper level passages with sandy-bottomed floors (an unusual cave habitat), trogloxenic cave crickets (*Hadenoecus subterraneus*) deposit their eggs. Since *H. subterraneus* regularly leave the cave to feed, their eggs are an important energy source. Predaceous beetles, *Neaphaenops tellkampfi*, specialize on predation on cricket eggs. Such specialization is unusual because of the scarcity of resources in most caves. Also noteworthy is the co-occurrence of five species of trechine beetles in the closely related genera *Neaphaenops* and *Pseudanophthalmus*. The aquatic community is notable for its large population of the fish *Typhlichthys subterraneus* and the crayfish *Orconectes tellkampfi*.

#### B. Postojna–Planina Cave System, Slovenia

Consisting of 17 and 6 km of passages, respectively, connected by 2 km of flooded corridors, the Postojna–Planina Cave system has more known species of stygobites and troglobites than any other cave or other subterranean site. The sinking river in the main passages is inhabited by a rich assortment of stygobites, stygophiles, and accidental surface species. Hydrologically inactive parts of the system contain other aquatic and terrestrial habitats. Among the 34 troglobites is the first described troglobite—the beetle *L. hochenwartii* (Fig. 1). Beetles and Collembola dominate, with 5 species of the Collembolan genus *Onychurus* co-occurring in the cave. Among the 48 stygobites is the European cave salamander *Proteus anguinus*. Both the snail (8 species) and crustacean (16 species) faunas are rich. The marine
origin of some of the stygobitic species is evident in the hydrozoan Velkovrhia enigmatica. The Postojna–Planina Cave is one of the best studied caves in the world, and parts of it have been heavily visited by tourists since 1818.

C. Gua Salukkan Kallang–Towakkalak, Indonesia
This immense river cave system, with more than 20 km of passage and a large bat population, is unusual among lowland tropical caves in the richness of its troglobitic and stygobitic fauna. Food resources are abundant in the cave. There are large amounts of flood debris along the riverbanks, and bat and swiftlet guano is scattered along the underground galleries. In addition to troglobites and stygobites, there are many species specialized on guano as well. Most of the 7 stygobitic species are crustaceans, and arachnids and Collembola dominate the 21 troglobitic species.

D. Walsingham Caves, Bermuda
This complex of anchialine caves (ones with a direct connection to the sea) is approximately 1 km long, with most of the passages submerged. With a freshwater lens and a redox boundary between fresh and saltwater, Walsingham Cave likely has considerable productivity as the result of chemautotrophic and heterotrophic production. The fauna is entirely aquatic, with 37 species, 29 of which are crustaceans. Among the crustaceans is a representative of the exclusively subterranean order Mictacea and rich ostracod and copepod faunas with 6 species each.

E. Movile Cave, Romania
Movile is a small, mostly water-filled cave near the coast of the Black Sea with no natural connection to the surface. Extensive chemosynthetic production in the form of sulfur oxidation occurs in the cave system. Nearly 50 stygobites and troglobites are known from the cave, and two-thirds of these are endemic to the cave or groundwater system that the cave intersects. Among the unusual components of the fauna is an endemic leech and an endemic water scorpion (Nepa anophthalma). Based on analysis of mitochondrial DNA sequences, N. anophthalma was isolated in the cave 2 million years ago. The high diversity is made possible by the high productivity, and the high endemism is a result of the lack of contact with the surface.

F. Edwards Aquifer, Texas
San Marcos Spring is an artesian spring that serves as an exit for the Edwards Aquifer, a large cavernous limestone reservoir that lies along the Balcones escarpment. The aquifer occupies an area of 10,000 km². Two-thirds of the 27 stygobites are crustaceans, and endemism is high as well. The subterranean amphipod fauna in this system may be the most taxonomically diverse of its kind in the world and includes 10 species of gammaridean amphipods, representing eight genera in four families. The vertebrate stygobitic fauna is rich with two catfishes and the Texas blind salamander Typhlomolge rathbuni. Oil and peat deposits above the aquifer as well as long-term and gradual input of organic material into the aquifer through numerous sinkholes on the adjacent Edwards Plateau are the primary inputs into the system. The aquifer is threatened by excessive draw down, both by agricultural interests and by the city of San Antonio.

G. Lobau Wetlands, Austria
Formed by a meander arm, the Lobau wetlands are part of the floodplain of the Danube River near Vienna. Extensive studies of the superficial gravel sediments, which are up to 20 m thick, using pumps and minivideo cameras in shallow wells revealed a complex habitat with areas of differing oxygen and permeability and a rich fauna. In a 1-km² area more than 100 species were found, one-third of which were stygobites. Stygobites were more common at depths of greater than 5 m and 100 m away from a surface water. The microcrustaceans—Copepoda and Ostracoda—were particularly diverse.

H. Rhone River Floodplain, France
After the Danube, the Rhone River is the most thoroughly studied river with respect to its interstitial fauna. Thirty-eight stygobitic species, dominated by 23 crustaceans and 10 oligochaetes, have been found in the interstitial waters of the Rhone. There are considerable faunal differences depending on the distance from the main channel of the river, depth, and upwelling and downwelling.

I. Flathead River Alluvial Aquifers, Montana
A tributary of the Columbia River, extensive aquifers occur in the Nyack and Kalispell Valleys. Utilizing an ex-
tensive network of shallow sampling wells, investigators discovered more than 200 species of invertebrates in the aquifer. Of these, 8 were stygobites and 6 were amphibiotic stenophiles. The amphibites, species whose immature forms only occur in interstitial water, included 4 species in the genus Isocapnia. Stygobites were concentrated far from the river, whereas other interstitial species were near to or directly below the river.

VII. GEOGRAPHY OF CAVE BIODIVERSITY

Although other subsurface habitats have not been sufficiently studied to analyze biogeographical patterns, caves have been. On a worldwide basis, more than 10,000 caves have been biologically investigated. Except for in the United States, there is no thorough analysis of regional patterns. However, there is an extensive amount of information about the number of stygobites and troglobites at individual caves and karst wells. Because of the restricted possibilities for dispersal and the restricted abilities of stygobites and troglobites to disperse, species numbers in any one cave or karst well are low relative to the regional fauna. On a worldwide basis, only 18 caves and two wells are known to have 20 or more stygobites and troglobites (Table II). Several patterns emerge. First, there is a remarkable concentration of sites (six) in the Dinaric karst of Slovenia and Bosnia and Herzegovina. Diversity of stygobites in this region is especially high, and this reflects the opportunities for both colonization from marine embayments of the ancient Mediterranean Sea and isolation resulting from the periodic regression of the Mediterranean. Second, sites with high productivity, especially chemosynthesis, compared to other subsurface sites are well represented; these include Movile Cave, Bayliss Cave, and Walsingham Caves. Gua Salukkang Kallang Towk-kalak and the Postojna–Planina Cave system can also be considered as caves with high (secondary) productivity. This accords with the widely held view that caves are usually resource poor and food limited. Third, caves and wells that intersect the permanent saturated zone are also well represented. These include Movie Cave, Shelta Cave, Edwards Aquifer, all five sites in France, and the

<table>
<thead>
<tr>
<th>Locality</th>
<th>Stygobites</th>
<th>Troglobites</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postojna–Planina system, Slovenia</td>
<td>48</td>
<td>36</td>
<td>84</td>
</tr>
<tr>
<td>Vjetrenica Jama, Bosnia and Herzegovina</td>
<td>39</td>
<td>21</td>
<td>60</td>
</tr>
<tr>
<td>Ponorica Novo, Romania</td>
<td>18</td>
<td>20</td>
<td>48</td>
</tr>
<tr>
<td>Kozina Jama, Slovenia</td>
<td>29</td>
<td>16</td>
<td>45</td>
</tr>
<tr>
<td>Logarkec Jama, Slovenia</td>
<td>28</td>
<td>15</td>
<td>43</td>
</tr>
<tr>
<td>Mammoth Cave, Kentucky</td>
<td>15</td>
<td>26</td>
<td>41</td>
</tr>
<tr>
<td>Walsingham Caves, Bermuda</td>
<td>37</td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td>Sea–Stika system, Slovenia</td>
<td>27</td>
<td>7</td>
<td>34</td>
</tr>
<tr>
<td>Toundon Wells, France</td>
<td>34</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>Bagar/Lacotri–Pyrene–Sainte Catherine system, France</td>
<td>14</td>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td>Salukkang Kallang–Towakkalak, Indonesia</td>
<td>7</td>
<td>21</td>
<td>28</td>
</tr>
<tr>
<td>Edwards Aquifer, Texas</td>
<td>27</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>Gourel di Her–Nerthe Trombe–Henne Morte, France</td>
<td>14</td>
<td>17</td>
<td>31</td>
</tr>
<tr>
<td>Bayliss Cave, North Queensland, Australia</td>
<td>0</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Shelta Cave, Alabama</td>
<td>12</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>Ceri–Jons, France</td>
<td>22</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Grotta dell’Arena, Italy</td>
<td>6</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Buso della Rana, Italy</td>
<td>15</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>Resurgence de Saint, France</td>
<td>20</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Grad (Ouspoka Jama), Slovenia</td>
<td>17</td>
<td>3</td>
<td>20</td>
</tr>
</tbody>
</table>

*Data assembled with assistance of Borut Sket, Ljubljana, Slovenia.*
and all six caves in the Dinaric karst. Finally, many of the caves are long caves. Seven of the caves listed in Table II (Mammoth, Buso della Rana, Goueil di Her, Krizna, Postojna–Planina, Vjetrenica Cave, and Gua Sulukkang Kallang–Towakkalak) are more than 5 km long, but less than 0.1% of all caves are this long. A longer cave usually means a greater number of habitats.

More complete information on cave biodiversity patterns for the 48 contiguous states of the United States is available. Nearly 45,000 caves are known and their distribution by county is shown in Fig. 7. More than one-third of the counties (1144) have at least 1 cave. The major karst regions of the United States are apparent—the Appalachians (Pennsylvania to Alabama), the Interior Low Plateau (Kentucky, Tennessee, and Alabama) and associated areas immediately to the west, the Florida lime sinks, the Ozarks (Missouri, Arkansas, and Oklahoma), the Driftless Area (Illinois, Iowa, and Wisconsin), the Edwards Plateau (Texas), the Guadalupe Mountains (New Mexico), the Black Hills (South Dakota), and Mother Lode karst (California). Figure 8 shows the distribution of stygobites and troglobites by county. Some of the same regions are obvious: the Appalachians, the Interior Low Plateau, the Edwards Plateau. Cave areas to the north and west are not well represented by stygobites and troglobites. Diversity decreases to the north primarily as a result of covering of caves by the Pleistocene ice sheets, reducing the time available for colonization. Diversity decreases to the west primarily as a result of the fragmented nature of karst, the lack of suitable colonists due to the absence of forest litter, and progressive aridity in late Tertiary and Recent times.

VIII. CONSERVATION AND PROTECTION OF SUBTERRANEAN HABITATS

Subterranean sites present special conservation and protection problems. Literally out of sight, they are often neglected in decisions concerning conservation priorities. When they are considered, the high levels of endemism make it difficult to set priorities among endemic species. Once the decision has been made to protect a cave or other subsurface site, several principles apply. The protection of bat sites is most effectively done by gating the cave. However, there is considerable risk to the bats if the gate is not properly designed. Improperly
designed gates can change environmental conditions in the
cave, making it unsuitable for bats, and if the bars on
the gate are not properly spaced bats can suffer consider-
able predation at the gate. Second, protection of the sub-
terranean aquatic fauna requires protection of the sur-
face riparian habitat. In the case of cave streams, their
upstream sources, whether they are sinking streams or
sinkholes, must be protected. In the case of epikarstic
species, it is the surface immediately above the cave that
must be protected. Similar protection schemes are
needed for the terrestrial riparian fauna. Third, protec-
tion of deep-groundwater species requires protection
from both excessive draw down of the aquifer and con-
tamination of the aquifer. Fourth, protection of troglo-
bites that rely on the flow of organic matter into the cave
entrance, such as those Neaphaenops beetles that are
predators of cricket eggs, must include protection of the
entrance area and any foraging area of the species respon-
sible for bringing organic matter into the cave. Finally,
conservation and protection of subsurface sites requires
greater public awareness and appreciation.

See Also the Following Articles
SOIL BIOTA, SOIL SYSTEMS, AND PROCESSES •
THERMOPHILES, ORIGIN OF • VENTS

Bibliography
Netherlands.
Wiley, New York.
Culver, D. C., Kane, T. C., and Fong, D. W. (1993). Adaptation and
Natural Selection in Caves. The Evolution of Gammarus minus.
Harvard Univ. Press, Cambridge, MA.
Culver, D. C., Hobbs III, H. H., Christman, M. C. and Master, L. L.
(1999). Distribution map of caves and cave animals in the United
States. J. Cave Karst Stud. 61, 139–140.
II. Societe de Biospeologie, Moulis, France.
I. Societe de Biospeologie, Moulis, France.
Univ. Press, Baltimore.
Ponina phalangide brelens in the southeastern United States (Coleoptera:
62, 710–740.
Midland Nat. 70, 257–290.
I. Introduction
II. Organic Explanations of Ecological Succession
III. Mechanistic Explanations of Ecological Succession
IV. Succession and Biodiversity

GLOSSARY

autotrophic successions These generate energy from internal processes (photosynthesis).

climax community According to some theories of succession, the end result of succession in which successional change ends with a community that does not change and which is in equilibrium with the climate.
disturbance Major alternations of vegetation due to events such as wildfires, hurricanes, landslides, and human clearing.
heterotrophic successions Dependent on already fixed energy, such as the successional of communities associated with decomposition of dead logs.
individualistic view of succession Concept that succession is a consequence of species interacting with one another and their environment.
primary succession Succession on newly exposed substrates such as a sandbar or rubble at the foot of a receding glacier.
progressive succession Successions in which the dynamic changes are in the directions of increasing species diversity, structural complexity, greater biomass, and increased stability. Retrogressive successions are in the opposite directions.
secondary succession Succession on existing substrate (soil) following a disturbance.
succession The pattern of change expected in a community over time after a disturbance or after new substrate has been exposed.

ECOLOGICAL SUCCESSION is an ordered progression of structural and compositional changes in ecosystems toward an eventual stable condition. Descriptions of succession involve the nature of the changes and the factors that cause the changes. Ecologists have debated whether the succession is a community process or the summation of the consequences of individual species and their interactions with each other and the environment. Most ecologists doing research in this area currently favor the latter view. Investigations of the types of interactions among species have led to an increased interest in the mosaic nature of vegetation and to the application of computer models to project the expected patterns of change in vegetation over time. The biodiversity of landscapes may be highest when there is an intermediate level of disturbance. The appearance of species at different times in succession appears to be idiosyncratic to the particular vegetation; the rate of
loss of species from communities seems to decrease logarithmically.

I. INTRODUCTION

Ecological succession is an ordered progression of structural and compositional changes in ecosystems toward an eventual stable condition. Primary successions are initiated on new substrates, such as a new volcanic island, a new sandbar in a river, or rubble fields at the foot of a receding glacier. Secondary successions involve the recovery of vegetation on established soils from land abandonment and from disturbances such as wildfires, hurricanes, or human alterations to vegetation. Other significant dichotomies used to categorize succession involve whether a given successional sequence is

1. Progressive (dynamically changing in the directions of increasing species diversity, structural complexity, greater biomass, and increased stability) or regressive (in the opposite directions)

2. Autotrophic (generating energy from internal processes) or heterotrophic (dependent on already fixed energy, such as the successional of communities associated with decomposition of dead logs)

3. Autogenic (changing due to interactions from inside the system) or allogenic (changing in response to changes in external variables)

Ecological succession is an early concept in ecology and was essential in the early definitions of ecological communities. Although the basic concepts of succession are easily understood, debates about succession (Box 1) have spawned considerable confusion and discussion. The mechanisms that drive ecological succession and its very existence as a natural phenomenon have been the subject of continual debate among ecologists. McIntosh (1985) identifies two contrasting views that typify traditional natural history and also can be found in current discussions of theoretical ecology. This dichotomy, which can be used to organize theories about ecological succession, contrasts mechanistic explanations and organic (holistic) explanations of the causes of succession. For mechanistic theories about succession, known laws explain the actions of the individual parts of a system and the whole system is the sum of these parts and their interactions. In the case of organic or holistic explanations, the whole system, its existence and design, explains the actions of the parts. Mechanistic explanations often are taken as the more modern interpretation of succession, but these "modern" views may have been the first developed and certainly developed as early as the organic views. Organic explanations of succession were most popular, at least in the United States, between the 1920s and early 1950s and have had a considerable impact on land use and conservation policies.

II. ORGANIC EXPLANATIONS OF ECOLOGICAL SUCCESSION

Organic explanations seek to understand succession in terms of principles that operate at the level of the whole system. In the 1920s and 1930s, F. E. Clements and particularly John Phillips were ascribing to ecological communities the attributes of a superorganism—a highly organized and coevolved assemblage of plants and animals interacting in a dynamic system. The ecosystem concept had its roots in debates regarding the organization and dynamics of natural systems. It was Tansley's negative view of Clements' and Phillips' interpretation of the community as a superorganism that in 1935 inspired his development of the ecosystem concept as an alternative to the organic term 'community.'

A. Clementsian Concepts of Ecological Succession

Between 1905 and 1935, F. E. Clements promoted a dynamic plant ecology built around a "supraorganismic" view of the ecological community. At the time of their conception, Clements' ideas represented a significant emphasis on the dynamics of vegetation and were an important early attempt to develop a formal theory predicting the pattern and expected change in ecological communities. The underlying conviction was that evolution and internal interactions would produce a homogeneous regional "climax" vegetation or community of regular species composition. The development of this "climatic climax" community was ecological succession, which was viewed as the community analog of the embryological interactions that produce an organism. In this view, succession was typified by a progressive sequence of serial stages or seres, communities that sequentially replaced one another over time until the climax stage was reached. Successional development was the result of a set of processes:

1. Nudation—the creation of a bare area (or partially bare area) to initiate succession
Box 1

Complex Debates Arising in a Simple Definition

Is the pathway of successional change ordered because the mechanisms that cause succession need to occur in a proper sequence or can the steps in the order be skipped in some cases? Is there more than one ordered progression? How much variability can be tolerated before the apparent order in the progression becomes recognized as disordered or chaotic?

Ecological succession is an ordered progression of structural and compositional changes in ecosystems toward an eventual stable condition.

In 1935, A.G. Tansley coined the term “ecosystem” to replace “community” because he felt the trend of thinking of the community as a direct analogue to a living organism — with succession an equivalent to embryological development — was scientifically and conceptually unwarranted.

Classically, the eventual target of successional development is the climax community - a stable community associated with a given climate. Can this condition be reached before the dynamic climate changes? If so, is there more than one stable climax community?

2. Migration—the arrival of organisms at the location
3. Ecesis—the establishment of organisms at the location
4. Coaction—the interactions, particularly competition, among the organisms
5. Reaction (or facilitation)—the modification of the site by the organisms and the subsequent change in the relative abilities of the organisms to establish and survive
6. Stabilization—the development of a stable community called the climax community
Stabilization is less a process and more a consequence of the iterative reapplication of the migration, ecesis, coaction, and reaction processes until a stable community is reached. Clements’ concept of the climax community was that there was only one stable vegetation type that was in equilibrium with the regional climate. Succession was the orderly, predictable, progressive, and linear development toward this climax community.

B. Application of Clementsian Concepts
The Clementsian succession paradigm had a pronounced effect on ecology in the United States in the first half of the twentieth century (much less so in Europe), and it shaped many of the laws and policies on the use of public lands. Earlier ecologists developed some of these ideas, and others evolved over Clements’ and associates’ scientific careers. Important aspects in the Clementsian paradigm, many of which continue to be a part of natural land management today, are

Community-level management: The idea that one could use the state of ecological communities to evaluate their past and present conditions and to predict their future is a significant contribution of Clements and associates. Although it has been established by paleoecological studies that communities are not necessarily stable over century and millennial timescales, short-term changes and disturbances are often seen as changes in communities. Conservation groups often make considerable effort to preserve unique communities (as well as unusual or important species). Wildlife and endangered species management is often based on maintaining particular communities or habitat types appropriate to the survival of focal species of animal or plants.

Indicator species concept: Clements believed that the presence or absence of particular species could be used to assess the state of a community and its potential for agricultural conversion. For example, the presence of poisonous or distasteful plants on a range indicates overgrazing, the occurrence of certain species of Lupine (Lupinus platensis) in a Nebraska prairie connotes a deep soil suitable for tilling and agriculture, and land with plants such as Salicornia might be so loaded with salt as to be unreclaimable. One could use indicator species to determine the past history of a landscape and to predict its future changes and potential uses.

Climax community concept: The climax community is the community that is stable in a given climate condition and is the ultimate product of successional processes. Current vegetation maps, particularly for large regions, display the expected climax community or potential natural vegetation expected in a region rather than the actual vegetation found there. Parks and wilderness areas are managed to maintain the natural climax community that is typical of the region. As an issue in conservation of diversity, often considerable effort is made to preserve unique communities (as well as species).

Progressive nature of succession: Clements viewed succession as being progressive (moving in a positive direction) toward the climax community. The succession progressed toward more diverse, more stable, and more desirable communities. This is a concept associated with eighteenth-century intellectuals and tied to other concepts such as the divine design of natural systems and ideas about the ‘balance of nature’ and the antiquity of certain ecological communities. Whether from Clements or earlier sources, these ideas have considerable influence on the aesthetics of conservation in the valuation of wilderness and the assessment of the importance of preserving certain species rather than others.

Perhaps the most enduring legacy from Clements is his pioneering attempt to synthesize quantitative observations about succession into a unified theory. Even strong detractors of the details of Clements’ concepts are still involved in understanding succession as a general phenomenon.

C. Alternative Organic Theories on Succession
Clements' theory is the most frequently presented organic succession theory, so much so that textbook writers often term Clementian theory to be “classic” succession theory. However, all American and British ecologists did not accept these ideas. This rejection was certainly the case for ecologists from continental Europe. Some of these ecologists emphasized a more individual species-oriented mechanistic description of succession. Others shared an interest in the holistic causes of successional change but differed from Clements in significant details. H. C. Cowles, who in 1899 was one of the first American ecologists to study ecological succession, characterized succession as “a variable approaching a variable.” Succession was considered the change of a system perturbed away from—but moving toward—an equilibrium that is itself changing. Cowles believed that the climax community was never reached. In 1935, A. G. Tansley contrasted Clements' climatic
climax or monochimax theory with a polycimax theory that had the climax vegetation of a region as a mosaic of local vegetation climaxes related to local conditions and disturbance history. In 1913, W. S. Cooper studied the forest succession on Isle Royale (Michigan) and found that the mature forest was a mosaic of patches of different ages and not the uniform climax community expected in Clementsian succession. He also believed that succession took multiple pathways and was not a linear progression of changes in seral stages. Similarly, in 1901, Cowles noted, "Succession is not a straight-line process. Its stages may be slow or rapid, direct or tortuous and often they are retrogressive" (Bot. Gazette 31, 73–108, 145–182). Recognizing that there is and has been considerable difference in opinion among ecologists who use organic explanations of ecological succession, there is an even stronger contrasting of concepts between Clementsian succession and mechanistic explanations of ecological succession emphasized by some ecologists.

III. MECHANISTIC EXPLANATIONS OF ECOLOGICAL SUCCESSION

Some of the debate about the nature of succession concerns the degree to which the vegetation can be arranged into communities that are natural units of biological organization. Introductions of modern biology or ecology texts often have diagrams of biological organization that illustrate an organizational hierarchy from cells to tissues to individuals to populations to communities, etc. Such progressions convey the idea that the community is a unit of organization as demonstrable at a level such as the existence of the liver or spleen as an organizational unit is at some other level.

The American H. A. Gleason (in 1926) and the Russian L. G. Ramensky (in 1924) emphasized that vegetation was mostly the consequence of the chance arrival of species at a location and the subsequent interactions among the available species to produce the observed pattern of relative species abundance. These interactions did not produce distinct unit communities. The vegetation was believed to vary continuously with changes in the underlying environmental conditions. Under this "individualistic" view of ecological succession, the process of succession was a consequence of species interacting with one another in the context of the environment to produce vegetation dynamics or successional change. Today, most ecologists subscribe to this individualistic view of ecological succession.

A. Descriptive Mechanistic Models of Succession

When one compares modern descriptive models of succession with the Clementsian model, it is apparent that there is a substantial difference in the spatial scale considered by the two schools. Clements' climax community was considered by him to be a phenomenon that occurred over large areas. This is evident by the fact that Clementsian succession proceeded toward a regional-scale climax. Also, the union of the climax community (vegetation) and associated animals was a "biome" of which there were thought to be only 13 in non-tropical North America. Clements believed that when the community in a given location was too small to have all of the species represented, succession might take different courses. However, he also believed that the succession processes he had described (nudation, ecesis, coaction, and reaction) are still represented by a mechanistic explanation of succession, albeit with different names and differing emphases. The differences are most pronounced with regard to the importance of the reaction (facilitation). Certainly in many primary successions, the changes produced by one set of species appear necessary for the success of the next. For example, the rapidly growing willows (Salix sp.) that stabilize sandbars in rivers seem to be a necessary preamble to the success of subsequent species, the ability of alder (Alnus sp.) to symbiotically fix nitrogen increases the fertility of sites uncovered by receding glaciers; and the organic acids produced by blue-green algae, lichens, and mosses speed the breakdown of granite and the development of a thin soil to support grasses, herbs, and even trees in primary successions on granite outcrops. However, some species appear to block the success of others and hold sites against species that might otherwise succeed them. In some secondary successions, all the species involved in the succession are present as seeds or other propagules from the initiation. In these cases, the familiar successional sequence of grasses and herbs yielding to shrubs and then to trees may reflect a difference in rate of growth of individuals present from the start. Evolutionarily, it is difficult to explain why a species would evolve to help another take over a site it could otherwise occupy.
Connel and Slatyer (1977) developed a descriptive model of the succession processes based on a mechanistic understanding of succession (Fig. 1). They used the reaction/facilitation issue to frame three models of succession based on mechanisms of interaction (the facilitation model, tolerance model, and inhibition model). In Fig. 1, the facilitation model is most like Clementsian succession as typically interpreted. The three models are different and have different implications for land management and particularly land reclamation. If one had the objective of restoring degraded landscapes, then one might speed the restoration by eliminating established species in the case of the inhibition model, but this would be ill advised in the facilitation model (Fig. 1).

B. The Landscape as a Mosaic

In addition to an expansion of the facilitation concept associated with Clementsian succession, there has also been an emphasis on understanding the mosaic nature of vegetation. Historically, this emphasis derives from earlier concepts of ‘cyclical succession’ (exemplified by the cyclical replacement series involving a forest canopy gap) and the ‘polyclimax’ (the mature forest as a mosaic trees, gaps, and recovering gaps). Theories about the dynamics of landscape mosaics have been developed in forests to a significant extent, probably because, regardless of other sources of spatial heterogeneity, a forest canopy is a mosaic of tree crowns (Fig. 2).

Starting with a small plot of land in a mature forest dominated by a single large tree, the large tree shades the ground and reduces the survival of smaller trees and seedlings below. There may be a few smaller shade-tolerating trees that survive under the large tree, but these are strongly suppressed in their rate of growth. The large tree dominates the resources (light, water, and nutrients) that are available at the site and blocks other trees from growing at the location. When this tree dies, the forest floor (where there had previously been little chance of a young tree’s survival) becomes a nursery for small seedlings and saplings. There is adequate light and other resources, and hundreds of small trees survive and begin to grow toward the canopy. As these trees grow, they begin to compete with one another. Some of the trees lose to more vigorously growing competitors. Eventually one tree manages to win the race to be the local canopy tree and begins to
eliminate the others. This represents the closure of the cycle with a large tree again dominating the site. Over the course of time this cycle is reiterated by the death of the new dominant replacement tree. The implications of the cyclical nature of small-scale forest dynamics were clearly elucidated by Cooper in 1913 and by A. S. Watt in a classic paper in 1947.

The expected changes in the amount of living material (biomass) over multiple iterations of a gap generation and filling should create a "saw-toothed"-shaped curve. This curve drops abruptly with the death of a canopy dominant and then builds biomass as the regenerating trees grow, compete, and occupy the site (Fig. 3, top). The distances between the "teeth" in the saw-toothed, small-scale biomass curve are determined by how long a particular tree lives and how much time is required for a new tree to grow to dominate a canopy gap.

FIGURE 1. Mechanistic models of ecological succession (reproduced with permission from Connell and Slatyer, 1977).
The larger scale biomass dynamics (Fig. 3, bottom) is a simple statistical consequence of summing the dynamics of the parts of the landscape mosaic. If there has been a synchronizing event, such as a clear-cutting or other disturbance, one would expect the landscape mosaic biomass curve to increase as all of its parts are simultaneously covered with growing trees (Fig. 3,a). If the trees over the area have relatively similar longevities, there is also a subsequent period when the deaths (and biomass) on plots where a canopy tree happens to die are balanced by those on plots where large trees are still growing. During this period, the loss balances gain in biomass and the curve levels (Fig. 3,b). If there is sufficient synchronization in the sawtooth curves of the component plots, this is followed by a period during which many of the pieces that comprise the forest mosaic all have deaths of the canopy-dominant trees (Fig. 3,c) and landscape biomass decreases. Over time, the local biomass dynamics become desynchronized and the biomass curve varies about some level (Fig. 3,d). This mosaic of repairing gaps with all different stages of recovery represented on the landscape can be taken as the mature forest.

The mature forest should have patches with all stages of gap phase dynamics and the proportions of each should reflect the proportional duration of the different gap replacement stages. T. C. Whitmore believed that this was the expected pattern and process for all forests and asserted in 1982 that forests of the world are fundamentally similar, despite great differences in structural complexity and floristic richness, because processes of forest succession and many of the autecological properties of tree species worked out long ago in the north temperate region, are cosmopolitan. There is a basic similarity of patterns in space and time because the same processes are at work.
Whitmire is referring to mosaic forest canopies as a consequence of gap replacement processes. The occurrence of such patterns has been documented for several different kinds of forests. The presence of shade-intolerant trees occurring in patches in mature undisturbed forest is another observation consistent with the mosaic dynamics view of mature forests. The scale of the mosaics in many natural forests is somewhat larger than one would expect from gap filling of single tree gaps, indicating an importance of phenomena that cause multiple tree replacements. Also, relatively long records (approximately 40 years in most cases) of forest structure and composition indicate a tendency for the forest composition to fluctuate with species showing periods of relatively weak recruitment of individuals to replace large trees and strong recruitment in other periods.

C. Quantitative Mechanistic Models of Succession

A consequence of recognition of the mosaic nature of vegetation and an emphasis on more mechanistic representation of the succession process has been the development of quantitative models that can predict changes in vegetation structure (Shugart, 1998). Several of these models simulate the successional dynamics by accounting for the birth, growth, death, and interactions with the environment for the hundreds of individual plants living on a small plot of land. The predictions of hundreds of these plots are then combined to obtain a prediction of the change in ecological landscapes. Because they simulate the fates of each of the millions of plants involved in a landscape succession, these models are called “individual-based” model. These models require considerable computational power of modern computers.

An advantage of individual-based models is that the following implicit simplifying assumptions associated with other modeling approaches (e.g., the Markov process or differential equation-based models) are not necessary: (i) The unique features of individuals are sufficient to the degree that individuals are assumed to be identical, and (ii) the population is “perfectly mixed” so that there are no local spatial interactions of any important magnitude. Most ecologists are interested in variation in individuals (a basis for the theory of evolution and a frequently measured aspect of plants and animals) and appreciate spatial variation as being quite important. These assumptions seem particularly inappropriate for trees which are sessile and which vary greatly in size over their life span. This may be one of the reasons that tree-based forest models are among the earliest and most widely elaborated of this genre of models.

One group of individual-based models simulates the establishment, diameter growth, and mortality of each tree in an area the size of a gap left by the death of a canopy tree and is called gap models. Gap models can be used as an example of the more general individual-based modeling approach. In most gap models, calculations are on a weekly to annual time step. Early gap models were developed for a size unit (approximately 0.1 ha) approximately that of a forest canopy gap. Gap models feature relatively simple protocols for estimating the model parameters. For many of the more common temperate and boreal forest trees, there is a considerable amount of information on the performance of individual trees (growth rates, establishment requirements, and height/diameter relations) that can be used in estimating the parameters of such models. Gap models have simple, general rules for interactions among individuals (e.g., shading and competition for limiting resources) and equally simple rules for birth, death, and growth of individual trees (based on the natural history of each species).

Gap models differ in their inclusion of processes that may be important in the dynamics of particular sites being simulated (e.g., hurricane disturbance, flooding, and formation of permafrost) but share a common set of characteristics. These latter characteristics involve an emphasis on the demography and natural history of plant species, relatively general rules for physiological tradeoffs among species, and an emphasis on the understanding of successional processes at the whole plant level. Each individual plant is simulated as an independent entity with respect to the processes of establishment, growth, and mortality. This feature is common to most individual tree-based forest models and provides sufficient information to allow computation of species- and size-specific demographic effects. Gap model structure emphasizes two features important to a dynamic description of vegetation pattern: (i) the responses of the individual plant to the prevailing environmental conditions and (ii) how the individual modifies these environmental conditions. The models are hierarchical in that the higher level patterns observed (i.e., population, community, and ecosystem) are the integration of plant responses to the environmental constraints defined at the level of the individuals.
IV. SUCCESSION AND BIODIVERSITY

Since the initial formulation of a progressive, holistic concept of ecological succession, there has been a tendency to associate positive attributes to mature communities. Hence, there is an expectation for increasing successional age to be associated with increased biotic diversity. There are certainly ecological systems (notably many forest systems) that demonstrate this pattern, but there are other examples of ecological successions that show highest levels of species diversity (measured as the number of species or by standard indices of diversity) at intermediate successional ages (e.g., shortgrass prairie in Colorado) or even at initial successional stages (e.g., boreal forests in areas of Canada). Succession on sand dunes on the coast of Queensland, Australia, has the highest species richness in shrub-dominated communities at the beginning and end. The pattern of diversity in different successions seems to be an idiosyncratic consequence of the attributes of the participating species and the environmental conditions at a given location.

A. Biodiversity on Disturbed Landscapes

At the landscape level, the overall biodiversity can be related to the frequency and intensity of disturbance and there is evidence from both theoretical work and observations that an intermediate level of ecological disturbance can produce the most diverse landscapes (Huston and Smith, 1987). This occurs in part because disturbed landscapes have a mixture of species able to successfully occupy the differently aged patches created by the disturbance history of a given landscape. One would generally expect highly heterogeneous landscapes to be more diverse. Disturbances also prevent particularly well-adapted species from occupying the entire landscape.

In 1979, J. P. Grime developed a “triangle” based on three primary plant response strategies that can be used to develop rules that can be used to predict the proportions of each strategy (and associated life-forms) expected under a particular environmental regime. Grime recognized two types of external factors limiting the biomass of plants. The first was stress, involving the conditions that restrict plant productivity (shortage of light, $H_2O$, mineral nutrients, etc.). The second was disturbance, involving partial or total destruction of plant biomass (activities of herbivores, diseases, fire, frosts, etc.). These two external factors can operate independently so that there are four possible combinations of high or low stress and high or low disturbance. Grime reasoned that the combined action of high stress and high disturbance created a condition from which the vegetation could not regenerate. Low-stress and low-disturbance environments would ultimately favor species that were able to compete effectively against other species (competitor strategy), high-stress and low-disturbance environments should favor short-lived, fast-growing species (ruderal strategy). The general problem exemplified by Grime’s work in development of these primary plant strategies is one of identifying plant functional types. An essential basis of this as well as other functional classifications of plants is that of “tradeoffs”—the idea that, due to underlying rules that derive from the species physiology and natural history, a single species cannot simultaneously be the best as a stress tolerator, a competitor, and a ruderal species.

B. The Gain and Loss of Species with Succession

One can view the richness of species on a parcel of land with the same disturbance history as a consequence of the gain and loss of species. The gain of species involves factors such as the migration and establishment of species (or the species gaining sufficient abundance or size so it can be sampled). There has been debate as to whether the gain of species at a site over succession is a consequence of the different species present at the site growing and developing at different rates or is actually due to species establishing themselves in an ordered sequence. The succession is a consequence of different growth rates of an initial inoculum of seeds of all the species involved germinating and growing at different rates and is sometimes called the initial composition model (Egler, 1994). The idea that one set of species is added to the community after a previous set has modified the site and lost the site has been termed the relay floristics model (Egler, 1994). To separate these two “models,” a particular succession study has to be sampled intensively enough to actually detect all the species at a location, which rarely occurs. Considering a wide range of successions of plant communities, animals associated with succession, and heterotrophic successions (such as the progression of changes in a decaying log), one finds that the appearance of species with abundances sufficient to be counted varies with the particular collection of species and with the changes in the physical environment associated with the succession.
The loss of species from successional communities tends to be somewhat more regular in its pattern of variation. Considering a wide range of communities (both autotrophic and heterotrophic successions), one finds that the rate of species local extinction through succession tends to decrease logarithmically over successional time. Successional sequences are often sampled and reported using a more or less logarithmic sampling regime (e.g., communities in a successional study might be sampled at 1, 2, 5, 10, 17, 33, 60, and 100 years). This reflects the pattern that one tends to sample successional sequences using designs where a proportion of the species in a site of a given age were found in the younger sites. Figure 4 illustrates this pattern for five different successional sequences in different locations. The initial loss rates of species are on the order of approximately 10% of the species found on a plot disappearing each year. In the later successional stages, this decreases to less than 0.1% species lost per year (Fig. 4). Because this pattern occurs across a wide range of successional sequences, this does not appear to be a consequence of the later successional species living longer (as is often the case in forest successions).

See Also the Following Articles
DISTURBANCE, MECHANISMS OF • ECOSYSTEM, CONCEPT OF • ECOSYSTEM FUNCTION, PRINCIPLES OF • INDICATOR SPECIES • LANDSCAPE DIVERSITY

Bibliography
I. A Concept in Search of Practice
II. Trends and Transitions
III. Indicators

GLOSSARY

anthropogenic disturbances Disturbances of human origin, including modifications of ecosystem structure or function and displacement or removal of species from habitats.
demographic transition Transition of human populations from conditions of high birth rates and high death rates to low death rates, followed by low birth rates; in progress since the seventeenth century, with a large increase in the number of humans—projected to be completed in the twenty-first century.
ecosystem services Flows of services of natural origin which are valuable to human users and occupants of an ecosystem, e.g., water purification by flowing streams and crop production from fertile soils.
indicators Quantitative measurements of environmental and social variables that provide time series describing long-term trends; some of the trends may indicate a transition toward sustainability.
maximum sustainable yield Estimate of the size or proportion of standing stocks of a population that may be harvested without altering the long-term abundance of characteristics of the stocks.
sustainability transition A search for sustainable development, through action and research, pursued during the remaining decades of the demographic transition. A sustainability transition would be shaped by normative goals for human well-being and preservation of the life-support systems needed by human populations.
sustainable development A pattern and path of economic and social development compatible with the long-term stability of environmental systems, particularly those essential to human well-being.

Biodiversity is dependent on sustainable utilization of the natural world by humans, and the conservation of biodiversity may well be essential to the durability of the human species. An increase in human-caused impacts on the natural world during the past two centuries is the driving force behind a great extinction, a large-scale reduction of the diversity of biota on the planet that has occurred only five times in the geological record. This sixth great extinction is the first to be caused by a living species.

Is a sustainable economy possible? Recent scientific appraisals suggest that it is but that a transition toward sustainability will require significant social, political, and technological changes during the next two generations. This is also the time period in which human population seems likely to level off; hence, it is possible
to think of a sustainability transition on the timescale of the demographic transition drawing to a close during the twenty-first century.

A sustainable economy is not a well-defined objective, so a transition toward sustainability must be a search rather than a march. Awareness of long-term trends and transitions together with indicators to inform our searches are important contributions that science can provide in addition to developing means for reconnecting human prosperity to the diverse and essential riches of the natural world.

I. A CONCEPT IN SEARCH OF PRACTICE

The idea of sustainability—that the fruits of nature, if harvested at moderate rates, may be reaped indefinitely—is ancient wisdom. However, translating that verity into workable policies is difficult and elusive. As economies and human population have grown rapidly during the past 200 years, exploitation of ecosystems for human gain has usually ignored sustainability and often depleted biodiversity. This may change in the next several decades as land transformation and human appropriation of ecosystem services surge toward natural limits and the growth rate of the human population declines toward zero. Still, more than 800 million people face hunger during at least a part of each year, and two-thirds of the human race live in developing countries, a label that implies priority for human welfare over the conservation of species and ecosystems. Is human well-being dependent on the survival of biodiversity, and if so how? Should attempts to improve the material conditions of human life be constrained by attempts to ensure the long-term survival of habitats and species? Species extinction is one of the only indicators of irreversible environmental loss widely accepted by laypersons; what is its practical significance, both biologically and socially? The connection between sustainability and biodiversity is neither conceptually clear nor practically straightforward, but it is of fundamental significance.

A sustainable material economy may be defined as one in which anthropogenic disturbances to ecosystems—those of human origin—are smaller than or similar to those due to other causes. The magnitude and scope of anthropogenic change today are unprecedented and in many important dimensions have been accelerating during the past two centuries. Although disturbances of human origin raise obvious concerns about their sustainability, a practical operational definition of sustainable practices has been difficult to articulate or to implement.

In nature, unsustainable behavior is self-limiting, but material constraints on human welfare have been slow to appear. Humans live and work in virtually all parts of the planet, exchanging goods in a global economy. In some respects, richer societies can transfer the burdens of their unsustainable practices onto poorer ones; some of the best agricultural land in the tropics has been committed to the production of sugar, tea, coffee, chocolate, and other luxury goods for temperate-zone markets for centuries. However, reallocation of the costs of unsustainable activities is not the only process at work. Remarkably, indicators of well-being, such as life span, income, and education, have increased for nearly all human populations during the past several decades, when reliable data series have been collected. As a result, direct signals that the human situation cannot be maintained remain fragmentary and inconclusive, and they may have been obscured by technological innovation and changes in social organization. Although many species and ecosystems have been lost and transformed, there are few instances known to the general public in which changes have been forced on people by the destruction of biodiversity.

This article approaches sustainability as a long-term phenomenon reflected in local and national cultures as well as in a changing global sensibility about the place of humans in the natural world. From this perspective, the survival of today’s biodiversity is likely to be determined by a larger historical process: whether the human economy can develop processes and governing mechanisms to achieve material sustainability. If not, our species will likely prove to be shorter lived than most in the geological record; however, we shall wreak havoc, as we are now doing, on a scale seen only a handful of times before in this record. Even if we are able to achieve material sustainability, this will not happen soon enough to save all that is endangered today. However, what would remain—under long-term human care—is likely to be a monument to our species worth contemplating seriously.

A. Science and Sustainability

The related ideas of sustainability and sustainable development have been influential throughout the rise of ecological science and environmental policy. Sustainability has in the course of this development acquired social significance in ways that were not in-
tended by the scientific community but that matter nonetheless.

That ecosystems have a finite capacity to support any population is evident from the uneven production of crops throughout human history. As human numbers and consumption of natural resources increased in the nineteenth century, so did concern that the impact of our species on the natural world would have irreversible, ultimately self-destructive effects. That such fears had substance was clear from the extirpation of valued species such as the passenger pigeon; thoughtful writings such as George Perkins Marsh’s *Man and Nature* (1864); the establishment of national parks beginning with Yellowstone (1872); and the founding of citizen groups as early as the British Commons, Open Spaces, and Footpaths Preservation Society, organized in 1863. However, it was also evident that farming, fishing, forestry, and manufactures had persisted for centuries in many places, often without noticeable decline in landscapes utilized and inhabited by humans.

The answer to this paradox seemed to lie in the idea of sustainable yield—the notion that most natural populations could be harvested, to a degree, without reducing their capacity to reproduce abundantly. Exceed the sustainable level and populations would dwindle even in the long run; harvest less, and part of the population would live out its life without benefiting humans. Therefore, there is an optimum (for humans) for each population, the maximum sustainable yield. This concept reinforced ideas of efficiency that were influential at the turn of the century in the United States when Gifford Pinchot, founder of the Forest Service, brought the concept of sustainable yield into public policy.

Sustainability held the promise of perpetual income, a flow of returns that humans might receive from nature in return for intelligent management. In this way, the stewardship ideals of an agricultural society might be extended to wild lands and sea. However, the promise was incomplete in significant ways. First, unlike the cultivation of annual crops, forests and fisheries were composed mainly of slower growing populations, in which the effects of habitat management or harvest on population size were delayed for years or decades. Moreover, the species of interest lived in ecosystems that humans could not manage as decisively as they did farmland or plantation. Third, mobile species such as fish were often exploited under open-access conditions, in which the short-term incentives facing harvesters ran counter to the interest of the community of harvesters or even themselves over the long term. Where these complications prevailed—as they often do even now—sustainable yield often proved to be a way to rationalize unsustainable behavior. Estimates, inevitably clouded by uncertainties and environmental fluctuations, would “inform” decisions that often led to overharvest or to irreversible habitat transformations such as the building of dams. Even the lessons of mismanagement are shaped by the analyst’s understanding of ecology and sustainability. As Cronon (1992) has shown, the American Dust Bowl of the 1930s can be seen as a tragic ecological error or as a heroic struggle against nature, such contrasting narratives imply contrasting recommendations for current and future management.

That a scientific idea such as sustainable yield could be misused reinforces an implicit but important message: Science has normally been the servant of those who would transform natural ecosystems for human gain. Although the content and methods of science might be neutral, both the scientists who inferred practical implications and the way science was used often advanced or legitimated an extractive, exploitative agenda. Proponents of sustainability, like those advocating conservation and environmental protection, have found this imperialist mantle difficult to recognize and even more difficult to shed. Taking this burden into account is of instrumental significance in debates and activities in the future.

Science plays another role in sustainability as the wellspring of technology—the methods, tools, and means humans have used to extend their control over the natural world. Guided by the idea of carrying capacity, ecologists have estimated that humans now appropriate slightly less than half the net primary production on land and approximately the same fraction of fresh water. Such estimates indicate that further economic growth or improvements in human well-being—even if they occur—may not continue to increase the size of the material economy on which life depends. Indeed, it seems likely that achieving sustainable economies will require decreasing the burdens of wasted energy, discarded materials, and pollution that are now imposed on the environment. Technology plays a strategic role in this aspiration.

Surprisingly, the definition of a sustainable material economy remains problematic: Which stocks must be preserved? Which flows must be conserved and at what levels? Is there a single numerate that can indicate sustainability or its opposite? There are no definitive answers to these questions. The following well-known statement was put forward in 1987 by the World Commission on Environment and Development chaired by Gro Harlem Brundtland:
Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs. It contains within it two key concepts:

- The concept of "needs," in particular the essential needs of the world's poor, to which overriding priority should be given; and,
- The idea of limitations imposed by the state of technology and social organization on the environment's ability to meet present and future needs.

This definition of sustainable development describes environmental limitations in terms of technology and social organization and not the carrying capacity of nature in an objective sense. This language reflects an important conclusion: There is no consensus on which material conditions do or will limit activities that obviously cannot continue to expand in the long term. Instead, social constraints, in the form of decisions to follow some paths rather than others, may avoid some undesirable consequences even as new technologies enable Homo sapiens to continue to evade others. It should be noted that many people, including many natural scientists, doubt that constraints can emerge in a voluntary and humane fashion. If not, the needs of the poor may not be the driving force of a sustainable future, as proposed by the Brundtland Commission.

Since we lack a crisp operational definition of sustainability, the topic is best organized around questions. What is the evidence of recent trends on the magnitude and tendency of human-caused disturbances? Is there a way to define indicators of sustainability that could guide social constraints and the search for new technology? Is there a way to link such indicators to a process of social learning? In the following sections, the surprisingly mixed evidence of human impacts is summarized. The picture they provide is fleshed out with a wide-ranging review recently completed by the U.S. National Research Council. These set a context for the notion of a sustainability transition.

II. TRENDS AND TRANSITIONS

The prospects for sustainability should be judged against the situation of the present and the trends that have carried us to this point. Such a review demonstrates the grave risks faced by biodiversity in the next two or three human generations but also indicates important opportunities. Three trends—population, economic growth and consumption, and biodiversity—are reviewed in the following sections. The picture they provide is fleshed out with a wide-ranging review recently completed by the U.S. National Research Council. These set a context for the notion of a sustainability transition.

A. Population

Human population at the beginning of the twenty-first century is 6 billion people, more than triple the 1.65 billion alive in 1900. About four-fifths of today's population lives in the less developed areas of the world. With an annual growth of 1.3%, about 80 million people are being added to the planet each year, nearly all of the net increases occurring in the developing world. Global growth rates are declining and have been doing so since the peak rate in modern history, about 2.2% per year, occurred in the early 1960s. Because this slowly declining growth rate is applied to a large base, absolute population growth will remain high for the next few decades. Human population is expected to increase by 2 billion between 2000 and 2025, the same amount as in the last fourth of the twentieth century.

The deceleration of population growth, the closing phase of a process referred to as the demographic transition, first became apparent in studies of European demography. This scenario is now believed to be operating globally, although with considerable variations from the simple pattern described here. Within two centuries the European population went from conditions of high births and deaths to the current situation of low births and low deaths. Initially, deaths declined more rapidly than births, and population grew rapidly. Later, birth rates decreased to match or even exceed the decline in the death rates, and population stabilized and sometimes declined. We are now in the midst of a global demographic transition that is more rapid than its European prototype. Birth and death rates in developing countries have decreased at unexpectedly rapid rates. The average number of births for each woman of reproductive age has declined to three compared to six at the post-World War II peak of population growth. The mortality transition has also proceeded very rapidly,
with life expectancy at birth having increased from 40 years in 1950 to about 64 years today. Whether these changes will lead to an increasing, decreasing, or stable population is unknown, although the rate of increase has clearly declined during the past three decades. Today's population growth has immense momentum because large new generations of young people are reaching reproductive age. How much population will increase depends on their choices of family size and their ability to implement these choices. Policies designed to encourage such implementation may be able to slow growth considerably. Indeed, recent rates of decline in fertility have outpaced earlier projections, and the United Nations (UN) reduced its medium expectation of global population in 2050 from almost 9.8 billion in the 1994 projection to 9.4 billion in the 1996 projection. By the end of the twenty-first century, the world's human population is projected to reach 10.4 billion, a level that seems likely to be subject to significant change by policies or by inaction.

### B. Economic Growth and Consumption

The growth of wealth has been as persistent as the growth of population. Moreover, barring severe disruptions to the global economy, income and consumption will continue to grow faster than human numbers. Supplying the energy and materials needed to support increasing consumption and addressing the environmental problems attendant on their extraction, consumption, and disposal may be the most significant challenge to sustainability, especially as more people adopt the materials-intensive, consumptive lifestyle now enjoyed by most people in industrialized nations. There have been dramatic changes in human well-being since the early nineteenth century, when the modern pattern of industrial and information-intensive economic growth became clear in the historical record. Trends in gross domestic product (GDP)—a measure of the total economic activity in a nation's markets—reflect a nation's production and wealth per capita and hence give an indication of the well-being of that country's people.

There has been an average worldwide gain in GDP per person by a factor of 7.9 between 1820 and 1992; in the four “Western offshoots”—Australia, Canada, New Zealand, and the United States—economic growth has resulted in a gain of more than 17-fold over this span of approximately six generations, doubling economic output within each human life span. Even in Africa, the region with the weakest record, economic output per capita had tripled by 1980. Indeed, contemporary Africa is approximately at the level of the United States in the 1840s, when Henry David Thoreau undertook his famous sojourn to Walden Pond to escape his countrymen's materialism. It is important to bear in mind that economic statistics such as GDP provide only a partial measure of human well-being. Pollution, which diminishes the value of ecosystem services or valued activities and assets, is excluded from conventional GDP accounts, whereas the costs of abating or repairing the damage caused by pollution are counted as contributions to economic output.

### TABLE I

<table>
<thead>
<tr>
<th></th>
<th>Actual—1950 compared to 1993</th>
<th>Scenario projection—1995 compared to 2050</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>2.2%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Food (gram)</td>
<td>2.7%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Energy</td>
<td>4.4%</td>
<td>2.8%</td>
</tr>
<tr>
<td>Economy (GDP)</td>
<td>3.1%</td>
<td>4.3%</td>
</tr>
</tbody>
</table>

*Source: National Research Council (1999).*

Averages also fail to account for disparities in the distribution of wealth. Disparities in incomes are widening and are likely to continue to do so in the absence of strong remedial actions. The gap is growing between rich and poor countries as a whole and between the rich and poor within many countries. On a global basis, the ratio of the income share of the richest 20% to the poorest 20% doubled during the past 30 years from 30:1 to 60:1.

Demand for energy and materials has approximately tracked growth in total economic output. As consumption has increased, however, use of energy and materials has become more efficient on average. Table II provides a description of consumption changes during the past 50 years and a scenario-based projection for the next 50 years. The scenario projection is intended as a reference case against which to test alternative sets of expectations; the reference scenario is not more likely than other forecasts, but it does provide one consistent set of assumptions as a benchmark.

### C. Species and Ecosystems

In the geologic record, paleontologists have found five mass extinction events, each of which drastically re-
duced the number of species on Earth. Each time, enough life-forms survived to repopulate the waters and lands of Earth. The impact of human activities on the planet has accelerated the loss of species and ecosystems to a level comparable to a sixth mass extinction, the first driven by a living species.

Rates of species extinction have been estimated to be 100–1000 times higher than they were before large-scale human dominance of ecosystems. These rates of loss are driven primarily by the alteration of natural habitats. The extirpation can be much higher than the proportional loss of land area if habitats are fragmented by roads or other human clearings that reduce unconverted land to isolated patches. Depletion of species in these remaining patches can be delayed but will occur in the absence of human action to restore and re-connect habitat.

In addition to causing extinction, human activities also introduce species into ecosystems in which they have not been present. In some situations, exotic species proliferate, and the introduction of a new species can transform the ecological relationships of these habitats and further stress endangered species.

Accelerating rates of depletion and change cannot continue for long. Habitats are finite and the loss of species is irreversible. Efforts to slow the rate of depletion and change require some recognition of the value of the system, whether in terms of provision of resources and services or on aesthetic or ethical grounds. However, the value placed on species and ecosystems has been uneven historically. A small suite of edible and useful species have been valued, conserved, and propagated by humans. These number in the thousands out of tens of millions of species. However, for every species such as cacao, the far more complex ecosystems within which these valuable species live have been neither understood nor managed until very recently. In some cases, including cacao (which grows better in tropical forests than in plantations), the environing ecosystems play an important role in the economic return to human cultivation.

In a small number of cases, endangered species have recovered when protected by human efforts. For example, the bald eagle and some marine mammals in North America have rapidly rebuilt their numbers. These recoveries have taken place under conditions in which habitats required by the species were either intact or readily protected, the species had not been damaged beyond recovery, and the public will to preserve or conserve them was strong. Such circumstances do not currently apply to most regions of the world where species are threatened.

Land conversions and land degradation also degrade or destroy ecosystems and the services they provide to humans. Freshwater, coral reef, and forest ecosystems have suffered enormous assault from human activities. Covering less than 1% of the earth’s surface, freshwater ecosystems have lost the largest proportion of species and habitats when compared with other ecosystems on land or with the oceans. Continuing overfishing, dam building and river development, and contamination will continue to place greater threats on freshwater ecosystems. Many estuaries and bays have deteriorated because of activities associated with land development and fishing pressure, undermining ecosystem services. For example, the oyster population of Chesapeake Bay once filtered a volume of water equal to that of the entire bay about once a week; overfished, they now filter that volume in about 1 year’s time. This has adverse effects on the water quality of the estuary and on the many species that live in it. More than half of the world’s coral reefs face changes in species composition, obliteration, and other major ecosystem effects. These losses in turn affect the livelihood of local communities that depend on the reef for food, tourism, and protection against damaging storms. Increasing need for fuelwood and land for agriculture, together with industrial logging, resulted in a net loss of approximately 180 million ha of forest between 1980 and 1995.

Losses of freshwater ecosystems, coral reefs, and forests inflicts large losses of ecological services on local human communities, many of which are poor. These services include water purification, flood control, recycling of nutrients, mitigating climate and temperature extremes, and the production of crops and forest and marine species.

The scientific understanding of conservation biology has taken clear form only in the past 15 years, although humans motivated by aesthetic and cultural considerations have been seeking to preserve places for millennia. As these two currents of human activity converge, there has been increasing awareness of the need to conserve ecosystem processes, including evolution. This implies working at the scale of whole landscapes, with explicit attention both to the preservation of critical ecosystems and to the interactions between human activities and the managed and uninhabited ecosystems among them. It is an open question whether a reformulated conservation of this kind will prove workable in enough places to salvage the biological richness of the planet. Moreover, the abilities to assess and monitor the well-being of Earth’s living resources and the services they provide are far from proven, but efforts in these areas will be
essential to understand the role of species and ecosystems for a sustainable future.

D. An Overview of Trends

A broad review of long-term trends affecting sustainability was completed in 1999 by the Board on Sustainable Development of the U.S. National Research Council (NRC). The survey’s findings included many that shape the social context in which a search for sustainability will take place:

- Although humans have modified their habitats for more than 10,000 years, during the past three centuries humankind has developed the capacity to change the environment on a scale that equals or exceeds natural rates of change. Most of this human-caused disturbance has occurred within the past two generations.
- The twentieth century was a time of notable transitions: The demographic transition was described previously; governance and politics have been changed by decolonization from 1945 to the dissolution of the Soviet Union, together with the rise of nongovernmental organizations; and several important biogeochemical cycles have moved from control by natural forces to anthropogenic dominance. These are historic changes in driving forces, with opportunities for sustainable governance as well as risks of catastrophe.
- The past generation has seen notable improvements in human well-being. Since 1960 life expectancy worldwide has increased by 17 years, infant mortality has been cut in half, access to safe drinking water has approximately doubled, primary school enrollments have increased by two-thirds, and per capita income has more than tripled. The developing world has gained as much ground in a single generation, in material terms, as the developed economies achieved in a century. These averages are misleading if reported without acknowledgment of continuing, worsening inequalities; the number of people living in absolute poverty (less than $1 a day using 1985 purchasing power as a standard) is more than 1 billion and increasing.
- The pattern of continued growth hides a sequence of technological transitions, each lasting two or three generations, that have been discerned by economic historians. Coal, the dominant energy source of the late nineteenth century, was displaced by oil and natural gas. Some national economies have shifted their centers of gravity from agriculture to manufacturing and then to services. Increasingly, products are transported throughout the global marketplace, whereas land and ecosystems remain fixed. The causes and details of these transitions are poorly understood, but their existence implies opportunities for sustainability, as suggested by the effective control of some pollutants in many developed countries. Technological transitions have been accompanied by dislocations of human communities, such as the displacement of rural African Americans into the industrial cities of the United States as mechanization transformed agriculture beginning in the 1920s. Such dislocations produce both permanent and transient social changes, exacerbate conflict, and affect attitudes toward the environment and sustainability.
- A majority of the human race live and work in urban centers, a shift that seems likely to persist. In the remainder of the demographic transition, approximately 100 million new urban dwellers will appear each year, almost all of them in developing countries. The challenges and opportunities of meeting their needs for ecosystem services such as clean water is likely to play a major role in the future of biodiversity, both in metropolitan regions and in agricultural areas, coastal zones, and forests.
- Human cultures are increasingly connected and increasingly aware of their diversity. Since 1930, trade between nations has grown more than twice as fast as their economies, with concomitant increases in the power of transnational firms and financial institutions. The diffusion of ideas through low-cost communications has spread Western images of high consumption and also environmental concern. A notable result of increasing connection has been the rise of the nongovernmental organization as an influential vehicle for articulating identity, demands, and pressures for institutional change and innovation. The notion of sustainability, and the expectation that a substantial degree of material equity must be part of a sustainable future, is likewise a product of this more connected global society. Diversity also fosters conflict under some conditions, bringing risks of terrorism, war, and civil disorder.
- The intensification of agriculture in the past 50 years has transformed the character of human dependence on the land. Increasing agricultural outputs have outpaced human population growth and dramatically reduced the incidence of famine. If increases in yield can be continued as the growth of human numbers slows, it may be possible to release a substantial fraction of land committed to crops today, permitting regrowth of forests similar to that observed in some developed countries. Intensification of agriculture also brings costs and risks to both nature and society, of course, and it

SUSTAINABILITY, CONCEPT AND PRACTICE OF
remains an open question whether the advances of recent decades are sustainable.

The scale and intensity of anthropogenic disturbances threaten to undermine the sustainability of human populations in some regions. Kasperson et al. (1995) performed a comparative study of nine regions that are under severe pressure using criteria that defined unsustainable damage in terms that encompassed the social capabilities of the human population. Six of these regions were found to be likely to be environmentally unsustainable in the future, but only one, the Aral Sea, was already unsustainable. More broadly, the global-scale analysis of Turner et al. (1990) concluded that the rates of anthropogenic disturbance are accelerating in some areas but decelerating in others; these judgments are summarized in Table II.

Environmental pressures do not always increase with consumption and economic output. An interesting array of patterns has been found as shown in Fig. 1, which plots various indicators of environmental pressure against per capita GDP on a logarithmic scale. Economic growth in cities requires that people living in dense settlements avoid one another’s pathogens; this seems to correspond to the availability of clean water, shown in the top two panels in Fig. 1. The two bottom panels indicate that material consumption, as measured by the production of trash and carbon dioxide, shows no sign of abating with income. However, the middle panels indicate that, in some cases, pollution increases with income until a demand for environmental quality appears and is met by governments. These patterns, which have also been found in other measures of environmental and resource exploitation, suggest the possibility of a long-term transition in environmental problems: from dirty water to industrial pollution to climate change and from public health to government regulation to technological transformation. Thus, the attention directed to managing municipal wastes and greenhouse gases may foreshadow effective action even in the lower panels of Fig. 1. The social dynamics driving these correlations have yet to be elucidated, however, and the policy implications are debated more than they are acted on. It seems imprudent, in any case, to rely solely on increasing income to produce effective environmental control before irreversible damage is done.

The lower right panel of Fig. 1 shows a local effect, emission of carbon dioxide, whose impact is global. The climate changes due to this anthropogenic disturbance are already occurring, with time lags that are not well understood. Changes in the composition of the atmosphere from human activities have been small in percentage terms but exert a large effect because of the role of greenhouse gases as thermal gatekeepers in the heat balance of the planet. These changes in greenhouse gas concentrations suggest that the course on which the world economy is embarked is unsustainable at the global level, affecting all regions and ecosystems. The 1992 Framework Convention on Climate Change commits the nations of the world to a goal of stabilizing "atmospheric concentrations of greenhouse gases at a level which would prevent dangerous human interference in the climate system." However, agreements under this convention during the twentieth century named only at stabilizing emissions of greenhouse gases. Even if this goal were achieved, the inventory of climate-altering gases would continue to increase with continuing emissions. Climate change is likely to occur at rates far more rapid than vegetation, in particular, can migrate; therefore, the implications for biodiversity from continued anthropogenic disturbance to the atmosphere are large and perhaps grave for many ecosystems.

In the marine environment of the United States, one-third of the species whose status is known are overfished. The prevalence of unsustainable exploitation of marine species used by humans is similar elsewhere. Forests continue to be lost in the tropics, where deforestation exacts much higher costs in biodiversity than it does in temperate forests. As with fisheries, the institutional means and political will to manage for multiple species and purposes on a sustainable basis remain elusive.

---

**TABLE II**

<table>
<thead>
<tr>
<th>Magnitude, Recovery, and Rate of Change in Human-Induced Transformation of Environmental Components</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Magnitude of change since 10,000 years before present</strong></td>
</tr>
<tr>
<td>50% total change reached in nineteenth century</td>
</tr>
<tr>
<td>Rates decelerating since 1990</td>
</tr>
<tr>
<td>Carbon tetrachloride, lead, sulfur releases</td>
</tr>
<tr>
<td>Marine mammals</td>
</tr>
<tr>
<td>Rates accelerating since 1990</td>
</tr>
<tr>
<td>Carbon, nitrogen, phosphorus releases</td>
</tr>
<tr>
<td>Forest diversity</td>
</tr>
<tr>
<td>Sediment flows</td>
</tr>
<tr>
<td>Water withdrawals</td>
</tr>
</tbody>
</table>

*Source: Turner et al. (1990).*
• Fresh water, a basic necessity that cannot be economically transported over long distances, is in short supply or contaminated in many parts of the world. Many urbanizing areas are likely to face severe problems as their human populations grow.

• The NRC survey did not include trends in warfare, terrorism, or civil violence, but as events in the Congo River basin and the Andes demonstrate, civil disorder can affect biodiversity and the prospects for managing land substantially. The limited ability of governments to manage violent conflict must accordingly be borne in mind.

• Disease organisms and vectors may be expanding in response to anthropogenic disturbances. These disturbances take various forms—artificial selection of drug-resistant strains of pathogens, diseases jumping across species as humans unwittingly invade or enhance the habitats of pathogens and carriers of disease, and expansion of the ranges of some disease-bearing organisms with climate change and international trade.

Overall, sustainability is not a single or simple objective. Rather, sustainable human economies are likely to
emerge, if they do at all, through social change and institutional adjustments that are only partially determined by rational processes and scientific knowledge. Rationality and science provide necessary guidance but are not likely to be sufficient in themselves.

In all, the NRC review demonstrated the abundant hazards and barriers to a sustainable future. However, the trends of the past have often elicited technological and social adaptations, turning a challenging hazard into a sometimes wrenching transition. The NRC study concluded that a determined search for a sustainability transition is worthwhile, as long as there is a commitment to recognizing inevitable surprises and to learning from experience and science.

E. A Sustainability Transition?

Using the demographic transition as a model, one can ask whether it may be possible, on scientific and institutional grounds, to achieve a sustainability transition. The demographic transition marks a long-term shift from high birth rates and high death rates to a society characterized by low birth rates and low death rates. Because death rates have been lowered first, there is a large net increase of population. A sustainability transition, by analogy, would be a change from today's patterns of highly dissipative use of energy and materials, together with severe and widening inequality, to a future in which use of energy and materials is far more conservative and material inequalities have been moderated, meeting the basic needs of all. Since the wasteful phase of growing consumption and high inequality came first, even a successful sustainability transition would entail significant losses of biological diversity and suffering in human populations; both are currently under way, of course. The NRC study summarized previously concluded that a sustainability transition is scientifically possible, although there remain many technological problems to be solved. It is unclear, however, whether the political will and leadership to assemble the needed social and institutional changes will be forthcoming.

A sustainability transition would pursue the following themes, inferred from the trends reviewed previously:

1. Completion of the demographic transition, with efforts to decrease the total population level by further facilitating choice of family size and additional moderating of population momentum
2. Conserving and rebuilding biodiversity using landscapes that reach beyond existing protected areas
3. Major improvements in the efficiency and end products of energy use, including alternatives to fossil fuels and mitigations and adaptations to climate change
4. Major improvements in industrial ecology: the ability to do more with less, in every material dimension, closing the loops of materials flows and increasing the efficiencies of production and use within these loops
5. Rethinking and changing the way cities are built, particularly in developing countries, so that the ecosystems services needed by urban populations can be affordable and accessible as cities continue to grow
6. Creating human and institutional capital to enlarge the existing capability to recognize, debate, and cross the hurdles that block the path to sustainability— to learn to negotiate a sustainability transition
7. Fostering social and political choices that can make sustainability transitions credible and feasible in democratic discourse

This is an agenda for scientific research, technological innovation, and social experimentation as well as activism. The first four items listed previously have long been staples of environmental studies and political action, whereas the last three raise new opportunities or re-formulate existing ones. These themes entail social and behavioral changes that challenge widely held values. Like all social changes, these will be contested and outcomes will be shaped by struggle as well as technical and economic feasibility.

It is far from clear that a sustainability transition is socially feasible. However, the pace at which humans are disturbing natural processes and ecosystems leaves few options other that pursuing change as rapidly as practicable. The idea of a sustainability transition proposes a linked set of goals and strategies.

III. INDICATORS

Sustainability is clear in principle but obscure in practice. The search for a sustainability transition needs to take this difficulty seriously for several reasons. First, the histories of environmental science and environmental politics demonstrate the ubiquity of surprise. This is likely to be characteristic of a path toward sustainability as well. Sustainability, like environmental quality, demands the reconciliation of multiple goals that are both partially incompatible and largely incommensurate. Second, the momentum of economic development and the social aspirations that propel it imply that those pursuing sustainability need to make good use of opportunities rather than await crystallization of a social con-
In day-to-day life we use prices, news and weather reports, and other routine methods of monitoring to guide our behavior and expectations. Indicators perform parallel functions for long-term changes and large-scale actions. As the members of the European Community prepared to institute a common currency, the nations agreed to meet numerical guidelines for their budget deficits as a fraction of GDP. This is a striking instance of the influence of indicators: For some time, these fractions superseded the electoral mandates of the national governments. Sustainability cannot be tracked by a single indicator; therefore, sets of parameters have been proposed to sense trends in social and environmental change.

These parameters have been selected by experts and citizens in quasi-governmental settings, usually using a pressure–state–response (PSR) framework. PSR posits linkages between human action and environmental consequences. Human activities exert pressures, such as burning gasoline in cars, that alter the state of environmental variables, such as the quality of air in a city. Impaired states in turn elicit responses, such as regulations governing pollution-control technology in new vehicles. These three classes of variables identified by PSR can be measured often using data already collected for administrative purposes. Combining these data with a simple but flexible scenario captures a central dynamic of sustainability: Humans can impair the life-support systems of the natural world, calling forth responses intended to protect environmental quality.

Using the PSR framework, governments and nongovernmental organizations have compiled numerous sets of indicators for sustainable development using various measurement regimes. However, because there is no widely accepted operational definition of the term “sustainable development,” proposed indicators are often scrutinized more for their moral, economic, and political implications than for their scientific substance. To date, no single set of available indicators has gained wide acceptance.

Despite these difficulties, there are numerous efforts under way to assemble indicators of sustainable development. These efforts range on a governmental scale from municipal to international and on an ecological scale from watersheds to the planet as a whole. Hundreds of indicators and numerous schemes to collect, analyze, and aggregate the information needed to form sets of indicators have been proposed. Two efforts by the UN Commission on Sustainable Development (UNCSD) and the World Bank demonstrate the way in which different approaches to indicators complement one another. The UN indicators were assembled using the PSR approach. Selected through a consensus process without an agreed operational definition of sustainable development, the 134 indicators are numerous, diverse in the methods used to measure development or sustainability, and include many indicators where reliable measurements do not exist.

The World Bank, in contrast, focused on estimating only three indicators, which it called capital accounts. Each attempt to capture the value to national economies of a vital aspect of the world. The most familiar account, “produced” capital, is normally called national wealth—physical capital and financial claims. A second account measures natural capital, the resources and capitalized value of services provided by the natural world. In principle, this would include standing timber,
These efforts bring together many sources of illumination to human welfare and the condition of the environment. The projects carried out during the decade since the agreement on the meaning of sustainable development are limited in their effectiveness by lack of indicators but does not provide a framework for monitoring. SCOPE effort outlines a broad research agenda for indicators useful for a sustainability transition. That there is no consensus on the end point of sustainable development is different, however, from seeking indicators useful for a sustainability transition. Like the slow loss of wetlands or the buildup of carbon dioxide in the atmosphere, a sustainability transition may well require quantitative indicators to be detected—that is, sensing the degree to which human needs are met and the degree to which life-support systems are put under stress. Human welfare, monitored since the depression of the 1930s, illuminates the role of governments in addressing hunger and poverty. Appropriate quantitative indicators have been developed, but at the turn of the twenty-first century they were not yet being reported regularly in a forum, such as the annual report of the UN Development Programme, in which crises could be highlighted and emergent transitions discerned. Environmental indicators are numerous and need conceptual organization. One approach is geographic (NRC, 1999), focusing attention on four scales of significance to a long-term transition. These could be reported regularly by the UN Environment Programme, for example:

- Global circulatory systems, including the atmosphere, climate and ocean circulation, trade and travel, and the spread of exotic species and diseases, are affected by human activity. Trouble in the circulatory systems is important because the scale of circulation can implicate the entire planet more rapidly or persistently than governments can regulate. Although greenhouse gases and ozone-depleting substances had been carefully studied by the end of the twentieth century, biological and information-intensive flows were not assessed as components of a global circulatory system. In particular, practical ways to monitor the flow of invasive species and exotic pathogens still had to be developed, implemented, and reported together with the other circulatory systems.

- Regions vulnerable to critical environmental damage are being identified. As described previously, the Aral Sea has experienced severe, possibly irreversible damage; the Valley of Mexico appears to be headed in
SUSTAINABILITY, CONCEPT AND PRACTICE OF

the same direction (Kasperson et al., 1995) as suburbanization continues apace. Because the governing factors, including a vulnerable natural setting, lack of feasible alternatives, and governing institutions unable to take effective action, vary among regions, there is not a single set of indicators that can monitor the combination of social and natural factors that lead to irreversible damage.

• Ecosystem services are only beginning to be monitored. The need is especially great in developing countries, where urbanization is continuing to accelerate, resulting in greater stress on geographically limited supplies, ecosystem services, and their associated infrastructures. Because many ecosystem services are not priced (e.g., flood control) or allocated by monopolies (e.g., water supply), market signals are absent or misleading. However, ecosystem services tend to be irreplaceable when damaged and inflexible because they are supplied by costly infrastructure.

• Parks and protected areas, created to enable their biota to persist indefinitely, have been identified on a place-by-place basis rather than through a consistent set of appraisals of their long-term sustainability. Attempts to conserve biodiversity on larger scales are just beginning. There are compilations at the species level of habitats with high levels of biodiversity but no regular comparative monitoring approach that could begin to build data series in which trends could be discerned.

In each of these settings, indicators form an indispensable but incomplete part of the intelligence needed to perceive and encourage a sustainability transition in the future.

A complementary function of indicators is to inform public policy and other decisions, large and small. Some navigational aids are available, although it seems likely that others may also come into wide use:

• National capital accounts, emerging from the World Bank’s work, could provide a simple but crude measure of sustainability: Each of the three kinds of capital might be transformed by human activity, but as long as a nation’s total capital increases over time its trajectory may lie in roughly the direction of a sustainability transition. The word “roughly” is important: A utilitarian metric is inadequate to assess sustainability. One should accordingly expect spirited debates over the value, biases, and meanings of indicators. In the related sphere of economic policy, one can observe that during the past 50 years there have been major disputes in the United States regarding economic growth, the incidence of poverty, unemploy-

C. Indicators and Social Learning

There is no agreement about which indicators to use to measure the movement of economies toward or away from sustainability. One should accordingly expect spirited debates over the value, biases, and meanings of indicators. In the related sphere of economic policy, one can observe that during the past 50 years there have been major disputes in the United States regarding economic growth, the incidence of poverty, unemploy-

ment, and inflation. All these are indicators that Ameri-
can politicians think will influence voter behavior. Remarkably, the independence of the data-gathering and analytical organizations has survived, despite their location within government agencies.

This is an important lesson: The independence of science is central to the social value of scientific information. Another lesson is that surprise is a valuable indicator. Governments and societies should anticipate unexpected things to happen. In a policy context, this idea is a kind of precautionary principle: Because surprise is likely, action should be undertaken with thought, humility, and caution.

Indicators used to report on a sustainability transition are likely to be biased, incorrect, inadequate, and indispensable. Obtaining the correct indicators is likely to be impossible in the short term. However, not trying to obtain the correct indicators will surely compound the difficulty of discovering, inventing, and achieving sustainable development.

1. Social Learning

As advisers to colonial powers and national governments, naturalists and ecologists have exercised influence on biodiversity and rare species through most of the history of ecological science. As the magnitude of anthropogenic disturbances increased during the past century, scientific understanding of the intricacy and vulnerability of ecosystems increased dramatically as well. The established means of preserving biological diversity—seed banks, zoos and captive breeding, and protected areas—do not suffice to safeguard the evolutionary heritage of the planet. A goal of environmentalists is to conserve biodiversity in more ambitious ways, working at the scale of landscapes that continue to be used and inhabited by people. These efforts should be viewed as part of a broader search for sustainable economies.

The search for sustainability, in turn, is a search rather than a march to a known destination. Salvaging the world’s dwindling biodiversity is a race against anthropogenic disturbance. It will be lost in some places, as it has been in many already. The challenge is to learn from these experiences how better to provide a social environment in which people may meet the needs of the current generation while ensuring future generations the biological heritage with which to choose how better to be human in a natural world.

Acknowledgments

This article draws on work done by me and others for the Board on Sustainable Development (1999). I am indebted to Robert Kates, William Clark, Darby Jack, Jerry Mahlman, and Sherburne Abbott for guidance, assistance, and encouragement.

See Also the Following Articles

| AGRICULTURE, SUSTAINABLE | ECONOMIC GROWTH AND THE ENVIRONMENT | ECOSYSTEM FUNCTION, PRINCIPLES OF | ECO SYSTEM SERVICES, CONCEPT OF | HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW | MASS EXTINCTIONS, NOTABLE EXAMPLES OF POPULATION STABILIZATION, HUMAN |

Bibliography


Turner, B. L., Clark, W. C., Kates, R. W., Richards, J. F., Mathews,
SYSTEMATICS, OVERVIEW

Quentin D. Wheeler
Cornell University

I. What Is Systematics?
II. Elements of Biodiversity
III. Predictive Classifications
IV. The Missions of Systematics
V. Role of Taxonomy in Biodiversity Studies and Conservation

GLOSSARY

adjacency Relative position of alternative states of characters prior to any hypothesis of polarity.
allopatry Species or populations occupying separate areas.
anagenesis Origin of species (or characters) within a lineage without splitting (cladogenesis).
apomorphic Derived, relative to sister group.
area cladogram A branching diagram showing relative divergence of geographic areas.
autapomorphy A derived character unique to a single species or clade.
binomial (= bionominal) nomenclature Naming species by combination of a specific epithet and the genus to which the species belongs.
bio-geography, historical The study of spatial distribution of taxa through time.
character An attribute that is constantly distributed among all members of a species or a clade.
clad Branch on cladogram.
cladogenesis Origin of branch or clade.
cladogram A general graphic representation of a cladistic hypothesis; more general than a phylogenetic tree.
convergent characters Resemblance due to independent evolutionary events.
dendrogram Cladogram or tree; sometimes used specifically for phenogram.
dichotomous tree A cladogram in which all nodes are bifurcate.
dichotomy Two taxa arising from one node.
endemic taxa Species or clades unique to one geographic place.
holomorphology Totality of characters indicative of phylogeny (including molecular, behavioral, etc.).
homology Two structures hypothesized to share a common ancestry.
homoplasy Similar attributes not due to common ancestry but rather to convergent evolution, reversals, or mistaken interpretation.
macroevolution Evolution above the species level.
microevolution Evolution within species.
monophyletic group An ancestral species and all of its descendant species; evidenced by synapomorphy.
node Branching point on cladogram corresponding to hypothetical common ancestor.
ontogeny Sequence of embryonic and postembryonic changes in characters of an organism in the course of its life.
paraphyletic group A group including an ancestral species and some but not all of its descendants; evidenced by synaplesiomorphy.
parsimony principle  Seeks to minimize the number of character transformations (a special application of Occam’s razor).

phenogram  A branching diagram showing overall levels of phenetic resemblance rather than clado-genesis.

phylogenetic tree  Common synonym for cladogram; or a more specific interpretation of phylogeny identifying ancestors and descendants.

phylogram  cladogram; strictly, a cladogram interpreted to show terminal taxa and ancestral nodes (intermediate specificity between generalized cladogram and specific phylogenetic tree).

plesiomorphic  Primitive, relative to sister group.

polarity  Relative apomorphy/plesiomorphy of character states.

polyphyletic group  A taxon for which all included species (or taxa) do not share a most recent common ancestor; based on homoplasies.

polytomy  Many branches arising from a single node or an unresolved (portion or whole) cladogram.

Popperian  A follower of philosopher of science Sir Karl Popper who accepts his falsificationist views.

semaphoront  Individual organisms at a particular point in their developmental (ontogenetic) history.

stem  Branch on cladogram between node and term.

sympatry  Species or populations co-occurring in same area.

symplesiomorphy  Shared, primitive similarity.

synapomorphy  Shared, derived similarity.

taxon (plural taxa)  Classification of species and higher taxa, ultimately based on phylogeny.

term  A terminal on a cladogram, either a taxon (in systematics) or a geographic area (in biogeography).

tree  Commonly a synonym for cladogram; or a cladogram interpreted in more evolutionary detail.

tritomy (“trichotomy,” in error)  Three undifferentiated branches or clades.

typification  The practice of designating a single specimen (or, formerly, a series) as the name-bearer for a species; or a species for a higher taxon, especially genera.
braced instead a wholesale focus on population biology. The widespread acceptance of this notion within biology assisted in the necessary growth of modern genetics but set the advance of taxonomy back by decades. Although many dedicated, ingenious taxonomists continued to make theoretical and practical progress against considerable odds, it was not until progress was made in phenetics ("numerical taxonomy" in the narrow, historic sense) and Hennig delivered the necessary theoretical justification that studies of phylogeny became unequivocally accepted again as solid science.

A predictive classification and the nomenclature that expresses it are invaluable for several reasons. First, a phylogenetic classification permits the retrieval of what is known about a taxon and predictions about attributes of taxa not yet observed. Second, formal nomenclature provides the vocabulary for biologists to communicate ideas and facts about biodiversity. With nearly 2 million species known and five or more times that number predicted, precision and accuracy in communication, retrieval of billions of facts about Earth's species, and the ability to predict where organisms with a particular property may be found will become immensely more important than they are already.

Taxonomy has been described as one of two fundamental divisions of modern biology. General biology is concerned with processes that are true for many kinds of living things or all of life. Experimental biologists fit within this division most of the familiar disciplines, i.e., ecology, ethology, physiology, and any others asking primarily "why?" and "how?" questions. A process is characterized in order to make predictions about the outcome of similar events in the same and as many other life-forms as possible, slicing through biodiversity in a functional or causal dimension. Taxonomists provide balance to this view through a broadly comparative and historically organized view of biological diversity. Given the immense variation in the ecology, behavior, physiology, and biochemistry of living things, there is in fact a single factor shared by all species: evolutionary history or phylogeny. Thus, a phylogenetic classification is the logical basis for the general reference system for biology. Taxonomy is widely misunderstood, largely because of its historicity. Evolutionary history, not unlike the sequence of events in human society, is not replaceable. The myriad of existing circumstances and preconditions cannot be enumerated, let alone recreated. Thus, the most fundamental assumptions underlying experimentation are violated by phylogeny. What is the universe of outcomes in the context of historical patterns among which one, at best, may be correct? Contributory processes and ancestral contingencies are so unpredictable a priori that efforts at randomization and modeling are little more than wishful thinking in most cases. Another effort has been made to force phylogenetic problems into a more familiar, population biology model. Taxonomy is unique among the biological sciences for both theoretical and practical reasons. Most significant among these are the following:

**Historical perspective:** Taxonomy is not infrequently viewed with suspicion because of its nonconformity to standard practices of experimental biology. There is a good reason for this: Taxonomy is not experimental biology. Phylogenetic history's unrepeatable sequence of unique events cannot be seen by rolling back time or repeating evolutionary trajectories. Fortuitously, and unlike human history, evidence of phylogeny is preserved in the genetic makeup of organisms belonging to descendant species. Taxonomy is no less scientific than experimental biology. However, it does require a different ontology and epistemology for its justification.

**Comparative method:** Taxonomy is comparative at virtually every level. Whether assessing species boundaries, hypothesizing homologous characters, or choosing among several equally parsimonious cladograms, taxonomists are comparing the characters of as many species in the groups under study as possible. Few other areas of biology are comparative, and none to this same degree.

**Worldwide collections:** This comparative aspect of taxonomy means that taxonomists must have ready access to specimens for every species of a group being studied regardless of where on Earth these species might live. To make such work practical, taxonomists have assembled vast synoptic collections of certain taxa in major museums and herbaria. Whereas many biodiversity scientists need to study the same organisms or phenomena in the same study sites for extended periods of time, the same constraints do not exist for taxonomy. Just the opposite is true: Taxonomists require samples from the most disparate reaches of the geographic range of a species and representatives of as widely differing forms within a species or higher taxon as possible. In order to unambiguously sort out species, the taxonomist needs to see as much infraspecifc variation as possible as well as to study specimens of closely related species whether or not they occur in the same places. In order to complete as comprehensive a study of an entire group as possible, the taxonomist needs to examine every species possible within that group regardless of where or when they live or lived. This is a very different requirement than that which faces the experi-
mental biologist, and it demands access to collections from wherever in the world members of a specialist’s taxon occur. Such collections are costly in time and money to be assembled and housed, laborious to maintain, and may only be accumulated with the combined efforts of many students and professionals over very long periods of time. The great natural history collections of the world have been built up over a period of one or two or more centuries and even the best of them contain but fractions of all the world’s species.

Taxon inventories: Building taxonomic research collections demands that scientists with taxon-specific knowledge be encouraged and permitted to collect in as many locations within the range of their group as is possible. Long series of specimens are needed to address infraspecific and transgeographic variation. Two centuries of taxonomic practice has taught us that making species decisions in isolation (i.e., describing species based on material restricted to one locality, country, or in many cases even single continents) leads to redundant species descriptions, synonymy, and taxonomic confusion. Such collection requirements, and the minimum useful data associated with specimens, vary sometimes greatly from the requirements of a population geneticist or ecologist.

Taxon scholars: Taxonomists spend their entire careers working on one or a few higher taxa. Because there are genera of arthropods with as many as 2000 described species, it is not uncommon for a specialist in the megataxa (e.g., fungi and insects) to learn hundreds of species in the course of his or her career. This taxon focus creates a very different perspective than one centered on a phenomenon or ecosystem.

In the philosophy of science, the demarcation problem involves how we can distinguish science from nonscience. Experimental biologists sometimes confuse the “scientific (experimental) method” with the limits of science. The failure of taxonomic theories to fit the requirements of experimental biology to be quantifiable by one of a handful of statistics—or, more precisely, the failure of experimental biology to capture that which experiments in the narrow sense are but one fraction of all the world’s species.

Stripped of the usual gauges of confidence limits in experimental biology, how can the historical patterns constructed by taxonomists be scientific? Phylogeneticists have found a powerful epistemic argument for their work within the hypotheticoeductive philosophy of science espoused by Sir Karl Popper. Knowledge does not exist with absolute certainty. Rather, “confidence” (in the common sense rather than statistical sense) increases in our ideas in proportion to the number of times that they are subjected to and survive critical testing. All-inclusive statements such as “all swans are white” are highly testable. Such law-like, all-or-nothing statements are in fact potentially falsified by a single observation. Aristotle took note of the hardened forewings of beetles called elytra. An implicit (and later explicit) hypothesis that “all beetles share elytra” has been corroborated approximately 400,000 times in the past 2000 years, but a single observation of a beetle without elytra tomorrow has the potential to refute it. Falsified hypotheses are discarded and new or modified hypotheses forwarded and tested in their place.

For the Popperian, the demarcation between science and nonscience is precisely that between the testable or potentially refutable assertion and the untestable claim. Unless it can be shown how a hypothesis is reasonably falsified, it is simply outside the bounds of science. This view of science is a general one, consistent with both experimental and comparative biological programs. Much of biology has become homogenized so that the protocols of experimental biology are viewed as more or less equivalent to the scientific method and not merely examples of it. As shown, Popper’s hypotheticoeductive approach explains the significance of congruent character distributions and refutational role of incongruent ones. The inductive reasoning so popular in the eighteenth and nineteenth centuries does not suffice for taxonomy. The failure of phenetics demonstrated that large amounts of data alone does not suffice to demonstrate relationships.

A Popperian approach to science proceeds approximately as follows:

1. Hypothesis: An initial hypothesis can derive from anywhere—observations of the world or merely conjecture. Hypotheses not founded on prior observations or based on few or poor observations, of course, seem less likely to stand when subjected to critical tests. The point is that where a hypothesis comes from is secondary to its formulation in a falsifiable form.

2. Predictions: From the initial hypothesis logical predictions are made about the natural world. Cladograms predict the pattern which synapomorphies will follow as discovered.
3. The successful test: When these predictions are realized, the hypothesis is corroborated. Given many such tests and much accumulated corroborative support, we feel increasing confidence in the accuracy of our assertion.

4. The unsuccessful test: When these predictions are not realized—even after many successful corroborative tests—then the hypothesis is called into question and the incongruent observations must be explained or the hypothesis rejected and modified or replaced with another.

5. Repeated tests: Regardless of whether it is corroborated (step 3) or replaced (step 4), the prevailing hypothesis is continually subjected to the rigors of further observations and continued critical tests.

Cladistics or phylogenetic systematics is the modern form of systematics that seeks to reconstruct patterns of evolutionary history. Darwin acknowledged the pioneering work on phylogenetic "trees" by Ernst Haeckel in Germany (Fig. 1), but a century would pass before a rigorous and cohesive theory of phylogenetic systematics emerged in Hennig's 1966 tome. Piecing together descent with modification is obviously complex, routinely encompassing worldwide patterns spanning millions of years. Despite the immense challenge of phylogenetic analysis, its fundamental tenets are surprisingly simple: Phylogenetic systematics was a logical extension and fully capitalized on centuries of steady growth in our knowledge of taxonomic diversity, merging as it did phylogeny and taxonomy.

Taxonomists study the patterns inherent among species due to relative recency of shared common ancestors. The distinction between pattern and process in evolutionary studies has evoked much controversy. Although different approaches and evidence might be required, all these varied ways of studying evolution are appropriate and ultimately necessary. Because taxonomists have sought to minimize the number of specific assumptions made about evolutionary processes, many microevolutionists have interpreted this to be a rejection or attack on their science. On the contrary, minimized process dependency of taxonomic hypotheses ensures that they are independent of any particular process, thus giving to microevolutionists a historical chronology within which any process of interest to them may be examined. The effort to purge phylogenetic analyses of unnecessarily specific assumptions does not mean that phylogenetic systematics is theory free or neutral. On the contrary, as most contemporary philosophers of science would agree, systematics, like all scientific assertions, is theory laden. Phylogeneticists have a detailed ontology and rich set of background assumptions that subdivide their science, as in their concept of characters discussed later.

The discovery of phylogenetic patterns not only can but also must precede insights into process. This is well established in respect to both how taxonomy is done and how it has progressed in the history of biology. Chapter 13 of the Origin of Species makes clear that Darwin was convinced of the fact of evolution from the patterns of character distribution documented by taxonomists. The existence of these character regularities led Darwin to seek an explanatory theory. Many taxa have been recognized for centuries, through changing ideas and social milieus. The same pattern has been visible to taxonomists whether they were creationists, microevolutionists, or cladists. A distinction exists be-

![FIGURE 1 Treelike diagram depicting Ernst Haeckel’s hypothesized phylogeny of life, cited for its genius by Charles Darwin (reproduced from Haeckel’s Generelle Morphologie der Organismen published in 1866).](image-url)
Hennig used the term molecular biology, biochemistry, ethology, and ecology. As expected, including anatomy, embryology, paleontology, characters. Character sources are as diverse as might be excepted; those defined by lacking such synapomorphies ("not-A groups"). The history of taxonomy may be told in the search to eliminate not-A groups and recognize A groups in their place.

A. The Evidential Basis of Taxonomy

Taxonomy is comparative in nature, and those attributes of species and higher taxa that are observable and comparable are characters. For the taxonomist, character is a precise term that is central to hypothesis testing and at the heart of his or her epistemology. A character is an attribute of a species that is constantly distributed among all members of that species. In the context of phylogeny, a character is such an attribute present in an ancestral species and in all of its descendant species either in its original or in some subsequently modified state. Attributes shown to be inconsistent in their distribution within terms on a cladogram (whether species or higher taxa), for example, are not characters in this strict sense but rather traits. Traits within a single species provide crucial evidence for the population geneticists but do not provide reliable indicators of common ancestry at the level of phylogeny. Traits at higher levels result in polymorphism in cladogram terms and the range of confusion associated with such. Considerable confusion persists in the literature that is actually nothing more than semantic misunderstandings spawned by not distinguishing these two concepts.

Any attribute that is heritable and that can differ constantly between species or clades is a potential character. Character sources are as diverse as might be expected, including anatomy, embryology, paleontology, molecular biology, biochemistry, ethology, and ecology. Hennig used the term holomorphology to express the totality of character sources, not limited in Hennig’s time or since to morphology. Hennig emphasized that the source of character data is of less significance to taxonomy than how that evidence is analyzed. Simple observations with a hand lens can be rigorously analyzed, and the most expensive molecular techniques can render data that are rendered useless by inappropriate analyses. Data that are not available for most taxa in an analysis or those of uncertain fidelity in their inheritance are frequently mapped onto a cladogram a posteriori. Recent arguments have been made in favor of “total evidence” analyses that combine all available sources of data. In practice, this has principally involved merging molecular and morphological data sets.

Mature sciences develop theories that are scientific (testable) yet not directly derivable from observations by the senses. Subatomic physics is based on the existence of particles never seen by anyone, but these ideas have been shown to map precisely onto predicted outcomes in power plants and atomic bombs. In taxonomy, characters represent a continuous intellectual tradition traceable to the notion of homology articulated by Richard Owen in 1842 and made consistent with evolutionary theory by Lankester in 1870, who differentiated homology from homoplasy. Today, homologies are theory-rich hypotheses of sameness that accommodate the “noise” of trait-level variation, incorporate existing knowledge from molecular and developmental genetics, and are more sophisticated than the simple phenetic idea that “they look the same.” Even the choice regarding what specimens to compare to one another has relevance since some characters are only observable during particular periods in the life history of the organism. These character-bearing periods are semaphoronts; thus, it may be necessary to compare a third-instar larva with the same, a neonatal female with the same, and so forth.

Consider another Popperian statement that “all reptiles have scales” and assume that we have a study set with a few species each of lizard, snake, and birds. We wish to use scales and feathers to sort out their relationships. A simple phenetic view of these characters (using characters as mere adjectival words rather than as evolutionary hypotheses and claiming thereby to be “objective”) would support the conclusion that lizards and snakes have scales and birds have feathers. This appears to be a matter of common sense and coding birds as lacking scales seems obvious. A more detailed consideration of the situation, taking into account phylogenetic insights as well as ontogenetic and genetic information, reveals that feathers are scales expressed in a highly modified form. The character “scale” in order to serve as a special similarity in the sense of Hennig must be interpreted to include its original state and all its subsequent modifications. Phylogeneticists would agree with pheneticists that scales are characteristic of a monophyletic group that we might call Reptilia. However, in order to be a monophyletic group Reptilia must be expanded to include birds. Snakes and lizards share nothing in common with one another not shared also by birds, either in an original or in a modified condition.

Hennig’s evolutionary novelties or apomorphies may be hypothesized at the outset, as in Hennig’s “search for the sister-group” hand method. Rationale used to determine the relative polarity of alternative conditions include outgroup comparisons, fossil precedence, and
ontogenetic sequence. Alternatively, global parsimony among all available characters may be used to infer the relative status of a condition as apomorphic or plesiomorphic and the polarity read off the cladogram post hoc. As data sets increase in size (number of taxa and/or number of characters) or in complexity, the latter approach implemented by any of a number of computer algorithms will prove more satisfactory. For relatively small data sets, it is possible to seek exact solutions with some of these programs. For large, messy data sets, options are used that approximate the shortest trees through random movement of branches and other procedures.

The simplest phylogenetic hypothesis is the three-taxon statement, which states that two taxa share a more recent common ancestor with one another than either does with the third. Implicit in this simple statement is the prediction that all synapomorphies discovered to vary among these three taxa shall be distributed so as to be consistent with the three-taxon statement.

B. Historical Context

The history of taxonomy is an extremely long one, with its roots lost in prehistory. The ability to distinguish among species in nature is not unique to humans and is part of even simple forms of animal cognition. Our concern here, of course, is the more formal human efforts to classify and name the various kinds of living things in a disciplined way. People living close to the land know and have words to refer to a surprising number of species, and it is estimated by ethnobiologists that several tens of thousands of species are in daily use by people throughout the world. Early hominids, at least those who survived to become our ancestors, had to distinguish poisonous plants from edible ones and prey items from potential predators of humanoids. Not coincidentally, the rise of civilization during the past few thousand years has been correlated with the increase in our knowledge of "other" species.

In the third century B.C., Aristotle had a remarkable conspectus of biodiversity and distinguished many taxa, some of which are considered valid to this day (e.g., Coleoptera). After the subjugation of Greece by the Roman Empire, however, these early flowerings of formal taxonomy ceased and were not resumed for another 1500 years. During these dark centuries, Aristotle’s insights were mongrelized with superstition into the essentialism attributed to him but not doing justice to his brilliance. An increasing interest in plants of medicinal value, published in ever-larger books known as herbals, helped fuel a renewed interest in classification. Cesalpino’s rudimentary species concept in De Plantis Libri, based on his observations that plant seeds give rise to like kinds, was a significant return to keen observation in taxonomy echoed in the animal world by John Ray in the middle of the seventeenth century.

The world of taxonomists rapidly expanded in the seventeenth century. Sailing ships were bringing many new kinds of plants and animals back to museums in Europe, and van Leeuwenhoek perfected simple microscopes revealing microbes to astonished biologists. Even as the known world increased in size, so too did efforts to learn more completely about the flora and fauna close to home. As naturalists sought to document species known to them, the absence of rules led to chaos, synonymy, and misunderstanding. Order came in the works of a Swedish taxonomist, Caroli Linnaeus (Fig. 2). In the writings of Linnaeus were deceptively simple criteria that, when applied with rigor, gave the foundations for all modern classifications. Among many contributions by Linnaeus, some deserve special mention. Linnaeus devised a stable system that avoided the rampant redundancy that had existed in names in the past. He adopted binomial nomenclature, the genius of which involved conveyance of ideas of relatedness in species names as well as maintenance of descriptive adjectives for use in multiple higher taxa. He proposed higher taxa that were ranked one within another, logically preadapted to meet needs for phylogenetic infor-
mation contained two centuries later. Also, he emphasized a character-based approach that made his ideas explicit and portable to the arrangement of specimens in collections and observations in the field.

The epistemic importance of a character-based system, and its logical justification, emerged from the comparative anatomical studies of Cuvier and the distinction by Richard Owen of analogy and homology. Darwin's *Origin* had little obvious direct impact on the practice of taxonomy, just as Darwin predicted. This was hardly surprising given the fact that Darwin derived his process ideas from the patterns so convincingly set out by taxonomists from Aristotle to Linnaeus and Darwin himself. On the other hand, the transformation of homology into evolutionary hypothesis by Lankester and his explicit recognition of homoplasy advanced the inevitable conceptual relationship between the classification of species and characters, the evidence for taxonomy.

Taxonomy entered a new period with the wide acceptance of evolutionary theory. Darwin boldly predicted that our classifications would come to be genealogies of species. Taxonomists almost immediately sought to realize this vision. John Henry Comstock, founder of the entomology department at Cornell University, wrote in 1893 that "the description of a species, genus, family, or order, will be considered incomplete until its phylogeny has been determined so far as is possible with the data at hand." Although the goal of phylogenetic systematics was soon clear to most post-Darwinian taxonomists, the method for achieving such phylogenetically informed classifications and names would not come together completely until the middle of the twentieth century. Comstock offered a method based on infusing functional considerations into the study of characters and taking on the study of character systems one at a time in sequence. Comstock's method foreshadowed the evolutionary taxonomy championed later by Mayr in the 1960s distinguished by (i) recognition of paraphyletic groups and (ii) desire to express both cladogenesis and the degree of divergence between branches. Because these neo-Darwinists wanted to express both degree of divergence and branching patterns in classifications, they ended up with many compromises and an excess of subjectivity from species limits to the highest taxa.

In part a reaction to this subjectivity—intellectual property of combining unnecessary evolutionary process assumptions with the act of classification—pheneticians or numerical taxonomists developed explicitly objective alternatives. Early pheneticians argued that community of similarity equated to community of descent and went to excessive lengths to compile as much raw data as possible. This quasi-inductive approach proved incorrect. Overall similarity did not reveal phylogenetic relationships. Rather than abandon phenetics, its practitioners admitted their failure but added that they now believed evolution to be unknowable. Many of the massive phenetic data sets were subsequently shown to be suspect; morphometrics were used extensively to generate a large number of "characters." Unfortunately, independence among such evidence was never shown and in some studies the same ratios were being measured many times. More damning to the pheneticists, however, were two additional observations. First, their results were only repeatable given the same matrix (i.e., given no addition or subtraction of characters or taxa between analyses) and the same algorithm. By this time, there were scores of algorithms, any of which might give a different result. Although each was "objective," the choice among them was purely subjective. Second, J. S. Farris demonstrated in 1979 that the phylogenetic system was superior in its information content, the very criteria utilized to promote phenetics. Farris had delivered a death blow from which phenetics would never recover.

Hennig's 1950 *Grundzüge Einer Theorie der Phylogenetischen Systematik* had been read and cited by a handful of North American taxonomists, a few of which, such as Pedro Wygodzinski, had applied his principles. Almost immediately after the publication of Hennig's *Phylogenetic Systematics* in 1966 (a new, English-language book), a theoretical revolution was afoot. Phylogenetic systematics developed rapidly along two parallel lines of scholarship that would converge years later. James S. Farris was transforming "numerical taxonomy" from the trivially objective phenetic methods that admitted to being uninformative about evolutionary history to sophisticated numerical approaches to phylogenetic analyses that took advantage of emerging computer technology to make possible the search for the shortest trees and the simultaneous analysis of large numbers of taxa and characters. Farris developed a rigorous quantitative taxonomy that was true to Hennig's theories and to the parsimony principle that he forged into the bedrock of phylogenetics. In so doing, Farris also laid the groundwork necessary to cope with massive numbers of characters that would later emerge from molecular sequence techniques. Simultaneously, Gareth Nelson, Norman Platnick, Eugene Gaffney, Donn Rosen, Joel Cracraft, E. O. Wiley, and other scientists associated primarily with the American Museum of Natural History were exploring the connection between the philosophy of science promulgated by Sir Karl Popper and Willi Hennig's theories and methods. The result...
was a powerful epistemic justification of Hennig and the transformation of cladistics to a rigorous science in its own right distanced from some of the evolutionary process assumptions common to taxonomic writings before Hennig and, to a lesser extent, inherent in Hennig’s writings.

C. Linnaean Nomenclature
Modern taxonomy is generally taken to refer to classifications beginning with the work of Linnaeus in the eighteenth century. In zoology, for example, the earliest available names for all groups (except spiders) are those printed in the 10th edition of Linnaeus’ Systema Natuarum, arbitrarily assumed to have been published on January 1, 1758. Any names proposed prior to that date are set aside. All those contained therein or published thereafter meeting applicable rules are available and are evaluated with regard to their appropriateness following an international code of rules, the most recent edition of which was published in January 2000 by the International Commission on Zoological Nomenclature.

Efforts to reject and replace the Linnaean system of names are proposed occasionally. To date, no such efforts have succeeded and the Linnaean system continues to serve as the backbone of information storage and communication about biological diversity. The most recent such proposition claims to be “phylogenetic” and would replace the familiar Linnaean categories with a rankless system in which names would be galvanized by reference to two common ancestral species. Proposers of rankless nomenclature emphasize putative stability of names in their arguments. This is ironic since phylogenetic theory demands names be sufficiently agile to reflect improved knowledge of monophyly, a necessary result of progress in cladistic hypothesis testing. Because this system cannot adapt to improvements in knowledge, from the perspective of cladistic information it is less stable than the existing Linnaean alternative.

The fact that the Linnaean system has been used continuously by taxonomists is not, of course, an accident nor a slavish adherence to tradition. Pre-Darwinians recognized patterns evident in character distributions that were used by Linnaeus and his students to group “related” animals and plants together. For Linnaeus, these relationships were the result of choices made by God in his Creation, but regardless of the causal process assumed, the evidence was expressible in hierarchic form. As the evolutionary implications of such patterns of resemblance became clear (as, for example, to Haecikel, Lankester, and other nineteenth-century biologists), improved explanations for the same undeniable patterns became available. Given a rigorous method for the study of phylogeny by Hennig, a classification system that was hierarchic in nature was a necessity to convey such relationships. The Linnaean system, due to its hierarchic logical structure and nested ranks, ideally met this need.

II. ELEMENTS OF BIODIVERSITY

We speak of biodiversity in quantitative terms, as in reference to one ecosystem having greater diversity than another, but sometimes fail to explicitly consider what the elements of biodiversity logically should be. In a general sense, of course, there are inescapable candidates: demes, ecosystems, populations, semaphoronts, genes, amino acids at particular sites, specific behavioral patterns, and so forth. Because species have traditionally served as the fundamental elements of nomenclature and of phylogeny, they clearly have special status in such attempts to quantify biodiversity. Ask any biologist about biological diversity of a place and his or her first response is likely to be some census or estimate of species numbers. Considering the necessary distinction between tokogeny and phylogeny, species are a given in such counts. However, species alone do not capture what most of us intuitively wish to express when comparing the diversity of areas.

Consider two islands of equal size with 10 species each. On one island are 10 closely related species of a single insect genus. On the other island are one each of species representing 10 higher taxa. In one trivial sense, the islands have equal diversity with 10 species each. In a deeper sense, the second island has more diversity, certainly in a qualitative sense but perhaps also in a quantitative sense. This is so if and when we take into account not only how many species live in a place but also what their relationships are to one another and in comparison with life on Earth elsewhere. Were all 20 of these hypothetical species unique endemics and we could save only one of the islands, I suspect that most of us would find this an easy decision.

In recent years, several numerical indices have been proposed to account for biodiversity at and above the species level. When biodiversity is seen through the lens of taxonomy, in terms of both species and phylogenetic diversity, needs and priorities for research and conservation can appear very different. Combined with research results from ecology, population biology, migration, and assessments of ecoeconomics and so-called biodiversity hot spots, these indices create a
powerful, science-centered basis for decision making in natural resource management and conservation biology.

Using the imagery of the world to represent broad scaled geographic models of biodiversity distribution on Earth, we can see three phases in the twentieth century. First was what Nelson and Platnick called the 'Sherwin–Williams' model, with biodiversity originating in Northern Hemisphere centers of origin and spilling down the earth in decreasing abundance. This traditional view was biased by many factors, including a focus on certain vertebrate taxa but especially the location of major universities, museums, and herbaria in Europe and North America.

As tropical forests were finally explored in greater detail, it soon became obvious that there were far more species in these than in any of the supposed centers of origin to the north. From this species counting exercise, a tropical biology mythology arose that posited that the tropics are most diverse, with biodiversity levels diminishing as one moves either northward or southward away from the equator. There are many taxa with enormous numbers of tropical species, and based on existing species counts the tropics unequivocally house more species than any other region in the world. This tropical view morphed the world into one with a very exaggerated equatorial region. For an increasing number of taxa, however, the largest number of species are found in southern temperate regions (as, for example, with many spiders as shown by recent studies by N. I. Platnick). Whether the southern temperates could ever compete for raw numbers of species seems doubtful, but Platnick's work led to another major shift in our conception of the distribution of biodiversity on Earth (Fig. 3).

Taking into account clades and subclades, Platnick argued convincingly that the southern temperates are home to a disproportionate number of relictual early clades. Just as marsupial diversity in Australia challenges the traditional eurocentric view of mammalian diversity, early arthropod clades make the southern end of the world diverse in amazing and unparalleled ways. According to Platnick's view, the world is most diverse in the tropics and in the southern temperates, resulting in a pear-shaped globe.

Taking into account clades and subclades, Platnick argued convincingly that the southern temperates are home to a disproportionate number of relictual early clades. Just as marsupial diversity in Australia challenges the traditional eurocentric view of mammalian diversity, early arthropod clades make the southern end of the world diverse in amazing and unparalleled ways. According to Platnick's view, the world is most diverse in the tropics and in the southern temperates, resulting in a pear-shaped globe.

Moving beyond species and clades, individuals differ one from another in genetic and other respects and any effort to draw a line across levels of biological organization is arbitrary. In this view, biodiversity is to be found down to the molecular level, where individual amino acids ultimately are characterized by physical properties rather than by meaningful individuality. Such a reductionist argument suggests that it is impossible to accurately describe or meaningfully contrast levels of biodiversity. The experience and practice of biologists and laypersons, however, suggest otherwise.

Most conservation literature refers to an ecological 'hierarchy' progressing from individual to deme to population to species to ecosystem. Such functional units of organization are necessary concepts for understanding the workings of living systems. They are not, strictly speaking, hierarchical. Because individuals may migrate into an area, die, or emigrate, the notion of a deme, population, or ecosystem is illusory. Actual combinations and numbers of organisms are in constant fluctuation and may change dramatically over geologic spans of time. Reference to key species is useful for this reason but does not negate the ephemeral nature of such assemblages. From the taxonomist’s perspective, biodiversity is organized otherwise. Individuals belong to populations (at least in sexually reproductive forms) and populations to species. Subsequently, species belong to
monophyletic clades of varied inclusivity. Because unambiguous indicia of common ancestry may only be retrieved in the realm of phylogeny rather than tokogeny, the elements of biodiversity are clearly species.

The hierarchic relationships that result from a shared evolutionary history may only be reconstructed among species. Higher taxa can be subdivided into less inclusive monophyla all the way down to pairs of sister species. Species, in contrast, cannot be divided into smaller units that retain the potential for unequivocal phylogenetic pattern recovery. Species, therefore, emerge as the logical elemental units for describing and discussing biodiversity. This also explains the intense interest in species concepts among biologists.

A great deal of confusion regarding systematic biology derives from frequent confounding of the kinds of genetic relationships described by Hennig as "tokogeny" and "phylogeny" (Fig. 4). Tokogeny refers to birth relationships, specifically the immediate genetic relations shared among actually or potentially interbreeding individuals. Tokogeny, then, is concerned with organisms within populations and the interchange of individuals (and thereby genes) among populations. In other words, tokogeny describes relationships below the species level. Phylogeny, in contrast, describes genetic relationships among species. Phylogeny refers to the pattern of descent with modification above the level of species. To some, species concepts have played an inordinately important role in biology. Because species recognition is a prerequisite for research in either tokogeny or phylogeny, it is not surprising that so much emphasis has been given to the so-called "species problem."

Since the early twentieth century, geneticists have focused so intensely on population, then molecular, mechanisms that it is widely accepted that the definition of evolution is changes in gene frequencies within and among populations through time. This is true for that part of evolution called microevolution but not for macroevolution. Phylogeneticists seek to unravel the pattern of evolutionary history above the level of species. For clonal, asexual kinds of organisms, mutations mark new beginnings from which all (not subsequently mutated) offspring will be essentially exact copies. For sexually reproductive organisms, however, geneticists have demonstrated that patterns within species are reticulate (due to virtually random episodes of reproduction, even at very low frequencies) rather than hierarchic. Our ability to retrieve historical patterns depends on the availability of evidence reasonably interpretable as hierarchic. Certainly there are histories within and among populations, but this complex mix of parent–offspring relationships (called tokogenetic relationships by Hennig) obfuscates the history signal. What, then, are the elements of phylogeny? That is, the units about which such history is retrievable and below which it is not? The answer appears to be the traditional one—species—even though no single species concept is agreed on by all biologists.

III. PREDICTIVE CLASSIFICATIONS

Cladistic hypotheses are most succinctly expressed in graphic form as branching diagrams known as cladograms (also phylogenetic trees, trees, phylograms, or dedrograms, sometimes with special meanings). Species or the least inclusive higher taxa included in an analysis are plotted as terms (terminals) at the tips of the finest branches of the tree structure. They are grouped together by the observed pattern of distribution of shared-derived similarities or synapomorphies. Synapomorphies, or shared apomorphies, represent evolutionary novelties. Such novelties are heritable characters, and as such are constantly distributed within the terminals defined by them, whether species or higher taxa. Attributes that vary within a terminal, in contrast, are termed traits even at higher levels. Characters are shared by all members of a taxon, either in their original (ancestral) condition or in a subsequently modified form, the mod-
ern interpretation of character formalized by Norman Platnick in 1979. Synapomorphies may occur once on a cladogram or more often if the overall parsimonious distribution of characters suggests multiple origins (convergence), losses, or losses plus gains (reversals). Hennig defined monophyletic groups as those including a common ancestral species and all of its descendant species. The goal of a phylogenetic classification is to make all groups monophyletic. Although monophyletic groups had long been recognized, Hennig's method was the first to require monophyly. Synapomorphies are taken as evidence of monophyly. Sympleiomorphy or shared primitive similarity erroneously leads to groups that include ancestral species and some but not all of its descendant species; these groups are known as paraphyletic. Errors are sometimes made while interpreting characters, and states are mistakenly grouped together as the same that have actually arisen from independent evolutionary events. Such instances of convergent evolution, parallelism, or reversal to ancestral character states are known as homoplasy. When groups are based on shared homoplasy, they include distantly related species that do not share a most recent common ancestor and are known as polyphyletic. Figure 5 shows examples of such groups referring to a flightless silverfish, a damselfly, a beetle, and a fly. The best available insect classification suggests that fly + beetle constitutes one monophyletic group whose sister group is the damselfly. The silverfish is then sister to the other three combined. Were the silverfish and damselfly grouped together based on their comparatively simple metamorphosis (lacking a pupal stage), they would be paraphyletic because simple metamorphosis is a shared, primitive similarity in comparison to holometaboly. Were the small, aristate antennae of the damselfly and fly taken as similarity, the group might be described as polyphyletic since it is based on convergent similarity and not on common ancestry.

Hennig was concerned specifically with historical patterns above the level of species, and the concepts described previously pertain to how species are grouped together into higher taxa and clades. Such ideas, of course, presuppose that there is some agreement with regard to the definition of species. Since the rise of modern genetic theory, biology in general has focused on the mechanistic aspects of species formation, basically asking "Why do species exist?" and "How do species originate and maintain their uniqueness?" Although the importance of such questions is obvious, another equally important and even more fundamental question was sometimes neglected: "What are species?" This "what" question differs fundamentally from the aforementioned questions in that it asks about a pattern rather than a process. This question, usually framed in terms of competing species concepts, was traditionally and is still logically within the purview of taxonomy. The answer to this question has direct and significant bearing on answers to related "why" and "how" questions.

The most elementary questions about biodiversity (How many species are there in a particular clade, place, or ecosystem?) require a scientific concept of species,
as do successful strategies, policies, and laws to protect biological diversity. Because taxonomists are primarily concerned with the discovery, description, naming, and classification of species, they have traditionally provided the answer to the “what” question. More literature has probably been devoted to species and speciation than to any other single topic in biology, but species concepts remain highly contentious. Given so much attention, one might expect that the species “problem” is solved. On the contrary, there are more species concepts vying for adoption today than there were a century ago, and the debate among competing concepts rages on. Even if biologists could agree on one concept of species, it would remain true that we have only begun the process of exploring the species of Earth. Estimates of the total number of species living today vary tremendously from a few million to as many as 100 million. Although our ignorance about biodiversity is due largely to inadequate support for taxonomists to discover and describe the world’s species, this work is impossible in the absence of an agreement regarding the concept of species to be applied.

Closely associated with the new synthesis was the biological species concept (BSC). The BSC was built on observations in the nineteenth century that sister species often lived in adjacent but separate areas, suggesting that allopatry was important in some way to species formation, particularly in animals. These ideas, combined with a shifting emphasis in biology to population genetic questions, made Ernst Mayr’s persistent advocacy for the BSC extremely effective among zoologists before they were set aside in favor of the BSC, although the number of studies providing the kind of interbreeding information required by the concept are few. Botanists never accepted the BSC in large numbers, in part due to the incredibly diverse and complex genetic mechanisms in angiosperms and rampant polyploidy in pterydophytes.

With some level of discontent already in place, Hennig’s writings forced a deeper consideration of species in the context of phylogenetic theory. Hennig pointed out that the BSC was at odds with evolutionary history since there were no clear breaks in the potential to interbreed among populations through geologic time. Projecting breeding patterns backwards through the geological record, there were no obvious places to demarcate one species from another. Hennig provided a fix for this dilemma and advocated a concept that in many respects resembled the BSC.

Donn Rosen noted that ancestral interbreeding, relative to extant populations, was plesiomorphic and therefore of little consequence to phylogeny; it was the loss of interbreeding that was of importance but which is indistinguishable from the absence of interbreeding. This, combined with a sincere desire to apply Hennigian theory to the species problem, led Rosen to develop an alternative species concept that attempted to apply cladistic analysis to populations. Rosen sought to ensure that species had novel status by making them equal to the smallest demonstrably autapomorphic units. Rosen’s was the first concept explicitly couched in terms of phylogenetic theory, but it suffered from several problems: Ancestors were impossible to positively recognize given this concept, even though they clearly had existed. When two or more species arose from one polymorphic ancestral population, it was not clear that one or the other daughter was more or less apomorphic than the other. Also, this use of the idea of monophyly was at odds in logic and intent with Hennig’s theories that dealt explicitly with supraspecific groupings. Despite these evident problems, some authors still advocated autapomorphic species.

George Gaylord Simpson developed an evolutionary species concept based on his view as a vertebrate paleontologist. His concept was revised and expanded by E. D. Wolfe nearly 30 years later, bringing the arguments for evolutionary species explicitly in line with phylogenetic theory. This theory has no overt conflict with either phylogenetic theory or known evolutionary processes, but it is not clear how this concept is put into practice in an empirical sense.

About 20 years ago, a second generation of phylogenetic species concept emerged that was independent of cladistic analysis (so that it could provide the elements of phylogeny to be analyzed prior to such an analysis) and fully compatible with phylogenetic theory. Working simultaneously, two pairs of authors produced nearly identically worded versions of such a concept: Eldredge and Cracraft (1980) and Nelson and Platnick (1981). The phylogenetic species concept is very similar to the morphological concepts in broad use by taxonomists before they were set aside in favor of the BSC, but it is formulated specifically in a phylogenetic framework. According to the phylogenetic species concept, a species is simply the smallest aggregation of (sexual) populations or (asexual) lineages diagnosable by a unique combination of character states.

Debate continues regarding the best alternative. This battle over species concepts is far from a mere academic exercise (Table I). Any effort to inventory or document biodiversity, to compare the biodiversity present in one habitat or geographic place versus another, and any wording in regulations aimed at natural resource management or biodiversity conservation will inevitably be
and made urgent by the biodiversity crisis. \textcolor{red}{gized by recent theoretical and technological advances to three traditional core missions that are ener-

IV. THE MISSIONS OF SYSTEMATICS

A. Mission 1: To Discover, Describe, and Inventory Earth’s Species

This mission seeks to answer several of the most fundamental of all biodiversity questions, such as “What and how many species exist?” “How are they related?” “What properties do they share with other species?” and “Where do they occur?” In slightly more than two centuries, taxonomists have documented nearly 2 million species and today estimate that the total number of living species is between approximately 10 and 100 million species. The completion of a project begun in the eighteenth century by Linnaeus to make an inventory and description of every species on Earth is now complicated by projections of the extinction of thousands or millions of species. E. O. Wilson estimated that just the minimal description of a flora and fauna with 10 million species would require the lifetime efforts of an army of taxonomists numbering 25,000.

Deceptively simple in its questions, this taxonomic inventory is in itself “big science” on a scale that rivals the impressive ambitions of astronomers and physicists. Stated another way, the magnitude of this project becomes clear: the complete exploration, characterization, analysis, and documentation of every life-form on an entire planet. Taxonomy is rarely conceived on a planetary scale, but the evolutionary history that provides its conceptual framework requires approaching the task on a global basis. Earth is the only known planet on which biodiversity exists with certainty. This biodiversity inventory is in itself “big science” on a scale that rivals the impressive ambitions of astronomers and physicists.

Although emerging technologies to manage this immense volume of data promise to facilitate the process, credible taxonomic work rests on rigorous theories and methods, worldwide collections of specimens, and the accumulated knowledge of specialists.

In a sense, every species is a part of a larger entity, and every group of species is nested within a larger assemblage.\textcolor{red}{tary and description of every species on Earth is now

Four centuries, taxonomists have documented nearly 2 million species. The completion of a project begun in the eighteenth century by Linnaeus to make an inventory and description of every species on Earth is now complicated by projections of the extinction of thousands or millions of species. E. O. Wilson estimated that just the minimal description of a flora and fauna with 10 million species would require the lifetime efforts of an army of taxonomists numbering 25,000.

Deceptively simple in its questions, this taxonomic inventory is in itself “big science” on a scale that rivals the impressive ambitions of astronomers and physicists. Stated another way, the magnitude of this project becomes clear: the complete exploration, characterization, analysis, and documentation of every life-form on an entire planet. Taxonomy is rarely conceived on a planetary scale, but the evolutionary history that provides its conceptual framework requires approaching the task on a global basis. Earth is the only known planet on which biodiversity exists with certainty. This biodiversity inventory is in itself “big science” on a scale that rivals the impressive ambitions of astronomers and physicists.

Although emerging technologies to manage this immense volume of data promise to facilitate the process, credible taxonomic work rests on rigorous theories and methods, worldwide collections of specimens, and the accumulated knowledge of specialists.

In a sense, every species is a part of a larger entity, and every group of species is nested within a larger assemblage.\textcolor{red}{tary and description of every species on Earth is now

Two models for taxonomic biodiversity inventories have been proposed. The All-Taxon Biodiversity Inventory (ATBI) seeks to enumerate every species living in a manageable-sized study site, e.g., 50–100 ha. The intent of the ATBI, conceived by Daniel Janzen, is to provide readily accessed identification information so that land and resource managers can make appropriate decisions and so that wild areas can be economically and socially valued possessions. This model is structured to

\textcolor{red}{TABLE I

<table>
<thead>
<tr>
<th>Alternative Species Concepts*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological species concept (Mayr, 2000, p. 17): “groups of inter-breeding natural populations that are reproductively isolated from other such groups.”</td>
</tr>
<tr>
<td>Hemigene species concept (Meier and Willmann, 2000, p. 33): “reproductively isolated natural populations or groups of natural populations. They originate via the dissolution of the stem species in a speciation event and cease to exist through either extinction or speciation.”</td>
</tr>
<tr>
<td>Autapomorphic species concept (Mohler and Thorton, 2000, p. 44): “the least inclusive taxon recognized in a formal phylogenetic classification grouped into species because of evidence of monophyly” (see original for more extensive wording).</td>
</tr>
<tr>
<td>Evolutionary species concept (Wiley and Mayden, 2000, p. 73): “an entity composed of organisms that maintains its identity from other such entities through time and over space and that has its own independent evolutionary fate and historical tendencies.”</td>
</tr>
<tr>
<td>Phylogenetic species concept (Wheeler and Platnick, 2000, p. 98): “the smallest aggregation of (sexual) populations or (asexual) lineages diagnosable by a unique combination of character states.”</td>
</tr>
</tbody>
</table>

* Citations from Wheeler and Meier (2000).
facilitate long-term ecological, economic, and conservation work. These undeniable laudable goals, however, may be better achieved through an approach structured to take full advantage of taxonomic research, collections, and expertise.

A taxonomic alternative has been suggested: The All-Biota Taxon Inventory (ABTI) simply selects a taxon or monophyletic group as the object of study rather than a place and seeks to document all species of that group regardless of where they live. Advantages of the ABTI, proposed by Q. D. Wheeler, relate to both efficiency and science. The end product of an ABTI is monographic knowledge of all species of a taxon and a predictive (phylogenetic) classification for the group. Within this conceptual framework, ecologists and resource managers anywhere in the world can access what is known of these organisms, predicting what properties new or poorly known species are most likely to possess. Such predictivity, the hallmark of science, is lacking in the site-specific alternative. Furthermore, every aspect of taxonomy is made more difficult through the ATBI model. Species limits and nomenclatural decisions are equivocal when divorced from knowledge of species occurring outside the study area. Classifications are denied access to extralimital species that can make character coding and cladograms more reliable. Also, detailed knowledge and experience of the taxon specialist and synoptic collections are neither used nor built to their full capacity. Recall the unique comparative and historical requirements of taxonomy that cannot be met at single study sites. Taxonomic knowledge is complementary to ecological knowledge, but it is most efficiently and effectively derived from a very different approach to fieldwork and inventory. Historically, much of the time of taxonomists learning a particular group has been occupied with correcting mistakes that resulted from geographically isolated descriptions of species.

Ignorance about the nature of taxonomy and its attendant philosophy has contributed to an academic version of political correctness that pressures taxonomy to conform to the standards of other disciplines. Ecologists often do not understand the value of general collecting or trapping of specimens and insist that "modern" standards be applied that can be used to quantify data. In order to answer taxonomic questions, however, it is not necessary to know how many individuals occur per square meter or to gather data on relative humidity or feeding habits. Were taxonomists to uniformly adopt such data collection standards, they would cease to do taxonomy and spend much of their time doing second-rate ecology. Even extremely crude data can sometimes answer important questions. A primitive tribe of leiodid beetles (Neopeleatopini) has been found in Chile, Argentina, and Australia. Does this clade show the often repeated southern temperate pattern of distribution? If so, we would expect one or more undescribed species to be found in South Africa. A neopeleatopine with a label as vague as "South Africa" would be sufficient to answer this question. Additional information is desirable, of course, but must be weighed in the context of its cost to limited time and expertise.

If you are unconvinced, reverse the roles. Good ecology requires a great deal of work at the same study site. For the taxonomist, however, the goal is to see examples of a taxon from as many parts of its range as possible. Were taxonomists to insist that ecologists spend no more than 1 week at any particular site and then move on to another, I can only imagine the indignant outrage. If we are to understand biodiversity, it is critical to support all the biodiversity sciences. These fields are complementary to one another. Knowledge from all are needed for a complete understanding, but the best and most cost-effective data will come from each when its unique needs are met.

B. Mission 2: To Analyze Phylogeny and Predictively Classify Species

Despite the fact that Charles Darwin and taxonomists who came after him all recognized the importance of having classifications reflect phylogeny, it would be a century before Hennig gave us the means to do so in a rigorous and testable way. Confronted with the urgent need for taxonomic knowledge in the face of the biodiversity crisis, we are fortunate to have phylogenetic systematic theory, methods, and computer algorithms. As the century between The Origin and Phylogenetic Systematics attests, such theoretical advances could not be produced at will. With this necessity met, however, the taxonomic community is poised to meet the need for predictive classifications. Remaining obstacles are very practical ones related primarily to a shortage of taxonomic specialists, insufficient funds for their work, and inadequate resources for the maintenance and growth of natural history collections. In other words, we know what needs to be done and have the scientific foundation to do it; we need only place a sufficiently high priority on taxonomic revisions and monographs to get on with the analysis of phylogenetic relationships and the production of optimally useful and predictive classifications.
C. Mission 3: To Disseminate Data, Information, and Knowledge

Because good taxonomy is omnispective with regard to members of higher taxa, it produces a large volume of data, information, and, ultimately, knowledge. It can be critically important to know what labels are placed on a type specimen and the geographic location of every specimen studied in a monograph, not infrequently numbered in the thousands or even tens of thousands—raw data that can occupy many expensive pages in publications. The advent of computer technology and particularly the Internet has the potential to revolutionize the way in which and the speed with which these data are made accessible to the community.

In our rush to embrace this technology, many erroneous records have been typed into computer databases and foisted upon the world. The costs of identifying and correcting this disinformation will be immense, undoubtedly more than the cost of obtaining the correct information the first time. For relatively well-known taxa, it is appropriate to obtain existing museum data on the Internet in a searchable format. For poorly known taxa, such as most arthropods, fungi, microbes, and many plants, the blind typing of questionable data from museum specimens seems of questionable worth. As a minimum, collections need to capture data as new specimens are accessioned and, even more important, to capture taxonomic information as taxonomists return borrowed specimens following the completion of a monograph or revision. In large groups in which particular taxa are revisited at most once per century, such data should be obtained when they are revised because these data are as good as they will likely be for the foreseeable future.

Credible taxonomic data are closely associated with specimens deposited in collections. As a result, there are considerable benefits to the maintenance of both collections and databases at a location where there are also scientists with taxon-specific knowledge and appropriate libraries. Because taxonomists ask geographically and temporally broad questions, reviewing an entire clade from its origin to the present anywhere on the earth that it has lived or currently exists, a typical study deals with thousands, often tens of thousands, of specimens and an enormous number of bits of data.

For a few taxa, it is possible for anyone sufficiently interested to quickly become adept at species identification. As a result, thousands of important records about the distribution, migration, and nesting of birds are accumulated in the absence of voucher specimens. For most taxa, and especially the largest ones (e.g., arthropods and fungi), such is not the case. Identifications are often only as believable as the credentials of the identifier are known. Major genera of insects are revised no more frequently than once per century, suggesting that at any given moment there may be no more than one or a few people who are fully competent to make an identification. Even such highly experienced specialists may require access to identified sets of specimens to verify their identifications.

The emergence of high-speed, high-capacity computers makes the dissemination and maintenance of biodiversity data at least feasible. It is possible to facilitate the exchange of information internationally on a time-scale never before dreamed of. Also, information can be shared in such a manner that each nation need not support an expert and a collection for every taxon that may be of importance to its agriculture or commerce. Traditional ‘products’ of taxonomy can be done more rapidly and cost-effectively, and entire new kinds of output of data and knowledge are possible. The following are among the products of taxonomy:

Reference specimens. In many respects, the most important output of taxonomic research is in the form of authority-identified specimens. These specimens provide a permanent, three-dimensional, character-rich record of the concept of particular scientists with regard to the status of species and clades. These specimens in museums and herbaria are of daily value in accurate identification work and of long-term, historical importance in taxonomic knowledge dissemination. In a single writing, one finds a review of essentially all that is known of a taxon (minimally an exhaustive review of every species ever thought to belong in it) and available new species and associated (geographic, ecological, etc.) information. A modern monograph includes a classification, expressed in carefully documented Linnaean names, based on the best corroborated existing cladistic hypothesis. Monographs have the decided advantage of taking into consideration every known member of a taxon at once, inevitably arriving at a better classification (or at least one that is better documented as such). Once a monograph is completed, it can be used for an incredibly diverse range of purposes and many more specific publications may be derived from it. For example, one might extract a checklist of the species of a particular country, generate a diagnostic key for their identification, or interpret some unusual biological attribute of an organism in its broadest evolutionary (historical) context. With an exiting monograph as a baseline, one can more confidently identify species from anywhere in the world and add new species descriptions.
or distribution records with great confidence. Even when a provincial study is done well (e.g., the species of a taxon for Mexico), there are equivocal species limits and nomenclatural uncertainties that can only be addressed when all included subgroups are included in the study regardless of where (or when) on Earth they occur.

Databases: Computer databases are an essential tool in dealing with billions of facts currently stored on labels attached to millions of specimens. These data, however, are of only great value when they are associated with a voucher specimen or specimens permanently deposited in a natural history collection. Also, the data rapidly degrade in importance and credibility if they are not routinely examined by taxonomists with knowledge of that taxon. As revisions are done, names changed, and so forth, it is essential that databases be brought up-to-date. This, however, is not always easy for the nonspecialist. Raw taxonomic data are of limited value. Far more important are the interpretive works of taxonomists that use such data to make both printed and electronic works which transmit instead information or knowledge.

Special-purpose publications: From a monographic or revisional study—a taxonomic study comprehensive enough to study all known and new members of a group—it is possible to extract less comprehensive works for special needs and purposes. For example, it will soon be possible to extract maps, checklists, or diagnostic keys for only those species of a particular country, park, or ecosystem from an electronic monograph.

Specialists: New generations of taxonomists are most easily and efficiently trained when students can have an established specialist as a mentor. An alarming number of taxa, some quite large and difficult to master, have few or no living authorities who can mentor a new generation of specialists.

Predictive classifications: As discussed previously, predictive classifications are perhaps the most useful of all taxonomic products.

V. ROLE OF TAXONOMY IN BIODIVERSITY STUDIES AND CONSERVATION

The significance of taxonomic knowledge in the study and conservation of biological diversity has not been sufficiently recognized nor supported. Recognition of the biodiversity crisis virtually requires a major shift in the way in which we think about biodiversity exploration and conservation. It is no longer (if it ever was) realistic to set out to maintain the status quo. Even the most optimistic among us no longer project a rosy future for all species. The prime question for conservationists is no longer “How can we save species threatened by extinction?” but rather “Which species shall we attempt to save?” Because every species, habitat, and ecosystem is unique in one or more valuable ways, it is difficult to prioritize conservation decisions based on ecological interactions or population genetic dynamics, even though success in any particular conservation effort depends heavily on the theory and practice of both ecology and genetics. One could crassly rank ecosystems according to their relative productivity or cash crop monetary value to current world economies, but this arbitrarily minimizes many valuable components of biodiversity and is based on an unwarranted arrogance that we know enough about biodiversity to predict which pieces are indispensable.

A more prudent approach might be to maximize the planet’s biodiversity as measured in species and their relative contribution to overall biological diversity (i.e., an assessment of their unique contributions to clade diversity). Since choices among various clades and species seem increasingly inevitable, the obvious taxonomic question is whether we can succeed in conserving evidence of as many clades of the tree of life as possible along with all the benefits that would go with such a crosscut of biodiversity. As efforts proceed to understand and conserve biodiversity, regardless of specific goals, taxonomy will contribute in many ways.

A. Conceptual Framework

Taxonomy gives to biodiversity a historic framework within which every aspect of life on Earth may be understood and interpreted. Regardless of spatial, ecological, or genetic complexities, the only thing that every species on Earth shares in common is its organic history of diversification. Although it is essential to understand biotic and abiotic ecological parameters affecting spatial and seasonal distributions, it is no less important to understand the contribution of shared evolutionary history in the formation of observable current or past patterns of molecular, morphological, ecological, or geographic similarity.

B. Language of Biodiversity

Taxonomy’s plant, animal, and microbial nomenclatures provide the language for communication about biodiversity. Because Linnaean names do not fully con-
very what we know about phylogeny, various conventions have been proposed to supplement its cladistic information content. Even before such augmentation, however, the Linnaean language of taxonomy is a powerful one. If I mention two staphylinoid families, for example, Leiodidae and Agyrtidae, you know with these three words several things. For example, the two families belong to the same superfamily (Staphylinoidea) and therefore have common remote ancestry, and they are two distinct groups (i.e., because each is ranked as a family, discernible by the “idae” ending, you may conclude that one is not subsumed within the other). Because taxonomists have worked diligently for a long time to make taxa monophyletic, you may assume also that the scientist who proposed this classification believed that all leiodids are more closely related to one another than any of them are to species of agyrtids.

C. Information Storage and Retrieval
Handling billions of pieces of data about millions of species is no small job. Although countless artificial, special-purpose classifications could be conceived (e.g., all fungi fruiting above 2725 m, all herbivores in forests, or all organisms that are green), the one thing shared in common by all forms of life on Earth is an evolutionary history or phylogeny. For this and more technical reasons, it may be stated that cladistic classifications offer the optimal available system in which to store or retrieve what is known of every aspect of biodiversity, the most logical general reference system for all biologists, and the best basis from which to predict what attributes species are likely to possess that have not yet been studied in sufficient detail.

D. Scale
Taxonomy provides a view of biodiversity on a scale not replicated by any other discipline. Many of taxonomy’s questions are binary: Does species X occur in region A or not? Answers to such fundamental questions have far-reaching implications but require data that may not appear particularly sophisticated or detailed. A nineteenth-century survey performed by John LeConte for a railroad may record a particular species of beetle from “Lake Superior.” Given some detective work, we might be able to reconstruct in greater detail where Dr. LeConte was working on a particular date, but the data would never pass for anything resembling standards for modern ecology. For a mostly tropical genus such as Aglyptinus, however, such data may be sufficient for these kinds of basic inquisitions. If we wish to determine whether the genus Aglyptinus occurs in North America or perhaps in Michigan, LeConte’s data may be quite sufficient. For an ecologist asking about population density, it is obviously of no more use than a marker perhaps of a vague geographic area within which to begin to take appropriately quantified samples. There is a mistaken belief, however, that taxonomists’ samples should conform to ecological standards. The argument goes more or less as follows: By carrying around a square-meter devise, the taxonomist could easily quantify his or her data and thereby make it useful to many scientists beyond taxonomy. Were this proposal without hidden costs, this argument would clearly prevail. In reality, any rigid sampling suitable to meet the constraints of serious ecological experimentation does incur substantial costs to taxonomy.

E. Evidential Record
The natural history museums (and botanical gardens and other institutions that care for collections) play a role in the documentation of biodiversity that can scarcely be overstated. Conservation organizations that have created databases for threatened species often include large numbers of observations not verifiable by reference to museum specimens. Such records cannot be critically questioned or trusted and are of far less value to science or society than those records that are evidenced in the form of permanently preserved museum specimens.

F. International Standards
Type specimens in museums and herbaria are the equivalent of an international bureau of standards for scientific names. The word “type” is sometimes naively used to vilify typification as a modern form of typology (in the Aristotelian sense). In fact, the type is used pragmatically to ground the name to observable features of one representative. Species in taxonomy complex hypotheses and take into account ideas about a wide range of genetic, geographic, and morphological variation. When two or more species are hypothesized to exist where a single one did before, the type merely ties the name of the existing species to whichever of the two that the type specimen falls within. No one presumes that another representative will necessarily match the type in detail, except for those characters hypothesized to be diagnostic for the species. The insistence on types by the international codes of nomenclature has done much to establish and maintain stability in what is intended by species names.
G. Record Truthing

Information published in scientific journals or released in the context of a database is only credible if the identity of the species may be verified. In any taxon, especially in megadiverse taxa such as insects, a species that is believed to be well-known today may suddenly prove to be misidentified from some other geographic region, to be part of a complex of previously undistinguished sibling species, to be synonymous with a comparatively senior or junior name, or otherwise nominally changed in status. The only way to guard against such future uncertainties is to deposit a voucher specimen in a museum for permanent housing. In instances in which this has been done, it is possible to later revisit the question of identity and confirm or correct what was first reported in the literature. Literally thousands of scientific journal article are brought into question or must be summarily rejected due to the absence of a voucher. Despite this fact, few journals require such deposition of their authors.

H. Predictivity

Given a cladogram, it is possible to make predictions about the distribution of properties of organisms not yet studied. Even the simplest three-taxon statement makes a bold prediction. If A and B are hypothesized to share a more recent common ancestor than either one does with C, then the implicit prediction is that every apomorphy discovered in the future will be so distributed. The power of such predictivity when biologists are confronted with tens of millions of species can hardly be overstated. Where does a pharmacologist begin to look for the same or related chemical compounds found serendipitously in one species? The answer is obvious given a cladogram.

I. Authoritative Basis for Biodiversity Information

Although a taxonomic monograph may be too exhaustive or too technical for every user, it provides the authoritative comparative overview in which we may have confidence about species numbers and limits. As computer-based monography matures, it will soon be possible to computer generate user-friendly field guides, checklists, and other documents directly from the monographic database but limited to those species and that information necessary for a special purpose.

J. Efficient Inventories of Taxa

Taxon-specific knowledge and experience held by taxon specialists make the inventory of the species of a taxon more efficacious than alternative approaches to determine what and how many species live on Earth.

K. Epistemic Privilege

Taxonomists have a unique view of the world, necessitated by the comparative, historical, and parsimony-based nature of their science. Experimental biologists whose work does not depend on (and is not answerable by reference to) collections of specimens can hardly be expected to take the lead in building vast herbaria and museums. Not can they be expected to place a major emphasis on the inventory of species and clades so close to the brink of extinction that only (not forthcoming) Herculean efforts could pull them back from the brink—those taxa referred to as the “living dead.” This vision of the importance of phylogenetic hypotheses and taxonomic inventories places a strong burden on the taxonomic community to ensure that these essential aspects of biodiversity study are continued.

Acknowledgments

I thank Roberto Keller for a detailed reading of the completed manuscript, an illuminating argument over metaphysics, and assistance in the preparation of the figures. I also thank Nico Franz, Kelly Miller, and Kevin Nixon for reading and commenting on the manuscript.

See Also the Following Articles

BIODIVERSITY, EVOLUTION AND CLADOGENESIS • DARWIN, CHARLES • NOMENCLATURE, SYSTEMS OF • PHYLOGENY • SPECIES, CONCEPTS OF • SPECIES DIVERSITY, OVERVIEW

Bibliography

TAXONOMY,
METHODS OF

R. I. Vane-Wright
The Natural History Museum London

1. The Tasks of Taxonomy
2. Building Blocks: Individuals and Characters
3. Special and General Classifications
4. Differing Philosophies and Methods of Taxonomy
5. General Procedures
6. From System to Classification
7. Current Practice: Variations on a Cladistic Theme
8. Conclusions

GLOSSARY

clad A complete ancestor–descendant lineage (cf. monophyletic group).
cladistics Production of taxonomic system based on hierarchical patterns of homologous characters, expressed as cladograms.
cladogram A dendrogram expressing estimated cladistic relationships among taxa; a cladogram has no direct connotation of ancestry and the long axis does not connote time.
dendrogram A branching, nonreticulate diagram expressing nested hierarchical relationships or similarities (or both) between entities (e.g., taxa) such that the entities only appear at the tips of terminal branches.
diagnosis A set of attributes sufficient to define, characterize, or identify a given taxonomic group.
distance A measure of dissimilarity between two taxa based (normally) on the number of mismatches within a large set of characters or attributes compared for both taxa.
gap A large or relatively large difference in overall similarity between two taxa.
grade A group based on, for example, the functional level of organization or overall similarity rather than ancestor–descendant relationships.
ground plan (archetype, bauplan) A basic plan or general type, or a hypothetical ancestor.

jizz Characteristic, instantly recognizable appearance of an organism.

monophyletic group A group comprising a given (hypothetical) ancestor and all its descendants; in a restricted sense, within a given set of terminal taxa, all the members of a subset arising from a common ancestor that has given rise to no other member(s) of the whole set.

paraphyletic group A group comprising a given (hypothetical) ancestor and only some of its descendants.

phenetics Production of a taxonomic system based on overall similarity.

phenogram A dendrogram expressing overall similarity (long axis) between terminal taxa and sequentially linked groups of terminal taxa; the nodes do not connote specifiable characters and the long axis does not connote time.
polyphyletic group A group that does not include the most recent common ancestor of all of its members.

polythetic group A natural taxonomic group in which the terminal taxa are not known to share any universal character(s) but are nonetheless united by overall similarity (phenetics) or global parsimony (cladistics).

rank A specified categorical level in the taxonomic hierarchy (e.g., species, genus, family, and class) made coordinate in classification by definition but very frequently not coordinate in the taxonomic system.

similarity A measure of coincidence (matches) among a (large) set of characters compared for any two taxa.

sister groups Two terminal taxa or monophyletic groups that share a common ancestor that has not given rise to any other taxon under consideration.

taxon Any formally named or recognizable group in a taxonomic system (e.g., order, family, and genus), including all the particular terminal taxa (species).

Taxonomy is a highly controversial subject, and the issues are inextricably bound up with philosophical disputes which have endured for centuries. The problems are so important that no biologist can totally avoid facing them.—Michael T. Ghiselin (The Triumph of the Darwinian Method, 1969, p. 79)

No one would think much of a chemist who confused water and benzene "because they look alike."—Arthur J. Cain (Animal Species and Their Evolution, 1954, p. 11)

TAXONOMY plays the central role in the scientific discipline of systematics and, according to some authors, the two are largely or even entirely synonymous. Systematics makes an essential contribution to our understanding of biological diversity, including its origins, distribution, and maintenance. The primary task of taxonomy is systematization: to establish and give an account of biological order among the diversity of organisms. This involves enumerating the kinds of living things that exist and have existed in the past and determining the patterns of difference and connection among them. By giving expression to these patterns through naming inclusive sets, subordinate sets and least included entities (e.g., higher classes, genera, species, subspecies, and varieties), taxonomists have produced the general biological classification: a categorical arrangement of named, diagnosable groups and components (taxa) that can be used to refer to all known living and extinct organisms.

Although effective taxonomy predates Darwinism, systematics is now closely linked to the theory of evolution. Descent with modification gives justification to the general form of classification adopted: hierarchical, nonoverlapping sets rather than fuzzy sets or periodic tables. The fundamental methods of taxonomic systematization are few. Ultimately, there may prove to be only two—overall similarity methods (grouping by degree of genetic or phenetic similarity) and hierarchical methods (grouping by inclusive phylogenetic or cladistic relationship)—but the theoretical bases of taxonomy are various. This has led to a proliferation of approaches, often with variants. To understand alternative methods of taxonomy it is necessary to appreciate the philosophical differences between these approaches, and much of this article is devoted to exploring these differences rather than describing techniques. The methods of cladistics, currently the dominant approach, are the subject of a separate review.

I. THE TASKS OF TAXONOMY

Taxonomists perform five main functions: discrimination (discovery or primary recognition of taxa; also entailing formal description and diagnosis); comparison (assessing similarities, differences, and relationships among taxa); classification (production of summary schemes that encapsulate current knowledge of taxa and their main interrelationships); symbolization (application of names to taxa and classes: technical nomenclature); and identification (secondary recognition of taxa matching unidentified material to the established system).

This article focuses on different approaches to primary discrimination and comparison of taxa (systematization) and the ways in which this knowledge can be expressed as summary schemes (classification). The procedures of taxonomy as an information retrieval system, which includes making checklists, catalogs, and databases, identification (naming specimens), and technical nomenclature (application and regulation of names: naming taxa), are not dealt with here. Nor, explicitly, are the means of gathering taxonomic data, e.g., from anatomy, karyology, biochemistry, and nucleic acid chemistry, and their corresponding subdisciplines (morphotaxonomy, cytotaxonomy, chemotaxonomy, molecular systematics, etc.). Although such subdisciplines are often described as taxonomic methods, in the wider context of systematics and taxonomy
as a whole they are subsidiary techniques. Much the same methods and problems of interpretation apply, whatever the source of empirical data. This also applies to experimental taxonomy (the investigation of taxa based on predictions from alternative systems in an attempt to gather data bearing directly on particular taxonomic problems) and biometrics (quantitative comparison of related taxa).

Disagreements regarding taxonomic methods can lead to major differences in classification. Such discrepancies range from disputes over the validity of species, subspecies, and infrasubspecifics at one end of the scale to the extreme opposite where composition of even the most inclusive categories (domains and their component kingdoms) remains uncertain. By classifying the classifiers it may be possible to identify some of the fundamental reasons, but not all taxonomic disagreements are due to method alone because historical precedent and subjectivity still intervene. Moreover, in practice a great deal of constructive taxonomic work is done with little reference to philosophy or explicit method, being achieved by pragmatic intervention, notably the extension or modification of existing parts of the system (e.g., by description of new species and establishment of synonymy).

II. BUILDING BLOCKS: INDIVIDUALS AND CHARACTERS

Taxonomy is an empirical activity. On the basis of characters derived from sensory data, individual organisms or life cycles can be discriminated and then divided among or gathered into groups, and groups within groups, and so on to produce a taxonomic system. This can be done either top-down (divisive methods, as in Linnaean divisions) or bottom-up (agglomerative methods, as in most numerical taxonomy). Although individual organisms and characters are the basic elements in this process, their definitions are not straightforward.

A. Individual Organisms

Organisms multiply, typically from spores or fertilized eggs, to reproduce the corresponding parental stage through a cycle of growth and differentiation. At any point along such an ontogenetic pathway, a particular organism can be referred to as an individual: fertilized egg, embryo, larva, subadult, adult, or postreproductive adult. However, some organisms multiply by splitting or propagation, rendering the parent–offspring distinction uncertain. Thus, land plants can spread by “runners” so that what appears to be a field of separate individuals can be identical clones of one original plant. In the case of social organisms (e.g., ants and termites) in which very large numbers of workers, guards, or other castes may be produced that never have an opportunity to reproduce, the entire colony has some of the properties of an individual. These distinctions are important because the reliability of the taxonomic process, other factors being equal, is dependent on the number of individual organisms sampled as well as the number of characters recognized and recorded. Multiple sampling from essentially the same individual organism can distort our views of natural variation or lack of it and lead to erroneous conclusions (Wiens, 1999).

B. Attributes and Characters

Taxonomists compare organisms by means of characters. Characters are abstractions derived from the detectable attributes of individual organisms or social groups (e.g., “large, two pointed prongs on head, color vision, always herd with tails erect”). To be informative, it is obvious that characters observed must not be universal throughout the organisms under investigation. The distribution of characters is the primary interest, and the degree to which they differentiate, coincide, and conflict will largely determine their usefulness for taxonomy.

In general, characters can be divided into two sorts: continuous (e.g., height, from tall to short) and discrete (e.g., paired horns versus no horns). However, this is not absolute but more a matter of scale. Thus, different individuals of a particular kind of fly might bear every possible combination of 1–20 spines on the thorax: Is this continuous or discrete variation? In addition, coding can be arbitrary (how tall is “tall”?), and subdivision can be ambiguous. If we observe different individuals with straight, curved, and forked horns, do we simply regard these as three separate characters, three states of one character (form of horn), or the intersection of two binary alternatives—horns straight/curved and horns simple/forked (with the expectation of being able to distinguish, in theory at least, two sorts of forked horns)? Moreover, there are other possibilities for coding such variables. The decision we make can affect analysis (e.g., by encoding spurious information derived from inapplicable characters). Thus, if an animal lacks horns, and this is coded as three separate pieces of data (not forked, not curved, and not straight), such redundancy can have undesirable effects on analysis (Strong and Lipscomb, 1999).
In theory, unit characters must not only be nonredundant but also homologous and independent (Pimentel and Riggs, 1987). If different parts of the body are functionally interdependent or developmentally linked, it will be misleading to count these attributes as separate characters. This can occur when a single gene has a pleiotropic effect, influencing, for example, the color of one organ and the form of another. At the extreme, it is clear that multiple characters should not be created by logical correlation (e.g., treating both the circumference and the diameter of a circular organ as two unit characters). Some organs (e.g., parts of the male genitalia of insects) may appear so complex that counting them as one character seems unreasonable, but any stopping rule for subdivision may be uncertain, whereas other characters that appear simple may prove to be complex (e.g., a single functional bone formed by fusion during development, representing several characters in comparison to related taxa in which the equivalent bones have not fused).

III. SPECIAL AND GENERAL CLASSIFICATIONS

On the basis of characters held in common, individual organisms can be grouped into a large number of classes, which are of two general kinds. On the one hand, individuals can be grouped in terms of a particular attribute (e.g., green, round, four-legged, marine, planktonic, nocturnal, and pollinating) or by small combinations of attributes to give prescriptions such as “plankton-feeding marine organisms” or “nonflowering, epiphytic semiparasitic plants.” Alternatively, they can be placed into categories of species, genera, families, orders, and so on. The former are regarded as artificial or special classes and are generally defined by reference to given attributes, whereas the latter are viewed, ideally, as natural kinds or natural groups which are discovered but cannot be defined a priori (although they can be diagnosed a posteriori).

Special classes frequently overlap, such as the overlay of plankton feeders, not all of which are marine, and marine organisms, not all of which feed on plankton, to define the special class discussed previously; also, the same organism may recur in many different special classifications (e.g., eagle as flying organism, predator, or nest builder and in the conjunction of all three). In contrast, natural groups typically form nested, nonoverlapping sets in which each kind of organism or group only appears once. Thus, within the general classification of birds, all eagles are included as members of the family Accipitridae, which also includes vultures, buzzards, hawks, and kites.

A. Evolution, Genetic Relationships, and the Natural Hierarchy

Natural groups comprise individuals with many attributes in common, whereas individuals belonging to special classes have relatively few shared characters. In practice, the difference between, for example, “large marine animals” and Cetacea is simply that individuals of the latter natural class have far more in common than those of the former. Even so, the word “natural” has had many connotations in the context of taxonomy, including the essence of things classified (Aristotle), rationality (e.g., as in God’s design), similarity, and explanatory power (Gilmour, 1940). Biology, however, has a unique theory that underpins its totality: the theory of organic evolution.

Ideas about evolution can be divided into a general theory of descent with modification and special theories about the processes affecting that descent (e.g., orthogenesis; natural, group, kin, sexual, and species selection; adaptive radiation; molecular drive; and molecular clocks). Most systematists agree that the general theory of evolution not only provides a compelling justification for seeking one natural, general classification of living organisms but also suggests the basis on which this classification is most securely founded. A hierarchical pattern of exclusive sets and inclusive subsets can reflect the primarily divergent sequence of ancestor-descendant relationships.

Hennig (1966) used the term “tokogenetic” relationships for the reticulate genealogical links that occur between parents and offspring within an evolving, sexually reproducing population. When a whole population splits to form two or more divergent subsystems within each of which tokogenetic relationships are maintained but between which processes of genetic recombination are largely or entirely discontinued, speciation (or phylogenesis) occurs. If we describe two extant taxa as phylogenetically more closely related to each other than to any third taxon (regarding them as sister taxa), we imply that the two do not share tokogenetic relations now but did so in the past within a common ancestor that they do not share with any other living taxon. Such ideas were basic to Hennig’s concept of phylogenetic systematics. However, many organisms reproduce without genetic recombination (sex) and thus lack tokogenetic relationships. Moreover, the great variety of sexual processes suggests that tokogenetic relationships are
not only nonuniversal but also may differ fundamentally in the many lineages where they do occur (Margulis and Sagan, 1984).

Species and other taxa grouped to reflect their phylogenetic relationships are expected, by virtue of their historical connections, to share far more characters in common than members of artificially formed groups. This expectation extends to unknown attributes, making phylogenetic classifications highly predictive. However, another type of generic relationship is increasingly recognized as important in organic evolution: lateral gene transfer. The chimerical nature of lichens as fungoidalgal symbionts has long been recognized. Following the work of Lynn Margulis, it is accepted that at least two of the cellular organelles found in all eukaryotes originated through symbiosis of fundamentally separate organisms. Evidence is accumulating from molecular phylogenetics that the genomes of many, perhaps all, major life-forms are chimeras formed from multiple original sources. The extent to which lateral gene transfer undermines current approaches to natural classification is uncertain, but at the domain level at least, a nonrepetitive hierarchy based exclusively on divergence appears to be unrealistic (Doolittle, 1999).

IV. DIFFERING PHILOSOPHIES AND METHODS OF TAXONOMY

Taxonomic methods have developed over time. The massive edifice of taxonomic classification, involving millions of terminal and higher taxa, on which the study, scientific use, and conservation of biodiversity depends has been built up over centuries. This system has been produced by thousands of different minds using different methods, working with different knowledge, under different influences, and often seeking different goals. Some parts have been revised and reworked repeatedly, others hardly at all. To understand the strengths, weaknesses, and limitations of the taxonomic system, it is necessary to appreciate the ways in which systematists of contrasting persuasions have sought order in nature and tried to reflect that order in biological classification. In the sections that follow, the various approaches and methods are reviewed in a historical sequence.

A. Essentialism, Idealism, and Prevolutionary Taxonomy

In his early writings, it is clear that Linnaeus, the founding father of modern taxonomy, was trying to detect the pattern of Creation in the classes and species that he recognized and named. Linnaeus' system was, superficially at least, very simple. Having divided all life-forms into animals and plants, these kingdoms were then divided successively, on the basis of one or a few defining characters, into a series of smaller units down to the level of the genus. Within each genus a large number of terminal species were usually recognized, and each was given a short, diagnostic description. For each species, he also accepted that nameable variations could occur but considered that these did not represent the fundamental plan of God's work.

It can be argued that Linnaeus started out as an essentialist, belonging to an intellectual tradition founded on Aristotle's methods of logical division (Linnaeus later modified his views considerably). Aristotle was perhaps more concerned with the classification of our knowledge about living things, or even the generation of knowledge through the process of classification, than he was with the classification of organisms as such. The young Linnaeus was trying to discover or reveal deep knowledge: the natural order of God's Creation. Through his search for an external criterion of verity, Linnaeus' concepts of species and higher taxa were fundamentally removed from those of the nominalists (see Section IV.C).

Essentialism and related but distinct ideas of the early Greek philosophers, notably Plato's idealism, continue to have influences on both the practice and understanding of taxonomy. According to the typological species concept, every species was thought to have its own idealized plan or design. The task of the taxonomist was then to recognize each of these theoretical designs, and describe, divine, or define the essential features of these "types" so that individual, real organisms could then be assigned to them. Thus, for a Platonist, "taxonomic names are the names, not of organisms, but of concepts" (Ghiselin, 1969). Even today, many taxonomists consider given species as concepts (e.g., fulfilling some ideal as gene pools, potential interbreeding units, or mate recognition systems) rather than empirical entities.

Idealism also had a great influence on concepts regarding major groups of organisms, especially animal phyla, which many morphologists believed could be formulated as a series of fundamentally different body plans (ground plans or bauplans). The notion of the ground plan was strongly developed by the ideal morphologists, who evidently affected the developing ideas of Willi Hennig (see Section IV.F) in the 1930s and 1940s, even though he later rejected many of their notions. According to Hull (1965), by supposedly rid-
B. Empirical and Inductive Taxonomic Methods

According to Mayr (1969), empiricism should be included as an atheoretical approach embracing the plausible, commonsense view that once sufficient knowledge has been gathered about organisms, a natural system of classification will simply emerge or become self-evident. Although it seems difficult to accord such a process the status of a general method, particular empirical observations can and often do render theoretical disputes irrelevant.

When Vaughan Thompson discovered in 1829 that barnacles develop from a nauplius larva, they were readily transferred from the Mollusca (where they had often been placed) to the Crustacea. Although this change in classification could in retrospect be justified by appeal to arguments about evolution, overall similarities, bauplans, or synapomorphies, to the empiricist this would all appear unnecessary. To raise a brood of insects and find that all the males belong to one genus and all the females to another leads to an instant appreciation of generic synonymy, without any need for appeal to theory.

The most complete expression of empirical classification occurs outside science in the form of folk taxonomy. Because folk taxonomy is well developed even in illiterate cultures, instead of inquiring into method or philosophy, it can only be understood by description and comparison. Berlin (1992) identified seven major features of ethnobiological categorization: recognition is given to the most distinctive local species; classification is based on “affinities that humans observe” (not on cultural significance); systems have a limited hierarchic structure; recognized taxa are distributed among a few mutually exclusive universal ranks, approximately equivalent to kingdom, life-form (e.g., plant), family, genus, species, and variety; in different ethnic systems, taxa at each rank show striking similarities regarding their number of subordinate taxa; taxa at generic and specific levels have an internal structure in which some included members are viewed as “prototypical” and others as reminiscent of other taxa at the same rank, giving rise to a type of fuzzy set classification; and a large majority of the taxa differentiated, most notably at the generic level, correspond to groups recognized in formal taxonomy.

These regularities and correspondences are highly suggestive that there is a natural biological classification that is almost literally self-evident and independent of the observer, and that a significant number of natural elements and groupings can be recognized simply through extensive knowledge and contact with nature. Although the young Linnaeus thought he was discovering the handiwork of God, it seems possible he was also involved in formalizing preexisting European ethnobiological knowledge.

Thus, as empirical knowledge accumulates, a common sense or consensus view of certain issues arises, and in some cases such a view may seem irrefutable. Even so, there are many problems in taxonomy that cannot be decided (e.g., questions of relative rank or the status of paraphyletic groups) without an explicit theoretical framework.

C. Nominalism

One view of taxonomy is that, even though we may desire a general system, there are no independent means for assessment. The only objective realties are individual organisms, and all taxonomic groups are man-made abstractions (“categories of thought” according to Louis Agassiz). Named groups are convenient for collective reference but have no independent basis separate from the human mind, simply being useful pigeonholes for dividing up or handling diversity. Many biologists have held similar views (apparently including Darwin, who once commented on species as a “term . . . arbitrarily given for the sake of convenience to a set of individuals closely resembling each other”; see Mayr, 1963, p. 14; also see Ghiselin, 1969).

By embracing the idea that taxonomic groups are established for convenience, nominalism has links with special classifications. If special classifications such as “trees” or “four-footed land vertebrates” are convenient, then there is no obvious reason why the nominalist should reject them, except insofar as it can be demonstrated that some comparable grouping has more heuristic or predictive power (e.g., trees extended to Tracheophyta and lizards Tetrapoda extended to include snakes, birds, whales, and bats); in other words, they are in some sense more convenient. Convenient for who then becomes the question.

Berlin's (1992) review suggests that folk taxonomies are not nominalist, which perhaps might have been expected, but seminatural and based on extensive em-
Folk systems often have terms for conspicuous natural groups such as Mammalia, but it is not clear to what extent paraphyletic or even polyphyletic taxa are also included. Because folk taxonomies have terms for life-forms, which are in effect grades, this suggests that humans may possess an innate mixed strategy for classification, sometimes grouping on pal-pable characters, whereas at other times forming groups on the basis of general resemblance, Gestalt, or jizz.

Although Panchen (1992) suggests that there are no extant practitioners of nominalism, the view that spe-cies are real while all higher taxonomic groups are artificial is widely expressed (even though quite mis-guided: contrast the difficulty of even "defining" many species of mice with the ease with which the class Mammalia can be recognized by many consistent fea-tures). There are major disagreements regarding the existence of fundamental differences between terminal taxonomic components (e.g., species) and higher groups (e.g., polytypic genera and families), as appar-ently accepted by Hennig, for example (and probably Mayr), and rejected by Gilmour (1940) and Nelson (1989).

D. Evolutionary Systematics:
Grades and Clades

The Darwinian revolution provided a rationale "exter-nal" to the human mind for the basis of taxonomy: Hierarchical relationships among living things are ex-plicable as the result of organic evolution, and taxono-mists should strive to reflect these patterns as the logical basis of a general classification. However, if taxonomists looked to Darwin for a more detailed lead, what did they find? Unfortunately, at various places in the Origin, Darwin appeared to shift between general evolutionary statements ("the natural system is founded on descent with modification"), phylogenetic assertions ("the ar-rangement of groups within each class . . . must be strictly genealogical in order to be natural"), and nomi-nalism ("I look at the term species, as one arbitrarily given for the sake of convenience"). Moreover, the pre-evolutionary taxonomic hierarchy had served the Dar-winians well, presaging a continuing debate to this day: Is the theory of evolution relevant to the pursuit of taxonomy only as justification for seeking a single gen-eral system, or should we build into the taxonomic method theories about the way in which evolution has occurred?

Whatever the precise reason, the immediate impact of Darwinism on method was very limited. Taxonomy remained highly individualistic, and essentialist, nomi-nalist, empiricist, and even creationist views all contin-ued, although often cloaked in evolutionary language. The first move to formulate a general method post-Darwin did not occur until the 1920s, later epitomized by two books by Julian Huxley: The New Systematics (1940) and Evolution: the Modern Synthesis (1942). This emergent approach was developed and defended most notably by George Gaylord Simpson and Ernst Mayr, and it became evolutionary systematics or evolution-ary taxonomy.

With respect to Darwin, the evolutionary system-atists argued that the natural system (Fig. 1) should reflect descent with modification, to include the pro cesses by which evolutionary change occurs, the mea surable degree of modification (anagenesis), and the temporal sequence of divergence (cladogenesis). Ac cording to Simpson (1981, p. 107), evolutionary sys tematics requires that a natural classification should be "consistent with all that can be learned of the phylogeny of the group classified," but he was emphatic that this does not mean that natural classification be based on phylogenetic relationships alone [a view also clearly stated by Gilmour in The New Systematics (1940)]. This led Simpson to adopt a very broad notion of monophyly as "the derivation of a taxon through one or more lin-eages . . . from one immediately ancestral taxon of the same or lower rank."

![FIGURE 1 Evolutionary systematization. Horizontal axis, anagenesis; vertical axis, time. In this hypothetical case, the slope of each branch corresponds to the rate of anagenesis, with the percentage values representing the degrees of difference in the three lineages from the ancestral species (A): B is grouped with C because the two are more like their common ancestor than either is to D (even though D shares a common ancestor with C at X that it does not share with B). Note that this method requires a means of measuring similarity or distance (phenetics), estimating phylogenetic relationships (cladistics), and knowledge of ancestors (paleontology) [based on Mayr (1974) and Patterson (1982)].
To be natural, according to Mayr (1969), a classification must have explanatory, predictive, and practical values but also be emendable in the light of new evidence or understanding. Mayr then proposed that the fundamental basis for naturalness is the proportion or number of genes held in common by any two taxa: "If we knew the entire genotype of each organism, it would be possible to undertake a grouping of species that would accurately reflect their 'natural affinity'" (p. 81).

The empirical approach closest to this new ideal of evolutionary systematics is the use of pairwise distance data representing, in some way, the entire genome (e.g., immunological data and notably DNA–DNA hybridization data used by Sibley and coworkers for bird classification). In addition, complete nucleotide sequences for the entire genomes of various organisms are now becoming available for direct comparison. These approaches, however, founder on some of the fundamental problems of numerical taxonomy. Mayr's proposal that 'genes-in-common' provides the ultimate arbiter of natural classification, however, is an important concept because it encapsulates the only well-articulated rival to the phylogenetic nexus idea first suggested by Darwin ('the arrangement of groups . . . must be strictly genealogical in order to be natural'). If lateral gene transfer undermines a strictly hierarchical approach, then the estimation of genes-in-common will certainly increase in importance as the basis of a natural system.

In practice, evolutionary systematics became a syncretistic, all-embracing method that included a regard for the absence of characters as informative and insisted on the primacy of paleontology for revealing phylogenetic sequences (Fig. 1). The method's most striking characteristic involved conflation of the two methods that soon sought to replace it: a desire to give expression to genetic distances (grades or anagenesis, as reflected in numerical taxonomy) and, at the same time, to ancestor–descendant relationships (clades, or cladogenesis, as reflected in cladistics). Because these two methods stem from fundamentally different philosophies, this led to an inevitable arbitrariness (Fig. 1). Thus, Simpson (1961, p. 107) was happy to write that although "taxonomy is a science . . . its application to classification involves a great deal of human contrivance and . . . there is a leeway for personal taste, even foibles." This lack of explicitness led to the demise of evolutionary systematics as the leading method in taxonomy. During its development, however, Ernst Mayr in particular made an enormous contribution, especially to ideas on the taxonomy of species, with which he was preoccupied as "the basic unit of classification" (Mayr, 1963, p. 11).

E. Numerical Taxonomy and Operationalism

Numerical taxonomy (or phenetics) emerged in the late 1950s, its origin associated with, among others, Charles Michener, Arthur Cain, and especially Robert Sokal and Peter Sneath. In Sokal and Sneath's original 1963 manifestation of the Principles of Numerical Taxonomy, any evolutionary approach is avoided in favor of an operational method based on direct comparison of phenotypes. As many characters as possible of the organisms to be compared, both continuous and discontinuous, are measured and counted from operational taxonomic units (OTUs), which can be individuals or samples from conventionally recognized taxa (typically species). On the basis of a matrix of variation in all features across all OTUs, the OTUs are then compared by overall similarity, affinity, or phenetic distance. Such measures can be obtained by transforming the raw matrix (Table I) to give the proportion of all character matches (affinity) or mismatches (distance: Table II) for every pairwise combination of OTUs. The results are displayed by means of a network (Fig. 2a), or OTUs are linked to each other by a clustering algorithm to produce a phenogram (Fig. 2b).

Although such a procedure appears objective at first, many different ways have been proposed to measure pairwise similarity or dissimilarity, and many different clustering methods have also been devised. Most clas-

**Table I**

<table>
<thead>
<tr>
<th>Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUT</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

*Single letters stand for particular amino acids and dashes for positions not represented based on alignment of the four complete myoglobin sequences (from Patterson, 1980, p. 237).
TABLE II
Data of Table I Transformed Into a Distance Matrix, Based on the Proportion of Mismatches Summed across All Eight Positions, for All Six Pairwise Comparisons of the Four OTUs

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.375</td>
<td>0.625</td>
<td>0.75</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0.75</td>
<td>0.875</td>
<td>0.62</td>
</tr>
<tr>
<td>C</td>
<td>0.25</td>
<td>0.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These distances are replicated in the branch lengths of an unrooted network (Fig. 2a). Taken as reciprocals to give measures of similarity, the data can also be used to produce a phenogram (Fig. 2b).

**FIGURE 2**
Numerical systematization. (a) The values in Table II between the four OTUs A–D represented by a distance network. The scaled lengths of the line segments are such that all the relative values in Table II are satisfied, but the angles of the four branches leading to the terminals have no special meaning. (b) Horizontal axis, linkages; vertical axis, overall similarity. By taking reciprocals of the values in Table II, the data give a similarity matrix—the basis of a phenogram. With a value of 75% similarity, (C/D) are the first OTUs clustered. Comparing A with B, and (C/D) with A and B separately, (A/B) form the next most similar cluster, linked at 62.3% similarity. The two clusters can then be linked together at 37.5% similarity based on three of the eight attributes occurring in one or both members of the two groups. Other linking procedures could be adopted.

Regrettably for phenetics, as discussed by Panchen (1992) and even acknowledged by Sneath and Sokal as early as 1973, these hypotheses, apparently basic to the validity of the methods, are now largely discredited (in particular the all-important nonspecificity hypothesis). More generally, phenetics offers no justification for choosing a hierarchical system unless it is accepted that the general theory of evolution dictates this as the most efficient representation. However, because numerical taxonomy conflates homologous and nonhomologous characters, and neither nodes nor branch lengths can be directly related to hypotheses about particular char-
acters, the origin of characters, or even the degree of
genetic change, the notion of overall similarity lacks
analytical power.
Moreover, although it was originally expected that
its procedures would lead to stability in classification,
several factors preclude this (Eldredge and Cracraft,
1980, p. 176). In addition to the technical ambiguities
already noted, every new taxon requires complete re-
analyses, with the unique attributes and combinations
of attributes usually affecting every branch length and
often many branching points in the phenogram. Of
course, other taxonomic methods, including cladistics,
are not invulnerable to change due to new discoveries,
but in cladistic analyses the effects are interesting (be-
cause they can be related to hypotheses about homol-
ogy) and cladists never laid serious claim to the idea that
stability in the face of new evidence was an important
justification for the method.
Cladistics also suffers from its own algorithmic prob-
lems, but it is always possible to work out what different
procedures are doing with respect to the data and the
inferences drawn. With phenetic methods, the nature of
the algorithms and the form of the results are not
separable. Thus, a basic problem of numerical phenetics
is the lack of a clear criterion of choice (such as parsi-
mony, or even a model of evolution) by which one result
can be judged against another. Even Mayr’s concept of
genes-in-common will not repair the difficulty because
of innumerable problems related to the definition of a
gene, duplication, alignment, position effects, and so
on. The only other external arbiter is the indefinable
standard of naturalness (‘‘a system of classification is the more natural the more
propositions can be made regarding its constituent classes’’), enthusiastically embraced by Sneath and So-
kal (1973). This led Panchen (1992) to suggest that
the original philosophy of numerical phenetics is closer
meritocratic than empiricism (not nationalism, as suggested by Mayr).
Given enough observations and an appropriate tech-
nique, the numerical taxonomists expected that a single,
stable, and predictive general system would simply
emerge.
Although numerical taxonomy has been judged
wanting, it has had a lasting and positive influence on
current taxonomic methods and still flourish in areas
such as bacterial taxonomy, in which the general ab-

F. Phylogenetic Systematics and Cladistics

The fundamental methods of phylogenetic systematics
were established by Willi Hennig, who wanted to base
systematization and classification directly on the histor-
ical branching patterns of the phylogenetic nexus. He
realized that it would never be possible to know the
course of history precisely, and that the system would
always be provisional. His goal was, by means of appro-
priate methods, to produce a ‘‘phylogenetic system’’ that
would ‘‘approximate more closely than any other the
ideal system’’ (Hennig, 1966, p. 29).
Hennig’s method was ‘‘the search for the sister
group,’’ epitomized as the ‘‘three-taxon problem.’’ For
any group of three natural taxa, the expectation is that
two have a common ancestor not shared by the third.
Thus, for the trio shark, tuna, and human, of the three
possible combinations of two from three, the one with
tuna and human grouped to the exclusion of shark
accords with what we know of ancestry (see Table III).
According to Hennig’s view, we should then recognize
a taxon linking tuna and human as sister groups

| Table III |
| Data from Table 1 Transformed to Show only Coincidences of Positive Attributes among the Four OTUs |

<table>
<thead>
<tr>
<th>Character</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon 1</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Taxon 2</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Taxon 3</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Taxon 4</td>
<td>C</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

* Characters 2–4 link (A + B), and character (11) links (A + B + C). These data are most efficiently represented by the nested-set pattern (D (C (A + B))); compare with the pattern ((D (C (A + B))) implied by the phenetic analysis (Fig. 2). In this analysis, four characters (0, 1, 3, 10) in Table I are uninformative because there is no positive coincidence (3), only shared absence (0), both (1), or only shared coincidence (10).
(Osteichthyes) but not one linking shark and tuna. From a phylogenetic perspective, Pisces are paraphyletic, or a nongroup. Thus, Hennig defined phylogenetic relationships solely in terms of common ancestry (pro- pinquity of descent) and did not include similarity, distance, grade, adaptive zone, or any comparable concept in his assessments.

This line of argument caused Mayr and other evolutionary systematists to object very strongly to Hennig's approach, accusing him of an unjustifiable restriction of the concept of relationship. For this reason, Mayr referred to Hennig and his followers as 'cladists' (from clade or cladogenesis) to criticize them for their narrow approach—with the unexpected result that the term, virtually all the characters that might seem to unite shark and tuna to the exclusion of human (e.g., gills and fins) are actually characters that relate to more inclusive levels of the hierarchy (e.g., Chordata and Vertebrata). Observations such as 'absence of mammary glands' or 'absence of hair' are conflicts characters linking 'fish' but the counter conditions of autapomorphies (of the Mammalia).

To see this in practice, the data in Table I have been transformed (Table III) to show only potential synapomorphies, i.e., the positive attributes shared by two or three of the four taxa, and thus able to provide evidence of grouping within (A–D). In this simple example there are no conflicting distributions, and there is just one optimally efficient solution. Efficiency means the hierarchical arrangement which captures all (or, where there is conflict, the largest number) informative characters (putative synapomorphies). In Fig. 3, all 15 possible fully resolved groupings of terminal taxa A–D are presented, with the characters that they can summarize marked at the relevant nodes. Arrangements 1–7, 9, 14, and 15 do not reflect any of the characters. Arrangements 10 and 12 capture one, and 8 and 13 capture three, but only arrangement 11 captures all four. The best cladistic arrangement, on the available data, is therefore (((A + B) C) D).

Colin Patterson (1980) summarized "the axioms of cladistics" as homologous characters having a hierarchical pattern in nature in which this pattern is efficiently expressed by cladograms, and the nodes connote the homologues shared by the organisms so grouped. The search for the sister group therefore reduces to finding the cladogram that summarizes the potentially homologous characters as parsimoniously as possible. Once this best fit cladogram has been found, the attributes at each node are hypothesized as homologous characters shared by the taxa subordinated at that node.

One of the most challenging features of cladistics (or at least 'transformed cladistics'—see Section VII) is the link between attributes, grouping, parsimony, characters, and homology. An attribute is accepted as a character if the weight of evidence from all other attributes under scrutiny suggests that it is a homologous feature peculiar to a particular group. In Table IV, two potential synapomorphies (x and y) have been added to the distribution of characters in Table III. Given these additional data, and presupposing independence of the attributes, the most efficient grouping of the four taxa based on positive shared features now changes and highlights a problem with the interpretation of attribute 11. Either it is not relevant as a character within the group and has undergone reversal in taxon D or it is not a character at all (it is nonhomologous).

Information about this could be sought, for example, by studying the ontogeny or some more fundamental quality of attribute 11 in A–C, revealing a particular scientific strength of the cladistic method: Conflicts can be resolved by character analysis (Kitching et al., 1998) involving recourse to additional data derived from previously unsampled or unused attributes or by investigation into the homology of the conflicting characters (they are not simply aggregated at face value, as in numerical taxonomy).

If characters x and y in Table IV were found to be due to pleiotropic effects of a single gene, then rejection of attribute 11 as a character relevant to resolving the relationships within the group would be premature. It, on the other hand, character 11 was a certain color produced, for example, by a pigment shared by taxa A and B, but the same hue was produced by a totally
FIGURE 3  Cladistic systematization. Horizontal axis, convention; nodes, pattern of homologous characters. The putative synapomorphies in Table III plotted on all 15 possible arrangements (as fully resolved cladograms) for four terminal taxa. Numbers at given nodes indicate the shared attributes from Table III; arrangement 11 is the only one that reflects all four attributes, all of which can then be construed as homologous characters (based on Patterson, 1980).

A different type of pigment in taxon C, this would be consistent with the idea that attribute 11 as originally formulated (color) was a noncharacter. The similarities and differences between taxa A–C in this regard should then be rescored as two new attributes (e.g., based on chemistry), giving further support to the (A + B) grouping and removing conflict with the (C + D) grouping. The convergence in color between (A + B) and C could then be seen to be in the eye of the human observer. Such convergences are of intense interest for the study of evolution, but once detected they play no part in natural classification.

Thus, the relativistic nature of cladistics is emphasized. Based on the whole (attribute × taxon) matrix, the cladograms chosen and attributes thereby specified as homologous characters are probability statements: “Common ancestry versus convergence is tested by topographical correspondence [on the cladogram]. The resulting explanation is a statement of maximal likelihood rather than a denotation of lawful relations or processes” (Rieppel, 1988, p. 166; see Hennig, 1966, p. 29).

V. GENERAL PROCEDURES

The main operations involved in taxonomic research are listed as follows in an idealized order. In practice, this order is rarely followed precisely, and certain steps may be omitted in part or even altogether:

1. Individual specimens, or individuals representing selected taxa (ideally at least all immediately subordi-
that the very act of cladistic analysis gives the primary
1. The second method of systematization is to produce a cladogram
2. A selection of attributes to be scored across all
3. The manifestations of these attributes are system-
4. Systematization of the data is carried out by cladis-
5. Inclusive monophyletic groups inferred from the
6. The homologous characters are used to help de-
7. Publication of the results will ideally include the

The manifestations of these attributes are system-

VI. FROM SYSTEM TO

Classification involves translating a systematization
scheme into words (or numbers). Based on the methods
of logical division, Linnaeus (who established many
of the conventions of formal classification) placed all
known organisms within a descending series of fully
nested hierarchical categories. The major ranks in the
Linnean system were kingdom (e.g., plants and ani-
mals), class (e.g., Mammalia, Aves, Pisces, and Insecta),
order (e.g., Hemiptera, Lepidoptera, Coleoptera, and
Diptera), genus (e.g., Calix, Tipula, and Musca), and
species (e.g., Musca domestica [housefly] and Musca
vomitaria [bluebottle]). For Linnaeus, systematization
and classification were the same. Modern classifications
attempt to summarize far more complex schemes. In
this brief review, only the major differences between
the major twentieth-century methods are considered.

According to the evolutionary method, classification
should be viewed as a useful art, meaning that in addi-
tion to scientific analysis, human ingenuity is also
needed to produce a practical classification. Simpson
(1961) recognized three principles: A classification
should reflect the most ‘biologically significant’ rela-
tionships among the organisms, it should be consistent
with the relationships on which it is based, and it should
be as stable as possible without contravening the first
two principles. These ideas were elaborated to embrace,
most notably, grades and clades (anagenesis and clado-
gensis), both of which were regarded as significant for
classification. This led to the idea that, even if literally
everything relevant were known about the connections
among the members of a group, ‘innumerable different
classifications could be made consistent with those in-
terrelationships.’ . . . Selection among those alterna-
tives is decidedly an art’ (Simpson, 1961, p. 110).

To appreciate the difficulties resulting from such a
view, consider Fig. 4. In cladistic terms, three paraphy-
letic groups (basal groups in A–C) are delimited, to-
gether with just 4 of the 34 subclades depicted. Why
were these particular groupings chosen? Such a mysteri-
Evolutionary classification. The original caption reads: "Three hypothetical phylogenies ... divided into taxa as shown by the broken lines. A is evidently less arbitrary ... than B, and B somewhat less than C" (redrawn from Simpson, 1961, Fig. 7).

ous art proved difficult to follow (especially when the prescription for creating such trees in the first place is also imprecise). As discussed previously, Ernst Mayr urged that classifications be based on genes-in-common. Because it is a distance method, this transformation of evolutionary systematics encounters the same problems of classification faced by numerical taxonomists. Even so, the recent estimation that man and chimpanzee share 98.4% of their genome in common has led, for example, to the long-overdue abandonment of the paraphyletic or grade family Pongidae in favor of an expanded Hominidae, in contradistinction to the position adopted for decades by the evolutionary systematists, who argued repeatedly that the "large gap" between Homo and the rest of the great apes was reason enough to place them in separate families (Simpson, 1961; Mayr, 1963).

Sneath and Sokal (1973) proposed that good classifications have three desirable properties: naturalness, ease of manipulation, and practicality for information retrieval. Their concept of naturalness, as already discussed, was based on the ideal of incorporating all codified information about a very large numbers of characters. Manipulation related to the practical relationship between a classification and its degree of hierarchical structure (useful for memorizing) and its utility (e.g., for the construction of identification keys). Convenience for information retrieval was also viewed as important, but not if it conflicted with natural classification.

The typical product of systematization in numerical taxonomy is the phenogram. How is such a dendrogram to be turned into a summary classification of named groups? An early proposal was the use of phenon lines (Fig. 5). A phenon line cuts across the phenogram at a particular similarity level. The lines must be straight and are not allowed to "bend up and down according to ... whim" (Sneath and Sokal, 1973), not only delimiting groups but also determining their rank. Thus, in Fig. 5, if OTUs 1–10 are species, the 80% phenon line could indicate seven subgenera, the 75% line four genera, the 65% line three subfamilies, and so on.

Such a procedure is not objective. The most fundamental problem is simply that the choices of percentage similarity levels and ranks are arbitrary, both within groups and between them. Other procedures, such as McNeill's method based on the assignment of rank levels to each node, also involve arbitrary limits and adjustments. The difficulty for phenetics is that "there is no necessary structural relationship between the pattern of clusters in hyperspace and the inclusive Linnean hierarchy into which they are converted" (Panchen, 1992, p. 151). If this is so, then there can be no nonarbitrary method for converting a phenogram (or any other distance-based method of systematization, including DNA–DNA hybridization trees) into a ranked summary classification.

FIGURE 5 Phenetic classification. "The formation of phenons from a phenogram" (redrawn from Sneath and Sokal, 1973, Fig. 3.33).
Cladistic classification is based on the tenet that every putative monophyletic group can be named, but polyphyletic and paraphyletic groups should not. In a grand, top-down cladistic classification of all life, all coordinate monophyletic taxa (sister groups) could be placed at the same rank. Alternatively, taxa of the same geological age could be given the same rank (Hennig, 1966). Also, classifications could be formed bottom upwards by linking all terminal sister species starting with the lowest ranking pair or pairs and building up classes and ranks based on sister groups and ascending ranks held coordinate across the entire hierarchy. Such grand visions remain almost wholly impractical due to our very imperfect knowledge. Moreover, such systems would cause massive proliferation of named groups and ranks and be unstable in the face of most newly discovered taxa and every older geological find. In terms of producing a useful summary, cladistic classification thus faces potential problems of impracticality as well as arbitrariness.

In practice, when classifications are based on cladograms an attempt is usually made, within conventional ranks (e.g., order, family, and genus), to give expression to major monophyletic groups by naming inclusive taxa at intermediate levels. However, this usually means abandoning any equation of classificatory rank with clade level and any attempt to give all groups in the cladogram formal recognition. Thus, in a labeled cladogram of the main lineages of the order Lepidoptera (Fig. 6), not all putative monophyletic groups are named, and all terminals (despite being almost completely non-coordinate) are classified at the same superfamily rank except one of the cladistically most inferior groups, the crown group Ditrysia, which appears to retain its subordinal status accorded in earlier, precladistic classifications.

One partial solution to these difficulties is the "sequencing" convention, first proposed by Gary Nelson. This attempts to combine the practicalities of a limited Linnean hierarchy with a listing that enables the entire (cladistic) hierarchy to be recovered. A sequence of taxa named at the same rank indicates that the first is sister to all following at that rank, the next is sister to the remainder, and so on. Taxa of uncertain position can be annotated as incertae sedis and those comprising unresolved polytomies as sedis mutabilis. In presenting these sequences, comprehension is greatly facilitated by indentation. Thus, the cladogram for the major groups of Lepidoptera in Fig. 6 can be converted into a written classification (Table V).

In conclusion, the idea that classification can simply be equated with systematization is a vestige of preevolutionary taxonomy and should be abandoned (Minelli, 1993, p. 14). In practice, reflecting a Simpsonian view,
TABLE V
Sequenced, Tabular Classification of the Lepidoptera, Reflecting Scoble’s Cladogram (Fig. 4)

<table>
<thead>
<tr>
<th>Order Lepidoptera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suborder Zeugloptera</td>
</tr>
<tr>
<td>Suborder Aglossa</td>
</tr>
<tr>
<td>Suborder Heterobathmida</td>
</tr>
<tr>
<td>Suborder Dacnonypha</td>
</tr>
<tr>
<td>Suborder Lophocoronina incertae sedis</td>
</tr>
<tr>
<td>Suborder Neopseustina</td>
</tr>
<tr>
<td>Suborder Exoporia</td>
</tr>
<tr>
<td>Infraorder &quot;Mnesarchaeoidea&quot;</td>
</tr>
<tr>
<td>Infraorder &quot;Hepialoidea&quot;</td>
</tr>
<tr>
<td>Infraorder Heteroneura</td>
</tr>
<tr>
<td>Infraorder &quot;Incurvariioidea&quot; sedis mutabilis</td>
</tr>
<tr>
<td>Infraorder &quot;Nepticuloidea&quot; sedis mutabilis</td>
</tr>
<tr>
<td>Infraorder &quot;Tischerioidea&quot; sedis mutabilis</td>
</tr>
<tr>
<td>Infraorder Ditrysia sedis mutabilis</td>
</tr>
</tbody>
</table>

Classifications are made consistent, as far as possible, with current systematization. Named para- and polyphyletic assemblages are rejected whenever strong evidence of nonmonophyly becomes apparent, with such groups being very misleading for comparative biology, biogeography, ecology, and extinction studies (although “dicots,” “fish,” “reptiles,” and “invertebrates” seem impossible to eliminate). Restrained use of consensus trees may be helpful here) and overelaborate nomenclature for the numerous hierarchical levels. With the real prospect of international consensus systems covering all major groups of organisms being developed on the Internet, a balance between stability and universally accessible periodic revision to reflect fundamental advances can be anticipated. Such IT-based bioinformatic systems should also solve most problems of information retrieval, including the ever-present difficulties of alternative classifications, synonymy, and misidentifications (the most pernicious of all classification problems affecting information retrieval).

VII. CURRENT PRACTICE: VARIATIONS ON A CLADISTIC THEME

Numerical taxonomy was afflicted by a proliferation of clustering algorithms, each tending to give different results for which no logical criterion of choice was available. The vaunted objectivity of phenetics dissolved, leaving the field of systematization open to the apparently more consistent and decisive methods of phylogenetic systematics. However, matters are rarely this simple.

In the 1970s, a partial schism opened between those phylogeneticists who followed directly in Hennig’s footsteps, approaching character transformation from an evolutionary perspective, and the so-called “transformed cladists” (Patterson, 1982), who held that particular theories of evolution were unnecessary for cladistics, taking instead a “taxic approach” to cladistic analysis (Kitching et al., 1998). The cladists embraced computer algorithms, character matrices, and global parsimony, leaving polarization of characters to the algorithms and outgroup choice. For Hennig, the nodes in a cladogram were hypothetical ancestors; for the transformed cladists the nodes indicate the hierarchical pattern of homologous characters. These differences have emerged in debate over congruence versus “total evidence” methods, now treated formally as partitioned versus simultaneous data analysis. Because many cladists claim to have demonstrated the superiority of simultaneous analysis for making cladograms, this debate links in turn to a seemingly fundamental disagreement over whether or not to employ Fisher’s maximum likelihood statistics in phylogeny reconstruction, as first proposed by Edwards and Cavalli-Sforza and later implemented by Joseph Felsenstein as a general method (Huelsenbeck and Crandall, 1997).

The real source of these disagreements seems to be one of different scientific agendas. Those in favor of model-based likelihood methods and partitioned data sets are mainly seeking insights into evolutionary processes, most notably those affecting molecular evolution within different lineages or under different constraints. Those rejecting likelihood models in favor of global parsimony and simultaneous analysis of heterogeneous data sets are primarily concerned with tests of homology and the construction of cladograms (but see Wiens, 1999). Both approaches have different strengths and weaknesses, and both make simplifying assumptions in pursuing these different goals (Kitching et al., 1998, p. 165). Because most work affecting taxonomy requires use of morphological data, or a combination of morphological and molecular data, for which plausable process models cannot be elaborated for simultaneous analysis using likelihood, the significance of likelihood methods for taxonomy per se (rather than understanding phylogenetic processes; Huelsenbeck and Crandall, 1997) remains to be demonstrated.
However, with respect to the reconstruction of phylogeny there are several other methods, global parsimony and maximum likelihood being only main contenders. These alternatives include James Lake’s rate-invariant method for molecular data and pairwise distances (e.g., the DNA–DNA hybridization method noted previously). Even within the parsimony techniques generally adopted for cladogram construction, at least one fundamentally different challenge to global parsimony has emerged—the three-item statement analysis, first proposed by Gary Nelson and Norman Platnick. Three-item analysis discards any presupposition that characters undergo evolutionary transformation and, by decomposing a taxon × attributes matrix into an analysis of all possible three-taxon statements supportable by the data, focuses entirely on evidence for sister group relationships rather than character transformations. Although not widely acclaimed, this method is operational and often gives more or less different results from those of conventional parsimony methods, thus raising many fundamental issues. For a brief but informative discussion of the current debate, see Kitching et al. (1998, Chapter 9).

VIII. CONCLUSIONS

Taxonomy involves the search for a general pattern of order among living and extinct organisms, from which a universal reference system (or classification) is derived. As such, taxonomy is the primary discipline of biodiversity. Fundamental philosophical disagreements about the nature of human knowledge have given rise to several basic methods. All recent approaches agree that systematics is empirical and should be based largely (evolutionary systematics and maximum likelihood molecular systematics) or entirely (numerical taxonomy and cladistics) on the comparison of data in the form of characters observed and abstracted from specimens. The methods differ somewhat with respect to the type of data preferred, and fundamentally with regard to methods of analysis used to reveal pattern, and how the system, once obtained, is translated into a written classification.

Cladistics has become the dominant method of taxonomy (where a preferred method is made explicit). The research program of cladistics is based on the view that the ideal general system should reflect the phylogenetic nexus. However, the true nexus cannot be known. If the precise course of evolution, taking account of all the relevant details of every lineage, is real but unknowable, then the natural system cannot be discovered but only approximated by an indirect process of estimation. Because cladistic classification can only admit monophyletic groups (based on the selection of a particular cladogram), all admissible groups are necessarily defined because they must have at least one unique (presumptively) homologous character (or a unique loss within a more inclusive group that is specified by some other homologous feature). Thus, as noted by both Rieppel and Panchen, the cladistic method is in practice essentialist because the search for the sister group depends on discovery of locally unique and thus defining shared characteristics for each and every group. It has long been understood that although evolution is not bound by parsimony, scientific method admits nothing better than parsimony as the criterion of choice among competing hypotheses. Insofar as the divergent model of evolutionary history which justifies searching for a hierarchical pattern is flawed (Doolittle, 1999), the methods of cladistics may need further modification, notably to deal more adequately with hybrid origins.

In conclusion, we may have to accept the seeming paradox that although the principle to which cladistics aspires is natural (i.e., the groups we seek to recognize exist regardless of human perception; Rieppel, 1988, p. 163), its empirical methods are inescapably essentialist. The basic method of taxonomy may be equivalent to the construction of a universal, never to be experienced directly but derived from the sensory experience of particulars (samples of individuals and characters) by means of an unending iterative sequence of analysis, hypothesis formation, testing, and reanalysis. This is a manifestation of the “two ways of seeing” explored by Rieppel (1988): The world of being owes its form to the processes of becoming, but those evolutionary processes can only be inferred indirectly from the patterns we observe in nature.

Finally, it is notable that current approaches to the measurement of biodiversity that seek to maximize the number of expressible genes held in networks of germplasm banks or functional ecosystems depend on models incorporating information about branching sequences (cladistic relationships) and branch lengths (anagenesis). In practice, the taxonomic hierarchy and the raw data matrix (before removal of homoplasy and autapomorphies) from which it has been derived will often be the best or only available data for use in such models. For a review of this field, see Crozier (1997).

See Also the Following Articles

CLADISTICS • CLADOGENESIS • DIFFERENTIATION • EVOLUTION, THEORY OF • GENES, DESCRIPTION
Bibliography


TEMPERATE FORESTS

John A. Silander, Jr.
University of Connecticut

GLOSSARY

evapotranspiration The process of transferring moisture from the earth to the atmosphere by evaporation of water and transpiration from plants: actual evapotranspiration as observed at a locality or potential evapotranspiration, given unlimited water availability.

neogene The Miocene and Pliocene Epochs, accorded the status of a period when the Tertiary is considered an era.

physiognomy The outward appearance or morphology of a community as determined by the growth forms of the dominant plants present.

sclerophyll A plant with tough, leathery, evergreen leaves, usually associated with drought resistance.

sere The series of stages in an ecological succession sequence.

temperate region Any locality with at least 1 month of frost (for continental areas) or with one or more months with a mean temperature lower than 18°C (for maritime-influenced areas), and with at least 4 months with a mean temperature higher than 10°C.

tertiary The earlier part of the Cenozoic Era, occurring from about 65 to 2 million years ago.

A COMPELLING CASE CAN BE MADE that the temperate forests are afforded less protection and are more at conservation risk than all other forest systems of the world. They are more altered and reduced in extent globally than any other forest type. Temperate forests currently cover only about 30–35% of their current potential extent versus about 45 and 65% respectively for tropical and boreal forests. Moreover, temperate forests are afforded less conservation protection on average than are tropical forests. Very few if any remaining temperate forests have avoided human impact. Only about 1% of the remaining Northern Hemisphere temperate broad-leaved forests is substantially unaltered and old growth; the vast majority are either managed for wood production, are in plantations, or they reflect the pervasive, long-term effects of human land use practices. Because some of the highest human population densities are found in the temperate forest biome, some of the lowest forest areas per capita globally occur in
this biome. Only the most isolated, inaccessible patches of forest remain unaltered by humans.

I. OVERVIEW

The temperate forests are globally important and unique. They host the largest and oldest organisms in the world. They serve as the world's major source of timber and wood products and are perhaps the only forests with some proven potential for sustainable management. The biomass of at least some temperate forests stands exceeds that of any tropical forest. The temperate forests of the world also provide critical ecosystem services locally and globally. Recent evidence indicates the global importance of carbon sinks in the temperate forest zone, especially in eastern North America. On a landscape level, temperate forests are critical to modulating hydrological, nitrogen, and carbon cycles. Although the biodiversity of temperate forests is typically much lower than that of tropical forests, some temperate forests approach the biodiversity observed at larger spatial scales in the tropics. Temperate forest biodiversity hot spots with high levels of endemism rival in importance those anywhere. They have a unique evolutionary history divergent from either the tropics or the boreal regions. Moreover, Northern and Southern Hemisphere forests are as different from each other as either is from tropical or boreal forests. This contrast reflects striking differences in climate, biogeography, evolutionary history, and the impact of humans.

The objectives here are to provide an overview of the distribution of temperate forests globally; their structure and composition; evolutionary history; diversity, endemism, and rarity; ecosystems services provided; current conservation status; and current threats.

II. GLOBAL DISTRIBUTION PATTERNS OF TEMPERATE FOREST SYSTEMS

A global view of the forests of the world reveals everything from broad continuous expanses of trees to mosaics of small forest patches in the landscape and from
dense, closed-canopy stands to open wooded parklands. As one progresses from the equator through the humid zones at midlatitude to tree line in the polar regions, changes in forest structure and composition typically occur gradually. Demarcating where the temperate forests begin and end along this continuum is difficult. Indeed, defining “temperate” is difficult. Of the many published maps portraying the extent of the temperate forest biome, few agree on boundaries. The broadest, most arbitrary definition of temperate forests includes all forested areas north or south of the tropics of Cancer and Capricorn, respectively. More common, general macroclimatic factors have been used to define boundaries. In such cases, the results are maps of the potential distribution of forests. Because of the pervasive effects of human and natural disturbances, such maps are considerably more extensive than those depicting extant forests (compare Figs. 1 and 2). Maps of existing forests
Global Distribution of Current Forests

FIGURE 2 The current distribution of forests globally. The data on which this map is based were assembled by the World Conservation Monitoring Centre in collaboration with the World Wide Fund for Nature and published on the World Wide Web (http://www.wcmc.org.uk/forest/data/wffm.htm).
are generally derived from remotely sensed images, but even these vary in what is portrayed as forest. In consequence, Fig. 2 does contain acknowledged, inherent errors in misclassification and omission.

Figure 1a shows where globally temperate forested landscapes may potentially be found, delimited by the humid temperate domain or ecoregion. The boundaries are based on macroclimate, which distinguishes this zone from the polar ecoregion with boreal forests, the humid tropical region with tropical forests, and the dry ecoregions dominated by arid grassland/savanna or desert vegetation. The latitudinal boundaries are set by thermal regime as modified from Köppen and Trewartha, who developed the most commonly used climate classification scheme. These zones are similar to those of Holdridge and Walter but differ in detail. The temperate region defined in Fig. 1 includes any locality with at least 1 month with frost (for continental areas) or with 1 or more months with a mean temperature lower than 18°C (for maritime-influenced areas) and with at least 4 months with a mean temperature higher than 10°C. Moisture availability sets the remaining boundaries, with the humid temperate domain bounded by where precipitation equals or exceeds potential evapotranspiration. Most other maps of temperate forest biomes employ variations on this theme.

Biome or community boundaries are only approximate, reflecting the difficulty inherent in delineating features that are fuzzy spatially and temporally. Boundaries are often either broad ecotones or a mosaic of patches. Moreover, these patterns will change as the climate, biotic composition, or disturbance regimes inevitably change over time. Consequently, any map will be at best an abstract representation of reality.

Within any biome or ecoregion, subdivisions of convenience may be designated. In Fig. 1, the humid temperate domain can be subdivided into subclimatic zones: marine (areas with temperature fluctuation moderated by oceanic influences and with elevated moisture availability, in some cases producing rainforests) and continental (areas with comparatively greater temperature fluctuation and greater probability of drought). Temperature zones are delineated as well: subtropical (defined as having 8 or more months with mean temperatures higher than 10°C), continental hot temperate (4-7 months with temperatures higher than 10°C, warmest month higher than 22°C, and coldest month lower than 0°C), and continental warm temperate (as for hot temperate, but with warmest month lower than 22°C).

The excess of annual precipitation over evapotranspiration becomes less as one moves away from oceanic influences on continents at midlatitudes. Thus, temperate forests tend to be replaced by grasslands in central North America, central and eastern Europe, central eastern Asia, and eastern South America. This may be augmented by fire and grazers or browsers. The boundary between forest and grassland may be a broad transitional, open wooded parkland with scattered trees, or a broad mosaic of forest and grassland patches with forests restricted to favorable soils, sites with more moisture, or sites protected from fire.

Forests may also be replaced as aridity increases in Mediterranean climates with winter rainfall-summer drought regimes. As fire becomes a pervasive element in the landscape, forests tend to be replaced by sclerophyllous shrublands, thickets, or sometimes open woodlands. Again, the transition may be abrupt or a gradual mosaic, with forest patches restricted to favorable soils and/or moist sites protected from fires.

The boundaries between the temperate forested regions and the adjacent boreal or tropical regions are often imprecise. In both cases, there can be a very broad transition zone with considerable overlap in species composition. The subtropical-tropical boundary set by the 18°C mean monthly isocline is quite arbitrary. Some have attempted to show that this boundary approximately corresponds to the natural poleward limits of the distribution of palms.

Boreal forests are distinguished only for the Northern Hemisphere as typically conifer-dominated forests within specified climate regimes (i.e., monthly mean temperatures all lower than 22°C but 1–3 months with means higher than 10°C). This boundary approximately corresponds with the mean position of the summer polar front in the Northern Hemisphere.

Confusion remains regarding the classification of conifer-dominated forests. The boreal forests of the Northern Hemisphere and the conifer forests of the North American Pacific Northwest, the Asian Pacific northeast, and the Southern Hemisphere are treated differently by different authors. Using the previously mentioned temperature criteria, all of the Southern Hemisphere forests, except perhaps the southern tip of South America, would be classified as temperate, as would most of the Pacific forests of Asia and North America. The climatic definition for the temperate zone forests will be used here.

There are parallel inconsistencies in classifying certain subtropical forest regions. For example, northern Indian forests, which are climatically subtropical, are usually classed as tropical. The same is true for forests with some temperate affinities in subtropical regions of southern Brazil.
Dealing with mountainous areas is problematic since they typically contain multiple ecoregions (tropical or temperate to boreal and alpine depending on elevation and latitude). In the Americas, Africa, eastern Asia, and Australasia, one can find discontinuous bands of forest with temperate affinities extending from sea level in the temperate zone well into the tropics at higher elevations. The most common resolution is to classify all forest elements as part of unclassified mountainous regions, as part of the surrounding domain, or as separately classified biome units within the domain. One consequence of this inconsistent classification is to obfuscate biotic patterns in some of the most important global hot spots of biodiversity.

The temperate forest biomes, thus defined, comprise about 14,600,000 ha (estimates vary). By far the largest actual or potentially forested landscapes occur in the Northern Hemisphere (80% or more) (cf. Fig. 1 and Table 1). The regional biomes include (i) eastern North America from the Atlantic coast west to about 95° latitude and from about 45° latitude south to about 26°; (ii) western North America from about 35° north to about 60° (and mainly from the Sierra–Cascade ranges, west); (iii) western and central Europe from the Atlantic coast north to about 60° and east through eastern Europe, but excluding the Mediterranean coastal zone and much of Spain, and then extending in a narrow strip around 55° east across Russia to west central Asia; (iv) a small, discontinuous temperate forest zone in the Middle East, especially along the south coast of the Black Sea, to the southern Caucasus and to the southern Caspian Sea; (v) eastern Asia from about 50° south to about 25° in southern China and from Japan and the Pacific coast northwest to about 120° and southwest to about 100°; and (vi) northern south Asia (India) and adjacent areas.

In the Southern Hemisphere the extent of temperate forests is much more restricted in extent: (i) eastern coastal Australia from about 25° south to Tasmania, plus the southern tip of Western Australia; (ii) most of New Zealand; (iii) southern Chile and adjacent Argentina from about 40° south to about 55°; (iv) a small area of southern Brazil just below the tropic of Capricorn, plus adjacent Paraguay and Argentina; and (v) small patches of coastal and interior forest in south and southeastern South Africa.

Table 1 summarizes the potential and current extent of forested areas within each of the 11 temperate forest biomes, from World Conservation Monitoring Centre (WCMC, see Fig. 2), World Wide Fund for Nature (WWF), and the Food and Agriculture Organization data based on approximately 25 forest types. The more

<table>
<thead>
<tr>
<th>Forested region</th>
<th>Potential extent of temperate forest cover (km² × 10⁹)</th>
<th>Maximum potential forest extent (including dry forests, woodlands, and thickets) (km² × 10⁹)</th>
<th>Current estimated extent of temperate forest cover (km² × 10⁹)</th>
<th>Conservation areas (IUCN classes I–VI) (km² × 10⁹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe (including Mediterranean)</td>
<td>3.30</td>
<td>3.91</td>
<td>0.85–1.1</td>
<td>44.0–54.0</td>
</tr>
<tr>
<td>Russia</td>
<td>1.12</td>
<td>1.13</td>
<td>0.26–0.36</td>
<td>11.0–36.0</td>
</tr>
<tr>
<td>East Asia</td>
<td>3.21</td>
<td>3.79</td>
<td>0.62–0.75</td>
<td>24.0–26.0</td>
</tr>
<tr>
<td>North America</td>
<td>4.26</td>
<td>4.72</td>
<td>2.13–2.17</td>
<td>117.0–123.0</td>
</tr>
<tr>
<td>Eastern north</td>
<td>3.56</td>
<td>3.72</td>
<td>1.63–1.57</td>
<td>—</td>
</tr>
<tr>
<td>Western (Pacific)</td>
<td>0.7</td>
<td>1.0</td>
<td>0.5–0.6</td>
<td>—</td>
</tr>
<tr>
<td>Middle East</td>
<td>0.36</td>
<td>0.61</td>
<td>0.05–0.11</td>
<td>1.0–4.0</td>
</tr>
<tr>
<td>South Asia</td>
<td>0.87</td>
<td>1.34</td>
<td>0.20–0.31</td>
<td>20.0–34.0</td>
</tr>
<tr>
<td>South America</td>
<td>0.7</td>
<td>1.8</td>
<td>0.4–0.52</td>
<td>45.0–69.0</td>
</tr>
<tr>
<td>Southern (Chile and Argentina)</td>
<td>0.60</td>
<td>0.8</td>
<td>0.25–0.30</td>
<td>51.0–54.0</td>
</tr>
<tr>
<td>Southeastern (Southern Brazil and adjacent countries)</td>
<td>0.14</td>
<td>1.0</td>
<td>0.12–0.22</td>
<td>12.0–15.0</td>
</tr>
<tr>
<td>Australia</td>
<td>0.45</td>
<td>1.64</td>
<td>0.03–0.66</td>
<td>4.0–44.0</td>
</tr>
<tr>
<td>New Zealand</td>
<td>0.23</td>
<td>0.25</td>
<td>0.04–0.08</td>
<td>17.0–18.0</td>
</tr>
<tr>
<td>Southern Africa</td>
<td>0.1</td>
<td>0.4</td>
<td>0.01–0.1</td>
<td>4.0–6.0</td>
</tr>
<tr>
<td>Total</td>
<td>14.6</td>
<td>19.6</td>
<td>4.7–6.2</td>
<td>287.0–423.0</td>
</tr>
</tbody>
</table>
conservative estimate of potential forest (Table I, column 2) is for mesic forest coverage only. Maximum potential extent (column 3) includes all types of dry forest, woodlands, and thickets but not savannas or other sparsely treed landscapes (i.e., forest cover <30%). The corresponding ranges under current forest extent are shown in the fourth column. These data are only approximate. The base data were obtained from different sources, with different categories or forest type classes reported for different regions. In some cases, conflicting information on forest cover is reported from different sources.

Almost 80–90% of the temperate forest biome types are found in the Northern Hemisphere. For current forest cover, the data are only slightly less for northern dominance. The major northern forest biomes of Europe, eastern North America, and east Asia all cover about the same potential area. However, eastern North America has by far the largest cover remaining (40–45%). The Middle East has the smallest percentage of potential forest cover (13–18%), followed by east Asia (<20%) and Europe (25–30%); western North America has the highest percentage of remaining forests (60–75+%). Estimates for western North America are only approximate (estimated from WCMC forest type information and other data), with the eastern and southern boundaries for this system fuzzy and disjunct. Similarly, the estimates for forest cover in southern Asia are only approximate. For some regions these data can be deceptive. Much of the forest covering Europe, especially in western Europe, is intensively managed plantation or seminatural forest. The same is true for Japan, even though it retains more than one-third of its potential forest cover. China retains almost 12% of its original forest, but most of this is in the boreal to northernmost, mixed conifer forests.

Southern temperate forests are fragmented into six or more regional biomes, all much smaller in extent than their northern counterparts. The extreme ranges in forest cover tabulated for some of the southern biomes reflect the large contribution of dry forest systems within the regions. The largest of the southern biomes comprises the forests of southern Chile and adjacent Argentina. The smallest of all temperate forest biomes is that found in South Africa, perhaps rivaled in size only by the forests of Western Australia. Most significantly, the South African forests are the richest of all temperate forests in tree families, genera, and possibly species (certainly if the complete tree flora is included).

Compared to other forest system globally, the temperate forest biomes covers a slightly greater area than boreal forests. Boreal and temperate forests together approximate the tropical forests of the world (Table II).

III. GENERAL CHARACTERIZATION OF TEMPERATE FORESTS

A. Physiognomic Features

Each of the previously discussed 11 temperate forest regions may be characterized by general physiognomic features (i.e., the outward appearance and structure of the dominant vegetation):

1. Eastern North America: dominated by broadleaf deciduous forests, mixed with conifers to the north, locally dominated by conifers under drier, successional conditions or in fire-prone areas in the southeast and northwest, plus small patches of broadleaf evergreen forests in the south. Closed forest systems predominate.

2. Western North America: dominated by evergreen conifers with broadleaf trees contributing little to the forests; some northwestern coastal areas support rainforests (i.e., rainfall in excess of 2000–3000 mm per annum). The eastern edge of this biome is discontinuous and grades to open conifer woodlands and montane...
boreal forests. To the south it grades to a mosaic of conifer or broad-leaved evergreen forests and woodlands or shrublands.

3. Europe: mainly broadleaf deciduous forests, mixed with conifers to the north and in mountainous areas, and to the south grading to a mosaic of broadleaf evergreen forests (many sclerophyllous leaved), conifer-dominated forests, and shrublands under Mediterranean climate influence.

4. Middle East: mainly broadleaf deciduous forests with a mosaic of broadleaf evergreen, sclerophyllous, and conifer forests and woodlands. Rainforests occur very locally near the southeastern Black Sea coast.

5. Eastern Asia: mainly broadleaf deciduous forests mixed with conifers to the north and broadleaf evergreen forests to the south; mostly mesic closed forests. Evergreen broad-leaved rainforests occur locally in southeastern Japan, one of the rarest forest types found in the Northern Hemisphere.

6. South Asia: dominated by broad-leaved evergreen to semievergreen monsoonal forests; temperate mixed forests (locally rainforests) occur in the foothills of the Himalayas.

7. Australia: dominated by broadleaf evergreen forests (mostly sclerophyllous); small patches of closed forest (<70% cover) and extensive areas of open forests grading to woodlands. Very restricted patches of rainforest are found in southeastern Australia and western Tasmania.

8. New Zealand: dominated by conifers and mixed with broadleaf evergreen forest patches, especially on the north island and locally on the south island. The western forests in New Zealand are rainforests.

9. Southern South America: dominated by broadleaf evergreen forests with some conifers. Broadleaf deciduous forests prevail in the southernmost areas and at higher elevation in the Andes; rainforests are restricted to the Pacific slope south of about 40° latitude.

10. Southern Brazil: characterized by the presence of southern conifer forests; elsewhere in this zone the forests are dominated by evergreen or semideciduous angiosperms with patches of open forest and thicket to the west.

11. South Africa: mixed broadleaf evergreen (many sclerophyllous leaved) with some conifers forming a very patchy mosaic in the landscape with thickets, shrublands, and savannas.

**B. Dominant Floristic Features**

The community dominance and floristic affinities for these regions as they exist today can be characterized very broadly. Detailed descriptions and characterizations may be found elsewhere. There are broad floristic affinities among the forested biomes in the Northern Hemisphere with many shared families and genera both now and in the fossil record. These include older lineages from the Tertiary flora of Asia and recent lineages that evolved under cooler and/or drier climates. The Southern Hemisphere forests are very different floristically, with few important families or genera shared with the north. There are some Gondwanian lineages in common now or in the fossil record. However, the floras today reflect considerable divergence with many tropical affinities and many fewer common links than are seen in the north.

In eastern North America, in the northeast birch, maple, beech, and hemlock (Betula, Acer, Fagus, and Tsuga) dominate the landscape, with the latter two tending to form monodominant stands late in succession. Farther south and west the forests tend to be dominated by oaks (Quercus spp.) or hackberries (Carya spp.). The central sections, especially in the central to southern Appalachians, tend to be the richest, with a diminished tendency toward dominance by one or a few canopy species, although the highest regional tree diversity is found in the southeast. In the extreme south, small patches of evergreen forest are found in protected areas dominated by evergreen oaks and Magnolia. Much of the southeast is now dominated by pine (Pinus taeda) plantations. Fire successional pines also dominate parts of the northwest. Drier sites throughout tend to be dominated by oaks or conifers, especially pines; wetter sites are dominated by conifers (e.g., Tsuga to the north and Taxodium to the south) or by locally adapted broad-leaved deciduous species (e.g., Ulmus, Nyssa, and Acer).

Western North American forests are dominated by a small number of large, long-lived conifer species; deciduous angiosperms are only minor components. The Pacific Coast rainforests are dominated by hemlocks (Tsuga), firs (Abies), spruce (Picea), and/or cedar (Thuja) from Alaska south to Washington State. Douglas fir (Pseudotsuga) becomes important in the central coast. Timber industries in these zones tend to actively manage the landscape for monospecific stands of native species (e.g., Douglas fir). Fragmented stands of redwoods (Sequoia), which globally are the tallest trees, occur more southerly extending to central California. On the drier eastern slopes of the coastal mountains (central sections), pines, Douglas fir, and sometimes poplars (Populus) dominate the landscape. Further south in the Sierras one finds local dominance by Sequoiadendron, the most massive tree globally, and Pinus aristata, the longest lived tree. Drier, lower eleva-
tion landscapes in the south (California) may be dominated by broad-leaved evergreen forests of tanoaks (Lithocarpus) and madrone (Arbutus) or open oak, pine, or mixed woodlands, which grade into Mediterranean shrublands or grasslands.

European forests are highly disturbed and fragmented following centuries of human habitation. Central European forests tend to be dominated by beech (Fagus) on many intermediate sites and by various oaks (Quercus) on drier and slightly wetter sites or very acidic sites. The wettest sites tend to be dominated by birches (Betula). Many of these forests have been converted to Picea or Pinus plantations. For example, German forests have gone from 90% deciduous broadleaf domination to 80% Picea plantations. To the south there is a transition to dominance by evergreen oaks and pines under Mediterranean influence. To the north and throughout much of the central uplands, conifers (Pinus, Picea, Abies, and Larix) can locally dominate the landscape or there are mixed forests with beech and birch. Many of the European forests not in plantation are still actively managed for timber products.

The Middle Eastern forests of northern Turkey to the Caucasus and the southern Caspian Sea are probably the poorest known and least studied of temperate forests. In the western Mediterranean-influenced zone, one finds sclerophyllous forests and woodlands dominated by pines and oaks. Along the coast of the Black and Caspian Seas and the southern Caucasus (Colchian and Hycanian regions) one finds highly diverse, mesic deciduous forests with oaks, maples, beech, chestnut, and many other species. At higher elevations the forests are dominated by beech with conifer-dominated or mixed stands above. Drier sites in the landscape are dominated by open oak forests or woodlands. Many of these forests have experienced a long history of human occupation and associated agriculture with overgrazing.

East Asia has the most diverse forests in the Northern Hemisphere. Along the Pacific coast in northern Japan, southern Russian, and China, one finds forests similar to those of the North American Pacific Northwest, with conifer dominance but also mixed with broad-leaved species of maples, birches, limes (Tilia), and elms. Central China has been extensively cultivated for centuries, and there is very little forest cover left. Remnant tracts here indicate a rich diversity of mixed deciduous trees mentioned previously plus many other genera, including oaks, elms, poplars, ash (Fraxinus), and rowan (Sorbus), with a rich understory. Locally, the understory may be dominated by bamboo. The same pattern is seen in Korea and Japan. In southern Japan, much of the forest is managed for native Cryptomeria, and in the north it is managed for Abies or Picea. South of the Yangtze River in China and in eastern and southern Japan the broad-leaved evergreen species increase in dominance as one approaches subtropics. These are the most diverse temperate forests in the Northern Hemisphere, with many of the same genera listed previously. However, these have become highly fragmented through human disturbance. In southern China there is a shift to forests with strong tropical affinities.

The temperate forests of south Asia are difficult to categorize. Most of the lowland and montane remnant forests of northern India are climatically subtropical but the flora has strong tropical affinities. There is local strong dominance by Sal (Shorea robusta) and bamboos. In the northern hill forests there is a transition to strong temperate affinities with high species diversity. Oaks mixed with Lauraceae dominate the forest, but maples, Castanopsis, and Magnolia occur. In the montane zone diverse oak forests are mixed with conifer (Abies, Picea, and Pinus) forests and patches of Rhododendron. As is the case in much of Asia, these forests have been long affected by human disturbance. Throughout the lower elevations small stands of Eucalyptus, teak, pine, or Populus plantations are common.

The subtropical forests of the eastern mountains and coastal areas of Australia and the small areas of forest in Western Australia are dominated by the numerous sclerophyllous Eucalyptus species. Locally, stands tend to be dominated by one or only a few species. The eucalypt forests tend to form open-canopy stands grading to woodlands. The tallest angiosperm trees are found here. Only on Tasmania and in scattered pockets along the eastern mountain chain is there local dominance by closed-canopy or temperate rainforest species, including the southern beech (Nothofagus) and various southern conifers (Dacrydium, Phyllocladus, Arthrotaxis, or Araucaria). Throughout much of southeastern Australia plantations of Monterey pine (P. radiata) have become a pervasive component of the landscape.

The forests of New Zealand tend to be either multi-storied mixed conifer—broadleaf with species composition varying across the landscape or low-diversity southern beech (Nothofagus) forests. The mixed forests dominating in the lowlands have a scattered overstory of Agathis in the north or various podocarps (e.g., Podocarpus, Dacrycarpus, and Phyllocladus), with a sub-canopy of Lauraceae (Bellschmeedia), Myrtaceae (Metrosideros), Cunoniaceae (Weinmannia), and many other families and genera. Evergreen Nothofagus forests may form pure dense canopies in subalpine areas and may be a component in the lowland forests along with other broad-leaved species.
The forests of southern South America are confined to Chile and adjacent areas of Argentina. They vary from small remnants of sclerophyllous forests and woodlands in the Mediterranean zone to the species-rich Valdivian rainforests, the species-poor but still extensive north Patagonian and Magellanic forests, and depauperate deciduous Nothofagus forests at higher elevations and the interior south. The sclerophyll forests were dominated by Acacia caven and other species, with deciduous Nothofagus forests at high elevation. The Valdivian forests may be either broadleaf dominate, with Nothofagus, Eucryphia (Eucryphiaceae), Laurelia (Monimiaceae), Weinmannia, and other species, or mixed with conifers [Podocarpus, Araucaria, or Fitzroya (Cupressaceae)]. The north Patagonian/Magellanic forests are dominated by evergreen Nothofagus mixed with Podocarpus, Weinmannia, and Drimys (Winteraceae). To the south, one of the deciduous Nothofagus species (N. pumilo) forms a pure stand or is mixed with N. betuloides at timberline. Many of the remaining Chilean forests are being clear-cut for chips and converted to P. radiata or Eucalyptus plantations.

The southeastern forests of Brazil have many south temperate affinities. These forests are characterized by the presence of Araucaria. However, other south temperate components include Podocarpus, Weinmannia, and Drimys, plus Sapindaceae, Proteaceae, and Myrtaceae. These forests have been largely cleared for agriculture. The subtropical forests to the west in adjacent Paraguay and Argentina have more tropical affinities.

The forests of South Africa are some of the richest in tree species of any in the temperate zone. However, this is also the smallest of all temperate forest biomes, and it is highly fragmented into many small forest patches. It is not clear how forest cover has changed during the Pleistocene. These forests do have south temperate affinities, with the presence of Podocarpus, Cunoniaceae, and Proteaceae, but the majority of the temperate forest flora have tropical affinities. Despite these tropical affinities, the level of tree endemism is high for a continental area contiguous with tropical forests. The "Afromontane" forest elements extend from southernmost South Africa at sea level to the mountains through northeastern Africa. The coastal, Indian Ocean (Maputaland/Pondoland) forests are quite different floristically, with high diversity and many local endemics. The vast majority of the forested landscape is now in P. radiata or Eucalyptus plantation.

### IV. TRENDS IN BIODIVERSITY

Comparing trends in biodiversity within and among regions is fraught with difficulties. The results can differ depending on the scale of the sample unit compared (i.e., 0.001 vs 1, 1000, or 100,000,000 ha). For many regions of the world, data are available only for a limited scale range. For example, in east Asia there are very few accessible data records for small plots (0.1–100 ha). In other regions (e.g., the Middle East), species diversity numbers are either estimates for large areas or entirely lacking. For many records the number of tree species may be accurately reported but the number of herbaceous, especially ephemeral, plants may be significantly underestimated or not reported at all. Even simply listing tree species numbers can be misleading because authors vary widely in delineating the threshold size for what constitutes a tree. Likewise, authors vary in classifying vegetation type with which tree taxa may be associated (closed forest to sparsely treed parkland). Nevertheless, some trends appear robust.

Table III summarizes regional tree taxon richness from a variety of different sources, and Fig. 3 plots tree richness tallies against area for 75 forest sample sites throughout the temperate zone (spanning 10^-2 to 10^6 ha). Despite this large range in areas and the inherent variation in estimates and counts, there are significant differences in tree richness among biomes and between hemispheres. For these data, greater tree species richness occurs in the Southern Hemisphere across the full range of areas surveyed (Fig. 3). This is also reflected

![Figure 3](image)

**Figure 3** The relationship between tree species numbers and area sampled is shown, log transformed for both variables. The samples are from a wide range of different sources for most of the temperate zone forest biomes. These are simply classified here as Northern (n) and Southern (s) Hemisphere localities with best fit linear regression lines. There were significant differences between hemispheres ($p = 0.024, n = 75, r^2 = 0.695$) and among regions ($p < 0.001, r^2 = 0.916$).
in the absolute number of families and genera. For species, the totals are similar, but the tabulated survey covers only a small part of the diverse tree flora for Australia and South Africa, respectively, and only part of southeastern Australia and Tasmania and KwaZulu-Natal province and adjacent Transkei. If all the tree species are included for South Africa alone (well over 1000 species in 370 genera and 97 families), the species numbers would be higher in the Southern Hemisphere despite the fact that southern forests only cover 10–20% of the area of northern forests. On an area basis, the taxon richness of the southern forests is at least an order of magnitude greater than that of the north.

This high southern diversity is contributed largely by the flora of South Africa, arguably one of the richest per unit area of any biome globally. Australia and New Zealand also contribute to this southern richness, each having a taxon richness per unit area of 4–10 or even 100 times that of northern forested regions. Just the rainforests of New South Wales and Victoria, covering less than 200,000 ha, have more than 250 tree species (not all included in Table III). Even the Chilean forests, perhaps the most depauperate in tree species of any temperate forest biome, have more taxa per unit area than any northern biome (except for total species in east Asia).

Among northern temperate forest biomes, east Asia has by far the richest tree flora. Europe and western North America are the most depauperate, with Europe having the lowest tree taxon diversity per unit area globally.

Figure 4 shows analogous species area data for all vascular plants tallied for about 200 plots or regions (spanning $10^{-6}$ to $10^8$ ha) across all temperate forest biomes. Here, the hemisphere trends are reversed. The north has a slight but significantly higher total vascular flora than the south across the range of sample areas. East Asia and the Middle East (the latter, a smaller sample size) tend to have the highest vascular plant diversity, and New Zealand, South America, and western North American forests tend to have the lowest vascular plant diversities. Forested systems in eastern North America, Europe, Australia, and South Africa tend to be intermediate across areas sampled. There is a dramatic decrease in species diversity toward the pole, more so in the Southern Hemisphere. In pairwise comparisons at intermediate sample areas the trends tend to hold. The vascular plant flora for east Asia at 0.5–10

### Table III

<table>
<thead>
<tr>
<th>Region</th>
<th>Families</th>
<th>Genera</th>
<th>Species</th>
<th>Genus:family</th>
<th>Species:genus</th>
<th>Forest biome maximum extent (km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>21</td>
<td>43</td>
<td>124</td>
<td>2.0</td>
<td>2.9</td>
<td>3300–3910</td>
</tr>
<tr>
<td>East Asia</td>
<td>67</td>
<td>177</td>
<td>876</td>
<td>2.6</td>
<td>4.9</td>
<td>3210–3790</td>
</tr>
<tr>
<td>Eastern North America</td>
<td>46</td>
<td>90</td>
<td>253</td>
<td>2.0</td>
<td>2.8</td>
<td>3560–3720</td>
</tr>
<tr>
<td>Western (Pacific) North America</td>
<td>24</td>
<td>47</td>
<td>131</td>
<td>2.0</td>
<td>2.8</td>
<td>700–1000</td>
</tr>
<tr>
<td>Chile</td>
<td>29</td>
<td>40</td>
<td>83</td>
<td>1.4</td>
<td>2.1</td>
<td>330–370</td>
</tr>
<tr>
<td>Southern Brazil</td>
<td>25</td>
<td>45</td>
<td>77</td>
<td>1.8</td>
<td>1.7</td>
<td>&lt;100</td>
</tr>
<tr>
<td>Southeast Australia</td>
<td>37</td>
<td>78</td>
<td>331</td>
<td>2.1</td>
<td>4.2</td>
<td>300–700</td>
</tr>
<tr>
<td>New Zealand</td>
<td>47</td>
<td>74</td>
<td>212</td>
<td>1.6</td>
<td>2.9</td>
<td>230–230</td>
</tr>
<tr>
<td>Southeast South Africa</td>
<td>88</td>
<td>280</td>
<td>598</td>
<td>3.2</td>
<td>2.1</td>
<td>20–50</td>
</tr>
</tbody>
</table>

![Figure 4](image_url)

**Figure 4** The relationship between total vascular plant diversity and area samples is shown, log transformed for both variables. The samples are from a variety of different sources covering most of the temperate zone forest biomes. The samples are classified by the biome region from which they were obtained. Significant differences are found between hemispheres ($p = 0.005$, $r^2 = 0.924$, $n = 198$) and among regions ($p < 0.001$, $r^2 = 0.938$).
ha is significantly richer than that of Europe or eastern North America. However, few differences can be seen among regions at the smallest plot sizes (<0.01 ha).

Trends in alpha, beta, and gamma diversity vary considerably within and among regions. Some of the large- to intermediate-scale patterns in species richness are undoubtedly related to spatial landscape heterogeneity. Regionally, high levels of species richness are associated with mountains (e.g., the Smoky Mountains (east North America), the Pyrenees, Alps, or Balkan Mountains (Europe), and the Sichuan Mountains or Mount Halla (east Asia)). Complex mosaics of vegetation types (e.g., southeast Australia or Pondoland (South Africa) or multiple successional states (e.g., Indiana Dunes National Seashore) all tend to support significantly higher diversity than adjoining areas. Even for the United Kingdom, with an exceptionally small flora overall, regional or county floras that encompass a diversity of habitats and seral stages may have as many species as are found in many similar-sized temperate regions elsewhere. Within Europe there many more tree species in the southern, Mediterranean-influenced region with complex spatial heterogeneity (almost 100 versus 12 common and 25 uncommon species in northern and central Europe). Perhaps the high tree and overall species diversity in South Africa is related to the extreme spatial heterogeneous and natural fragmentation of the vegetation (forest and otherwise). Locally or regionally, lower species diversity tends to be found under closed conifer forests and late successional stands dominated by trees that cast deep shade. Often, the highest diversity will be found in intermediate successional stage forests or in a landscape with a mosaic of seral stages. Locally, higher diversity will tend to be found on richer, fine-textured soils with circum neutral pH with higher cation exchange capacity and more humus. This contrasts with diversity trends elsewhere. For example, higher diversity tends to be found on poorer soils in Mediterranean shrubland biomes.

Compared to tropical forests, the species richness (of trees or all vascular plants) in the temperate zone will usually be smaller than that of comparable areas in the tropics, especially at the 0.05- to 10-ha scale. Most temperate forest stands tend to be dominated by one or a few species, with the other tree species being uncommon or rare. In the majority of tropical forest stands local dominance by one or a few species is rare; greater evenness is common. Perhaps only on a regional basis may floristic diversity of temperate forests (e.g., southern China and South Africa) begin to approach those observed in the tropics.

An alternative measure of global richness patterns may be obtained by focusing on clades or higher taxonomic richness. The data from Table III show more genera and families in the Southern Hemisphere forests sampled. Also, the ratios of species to genus and genus to family tend to be lower in the south. These trends indicate a potentially richer phylogenetic diversity (using higher taxa in lieu of cladistic information) in the south. A complementary, global perspective is provided by the British Natural History Museum (NHM) global mapping of plant family richness (395 in total) on 10° latitude/longitude grids. The trends reveal higher family diversity on average in the tropics, with the richest cell in Southeast Asia, and a dramatic decline poleward. However, family richness is as high in southern China as it is anywhere else in the world (other than Southeast Asia), and about as high in the central southern United States as in most of South America (including Amazonia) and southern Africa versus tropical Africa. Centers of plant family richness in the temperate zone at the 10° grid cell included southern east Asia, southern North America, and southern Africa, with lower diversity in Europe and temperate Australia and South America.

V. ENDEMISM AND RANGE SIZE RARITY

Endemism (species restricted to specific localities) and taxon rarity have received a similar amount of attention as has biodiversity. How do levels of plant endemism compare among temperate forested regions and with other biomes? Statistics on plant endemism tend to be available for only a limited range of spatial scales (often countywide, occasionally for states or provinces, and infrequently for localities) and are often incomplete or lacking for poorly known regions such as the tropics. Nevertheless, there are some trends in the data that are available. Boreal and cold temperate forested regions in the Northern Hemisphere tend to have very low levels of endemism based either on absolute numbers or the percentage expected on average per unit area [e.g., northern and western European countries (0-2%) and the eastern United States (0 to <1%)]. Regions that include warm temperate to subtropical regions with topographically heterogeneous landscapes (and therefore heterogeneous climates and vegetation types), and especially with isolated mountain ranges, tend to have higher than expected endemism per unit area [e.g., Bulgaria (9%), Turkey (31%), western North America, and countrywide for China (56%)]. Larger islands and
peninsulas tend to have higher than expected species diversity (e.g., Korea (14%), Florida (12%), Japan (37%), and New Zealand (82%)). The Southern Hemisphere forested regions tend to have some of the highest levels of floristic endemism globally (50–80% for Australia, New Zealand, Chile, and South Africa). In part this reflects their geographic isolation. The previous data are for species endemism for the entire flora. Tree species endemism is more difficult to determine and likely to be much lower. For example, in South Africa about 25% of the more than 1000 tree species are confined approximately to the borders (and enclaves) of the country and about 40% to southern Africa (versus 70–80% for the entire flora). Regional tree endemism is much lower (e.g., about 8% for Pondoland and 15% for Maputaland forest regions). However, very high levels of tree species continental endemism are found in southern South America: 85% of woody species and 34% of genera are endemic.

Compared to tropical forests, the absolute numbers of endemics are generally lower in the temperate zone, except perhaps for some of the Southern Hemisphere temperate zone. This reflects the larger floras in the tropics, which are still poorly known for many localities. However, on the basis of expected percentage endemism per unit area the trends are less clear. For example, although Venezuela (38%), Panama (13%), and the Congo (29%) have higher than expected levels of plant endemism, Columbia (4%), Nicaragua (<1%), and Nigeria (4%) have lower than expected endemism. On a subregion or local basis the tropics may well have substantially higher levels of endemism than equivalent localities in the temperate zone, but this pattern remains to be shown conclusively.

Species endemism is only one way to assess rarity or the geographical restrictions of taxa for conservation purposes. Range size rarity of taxa or clades evaluated in the absence of political or other arbitrary boundaries may be a more consistent and robust means of judging rarity. A global survey by the British NHM of range size rarity was done on a 10° grid for selected plant taxa. This showed that there were no clear trends between warm temperate and tropical regions in range size rarity. Many, perhaps the majority of the prime hot spots, are in temperate grids that include forested regions (e.g., southern South Africa, central China, southeastern Australia, and central to south-central Chile, with secondary centers in southern Chile, northern New Zealand, elsewhere in China, and the southeastern United States). Clade rarity and endemism need evaluation as an alternative to enumerating species endemism or range size rarity.

VI. EXPLANATIONS FOR PATTERNS OF DIVERSITY AND ENDEMISM

A variety of hypotheses have been put forward to explain local and global patterns of species distributions and therefore richness. One author lists as many as 120 named hypotheses for variation in species richness. Geological and biogeographic history provides one set of important related explanations for observed trends in richness and endemism. This may well explain some of the patterns within and between hemispheres.

A. Historical and Geological Explanations

In the Triassic period global plate tectonic activity had united major landmasses of the world to form a single supercontinent (Fig. 5) which created the potential for a common biota. However, by the beginning of the Cretaceous this landmass began to break up, forming northern and southern landmasses. This, together with the formation of a broad sea, effectively isolated the northern Asiamerica continent from the southern Gondwanaland. The angiosperm and gymnosperm floras on these separate landmasses thus began to evolve independently. This history is reflected in both the Tertiary floras of the fossil record and the floras we see today. Until approximately 40 million years into the Tertiary the climate was warm, moist, and relatively stable. In the Northern Hemisphere forests spanned North America and Eurasia into the present-day Arctic. In the mid-Cretaceous (100 million years ago (mya)) these forests were dominated by both conifers (Pinaceae, Gymnospermae, and Taxodiaceae) and angiosperm taxa. Early examples of widespread flowering plant genera include Magnolia, Betula, and Platanus. Many modern northern genera soon followed in the fossil record (e.g., Quercus, Castanea, Carya, Ulmus, Juglans, and Acer), also spanning the Northern Hemisphere.

By the beginning of the Neogene, approximately 25 mya, the climate began to change, becoming cooler and drier in certain regions as a consequence of mountain building and shifts in ocean currents. In the cold northern regions, conifers were favored, giving rise to the boreal forests. At midlatitude, summer drought and/or winter cold favored deciduous angiosperms. Midcontinent aridity also gave rise to the temperate steppe grasslands and deserts. This, together with Pleistocene glaciation, undoubtedly contributed to divergence in the north temperate forest floras and their biodiversity.

In Europe and central Asia, east–west trending mountain ranges and arid zones blocked the southward
FIGURE 5  Changes in the locations of continental landmasses are shown from the Mesozoic to the early Tertiary in Lambert equal area projections. Areas on the back projection of the spheres are folded out for view. Dashed lines indicate shoreline of modern continents, hatched areas are epicontinental seas, and stippled areas are oceans (reproduced with permission from Cox and Moore, 1985).
retreat of the forests as the glaciers advanced. In consequence, there was apparently greater species extinction here than elsewhere in the Northern Hemisphere, and geographic isolation prevented recolonization from temperate or tropical locations elsewhere. In contrast, in east Asia, continental glaciation was less extensive and continuous connections with tropical wet forests permitted temperate forest species to retreat south and advance north with climate oscillations; there were many fewer extinctions and thus greater retention of an older phylogenetic diversity. This tropical connection undoubtedly accounts for the greater tropical affinities seen in the flora of east Asia today. In eastern North America the mountains trend north–south and thus migrations with the glacial oscillations were possible. However, the tropical connections were largely blocked by seas or arid zones. Compared to Asia, there were more extinctions with fewer opportunities for recolonizations by taxa with tropical affinities. In eastern North America the phylogenetic lineages have a more recent and more northerly balance. Western North America suffered the greatest extinctions. Extensive mountain building, accompanying aridity, and the development of a Mediterranean climate to the south favored conifers or sclerophyllous trees with divergent evolutionary lineages from those of the mesic Tertiary forests. To the north, cool, wet climates favored conifer forests (many with tertiary affinities) at the expense of angiosperms.

The forests of the Southern Hemisphere began similarly in the Cretaceous. Warm, mesic, mixed conifer and angiosperm forests spanned Gondwanaland. These were dominated by southern conifers, including Araucariaceae and Podocarpaceae, and southern angiosperms, including Cunoniaceae, Proteaceae, Myrtaceae, and Sapindaceae. In addition to these, Casuarinaceae and the genus *Nothofagus* were present throughout, except on what became Africa. Gondwanaland began to break up early in the Cretaceous (Fig. 5) well before Asiamerica. Southern Africa and India had become isolated by the early Cretaceous. This undoubtedly accounts for Africa's greater dissimilarity with the rest of the southern temperate flora. India eventually joined Asia and today it has primarily north temperate or Asian tropical floristic affinities. Not until early in the Tertiary did South America, Antarctica, and Australia finally separate. Climate changes in the Southern Hemisphere were less dramatic during the Neogene and Pleistocene than they were in the north. Aridity and glaciation were more localized; the climate remained less continental. In consequence, the forest flora retains more characteristics of the warm, mesic Tertiary flora than is the case in the north.

During the Neogene, temperate southern Africa experienced increased aridity. With climatic oscillations the forests expanded and contracted. Many of the Gondwanan elements became extinct, leaving only a couple of species of Podocarpaceae and a minor contribution of Cunoniaceae and Proteaceae to the forest flora today. Connections along the east coast of southern Africa allowed connections with the tropical forest flora, and today the temperate South African forests have the strongest tropical affinities of any temperate region and the greatest tree diversity. Exceptionally high levels of spatial and temporal environmental heterogeneity may contribute to speciation here as well. With substantial tropical affinities also found in Australia and New Zealand, the southern forests as a group have a much stronger representation of tropical plant families than do the northern forests.

During the Pliocene the southern temperate forest region of Chile and adjacent Argentina became isolated from the rest of South America by the Andes and arid zones to the north and east. The isolation and relatively small extent of these forests undoubtedly account for the low diversity and high endemism found here. Only in the extreme south of Chile and at high elevations with lower mean and absolute temperatures are deciduous (*Nothofagus*) forests found in the Southern Hemisphere.

Australia and New Zealand retain much of their Tertiary flora. The maritime climate of New Zealand has changed relatively little during the Tertiary. Its long isolation from other landmasses is responsible for the high level of endemism. Increased aridity in Australia in the Neogene substantially affected the temperate forest structure and composition. Explosive radiation occurred in the genus *Eucalyptus* (Myrtaceae) and *Acacia* (Mimosaceae). Unique among temperate forests, there are more than 500 species of *Eucalyptus* in Australia (many restricted to temperate forests). These are sclerophyllous-, drought-, and fire-tolerant tree species that dominate most of the temperate forest and woodlands in the east and west. Only in moist, cool patches in the southeast or at higher elevations are remnants of the tertiary flora found. *Nothofagus*, Araucariaceae, Podocarpaceae, and other groups are part of the rich tree flora.

The previous account indicates that it is not simply history or geology but a complex process that incorporates the role of temporal and spatial heterogeneity, geographic isolation, recent Pleistocene effects, and the role of climate stability that jointly affect the patterns of taxon distributions, diversity, and endemism we see today. What about other explanatory variables?
B. Productivity

The differential effects of solar radiation figure prominently as an explanatory hypothesis, measured either directly or indirectly via actual or potential evapotranspiration (i.e., AET or PET) or productivity. There is conflicting evidence on the relative importance of this as an explanatory variable versus historical factors, at least at the regional scale. Examined at the plot level (~1 ha) across regions, correlations are found between diversity and AET. There are also correlations at the regional and latitudinal scales (boreal vs temperate vs tropical forests). However, other evidence shows that differences at regional and local scales can be well explained by historical factors after taking into account AET. Global joint correlations between species richness and productivity were examined by the British NHM for plant family diversity on a 10° grid scale. This study showed higher family richness than expected from productivity in southern Africa, southern North America, and southern east Asia and lower or expected diversity elsewhere in the temperate zone. Undoubtedly both sets of factors may be involved, but these remain only correlates; cause-effect relationships have not been demonstrated.

C. Spatiotemporal Heterogeneity and Other Explanations

The roles of disturbance regimes and environmental or habitat heterogeneity at local and regional levels have figured prominently as explanatory hypotheses for richness patterns. Higher diversity is predicted under intermediate levels of disturbance. On a local or landscape level there is considerable evidence to support this idea. Midsuccessional forests and successional diverse landscape support a richer flora. For example, in northeastern North America many of the extinct or threatened species are those that occurred in open or early successional landscapes that were common 100 years ago but have since disappeared with the homogenizing reforestation of the landscape.

The predominant natural disturbances in temperate forest zones include fire and windstorm. Fires are most prevalent in the drier forests of western North America, Australia, South Africa, and the Mediterranean Basin. In many of these systems, periodic fires contribute to higher landscape biodiversity. The role of fires in other forest systems is less clear. Large cyclonic storms affect mainly eastern coastal, midlatitude regions in the Northern Hemisphere (eastern North America and Asia). Tornadoic events are most prevalent in midcontinental North America. Such storms contribute to heterogeneous successional stages in the landscape and thus higher diversity. Occasionally, human activity can contribute to higher landscape diversity. Historically, Europeans husbanding the landscape for diverse forest products undoubtedly increased the local diversity over what it would have been naturally. Humans creating a moderate spatially and temporally heterogeneous landscape ironically may contribute slightly to increased regional diversity. The higher regional species richness in mountainous areas associated with the manifold habitats was discussed previously. Another example is the extremely rich South African forests (and other biomes). The high local and regional richness has been largely attributed to the environmental heterogeneity found at this scale.

In addition, there are many other explanatory variables for patterns in species richness, some of which were intimated or included in the topics discussed previously: environmental stability or predictability, abiotic rafication, land area, seasonality, aridity, range limits and geometric constraints, and many more. For many of these factors it is easy to establish correlation and much more difficult to establish cause-effect relationships.

VII. ECOSYSTEM SERVICES

Temperate forests provide important ecosystem services globally, regionally, and locally. Temperate forests contribute about 17% to global net primary productivity (versus about 49% for tropical forest systems and 8% for boreal forests). However, recent evidence from atmospheric and oceanic CO₂ data point to temperate forests, especially those in eastern North America, as globally important carbon sinks. The magnitude and causes of the net carbon uptake by temperate biomes are uncertain. However, it may be related to the reforestation that has occurred during the past century, especially in North America. CO₂ fertilization, anthropogenic N deposition, and global warming may contribute as well.

At local to landscape levels there are tight links among forest structure, composition, and species richness; soil attributes; mineral and hydrological cycles; and human disturbances. In the moist temperate zone species richness tends to be higher on soils that are
better drained, warmer, and finer textured, with greater NO3-N and P availability and higher cation saturation and lower Al+ (toxic) levels, all associated with higher pH. These attributes are associated with many calcareous soils. Thinner calcareous soils, which apparently create greater spatial heterogeneity in forest canopies, are also associated with higher local richness. Hence, soil attributes are intimately related to pattern in species diversity. The extent to which these features are altered will directly affect biodiversity, the extent to which the community structure and composition are changed will feed back on these variables.

The strongest correlates of productivity tend to be soil texture and N cycling, which in turn are correlated with moisture retention, cation exchange, and maximum levels of humus accumulation. These correlations are strongest in undisturbed forests. Also, ecosystem control by soil texture extends over long time frames. Human disturbances can profoundly affect these links. Soil compaction with intensive forest management will have a cascading and long-term effect on nutrient and hydrological cycles. As the studies at Hubbard Brook Experimental Forest have demonstrated, forest clearing will dramatically alter watershed hydrology (increased water loss and sediment loss), nutrient cycling (with elevated nutrient loss), microclimate, and species composition for years or decades. Changes in forest species composition such as with conifer plantations will have an equally large effect. Substituting homogeneous conifers stands for mixed broad-leaved forests will increase C/N and lignin/C ratios, which in turn reduces decomposition rates, N mineralization rates, and pH, with cascading effects on cation exchange capacity and diversity in the soil flora and fauna. For example, in German *Picea* plantations, N cycling between canopy and forest soils is reduced by 75% from that observed in native beech forests. Moreover, with greater stem interception of precipitation, the conifer forest soils also become drier. In consequence, it may not be possible to successively reintroduce beech forests on these sites without large-scale soil amendments.

Even small changes in the broadleaf forest composition can alter decomposition rates and nutrient status. For example, sugar maple (*Acer saccharum*) and ash (*Fraxinus*) promote higher N mineralization, and sugar maple accumulates calcium. In contrast, oak-dominated forests have lower N mineralization. Substituting exotic species in plantations can dramatically alter ecosystem processes. *Eucalyptus* and *Melaleuca* tend to significantly lower water tables where they have been introduced. In the northeastern United States, Japanese barberry (*Berberis thunbergii*) has become a seriously invasive exotic, forming a continuous shrub layer. Under barberry canopies soil ammonium N levels are elevated, the soil flora and fauna are altered, and native species richness is depressed.

### VIII. CONSERVATION STATUS

Table I shows the approximate extent of protected forests in each of the regional biomes based on WCMC global data for International Union for the Conservation of Nature (IUCN) conservation protection categories I–VI. Globally for temperate forests, 6 or 7% of the remaining forests receive some level of protection. This represents about 1 or 2% of the total temperate forest biome extent. Compared to other forest systems, temperate forests are slightly better protected than boreal forests (6 or 7% versus 5 or 6%) but apparently less well protected than tropical forests as a whole (10–12%).

There is considerable variation both within and among regions in the level of forest protection as well as among forest types. For example, temperate freshwater swamp forests are afforded the least protection (2.7% globally), whereas Southern Hemisphere evergreen broadleaf forests receive the highest level of protection (22.6% globally). Overall, Southern Hemisphere forests are afforded significantly better protection as a group than are northern forests.

The level of protection varies considerably among regions and countries within the temperate zone: Less than 3% of east Asian temperate forests are protected, but this varies with forest type from about 5% for cool to cold temperate and subtropical forests to 1% or less for most warm to hot temperate forests types (except for warm temperate rainforests, which only cover a tiny area and are about 15% protected). There is considerable variation among the countries of east Asia in the amount of forest afforded protection: almost 10% for Japan to less than 5% of China and less than 1% of North Korea. For Europe (excluding Russia) about 8% for the various forest types are given some protection. This also varies considerably among countries: As much as 25% of the broad-leaved deciduous forests in Germany are protected, whereas <1% in Bosnia and 2% in Russia are protected. For North America, the extent of forest protection varies from about 15% for conifer-dominated temperate rainforests to 7% for the northern cool or cold temperate forests and about 2% for the warm to hot temperate and subtropical forests versus more than 9% for North American boreal forests. The
forests of the Middle East, which are the most reduced in extent, are afforded the least protection overall. Less than 3% of the forests are protected in this zone. This varies from <1% in Georgia and Azerbaijan to 1% in Turkey and perhaps as much as 12% in Iran. About 8% of India's temperate forests have some level of conservation protection. For temperate Australia about 9% of the forests are protected; for New Zealand the figure is about 43%. For South Africa, about 24% of its southernmost temperate forests are protected. For South America, about 23% of the temperate forests are afforded some protection, but the vast majority (more than 90%) of these are the wet evergreen forests of the south, mostly located in Chile. Less than 2% of the sclerophyllous and dry temperate forest are protected. An estimated 8% of Brazil's temperate mixed conifer forests are protected.

Overall, the most well protected forests, on a relative scale, are the Southern Hemisphere temperate rainforests, the Pacific wet forests of western North America and east Asia, the northernmost temperate forest (mixed with boreal elements), and mixed temperate forests in mountainous areas. The most poorly protected are dry and sclerophyll leafed forests of the Northern Hemisphere, wet and rainforest broadleaf deciduous and evergreen forests in the Northern Hemisphere (very limited in extent), and moist temperate deciduous and evergreen broad-leafed forests of the Southern Hemisphere. East Asia has the highest percentage of temperate forest types receiving no protection, followed by the Middle East and South America.

In some ways these data are misleading. Many of the forests classified as receiving protection are plantation or otherwise highly managed forests. These are largely native species forests in Europe, North America, and Asia, but they tend to be managed as homogenous, even-aged monocultures with consequent reduced diversity and ecosystem services. Natural, undisturbed, or old-growth forests comprise only a small fraction of the remaining forests: western Europe, 1%; eastern North America, 1%; eastern Asia, 1%; Australia, about 5% (but probably <1% is unlogged); South Africa, <1%. Somewhat better off are New Zealand with about 25%, northwest (Pacific) North America with about 13%, and southern Chile with about 45%.

Some temperate forests are unique in that significant reforestation has occurred during the past century (about 1 or 2% increase on average per annum), primarily in Europe and eastern North America. However, most of these net gains in forest cover are intensively managed (in Europe and the southeastern United States). In northeastern North America, during the past century the landscape has reverted naturally from 50–90% agricultural to mostly forested, following the extensive abandonment of agriculture. Most of this forest is highly fragmented and fairly homogeneous, being of similar age. Consequently, the ecological functionality of these forests is limited. Elsewhere, temperate forest cover continues to decline at rates that vary from 1 to 10% pa.

**IX. CONSERVATION AND PROTECTION STRATEGIES**

How does one develop a strategy for conserving or protecting the remaining temperate forests of the world? Some authors focus on "hot spots" of diversity or endemism. Globally, most attention has been focused on tropical systems and recently Mediterranean systems. Temperate forest-dominated systems have received less attention. For example, of Conservation International's (CI's) 24 hot spots, only New Zealand is a temperate zone forest-dominated locality. However, in his most recent iteration of "megadiversity countries," Mittermeier highlights 4 of 17 countries that contain important temperate forest biomes: China, the United States, Australia, and South Africa. These of these (Australia, China, and South Africa) rank among the top 12 countries worldwide in species richness across phyla and species endemism. The four largely temperate countries also rank among the top globally in the number of IUCN Red Data Book (RDB) "threatened" plant species. Although these predominantly temperate florals are better known, as a group they have more RDB species than all of the tropical countries together. The WWF and the IUCN have also identified approximately 240 centers of plant diversity globally. Of these, about 15% represent temperate forest-dominated systems and another 10% represent other temperate biomes (e.g., shrublands, grasslands, and arid lands). Only 1 or 2% are boreal or polar, and the balance (~74%) are tropical.

There have been similar attempts to identify hot spots regionally. For example, WWF confirmed more than 100 forest hot spots across Europe, the Mediterranean Basin, and the Middle East. Regional centers of diversity can also be detected from taxon turnover. Data from the Atlas Florae Europaeae project, which compared the joint diversity of Pinaceae and Fagaceae across Europe at a 50-km grid scale, show highest joint species diversity in these two families in the Balkans, the southern Alps, the Carpathians, and the southern Pyrenees. This obviously reflects the high spatial heterogeneity in these localities.
Alternatively, conservation assessments may include such factors as taxon or clade irreplaceability, minimum area sets, minimum viable niche space, or ecosystem integrity. A study was done by the British NHM to select conservation priority areas for selected plant groups. This exercise was done on a 10° grid globally and regionally for all of Europe on a 50-km grid. The global analysis revealed top-priority sites in eastern China, southeastern Australia, and central Chile along with six tropical areas, plus secondary centers in southern Chile, the southeastern and western United States, central and southern China, northern Japan, southwestern Europe, eastern and southeastern Australia, and southern Africa along with nine secondary centers in the tropics. Within Europe, many priority sites were identified in the Balkan Peninsula and margins of the Mediterranean (which include forested lands). However, there are prioritized areas elsewhere throughout Europe at lower densities.

World Resources Institute (WRI) provides another perspective on priority sites for forest conservation. They recognize “frontier forests”—large tracts of intact forested ecosystems sufficient to maintain viable populations of all indigenous species. This perspective includes large, wide-ranging predators and migratory species and takes into account the prevailing natural disturbance patterns. These are thus considered intact, fully functional forested ecosystems or landscapes. Very few of these are in the temperate zone, and most are under medium or high threat. This contrasts with the tropical and boreal zones, in which there are many more large tracts identified as frontier forest, even if there are substantial areas at risk. No intact forested landscapes are found in temperate Europe or Africa. In North America only one small patch of transitional boreal forest in central Ontario and discontinuous tracts of conifer forests in coastal British Columbia and Alaska (all under threat) can be considered intact landscapes. In Asia, only a few small forest patches in the inaccessible mountains of central and south-central China, boreal transition forest patches along the border between China and Russia, and the Primorski Krai region of Pacific Russia are considered intact temperate forest landscapes. All but one small patch are under threat. In South America the only frontier temperate forest is in southern Chile and adjacent Argentina. Most of this region is considered to be at risk. In Australia, only a small rainforest patch in Tasmania is classified as intact. Small patches of frontier forest occur on the west coast of the south island of New Zealand and one patch in the central north island. All are at risk except for the Tasmania rainforest.

X. MAJOR THREATS TO CONSERVATION

Many major threats exist for conserving the biodiversity and ecosystems services provided by temperate forests locally, regionally, and globally. Certainly a major threat is homogenization of the landscape. This is a result of intensively managing forests as near monocultures of either native or exotic species, managing landscape for similar forest age and size classes, introducing invasive exotic species, and relying globally on only a few taxa for forest plantations (e.g., P. radiata or Eucalyptus). The consequences of these landscape management strategies are reductions in local or regional diversity and alternation of many ecosystem processes (e.g., nutrient and hydrological cycles) and soil attributes. Large-scale increases in timber harvesting may well have a negative impact on carbon source–sink relationships and atmospheric CO₂ levels.

Increasing urbanization of the landscape and continued harvesting of ever more remote forests will lead to a more fragmented landscape. The likelihood of these fragments maintaining ecosystem functionality is ever diminished. Within the next few decades, it is likely that no fully intact, functional forested landscapes will remain in the temperate zone, given current trends.

Atmospheric pollution from acid precipitation, and the associated ecosystem N and S loadings, remain prevalent in the north temperate zone and will have long-term effects. There is evidence that calcium is being rapidly depleted in acidic soils as a consequence of long-term acid precipitation. This will have cascading effects on forest soils and on major tree species performance, particularly those most sensitive to calcium levels such as sugar maple.

Projected global warming will have a major effect on northern temperate forests. The main direct effect of global warming will likely be seen in the boreal forest zone, with decreased periods of snow cover and hence changes in surface albedo. This will likely favor the expansion of temperate forests well north into the boreal zone. The community response patterns will be determined by dispersal characteristics of the biota and the availability of propagules in source populations. This in turn will be dictated by the configuration of forest fragments in the landscape.

In the Southern Hemisphere, the effects of global warming will be much smaller due to the moderating maritime influences. Atmospheric pollution effects are very negligible or very localized. The major threats to temperate forests here are deforestation, especially in
South America, and the pervasive effects of introduced exotic species in plantation or as escapes. *Eucalyptus* and northern pines (especially *P. radiata*) are planted in monocultures across the Southern Hemisphere, with negative effects on biodiversity, soil attributes, and hydrology.

**XI. CONSERVATION OBJECTIVES AND RESEARCH NEEDS**

Where do we need to focus attention to improve our understanding of temperate forests and develop a more effective conservation plan? Clearly we need better and more accurate means of mapping temperate forests and assessing their conservation status. There are large discrepancies and errors in assessing the current and potential forest cover for all forest types. We need more information on patterns of species occurrences (simply the presence/absence is sufficient) at a variety of scales throughout the temperate zone. Because distribution (and hence richness) patterns vary across spatial scales, one cannot simply rely on particular, arbitrary sample sizes (e.g., 0.04-, 0.1-, 1-, or 50-ha plots). Effort should be placed on inventorying poorly known areas, such as the Middle East. We need to have a better, predictive understanding of the links between forest structure and composition, ecosystem functions, and the effects of human disturbances.

There needs to be a concerted, manifold effort at developing a variety of different conservation strategies for temperate forests. These include targeting more protection for forested biomes with high ratios of people to forest area and forest biomes or forest types that are poorly protected. Examples include east Asian forests in general, north temperate broad-leaved rainforests, and subtropical dry forests. Moreover, concerted effort should be placed on protecting the few remaining intact frontier forests, especially those at high risk, and old-aged forest stands particularly where these comprise miniscule components of the landscape. However, the strategy needs to be inclusive, focusing on conserving the complete spatial and temporal heterogeneity of the landscape, even if this necessitates some human husbandry of the landscape.

Strategies for identifying priority conservation sites should not simply target sites with high species diversity and/or endemism. Alternative means of evaluation need to be incorporated as well, such as phylegetic richness and irreplaceability, range rarity, ecosystem or landscape integrity, landscape heterogeneity that incorporates migration in the face of natural and human-made disturbances, and minimum niche space for all biotic components of the landscape. Only with a manifold creative approach can we hope to come to grips with the conservation of this critical suite of temperate forest biomes.

**See Also the Following Articles**

- BOREAL FOREST ECOSYSTEMS
- DEFORESTATION AND LAND CLEARING
- ENDEMISM
- FOREST ECOSYSTEMS
- HOTSPOTS
- TEMPERATE GRASSLAND AND SHRUBLAND ECOSYSTEMS
- TROPICAL FOREST ECOSYSTEMS

**Bibliography**


TEMPERATE GRASSLAND AND SHRUBLAND ECOSYSTEMS

Osvaldo E. Sala, Amy T. Austin, and Lucia Vivanco
University of Buenos Aires and IFEVA-CONICET

I. Grassland Distribution
II. Extent of Biodiversity in Temperate Grasslands
III. Biodiversity and Ecosystem Functioning in Grasslands
IV. The Future of Biodiversity in Grasslands

GLOSSARY

convention on biological diversity The convention was first enacted in June 1992, and it has been signed by many countries. Its objectives are the conservation of biological diversity, the sustainable use of its components, and the fair and equitable sharing of the benefits arising from the utilization of genetic resources, including by appropriate access to genetic resources and by appropriate transfer of relevant technologies, taking into account all rights over those resources and technologies, and by appropriate funding.

functional type A group of species that share morphological and physiological characteristics that result in a common ecological role.

global biodiversity assessment The Global Biodiversity Assessment is an independent peer-reviewed analysis of the biological and social aspects of biodiversity commissioned by the United Nations Environment Programme.

international geosphere biosphere programme A scientific program that is part of International Council of Science and provides an international and interdisciplinary framework for the conduct of global change science.

niche complementarity Refers to how the ecological niches of species may not fully overlap and complement each other. Consequently, an increase in the number of species that complement each other may result in a larger volume of total resources utilized and in a higher rate of ecosystem processes.

sampling effect Refers to the phenomenon in which increases in the number of species increase the probability of including in the community a species with a strong ecosystem effect. This phenomenon yields an increase in ecosystem processes with increases in diversity without invoking niche complementarity.

scientific committee on problems of the environment An international organization that is part of the International Council for Science and is charged with synthesizing current scientific understanding associated with environmental issues.

GRASSLANDS are water-limited ecosystems, and water availability defines the distribution of grasslands. Grassland ecosystems occur in areas of the world that have an annual precipitation between 150 and 1200 mm.
and mean annual temperature between 0 and 25°C. Temperature controls the distribution of grasslands mainly indirectly by modulating water demand and consequently water availability. Increases in temperature result in increases in soil evaporation and plant transpiration; consequently, for a similar precipitation regime, the water balance becomes more negative as temperature increases. In contrast with most biological phenomena, primary production of North American grasslands decreases with increasing temperature, highlighting the indirect mechanism of the temperature control on the distribution of grasslands.

I. GRASSLAND DISTRIBUTION

How does the distribution in temperature and precipitation space translate into the distribution of grasslands in geographical space? Along precipitation gradients grasslands are located between forests and deserts. Several of the International Geosphere Biosphere Programme terrestrial transects (Koch et al., 1995) are located along precipitation gradients and they intersect forests, grassland–forest ecotones, grasslands, and deserts. This pattern is repeated in the North Eastern China transect, in the Great Plains transect in North America, and in the Patagonian transect in South America (Fig. 1).

Although temperature and precipitation are the major determinants of the distribution of grasslands, fire also may play an important role. Fire becomes particularly important in the grass–forest ecotones where the dominance of grasses or woody plants in many cases is determined by the frequency and intensity of fires. For example, in North American tallgrass prairie, the area covered with woody plants has increased dramatically in the past 100 years and the human intervention in reducing fire frequency is largely responsible for the change (Briggs et al., 1998). Similarly, data from pollen profiles, tree ring analysis, and photographic sources documented a shift in the grassland–forest ecotone in northern Patagonia with woody vegetation invading the grassland (Vehlen and Markgraf, 1988). Again, changes in the human-induced fire regime were responsible for the forest expansion.

This article focuses exclusively on climatically determined grasslands, in contrast with grasslands resulting from human intervention. Anthropogenically determined grasslands are located in areas where potential natural vegetation is forest. Humans, in an attempt to produce forage for domestic animals, have logged forests and have maintained these plots as grasslands by mowing them periodically.

Worldwide, temperate grasslands are represented in all continents and potentially cover a vast area of $49 \times 10^6$ km$^2$ that represents 36% of the earth's surface.
TEMPERATE GRASSLAND AND SHRUBLAND ECOSYSTEMS

(Shantz, 1954). Our definition of temperate grasslands excludes tropical and subtropical savannas but includes grass and shrub deserts. In North America, grassland is the potential natural vegetation of most of the Great Plains and it reaches from the Chihuahuan desert in the south to the deciduous forests of Canada in the north and from the Rocky Mountains in the west to the deciduous forest of the eastern United States (Fig. 1). In South America, grassland is the potential vegetation of the vast pampas and most of the Patagonian steppe. Finally, in Asia, grassland ecosystems cover a huge area from the Ukraine to China.

II. EXTENT OF BIODIVERSITY IN TEMPERATE GRASSLANDS

Biodiversity can be examined in many different ways, and multiple definitions exist for what constitutes "biodiversity." Nevertheless, the authors of the Global Biodiversity Assessment, using definitions originally proposed by the Convention on Biological Diversity, defined biological diversity as "variability among living organisms from all sources" (Heywood and Baste, 1995). Here, we focus on biodiversity in terms of taxonomically defined species, and the vast majority of studies quantifying ecosystem variation have used this measure. However, genetic biodiversity (genetic variation among a single species) and ecological diversity (including landscape diversity and functional group diversity) are also important components of biological diversity. The definition of biodiversity, therefore, will depend to a certain extent on one's objective and scale of interest, ranging from the gene to the ecosystem.

A. Diversity of Plants

Floristic diversity in grasslands varies broadly, with many natural grasslands having a very high level of plant diversity, at times approaching that seen in mainland tropical forests (Groombridge, 1992). For example, the Pampa region in Argentina represents some of the highest diversity grassland, with more than 400 species of grasses (Cabrera, 1970). In North America, more than 250 native species are found in tallgrass prairie (Freeman, 1998), the vast majority of which are perennial grasses.

Plant species of grasslands can be categorized into four functional types: grasses, shrubs, succulents, and herbs (Sala et al., 1997). The classification of plant species into functional types only has epistemological value and serves the purpose of facilitating the study of grasslands. This classification can be divided into many new subcategories or aggregated into fewer units depending on the needs of the analysis. The relative contribution of the four functional types depends on the seasonality of precipitation and the soil texture, which are the factors controlling the distribution of water availability in the soil profile. Water penetrates deeper into the soil profile in coarse-textured soils than in fine-textured soils. Similarly, water penetrates deeper into the soil in regions in which most of the precipitation occurs during the cold time of the year when evaporative demand is low. In general, grasses and shrubs have contrasting rooting patterns, with grasses having predominantly shallow roots and shrubs having deep roots (Jackson et al., 1996). Consequently, grasses dominate in regions in which the wet and warm seasons are in synchrony and in areas with predominantly fine-textured soils (Sala et al., 1997, Fig. 2).

Grassland plant species can also be classified according to their photosynthetic pathway into C3 and C4 species. The two groups of species have differences in the physiology of photosynthesis and in the morphology of leaves that result in different ecological characteristics that separate them in time and space. Regional analyses of the distribution of these two types of grass species showed that C3 species decrease southward in North America and northward in South America and C4 species show the opposite pattern (Paruelo et al., 1997).

---

**FIGURE 2** Conceptual model explaining the distribution of the functional groups grasses and shrubs with changes in seasonality of temperature and precipitation (congruence of warm and wet seasons) and soil texture. The gray shaded area represents conditions that are likely to favor the persistence of grasses, whereas the speckled area represents conditions that favor the shrub functional group. The intersection of the two areas represents points at which biotic influences are likely to be most pronounced (adapted from Sala et al., 1997).
Similarly, the abundance of C4 species increases whereas that of C3 decreases along an altitudinal gradient (Cavagnaro, 1988). These biogeographical analyses correlate with ecophysiological studies showing that C4 species have photosynthesis optima at higher temperature, have higher water use efficiency, and are better adapted to low water availability conditions (Kemp and Williams, 1980).

B. Aboveground Diversity of Animals

All major taxonomic groups are represented in grasslands, but despite their large areal extent (36% of the area of terrestrial ecosystems), overall faunal diversity is lower than in many other biomes. The number of bird and mammalian species that are found primarily in grasslands are estimated to be 477 and 249, respectively, representing only 5% of the world's species for each taxonomic group (Groombridge, 1992). Local diversity can be high in specific areas (e.g., there is an estimated 208 avian species for tallgrass prairie; Kauffman et al., 1998), but general patterns show lower diversity for most taxonomic groups compared with other ecosystems.

One of the striking features of grasslands in terms of animal diversity is the presence of large herbivores as a prominent component of secondary production. These large grazing mammals have an important impact on the functioning of grasslands, altering patterns of nutrient cycling, primary production, and plant species composition (McNaughton, 1993), although their presence and diversity varies across different continents. For example, in the Great Plains of North America, nearly all the large grazing mammals went extinct during the glaciation of the Pleistocene, but the proliferation of a very few species, particularly Bison bison (plains bison), dominated the plant–herbivore interactions until the introduction of domestic cattle at the beginning of the twentieth century (Lauenroth and Milchunas, 1992). In contrast, African grasslands contain a very high level of mammalian diversity of grazers, with up to 20 species coexisting in a single reserve (Cumming, 1992). Finally, many South American grasslands evolved without the presence of large grazers, and their primary herbivory prior to the introduction of sheep and cattle was due to insect species (Bucher, 1982). Thus, although there is variation in the diversity of the large herbivores, their presence and importance is a distinctive characteristic of grassland ecosystems.

Small mammals, birds, reptiles, amphibians, and insects also play an important role in the functioning of grasslands. Species richness of small mammals is actually higher than that for large mammals (168 vs 77 species overall), and they are mostly granivores or omnivores (Groombridge, 1992). In contrast, in Australian deserts small mammals are mostly insectivorous. Fluctuations in seed supply caused by unpredictable environmental conditions and the infertile soils could be an explanation for these differences (Morton, 1993).

Avian diversity in grasslands represents 5% of the total species of the world species diversity, and again the fluctuating climate has an important control on this distribution. In this case, birds can migrate to remote areas outside of the grassland biome to seek alternative resources in periods of unfavorable conditions. In North American grasslands, which demonstrate a strong seasonality, there are large annual variations of passerines in response to climatic conditions. Additionally, within the grassland ecosystems, there exists a gradient of avian biomass that decreases with precipitation and primary production from the tallgrass prairie to the mixed prairie and shortgrass steppe (Lauenroth and Milchunas, 1992).

Reptiles in grasslands are less diverse than mammals and birds, and amphibians are less diverse than reptiles in the tallgrass prairie of North America (Kauffman et al., 1998). Latitude has an effect on the biodiversity of reptiles and amphibians because they are ectothermic organisms, with an increase in the number of species from north to south of the tallgrass prairie (Kucera, 1992).

Insects are a diverse element of the terrestrial macrofauna of tallgrass prairie (Kauffman et al., 1998), reflecting general patterns of diversity for terrestrial ecosystems in which insects represent more than 50% of the species (Strong et al., 1984). They have a very important role as herbivores, pollinators, predators, parasitoids, and decomposers. Herbivorous insects are probably the most conspicuous functional group in tallgrass prairie (Kauffman et al., 1998) and may replace large grazing mammals as the primary consumer in some South American grasslands (McNaughton et al., 1993).

C. Diversity of Soil Organisms

The diversity of belowground organisms (bacteria, fungi, and micro-, meso-, and macrofauna) is known in much less detail than that of plants and animals aboveground, but it may constitute a very important component of the biota, often equaling or exceeding the aboveground biomass in grasslands (Paul et al., 1979). The large biomass of roots and other underground organs in grasslands and the high concentration of organic matter provide substrate for a large variety
III. BIODIVERSITY AND ECOSYSTEM FUNCTIONING IN GRASSLANDS

The relationship between biological diversity and the functioning of ecosystems has been central in the development of ecological ideas, and grassland ecosystems have been crucial in testing these ideas. McNaughton (1993) traced the idea of the relationship between biodiversity and ecosystem functioning to statements by Charles Darwin about how increasing species diversity in a plot might result in higher productivity as a result of niche complementarity. Most species in Darwin’s plot supposedly exploited different resources, and consequently the larger the number of species, the larger the volume of resources exploited. For example, plots containing only shallow-rooted or deep-rooted species should have lower productivity than plots containing both groups of species that jointly have access to water and nutrients stored in both upper and lower layers of the soil.

Species not only differ in their ability to exploit resources but also in their response to the environment. Species show a variety of responses to abiotic factors such as temperature or water availability as well as biotic factors such as predation or competition. These relationships led to the diversity-stability hypothesis or portfolio hypothesis that stated that higher species diversity results in greater ecosystem stability (McNaughton, 1993). Consequently, ecosystem processes will vary more in space or time in less diverse than in more diverse communities.

The basic theory of the relationship between biodiversity and ecosystem functioning has been established for decades, but its empirical support has been scarce and fragmentary. Pimm’s (1989) review of experimental results and later the Scientific Committee on Problems of the Environment (SCOPE) project led by Hal Mooney (Schultz and Mooney, 1993) reinvigorated the field and highlighted the gaps in our understanding. The value of possessing a quantitative understanding of the relationship between biodiversity and ecosystem functioning is underscored by the fact that we are in the midst of the sixth major extinction event in the history of life (Chapin et al., 1998). The major difference between this and previous extinction episodes is that the current wave of extinction is the result of human activity. Society is urged to identify the consequences of the human-driven changes in biodiversity for the functioning of ecosystems that provide so many goods and services, including grasslands, that are at the basis of human well-being (Sala and Paruelo, 1997).

A large SCOPE project had the responsibility of assessing and evaluating the vast number of observations and experiments that addressed in different ways the role of biological diversity on the functioning of ecosystems. An initial meeting and corresponding volume synthesized the theory and formulated a plan of action (Schulze and Mooney, 1993). The second step was to address the same questions on a biome by biome basis. Finally, a large synthesis assessed the differences and commonalities among biomes in the relationship between biodiversity and ecosystem functioning (Mooney et al., 1995a, b, 1996). Results for the grassland biome emphasized the effects of changes in biodiversity on primary productivity, decomposition, water balance, atmospheric properties, landscape structure, and species interactions (Sala et al., 1995, 1996). This effort synthesized our knowledge and it also highlighted the gaps in our understanding and the need for further experimentation.

Grassland ecosystems played a key role in the experimentation regarding the functional roles of biodiversity mostly because of the small size and short life span of grasses, which made manipulative experiments feasible with few resources and in short periods of time. The most widely accepted hypothesis indicates that the rate of ecosystem processes, such as primary productivity or nutrient cycling, might increase linearly as species richness increases and that this relationship eventually saturates as ecological niche overlap increases (Fig. 3) (Vitousek and Hooper, 1993):
FIGURE 3  Theoretical relationship between biodiversity and ecosys-
tem functioning. Species richness (number of species) is the most
common indicator of biodiversity, although this axis term could
also include genetic diversity (diversity within species) to landscape
diversity. Ecosystem process could signify processes occurring at the
ecosystem level, including primary production and decomposition,
and components of nutrient cycling (adapted from Vitousek and
Hooper, 1993).

The first experiment specifically designed to test this
hypothesis was carried out under controlled conditions
at the Imperial College Ecotron facility (Naeem et al.,
1994). The experiment used synthetic communities
with three trophic levels and three levels of species
diversity, from the simplest with 1 species of a second-
ary consumer, 3 primary consumers, 2 producers, and
3 decomposers to the most diverse that had 31 species.
The results of this experiment supported the hypothesis
described in Fig. 3 in that carbon fixation (an indirect
measure of primary productivity) increased with diver-
sity. The first large-scale field experiment specifically
designed to test this hypothesis was located in the North
American tallgrass prairie (Tilman et al., 1996). The
experiment consisted of sowed plots with seven levels of
plant diversity and 20 replicates per level. Each replicate
was a random draw from a pool of 24 native species. Consequently, replicates had the same diversity level
but could have a different or equivalent species compo-
sition. Results follow a pattern similar to that of Fig.
3, with total plant cover (a nondestructive way of esti-
mating light interception and production) and nutrient
uptake increasing with species richness up to a level
of approximately 10 species. Beyond 10 species, the
different estimates of ecosystem functioning did not
change. Recently, a large consortium of scientists orga-
nized a large-scale field experiment across Europe with
the purpose of testing the same hypothesis (Hector et
al., 1999). The same experimental design was repeated
in eight locations in Europe from Sweden to Greece and
it consisted of five diversity treatments with richness
ranging from 1 to 32 species. Species identity varied
among the sites since species for each treatment were
always drawn from a pool of species adapted to local
conditions. A single model represented the loglinear
increase of aboveground biomass with increasing spe-
cies richness for all sites.

The different experiments designed to test the effect
of diversity on ecosystem functioning yielded results
that are similar to those predicted from the hypothesis
described in Fig. 3. These results can be interpreted as
evidence of niche complementarity; that is, the higher
the number of species with niches that do not overlap,
the larger the total volume of resources exploited (Til-
man et al., 1997). Alternatively, the same results can
be interpreted as resulting from the sampling effect
(Huston, 1997; Tilman et al., 1997). It indicates that
as the number of species increases, the probability of
including in the mix a species with strong ecosystem
effects increases. This species may be a nitrogen fixer,
a deep-rooted species, or simply a species with a combi-
nation of characters that maximizes production in these
circumstances.

IV. THE FUTURE OF BIODIVERSITY
IN GRASSLANDS

Biodiversity in grassland ecosystems is seriously threat-
eted by human activity. Grassland ecosystems, in two
of three possible scenarios of biodiversity change for
the Year 2100, appear to be the most threatened biome;
in the third scenario, grasslands appear behind only
tropical forests, arctic ecosystems, and southern tem-
perate forests (Sala et al., 2000). What makes biodiversity in grassland ecosystems so
vulnerable to human impacts? Are grassland ecosystems
particularly sensitive? Are they located in areas that
will be affected the most? A recent study identified
the most important drivers of biodiversity change in
grasslands as changes in land use, climate, nitrogen
deposition, biotic exchange (accidental or deliberate
introduction of plant or animal species to an ecosys-
tem), and atmospheric CO2, and studied their expected
change and the sensitivity of each biome (Sala et al.,
2000). The study concludes that grasslands are both
quite sensitive ecosystems and are located in parts of
the world where ecosystems are going to be affected
most by human activity (Fig. 4).

The expected change of each of the drivers was as-
sessed using other scenarios, such as climate change or
land use change, that were developed independently
(Alcamo, 1994; Haxeltine and Prentice, 1996). The dif-
different drivers of biodiversity change were originally expressed in different units from parts per million to hectares and degrees Celsius. In order to compare the rate of change among drivers, they were all converted into a 1 to 5 arbitrary scale. Land use, nitrogen deposition, and biotic exchange are expected to be the drivers that are going to change the most in grassland ecosystems (Sala et al., 2000) (Fig. 4A). The most dramatic changes in land use in grasslands are those that result in changes in land cover, such as the conversion into croplands. Grasslands are among the biomes that are going to experience the largest conversion in land use because of their mild climate and favorable soil conditions that made them quite suitable for agriculture. The conversion into agricultural land is not expected to be even across the world but driven by the patterns of food demand and population growth that indeed are quite idiosyncratic. For example, the IMAGE2 model (Alcamo, 1994) predicts for the Year 2100 a large increase in agricultural area in Africa and a reduction in North America resulting from an increase in demand and an increase in intensification, respectively. Biodiversity losses resulting from conversion to agriculture in one part of the world are not offset by a similar area that will be abandoned and reverts to grassland but that is located in a different part of the world. Therefore, total change in grassland area grossly underestimates the impact of land use change on biodiversity. Finally, the expected change of nitrogen deposition and biotic exchange in grasslands are among the highest of all biomes. Densely human-populated regions are predominantly located in temperate regions where the potential native vegetation is that of grasslands. Nitrogen deposition is associated with industrialization, whereas biotic exchange is associated with trade patterns and ultimately with human density.

Sensitivity to changes in each driver is the magnitude of change in biodiversity due to a change in a unit of driver. Sensitivity varies among biomes, and grasslands are quite sensitive to changes in land use, climate, and CO2 concentration in the atmosphere (Fig. 4B). Grasslands are most sensitive to land use change, which means the conversion into croplands that implies plowing of native grasslands and sowing of a monospecific crop. This activity clearly results in the local extinction of all the native plant species that, in turn, determine the major characteristics of the habitat of animals and microorganisms. Consequently, land use change drives all plant species to local extinction and drastically affects the diversity of other organisms (Anderson, 1995). Biodiversity may be quite sensitive to changes in CO2 concentration in ecosystems that are limited by water availability such as grasslands. There are well-known differences in the species response in water use efficiency due to changes in CO2 (Jackson et al., 1994). Consequently, changes in CO2 will first affect the com-
 touted balance and then the relative abundance, and they may result in the local extinction of species. The biodiversity change resulting from each driver (Fig. 4C) can be calculated as the product of the expected change (Fig. 4A) multiplied by the biome sensitivity to that driver (Fig. 4D). The largest biodiversity change for the Year 2100 in grasslands is expected to occur due to changes in land use. The second largest effect will result from the expected increase in CO₂ because although the atmospheric concentration of this trace gas will increase uniformly throughout the world (Fung et al., 1987), grassland biodiversity may be particularly sensitive to it.

Finally, the total biodiversity change in grasslands expected for the Year 2100 can be calculated as the sum of the individual effects of each driver (Fig. 4C). This calculation assumes no interaction among drivers. Comparison of the total biodiversity change in grasslands against that of all the other terrestrial biomes of the world indicates that grasslands and Mediterranean ecosystems appear to be the biomes that will experience the largest change and will be affected the most (Sala et al., 2000).

Other scenarios consider the assumption that there are synergistic interactions among the drivers of biodiversity change. We can envision several examples of the synergistic interactions among drivers. For example, the effect of biotic exchange will be amplified if combined with a land use change that results in fragmentation. Similarly, the effects of elevated CO₂ on biodiversity when occurring in conjunction with increased nitrogen deposition will be much larger than the sum of the individual effects. The scenario developed using the assumption of a synergistic interaction among drivers also indicated that grasslands and Mediterranean ecosystems are the most vulnerable (Sala et al., 2000). The only scenario in which grasslands are not the most vulnerable ecosystem is when an antagonistic interaction was assumed and the total biodiversity change was equated with the change resulting from the driver with the maximum value. This scenario is plausible only under extreme conditions, such as those of slash-and-burn that destroy most of the plant and animal species, and further change due to the other drivers is not possible. Even in the antagonistic scenario, grasslands are among the most vulnerable biomes, following tropical, arctic, and southern temperate ecosystems.

Acknowledgments

We thank J. P. Gauchenschm for assistance with the figures. Research by OES reported in this chapter was supported by IAI, CONICET, FOCEyT and the University of Buenos Aires. ATA was supported by Grant CRN 612 from the Tucan-American Institute for Global Change Research. LV was supported by a fellowship from the University of Buenos Aires.

See Also the Following Articles

CARRON CYCLE • PLANT BIODIVERSITY, OVERVIEW • SLASH AND BURN FARMING, EFFECTS OF • TEMPERATE FORESTS

Bibliography


J. F. Guebertschm, for assistance with the figures. Research by OES reported in this chapter was supported by IAI, CONICET, FOCEyT and the University of Buenos Aires. ATA was supported by Grant CRN 612 from the Tucan-American Institute for Global Change Research. LV was supported by a fellowship from the University of Buenos Aires.

See Also the Following Articles

CARRON CYCLE • PLANT BIODIVERSITY, OVERVIEW • SLASH AND BURN FARMING, EFFECTS OF • TEMPERATE FORESTS

Bibliography


Kemp, P. R., and Williams, G. J. (1980). A physiological basis for niche separation between Agropyron smithii (C3) and Bouteloua gracilis (C4). Ecology 61, 848–858.


Primary Productivity of the Biosphere. Springer-Verlag, New York.


I. The Earth’s Terrestrial Ecosystems

A. Ecosystem Classifications: Overview

Terrestrial ecosystems cover approximately 148 million km², corresponding to 29% of the total surface area of the earth. They include such diverse habitats as the frigid regions around the poles, the searing heat of tropical deserts, and lush temperate and tropical rainforests. This enormous variety poses great difficulties for the study of ecosystems at regional to global scales, and it has led to the development of many classification systems which seek to reduce this complexity to a manageable level. Indeed, the development of such classification systems, and the identification of the spatial extent of the Earth’s major ecosystem types, is a fundamental first step in managing the world’s biological systems.

Despite the importance of this problem, the classification of the world’s ecosystems into a consistent and useable framework is not without its problems. First,
ecosystems usually grade into one another, hence mak-
ing it difficult to mark ecosystem boundaries on a map. 
Second, there is no single agreed upon definition of an 
“ecosystem.” To some researchers it represents an area 
within which the same broad mix of species can be 
found. To others it is merely an arbitrary area delimited 
solely for practical purposes. Third, different criteria 
can be used to generate different kinds of classifications. 
For example, human-induced land-use change is in-
cluded in some classifications but not in others. Ecosys-
tem classification systems are therefore approximations 
of reality and should be viewed as a set of tools which 
provide a simplified overview of the relative extents 
and spatial arrangements of the world’s major ecosys-
tem types.

Ecosystems are usually classified according to the 
dominant vegetation type for a region, which is in turn 
determined largely by geographical and environmental 
factors. Vegetation is considered the most appropriate 
criteria for classifying ecosystems because it constitutes 
the majority of biomass and also because of its role, 
through the process of photosynthesis, as the primary 
fixer of energy on which all other life depends.

B. The Biome Classification

The broadest units of ecosystem classification are called 
biomes. Biomes are complex mixtures of a large number 
of similar ecosystem types which share a similar physi-
ognomy (vegetation structure) and climate but do not 
necessarily share the same species. Indeed, one of the 
strengths of the biome approach is that ecosystems with 
different species compositions, but with similar vegeta-
tion structures and ecosystem processes, can be com-
pared. For example, the tropical rainforests of Central 
America and northern Australia are both recognized as 
belonging to the same biome “tropical rainforest,” even 
though there would be little overlap in species composi-
tion between the two regions. The major climatic vari-
able on which the biome classification is based are 
average annual temperature and average annual precipi-
tation.

Many biome classifications have been suggested. 
These have classified the world into as little as 8 and 
into as many as 18 separate biomes. Figure 1 shows 
the classification of Cox and Moore, which maps 10 
major terrestrial biomes. One feature of the spatial dis-
tribution of the world’s biomes is that many are frag-
mented over the surface of the earth. For example, the 
“Mediterranean” biome is found in the areas sur-
rounding the Mediterranean sea and also in Australia, 
South Africa, South America, and North America. 
A brief description of the major terrestrial biomes is 
given in the following sections. More complete accounts 
of each can be found in the appropriate sections of this 
encyclopedia.

1. Tundra

Tundra occurs around the Arctic Circle at latitudes 
above where trees can survive and also at high altitudes 
above the tree line in other regions. The vegetation is 
of a very low stature, typically comprising low shrubs, 
seeds, grasses, lichens, and mosses. The climate is 
characterized by long, cold winters, during which aver-
age monthly temperatures seldom exceed 0°C and, at 
the highest latitudes, there is constant darkness for 
weeks at a time. Soils are permanently frozen (perma-
frost), making water unavailable to plant growth. Also, 
liquid water is available for only short periods at a 
time, thus making these ecosystems effectively “polar 
deserts.” Due to these harsh climatic conditions and 
brief growing season (usually less than 60 days), plant 
growth and decomposition is slow and ecosystems are 
very slow to recover from disturbance. Animal diversity 
is similarly low, with a small number of permanent 
bird and mammal species and a number of seasonal insects. 
Many bird species and some mammals migrate to the 
tundra during the brief summer.

2. Taiga (Boreal Forest and Northern 
Coniferous Forests)

The taiga forms a broad band adjacent to the tundra 
biope, extending across North America, northern Eu-
rope, and northern Asia. The dominant vegetation con-
sists of evergreen conifers, with some representation 
of deciduous broadleaved species in some environments. 
Within this biome it is common to find large areas of 
land dominated by only one or a few coniferous tree 
species. Winners are typically long and cold, contrasting 
with summers that, although brief, can be both warm 
and wet. The soils are characterized as being nutrient 
poor and acidic. Characteristic animals include insect-
and seed-eating birds, small mammal herbivores such 
as the snowshoe hare, larger browsers such elk and 
moose, and carnivorous mammals such as the lynx, 
wolf, and grizzly bear. Many bird species migrate to 
the taiga during the summer to feed on swarms of 
seasonally breeding insects.

3. Temperate Forests

The temperate forest biome includes a wide range of 
forest types, including the temperate rainforests of the 
Pacific Coast of North America, southern Chile, and 
New Zealand; the temperate deciduous forests of east-
ern North America, eastern Asia, and Europe; and tem-
perate evergreen forests such as the sclerophyllous eu-
calyptus forests of Australia and the southern beech forests of Chile and New Zealand. The forest structure is generally well developed, with various strata including emergent trees, canopy trees, small trees, shrubs, and a herbaceous layer. Along with an enhanced diversity of plant species and habitat structure, there is an associated general increase in the diversity of the fauna. The growing season is approximately 6 months. Because of the wide range of forest types and environments represented in this biome, soil nutrition, precipitation, and water availability are also highly variable.

4. Tropical and Temperate Grasslands

The grassland biome includes ecosystems from a wide range of geographical regions and environments. In the tropics these include the savanna, which can also be associated with an open cover of shrubs or trees. Examples of temperate grasslands can be found in North America (prairies), Europe and Asia (steppe), South America (pampas), and South Africa (veld). A common feature is the importance of herbivory, particular in the savanna, which has a high diversity of ungulate browsers. In comparison with tropical grasslands, temperate grasslands support a much lower faunal diversity. Fires are also an important component in many grassland ecosystems, in which repeated fires can prevent the establishment of woody species, thus preventing the grasslands from developing into shrublands or forests. Climate and soils vary with grassland type and region, with the tropical savanna experiencing distinct wet and dry seasons and temperate grasslands experiencing cold winters and hot summers.

5. Mediterranean Vegetation (Chaparral)

Examples of the Mediterranean biome can be found between approximately 30° and 40° latitude in both hemispheres and are associated with wet winters and dry summers. Mediterranean vegetation occurs most extensively on the coast surrounding the Mediterranean Sea, but it is also found as the “chaparral” in California and northwest Mexico, the “matorral” in Chile, the “fynbos” around the Cape of South Africa, and also in the southern regions of Australia. Because precipitation in this biome is generally less than in grasslands, and there are regular periods of drought, the vegetation is characterized by shrubs and small trees which have small, leathery “sclerophyllous” leaves adapted to drought stress. Fire is also a common feature of Mediterranean vegetation, and many plant species share adaptations to persist under a regime of repeated fire, e.g., resprouting from belowground and seeds that require fire to germinate. In general, soils are nutrient poor.
One striking feature of this biome is that high levels of endemism are common, but adaptations to the Mediterranean environment in geographically separate regions are very similar. This has been seen as evidence for convergent evolution, in which plant and animal species from different taxonomic groups respond independently, but similarly, in response to the same environmental pressures.

6. Deserts

Deserts occur in climates characterized by extreme aridity. Water limitation in these areas can arise due to many factors, for example, being located in hot tropical regions in which evaporation exceeds precipitation such as the Sahara desert; rainshadow regions in which the presence of a mountain range influences precipitation patterns, resulting in dry conditions on the leeward side of the range, such as in the Patagonian desert in Argentina; and coastal deserts in which prevailing winds prevent precipitation occurring, such as in the Namib desert in southern Africa. Continental-scale effects can also be important, where the interiors of major continents are too distant from moist airflow to receive sufficient rainfall. Vegetation in deserts is sparse, with shrubs being the most common life-form. Deserts are also characterized by succulents (e.g., cacti) and many ephemeral annual and perennial species which have dormancy as some part of their life cycle, enabling them to actively grow and reproduce when rain does occur. The animals of desert regions have also evolved in response to the harsh conditions. Examples include nocturnal activity to avoid the heat of day and dormancy during the periods between rainfall events.

7. Tropical Forests

This biome includes tropical seasonal forests, which occur in humid tropical climates with a pronounced dry season, as well as tropical rainforests. Tropical seasonal forests are characterized by a mix of evergreen and deciduous species, with some forest types dominated solely by deciduous species. In contrast, tropical rainforest thrives in regions in which rainfall is abundant and evenly distributed throughout the year and temperatures are uniformly warm. Tropical rainforests are noted for supporting the highest recorded diversities of both plants and animals. The soils of tropical forests are typically nutrient poor, with decomposition being rapid and with the nutrients stored mostly within the plant biomass.

C. Other Classification Systems

Although large-scale "biome" classification systems are useful in gaining an appreciation of the spatial extent of the world's major ecosystem types, for many applications a finer resolution is often required. Two approaches which aim to increase the detail but still retain a manageable number of categories are considered.

First, Olson's classification is an attempt to map the world's ecosystems as they were in 1980. Included in the classification are natural or seminatural ecosystems (e.g., "Tundra" and "Tropical dry forests") as well as those ecosystems that have undergone human modification (e.g., "Warm crops" and "Paddyland"). Olson's classification recognizes 44 different ecosystem complexes compared with the 10 major biomes in Fig. 1.

A different approach was adopted by Holdridge in his "life zone" classification. This classification aims to predict the potential vegetation that an area could support rather than what ecosystems might actually be occurring there. The classification is based on two main variables, labeled "Biotemperature" and "Annual precipitation." Annual precipitation is the average annual amount of rainfall a region receives, without taking into account its seasonal pattern. Biotemperature is the average annual temperature, excluding temperatures below 0°C and above 30°C. Note that these two climatic variables are essentially the same as those used to define the biomes discussed previously. A third variable in the Holdridge classification, "Potential evapotranspiration ratio," combines precipitation and biotemperature into an index which summarizes the water availability for a particular region and quantities the balance between the amount of water entering the system through precipitation and that lost through the combined processes of evaporation and transpiration. Evaporation is the physical process whereby water is returned directly to the atmosphere from plant and soil surfaces, and its rate is directly influenced by the amount of incoming solar radiation. Transpiration is the biological process whereby soil water is taken up by plant roots, transported through the plant's vascular systems, and evaporated from the leaf surfaces through pores called stomata. Potential evapotranspiration ratios less than 1.0 indicate an excess of water in the ecosystem, and ratios greater than 1.0 indicate a deficit.

The classification produces 38 major life zone categories, ranging from hot-wet tropical forests and hot-dry temperate and tropical deserts to cold-dry polar deserts (Fig. 6).

The Holdridge classification has been used in global vegetation models to predict major changes in ecosystems types in response to global climate change, and Olson's classification has been similarly used to model the effects of global change, and in particular the dynamics of the global carbon cycle and carbon storage.
II. GLOBAL PATTERNS OF BIODIVERSITY ACROSS TERRESTRIAL ECOSYSTEMS

A. Description of Known Patterns

One of the best documented patterns of biodiversity is the tendency for tropical ecosystems to harbor more species than temperate or polar ecosystems, i.e., there is a gradient in diversity stretching from lower latitudes (high diversity) to higher latitudes (low diversity). Some examples of different terrestrial organisms which show this pattern are shown in Fig. 2. The trend is also seen in a wide range of marine organisms.

It is worth noting that these examples include only a subset of the total biodiversity for any given region, e.g., land-breeding birds or swallowtail butterflies. Furthermore, none of these examples are global in extent, with two of the four (landbirds and mammalian quadrupeds) being based on data from only a single continent.

The reason for the focus on small groups of species, at subglobal scales, is simply because the data required to perform global-scale, multispecies studies of biodiversity are still incomplete. Nevertheless, despite some minor exceptions, the majority of studies that have investigated latitudinal trends in diversity have shown that tropical regions are indeed characterized by high levels of species diversity.

Although distributional information for groups of individual species at the global level is in general poorly known, that for the higher taxonomic groupings (e.g., families) is considered more complete. One example of this approach, compiled by Williams and coworkers, is shown in Fig. 3. It shows the global distribution of the combined number of families of flowering plants, amphibians, reptiles, and mammals. Together, these four groups of families are considered to be the most completely known. It has been estimated that for the chordates 90% of all living species have been identified and for plants 84.4%. These values are far in excess of any other taxonomic group. Figure 3 shows that the trends at the family taxonomic level mirror those of individual species, with a concentration of families in the tropics, particularly around Central America and Southeast Asia, decreasing toward the higher latitudes in each hemisphere.

Figure 4 shows the relationship between the distribution of the family richnesses in Fig. 3 relative to the life zone classification of Holdridge. There is a consistent pattern, with warm/wet tropical ecosystems containing many more families than either the hot-dry (desert) or cold-wet (polar) ecosystems. Incorporating the spatial information of Fig. 3 into the Holdridge classification allows the trends in family richness across different ecosystem types to be displayed visually. The advantage of this approach, over traditional richness vs latitude relationships, is that different ecosystem types occurring at the same latitude but with different richnesses are able to be separated. For example, tropical rainforests and the Sahara desert both occur at tropical latitudes, but the latter is conspicuously lower in diversity.

A second major trend in terrestrial biodiversity is a decrease in diversity with increasing altitude. As altitude increases there is an associated change in the environment that is analogous to the changes that occur when moving from tropical to temperate latitudes. For example, temperature decreases with increasing altitude, and the total land surface area also tends to decrease with altitude (simply because mountains tend to be broader at the base than at the peaks). Because these changes appear to mirror those observed in the tropical to temperate gradient, many researchers have suggested that the factors controlling the pattern of lower diversity at higher altitudes are likely to be the same as those controlling the latitudinal gradient.

B. Possible Explanations

Although many plausible hypotheses have been suggested to explain the observed latitudinal trend in diversity, many of them have proved to be controversial and there is no consensus on which of the explanations, if any, are of primary importance in determining actual diversity gradients. A major source of the difficulty is that most of the explanations are based on correlative evidence, and it is well-known that correlation does not necessarily imply causation. Correlative evidence predominates because appropriate experimental tests are extremely difficult to devise at the spatial and temporal scales required.

For any given area on the earth, the number of species present (S) is due to many processes acting over a range of spatial and temporal scales. These processes can be summarized into three broad categories and combined into the following simple formula:

\[ S = \text{species which have immigrated from outside the area (dispersal)} + \text{species which have arisen through speciation} - \text{species which have been lost due to extinction} \]

At large spatial scales (e.g., regional or between biomes), dispersal can be considered of limited importance compared with speciation or extinction. Therefore, the
hypotheses which aim to explain the latitudinal gradient in diversity must have, as a basis, processes which enhance speciation rates and/or reduce extinction rates in the tropics or, alternatively, which reduce speciation rates at higher latitudes and/or increase extinction at those latitudes. The hypotheses that have been suggested to explain the enhanced diversity of the tropics are described within this broad framework.

1. Time
One of the earliest explanations for the latitudinal diversity gradient is that tropical regions are more ancient,
and therefore less disturbed, than their temperate counterparts. The greater age of the tropics implies that there has been more time for speciation to occur, hence a longer time for species to accumulate. Particular emphasis was placed on glaciations because it was assumed that temperate regions are impacted upon by glacial activity more so than the tropics. However, the fossil record does not support this view. For example, there is evidence that the latitudinal gradient for angiosperms has existed for at least 110 million years. There is also evidence that during the most recent glaciation both temperate and tropical forests decreased substantially, with tropical forests being replaced by grassland and savanna. This suggests that tropical regions may not be as ancient or stable as once thought.

However, if the tropical rainforest remnants were more numerous, and their destruction less complete than in the temperate zone, then this continual isolation and reforming of tropical communities might promote diversity through the mechanism of allopatric speciation. Allopatric speciation is considered by many to be the most important of the speciation-promoting mechanisms. It occurs when a geographical barrier subdivides a population, such that dispersal and genetic exchange between the subpopulations is prevented. These subpopulations continue to evolve in isolation, eventually diverging enough so that two new species are produced. In this scenario, higher diversity in the tropics might be promoted due to enhanced allopatric speciation due to historical geological events, possibly combined with an increased extinction rate in the temperate regions.

2. Area

Calculated on a land area basis, tropical and subtropical ecosystems are approximately five times more extensive than temperate ecosystems and approximately twice as extensive as boreal and tundra ecosystems. The reason is due to the earth being approximately spherical so that a zone of a constant width running around the equator (the tropics) will contain much more area than a zone of equivalent width running around the earth’s surface closer to the poles. Therefore, tropical ecosystems, which contain the highest diversities, are also those which cover the largest land area.
Rosenzweig suggested three reasons why a greater land area in the tropics might result in greater diversity. They are all based on the assumption that the geographical ranges of tropical species are, or were in the past, larger than the ranges of species in other latitudes. First, a larger geographical range leads to a larger total population size, and a larger population size would be expected to buffer the species against accidental extinction. Second, a larger geographical range contains a greater number of potential refuges where the species might survive following catastrophic disturbances or climatic events, again buffering the species against accidental extinction. Third, a larger range has more chance of being subdivided by a geographical barrier, hence increasing the probability of allopatric speciation.

3. Habitat Heterogeneity

All other things being equal, larger areas also contain a greater variety of habitats than smaller areas, and hence they are capable of supporting more species, i.e., there is opportunity for the evolution of species to fill these available habitats. Because tropical ecosystems tend to be larger in extent than those in other regions, it follows that they should contain a greater number of species.

It has also been suggested that the increase in the diversity and complexity of vegetation as one approaches the tropics can enhance diversity by providing animals with a greater variety of arboreal habitats. This increased faunal diversity can enhance diversity because many of these species can become a resource on which species higher in the food chain can be supported. Although the increased vegetation complexity in the tropics potentially explains the increased diversity in animal species, it cannot explain the high plant diversity because there is no evidence to suggest that plant habitats are more diverse in tropical regions.

Despite many studies which illustrate a positive rela-
tionship between habitat heterogeneity and diversity, care must be taken with the interpretation of these results because habitats are often defined by the species that occur within them, hence leading to circular reasoning. To be a valid hypothesis, habitat heterogeneity must be defined independently from the species occupying those habitats.

4. Solar Energy
Tropical regions receive, per unit area and per unit time, greater amounts of solar radiation than any other ecosystems. This is again due to a spherical Earth, whereby light energy at higher latitudes intercepts the earth’s surface at a more oblique angle compared with the tropics. Although the amount of incoming solar radiation directly affects productivity, and relatively, the water balance of a region, it also has the potential to accelerate the rate at which new species evolve.

An acceleration of the evolutionary process in tropical regions can arise due to two factors. First, the greater amount of available energy might explain the tendency for tropical species to have shorter generation times. A shorter generation time means more offspring per unit time, hence faster evolution. Second, increased solar radiation would be expected to increase mutation rates, thereby increasing the genetic variability on which natural selection acts, leading to an increase in the rate of speciation. These two factors combine so that over evolutionary time more species would be expected to have evolved in tropical regions rather than those at higher latitudes.

5. Species Interactions
Species interactions, and in particular competition, are central to many hypotheses addressing the latitudinal diversity gradient. A common ingredient to these hypotheses is that there are factors in the tropics which allow competitive exclusion to be avoided or which allow shifts in the “ecological niche” of a species so that competition is no longer intense, enabling the coexistence of many species. The ecological niche is a multidimensional summary of the environmental ranges and resource tolerances of a species. Species with broad niches are considered generalists, and species with narrow niches are considered specialists. Most important, the amount of niche overlap between species is a measure of the amount of competition experienced. For example, two species with overlapping niches require the same range of resources and environmental conditions; hence, they would be expected to compete strongly, and in the absence of any other factors the weaker competitor of the pair would be expected to be eliminated by the stronger. This is known as the “competitive exclusion principle.” It has been suggested that competition between species in the tropics is of greater importance, and is potentially more intense, than that in temperate regions. The reason given is that the tropical environment is more predictable and less prone to large disturbance events which cause indiscriminate mortality; therefore, population sizes are greater so that species interactions are more likely. High competitive pressure and a stable environment allow natural selection to narrow the niche requirements of the species, allowing the environment and resources for which they are competing to be partitioned to a finer degree than otherwise possible. This would result in more species coexisting per unit area of habitat in the tropics than in temperate areas. This hypothesis assumes that competition is not as important in temperate and higher latitudinal regions, and that in these regions the physical environment plays a greater role in forcing natural selection and evolution. The validity of this assumption remains an open question.

A second hypothesis assumes there are more predators (including herbivores) in the tropics. Predators remove prey biomass, thereby reducing competition among the prey species by reducing the demand on available resources. Under this hypothesis, competition is also less intense in the tropics, not because the species have evolved to avoid it through partitioning the environment more efficiently but because prey biomass is reduced to such an extent that competition is significantly reduced. The assumption that predation is more important in tropical ecosystems also remains an open question.

6. Productivity
Productivity is defined as the rate at which biomass is accumulated per unit area. Many hypotheses have attempted to relate the productivity of different ecosystems to diversity; however, they are not always easy to disentangle from one another and from other hypotheses (e.g., the solar energy hypothesis).

One line of reasoning suggests that low-productivity environments (e.g., those at higher latitudes) can support only a limited resource base (e.g., a low-productivity environment might be able to support a herbivore species but at a population density too low to support a predator higher in the food chain). In contrast, in a high-productivity tropical environment a herbivore population might be large enough to support several predatory species. Also, in high-productivity environments food might be more abundant, allowing the opportunity for predators and herbivores to specialize more finely, resulting in a greater number of species
Another suggestion is that although total productivity over the course of a year might be higher in the tropics than that in other regions, instantaneous rates at other latitudes are much higher (e.g., the flush of plant growth during spring at more temperate latitudes). This effect, combined with low nutrient availability in the tropics, would result in overall lower population growth rates in tropical species, which would tend to slow rates of competitive exclusion and hence extinction. Although slow rates of competitive exclusion cannot explain enhanced diversity on their own, when population sizes are continually impacted upon by external factors such as disturbances, under certain conditions the slow competitive exclusion can be converted into species coexistence, thus promoting higher diversity.

Because the latitudinal diversity gradient appears to be a universal pattern across a wide range of different kinds of organisms, it is likely that there must be some common explanation. The challenge for the future is to identify which factors, particularly which combinations of factors, are the primary causes of this pattern.

See Also the Following Articles
ECOSYSTEM FUNCTION MEASUREMENT, TERRESTRIAL COMMUNITIES, ENDANGERED TERRESTRIAL INVERTEBRATES, INVERTEBRATES, TERRESTRIAL, OVERVIEW, MARINE ECOSYSTEMS

III. SUMMARY

There are consistent trends in global biodiversity across the various ecosystem types which broadly correlate with latitude. Many hypotheses have been proposed to explain this gradient; however, the spatial and temporal scales over which the processes of speciation and extinction take place make rigorous testing of these hypotheses extremely difficult. As a result, there is still no consensus on which of the hypotheses might be the most important in determining actual gradients. A further difficulty in identifying the primary causes of the latitudinal diversity gradient is that many of the hypotheses are intimately related, with the potential for many of them to be acting in unison to determine the diversity of a particular ecosystem. It is also likely that the relative importance of different hypotheses varies both across different ecosystems and through time.

Bibliography
THERMOPHILES, ORIGIN OF

Anna-Louise Reysenbach and Margaret L. Rising
Portland State University

I. Thermophiles and Hyperthermophiles Defined
II. Ecological Niche and Metabolic Diversity of Hyperthermophiles
III. Molecular Mechanisms for Thermophily
IV. Origin of Thermophiles: Fossil Record and Universal Phylogenetic Tree
V. Theories of the Origin of Life
VI. Conclusion

GLOSSARY

Archaea One of three domains of life; from the Greek archaios (ancient, primitive); prokaryotic cells; membrane lipids predominantly isoprenoid glycerol diethers or diglycerol tetraethers; formerly called archaebacteria.

Bacteria One of three domains of life; from the Greek bacterion (staff, rod); prokaryotic cells; membrane lipids predominantly diacyl glycerol diesters; formerly called esubacteria.

biomarker A macromolecule unique to a particular organisms or group of organisms such that its detection alone would suggest the presence of the organism or group of organisms.

chemolithotroph Organism deriving its energy from the oxidation of inorganic compounds.

Eukarya One of three domains of life; from the Greek eu- (good, true) and karion (nut; refers to the nucleus); eukaryotic cells; cell membrane lipids predominantly glycerol fatty acyl diesters.

heterotroph Organism deriving its energy from the oxidation of organic compounds.

hyperthermophiles From the Greek hyper- (over), therme- (heat), and philos (loving); includes organisms that grow best at temperatures warmer than 80°C.

photoautotroph Organism that uses light as its source of energy (photosynthesis) and inorganic carbon (CO2) as its sole carbon source.

planetesimals Small, solid bodies similar to meteors in composition but revolving in orbit around a central gaseous nucleus, as do planets around the sun.

small subunit ribosomal RNA RNA (about 1500 bases in prokaryotes) that functions as part of the ribosome and the sequence permits the inference of evolutionary relationships among organisms.

thermophiles From the Greek therme- (heat) and philos (loving); includes organisms that grow best at temperatures between 50° and 80°C.

AT THE TIME when liquid water—a prerequisite for life as we know it—appears in the geological record (3.8 billion years ago), Earth was a hot, anoxic environment and under constant bombardment by meteors, many of which could have virtually vaporized the oceans. Early Earth, therefore, would have been an attractive home to heat-loving thermophiles and their extreme cousins, the hyperthermophiles, where thermophily would have offered a great selective advantage.
The geochemical and thermal characteristics of deep-sea hydrothermal vents and terrestrial hot springs are thought to approximate conditions on Earth at the earliest possible time that it could have supported life, providing modern analogs for testing early evolution of life hypotheses. In addition, the relatively recent use of the evolutionarily conserved molecules such as the small subunit ribosomal RNA (ss rRNA) sequences in phylogenetic analyses of all life has placed the thermophiles closest to the root of the universal tree of life (Fig. 1). On this tree, hyperthermophiles decidedly monopolize the lowest and shortest branches, suggesting that they may be related to the earliest microbes to inhabit Earth. However, the proposal that thermophiles may have originated early in Earth’s history is much debated. Therefore, this article also considers alternative scenarios for the origins of thermophiles.

I. THERMOPHILES AND HYPERTHERMOPHILES DEFINED

All life can be placed in one of the three domains of life: the Archaea, Bacteria, or Eukarya. Archaea and Bacteria are prokaryotes, and the Eukarya are eukaryotes. However, phylogenetically, the Archaea are as different from the Bacteria as the Bacteria are from the Eukarya. Thermophiles are found in all domains as multicellular and unicellular organisms, such as fungi, algae, cyanobacteria, and protozoa, and they grow best at temperatures higher than 45°C. In contrast, the extreme thermophiles, or hyperthermophiles, grow best at temperatures higher than 80°C and are almost exclusively restricted to the Archaea, with only two hyperthermophilic orders in the Bacteria, namely, the Thermotogales and Aquificales. Commonly, thermophiles and hyperthermophiles are found associated with deep-sea and terrestrial hydrothermal vents. Thermophiles have also been obtained from deep (up to 3500 m), hot, subterranean areas, Jurassic oil-bearing sandstone and limestone formations, and suitable manmade environments such as smoldering coal refuse piles, coal-containing uranium mines, and boiling wastewaters from geothermal power plants.

II. ECOTOLOGICAL NICHE AND METABOLIC DIVERSITY OF HYPERTHERMOPHILES

Deep-sea hydrothermal and terrestrial hydrothermal vents form primarily as the result of plate tectonics, either as a result of seafloor spreading or as a tectonic plate moves across a hot spot. In both scenarios, fissures and faulting occur in the earth’s crust, permitting water (seawater, groundwater, or rainwater) to percolate downward. As the fluid moves through the earth’s crust and approaches the magma chamber, the fluid heats and reacts with the surrounding rocks, adding some minerals and gases (e.g., iron, manganese, carbon dioxide, hydrogen, and hydrogen sulfide) and removing others (e.g., sulfate and magnesium in seawater). The
fluid is finally forced back to the surface as a highly altered hot fluid, its chemistry representing a history of its travels through the earth's crust. At deep-sea hydrothermal vents, this fluid can reach temperatures of 400°C, remaining in a liquid state due to the hydrostatic pressures at these depths (>2000 m). It rapidly mixes with cold, oxygenated water, and the minerals rapidly precipitate, giving the fluid a smoky appearance (Fig. 2). These vents are aptly named “black smokers.” As the minerals precipitate they create porous sulfide–mineral structures called “chimneys.” At terrestrial hot springs, the fluid may be ejected forcefully as a geyser or simply bubble into a thermal spring or mudpot (Fig. 3).

The high temperatures and unusual geochemical characteristics of deep-sea and terrestrial hydrothermal vents create ecological niches that are exploited by hyperthermophilic Bacteria and Archaea. The geochemical milieu provides different energy sources (electron donors) and carbon dioxide for a chemolithotrophic existence. The porous chimney structures provide temperature gradients along which thermophiles can situate themselves, and thermal springs create temperature gradients as the fluid moves away from the source. Comparable to photosynthetic organisms that harvest light energy to fix inorganic carbon, these chemolithotrophs harvest inorganic chemical energy to fix inorganic carbon. It has been calculated that the mixing of superheated thermal fluids at deep-sea vents with cold, oxygenated seawater causes a significant geochemical disequilibrium, significant enough to lower the Gibbs free energy and help drive biologically mediated redox reactions (Shock, 1996). The food chain does not stop there, though. Many thermophiles are heterotrophs, consuming organic carbon produced from biological activity. Additionally, at terrestrial vents, due to the available light energy, the springs are often colonized by a rich diversity of thermophilic phototrophs, including the cyanobacteria, which are oxygenic phototrophs that use water as the reductant and evolving oxygen in the process. Anoxygenic phototrophs are also found at terrestrial vents, primarily using the readily available hydrogen sulfide as the reductant in photosynthesis, releasing elemental sulfur in the process. Both of these types of phototrophs are part of a complex community with heterotrophs, and together they form thick green, purple, and orange microbial mats typical of many terrestrial thermal areas throughout the world.

The majority of hyperthermophiles are anaerobes, with a few exceptions that are microaerophilic, requiring low oxygen concentrations for growth. In contrast, aerobiosis is more common among the thermophiles because oxygen dissolves more readily in the lower temperatures of the cooler thermal springs. The thermophiles that have been isolated from deep-sea vents generally grow at pH values near neutrality. However, thermophiles from terrestrial hot springs have been grown at pH values from 1 to 10.

Although the highest temperature for life has not been determined, it has been suggested that 150°C may represent that threshold. In contrast, the lower temperature limit for most hyperthermophiles is 60°C, a temper-
THERMOPHILES, ORIGIN OF

ature fatally hot for most other organisms. Even 90°C is too low for one of the more extreme hyperthermophiles, *Pyrolobus fumarii*, which has a maximum temperature for growth of 113°C. However, for these thermophiles, the inability to grow does not result in death; they are capable of entering a dormant state when exposed to cold conditions and will thrive again when returned to favorable temperatures. Thermophiles are therefore very well suited to the deep-sea hydrothermal vent environment, in which the hot fluids are continually being mixed with cold, oxygenated seawater.

III. MOLECULAR MECHANISMS FOR THERMOPHILY

On a molecular level, thermophily is accomplished in many ways, contributing together to overall stability and success of life at high temperatures. In general, the amino acid composition of hyperthermophilic enzymes is surprisingly similar to that of homologous mesophilic enzymes. Because of the amino acid sequence similarity, it has been proposed that heat resistance is explained, instead, by the manner in which the polypeptide chains are folded on themselves, forming their tertiary conformational stability. It is a polypeptide's three-dimensional or tertiary shape in critical locations on the polymer that permits an enzyme to function. Proteins do not spontaneously fold into their active form after synthesis. Instead, many require the assistance of molecular chaperones for proper folding. A specialized form of molecular chaperone, called a thermosome, is of particular importance to hyperthermophiles because it assists proteins in refolding correctly after denaturation due to heat exposure. Also called heat shock proteins, thermosomes effectively increase heat tolerance of organisms so they can function at higher temperatures. For instance, at 108°C, approximately 80% of the protein of *Pyrodictium occultum* consists of a heat-induced thermosome, allowing it to survive 1 h of autoclaving at 121°C (Stetter, 1998).

Additional thermal resistance of DNA is conferred by the presence of DNA topoisomerases and histones and, to a lesser extent, by the base composition of the genome. All hyperthermophiles known to date have been shown to possess reverse gyrase, a unique type 1 DNA topoisomerase that causes stabilizing, positive supertwists in the DNA helix. Additionally, archaeal (but not bacterial) hyperthermophiles possess histones which significantly increase the temperature at which DNA denatures.

Because an increased guanine + cytosine (G + C) nucleic acid content increases the melting temperature of DNA, one would predict that hyperthermophiles have a higher G + C content in their genomes. In many cases, the opposite is true. However, analysis of the genome reveals that certain genes essential to survival, such as the 16S rRNA gene, have higher G + C contents.

The presence of ethers instead of esters in cellular membranes also contributes to the stability of hyperthermophiles at high temperatures. Cell membranes of *Bacillus* and *Archaea* are generally distinguished by the presence of glycerol-linked ether lipid moieties in *Archaea* and ester lipid moieties in *Bacteria*. An exception to this rule is the presence of a glycerol ether lipid in *Thermotoga maritima*, one of the few bacterial hyperthermophiles. The presence of the ether lipids probably increases stability of cellular membranes against hydrolysis at high temperatures. Membranes of the archaeal hyperthermophiles contain lipids derived from diethers or tetraethers and are highly resistant against hydrolysis at high temperatures.

IV. ORIGIN OF THERMOPHILES: FOSSIL RECORD AND UNIVERSAL PHYLOGENETIC TREE

Were thermophiles the original ancestors to all life? Much of the debate about origins of life and early evolution of life on Earth centers on this question. Is thermophily an acquired feature derived from adaptation? As more information accumulates, the resolution of these questions becomes less clear. However, many different scenarios are emerging based on fossil record analyses, RNA and protein phylogenies, and our understanding of the conditions that prevailed on early Earth.

A. Early Earth and Its Hyperthermophilic Niche

Perhaps the most compelling evidence supporting the proposal that hyperthermophiles are living fossils of the first life that arose on Earth is that modern ecological niches of hyperthermophiles fit our current view of the conditions on primitive Earth, which was anoxic, hot, and volcanically active. During its first half billion years of existence (+6–4.0 billion years ago), the temperature of the surface of Earth probably exceeded 100°C. From recent studies of ancient sedimentary rocks of the Isua formation in Greenland, traces of biologically produced carbon were detected, suggesting that the invention of
life had already taken place 3.8 billion years ago. Thus, sometime prior to 3.8 billion years ago, it is likely that the first living organisms appeared at a time when Earth, although cooling, was much hotter than it is now, and all the essential geochemical energy sources were present for chemolithotrophic and thermophilic existence. Heat-tolerant organisms such as hyperthermophiles would have been at a distinct advantage over heat-sensitive organisms.

In addition to the generally hot and anoxic conditions on early Earth, the bombardment of early Earth by meteorites may have caused a bottleneck event that favored survival of hyperthermophilic organisms. An examination of the impact craters on Earth’s geologically inactive moon strongly suggests that the earth was heavily bombarded by meteorites from its formation 4.6 billion years ago until about 3.6 billion years ago. A large meteor impact, of which there were many, would have caused intense local heat and sent sufficient debris into the atmosphere to cause global cooling and a reduction in light intensity, detrimentally impacting photosynthesizing organisms. There is also evidence of extremely large meteor impacts that would have generated enough energy to heat the earth’s surface to 2000 K and virtually vaporize the world’s oceans. In either scenario, hyperthermophiles would have been uniquely adapted to survive. They could have tolerated lower light levels because many are chemolithotrophs, making a living independent of photosynthesis. Any life adapted to live near ocean-floor volcanic centers would have had the greatest chance of surviving one of these ocean-evaporating events.

B. The Fossil Record

Additional evidence for early ancestors of microbial life can be found in the rock record. Rocks older than 3.5 billion years are highly metamorphosed and deformed, precluding the preservation of morphological fossils. However, the 3.8-billion-year-old Isua rocks from Greenland, mentioned previously, offer indirect evidence that life existed earlier than 3.5 billion years ago. These sedimentary rock formations are associated with liquid water, a prerequisite for life as we know it. The rocks contain geochemical evidence of past biotic activity preserved within the minerals. Grains of apatite (calcium phosphate) contain a significant portion of carbonaceous inclusions that are isotopically “light,” suggesting biological activity. Microorganisms preferentially incorporate the lighter carbon-12 isotope over the carbon-13 isotope when organic carbon is fixed into organic carbon using the enzyme ribulose bisphosphate carboxylase/oxygenase (Rubisco). This isotopic evidence suggests these putative life-forms were carbon-fixing chemolithotrophs or phototrophs and not heterotrophs.

Three hundred million years later, the first microfossils appear in rocks (about 3.5 billion years old) from Western Australia and South Africa. The fossils are simple rod-shaped and filamentous bacteria, indistinguishable morphologically from any similar-shaped bacteria today. Many are reminiscent of photosynthetic bacteria, such as modern cyanobacteria. However, there is some doubt that much oxygen evolved from photosynthesis into early Earth’s atmosphere. Isotopic signatures associated with these fossils do suggest a lighter carbon preference, similar to that observed in the 3.8-billion-year-old rocks. Were these first fossils perhaps thermophiles? Unless we are able to identify specific thermophilic biomarkers in these fossils, we may never know the answer.

C. The Molecular Fossils

Although they are not a substitute for the existence of microfossils, evolutionarily conserved macromolecules within living organisms function as “molecular fossils,” permitting an inference of relatedness to other organisms and of evolutionary distance from a hypothetical common ancestor. Carl Woese’s pioneering work with the ss rRNA molecule (16S rRNA in prokaryotes) led to the generation of the universal phylogenetic tree (Fig. 1), in which life falls within three domains: the Bacteria, Archaea, and Eukarya. The suitability of ss rRNA as the macromolecule for evolutionary comparison is based on several critical reasons. The ss rRNA exists in every living organism, its function in every living organism is constant, it is an ancient molecule (assuming protein synthesis was necessary in the earliest of life), it has undergone only moderate changes in nucleic acid sequence when compared between diverse biological domains, and the size of the molecule is large enough to contain considerable information but small enough to be manageable.

1. Molecular Phylogenies

The relatedness of organisms to each other as suggested by the universal phylogenetic tree is markedly different than the previously proposed phylogenies. Most notably, the three domains of biological life (the Bacteria, Archaea, and Eukarya) replace the five kingdoms (animals, plants, fungi, protists, and monera). The Eukarya domain includes the multicellular animals, plants, fungi, and protists. Distributed throughout the Bacteria...
and Archaea are the unicellular prokaryotes. Using alternative markers, such as 23S rRNA, DNA polymerase, elongation factor Tu, F$_{1}$ATPase $\beta$-subunit, RecA protein, and HSP60 heat shock protein, reveals general agreement with the major lineages in the ss rRNA tree. However, some discrepancies are evident.

One can root the ss rRNA tree by comparing it with a paralogous marker that arose from gene duplications (such as EF-Tu or ATPase) prior to the diversification of the three primary domains. This rooted tree reveals that all hyperthermophiles, whether bacterial or archaeal, comprise the deepest and earliest lineages within the tree (Fig. 1). In other words, they may be the closest living relatives to a common universal ancestor. However, mutational rates of ss rRNA and their associated base sequence changes are not consistent enough through time not within lineages to assign a specific date to branching events, and so one cannot assign a clock to this tree. Nonetheless, the shorter branches leading to the hyperthermophiles indicate a slower evolutionary rate, whereas the longer branches leading to their mesophilic relatives suggest a faster evolutionary rate. It is possible that the hyperthermophiles have short lineages because they are so highly adapted to their ecological niche than they have no need for further adaptation. It is also possible that certain evolutionary constraints are imposed on thermophiles due to the extreme selectivity of the high-temperature environment. Nevertheless, the placement of thermophiles at the base of the universal phylogenetic tree strongly suggests that they are most closely related to the common ancestor of all life. In this sense, they are not unlike other “living fossils,” or organisms whose morphology has changed very little based on comparisons of modern and ancient fossilized specimens, such as horseshoe crabs, club mosses, Welwitschia, and Gingko biloba.

Although hyperthermophiles may be most closely related to the universal ancestor of all life, their metabolic machinery is anything but primitive. Viewed independently of the rooted ss rRNA universal phylogenetic tree, their heat-tolerant adaptations appear sophisticated or even highly evolved. If an analysis of another universal and highly conserved macromolecule placed hyperthermophiles away from the root of its phylogenetic tree, it would severely shake the topology of the phylogenetic tree discussed previously. A phylogenetic analysis of the conserved DNA-directed RNA polymerase has done just that (Klenk et al., 1999). Although overall the RNA polymerase-based phylogeny corresponds very well with the ss rRNA-based phylogeny, the positions of the hyperthermophilic bacteria A. pyrophilus and Thermotoga maritima are in the middle of the RNA polymerase tree instead of in the most deeply rooted branch in the Bacteria as suggested by ss rRNA phylogeny. Other examples also exist in which the ss rRNA tree does not quite hold true. Additional analyses of universal and highly conserved macromolecules such as DNA-directed RNA polymerases or whole genomes may one day provide a comprehensive universal phylogenetic tree that may or may not place hyperthermophiles near its root.

The debate continues, and evidence against an early origin of thermophily accumulates. If mesophily preceded thermophily, then versions of heat-tolerant mechanisms that characterize hyperthermophiles would have evolved first in mesophiles and been exploited by hyperthermophiles to open up new ecological niches. Structural analysis of reverse gyrase and Taq polymerase support such a proposal, suggesting they evolved from mesophilic ancestors (Forterre, 1996). Additionally, the lipid moieties in hyperthermophilic Bacteria are not homologs of the lipid moieties in hyperthermophilic Archaea. Instead, they are analogs with opposite stereochemistry, suggesting the ether-based lipid moiety feature evolved independently in the Bacteria and Archaea instead of from a common ancestor (Forterre, 1996).

2. Lateral Gene Transfer and the Genetic Annealing Model

Lateral (horizontal) gene transfer and variable rates of evolution and mutation may have played a significant role in the evolution of life and may offer an explanation for the confusing results of the analyses of non-ss rRNA molecules (Woese, 1998; Doolittle, 1999). Lateral gene transfer refers to the exchange of genetic material from one organism to another. It is contrasted against vertical gene transfer, in which genetic information is passed vertically from parent to offspring. Bacteria and Archaea, which as a general rule have simpler cell designs, exhibit considerable horizontal gene transfer. In contrast, the Eukarya, which contain highly evolved cell designs, generally do not engage in horizontal gene transfer. Woese proposed in his “genetic annealing model” that the universal ancestor was not a discrete entity but rather a diverse community of primitive cells that evolved as a unit, engaging in horizontal gene transfer on a scale even greater than that occurring in Bacteria and Archaea today and developing into three different communities, which in turn gave rise to the three primary lines of descent as defined by the ss rRNA tree (Woese, 1998). The ss rRNA universal tree, then, is not a conven-
tonal organismal phylogenetic tree but rather a history of the evolution of central components of the ribosome, with the deeply rooted branches of the universal tree representing a ‘gene tree’ and not an ‘organismal tree.’ In other words, by the time the three primary lines of descent emerged, and the tree started to take form, self-replicating organisms had not yet taken form. Instead, ‘life,’ with its associated exchange of genetic information, existed in communal entities by way of lateral gene transfer. Additional cell complexity and function needed to evolve before there was life as we envision it today. Nonetheless, these communal entities could have evolved in the conditions on early Earth.

V. THEORIES OF THE ORIGIN OF LIFE

Theories of the origin of life are highly debated. Did life originate here on Earth, or did it develop on another planet, such as Mars, which shared similar planetary conditions 4.0 billion years ago? Did life originate in environments analogous to deep-sea hydrothermal vents, or did the first biological molecules form as a result of the reaction of electrical discharges within a prebiotic soup of chemicals? Closely tied to much of this debate are models pertaining to early Earth’s atmosphere and the possible energy carbon sources (organic and inorganic). Given that this article deals with the origins of thermophiles, we focus our discussion on a possible high-temperature origin of life, highlighting some arguments that do not support this hypothesis.

In light of Earth’s hot, reduced, and anoxic origins, it has been proposed that life may have arisen in environments very similar to present-day deep-sea hydrothermal vents. Here, life could take refuge and survive the planetary bombardment of early Earth in an environment rich in redox energy and inorganic carbon. The first biological entities could evolve rapidly as anaerobic, hyperthermophilic chemolithotrophs. The rapid mixing of superheated hydrothermal fluid, rich in reduced minerals, with cold oxygenated sulfate-rich seawater at deep-sea hydrothermal vents creates thermodynamic disequilibrium conditions that favor the production of organic molecules (Shock, 1996). The abundance of charged minerals associated with deep-sea vents led to the proposal by Gunter Wächtershäuser that the original source of reducing power for carbon fixation (and therefore a chemolithotrophic origin of life) may have come from exergonic ‘pyrite-pulled reactions’—the oxidative formation of pyrite (FeS₂) from ferrous sulfide (FeS) and hydrogen sulfide (H₂S or SH⁻). Charged surfaces such as pyrite would attract and bind any negatively charged molecule in solution, such as carbonate, phosphate, and sulfide. These molecules would be maintained in sufficient proximity for subsequent metabolic interactions by ‘surface bonding,’ or anionic bonding to the positively charged pyrite surface, resulting in the formation of the first biomolecules on a charged surface. Many are skeptical of these high-temperature scenarios for the origin of life, offering evidence that many of the biomolecules of life, such as RNA, are not stable at high temperature and therefore would not support the proposal that RNA arose prior to DNA in an early RNA world (Miller and Bada, 1988).

Furthermore, others have generated models to show that the ancestral rRNA would have a moderate G + C content, contrary to all thermophilic rRNAs which are generally rich in G + C contents (Galtier et al., 1999).

The ‘panspermia’ hypothesis offers another scenario for the origin of life. It holds that life originated elsewhere in the galaxy and that microorganisms were propelled through space to Earth or, alternatively, that exogenous organic carbon arriving on planetesimals fueled a heterotrophic origin of life on Earth. Chyba and Sagan (1992) estimate that approximately 4 billion years ago about 100,000,000 kg/year of organic carbon was delivered by interplanetary dust particles, and about 10,000,000,000 kg/year of organic carbon was produced by postimpact plumes caused by meteor, asteroid, and comet bombardment. From what is known about the hostile environment on early Earth and the timing of the appearance of the first microfossils, life became considerably complex within a relatively short period of time. Therefore, life either evolved very rapidly after its inception or, consistent with the panspermia hypothesis, it was raining down on Earth from elsewhere in the galaxy. Because the panspermia hypothesis involves high temperatures as particles enter the earth’s galaxy, it embraces, at the very least, a thermotolerant origin of life.

VI. CONCLUSION

Morphological features of early microfossils generally offer little information about their relatedness to extant microbes. Consequently, microbiologists have come to rely on the molecular record to define phylogenetic relationships and infer the history of microorganisms. The absence of a good microbial fossil record prevents the assignment of a time line on the molecular record. The ss rRNA universal phylogenetic tree places hyperthermophiles in the deepest and shortest branches of the tree, implying that they are the closest living relative
to the common universal ancestor of life. However, as we sequence more microbial genomes and create detailed phylogenies from other molecules, the situation becomes increasingly confusing, with lateral gene transfer perhaps "muddying" the phylogenetic record. However, high-temperature conditions that prevailed on early Earth and its subsequent cooling favor the likelihood that at some time during Earth's early history thermophiles took advantage of the geochemical energy supplied by the hydrothermal fluid and evolved into the highly adapted chemolithotrophic and heterotrophic life that is found at high temperatures.

See Also the Following Articles

ARCHAEA, ORIGIN OF • BACTERIAL BIODIVERSITY • EUKARYOTES, ORIGIN OF • HIGH-TEMPERATURE ECOSYSTEMS • ORIGIN OF LIFE, THEORIES OF • PSYCHROPHILES • VENTS

Bibliography


I. Main Features of Boreal Forests, with Special Reference to Finland
II. Dynamics of Boreal Forest Ecosystems and Management Implications
III. Management of Forests for Optimal Productivity and Timber Production
IV. Impacts of Timber-Oriented Management on Biodiversity: Comparison between Managed and Natural Forest Areas
V. Conclusions

GLOSSARY

biodiversity of forests Variability in genetic structures, species composition, and/or habitat properties in forest ecosystems.
boreal forests A forest zone dominated by coniferous species, covering the northern latitudes of North America, the Nordic countries, and Russia.
forest management Manipulation of the properties of tree populations and/or modification of the properties of the habitats occupied by these populations.
sustainable forestry The management and utilization of forest resources in a way that balances human needs with undisturbed functioning of the forest ecosystem, and a long-term maintenance and sustainability of the resources and biodiversity of forest ecosystems.
timber industry Any kind of forest-based production using timber as a raw material.

THE MANAGEMENT OF FOREST RESOURCES can substantially change the structure and functioning of forest ecosystems at both the local and the regional level, with potential impacts on forest biodiversity. This article exemplifies these impacts in the context of boreal conditions with the aim of outlining how to maintain the biodiversity of boreal forest ecosystems. The management of forest resources and the impacts of this management on biodiversity are contrasted with the successional dynamics of boreal forests, which are also disturbance driven under natural conditions. The findings of this discussion are related to the structure of forests on opposite sides of the Finnish–Russian border, and the management of boreal forests in Finland and Russia is addressed against a background of the natural succession in such forests. Furthermore, these two forest areas offer a unique opportunity to study the effects that intensive management may have on the biodiversity of boreal forests. In this discussion, biodiversity is illustrated by reference to selected beetle populations. Outlines for the
management of boreal forests with a view to preserving biodiversity are discussed.

I. MAIN FEATURES OF BOREAL FORESTS, WITH SPECIAL REFERENCE TO FINLAND

Forests currently cover about 3,440 million ha, or 27% of the world’s land surface, and represent the largest land-based ecosystem types in the world. Forest species, along with related species communities and ecosystems, represent a large proportion of the global biological diversity and are of great importance in maintaining the functioning and structure of the biosphere. The boreal forests alone represent an area of 600 million ha and account for about 20% of the world’s industrial timber supplies. Commercially, the most important coniferous species are pine (Pinus), spruce (Picea), fir (Abies), larch (Larix), juniper (Juniperus), thuya (Thuja), and hemlock (Tsuga), whereas the most common deciduous species in these forests are aspen (Populus), birch (Betula), willow (Salix), and alder (Alnus).

The boreal forests in northern Europe or Fennoscandia (including Norway, Sweden, Finland, and northwestern Russia) are probably the most intensively utilized of all; that is, they form 85% of the total forest area in Europe (956 Mha) and provide about 40% of the timber used in Europe (Table I). The stocking (68 m$^3$ ha$^{-1}$) and growth rate (1.1 m$^3$ ha$^{-1}$ year$^{-1}$) in these forests are substantially lower than in more southerly regions, however, and consequently the timescale in forest production is long, with rotations of 40–160 years according to species and region. Growth and felling are currently well balanced, and the growth may even exceed the needs of the timber industry.

Forestry in northern Europe is mainly based on native tree species which invaded this region postglacially. In Finland, for example, Scots pine (Pinus sylvestris) is the dominant species throughout the country (about 45% of total stem volume) wherever nutrient and water supplies are limited at the site. Scots pine can also dominate at more fertile sites, but both Norway spruce (Picea abies) (about 37% of total stem volume) and birch (Betula pendula and Betula pubescens) can win out over it under such conditions. Birch species (about 18% of total stem volume) are most common on the most fertile soils.

In contrast with many developing countries, forest coverage in Fennoscandia has remained at the current level for decades or even increased slightly. Forest utilization in Finland has been occurring for centuries. Toward the late 1800s Finnish forests were intensively used for slash-and-burn cultivation and tar production without any actions to promote new tree generation. As a result, large areas, especially in southeastern Finland, became dominated by relatively young forests 100 years ago. Effective prevention of slash-and-burn cultivation and forest fires together with active timber-oriented management have led to an increase in the areas of closed-canopy forest during the past century.

In general, the number of species per unit area is low at northern latitudes. This also holds for Finland, where the estimated total number of species is about 50,000, of which about 43,000 are known. The reason for the relatively low total number of species is the short time that has elapsed since the last glaciation (10,000 years), with the consequence that immigration is still occurring. These species and the subsequent biodiversity involve a large contribution from natives of the eastern taiga (e.g., flying squirrel, Ural owl, and Siberian jay). Most of these taiga species are connected with spruce forests. Furthermore, the amount of dead or decaying wood is generally high in boreal forests.

<table>
<thead>
<tr>
<th>Countries representing boreal forests in Europe</th>
<th>Examples of tree species of importance in forestry</th>
<th>Main parameters for forests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finland, Sweden, Norway excluding south-west part, north-western Russia</td>
<td>Pinus sylvestris, Picea abies, Betula pendula,</td>
<td>Area (Mha)</td>
</tr>
<tr>
<td></td>
<td>Pinus contorta, Larix sibirica</td>
<td>Total volume (Mm$^3$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total growth (Mm$^3$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total cut (Mm$^3$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Balance (Mm$^3$)</td>
</tr>
</tbody>
</table>

especially in the early phase of the succession after a fire or windstorm. Consequently, a high proportion of the forest species (20–23%) are dependents on dead wood (800 coleopterans, 1000 dipterans, 1000 fungi, 200 lichens, etc.). Many of these species are also specialized in living on recently burned tree material, whereas a high proportion live in peatland forests or on mires. Most of these species need a moist and/or shady habitat.

II. DYNAMICS OF BOREAL FOREST ECOSYSTEMS AND MANAGEMENT IMPLICATIONS

A. Management: Control of the Structure and Functioning of the Forest Ecosystem

Forests are ecosystems in which trees and other green plants occupy the site and intercept solar energy under the control of climatic and edaphic factors (Fig. 1). The interaction of populations with their environments forms complex food webs, in which the solar energy flows from producers to consumers. The links between the organisms are the key to the management of a forest ecosystem. Proper manipulation of forest dynamics will allow the production of timber or other items, with the management needs being related to the management objectives and to the changes in the structural properties of the forest ecosystem occurring with time. The management history and past structure of a forest ecosystem may also have a crucial impact on future management. Production in a forest can be maintained and/or increased by manipulation of the genetic properties of the tree populations and/or modification of the properties of the habitats occupied by these populations. This takes places by controlling the long-term functional and structural development of the forest ecosystems (the forest succession) in order to induce them to produce the items defined in the management goals. In other words, any kind of forest-based production is a result of environment, genotype, and the interaction between these.

B. Successional Dynamics of Forest Ecosystems and Management Implications

The long-term growth and development of tree populations are subject to disturbances due to wind, fire, and attacks by pests and pathogens. Under boreal
conditions in particular, such disturbances cause cycling in the dynamics of the forest ecosystem by returning the succession to an earlier phase. For example, a fire that kills old trees provides space for regeneration. For larger forest areas, this implies that forest landscapes contain a mosaic of tree stands of varying species composition and age, from open areas and saplings to mature trees.

During the succession, each stand or part of the mosaic undergoes a gradual change involving alterations in species composition and the accumulation of organic matter in trees and soils (Fig. 2). Many broad-leaved tree species can make better use of the abundant supply of resources at fertile sites than can coniferous trees. Broad-leaved species typically dominate the initial phase of the succession after a fire, and wherever site fertility is sufficiently high the changing conditions will later start to support shade-tolerant conifers more than the broad-leaved trees. This brings an invasion of conifers such as Norway spruce to the site, which finally replace the broad-leaved species unless a disturbance causes the succession to return to an earlier phase. These disturbances and the dynamics of the tree species result in substantial variability in the mass of organic matter in forest ecosystems. This cycling determines the management of the forest ecosystem and the timber yield obtained.

In terms of the energetics of the ecosystem, the energy intercepted by the trees and other green plants and the energy used at different trophic levels in the ecosystem's food webs are balanced by means of enhanced formation of dead and decaying wood and other organic matter. In this context, the food webs provide complex feedbacks through which the populations of the various species interact, with consequent control of the ecosystem dynamics. The food webs tend to increase in complexity during the succession.

C. Disturbance Dynamics of Forest Ecosystems and Management

The disturbance dynamics of forest ecosystems provide the framework in which management occurs. Two categories of disturbance may be distinguished which drive the succession: autogenic and allogenic (hence, we may speak of an autogenic and/or allogenic succession). Allogenic and autogenic successions are typical processes operating simultaneously in boreal forests.

In the autogenic succession, the structural dynamics of the forest ecosystem are driven by the regeneration, growth, and death of single trees, i.e., minor disturbances related to their life cycle (gap-phase dynamics). The deaths of single trees and the consequent gaps in the canopy release resources and enhance regeneration and growth. The concept of autogenic succession is applied in the form of selection forestry or uneven-aged management, in which single trees are the basic object of management. This system is widespread in the temperate zone, in which autogenic succession predominates in the dynamics of forest ecosystems.

Allogenic succession refers to dynamics driven mainly by major disturbances induced by fires, gales, and excessive snowfall. Under these conditions many trees may die simultaneously, releasing space for new trees to regenerate and grow to form a population with individuals of more or less the same age (cohort-phase dynamics). In allogenic succession the stand is the basic unit driving the structural dynamics of the forest ecosystem. The concept is applied to management in the form...
TABLE II
Schematic Presentation of the Effects of Given Management Measures on Site Properties and Tree Populations

<table>
<thead>
<tr>
<th>Measure</th>
<th>Nutrients</th>
<th>Moisture</th>
<th>Temperature</th>
<th>Soil physics</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Allogenic measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site preparation</td>
<td>+</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Ditching</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td>Prescribed burning</td>
<td>+++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td>Fertilization</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Regenerative cutting</td>
<td>+++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td><strong>Autogenic measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection for regeneration</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Precommercial thinning</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Commercial thinning</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td>Pruning</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

*+, slight effect; +++, moderate effect; +++++, substantial effect.

of standwise forestry, or even-aged management, in which single stands form the basic object of management. This system is commonly employed in the boreal zone, in which allogenic succession predominates in the dynamics of forest ecosystems.

Management involves the use of a variety of measures to control the supply of resources and the properties of tree populations. The measures used to control the supply of resources may be classified into two categories in relation to the disturbances driving the succession methods liberating resources to a large extent (allogenic measures related to the allogenic succession) and methods liberating resources to a lesser extent (autogenic measures related to the autogenic succession) (Table II).

Allogenic measures mainly control the properties of the soil system, and subsequently the nutrient cycle and soil moisture, with the greatest enhancement of growth and regeneration of pioneer species being achieved through disturbance of the soil surface. Site preparation refers to mechanical measures used to reduce the effects of the ground vegetation on saplings and to adapt the physical and chemical properties of the soil. In this context, ditching is aimed at lowering the groundwater level and reducing excess soil moisture. Prescribed burning affects the chemical properties of the soil and eliminates ground vegetation interference with sapling growth. Fertilizing implies the addition of nutrients at the site, leading to a substantial increase in the availability of nutrients for tree growth.

Regenerative felling refers to terminal felling aimed at promoting reforestation of the site through natural regeneration by means of natural seeding or artificial regeneration by means of the sowing of seed or planting. Autogenic measures are mainly represented by spacing, leading to enhancement of growth and regeneration among both pioneer and climax species. This is achieved through the thinning of sapling stands (precommercial thinning) and more mature stands (commercial thinning). The tending of sapling stands is aimed at proper spacing, as is also true of thinnings, and it also implies the elimination of unwanted tree species. Pruning involves the removal of dead or living branches from the lower crown in order to increase the amount of knot-free wood obtainable from the stem. Selection for regeneration refers to the removal of single mature trees or small groups of mature trees in order to create canopy gaps in which seedlings can become established and grow.

III. MANAGEMENT OF FORESTS FOR OPTIMAL PRODUCTIVITY AND TIMBER PRODUCTION

Among the key issues to be addressed in management is the optimal distribution of tree species and stand age (or tree size) over the forest area in order to maximize long-term growth. Assume a forest area of 100 ha to
be divided into six compartments (stands) allocated to initial age classes (i.e., 2, 10, 22, 38, 82, and +82 years) in different ways. Calculations by Kellomäki (1998, pp. 225–226) show growth during the next 10 years to be highest in the stand with an initial age of 38 years, with this growth being five times that in the stand with an initial age of 85 years. Consequently, a distribution skewed toward dominance by old stands represents the highest level of stocking but about 13% less growth than that for a normal distribution (Fig. 3). Where young stands are dominant, the total growth is only one-third of that for a normal distribution, whereas an even distribution gives 70–80% of the total growth achieved with a normal distribution.

In the short term, the maximization of timber production would thus require an age structure representing a normal distribution, i.e., a forest area that is fragmented into single stands of variable age and stocking with the inevitable impact that this will have on species with a habitat corresponding to mature or old-growth forests. In the long term, this age distribution will gradually lead to dominance by old stands, with increasing stocking and decreasing growth—a situation that is less than optimal for timber production. Terminal felling of stands previously of intermediate age but now mature will lead to a shift in the age distribution toward dominance by young stands with low stocking and absolute growth but high relative growth. The maximization of timber production over several rotations in a sustainable manner thus requires that the forest area should be divided into sections representing a mix of sapling, pole, and mature stands in such a way that the overall age distribution implies dominance by pole stands with high stocking and high growth. The forest area will still
be fragmented in this case, which will mean major impacts on species with a preference for mature or old-growth forests.

IV. IMPACTS OF TIMBER-ORIENTED MANAGEMENT ON BIODIVERSITY: COMPARISON BETWEEN MANAGED AND NATURAL FOREST AREAS

A. Impact Mechanisms

Among the key issues in managing biodiversity in commercial forests is the maintenance of the functional and structural properties of the forest ecosystem, which will create and maintain heterogeneity among the habitats available for species to occupy. Two levels of heterogeneity may be distinguished: allogenic heterogeneity between stands and autogenic heterogeneity within stands. These concepts are related to allogenic and autogenic succession. Allogenic heterogeneity creates a landscape mosaic, whereas autogenic succession creates further heterogeneity within each element of the mosaic in terms of an even-aged stand structure or multilayered tree canopies (Table III). The contribution of allogenic and autogenic heterogeneity to the total heterogeneity of the landscape is probably multiplicative rather than additive.

The total impact of forestry on the availability of suitable habitats is difficult to predict. The optimization of age distribution for timber production may increase allogenic heterogeneity and fragmentation of the landscape. In particular, the remaining patches of mature or old-growth forests may be too small to sustain viable populations or too isolated for recolonization of areas where a population has temporally become extinct. Simultaneously, management may reduce autogenic heterogeneity in two ways. First, it may avoid or even curb successional processes that give rise to within-stand structural components and spatial patterns that are important for forest biodiversity (Table II). Second, regeneration practices and precommercial and commercial thinnings can reduce or even remove autogenic heterogeneity. This tendency is quite obvious in stands where a preference for forest habitats are dependent on dead wood (e.g., there are about 1000 fungi and about 800 beetles whose occurrence is related to the presence of such material). This implies that 36% of the total number of species threatened by forestry practices are ones whose occurrence is related to the presence of dead wood.

Table III. Many of these elements are interrelated, and practically all are somehow related to different types of disturbances and the subsequent succession. The amount of coarse woody debris is highest during the early successional phases in natural forests, whereas in managed forests no dead or dying trees are left at this stage. The scarcity of such debris is generally accepted as the main reason for the disappearance of many species from the boreal forests that are being managed for timber production. In Finland, for example, more than 8% of the species with a preference for forest habitats are dependent on dead wood (e.g., there are about 1000 fungi and about 800 beetles whose occurrence is related to the presence of such material). This implies that 36% of the total number of species threatened by forestry practices are ones whose occurrence is related to the presence of dead wood.

It was not known for certain until recently whether species depending on dead wood require closed-canopy old-growth habitats or whether they could make use
of younger successional phases if a sufficient amount of dead wood was available. Recent evidence suggests (Fig. 4) that the latter may be possible and that many threatened species may survive in managed forests if certain properties typical of natural forests are maintained or even increased. On the other hand, many threatened species have highly specialized habitats, e.g., sand ridges, riverbanks, or large, overmature aspen trees (Populus tremula). The majority of these key habitats are of marginal importance for timber-oriented forestry since they represent sites of low productivity or are quite small in extent. Given careful planning and management, these key habitats could have a major impact on the conservation of threatened species which prefer forest habitats. Along with measures to increase dead wood in commercial forests, this would provide new possibilities for reducing the impact of timber-oriented management on threatened species.

B. Properties of Forests under Intensive and Extensive Management: An Example

The impacts of timber-oriented management are exemplified in the following by referring to a comparison of selected beetle populations in an intensively managed forest area and an extensively managed one with a quite different management history (Siitonen and Martikainen, 1994; Siitonen et al., 1995). This material also provides a unique opportunity to contrast our current understanding with experimental data representing the long-term effects of forestry on the properties and biodiversity of boreal forests.

The two areas are located on either side of the border between Finland and Russia (62 or 638 N, 30–328 E) (Fig. 5). Before World War II, both areas belonged to Finland and were utilized and managed in similar ways. The border, created in 1944, split them in a biogeographically random manner.

The forests on the Finnish and Russian sides of the border have been used and managed quite differently for the past 50 years. Those in Finland have been subject to intensive use and a form of management that has included balancing growth with felling, systematic regeneration in the form of plantations established in clear-felled areas, or natural regeneration by means of seed trees or shelter wood fellings in favor of coniferous species. Precommercial thinnings and intermediate thinnings have been used to reduce the role of deciduous trees, and the latter have also been used to harvest the growth and yield that would otherwise have been lost in the natural self-thinning associated with the succession. Furthermore, the length of the rotation has been determined with the aim of maximizing the timber yield (80–100 years), which has further reduced the formation of dead and decaying wood. Additional reductions in dead wood have been achieved by effective fire fighting and the logging of trees blown down by storms or broken by heavy snowfall. The small-scale ownership of forest plots has further enhanced the fragmentation of the forests as each owner has aimed to maximize her or his timber yield and income.

On the Russian side of the border, clear-felled areas...
were not systematically regenerated, and thus fertile sites were occupied by mixtures of coniferous and deciduous trees. Since no precommercial thinnings or intermediate thinnings took place, the dominance of deciduous trees throughout the rotation was substantially higher than on the Finnish side. Later, dying trees were not logged in connection with thinnings, with the consequence that there is now a large amount of dead and decaying wood present. Furthermore, a low utilization rate of stems and the leaving of deciduous trees of low commercial value in the clear-felled areas were characteristic of logging practices. Simultaneously, uncontrolled forest fires were common and timber destroyed in other natural disturbances (e.g., storms and heavy snowfalls) remained unharvested. In other words, human intervention in the regeneration and growth of the forests has been substantially less than that on the Finnish side of the border.

C. Landscape and Local Patterns in Relation to Management

Systematic field samples from both sides of the border and analyses of satellite images indicated that the mean area of individual tree stands was quite similar but old-growth stands were substantially larger on the Russian side. Furthermore, the largest areas covered by homogeneous tree stands were larger on the Russian side. Most of the large stands on the Russian side were the result of extensive forest fires. On the other hand, the number of separate stands relative to the total area was larger on the Finnish side, i.e. the Finnish forest landscape is more fine grained and fragmented. This difference is quite evident even on the satellite image, which indicates the dominance of a coarse-grained landscape on the Russian side (Fig. 5).

Closer analysis of the properties of the forests shows that deciduous tree species are more common in the Russian forests than in the Finnish ones throughout the area (Fig. 6). Mature aspen in particular has been effectively eradicated from the Finnish forests because of its tendency to host a serious fungus disease that affects sapling stands dominated by pine. The most striking differences, however, concern sapling stands on fertile sites, which are dominated by Scots pine and Norway spruce in Finland, but both Pendula and Pubescent birch in Russia. Coniferous tree species naturally dominated the sapling stands on poor sites in Finland, but they also did so on fertile sites, whereas in Russia deciduous species exceeded 20% of the total number of stems per hectare even on poor sites in many cases. In other words, the forest cover more frequently represented a mixture of tree species on the Russian side than on the Finnish side.

In terms of dead and decaying wood, the difference between the two areas was clear; that is, the mean amount of dead wood in the Finnish forests was 4 m$^3$/ha$^{-1}$ (range 0.1–26 m$^3$/ha$^{-1}$) and that in the Russian forests was 29 m$^3$/ha$^{-1}$ (range 0.1–213 m$^3$/ha$^{-1}$) (Fig. 7). It is worth noting that the amount of coarse woody debris is high even in sapling stands in the Russian forests (up to 79 m$^3$/ha$^{-1}$), mainly as a result of forest fires. Here, the dead wood represents mainly dead trees (standing and lying), whereas in the Finnish forests it mainly comprises logging residues (stumps, branches, and tops of stems). These patterns held good over the whole range of variability in fertility and the age of sites and stands.

D. Beetle Fauna in Forests under Intensive and Extensive Management

Siitonen et al. (1995) assessed the impact of forest management on biodiversity in terms of the number of beetle species caught with window traps operating over one growing season. The total yield was about 29,000 individual beetles representing 623 species. There were
FIGURE 7 Volume of dead wood as a function of tree species and maturity of tree stands on the Finnish and Russian sides of the border. Spruce forest represents higher fertility than pine forest. No major differences in the total number of species or in the main groups of species, except for the species preferring dead wood, with the consequence that the total number of saproxylic species was 179 in the Finnish forests and 213 in the Russian ones (Table IV).

It was found that 17 of the 20 most frequent species with a preference for dead wood were more frequent on the Russian side of the border. These represented mainly species with a preference for decaying birches (e.g., Trichius fasciatus, Hylocoetus dermestoides, Leptura quadrifasciata, and Anaspis arctica) or for the final phases of decay in coniferous trunks (e.g., Anoplodera virens, Ampedus balteatus, Ampedus tristis, and Hadro- bregmus pertinax). The numbers of common and fairly common species in both countries (found consistently less than 25 times in Finland) were very similar, but the number of the rarest species (found consistently less than 12 times in Finland) was substantially greater on the Russian side (Table V).

The occurrence of species with a preference for dead wood was related to the amount of dead wood present in the tree stands, as shown in Fig. 8. On the other
hand, there was no correlation between the occurrence of other species and the amount of dead wood. The numbers of specimens of other species were inversely correlated with the basal area of trees in the stand, however, which implies that there are more beetle species in sapling stands and pole stands than in mature or old-growth stands, regardless of tree species and site fertility.

V. CONCLUSIONS

Forests are a natural resource of prime importance, and they provide human communities with tangible items such as timber and wildlife and intangible benefits such as scenic beauty, wind reduction, and urban noise abatement. A unique feature of forest production is that its control requires only a small external input. Natural processes exert the greatest control, in the form of the biological diversity that controls the functioning and structure of forested ecosystems. The management and utilization of forest resources in a way that balances human needs with the undisturbed functioning of the forest ecosystem is the key to a form of sustainable forestry which provides ample space for maintaining biodiversity in the world’s forests.

The current biodiversity of the boreal forests is a result of long-term evolution and changes in the environment and is thus related to the successional dynamics of the forest ecosystem. The long-term growth and development of tree populations is subject to disturbances due to wind, fire, and attacks by insects and pests, and under boreal conditions in particular, these disturbances cause a cyclic effect in the dynamics of the forest ecosystem by returning it to earlier phases in the succession. For example, a fire that kills old trees provides space for regeneration, enhancement of growth, and the accumulation of mass. Considered over large areas, this means that under natural conditions the boreal forests represent a mosaic of tree stands of varying age from saplings to mature trees and of varying tree species composition. The maximization of growth and timber yield implies that the area of mature or old-growth forests may be smaller than might be desirable in order to conserve the most specialized species.

The forest resources of northern Europe have been utilized intensively during the past 100 years, and this has inevitably had an impact on the current diversity of the forests. The management goal of maximizing the timber yield will inevitably lead to fragmentation of the forests and structural properties which deviate substantially from those arising under the influence of major disturbances. An increasing harvesting rate in terms of regular thinnings, together with a preference for coniferous species and a shorter rotation, will reduce the proportion of deciduous trees and the occurrence of dead and decaying wood. Furthermore, effective fire fighting and measures aimed at increasing the capacity of forests to resist the force of winds and heavy snowfalls will reduce the formation of dead wood and the occurrence of burnt wood, for which many species are specialized.

The previously mentioned tendencies will detract from the conditions which many rare and endangered species need in order to survive, as indicated by the positive correlation between the occurrence of these species and the amount of dead wood in the boreal forests. The increasing amount of dead wood probably indicates that there are more specialized microhabitats available under these conditions than in forests with a small amount of dead wood. These highly specialized species seem to be those affected most by the utilization and management of forest resources. On the other hand, the occurrence of species other than those confined specifically to dead wood seems to be quite similar regardless of the management history of the forests.
The protection of threatened biota may be based solely on a network of strictly protected areas. Currently, only 3.6% of the productive forestland in Finland is strictly protected, and opportunities to increase this within a short time span seem to be limited. Most of the valuable old-growth areas are already protected. Perhaps a better potential for protecting forest biodiversity and threatened species can be found in managed forests, in which the application of silvicultural practices that aim at restoring the natural properties of younger forests can be developed. Natural characteristics can be promoted quite rapidly in young managed forests since such forests are created continuously along with timber harvesting. On the other hand, most of the threatened species have highly specialized habitat requirements. By preserving these habitats (key habitats) in forest management, viable populations of these species can be maintained. Consequently, strictly protected areas and biodiversity-oriented management should be regarded as important complementary elements rather than alternatives when planning and developing new forestry practices.

See Also the Following Articles

Bibliography


TOURISM'S ROLE IN BIODIVERSITY might be to both build public support for biodiversity and help fund its conservation. However, the partnership has yet to be forged and current emphasis is on impacts.

I. INTRODUCTION

A. What Is Tourism?

Tourism can be defined as the sum of the phenomena and relationships arising from the interaction of tourists, business suppliers, host governments, and host communities in the process of attracting and hosting of these tourists and other visitors. Industries are typically thought of in terms of their output (for example, bales of wool), and measuring and assessing them involves taking into account activities that produce similar products. However, tourism does not conform to this model. It is not possible to reduce tourism to a simple relationship between inputs and the supply of a particular product. Tourism supply is defined in terms of the demand (of the buyer) rather than, as usual in other industries, in terms of the properties of the product and its production. The industry is dominated by marketing, and research has largely focused on estimating demand. It is part of the large and growing service sector of industry. While tourism does involve selling normal commodities, the core product that people are ultimately seeking is an experience. Experience is intangible, consists of activities rather than things, is produced and consumed...
TOURISM, ROLE OF

simultaneously, and the customer has to be present and participate in the production process. A region must have an attraction or attractions to be a tourist destination. Attractions are a necessary condition and they are arguably the most important component in any tourism product. Without attractions there would be no need for other tourism services. They not only entice, lure, and stimulate interest in travel, but they provide visitor satisfaction, the rewards from travel—the pure travel product. These attractions can be natural, manufactured, or of sociocultural character.

In one view, the ‘tourism system’ consists of four components:

- **Market.** The internal and external factors affecting travel, the market inputs and the process by which purchasers of tourism products select a destination.
- **Travel.** The market segments, the passenger flows and forms of transportation.
- **Destination.** Procedures that a destination should follow to develop and service tourism activity.
- **Marketing.** The processes by which suppliers and destinations market products and services.

In essence, marketing sells the destination to the market while travel allows the market to get to the destination. Another way of conceptualizing tourism is to look at the factors that contribute to a long-term profitable industry in a local area. Obviously not all of these conditions need be met to develop a successful local industry, but the more that are met the more likely the tourism destination is to be sustainable. Many of the insights represented in Figure 1 are derived from a large number of interviews done with tourism industry experts from regions across eastern Australia in 1997.

Different parts of this system are given different emphasis by different interests. Industry people tend to overemphasize the importance of marketing. Local government tends to focus on community support and local revenue from tourism. Tourism academics and management experts emphasize the chain of value represented by the series of experience points for the customer, of which the immediate experience of the attraction that the customers thinks they are purchasing is but one. Town planners focus on infrastructure and transport. Little, of course, would happen unless investors made money. Conservationists, in turn, focus on protection of the environment. Few people would have an overview of all the factors.

In the context of this chapter, it is important to recognize that protection of the environment is seen as only one of a range of important issues for tourism. When biologists talk of conducting research on the impact of tourism on biodiversity, many industry people feel threatened and hostile. As important as the impact issues are, a more effective approach is to package impact research in projects that deal with a wider range of issues. In particular, the quality of the tourism experience and using tourism as a vehicle for conservation education are two areas that naturally fit. However, they require social research and may need multidisciplinary collaboration with social scientists, a step many biologists are still unwilling to make.

It is also worth pointing out that the push for environmental care is mainly external. Governmental responsibility rests with sections of government outside the tourism system. This tends to be a recipe for tension and ineffectiveness. Educating the public in conservation, in turn, is unlikely to be high on the tourism industry agenda. It is high on the conservation agenda but will only be high on the tourism agenda if tourists are willing to pay for it.

**FIGURE 1**

**B. Growth in Tourism and Its Contribution to Economic Development**

Although people have always traveled to new places, the motivation of recreation rather than work is only a century or two old. The widespread phenomenon of today, in which most of the population of economically developed countries are tourists for part of the year, is a post–World War II development. Long-haul international destinations, such as many of the ‘megadiverse’ countries, were greatly enhanced as destinations by the reduction in airfares due to the widespread intro-
Tourism, Role of

The beach is still popular, but it is now much
important for the future of biodiversity. First, the pro-
important, tourism regions tend to be less developed
and more peripheral regions. Indeed, one of their at-
tractions is often an area's relatively unspoilt nature.
Most communities, however, can only hope for small-
scale economic advantages from tourism, and in such
areas tourism can be no more than one element in a
wider development strategy.
Critics of nature tourism relative to other forms of
tourism believe that nature tourists are young and on
low incomes, and therefore spend little and contribute
little to the local economy. Recent World Wide Fund
for Nature surveys in Caribbean countries on the impor-
tance attached by visitors to parks and protected areas
show that when people classified their motivations as
"main reason," "very important," or "not important,"
total expenditure per head was $2588, $1638, and
$1531, respectively. The average lengths of stay were
13.0, 13.8, and 14.7 days.

Some major patterns in the growth of tourism are
important for the future of biodiversity. First, the pro-
portion of tourists traveling to the undeveloped world
is increasing. It was 18.7% in 1980 but 21.8% by 1989.
However, the percentage of receipts (i.e., income) over
that period declined from 26.7% to 23% (Jenner and
Smith, 1992). Second, tourists are becoming more ac-
lative use of wildlife for recreation such as hunting and
fishing were more important than they are now. A report for Alberta Tourism
suggested that people involved in consumptive wildlife
use were mainly male (90%) and few held degrees.
Nonconsumptive use were mainly male (90%) and few held degrees
than they are now. A report for Alberta Tourism
suggested that people involved in consumptive wildlife
use were mainly male (90%) and few held degrees.
Consumer use is likely to be declining relative to
nonconsumptive use of nature. In Canada, noncon-
sumptive use of wildlife is growing more rapidly than
in the United States.
Concern about environmental issues has grown rap-
idly since the mid-1960s. In the United States, it peaked
in 1970 with Earth Day and then declined until the
1980s when spectacular mishaps like the Exxon Valdez
oil spill and the Chernobyl nuclear meltdown focused
media attention on environmental matters. Concern
continues to grow. In Europe, survey research indicates
that public concern was more constant in the 1970s,
rose steadily in the 1980s, and peaked in 1989 about
the time of the European elections (Macnaghten and
Urry, 1988). In Australia, support grew steadily from
the mid-1960s and was boosted by successive Australian
conservation issues, including the proposed mining of
the Great Barrier Reef, Fraser Island, and in Kakadu
duction of the long-haul Boeing 747 in the 1970s. It
was largely this technological development that made
tourism the global economic force it is today.
According to the World Tourism Organization, the
top three exports of world trade are crude petroleum
products, motor vehicles and parts, and machinery;
therefore, tourism, each between 8 and 9% of the total. Tourism
is the fastest growing sector of world trade, with an
average annual growth rate of 9.6% in the 1980s. This
comparisons with growth rates of 5.3% in merchandise
exports and 7.5% in commercial services. From 1950
to 1991, international visitor arrivals worldwide in-
creased 22 fold, from 25 million to 550 million, and
this number was expected to increase by an additional
200 million, or 36%, in the 1990s. It is estimated that
nature-based tourism might account for somewhere be-
tween 10 and 30% of all tourism, depending on whether
direct or more indirect criteria are applied. This would
indicate a contribution of between US$600 billion and
US$3 trillion in 1996.
In countries with a low population density like Aus-
tralia, the tourism industry is primarily based on the
country's natural attractions, particularly its flora and
fauna. While large city and other forms of urban tourism
are important, tourism regions tend to be less developed
and more peripheral regions. Indeed, one of their at-
tractions is often an area's relatively unspoilt nature.
C. What Do Tourists Want?
In the past, consumptive use of nature for recreation
such as hunting and fishing were more important than
they are now. A report for Alberta Tourism
suggested that people involved in consumptive wildlife
use were mainly male (90%) and few held degrees.
Consumer use was likely to be declining relative to
nonconsumptive use of nature. In Canada, noncon-
sumptive use of wildlife is growing more rapidly than
in the United States.
Concern about environmental issues has grown rap-
idly since the mid-1960s. In the United States, it peaked
in 1970 with Earth Day and then declined until the
1980s when spectacular mishaps like the Exxon Valdez
oil spill and the Chernobyl nuclear meltdown focused
media attention on environmental matters. Concern
continues to grow. In Europe, survey research indicates
that public concern was more constant in the 1970s,
rose steadily in the 1980s, and peaked in 1989 about
the time of the European elections (Macnaghten and
Urry, 1988). In Australia, support grew steadily from
the mid-1960s and was boosted by successive Australian
conservation issues, including the proposed mining of
the Great Barrier Reef, Fraser Island, and in Kakadu
National Park, hydro power schemes drowning Lake Pedder, and clearing of native forests in New South Wales. Some development proposals were successfully stopped and others were not. However, the public debate generated further support for conservation issues generally. Support probably peaked around the 1990 federal election, which was fought primarily over green issues. Since then neither of the major parties has chosen to make it an issue and both promote a strong pro-conservation image.

U.S. polls have shown an increase in the number of people who feel that “we must sacrifice economic growth in order to preserve and conserve the environment” from 38% to 64% between 1976 and 1990. The “Canada in the World” survey in 1996 shows that only Germany, Japan, Australia, and the United Kingdom have a majority of people saying that economics should give way to environment when there is serious conflict. Generally the surveys show there is most concern about pollution issues that affect everyday quality of life (Macnaghten and Urry, 1998). At a local level, people are concerned about pieces of remnant vegetation. Obviously campaigns to save the tiger and other high-profile species also attract strong support. The origin of national parks lay in the recognition of the importance of encounters with nature. Philosophers tell us that what is actually valued is not wilderness, but the experiences it produces: feelings of awe, experiences of beauty. The issue is complex.

There are many surveys about why people choose the holiday experiences they do. They tend to show the growing importance of environmental issues in such decisions. The environmental factors influencing the choice of holiday destination of tourists in Spain were beautiful landscape (51%), unspoilt nature (43%), water quality (27%), air quality (22%), old customs (16%), and architecture (13%). The purpose of Japanese overseas trips were found to be to enjoy nature (72%), see famous tourist attractions, (56%), taste local food (48%), enjoy shopping (43%), rest and relax (38%), experience a different culture (36%), visit museums (31%), and stay in famous hotels (22%) (Jenner and Smith, 1992). Foreign holiday makers in Germany showed increasing environmental concerns, with the percentage of people noting serious environmental problems increasing from 21.9% in 1985 to 46.8% in 1988. The percentage of people noting problems at least doubled, whether seas, lakes, streams, rubbish, dirty beaches, dead trees, degraded landscape, or noise were under consideration.

In Europe, biodiversity issues are relatively few in the main communication media, but there is great interest in local birds among bird watchers. In all countries, there are nature-based tourism experiences available. Examples of the diversity may be found in compilations such as Geffen and Berglie (1993). However, in some countries, nature-based tourism is promoted more heavily than it is in other countries. For example, in Australia any hotel lobby has brochures promoting a range of nature-based tours and attractions, whereas one does not find this in the United Kingdom.

While tourists are increasingly seeking good quality nature-based experiences, there is also growing concern about the impact of tourism on host communities and environments. Activist groups, such as Tourism Concern in the United Kingdom, publicize examples of negative practices by their country’s outbound operators in Third World countries. Such groups exist in many countries and primarily focus on the displacement of Third World peoples by tourism developments for people from the more developed countries. They frequently also attack conservation initiatives for the same reason. A range of somewhat anecdotal and non-quantitative assessments of the impacts of tourism are available for different regions (e.g., Croall, 1995; Patullo, 1996).

Most countries conduct surveys to determine where tourists come from, what they do, and where they go. They also ask why the tourists made the decisions they did. The interpretation of these results is complex (e.g., Ryan, 1993). In this section, I have tried to establish that concern about the environment and the desire or need for good-quality, nature-based experiences are strong drivers for tourism. In the next section I examine the attitude to the environment and biodiversity by the tourism industry.

D. Engagement of Tourism Industry Sectors with the Environment

1. General Industry Leaders

Clearly, this concern reflects a much broader societal alarm about the degradation of our physical environment. Tourism, however, is perhaps even more sensitive and more dependent on a high-quality environment for its long-term success than are many other sectors. There are many quotes by industry leaders recognizing the importance of the environment. The leaders of the industry recognize that tourism is much more than recreational sites or facilities for accommodation and entertainment. Tourists also consider scenery, weather, transport, the hosts, and many intangible factors, any
of which—singly or in combination—can be limiting factors.

A number of special features of the tourism industry offer peculiar challenges to our society and government. It is often said that the industry is fragmented and does not have a strong sense of itself. This fragmentation likely results because many of the major industry sectors are only partly in the tourism industry. For example, a restaurateur for whom 20% of customers are tourists and 50% are business people is likely to focus on the business trade. Many hoteliers, wine-makers, and restaurateurs do not regard themselves as part of the tourism industry for the purposes of devising their business strategy. Economically, of course, a not inconsiderable part of their income is derived from tourism. This diversity of viewpoints perhaps inhibits a strong common stand on issues.

The Pacific Asia Travel Association claims that the tourism income derived from what international visitors spend on sightseeing is only 3.3% in Hong Kong and 21.4% in the Philippines. In Australia, about 8% of the tourism industry income from international tourists is directly from tours, entertainments, and the like—that is, the sorts of things that people actually came to see. Perhaps half of that is strongly related to the natural environment and biodiversity. Put conversely, the tourists certainly did not come to experience the marvellous flight from another continent, to stay in a hotel, and to eat at a restaurant that could be in their home country. All industry sectors are dependent on the core product and are part of the "chain of value." If the core product fails, all sectors fail. Developing a sense of responsibility for the core product from all sectors may well prove to be the key to the development of a sustainable industry.

The results of a Tourism Council Australia survey of 208 Australian operators (accommodation, attractions, resorts, tours, and transport) asking about the attributes of the Australian tourism product are encouraging. The respondents gave 910 votes out of 1638 to natural environment categories ahead of hospitality/dining, service, price, entertainment, cultural/heritage attractions, special events, indigenous culture, and shopping.

2. Developers

Of the various components of the tourism industry, the people who build the resorts, hotels, shops, attractions, and so on seem least concerned about the natural environment. They build facilities that are usually sold and then managed by someone else as a tourism operation. Their main aim is to make money on their building activities. Architects generally try to produce an aesthetic building. Those that capture the "sense of place" are less common. That is, the buildings and environs become part of the environment in a harmonious and appropriate way. Such buildings are different and memorable. The landscaping around the buildings can not only enhance the appearance of the building but also greatly add to the amenity value of the development. Obviously, the use of local native plants or the recreation of local ecosystems can be a strong contribution to generating a sense of place in a local community. This can make the place special, interesting, and thereby more attractive to tourists.

3. Attraction Operators

The people who run the attractions that tourists come to see are generally closest to the natural environment. In my experience, the majority regard themselves as conservationists. However, they may find themselves excluded from natural areas by conservation agencies. Sometimes this is due to overzealous application of the precautionary principle, and operators then call for research to provide more impartial information. Wildlife can be an enormous long-term tourist drawcard, far outweighing the short-term profits to be made from transitory hallmark events. However, we need to get a lot better at managing wildlife tourism in a sustainable and holistic way. Tourism has the ability to encourage vast numbers of people to take a more caring approach to the natural environment. It also has the ability, like any industry, to simply see the natural environment as a resource to be appropriated for corporate profit irrespective of environmental impacts. It is important to generate a succouring social environment that facilitates the former option. Very often this falls to protected area managers. They need to encourage appropriate behavior in operators rather than show an antagonistic attitude through officious behavior.

Ecotourism was one of the buzzwords of the 1990s (e.g., Lindberg et al., 1998). It is an idealized form of nature-based tourism encompassing holistic education of the tourist, maximizing benefits to the local community, and encouraging dedication to the pursuit of sustainability. The substantial literature on ecotourism seems to focus on the important issue of benefits to the local community. The environmental side is still poorly researched and discussions of it tend to be vague and nonquantitative. The word has been co-opted as an advertising gimmick by a range of operators and seems to have a confused image with many consumers. For the tourism industry as a whole, ecotourism appears to have been a distraction from the pursuit of
As noble as the aims might be, the use of the term does seem to have had the effect of allowing other segments of the tourism industry to be less vigorous in pursuing sustainability than might otherwise have been the case. Sustainability should be everybody’s business.

4. Accommodation
There has been considerable attention given to environmentally sound behavior by hotel chains. Some have attempted to develop a reputation in this area for marketing purposes. For example, in the early 1990s, the Canadian Pacific and Hotels Corporation conducted a thorough environmental audit of all its operations and set goals in waste reduction, reuse and recycling, increasing energy efficiency, as well as purchase of “nature friendly” hotel supplies. Various governments and industry groups have also produced publications to encourage best practice. However, while operator interest is substantial, application is patchy due to both economic and regulatory restrictions. In general, policies of this sector of the tourism industry appear to have little impact on biodiversity.

5. Restaurants
In the quest for tourism, regions are encouraged to express their regional identity in a variety of ways including the presentation of local foods. In Australia, some restaurants now prepare dishes incorporating indigenous species. These locally based foods are generally not foods prepared or presented in a traditional way (“bush tucker” in Aboriginal vernacular), but are usually the creative integration of Australian biodiversity products with more traditional European and Asian foods to produce new foods. Native fish and waterfowl species have long been used, but kangaroo, emu, and crocodile meat, insects, and a range of fruit, seeds, and leaves of Australian plants are now also used. To a fair extent they are a tourist curiosity, and most larger centers of population have at least one restaurant that specializes to some extent on modifications of “bush tucker.”

6. Transport
“The business of tourism is, in reality, the renting out of the environment. It is therefore imperative for the industry to ensure that its ‘product’ is kept safe, unsullied and fresh not just for the next day but for every tomorrow” (Sir Colin Marshall, British Airways). Most of the biggest tourism operators are airlines. Because of the long lead times with purchase of aircraft, there is a greater interest in the long term. While airlines like Qantas and Air New Zealand promote themselves as environmentally aware, their interest has been focused on their own airline operations. Cathay Pacific, however, has focused its conservation effort on funding ecological studies of local high-profile species near the major destinations it serves. Perhaps because of the media stories on the impact of affluent European tourists on the society and environment of Third World countries, British Airways has funded environmental and social audit studies in the Seychelles and more recently St. Lucia. The studies include an assessment of a range of environmental and social issues. They take a matrix approach in examining how water pollution, air pollution, noise impacts, aesthetic impacts, habitat loss (land), natural resource exploitation, and disruption of natural cycles are affected by tourism infrastructure, infrastructure operations, transport, use and consumption, and waste at a location.

II. NEGATIVE IMPACTS OF TOURISM
Tourists go to the Caribbean for its climate, sea and beaches . . . and this has put its coastlines under enormous pressure . . . . The large concrete hotels have been built close to the high water mark, groynes and piers erected, marinas for yachts and deep-water harbours for cruise ships constructed . . . . The great wetlands of the Caribbean have been grubbed out by developers eyeing their proximity to some of the regions best beaches . . . . In a generation the land and seascape have been transformed. (Patullo, 1996)

There are a number of places around the world where tourism has seemingly run out of control and has come to generate a lot of problems and very few benefits for the local community. The negative impacts of tourism may be divided into the direct impacts of tourists experiencing nature and the indirect ones resulting from the impact of providing facilities for tourists. While there are reviews of the negative impacts of particular plants and animals on tourists (e.g., Edington and Edington, 1986; Mieczkowski, 1995), the emphasis of this brief review is on the impact of tourists on biodiversity. A number of frameworks have been used in the literature. For example, Wall and Wright (1977) describe the main ecosystem characteristics of impacts on water, soil, vegetation, and wildlife, while Edwards (1987) provides a matrix of habitats and direct tourism activi-
ties. There are many reviews of the negative impacts, but perhaps the most comprehensive are the 600-page tomes of Mieczkowski (1995) and Liddle (1997). There are also comprehensive treatments of specific ecosystem types (e.g. seacoasts, German Federal Agency for Nature Conservation, 1997).

A. Harvesting

The collecting, shooting, and trapping of wild food plants, wildlife, and fish, has the potential to be detrimental to populations of plants and animals. These consumptive activities can either be a tourist activity themselves or can arise from the goal of feeding or providing souvenirs for tourists.

There are many examples of overexploitation of resources by hunters and fishermen, some of this in a tourism context. For example, the Great Barrier Reef Marine Park Authority found that surveys conducted 10 years apart revealed that anglers' perceived catches had declined in numbers of fish and sizes of fish caught in four regions of the park.

Hunting and fishing, which are well regulated, are generally preferable to prohibition accompanied by illegal poaching (see Edington and Edington, 1986). The conservation of game species is enhanced in South Africa where well-managed game ranches have been established for decades and allow regulated hunting by residents and tourists. However, areas like Kenya have found nonconsumptive tourism more profitable (Whelan, 1991).

In Australia, the food used by indigenous Australians is called “bush tucker.” Bush tucker tours are usually based on a range of bush products, thereby increasing the risk of nonsustainable use of particular resources. There are publications on bush tucker, which may encourage usage by a wider range of people.

B. Small-Scale Physical Impacts

The impacts of tourists and their vehicles crushing and breaking plants has been well studied (Liddle, 1997). In addition to physical abrasion and breaking of the plants, trampling compacts the soil, particularly with some soil types, and renders it unsuitable for plant growth primarily due to lack of water capacity of the soil. Unbordered paths tend to widen laterally over time as people walk out to the expanding limit of ground vegetation. People sometimes create their own paths where there is crowding, where the established pathway is overly circuitous, or where there are vantage points unserviced by paths. Poorly sited and designed paths may have the effects of trampling exacerbated by water-driven erosion during rainy periods. The number of passages by walkers to reduce the vegetation by 50% can vary between 12 and 1412 for different ecosystems (Liddle, 1997). The type of usage is also important. For example, per unit passage, horses have a greater impact on trail width than motorcycles, which in turn have a greater impact than hikers.

Campsites experience the same sort of trampling impacts as paths with a gradient of impact extending out from the most intensely used center (Cole, 1992). However, campsites are also affected by the collection of dead and living wood for campfires. Snow sports fragment habitats and destroy lichen cover (Jenner and Smith, 1992).

Boats can have a range of impacts through direct contact, wash, and propeller action, which can cause bank erosion, washing out of roots, and turbulence. These impacts cause redistribution of nutrients and increased turbidity with consequent effects on aquatic plant and animal life (Liddle and Scorgie, 1980).

There have also been studies of the impact of boats and divers on coral. Research done by the Great Barrier Reef Marine Park Management Authority indicates that coral breakages occur, but at levels that are statistically indistinguishable from the natural background of breakages from wave-action and storms. Floating hotels and day-trip pontoons are larger and have a wider range of impacts. The largest day-use pontoons may cater to in excess of 50,000 snorkellers per year with up to 700 at one time. Impacts are greater at low tide when visitors are able to stand on the coral. In Malaysia, beach closures occur at such times to protect the coral. In the Red Sea, it has been shown that careful briefing of divers decreases coral breakages.

C. Spreading Exotic Plants and Animals

In many parts of the world, local species have existed in isolation from the major land masses. Their biotas have developed high endemicity and consequential vulnerability to exotic invasion. Modern communication allows goods and people to move around the world in vast quantities. With the large volume of traffic moving between countries, there are numerous opportunities for organisms to be accidentally or illegally moved from one part of the world to another. For islands generally, with their long isolation, many potential pests have not arrived, but conversely some of those that have survived the journey have had a devastating impact. This is best documented for Hawaii and Australia.

It is likely that tourists have inadvertently moved...
small plants and animals between countries. Certainly the spread of weeds into natural areas is well documented. For example, Lonsdale and Lane (1994) collected 1832 seeds of 76 species from 304 tourist vehicles entering Kakadu National Park; they found that 13 species had not previously been recorded in the park and 9 species were known tropical weed species. While these authors found most of their seeds trapped in the vehicles’ radiator, in an African area with poorer roads many more were in mud stuck to the undersides of the vehicles where the seeds were also more likely to detach in a new location.

The transmission of diseases by tourists is a less common problem but is of great importance in a few situations. For example, direct transmission of disease is a long-standing concern for mountain gorillas, which are highly susceptible to human viruses and bacteria. These include tuberculous, measles, and pneumonia, all of which could potentially wipe out a population of this highly endangered species (Roe et al., 1997). There are similar concerns that the introduction of Newcastle Disease to Antarctica through infected poultry products could wipe out much of the bird life there.

D. Disturbance of Wildlife

There is a substantial North American literature on disturbance to wildlife by recreationists (Knight and Gutzwiller, 1995), and most of this is relevant to tourism. There are four broad causes of impact: harvest, habitat modification, pollution, and disturbance. Separately or in tandem they may cause death or modify the behavior of individuals. In the longer term, behavior, vigor, and productivity may be altered. Again death may result. As a population level, abundance, distribution, and demographic structure may be altered. Finally at a community level, interrelations between species (e.g., predation, competition) may be altered and species composition changed.

Sometimes tourism-derived impacts on a species can be diverse. For example, boats, sandcastle diggers, cars, motorbikes, beach umbrellas, and plastic bags are all hazards to loggerhead turtles at different stages of their life cycle on the Greek island of Zakynthos (Ryan, 1995). Conversely, some activities can have a range of impacts. For example, snowmobiles seem to have a range of impacts, but disturbance of wildlife during winter from this source can be particularly severe (Jenner and Smith, 1992). The Black Grouse (Tetrao tetrix) has been the subject of various studies in relation to the impact of skiing in the French Alps and is in some danger as a species. Another impact has been the increase in litter and waste due to tourist presence, which has attracted predators and severely reduced the success of egg laying (Jenner and Smith, 1992).

In northern Australia’s Kakadu National Park, a study showed a range of species’ responses to tourists on tour boats (Brinkwate et al., 1996). Some species habituate quickly while others remain very sensitive to approach by humans. Each year as the waterbody contracts through the dry season the animals rehabilitate to close contact with humans. Thirteen species were particularly sensitive to disturbance by the boats of tourists, and it was suggested that these be used as indicators of disturbance.

Currently, much of this wildlife is accustomed to the boats. Some neither move away nor move toward the boat, but allow the boat to come close. They behave this way only if they do not feel too threatened. If they feel threatened, some “freeze” and appear to hope not to be seen, but most move away. In moving away they experience stress, stop feeding, and expend considerable energy. If they do this often enough they are unable to breed. If they do it more they may not obtain enough energy to survive. Before these things happen, animals usually abandon the area. The relationship between the visitor and the fauna is often largely determined by a guide or operator.

One particularly difficult issue is that a crocodile leaping into the water or a large flock of magpie goose flying off is spectacular. Many tourists enjoy that spectacle. This means that the tourist may be more satisfied with their experience if the boat operator/guide regularly and severely disturbs the wildlife. We suspect that the right atmosphere is not being created in such cases. Some tourists will go to extraordinary lengths to get best photograph possible. Is there a compromise, a level of disturbance that satisfies the tourists, but is also acceptable to the animals?

A similar situation is found with wildlife tours in Nepal. Sloth bears and tigers are most easily disturbed with their experience if the boat operator/guide regularly and severely disturbs the wildlife. We suspect that the right atmosphere is not being created in such cases. Some tourists will go to extraordinary lengths to get best photograph possible. Is there a compromise, a level of disturbance that satisfies the tourists, but is also acceptable to the animals?

A study for the Great Barrier Reef Marine Park Management Authority showed that large numbers of predatory fishes aggregated when the boats came to pontoons on the Great Barrier Reef. Regular fish feeding seemed to encourage such a buildup. However, there was no evidence that the feeding aggregations depleted fishes from other areas of the reef or that they affected local populations of fishes and invertebrates.

Another disturbance increased by tourists is wildfire.
Accidental ignition of wilderness areas is increased by visitation. Such changes in the frequency and time of year of these fires can have important impacts on both fauna and flora. Such problems are likely to be most common in fire-prone regions of the world such as those with Mediterranean type and wet-dry tropical climates.

E. Land Clearing as Part of Development

The primary cause of the loss of biodiversity is the habitat destruction that inevitably results from the expansions of human populations and activities. Obviously tourism developments and any facilities that even a few tourists use have often caused removal of natural vegetation. A substantial part of world tourism results from urban people going to less urban places in warmer climates. Many tourism developments occur close to attractive natural areas that are rich in biodiversity. Even in Europe, land use for tourism has been growing significantly, particularly in the Alps, the Mediterranean coast, and the North Sea coast. The European Union (EU) expects this use and the more extensive forms of tourism such as hiking, cycling, fishing, and hunting to increase at the expense of land formerly dedicated to nature or agricultural purposes.

In some cases, tourism developments are less destructive of biodiversity than the extractive industries such as forestry and agriculture that they replace. For example, in Bali and other parts of Asia, the gardens created around hotels, resorts, and golf courses offer some improvement for biodiversity over the previous intensive rice agriculture. Major landscape modification is an extreme form of land clearing. Two particular forms associated with tourism are building marinas and golf courses. The direct physical impacts of building marinas include loss of marine habitat from land clearance, reclamation and dredging, changes in water flow within marina basins, and reduced water quality associated with waste disposal, refuelling, antifoulants, and effluent discharge (MacMahon, 1989).

With more than 50 million active golfers worldwide, golf is possibly the fastest growing sport, and golf course development has become the fastest growing property sector in the world. Improved technologies have enabled turf-sward treatment and large-scale removal of soil to change radically the landscape for golf courses. A scorched earth school of golf-course design has emerged whereby the land is flattened and recontoured from scratch. With a water demand of 3000 cubic meters a day, pumping up large amounts of underground water is depleting water in springs in natural areas. The fairways are increasingly being made broader to accommodate more people. Large amounts of fertilizers, pesticides, herbicides, fungicides, and other chemicals are needed to maintain the semblance of the Scottish coastal landscape at St Andrews. Japanese golf courses use 8.5 times as much pesticides as rice paddies, two tons per golf course annually. Much of it makes its way into surrounding systems (Pleumaron, 1992). The biodiversity impacts of these golf course developments are likely to be substantial but do not appear to be well documented. How can golf courses be made more environmentally benign? Clearly this is an important issue for biodiversity.

F. Pollution and Resource Use

If a country has a positive balance of payments for tourism, then the excess of inbound over outbound tourism effectively represents an increase in population size for such places. Such increase is a particularly heavy user of transport and all motorized transport contributes to pollution of the air and often water as well. The full range of human impacts, usually at the extravagant end of the per capita level of impact, are added to that of the resident population increasing the ecological footprint disproportionately.

In Hawaii, the discharge of partially treated sewerage effluents into Kaneohe Bay stimulated growth of the alga Dictyosphaerium crenata such that it overgrew and killed large sections of the reef (Johannes, 1975). The use of water by tourism ventures is in competition with other users, potentially domestic, industrial, and agricultural. In areas of the world where there is heavy dependence on groundwater, tourism activity will contribute to lowering the level of aquifers. This affects dependent fauna and flora as water no longer flows out on the surface. In drier parts of the world, such resources support major components of the biota, including many specialized species.

G. Perception of Impact by Tourists

Visitors to natural areas are most observant of, and regard as important, the direct impacts of other users (trail use for more than one activity, litter) but are becoming more aware of other impacts on the environment such as soil erosion (Hammitt et al., 1996). In central Australia, M. Hillery and colleagues demonstrated a positive relationship between annual visitor numbers and level of measurable environmental impact, despite a relatively small level of impact. Fifteen percent of visitors were able to rate two or more sites for impact.
Over half of the visitors in the study identified concerns relating to tourism and introduced species. Only 8% identified other environmental issues. Surprisingly, some visitors were able to rate the area they were in against other parts of Australia. These results suggest that environmental quality and relative lack of impact are going to be commercially important in that consumers not only regard them as significant but also are able to discriminate with increasing accuracy. This bodes well for an improving relationship between tourism and conservation.

III. POTENTIAL BENEFITS OF TOURISM

The ugly face of tourism is often evident to travelers. This sometimes leads people to ignore the positive impacts of tourism and thinking about how the positive impacts might be maximized. Much of the writing on the positive effects is by social scientists and thus is different from the style of information accumulated by biologists on the negative impacts.

A. Building Public Support for Conservation

Many tourism destinations owe their popularity to conservation controversies. In Australia, Kakadu, South-West Tasmania, Fraser Island, and the Daintree River owe much of their tourism activity to all the free publicity on national media during conservation battles. Night after night, images of these areas are shown on prime-time television. They become very fashionable places to visit. A whitewater rafting operator in north Queensland suggested to me that announcement of a government intention to build a dam on the river he uses would be great for his business.

Conversely, travel to a beautiful place generally builds some bond with that place. If tourists have a positive experience they will care about it not being destroyed. Even the most sensitive new development destroys the place as people have it in their memories and is resented at least to some extent. The biological essence of a place needs to be emphasized as part of the tourism experience. It is not just identifying plants and animals, but natural history stories and, most important, big ideas about ecology and biogeography. Generally most tour guides in most places cover a small part of the range of information available for interpretation.

Obviously the more the quality of the interpretation improves, the better. Recent surveys of tourists visiting Kakadu National Park reveal that the "perceived needs for additional facilities and services" are largely to do with the desire for more information. The rankings of percentages were, in descending order, more information on native plants and animals, more opportunities to have direct contact with Aboriginal people, more information on Aboriginal culture, more information on the geology of the area, more information signs, places to buy Aboriginal arts and crafts, better maps of the park, more toilet facilities, more guided tours to places of interest, more places to buy food and drinks, and finally more picnic facilities. Clearly the demand is there. A huge amount of information is available but relatively little of it is accessible to tourists. The problem is getting it to the operators in contact with the tourists.

B. Funding for Conservation Management

In the era of user pays, protected areas are already under great pressure to extract money directly from the public for the use of resources, both artificial and natural. Currently, the level of self-funding of protected areas ranges from zero to greater than 100%, with most less than 10%. The highest levels of self-funding are seen in Africa.

In the world of international trade where different traditions of subsidies have arisen in different countries, countries with large estates of publicly funded protected areas are already being accused of unfairly subsidizing their tourism industry. If world trade liberalization continues, international diplomatic pressure will add to the internal economic rationalist push. Others argue that public good requires ongoing subsidization of natural areas.

A key issue is how to do this in the most functional way. Some observers argue that resources are managed better in agencies that are funded out of fees rather than tax dollars and that full market prices should be charged. There is the issue of whether private industry, governments, or a parastatal organizations (i.e., corporate bodies within government) would work best for different aspects (entrance fees, accommodation, equipment rental, food sales, merchandise sales, etc.). Further, a range of financial mechanisms is available including user fees (either for general admission or for a specific activity), concession fees (charges to provider per visitor or as a general license fee on the operator), royalties on sales, special taxes, fines, and donations.
Tourism, Role of

There seems to be much greater public acceptance of fees if it is clear that all the money goes to fund conservation rather than into "consolidated revenue." Multitiered systems with some groups paying more than others are also an important consideration. The most efficient and most acceptable method will vary with circumstances within and between countries. For example, some government agencies are reluctant to allow private concessions because the mechanisms for enforcement of standards of environmental care are inadequate or too prone to political interference.

C. Sustainable Regional Development

Across the world there have been many government schemes for redevelopment of rural areas in population and economic decline. The main problem has been the decline in value of agricultural commodities. However, in some cases the depletion of resources through use has also been a factor. Tourism is widely seen as an important option for economic recovery. The European Union increasingly views tourism as a better option than agricultural subsidy for adjusting for prosperity between regions. The EU’s regional policy is increasingly focused on the role of tourism in socioeconomic development and the need to encourage responsible use of local resources for tourism. Economically disadvantaged rural areas are encouraged to develop ecotourism activities. Approximately 40% of the total budget of the ECU of $1755 million (Community contribution) for the community initiative called Leader II will be made available to develop rural tourism. Many Leader projects combine rural tourism activities with the protection of indigenous species and marketing based on protected areas of diversity.

In the United States, the Forestry Service has policies for aiding rural economic diversification through promoting the assessment of amenity resources. These refer to the aspects of the rural environment in which residents and visitors may find beauty, pleasure, and experiences that are unique to their locales. Such a process promotes a feeling of local pride and is more likely to result in a successful local tourism industry based on the local assets than might result from asking how the region can make money out of tourism.

D. Tourism in the Third World

Increasingly the nature-based tourism of people from more affluent countries is located in poor Third World countries where substantial natural areas are still available. In many of these less-developed countries, tourism has proved to be more valuable than traditional industries like agriculture. However, the key to the long-term success of such ventures is the level of returns to the local people. For example, in the Galapagos, limited returns to local people from international tourism forces them into unsustainable exploitation of sea cucumber and other natural resources, which may in turn have large impacts on the dependent species in the ecosystems that are the basis of the tourism industry there (Southgate, 1998).

Many international aid projects now attempt to develop tourism sustainably in the poorer countries of the world. In fact, tourism is probably the only long-term prospect for funding conservation in much of the world. However, unless the local economic and social benefits are substantial, there is no prospect of success. Part of this problem is the lack of understanding by governments of the scale of economic benefits to be had from the tourism activity. For example, a study in the Virgin Islands National Park shows that the contribution by tourists to the island’s economy was 10 times that of the annual management expenditure (Heywood and Watson, 1996). It is also essential that the solutions to the biodiversity-tourism nexus are harmonious with the local culture as well as consistent with local environmental and social conditions. The models for success are likely to be various.

Former WTO Secretary-General Antonio Enriquez Savignac attended the Rio Earth Summit in 1992 and was instrumental in getting tourism included in Agenda 21 as one of the only industries capable of providing an economic incentive for preservation of the environment. Secretary-General Francesco Frangialli renewed WTO’s commitment to the goals of sustainable development in 1997 at the United Nations Earth Summit II in New York.

IV. THE BIG ISSUES

There are some difficult issues concerning biodiversity and tourism, which will require research and careful negotiation to resolve.

A. Engagement in Conservation Issues

While the tourism industry often has a bigger stake in environmental issues than traditional industries, it has yet to involve itself as a political protagonist in this area. The parts of the industry closest to the natural environment like tour operators are often ardent conservationists. Many are in it for the lifestyle rather than...
the money. In Australia, 85% of the industry is in small businesses and must focus strongly on the bottom line to survive financially. The larger businesses tend to be more remote from environmental issues and are more interested in other issues. As the industry matures, it is likely to take a higher profile on conservation issues in which it has a substantial stake.

B. Protected Areas and Tourism

The commercial world of the tourism industry and the governmental approach by park managers do not mesh easily. Control tends to rest with the park managers and to build a synergistic relationship with tourism they need to do the following:

• Develop empathy for the difficulties most small businesses have to survive financially.
• Respect third party legal obligations to tourists (i.e., not changing the rules at short notice).
• Sell the value of enlightened regulation to the tourism industry.
• Pursue policies and provide assistance, which will help to improve the quality of the tourism experience.

C. Carrying Capacity

The idea that beyond some point the resource degrades with greater usage, and the quality of the experience for visitors declines, seems simple enough. Many in the industry accept this in theory, but they do not like the idea that there are limits to the growth of their region or their business. The marketing paradigm suggests that the demand for a product is only limited by the quality of the salesmanship. The dominant view seems to be that the supply or carrying capacity can always be increased by appropriate “site-hardening.” Supply can, however, exceed demand leading to cost cutting and a decline in quality and sometimes economic nonviability. On the other hand, low profitability may be offset at a community level by substantial employment. Research on the environmental, economic, and social tradeoffs in different situations is needed to develop a more sophisticated approach to this difficult issue.

D. Wildlife Tourism

Tourism based on the exposure to wildlife (animals) possibly offers the most satisfying of nature-based experiences but presents some special problems. For example, the experience may be intensified by severely disturbing the wildlife and the impacts may be more widespread and significant for conservation than with other nature-based experiences. On the other hand, the challenge is to increase the probability of the marvelous experiences that field biologists occasionally have in the course of their work.

E. Interpretation and Education

Visitors can gain from experiences physically, emotionally, and intellectually. However, satisfaction with a tourist experience, vacation, or holiday is the fulfillment of motivations. The motivations become goals, goals determine behavior, and in the search for a satisfying holiday, holiday makers engage in adaptive behaviors to secure the success of a satisfying vacation (Ryan, 1999). How does one meet expectations? How can people be helped to find the right places for them to do the things that they will most enjoy? What interpretative/educational material should be provided that will both meet the visitors’ needs and also raise awareness and support for conservation issues? An example of this sort of research was done on the Skyrail chairlift over rainforest in north Queensland. The research showed visitors valued the interpretive computer interactives over all other components of the experience, including the interpretative center as a whole (Moscardo and Woods, 1998).

Much of the world of biodiversity is inaccessible to most tourists. Innovative ideas can allow a wider range of experiences to be shared with a wider range of people. For example, at the Naracoorte Caves Conservation Park in South Australia a new interpretative center substantially enriches the normal cave visits. Remote control cameras, infrared lighting, and image enhancement technology are used to enable visitors to see a range of bat behavior, including roosting, flocking, birthing, and feeding. Such visitor-driven systems could be applied to a wide range of sensitive habitats.

Part of the problem is that people often have a limited image of nature-based tourism. For example, the Penguin Parade at Phillip Island in southern Australia caters to a maximum of 4000 visitors per night. This well-managed attraction allows a large number of people to see hundreds of penguins come from hunting in the sea to return to their burrows in the sand hills. The facility illuminates the area, allows close proximity, and provides good quality interpretation, but it also monitors the welfare of the penguins closely. Run by a local trust, the facility provides an intimate nature experience to a large number of people with minimum impact on the animals. The admission fees, in turn, fund a major
research program on penguins and the facility is a tourism icon, drawing visitors to a range of lesser known attractions in the region. It is a fine example of what can be done.

Tourism should be seen as entertainment, striving to satisfy tourist needs. To be successful, and therefore commercially viable, the tourism product must be packaged in a way that is attractive to the consumer. To assume tourists wish to be educated about biodiversity—or anything else—would be a mistake. Thus the tourism educational agenda for conservation must be creatively developed if it is to be successful.

See Also the Following Articles

CONSERVATION EFFORTS, CONTEMPORARY • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • EDUCATION AND BIODIVERSITY • EDUCATION, BIODIVERSITY AND • ENVIRONMENTAL MOVEMENT • HUMAN IMPACT ON BIODIVERSITY • NATURAL RESERVES AND PRESERVES • RECREATIONAL USES AND ISSUES • ZOOS AND ZOLOGICAL PARKS

Bibliography


TRADITIONAL CONSERVATION PRACTICES

Carl Folke and Johan Colding
Stockholm University and Beijer International Institute of Ecological Economics

I. Introduction
II. Ecological Monitoring and Practices Framing Access to and Use of Species and Habitats
III. Practices of Multiple-Species Management, Resource Rotation, and Succession
IV. Practices Related to the Dynamics of Complex Ecosystems
V. Concluding Remarks

GLOSSARY

institutions Humanly devised constraints that shape human interaction and the way societies evolve through time; made up of formal constraints (rules, laws, constitutions), informal constraints (norms of behavior, conventions, self-imposed codes of conduct), and their enforcement characteristics.
natural disturbance Any relatively discrete event in time that disrupts ecosystem community or population structure and changes resources, substrate availability, or the physical environment; key for structuring biological communities and for maintaining resilience in ecological systems.
resilience The system's capacity to absorb disturbance and conserve opportunity for self-organization and evolution. Resilience has to do with how resistant the system is to fundamental reorganization such as a phase shift into another stability domain.
social taboo A prohibition imposed by social custom or as a protective measure.
traditional ecological knowledge A cumulative body of knowledge, practice, and belief, evolving by adaptive processes and handed down through generations by cultural transmission, about the relationship of living beings (including humans) with one another and with their environment.
traditional peoples Variously referred to as indigenous peoples, native peoples, or tribal peoples; peoples with cultures that differ from those of the mainstream of a national population, having distinct ethnic languages with locally evolved resource management systems; currently some 5000 distinct indigenous peoples exist in the world.

PEOPLE HAVE INHABITED terrestrial ecosystems of the world for thousands of years. Both resource management systems and cosmological belief systems have evolved and continue to develop. In fact, most, if not all, ecosystems and biodiversity have been altered by humans to various degrees (Nelson and Serafin, 1992). The human imprint has in many cases wiped out species and caused substantial land use change (e.g., Turner et al., 1990; Wilson, 1992). However, there are practices of local peoples of both traditional and contemporary society that contribute to biodiversity conservation, practices that are more common than generally recog-
I. INTRODUCTION

Even traditional peoples with a relatively high human population number were able to utilize rain forest areas without destroying these environments and surrounding biological communities (Primack, 1993). However, the specific objective of these local practices is not necessarily directed toward the conservation of species, their habitats, and ecosystems. Rather, such practices are often geared for sustainable use of local resources and ecosystems, with biological conservation resulting as an indirect outcome. They are often tied to cultural belief systems, which makes it difficult to separate the belief component from actual management practices and the ecological knowledge system on which they build. Knowledge, practices, and beliefs tend to intermingle in most traditional management systems (Gadgil et al., 1993). This constitutes the basis behind traditional ecological knowledge (TEK), denoting that resource management patterns are the products not only of a people’s physical environment and its resources but also of their cultural perceptions of the environment and its resources (Ruddle, 1994).

Cultural belief systems of various peoples have protected species, their habitats, and even smaller ecosystems. For example, several verses in the Vedas and Upanishads mention conservation and protection of plants and animals, indicating that traditional conservation practices of many rural and indigenous groups of India go as far back as the Vedic period (3000 B.C.). In fact, a great deal of social mechanisms, such as social taboos, may be highly adaptive from an ecological perspective and contribute to biodiversity conservation (Colding and Folke, 1997). The term conservation is here used in the sense of sustainable use of natural resources for human benefit, without compromising the interests of future generations (WCED, 1987).

This article will provide examples of a diverse set of traditional management practices and institutions that exist among local resource users. We will illustrate that such practices play an important role for in situ conservation of biological resources. We deal with traditional management practices that lead to the sustainable use of biological resources and local institutions and belief systems that impose regulations on the use of species and ecosystems.

The word traditional refers here to the historical and cultural continuity of resource management, recognizing that societies are constantly redefining what is considered “traditional.” The term ‘local resource users’ refers to both traditional peoples and small-scale societies of Western countries with locally evolved management systems. Local resource users generally depend on a rather limited resource procurement base to provide them with a wide diversity of resources (Gadgil et al., 1993). Many do not have access to fossil fuel dependent technology or capital markets. Their day-to-day survival depends on proper interpretations and knowledge of the dynamics of their local resources and ecosystems. Thus, they often have a stake in managing their resources for long-term endurance.

Many of the examples of management practices and local institutions presented in this article derive from the case study anthology edited by Berkes and Folke (1998). This anthology predominantly deals with local resource users that display success of long-term environmental management by using practices that contribute to building resilience in local ecosystems. A diverse set of management practices and social mechanisms from temperate and tropical regions and traditional and contemporary society were identified and analyzed in recognition of their importance for building resilience in combined social–ecological systems (Levin et al., 1998). The focus of the analysis is presented in Fig. 1.

The examples stem from a wide range of local resource users, including hunters and gatherers, herders, fishers, agriculturists, and small-scale communities of industrial nations. This article describes these practices in relation to their nature conservation functions. Most of the practices described are interrelated and have multiple ecological functions. Nonetheless, we have categorized them in the manner outlined in Table I.

As indicated in Table I, the first category of practices includes ecological monitoring and regulation of the use of biological resources and ecosystems. They are dealt with in Section II of this article. The second category mainly concerns traditional agroforestry practices, with an emphasis on their effect for biological conservation. They are dealt with in Section III. The third category describes management practices that are related to the dynamics of complex ecosystems, in particular the processes that structure ecosystems at different temporal and spatial scales. These practices are dealt with in Section IV. Such conservation practices, often at the watershed and landscape levels, are important in secur-
TRADITIONAL CONSERVATION PRACTICES

In the concluding section, we stress lessons that can be learned from traditional conservation practices in building resilience through improved management of biodiversity. Resilience conserves options and opportunity for ecosystem renewal and evolutionary change, and its dynamic maintenance is a prerequisite for all forms of biological conservation.

II. ECOLOGICAL MONITORING AND PRACTICES FRAMING ACCESS TO AND USE OF SPECIES AND HABITATS

The practices described below are often determined by local social institutions. These are often informal, based on traditional norms and conventions. Some appear to be more directly associated with resource management activities, while others appear to be more closely associated with cultural belief systems.

A. Monitoring the Status of the Resource

Monitoring the status of the resource is very common among local resource users (Folke et al., 1998). Monitoring leads to the acquisition of local ecological knowledge and helps local users respond to resource and ecosystem dynamics. Both qualitative and quantitative indicators develop as the result of experience and ecological knowledge based on monitoring (Johannes, 1998). Monitoring may provide information about location and timing of target resources. Ecological monitoring also provides the basis for the spatial and temporal regulations of resource use in local communities. Thus, various social response mechanisms evolve as a result of ecological monitoring. These are often in the form of local institutions, such as social taboos, area rotation regulations, and seasonal closures. The list of the importance of ecological monitoring for local resource users can be made very long. We will provide a few examples.

For resource procurement reasons, the Cree of James

---

TABLE I

Management Practices of Local Resource Users

<table>
<thead>
<tr>
<th>Number</th>
<th>Practice Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Ecological monitoring, and practices, framing access to and use of species and habitats</td>
</tr>
<tr>
<td></td>
<td>Monitoring the status of the resource</td>
</tr>
<tr>
<td></td>
<td>Total protection of certain species</td>
</tr>
<tr>
<td></td>
<td>Protection of vulnerable life histories of species</td>
</tr>
<tr>
<td></td>
<td>Protection of specific habitats</td>
</tr>
<tr>
<td></td>
<td>Temporal restrictions of harvest</td>
</tr>
<tr>
<td>2.</td>
<td>Practices of multiple-species management, resource rotation, and succession</td>
</tr>
<tr>
<td></td>
<td>Multiple-species and integrated management</td>
</tr>
<tr>
<td></td>
<td>Resource rotation</td>
</tr>
<tr>
<td></td>
<td>Management of succession</td>
</tr>
<tr>
<td>3.</td>
<td>Practices related to the dynamics of complex ecosystems</td>
</tr>
<tr>
<td></td>
<td>Management of landscape patchiness</td>
</tr>
<tr>
<td></td>
<td>Watershed management</td>
</tr>
<tr>
<td></td>
<td>Managing ecological processes at multiple scales</td>
</tr>
<tr>
<td></td>
<td>Responding to and managing pulses and surprises</td>
</tr>
<tr>
<td></td>
<td>Nurturing sources of ecosystem renewal</td>
</tr>
</tbody>
</table>

*a Source: Folke et al. (1998).*
Bay, Canada, monitor geographic distribution, migration patterns, and individual behavior of caribou, as well as the sex and age composition of the caribou herd, including the presence or absence of predators. They also predict herd size of caribou by examining the fat content of caribou (Berkes, 1998). The Inuit of Quebec and the Inuit and Innu of Labrador also use fat content as an indicator of health of both individual animals and the herd. Such monitoring may provide useful information for resource procurement and in the enforcement of hunting regulations when prey is low in abundance.

Similarly, the Cree of James Bay monitor the interaction between beaver and vegetation (Berkes, 1998). Such monitoring leads to knowledge about beaver populations and their ecological interaction with other biological resources. For example, when aspen is low in a particular area, trappers know that the beaver population is low in abundance.

Knowledge about the effect of natural disturbance regimes on target resources and other species is also based on ecological monitoring. Such knowledge requires monitoring over extensively long time periods and may be passed on from one generation to the next. For example, Cree beaver trappers know that 3–4 years after a fire occurs in an area, beavers will start to inhabit the area again (Berkes, 1998).

Also, monitoring combined with ecological knowledge may lead to resource procurement methods that do not deteriorate the habitats of target resources and secure a sustainable use of target resources. For example, in Tokelau in the South Pacific, islanders use, based on detailed knowledge of octopus behavior, a special octopus stick to extract the animal, which obviates the need for the destructive crushing of the coral or the use of poison (Ruddle, 1994). Another example is the use of the clam rake in the Maine soft-shell clam fishery, which is the only tool allowed for clam harvesting (Hanna, 1998). Such methods reduce the risk that a resource is overexploited, as compared to more advanced forms of technological methods.

Local institutions for management, based on ecological monitoring, may either be permanent in time such as in the two examples above, or ‘kick in’ occasionally in times of resource scarcity. For example, ecological monitoring underlies the Cree’s decision when to rotate beaver trapping grounds and fishing areas. Cree fishers rotate fishing grounds based on a declining catch per unit of effort and rest these sites when needed (Berkes, 1998). In this way, overfishing is limited. Coastal communities in Maine monitor clam populations to help determine the areas needing enhancement (Hanna, 1998). Such response mechanisms represent fine-tuned ways of responding to ecological feedback and are employed in order to avoid overexploitation of subsistence resources.

The widespread uses of closed seasons in Oceania are based on observations founded on local ecological knowledge about the spawning periods of key fish species, and fishing is prohibited during such periods. Pacific island groups often know when and where fish aggregate to spawn. At such aggregation sites, fishers monitor yearly changes of fish stock size and composition and reduce their fishing effort when stocks seem to be low (Johannes, 1978; Hviding, 1988).

Similarly, the use of fishing taboos among Pacific islanders at particular sea areas is based on local knowledge about the fluctuations of reef fish populations. On Satawal Island in the Central Caroline Island of Micronesia, fishing off one reef section is prohibited by taboo in order to conserve a breeding ground to supply the rest of the reef with resources (Ruddle, 1994). On Tokelau, the lafu (fishing taboo) is the most explicit conservation measure. Lafu is invoked by the Council of Elders. It bans all fishing in specific areas on the main reef and is announced to permit stock recovery. Such examples allowing for ecosystem renewal is evident among a number of local resource users. For example, pastoral groups in arid and semiarid Africa monitor ecosystem change by way of tracking to determine daily movements of herds. This includes tracking ecological processes, such as climate, soil content, groundwater availability, forage availability, temporal environmental variability, and environmental degradation. Various forms of indicators are used for this, such as the behavior of fauna, specific plant indicators, soil color and texture, forage quantity and quality, and vegetation/plant composition (Niamir-Fuller, 1998).

Due to the proximity of local resource users to their resource base, ecological monitoring is often feasible on a daily basis. This facilitates detection and response to ecological change. Quite often, monitoring activities may be ascribed to particular individuals in local communities, such as resource stewards, elders, or shamans (Berkes and Folke, 1998).

B. Total Protection of Certain Species

Total protection of certain species is another practice that is particularly common among traditional peoples. For example, flora and fauna in India have been protected by indigenous belief systems for millennia and have been revered as the vehicles of Gods and Goddesses. Of particular interest are the different kinds of social taboos imposed on species by traditional societies...
throughout the World. Colding and Folke (1997) found that specific-species taboos protect threatened species as well as species considered key species and/or endemic by ecologists. It was estimated that about 30% of the identified taboos protect species listed as threatened by IUCN. Table II displays a number of ecologically important and threatened species protected through taboos. This indicates the role that traditional local institutions, such as taboos, may have in biological conservation.

C. Protection of Vulnerable Life

Histories of Species

This practice reduces the danger of overharvesting and the depletion of a population of target resources. For example, in the Maine fisheries, it is prohibited to gather lobsters with eggs (Acheson et al., 1998). Traditional fishing castes in the Bhandara district of Maharashtra, India, never disturb the spawning aggregations of freshwater fish in hill streams, and the Phaseaparids of Harar traditionally let loose pregnant does caught in their snares (Gadgil, 1987). The Cree of James Bay never kill or disturb nesting geese (Berkes et al., 1995). The Tukano Indians of Colombia impose taboos on the collection of bird eggs and avoid the collection of reptiles during their breeding season as well as protect fish spawning aggregation sites in rivers (Reichel-Dolmatoff, 1976). Size restrictions are sometimes employed among South Pacific islanders on slow-moving or sessile marine species that are particularly susceptible to overharvesting (Johannes, 1978).

D. Protection of Specific Habitats

This practice is commonly found among local resource users. For example, pastoralists of arid and semiarid Africa use buffer zone areas of Sahelian rangelands

<table>
<thead>
<tr>
<th>Species</th>
<th>Popular name</th>
<th>Local resource users/locality</th>
<th>IUCN status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rinosternum oaxacae</td>
<td>Oaxaca mud turtle</td>
<td>Pima Bajo, Papago, Yuman, U.S.A./Mexico</td>
<td>I</td>
</tr>
<tr>
<td>Chelonia mydas</td>
<td>Green sea turtle</td>
<td>Inhabitants of Buzios Island, Brazil</td>
<td>E</td>
</tr>
<tr>
<td>Naja oxiana</td>
<td>Oxus cobra</td>
<td>Local protection in vicinity of temples in India</td>
<td>K</td>
</tr>
<tr>
<td>Melanochirus niger</td>
<td>Black caiman</td>
<td>Pito de Peru</td>
<td>V</td>
</tr>
<tr>
<td>Heloderma suspectum</td>
<td>Gila monster</td>
<td>Riverine Pima, Papago, U.S.A./Mexico</td>
<td>V</td>
</tr>
<tr>
<td>Paro masticus</td>
<td>Green pread</td>
<td>Tamol-Huacho, Rajastan, Guajara, India</td>
<td>V</td>
</tr>
<tr>
<td>Gorilla gorilla</td>
<td>Gorilla</td>
<td>Edo state, Nigeria</td>
<td>V</td>
</tr>
<tr>
<td>Colobus polykomos</td>
<td>Black and white colobus</td>
<td>Villages of Ambilawo and Fiuma, Ghana</td>
<td>V</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>Chimpanzees</td>
<td>Edo state, Nigeria</td>
<td>V</td>
</tr>
<tr>
<td>Thomomys umbrosus mexicanus</td>
<td>Southern pocket gopher</td>
<td>Riverine Pima, Papago, Martha, U.S.A./Mexico</td>
<td>V</td>
</tr>
<tr>
<td>Perognathus altissica</td>
<td>White-eared pocket mouse</td>
<td>Riverine Pima, Papago, Martha, U.S.A./Mexico</td>
<td>V</td>
</tr>
<tr>
<td>Dipodomys mexicanus</td>
<td>San Quintin kangaroo rat</td>
<td>Riverine Pima, Papago, U.S.A./Mexico</td>
<td>E</td>
</tr>
<tr>
<td>Canis lupus</td>
<td>Grey Wolf</td>
<td>Bishnois of Thar desert, Rajasthan, India</td>
<td>V</td>
</tr>
<tr>
<td>Teramiconis cruzi</td>
<td>Spectacled bear</td>
<td>Ache of Ecuador/Peru</td>
<td>V</td>
</tr>
<tr>
<td>Panthera tigris</td>
<td>Tiger</td>
<td>Local protection in vicinity of temples in India</td>
<td>E</td>
</tr>
<tr>
<td>Felis bengalensis</td>
<td>Puma</td>
<td>Mantoca, Yuman speakers, U.S.A./Mexico</td>
<td>E</td>
</tr>
<tr>
<td>Tapirus indicus</td>
<td>Central American tapir</td>
<td>Coshorewa, Pata de Brazil/Venezuela, Ache of Ecuador</td>
<td>V</td>
</tr>
<tr>
<td>Myrmecophaga tridactyla</td>
<td>Giant anteater</td>
<td>Coshorewa, Pata de Brazil/Venezuela, Ache of Ecuador/Peru</td>
<td>V</td>
</tr>
<tr>
<td>Pecari tajacu</td>
<td>Giant armadillo</td>
<td>Achuar of Ecuador/Peru</td>
<td>V</td>
</tr>
<tr>
<td>Antilope cervicapra</td>
<td>Blackbuck</td>
<td>Bishnois of Thar desert, Rajasthan, India</td>
<td>V</td>
</tr>
</tbody>
</table>

* Source: Colding and Folke (1997).
* IUCN status: E, endangered; I, indeterminate; V, vulnerable; K, insufficiently known; R, rare.
which are protected from grazing except in the case of emergencies (Niamir-Fuller, 1998).

Also, whole forests, forest patches, coast stretches, rivers, or ponds may to various degrees be protected for human resource use. Usually, such areas are set aside by religious taboos and considered sacred to community members. In India, sacred groves were once extremely common. A sacred grove is a small part of a forest set aside for spiritual or religious purposes. The sizes of such protected areas vary from a clump of 3–10 trees to as much as 30 ha or more (Gadgil and Vartak, 1976). Sacred groves still exist in many parts of contemporary India—for example, in the Khasi Hills in Assam, in the Arvalhi ranges of Rajasthan, all along the Western Ghats in the southern peninsula, in the districts of Bastar and Sarguja in Madhya Pradesh, and in the Chanda district in Maharashtra. Gadgil and Chandran (1992) indicate that the Indian traditional shifting cultivation system, jhum (described in Section III.C.), is associated with sacred groves.

Kenya has sacred groves all along its coast, known as hayas (homesteads), used for ceremonies and burials (Wilson, 1993). So do the Yoruba of Ara in southwestern Nigeria (Warren and Pinkston, 1998). In South America, the Kuna Indians of Panama have spirit sanctuaries, places where spiritual animals, plants, or demons are believed to reside (Chapin, 1991). According to cosmological beliefs, the Kuna must respect these areas often located on choice agricultural land. The Cocnucos and Yanaconas of Colombia have similar sanctuaries (Redford and Maclean Stearman, 1993). The Yukano of the Uaupés basin on the Brazil–Colombia border reserve the forested river margin for fish and fishing. Fishing may be restricted to as little as 38% of the total river margin available (Chernella, 1987). The result is a management system that allows for, yet distinguishes, human use areas and animal refuge areas. Any deforestation of the river edge is prohibited.

The importance of such cultural beliefs for preserving patches of ecosystems should not be underestimated. New species of plants, and species that have disappeared from other areas, are still being discovered within sacred groves (Mohannan and Nair, 1981). A botanical survey in a Nigerian sacred grove yielded 330 plant species as compared to only 23 in surrounding nonprotected areas (Warren and Pinkston, 1998). Sacred groves are also important for maintaining ecological services, such as preserving local hydrological cycles, preventing soil erosion, serving as firebreaks, and serving as areas of recruitment of species, allowing for ecosystem renewal in face of various disturbances.

E. Temporal Restrictions of Harvest

This practice is adopted among some local resource users. The idea is that certain biological resources are protected from exploitation for certain periods of time. Among local resource users, the imposition of temporal taboos regulates access to resource(s) on either a sporadic, daily, weekly, or monthly basis (Colding and Folke, 2003).

For example, clans of Tiopka in the Solomon Islands impose sporadic taboos on particular foodstuffs they are associated with (Chapman, 1985). Several different durations of closed seasons exist among Vamatu fishing villages, ranging from 1 month to 5 years (Johannes, 1998).

At the Sakumo and Djange Lagoons in Ghana, taboos are imposed on fishing during a particular day every week (Ntiamoah-Baidu, 1991). In India, taboos imposed on a monthly basis appear quite widespread. For example, many castes abstain totally from consumption of fish, poultry, and meat, and suspend all hunting as well, in the Hindu month of Shavana (roughly August), which is the peak of the main rainy season over most of India (Gadgil, 1987). Similarly, taboos on hunting certain animals from July to October exist in many Indian villages. The Oraons, the fourth largest tribal group in India, observe a taboo of hunting any wild animal or bird during the months of June and July (Xasa, 1992).

In Papua New Guinea, the Maspa people in the Marshall Lagoon impose a seasonal ban on hunting (Kwapan, 1984). Every 3–4 years at the end of their cultivation work a bush area is set on fire to chase and hunt any animals found. This traditional hunting ritual lasts for a week. Once a particular animal-rich area is selected and hunted on, it will be left alone for another 3–4 years. This ritualistic way of hunting allows for hunting areas to regenerate. It may also mimic natural disturbance regimes. Groups of Canadian Amerindian hunters use a similar system for rotating hunting, fishing, and trapping areas by their periodic resting (Berkes, 1998).

In the Maluku Islands of eastern Indonesia, temporal prohibitions on gathering are imposed on terrestrial plant and animal resources, such as coconut, nutmeg, areca, and cuscus, as well as on marine species, including fish, trochus, shell, sea cucumber, and seaweed. This traditional management institution, known as sasi, is continually changing, depending on changing social and ecological conditions (Soselisa, 1998).

Locally decided and implemented closed periods is one of the measures used for managing kombu kelp (Laminaria augusta) at Hokkaido Island, Japan. Regul-
tions initiated for kelp harvesting as early as 1881 stated that villagers could not harvest kombu in the evening or on a rainy day, when good quality kombu cannot be produced (Iida, 1998).

III. PRACTICES OF MULTIPLE-SPECIES MANAGEMENT, RESOURCE ROTATION, AND SUCCESSION

This category constitutes management practices that largely can be described as agroforestry practices, although hunter and gatherers may employ methods that draw on the same principles. Agroforestry is the generic name used to describe an old and widely practiced land use system in which trees are combined spatially and/or temporally with agricultural crops and/or animals (Farrell, 1987). They are common in traditional agricultural systems throughout the world, especially in the tropics, but were once common in temperate regions as well (Beets, 1990).

The three types of practices described in the following are often related to the efficient use of local ecosystems in terms of space, nutrient availability, climatic conditions, and soil conditions. These practices serve to actively manage species diversity and intensity of use of subsistence resources for ecosystem renewal and species recruitment. Such practices often create habitat heterogeneity on the landscape scale, affecting forest structure and species composition by creating a mosaic of forest patches of different ages (Primack, 1993). Thus, the outcome of these practices often enhances diversity of biological resources at the local level (Oren-yuela, 1992).

A. Multiple-Species and Integrated Management

The cultivation of several species of crops on the same piece of land is often referred to as polycultures or intercrops. There is an enormous variety of this type of practice, and they may involve the mix of annual crops with other annuals, annuals with perennials, or perennials with perennials. Crops may be sown in different spatial arrangements and may range from the growing of two crops in alternate rows to complex assemblies of a dozen or more species (Altieri, 1987). Integration of crop cultivation with animal species is also common in traditional agroforestry systems.

Polyculture cropping systems constitute at least 80% of the cultivated area of West Africa and predominate in other parts of Africa as well as in the Latin American tropics and many Asian countries (Liebman, 1987). Frequently, more yield can be harvested from an area sown in polyculture than from an equivalent area of monoculture without the application of synthetic fertilizers, pesticides, and field machinery. Polyculture cropping systems may provide several social–ecological benefits as compared to monocultures (Liebman, 1987)—for example, by reducing the risk of total crop failure and by contributing to an increase of nutritional returns to local resource users.

Polycultures may also promote soil water conservation and nutrient recycling as well as reduce soil erosion and insect pests. Some polyculture systems may also reduce crop diseases and increase weed control. Polycultures may thus provide for many ecological services, including the maintenance and enhancement of biological diversity. In fact, diversity is used to enhance productivity and resilience (Berkes et al., 1993).

The two most common polyculture systems in Java, Indonesia, are Talun-kebun (rotation between mixed gardens and tree plantation) and pekaranan (home garden intercropping system including animals) (Altieri, 1987). A Talun-kebun is an indigenous Sudanese agricultural system, consisting of three stages that each serves a different function. In the first stage, heban, a mixture of annual crops is usually planted, mainly for cash income. After 2 years, tree seedlings begin to grow in the field, and the heban gradually evolves into a keban-campuran, where annual crops are mixed with half-grown perennials. This stage promotes soil and water conservation. After the annuals are harvested, the field is usually abandoned for 2 or 3 years and becomes dominated by perennials, the talun stage (perennial crop garden). The talun can be turned back into a heban after the forest is cleared or be planted to rice paddy, depending on whether irrigation water is available. A Talun-kebun may be composed of up to 112 species of plants, of which about 40% provide for building materials and fuel wood, 20% are fruit trees, 15% are vegetables, and the remainder is used for ornamentals, medicinal plants, spices, and cash crops (Altieri, 1987).

The integration of animals in polyculture systems is a common practice in diverse regions of the world. In Southeast Asia, rice ecosystems include diverse animal species. For example, domestic ducks may be integrated in paddy rice cultivation to allow for insect and weed control, and fish species, such as common carp, may be integrated and harvested at the end of the rice-growing season. The rice/fish culture differs considerably from country to country and from region to region (Altieri, 1987).
In China, as in Southeast Asia, crops, chicken, duck, and fish are often integrated to provide for a very high overall production through waste recycling and use of residues (Yan and Yao, 1989). Also, natural control of insects and weeds is provided for in these systems (Beets, 1990). In Indonesia, traditional systems combined rice and fish culture, and wastes from this system often flowed downstream into brackish water aquaculture systems (tambah). The tambahs were polyculture ponds, often combining fish, vegetation, and tree crops (Gadgil et al., 1993).

In Rajasthan, India, local farmers grow crops and rear livestock under and among trees. The Bishnoi, inhabiting one of the driest climatic zones on Earth, integrate the cultivation of food crops and animal rearing among islands of khejri trees, Prosopis cineraria, which are strongly protected by religious taboo (Sankhala, 1993). The Bishnoi keep as many khejri trees as possible on their farms, providing them with material for fencing, firewood, and fodder and pods for cattle as well as providing excellent microclimatic conditions for crop cultivation. Even threatened species, such as the blackbuck and gray wolf, benefit from these farming systems (Colding and Folke, 1997).

In many parts of Africa, "farm trees" can also be found within and adjacent to farm fields. The trees are actively protected, harvested, and managed by farmers to yield construction poles, fuel wood, fodder, edible fruits, nuts and leaves, medicines, and other products without unduly competing with associated annual crops (Beets, 1990).

B. Resource Rotation

Resource rotation exists among several communities of local resource users (Folke et al., 1990). Chisasibi Cree hunters rotate trapping areas (ideally) of beavers on a 4- or 5-year cycle to reduce disturbance to feeding and resting geese and to harvest for subsistence needs with a minimum disruption of the large population that passes through an area (Gadgil et al., 1993).

The Awa Indians of southwestern Colombia and neighboring Ecuador hunt echimid rodents on a cyclical basis, which allows sufficient time for the recovery of wild populations (Orejuela, 1992). Arid and semiarid African pastoralists migrate seasonally in accordance with plant availability, determined by precipitation and lunar cycles (Niamir-Fuller, 1998). The yearly cycle of nomads and their cattle provides a rotational management system that enables the recovery of heavily grazed rangelands. For example, the Wodaabe Fulani follow the lunar cycle when moving to new pastures, which means that the camp is moved every 2–3 days. By contrast, the Rufa’a al Hoi of Sudan move to a new pasture every 204 days. The Fulani of northern Sierra Leone once practiced "shifting pastureage," whereby they heavily stocked an area for 2–3 years and then moved elsewhere and rested the first area for 15–20 years. These are all examples of small-scale movement, or micromobility, practiced by many Sahel pastoralist tribes (Niamir-Fuller, 1998). Local resource users enforce grazing rotation in parts of the Hindu–Kush Himalayas in a similar fashion (Jodha, 1998).

C. Management of Succession

Polycultures and crop rotation often involve management of succession, as exemplified by the different systems of shifting cultivation that exist in the world. Shifting cultivation is defined by FAO (1982) as "a farming system in which relatively short periods of cultivation are followed by relatively long periods of fallow." Although part of polycultures, it can be distinguished by fallow periods that are ideally very long.

Shifting cultivation, or "slash-and-burn" cultivation or "swiddens," involves the clearing of a plot of land, usually a forest area, its use for a few years, and, as soil fertility declines, its abandonment in favor of another plot of land to be cleared in the same fashion. It is one of the oldest forms of agriculture and most present-day agricultural systems have evolved from it. In shifting cultivation, agriculture becomes a sequential cropping of crops and noncrops. Presently, about 3–500 million people, or about 6% of the total agricultural population of developing nations, depend on shifting cultivation for their daily livelihood. In total, shifting cultivation covers about 30% of the world's exploitable soils (Beets, 1990).

Shifting cultivation is common in all tropical areas and was once common in temperate regions as well. It has received much attention as one of the major degrading processes in tropical forest areas due to human population increase that greatly has led to shortages in fallow periods with subsequent loss of soil fertility. Generally, if fallow is less than 3–7 years, land degradation occurs and species diversity may be greatly reduced (Berkes et al., 1995). However, if adequate long fallow...
is allowed for, shifting cultivation may be highly sustainable (e.g., Posey, 1985).

Extremely high crop diversity is a characteristic of many traditional shifting cultivation systems. For example, Philippine swidden cultivators can distinguish over 600 plant species (Beets, 1990). In general, swiddens do not compare in complexity to the surrounding forest. However, when shifting cultivation is analyzed as an agroforestry system, i.e., the use of trees is also taken into account, then the overall result of managing forest patches can lead to an enhancement of biodiversity (Berkes et al., 1995).

For example, the Runa Indians in the Ecuador Amazon increase species diversity in 5-year-old fallows as compared to unmanaged swiddens. Between 14 and 33% of this enhanced species diversity was attributed to direct planting and protection of secondary species.

Different forms of swiddens and related fallow systems exist in the world. In tropical Mexico, as well as in other Mesoamerican countries, milpa (maize fields) is widely practiced (Alcorn and Toledo, 1998). Milpa involves the clearing of new fields in high forests, or secondary regrowth forests, for maize cultivation. If more than one successive crop of maize is taken in a short fallow milpa, weedy species come to dominate the plot and forest regeneration may not occur. Milpa is governed by strong informal institutions, which reinforces reciprocity and community-based control of natural resources. Farmers who mismanage their milpa are labeled as witches and punished by social pressure.

Each stage of the milpa cycle is also named and marked by ritual activities, rendering milpa a strong sociocultural practice. The Awa Indians of southwestern Colombia and neighboring Ecuador practice a shifting agriculture known as jhum (Orejuela, 1992). It involves the cutting of natural vegetation and the mulching of this material for a temporary agricultural field. Maize and/or short-cycle varieties of red and white beans are planted by the slash-mulch method. Once yields decline after several cycles of harvest, the field is left fallow and permitted to regrow as an enriched secondary forest. A similar system involving mulching is used to clear lots to raise cattle.

In northeastern India, detailed studies on shifting agriculture (jhum) have described multispecies systems involving 4 to over 35 crop types based on locally adapted native strains (Berkes et al., 1995). This practice requires sophisticated local ecological knowledge, including the use of soil nutrients by adequate changes in the crop mixture depending on the length of jhum cycles and the consequent availability of soil nutrients.

On hill slopes, farmers combine r-strategist species (cereals and legumes) with K-strategists, with emphasis on vegetative growth, such as leafy vegetables (Berkes et al., 1995).

An interesting form of shifting cultivation by the Kayapo Indians of Brazil constitutes the creation of forest islands known as apete (Posey, 1985). This begins as small mounds of vegetation about 2 m in diameter (apete-ns). As planted crop and tree seedlings grow and the planted area expands, the taller vegetation in the center of the mounds is cut to allow light to enter. A full-grown apete has an architecture that creates zones that vary in shade and moisture (Fig. 2). The species mix includes medicinal plants, palms, and vines that produce drinking water. Of a total of 120 species found in 10 apete, Posey (1985) estimated that 75% may have been planted. Apete constitutes the manipulation of semi-domesticated plants on which the Kayapo could survive during times of warfare. These old forest islands have been scattered for millennia in known spots throughout the forest and savanna.

Practices of multiple species, resource rotation, and succession management contribute to biodiversity conservation relative to more modern and technology-dependent ways of exploiting resources and ecosystems. By actively using biodiversity in production, they contribute to biodiversity conservation in areas outside protected areas and reserves.

IV. PRACTICES RELATED TO THE DYNAMICS OF COMPLEX ECOSYSTEMS

Apete is also a good example of traditional practices that actively promote patchiness and heterogeneity at the landscape level. Many agroforestry systems make use of patchiness for multifunctional benefits, although patchiness may also be used in smaller cultivation systems to provide for crop protection and natural pest management (Altieri, 1987).

A. Management of Landscape Patchiness

Such management is employed by Sahelian pastoralists in order to mimic the variability and unpredictability of the landscape. For example, an appropriate mix of herding animals is used to utilize different vegetation types and patches in a dynamic fashion (Naumir-Fuller, 1998). The progressive widening of grazing radius around wells as the wet season advances is also an example of active management at the landscape level.
Among the Kayapo Indians of Brazil, enhancement of biodiversity is facilitated by the creation of forest islands, *apete*. Through a number of management methods, this behavior promotes patchiness and heterogeneity in the landscape in time and space. Source: Gadgil et al., 1993.

This practice is employed by the Maasai of Kenya to leave enough forage around wells for the dry season (Niamir-Fuller, 1998).

Fire management among Australian aborigines and northern Canadian Amerindians was practiced widely to open up clearings (meadows and swales), corridors (trails, traplines, ridges, grass fringes of streams and lakes), and windfall forests. These clearings provided improved habitats for ungulates and waterfowl, thus increasing hunting success, and the corridors and windfall areas improved accessibility (Gadgil et al., 1993).

The use of different elevation zones for cultivators in the eastern Himalayas and other parts of the world represents practices that make use of habitat heterogeneity and create patchiness at the landscape level. The ancient *ahupua‘a* system of Hawaii utilizes different elevation zones, determined by precipitation, that are used for various integrated farming practices (Costa-Pierce, 1987). In the ancient *ahupua‘a*, both freshwater and seawater fish ponds were integrated with agriculture, and river valleys were managed as integrated systems, from the upland forest (left uncut by taboo) all the way down to the reef (Costa-Pierce, 1987).

**B. Watershed Management**

The *ahupua‘a* also constitutes an example of traditional watershed management systems. Southeast Asia and Oceania had, and to lesser extent still have, a number of these prescientific ecosystem management practices. Examples include the Yap tabinau, the Fiji vanua, and the Solomon Islands puava (Ruddle et al., 1992). These all refer to generically similar watershed-based management systems. For example, the vanua concept is an integrated human–nature component that regards the land, water, and human environment as one unit, one and indivisible (Ruddle et al., 1992). Similarly, the puava includes all resources and land in a watershed, from the top of the mainland mountains to the open sea outside the barrier reef (Hviding, 1990). In all these cases, the social group inhabiting the ecosystem unit is considered to be part of the system.

**C. Managing Ecological Processes at Multiple Scales**

There is some evidence that traditional management systems may be useful in managing ecological processes at multiple scales. In *milpa*, described earlier in Section III.C, food crops are managed on a 1- to 3-year scale, and some tree crops and products on a 30-year scale (Alcorn and Toledo, 1998). James Bay Cree hunters seem to be simultaneously managing beaver populations on a 4- to 6-year scale, lake fish on a 5- to 10-year scale, and caribou on an 80- to 100-year scale (Berkes, 1998). The holistic forestry of the Gitkian people of northern British Columbia simultaneously manages the production of fiber over several square kilometers with ecological processes involving soil bacteria at the spatial scale of a few square meters (Pinkerton, 1998).
D. Responding to and Managing Pulses and Surprises

Natural disturbances, such as fire, hurricanes, pest outbreaks, and heavy grazing, are inherent to the internal dynamics of ecosystems and often set the timing of ecosystem renewal processes (Holling et al., 1995). An example of a practice that responds to disturbance and manages pulses and surprises is the establishment of range reserves within the annual grazing areas of African herders. These reserves provide an emergency supply of forage that serves as a buffer when disturbance, such as drought, challenges the process and function of the dryland ecosystem (Niamir-Fuller, 1998). At the landscape level, this function to maintain the resilience of both the ecosystem and the social system of the herders. Such practices may be considered ecological adaptations to unpredictable, low-rainfall environments and provide “ecological insurance” to the local communities.

Sacred groves in India absorb disturbance by serving as firebreaks for cultivated areas and villages (Gadgil et al., 1998). At the same time, these sacred habitats function as recruitment centers for species regeneration and ecosystem renewal. In times of natural disturbances and other emergencies, such areas of refugia are essential for providing social-ecological resilience.

Many of the traditional forms of polyculture previously described may provide for pest and insect protection that contributes to building resilience (Altieri, 1987). The Warlis of India control pests by placing certain kinds of tree branches in their paddy fields. This practice serves to attract birds for insect control and buffers against outbreaks of various pest populations (Pereira, 1992).

E. Nurturing Sources of Ecosystem Renewal

Many traditional societies seem to actively nurture sources of ecosystem renewal by creating small-scale disturbances. Traditional agroforestry practices such as shifting cultivation create forest gaps and enable people to produce crops or enhance wild foods without disrupting natural renewal processes (Berkes et al., 1995). Such practices may even enhance genetic and ecological diversity in the benign mosaic landscape of forest, fallow, and gardens (Orejuela, 1992). Also, traditional rotational cycles employed by hunters and gatherers, described in Section III, allow for species recovery of wild populations.

Aboriginal use of fire in as geographically diverse areas as Canada, Australia, and California had many elements and principles in common (Berkes et al., 1995). Until the late 1940s, Amerindians of northern Alberta, Canada, regularly used fire to create clearings (meadows and swales), corridors (trails, traplines, ridges, grass fringes of streams and lakes), and windfall forests (Lewis and Ferguson, 1988). These clearings provided improved habitat for species such as ungulates and waterfowl. Australian aborigines possessed detailed technical knowledge of fire and used it to improve feeding habitat for game and to assist in the hunt itself (Lewis, 1989).

African herders behave like a disturbance by following the migratory cycles of the herbivores from one area to another (Niamir-Fuller, 1998). Pulses of herbivore grazing contribute to the capacity of the semiarid grasslands of Africa to function under a wide range of climatic conditions. If this capacity of the ecosystem to deal with pulses is reduced, an event that previously could be absorbed can tip the grassland ecosystem into a relatively unproductive state, dominated and controlled by woody plants for several decades (Walker, 1993).

V. CONCLUDING REMARKS

One of the 10 principles of the Global Biodiversity Strategy concerns the linkage of biodiversity with cultural diversity and the conservation of the two together. At a broader level, traditional conservation practices as a part of cultural diversity may offer benefits in terms of local biological and ecological understanding, sustainable resource management systems, implementation of protected areas, development planning, and environmental impact assessment. Each of these potential benefits is related to biodiversity conservation, either directly or indirectly. Biodiversity is recognized as an important component of sustainable use. It is in this sense that it is being conserved through traditional conservation practices as illustrated in this article through practices framing access to and use of species and habitats, practices of multiple-species management, and practices related to biodiversity conservation in complex dynamic ecosystems. Such practices seem to contribute to maintaining and building ecological resilience (Holling, 1986).

Traditional conservation practices rely on the accumulation of ecological knowledge and understanding over many generations, and knowledge is embedded in
local institutions and transmitted culturally. Practices are generally site and context specific although the ecological knowledge embedded in them may be spatially and temporally transmitted.

The locally generated knowledge and associated conservation practices are not necessarily complete in the sense of understanding of all aspects of an ecosystem and its dynamics. However, many traditional conservation practices are in line with the shifting scientific view on the nature of ecosystems as nonlinear, multiequilibrium, and full of surprises, threshold effects, and system flips. Predictability and controllability are not limited by the scientific data available but by the very nature of ecological systems.

Traditional ecological knowledge systems, based on detailed observations of the dynamics of the natural environment, feedback learning, social–ecological system linkages, and resilience-enhancing mechanisms, seem akin to adaptive management (Berkes and Folke, 1998). Traditional conservation practices parallel adaptive management in their reliance on learning-by-doing and the use of feedback from the environment to provide corrections for management practice.

The parallels between adaptive management and indigenous management systems are probably not accidental. Flexible social systems that proceed along learning-by-doing are better adapted for long-term survival than are rigid social systems that have set prescriptions for resource use. In light of this, adaptive management in modern society could be seen as a replication of traditional conservation in the South Pacific region. (Berkes and Folke, 1998: 10)

Bibliography


TRADITIONAL CONSERVATION PRACTICES


I. Food Chains and Trophic Levels
II. The Utility of Trophic Levels
III. Food Webs and Trophic Levels

GLOSSARY

basal species A species that eats no other species.
ecological transfer efficiency The ratio of energy ingested by a population’s predators to energy ingested by the population.
food chain A sequential relationship of the form $x_1$ is eaten by $x_2$ which is eaten by $x_3$ which is eaten by . . . which is eaten by $x_n$.
food web A specification of which species eat which in an ecosystem.
trophic level of a species $1/\sum L$ a weighted average of the lengths of all food chains linking that species to basal species. Different weightings may be appropriate for addressing different questions.

TROPHIC LEVELS

Peter Yodzis
University of Guelph

I. FOOD CHAINS AND TROPHIC LEVELS

Trophic ecology has to do with feeding relations—for instance, weasels eat mice, mice eat herbs—and is among the most basic organizing principles underlying biodiversity in natural ecosystems. One of the earliest attempts to identify ecosystem structure was based on trophic ecology, as follows. Some species in nature (mostly plants) do not eat anything; instead they utilize solar energy through photosynthesis and are called primary producers or autotrophs. Other species eat autotrophs and are called herbivores (Fig. 1). Carnivores eat herbivores, secondary carnivores eat carnivores, tertiary carnivores eat secondary carnivores, and so on (Fig. 1). These are, of course, highly aggregated entities: each of the categories just mentioned contains many species in any given ecosystem.

A food chain is a sequential relationship of the form $x_1$ is eaten by $x_2$ which is eaten by $x_3$ which is eaten by . . . which is eaten by $x_n$, where the entities $x_i$ might be individual species, or they might be aggregations such as those just defined.

Early trophic ecology (approximately the 1930s...
The IBP conception of an ecosystem as a linear sequence of trophic levels.

The International Biological Program (IBP) of 1964–1974 was very much organized around this concept of ecosystem structure.

The concepts of food chain and trophic level are very closely related. Thus, in the formulation sketched in Fig. 1, species in the $n$th trophic level are linked to primary production (autotrophs) through a food chain of length $n - 1$.

There is another food chain that is very important in a great many ecosystems, namely, the one that is based on detritus. Detritus is certainly eaten by a wide variety of organisms, but it is not itself a living organism, so it cannot exactly be said to "eat" anything. However, biomass does move from living organisms into detritus as they decay after dying. If one is concerned to follow the recycling of specific nutrients, one needs to keep track of this entire dynamic. For purely trophic studies it is generally adequate (and far simpler) to treat detritus as though it were another organism that, like an autotroph, does not eat anything. Then autotrophs and detritus together are called basal species, and the detritus-vores are lumped together with the herbivores in trophic level number 1.

II. THE UTILITY OF TROPHIC LEVELS

The trophic level concept has been exceptionally durable: it has been one of the basic concepts of ecology for six decades and is one of the few ecological concepts contained in the vocabulary of most educated people. The reason for this distinguished place in the scheme of things is that the concept is both simple and useful. Furthermore, it is universal: it applies to all ecosystems.

Because of this universality, trophic levels enable us to compare the role of vastly different species in vastly different systems. For instance, we can discuss and understand a lake and the surrounding forest with a common language: the forest has its vegetation and its leaf litter; the lake has its phytoplankton and its dissolved organic matter (basal species). The forest has herbivorous insects, birds, and mammals; the lake has zooplankton (herbivores). And so on. We can use the same language to compare these two systems with any other ecosystem anywhere in the world.

This categorical and conceptual role can be made more quantitative and detailed, revealing important similarities and important differences among systems, by adopting a bioenergetic viewpoint, as follows.

Biological organisms contain caloric energy, which is transferred to organisms in the next step up a food chain: herbivores gain energy from consuming basal species, carnivores gain energy by consuming herbivores, and so forth. Each organism, or set of organisms such as a trophic level, produces energy at a certain rate. This is the maximum rate at which the next trophic level up the food chain could in principle ingest energy. The rate of energy production by a trophic level must necessarily be less than the rate of energy ingestion by that trophic level. First, not all energy ingested by an organism is available to be metabolized by that organism. Some of it will be lost to excretion. The ratio of metabolizable energy to ingested energy is called assimilation efficiency and is typically about 0.45 for herbivores and 0.85 for carnivores. Of the metabolizable energy, some is lost to respiration, being used up by the organism to carry out its various activities and also simply to live, and the remainder is available for the production of new tissue, which can in principle be consumed by the next trophic level. The ratio of energy production to metabolizable energy is called production efficiency and ranges from about 0.1 to 0.4 for invertebrate ecotones to about 0.01–0.03 for endotherms.
TROPHIC LEVELS

FIGURE 2 The pyramid of production in an ecosystem. The width of each layer is proportional to the rate of production of biomass in the corresponding trophic level. Because energy is dissipated in each transfer from one trophic level to the next, production decreases at an approximately geometric rate as trophic level increases.

Furthermore, not all of the energy produced at a trophic level will actually be ingested by the next higher trophic level; much of it will be missed and end up as detritus. There is no generally agreed term for this form of energy loss, nor is there a great deal of quantitative data for it.

The ecological transfer efficiency of a trophic level is the ratio of (energy ingested from that trophic level by the next highest trophic level) to (energy ingested by that trophic level). It is the product of the three efficiencies explicated in the preceding paragraph. Ecological transfer efficiency ranges from 0.001 or smaller (depending upon losses to detritus) up to a maximum of about 0.3.

It quickly becomes apparent that energy is dissipated quite rapidly as we ascend a food chain. Suppose, for example, that each ecological transfer efficiency in a food chain is 0.1, which is rather high. Then of the energy produced by basal species (primary production), one-tenth is produced by herbivores. One-tenth of that, or one-hundredth of primary production, is produced by carnivores. One-tenth of that, or one-thousandth of primary production, is produced by secondary carnivores, and so on up the food chain. Less and less energy is available to higher trophic levels as we move up a food chain. One can visualize this phenomenon as a ''pyramid of production'' for a food chain (Fig. 2).

This geometrically decreasing production as we move up a food chain, hence the rapidly decreasing available energy flow for the next higher trophic level, has been offered as one possible explanation for the apparently limited number of trophic levels in natural systems. Eventually, as we move up a food chain, the small fraction of primary production available to a putative next highest trophic level simply will not be enough to support a viable biological population.

The pyramid of production is an inescapable consequence of the dissipative processes, sketched above, that lead to ecological transfer efficiencies less than 1. There are a couple of similar ''pyramids'' that, while not universal in this way, are fairly typical of trophic levels. They follow from the circumstance that, for the most part, predators tend to be larger than their prey. For a predator to be larger (hence also faster and stronger) than its prey greatly facilitates the capture and consumption of prey.

Thus, as we move to higher trophic levels, we will, generally speaking, see larger animals. And yet, moving to higher trophic levels, these larger animals need to live on smaller energy production from the next trophic level down. As a result, there will usually be fewer animals at higher trophic levels. This ''pyramid of numbers'' is frequently, though not necessarily always, observed.

An obvious exception to the pyramid of numbers emerges if we treat parasites and parasitoids as ''predators''; they are almost always smaller than their ''prey.'' Even though parasitism is tremendously widespread in nature, these are not really trophic relationships, and so most trophic studies do not include parasites or parasitoids.

What about total biomass \( \text{[\text{number of animals}] \times \text{(weight of each animal)}} \) at each trophic level? The number of animals tends to decrease as trophic level increases, while the weight of each animal tends to increase. The result is equivocal. Particularly in aquatic systems, where very small organisms at low trophic levels have very rapid rates of biomass turnover and can be grazed to quite low levels, one frequently (but not always) sees "inverted pyramids" of biomass, with more biomass at higher trophic levels. But terrestrial systems typically (though by no means always) display pyramids of biomass, with less biomass at higher trophic levels.
There are exceptions to this scheme, but they prove the rule; that is, they make sense in terms of the ideas underlying the scheme. For instance, some of the very largest animals, such as elephants and big ungulates, are herbivores. These animals are so large that they could not possibly range far enough to live by eating, say, lions. The only way to get a high enough energy density to support such large animals is by feeding directly on plants.

Just as energy propagates upward through food chains, so may chemical substances contained in organisms. This becomes particularly interesting when toxic contaminants are present. If those toxic substances are absorbed and/or ingested by animals at some trophic level, then, depending upon the rate at which they are excreted, there may be residues in the tissue consumed by higher trophic levels. Under some circumstances, the concentration of toxins may increase as trophic level increases, which is called biomagnification.

III. FOODwebs AND TROPHic LEVELS

The conception of an ecosystem as a linear chain of "trophic levels" (Fig. 1) is a useful starting point, but if we examine trophic relations with a higher degree of taxonomic resolution—that is, not lumping so many biological species together as we did in motivating Fig. 1—we find quite a different trophic structure. A food web is a specification of which species eat which in an ecosystem. For instance, Fig. 3 is a food web for Wytham Wood, a forest near Oxford in England. An arrow from one kind of organism to another indicates that the organisms at the head of the arrow eat the organisms at the other end.

One can detect something like "trophic levels" here (partly because of the way I chose to draw the picture), but there is certainly not a simple flow of energy through a linear sequence of levels as in Fig. 1; this picture is more "webby." For instance, we might put weasels at "trophic level" 5, because they eat titmice, which eat spiders, which eat insects, which eat herbs. But weasels also eat voles and mice, which eat herbs; this would put weasels at "trophic level" 3.

We do not want to throw away the trophic level concept altogether—it is too useful for that—but in the light of more refined data such as Fig. 3, we need to refine our concept of trophic level. In fact, there are a number of different ways that we may define the term "trophic level," and it seems imprudent to insist that any one definition is "The Right" one. Rather, different trophic level concepts may be appropriate for different purposes.

The constant theme linking all trophic level concepts together is the idea that trophic level has to do with the lengths of food chains linking a species to basal species. We just have to bear in mind that a species will generally be linked to basals through several food chains, which might be of different lengths. For instance, there are 13 food chains that link weasels to basal species in the Wytham Wood food web of Fig. 3. Three of these have length 2, 9 of them have length 3, and 1 of them has length 4. The relative importance assigned to these 13 food chains distinguishes several different definitions of "trophic level." Five commonly used definitions are listed here; following discussion of these, a sixth, in a considerably different spirit, will be addressed:

1. 1 + the length of the shortest food chain linking a species to some basal species.
2. 1 + the length of the longest food chain linking a species to some basal species.
3. 1 + the mean length of food chains linking a species to some basal species.
4. 1 + the weighted mean length of food chains linking a species to some basal species, where the weighting reflects energy flow through each food chain.
5. A + \( \left( \Delta X_{\text{reference level}} - \Delta X_{\text{reference level}} \right) E \), where \( E \) is the average enrichment of a heavy isotope and \( \Delta X = \left( \frac{\text{isotope in organism's tissue}}{\text{isotope in diet}} \right) \), for instance, \( ^{13}\text{C}/^{12}\text{C}, ^{15}\text{N}/^{14}\text{N}, \text{or } ^{34}\text{S}/^{32}\text{S} \). The trophic level of the reference level organisms is \( A \); this might be basal species (\( A = 1 \)), or perhaps herbivores (\( A = 2 \)).

Definition 5 requires explanation. Certain heavy isotopes are enriched relative to the light isotope in biochemical reactions. As a result, the ratio of heavy to light isotope in an organism's tissue may bear a systematic relation to the ratio in that organism's diet. For instance, the ratio \( ^{15}\text{N}/^{14}\text{N} \) appears to be enriched by \( E = 3.4\% \) in a wide variety of organisms. Therefore, if we know the isotope ratio for organisms at some reference trophic level, we can use Definition 5 to calculate a trophic level from measurements of the isotope ratio in other organisms. The resulting numbers, which are a weighted average over all food chains from the reference level to the organisms in question, are probably fairly close to what we would get from Definition 4. This method requires far less effort than a direct calculation.
of Definition 4 would, for instance, there is no need to measure dietary proportions. It does require careful calibration of isotope ratios at the reference level, including allowance for possible differences among species at that level. In practice, use of the $^{13}\text{C}/^{12}\text{C}$ ratio has not been fruitful, in part because so little $^{13}\text{C}$ is enriched at each trophic transfer, but as of this writing, the $^{15}\text{N}/^{14}\text{N}$ ratio appears to hold promise as a tool for trophic studies.

Table I shows, for each nonbasal species in the Wytham Wood food web, the frequency distribution of food chains linking that species to basals and the consequent trophic level according to Definitions 1–3. (Definitions 4 and 5 require more data than are available for the Wytham Wood food web.) For instance, depending upon the relative importance attached to the 13 food chains linking weasels to basal species, we may put weasels at trophic level 3 (shortest chain; Definition 1), 5 (longest chain; Definition 2), or 3.8 (mean length; Definition 3).

Generally speaking, one would like to use something like Definition 4, even though it means replacing the notion of discrete trophic levels with a trophic continuum. However, calculating trophic level in this way requires a tremendous amount of data. Definition 5 may be a good surrogate, but it still requires data beyond the food web itself. One is tempted to regard Definition 3 as a reasonably good substitute, but if we are thinking energetically, then because of the dissipation of energy as we move up a food chain, the shorter chains may well be more important energetically than the longer ones, so the equal weighting of Definition 3 may be deceptive. Definition 2 is what one has in mind implicitly when one draws tidy pictures such as Fig. 3, but the existence of very long food chains can be deceptive. Animals that have a food chain to basals with 8, 9, or even 10 links exist, but they invariably also have chains no longer than 3 links, and these shorter chains are likely more important energetically. However, very long food chains may be particularly significant if one is concerned with biomagnification of toxin concentrations.

Another viewpoint, which ought to produce a sixth trophic level definition if it could be articulated pre-
Table I
Food Chains and Trophic Levels in the Wytham Wood Food Web of Fig. 3

<table>
<thead>
<tr>
<th>Species</th>
<th>Shortest chain</th>
<th>Longest chain</th>
<th>Mean length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insects</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Winter moth</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tarves</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Leaf feeders</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Earthworms</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fungus</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Voles, mice</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Spiders</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Titjme</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Cynips</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Philanthus, thys.</td>
<td>0</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Soil insects, mice</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Owls</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Wasps</td>
<td>0</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Shrews</td>
<td>0</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Molest</td>
<td>0</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

Cisely enough, emphasizes the top-down aspect of trophic relationships. This viewpoint, which has been put forward by S. Fretwell and L. Oksanen, counts a predator as one trophic level higher than its prey only if it significantly controls the biomass or dynamics of the prey species. Oksanen suggests that on this basis the distinction among carnivores, secondary carnivores, tertiary carnivores, and so on largely evaporates, leaving only three true trophic levels: basal species, herbivores, and carnivores—except in pelagic systems, where, due to the very small size of the primary producers and the consequent small size of zooplankton, planktivory and piscivory emerge as truly distinct trophic roles, permitting four trophic levels. Thus, the term "trophic level" needs to be used with caution, and if we are to speak quantitatively of trophic levels, we need to specify exactly which definition we are using and to choose a definition that sheds the most light on the particular issues of concern.

See Also the Following Articles
ECOTOXICOLOGY • ENERGY FLOW AND ECOSYSTEMS • FOOD WEBS • PREDATORS, ECOLOGICAL ROLE OF • SPECIES INTERACTIONS

Bibliography
I. Introduction

Though covering only 9% of the Earth's land, tropical forest ecosystems are famous for their exuberant green-ness, impressive structure, and striking diversity of fauna and flora—and how poorly studied they are. De-
spite their attraction to nineteenth-century explorers such as Charles Darwin, Alexander von Humboldt, and Alfred Russell Wallace and the development of permanent field stations in the twentieth century, we still know very little about the biodiversity in tropical forests and how species function and interact in these ecosystems. For example, the huge uncertainty about the total number of species on our planet is largely due to how little is known about the invertebrates living in tropical forest ecosystems. Even though the phrase "tropical rain forest" still conjures up vivid images and myths about this signature ecosystem, there is a great amount of variety and heterogeneity in the types and extent of tropical forest ecosystems. This essay first provides an overview of the principal physical factors influencing the distribution and variation in tropical forest ecosystems and then describes the major types and features of tropical forest ecosystems.

II. PHYSICAL FACTORS
A. Geography
Tropical forest ecosystems are usually differentiated into neotropical (New World Tropics) and paleotropical (Old World Tropics). In contrast to the Amazon's dominance of the Neotropics, the Paleotropics include two major regions—Africa and Asia. African tropical forest ecosystems occur primarily in the Congo basin and the higher lands that border the basin. Outliers occur in a narrow coastal band of West Africa and in the Eastern Arc Mountains of East Africa. A narrow strip of tropical forests occurs along the eastern coast of Madagascar as well as in patches on other islands of the Indian Ocean. Because of striking biogeographic differences (e.g., Wallace's Line), Asia is further differentiated into the Indo-Malayan and Australasian regions. Tropical biogeography is profoundly influenced by the degree of insularity of major landmasses. For example, the lengthy isolation and connections of Australia, New Guinea, and other islands of Australasia facilitated the striking evolution and diversification of marsupials (e.g., tree kangaroos) in the absence of Indo-Malayan carnivores. The tree family Dipterocarpaceae dominates tropical forest ecosystems in the Indo-Malayan region, but this family has far less importance in Australasian tropical forests. Similarly, the long geological isolation of South America led to a distinctive fauna (e.g., anteaters, sloths, and tapirs) but many of these species were wiped out during the Great Faunal Interchange via the Central American isthmus. In contrast, the continued isolation of oceanic islands (e.g., Galapagos and Madagascar) has enabled some of the unique fauna and flora to survive.

The latitudinal extent of tropical forest ecosystems is not defined by the astronomical limits of the tropics (Capricorn (23°30' S) and the tropic of Cancer (23°30' N)). The ecological limits are determined by winter cold (which may not mean freezing temperatures) and inadequate rainfall to support forest vegetation. Generally, tropical forest ecosystems extend farther poleward where moisture is adequate, e.g., in the moist lowlands of Australia, Madagascar, and Mexico. Where aridity precludes the presence of lowland forest (e.g., Argentina and India), tropical forest ecosystems may extend to higher latitudes on the flanks of cooler, moister mountains. Though approximately 85 countries contain tropical forest ecosystems covering 14,076,490 km², equivalent to 58% of North America, the top 10 countries comprise 73% of this total (see Table I). The Amazon basin, shared by eight countries, has by far the largest area of tropical forest ecosystems, with 5 of the top 10 countries by extent of forest. Vast tropical forest ecosystems also occur in the Congo basin and on the mega-islands of the tropical Far East (Borneo, New Guinea, Sumatra, and Sulawesi). Smaller, but significant tropical forest ecosystems occur in the Caribbean, Mesoamerica, trans-Andean South America, coastal Brazil, West Africa, East Africa, Madagascar, South Asia, Southeast Asia, southern China, the Philippines, and many tropical Pacific islands.

B. Physiography
Regional features such as mountains and water bodies play an important role, influencing climate, soils, and vegetation in addition to ecosystem functions. Because major mountain chains (above 2500 m) block the normal flow of air masses, they are the most prominent physiographic feature influencing tropical forest ecosystems. The best example is the Andes, running along the western coast of South America. On the eastern (windward) side, copious rains fall usually throughout the year, whereas on the leeward (i.e., rain shadow) side, coastal deserts dominate—especially in Peru and Chile with the oceanic upwelling of the cold Humboldt Current. Many other mountain ranges (e.g., New Guinea, western India, East Africa, and Mesoamerica) have similar windward–leeward effects on climate and vegetation. Though lesser ranges may not create such striking differences, they provide impressive altitudinal
TABLE I
Top 10 Countries with Tropical Forests

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Forests (km²)</th>
<th>Cover(%)</th>
<th>Annual loss (km²)</th>
<th>% loss/year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Brazil</td>
<td>4,427,200</td>
<td>52.4</td>
<td>25,544</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>Democratic Republic of the Congo</td>
<td>1,352,650</td>
<td>60.0</td>
<td>7,402</td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>Indonesia</td>
<td>931,340</td>
<td>50.3</td>
<td>10,844</td>
<td>1.0</td>
</tr>
<tr>
<td>4</td>
<td>Peru</td>
<td>782,060</td>
<td>61.2</td>
<td>2,168</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>Bolivia</td>
<td>686,380</td>
<td>63.3</td>
<td>5,814</td>
<td>1.2</td>
</tr>
<tr>
<td>6</td>
<td>Venezuela</td>
<td>602,330</td>
<td>68.3</td>
<td>5,074</td>
<td>1.1</td>
</tr>
<tr>
<td>7</td>
<td>India</td>
<td>540,140</td>
<td>51.6</td>
<td>2,622</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td>Colombia</td>
<td>535,540</td>
<td>31.6</td>
<td>2,708</td>
<td>0.5</td>
</tr>
<tr>
<td>9</td>
<td>Mexico</td>
<td>457,650</td>
<td>24.0</td>
<td>5,080</td>
<td>0.9</td>
</tr>
<tr>
<td>10</td>
<td>Angola</td>
<td>373,640</td>
<td>30.1</td>
<td>2,366</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Subtotal top 10: 10,671,830

Total 85 countries: 14,076,490


C. Climate

Tropical climates are characterized by predictable temperature patterns with modest seasonal variation and precipitation regimes that are much more unpredictable and highly variable from year to year. Daily temperature range (on a sunny day) is far greater than the seasonal variation between the warmest and coolest months. In mountainous terrain, the daily temperature variation may be as large as 20°C. Seasonal differences are usually determined by rainfall patterns, such that the drier season is called "summer" and the wet season is "winter." The rainy season(s) track the movement of the thermal equator, also known as the intertropical convergence zone (ITCZ), which is about 1 month behind the passage of the sun. The ITCZ brings a daily pattern of afternoon convectional rains, often via thunderstorms or the monsoon. When the ITCZ is seasonally away from the region, it is usually the dry season. For example, when the sun is in the Southern Hemisphere, the northeast tradewinds dominate the Northern Hemisphere subtropics, e.g., the Caribbean and Africa. In contrast to the persistent tradewinds, the Asian monsoon involves a reversal of winds, with a summer inflow of warm, moist air and a winter outflow of cool, dry air.
There are two major types of volcanic ash—andesitic prevalent than lava flows (Hawaii is an exception). Volcanic ash deposits and mudflows tend to be more nary, especially in mountainous regions of the Tropics. They are so typical of much of the Amazon basin.

Andes, thus juxtaposing annually renewed, fertile floodplain soils with the infertile terra firme soils of the Amazon basin were eroded from the geologically ancient Brazilian shield. The active floodplains of several major western Amazonian rivers receive sediments eroded from the geologically young Andes, thus juxtaposing annually renewed, fertile floodplain soils with the infertile terra firme soils that are so typical of much of the Amazon basin.

Much volcanic activity has characterized the Quaternary, especially in mountainous regions of the Tropics. Volcanic ash deposits and mudflows tend to be more prevalent than lava flows (Hawaii is an exception). There are two major types of volcanic ash—andesitic and rhyolitic; the former is rich in certain nutrients, whereas the latter is high in silica. Soils developed from andesitic ash are termed “andepts” and are famous as the best soils for coffee or tea (e.g., Assam, Colombia, Central America, Hawaii, Jamaica, Java, and Kenya). The natural fertility of andepts is the foundation of the highly productive paddy rice cultures of Southeast Asia.

In contrast, extensive soils derived from rhyolite in Honduras and Nicaragua have low agricultural potential and are dominated by open pine forests.

Tropical soils vary greatly in their principal chemical and physical properties, such as parent material, soil structure, drainage regime, and age. Variations in soil properties play an important role in the different types of tropical forest ecosystems; however, we do not know much about their controlling influence on the distribution of most tropical tree species. The more restrictive a site is, the more distinctive the vegetation may be. For example, the combination of nutrient-poor, white sand soils and a perched water table in southern Venezuela gives rise to a dwarf shrubly vegetation (“bana” or “caatinga”) in a tropical rain forest region.

III. MAJOR FOREST ECOSYSTEMS

Tropical forest ecosystems are extremely rich in species of both fauna and flora. The highest tree species richness numbers—approximately 300 species per hectare—have been documented from the western Amazon and northern Borneo. Lowland forests of the tropical Far East and Amazonia are far richer in species than is the Congo basin. Even more striking is the fact that if one excludes tree species from the total floristic richness, tropical forest ecosystems are still the richest ecosystems in the world, due to the species diversity of non-trees such as epiphytes and climbers. Given similar rainfall regimes, central Amazonian forests tend to have fewer epiphytes than Mesoamerican forests, whereas tree species richness may be three times greater in the former.

There are many classification systems for differentiating tropical forest ecosystems that use bioclimatic, structural, or floristic criteria. Each has its purposes and merits, but they are too detailed for general overview purposes such as this chapter. Rather, a simple artificial system is used that first recognizes three elevational zones—lowlands (<500 m), low mountains (500–2000 m), and high mountains (>2000 m)—and then differentiates three major moisture regimes—perhumid (wet), humid (intermediate), and subhumid (seasonally dry). Brief overviews are also included for some distinc-
tive tropical forest ecosystems such as mangroves, flooded forests, savannas, and páramo. The interested reader must keep in mind that these are arbitrary groupings along continuous gradients, analogous to differentiating the colors of the rainbow, and that it is impossible to assign regional altitudinal or rainfall limits for major types of tropical forest ecosystems.

A. Lowlands (<300 m)

Classic lowland tropical forest ecosystems extend onto the foothills and lower slopes, but it is difficult to define an upper ecological limit for the lowlands. Some authors prefer an elevation (e.g., 1300 m), whereas others use the upper limit of typical lowland crops (such as the ‘‘coffee line’’) or more conventional bioclimatic, structural, or floristic criteria. In order to include the very representative lowlands of the Amazon and Congo basins (with base elevation of —300 m) as well as extensive foothills, the arbitrary upper limit of 500 m for lowlands is used here.

1. Mangroves

Probably the best known tropical forest ecosystem, mangrove forest ecosystems are characteristic of most tropical coastal zones due to the ability of several unrelated tree species to grow in saline conditions. Mangroves are typical of estuaries, deltas, and low-energy coasts lacking strong wave action. Due to greater tidal fluctuation and lower base flow of rivers associated with longer and more severe dry season, mangrove forests extend substantially farther inland on the western side of landmasses than on eastern coasts. For example, mangroves extend 190 km inland along the Gambia river in West Africa. Neotropical and West African mangrove forests have far fewer tree species than the mangroves of the Indian Ocean and the tropical Far East. Most mangrove species have evolved effective ways of dealing not only with their salty environment (e.g., leaves that exude salt through specialized glands) but also with the anaerobic sediments in which they grow. Mangrove ecosystems are important nurseries for shellfish as well as fish and offer frontline buffering against tropical storms.

2. Flooded Forests

In the wet lowlands, flooded forests occur behind the mangroves (where salinity does not influence the floristic composition of the forests), along many major rivers such as the lower Amazon, and in low-lying areas with poor drainage. The frequency, depth, and length of flooding appear to be important physical factors determining the composition and dominance of flooded forests. Many flooded forests are pure stands of a single tree species, such as *Campanopsis brevipes* (New Guinea), *C. panamensis* (Panama and Costa Rica), *Mauritia flexuosa* (Amazon basin), *Manicaria succifera* (Neotropics), *Musa oleifera* (circum-Caribbean), *Nypa fruticans* (New Guinea), *Prioria capriflora* (Central America), *Pterocarpus officinalis* (Costa Rica), *Raphia taedigera* (Neotropics and West Africa), and *Shorea albiada* (Borneo). We do not have a good understanding of why any one of these species is able to dominate a particular ecosystem.

In contrast to flooded forests dominated by a single species, more heterogeneous flooded forests occur farther from the sea or adjoining the pure stands. The dominant tree species of pure stands also occur in the more mixed forests, usually with far less dominance. Swamp forests may have very large trees, but tree density tends to be lower in swamp forests than in adjoining terra firme forests. Most tree species are not restricted to swamp or flooded forests, but may occur on terra firme, where rainfall is adequate and especially where there is no effective dry season. Even with a mixture of tree species, these flooded forests are not nearly as species rich as nearby terra firme forests. Many tree species that occur in flooded forests have large seeds that float, as can be seen in the flotsam on tropical beaches. One of the most characteristic trees of tropical beaches, the coconut (*Cocos nucifera*, Palmae), has such a wide pantropical distribution that it is difficult to determine its geographic origin. Several tree genera with large, buoyant seeds have the same or a sibling species across the Atlantic Ocean, such as *Elaeis* (Palmae), *Pentaclethra* (Mimosaceae), *Raphia taedigera* (Palmae), and *Sacoglottis* (Humiriaceae).

3. Perhumid Forests

Lowland perhumid forests are well known as classic tropical rain forests. Curiously, these forest ecosystems are more typical of the foothills and lower slopes of front ranges than of the extensive Amazon or Congo lowlands. The absence of major mountain ranges in tropical Africa may be a factor in the comparative scarcity of perhumid forest ecosystems there, especially in the Congo basin. Coastal West Africa from Guinea to Liberia is one of the few regions on that continent with average annual rainfall exceeding 3000 mm. Other lowland perhumid forests (e.g., southeastern Mesoamerica, Colombian Chocó, Gulf Province of Papua New Guinea, north Borneo, and Chittagong in Bangladesh) may have annual rainfall of 6000–10,000 mm.

Lowland perhumid forests on adequately draining
soils have high tree diversity, consistently exceeding 100 species per hectare for trees 10 cm or more in diameter (dbh) in heterogeneous old-growth forests. The latter genus (with sticky, bird-dispersed seeds) is the only member of this neotropical plant family present in Africa. Lianas (woody climbers) are one of the most striking components of tropical forests because of their growth form, stem structure, and impressive size; they may be 35 cm in diameter and up to 100 m long as they drape over the crowns of canopy trees. Many plant families are represented by lianas, but they are especially common in the Apocynaceae, Bignoniaceae, Dilleniaceae, Fabaceae, and Marcgraviaeae. In contrast to robust lianas, slender herba- ceous climbers are typical of the Araceae, with hundreds of species in several genera.

Although some stranglers belong to the genus Clusia (Clusiaceae), most are figs (Ficus, Moraceae). A strangler begins life as an epiphyte, sending down aerial roots, eventually forming an interconnected network of roots that embrace the host tree. A strangler may survive as an independent tree for many years after the host tree dies and decomposes, leaving a "hollow cylinder" as its testimony. Interestingly, the strangling part of this relationship has not been confirmed; death may be associated with the inability of the host tree to tolerate shading by the strangler. Each fig species is pollinated by a unique species of tiny fig wasps, representing one of the best examples of coevolution. Figs are important keystone food resources for vertebrate fruit eaters in many tropical forests.

Lowland humid forests are the most widespread of tropical ecosystems, covering much of the Amazon and Congo basins, Mesoamerica, and the tropical Far East. Mean annual rainfall typically exceeds 1300 mm, but the upper limit is more related to the seasonality of rainfall. Whereas lowland humid forests have only a short or ineffective dry season, lowland humid forest ecosystems are characterized by a significant dry season. These latter forests are almost as equally impressive as the former (e.g., the tallest trees in Ghana reach 60 m); however, the number of deciduous trees is much greater in the seasonal forests. Epiphytic orchids and gesneriads are less abundant in the canopy of humid forests than in perhumid forests. Smaller, thin, supple woody vines are more common in humid forests, perhaps related to the greater proportion of deciduous species in the forest canopy. Large palms are common subcanopy compo-
ments in humid forest ecosystems. Because of fire tolerance and the adaptive strategy of building the trunk belowground, some of these palms (e.g., Attalea speciosa, Copernicia spp., and Schelea rostrata) are aggressive colonizers of pastures created from humid forests and maintained by fire.

Seasonal rains and drought not only determine the ecological functions and processes but also have strong influence on the populations of animals and plants. The synchronization of flowering with the dry season is truly impressive. Flowering during the dry season when tree crowns and lianas are deciduous facilitates visitation by pollinators such as large bees. The alternation of rainy and dry seasons and the use of seasonal changes to trigger physiological responses (i.e., flowering, insect emergence, nesting, and rearing of young) make these ecosystems unusually susceptible to climatic fluctuations such as ENSO events. The best example is the ENSO-caused rains during the dry season that occasionally limit flowering and cause famine among fruit-eating animals on Panama's Barro Colorado Island (BCI).

The pioneering studies of the spatial distribution of trees by Hubbell and associates on BCI indicate that very few tree species are hyperdispersed as suggested in the classic literature on tropical forests. Many species have the adults randomly distributed, while some species show clumping. For example, several caesalpinioideae legumes (Brachystegia laurentii, Cynometra alexandri, Gilbertiodendron dewevrei, and Julbernardia seretii) form single-species stands in the Congo basin. A long-term study on the 50-ha plot at BCI has elucidated evidence for a strong density dependence within species, suggesting an important role for interactions between plants and their predators and pathogens. The spatial distribution of plants has important consequences for pollination, breeding systems, herbivory, and successful regeneration.

One of the most famous tropical trees is the Brazil nut (Bertholletia excelsa, Lecythidaceae), which has a patchy distribution in Amazonian humid lowland forests. Because adult trees occur in clumps, there has been considerable conjecture about indigenous influence on the abundance of this valuable species. Detailed studies document that the clumping is natural, probably due to the scatter-hoarding behavior of forest agoutis (Dasyprocta leporina, Heteromyidae) that cache seeds for later use. Another neotropical tree species with an unusual distribution is the true mahogany (Swietenia macrophylla, Meliaceae), which is reasonably abundant in the subtropical humid forests of northern Mesoamerica and the southern Amazon basin but quite scarce in the intervening tropical forests. Because of its high-value timber, there have been numerous attempts to promote or increase its abundance in natural forests, but they have mostly failed. Recently, it has been suggested that the mahogany requires large-scale disturbance such as by fire or river meanders, creating the high light regime needed for establishment and growth of young trees.

One of the more fascinating historical aspects of this ecosystem is the development of truly impressive human civilizations such as the Maya in northern Mesoamerica and the Khmer in Kampuchea (Cambodia). Both cultures developed intensive agriculture based on sophisticated engineering and management of water during the dry season through the use of raised fields, irrigation canals, and paddies. The Maya are believed to have also favored particularly useful plants such as mahogany, Manilkara zapota ( Sapotaceae), and Brosimum alniatum (Moraceae). About A.D. 900 the classic Maya civilization collapsed, perhaps due to overpopulation and/or a prolonged, severe drought. Given our current understanding of tropical forest dynamics, it is quite unlikely that the humid lowland forests now occupying the classic Maya region are successional forests.

5. Subhumid Forests

Lowland subhumid tropical forest ecosystems are much more extensive in the Paleotropics (e.g., Madagascar and India) than in the Neotropics. The high-value teak (Tectona grandis, Verbenaceae) forests of Burma, Thailand, and Java are representative of this ecosystem, as are the vast sal forests (Shorea robusta, Dipterocarpaceae) of northern South Asia. These ecosystems also occupy significant areas in the Caribbean, along the Pacific coast of Mesoamerica and northeast Brazil. Mean annual rainfall seldom exceeds 1300 mm, but the determining ecological factor is the long and severe dry season. Because of the historic pastoralist tradition in this ecosystem and the prevalence of nearly annual fires, most lowland subhumid forests have been destroyed or severely degraded by centuries of use. One of the most notable exceptions is the well-studied dry forest of Guaicia on the south coast of Puerto Rico. This island is unusual in that the economic pressures to use land have not been nearly as intense this century compared to most developing countries.

Lowland subhumid forest ecosystems have far lower tree species richness than humid and perhumid forests. They are also much shorter in height and simpler in structure, but virtually all woody species are deciduous. In low-stature forests, the shrub layer is often dense and very spiny; thin lianas are abundant (including the Vanilla orchid), but epiphytes are less rich in species.
and abundance. Terrestrial cacti and bromelids are more common, especially on extreme sites such as rock outcrops. At the latitudinal limits of the Tropics, lowland subhumid forest ecosystems include chaparral, open pine forests, and mixed oak–pine forests—depending primarily on moisture availability. Evergreen forests occur in riparian situations and in other wet sites such as in karst terrain, where underground moisture is available during the long dry season.

In contrast to the pure stands in some freshwater swamps mentioned earlier, pure stands in lowland subhumid forest ecosystems usually can be attributed to edaphic factors, such as rhodolic ash (tuff or pumice) or montmorillonite clay (black cotton soils). One of the more distinctive forests is the pure stands of the only lowland oak species, Quercus oleoides, that has a patchy distribution on rhodolic ash deposits from southern Mexico to northwestern Costa Rica. Appreciable areas of lowland subhumid forest ecosystems in Mesoamerica exhibit single-species dominance, such as Ceratandra mexicana (Euphorbiaceae); Crescentia alata (Bignoniaceae); Erythrina fusca (Fabaceae), and Parkinsonia aculeata (Caesalpiniaceae). Despite the attribution of these single-dominant stands to restrictive soil factors, there is precious little clarity as to why a species is able to attain such dominance.

6. Savannas

Savannas are defined as having a continuous cover of grass, but trees may be conspicuous components of savanna landscapes. Though vast areas in the major tropical regions meet the criteria for savannas, it is much more difficult to ascertain if it is natural or derived through human activities (e.g., burning). Tropical Africa epitomizes the image and grandeur of savannas with vast expanses of grass, sparse trees, migratory herds of herbivores, and their predators. Recently, the role of elephant herds in transforming woodland to savanna has been well documented. Extensive savannas also occur in the tropical Far East and the Neotropics (the Ilosos of Colombia and Venezuela, Rupununi in Guyana, and Beni in Bolivia).

The extensive pastures of Mesoamerica are considered to be derived savannas created by almost annual burning to “freshen” pastures and suppress woody invaders. Neotropical savannas have some very characteristic, fire-resistant tree species, such as Acrocomia vinifera (Palmae), Byrsonima crassifolia (Malpighiaceae), and Carattella americana (Dilleniaceae). Lowland pine (Pinus caribaea var. hondurensis, Pinaceae) savannas occur in northeastern Nicaragua, much of lowland Honduras, part of the Guatemalan Petén, and east central Belize. In contrast to the derived savannas of Mesoamerica, the pine savannas appear to be natural ecosystems maintained by frequent fire. The flat pine savannas of Belize occur on sandy soils that flood during the rainy season and undergo severe drought in the dry season.

B. Low Mountains

For purposes of this article, low-mountain ecosystems occur between 500- and 2000-m elevation. The rather arbitrary upper limit represents the maximum extent of tropical lowland flora near the equator. In many regions, the transition to montane flora more typically occurs by 1300 m, while nearer the latitudinal limits of the Tropics, this transition can be as low as 500 m. This elevational limit for tropical plants is more noticeable with crops such as highland coffee (Coffea arabica, Rubiaceae) or tea (Camellia sinensis, Theaceae). Low-mountain ecosystems typically have a mixture of tropical and temperate fauna and flora; for example, pines and oaks are conspicuous trees in northern Mesoamerican forests. Some tree genera have pairs of closely related species, with one occurring at lower elevations and the other in higher elevations in Mesoamerica, e.g., Billia (Hippocastanaceae), Bruneilia (Brunelliaceae), Hedysosmum (Chloranthaceae), and Smplocos (Smplocaceae). In some ways, low-mountain ecosystems between 500 and 2000 m are transitional between lowland tropical and tropical montane floras. These ecosystems are typical of the major mountain ranges as well as the many smaller chains and high plateaus (e.g., Brazilian Shield, Serengti Plain, and Great African Plateau).

1. Subhumid Forests

Low-mountain subhumid forest ecosystems are common on leeward slopes and in valleys and plateaus protected from orographic rain. Because of favorable climate, this ecosystem has had a long history of human activities, leaving little natural vegetation. Subhumid forests are relatively open and low in stature, with an abundance of spiny trees and shrubs, especially Mimosaceae. In northern Mesoamerica and continental Asia, conifers (Pinus spp.) are common in subhumid forest ecosystems. P. kesiya and P. merkusii occur in the Malay archipelago, while P. wallichiana and P. rosburgiana are common on the lower flanks of the Himalayas.

This ecosystem is extensive on the plateaus that border both the Amazon and Congo basins. The Great African Plateau is dominated by miombo woodland (dominated by Brachystegia, Julbernardia, and Isoberyllia, Caesalpiniaceae), while the Brazilian shield is
characterized by similar open forest ("cerrado"). The Brazilian cerrado are low, open forests or woodlands with characteristically crooked trees that occur on soils high in aluminum and low in pH. Cerrado vegetation exhibits considerable variation, ranging from grass-dominated savanna to fairly dense forest; however, most of the plant species are of tropical origin. Where soils are less restrictive, more typical tropical forest exists that is surprisingly tall (to 40 m), fairly open canopy, with a high proportion of deciduous species.

2. Humid Forests

Humid forest ecosystems often ring lower, more arid regions on low mountains, because of more moisture, cooler temperatures, and/or less severe drought. Low-mountain humid forests show enormous variation of forest type, structure, and floristic composition due to their widespread distribution and great heterogeneity of physiography, soils, and rainfall patterns. An excellent example of this variation occurs in northern Mesoamerica, where pine forests dominate the upper slopes and broad-leaved forests occur in the valleys. Pines are characteristic of low-fertility soils and frequently burned sites, with *P. occarpa* often dominant. The broad-leaved forests usually have a mixture of tropical (e.g., *Alchornea*, Euphorbiaceae) and temperate tree genera. *Parinari excelsa* (Chrysobalanaceae) has an unusually broad distribution in Africa, dominating humid forests in both the Congo lowlands and the Guinea highlands. Impressive tropical forests occur on fertile soils (i.e., andepts) in Mesoamerica with a high density of canopy trees (typically in the Lauraceae and Sapotaceae) plus occasional tall (>50 m), emergent trees of *Ulmus mexicana* (Ulmaceae). Also occurring in these mid-elevation humid forests are canopy trees of Podocarpaceae (Southern Hemisphere conifers), Juglandaceae, and Proteaceae. The latter tree family (that includes Macademia) is especially diverse and abundant in the Queensland Wet Tropics of northeastern Australia.

3. Perhumid Forests

These high-rainfall ecosystems are common on the eastern flanks of low mountains and ranges that receive moisture-laden clouds from the east. When the mountains are oriented more or less perpendicular to the prevailing air currents, they nurture the famous cloud forests of the Caribbean, Mesoamerica, northern and eastern Andes ("ceja de selva" and "yungas"), and even on isolated volcanic masses. Perhumid forests on low mountains may have tree species richness comparable to lowland perhumid forests. They also exhibit complex vertical structure and fairly tall trees; the latter may be related to the better drainage and anchorage typical of dissected topography. Canopy trees are usually evergreen (with abundant epiphylls) and with heavy epiphyte loads, including bryophytes, shrubs, and even small trees. The abundant mosses in the canopy give these perhumid forests a distinctive brownish hue visible from small planes. Epiphytic Ericaceae shrubs are usually very abundant in cloud forests, providing year-round nutrition to nectar-feeding birds. The sizeable and heavy (when wet) loads of organic material in tree crowns cause frequent branch falls as well as stimulating the host tree to produce feeding roots on the branches to take up nutrients. In perhumid forests on nutrient-poor soils, numerous species of insect-trapping, epiphytic pitcher plants (Utricularia, Lentibulariaceae) are quite abundant.

C. High Mountains (>2000 m)

Because several tropical lowland species and crops may grow on protected sites up to 2000-m elevation, we use this as the lower limit of high-mountain forest ecosystems. High mountains in the Tropics have a temperate climate characterized by warm diurnal temperatures (on a sunny day) and cold nocturnal temperatures, including occasional freezes. In Africa, the fairly isolated mountains are defined as the Afromontane zone. Treeline varies considerably between about 3300 and 3800 m, depending on moisture, disturbance by fire, and proximity to the equator. Despite the temperate climate, most forest ecosystems have distinctive tropical features such as epiphytes and lamas. Nevertheless, the high-mountain tropical flora has much stronger affinities with temperate zone flora.

1. Subhumid Forests

In the Afromontane region of East Africa, *Juniperus procera* (Cupressaceae) forms pure stands; however, it is unclear if human disturbance is a contributing factor to this species' success. It is possible that high-mountain subhumid forests once existed on the Andean Altiplano, but centuries of human activities have destroyed most natural vegetation. The only remnants may be the small stands of *Polylepis* (Rosaceae) that occur up to 4000-m elevation in the central Andes—the highest forests in the world. The forest patches persist on rocky slopes that may provide natural protection from frequent fires and soil temperatures may be slightly higher. *Polylepis* trees are small, twisted, and gnarled, with considerable trunk and branch dieback, especially on stand edges.
TROPICAL FOREST ECOSYSTEMS

2. Humid and Perhumid Forests

The high elevations and cool temperatures tend to produce a preponderance of humid and perhumid forest ecosystems because of frequent cloud cover and low evapotranspiration. These ecosystems usually have the canopy dominated by holarctic genera (e.g., Pinus, Quercus, and Lauraceae), whereas the understory is more of a mixture of holarctic and tropical genera. Northern Mesoamerican forests are the center of diversity for oaks (Quercus, Fagaceae) and conifers (Abies, Cupressus, Juniperus, Pinus, Podocarpus, Taxodium, and Taxus). Except for Podocarpus, these are all holarctic genera. On poor soils, the coniferous forests tend to be open and frequently degraded by grazing and fire. More fertile and/or wetter sites also have conifers, but the forests are more of a mixture with oaks. Northern Hemisphere conifers do not occur naturally south of the Nicaraguan lowlands; thus oaks dominate the high-mountain forests of southern Mesoamerica and Colombia. The bamboos Chusquea in the Neotropics and Arundinaria alpina (both Poaceae) in East Africa are typically very abundant in the understory, where they may be impeding natural regeneration of trees. As with lower elevation perhumid forests, the large canopy trees carry huge epiphyte loads of mosses, tank bromeliads, and ericaceous shrubs.

3. Paramo

Although not strictly a forest ecosystem, high-mountain paramo merits brief mention as a unique high-mountain tropical ecosystem. Subalpine paramo occurs above treeline on high mountains in the Neotropics (Ecuador to Costa Rica) and East African volcanic massifs. Neotropical paramo is characterized by a high density of shrubs, including Chusquea, Hypericum (Hypericaceae) and Vaccinium (Ericaceae) and occasional emergent trees of Clusia (Clusiaceae). Where drainage is poor, bogs and a distinctive flora are frequent. But the most striking feature in Andean paramo is the abundant, postlike Espeletia (Compositae) that are very similar in growth form to Espeletia. The thick trunk is fire-resistant and the rosette of leaves may offer some protection from nightly freezing temperatures.

Bibliography

TRUE BUGS AND THEIR RELATIVES, DIVERSITY OF

Carl W. Schaefer
University of Connecticut

I. Suborders of Hemiptera
II. Heteroptera: General
III. Heteroptera: Specific
IV. Conclusions

GLOSSARY

beak (or rostrum) The thin elongate mouthparts, adapted for piercing the organisms upon which the bugs feed (the mouthparts of bugs are similar in function to, but quite different in structure from, those of mosquitoes).

food pump The sclerotized pump used by hemipterans to suck up the juices (plant or animal) upon which they feed.

hemelytron The heteropteran’s forewing; so called because the anterior part of the wing is opaque and the posterior part is clear (membranous).

salivary pump The sclerotized pump used by hemipterans to inject saliva into the organisms upon which they feed.

scent glands Glands which in heteropterans secrete various compounds for various purposes, including defense, attraction of mates, and attraction of both sexes. Adult heteropterans have scent glands on their metathoraces, and nymphal heteropterans as well as some adults have scent glands on the dorsa of their abdomens.

scutellum The dorsal portion of the mesothorax, in most heteropterans triangular or shield-shaped (hence “little shield”).

HETEROPTERA IS ONE of the most diverse groups of insects. Heteroptera (“true bugs”) range in length from a single millimeter to more than 8 cm and in width from a millimeter to several centimeters. They occur in all colors and combinations of colors, including jewel-like combinations of gold or green on an iridescent purple background. They live in habitats as varied as spider webs, anthills, termite mounds, grain, the bodies of bats, under the bark of trees (feeding on fungi), on the bark of trees (feeding on other arthropods), high in the canopies of tropical trees, deep in the soil, on the surface of water (without getting wet), and in the homes and beds of people. Some use tools (to capture termites), the males of some carry the eggs about to keep them safe, others squirt foul liquids (to repel enemies), yet others eat foul liquids (milkweed sap) and store the bitterness. Heteropterans feed on everything from fungi and crop plants to other insects and vertebrates, including people. One group (triatomine bugs) is the vector of one of the most serious diseases in Latin America (Chagas’ disease). Heteroptera is a suborder in the insect order Hemiptera. Like other hemipterans, heteropterans have narrow elongated mouthparts that pierce plants or animals and suck up the sap or the body fluids. The insects have incomplete metamorphosis (i.e., there is no pupal stage.
and the immatures look very much like the adults). The order is an old one: Hemipterans occurred in the Lower Permian, and heteropterans assignable to an extinct infraorder have been found from the Upper Permian.

I. SUBORDERS OF HEMIPTERA

The number of actual suborders has recently come into question, because of the uncertain validity of Homoptera. However, three suborders are frequently accepted in most entomology texts: Homoptera, Heteroptera, and Coleorrhyncha. Nevertheless, many homopterists recognize two orders, Homoptera and Hemiptera (= Heteroptera). The reasons for this confusion are themselves confusing and have to do with the importance that heteropterists and homopterists give to the groups they study: this importance in turn is based on the extent to which the students of one group understand the work of the other group's students. Those who have undertaken the most thorough study of all hemipteran groups conclude that they share too many characteristics to be treated as separate orders.

A. "Homoptera"

This group includes the cicadas, leafhoppers, aphids, scale insects, whiteflies, and their relatives. Nearly all feed on plants (a few on fungi), and many are important pests of crops and ornamentals. Recent work on the origins and (especially) the evolutionary relationships of groups within "Homoptera" has led to the conclusion that the group did not have a unique ancestor common to all homopterans and to no other hemipterans. As a result, the higher classification of "Homoptera," and the systematic status of the group itself, are deeply in question. That is why I place the term "Homoptera" in quotation marks: The term is useful in grouping together a number of insects, but it is not useful in suggesting they compose a single evolutionary unit.

B. Coleorrhyncha

Members of the single family (Peloridiidae) have a classic Gondwana distribution: Australia, New Zealand, southern South America. They are associated with moss in Nothofagus (southern beech) forests and probably feed on the moss. Small, somewhat flattened, often with lovely semitransparent expansions of the head and thorax, these insects are poorly known and deserve much more study. Some believe Coleorrhyncha (formerly included in Homoptera) to be the sister group of Heteroptera—that is, that these two groups had a common ancestor. The evidence for this conclusion is stronger than that for the earlier inclusion of coleorrhyncha in Homoptera, where they were thought to be a sort of "nonmissing link" between Heteroptera and Homoptera.

II. HETEROPTERA: GENERAL

A. Why "True" Bugs?

Heteropterans are the only group of insects that entomologists agree may legitimately be called "bugs"; hence the name entomologists themselves use, "true bugs." Nonentomologists call all insects (and other arthropods) "bugs," and, although entomologists disapprove, they are resigned to this use.

Why are heteropterans called "bugs"? The reason almost certainly lies in the original Middle English meaning of the word: "bug" (as bugge) meant originally a wraith or specter, something unseen that emerges from the shadows and does harm. This is exactly what the best known heteropteran does: the bedbug (Cimex lectularius L.) sucks blood from the sleeping victim and disappears; in the morning only the irritating welt from the bite remains, its cause and origin unknown. Clearly, an unseen nocturnal shadow has caused this damage: a bug. (This original meaning of bugge is retained in Modern English in such words as bugbear, bugaboo, and bogeyman; one may also wonder if it is retained in the verb "to bug," meaning (1) to irritate and (2) to use a small and secret device for spying—a form of damage.)

B. Characteristics

Heteropterans share many structural features with other insects, of course; other features with other insects with
TRUE BUGS AND THEIR RELATIVES, DIVERSITY OF

incomplete metamorphosis; and of course some characteristics with other hemipterans. Still more aspects of their morphology are constrained by the fact one body part must move and be in contact with an unmoving part.

Like all insects, heteropterans have three basic body regions: head, thorax, abdomen; and, like those of nearly all insects, these regions are specialized: the head for grasping and preparing food and for exploring the environment, the thorax (with legs and wings) for locomotion, and the abdomen for processing food and for reproduction. Like most insects, heteropterans have three pairs of five-jointed legs and two pairs of wings; a pair of compound eyes (usually large, because true bugs are mostly diurnal and therefore sight is an important sense), a pair of antennae, and an aedeagus (or penis) for the introduction of sperm into the female, who in turn has an ovipositor for laying eggs.

Like other insects with incomplete metamorphosis, immature bugs look like their adults, except for the lack of functional wings and genitalia and for different body proportions. As the immatures develop, wings appear as small pads, which enlarge at each molt; later, the genitalia too may be discerned, and often the last immature stage can be sexed.

Heteropterans share more specific features with other hemipterans: The mouthparts are modified into a long sharp tube, with two internal openings. Through one of these, salivary fluid is pumped into the plant or animal (plant only, in the case of nonheteropteran hemipterans), and through the other opening the fluids are sucked up. A set of pumps within the head accomplishes these activities. This mechanism and method of feeding are perhaps the best evidence that all hemipterans are closely related, and this in turn is why most entomologists consider all hemipterans to compose a single order.

Heteropterans, like other hemipterans, have certain characteristic features of the digestive system and the excretory system (although heteropterans lack the “filter chamber” independently evolved in several homopteran groups). The primitively 12–13 ganglia of the ventral nervous system are fused to a single one (or to four in a few homopteran groups). The wing venation is reduced (less so in many homopterans); the chromosomes are holocentric (diffuse centromere). Many hemipterans can produce sounds, usually by scraping a “pick” over a corrugated surface. Often these structures appear the same and occur on the same parts of the body. But in most cases, these sound-producing structures have evolved independently; the appearance is constrained by their function, and their occurrence on the same body parts is constrained by the fact one body part must move and be in contact with an unmoving part.

Heteropterans themselves are more diverse in habit and habitat than the other hemipteran groups. Heteroptera contains the only hemipterans that live in, and sometimes on, the water; the only hemipterans that are partly marine; the only carnivorous hemipterans; and the only hemipterans that feed on vertebrate (including human) blood. It is not surprising, then, that within the suborder Heteroptera is a wide variety of sizes and shapes, nor is it surprising that structures of heteropterans (legs, etc.) are highly modified in some groups.

The beak, or rostrum, characteristic of Hemiptera in general, in some plant-feeding heteropterans may be longer than the body, and in some predacious heteropterans be so short as not to reach the thorax. It may be slender or stout, and the number of segments may be reduced from the original five. The beak arises closer to the back of the head in heteropterans than in homopterans, and thus a region of the head lies just in front of the rostrum’s base; the presence of this region in Coleorrhyncha is one reason why Coleorrhyncha and Heteroptera are thought to have had a common ancestor.

Highly characteristic of Heteroptera alone is the division of the forewing into two regions, the posterior one, clear and membranous (like the hindwing), and the anterior portion, harder, often leathery, and opaque: hence the name “heteroptera” (“different wings”). The two wings on each side are coupled together with a device on the forewing; this is found only in Heteroptera and Coleorrhyncha. Heteropteran sperm, too, has some unique features; Coleorrhyncha sperm has not been examined.

Characteristic of nearly all heteropterans is a triangular (pointing backward) part of the top of the thorax, the scutellum (“little shield”); in a few heteropterans, especially in the group Pentatomoida, the scutellum covers all or most of the dorsal surface of the insect, except the head (but including the wings).

Unique to Heteroptera are scent glands. These are found on the abdomen of immature heteropterans (from one to four pairs on either side of the midline). These glands are usually nonfunctional in adults, although in many species they do produce scent; the known number of such species is increasing with increasing study, and the function of the scent from these dorsal abdominal glands is becoming more clear. Adult heteropterans have a pair of glands (regardless of their possession of functional dorsal glands) on the last (third) thoracic segment. These lateral metathoracic scent glands are thought to repel predators (and have...
been shown to repel ants), but they certainly have other functions as well, including the attraction of conspecifics of the opposite (and sometimes of both) sex(es). Heteroptera do not differ biologically from other insects and other hemipterans. They do the three basic things that all organisms do—feed, reproduce, and disperse—and, because these things are basic to survival as a species, bugs do them much as do other winged insects. However, bugs are remarkably more diverse than other hemipterans (and indeed than other insects with incomplete metamorphosis) in their choices of food and of habitats and in their interactions with their environments (for the latter, see below).

Heteroptera feed like other hemipterans. Their sharp elongated mouthparts are adapted for piercing; one internal pump pushes saliva into the food source, and another sucks back the food. In most cases, this food has been all or partly digested by salivary enzymes. Few bugs feed on the watery contents of plants’ xylem, and so bugs lack the filter chamber that in many homopterans removes and excretes excess water (and sugars). (Diuresis does occur in blood-sucking bugs, however: see below.)

The salivary enzymes and substances of plant-feeding bugs serve other functions than digestion alone. Many phytophagous bugs form a small feeding cone on the surface of the plant where the mouthparts are inserted. This cone, made from salivary secretions, helps stabilize the mouthparts as they penetrate to the source of food. Other salivary products help break down cell walls and release cells’ contents; other products mimic plant auxins and ‘‘fool’’ the plant into mobilizing nitrogen-rich materials at the point of feeding.

As a general rule, with many exceptions, plant-feeding heteropterans feed on the highly nutritional reproductive parts of plants—flowers, developing fruits and seeds, ripened fruits and seeds, and fallen fruits and seeds. This may be a matter of size: Many plant-feeding bugs are relatively large (i.e., larger than the majority of homopterans); larger hemipterans are better able to penetrate plant parts to the seeds within and to disperse to find flowers, fruits, and seeds—which are transitory resources. Smaller heteropterans, like many homopterans, feed on leaves, a resource more permanent, easier to find flowers, fruits, and seeds. This may be a matter of size: Many plant-feeding bugs serve other functions than digestion alone. Many phytophagous bugs form a small feeding cone on the surface of the plant where the mouthparts are inserted. This cone, made from salivary secretions, helps stabilize the mouthparts as they penetrate to the source of food. Other salivary products help break down cell walls and release cells’ contents; other products mimic plant auxins and ‘‘fool’’ the plant into mobilizing nitrogen-rich materials at the point of feeding.

As a general rule, with many exceptions, plant-feeding heteropterans feed on the highly nutritional reproductive parts of plants—flowers, developing fruits and seeds, ripened fruits and seeds, and fallen fruits and seeds. This may be a matter of size: Many plant-feeding bugs are relatively large (i.e., larger than the majority of homopterans); larger hemipterans are better able to penetrate plant parts to the seeds within and to disperse to find flowers, fruits, and seeds—which are transitory resources. Smaller heteropterans, like many homopterans, feed on leaves, a resource more permanent, easier to penetrate, but of less nutritional value.

The flat bugs (Aradomorpha—see below) have very long, very slender mouthparts with which they feed on the linear array of fungal mycelia below the bark of dead or dying trees. A few of these bugs live in termite nests, where they probably feed also on fungi. With very few exceptions, these are the only bugs that feed on fungi.

Indeed, the great majority of phytophagous heteropterans feed on flowering plants, angiosperms. Several small groups feed on gymnosperms, and a very few bugs feed on mosses or ferns. It seems likely that the great radiation of heteropterans—like the radiations of so many other insect groups—occurred at the same time as, and perhaps in response to, the radiation of flowering plants.

Within the angiosperms, some groups are less preferred than others: grasses, for example, and legumes. Each of these groups has developed defenses against being eaten, and in each case one or more groups of bugs has overcome the defense and radiated on the plant group. It is significant that these two groups (grasses and legumes) have few individual species feeding on them but, rather, a few groups of species (subfamilies, families). This suggests that an early ancestor broached the defense (of grass or legume) and its descendants radiated on this newly available food source.

Many bugs are predacious. Indeed, it has been argued that this was the original way of life of Heteroptera (an argument that some dispute). The prey is other insects and often their eggs or other immature stages. Some of the smallest heteropterans are predacious, on arthropod eggs, mites, and scale insects and aphids. It was long thought that predacious bugs inserted their mouthparts and sucked up fluids inside the prey, much as plant feeders suck sap. Recently, it has been demonstrated how much more complex, and interesting, the situation really is (Cohen, 1998). A predacious bug actively pumps digestive enzymes and fluid into the prey. The enzymes digest the soft internal organs of the prey, and the fluid and digested food are sucked up. This ‘‘solid-to-liquid feeding’’ process is repeated over and over, the mouthparts within the prey moving throughout its body, probing even into the legs and antennae to digest the muscles there. There remains but a husk, a shell, the exoskeleton of the prey, all internal organs having been digested, liquefied, and sucked into the predator. As a consequence, a predacious bug need feed on only a few prey, because all of the prey except its exoskeleton is eaten. The implications of this for biological control using heteropterans are considerable (see below).

The predacious habit has arisen often in the Heteroptera, secondarily, probably more often than phytophagy. This suggests that the search for more concentrated and usable nitrogen is as important in this group as it is in others.

A few bugs feed on vertebrate blood. Heteropterans’ mouthparts, like those of mosquitoes, are long and sharp, and so adapted for sticking into liquid-filled organisms. It seems inevitable that some would take verte-
brate blood, although few feed on birds and none on cold-blooded vertebrates (with one exception). With few exceptions, the vertebrates fed upon are small mammals and bats—and man. Members of two closely related families (Polycentrinae and Cimicidae) feed on bats, and a few members of the latter family feed on humans: the bedbugs, some of whose relatives also feed on poultry. Members of the assassin bug subfamily Triatominae feed on small mammals, and some of them on humans; these spread the protozoon trypanosome which causes Chagas' disease.

The problem of feeding on vertebrate blood includes the problem of ingesting red blood cells through a narrow tube. The physics of this type of feeding have been studied, and maximal feeding is related to the inside diameter of the feeding tube, the concentration of red blood cells, and the duration of feeding. Such feeding also results in the accumulation of blood fluids. These are excreted (diuresis) in a manner roughly analogous to the excretion of plant juices by homopterans ("honeydew").

Two other heteropterans feed on vertebrates: a bird-nest-living lygaeid (the "seed bug" family) apparently feeds only on the blood of the host bird; another group of lygaeids apparently feeds on small mammals in Africa (and may there occupy the niche occupied in Latin America by Triatominae). Giant water bugs (Belostomatidae) have been reported to attack baby ducklings (this may be a legend), and some species specialize in taking the fluids of small fish.

No heteropterans feed exclusively on dead or decaying organic matter (i.e., none is a saprophage). However, there are many anecdotal accounts of various bugs found feeding on bird droppings, carrion, and dead insects. It is likely these bugs sought water or a source of concentrated soluble nitrogen.

The great majority of heteropterans reproducers like other insects: Sexes find one another, there may be a brief period of courtship, mating occurs (fertilization is internal), cigarettes are smoked and promises made, and the fertilized eggs are laid near or on the appropriate food for the young.

The sexes find one another probably via sex attractants given off by one or the other sex, or perhaps by both (depending on the species; however, very few species have been studied). Courthship, if it occurs at all, is brief. In a few species, males fight for females, the enlarged legs of certain males are used to knock one another off a leaf to which sex attractants are drawing females. However, there seems to be very little intrasexual rivalry in Heteroptera. Mating is often end-to-end, or in many species the male lies at an angle across the female; in all cases both sexes are ventrally side down. In some end-to-end species, the female, larger than the male, may be spotted moving around dragging her mate with her; or she may feed casually on a leaf, with the smaller male dangling helplessly in the air, attached only by the genitalia. Soon after mating occurs, eggs are laid singly, or in characteristic batches, sometimes in several layers, depending on the species and, often, on the family. A female lays several batches of eggs in her lifetime, each one usually requiring a separate mating. The eggs of many species have tiny openings (micropyles) through which the sperm enter; other small openings (aeropyles) provide gas exchange; and some species' eggs have a round or oval line of weakening, through which the insect will hatch.

Eggs may be round, oval(!), cylindrical, or barrel-shaped. Many are brightly colored. The egg of Rhopalidae ("scentless plantbugs") is mounted on an elongate pedicel (like eggs of some neuropterans), presumably to protect it from predators. Parental care of the eggs occurs in a few species. As is well known, the female of some giant water bugs (Belostomatidae) glues her eggs onto the back of the male, who carries them until hatch. By sweeping water across them, he keeps them aerated and prevents the attack of fungi. The females of some stinkbugs (Acanthosomatidae) and some lace bugs (Tingidae) protect their eggs and early immature stages, guarding them from attack from parasites and such predators as ants. Experiments have shown that in all these instances, the protection is effective.

I stress, however, that with respect to reproduction, as with respect to feeding and all other aspects of heteropterans' lives, we know very little about the lives of some bugs, and nothing about the lives of most. One strange exception, unique to a few heteropterans, is "traumatic insemination." One of the male's claspers is sharp and sometimes scimitar-shaped. With it he slashes the female's abdomen and sperm is deposited via a duct running along the clasper. The sperm finds its way to the storage receptacle of the female. In the evolutionarily more advanced members of this group, a special pad of tissue on one side of the female's abdomen receives the sperm, which then moves to the storage site, passing on its way through the cells. This is the situation in the common bedbug. In evolutionarily less advanced species (certain members of the bedbug's family and some Anthocoridae), there is no such tissue, and the sperm are simply deposited in the abdomen. Counting the resulting mating scars allows one to learn how often a female has been mated.
Like those of most insects, the eggs of heteropterans hatch within the year; often they wait through cold seasons (in temperate regions) or dry ones (in tropical regions). Hatching often occurs through a weakened area of the egg and is aided by an "egg burster" on the hatching, which is soon lost.

After hatching, the immature bug molts several times, the last time to adult; five immature stages occur in nearly all bugs, and four in the rest. The first immature of several plant-feeding bugs does not feed. Blood-feeding bugs require a bloodmeal to molt to the next stage; molting may require a full meal in other bugs (especially predacious ones), but in fact the stimulus to molt in nonbloodfeeders is unknown.

Each successive stage is larger than its predecessor, of course, and in each the wings are more fully developed. Development of the wings is so consistent in Heteroptera, that the immature stage can often be determined by the relative development of the wings. The sex can often be detected in fourth-stage, and always in fifth-stage, immatures, as the outlines of the external genitalia become discernible.

In general, the immatures feed on the same organisms as the adults. Predacious immatures may take smaller prey, or they may collaborate in taking larger prey. The young of some plant-feeding bugs feed on different plants than the adults, perhaps in response to differing seasonal availabilities of different plants; perhaps also some plants provide better food for growth and others better food for reproduction.

In temperate regions, most bugs go through a single life cycle, overwintering as egg or adult (there being no pupal stage). Eggs overwinter embedded in twigs or in leaf litter; diapausing adults overwinter in litter or just below the soil surface, near their host plants. In spring, overwintered adults mate, lay eggs, and the hatchlings actively seek food (those of some species do not feed, however). Overwintered eggs hatch usually in midspring, and become adult in midsummer or late summer, mating and laying eggs that will overwinter.

In warmer regions (and, for some species, in temperate regions), there may be more than one generation in a year, and these may overlap. In regions with dry seasons, adults or eggs may estivate. Aquatic heteropterans may overwinter as eggs in or near the water; several species are active below the ice in temperate-region winter; and adults of water striders and their relatives often overwinter on land, rather far from water.

Heteropterans vary widely in their ecological adaptations, both to the biotic and the abiotic aspects of their habitats. Their relationship with their sources of food is of particular significance, both ecologically and because it is this relationship that determines their impact on humans.

Because heteropterans have sucking mouthparts, they must take in liquid or semiliquid food and can accommodate only the smallest of particulate matter. This constraint lies at the base of their relationships to their food.

Plant-feeding heteropterans vary in their degree of host specificity. Some species feed on only a single species of plant, others on members of the same genus; others on many members of the same plant family; and many bugs feed very widely indeed. One cored is listed from plants in 90 genera representing 35 plant families! It is of interest that a single genus may contain both very highly host-specific species and species feeding on a great variety of plants; this phenomenon is not restricted to Heteroptera.

Not surprisingly, the more host specific a bug is, the more closely its life cycle is tied to the host's. Eggs hatch when the preferred parts of the preferred host plant are available. Hatch occurs later in the plant's life cycle if the bug prefers reproductive parts, earlier if the bug prefers somatic parts. The triggers for hatch are not known in nearly all cases. The eggs of some mirids are laid in woody tissue of the host and prepare to hatch when the spring uptake of moisture causes the wood to swell and exert pressure. Several other species of Miridae specialize on grasses, and as these different grasses succeed one another in a field over several weeks, so do the mirids. Again, the dynamics and pheno- nology of this succession remain to be studied.

Mirids in particular, many of which feed on short-lived annuals, are closely tied to their hosts' life cycles. The adults of several species of heteropterans feed on different plants than do the immatures. This may reflect differing needs of the stages: immatures need nutrients for growth, adults need nutrients for reproduction; or, the switch from one host to another may merely reflect that fact that these bugs live longer than the plants preferred by the immatures.

The newly hatched immatures of many phytophagous bugs do not feed at all, or take only water; some species take up symbionts from the egg-shells. This period of time when feeding does not occur may be a "hedge" against a delayed appearance of the host plant. It would be interesting to seek a correlation between species with nonfeeding hatchlings and the uncertainty of their host plants' appearance.

Many heteropterans prey on other insects; a few feed on vertebrates; and only a very few feed, occasionally, on noninsect arthropods. In general, predacious bugs feed on prey of appropriate size: very small predators...
feed on small prey, such as insect (and mite) eggs, mites, scales, and aphids. Large predators feed on larger prey. However, many larger prey organisms (such as caterpillars) are slow-moving and have defenses against predators on their body surfaces; such organisms are available to predators that avoid the body’s surface (by penetrating it with their beaks). And so many small predacious bugs will attack large prey, often collectively; such bugs include small adults and the small early immatures of predators whose adults are large. There is some evidence that the collective effect of the predators’ injected digestive enzymes benefits all the predators feeding on the single prey. These collective associations are all of the same species. Aquatic and semi-aquatic bugs are nearly all predacious. Those bugs that live beneath the water’s surface prey upon other small aquatic invertebrates (although one group of water bugs feeds on snails, and another on small fish). Bugs that live on the water’s surface (semi-aquatic) prey upon land invertebrates that fall into the water and become trapped on its surface. Predacious bugs—whether land, aquatic, or semi-aquatic—are either sit-and-wait predators or stalk-and-pounce predators. The great majority attacks its prey on foot, not in flight. Bugs that feed upon vertebrate blood live associated with their hosts, not on them. The hosts are vertebrates that live in enclosed habitats: burrows, nests, caves—and houses. The bugs spend the day in cracks and crevices in these habitats and, at night, venture forth to feed. When feeding, the bug injects an anti-irritant, so that only later is the fact of feeding known. The host list is rather limited: a few birds that live communally (swallows, domestic poultry, and a few others), humans (which also live communally), and especially bats—two families of heteropterans (Cimicidae and Polyctenidae) appear to have arisen as specialists on bats. One may speculate that the bedbug itself (family Cimicidae) became acquainted with humans when the latter took shelter from the cold in caves, during early Ice Ages; indeed, the only wild bedbug collected came from a bat cave in the northern Middle East. Surprisingly little is known about the influence of abiotic factors on the individual lives of heteropterans or on bugs’ population or community ecologies. The effects of, and the reactions of bugs to, light and humidity are very poorly known. A few species disperse (i.e., large numbers move from one place to another, although not at regular intervals), and this movement in a few species is influenced by amount of moonlight. In other cases, dispersal is a response to deteriorating environment, whether physical (drying, for example) or biological (dearth of food). Several such bugs (e.g., species of cotton stainers, Dysdercus) lose their flight muscles after such dispersal, presumably to divert resources to reproduction. The existence of circadian rhythms has been documented as influencing a very few physiological phenomena of a very few bugs. The greatest enemy of heteropterans is other arthropods—this is of course true of all groups of insects. Small heteropterans are fed upon by small predacious arthropods, and large heteropterans by large ones. Large heteropterans are also eaten by insectivorous vertebrates—fish, reptiles, birds, and small mammals. All bugs are attacked by ants. It has been thought that the scent glands, whose possession is a unique characteristic of heteropterans, are defensive against predators. This is probably so, but it is certainly not their only function. Nevertheless, the few actual experiments and observations that have published support the idea that predators—especially ants and vertebrates—are repelled by the glands’ secretions. More work on this is needed. Many true bugs are brightly colored, arrayed in contrasting patterns of black with red, orange, yellow, or white. These bugs, said to be warningly colored, are assumed to be distasteful to predators (at least, to visually orienting predators, like vertebrates) and to warn such predators away. Experiments with the large milkweed bug (family Lygaeidae) confirm that this is the case. The bug feeds on the intensely bitter sap of milkweed and sequesters the cardenolides (which are also strong ATPase inhibitors); these have no effect on the bug itself. There are many other bugs with conspicuous patterns like the milkweed bug. Many of these bugs are also had-tasting, but certainly many are not. Batesian mimicry (where a palatable organism looks like an unpalatable one, and thereby gains protection) is common in Heteroptera. So is Mullerian mimicry, where two unpalatable species resemble one another and presumably each gains from the protection afforded by the other. The difference between Batesian and Mullerian mimicry is graded, not absolute. Also, we need experimental evidence to show what type of mimicry two phylogenetically different but similar bugs represent. In making these assessments, one should also realize that the types of patterns, and colors, “available” (in the ecological-evolutionary sense) to these bugs are limited. The dark parts of the pattern must be black or near black, and the colored parts red to yellow, because of pigment limitations. The types of patterns are also limited by the smallness of the canvas (the insect’s
unlikely. More probably, these are Batesian (or Mullerian) mimicry complexes, and what the prey loses in associating with the predator, it more than gains by being protectively—warningly—colored like the predator. This suggestion needs experimental testing. Such complexes often involve large populations. Other (nonheteropteran) warningly colored species also live in large populations. Perhaps the mass of bright color warns motile predators (like birds) away from a distance.

Other bugs protect themselves by mimicking inedible bits of their environment. At least seven distantly related families and superfamilies contain species that mimic ants; and within several of these families such mimicry has evolved several times independently. Ants are unattractive as food, because of their high content of formic acid, and many other insects mimic ants. In the Heteroptera there are species that look so much like ants that even experts may be fooled. When mimics look like particular species of ants (rather than like ants in general), they look like local species. The immatures of some Alydidae look like small species of ants when young and like larger species of ants as they grow. The females of some lace bugs (Tingidae) ward off parasites than (for example) the Lepidoptera. And elongate. This is probably why ant mimicry has not developed in the very large superfamilies Pentatomomorpha, whose members are large and squat. Third, ants are characterized by a constriction between thorax and abdomen (the “waist”). Many bugs can defend themselves more actively, as anyone who has been “bitten” by a backswimmer or giant water bug knows. Predacious bugs especially, with their stout beaks designed to be driven through the hard exoskeleton of another insect, can defend themselves effectively against vertebrate predators—or against insect collectors.

Aquatic bugs would seem to be delightful prey for fish. Yet they are not so frequently eaten; indeed, even though trout live in the same streams as water striders, stomach contents indicate the striders are rarely eaten. Many water bugs “bite” fiercely (as most aquatic biologists know), and it seems likely that their scent glands help protect them, although the evidence for this is very slim.

Many unrelated heteropteran ant mimics have one or more dorsal spines, presumably to harm the tender mouth tissues of a vertebrate not deterred by the antlike appearance.

Less common is the mimicry of wasps, which occurs in some adult Alydidae, and doubtless in a few other bugs. Other heteropterans look like twigs, or leaves, or other things of no interest to a predator. Many run or fly rapidly and erratically. Many bugs can defend themselves more actively, as anyone who has been “bitten” by a backswimmer or giant water bug knows. Predacious bugs especially, with their stout beaks designed to be driven through the hard exoskeleton of another insect, can defend themselves effectively against vertebrate predators—or against insect collectors.

Aquatic bugs would seem to be delightful prey for fish. Yet they are not so frequently eaten; indeed, even though trout live in the same streams as water striders, stomach contents indicate the striders are rarely eaten. Many water bugs “bite” fiercely (as most aquatic biologists know), and it seems likely that their scent glands help protect them, although the evidence for this is very slim.

Like many other animals, some true bugs break up their body outline by using spots or lines of white against a dark background (the “zebra effect”); indeed, the antlike appearance of some bugs may have arisen this way. Many Coreidae have a white line, straight or zigzagged, across the dark wings; many Pyrrhocoridae have white spots in the same place, as do some members of other groups. Like the zebra’s white stripes, these bits of white break up the bug’s outline and make it more difficult to be seen—especially when moving in an erratic way.

Heteropterans seem not to be heavily parasitized. Some groups of parasites are specialized for parasitizing certain bugs (the subfamily Phasinae of the dipteran family Tachinidae, for example); the eggs of heteropterans may at times be parasitized by hymenopterans; and certainly many species of heteropterans are quite heavily parasitized. But overall, the group seems less afflicted with parasites than (for example) the Lepidoptera.

The females of some lace bugs (Tingidae) ward off predators and parasites by protecting their eggs and young, and this behavior occurs in a few other heteropteran groups. Some heteropterans bury their eggs partly...
or completely in the substrate, probably to prevent their being parasitized. Many lace bugs, for example, embed their eggs on the lower surface (less conspicuous) of leaves and cover them with frass, thus concealing them from sight and (perhaps) from scent.

Finally, like all arthropods, heteropterans wear their skeletons on the outside. This makes their bodies hard and often smooth. As a result, several kinds of potential predators (especially ones with jaws) simply cannot grasp or crush heteropterans (or many other arthropods). For example, there are few records of predacious beetles feeding on bugs.

C. Economic Importance

Many heteropterans are serious pests of crops, and one group transmits the trypanosome protozoan that causes Chagas’ disease. This group is the Triatominae, a subfamily of assassin bugs (Reduviidae) that lives in the New World Tropics (with one exception) and in which Trypanosoma cruzi develops. These bugs feed on vertebrate blood. Several live in association with humans and can transmit Trypanosoma cruzi to them. The result is Chagas’ disease, which afflicts millions of people a year in Latin America, often killing its victims and costing several billions of dollars in lost lives, lost productivity, and disease prevention and treatment (see Schofield, 1994). A genus of Triatominae lives in India, but so far Chagas’ disease does not occur there. Very rarely, the disease occurs in the southern United States, where the bugs, the trypanosomes, and the wild rodent hosts all may be found.

More widespread is the damage caused to plant crops by heteropterans. Several species of stinkbugs, called collectively Sunn pests, are the chief destroyers of wheat and barley in a broad swathe from eastern Europe through the Middle East into Pakistan. In parts of this range, these bugs may destroy up to 70% of the crop (see Javahery, 1995). The southern green stinkbug is a major pest of many crops throughout the world. In the United States, the chinch bug Blissus leucopterus (Lygaeoidea: Blissidae) is sometimes a serious pest of wheat. Cotton stainers are damage cotton and cotton seed in many semitropical parts of the world (but not the United States). Lygus bugs (Miridae) are very serious pests of a wide variety of crops, especially in the temperate parts of North America. Many other heteropterans are general or specific pests, locally or worldwide, on many crops: A detailed account of these pest insects may be found in Schaefer and Panizzi (2000).

A few heteropterans are important in biological control. A very few feed on weeds and provide some control. Many are predacious and feed on other insects. However, like most predators, most heteropterans are not prey-specific and cannot usually be depended upon to control completely pests on a crop; moreover, they do not feed on many prey items. These insects would serve better as components of an Integrated Control Program. Among the families that have been studied are Reduviidae, Berytidae, Nabidae, Anthocoridae, (some) Miridae, Geocoridae, and Pentatomidae: Asopinae; members of several aquatic families have been tried in controlling mosquitoes, without much success (see Schaefer and Panizzi, 2000).

As noted above, predacious heteropterans feed on only a few prey. When it was believed that these bugs removed from their prey only the latter’s liquids, it was also thought that each predator must feed on many prey in order to get enough nourishment. Now we know this is not so. As a result, those using these bugs in biocontrol programs must rethink the bugs’ efficiency, because each bug actually feeds on far fewer prey items than had been thought. Many prey-predator models, and considerable biocontrol planning, have been rendered invalid by the recent discovery of this mode of feeding (Cohen, 1998).

III. HETEROPTERA: SPECIFIC

The suborder Heteroptera contains eight infraorders: Eucuculiphalomorpha, Diplocuculipomorpha, Gerromorpha, Leptopodomorpha, Nepomorpha, Cimicomorpha, Pentatomomorpha, and Aradomorpha. The “-morpha” means “in the form of,” and each infraorder is derived from the generic name of a member. Hence the order Pentatomidae was thought to have at least two suborders: Pentatomomorpha and Aradomorpha. The first two contain relatively few species (130 and 210, respectively), of poorly known insects; these are all predacious and many live in or on ground debris; it is certain that many more species will be discovered when this type of habitat is explored in the Tropics. They are of interest because they are phylogenetically primitive.

The Gerromorpha contain about 1300 species, most of them in the Gerridae, the water striders, or water skaters, or “Jesus bugs,” which are common everywhere skimming the surface of still waters or the edges of moving waters. Here they capture insects that fall into the water; they rarely if ever capture aquatic organisms. One group of water striders is among the very few
insects that live on or in saltwater: the marine water striders live on the ocean’s surface, often far from land. Because they live on the water, not in it, Gerridae are called “semi-aquatic” bugs, or, in an earlier terminology, Amphibicaudidae.

The Leptocoridae too are semi-aquatic, most of the 300 or so species living near water but neither on it nor in it; a few species live in the splash zone of the sea. All are predacious. Most belong to the family Saldidae, the shore bugs, which are fairly common on rocks in or near water, where their ability to jump and conceal themselves in crevices frustrates the best of collectors.

The 2000 species of Nepomorpha, aquatic bugs (Hydrocoridae), live below the water’s surface. Here are the familiar backswimmers (Notonectidae), water boatmen (Coridae), giant water bugs (Belostomatidae; some Indian species may reach 4 in. in length), and water scorpions (Neptidae); there are several other nepomorphan families that have more tropical species than temperate. All of these are predacious, although water boatmen feed also on freshwater algae. Some giant water bugs specialize on freshwater snails and may be useful in controlling those harboring the schistosomiasis platyhelminth. Others sometimes cause some damage in fisheries, where they may attack young fish. From time to time, someone suggests using some of these aquatic bugs to control mosquitoes, but the attempts have yet to succeed.

The Cimicomorpha, with more than 19,000 species, is the largest of the heteropteran infraorders. This is because it contains the Miridae, or plant bugs, which, with almost 10,000 species, is the largest heteropteran family. Cimicomorpha also hold the Reduviidae, whose 6700 species make it the second-largest heteropteran infraorder. Reduviidae, are the most important. Some are specialists (unusual for predators), one group feeding on millipedes, another living in spider webs and stealing food from the spider; yet another feeds on termites and sometimes lures them from the mound. The group’s biocontrol potential has been little studied, except recently in southern India (see Schaefer and Panizzi, 2000). As mentioned above, members of the subfamily Triatominae feed on vertebrate blood, and the neotropical members transmit the Chagas’ disease trypanosome to humans.

All 80 species of another cimicomorphan group, Cimicidae, feed on vertebrate blood, especially on that of cave-dwelling vertebrates such as bats, communally nesting birds—and man. The bedbugs (Cimex lectularius L. and C. hemipterus Fabricius) are parasites of man. C. lectularius feeds only on man and is one of a very few insects that occur only with man (the housefly and the head and body louse are among the others). Cimex hemipterus is as times a pest of poultry. C. lectularius is primarily a temperate-zone species, although it extends into the Tropics; C. hemipterus is tropical.

Other important cimicomorphan groups include the damsel bugs (Nabidae), a worldwide family of nearly 400 species, common in crop fields in North America and often studied for their biocontrol potential; and Anthocoridae, minute flower bugs, a group (600 species), also worldwide, of very small and often strikingly patterned insects much studied for their control of pests, especially in greenhouses.

The seventh infraorder, Pentatomomorpha, is also large, with at least 11,600 species. Two of its families are tied for third largest in Heteroptera: Pentatomidae and Lygaeidae, each with about 4100 described species (Note: Recent work has shown that many of the subfamilies of Lygaeidae are worthy of family rank). Pentatomomorpha is one of only two infraorders that appear to have arisen as plant feeders, although one important subgroup, the subfamily Asopinae (family Pentatomidae), is secondarily predacious and important in biocontrol programs. Most pentatomomorphs feed on the nitrogen-rich reproductive parts of plants, especially on their ripe and ripening seeds. Pentatomomorpha is also the most difficult infraorder to spell correctly.

A major pentatomomorph group is the Lygaeidae, or seed bugs and milkweed bugs; most of these feed on seeds, although a few groups feed on grasses and members of one of these, Liliidae or chinch bugs, sometimes become pests on wheat. Another major group is the Pentatomidae (stinkbugs), which also feed on reproductive parts, although many others feed on somatic tissue. Included here is the southern green...
sinkling, *Nezara viridula* (L.), a major pest of many crops across the world. Here also are the Sunn pests, a group of pentatomids and some Scutelleridae (a family closely related to Pentatomidae) that ravage wheat and barley throughout the Middle East and surrounding areas. Several small families are related to Pentatomidae, including the Cydnidae, many of which live in the soil and suck from the roots of plants. Members of Scutelleridae are often brilliantly and iridescently colored (although Sunn pests are not). Females of many Acanthosomatidae guard their eggs from parasites and their young from predators.

The Coreidae (leaf-footed bugs) (1300 species) and Pyrrhocoridae (cotton stainers) (400 species) are related to Lygaeidae. The former includes the squash bug (*Anasa tristis* (De Geer)) and a group of bugs with expansions on their legs ("leaf-footed" bugs); some of the latter (in the Neotropics) are brilliantly colored, and others (in North America) come to houses for warmth when winter sets in. The largest genus in Pyrrhocoridae is the tropical *Dysdercus*, many of whose species are serious pests of cotton, which they damage in part by direct feeding on the seeds. Greater damage is caused by their feeding punctures providing entry into the cotton boll of boll rot disease organisms; these organisms, and the bugs' excreta, destroy and stain the cotton fibers (hence "cotton stainers"). One other species, the European *Pyrrhocoris apterus* L., is famous for its role in the discovery of "paper factor," a discovery that led to the development of natural and artificial juvenile hormone analogs, useful in insect control.

The last (eighth) infraorder is Aradomorpha, until recently included as a member (superfamily Aradoidea) of the Pentatomomorpha, from whose other members it differs in several significant respects. Among these are the long mouthparts and the lack of long sensory hairs arranged in characteristic patterns on the underside of the abdomen; pentatomomorphans have shorter mouthparts (like those of other heteropterans) and have these abdominal sensory hairs (unlike most other heteropterans).

The Aradomorpha are a small group of brownish to grayish flattened bugs, nearly all of which live under the bark of dead or dying trees, where they feed on the long mycelia of fungi; adapted for this are the very long very slender mouthparts, which at rest are kept coiled in a special pouch within the head. These bugs range from about 3 to 10 mm long and are often wingless; the upper surface of the body is frequently "bumpy" or pebbly and the legs and antennae are short. One species, *Aradus cinnamonus* Panzer, differs from the others and feeds on the sap of several pine species; from time to time it becomes a serious pest of commercial pines in northern Europe. It has a 2-year life cycle, and the discovery that some populations reproduce in odd-numbered years, and others in even-numbered years, has led to interesting biogeographical and population studies. Two other species (family Termitaphididae), distantly related to all the others (family Aradidae), live in termite nests, where they lay their eggs among the termites. Only a few millimeters long, these bugs lack an ovipositor, eyes, and wings and probably feed on fungi within the nests. Very little indeed is known about them.

**IV. CONCLUSIONS**

Heteroptera share fundamental properties of structure and biology, especially of feeding: their elongated mouthparts so well adapted for the sucking up of the fluids of other organisms. Yet the diversity of heteropterans is remarkable, in size, form, habitat, and structure. Groups now considered rare are probably far more common than we believe, but live in soil and debris habitats in the Tropics, habitats no one studies. Recent work on tree-canopy habitats in the Tropics has revealed assemblages of unknown or poorly known heteropterans found (apparently) nowhere else. Too many entomologists have for too long scorned insects that cannot be collected with nets. Once such other collecting methods as searching the ground's surface became common, so did groups (like Lygaeoidea) once thought relatively scarce.

It follows that the diversity of Heteroptera, like the diversity of all other groups, is constrained not by evolution but by man's knowledge and by man's willingness to seek diversity in different habitats: For the diversity is there. What I have sketched above is based in truth on a remarkably small number of observations and experiments on a remarkably small number of heteropterans. One hopes those observations are accurate and those heteropterans are representative. Much more work, on many more heteropterans, as well of course as much more collecting, will make even more clear just how diverse and fascinating true bugs are.

See Also the Following Article INSECTS, OVERVIEW
Bibliography


I. PROPERTIES OF LIGHT

Light is a natural phenomenon that allows us to see objects, shapes, and colors. It radiates from the sun, the stars, a flame, or a lightbulb. Light encompasses a broad spectrum of radiation that includes gamma rays, ultraviolet rays, infrared rays, microwaves, and radio waves (Fig. 1). These types of radiation are all classified as nonvisible light and consist of the longest and shortest wavelengths and frequencies. A small section of the spectrum is the visible light spectrum. The wavelengths of this region range from 380 to 750 nanometers (nm) and are the only part of the electromagnetic spectrum (Fig. 1) that can be detected by the human eye.

When light shines on an object, some of the light is absorbed and converted to heat; another percentage is scattered or dispersed in various directions; some is transmitted, and the rest may be reflected depending on the material on which the light strikes.

II. ULTRAVIOLET RADIATION

The electromagnetic spectrum is continuous but the types of electromagnetic radiation do not begin or end at precise points along the spectrum. For example, red light shades into invisible infrared (below red) radiation...
and violet light shades into invisible ultraviolet (beyond violet) radiation. The major source of ultraviolet (UV) radiation for the earth is the sun. UV is responsible for producing suntans and vitamin D in the human body. In humans, overexposure to UV can lead to serious and sometimes irreparable harm. It can cause mutations in cellular DNA, which can ultimately cause significant alterations in cells, the main cause of cancer. Other damaging effects of UV include premature aging, blindness, and sterilization. Ultraviolet radiation also has the ability to kill microorganisms, plants, and animals.

UV can be divided into four wave bands. These are Vacuum UV (<200 nm), UV-C (200–280 nm), UV-B (280–315 nm), and UV-A (315–400 nm). At the earth’s surface, vacuum UV and UV-C are not present because of their absorption by various gases such as oxygen and ozone. The formation of atmospheric oxygen and a stratospheric ozone layer was essential for the evolution of life on earth. The ozone layer shields the terrestrial surface from harmful UV radiation. Unfortunately, through anthropogenic emissions of chlorofluorocarbons (CFCs) and other gases, the ozone layer has been adversely affected. It has thinned and has developed ‘holes’ in polar regions. Thus, there is potential for increased UV radiation hitting the earth’s surface.

The ozone hole over Antarctica has dramatically increased since its discovery in the 1970s. At midlatitudes, ozone levels have also continued to decrease. Stratospheric ozone levels are at their lowest point since measurements began, so current UV-B radiation levels are thought to be close to the maximum. Global ozone measurements from satellites from 1979 to 1993 show increases in UV-B radiation at high and midlatitudes of both hemispheres, but only small changes in the tropics. These estimates assume that cloud cover and pollution have remained constant over the years. Increases in surface erythemal (sun-burning) UV radiation relative to values obtained in the 1970s are estimated to be about 7% at Northern Hemisphere midlatitudes in winter and spring and 4% in summer and fall, 6% at Southern Hemisphere midlatitudes throughout the year, 130% in the Antarctic in spring, 22% in the Arctic in spring. The resulting increases in UV radiation threaten humans, animals, plants, and microorganisms in both terrestrial and aquatic ecosystems.

The pioneering work of Sherwood Rowland and Mario Molina showing that CFCs were responsible for the decrease of stratospheric ozone caused great concern among atmospheric and environmental scientists. They reasoned that CFCs in the stratosphere would be subjected to intense UV radiation, which would break them apart, releasing free chlorine atoms.

All the chlorine of a CFC molecule would eventually be released due to further photochemical breakdown. The free chlorine atoms could then damage stratospheric ozone forming chlorine monoxide (CIO) and molecular oxygen. Molecules of CIO can react to release more chlorine and an oxygen molecule. These reactions are part of the chlorine cycle because of the continuous
regeneration of chlorine as it reacts with ozone. Chlorine is the catalyst in the reaction because it promotes the chemical reaction without being used up. Because chlorine can last for 40 to 100 years, every chlorine atom in the atmosphere can potentially breakdown 100,000 molecules of ozone. CFCs are especially damaging because they transport agents that constantly move chlorine atoms into the stratosphere and because chlorine atoms are removed from the stratosphere very slowly.

By 1978 most Scandinavian countries and the United States banned the use of CFCs in spray cans. In 1983, most European countries proposed a voluntary reduction in the use of CFCs. As a result of increasing concern over the effects of ozone depletion, an international agreement was developed to reduce and eventually eliminate certain anthropogenic ozone destroying substances. A treaty, known as “The Montreal Protocol” was signed by a number of nations on September 16, 1987, in Montreal, Quebec. The protocol, which consisted of a plan that would dramatically cut CFC production, was recognized as the first worldwide effort to solve a massive environmental problem. Since 1987, more than 150 countries have signed the agreement to phase out all use of CFCs by 2000. Despite this agreement, the worsening environmental news about CFCs and ozone depletion in the 1990s led certain nations to adopt stricter measures limiting CFC production.

Several industrial companies have developed CFC substitutes. CFCs and several other chemicals that contribute to ozone depletion have been phased out in the United States and several other countries. Nevertheless, existing stockpiles of these chemicals can be used until the deadline. Moreover, developing countries are on a different time course and will phase out CFCs by 2006. Unfortunately, CFCs are very stable and estimates suggest that those in use today will continue to deplete stratospheric ozone for 50 to 150 years. Atmospheric scientists believe that the Antarctic ozone hole will reappear each year until about 2050. However, because of the Montreal Protocol and other measurements to limit ozone depleting substances, there is some optimism about ozone depletion being curtailed. Bromine (Br) is another important source of ozone depletion. Bromine is present in smaller quantities than chlorine but is more destructive on an atom to atom basis. The largest source of Br is methyl bromide that naturally comes from the oceans and wildfires. However, a large portion of methyl bromide is human made and is used as a fumigant. Another important source of bromine is in fire extinguishers. Like CFCs, methyl bromide has a long atmospheric lifetime (approximately 72 years). Regulations on methyl bromide use are being debated.

III. MEASUREMENT OF ENVIRONMENTAL UV RADIATION

Although significant advances have been made, measuring UV radiation is difficult. There are significant differences in how various devices measure UV radiation. Moreover, a number of factors affect UV radiation. For example, there are geographical and seasonal variations in UV. Thus, recent studies have confirmed that there are generally higher UV-B levels at lower latitudes in the United States. Other measurements confirm latitudinal differences in Europe, Asia, and New Zealand.

Spectral measurements show higher summer values of both UV-A and UV-B radiation in New Zealand and Australia compared with Germany due to the yearly cycle of the sun-earth distance and to lower stratospheric ozone levels in the Southern Hemisphere and higher air pollution levels in Germany. Additional UV data are being accumulated for midlatitudes and eventually we will gain a more complete picture of the UV situation at these latitudes.

Cloud cover also affects UV-B radiation levels at the surface. Data from several locations in the United States suggest that monthly average UV levels are reduced by 10 to 50% by cloud cover, depending on the location and season. Aerosols (small particles suspended in the air) may also reduce UV levels in polluted areas. The magnitude of this effect is highly variable and may depend on the number of particles and their chemical and physical composition.

UV levels are expected to increase with increasing surface elevation above sea level due to a thinner atmosphere overhead. Measurements in a remote area of Chile, for example, showed increases of 4 to 10% per kilometer. Other locations showed larger vertical gradients of up to 40% per kilometer near Santiago, Chile, and 9 to 23% per kilometer in the Swiss Alps.

Because of the high spatial and temporal variability of surface UV radiation and the difficulty of maintaining calibration within instruments, it will be very difficult to provide global UV climatology and to represent long-term UV trends based on ground-monitoring stations alone. On the other hand, satellite-based measurements can provide global coverage and continuous long-term monitoring. Yet it is difficult to use remote satellite data in an attempt to estimate UV levels from specific microhabitats on earth. The derivation of surface UV
irradiance from satellites is indirect because satellites detect radiation reflected by the atmosphere and the surface of the earth. The use of radiative transfer models is necessary to relate transmission, reflection, and atmospheric absorption. These models have been useful in showing general changes in UV radiation reaching the surface, computed for clear skies using satellite ozone measurements.

IV. EFFECTS OF UV RADIATION ON BIOLOGICAL SYSTEMS

A. DNA Repair
At the terrestrial surface, most UV radiation of biological concern is in the 280 to 315 nm range (UV-B band). UV light induces a variety of photoproducts in DNA. UV photoproducts impede gene expression by blocking transcription and they can cause cell death or mutagenesis. There are a number of responses to DNA damage by living cells. The two fundamental processes for repairing UV-induced DNA damage are direct reversal of DNA damage or its excision. These processes can be explained by analogy. If we consider a rope with a knot, we can either undo the knot (direct reversal of damage) or we can cut the knot out and replace the rope with a new piece (excision of the knot). This descriptive analogy roughly represents how photoreactivating enzyme, photolyase, and excision repair, respectively, work to undo DNA damage.

Thus, harmful photoproducts can be removed by photoreactivating enzyme, photolyase. In some species, this is the most important mechanism for DNA repair. In excision repair, when DNA molecules are damaged, a segment of the strand of DNA containing the damage is cut out by one repair enzyme, and the resulting gap is filled in with nucleotides (building blocks of DNA) properly paired with nucleotides in the undamaged strand.

B. Effects of Increased Solar Ultraviolet Radiation on Terrestrial Plants
Obviously, it is extremely important to understand the effects increasing solar ultraviolet radiation has on terrestrial plants. Plants comprise most of the living matter in terrestrial ecosystems and in agricultural systems can have far-reaching consequences to other organisms, including humans.

Both physiological and developmental processes of plants are affected by UV-B radiation. Several studies have shown that some plant species in greenhouses, in growth chambers, and in the field have reduced growth with reduced leaf area under close to ambient UV-B conditions compared with plants grown under reduced UV-B levels. Increased UV-B radiation may greatly affect photosynthesis. In certain species, such as soybean, sunflower, and corn seedlings, solar UV-B radiation reduced photosynthesis by about 15% when a 12% ozone depletion was simulated. UV-B radiation can also alter the time of flowering and the number of flowers in certain species. Alteration of flowering time can have a severe impact on plants because pollinator availability may be subsequently affected. Anther walls can absorb more than 98% of incident UV-B radiation and pollen walls contain UV-B-absorbing compounds. However, after the transfer to the stigma, pollen may be susceptible to UV-B radiation. Several experimental studies using Mylar or glass filters that shielded plants from UV-B radiation, showed that flowering was enhanced under these regimes. Different studies have shown different effects of UV-B on flowering in different species. For example, flowering is inhibited in such plants as Melilotus and Trifolium but 2ea mays and Sorghum were not affected. In sexually reproducing populations of a desert plant, the effects of UV-B radiation on growth and biomass seem to accumulate in subsequent generations that were exposed to UV-B radiation. Thus, the effects of exposure to UV-B radiation may be amplified.

The yields of certain crop plants may be greatly affected by increases in UV-B radiation. The available data on the effects of UV-B radiation on yield illustrate significant interspecific variability and variability among cultivars complicated by differences in how the experiments were performed. Nevertheless, some species and certain cultivars seem to be more tolerant of UV-B effects than others. For example, results from greenhouse and field tests suggest that the soybean cultivar ‘Forrest’ was more tolerant to UV-B radiation than the cultivar ‘Shore.’ A study of 10 crop species in Florida showed that under UV-B radiation, yields were reduced by 3% and 90% in half of them, including wheat (3% reduction), potato (21% reduction), and squash (90% reduction). In this study, rice, peanut, and corn were not affected.

Several studies have shown that plants may be more susceptible to pathogens and may affect insect pests when exposed to UV-B radiation. Thus, a number of studies have shown that ambient UV-B radiation can reduce insect herbivory of agricultural pests and native plants. Supplementation of solar UV-B radiation in field studies can reduce the population of herbivorous insects in certain systems. It is unclear why changes in
herbivory occur. It is possible that the secondary com-

pound plant defenses may be altered. Most of these

studies suggest that the changes in insect herbivory are
due to changes in host plant tissues.

Experiments conducted in greenhouses and in the

laboratory suggest that viral and fungal pathogens react
differently to UV-B radiation. In some studies UV-B
radiation promotes the severity of disease, whereas in
other studies, it seems to prevent the severity of the
disease. For example, cucumber plants first exposed to
UV-B radiation were more susceptible to subsequent
infection by fungal pathogens. However, if exposed to
UV-B radiation after infection, there was no effect on
the severity of the disease. Other studies have shown
that when UV-B radiation is removed, there is increased
incidence of fungal infections.

Even roots of plants whose shoots were exposed to
higher UV-B radiation can be affected. For example, the
microorganisms associated with the roots of sugar
maple trees were altered when the shoots of the trees
were exposed to UV-B radiation. Because there are differences in tolerance to UV-B
radiation in different species, it is suggested that a
reduction of primary productivity in one plant species
may lead to an increase in primary productivity in an-
other more UV-B-tolerant species. Thus, it is possible
that the overall productivity within an ecosystem may
change, but the species composition of the system may
not change. Even if the plant species composition does
not change, individual plant form may change, which
could affect how these plants compete for sunlight,
motion, and nutrients. This could lead to significant
changes in the overall characteristics of ecosystems.

C. Effects of Radiation on
Aquatic Systems

1. Plankton

The attenuation of solar radiation on the water column
is dependent on a variety of factors. The transparency
of the water to ultraviolet radiation is dependent on
the type of water. For example, there can be highly
turbid coastal waters that do not allow much UV-B to
penetrate very deeply. In comparison, some ocean wa-
ters are clear enough so that UV-B penetration can be
dozens of meters. In the Antarctic, 1% of the solar UV-
B hitting the surface has been measured at a depth of
639m. Solar UV-B has been shown to degrade dissolved
organic carbon (DOC). Increased breakdown of DOC
and subsequent consumption by bacteria increases the
UV-B penetration in the water column.

Globally, phytoplankton is the most important pro-
ducer in aquatic ecosystems. Thus, damage to phyto-
 plankton populations will affect higher trophic levels.
A number of recent studies in a variety of aquatic ecosys-
tems have shown that UV-B radiation affects the growth,
survival, and distribution of phytoplankton. Phyto-
 plankton exist on the top layers of the oceans and fresh-
water aquatic systems. The Antarctic is especially pro-
ductive in phytoplankton and the region is significantly
impacted by UV-B because of the Antarctic ozone hole.
Therefore, a number of studies on the effects of UV-B
on phytoplankton have been conducted in that region.
In certain experiments, productivity was two to four
times higher in tanks where UV-A and UV-B were ex-
cluded. Pigmentation was also affected. In situ incuba-
tions of natural phytoplankton assemblages in Antarctic
waters indicated that photosynthesis was impaired by
about 5% under the ozone hole. A similar result was
found in the tropics. Screening of most UV <378 nm
resulted in a 10 to 20% increase in photosynthesis.
However, no significant decreases in stratospheric
ozone have been observed in the tropics.

Many phytoplankton can actively move to different
positions within a particular habitat. They may do this
via flagella, cilia, or by utilizing buoyancy to adjust
their position in the water column. Various chemical,
magnetic, light, and gravity cues influence movement
so that plankton can maintain specific positions within
the water column. To cope with constantly changing
environmental conditions, these organisms must con-
stantly adjust their positions. If UV radiation affects
motility, or the ability of phytoplankton to respond to
external cues, this may negatively affect their growth
and survival. There is growing evidence that many phy-
toplankton species are under stress from ambient levels
of UV radiation.

2. Macroalgae and Seagrasses

In contrast to the motile phytoplankton, macroalgae
and seagrasses are attached to their growing sites. Thus,
they are restricted to certain depth zones. It is thought
that this zonation is caused, at least in part, to limits of
visible light penetration at various depths. Some species
may be more tolerant to solar radiation than others.
Thus, increased levels of UV-B radiation may expose
algae and seagrasses to levels that they have not encoun-
tered, perhaps affecting growth, or photosynthesis. Sev-
eral studies have shown that UV-B radiation inhibits
photosynthesis in many red, brown, and green algae.
Deep-water algae were most affected whereas intertidal
algae were the least sensitive.

The DNA of algae appears to be poorly shielded
when compared to higher plants. For example, doses
of UV-C radiation (at 294 nm) necessary to kill leaves

ULTRAVIOLET RADIATION
of higher plants appear to be about four orders of magnitude greater than that necessary to kill highly resistant algae. Flavonoids, highly effective UV screening compounds found in higher plants, have not been found in algae. In higher plants, flavonoids exist in high enough concentrations in the epidermis of leaves so that in combination with cuticle waxes and other cell wall components, the incidence of UV radiation is reduced by one or two orders of magnitude. However, algae do produce other UV-absorbing substances, including a yellow protein-carotenoid complex in some species and other substances that afford some protection from UV radiation.

A recent experimental study conducted in Canada illustrates how ecosystems may have complex responses to increased solar UV-B radiation. Solar UV radiation can reduce photosynthesis and growth in bottom-dwelling algal communities in shallow freshwater. However, in this study, greater amounts of algae accumulated in UV-exposed habitats than in UV-protected environments. UV-A and UV-B radiation inhibited insects that feed on algae. Because larval algal consumers are more sensitive to UV radiation than algae, algal abundance increased.

3. Invertebrates

Sunlight can be lethal to a wide variety of marine and freshwater plankton following exposure for just a few hours or a few days. Mortality rates are usually lower in zooplankton that contain photoprotective compounds including pigments derived from dietary plant carotenoids, melanin, and substances known as mycosporine-like amino acids that absorb in the UV-A and UV-B range. The presence of these photoprotective compounds suggests that there is significant selection pressure associated with the harmful effects of UV radiation.

A number of experimental studies have shown that natural levels of UV radiation are lethal to zooplankton. One recent study showed that zooplankton communities exposed to ambient levels of UV-B for three days in Pennsylvania experienced significant mortality when they were not shielded from UV-B radiation. However, mortality was significant only in an oligotrophic lake, not in a eutrophic lake where light penetration is not increased.

Marine invertebrates differ greatly in their sensitivity to UV-B radiation. For example, while one species of crustacean may suffer about 50% mortality at current levels of ambient UV-B radiation at the sea surface, some shrimp can tolerate irradiances higher than those predicted for a 16% ozone depletion. Bottom-dwelling invertebrates of the ocean may also be affected by UV-B radiation. For example, cleavage in sea urchins is impaired by UV radiation. Marine organisms associated with coral reefs, such as sponges, bryozoans, and tunicates, are also adversely affected by UV-B radiation. Corals are affected by UV radiation in a number of ways. Depending on the species and the particular ecosystem, a number of studies illustrate that UV radiation can cause death, inhibit growth, and contribute to coral bleaching. Bleaching, a well-recognized phenomenon, occurs principally via the loss or expulsion of symbiotic algae from the coral host tissue. When the algae are lost, the coral loses its characteristic color and the remaining white skeleton is most noticeable. Although a number of events, including elevated seawater temperature and heavy rains, can contribute to bleaching, there is some evidence that UV radiation may also play a role in certain bleaching events. UV radiation can inhibit photosynthesis in symbiotic algae. UV radiation affects respiration among corals and their symbiotic algae, but not in a consistent way. In some species, respiration increases under UV radiation whereas in other species it may decrease and in others respiration may be unaffected.

UV radiation inhibits the growth of symbiotic algae in culture. Reproduction in corals may be affected by UV-B radiation. Broadcast spawning at night, a nearly universal phenomenon among many reef invertebrates and corals, might be related to avoiding UV-induced DNA damage as well as to reduce predation. In at least one study the larvae of reef corals from shallow water were more resistant to UV-B radiation than those from deeper water.

In the Antarctic, a number of marine invertebrates sustain UV-induced DNA damage. These include worms and crustaceans. Krill, copepods, and gelatinous zooplankton generally have transparent eggs, larvae, and or adults stages that are pelagic, planktonic, and are often found in surface water for several months. Thus, these species are especially vulnerable to DNA damage from elevated UV-B exposure. These species are important components of the ecosystem and damage to them could have significant ecosystem consequences.

4. Fishes and Amphibians

A recent study of Antarctic zooplankton, including larval fish, showed that they sustain DNA damage during periods of increased UV-B flux. Fish larvae in Antarctic marine ecosystems sustained UV-B-induced DNA damage greater than the lethal limit determined for Antarctic diatoms and comparable to the lethal limit of damage.
for cultured goldfish cells. DNA damage has been shown to be especially correlated with daily UV-B flux in icefish eggs. Icefish larvae, however, showed patterns of DNA damage that correlated less well with daily UV-B flux. Antarctic fish appear primarily to use photolyase to remove harmful UV-B-induced photoproducts.

Other fish species may also be affected by UV-B radiation. For example, in the Arctic ecosystem, many economically important fish species, including cod, pollock, herring, and salmon, spawn in open shallow water and are subjected to increased solar UV-B radiation. Because many of their eggs are found near the surface, it is possible that marine fish productivity could decline in this region due to increased UV-B radiation. At this time, however, it is difficult to assess the impact of UV-B radiation on Arctic fish productivity. However, one recent laboratory has shown that salmon exposed to UV-B radiation are more prone to fungal infections and skin lesions.

Several laboratory studies have shown that UV radiation affects the growth, development, and hatching success of certain amphibian species. These studies have shown that under simulated UV light of various intensities, amphibians may develop skin lesions, edema, eye damage, curvature of the body, and behavioral abnormalities. Under relatively low-level but prolonged doses of UV radiation, the mortality of embryos increases compared with controls that were shielded from UV radiation.

a. Amphibian Declines and Ultraviolet Radiation

Amphibian populations are in serious decline in various areas of the world. Unfortunately, the causes for amphibian population declines have been difficult to assess. Much of the information on amphibian declines comes from observational or anecdotal accounts. Hypothesized causes for the declines include habitat destruction (the most obvious cause), pathogens, introduced exotic species, pollution, and increased ultraviolet radiation. These agents may act alone or in combination to contribute to the decline of amphibian populations.

The diversity of locations where amphibian populations have declined prompted consideration of atmospheric factors such as increased ultraviolet irradiance associated with depletion of stratospheric ozone. Several investigators have used field experiments to examine the potential role of ultraviolet-B radiation in amphibian population declines by measuring the mortality of embryos that were shielded from UV radiation compared to embryos that were unshielded. Continuous high mortality in early life stages may ultimately contribute to a decline at the population level.

Recent field experiments from North America, Europe, and Australia show that ambient UV-B damages the embryos of certain amphibian species but not others. Results of these experiments by several different investigators strongly indicate that the hatching success of at least nine species of amphibians, from widely separated locales, is reduced under ambient UV-B radiation. This includes a diverse group taxonomically of two frog species, one toad species, two salamander species, and a newt from North America, two frog species from Australia, and a species of toad from Europe. Some of these species are found in montane areas, others at sea level. A key characteristic shared by these species is that they often lay their eggs in shallow water, where they are exposed to solar radiation.

Hatching success of several other amphibian species in North America, Australia, and Europe were not affected by UV-B radiation. This is not surprising because many studies have demonstrated differential sensitivity of amphibians to various abiotic factors. There may be variation in response to UV-B radiation, perhaps even within a species at different locations. For example, embryos of western toads in Oregon are sensitive to ambient levels of UV-B while those of a different subspecies in Colorado are unaffected.

Based on a limited sample, there is a correlation between resistance to UV-B and the activity of photoreactivating enzyme, photolyase. Species with the highest photolyase activities seem to be more resistant to UV-B radiation. Furthermore, of the species examined, frogs and toads generally have more photolyase activity than salamanders. It is also possible that nuclear excision repair is also being used to counter UV-induced DNA damage, but this has not been measured in amphibian species taken from the wild.

In nature, more than one environmental agent may affect an animal as it develops. This seems also true for amphibians developing at their natural field sites. Field experiments have been used to examine at least three factors that seem to interact synergistically with UV-B: a pathogenic fungus, low pH, and fluoranthene, a polycyclic aromatic hydrocarbon that may pollute aquatic environments impacted by petroleum contamination. In certain combinations with UV radiation, these three agents increase mortality to levels greater than that contributed by UV radiation alone.

Obviously, UV-B cannot be invoked as contributing to all amphibian population declines. For example, UV-B radiation is unlikely to contribute to morality in amphibians that are primarily nocturnal, live under
b. Amphibian Deformities and UV Radiation

Reports of deformed amphibians have been given wide media attention and have been the subject of several recent scientific workshops. The most common reports of deformities are frogs and toads with extra or missing limbs. Three major agents are being examined: pesticides, UV radiation, and parasitic trematode infection. However, there are very few available data on this subject based on experimental work. One recent laboratory study of northern leopard frogs showed that upon exposure to 24 hr of simulated ambient UV light, frogs developed hindlimb malformations. It remains unclear as to whether UV radiation can cause extra limbs in wild amphibian species. However, other deformities in wild species have been observed. These include edema, curvature of the spine, and lesions in larvae and newly metamorphosed amphibians. A recent report showed severe retinal abnormalities consistent with UV damage in a hazing frog species.

V. EFFECTS OF UV RADIATION ON BIOGEOCHEMICAL CYCLES

The effects of increased UV-B radiation on biogeochemical processes may be complex. For example, increased UV-B radiation may affect the magnitude and direction of trace gas on emissions and mineral nutrient cycling in terrestrial systems. Moreover, the effects on these processes may be species specific and may vary in different ecosystems. Increased UV-B radiation may alter the chemical composition of plant tissue, the photodegradation (breakdown) of dead plant matter, the release of carbon dioxide, the microbial decomposers, and nitrogen fixing organisms.

In aquatic ecosystems, increased UV-B radiation may affect the processes that produce organic matter and the processes that degrade organic matter. There may be photodegradation of dissolved organic matter, which may lead to the production of organic acids and ammonium. Photoinhibition of surface aquatic organisms may also affect biogeochemical processes.

The potential effects of increased UV-B radiation on terrestrial and aquatic carbon, nitrogen, sulfur, oxygen, and metal cycles have been explored in a number of studies. These effects and similar effects on biogeochemical cycles in the atmosphere may aid or help impede the buildup of greenhouse gases and aerosols in the atmosphere.

VI. EFFECTS OF UV RADIATION ON HUMANS

UV-B radiation does not penetrate far into the body because most of it is absorbed in the superficial tissue layers. Therefore, much of the UV damage affects the skin and eyes. However, there are also systemic effects. UV-B is the main cause of sunburn and tanning and the formation of vitamin D3 in the skin. UV-B also affects the immune system. UV-B can cause snow blindness and is a significant factor causing cataracts. It also contributes significantly to the aging of the skin and the eyes and is effective in causing skin cancer.

A. Eyes

Photokeratitis is the effect most attributable to exposure to UV radiation. It is similar in effect to a sunburn and often occurs after short-term exposure to UV radiation. The eyeball becomes inflamed and red, often accompanied pain and photophobia (fear of light). This is frequently diagnosed in skiers as snow blindness. Exposure of the eye to UV radiation can also have effects on the cornea. Thus, exposure to UV radiation can contribute to degeneration of the fibrous layer of the cornea. Under certain conditions, exposure to UV radiation can cause an outgrowth of the outermost mucous layer over the cornea, which results in the loss of transparency. Squamous cell carcinoma of the cornea, a malignant neoplasm, can also be found after exposure to solar radiation. These diseases are associated with outdoor living or working near areas of high reflectance (e.g., near water, concrete, or sand).

Cataracts are the leading cause of blindness in the world. They are characterized by a gradual loss in the transparency of the lens of the eye due to oxidized lens proteins. This can lead to blindness unless the affected lens is removed. There is a correlation between certain types of cataracts and exposure to UV-B radiation. Several studies have suggested that the relative risk associated with increased exposure to sunlight and cortical cataracts (those that develop in the outer layer of lens protein) is between about one- and three-fold. One model suggests that a sustained 10% loss of ozone worldwide would lead to an additional 30,000 blind people per year.

B. Sunburn

Sunburn is the most common effect of exposure to UV radiation. This results in the reddening of the skin and possibly blistering. Sensitivity to sunburn and tanning
cumulative lifetime exposure to UV radiation. Further-
to be a clear relationship between developing CM and
most of the mortality. Like BCC, there does not seem
associated with solar radiation, but is also accounts for
It accounts for only about 2 to 3% of the skin cancers
than the number of sunburns per se.

care more pigmented after exposure to the sun.

C. Photoaging
Exposure to sunlight ages the skin. This is known as
photoaging and is characterized by wrinkles, altered
skin pigmentation, and an overgrowth of abnormal elas-
tic fibers in the dermis.

D. Skin Cancer
There are various types of skin cancers. These are basal
cell carcinoma (BCC), squamous cell carcinoma (SCC),
and cutaneous melanoma (CM). The carcinomas of the
skin are often referred to as the "nonmelanoma skin
cancers" (NMSC). The nonmelanoma skin cancers are
clearly correlated with sunlight. They occur primarily
in light-skinned people and then usually on the areas
of the body most exposed to sunlight.

BCC is the predominant form of NMSC in light-
skinned people. The most susceptible people to BCC
are those with the lightest skin and poor tanning ability.
The incidence of BCC in light-skinned populations has
recently been increasing in certain regions. Although
it was originally thought that cumulative lifetime expo-
sure to sunlight was directly related to developing BCC,
recent information suggests that this may not be the
case.

SCC is much less common than BCC but is much
more common than CM in the United States. More than
the other skin cancers, epidemiological data suggest
that cumulative lifetime exposure to UV is a critical
risk factor for developing SCC. Several studies found
an increase in the risk of developing SCC with incidence
of childhood sunburns. However, this may be related
to a high level of childhood exposure to sunlight rather
than the number of sunburns per se.

CM is relatively rare compared with BCC or SCC.
It accounts for only about 2 to 3% of the skin cancers
associated with solar radiation, but is also accounts for
most of the mortality. Like BCC, there does not seem
to be a clear relationship between developing CM and
cumulative lifetime exposure to UV radiation. Further-
more, CM often appears on areas of the body that are not
the most heavily sun-exposed. Several epidemiological
studies have shown that exposure to sun during child-
hood increases the risk of developing CM. An additional
risk factor is the appearance of freckles or moles.

E. Effects on the Immune System
In humans, the skin is the first line of defense against
foreign bodies that may threaten an individual's health.
Thus, the immune system helps maintain health against
infectious diseases, cancers, and parasites. The skin in-
corporates a number of cells from the immune system
that can mount or influence immune responses to for-
eign substances. Substances entering the body, such as
viruses, have to be "recognized" by the immune system
as either "self" or "nonself" (foreign) entities. UV radia-
tion can induce photochemical changes in the skin and
potentially alter cell surface proteins that are used to
determine "self" from "nonself" entities. Thus, UV radia-
tion can act as an immunosuppressive. The immuno-
suppressive effects of UV-B radiation can influence the
outcome of melanoma and nonmelanoma skin cancers,
certain infectious diseases, some forms of autoimmu-
nity, and allergy. For example, implants of UV-induced
tumors between genetically identical mice are rejected
in unexposed hosts but fail to be rejected by UV-ex-
posed mice. UV-B radiation can inhibit local inflamma-
tory responses within UV-irradiated skin. Thus, the
response elicited by injection of an antigen into the
skin of sensitized individuals may be diminished in UV-
irradiated skin.

Cellular immune responses are of great importance
in the defense of a variety of infectious agents. Some
infectious agents can be harmed by exposure to UV
radiation, whereas others are unaffected. Immuno-
pression may also decrease an individual's resistance to
certain infectious diseases. In animal models, human
infectious diseases have been shown to be influenced
by exposure to UV-B radiation. These diseases include
herpes, tuberculosis, trichinella, candidiasis, leishmani-
asis, listeriosis, and Lyme disease. Effects include sup-
pression of immune responses to the organisms or their
antigens, reactivation of latent infections, increased
body loads of infectious organisms, decreased resistance
to reinfection, and reduced survival. The impact of ex-
posure on antigen presenting cells that are in the skin
and act essentially as the skin's defense system suggests
the possibility that UV-B radiation may exacerbate or
ameliorate autoimmune diseases such as Lupus or HIV.
At the very least, UV-induced immune suppression
could affect the course of certain diseases within hu-
man populations.
See Also the Following Articles

BIOGEOCHEMICAL CYCLES • ENDANGERED REPTILES AND AMPHIBIANS • GREENHOUSE EFFECT • PLANKTON, STATUS AND ROLE OF • SEAGRASSES

Bibliography


URBAN–SUBURBAN
ECOLOGY

Ann P. Kinzig* and J. Morgan Grove†
*Arizona State University and †United States Department of Agriculture Forest Service

I. Introduction
II. Definitions
III. The Founding of Cities and Affected Biomes
IV. The Biogeophysical Determinants of Biodiversity
V. Human Ecology and Social Differentiation
VI. Human Impacts on Biodiversity in and around Urban Areas
VII. Cities in the Future

GLOSSARY

human ecology A type of ecology focused on a specific species, Homo sapiens. Human ecology may be thought of as a subset of social ecology, which is a life science focusing on the ecological study of various social species such as ants, bees, wolves, dolphins, or orangutans.
social differentiation A term taken from biology to describe the specialization of functions in a society and to characterize societies over time. Theories of social change propose that increased social differentiation emerges as societies increase in size and complexity. Differentiation is frequently accompanied by a need for increased coordination and increased interdependence in larger and more complex societies.
urban An area characterized by high human population densities, or significant commercial or industrial infrastructure. The boundaries between urban, suburban, and rural are not sharp and can be difficult to characterize.

urban ecology The study of urban systems from an ecological perspective, an emerging field within ecology that strives to understand human interactions in ecological systems in and around urban areas and to develop theories and analyses that include human communities as fundamental components of ecological systems.

ALTHOUGH CITIES MAY be the glory of humanity, Homo sapiens has lived, through most of its history, as relatively isolated bands of hunter–gatherers, migratory herders, or agriculturists in scattered farming villages, farmsteads, or small trading centers. It is only recently that large cities have become commonplace, and only recently that a large fraction of the world’s people has found its home in them. This urbanization trend has significant—and global—ecological consequences. The land use and land cover changes that accompany urbanization can substantially alter biogeochemical cycles, biodiversity levels, and disturbance patterns, among other things.

1. INTRODUCTION

In spite of the evidence that humans are an increasingly global force in determining the structure and functioning of ecosystems, ecologists have in general been reluctant to study those areas in which a human presence—
and thus potentially their impacts on their surroundings—is most intense. Instead, ecologists have been drawn more to the “pristine” ecosystems where humans, if considered at all, are treated as an exogenous perturbing force. Nonetheless, some threads in twentieth-century scholarship offer a basis for studying the ecology of urban systems, beginning with human ecology in the 1920s and advancing to more recent ecological research programs that have begun to fully integrate humans into their ecological observations and explanations (such as the new Urban Long-Term Ecological Research sites in Baltimore and Phoenix, funded by the National Science Foundation).

Our purpose in this article is to examine the impacts of urban settlements on biodiversity. In Section II, we offer definitions that will serve as guides in subsequent sections. Section III gives a general history of urban impacts on biodiversity—the preferential settlement of some environments over others, and thus the differential impacts on homes and species. In Section IV, we discuss determinants of biodiversity in the absence of human influences, and in Section V we examine those aspects of human social organization and activities needed to understand the patterns of biodiversity in urban systems. In Section VI we examine how human activities in urban settings have altered distributions and abundances of species. Finally, in Section VII we briefly discuss what some future trajectories of those alterations might be.

II. DEFINITIONS

There are many possible definitions of what might constitute an urban environment. We might begin with an examination of human population density and define urban as those areas with high, suburban as those areas with moderate, and rural as those areas with low human population densities. We might forego counting humans in favor of counting structures—commercial, manufacturing, and residential—that could define urban and suburban boundaries. As with any definition, the ‘dividing lines’ between urban, suburban, and rural categories would be somewhat arbitrary. We could always resort to using the political boundaries that define cities but many political boundaries do not derive from the contours or features of the landscape, and therefore they do not tend to capture the boundaries one would want to use in an ecological assessment.

We will not attempt or employ an exact definition of urban, suburban, rural, and “pristine” ecosystems here. We will allow some measure of intuition to guide the reader with respect to where those boundaries might lie. We propose that urban ecology in general—and the study of patterns of biodiversity in urban settings in particular—cannot stop at the city boundary. The high human population densities that characterize cities impact their less populated surroundings through flows of resources and wastes, through increased temperatures and other climatic influences, and through the extension of city infrastructure. Thus, assessments of patterns of “urban biodiversity” must extend to what would be considered rural or even pristine environments.

We have and will use “pristine” ecosystems to define those areas that appear untouched or least affected by humans and human activities. The advance of global warming and the global dispersal of such pollutants as DDT mean that such systems no longer exist. Nonetheless, they exist in our minds—for many ecologists and environmentalists, they represent a “desirable” state prior to human interference, one to which we should strive to return ecosystems. The notion that one could—through conservation or restoration—re-create a pristine ecosystem itself contains a fundamental irony or philosophical absurdity; nonetheless, it is the idea of pristine, the idea of a wilderness, that in many ways defines our relationship to Nature. But, again, the boundary line between impacted and pristine ecosystems is fuzzy, and the latter category may no longer exist.

Similarly, defining “natural” ecosystems presents a special challenge to urban and human ecologists. Are humans to be seen as an intrinsic part of urban or other ecosystems—as natural in their impacts as beaver dams or termite mounds? If so, what becomes of the notion of a “natural” system or Nature apart from humans? Is the idea of a “natural” system or Nature apart from humans—a pristine ecosystem itself contains a fundamental irony or philosophical absurdity; nonetheless, it is the idea of pristine, the idea of a wilderness, that in many ways defines our relationship to Nature. But, again, the boundary line between impacted and pristine ecosystems is fuzzy, and the latter category may no longer exist. Philosophical considerations aside, we will sometimes—for want of a better vocabulary—use the word “natural” in a manner similar to “pristine”—relatively untouched by human impacts. Perhaps the most precise definition would be an ecosystem whose structure and functioning can be understood largely without reference to humans or human activities (except that it is humans who must examine and write about them). Nonetheless, we do believe that humans can, to a degree, be studied as other organisms are studied, in terms of their resource needs, interactions with other organisms, and impacts on their surroundings. From our perspective, there may be quantitative and even qualita-
tive differences in these interactions and impacts, but there is nothing "unnatural" about the ecosystems humans have constructed for themselves.

Finally, we use the term "biodiversity" somewhat loosely in the coming pages. We recognize that biodiversity encompasses biological variability across many scales—from the level of the gene to the individual to the population, species, community, habitat, and landscape. We will frequently use species number as our measure of biodiversity, not because we believe that unit is necessarily the "proper" one for discussing distributions of biological resources or human impacts on those resources. It is, however, the measure frequently used in discussions of biological resources, and the measure frequently used for analysis in the literature.

III. THE FOUNDING OF CITIES AND AFFECTED BIOMES

Humans, through their creation of urban and near-urban environments, can profoundly affect the distribution and abundance of populations, species, communities, and ecosystems. They do so in many ways—through fragmentation, the mobilization of formerly recalcitrant or scarce resources, preferential cultivation, or destruction. We will examine these on-site or near-site disturbances more systematically below. But before turning to that discussion, it is worth noting that not all biomes are equally susceptible to the disruptions caused by urban settlements. There are certain characteristics that make sites more or less desirable for dense human settlements. These characteristics have changed over time as our technologies and societies have advanced, allowing acquisition of resources from afar or the defense of even physically vulnerable locations. Nonetheless, we are more apt to find cites nestled in valleys than perched on mountaintops; more likely to find ancient cities on fertile rather than barren soils; more likely to find cities situated on ecotones or "transition zones"—for instance, between the piedmont and coastal plains—because those locations allow access to more diverse resources.

Consider the appearance of the very first urban centers. They would have begun, in many cases, as villages—and later fortresses—around which agricultural fields were established. There would have to be a favorable climate and soil and a ready supply of water. The labor required for clearing would have to be minimal. There would also have to be locally available productive grains and legumes that were a part of the first "packages" of cultivatable materials. Frequently (but not always), the emergence of agriculture as a dominant means of food procurement over hunting and gathering was aided by the availability of domesticatable animals, to be used both as draft animals and protein supplements. It is only in sites sharing these characteristics that enough food could have been produced to allow population densities to become high enough to form settlements somewhat larger than villages, on their way to being urban centers. We know, in fact, many of the places where agriculture first arose—in southwest Asia by 8000 B.C.; Egypt by 6000 B.C.; Greece, Cyprus, and the Indian subcontinent by 6300 B.C.; and Britain by 3500 B.C. The thick forests of the wet tropical regions and the unfavorable climate of the boreal regions made these areas less favorable for the early emergence of intensive agriculture, and thus these biomes were not initially impacted by urban centers.

Later, as civilizations advanced, the need for local agricultural self-sufficiency may have been offset to some extent by the emergence of trade routes. Some urban centers would have been established in areas of particularly favorable mineral or gemstone deposits. Similarly, circumstances of chance or talent would have allowed the emergence of an artisan class within an urban center; both types of sites would have served to anchor the locations of trade routes. Other urban centers may have emerged merely because they were particularly favorable stopping-off places—just a day's journey up or down the road from a final destination—or because they were at the crossroads of two routes. Nonetheless, there would have been certain characteristics of these sites as well that meant some biomes were more susceptible to disruption from urban centers than were other biomes—those along waterways or coasts or valley bottoms, for instance.

In current times, technology has allowed establishment of urban centers in new locations and biomes. Near-global transportation of food products, improvements in our ability to clear land, and our ability to control indoor climate have extended urban centers into deserts, wet tropical forests, and boreal regions. Nonetheless, many of the large urban centers remain in areas of the highest terrestrial productivity. A high percentage continue to be located along waterways or coasts. Much of the growth in urban centers in the coming century will come in those countries with the fastest growing populations; these countries are largely located in the subtropical or tropical regions of both the Northern and Southern Hemispheres. Thus, when we think about the types of biomes—and thus the types
of flora and fauna—likely to be most significantly impacted by urban centers, we should concentrate on temperate forests and grasslands, coastal and riparian zones, savanna/woodland systems, and wet and dry tropical forests. Boreal forests and montane or alpine ecosystems are less likely to be impacted. Desert and arid systems will become increasingly vulnerable as populations grow in the northern regions of sub-Saharan Africa and the North American Southwest, among other locations.

In addition to significant spatial differentiation in the distribution of cities and consequent impacted biomes, there is temporal differentiation as well. Technology not only allows the founding of urban centers in previously unsettled areas but also changes the interactions of a city with its surroundings and the geographic region over which resources are acquired or wastes dispersed. Advances in transportation allow housing expansion in areas further removed from industrial or commercial centers; advances in architecture allow vertical rather than horizontal expansion of the city center. Figure 1 shows the development of Baltimore over the years and illustrates the impacts of technological advancement on that development.

Finally, the situation with respect to regional and global levels of development and trade can also profoundly influence the growth and structure of cities. Older cities can display the residual characteristics of their structures of a century ago—residential areas close to the city center, the robust industrial base—or skeleton of that industrial base—that fueled its urban growth. Newer cities, or newer sections of cities, can

**FIGURE 1** Perspective view of urban growth in Baltimore, MD over 200 years (1792–1992). Yellow polygons (light shading) are built-up areas as determined from historical maps and satellite imagery; green areas (medium shading) are forests and blue (dark) areas are water (Chesapeake Bay). Source: Penny Masuoka, UMBC, NASA Goddard Space Flight Center and William Acevedo, USGS, NASA Ames Research Center. See also color insert, Volume 1.
IV. THE BIOGEOPHYSICAL DETERMINANTS OF BIODIVERSITY

Before looking at the impacts of urban systems on biodiversity, it helps to first have at least a rudimentary understanding of what allows the coexistence and persistence of different types of organisms. Why are there so many species? Why aren’t there more? Why do we see the types of species we do in a particular location? Ecologists do not have a “unified theory” that explains patterns of biodiversity in all systems. Thus it is difficult to offer a complete set of causal mechanisms underlying those patterns. Nevertheless, we review some basic principles that influence the distribution and abundance of organisms in terrestrial systems, and then discuss in later sections the ways in which the organization and structure of urban systems—and the types of human activities taking place within them— influence these determinants of biodiversity and thus biodiversity itself. The following pages discuss in more detail six key concepts: (1) performance trade-offs; (2) environmental and resource heterogeneity; (3) disturbance regimes; (4) invasion dynamics; (5) interspecific interactions; and (6) spatial relationships.

Organisms face several different life “tasks”—to gather resources, to survive, to reproduce. Performing these tasks means trade-offs—resources allocated to seed cannot be allocated to leaves, for instance, and time spent hiding from predators cannot be devoted to gathering food. The ways in which organisms balance these trade-offs leads to a variety of different strategies for survival and reproduction, and thus different types—or species—of organisms. Consider the serpentine grasslands of Jasper Ridge, California. These hilltop systems are found in a Mediterranean climate, characterized by summer droughts and winter rains. Unlike most of the ecosystems of temperate North America, which exhibit growth in the summer and dormancy in the winter, the Mediterranean grasslands grow, flower, and set seed from late autumn to late spring and are dormant during the dry summer months. Serpentine soils are thin, and many of the plants rise a mere 10 cm from the grassland floor. Most of the 100 or more species found there are annual plants—growing, reproducing, and dying within a year. Lepidium nitidum is an early bloomer. As soon as the autumn rains come, it uses the resources derived from photosynthesis to send out shallow roots, extracting water and nutrients needed for further growth from the near-surface layers of the soil. But it soon diverts those acquired resources to seed production, dispersing progeny, and dying a mere 3–4 months after the rains begin.

Contrast this evolved strategy to that of some of the longer-lived annuals. These send out deeper roots, waiting for the drier periods when most of its annual competitors—after drawing resources from shallow and intermediate layers of the soil to use for growth and reproduction—have already died. The deeper roots allow acquisition of moisture and nutrients in locations not reached by the more shallow-rooted plants and continue to garner resources for reproduction even during dry periods. These deep-rooted individuals set seed and die nearly a year after the rains begin. The long-lived annuals may acquire more resources over their longer lifetimes than do the short-lived annuals, but they also must allocate more resources to rooting structures. Early senescence may be more favorable in a particularly dry year, while late senescence could flourish under generous late summer rains, but both strategies—and all the strategies in between—allow for successful competition for the resources required for reproduction, and thus coexistence.

These “performance trade-offs” and the resulting diversity of types are a feature of all ecosystems. Performance trade-offs—and the multitude of strategies that can evolve to balance these trade-offs—allow many different species to coexist in one ecosystem. But not all strategies are equally viable under all conditions. The needlelike leaves of the temperate and boreal evergreen forests would not necessarily serve their wearers well in the Sonoran Desert. The water-loving plant that establishes in riparian zones would not do well under the more arid conditions that exist slightly uphill. Thus, performance trade-offs may allow many different species to emerge and coexist, but different conditions—from place to place around the globe,
or from place to place within an ecosystem or landscape—impose constraints on the types of strategies and species that can persist there. Thus, heterogeneity in resources or environments influence patterns of biodiversity by creating conditions under which different types of organisms—employing different resource acquisition and reproduction tactics—are best suited to thrive.

We can see variations in resources and environments over both broad and small scales. Macroscopic features like mountains can affect climate by forcing air upward, causing the air to get colder and denser and thus “wringing out” the water vapor held within. This makes the upwind sides of mountains moist, the downwind side dry. But topology can play a significant role on smaller scales as well—the footprint that serves as a temporary wetland and home for water-loving bacteria, for instance. Similarly, soils and the nutrients they hold can vary from tropical to boreal regions, but can also vary as we move from tree trunk to root to root-free soil.

In addition to variability in resources and environments, disturbance regimes influence patterns of biodiversity as well. Frequent fires in the prairie, for instance, favor grasses that have allocated significant resources and reproductive structures belowground, where they can remain relatively unscathed by the fire. In the absence of fire, however, those grasses that have allocated more to aboveground reproductive or photosynthetic capacity might thrive. Disturbances also operate over a variety of spatial and temporal scales—from the 100-year floods or storms that devastate huge areas to the tree-fall or worm track that opens up new habitat on much smaller scales. (These examples illustrate another important principle—spatial and temporal scales are frequently related. Thus, large-scale disturbances occur relatively infrequently, small-scale disturbances more frequently.) In addition, disturbance—particularly large-scale disturbance—is frequently patchy. We can see this in large forest fires, which consume some patches while leaving others untouched. Thus, disturbance itself can create patterns of environmental heterogeneity that will influence patterns of biodiversity.

Invasion dynamics also shape communities and ecosystems. Ecosystems are continually “tested” by invading organisms—can an invader find a “niche” either by coexisting within the original complement of species or by outperforming another species, thus driving it to local extinction and creating a niche? The communities we see today are a result of these continuous invasions, a testing and retesting to find or create situations in which a species can survive and reproduce. Those systems subjected to longer or more continuous invasions may have filled more niches—“squeezing in” more types of organisms and strategies. Isolated systems, on the other hand, may contain organisms maladapted to the current environment but able to persist because they have never been tested against a more efficient, invading organism. In addition, the tests invading organisms face may change over time as community composition changes or as climate or other aspects of the disturbance regime change. Thus, a species that originally could not find a foothold may appear later, and one that thrived in an initial invasion may find itself the victim of changing circumstances. The impact of an invasive species on biodiversity levels will depend, in large part, on the interactions the invader encounters within its new home. Impacts can frequently be large, because invasive species often arrive without the diseases, pests, or competitors that keep them in check in their native environments.

Interspecific interactions can affect the complement of species in a given ecosystem. “Keystone” species are those that have a disproportionate influence on community composition and structure—their removal bodes dramatic consequences for biodiversity and the functioning of ecosystems. Less dramatic, but equally important for patterns of biodiversity, are those species that create niches for others—the beetles that burrow into bark, the mice that live on the feet of grasshoppers, the ferns freed from the surface when they colonize towering trees. Antagonistic interactions can influence biodiversity as well—the presence of a predator engenders new strategies for predator evasion; the evolution of a toxin-producing bacterium promotes evolution of strategies of defense.

Finally, the area available for species can influence levels of biodiversity. Ecologists have long observed that larger areas have more species and have been able to quantify this areal influence on species numbers through the species–area relationship. (The species–area relationship basically states that species number is proportional to the area available raised to the power 2, where 2 usually falls between 0.1 and 0.35 and depends on the habitat and the taxonomic group of interest.) Thus, small islands in an island chain tend to have fewer species than large islands in the same chain, and small lakes in a region tend to have fewer species than large lakes in the same region. Connectivity among habitat patches can be important as well, allowing, for instance, replenishment of existing populations from other, separated populations that may serve as an important source of progeny or genetic diversity.

Humans have altered all of these biodiversity-
determining conditions in urban, suburban, and rural systems, with consequent impacts on the patterns of biodiversity we see within and surrounding urban environments. A catalog or “snapshot” of those patterns of biodiversity is certainly possible—we could map the existing patterns of species from the city center to more pristine surroundings, and even compare the species lists and communities with those we surmise existed prior to the founding of the human settlement. Understanding those patterns, however, and developing predictions regarding past and future dynamics—and a comprehension of the types of policies or institutions capable of influencing those patterns—require much more than this mapping or snapshot approach. Instead, we must understand how the pre-settlement landscape influenced human settlement patterns, how human activities altered the landscape, and how these alterations further influenced the decisions and activities of the human inhabitants. We would also need to understand what changes in species distributions might have occurred even in the absence of human settlements—due to natural extinction or speciation events, for instance, or climatic variations. In other words, we need an integrated understanding of ecological and human-social dynamics. Moreover, just as heterogeneity in climate, disturbance, soils, and community composition can affect patterns and dynamics of biodiversity, heterogeneity in human culture and community influences the evolving patterns of urban biodiversity. There can be significant differences in the ways in which city inhabitants and urban communities purposefully manipulate biodiversity or respond to changing conditions in ways that indirectly influence biodiversity. Understanding these differences requires that we further integrate knowledge of diverse socioeconomic and cultural conditions in our human ecological system, just as we must account for the diversity of strategies, species, and conditions inherent in the non-human ecological system.

V. HUMAN ECOLOGY AND SOCIAL DIFFERENTIATION

The study of cities from an ecological perspective in the United States began with Park et al.’s landmark publication, *The City*, in 1925, which formally introduced human ecology as a new research agenda for sociology and the study of cities in America. Their research focused on many of the social changes that had resulted at that time from the rapid expansion of America’s urban areas due to the mass immigration of people from Europe and rural America. The explosive growth of the city, the confluence of people from diverse backgrounds, the breakdown of old ways, and the changes that were necessary for a viable new urban life caught the imagination of the authors. Although Park et al. drew upon the work of European social and biological scientists such as Malthus, Darwin, and Spencer, the initial development of human ecology in America was influenced significantly by and contemporaneous with the emerging fields of plant and animal ecology in America. The Chicago School—as it came to be known—conceived of human ecology as an extension of the developing fields of plant and animal ecology.

The Chicago School articulated and developed an approach to human ecology that drew upon and paralleled earlier ecological work. First, Park and his colleagues applied a community ecology approach to the complexities of urban society in order to uncover a set of regular social patterns and processes in the apparent confusion of the urban melting pot. For instance, Park et al. employed ecological concepts such as succession, competition, and metabolism to describe stages of human community structure (organization) and function (processes): specifically, indicators of social disorganization such as disease, crime, vice, insanity, and suicide. Second, the Chicago School conceived of the city as a closed and functional system (community) that could be treated as an organism or “superorganism.” Park and his colleagues also focused on the spatial and temporal dimensions of the city.

A significant product of this work was Burgess’s ideal model of the city which the Chicago School used to describe and measure the city’s spatial differentiation and development into zones and areas-within-zones through processes of concentration, centralization, segregation, invasion, and succession.

The Chicago School’s conception of human ecology was criticized strongly by social scientists for several reasons, including opposition to the use of biological factors to explain individual human behavior and social structures; a singular reliance upon competition as the mechanism for explaining the organization of economic functions and the spatial distribution of human populations and services; and the use of macroscale processes and functional approaches to explain individual behavior from both a conceptual and a statistical point of view. Many of these criticisms were addressed in later studies in human ecology. It is not enough, however, to treat human social systems as monolithic blocks. Just as performance...
trade-offs and the emergence of different strategies for resource acquisition and reproduction can influence the patterns of biodiversity on the landscape. Differences in the structure and organization of human social systems need to be considered in order to understand how humans impact and manipulate the biodiversity surrounding them.

All social species are characterized to varying degrees by patterns and processes of social differentiation. In the case of Homo sapiens, social differentiation or social morphology has been a central focus of sociology since its inception. In particular, social scientists have used concepts of social identity (age, gender, class, caste, and clan) and social hierarchies (wealth, power, status, knowledge, and territory) to study how and why human societies become differentiated.

Social hierarchies—or social differentiation—is an important concept for understanding human ecological systems because it affects the allocation of critical resources (natural, socioeconomic, and cultural) and the patterns of and impacts on biodiversity. In essence, social differentiation determines "who gets what, when, how, and why." This allocation of critical resources is rarely equitable. Unequal access to and control over critical resources is a consistent fact within and between households, communities, regions, nations, and societies.

Wealth is access to and control over material resources in the form of natural resources, capital (money), or credit. The unequal distribution of wealth is a central feature of human ecological systems, and one that must be accounted for when examining the driving forces behind past changes in the environment and predicting the impacts of future policies and conditions. Power is the ability to alter others' behavior through explicit or implicit coercion. The powerful (often elites with political or economic power) typically have access to resources that are denied the powerless. Status is access to honor and prestige and the relative position of an individual (or group) in an informal hierarchy of social worth. Status is distributed unequally, even within small communities, and high-status individuals may or may not have access to either wealth or power. For instance, a minister or an imam may be respected and influential in a community even though he or she is neither wealthy nor has the ability to alter coercively other people's behavior. Knowledge is access to or control over specialized types of information (technical, scientific, religious, and so forth). Not everyone within a social system has equal access to different types of information. Knowledge often provides advantages in terms of access to and control over the critical resources and services of social institutions. Finally, territory is access to and control over critical resources through formal and informal property rights. These types of hierarchies are important constraints that affect the allocation or supply of critical resources. It is also important to examine and understand variations in demand. For instance, additional social characteristics that may affect the demand for critical resources are often related to demographics or ethnic or religious backgrounds, since different age groups and various cultures may have different needs from or attitudes toward their environment and each other.

Processes of social differentiation of human ecological systems also have a spatial dimension that is usually characterized by patterns of territoriality that lead to spatial heterogeneity on many scales. This spatial understanding of social differentiation in an ecological context enables researchers to ask "who gets what, when, how, why, and where?" and, subsequently, to ask about the reciprocal relationships between spatial patterns and sociocultural and biophysical patterns and processes of a given area (Grove and Burch, 1997).

Today, it is increasingly difficult to determine where traditional biogeophysical ecology ends and human ecology begins (Golley, 1993). The articulation of a biosocial approach to human ecological systems has occurred over the past 20 years and continues from the development of and discourse between plant, nonhuman animal, and human ecologies and social sciences. Indeed this integration among ecological and social sciences is absolutely essential if we are to understand the functioning and dynamics of cities in an ecological landscape, and if we are to be able to both understand and predict the impacts of humans in urban centers on biodiversity.

VI. HUMAN IMPACTS ON BIODIVERSITY IN AND AROUND URBAN AREAS

Once an urban center has been established, human population densities and activities can affect biodiversity in many ways—both increasing and decreasing the variety of flora and fauna. These impacts span a variety of spatial scales. Thus, we can discuss the alteration of biodiversity in the urban centers or surrounding suburban regions, or examine the impacts at the landscape level due to the import and procurement of resources from surrounding rural regions to urban areas or the export of pollution or heat. Moreover, the effects
on biodiversity span both ecological and evolutionary time scales. Finally, the intensity of impacts can vary—from the complete loss of species or the creation of entirely new ecosystem complexes to less obvious alterations in the genetic makeup of populations.

Recall those characteristics of the environment that could vary within and across ecosystems and influence community composition—resource availability, environmental heterogeneity, disturbance regimes, invasion dynamics, interspecific interactions, and habitat area and connectivity. Humans alter all of these determinants of biodiversity, and thus alter patterns of biodiversity within and around urban systems. In addition, humans create, alter, or destroy habitats in urban centers. Thus human influences on biodiversity fall into at least four categories:

1. Alteration of habitat
2. Alteration of resource flows
3. Alteration of disturbance regimes
4. Alteration of species composition (invasion dynamics and interspecific interactions)

Space precludes an exhaustive analysis of the extent and patterns of each of these impacts. Instead we briefly discuss each category below; a summary of these various categories is presented in Table I.

### Table I

A Classification of Human Impacts That Affect Biodiversity in Urban Settings

<table>
<thead>
<tr>
<th>Impact Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alteration of habitat</td>
<td>Destruction of habitat, Creation of habitat, Fragmentation of habitat</td>
</tr>
<tr>
<td>Alteration of resource flows</td>
<td>Reduction in net primary production, Mobilization and concentration of nutrients, Dispersion of toxins</td>
</tr>
<tr>
<td>Alteration of disturbance regimes</td>
<td>Increase in small-scale disturbance, Unintentional shifts in disturbance</td>
</tr>
<tr>
<td>Alteration of species composition</td>
<td>Preferential cultivation or destruction of native species, Introduction of non-native species</td>
</tr>
</tbody>
</table>

### A. Alteration of Habitat

Humans alter habitats in several different ways in and around urban settings. These alterations can roughly be grouped into the three categories of

1. Destruction of habitats
2. Creation of habitats
3. Fragmentation of habitats

#### 1. Destruction of Habitats

Perhaps the most obvious destruction of habitat comes from the nearly complete elimination of the ecological community that existed prior to human settlement as the result of the paving of land or clearing for building or agriculture. Any species endemic to the converted areas would be lost and, by the species-area relationship, a reduction in the overall area of habitat would lead to a reduction in the number of species the remaining habitat could support. Habitat destruction could also lead to the loss of different populations within a species, and these lost populations may carry unique genetic information.

But human activities in and around urban settings destroy habitats in other ways. The vegetation structure in urban areas, for instance, can differ from the ecosystem the city replaced. Thus orchards can be depauperate in understory vegetation; similarly, many urban gardens lack primary producers that rise more than a few feet in height. This change in structure can affect the habitats available to small animals and birds, among other organisms. Similarly, the loss of many large predators from urban systems means the loss of any companion species—the mites, insects, or bacteria that may preferentially or obligatorily colonize the host.

The extent and pattern of habitat destruction will depend on the socioeconomic characteristics of the community in which land conversion is taking place and the power relationships between that community and nearby communities. Wealthy communities—frequently located at the urban fringes or suburban areas—are more likely to have the resources required to own and manage large residential lots and, in some cases, therefore allow natural vegetation to remain somewhat intact. These same resources, however, may be associated with increased applications of fertilizers and pesticides. Poorer communities frequently find themselves—in contrast—on the most infertile or poorly drained soils (the latter making these communities historically more susceptible to diseases such as cholera, malaria, and dysentery).
There have been few studies on how the rate of destruction of habitat depends on such characteristics as the age of the city, the stage of development of regional or global markets, the institutional arrangements for managing natural resources or urban expansion, or the biogeophysical characteristics of the region. Some ecologists working in urban settings, or along urban to rural gradients, have examined the differences in ecosystem structure that urban development can bring, but we do not yet have a comprehensive picture of the general patterns of those changes in structure or what those changes might mean for preferential selection or suppression of certain types of organisms.

2. Creation of Habitats

The transformations that accompany the creation of urban centers not only destroy habitat, they create habitats. Some of these new habitats are not necessarily unique to the urban environments—they are merely the result of any human settlement or presence. These types of habitats would include agricultural fields, gardens, barns, and houses. The impacts of these kinds of transformations and structures on the biological resources of the planet have been around for millennia and have, in some cases, led to the creation of new species or subspecies that are obligates for a human presence. Other urban habitats—such as sewage lines, landfills, and factories—have emerged in more recent times. Humans also create new ecological communities by mixing native and nonnative species in residential or commercial lots or parks and botanical gardens.

Some created “urban habitats” disappear as urban development progresses. Thus, raw surface sewage was not uncommon in English cities of the early nineteenth century; today, this habitat would be found primarily in the cities or urban fringes of those portions of the world lacking the resources required to ameliorate this problem. Other habitats, previously absent, appear in ever increasing abundance—witness the rise of golf courses over the latter half of this century. Finally, the economic development of a city and its surroundings can allow the return of previously destroyed habitat. Thus, we see regrowth of forests in the northeastern United States and in Europe, in part driven by more efficient agriculture and in part driven by a shift from agricultural to industrial to service economies in some regions.

3. Fragmentation of Habitats

Urban development fragments the landscape into patches of varying sizes, qualities, and land uses. The size and configuration of the patches will have important implications for organisms in the urban environment. Perhaps the most direct impact of that fragmentation is the elimination of species that require large contiguous ranges in which to obtain the resources required for survival. Thus, we see the elimination of much of the large-to-medium fauna in city centers and urban environments (except for humans and their associated pets). Further, some species must rely on dispersal to maintain populations in smaller or more widely separated patch types, as some patches may be too small to maintain indigenous populations. Thus, fragmentation may lead to shifts in species abundance that are related to dispersal characteristics. Migration routes may be adversely influenced by fragmentation of habitats as well.

Fragmentation also introduces the urban equivalent of “ecotones”—transition zones between pavement and yard, or a lawn and a garden, or a park and a commercial mall. The preponderance of edges in a highly fragmented landscape increases the abundance of edge-dwelling species and decreases the abundance of those species particularly sensitive to edge effects. Finally, fragmentation can affect the exchange of water, nutrients, and energy both within and across patches, leading to changes in the competitive outcomes among resident species.

Again, there have been few studies of the patterns of patch size, distribution, type, and quality in urban environments, how those patterns change over time and across urban to rural gradients, or how those patterns depend on the biogeophysical or socioeconomic characteristics of the region. Fragmentation may in general be more intense within the city proper and diminish as we move toward rural settings, but how does the city center compare to an urban–suburban transition zone with respect to distribution and quality of patches? Does fragmentation increase over time as more activities and structures are accommodated within the city boundaries, or does it decrease over time as economic growth allows greater landholdings and more preservation? And how will communities resemble as patterns of fragmentation change—particularly if we have lost many species in response to fragmentation, climatic change, or other urban disturbances?

B. Alteration of Resource Flows

The availability of resources such as energy, temperature, nutrients, and water has profound influences on the distribution and abundance of species. Humans can and do alter all of these resource flows in and around
the urban center. Biodiversity-impacting activities generally fall into six categories, as follows:

1. Reduction in net primary production
2. Increase in regional temperature
3. Mobilization and concentration of nutrients
4. Dispersion of toxins
5. Diversions of water and changes in the timing of water availability
6. Degradation of water quality

With respect to (1), conversion of habitat in urban regions frequently leads to a reduction in net primary productivity as vegetation is eliminated or thinned. Quite apart from the impacts of lost habitat area, the reduction in net primary productivity will impact the complexity of the plant and animal community and the number of species that can be supported given the available energy. Since all food transfers involve some loss of energy that cannot thus be put to productive purposes, the net primary productivity—or the amount of energy fixed by photosynthesizers and available to support nonphotosynthetic life forms—will determine the extent to which multiple trophic layers can be sustained within a food web. On the other hand, in some environments—such as the Sonoran desert of central Arizona—net primary production in irrigated and managed yards may exceed the primary production of the desert that used to be there.

The original capacity of the land to fix energy through photosynthetic activity is instead replaced or supplemented with the energy plants and combustion engines required to power residential, commercial, and industrial activity. There is also “waste heat” associated with these energy conversions, and this, in conjunction with the preponderance of heat-absorbing dark surfaces, leads to the well-known urban heat island effect, which can elevate city temperatures as much as 3°F above the surrounding countryside. Such an increase in average temperatures could prove detrimental to those nonnonsissile organisms living at the very edge of their optimum temperature range.

Human activities also mobilize and concentrate nutrients. The disruption of soils that accompanies land conversion and building can stimulate nutrient cycles and promote soil erosion—both can increase the nutrient loading of the river or groundwater draining the watershed in which a city is located. Further, there is a substantial import of nutrients from regions of agricultural productivity to the mouths of the city inhabitants. The waste generated by these inhabitants—and the nutrient loads they contain—can have significant impacts on the nutrient dynamics and balances of surrounding biomes. Similarly, the burning of fossil fuels for energy can lead to a “fixing” of nitrogen—conversion of atmospheric gas in the atmosphere (N₂) to nitrogen oxides, which are then deposited on surrounding ecosystems, increasing the nitrogen availability in those systems. In addition, through these wastes and through industrial processes, areas surrounding cities may experience elevated levels of elements usually present only in trace amounts—such as mercury—or that are completely foreign—such as plastics or manufactured pesticides.

Significant portions of the available flow of surface water can be diverted to support activities in urban regions. In the western portions of the United States, for instance, there are protracted and complicated political and social negotiations to determine allocation of water to cities versus agriculture versus “the environment”—basically all other flows that had previously occurred in the absence of human occupation. (Indeed, the water politics of the American West is a prime example of how power and wealth can determine ecological impacts within the human habitat.) The decisions made about water allocation may affect not only the quantity of water delivered to systems but the seasonal or diurnal timing of water availability as well.

Finally, water quality can also be degraded, through increased loading of sediments, nutrients, or toxics or through elevated temperatures (the result of first using the water for cooling in power plants or factories before returning it to downstream flow).

The impacts of these resource diversions or perturbations will depend on how they are distributed across the landscape. Biological communities simultaneously experiencing elevated levels of toxins in the soils, a reduction in net primary productivity, elevated temperatures, and increased nitrogen deposition will likely suffer more profound consequences than one in which, perhaps, only net primary productivity has been altered. Multiple stresses can have impacts that are more severe than one might expect from a consideration of the impacts of single stresses.

C. Alteration of Disturbance Regimes

On small scales, human activities in urban environments lead to a general increase in the frequency of disturbance. Lawns are clipped weekly, vegetation in parks is plucked or trampled daily, and buildings are razed and raised over the years. This increased disturbance can lead to selection of “weedier,” colonizing species at the expense of late-successional, slower-
growing organisms. Increased frequency of disturbance can also favor species with shorter generation times. (Humans do also, however, tend longer-lived trees and perennials, thus potentially reducing the presence of weedier species—this influence is discussed below.)

On larger scales, in contrast, humans attempt to reduce the frequency and intensity of disturbances in and around urban settings. Thus, urban and suburban wooded or forested parks are frequently managed so as to minimize large fire events. Similarly, there are attempts to shelter coastal cities against the worst ravages of storms by building seawalls or planning the orientation of buildings and other structures. Floods are controlled, with varying success. This change in the disturbance regime can shift community composition relative to that which would emerge in the absence of the urban center—leading, for instance, to establishment of less fire-tolerant trees in urban forests.

Urban design can also change disturbance regimes in unintended ways. In Baltimore, the local climate has been changing—the precipitation from thunderstorms is occurring further from the city boundary, since storm fronts coming from the northwest are encountering warm air earlier—a result of the urban heat island. Similarly, the buildup of particulate matter in the lower atmosphere as a result of five continuous days of commuting has led to a greater frequency of weekend storms along the Northeastern seaboard—the particulate matter peaks by the end of the commute week and serves as a condensation medium for clouds.

D. Alteration of Species Composition

Humans intentionally and unintentionally manipulate the species compositions in urban settings. They do so indirectly through all of the other activities listed above, which can promote some species at the expense of others. In addition, however, there is

1. Preferential destruction or promotion of native species
2. Introduction of nonnative species

Humans have generally preferentially cultivated or destroyed species in urban settings based on aesthetics, values, or comfort. In hot environments, for instance, city residents might preserve and cultivate shade trees. There is also a clear aesthetic preference for flowering plants in most residential or commercial gardens. Residents may value preservation of native plants or feel strongly about providing habitat for songbirds. In contrast, most urban environments are depauperate in poisonous or prickly plants (the yards of cactus lovers aside), and there is in addition a general attempt to control "unappealing" organisms through the use of exterminators and traps. People's childhood homes, socioeconomic class, racial identification, and recreational preferences can all have a profound impact on those species that are deemed "desirable" versus "undesirable."

To understand these patterns of species portfolio management, ecologists will have to join with sociologists to examine how and why urban residents manipulate the neighborhood species complement the way they do. Such studies have been conducted in subsistence villages, where species manipulation is frequently a matter of survival, but fewer studies have examined the dynamics of aesthetic or cultural choices in high-population-density urban and suburban landscapes.

In addition to manipulating the indigenous suite of species, humans alter "invasion dynamics" through the intentional or unintentional introduction of nonnative species. On small, residential scales, aesthetic choices about landscaping can lead to significant introduction of nonnatives; similarly, the affinity for family pets can expose the small prey of a region to previously unknown predators. These aesthetics and affinities are very much influenced by culture and socioeconomic considerations. Immigrant communities may prefer vegetation reminiscent of their native homeland; families with children would prefer lawns to a more prickly groundcover; small lots are less likely to support large and complex vegetation structures than are large lots.

Finally, the ways in which humans manipulate species composition—or destroy or preserve habitat—depend on the value they place on biodiversity in general, or particular types of biodiversity in particular. Are habitats and species that reflect the prehuman landscape of a region prized most highly? Those that evoke memories of the childhood home? Are birds more highly valued than arthropods? Are rare species to be most highly valued, or certain combinations of species? These values will vary from person to person, town to town, country to country. They will play out differently at different levels of organization—from decisions made about species or habitat preservation at the federal level, to protection of native diversity at county levels, to recreation of preferred landscapes at the individual level.

VII. CITIES IN THE FUTURE

We cannot predict with certainty what cities will look like in the future. How will the economic structures
of cities change in the context of regional and global markets? To what extent will changes in economy affect our stewardship of the biosphere? Will development bring with it not only the means for further destruction of biodiversity but also the means for more effective protection? How will information flow and people move on a daily basis with the advent of new technologies and changes in the availability of renewable and nonrenewable resources? How will ideologies change to either promote or hinder the protection of the other species with whom we share the planet? Will the political and social forces that determine the dynamics of a city and the impacts on the environment evolve to encompass a more integrated understanding of the city as part of a broader ecosystem?

It is difficult to predict how such changes will affect human settlement patterns; but it is evident that urbanization is a clear and dominant trend that will continue to characterize our species and the biodiversity of Earth. Ecologists have traditionally treated human social systems as separable from ecological systems and have traditionally treated anthropogenic impacts on biodiversity as an exogenous, rather than endogenous, perturbing force. But as an increasing number of the Earth’s ecosystems come under the influence of these highly populated human habitats, and as an increasing number of species are influenced by the manipulations (either intentional or unintentional) of humans as they shape and respond to their surroundings, the scientific community will—predict—discover that understanding the future of biodiversity on the planet—not just in cities—will require treating Homo sapiens as an interacting member of ecological communities, and not as a species that stands apart from them. Cities are an ideal environment for examining humans as integral parts of ecological systems, and therefore serve as an ideal proving ground for the theories of how human activities will influence biodiversity patterns in the future.

Acknowledgments
The authors thank James Collins, Simon Levin, and Gary Wade for their comments on the manuscript. For Morgan Grove, research support on the Baltimore Ecosystem Study was provided by the Burlington Laboratory (4454) and Global Change Program, Northeastern Forest Research Station, USDA Forest Service, the National Science Foundation (NSF Grant #DEB-9714833), and the Environmental Protection Agency (EPA Grant #R-825792-01-0).

See Also the Following Articles
ECOLOGICAL FOOTPRINT, CONCEPT OF • ENERGY USE, HUMAN • HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW • KEYSTONE SPECIES • LAND-USE PATTERNS, HISTORIC

Bibliography
Glossary

autogenic species A species that creates habitat for other organisms; this habitat would not exist if the autogenic species was absent.

chemoautotrophic Primary production of organic matter using energy derived from chemical reactions rather than from photons.

dyke intrusion Upwelling magma, may reach the surface of the crust to flow out as lava and solidify as basalt.

end-member fluid In black smokers, the hot, chemically modified fluid that develops at depth in the crust and exits undiluted by seawater.

euphotic zone Sunlit upper layer of the ocean that can support photosynthesis.

heterotroph Organism that relies on organic material (other organisms, pieces of organisms, dissolved organic material) for its nutrition.

hydrothermal plumes Created by rising black smoker effluents that mix turbulently with the surrounding seawater, becoming dilute and eventually reaching neutral buoyancy (typically ~200 m above the black smoker source); once neutral, they spread laterally and are detectable by chemical and physical tracers for kilometers.

in the catalog of habitats on earth, chemo-
synthetic ecosystems at deep-sea hydrothermal vents are singularly decoupled from climatic variations and from anthropogenic activities. Furthermore, chemosyn-
thetic ecosystems are tightly coupled to the geophysical properties of tectonism and volcanism. as a conse-

mid-ocean ridge Crustal accretion boundaries between major tectonic plates that make up the ocean crust.

phylootypes Molecularly differentiated taxa.

plume prospecting Use of chemical and physical anomalies in hydrothermal plumes to locate hydro-
thermal vents.

spreading rate The rate at which tectonic plates move apart from each other.

symbiont One of a pair of coexisting, mutually supportive organisms; in vent invertebrate–bacteria symbioses, the "symbiont" colloquially refers to the bacterium associated with the host invertebrate tissue; endosymbiont—within the host invertebrate (intra- or extracellularly); epipsymbiont—on the outer surface of the host invertebrate.

tectonism Forces that disturb and dislocate the Earth's crust (earthquakes)

vicariant event A historical event that isolates species' populations and leads to speciation (e.g., the closing of the isthmus of Panama); usually considered in a geological context, but human activities also generate vicariant events.

I. Mid-ocean Ridges and the Hydrothermal Setting
II. Microbial Diversity
III. Invertebrate Diversity
quence, one can look to tectonic history and volcanic periodicity as fundamental predictors of the boundaries of biogeographic provinces and of species richness within provinces. Chemosynthetic systems support endemic faunas, are globally distributed, and are insular habitats. In such putatively simple systems, hydrographic and topographic controls on biodiversity and biogeography of microbial and metazoan communities may be much more readily resolved than in systems where climate and human activities obscure their role. Systematic study of diversity in chemosynthetic systems within the context of planetary-scale processes will lead to an increased understanding of fundamental controls on biodiversity and biogeography within oceans (on Earth and other planetary bodies) where propagules are waterborne and subject to dispersal in an open system.

I. MIDOCEAN RIDGES AND THE HYDROTHERMAL SETTING

Although deep-sea hydrothermal vents occur in a variety of geological settings, including isolated back-arc spreading centers (such as the one that lies behind the Mariana Trench) and seamounts (e.g., Loihi, the incipient Hawaiian island), most hydrothermal systems lie along the network of volcanic midocean ridges that girdle the globe. Midocean ridges are the boundaries between the tectonic plates that make up the ocean crust; as such, they are the locus of episodic volcanism and earthquake activity associated with magmatic dyke intrusions and seafloor spreading. Most of the midocean ridge (and thus its hydrothermal systems) lies well below the euphotic zone, at depths greater than 2000 m. Where ridges intersect continents (e.g., at the head of the Gulf of California) or shoal to volcanic islands (e.g., Iceland and the Azores), hydrothermal systems may be found at shallower depths and even intertidally.

The geological pavement of deep-sea vents is typically hard basalt. Local relief of the basalt varies from flat, featureless sheets to smooth lobate surfaces a few tens of centimeters high to rounded pillows of a meter or more relief. There are a small number of locales where sedimentation rate is in excess of volcanism, resulting in soft-sedimented bottoms with intruding basalt sills. The two best known sediment-hosted hydrothermal systems are in Guaymas Basin (Gulf of Mexico), where terrestrial runoff and pelagic production rates are high, and portions of the Gorda Ridge (off Oregon) and the northern Juan de Fuca Ridge (off Vancouver, Canada), where terrigenous inputs from the continental margin are significant.

Hot springs develop on midocean ridges where seawater can penetrate ~2 km into the porous basalt crust to reach hot rock. In this “reaction zone” at depth, seawater reacts with the basalt, losing all of its magnesium and oxygen and becoming enriched in metals and reduced compounds, including hydrogen sulfide. The resulting acidic, hot water is thermally buoyant and rises to the seafloor to exit as high-temperature (typically 350°C) fluids. As the vent fluid exits, it mixes with the cold, nearly neutral seawater and precipitates mineral sulfides of copper, iron, and zinc, forming substantial mineral chimneys known as black smokers. Where venting is persistent over time, chimneys coalesce to form larger, hydrodynamically and topographically complex sulfide edifices tens to hundreds of meters in diameter. Plumes emitted from black smokers rise ~200 m, until they become neutrally buoyant and spread out laterally. Hundredths of a degree deviations from ambient temperatures, elevated particulate concentrations, and modified chemistry allow water column geochemists to undertake “plume prospecting” by towing sensor packages at the depth of the neutrally buoyant plume. This prospecting method permits extensive mapping of ridge segments to yield integrated measures of hydrothermal output. Plume prospecting is also critical in locating new hydrothermal fields in unexplored ocean basins.

Not all of the vent fluid escapes as high-temperature fluid. Large volumes of warm, sulfide-laden water are emitted as “diffuse flow” vents where ambient seawater has mixed with the end-member fluid. It is this warm-water flow that sustains productive populations of free-living chemosynthetic, thermophilic microorganisms (up to 115°C) and dense invertebrate populations (typically <40°C). These diffuse flows may issue from porous surfaces of sulfide-mineral structures or directly from cracks and fissures in basalt lavas and can contribute an order of magnitude more flow than the focused output of black smokers at some sites. Because there are large gradients in chemistry and temperature within vent ecosystems, there is a large potential for diverse microbial and invertebrate types.

Hydrothermal vent fields, typically consisting of multiple discrete zones of focused and diffuse flows, are the basic unit of hydrothermal activity on a ridge axis. Vent fields vary in size from hundreds to thousands of square meters and represent islands of productive habitat within an otherwise relatively barren, hard-substrate environment.
The tectonic plates that make up the crust do not all move apart at the same rate. Spreading rate is a function of magma supply, with fastest spreading centers (100–170 mm year$^{-1}$) associated with well-developed magma chambers and supporting a higher spatial frequency of hydrothermal systems than slow-spreading systems (moving apart at 10–50 mm year$^{-1}$). As will be seen below, spreading rate may have a profound influence on diversity and biogeography of vent invertebrate communities because of its control on the spacing between vents. Other features vary with spreading rate, including the relative longevity of venting sites and bathymetric isolation of near-bottom waters. The two end members of the spreading rate continuum include the slow-spreading Mid-Atlantic Ridge, characterized by long-lived (thousands of years) vent systems and deep, basinalike valleys within which vents are located, and the fast-spreading East Pacific Rise, where vents are relatively short-lived (months to years to decades) and occur on topographic highs.

II. MICROBIAL DIVERSITY

A. Free-Living Microorganisms

Deep-sea hydrothermal vent ecosystems are celebrated as sites where chemosynthesis by microorganisms sustains primary production and a huge biomass of primary and secondary invertebrate consumers. The generation of biomass in the system is driven primarily by the microbial oxidation of sulfide, which yields biochemical energy for the fixation of CO$_2$. As much as the oxygen used in sulfide oxidation comes from seawater and ultimately from photosynthesis, the larger part of the vent ecosystem cannot be said to be totally independent of sunlight.

The potential metabolic diversity of vent microorganisms is large, with both aerobic and anaerobic isolates. Reduced sulfur substrates, iron, manganese, ammonia, methane, and hydrogen can all serve as electron donors, leading to sulfide, sulfur, and thiosulfate oxidation, iron and manganese oxidation, nitritation, methanogenesis, and hydrogen oxidation, denitrification, sulfur and sulfate reduction, and methanogenesis. While microbial isolates may be induced to be strictly autotrophic in the laboratory, mixotrophy may be the rule in the natural environment, with assimilation of simple organic compounds (heterotrophy) whenever they are available. Physiological diversity of microorganisms may also be great, with microbial types optimized for life in extreme temperatures (perhaps to $>115^\circ$C; hyperthermophiles), high temperatures (30–115$^\circ$C; hyperthermophiles and thermophiles), intermediate temperatures (10–30$^\circ$C; mesophiles), and cold temperatures (<10$^\circ$C; psychrophiles). Microorganisms can be pressure tolerant or barophilic (unable to survive low pressure). Microbial habitats include low-temperature diffuse-flow environments, buoyant plumes, interstices of sulfide structures, and surfaces of rocks and organisms. There is also growing evidence for an expansive, deep subsurface biosphere within the porous upper layer of the ocean crust, wherever temperatures are compatible with life.

Diversity of microbial communities at the level of species or phylotype is one of the outstanding issues in microbial ecology. Isolates cultured from deep-sea samples may only deliver a few percent of the actual diversity within the system. Current research combines traditional culture methods with molecular characterization of natural communities. Within vent systems, microbial communities may be dominated by a single phylotype, as in the case of bacteria associated with mineral sulfides at black smoker chimneys on the Mid-Atlantic Ridge. Dominance by a small number of phylotypes is also observed at the Loihi Seamount hydrothermal vents. The dominance of microbial communities by a small number of taxa is reminiscent of species-abundance characteristics observed in studies of diversity in marine invertebrate communities from “extreme” environments. It is, however, too soon to conclude that there is any single pattern of microbial diversity in the free-living microbial populations at hydrothermal vents. An even greater unknown is the degree to which vent microorganisms express biogeographic patterns. The current dogma is that free-living bacteria have unlimited dispersal capabilities in the open ocean because they can drift indefinitely in an inactive state. Hydrothermal vents may prove to be an ideal system in which to test this hypothesis with rigor at a global scale.

Biotechnology industries have embraced the vent ecosystem as a source of novel microbial organisms and enzymes that operate under extreme conditions. The prospect of thermostable or barophilic enzymes in particular has led to systematic assays of deep-sea microbial communities for industrial applications. Where microbial organisms live at mass- and thermophilic temperatures in organic-rich sediments, they are implicated in transformation and decomposition of freshly cracked organic material. Microbial hydrocarbon processors are obvious targets for extraction and design of enzyme systems by petroleum and environmental waste management.
There is physiological diversity in microbial endosymbions of vent taxa. Sulfide oxidizers predominate in vestimentiferan tubeworms, but there is at least one instance of a tubeworm that hosts a methanotrophic symbiont. Within bivalves, some symbionts use sulfide; others use the more oxidized and less toxic forms of sulfur, including thiosulfate or elemental sulfur. Several vent mussel species exhibit a dual symbiosis, harboring both methanotrophs and sulfur oxidizers. This diversity of symbiont capabilities within individuals increases metabolic flexibility and thus the range of environmental conditions under which the mussels may thrive.

Episymbiotic bacteria also occur in well-known associations with certain invertebrates. One of the best examples is the bacterial "fur coat" of the pompeii worm, *Alvinella pompejana*. This large polychaete worm lives in galleries on the sides of black smoker chimneys and experiences one of the steepest gradients of temperature along its body length of any invertebrate (40–50°C). The dorsal surface of *A. pompejana* is colonized by a morphologically diverse consortium of microorganisms, including large filamentous bacteria visible to the naked eye. The anatomical relationship between bacteria and worm is exquisite in its complexity, but the functional role of the members of the consortium remains unclear. Both trophic and detoxifying roles have been suggested. Episymbiotic, chemoautotrophic bacteria are also known from the chitinous cuticle of vent shrimp, *Rimicaris exoculata*. In this instance, the bacteria are thought to contribute to the diet of the shrimp; they are also likely to play some role in sulfide detoxification. Molecular characterization of the shrimp endosymbions shows them to be a single phylotype that matches the phylotype of the dominant free-living microbial species associated with the sulfide minerals on which the shrimp lives.

While there is much known and far more to be learned about microbial diversity at the interface between vent fluid and seawater, recent speculation about a vast, deep subsurface biosphere has fueled efforts to tap into the fluid-filled voids of the upper ocean crust. Efforts to understand microbial abundance, diversity, and function within such a biosphere may yield ancient lineages or otherwise novel species adapted to unusual conditions.

**III. INVERTEBRATE DIVERSITY**

It is the lush, gardenlike accumulations of invertebrate biota that make hydrothermal vent ecosystems a prized subject for television documentaries. Tubeworms often
star in the title role, with clams and mussels as supporting actors, but this view of a simple ecosystem is little more than a Hollywood gimmick for audiences with short attention spans. The idea that vent systems have low invertebrate diversity is both pervasive and potentially misleading and is in large part due to the paucity of quantitative studies of diversity at vents. There is also a widespread failure to scale diversity measures at vents to the truly minute global acreage of vent habitat compared to extensive areas of soft-sediment deep sea or terrestrial rain forest, where diversity is so celebrated. These cautionary notes aside, there is an a priori reason to expect that diversity at vents might be low: Sulfide is a potent toxin, poisoning a major enzyme system of cellular respiration (cytochrome c oxidase) and bringing to a halt the aerobic production of ATP. Multicellular organisms that live at vents all must have some means of avoiding sulfide toxicity. This requirement may be a fundamental determinant of higher order invertebrate diversity at vents—only those groups that have efficient sulfide detoxification systems can occur where sulfide is present.

If we consider diversity at the phyletic level, vents are depauperate in several marine phyla normally found in the deep-sea benthos, including Porifera, Echinodermata, and the lophophorates (Phoronida, Bryozoa, Brachiopoda). The scarcity of sponges (Porifera) and echinoderms is attributed to the lack of excretory and blood vascular systems in these groups and a consequent inability to tolerate elevated levels of sulfide or other toxic compounds associated with hydrothermal vents; where these phyla are represented at vents, the animals are associated with the periphery of a vent field or other circumstances where sulfide levels are likely to be low (as in a waning vent site). The colonial motif and asexual reproduction are also apparently unsuccessful at vents. Colonial coelenterates such as gorgonians, antipatharians, and hydroids are common animals of the deep sea but so far are virtually unknown at vents. An exception is a colonial siphonophore (dis-tantly related to the Portuguese Man o’ War) that is often found in large numbers at dying vents.

A number of hypotheses have been put forth to explain the absence of colonial organisms at vents, including intense competition and the ephemerality of vents, which may preclude any advantage of vegetative reproduction. Colonial organisms would also be at a disadvantage by virtue of the connectivity between individuals, so exposure of one individual within the colony to sulfide jeopardizes the aerobic metabolism of the entire colony. Encrusting colonial species are difficult to posit, because the vent effluent is buoyant and would tend to lift off recruiting polyps. Further, if established, an encrusting species could cap the flow, resulting in either diversion of the flow to less resistant paths or pooling of the flow with concomitant, life-threatening elevation of sulfide concentrations and temperature. While strategies to cope with these kinds of physical and chemical challenges to colonial organisms are not inconceivable, they have so far not been observed.

A. Origins of Vent Invertebrate Taxa

The majority of species that occur at vents (~90%) are known only from vents. This high degree of endemicity is difficult to prove, given the lack of effective means of (and interest in) collecting organisms from nonvent, deep-sea basaltis. Nevertheless, it is a distinctive fauna. Where does it come from? There is no single source. Some species belong to nonvent deep-sea genera (e.g., the squat lobster, Munidopsis lentigo), others have closest relatives among the shallow-water invertebrates (e.g., the polychaete Ophryotrocha), Several species, notably symbiont-containing invertebrates (tubeworms, clams, and mussels), have closest alliances to species known from other chemosynthetic ecosystems, such as brine and hydrocarbon seeps and whale skeletons. Some vent families (e.g., scale worms in the polychaete family Polynoidae) are broadly represented in the marine environment. Still others appear to be specialized taxa, known only from hydrothermal systems—the polychaete family Alvinellidae is a good example. Finally, some species appear to be most closely related to ancient Paleozoic or Mesozoic taxa, relict lineages that have found refuge in hydrothermal systems. These include several stalked barnacle species and an archaeogastropod limpet. Some taxa routinely found at vents have been hugely successful, undergoing substantial radiation. These include the siphonostome copepods, the archaeogastropod limpets, and the polynoid polychaetes, within each of these groups there are now dozens of described species.

B. Biogeography

When vents were first discovered in the eastern Pacific, there was much speculation about how many vents there might be, where they were, and whether the fauna would be cosmopolitan. We know now that there are likely to be hundreds, if not thousands, of hydrothermal vent fields along the midocean ridge system and that they can be found in every ocean basin. Furthermore, we know that the fauna is not cosmopolitan. Distinct biogeographic provinces occur, defined in part by the
tectonic fabric and history of the ocean plates. Along a continuous ridge system such as the East Pacific Rise, which stretches from the head of the Gulf of California down to high latitudes in the Southern Hemisphere, the fauna of the hydrothermal vents undergoes subtle changes. The same species of tubeworms, mussels, and clams generally dominate the benthos (mussels are absent from the northernmost locales), but the details of their relative abundance change from site to site. These dominant taxa (especially the tubeworms and mussels) are autogenic bioengineers, creating complex three-dimensional habitats that become occupied by countless individuals of crinoids, molluscs, and polychaete species. We know little about how the ranges of the hundreds of smaller invertebrate species that live in association with these autogenic taxa vary along the length of the East Pacific Rise or of how these ranges relate to the life history characteristics of each species.

At its northern limit in the Gulf of California, the East Pacific Rise goes terrestrial as a strike-slip fault (the San Andreas), reemerging in the submarine environment off northern California as the Mendocino fracture zone. From there, the spreading ridge axis between the Pacific and North American plates heads northward as the Gorda, Juan de Fuca, and Explorer Ridge systems. The vent fauna of these northern ridges is distinct from that of the East Pacific Rise at the species level, but there are alliances at the generic and familial levels. Verena Tunnicliffe (University of Victoria) notes that at one point in the geological history of these ocean plates, the East Pacific Rise and northern ridges were one continuous system. Overriding by the North American Plate was a vicariant event, bisecting the ridge system and resulting in subsequent development of the vent faunas in isolation and of two distinct, formerly related, biogeographic provinces. Other biogeographic provinces are known: Deep-sea hydrothermal vents at back-arc spreading centers associated with western Pacific microplates support a distinctive invertebrate fauna at the species level, although at higher taxonomic levels, affiliation to the two eastern Pacific biogeographic provinces is patent. This affiliation may be related to the direct link that existed between the eastern and western Pacific via the now extinct Kula Ridge more than 40 million years ago. Deep-sea vents along the Mid-Atlantic Ridge support an even more disparate group of species and genera. Based on biogeographic characterization of these four major geographic regions, we expect that ridge systems in unexplored ocean basins (the Arctic and the Indian Oceans in particular) will yield hundreds of new species. Exploration and discovery remain hallmarks of the study of diversity and biogeography at vents.

C. Gene Flow and Genetic Diversity

Deep-sea hydrothermal vents on midocean ridges are distributed as linear, insular, and ephemeral arrays of endemic species. To persist in the face of certain local extinction, survival of a species in any region is only possible where propagules can disperse to other vents. An understanding of gene flow along these arrays is essential if we are to understand species distributions and speciation processes at vents. Two basic models of gene flow, borrowed from island biogeography, have been applied to genetic data from vent populations. In the “stepping-stone” model, gene exchange occurs locally, between neighboring populations, and genetic exchange decreases with increasing distance. In the “island” model, there is thorough mixing of gene pools over a regional scale. Vent taxa conform to no single model of gene flow. Tubeworms on the East Pacific Rise, for instance, appear to follow a stepping-stone model, while mussels on the same ridge follow the island model. Still other species display a ridge-based isolation model, where topography of the ridge system (e.g., deep east–west fracture zones that offset the shallower north–south-trending ridge) creates a genetic filter. Studies of genetic diversity in vent systems by Robert Vrijenhoek (Monterey Bay Aquarium Research Institute) and his students highlight the significance of the number of populations in sustaining genetic diversity: vent species with abundant populations have higher genetic diversity than species with fewer populations. Genetic diversity in vent organisms (measured as percentage of polymorphic loci within a population) is also correlated with the order of establishment of species at nascent vents, with early colonists having double the genetic diversity of later colonizing species.

D. Biodiversity

For two decades, basic community descriptors of species richness and diversity eluded vent ecologists. Sampling was qualitative, with uneven efforts, and typically resulted in lists of species for vent sites without regard to distributions within specific microhabitats. An exception is the work of Fred Grassle (Rutgers University), who compared diversity of soft-sediment vent fauna with infaunal diversity in the surrounding soft-sediment, nonvent environment. His quantitative data demonstrated that diversity in vent muds is low compared to the relatively high species diversity of the “normal” deep sea. Further, at vents, one or two species account for 70–90% of the infaunal individuals, whereas in the nonvent deep sea, abundances are more evenly distributed among species, with the most com-
mon species making up less than 20% of the total. Similar results were found in studies by others of soft-sediment meiofauna. But soft sediments are the exception rather than the rule for deep-sea hydrothermal vents.

What about hard-substrate biodiversity at vents? Species richness measured at a site on the Mid-Atlantic Ridge based on replicate samples within a single vent habitat (mussel bed) and species–effort curves has been compared to values obtained using a comparable sampling effort at an intertidal mussel bed in south-central Alaska. The intertidal, photosynthetic site supported more than twice as many species than the deep-sea chemosynthetic site, consistent with the idea that diversity at vents is low. But diversity in intertidal mussel beds is well-known to be geographically variable; some intertidal mussel beds (e.g., along the coast of Japan) are reported to have species richness values comparable to what is found at the Mid-Atlantic Ridge vent site.

Vents allow us to study patterns of regional diversity over different scales of spatial frequency of habitat and in the absence of confounding climatic and anthropogenic effects. The Mid-Atlantic Ridge is a slow-spreading ridge system, with vents spaced relatively far apart (hundreds of kilometers). Diversity is predicted to be higher on faster spreading centers where vents are more closely spaced. With the same sampling methods as used at the Mid-Atlantic Ridge vent site, diversity has been measured at two vent fields on the ultrafast spreading southern East Pacific Rise. Diversity at these vents is comparable to that of the Alaskan intertidal mussel bed. Thus, whether one considers diversity at vents to be high or low depends only on which end members one chooses to compare.

Of theoretical importance is the fact that patterns in diversity at midocean ridge hydrothermal vents may be controlled by fundamental properties associated with plate tectonics and volcanism that determine the spatial frequency of vent habitat. Where vents are closely spaced, the likelihood of extinction is relatively low, and species tend to accumulate. Where vents are far apart, a premium is placed on effective dispersal, extinction becomes much more likely, and allopatric speciation is enhanced. This hypothesis leads to the prediction that slow-spreading systems, with distant vents, support a greater number of biogeographic provinces, each with lower species diversity than fast-spreading systems with closely spaced vents.

Area–spreading rate relationships have also been proposed as controls on species diversity at vents. In general, the well-known area–species relationship is driven by increasing numbers of habitats with increasing area. While presumably relatively unimportant when comparing diversity within habitat types (e.g., mussel beds), the increasing vent area (and, by inference, habitat diversity) with spreading rate should contribute to increasing species diversity within vent fields.

Quantitative assessment of species diversity within all representative microhabitats of vent fields and measures of areal extent of fields along ridges of different spreading rates is not a tractable approach at this time.

See Also the Following Articles
- ARCHAEA, ORIGIN OF
- • HIGH-TEMPERATURE ECOSYSTEMS
- MARINE ECOSYSTEMS
- MICROBIAL BIODIVERSITY
- THERMOPHILES, ORIGIN OF

Bibliography


VERTEBRATES, OVERVIEW

Carl Gans* and Christopher J. Bell†
*Department of Integrative Biology, University of Texas at Austin and †Department of Geological Sciences, University of Texas at Austin

I. Introduction
II. General Vertebrate Characteristics
III. Early Chordate and Vertebrate History
IV. Vertebrate Classification
V. Definitions and Diagnoses of Major Chordate Groups

GLOSSARY

chordate A member of the group Chordata. The Chordata includes the most recent common ancestor of tunicates and cephalochordates and all of that ancestor's descendants. Tunicates, lancelets, hagfishes, and vertebrates are all chordates.

ectoderm An embryonic tissue that provides the future outside layer of the animal.

ectothermy A method of body temperature control in which the animal utilizes external sources for gaining and giving up heat, thus achieving temperature control without affecting metabolic rate.

endothymery A method of body temperature control in which the animal modifies its metabolic rate to achieve the desired body temperature.

neural crest An embryonic tissue intermediate between neurectoderm and ectoderm, with cells migrating widely to their final destination. This tissue gives rise to anterior skeletal elements, many portions of the future head and pharynx, and all pigment cells. Sometimes also referred to as mesectoderm.

neurectoderm An embryonic tissue that gives rise to the central tube of the nervous system.

notochord A stiff, flexible, longitudinal rod running along the middorsal portion of the chordate body. It is situated dorsal to the coelom and ventral to the central tube of the nervous system.

pharynx The anterior portion of the alimentary canal, characterized by lateral buds that provide skeletal support for the gill region.

tuberculum interglenoidenum An anterior projection of the first (cervical) vertebra in salamanders. The tuberculum interglenoidenum bears articular facets that insert into the foramen magnum of the skull and provide additional articulation points between the skull and the vertebral column.

VERTEBRATES INCLUDE ALL the fishes, amphibians, reptiles, birds, and mammals. These animals are united in a more inclusive group, the Chordata, that includes the closest living relatives of vertebrates, the hagfishes, lancelets, and tunicates. There are approximately 54,450 known species of chordates, over 51,000 of which are classified as members of Vertebrata. Nearly half (approximately 24,000) of the known species of vertebrates are members of a single group of ray-finned fishes, the actinopterygians. The vertebrates are found on all major land masses and in all major oceans and seas on Earth.
I. INTRODUCTION

The vertebrates include most of the major groups of animals that humans encounter and interact with on Earth, and in common parlance, the term “vertebrate” is often equivalent to “animal.” Vertebrates include all the fishes, amphibians, reptiles, birds, and mammals; they are found on every major landmass and in all of the world’s oceans and major seas. The fossil record of the group is excellent and extends back to the late Cambrian period of the early Paleozoic, more than 500 million years ago. The vertebrates are members of a nested series of more inclusive groups, extending all the way down the tree of life to a common ancestor they share with bacteria, the most primitive life forms known. In order to properly understand the vertebrates, it is necessary not only to briefly discuss the characteristic features unique to the group but also to evaluate the features they share with their closest living relatives, the hagfishes, lancelets, and tunicates. These organisms plus vertebrates constitute the group Chordata. In the pages that follow, we summarize some generalized characteristics of chordates and vertebrates, examine the early history of these groups as it is revealed from the study of the fossil record, and provide explicit definitions and diagnoses for all of the major chordate groups.

II. GENERAL VERTEBRATE CHARACTERISTICS

The diversity among vertebrate groups makes it possible to utilize functional analyses and paleontological data to reconstruct a phylogenetic sequence of vertebrates and to place the currently surviving (extant) vertebrate groups into a historical sequence. This sequence permits the development of explanatory schemes for some conditions observed in humans. Consequently, the study of the comparative morphology and physiology of vertebrates has proved important for the understanding of the human condition. These studies in turn led to the development of curricular variants for students training themselves for careers in human medicine. Examples are approaches to applied physiology and morphological of the precursors’ conditions. Such instructional variants may be presented as the sequential development of the history of a topic such as gas exchange, digestion, or embryology. Alternatively, the study may involve a series of steps, each explaining the condition in one of the subgroups of animals representing putative precursors for the vertebrate condition (i.e., basal chordates or echinoderms), or historical stages in the development of groups of vertebrates (i.e., fishes, reptiles, and mammals).

Although a few arthropods and mollusks achieve large, indeed giant, sizes, most of the largest species, whether terrestrial, aquatic, or aerial, are vertebrates, and this latter group includes the largest animals known. However, some vertebrates are small, almost tiny, and this poses interesting questions of how they perform the basic functions of reproduction, locomotion, respiration, food processing, and waste discharge. Vertebrate size reduction involves two strategies: (1) maintenance of the adult pattern at a reduced body size and (2) the acquisition of metamorphosis, meaning that the life processes, and the structures that facilitate them, change markedly at one or more stages and sizes of development.

Vertebrates may be visualized as elongate, tubular creatures possessed of a central tubule, this being a gut or alimentary canal. The anterior end of the gut forms a mouth and buccal structures, whereas the posterior end is modified for the storage and ultimate elimination of waste. The central tubule is placed in a lined cavity, the coelom, that also houses various diverticulae of the gut that support aspects of such visceral functions as gas exchange (i.e., gills and lungs, heart, and vessels), digestion (pancreas and liver), fluid balance (kidneys), and reproduction (gonads). The entire animal is coated by a multilayered epidermis that protects the internal tissues from mechanical stresses and can serve as a physiological barrier. The space between coelom and epidermal layer is occupied by the somatic musculoskeletal system that facilitates locomotion and supports and maintains the shape of the coelom and hence of the entire animal.

The musculoskeletal system involves the cartilages and bones of the internal skeleton and the striated somatic musculature. Externally, the system underlies the epidermis, this portion forming a dermis having skeletal properties. Two additional median and tubular structures complete this architectural scheme. The first tube, the notochord, lies just dorsal to the gut and is filled with vacuolated cells, their turgor pressure turning the notochord into a stiffened skeletal rod. The second, still more dorsal, tube consists of neuronal cells that involve sensory function and transmission of signals to the musculature.

The entire animal may have various appendages such as fins or limbs, each having characteristic internal skeletal supports and motile capacity developed by the contraction of the attaching striated musculature, which is subdivided into myotomes, paired blocks that parallel...
the notochord. The notochord provides a capacity for resisting longitudinal stresses, while the appendages involve support of the animal above the substrate and the transmission of laterally directed force patterns.

Whereas this general pattern seems to characterize vertebrate structure, the group shows major curiosities. The various groups of nonvertebrate creatures can generally be diagnosed by a relatively few characteristics. In contrast, there are no fewer than 25 new characteristics shared by vertebrates (Table I). These involve multiple and major physiological systems, for instance the multiple senses, the head skeleton, the vertebral column, and the digestive system. This clear diversity of vertebrate specializations has long raised questions about the adaptive reasons to explain why these comparable characteristics are modified in vertebrates and why and how a multiplicity of such characteristics might be associated and consequently subject to possibly simultaneous selection. Actually, the answer to this conundrum has been available for some 30 years. Rather than looking at the functions of the many “vertebrate characteristics,” namely, aspects seen in the adult condition, it is more informative to base comparisons on the stages of vertebrate embryology or development. The traditional rules that postulated firm conditions, such as that the ectoderm formed all the epidermis and the mesoderm all of the striated musculature, soon broke down. In particular, the neural crest, an embryonic tissue intermediate between the ectoderm and the neuroectoderm, proved to be the source of many diverse tissues. Similarly, the neurogenic epidermal placodes formed a set of embryonic tissues that provided a source for many shared vertebrate characteristics (Table I). The most interesting aspect was the discovery that many otherwise indistinguishable adult tissues could be derived from different embryonic sources, depending on their position in the embryo. Thus cephalic striated muscle has a histological pattern equivalent to that of the striated muscle of the trunk but is of neural crest rather than mesodermal origin. Similarly, the bones of the skull are derived from neural crest and the more posterior and serially equivalent vertebral bones are of (sclerodermal) mesodermal origin.

Consequently, the vertebrate characteristics either are derived from neural crest of neurogenic epidermal placodes or are induced by these tissues. In contrast to searching for 25 functional issues, the developmental approach makes it possible to recognize two or three associated aspects and note that these provided the cause of the vertebrate condition.

### III. EARLY CHORDATE AND VERTEBRATE HISTORY

Vertebrates are members of a more inclusive group of animals, the chordates. The history of both chordates and vertebrates extends back into the Cambrian period in the early Paleozoic. Although the fossil record of early chordates is generally quite poor, several fossil localities have produced fossils that are likely to repre-

<table>
<thead>
<tr>
<th>Derived Characters of Vertebrata and Their Embryonic Origin*</th>
<th>Embryonic origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nervous systems</td>
<td></td>
</tr>
<tr>
<td>Sensory nerves with ganglia, cranial</td>
<td>NC, P</td>
</tr>
<tr>
<td>Sensory nerves with ganglia, trunk</td>
<td>NC</td>
</tr>
<tr>
<td>Peripheral motor ganglia</td>
<td>NC</td>
</tr>
<tr>
<td>Second- and higher-order motor neurons</td>
<td>NC</td>
</tr>
<tr>
<td>Forebrain</td>
<td>NC?</td>
</tr>
<tr>
<td>Chromatophores</td>
<td>NC</td>
</tr>
<tr>
<td>Paired special sense organs</td>
<td></td>
</tr>
<tr>
<td>Nose</td>
<td>P</td>
</tr>
<tr>
<td>Eyes (accessory organs)</td>
<td>NC?, (P)</td>
</tr>
<tr>
<td>Ears</td>
<td>P</td>
</tr>
<tr>
<td>Lateral-line mechanoreceptors</td>
<td>P</td>
</tr>
<tr>
<td>Lateral-line electroreceptors</td>
<td>P</td>
</tr>
<tr>
<td>Gustatory organs</td>
<td>NC?, P</td>
</tr>
<tr>
<td>Pharyngeal and alimentary modifications</td>
<td></td>
</tr>
<tr>
<td>Cartilaginous bars</td>
<td>NC</td>
</tr>
<tr>
<td>Branchiomeric muscle</td>
<td>ME</td>
</tr>
<tr>
<td>Smooth muscle of gut</td>
<td>MH</td>
</tr>
<tr>
<td>Calcitonin cells</td>
<td>NC</td>
</tr>
<tr>
<td>Chromaffin cells, adrenal cortex</td>
<td>NC</td>
</tr>
<tr>
<td>Circulatory system</td>
<td></td>
</tr>
<tr>
<td>Giff capillaries, endothelium</td>
<td>ME</td>
</tr>
<tr>
<td>Major vessels, trunk</td>
<td>M</td>
</tr>
<tr>
<td>Wall of aortic arch</td>
<td>NC</td>
</tr>
<tr>
<td>Muscular heart</td>
<td>MH</td>
</tr>
<tr>
<td>Skeletal system</td>
<td></td>
</tr>
<tr>
<td>Anterior neuromuscle and sensory capsules</td>
<td>NC</td>
</tr>
<tr>
<td>Cephalic armor and derivatives</td>
<td>NC</td>
</tr>
<tr>
<td>Armor of trunk</td>
<td>NC?, ME3</td>
</tr>
</tbody>
</table>

* M, mesoderm; ME, mesoderm, epimere; MH, mesoderm, hypomere; NC, neural crest; NC?, intermediate between neuroectoderm and neural crest; P, placodes; (P), placodes provide larval and peripheral component, but not sensory tissues. Origin of the various trunk armors is uncertain. Table modified from Gans (1993).
sent the remains of early chordate animals. Controversy over the interpretation of anatomical features of these animals is in part due to the fact that the important characteristics by which we recognize chordates are soft anatomical features that are not easily fossilized. Exquisite preservation of soft-bodied organisms in the Burgess Shale in Canada (middle Cambrian, approximately 530 million years ago) resulted in the discovery of several specimens of a chordate-like animal named Pikaia gracilans. These important fossils appear to display major chordate structural features, including myotomes (segmented muscle blocks) and a stiff, longitudinal rod running along the middorsal portion of the body, the notochord. Some controversy regarding the precise phylogenetic position of Pikaia persists due to a lack of adequate description of these fossils in the primary literature. Recent discoveries in early Cambrian deposits in China stimulated renewed interest in the origin and diversification of chordates. Yunnanozoon lividum was described as an early representative of the cephalochordates (Chen et al., 1995; Chen and Li, 1997), and Haikouela lanceolata was recently identified as a potential crustate from the early Cambrian (Chen et al., 1999) but these interpretations are controversial (Shu et al., 1996). The earliest fossils attributed to the Vertebrata are enigmatic bits of bone from the late Cambrian described under the name Anatolepis hennegyi. The vertebrate affinity of these fossils was challenged on various grounds, but recent analyses showed they contain dentine, a form of the calcium phosphate mineral apatite (the major mineral component of bone) that is found only in vertebrate skeletal tissues (Smith et al., 1996).

IV. VERTEBRATE CLASSIFICATION

The classification of chordates has a long and complicated history. Traditionally, the vertebrates were divided into fishes, which occupied aquatic environments, and the tetrapods, which mainly occupied terrestrial habitats. At present, there are two major classification systems in use. The Linnaean classification system was first published by Linnaeus about 1758 and organizes major divisions of life into kingdoms (Plantae, Animalia, Fungi, etc.), with major body plans organized into phyla (singular: phylum). The basic ranks descend from phylum to class, order, family, genus, and species. This classification system is familiar to most teachers and students and remains the dominant classification taught to elementary and high school students in the United States (Fig. 1). In the past several decades, a nonranked classification system was described by Hennegyi, a European entomologist. This classification system, termed “phylogenetic systematics” or “cladistics,” organizes taxa into nested sets of monophyletic groups based on common ancestry (a monophyletic group contains an ancestor and all of that ancestor’s descendants). A classification system based on natural groupings of taxa (based on ancestor–descendant relationships) is thus achieved. Phylogenetic systematics is rapidly becoming an important alternative classification.
system and is increasingly being utilized in biological sciences (including paleontology). Although we provide ancestry-based definitions for all the major chordate groups discussed below, we include three figures to illustrate the differences in classification systems now in use. Figure 1 illustrates a version of a Linnaean classification system, in which taxa are placed in equivalent ranks. Figures 2 and 3 are graphic representations of cladistic classifications, showing the hierarchical arrangement of nested taxa (Fig. 2 shows the relationships of the major chordate groups, and Fig. 3 shows the relationships of the major groups nested within the taxon Tetrapoda).

V. DEFINITIONS AND DIAGNOSES OF MAJOR CHORDATE GROUPS

The major chordate groups discussed below are defined in terms of their ancestry; in other words, membership in a given taxon is dependent upon shared ancestry. Taxa are diagnosed by a list of characteristics hypothesized to have first appeared in the common ancestor of the members of the group; this assemblage of characters can be used to identify a taxon (see Rowe, 1987, for further discussion).

A. Chordata

Chordata is defined as the most recent common ancestor of tunicates and cephalochordates, and all of that ancestor’s descendants. Chordates share a common, generalized body plan and can be diagnosed by four features shared by members of all the major groups: a notochord, pharyngeal slits, a hollow dorsal nerve tube, and a postanal tail (an extension of the body posterior to the anus). The notochord is a stiff, flexible rod running dorsal to the coelom along the length of the body beneath the central nervous system. In many vertebrates, the notochord is almost completely replaced by blocks of bone forming vertebrae. Pharyngeal slits are lateral openings from the pharynx (an organ situated behind
FIGURE 3. A graphic representation of a cladistic classification of tetrapods. The major groups are arranged in a nested hierarchy, and taxon names are defined based on shared ancestry. See text for definitions and diagnoses.

the mouth that constitutes part of the digestive tract). These features are found in all chordates at some point during their lifetime, but they are not retained as fully functional units throughout life in all chordates.

The tunicates (urochordates) are a group of basal chordates represented by approximately 3000 species. They are entirely marine. Some tunicates spend their entire life as pelagic organisms, floating in the water column. One group of tunicates, the sea squirts, undergo a dramatic metamorphosis from a planktonic larval form to a sessile adult (attached to a substrate); a notochord is present in the tail of the larval stage but is resorbed in the adult during metamorphosis. The pharynx enlarges in the adult and expands to form a barrel-shaped branchial basket through which a constant flow of water passes as the animal filter-feeds.

B. Cephalochordata
Cephalochordata is defined as the most recent common ancestor of lancelets and Craniata, and all of that ancestor's descendants. Cephalochordates are diagnosed by the presence of segmented muscle blocks. The lancelets are marine organisms that live in shallow, tropical and temperate seas surrounding all the continental land masses except Antarctica. Two major groups are recognized and are classified under the names Branchiostoma (23 species) and Epigonichthyes (7 species; Poss and Boschung, 1996). Lancelets lack a well-differentiated head region. A rudimentary brain is present, but significant regions of the brain found in vertebrates appear to be absent. A recent volume addressing many aspects of lancelet biology (Gans et al., 1996) provides an important foundation for future research on this group.

C. Craniata
Craniata is defined as the most recent common ancestor of hagfishes and Vertebrata, and all of that ancestor's descendants. Craniates are diagnosed by the presence of a differentiated head, with well-developed paired sense organs designed to act as distance receptors.
are 43 described species of hagfishes; all are marine organisms, and the group has a worldwide distribution. Hagfishes are scavengers, feeding on dead and moribund fishes, worms, mollusks, and crustaceans. They lack recognizable vertebral structures but have weakly developed, irregularly shaped plates that are positioned almost randomly along the nerve chord in the tail region. Their eyesight is poor, but they have well-developed senses of smell and touch. Reproductive biology is poorly understood, only a few embryos are known. Hagfishes were not known as fossils until 1991. The first reported fossil was found in Pennsylvanian age rocks in Illinois (Bardack, 1991). It differs from living forms in at least a few features (better developed eyes, different position of the gills, and some aspects of the feeding apparatus).

D. Vertebrata

Vertebrata is defined as the most recent common ancestor of lampreys and Gnathostomata, and all of that ancestor's descendants. As noted above, Vertebrata is diagnosed by a suite of characters (see Table I). In all vertebrates there is at least some development of vertebral elements. The basal members of Vertebrata, the lampreys, have only rudimentary development of these structures. Lamprey vertebral elements consist of tiny distinct cartilages, situated above the dorsal nerve cord. The homologies of these elements with vertebral units in other vertebrates are not known. Many adult lampreys are parasites on other fish, and sometimes act as scavengers. They have a circular mouth with a rasping tongue; the mouth is used like a suction cup to attach the lamprey to a fish, and the tongue is used to open a hole in the body cavity of the fish, providing access to the body fluids on which the lampreys feed. Some lampreys spend significant portions of their adult lives in marine habitats, but all lampreys return to freshwater to breed. There are approximately 32 living species of lampreys.

Many earlier classifications placed lampreys, hagfishes, and a diverse suite of extinct jawless fishes into the class Agnatha ("without jaws"). A great diversity of jawless marine vertebrate groups populated Ordovician, Silurian, and Devonian seas. One enigmatic group, the conodonts, were first discovered in 1856 and subsequently were found to be remarkably abundant in marine rocks dating from the late Cambrian through the Triassic. For over a century, conodonts were known only from abundant, tiny, phosphatic, toothlike fossils that showed a remarkable diversity in form. These fossils were, and remain, important tools for establishing correlation and relative chronology of Paleozoic and early Mesozoic rocks. The systematic position of conodonts was debated for decades, and various authorities suggested affinities with many different animal groups. In 1983, a conodont fossil with soft-body preservation was reported for the first time (Briggs et al., 1983). Subsequent finds that preserve greater anatomical detail revealed several features that can be interpreted as a notochord, a dorsal nerve cord, myotomes, and a distinct head region with very large eyes. These features suggest that conodonts may well have been chordates, or possibly vertebrates, but the relationships of conodonts and other jawless fishes are still quite controversial (for reviews, see Benton, 1997; Zimmer, 2000).

E. Gnathostomata

Gnathostomata is defined as the most recent common ancestor of Chondrichthyes and Osteichthyes, and all of that ancestor's descendants. Gnathostomata is diagnosed by the presence of jaws, an enlarged forebrain, and a skeleton in the paired pectoral and pelvic appendages. The evolution of jaws provided new opportunities for vertebrates by opening the door to truly predatory habits, with new food-handling capabilities and a broader diversity in potential diet. Among living vertebrates the chondrichthians constitute the basal members of Gnathostomata. There are approximately 850 species of living chondrichthyans, including the sharks, skates, rays, and chimaeras. The majority are marine, but a few species inhabit freshwater environments. There is no bony skeleton; the skeleton is predominately cartilaginous, with only minor calcification occurring throughout. Males have special structures called claspers that help facilitate internal fertilization. The sharks first appear in the fossil record in the late Devonian, chimaeras appear in the Carboniferous, and rays appear in the Jurassic. The whale shark (Rhincodon typus) is the largest living fish, achieving lengths of up to 13 meters.

F. Osteichthyes

Osteichthyes is defined as the most recent common ancestor of Actinopterygii and Sarcopterygii, and all of that ancestor's descendants. Osteichthyes is diagnosed by the presence of lungs (see below) and by most of the bony skeleton being formed from cartilaginous precursors. Many bony fishes have an elongated sac, the gas bladder, situated dorsal to the digestive tract. Gas bladders can serve either a buoyancy control function (swim bladder) or a gas exchange function (respiratory
gas bladder). In the latter condition, they are highly vascularized and serve as a supplementary reserve for oxygen to be used in respiration. Although gas bladders and lungs differ in their position relative to the digestive tract (lungs are ventral to the digestive tract) and lungs tend to be paired while gas bladders are single, their developmental origins and gross anatomical features are quite similar. These similarities provide some evidence that the two structures are homologous. Actinopterygii includes a remarkably diverse assemblage of fishes, including more than half of the living species of known vertebrates; over 24,000 valid species are known (see Eschmeyer, 1998). The skeleton of actinopterygians is extensively ossified and the internal pectoral skeleton has broad bony plates that facilitate support of fin rays.

G. Sarcopterygii

Sarcopterygii is defined as the most recent common ancestor of Actinistia and Choanata, and all of that ancestor's descendants. Sarcopterygii is diagnosed by the presence of fleshy, lobed appendages, with differentiated proximal limb bones in the pectoral and pelvic fins, and by the presence of a ring of four or more bony plates in the eye (the sclerotic, or scleral, ring). Actinistia includes at least one, and possibly two, species of living coelacanths. These fishes first appear in the fossil record in the Devonian and make their last appearance in the fossil record near the end of the Cretaceous, approximately 70 million years ago. The discovery of a extant species of coelacanth off the coast of South Africa in 1938 was therefore quite surprising. The South American and African lungfishes can survive in stagnant pools by coming to the surface to gulp air. The Australian lungfish is the least derived of the three; it cannot live out of water but can survive on dry land for a short time. The southern continents of Africa, Australia, and South America. The Australian lungfish is the least derived of the three; it cannot live out of water but can survive in stagnant pools by coming to the surface to gulp air. The South American and African lungfishes can survive out of water for relatively long periods of time.

I. Tetrapoda

Tetrapoda is defined as the most recent common ancestor of Amphibia and Amniota, and all of that ancestor's descendants. Tetrapoda is diagnosed by the presence of differentiated fingers and toes and by a bony joint formed between the occipital condyles of the skull and the anterior vertebral elements. Extant amphibians share three morphological characteristics: (1) loss of several skull bones found in extinct amphibian groups; (2) short or absent ribs; (3) pedicellate dentition (an upper tooth crown resting on a lower pedicle). The relationships among the three major extant amphibian groups are not well supported, but combined analyses of morphological and molecular data favor a sister taxon relationship between Anura (frogs and toads) and Caudata (salamanders). Batrachia is defined as the most recent common ancestor of Anura and Caudata, and all of its descendants. In this arrangement, Batrachia and Gymnophiona (the caecilians) are sister taxa (Fig. 1; see also Cannatella and Hillis, 1993).

There are approximately 165 species of living caecilians. All extant caecilians are limbless, and their distribution encompasses tropical and subtropical regions of Central and South America, Africa, India, and Southeast Asia. Most caecilians are fossorial, spending the majority of their life underground. One South American group (Typhlonectes) is secondarily aquatic. The otic capsules, exoccipital, basioccipital, and parapophyseal ossifications of the braincase are fused into a solid unit (the os basale). Caecilians possess a pair of sensory organs called tentacles that are unique among vertebrates. The tentacles are protrusible organs, usually situated between the external nares and the orbit, and probably function in chemoreception. Although the orbits of caecilians are often covered with skin (and sometimes bone), the eyes are photoreceptive in most species that have been examined. In some caecilians, the eye is closely associated with the tentacle and in at least one species, they eye is protruded from the skull during protrusion of the tentacle (Nussbaum, 1992). The derived groups of caecilians fuse many elements of the skull and lack a tail. The fossil record of caecilians is notably poor but extends back at least as far as the Cretaceous. The Jurassic Eocaecilia shares many characteristics with extant caecilians (including the reported
There are approximately 415 described species of extant salamanders, the majority of which are found in the Northern Hemisphere. A significant radiation of tropical South American salamanders extends as far south as southern Brazil and Bolivia. Caudates share a number of morphological features of the skeleton, including the loss of the quadratojugal ossification and the presence of intravertebral spinal foramina in the atlas vertebra. Most salamanders also have a reduced number of bones in the pectoral girdle relative to extinct amphibian groups and frogs (caecilians lack a pectoral girdle), and they have an anterio-laterally projecting process on the atlas (the first vertebra), the tuberculum inter-glenoidicum, which forms accessory articulations between the skull and the vertebral column. Members of one large assemblage of caudates (the Plethodontidae) lack lungs, and this condition is also found in one gymnophionan. The early history of salamanders is poorly known, but fossils belonging to Caudata (as defined above) are known from the Mesozoic. Stem-group salamanders (outside of Caudata as defined above, but on the evolutionary stem leading to modern salamanders) are reported from the Jurassic. A recent review of Mesozoic taxa was provided by Evans and Milner (1996).

There are more than 4100 described species of extant anurans, distributed on all the major continental land-masses except Antarctica. Skeletal features shared by frogs include the fusion of forearm bones (radius and ulna), fusion of the lower leg bones (tibia and fibula), elongation of the ankle bones, reduced number of trunk vertebrae (10 or fewer among extant frogs), and presence of a urostyle (a rodlike, posterior bony extension of the vertebral column). Many of these skeletal features are associated with jumping ability and evolved as early as the early Jurassic (Shubin and Jenkins, 1993). The fossil record of the group is extensive (see recent review by Sanchez, 1998) and at least two stem-group frogs are now known from the early Triassic (Evans and Borsuk-Bialynicka, 1998).

Many amphibians have a larval stage followed by metamorphosis into the adult form. Metamorphosis was defined by Duellman and Trueb (1986, p. 173) as "a series of abrupt postembryonic changes involving structural, physiological, biochemical, and behavioral transformations." The transformations that take place in frogs are the most dramatic of any within the Tetrapoda. The majority of frog species have a larval stage in the life cycle (following hatching from an egg and preceding development of the mature adult body form) known as a tadpole. The tadpole has a more or less oval-shaped body with a tail; there is no distinct head region, but the mouth includes a beak and a number of rows of chitin toothlike structures. During metamorphosis many morphological and physiological changes take place that completely transform the animal's appearance and the way it interacts with its environment (e.g., the tail is resorbed, limbs grow, the axial and appendicular skeleton is dramatically transformed, the digestive tract shortens, gills disappear, and respiration function is transferred to lungs). Many frogs show variants of direct development.

J. Amniota

Amniota is defined as the most recent common ancestor of Reptilia and Mammalia, and all of that ancestor's descendants. Amniota is diagnosed by a suite of characters, including presence of an amniote egg (see below), caniniform teeth, two or more vertebrae in contact with the pelvic girdle, internal fertilization, and keratin (a protein that acts as the building block for scales, nails, hooves, hair, and feathers). The amniote egg includes a series of extramembranous membranes that surround the developing embryo and provide all the nutritional, waste disposal, and gas exchange requirements during development. In reptiles (including birds) and the monotremes (basal mammals), the embryo and extramembranous membranes are encapsulated in an egg with either a leathery exterior or a hard shell. Because this is a self-contained fluid-filled system, eggs can be laid away from water. Although fishes, amphibians, and most reptiles are often referred to as "cold-blooded," many species in these groups manage to maintain elevated body temperature at a constant level. This is achieved by shuttling between heat sources and heat sinks (ectothermy). However, birds and mammals generally maintain a constant temperature by changing the basal metabolic rate (endothermy). There is substantial evidence suggesting that some of the fossil reptiles did maintain body temperatures above the level of the environment. There has been much argument about the question of whether the mechanism by which this was achieved was ectothermy or endothermy. The earliest known members of the Amniota appear in the fossil record during the Pennsylvanian period.

K. Reptilia

Reptilia is defined as the most recent common ancestor of Chelonia and Sauria, and all of that ancestor's descendants. Reptilia is diagnosed by several features of the skull, including the presence of suborbital fenestrae.
populations are classified in the genus *Sphenodon* (turtles). They are found on all continental landmasses except Antarctica and also occupy all of the world’s tropical and temperate oceans and seas. All turtles lack marginal teeth, but the earliest known stem-group turtles (*Prenosuchus*) from the Triassic have a few small teeth on the palatal portion of the skull; all extant cheloniens lack palatal teeth. The palate is firmly attached to the braincase. The presence of a hard, bony shell is one of the most distinguishing features of turtles. The upper portion of the shell (the carapace) is formed in part from the vertebral column and ribs. During development these elements fuse with dermal ossifications to form a generally continuous, hard bony shell. The generally flat, ventral plate (the plastron) of the shell is formed anteriorly by dermal bones and skeletal elements that are probably homologous with elements of the pectoral girdle in other vertebrates. The bony shell of most turtles is covered by a series of keratinous scutes; in some forms, these scutes are lacking and a leathery skin covers the carapace and plastron. Turtles are unique among vertebrates in having the pectoral girdle enclosed by the rib cage. Turtles first appear in the fossil record during the Triassic.

L. Sauria

Sauria is the most recent common ancestor of Lepidosauria and Archosauria, and all of that ancestor’s descendants. Sauria is diagnosed by the presence of a specialized ankle joint, long and generally slender hind limbs, and relatively larger leg muscles. Lepidosauria can be defined as the most recent common ancestor of Squamata and the extant rhynchocephalians (*Sphenodon*; see below), and all of that ancestor’s descendants. Lepidosauria is diagnosed by various features of the skull and the presence of a transverse clausal slit, a distally notched tongue, and a modified middorsal scale row. Rhynchocephalians were a very diverse and widespread group of lepidosaurs during the Mesozoic era, but by approximately 60 million years ago, they were extinct everywhere except in New Zealand. The last surviving rhynchocephalians are restricted to a few islands in New Zealand. These relict populations are classified in the genus *Sphenodon*, and two species are currently recognized. Squamata includes all the lizards and snakes as well as the fossorial amphibiaenians. A number of anatomical features are shared by these three groups, including aspects of the skull, postcranial skeleton, musculature, and tongue. All squamates have paired, evertible hemipenes. Many squamate species lack, or have reduced, limbs and girdles.

There are approximately 4130 species of extant lizards, widely distributed on all the continents except Antarctica. A number of extinct groups were fully marine, but today only one species utilizes a marine habitat (the marine iguana of the Galapagos Islands feeds on marine algae). Many groups of lizards have lost or reduced limbs. The two North American species of *Heloderma* are venomous. Lizards first appear in the fossil record in the Jurassic. Both morphological and molecular data indicate that snakes evolved from within lizards, but their precise phylogenetic position among other squamates is unresolved. There are approximately 2700 species of extant snakes, with a distribution encompassing tropical oceans and seas and all continents except Antarctica. Several species spend most or all of their lives in open ocean habitats. Most snakes are limbless, but some basal members retain vestigial pelvic girdle and limb elements. A few poorly known Mesozoic snakes had well-developed limbs. All snakes are carnivorous, and several lineages of advanced snakes evolved potent venom that lets them subdue prey. Most snake species have an extremely modified and flexible skull structure that permits them to eat animals much more bulky than themselves and to ingest prey that is of a greater diameter than their own head. Snakes first appear in the fossil record during the Cretaceous, but their fossil record as a whole is relatively poor and consists predominantly of isolated vertebrae.

Amphibiaenas is a small group of squamates with approximately 160 extinct species. They are found in Africa, the Iberian Peninsula, the Middle East, Mexicofrom Panama to Argentina in South America, and Georgia and Florida in the southeastern United States. As a result of their subterranean habits and the fact that almost one-third of the described species are known from only one specimen, many aspects of their biology and systematics are not well understood. Three species of *Bipes* retain anterior limbs, but all other species are limbless. The compact and well-ossified skull is the primary burrowing tool and is covered with relatively large scales. The scales of the skin are usually rectangular and are arranged in rings that encircle the body and allow rectilinear locomotion. Amphibiaenans first appear in the fossil record in the late Paleocene of North America.

Archosaurs is defined as the most recent common ancestor of Crocodylia and Aves, and all of that ancestor’s descendants. Archosaurs are diagnosed by a num-
M. Mammalia

Mammalia is defined as the most recent common ancestor of Monotremata and Theria (marsupials plus placental mammals), and all of that ancestor’s descendants. Mammalia is diagnosed by a large number of anatomical and physiological features, including the presence of mammary glands and a unique jaw joint. Mammary glands in females produce milk to nourish the rapidly growing young during postnatal development. In most vertebrates the jaw joint is formed between the quadrate bone of the skull and the articular bone of the lower jaw, but in mammals, the jaw joint is formed between the dentary bone (the only bone in the lower jaw of mammals) and the squamosal bone of the skull. Bones homologous with the quadrate and articular are still present in mammals, but their positional relationship and function has been transformed. The incus and malleus in the middle ear of mammals are homologous with the quadrate and articular (respectively) in other vertebrates. The 4810 species of extant mammals can be found on all continents and major bodies of water on Earth.

Monotremata is represented by three living species (duck-billed platypus and two species of echidna) that are found only in Australia and New Guinea, but the fossil record of this group extends back to the Cretaceous and reveals that monotremes were once found in South America. The monotremes are the only mammals that lay eggs. The marsupials are found today only in Australia and surrounding islands, North America, and South America. Their earliest appearance in the fossil record is in the Cretaceous of North and South America, and fossils are known from all continental landmasses. Marsupials differ from other mammals in a number of physiological and anatomical features, especially those related to reproduction. The gestation period in marsupials is very short compared with that of other mammals and the young are born in a nearly embryonic state. After birth, development takes place in a pouch (or sometimes under a fold of skin) on the mother’s abdomen. The Placentalia includes a diverse assemblage of mammals that include many of the most familiar mammals and the only mammals to adopt a completely or predominantly marine lifestyle (whales, dolphins, porpoises, seals, sea lions, walruses) and to achieve powered flight (bats). Human beings are members of a placental mammal group, the Primates.

See Also the Following Articles

AMPHIBIANS, BIODIVERSITY OF • BIRDS, BIODIVERSITY OF • ENDANGERED BIRDS • ENDANGERED MAMMALS • ENDANGERED REPTILES AND AMPHIBIANS • FISH BIODIVERSITY OF • MAMMALS, BIODIVERSITY OF • REPTILES, BIODIVERSITY OF

Bibliography


Cannatella, D. C., and Hillis, D. M. (1993). Amphibian relationships:
I. Historical Biogeography
II. Vicariance, Dispersal, and Extinction
III. Examples of Vicariance and Dispersal
IV. Plate Tectonics
V. Cladistic Methods in Biogeography
VI. The Progression Rule
VII. Vicariance Biogeography
VIII. Poeciliid Fish in Middle America
IX. Component Analysis
X. Heterandria and Xiphophorus
XI. Conclusions

GLOSSARY

area of endemism An area recognized by congruent distributions of two or more groups of organisms.
biogeography The study of the distributions of organisms on Earth and why the patterns exist.
cladistic biogeography The combination of cladistics (phylogenetic systematics) with vicariance biogeography. A method that searches for patterns of relationship among areas of endemism.
cladistics (phylogenetic systematics) A method of phylogeny reconstruction concerned with branching patterns. Sister group relationships are hypothesized on the basis of shared derived, or synapomorphic, characters.
components Elements of groups of areas, or taxa, as determined by the branching pattern of a cladogram. For example, in a group comprising three taxa A, B, and C, when B is more closely related to C, there are two components, an ABC component and a BC component.
dispersal The movement, or spread, of an organism from one area to another independent from other organisms and of historical events that change the natural distribution of the organism (see Fig. 1).
ecological biogeography Spatial patterns of organisms considered in terms of their interactions with other organisms and the environment.
endemic A taxon, or group of taxa, restricted to a prescribed area and found nowhere else on Earth.
historical biogeography Spatial patterns of organisms interpreted in terms of their concordance to patterns of Earth history.
panbiogeography The examination of distributions and relationships on a worldwide scale; term introduced by Croizat (1964).
plate tectonics (including continental drift) The theory that the Earth's crust is composed of plates that move relative to each other through seafloor spreading and subduction. As continents are found on plates, continental drift is one result of this movement.
vicariance The existence of closely related taxa or biota in separate, disjunct areas, which have been separated by the formation of a natural barrier (vicariance event).
vicariance event The splitting of a taxon or biota into two or more geographical subdivisions by the formation of natural barriers, e.g., mountain building, glaciation, or stream capture (see Fig. 1).
VICARIANCE BIOGEOGRAPHY, in the strict sense, is the study of repeated patterns of disjunct distributions within many members of a biota that may be explained by vicariance (or splitting) and other historical events. However, in the past 20 years or more, it has become a metaphor for historical biogeography, involving the study of life and Earth history using cladistic methods. The subject has developed in fits and starts but has today come to represent a method to classify areas of endemism on a global, regional, and local scale in terms of their historical relationships.

I. HISTORICAL BIOGEOGRAPHY

Biogeography means many different things to biologists and geographers of different persuasions depending on their outlooks and purposes of enquiry. It has been clearly noticed and well documented through accumulation of systematic distribution data for most groups of organisms over the past two centuries of global exploration that there are many organisms that exhibit particular distribution patterns on Earth. In efforts to interpret these patterns, biogeographers pose the question: "What lives where and why?" The usual answers invoke either ecological or historical explanations, or a combination of both (Nelson and Platnick, 1981; Craw et al., 1999; Humphries and Parenti, 1999).

For the most part ecological biogeographers classify distribution patterns as communities in a general hierarchy of organization. Populations and species live in habitats that belong to different land classes in ecosystems within biomes. Few would dispute that many similarities of climate, topography, habitat, and landscape exist between, for example, the rain forests of South America, Africa, and Southeast Asia, but, when examined closely, the resident organisms in each rain forest are specifically and generically quite different. Similar species in similar ecological niches belong to phylogenetically disparate taxa that have particular phylogenetic and distribution histories that have occurred through geological time. Historical biogeographers focus on the older developments of global biotas in an effort to study the history of the Earth rather than relatively more recent events such as short-term species interactions and recent perturbations in distribution due to climatic changes.

Data for historical biogeography come from comparative biologists and systematists who include distribution information in their monographs and revisions. A major synthesis is found in the works of Croizat, especially in *Space, Time, Form: The Biological Synthesis* (1964; see Craw et al., 1999). Croizat considered geological events and changes in Earth history and the "form-making" of species as two parts of a continuous historical process. To him, evolution of the Earth and the evolution of organisms are part and parcel of one series of historical events. Recently, greater attention has been given to these ideas, kindled by the general acceptance of plate tectonics and the notion that the continents are continually reshaped and changing in their spatial geometry (see Craw et al., 1999). At the same time, there has been a widening interest in the use of cladistics for systematic investigation of organisms (Hennig, 1966). Today, cladistics is considered as the primary means of reconstructing the relationships of organisms. Comparing different groups of organisms sharing similar distribution patterns and occupying similar areas of endemism with historical patterns in geology and Earth history allows biogeographers to provide retrodictive hypotheses of divergence and explanations of past and present diversity of organisms (Humphries and Parenti, 1999).

II. VICARIANCE, DISPERSAL, AND EXTINCTION

Particular species are endemic to certain areas today either because their ancestors originated there and their descendents have survived to the present day or because the ancestors occurred elsewhere and later some of the descendents dispersed into new areas. Such a difference invokes two, very different, historical explanations which in a shorthand way are described as vicariance or dispersal modes of diversification. In a vicariance explanation, an ancestral species (Fig. 1a, 1) divides into two separate populations when a physical barrier...
is formed across the population such that neither of the two new populations can cross the barrier (Fig. 1a, 2). With time, the two separated populations evolve into two closely related species, shown here as black-and-white ovals (Fig. 1a, 3). In a dispersal explanation, the range of one species is already bounded at least in one direction by a barrier (Fig. 1b, 1), which is later crossed by some members of the population (Fig. 1b, 2). If the organisms of the new population survive, successfully interbreed, and yet remain isolated from the "ancestral" population, they eventually evolve into a different species (Fig. 1b, 3). Thus, in the vicariance explanation the appearance of the barrier is seen as the cause of the disjunction of the two species, whereas in the dispersal explanation the barrier is considered to be older than the disjunction.

III. EXAMPLES OF VICARIANCE AND DISPERSAL

By Victorian times and the emergence of evolutionary theory, it was clear that the distinction of dispersal and vicariance explanations was emerging. As described by Nelson and Platnick (1984), the English botanist Joseph Hooker (1817–1911) provided perhaps the earliest coherent vicariance explanations of plant distributions. Hooker studied the plants of the Southern Hemisphere in South Africa, Australia, southern South America, and New Zealand and concluded that their pattern of distribution suggested a former, more extensive flora, which had broken up by geological and climatic causes. Hooker compared the southern floras with those of the Northern Hemisphere and also concluded that the various southern floras were more closely related to each other than any were to the floras of the Northern Hemisphere. As a consequence, he also boldly concluded that divergence of the northern and southern floras must have occurred before the fragmentation of the southern floras. His vicariance hypothesis considered that initially a worldwide flora was first subdivided by a barrier into the northern and southern floras and then the latter were further subdivided by expanses of ocean—barriers severing once continuous floras that no longer exist today.

Nelson and Platnick (1984) compared Hooker’s views with fellow Englishman Alfred Russell Wallace, another great biologist of the late nineteenth century. Wallace disagreed entirely with Hooker’s hypothesis. In 1876, he said that the north and south division of biotas was not a vicariance pattern but represented the fact that "... the northern continents are the seat and birthplace of all higher forms of life, while the southern continents have derived for the greater part, if not the whole, from the north; ... it implies the erroneous conclusion that chief southern lands—Australia and South America—are more closely related to each other than to the northern continent."

At that time the prevailing geological theory was one of fixed continents and slow, gradual changes in both Earth history and biogeographical developments. Hooker saw a close connection between cause and effect and said that the basic pattern of relationships divided the northern from the southern taxa, and the pattern was explained by vicariance events. Wallace, heavily influenced by Darwinian selection theory on the other hand, did not see a necessary connection between distribution and Earth history. For him, the basic pattern was east–west and the pattern was explained by dispersal from northern centers of origin.

IV. PLATE TECTONICS

It has become clear that the nineteenth century was characterized by the dominance of British Empire and European colonialism throughout the world. It was no accident that the "favored races" were of northern origin as indicated in the title of Darwin’s (1859) famous book: On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggles for Life. Although throughout his life his ideas changed, Wallace’s earlier views on biogeography were part of Darwinian orthodoxy (Nelson and Platnick, 1984). Many groups of animals and plants were seen created in "centers of origin" and by dispersal to other areas. Following the central dogma of Darwinism, the colonizing organisms were seen as the surviving fittest and superior competitors to all others that lay before them. The less fit southern forms became thrust aside and eventually extinct. This rather fanciful scenario has been repeatedly described in textbooks, even within the 1990s, and is sometimes known as the "monoboreal relict hypothesis" (Nelson and Platnick, 1984).

The origin of taxa from "centers of origin" and by dispersal to other areas is heavily entrenched in bioge-
percontinent, Laurasia. Laurasia contains those areas that subsequently crystallized out into Europe, North America, and Asia, but exclude the Indian subcontinent. From the Cretaceous onward, the southern end of the world, Gondwanaland, broke up into those areas we now know as Antarctica, Australia, India, Africa, and South America. The breakup of Gondwanaland lends credence to vicariance biogeography and hence Hooker’s theory. The similarities among floras of the Southern Hemisphere suggest former land connections. It also revived the theory that Hooker’s hypotheses of north–south connections exist for plants and contradicted Wallace’s theory of east–west connections based on animals. As Nelson and Platnick (1984) asked, how can we determine which hypothesis is correct? Are both theories correct to some degree, or is there another hypothesis that should be erected which subsumes both into one explanation? The interesting point is that if the explanation is a general one, what process caused it and what methods are required to uncover the historical pattern? Paradoxically, through attempts to resolve the relationship between space, form, and time in terms of vicariance and dispersal, including application of track analysis in panbiogeography (Croizat, 1964; Craw et al., 1999) and cladistics in vicariance biogeography (Nelson and Platnick, 1981), it was clear that a more general issue was at stake. This was the question of how to interpret patterns of plant and animal distributions in a systematic way to provide independent theories of biogeography, rather than rely on geophysical and geological theories as prior explanations for present-day distributions (see Humphries and Parenti, 1999). To give some flavor of the rather fitful debate this idea created, the stages of application of cladistic methods that grew out of vicariant biogeography are briefly described.

V. CLADISTIC METHODS IN BIOGEOGRAPHY

The analysis of patterns in cladistic biogeography starts with finding the species or higher taxa that exist in the different biogeographic areas and to discover their relationships. Cladograms, or branching diagrams, are used to express hypotheses of relationship between organisms. An example is shown in Fig. 2a, where among the four species, C and D are more closely related to each other than either are to B, and also species B, C, and D are more closely related to each other relative to A. The cladogram specifies two inclusive informative
Cladograms are produced by the distribution of characters or features shared by the different species or taxa. Thus, in Fig. 2a, character 3 specifies the group CD, character 2 specifies group BCD, and character 1 specifies ABCD as the overall group. Characters 2 and 3 then specify unique groups that have characters not shared by any other organism. If such features are taken to be present in the ancestors of each group, then the cladogram can be taken as specifying the pattern of evolution or common descent of that group. Seen in this way, the cladogram of Fig. 2a may be taken to specify that taxa C and D have a unique ancestor Z that is not the ancestor of A or B and also that taxa B, C, and D have an ancestor Y that is not the ancestor of A. CD and BCD are monophyletic groups, compared, for example, with A and D, which is a polyphyletic group, and A and B, which comprise a paraphyletic group.

Cladistics was applied to biogeography in the 1960s at about the same time when theories of plate tectonics and shifting continents were becoming respectable. However, dispersal hypotheses to explain distribution patterns were still in vogue and thus the earliest applications in biogeography used cladograms to determine “centers of origin” of monophyletic groups. The German dipterist Willi Hennig (1966) reasoned that there is a recognizable, close relationship between a species and the space each occupies. He assumed that dispersal patterns are unique for each taxonomic group and that each had an independent history.

Central to Hennig’s (1966) method to find a center of origin for a group of taxa from a particular cladogram was the idea that phylogenetically primitive members of a monophyletic group will be found near the center. That is, they are occupied by the species nearest to the basal nodes of cladograms. Within a continuous range of species of a monophyletic group, it was plausible that a series of characters transform and run parallel with distribution in space, such that the youngest members would be on the geographic periphery of the group.

To apply the “progression rule,” one works backward considering the areas in the terminal positions of the cladogram and then gradually working down to the base of the cladogram, assigning distributions to the internal nodes or the common ancestors. If the areas occupied by the descendents do not overlap, they are combined together in assigning the probable distribution of the common ancestor. If they do overlap, the unshared element is eliminated. Consider the four species in Fig. 2b. Species A and B occur in South America, C occurs in eastern Australia, and species D occurs in New Guinea. Working from the most derived species, C and D, the distributions clearly do not overlap; hence they are added together to give the area for ancestor Z as eastern Australia plus New Guinea. For ancestor Y the descendant taxa do not overlap, with B occurring in South America and ancestor Z in eastern Australia and New Guinea. The two distributions do not overlap so the area for ancestor Y is South America, eastern Australia, and New Guinea. Moving down the cladogram to ancestor X, descendent species A occurs in South America and ancestor Y in South America, eastern Australia, and New Guinea. The two distributions thus overlap in South America, and eastern Australia and New Guinea are eliminated as the unshared area. Working back up the cladogram, the simplest hypothesis is to consider that South America is common to all three ancestral species, X, Y, and Z, and that species C and D dispersed to eastern Australia and New Guinea from South America. The conclusion from analysis of this biogeographic pattern is that the center of origin for the group is South America.

To demonstrate this early use of cladograms in biogeography, a simple example can be seen in the caddis fly distribution of Wormaldia (Ross, 1974; see Humphries and Parenti, 1999). Wormaldia and other caddis flies have been a lifelong study of Ross because they occur in almost all freshwater systems throughout the globe, show a high degree of endemcity in all areas, and are very speciose. Ross considered the geographical
distribution of the nine species of the *Wormaldia kisoensis* complex of caddis flies. Eight species are present in the western Pacific from northwestern Borneo in the south and Japan in the north and one species occurs in the Smoky Mountains of eastern North America (Fig. 3). The cladogram of phylogenetic relationships is shown diagrammatically in Fig. 4. Like Hennig, Ross (1974) assumed that the species nearest to the base of the stem of the cladogram denotes the ancestor of the group. He applied also the progression rule to the analysis to determine the center of origin of the group. In *Wormaldia* the most derived species pair occurs in Japan and eastern North America, which led Ross to propose the dispersal hypothesis of origin of the group in the western Pacific and a single dispersal of one species across the Bering straits to North America (Figs. 3 and 4).

Brundin’s classic studies of chironomid midges (Brundin, 1966) showed that the southern temperate areas of South America, Southern Africa, Tasmania, southeast Australia, and New Zealand are inhabited by 600–700 species. Trans-Antarctic relationships are a recurring phenomenon throughout the group, so by way of an example consider the midges of subfamily Diamesinae, which show a double distribution: two major groups present in both Northern and Southern Hemisphere temperate areas but absent from the Tropics. The largest and most widespread is the relatively generalized tribe Heptagymi, represented by eleven species in Andean South America, two species in southeastern Australia, and five species in New Zealand (Fig. 5). Its sister group is the relatively more apomorphic and monotypic tribe Lobodiamesini of New Zealand.

The cladogram in Fig. 5 shows that there are a total of 25 terminal taxa in the Southern Hemisphere areas of South America, New Zealand, Australia, and South Africa and three groups in Laurasia. The monotypic genus *Heptagyia* occurs in South America and *Paraheptagyia* has five South American species. According to...
Brundin, the southeastern Australian subgroup of two species is a younger evolved offshoot of the older South American group including *Heptagyia*. Brundin considered the Australian taxa to have dispersed from Patagonia or east Antarctica at stem 6 by the end of the Paleocene because they have derived characters and because the stem species (indicated by 1, 2, and 4) never occurred in Australia. The other stem (2a) includes the genera *Reissia*, with three species in South America, *Limaya*, with two species in South America, and *Maoridiamesa*, with five species in New Zealand. Brundin (1981) considered that *Maoridiamesa*, on a different stem from the *Paraheptagyia* group, agreed well with plate tectonic theory for an early separation of New Zealand from western Antarctica in the Upper Cretaceous. The fact that *Maoridiamesa* is a comparatively younger, derived offshoot of an older group in South America was, according to Brundin, evidence of long-distance dispersal from South America via west Antarctica to New Zealand of stem species 4a rather than a vicariance event. In other words, Brundin thinks that the *Maoridiamesa* group is younger than the areas in which it occurs.

The progression rule can be interpreted as one particular optimization procedure applied to cladograms. It has problems in the sense that it can specify a different center of origin when applied to different data. Changes in data occur for a host of different reasons, but largely through the discovery of known taxa in new places, the discovery of new taxa, and reinterpretation of relationships. Patterson (1981) showed that the progression rule was very sensitive to the discovery of new taxa, and particularly fossils. The discovery of fossils simply added more areas to the problem and thus increased the number of possibilities for finding new centers of origin. Thus, for example, if we added two new fossils, V and W, to the example described earlier (see Figs. 2 and 6), then the possibility of a center of origin in North America and in turn Europe would alter the interpretation significantly. Patterson considered that for groups of organisms with a significant fossil record, the progression rule is heavily influenced by fossils and this would be especially true in the areas where the best fossils occurred. The progression rule is dependent on the notion that organisms occur in a center of origin. The problem is that when the progression rule is applied to a cladogram, it can always specify a center whether it may, or may not, have occurred. Suppose, for the sake of argument, the species of Fig. 6b in Europe, North America, South America, eastern Australia, and New Guinea were caused by the breakup of continents, a vicariance explanation; then the progression rule would provide a spurious hypothesis impossible to test.
VII. VICARIANCE BIOGEOGRAPHY

The problem of interpreting cladograms as dispersal scenarios is the same as treating them as phylogenetic trees rather than as hierarchical classifications to show groups with relatively more inclusive groups of set membership. Such a procedure requires making ad hoc assumptions not based on the information on which they are based. Furthermore, interpreting individual cladograms as having individual histories leads to certain conceptual difficulties. A crucial one is how do we interpret repetitious distribution patterns? If we have many unrelated taxonomic groups repeating a pattern of distribution between major continents, such as South America, eastern Australia, and New Zealand (Fig. 7), it is improbable that the pattern was caused by separate unique dispersal events. This implies that species in each different group separately makes its own way from one continent to the other. A logical and more robust conclusion would be to suggest that at one time, the continents were in contact and that the present-day pattern was due to the breakup of a formerly continuous biota.

The breakthrough in application of cladistic reasoning to biogeography came with the realization that disjunct distributions could occur because of such vicariance events. This is based on the idea that present-day endemic taxa occur in the same areas as their ancestors and the Recent taxa evolved in situ. Where dispersal models explain disjunctions by dispersal across pre-existing barriers, vicariance models explain them by the appearance of barriers fragmenting ancestral species ranges (Fig. 1). To historical biogeographers what became particularly clear was the important idea that distribution data are insufficient to resolve decisively either dispersal or vicariance as the cause of disjunct distribution patterns.

Platnick and Nelson (1978) argued that one should not worry about the cause of a disjunct distribution pattern between different related areas of endemism but whether or not it conforms to a general pattern of relationships shown by other groups of taxa endemic to the areas occupied. Thus, there is an analogy to the three-taxon statements in systematics as the most basic units for expressing relationships. In biogeography three-area statements are the most basic units for expressing relationships between areas of endemism. The generality of the area cladograms can be examined by comparison with other unrelated taxonomic groups endemic to the relevant areas and corroboration of a particular pattern is equivalent to a general statement for the relative recent ancestry of the biotas under scrutiny. However, there has been a long, drawn out saga of trying to find ways for applying the principles of cladistic biogeography because initial applications of the method (e.g., Rosen, 1976) encountered problems of incongruence between different groups of organisms. Theoretically, it should be possible to connect every area of endemism of the world into one huge general statement of interrelationships. However, our perception of the world is less than perfect for a variety of reasons—extinction, dispersal of widespread taxa, and restricted distributions of taxonomic groups. The significance of
these problems can be demonstrated with poeciliid fish in Middle America.

**VIII. POECILIID FISH IN MIDDLE AMERICA**

To discover relationships of areas, at least two groups of taxa must be available for the same or similar set of areas. Rosen (1978, 1979) examined two groups of fishes, *Heterandria* and *Xiphophorus* from the Mesoamerican region. Both genera have close relatives elsewhere but each has a monophyletic subgroup inhabiting 11 areas in southern tropical Mexico, south to eastern Honduras in *Xiphophorus*, and further south to eastern Nicaragua in *Heterandria* (Fig. 8). Area cladograms for the two genera, which indicate areas occupied by particular species, are shown in Figs. 9 and 10.

**IX. COMPONENT ANALYSIS**

Cladistic biogeography would be uncomplicated if all groups of organisms were each represented by one taxon in each of the smallest identifiable areas of endemism. This is not the case. Unique patterns may be meaningful and cannot at the same time be incongruent with patterns determined from other taxa. Nevertheless, incongruence between two or more cladograms can occur for a variety of historical reasons. Particular groups of organisms can exhibit older or younger patterns than the groups to which they are being compared. Also, some taxa have dispersed and become widespread and by so doing have obscured the historical signal in the area patterns. Comparing several different groups of organisms is redundant in the sense that the same pattern is repeated over and over or areas are represented by two or more taxa. With extinction the observed pattern creates a spurious historical signal. All of these problems effectively cause sampling errors and lead to wrong predictions of the general patterns of area interrelationship. Nelson and Platnick (1981) used cladistic logic and component analysis to make comparisons between area cladograms to yield the maximum resolution in the general-area cladograms (see Box 1).
Box 1

Component Analysis

To determine the degree of congruence between different groups of organisms occupying similar areas, the problems of “missing areas,” widespread taxa, and redundancy are taken into consideration using component analysis and consensus.

Components are the elements of a group of areas, or group of taxa, as determined by the branching pattern of a cladogram. For example, in a group comprising three taxa (or areas) A, B, and C, when B is more closely related to C, there are two components, a general uninformative ABC component and an informative BC component.

Components of cladograms in vicariance biogeography can be manipulated to extract historical signal from even the most recalcitrant cases (Fig. 11). For example, consider the four areas of endemism, South Africa (SA), Pacific South America (PSA), New Zealand (NZ), and eastern Australia (EA). Now consider three groups of organisms distributed in these four areas: three species of lizards (L1–L3) in SA, PSA, and NZ, and three species each of birds (B1–B3) and fish (F1–F3) in PSA, NZ, and EA. The three numbered informative nodes (1–3) on the cladograms express the relationships for each species in the lizards, birds, and fishes. Notice that each group of organisms occurs in three areas, and in each case there is one area not present (EA for the lizards and SA for the birds and fish). Also, for each group there is one species endemic to an area, i.e., L2 in PSA, B2 in NZ, and F2 in EA. The remaining species are more widespread, occurring in either two or three areas (respectively L1, B1, and F1 in two areas and L3, B3, and F3 in three areas).

Because of widespread species and the lack of representation of one area in each of the area cladograms, at first sight relationships between the four areas seem difficult to resolve (Fig. 11). However, it can be assumed that each species originated in one area with extinction, dispersal, and failures to vicariate, causing the more complex patterns. Taking this view, we can examine the informative three-area relationships in the area cladograms to see if they tell us anything about the general-area relationships. Thus, in the lizard cladogram, L1 in SA, L2 in NZ, and L3 in PSA is the only informative statement that can be represented as SA(PSA, NZ) and scored as 0(1, 1) in a matrix (Table I, column 1). The “missing” area (EA) is scored as a dash. Informative area statements for birds are PSA(NZ, EA) and EA(NZ, PSA), and for fishies PSA(EA, NZ) and NZ(EA, PSA) similarly scored in the matrix (Table I, columns 2–5). Consequently, for the three cladograms there are five informative three-area statements (Table I). By using a standard parsimony

![Figure 11](image-url)
FIGURE 12 General-area cladogram for lizards (species of lizards (L1–L3), birds (B1–B3), and fishes (F1–F3)). The acronyms for areas are as in Fig. 12. The five informative components (1–5) identify area relationships.

analysis, the best fitting tree for the five statements is shown in Fig. 12. Thus, despite the complexity of the patterns in the original three groups (Fig. 11), if there is one history that best explains the data, then it might be a vicariance hypothesis for the four areas as offered in Fig. 12—(SA, PSA, NZ, EA) (SA(PSA, NZ, EA)) (SA(PSA(NZ, EA))))—and shown as an area cladogram in Fig. 13, indicating the relationships of the four areas of endemism.

X. HETERANDRIA AND XIPHOPHORUS

To illustrate just how the methods to extract the best possible patterns of relationship work, consider the two groups of poeciliid fish genera Heterandria and Xiphophorus (Figs. 8–Fig. 10) based on the work of the American ichthyologist Donn Rosen (1978, 1979). The 11 identifiable areas of endemism, each determined by substituting species of Heterandria or Xiphophorus for the areas in which each occurs, include areas 1, 2, 4, 5, 8, 9, and 10 common to both groups. Areas 4 and 5 are occupied by one species from each group and thus treated as one area, [4–5]. An area 11 was described by Rosen as a putative hybrid area between areas 4–5 and 2, but to simplify matters, it is excluded here. A comparison of the two cladograms shows that Xiphophorus is less informative than Heterandria because it has two widespread species in areas [4–5], 6, 9, and 10 and is absent from area 7. In Heterandria areas [4–5], 6, 9, and 10 are all occupied by recognizable endemic species. However, the pattern is not quite complete in that there is no species present in area 3.

By using the assumption that one sequence of historical events can explain the patterns observed in both Heterandria and Xiphophorus and then using the logic applied for the hypothetical example used for the breakup of Gondwana (Box 1), a single result can be obtained. Platnick (1981) noted that if widespread taxa are uninformative, they cannot at the same time be incongruent. We can assume that the information on areas 6 and 9 in Heterandria is correct. Thus the incongruent information in the same areas for Xiphophorus (areas 6 or 4–5, but not both, and 9 or 10, but not both) is due either to dispersal or a failure to speciate in response to vicariance events that gave the pattern.
in Heterandria. Taken on their own, widespread taxa are uninformative. However, when considered with other cladograms, informative results are possible. Similarly, “absence” data (e.g., area 3 in Heterandria and area 7 in Xiphophorus) can never be incongruent with information at hand so unique areas should be placed in terms of whatever pattern exists (i.e., area 7 by Heterandria and area 3 by Xiphophorus). By using these assumptions, the Xiphophorus cladogram allows the populations in areas 9 and 6 to occur in any of 12 positions and the Heterandria cladogram allows area 3 to float onto any of the branches of the cladogram. Thus, of the 425 allowable trees for Xiphophorus and the 27 for Heterandria, the analysis yields three possible intersections. These are summarized as a consensus tree (Fig. 14).

By using this method, we have an area hypothesis that accounts for all 10 areas of endemism (i.e., excluding the hybrid area 11) that can be recognized from the two genera of fishes Heterandria and Xiphophorus. If the pattern has been created due to changes in Earth history, the question that we could ask now is: “What might have been the historical factors in Mesoamerica to cause this pattern and how might these be compared with the given biological distribution?” Ideally, we would require that geological information be assembled into cladograms, in the same way as biological cladograms, so that biotic and historical patterns can be compared. Until such time as geological data can be ordered for a more informative comparison, one thing we can say is that the observed patterns in species relationships of Heterandria and Xiphophorus in Mesoamerica have been formed over a period of at least the past 80 million years and are most likely to have been brought about by vicariance events where life and Earth evolved together (Humphries and Parenti, 1999; Rosen, 1978).

XI. CONCLUSIONS

At this time, methods of vicariance or cladistic biogeography have proved useful to analyze and compare biotic patterns at the highest resolution and organized in such a way as to compare them to independent sources of data such as geological patterns. Cladistics is a general method of determining class and subclass relations whatever the source of data without recourse to evolutionary narrative. A cladistic approach to Earth history makes it possible to express biogeographic interrelationships as hierarchical relations from biotic information. The development of cladistic methods that cater for the possibilities to consider complicating events but without ever using such information prior to analysis provides a system that allows the possibility of generating general biogeographic hypotheses even from seemingly ambiguous patterns. In other words, vicariance biogeography is a general empirical procedure without any prior assumptions in the analysis about dispersal, vicariance, redundancy, or extinction events but which at the same time never denies that they occur.

See Also the Following Articles

BIOGEOGRAPHY, OVERVIEW • CLADISTICS • DARWIN, CHARLES • DIVERSITY, COMMUNITY/REGIONAL LEVEL • ENDEMISM • ISLAND BIOGEOGRAPHY

Bibliography


Croat, L. (1964). Space, Time, Form, the Biological Synthesis. Published by the author, Caracas.
I. Diversity of Wetland Ecosystems
II. Diversity within Wetland Ecosystems
III. Functional Diversity
IV. Threats to Wetland Diversity
V. Restoring and Maintaining Wetland Diversity
VI. Conclusions

GLOSSARY

benthic Located on the bottom of a water body, or pertaining to bottom-dwelling organisms.
biological diversity The variety of different forms of life on earth, usually considered to include variety at all levels of biological organization from genetic variation to variety seen in species, populations, communities, and ecosystems.
bog An acidic peatland that receives all water inputs from the atmosphere and hence has low pH (usually 3.4–4.6), low alkalinity, and negligible concentrations of base cations in surface waters; typically dominated by mosses of the genus *Sphagnum* and other plants able to tolerate low pH and low nutrient availability—that is, members of the heath family (Ericaceae).
calcareous fen A term variously used for wetlands with high calcium carbonate in water and soils. In the strict sense, the term refers to peat-accumulating wetlands with surficial deposits of calcium carbonate and a distinctive flora of calciphilic species. These wetlands typically develop where substantial amounts of cold, calcium-carbonate rich groundwater discharge to the surface.
detritivores Organisms that consume dead plant or animal material.
emergent vegetation Rooted plants that tolerate saturated or flooded soil but not extended periods of submergence.
floodplain Broad, flat areas adjacent to rivers and streams that become inundated during floods.
hydrologic regime The aggregate set of variables describing the behavior of water in wetlands, including water depth, magnitude and frequency of water level fluctuations, seasonal pattern of water fluctuations, direction and velocity of water flows, and water residence time.
hydrophytes Plants capable of growing and persisting in saturated or flooded soils.
littoral The shoreline zone where sunlight penetrates to a wetland’s substrate and supports rooted plant growth.
macroinvertebrates Organisms lacking backbones and generally visible to the human eye.
mash A wetland dominated by herbaceous species of plants (e.g., cattails, sedges, grasses, floating ferns) and typically occurring on mineral soils. Floating-leaved and submerged species of plants dominate the parts of marshes that grade into open water. Although the surface layer of some marsh soils may
WETLANDS ECOSYSTEMS

consist of shallow accumulations of organic matter over mineral soils, the depth of this layer is usually less than the minimum required for the site to be classified as peatland.

peat A type of soil formed from accumulation of the remains of dead plants, with high organic matter content (minimally 20% and as high as 90% by dry weight) and depth greater than 30 to 40 cm. Peat forms in some wetlands because anoxic conditions, low pH, low temperatures and short growing seasons, or plant tissues highly resistant to decay prevent rates of decomposition from keeping up with rates of organic matter production.

peatlands Wetlands in which dead organic matter accumulates to form peat more than 30 to 40 cm deep; typically includes bogs, poor fens, and rich fens but some swamps also occur on peat.

poor fen An acidic peatland similar to a bog but having a somewhat higher surface water pH (4.3–5.8) because it receives some groundwater or surface water in addition to precipitation; typically dominated by Sphagnum mosses or sedges. But for the presence of distinctive poor-fen indicators, vegetation may superficially resemble bog vegetation or sedge-dominated rich fens that lack Sphagnum.

rich fen An alkaline peatland that receives significant inputs from groundwater or surface water; typically having surface water pHs >6.0, higher alkalinity, and distinctly higher concentrations of base cations than bogs and poor fens; dominated by brown mosses (true mosses, largely of the Amblystegiaceae) and small sedges.

species richness The number of different species in a specified area.

submerged vegetation Plants that grow and reproduce while completely submerged.

swamp Wetlands dominated by woody species. Although the term has been variously used, it is used here in the restricted sense of forested wetland, either on mineral soils or peat. However, while many peatland ecologists classify forested fens on deep peat as swamps, they would not call a bog forest a swamp.

THE BIOLOGICAL DIVERSITY OF WETLANDS, or wetland biodiversity, encompasses the immense variety of life supported by the high number of different types of wetlands occurring worldwide. This chapter describes the diversity of wetland ecosystems occurring across major geographic regions, and the diversity of plant and animal species found within particular types of wetlands, including factors believed to control observed patterns in species diversity. The functional diversity of wetlands is discussed at the organismal level (i.e., in terms of various adaptations of plants and animals to the wetland environment), because these adaptations often allow a relatively large number of species to coexist within a wetland. The chapter also identifies threats to wetland diversity and discusses methods of managing to reverse trends in the loss of wetland diversity. Examples and data are drawn primarily from North America, with some reference to wetlands on other continents.

I. DIVERSITY OF WETLAND ECOSYSTEMS

Wetlands, and the organisms adapted to these unique environments, are as diverse as the myriad combinations of climate, topography, and geology that cause different types of wetlands to form. By definition, wetlands are ecosystems whose physical, chemical, and biological characteristics are determined by the constant or recurrent presence of water at or near the soil surface. Beyond this commonality, however, lies a seemingly infinite number of ways in which climate and landscape features combine to create wetlands with different magnitudes and seasonal patterns of water level fluctuation, paths and rates of water movement, soils, water chemistry, nutrient status, and species composition. Out of this landscape diversity rises the large number of different types of wetland ecosystems seen worldwide.

Wetlands occur on every continent except Antarctica and in every climatic regime. The best available estimates indicate that they cover only between 4 and 6% of the Earth’s land surface. Within this relatively small area, however, an astonishingly high number of different wetland ecosystem types occur. A recent global summary (Mitsch, 1994) hints at this diversity when it describes the world’s wetlands in terms of coarse-scaled differences in climatic setting, geomorphologic setting, and major wetland type. According to that summary, wetlands “are found (a) in humid, cool regions as bogs, fens, and tundra; (b) along rivers and streams as riparian wetlands, seasonally flooded forests, and back swamps; (c) in the deltas of the world’s great rivers; (d) along temperate, subtropical, and tropical coastlines as salt marshes, mudflats, and mangrove swamps; and (e) in arid regions as inland salt flats, seasonal playas, and vernal pools.” They also are found
WETLANDS ECOSYSTEMS

FIGURE 1 Southern deepwater swamp, Congaree National Monument, South Carolina, USA. Photo by Donald J. Leopold.

FIGURE 2 Forested wetland in upstate New York, USA. Photo by Donald J. Leopold.

as (f) marshes, wet meadows, and swamps in depressions that occur in a variety of landscapes and climatic settings; (g) marshes, wet meadows, and swamps on extensive flat areas where the land surface slope is minimal and climate is humid; and (h) marshes, bogs, and fens along the shores of lakes in many climatic regions. Wetlands even occur on (i) slopes as blanket bogs, seepage fens, and vegetated springs.

Finer-scaled differences in climate, geomorphologic setting, and biogeographic region yield an even greater diversity of wetland ecosystem types. Mangrove swamps, for instance, do not exhibit uniform species composition worldwide. The mangrove ecosystems of the Indo-western Pacific have a far greater number of plant species than do the mangrove swamps of Central America, and they certainly should be considered different ecosystems for conservation purposes. The nutrient-poor sawgrass (Cladium jamaicense) marshes of the Everglades differ profoundly from the nutrient-rich giant reed grass (Phragmites australis) marshes of the Camargue in France’s Rhone delta in land form, water chemistry, vegetation, and animal life. Neither of these marsh ecosystems bears much similarity to the thousands of small marshes that dot the prairie pothole region of Canada and the north central United States.

The montane cushion-plant herbaceous bogs of Chile contain flora and fauna quite distinct from that of the Sphagnum moss-dominated bogs of the Northern Hemisphere, and from the domed bogs of Sarawak. The shore pine (Pinus contorta) bogs of Vancouver Island’s Pacific coast contain many plant species not seen in Canada’s continental bogs. Seasonally flooded forests along rivers include ecosystems as dissimilar as the sometimes monotypic palm swamps of South America, the moderately species-rich hardwood swamps of the southeastern United States, and the highly species-rich varzea forests of the Amazon basin. Forested wetlands also include those occurring on steep windward slopes of Puerto Rico’s Luquillo Mountains where annual precipitation as high as 5000 mm keeps the clayey soils continuously saturated. In Finland, where about one-third of the land is peatland, at least 30 different mire types are recognized. Some of the bogs and fens of Ireland and Scotland, Scandinavia, Finland, Alaska and Canada, the former Soviet Union, and Maine form intricately patterned peatlands, some of the most unique land forms on earth.

A. Classification of the Major Wetland Ecosystems of the World

Because wetlands take so many forms, occur in many different physiographic and climatic settings, and form part of the history and culture of many regions of the world, their diversity is not easily cataloged. Different regional and colloquial names for them abound, as do formal classification systems. Different common names may refer to the same type of wetland with similar species, while the same name may be applied in different parts of the world to wetlands with quite different species. For example, the terms mire, moor, muskeg, and peatland all refer to similar types of wetlands with similar plant growth forms and many common species. On the other hand, the term swamp is used for a variety of forested wetlands in the United States but for areas dominated by giant reed grass (Phragmites) and other herbaceous species in parts of Europe and Africa. Fens dominated by woody species may be called swamps, forested fens, or forested wetlands. In the strictest sense, bogs are peat-accumulating wetlands that receive water
The high diversity of wetland ecosystems also is revealed by considering the many different bases for wetland classification: geomorphologic setting in the landscape, genera (how they formed), shape or form (e.g., surface morphology, basin morphometry, morphology of the underlying mineral terrain), hydrology (e.g., flooding frequency and duration), sources of water supply, physiognomy of the vegetation, species composition of the vegetation, soil type, water chemistry, and various combinations of these factors. Given that each of these variables represents not a single factor but rather a continuum, the extremely large number of possible combinations becomes apparent. Thus, most classifications in wide use today are broadly based, interdisciplinary, and hierarchical (i.e., consisting of several nested levels with increasing detail only at lower levels in the hierarchy). Species composition, the basis for diversity within wetlands and the focus of many older classifications, is usually not a defining characteristic except at the lowest levels in current classifications. Rather, similarities in hydrology and geomorphology are more likely to form the highest levels of classification, reflecting the importance of the interaction of water with the landscape in determining the biotic characteristics of wetlands.

Nonetheless, the terms bog, fen, marsh, and swamp persist in both scientific and popular literature because of their long history of use and the powerful images they evoke. These terms are not distinguished primarily on the basis of geomorphology or hydrology; nor are they distinguished on any single criterion, nor used consistently. Perhaps the key distinction among them in terms of species diversity is their source or sources of water, which determine hydrologic regime and water chemistry. Greater surface water inputs generally lead to greater water level fluctuations and higher concentrations of base cations and nutrients in marshes and swamps on mineral soils than in bogs and fens. In this chapter, these terms follow current American usage (see glossary).

The current classification systems used in the United States (Cowardin et al., 1979) and Canada (National Wetlands Working Group, 1997) reveal much about the diversity of wetlands in these two countries. For example, the U.S. classification is hierarchical, recognizing 5 major wetland systems and 10 subsystems, each of which is divided further into 4 to 8 classes. A system is defined as "a complex of wetlands and deepwater habitats that share the influence of similar hydrologic, geomorphologic, chemical, or biological factors" (Cowardin et al., 1979, p. 4). The five systems are the marine, estuarine, riverine, lacustrine, and palustrine. The term class "describes the general appearance of the habitat in terms of either the dominant life forms of the vegetation or the physiography and composition of the substrate—features that can be recognized without the aid of detailed environmental measurements." Classes can be further divided into dominance types based on dominant plant species for vegetated sites or on the dominant sedentary or sessile animal species where vegetation is not the dominant cover. Special modifiers are used for water regime (type and duration of flooding), water chemistry, soils, and human alterations (e.g., excavated, impounded, diked). Fifty-five classes emerge before one ever gets to the level of dominant species, water regime, or water chemistry.

Most of Canada's wetlands fall within just one system (palustrine) of the U.S. classification. Canada's classification differentiates many types of palustrine wetland, especially peatlands. This classification also is hierarchical but with just three levels: class, form, and type. Five classes are recognized: bog, fen, marsh, swamp, and shallow water. Forms are defined on the basis of surface morphology of the wetland, position in the landscape, morphology of the underlying terrain, tidal effects, and proximity to surface water bodies. The Canadian system recognizes 16 bog forms and 4 subforms, 12 fen forms and 7 subforms, 8 marsh forms with 18 subforms, 8 swamp forms with 19 subforms, and 5 shallow water forms with 24 subforms. The 18 types based on the general physiognomy of the vegetation (e.g., hardwood treed, coniferous treed, low shrub) modify these 121 forms and subforms. Differences in species composition are left to individual users.

Only the wetlands of western Europe are as well documented as those of the United States and Canada. Specific sites on other continents have received consid-
erable attention but, with few exceptions, comparative regional treatments began to emerge in the English language only 15 years ago, with most appearing since 1980 (see Bibliography). The exceptions include Chapman’s books—for example, Salt Marshes and Salt Deserts of the World (1960), Mangrove Vegetation (1976), and Wet Coastal Ecosystems (1977). However, despite the general lack of scientific studies in many areas of the world, as of 1991, 62 countries from all continents had signed the Convention on Wetlands of International Importance Especially as Waterfowl Habitat (Ramsar Convention) and had designated 527 sites covering more than 30 million ha as being of international importance. Five international wetland symposia have been organized since 1980 by INTECOL (The International Society of Ecology), with a sixth planned for the year 2000 in Quebec City.

B. Major Regional Wetland Ecosystems

Efforts to classify wetlands into unique entities often obscure, either spatially or temporally, critical connections among parts of what might be viewed as integrated ecological systems. For example, various classification systems segment the great delta of the Mississippi and Atchafalaya rivers into many types of floodplain forest, freshwater marsh, salt marsh, streams, and lakes. They assign each prairie pothole in the prairie region of the United States and Canada to one of about 20 classes of prairie wetlands. The mangrove forests of Central America are made to seem distinct from the adjacent beds of sea grasses. Functionally, however, these diverse wetlands are linked and might be better viewed as major regional wetland ecosystems. Water, sediments, plant propagules, fish, birds, other animals, genetic material, and pollutants move among them. One type changes through time to become another. Diversity emerges in both space and time from the components as well as their interconnectedness.

In fact, prior to human modification of them, most wetlands existed as complexes of several different types functioning as integrated systems linked by movements of water, sediment, nutrients, plants, and animals. The vast rain-fed and riverine floodplains that form the Sudd in southern Sudan, for example, consist of a complex and dynamic mix of different types of grassland (e.g., Echinochloa pyramidalis or Hyparrhenia rufa dominated), wild rice marshes (Oryza longistaminata), woodlands (Aaavia seyal or flanmites aegyptiacae dominated), and open water that cover more than 45,000 km². The great floodplain of the Orinoco River and its tributaries form a nearly 10-million ha wetland complex of streams, ponds, shallow oxbow lakes, riverine forest, marshes, and seasonally-flooded palm (Copernicia tectorum) savannas and grassland in Venezuela and eastern Columbia. The eastern shore of Lake Ontario, New York, is bordered by a 17-mile long interconnected system of streams, shallow ponds, marshes, fens, and shrub swamps. Most temperate zone estuaries include freshwater, brackish, and saltwater marshes along with tidal creeks and mudflats. Mangrove swamps, saline lagoons, sea grass beds, and coral reefs form interconnected systems along shallow coasts in subtropical and tropical regions.

These complexes, or assemblages of wetlands within regions, represent another level of wetland diversity that only partially remains today. Current classification systems do not capture such regional ecosystems, though clearly they exist. Many people recognize wetland complexes such as the Everglades and the Prairie Pothole Region as regional ecosystems. Classification systems, however, only recognize hierarchy in types, not in assemblages of wetlands within regions. Current efforts by conservation organizations and government agencies to move to ecoregional mapping attempt to overcome this limitation.

C. Continental and Regional Diversity

Climate and geomorphology, along with biogeographic factors, control the diversity of wetlands on different continents and in different regions. Wetland development is promoted where precipitation annually exceeds evapotranspiration on a long-term basis. In such climates, wetlands develop in a number of landscape settings, including depressions in the land surface, areas where minimal land surface slope slows surface water
runoff, along breaks in the slope of the land surface, and even on slopes if precipitation greatly exceeds evaportranspiration. In semiarid, arid, and desert climates, wetlands occur far less frequently, but they are found (a) in areas with depressions or lowlying land along rivers or coasts where runoff converges from extensive uplands or mountains, (b) along coasts regularly inundated by tides, (c) in areas where permafrost or other impermeable layers below ground prevent or slow infiltration of surface water, and (d) in spring-fed oases.

In addition to topographic variation, variation in the composition and stratigraphy of surficial geological deposits promotes diversity in the types of wetland that occur within a region. The composition of subsurface materials plays a major role in determining the chemistry of water entering wetlands, and thereby the plant species composition of wetlands. For example, wetlands occurring in regions dominated by limestone deposits tend to have high-pH waters rich in base cations that promote the growth of distinctive rich fen vegetation. In contrast, waters draining areas with granitic bedrock tend to be low in pH and base cations; such areas more frequently contain bog-like vegetation. Areas with thick deposits of glacial till can contain a mix of rock types that, in turn, may support a variety of wetland types. A diversity of types in a given region also is promoted where surface topography and subsurface stratigraphy of geological deposits produce groundwater flow paths of varying lengths. Wetland water chemistry is strongly influenced not only by water source but also by the length of groundwater flow paths that enter the wetland and the number of different groundwater flow systems that feed the wetland. Different flow systems may contribute water with different chemistry because the water originates from different geological deposits: If the water originates from the same geological materials, longer flow paths may yield higher concentrations of solutes because longer flow paths generally correlate with longer groundwater residence time (i.e., longer contact time with geological materials).

Biogeographic factors also influence the number of wetland types occurring in different regions and continents. Aquatic plants generally are known for their remarkably wide geographic range, but many species have smaller ranges determined in part by biogeographic factors (i.e., the climatological, geological, glacial, and floristic history of a region). The distributions of these more restricted species help define distinctive biogeographic regions that contain wetlands with species compositions unlike those from other biogeographic regions. For example, shallow ponds that support many species of the Atlantic Coastal Plain flora might be classed broadly as emergent marshes, but this would miss their unique contribution to the diversity of North American wetlands. The flora of these coastal plain ponds, which has an unusual abundance of rare and uncommon species, bears little resemblance to the emergent marshes of other biogeographic regions.

While no systematic comparison has been made of wetland ecosystem diversity in different regions, the factors controlling regional wetland diversity are understood. All other things being equal, regions with the greatest climatic and geomorphologic variation are likely also to contain the highest diversity of wetland types. If wetland types are defined to the level of plant species composition, then biogeography enters into the equation. Thus, continental diversity will vary as a function of the number and combination of different climatic, geomorphologic, and biogeographic regions occurring on that continent.

II. DIVERSITY WITHIN WETLAND ECOSYSTEMS

A. Plant Species Richness within Wetlands

At the regional and global scale, wetlands clearly exhibit high diversity of types and an associated high diversity of plant species. Within-habitat (or beta) diversity, and between-habitat (or alpha) diversity, however, is highly variable among wetlands. Wetlands include some of the most species-rich plant communities in the world, as well as numerous examples of extensive stands dominated by a single species. Familiar examples of single-species wetlands include mangrove swamps, tidal marshes of salt marsh cordgrass (Spartina alterniflora or Spartina patens), cattail (Typha spp.) marshes, and papyrus (Cyperus papyrus) swamps. Species-rich herbaceous communities include calcareous fens with as many as 20 to 30 species of vascular plants and bryophytes per square meter. Intermediate examples include abandoned beaver ponds with 5 to 12 species per square meter and a range of coastal wetlands (bottomland hardwoods, wet prairies, fresh marsh, and salt marsh) from the southern United States with 1 to 17 species per square meter.

Change in plant species composition across gradients within wetland sites depends on the amount of change that occurs within a site in the physical and chemical variables controlling species distributions. A single small depression may encompass a wide range of water depths and the associated changes in species composition. Although adjacent zones may share some species, the deep-water end of the gradient, which may
be dominated by floating or submerged species, shares no species with the sedge- and forb-dominated wet meadow zone at the other end of the gradient. Depending on the slope and width of the floodplain, wetlands along major rivers may include areas that experience very different frequencies of flooding and consist of very different plant communities, from the open water and marshes of backwater sloughs to hardwood-dominated forests along natural levees. Erosion and deposition of sediments within the floodplain create surfaces of varying ages that support plant communities of different successional status and composition. Much of the species diversity in the Peruvian Amazon has been attributed to the heterogeneity created by long-term patterns of erosion, deposition, and plant succession.

In other wetlands, environment changes little over relatively large distances, for example, where the accumulation of peat has filled in basins and minimized topographic changes. Though the presence of hummocks and hollows introduces small-scale heterogeneity, mean water depth, pH, and nutrient availability change little as one moves across the bog surface. Plant species composition in plots spaced at great distances, but at equivalent microtopographical elevation, may be quite similar.

Although many studies have examined plant species richness in small plots, systematic comparisons are difficult to make because plot size and sample number vary among studies. Given that species richness is known to increase with increasing sample area, conclusions cannot be drawn about relative richness among samples if sample areas are not equal. Nonetheless, some general statements can be made. First, the number of plant species present generally decreases as water depth and frequency of saturation or flooding increase. For example, in his study of Wisconsin plant communities, Curtis (1959) reported that the average number of species per stand ranged as follows: 7 in stands of submerged aquatics, 11 for emergent aquatics, 28 to 29 in sedge meadows, 44 in wet prairie, and 62 in wet-mesic prairie. Second, the effect of moving water on diversity depends on flood velocity, magnitude, and frequency. Some degree of disturbance from flooding increases diversity by removing live biomass and litter, thereby creating openings for germination of new species. Extreme or very frequent flooding, however, lowers diversity. Few species can withstand both the high stress of extended periods of anoxia and the frequent disturbance of biomass removal.

Third, the vegetation of saline wetlands tends to be less species-rich than freshwater wetlands. For example, coastal salt marshes of the eastern United States rarely contain more than two species in 0.25-m² plots. Plots of comparable size in freshwater herbaceous communities frequently contain 10 or more species. Increasing the plot size in salt marshes adds few species but is likely to significantly increase species number in many types of freshwater wetlands. For example, the number of vascular and bryophyte species in rich fens in New York increases from about 20 in 1-m² plots to 40 in 25-m² plots to 65 in 100-m² plots.

Fourth, in the temperate and boreal zones, alpha diversity generally increases across the range of peatland types from bog to rich fen for vascular species. Continental raised bogs in eastern North America may have only 20 to 25 vascular species per site, one of the most impoverished vascular floras in North America. Rich fens may contain as many as 140 vascular species at a single site. The bryophyte diversity also tends to increase from bogs to rich fens in eastern North America. Continental raised bogs usually have fewer than 20 species of bryophytes in 100 m² while rich fens usually have between 20 and 60 bryophytes in a similar-sized plot. However, Vitt and others (1995) reported that bryophyte diversity in peatlands of continental western Canada could be similar for bogs and fens. Species richness in larger plots (2500 m²) ranged from about 5 to 24 in bogs and poor fens and from about 5 to 28 in rich fens. Thus, even though the highest species richness was found in extreme rich fens, any individual bog could contain as many species of bryophytes as a rich fen. At the landscape scale, rich fens are distinctly more species rich than bogs and poor fens. Because their species composition varies more among sites, rich fens have greater diversity when an entire region is considered. In both Boreal and continental western Canada, bogs have only about half the number of bryophyte species as fens.

Fifth, relative to the landscapes in which they occur, many types of wetlands appear to be islands of diversity. Fens in northeastern Iowa support 28% of the regional and 18% of the total state flora even though they currently cover only 0.01% of the land surface. Boreal old-growth swamp forest communities constitute only 5% of Sweden's forest land area but harbor more than half of all vascular forest plant species found in Sweden and about a third of the total Swedish boreal flora of bryophytes. Almost a third of the federally listed threatened and endangered plant species in the United States are wetland dependent. Furthermore, for some types of wetlands, rare and uncommon plant species tend to be associated with the most species-rich sites.

Finally, latitudinal patterns in plant species diversity of wetlands do not appear to parallel those for terrestrial environments. In sharp contrast to terrestrial habitats,
wetlands in the tropics are not necessarily more species rich than those in the temperate and boreal zones. According to Ellison (in press), the mangrove swamps, palustrine forested wetlands, palm swamps, and hardwood swamp forests of Central America all exhibit low species diversity. On the basis of a comparison of representative aquatic families and aquatic habitats, Crow (1993) also concluded that aquatic plant diversity was higher in both warm and cool temperate latitudes than in the tropics. On the other hand, some tropical wetlands can be highly diverse. Ellison (in press) reported high plant species richness (292 species of 0.5 m in height or greater) in riparian forest fragments in Belize. Additional systematic and standardized studies are needed to address this issue.

B. Major Groups of Wetland Animals

In terms of abundance, biomass, and species richness, macroinvertebrates are the most important wetland-dependent animals in freshwater wetlands. Four groups—the insects, mollusks, crustaceans, and annelids—make up the majority. The aquatic insects are the most functionally and taxonomically diverse, and include forms that occur throughout all habitats (the bottom, deep waters, and shallow, vegetated areas) in most wetlands. The midges (Chironomidae, Diptera) constitute perhaps the most widespread group of aquatic insects, and larval forms of midges typically represent the most abundant macroinvertebrates in wetlands. Other ecologically important, widespread groups include the mosquitoes, dragonflies and damselflies (Odonata), mayflies (Ephemeroptera), caddisflies (Trichoptera), stoneflies (Plecoptera), and a large number of other groups. Among these groups, only the water beetles are entirely aquatic; members of other groups typically have aquatic larval stages that undergo a synchronized transformation (“hatch”) into adult forms, which may live for periods ranging from 1 to 2 days (some mayflies) to weeks or months (many dragonflies).

The mollusks include the many species of filter-feeding clams and herbivorous snails. The most widespread and common are fingernail clams. Larger clams and mussels are equally widespread but less common, although they are sometimes abundant along large rivers and lake fringes. Most large clams dwell on the bottom of wetlands, whereas the fingernail clams often occur in submerged vegetation. Snails also occur in many wetlands and live upon stems, stalks, and leaves of aquatic vegetation, and graze on epiphytes.

The crustaceans, like the aquatic insects, are commonly encountered throughout most freshwater wetlands. Especially common in open-water and upon aquatic vegetation are isopods and amphipods. In protected open water areas, small zooplankters, including cladocerans and copepods are abundant. Crayfish are common bottom-dwelling crustaceans in many wetland ecosystems. The last major group of aquatic invertebrates, the annelids (mainly oligochaetes), are not particularly diverse taxonomically. However, they can be extremely abundant in some wetlands, where they burrow in the wetland bottom or attach to aquatic vegetation.

Vertebrate animals represent another important component of the wetland fauna and include the amphibians, reptiles, fishes, birds, and mammals. The amphibians are composed of two main groups, the frogs and toads, and the salamanders. Most species of frogs and toads have complex life cycles that include an aquatic, usually herbivorous, larval stage adapted for rapid growth (the tadpole stage), and a terrestrial, carnivorous adult stage adapted for dispersal among wetlands. Their relative use of wetlands versus uplands varies among groups. For example, the major group of frogs in North America, the Ranidae, includes (a) members that may be mostly terrestrial but return to wetlands to breed and sometimes to overwinter (e.g., members of the leopard frog complex, e.g. *Rana pipiens*) and (b) other species (e.g., bullfrogs, *Rana catesbeiana*) that are primarily aquatic and typically leave wetlands only to disperse to new areas. Among North American salamanders, all are wetland dependent except for members of the Plethodontidae, although even this family includes the Desmognathinae, a large group of stream-
associated species. In salamanders, wetland dependency ranges from strictly aquatic (e.g., mudpuppies [Necturus]), hellbenders [Cryptobranchus], and pickerel (Esox) to semiaquatic in groups that spend much of the growing season in wetlands but overwinter on land (e.g., newts [Salamandridae]), to primarily terrestrial species that return to wetlands only to breed (e.g., the vernal pool-breeding mole salamanders [Ambystomidae]).

In terms of numbers, species richness, and biomass, the majority of wetland fishes are small forage fishes. In North America, the primary groups are the killifishes (Fundulus), shiners (Notropis), sunfishes (Lepomis), and mosquito fishes (Gambusia). In deeper-water wetlands or wetlands connected to permanent water bodies, larger, bottom-feeding species occur, such as Ictalurid catfish (e.g., bullheads) and carp (Cyprinus carpio), as well as carnivorous species such as pickerel (Esox), perch (Perca), and bass (Micropterus, Morone).

In North America, the primary groups are the killifishes (Fundulus), shiners (Notropis), sunfishes (Lepomis), and mosquito fishes (Gambusia). In deeper-water wetlands or wetlands connected to permanent water bodies, larger, bottom-feeding species occur, such as Ictalurid catfish (e.g., bullheads) and carp (Cyprinus carpio), as well as carnivorous species such as pickerel (Esox), perch (Perca), and bass (Micropterus, Morone).

Turtles are the most diverse group of wetland reptiles. Most turtles are highly aquatic and leave water only to lay their eggs and to disperse to new habitats. Commonly encountered, highly aquatic turtles in North America include the snapping turtle (Chelydra serpentina), mud and musk turtles (Kinosternidae), and softshells (Trionychidae). The Family Emydidae includes many other highly aquatic turtles, such as cooters, sliders, painted turtles (Pseudemys and Chrysemys), map turtles and sawbacks (Clemmys), and terrapins (Malaclemys), as well as other partially aquatic species that travel regularly among disjunct wetlands (e.g., spotted turtle [Clemmys gutata]) or that use wetlands only for hibernation (e.g., wood turtles [Clemmys insculpta]). Other wetland-dependent reptiles include the water snakes (e.g., Nerodia) that use wetlands primarily for feeding on fish, frogs, and crayfish, and rest above water on hydrophytes or on land at other times. Most wetland-dependent snakes are viviparous, that is, bear live young, and therefore do not undertake migrations to terrestrial habitats to lay eggs, as do turtles. Many other snakes, however, and some lizards, live primarily in uplands but feed opportunistically in wetlands (e.g., king snakes [Lampropeltis], rat snakes [Elaphe], and garter snakes [Thamnophis]).

Several large groups of birds occur nearly exclusively in wetlands. These include many carnivorous (mainly fish-eating) species, such as the loons (Gaviidae), herons and bitterns (Ardeidae), several waterfowl species (Anatidae), the kingfishers (Alcedinidae), terns (Sternidae), and several raptores (Accipitridae, e.g., ospreys Pandion haliaetus). Omnivorous groups include the grebes (Podicipedidae), many species of diving and dabbling ducks (Anatidae), cranes (Gruidae), rails (Rallidae), and gulls (Laridae). Primarily insectivorous species include the shorebirds (Charadridae) and many songbirds. Unlike other wetland inhabitants, birds are highly mobile and often use disjunct wetlands seasonally or even daily.

The last group of wetland-associated vertebrate animals, the mammals, include few, wholly wetland-dependent species. Most prominent are rodents such as the muskrat (Ondatra zibethicus) and beaver (Castor canadensis). Other obligate wetland mammals include several primitive, carnivorous shrews (‘‘water shrews’’ in the genus Sorex), some lagamorphs such as the swamp rabbit (Sylvilagus aquaticus) and marsh rabbit (S. palustris), and muskrats such as the river otter (Lutra canadensis). Wetlands are a critical resource, however, for many otherwise terrestrial species, which use wetlands for feeding and cover, such as moose (Alces alces), raccoon (Procyon lotor), and mink (Mustela vison).

Animal communities of saltwater wetlands can differ substantially from those in freshwater systems because of the effects of salinity on animal metabolism. Generally speaking, analogous taxa occur in freshwater and saltwater wetlands, such that similar niches are occupied, but species richness of animals generally declines along a freshwater-brackish-saltwater gradient. Reptiles from the Gulf of Mexico region provide an example: about 24 species occur in fresh marsh, 16 in brackish marshes, but just 4 in salt marshes. Of the salt marsh species, few are salt marsh specialists. Some terrapins are largely restricted to salt marsh habitats, but other regularly encountered salt marsh inhabitants, such as cooters and alligators, actually prefer freshwater situa-
tions. Amphibians show a particularly sharp trend along the wetland salinity gradient. Owing to their highly permeable skins and inability to counteract the drying conditions produced by high osmotic pressures of salt water, very few species can survive in saline wetlands and even then occur only infrequently (e.g., some toads). Relatively few breeding birds specialize on salt marsh habitats, although sparrows and rails are well represented. However, salt marshes are noted for their importance as stopover sites for the vast populations of migrant species that breed in freshwater wetlands that freeze in winter. Waterfowl, shorebirds, songbirds, and owls can occur in brackish and salt marsh habitats in large numbers during the colder months resulting in very diverse, if unstable, assemblages during the winter period. Thus, though occupied for relatively short periods, and not by a large number of breeding species, coastal marshes nevertheless are a critical habitat in the annual cycle of many bird species.

Mammal communities in saltwater wetlands are generally quite species poor. However, the occasional effects of high densities of aquatic rodents (e.g., muskrats and nutrias) can be of ecological significance. As in freshwater systems, these species can radically alter the wetland environment through so-called "eat-outs" of the vegetation. Otherwise, saline wetlands are infrequently visited by species more dependent on freshwater wetlands, such as otter, mink, or raccoon.

Overlap of freshwater and marine species, in conjunction with resident species that can tolerate the substantial fluctuations in salinity, produces highly complex, species-rich assemblages of fish and crustacean species in saltwater wetlands. Though characteristically species rich, these assemblages are temporally quite unstable because of daily tidal cycles, with freshwater-associated species descending into salt marsh systems on outgoing tides and marine species arriving on incoming tides. In particular, many marine fishes and crustaceans, especially shrimp, use salt marshes as nursery areas, particularly for post-larval and juvenile forms. Here they take advantage of the lower numbers of predators in salt marshes relative to the open ocean and have access to the high productivity of salt marshes.

C. Controls of Species Diversity

Within particular wetland ecosystems, plant and animal diversity reflects the interacting influence of a number of abiotic and biotic variables that operate across a range of spatial and temporal scales. Diversity in wetlands cannot be understood without attention to these gradients, the scales at which they operate, and their interactions. Of the physical and chemical variables, hydrologic regime, mineral ion concentrations, pH, nutrient availability, salinity, and disturbance have been shown to play particularly strong roles. They largely determine the number of species that are physiologically capable of living in a site. Important biotic variables include primary productivity and interspecific competition. Along with abiotic factors, they influence which plant and animal species actually occur within the site. Because the composition, structure, and density of vegetation form key elements of habitat for animal species, these gradients also strongly influence animal diversity. Depending on the type of wetland, different abiotic and biotic variables assume greater or lesser importance in determining species distributions. Hydrologic regime is important in all wetlands for both plant and animal species. Saturation with water severely restricts diffusion of oxygen into the soil while microbes rapidly deplete available soil oxygen. Hydrologic regime, therefore, largely determines the distribution and frequency of the most universal stress to which wetland plants and animals must adapt—oxygen in the water column, shallow soils, and rooting zone of plants. Salinity generally assumes importance only in coastal wetlands and in the saline wetlands of semiarid and arid regions. However, inland salt marshes occur in the Great Lakes region and some of the wetlands of the prairie pothole region of North America are saline. Mineral ion concentrations and pH affect species composition in all types of wetlands but have particular significance in peatlands that exhibit both extremes of the gradient in these factors. The combination of low nutrient availability, low pH, and low concentrations of calcium and magnesium makes bogs inhospitable environments for most plant and animal species. The extremely high pH and calcium concentrations of some rich fens are associated with high species richness and a large number of rare and uncommon species of plants and snails. Nutrient availability, which is lower in all undisturbed peatlands than in mineral soil wetlands, has become an issue for many types of wetlands as humans have increased nutrient inputs to wetlands.

Flowing water, especially at faster rates and larger volumes, exerts an additional mechanical stress on wetland species in some environments. Plants growing in river floodplains, tidal marshes, and other wetlands subject to flooding with rapidly flowing water must be able to withstand the force of moving water or must have life histories that allow them to germinate and reproduce in periods between floods. Well-developed root systems and flexible stems are essential for perennial plant species. Animals in these environments must be able to move out of the reach of flood waters, either into
Many types of disturbance—factors that remove plant biomass—affect wetland diversity. Extended periods of inundation with deep water kill most emergent plants, which allows new species to germinate when water levels drop. Muskrats and nutria eat marsh vegetation. Beavers fell trees, alter hydrologic regimes with their dams, and destroy existing vegetation in digging for mud to strengthen the dam. The flooding created by beaver ponds destroys flood-intolerant vegetation. Tides and storm surges deposit patches of debris that bury vegetation in coastal wetlands. Hurricanes rip trees out of swamps and deposit sediment and other debris in wetlands. Humans harvest timber from swamps in many parts of the world. In Africa, southeastern Asia, and parts of South America, humans harvest other wetland plants for various uses, including firewood, food, and shelter. Reeds (Phragmites communis) are still harvested in The Netherlands and England as thatching for roofs.

Change over time as well as space is the rule in wetlands. Because climate drives hydrology, seasonal, annual, and longer-term changes occur in water depth and all the chemical and physical factors it regulates. Extended periods of wet or dry years dramatically change the distribution of both plant and animal species. Some wetlands systems, such as the prairie potholes of North America, have 10- to 20-year cycles of wet to dry climate. In other regional wetlands, such as the deep-water cypress swamps of the southeastern United States, wet to dry cycles occur over much longer time periods. The effect of this temporal variation at many scales is to increase overall species diversity by promoting a dynamic community composition and spatial distribution among years. In wetlands where organic matter accumulates as peat, even more dramatic changes occur with time; the landscape itself is transformed. Over hundreds to thousands of years, a mineral-rich, deep-water pond with high pH and relatively high nutrient availability can change to a raised bog with low pH, low nutrient availability, and low mineral ion concentrations. The plants and animals of a pond are replaced with those of the bog. Long-term deposition of mineral sediments in estuaries can create similarly dramatic changes in topography, landform, and the hydrologic regime for wetland organisms.

1. Controls of Plant Species Richness

The question of what controls patterns in plant species richness is an old one that has received renewed attention with observed declines in species diversity over much of the globe. In wetlands, the question has been resolved only in part. The effects of many single factors are relatively well understood because of the over-riding importance of some physical and chemical variables in determining the pool of species physiologically capable of living in a site. Spatial and temporal heterogeneity, long recognized as critical determinants of species richness in other systems, are also positively correlated with species richness in wetlands. However, the relative importance of interacting sets of variables, operating in different types of wetlands and over different spatial and temporal scales, is only beginning to be understood. Because humans have increased nutrient inputs to many wetlands, the question of the relationship of plant species richness to nutrient availability, community productivity or biomass, and competition continues to motivate many wetland studies. Controls operating at regional and global scales have been little studied.

Gradients in several physical and chemical factors show strong correlation with species richness in wetlands, especially at the ends of the gradients where stress on plant growth is high. While more than 6700 species, approximately one-third of the total United States vascular flora, grow in wetlands, few species have evolved to persist in wetlands with low pH, high salinity, continuous inundation with shallow water, or high flooding frequency. Hence, bogs (low pH), the low marsh zone of salt marshes (high salinity), deep-water marshes (continuous inundation), and wetlands in the active channels of rivers (high flooding frequency) are all species-poor. In general, species richness increases with decreasing stress due to abiotic factors.

Numerous studies have shown that water level fluctuations are critical to maintaining species diversity in a wide variety of wetlands. Alternating periods of inundation and exposure of the soil create temporal...
Shot and spatial gradients of water depth that provide conditions for germination of the largest number of species. Species diversity always declines when water levels are artificially stabilized. Several studies have shown that coastal plain ponds, Great Lakes shoreline vegetation, prairie wetlands, and Carolina bay wetlands all require fluctuations in water depth to maintain overall species richness.

Spatial heterogeneity increases the number of species co-occurrences in wetlands through a number of mechanisms. The presence of hummocks and hollows in peatlands creates small-scale gradients in pH and moisture that allow species with a wider range of tolerances to coexist in relatively small areas. The presence of mounds and tree bases in floodplain swamps provides sites for germination of a number of species that do not germinate or persist in standing water. Old muskrat mounds serve the same function in cattail marshes. Microtopographic variation in riparian wetlands causes spatial variation in flood frequencies that results in higher species richness. Deposition of wrack (dead plant debris) by tides and storms increases species richness in salt marshes by producing openings in dense stands of Spartina; these openings have higher salinity than the surrounding vegetation and are colonized by more salt-tolerant species (e.g., Distichlis spicata). High variation in substrate pH, type, texture, and age creates microhabitat variation along arctic, boreal, and tropical rivers and increases the number of plant species occurring in riparian vegetation.

Some of the spatial heterogeneity associated with higher species richness in wetlands results from various types of disturbance. In general, richness first increases and then decreases along gradients of increasing disturbance so that richness is highest at intermediate points along the gradient. In many environments, lack of disturbance allows a few highly competitive species to displace other species through the process of competitive exclusion. Disturbance disrupts this process and allows other species to persist. However, richness decreases if disturbance is too frequent because few species are adapted to very high disturbance frequencies. Fertility, or nutrient availability, appears to be another strong control on plant species richness. However, few studies have actually measured nutrient availability because of the difficulty of doing so. Rather, nutrient availability has been inferred from surrogates such as community biomass, percent soil organic matter, or substrate texture. Use of these surrogates generally suggests that species richness is highest at moderate levels of community biomass and presumably, therefore, at moderate levels of nutrient availability. Close inspection of the data shows that richness can be low or high at low to intermediate levels of community biomass but almost always decreases above some threshold of high biomass. The amount of biomass at which threshold occurs, however, varies greatly among wetland types. Furthermore, the relationship between richness and biomass appears to be scale-dependent. When data from wetlands with different vegetation types are examined, richness follows a unimodal relationship with community biomass. Within small plots within types, the relationship breaks down.

The complex interplay of abiotic factors with biophysical variables such as competition and productivity has been examined in only a few freshwater wetlands and seldom through experimental manipulation of variables. Salt marshes have received more attention. Results from these studies emphasize that the relative importance and direction of the effect of biotic variables on species richness change along gradients of stress. Abiotic variables play the strongest role where stress is high but the presence of some species can ameliorate the stress for other species (e.g., by lowering salinity through shading, which decreases concentration of salts due to evapotranspiration). Species richness is thereby increased above what abiotic conditions might otherwise allow through positive interactions among species. Nonetheless, the number of species tends to decline monotonically with increasing stress from abiotic factors. Where stress in minimal, biotic control of species richness is greater. Here, variables related to density effects (e.g., productivity, biomass, light) are more likely to control variation in species richness, probably through competitive effects of dominant species on subordinate ones. In such environmentally benign wetlands, productivity is likely to be high and a few species are likely to dominate unless the process of competitive displacement is disrupted by disturbances that remove plant biomass and allow other species to establish.

The interplay of these local biotic and abiotic factors is played out against the background of regional and historical controls on the species pool (i.e., the number of species available to colonize a site). While local and current conditions determine which plant species do colonize a site, the regional species pool sets the upper limit to the number of species that could colonize the site. The geological, evolutionary, and ecological history of a region, along with its area, controls this pool. Geological age, processes of speciation and extinction, geographic dispersal, and chance historical events all leave their imprint on the species pools of particular regions. Few attempts have been made to relate the species pool to the number of species actually co-
occurring in a site because of the difficulty of obtaining adequate data sets.

2. Controls of Animal Species Diversity
Patterns of diversity and productivity in wetland animals follow gradients strongly related to hydrologic regime, primarily because of its effects on the composition, structure, and productivity of wetland vegetation and on oxygen availability within the water column and wetland sediments. Oxygen levels most influence the distributions of wholly aquatic, nonmigratory wetland animals (i.e., the macroinvertebrates and fishes). The productivity and diversity of wetland invertebrates and fish, and, consequently, many water birds and mammals, are highest in semipermanent wetlands (e.g., those flooded during most of the growing season but dry otherwise). Most amphibians, however, occur in transient pools. Trends in reptile diversity and productivity along this inundation gradient are not clear.

Semipermanent and permanent water bodies are the primary habitats of fish, which generally lack the ability of many other wetland inhabitants to disperse overland or enter dormancy when wetlands dry. Thus, areas of sufficiently oxygenated water must occur in wetlands throughout the year to support fish populations, especially those of large-bodied species. Large, seasonally flooded wetlands with pools or channels that remain permanently flooded, and lakes with well-developed littoral vegetation tend to support the most diverse fish communities. Similarly, fish-eating birds and mammals frequent the habitats occupied by their fish prey. In contrast, temporary, unvegetated pools are the primary breeding habitats of amphibians, offering two main advantages. First, amphibian larvae are able to exploit a pulse of primary productivity that occurs shortly after pools fill with water. Second, temporary pools will not support fish, major predators on amphibian eggs and larvae.

Many other wetland animals exploit intermittently flooded wetlands. Waterfowl, particularly egg-laying females, frequently travel among shallow pools, ditches, and puddles to feed on the pulse of invertebrate productivity along this inundation gradient. Many water birds and mammals follow gradients strongly related to hydrologic regime, primarily because of its effects on the composition, structure, and productivity of wetland vegetation and on oxygen availability within the water column and wetland sediments. Oxygen levels most influence the distributions of wholly aquatic, nonmigratory wetland animals (i.e., the macroinvertebrates and fishes). The productivity and diversity of wetland invertebrates and fish, and, consequently, many water birds and mammals, are highest in semipermanent wetlands (e.g., those flooded during most of the growing season but dry otherwise). Most amphibians, however, occur in transient pools. Trends in reptile diversity and productivity along this inundation gradient are not clear.

Semipermanent and permanent water bodies are the primary habitats of fish, which generally lack the ability of many other wetland inhabitants to disperse overland or enter dormancy when wetlands dry. Thus, areas of sufficiently oxygenated water must occur in wetlands throughout the year to support fish populations, especially those of large-bodied species. Large, seasonally flooded wetlands with pools or channels that remain permanently flooded, and lakes with well-developed littoral vegetation tend to support the most diverse fish communities. Similarly, fish-eating birds and mammals frequent the habitats occupied by their fish prey. In contrast, temporary, unvegetated pools are the primary breeding habitats of amphibians, offering two main advantages. First, amphibian larvae are able to exploit a pulse of primary productivity that occurs shortly after pools fill with water. Second, temporary pools will not support fish, major predators on amphibian eggs and larvae.

Many other wetland animals exploit intermittently flooded wetlands. Waterfowl, particularly egg-laying females, frequently travel among shallow pools, ditches, and puddles to feed on the pulse of invertebrate productivity that occurs following flooding. These feeding pools often are remote from nesting areas at more permanent wetlands, where deep waters and dense vegetation protect nesting birds and their eggs from predators. Similarly, snakes and turtles often can be found in temporarily flooded wetlands, and seek refuge from subsequent low-water periods by migrating or entering dormancy in a wetland's sediments.

Much of the variation in cover and food for wetland fauna, and hence in animal diversity, is associated with variation in the species diversity, density, and structural characteristics of wetland plants. Many wetland animals show affinities for particular life-forms of hydrophytes (e.g., submerged plants, floating-leaved plants, or emergents), and some are associated with specific plant taxa. For example, most aquatic invertebrates seek protection from predation and safe oviposition sites in macrophytes. In general, plants with dissected leaves support more invertebrates than plants with broad leaves. Macroinvertebrate diversity and density generally are greater in vegetated than open water areas, and, within vegetated areas, are greatest in areas of mixed emergent and submerged vegetation. Shading by erect hydrophytes also influences macroinvertebrate abundance; unvegetated areas of submerged vegetation often support dense populations owing, in part, to elevated water temperatures in unvegetated areas. Similarly, shallow, vegetated wetland habitats fringe deep-water areas as critical nursery habitats for fishes, many species of which use stands of particular wetland plant species as spawning habitats. These areas provide protective cover for eggs and larval fishes and food resources for young fishes. Like aquatic insects, many fishes prefer unvegetated, shallow areas where warm water temperatures hasten the development of eggs and larvae.

Two key water chemistry parameters that influence distributions and diversity of wetland animals are acidity and salinity. The relationship of acidity to the distribution of wetland animals can be complex. For example, many aquatic, fish-eating birds frequent wetlands with moderately acidic waters (pH 5.5–6.0). Although fish populations often are reduced in wetlands with waters of pH <6.0, suspended material may precipitate from the water column at low pH. The resulting increase in water transparency may improve encounter rates with prey and capture success for birds that pursue their prey underwater, and thereby compensate for decreased fish density. The reduction or loss of fish from moderately and strongly acidified waters also reduces competition between fish and insectivorous birds for macroinvertebrate prey. This may explain why dabbling ducks frequent waters of relatively high acidity that fish-eating ducks avoid. Highly acidic waters (i.e., pH <5.0) generally support few fish or amphibians and have low diversity of aquatic invertebrates. Water bird and mammal use of such wetlands is limited.

Salinity strongly limits animal diversity in wetlands, primarily through effects on the composition of plant and invertebrate communities and through its direct physiological effects on wetland animals. Difficulty in maintaining water balance leads many dabbling ducks, particularly those brooding young, to avoid wetlands
with highly saline waters. Owing to their highly permeable skin, amphibians are intolerant of even moderately saline conditions. For similar reasons, highly saline conditions often limit the survival and hatching success of fish eggs and survival of larval fish. Many mammals, such as muskrats, are tolerant of highly saline conditions, but populations often are limited by the low quality of forage plants that grow in saline wetlands.

Three spatial characteristics of wetlands and their distribution in the landscape—size, connectivity, and isolation—also influence the diversity of wetland animals within and among sites. For example, small wetlands often support fewer species of birds than large wetlands, and isolated wetlands support fewer species than wetlands in complexes. Aquatic, fish-eating birds are particularly sensitive to wetland area and often shun small wetlands (<10 ha). Small wetlands may not provide a sufficient quantity of habitat or prey for many large-bodied species, and short inter-wetland distances may provide certain species with alternate foraging sites while minimizing their time in flight. Larger wetlands generally support a wider diversity of vegetative life-forms and, therefore, may provide a greater variety of habitats for wetland species than smaller wetlands. This may account for the frequently observed pattern of increased fish species diversity in larger than smaller wetlands.

The proximity of diverse habitat types within a wetland, and connectivity among wetland environments and with upland habitats, is critical to maintaining high diversity of wetland animals, particularly amphibians, some birds, and fish. Aquatic breeding and terrestrial nonbreeding habitats must be contiguous for amphibian populations to persist. Local breeding populations undergo frequent extinction and recolonization events, and individuals often are exchanged among populations. The importance of migration in amphibian population ecology is indicated by the presence in many species of a juvenile stage adapted strictly for dispersal. Destruction of larval or adult habitats, or the connection between them (e.g., riparian corridors) often drives local amphibian populations to extinction.

Proximity of diverse habitat types within a wetland is critical to many breeding water birds and fish. Some water birds, such as rails, dwell within dense stands of emergent vegetation, while others, such as bitterns, herons, and shorebirds, forage mainly along the interface of wetland vegetation and open water. Still others, such as grebes and many dabbling ducks, are associated with floating and submerged vegetation, while loons and cormorants forage in deep, open waters. Most of these birds, however, build their nests within dense stands of emergents, on small islands, or in trees or shrubs.

Fish assemblages provide another example of the importance to wetland animals of spatial heterogeneity and connectivity in wetland habitats. Most large species of bottom-feeding and predatory species undergo small-scale migrations between feeding and spawning areas in shallow wetlands and resting areas in deeper, open-water habitats. These migrations often occur nightly, with fishes returning to deep waters to rest during the day. Critical to these migratory movements are the links established between shallow, vegetated areas and deep-water zones by seasonal regimes of wetland flooding. Finally, many fishes undergo seasonal migrations over hundreds of kilometers into river floodplains and return to the river channels as waters recede.

Some wetland animals create some of the spatial and temporal heterogeneity that promotes animal diversity in wetlands. Beavers alter stream channels and transform riparian forests into freshwater marshes and ponds with stands of dead timber. Beaver activities also permanently modify soils beneath dam sites and even change local topographies; beaver dams reduce current velocities and cause massive silt depositions, which remain long after the dams disintegrate. High levels of herbivory by muskrats and geese can result in “eat outs,” whereby essentially all emergent plant material in a marsh is consumed over a few years. Such drastic, animal-caused modification of the wetland environment not only affects habitat availability for other wetland animals but can alter patterns of energy flow and biotic productivity within wetland ecosystems. Other, less drastic examples of habitat modification include the activities of alligators, which construct small basins that serve as critical refuges for many fishes, and important feeding areas for fish-eating birds, during low water periods. Crayfish burrows serve in many areas as moist refuges for water snakes, salamanders, and frogs.

### III. FUNCTIONAL DIVERSITY

Wetland plants exhibit a diverse array of adaptations that allow many species to tolerate stresses associated with soil saturation, alternating periods of flooding and drought, high salinity, low pH, and low nutrient availability. Other adaptations assist wetland plants in dispersal and establishment. Animals also exhibit an array of adaptations to wetland environments, although un-
like plants, many simply can move to escape stresses. Collectively, these adaptations explain much of the high species richness and diversity of many wetlands, especially those with strong hydrological and chemical gradients.

A. Adaptations to Anoxic Environments

Many wetland plants have one or more morphological and anatomical adaptations that allow them to tolerate soil saturation and anoxia for short to long time periods, primarily by allowing more oxygen to reach the plant root system. Key morphological adaptations include (a) aerenchyma, air spaces in roots and stems that allow oxygen diffusion from stems above water to roots; (b) hypertrophied lenticels, enlarged openings in stems and roots that allow gas exchange between internal plant tissue and the atmosphere; (c) adventitious or stem roots that allow gas exchange between internal plant tissue and the atmosphere; and (d) the ability to grow new roots under anoxic conditions. In black mangrove (Avicennia), pneumatophores, vertically growing air roots, absorb oxygen that is transported to the connected, submerged, lateral growing roots. Prop roots of red mangrove (Rhizophora) function in much the same way. In both mangrove species, oxygen enters the plant through lenticels exposed above water and diffuses to roots in anoxic sediments. Many woody species of alluvial floodplains have extensive, shallow root systems placed where sediments are least likely to experience oxygen deficits.

Plant physiological adaptations generally involve tolerance to low soil oxygen and specialized chemical reactions. Specialized reactions include an accumulation of malate instead of ethanol, the production of high levels of nitrate reductase, and a reduction in ethanol production by reducing alcohol dehydrogenase activity. Additionally, rhizospheric oxygenation is an important mechanism for many species and can lessen toxic concentrations of some anaerobic soil compounds.

Wetland animals, with their characteristically high metabolic rates, have developed a variety of adaptations to low levels of oxygen and carbon dioxide. Perhaps most obvious is development of specialized regions of the body for gas exchange. Examples include gills in fish and crustacea, parapodia in polychaetes, and highly vascularized tissues on the lower lips of some tropical fishes or in the clypeus (uro-genital openings) of turtles. Other physical adaptations include modification of respiratory pigments to improve oxygen-carrying capacity in invertebrates. For example, midge larvae (Chironomus) are often colored brightly red, indicating the unusually high concentrations of hemoglobin present in these organisms, permitting them to survive long periods of hypoxia. Some vertebrates, particularly fishes, also increase densities of circulating red blood cells and thereby their oxygen-holding capacity.

Physiological adaptations of animals primarily involve shifts in metabolic pathways. For example, bi-valves use alternative biochemical pathways, primarily a switch to glycolytic fermentation, to increase energy production under anoxic conditions. Some invertebrates also diversify the by-products of glycolysis to avoid toxic accumulation of any single compound, particularly ethanol. Turtles are remarkable for the ability of these lung-breathe to remain under water submerged in sediments for months during the winter season. Biochemical adaptation for natural anoxia tolerance in turtles includes well-developed antioxidant defenses that minimize or prevent damage by reactive oxygen species during the reoxygenation of organs after anoxic submergence. Last, many invertebrates store large quantities of respirable carbohydrate, usually glycogen, for breakdown and oxygen-liberation during periods of anoxia.

Behavioral adaptations also are critical and widespread, including dormancy or low locomotor activity during periods of oxygen stress, and migration from hypoxic to oxygen-rich environments. Many animals in low-oxygen situations have developed means of moving water more rapidly across respiratory surfaces. For example, benthic animals often use a variety of behavioral means (lanning, retreating into and out of burrows) to ventilate their burrows and increase the flow of water across membranes during times of hypoxia. These adaptations combine in interesting ways in particular taxa. Some fishes highly adapted to mud and shallow-water swamps cope by aestivating in mucous cocoons or by migrating overland while air breathing through modified swim bladders (e.g., lungfish). Some aquatic microinvertebrates, such as Colembola (springtails) and mites (Acari), have developed a "physical gill" that traps air gathered at the water surface in body surface hairs. Breathing of the trapped air, while under water, occurs via a tracheal system, which opens to the body surface. Some aquatic insects, including mosquito larvae and chrysomelid beetles, tap the air within the aerenchyma of plant roots using a highly specialized, spinelike siphon attached to their abdomens. Last, some fly larvae use snorkel-like devices that extend above the surface of liquid mud or anoxic water and that permit the animal to air-breathe while remaining submerged in the anoxic substrate.
B. Adaptations to Salinity and Drought

Although halophytes (i.e., plants capable of persisting in saline environments) are not restricted to wetlands, these species dominate saline wetlands, such as inland and coastal salt marshes, and coastal fringe forests. Some species, such as various mangroves (Avicennia, Laguncularia, and Rhizophora spp.), are facultative halophytes and have mechanisms to inhibit salt absorption by their roots and to excrete salt from their leaves. Because nonhalophytes ("glycophytes") exhibit a range of tolerance to a salinity gradient, salinity and flooding regime determine plant species composition and richness along a gradient from salt marsh to tidal freshwater marsh. Plant species that use the C₃ pathway of photosynthesis are well adapted to habitats subjected to drought stress. Saline wetlands, therefore, often are dominated by C₃ plant species because the salt content makes water uptake more difficult. Plant species with C₄ photosynthesis are much more efficient in fixing carbon and in water use and have higher rates of net photosynthesis, growth, and dry matter production than C₃ plants.

Little is known about adaptations of wetland animals to salt stress. The fundamental problem of salt stress is its effect on water movement at the cellular level, which interferes with metabolic processes. Most simple microinvertebrates are osmo-conformers whose internal salt concentration tracks that of the external environment. More complex wetland animals, particularly vertebrates, typically osmoregulate—that is, retain internal salt concentrations independent of those external to the body. Osmoregulation primarily involves moving ions across body membranes against the concentration gradient and is thus an energy intensive process. This process is usually associated with specialized renal organs (kidneys and antennal glands) and salt-secreting nasal or rectal glands. Behavioral adaptations, particularly short-term migrations, are common in mobile species. In such a manner, both freshwater and marine species can occur in salt marshes, retreating during unfavorable periods either upstream (freshwater species) or downstream (marine species) during the course of the daily tidal cycle.

C. Adaptations to Low Nutrient Availability and High Acidity

Nitrogen fixation is relatively uncommon in wetland vascular plants but the few examples are noteworthy. Nitrogen fixation by sweet gale (Myrica gale) can contribute 3 to 4 g N m⁻² yr⁻¹ to Sphagnum bogs. Nitrogen fixation rates of speckled alder (Alnus rugosa = A. incana ssp. rugosa) have been measured at 80 to 167 kg N₂ ha⁻¹ yr⁻¹. Some nonsymbiotic, aerobic and anaerobic bacteria, and blue-green algae are also significant nitrogen fixers in wetlands.

Most of the carnivorous species in the world occur in wetlands and other oligotrophic habitats. The relationships between insect capture, mineral nutrition (particularly nitrogen and phosphorus), and plant growth and reproduction have been investigated for numerous species. The relatively large diversity of organisms that can survive in the "pitchers" of some carnivorous plants also has been studied. Although benefits to carnivorous plants have been demonstrated, there is limited evidence that carnivory is necessary for most species to survive.
in wetlands is uncertain, but it has been suggested that plant species with evergreen leaves may be more efficient in using nutrients than species with deciduous leaves. These leaves, which persist for up to four years, extend the period that the plant can photosynthesize, which conserves energy in the plant because all of the leaves do not have to be renewed each year. Furthermore, evergreen species tend to respire more nutrients from senescing leaves than other species, thus increasing nutrient use efficiency. High nutrient resorption and long retention of nutrients in evergreen sclerophyllous leaves may be especially important in phosphorus-poor habitats such as bogs.

One of the most important groups of plants in colder, generally acidic wetlands of the world are Sphagnum mosses, a large genus consisting of more than 135 species. These mosses can thrive in conditions highly stressful to most plants (e.g., low pH, low nutrient availability, and saturated soils). They are the primary peat producers worldwide and have the capacity to modify the environment in ways that favor their own growth. Once established, Sphagnum mosses can further acidify their environment through cation exchange processes and by excretion of polygalacturonic acids. They also paludify their environment through their high water-holding capacity and their role in peat accumulation which can block water drainages. Sphagnum species also can tolerate great desiccation. Three major gradients exert the greatest influence on the bryophyte flora of wetlands: acidity–alkalinity, dry–wet, low light–high light. Numerous Sphagnum species can occur even within a relatively small peatland because of the affinities of particular species to very specific pHs, other water chemistry variables, height above water table, and light.

D. Adaptations to Reproducing in Wetlands

Seeds of many wetland species have features that enable their dispersal by water (hydrochory). Red mangrove (Rhizophora mangle) produces seeds that germinate while on the parent tree, resulting in viviparous seedlings. The germinated seed can root if it lands on sediment. Otherwise, the seedling floats until a suitable substrate is reached.

Seed banks, the store of seeds that accumulate in soils, play a critical role in the long-term vegetation dynamics of many types of wetlands worldwide. Seed banks allow wetland vegetation to adjust to rising and falling water levels and help maintain compositional diversity across water depth gradients and from year to year and over longer climatic cycles. Most wetland species do not germinate under water and only a few survive more than a year or two of continuously deep water. A series of wet years can eliminate most vegetation, but when water levels fall, many species that existed only as seeds in the seed bank germinate in the wet exposed mud. Thus, fluctuating water levels and seed banks drive wetland succession. Seed banks are particularly important in maintaining the species composition and zonation of wetlands that have a large number of annual species (e.g., tidal freshwater marshes). An understanding of the seed bank is essential in wetland management, especially in restoration efforts.

Seed density in wetland soils, usually determined by germination tests, typically ranges from 1000 to 14,000 seeds m⁻² with density and viability decreasing with increasing depth in the substrate. Graminoids (grasses and sedges) are the most abundant group in wetland seed banks, with seeds of broad-leaved herbaceous and woody species far less abundant. Although the composition of the seed bank and plant species present in the wetland can be highly correlated, in many cases the seed bank harbors a much greater species richness than found in the extant vegetation. Seeds of some species are transient, while others can persist for decades.

The greatly enlarged, rounded bases of water tupelo (Nyssa aquatica) and buttressed bases of bald- and pond cypress (Taxodium spp.) develop especially in deeply flooded conditions. These swollen bases likely aid in anchoring the tree. The “knees” of Taxodium also may function in this manner. Taxodium buttresses appear to provide a refugium above water for numerous species of herbaceous and woody plants that become established in the crevices where organic materials have accumulated over time. In some deep-water swamps, alternative colonization sites are limited for many plant species.

IV. THREATS TO WETLAND DIVERSITY

Worldwide, the loss of wetlands is estimated at about 50%. Functionally, the losses are much greater because extensive areas of boreal wetlands have been little altered while many different types of regional wetlands have been reduced by 80 to 90% elsewhere (e.g., California, southern Ontario, southwestern France, New Zealand). Wetland diversity and numerous plant and animal species have been impacted by a host of anthropogenic activities and natural disturbances for centuries. While single and cumulative effects of these activi-
ties have been assessed for many wetland species (primarily wildlife) and functions, few studies have examined the effects of these activities on species diversity. Because human activities often modify wetland hydrology and water chemistry, one should expect significant changes in diversity given the well-established relationships between wetland hydrology, water chemistry, and species distributions.

Despite the relatively recent appreciation of wetland functions and values, wetlands continue to be destroyed. A notable example is the Pantanal of Brazil, the world’s largest wetland and home to many endangered or threatened plant and animal species. Various development plans, including dam and highway construction, threaten the diversity of this vast wetland. Although some human activities may be able to be carried out in a sustainable manner in wetland habitats, habitat preservation is the key to preserving wetland diversity around the world.

A. Direct Habitat Destruction

By far, the majority of palustrine wetland loss in North America has been due to agricultural activities, especially direct conversion of wetlands to agricultural land. Agricultural uses have converted over 80% of the Mississippi River bottomland hardwood forest. About 90% of the wetlands in California, Ohio, and Iowa have been destroyed, primarily for agriculture. Wetlands adjacent to agricultural lands also can be adversely affected because drainage effects can extend a significant distance into the wetland. Coastal wetlands, most first altered by agricultural activities, including grazing and salt production, are now impacted by draining and filling for urban and industrial development. Drainage of wetlands on organic substrates leads to accelerated decomposition, erosion of organic matter, and release of carbon dioxide (a “greenhouse” gas) into the atmosphere.

Riparian wetlands in the western United States can be seriously degraded by cattle grazing, which adversely affects the adjacent aquatic ecosystem. Wildlife species composition and diversity are modified because heavy grazing reduces the structural complexity of riparian wetlands.

Timber harvest, when done without extensive drainage and total land clearing, generally converts forested wetland to earlier successional stages, or changes species composition and structure. Logging, specifically the extraction of cut trees through mechanical skidding and transport of heavy log loads, adversely modifies soil physical properties. Early logging of southern forested wetlands was typically done with no regard for sustainability. Consequently, forest composition and structure were dramatically altered. Logging roads can modify wetland hydroperiods such that productivity is raised or lowered. Occasionally, especially in northern forested wetlands, a heavy timber harvest can raise the water table, flooding already established species and decreasing the regeneration sites for new plants. Whole-tree timber harvest of oligotrophic peatlands can greatly reduce a site’s nutrient pool.

Tropical intertidal zones are being increasingly altered for mariculture, especially mangrove wetlands for shrimp farming. Many salt marshes have been ditched and diked, and often converted to impoundments, to favor waterfowl.

Peat extraction, primarily for fuel or horticultural uses, has destroyed millions of hectares of peatlands in North America, Great Britain, and northern Eurasia. Where the mined peatlands have been abandoned or actively managed for conservation values, the vegetation can slowly recover. Which species come to occupy the site depends on the hydrology of the site, availability of propagules, and time since abandonment.

Gravel and sand mining can adversely affect wetlands directly by habitat destruction, and indirectly by changing the hydrology and water chemistry of an adjacent wetland. These mining activities can be especially damaging to ombrotrophic and minerotrophic peatlands (e.g., bogs and fens) because many of the characteristic species occur at very specific water levels and chemistry parameters.

B. Direct Alteration of Wetland Hydrology

Humans alter wetland hydrology primarily by building dams at wetland outlets to regulate the depth and duration of inundation, and by constructing channels and ditches to drain wetlands and prevent prolonged flooding. Artificially regulated water levels often result in impoverished animal communities because few species of wetland animals have life histories adapted to the stable water regimes that result. This decrease in diversity occurs even where water control structures were originally built with the intention of benefiting wetland animals (e.g., at wildlife refuges). Management strategies for many water impoundments now seek to emulate the “natural” flooding regimes that were present before impoundments were built. Wetland drainage has obvious and drastic effects on wetland animal communities and has resulted in a loss of about half of the original habitats available to the wetland fauna of the United States. Among the most imperiled wetland habitats are the easily drained, intermittently flooded basins and
pools that are critical to a large proportion of the wetland fauna.

Dam construction and water diversion projects also have significant effects on downstream riparian vegetation by altering peak and minimum flows, decreasing erosion and deposition of sediment, and lowering floodplain water tables. Altering the natural variability of river flow can adversely affect the species composition and structure of riparian wetland vegetation. Dredging, channelization, and levee construction have greatly impacted extensive wetland areas in the Mississippi River Delta.

Groundwater extraction is a primary cause of wetland modification because it lowers the regional water table and alters recharge and discharge patterns within the wetland and surrounding landscape. Irrigation with water supplied by aquifers that are recharged in part by wetlands has led to declines in the area of major regional wetlands in North America (e.g., the prairie pothole and Nebraska sandhills regions).

Although plant species of coastal wetlands may be well adapted to flooding, increases in the level or duration of flooding above that at which species have long persisted can cause serious deterioration of these wetlands. Sea-level rise and saltwater intrusion are important factors in the degradation and loss of wetland forests and coastal marshes in the southeastern United States. Increased flooding and salinity have been particularly damaging to forests dominated by baldcypress.

C. Landscape Fragmentation

Other human activities impoverish the wetland fauna in more subtle ways, even where wetlands receive nominal legal protection. Destruction of the uplands adjacent to wetlands destroys the habitat interface critical for amphibians and reptiles that migrate between wetland and upland habitats. Reductions in the connectivity among wetlands (e.g., by construction of roads and levees) blocks the migration routes of amphibians, fish, and birds, and can lead to local population extinctions. In contrast, dredging of canals that link major river systems causes sudden mixing of fish and reptile faunas that may have evolved in isolation for thousands of years.

D. Other Threats

Anthropogenic inputs of nutrients and toxins drastically alter the chemical environment of wetlands and render many sites unsuitable for wetland plants and animals. Atmospheric deposition of nitrogen and sulfur is particularly high in parts of western Europe but is also high in the northeastern United States. Atmospheric deposition may be especially damaging to peatlands dominated by bryophytes (mosses and lichens), because these species are very sensitive to changes in nutrients and acidity-alkalinity. Wetlands in agricultural landscapes receive excess nitrogen via ground water draining from agricultural fields and excess phosphorus primarily via surface water runoff. Residential development also contributes excess nitrogen and phosphorus to wetlands. Coastal oil spills can adversely affect the vegetation of salt marshes and tidal freshwater wetlands.

Invasive, exotic wetland species typically cause a substantial loss of native wetland species and greatly alter wildlife habitat. In otherwise arid regions, invasive species use substantial amounts of water and can desiccate watercourses. Invasive species generally respond favorably to altered hydrologic regimes, substrate disturbance, or changes in water quality, especially eutrophication. Introduction of exotic wetland animals, for example, carp, murrea (Myocaster coypus), and bullfrogs, outside their native range have greatly altered wetland environments throughout much of North America, mostly to the detriment of the native wetland fauna.

Purple loosestrife (Lythrum salicaria), an emergent, herbaceous species of Eurasian origin, has become a serious invasive species of open wetlands in eastern North America. A single individual of this species can produce an average of 2,700,000 seeds. Another very serious invasive species in wetlands of the eastern United States is the common reed (Phragmites communis), which can produce 200 to 300 culms m⁻² in fresh and brackish wetlands through an extensive network of rhizomes. Recent biological control methods (i.e.,
leaf- and root-feeding beetles) appear promising for
*L. salicaria*.

Salt cedars (*Tamarix* spp.) are rapid, arborescent in-
vaders of riparian wetlands in the southwestern United
States, especially downstream from dams where flood-
ing is minimal. Another invasive tree species of riparian
wetlands in the western United States is Russian-olive
(*Elaeagnus angustifolia*). In southern Florida, the inva-
sive tree, melaleuca (*Melaleuca quinquenervia*), has been
aggressively colonizing shallow wetlands. Two other
tree species that seriously threaten southern Florida
wetlands are Australian pine (*Casuarina*) and Brazilian
pepper (*Schinus terebinthifolius*). Glossy buckthorn
(*Rhamnus frangula*) is becoming a serious problem in
wetlands around the eastern Great Lakes basin.

Some native species become invasive following some
hydrologic changes or an increase in nutrients and dom-
inate wetlands previously occupied by other, more de-
sirable species. Cattail is regarded as a threat to southern
Florida marshes that have been dominated by sawgrass
(*Cladium jamaicense*). The spread of cattail is believed
due mostly to increased phosphorus input from agricul-
tural lands and anthropogenic changes to the regional
hydrology. Cattail is also considered a weed in many
wetlands and aquatic habitats that are used for agricul-
ture or to provide water resources (e.g., rice fields,
irrigation canals, recreational lakes, and reservoirs).
Mechanical (e.g., cutting, water level modification, fire,
shading), chemical (selective versus nonselective herbi-
cides), and biological control methods vary in their
effectiveness, which varies widely among species. Main-
taining water levels over 0.5 m is effective in controlling
spread of cattail.
E. Natural Disturbances

Severe flooding, hurricanes, and fire are the major natural disturbances that greatly affect wetlands. Although wetland plant and animal communities are adapted to each of these disturbances, events of exceptional magnitude and duration, or that occur during unusual times of the year, can have profound effects. A serious concern following any of these severe events is the spread of invasive species.

During the 1993 flooding of the Mississippi River, submerged aquatic, emergent wetland, and floodplain species were adversely affected by extreme sediment deposition, uprooting due to wave action, and probably increased inputs of agricultural chemicals. Recolonization of open spaces left by plant mortality can happen quickly by germination of local or upstream plant propagules.

Hurricanes especially affect the mangrove forests of the Gulf of Mexico and Caribbean, and the deep-water and alluvial swamps of the southern United States. These coastal forests, where many species have shallow root systems, are often subjected to the highest winds and succumb to windthrow. Trees that are especially resistant to windthrow (e.g., *Taxodium distichum*) may lose most of their crown but are able to reestablish a new canopy quickly by extensive sprouting along the bole. Seedling regeneration is generally higher after such disturbances because of increased light and microsite availability. Storm surges of saltwater occasionally cause greater damage than high winds to species intolerant of high salinity.

Forested wetlands on substantial peat deposits, such as the black spruce (*Picea mariana*) and tamarack (*Larix laricina*) swamps of the boreal forest and baldcypress swamps of the southeastern United States, have naturally burned during periods of extreme drought. In northern climates, fires can convert the forest to a more open peatland. In the southeastern swamps (e.g., Okefenokee Swamp), fires can convert forested wetlands to prairie-like, tree-less wetlands. Fire can favor the regeneration of some tree species, like *Chamaecyparis thyoides* (Atlantic white-cedar), which regenerate best under full sunlight conditions on saturated peat. In the Everglades of southern Florida, open sawgrass marshes naturally burned every 1 to 5 years, ignited by lightning strikes from May to October when water levels are highest and substrates are saturated. However, human-set fires are most common from November to May when soils can be dry enough to ignite. These fires, to which the vegetation is not well adapted, can cause substantial ecological damage.

V. RESTORING AND MAINTAINING WETLAND DIVERSITY

Because of extensive loss and degradation of wetlands throughout the world, and continuing loss and degradation in much of the world, maintaining wetland diversity will require a strategy that includes restoration. The ultimate goal in any specific wetland restoration project is to have a self-sustaining, functioning ecosystem. The following discussion focuses on techniques to achieve this goal. However, in terms of restoring wetland diversity, the goal must be a broader one—best addressed at the landscape, regional, and continental scale—to restore the structure, function, and self-sus-
taining properties of a wide diversity of wetland types. Without attention to restoration of the full complement of wetland types, wetland diversity cannot be restored at any but the smallest scales.

Restoring self-sustaining wetland ecosystems initially, and at later times, may require various engineering, horticultural, and wildlife management activities (National Research Council, 1992). The intensity of these activities depends on whether the site was initially an upland (thus requiring wetland "creation"), a wetland that has been drastically disturbed (e.g., through conversion to agricultural crop production), or disturbed to lesser degrees. In some cases, a wetland may have reached a successional stage that does not support a species or group of species of concern, and manipulation is required to maintain these species. Nontidal emergent wetlands are generally easiest to restore. Long-term monitoring of site hydrology and vegetation is required to determine if and when project objectives are not being met.

Because wetlands serve many functions, and seldom can all functions be restored in a single site, restoration objectives must be clearly defined for every project. Restoring wetlands as habitat for rare and uncommon species, or for overall biological diversity, especially requires clear statements of objectives, as well as thorough knowledge of the ecology of the species of concern. Wetlands that could be correctly described as functioning wetlands have failed to attract the species for which they were restored. Consider the restoration of a bottomland hardwood forest. The restoration could be viewed as successful if a certain percentage of planted oak (*Quercus*) species and individuals were to become established after a given period of time. However, such a planting would be a failure in terms of restoring the diversity of neotropical migratory birds, which would fail to use the site because they require a more complex forest structure. In general, a higher overall plant and animal diversity is achieved by planting an array of fast- and slow-growing plant species, and where appropriate a mix of woody and herbaceous species tolerant of the range of hydrological and chemical conditions at a site.

The first step to restoring wetland plant composition, structure, and function is to ensure a range of hydrological conditions by implementing various techniques. Depending on the existing condition of the site and restoration goals, these techniques may include reestablishing stream flow or tidal fluctuations, restoring flood regimes, halting drainage, and reestablishing topography. These conditions can be attained by excavating basins or channels, constructing dikes, grading an undulating topography, installing water control structures, plugging ditches, and removing drainage tile lines. If the site is to be used for the reintroduction of rare species, there are additional considerations, especially genetic, political, and legal issues.

Once a range of hydrological conditions is established, plant species best adapted to each specific hydrological condition can be selected. Many wetland restorations fail because the relatively few plant species selected grow poorly or die because conditions are too wet or dry for too long. If the goal of a particular restoration is to restore functions such as flood control or nutrient retention, then relatively few plant species need to be planted. If the goal of a project is to restore the natural heritage or educational values of a wetland, or to reestablish the compositional and structural complexity of highly diverse wetlands, then many plant species should be established.

Sources of native trees, shrubs, sedges, grasses, and herbaceous species include plant nurseries and other wetlands, especially those that are planned for future, permitted alteration. Seeds of many wetland species can be collected and grown for future transplanting or direct seeding. Some woody species can be established from stem cuttings. Many herbaceous species can be propagated from divisions off the parent. Spreading soil collected from a donor wetland, especially those planned for future alteration, will provide the greatest diversity of plant and animal organisms. Numerous woody and herbaceous plant species can arise from transplanted soil, via whole plants, seeds, roots, rhizomes, and stolons. Including leaf litter, detritus, and, for forested wetlands, large amounts of coarse woody debris are beneficial.

Herbivores cause many wetland restoration projects to fail, directly (by feeding) and indirectly (by burrowing, by dam building). Invasive species, such as *Lythrum salicaria, Phragmites australis*, and *Tamarix* spp., also must be controlled if plant and animal diversity are desired.

Understanding the ecological factors that control the diversity of animals in wetlands is key to altering these factors to restore, enhance, or create wetlands for the benefit of wetland fauna. Because hydrology underpins so much of wetland ecology, manipulation of wetland hydrology through water-level regulation is the most common and cost-effective way to manage wetland habitats for wetland fauna. More specifically, water-control structures can be used to flood or drain wetlands and thereby alter plant communities and the habitats available for wildlife. By drawing down water levels, a germination phase can be produced, which can be followed...
by gradual reflooding, once plant communities become established. Timing, duration, and degree of drawdown (for example, shallow, growing season drawdowns versus complete overwinter drawdowns) influence subsequent plant species composition and the types of wildlife later attracted. Several years of stable water levels of moderate depth following drawdown are often required to establish the submerged plant communities that are important to many wetland animals (invertebrates, ducks, fish). Manipulation of wetlands to produce early stages of plant succession can create a diversity of habitat niches for wetland animals, particularly if practiced within a wetland complex in which management of different units is staggered for different successional stages. This practice permits the wetland fauna, elements of which can be quite mobile, to track habitat conditions over time and persist over the long-term within the same local area. Manipulation of wetland hydrology also has been used to manipulate salinity in coastal areas by blocking natural drainages, thereby permitting the fresh water that accumulates to leach out salt. The resultant shift in plant communities to more salt-intolerant species also changes wildlife communities.

In addition to water-level manipulation, a number of artificial procedures can be used to modify wetlands to enhance wildlife habitats. These include creation of openings in dense emergent vegetation through cutting, application of herbicides or, more permanently, by blasting. Planting of select forage species is sometimes attempted but is usually expensive and rarely successful. Provision of artificial nest sites for select species, especially ducks, has been one of the most common and visible forms of wetland management to benefit wildlife. Artificial nesting and loafing sites can be constructed for birds. Large, whole tree boles, including partially hollowed-out trees, can be placed within or adjacent to the restoration site. While perhaps aesthetically pleasing, the overall contribution of such efforts to boost local populations is at best modest given the small fraction of regional populations that can be supported by such structures. These artificial procedures, in general, are expensive and time-consuming and should be considered a minor complement to a larger strategy of natural management that uses hydrological manipulation to produce habitat changes over large areas in a cost-effective manner.

The tradeoffs between community types must be articulated and considered in management and planning. For example, flooding bogs and meadows to produce marshes, a not uncommon occurrence in North America, entails losses of certain species, many rare or unusual, and gains of others, many common and sought after by recreational users. In general, maintaining wetlands in a productive and natural state that meets the needs of the entire wetland fauna is more likely to meet diverse public needs than is species-specific management (e.g., for waterfowl only).

VI. CONCLUSIONS

Despite lack of systematic comparisons, the existing scientific literature on wetland ecosystems is unambiguous in conveying the high level of diversity supported by wetlands. The source of this high diversity lies in the exceptionally large number of different types of wetlands that occur worldwide. Because wetlands occur on every continent except Antarctica, in every climatic region, in most biogeographic regions of the globe, and in a wide array of geological and topographic settings, the number of different types of wetlands created by the myriad combinations of these factors is very large and has yet to be catalogued. The extraordinary richness of plant and animal species that depend on wetlands arises from this diversity of types. Relative to the area they occupy, wetlands support a disproportionately large fraction of the world’s rare, endangered, and threatened plant and animal species. Maintaining the high diversity of wetland species means maintaining a high diversity of wetland types.

See Also the Following Articles
ESTUARINE ECO SYSTEMS • LAKE AND POND ECO SYSTEMS • RIVER ECO SYSTEMS • ECOSYSTEMS WETLANDS: RESTORATION

Bibliography


I. What Is a Wetland?
II. The Value of Wetlands—Functions and Biodiversity
III. Wetlands—Threats and Losses
IV. What Is Wetland Restoration?
V. Restoration Planning and Implementation
VI. Wetland Restoration Case Studies
VII. Conclusion and the Way Forward

GLOSSARY

bund An artificial embankment keyed into the substrate and used to retain water.
groundwater Underground water that is held in the soil and in pervious rocks.
hydraulic conductivity A measure of the potential water flow through a soil.
ox-bow A crescent-shaped waterbody occurring on a river floodplain having once been part of a river meander that has been cut through and abandoned.
rehabilitation The partial return of structure and functions found in a predisturbance state to a disturbed habitat.
restoration The process of reinstating some or all pre-existing functions to a lost wetland.
riparian Of or inhabiting a riverbank or margin.
stakeholder A person who has an interest or concern with a particular issue or business.
watershed The total area from which a single river collects water.
water table The upper surface of the zone of saturation in a soil or rock formation.

EFFECTIVE WETLAND RESTORATION is not only vital for the maintenance of biodiversity, but also for a number of other valuable functions. The chapter defines the term wetland and describes the reasons and extent of habitat loss. The nature and value of natural wetland functions is described. After briefly discussing project planning and implementation, a series of restoration case studies representative of a broad cross-section of wetland types from different continents are presented. The chapter concludes with an overview of current progress on wetland restoration and indications for future direction.

I. WHAT IS A WETLAND?
The Ramsar Convention on Wetlands defines wetlands as ‘areas of marsh, fen, peat, and or water, whether
WETLANDS RESTORATION

natural or artificial, permanent or temporary, with water that is static or flowing, fresh, brackish or salt, including areas of marine water the depth of which at low tide does not exceed six meters.” This Ramsar definition is perhaps the broadest of the many definitions of what constitutes a wetland and it is the one used throughout this chapter. Wetlands therefore include the following:

• marine—coastal wetlands such as coastal lagoons, rocky shores, and coral reefs
• estuarine—for example, deltas, tidal marshes and mangrove swamps
• lacustrine—wetlands associated with lakes
• riverine—wetlands along rivers and streams
• palustrine—marshes, swamps, and bogs
• human-made wetlands such as reservoirs, fish ponds, flooded mineral workings, saltpans, sewage farms, and canals

Because wetlands are dynamic systems and occupy the transitional zone between aquatic and terrestrial habitats there is often controversy over dictating boundaries and precise definitions. More restrictive definitions than the one presented earlier and methods for boundary delineation certainly exist and are the subject of several manuals published in North America where this is an important and politically charged issue (Mitsch et al., 1994).

II. THE VALUE OF WETLANDS—FUNCTIONS AND BIODIVERSITY

Aside from their importance for wildlife, wetlands are important, and sometimes essential, for the well-being of the people who live in or near them. In recent years there has been increasing awareness of the value of the hydrological and chemical functions of wetlands. To summarize, wetlands perform the following functions:

• flood alleviation—wetlands naturally regulate water flows within a watershed
• shoreline stabilization and erosion control
• storm protection—many wetlands, especially mangroves, act as windbreaks during storms, protecting coastal habitat and property
• groundwater recharge—wetlands retain water within a watershed enabling groundwater recharge
• water quality improvement—wetlands can retain nutrients, toxic substances and sediment
• climate—wetlands can influence local climatic conditions
• carbon storage—some wetlands such as peatlands store carbon
• biomass export
• habitat for wildlife—wetlands are crucial for the maintenance of biodiversity

Economic benefits accrue from the functions just described, therefore when they are lost through the destruction of wetlands they often have to be replaced at enormous financial cost to society. These benefits include the following:

• water supply—wetlands can influence both water quantity and quality
• fisheries—nearly three-quarters of the world's fish harvest is linked to the health of wetland areas
• agriculture—floodplains provide some of the most fertile agricultural land in the world
• timber production
• energy resources, such as peat and plant material
• wildlife resources
• transport
• recreation and tourism opportunities

Additionally many wetlands have special cultural attributes; being part of religious, cosmological and folklore beliefs, as well as providing aesthetic inspiration and wildlife sanctuaries. From a biodiversity perspective, wetlands are very important; most wetland ecosystems have high biodiversity and support suites of specialist species (Fig. 1).

III. WETLANDS—THREATS AND LOSSES

Wetlands have been lost at a phenomenal rate, especially in the developed countries of the Northern Hemisphere. Reasons for the destruction of wetlands center on the fact that their functions and values are often not directly “harvestable” by their owners for profit. Flood alleviation functions benefit those living downstream, fish and wildlife produced may migrate outside the system and be harvested by others, and so on. Other valuable functions such as groundwater recharge and sediment and nutrient removal are incredibly important but also cannot be exploited commercially.

Threats faced by wetlands are summarized in Table 1. These threats have led to catastrophic declines in the
coverage of wetlands around the world (Fig. 2). Any such statistics belie the real situation, however, as the remaining wetlands in the areas that have suffered serious loss are usually far from pristine. The extent of the loss of wetlands and the resultant financial cost incurred due to the loss of important wetland functions have created the need for restoration in some areas. To date, restoration attempts have largely taken place in Europe and North America, regions which have seen the most devastating declines in wetland coverage.

IV. WHAT IS WETLAND RESTORATION?

Strictly speaking, restoration is the process of reestablishing a naturally functional, self-sustaining system, implying a return to an original predisturbance state. In practice, this is not possible as restored habitats rarely mimic all the properties of the original. Replacing or restoring preexisting biological resources is likewise not possible. Restoration is more realistically defined as the reinstatement of some or all preexisting functions to “lost” wetlands (Hollis, 1993). At this point, it is also worth defining rehabilitation, which is the partial return of structure and functions to degraded wetlands. The process often involves the selection of desirable features and functions only. For the purposes of this chapter, restoration and rehabilitation will be treated as synonymous. Restoration and rehabilitation, together with management, form a continuum. Management can be applied to existing wetlands to develop certain functions artificially (Hollis, 1993).

Restored wetlands are unlikely to support diverse assemblages or rare species. The process of restoration can, however, rapidly enhance the biodiversity of an impoverished area. Restoration of wetland ecosystems is becoming increasingly important for the maintenance of biodiversity, particularly in developed countries because (a) remaining wetland sites are becoming increasingly fragmented and isolated making it difficult or impossible for biota to move from one area to another and (b) loss of natural processes (e.g., flooding) means that natural habitats are no longer dynami-
These factors are compounded by institutional and legal factors such as lack of legislation or its enforcement, destructive policies, and livestock pressure. Social factors such as poverty of wetland users, inadequate control by users of marketing of wetland products, low economic return,土地 ownership, official attitudes toward wetlands, lack of awareness of sustainable alternatives, malnutrition, and outside vested interests.

These factors are compounded by institutional and legal factors such as lack of legislation or its enforcement, destructive policies, and livestock pressure. Social factors such as poverty of wetland users, inadequate control by users of marketing of wetland products, low economic return, land ownership, official attitudes toward wetlands, lack of awareness of sustainable alternatives, malnutrition, and outside vested interests.

Those factors are compounded by institutional and legal factors such as lack of legislation or its enforcement, destructive policies, and livestock pressure. Social factors such as poverty of wetland users, inadequate control by users of marketing of wetland products, low economic return, land ownership, official attitudes toward wetlands, lack of awareness of sustainable alternatives, malnutrition, and outside vested interests.

Those factors are compounded by institutional and legal factors such as lack of legislation or its enforcement, destructive policies, and livestock pressure. Social factors such as poverty of wetland users, inadequate control by users of marketing of wetland products, low economic return, land ownership, official attitudes toward wetlands, lack of awareness of sustainable alternatives, malnutrition, and outside vested interests.

Those factors are compounded by institutional and legal factors such as lack of legislation or its enforcement, destructive policies, and livestock pressure. Social factors such as poverty of wetland users, inadequate control by users of marketing of wetland products, low economic return, land ownership, official attitudes toward wetlands, lack of awareness of sustainable alternatives, malnutrition, and outside vested interests.

Those factors are compounded by institutional and legal factors such as lack of legislation or its enforcement, destructive policies, and livestock pressure. Social factors such as poverty of wetland users, inadequate control by users of marketing of wetland products, low economic return, land ownership, official attitudes toward wetlands, lack of awareness of sustainable alternatives, malnutrition, and outside vested interests.

Those factors are compounded by institutional and legal factors such as lack of legislation or its enforcement, destructive policies, and livestock pressure. Social factors such as poverty of wetland users, inadequate control by users of marketing of wetland products, low economic return, land ownership, official attitudes toward wetlands, lack of awareness of sustainable alternatives, malnutrition, and outside vested interests.
V. RESTORATION PLANNING AND IMPLEMENTATION

The framework that follows describes the essential areas that should be addressed when developing restoration proposals. Suggested contents for a restoration plan are outlined in Table II. Restoration planning is required to do the following:

- Clearly state the aim and objectives of the project.
- Enable the safeguarding of any existing conservation interest.
- Identify appropriate target habitat and communities.
- Identify appropriate techniques for restoration.
- Identify project constraints (e.g., legal problems, access difficulties).
- Ensure that restoration operations are well co-ordinated.

### TABLE II

<table>
<thead>
<tr>
<th>Essential Components of a Restoration Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>• An outline and details of the scope of project</td>
</tr>
<tr>
<td>• Site survey and appraisal—existing site conditions should be assessed in detail to identify topographical features, land use and classification, climate, water levels and movements, ownership boundaries, right of way, existing vegetation, and special factors (e.g., prevailing winds, views, and existing nature conservation interest)</td>
</tr>
<tr>
<td>• Formulation of restoration objectives, which should identify final land form and water levels, new positions of rights of way, planting details and management requirements of new vegetation, and treatment of unstable slopes</td>
</tr>
<tr>
<td>• Design of operational details and working plans to give precise details for implementation phase</td>
</tr>
<tr>
<td>• Acknowledgment of limitations</td>
</tr>
</tbody>
</table>
WETLANDS RESTORATION

Initial idea and establishing aim(s)
Undertaking pilot study or studies
Contacting landowners and gaining provisional acceptance
Integrating the interests and aspirations of landowners, the public and relevant authorities

Project design including:
• reviewing knowledge about the site/reconstructing past wetland
• determining elements which need to be restored, e.g., raised water table
• setting criteria for success

Achieving approval and processing by authorities

Agreement of costs and funding
Implementation of restoration

monitoring
modify actions if necessary

Monitoring
Post-project appraisal

FIGURE 3  Essential stages of any restoration project. Adapted from Benstead et al. (1999).

• Detail monitoring/aftercare programs/post-project appraisal.

The stages which typically make up a restoration project are shown in Fig. 3. The timescale for achieving restoration will depend on a number of factors, such as the following:

• how degraded the site is (e.g., high levels of nutrients can take a long time to deplete)
• how isolated the site is (e.g., a site in an active river corridor is more likely to recolonize rapidly)
• the ability to manage the site for commercial/artisanal purposes, such as management for pastoral agriculture or regular cutting for roofing material, where such management is desirable
• climatic factors (e.g., low rainfall will prolong restoration and dry conditions will encourage invasion of ruderal, weedy plant species)
• the extent and knowledge of any previously installed land drainage system (e.g., location of under-drainage)
• nature of adjoining land use (e.g., risk of flooding of neighboring residential and agricultural land)

Achieving project objectives depends on a sound understanding of hydrological and ecological processes. The following section describes the key areas and techniques that need to be addressed when implementing a restoration plan: hydrology, soils, and vegetation establishment/management.

A. Hydrology

Complete understanding of site hydrology and the feasibility of restoring a target hydrological regime is key to achieving restoration goals. Failure to provide suitable hydrological conditions is the likeliest cause of project failure. Restoration of a hydrological regime should aim to mimic natural conditions wherever possible. Planning and site assessment should include careful evaluation of the target water regime (levels, seasonality, etc.) and whether these are likely to be achievable and will meet the requirements of target habitats and species. Calculating the water budget for the site by assessing likely inputs and outputs is essential. It is possible to calculate water budgets using just climatic information to give a rough estimate of site water demand. However, in reality surface wetness is influenced by a range of other factors such as the relative hydraulic conductivity of different soil types and topography (Benstead et al., 1997). Hydrological restoration may include a range of options from reinstatement of natural flooding by setting back flood embankments, using existing drainage
infrastructure in reverse, installing water retention features (see the Hornborgasjön case study, presented later), and removing or destroying existing drainage infrastructure (see the Abernethy case, presented later).

B. Soils
In all forms of restoration ecology particular attention should be paid to soils. Creating a functional soil profile with balanced aerobic, anaerobic, organic, and inorganic components is essential for the maintenance of wetland functions that are related to nutrient transformations, food chain support, and fish habitats.

Mineral soils that have been dewatered are likely to respond to rewetting. However, organic soils that have dried out are likely to have suffered irreversible physical and chemical alteration. Detailed investigation of soil properties (such as nutrient status, structure, moisture retention, and porosity) should be the subject of early site assessment (Allen, 1989). Soil invertebrate fauna of areas that have been drained are usually terrestrial in nature and are unlikely to survive “rewetting.” Complete loss of earthworm fauna is possible and this can drastically lengthen the time it takes soil to recover from rewetting where structure has been lost. The loss of soil invertebrates also has implications for species that depend on them for prey. Recolonization of soil fauna can, in some cases, take decades and “inoculation” with cultures of suitable invertebrates may have to be considered (Butt and Fredericksen, 1995).

C. Vegetation Reestablishment
Where sites are adjacent to existing wetlands, restoration can sometimes be left to natural regeneration. This occurs through seed drift from nearby areas, germination from the seed bank and the arrival of flood borne diaspores, and rhizome fragments. A number of techniques can be used to increase the rate at which species recolonize a wetland site where natural sources of material are not present (see the Upper Thames case study, presented later); these involve the introduction of biological material using rhizomes, seed, seedlings, mature plants, and cuttings.

VI. WETLAND RESTORATION CASE STUDIES
The following 13 case studies (Fig. 4) illustrate how restoration has been successfully undertaken on a range of habitats with a broad geographical spread. Initial case studies highlight the importance of project planning. Subsequent case studies work up from simple small-scale restoration projects to larger, more complex projects.

A. Advocacy and Planning
Solutions—Creating the Framework for Successful Restoration at Tasek Bera
Tasek Bera, Peninsular Malaysia, is a 6150 ha freshwater lake system with open water and fringe vegetation surrounded by freshwater swamp forest. This habitat diversity coupled with relative isolation has combined to produce a diverse ecosystem with an unusually high degree of endemism. The biological community supported is unique within Malaysia, and in recognition of this Tasek Bera was recently designated as the country's first Ramsar site.

In the past 30 years Tasek Bera has come under threat from a number of directions:
- the forested watershed has been cleared for agriculture, increasing the sediment and nutrient loading of water entering the lake and also reducing the volume of water entering the system especially during dry periods
- pollution in the form of sewage from nearby rivers during high river levels (when flow reversal occurs)
- increased access due to new logging tracks and highways
- damage to hydrology by logging infrastructure, especially roads, which leads to altered drainage patterns and increased siltation

Because of the importance of the site and the problems it faces, the area is the subject of a restoration project. This project is in the early stages and before the required physical restoration action can be taken it is necessary to ensure that policy, planning and institutional arrangements are in place to ensure its ultimate success. This work includes the following:
- defining the site boundaries
- identifying an appropriate buffer zone and developing guidelines for neighboring land managers within this area
- baseline surveys on all aspects of the site
- writing the management plan for the site
- providing training for local government officers
education work with local schools
• writing and implementing a community develop-
ment strategic plan
• undertaking community pilot projects focusing on
handicraft production and ecotourism involving the
local aboriginal people
• planning ecotourism use of the site

This level of planning, site assessment, communica-
tion, and community involvement is essential if large
projects are to achieve their objectives.

B. Simple Solutions—Reinstating
Agricultural Management at Tipperne
(Thorup, 1998)

Tipperne, Denmark, is a 700 ha nature reserve con-
sisting mainly of brackish meadows, reedswamps, and
dunes. A wealth of information on its bird populations
is available for this century.

During the period 1928–1945 agricultural utiliza-
tion of the site was fairly constant; the entire area was
mown once or twice a year for hay in August/Septem-
ber, followed by aftermath grazing until November.
After 1946 agricultural use declined and by the early
1960s the area had been abandoned. Invasion by rank
vegetation and common reed Phragmites australis fol-
lowed over the next 10 years. The loss of wet grassland
habitat impacted on the numbers of breeding shorebirds
using the site until only a handful remained.

In 1973 a management plan was initiated with the
aim of restoring the wet grassland areas. During the
first 10 years part of the reserve was grazed and quite
dense populations of breeding waders and wildfowl
reestablished in these areas. From 1984 management
was further improved by mowing the remaining areas
of the site; this increased the area of restored wet grassland
dramatically. Breeding wet grassland birds increased
again in response (see Fig. 5).

C. Reinstating Management Beneficial to
the Maintenance of Biodiversity after
Intensive Agricultural Use

Damage to wetlands, through intensification or neglect,
may be reversible, although it can take many years for a
characteristic wetland community to recover. Dramatic reversals can be achieved after relatively short periods of time, however. Five years after reinstating cutting management to a wet grassland that had been neglected for 20 years in the Luznice river floodplain (Czech Republic), plant species diversity and composition were restored to a quality comparable with adjacent grasslands that had received uninterrupted management (Benstead et al., 1999).

A similar time span was found for the restoration in 1991 of 141 ha of wet grazing pasture on a nature reserve in the North Kent Marshes, United Kingdom (Benstead et al., 1997). The project involved manipulating the sward either by reseeding or through allowing natural regeneration and by raising water levels in the drainage channels using sluices, the creation of bunds, and blocking off piped drainage. The site was grazed by sheep owned by grazing tenants (April-December) followed by mowing to control persistent weed species. This simple process resulted in an increase in the number of shorebirds breeding on the site within 5 years (Table III).

In the Drenthe Aa wet grasslands of the Netherlands, the mean number of plant species increased from 18 to 26 (per field of 500 m²) within 16 years after the cessation of fertilizer application (Bakker, 1989). These fields were adjacent to diverse unimproved meadow communities so that colonization from these populations could occur.

Restoration of agricultural land to wetland can be difficult because the sites may have any of the following:

- high nutrient availability (cropping or topsoil removal may have to be considered to lower nutrient levels)
- impoverished seed banks, repeated cropping, and the use of herbicides will have removed most of the naturally occurring plant material, and this will have been replaced by many undesirable ruderal species
- altered hydrological regimes (soil structures may have altered irreversibly)

Reversion of arable land to wetland habitats is therefore more easily undertaken on sites that have been under cultivation for a short time, that are adjacent to existing wetlands which can act as a seed source, and where appropriate hydrological conditions can be reestablished. Where livestock are used to manage sites, moving stock from an existing wetland to a site which is being restored can help. The livestock will transport both diaspores and invertebrates between the sites. Sheep are considered to be better at transporting biological material than other livestock: in a study in Germany, examinations of one sheep yielded 8500 diaspores of 85 plant species, and the sheep was also found to transport 13 species of Orthoptera (Fischer et al., 1996). This transport process should be borne in mind when reintroducing grazing stock to restored sites as it may have a negative effect by introducing undesirable species.

D. Reintroduction—Accelerating Target Vegetation Community Development in the Floodplain Wet Grasslands of the Upper Thames Tributaries

Restoration of agricultural land to floodplain wet grassland has been investigated, along the River Ray (a tributary of the River Thames in the United Kingdom). Drainage improvements in the 1970s led to the conversion of traditional hay meadows to arable land. However, the combination of periodic flooding and a heavy clay soil make cultivation problematic and the recent introduction of agri-environment mechanisms (Countryside Stewardship and Environmentally Sensitive Areas schemes) aimed at restoring hay meadows has been well received by farmers. Restoration best-practice techniques have been the subject of experimentation.

The investigation looked at four arable reversion techniques: natural regeneration, use of cut plant material transported from another site, sowing seed mixes,
and the use of nurse crops (Manchester et al., 1998). The study revealed the following results:

- On ex-arable sites, natural regeneration would take decades to restore a seminatural vegetation type. Analysis of the available seed bank found that small quantities of desirable species were represented, but not in great enough abundance or reasonable diversity to restore target communities. "Seed rain" or floodwater seed dispersal from neighboring herb-rich areas appeared to be limited. However, several species did colonize the sward by natural regeneration.
- Scattering hay bales has several compelling advantages; it ensures that the supply of seeds are of known provenance and is very cheap. This technique was, however, unpredictable and less effective than sowing seed.
- Sowing seed mixtures resulted in significantly greater numbers of species in the resultant sward than either natural regeneration or hay bales. The most comprehensive seed mix produced the sward with the highest species richness and also the highest number of desired/potential species. Sowing seed in this way is far cheaper than transplanting mature plants or turfing areas, neither of which were considered in the experiment.
- Nurse cropping was ineffective in boosting the establishment of other species.

E. Rehabilitation of Damaged Peat Bogs at Abernethy by Removing Drainage Systems

(Brooks and Stoneman, 1997)

Peat bogs have suffered around the world from damage caused by peat extraction, grazing and conversion to improved grassland, and afforestation. Peat extraction results in drainage of peat bogs and often leads to the invasion of scrub. Functional restoration of peat bogs following extraction therefore tends to involve around hydrological control and vegetation management. Hydrological control techniques employed include blocking off drainage systems using dams and sluices, retaining water using bunds, and infilling drainage channels.

Vegetation control often involves the physical removal or chemical treatment of invasive and unwanted plant species, the re-introduction or transplanting of desirable plant species, and subsequent appropriate management of the desired vegetation community. At Abernethy Forest (a reserve managed by The Royal Society for the Protection of Birds), areas of valley mire were threatened by an extensive forestry drainage network installed prior to purchase. Mire vegetation was slowly being replaced by drier heath communities. To counter this the drainage channels were initially blocked with peat dams, using hydraulic diggers. Peat was removed from borrow pits, either within or adjacent to the drainage channels. All vegetation and humified peat was removed, and as each successive layer of the dam was built up, the structure was compacted to ensure a watertight barrier. This method worked well where water levels were high but if they were low then the exposed peat dried out, cracked, and the dams leaked. The ploughlines were not touched and continue to have a limited drainage function.

An alternative method was tried later in another intensively drained mire on the site. This involved completely destroying the drainage network. The plough ridges were bulldozed into the drainage channels, filling them in, and then compacted down to reduce water flow. This method is useful in that it is simple and effectively removes all functioning components of the drainage system.

F. Biomanipulation—Fish Removal in Lake Vaeng, Denmark

(Moss et al., 1996)

Eutrophication of shallow freshwater lakes initially results in a flush of plant growth and an increase in

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanellus vanellus</td>
<td>3</td>
<td>2</td>
<td>16</td>
<td>35</td>
<td>56</td>
<td>26</td>
<td>15</td>
<td>26</td>
</tr>
<tr>
<td>Gallinago gallinago</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Tringa totanus</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>27</td>
<td>35</td>
<td>5</td>
<td>10</td>
<td>32</td>
</tr>
<tr>
<td>Avocetta recurvirostra</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>33</td>
<td>26</td>
<td>14</td>
<td>20</td>
</tr>
</tbody>
</table>

TABLE III

Number of Pairs of Breeding Waders at Northward Hill, North Kent Marshes, UK, after Restoration in 1991
phytoplankton density. But eventually aquatic macrophytes disappear and algae take over. There is a net loss of biodiversity. This can be a continuing problem even when the original polluting element has been dealt with. The core of the problem appears to be that once open waters are created, fish find it easy to remove zooplankton, which are important grazers of phytoplankton. Algal dominance is further ensured by factors such as shading (algae begin growing earlier in the season and quickly create unsuitable light conditions for colonizing plants) and sediment changes (algae lay down a fine sediment, which contrasts markedly with the more fibrous deposits laid down by plant beds; this increases turbidity and decreases light values further disadvantaging plants).

One way to reverse this problem after preventing the source of nutrient pollution is by biomanipulation. This usually involves the removal of zooplanktivorous fish or the addition of piscivorous ones. Following on from this it may be necessary to protect reestablishing plants against physical disturbance and grazing. Upon successful reestablishment of aquatic plant communities, fish and grazing animal populations can be reinstated.

This technique was successfully employed at Lake Vaeng in Denmark. The 15 ha, shallow lake was polluted by sewage effluent and devoid of plants. Diversion of the sewage effluent in 1981 reduced nutrient levels but failed to restore water clarity and aquatic plant growth. During 1986-1987 roughly half the zooplanktivorous fish were removed. Successive removal in the following two seasons took the fish density down to 8 g live weight per square meter (Fig. 6). The results were rapid, within two years water clarity improved, the zooplankton community acquired Daphnia species, which are important algae grazers, blue-green algae disappeared, and nitrogen and phosphorous levels halved. Plants, however, were slower to respond but 80 to 90% plant coverage of the lake bed was achieved by 1991. This delay was attributed to grazing by birds.

G. Sediment Removal and Isolation as a Method for Restoring Lacustrine Systems in the Norfolk Broads (Moss et al., 1996; Tickner et al., 1991)

The Norfolk Broads (United Kingdom) are a system of shallow-flooded, medieval peat diggings connected by a set of rivers flowing through low-lying ground in East Anglia. This area supports high biodiversity but this has declined in recent years (George, 1992). Drainage improvements, the declines in traditional wetland management practices, and the pollution of rivers have led to widespread degradation of habitat.

Many of the broads were open to the river and subject to tidal flooding with eutrophic freshwater. Situations gradually left few open water areas remaining (e.g., at Strumpshaw Broad) or greatly reduced the depth of the waterbodies (e.g., Cockshoot Broad). Aquatic macrophytes were lost and the remaining wetland plants were all indicators of eutrophic conditions. River drainage improvements also reduced the extent, depth, and duration of regular flood events, leading to the encroachment of undesirable trees species (Salix and Alnus spp).

In some cases isolation was achieved by sealing the channel connecting the broad to the river and building up the river bank to prevent flooding in all but the highest flood events. With no other inputs, rainwater alone was quickly found to be insufficient for the wetland requirements at Strumpshaw and a borehole was sunk into the underlying aquifer and up to 183,000

FIGURE 6: Biomanipulation is one way to deal with the effects of algal dominance in lake ecosystems resulting from eutrophication.
TABLE IV

Drainage Channel Flora Recovery: The Presence of Pollution-Tolerant and Pollution-Intolerant Aquatic Plants in Drainage Channels at Strumpshaw, 1976–1990

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Potamogeton pectinatus</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Callitriche stagnalis</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Nuphar lutea</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Filamentous algae</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Elodea canadensis</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ranunculus aquatilis</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Potamogeton crispus</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ceratophyllum demersum</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Hydrocharis morsus-ranae</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Chara spp.</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Utricularia minor</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Utricularia vulgaris</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lemna minor</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Littorella palustris</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Myriophyllum aquaticum</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Myriophyllum verticillatum</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Zannichellia palustris</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Nymphaea alba</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Rumex hydrolapathum</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Rumex acetosella</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Polygonum amphibium</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Polygonum cuspidatum</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Sparganium erectum</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Sparganium robustum</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Sparganium lanceolatum</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha angustifolia</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha latifolia</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha orientalis</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha angustifolia</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha latifolia</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha orientalis</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha angustifolia</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha latifolia</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha orientalis</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha angustifolia</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha latifolia</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Typha orientalis</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Adapted from Tickner et al. (1991).

This trial work has shown considerable benefits to biodiversity. At Strumpshaw Broad, aquatic plants have made a steady recovery and pollution intolerant plants are making a comeback (Table IV). Invertebrate diversity has likewise increased, with a rise in the number of dragonfly species using the site. The increase in the area of open water has benefited wintering wildfowl. The restoration has certainly succeeded in improving water quality on the reserve and made the site more attractive to target bird species, but the current management regime is not sustainable, relying as it does on pumping water. Clearly isolation is not a long-term option either. Connectivity with the rivers in the system is ecologically desirable and therefore sources of pollution in main rivers must eventually be addressed. At Cockshoot Broad, simply removing the sediment and associated nutrients was not enough and biomanipulation techniques (such as those described in the Lake Vaeng case study) may have to be used in the future.

H. Removing Alien Species—The Melaleuca Problem in Southern Florida

(Maltby, 1997)

Introductions of the tree *Melaleuca quinquenervia* from Australasia to the United States has been an ecological disaster. Having wide ecological tolerances the tree is able to survive in swamp environments and today the trees cover nearly 10% of the southern half of Florida. *Melaleuca* has much higher water requirements than the natural sawgrass vegetation it quickly replaces. The subsequent loss of swamp vegetation and drying effect threatens wildlife and may even lower water tables and affect public water supplies. The only solution in cases such as this is the elimination of the problem species. Mechanical removal of *Melaleuca* will cost $370–2000 ha⁻¹. Authorities in Florida are pinning their hopes on a biological control in the form of a weevil *Oxyops vitiosa* imported from Australia.

I. Engineering Solutions—Water Level and Vegetation Management at Hornborgasjön

(Hertzman and Larsson, 1999)

Hornborgasjön, Sweden, is a large post-glacial lake, some 10,000 years old. Successive agricultural drainage projects between 1805 and 1935 lowered the lake water level and reduced the extent of wetland around the lake. The remaining shallow, open water was steadily invaded by *Phragmites australis*, sedge *Carex*, and willow *Salix* scrub until little open water remained by the mid-1960s. Drainage and land-use changes reduced the value of the lake for wildlife. Ambitious plans to restore the lake were proposed in 1965 and have recently been completed.

The restoration plan called for the raising of the lake water level by 0.85 m. In order to achieve this, a 3-km cubic meters of water were abstracted each year. The largest management operation involves the removal of the underlying, nutrient-rich sediment from the former basins of the Broads (Moss et al., 1990; Tickner et al., 1991).

This trial work has shown considerable benefits to biodiversity. At Strumpshaw Broad, aquatic plants have made a steady recovery and pollution intolerant species are making a comeback (Table IV). Invertebrate diversity has likewise increased, with a rise in the number of dragonfly species using the site. The increase in the area of open water has benefited wintering wildfowl. The restoration has certainly succeeded in improving water quality on the reserve and made the site more attractive to target bird species, but the current management regime is not sustainable, relying as it does on pumping water. Clearly isolation is not a long-term option either. Connectivity with the rivers in the system is ecologically desirable and therefore sources of pollution in main rivers must eventually be addressed. At Cockshoot Broad, simply removing the sediment and associated nutrients was not enough and biomanipulation techniques (such as those described in the Lake Vaeng case study) may have to be used in the future.
bund and sluice was built around the western boundary to protect neighboring arable farmland. Water levels were slowly raised between 1992 and 1995, restoring some 4100 ha to wetland. Approximately 800 ha of invasive birch Betula and alder Alnus woodland and Salix scrub were cut and 1200 ha of reedbed was burnt or chopped using specially designed amphibious machinery.

The open water character of the lake has been restored (Fig. 7). The reserve supports the only remaining Swedish breeding colony of black-necked grebe Podiceps nigricollis (110 pairs in 1997). Additionally, although maximum water levels have only just been achieved, black tern Chlidonias niger have returned to breed on the lake (65 pairs in 1997). The remaining reedbeds provide improved feeding conditions for many birds and support increased numbers of many species such as bittern Botaurus stellaris, marsh harrier Circus aeruginosus, and reed warbler Acrocephalus arundinaceus, despite a marked reduction in areal cover of this habitat.

Total costs exceeded $10 million (in 1992), including $2.7 million in compensation payments, $2.0 million in bund and sluice construction, and a similar sum spent on vegetation clearance.

J. Restoring Socioeconomic Functions—The Senegal River Valley (Wit, 1997)

In developing countries alteration of natural hydrological regimes can have drastic and far-reaching effects on the human populations relying on wetland resources. Construction of dams along the River Senegal, at Manatali in Mali and Diama at the river mouth, was originally undertaken without appreciation of the reliance of the entire socioeconomic system of a large region on natural river processes. These developments resulted in alteration of 73% of floodplain wetlands; 95% loss of the seasonal grazing land; 90% reduction in fish production in the river and estuary; salinization of half of the newly

FIGURE 7 Extent of major vegetation communities at Hornborgasjön before and after restoration. Adapted from Benestad et al. (1998).
created, irrigated land; stagnation, which has led to a high incidence of disease and pollution; and threats to the biodiversity of two National Parks in the delta area.

To counter the threat to local communities and biodiversity, a trial restoration of 60,000 ha was undertaken. New sluices and embankments were installed to bypass the Duma dam and to allow fresh water to flood former estuarine wetlands connected to the sea. New floodgates allowed migratory shrimp and fish access between the former lakes and sloughs, allowing them to breed. Fish harvests have increased dramatically as a result. Results are positive and there are plans to increase the scope of the restoration.

K. Local Community Involvement in the Restoration of Mangrove Forest in Pattani Bay, Thailand (Erftemeijer and Bualuang, in press)

Development driven processes have resulted in the loss of over half of Thailand’s mangrove resource since 1970. Fish catches, coastal biodiversity and incomes of coastal-dwelling communities have declined as a result and now restoration has a high priority. While the techniques for mangrove restoration are well documented, securing the long-term future of a restored area through the wise use of natural resources must involve the local community.

A 3-year project in Pattani Bay, an internationally important wetland for migratory shorebirds, aimed to demonstrate the ability of mangroves to strengthen the regional prosperity of those managing the resource. This was achieved by promoting community organization and partnership between all stakeholders, physical rehabilitation of degraded mangrove forest, supporting alternative livelihood initiatives, and effective communication.

Progress in any project of this nature is slow and incremental. Organization of village committees and generating support from key community leaders and the local administration was a lengthy process fraught with internal “political difficulties.” Project staff were trained in conflict resolution. Early planting trials failed due to overgrazing and drought, but then the villagers drew up a restoration plan and microtopographical improvements were made to ensure survival of future plantings. To date 30 ha of forest have been replanted.

Lessons learned from this project highlight the need to ensure that community-led restoration projects:

- Do not shift objectives away from broader long-term solutions to activities with immediate benefits.
- Ensure involvement of all stakeholders in the process is long-term as progress is slow and incremental. Promote and maintain community ownership and commitment. This is essential for long-term sustainability.
- A similar approach has been used by the Yadfon Association in nearby Trang Province, where, over 13 years, 512 ha of mangrove forest have been restored and are being managed as a community forest.

L. Effects of Stream Restoration in Denmark (Iversen et al., 1995)

European rivers are often highly eutrophic being affected by excessive discharges of organic matter and nutrients. Recently, improved sewage treatment has reduced the transport of organic matter and phosphates leading to improved ecological conditions. However, nitrogen levels remain high where leaching from agricultural land is a problem. Several channelized Danish streams have been physically restored since 1989. Restoration involves remeandering, channel bed raising, replacement of riffle-pool sequences, replacement of gravel fish spawning beds, and removal of obstacles to fish migration.

Before restoration the straight channels provided little physical variety and low biodiversity. After remeandering, the new river course varied in respect to stream velocity, water depth, and substrate, improving habitat for wildlife (Fig. 8). Submerged and emergent macrophytes have been shown to increase following restoration (Iversen et al., 1993). Macroinvertebrate density and diversity also increased. Specialist species favoring stone and gravel beds were able to return. Fish populations benefited from the provision of suitable spawning habitat, increased shelter opportunities provided by the more diverse riverine topography, and the increase in prey density.

Additional benefits occur. Restoration produced higher water table levels resulting in wetter riparian corridors and a higher frequency of flooding. Wet riparian meadows facilitate denitrification and can play a valuable role in treating agricultural runoff before it enters streams and rivers. Typical nitrate removal rates of 600 kg ha⁻¹ yr⁻¹ have been recorded (Behdorof et al., 1994). Flood events allow sedimentation of particulate matter and associated nutrients over riparian corridors. Sediment can sequester phosphorus transported by rivers and streams. Restoring this nutrient removal process is important in countries like Denmark, where nonpoint source pollution is leading to excessive eutrophication of shallow coastal waters.
The restoration of streams benefits both the individual landowner and society as a whole. Experience from Denmark has shown that conflicts between landowners and environmental interest can be reduced. Landowners involved have indicated that it is particularly important that land redistribution is considered in such schemes. With remeandering, land on either side of a previously straightened river may have to be swapped so that the river remains the property boundary. If a land holding is primarily within the previously drained floodplain, land swaps to include some higher ground may be needed to ensure farming viability. Once again communication with, and involvement of, all user groups is essential for success.

M. Thinking Big on the Kissimmee River, Florida—The Ecosystem Management Approach (Dahm et al., 1995)

The Kissimmee River project in Florida’s Everglades is one of the largest wetland restoration project conceived to date. The project aims to remeander 70 km of river and restore natural flooding conditions to 11,000 ha of floodplain wetland. This project, at a cost of $700 million, will reverse the drainage work and river channelization undertaken in the 1950s and 1960s. The detailed project aims are to recover natural ecological functions of the river system, including reflooding drained wetlands; maintain the physical, chemical, and biological integrity of the river; remeander the river and block all drainage channels; and expand the area of the Everglades National Park by 44,600 ha.

So far a 550-ha area of drained floodplain wetland has been used as a pilot study area where the following has been undertaken:

- The flow of water through remnant sections of the old river has been increased.
- New flow fluctuation schedules have been adopted.
- Hydrological and hydraulic modeling studies have been conducted.
- Monitoring has been established to discover the impact of the hydrological changes on wetland vegeta-
WETLANDS RESTORATION

...tion, fish, secondary productivity, benthic invertebrates, and river channel habitat characteristics.

The results are promising and show that reflooded areas still have viable seedbanks, even decades after drainage; riverine ecosystems responded favorably to the reinstatement of natural flow regimes; restoration of water flow to old river sections reestablished channel morphology, natural substrate characteristics, and benthic invertebrate species diversity; and increased flow through remnant channels cleared organic debris.

VII. CONCLUSION AND THE WAY FORWARD

The growing realization that wetland degradation has led to the loss of important economic and ecological functions has provoked many governments and organizations to consider the restoration of wetlands in order to reinstate beneficial functions. This chapter has presented a broad overview of the value of wetlands and the threats they face, as well as examples of a range of wetland restoration projects from different regions of the world.

Wetland restoration is costly and not guaranteed to succeed. The costs are so great in fact that currently only the developed countries can consider restoration. Globally this is where the bulk of the damage to wetlands has taken place. Where successful, wetland restoration can undoubtedly offset wetland losses, improve degraded sites, reestablish valuable functions, and increase biological capital.

Wetland restoration projects that achieve their objectives often have the following attributes:

• aims that address causes of damage and not symptoms
• good planning and legislative frameworks, which includes watershed planning and the provision for sympathetic land use in and around wetlands
• good planning of project implementation, aftercare, and subsequent management
• identification and involvement of all stakeholders from an early stage; successful projects depend on the cooperation between landowners, managers, user groups, public authorities and politicians at different levels, technical and scientific consultants, and non-governmental organizations
• communications networks that keep people informed throughout all the phases of the project
• successful pilot projects that engender interest in wider schemes and clarify methodologies
• an appropriate and sustainable hydrological regime
• close proximity to, or surviving sources of, biological material suitable for recolonization
• compliance monitoring during construction phase
• monitoring systems that allow management prescriptions to be modified beneficially in the light of information received
• management systems installed that ensure long-term survival of the restored habitat—ensuring low revenue costs is important
• large areas; this is especially important for biodiversity conservation as large areas hold more species and contained populations are less vulnerable to external influences

Failure of wetland restoration projects often results from inadequate planning and typically arises for a variety of reasons:

• lack of scientific knowledge and the dearth of information in the field
• inadequate consideration of the site’s hydrology
• inadequate consideration of the site’s topography (actual or designed)
• lack of practitioner expertise in the two key fields of hydrology and soil science
• inadequate supervision of plant and site personnel during project implementation
• failure to adequately implement project design
• invasion of ‘weed’ species
• inappropriately high grazing pressure on newly developed vegetation

A growing concern is that the increasing availability of restoration technology and improved best-practice must not be used by developers to justify the further damage of natural wetlands around the world. There is no replacement for in situ conservation of natural and seminatural habitats. Conservation of wetland biodiversity, especially in the Northern Hemisphere, can only be assured with a continued investment in restoration, however. Potential restoration sites must therefore be identified and safeguarded by the planning process to ensure that they remain undeveloped and available for restoration in the future.

It must be stressed that wetland restoration is more than the provision of habitat for waterfowl. Restoration of wetlands should be seen as a vital component of integrated watershed management planning. In the past many restoration projects have relied on quick techno-
logical fixes not the creation of self-maintaining, sustainable wetland ecosystems. In the future the scale of projects will increase to take in entire floodplains and the emphasis will be on restoring near-natural functions. Natural processes and ecosystem management will be embraced more widely. This is already apparent in the bold restoration plans for the Kissimmee River in southern Florida and other projects in North America, and this approach should become more commonplace in the future.

Restoration ecologists must seize every opportunity. Sea-level rise is ongoing, and while coastal wetlands will be threatened by marine incursion and the improvement of sea defenses, there is also the opportunity for wetland restoration that will provide flood defense functions and create habitat. The task for environmentalists is to lobby and work with coastal defense bodies to ensure that flood protection and managed retreat schemes result in no net loss of wetland habitat.

Finally, it must be stressed that setting clear objectives, monitoring results, and undertaking research must be undertaken to accurately appraise restoration projects and better the current understanding of the processes involved. The publication of both positive and negative results will ensure that future projects succeed in restoring lost functions and maximizing biodiversity conservation opportunities.

Acknowledgments

The authors thank the following for providing information for this chapter: Dr Paul Erftemeijer (Wetlands International), Tomas Hertzman (Swedish Environmental Protection Agency), Dr Chris Joyce (Loughborough University), Hanne Skotte Maer (Ministry of Environment and Energy, Denmark), Magne Bjune Nielsen (County of Sønderjylland), Crawford Prentice (Wetlands International—Asia Programme), Ole Thorup, and Professor Max Wade (University of Hertfordshire). Finally, we thank the anonymous reviewers of this chapter for their input.

See Also the Following Articles

RESTORATION, CHARACTERISTICS AND REQUIREMENTS • RESTORATION OF BIODIVERSITY, OVERVIEW • WETLANDS: ECOSYSTEMS

Bibliography


I. Introduction

II. Harvest and the Concept of Sustained Yield

III. Endangered Species and Reducing the Risk of Extinction

IV. Managing Wildlife for Conserving the Integrity of Ecosystems and Landscapes

V. Control of Wildlife

VI. Wildlife Management Techniques

GLOSSARY

density dependence Change in the birth and death rate of a population as a result of changes in population density; usually in the form of increased mortality rate and decreased fecundity as density increases.

focal species A species that can be monitored and managed to maintain the integrity of the ecological system of which they are a part.

intermediate disturbance hypothesis A hypothesis claiming that some level of disturbance is necessary to maintain biodiversity.

minimum viable population (MVP) A method for assessing the minimum size a population must be to bring its extinction over a given time span to below a predetermined level.

population viability analysis (PVA) A method of assessing the probability of extinction for a given population within a specified time span.

stochastic processes Processes that affect the dynamics of a population through random chance alone.

sustained yield The removal of a constant number of animals from a wild population that can be maintained through time while keeping the population at a predetermined size.

WILDLIFE MANAGEMENT is the science of manipulating wild populations to achieve a specific goal.

I. INTRODUCTION

Although the term wildlife can be expanded to include all living organisms, the science of wildlife management, as an academic program offered and taught in universities and as written in textbooks, generally refers only to mammals and birds.

Four major goals of wildlife management can be recognized:

1. Maximizing harvest/yield over time.
2. Preventing extinction and increasing population survival probability.
3. Maintaining and managing the integrity of ecosystems and landscapes of which wildlife populations are a part.
4. Controlling wildlife to minimize damage to human crops and assets caused by wild populations or to return the ecosystem to some predetermined state.
As a science, wildlife management relies heavily on the understanding of ecological theory and processes, but differs from the science of ecology in its research objectives. Ecological research is targeted at formulating new hypotheses and providing the data to support or reject them. Thus, the species or setting selected to perform the research is that best suited to test a specific ecological theory. In contrast, research in wildlife is often dictated by real needs of specific problem-species or specific environments. The management of wild populations requires four processes: assessing the problem and defining the management objectives for the target population, formulating and implementing management actions based on the identified problems and ecology of the target population, and evaluating the implemented methodology. These processes can be carried out on a local or global scale. In addition, because wildlife management is an applied science involving what is often a national asset, public opinion strongly influences its goals. Thus, education, law enforcement, and public relations are also legitimate aspects of wildlife management.

II. HARVEST AND THE CONCEPT OF SUSTAINED YIELD

The science and profession of wildlife management have evolved over time based on perceived needs and ecological understanding. Originally, the field of wildlife management addressed only game species and harvest issues, recognizing wildlife as a crop that is a product of the land. The realizations that populations may go extinct due to over-harvesting and that past harvest rates affect the future performance of wild populations were the main forces driving the development of the science (Leopold, 1933). The recognition that populations can be over-harvested is not new (a review of past examples of game management can be found in Aldo Leopold’s book ‘Game Management’). As far back as biblical times, harvesting wildlife intelligently was considered important (Deuteronomy 22:6). According to Marco Polo’s writings, Kublai Khan imposed hunting regulations throughout his Mongol Empire. The first refuge in the modern western world was established by the British Parliament in 1869. However, the formulation of wildlife management as a quantitative science occurred only during the early part of the 20th century, driven mostly by the concept of ‘sustained yield’ and its maximization (e.g., maximum sustained yield; MSY). Leopold’s ‘Game Management’ was a keystone to this process, as was president Roosevelt’s doctrine of conservation.

Sustained yield refers to a constant level of harvest that can be applied that will allow the harvested population to stabilize at a given size. This can be achieved by removing an amount that is equal to the population’s rate of increase. In the absence of limiting factors, the rate of increase of a population should remain unchanged, and, therefore, absolute growth will increase with population size (more animals have more offspring). However, populations do not grow indefinitely, which suggests that at some point the population rate of increase begins to decline as the population increases. This ‘density-dependent’ response is due to a decline in the amount of resources available per individual, which causes a subsequent reduction in reproductive success and increased mortality. The classic ‘logistic growth equation,’ introduced by P. F. Verhulst in 1838, describes the growth curve of such a population, depicting an S-shaped curve of population size against time (Fig. 1a). The rate of increase of the population at time \( t \) is the first derivative \( \frac{dN}{dt} \) and is given by

\[
\frac{dN}{dt} = r_{\text{max}} \left(1 - \frac{N}{K}\right)
\]

where \( N \) is population size at time \( t \), \( r_{\text{max}} \) is the maximum growth rate per individual when resources are not limiting, and \( K \) is the maximum size of the population that can be supported over time in a particular environment (\( K \) is often termed “carrying capacity”). As can be clearly seen, there are two cases when population growth is 0: when \( N = 0 \) and when \( K = N \). Thus a sustained yield requires that the population be kept below \( K \). Between \( N = 0 \) and \( N = K \), growth rate initially increases with population size and then decreases (Fig. 1b). Each point on this curve represents the sustained yield for a given population size, and the highest point is the MSY.

We calculate MSY by finding the maximum value of \( \frac{dN}{dt} \). This can be done analytically by taking the second derivative of the equation depicting the growth curve of the population and equating it to 0. In the case of the logistic equation,

\[
\frac{d^2N}{dt^2} = r_{\text{max}} \left(1 - \frac{2N}{K}\right) = 0.
\]

This equation equals 0 when \( r_{\text{max}} = 0 \) (which is meaningless) or when \( \frac{2N}{K} = 1 \). Rearranging the latter, we find that \( N = \frac{K}{2} \); that is, the population grows the fastest when it is \( \frac{K}{2} \). To find the rate of increase at that point, we simply substitute \( \frac{K}{2} \) for \( N \) in Eq. (1) and find
that the maximum growth rate and, therefore, MSY, is 

\[ r_{\text{max}} K \]

Note that except for the point of MSY, a given yield (growth rate) can be achieved at two different population sizes, one on either side of MSY (\( \frac{K}{2} \))—the upper and lower sustained yields. Thus it appears that one may achieve the same level of sustained yield with a small population as with a large one. However, the two points differ considerably in terms of stability. The points to the right of MSY are stable because moderate over-harvesting will increase the growth rate of the population, thus compensating for the excessive harvest (and vice versa). In contrast, the points to the left of MSY are unstable because over-harvesting will cause a decrease in the rate of increase. As a result, the next harvest will further decrease the population and rate of increase and so on. Initially the impact may appear minor, but this "snowballing" effect can very quickly lead to the demise of the population.

Clearly, real populations do not behave deterministically as in as the logistic model. Density-dependent responses may vary depending on the species and environment skewing the \( \frac{dN}{dt} \) curve; the age structure of the population must be considered; the environment (carrying capacity) constantly changes; other species (predators, competitors, etc.) interact with the population; and hunters may select for specific animal traits (sex, age group, size, etc.). Thus, much of the research on harvested populations has been directed at developing more sophisticated models with better predictive capabilities and estimating the parameters needed for them (Getz and Haight, 1989).

Although many advances have been made in this field, to date, harvest strategies in many places around the world are still implemented on a trial-and-error basis. However, even under such protocol, harvest theory has made an important contribution in demonstrating that uncontrolled harvest can very quickly lead to the demise of populations. Furthermore, while the logic of a sustained yield set somewhat above MSY is obvious, nonecological factors may have a strong influence on harvest policy. In many species, the long-term monetary income from a sustained yield may be less than the yield offered by the economic market on regular investments (Clark, 1976). In other words, economically, a greater profit can be realized by selling licenses to hunt the entire population and investing the profits than by limiting the number of licenses to a sustained yield that will ensure the continued existence of the population. Thus, in recent years effort has been directed at assessing the economic value of wildlife for nonconsumptive uses such as tourism and conservation.

III. ENDANGERED SPECIES AND REDUCING THE RISK OF EXTINCTION

With the biodiversity crisis coming into focus during the second half of the 20th century, the science of conservation began to evolve. Initially, conservation and wildlife management were treated as separate sciences. Eventually, the sciences merged in the area where they overlap (wildlife), and managing wildlife populations for conservation became an integral part of wildlife management. Topics such as management for recovery and sustainability, the dynamics of small population (especially as impacted by demographic, genetic, and environmental stochastic processes), fragmentation, and the importance of populations to the integrity of the ecosystem and biodiversity became a central focus of wildlife management. In terms of management for conservation, Caughley and Gunn (1996) recognize two central paradigms: the small population paradigm and the declining population paradigm.
The sensitivity of small wildlife populations to extinction was first recognized and expressed in 1938 in a book written by the famous American ecologist W. C. Allee. In his book, Allee writes: “The general conclusion seems to be that different species have different minimum populations below which the species can not go with safety, and that in some instances this is considerably above the theoretical minimum of one pair.” In 1981, Shaffer defines four stochastic elements that play an important role in the dynamics of small populations and contribute to their increased risk of extinction: demographic stochasticity, genetic stochasticity (including loss of genetic variance and inbreeding depression), environmental stochasticity, and catastrophes.

The accelerated loss and fragmentation of wildlife habitat during the second half of the 20th century restricted many wildlife populations and required that management steps be taken to minimize these stochastic effects. The size of refuges and other areas set aside for conserving wild populations must be large enough to support a population equal to or larger than the minimum viable population (MVP) (Belovsky, 1987). Corridors connecting small populations are currently considered important (Bennett, 1999), but their effectiveness has not yet been fully demonstrated. Where the establishment of corridors is not feasible, translocation has been implemented. For instance, in South Africa, mountain zebras are translocated by airlift between isolated populations. Habitat improvement and resource supply during limiting periods of harsh environmental conditions are also being used to minimize population declines and reduce the impact of stochastic processes and risk of extinction.

The Management of declining species requires that the cause of decline be identified and managed directly or indirectly. Human-related decline in wildlife populations in the 20th century can be attributed to three main factors: over-harvesting, loss and fragmentation of habitat, and the introduction of alien species. With the improved control of harvest by science and legislation, loss of habitat and fragmentation became the major causes of wildlife population declines. This process has highlighted the importance of considering space in wildlife management. Management efforts have been concentrated on land allocation for conserving wildlife and securing refuges and linkages (corridors) between patches (fragments) of adequate habitat. Roads and fencing are some of the major causes of fragmentation, and special overpasses, underpasses, and gates have been developed to ensure wildlife movement.

Wildlife populations are important for maintaining biodiversity in the system of which they are a part. Many wild populations are critical for the functioning of the entire community. Beavers (Castor canadensis) are a classic example. Thus, in many instances, populations are managed in order to sustain species richness of the ecosystem of which they are a part. For example, research has shown that both overgrazing and undergrazing will reduce plant species richness. This is known as the intermediate disturbance hypothesis (Connell, 1978). Thus, in certain cases, herbivore populations are managed to maintain the diversity of the plant community. Elk (Cervus elaphus) populations were regularly controlled to maintain the existing landscape in Yellowstone National Park. Wolves (Canis lupus) were reintroduced in 1995 (Fritts and Carbyn) into Yellowstone with the hope that they would control the elk population. However, this is yet to be seen.

The use of focal species as a tool to manage ecosystems has gained popularity in recent years. The underlying assumption is that by managing the focal species, the entire system can be secured. Indicator, keystone, and umbrella species are the main types that have been suggested as focal species through which systems can be monitored, conserved, or managed. Indicator species are species that testify to the well-being (health) of the entire ecosystem. These species are the first to respond to deterioration of the ecosystem, since they are more sensitive. By monitoring and managing the indicator species population only, we can estimate and protect the welfare of the entire ecosystem. We assume that as long as the status of the indicator species is satisfactory, the entire system is operating adequately.

Keystone species are species that play an important role in the ecosystem. These may be animals that hold most of the biomass in the ecosystem or that influence many of the other species or functions of the ecosystem. In this way, if the population of a keystone species is removed from the ecosystem, the ecosystem changes dramatically. Therefore, by managing only the population of the keystone species, we are managing most of the ecosystem.

Umbrella species are species that require a large area containing many types of habitats to sustain a viable population. Often these are large bodied homeotherms with large home ranges. By securing a tract of land large
enough to sustain a viable population of these species, many others will come under the same protection. The focus here is on leaving enough area for the umbrella species and other members of the ecosystem, but there are no direct management implications for this approach.

V. CONTROL OF WILDLIFE

The ever-expanding human population ensures that human–wildlife conflicts will continue. Most of these conflicts are within the realm of damage to crops, equipment, and other assets. Thus, a major component of wildlife management is wildlife control. A good example is managing habitats around airports to reduce gull activity near runways that endangers aircraft during takeoff. Often, damage caused by wildlife is the result of human manipulation of the environment, such as outbreaks of herbivores following the eradication of predators and high concentration of canids around human waste sites leading to outbreak of rabies. Wildlife management strives to implement control measures in a manner enabling the other goals of wildlife management (sustained yield or maintaining biodiversity) to be met.

The four processes necessary for wildlife management generally apply for wildlife damage control, with some differences. These processes are the following: (1) defining the problem. The species causing the damage must be identified, the type of damage determined, and extent of the damage evaluated. (2) knowledge of the general ecology, behavior, and dynamics of the species causing the damage must be obtained and evaluated with special regard to its response to various control techniques. (3) The combination of (1) and (2) defines the methods that are feasible and economical to control the damage. (4) The methodology used must be evaluated. Unless the cost of control is less than the losses due to the damage, management should not be implemented.

VI. WILDLIFE MANAGEMENT TECHNIQUES

Because wildlife management is an applied science, methods and techniques for both research and implementation are important. As a result, a considerable amount of research has been directed at developing, improving, and assessing techniques. Techniques for managing wildlife are aimed at studying, reducing, increasing, or maintaining the population at its current level while securing its integrity (i.e., preventing loss of genetic diversity, enhancing long-term survival, preventing epizootics, etc.) and the integrity of its ecosystem. The techniques can be classified generally into two categories:

1. Applied techniques for manipulating populations by impacting survival, reproductive success, or distribution.
2. Techniques for studying, analyzing, and assessing population.

Many of the more common techniques are published in a series of “Wildlife Techniques” published by the Wildlife Society. The first of this series, edited by H. S. Mosby, was published in 1960. The most recent issue (the fifth edition) was published in 1994 and edited by T. A. Bookhout. The book covers four main topics: experimental design, laboratory techniques, population analysis and management, and habitat analysis and management.

A. Applied Techniques for Managing Wildlife

Applied techniques may be resource (habitat) manipulation, mechanical, behavioral, immunization and immuncontraception, biological, and direct population control.

Habitat and community manipulation. This refers to managing target species by impacting important resources and other species that interact with them (i.e., habitat manipulation, or impacting predator–prey relationships and competing species). By controlling key elements that affect the species’ abundance (water, shelter, predators, etc.), it is possible to manipulate its densities and distribution. The increase in computer power enabled the use of spatially explicit models and Geographic Information Systems techniques to project the impacts of various habitat manipulations (Verner et al., 1986).

Mechanical. The simplest and most successful method of controlling wildlife damage is exclusion by fencing. The advantages of fencing are that it is relatively fail-proof and nonlethal. However, it is expensive and labor intensive, demanding constant upkeep. Furthermore, fencing over large tracts of land is a major cause of fragmentation.

Behavioral. These are modern forms of scarecrows that operate on the animal’s senses and are, therefore, acoustic, olfactory, or visual. A good example of visual
repellents, other than the classic scarecrow, is the silhouettes of birds of prey that are placed on large glass windows to prevent other birds from slamming into them. Examples of acoustic deterrents are gas cannons that produce explosions at irregular time lapes or recordings of alarm calls of other animals. Ultrasonic deterrents work by using sounds at a high decibel level (above 120 dB), outside the range of human hearing, to cause a painful stimulus to animals with sensitive hearing, such as canids. Olfactory repellents are chemical compounds. Two types of chemical compounds can repel animals: (1) compounds that link a food source with induced sickness through a behavioral process called conditioned taste aversion and (2) compounds that repel through unpleasant stimuli of the nervous system. The first type is properly referred to as aversive agents, while the second type are true sensory repellents.

Immunization and immunocontraception. These techniques require the administration of a vaccine either directly, by capturing and injecting, or indirectly, by oral administration with bait. A major advancement in wildlife management in the past decade is the control of rabbits by oral immunization. Immunocontraception is a very attractive method of control because it is potentially highly species specific, but so far has had limited success.

Biological. Biological control has been mostly effective against insects. There are only a few cases in which biological control was effectively applied to wildlife. Of these, the introduction of the myxoma virus to control rabbits introduced into Australia is the most famous. Although complete eradication was not achieved, the virus now holds the rabbit population at 20% of its original size.

Direct population control. Direct population control refers to the removal (harvesting) or the addition (reintroduction and translocation) of animals. Removal is usually used when the animal is a game species and can also be controlled by hunting license quotas. Its main advantages are that it is species specific, the number of animals removed can be closely controlled, and it has limited impact on the environment. Reintroductions and translocations are important methods for enhancing the viability of existing populations and reducing the species risk of extinction.

B. Studying, Analyzing, and Population-Assessing Techniques

A major prerequisite of sound management is the accurate estimate of population size and dynamics, and the determination of how these are influenced by other factors. Consequently, a significant part of the science today is directed at the development and improvement of estimation techniques for assessing population density, survival, reproduction success, well-being, age structure, density-dependent responses, and interspecific interactions. Such data are the basic requirements for estimating sustainable harvest quotas in different and variable environments and assessing extinction probability and other risks.

Research and management of wildlife are carried out over large geographic areas, often with limited access and visibility. Considerable effort has been directed toward developing remote data collection techniques, such as radio telemetry and other marking methods, as well as indirect methods of assessing population condition by using individual animal physiological indicators (Harder and Kirkpatrick, 1994). Rigorous statistical methods, based on maximum likelihood estimation techniques, are continuously being refined. Computer software for assessing density and survival, such as DISTANCE, SURVIV, ESTIMATE, JOLLY, BROWNIE (Lancia et al., 1994), and most recently MARK, were developed specifically for wildlife research and management.

Methods for assessing the viability of small populations (population and habitat viability analysis, PHVA) and for estimating the minimum population size necessary to ensure its long-term existence (MVP) have been developed based on stochastic theory (Burgman et al., 1993). Commercial software such as RAMAS and Vortex are available for PHVA and MVP analysis; however, in most cases and due to the detailed knowledge required, the data are insufficient for a reliable analysis.

See Also the Following Articles

ENDANGERED BIRDS • ENDANGERED ECOSYSTEMS • ENDANGERED MAMMALS • INDICATOR SPECIES • KEYSTONE SPECIES • SUSTAINABILITY, CONCEPT AND PRACTICE OF

Bibliography


I. Morphology

II. Reproductive Biology

III. Systematics, Diversity, and Phylogeny

GLOSSARY

chaetae Bristles composed of slender chitin cylinders glued together by scleroprotein, emerging from parapodia or body wall in most polychaetes.

nuchal organ Sensory organ with nerves coming from the posterior part of the brain, or the nerve ring, usually present as paired ciliated pits or short ridges at the posterior end of the prostomium; present in most polychaete species.

parapodium Segmentally arranged outpocket from the body wall; usually divided into two branches (rami), the dorsal notopodium and the ventral neuropodium; supports the chaetae; extremely variably developed and may be absent.

peristomium Region around the mouth of the larvae; if distinct in adults may carry peristomial ciri.

prostomium Morphologically anterior-most part of the body; may carry antennae and eyes and contain at least part of the brain.

segment In annelids a section of body set off from the rest of the body by septae at both ends containing a separate part of the secondary body cavity (coelom), with paired ventral ganglia, and usually nephridia,

gonads, transverse muscles, and segmental blood vessels; carrying externally parapodia and gills.

trochophore Larva characteristic especially of annelids and molluscs consisting of an episphere, ciliated griddle around the middle called the prototroch, and the hyposphere. In feeding trochophores the mouth is located behind the prototroch and the anus is terminal, usually surrounded by a second ciliated band, the telotroch. Segmentation takes place between the prototroch and the telotroch and is initiated immediately in front of the telotroch.

POLYCHAETA is a commonly encountered group of annelids (segmented worms) best represented in the marine environment, although several species are present in freshwater and a few are known from moist soil on land. In addition to the polychaetes, the Annelida contains one additional species-rich group, the Clitellata, which includes the earthworm and their aquatic relatives and the leeches. The clitellates evolved from the polychaetes, but the exact relationship has not yet been diagnosed. The Polychaeta, the many-bristled worms, includes about 15,000 described species. Estimates of the total numbers of species varies, but judging from the rates of description of new species it is likely that more than 20,000 species will eventually be recognized, based on morphological characters in addition to the study of all genetically fixed properties of the worms. The shape of these worms varies a great deal.
This is reflected in the description given later. No single feature can be used to diagnose the taxon uniquely; until recently the presence of the nuchal organ was considered such a feature, but even this structure is missing in several polychaetes and present in several nonpolychaete taxa.

I. MORPHOLOGY

The polychaetes are composed of three body-regions derived from structures present in the larvae. The pre-segmental prostomium and peristomium are followed by a segmented body and a post-segmental pygidium that carries the anus. Most polychaetes are less than 100 mm in length, with 100 or fewer segments; many are very small, with as few as 10 segments, and are adapted to life between sand grains in sandy beaches. Some species may become very large, with 2 to 3000 segments and lengths between 1 and 3 m; the longest recorded specimen was 6 m long when collected in Port Jackson, Australia. These very large worms may be as old as 90–100 years, but the documentation for this is rather poor. In general, small species are short-lived compared to larger relatives. Perhaps most polychaetes live to 12–18 months, while others go through three or four generations in a year. The general body shape may be earthworm-like with cylindrical bodies and very little in the way of external appendages, but more usually the protruding parapodia and chaetae makes the worms appear ragged. The ventral side may be flattened with two longitudinal ridges on either side of the mid-
WORMS, ANNELED

A. The Head

The head consists of the prostomium and peristomium in adults, but one or more of the first segments may be fused to this structure (see Figs. 1–3). The prostomial part may carry antennae and eyes, and the fused anterior segments may have one pair of tentacular cirri per fused segment; these cirri are the dorsal and ventral cirri of each segment participating in the head structure.

B. Segments

Segments are always added just in front of the pygidium; that is, from the posterior end. They are often seen first as a low ridge in front of the pygidium; parapodia, chaetae, and other externally obvious structures are differentiated as more segments are added and, as a consequence, each segment no longer has the position

ventral nerve chord; the dorsal side is usually convex. This description fits best for free-living, medium-sized worms; the tubicolous worms are often cylindrical with inconspicuous parapodia, the very small species are often very slender, and the parasitic taxa may be flattened and disc-shaped.

In Canalipalpata, the head end can be complex and the numbers of segments included in the head may be difficult to count. Some of these worms have a tentacular crown (e.g., the sabellids and serpulids) or a very large number of antennae (e.g., terebellids). Certain polychaetes lack anterior appendages entirely (e.g., capitellids and maldanids).

FIGURE 2 Polychaete anterior ends. (A) Eunicid, dorsal view. (B) Flabelligerid, dorsal view. (C) Hesionid, dorsal view. (D) Opheliid, lateral view. (E) Orbiniid, dorsal view. (F) Oweniid, lateral view.

(Modified from Rouse and Fauchald, 1997.)
of the last segment. The total number of segments may be fixed and limited (e.g., 20, 25, etc., in certain polynoids, maldanids, and opheliids). In many species the number of segments varies with relatively narrow ranges; for example, many nereidids species have 90–120 segments. Some species have an unlimited number of segments and the worm keeps adding segments throughout life. Normally each segment has a pair of parapodia (see Fig. 4). In many species the parapodia are poorly developed and marked only by the position of the chaetae; in other taxa the parapodia are complex structures with a variety of lobes and lamellae. In well-developed parapodia, the two rami often differ. The notopodia are smaller (or shorter) than the neuropodia in many species; when they are completely missing, the parapodia are called uniramous. In the identificatory literature a number of terms (e.g., sesquiramous) are used to characterize the level of reduction of the notopodia. These terms are used inconsistently, and it is better to specify the structures observed than to apply a poorly understood term. When the parapodia are best developed each ramus has a chaetal lobe supporting one or more bundles (or fascicles) of chaetae and pre- and postchaetal lamellae, often covering the bases of the emerging chaetae. The chaetal lobe is supported by one or more aciculae in many species. A dorsal cirrus is often found on the dorsal edge of the notopodium or on the adjacent body wall and a ventral cirri is along
the ventral edge of the neuropodium; both cirri are mainly sensory structures. In addition to the pre- and postchaetal lamellae, various other lamellar or finger-shaped structures may be associated with each ramus, and a complex terminology is associated with the position and shape of each structure.

C. The Pygidium

The Pygidium may be a rounded or tapering simple structure with the anus opening either dorsally or terminally, but frequently it carries anal (or pygidial) cirri. Commonly one or two pairs of cirri are present, but
in some taxa the pygidium may be ornamented with numerous anal cirri or flattened lamellae (e.g., maldanids), and in the sipunculids it may have an open hood with many papillae attached both inside the hood and along the margin.

D. Chaetae

Chaetae, the bristles, may be in bundles; however, in most cases the “bundles” consist of two or more short, oblique rows of a few chaetae each. Other polychaetes have chaetae in flattened fascicles usually arranged dorsoventrally. Each chaeta is produced from an invaginated epidermal cell (chaetoblast). The outer cell membrane of each chaetoblast is covered with microvilli, each of which produces a slender cylinder of chitin. These cylinders are glued together by scleroproteins produced by cells lining the invagination. All chaetae are formed the same way, through complex interactions between production of cylinders and glue secreted to keep the chaetae together and give the external shape. The chaetae may be tapering capillaries, but most chaetae are much more complex (Fig. 5). A species may have a single kind of chaetae or two or more kinds; in some parapodia each chaeta may be different, but more usually groups of chaetae with a similar structure are present. New chaetae are formed throughout the life of the worms, and distribution of different kinds of chaetae along the body is carefully controlled and changes as the worm grows. In rows of identical chaetae, new chaetae are produced at one end of the fascicle and migrate through the fascicle and drop out at the other end (e.g., hooks in terebellids). In syllids, which often have few chaetae, and in which each chaeta may differ from all other chaetae in a parapodium, replacement takes places in a one-for-one fashion. Special kinds of chaetae, such as aciculae, which are rods supporting long parapodia, and various large spines often show signs of wear, an indication that they are not replaced or at least not replaced as often as other chaetae. The base of the chaetae (chaetal sac) is anchored with muscles running across the body cavity, making it possible to protrude and retract the whole fascicle or bundle since in most cases only a single set of muscles is present. Aciculae and large spines often have their own investment of muscles and are independently movable.

E. The Nervous System

The nervous system consists of a dorsal brain connected by a double nerve ring running around the esophagus to double ventral nerve chords. The anterior part of the brain is located in the prostomium, but penetrates well into the peristomium or even into the next several segments in many species. In most groups the two parts of the ventral nerve chord are wholly or partially fused into a single entity; each segment always contains a pair of ganglia from which the segmental nervous system runs to the parapodial muscles, nephridia, gonads, and circulatory system. The gut is invested from a separate set of nerves from the lower part of the brain (stomatogastric nerves). Antennal nerves run from the anterior part of the brain, and a pair of nerves associated with the posterior part of the brain runs to the palps. Indeed, Orrhage has found a pattern in the emergence of nerves and internal connections in polychaete brains, showing that the limited number of nerves can be identified from one group to the next based on position and function. The ventral nerve chord may contain two or more giant fibers associated with rapid retraction into burrows or tubes consisting of very thick nerves running from the anterior end innervating the longitudinal muscles directly with only a minimum of synapses. Other motor fibers are segment-to-segment loops passing along undulatory locomotory waves.

F. Sense Organs

Eyes are mostly eye-spots located on the prostomium or even directed on the upper surface of the brain. In some polychaetes the eyes become much larger when the species are sexually mature (e.g., neredidids), and the epidermis over the eye becomes a translucent lens. In a group of pelagic polychaetes (e.g., alciopin phyllocodiids) the eyes are very large camera constructions with a distinct lens and with focal capabilities of considerable complexity. In the polychaete eyes, the sensory cells penetrate through a pigment layer (retina) into the space in which the lens is supported, and the optical nerve is collected behind the pigment layer; thus the retina is reversed compared to in vertebrate eyes.

The nuchal organs are usually located in pits, in short grooves, or on short ridges near the junction between the prostomium and the first segment. The nuchal organs may be very large, complexly folded structures (caruncles) in the euspriomnids and amphipomnids, and the epaulettes present in some syllids are also nuchal organs. In some spinoids the nuchal organs stretch along most of the anterior region of the worm as narrow, ciliated attached ridges. The presence of nuchal organs has been considered diagnostic for the polychaetes in relation to the clitellates, which lack these structures. However, some polychaetes lack nu-
WORMS, ANNELIDA

FIGURE 5 Polychaetae chaetae. (A–E) Compound chaetae showing a variety both of hinges and of the dorsal ends of the chaetae (A, sigalionid; B, nereidid; C, syllid; D, flabelligerid; E, eunicid). (F) Capillary chaeta of hesionid. (G) Spine of a polynoid. (H) Hook of a spionid; note the two teeth at the end. (I) Hook of a malacostracan; note the many teeth at the end and the spray of bristles under the main tooth. (J) Uncus of a chaetopterid; note the straight edge with the many small teeth along it. (Modified from Rouse and Fauchald, 1997.)

chal organs, and sensory structures closely similar to the muscal organs are present in more distantly related worms (e.g., sipunculans, the peanut worms), so the presence of these sensory organs may be a shared feature of a large group of invertebrates and the loss among certain annelids may be a secondary phenomenon.

Ciliated lateral organs are located between the parapodial rami in many polychaetes; they are usually tufts of cilia in a shallow pit, but may form eversible papillae in some taxa. Similarly constructed patches of cilia have been reported from a variety of polychaetes and may form patterns of sensory organs characteristic of each of the major groups. All of these ciliated structures are assumed to be at least partially chemosensory, but the experimental evidence is usually missing.

G. Musculature

Musculature includes two pairs of longitudinal muscle bundles; these bundles are usually ovate in cross-section, but especially in species with massive longitudinal muscles, they are folded over presumably to avoid buck-
WORMS, ANNELIDA

ling during strong contraction. The ventral pair is usually heavier than the dorsal pair, which in some worms (e.g., terebellids) forms a thin layer between the upper edge of the notopodia and the dorsal mid-line. Circular muscles are usually considered to be present; however, at best they are present near the septal boundaries. Most of the other segmental muscles are associated with the parapodia, the chaetae, or are dorsoventral transverse muscles. Whether these antagonists to the longitudinal muscles represent the remnants of a former complete layer of circular muscles has yet to be determined.

H. Septa

Septa separating the segments are rarely complete, but in many cases they are sufficiently muscularized, with radiating and circular muscles to allow each septum to clamp down around the digestive tract and at least temporarily isolate each segment. In certain families one or more of the anterior septa may be complete (gular membranes) and may isolate an anterior part of the body cavity from the rest of the body.

I. Digestive Tract

The digestive tract is usually a more-or-less straight tube running from mouth to anus. The larval stomo-
daeal region (infolded ectodermal region in the mouth) may be elaborated into an eversible pharynx (probos-
cicus). This pharynx may be axial or may include only the ventral side; it may be sac-like or muscular. Jaws are present in some taxa; they are thickened regions of the cuticle and as such consist primarily of sclerotinized protein in which a variety of metal ions may be imbed-
ded, or they may be calcified. Paired salivary glands are often present. The middle of the gut may be invested with a large blood sinus or is at least well supplied with segmental blood vessels. In postlarval siboglinids (Pogonopora + Vestimentifera auctores) the gut is blind-ending at both ends and the lumen of the gut is largely obliterated, but a very inflated specialized gut wall is present. The cells of this gut wall contain numerous commensal bacteriae. The scale worms have numerous segmentally arranged blind sacs attached to the mid-gut; at least some part of the intermediary metabo-
lism takes place in these sacs, which also in part supple-
ments the circulatory system in bringing nutrients to the tissues in which they are needed. Some polychaetes have a looped gut; in these taxa the septation is missing in the region that contains the loop. Looping allows for an increased length of gut in relation to the size of the worms.

J. Gills

Gills are present in a variety of taxa. They are usually associated with the notopodia but may be present any-
where along the body; they are recognizable as thin-
walled extensions from the body wall containing a vas-
cular loop in which the two limbs of the loop are linked through capillaries. This definition excludes the struc-
tures called gills in the capitellids and glycerids, both of which lack a circulatory system; these structures cannot at this time be considered homologous with the gills present in taxa with closed circulatory systems. Functionally, any thin-walled extension of the body wall will have a gas-exchange function whether it struc-
turally belongs to a given category of structure or not. This issue is of considerable phylogenetic interest in that gills must have evolved independently several times, while in other clades, morphologically different structures, including the notopodial cirri, have taken on respiratory functions.

K. Circulatory Systems

Circulatory systems may be complete, consisting of a dorsal, often contractile vessel, in which the blood flows anteriorly, one or two major ventral vessels, and capil-
laries connecting these vessels to the gut and to the rest of the organs. Some polychaetes lack the capillary beds and thus have an open circulatory system, and small species often lack a circulatory system entirely, presumably relying on diffusion for transportation of oxygen, carbon dioxide, and so forth. Shared absence of a circulatory system is not considered a feature of phylogenetic importance; it seems to be linked to a reduced body size.

L. Nephridia

Nephridia may be present already in the larvae as proto-
nephridia; the postlarval nephridia may be either proto-
nephridia or metanephridia. Protonephridia open into the body cavity with intracellular slits; metanephridia open with a funnel. In all postlarval polychaetes the inner end of the nephridia is located in the segment preceding the one in which the duct opens to the exte-
rrior. About 100 years ago, Goodrich identified a series of possible fusions between the nephridia and the exit ducts for the gonads (e.g., protonephromixia, meta-
nephromixia, mixonephridia). New information sug-
gests that this classification may not represent the situa-
tion accurately, and that it is more complex than previously understood. Nephridia are often present in many segments, especially when the segments are uni-
lar in structure; however, in many polychaetes the number of pairs of nephridia may be limited to four to five. Excretory nephridia are always located anterior to nephridia, which functionally become gonoducts when the worms are sexually mature.

M. Gonads

Gonads may be present in many segments, or they may be limited to several or even just a few segments in the middle of the body. Sexes are most frequently separate, but hermaphroditic taxa are scattered throughout the group. The gonads are usually very simple, part of the peritoneal lining of each segment, sometimes attached to the transverse or oblique muscles. In many polychaetes only the first few cell cleavages leading to eggs and sperm take place in the gonad proper; the rest of the development takes place in the coelomic fluid in which the eggs and sperm float. The developing spermatocytes may have incomplete early mitotic cleavages, leading to flattened sheets or rounded balls of cells in which the head end of the sperm is embedded in the shared structure; these structures are called sperm modules. The yolk of the developing eggs may be furnished by nurse cells, either in a layer or as two strings. However, in many species no nurse cells are present, but the coelomocytes function at least in part to furnish the building blocks of yolk for the eggs and the general nutrients to the developing sperm.

II. REPRODUCTIVE BIOLOGY

Most species spawn in the water, the site of fertilization as well as development up to settlement; indeed, several species swarm on the surface, such as the palolo-worms of tropical waters, sometimes while emitting luminescence (e.g., Odontosyllis). A few species have modified some of the parapodia for copulation, but this is rare and limited to certain small species. However, we are increasingly finding that females may store sperm in spermathecae, which may be segmentally arranged or at or near the anterior end. Sperm morphology has been related to the modes of fertilization. The round-headed sperm with a short acrosome and a long tail has been considered primitive since it is present in taxa that spawn into the water column. When fertilization takes place elsewhere, the sperm usually has another structure, often with greatly elongated heads and short tails.

In taxa with both feeding and nonfeeding (or brooded) larvae, eggs are much larger in the taxa in which the larvae do not feed. Nonfeeding larvae are present in all major groups of polychaetes and in some families are the only kind of larvae present. The shift between feeding and nonfeeding larvae must have evolved more than once; the direction in which this shift has taken place cannot be considered settled in many cases. Early development may take place in open water, in deposited egg masses, in brood-chambers of various sorts, or internally, in the body of the females. After the early cleavages, which follow the pattern called spiral cleavage as in many related phyla, and gastrulation, a trochophore larva develops. The trophophore consists of an upper episphere with a sensory tuft of cilia on the top, and often a pair of small eyes. The episphere is separated from the hyposphere by an encircling prototroch, a ciliary band which in the polychaetes may consist of anything from a single row of ciliated cells to a broad densely ciliated girdle. The mouth is located on the ventral side, immediately posterior to the prototroch and is often located at a peristomal cirrall band, the gastrotroch. At the posterior end of the larva, a third cirral band, the telotroch, is often also present. In feeding larvae the gut is fully developed and a pair of larval nephridia are present. When metamorphosis starts, segmentation starts posteriorly, just in front of the telotroch; internally the segments are first visible as paired mesoderm blocks on either side of the gut. These blocks eventually become hollow, and the mesoderm forms the peritoneal lining and muscles and most other structures in the developing segment. The space inside each block becomes the secondary body cavity, the coelom, which eventually contacts the exterior through nephridial ducts. Metamorphosis in most polychaetes is a gradual process of adding more segments with a progressive reduction in the prototroch and other larval structures; the body proportions of a newly settled juvenile may reflect what the adult worm will eventually look like, but in many cases considerable change in relative proportions of various body regions takes place during the later development. The body of the polychaetes may appear to consist of a series of repeated units, but as we become more familiar with postmetamorphic development it is becoming increasingly obvious that the polychaetes maintain integrated bodies with distinct proportionality between different regions.

III. SYSTEMATICS, DIVERSITY, AND PHYLOGENY

Traditionally the polychaetes were separated into errants and sedentaries (two equivalent groups); since
WORMS, ANNELIDA

FIGURE 6 Diagram showing the current systematic subdivision of the Polychaeta. Only 54 of the 80 families have been included; most of the remaining families can be readily associated with the taxa where they fit but are sufficiently poorly known or lack so many features that including them would have distorted the current analysis. Eventually we expect to be able to incorporate all families. (Modified from Rouse and Fauchald, 1997.)

In addition to the approximately 80 families traditionally included among the polychaetes, a major change was proposed by Rouse and Fauchald. The pogonophorans (including the vestimentiferans) have been considered a separate phylum (phyla). Their major features match the polychaetes well, and they are included here as a single family of polychaetes, Siboglinidae. Some of the unusual features, such as the "absence"
of a gut is incorrect: the siboglinids have most tissues associated with the gut, but in the adults the gut is closed both anteriorly and posteriorly, and the gut lumen is nearly obliterated. Evidence from molecular systematics furnishes good criteria for including the siboglinids among the polychaetes, even if the exact placement among the canailipalpates is still uncertain. Rouse and Fauchald included the myzostomids among the polychaetes; this group has been considered distinct for more than 100 years, and recently it has been shown that the apparent similarity to the polychaetes is incidental (see especially papers by Eeckhaut and collaborators). The myzostomids are associated with echinoderms and have a fossil record stretching back to the paleozoic, so they certainly have a long separate evolutionary history.

The fossil record of the polychaetes is, not surprisingly, uneven, considering that they are soft-bodied with very few readily fossilizable structures. They were present in the middle Cambrian (in the Burgess Shale) in taxa resembling, in certain respects, recent canalipalpates. Whole body fossils have also been reported from other paleozoic strata, and the presence of recent families has been suggested. Many records lack the diagnostic characters of recent taxa so the presence is often difficult to document. Calculated tubes similar to those present among the serpulids are present from early paleozoic strata on; some of the tubes may have belonged to nonannelid taxa, but certainly some of them were serpulids. The upshot of the scattered records shows that most major taxa of polychaetes must have been present by the Ordovician and some clades may be even older than that. This suggests that each group has a long separate evolutionary history. Most of the groups were thus present before Pangaea was formed, which agrees with the very wide dispersal of all major polychaete taxa: they have been rafting on continents. On the other hand, there is much evidence to show that individual species are capable of very rapid dispersal as invasive species, probably mostly through transportation of adults.

Polychaetes are important components of the bottom fauna at all depths of the oceans from beach sands to the soft muds of the deepest trenches. Recently a rich fauna of very unusual polychaetes have been found in both hot vents and cold seeps. Various polychaetes are also present in the pelagic environment throughout their life; more commonly the polychaetes have benthic adults but are present in the plankton as larvae for a varying period of time. In certain areas the polychaetes may make up as much as 90% of the standing crop of benthic macrofauna, so any account of life in and on the sea bottom must take into account the activities of these worms.

Polychaetes often appear to be very widely dispersed geographically, and some species are considered cosmopolitan. Careful studies have in many cases shown that the records on which these distribution maps were based are erroneous. More than one species have been called by the same name; later investigations have shown that what was originally thought to be a single taxon often represented several distinct taxa. This correction of the geographical records is important for our understanding of the biodiversity of many areas, and is in part the result of increased quality of analysis based on the development of statistical techniques and the emergence of molecular systematics or has been demonstrated in experimental studies. Many polychaetes are genuinely widely dispersed, however, possibly because they can be transported on ship bottoms and in ballast water. Other forms of accidental introductions also have been demonstrated. Importation of a species of abalone from South Africa to California for use in aquaculture brought along a small sabellid polychaete that burrows into the shell of the South African abalone; the sabellid escaped and is now doing damage to native abalones not only in aquaculture but also in the wild. Other polychaetes, especially members of the family Spionidae, have been transported with importation of oysters; they burrow into the shells of the oysters, forming mud blisters, which may have serious consequences for the oyster industry in certain areas.

Polychaetes may move around at the surface of the substrate on which they live, but even the most active species are usually cryptic, nesting in cracks and crevices, under the cover of sea weed or debris or along the edge of rocks where these meet sand or mud. Many polychaetes from a variety of higher taxa form tubes or burrows in which they live more or less permanently. Some, such as the onuphid Hyalinoecia are capable of moving their quill-like tubes along, but others, such as the related Diopatra are sessile; in this genus the tube consists of a flaky grey matrix which, where it reaches above the substrate, is decorated with shell fragments, wood or sea grass debris, and sea weed. The sabellids,
especially the larger species, form tough but pliable tubes, but the closely related serpulids have their tubes impregnated with calcium carbonate. Some species, such as certain glycerids, live in complex galleries with several openings on the surface; the worm moves around in the galleries so that they can intercept small prey-organisms moving around on the surface of the mud or sand.

The presence of polychaete tubes creates a microenvironment in many areas so that the sheer presence of these worms modifies sedimentation patterns and influences the relationship between the surface of the sediment and the overlying water.

In general, polychaetes feed on anything organic. Some taxa are obligate carnivores, but most species will feed basically on any kind of plant or animal fragments present, live or dead. Several species take food particles out of the water column, mostly by creating small back eddies in flowing water so that particles in transport in the water will drop out of the flow. Only the chaetopterids are truly filtering the water for contained food and are specialists of that mode of feeding. The neridid polychaetes have been shown to be capable of filtering as well, but they are not morphologically structured for this activity. This may also be true for several other polychaetes that are capable of building temporary tubes and driving water through the tube with undulatory motions. Incomplete mucus plugs will catch particles in the stream. Many polychaetes feed by capturing small, light-weight particles in mucus strands and transporting them to the mouth along the palps, feeding mainly on the light fraction of particles in bottomload transport at the water–sediment interface, often creating or taking advantage of turbulence developing around irregularities in the bottom. Some polychaetes feed at depth at the bottom of their tubes (maldanids) and are thus capable of turning over sediments. The areniculids are especially well known for their feeding in the inerridal sands and mud flats; these worms are large and as a consequence have considerable effect on the life conditions for all other inhabitants in the bottom. They are buried sufficiently deeply to avoid being the prey of plovers and other birds.

Certain polychaetes, such as the capitellids, include some of the most pollution-resistant species in the sea, capable of surviving in inner harbor waters where the organic pollution load is very heavy. Other polychaetes such as some glycerids are extremely sensitive to change in sediment composition and oxygen levels in the water. This range of reactions to disturbance has made polychaetes much used in environmental research. Most of the taxa used in this kind of research can be maintained under very simple conditions and at most very short larval life, making them ideal subjects for laboratory studies of heavy metal ion toxicity. Most of the species used in this kind of research can be maintained under very simple conditions and have contributed to giving the polychaetes a reputation for being generally pollution resistant, but it is likely that this is a sampling error. These species used in research easily go into culture. The more sensitive taxa are much more difficult to keep in culture, but may be more generally characteristic of the group as a whole in terms of physiological requirements.

**See Also the Following Articles**

INVERTEBRATES, MARINE, OVERVIEW • WORMS, NEMATODA • WORMS, PLATYHELMINTHES

**Bibliography**


WORMS, NEMATODA

Scott L. Gardner
The University of Nebraska, Lincoln

I. What Is a Nematode? Diversity in Morphology
II. The Ubiquitous Nature of Nematodes
III. Diversity of Habitats and Distribution
IV. How Do Nematodes Affect the Biosphere?
V. How Many Species of Nemata?
VI. Molecular Diversity in the Nemata
VII. Relationships to Other Animal Groups
VIII. Future Knowledge of Nematodes

GLOSSARY

anhydrobiosis A state of dormancy in various invertebrates due to low humidity or desiccation.
cuticle The noncellular external layer of the body wall of various invertebrates.
gubernaculum A sclerotized trough-shaped structure of the dorsal wall of the spicular pouch, near the distal portion of the spicules; functions in guiding the spicules.
hypodermis The cellular, subcuticular layer that secretes the cuticle of annelids, nematodes, arthropods (see epidermis), and various other invertebrates.
pseudocoelom A body cavity not lined with a mesodermal epithelium.
spicule Blade-like, sclerotized male copulatory organ, usually paired, located immediately dorsal to the cloaca.
stichosome A longitudinal series of cells (stichocytes) that form the anterior esophageal glands in Trichuri.
stoma The buccal cavity, just posterior to the oval opening or mouth; usually includes the anterior end of the esophagus (pharynx).
synlophie In numerous Trichostrongyliidae, an enlarged longitudinal or oblique cuticular ridge on the body surface that serves to hold the nematodes in place on the gut wall.
vermiform Worm-shaped with tapering form both posteriorly and anteriorly.

NEMATODES are the most speciose phylum of metazoa on earth. Not only do they occur in huge numbers as parasites of all known animal groups, but also they are found in the soils, as parasites of plants, and in large numbers in the most extreme environments, from the antarctic dry valleys to the benthos of the ocean. They are extremely variable in their morphological characteristics, with each group showing morphological adaptations to the environment that they inhabit. Soil-dwelling forms are extremely small, many marine species have long and complex setae, and parasitic species manifest amazingly great reproductive potential and large body size. Nematodes are one of the major synanthropic metazoa, with some species such as pinworms having coevolved with humans and their relatives since the beginning of the lineage of the primates. While estimates of the numbers of known species hover around 20,000 actual numbers of taxonomists/systematists with expertise in this group are decreasing yearly. This is despite the fact that the Nemata are probably the last
great group of Metazoa to be well-documented and described. Estimates of the actual number of species of nematodes that remain to be described include several thousand from insects and millipedes, several thousand from vertebrates, and perhaps millions from marine habitats.

1. WHAT IS A NEMATODE? DIVERSITY IN MORPHOLOGY

A. General Characteristics and Synapomorphies

Despite numerous assertions that appear in the literature stating that nematodes are morphologically conservative, the contrary is actually true: species of the phylum Nemata are extremely variable in their morphological characteristics. Because of this diversity, almost any broad statement regarding their anatomy probably should be tempered or qualified. Nevertheless, nematodes are nonsegmented worms that generally lack external appendages. Most are vermiform, with tapering anterior (Fig. 1a) and posterior (Fig. 1b) ends, cylindrical in cross-section, and covered with a usually translucent, flexible, acellular cuticle (Fig. 1c) secreted by an underlying cellular hypodermis. The cuticle may be smooth (Fig. 2) or ornamented with rings (Figs. 3a–3c), longitudinal striations, spines (Fig. 3b), or spikes, or it may have well-developed wing-like structures called lateral alae that are very common on the externo-lateral surfaces (Fig. 4). Some marine species have long setae (genus Dracontoma) (Figs. 5a, 5b, and 6), modified sensory papillae that are probably used in movement and in detecting their environment. In contrast to all other nematodes, some marine species have eye spots that enable them to detect light in their environment (e.g., genus Thoracostoma) (Fig. 7).

The fine structure of the external body-cuticle is complex, acting to protect the animal from the external environment and allowing it to remain homeostatic inside. The cuticle can be extremely resistant, and depending on the nematode species and its life-history attributes. The cuticle of the nematode may be able to resist digestion in the most inhospitable stomachs in the vertebrate world. On the other hand, nematodes may be extremely delicate, able to exist intact only within the osmotically balanced tissues of other animals (e.g., members of the order Filaroida), and if moved from the isosmotic solution to one with fewer salts, they may explode and die in a most amazing display.

Nematodes are called “pseudocoelomates” because in most forms, their coelom is not completely lined with mesodermally derived cells and they are triploblastic. Possessing no circular body muscles, movement is accomplished by contraction and relaxation of longitudinal muscles in apposition to a hydrostatic skeleton. The body wall is flexible and very strong. All nematodes maintain their form because their body fluids (in the hydrocoel) are under a positive pressure relative to their environment, analogous to the way a water balloon maintains its shape. The cross-section in Fig. 8 shows Vexillata and the round shape of the body under the cuticle that supports spines or aretes (the whole structure is called the synlopethe). Nematodes have a complete digestive system with an anterior stoma just behind the mouth and usually a tri-radiate or triple muscle-pumping type of esophagus (Figs. 1a, and 1c) that can be muscular or glandular in structure (or both). The intestine, tubular in form (Fig. 1c), is usually a single cell in thickness lined on the peritoneal side with a thin collagen-like material, internally lined with micro villi (Grasse, 1965; Maggenti, 1981, 1991a). The tube extends from the esophagus straight to the anus or cloaca, sometimes showing out pouches cece or diverticulae near the esophageal end.

Most nematodes are sexually dimorphic with separate sexes (dichious or amphigonic) and are oviparous; however, some are ovoviviparous and some are viviparous. Males usually have a curlicularized spicule or pair of spicules that are used to assist in the transfer of sperm to the females. Many species also have a gubernaculum (seen extended from the cloaca of a species of an aspidodrilid in Fig. 23) that guides the spicule during copulation and spicular eversion. Some are hermaphroditic, in these cases the nematode produces both sperm and ova from the ovotestis of the same individual at different times during ontogeny (Maggenti, 1991a; Malakhov, 1994).

A synapomorphy of the Nematida is the presence of noncontractile axon-like myoneural processes or extensions that run from the contractile or body portion of muscle cells to the neural junctions of the nerve cords. The width of nematodes is usually less than 2 mm, even when extremely long. However, even here there are exceptions, with the giant kidney worm, Dicrocoelium dendriticum, of canids attaining a size of 15 mm long by 20 to 30 mm wide, with an exception being the antarctic marine nematode Deonostoma timmerchioi (40 mm in length of body) that has eggs that range from 870 to 1100 mm long by 240 to 350 mm wide (huge by nematode standards).

The fine structure of the external body-cuticle is complex, acting to protect the animal from the external environment and allowing it to remain homeostatic inside. The cuticle can be extremely resistant, and depending on the nematode species and its life-history attributes. The cuticle of the nematode may be able to resist digestion in the most inhospitable stomachs in the vertebrate world. On the other hand, nematodes may be extremely delicate, able to exist intact only within the osmotically balanced tissues of other animals (e.g., members of the order Filaroida), and if moved from the isosmotic solution to one with fewer salts, they may explode and die in a most amazing display.

Nematodes are called “pseudocoelomates” because in most forms, their coelom is not completely lined with mesodermally derived cells and they are triploblastic. Possessing no circular body muscles, movement is accomplished by contraction and relaxation of longitudinal muscles in apposition to a hydrostatic skeleton. The body wall is flexible and very strong. All nematodes maintain their form because their body fluids (in the hydrocoel) are under a positive pressure relative to their environment, analogous to the way a water balloon maintains its shape. The cross-section in Fig. 8 shows Vexillata and the round shape of the body under the cuticle that supports spines or aretes (the whole structure is called the synlopethe). Nematodes have a complete digestive system with an anterior stoma just behind the mouth and usually a tri-radiate or triple muscle-pumping type of esophagus (Figs. 1a, and 1c) that can be muscular or glandular in structure (or both). The intestine, tubular in form (Fig. 1c), is usually a single cell in thickness lined on the peritoneal side with a thin collagen-like material, internally lined with micro villi (Grasse, 1965; Maggenti, 1981, 1991a). The tube extends from the esophagus straight to the anus or cloaca, sometimes showing out pouches cece or diverticulae near the esophageal end.

Most nematodes are sexually dimorphic with separate sexes (dichious or amphigonic) and are oviparous; however, some are ovoviviparous and some are viviparous. Males usually have a curlicularized spicule or pair of spicules that are used to assist in the transfer of sperm to the females. Many species also have a gubernaculum (seen extended from the cloaca of a species of an aspidodrilid in Fig. 23) that guides the spicule during copulation and spicular eversion. Some are hermaphroditic, in these cases the nematode produces both sperm and ova from the ovotestis of the same individual at different times during ontogeny (Maggenti, 1991a; Malakhov, 1994).

A synapomorphy of the Nematida is the presence of noncontractile axon-like myoneural processes or extensions that run from the contractile or body portion of muscle cells to the neural junctions of the nerve cords. The width of nematodes is usually less than 2 mm, even when extremely long. However, even here there are exceptions, with the giant kidney worm, Dicrocoelium dendriticum, of canids attaining a size of 15 mm long by 20 to 30 mm wide, with an exception being the antarctic marine nematode Deonostoma timmerchioi (40 mm in length of body) that has eggs that range from 870 to 1100 mm long by 240 to 350 mm wide (huge by nematode standards).
As mentioned previously, nematodes are extremely diverse in size, shape, and structure. The following will only briefly touch on the expansive subject of nematode morphological diversity. For additional reading and exploration see the essential works and references therein of Nickle (1991), Maggenti (1981), and Grasse (1965).

B. External Covering: The Cuticle

An example of diversity in shape is that of the complex cuticular aretes found in species of the O. Strongylida: Trichostrongyloidea (parasitic in vertebrates) in which the exocuticle is modified into a series of cuticular aretes called the synlophe. In these forms, it is thought that the ridges running down the length of the body of the nematode are used in maintaining their position in the intestine of their hosts (Fig. 8). Other forms have an exocuticle composed of serrated ridges (Figs. 3a, 3b, 9, 10, 11), bumps, or may be very smooth (Fig. 2). Some groups (Heterakoidea) possess large cuticularized suckers just anterior to the cloaca that are surrounded by sensory papillae that evidently allow the male to find the female in the intestinal tract of the host as in Paraspidodera (Figs. 2, 12, 13) and other aspidoderids (Fig. 23). Marine nematodes of the genus Draconema (Figs. 5a, and 6) and Desmoscolex (Fig. 3c) have a wildly modified cuticle relative to most nematodes and very large hair-like and segmented cuticular setae that the nematodes use for both movement and detecting its environment (Figs. 5a and 6). One curious structure
that occurs in all Nemata is the amphid, a highly variable sensory organ that can be very obvious as in *Desmodera* (Fig. 10) or very inconspicuous as in *Chamberstella* (Fig. 18).

C. The Alimentary Canal

The mouth of the nematode (Fig. 14a) provides the anterior opening to the external environment that connects to the stoma leading posteriad to the muscular esophagus (Figs. 15, 16, 17, 18). The stoma may be quite reduced or absent as in the case of members of the Trichostrongyloidea (O. Strongylida) or they may be well developed such as that found in *Clarbus* (Fig. 16) and *Mononchus* (Fig. 1b) and capable of inflicting damage on their prey, or it may be modified into horn tooth-like structures that are used to attach to the intestinal villi of the host animal (e.g., *Ancylostoma*) (Fig. 19). In mammalian gut parasites of the genus *Trichuris* (Adenophorea: Trichuridae), the stoma is lacking and the esophagus is formed of stichocytes comprising a stichosome that is glandular in function, but many other plant parasites such as the Adenophorean ectoparasitic, below-ground root-feeders (capable of transmitting viruses between and among plants *Xiphinema*, *Longidorus*, or *Dorylaimus*; O. Dorylamida: Longidoridae) have a tubular stoma with the posterior parts of the stoma modified into a spear that is used to penetrate plant cell walls as in *Dorylaimus stagnalis* (Fig. 20). The Secernentean plant endoparasitic nematode of the genus *Pratylenchus* (Fig. 21a) also has a modified stomatal spear (much more delicate than those found in *Dorylaimus* sp.) with which it penetrates plant cells as the nematode moves through the tissues and cells of the plant (usually below-ground).

D. The Reproductive System

The reproductive system of most animal parasitic nematodes is adapted to produce extremely large numbers
of eggs (e.g., *Ascaris*). Some characteristics of larger animal parasitic nematodes include very large body, very two large ovaries, and an equally large uterus. Free-living nematodes generally have much smaller bodies and are therefore individually less prolific. However, the reproductive systems of most nematodes have the following basic structural similarities: Most females have two ovaries, one anterior (as in *Pratylenchus*, Fig. 21b) and one posterior, each connected to an oviduct and uterus with the uteri connecting to the vagina, terminating in the vulva. Males usually have one testes as in the Secernentea or two as in the Adenophorea, a seminal vesicle, and a vas deferens (see Maggenti, 1999a) connecting to the outside via the cloaca (Fig. 13) which is a joining of the reproductive system and the rectum. Secondary sexual organs in male nematodes are usually much more pronounced and variable than that of the female (Figs. 2, 6, 11, 12, 13, 22, 23, 24, 25) and most males have one or two spicules (Fig. 22), a spicular pouch, sometimes with a spicular sheath (Fig. 24), and a gubernaculum (Fig. 23). Some males of the O. Strongylida have a well-developed copulatory bursa (Fig. 22).

Some plant parasitic nematodes produce eggs directly into the plant where the nematode lives (in the case of species of the genus *Pratylenchus*, the female produces eggs that hatch within the tissues of the plant and the juveniles begin feeding), whereas in others such as *Heteroderda* (which are ectoparasites) the body of the female fills with eggs, forming a sac that eventually dries and transforms into a resistant cyst, with the juveniles within capable of resisting environmental extremes.

II. THE UBQUITOUS NATURE OF NEMATODES

There are few habitats on earth unoccupied by nematodes. More than a century of both biological surveys and informal collecting has led many biologists to believe that the phylum Nemata is probably the most ubiquitous of all animal groups. Early in the history of scientifically based biological investigations, pioneers of microscopy opened a new window into the previously unseen microscopic world of the soils.
FIGURE 6 Drawing of a male of *Draconema* sp. illustrating the large setae, spicules, testis, and large expanded stoma and head (after Cobb, 1914).

FIGURE 7 Cuticularized “eye spots” in the marine nematode *Thoracostoma*. Located at about the mid-part of the esophagus, this nematode can detect light in its environment.

FIGURE 8 Cross-section at midbody of *Vesicula* armandae, a species of nematode parasitic in rodents of the genus *Perognathus* in New Mexico. This photograph shows the spines (arterae) in the cuticle, called the synlophe in *Trichostrongylidae*. Inside is the hypodermis lining the body of the nematode. The reproductive tract and the intestine are not visible in this photograph.

Descriptions of the “invisible” life seen by pioneers of microscopy such as Antony van Leeuwenhoek evoked images of a wonderfully diverse and dynamic community of worms and other organisms. Subsequent investigations by other early researchers began to open up the unseen world of the nematodes.

FIGURE 8 Cross-section at midbody of *Vesicula* armandae, a species of nematode parasitic in rodents of the genus *Perognathus* in New Mexico. This photograph shows the spines (arterae) in the cuticle, called the synlophe in *Trichostrongylidae*. Inside is the hypodermis lining the body of the nematode. The reproductive tract and the intestine are not visible in this photograph.

FIGURE 9 Anterior end of a species of *Criconemoides*, an external root feeding plant parasite. The rings of the cuticle can be seen from the head end posterior. The strong cuticularized stomatal spear is clearly visible in this photograph.
A. Early Views of Nematode Diversity: The Human Perspective

Humans have been parasitized by nematodes from the earliest times. Eggs of the pin-worm Enterobius vermicularis and the whip-worm Trichuris trichiura occur in coprolites dated to about 7000 years old (y.o.) from dry areas of Peru, and eggs of the hookworm Ancylostoma duodenale have been reported from coprolites dated to around 7230 y.o. In eastern Brazil, Ascaris lumbricoides has been positively identified from human coprolites dated to about 28,000 y.o. from caves in France, but this is the only occurrence of a record this old. The seeming dearth of other nematodes from human remains older than about 7000 y.o. appears to be due to the fact that organic material comprising the coprolites themselves do not preserve well enough to last that long (Karl J. Reinhard, personal communication).

In what is thought to be the oldest surviving written account of Ascaris in humans (dated to approximately 4700 y.o.) (in China), foods to avoid and a description of the symptoms of humans infected with these worms was accurately given (Maggenti, 1981). In the area of the Nile River Valley, early Egyptian physicians recorded the presence of both Ascaris and Dracunculus (the “Guinea worm”) in an ancient papyrus manuscript (written by Egyptian physicians around 3552–3550 years ago) that was obtained and translated by the
FIGURE 14 (a) Anterior end of a species of Paraspidodera (a parasite of the cecum of rodents of the genus Ctenomys in Bolivia) showing three huge lips, the stomatal opening (mouth) in the middle of the lips, and large anteriorly directed sensory papillae on lips 2 and 3 near the outer part of the photograph. (b) Head-on view of a predaceous nematode as might be seen from the perspective of a mild-mannered bacterial feeding form such as Caenorhabditis just before it is devoured (after Cobb, 1914).

Egyptologist ‘Ebers’ in 1872 (see Chitwood and Chitwood, 1977; Maggenti, 1981). In the extant literature, the first mention of a nematode from a nonhuman animal was by Hippocrates about 2430 years ago, he described the occurrence of pinworm nematodes of horses and human females. From that time, little more was discovered until Albertus Magnus and Demetrius Pape-
FIGURE 18 Delicate but well-cuticularized stoma of Chambersiella.

FIGURE 19 Scanning electron micrograph looking into the mouth of a species of mammalian parasite of the genus Ancylostoma. The teeth are used to attach to the villi of the host mammal.

FIGURE 20 Anterior end of Dorylaimus stagnalis, a plant root ecto-parasite showing the well-developed stomatal spear that is used to penetrate plant root cells.

FIGURE 21 (a) Confocal image of Pratylenchus sp. showing the well-developed spear that the nematode uses to penetrate plant cells while moving through the roots of the plant. The muscles and glandular part of the esophagus are also visible. (b) Confocal image of Pratylenchus showing the anteriorly directed ovary with individual oocytes and their nuclei visible.

gomenos (in the 13th century, cited from Rausch, 1983) recorded nematodes from falcons (also see Chitwood and Chitwood 1977). With the development of the microscope and the emergence of Europe from the dark ages, knowledge of nematodes as parasites of plants and animals and of free-living forms expanded rapidly. It

FIGURE 22 Posterior end of a male trichostrongylid showing the well-developed copulatory bursa, bursal rays, and the long thin paired spicules.
was found that nematodes occurred everywhere people looked; in fact, Anton van Leeuwenhoek first recorded the presence of vinegar eels (Anguillula aceti) in his vinegar stored for personal use in a letter dated 21 April 1676, although he was not aware that others had reported finding nematodes some time earlier (Dobell, 1932).

Estimates of infections of people with common human parasitic nematodes give the following numbers (from Crompton, 1999): of a total human population of about 6 billion individuals (in the year 2000) the strongyloid hookworms Ancylostoma duodenale and Necator americanus infect about 1,298,000,000 (22%), and the large intestinal nematode Ascaris lumbricoides occurs in about 1,472,000,000 (25%) people at any one time in the world. Obviously, many people harbor more than one species of nematode at a time, and it is common for people to sport Ascaris, Necator, Trichuris, and Enterobius simultaneously. I provide the following estimate to indicate just how important these organisms are in the web of life on Earth. To put the number of infections of humans in perspective, I made the following extrapolations: An adult female Ascaris produces approximately 200,000 eggs per day at an average rate of about 5 grams of eggs per year. Actual data for Ascaris in humans that are infected show an average of 18 worms per infected person. Given that half of these are females, I calculate that 9 worms/person will produce about 45 grams of eggs in the feces of the host per year. In one year the total population of Ascaris in humans worldwide is conservatively estimated to produce 66,240,000 kg or 66,240 metric tons (72,864 tons [English]) of eggs; this is equal in weight to about 348 large adult blue whales, 8,832 adult male elephants, or 364 fully loaded railroad coal cars.

Estimates of the number of human infections in the year 2000 by other species of parasitic nematodes are shown in Table I. At the present time, it is estimated that approximately 138 species of nematodes have been reported from humans (Crompton, 1999) with from 32 to 36 being host-specific.

B. The Science of Nematode Diversity

Nathan A. Cobb, often considered the father of nematology in North America, was a student of the renowned German zoologist Ernst Haeckel. After just a few years...
WORMS, NEMATODA

TABLE I

<table>
<thead>
<tr>
<th>Species of nematode</th>
<th>Numbers infected</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ancylostoma duodenale</strong> and <strong>Necator americanus</strong></td>
<td>1,298,000,000</td>
<td>Worldwide</td>
</tr>
<tr>
<td><strong>Ascaris lumbricoides</strong></td>
<td>1,472,000,000</td>
<td>Worldwide</td>
</tr>
<tr>
<td><strong>Brugia malayi</strong> and <strong>B. timori</strong></td>
<td>13,000,000</td>
<td>South Pacific, SE Asia, India</td>
</tr>
<tr>
<td><strong>Dracunculus medinensis</strong></td>
<td>80,000</td>
<td>Sub-Saharan Africa and Yemen</td>
</tr>
<tr>
<td><strong>Loa loa</strong></td>
<td>13,000,000</td>
<td>West and Central Sub-Saharan Africa and Yemen</td>
</tr>
<tr>
<td><strong>Onchocerca volvulus</strong></td>
<td>17,660,000</td>
<td>Central and South America and Sub-Saharan Africa</td>
</tr>
<tr>
<td><strong>Strongyloides stercoralis</strong></td>
<td>70,000,000</td>
<td>Temperate regions</td>
</tr>
<tr>
<td><strong>Trichuris trichiura</strong></td>
<td>1,049,000,000</td>
<td>Worldwide</td>
</tr>
<tr>
<td><strong>Enterobius vermicularis</strong></td>
<td>400,000,000</td>
<td>Temperate regions</td>
</tr>
</tbody>
</table>

* Data from Crompton (1999).

As if challenged by this assertion, scientists have tested Cobb’s hypotheses by examining the extremes of the biosphere on Earth to evaluate the limits of nematode life. Through these investigations, biologists have now shown that nematodes are living and reproducing everywhere on Earth that water exists in a liquid state even for short periods of time annually.

III. DIVERSITY OF HABITATS AND DISTRIBUTION

A. General Distribution

The most obvious ecological characteristic that defines habitats for members of the phylum Nemata is that they all are aquatic animals—to move, live, eat, and reproduce, nematodes must exist in an aqueous environment. This environment includes soils, muds, sands, plants, and animals. They can be found living in soils with moisture contents as low as 5 to 10%, but in the majority of these cases, the nemas are associated with the roots of plants. It is evident that the environment in which nematodes live constrains their ultimate size. Soil-dwelling species live in the water film of the interstices of soil particles. Nematodes must live and carry out all functions of life in these spaces, thus free-living or plant-parasitic soil-dwelling forms are usually extremely small. Morphological diversification of nematodes can operate only within the constraints of their life history parameters, and evolutionary pathways for nematodes living in an interstitial–soil or sediment–sand environment are limited. Thus these forms have limited abilities to diversify into the nonaquatic regions of research spanning the globe from Europe, Australia, and North America, Cobb amassed a huge amount of knowledge and came to have a deep appreciation for the immense number of species that existed. With scientific knowledge based on keen observational skills, he understood the nature of both the great numerical density and species diversity of nematodes in all habitats of the globe that he examined. Thus armed, he wrote the following:

In short, if all the matter in the universe except the nematodes were swept away, our world would still be dimly recognizable, and if, as disembodied spirits, we could then investigate it, we should find its mountains, hills, vales, rivers, lakes, and oceans represented by a film of nematodes. The location of towns would be decipherable, since for every massing of human beings there would be a corresponding massing of certain nematodes. Trees would still stand in ghostly rows representing our streets and highways. The location of the various plants and animals would still be decipherable, and, had we sufficient knowledge, in many cases even their species could be determined by an examination of their erstwhile nematode parasites.

We must therefore conceive of nematodes and their eggs as almost omnipresent, as being carried by the wind and by flying birds and running animals, as floating from place to place in nearly all the waters of the earth, and as shipped from point to point throughout the civilized world in vehicles of traffic.

COBB (1914)
of the Earth, and their size is constrained by this fact. A lingering question remains, why do we not see extremely large marine nematodes? Perhaps it is because we have not yet looked carefully in the marine environment.

In contrast to the small sizes of plant-parasitic, free-living marine, freshwater, or soil nematodes (some adults can be as small as Criconemoides with an adult length of around 250 μm), species that occur as parasites of animals are free from many of the physical constraints on their size, since the bodies of species in mammals are, therefore, relatively large. In fact, the largest nematode thus far recorded is Placentonema gigantissima from the blue whale with a length of more than 8 m (Maggenti, 1981). The ratio between the smallest adult nematode to the largest can be calculated as \(0.239 \, \text{mm} / \text{Criconemoides} / 0.000031\), compared to the ratio between a shrew and a blue whale (30 mm / 24,400 mm = 0.12), indicating that the size differences in nematodes are three orders of magnitude greater than the mammals.

In the marine-benthic environment, Lambshead (1993) has estimated (based on transect data from deep sea benthic samples) that there may be as many as 100 million species of marine infaunal nematodes. As exorbitant as that estimate seems, deep-sea nematodes have been shown to be extremely rich in species diversity. For example, one marine sediment sample from the east Pacific benthos was reported to contain 148 species from a total of only 216 individual specimens examined (Lambshead, 1993). At the present time the true extent of species diversity in the marine benthos can only be imagined, because fewer than 20 studies of nematode community structure from marine benthic habitats have been reported (Boucher and Lambshead, 1995). Data from Boucher and Lambshead (1995) shows that in marine environments the highest diversity in nematodes occurs in abyssal benthic sediments.

We know that great numbers of individuals and species of nematodes live in sediments on the ocean floor distributed from the intertidal continental margins to the benthos of the abyssal zones, even though these marine environments remain mostly unexplored. Nematodes occur in tissues and organs of all species of vertebrates that have been studied, and some, such as Physaloptera spp. live, feed, and reproduce in the strongest stomach acids of mammals, birds, and reptiles. Desiccated specimens from both the Arctic and Antarctic have been rehydrated to form viable colonies, and living nematodes have been found in the limited meltwater in some of the dry valleys of Antarctica, an area that is probably one of the most extreme biotopes on the earth.

B. Extreme Biotopes

Some of the most extreme soil habitats on the Earth exist in the dry valleys of the Antarctic, where the annual mean air temperature is \(-20^\circ\text{C}\), and soil temperatures at a 5-cm depth for the two “summer months” range from \(-2.7\) to 15.9°C. In this area, no vascular plants grow and mosses and lichens are rare; this is the only terrestrial soil system known where nematodes are the final consumers and are at the apex of the food chain (in this case it seems more of a chain than a web). Three nematode species exist in these dry soils: Scottonema lindsayae, a microbivore (feeding on bacteria and yeast), Plectus antarcticus (a bacterial feeder), and Endorylaimus antarcticus, an omnivorous predator that presumably feeds on individuals of the other two species (Powers et al., 1998). Another biological extreme, Death Valley in California, has recorded some of the highest temperatures in North America, and the soils of the valley are teeming with nematodes, many of which have been discovered to possess similar adaptive traits to those found in the cold deserts of Antarctica.

Individuals of some species of Nemata are capable of resisting extended periods of desiccation, for example, 3rd stage juveniles of Anguina tritici have been dried for more than 20 years in a state of anhydrobiosis in which all metabolic activities are shut down (Maggenti, 1981). These nematodes have been shown to have specialized proteins that fold into stable/preserved structures as the organism dries. In this state of crypto- or anhydrobiosis, individual nematodes can remain viable through incredible extremes of temperature, desiccation, hypoxia, and even synthetic nematicides designed to kill nematodes (of course the regular biochemical processes that nematicides interfere with are not operational, so the nematode does not notice this particular assault). When water again becomes available, the animal comes back to life when the molecular structures rehydrate and the proteins and enzymes spring back into normal operation.

C. Habitat Diversity

Nematodes occupy every conceivable life history niche. There are benthic deepwater marine forms that appear to consume mostly diatoms and others such as species of the genus Draconemoides (Figs. 5a and 6) that are mostly associated with marine algae, but it is still unclear what they actually eat. Nematodes of the genus Dirofilaria live in the aorta and left ventricles of canids and are transmitted by mosquitoes from dog to dog. Species of nematodes live in the hearts of sharks, and other species...
can occur by the thousands in the stomachs of pilot whales (S.I.G., personal observation). As the human consumption of raw marine fish (ceviche and sashimi) has increased, transfer of juvenile Anasakis, Terranova, and other anasakines from the fish intermediate host to humans is occurring more commonly, and these nematodes are turning up as parasites in the stomachs of humans.

Some species feed on fungi in the soil while others are trapped and are themselves consumed by different species of fungi. Still other nematodes such as Mononchus (Fig. 1b) and Clarkus (Fig. 16) are predatory and hunt and eat other nematodes in the soil environment. Free-living bacterial feeding nematodes have been shown to be integral parts in the carbon and nitrogen cycles in healthy soils (Ferris et al., 1998), and recent survey work has shown that undisturbed or natural soils in noncultivated habitats can have as many as 20 times more species than soil from similar areas that have been under cultivation (Al Banna and Gardner, 1996).

D. Abundance: Estimates and Facts

In addition to the large numbers of species that may occur in any given habitat, nematodes also occur in very great densities. For instance, in shear numerical density of individuals in any given environment, nematodes exceed even the mites and beetles combined. More than 90,000 nematodes were recorded from a single decomposing apple, and one report showed that 1 cc of marine mud contained 45 nematodes representing 19 species. Nematodes in marine estuaries occur at high numerical densities with reports of 4,420,000/m2 in surface mud and 527,000,000/acre in the top 3 inches of sand on the Massachusetts coast. Counts and extrapolations for relatively moist soils (10 to 70% moisture content) worldwide show that in the uppermost levels, nematodes occur in mind-boggling abundance: 7 to 9 billion/acre in undisturbed soil in North China; from 800,000,000 to 2.3 billion/acre (representing just 35 species) in Utah and Idaho, and around 3 billion/acre in low-lying alluvial soils of Europe and other areas of North America.

IV. HOW DO NEMATODES AFFECT THE BIOSPHERE?

Because of the huge numbers of nematodes that have been shown to occur in plants, animals, soils, and the benthos, there has been much speculation about the role of nematodes in basic biological processes occurring in the soils of the Earth. Cobb (1914) speculated that some nematodes “are beneficial,” however, he also noted that this area of study was still in its infancy. In fact, in the year 2000, this area of study is still just developing, and recent work has shown that nematodes can be good indicators of biodiversity (Bongers and Ferris, 1999). Gardner and Campbell (1989) showed that mammalian parasites with complex life cycles may serve as excellent indicators of areas of high biological diversity. Because of the multifarious nature of parasitic nematodes in mammals, it is expected that this kind of species may provide biologists with additional tools for identification of areas of high biological diversity.

The fact that parasitic nematodes occur in such high prevalences and numerical densities in mammals should give us pause. There is obviously a huge energy drain on any population of mammal that we should care to analyze, and this energy drain probably causes significant decreases in the number of offspring in any given population over long periods of time.

A. Soils and Plants

Studies have indicated that nematodes play a substantial role in the cycling of carbon and nitrogen in the soil environment (Bongers and Ferris, 1999), and it has been shown that the number of bacterial-feeding nematodes increases as the bacteria increase with annual warming of the soils. In the rainforest of Cameroon, average nematode abundances of $2.04 \times 10^7/m^2$ of rainforest soil were found, indicating that these nematodes play a significant role in carbon flux (CO2 and CH4) in forest soil (Ferris et al., 1998), and recent survey work has shown that undisturbed or natural soils in noncultivated habitats can have as many as 20 times more species than soil from similar areas that have been under cultivation (Al Banna and Gardner, 1996).

The fact that parasitic nematodes occur in such high prevalences and numerical densities in mammals should give us pause. There is obviously a huge energy drain on any population of mammal that we should care to analyze, and this energy drain probably causes significant decreases in the number of offspring in any given population over long periods of time.

IV. HOW DO NEMATODES AFFECT THE BIOSPHERE?

Because of the huge numbers of nematodes that have been shown to occur in plants, animals, soils, and the benthos, there has been much speculation about the role of nematodes in basic biological processes occurring in the soils of the Earth. Cobb (1914) speculated that some nematodes “are beneficial,” however, he also noted that this area of study was still in its infancy. In fact, in the year 2000, this area of study is still just developing, and recent work has shown that nematodes can be good indicators of biodiversity (Bongers and Ferris, 1999). Gardner and Campbell (1989) showed that mammalian parasites with complex life cycles may serve as excellent indicators of areas of high biological diversity. Because of the multifarious nature of parasitic nematodes in mammals, it is expected that this kind of species may provide biologists with additional tools for identification of areas of high biological diversity.

The fact that parasitic nematodes occur in such high prevalences and numerical densities in mammals should give us pause. There is obviously a huge energy drain on any population of mammal that we should care to analyze, and this energy drain probably causes significant decreases in the number of offspring in any given population over long periods of time.
environment as are normal seeds (Maggenti, 1981). Nematodes of the genus Pratylenchus use their spear to move through the roots of the plants that they infect, penetrating the plant cells with repeated jabs of their stomatal armature.

B. Predators, Entomopathogenic Forms, Fungal and Bacterial Feeders

Many nematodes that are found in the soil are either predaceous forms eating other nematodes or forms that prey on mites or other soil macroorganisms. The more spectacular predators such as Mononchus and Clarkus have specialized buccal structures with which to puncture the cuticles of other nematodes (Figs. 1b and 16). Microbivorous fungal and bacterial feeding nematodes are also extremely abundant, with species specializing in their feeding habits on bacteria, fungi, diatoms, and other microscopic organisms. Some, such as juveniles of species of the genera Heterorhabditis and Steinernema, carry bacteria of the genus Photorhabdus in their digestive system. The 3rd stage juveniles wait in the soil until an unwary insect passes nearby. The nematode then homes in on and penetrates the hapless insect, making its way into the hemocoel where it releases the bacteria, which proliferate, killing the host. The nematode then feeds on the bacterial colony and reproduces in the insect, eventually again producing 3rd stage juveniles that leave the carcass of the insect and disperse into the soil, waiting there for another insect to invade (for more specific details of the life-cycle of these entomopathogenic nematodes, see Gaugler and Kaya, 1990).

Bacterial-feeding forms occur in the soil in extremely high numbers. In soils that have not been disturbed and that have a good layer of organic matter, large numbers of all kinds of nematodes occur. One of the most well-known groups is the Rhabditida, or the rhabditid nematodes. These forms feed on bacteria and yeasts growing in the soil and are typically found in high numbers in moist soils with high organic content. The most famous of these forms are members of the genus Caenorhabditis, of which the complete genome of C. elegans has been sequenced.

C. Aquatic and Marine Nematodes

This is an area that is wide open for future biologists. How do the trillions upon trillions of individuals and the millions of species that occur in the oceans really affect the biosphere? Nothing is known on the subject at the present time.

V. HOW MANY SPECIES OF NEMATA?

A. Estimates of Number Described

Estimates of the numbers of species of nematodes that are known (i.e., described species) vary widely. However, in 1819, Rudolphi summarized what was known of the nematodes, recording 11 genera and about 350 species. Just 115 years later, in 1934, Filipjev reported that 4003 species of nematodes had been described, with about half free-living and the other half parasitic. By 1950, Libbie Hyman estimated that approximately 9000 species were described (based on her analysis of the zoological record with descriptions being recorded at a rate of about 200 new species described per year). In 1981, Maggenti’s summary showed around 15,000 described species. My analysis from counting additions to the zoological record shows that in the 5-year period from 1992 through 1996, numbers of descriptions were relatively stable, with approximately 776 new species described (based on her analysis of the zoological record with descriptions being recorded at a rate of about 200 new species described per year). From 1996 through 1998 the numbers of descriptions decreased to 118 per year, most likely due to the continued retirement and expiration of knowledgeable taxonomists.

B. How Many Species of Vertebrate Parasitic Nematodes Exist?

All species of vertebrates examined thus far serve as hosts for at least one species of parasitic nematode. Some mammals host many species of nematodes that are distributed through several orders and families. Some of these nematodes are highly host-specific, surviving and reproducing successfully only in host individuals comprising a single species or perhaps a closely related group of species. Other nematodes show less specificity, being much more likely to jump from one suitable vertebrate host to another during opportune times during their life history (Brant and Gardner, 2000).

Within a host, many different types of habitats may be occupied by nematodes. Some species, Homo sapiens harbors approximately 33 species of host-specific parasitic nematodes (Chitwood and Chitwood, 1977). To illustrate the diversity of habitats in a single animal host, humans will be used as an example. In a human, nematodes can occur as juveniles in muscle tissues (usually smooth muscle such as the diaphragm or the tongue [Trichinella]) and in the mucosa of the intestine (Strongyloides); as migrating forms in blood and lungs (Ascaris); as microfilariae in blood or lymph (filarioids...
WORMS, NEMATODA

857

C. Comparative Nematode Diversity of New World Subterranean Rodents: Geomyidae

Papers on nematode parasites from rodents of the Nearctic family Geomyidae covering the dates from 1857 to 1999 were reviewed. Combined with field-collected specimens from the early 1970s up to the present time, I discovered that six of the approximately 11 nematode parasites reported from pocket gophers in North America are host specific to only the Geomyidae (Table II).

Of members of the vertebrate class Mammalia, one of the most complete sets of nematode parasite data exists for rodents of the family Geomyidae. Of the approximately 35 known species of pocket gophers (Wilson and Reeder, 1993), only 15 species have been surveyed for parasitic nematodes. From those 15 species, six species of nematodes are known to be host-specific only to geomyids. Some nematodes such as the stron-
TABLE II

<table>
<thead>
<tr>
<th>Nematode species</th>
<th>Classification and location in host</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trichuris fossor</td>
<td>Cecum and large intestine</td>
</tr>
<tr>
<td>Ransomus rodentorum</td>
<td>Cecum/Small intestine</td>
</tr>
<tr>
<td>Vexillata vexillata</td>
<td>Duodenum</td>
</tr>
<tr>
<td>Vexillata convoluta</td>
<td>Duodenum</td>
</tr>
<tr>
<td>Heligmosomoides thomomyos</td>
<td>Duodenum</td>
</tr>
<tr>
<td>Litomosoides thomomydis</td>
<td>Peritoneal cavity and mesenteries</td>
</tr>
<tr>
<td>Litomosoides westi</td>
<td>Mesenteries</td>
</tr>
</tbody>
</table>

D. Nematodes of Tuco Tucos (Ctenomyidae)

A review of the nematode parasites occurring in Neotropical rodents of the genus Ctenomys (Table III) indicates a considerably more depleted fauna of nematodes as compared with the nearctic Geomyidae. Data collected from 1984 on indicate that nematodes of the genera Trichuris and Paraspidodera have cospeciated with their hosts and exhibit different levels of phylogenetic congruence relative to their hosts. In addition, nematodes of the trichostrongyloid (O. Strongylida) genus Pudica were encountered only two times from the same species of Ctenomys in one locality (from a sample of more than 500 individuals and more than six species of hosts examined). The occurrence of A. caninum in Ctenomys appears to be a capture, as it only occurred in areas where dogs, humans, and ctenomyids lived in relatively close proximity (banana fields in lowland Santa Cruz, Bolivia).

Whereas most of the pocket gophers examined carefully generally harbor from one to several species of nematodes in the small intestine, very few nematodes are found in samples of tuco tucos. Even though the genus Ctenomys contains almost 40 species, comparatively few species of nematodes from them have been described or reported. This lack of parasites in a wide-open group of mammals might be a result of rapid speciation in the mammal group, with parasites failing to keep up with the speciation rate of the mammals themselves and actually losing parasites through time; there is some evidence that the ctenomyids have speculated rapidly in the recent past. The lack of a diverse fauna of nematodes in these mammals could also be due to historical accident, whereby the ancestor of the ctenomyids had a low diversity of nematode parasites (for whatever reason) thus giving rise to a phylogenetic lineage of mammals lacking a diverse fauna of parasites. But why have they not picked up more parasitic nematodes from other sympatric species of mammals?

There is some evidence of host-switching in nematodes as compared with the nearctic Geomyidae. Data collected from 1984 on indicate that nematodes of the genera Trichuris and Paraspidodera have cospeciated with their hosts and exhibit different levels of phylogenetic congruence relative to their hosts. In addition, nematodes of the trichostrongyloid (O. Strongylida) genus Pudica were encountered only two times from the same species of Ctenomys in one locality (from a sample of more than 500 individuals and more than six species of hosts examined). The occurrence of A. caninum in Ctenomys appears to be a capture, as it only occurred in areas where dogs, humans, and ctenomyids lived in relatively close proximity (banana fields in lowland Santa Cruz, Bolivia).

Whereas most of the pocket gophers examined carefully generally harbor from one to several species of nematodes in the small intestine, very few nematodes are found in samples of tuco tucos. Even though the genus Ctenomys contains almost 40 species, comparatively few species of nematodes from them have been described or reported. This lack of parasites in a wide-open group of mammals might be a result of rapid speciation in the mammal group, with parasites failing to keep up with the speciation rate of the mammals themselves and actually losing parasites through time; there is some evidence that the ctenomyids have speculated rapidly in the recent past. The lack of a diverse fauna of nematodes in these mammals could also be due to historical accident, whereby the ancestor of the ctenomyids had a low diversity of nematode parasites (for whatever reason) thus giving rise to a phylogenetic lineage of mammals lacking a diverse fauna of parasites. But why have they not picked up more parasitic nematodes from other sympatric species of mammals?

There is some evidence of host-switching in nematodes as compared with the nearctic Geomyidae. Data collected from 1984 on indicate that nematodes of the genera Trichuris and Paraspidodera have cospeciated with their hosts and exhibit different levels of phylogenetic congruence relative to their hosts. In addition, nematodes of the trichostrongyloid (O. Strongylida) genus Pudica were encountered only two times from the same species of Ctenomys in one locality (from a sample of more than 500 individuals and more than six species of hosts examined). The occurrence of A. caninum in Ctenomys appears to be a capture, as it only occurred in areas where dogs, humans, and ctenomyids lived in relatively close proximity (banana fields in lowland Santa Cruz, Bolivia).

Whereas most of the pocket gophers examined carefully generally harbor from one to several species of nematodes in the small intestine, very few nematodes are found in samples of tuco tucos. Even though the genus Ctenomys contains almost 40 species, comparatively few species of nematodes from them have been described or reported. This lack of parasites in a wide-open group of mammals might be a result of rapid speciation in the mammal group, with parasites failing to keep up with the speciation rate of the mammals themselves and actually losing parasites through time; there is some evidence that the ctenomyids have speculated rapidly in the recent past. The lack of a diverse fauna of nematodes in these mammals could also be due to historical accident, whereby the ancestor of the ctenomyids had a low diversity of nematode parasites (for whatever reason) thus giving rise to a phylogenetic lineage of mammals lacking a diverse fauna of parasites. But why have they not picked up more parasitic nematodes from other sympatric species of mammals?

There is some evidence of host-switching in nematodes as compared with the nearctic Geomyidae. Data collected from 1984 on indicate that nematodes of the genera Trichuris and Paraspidodera have cospeciated with their hosts and exhibit different levels of phylogenetic congruence relative to their hosts. In addition, nematodes of the trichostrongyloid (O. Strongylida) genus Pudica were encountered only two times from the same species of Ctenomys in one locality (from a sample of more than 500 individuals and more than six species of hosts examined). The occurrence of A. caninum in Ctenomys appears to be a capture, as it only occurred in areas where dogs, humans, and ctenomyids lived in relatively close proximity (banana fields in lowland Santa Cruz, Bolivia).

Whereas most of the pocket gophers examined carefully generally harbor from one to several species of nematodes in the small intestine, very few nematodes are found in samples of tuco tucos. Even though the genus Ctenomys contains almost 40 species, comparatively few species of nematodes from them have been described or reported. This lack of parasites in a wide-open group of mammals might be a result of rapid speciation in the mammal group, with parasites failing to keep up with the speciation rate of the mammals themselves and actually losing parasites through time; there is some evidence that the ctenomyids have speculated rapidly in the recent past. The lack of a diverse fauna of nematodes in these mammals could also be due to historical accident, whereby the ancestor of the ctenomyids had a low diversity of nematode parasites (for whatever reason) thus giving rise to a phylogenetic lineage of mammals lacking a diverse fauna of parasites. But why have they not picked up more parasitic nematodes from other sympatric species of mammals?
WORMS, NEMATODA

859

todes of the genus *Litomosoides*, in that two species occur in *Ctenomys opimus* in high altitude western Bolivia, but these nematodes have not been reported from any other species of *Ctenomys* from throughout the Neotropics. Superficially this indicates a host-capture event from some other lineage of mammals (Brant and Gardner, 2000). Another example is the fact that nematodes of the genus *Pudica* found in *Ctenomys* have diverse relatives in other species of muroid rodents in South America but none in ctenomyids, leading to the conclusion that most are now found in the tucos because of host-switching events and not phylogenetic coevolution.

Nematodes of the genus *Paraspidodera* are found only in Hystricognath rodents in the Neotropical region and these nematodes appear to have had a long historical, coevolutionary association with ctenomyids (Gardner, 1991), showing varying levels of both coexistence and host switching. The adenophorean whip worm genus *Trichuris* occurs in many diverse groups of rodents in the neotropics, but no analyses have yet been done to examine the levels of coevolution with ctenomyids.

The multilatirous nature of nematode diversity in subterranean mammals in the Nearctic and Neotropical regions requires at the minimum that phylogenetic hypotheses for each group of mammals and their nematodes be developed so each host group can be compared with each parasite group. Much more detailed work must be paid to collecting parasites from some of the unknown species of *Ctenomys* throughout the Neotropics. The same can be said about the level of knowledge of parasitic nematodes in the Geomyidae in the northern Neotropics and southern Nearctic regions.

VI. MOLECULAR DIVERSITY IN THE NEMATA

Several molecules have been used to begin to assess phylogenetic and genetic diversity within the Nemata, and the number of investigations using molecular methods to try to quantify the diversity of the nematodes is rapidly increasing (see references in Dorris et al., 1999; Blaxter et al., 1998; Adams, 1998; Al-Banna et al., 1997; Nielsen, 1996). However, because of the extremely large number of species that may exist, examination of levels of molecular or genetic diversity in representatives of the group as a whole is just beginning even though massive amounts of molecular data on nematodes are now pouring into the literature stream. The summary papers by Blaxter et al. (1998) and Dorris et al. (1999) indicate the utility and power of estimating the molecular-phylogenetic relationships among the Nemata using ribosomal DNA and other molecular sequence data. Initial studies of the molecular diversity within and among several lineages of the Strongylida show that the genetic diversity among these taxa is relatively low (Chilton et al., 1997). From these works, it is clear that molecular phylogenies will provide robust tests of the hypotheses of morphological relationships among the Nemata. As more regions of DNA are used to examine the relationships among the nematodes, we expect a clarification of both the deep phylogenetic branches that are relatively obscure in the molecular phylogeny of the Nemata and the more rapidly evolving branch tips that represent extant species with valuable genetic information.

VII. RELATIONSHIPS TO OTHER ANIMAL GROUPS

An analysis grouped the nematodes, gastrotrichs, priapulids, kinorhynchs, and the foraminifera into a group (superphylum) called the Cycloneuralia (Fig. 26) based on the circular shape of the brains in these groups (Nielsen et al., 1996). The aforementioned study and at least one other (Zrzavy et al., 1998) using 18s rDNA sequences showed that the Nemata share a common ancestor with members of the phylum Nematomorpha.
but there was a shuffling of the other groups out of the “Cyclonerealia” (Fig. 27).

As mentioned previously, nematodes are soft-bodied and mostly very small organisms. Any larger forms that existed were probably parasites of vertebrates; however, these left no fossil traces. The only fossil nematodes that are known are insect parasitic or plant parasitic forms that occur very rarely in amber inclusions. Because there are no fossil records of nematodes of Cambrian or Precambrian ages, estimates of the age of the Nemata have been only speculative, and without fossils, it is difficult to calibrate molecular clocks for the nematodes. However, through application of various models of molecular evolution and molecular clock theory, estimates of the time of divergence of the nematodes from the rest of the animal groups appears to be about 1177 ± 79 million years (Wang et al., 1999) (this study showed a basal origin of the nematodes on a phylogenetic tree, in contrast to the relatively derived placement in the analyses shown in Figs. 26 and 27).

A. Groups of the Nemata

Above the level of the order, confusion reigns relative to the classification and systematic arrangement of the nematodes. Maggenti (1991a) is usually followed in this regard, and his analyses followed corroborated historical analyses in recognizing the two main subphyla: The Secernentea and the Adenophorea. Recent work shows that these groups are substantiated both in morphological and in molecular analyses, although competing phylogenetic hypotheses and associated classifications have also been proposed (Dorris et al., 1999; Blaxter et al., 1998; Brooks and McLennan, 1993; Adamson, 1989).

B. Classification

Since 1949, at least eight authors have provided classifications for members of the phylum Nematota (see references in Malakhov, 1994; Brooks and McLennan, 1993; Maggenti, 1991). Of these, the classifications of Maggenti (1981, 1991a) have proven to be the most useful summary of all nematodes (free-living and parasitic); however, phylogenetic hypotheses have been proposed based on both molecular and morphological characteristics. This does not necessarily mean that a classification will be developed from a proposed phylogenetic tree (see Maggenti, 1991a; Brooks and McLennan, 1993).

VIII. FUTURE KNOWLEDGE OF NEMATODES

I hope that this summary treatment provides readers with sufficient knowledge to allow more in-depth research on nematodes. The group is so large and so ecologically, morphologically, and phylogenetically diverse that to attempt to discuss the diversity of the group in such an abbreviated way is practically futile at best. N. A. Cobb (1914) stated this clearly on the last page of his famous “Nematodes and Their Relationships” (p. 490):

The foregoing fragmentary sketch may indicate to the student, as well as to the general reader, the vast number of nematodes that exist, the enormous variety of their forms, and the intricate and important relationships they bear to mankind and the rest of creation.

As more data on distribution and function of nematodes throughout the biosphere are obtained, the importance of this group of worms will surely be realized. We are just beginning to explore the oceans, and we...
are probably losing more species of nematodes from rainforest clearing than will ever be ultimately found, described, and classified.

I hope that mankind will generate more interest in the microscopic world of the Nematata, and I hope that this article will provoke the reader into action and dispel what Van Leeuwenhoek stated about his fellow man on 28 September 1715: “And over and above all, most men are not curious to know: nay, some even make no bones about saying, What does it matter whether we know this or not?” (Dobell, 1932, pp. 324–325).

Acknowledgments

Special thanks to Tom Powers for encouragement during this writing blitz and to the staff and students of the Manter Laboratory. Also thanks to Sue Ann, Teal, and Hudson for their help with reference materials and support.

See Also the Following Articles

PARASTOMI•WORMS, ANNELIDA•WORMS, PLATYHELMINTHES

Bibliography


soil invertebrates along an elevational gradient in Taylor Valley, Antarctica. Arctic Alpine Res. 30, 133–141.


WORMS,
PLATYHELMINTHES

Janine N. Caira* and D. Timothy J. Littlewood†
*University of Connecticut and †The Natural History Museum London

I. General Features
II. Major Groups
III. Nervous System
IV. Feeding and Digestion
V. Excretion and Osmoregulation
VI. Reproduction
VII. Ontogeny
VIII. Phylogenetic Relationships
IX. Host Associations
X. Medical Importance

GLOSSARY

acetabulum Ventral sucker used for attachment and/or locomotion in digeneans.
acoelomate Lacking a body cavity; see parenchyma.
cercomer Posterior parenchymous and/or muscular extension of body; homology among groups is controversial; some view both the hook-bearing regions of monogeneans and some larval cestodes to be modifications of the cercomer.
cirrus Male intromittent copulatory organ consisting of distal, invaginable portion of ejaculatory duct that is evaginated (by turning inside out) through the genital pore during copulation; may or may not be armed with spines, microtriches, etc.; see penis.
cortex The outermost region of parenchyma in taxa in which two distinct regions are present; inner boundary often marked by conspicuous circular muscle fibers; see medulla.
duogland adhesive system Combination of a cement, a viscid gland producing adhesive material, and a releasing gland; usually associated with cells of ventral epidermis; provides temporary adhesion to the substrate for many interstitial flatworm species.
haptor The posterior attachment organ of monogeneans.
medulla The innermost region of parenchyma in taxa in which two distinct regions are present; outer boundary often marked by conspicuous circular muscle fibers; see cortex.
monozoic Possessing only a single set of reproductive organs; see polyzoic.
microthrix A specialized surface extension of the neodermis of cestodes, characterized by an electron-dense cap composed of numerous microtubules separated from distal cytoplasm by a base-plate (plural microtriches).
neodermis A syncytial, nonciliated epidermis with cell bodies and nuclei (often referred to as cytons or pericaryon) laying below other body wall elements; unique to the Neodermata; completely replaces epidermis, usually at metamorphosis when larva encounters the first host.
parenchyma The connective tissue between cells of ectoderm and endoderm layers that fills spaces among organ systems; as platyhelminths are acoelomates, the space among internal organs is completely filled with parenchyma.
penis Male intromittent copulatory organ consisting of a muscular, often papilliform, retractable structure that is protruded (without turning inside out) through the genital pore during copulation; may or may not be armed with spines, etc.; see cirrus.

polyzoic Possessing multiple serial sets of reproductive organs; controversial as to whether proglottids are homologous with segments of coelomates.

proglottid Compartmentalized regions of cestode body; each proglottid contains one or more sets of reproductive organs; distinctive, potentially monophyletic groups: Digenea, Monogenea, and Cestoda. A number of these are of significant medical or veterinary importance. There exists little evidence to suggest that the remaining platyhelminths, some of which are free-living and some of which exist as either commensals or parasites of vertebrates, form a single cohesive group, although collectively they are often referred to as “turbellarians.” Present knowledge suggests that the parasitic groups are more diverse than the nonparasitic groups. However, there is no question that numerous free-living and parasitic taxa await discovery and formal description.

I. GENERAL FEATURES

A glance at the defining features of each of the major platyhelminth groups reveals that the characteristics that vary most among, and even within, these groups are primarily those associated with the digestive and reproductive systems. The digestive system generally lacks an anus. Thus, there is no unidirectional flow of food materials through the system; in many species food enters and solid wastes exit through the oral opening. Several major features of the female reproductive system are of particular importance. The female gonad of platyhelminths is either homocellular, in that it consists of only a single type of cell, or it is heterocellular, consisting of cells that produce either ova or vitelline cells. The former configuration results in the production of endolecithal eggs because vitelline (yolk) material is incorporated directly into the cytoplasm of the ova. Heterocellular platyhelminths usually conduct these functions in separate organs, possessing an ovary (or occasionally several ovaries) that produces ova and a vitellarium that produces vitelline cells containing the nutritive yolk. This configuration results in the production of ectolecithal eggs, as vitelline cells are packaged along with, but independent of, the zygote in the egg.
In addition, most platyhelminths possess a formalized excretory system consisting of generally two (rarely one or three) sets of pronephridia. The nervous system and sense organs are present to a greater or lesser extent among different taxa. Specialized organs functioning in respiration and circulation are entirely lacking in many groups of platyhelminths. Thus, respiration generally occurs by means of diffusion of gases directly through the surfaces of the body, a process that is greatly facilitated and made effective by the relatively flattened nature of the platyhelminth body. The “circulation” of materials throughout the various parts of the platyhelminth body is also accomplished by diffusion. However, a number of digeneans are known to possess a “lymphatic system” with contractile vessels and cellular inclusions that may have a circulatory function. The details of most organ systems are difficult to observe without the use of specialized staining and microscopic techniques.

One final, rather unusual feature of the platyhelminths is that the differentiated nonreproductive (somatic) cells of the body are unable to divide. Thus, mitosis does not occur in the differentiated somatic cells of platyhelminths; rather, these cells are replaced as necessary by undifferentiated stem cells located below the outer layers of the body, which undergo mitosis. This phenomenon helps to explain the remarkable regenerative abilities of certain platyhelminth groups.

II. MAJOR GROUPS

Diversity in the group is generally most conveniently represented at the level of the taxonomic rank of order. Within the platyhelminths this rank is relatively informative and much more numerically manageable than lower ranks such as family or genus (for example, there are 42 orders versus 401 families and 4241 genera). However, it is important to recognize that traditionally, taxonomists working on each of the three major groups exhibiting obligate parasitism (Digenea, Monogenea, and Cestoda) have generally worked independently of one another, and also independently from the taxonomists studying the platyhelminth groups exhibiting more free-living life history strategies. Thus, the concepts of higher ranks, such as order, are not necessarily comparable among the major groups. For example, the 243 genera of monogeneans are currently distributed among 12 orders, whereas the 2549 genera of digeneans are distributed among only three orders. Diagnoses are presented later for the orders currently recognized in all groups except the digeneans, which are treated at the level of superfamily. Numbers of families and genera for each of the orders (or superfamilies in the case of the digeneans) are given in Table 1.

For more detailed information on specific groups, readers are referred to the works of Cannon (1986) and Rieger (1998) for the non-neodermatan groups; Yamaguti (1971) and Schell (1985) for the digeneans; Yamaguti (1963), Schell (1985), and Boeger and Krisky (1993) for the monogeneans; and Schmidt (1986) and Khalil et al. (1994) for the cestodes. However, the treatments of these groups by Hyman (1951) and the five sections covering the platyhelminths in the Traite de Zoologie (Grasse, 1961) are classic and remain among the more comprehensive sources of detailed information on the platyhelminths. The many works of Klaus Rohde (for example, 1997), Ulrich Ehlers (for example, 1985), and Peter Ax (for example, 1996), among others, have done much to further our understanding of the morphology and phylogenetic relationships among these groups.

Discussions of diversity are most effectively organized around hypotheses of the phylogenetic relationships among the groups involved. Although investigators have not yet come to a consensus about the phylogenetic relationships among all of the major platyhelminth groups, the relationships among some of these groups are fairly well accepted. The tree in Fig. 1 illustrates some of the aspects of these relationships for which supporting morphological and molecular data are starting to accumulate. The presentation of diversity that follows is generally organized around the topology of this tree; the non-neodermatan orders are treated in the sequence in which they appear from left to right on this tree. A more detailed treatment of the phylogenetic relationships among the various groups is provided in Section VIII.

An illustration is included of a representative of each platyhelminth order (or superfamily for the digeneans). A consistent stippling pattern has been used throughout these figures for each different organ system so that the conditions of the major organs are readily visible and comparable among figures. The organs represented by each stippling pattern are labeled in Figs. 12 and 18.

1. Acoela

See Fig. 2. Glandular frontal organ present; lamellated rhabdites absent, some with unlamellated rhabdoids; duogland adhesive system absent; epidermis ciliated throughout; statocyst present, with single statolith and two parietal cells; some with paired ocelli; mouth usually ventral, occasionally subterminal, opens directly into parenchyma or into simple or occasionally strongly muscular pharynx; gut lumen absent; digestion in tem-
<table>
<thead>
<tr>
<th>Text box no.</th>
<th>Figure no.</th>
<th>No. of families</th>
<th>No. of genera</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>13</td>
<td>87</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>37</td>
<td>146</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>7</td>
<td>68</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>10</td>
<td>101</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>11</td>
<td>26</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>16</td>
<td>124</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>8</td>
<td>104</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>8</td>
<td>65</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>17</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>16</td>
<td>18</td>
<td>3</td>
<td>179</td>
</tr>
<tr>
<td>17</td>
<td>19</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>18</td>
<td>20</td>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>19</td>
<td>21</td>
<td>10</td>
<td>158</td>
</tr>
<tr>
<td>20</td>
<td>22</td>
<td>21</td>
<td>213</td>
</tr>
<tr>
<td>21</td>
<td>23</td>
<td>16</td>
<td>358</td>
</tr>
<tr>
<td>22</td>
<td>24</td>
<td>12</td>
<td>176</td>
</tr>
<tr>
<td>23</td>
<td>25</td>
<td>10</td>
<td>162</td>
</tr>
<tr>
<td>24</td>
<td>26</td>
<td>3</td>
<td>85</td>
</tr>
<tr>
<td>25</td>
<td>27</td>
<td>7</td>
<td>76</td>
</tr>
<tr>
<td>26</td>
<td>28</td>
<td>8</td>
<td>236</td>
</tr>
<tr>
<td>27</td>
<td>29</td>
<td>15</td>
<td>162</td>
</tr>
<tr>
<td>28</td>
<td>30</td>
<td>29</td>
<td>462</td>
</tr>
<tr>
<td>29</td>
<td>31</td>
<td>3</td>
<td>84</td>
</tr>
<tr>
<td>30</td>
<td>32</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>31</td>
<td>33</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>32</td>
<td>34</td>
<td>6</td>
<td>110</td>
</tr>
<tr>
<td>33</td>
<td>36</td>
<td>8</td>
<td>74</td>
</tr>
<tr>
<td>34</td>
<td>37</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>35</td>
<td>38</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>36</td>
<td>39</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>40</td>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>38</td>
<td>41</td>
<td>5</td>
<td>24</td>
</tr>
<tr>
<td>39</td>
<td>42</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>40</td>
<td>43</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>41</td>
<td>44</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>42</td>
<td>45</td>
<td>27</td>
<td>89</td>
</tr>
</tbody>
</table>

### Neodermata

<table>
<thead>
<tr>
<th>Text box no.</th>
<th>Figure no.</th>
<th>No. of families</th>
<th>No. of genera</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>47</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>44</td>
<td>48</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>49</td>
<td>4</td>
<td>41</td>
</tr>
<tr>
<td>46</td>
<td>50</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>47</td>
<td>51</td>
<td>6</td>
<td>56</td>
</tr>
<tr>
<td>48</td>
<td>52</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>49</td>
<td>53</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>54</td>
<td>19</td>
<td>45</td>
</tr>
<tr>
<td>51</td>
<td>55</td>
<td>8</td>
<td>53</td>
</tr>
<tr>
<td>52</td>
<td>56</td>
<td>2</td>
<td>58</td>
</tr>
<tr>
<td>53</td>
<td>57</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>54</td>
<td>58</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>55</td>
<td>59</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>56</td>
<td>60</td>
<td>15</td>
<td>364</td>
</tr>
</tbody>
</table>

### Cestoda

<table>
<thead>
<tr>
<th>Text box no.</th>
<th>Figure no.</th>
<th>No. of families</th>
<th>No. of genera</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>649</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Habitat:** Predominantly marine; predominantly free-living, but some symbiotic in digestive system or body cavity of echinoderms (primarily holothuroids and echinoids); a number of species of Convolutidae and Sagittiferidae symbiotic with green algae.

2. Nemertodermatida

See Fig. 3. Glandular frontal organ present; lamellated rhabdites absent, some with unlamellated rhabdoids; duogland adhesive system generally absent; epidermis ciliated throughout; statocyst usually with two (occasionally one, three, or four) statoliths and several parietal cells; usually with, occasionally without, ventral mouth; usually without, occasionally with, simple pharynx; but generally with some form of lumen, with inter-
FIGURE 1  Hypothesized phylogenetic relationships among major platyhelminth groups. State “0” should be assumed prior to point at which character is marked on tree; state “1” should be assumed if character is marked on tree but no state is given. Topology is modified from Littlewood et al. (1999).
tinal epithelium and glandular club cells; protonephridial system absent; gonad hermaphroditic; sperm filiform, unflagellated, flagellum not incorporated into body of sperm, 9 + 2 microtubule arrangement; male antrum shaped as simple ciliated invagination of epidermis, eversible penis or hard parts absent; male pore posterior; vitellarium absent; oviducts absent, female genital pore usually absent, occasionally dorsal female pore present, eggs possibly released into lumen of gut; eggs endolecithal.

**Habitat:** Entirely marine; predominantly free-living, some symbiotic in body cavity or intestine of holothuroidean echinoderms.

3. Catenulida

See Fig. 4. Some with tail; lamellated rhabdites absent, unlamellated rhabdoids present; duogland adhesive system absent; epidermal ciliation sparse; some with ventrolateral ciliated furrow; brain lobed or not; statocyst often present, usually with one but occasionally with up to six statoliths and several parietal cells; mouth ventral, opening into simple pharynx; intestine in form of simple elongate sac; protonephridial system unpaired; single hermaphroditic gonad present; sperm aflagellated; penis present or absent, male and female genital pores dorsal; vitellarium absent; eggs endolecithal; some undergo asexual multiplication via paratomy or parthenogenesis.

**Habitat:** All species are free-living, primarily in freshwater, but some in marine environments.

4. Haplopharyngida

See Fig. 5. With protrusible anterior proboscis not connected to digestive system; lamellated rhabdites present; duogland adhesive system present; brain enclosed in connective tissue capsule (encapsulated), bilobed; statocyst absent; mouth opens into simple muscular pharynx; intestine in form of simple elongate sac; protonephridial system paired; single testis present; sperm aflagellated; penis present, with circle of hard, straight stylets; single ovary; vitellarium absent; eggs endolecithal; male pore anterior to female pore.

**Habitat:** All species are free-living in marine environments.

5. Lecithoeotheliata

See Fig. 6. Some with weak posterior adhesive disc; some with anterior pigmented girdle; lamellated rhabdites present; duogland adhesive system present; statocyst absent; mouth anterior; pharynx usually muscular, anterior; intestine in form of simple elongate sac; protonephridial system paired; testes compact or follicular, one or many; sperm biflagellated, flagella not incorporated into body of sperm, 9 + “1” microtubule arrangement; seminal vesicle present or absent; penis present, with or without stylet; muscular copulatory bulb present or absent; accessory male organ present or absent, if present opening into male atrium, prostaticvesicle present or absent; male pore opening to outside or into pharyngeal cavity; female organ single or tetrapartite, in form of heterocellular germovitelarium (vitelline-producing cells surrounding ovum-producing cells); seminal receptacle present or absent; genitointestinal canal sometimes present; female pore ventral; eggs ectolecithal (yolk not produced in oocytes).

**Habitat:** All species are free-living in marine, or terrestrial environments.

6. Macrostomida

See Fig. 7. Frontal glands sometimes present; some with spatulate posterior region bearing adhesive papillae; lamellated rhabdites present; duogland adhesive system present; brain not encapsulated; statocyst absent; simple pharynx present, sometimes tubular; intestine in form of simple elongate sac, sometimes with pre-oral blind sacs; with protonephridial system paired; testes compact, single or paired; sperm aflagellated; seminal vesicle present or absent; prostatic vesicle present or absent; penis present or absent; penis stylet present or absent; ovary single or paired; vitellarium absent; eggs endolecithal; genital pores separate or combined; ability to asexually produce zoids present or absent.

**Habitat:** All species are free-living in marine, euryhaline, or freshwater environments.

---

**FIGURES 2–15** 2 Acoela: adult of Otocelis sp. (modified from Cannon, 1986); 3 Nemertodermatida: adult of Nemertodermatula sp. (modified from Cannon, 1986); 4 Catenulida: adult of Retronecrista sp. (modified from Cannon, 1986); 5 Haplopharyngida: adult of Haplopharynx sp. (modified from Cannon, 1986); 6 Lecithoeotheliata: adult of Gnosenesima sp. (modified from Cannon, 1986); 7 Macrostomida: adult of Bradylocera sp. 8 Polycladida: adult of Cestiplana sp. (redrawn from Beauchamp in Grasse, 1961); 9 Proseriata: adult of Bothrioplana sp. (modified from Beauchamp in Grasse, 1961); 10 Tricladida: adult of Dopoeia sp. (modified from Cannon, 1986); 11 Prolecithoplana: adult of Protelecomerella sp. (from several literature sources); 12 Kalypocythara: adult of Cystiplana sp. (modified from Cannon, 1986); 13 Typhloplanida: adult of Promesostoma sp. (redrawn from Ehlers, 1974); 14 Dalayliida: adult of Syndesmida sp. (modified from Cannon, 1986); 15 Temnocephalida: adult of Temnocepha sp. (modified from Baer in Grasse, 1961). (See abbreviations on pages 897–898.)
7. Polycladdida
See Fig. 8. Body oval or elongate, or thick and broad with folded margins; with (Cotylea) or without (Acotylea) pseudosuckers posterior to female genital pore; anterior or nuchal (neck) tentacles present or absent; ventral prostomatoid-like organs present or absent; dorsal surface smooth or papillate; body margins of some with hard spines; may be multicolored with complex patterns; lamellated rhabdites present; duogland adhesive system present; statocyst absent; numerous ocelli usually present, marginal or scattered anteriorly, some arranged in one or two pairs of clusters (cerebral and tentacular); pharynx variable in position, ruffled and vertically oriented or sometimes consisting of anteriorly directed tube, rarely consisting of multiple tubes arranged ventrally; intestine usually with numerous radiating branches, branches sometimes anastomosing protonephridial system paired; testes numerous, scattered between branches of gut; sperm biflagellated, flagella not incorporated into body of sperm, 9 + “1” microtubule arrangement; true seminal vesicle present or absent; intromittent organ absent or in form of a penis or cirrus; penis stylet present or absent; cirrus armed or not; prostatic vesicle present or absent; bursa copulatrix present or absent; some with multiple copulatory complexes; ovaries numerous, scattered between branches of gut; seminal bursa (or Lang’s vesicle) opening into vagina present or absent; uterus present or absent (in form of expanded oviducts in some); vitellarium absent; eggs endolecithal; male and female gonopores usually separate (male anterior to female) but sometimes combined; some with hypodermic impregnation.

**Habitat:** Predominantly marine, a few freshwater or euryhaline; mostly free-living, some symbiotic with other invertebrates, for example, wrapped around the abdomen of hermit crabs, associated with gorgonian corals in mantle cavities of gastropod or lamellibranch molluscs, in genital bursa of ophiuroid or echinoid echinoderms; some live in association with their food, for example, in association with tunicate colonies feeding on zooids, or in association with oyster colonies feeding on oyster tissue; some seek shelter in empty shells.

8. Proseriata
See Fig. 9. Epithelial nuclei insunk or not; body of some with feeble or prominent tactile bristles; body uniformly ciliated (sometimes with exception of caudal tip) or cilia restricted to ventral surfaces and head; head with or without ciliated pits; lamellated rhabdites absent; duogland adhesive system present; brain encapsulated or not; statocyst usually present; pigmented ocelli present or absent; pharynx tube horizontal or vertical; intestine linear or tripartite, dorsal to pharynx; protonephridial system paired or tripled, occasionally single; testes paired or not, usually compact, rarely follicular; sperm biflagellated, flagella not incorporated into body of sperm, 9 + “1” microtubule arrangement; usually with single male copulatory organ but occasionally with many; occasionally with accessory (prostatoid) organ; bursa present or absent; with one pair of compact ovaries, ovary subdivided into smaller follicles consisting of oocytes surrounded by accessory cells in some; vitellarium follicular, follicles in paired lateral rows; vagina externa present or absent; genitointestinal canal usually present; eggs ectolecithal; genital pores common or separate.

**Habitat:** Predominantly free-living in marine environments; a few occur in freshwater; occasionally ectocommensal on marine crustaceans.

9. Tricladida
See Fig. 10. Usually large, flattened, elongate; body sometimes pigmented, terrestrial forms often colorful; some with triangular, semi-lunate, or pointed anterior region of body; caudal adhesive disc present or absent; lamellated rhabdites present in some; duogland adhesive system present; cilia uniform throughout all surfaces of body or restricted to ventral surface; statocyst absent; with many ocelli or one pair of ocelli (cave-dwelling species lack ocelli); ocelli sometimes with distinct lens; mouth in middle or posterior half of body; pharynx plicate or tubular, posteriorly directed, usually one, occasionally more pharynges; intestine tripartite with one anterior and two posterior branches, posterior branches may or may not fuse, diverticulae sometimes present; anterior branch sometimes anterior to brain; protonephridial system paired; testes, two or many; sperm biflagellated, flagella not incorporated into body of sperm, 9 + “1” microtubule arrangement; seminal vesicle present or absent; male copulatory structures complex; penis usually pyriform, usually unarmed, but sometimes armed with spines or stylet; prostate organ usually absent; one pair of small ovaries usually at anterior of body; female copulatory structures variable, with or without copulatory bursa; accessory genital organs may be present in form of extra bursae each with separate pore; vitellarium follicular, extensive; genitointestinal canal sometimes present; eggs ectolecithal; gonopore usually single, rarely two or three; temporary embryonic pharynx and intestine replaced by definitive pharynx during development to adult stage.
Habitat: Predominantly free-living in marine, euryhaline, freshwater, or moist terrestrial environments (suborders associated with habitats); one family known only from freshwater caves; some symbiotic with other invertebrates, for example, in mantle cavity of gastropods, or gill lamellae of horseshoe crabs, or on dorsal surface of skates.

10. Prolecithophora
See Fig. 11. With or without two ciliated grooves around anterior body; anterior adhesive disc usually absent (present in Hypotrichinidae); extensive epidermal spicules present or absent; cilia uniform throughout all surfaces of body or restricted to ventral surface; brain encapsulated or not; statocyst lacking; one to three pairs of ocelli present or absent; mouth anterior, midventral or posterior; with simple, plicate or weakly bulbous pharynx; intestine in form of simple elongate sac, well defined (surrounded by tunic of connective tissue) in most; paired prothonephridial system present; testes, one or two, compact or diffuse; sperm biflagellated, flagella not incorporated into body of sperm, mitochondria with elaborate outer cell membrane foldings; seminal vesicle in some; male copulatory stylet present or absent; ovary and vitellarium not fully separated in some, but separated in others; ovary compact or diffuse, paired or not; vitelliniferous diffuse; accessory female openings present or not; eggs ectolecithal; male pore opening anterior to female opening but into common atrium; combined mouth and gonopore (Combinata) or gonopore opening separate from mouth (Separata).

Habitat: Predominantly marine, some inhabiting freshwater; free-living, or symbiotic (commensal or parasitic) on gills, on body surfaces, in mantle cavity of gastropod and lamellibranch molluscs, or on surface of small crustaceans; rarely parasitic on anal and branchial regions of fish; some of these associations may be phoretic, with the worm using the host merely for transport.

11. Kalyptorhynchia
See Fig. 12. Body usually ciliated throughout; rarely with epidermal spicules of crystalline calcium carbonate; rarely with anterior girdle of highly vacuolated epidermal cells; anterior sheathed proboscis present or absent, divided or not, with or without hooks, gland cells, or specialized apex; lamellated rhabdites present in some; duogland adhesive system present; statocyst absent; ocelli present or absent; pharynx variable in position (anterior or posterior), bulbous, usually oriented ventrally, with or without sphincter or hard knobs, not associated with proboscis; gut in form of simple elongate sac; prothonephridial system paired; testes single or paired, or in form of median row of follicles; sperm biflagellated, flagella not incorporated into body of sperm, 9 + 1 microtubule arrangement; internal seminal vesicle in copulatory bulb or not; copulatory organ with stylet or not; or three separately opening prostatic organs with their own spiny cirri, or one cirrus, or copulatory organ with no penial or cirrus structures; prostate vesicle incorporated in testis or not; prostate with spiny structures or not; ovary and vitellarium separated or not (ovo-vitellarian), paired or single; uterus generally present, with or without accessory piece; seminal receptacle present or absent, paired or not; bursa present or absent; eggs ectolecithal; usually 1 but sometimes 2 gonopores.

Habitat: All species free-living, primarily in marine, a few in euryhaline, several in freshwater environments.

12. Typhloplanida
See Fig. 13. Anterior usually modified with prominent rhabdoid tracts or rarely with an impermanent sheathed proboscis, or with a permanent terminal invagination with proboscis glands; dermal and adenal rhabdites present; duogland adhesive system present; cilia uniform throughout all surfaces of body or restricted to ventral surface; brain undivided, encapsulated; statocyst absent; paired ocelli present or absent; tubular spinous, buccal region occasionally present; mouth usually midventral or posterior; pharynx usually ventrally oriented, in form of simple tube with annular fold or bulbous; intestine in form of simple elongate sac; paired prothonephridial system present; excretory pore opens into mouth or genital atrium; testes single or paired; sperm biflagellated, flagella not incorporated into body of sperm, 9 + 1 microtubule arrangement; single or paired seminal vesicles sometimes present; male system often with funnel-shaped stylet; sheathed, unarmed cirrus present; penis papilla sometimes present; ovary and vitellarium usually separate; ovary single, paired, or unpaired; vitellarium paired or a column of serially arranged follicles; usually with bursa copulatrix and seminal receptacle; vagina absent; eggs ectolecithal; male and female pores separate or combined, male anterior to female; common oro-genital pore present or absent.

Habitat: Species known from freshwater, terrestrial, and marine environments; mostly free-living; a few known from the body surfaces of polychaete annelids.

13. Dalyelliida
See Fig. 14. Lamellated rhabdites present in some; duogland adhesive system absent; cilia completely covering
body, or just ventral; statocyst usually lacking; paired ocelli present or absent; mouth, pharynx, and intestine present or occasionally absent; pharynx bulbous, oriented anteriorly if present; intestine in form of simple elongate sac if present; paired protonephridial system present or occasionally absent; testes paired, compact or lobed; sperm biflagellated, flagella not incorporated into body of sperm, 9 + “1” microtubule arrangement; with or without male copulatory organ; if present male organ with or without stylet; ovary and vitellarium separate or occasionally combined; ovary single or paired, lobed; vitellarium single or paired, branched or unbranched; seminal receptacle present or absent; uterus present; copulatory bursa present or absent; single genital pore present or absent; at least one taxon is dioecious and sexually dimorphic (Kronborgia).

**Habitat:** Some species free-living in marine, freshwater, or terrestrial environments; many species commensals or parasitic: found, for example, in the connective tissue of tube feet of asteroid echinoderms, in the hemocoel of crabs, shrimp, amphipods, and isopods, in the mesenchyme of myzostomid annelids, in kidneys and gonoducts or mantle cavity and gut of gastropod molluscs, in stomach of lamellibranch molluscs, in mesenchyme of the turbellarian Plagiostomum, in either the coelom or the digestive tract of holothuriod, echinoid, or crinoid echinoderms; some also in sipunculans.

### 14. Temnocephalida

See Fig. 15. Usually with two to 12 anterior tentacles, tentacles usually adhesive; with posterior adhesive region, usually circular but occasionally crescentic or divided into two or more smaller regions, often pedunculate; some with lamellated rhabdites, if present restricted to anterior region of body; duod gland adhesive system absent; locomotory cilia usually absent; outer body layer divided into a series of syncytial epidermal plates; statocyst absent; paired ocelli usually present; mouth terminal or ventral; pharynx tubular or bulbous, usually oriented anteriorly, sometimes reduced; intestine in form of simple sac or with weak lateral diverticulae; paired protonephridial system present, usually with two lateral excretory vesicles; with one to 10 pairs of testes; sperm biflagellated, flagella not incorporated into body of sperm, 9 + “1” microtubule arrangement; male organ usually armed, usually with eversible cirrus mounted on tubular or conical shaft; ovary single; vitellarium lateral or scattered in form of large follicles over intestine; eggs ectolecithal.

**Habitat:** Entirely freshwater; ectocommensal, primarily on external body surfaces, gills, or lining of branchial chambers of decapod crustaceans, occasionally found on amphipods, molluscs, turtles, and hemipterans.

### A. Neodermata

Ehlers championed the recognition of the remaining platyhelminth groups as the neodermatans because they share an outer body covering in the form of a neodermis, which is sometimes also referred to as a tegument (Fig. 16). Neodermatans generally begin life with a single-layered, cellular, ciliated epidermis, like most of the nonneoderman platyhelminth groups, but as they mature they shed this outer epidermal layer and replace it with the neodermis. This transformation usually occurs when the first larval stage encounters its host. The cytoplasm of the neodermal cells is continuous throughout the surface of the animal, as there are no cell membranes separating adjacent cells; this layer is a syncytium. The nuclei of these cells are generally located below the basal lamina and outer muscle layers of the body, in structures called cytons or pericarya. These are, in fact, the nuclei of the stem cells, the cytoplasm of which grew outward to form the syncytial layer of the neodermis. The cytons are connected to the outer syncytial layer by thin cytoplasmic extensions.

The neodermatans share several other rather unique features. In these platyhelminths, the filters of the flame bulbs of the protonephridial system are in the form of a two-celled weir. Each filter is constructed from longitudinal cytoplasmic rods of the terminal cell that alternate with longitudinal cytoplasmic rods of the first canal cell. In addition, rather than possessing two ciliary...
rootlets, one directed anteriorly (the rostral rootlet) and one directed posteriorly (the caudal rootlet), the cilia of neodermatans appear to have lost the caudal rootlet, and possess only the anterior rootlet. Finally, the flagella of the sperm in neodermatans, rather than remaining separate from the cytoplasm of the main sperm body, are incorporated into the main sperm body, arranged in a proximal—distal direction along its lateral margins.

There are four major neodermatan groups: the Aspidogastrea and the Digenea, collectively referred to as the Trematoda, and the Monogenea and the Cestoda, collectively referred to as the Cercomeromorpha. The orders or superfamilies of each of these 4 groups are treated separately below.

15. Aspidogastrea

See Fig. 17. With ventral holdfast organ in form of linear series of 20-100 suckers, numerous transverse rugae, or extensive adhesive disc subdivided by septa into numerous loculi; disc with or without papillae; neodermis with microvilli in form of hemispherical microrugae; buccal funnel at anterior end of body; oral sucker usually absent; oral lobes occasionally present; with one or two blind intestinal ceca; excretory vesicle present, V-shaped, excretory pore near posterior of body; with one, two, or many testes; sperm biflagellated, flagella incorporated into body of sperm, 9 + "1" microtubule arrangement; cirrus present; cirrus-sac present or absent; ovary single; vitellaria follicular or tubular, if follicular usually with multiple lateral follicles, rarely single median tube flanked by follicles, if tubular, two tubes; uterus generally long; Laurer's canal present or absent; Mehlis' gland present; eggs ectolecithal; single common genital pore.

**Habitat:** Adults endoparasites in kidneys, pericardial chamber, gill lamellae, or intestine of freshwater and occasionally marine lamellibranch molluscs, kidneys (renal cavities) and pericardial chamber of marine and freshwater gastropod molluscs, occasionally guts of turtles or freshwater and marine teleosts, gall bladder and bile ducts of elasmobranchs, or in lumen of rectal glands of Holocephali.

**B. Class Digenea**

Almost certainly monophyletic, the digeneans are most easily characterized by their possession of a unique series of larval stages including miracidia, sporocyst, redia, cercaria, and metacercaria. Variation in life cycles seen among digenean species suggests that perhaps only a subset of these ontogenetic stages should be considered to distinguish the class. Most digeneans possess an anterior sucker (or oral sucker), which is usually (Fig. 19), but not always (Fig. 24), associated with the mouth. Most also possess a second sucker (or acetabulum) that is usually midventral (Fig. 30), but occasionally posterior in position (Fig. 29). The body of some species is divided into two parts (Fig. 21). Other major variations in adult body form include a posterior telescoping ecsoma (Fig. 23), a sucker-like ventral adhesive tribocytic organ (Fig. 21), one or more anterior collars of large spines (Fig. 22), and paired anterior retractile proboscides. Most species possess an anterior mouth that opens into a bulbous pharynx and an intestine consisting of two or occasionally one blind ceca, with or without diverticulae. However, the ceca in some species are connected to one another posteriorly (Fig. 20) or open into the excretory bladder (Fig. 21). The excretory system generally consists of two sets of protonephridia, both of which empty into a common excretory vesicle. This vesicle is variable in shape among taxa and usually opens at the posterior end of the body. With the exception of the schistosomes (Fig. 31) and some didymozoids, all digeneans are hermaphroditic. The male intromittent organ is usually a cirrus, which may or may not be armed with spines, and may or may not be surrounded by a cirrus-sac. Testes are usually oval, but occasionally lobed or tubular; testes can be single, paired, or numerous. Sperm may be stored inside or outside the cirrus-sac in an internal or external seminal vesicle respectively; some taxa possess both vesicles. Most digeneans possess a single ovary that may be round, lobed, or tubular in form. Digeneans lack a vagina; instead, sperm enters the female system through the uterus, which opens, generally in combination with the male system, into a common genital pore. Some species possess a seminal receptacle in the form of a basal expansion of the uterus or as diverticulae of the Laurer's canal. A Laurer's canal, which may or may not open to the outside, may be present or absent. The function of the Laurer's canal is unknown, but it is considered by some to represent the remnant of a vagina. The vitellarium consists of columns of numerous follicles along both sides of the body in most taxa, but is occasionally present as one or a pair of compact follicles or vitelline chords. Almost all digeneans are endoparasites as adults in various organ systems of vertebrates.

With over 2500 nominal genera, the digeneans are by far the most diverse of the platyhelminth groups. Unlike all of the other major platyhelminth groups, family rather than order is the highest taxonomic category of general discussion and relative stability within the class. Unfortunately, the relationships among the
approximately 150 families remain poorly understood, and as a consequence, there is currently no generally accepted higher level classification scheme for the digeneans (for example, see Brooks et al., 1985; Pearson, 1992). As a consequence, this group is treated here following the superfamilies recently recognized by David Gibson and Rod Bray of The Natural History Museum in London. No tree of the relationships among these groups is presented; the superfamilies are treated below in alphabetical order. This scheme was selected primarily to demonstrate digenese diversity; the monophyly of many of these groups remains uncertain. It is important to note that the concepts of the superfamilies presented here are based on a combination of life history, adult morphology, and host association characters; thus as memberships in the groups change, so will the generalized characteristics and host generalizations for the groups. The fact that details of the life histories of many species remain unknown presents obvious problems with a classification scheme based to some extent on larval features. Not to be ignored is the diversity of hosts and sites within the hosts that these animals parasitize.

To make the following diagnoses more concise, unless mentioned it should be assumed that the superfamily possesses no specialized body regions or structures, both suckers are present, the mouth opens into the oral sucker, a muscular pharynx is present; the gut consists of two blind intestinal ceca, the organisms are hermaphroditic, the testes are paired, the sperm are biflagellated and the flagella are incorporated into the body of the sperm, the flagella of the sperm possess a 9 + “1” arrangement of microtubules, a cirrus and cirrus sac are present, the ovary is single, the vitellarium is follicular and arranged in two lateral fields, there is a common genital pore, the eggs are ectolecithal and the epidermal cells of the miracidium are covered with cilia.

16. Allocreadioidea

See Fig. 18. Oral sucker with or without one or three pairs of muscular papillae; body surfaces usually in-spined; cirrus sac usually present; genital pore opens near acetabulum. Larvae: Cotylocercous xiphidiocercaria, ophthalmoxiphidiocercaria, trichocercous or homolometrine cercaria; metacercaria usually encysts in invertebrate hosts, rarely on substrate.

**Habitat:** Adults parasitic in digestive tract and occasionally gall bladder of marine and freshwater teleosts, or occasionally gall bladder of freshwater turtles, or intestine of chameleons; some species progenetic in insects.

17. Clinostomoidea

See Fig. 19. Oral sucker weak, enveloped in collar-like anterior end of body, pharynx rudimentary or small; genital pore opens in posterior third of body. Larvae: Brevifurcate cercaria; metacercaria encysts on vertebrate hosts.

**Habitat:** Adults parasitic in mouth and esophagus of alligators and piscivorous birds.

18. Cyclocoeloida

See Fig. 20. Oral sucker usually absent, occasionally present; vertral sucker absent or, if present, weakly developed; intestinal ceca usually fused posteriorly to form cyclocoel; genital pore opens near pharynx. Larvae: Cercaria generally tailed or with rudimentary tail; metacercaria encysts within sporocyst or redia in mol luscan host.

**Habitat:** Adults parasitic in nasal sinuses, nasolacrimal sinus, intraorbital sinuses, respiratory tract, and occasionally coelom or intestine of birds.

19. Diplostomoidea

See Fig. 21. Body generally divided into anterior forebody and posterior hindbody; tribocytic organ present posterior to acetabulum; reproductive organs generally concentrated in hindbody; genital pore usually opens in posterior of body. Larvae: Cercaria variable; metacercaria generally encysts on vertebrate hosts, very rarely invertebrate hosts.
Habitat: Adults parasitic in digestive system of birds and reptiles, especially crocodilians and snakes, sometimes amphibians, sometimes mammals including dogs, cats, dolphins, and some monotremes.

20. Echinostomatoidea
See Fig. 22. Spiny head collar often present; two anterior, retractile, spiny proboscides rarely present; body usually spined; ceca sometimes with lateral diverticula; hermaphroditic sac sometimes present, often with cirrus sac, cirrus occasionally absent; genital pore variable in position. Larvae: Cercaria echinostome-like, megulaorous, gymncephalous megaperid or haplospanchnid; metacercaria usually encysts on aquatic vegetation, sometimes on vertebrates, rarely in invertebrates.

Habitat: Adults parasitic in the intestinal tract, gall bladder, bile duct, liver, lungs, nasal cavity, conjunctival sac, bursa fabrici, and orbit of vertebrates including a variety of marine and freshwater teleosts, birds, crocodiles, turtles, snakes, lizards, marine mammals including pinnipeds and cetaceans, and a variety of terrestrial herbivorous mammals.

21. Hemiuroidea
See Fig. 23. Body generally elongate; sometimes with posterior telescoping ecsoma; body surface aspinose but often with transverse plications; ceca sometimes fused to one another posteriorly, or fused posteriorly with excretory vesicle; usually hermaphroditic (monocious) but one group (didymozoids) occasionally dioecious (if dioecious adult sexually dimorphic, male and female usually encysted in pairs); hermaphroditic duct almost always present; cirrus sac usually absent; vitellarium often in form of rosette or tubular, or in two compact masses. Larvae: Miracidium spinous, lacking cilia; cercaria cystophorous; metacercaria usually unencysted in invertebrate or vertebrate hosts, occasionally within redia in molluscan host.

Habitat: Adults parasitic in digestive tract or swim bladder of freshwater and marine teleosts and rarely elasmobranchs; some associated with branchial cavities of marine teleosts and elasmobranchs; occasionally in mouth of frogs or stomach of sea snakes.

22. Gymnophalloidea
See Fig. 24. Mouth opening into oral sucker or ventral sucker (bucephalids); cirrus sac present or absent; cirrus sometimes absent; genital pore variable in position. Larvae: Cercaria often produced in branched sporocysts; cercaria often with rudimentary tail or tailless, often gasterostomous; metacercaria encysts on vertebrate hosts or within sporocyst in molluscan host; very rarely encysts in invertebrates.

Habitat: Adults parasitic in liver, bile duct, gall bladder, digestive tract of freshwater and marine teleosts, or large intestine, cloaca, bursa fabricius of birds, occasionally mammals (such as, for example, hedgehogs, rabbits, ruminants).

23. Lepocreadioidea
See Fig. 25. Spined or not; with or without circumoral spines; oral sucker occasionally with papilliform extensions; acetabulum ventral or rarely at posterior end of body; ceca open into excretory vesicle or not; one or usually two intestinal ceca present; two to many testes; cirrus sac usually present, sometimes absent; hermaphroditic duct in some; external seminal vesicle in some; internal seminal vesicle bipartite in some. Larvae: Gymnophalous cercaria, produced in redia; metacercaria encysts on invertebrates, or on substrate, rarely in sporocyst or redia in molluscan host.

Habitat: Adults parasitic almost exclusively in digestive tract, rarely gall bladder or urinary bladder of marine teleosts; one record from lungs of reptile.

24. Microphalloidea
See Fig. 26. Body surfaces usually spined; oral sucker may be surrounded by semicircular collar of spines; gut lacking or poorly developed (with short ceca) in some; vitellarium usually in restricted symmetrical clusters of large follicles; genital atrium often with papilla. Larvae: Xiphidiocercaria of ubiquita, microphallous, or virgulake type; metacercaria encysts on invertebrates or vertebrates, or occasionally in sporocyst in the molluscan host.

Habitat: Adults parasitic in intestine or ceca of birds most commonly, but also in digestive system of bats, snakes, varanid lizards, marine and freshwater teleost fishes, and mammals; occasionally bile ducts of mammals.

25. Notocotyloidea
See Fig. 27. Ventral sucker generally lacking (monostomes); often with adhesive papillae, ridges, or spines throughout ventral surface of body; head collar present or absent; male and female pores occasionally separate, pore(s) variable in position. Larvae: Cercaria monostome, bi- or trioculate, with posterior adhesive glands at posterior of body; metacercaria encysts on substrate or on invertebrate hosts, rarely in sporocyst or redia within molluscan host.

Habitat: Adults parasitic in digestive system of marine turtles, marine iguanids, birds, mammals (includ-
26. Opisthorchioida
See Fig. 28. Body sometimes spined; oral sucker sometimes with ring of spines; cirrus sac usually lacking; hermaphroditic duct sometimes present opening into genital atrium; ventral sucker sometimes enclosed in ventrogenital sac, with or without gonotyl; seminal vesicle when present often bipartite. Larvae: Cercaria generally with ocelli, pleurolophocercous, parapleurolophocercous or gymnocephalus cercaria; metacercaria encysts on vertebrates, rarely invertebrates.

**Habitat:** Adults parasitic in stomach, intestine, occasionally bile ducts of marine teleost fishes, snakes, lizards, birds, and mammals; occasionally ovary of freshwater teleosts; occasionally progenetic encysted in musculature of frogs.

27. Paramphistomoidea
See Fig. 29. Usually unspined; oral sucker often with paired lateral diverticula, occasionally one unpaired diverticulum; acetabulum posterior or subterminal, sometimes with muscular lip, occasionally lacking; body occasionally with caudal appendages; usually with lymphatic system (usually with one to three longitudinal trunks); genital sucker sometimes present; ventral pouch sometimes present; cirrus sac present or absent; hermaphroditic sac in some. Larvae: Cercaria lacking acetabulum or with posterior acetabulum, develops from redia or branched sporocyst; metacercaria usually encysts on vegetation, but occasionally on frogs, metacercarial stage occasionally lacking.

**Habitat:** Adults parasitic in digestive tract of marine and freshwater teleost fishes, snakes, marine turtles, amphibians, birds and a wide diversity of marine and terrestrial mammals (including humans), or in lungs and trachea of turtles and tortoises.

28. Plagiorchioidea
See Fig. 30. (Perhaps most problematic of digenean superfamilies, clearly paraphyletic as currently circumscribed.) Body usually spined; oral sucker occasionally with pair of muscular lateral papillae; acetabulum rarely vestigial; fields of vitellarium usually restricted anteriorly and posteriorly; excretory vesicle Y-shaped. Larvae: Cercaria often xiphidiocercaria of leptocercous, microcercous, cysto cercus, armatao, ornatae, virgulate, ubiquita, or or caecariaeum type; often developing in daughter sporocysts in gastropod or lamellibranch molluscs; metacercaria encysts in terrestrial arthropods or gastropods.

**Habitat:** Adults parasitic in digestive system of mammals (especially bats), turtles, amphibians, reptiles, and occasionally fishes; liver and bile ducts of mammals, birds, and reptiles; cloaca and oviducts of birds and mammals; bursa fabricus of birds; urinary bladder of fishes and amphibians; coelom or pericardial chamber of elasmobranchs; occasionally progenetic in leeches and crustaceans.

29. Schistosomatoidea
See Fig. 31. Dioecious or monoecious, if dioecious male with ventral gynecophoral canal in which female resides for much of her life; surface of male body often with tubercles, with or without spines; pharynx usually lacking; both oral sucker and acetabulum present, or one or both lacking; ceca of gut often anastomosing at several points along length; one, two or many testes; genital pores separate or combined. Larvae: Cercaria apharyngeate, brevifurcate, sometimes also lophophorate; metacercarial stage lacking; cercaria develops into schistosomula; first host usually a mollusc but rarely an annelid.

**Habitat:** Adults parasitic in blood vessels and/or heart of freshwater and marine teleosts, elasmobranchs, holocephalans, usually freshwater (but sometimes marine) birds and mammals (including humans), crocodiles, and marine and freshwater turtles.

30. Transversotrema toidea
See Fig. 32. Body transversely elongated; spined; ceca united posteriorly to form transverse cyclocoel; excretory vesicle curved to right, receiving collecting vessels from both sides of body; cirrus absent; genital pore anterior. Larvae: Cercaria brevifurcate with pair of anterior lateral appendages; metacercarial stage lacking; cercaria infects definitive host cutaneously and develops directly into adult stage.

**Habitat:** Adults parasitic under scales of freshwater, euryhaline, and marine teleosts.

31. Troglo trema toidea
See Fig. 33. Body spined or not; oral sucker and pharynx sometimes weak; cirrus sac usually lacking; vitellarium follicular or often in two lateral dendritic clumps; genital pore anterior or posterior. Larvae: Microcercous xiphidiocercaria; metacercaria encysts on invertebrates or vertebrates.

**Habitat:** Adult parasitic in lungs of mammals (including humans) and frontal sinus, intestine and kidneys of some insectivores; also known from nasal fossa
of foxes; abscesses in skin of cats; trachea and esophagus, or bursa fabric of passeriform birds.

32. Zoogonoidea
See Fig. 34. Body usually spined; 1–11 testes; cirrus sac usually present; cirrus armed with spines in some; vitelline follicles usually few, compact or often arranged in lateral bunches. Larvae: Cercaria chaetomicro cercous or tailless; metacercaria encysts in marine lamellibranchs, freshwater gastropods, turbellarians, aquatic annelids, or echinoderms; occasionally second intermediate host lacking.

**Habitat:** Adults parasitic in stomach or intestine of marine or freshwater fishes, or ovary of marine fishes, rarely coelom, gall bladder or urinary bladder of teleost fishes.

C. Class Monogenea

The Monogenea (or Monogonoidea) generally possess both anterior and posterior attachment structures. The form of the anterior attachment structure varies among groups, consisting of one or two suckers, pseudosuckers, or grooves. In some cases this region is equipped with gland cells. In the past this anterior structure has been referred to as a prohaptor. The posterior attachment structure also varies widely in morphology among taxa. This structure is generally now referred to as a haptor, although in the past the term opisthaptor has also been used. The haptor may be symmetrical or asymmetrical, may bear glands, suckers, and/or may be subdivided into numerous loculi. In addition, the haptor of most monogeneans is armed with a series of hardened structures greatly facilitating attachment to the host. The terminology of these structures is based to some extent on their morphology, but also to a large extent on when they appear in the course of development. For example, the small hooked structures that are usually found around the periphery of the haptor of the oncomiracidium and that persist into the adult stage of some monopisthocotyleans are termed hooklets or uncinuli. The larger, more centrally located hooked structures are generally termed hamuli or anchors, depending on whether or not they are present in the oncomiracidium. The more complex structures consisting of multiple sclerites, and which are found only in polyopisthocotyleans, are termed clamps. Adults of some, but not all, monogeneans possess one or more pairs of anterior ocelli, but monogeneans generally possess four rhabdometric ocelli in at least one stage of their development. The mouth is almost always located in the anterior region of the body. Most monogeneans possess a conspicuous muscular pharynx and an intestine that is usually divided into two, rarely one, blind ceca, which are occasionally branched or anastomose posteriorly. All monogeneans are hermaphroditic. The male and female genital pores are usually combined. The number of testes ranges from one to many among different groups. Most groups possess a single ovary that varies somewhat in form from compact to double inverted U-shaped. The male intromittent organ can be a penis or a cirrus, but the exact form of this feature is not well known in some groups. The terminal genitalia are often complex. The excretory system consists of a pair of prothonephridial systems that open separately to the outside, usually in the anterior of the body. Adult monogeneans are generally external parasites of cold-blooded vertebrates, but a few occur internally in the body cavity or very rarely the digestive system of their vertebrate hosts. The majority of the monogenean groups possess a ciliated larval stage known as an oncomiracidium, although the larvae of several groups are unciliated. The oncomiracidium bears a posterior haptor that develops into the haptor of the adult form.

The monogenean orders are treated below in sequence according to their position from left to right on the tree in Fig. 35. Membership in the families and even orders is quite stable although the monophyly of some orders remains uncertain.

To make the following diagnoses more concise, unless otherwise mentioned in the diagnosis of a group, it should be assumed that the order possesses a posterior, discoidal, undivided, symmetrical haptor, a conspicuous muscular pharynx and an intestine that is usually divided into two, rarely one, blind ceca, which are occasionally branched or anastomose posteriorly. All monogeneans are hermaphroditic. The male and female genital pores are usually combined. The number of testes ranges from one to many among different groups. Most groups possess a single ovary that varies somewhat in form from compact to double inverted U-shaped. The male intromittent organ can be a penis or a cirrus, but the exact form of this feature is not well known in some groups. The terminal genitalia are often complex. The excretory system consists of a pair of prothonephridial systems that open separately to the outside, usually in the anterior of the body. Adult monogeneans are generally external parasites of cold-blooded vertebrates, but a few occur internally in the body cavity or very rarely the digestive system of their vertebrate hosts. The majority of the monogenean groups possess a ciliated larval stage known as an oncomiracidium, although the larvae of several groups are unciliated. The oncomiracidium bears a posterior haptor that develops into the haptor of the adult form.

FIGURE 35  Hypothesized phylogenetic relationships among monogenean orders (after Boeger and Kritsky, 1993).
ous muscular pharynx, a gut consisting of two blind intestinal ceca, paired testes, biflagellated sperm, flagella incorporated into body of sperm, 9 + "1" microtubule arrangement, a single ovary; a vitellarium arranged in two extensive lateral fields of follicles, ectolecithal eggs, and a single, ventral genital pore.

33. Dactylogyridea
See Fig. 36. With two or more pairs of head organs, with or without cephalic glands; posterior half of body covered with anteriorly directed cuticular spines or not; haptor with one or two pairs of anchors supported by one to four transverse bars, with or without accessory plaques, occasionally with two suckers and median terminal anchor complex, usually 14 or 16 marginal hooklets; mouth subterminal, ceca fused posteriorly or not, with lateral diverticulae or not; testes, one to many; sperm with one axoneme; vas deferens encircling left caecum; seminal vesicle usually present; prostatic complex present; male copulatory organ sclerotized, elongated, muscular, ovary compact, lobed or tubular; vitellarium follicular or occasionally divided into tubular lobules in form of frond-like sprays spreading laterally in two or three groups (Protogyrodactylidae), coextensive with intestinal ceca; seminal receptacle and vagina present or absent; genito-intestinal canal absent; oviparous; eggs oval or tetrahedral; genital aperture median or marginal.

Habitat: Adults parasitic on marine copepods that are themselves parasitic on marine teleosts or elasmobranchs.

34. Gyrodactylinea
See Fig. 37. With two head organs; haptor usually with one pair of anchors often supported by dorsal and ventral transverse bars with 16 marginal hooklets; mouth ventral, intestinal ceca, one or two, usually not fused posteriorly; testis single; sperm microtubules absent; vas deferens encircling left caecum; male copulatory organ sclerotized or muscular, spines present or absent; accessory piece present; vitellarium symmetrical, near posterior part of intestinal ceca; vagina absent or present, opening ventrally near left margin of body; genital pore marginal; many species viviparous; eggs oval or tetrahedral.

Habitat: Adults parasitic on gills and skin of marine and freshwater, occasionally euryhaline teleosts; rarely on skin of elasmobranchs.

35. Udonellidea
See Fig. 38. Pair of head organs or pseudosuckers usually present; haptor unarmored in all developmental stages, often glandular, mouth terminal, intestine saciform; with single testis; prostatic complex present; ovary compact; vagina absent; ciliated larval stage completely lacking. It should be noted that recent evidence suggests that the udonellids may represent an unusual group of gyroactylineans.

Habitat: Adults parasitic on marine copepods that are themselves parasitic on marine teleosts or elasmobranchs.

36. Montchadskyellidea
See Fig. 39. Haptor with one pair of anchors lacking supporting bars, with 14 marginal hooklets; mouth subterminal, gut diverticulae present; testis single; male copulatory organ sclerotized; vas deferens encircling left caecum; accessory piece present; ovary encircling right caecum; one midventral vagina present; genital pores separate, opening on separate ventral papillae.

Habitat: Adults parasitic in gut of marine teleosts.

37. Capsalidea
See Fig. 40. Anterior end with paired petaloid head organs or paired glandular pseudosuckers; haptor sometimes subdivided by septa into central area and multiple peripheral loculi, with 14 marginal hooklets and two or four central anchors and two anterior sclerites; mouth ventral; intestinal ceca fused posteriorly, with lateral and axial diverticulae; testes, two to many; sperm microtubules absent; vas deferens forming bipartite external seminal vesicle and simple internal vesicle, or seminal vesicle absent; penis sac present or absent; penis present or absent; penis elongate, muscular, spines absent; prostatic complex present; vagina present or absent; seminal receptacle present or absent; genital apertures usually combined, marginal.

Habitat: Adults parasitic on skin or gills of marine teleosts (including remoras), or on skin or gills of elasmobranchs.

38. Monocotylidea
See Fig. 41. With single anterior sucker or several sucker-like depressions or paired head organs (lobes) with cephalic glands; haptor aspetate or divided by septa into loculi, often with one pair of anchors, with or without marginal papillae, usually with 14 marginal hooklets; ocelli present or absent; mouth ventral, intestinal ceca occasionally fused posteriorly, occasionally with lateral diverticulae; with 1 to several testes; sperm microtubules lying along 1/4 of cell periphery, with 2 axonemes 1 of which is reduced; male copulatory organ sclerotized; ovary tubular, encircling right intestinal
WORMS, PLATYHELMINTHES
caecum; ventral vaginal pore sometimes present; genital pore usually median.

**Habitat:** Adults parasitic on skin, gills and sometimes nasal bulbs of elasmobranchs; occasionally in coelom, oviducts and cloaca of elasmobranchs; occasionally on skin of holocelphants.

39. Polystomatidea

See Fig. 42. With anterior muscular oral sucker; haptor with three but occasionally one pair of suckers, one to two pairs of hamuli usually present, accessory spines or spurs sometimes present, marginal hooklets often present, medial appendix-like prolongation sometimes present; ocelli usually absent; mouth terminal or subterminal; ceca occasionally fused posteriorly, often with lateral diverticulae; testes, one to many; seminal vesicle present; male intromittent organ usually armed with hooks or spines; genito-intestinal canal present; ovary compact; vitellarium rarely compact; uterus anterior or opposite ovary or reaching posterior of body; with two lateral vaginas.

**Habitat:** Adults parasitic in mouth, pharynx, esophagus, or urinary bladder of batrachian and anuran frogs; mouth, esophagus, nasal passages, or urinary bladder of turtles; urinary bladder of urodeles; eyes of hippopotamus; or skin and gills of batrachians.

40. Chimaericolidea

See Fig. 43. Single, weak circumoral sucker usually present; body with highly contractile, sometimes with very long, posterior peduncle; haptor with four pairs or four alternating rows of two short-stalked symmetrical clamps, and terminal lappet-bearing hooks; ceca with lateral diverticulae, extending posteriorly into haptor; with numerous testes; seminal vesicle present or absent; cirrus muscular, armed or not; prostatic complex present or absent; ovary lobed or branched; with two lateral vaginas opening into vitelline duct; seminal receptacle present or absent; uterus with numerous longitudinal loops, with or without lateral branches; genito-intestinal canal present.

**Habitat:** Adults parasitic on gills or in gill chamber of holocelphants.

41. Dicybothriidea

See Fig. 44. With anterior muscular, paired ventral bothria or oral sucker; haptor with three pairs of sessile suckers, each with single, large, curved, hooked sclerite, with medial, haptoral appendix bearing three pairs of hamuli or one pair of terminal suckers and two small anchors; mouth terminal or nearly so; ceca united posteriorly, with lateral diverticulae; testes numerous; male intromittent organ usually armed; ovary tubular, convoluted; with two lateral vaginas; uterus prevovarian; genito-intestinal canal present; eggs operculate, with or without polar filaments, occasionally forming chains.

**Habitat:** Adults parasitic on gills of sturgeon, paddlefish (chondrosteans) and elasmobranchs.

42. Mazocraeidea

See Fig. 45. With two anterior oral suckers, occasionally enlarged in form of discoid adhesive organ with conical papillae; haptor symmetrical or asymmetrical, set off from body by distinct peduncle or not, distinctly bilobed or not, usually with four to numerous pairs of clamps, without marginal hooklets, terminal lappet usually present; terminal lappet with one to three pairs of similar or dissimilar hook-like sclerites, bifid or not; testes numerous; copulatory apparatus consisting of sheaf of long spines or longitudinally striated sheath; male intromittent organ present or absent; armed or unarmed; ovary tubular, winding, convoluted, rarely compact or inverted U-shaped; vitellarium sometimes extending posteriorly into haptor; seminal receptacle present or absent; vagina present or absent, paired or not; vaginal aperture spined or not, conical sclerite in front of vaginal pore or not; eggs with two filaments; genito-intestinal canal present; genital pores common or separate, genital pore sometimes sucker-like; in rare cases (Diplozoidae) two adults permanently fused in form of letter X.

**Habitat:** Adults parasitic on gills and skin, or occasionally mouth cavity of marine, or sometimes freshwater teleosts; sometimes on crustaceans in mouth cavity of teleosts.
D. Class Cestoda

The cestodes, or tapeworms, conspicuously differ from almost all other platyhelminth groups in their lack of all organs of a digestive system, the only exceptions perhaps being the amphilinideans and gyrocytolyideans, the anterior attachment organ of which is considered by some to represent a vestige of a digestive system. Thus, in general, the cestodes have no mouth, pharynx, esophagus, or intestine at any stage in their development. Rather, they absorb nutrients, in the form of very small molecules, directly through the outer layer of their body. There is evidence to suggest that neodermal extensions known as microtriches assist in this process. Microtriches appear to be unique to tapeworms. These structures typically consist of an outer electron dense cap composed of multiple microtubules, a baseplate, and a base consisting of an extension of the cytoplasm of the tegument (Fig. 1). The entire structure is surrounded by a plasma membrane. The form and height of the electron dense cap varies rather significantly among different tapeworm groups. These structures may also assist with attachment.

The majority of tapeworms have a body consisting of an attachment region called the scolex, followed by a series of proglottids collectively referred to as the strobila. Scolex morphology varies significantly among the various tapeworm groups; indeed, recognition of the major subgroups of tapeworms is based to a large extent on scolex form. The scolex is simple in form in some groups, but it is more commonly equipped with attachment structures such as hooks, suckers, or hooked tentacles. It is also common for the scolex to be subdivided into two flexible bothria or four muscular bothridia. Bothridia can be distinguished from bothria in their possession of discrete inner and outer membranes bounding their musculature, visible most readily in histological section.

The proglottids of most tapeworms are hermaphroditic, each containing one or more sets of male and female reproductive organs. A few groups of tapeworms (specifically the amphilinideans, gyrocytolyideans, and caryophyllideans) lack proglottids and possess only a single set of male and female reproductive organs rather than multiple sets (i.e., they are monozoic rather than polyzoic). One group of tapeworms, the spathebothridideans, lack distinct proglottids but possess multiple sets of reproductive organs arranged in a linear series (i.e., they are nonproglottid but polyzoic). The majority of tapeworms, however, are proglottized, and each proglottid bears one or more sets of reproductive organs (i.e., they are proglottized and polyzoic). In proglottized tapeworms, a germinative region located directly behind the scolex consecutively produces the proglottids of the strobila. Thus, the proglottid nearest the scolex is the youngest, and that furthest from the scolex is the oldest. In some groups of tapeworms, the posterior-most proglottids, upon reaching a certain level of maturity, drop from the strobila (apolytic) and may continue their development independent of the main body of the worm. In other groups, the proglottids are retained on the strobila essentially throughout the life of the animal (anapolytic). In the former instance the proglottids may complete the production of eggs independent of the strobila; in the latter case proglottids on the strobila produce eggs that are released through uterine pores or tears. Proglottids filled with eggs are termed gravid. Consecutive proglottids either overlap (craspedote) or not (acraspedote). A very few species exhibit proglottids that are either male or female. Because these proglottids are found in the same strobila, individuals of these species are considered to be hermaphroditic, but the proglottids themselves are not. There is, however, evidence of dioecy in individuals of a few tapeworm species in which all of the proglottids of the strobila are either entirely male or entirely female.

With almost 650 recognized genera, the cestodes are the second most diverse group of neodermatans. The most recent comprehensive systematic treatment of the group is that of Khalil et al. (1994), in which 14 orders were recognized. The tree in Fig. 46 summarizes some of the ideas of cestode interordinal relationships that are more widely accepted. Membership in some orders remains, to a large extent, based on the vertebrate group parasitized by the adult worms, complicating ordinal circumscription.
The terms Cestoidea, Cestodaria, and Eucestoda are also sometimes used in reference to tapeworms. The term Cestoidea is synonymous with the term Cestoda; the terminology of Hyman (1951) and Grassé (1961) is followed here. In the past, the amphilinids and gyrocothylideans have been collectively referred to as the Cestodaria. Evidence against the monophyly of this group is mounting, and, thus, this term will not be used here. The term Eucestoda is used to collectively refer to the cestodes other than the amphilinids and gyrocothylideans. The eucestoda are distinguished from these groups, for example, by their possession of sperm that lack mitochondria.

To make the following diagnoses more concise, unless otherwise indicated in the diagnosis of a group, it should be assumed that the members of an order possess an anterior scolex of some form, lack all vestiges of a digestive system, and possess a strobila consisting of numerous proglottids that each contain one full set of both male and female reproductive organs (polyzoic). These proglottids contain numerous, compact testes filling much of the region anterior to the ovary, sperm lacking a mitochondrion, flagella that are incorporated into the main body of sperm, 9 + 1 microtubule arrangement in the flagella of the sperm, a cirrus armed with microtriches, a cirrus sac, a single posterior ovary, a vitellarium consisting of numerous follicles that extend in two lateral bands, one down each side of the proglottid, a Mehlis' gland located posterior to the ovarian bridge, a uterus that is saciform, and male and female systems that open through a combined lateral genital pore, with pores of consecutive proglottids alternating irregularly down the length of the strobila.

43. Amphilinidea
See Fig. 47. Distinct scolex absent but with anterior coniform “sucker”; one set of male and female reproductive organs (monozoic) present; testes in two lateral bands; sperm with mitochondrion; cirrus sac absent; ovary varying in shape; uterine pore anterior, uterus often N-shaped and/or sinuous; seminal receptacle present; genital pores usually separate, posterior.

Habitat: Adults parasitic in body cavity of chondrostens, some freshwater and marine teleosts, and Australian freshwater turtles.

44. Gyrocothylidean
See Fig. 48. Distinct scolex absent but with anterior “sucker,” rosette usually present at posterior end of body; body usually “spined” (spines are likely to be microtriches), “spines” limited in distribution or not; excretory system reticulate; one set of male and female reproductive organs (monozoic) present; sperm with mitochondrion; cirrus sac absent; ovary follicular, in V- or U-shaped band around seminal receptacle; seminal receptacle present; uterus sinuous, in median region of anterior part of body, terminal portion expanded into sac, opening into ventral uterine pore; genital pores separate, both anterior.

Habitat: Adults parasitic in spiral intestine of holocephalans.

45. Caryophyllidea
See Fig. 49. Scolex usually with one to three pairs of weakly developed shallow depressions or grooves (considered to be loculi if temporary, bothria if more permanent), occasionally with terminal introvert or disc; with one set of male and female reproductive organs (monozoic); ovary H- or dumbbell-shaped; vitellarium usually preovarian, occasionally postovarian follicles also present; seminal receptacle present or absent; uterus often sinuous; female pore a combined utero-vaginal pore; genital pores separate or combined, median, ventral.

Habitat: Adults parasitic in spiral intestine of siluriform and cypriniform freshwater fishes.

46. Spathébothriidea
See Fig. 50. Scolex usually bearing one or two sucker-like attachment organs, but occasionally absent; osmoregulatory system in form of canals on each side of strobila, with irregular commissures and lateral pores; body proglottization absent; with multiple sets of male and female reproductive organs serially repeated down body (polyzoic); testes in two irregular lateral bands that extend throughout the length of the body; ovary rosette-shaped or bilobed, median, posterior to genital pores; small seminal receptacle occasionally present; uterus often sinuous; male and female pores separate, opening close to one another along median line, usually irregularly alternating in adjacent proglottids between opening dorsally and ventrally.

Habitat: Adults parasitic in intestine of freshwater, euryhaline and marine chondrostes and teleosts.

47. Pseudophyllidea
See Fig. 51. Scolex usually with two bothria (one dorsal and one ventral), margins of bothria occasionally crenulated or rolled, each bothrium occasionally armed with one pair of hooks or apical series of hooks, occasionally with posterior appendages, occasionally divided into two or several vertical loculi, with or without muscular apical disc, scolex occasionally replaced by scolex deformatus or pseudoscolex, scolex occasionally lacking; polyzoic, proglottization usually complete, proglottids
craspedote or acraspedote, generally anapolytic; usually with one set of male and female reproductive organs per proglottid, but occasionally with up to eight sets per proglottid; testes usually arranged in two lateral bands throughout length of proglottid; cirrus-sac occasionally absent; cirrus with or without microtriches; ovary compact or dendritic, vitellarium rarely in single compact postovarian mass, vitelline follicles lateral or encircling proglottid; uterus tubular or saciform with lateral diverticulae, coiled or not; preformed ventral uterine pore present; genital pores usually common (one for each set of reproductive organs per proglottid), usually median, occasionally marginal.

**Habitat:** Adults parasitic in intestines of freshwater and marine fishes, sometimes piscivorous mammals (such as for example, bears, seals, cetaceans, felids), piscivorous birds, rarely varanid lizards, boid snakes, anurans and caudates.

48. Haplobothriidea
See Fig. 52. Body bearing primary and secondary scolexs and strobila; primary scolex with four tentacles with spiniform microtriches at bases, each tentacle with muscular sac into which it can be withdrawn, bothria, bothridia and suckers absent; primarily proglottid proglottized, but primary proglottids never develop reproductive organs on primary strobila; proglottized regions of primary strobila separate from remainder of primary strobila and develop into secondary strobilae that produce proglottids, each proglottid bearing one set of reproductive organs; anterior-most proglottid of each secondary strobila modified as secondary scolex; secondary scolex with four shallow anterior indentations surrounding central, apical dome; cirrus-sac median, anterior, cirrus armed with microtriches; testes numerous, arranged in two lateral fields; ovary inverted U-shaped; vagina posterior to cirrus-sac; seminal receptacle present; follicles of vitellarium arranged in two lateral fields, confluent in midline at anterior and posterior regions of proglottid; uterus with dilated uterine sac; preformed uterine pore present or absent; genital pores separate, median, anterior; eggs operculate.

**Habitat:** Adults parasitic in intestine of the bowfin *Amia calva.*

49. Diphyllidea
See Fig. 53. Scolex with two bothria, proximal surfaces of bothria occasionally with spines; scolex with apical organ; apical organ usually armed with one row of large hooks on dorsal and ventral surfaces, often with row of smaller hooks on lateral margins, occasionally with small spines; bothria supported by cephalic peduncle; cephalic peduncle often armed with multiple columns of straight hooks with trifid bases; proglottids acraspedote; generally anapolytic; testes usually filling proglottid anterior to ovary; ovary bilobed; lateral bands of follicles of vitellarium occasionally extending medially; vagina posterior to cirrus-sac; uterine pore lacking; common genital pore median, ventral.

**Habitat:** Adults parasitic in spiral intestine of rays, skates, and sharks of families Hemiscyliidae, Scyllorhinidae, and Triakidae (all marine).

50. Trypanorhynchida
See Fig. 54. Scolex with four retractable, hollow tentacles each armed with numerous spiral rows of hooks (hook patterns of enormous taxonomic importance); each tentacle everted and retracted by rhynechal apparatus consisting of tentacle sheath, and retractor muscle originating within posterior muscular bulb (tentacles and rhynechal apparatus rarely absent); scolex with two bothria and elongated peduncle housing rhynechal apparatus; proglottids craspedote or acraspedote, apolytic or anapolytic; usually with one set of male and female reproductive organs per segment, but occasionally with two sets; cirrus-sac or hermaphroditic-sac usually present; external, internal and accessory seminal vesicles present or absent; ovary bi-, tetra-, or multi-lobed in cross-section; follicles of vitellarium lateral or completely encircling proglottid; vagina usually posterior to cirrus-sac, occasionally dorsal or ventral to cirrus-sac; uterus tubular or branched, with or without preformed uterine pore; common genital pore often associated with muscular lips.
**Habitat:** Adults parasitic in spiral intestine, stomach, rarely gall bladder, of elasmobranchs or very rarely holoccephalans.

51. Tetraphyllidea
See Fig. 55. Scolex usually divided into four muscular sessile or pedunculated bothridia of a variety of forms, bothridia occasionally absent; bothridia with or without apical muscular pad with or without one or three accessory suckers, with or without one or two pairs of uni-, bi-, or tripronged hooks; bothridia may or may not be further subdivided into loculi, metascolex occasionally present; proglottids craspedote or acraspedote, generally apolytic; rarely with proglottids that are either male or female (or female proglottids retain male copulatory apparatus); usually with one strobila per scolex, rarely with multiple strobilae per scolex; ovary H-, dumbbell-, A-, or inverted U-shaped; vagina anterior to cirrus-sac, crossing male duct (vas deferens); follicles of vitellarium sometimes completely encircling proglottid, rarely condensed around ovary; uterine pores sometimes present or uterus opening through median slits in uterine wall; genital pores rarely separate; genital pores rarely unilateral along length of strobila.

**Habitat:** Adults parasitic in spiral intestine of marine and freshwater elasmobranchs.

52. Proteocephalidea
See Fig. 56. Scolex usually with four round suckers or bothridia; suckers or bothridia uni-, bi-, tri-, or tetriloculate; apex of scolex sometimes with armed rostellum, or large sucker, or occasionally with apical organ in form of two lappets; metascolex sometimes present as wrinkled region posterior to suckers that may or may not cover suckers of scolex; scolex with or without gland cells; proglottids craspedote or acraspedote, usually anapolytic; longitudinal muscle bundles usually conspicuous, occasionally inconspicuous; ovary usually dumbbell-shaped in dorso-ventral view; uterus usually with lateral branches (diverticulae) when gravid, opening through one or more longitudinal apertures; vagina usually anterior to, but occasionally posterior to, cirrus-sac; genital pores rarely separate; distinction between cortex and medulla usually marked.

**Habitat:** Adults parasitic in intestine of freshwater teleosts, occasionally in amphibians, terrestrial lizards (such as, for example, Varanus, monitor lizards) and snakes.

53. Lecanicephalidea
See Fig. 57. Scolex usually divided into two parts: an anterior region (pars apicalis) quite variable in form, elongate (myzorhynchus), pad-like (metaporhynchus), sucker-like, or consisting of a crown of nonhooked tentacles; posterior region (pars basalis) usually cushion-like, often bearing four circular suckers or bothridia; proglottids, craspedote or acraspedote, usually apolytic; testes, three to many, arranged in a single or multiple columns; external seminal vesicle usually present, often conspicuous, extending length of proglottid; ovary bi-lobed, multi-lobed or irregular in dorso-ventral view; follicles of vitellarium sometimes encircling proglottid; vagina almost always posterior to cirrus-sac.

**Habitat:** Adults parasitic in spiral intestine of marine elasmobranchs, primarily rays.

54. Nippotaeniidea
See Fig. 58. Scolex in form of single apical sucker; proglottids craspedote, anapolytic or apolytic; testes restricted to previtelline medulla of proglottid; cirrus-sac thin-walled; convoluted ejaculatory duct present; seminal vesicle present or absent; ovary dumbbell-shaped in dorso-ventral view, vitellarium consisting of two symmetrical compact masses, anterior to ovary; uterus in form of transverse coils in medulla; vagina posterior to cirrus-sac.

**Habitat:** Adults parasitic in digestive system of freshwater teleosts.

55. Tetrabothriidea
See Fig. 59. Scolex usually with four muscular bothridia; bothridia round, rectangular or occasionally triangular in form, often with auriculate muscular appendages; auricles sometimes fused to form apical “organ”; apical rostellum present or absent; proglottids craspedote; anapolytic; testes, few to many; ovary bi-lobed or highly lobed in dorso-ventral view; vitellarium in form of single compact mass, usually posterior to ovarian bridge; uterus transverse, with one or more dorsal pores; seminal receptacle present or absent; vagina occasionally armed with “spines”; genital pores combined or not; genital atrium often muscular, sometimes with distinct male atrial canal, or genital papilla, occasionally with “spines”; genital pores usually unilateral along length of strobila; vagina usually posterior to cirrus-sac.

**Habitat:** Adults parasitic in intestine of seabirds, pinnipeds or cetaceans.

56. Cyclophyllidea
See Fig. 60. Scolex usually with four round suckers; suckers armed with spines or not; often with apical organ in form of single sucker (rarely four apical suck-
WORMS, PLATYHELMINTHES

ers), some with apical, often protrusible rostellum; rostellum armed with one or two rings of hooks or not, with or without rostellar pouch; scolex divided into scolex and metascolex, or with pseudoscolex in some; proglottids craspedote or acraspedote; apolytic or anapolytic; usually with one but rarely two strobilae per scolex; usually with one but occasionally two sets of male and female reproductive organs per proglottid; strobila rarely with proglottids that are entirely either male or female, or male and female proglottids that alternate regularly along strobila; testes, one to many, usually anterior to ovary, occasionally posterior to ovary; cirrus-sac usually present; cirrus usually armed with spiniform microtriches, occasionally with stylet, occasionally with accessory sac; external and internal seminal vesicles present or absent; ovary dumbbell-shaped, or asymmetrical in dorso-ventral view; vitellarium usually single compact mass, occasionally bilobed mass, usually posterior to ovary; vagina usually present or absent, replaced functionally by cirrus-sac and modified male duct, or replaced by supplementary ducts running between seminal receptacles in consecutive proglottids in some taxa; uterus ring-shaped, or saccate, with or without lateral branches (diverticulae), or reticular, transverse in some, replaced by parateratine organ(s) in some, replaced by uterine capsules or parenchymatous egg capsules in some, uterus single or double, usually with preformed uterine pore; common genital pore rarely midventral; genital pores occasionally regular or unilateral along strobila; genital atrium conspicuous or not, with or without prominent radial musculature; vagina usually posterior to cirrus-sac, occasionally dorsal to ventral to cirrus-sac.

Habitat: Adults parasitic in intestine of amphibians, snakes, lizards, birds or mammals.

III. NERVOUS SYSTEM

A bewildering variety of neural organization and complexity characterizes the platyhelminths (for example, see Halton and Gustafsson, 1996). The basic plan involves an apical subepidermal brain consisting of multiple neurons, which can be uni-, bi- (especially in Neodermata), or multilobed. This is a true brain because it controls reflexes in the peripheral nerve net. The number of nerve cells (neurons) in the brain varies between 50 and 550 in free-living species, but this number is less well known for the neodermatans. The brain is generally connected to two, or sometimes more, main longitudinal nerve cords, each consisting of axons with cell bodies distributed at irregular intervals along the length of the cords. These main nerve cords are usually connected to one another by numerous transverse commissures in a ladder-like arrangement. In addition, a peripheral array of nerve plexuses (or nerve-nets) is found throughout the body. The attachment organs and the components of the female reproductive system involved in egg formation are especially well supplied with nerve plexuses, but plexuses are also found associated with the pharynx and intestine and below the surface and below the muscle layers of the body. There are separate plexuses for sensory, integrative and motor activities. The central nervous system (brain plus main cords) is more emphasized in neodermatan groups than in non-neodermatan groups. Cestodes may exhibit regional differences in the complexity of the subtegumental plexuses along the length of the body. Transmission electron microscopy suggests that platyhelminths exhibit an extensive diversity of types of sensory receptors.

Several different neuronal cell types are present, including unipolar, bipolar, and multipolar cells. Whereas unipolar neurons are generally confined to the ganglia of the brain and/or longitudinal nerve cords, cells of plexuses are dominated by bipolar cells. Individual neurons are highly secretory and contain a variety of types of vesicles in their cytoplasm. Platyhelminths utilize a diversity of signaling molecules and at least two different groups of neuropeptides. The nervous systems of platyhelminths appear to possess compounds similar to all of the major neurotransmitters known in taxa as complex as vertebrates. In addition, there are two groups of neuropeptides that appear to be unique to flatworms: neuropeptide F and FMRPamide-related peptides (FaRPs) both of which are fairly extensively distributed among the various platyhelminth groups. The existence of glial or glial-like cells in platyhelminths remains somewhat controversial.

IV. FEEDING AND DIGESTION

Other than acoels, a few dalyelloids, and the cestodes, flatworms generally have a digestive system divisible into a foregut and a cecum. The foregut leads from the mouth to the blind-ended sac-like caecum that is lined with a thin gastrodermis. Free-living species tend to be microphagous or predatory and utilize bacteria, unicellular alga, protists, and/or almost all types of invertebrates as sources of nutrition. Characteristically, the free-living species have a highly muscular, bulbous, suckorial pharynx. Depending on the species, the phar-
ynx can be protruded, everted, inserted, applied to, extended over, or used to envelop and swallow whole the various foodstuffs they actively seek or encounter. Food passes from the mouth via the pharynx to the esophagus and then to the caecum. The gastrodermis has glandular and phagocytic cells that digest and incorporate the nutrients.

Symbiotic and parasitic platyhelminths vary their feeding habits according to the hosts on or in which they live. Groups that possess suckers may have numerous secretory cells and sensory structures surrounding the mouth and sucker. Although similar in basic design to the foregut of free-living groups, the parasitic groups tend to exhibit distinct cell type configurations in the gastrodermis. In these groups, the junction between the foregut and caecum is invariably abrupt. Aspidogastreans have a single cell type in their gastrodermis that alternates cyclically between a secretory-absorptive phase and an autophagic-exocytic phase. Digeneans have a syncytial or cellular gastrodermis with only one cell type, and digestion takes place extracellularly in the lumen of the caecum. There is evidence, however, that some digeneans appear capable of acquiring nutrients directly through microvillar extensions of their body surfaces. In contrast, digestion in the caecum of monogeneans is intracellular. The polyopisthocotylean monogeneans tend to be blood-feeders. The nonpigmented cells in these taxa are thought to support and protect the pigmented digestive cells in the gastrodermis. In addition, at least some of these taxa possess a bucco-esophageal canal between the buccal cavity and the esophagus that permits regurgitation of intestinal contents. The monopisthocotyleans tend to feed on the epidermal cells and mucus of their host and have only a single gastrodermal cell type involved with endocytosis and digestion.

In the Cestoda and some parasitic dalyellids (for example, *Fecampia* and *Kronborgia*) all vestiges of a digestive system have been lost. These species rely solely on acquiring nutrients through their outer body wall. Among the Cestoda, and indeed all other neodermatan groups, the neoderm plays an active role in the biochemical transport of nutrients. In cestodes the surface area of the neoderm is markedly increased by the presence of microtriches, which enhance the availability of readily available nutrients to be actively absorbed into the parasite's body. Acoels have no gut lumen and digestion takes place in temporary vacuoles within the syncytial endoderm. Some acoels harbor Zooxanthellae and subsequently derive nutrients from the products of photosynthesis of these symbionts.

### V. Excretion and Osmoregulation

The protonephridial system is thought to function primarily in osmoregulation, whereas the removal of metabolic wastes generally occurs by way of diffusion through the outer body layer. However, removal of excess water can also be achieved through the gut, and in free-living species it is not uncommon to find excretory products stored in various tissues. Each protonephridial system is composed of a series of collecting ducts and ciliated flame bulbs. The beating of the cilia in the flame bulbs and collecting ducts directs and drives body fluids through filters formed by the terminal cells and/or the distal canal cells (of various forms depending on the taxon) into minor collecting ducts or canals, which empty into one or more major canals (nephrioducts) that either open to the outside through a nephriodipore or empty into an excretory vesicle that subsequently opens to the outside through a nephriodipore. This system is present in all platyhelminths except the acoelomorphs and a few catenulids in which excretion is thought to take place solely through the digestive syncytium. The protonephridial system is most developed in freshwater species. Indeed, this system is considered to have played a major role in allowing platyhelminths to invade freshwater habitats.

The protonephridial system of most taxa is paired, consisting of two nephrioducts, one running along the lateral margin of each side of the body; these nephrioducts connect to many flame bulbs throughout their length. Some of the catenulids are exceptional in their possession of a single dorsal nephrioduct that opens through a single posterior nephriodipore; some prosérianans also possess a single system, but some possess three protonephridial systems. In most of the non- neodermatan groups, the two nephrioducts open to the outside of the body through separate nephriodipoors that are found in the anterior half of the body, generally on the ventral surface. In most temnocephalideans, the two nephrioducts open into separate saccate excretory vesicles, located in the anterior of the body on either side of the pharynx, and the left and right protonephridial systems are generally connected by two transverse ducts, one at the anterior and one at the posterior aspect of the system.

The aspidogastreans possess a protonephridial system much like that of the temnocephalans but without the transverse ducts. The excretory system of digeneans differs somewhat from that of the aspidogastreans in
that the two nephridioducts generally open into a common posterior excretory vesicle. This vesicle takes on a number of different forms in different taxa; for example, it can be V-, Y-, or I-shaped. Unlike other platyhelminth taxa, the arrangement of the flame bulbs relative to the nephridioducts is regular and consistent in the cercarial stage of digeneans among individuals of the same species, but differs among individuals of different taxa, and thus this “flame cell formula” is often of significant taxonomic value. The larger number of flame bulbs and the difficulty of seeing most of the elements of the protonephridial system in the more robust bodies of the adult digeneans reduce the taxonomic utility of the flame bulb arrangement in adults.

The protonephridial system of monogeneans is also similar to that of temnocephalans, but the anterior transverse duct is absent in the polyopisthocotylans. The amphilinideans possess an excretory system consisting of a network of collecting ducts, rather than just two lateral nephridioducts. This network empties into an elongated excretory vesicle that opens to the outside through a posterior nephriodiopore. The flame bulbs are usually found in pairs, grouped in clusters throughout the length of the body. The gyrocotylidean system also includes two longitudinal nephriodiopods that are connected to one another by a number of secondary canals. The main nephriodiopods open into a nephriodiopore on the dorsal surface in the anterior half of the body. In most of the other cestode groups the protonephridial system consists of two dorsal and two ventral longitudinal nephriodiopods, the ventral ducts generally being larger in diameter than the dorsal ducts; the nephriodiopods are connected to multiple flame bulbs throughout the length of the body. Some cestodes do, however, have up to 12 longitudinal ducts. The pairs of dorsal and ventral nephriodiopods are connected to one another in the scolex, where they may undergo extensive winding. The left and right nephriodiopods are often connected to one another by way of transverse excretory ducts located in the posterior end of each proglottid; this condition is most commonly found in anapolytic eucestodes species. In some groups, the longitudinal vessels are connected to one another with numerous lateral nephriodiopods, resulting in a network-like arrangement of vessels. In the first proglottid formed (located at the posterior end of the strobila) the ventral pair of nephriodiopods often empty into a small vesicle at the posterior end of this proglottid. However, once this proglottid drops from the strobila, the ventral vessels open to the outside separately in the posterior end of the terminal proglottid of the strobila. Rohde (1991) provides a useful overview of the evolution of the protonephridial system in platyhelminths.

VI. REPRODUCTION

All platyhelminth species undergo sexual reproduction. Given the predominance of hermaphroditism in this phylum, there is some question about the relative frequency and importance of self-versus cross-fertilization in most species. In the few taxa in which this issue has been investigated in detail, cross- rather than self-fertilization appears to be the dominant form of sexual reproduction. Dioecy does, however, occur in some acoels, in some triclads, in some lecithoepitheliatans, in a few dalyelloids (for example, *Kronborgia*), within the digenean family Schistosomatidae, and in a few tetrathyridial and cyclophyllidean cestodes. To date, no instance of dioecy in aspidogastreans or monogeneans is known. Most platyhelminths are simultaneous hermaphrodites, each individual possessing a full complement of male and female reproductive systems. There is a general tendency for the male system to mature before that of the female (protandry). Sperm transfer usually occurs during copulation. But, in a few species of triclads, polyclads, rhabdocoels, and cestodes, sperm is injected directly through the body wall and then migrates through the body to the female reproductive organs. Fertilization is always internal.

Asexual multiplication in one form or another is also found in a number of platyhelminths, either in the adult or, most commonly, in one or more of the larval stages in neodermatans. Adults of many of the catenulids, acoels, and macrostomids undergo forms of asexual multiplication either as paratomy (regeneration before fission) or as architomy (regeneration after fission). These phenomena result in a linear chain of individuals that eventually separate and each develop into a mature worm.

Some adult triclads possess the ability to regenerate complete individuals from, in some cases, a very small portion of the body of the original individual. For example, it is possible to divide an adult freshwater triclad posterior to the pharynx and produce two complete adult individuals. More dramatic is the ability of adults of species belonging to the triclads genus *Phagocata* to divide into a number of separate fragments, each of which subsequently encysts and develops into a complete adult worm. Their ability to regenerate has made freshwater triclads favored animals for studies in wound healing and developmental genetics.
With the exception of the aspidogastreans, asexual multiplication is found in at least some members of all of the neodermatan groups. Digeneans typically undergo asexual multiplication in the larval stages that parasitize the first, usually molluscan, intermediate host. All digeneans possess one or more generations of sporocysts and/or one or more generations of redia. Asexual multiplication occurs in the generation of all larval stages developing within any generation (mother or daughter) of both sporocysts and redia. This form of asexual multiplication, termed polyembryony, results in multiple genetically identical individuals of the subsequent larval stage that develop within the saccate body of the parent generation. Thus, first generation sporocysts (mother sporocysts) produce either multiple second generation sporocysts (daughter sporocysts) or multiple first generation redia (mother redia), which in turn produce multiple first generation redia (mother redia) or multiple second generation redia (daughter redia), respectively. Second generation sporocysts produce either multiple cercariae or multiple redia that then subsequently produce multiple cercariae.

Asexual multiplication is uncommon among the monogeneans. It is, however, considered to be present in the form of sequential polyembryony (hypermiviviparity) in some members of the monogenean order Gyrodactyliidea. Rather than releasing eggs containing oncomiracidia, an adult can produce a second generation juvenile in utero (without sexual reproduction) that in turn can produce a third generation juvenile in utero, until one single individual may yield in excess of 20 individuals of the subsequent generation from within its own body, apparently without the intervention of a second individual or of gametes. As a result, the body of such asexually multiplying individuals in effect resembles a series of Russian dolls. There is evidence, however, of sperm in the seminal receptacle of some of these embryos, which has led some to question this as a truly asexual form of multiplication.

In tapeworms, asexual multiplication occurs in the terrestrial members of the order Cyclophyllidea that develop through larval stages known as coeneri and hydatid cysts. The fluid-filled cavity of these larvae is lined with a layer of germinative tissue that produces multiple protoscoleces, in the case of coeneri, and daughter hydatids that in turn are lined with germinative tissue that produces multiple protoscoleces, in the case of hydatid cysts. Each protoscolex has the potential to become an adult worm. The reproductive potential in these groups is enormous, with the hydatid cysts of some species producing hundreds of thousands, perhaps even millions, of protoscoleces.

FIGURE 61 Generalized digenean life cycle. (a) Adult in definitive host, undergoes sexual reproduction and produces eggs; (b) eggs released with faeces of host; (c) miracidium hatches from egg and penetrates first intermediate host (usually a mollusc); (d) miracidium develops into sporocyst containing numerous redia; (e) redia leave sporocyst and produce numerous cercariae; (f) cercaria escapes from first host and finds second intermediate host; (g) cercaria sheds tail, encysts, and develops into metacercaria; (h) second intermediate host containing metacercaria is eaten by definitive host.

VII. ONTOGENY

There are a remarkable variety of larval forms found among the platyhelminths, especially within and among the classes of neodermatans. In general, the life cycle of each class of neodermatan is characterized by one or more unique larval stage (see Figs. 61–63). This diversity reflects the great range of habitats invaded and life history strategies employed. Most non-neodermatans undergo direct development such that the zygote develops directly into a form resembling a young adult. As discussed by Ruppert (1978), however, there are several notable exceptions to this pattern. One species of catenulid develops from a zygote into a pelagic larval stage known as a Luther’s larva (Fig. 64); the larva is vermiform in shape, ciliated throughout, but in addition possesses several anterior rings of elongate cilia that aid in swimming. The zygotes of some polyclads develop into free-swimming, ciliated larval stages bearing arms. These larvae are generally referred to as Götte’s
FIGURE 62 Generalized monogenean life cycle. (a) adult on gills of definitive host (usually an aquatic vertebrate), undergoes sexual reproduction, and produces eggs; (b) eggs released into water; (c) oncomiracidium hatches from egg; (d) oncomiracidium finds definitive host.

FIGURE 63 Generalized life cycle of aquatic cestode. (a) adult in definitive host, undergoes sexual reproduction, and produces eggs; (b) eggs released with faeces of host; (c) coracidium larva hatches from egg and penetrates first intermediate host (usually a copepod); (d) coracidium develops into procercoid; (e) first host eaten by second intermediate host; (f) procercoid develops into plerocercoid; (g) second host eaten by definitive host (usually a piscivorous mammal).

Prior to developing into the adult form, aspidogastreans pass through a larval stage called a cotylocidium (Fig. 67), which is often but not always ciliated. This larval stage is free-swimming in at least some species. However, the life cycles of many species are not yet fully understood. In some cases it appears that the cotylocidium hatches from the egg and parasitizes the same host as the parent individual from which it was released. In other cases it appears that the cotylocidium hatches from the egg and moves into the water column to find a new host. In yet other cases the cotylocidium never hatches from the egg; rather, the egg is ingested by the molluscan host. In some species the cycle is direct, involving only a single host. In other species, the cycle may involve a second host. This host may be infected by ingesting infected molluscs.

Digenea undergo complex development in which certain intermediate life cycle stages are able to produce multiple copies of new generations before becoming sexually reproductive adults. Consequently the term “larva” is used loosely when discussing intermediate life cycle stages in the Digenea. The digeneans pass through a series of larval stages that are either parasitic or free-living, depending on their association with this sequence of hosts. A generalized digenean life cycle is illustrated in Fig. 61. As can be seen, the basic sequence of larvae is as follows: miracidium, sporocyst, redia, cercaria, and metacercaria. The miracidium (Fig. 68) possesses cilia on epidermal plates, or occasionally on tufts or bars. The neodermis is formed when the miracidium sheds its epidermal cell plates, thus all stages following the miracidium have an outer layer that is a neodermis. In most taxa the miracidium develops into a sporocyst, but occasionally may develop into a redia. Sporocysts and redia are both sacciform. The main difference between them is that the sporocyst has none of the components of a digestive system (Fig. 69), whereas the redia usually possesses at least a mouth, a pharynx, and often a single intestinal caecum (Fig. 70). These stages vary morphologically among species in their possession of lateral lobes, and in some cases birth pores;
the sporocysts of some species are branched. It is quite common for digeneans to possess more than a single sporocyst and/or redial generation in their life cycle. A first generation sporocyst (primary sporocyst) is essentially a miracidium that has lost its epidermal plates, whereas the second generation sporocysts (secondary sporocysts) are produced through asexual multiplication within the primary sporocyst. A first generation redia (primary redia) is most commonly produced through asexual multiplication within a primary or secondary sporocyst, but exceptionally develops directly from a miracidium that has shed its epidermal plates. In most taxa, the final result of several generations of sporocysts and/or redia is the development of a number of cercariae. Each cercaria generally possesses an anterior body and a posterior tail. The cercaria is by far the most morphologically diverse of the digenean larval forms. Indeed, classification of the digenean families is based to a large extent on cercarial morphology. An elaborate nomenclature based on morphology exists to describe these cercarial types (for example, see Cable, 1956). Examples of some of the more distinctive types include ophthalmoxiphidiocercaria (Fig. 71), furcocercous (Fig. 72), microcercous (Fig. 73), and cystocercous cercariae (Fig. 74). The metacercarial stage is usually achieved when the cercaria drops its tail and secretes a protective cyst around itself. The metacercaria is essentially a juvenile version of the adult worm (Fig. 75). However, variations on this general developmental pattern occur: the metacercarial stage may be lacking or it may be preceded by an unencysted mesocercarial stage.

Associations of these larval forms with their hosts are quite predictable. In general, the miracidium and cercaria are free-living forms. The miracidium hatches from the egg and seeks the first (molluscan) host; the cercaria escapes from the mollusc and seeks either the second intermediate or the definitive host. The sporocyst and redial stages generally parasitize the first (molluscan) host. The metacercarial stage encysts either on aquatic vegetation or in or on the tissue of a second intermediate host, which can be either an invertebrate or vertebrate, depending on the digenean species. Exceptions to this general pattern do occur. In some species, the miracidium is not free-swimming; rather, it remains within the egg, which is subsequently eaten by the first host. In some species, the cercaria does not leave the mollusc to swim freely; rather, it continues its development into a metacercaria, encysting within the body of the sporocyst or redia within the molluscan host. In species that lack metacercaria, such as the schistosomes, the cercaria penetrates the tissues of the definitive host directly. This stage is referred to as a schistosomula once it drops its tail and enters the host.

In general, the monogenean life cycle consists of two stages, the adult and a free-living, usually ciliated, larval stage known as the oncomiracidium. This larva possesses one or two pairs of ocelli, and is conspicuous in its possession of a posterior muscular attachment structure known as a haptor, which is usually armed with hooklets (uncini) and/or hamuli. The oncomiracidium differs from the ciliated larvae of other neodermatans in that its cilia are often restricted to three distinct regions. When the oncomiracidium hatches from an egg, which is sometimes anchored to the substrate, it swims or creeps to find its vertebrate host. An oncomiracidial stage is lacking from monogeneans such as the gyrocytuleans, that undergo sequential polyembryony. A generalized monogenean life cycle is shown in Fig. 62.

Both amphilinideans and gyrocytuleans possess a ciliated larval stage known as a lycophore larva (Figs. 77, 78 respectively). This larva possesses 10 hooks at the posterior end of its body; these hooks are of similar size and shape in gyrocytuleans but consist of six large and four small hooks, which also differ in shape, in the amphilinideans. This larval stage is also sometimes referred to as a decacanth. The life histories of many species in these two cestode groups are not well known. It appears that, in at least some amphilinideans, the lycophore does not escape from the egg until it is eaten by a small crustacean. In this intermediate host, the lycophore burrows through the gut and comes to rest in the hemocoel, where it develops into a proceroid by shedding its ciliated epidermal cells, elongating, and developing a posterior tail. It subsequently develops into a plerocercoid by shedding its tail and undergoing further elongation. In another case, the lycophore larva hatches and actively penetrates through the cuticle of the intermediate crayfish host. Development into the adult stage occurs only when the intermediate host is eaten by an appropriate vertebrate, at which point the plerocercoid sheds its tail and becomes sexually mature. It is possible that gyrocytuleans do not require an intermediate host to complete their development.

Most eucestodes require at least two hosts to complete their life cycles, but it is also common for additional paratenic hosts to occur in the life cycles of these platyhelminths. Paratenic hosts are not required, but serve to bridge food chain gaps, thereby facilitating infection of appropriate vertebrate hosts. A generalized aquatic eucestode life cycle is illustrated in Fig. 63. In general, eucestode zygotes initially develop into a larval
stage known as a hexacanth embryo (Fig. 79). This embryo characteristically possesses three pairs of hooks. The hexacanth may or may not be surrounded by a layer termed an embryophore. A hexacanth surrounded by an embryophore is referred to as an oncosphere. An oncosphere in which the outer region of the embryophore is ciliated is referred to as a coracidium and is usually free-swimming. Some cestode hexacanths or oncospheres are surrounded by an egg shell; this shell possesses a cap-like operculum in some groups. A hexacanth enters the first intermediate host either when it is eaten or by actively swimming to and then penetrating this first host. The first intermediate host is an invertebrate in most eucestode groups, but some terrestrial cyclophyllideans utilize vertebrate first intermediate hosts. The developmental fate of the hexacanth remains to be discovered in many cestode species; this is especially true of the trypanorhynchs and tetraphyllideans. However, existing data suggest that, in general, the larval stages that follow the hexacanth vary somewhat in form among the different cestode orders. These forms are generally given different names depending on their morphology. The key features for distinguishing among larval types are the presence or absence of a scolex, the number of scoleces, whether the scolex (or scoleces) are invaginated (turned inside out) or retracted (withdrawn) into the main body of the larva, and whether a cavity known as a primary lacuna is present. Although current terminology does not accommodate all known combinations of these features, some common larval forms are given (see following). For example, procercoids are solid-bodied, lack a scolex, and usually possess a posterior extension of the body considered a cercomer by some (Fig. 80). Plerocercoids possess a scolex that is neither invaginated nor retracted (Fig. 84). Cysticercoids possess a conspicuous scolex that is retracted into the tissue of the neck alone, or, in some cases, the neck is then retracted into the main body of the larva (Fig. 81). Cysticerci possess one, or more invaginated scoleces. Cysticerci with multiple (Fig. 82) scoleces are referred to as either hydatid cysts (Fig. 83) or coeneri, depending on whether they have the ability to generate daughter cysts or not, respectively. The reproductive potential of hydatid cysts is enormous as each scolex (termed a protoscolex in hydatids) can go on to become either an adult worm or another protoscolex, if released from the original cyst within the first host. The numerous marine eucestode larvae referred to as “Scolex plerotypus” or “Scolex polymorphus” are likely to be the plerocercoid larvae of tetraphyllideans.

VIII. PHYLOGENETIC RELATIONSHIPS

The tree shown in Fig. 1 provides an overview of the general understanding of the phylogenetic relationships among the major platyhelminth groups. The topology of this tree has been modified from several sources (e.g., Littlewood et al., 1999). The characters that have been mapped onto the tree include some of the more conspicuous characters that support these relationships, as well as some of the useful distinguishing characteristics of major groups. As this tree includes only a mapping of some useful characters it should not be interpreted as a cladogram.

Few unique features (synapomorphies) have been identified for the phylum Platyhelminthes. None of the characters normally used to describe the group (i.e., dorsoventral flattening, acelolate, bilateral symmetry) is unique to this group. Multiciliated epidermal and gastrodermal cells are candidates as diagnostic features for the phylum. However, multiciliated epidermal cells are also found in the phylum Nemertini, suggesting that this character may be more appropriately considered as diagnostic at a higher taxonomic level. The fact that all of the cestodes and some parasitic dalyelliids lack a gastrodermis (and in fact a digestive system) means that gastrodermal ciliation is not universally diagnostic for the platyhelminths either. Nonetheless, transmission electron microscopy indicates that the lack of accessory centrioles from the base of the cilia of the protonephridial flame bulbs and epidermis may be truly synapomorphic for the phylum as a whole, as may be the presence of more than one cilium in the flame bulbs of the protonephridial system.

Some morphological (Smith et al., 1986) and recent molecular data tend to confirm the monophyly of many of the major platyhelminth groups but have highlighted the need for further systematic research. For instance, there is increasing evidence that the Acoela and Nemertodermatida (collectively known as the Acoelomorpha) may not belong in the Platyhelminthes. The nemertodermatidans are unique among the platyhelminths in their possession of sperm that is monoflagellated rather than biflagellated. The acoels show unique neural organization and developmental patterns. Typically appearing as basal bilaterians and not platyhelminths in molecular phylogenetic analyses, the acoelomorphs lack a protonephridial system and are the only flatworms with glandular frontal organs. In addition, these two groups lack a well-defined gut and possess nonla-
mellated rhabdoids rather than lamellated rhabdites. However, they have been included here until their phylogenetic position is more formally addressed. Relationships among the remaining nonneodermatan platyhelminth taxa are even less well understood. There is some agreement that the catenulids are among the more basal platyhelminths. The members of this order are exceptional among the platyhelminths in their possession of aflagellated sperm and one rather than two protonephridial systems as well as genital pores that are dorsal rather than ventral in position. With the exception of the acelomorphs, catenulids, and proseriatans (and at least some dalyelliidans and temnocephalidans) all of the nonneodermatan groups possess a unique duogland adhesive system consisting of two different types of gland cells, one that produces an adhesive substance and one that produces a substance that reverses (releases) the action of the adhesive substance. In addition, all of the nonneodermatan taxa, with the exception of the acelomorphs, catenulids, proseriatans and some dalyelliidans, possess lamellated rhabdites. Yet there is little evidence to suggest that these orders belong together in a distinct, monophyletic group. Rather, additional characters unite a subset of these orders with the neodermatans. Unfortunately, the affinities suggested by these characters are inconsistent. For example, all platyhelminths except the acelomorphs and catenulids possess flame bulbs with more than three cilia. All platyhelminths except the acelomorphs, catenulids, macrostomids, and prolecithophorans possess biflagellated sperm. All platyhelminths except the acelomorphs, catenulids, macrostomids, and prolecithophorans possess biflagellated sperm. All platyhelminths except the acelomorphs, catenulids, macrostomids, and prolecithophorans possess biflagellated sperm. All platyhelminths except the acelomorphs, catenulids, macrostomids, and prolecithophorans possess biflagellated sperm.

Unique features support the individual monophyly of many of the remaining nonneodermatan orders: the polyclads possess intestinal ceca that are highly branched, the triclads possess a triradiate intestine and an unusual embryonic intestine that does not appear to be homologous to that of the adult, all but one of the temnocephalans possess two to 12 anterior tentacles, and the haplopharyngids and kalyptorhynchs both possess an anterior proboscis independent of their digestive systems. The probosci of the two groups are not thought to be homologous.

The monophyly of the Neodermata, consisting of the aspidogastreans and digeneans (collectively the Trematoda) along with the monogeneans and cestodes (collectively the Cercomeromorpha), is now one of the most well-established aspects of platyhelminth relationships. The members of this group are characterized by their unique syncytial outer body layer known as the neodermis, which generally replaces the cellular epidermis when the first larval stage encounters the first host. In addition, the neodermatans possess protonephridial filters in the form of a two-celled weir. The sperm of neodermatans is unusual in that the two flagella are incorporated into the cytoplasm of the sperm cell and are arranged in a proximal–distal direction, rather than remaining separate. All neodermatans also lack lamellated rhabdites and the duogland adhesive system and possess cilia with only one rather than two rootlets. All neodermatans are obligate parasites at some time during their lives.

The Trematoda are characterized by their possession of epidermal cells in their first larval stage (miracidia and cotylocidia) that are separated from one another by portions of the neodermis. In addition, the first or primary host in the life cycle of the trematodes is a mollusc. The Aspidogastrea are unique in their possession of oviducts separated into chambers by septa and neodermal microvilli in the form of hemispherical microtubercles. The Digenea are distinctive in their possession of several unique larval stages, the miracidium, sporocyst and/or redia, and cercaria. Perhaps in part because they are by far the most diverse of the major

FIGURES 64–84 64 Catenulida: Luther’s larva (modified from Ruppert, 1977); 65 Polycladida: Gottle’s larva (modified from Ruppert, 1977); 66 Polycladida: Müller’s larva (modified from Ruppert, 1977); 67 Aspidogasteridea: cotylocidium; 68 Digenea: miracidium; 69 Digenea: sporocyst; 70 Digenea: redia; 71 Digenea: ophthalmoxiphidiocercaria; 72 Digenea: furcocercous cercaria; 73 Digenea: microcercous cercaria; 74 Digenea: cystoercous cercaria; 75 Digenea: unencysted metacercaria; 76 Monogenea: oncomiracidium; 77 Cestoda: gyrocotylidean lycophore larva (redrawn from Beauchamp in Grasse, 1961); 78 Cestoda: amphilidean lycophore larva (redrawn from Beauchamp in Grasse, 1961); 79 Cestoda: coracidium; 80 Cestoda: procercoid (modified from Joyeux and Baer in Grasse, 1961); 81 Cestoda: cysticercoid (scolex retracted); 82 Cestoda: cysticercus (scolex retracted and invaginated); 83 Cestoda: hydatid cyst (modified from Hyman, 1951); 84 Cestoda: plerocercoid. (See abbreviations on pages 897–898.)
platyhelminth groups, there is currently no generally accepted, higher level classification scheme for the digeneans. A preliminary attempt at the generation of a tree for the families in this subclass was presented by Brooks et al. (1985). However, it is clear from the work of others (for example, Pearson, 1992) that a more detailed analysis of characters is required before these relationships can be resolved with any confidence.

The Cercomeromorpha are unique in their possession of, at least at some time in their development, a posterior extension of the body that is equipped with a number of sickle-shaped hooklets. The name for this group comes from the notion that this posterior extension of monogeneans and cestodes is homologous, an idea that is not universally accepted. The monophyly of both the monogeneans and the cestodes is fairly well established. The Monogenea are unique among the platyhelminths in their general possession of four rhabdomeric ocelli at some point in their lives, a ciliated larval stage known as the oncomiracidium (with the exception of the few viviparous forms), and the retention of the posterior haptor in the adult stage. The results of the most recent extensive cladistic treatment of the group (see Fig. 35 and Boeger and Kritsky, 1993) support recognition of two major subgroups within the class: the Polyopisthocotylea, which possess a haptor that is subdivided into multiple suckers and/or bears distinctive attachment structures known as clamps, and the Monopisthocotylea, which possess a haptor that, although it may bear loculi, is undivided and, although it may bear anchors or hamuli, never bears clamps. Although the monophyly of the Monogenea has been called into question by some molecular phylogenetic analyses, it seems likely that this reflects a rapid divergence of the two major subgroups rather than lack of support for monogenean monophyly. The interrelationships among the monogenean orders within these two major subgroups are not universally agreed upon. The phylogenetic position of the udonellids is particularly uncertain.

There is little doubt that the cestodes are monophyletic; all cestodes lack a gut and thus also a ciliated gastrodermis. All cestodes possess unique extensions of the neoderm called microtriches (Fig. 16). The relationships among the cestode orders remain controversial despite the fact that this issue has received much recent attention (for example, Hoberg et al., 1997; Mariaux, 1998; Olson and Caira, 1999). The tree in Fig. 46 summarizes some of the more generally accepted ideas on these relationships. There is evidence to suggest that many, but not necessarily all, of the individual cestode orders are monophyletic.

**IX. HOST ASSOCIATIONS**

It is useful to consider host associations in general terms. The tendencies of platyhelminths to live in association with other organisms are summarized on the tree in Fig. 1. One of the most striking aspects of this mapping is that it becomes apparent that all but five of the major platyhelminth groups (catenulids, haplopharyngids, lecithoepitheliata, macrostomids, and kalyptrateans) include at least some species that live in association with other organisms. Thus, symbiosis, in its broadest sense, is a widely distributed feature among the platyhelminths. Within the major platyhelminth groups, however, symbiosis varies in its occurrence. Only the four neodermatan groups and the tennocephalids consist entirely of obligate symbionts. The dalyellidans are predominantly symbiotic, although a few free-living dalyellidan species are known. The remaining groups (acoelans, nemertodermatidans, polyclads, proseriatans, tricladans, proleciophorans, and typhloplanidans) are predominantly free-living, but all of these groups also include at least some species that are symbiotic.

The hosts of almost all of the symbiotic nonneodermatans are invertebrates; molluscs, arthropods, and echinoderms figure prominently in this role. The neodermatans are unlike almost all of the other symbiotic platyhelminths in that they are generally associated with at least one vertebrate host group at some point in their lives. The aspidogastreans are somewhat exceptional in that some species are known only from freshwater clams and snails; however, records of aspidogastreans from the digestive system of turtles, elasmobranchs, and teleosts are not uncommon. Although uncharacteristic of the group as a whole, some monogeneans are found associated with cephalopods, and others are found associated with crustaceans that parasitize teleosts or elasmobranchs.

The neodermatan groups are also unique in that their life cycles generally involve more than a single host species. All digeneans possess at least two different hosts in their life cycles. In general, the first (intermediate) host is a mollusc, while the final (definitive) host is a vertebrate. In species possessing life cycles involving more than two hosts, there is one (or sometimes more) additional intermediate (or paratenic) host that is either an invertebrate or a vertebrate depending on the digen-
ean group. In such taxa the host that precedes the vertebrate is generally an invertebrate, often some sort of mollusc or arthropod. The monogeneans are exceptional in this respect as their life cycles involve only a single host.

Platyhelminths are found either in or on their hosts. The nonneodermatan groups are often found on, rather than in, their hosts. The exceptions are the symbiotic acoels, nemertodermatidans, and dalyelliidans, which are most commonly found inhabiting the coelom or digestive tract of their hosts. Among the neodermatans, the monogeneans are unique in that they are generally found on, rather than in, their hosts; the remaining neodermatan groups are most commonly found as endoparasites of their respective hosts.

X. MEDICAL IMPORTANCE

The medical and economical importance of the various platyhelminth groups is closely tied to the type and intimacy of the host associations of the various groups. Species that use vertebrates as hosts are obviously of much greater concern than those that do not. Thus, the neodermatans are by far the most medically and economically important platyhelminth groups. Space limitations do not permit discussion of the economic consequences of infections with neodermatans, but some of the medically important species are briefly discussed below.

Within the neodermatans only the digeneans and cestodes include species that infect humans. Among the 17 superfamilies of digeneans, human parasites are essentially restricted to four. In the majority of these species it is the adult stage that parasitizes humans. Many of these species are zoonotic; that is, their normal hosts are usually vertebrates other than humans. Humans acquire infections with such species when they unwittingly ingest the metacercarial stage. These zoonotic groups infect a variety of different organs in their human hosts. They include a total of eight species of lung digeneans in the superfamily Troglocrematoidae, five species of liver digeneans in the superfamily Opisthorchoidea, and seven species in the superfamilies Echinostomatoidea, Opisthorchioidae, and Troglocrematoidae that inhabit the digestive tract of their hosts. Finally, human infection with cercariae of species of the superfAMILY Schistosomatoidea that normally infect birds and aquatic mammals is responsible for an irritating condition in humans known as cercarial dermatitis or swimmer’s itch.

Only six species of digeneans parasitize humans as their normal final (definitive) hosts. These include one species of liver fluke in the superfamily Opisthorchoidea and five species of blood flukes in the superfamily Schistosomatoidea. The unusual life-cycle of the schistosomes requires that humans come into contact with the free-swimming cercariae of these species to acquire an infection. Schistosomes, which cause a disease known as schistosomiasis in humans, are without question the most deadly of the digeneans. Hundreds of millions of people globally are infected with these species. Schistosomes are responsible for the deaths of tens of thousands of people each year.

Species that parasitize humans belong to only two of the 14 orders of tapeworms. The adults of at least four species of pseudophyllideans (all of which are members of the genus Diphyllobothrium) have been reported from the digestive system of humans; evidence suggests that all of these species are zoonotic in humans, as they normally parasitize other fish-eating vertebrates as adults. Plerocercoids of some species of pseudophyllideans of the genus Spirometra are known to parasitize the musculature and subcutaneous tissues of humans, causing a condition known as sparganosis. By far the most pathogenic of the human tapeworms, however, are the cyclophyllideans. Nine species in five genera parasitize humans. These species are known from humans as either larval or adult stages. Larval stages are generally much more pathogenic than adult stages. The hydatid cysts, coeneri, and cysticerci of Echinococcus, Multiceps, and Taenia species, respectively, are arguably the most life-threatening of the human-infecting cestodes. These larval stages often develop in difficult-to-treat extraintestinal sites, such as the brain, or grow extremely large and/or invasively into one or more organs of the body, making surgical removal of the parasite difficult. Thus, infection with any of these larval stages can be fatal. Infection with an adult tapeworm is generally debilitating rather than life-threatening, because these stages remain in the digestive system and thus are relatively easy to treat. With the exception of Taenia solium, it appears that tapeworm infections in humans are also all zoonotic. Infection with adult tapeworms can be prevented by properly cooking food. Infection with larval stages usually involves contact with fecal matter from humans or other animals.

Abbreviations: A, adhesive disc; AC, acetabulum; AD, adhesive glands; AG, accessory gland cells; AN, anchor; AO, accessory male organ; AP, adhesive papillae; APO, apical organ; AR, arm; AS, apical sucker; B, brain; BA, basal lamina; BDY, body; BF, buccal funnel;
BL, bladder; BO, bothrium; BOD, bothridium; BP, base plate; BR, brood sac; C, cirrus; CB, copulatory bulb; CE, cercomer; CGP, common genital pore; CL, clamp; CM, circular muscle; CO, copulatory organ; CP, ciliated pits; CPE, cephalic peduncle; CR, ciliary ring; CS, cirrus-sac; CT, ciliary tuft; CY, cyton; CYC, cyclocoel; CYE, cytoplasmic extension; DC, developing cercaria; DCY, distal cytoplasm; DE, developing embryo; DH, daughter hydatid cyst; DR, developing reida; EC, ecorsom; ED, excretory duct; EDC, electron dense cap; EM, embryophore; EP, epidermal cell; ES, esophagus; ESV, external seminal vesicle; EV, excretory vesicle; EXP, excretory pore; F, forebody; FI, filament; FO, frontal organ; FP, female genital pore; FVP, female ventral papilla; GC, genitointestinal canal; GCO, genital corona; GD, gut diverticula; GE, germ cells; GEL, genermal layer; GL, gland cells; GP, gravid proglottid; GV, germovitellarium; GY, gyrocynocyl central; H, hook; HA, haptor appendix; HB, hindbody; HE, hexacanth embryo; HG, hermaphroditic gland; HO, hooklet; HOR, head organ; HP, haptor; I, intestine; IMP, immature proglottid; IE, inner envelope; IS, internal seminal vesicle; L, loculus; LD, lateral diverticula of oral sucker; LL, laminated layer; LM, longitudinal muscle; M, mouth; MA, muscular appendage; MB, muscular bul; MP, male genital pore; MP, mature proglottid; MT, microthrix; MUP, muscular pad; MVP, male ventral papilla; N, nucleus; NE, neoderms; NK, neck; O, ovary; OC, ocellus; OE, outer envelope; OGO, orogenital opening; OS, oral sucker; P, penis; PA, pars apicalis; PAP, papilla; PB, pars basal; PC, parenchymal cell; PEG, penetration glands; PG, proboscis glands; PH, pharynx; PI, posterior isthmus; PN, protonephridium; PR, proboscis; PRG, proglottid; PRO, prostatic reservoir; PRT, protoscolex; PS, pseudosucker; PSC, primary scolex; PST, primary strobila; PV, prostatic vesicle; R, rosette; RO, rostellarium; RT, rhabdoid tracts; S, sucker; SC, statocyst; SCL, sclerite; SCO, scolex; SET, setae; SH, spiny head collar; SL, statolith; SR, seminal receptacle; SSC, secondary scolex; ST, strobila; SV, seminal vesicle; SY, syncytium; T, testis; TA, tail; TB, transverse bar; TE, tentacle; TL, terminal lappet; TO, tribocytic organ; TP, transverse plications; TS, tentacle sheath; TU, tubercles; U, uterus; Ug, uterine gland; UP, uterine pore; V, vitellarium; VA, vagina.

See Also the Following Articles
PARASITISM • WORMS, ANNELIDA • WORMS, NEMATODA

Bibliography


I. The Changing Role and Façade of Zoos
II. Choosing Animals to Promote Biodiversity Conservation
III. Keeping Biodiversity
IV. Supporting Diversity in Nature

GLOSSARY

exotic animals Animals that have not been domesticated for use by humans in agriculture or as pets; also connotes that the wild animals are from another region of the world.
ex situ Outside the natural range. Hence, a captive breeding program in Europe for a South American species would be an ex situ program. (Conversely, a Chinese program to breed native panda bears in captivity is an in situ breeding program because it is within the natural range.)
founders Animals that were collected from their natural range and that have produced progeny in captivity (and are assumed to be unrelated). Thus, these animals are the founding stocks of the captive population. Also generalized to include animals with wild parents that are collected in order to supplement the genetic diversity of an established captive population.
International Species Information System (ISIS) A central repository for zoo animal records founded in 1973; has also developed the computer programs that most zoos use for keeping their animal records, as well as software used by zoos for studbooks. Studbooks Also called genealogies, registries, or pedigrees, a database of historic and living animals intended to go beyond passively recording bloodlines and records, typically kept in computer databases that are also useful for actively managing the genetics, demography, and husbandry of captive populations.

ZOOS is a handy three-letter word, and thus in this article “zoos” shall be used in a very general way to include any public or private institutions that house live exotic animals, largely for the purposes of exhibition to the public. Zoos are permanent establishments that are open to the public to provide education, recreation, and cultural enjoyment. Zoos, for purposes of this discussion, are not transient shows such as circuses or petting zoos. Nor are zoos private collections of exotic animals. “Zoo,” defined thus, include the facilities that are traditionally considered zoos, such as zoological parks and aquariums, and also many wildlife centers and “bioparks,” as well as some insect houses and ex situ breeding centers. There are currently over 1,200 well-established zoos worldwide.

Zoo reach a vast audience; for example, North American zoos attract more attendees than the major
professional sports combined. The best modern zoo exhibits are sublime, invoking in visitors a sense of awe and respect for nature. Zoos are increasingly operated by specialized, professional staff in partnership with government and private interests, with the express goal of conserving wildlife. The main focus of this article is on how zoos are evolving into institutions for wildlife conservation, and also on the art and science of managing biodiversity in zoos.

I. THE CHANGING ROLE AND FACADE OF ZOOS

A. Zoos from Ancient to Recent Times

Zoos probably arose as places of spectacle and entertainment shortly after civilization became urbanized. The first recorded zoological collection dates to 2,500 B.C., when several thousand wild animals are documented in Saqqara, Egypt. Ancient Egyptians kept many wild species for use in religious ceremonies, or perhaps because they hoped to tame them. These ancient animal collections were acquired as tribute gifts or via expeditions that were mounted specifically to collect wildlife. Zoo collections have sprung up independently on many continents and in many civilizations. For example, Chinese emperors as far back as 1150 B.C. had walled animal collections they called “parks of knowledge.” When Cortés reached Mexico City in 1519, he witnessed, and then destroyed, the Aztec ruler Montezuma’s immense zoo and gardens. Montezuma’s zoo was so large that 300 keepers were employed for the avatars alone, and it even housed malformed humans as part of the collection. Most of the zoological collections of antiquity were imperial and exclusive—symbols of wealth and power. They were menageries in the sense that they were collections of living trophies to be managed.

Collections during Roman times were also for the amusement of citizens. Large numbers of wild animals were used for blood sport and spectacles; holding areas that also allowed public viewing, called vivaria, were associated with arenas. After the fall of Rome, collections of wildlife were not so common in Europe. During Europe’s Middle Ages, imperial menageries once again became more significant. Wild animals were used as diplomatic gestures—as tributes and ceremonial gifts.

The roots of public zoos are somewhat contemporary with the emergence of Europe’s great cities. London gained a zoo, of sorts, in 1232 when Henry III transferred his menagerie to the Tower of London; London’s citizens were pressed to support the menagerie, and in turn the public could view some of the animals, such as a polar bear and an elephant. Urban menageries, under the sponsorship of merchants and other citizens, began to emerge, first in Frankfurt in the late 1300s, then in the 1300s at The Hague and in Augsburg.

The rising profile of zoos can be traced to the Renaissance, when trade and exploration brought more animals into Europe. Colonial officers in far-flung empires served as conduits for animal requests. Public interest was most piqued by the wildlife of Africa and the Far East. The first zoological garden, which combined displays of plants and animals, was established in Versailles by Louis XIV in 1665; in 1794 this collection became a division of Paris’ natural history museum, creating a formal link to zoological research that was later emulated by many other zoos and museums. The first “modern” zoo that survives to this day is probably the Schönbrunn Zoo in Vienna, originally established as a private collection in the 1770s, but to which the public were intermittently admitted.

The 19th century has been hailed as the century of science and the golden age of museums. This century witnessed an explosive increase in the numbers of zoos and museums. Scientific societies sprang up, including zoological societies. Zoological research was nominally based in 19th century zoo, but the inquiry was mostly limited to collecting and classifying. Local governments tended to operate zoos of this era with the aim of providing recreation and edification for the increasingly better educated citizenry that arose with the Industrial Revolution. Zoos were sources of civic pride, and they competed to exhibit as many kinds of animals as possible. The rise of print media made it possible for civic leaders to compare collections more readily with rival zoos. Natural social groupings were largely ignored, and frequently solitary animals were displayed. Generally species were grouped to correspond with scientific classification, for instance, into bird, reptile, or carnivore houses. These animal collections were viewed as analogous to collections of art, and even the zoo architecture was considered art.

B. Zoos in the Mid-20th Century

A major shift occurred in the mid-20th century when zoos began to emerge, first in Frankfurt in the late 1300s, then in the 1300s at The Hague and in Augsburg. The rising profile of zoos can be traced to the Renaissance, when trade and exploration brought more animals into Europe. Colonial officers in far-flung empires served as conduits for animal requests. Public interest was most piqued by the wildlife of Africa and the Far East. The first zoological garden, which combined displays of plants and animals, was established in Versailles by Louis XIV in 1665; in 1794 this collection became a division of Paris’ natural history museum, creating a formal link to zoological research that was later emulated by many other zoos and museums. The first “modern” zoo that survives to this day is probably the Schönbrunn Zoo in Vienna, originally established as a private collection in the 1770s, but to which the public were intermittently admitted.

The 19th century has been hailed as the century of science and the golden age of museums. This century witnessed an explosive increase in the numbers of zoos and museums. Scientific societies sprang up, including zoological societies. Zoological research was nominally based in 19th century zoo, but the inquiry was mostly limited to collecting and classifying. Local governments tended to operate zoos of this era with the aim of providing recreation and edification for the increasingly better educated citizenry that arose with the Industrial Revolution. Zoos were sources of civic pride, and they competed to exhibit as many kinds of animals as possible. The rise of print media made it possible for civic leaders to compare collections more readily with rival zoos. Natural social groupings were largely ignored, and frequently solitary animals were displayed. Generally species were grouped to correspond with scientific classification, for instance, into bird, reptile, or carnivore houses. These animal collections were viewed as analogous to collections of art, and even the zoo architecture was considered art.

An innovation at the beginning of the 20th century was habitat exhibits, as pioneered by Carl Hagenbeck at Tierpark, outside Hamburg, Germany. In these zoos of living panoramas, it became increasing common to display animals along geographic lines rather than strictly taxonomic ones. Habitat exhibits were well suited to zoo designs that intended to teach ecological principles, and thus led to the display of animals in

I. THE CHANGING ROLE AND FACADE OF ZOOS

A. Zoos from Ancient to Recent Times

Zoos probably arose as places of spectacle and entertainment shortly after civilization became urbanized. The first recorded zoological collection dates to 2,500 B.C., when several thousand wild animals are documented in Saqqara, Egypt. Ancient Egyptians kept many wild species for use in religious ceremonies, or perhaps because they hoped to tame them. These ancient animal collections were acquired as tribute gifts or via expeditions that were mounted specifically to collect wildlife. Zoo collections have sprung up independently on many continents and in many civilizations. For example, Chinese emperors as far back as 1150 B.C. had walled animal collections they called “parks of knowledge.” When Cortés reached Mexico City in 1519, he witnessed, and then destroyed, the Aztec ruler Montezuma’s immense zoo and gardens. Montezuma’s zoo was so large that 300 keepers were employed for the avatars alone, and it even housed malformed humans as part of the collection. Most of the zoological collections of antiquity were imperial and exclusive—symbols of wealth and power. They were menageries in the sense that they were collections of living trophies to be managed.

Collections during Roman times were also for the amusement of citizens. Large numbers of wild animals were used for blood sport and spectacles; holding areas that also allowed public viewing, called vivaria, were associated with arenas. After the fall of Rome, collections of wildlife were not so common in Europe. During Europe’s Middle Ages, imperial menageries once again became more significant. Wild animals were used as diplomatic gestures—as tributes and ceremonial gifts.

The roots of public zoos are somewhat contemporary with the emergence of Europe’s great cities. London gained a zoo, of sorts, in 1232 when Henry III transferred his menagerie to the Tower of London; London’s citizens were pressed to support the menagerie, and in turn the public could view some of the animals, such as a polar bear and an elephant. Urban menageries, under the sponsorship of merchants and other citizens, began to emerge, first in Frankfurt in the late 1300s, then in the 1300s at The Hague and in Augsburg.

The rising profile of zoos can be traced to the Renaissance, when trade and exploration brought more animals into Europe. Colonial officers in far-flung empires served as conduits for animal requests. Public interest was most piqued by the wildlife of Africa and the Far East. The first zoological garden, which combined displays of plants and animals, was established in Versailles by Louis XIV in 1665; in 1794 this collection became a division of Paris’ natural history museum, creating a formal link to zoological research that was later emulated by many other zoos and museums. The first “modern” zoo that survives to this day is probably the Schönbrunn Zoo in Vienna, originally established as a private collection in the 1770s, but to which the public were intermittently admitted.

The 19th century has been hailed as the century of science and the golden age of museums. This century witnessed an explosive increase in the numbers of zoos and museums. Scientific societies sprang up, including zoological societies. Zoological research was nominally based in 19th century zoo, but the inquiry was mostly limited to collecting and classifying. Local governments tended to operate zoos of this era with the aim of providing recreation and edification for the increasingly better educated citizenry that arose with the Industrial Revolution. Zoos were sources of civic pride, and they competed to exhibit as many kinds of animals as possible. The rise of print media made it possible for civic leaders to compare collections more readily with rival zoos. Natural social groupings were largely ignored, and frequently solitary animals were displayed. Generally species were grouped to correspond with scientific classification, for instance, into bird, reptile, or carnivore houses. These animal collections were viewed as analogous to collections of art, and even the zoo architecture was considered art.

An innovation at the beginning of the 20th century was habitat exhibits, as pioneered by Carl Hagenbeck at Tierpark, outside Hamburg, Germany. In these zoos of living panoramas, it became increasing common to display animals along geographic lines rather than strictly taxonomic ones. Habitat exhibits were well suited to zoo designs that intended to teach ecological principles, and thus led to the display of animals in...
more natural social groupings. Eventually mixed-species exhibits became more common, and these conglomerations might even be grouped by ecological biomes such as "polar zone" or "tropical zone," rather than along strictly geographical or taxonomic lines.

Later in the 20th century, world culture became increasingly electronic and graphic. Zoos found themselves in competition with wildlife film documentaries. No longer would the animal displayed in a sterile cage suffice for an increasingly sophisticated audience. "Immersions" exhibits became the ideal for most new zoo development. Immersion exhibits are ones that seek to make visitors feel that they are in the habitat with the animals. This may be accomplished by incorporating the same details of rocks, logs, or plants into both the animal and the public space, while using these same details to hide any cage elements, and also by special effects such as mist machines and animal call background soundtracks that help create an illusion. Barriers to separate animals and people may be relatively invisible (e.g., glass) or carefully placed so they are unobtrusive. Special feeders and techniques of animal management may be used to encourage animals to be active and visible, for example, hidden heating pads for animals to lounge upon or feeders that intermittently spray out small amounts of food near a viewing station to encourage visible foraging behavior. Zoos often try to craft the visitor space so that visitors do not see other visitors while observing the animals. In many instances, the designers' intent is to invoke an almost religious experience of awe and respect for nature, much as the grand cathedrals did in early eras.

B. Rise of the Zoo Conservation Ethic

Zoos, as we have seen, arose as places of spectacle, not as conservation organizations. But many modern zoos have chosen wildlife conservation as their primary mission. The conservation ethic has developed in zoos over the last century in parallel with the non-zoo world, and in many cases one can point to zoos as the places that incubated the people and ideas that have led to important conservation developments. The National Zoo in the United States, founded in 1887, is considered to be the first zoo created to preserve endangered species.

Conservation of wildlife and wildlife habitat has become the major goal of most of the prominent zoos and zoo professional organizations. Although altruism is part of the reason for the shift, compelling utilitarian reasons also drive these lofty goals. The supply of wild animals to stock zoos is drying up. As humanity expands and wild places contract, exotic animals are getting scarcer. As regulators seek to simultaneously protect their natural resources and their agricultural sectors, governments make it ever more difficult to import and export wildlife. Wild animals may carry diseases that will harm livestock and perhaps even spread to the human populace; increasing understanding of disease has led to import restrictions on quarantine. For example, the bird-importing era, at least in the United States, slowed markedly in the early 1970s, when quarantine laws were enacted to guard against Newcastle's disease. In the 1990s the Wild Bird Trade Act was passed in the United States to restrict importation of any birds protected by CITES, the Convention on International Trade in Endangered Species of Wild Fauna and Flora. Most professional animal dealers have gone out of business due to the increasing restrictions, and zoos cannot easily collect wild animals.

Simultaneously, zoos have needed to import fewer animals because they became more proficient at keeping and breeding wildlife. Zoo keeping is a complicated art and science, and zoo keepers and managers have found it useful to pool their knowledge. Furthermore, single zoos are rather like islands from a population biology point of view, in that they are not big enough to sustain viable populations. So cooperative breeding programs for animals have become key, while cheap, fast transport, usually by air, makes it feasible for distant zoos to exchange breeding stocks. In many regions of the world, the leading zoos and their staff now belong to professional organizations. There are also world zoo organizations that help link the regions. These organizations have codes of ethics and accreditation (certification) programs. For example, North America's American Zoo and Aquarium Association was founded in 1924 and now includes nearly 200 institutional members (this is about one-tenth of the number of animal exhibitors that are licensed with the U.S. government).

Public support is yet another motive for many zoos to become conservation organizations. Zoos are having difficulties in the competition for public funding. Modern zoos are costly, with expensive professional staffs, sophisticated veterinary medicine facilities, and a public that demands complex, educational, and artistic exhibits. Municipal governments are increasingly burdened by social programs and other competing demands for funds and now only fund a fraction of zoo operations compared to earlier times (on average 21% of the zoo budget in North America). But while the funding situation is tough in the richer nations, it is often quite desperate in poorer regions of the world. Most zoo...
based conservation is being done by the "haves" while much of the need for conservation and access to endan-
gered species is in the "have-not," largely tropical-ly located zoos of the world.

There is enormous public support for environmental causes and grants available for conservation programs. However, zoos face some challenges in getting or gener-
ating environmental funds, in part because of public ambivalence about the ethics of keeping animals in zoos.

C. Animal Welfare in Zoos

Advocates for animals are becoming increasingly stri-
dent, even to the extent of using terrorist tactics. Their ire is more typically directed against laboratories that use animals, but zoos also have come under fire. Some animal rights extremists have expressed a goal of put-
ting zoos out of existence. Extreme though these views may be, they signal a shift in public attitudes towards animal welfare. Many people now consider animals as fellow creatures with needs and feelings that should be respected. Zoos that have not become attuned to perceptions of animal welfare have often experienced an erosion of public support. Some of the poorer, old-
-fashioned zoos have closed, and even some prominent zoos have almost succumbed.

These animal advocates often confuse animal rights with animal welfare, and their anthropomorphic ideas about what animals need are often at odds with what is best for animals. For example, well-meaning protestors tried to prevent the transfer of a gorilla named Timmy, who had never procreated, to a zoo in New York that had considerable success in breeding gorillas as well as many other animal species. The protectors did not want Timmy to be separated from his current, infertile, female partner. However, in nature gorillas do not form monogamous bonds and they do change partners. Eventually, after court battles, Timmy was moved, formed bonds with new females, and has begat numerous offspring.

A related problem that we shall return to in Section III is the problem of limited space in zoos. Thus, breed-
ing unwanted, "surplus" animals is a problem for zoos because surplus animals inhabit valuable space. Some zoos took the apparently logical stance, from a utilitar-
ian animal management vantage, of culling their surplus stocks or selling them to the private collectors. In pri-
ivate hands, some animals have ended up on game ranches where they are shot for sport. This has caused major public relations imbroglios, as a result of which zoos have had to examine their ethical policies and restrict their animal trade to within accredited zoos. Zoos have developed contraceptive and husbandry techniques to prevent unplanned breeding, and they have developed scientific breeding plans to keep popu-
lation demographics aligned with available space. Fur-
thermore, there has been a trend of exchanging animals without fees as they become increasingly priceless. Improved contraception for animals is just one ex-
ample of how health care for exotic animals is becoming increasingly sophisticated. Improved veterinary care de-
serves much of the credit for allowing zoos to shift from relatively sterile concrete and steel cages to more natural housing, which is often, literally, dirty (as the substrate is frequently dirt or other organic matter). Most zoos now have their own veterinarian and veterinary hospi-
tal. Some even have pathologists, endocrinologists, and nutritionists providing preventative care, better drugs, and diagnostics. The understanding of wild animals' nutritional needs is advancing rapidly, and there are now numerous commercial brands of food for exotic animals, such as "Flamingo Fare." Nutritionists are even improving on fish diets (which are expensive and be-
come depleted in certain vitamins during shipping and storage) for animals such as penguins and sea lions by devising a gelled, artificial fish in a fish shape. Most zoos have "browse" programs to grow and harvest fresh plant foods for their denizens.

There are numerous technical periodicals, books, and other publications related to zoo husbandry, in-
cluding "Zoo Biology," "International Zoo News," the "International Zoo Yearbook," "The Shape of Enrich-
ment," and "The Dodo." As a testament to the improving welfare of zoo animals, zoos are now recording many longevity records, and zoo animals frequently senesce and die of old age—an almost unheard of occurrence in nature. As zoos have become more successful at keeping exotic animals alive and meeting their physical needs, the attention of zoo biologists has begun to shift toward behavior.

Zoos have also added applied psychology to their toolkits. They now must reassure their public that ani-
mals are not distressed but are potentially "happy." New techniques are being developed to stimulate zoo animals, reduce psychological stress, and elicit coopera-
tion rather than coerce desired behaviors. Some zoos train animals to perform in shows, which is not a new phenomenon. But modern shows demonstrate natural behaviors to the public and keep the performing animals occupied, rather than emphasizing cute tricks or an animal's fierce nature. These new techniques range from
behavior management and operand conditioning to environmental enrichment. Behavioral techniques aim to banish boredom, eliminate stereotypic tics, and otherwise encourage behavioral patterns that approximate those of animals in nature.

D. Emerging Trends

Zoos, as we have seen, have evolved from imperial menageries to the modern habitat and immersion exhibits that featured animals in more natural settings and social groupings. In the 21st century, we expect this trend toward more ecological themes to continue, with more zoos featuring animals in diverse, mixed-species exhibits. Zoos that predominantly have these diversity-designed exhibits have been given the moniker "biodome," or "biopark." This sort of zoo may feature botanical collections housed with terrestrial and aquatic vertebrates and invertebrates. There is a considerable art in developing these mixed species exhibits, because zoos must ensure that co-housed animals do not decimate the foliage or each other. Animals that do wreak havoc with plants or other animals are either excluded from the biopark or tend to be isolated, although their cage boundaries may be crafted carefully so that this is not obvious. Some zoos are extending this idea to focus on the ecology of a certain region using both zoo exhibits and more museum-like exhibits; typically the focus may be the region in which they are located, for example, the "living desert museum" in the American southwest near Tucson, Arizona.

Also emerging at the turn of the 21st century seems to be a new trend, which we shall call "conservation education." These thematic exhibits have a storyline to which visitors may be introduced, sometimes through videos or costumed actors, as they wait to enter the exhibit. This visit may even include an amusement-style ride, for example, Sea World visitors to the polar region exhibit experienced a "tundra buggy" simulated ride before viewing polar bears and other creatures. Some call this the "Disneyfication" of zoos, alter a new zoo that the Walt Disney Company opened in Florida in the late 1990s; one feature is a drive-through, safari-style bus ride, where the storyline is that visitors are joining a patrol to catch poachers.

The actual educational value of these new style zoo exhibits is debatable, and a still greater unknown is whether the lessons learned by visitors will translate into conservation actions. Conservation education, zoos have argued, is one of their major contributions, and one that is becoming increasingly important as citizens become further distanced from nature. Zoos are trying to justify these claims, and also to do a better job of conservation education. Thus, zoos are beginning to scrutinize their exhibits and programs using evaluation techniques. Evaluation methods, including surveys, interviews, or observation of visitors, are sometimes employed during the design phase to improve the message before expensive construction begins. They may also be used at other times to help fine-tune exhibits or otherwise inform about how the zoo is working or what the audience is like. The science of evaluation is rapidly developing in zoos and museums, and it will probably become an increasingly important part of zoo planning.

As the world becomes more wired to the Internet, zoos are following suit. Many zoos now have Web sites where visitors can find information about the zoo. Electronic mail is becoming a valuable communication tool for sharing the specialized knowledge that zoo keeping requires. Some zoo databases, including the vast ISIS (International Species Information Service) registry of zoo animals, are now available at Internet Web sites. Some zoos are incorporating electronic media into their educational efforts, either integrated with exhibit graphics or separately. Cooperative breeding plans, and the data needed to manage them, will increasingly be shared via electronic means. And some Web site programs, for example, one for trumpeter swans, allow school children and others to dynamically follow the movements of reintroduced animals via satellite tracking. Virtual zoos will not keep real animals, which is the primary business of zoos, so zoos will probably continue to find ways to use the Internet that enhance rather than radically change what they do.

As the supply of wild-born founders dwindles, and more amenities are expected for the animals in hand, zoos are abandoning the one-of-each-kind collecting approach. Within-zoo species diversity is decreasing in order to make room for zoos to house a breeding nucleus of at least several pairs and their young offspring. Many zoos are trying to allocate a proportion of their animal space for off-exhibit breeding efforts, and some are operating special facilities devoted to breeding of endangered species—mostly ex situ but also some in situ. Many zoos are developing specializations, and concentrating their efforts on the specific groups of animals that they can keep best and exhibit most dramatically. Zoo workers are also becoming more specialized, both along taxonomic lines and also professionally into fields such as educators, veterinarians, nutritionists, population biologists, behaviorists, life-support system opera-
tors, pathologists, and horticulturists. With the increasing need for cooperation, more formal committees and other structures are developing through the professional organizations.

II. CHOOSING ANIMALS TO PROMOTE BIODIVERSITY CONSERVATION

A. Space, the Final Limit

Modern zoos are expensive to build and upgrade. In North America alone, zoos spend around two million dollars per year on capital or physical improvements, with much larger expenditures for major new exhibits or zoo facilities. A major exhibit can cost in the tens of millions of dollars and typically takes at least three years to plan and build. Maintaining a zoo and its animals is also very expensive. Therefore, there are practical limits to how many zoos can be supported and thus to how many animals can be maintained by zoos. For example, the number of animals that are presently kept in the 181 accredited zoos of North America is approximately 800,000.

Thus, conservation-oriented zoos choose their inhabitants with care in order to maximize the biodiversity that they preserve in their finite space. A biodiversity trade-off must be made between the number of species that are maintained by zoos versus the genetic diversity or size of the captive populations that are maintained. Zoo planners have to think long-term because they are trying to maintain viable populations for decades or centuries. In many cases, zoo populations are, essentially, closed gene pools because it is difficult or impossible to acquire fresh bloodlines or founders.

Extinction in nature is the most obvious cause for a closed captive population, and zoos work with several “extinct” species such as the Prezwalski’s horse. But species that are extant in nature can also be effectively isolated in zoos by the genetic diversity or size of the captive populations that are maintained. Zoo planners have to think long-term because they are trying to maintain viable populations for decades or centuries. In many cases, zoo populations are, essentially, closed gene pools because it is difficult or impossible to acquire fresh bloodlines or founders.

In order to prevent domestication of their wild animals, or, in other words, they seek to freeze evolution or at least retard inbreeding. Arguably, this is a prudent and responsible strategy should the animals be needed someday for restocking wild habitats.

Hence, zoo population plans usually strive to breed animals better than would be done randomly to retard natural selection, and they try to do this using as few animals as possible. This usually requires a captive population of several hundred animals. In this effort, attention must often be paid to picking the right unit for conservation, or the evolutionarily significant unit. Zoos have funded a considerable amount of applied molecular genetic research in order to determine subspecies and species characteristics, so that they are breeding animals that should interbreed, and so that they are not maintaining subviable numbers of several subspecies or species rather than a viable population of one.

B. Metamanagement of Zoo Biodiversity

Each zoo is an island too small to do much conservation breeding on its own. To facilitate cooperation, organized conservation programs involving captive animals are beginning to develop around the world through regional zoo professional associations. Most of the labor that goes into these programs is pro bono or volunteer. Much of the work is accomplished through workshop meetings and via committee, at considerable personal and institutional expense.

The basic unit of these programs is the studbook, which is the database of historic and living animals in a particular captive population. Studbooks are usually maintained through the efforts of a “studbook keeper.” Populations that are monitored with studbooks can be managed through regional or even global captive breeding plans, usually overseen by a committee, which specifies where each animal should be and whether it should be bred.

In order to prioritize how much captive space various populations should be allocated, zoo specialist groups for each kind of animals have formed, and are known as “taxon (for taxonomic) advisory groups.” Taxon advisory groups typically work closely with the appropriate taxonomic and scientific specialist groups of the World Conservation Union (IUCN), as well as with scientific advisors. Regional taxon advisory groups tend to develop priorities that are complementary to those in other areas; for example, Europe may choose to allocate space to breed a different species of penguin than North America. Taxon advisory group priorities are formalized into “regional collection plans,” which ideally include a justification for including or excluding each species or subspecies in the taxonomic group, as well as setting a target population size for each managed population. These plans may also specify areas where research is needed, or where field conservation should be a priority.

Tools are being developed to facilitate regional planning, including meta-collection software that has been developed by Australian zoos.
The representation of biodiversity in zoos is biased toward animals that are furry, large, brightly colored, or otherwise captivating for people. In general, the more closely related to humans, the greater the representation. So, most kinds of primates are kept in zoos and managed through cooperative programs, whereas only a fraction of rodents or bats are represented. Similarly, most kinds of flamingos can be found in captivity, whereas only a fraction of the small perching birds are kept. A smaller proportion of birds are maintained in organized programs than mammals, but only a fraction of the scaly creatures (fish, reptile, and amphibian) are kept. Very few invertebrates are represented in captivity, but this seems to be changing as butterfly exhibits and insect displays are gaining in popularity. Invertebrate husbandry is relatively primitive for both aquatic and terrestrial species; the public generally considers the collection of wild invertebrates acceptable so there has been less pressure to develop husbandry techniques. The situation for fish, and most other nonmammalian aquarium denizens, is similar to that of invertebrates. Much work is needed to develop breeding programs for aquariums and invertebrate collections.

C. Institutional Planning

Zoos as individual institutions are also becoming very mission-oriented. Many zoos have developed mission statements and master plans for growth and development. These plans provide the overall organization of the zoo exhibits and the visitor traffic patterns as well as locations of facilities for merchandise, refreshments, and bathrooms. (Bathroom location is amazingly important for a positive visitor experience!) Master plans usually designate the kind of educational messages the zoo will try to convey and the themes of the exhibits. They may be informed by practical considerations of what animals and plants will do best in the zoo, and these may be based on local climate or the experience and expertise of the zoo’s staff.

Within the context of the master plan, zoo planners, typically curators, will develop an animal collection plan. Ideally, curators will develop these plans after studying the regional collection plans that have been developed for metamanagement of zoo collections. This collection plan will usually list each species that the zoo keeps or wishes to obtain, along with a justification for what the species will contribute to the zoo and to conservation of the species. The plan should say how many individuals there are of each sex, and how many are desired. The plan will specify where these animals will be housed. If the species is managed by a cooperative breeding plan, then the institutional collection plan should include the recommendations for which animals are to be bred and which not, if not part of a cooperative plan, the curator should have a breeding and possibly a research plan of their own. Institutional collection plans usually need to be updated annually.

III. KEEPING BIODIVERSITY

A. Zoo Animal Records Keeping

Basic to managing animals and improving husbandry is keeping track of the inventory and gathering data. The memory of zoo staff is no longer sufficient to keep these records, as animals live longer and breed better, the government expects more information, and both animals and keepers are more mobile. Zoos are tracking an ever-increasing amount of information about their animals. Each animal in the inventory, current and historic, is accessioned, which means it is assigned a unique identification number. The animal also must be physically identifiable, so most are tagged or marked; this can be done with traditional bands and tags or with more high-tech injectable microchips. Records are kept for diverse variables, including characteristics that allow an animal to be individually identified, dates of birth and death and other life-stage transitions, sex, parents, offspring, current and previous locations, health notes, cage mates, previous owners, physical parameters such as measurements, species, and subspecies, and behavioral observations. Zoos are standardizing their data collection and records so that all this information can be pooled.

Keeping these records has turned into a specialty of its own, and the keepers of the data are usually called “registrars.” Registrars may also be responsible for some of the government permits and paperwork associated with modern zoo keeping. These data are generally organized into a computer database. The most widely used program is ARKS, the Animal Records Keeping System that was developed by ISIS in 1985. ARKS generates a variety of standard reports. In addition, ARKS is useful for looking up simple questions about individual animals, for example, potential exposure to disease, as well as for mining the data to seek trends or clues to larger questions.

Most zoos send their ARKS data to ISIS at regular intervals so that it is backed-up and also so that parts of the central database can be combined to create an international inventory. This international data is useful for various other uses such as cooperative breeding.
grams are the current standards. The first is a program population management. Graphic and genetic analyses that are essential to zoo computer software that performs much of the demonstr
tows use a similar program called "Zoobook." These studbook programs integrate with computer software. The analyses are used to estimate parameters, genetic and demographic analyses are done with the help of analytical tools that consider the conservationists for taking animals from the wild for exhibition, including charges that this has led to the weakness of ARKS is of little use for managing animals such as schooling fish, many invertebrates, colony-nesting birds, or other animals that are kept in groups with multiple adult males and females. Similarly, animals that breed in relatively large enclosures (such as ducks in large ponds or fish in large tanks) are difficult to keep track of with individual records. Because of this weakness, animal records for aquaria and invertebrate collections are not standardized.

B. Zoo Population Management Strategies

Zoos have historically made much of their “babies,” as births used to be relatively rare and exciting events. But as zoos have become more cooperative, an unfortunate cycle has developed. Rare breeding successes for a particular species tended to be replicated and replicated again. Then, zoo curators would notice that there was a surplus of the species and they could no longer find good homes for the babies. So all zoos would stop breeding the species. Some years later, as the stocks of the animal were aging and starting to die off, there would again be room and demand for babies, and curators would try to resume breeding. But all too often, it would be too late for the aging stocks and the captive population would be lost. To prevent these boom-and-bust breeding cycles, and also to avoid inbreeding and subsequent loss of vigor, zoo biologists have developed population management strategies and tools. Key to zoo population management is gathering information about all the animals in the managed, captive population (the managed population often is not identical to the captive population because some zoos may opt out of the cooperative program and some individual private breeders may participate). The population data are kept in studbooks, which contain genealogical information as well as demographic information. As in zoo records keeping, ISIS distributes the most popular database for studbooks, a program called SPARKS, which stands for single population analysis and records keeping system. Many European studbook keepers use a similar program called "Zoobook." These studbook programs integrate with computer software that performs much of the demographic and genetic analyses that are essential to zoo population management.

Two principle kinds of conservation breeding programs are the current standards. The first is a program that participants essentially sign a contract to join, and agree to abide by the decisions of an elected committee. This type of program was first introduced in the United States, and was named the species survival plan or SSP; nearly identical programs are known by other names in other regions of the world. The second kind of program is a less intensive one, in which recommendations are issued but individual zoos elect whether or not to cooperate. In North America this second kind of program is known as a population management plan or PMP. These two kinds of programs follow similar population management strategies but differ more in the intensity and politics of the actual management. The SSP also has conservation, research, and education components, whereas the PMP is strictly for population management.

Planning for conservation breeding programs generally begins by determining the purpose of the program. Will animals from the population eventually be returned to the wild, or will the primary purpose of the population be for exhibition in zoos? The managers should decide what the appropriate population is for management purposes, for example, whether to limit the management to a subset of the animals that are known to derive from founders of the same subspecies, while excluding those that are of more uncertain ancestry. Once these decisions are made, demographic and genetic goals are set. Genetic and demographic analyses are done using the studbook data and specially designed software. The analyses are used to estimate parameters, which are needed for the long-range population forecasts that are used to guide the goals. Goal setting is done with the help of analytical tools that consider the number of founders (amount of genetic diversity that was likely to be captured), the effective population size N_e, the current population size, the rate of population growth, the number of years since the program was founded, and the generation time. Typically, conservation-breeding programs aim to preserve greater than 90% of the original diversity for at least 100 years. Sometimes it becomes apparent that demographic intervention is needed, perhaps to increase the population growth rate, lengthen the generation time, or increase the number of spaces available so that the population can grow. At other times managers realize that more founders must be imported, if possible from zoos in another region of the world, but sometimes from nature. As an aside, zoos have sustained much criticism from conservationists for taking animals from the wild for exhibition, including charges that this has led to the
depletion of wild stocks. Zoos rebut that the numbers of animals that they take are trivial compared with animals that die through habitat loss, hunting, or other human activities. Good captive management reduces the number of founders, or wild-born breeding stock, needed for long-term population viability. Theoretically, the required number of founders is 25 individuals picked at random from a wild population, but many captive populations start with only a fraction of this number. Even with good management, new founders may need to be added, since not all founders will produce the ideal number of offspring. Founders should be collected with minimal impact on the wild population. The way that this is best accomplished depends upon the natural history of the species and the specifics of the program. For example, one low-impact approach to collecting founders is to salvage orphaned or injured nonreleasable wild animals; this opportunistic approach has the disadvantage that it may be slow and that disabled wildlife can constitute poor breeding stock.

Once the goals of the population are determined, and plans are made to redress problems such as too few founders, a detailed conservation-breeding plan is developed. These plans give a recommendation for each animal in the managed population, which will include whether the animal should be moved, if it should be bred, which animal it should be paired with, and whether the animal should be moved, if it should be bred, and whether any special actions are needed, such as fertility testing.

Scientific principles are given a heavy weight in the planning. Zoo biologists tend to start with demography, using a Leslie matrix approach that incorporates both current and historic data. The demographic software tools were developed at the National Zoological Park by Jon Ballou and Laurie Bingaman-Lackey. Demographic data sets for zoos are often small, but contain highly detailed information compared to wild populations. Therefore zoo demographic estimates are often very crude and may require considerable tweaking. Early in the analysis, zoo biologists look at the population’s age structure and sex ratio. If these are skewed, efforts may be needed to create a better demographic balance. The next consideration is rates of fertility, infant and adult mortality; these estimates are used to project the number of breeding pairs that will be required in order to end up with the desired number of births that will balance deaths and also generate the recommended population growth or shrinkage.

Genetics usually are the primary variables considered in deciding the animal-by-animal breeding priorities, except that aging animals with only a few breeding years left may be given special breeding priority. Genetic algorithms, largely developed by Brookfield Zoo’s Robert Lacy, use pedigree-based genetic simulations to assess the genetics of each animal and potential pairing in the population. Managers attempt to simultaneously maximize three parameters when recommending breeding matches: (1) good matches will improve the genetic diversity of the population (i.e., by increasing underrepresented lineages), (2) good matches will attempt to balance the genetic value of male and female, and (3) good matches will also avoid inbreeding with close relatives. Balancing the genetic value, which is done by looking at a parameter called mean kinship, or the effective number of close kin, is important because it breeds big families with other big ones and little families with other little ones. It is detrimental to mix overrepresented and underrepresented lineages because pedigree webs become increasingly hard to untangle through time, and it is easier to balance representation over time when the lines are kept separate.

Although the science of conservation breeding has advanced considerably, practical considerations also must enter into planning. Wild animal keeping in practice is a complex art, and it would be foolish to ignore details such as the distance of a proposed transfer and behavioral and physiological observations. For example, there is no point in recommending a pairing of two gorillas that are incompatible. Similarly, if an animal has not bred after many attempts at several zoos, there may be a physiological problem or perhaps managers have not yet tried the right trick.

Group breeding species are a special challenge to zoo population biologists as they are for zoo animal records keepers. The methods outlined above work poorly for populations in which a significant proportion of the pedigree is unknown, and also when managers cannot control which individuals mate with which, as is generally the case with group breeding animals. An approach that is sometimes used is to completely ignore the pedigree and manage the species as a subdivided population, with occasional transfers of animals between demes or subpopulations. But the pedigree-free approach often throws out large amounts of information about a population and results in overly conservative management that will require larger populations than might otherwise be needed. Similarly, species whose life history is more size-dependent than age-dependent are not well served by current methods, and may be better managed using stage-based population management. As zoos move beyond management of animals with relatively simple breeding biology, such as tigers,
and tackle species with more complex life histories, new population approaches will be needed. Similarly, as the “brave new world” of reproductive technology develops, frozen zoos will need to become integrated into population planning.

C. Frozen Zoos and Other High-Tech Solutions

An alluring prospect is that someday we may be able to preserve biodiversity in bottles on a warehouse shelf. Another enticing prospect is the use of reproductive technology to propagate animals that have not birthed through conventional techniques. New technologies are making these prospects possible in some cases.

“Assisted reproduction” is a catchall phrase for a variety of veterinary or endocrine interventions used to promote breeding. Artificial insemination, using fresh or frozen sperm, has been in use for many years in the domestic livestock industry and also for some zoo animals. The techniques can be readily adapted to zoo animals if there is a closely related domestic model in which the technique has been developed; for example, rare species of cows can benefit from research with domestic cows. Zoos have also employed other techniques, such as in vitro fertilization with harvested eggs, or embryo transfer both within and between species. The idea of growing a rare species in the womb of a common one is appealing, and has been achieved in some instances.

“Frozen zoos” are being developed by several zoos. In these super-chilled repositories, sperm, eggs, and embryos are committed to long-term storage. Now that cloning is a real possibility, frozen zoos of nonreproductive tissue are also being developed. Saving biodiversity in freezers is tantalizing, but there are many limitations to this approach and hurdles to be overcome. Assisted reproductive techniques are difficult and costly to adapt to animals that do not have a closely related model, and also for those with complex reproductive physiology, for example, animals with induced ovulation. By and large, the techniques have not been cracked for animals with shelled eggs. We cannot, as of yet, grow embryos in a test tube, so those derived from a frozen zoo must be grown in a real womb (or egg). This means a real population of surrogates will be needed to mother the frozen zoo, although in some cases these surrogate mothers can be of a different species. Surrogate mothers may produce animals with species-inappropriate behaviors and possibly even a hybrid phenotype.

Frozen zoos will not preserve learned animal behaviors (culture), and they will not eliminate the need to keep real populations. Frozen zoos, cloning techniques and the like, as San Diego’s Oliver Ryder predicts, will probably become powerful adjuncts to the management of populations. So far, genetic and demographic models for managing zoo populations have not factored in frozen resources, but someday they probably will. This will allow managers to keep smaller numbers of living animals. In addition, relative to the transfer of whole animals, the transfer of sperm, egg, and embryos between wild and captive populations can be done with far less risk of cotransferring disease-causing pathogens. Thus, these techniques may be highly useful in importing new gene lines into small populations, both wild and captive.

IV. SUPPORTING DIVERSITY IN NATURE

A. The Ark Paradigm

Enamored, perhaps, with their newfound ability to breed and manage exotic animals, zoos and their commentators were for a time fixated upon the “ark” paradigm of conservation. That is, like the biblical Noah, they talked about saving species in captivity until some time in the future when the flood of human population shall recede and there will again be the space, and also the will, to restore Nature. Zoos have actually had some success saving species this way. The list of species that have been rescued from extinction by taking refuge for some time in zoos is ever-growing and includes black-footed ferret, California condor, European bison, Mongolian wild horse, Arabian oryx, Partula snails, and Guam rails. However, some species have become extinct during their captive sojourn and no opportunity has arisen to restore some others, for example, the birds of Guam. We are seeing that restoration of a fauna may be only a remote possibility and the utility of reintroduction overstated in some ways.

B. The Post-Ark Paradigm

Maintenance of captive stocks has several functions, but the primary purpose is not, as it is popularly misperceived, for future reintroduction efforts. Zoos may indeed rescue individual animals and sequester them in...
the safety of captivity. As zoo professionals look beyond the ark, they are developing many other ways to contribute to the conservation of exotic animals. Exciting, educational exhibits are potentially zoos' preeminent contribution to conserving animals and their habitats. Worldwide, zoos are estimated to attract over 600 million visitors annually. The majority of these visitors live in urban areas and otherwise have little contact with wildlife. Zoos are beginning to study and develop approaches for mobilizing visitors for conservation through innovative exhibit and educational programs. North America’s accredited zoos report that each year 9 million students visit and enjoy onsite education programs, and in addition 85,000 teachers profit from teacher training. Some zoos are even developing “museum schools” in which a zoo or museum actually serves as an alternative school for a small number of students who typically combine rigorous college preparatory curricula with hands-on or applied learning in the zoo setting (more information about museum schools can be found at http://www.astc.org). Zoo-based conservation education programs need not end at the exit gates. Zoo educators and graphic artists can also develop conservation curricula that can be used in other parts of the world, including the native range of some endangered species. There is a remarkable dearth of educational materials about most endangered species within their native ranges, and any effort that zoos can make to redress this problem will be most helpful. Zoos may also need to develop curricula that more directly address the biodiversity crisis, and they may need to reach out more effectively to decision makers (politicians).

Zoos can serve as refuges for displaced, injured, or confiscated wildlife. Thus, zoos help wildlife authorities enforce regulations by providing homes for seized wildlife. In addition, many zoos maintain patches of artificial or natural habitats for native wildlife and can provide significant breeding refuges for some birds and butterflies, and for fish and small animals that need relatively little space for a viable population.

Zoos sponsor significant basic research, often studies that would be difficult or impossible to accomplish with animals in the wild. Furthermore, zoos invest a considerable amount of money in research; for example, accredited zoos in North America spent $1 million dollars on scientific research in 1998. While developing technology and methodology for captive animals under relatively controlled conditions, zoos create expertise that is useful in situ, for example, guidelines for medication dosages.

Direct financial support of field conservation is another way that zoos are helping preserve biodiversity. In 1998, accredited North American zoos were involved in more than 700 conservation projects in 80 countries, including providing grants to non-zoo researchers. Another form of in situ conservation funding is for zoos in the wealthier nations to help improve the conservation impact of zoos in lesser developed nations. Many zoos also sponsor member tours to see wildlife in nature, and they try to make arrangements that will have a positive impact on conservation.

C. The Ark Revisited

Human populations are still increasing and wildlife will continue to disappear. More direct conservation measures will be needed from all zoos. Zoos will be called upon to provide refuge for more species as the wild populations dwindle. Shrinking wild populations in postage-stamp parks may need to be managed more like zoo populations, with infusions of unrelated blood, potentially from zoo stocks. Intensive, joint, meta-population management of in situ and ex situ wild animals will require both scientific and political innovation. Zoos may indeed help to save some wildlife from the flood of humanity, but they probably will not be acting as closed vessels. Rather, expect to see zoo-based conservation continue to become increasingly dynamic and interactive.

See Also the Following Articles

BREEDING OF ANIMALS • CAPTIVE BREEDING AND REINTRODUCTION • CONSERVATION EFFORTS, CONTEMPORARY • EDUCATION AND BIODIVERSITY • ETHICAL ISSUES IN BIODIVERSITY PROTECTION • EX SITU, IN SITU CONSERVATION • NATURAL RESERVES AND PRESERVES

Bibliography


Web site resources:
- American Zoo and Aquarium Association: http://www.aza.org
- ISIS (International Species Inventory System): http://www.isis.org
- IUCN Captive Breeding Specialist Group: http://www.cbsg.org
- ZooLex (animal exhibit design): http://www.zoolex.org
Plate 1

Acid Rain. Depiction of global emissions of sulfur and nitrogen. Annual sulfur oxide emissions as sulfur on a $1^\circ \times 1^\circ$ latitude/longitude grid (1000 kg/year) (Canadian Global Emissions Interpretation Centre; from Voldner et al., 1994: Benkovitz et al., 1996). See article Acid Rain and Deposition.
Domesticated Plants. A diverse assemblage of fruits and flowers of wild and domesticated tomatoes. Such diversity was the basis for the development of tomato as a multi-use crop. It also will be the foundation for improvements in the crop in years to come. (Photo courtesy of North Central Regional Plant Introduction Station, Ames, Iowa.) See article Breeding of Plants.
Domesticated Plants. Some examples of the great variety of colors, sizes, and shapes of domesticated plants selected by Amazonian Indians: Center left: Peach palm (Bactris gasipaes Kunth). Bottom left: abiu (Pouteria caimito Radl.) var. Tikuna (Photo by Charles R. Clement.) Bottom right: cubiu (Solanum sessiliforum Dunäl). See article Crop Improvement.
Arachnids. Some examples of arachnid diversity. Scorpion (Scorpiones: Vaejovidae, Uroctonus sp.). See article Arachnids
Plate 5

Some examples of arachnid diversity. Microwhipscorpion (Palpigradi: Xxidae). See article Arachnids.
Archnids. Some examples of arachnid diversity. Whip spider or tailless whip scorpion (Amblypygi: Phrynidae, *Phrynus sp*.). See article Archnids.
Archnids. Some examples of arachnid diversity. Pseudoscorpion (Pseudoscorpiones: Xidae). See article Archnids.
Desertification. Depictions of the process in which other land types (e.g., semiarid croplands) undergo degradation to desert. Global distribution of hyperarid, arid, semiarid, dry subhumid, and humid areas. (UNEP, 1997.) See article Desertification.
Desertification. Depictions of the process in which other land types (e.g., semiarid croplands) undergo degradation to desert. Dynamics of acacia forests in the Serengeti grasslands are a function of complex interactions between ecological factors (e.g., the life span of the acacia trees and the number of wildebeest), human dimensions (cattle grazing and fire setting), and meteorological factors (precipitation patterns). See article Desertification.
Functional Diversity. Functional diversity, those components of biodiversity that influence how an ecosystem operates or functions, has been examined in a notable set of ongoing field experiments in Minnesota. The smaller of the Minnesota biodiversity experiments, has demonstrated that plant diversity has a strong effect on ecosystem productivity and nutrient dynamics. The experiment has 147 plots, each being 3 m x 3 m (about 10 feet by 10 feet) in size. See article Functional Diversity.
Functional Diversity. Functional diversity, those components of biodiversity that influence how an ecosystem operates or functions, has been examined in a notable set of ongoing field experiments in Minnesota. The larger of the Minnesota biodiversity experiments uses about 270 of the 342 plots visible here. It has shown strong effects of plant species richness, plant functional group richness, and plant functional group composition on ecosystem processes. Each plot is 13 m × 13 m (about 40 feet by 40 feet). See article Functional Diversity.
Fungi. Examples of fungal biodiversity. Members of the phylum Basidiomycota bear their sexual spores externally on a club-shaped structure called a basidium. Shown here, Members of the phylum Ascomycota bear their sexual spores (ascospores) in sacs called asci. Shown here is *Morchella esculenta*, the morel, a prime edible mushroom. See article Fungi.
Fungi. Examples of fungal biodiversity. Members of the phylum Basidiomycota bear their sexual spores externally on a club-shaped structure called a basidium. Shown here, The deuteromycetes (molds) have no known sexual state and reproduce only by producing spores via mitosis. Shown here, *Aspergillus conidia*. See article Fungi.
Fungi. Examples of fungal biodiversity. Members of the phylum Basidiomycota bear their sexual spores externally on a club-shaped structure called a basidium. Shown here, Formerly considered to be true fungi, the Myxomycota (slime molds) exit in nature as a plasmodium—a blob of protoplasm without cell walls, encomposing only by a cell membrane. See article Fungi.
Land Use Issues. Land use issues concern the process by which human activities determine land cover. Important issues in land use include agricultural development and intensification, human settlement, and extraction of natural resources. Shown here, fragmentation of the world’s forests at present time (bottom) compared to in preagricultural time (top). The original distribution of forests is estimated from climatic models. Current distributions are based on a variety of sources compiled from the 1980s to early 1990s. (From The World Conservation Monitoring Center, Cambridge, United Kingdom, www.wcmc.org.uk.) See article Land Use Issues.
Invertebrates, Freshwater. Since many freshwater invertebrates require specific habitats and flood regimes, their abundance and diversity in many parts of the world is at risk because habitats are being lost and natural flow regimes altered. Shown here, a schematic illustration of major freshwater habitats. Note that all water bodies, including lakes, streams and rivers, wetlands, and the oceans, are connected via the groundwaters. (Modified from Palmer [in press]). See article Invertebrates, Freshwater.
Isoptera (Termites). Termites are ecosystem engineers, constructing nests of various shapes and sizes that sometimes reach to 5 m in height. Shown here, nests of *Nasutitermes triodiae* in Australia. See article Isoptera.
**Insecticide Resistance.** The phenomenon of insecticide resistance reflects a genetic adaption enabling arthropod pests to survive exposure to otherwise lethal amounts of insecticide. Noted examples of this phenomenon include: Left: the bollworm, *Helicoverpa armigera*, a major pest of cotton and vegetables in the Old World. Center: the tobacco or cotton whitefly, *Bemisia tabaci* which threatens a wide range of crops through direct-feeding damage and the transmission of virus disease. Right: the two-spotted spider mite, *Tetranychus urticae*, a cosmopolitan pest of fruit, vegetable, and ornamental crops. See article Insecticide Resistance.
Introduced Plants. Examples of the effect of introduced plants. The Asian laminarian kelp (Undaria pinnatifida), native to Japan, Korea, and China, has been introduced and spread in other northern hemisphere waters and recently in the southern hemisphere as well, making it the first large kelp to cross tropical waters successfully. It is highly invasive, has significant negative impacts on the local native shallow marine fauna and flora, and may be impossible to eradicate. See article Introduced Plants.
Introduced Plants. Examples of the effect of introduced plants. *Salvinia molesta* (kariba weed, African pyle, Australian azolla, water fern, giant azoll), a free-floating perennial aquatic fern native to tropical and subtropical areas of South America, has recently been spread worldwide, becoming highly invasive in fresh waters and causing major disruptions to the utilization of important water resources. See article Introduced Plants.
Introduced Plants. Examples of the effect of introduced plants. A garden ornamental that has become an invasive weed is *Fallopia japonica* (Asiatic knotweed), a large perennial herb native to the Far East. It has infiltrated much of central and western Europe, North America, and several southern temperate countries, where it excludes native plant species in artificial and highly disrupted habitats, riparian areas, and open woodlands. In the United Kingdom, it is presently the most aggressive common herbaceous species. See article Introduced Plants.
Mangrove Ecosystems. Mangrove are a group of trees and shrubs, mostly evergreen, that often dominate large areas of tropical coastline or estuary. Aerial roots of *Rhizophora* one of the most widespread mangrove trees, in a Malaysian forest. See article Mangrove Ecosystems.
Mangrove Ecosystems. Mangrove are a group of trees and shrubs, mostly evergreen, that often dominate large areas of tropical coastline or estuary. Mangrove propagules on a *Rhizophora* tree, Indus Delta, Pakistan. See article Mangrove Ecosystems.
Mangrove Ecosystems. Mangrove are a group of trees and shrubs, mostly evergreen, that often dominate large areas of tropical coastline or estuary. Fiddler crabs are widespread throughout mangrove habitats of the Old and New World. (Photo courtesy of D. Barnes). See article Mangrove Ecosystems.
Origin of Life. A theoretical representation of the emergence of stages of life, from a prebiotic chemical state to cellular forms. See article Origin of Life, Theories of.
**Pelagic Ecosystems.** Phytoplankton species illustrating the taxonomic diversity as well as the shape and size of these organisms, which are an important component of the pelagic food web. (A) Diatomophyceae, *Skeletonema costatum*, a worldwide diatom species with distinctive features such as long chains of small cells with long external tubes. Diameter varies between 2 and 21 μm. (B) Dinophyceae, *Ceratium furca*, a worldwide dinoflagellate with a solitary or paired life-form, length between 70 and 200 μm and in width between 30 and 50 μm. (C) Dinophyceae, *Dinophysis norvegica*, a solitary dinoflagellate, normally found in cold waters, between 48 and 67 μm and width between 39 and 53 μm. It can be regarded as a toxic species since it produces a toxin that causes diarrhoeal shellfish poisoning. See article Pelagic Ecosystems.
Plate 34

Seagrasses. A mixed (seven species) seagrass meadow in Bolinao (Pangasinan, the Philippines). See article Seagrasses.
Seagrasses. *Halophila ovalis* population in a large intertidal flat visited by dugongs in Con Dao Island (Vietnam). See article Seagrasses.
Urban Ecology. A depiction of land-use change for the period 1792–1992, metropolitan region of Baltimore, Maryland, USA. The yellow areas indicate urbanization. See article Urban/Suburban Ecology.
Remote Sensing. A land cover map produced for Wyoming, using visual interpretation of Landsat data. (Reproduced from Driese et al., 1997, with permission from OPULUSPRESS.) See article Remote Sensing.
Remote Sensing. A probability map of buzzard nesting sites, based on vegetation variables derived from Landsat data. Observed nesting sites are indicated as B. Adapted from Austin et al. 1996, with permission from Blackwell Science Ltd. See article Remote Sensing.
Plate 40

Acid Rain. Depiction of global emissions of sulfur and nitrogen. Annual nitrogen oxide emissions as nitrogen on a $1^\circ \times 1^\circ$ latitude/longitude grid (1000 kg/year). (Canadian Global Emissions Interpretation Centre; from Voldner et al., 1994; Benkovitz et al., 1996). See article Acid Rain and Deposition.